Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore for additional books and resources.
## Contents

*What’s New In SAS LASR Analytic Server* .............................................. ix

**Chapter 1 • Introduction to the SAS LASR Analytic Server** .......................... 1
- What is SAS LASR Analytic Server? .................................................. 2
- How Does the SAS LASR Analytic Server Work? ............................. 3
- About the SAS High-Performance Deployment of Hadoop .................. 5
- Benefits of Using the Hadoop Distributed File System ...................... 5
- Components of the SAS LASR Analytic Server ................................. 6
- Administering the SAS LASR Analytic Server .................................. 7
- Memory Management ........................................................................... 12
- Data Partitioning and Ordering ....................................................... 14
- SAS LASR Analytic Server Logging .................................................. 15

**Chapter 2 • Non-Distributed SAS LASR Analytic Server** .......................... 19
- About Non-Distributed SAS LASR Analytic Server ......................... 19
- Starting and Stopping Non-Distributed Servers ............................... 19
- Loading and Unloading Tables for Non-Distributed Servers ............ 21

**Chapter 3 • LASR Procedure** ................................................................. 23
- Overview: LASR Procedure ............................................................... 23
- Syntax: LASR Procedure .................................................................. 24
- Examples: LASR Procedure ............................................................. 32

**Chapter 4 • IMSTAT Procedure (Analytics)** .............................................. 41
- Overview: IMSTAT Procedure (Analytics) ........................................ 42
- Syntax: IMSTAT Procedure (Analytics) ............................................ 42
- Examples: IMSTAT Procedure (Analytics) ...................................... 172

**Chapter 5 • IMSTAT Procedure (Data and Server Management)** ............... 187
- Overview: IMSTAT Procedure (Data and Server Management) ........ 188
- Concepts: IMSTAT Procedure .......................................................... 188
- Syntax: IMSTAT Procedure (Data and Server Management) ............ 194
- Examples: IMSTAT Procedure (Data and Server Management) ....... 235

**Chapter 6 • IMXFER Procedure** ............................................................... 247
- Overview: IMXFER Procedure .......................................................... 247
- Syntax: IMXFER Procedure .............................................................. 247
- Examples: IMXFER Procedure .......................................................... 251

**Chapter 7 • OLIPHANT Procedure** ......................................................... 253
- Overview: OLIPHANT Procedure ..................................................... 253
- Concepts: OLIPHANT Procedure .................................................... 254
- Syntax: OLIPHANT Procedure ......................................................... 255
- Examples: OLIPHANT Procedure .................................................... 258

**Chapter 8 • RECOMMEND Procedure** .................................................... 261
- Overview: RECOMMEND Procedure ............................................... 261
- Syntax: RECOMMEND Procedure .................................................... 262
- Examples: RECOMMEND Procedure ............................................... 276
Chapter 9 • VASMP Procedure .................................................. 287
  Overview: VASMP Procedure ........................................... 287
  Syntax: VASMP Procedure ............................................... 287
  Example: Copying Tables from One Hadoop Installation to Another 292

Chapter 10 • Using the SAS LASR Analytic Server Engine .......... 295
  Understanding How the SAS LASR Analytic Server Engine Works 295
  Understanding Server Tags ............................................. 296
  Comparing the SAS LASR Analytic Server Engine with the LASR Procedure 296
  What is Required to Use the SAS LASR Analytic Server Engine? 297
  What is Supported? ................................................... 297

Chapter 11 • LIBNAME Statement for the SAS LASR Analytic Server Engine .................................................. 299
  Dictionary ................................................................ 299

Chapter 12 • Data Set Options for the SAS LASR Analytic Server Engine ............................................................... 309
  Dictionary ................................................................ 309

Chapter 13 • Using the SAS Data in HDFS Engine .................. 315
  What Does the SAS Data in HDFS Engine Do? .................... 315
  Understanding How the SAS Data in HDFS Engine Works .... 315
  What is Required to Use the SAS Data in HDFS Engine? ...... 316
  What is Supported? ................................................... 316
  Common HDFS Commands ........................................... 316

Chapter 14 • LIBNAME Statement for the SAS Data in HDFS Engine ................................................................. 319
  Dictionary ................................................................ 319

Chapter 15 • Data Set Options for the SAS Data in HDFS Engine ................................................................. 325
  Dictionary ................................................................ 325

Chapter 16 • Programming with SAS LASR Analytic Server .... 335
  About Programming ..................................................... 335
  DATA Step Programming for Scoring In SAS LASR Analytic Server 335

Chapter 17 • Text Analytics in SAS LASR Analytic Server .... 341
  SAS Linguistic Files .................................................... 341
  Language Processing Concepts ....................................... 342
  Data Preparation ....................................................... 344
  Output Tables for the TEXTPARSE Statement .................... 344

Appendix 1 • SAS In-Memory Statistics for Hadoop ................. 355
  About SAS In-Memory Statistics for Hadoop ..................... 355
  Writing and Running SAS Programs ................................ 356
  Deploying SAS In-Memory Statistics for Hadoop ............... 359

Appendix 2 • Removing a Machine from the Cluster ................. 365
  About Removing a Machine from the Cluster .................... 365
  Which Servers Can I Leave Running When I Remove a Machine? 365
  Remove the Host Name from the grid.hosts File ............... 366
  Remove the Host Name from the Hadoop Slaves File .......... 366
  Restart Servers ....................................................... 366

Appendix 3 • Adding Machines to the Cluster ....................... 367
  About Adding Machines to the Cluster ........................... 367
Which Servers Can I Leave Running When I Add a Machine? .................. 368
Configure System Settings ............................................................ 368
Add Host Names to Gridhosts ...................................................... 369
Propagate Operating System User IDs .......................................... 370
Configure SAS High-Performance Deployment of Hadoop ................. 371
Configure SAS High-Performance Analytics Infrastructure .............. 374
Restart SAS LASR Analytic Server Monitor .................................. 375
Restart Servers and Redistribute HDFS Blocks .............................. 375
View Explorations and Reports ................................................... 375

Glossary .................................................................................. 377
Index ..................................................................................... 379
What’s New In SAS LASR Analytic Server

Overview

SAS LASR Analytic Server 2.3 includes the following changes:

- Analytic features for the IMSTAT procedure
- The RECOMMEND procedure and support for recommender applications
- Support for SAS In-Memory Statistics for Hadoop
- Support for in-memory text analytics
- Enhancements to data and server management for the IMSTAT procedure
- Documentation enhancements

Analytic Features for the IMSTAT Procedure

The IMSTAT procedure is enhanced with numerous statements that enable in-memory analytics. These statements are included in Chapter 4, “IMSTAT Procedure (Analytics),” on page 41. They are licensed separately from the data and server management statements.

The GLM, LOGISTIC, and GENMODEL statements are available for performing a variety of modeling techniques. Each of these statements have the following features:

- You can specify the CODE option so that the server generates and saves SAS scoring code.
- You can assign a role variable to divide the data into training and validation sets. Alternatively, you can request that the server divides the data at random into these sets by specifying the proportion for validation data. See the ROLEVAR=, SEED=, and VALIDATE= options in these statements.
- You can apply group-by filtering by passing the name of a temporary table generated by the GROUPBY statement and you can define rules for extracting groups from the group-by set. This enables you to restrict model fitting to groups of interest. For example, groups in which the average of some measure exceeds a particular value, or fitting models across groups in stages to prevent overwhelming the SAS session.
- You can specify the INFORMATIVE option to account for missing values in the data with an informative missing algorithm. The option adds additional dummy effects to the model and replaces missing values with the effect mean. The
coefficients for the dummy effects estimate the difference between the response predicted by the missing value grouping and the predicted response if the effect is evaluated at its mean. This enables you to use all the data in estimating and scoring a model, without additional imputation steps.

The RANDOMWOODS statement generates a collection of decision trees. Each tree is built from a bootstrap sample of the data and each tree is based on a random selection of variables. The collection of trees can be used generally as classification and regression trees. The current implementation is for use in classification problems. You can also specify a CODE option to generate and save SAS scoring code.

The ASSESS statement is used to assess the quality of one or more statistical models. For a set of classification models, you can compute model lift, receiver-operating characteristic (ROC), and concordance statistics. You can also apply the ASSESS statement to regression-type models.

---

**RECOMMEND Procedure**

The RECOMMEND procedure is a new procedure that enables you develop a recommender system. A common goal for a recommender system is to make personalized recommendations to individuals who lack the capacity, experience, or resources to select from a potentially overwhelming list of choices.

The procedure enables building content-based recommender systems and collaborative filtering systems.

---

**SAS In-Memory Statistics for Hadoop**

SAS LASR Analytic Server and the analytic statements for the IMSTAT procedure provide the core features that are available with SAS In-Memory Statistics for Hadoop. SAS/ACCESS Interface for Hadoop is included and enables you to access your data that is stored in HDFS in a variety of formats.

The bundle includes other products, such as SAS Studio and SAS/STAT. This document covers the features available in the server and the IMSTAT procedure.

---

**Support for Text Analytics**

SAS LASR Analytic Server can work on unstructured text, such as social media feeds, news articles, and arbitrary collection of documents. The server uses text analytics to turn the unstructured text into numbers and counts and weights, and terms and topics with relationships, suitable for visualization and exploration. For more information, see the “TEXTPARSE Statement” on page 164 and “Text Analytics in SAS LASR Analytic Server” on page 341.
Enhancements to Data and Server Management for the IMSTAT Procedure

The statements that were introduced in previous releases are included in Chapter 5, “IMSTAT Procedure (Data and Server Management),” on page 187. Some of the enhancements are as follows:

- A BATCHMODE option is added to the procedure. When the option is enabled and an error occurs, the procedure terminates and sets the SYSERR macro variable.

- The SCORE statement is enhanced as follows:
  - You can specify the names of additional in-memory tables that contain information about key-value pairs for DATA step hash objects. For more information, see the HASHDATA option on page 214.
  - The DSRETAI N option is added. When the option is specified, scoring code behaves like the DATA step with respect to retention of output symbols.

- The STORE statement is enhanced to support complex expressions. For example, you can build a string for a WHERE clause from the contents of IMSTAT procedure result tables.

- The FETCH statement is enhanced as follows:
  - You can specify formats for the variables to retrieve.
  - You can specify an ORDERBY= option to retrieve rows from the server that are sorted.
  - You can specify format instructions for ORDERBY= variables and you can choose whether each variable is sorted by unformatted or formatted values. You can also specify whether the sort order is ascending or descending.

The TABLEINFO statement is enhanced with the PARTVARS option. This option enables displaying the names of the partition variables and order-by variables for tables.

Documentation Enhancements

Information about working with HDFS has been added. See “Common HDFS Commands” on page 316.
Chapter 1
Introduction to the SAS LASR Analytic Server

What is SAS LASR Analytic Server? ............................................. 2
How Does the SAS LASR Analytic Server Work? ....................... 3
  Distributed SAS LASR Analytic Server .................................. 3
  Non-Distributed SAS LASR Analytic Server .............................. 4
About the SAS High-Performance Deployment of Hadoop ................. 5
Benefits of Using the Hadoop Distributed File System .................. 5
Components of the SAS LASR Analytic Server............................ 6
  About the Components ....................................................... 6
  Root Node ......................................................................... 6
  Worker Nodes ..................................................................... 6
  In-Memory Tables ............................................................... 6
  Signature Files ..................................................................... 6
  Server Description Files ....................................................... 7
Administering the SAS LASR Analytic Server ............................. 7
  Administering a Distributed Server ......................................... 7
  Administering a Non-Distributed Server .................................. 7
  Common Administration Features ......................................... 8
  Features Available in SAS Visual Analytics Administrator ............ 8
  Understanding Server Run Time ............................................. 8
  Distributing Data ................................................................... 9
Memory Management ............................................................... 12
  About Physical and Virtual Memory ...................................... 12
  How Does the Server Use Memory for Tables? ......................... 12
  How Else Does the Server Use Memory? .................................. 13
  Managing Memory ............................................................... 13
Data Partitioning and Ordering ............................................... 14
  Overview of Partitioning ....................................................... 14
  Understanding Partition Keys ............................................... 14
  Ordering within Partitions ..................................................... 15
SAS LASR Analytic Server Logging ......................................... 15
  Understanding Logging ....................................................... 15
  What is Logged? ............................................................... 16
  Log Record Format ............................................................ 16
What is SAS LASR Analytic Server?

The SAS LASR Analytic Server is an analytic platform that provides a secure, multi-user environment for concurrent access to data that is loaded into memory. The server can take advantage of a distributed computing environment by distributing data and the workload among multiple machines and performing massively parallel processing. The server can also be deployed on a single machine where the workload and data volumes do not demand a distributed computing environment.

The server handles both big data and smaller sets of data, and it is designed with a high-performance, multi-threaded, analytic code. The server processes client requests at extraordinarily high speeds due to the combination of hardware and software that is designed for rapid access to tables in memory. By loading tables into memory for analytic processing, the server enables business analysts to explore data and discover relationships in data at the speed of RAM.

The server can also perform text analysis on unstructured data. The unstructured data is loaded to memory in the form of a table, with one document in each row. The TEXTPARSE statement in the IMSTAT procedure can then provide similar analysis to what is available with the HPTMINE procedure.

Another use for the analytic platform that the server provides is to create a recommender system. Creating recommender systems introduces the concept of an application in the server. The recommender system contains the application and might contain four or five tables. Each of the tables can be used in different ways, depending on the task and which method you apply. For example, making an item-based prediction for a nearest-neighbor method requires different data structures than a singular-value decomposition. You can associate a particular method or a set of methods with the application. You can execute one method or an ensemble. The flexibility provided by the server enables you to add and drop methods from the application. As a modeler, you want to explore and evaluate with different methods and different parameter configurations for the methods until you have optimized the system for your purposes. Then, you can deploy the recommender system in an online scoring environment.

The architecture for the server was originally designed for optimal performance in a distributed computing environment. A distributed server runs on multiple machines. A typical distributed configuration is to use a series of blades as a cluster. Each blade contains both local storage and large amounts of memory. Local storage is used to store large data sets in distributed form. Data is loaded into memory and made available so that clients can quickly access that data.

For distributed deployments, having local storage available on machines is critical in order to store large data sets in a distributed form. The server supports the Hadoop Distributed File System (HDFS) as a co-located data provider. HDFS is used because the server can read from and write to HDFS in parallel. In addition, HDFS provides replication for data redundancy. HDFS stores data as blocks in distributed form on the blades and the replication provides failover capabilities.

In a distributed deployment, the server also supports some third-party vendor databases as co-located data providers. Teradata Data Warehouse Appliance and Greenplum Data Computing Appliance are massively parallel processing database appliances. You can install the SAS LASR Analytic Server software on each of the machines in either appliance. The server can read in parallel from the local data on each machine.

For the SAS LASR Analytic Server 1.6 release (concurrent with the SAS Visual Analytics 6.1 release) the server supports a non-distributed deployment. A non-
How Does the SAS LASR Analytic Server Work?

Distributed SAS LASR Analytic Server

The server provides a client/server environment where the client connects to the server, sends requests to the server, and receives results back from the server. The server-side environment is a distributed computing environment. A typical deployment is to use a series of blades in a cluster. In addition to using a homogeneous hardware profile, the software installation is also homogeneous. The same operating system is used throughout and the same SAS software is installed on each blade that is used for the server. In order for the software on each blade to share the workload and still act as a single server, the SAS software that is installed on each blade implements the Message Passing Interface (MPI). The MPI implementation is used to enable communication between the blades.

After a client connection is authenticated, the server performs the operations requested by the client. Any request (for example, a request for summary statistics) that is authorized will execute. After the server completes the request, there is no trace of the request. Every client request is executed in parallel at extraordinarily high speeds, and client communication with the server is practically instantaneous and seamless.

There are two ways to load data into a distributed server:

- **load data from tables and data sets.** You can start a server instance and directly load tables into the server by using the SAS LASR Analytic Server engine or the LASR procedure from a SAS session that has a network connection to the cluster. Any data source that can be accessed with a SAS engine can be loaded into memory. The data is transferred to the root node of the server and the root node distributes the data to the worker nodes. You can also append rows to an in-memory table with the SAS LASR Analytic Server engine.

- **load tables from a co-located data provider.**
  - Tables can be read from the Hadoop Distributed File System (HDFS). You can use the SAS Data in HDFS engine to add tables to HDFS. When a table is added to HDFS, it is divided into blocks that are distributed across the machines in the cluster. The server software is designed to read data in parallel from HDFS. When used to read data from HDFS, the LASR procedure causes the worker nodes to read the blocks of data that are local to the machine.
  - Tables can also be read from a third-party vendor database. For distributed databases like Teradata and Greenplum, the SAS LASR Analytic Server can access the local data on each machine that is used for the database.

The following figure shows the relationship of the root node, the worker nodes, and how they interact when working with large data sets in HDFS. As described in the previous list, the LASR procedure communicates with the root node and the root node directs the
worker nodes to read data in parallel from HDFS. The figure also indicates how the SAS Data in HDFS engine is used to transfer data to HDFS.

**Figure 1.1 Relationship of PROC LASR and the SAS Data in HDFS Engine**

![Diagram of PROC LASR and SAS Data in HDFS Engine](image)

*Note:* The preceding figure shows a distributed architecture that uses HDFS. For deployments that use a third-party vendor database, the architecture is also distributed, but different procedures and software components are used for distributing and reading the data.

After the data is loaded into memory on the server, it resides in memory until the table is unloaded or the server terminates. After the table is in memory, client applications that are authorized to access the table can send requests to the server and receive the results from the server.

In-memory tables can be saved. You can use the SAS LASR Analytic Server engine to save an in-memory table as a SAS data set or as any other output that a SAS engine can use. This method of using an engine transfers the data across the network connection. For large tables, saving to HDFS is supported with the LASR and IMSTAT procedures. This strategy saves the data in parallel and keeps the data on the cluster.

**Non-Distributed SAS LASR Analytic Server**

Most of the features that are available with a distributed deployment also apply to the non-distributed deployment too. Any limitations are related to the reduced functionality of using a single-machine rather than a distributed computing environment.

In a non-distributed deployment, the server acts in a client/server fashion where the client sends requests to the server and receives results back. The server performs the analytic operations on the tables that are loaded in to memory. As a result, the processing times are very fast and the results are delivered almost instantaneously.

You can load tables to a non-distributed server with the SAS LASR Analytic Server engine. Any data source that SAS can access can be used for input and the SAS LASR Analytic Server engine can store the data as an in-memory table. The engine also supports appending data.

You can save in-memory tables by using the SAS LASR Analytic Server engine. The tables can be saved as a SAS data set or as any other output that a SAS engine can use.
About the SAS High-Performance Deployment of Hadoop

SAS offers the SAS High-Performance Deployment of Hadoop that includes the following:

• Hadoop software that is provided by Apache
• two JAR files that provide services that run inside Hadoop
• an executable file, saslasrfd, that facilitates reading data in parallel into SAS LASR Analytic Server
• an installation program that simplifies installation and configuration on the cluster

The JAR files and the executable are installed automatically when you install SAS High-Performance Deployment of Hadoop. As an alternative, you can manually configure several commercially available Hadoop distributions with the JAR files and executable. After you have performed the manual configuration, the commercially available distribution is functionally equivalent to SAS High-Performance Deployment of Hadoop. The cluster then supports parallel I/O with SAS LASR Analytic Server as well as operating with the SAS Data in HDFS engine.

See Also

SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide

Benefits of Using the Hadoop Distributed File System

Loading data from disk to memory is efficient when the SAS LASR Analytic Server is co-located with a distributed data provider. The Hadoop Distributed File System (HDFS) provided by SAS High-Performance Deployment of Hadoop acts as a co-located data provider. HDFS offers some key benefits:

• **Parallel I/O.** The SAS LASR Analytic Server can read data in parallel at very impressive rates from a co-located data provider.

• **Data redundancy.** By default, two copies of the data are stored in HDFS. If a machine in the cluster becomes unavailable or fails, the SAS LASR Analytic Server instance on another machine in the cluster retrieves the data from a redundant block and loads the data into memory.

• **Homogeneous block distribution.** HDFS stores files in blocks. The SAS implementation enables a homogeneous block distribution that results in balanced memory utilization across the SAS LASR Analytic Server and reduces execution time.
Components of the SAS LASR Analytic Server

About the Components

The following sections identify some software components and interactions for SAS LASR Analytic Server.

Root Node

When the SAS client initiates contact with the grid host to start a SAS LASR Analytic Server instance, the SAS software on that machine takes on the role of distributing and coordinating the workload. This role is in contrast to a worker node. This term applies to a distributed SAS LASR Analytic Server only.

Worker Nodes

This is the role of the software that receives the workload from the root node. When a table is loaded into memory, the root node distributes the data to the worker nodes and they load the data into memory. If you are using a co-located data provider, each worker node reads the portion of the data that is local to the machine. The data is loaded into memory and requests that are sent to root node are distributed to the worker nodes. The worker nodes perform the analytic tasks on the data that is loaded in memory on the machine and then return the results to the root node. This term applies to a distributed SAS LASR Analytic Server only.

In-Memory Tables

SAS LASR Analytic Server performs analytics on tables that are in-memory only. Typically, large tables are read from a co-located data provider by worker nodes. The tables are loaded quickly because each worker node is able read a portion of the data from local storage. Once the portion of the table is in memory on each worker node, the server instance is able to perform the analytic operations that are requested by the client. The analytic tasks that are performed by the worker nodes are done on the in-memory data only.

Signature Files

SAS LASR Analytic Server uses two types of signature files, server signature files and table signature files. These files are used as a security mechanism for server management and for access to data in a server. When a server instance is started, a directory is specified on the PATH= option to the LASR procedure. The specified directory must exist on the machine that is specified as GRIDHOST= environment variable.

In order to start a server, the user must have Write access to the directory in order to be able to create the server signature file. In order stop a server, the user must have Read access to the server signature file so that it can be removed from the directory.

In order to load and unload tables on a server, the user must have Read access to the server signature file in order to interact with the server. Write permission to the directory
is needed to create the table signature file when loading a table and to delete the table signature file when unloading the table.

Server Description Files

Note: Most administrators prefer to use the PORT= option in the LASR procedure rather than use server description files.

If you specify a filename in the CREATE= option in the LASR procedure, then you start a SAS LASR Analytic Server instance, the LASR procedure creates two files:

- a server description file
- a server signature file (described in the previous section)

The server description file contains information such as the host names of the machines that are used by the server instance and signature file information.

In the LASR procedure, the server description file is specified with the CREATE= option. The server description file is created on the SAS client machine that invoked PROC LASR.

Administering the SAS LASR Analytic Server

Administering a Distributed Server

Basic administration of a distributed SAS LASR Analytic Server can be performed with the LASR procedure from a SAS session. Server instances are started and stopped with the LASR procedure. The LASR procedure can be used to load and unload tables from memory though the SAS LASR Analytic Server engine also provides that ability.

The SAS Data in HDFS engine is used to add and delete tables from the Hadoop Distributed File System (HDFS). The tables are stored in the SASHDAT file format. You can use the DATASETS procedure with the engine to display information about tables that are stored in HDFS.

The HPDS2 procedure has a specific purpose for use with SAS LASR Analytic Server. In this deployment, the procedure is used to distribute data to the machines in an appliance. After the data are distributed, the SAS LASR Analytic Server can read the data in parallel from each of the machines in the appliance.

Administering a Non-Distributed Server

A non-distributed SAS LASR Analytic Server runs on a single machine. A non-distributed server is started and stopped with the SAS LASR Analytic Server engine. A server is started with the STARTSERVER= option in the LIBNAME statement. The server is stopped when one of the following occurs:

- The libref is cleared (for example, libname lasrvr clear;).
- The SAS program and session that started the server ends. You can use the SERVERWAIT statement in the VASMP procedure to keep the SAS program (and the server) running.
- The server receives a termination request from the SERVERTERM statement in the VASMP procedure.
A non-distributed deployment does not include a distributed computing environment. As a result, a non-distributed server does not support a co-located data provider. Tables are loaded and unloaded from memory with the SAS LASR Analytic Server engine only.

**Common Administration Features**

As described in the previous sections, the different architecture for distributed and non-distributed servers requires different methods for starting, stopping, and managing tables with servers. However, the IMSTAT procedure works with distributed and non-distributed servers to provide administrators with information about server instances. The statements that provide information that can be of interest to administrators are as follows:

- SERVERINFO
- TABLEINFO

Administrators might also be interested in the SERVERPARM statement. You can use this statement to adjust the number of requests that are processed concurrently. You might reduce the number of concurrent requests if the number of concurrent users causes the server to consume too many sockets from the operating system.

**Features Available in SAS Visual Analytics Administrator**

SAS LASR Analytic Server is an important part of SAS Visual Analytics. SAS Visual Analytics Administrator is a web application that provides an intuitive graphical interface for server management. You can use the application to start and stop server instances, as well as load and unload tables from the servers. Once a server is started, you can view information about libraries and tables that are associated with the server. The application also indicates whether a table is in-memory or whether it is unloaded.

For deployments that use SAS High-Performance Deployment of Hadoop, an HDFS explorer enables you to browse the tables that are stored in HDFS. Once tables are stored in HDFS, you can load them into memory in a server instance. Because SAS uses the special SASHDAT file format for the data that is stored in HDFS, the HDFS explorer also provides information about the columns, row count, and block distribution.

**Understanding Server Run Time**

By default, servers are started and run indefinitely. However, in order to conserve the hardware resources in a distributed computing environment, server instances can be configured to exit after a period of inactivity. This feature applies to distributed SAS LASR Analytic Server deployments only. You specify the inactivity duration with the LIFETIME= option when you start the server.

When the LIFETIME= option is used, each time a server is accessed, such as to view data or perform an analysis, the run time for the server is reset to zero. Each second that a server is unused, the run timer increments to count the number of inactive seconds. If the run timer reaches the maximum run time, the server exits. All the previously used hardware resources become available to the remaining server instances.
Distributing Data

**SAS High-Performance Deployment of Hadoop**
SAS provides SAS High-Performance Deployment of Hadoop as a co-located data provider. The SAS LASR Analytic Server software and the SAS High-Performance Deployment of Hadoop software are installed on the same blades in the cluster. The SAS Data in HDFS engine can be used to distribute data to HDFS.

For more information, see “Using the SAS Data in HDFS Engine” on page 315.

**PROC HPDS2 for Big Data**
For deployments that use Greenplum or Teradata, the HPDS2 procedure can be used to distribute large data sets to the machines in the appliance. The procedure provides an easy-to-use and efficient method for transferring large data sets.

For deployments that use Greenplum, the procedure is more efficient than using a DATA step with the SAS/ACCESS Interface to Greenplum and is an alternative to using the gpfdist utility.

The SAS/ACCESS Interface for the database must be configured on the client machine. It is important to distribute the data as evenly as possible so that the SAS LASR Analytic Server has an even workload when the data is read into memory.

The following code sample shows a LIBNAME statement and an example of the HPDS2 procedure for adding tables to Greenplum.

```
libname source */data/marketing/2012;

libname target greenplm
    server = "grid001.example.com"
    user = dbuser
    password = dbpass
    schema = public
    database = template1
    dbcommit=1000000;

proc hpds2 data = source.mktdata
    out = target.mktdata (distributed_by = 'distributed randomly'); 1
    performance host = "grid001.example.com"
    install = "/opt/TKGrid";

    data DS2GTF.out;
    method run();
    set DS2GTF.in;
    end;
    enddata;
    run;

proc hpds2 data = source.mkdata2
    out = target.mkdata2 (dbtype=(id='int')
        distributed_by='distributed by (id)'); 2
    performance host = "grid001.example.com"
    install = "/opt/TKGrid";
```
The rows of data from the input data set are distributed randomly to Greenplum.

The id column in the input data set is identified as being an integer data type. The rows of data are distributed based on the value of the id column.


**Bulkload for Teradata**

The SAS/ACCESS Interface to Teradata supports a bulk loading feature. With this feature, a DATA step is as efficient at transferring data as the HPDS2 procedure.

The following code sample shows a LIBNAME statement and two DATA steps for adding tables to Teradata.

```
libname tdlib teradata
   server="dbc.example.com"
   database=hps
   user=dbuser
   password=dbpass
   bulkload=yes;

data tdlib.order_fact;
   set work.order_fact;
run;

data tdlib.product_dim (dbtype=(partno='int')
   dbcreate_table_opts='primary index(partno)');
   set work.product_dim;
run;

data tdlib.salecode(dbtype=(_day='int' fpop='varchar(2)')
   bulkload=yes
   dbcreate_table_opts='primary index(_day,fpop)');
   set work.salecode;
run;

data tdlib.automation(bulkload=yes
   dbcommit=1000000
   dbcreate_table_opts='unique primary index(obsnum)');
   set automation;
   obsnum = _n_
run;
```

1 Specify the BULKLOAD=YES option. This option is shown as a LIBNAME option but you can specify it as a data set option.

2 Specify a data type of int for the variable named partno.
Specify to use the variable named `partno` as the distribution key for the table.

Specify to use the variables that are named `_day` and `fpop` as a distribution key for the table that is named `salecode`.

Specify the `DBCOMMIT=` option when you are loading many rows. This option interacts with the `BULKLOAD=` option to perform checkpointing. Checkpointing provides known synchronization points if a failure occurs during the loading process.

Specify the `UNIQUE` keyword in the table options to indicate that the primary key is unique. This keyword can improve table loading performance.

**Smaller Data Sets**

You can use a DATA step to add smaller data sets to Greenplum or Teradata. Transferring small data sets does not need to be especially efficient. The SAS/ACCESS Interface for the database must be configured on the client machine.

The following code sample shows a LIBNAME statement and DATA steps for adding tables to Greenplum.

```sas
libname gplib greenplm server="grid001.example.com"
  database=hps
  schema=public
  user=dbuser
  password=dbpass;

data gplib.automation(distributed_by='distributed randomly'); 1
  set work.automation;
run;

data gplib.results(dbtype=(rep='int') 2
  distributed_by='distributed by (rep)'); 3
  set work.results;
run;

data gplib.salecode(dbtype=(day='int' fpop='varchar(2)') 4
  distributed_by='distributed by day,fpop'); 5
  set work.salecode;
run;
```

1. Specify a random distribution of the data. This data set option is for the SAS/ACCESS Interface to Greenplum.
2. Specify a data type of `int` for the variable named `rep`.
3. Specify to use the variable named `rep` as the distribution key for the table that is named `results`.
4. Specify a data type of `int` for the variable named `day` and a data type of `varchar(2)` for the variable named `fpop`.
5. Specify to use the combination of variables `day` and `fpop` as the distribution key for the table that is named `salecode`.

The following code sample shows a LIBNAME statement and a DATA step for adding a table to Teradata.

```sas
libname tdlib teradata server="dbc.example.com"
  database=hps
  user=dbuser
  password=dbpass;
```
data tdlib.parts_dim;
  set work.parts_dim;
run;

For Teradata, the SAS statements are very similar to the syntax for bulk loading. For more information, see “Bulkload for Teradata” on page 10.

See Also

SAS/ACCESS for Relational Databases: Reference

Memory Management

About Physical and Virtual Memory

The amount of memory on a machine is the physical memory. The amount of memory that can be used by an application can be larger, because the operating system can provide virtual memory. Virtual memory makes the machine appear to have more memory available than there actually is, by sharing physical memory between applications when they need it and by using disk space as memory.

When memory is not used and other applications need to allocate memory, the operating system pages out the memory that is not currently needed to support the other applications. When the paged-out memory is needed again, some other memory needs to be paged out. Paging means to write some of the contents of memory onto a disk.

Paging does affect performance, but some amount of paging is acceptable. Using virtual memory enables you to access tables that exceed the amount of physical memory on the machine. So long as the time to write pages to the disk and read them from the disk is short, the server performance is good.

One advantage of SASHDAT tables that are read from HDFS is that the server performs the most efficient paging of memory.

How Does the Server Use Memory for Tables?

When you load a table to memory with the SAS LASR Analytic Server engine, the server allocates physical memory to store the rows of data. This applies to both distributed and non-distributed servers.

When a distributed server loads a table from HDFS to memory with the LASR procedure, the server defers reading the rows of data into physical memory. You can direct the server to perform an aggressive memory allocation scheme at load time with the READAHEAD option for the PROC LASR statement.

Note: When a distributed server loads a table from either the Greenplum Data Computing Appliance or the Teradata Data Warehouse Appliance, physical memory is allocated for the rows of data. This is true even when the data provider is co-located.
How Else Does the Server Use Memory?

Physical memory is used when the server performs analytic operations such as summarizing a table. The amount of memory that a particular operation requires typically depends on the cardinality of the data. In most cases, the cardinality of the data is not known until the analysis is requested. When the server performs in-memory analytics, the following characteristics affect the amount of physical memory that is used:

• Operations that use group-by variables can use more memory than operations that do not. The amount of memory that is required is not known without knowing the number of group-by variable combinations that are in the data.
• The memory utilization pattern on the worker nodes can change drastically depending on the distribution of the data across the worker nodes. The distribution of the data affects the size of intermediate result sets that are merged across the network.

Some requests, especially with high-cardinality variables, can generate large result sets. To enable interactive near-real-time work with high cardinality problems, the server allocates memory for data structures that speed performance. The following list identifies some of these uses:

• The performance for traversing and querying a decision tree is best when the tree is stored in the server.
• Paging through group-by results when you have a million groups is best done by storing the group-by structure in a temporary table in the server. The temporary table is then used to look up groups for the next page of results to deliver to the client.

Managing Memory

The following list identifies some of the options that SAS provides for managing memory:

• You can use the TABLEMEM= option to specify a threshold for physical memory utilization.
• You can use the EXTERNALMEM= option to specify a threshold for memory utilization for SAS High-Performance Analytics procedures.

By default, whenever the amount of physical memory in use rises above 75% of the total memory available on a node of a distributed server, adding tables (including temporary ones), appending rows, or any other operation that consumes memory for storing data fails.

If the machine has already crossed the threshold, your requests to add data are immediately rejected. If you attempt to add a table and the server crosses the threshold as the data is added, the server removes the table you attempted to add and frees the memory. Similarly, if you attempt to append rows and the server crosses the threshold during the request, the entire append request fails. The table remains as it was before the append was attempted.

You can specify the threshold when you start a server with the TABLEMEM= option in the PROC LASR statement or alter it for a running server with the SERVERPARM statement in the VASMP procedure. By default, TABLEMEM=75 (%).

Note: The memory that is consumed by tables loaded from HDFS do not count toward the TABLEMEM= limit.
Be aware that the TABLEMEM= option does not specify the percentage of memory that can be filled with tables. The memory consumption is measured across all processes of a machine.

A separate memory setting can be applied to processes that extract data from a server on a worker node. SAS High-Performance Analytics procedures can do this. If you set the EXTERNALMEM= option in the PROC LASR statement or through the SERVERPARAM statement in the VASMP procedure, then you are specifying the threshold of total memory (expressed as a percentage) at which the server stops sending data to the high-performance analytics procedure.

**See Also**
- “TABLEMEM=pct” on page 28
- “EXTERNALMEM=pct” on page 25

---

## Data Partitioning and Ordering

### Overview of Partitioning

By default, partitioning is not used and data are distributed in a round-robin algorithm. This applies to SAS Data in HDFS engine as well as SAS LASR Analytic Server. In general, this works well so that each machine in a distributed server has an even workload.

However, there are some data access patterns that can take advantage of partitioning. When a table is partitioned in a distributed server, all of the rows that match the partition key are on a single machine. If the data access pattern matches the partitioning (for example, analyzing data by Customer_ID partitioning the data by Customer_ID), then the server can direct the work to just the one machine. This can speed up analytic processing because the server knows where the data are.

However, if the data access pattern does not match the partitioning, processing times might slow. This might be due to the uneven distribution of data that can cause the server to wait on the most heavily loaded machine.

*Note:* You can partition tables in non-distributed SAS LASR Analytic Server deployments. However, all the partitions are kept on the single machine because there is no distributed computing environment.

### Understanding Partition Keys

Partition keys in SASHDAT files and in-memory tables are constructed based on the formatted values of the partition variables. The formatted values are derived using internationalization and localization rules. (All formatted values in the server follow the internationalization and localization rules.)

All observations that compare equal in the (concatenated) formatted key belong to the same partition. This enables you to partition based on numeric variables. For example, you can partition based on binning formats or date and time variables use date and time formats.

A multi-variable partition still has a single value for the key. If you partition according to three variables, the server constructs a single character key based on the three
variables. The formatted values of the three variables appear in the order in which the variables were specified in the PARTITION= data set option. For example, partitioning a table by the character variable REGION and the numeric variable DATE, where DATE is formatted with a MONNAME3. format:

```sas
data hdfslib.sales(partition=(region date) replace=yes);
  format date monname3.;
  set work.sales;
run;
```

The partition keys might resemble EastJan, NorthJan, NorthFeb, WestMar, and so on. It is important to remember that partition keys are created only for the variable combinations that occur in the data. It is also important to understand that the partition key is not a sorting of Date (formatted as MONNAME3.) within Region. For information about ordering, see “Ordering within Partitions” on page 15.

If the formats for the partition keys are user-defined, they are transferred to the LASR Analytic Server when the table is loaded to memory. Be aware that if you use user-defined formats to partition a SASHDAT file, the definition of the user-defined format is not stored in the SASHDAT file. Only the name of the user-defined format is stored in the SASHDAT file. When you load the SASHDAT file to a server, you need to provide the XML definition of the user-defined format to the server. You can do this with the FMTLIBXML= option to the LASR procedure at server start-up or with the PROC LASR ADD request.

**Ordering within Partitions**

Ordering of records within a partition is implemented in the SAS Data in HDFS engine and the SAS LASR Analytic Server. You can order within a partition by one or more variables and the organization is hierarchical—that is ordering by A and B implies that the levels of A vary slower than those of B (B is ordered within A).

Ordering requires partitioning. The sort order of character variables uses national language collation and is sensitive to locale. The ordering is based on the raw values of the order-by variables. This is in contrast to the formation of partition keys, which is based on formatted values.

When a table that is partitioned and ordered in HDFS is loaded into memory on the server, the partitioning and ordering is maintained. You can append to in-memory tables that are partitioned and ordered. However, this does require a re-ordering of the observations after the observations are transferred to the server.

---

**SAS LASR Analytic Server Logging**

**Understanding Logging**

Logging is an optional feature that can be enabled when a server instance is started with the LASR procedure. In order to conserve disk space, the default behavior for the server is to delete log files when the server exits. You can override this behavior with the KEEPLOG suboption to the LOGGING option when you start the server. You can also override this behavior with a suboption to the STOP option when you stop the server.

The server writes logs files on the grid host machine. The default directory for log files is `/tmp`. You can specify a different directory in the LOGGING option when you start
the server instance. The log filename is the same as server signature file with a .log suffix (for example, LASR.924998214.28622.saslasr.log).

**See Also**

- LOGGING option for the LASR procedure on page 26
- “Example 2: Starting a Server with Logging Options” on page 33
- “Starting and Stopping Non-Distributed Servers” on page 19

**What is Logged?**

When a server is started with the LOGGING option, the server opens the log file immediately, but does not generate a log record to indicate that the server started. As clients like SAS Visual Analytics Explorer make requests to the server for data, the server writes a log record.

The server writes a log record when a request is received and completed by the server. The server does not write log records for activities that do not contact the server (for example, ending the SAS session).

A user that is configured with passwordless SSH to access the machines in the cluster, but who is not authorized to use a server instance is denied access. The denial is logged with the message **You do not have sufficient authorization to add tables to this LASR Analytic Server**. However, if a user is not configured correctly to access the machines in the cluster, communication with the server is prevented by the operating system. The request does not reach the server. In this second case, the server does not write a log record because the server does not receive the request.

**Log Record Format**

The following file content shows an example of three log records. Line breaks are added for readability. Each record is written on a single line and fields are separated by commas. Each field is a name-value pair.

```plaintext
File 1.1 Sample Log File Records

ID=1,PID=28622,SASTime=1656782485.36,Time=Tue Jul 24 20:54:45 2012,User=sasdemo, Host=grid001,LASRServer=/tmp/LASR.924998214.28622.saslasr,Port=56925, RawCmd=action=ClassLevels name=DEPT.GRP1.PRDSALE "NlsenCoding=62", ExeCmd=action=ClassLevels name=DEPT.GRP1.PRDSALE "NlsenCoding=62",JnlMsg=, StatusMsg=Command successfully completed.,RunTime= 2.17

ID=2,PID=28622,SASTime=1656782593.09,Time=Tue Jul 24 20:56:33 2012,User=sasdemo, Host=grid001,LASRServer=/tmp/LASR.924998214.28622.saslasr,Port=56925, RawCmd=action=BoxPlot name=DEPT.GRP1.PRDSALE, ExeCmd=action=BoxPlot name=DEPT.GRP1.PRDSALE,JnlMsg=, StatusMsg=Command successfully completed.,RunTime= 0.12

ID=3,PID=28622,SASTime=16568825361.76,Time=Wed Jul 25 08:49:21 2012,User=sasdemo, Host=grid001,LASRServer=/tmp/LASR.924998214.28622.saslasr,Port=56925, RawCmd=action=APPEND_TABLE ,ExeCmd=action=APPEND_TABLE ,JnlMsg=, StatusMsg=Command successfully completed.,RunTime= 0.09
```
Table 1.1  Log Record Fields

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>specifies a unique identifier for the action.</td>
</tr>
<tr>
<td>PID</td>
<td>specifies the operating system process identifier for the server.</td>
</tr>
<tr>
<td>SASTime</td>
<td>specifies the local time of execution in SAS datetime format.</td>
</tr>
<tr>
<td>Time</td>
<td>specifies the local time of execution as a date and time string.</td>
</tr>
<tr>
<td>User</td>
<td>specifies the user ID that started the server.</td>
</tr>
<tr>
<td>Host</td>
<td>specifies the host name of the grid host machine.</td>
</tr>
<tr>
<td>LASRServer</td>
<td>specifies the server signature file.</td>
</tr>
<tr>
<td>Port</td>
<td>specifies the network port number on which the server listens.</td>
</tr>
<tr>
<td>RawCmd</td>
<td>specifies the request that is received by the server.</td>
</tr>
<tr>
<td>ExeCmd</td>
<td>specifies the command that the server executes. This value can include default substitutions or adjustments to the RawCmd (for example, completion of variable lists).</td>
</tr>
<tr>
<td>JnlMsg</td>
<td>specifies an error message that is buffered in a journal object.</td>
</tr>
<tr>
<td>StatusMsg</td>
<td>specifies the status completion message.</td>
</tr>
<tr>
<td>RunTime</td>
<td>specifies the processing duration (in seconds).</td>
</tr>
</tbody>
</table>

The server uses a journal object to buffer messages that can be localized. The format for the JnlMsg value is \textit{n-m:text}.

- \textit{n} is an integer that specifies the message is the \textit{n}th in the journal.
- \textit{m} is an integer that specifies the message severity.
- \textit{text} is a text string that specifies the error.

**Sample JnlMsg Values**

\texttt{JnlMsg=1-4:ERROR: The variable c1 in table WORK.EMPTY must be numeric for this analysis.}
Jn1Msg=2-4:ERROR: You do not have sufficient authorization to add tables to this LASR Analytic Server.
Chapter 2
Non-Distributed SAS LASR Analytic Server

About Non-Distributed SAS LASR Analytic Server

In a non-distributed deployment, the SAS LASR Analytic Server runs on a single machine. All of the in-memory analytic features that are available for the distributed deployment are also available for the non-distributed server.

One key difference has to do with reading and writing data. Because the server does not use a distributed computing environment, the server cannot be co-located with a data provider. The server does not read data in parallel and does not write SASHDAT files to HDFS.

Starting and Stopping Non-Distributed Servers

Starting Servers

Non-distributed servers are started and stopped with the SAS LASR Analytic Server engine. Starting a server requires the STARTSERVER= LIBNAME option.

To start a server:

Example Code 2.1 Starting a Non-Distributed Server

```sas
libname server1 sasiala;
startserver=(
  path="c:\temp"
  keeplog=yes maxlogsize=20
)
host=localhost
port=10010
tag='hps';
```
1 The STARTSERVER= option indicates to start a server. For information about the options, see “STARTSERVER= YES | NOSTARTSERVER = (non-distributed-server-options)” on page 301.

2 The KEELOG= option implies the LOGGING option and prevents the server from removing the log file when the server exits. The MAXLOGSIZE= option specifies to use up to 20 MB for the log file before the file is rolled over.

3 The HOST= specification is optional.

4 If you do not specify a PORT= value, then the server starts on a random port and sets the LASRPORT macro variable to the network port number.

Submitting the previous LIBNAME statement from a SAS session starts a server and the server remains running as long as the SAS session remains running. In a batch environment where you want to start a server for client/server use by other users, follow the LIBNAME statement with the following VASMP procedure statements:

**Example Code 2.2 SERVERWAIT Statement for the VASMP Procedure**

```sas
proc vasmp;
  serverwait port=10010;
quit;
```

The SERVERWAIT statement causes the server to continue running and wait for a termination request.

When a non-distributed SAS LASR Analytic Server is used in a metadata environment like SAS Visual Analytics, the SIGNER= option enables the server to enforce the permissions that are set in metadata. The values for the HOST= and PORT= options must match the host name and network port number that are specified for the server in metadata.

```sas
libname server1 sasiola startserver=(path="/tmp") host="server.example.com" port=10010 tag='hps'
signer="http://server.example.com/SASLASRAuthorization";
```

For information about using SAS LASR Analytic Server in a metadata environment, see *SAS Visual Analytics: Administration Guide*.

If you want to use a script for starting a server, then include the STARTSERVER=LIBNAME option and the SERVERWAIT statement for the VASMP procedure in the program. Start one server only and do not include additional SAS statements after the QUIT statement for the VASMP procedure. If additional statements are included, they can prevent the SAS session from terminating (after receiving a SERVERTERM request). This can prevent the SAS session from freeing memory resources that were used by the server. It is best to restrict the program to starting the server only.

**Stopping Servers**

Stopping a server is performed by clearing the libref that was used to start the server (if you start the server from a SAS session and keep the session running) or with the SERVERTERM statement.

To stop a server from the same SAS session that started it:

**Example Code 2.3 Stopping a Non-Distributed Server with the LIBNAME CLEAR Option**

```sas
libname server1 clear;
```
To stop a server from a different SAS session, use the SERVERTERM statement:

**Example Code 2.4  SERVERTERM Statement for the VASMP Procedure**

```sas
proc vasmp;
  serverterm host="server.example.com" port=10010;
quit;
```

*Note:* Exiting the SAS session that started the server also terminates the server because all librefs are automatically cleared at the end of a SAS session.

---

### Loading and Unloading Tables for Non-Distributed Servers

Tables are loaded into memory in a non-distributed server with the SAS LASR Analytic Server engine. A DATA step can be used. The following example demonstrates loading the Prdsale table into memory after starting a server on port 10010.

To load a table to memory:

**Example Code 2.5  Loading a Table to Memory for Non-Distributed Servers**

```sas
libname server1 startserver port=10010 tag='hps';

data server1.prdsale;
  set sashelp.prdsale;
run;
```

You can unload a table from memory with the DATASETS procedure:

**Example Code 2.6  Unloading a Table with the DATASETS Procedure**

```sas
proc datasets lib=server1;
  delete prdsale;
quit;
```
Overview: LASR Procedure

What Does the LASR Procedure Do?

The LASR procedure is used to start, stop, and load and unload tables from a distributed SAS LASR Analytic Server. The LASR procedure can also be used to save in-memory tables to HDFS.

Data Sources

The LASR procedure can transfer data from any data source that SAS can read and load it into memory on the SAS LASR Analytic Server. However, the LASR procedure can also be used to make the server read data from a co-located data provider. The HDFS that is part of SAS High-Performance Deployment of Hadoop provides a co-located data provider. Some third-party vendor databases can also act as co-located data providers.
Two examples of third-party vendor databases are the Greenplum Data Computing Appliance (DCA) and Teradata Data Warehouse Appliance. When the data is co-located, each machine that is used by the server instance reads the portion of the data that is local. Because the read is local and because the machines read in parallel, very large tables are read quickly.

In order to use a third-party vendor database as a co-located data provider, the client machine must be configured with the native database client software and the SAS/ACCESS Interface software for the database. The database is identified in a LIBNAME statement. The LASR procedure then uses the SERVER= information from the LIBNAME statement and the host name information in the PERFORMANCE statement to determine whether the data is co-located. If the host information is the same, then the data is read in parallel.

## Syntax: LASR Procedure

**PROC LASR** server-options;

PERFORMANCE performance-options;

REMOVE table-specification;

SAVE table-specification / save-options;

<table>
<thead>
<tr>
<th>Statement</th>
<th>Task</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROC LASR</td>
<td>Start a server.</td>
<td>Ex. 1</td>
</tr>
<tr>
<td>PROC LASR</td>
<td>Start a server with logging.</td>
<td>Ex. 2</td>
</tr>
<tr>
<td>PROC LASR</td>
<td>Using the SAS Data in HDFS engine.</td>
<td>Ex. 3</td>
</tr>
<tr>
<td>PROC LASR</td>
<td>Load a table from Teradata to memory</td>
<td>Ex. 4</td>
</tr>
<tr>
<td>PROC LASR</td>
<td>Load a table from Greenplum to memory</td>
<td>Ex. 5</td>
</tr>
<tr>
<td>REMOVE</td>
<td>Unload a table from memory.</td>
<td>Ex. 6</td>
</tr>
<tr>
<td>PROC LASR</td>
<td>Stop a server.</td>
<td>Ex. 7</td>
</tr>
<tr>
<td>PROC LASR</td>
<td>Working with user-defined formats.</td>
<td>Ex. 8</td>
</tr>
<tr>
<td>PROC LASR</td>
<td>Working with user-defined formats and the FMTLIBXML= option.</td>
<td>Ex. 9</td>
</tr>
<tr>
<td>SAVE</td>
<td>Save a table to HDFS.</td>
<td>Ex. 10</td>
</tr>
</tbody>
</table>

## PROC LASR Statement

Controls the SAS LASR Analytic Server.
Syntax

PROC LASR server-options;

Server Options
These options control how the server starts, stops, and operates with data.

ADD
specifies to load a table to the SAS LASR Analytic Server. The data to load is identified by the DATA= option or the HDFS= option.

You can also add tables to memory with the SAS LASR Analytic Server engine. An important difference between using the LASR procedure and the engine is that the procedure has the ability to load data in parallel.

CONCURRENT=maximum-requests
specifies the number of concurrent requests that can execute in the server. This option does not reject connections or requests that exceed maximum-requests. When maximum-requests is reached, the additional requests are queued and then processed in first-in-first-out order.

After the server is running, you can adjust this value in a SERVERPARM statement with the VASMP procedure.

Alias NACTIONS=
Default 20

CREATE <="server-description-file">
specifies to start a server. The optional server-description-file argument specifies the fully qualified path to a file. The file is created when the server starts. Enclose the value in quotation marks. The fully qualified path is limited to 200 characters. The server description file is assigned to the LASRLAST macro variable.

If you do not specify a server description file, then you can use the PORT= option to specify the network port number. In either case, the LASRPORT macro variable is updated with the network port number that the server uses for communication.

DATA=libref.member-name
specifies the data to load into the SAS LASR Analytic Server.

DETAILS= TABLES | ALL
specifies the information to return. Use TABLES to retrieve the table names, NLS encoding, row count, owner, and the table load time. The ALL value provides the previous information and adds the MPI rank and host name for each machine in the server.

The information always includes the performance information. This information includes the host name for the grid host, the grid installation location, and the number of machines in the server.

EXTERNALMEM=pct
specifies the percentage of memory that can be allocated before the server stops transferring data to external processes such as external actions and the SAS High-Performance Analytics procedures. If the percentage is exceeded, the server stops transferring data.

Default 75
FMTLIBXML
specifies the file reference for a format stream. For more information, see “Example 8: Working with User-Defined Formats” on page 37.

FORCE
specifies that a server should be started even if the server description file specified in the CREATE= option already exists. The procedure attempts to stop the server process that is described in the existing server description file and then the file is overwritten with the details for the new server.

Restriction Use this option with the CREATE= option only.

HDFS(HDFS-options)
specifies the parameters for the SASHDAT file to load from HDFS.

TIP Instead of specifying the HDFS option and parameters, you can use the ADD= option with a SAS Data in HDFS engine library.

FILE=
specifies the fully qualified path to the SASHDAT file. Enclose the value in quotation marks. The filename is converted to lowercase and the SASHDAT file in HDFS must be named in lowercase.

Alias PATH=

LABEL=
specifies the description to assign to the table. This value is used to override the label that was associated with the data set before it was stored in HDFS. If this option is not specified, then the label that was associated with the data set is used. Enclose the value in quotation marks.

DIRECT
specifies that the data is loaded directly from HDFS into memory. This option provides a significant performance improvement. With this option, the user account ID that is used to start the server process is used to create the table signature file.

Alias HADOOP=

LIFETIME=maximum-runtime<active-time>
specifies the duration of the server process, in seconds. If you do not specify this option, the server runs indefinitely.

maximum-runtime
When the maximum-runtime is specified without an active-time value, the server exits after maximum-runtime seconds.

active-time
When the maximum-runtime and active-time values are specified, the server runs for maximum-runtime seconds and then starts a run timer with an inactivity timeout of active-time seconds. When the server is contacted with a request, the run timer is reset to zero. Each second that the server is unused, the run timer increments to count the number of inactive seconds. If the run timer reaches the active-time, the server exits.

LOGGING <(log-options)>
The log file is named lasr.log.
CLF
specifies to use the common log format for log files. This format is a
standardized text file format that is frequently analyzed by web analysis
software. Specifying this option implies the LOGGING option.

KEEPLOG
specifies to keep the log files when the server exits instead of deleting them. By
default, the log files are removed when the server exits. If you did not specify this
option when the server was started, you can specify it as an option to the STOP
option.

MAXFILESIZE=
specifies the maximum log file size, in megabytes, for a log file. When the log
file reaches the specified size, a new log file is created and named with a
sequentially assigned index number (for example, .log.1). The default value is
100 megabytes.

Tip Do not include an MB or M suffix when you specify the size.

MAXROLLNUM=
specifies the maximum number of log files to create. When the maximum has
been reached, the server begins to overwrite existing log files. The oldest log file
is overwritten first. The default value is 10.

OSENCODING
specifies that the log file is produced with the operating system encoding of the
SAS LASR Analytic Server root node. This option is useful when the server is
run in a different encoding than the operating system, but you want a log file that
is readable in the server operating system.

PATH='log-file-directory'
specifies the fully qualified path to the directory to use for server log files. The
default value is /tmp.

MERGELIMIT=n
specifies that when the number of unique values in a numeric GROUPBY variable
exceeds n, the variable is automatically binned and the GROUPBY structure is
determined based on the binned values of the variable, rather than the unique
formatted values.

For example, if you specify MERGELIMIT=500, any numeric GROUPBY variable
with more than 500 unique formatted values is binned. Instead of returning results
for more than 500 groups, the results are returned for the bins. You can specify the
number of bins with the MERGEBINS= option.

NOCLASS
specifies that all character variables are not to be treated implicitly as classification
variables. Without this option, all character variables are implicitly treated as
classification variables. The performance for loading tables is improved when this
option is used.

PATH="signature-file-path"
specifies the directory to use for storing the server and table signature files. The
specified directory must exist on the machine that is specified in the GRIDHOST=
environment variable.

PORT=integer
specifies the network port number to use for communicating with the server. You
can specify a port number with the CREATE option to start a server on the specified
port.
Interaction  Do not specify the PORT= option in the LASR procedure statement with a LASRSERVER= option in the PERFORMANCE statement.

**READAHEAD**
specifies for the server to be more aggressive in reading memory pages during the mapping phase when tables are loaded from HDFS. Loading the table takes more time with this option, but the first access of the table is faster.

**Engine**  SAS Data in HDFS engine

**SERVERPERMISSIONS=mode**
specifies the permission setting for accessing the server instance. The mode value is specified as an integer value such as 755. The mode corresponds to the mode values that are used for UNIX file access permissions.

<table>
<thead>
<tr>
<th>Alias</th>
<th>SERVERPERM=</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
<td>600 to 777</td>
</tr>
</tbody>
</table>

Interaction  You can use this option with the CREATE option when you start a server.

**SIGNER="authorization-web-service-uri"**
specifies the URI for the SAS LASR Authorization web service. The web service is provided by the SAS Visual Analytics software. For more information, see *SAS Visual Analytics: Administration Guide*.

Example  SIGNER="https://server.example.com/SASLASRAuthorization"

**STOP <(stop-options)>**
terminates a SAS LASR Analytic Server. The server instance is specified in the LASRSERVER= option that identifies a server description file, or it is determined from the LASRLAST macro variable. Once the server instance receives a request to stop, the server does not accept new connections.

**IMMEDIATE**
specifies to stop the server without waiting for current requests to complete. Without this option, termination requests are queued and can be queued behind a long-running request.

| Alias | NOW |

**KEEPLOG**
specifies to keep log files that are created with the LOGGING option.

| Alias | TERM |

**TABLEMEM=pct**
specifies the percentage of memory that can be allocated before the server rejects requests to add tables or append data. If the percentage is exceeded, adding a table or appending rows to tables fails. These operations continue to fail until the percentage is reset or the memory usage on the server drops below the threshold.

This option has no effect for non-distributed servers. For non-distributed servers, the memory limits can be controlled with the MEMSIZE system option.

**Note:** The specified pct value does not specify the percentage of memory allocated to in-memory tables. It is the percentage of all memory used by the entire server.
machine that—if exceeded—prevents further addition of data to the server. The memory used is not measured at the process or user level, it is computed for the entire machine. In other words, if operating system processes allocate a lot of memory, then loading tables into the server might fail. The threshold is not affected by memory that is associated with SASHDAT tables that are loaded from HDFS.

**Alias**

MEMLOAD=

**Default**

75

**TABLEPERMISSIONS=mode**

specifies the permission setting for accessing a table. The mode value is specified as an integer value such as 755. The mode corresponds to the mode values that are used for UNIX file access permissions.

**Alias**

TABLEPERM=

**Range**

600 to 777

**Interaction**

You can use this option with the ADD option when you load a table to memory.

**VERBOSE**

specifies to request additional information about starting a server or connecting to a server in the SAS log. This information can be helpful to diagnose environment configuration issues.

**Alias**

GRIDMSG

---

**PERFORMANCE Statement**

The PERFORMANCE statement defines performance parameters for multithreaded and distributed computing.

**Examples:**

“Example 4: Load a Table from Teradata to Memory” on page 34

“Example 6: Unload a Table from Memory” on page 36

**Syntax**

PERFORMANCE performance-options;

**Performance Statement Options**

**COMMIT=**

specifies that periodic updates are written to the SAS log when observations are sent from the client to the server instance. Whenever the number of observations sent exceeds an integer multiple of the COMMIT= size, a message is written to the SAS log. The message indicates the actual number of observations distributed and not an integer multiple of the COMMIT= size.

**DATASERVER=**

specifies the host to use for a database connection. This option is used in Teradata deployments so that the LASR procedure compares this host name with the host
name that is specified in the SERVER= option in the LIBNAME statement. If you do not specify the DATASERVER= option, the host to use for the database connection is determined from the GRIDDATASERVER= environment variable.

**HOST=**
specifies the grid host to use for the server instance. Enclose the host name in quotation marks. If you do not specify the HOST= option, it is determined from the GRIDHOST= environment variable.

*Alias GRIDHOST=* 

**INSTALL=**
specifies the path to the TKGrid software on the grid host. If you do not specify this option, it is determined from the GRIDINSTALLLOC= environment variable.

*Alias INSTALLOC=* 

**LASRSERVER=**
specifies the server to use. Provide the fully qualified path to the server description file.

*Alias LASR=* 

**MODE= SYM | ASYM**
specifies whether the server runs in symmetric (SYM) mode or asymmetric (ASYM) mode. Asymmetric mode is useful when you want to read data in parallel from a massively parallel processing (MPP) data appliance that is not co-located with the server.

*Alias GRIDMODE=* 

*Default SYM*

**NODES=**
specifies the number of machines in the cluster to use for the server instance. Specify ALL to calculate the number automatically.

*Alias NNODES=* 

*Restriction* This option has no effect when you use a third-party vendor database as a co-located data provider and you specify the CREATE= and DATA= options in the PROC LASR statement. When you use a third-party vendor database as a co-located data provider, you must use all of the machines to read data from the database.

**NTHREADS=**
specifies the number of threads for analytic computations and overrides the SAS system option THREADS | NOTHREADS. By default, the server uses one thread for each CPU core that is available on each machine in the cluster. Use this option to throttle the number of CPU cores that are used on each machine.

*Note: The SAS system options THREADS | NOTHREADS apply to the client machine that issues the PROC LASR statement. They do not apply to the machines in the cluster.*
TIMEOUT=
specifies the time in seconds for the LASR procedure to wait for a connection to the
grid host and establish a connection back to the client. The default value is 120
seconds. If jobs are submitted through workload management tools that might
suspend access to the grid host for a longer period, you might want to increase the
value.

REMOVE Statement
The REMOVE statement is used to unload a table from memory.

Syntax

REMOVE table-specification;

Required Argument
table-specification
specifies the table to unload from memory. For a table that was loaded from a SAS
library, the table specification is the same libref. member-name that was used to load
the table. For a table that was loaded from HDFS, the table specification is the same
as the HDFS path to the table, but is delimited with periods (.) instead of slashes (/).
For a table that was loaded from the / directory in HDFS, the table specification is
HADOOP.TABLENAME.

SAVE Statement
The SAVE statement is used to save an in-memory table to HDFS.

Syntax

SAVE table-specification / save-options;

Required Arguments
table-specification
specifies the table that is in memory. For a table that was loaded from a SAS library
with the procedure, the table specification is the same libref. member-name that was
used to load the table. For a table that was loaded from HDFS, the table specification
is the same as the HDFS path to the table, but is delimited with periods (.) instead of
slashes (/). For a table that was loaded from the / directory in HDFS, the table
specification is HADOOP.TABLENAME.

save-options
specifies the options for saving the file in HDFS.
BLOCKSIZE=
specifies the block size to use for distributing the data set. Suffix values are B
(bytes), K (kilobytes), M (megabytes), and G (gigabytes). The default block size
is 32M.
Alias BLOCK=

COPIES=n
specifies the number of replications to make for the data set (beyond the original blocks). The default value is 1.

FULLPATH
specifies that the value for the PATH= option specifies the full path for the file, including the filename.

PATH=’HDFS-path’
specifies the directory in HDFS in which to store the SASHDAT file. The value is case sensitive. The filename for the SASHDAT file that is stored in the path is always lowercase.

Note: If the PATH= option is not specified, the server attempts to save the table in the /user/userid directory. The userid is the user ID that started the server instance.

REPLACE
specifies that the SASHDAT file should be overwritten if it already exists.

Examples: LASR Procedure

Example 1: Start a Server

Details
This PROC LASR example demonstrates starting a server instance on network port number 10010. Once the server instance is started, the LASRPORT macro variable in the SAS session is set.

Program

```sas
option set=GRIDHOST="grid001.example.com";
option set=GRIDINSTALLLOC="/opt/TKGrid";

proc lasr create port=10010 path="/tmp" noclass;

performance nodes=all;
run;
```

Program Description
1. The GRIDHOST= and GRIDINSTALLLOC= environment variables are used to identify the machine to connect to and the location of the SAS High-Performance Analytics components.
2. The CREATE option is required and the PORT= option specifies the network port number to use.
Example 2: Starting a Server with Logging Options

Details
This PROC LASR example demonstrates how to start a server instance and specify logging options.

Program

```
option set=GRIDHOST="grid001.example.com";
option set=GRIDINSTALLLOC="/opt/TKGrid";

proc lasr
  create port=10010
  path="/tmp"
  noclass
  logging(path="/opt/logs" maxfilesize=5 keeplog clf);
  performance nodes=all;
run;
```

Program Description
The logging statement modifies the default logging behavior. Log files are written to `/opt/logs` instead of the default directory, `/tmp`. The log files are rolled over when they reach five megabytes. The KEEPLOG option is used to keep the log files when the server exits rather than delete them.

Example 3: Using the SAS Data in HDFS Engine

Details
The LASR procedure can load tables to memory from HDFS with the SAS Data in HDFS engine. This use is similar to using the HDFS option with the procedure, but has the advantage that you can use FORMAT statements and data set options.

Program

```
option set=GRIDHOST="grid001.example.com";
option set=GRIDINSTALLLOC="/opt/TKGrid";

libname grp1 sashdat path="/dept/grp1";
libname grp1 sashdat path="/dept/grp1";

proc lasr create port=10010 noclass;
  performance nodes=all;
run;

proc lasr add data=grp1.sales2012 port=10010;
  format predict $dollar20.
  actual $dollar20.;
```
run;

proc lasr add data=grp1.sales2013(where=(region="West")) port=10010;  
run;

Program Description

1. The GRIDHOST= and GRIDINSTALLLOC= environment variables are used by the
   LASR procedure and the GRIDHOST= option is also used by the LIBNAME
   statement.

2. The SAS Data in HDFS engine uses the GRIDHOST= environment variable to
determine the host name for the NameNode. The PATH= option is used to specify
   the directory in HDFS.

3. The FORMAT statement is used to override the format name in HDFS for the
   variable.

4. The WHERE clause subsets the Sales2013 table. Only the rows with Region equal to
   "West" are read into memory. The WHERE clause is useful for subsetting data, but it
does not take advantage of the memory efficiencies that are normally used with
   SASHDAT tables.

If the table in HDFS has variables that are associated with user-defined formats, then
you must have the user-defined formats available in the format catalog search order.

Example 4: Load a Table from Teradata to Memory

Details

This PROC LASR example demonstrates how to load a table to memory from Teradata.
The native database client for Teradata and SAS/ACCESS Interface to Teradata must be
installed and configured on the client machine.

libname tdlib teradata server="dbccop1.example.com"  
database=hps user=dbc password=dbcpass;

proc lasr create port=10010
   data=tdlib.sometable
   path="/tmp";

   performance host="tms.example.com"
   install="/opt/TKGrid"
   dataserver="dbccop1.example.com";
run;

proc lasr add data=tdlib.tabletwo [label = "Table description"]  
   port=10010;
   format revenue dollar20.2
       units comma9;
run;
Program Description

1. The `SERVER=` option in the `LIBNAME` statement specifies the host name for the Teradata database.
2. The `HOST=` option in the `PERFORMANCE` statement specifies the host name of the Teradata Management Server (TMS).
3. The `DATASERVER=` option in the `PERFORMANCE` statement specifies the same host name for the Teradata database that is used in the `LIBNAME` statement.
4. The input data set option, `LABEL=`, associates the description with the data in the server instance. This option causes a warning in the SAS log because the SAS/ACCESS Interface to Teradata does not support data set labels.
5. SAS formats are applied with the `FORMAT` statement. Specifying the variable formats is useful for DBMS tables because database systems do not store formats.

Example 5: Load a Table from Greenplum to Memory

Details

This PROC LASR example demonstrates how to load a table to memory from Greenplum. The ODBC drivers and SAS/ACCESS Interface to Greenplum must be installed and configured on the client machine.

```
libname gplib greenplm server="mdw.example.com"  
database=hps user=dbuser password=dbpass;
```

```
proc lasr create port=10010  
data=gplib.sometable  
path="/tmp";

performance host="mdw.example.com"  
install="/opt/TKGrid";
run;
```

```
proc lasr add data=gplib.tabletwo [label = "Table description"]  
port=10010;  
format y x1-x15 5.4  
dt date9.;
run;
```

Program Description

1. The `SERVER=` option in the `LIBNAME` statement specifies the host name for the Greenplum database.
2. The `HOST=` option in the `PERFORMANCE` statement specifies the host name of the Greenplum master host.
3. The input data set option, `LABEL=`, associates the description with the data in the server instance. This option causes a warning in the SAS log because the SAS/ACCESS Interface to Greenplum does not support data set labels.
4. SAS formats are applied with the `FORMAT` statement. Specifying the variable formats is useful for DBMS tables because database systems do not store formats.
Example 6: Unload a Table from Memory

Details
This PROC LASR example demonstrates how to unload tables from memory. The first REMOVE statement applies to tables that were loaded from HDFS. The second REMOVE statement is typical for tables that are loaded from SAS libraries.

Program

```
libname finance '/data/finance/2011/';

proc lasr port=10010;
  remove user.sales.2011.q4; 1
  remove finance.trans; 2
  performance host="grid001.example.com"
       install="/opt/TGGrid";
run;
```

Program Description
1. This REMOVE statement specifies a table that was loaded from HDFS.
2. The libref and member name for a SAS data set are specified in this REMOVE statement example.

Example 7: Stopping a Server

Details
This PROC LASR example demonstrates stopping a server instance.

Program

```
option set=GRIDHOST="grid001.example.com";
option set=GRIDINSTALLLOC="/opt/TGGrid";

proc lasr term port=10010;
run;
```

Program Description
The server instance listening on port 10010 is stopped.
Example 8: Working with User-Defined Formats

Details
By default, when user-defined formats are used with the server, the LASR procedure automatically uses these formats. The formats must be available in the format catalog search order. You can use the FMTSEARCH= system option to specify the format catalog search order. The LASR procedure converts the formats to an XML representation and transfers them to the server with the data.

Program

```plaintext
proc format library=myfmts;
  value YesNo 1='Yes' 0='No';
  value checkThis 1='ThisisOne' 2='ThisisTwo';
  value $cityChar 1='Portage' 2='Kinston';
run;

options fmtsearch=(myfmts);

proc lasr add data=orsdm.profit_company_product_year port=10010;
  format city $cityChar.;
.performance host="grid001.example.com"
     install="/opt/TKGrid"
     nodes=ALL;
run;
```

Program Description
1. The user-defined formats are available to the LASR procedure because they are added to the format catalog search order.
2. When the $cityChar. format is applied to the city variable, the LASR procedure converts the formats to XML, and transfers the format information and the data to the server.

Example 9: Working with User-Defined Formats and the FMTLIBXML= Option

Details
As explained in the previous example, the LASR procedure can use any format so long as the format is in the format catalog search order. The procedure automatically converts the format information to XML and transfers it to the server with the data. However, if the same formats are used many times, it is more efficient to convert the formats to XML manually and use the FMTLIBXML= option.

You can use the FORMAT procedure to write formats to an XML fileref. Then, you can reference the fileref in the FMTLIBXML= option each time you use the LASR.
procedure to load tables. This improves performance because the conversion to XML occurs once rather than each time LASR procedure transfers the data.

Formats are created with the FORMAT procedure. The following SAS statements show a simple example of creating a format and using the XML fileref in the LASR procedure.

**Program**

```sas
proc format library=gendrfmt;
  value $gender 'M'='Male' 'F'='Female';
run;

options fmtsearch=(gendrfmt);

filename fmtxml 'genderfmt.xml';
libname fmtxml XML92 xmltype=sasfmt tagset=tagsets.XMLsuv;
proc format library=gendrfmt cntlout=fmtxml.allfmts;
run;
proc lasr add data=sashelp.class fmtlibxml=fmtxml;
  format sex $gender.;
  performance host="grid001.example.com"
    install="/opt/TKGrid"
    nodes=ALL;
run;
```

**Program Description**

1. The user-defined formats are available to the LASR procedure because they are added to the format catalog search order.

2. An XML stream for the formats in the file GenderFmt.xml is associated with the file reference FmtXml. The formats are converted to XML and stored in the file.

3. The file reference FmtXml is used with the FMTLIBXML= option in the PROC LASR statement. For subsequent uses of the LASR procedure, using the FMTLIBXML= option to reference the fileref is efficient because the formats are already converted to XML.

4. The $gender. format information is transferred to the server in an XML stream and associated with the variable that is named sex. However, the format must be available to the SAS session that runs the LASR procedure.

---

**Example 10: Saving a Table to HDFS**

**Details**

The server can save in-memory tables to HDFS. Use the SAVE statement to provide a table specification and the save options.

```sas
option set=GRIDHOST="grid001.example.com";
option set=GRIDINSTALLLOC="/opt/TKGrid";
```
proc lasr port=10010;
  save sales.sales2012 / path="/dept/grp1/" copies=1 blocksize=32m;  
  save sales.avg2012 / fullpath path="/dept/grp1/avg/y2012" copies=1;  
run;

Program Description
1. The table that is named sales2012 is saved to HDFS as /dept/grp1/sales2012.sashdat.
2. The table that is named avg2012 is saved to HDFS as /dept/grp1/avg/y2012.sashdat.
   The FULLPATH option is used to rename the file.
Chapter 4
IMSTAT Procedure (Analytics)

Overview: IMSTAT Procedure (Analytics) ............................................ 42
What Do the Analytic Statements for the IMSTAT Procedure Do? .......... 42

Syntax: IMSTAT Procedure (Analytics) ............................................... 42
PROC IMSTAT (Analytics) Statement ........................................... 43
ARM Statement ................................................................. 45
ASSESS Statement ............................................................ 57
BOXPLOT Statement .......................................................... 60
CLUSTER Statement ........................................................... 63
CORR Statement ............................................................... 69
CROSSTAB Statement ......................................................... 69
DECISIONTREE Statement .................................................... 74
DISTINCT Statement .......................................................... 80
FORECAST Statement .......................................................... 87
FREQUENCY Statement ........................................................ 90
GENMODEL Statement ........................................................ 92
GLM Statement ................................................................. 103
GROUPBY Statement .......................................................... 118
HISTOGRAM Statement ....................................................... 122
KDE Statement ................................................................. 124
LOGISTIC Statement ............................................................ 125
MDSUMMARY Statement ...................................................... 140
OPTIMIZE Statement .......................................................... 142
PERCENTILE Statement ....................................................... 147
RANDOMWOODS Statement .................................................. 151
REGCORR Statement .......................................................... 156
SUMMARY Statement ........................................................ 158
TEXTPARSE Statement ........................................................ 164
TOPK Statement ................................................................. 169
QUIT Statement ................................................................. 171

Examples: IMSTAT Procedure (Analytics) ......................................... 172
Example 1: Calculating Percentiles and Quartiles ................................ 172
Example 2: Retrieving Box Values ............................................. 174
Example 3: Retrieving Box Plot Values with the NOUTLIERLIMIT= Option ... 175
Example 4: Performing a Cluster Analysis ..................................... 176
Example 5: Performing a Pairwise Correlation ............................... 177
Example 6: Crosstabulation with Measures of Association and Chi-Square Tests 178
Example 7: Training and Validating a Decision Tree .......................... 180
Example 8: Storing and Scoring a Decision Tree ............................ 182
Example 9: Performing a Multi-Dimensional Summary ...................... 184
Example 10: Fitting a Regression Model ....................................... 185
Overview: IMSTAT Procedure (Analytics)

What Do the Analytic Statements for the IMSTAT Procedure Do?

The analytic statements of the IMSTAT procedure are used to perform in-memory analytics with a SAS LASR Analytic Server. All analyses are performed on in-memory tables.

Syntax: IMSTAT Procedure (Analytics)

Tip: For information about the data and server management statements, see Chapter 5, “IMSTAT Procedure (Data and Server Management),” on page 187.

PROC IMSTAT <options>;
  ARM ITEM=item-variable TRA=transaction-variable </ options>;
  ASSESS <variable-list> / Y=response-variable </ options>;
  BOXPLOT <variable-list> </ options>;
  CLUSTER <variable-list> </ options>;
  CORR <variable-list> </ options>;
  CROSSTAB row*column </ options>;
  DECISIONTREE target-variable </ options>;
  DISTINCT <variable-list> </ options>;
  FORECAST timestamp-variable </ options>;
  FREQUENCY variable-list </ options>;
  GENMODEL dependent-variable <(class-variables)= model-effects </ options>;
  GLM dependent-variable <(class-variables)= model-effects </ options>;
  GROUPBY <variable-list> </ options>;
  HISTOGRAM <variable-list> </ options>;
  KDE variable-list </ options>;
  LOGISTIC dependent-variable <(class-variables)= model-effects </ options>
  MDSUMMARY variable-list </ set-specification,...> options>;
  OPTIMIZE <options>;
  PERCENTILE <variable-list> </ options>;
  RANDOMWOODS target-variable </ options>;
  REGCORR <variable-list> </ options>;
  SUMMARY <variable-list> </ options>;
  TEXTPARSE TXT=text-variable ID=document-ID </ options>;
  TOPK <variable-list> </ options>;
QUIT;
PROC IMSTAT (Analytics) Statement

Performs in-memory analytics with a SAS LASR Analytic Server.

Syntax

PROC IMSTAT <options>;

Optional Arguments

BATCHMODE
By default, the IMSTAT procedure operates in interactive mode. If your program contains errors that prevent SAS from parsing or executing statements, the errors are reported in the SAS log, but they do not stop the procedure. If the errors are fatal errors such as running out of memory on the SAS client, the procedure stops.

In contrast, when the BATCHMODE option is specified in the PROC IMSTAT statement, the procedure behaves with respect to error handling as if it were not an interactive procedure. Whenever an error occurs, the procedure terminates and sets the SYSERR macro variable.

Alias BATCH

DATA=libref.member-name
specifies the table to access from memory. The libref must be assigned from a SAS LASR Analytic Server engine LIBNAME statement.

FMTLIBXML=file-reference
specifies the file reference for a format stream. For more information, see “FMTLIBXML” in the LASR procedure.

IMMEDIATE
specifies that the procedure executes one statement at a time rather than accumulating statements in RUN blocks.
NODATE

specifies to suppress the display of time and date-dependent information in results from the TABLEINFO statement.

NOPREPARSE

prevents the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects problems with the code, the error messages might not to be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

When you specify this option in the PROC IMSTAT statement, the option applies to all statements that can generate code. You can also exclude specific statements from pre-parsing by using the NOPREPARSE option in statements that allow temporary columns or the SCORE statement.

NOPRINT

This option suppresses the generation of ODS tables and other printed output in the IMSTAT procedure. You can use this option to suppress printed output during execution of the actions, and then use the REPLAY statement to print the tables at a later point in the procedure execution.

NOTIMINGMSG

When an action completes successfully, the IMSTAT procedure generates a SAS log message that contains the execution time of the request. Specify this option to suppress the message.

PGMMMSG

specifies to capture messages associated with user-provided SAS statements on the server in a temporary table. Messages are produced when parsing errors occur, when code generation fails, or by PUT statements in a SAS program.

You can use this option as a debugging feature for SAS code that you submit through temporary column expressions. The macro variable _PGMMMSG_ is used in the IMSTAT procedure to capture the name of the table. The _TEMPLAST_ macro variable is also updated in case this temporary table is the most recently created temporary table.

SIGNER="authorization-web-service-uri"

specifies the URI for the SAS LASR Authorization web service. For more information, see SAS Visual Analytics: Administration Guide.

Example SIGNER="https://server.example.com/SASLASRAuthorization"
TEMPTABLEINFO
specifies to add additional information for temporary tables to the ODS table that is
created on the SAS client. The information includes the time at which the temporary
table was created in the server, the number of rows, and the number of columns.

Alias TEMPINFO

UCA
specifies that you want to use Unicode Collation Algorithms (UCA) to determine the
ordering of character variables in the GROUPBY= operations and other operations
that depend on the order of formatted values.

Alias UCACOLLATION

ARM Statement
The ARM statement is used to perform associative rule mining (ARM). You can use it to derive frequent
itemsets, perform association rule mining, and sequence mining.

Syntax
ARM ITEM=item-variable TRAAN=transaction-variable </options>;

Required Arguments
item-variable
specifies the name of the variable in the active table that identifies items.

transaction-variable
specifies the name of the variable in the active table that identifies transactions.

Optional Argument
variable-list
specifies one or more numeric variables. If you do not specify this option, then all
numeric variables in the table are used.

ARM Statement Options
AGGREGATE=aggregation-method
lists the aggregator for which the score of an itemset at each occurrence in a data set
is aggregated into a final score of such itemset. If the WEIGHT= variable is not
specified, then the aggregator specification is ignored.

The available aggregation methods are as follows:

MAX maximum value
MEAN arithmetic mean
MIN minimum value
SUM sum of the nonmissing values

Alias AGG=
FREQ=variable

specifies the numeric frequency variable to use for computing the score of each frequent itemset along with WEIGHT= option. When the FREQ= variable is not specified, the score of a frequent itemset equals the value of the WEIGHT= variable scaled by 1. Negative values for the specified variable are considered missing.

ITEMAGG=aggregation-method

lists the aggregator for which the values of the WEIGHT= variable, and optionally the FREQ= variable, are rolled up into the score of an itemset at each occurrence in the data set, provided that a WEIGHT= variable is specified. If the WEIGHT= variable is not specified, then the aggregator specification is ignored. The aggregation methods are identical to the list in the AGGREGATE= option.

The ITEMAGG= and AGGREGATE= options work together to derive the final score of an itemset. Given an itemset, the score \( S \) is first aggregated over the frequency, \( f \), and weight, \( w \), variables associated with each item at each occurrence among a transaction by item aggregator \( \Phi_{\text{item}}(\cdot) \). Then, the intermediate scores of such itemset among all the occurrences are then aggregated again by set aggregator \( \Phi_{\text{set}}(\cdot) \) See the following equation.

\[
S = \Phi_{\text{set}}(\Phi_{\text{item}}(fw))
\]

ITEMFMT="format-specification"

specifies the formats for the ITEM= variable. If you do not specify the ITEMFMT= option, then the unformatted values of the ITEM= variable are used.

Enclose the format specification in quotation marks.

ITEMSTBL

specifies to save the derived frequent itemsets in a temporary table. By default, the frequent itemsets are not saved.

MAXITEMS=n

specifies the maximal number of items to allow in a frequent itemset. The value must be greater than or equal to 1. If an invalid value is specified, then it is replaced with 1, the default value.

Default 1

MINITEMS=n

specifies the minimal number of items to allow in a frequent itemset. The value must be greater than or equal to zero. If an invalid value is specified, then it is replaced with 0, the default value.

If you specify MAXITEMS= < MINITEMS=, the server swaps the values. If you specify MAXITEMS = MINITEMS, the server assigns MAXITEMS = MINITEMS + 1.

Default 0

NOMISSING

specifies that missing values of the ITEM= and TRAN= variables are excluded from analysis. By default, missing values of the ITEM= variable are considered a separate item. Missing values of the TRAN= variable are considered a separate transaction.
PARTITION \langle=\text{partition-key}\rangle

specifies to use partitioning variables. When only PARTITION is specified and the
table is partitioned first by the TRAN= variable, and the TRANFMT= option is
specified, the associative rule mining is performed separately for each value of the
partition key. If a value for \text{partition-key} is specified, then the associative rule mining
is performed on that partition only.

RELSUPPORT

specifies that the values for LOWER= and UPPER= in the SUPPORT option are
relative to the most frequent itemset. For example, if 500 is the support of the most
frequent itemset, then specifying RELSUPPORT SUPPORT(LOWER=0.1
UPPER=0.5) means the minimum and the maximum supports for the analysis are 50
and 250, respectively. When using this option, the values for LOWER= and
UPPER= must be between 0 and 1. Otherwise, they are set to the default values 0.05
and 1.0, respectively.

RULES(<\text{sub-options}>)

specifies the requirements for how association rules are generated from frequent
itemsets. The following sub-options are available:

\begin{itemize}
  \item \text{AGGREGATE=}aggregation-method
    lists the aggregator for which the score of a rule at each occurrence in a data set
    is aggregated into a final score of such rule. If the WEIGHT= variable is not
    specified, then the aggregator specification is ignored.
\end{itemize}

The available aggregation methods are as follows:

\begin{itemize}
  \item MAX maximum value
  \item MEAN arithmetic mean
  \item MIN minimum value
  \item SUM sum of the nonmissing values
\end{itemize}

\begin{itemize}
  \item Alias \text{AGG}=
  \item Default \text{SUM}
  \item Interaction You must specify the WEIGHT= option to use this option.
\end{itemize}

\text{CONFIDENCE}(<\text{LOWER}=\text{lower-value}> <\text{UPPER}=\text{upper-value}>)

specifies the minimal and maximal confidence values of the association rules
have to fulfill. The default value for LOWER= is 0.5.

If you specify UPPER= \text{< LOWER=}>, the server swaps the values. If you specify
the same value for LOWER= and UPPER=, the server adds \varepsilon
(0.111023024625157e-12) to value and uses the result for UPPER=.

\begin{itemize}
  \item Range 0 to 1
\end{itemize}

\text{FREQ=}\text{variable-name}

specifies the numeric frequency variable to use for computing the score of each
association rule along with ORDER= option. When a FREQ= variable is not
specified, the score of an association rule equates the value of the ORDER= variable scaled by 1. Negative values for the specified variable are considered missing.
ITEMAGG=aggregation-method
lists the aggregator for which the values of the WEIGHT= variable, and optionally the FREQ= variable, are rolled up into the score of a rule at each occurrence in the data set, provided that a WEIGHT= variable is specified. If you do not specify a WEIGHT= variable, then the aggregator specification is ignored. The aggregation methods are identical to the list in the AGGREGATE= option.

NUMLHS<LOWER=lower-value> <UPPER=upper-value>)
specifies the minimum and maximum number of items in the left-hand side (LHS) of a rule to allow. If you specify UPPER= < LOWER=, the server swaps the values.

NUMRHS<LOWER=lower-value> <UPPER=upper-value>)
specifies the minimum and maximum number of items in the right-hand side (RHS) of a rule to allow. If you specify UPPER= < LOWER=, the server swaps the values.

SCORE<LOWER=lower-value> <UPPER=upper-value>)
specifies the minimum and maximum scores of the association rules that are derived. If you specify UPPER= < LOWER=, the server swaps the values. If you specify the same value for LOWER= and UPPER=, the server adds \( \epsilon \) \((0.1110223024625157e-12)\) to value and uses the result for UPPER=.

WEIGHT=variable-name
specifies the numeric weight variable to use for computing the score of each association rule, along with FREQ= variable. If you do not specify a WEIGHT= variable, then the AGGREGATE=, FREQ=, and ITEMAGG= options are ignored.

SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

SCORE<LOWER=lower-value> <UPPER=upper-value>)
specifies the minimum and maximum scores of the frequent itemsets that are derived. If you specify UPPER= < LOWER=, the server swaps the values. If you specify the same value for LOWER= and UPPER=, the server adds \( \epsilon \) \((0.1110223024625157e-12)\) to value and uses the result for UPPER=.

SEQUENCES(TIME=t <sub-options>)
specifies the requirements for how sequences are generated from the original table. The sequences do not necessarily depend on previously generated frequent itemsets. You can specify the following sub-options in SEQUENCES option:

AGGREGATE=aggregation-method
lists the aggregator for which the score of a sequence at each occurrence in a data set is aggregated into a final score of such sequence. If the WEIGHT= variable is not specified, then the aggregator specification is ignored.

The available aggregation methods are as follows:

- MAX maximum value
- MEAN arithmetic mean
- MIN minimum value
- SUM sum of the nonmissing values

Alias AGG=
FREQ=variable-name

specifies the numeric frequency variable to use for computing the score of each sequence along with WEIGHT= option. When a FREQ= variable is not specified, the score of a sequence equates the value of the WEIGHT= variable scaled by 1. Negative values for the specified variable are considered missing.

INCLUDEMISSTIME

indicates that records with a missing value for the TIME= variable are considered for sequence analysis. If this option is specified, then the missing value for the TIME= variable is treated as the smallest value in a sequence.

ITEMAGG=aggregation-method

lists the aggregator for which the values of the WEIGHT= variable, and optionally the FREQ= variable, are rolled up into the score of a rule at each occurrence in the data set, provided that a WEIGHT= variable is specified. If the WEIGHT= variable is not specified, then the aggregator specification is ignored. The aggregation methods are identical to the list in the AGGREGATE= option.

ITEMSETFILTER=SINGLETONS | ALLITEMS | NONE

specifies how the sequences are filtered by frequent itemsets. The SINGLETONS setting means that each item of any sequence has to be a frequent singleton. The ALLITEMS setting means the set of all distinct items of any sequence have to be a frequent itemset. The NONE setting indicates that sequences are not influenced by what frequent itemsets are derived.

The following IMSTAT statements specify that all frequent itemsets must have their support greater or equal than 150. The support ranges specified on sequences are LOWER=110 and UPPER=140. The support boundaries on frequent itemsets and sequences are disjoint. However, with FILTER=NONE, the sequences that are generated do not depend on the frequent itemsets.

```
proc imstat data=example.assocs;
arm item=Product tran=Customer / maxItems=6 support(lower=150) itemsTbl sequs(time=Time minItems=4 maxItems=6 minWindow=0 support(lower=110 upper=140) filter=none);
```
MAXWINDOW=$t$

specifies the maximum difference to allow between the onset of any two adjacent items in a sequence. If the difference is greater than $t$, then the two items cannot be part of the same sequence. The value must be greater than or equal to zero.

MINDURATION=$t$

specifies the minimum difference to allow between the onset time of the last item and the first item in a sequence. If the difference is less than $t$, then the sequence is excluded from the result set. The value must be greater than or equal to zero. If you specify a value for MAXDURATION= that is less than MINDURATION=, the server swaps the values.

MINITEMS=$n$

specifies the minimal number of items to allow in a sequence. The value must be greater than or equal to 1. If an invalid value is specified, then it is replaced with 1, the default value.

If you specify MAXITEMS= < MINITEMS=, the server swaps the values. If you specify MAXITEMS = MINITEMS, the server assigns MAXITEMS = MINITEMS + 1.

Default 1

MINWINDOW=$t$

specifies the minimum difference to allow between the onset of any two adjacent items in a sequence. If the difference is less than or equal to $t$, then the two items are treated as happening at the same time. The value must be greater than or equal to zero.

If you specify MAXWINDOW= < MINWINDOW=, the server swaps the values.

NODUP

specifies that duplicated items within a sequence are not allowed.

SCORE(<LOWER=lower-value> <UPPER=upper-value>)

specifies the minimum and maximum scores of the sequences that are derived. If you specify UPPER= < LOWER=, the server swaps the values. If you specify the same value for LOWER= and UPPER=, the server adds $\varepsilon$ (0.1110223024625157e-12) to the value and uses the result for UPPER=.

SUPPORT(<LOWER=lower-value> <UPPER=upper-value>)

specifies the minimum and maximum support of one sequence allowed in the analysis. By default, LOWER=1 and UPPER= is not set. Valid values for LOWER= and UPPER= are integers greater than 0. If you specify an invalid value for LOWER= or UPPER=, the server sets LOWER=1. The value for LOWER= must be less than or equal to the UPPER= value. If you specify UPPER= < LOWER=, the server swaps the values. Note that

Default LOWER=1

Note This option does not overwrite the SUPPORT option that is specified for deriving frequent itemsets.

TIME=$t$

specifies the numeric variable to use for sorting the items in a sequence. This option is required for sequence analysis.

WEIGHT=variable-name

specifies the numeric weight variable to use for computing the score of each sequence, along with FREQ= variable. If you do not specify a WEIGHT=
variable, then the AGGREGATE=, FREQ=, and ITEMAGG= options are ignored.

**WINDOWAGG=aggregation-method**
is used with the MINWINDOW= and MAXWINDOW= options. It lists the aggregator for which the values of the TIME= variable to update the anchor time. The default value is MEAN.

The available aggregation methods are as follows:

- **MAX**  maximum value
- **MEAN**  arithmetic mean
- **MIN**  minimum value

For example, see the values in the following table:

<table>
<thead>
<tr>
<th>Transaction</th>
<th>Item</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>A</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>D</td>
<td>3</td>
</tr>
</tbody>
</table>

If MINWINDOW=0, then the following sequence is formed with a chain length of four because the time difference between two adjacent items is > 0.

A ⇒ B ⇒ C ⇒ D

If MINWINDOW=1 and WINDOWAGG=MIN, then the following sequence is formed with a chain length of two.

A & B ⇒ C & D

If MINWINDOW=1 and WINDOWAGG=MAX, then the following sequence is formed, with a chain length of one because after A and B are merged as concurrent items, the WINDOWAGG=MEAN setting defines the time value for A & B to be 0.5. Item C is then merged with A & B with a new merged time value of 1.0. Because the time value for item D is 3, then it is not merged with A & B & C.

A & B & C ⇒ D

If MAXWINDOW=1, then no two items in the transaction can form a sequence.

**SUPPORT(<LOWER=lower-value> <UPPER=upper-value>)**
specifies the minimum and maximum frequencies to allow for derived frequent itemsets. If RELSUPPORT is not specified, then LOWER= and UPPER= are the minimum and maximum frequencies of frequent items that appeared in the transactions. If RELSUPPORT is specified, then specify the two values as the ratios for the minimum and maximum frequencies of frequent itemsets to the frequency of the most frequent itemset.

By default, LOWER=1 when RELSUPPORT is not specified and LOWER=0.05 when RELSUPPORT is specified.
The values for LOWER= and UPPER= must be greater than or equal to zero. If you specify UPPER= < LOWER=, then the server swaps the values.

For example, the following statements derive and display frequent itemsets of sizes between MINITEMS=3 and MAXITEMS=4. The support for each frequent itemset must be between [LOWER=97, UPPER=100). LOWER= is inclusive and UPPER= is exclusive.

```plaintext
proc imstat data=example.assocs;
   arm item=Product tran=Customer / minItems=3 maxItems=4
       support(lower=97 upper=100) itemsTBL;
run;
   table example.&_tempARMItems_;
   fetch / orderby=(_SetSize_ _Count) desc=_Count_;
run;
```

The ARM statement produces the following display, indicating that 14 frequent itemsets were derived.

<table>
<thead>
<tr>
<th>Descr</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Source</td>
<td>WORK_ASSOCS</td>
</tr>
<tr>
<td>Number of Transactions</td>
<td>1001</td>
</tr>
<tr>
<td>Number of Frequent Itemsets</td>
<td>14</td>
</tr>
<tr>
<td>Frequent Itemset Table</td>
<td>_T_ACC833E3_7F8DF286E008</td>
</tr>
</tbody>
</table>

The FETCH statement displays the 14 frequent itemsets.

<table>
<thead>
<tr>
<th><em>SetSize</em></th>
<th><em>Count</em></th>
<th><em>Support</em></th>
<th>PRODUCT1</th>
<th>PRODUCT2</th>
<th>PRODUCT3</th>
<th>PRODUCT4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.000000</td>
<td>99.000000</td>
<td>0.098901</td>
<td>steak</td>
<td>herring</td>
<td>apples</td>
<td></td>
</tr>
<tr>
<td>3.000000</td>
<td>99.000000</td>
<td>0.098901</td>
<td>steak</td>
<td>olives</td>
<td>apples</td>
<td></td>
</tr>
<tr>
<td>3.000000</td>
<td>99.000000</td>
<td>0.098901</td>
<td>turkey</td>
<td>ice_crea</td>
<td>bourbon</td>
<td></td>
</tr>
<tr>
<td>3.000000</td>
<td>99.000000</td>
<td>0.098901</td>
<td>peppers</td>
<td>cracker</td>
<td>chicken</td>
<td></td>
</tr>
<tr>
<td>3.000000</td>
<td>98.000000</td>
<td>0.097902</td>
<td>comed_b</td>
<td>chicken</td>
<td>bourbon</td>
<td></td>
</tr>
<tr>
<td>3.000000</td>
<td>98.000000</td>
<td>0.097902</td>
<td>turkey</td>
<td>coke</td>
<td>bourbon</td>
<td></td>
</tr>
<tr>
<td>3.000000</td>
<td>97.000000</td>
<td>0.096903</td>
<td>peppers</td>
<td>baguette</td>
<td>apples</td>
<td></td>
</tr>
<tr>
<td>3.000000</td>
<td>97.000000</td>
<td>0.096903</td>
<td>peppers</td>
<td>comed_b</td>
<td>bourbon</td>
<td></td>
</tr>
<tr>
<td>4.000000</td>
<td>99.000000</td>
<td>0.098901</td>
<td>herring</td>
<td>baguette</td>
<td>avocado</td>
<td>artichok</td>
</tr>
<tr>
<td>4.000000</td>
<td>99.000000</td>
<td>0.098901</td>
<td>ham</td>
<td>cracker</td>
<td>avocado</td>
<td>artichok</td>
</tr>
<tr>
<td>4.000000</td>
<td>97.000000</td>
<td>0.096903</td>
<td>steak</td>
<td>herring</td>
<td>comed_b</td>
<td>apples</td>
</tr>
<tr>
<td>4.000000</td>
<td>97.000000</td>
<td>0.096903</td>
<td>steak</td>
<td>olives</td>
<td>comed_b</td>
<td>apples</td>
</tr>
<tr>
<td>4.000000</td>
<td>97.000000</td>
<td>0.096903</td>
<td>steak</td>
<td>olives</td>
<td>herring</td>
<td>apples</td>
</tr>
<tr>
<td>4.000000</td>
<td>97.000000</td>
<td>0.096903</td>
<td>olives</td>
<td>ice_crea</td>
<td>coke</td>
<td>bourbon</td>
</tr>
</tbody>
</table>
TEMEXPRESS="SAS-expressions"

TEMEXPRESS=file-reference

specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias TE=

TEMPNAMES=variable-name
TEMPNAMES=(variable-list)

specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMEXPRESS= option.

Alias TN=

TRANFMT="format-specification"

specifies the formats for the TRAN= variable. If you do not specify the TRANFMT= option, then the unformatted values of the TRAN= variable are used.

Enclose the format specification in quotation marks.

WEIGHT=variable

specifies the numeric weight variable to use for computing the score of each frequent itemset, along with FREQ= variable. If you do not specify a WEIGHT= variable, then the AGGREGATE=, FREQ=, and ITEMAGG= options are ignored.

Details

Overview

Frequent itemsets are the a priori information in order to mine association rules. These are widely used in market basket analysis, web usage mining, and bio-informatics. Association rules are popular for discovering relations among different values of a variable. Sequence mining aims to discover the causality relationship among items in transactions of customer purchasing habits or anti-money laundry, for instance.

By specifying ITEM=, TRAN=, and optionally TIME=, the server derives either the frequent itemsets, the association rules, the sequences, or any combinations of them. If TIME= is not specified, the server does not generate sequence results. The frequent itemsets, the association rules, and the sequences are stored in separate temporary tables in the server.

Frequent Itemsets Table

The frequent itemsets table is generated when you specify the ITEMSTBL option and it is accessed with the &_tempARMItems_ macro variable. See the following example:

data example.aggdata;
  input customer product $ time price amount product_id;
datalines;
1 e 0    2.49 2 1
1 t 1 2999.00 1 2
1 e 2    2.49 2 1
1 t 3  499.00 1 2
1 e 4    3.49 3 1
1 t 5  199.00 1 2
2 t 0  199.00 1 2
2 e 1    3.49 2 1
2 h 2  50.00 1 3
The preceding statements generate the following output for the sample data:

```
2 e 3    3.49 1 1
2 t 4    499.00 1 2
2 e 5    3.49 1 1
;
run;
```

```
proc imstat data=example.aggdata;
arm item=product tran=customer / maxitems=3 freq=amount
   weight=price itemagg=SUM agg=MIN itemstbl;
run;
table example.&_tempARMItems_
   fetch / orderby=(_SetSize_);
run;
```

The preceding statements generate the following output for the sample data:

```
<table>
<thead>
<tr>
<th><em>SetSize</em></th>
<th><em>Count</em></th>
<th><em>Support</em></th>
<th><em>Score</em></th>
<th>product1</th>
<th>product2</th>
<th>product3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.000000</td>
<td>1.000000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.000000</td>
<td>2.000000</td>
<td>1.000000</td>
<td>13.960000</td>
<td>e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.000000</td>
<td>2.000000</td>
<td>1.000000</td>
<td>698.00000</td>
<td>t</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.000000</td>
<td>1.000000</td>
<td>0.500000</td>
<td>50.000000</td>
<td>h</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.000000</td>
<td>2.000000</td>
<td>1.000000</td>
<td>711.96000</td>
<td>t</td>
<td>e</td>
<td></td>
</tr>
<tr>
<td>2.000000</td>
<td>1.000000</td>
<td>0.500000</td>
<td>63.960000</td>
<td>h</td>
<td>e</td>
<td></td>
</tr>
<tr>
<td>2.000000</td>
<td>1.000000</td>
<td>0.500000</td>
<td>748.00000</td>
<td>t</td>
<td>h</td>
<td></td>
</tr>
<tr>
<td>3.000000</td>
<td>1.000000</td>
<td>0.500000</td>
<td>761.96000</td>
<td>t</td>
<td>h</td>
<td>e</td>
</tr>
</tbody>
</table>
```

The columns in the frequent itemsets table are as follows:

- **_SetSize_**: Shows the number of items in the frequent itemset.
- **_Count_**: Shows the frequency for the frequent itemset in all the transactions.
- **_Support_**: Shows the ratio of the _Count_ value to the number of transactions.
- **_Score_**: Shows the aggregated values of the FREQ= and WEIGHT= values, when they are specified.

Consider the frequent itemset for product `t` (row 3 in the preceding table). It appears 3 times for customer 1 and 2 times for customer 2. First, the server performs item aggregation with each customer. Then, the server performs second stage aggregation to obtain the final score of a frequent itemset. In this case, the intermediate scores of itemset `t` are (1*2999.00 + 1*499.00 + 1*199.00) = 3697.00 and (1*199.00 + 1*499.00) = 698.00. The final score of this itemset is MIN(3697.00, 698.00) = 698.00.
PRODUCTn
Shows the name of an item in the frequent itemset. The column name is variable. The name is based on the column that is specified in the ITEM= option.

**Association Rules Table**
The association rules table is generated when you specify the RULES and RULESTBL options. It is accessed with the &_tempARMRules_ macro variable. For example, the following statements derive association rules of sizes between MINITEMS=3 and MAXITEMS=4. The support range of each frequent itemset is set at LOWER=125 and UPPER=130. The minimal confidence value permitted is 0.8. Each association rule's score has to be greater or equal than 2.

```r
proc imstat data=example.assocs;
    arm item=Product tran=Customer / minItems=3 maxItems=4 itemsTbl
        support(LOWER=125 UPPER=130) weight=TIME
        rules(confidence(LOWER=0.8) score(LOWER=1) weight=TIME) rulesTbl;
run;

table example.&_tempARMRules_;
    fetch _SetSize_ -- _Rule_ / to=10;
run;
```

*Note:* The FETCH statement in the preceding example does not include the values from the ITEM= column in the display.

The preceding statements generate the following output for the Assocs data:

<table>
<thead>
<tr>
<th><em>SetSize</em></th>
<th><em>SetCount</em></th>
<th><em>SetSupport</em></th>
<th><em>SetScore</em></th>
<th><em>Score</em></th>
<th>Confidence</th>
<th>ExpConf</th>
<th><em>Lift</em></th>
<th><em>NumHS</em></th>
<th><em>NumSHS</em></th>
<th><em>Rule</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>3.000000</td>
<td>127.000000</td>
<td>0.129637</td>
<td>578.000000</td>
<td>578.000000</td>
<td>0.030000</td>
<td>0.407512</td>
<td>1.790655</td>
<td>2.000000</td>
<td>1.000000</td>
<td>baguette, soda --&gt; cracker</td>
</tr>
<tr>
<td>3.000000</td>
<td>129.000000</td>
<td>0.130714</td>
<td>466.000000</td>
<td>466.000000</td>
<td>0.843137</td>
<td>0.405514</td>
<td>1.736955</td>
<td>2.000000</td>
<td>1.000000</td>
<td>baguette, soda --&gt; herring</td>
</tr>
<tr>
<td>3.000000</td>
<td>127.000000</td>
<td>0.129637</td>
<td>722.000000</td>
<td>722.000000</td>
<td>0.909743</td>
<td>0.472527</td>
<td>1.919767</td>
<td>2.000000</td>
<td>1.000000</td>
<td>bourbon, turkey --&gt; olives</td>
</tr>
<tr>
<td>3.000000</td>
<td>125.000000</td>
<td>0.125074</td>
<td>1137.000000</td>
<td>1137.000000</td>
<td>0.840000</td>
<td>0.590401</td>
<td>1.401400</td>
<td>2.000000</td>
<td>1.000000</td>
<td>cracker, ham --&gt; heineken</td>
</tr>
<tr>
<td>4.000000</td>
<td>125.000000</td>
<td>0.124975</td>
<td>1165.000000</td>
<td>1165.000000</td>
<td>0.976563</td>
<td>0.590401</td>
<td>1.626232</td>
<td>3.000000</td>
<td>1.000000</td>
<td>baguette, cracker, herring --&gt; heineken</td>
</tr>
<tr>
<td>4.000000</td>
<td>125.000000</td>
<td>0.124975</td>
<td>1165.000000</td>
<td>1165.000000</td>
<td>0.852068</td>
<td>0.495514</td>
<td>1.777578</td>
<td>3.000000</td>
<td>1.000000</td>
<td>baguette, cracker, heineken --&gt; herring</td>
</tr>
<tr>
<td>4.000000</td>
<td>126.000000</td>
<td>0.125574</td>
<td>1098.000000</td>
<td>1098.000000</td>
<td>0.911043</td>
<td>0.405712</td>
<td>1.672862</td>
<td>3.000000</td>
<td>1.000000</td>
<td>baguette, heineken, soda --&gt; cracker</td>
</tr>
<tr>
<td>4.000000</td>
<td>126.000000</td>
<td>0.125574</td>
<td>1098.000000</td>
<td>1098.000000</td>
<td>0.902126</td>
<td>0.590401</td>
<td>1.655197</td>
<td>3.000000</td>
<td>1.000000</td>
<td>baguette, cracker, soda --&gt; heineken</td>
</tr>
<tr>
<td>4.000000</td>
<td>126.000000</td>
<td>0.125574</td>
<td>1098.000000</td>
<td>1098.000000</td>
<td>0.956986</td>
<td>0.517682</td>
<td>2.739329</td>
<td>3.000000</td>
<td>1.000000</td>
<td>baguette, cracker, heineken --&gt; soda</td>
</tr>
<tr>
<td>4.000000</td>
<td>126.000000</td>
<td>0.125574</td>
<td>1098.000000</td>
<td>1098.000000</td>
<td>0.823629</td>
<td>0.369654</td>
<td>2.252330</td>
<td>2.000000</td>
<td>2.000000</td>
<td>baguette, soda --&gt; cracker, heineken</td>
</tr>
</tbody>
</table>

The columns in the association rules table are as follows:

**_SetSize_**
Shows the number of items in the frequent itemset.

**_SetCount_**
Shows the frequency for the frequent itemset that contain the rule in all the transactions.

**_SetSupport_**
Shows the ratio of the _SetCount_ value to the total number of transactions.

**_SetScore_**
Shows the aggregated values WEIGHT= values of all the frequent itemsets, when they are specified. The AGGREGATOR=SUM and ITEMAGG= option defaults to SUM.

**_Score_**
Shows the aggregated values of the WEIGHT= value in the association rule suboptions. The AGGREGATOR=MAX and ITEMAGG= option defaults to SUM.
**Confidence**
Shows the confidence for the association rule.

**ExpConf**
Shows the expected confidence for the association rule.

**Lift**
Shows the lift for the association rule.

**NumLHS**
Shows the number of items in the left-hand-side of a rule.

**NumRHS**
Shows the number of items in the right-hand-side of a rule.

**Rule**
Shows the full string of the rule.

### Sequences Table
The sequences table is generated when you specify the SEQUENCES and SEQUESTBL options. For example, the following statements derive association rules of sizes between MINITEMS=3 and MAXITEMS=4. The support range of each frequent itemset is set at LOWER=125 and UPPER=130. The minimal confidence value permitted is 0.8. Each association rule's score has to be greater or equal than 2.

```sql
proc imstat data=example.assocs;
  arm item=Product tran=Customer / maxItems=3
  sequences(time=time minItems=3 maxItems=3 support(lower=110 upper=120))
    sequestbl;
run;
```

The preceding statements generate the following output for the Assocs data:

<table>
<thead>
<tr>
<th><em>ChainLength</em></th>
<th><em>Count</em></th>
<th><em>Support</em></th>
<th><em>Probability</em></th>
<th><em>LiftProduct</em></th>
<th>PRODUCT1</th>
<th>Separator1</th>
<th>PRODUCT2</th>
<th>Separator2</th>
<th>PRODUCT3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.000000</td>
<td>111.000000</td>
<td>0.110899</td>
<td>0.166300</td>
<td>1.962692</td>
<td>artchok</td>
<td>==&gt;</td>
<td>heineken</td>
<td>==&gt;</td>
<td>ham</td>
</tr>
<tr>
<td>3.000000</td>
<td>111.000000</td>
<td>0.110899</td>
<td>0.363934</td>
<td>3.293705</td>
<td>avocado</td>
<td>==&gt;</td>
<td>artchok</td>
<td>==&gt;</td>
<td>ham</td>
</tr>
<tr>
<td>3.000000</td>
<td>112.000000</td>
<td>0.118544</td>
<td>0.198384</td>
<td>1.689384</td>
<td>avocado</td>
<td>==&gt;</td>
<td>heineken</td>
<td>==&gt;</td>
<td>ham</td>
</tr>
<tr>
<td>3.000000</td>
<td>114.000000</td>
<td>0.130696</td>
<td>0.190309</td>
<td>0.951203</td>
<td>baguette</td>
<td>==&gt;</td>
<td>cracker</td>
<td>==&gt;</td>
<td>heineken</td>
</tr>
<tr>
<td>3.000000</td>
<td>113.000000</td>
<td>0.12087</td>
<td>0.232510</td>
<td>1.949060</td>
<td>baguette</td>
<td>==&gt;</td>
<td>hering</td>
<td>==&gt;</td>
<td>artchok</td>
</tr>
<tr>
<td>3.000000</td>
<td>118.000000</td>
<td>0.117920</td>
<td>0.242718</td>
<td>1.709705</td>
<td>baguette</td>
<td>==&gt;</td>
<td>hering</td>
<td>==&gt;</td>
<td>avocado</td>
</tr>
<tr>
<td>3.000000</td>
<td>113.000000</td>
<td>0.112687</td>
<td>0.353346</td>
<td>1.961280</td>
<td>baguette</td>
<td>==&gt;</td>
<td>soda</td>
<td>==&gt;</td>
<td>cracker</td>
</tr>
<tr>
<td>3.000000</td>
<td>113.000000</td>
<td>0.112687</td>
<td>0.353346</td>
<td>1.513848</td>
<td>baguette</td>
<td>==&gt;</td>
<td>soda</td>
<td>==&gt;</td>
<td>heineken</td>
</tr>
<tr>
<td>3.000000</td>
<td>113.000000</td>
<td>0.112687</td>
<td>0.353346</td>
<td>1.869848</td>
<td>baguette</td>
<td>==&gt;</td>
<td>soda</td>
<td>==&gt;</td>
<td>hering</td>
</tr>
<tr>
<td>3.000000</td>
<td>111.000000</td>
<td>0.110899</td>
<td>0.227459</td>
<td>1.795377</td>
<td>bourbon</td>
<td>==&gt;</td>
<td>cracker</td>
<td>==&gt;</td>
<td>chicken</td>
</tr>
</tbody>
</table>

The columns in the association rules table are as follows:

**_ChainLength_**
Shows the number of items in the sequence.

**_Count_**
Shows the frequency of transactions that contain the sequence.
Support
Shows the ratio of the _Count_ value to the total number of transactions.

Probability
Is defined as \( \text{Pr}(A \rightarrow B \rightarrow C \rightarrow D) = \frac{N(A \rightarrow B \rightarrow C \rightarrow D)}{N(A, B, C, D)} \) where \( N() \) is the count function.

LiftProduct
Is defined as \( \text{Lift}(A \rightarrow B \rightarrow C \rightarrow D) = \frac{N(A \rightarrow B \rightarrow C \rightarrow D)}{N_{\text{trans}}(A) \cdot N_{\text{trans}}(B) \cdot N_{\text{trans}}(C) \cdot N_{\text{trans}}(D)} \) where \( N_{\text{trans}} \) is the number of transactions.

Separator
Shows the relationship of the items to the left and right. "==>" indicates that the item on the left occurs before the item on the right. "&" indicates that the two items are considered to happen at the same time.

### ASSESS Statement

The ASSESS statement is used to assess one model or several models. For a set of classification models, the ASSESS statement returns three types of assessments: lift-related assessments, assessments related to a receiver operating characteristic (ROC), and concordance statistics. For a set of regression models, the ASSESS statement returns the summary statistics of the response variable for each bin of the predictions after a quantile binning of the predictions.

#### Syntax

```
ASSESS <variable-list> / Y=response-variable <options>;
```

#### Required Argument

Y=response-variable
specifies the response variable for model assessment.

Alias RESPONSE=

#### ASSESS Statement Options

CUTSTEP=n
specifies a number between 0 and 1 that defines the step size in receiver operating characteristic (ROC) calculations.

Alias STEP=

DESCENDING
specifies that the levels of the GROUPBY variables are to be arranged in descending order.

Alias DESC

EPSILON=e
specifies the tolerance that is used in determining the convergence of the iterative algorithm for the percentile calculation.
EVENT="quoted-strings"

specifies the formatted value of the response variable that represents the event. When this option is not specified, the ASSESS statement performs model assessment for a regression model and the response variable must be numeric.

FORMATS=("format-specification",...)

specifies the formats for the GROUPBY= variables. If you do not specify the FORMAT= option, or if you do not specify the GROUPBY= option, the default format is applied for that variable.

Enclose each format specification in quotation marks and separate each format specification with a comma.

GROUPBY=(variable-list)

specifies a list of variable names, or a single variable name, to use as GROUPBY variables in the order of the grouping hierarchy. If you do not specify any GROUPBY variable names, then the calculation is performed across the entire table—possibly subject to a WHERE clause.

GROUPBYLIMIT=n

specifies the maximum number of levels in a GROUPBY set. When the software determines that there are at least n levels in the GROUPBY set, it abandons the action, returns a message, and does not produce a result set. You can specify the GROUPBYLIMIT= option if you want to avoid creating excessively large result sets in GROUPBY operations.

GROUPFILTER=(groupfilter-options)

specifies a section of the GROUPBY= hierarchy to include in the ASSESS computation.

MAXITER=i

specifies a positive integer that determines the maximum number of iterations for the percentile algorithm.

Default \(5 \times \) the number of bins (NBINS= option).

MERGEBINS=b

specifies the number of bins to create when a numeric GROUPBY variable exceeds the MERGELIMIT=n specification. If you specify a MERGELIMIT, but do not specify a value for the MERGEBINS= option, the server automatically calculates the number of bins.

MERGELIMIT=n

specifies that when the number of unique values in a numeric GROUPBY variable exceeds n, the variable is automatically binned and the GROUPBY structure is determined based on the binned values of the variable, rather than the unique formatted values.

For example, if you specify MERGELIMIT=500, any numeric GROUPBY variable with more than 500 unique formatted values is binned. Instead of returning results for more than 500 groups, the results are returned for the bins. You can specify the number of bins with the MERGEBINS= option.

NBINS=n

specifies the number of bins to use in the lift calculations.

NOMISSING

specifies that you do not want to include missing values in the determination of Group-By values. By default, levels with missing values are included.
PARTITION \(<=\text{partition-key}\>

When you specify this option and the table is partitioned, the results are calculated separately for each value of the partition key. In other words, the partition variables function as automatic GROUPBY variables. This mode of executing calculations by partition is more efficient than using the GROUPBY= option. With a partitioned table, the server takes advantage of knowing that observations for a partition cannot be located on more than one worker node.

If you do not specify a partition-key, the analysis is performed for all partitions. If you do specify a partition-key, the analysis is carried out for the specified key value only. You can use the PARTITIONINFO statement to retrieve the valid partition key values for a table.

You can specify a partition-key in two ways. You can supply a single quoted string that is passed to the server, or you can specify the elements of a composite key separated by commas. For example, if you partition a table by variables GENDER and AGE, with formats $1 and BEST12, respectively, then the composite partition key has a length of 13. You can specify the partition for the 11-year-old females as follows:

\[
\text{statement} / \text{partition}=\text{"F          11"}; /* passed directly to the server */
\text{statement} / \text{partition}=\text{"F","11"}; /* composed by the procedure */
\]

If you choose the second format, the procedure composes a key based on formatting information from the server.

Alias PART=

RAWORDER

specifies that the ordering of the GROUPBY variables is based on the raw values of the variables instead of the formatted values.

SAVE=\text{table-name}

saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

SETSIZE

requests that the server estimate the size of the result set. The procedure does not create a result table if the SETSIZE option is specified. Instead, the procedure reports the number of rows that are returned by the request and the expected memory consumption for the result set (in KB). If you specify the SETSIZE option, the SAS log includes the number of observations and the estimated result set size. See the following log sample:

\[
\text{NOTE: The LASR Analytic Server action request for the STATEMENT statement would return 17 rows and approximately 3.641 kBytes of data.}
\]

The typical use of the SETSIZE option is to get an estimate of the size of the result set in situations where you are unsure whether the SAS session can handle a large result set. Be aware that in order to determine the size of the result set, the server has to perform the work as if you were receiving the actual result set. Requesting the estimated size of the result set does consume resources on the server. The estimated number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.
TEMPEXPRESS="SAS-expressions"
TEMPEXPRESS=file-reference
  specifies either a quoted string that contains the SAS expression that defines the
temporary variables or a file reference to an external file with the SAS statements.

Alias  TE=

TEMPNAMES=variable-name
TEMPNAMES=(variable-list)
  specifies the list of temporary variables for the request. Each temporary variable
must be defined through SAS statements that you supply with the TEMPEXPRESS=
option.

Alias  TN=

TEMPTABLE
generates an in-memory temporary table from the result set. The IMSTAT procedure
displays the name of the table and stores it in the &_TEMPLAST_ macro variable,
provided that the statement executed successfully.

When the IMSTAT procedure exits, all temporary tables created during the IMSTAT
session are removed. Temporary tables are not displayed on a TABLEINFO request,
unless the temporary table is the active table for the request.

YFORMAT="quoted-string"
  specifies the format for the response variable. This format produces the event
specified in the EVENT= option.

Alias  YFMT=

Details
You can compare multiple models by specifying predicted values from those models in
the variable-list. You can compare models in different data segments with the
GROUPBY= option. Note that you must specify the response variable for the ASSESS
statement.

When variable-list is not provided, assessment statistics are computed for all numerical
variables in the active table.

BOXPLOT Statement
The BOXPLOT statement generates a table with the information that can be used to generate a box plot. It
does not generate the plot.

Examples:
  “Example 2: Retrieving Box Values” on page 174
  “Example 3: Retrieving Box Plot Values with the NOUTLIERLIMIT= Option” on page
175

Syntax

BOXPLOT <variable-list> </options> ;
Optional Argument

variable-list
specifies one or more numeric variables. If you do not specify this option, then all numeric variables in the table are used.

BOXPLOT Statement Options

DESCENDING
specifies that the levels of the GROUPBY variables are to be arranged in descending order.

Alias  DESC

FORMATS="(format-specification",...)
specifies the formats for the GROUPBY= variables. If you do not specify the FORMATS= option, or if you omit the entry for a GROUPBY variable, the default format is applied for that variable.

Enclose each format specification in quotation marks and separate each format specification with a comma.

Example
proc imstat data=lasr1.table1;
  boxplot x / groupby=(a b) formats=('8.3', '$10');
quit;

GROUPBY=(variable-list)
specifies a list of variable names, or a single variable name, to use as GROUPBY variables in the order of the grouping hierarchy. If you do not specify any GROUPBY variable names, then the calculation is performed across the entire table —possibly subject to a WHERE clause.

GROUPBYLIMIT=n
specifies the maximum number of levels in a GROUPBY set. When the software determines that there are at least \( n \) levels in the GROUPBY set, it abandons the action, returns a message, and does not produce a result set. You can specify the GROUPBYLIMIT= option if you want to avoid creating excessively large result sets in GROUPBY operations.

MERGEBINS=b
specifies the number of bins to create when a numeric GROUPBY variable exceeds the MERGELIMIT= \( n \) specification. If you specify a MERGELIMIT, but do not specify a value for the MERGEBINS= option, the server automatically calculates the number of bins.

MERGELIMIT=n
specifies that when the number of unique values in a numeric GROUPBY variable exceeds \( n \), the variable is automatically binned and the GROUPBY structure is determined based on the binned values of the variable, rather than the unique formatted values.

For example, if you specify MERGELIMIT=500, any numeric GROUPBY variable with more than 500 unique formatted values is binned. Instead of returning results for more than 500 groups, the results are returned for the bins. You can specify the number of bins with the MERGEBINS= option.

DESCENDING
specifies that the levels of the GROUPBY variables are to be arranged in descending order.
Alias  DESC

**NOUTLIERBINS=b**
specifies the number of bins for reporting outliers. The default number of bins is 10 if you do not specify an **NOUTLIERBINS=** value, but do specify the OUTLIERS option. Specifying a nonzero value for **NOUTLIERBINS=** implies the specification of the OUTLIERS option.

Alias  NOUTBINS=

Default  10

**NOUTLIERLIMIT=k**
specifies the largest number of outliers to be returned. If you request outliers with the OUTLIERS option, and you specify a **NOUTLIERLIMIT=** value, the actual outliers are being returned rather than the binned values. Specifying a nonzero value for **NOUTLIERLIMIT=** implies the specification of the OUTLIERS option.

Alias  NOUTLIMIT=

**OUTLIERS**
specifies to include outliers in computations and results. If the **NOUTLIMIT=n** option is specified, then the server returns up to *n* outliers on the high and low ends of the distribution. Otherwise, outliers are binned into **NOUTLIERBINS=b** bins.

**PARTITION <partition-key>**

When you specify this option and the table is partitioned, the results are calculated separately for each value of the partition key. In other words, the partition variables function as automatic GROUPBY variables. This mode of executing calculations by partition is more efficient than using the GROUPBY= option. With a partitioned table, the server takes advantage of knowing that observations for a partition cannot be located on more than one worker node.

If you do not specify a *partition-key*, the analysis is performed for all partitions. If you do specify a *partition-key*, the analysis is carried out for the specified key value only. You can use the PARTITIONINFO statement to retrieve the valid partition key values for a table.

You can specify a *partition-key* in two ways. You can supply a single quoted string that is passed to the server, or you can specify the elements of a composite key separated by commas. For example, if you partition a table by variables GENDER and AGE, with formats $1 and BEST12, respectively, then the composite partition key has a length of 13. You can specify the partition for the 11-year-old females as follows:

```
statement / partition="F          11"; /* passed directly to the server */
statement / partition="F","11";        /* composed by the procedure */
```

If you choose the second format, the procedure composes a key based on formatting information from the server.

Alias  PART=

**RAWORDER**
specifies that the ordering of the GROUPBY variables is based on the raw values of the variables instead of the formatted values.
SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

SETSIZE
requests that the server estimate the size of the result set. The procedure does not create a result table if the SETSIZE option is specified. Instead, the procedure reports the number of rows that are returned by the request and the expected memory consumption for the result set (in KB). If you specify the SETSIZE option, the SAS log includes the number of observations and the estimated result set size. See the following log sample:

NOTE: The LASR Analytic Server action request for the STATEMENT statement would return 17 rows and approximately 3.641 kBytes of data.

The typical use of the SETSIZE option is to get an estimate of the size of the result set in situations where you are unsure whether the SAS session can handle a large result set. Be aware that in order to determine the size of the result set, the server has to perform the work as if you were receiving the actual result set. Requesting the estimated size of the result set does consume resources on the server. The estimated number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.

TEMPEXPRESS="SAS-expressions"
TEMPEXPRESS=file-reference
specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias TE=

TEMPNAMES=variable-name
TEMPNAMES=(variable-list)
specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

Alias TN=

---

CLUSTER Statement

The CLUSTER statement can be used to perform a \( k \)-means cluster analysis that uses the Euclidean distance between values or it can use a density-based algorithm—DBSCAN—that was originally developed to discover clusters in large spatial databases with noise.

Example: “Example 4: Performing a Cluster Analysis” on page 176

Syntax

CLUSTER <variable-list> </options> ;
**Optional Argument**

`variable-list` specifies a list of variables. If you do not specify this option, then all the variables in the table are used.

For the $k$-means clustering, the variable list defines a vector for each observation that is used to compute the Euclidean distance between the observations. By minimizing the within-cluster sum of squares (WCSS) of this distance, a set of clusters and their centers are determined.

**CLUSTER Statement Options**

_BUBMAXPTS=*_n_

specifies the maximum number of points in each bubble and must exceed the value specified in the _BUBMINPTS= _option.

 Applies to   _METHOD=DBSCAN_

_BUBMINPTS=*_n_

specifies the minimum number of points in each bubble.

Default 1

 Applies to  _METHOD=DBSCAN_

_CLUSTINFO_

specifies to save the cluster center that each observation belongs to, and the distance between them, into the temporary table.

Interaction You must specify the TEMPTABLE option along with this option.

_CONV=*_c_*
specifies the convergence criterion $c$ for the $k$-means analysis. When the relative change in within-cluster sum of squares (WCSS) between successive iterations is less than $c$, the analysis is presumed to have converged.

Default $1e-05$

 Applies to _METHOD=KMEANS_

_DISP=(_variable-list)_

_DISP=variable-name_ specifies the variable or variables to include in the clustering results. Use this option with the _NSAMP=_ option to generate output that is suitable for graphing.

_DIST= EUC | SQUAREDEUC | MANHATTAN | MAXIMUM | COSINE | JACCARD | HAMMING_ specifies the distance measure used in the DBSCAN method. The $k$-means method uses DIST=EUC.

_DMAX=*_v_*
specifies the maximum diameter of bubbles with the specified DIST= distance measure.

Default 0
EPS=$r$
specifies the distance value for neighborhood querying in the DBSCAN method. You must specify a value for $r$ when METHOD=DBSCAN. There is no default value since $r$ is a distance measurement and depends on the range of the data.

The values of EPS= and MINPTS= are important for the number of clusters that DBSCAN can find. The EPS= value determines the minimal radius of the clusters in terms of the distance measurement. (See the DIST= on page 64 option.) The value of MINPNTS=$n$ suggests that data points in a cluster with less than $n$ observations are noise.

FREQ=(variable-list)
FREQ=variable-name
specifies the variable or variables that are used to perform the frequency analysis for each cluster. The procedure generates separate output tables for each variable.

INITMETHOD=FORGY | RAND | AVG
specifies the method for obtaining the initial estimate of cluster assignment. For the INIT=FORGY partition method, the initial mean centers are computed from NUMCLUS $\times n_t \times n_n$ observations where $n_t$ and $n_n$ are the number of threads and number of nodes used by the server.

When you specify INIT=RAND, all methods are assigned randomly to one of the NUMCLUS clusters. When you specify INIT=AVG, the initial centers are formed by averaging the observations on a thread-by-thread basis.

Alias INIT=
Default FORGY

MAXITER=$i$
specifies a positive integer that determines the maximum number of iterations for clustering.

Default 10

METHOD=KMEANS | DBSCAN
specifies the clustering method for the analysis.

Default KMEANS

MINPTS=$n$
specifies the minimum number of points required in one cluster for the DBSCAN method. The EPS= and MINPTS= options can have a dramatic effect on the clusters that are generated with the DBSCAN method. You should exercise care in specifying the values for these options.

Default 1, which means that any cluster should contain at least one data point.

NOCASE
specifies that the comparison between terms and the values of character variables is case insensitive. By default, comparisons are case sensitive.

NOIDF
specifies that only the term frequency is used to construct the vectors, and not the inverse document frequency.

NONORM
specifies that the term frequency-inverse document frequency (TF-IDF) vectors are not normalized.
NOPREPARSE
prevents the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects problems with the code, the error messages might not to be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

When you specify this option in the PROC IMSTAT statement, the option applies to all statements that can generate code. You can also exclude specific statements from pre-parsing by using the NOPREPARSE option in statements that allow temporary columns or the SCORE statement.

Alias NOPREP

NREP=k
specifies the number of representative points for each bubble.

Default 1

NSAMP=k
specifies the number of sample points that are returned for each cluster. Note that this returns the $k$ nearest and $k$ farthest points from the cluster centers including their distances.

Default 0

NUMCLUSTERS=number
specifies the number of clusters for the $k$-means analysis.

Alias NUMCLUS=

Default 2

NSAMP=k
specifies the number of sample points to return for each cluster. This option returns the $k$ nearest and $k$ farthest points from the cluster centers, including their Euclidean distances.

Default 0

SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

SAVETERMS
specifies to save the TF-IDF vectors into the temporary table along with clustering results.

Interaction You must specify the TEMPTABLE option along with this option.
SAVEVECTORS

specifies to save the distance vectors to the temporary table along with clustering results. When the VARS= option on page 68 already includes the variables used to obtain the distance vectors, these variables are not saved again.

**Alias**

SAVEVEC

**Interaction**

You must specify the TEMPTABLE option along with this option.

SEED=number

specifies the random number seed to use when the initialization methods INIT=FORGY or INIT=RAND are also specified. Specifying a nonzero number results in a reproducible random number stream for the specific combination of number of threads and number of worker nodes in the server.

SETSZE

requests that the server estimate the size of the result set. The procedure does not create a result table if the SETSIZE option is specified. Instead, the procedure reports the number of rows that are returned by the request and the expected memory consumption for the result set (in KB). If you specify the SETSIZE option, the SAS log includes the number of observations and the estimated result set size. See the following log sample:

```
NOTE: The LASR Analytic Server action request for the STATEMENT statement would return 17 rows and approximately 3.641 kBytes of data.
```

The typical use of the SETSIZE option is to get an estimate of the size of the result set in situations where you are unsure whether the SAS session can handle a large result set. Be aware that in order to determine the size of the result set, the server has to perform the work as if you were receiving the actual result set. Requesting the estimated size of the result set does consume resources on the server. The estimated number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.

TEMPEXPRESS="SAS-expressions"

specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

**Alias**

TE=

TEMPNAMES=variable-name

TEMPNAMES=(variable-list)

specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

**Alias**

TN=

TEMPTABLE

generates an in-memory temporary table from the result set. The IMSTAT procedure displays the name of the table and stores it in the &_TEMPLAST_ macro variable, provided that the statement executed successfully.
When the IMSTAT procedure exits, all temporary tables created during the IMSTAT session are removed. Temporary tables are not displayed on a TABLEINFO request, unless the temporary table is the active table for the request.

TERMS="term1", "term2", ...

specifies the terms used in computing term frequency. Each string represents one term. Specifying the TERMS= option triggers the use of TF-IDF to compute distance vectors for character variables. The TERMS= option is useful if you have a fairly small number of terms to pass to the server. If the number of terms is large, you might want to use the TERMDATA= option instead.

TERMDATA=table-name

specifies an in-memory table in the server that contains the term list. Specifying the TERMDATA= option triggers the use of TF-IDF to compute distance vectors for character variables. If you specify both the TERMS= and TERMDATA= options, the server uses the union of the two sets in the TF-IDF calculation. Note that the IMSTAT procedure assumes that the TERMDATA= table already exists in the server.

TIMEOUT=seconds

specifies the maximum number of seconds that the server should run the statement. If the time-out is reached, the server terminates the request and generates an error and error message. By default, there is no time-out.

TOKENS="token1", "token2", ...

specifies the tokens that separate terms when scanning character variables. If you do not specify the tokens (term delimiters), then the terms are compared against the full raw length of the character variable.

For example, if the term list is "better", "worse", and the variable Opinion contains "Recent news is better than last week," then without a token list that contains the " " (blank) delimiter, the term "better" is not counted. In other words, the absence of the blank token prevents the TF-IDF from scanning "better". If your token list is large, you might want to use the TOKENDATA= option instead.

TOKENDATA=table-name

specifies an in-memory table in the server that contains the tokens list.

VARS=variable-name

VARS=(variable-name1, variable-name2, ...)

specifies the names of the variables to transfer to a temporary table in the server. This option is ignored unless you score an in-memory table and the TEMPTABLE option is specified. The observations with these variables are copied to the generated temporary table.

Details

Two clustering methods are implemented in the CLUSTER statement:

- The default clustering method is $k$-means clustering. For this method, the optional list of variables defines a vector for each observation, which is used to compute the Euclidean distance between the observations. By minimizing the within-cluster sum of squares (WCSS) of this distance, a set of clusters and their centers can be determined.

- You can also use a density-based algorithm, DBSCAN. It was originally developed to discover clusters in large spatial databases with noise, and was published in the proceedings of KDD 1996. You do not need to supply the number of clusters for the DBSCAN method. The bubbling scheme is designed specifically to improve the DBSCAN performance for large data sets. Bubbling is disabled by default, because the default value of the DMAX= option is 0.
The statement also supports TF-IDF (term frequency-inverse document frequency) to compute distance vectors for character variables. For any observation with TF-IDF, the total length of the distance vector is given by the number of numerical variables plus the number of terms. The following options relate to TF-IDF calculations: TERMS=, TERMDATA=, TOKENS=, TOKENDATA=, NOCASE, NOIDF, NONORM, and SAVETERMS. You trigger TF-IDF calculations by specifying the TERMS= or TERMDATA= options in the CLUSTER statement.

**CORR Statement**

The CORR statement is used to calculate a matrix of pairwise correlations of numeric variables in an in-memory table.

**Example:**

“Example 5: Performing a Pairwise Correlation” on page 177

**Syntax**

CORR <variable-list> <options>;

**Optional Argument**

variable-list

specifies one or more numeric variables. If you do not specify this option, then all numeric variables in the table are used.

**CORR Statement Options**

SAVE=table-name

saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

TEMPEXPRESS="SAS-expressions"

specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias  TE=

TEMPNAMES=variable-name

TEMPNAMES=(variable-list)

specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

Alias  TN=

**CROSSTAB Statement**

The CROSSTAB statement is used to calculate one and two dimensional tables. You can use GROUPBY= variables, partitioned tables, a WEIGHT= variable to calculate multiple statistics for each table cell, and calculate marginal versions of the statistics as well.
Example: “Example 6: Crosstabulation with Measures of Association and Chi-Square Tests” on page 178

Syntax

CROSSTAB row-variable </ options>;  
CROSSTAB row*column </ options>;

Required Argument

row-variable
specifies to create a one-dimensional table using the specified variable.

Optional Argument

row
column
specifies to create a two-dimensional table based on the two variables.

CROSSTAB Statement Options

ACROSSBY
specifies that the levels of row and column variables are the same across the GROUPBY= variables. If you specify this option, then the tables in each group are shown with the same row and column layout. If the ACROSSBY option is not specified, then the particular row and column levels might be different among the groups based on which values of the row and column variables occur in each group.

Alias ACROSS

AGGREGATE=(statistic(s))

specifies the statistics to use as aggregation methods for which crosstabulations are computed when a WEIGHT variable is also specified. If no WEIGHT variable is specified, then the N aggregator is applied. In other words, the crosstabulation shows the frequency with which the values occur when no WEIGHT variable is specified. If you specify multiple aggregation methods, then the server computes a crosstabulation for each method.

The available aggregation methods are as follows:

CSS corrected sum of squares
CV coefficient of variation
MAX maximum value
MEAN arithmetic mean
MIN minimum value
N number of observations
PROBT p-value for the t-statistic
STD standard deviation
STDERR standard error
SUM sum of the nonmissing values
TSTAT t-statistic for the null hypothesis that the mean equals zero
USS uncorrected sum of squares
VAR sample variance

Alias AGG=

ASSOCIATION specifies to calculate the measures of association between the row and column variable of the cross tabulation. The option generates the following measures: the Gamma statistic, Kendall’s Tau-b, Stuart’s Tau-c, Somers’ measures, Lambda measures, and uncertainty measures.

Alias MEASURES

CHISQ computes Chi-square statistics for the test of independence of the row and column variables and their asymptotic p-values. The option calculates the Pearson Chi-square statistic as well as the likelihood-ratio test.

COLBINS=b specifies the number of bins to use in binning the column variable.

Alias XBINS=

DESCENDING specifies that the levels of the GROUPBY variables are to be arranged in descending order.

Alias DESC

FORMATS=("format-specification", ...) specifies the formats for the GROUPBY= variables. If you do not specify the FORMATS= option, or if you omit the entry for a GROUPBY variable, the default format is applied for that variable.

Enclose each format specification in quotation marks and separate each format specification with a comma.

Example proc imstat data=lasr1.table1;
   crosstab x*y / groupby=(a b) formats=("8.3", "$10");
   quit;

GROUPBY=(variable-list) specifies a list of variable names, or a single variable name, to use as GROUPBY variables in the order of the grouping hierarchy. If you do not specify any GROUPBY variable names, then the calculation is performed across the entire table—possibly subject to a WHERE clause.

GROUPBYLIMIT=n specifies the maximum number of levels in a GROUPBY set. When the software determines that there are at least n levels in the GROUPBY set, it abandons the action, returns a message, and does not produce a result set. You can specify the GROUPBYLIMIT= option if you want to avoid creating excessively large result sets in GROUPBY operations.

MARGINS= ROW | COL | ALL | NONE specifies whether to calculate marginal values in addition to the crosstabulation. The default is MARGINS=None, which specifies that no marginal values are calculated. The MARGINS=ROW setting calculates margins for the row variable
only. The MARGINS=COL setting calculates margins for the column variable only. The MARGINS=ALL setting calculates row margins, column margins, and the overall margin.

These calculations are repeated for each aggregate method in the CROSSTAB request.

Default  NONE

MERGEBINS=b
specifies the number of bins to create when a numeric GROUPBY variable exceeds the MERGELIMIT=n specification. If you specify a MERGELIMIT, but do not specify a value for the MERGEBINS= option, the server automatically calculates the number of bins.

MERGELIMIT=n
specifies that when the number of unique values in a numeric GROUPBY variable exceeds n, the variable is automatically binned and the GROUPBY structure is determined based on the binned values of the variable, rather than the unique formatted values.

For example, if you specify MERGELIMIT=500, any numeric GROUPBY variable with more than 500 unique formatted values is binned. Instead of returning results for more than 500 groups, the results are returned for the bins. You can specify the number of bins with the MERGEBINS= option.

NOEMPTY
specifies that empty cells are not returned to the SAS session (only full cells are returned). When this option is specified, the server arranges the results into a vector of nonzero values with row and column indices. This sparse storage is memory efficient when the table has many empty cells.

Alias  FULLCELL

DESCENDING
specifies that the levels of the GROUPBY variables are to be arranged in descending order.

Alias  DESC

NOPREPARSE
prevents the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects problems with the code, the error messages might not be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

When you specify this option in the PROC IMSTAT statement, the option applies to all statements that can generate code. You can also exclude specific statements from pre-parsing by using the NOPREPARSE option in statements that allow temporary columns or the SCORE statement.

Alias  NOPREP
NOTE MPPART

specifies that the temporary table generated by the TEMPTABLE option is not
partitioned by the GROUPBY= variables. When you request a temporary table with
the CROSSTAB statement, by default, the server creates a partitioned table. When
the number of groups is large, this can result in many small partitions, and requires
extra memory resources to store the partition information for the temporary table. By
specifying this option, the temporary table is organized similarly to the default table,
but is not a partitioned table.

Alias NOTP

PARTITION <partition-key>

When you specify this option and the table is partitioned, the results are calculated
separately for each value of the partition key. In other words, the partition variables
function as automatic GROUPBY variables. This mode of executing calculations by
partition is more efficient than using the GROUPBY= option. With a partitioned
table, the server takes advantage of knowing that observations for a partition cannot
be located on more than one worker node.

If you do not specify a partition-key, the analysis is performed for all partitions. If
you do specify a partition-key, the analysis is carried out for the specified key value
only. You can use the PARTITIONINFO statement to retrieve the valid partition key
values for a table.

You can specify a partition-key in two ways. You can supply a single quoted string
that is passed to the server, or you can specify the elements of a composite key
separated by commas. For example, if you partition a table by variables GENDER
and AGE, with formats $1 and BEST12, respectively, then the composite partition
key has a length of 13. You can specify the partition for the 11-year-old females as
follows:

```
statement / partition="F 11"; /* passed directly to the server */
statement / partition="F","11"; /* composed by the procedure */
```

If you choose the second format, the procedure composes a key based on formatting
information from the server.

Alias PART=

RAWORDER

specifies that the ordering of the GROUPBY variables is based on the raw values of
the variables instead of the formatted values.

ROWBINS=n

specifies the number of bins to use for binning the row variable.

Alias YBINS=

SAVE=table-name

saves the result table so that you can use it in other IMSTAT procedure statements
like STORE, REPLAY, and FREE. The value for table-name must be unique within
the scope of the procedure execution. The name of a table that has been freed with
the FREE statement can be used again in subsequent SAVE= options.

SETSIZE

requests that the server estimate the size of the result set. The procedure does not
create a result table if the SETSIZE option is specified. Instead, the procedure reports
the number of rows that are returned by the request and the expected memory
consumption for the result set (in KB). If you specify the SETSIZE option, the SAS
log includes the number of observations and the estimated result set size. See the following log sample:

NOTE: The LASR Analytic Server action request for the STATEMENT statement would return 17 rows and approximately 3.641 kBytes of data.

The typical use of the SETSIZE option is to get an estimate of the size of the result set in situations where you are unsure whether the SAS session can handle a large result set. Be aware that in order to determine the size of the result set, the server has to perform the work as if you were receiving the actual result set. Requesting the estimated size of the result set does consume resources on the server. The estimated number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.

TEMPEXPRESS="SAS-expressions"
TEAMPPRESS=file-reference

specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias TE=

TEMPNAMES=variable-name
TEMPNAMES=(variable-list)

specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

Alias TN=

TEMPTABLE

generates an in-memory temporary table from the result set. The IMSTAT procedure displays the name of the table and stores it in the &_TEMPLAST_ macro variable, provided that the statement executed successfully.

When the IMSTAT procedure exits, all temporary tables created during the IMSTAT session are removed. Temporary tables are not displayed on a TABLEINFO request, unless the temporary table is the active table for the request.

WEIGHT=variable-name

specifies the numeric weight variable to use for calculating the statistics in the table cell and the margins of the table. If no WEIGHT variable is specified, then the only aggregate method that is available to the CROSSTAB statement is N. In this case, then number of observations (frequency) is reported in each table cell.

**DECISIONTREE Statement**

The DECISIONTREE statement provides an implementation of a C4.5 decision tree method for classification. You specify a single column as the target variable when you generate the decision tree. You can also score against the generated tree.

Examples:  “Example 7: Training and Validating a Decision Tree” on page 180  “Example 8: Storing and Scoring a Decision Tree” on page 182
Syntax

**DECISIONTREE** target-variable */options*;

**DECISIONTREE Statement Options**

**ASSESS**  
specifies that predicted probabilities are added to the temporary result table for the event levels. You can use these predicted probabilities in an ASSESS statement.

**CODE */(code-generation-options)*/**  
requests that the server produce SAS scoring code based on the actions that it performed during the analysis. The server generates DATA step code. By default, the code is replayed as an ODS table by the procedure as part of the output of the statement. More frequently, you might want to write the scoring code to an external file by specifying options.

The scoring code computes the predicted value of the response variable on the data scale (the inverse link scale) and prefixes the name with "P_". For example, if the response variable is \( Y \), the generated code stores the predicted value as \( P_Y \). The name of the variable is truncated to fit within the SAS name length requirements.

**COMMENT**  
specifies to add comments to the code in addition to the header block. The header block is added by default.

**FILENAME='path'**  
specifies the name of the external file to which the scoring code is written. This suboption applies only to the scoring code itself. If you request that the server generate IMSTAT programming statements with the IMSTAT suboption, then these statements are saved as an ODS table.

**FORMATWIDTH=k**  
specifies the width to use in formatting derived numbers such as parameter estimates in the scoring code. The server applies the BEST format, and the default format for code generation is BEST20.

**IMSTAT**  
specifies to generate IMSTAT programming statements that reproduce the analysis in addition to the scoring code. For example, this option is helpful when you perform variable selection and you want to capture the modeling code that reflects only the selected variables.

**IMSTATONLY**  
specifies to generate the IMSTAT programming statements only. No scoring code is produced.

**LINESIZE=n**  
specifies the line size for the generated code.

**Alias**  
FILE=

**FMTWIDTH=**

**Range**  
4 to 32

**Alias**  
LS=

**Default**  
72
NOTRIM
requests that the comparison of the formatted values for class variables and
group-by variables is based on the full format width with padding. By default, the
leading and trailing blanks are removed from the formatted values.

REPLACE
specifies to overwrite the external file with the new contents if the file already
exists. This option has no effect unless you specify the FILENAME= option.

CFLEV=number
specifies a value between 0 and 1 that controls the aggressiveness of tree pruning
according to the C4.5 algorithm. Smaller numbers indicate to more aggressive
pruning. See the “PRUNE” on page 78 option.

DETAIL
requests detailed information about the classification results when scoring a table
against a previously calculated tree.

FORMATS=("format-specification", ...)
specifies the formats for the input variables. If you do not specify the FORMATS=
option, the default format is applied for that variable.

Example
proc imstat data=lasr1.table1;
    decisiontree x / input=(a b) formats=('8.3', '$10');
quit;

GAIN
specifies that the splitting criterion is changed to information gain. Typically, this
criterion intends to generate trees with more nodes than information gain ratio.

GREEDY
specifies how to perform splitting under specific circumstances.

Assuming that one variable has q levels, when binary splitting is performed and q is
less than 15, or option MAXBRANCH > 2 and q < 12, all possible binary splits are
enumerated and the split with the largest gain or gain ratio is chosen for the variable.

When q is less than 1024 and splitting is not just binary, local greedy searches are
applied to determine the optimum local split. Specifically, when the variable is
numeric, q levels (similar to q bins) are sorted by value.

When the variable is nominal, the q levels are ordered by random weights. The best
binary splitting is applied until the desired number of branches is reached. Only a
local optimum can be found with this technique.

For values of q ≥ 1024, the default k-means clustering algorithm is applied to
determine the splits.

IMPUTE
specifies how to treat observations with nonmissing values for the target variable
during scoring. When this option is specified, the observed values are used as the
predicted values. That is, the observed value is assumed to be known without error.
Only the observations with missing values for the target variable are then scored
against the decision tree based on their values for the input variables.

The IMPUTE option is useful if you want to replace missing values of a target
variable with classified values based on the decision tree.
INPUT=variable-name
INPUT=(variable-name1 <, variable-name2, …>)
specifies the variables to use for building the tree. You can add the target variable to
the input list if you want to assign a format to the target variable by using the
FORMATS= option. Any numeric variable that is not specified in the NOMINAL=
option is binned according to the NBINS= specification.

LEAFSIZE=m
specifies the minimal number of observations on each node. When the number of
observations on a tree node is less than m, the node is changed to a leaf during the
building of the decision tree.

Interaction  Specifying the LEAFSIZE option affects the pruning of the tree.

MAXBRANCH=n
specifies the maximum number of children (branches) to allow for each level of the
tree.

Default  2

MAXLEVEL=n
specifies the maximum number of the tree level.

Default  6

MULTVAR
specifies to allow a variable to appear multiple times when traversing the tree from
top to bottom.

NBINS=k
specifies the number of bins to use in the calculation of the decision tree. The
number of bins affects the accuracy of the tree. The accuracy increases as values of k
increase. However, computing time and memory consumption also increase as
values of k increase.

Default  2

NBINSTARGET=k
specifies the number of bins to use for a numeric target variable. The number of bins
affects the accuracy of the tree. The accuracy increases as values of k increase. However, computing time and memory consumption also increase as values of k
increase. When k is greater than zero, the numeric target variable is binned into
equally sized bins first and then the bins are used to perform the classification.

Default  0

NOMINAL=variable-name
NOMINAL=(variable-list)
specifies the numeric variables to use as nominal variables. Binning is not applied to
the specified variables. The target variable is always treated as a nominal variable
and does not need to be listed.

NOMISSOBS
specifies to ignore observations that have missing values in the analysis variables
when building a decision tree. When scoring a data set, any observations with
missing values in the analysis variables for the decision tree are ignored when this
option is specified.
When this option is not specified, the RANDOMWOODS statement builds a tree by applying the following policy for missing values:

- for an interval variable, the smallest machine value is assigned
- for a nominal variable, missing values are represented by a separate level

**NOPREPARSE**

prevents the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects problems with the code, the error messages might not be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

When you specify this option in the PROC IMSTAT statement, the option applies to all statements that can generate code. You can also exclude specific statements from pre-parsing by using the NOPREPARSE option in statements that allow temporary columns or the SCORE statement.

**Alias**  

NOPREP

**Nopruneobs**

specifies not to prune any observations when building a decision tree.

**NOSCORE**

suppresses the generation of the scoring temporary table when the TEMPTABLE option is specified. In this case, the server generates only one temporary table and the table contains the decision tree.

**PRUNE**

requests to prune the tree according to the C4.5 algorithm. Pruning can increase the error of misclassification. You can control the aggressiveness of pruning with the CFLEV= option. Smaller values for the CFLEV= option result in more aggressive pruning.

**PRUNEGROW**

specifies to enable C4.5 pruning when building a classification decision tree. The tree could have large a misclassification rate but the building process is performed quickly.

**REG**

specifies to build a regression tree. Minimal cost-complexity pruning is applied to prune the tree.

**SAVE=table-name**

saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

**SCORERADATA=table-name**

specifies the in-memory table that contains the scoring data. The table must exist in-memory on the server. The DECISIONTREE statement in the IMSTAT procedure does not transfer a local data set to the server.
SETSIZE requests that the server estimate the size of the result set. The procedure does not create a result table if the SETSIZE option is specified. Instead, the procedure reports the number of rows that are returned by the request and the expected memory consumption for the result set (in KB). If you specify the SETSIZE option, the SAS log includes the number of observations and the estimated result set size. See the following log sample:

```
NOTE: The LASR Analytic Server action request for the STATEMENT statement would return 17 rows and approximately 3.641 kBytes of data.
```

The typical use of the SETSIZE option is to get an estimate of the size of the result set in situations where you are unsure whether the SAS session can handle a large result set. Be aware that in order to determine the size of the result set, the server has to perform the work as if you were receiving the actual result set. Requesting the estimated size of the result set does consume resources on the server. The estimated number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.

STAT specifies to generate two additional tables that contain statistical information about the variables that are used in the decision tree. One table contains the variable importance information, which is determined by the total Gini reduction. The second table contains the variable splitting information for each node in the decision tree.

```
TEMPEXPRESS="SAS-expressions"
```

`TEMPEXPRESS=file-reference` specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

```
Alias   TE=
```

```
TEMPPNAMES=variable-name
```

TEMPPNAMES=(variable-list) specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

```
Alias   TN=
```

TEMPTABLE specifies to store the results in a temporary table. The type of information that is stored depends on whether you are building a decision tree or scoring a table with a decision tree.

When you are building a decision tree, the generated decision tree is stored in the server and the input table is automatically scored using this tree. The scoring details are saved in a temporary table. The _TEMPTREE_ macro variable stores the name of the temporary table for the tree. The _TEMPSCORE_ macro variable stores the name of the temporary table that has the scoring results of traversing the decision tree. You can suppress the generation of the scoring temporary table (_TEMPSCORE_) during the tree building phase by specifying the NOSCORE option.

When you are scoring a table using a decision tree, the TEMPTABLE option requests to store the scoring details in a temporary table in the server. The IMSTAT
procedure displays the name of the table and stores it in the _TEMPSMCRE_ macro variable. Be aware that the DETAIL option can generate a very large amount of scoring results when the in-memory table that is specified in the SCOREDATA= option is large. Observations from the scored data set can be transferred to the temporary table using the VARS= option.

**TIMEOUT=** specifies the maximum number of seconds that the server should run the statement. If the time-out is reached, the server terminates the request and generates an error and error message. By default, there is no time-out.<

**TREEDATA=libref.member-name**
**TREETAB=saved-table**
**TREELASRTAB=table-name**
specifies the saved table that contains the generated tree. In order to score a (validation) data set against the generated tree, you need the validation data and a representation of the tree. Specify these options as follows:

- The TREEDATA= option is used to specify the name of a SAS data set that stores the generated tree. The data set is local to the SAS client.
- The TREETAB= option is used to specify a table on the SAS client that stores the generated tree.
- The TREELASRTAB= option is used to specify a valid decision tree that is stored in an in-memory table.

The data set with the observations to score is specified in the SCOREDATA= option

Alias **SCORETAB=**

**VARS=variable-name**
**VARS=(variable-name1 <, variable-name2, ...>)**
specifies the variables to transfer from the input table to the temporary table in the server that contains the results of scoring a decision tree. This option has no effect unless you specify the TEMPTABLE option and you score a decision tree.

### DISTINCT Statement

The DISTINCT statement calculates the count of unique raw values of variables. You can specify the variables to calculate in the variable list. If no list is specified, the count of unique raw values is calculated for all variables.

**Syntax**

```
DISTINCT <variable-list> </options>
```

**DISTINCT Statement Options**

**FORMATS=("format-specification",...)**
specifies the formats for the GROUPBY= variables. If you do not specify the FORMAT= option, or if you do not specify the GROUPBY= option, the default format is applied for that variable.

Enclose each format specification in quotation marks and separate each format specification with a comma.
Example: proc imstat data=lasr1.table1;
    DISTINCT x / groupby=(a b) formats=("8.3", "$10");
quit;

GROUPBY=(variable-list)
specifies a list of variable names, or a single variable name, to use as GROUPBY
variables in the order of the grouping hierarchy. If you do not specify any
GROUPBY variable names, then the calculation is performed across the entire table
—possibly subject to a WHERE clause.

GROUPBYLIMIT=n
specifies the maximum number of levels in a GROUPBY set. When the software
determines that there are at least n levels in the GROUPBY set, it abandons the
action, returns a message, and does not produce a result set. You can specify the
GROUPBYLIMIT= option if you want to avoid creating excessively large result sets
in GROUPBY operations.

GROUPFILTER=(filter-options)
specifies a section of the group-by hierarchy to be included in the computation. With
this option, you can request that the server performs the analysis for only a subset of
all possible groupings. The subset is determined by applying the group filter to a
temporary table that you generate with the GROUPBY statement.

You can specify the following suboptions in the GROUPFILTER option:

DESCENDING
specifies the top or the bottom section of the groupings to be collected. If the
DESCENDING option is specified, the top LIMIT=n (where n > 0) groupings
are collected. Otherwise, the bottom LIMIT=n groupings are collected.

   Alias  DESC

LIMIT=n
specifies the maximum number of distinct groupings to be collected, where
integer n >= 0. If n is zero, then all distinct groupings (up to 2^31-1) that satisfy
the boundary constraints, such as LOWERSCORE=, are collected.

CAUTION
High Cardinality Data Sets Setting n to zero with high-cardinality
data sets can significantly delay the response of the server.

SCOREGT=f
specifies the exclusive lower bound for the numeric scores of the distinct
groupings to collect.

   Alias  SGT=

SCORELT=f
specifies the exclusive upper bound for the numeric scores of the distinct
groupings to collect.

   Alias  SLT=

VALUEGT=("format-name1" <, "format-name2" …>)
specifies the exclusive lower bound of the group-by variable’s formatted values
for the distinct groupings to collect.

   Alias  VGT=
VALUETL=("format-name1","format-name2" …)
   specifies the exclusive upper bound of the group-by variable’s formatted values
   for the distinct groupings to collect.

Alias VLT=

TABLE=table-with-groupby-results
   specifies the in-memory table from which to load the group-by hierarchy. If the
   TABLE= option is not specified, then all other GROUPFILTER= options are
   ignored.

The following program request all the groupings of State, City, and then
Trade_In_Model in the Cars_Program_All table. The groupings are ordered by
the maximum value of New_Vehicle_Msrp for each grouping:

proc imstat;
   table example.CARS_PROGRAM_ALL;
   groupby state city trade_in_model / TEMPTABLE
      WEIGHT=new_vehicle_MSRP
      AGG   =(MAX)
      ORDER =WEIGHT;
run;

The TEMPTABLE option in the GROUPBY statement directs the server to save
all the groupings in a temporary in-memory table. The following DISTINCT
statement requests the count of the distinct unformatted values of Sales_Type for
each of the selected groupings of State, City, and Trade_In_Model.

   table example.CARS_PROGRAM_ALL;
   distinct sales_type / GROUPFILTER=(
      table  =mylasr.&_TEMPLAST_
      scoregt=40000
      valuelt=("FL","Ft Myers","")
      limit  =20
      descending);
run;

This example only considers groupings that have maximum values of the
New_Vehicle_Msrp above 40,000 and with formatted values that are less than
State="FL" and City="Ft Myers." The empty quotation marks result in no
restriction on Trade_In_Model values. These groupings are ordered according to
the maximum values of New_Vehicle_Msrp. Because of the DESCENDING
option, this example collects the 20 top groupings within the specified group-by
range for the DISTINCT analysis.

Interaction
   If you specify the GROUPFILTER= option, then the GROUPBY= and
   FORMATS= options have no effect.

MAXNVALS=n
   specifies the maximum size that trees are allowed to consume during the calculation
   of distinct counts. If you execute a DISTINCT statement with a GROUPBY= or
   PARTITION= option, then the MAXNVALS limit applies within the groups or
   partitions.

   Default 6

NOMISSING
   specifies that you do not want to include missing values in the determination of the
   distinct count.
NOPREPARSE
prevents the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects problems with the code, the error messages might not be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

When you specify this option in the PROC IMSTAT statement, the option applies to all statements that can generate code. You can also exclude specific statements from pre-parsing by using the NOPREPARSE option in statements that allow temporary columns or the SCORE statement.

Alias NOREP

NOTEMPPART
specifies that the temporary table generated by the TEMPTABLE option is not partitioned by the GROUPBY= variables. When you request a temporary table with the DISTINCT statement, by default, the server partitions the table and the size of a partition is equal to the number of analysis variables in the variable-list of the DISTINCT statement. When the number of groups is large, this can result in many small partitions, and requires extra memory resources to store the partition information for the temporary table. By specifying this option, the temporary table is organized similarly to the default table, but is not a partitioned table.

Alias NOTP

ORDERBY=(variable-list)
specifies one or more variables by which to order the result set. The variables specified in variable-list are either one or more of the GROUPBY= variables or one or more of the analysis variables. If you specify an incorrect variable, the server returns an error and no result set. Separate multiple variables with a space.

When there are ties given the ordering of the ORDERBY= variable values, the server sorts the tied items by the GROUPBY= or PARTITION= variable values (unless neither the GROUPBY or the PARTITION option are specified). If the ORDERBY= option is not specified, the result set is ordered by the formatted values of GROUPBY= variables.

The following DISTINCT statement requests the number of the distinct raw values of Invoice, grouped by Type in the Cars table:

data example.cars; set sashelp.cars; run;

proc imstat data=example.CARS;
  distinct Invoice / GROUPBY=Type ORDERBY=Invoice MAXNVALS=32;
run;
The result set is ordered by Invoice values. The statement produces the following output:

<table>
<thead>
<tr>
<th>Type</th>
<th>Column</th>
<th>Number of Distinct Values</th>
<th>Number of Missing Values</th>
<th>Truncated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid</td>
<td>Invoice</td>
<td>3</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>Truck</td>
<td>Invoice</td>
<td>24</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>Wagon</td>
<td>Invoice</td>
<td>30</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>SUV</td>
<td>Invoice</td>
<td>32</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td>Sedan</td>
<td>Invoice</td>
<td>32</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td>Sports</td>
<td>Invoice</td>
<td>32</td>
<td>0</td>
<td>Yes</td>
</tr>
</tbody>
</table>

There are three items tied with the same distinct Invoice raw value, 32. These items are then ordered by the formatted values of Type.

**ORDERBYDESC**

specifies the sort order for the result set. The default is ascending order. Specifying this option arranges the results in descending order. This option has no effect unless you specify the ORDERBY= option.

**PARTITION <=partition-key>**

When you specify this option and the table is partitioned, the results are calculated separately for each value of the partition key. In other words, the partition variables function as automatic GROUPBY variables. This mode of executing calculations by partition is more efficient than using the GROUPBY= option. With a partitioned table, the server takes advantage of knowing that observations for a partition cannot be located on more than one worker node.

If you do not specify a partition-key, the analysis is performed for all partitions. If you do specify a partition-key, the analysis is carried out for the specified key value only. You can use the PARTITIONINFO statement to retrieve the valid partition key values for a table.

You can specify a partition-key in two ways. You can supply a single quoted string that is passed to the server, or you can specify the elements of a composite key separated by commas. For example, if you partition a table by variables GENDER and AGE, with formats $1 and BEST12, respectively, then the composite partition key has a length of 13. You can specify the partition for the 11-year-old females as follows:

```plaintext
statement / partition="F 11"; /* passed directly to the server */
statement / partition="F","11"; /* composed by the procedure */
```

If you choose the second format, the procedure composes a key based on formatting information from the server.

**RESULTLIMIT=k**

specifies that the number of items that are returned to the client is limited to \( k \) times the number of analysis variables if you also specify the GROUPBY= or ORDERBY= option.
The following DISTINCT statement requests the numbers of the distinct raw values of Invoice grouped by Type in the Cars table:

The following DISTINCT statement requests the numbers of the distinct raw values of Invoice and Cylinders grouped by Type, and limits the results to \((2 \text{ variables} \times 4) = 8\) rows.

data example.cars; set sashelp.cars; run;

proc imstat data=example.CARS;
  distinct Invoice Cylinder / resultlimit=4;
run;

The statement produces the following output:

<table>
<thead>
<tr>
<th>Type</th>
<th>Column</th>
<th>Number of Distinct Values</th>
<th>Number of Missing Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid</td>
<td>Invoice</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Hybrid</td>
<td>Cylinders</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>SUV</td>
<td>Invoice</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>SUV</td>
<td>Cylinders</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Sedan</td>
<td>Invoice</td>
<td>260</td>
<td>0</td>
</tr>
<tr>
<td>Sedan</td>
<td>Cylinders</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Sports</td>
<td>Invoice</td>
<td>49</td>
<td>0</td>
</tr>
<tr>
<td>Sports</td>
<td>Cylinders</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

Four groups of two rows each are displayed. Without the RESULTLIMIT= option, six groups of two rows are displayed.

**SAVE=table-name**

saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for `table-name` must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

**SETSIZE**

requests that the server estimate the size of the result set. The procedure does not create a result table if the SETSIZE option is specified. Instead, the procedure reports the number of rows that are returned by the request and the expected memory consumption for the result set (in KB). If you specify the SETSIZE option, the SAS log includes the number of observations and the estimated result set size. See the following log sample:

NOTE: The LASR Analytic Server action request for the STATEMENT statement would return 17 rows and approximately 3.641 kBytes of data.

The typical use of the SETSIZE option is to get an estimate of the size of the result set in situations where you are unsure whether the SAS session can handle a large result set. Be aware that in order to determine the size of the result set, the server has to perform the work as if you were receiving the actual result set. Requesting the
estimated size of the result set does consume resources on the server. The estimated number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.

SORTAGG=aggregation-method
specifies the aggregator for which the ordering of the result set is based, if the ORDERBY= option is specified.

The available aggregation methods are as follows:

N number of observations
NMISS number of missing observations

Interaction You must specify the ORDERBY= option to use this option.

TEMPEXPRESS="SAS-expressions"
TEMPNAMES=variable-name

specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias TE=

TEMPNAMES=(variable-list)
specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

Alias TN=

TEMPTABLE

generates an in-memory temporary table from the result set. The IMSTAT procedure displays the name of the table and stores it in the &_TEMPLAST_ macro variable, provided that the statement executed successfully.

When the IMSTAT procedure exits, all temporary tables created during the IMSTAT session are removed. Temporary tables are not displayed on a TABLEINFO request, unless the temporary table is the active table for the request.

Interaction The TEMPTABLE option requires a group-by analysis or a partitioned analysis.

VARFORMATS=("format-specification",...)
specifies the formats for the analysis variables. If you do not specify this option, the distinct count is based on the number of distinct unformatted values of a variable.

Note that the FORMATS= option controls the formatting of the GROUPBY= variables and the VARFORMATS= option controls the formatting of the analysis variables. It is possible to specify a different format for a variable if it appears as a GROUPBY variable and as an analysis variable.

You can specify a combination of formatted and unformatted value counts by submitting an empty string as the format for variables that you do not wish to format. For example, in the following code the distinct count of variable Invoice is based on the formatted values according to the user-defined format PRICE20. The distinct count of variable Msrp is based on its unformatted values.
Example

```plaintext
proc imstat data=example.cars;
    distinct msrp invoice / varformats=(' ', *PRICE20');
run;
```

**FORECAST Statement**

The FORECAST statement computes predicted values, measures of precision, and confidence limits for observed and future (forecast) values of a time series.

**Syntax**

```plaintext
FORECAST timestamp-variable / options;
FORECAST DATA=libref.member-name timestamp-variable / options;
```

**Required Arguments**

- `timestamp-variable` specifies the name of the SAS datetime variable to use.
- `libref.member-name` specifies the libref and table name of a SAS data set when you specify the DATA= option. The data set must contain the timestamp variable and one or more of the analysis variables. The procedure then sends these values to the server to request the forecast calculation.

**FORECAST Statement Options**

- `AGGREGATE=(list-of-aggregators)` specifies the aggregators on which the ordering of the result set is based. The following aggregators are valid:

  - CSS  corrected sum of squares
  - CV   coefficient of variation
  - MAX  maximum value
  - MEAN arithmetic mean
  - MIN  minimum value
  - N    number of observations
  - PROBT p-value for the t-statistic
  - STD  standard deviation
  - STDERR standard error
  - SUM  sum of the nonmissing values
  - TSTAT t-statistic for the null hypothesis that the mean equals zero
  - USS  uncorrected sum of squares
  - VAR  sample variance

```
Each analysis variable can be associated with a different aggregator. For example, the following statement forecasts the sum of expenses and the mean of revenue:

```plaintext
forecast ts / vars = (expenses revenue)
aggregate = (sum mean);
```

The default aggregation is the mean of the analysis variables within unique values of the timestamp variable.

**Interaction**
This option has no effect if you specify a data set with the DATA= option.

**FORMATS=("format-specification")**
specifies the format for the time stamp variable. If you do not specify the FORMATS= option, the default format is applied for the time stamp variable.

**Interaction**
This option has no effect if you specify a data set with the DATA= option.

**FRAME=LEAD | HORIZON**
**FRAME=TAIL | HISTORY**
**FRAME=BOTH**
specifies how to compose the main result table. The default is FRAME=BOTH and the result set contains the observed series (the history) as well as the forecast (the horizon). If you specify FRAME=LEAD (or FRAME=HORIZON), then only the future values are returned. You can control the length of the horizon with the LEAD= option.

If you specify FRAME=TAIL (or FRAME=HISTORY), then only the results for the historic values are returned. The returned values are the aggregated values, their predicted values, residuals, prediction standard errors, and confidence limits. You can control the number of the historical records with the TAIL= option.

**Alias**
**WINDOW=**

**Default**
**BOTH**

**HOST="host-name"**
specifies the machine to which you want to connect to produce the forecast when you specify the DATA= option in the FORECAST statement. If you do not specify the host information, it is determined from the active table.

**PORT=number**
specifies to use the server that is listening on that port to produce the forecast when you specify the DATA= option in the FORECAST statement. You can use this option with the HOST= option to use a specific server. If you do not specify a PORT= value, the behavior of the FORECAST statement depends on whether a table is active. If there is no active table, then the IMSTAT procedure tries to connect to the server using the LASRPORT macro variable. If a table is active, then a connection is made to the server that has the active table.

**INDEP=variable-name**
**INDEP=(variable-list)**
specifies the independent variables used in automatic modeling. When you specify one or more independent variables, the server performs model selection and determines the best-fitting time series model and the important independent variables. If any variables are selected, a table is generated to show the actual and predicted values for each variable. Specify the INFO option to view the Forecast Information table that displays the selected time series model.
INFO
specifies to display a forecast information table for each analysis variable. Each table provides informational details about the forecast. For example, you can learn from this table what time units were applied and which method was used to compute the forecast.

LEAD=n
specifies the forecast horizon (in number of time intervals).

Default 12

Interaction This option has no effect if you specify a data set with the DATA= option.

NOPREPARSE
specifies to prevent the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects problems with the code, the error messages might not be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

STAMPLIMIT=m
specifies a hard limit for the number of time stamps. If that number reaches \( m \), then execution stops and the server generates an error message. This option is useful to protect against the generation of very large result sets. You can also limit the number of time stamps used in the forecast with the TAIL= option. Using the TAIL= option also reduces the size of the result set.

SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for \( table-name \) must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

TAIL=k
specifies the number of most recent time intervals on which to base the estimation of the predicted and forecasted values. The TAIL= option enables you to restrict the length of the series that is used in the forecast.

For example, if the aggregation results in 500 unique values of the time stamp, then specifying TAIL=30 uses only the thirty most recent values in the estimation procedure. If you do not specify the TAIL= option, then all the aggregated time stamps are used in the estimation procedure. This option can also limit the size of the result set since at most \( k \) observations are used in the computation of the forecast.
Interaction

This option has no effect if you specify a data set with the DATA= option.

**TEMPRESS**="SAS-expressions"
**TEMPRESS**=file-reference

specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

**Alias**

**TE**=

**TEMPNAMES**=variable-name
**TEMPNAMES**=(variable-list)

specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the **TEMPRESS**= option.

**Alias**

**TN**=

**VARS**=variable-name
**VARS**=(variable-list)

specifies one or more numeric analysis variables to forecast. If you do not specify the **VARS**= option, a forecast is produced for all numeric variables in the active table. If you specify a data set with the DATA= option, you must specify the analysis variables in the **VARS**= option. If you do not, the server generates an error.

**Alias**

**DEPVARS**=

**Details**

There are two ways to use the FORECAST statement. You can use the active table or you can specify a data set with the DATA= option. The following paragraphs provide more information about these choices.

When you use the active table, the server forms a time series by aggregating the values of the analysis variables according to the unique (formatted) values of a numeric time stamp variable. The time stamp variable must be a SAS datetime type. The aggregate series (one for each analysis variable) are then used to compute predicted values of the series. The predicted values can cover the observed time interval or can apply to future observations. Measures of precision (standard errors of prediction and confidence limits) are also available. You can produce forecasts for multiple variables and you can vary the method for aggregating values on a variable-by-variable basis.

Alternatively, you can specify a SAS data set with the DATA= option. The data set must have a time stamp variable and one or more of the analysis variables. In this case, the data are sent to the server for the forecast calculation. In this case, there is no aggregation because the values read from the data set are assumed to constitute the series of interest. You can produce forecasts for multiple variables when you use the DATA= option, but you cannot specify the aggregation technique for the variables or specify the **TAIL**= and **HEAD**= options in the FORECAST statement.

**FREQUENCY Statement**

The FREQUENCY statement is used to calculate a frequency distribution for one or more variables.
Syntax

FREQUENCY variable-list < options >;

Required Argument

variable-list

specifies the numeric and character variables to use for calculating the frequency distribution. The distribution is calculated for the unique formatted values of the variables.

FREQUENCY Statement Options

DESCENDING

specifies that the levels of the GROUPBY variables are to be arranged in descending order.

Alias DESC

FORMATS="format-specification", ..."

specifies the format to apply to each variable. Specify the list as a comma-separated list and enclose each format specification in quotation marks. If you do not specify a format, then the default format for the variable is used.

MERGEBINS=b

specifies the number of bins to create when a numeric GROUPBY variable exceeds the MERGELIMIT=n specification. If you specify a MERGELIMIT, but do not specify a value for the MERGEBINS= option, the server automatically calculates the number of bins.

MERGELIMIT=n

specifies that when the number of unique values in a numeric GROUPBY variable exceeds n, the variable is automatically binned and the GROUPBY structure is determined based on the binned values of the variable, rather than the unique formatted values.

For example, if you specify MERGELIMIT=500, any numeric GROUPBY variable with more than 500 unique formatted values is binned. Instead of returning results for more than 500 groups, the results are returned for the bins. You can specify the number of bins with the MERGEBINS= option.

NOEMPTY

specifies that empty cells are not returned to the SAS session (only full cells are returned). When this option is specified, the server eliminates all levels with zero frequency from the result set.

Alias FULLCELL

NOMISS

specifies that missing values are excluded in the calculation of formatted values. By default, levels with missing values are included.

Alias NOMISSING

RAWORDER

specifies that the ordering of the GROUP BY value is based on the raw values of the variables instead of the formatted values.
SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

SETSIZE
requests that the server estimate the size of the result set. The procedure does not create a result table if the SETSIZE option is specified. Instead, the procedure reports the number of rows that are returned by the request and the expected memory consumption for the result set (in KB). If you specify the SETSIZE option, the SAS log includes the number of observations and the estimated result set size. See the following log sample:

NOTE: The LASR Analytic Server action request for the STATEMENT statement would return 17 rows and approximately 3.641 kBytes of data.

The typical use of the SETSIZE option is to get an estimate of the size of the result set in situations where you are unsure whether the SAS session can handle a large result set. Be aware that in order to determine the size of the result set, the server has to perform the work as if you were receiving the actual result set. Requesting the estimated size of the result set does consume resources on the server. The estimated number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.

TEMPEXPRESS="SAS-expressions"

TEMPEXPRESS=file-reference
specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias TE=

TEMPNAMES=variable-name
TEMPNAMES=(variable-list)
specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

Alias TN=

GENMODEL Statement
The GENMODEL statement is used to fit statistical models from the class of generalized linear models and some related models.

Syntax

GENMODEL dependent-variable <(class-variables)> = model-effects </ options>;
**Required Arguments**

`dependent-variable`

specifies the variable to model. This variable is also referred to as the response variable.

`model-effects`

specifies a list of variables to use for modeling the dependent variable.

**Optional Argument**

`class-variables`

specifies a list of variables to use as classification variables. The variables in this list take the place of the CLASS statement in traditional SAS procedures.

**GENMODEL Statement Options**

**ALLIDVARS**

requests that all variables in the input table are treated as ID variables when a scoring table is produced. In other words, if this option is specified, all variables from the input table, including computed columns, are transferred to the scoring table.

**ALPHA=number**

specifies a number between 0 and 1 from which to determine the confidence level for approximate confidence intervals of the parameter estimates. The default is \( \alpha = 0.05 \), which leads to \( 100 \times (1 - \alpha)\% = 95\% \) confidence limits for the parameter estimates.

Default 0.05

**CI**

specifies to add confidence intervals to the table of parameter estimates. The confidence level is \( 100\% \times (1 - \alpha) \) where \( \alpha \) is determined by the `ALPHA=` option. The default value is \( \alpha = 0.05 \). This value is equivalent to a 95% confidence limit.

Default 0.05

**CLASSFORMATS=("format-name1", "format-name2" ...)>**

specifies the formats for the classification variables in the model. If you do not specify the `CLASSFORMATS=` option, the default format is applied for the classification variable. That default format was determined when the table was originally loaded into the server. In the following example, the `CLASSFORMAT=` values apply to variables `x1` and `x2`.

Alias `CLASSFMT=`

Example `genmodel y (x1 x2) = x3-x7 / classformats="YN.", "F8.");`

**CODE <(code-generation-options)>**

requests that the server produce SAS scoring code based on the actions that it performed during the analysis. The server generates DATA step code. By default, the code is replayed as an ODS table by the procedure as part of the output of the statement. More frequently, you might want to write the scoring code to an external file by specifying options.

The scoring code computes the predicted value of the response variable on the data scale (the inverse link scale) and prefixes the name with "P_.". For example, if the response variable is `Y`, the generated code stores the predicted value as `P_Y`. The name of the variable is truncated to fit within the SAS name length requirements.
COMMENT
  specifies to add comments to the code in addition to the header block. The header block is added by default.

FILENAME='path'
specifies the name of the external file to which the scoring code is written. This suboption applies only to the scoring code itself. If you request that the server generate IMSTAT programming statements with the IMSTAT suboption, then these statements are saved as an ODS table.

  Alias  FILE=

FORMATWIDTH=k
  specifies the width to use in formatting derived numbers such as parameter estimates in the scoring code. The server applies the BEST format, and the default format for code generation is BEST20.

  Alias  FMTWIDTH=
  Range  4 to 32

IMSTAT
  specifies to generate IMSTAT programming statements that reproduce the analysis in addition to the scoring code. For example, this option is helpful when you perform variable selection and you want to capture the modeling code that reflects only the selected variables.

IMSTATONLY
  specifies to generate the IMSTAT programming statements only. No scoring code is produced.

LINESIZE=n
  specifies the line size for the generated code.

  Alias  LS=
  Default  72
  Range  64 to 256

NOTRIM
  requests that the comparison of the formatted values for class variables and group-by variables is based on the full format width with padding. By default, the leading and trailing blanks are removed from the formatted values.

REPLACE
  specifies to overwrite the external file with the new contents if the file already exists. This option has no effect unless you specify the FILENAME= option.

DIST=distribution
  specifies the distribution of the response variable. See the following list for the available values:
  •  BETA
  •  EXPONENTIAL | EXPO
  •  GAMMA | GAM
  •  GENPOISSON
  •  GEOMETRIC | GEOM
• INVGAUSS | IGAUSSIAN | IG
• NEGBINOMIAL | NEGBIN | NB
• NORMAL | GAUSSIAN | GAUSS
• POISSON | PO
• T | STUDENT
• WEIBULL

**EXCLUDE=(list-of-ODS-tables)**
specifies the result tables that you want to exclude from being generated on the server and from being sent to the client. The GENMODEL statement can generate the following tables:

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Table Alias</th>
<th>Description</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ModelInfo</td>
<td></td>
<td>Information about the model—constant across groups or partitions.</td>
<td>This table is shown by default.</td>
</tr>
<tr>
<td>ClassLevels</td>
<td>Class</td>
<td>Information about the classification variables, such as the number of levels and their values.</td>
<td>This table is shown when classification variables are present in the model.</td>
</tr>
<tr>
<td>ConvStatus</td>
<td>Convergence</td>
<td>Convergence status of optimization</td>
<td>This table is shown by default.</td>
</tr>
<tr>
<td>Dimensions</td>
<td>Dim</td>
<td>Model dimensions</td>
<td>This table is shown by default.</td>
</tr>
<tr>
<td>FitStatistics</td>
<td>Fit</td>
<td>Fit statistics customary for generalized linear models</td>
<td>This table is shown when it is requested with the SELECT= option.</td>
</tr>
<tr>
<td>IterHistory</td>
<td>IterHist</td>
<td>Iteration history</td>
<td>This table is shown when the ITDETAILS option is used or when the table is requested with the SELECT= option.</td>
</tr>
<tr>
<td>ParmEstimates</td>
<td>ParameterEstimates Pest</td>
<td>The solutions for the linear model coefficients</td>
<td>This table is shown by default.</td>
</tr>
<tr>
<td>Table Name</td>
<td>Table Alias</td>
<td>Description</td>
<td>Condition</td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Tests3</td>
<td></td>
<td>Type III tests of model effects</td>
<td>This table is shown when the distribution is in the exponential family, the effects contain classification variables, and the NOSTDERR option is not specified.</td>
</tr>
</tbody>
</table>

Whether a table is shown by default or not, you can request any table with the SELECT= option in the GENMODEL statement. The Condition column in the table identifies when a table is produced by default.

**FORMATS=("format-specification"<,...>)**

specifies the formats for the GROUPBY variables. If you do not specify the FORMATS= option, or if you omit the entry for a GROUPBY variable, the default format is applied for that variable.

Enclose each format specification in quotation marks and separate each format specification with a comma.

**Example**

```plaintext
proc imstat data=lasr1.table1;
    statement / groupby=(a b) formats=("8.3", "$10");
quit;
```

**FCONV=r**

specifies a relative function convergence criterion. For all techniques except NMSIMP, termination requires a small relative change of the function value in successive iterations. Suppose that Ψ is the p x 1 vector of parameter estimates in the optimization and the objective function at the kth iteration is denoted, f(Ψ). Then, the FCONV criterion is met if

\[
\frac{f(Ψ(k)) - f(Ψ(k-1))}{f(Ψ(k-1))} \leq r
\]

Default \( r=10^{-\text{FDIGITS}} \) where FDIGITS is -log\(_{10}(e)\) and e is the machine precision.

**FREQ=variable-name**

specifies the numeric variable that provides frequencies for the analysis. For example, if the FREQ= variable has the value 5, then it implies that the record represents five such observations with identical values for the modeling variables. If you specify a FREQ= variable, then only the observations with a value that is not missing and greater than zero for the variable are used in the analysis.

**GCONV=r**

specifies a relative gradient convergence criterion. For all optimization techniques except CONGRA and NMSIMP, termination requires that the normalized predicted function reduction is small. The default value is \( r=1e-8 \). Suppose that Ψ is the p x 1 vector of parameter estimates in the optimization with ith element Ψ\(_i\). The objective function, its p x 1 gradient vector, and its p x p Hessian matrix are denoted, f(Ψ), g(Ψ), and H(Ψ), respectively. Then, if superscripts denote the iteration count, the normalized predicted function reduction at iteration \( k \) is
The GCONV convergence criterion is assumed to be met if that value is less than or equal to \( r \).

Note that it is possible that the relative gradient reduction is small, even if one or more gradients is still substantial in absolute value. If this situation occurs, you can disable the GCONV criterion by setting \( r=0 \). If the optimization would have stopped early due to meeting the GCONV criterion, the iterative process usually takes one more step until the gradients are small in absolute value.

**GROUPBY=(variable-list)**
specifies a list of variable names, or a single variable name, to use as GROUPBY variables in the order of the grouping hierarchy. If you do not specify any GROUPBY variable names, then the calculation is performed across the entire table — possibly subject to a WHERE clause.

**GROUPFILTER=(filter-options)**
specifies a section of the group-by hierarchy to be included in the computation. With this option, you can request that the server performs the analysis for only a subset of all possible groupings. The subset is determined by applying the group filter to a temporary table that you generate with the GROUPBY statement.

You can specify the following suboptions in the GROUPFILTER option:

**DESCENDING**
specifies the top or the bottom section of the groupings to be collected. If the DESCENDING option is specified, the top \( LIMIT=\) \( n \) (where \( n > 0 \)) groupings are collected. Otherwise, the bottom \( \) \( LIMIT=\) \( n \) groupings are collected.

Alias **DESC**

**LIMIT=\( n \)**
specifies the maximum number of distinct groupings to be collected, where integer \( n \geq 0 \). If \( n \) is zero, then all distinct groupings (up to \( 2^{31}-1 \)) that satisfy the boundary constraints, such as LOWERSCORE\( =f \), are collected.

**CAUTION** **High Cardinality Data Sets** Setting \( n \) to zero with high-cardinality data sets can significantly delay the response of the server.

**SCOREGT=\( f \)**
specifies the exclusive lower bound for the numeric scores of the distinct groupings to collect.

Alias **SGT=**

**SCORELT=\( f \)**
specifies the exclusive upper bound for the numeric scores of the distinct groupings to collect.

Alias **SLT=**

**VALUEGT=("format-name1" <, "format-name2" ...)**
specifies the exclusive lower bound of the group-by variable’s formatted values for the distinct groupings to collect.

Alias **VGT=**
VALUELT="("format-name1" <,"format-name2" ...")"
specifies the exclusive upper bound of the group-by variable’s formatted values
for the distinct groupings to collect.

Alias   VLT=

TABLE=table-with-groupby-results
specifies the in-memory table from which to load the group-by hierarchy. If the
TABLE= option is not specified, then all other GROUPFILTER= options are
ignored.

The following program request all the groupings of State, City, and then
Trade_In_Model in the Cars_Program_All table. The groupings are ordered by
the maximum value of New_Vehicle_Msrp for each grouping:

```
proc imstat;
  table example.CARS_PROGRAM_ALL;
  groupby state city trade_in_model / TEMPTABLE
    WEIGHT=new_vehicle_MSRP
    AGG=(MAX)
    ORDER=WEIGHT;
run;
```

The TEMPTABLE option in the GROUPBY statement directs the server to save
all the groupings in a temporary in-memory table. The following DISTINCT
statement requests the count of the distinct unformatted values of Sales_Type for
each of the selected groupings of State, City, and Trade_In_Model.

```
table example.CARS_PROGRAM_ALL;
distinct sales_type / GROUPFILTER={
  table =mylasr.&_TEMPLAST_
  scoregt=40000
  valueLT="FL","Ft Myers","
  limit =20
  descending};
run;
```

This example only considers groupings that have maximum values of the
New_Vehicle_Msrp above 40,000 and with formatted values that are less than
State="FL" and City="Ft Myers." The empty quotation marks result in no
restriction on Trade_In_Model values. These groupings are ordered according to
the maximum values of New_Vehicle_Msrp. Because of the DESCENDING
option, this example collects the 20 top groupings within the specified group-by
range for the DISTINCT analysis.

Interaction If you specify the GROUPFILTER= option, then the GROUPBY= and
FORMATS= options have no effect.

IDVARS=(variable-list)
IDVARS=variable-name
specifies the variables from the active table to transfer to the temporary table that is
created by scoring the input table. This option has no effect unless the SCORE
option is also specified. (See the SCORE option for details about which variables are
added to the temporary table by default.) The IDVARS= option should be used to
transfer additional columns from the input table to the scoring table.

Alias   ID=
Tip Instead of this option, you can specify the ALLIDVARS option to transfer all variables from the input table to the scoring table.

ITDETAILS
requests to add details about the iterative model fitting process (an iteration history) to the ODS output tables.

Alias ITDETAIL

KEYORDER
requests that the results for a partitioned analysis are displayed in the order of the partition keys. If this option is not specified, then results are displayed by using the partitions on the first worker node followed by the partitions on the second node, and so on. Without this option, the results are likely to have random ordering of the partitions. The KEYORDER option makes result collection less efficient but produces a natural, predictable order.

LINK=function
specifies the link function to use for the model fitting process. If you do not specify a link function, the server selects the most appropriate function for the distribution of the data. See the following list for the available functions:

- IDENTITY
- LOGIT
- PROBIT
- LOG
- LOGLOG
- CLOGLOG
- RECIP
- POWMINUS2
- POWER(v) | POW(v) | POM(v)

MAXFUNC=n
specifies the maximum number n of function calls in the iterative model fitting process. The default value depends on the optimization technique as follows:

<table>
<thead>
<tr>
<th>Optimization Technique</th>
<th>Default Number of Function Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUREG, NRRIDG, and NEWRAP</td>
<td>125</td>
</tr>
<tr>
<td>QUANEW and DBLDOG</td>
<td>500</td>
</tr>
<tr>
<td>CONGRA</td>
<td>1000</td>
</tr>
<tr>
<td>NMSIMP</td>
<td>3000</td>
</tr>
</tbody>
</table>

Alias MAXFU=

MAXITER=i
specifies the maximum number of iterations in the iterative model fitting process. The default value depends on the optimization technique as follows:
### Optimization Technique

<table>
<thead>
<tr>
<th>Optimization Technique</th>
<th>Default Number of Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUREG, NRRIDG, and NEWRAP</td>
<td>50</td>
</tr>
<tr>
<td>QUANEW and DBLDOG</td>
<td>200</td>
</tr>
<tr>
<td>CONGRA</td>
<td>400</td>
</tr>
<tr>
<td>NMSIMP</td>
<td>1000</td>
</tr>
</tbody>
</table>

### MAXTESTLEV=n

specifies the maximum number of levels in an effect for which the server generates Type III tests. The idea behind the MAXTESTLEV= option is that testing effects for significance that have a large number of levels is typically not meaningful. The effects tend to be highly significant anyway, but determining the exact significance level is computationally intensive. The default value is 300 and implies that no test statistics are produced for any effect that has more than 300 levels.

Default 300

### NOCLPRINT <=n>

specifies the number of levels for each classification variables to show in the Class Level Information ODS table. If you do not specify the NOCLPRINT option, all unique values are shown in the order of the class variable levelization. If you specify NOCLPRINT=n, then the values are shown for those classification variables that have less than n levels only. The value for n must be at least 1.

If you specify the NOCLPRINT option without specifying a value for n, then n = 0 is assumed. This enables you to get a listing of the classification variables in the model. This might be useful if you did not identify classification variables explicitly—without listing their (possibly many) levels.

For example, the following Class Level Information table is displayed with NOCLPRINT=4. Because the number of levels for variable Smoking_Status exceeds 4, the values are not displayed.

### NOINT

suppresses the inclusion of an intercept in the model. By default, all models contain an intercept term.

### NOPREPARSE

prevents the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects
problems with the code, the error messages might not to be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

When you specify this option in the PROC IMSTAT statement, the option applies to all statements that can generate code. You can also exclude specific statements from pre-parsing by using the NOPREPARSE option in statements that allow temporary columns or the SCORE statement.

Alias NOPREP

NOSTDERR
prevents the computation of the covariance matrix and the standard errors of the parameter estimates. When you specify this option, the Type III tests for the model effects are also not available.

Alias NOSTD

OFFSET=variable-name
specifies the offset variable for the analysis. An offset variable can be thought of as a regressor variable whose regression coefficient is known to be 1. Offsets are used to shift the linear predictors by a certain amount. For example, an offset can be used to accommodate constants in the underlying model. In generalized linear models, offsets arise frequently when the data represents a value relative to some measure of size. For example, if you model the number of stops (Y) for each trip and the trips are of different length (t), then you are really interested in the random variable Y/t. The generalized linear model becomes as follows:

\[ g(\frac{\mu}{t}) = \eta \]

If you choose the log link function, this can be written as follows:

\[ \log \mu - \log t = \eta \]

\[ \log \mu = \eta + \log t \]

The value \( \log(t) \) is the offset of the model.

PARTITION<=partition-key>
specifies to fit the model separately for each value of the partition key. In other words, the partition variables function as automatic group-by variables for the request.

If you do not specify a value for partition-key, then the analysis is performed for all partitions. If you do specify a value, then the analysis is performed for the specified key value only. You can use the PARTITIONINFO statement to retrieve the valid partition-key values for a table.

Alias PART=

ROLEVAR=variable-name
specifies a variable in the in-memory table that defines whether an observation belongs to the training set, the validation set, or is to be excluded from the analysis. The role variable can have a numeric or character type, and it can be a temporary computed variable.
In case of a numeric role variable, the values of variable-name are interpreted as follows:

- value = 1: this observation is in the training set
- value = 2: this observation is in the validation set
- any other value: this observation is to be excluded from the analysis

In case of a character role variable, the values of variable-name are interpreted as follows:

- If the first non-blank character is 't' or 'T', then the observation is in the training set.
- If the first non-blank character is 'v' or 'V', then the observation is in the validation set.
- Any other value for the first non-blank character, including an all blank entry, leads to the exclusion of the observation from the analysis.

**SELECT**=(list-of-ODS-tables)

specifies the list of ODS tables that you want to display for the analysis. The specified list replaces the default tables that are generated by the server and displayed. See the **EXCLUDE** option for the list of default tables and the table names that you can display.

**TECHNIQUE**=value

specifies the optimization technique for the iterative model-fitting process. The valid values are as follows:

- CONGRA (CG): performs a conjugate-gradient optimization
- DBLDOG (DD): performs a version of the double-dogleg optimization
- NMSIMP (NS): performs a Nelder-Mead simplex optimization
- NONE: does not perform any optimization. This value can be useful to perform a grid search without optimization.
- NEWRAP (NRA): performs a (modified) Newton-Raphson optimization that combines a line-search algorithm with ridging
- NRRIDG (NRR): performs a (modified) Newton-Raphson optimization with ridging
- QUANEW (QN): performs a quasi-Newton optimization
- TRUREG (TR): performs a trust-region optimization

The factors that go into choosing a particular optimization technique for a particular problem are complex. Trial and error can be involved.

Default: NRRIDG
TEMPEXPRESS="SAS-expressions"

TEMPEXPRESS=file-reference

specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias TE=

TEMPNAMES=variable-name

TEMPNAMES=(variable-list)

specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

Alias TN=

TEMPTABLE

generates an in-memory temporary table from the result set. The IMSTAT procedure displays the name of the table and stores it in the &_TEMPLAST_ macro variable, provided that the statement executed successfully.

When the IMSTAT procedure exits, all temporary tables created during the IMSTAT session are removed. Temporary tables are not displayed on a TABLEINFO request, unless the temporary table is the active table for the request.

VALIDATE=f

specifies the proportion f in the validation data set.

Alias VALPROP=

Range 0 to 1

Interaction If you specify both the ROLEVAR= and the VALIDATE= options, then the ROLEVAR= setting supersedes the VALIDATE= option.

WEIGHT=variable-name

specifies the numeric variable to use as a weighing variable in solving the linear model.

Details

The GENMODEL statement offers the options that are necessary to formulate a model in the generalized linear model family, and to specify some other, closely related, models. For example, you can fit data to the generalized Poisson distribution for overdispersed Poisson counts. Or, you can use the shifted T distribution for symmetric data that is heavier in the tails than the normal distribution.

GLM Statement

The GLM statement is used to fit models that are similar to those handled by the GLM procedure. There are some important differences in syntax and functionality between the GLM procedure and the GLM statement in IMSTAT.

Syntax

GLM dependent-variable <(class-variables)> = model-effects </options>;
Required Arguments

dependent-variable
specifies the variable to model. This variable is also referred to as the response variable.

model-effects
specifies a list of variables to use for modeling the dependent variable.

Optional Argument

class-variables
specifies a list of variables to use as classification variables. The variables in this list take the place of the CLASS statement in traditional SAS procedures.

GLM Statement Options

ALLIDVARS
requests that all variables in the input table are treated as ID variables when a scoring table is produced. In other words, if this option is specified, all variables from the input table, including computed columns, are transferred to the scoring table.

ALPHA=number
specifies a number between 0 and 1 from which to determine the confidence level for approximate confidence intervals of the parameter estimates. The default is $\alpha = 0.05$, which leads to $100 \times (1 - \alpha)\% = 95\%$ confidence limits for the parameter estimates.

Default 0.05

CHISQ
requests that $p$-values in the table of parameter estimates and Type III tests are determined as probabilities under a $\chi^2$ distribution. This means that instead of two-sided $p$-values based on the $t$ distribution, the $p$-values are computed as two-sided probabilities under a standard normal distribution. Similarly, the assumption of $F$ distributions with finite denominator degrees of freedom is ignored in lieu of assuming infinite degrees of freedom.

CI
specifies to add confidence intervals to the table of parameter estimates. The confidence level is $100 \times (1 - \alpha)\%$ where $\alpha$ is determined by the ALPHA= option. The default value is $\alpha = 0.05$. This value is equivalent to a 95% confidence limit.

Default 0.05

CLASSFORMATS= ("format-name1", "format-name2" …)
specifies the formats for the classification variables in the model. If you do not specify the CLASSFORMATS= option, the default format is applied for the classification variable. That default format was determined when the table was originally loaded into the server. In the following example, the CLASSFORMAT= values apply to variables x1 and x2.

Alias CLASSFMT=

Example glm y (x1 x2) = x3-x7 / classformats= ("YN.", "F8.");

CODE <(code-generation-options)>
requests that the server produce SAS scoring code based on the actions that it performed during the analysis. The server generates DATA step code. By default, the code is replayed as an ODS table by the procedure as part of the output of the
statement. More frequently, you might want to write the scoring code to an external file by specifying options.

The scoring code computes the predicted value of the response variable on the data scale (the inverse link scale) and prefixes the name with "P_". For example, if the response variable is $Y$, the generated code stores the predicted value as $P_Y$. The name of the variable is truncated to fit within the SAS name length requirements.

**COMMENT**

specifies to add comments to the code in addition to the header block. The header block is added by default.

**FILENAME='path'**

specifies the name of the external file to which the scoring code is written. This suboption applies only to the scoring code itself. If you request that the server generate IMSTAT programming statements with the IMSTAT suboption, then these statements are saved as an ODS table.

Alias **FILE=**

**FORMATWIDTH=$k$**

specifies the width to use in formatting derived numbers such as parameter estimates in the scoring code. The server applies the BEST format, and the default format for code generation is BEST20.

Alias **FMTWIDTH=**

Range 4 to 32

**IMSTAT**

specifies to generate IMSTAT programming statements that reproduce the analysis in addition to the scoring code. For example, this option is helpful when you perform variable selection and you want to capture the modeling code that reflects only the selected variables.

**IMSTATONLY**

specifies to generate the IMSTAT programming statements only. No scoring code is produced.

**LINESIZE=$n$**

specifies the line size for the generated code.

Alias **LS=**

Default 72

Range 64 to 256

**NOTRIM**

requests that the comparison of the formatted values for class variables and group-by variables is based on the full format width with padding. By default, the leading and trailing blanks are removed from the formatted values.

**REPLACE**

specifies to overwrite the external file with the new contents if the file already exists. This option has no effect unless you specify the FILENAME= option.

**EXCLUDE=(list-of-ODS-tables)**

specifies the result tables that you want to exclude from being generated on the server and from being sent to the SAS session. The GLM statement can generate the following tables:
<table>
<thead>
<tr>
<th>Table Name</th>
<th>Table Alias</th>
<th>Description</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ModelInfo</td>
<td>Information about the model—constant across groups or partitions.</td>
<td>This table is shown by default.</td>
<td></td>
</tr>
<tr>
<td>ClassLevels</td>
<td>Class</td>
<td>Information about the classification variables, such as the number of levels and their values.</td>
<td>This table is shown when classification variables are present in the model.</td>
</tr>
<tr>
<td>Dimensions</td>
<td>Dim</td>
<td>Model dimensions</td>
<td>This table is shown by default.</td>
</tr>
<tr>
<td>FitStatistics</td>
<td>Fit</td>
<td>Fit statistics customary for generalized linear models</td>
<td>This table is shown when it is requested with the SELECT= option.</td>
</tr>
<tr>
<td>OverallAnova</td>
<td>GlobalAnova</td>
<td>Model, source, and error decomposition of variation</td>
<td>This table is shown when classification variables are present in the model.</td>
</tr>
<tr>
<td>ModelAnova</td>
<td>Anova</td>
<td>Variance decomposition with significance tests for all model effects</td>
<td>This table is shown when classification variables are present in the model.</td>
</tr>
<tr>
<td>ParmEstimates</td>
<td>ParameterEstimates Pest</td>
<td>The solutions for the linear model coefficients</td>
<td>This table is shown when there are no classification variables in the model.</td>
</tr>
<tr>
<td>Tests3</td>
<td></td>
<td>Type III tests of model effects</td>
<td>This table is shown when it is requested with the SELECT= option.</td>
</tr>
</tbody>
</table>

Whether a table is shown by default or not, you can request any table with the SELECT= option in the GLM statement. The Condition column in the table identifies when a table is produced by default. For example, if the model contains classification variables, the statement shows an OverallAnova and a ModelAnova table. If there are no classification variables, the statement shows a table of parameter estimates and no ANOVA tables.

**FORMATS=("format-specification"<,...,>)** specifies the formats for the GROUPBY variables. If you do not specify the FORMATS= option, or if you omit the entry for a GROUPBY variable, the default format is applied for that variable.

Enclose each format specification in quotation marks and separate each format specification with a comma.
Example

```plaintext
proc imstat data=lasr1.table1;
  statement / groupby=(a b) formats=('8.3', '$10');
quit;
```

**FREQ=variable-name**
specifies the numeric variable that provides frequencies for the analysis. For example, if the FREQ= variable has the value 5, then it implies that the record represents five such observations with identical values for the modeling variables. If you specify a FREQ= variable, then only the observations with a value that is not missing and greater than zero for the variable are used in the analysis.

**GROUPBY=(variable-list)**
specifies a list of variable names, or a single variable name, to use as GROUPBY variables in the order of the grouping hierarchy. If you do not specify any GROUPBY variable names, then the calculation is performed across the entire table—possibly subject to a WHERE clause.

**GROUPBYMODE= DATA | MODEL | LASR**
specifies the parallelization technique for group-by processing. The default is GROUPBYMODE=MODEL in which threads solve separate models following a lateral reconciliation of cross-product matrices. This mode is appropriate in situations with many groups and relatively small cross-product matrices. Model-parallel processing minimizes passes through the data.

Specify GROUPBYMODE=DATA to form the cross-product matrices in parallel across the data and one group at a time. This data-parallel technique is appropriate in situations with few groups and many observations per group or in applications with large cross-product matrices. Data-parallel processing consumes fewer resources than model-parallel processing but passes through the data more often.

If you specify GROUPBYMODE=LASR, then the server examines the data structure of the groups to select the parallelization mode.

Default  MODELMODEL

**GROUPFILTER=(filter-options)**
specifies a section of the group-by hierarchy to be included in the computation. With this option, you can request that the server performs the analysis for only a subset of all possible groupings. The subset is determined by applying the group filter to a temporary table that you generate with the GROUPBY statement.

You can specify the following suboptions in the GROUPFILTER option:

**DESCENDING**
specifies the top or the bottom section of the groupings to be collected. If the DESCENDING option is specified, the top LIMIT=n (where n > 0) groupings are collected. Otherwise, the bottom LIMIT=n groupings are collected.

 Alias DESC

**LIMIT=n**
specifies the maximum number of distinct groupings to be collected, where integer \( n \geq 0 \). If \( n \) is zero, then all distinct groupings (up to \( 2^{31}-1 \)) that satisfy the boundary constraints, such as LOWERSCORE=\( f \), are collected.

### CAUTION High Cardinality Data Sets
Setting \( n \) to zero with high-cardinality data sets can significantly delay the response of the server.
SCOREGT=f
specifies the exclusive lower bound for the numeric scores of the distinct
  groupings to collect.
  Alias SGT=

SCORELT=f
specifies the exclusive upper bound for the numeric scores of the distinct
groupings to collect.
  Alias SLT=

VALUEGT=("format-name1" <, "format-name2" …>)
specifies the exclusive lower bound of the group-by variable’s formatted values
  for the distinct groupings to collect.
  Alias VGT=

VALUELT=("format-name1" <, "format-name2" …>)
specifies the exclusive upper bound of the group-by variable’s formatted values
  for the distinct groupings to collect.
  Alias VLT=

TABLE=table-with-groupby-results
specifies the in-memory table from which to load the group-by hierarchy. If the
  TABLE= option is not specified, then all other GROUPFILTER= options are
  ignored.

The following program request all the groupings of State, City, and then
Trade_In_Model in the Cars_Program_All table. The groupings are ordered by
  the maximum value of New_Vehicle_Msrp for each grouping:

    proc imstat;
      table example.CARS_PROGRAM_ALL;
      groupby state city trade_in_model / TEMPTABLE
        WEIGHT=new_vehicle_MSRP
        AGG  =(MAX)
        ORDER =WEIGHT;
    run;

The TEMPTABLE option in the GROUPBY statement directs the server to save
all the groupings in a temporary in-memory table. The following DISTINCT
statement requests the count of the distinct unformatted values of Sales_Type for
each of the selected groupings of State, City, and Trade_In_Model.

    table example.CARS_PROGRAM_ALL;
    distinct sales_type / GROUPFILTER={
      table =mylasr.&_TEMPLAST_
        scoregt=40000
        valuegt="FL","Ft Myers",""
        limit  =20
        descending);
    run;

This example only considers groupings that have maximum values of the
New_Vehicle_Msrp above 40,000 and with formatted values that are less than
State="FL" and City="Ft Myers." The empty quotation marks result in no
restriction on Trade_In_Model values. These groupings are ordered according to
the maximum values of New_Vehicle_Msrp. Because of the DESCENDING option, this example collects the 20 top groupings within the specified group-by range for the DISTINCT analysis.

Interaction
If you specify the GROUPFILTER= option, then the GROUPBY= and FORMATS= options have no effect.

**IDVARS=(variable-list)**

**IDVARS=variable-name**

specifies the variables from the active table to transfer to the temporary table that is created by scoring the input table. This option has no effect unless the SCORE option is also specified. (See the SCORE option for details about which variables are added to the temporary table by default.) The IDVARS= option should be used to transfer additional columns from the input table to the scoring table.

**Alias**

ID=

**Tip**

Instead of this option, you can specify the ALLIDVARS option to transfer all variables from the input table to the scoring table.

**INCLUDEMISS**

specifies to treat missing values for classification variables as valid levels. If the INCLUDEMISS option is not specified, observations with missing values in the classification variables are not used in the analysis.

**INFORMATIVE**

requests that missing values are handled by modeling them through extra model effects. These effects consist of dummy variables that take on the value 1 when the value of a continuous model variable that is involved in the effect is missing. Otherwise, they are assigned the value 0. The missing value in the original model effect is replaced with the average value for the effect for the nonmissing values.

For continuous-by-class effects, such as A*x, where A is a classification variable and x is a continuous variable, informative missingness creates multiple dummy columns and substitutes the effect mean of x that corresponds to the respective level of A.

Specifying the INFORMATIVE option implies the INCLUDEMISS option. That is, when you choose to model informative missingness, then missing values for classification variables are treated as valid levels. For more information, see “Informative Missingness” on page 117.

**Alias**

INFORMMISS

**KEYORDER**

requests that the results for a partitioned analysis are displayed in the order of the partition keys. If this option is not specified, then results are displayed by using the partitions on the first worker node followed by the partitions on the second node, and so on. Without this option, the results are likely to have random ordering of the partitions. The KEYORDER option makes result collection less efficient but produces a natural, predictable order.

**MAXTESTLEV=n**

specifies the maximum number of levels in an effect for which the server generates Type III tests. The idea behind the MAXTESTLEV= option is that testing effects for significance that have a large number of levels is typically not meaningful. The effects tend to be highly significant anyway, but determining the exact significance level is computationally intensive. The default value is 300 and implies that no test statistics are produced for any effect that has more than 300 levels.
**NAME=SAS-name**

specifies the name to use for identifying the model in the server output and in the temporary table of results generated by the TEMPTABLE option. SAS name rules apply. For example, the following statements add the 'Model' entry to the ModelInformation table.

```sas
proc imstat;
  table hps.iris;
  glm sepalwidth = sepallength / name = FirstModel;
run;
```

### NOCLPRINT <=n>

specifies the number of levels for each classification variables to show in the Class Level Information ODS table. If you do not specify the NOCLPRINT option, all unique values are shown in the order of the class variable levelization. If you specify NOCLPRINT=n, then the values are shown for those classification variables that have less than n levels only. The value for n must be at least 1.

If you specify the NOCLPRINT option without specifying a value for n, then n = 0 is assumed. This enables you to get a listing of the classification variables in the model. This might be useful if you did not identify classification variables explicitly—without listing their (possibly many) levels.

For example, the following Class Level Information table is displayed with NOCLPRINT=4. Because the number of levels for variable Smoking_Status exceeds 4, the values are not displayed.

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight_Status</td>
<td>3</td>
<td>Normal Overweight Underweight</td>
</tr>
<tr>
<td>Smoking_Status</td>
<td>5</td>
<td>not printed</td>
</tr>
</tbody>
</table>

### NOINT

suppresses the inclusion of an intercept in the model. By default, the GLM statement adds an intercept as the first model effect to the model. Exclusion of the intercept is useful in certain models to achieve a desired interpretation of the model effects.

For example, the following code sample shows a cell-means model where the coefficients in the β vector estimate the means of Y in the groups associated with levels of A.

```sas
glm y (A) = A / noint;
```

**NOPREPARE**

prevents the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.
When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects problems with the code, the error messages might not be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

When you specify this option in the PROC IMSTAT statement, the option applies to all statements that can generate code. You can also exclude specific statements from pre-parsing by using the NOPREPARSE option in statements that allow temporary columns or the SCORE statement.

Alias NOPREP

PARTITION<=partition-key>

specifies to fit the model separately for each value of the partition key. In other words, the partition variables function as automatic group-by variables for the request.

If you do not specify a value for partition-key, then the analysis is performed for all partitions. If you do specify a value, then the analysis is performed for the specified key value only. You can use the PARTITIONINFO statement to retrieve the valid partition-key values for a table.

Alias PART=

ROLEVAR=variable-name

specifies a variable in the in-memory table that defines whether an observation belongs to the training set, the validation set, or is to be excluded from the analysis. The role variable can have a numeric or character type, and it can be a temporary computed variable.

In case of a numeric role variable, the values of variable-name are interpreted as follows:

- value = 1: this observation is in the training set
- value = 2: this observation is in the validation set
- any other value: this observation is to be excluded from the analysis

In case of a character role variable, the values of variable-name are interpreted as follows:

- If the first non-blank character is 't' or 'T', then the observation is in the training set.
- If the first non-blank character is 'v' or 'V', then the observation is in the validation set.
- Any other value for the first non-blank character, including an all blank entry, leads to the exclusion of the observation from the analysis.

Alias ROLE=

Interactions You can divide the data at random into training and validation sets by providing the VALIDATE= and SEED= options.
If you specify both the ROLEVAR= and the VALIDATE= options, then the ROLEVAR= setting supersedes the VALIDATE= option.

**SCORE <(score-statistic1 score-statistic2…)>**

requests that the active table be scored after the model is fit and the results be stored in a temporary table. The server automatically adds all model variables to the temporary table with the score results. These results include the response variable, the class variables, all explanatory variables from which effects are formed, and the WEIGHT=, and FREQ= variables.

In addition, if the active table is partitioned or ordered, the partition variables and order-by variables are transferred from the input table to the temporary table. The temporary table is partitioned and ordered in the same way as the active table.

If the analysis uses the GROUPBY= option, the variables in the group-by list are also transferred to the scoring table. If you want to transfer additional variables, you can specify them with the IDVARS= option.

If you do not specify the list of score statistics, default statistics are computed. These statistics are identified with Yes in the Default column in the table below. You can request that the following statistics be computed for each observation:

<table>
<thead>
<tr>
<th>Keyword and Aliases</th>
<th>Column Name</th>
<th>Description</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRED, PREDICTED, MEAN</td>
<td><em>PRED</em></td>
<td>Predicted value</td>
<td>Yes</td>
</tr>
<tr>
<td>RESID, RESIDUAL, R</td>
<td><em>RESID</em></td>
<td>Raw residual (observed - predicted)</td>
<td>Yes</td>
</tr>
<tr>
<td>STUDENT</td>
<td><em>STUDENT</em></td>
<td>Studentized residual</td>
<td>Yes</td>
</tr>
<tr>
<td>RSTUDENT</td>
<td><em>RSTUDENT</em></td>
<td>Studentized residual with the current observation removed</td>
<td>Yes</td>
</tr>
<tr>
<td>LEVERAGE, H</td>
<td><em>LEVERAGE</em></td>
<td>Leverage value of the observation</td>
<td>Yes</td>
</tr>
<tr>
<td>STDP</td>
<td><em>STDP</em></td>
<td>Standard error of the mean predicted value</td>
<td>No</td>
</tr>
<tr>
<td>STDR</td>
<td><em>STDR</em></td>
<td>Standard error of the residual</td>
<td>No</td>
</tr>
<tr>
<td>STDI</td>
<td><em>STDI</em></td>
<td>Standard error of the (individual) predicted value</td>
<td>No</td>
</tr>
<tr>
<td>LCLM, LOWERMEAN</td>
<td><em>LCLM</em></td>
<td>Lower confidence limit for the mean of the predicted value</td>
<td>No</td>
</tr>
<tr>
<td>Keyword and Aliases</td>
<td>Column Name</td>
<td>Description</td>
<td>Default</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------</td>
<td>---------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>UCLM, UPPERMEAN</td>
<td><em>UCLM</em></td>
<td>Upper confidence limit for the mean of the predicted value</td>
<td>No</td>
</tr>
<tr>
<td>LCL, LOWERPRED</td>
<td><em>LCL</em></td>
<td>Lower confidence limit for the predicted value</td>
<td>No</td>
</tr>
<tr>
<td>UCL, UPPERPRED</td>
<td><em>UCL</em></td>
<td>Upper confidence limit for the predicted value</td>
<td>No</td>
</tr>
<tr>
<td>COOKD, COOKSD</td>
<td><em>COOKD</em></td>
<td>Cook's $D$ influence measure</td>
<td>No</td>
</tr>
<tr>
<td>DFFITS</td>
<td><em>DFFITS</em></td>
<td>Standardized influence of the observation on predicted value</td>
<td>No</td>
</tr>
<tr>
<td>COVRATIO</td>
<td><em>COVRATIO</em></td>
<td>Standardized influence of the observation on the covariance matrix of the parameter estimates</td>
<td>No</td>
</tr>
<tr>
<td>LIKEDIST, LD</td>
<td><em>LIKEDIST</em></td>
<td>Displacement (distance) of log-likelihood when the observation is removed (assuming normal distribution)</td>
<td>No</td>
</tr>
</tbody>
</table>

If you specify SCORE(_ALL_), then the server calculates and adds all the possible output statistics to the temporary table. The confidence levels for the LCLM, LCL, UCLM, and UCL confidence bounds are determined from the significance level specified in the ALPHA= option as $(100 (1-\alpha)\%)$. The default is $\alpha = 0.05$.

The server determines the column names for the output statistics. This differs from many SAS procedures where you can specify the name for the statistic.

**SEED=number**

specifies the random number seed for generating random numbers. The random number is used to determine whether an observation belongs to the training or validation data set. The SEED= option has no effect unless you specify the VALPROP= option. If the specified number is negative or zero, the random number generation is based on the computer clock of the server—this generates a non-reproducible random number sequence.

**SELECT=(list-of-ODS-tables)**

specifies the list of ODS tables that you want to display for the analysis. The specified list replaces the default tables that are generated by the server and
displayed. See the EXCLUDE= option for the list of default tables and the table names that you can display.

**SHOWSELECTED**
requests that the server perform variable selection for the model. A backward selection method is used, where the significance level for an effect to remain in the model is determined by the SLSTAY= option. This option performs variable selection like the VARSEL option, but in contrast to the latter option, it displays output only for the selected effects.

Alias SHOWSEL

**SLSTAY=**\( \alpha \)
specifies the significance level used in determining whether effects should stay in the model during variable selection.

Default 0.1

Range 0 to 1

**TEMPEXPRESS=**"SAS-expressions"

**TEMPEXPRESS=**file-reference
specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias TE=

**TEMPNAMES=**variable-name
**TEMPNAMES=**(variable-list)
specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

Alias TN=

**TEMPTABLE**
generates an in-memory temporary table from the result set. The IMSTAT procedure displays the name of the table and stores it in the &_TEMPLAST_ macro variable, provided that the statement executed successfully.

When the IMSTAT procedure exits, all temporary tables created during the IMSTAT session are removed. Temporary tables are not displayed on a TABLEINFO request, unless the temporary table is the active table for the request.

**VALIDATE=**\( f \)
specifies the proportion \( f \) in the validation data set.

Alias VALPROP=

Range 0 to 1

Interaction If you specify both the ROLEVAR= and the VALIDATE= options, then the ROLEVAR= setting supersedes the VALIDATE= option.

**VARSELECTION**
requests that the server perform variable selection for the model. A backward selection method is used, where the significance level for an effect to remain in the model is determined by the SLSTAY= option. In contrast to the SHOWSEEL option, all effects are reported in the IMSTAT output.
Alias VARSEL

**WEIGHT=variable-name**

specifies the numeric variable to use as a weighing variable in solving the linear model.

When you specify a WEIGHT= variable, the normal equations

\[ X'Xb = X'Y \]

are replaced by

\[ X'WXb = X'WY \]

where \( W \) is a diagonal matrix with the values of the variable specified in the WEIGHT= option on the diagonal. Only the observations with a weight value that is not missing and greater than zero are used in the analysis.

**Details**

**Basic Syntax**

The basic syntax of the GLM statement requires that you specify the response variable (the dependent variable), an equal sign (=) and then the model effects. The dependent variable must be numeric. In contrast to the GLM procedure, you can specify only one dependent variable in the GLM statement of the IMSTAT procedure.

The underlying statistical model of the GLM statement is as follows:

\[ Y = XB + e \]

where \( Y \) is an \((n \times 1)\) random vector of the dependent variable, \( X \) is an \((n \times p)\) design matrix, \( B \) is a \((p \times 1)\) vector of coefficients, and \( e \) is an \((n \times 1)\) vector of random disturbances (errors). The key assumptions of a GLM type of model are that the errors \( e \) are uncorrelated, homoscedastic (have the same variance \( \sigma^2 \)), and have zero mean. If these assumptions are met, then the model is correct and the elements of the vector \( Y - XB \) are stochastically unrelated. The goals of the GLM analysis are as follows:

- to estimate the unknowns \( \beta \) and \( \sigma^2 \)
- to diagnose the appropriateness of the specified model
- to select appropriate variables and terms for the \( X \) matrix
- to predict the average response and unobserved values of the response variable with confidence
- to test hypotheses about the elements of \( \beta \) provided that the model is acceptable

A model effect is a syntactic expression of how one or more variables act together to define columns in the design matrix \( X \) of the linear statistical model. In other words, how you specify the model-effects on the right-hand side of the GLM statement affects how the server constructs the \( X \) matrix and how you interpret the results of the analysis pertaining to the contributing variables.

There are a few basic types of effects:

- the intercept is included by default in every model. It is then the leading effect in \( X \) and simply adds a column of ones to this matrix. The intercept can be approximately interpreted as an adjustment for the mean of the response variable.
- a continuous effect consists of only numeric non-class variables. The simplest continuous effect contains only one variable. For example, if you add the numeric
variable Age to the model, you are adding a continuous effect. If the variable Height is not a classification variable, and you add the term Age*Height to the model, you are adding a continuous interaction effect.

- a classification variable is a variable that is used in the model not through its raw values, but through an encoding of its unique (formatted) values. For example, if variable Gender is used as a classification variable in the model, then it represents two levels.

- a classification effect is a model effect that contains one or more classification variables. A "pure" classification effect comprises only classification variables, a continuous classification effect also involves some continuous variables.
  - If A and B are classification variables, and X and Z are non-classification variables, then the effects A, B, and A*B are pure classification effects, termed the A and B main effect and the interaction of A and B, respectively. The effect A*Z would be a continuous classification effect.
  - Effect A is said to be nested within effect B if levels of A within one level of B do not mean the same thing for other levels of B. Nested effects are expressed with parenthetical notation. For example, if City and State are classification variables, then City(State) represents the nested effect of cities within states. One example of appropriate nesting is when city #1 in Alaska refers to a different city than city #1 in Colorado.
  - Two effects are said to be crossed if the levels of one effect retain their interpretation across the levels of the other effect. If Married is a classification variable that groups individuals into married and unmarried status, and Gender is a two-level variable, the Gender*Married effect is crossed, because a man in the unmarried group is also a man in the married group.

Deciding which variables to involve in a statistical model and how the variables should act and interact is key in modeling. The following rules apply for the GLM statement in the IMSTAT procedure:

- A character variable that is used in a model effect is treated as a classification variable.
- A numerical variable that is used in a model effect is treated as a non-classification variable.
- All variables explicitly listed in the optional variable list that follows the specification of the dependent variable in the GLM statement are classification variables.
- The role of temporary computed variables is determined by the data type.
  - If the computed column is of character type, then it is automatically added to the model as a classification variable if it appears in a model effect.
  - If the computed column is of numeric type, then it is treated as a classification variable only if it is specified in the list of `class-variables`.

The following example of modeling the Sashelp.Class data set shows how variables act and interact. The following GLM statement models a student's height as a function of his or her weight and gender:

```plaintext
glm height = weight sex;
```

The Sex variable, because it has character type, is treated as a classification variable. The following GLM statement is equivalent, but it expresses the classification variables explicitly:

```plaintext
glm height(sex) = weight sex;
```
Computed columns can also be used. The following example uses the same variables that were used in the previous examples, but specifies them as computed columns to demonstrate the syntax:

```
table lasr.class(tempnames=(t1 t2));
glm height = t1 t2 / tn=(t1 t2) te="t1=weight; t2=sex;";
```

The analysis that uses the computed columns (T1 and T2) is identical to the previous GLM statement. This is because T2 would be discovered by the server to be of character type and would be added automatically to the list of classification variables.

**Informative Missingness**

The concept of informative missingness is one way to account for missing values in statistical analyses and, in particular, statistical modeling. Missing values are a problem because they reduce the amount of available data. When working with classification variables (factors, which are levelized variables), a missing value can be treated as an actual level of the variable and can participate in the analysis.

When continuous variables have missing values, however, the observation is removed from the analysis. In data with many missing values, this can reduce the amount of available data greatly, and the sets of observations used in one model versus another model can vary based on which variables are included in the model.

Of course there are many reasons for missing values and substituting values for missing values has to be done with caution. For example, the famous Framingham Heart study data set contains 5,209 observations on subjects in a longitudinal study that helped understand the relationship between smoking, cholesterol, and coronary heart disease. One of the variables in the data set is AgeCHDdiag. This variable represents the age at which a patient was diagnosed with coronary heart disease (CHD). If you include this variable in a statistical model, only 1,449 observations are available, since the value cannot be observed unless a patient has experienced CHD. Including this variable acts as a filter that reduces the analysis set to the subjects with CHD. We cannot impute the value for subjects where the variable has a missing value, because we cannot impute an age at which someone who has not had CHD would have contracted coronary heart disease.

With informative missingness, we are not as much substituting imputed values for the missing values, as we are modeling the missingness. Consider a simple linear regression model:

\[
y = \beta_0 + \beta_1 x + \epsilon
\]

Suppose that some of the values for the regressor variable \( x \) are missing. The fitted model uses only observations for which \( y \) and \( x \) have been observed.

In order to predict the outcome \( y \) for an observation with missing \( x \), we either assume that \( y \) is missing or substitute a value for the missing \( x \), such as the average value, \( \bar{x} \).

Because the estimate for the intercept is \( \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \) in the simple linear regression model, the predicted value would be the average response of the nonmissing values, \( \bar{y} \).

With informative missingness, we extend the model by adding extra effects for each effect that contains at least one continuous variable. In the simple linear regression model, we add one column to the model and slightly change the content of the \( x \) variable:

\[
y = \beta_0 + \beta_1 x^* + \beta_2 x^{**} + \epsilon_1
\]
The variable $x^*$ contains the original values of $x$ if these are not missing, and the average of $x$ otherwise:

$$
x^* = \begin{cases} 
  x & \text{if } x \text{ is not missing} \\
  \bar{x} & \text{otherwise}
\end{cases}
$$

The variable $x^{**}$ is a dummy variable with value 1 when $x$ is missing, and zero otherwise:

$$
x^{**} = \begin{cases} 
  1 & \text{if } x \text{ is missing} \\
  0 & \text{otherwise}
\end{cases}
$$

The fitted model is not the same model that results from substituting $\bar{x}$ for the missing values during training. This can be seen, since the model that simply substitutes $\bar{x}$ for the missing values is as follows:

$$
y = \beta_0 + \beta_1 x^* + \epsilon_2
$$

The informative missing model has an extra parameter, and unless all values of $x^{**}$ are zero—in which case there are no missing values—the informative missing model has a higher $R^2$ value, because it picks up more variation.

The parameter estimate for $\beta_2$ measures the amount by which the predicted value differs from a predicted value at $\bar{x}$.

---

**GROUPBY Statement**

The GROUPBY statement derives the grouping hierarchy of the distinct formatted values for the specified variables. If no list of variable names is specified, the grouping hierarchy is computed for all variables in the active table. The statement can return a section of all distinct groupings to the client or save the entire grouping set as a temporary table in the server.

**Syntax**

GROUPBY <variable-list> /options;

**GROUPBY Statement Options**

AGGREGATE=(aggregation-methods)

Lists the aggregator on which the ordering of the result set is based.

Lists the aggregator for which the values of the WEIGHT variable are rolled up into a rank order score, provided that a WEIGHT= variable is specified. If no WEIGHT= variable is specified, then the aggregator specification is ignored.

The available aggregation methods for the GROUPBY statement are as follows:

- MAX: maximum value
- MEAN: arithmetic mean
- MIN: minimum value
- N: counts the nonmissing values of the weight variable
- SUM: sum of the weight values
DESCENDING
specifies to arrange the returned grouping hierarchy of the variables in descending order of the item rankings. If this option is not specified, the returned items are arranged in ascending order. When combined with the LIMIT=n option, the GROUPBY statement can either return the top n or the bottom n distinct groupings.

Interaction
The DESCENDING option is ignored if the TEMPTABLE option is specified.

FREQ=variable-name
specifies the numeric frequency variable that is used to compute the ranking of a distinct grouping. When this option is specified, the AGGREGATE= and WEIGHT= options are ignored. The following GROUPBY statement requests the top 5 groupings of Region and then Product from the Prdsale table. The groupings are rank ordered by the sum of the Actual column:

Example
proc imstat data = mylasr.prdsale;
    groupby region product / freq = actual limit = 5;
run;

LIMIT=n
specifies the maximum number of distinct groupings to be returned. When combined with the DESCENDING option, the GROUPBY statement can either return the top n or the bottom n distinct groupings. The value for n must be a positive integer. For example, the commands below return the bottom 5 groupings according to their Score values:

Default
0

Interaction
This option is ignored if the TEMPTABLE option is specified.

Tip
If n is zero, then all distinct groupings are returned (up to $2^{31}-1$). With high-cardinality data sets, setting n to zero can significantly delay the response of the server.

Example
proc imstat data = mylasr.prdsale;
    groupby region product / weight = actual
        agg     = max
        valuegt = ("West", "Chair")
        limit=5;
run;

NOMISSING
specifies that missing values are excluded in the determination of GROUPBY values. By default, levels with missing values are included.

Alias
NOMISS

NOTEMPPART
specifies that the temporary table that is generated by the TEMPTABLE option is not partitioned by the group-by variables. When you create a temporary table with the GROUPBY statement, by default, the server partitions the table and each partition has a single row. When the number of groups is large, this results in many tiny
partitions and requires additional memory resources to store the partition information for the temporary table.

By specifying this option, the temporary table is organized similarly to the default table, but it is not partitioned. This also enables more efficient processing of the table in threaded computations. For example, it is more efficient if you were to add computed columns to the table that you want to use as dimension keys in subsequent SCHEMA statements.

**ORDER=rank-order-type**

specifies the rank ordering to use for sorting the distinct groupings. The following rank-order types are valid in the GROUPBY statement:

- **FREQ**  
  frequency count of the variables
- **VALUE**  
  formatted values of the variables
- **WEIGHT**  
  aggregate values of the WEIGHT= variables

Default **FREQ**

**PARTITION <=partition-key>**

When you specify this option and the table is partitioned, the results are calculated separately for each value of the partition key. In other words, the partition variables function as automatic GROUPBY variables. This mode of executing calculations by partition is more efficient than using the GROUPBY= option. With a partitioned table, the server takes advantage of knowing that observations for a partition cannot be located on more than one worker node.

If you do not specify a **partition-key**, the analysis is performed for all partitions. If you do specify a **partition-key**, the analysis is carried out for the specified key value only. You can use the PARTITIONINFO statement to retrieve the valid partition key values for a table.

You can specify a **partition-key** in two ways. You can supply a single quoted string that is passed to the server, or you can specify the elements of a composite key separated by commas. For example, if you partition a table by variables GENDER and AGE, with formats $1 and BEST12, respectively, then the composite partition key has a length of 13. You can specify the partition for the 11-year-old females as follows:

```
statement / partition="F          11"; /* passed directly to the server */
statement / partition="F","11";        /* composed by the procedure */
```

If you choose the second format, the procedure composes a key based on formatting information from the server.

Alias **PART=**

**SAVE=table-name**

saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for **table-name** must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent **SAVE=** options.

**SETSIZE**

requests that the server estimate the size of the result set. The procedure does not create a result table if the SETSIZE option is specified. Instead, the procedure reports the number of rows that are returned by the request and the expected memory consumption for the result set (in KB). If you specify the SETSIZE option, the SAS
log includes the number of observations and the estimated result set size. See the following log sample:

NOTE: The LASR Analytic Server action request for the STATEMENT statement would return 17 rows and approximately 3.641 kBytes of data.

The typical use of the SETSIZE option is to get an estimate of the size of the result set in situations where you are unsure whether the SAS session can handle a large result set. Be aware that in order to determine the size of the result set, the server has to perform the work as if you were receiving the actual result set. Requesting the estimated size of the result set does consume resources on the server. The estimated number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.

**SCOREGT=f**
specifies the exclusive lower bound of the numeric rank order scores of the distinct groupings to return. All distinct groupings with numeric rank order scores that are greater than \( f \) are returned.

Alias \( SGT= \)

Interaction This option is ignored if the TEMPTABLE option is specified.

**SCORELT=f**
specifies the exclusive upper bound of the numeric rank order scores of the distinct groupings to return. All distinct groupings with numeric rank order scores that are less than \( f \) are returned.

Alias \( SLT= \)

Interaction This option is ignored if the TEMPTABLE option is specified.

**TEMPEXPRESS=**

- **"SAS-expressions"**
- **file-reference**
specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias \( TE= \)

**TEMPNAMES=**

- **variable-name**
- **(variable-list)**
specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

Alias \( TN= \)

**TEMPTABLE**
generates an in-memory temporary table from the result set. The IMSTAT procedure displays the name of the table and stores it in the \&_TEMPLAST_ macro variable, provided that the statement executed successfully.

When the IMSTAT procedure exits, all temporary tables created during the IMSTAT session are removed. Temporary tables are not displayed on a TABLEINFO request, unless the temporary table is the active table for the request.
VALUEGT="("format-specification","…")
    specifies the exclusive lower bound of the variable’s formatted values of the distinct
groupings to return. All distinct groupings with formatted values for the variable that
are lexicographically greater than the specified bound are returned.

Alias     VGT=

Interaction This option is ignored if the TEMPTABLE option is specified.

VALUELT="("format-specification","…")
    specifies the exclusive upper bound of the variable’s formatted values of the distinct
groupings to return. All distinct groupings with formatted values for the variable that
are lexicographically less than the specified bound are returned.

Alias     VGT=

Interaction This option is ignored if the TEMPTABLE option is specified.

WEIGHT=variable-name
    specifies the numeric weight variable to use for computing the rank order score of a
distinct grouping.

Interaction The WEIGHT= and AGGREGATE= option have no effect unless you
specify ORDER=WEIGHT.

HISTOGRAM Statement
The HISTOGRAM statement calculates a histogram table for numeric variables.

Syntax
HISTOGRAM <variable-list> / options;

Optional Argument
variable-list
    specifies a single variable or a list of numeric variables. Separate each variable name
by at least one space. If you do not specify this option, a histogram table is calculated
for each numeric variable.

HISTOGRAM Statement Options
BINVALS=list-of-values
    specifies an array of NBINS lower bin boundaries as list-of-values. The histogram
binning then uses those values strictly and does not alter them so that they are
equally spaced (or “nice”). This option is useful to compute a histogram with bins
that are the same as those of another histogram so that the values can be compared or
overlaid. The bins do not need to be equally spaced.

EQUALFREQ
    specifies to create bins such that each bin contains the same fraction of the data.

Alias     EQUAL
MAX=number
specifies the upper end of the range to determine the histogram bins. By default, the maximum value is determined from the data (subject to the WHERE clause). The bins of the histogram can extend beyond the extreme values when the "nice-ing" algorithm places bin boundaries on numbers that are convenient to label on axes.

MIN=number
specifies the lower end of the range to determine the histogram bins. By default, the minimum value is determined from the data (subject to the WHERE clause). The bins of the histogram can extend beyond the extreme values when the "nice-ing" algorithm places bin boundaries on numbers that are convenient to label on axes.

NBINS=k
specifies the number of bins to use for calculating the histogram.

NOEMPTYBIN
prevents bins without observations from being displayed. The leading and trailing empty bins are trimmed. Any internal empty bins are combined into the first non-empty bin to the immediate right. The mid-value of the bin into which the empty bins are combined is not adjusted. If the mid-value is not missing, then you can use the asymmetry of a bin as an indicator that it was combined with empty bins.

NONICE
specifies that the "nice-ing" algorithm is suspended. The boundaries of the histogram are based on the actual range of the data (subject to the WHERE clause) or on the MIN= and MAX= values that you specify. The bin boundaries are not guaranteed to fall on "nice" values.

OUTLIERBIN
specifies that outliers are placed in special bins in the two tails. Outliers with values that fall below Q1 – 1.5*IQR are placed in the left-most bin. Outliers with a value that is above Q3 + 1.5*IQR are placed in the right-most bin. IQR is the inter-quartile range, which covers the central 50% of the distribution of the variable. The mid-value reported by the IMSTAT procedure can be used as an indicator whether a bin is an outlier bin. The mid-value is set to 1 for an outlier bin and set to missing otherwise.

Interaction  This option is ignored if you specify the EQUALFREQ option.

ROUNDINGFACTOR=value
specifies the factor to use for rounding up internal bin boundaries. The lower bound of the left-most bin and the upper bound of the right-most bin are not rounded. For example, when you work with prices in dollars, specifying ROUNDINGFACTOR=0.01 rounds the bin boundaries to cents. In the event that the specified rounding factor is greater than the bin width and multiple bins round up to the same number, the bins are collapsed into a single bin.

SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

TEMPEXPRESS="SAS-expressions"
TEMPEXPRESS=file-reference
specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias  TE=
TEMPNAMES=variable-name
TEMPNAMES=(variable-list)
specifies the list of temporary variables for the request. Each temporary variable
must be defined through SAS statements that you supply with the TEMPEXPRESS=
option.

Alias        TN=

KDE Statement

The KDE statement calculates kernel-density estimates of the distribution of one or more numeric variables
from an in-memory table. You can choose between normal, tricube, and quadratic kernel functions. The
default is to use a normal kernel function. The number of points returned are determined by the center
region of a multi-threaded, inverse finite Fourier transform.

Syntax

KDE variable-list </options>;

Required Argument

variable-list
    specifies a one or more numeric variables.

KDE Statement Options

BANDWIDTH=b
    specifies the standardized bandwidth of the kernel function. The default bandwidths
are optimal values that minimize the asymptotic mean integrated squared errors of
the kernel function. The actual bandwidth for the kernel estimator is a multiple of the
standardized bandwidth, the inter-quartile range of the data, and n^{-1/5}. Larger values
for bandwidth result in smoother density estimates. However, specifying a
bandwidth that is too large can result in density estimates that omit important aspects
of the distribution at finer granularity.

KERNEL= NORMAL | TRICUBE | QUADRATIC
    specifies the kernel function.

    Default        NORMAL

MAX=number
    specifies the largest value to consider in the density calculation. If a value is not
specified, then the largest value in the data range is used, subject to the WHERE
clause.

    Alias        UPPER=

MIN=number
    specifies the smallest value to consider in the density calculation. If a value is not
specified, the smallest value in the data range is used, subject to the WHERE clause.

    Alias        LOWER=
MULTIPLER=number
   specifies a scaling factor for the calculated density.

   Default  1

NPOINTS=n
   specifies the number of points from which to calculate the center region of the
   inverse finite Fourier transform. The value of the NPOINTS= option is adjusted to
   the largest integer of power of 2 that is equal to or smaller than n. For example,
   specifying NPOINTS=40 is adjusted to 32. The number of density points returned
   depends on the distribution of the data.

   Default  512
   Range   16 to 512

SAVE=table-name
   saves the result table so that you can use it in other IMSTAT procedure statements
   like STORE, REPLAY, and FREE. The value for table-name must be unique within
   the scope of the procedure execution. The name of a table that has been freed with
   the FREE statement can be used again in subsequent SAVE= options.

SCALE= PERCENT | COUNT | PROPORTION
   specifies the units in which the density is calculated.

TEMPEXPRESS="SAS-expressions"
   specifies either a quoted string that contains the SAS expression that defines the
   temporary variables or a file reference to an external file with the SAS statements.

   Alias  TE=

TEMPNAMES=variable-name
   TEMPNAMES=(variable-list)
   specifies the list of temporary variables for the request. Each temporary variable
   must be defined through SAS statements that you supply with the TEMPEXPRESS=
   option.

   Alias  TN=

LOGISTIC Statement

The LOGISTIC statement can model binary data with logit, probit, log-log, and complementary log-log link
functions. It can also model binomial data with the same set of link functions.

Syntax

LOGISTIC dependent-variable <(class-variables)> = model-effects </ options>
LOGISTIC event-variable / trial-variable <(class-variables)> = model-effects </ options>
Required Arguments

dependent-variable
specifies the variable to model. This variable is also referred to as the response variable.

TIP The LOGISTIC statement produces a response profile table that shows the ordered values of the dependent variable. By default, the smaller ordered value is the event that is modeled. For example, if the dependent variable has values 0 and 1, the statement models the probability that the dependent variable takes on the value 0. You can change the ordering of the dependent variable values with the DESCENDING option.

event-variable
specifies the name of the variable that indicates the count of positive responses.

model-effects
specifies a list of variables to use for modeling the dependent variable.

trial-variable
specifies the name of the variable that indicates the total number of trials.

Optional Argument

class-variables
specifies a list of variables to use as classification variables. The variables in this list take the place of the CLASS statement in traditional SAS procedures.

LOGISTIC Statement Options

ALLIDVARS
requests that all variables in the input table are treated as ID variables when a scoring table is produced. In other words, if this option is specified, all variables from the input table, including computed columns, are transferred to the scoring table.

ALPHA=number
specifies a number between 0 and 1 from which to determine the confidence level for approximate confidence intervals of the parameter estimates. The default is α = 0.05, which leads to 100 x (1- α)% = 95% confidence limits for the parameter estimates.

Default 0.05

CI
specifies to add confidence intervals to the table of parameter estimates. The confidence level is 100*(1-α)% where α is determined by the ALPHA= option. The default value is α = 0.05. This value is equivalent to a 95% confidence limit.

Default 0.05

CLASSFORMATS=("format-name1"<, "format-name2" …>)
specifies the formats for the classification variables in the model. If you do not specify the CLASSFORMATS= option, the default format is applied for the classification variable. That default format was determined when the table was originally loaded into the server. In the following example, the CLASSFORMAT= values apply to variables x1 and x2.

Alias CLASSFMT=

Example logistic y (x1 x2) = x3-x7 / classformats="YN.", "F8.");
CODE <(code-generation-options)> requests that the server produce SAS scoring code based on the actions that it performed during the analysis. The server generates DATA step code. By default, the code is replayed as an ODS table by the procedure as part of the output of the statement. More frequently, you might want to write the scoring code to an external file by specifying options.

The scoring code computes the predicted value of the response variable on the data scale (the inverse link scale) and prefixes the name with "P_.". For example, if the response variable is Y, the generated code stores the predicted value as P_Y. The name of the variable is truncated to fit within the SAS name length requirements.

COMMENT specifies to add comments to the code in addition to the header block. The header block is added by default.

FILENAME=path specifies the name of the external file to which the scoring code is written. This suboption applies only to the scoring code itself. If you request that the server generate IMSTAT programming statements with the IMSTAT suboption, then these statements are saved as an ODS table.

Alias FILE=

FORMATWIDTH=k specifies the width to use in formatting derived numbers such as parameter estimates in the scoring code. The server applies the BEST format, and the default format for code generation is BEST20.

Alias FMTWIDTH=
Range 4 to 32

IMSTAT specifies to generate IMSTAT programming statements that reproduce the analysis in addition to the scoring code. For example, this option is helpful when you perform variable selection and you want to capture the modeling code that reflects only the selected variables.

IMSTATONLY specifies to generate the IMSTAT programming statements only. No scoring code is produced.

LINESIZE=n specifies the line size for the generated code.

Alias LS=
Default 72
Range 64 to 256

NOTRIM requests that the comparison of the formatted values for class variables and group-by variables is based on the full format width with padding. By default, the leading and trailing blanks are removed from the formatted values.

REPLACE specifies to overwrite the external file with the new contents if the file already exists. This option has no effect unless you specify the FILENAME= option.
**EXCLUDE=(list-of-ODS-tables)**
specifies the result tables that you want to exclude from being generated on the server and from being sent to the SAS session. The GLM statement can generate the following tables:

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Table Alias</th>
<th>Description</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ModelInfo</td>
<td></td>
<td>Information about the model—constant across groups or partitions.</td>
<td>This table is shown by default.</td>
</tr>
<tr>
<td>ClassLevels</td>
<td>Class</td>
<td>Information about the classification variables, such as the number of levels and their values.</td>
<td>This table is shown when classification variables are present in the model.</td>
</tr>
<tr>
<td>ConvStatus</td>
<td>Convergence</td>
<td>Convergence status of optimization</td>
<td>This table is shown by default.</td>
</tr>
<tr>
<td>Dimensions</td>
<td>Dim</td>
<td>Model dimensions</td>
<td>This table is shown by default.</td>
</tr>
<tr>
<td>FitStatistics</td>
<td>Fit</td>
<td>Fit statistics customary for regression models</td>
<td>This table is shown when it is requested with the SELECT= option.</td>
</tr>
<tr>
<td>GlobalTest</td>
<td>Global</td>
<td>Test of the hypothesis that the model fits as well as a null model without explanatory variables</td>
<td>This table is shown by default.</td>
</tr>
<tr>
<td>IterHistory</td>
<td>IterHist</td>
<td>Iteration history</td>
<td>This table is shown when the ITDETAILS option is used or when the table is requested with the SELECT= option.</td>
</tr>
<tr>
<td>ParmEstimates</td>
<td>ParameterEstimates Pest</td>
<td>The solutions for the linear model coefficients</td>
<td>This table is shown when there are no classification variables in the model.</td>
</tr>
<tr>
<td>ResponseProfile</td>
<td>Resp</td>
<td>Information about the values of the binary response variable such as the level order and frequency</td>
<td>This table is shown when modeling binary data. (When the events/trials syntax is not used.)</td>
</tr>
</tbody>
</table>
Table Name | Table Alias | Description | Condition
--- | --- | --- | ---
Tests3 | Type III tests of model effects | This table is shown when the effects contain classification variables and the NOSTDERR option is not specified.

Whether a table is shown by default or not, you can request any table with the SELECT= option in the LOGISTIC statement. The Condition column in the table identifies when a table is produced by default. For example, if the response variable is binary, the server generates the ResponseProfile table.

**FORMATS=("format specification", ...)**
specifies the formats for the GROUPBY variables. If you do not specify the FORMATS= option, or if you omit the entry for a GROUPBY variable, the default format is applied for that variable.

Enclose each format specification in quotation marks and separate each format specification with a comma.

**Example**
```
proc imstat data=lasr1.table1;
  statement / groupby=(a b) formats=('8.3', '$10');
quit;
```

**FCONV=**
specifies a relative function convergence criterion. For all techniques except NMSIMP, termination requires a small relative change of the function value in successive iterations. Suppose that Ψ is the p × 1 vector of parameter estimates in the optimization and the objective function at the kth iteration is denoted, f(Ψ). Then, the FCONV criterion is met if
\[
\frac{f(Ψ(k)) - f(Ψ(k-1))}{f(Ψ(k-1))} \leq r
\]

**Default**
r = 10^-FDIGITS where FDIGITS is -log_{10}(e) and e is the machine precision.

**FREQ=**
specifies the numeric variable that provides frequencies for the analysis. For example, if the FREQ= variable has the value 5, then it implies that the record represents five such observations with identical values for the modeling variables. If you specify a FREQ= variable, then only the observations with a value that is not missing and greater than zero for the variable are used in the analysis.

**GCONV=**
specifies a relative gradient convergence criterion. For all optimization techniques except CONGRA and NMSIMP, termination requires that the normalized predicted function reduction is small. The default value is r = 1e-8. Suppose that Ψ is the p × 1 vector of parameter estimates in the optimization with ith element Ψ_i. The objective function, its p × 1 gradient vector, and its p × p Hessian matrix are denoted, f(Ψ), g(Ψ), and H(Ψ), respectively. Then, if superscripts denote the iteration count, the normalized predicted function reduction at iteration k is
The GCONV convergence criterion is assumed to be met if that value is less than or equal to \( r \).

Note that it is possible that the relative gradient reduction is small, even if one or more gradients is still substantial in absolute value. If this situation occurs, you can disable the GCONV criterion by setting \( r=0 \). If the optimization would have stopped early due to meeting the GCONV criterion, the iterative process usually takes one more step until the gradients are small in absolute value.

GROUPBY=(variable-list)
specifies a list of variable names, or a single variable name, to use as GROUPBY variables in the order of the grouping hierarchy. If you do not specify any GROUPBY variable names, then the calculation is performed across the entire table—possibly subject to a WHERE clause.

GROUPBY=(variable-list)
specifies the names of the Group-by variables in the order of the grouping hierarchy. If no variable names are specified, the model is fit across the entire table—possibly subject to a WHERE clause.

If you work on a partitioned table, you can also use the PARTITION option to fit the model for a specific partition or separately for all partitions. Operations on partitions are much more efficient than a group-by analysis.

Because fitting logistic models requires an iterative method, the Group-by analysis for these models is a data-parallel technique where the model in each group is fit separately, assigning different rows of the group to different threads.

GROUPFILTER=(filter-options)
specifies a section of the group-by hierarchy to be included in the computation. With this option, you can request that the server performs the analysis for only a subset of all possible groupings. The subset is determined by applying the group filter to a temporary table that you generate with the GROUPBY statement.

You can specify the following suboptions in the GROUPFILTER option:

DESCENDING
specifies the top or the bottom section of the groupings to be collected. If the DESCENDING option is specified, the top \( \text{LIMIT}=n \) (where \( n > 0 \)) groupings are collected. Otherwise, the bottom \( \text{LIMIT}=n \) groupings are collected.

Alias DESC

LIMIT=n
specifies the maximum number of distinct groupings to be collected, where integer \( n \geq 0 \). If \( n \) is zero, then all distinct groupings (up to \( 2^{31}-1 \)) that satisfy the boundary constraints, such as LOWERSCORE=\( f \), are collected.

CAUTION High Cardinality Data Sets Setting \( n \) to zero with high-cardinality data sets can significantly delay the response of the server.

SCOREGT=\( f \)
specifies the exclusive lower bound for the numeric scores of the distinct groupings to collect.

Alias SGT=
SCORELT=f
specifies the exclusive upper bound for the numeric scores of the distinct groupings to collect.

Alias  SLT=

VALUEGT="format-name1" <, "format-name2" ...>
specifies the exclusive lower bound of the group-by variable’s formatted values for the distinct groupings to collect.

Alias  VGT=

VALUELT="format-name1" <, "format-name2" ...>
specifies the exclusive upper bound of the group-by variable’s formatted values for the distinct groupings to collect.

Alias  VLT=

TABLE=table-with-groupby-results
specifies the in-memory table from which to load the group-by hierarchy. If the TABLE= option is not specified, then all other GROUPFILTER= options are ignored.

The following program request all the groupings of State, City, and then Trade_In_Model in the Cars_Program_All table. The groupings are ordered by the maximum value of New_Vehicle_Msrp for each grouping:

proc imstat;
  table example.CARS_PROGRAM_ALL;
  groupby state city trade_in_model / TEMPTABLE
     WEIGHT=new_vehicle_MSRP
     AGG=(MAX)
     ORDER=WEIGHT;
run;

The TEMPTABLE option in the GROUPBY statement directs the server to save all the groupings in a temporary in-memory table. The following DISTINCT statement requests the count of the distinct unformatted values of Sales_Type for each of the selected groupings of State, City, and Trade_In_Model.

    table example.CARS_PROGRAM_ALL;
    distinct sales_type / GROUPFILTER=(
             table =mylasr.&_TEMPLAST_
             scoregt=40000
             value lt="$FL","Ft Myers","
             limit =20
             descending);
run;

This example only considers groupings that have maximum values of the New_Vehicle_Msrp above 40,000 and with formatted values that are less than State="FL" and City="Ft Myers." The empty quotation marks result in no restriction on Trade_In_Model values. These groupings are ordered according to the maximum values of New_Vehicle_Msrp. Because of the DESCENDING option, this example collects the 20 top groupings within the specified group-by range for the DISTINCT analysis.

Interaction  If you specify the GROUPFILTER= option, then the GROUPBY= and FORMATS= options have no effect.
IDVARS=(variable-list)
IDVARS=variable-name

specifies the variables from the active table to transfer to the temporary table that is created by scoring the input table. This option has no effect unless the SCORE option is also specified. (See the SCORE option for details about which variables are added to the temporary table by default.) The IDVARS= option should be used to transfer additional columns from the input table to the scoring table.

Alias ID=

Tip Instead of this option, you can specify the ALLIDVARS option to transfer all variables from the input table to the scoring table.

ITDETAILS
requests to add details about the iterative model fitting process (an iteration history) to the ODS output tables.

Alias ITDETAIL

KEYORDER
requests that the results for a partitioned analysis are displayed in the order of the partition keys. If this option is not specified, then results are displayed by using the partitions on the first worker node followed by the partitions on the second node, and so on. Without this option, the results are likely to have random ordering of the partitions. The KEYORDER option makes result collection less efficient but produces a natural, predictable order.

LINK=function
specifies the link function to use for the model fitting process. See the following list for the available functions:

• LOGIT
• PROBIT
• LOGLOG
• CLOGLOG

Default LOGIT

MAXFUNC=n
specifies the maximum number \( n \) of function calls in the iterative model fitting process. The default value depends on the optimization technique as follows:

<table>
<thead>
<tr>
<th>Optimization Technique</th>
<th>Default Number of Function Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUREG, NRRIDG, and NEWRAP</td>
<td>125</td>
</tr>
<tr>
<td>QUANEW and DBLDOG</td>
<td>500</td>
</tr>
<tr>
<td>CONGRA</td>
<td>1000</td>
</tr>
<tr>
<td>NMSIMP</td>
<td>3000</td>
</tr>
</tbody>
</table>

Alias MAXFU=
MAXITER=i
specifies the maximum number of iterations in the iterative model fitting process. The default value depends on the optimization technique as follows:

<table>
<thead>
<tr>
<th>Optimization Technique</th>
<th>Default Number of Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUREG, NRRIDG, and NEWRAP</td>
<td>50</td>
</tr>
<tr>
<td>QUANEW and DBLDOG</td>
<td>200</td>
</tr>
<tr>
<td>CONGRA</td>
<td>400</td>
</tr>
<tr>
<td>NMSIMP</td>
<td>1000</td>
</tr>
</tbody>
</table>

Alias MAXIT=

MAXTESTLEV=n
specifies the maximum number of levels in an effect for which the server generates Type III tests. The idea behind the MAXTESTLEV= option is that testing effects for significance that have a large number of levels is typically not meaningful. The effects tend to be highly significant anyway, but determining the exact significance level is computationally intensive. The default value is 300 and implies that no test statistics are produced for any effect that has more than 300 levels.

Default 300

NAME=SAS-name
specifies the name to use for identifying the model in the server output and in the temporary table of results generated by the TEMPTABLE option. SAS name rules apply. For example, the following statements add the 'Model' entry to the ModelInformation table.

```sas
proc imstat;
   table hps.neuralgia;
   logistic pain = treatment sex duration / name = LogisModel
run;
```

<table>
<thead>
<tr>
<th>Model Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Data Source</td>
</tr>
<tr>
<td>Response Variable</td>
</tr>
<tr>
<td>Distribution</td>
</tr>
<tr>
<td>Link Function</td>
</tr>
</tbody>
</table>

NOCLPRINT <=n>
specifies the number of levels for each classification variables to show in the Class Level Information ODS table. If you do not specify the NOCLPRINT option, all unique values are shown in the order of the class variable levelization. If you specify NOCLPRINT=n, then the values are shown for those classification variables that have less than n levels only. The value for n must be at least 1.
If you specify the NOCLPRINT option without specifying a value for \( n \), then \( n = 0 \) is assumed. This enables you to get a listing of the classification variables in the model. This might be useful if you did not identify classification variables explicitly—without listing their (possibly many) levels.

For example, the following Class Level Information table is displayed with NOCLPRINT=4. Because the number of levels for variable Smoking_Status exceeds 4, the values are not displayed.

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight_Status</td>
<td>3</td>
<td>Normal Overweight Underweight</td>
</tr>
<tr>
<td>Smoking_Status</td>
<td>5</td>
<td>not printed</td>
</tr>
</tbody>
</table>

**NOINT**
suppresses the inclusion of an intercept in the model. By default, all models contain an intercept term.

**NOPREPARSE**
prevents the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects problems with the code, the error messages might not be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

When you specify this option in the PROC IMSTAT statement, the option applies to all statements that can generate code. You can also exclude specific statements from pre-parsing by using the NOPREPARSE option in statements that allow temporary columns or the SCORE statement.

Alias **NOPREP**

**NOSTDERR**
prevents the computation of the covariance matrix and the standard errors of the parameter estimates. When you specify this option, the Type III tests for the model effects are also not available.

Alias **NOSTD**

**OFFSET=variable-name**
specifies the offset variable for the analysis. An offset variable can be thought of as a regressor variable whose regression coefficient is known to be 1. Offsets are used to shift the linear predictors by a certain amount. For example, an offset can be used to accommodate constants in the underlying model. For example, a model for the probability of being seropositive is as follows:

\[
\pi = 1 - \exp (-\beta X)
\]

After applying the log function, the model on the linear scale is as follows:

\[
\log (1 - \pi) = \log (\beta) + \log (X)
\]
You can model this relationship with a complementary log-log link (LINK=CLOGLOG), and the offset variable log(X). The term log(β) is then estimated by the intercept of the model.

**PARTITION <=partition-key>**

When you specify this option and the table is partitioned, the results are calculated separately for each value of the partition key. In other words, the partition variables function as automatic GROUPBY variables. This mode of executing calculations by partition is more efficient than using the GROUPBY= option. With a partitioned table, the server takes advantage of knowing that observations for a partition cannot be located on more than one worker node.

If you do not specify a partition-key, the analysis is performed for all partitions. If you do specify a partition-key, the analysis is carried out for the specified key value only. You can use the PARTITIONINFO statement to retrieve the valid partition key values for a table.

You can specify a partition-key in two ways. You can supply a single quoted string that is passed to the server, or you can specify the elements of a composite key separated by commas. For example, if you partition a table by variables GENDER and AGE, with formats $1 and BEST12, respectively, then the composite partition key has a length of 13. You can specify the partition for the 11-year-old females as follows:

```plaintext
statement / partition="F          11"; /* passed directly to the server */
statement / partition="F","11";        /* composed by the procedure */
```

If you choose the second format, the procedure composes a key based on formatting information from the server.

**ROLEVAR=variable-name**

specifies a variable in the in-memory table that defines whether an observation belongs to the training set, the validation set, or is to be excluded from the analysis. The role variable can have a numeric or character type, and it can be a temporary computed variable.

In case of a numeric role variable, the values of variable-name are interpreted as follows:

- value = 1: this observation is in the training set
- value = 2: this observation is in the validation set
- any other value: this observation is to be excluded from the analysis

In case of a character role variable, the values of variable-name are interpreted as follows:

- If the first non-blank character is 't' or 'T', then the observation is in the training set.
- If the first non-blank character is 'v' or 'V', then the observation is in the validation set.
- Any other value for the first non-blank character, including an all-blank entry, leads to the exclusion of the observation from the analysis.
Interactions

You can divide the data at random into training and validation sets by providing the VALIDATE= and SEED= options.

If you specify both the ROLEVAR= and the VALIDATE= options, then the ROLEVAR= setting supersedes the VALIDATE= option.

**SCORE <(score-statistic1score-statistic2…)>

requests that the active table be scored after the model is fit and the results be stored in a temporary table. The server automatically adds all model variables to the temporary table with the score results. These results include the response variable, the class variables, all explanatory variables from which effects are formed, and the WEIGHT=, and FREQ= variables.

In addition, if the active table is partitioned or ordered, the partition variables and order-by variables are transferred from the input table to the temporary table. The temporary table is partitioned and ordered in the same way as the active table.

If the analysis uses the GROUPBY= option, the variables in the group-by list are also transferred to the scoring table. If you want to transfer additional variables, you can specify them with the IDVARS= option.

If you do not specify the list of score statistics, default statistics are computed. These statistics are identified with Yes in the Default column in the table below. You can request that the following statistics be computed for each observation:

<table>
<thead>
<tr>
<th>Keyword and Aliases</th>
<th>Column Name</th>
<th>Description</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRED, PREDICTED, LINP</td>
<td><em>PRED</em></td>
<td>Predicted linear predictor value</td>
<td>Yes</td>
</tr>
<tr>
<td>RESID, RESIDUAL, R</td>
<td><em>RESID</em></td>
<td>Raw residual (on a linear scale)</td>
<td>Yes</td>
</tr>
<tr>
<td>LEVERAGE, H</td>
<td><em>LEVERAGE</em></td>
<td>Measure of how extreme an observation is in the regressor space</td>
<td>Yes</td>
</tr>
<tr>
<td>ILINK, MEAN, PROB</td>
<td><em>ILINK</em></td>
<td>Inversely linked linear predictor, the predicted mean of the response</td>
<td>Yes</td>
</tr>
<tr>
<td>PEARSON, RESCHI</td>
<td><em>PEARSON</em></td>
<td>Pearson residual, also known as the Chi-square residual</td>
<td>Yes</td>
</tr>
<tr>
<td>DEVRESID, RESDEV</td>
<td><em>DEVRESID</em></td>
<td>Deviance residual</td>
<td>Yes</td>
</tr>
<tr>
<td>LIKEDIST, LD, RESLIKE</td>
<td><em>LIKEDIST</em></td>
<td>Likelihood displacement</td>
<td>Yes</td>
</tr>
<tr>
<td>STDRES, STDRESCHI</td>
<td><em>STDRESCHI</em></td>
<td>Standardized Pearson Chi-square residual</td>
<td>Yes</td>
</tr>
</tbody>
</table>
The calculation of these statistics agrees with the LOGISTIC procedure. You can consult the documentation for the procedure in the SAS/STAT User’s Guide for details about the calculations. The output statistics of the LOGISTIC procedure referred to as PREDICTED, LOWER, and UPPER are equivalent to the _ILINK_, _LCLM_, and _UCLM_ statistics.

If you specify SCORE(_ALL_), then the server calculates and adds all the possible output statistics to the temporary table. The confidence levels for the LCLM, LCL, UCLM, and UCL confidence bounds are determined from the significance level specified in the ALPHA= option.

The interpretation of the LCL/UCL and LCLM/UCLM bounds is slightly different in the LOGISTIC statement as compared to the GLM statement. In a GLM model, the distinction between LCLM and LCL bounds is that the former apply to predictions of the mean (expected value) of an observation, whereas the latter apply to prediction of a new observation that has not been used in modeling the data. In the LOGISTIC statement all confidence bounds are bounds for predicting an expected value. The difference between LCLM/UCLM and LCL/UCL bounds here relates to whether the bound applies to the mean scale (the data scale), or the scale of the linear predictor. The LCLM and UCLM bounds apply to the mean (=data) scale—these are confidence bounds for the predicted probability. The LCL/UCL bounds apply to the linear scale—these are confidence bounds for the linear predictor η.
SELECT=(list-of-ODS-tables)
specifies the list of ODS tables that you want to display for the analysis. The
specified list replaces the default tables that are generated by the server and
displayed. See the EXCLUDE= option for the list of default tables and the table
names that you can display.

SHOWSELECTED
requests that the server perform variable selection for the model. A backward
selection method is used, where the significance level for an effect to remain in the
model is determined by the SLSTAY= option. This option performs variable
selection like the VARSEL option, but in contrast to the latter option, it displays
output only for the selected effects.

Alias SHOWSEL

SLSTAY=\alpha
specifies the significance level used in determining whether effects should stay in the
model during variable selection.

Default 0.1
Range 0 to 1

TECHNIQUE=
specifies the optimization technique.
Valid values are as follows:

CONGRA (CG) performs a conjugate-gradient optimization.
DBLDOG (DD) performs a version of the double-dogleg optimization.
DUQUANEW (DQN) performs a (dual) quasi-Newton optimization.
NMSIMP (NS) performs a Nelder-Mead simplex optimization.
NONE specifies not to perform any optimization. This value
can be used to perform a grid search without optimization.
NEWRAP (NRA) performs a (modified) Newton-Raphson optimization
that combines a line-search algorithm with ridging.
NRRIDG (NRR) performs a (modified) Newton-Raphson optimization
with ridging.
QUANEW (QN) performs a quasi-Newton optimization. If you specify
this technique, but specify bounds for any parameter,
the server automatically performs DUQUANEW.
TRUREG (TR) performs a trust-region optimization.

The factors that go into choosing a particular optimization technique for a particular
problem are complex. Trial and error can be involved. For many optimization
problems, computing the gradient takes more computer time than computing the
function value. Computing the Hessian sometimes takes much more computer time
and memory than computing the gradient, especially when there are many
parameters. Unfortunately, first-order optimization techniques that do not use some
type of Hessian or Hessian approximation usually require more iterations than
second-order techniques that use a Hessian matrix. As a result, the total run time of
first-order techniques can be longer. Techniques that do not use the Hessian also tend
to be less reliable. For example, they can terminate more easily at stationary points than at global optima.

The TRUREG, NEWRAP, and NRRIDG algorithms are second-order algorithms. The server computes first and second derivatives of the objective function with respect to the parameters in analytic form wherever possible. Finite-difference approximations for derivatives are used only when the derivatives of functions are not known. In most cases, finite-difference approximations are not necessary.

For more information about the algorithms, see *SAS/STAT User's Guide*.

**Alias**

**TECH=**

**Default**

**NRRIDG**

**TEMPRESS=**"SAS-expressions"

**TEMPRESS=**file-reference

specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

**Alias**

**TE=**

**TEMPNAMES=**variable-name

**TEMPNAMES=(**variable-list**)

specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPPRESS= option.

**Alias**

**TN=**

**TEMPTABLE**

generates an in-memory temporary table from the result set. The IMSTAT procedure displays the name of the table and stores it in the &_TEMPLAST_ macro variable, provided that the statement executed successfully.

When the IMSTAT procedure exits, all temporary tables created during the IMSTAT session are removed. Temporary tables are not displayed on a TABLEINFO request, unless the temporary table is the active table for the request.

**VALIDATE=**f

specifies the proportion f in the validation data set.

**Alias**

**VALPROP=**

**Range**

0 to 1

**Interaction**

If you specify both the ROLEVAR= and the VALIDATE= options, then the ROLEVAR= setting supersedes the VALIDATE= option.

**VARSELECTION**

requests that the server perform variable selection for the model. A backward selection method is used, where the significance level for an effect to remain in the model is determined by the SLSTAY= option. In contrast to the SHOWSEL option, all effects are reported in the IMSTAT output.

**Alias**

**VARSEL**
WEIGHT=variable-name
specifies the numeric variable to use as a weighing variable in solving the linear model.

MDSUMMARY Statement
The MDSUMMARY statement calculates a multi-dimensional summary for numeric variables.

Syntax
MDSUMMARY variable-list < / <set-specification,…> options> ;

Optional Arguments
variable-list
specifies one or more numeric variables. If you do not specify this option, then all numeric variables in the table are used.

set-specification,…
specifies the three elements for generating a set. Separate each set-specification with a comma.

GROUPBY=variable-name
GROUPBY=(variable-list)
specifies the list of GROUP BY variables for this set-specification. A GROUPBY= specification is required.

FORMATS=("format-specification",…)
specifies the formats for the GROUPBY= variables. If you do not specify the FORMAT= option, the default format is applied for that variable. Enclose each format specification in quotation marks and separate each format specification with a comma.

You can omit the assignment of a format for a GROUPBY= variable by entering an empty string. For example, FORMATS=("$10.", "", "BEST4.") specifies to format the first variable, with $10 and the third variable with BEST4. The default format is applied to the second variable. The FORMATS= element of the set-specification is optional.

FILTER="expression"
specifies an optional WHERE clause for this set-specification. The filter is applied separately for each set and possibly in addition to an overall WHERE clause.

MDSUMMARY Statement Options
DESCENDING
specifies that the levels of the GROUPBY variables are to be arranged in descending order.

GROUPBYLIMIT=n
specifies the maximum number of levels in a GROUPBY set. When the software determines that there are at least n levels in the GROUPBY set, it abandons the action, returns a message, and does not produce a result set. You can specify the
GROUPBYLIMIT= option if you want to avoid creating excessively large result sets in GROUPBY operations.

DESCENDING
specifies that the levels of the GROUPBY variables are to be arranged in descending order.

Alias DESC

NOPREPARSE
prevents the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects problems with the code, the error messages might not be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

When you specify this option in the PROC IMSTAT statement, the option applies to all statements that can generate code. You can also exclude specific statements from pre-parsing by using the NOPREPARSE option in statements that allow temporary columns or the SCORE statement.

Alias NOPREP

RAWORDER
specifies that the ordering of the GROUPBY variables is based on the raw values of the variables instead of the formatted values.

SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

SETSIZE
requests that the server estimate the size of the result set. The procedure does not create a result table if the SETSIZE option is specified. Instead, the procedure reports the number of rows that are returned by the request and the expected memory consumption for the result set (in KB). If you specify the SETSIZE option, the SAS log includes the number of observations and the estimated result set size. See the following log sample:

NOTE: The LASR Analytic Server action request for the STATEMENT statement would return 17 rows and approximately 3.641 kBytes of data.

The typical use of the SETSIZE option is to get an estimate of the size of the result set in situations where you are unsure whether the SAS session can handle a large result set. Be aware that in order to determine the size of the result set, the server has to perform the work as if you were receiving the actual result set. Requesting the estimated size of the result set does consume resources on the server. The estimated number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.
TEMPEXPRESS="SAS-expressions"
TEMPEXPRESS=file-reference
specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias   TE=

TEMPNAMES=variable-name
TEMPNAMES=(variable-list)
specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

Alias   TN=

TEMPTABLE
generates an in-memory temporary table from the result set. The IMSTAT procedure displays the name of the table and stores it in the &_TEMPLAST_ macro variable, provided that the statement executed successfully.

When the IMSTAT procedure exits, all temporary tables created during the IMSTAT session are removed. Temporary tables are not displayed on a TABLEINFO request, unless the temporary table is the active table for the request.

OPTIMIZE Statement
The OPTIMIZE statement performs a non-linear optimization of an objective function that is defined through a SAS program. The expression defined in the SAS program and its analytic first and second derivatives are compiled into executable code. The code is then executed in multiple threads against the data in an in-memory table. Like all other IMSTAT statements, the calculations are performed by the server. You can choose from several first-order and second-order optimization algorithms.

Syntax
OPTIMIZE <options>;

OPTIMIZE Statement Options
ALPHA=number
specifies a number between 0 and 1 from which to determine the confidence level for approximate confidence intervals of the parameter estimates. The default is $\alpha = 0.05$, which leads to $100 \times (1 - \alpha)\% = 95\%$ confidence limits for the parameter estimates.

Default 0.05

BOUNDS=(boundary-specification<, boundary-specification...,>)
specifies boundary values for the parameters. A boundary-specification is specified in the following form:

$parameter-name\ operator\ value$

$parameter-name$ specifies the parameter

operator is one of $\geq$, $\leq$, $>$, $<$, $\lt$, $=$, $\neq$. 

142  Chapter 4 • IMSTAT Procedure (Analytics)
value
  specifies the boundary value

Alias  BOUND=

Example  BOUNDS=(s2 > 0, beta2 >= 0.2)

CODE=file-reference
  specifies a file reference to the SAS program that defines the objective function. The program must make an assignment to the reserved symbol _OBJFNC_. The server then minimizes the negative of that function (or maximize the function). In other words, you should specify _OBJFNC_ to be the function that you want to maximize across the in-memory table. The actual optimization is carried out as a minimization problem.

Alias  PGM=

DEFSTART=value
  specifies the default starting value for parameters whose starting value has not been specified. The default value, 1, might not work well depending on the optimization.

Alias  DEFVAL=

Default  1

DUD
  specifies that you do not want to use analytic derivatives in the optimization. The option name is an acronym for "do not use derivatives." Instead, the server calculates gradient vectors and Hessian matrices from finite difference approximations. Generally, you should not rely on derivatives calculated from finite differences if analytic derivatives are available. However, this option is useful in situations where the objective function is not calculated independently for each row of data. If derivatives of the objective function depend on lagged values, which are themselves functions of the parameters, then finite difference derivatives are called for.

Alias  NODERIVATIVES

FCONV=r
  specifies a relative function convergence criterion. For all techniques except NMSIMP, termination requires a small relative change of the function value in successive iterations. Suppose that \( \Psi \) is the \( p \times 1 \) vector of parameter estimates in the optimization and the objective function at the \( k \)th iteration is denoted, \( f(\Psi)^k \). Then, the FCONV criterion is met if

\[
\frac{|f(\Psi^k) - f(\Psi^{k-1})|}{f(\Psi^{k-1})} \leq r
\]

Default  \( r=10^{-FDIGITS} \) where FDIGITS is \(-\log_{10}(e)\) and \( e \) is the machine precision.

GCONV=r
  specifies a relative gradient convergence criterion. For all optimization techniques except CONGRA and NMSIMP, termination requires that the normalized predicted function reduction is small. The default value is \( r = 10^{-8} \). Suppose that \( \Psi \) is the \( p \times 1 \) vector of parameter estimates in the optimization with \( i \)th element \( \Psi_i^k \). The objective function, its \( p \times 1 \) gradient vector, and its \( p \times p \) Hessian matrix are denoted, \( f(\Psi) \), \( g(\Psi) \), and \( H(\Psi) \), respectively. Then, if superscripts denote the iteration count, the normalized predicted function reduction at iteration \( k \) is
The GCONV convergence criterion is assumed to be met if that value is less than or equal to \( r \).

Note that it is possible that the relative gradient reduction is small, even if one or more gradients is still substantial in absolute value. If this situation occurs, you can disable the GCONV criterion by setting \( r = 0 \). If the optimization would have stopped early due to meeting the GCONV criterion, the iterative process usually takes one more step until the gradients are small in absolute value.

**ITDETAIL**

requests that the server produce an iteration history table for the optimization. This table displays the objective function, its absolute change, and the largest absolute gradient across the iterations.

**MAXFUNC=** \( n \)

specifies the maximum number \( n \) of function calls in the iterative model fitting process. The default value depends on the optimization technique as follows:

<table>
<thead>
<tr>
<th>Optimization Technique</th>
<th>Default Number of Function Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUREG, NRRIDG, and NEWRAP</td>
<td>125</td>
</tr>
<tr>
<td>QUANEW and DBLDOG</td>
<td>500</td>
</tr>
<tr>
<td>CONGRA</td>
<td>1000</td>
</tr>
<tr>
<td>NMSIMP</td>
<td>3000</td>
</tr>
</tbody>
</table>

Alias **MAXFU=**

**MAXITER=** \( i \)

specifies the maximum number of iterations in the iterative model fitting process. The default value depends on the optimization technique as follows:

<table>
<thead>
<tr>
<th>Optimization Technique</th>
<th>Default Number of Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUREG, NRRIDG, and NEWRAP</td>
<td>50</td>
</tr>
<tr>
<td>QUANEW and DBLDOG</td>
<td>200</td>
</tr>
<tr>
<td>CONGRA</td>
<td>400</td>
</tr>
<tr>
<td>NMSIMP</td>
<td>1000</td>
</tr>
</tbody>
</table>

Alias **MAXIT=**

**MAXTIME=** \( t \)

specifies an upper limit of \( t \) seconds of CPU time for the optimization process. The default value is the largest floating-point double representation value for the
hardware used by the SAS LASR Analytic Server. Note that the time specified by the MAXTIME= option is checked only once at the end of each iteration. The time is measured on the root node for the server. Therefore, the actual running time can be longer than the value specified by the MAXTIME= option.

**MINITER=i**

specifies the minimum number of iterations.

Alias     MINIT=

Default    0

**NBEST=k**

requests that only the $k$ best points in the starting value grid are reproduced in the "Starting Values" table. By default, the objective function is initially evaluated at all points in the starting value grid and the "Starting Values" table contains one row for each point on the grid. If you specify the NBEST= option, then only the $k$ points with the smallest objective function value are shown.

Alias     BEST=

**NOEMPTY**

requests that result sets for optimizations without usable data are not generated.

**NOPREPARSE**

specifies to prevent pre-parsing and pre-generating the program code that is referenced in the CODE= option. If you know the code is correct, you can specify this option to save resources. The code is always parsed by the server, but you might get more detailed error messages when the procedure parses the code rather than the server. The server assumes that the code is correct. If the code fails to compile, the server indicates that it could not parse the code, but not where the error occurred.

Alias     NOPREP

**NOSTDERR**

specifies to prevent calculating standard errors of the parameter estimates. The calculation of standard errors requires the derivation of the Hessian or cross-product Jacobian. If you do not want standard errors, $p$-values, or confidence intervals for the parameter estimates, then specifying this option saves computing resources.

Alias     NOSTD

**PARAMETERS=(parameter-specification <, parameter-specification...>)**

specifies the parameters in the optimization and the starting values. You do not have to specify parameters and you do not have to specify starting values. If you omit the starting values, the default starting value is assigned. This default value is 1.0 and can be modified with the DEFSTART= option.

If you do not specify the parameter names, the server assumes that all symbols in your SAS program are parameters if they do not match column names in the in-memory table or are not special or temporary symbols. This might not be what you want and you should examine the "Starting Values" and "Parameter Estimates" table in that case to make sure that the server designated the appropriate quantities as parameters in the optimization.

In the first example that follows, Intercept is assigned a starting value of 6. The remaining parameters start at 0 because the DEFSTART= option is 0.

In the second example that follows, the server evaluates the objective function initially for the Cartesian product set of all the parameter vectors. The server
evaluates $1 \times 3 \times 2 \times 1 = 6$ parameter vectors. The optimization then starts from the vector associated with the best objective function value.

**Alias**

<table>
<thead>
<tr>
<th>Examples</th>
<th>DEFSTART=0; PARMS=(Intercept = 6, a_0, b_0, c_0, x_1, x_2, x_3);</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PARMS=(beta1 = -3.22, beta2 = 0.5 0.47 0.6, beta3 = -2.45 -2.0, s2 = 0.5);</td>
</tr>
</tbody>
</table>

**RESTRICT=(one-restriction <, one-restriction>)**

specifies linear equality and inequality constraints for the optimization. A single restriction takes on the general form

```
coefficient parameter ... coefficient parameter operator value
```

Inequality restrictions are expressed as constraints greater than (>) or greater than or equal (>=) than the right hand side value.

The first example that follows shows the restriction $\beta_1 - 2 \beta_2 > 3$.

The second example that follows shows how to use more than one restriction. Restrictions are separated by commas and the second example requests that the estimates for parameters dose1 and dose2 are the same, as well as the estimates for logd1 and logd2.

**Examples**

<table>
<thead>
<tr>
<th>Examples</th>
<th>RESTRICT=(1 beta1 -2 beta2 &gt; 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RESTRICT=(1 dose1 -1 dose2 = 0, 1 logd1 -1 logd2 = 0)</td>
</tr>
</tbody>
</table>

**SAVE=table-name**

saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for `table-name` must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

**SETSIZE**

requests that the server estimate the size of the result set. The procedure does not create a result table if the SETSIZE option is specified. Instead, the procedure reports the number of rows that are returned by the request and the expected memory consumption for the result set (in KB). If you specify the SETSIZE option, the SAS log includes the number of observations and the estimated result set size. See the following log sample:

```
NOTE: The LASR Analytic Server action request for the STATEMENT statement would return 17 rows and approximately 3.641 kBytes of data.
```

The typical use of the SETSIZE option is to get an estimate of the size of the result set in situations where you are unsure whether the SAS session can handle a large result set. Be aware that in order to determine the size of the result set, the server has to perform the work as if you were receiving the actual result set. Requesting the estimated size of the result set does consume resources on the server. The estimated number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.

**TECHNIQUE=**

specifies the optimization technique.
Valid values are as follows:

- **CONGRA (CG)** performs a conjugate-gradient optimization.
- **DBLDOG (DD)** performs a version of the double-dogleg optimization.
- **DUQUANEW (DQN)** performs a (dual) quasi-Newton optimization.
- **NMSIMP (NS)** performs a Nelder-Mead simplex optimization.
- **NONE** specifies not to perform any optimization. This value can be used to perform a grid search without optimization.
- **NEWRAP (NRA)** performs a (modified) Newton-Raphson optimization that combines a line-search algorithm with ridging.
- **NRRIDG (NRR)** performs a (modified) Newton-Raphson optimization with ridging.
- **QUANEW (QN)** performs a quasi-Newton optimization.
- **TRUREG (TR)** performs a trust-region optimization.

The factors that go into choosing a particular optimization technique for a particular problem are complex. Trial and error can be involved. For many optimization problems, computing the gradient takes more computer time than computing the function value. Computing the Hessian sometimes takes much more computer time and memory than computing the gradient, especially when there are many parameters. Unfortunately, first-order optimization techniques that do not use some type of Hessian or Hessian approximation usually require more iterations than second-order techniques that use a Hessian matrix. As a result, the total run time of first-order techniques can be longer. Techniques that do not use the Hessian also tend to be less reliable. For example, they can terminate more easily at stationary points than at global optima.

The TRUREG, NEWRAP, and NRRIDG algorithms are second-order algorithms.

The server computes first and second derivatives of the objective function with respect to the parameters in analytic form wherever possible. Finite-difference approximations for derivatives are used only when the derivatives of functions are not known. In most cases, finite-difference approximations are not necessary.

For more information about the algorithms, see *SAS/STAT User's Guide*.

#### Alias
**TECH=**

#### Default
**DUQUANEW**

---

**PERCENTILE Statement**

The **PERCENTILE** statement computes percentiles for one or more numeric variables.

**Examples:**
- “Example 1: Calculating Percentiles and Quartiles” on page 172
- “Example 8: Storing Temporary Variables” on page 245

**Syntax**

```
PERCENTILE <variable-list> </options>;
```
**Optional Argument**

**variable-list**

specifies one or more numeric variables. If you do not specify this option, then all numeric variables in the table are used.

**PERCENTILE Statement Options**

**DESCENDING**

specifies that the levels of the GROUPBY variables are to be arranged in descending order.

Alias DESC

**EPSILON=number**

specifies the tolerance used for determining the convergence of the iterative algorithm for the percentile calculation.

Alias EPS=

Default 1e-5

**FORMATS=("format-specification",...)**

specifies the formats for the GROUPBY= variables. If you do not specify the FORMATS= option, or if you omit the entry for a GROUPBY variable, the default format is applied for that variable.

Enclose each format specification in quotation marks and separate each format specification with a comma.

Example

```plaintext
proc imstat data=lasr1.table1;
   percentile x / groupby=(a b) formats=("8.3", "$10");
quit;
```

**GROUPBY=(variable-list)**

specifies a list of variable names, or a single variable name, to use as GROUPBY variables in the order of the grouping hierarchy. If you do not specify any GROUPBY variable names, then the calculation is performed across the entire table — possibly subject to a WHERE clause.

**GROUPBYLIMIT=n**

specifies the maximum number of levels in a GROUPBY set. When the software determines that there are at least n levels in the GROUPBY set, it abandons the action, returns a message, and does not produce a result set. You can specify the GROUPBYLIMIT= option if you want to avoid creating excessively large result sets in GROUPBY operations.

**MAXITER=i**

specifies the maximum number of iterations for the algorithm. The percentile algorithm is iterative. You can limit the number of iterations with the MAXITER= option. You can also control the computational demand with the EPSILON= option. That option affects the tolerance criterion by which the convergence of the iterative algorithm is judged. Whether the percentile calculation has converged is displayed separately for each of the calculated percentiles.

Alias ITER=

Default 10
MERGEBINS=b
specifies the number of bins to create when a numeric GROUPBY variable exceeds the MERGELIMIT=n specification. If you specify a MERGELIMIT, but do not specify a value for the MERGEBINS= option, the server automatically calculates the number of bins.

MERGELIMIT=n
specifies that when the number of unique values in a numeric GROUPBY variable exceeds n, the variable is automatically binned and the GROUPBY structure is determined based on the binned values of the variable, rather than the unique formatted values.

For example, if you specify MERGELIMIT=500, any numeric GROUPBY variable with more than 500 unique formatted values is binned. Instead of returning results for more than 500 groups, the results are returned for the bins. You can specify the number of bins with the MERGEBINS= option.

DESCENDING
specifies that the levels of the GROUPBY variables are to be arranged in descending order.

Alias DESC

NOPREPARSE
prevents the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects problems with the code, the error messages might not be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

When you specify this option in the PROC IMSTAT statement, the option applies to all statements that can generate code. You can also exclude specific statements from pre-parsing by using the NOPREPARSE option in statements that allow temporary columns or the SCORE statement.

Alias NOPREP

NOTEMPPART
specifies that the temporary table generated by the TEMPTABLE option is not partitioned by the GROUPBY= variables. When you request a temporary table with the PERCENTILE statement, by default, the server partitions the table and the size of a partition is equal to the number of analysis variables in the variable-list of the PERCENTILE statement. When the number of groups is large, this can result in many small partitions, and requires extra memory resources to store the partition information for the temporary table. By specifying this option, the temporary table is organized similarly to the default table, but is not a partitioned table.

Alias NOTP

PARTITION <=partition-key>
When you specify this option and the table is partitioned, the results are calculated separately for each value of the partition key. In other words, the partition variables function as automatic GROUPBY variables. This mode of executing calculations by
partition is more efficient than using the GROUPBY= option. With a partitioned table, the server takes advantage of knowing that observations for a partition cannot be located on more than one worker node.

If you do not specify a partition-key, the analysis is performed for all partitions. If you do specify a partition-key, the analysis is carried out for the specified key value only. You can use the PARTITIONINFO statement to retrieve the valid partition key values for a table.

You can specify a partition-key in two ways. You can supply a single quoted string that is passed to the server, or you can specify the elements of a composite key separated by commas. For example, if you partition a table by variables GENDER and AGE, with formats $1 and BEST12, respectively, then the composite partition key has a length of 13. You can specify the partition for the 11-year-old females as follows:

```
statement / partition="F          11"; /* passed directly to the server */
statement / partition="F","11";        /* composed by the procedure */
```

If you choose the second format, the procedure composes a key based on formatting information from the server.

**Alias PART=**

**RAWORDER**

specifies that the ordering of the GROUPBY variables is based on the raw values of the variables instead of the formatted values.

**SAVE=table-name**

saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

**SETSIZE**

requests that the server estimate the size of the result set. The procedure does not create a result table if the SETSIZE option is specified. Instead, the procedure reports the number of rows that are returned by the request and the expected memory consumption for the result set (in KB). If you specify the SETSIZE option, the SAS log includes the number of observations and the estimated result set size. See the following log sample:

```
NOTE: The LASR Analytic Server action request for the STATEMENT
statement would return 17 rows and approximately
3.641 kBytes of data.
```

The typical use of the SETSIZE option is to get an estimate of the size of the result set in situations where you are unsure whether the SAS session can handle a large result set. Be aware that in order to determine the size of the result set, the server has to perform the work as if you were receiving the actual result set. Requesting the estimated size of the result set does consume resources on the server. The estimated number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.

**TEMPEXPRESS=\"SAS-expressions\"**

**TEMPEXPRESS=file-reference**

specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.
TEMPNAMES specifications the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

TEMPTABLE generates an in-memory temporary table from the result set. The IMSTAT procedure displays the name of the table and stores it in the &_TEMPLAST_ macro variable, provided that the statement executed successfully.

When the IMSTAT procedure exits, all temporary tables created during the IMSTAT session are removed. Temporary tables are not displayed on a TABLEINFO request, unless the temporary table is the active table for the request.

VALUES= specifies the values for which to calculate the percentiles. The default is to calculate the 25th, 50th, and 75th percentile. These are also known as the first, second, and third quartile. The second quartile is the median.

Example

The following statement requests the 10th, 20th, ..., 90th percentile:

```
percentile invoice / values=(10, 20, 30, 40, 50, 60, 70, 80, 90);
```

RANDOMWOODS Statement

The RANDOMWOODS statement builds a random forest of decision trees. Each tree is constructed from a bootstrap sample of the data, drawn with replacement, and is constructed from only a subset of the variables specified in the INPUT= option.

Syntax

```
RANDOMWOODS target-variable </options>
```

Required Argument

```
target-variable
```

specifies a single column in the in-memory table as the target variable. The variable can be a temporary calculated column.

RANDOMWOODS Statement Options

```
ASSESS
```

specifies that predicted probabilities are added to the temporary result table for the event levels. You can use these predicted probabilities in an ASSESS statement.
ADDTTREES
requests that the temporary table that is generated by scoring a random forest is
enhanced with information about the votes of the individual trees. The process of
scoring a random forest means that each tree votes on the predicted value and the
predicted value for the forest is obtained by majority vote. This option adds the votes
for each tree in the forest. By default, only the overall vote is reported in the
temporary table.

BOOTSTRAP=f
specifies the fraction of the data in the bootstrap sample.

Default \( f = 1 - \exp(-1) \)

Range 0 to 1

CODE <(code-generation-options)>
requests that the server produce SAS scoring code for the random forest. The server
generates DATA step code. The server does not generate DS2 code. By default, the
code is replayed as an ODS table by the procedure as part of the output of the
statement. More frequently, you might want to write the scoring code to an external
file by specifying options.

The scoring code computes the predicted value of the response variable on the data
scale (the inverse link scale) and prefixes the name with "P_". For example, if the
response variable is \( Y \), the generated code stores the predicted value as \( P_Y \). The
name of the variable is truncated to fit within the SAS name length requirements.

COMMENT
specifies to add comments to the code in addition to the header block. The header
block is added by default.

FILENAME='path'
specifies the name of the external file to which the scoring code is written. This
suboption applies only to the scoring code itself. If you request that the server
generate IMSTAT programming statements with the IMSTAT suboption, then
these statements are saved as an ODS table.

Alias FILE=

FORMATWIDTH=k
specifies the width to use in formatting derived numbers such as parameter
estimates in the scoring code. The server applies the BEST format, and the
default format for code generation is BEST20.

Alias FMTWIDTH=

Range 4 to 32

LINESIZE=n
specifies the line size for the generated code.

Alias LS=

Default 72

Range 64 to 256
NOTRIM
requests that the comparison of the formatted values for class variables and
group-by variables is based on the full format width with padding. By default, the
leading and trailing blanks are removed from the formatted values.

REPLACE
specifies to overwrite the external file with the new contents if the file already
exists. This option has no effect unless you specify the FILENAME= option.

EVENT="event1" <, "event2" >...
specifies the event names of the target variable. This option is combined with the
WEIGHT= option to specify the weight for each specific event. Observations with
the specified event are reweighted with the value from the WEIGHT= option. This
option is useful for rare-event sampling.

FORMATS="format-specification",...
specifies the formats for the input variables. If you do not specify the FORMATS=
option, the default format is applied for that variable. Enclose each format
specification in quotation marks and separate each format specification with a
comma.

GAIN
specifies that the splitting criterion is changed to information gain. Typically, this
criterion intends to generate trees with more nodes than information gain ratio.

GREEDY
specifies how to perform splitting under specific circumstances.
Assuming that one variable has $q$ levels, when binary splitting is performed and $q$ is
less than 15, or option MAXBRANCH > 2 and $q < 12$, all possible binary splits are
enumerated and the split with the largest gain or gain ratio is chosen for the variable.
When $q$ is less than 1024 and splitting is not just binary, local greedy searches are
applied to determine the optimum local split. Specifically, when the variable is
numeric, $q$ levels (similar to $q$ bins) are sorted by value.
When the variable is nominal, the $q$ levels are ordered by random weights. The best
binary splitting is applied until the desired number of branches is reached. Only a
local optimum can be found with this technique.
For values of $q \geq 1024$, the default $k$-means clustering algorithm is applied to
determine the splits.

IMPUTE
specifies how to treat observations with nonmissing values for the target variable
during scoring. When this option is specified, the observed values are used as the
predicted values. That is, the observed value is assumed to be known without error.
Only the observations with missing values for the target variable are then scored
against the random forest, based on their values for the input variables.
This option is useful if you want to replace missing values of a target variable with
classified values that are based on the random forest.

INPUT=variable-name
INPUT=(variable-list)
specifies the variables to use for building the tree. You can add the target variable to
the input list if you want to assign a format to the target variable by using the
FORMATS= option. Any numeric variable that is not specified in the NOMINAL=
option is binned according to the NBINS= specification.
In random forest implementations, all of the input variables do not participate in the construction of the trees. Each tree is built from a subset of the input variables. You can use the M= option to affect the selection of these input variables.

**LEAFSIZE=m**

specifies the minimal number of observations on each node. When the number of observations on a tree node falls short of the specified leaf size m, the node is changed into a leaf during the building of the tree.

Interaction Specifying the LEAFSIZE option affects the pruning of the tree.

**M=k**

specifies the number of input variables used to build a tree. The k variables are selected at random from the list of input variables for each tree. If not specified, then k defaults to the square root of the number of input variables, rounded up to the nearest integer.

**MAXBRANCH=n**

specifies the maximum number of children (branches) allowed for each level of the tree.

**MAXLEVEL=n**

specifies the maximum number of tree levels.

**NBINS=k**

specifies the number of bins used in the calculation of the tree. The number of bins affects the accuracy of the tree and increases with k at the expense of computing time and memory consumption.

**NBINSTARGET=k**

specifies the number of bins to use for a numeric target variable. The number of bins affects the accuracy of the tree. The accuracy increases as values of k increase. However, computing time and memory consumption also increase as values of k increase. When k is greater than zero, the numeric target variable is binned into equally sized bins first and then the bins are used to perform the classification.

**NOERROR**

specifies that the out-of-bag error is not computed when building a random decision forest. This option is useful to speed up the building process.

**NOMINAL=variable-name**

**NOMINAL=(variable-list)**

specifies the numeric variables to use as nominal variables. Binning is not applied to the specified variables. The target variable is always treated as a nominal variable and does not need to be listed.

**NOMISSOBS**

specifies to ignore observations that have missing values in the analysis variables when building a decision tree. When scoring a data set, any observations with missing values in the analysis variables for the decision tree are ignored when this option is specified.
When this option is not specified, the RANDOMWOODS statement builds a tree by applying the following policy for missing values:

- For an interval variable, the smallest machine value is assigned.
- For a nominal variable, missing values are represented by a separate level.

NOPREPARE

prevents pre-parsing and pre-generating the program code that is referenced in the CODE= option. If you know the code is correct, you can specify this option to save resources. The code is always parsed by the server, but you might get more detailed error messages when the procedure parses the code rather than the server. The server assumes that the code is correct. If the code fails to compile, the server indicates that it could not parse the code, but not where the error occurred.

Alias NOPREP

NTREE=n

specifies the number of trees to build for the random forest.

Default 1

REG

specifies to build the random decision forest using regression trees. Minimal cost-complexity pruning is applied to prune the trees.

SAVE=table-name

saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

SCORERDATA=table-name

specifies the in-memory table that contains the scoring data. The table must exist in-memory on the server. The RANDOMWOODS statement in the IMSTAT procedure does not transfer a local data set to the server.

If you do not specify a table name for this option, the active table is used as the scoring input.

SEED=s

specifies the random number seed for the random number generator in the server. The default value, zero, implies that the random number stream is based on the computer clock. Negative seed values also lead to random number streams that are based on the computer clock. If you want a reproducible random number sequence between runs, specify a value that is greater than zero.

Default 0

TEMPEXPRESS="SAS-expressions"

specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias TE=
TEMPNAMES=variable-name
TEMPNAMES=(variable-list)

specifies the list of temporary variables for the request. Each temporary variable
must be defined through SAS statements that you supply with the TEMPEXPRESS=
option.

Alias  TN=

TEMPTABLE

specifies to save the results of the RANDOMWOODS statement in a temporary
table.

If you build a random forest, the temporary table contains information about the
forest.

If you score a random forest, the temporary table contains the predicted values for
the observations in the input table that you scored. The temporary table also contains
the variables that you specified to transfer from the input table and other statistics.
This option is required when you perform scoring so that you can access the
predictions for each observation. The IMSTAT procedure then displays the name of
the table and stores it in the _TEMPSCORE_ macro variable, provided that the
scoring action was successful. Observations from the table that you scored can be
transferred to the temporary table using the VARS= option.

TIMEOUT=s

specifies the maximum number of seconds that the statement should run in the
server. If the computation does not complete before the time-out is reached,
execution stops and the server generates an error message. There is no default time-
out.

TREEINFO

specifies to display information about individual trees, when you build a tree. For
example, the table that is shown can display which variables are used in each tree.
The option has no effect if you store the tree in a temporary table.

TREELASR=table-name

specifies the in-memory table that contains the information for the random forest if
you want to score a table against the random forest.

The data set whose observations are to be scored is specified in the SCOREDATA=
option. If you do not specify the SCOREDATA= option, the active table is used as
scoring input.

Alias  LASRTREE=

VARS=variable-name
VARS=(variable-name1 <, variable-name2, ...>)

specifies the variables to transfer from the input table to the temporary table in the
server that contains the results of scoring a decision tree. This option has no effect
unless you specify the TEMPTABLE option and you score a decision tree.

WEIGHT=

specifies the weight for each corresponding event in the EVENT= option.

REGCORR Statement

The REGCORR statement calculates and returns the results for linear, quadratic, or cubic polynomial
regression models.
**Example:**

“Example 10: Fitting a Regression Model” on page 185

---

**Syntax**

```
REGCORR <variable-list> </ options> ;
```

**Optional Argument**

`variable-list`

specifies one or more numeric variables. If you do not specify this option, then all numeric variables in the table are used.

**REGCORR Statement Options**

**NBEST=n**

specifies that results are returned only for the `n` regression with the highest R-square value (the highest coefficient of determination). If `n` is smaller than the number of regressions computed by the statement, then the actual number of computed regression is returned.

**ORDER=1 | 2 | 3**

**ORDER=-1 | –2 | –3**

specifies the highest polynomial degree in the regression model. By default, ORDER=1, and the model is a simple linear regression. Specify ORDER=2 for a quadratic model and ORDER=3 for a cubic model.

If you specify a negative value for the ORDER= option, the server evaluates the best model for each variable combination based on statistical principles. For example, if you specify ORDER=–2, the server returns results for a linear regression provided that the removal of the quadratic term does not result in a poorer model—as judged statistically. Similarly, with ORDER=–3, you might get results for a cubic, quadratic, or a linear regression. The results depend on which model is deemed to fit best. The evaluation of the models is done by the same rules that apply for the backward selection method in the REG procedure—that is, coefficients that are not significant at the 0.1 significance level are removed. Furthermore, if a higher-order term remains in the model, the lower-order polynomials are not being evaluated (for example, if the quadratic term is needed, the server does not try to remove the linear term).

Default 1

**SAVE=table-name**

saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

**TEMPEXPRESS="SAS-expressions"**

**TEMPEXPRESS=file-reference**

specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias TE=
TEMPNAMES=variable-name
TEMPNAMES=(variable-list)
specifies the list of temporary variables for the request. Each temporary variable
must be defined through SAS statements that you supply with the TEMPEXPRESS=
option.

Alias TN=

VARWITH
if this option is specified, the variable-list of size k is interpreted to consist of one
response variable and k–1 regressors. Otherwise, the variable-list is used to compute
all pairs of regressions where the response variable cycles through the left-hand side
of the list. For example, if variable-list is \( x_1, x_2, x_3, \) and \( x_4 \), then the REGCORR
statement computes the following regression models:

<table>
<thead>
<tr>
<th>Response Variable</th>
<th>Regressor Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( x_2 )</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>( x_3 )</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>( x_4 )</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( x_3 )</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( x_4 )</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>( x_4 )</td>
</tr>
</tbody>
</table>

If the VARWITH option is specified, the list of regressions models changes as
follows:

<table>
<thead>
<tr>
<th>Response Variable</th>
<th>Regressor Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( x_2 )</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>( x_3 )</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>( x_4 )</td>
</tr>
</tbody>
</table>

**SUMMARY Statement**
The SUMMARY statement is used to calculate descriptive statistics such as the sample mean, sample
variance, number of observations, sum of squares, and so on. If you specify one or more variables in the
GROUPBY= option, the results are produced separately for each combination of the GROUPBY variables.

**Examples:**
"Example 1: Partitioning a Table into a Temporary Table" on page 235
"Example 4: Deleting Rows and Saving a Table to HDFS" on page 240
Syntax

**SUMMARY** <variable-list> </options>;

**Optional Argument**

*variable-list*

specifies one or more numeric variables. If you do not specify this option, then all numeric variables in the table are used.

**SUMMARY Statement Options**

**AGGREGATE=(aggregation-methods)**

lists the aggregator on which the ordering of the result set is based.

The available aggregation methods are as follows:

- **CSS**  corrected sum of squares
- **CV**  coefficient of variation
- **MAX**  maximum value
- **MEAN**  arithmetic mean
- **MIN**  minimum value
- **N**  number of observations
- **NMISS**  number of missing observations
- **PROBT**  $p$-value for the t-statistic
- **STD**  standard deviation
- **STDERR**  standard error
- **SUM**  sum of the nonmissing values
- **TSTAT**  t-statistic for the null hypothesis that the mean equals zero
- **USS**  uncorrected sum of squares
- **VAR**  sample variance

Alias  **AGG=**

**DESCENDING**

specifies that the levels of the GROUPBY variables are to be arranged in descending order.

Alias  **DESC**

**FORMATS=("format-specification"....)**

specifies the formats for the GROUPBY= variables. If you do not specify the FORMATS= option, or if you omit the entry for a GROUPBY variable, the default format is applied for that variable.

Enclose each format specification in quotation marks and separate each format specification with a comma.

Example

```
proc imstat data=lasr1.table1;
    summary x*y / groupby=(a b) formats=("8.3", "$10");
quit;
```
GROUPBY=(variable-list)
 specifies a list of variable names, or a single variable name, to use as GROUPBY variables in the order of the grouping hierarchy. If you do not specify any GROUPBY variable names, then the calculation is performed across the entire table—possibly subject to a WHERE clause.

GROUPBYLIMIT=n
 specifies the maximum number of levels in a GROUPBY set. When the software determines that there are at least n levels in the GROUPBY set, it abandons the action, returns a message, and does not produce a result set. You can specify the GROUPBYLIMIT= option if you want to avoid creating excessively large result sets in GROUPBY operations.

GROUPFILTER=(filter-options)
 specifies a section of the group-by hierarchy to be included in the computation. With this option, you can request that the server performs the analysis for only a subset of all possible groupings. The subset is determined by applying the group filter to a temporary table that you generate with the GROUPBY statement.

You can specify the following suboptions in the GROUPFILTER option:

- DESCENDING
  specifies the top or the bottom section of the groupings to be collected. If the DESCENDING option is specified, the top LIMIT=n (where n > 0) groupings are collected. Otherwise, the bottom LIMIT=n groupings are collected.

  Alias DESC

- LIMIT=n
  specifies the maximum number of distinct groupings to be collected, where integer n >= 0. If n is zero, then all distinct groupings (up to 2^31-1) that satisfy the boundary constraints, such as LOWERSCORE=f, are collected.

  CAUTION High Cardinality Data Sets Setting n to zero with high-cardinality data sets can significantly delay the response of the server.

- SCOREGT=f
  specifies the exclusive lower bound for the numeric scores of the distinct groupings to collect.

  Alias SGT=

- SCORELT=f
  specifies the exclusive upper bound for the numeric scores of the distinct groupings to collect.

  Alias SLT=

- VALUEGT=("format-name1" <, "format-name2" ...>)
  specifies the exclusive lower bound of the group-by variable’s formatted values for the distinct groupings to collect.

  Alias VGT=

- VALUELT=("format-name1" <, "format-name2" ...>)
  specifies the exclusive upper bound of the group-by variable’s formatted values for the distinct groupings to collect.

  Alias VLT=
TABLE=table-with-groupby-results

specifies the in-memory table from which to load the group-by hierarchy. If the TABLE= option is not specified, then all other GROUPFILTER= options are ignored.

The following program request all the groupings of State, City, and then Trade_In_Model in the Cars_Program_All table. The groupings are ordered by the maximum value of New_Vehicle_Msrp for each grouping:

```
proc imstat;
  table example.CARS_PROGRAM_ALL;
  groupby state city trade_in_model / TEMPTABLE
    WEIGHT=new_vehicle_MSRP
    AGG=(MAX)
    ORDER=WEIGHT;
run;
```

The TEMPTABLE option in the GROUPBY statement directs the server to save all the groupings in a temporary in-memory table. The following DISTINCT statement requests the count of the distinct unformatted values of Sales_Type for each of the selected groupings of State, City, and Trade_In_Model.

```
table example.CARS_PROGRAM_ALL;
distinct sales_type / GROUPFILTER=
  table =mylasr.&_TEMPLAST_
  scoregt=40000
  valuelt=(*"FL","Ft Myers",""")
  limit =20
  descending);
run;
```

This example only considers groupings that have maximum values of the New_Vehicle_Msrp above 40,000 and with formatted values that are less than State="FL" and City="Ft Myers." The empty quotation marks result in no restriction on Trade_In_Model values. These groupings are ordered according to the maximum values of New_Vehicle_Msrp. Because of the DESCENDING option, this example collects the 20 top groupings within the specified group-by range for the DISTINCT analysis.

Interaction

If you specify the GROUPFILTER= option, then the GROUPBY= and FORMATS= options have no effect.

LIMIT=n

limits the size of the result set returned to the SAS client. For example, the following SUMMARY statement returns the size (in number of records) of the largest partition for Table1.

```
proc imstat data=mylasr.Table1;
  summary Amount / partition orderby=(Amount) desc
    aggregate=(N) limit=1;
run;
```

Because of the PARTITION option, the rows are processed by partition and the summary request returns one row per partition. The ORDERBY= specification requests that the results are ordered by the variable Amount, and that the result is sorted in descending order with respect to the number of observations (AGGREGATE=N). Once the results are arranged this way, the first observation in the set is returned (LIMIT=1).
MERGEBINS=b
specifies the number of bins to create when a numeric GROUPBY variable exceeds
the MERGELIMIT=n specification. If you specify a MERGELIMIT, but do not
specify a value for the MERGEBINS= option, the server automatically calculates the
number of bins.

MERGELIMIT=n
specifies that when the number of unique values in a numeric GROUPBY variable
exceeds n, the variable is automatically binned and the GROUPBY structure is
determined based on the binned values of the variable, rather than the unique
formatted values.

For example, if you specify MERGELIMIT=500, any numeric GROUPBY variable
with more than 500 unique formatted values is binned. Instead of returning results
for more than 500 groups, the results are returned for the bins. You can specify the
number of bins with the MERGEBINS= option.

DESCENDING
specifies that the levels of the GROUPBY variables are to be arranged in descending
order.

Alias DESC

NOPREPARSE
prevents the procedure from pre-parsing and pre-generating code for temporary
expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects
problems with the code, the error messages might not be as detailed as the
messages that are generated by SAS client. If you are debugging your user-written
program, then you might want to pre-parse and pre-generate code in the procedure.
However, if your SAS statements compile and run as you want them to, then you can
specify this option to avoid the work of parsing and generating code on the SAS
client.

When you specify this option in the PROC IMSTAT statement, the option applies to
all statements that can generate code. You can also exclude specific statements from
pre-parsing by using the NOPREPARSE option in statements that allow temporary
columns or the SCORE statement.

Alias NOPREP

NOTEMPPART
specifies that the temporary table generated by the TEMPTABLE option is not
partitioned by the GROUPBY= variables. When you request a temporary table with
the SUMMARY statement, by default, the server partitions the table and the size of a
partition is equal to the number of analysis variables in the variable-list of the
SUMMARY statement. When the number of groups is large, this can result in many
small partitions, and requires extra memory resources to store the partition
information for the temporary table. By specifying this option, the temporary table is
organized similarly to the default table, but is not a partitioned table.

Alias NOTP

ORDERBY=(variable-list)
specifies the variables to use for ordering the result set. If a variable is not one of the
numeric variables in the variable-list for the SUMMARY statement, it is assumed to
be one of the GROUPBY variables.
ORDERDESC
specifies the sort order for the result set. The default is ascending order. Specify the
ORDERDESC option to sort in descending order. Note that the ORDERDESC
option is different from setting the DESCENDING option. The DESCENDING
option affects the order of the values for the GROUPBY variables.

PARTITION <=partition-key>
When you specify this option and the table is partitioned, the results are calculated
separately for each value of the partition key. In other words, the partition variables
function as automatic GROUPBY variables. This mode of executing calculations by
partition is more efficient than using the GROUPBY= option. With a partitioned
table, the server takes advantage of knowing that observations for a partition cannot
be located on more than one worker node.

If you do not specify a partition-key, the analysis is performed for all partitions. If
you do specify a partition-key, the analysis is carried out for the specified key value
only. You can use the PARTITIONINFO statement to retrieve the valid partition key
values for a table.

You can specify a partition-key in two ways. You can supply a single quoted string
that is passed to the server, or you can specify the elements of a composite key
separated by commas. For example, if you partition a table by variables GENDER
and AGE, with formats $1 and BEST12, respectively, then the composite partition
key has a length of 13. You can specify the partition for the 11-year-old females as
follows:

statement / partition="F 11"; /* passed directly to the server */
statement / partition="F","11"; /* composed by the procedure */

If you choose the second format, the procedure composes a key based on formatting
information from the server.

Alias PART=

RAWORDER
specifies that the ordering of the GROUPBY variables is based on the raw values of
the variables instead of the formatted values.

SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements
like STORE, REPLAY, and FREE. The value for table-name must be unique within
the scope of the procedure execution. The name of a table that has been freed with
the FREE statement can be used again in subsequent SAVE= options.

SETSIZE
requests that the server estimate the size of the result set. The procedure does not
create a result table if the SETSIZE option is specified. Instead, the procedure reports
the number of rows that are returned by the request and the expected memory
consumption for the result set (in KB). If you specify the SETSIZE option, the SAS
log includes the number of observations and the estimated result set size. See the
following log sample:

NOTE: The LASR Analytic Server action request for the STATEMENT
statement would return 17 rows and approximately
3.641 kBytes of data.

The typical use of the SETSIZE option is to get an estimate of the size of the result
set in situations where you are unsure whether the SAS session can handle a large
result set. Be aware that in order to determine the size of the result set, the server has
to perform the work as if you were receiving the actual result set. Requesting the
estimated size of the result set does consume resources on the server. The estimated
number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.

**TEMPEXPRESS=“SAS-expressions”**

**TEMPEXPRESS=file-reference**

specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias   TE=

**TEMPNAMES=variable-name**

**TEMPNAMES=(variable-list)**

specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

Alias   TN=

**TEMPTABLE**

generates an in-memory temporary table from the result set. The IMSTAT procedure displays the name of the table and stores it in the &_TEMPLAST_ macro variable, provided that the statement executed successfully.

When the IMSTAT procedure exits, all temporary tables created during the IMSTAT session are removed. Temporary tables are not displayed on a TABLEINFO request, unless the temporary table is the active table for the request.

**TWELVEBIN**

specifies to augment the summary results with a 12-bin histogram. This option has no effect when the summaries are computed in GROUPBY or partitioned mode.

---

**TEXTPARSE Statement**

The TEXTPARSE statement performs text analytics on the active in-memory table. You can separate the documents in the table into terms, derive topics based on weighted term frequencies, and project the active table onto the latent space defined by the topic-discovered topics.

---

**Syntax**

TEXTPARSE TXT=text-variable ID=document-ID <options>;

**Required Arguments**

**TXT=text-variable**

specifies the name of the variable that contains the text to analyze.

**ID=document-ID**

specifies the name of the variable that identifies the documents in the table uniquely. The values are typically a row number or other value that identifies the rows. The document ID is important to perform joins of the result tables.

Alias   DOCID=
TEXTPARSE Statement Options

**CELLWGT=** NONE | LOG
specifies how elements in the term × document matrix are weighted. Elements in the matrix are assigned weight $w_i \times g(f_{ij})$, where $w_i$ is the term weight for the $i$th term, $f_{ij}$ is the frequency of appearance of this term in document $j$.

If CELLWGT=LOG, then $g(f_{ij}) = \log_2(f_{ij}+1)$. The logarithmic function tempers the influence of very frequent terms.

**Default** LOG

**ENTITIES=** NONE | STD
determines whether the entity extractor should use the standard list of entities. When ENTITIES=STD, entity extraction is enabled and standard entities are used. Terms such as "George W. Bush" are then recognized as an entity and given the corresponding entity role and attribute. For this example, the entity role is PERSON and the attribute is Entity. Although the entity is treated as the single term, "george w bush," the individual tokens "george," "w," and "bush" are also included.

**Default** NONE

**EXACTWEIGHT**
specifies not to round the weights that are aggregated during topic derivation. By default, the calculated weights are rounded to the nearest .001.

**Alias** NOWTRND

**KEEP=**(variable-list)
**KEEP=**variable-name
specifies one or more variables to transfer from the input data to the temporary table with the document projection. You can use _ALL_ for all variables, _NUMERIC_ for all numeric variables, and other valid variable list names. By default, only the document ID (ID=) is transferred to the projected document table so that it can be used to join with the active table.

**NONOUNGROUPS**
specifies not to use the noun group extractor. By default, the server extracts noun groups and returns maximal groups and subgroups (which do not include groups that contain determiners or prepositions). If stemming is turned on, then noun group elements are also stemmed.

**Alias** NONG

**NOSTEMMING**
specifies not to stem words. By default, words are stemmed and terms such as "advises" and "advising" are mapped to the parent term "advise."

**Alias** NOSTEM

**NOTAGGING**
specifies not to tag terms. By default, terms are tagged and the server identifies a term's part of speech based on context clues. The identified part of speech is provided in the Role variable of the TERMS table.

**NUMLABELS=**n
specifies the number of terms to use in labeling a topic. By default, the $n = 5$ terms with the largest weight are used in constructing a label for the topic.
Alias NLABELS=

Default 5

REDUCEF=n
specifies the minimum document frequency of terms. By default, n = 4 and implies that a term is not kept for analysis unless it occurs in at least four documents.

Default 4

SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

SELECT <= (list-of-temporary-tables)
specifies the results the server should store as temporary tables. By default, the server generates the Terms table, which contains terms, their parent-child relationships, and weights. If you specify the NUMTOPICS= option, the server also generates the Topics table. You can specify SELECT=(_ALL_) to generate all of the tables.

The possible values for the list specification are shown in the following table:

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Table Alias</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TERMS</td>
<td>TERM</td>
<td>Contains summary information about the terms in the document collection.</td>
</tr>
<tr>
<td>TERMDOC</td>
<td>BAGOFWORDS</td>
<td>Contains a compressed representation of the sparse term-by-document frequency matrix in transactional style. The matrix is represented as a set of (row, column, value) triples.</td>
</tr>
<tr>
<td>V</td>
<td>SVDV</td>
<td>Contains the V matrix of the singular-value decomposition.</td>
</tr>
<tr>
<td>U</td>
<td>SUDV</td>
<td>Contains the rotated U matrix of the singular-value decomposition.</td>
</tr>
<tr>
<td>Table Name</td>
<td>Table Alias</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>PROJECTION</td>
<td>DOCPRO, PROJ</td>
<td>Contains the projections of the columns of the term-by-document frequency matrix onto the columns of U. Because each column of the term-by-document frequency matrix corresponds to a document, the output forms a new representation of the input documents in a space with much lower dimensionality.</td>
</tr>
<tr>
<td>TOPICS</td>
<td></td>
<td>Contains the topics and a label constructed from the most highly weighted terms. This is typically a small table, as the number of topics is limited by ( k ), the value of the singular-value decomposition or by the value specified in the NUMTOPICS= option.</td>
</tr>
<tr>
<td>TERMTOPICS</td>
<td>TERMBYTOPICS</td>
<td>A sparse representation of the terms by topic using the term ID and topic ID. This table might be useful in joins involving terms or topics.</td>
</tr>
</tbody>
</table>

**START=table-name**

specifies the name of the in-memory table that contains the terms that are to be kept for the analysis. These terms are displayed in the Terms result table with a keep status of "Y." The START= table must have a variable that is named Term and can also have a variable that is named Role.

**STOP=**

specifies the name of the in-memory table that contains the terms to exclude from the analysis. The STOP= table must contain a variable that is named Term and can also have a variable that is named Role.

**SVD(singular-value-decomposition-options)**

specifies how to perform the singular-value decomposition (SVD). The server carries out this decomposition whenever you request a result table that depends on topics, or if you request to save the V or U matrix of the decomposition. You can specify the following SVD options inside the parentheses:
K=k

specifies the number of dimensions to be extracted by SVD. This number is equal to the number of topics for topic generation. If you specify the TOPICS= (NUMTOPICS= ) option, then the value of $k$ is automatically set to match the value given in the TOPICS= option.

If the value of $k$ is too large, then the server might process for an unnecessarily long time.

Default

If you request topic generation and do not specify the K= or MAXK= option, then $k = 10$.

Interaction

If you specify both the K= and MAXK= options, the K= option takes precedence.

MAXK=m

specifies the maximum value that the server should return as the recommended value of $m$. If the RESOLUTION= option is specified to recommend the value of $k$, then this option limits that value to at most $m$. The HPTMINE procedure attempts to calculate (as opposed to recommends) $k$ dimensions when it performs the singular-value decomposition.

Interaction

If you specify both the K= and MAXK= options, the K= option takes precedence.

RESOLUTION=LOW | MED | HIGH

specifies the recommended number of dimensions (resolution) for the singular value decomposition. If you specify this option, you must also specify the MAXK= option. A low-resolution singular value decomposition returns fewer dimensions than a high-resolution singular value decomposition. This option recommends the value of $k$ (the number of topics) heuristically based on the value specified in the MAXK= option.

Assume that the MAXK=$n$ option and the singular value decomposition with $n$ dimensions accounts for $r\%$ of the total variance. If you specify RES=HIGH, the server always recommends the maximum number of dimensions. That is, $k=n$. If you specify RES=MED, the server recommends a value for $k$ that explains $(5/6) \times r\%$ of the total variance. If you specify RES=LOW, the server recommends a value for $k$ that explains $(2/3) \times r\%$ of the total variance.

Alias

RES=

TOL=\varepsilon

specifies the maximum allowable tolerance for the singular value.

Default

The value of epsilon on the machine where the server is running.

SYNONYMS=table-name

specifies the name of an in-memory table that contains user-defined synonyms to use in the analysis. The table specifies parent-child relationships that enable you to map child terms to a representative parent. The synonym relationship is indicated in the Terms result table and is also reflected in the term-by-document result table known as the Termdoc or Parent table.

The specified table must have either the two variables Term and Parent, or the four variables Term, Parent, Termrole, and Parentrole. When stemming is enabled (the default), the relationships provided by the SYNONYMS= table take precedence over relationships that are identified through term stemming.
TERMWGFT=ENTROPY | MI | NONE
specifies how terms are weighted. TERMWGFT=ENTROPY specifies that terms are weighted using the entropy formulation. If you specify TERMWGFT=MI, then terms are weighted using the mutual information formulation. Specifying TERMWGFT=NONE suppresses term weighting. See the documentation for the HPTMINE procedure for the details about computing term weights.
If you specify TERMWGFT=MI, then you must specify a target variables with the TARGET= option.

Default ENTROPY

TOPICS=n
specifies the number of topics to generate. When you specify n, the server automatically produces a table of topics with up to n entries. You can also request the Topics table with the SELECT= option. Specifying TOPICS=n is equivalent to requesting topics based on a singular-value decomposition with n=k factors.

NUMTOPICS=

You can use the NUMLABELS= option to control the number of terms to use in labeling the topic.

TARGET=target-variable
specifies a variable that contains information about the category that a document belongs to. If specified, the target variable is used in computing term weights. For example, it is used with TERMWGFT=MI.

TEMPEXPRESS="SAS-expressions"

specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

TEMPNAMES=variable-name

specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

TOPK Statement
The TOPK statement calculates and selects the top-k and bottom-k distinct values of a variable based on a user-specified ranking order. The distinct values can be reported as raw or formatted values. The ranking can be based on the raw value, the formatted value, the frequency count, or based on a calculated score derived from the values of a weight variable. You can also specify aggregate functions to roll up multiple weight values into a single score for a distinct value.
Syntax

**TOPK** <variable-list> / options;  

**Optional Argument**

**variable-list**

specifies one or more numeric variables. If you do not specify this option, then all numeric variables in the table are used.

**Topk Statement Options**

**AGGREGATE=(aggregation-methods)**

specifies the aggregation methods for which WEIGHT= variable values are rolled up into rank order score for distinct values. If no WEIGHT= variable is specified, then this option is ignored.

The available aggregation methods are as follows:

- **MAX** specifies to use the maximum value of the weight values
- **MEAN** specifies to use the arithmetic mean of the weight values
- **MIN** specifies to use the minimum value of the weight values
- **SUM** specifies to use the sum of the weight values

**FORMATS=("format-specification", ...)**

specifies the formats for the variables. If you do not specify the FORMATS= option, or if you omit the entry for a variable, the default format is applied for that variable. Enclose each format specification in quotation marks and separate each format specification with a comma.

**Example**

```
proc imstat data=lasr1.table1;
    topk x1 x2 / formats=("10.2", 10.2");
quit;
```

**FREQ=variable-name**

specifies the numeric frequency variable to use for calculating the rank order score for distinct values. This option is valid when ORDER=FREQ or when AGGREGATE= is N, SUM, or MEAN only.

**K1=n**

specifies the maximum number of distinct values to include in the top-k list.

**K2=n**

specifies the maximum number of distinct values to include in the bottom-k list.
DESCENDING
specifies that the levels of the GROUPBY variables are to be arranged in descending order.

Alias DESC

ORDER= FREQ | VALUE | WEIGHT
specifies the rank ordering to apply to the distinct values when no WEIGHT= variable is specified. The following rank orders are valid in the TOPK request.

The available ordering methods are as follows:

FREQ specifies to order by frequency count
VALUE specifies to order by raw or formatted values of the variable
WEIGHT specifies to order by the aggregate values of the WEIGHT= variable

Default FREQ

WEIGHT=variable-name
specifies the numeric weight variable to use for calculating the rank order score. If you specify ORDER= and WEIGHT=, then the WEIGHT= variable takes priority over ORDER.

SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

TEMPEXPRESS="SAS-expressions"
TEMPEXPRESS=file-reference
specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias TE=

TEMPNAMES=variable-name
TEMPNAMES=(variable-list)
specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

Alias TN=

QUIT Statement
The QUIT statement is used to end the procedure execution. When the procedure reaches the QUIT statement, all resources allocated by the procedure are released. You can no longer execute procedure statements without invoking the procedure again. However, the connection to the server is not lost, because that connection was made through the SAS LASR Analytic Server engine. As a result, any
subsequent invocation of the procedure that uses the same libref executes almost instantaneously because the engine is already connected to the server.

**Interaction:** Using a DATA step or another procedure step is equivalent to issuing a QUIT statement. If there is an error during the procedure execution, it is also equivalent to issuing a QUIT statement.

---

**Syntax**

QUIT;

---

**Examples: IMSTAT Procedure (Analytics)**

**Example 1: Calculating Percentiles and Quartiles**

**Details**

If you specify the PERCENTILE statement without variables or options, you obtain results for the 25th, 50th, and 75th percentile. These are also known as the first quartile, the median, and the third quartile. This is done for all numeric non-CLASS variables in the table.

This PROC IMSTAT demonstrates the default behavior for calculating percentiles for a single variable and then demonstrates using GROUPBY= variables and generating results for nonstandard percentiles.

**Program**

libname example sasiola host="grid001.unx.sas.com" port=10010 tag='hps';

data example.prdsale; set sashelp.prdsale; run;

proc imstat data=example.prdsale;
   percentile actual;  
run;
   percentile actual / groupby=(region division);  
run;
   percentile actual / values=(3 5 10 90 95 97);  
quit;

**Program Description**

1. This PERCENTILE statement generates the default output for the Actual variable.
2. The quartiles for the Actual variable are calculated for the groups of Region and Division.
3. The VALUES= option is used to specify the percentiles to calculate.
Output

The following results include the column that is named Converged. This column indicates whether the iterative percentile algorithm converged for the variable and percentile. It is possible that some percentiles can fail to converge while others do converge. The converged percentiles match those computed with the UNIVARIATE or MEANS procedures using their default definition of a quantile.

Output 4.1  Default Output for a Single Variable

| Percentiles and Quantiles for Table WORK.PRDSALE |
|----------------|----------------|----------|--------|
| Column         | Percentile | Value    | Converged |
| ACTUAL         | 25         | 261.0000000 | Yes    |
| ACTUAL         | 50         | 503.0000000 | Yes    |
| ACTUAL         | 75         | 756.5000000 | Yes    |

Output 4.2  Results for Actual Grouped by Region and Division

| Percentiles and Quantiles for Table WORK.PRDSALE |
|----------------|----------------|----------|--------|
| REGION         | DIVISION       | Column   | Percentile | Value    | Converged |
| EAST           | CONSUMER       | ACTUAL   | 25         | 260.5000000 | Yes    |
| EAST           | CONSUMER       | ACTUAL   | 50         | 494.5000000 | Yes    |
| EAST           | CONSUMER       | ACTUAL   | 75         | 745.5000000 | Yes    |
| EAST           | EDUCATION      | ACTUAL   | 25         | 277.5000000 | Yes    |
| EAST           | EDUCATION      | ACTUAL   | 50         | 548.5000000 | Yes    |
| EAST           | EDUCATION      | ACTUAL   | 75         | 759.5000000 | Yes    |
| WEST           | CONSUMER       | ACTUAL   | 25         | 243.5000000 | Yes    |
| WEST           | CONSUMER       | ACTUAL   | 50         | 508.5000000 | Yes    |
| WEST           | CONSUMER       | ACTUAL   | 75         | 750.0000000 | Yes    |
| WEST           | EDUCATION      | ACTUAL   | 25         | 225.0000000 | Yes    |
| WEST           | EDUCATION      | ACTUAL   | 50         | 465.5000000 | Yes    |
| WEST           | EDUCATION      | ACTUAL   | 75         | 774.0000000 | Yes    |
Example 2: Retrieving Box Values

Details
This PROC IMSTAT example demonstrates retrieving the statistics for a box plot. The BOXPLOT statement does not generate a plot, it generates values that can be used to create a plot.

Program
```
libname example sasiola host="grid001.example.com" port=10010 tag='hps';

data example.cars;
set sashelp.cars;
run;

proc imstat data=example.cars;
boxplot;
quit;
```

Program Description
1. The sashelp.cars data set is loaded to memory on the SAS LASR Analytic Server.
2. The in-memory table is referenced with the DATA= option and then the BOXPLOT statement is used.

Output
The arithmetic mean is reported because the center line of the box plot is sometimes drawn at the mean and not the median. In some displays the median is shown as a line and the box is augmented with a graphic symbol at the position of the mean. The low and high whiskers are values of actual observations in the data set. These values might be the minimum or maximum values in the data set if the value for that observation equals the value nearest 1.5 times the inter-quartile range from the edge of the box. The box in box plot is drawn from the first quartile to the third quartile.
Example 3: Retrieving Box Plot Values with the NOUTLIERLIMIT= Option

Details

When you specify the NOUTLIERLIMIT= option, the IMSTAT procedure requests outlier information for the variables. When outliers are reported for a variable, pay attention to the last two columns of the display (columns Lo Bin and Hi Bin). These two columns let you know whether the values displayed in the outlier columns are actual data values, or counts in bins. For more information, see the information in the Output section.

Program

```
libname example sasiola host="grid001.example.com" port=10010 tag='hps';

data example.cars;
  set sashelp.cars;
run;

proc imstat data=example.cars;
  boxplot / noutlierlimit=7;
  ods output boxplot=outliers;
quit;

proc print data=outliers noobs;
  var column outlo1-outlo5 outhi1-outhi10 binlo binhi;
run;
```

Program Description

1. The program example requests that the raw values for up to seven high outliers and up to seven low outliers are retrieved. If there are more than seven outliers, the procedure returns the binned values for the outlying values.
The ODS statement and the PRINT procedure that follows are display purposes only.

Output

In example that follows, using the CARS data set, several variables exhibit outliers on the low end. For example, there are two outlying values for the MPG_City variable. Since the Lo Bin column of the result table is set to **No** for this variable, the values, 10 for OutLo1 and 10 for OutLo2, are actual values in the data. Note that these values are smaller than the lower whisker value of 12. (See the previous example.) Similarly, the Horsepower variable shows several outliers on the high end of the distribution and the Hi Bin column is set to **No**. This lets you know that the values 493, 450, 500, and so on, represent actual values in the CARS table.

On the other hand, the Hi Bin column for the MSRP variable is set to **Yes**. This lets you know that more outliers were found than the specified NOUTLIERLIMIT= limit of 7. The outliers are then placed in bins and the binned counts are reported. For example, there are 14 values in the first bin of MSRP outliers, 8 values in the second bin, 1 value in the fourth bin, and no value in the fifth bin, and so on.

Output 4.5  **BOXPLOT Statement Results with the NOUTLIERS= Option**

<table>
<thead>
<tr>
<th>Column</th>
<th>OutLo1</th>
<th>OutLo2</th>
<th>OutLo3</th>
<th>OutLo4</th>
<th>OutHi1</th>
<th>OutHi2</th>
<th>OutHi3</th>
<th>OutHi4</th>
<th>OutHi5</th>
<th>OutHi6</th>
<th>OutHi7</th>
<th>OutHi8</th>
<th>OutHi9</th>
<th>BinLo</th>
<th>BinHi</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSRP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EngineSize</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cylinders</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horsepower</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPG_City</td>
<td>10</td>
<td>10</td>
<td>1</td>
<td>16</td>
<td>14</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MPG_Highway</td>
<td>16</td>
<td>16</td>
<td>13</td>
<td>1</td>
<td>9</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Weight</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wheelbase</td>
<td>90</td>
<td>99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length</td>
<td>153</td>
<td>144</td>
<td>143</td>
<td>150</td>
<td>153</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Example 4: Performing a Cluster Analysis

Details

You can perform a clustering analysis for all variables in an in-memory table by simply issuing a CLUSTER statement. However, specifying the variables to analyze and options can be specified to provide more a meaningful analysis.

The following SAS statements load the famous Iris flower data of R.A. Fisher to memory, and then perform \( k \)-means clustering on four of the variables.

Program

```sas
libname example sasiola host="grid001.example.com" port=10010 tag='hps';

data example.iris;
  set sashelp.iris;
run;

proc imstat data=example.iris;
  cluster SepalLength SepalWidth PetalLength PetalWidth / x
    maxiter=50
    numclus=3
    nsamp =2
run;
```
conv = 1.e-06
init = rand
freq = Species;
quit;

Program Description

1. The four variables to analyze are specified in the CLUSTER statement.
2. Species is specified as the frequency variable and is used to cluster the four variables.

Output

Output 4.6  CLUSTER Statement Results for the Iris Data Set

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Number of Obs</th>
<th>Root Mean Square of STD</th>
<th>Maximum Distance from Seed to Obs.</th>
<th>Minimum Distance from Seed to Obs.</th>
<th>Within Sum of Cluster Square</th>
<th>Nearest Cluster</th>
<th>Distance between Centroids</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>61</td>
<td>3.9543</td>
<td>16.4650</td>
<td>2.3571</td>
<td>3129.06</td>
<td>1</td>
<td>17.842</td>
</tr>
<tr>
<td>1</td>
<td>39</td>
<td>4.0200</td>
<td>15.6156</td>
<td>2.3945</td>
<td>2541.38</td>
<td>0</td>
<td>17.842</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>2.7603</td>
<td>12.4603</td>
<td>0.6615</td>
<td>1515.19</td>
<td>0</td>
<td>33.424</td>
</tr>
</tbody>
</table>

Example 5: Performing a Pairwise Correlation

Details

This PROC IMSTAT example demonstrates how to perform a pairwise correlation for all the numeric variables in the Iris data set.

Program

```
libname example sasiola host="grid001.example.com" port=10010 tag='hps';

data example.iris;
    set sashelp.iris;
run;

proc imstat data=example.iris;
    corr;
quit;
```
Output
The output does not display a correlation matrix in the statistical sense. It is a collection of pairwise correlations.

<table>
<thead>
<tr>
<th>Column</th>
<th>Row</th>
<th>SepalLength</th>
<th>SepalWidth</th>
<th>PetalLength</th>
<th>PetalWidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>SepalLength</td>
<td>1</td>
<td>1.0000</td>
<td>-0.1176</td>
<td>0.8718</td>
<td>0.8179</td>
</tr>
<tr>
<td>SepalWidth</td>
<td>2</td>
<td>-0.1176</td>
<td>1.0000</td>
<td>-0.4284</td>
<td>-0.3661</td>
</tr>
<tr>
<td>PetalLength</td>
<td>3</td>
<td>0.8718</td>
<td>-0.4284</td>
<td>1.0000</td>
<td>0.9629</td>
</tr>
<tr>
<td>PetalWidth</td>
<td>4</td>
<td>0.8179</td>
<td>-0.3661</td>
<td>0.9629</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Example 6: Crosstabulation with Measures of Association and Chi-Square Tests

Details
To compute measures of association of the row and column variable, you can add the MEASURE option to the CROSSTAB statement. ASSOCIATION is an alias for the MEASURE option. You can also request Chi-Square statistics for the test of independence between row and column variable with the CHISQ option.

The following statements request a crosstabulation of the Cylinders and Origin variables for the CARS data set. The statements also request measures of association and Chi-Square statistics. The NOMISS option is used to exclude levels of the variables that correspond to missing values.

```plaintext
libname example sasiola host="grid001.example.com" port=10010 tag='hps';

data example.cars; set sashelp.cars; run;

proc imstat data=example.cars;
    crosstab cylinders * origin / measures chisq nomiss;
    quit;
```

Program Description
1. The variables to use for the columns and rows are specified in the CROSSTAB statement.
2. The NOMISS option excludes levels that have missing values. Note that by default the FREQ procedure excludes levels with missing values, whereas the IMSTAT procedure includes those as valid levels.
Output

Output 4.7  Crosstabulation Results

<table>
<thead>
<tr>
<th>Cylinders</th>
<th>Asia</th>
<th>Europe</th>
<th>USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>74</td>
<td>25</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>69</td>
<td>54</td>
<td>67</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>34</td>
<td>41</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Output 4.8  Measures of Association Results

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>Asymptotic Standard Error</th>
<th>Lower 95% Confidence Limit</th>
<th>Upper 95% Confidence Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma</td>
<td>0.3433</td>
<td>0.0580</td>
<td>0.2295</td>
<td>0.4570</td>
</tr>
<tr>
<td>Kendall's Tau-B</td>
<td>0.2316</td>
<td>0.0404</td>
<td>0.1525</td>
<td>0.3108</td>
</tr>
<tr>
<td>Stuart's Tau-c</td>
<td>0.2284</td>
<td>0.0402</td>
<td>0.1506</td>
<td>0.3082</td>
</tr>
<tr>
<td>Somers' D CJR</td>
<td>0.2227</td>
<td>0.0403</td>
<td>0.1538</td>
<td>0.3117</td>
</tr>
<tr>
<td>Somers' D RJC</td>
<td>0.2305</td>
<td>0.0405</td>
<td>0.1511</td>
<td>0.3100</td>
</tr>
<tr>
<td>Lambda Asymmetric CJR</td>
<td>0.1519</td>
<td>0.0275</td>
<td>0.0960</td>
<td>0.2059</td>
</tr>
<tr>
<td>Lambda Asymmetric RIC</td>
<td>0.0212</td>
<td>0.0501</td>
<td>0.0000</td>
<td>0.1194</td>
</tr>
<tr>
<td>Lambda Symmetric</td>
<td>0.0509</td>
<td>0.0268</td>
<td>0.0364</td>
<td>0.1435</td>
</tr>
<tr>
<td>Uncertainty Coefficient CJR</td>
<td>0.0776</td>
<td>0.0156</td>
<td>0.0469</td>
<td>0.1082</td>
</tr>
<tr>
<td>Uncertainty Coefficient RIC</td>
<td>0.0713</td>
<td>0.0134</td>
<td>0.0451</td>
<td>0.0974</td>
</tr>
<tr>
<td>Uncertainty Coefficient Symmetric</td>
<td>0.0743</td>
<td>0.0144</td>
<td>0.0461</td>
<td>0.1025</td>
</tr>
</tbody>
</table>
Example 7: Training and Validating a Decision Tree

Details

This PROC IMSTAT example demonstrates how to use the DECISIONTREE statement to generate a decision tree and then use a validation data set for scoring against the tree.

The data for this example is available from the Machine Learning Repository of the University of California at Irvine.


Program

```plaintext
libname mylib 'path-to-datasets';
libname example sasiola host="grid001.example.com" port=10010 tag='hps';

data example.bank_train_1; set mylib.bank_train_1;
data example.bank_valid_1; set mylib.bank_valid_1;run;

proc imstat data=example.bank_train_1;
decisiontree subscribe_term_deposit /
nbins      =10
maxlevel   =7
maxbranches=4
input      =(age job marital_status education
default balance housing loan contact
day month duration campaign previous
poutcome)
nominal    =(contact default education housing
job loan marital_status month
poutcome)
multvar
prune
leafsize   =5
save       =DTreeTab; 1
/*   ods output dtree=example.banktree_train_1; 2 */
run;

decisiontree subscribe_term_deposit /
treetab    =DTreeTab
```

Output 4.9 Chi-Square Statistics Results

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>12</td>
<td>6.7854</td>
</tr>
</tbody>
</table>
| Likelihood Ratio Chi-Square| 12  | 72.3057 | <.0001

Chi-Square Statistics of Cylinders by Origin for Table WORK.CARS
Program Description

1. The SAVE= option stores the result table so that it can be used in subsequent statements. It is named DTreeTab.

2. As an alternative to the SAVE= option, the ODS OUTPUT statement can also be used to save the result table.

3. To use the table that was stored with the ODS OUTPUT statement, the TABLE statement switches the active table to bank_valid_1.

4. The TREEDATA= option specifies the decision tree that was saved with the ODS OUTPUT statement.

Output

Output 4.10  Partial Results for the Decision Tree Created from the Training Data
Example 8: Storing and Scoring a Decision Tree

Details

You can store the representation of a decision tree in an in-memory table on the server and at the same time score the input table. This process generates two temporary tables: the temporary table with the tree representation and the temporary table with the scoring results.

This enables you to compute decision trees for high-cardinality problems. The results from tree building and tree scoring are available to you without transferring large amounts of data between the SAS client and the server. You can query the tree (for a drill-down, for example) and query the scoring information efficiently, using the storing and querying features of the server. Also, by storing them as temporary tables, you can process them with other IMSTAT procedure statements.

Program

```sas
libname example sasiola host="grid001.example.com" port=10010 tag='hps';

data example.heart;
set sashelp.heart;
run;

proc imstat;
  table example.heart;
  decisiontree Weight / input=(Sex DeathCause Chol_Status BP_Status Weight Smoking_Status)
   nbinstarget=5
   temptable 1
   vars=(Sex DeathCause Chol_Status BP_Status Weight Smoking_Status)
   nomissobs;
run;
```
Table Example 3: Storing and Scoring a Decision Tree

1. The TEMPTABLE option specifies to save the decision tree and the scoring results in in-memory tables on the server.

2. The &TEMPTREE macro variable is used to access the representation of the decision tree and the following FETCH statement prints a subset of the variables from the first ten rows.

3. The &TEMPSCORE macro variable is used to access the scoring table. The following FETCH statement prints the first five rows.

4. The TREELASR= option demonstrates how to score an input table explicitly that is already in memory.

Program Description

1. The TEMPTABLE option specifies to save the decision tree and the scoring results in in-memory tables on the server.

2. The _TEMPTREE_ macro variable is used to access the representation of the decision tree and the following FETCH statement prints a subset of the variables from the first ten rows.

3. The _TEMPSCORE_ macro variable is used to access the scoring table. The following FETCH statement prints the first five rows.

4. The TREELASR= option demonstrates how to score an input table explicitly that is already in memory.

Output

Output 4.12 Partial Results for the Decision Tree Created from the Heart Data Set

<table>
<thead>
<tr>
<th><em>CI0</em></th>
<th><em>CI1</em></th>
<th><em>Val0</em></th>
<th><em>Val1</em></th>
<th>Parent</th>
<th><em>ParentName</em></th>
<th><em>NodeType</em></th>
<th>NodeName_</th>
<th>Gain_</th>
<th>NumObs</th>
<th><em>TargetLowerbd</em></th>
<th><em>TargetUpperbd</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5002505172</td>
<td>0.3875049667</td>
<td>0.3529411765</td>
<td>0.3529411765</td>
<td>Desirable</td>
<td>15</td>
<td>Chol_Status</td>
<td>CLASS</td>
<td>Sex</td>
<td>0.1013667062</td>
<td>1618</td>
<td>113.6</td>
</tr>
<tr>
<td>0.6</td>
<td>0.4</td>
<td>High</td>
<td>Desirable</td>
<td>10</td>
<td>Chol_Status</td>
<td>LEAF</td>
<td>Weight</td>
<td>0</td>
<td>17</td>
<td>113.6</td>
<td>160.2</td>
</tr>
<tr>
<td>0.510037359</td>
<td>0.4023101087</td>
<td>0.5510204082</td>
<td>0.2244975769</td>
<td>High</td>
<td>Borderline</td>
<td>15</td>
<td>Chol_Status</td>
<td>CLASS</td>
<td>BP_Status</td>
<td>0.0359165052</td>
<td>160</td>
</tr>
<tr>
<td>0.634032934</td>
<td>0.2259037296</td>
<td>0.4295714296</td>
<td>0.3571425717</td>
<td>Female</td>
<td>0</td>
<td>Sex</td>
<td>CLASS</td>
<td>BP_Status</td>
<td>0.055274465</td>
<td>63</td>
<td>113.6</td>
</tr>
<tr>
<td>0.571666667</td>
<td>0.3216666667</td>
<td>0.9354615385</td>
<td>0.423076923</td>
<td>High</td>
<td>1</td>
<td>BP_Status</td>
<td>CLASS</td>
<td>Death_Cause</td>
<td>0.0139436036</td>
<td>60</td>
<td>113.6</td>
</tr>
<tr>
<td>0.5394518285</td>
<td>0.423076923</td>
<td>0.423076923</td>
<td>0.423076923</td>
<td>Female</td>
<td>15</td>
<td>Chol_Status</td>
<td>CLASS</td>
<td>LEAF</td>
<td>0</td>
<td>52</td>
<td>113.6</td>
</tr>
<tr>
<td>0.580696522</td>
<td>0.4434702069</td>
<td>0.5002505172</td>
<td>0.3875049667</td>
<td>Normal</td>
<td>1</td>
<td>BP_Status</td>
<td>CLASS</td>
<td>Smoking_Status</td>
<td>0.030616307</td>
<td>460</td>
<td>113.6</td>
</tr>
</tbody>
</table>
When the tree is scored with the DECISIONTREE statement (the last statement in the example), the misclassification rate information is printed to the SAS log.

**Output 4.14  Misclassification Rate Information**

NOTE: The misclassification rate for scoring the decision tree is 0.368401 using table EXAMPLE.HEART with 5209 records out of 5209.

---

**Example 9: Performing a Multi-Dimensional Summary**

**Details**

This PROC IMSTAT example demonstrates creating multi-dimensional summaries of the Prdsale data set. Three set specifications are shown in the example. There is no limit to the number of set specifications that you can specify.

**Program**

```sas
libname example sasiola host="grid001.example.com" port=10010 tag='hps';

data example.prdsale; set sashelp.prdsale; run;

proc imstat data=example.prdsale;
  mdsummary actual / 1
    groupby=(region prodtype)
    formats=('$', "$")
    filter="(NOT (REGION='WEST'))",

    /* 2 */  groupby=(region prodtype division)
    formats=('$', "$", "$")
    filter="(NOT (REGION='WEST')) AND (NOT (PRODUCT='SOFA'))"

    /* 3 */  groupby=(region prodtype division year)
    formats=('$", "$", "$", "f4."")
    filter="(NOT (REGION='WEST'))";
run;
```
Program Description

1. This MDSUMMARY statement uses the variable that is named actual. You can analyze more than one variable by adding the variable names before the slash.

2. Formats are enclosed in quotation marks and separated by commas. Numeric formats are also enclosed in quotation marks.

Output

Output 4.15  Results for Summarization of Actual by Region and Prodtype

<table>
<thead>
<tr>
<th>REGION</th>
<th>PRODTYPE</th>
<th>Column</th>
<th>Min</th>
<th>Max</th>
<th>N</th>
<th>Sum</th>
<th>Mean</th>
<th>Std Dev.</th>
<th>Coefficient of Variation</th>
<th>Number Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAST</td>
<td>FURNITURE</td>
<td>ACTUAL</td>
<td>15.0000</td>
<td>990.00</td>
<td>72</td>
<td>37065</td>
<td>528.33</td>
<td>290.32</td>
<td>55.0334</td>
<td>0</td>
</tr>
<tr>
<td>EAST</td>
<td>OFFICE</td>
<td>ACTUAL</td>
<td>13.0000</td>
<td>1000.00</td>
<td>216</td>
<td>109066</td>
<td>503.10</td>
<td>266.72</td>
<td>56.9812</td>
<td>0</td>
</tr>
</tbody>
</table>

Output 4.16  Results for Summarization of Actual by Region, Prodtype, and Division

<table>
<thead>
<tr>
<th>REGION</th>
<th>PRODTYPE</th>
<th>DIVISION</th>
<th>Column</th>
<th>Min</th>
<th>Max</th>
<th>N</th>
<th>Sum</th>
<th>Mean</th>
<th>Std Dev.</th>
<th>Coefficient of Variation</th>
<th>Number Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAST</td>
<td>FURNITURE</td>
<td>CONSUMER</td>
<td>ACTUAL</td>
<td>15.0000</td>
<td>990.00</td>
<td>72</td>
<td>37065</td>
<td>528.33</td>
<td>290.32</td>
<td>55.0334</td>
<td>0</td>
</tr>
<tr>
<td>EAST</td>
<td>FURNITURE</td>
<td>EDUCATION</td>
<td>ACTUAL</td>
<td>13.0000</td>
<td>1000.00</td>
<td>216</td>
<td>109066</td>
<td>503.10</td>
<td>266.72</td>
<td>56.9812</td>
<td>0</td>
</tr>
</tbody>
</table>

Output 4.17  Results for Summarization of Actual by Region, Prodtype, Division, and Year

<table>
<thead>
<tr>
<th>REGION</th>
<th>PRODTYPE</th>
<th>DIVISION</th>
<th>YEAR</th>
<th>Column</th>
<th>Min</th>
<th>Max</th>
<th>N</th>
<th>Sum</th>
<th>Mean</th>
<th>Std Dev.</th>
<th>Coefficient of Variation</th>
<th>Number Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAST</td>
<td>FURNITURE</td>
<td>CONSUMER</td>
<td>1993</td>
<td>ACTUAL</td>
<td>5.0000</td>
<td>990.00</td>
<td>72</td>
<td>37065</td>
<td>528.33</td>
<td>290.32</td>
<td>55.0334</td>
<td>0</td>
</tr>
<tr>
<td>EAST</td>
<td>FURNITURE</td>
<td>CONSUMER</td>
<td>1994</td>
<td>ACTUAL</td>
<td>63.0000</td>
<td>991.00</td>
<td>72</td>
<td>34918</td>
<td>404.97</td>
<td>272.17</td>
<td>56.9812</td>
<td>0</td>
</tr>
<tr>
<td>EAST</td>
<td>FURNITURE</td>
<td>EDUCATION</td>
<td>1993</td>
<td>ACTUAL</td>
<td>4.0000</td>
<td>999.00</td>
<td>72</td>
<td>39329</td>
<td>526.79</td>
<td>291.69</td>
<td>55.0334</td>
<td>0</td>
</tr>
</tbody>
</table>

Example 10: Fitting a Regression Model

Details

This IMSTAT procedure example demonstrates using higher-order polynomial models and also model selection.

Both REGCORR statements in the example request fitting a regression model of the response variable, Sales, and the regressor variable, Inventory. If no variables are specified, then the procedure fits a model for each pair of numeric non-CLASS variables in the table.
Program

```
libname example sasiola host="grid001.example.com" port=10010 tag='hps';

data example.shoes; set sashelp.shoes; run;

proc imstat data=example.shoes;
   regcorr sales inventory / order=2; 1
run;

   regcorr sales inventory / order=-3; 2
run;
quit;
```

Program Description

1. This REGCORR statement uses the ORDER=2 option to specify a quadratic model. ORDER=1 requests a linear model and ORDER=3 requests a cubic model.

2. Specifying -3 for the ORDER= option indicates that the procedure perform model selection from linear, quadratic, and cubic models. The procedure finds the best-fitting polynomial that is most appropriate, according to statistical principles.

Output

The following display shows the results for the two REGCORR statements. The first display shows the results for the quadratic model that was requested with ORDER=2. The second display shows the results for letting the procedure determine the best fit of the model. In this case, because the Quadr column has a nonzero value, that indicates the procedure determined that the quadratic model fits the data best. Had the Cubic column been nonzero, that would indicate that the procedure selected the cubic model as the most appropriate.

```
Linear Regression for Table WORK.SHOES

<table>
<thead>
<tr>
<th>Y</th>
<th>X</th>
<th>Intercept</th>
<th>Slope</th>
<th>Quadr</th>
<th>SS(Error)</th>
<th>SSTotal</th>
<th>R-Square</th>
<th>Correlation Coefficient</th>
<th>Number of Obs</th>
<th>Mean(X)</th>
<th>Mean(Y)</th>
<th>Std(X)</th>
<th>Std(Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>Inventory</td>
<td>1259.40</td>
<td>0.3257</td>
<td>1.6E-6</td>
<td>5.755E11</td>
<td>6.957E12</td>
<td>0.9124</td>
<td></td>
<td>396</td>
<td>250899</td>
<td>85700</td>
<td>351515</td>
<td>129107</td>
</tr>
</tbody>
</table>

Linear Regression Coefficient Table WORK.SHOES

<table>
<thead>
<tr>
<th>Y</th>
<th>X</th>
<th>Intercept</th>
<th>Slope</th>
<th>Cubic</th>
<th>SS (Error)</th>
<th>SSTotal</th>
<th>R-Square</th>
<th>Correlation Coefficient</th>
<th>Number of Obs</th>
<th>Mean(X)</th>
<th>Mean(Y)</th>
<th>Std(X)</th>
<th>Std(Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>Inventory</td>
<td>1259.40</td>
<td>0.3257</td>
<td>1.6E-6</td>
<td>5.755E11</td>
<td>6.957E12</td>
<td>0.9124</td>
<td></td>
<td>396</td>
<td>250899</td>
<td>85700</td>
<td>351515</td>
<td>129107</td>
</tr>
</tbody>
</table>
```
Chapter 5
IMSTAT Procedure (Data and Server Management)

Overview: IMSTAT Procedure (Data and Server Management) ............................................ 188
What Do the Data and Server Management Statements for the IMSTAT Procedure Do? .......... 188
Concepts: IMSTAT Procedure .................................................................................................. 188
Partitioned Tables ...................................................................................................................... 188
RUN-Group Processing .............................................................................................................. 189
WHERE Clause Processing ....................................................................................................... 189
Temporary Tables .................................................................................................................... 190
Creating Temporary Tables with GROUPBY Variables .......................................................... 192
Creating Temporary Variables ................................................................................................. 192
Syntax: IMSTAT Procedure (Data and Server Management) .................................................... 194
PROC IMSTAT (Data and Server Management) Statement ..................................................... 195
BALANCE Statement ............................................................................................................... 197
COLUMNINFO Statement ....................................................................................................... 197
COMPUTE Statement .............................................................................................................. 198
DELETEROWS Statement ........................................................................................................ 199
DISTRIBUTIONINFO Statement ............................................................................................. 200
DROPTABLE Statement .......................................................................................................... 200
FETCH Statement .................................................................................................................... 200
FREE Statement ..................................................................................................................... 203
IMPORTCUBE Statement ........................................................................................................ 204
LIFETIME Statement ............................................................................................................... 205
NUMROWS Statement ............................................................................................................ 205
PARTITION Statement ............................................................................................................ 206
PARTITIONINFO Statement .................................................................................................... 207
PROMOTE Statement .............................................................................................................. 208
PURGETEMPTABLES Statement .............................................................................................. 209
REPLAY Statement ................................................................................................................ 209
SAVE Statement ..................................................................................................................... 210
SCHEMA Statement ............................................................................................................... 210
SCORE Statement ................................................................................................................ 213
SERVERINFO Statement ....................................................................................................... 220
SERVERPARM Statement ....................................................................................................... 221
SERVERTERM Statement ....................................................................................................... 222
SERVERWAIT Statement ....................................................................................................... 223
SET Statement ....................................................................................................................... 223
STORE Statement .................................................................................................................. 224
TABLE Statement .................................................................................................................. 231
TABLEINFO Statement ....................................................................................................... 232
UPDATE Statement ............................................................................................................... 232
Overview: IMSTAT Procedure (Data and Server Management)

What Do the Data and Server Management Statements for the IMSTAT Procedure Do?

This portion of the IMSTAT procedure is used to manage in-memory tables and SAS LASR Analytic Server instances.

Concepts: IMSTAT Procedure

Partitioned Tables

A SAS LASR Analytic Server table can be partitioned, and it can also be ordered within each partition. A partition is a collection of observations that share the same key value and are located on the same worker node. The key value can be constructed from one or more variables. The partitioning keys are created according to the formatted values of the specified variables. Partitioning of data is an important tool in managing distributed data sources. The process of partitioning can consume computing and network resources as well as create greater imbalance in the data compared to a round-robin distribution. However, partitioned data can be accessed more quickly and subsequent procedure statements can execute more quickly.

You can achieve a partitioned table as follows:

- load a table with the SAS LASR Analytic Server engine using the PARTITION= data set option
- load a SASHDAT file that has been previously partitioned into HDFS by using the PARTITION= data set option of the SAS Data in HDFS engine
- re-partition data from one table into a temporary table (you can then make it a regular table with the PROMOTE statement)
- create a temporary table with a GROUPBY= option. The temporary table is partitioned by the formatted values of the GROUPBY= variables.

For more information, see “Data Partitioning and Ordering” on page 14.
**RUN-Group Processing**

The IMSTAT procedure supports RUN-group processing. RUN-group processing enables you to submit RUN groups without ending the procedure. This feature is particularly useful for running SAS interactively. You can start the procedure with the PROC IMSTAT statement and then execute statements like SUMMARY and FREQUENCY. Each statement runs when it reaches a RUN statement.

To use RUN-group processing, you start the procedure and then submit multiple RUN-groups. A RUN-group is a group of statements that contains at least one action statement and ends with a RUN statement. As long as you do not terminate the procedure, it remains active and you do not need to resubmit the PROC statement.

To end RUN-group processing, submit a RUN CANCEL statement. Statements that have not been submitted are terminated. To stop the procedure, submit a QUIT statement. Statements that have not been submitted are terminated as well.

**WHERE Clause Processing**

There are two important features for WHERE clause processing that are related to the IMSTAT procedure. The first is that the WHERE clause is applied to the data by the server. When you use a WHERE clause to subset data, the subsetting is performed by the server. Only the rows that meet the WHERE clause criteria are affected by subsequent operations.

The second important feature for WHERE clause processing is related to the RUN-group processing that the IMSTAT procedure supports. You can modify the WHERE clause between statements. Unless a WHERE clause is specified in a RUN block, no subsetting of rows occurs. In the following code example, the SUMMARY statement in the first RUN-group is not subject to a WHERE clause. The FREQUENCY statement in the second RUN-group applies only to observations for which Division='EDUCATION.'

```sas
proc imstat data=example.prdsale(tag=sashelp);
  summary actual predict / groupby=(region);
  run;
  where division='EDUCATION';
  frequency prodtype;
  run;
```

If you specify WHERE clauses in different RUN blocks, the clauses replace each other. A note is written to the SAS log to indicate the change. For example, the SUMMARY statement in the following code example applies to observations for which Division='CONSUMER.' The FREQUENCY statement applies to observations for which Region='EAST.'

```sas
proc imstat data=example.prdsale(tag=sashelp);
  where division='CONSUMER';
  summary actual predict / groupby=(region);
  run;
  where region='EAST';
  frequency prodtype;
  run;
```

When the FREQUENCY statement runs, the following line is added to the SAS log.

```
NOTE: WHERE clause has been replaced.
```
WHERE clauses can remain active across RUN statements. The following example is the same as the previous example, except that the second WHERE clause is not submitted.

```sas
proc imstat data=example.prdsale(tag=sashelp);
    where division='CONSUMER';
    summary actual predict / groupby=(region);
run;
    /* where region='EAST'; */
    frequency prodtype;
run;
```

In this case, the SAS log includes the following note.

```
NOTE: A WHERE clause remains active from a previous RUN block: '(division='CONSUMER')'.
```

You can clear a WHERE clause by submitting `WHERE;`.

Each time you access a different table with the TABLE statement, the WHERE clause is cleared. In following example, the second FREQUENCY statement is not restricted to observations for which Region='EAST' because the TABLE statement that accesses Prdsal2 clears the WHERE clause.

```sas
proc imstat;
    table example.prdsale(tag=sashelp);
    where region='EAST';
    frequency prodtype;
run;
    table example.prdsal2(tag=sashelp);
    frequency prodtype;
run;
```

The SAS log indicates that the WHERE clause is no longer applied.

```
NOTE: The WHERE statement is cleared when you open a LASR Analytic Server table with the TABLE statement.
```

### Temporary Tables

A temporary table is an in-memory table that contains the result set of a procedure statement. Instead of transferring the results to the client SAS session, the results remain in the server and only the name of the temporary table is transferred to the client. You can then use other procedure statements with the temporary table.

Temporary tables can be partitioned and the SUMMARY, CROSSTAB, DISTINCT, and PERCENTILE statements perform this action. For non-partitioned data, you can also generate temporary tables with the SUMMARY and CROSSTAB statements, provided that you request a GROUPBY analysis.

The following DATA step shows how to create a partitioned table on the variables Country and Region.

```sas
data lasr.prdsale(partition=({country region)});
    set sashelp.prdsale;
run;
```
The following statements generate a summary analysis for variables Actual and Predict in each of the partitions.

```sas
proc imstat;
  table lasr.prdsale;
  summary actual predict / partition;
run;
```

The output for the previous statements is as follows:

As an alternative, you can leave the result set in an in-memory table by adding the TEMPTABLE option to the SUMMARY statement:

```sas
summary actual predict / partition temptable;
run;
```

The previous SAS statements generate the following output in the SAS session.

```
<table>
<thead>
<tr>
<th>Country</th>
<th>Region</th>
<th>Column</th>
<th>Min</th>
<th>Max</th>
<th>N</th>
<th>Sum</th>
<th>Mean</th>
<th>Std Dev.</th>
<th>Std Error</th>
<th>Coefficient of Variation</th>
<th>Number Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
<td>East</td>
<td>Actual</td>
<td>6.0000</td>
<td>999.00</td>
<td>240</td>
<td>127485</td>
<td>531.19</td>
<td>828.61</td>
<td>18.1781</td>
<td>18.0.1781</td>
<td>63.0160</td>
</tr>
<tr>
<td>Canada</td>
<td>East</td>
<td>Predict</td>
<td>0</td>
<td>996.00</td>
<td>240</td>
<td>126546</td>
<td>502.69</td>
<td>284.29</td>
<td>18.3610</td>
<td>18.3610</td>
<td>66.6540</td>
</tr>
<tr>
<td>Canada</td>
<td>West</td>
<td>Actual</td>
<td>3.0000</td>
<td>1000.00</td>
<td>240</td>
<td>119505</td>
<td>497.94</td>
<td>296.53</td>
<td>19.1400</td>
<td>19.1400</td>
<td>69.6567</td>
</tr>
<tr>
<td>Canada</td>
<td>West</td>
<td>Predict</td>
<td>6.0000</td>
<td>1000.00</td>
<td>240</td>
<td>112373</td>
<td>468.22</td>
<td>275.99</td>
<td>17.8150</td>
<td>17.8150</td>
<td>68.9443</td>
</tr>
<tr>
<td>Germany</td>
<td>East</td>
<td>Actual</td>
<td>13.0000</td>
<td>1000.00</td>
<td>240</td>
<td>124547</td>
<td>510.95</td>
<td>297.71</td>
<td>18.5714</td>
<td>18.5714</td>
<td>65.4406</td>
</tr>
<tr>
<td>Germany</td>
<td>East</td>
<td>Predict</td>
<td>4.0000</td>
<td>983.00</td>
<td>240</td>
<td>117579</td>
<td>489.81</td>
<td>292.43</td>
<td>19.0752</td>
<td>19.0752</td>
<td>69.6601</td>
</tr>
<tr>
<td>Germany</td>
<td>West</td>
<td>Actual</td>
<td>3.0000</td>
<td>996.00</td>
<td>240</td>
<td>121451</td>
<td>506.05</td>
<td>289.17</td>
<td>18.6503</td>
<td>18.6503</td>
<td>57.1429</td>
</tr>
<tr>
<td>Germany</td>
<td>West</td>
<td>Predict</td>
<td>0</td>
<td>981.00</td>
<td>240</td>
<td>113975</td>
<td>474.00</td>
<td>280.49</td>
<td>18.1068</td>
<td>18.1068</td>
<td>69.0637</td>
</tr>
<tr>
<td>U.S.A.</td>
<td>East</td>
<td>Actual</td>
<td>4.0000</td>
<td>984.00</td>
<td>240</td>
<td>118229</td>
<td>492.62</td>
<td>282.26</td>
<td>18.2200</td>
<td>18.2200</td>
<td>67.2983</td>
</tr>
<tr>
<td>U.S.A.</td>
<td>East</td>
<td>Predict</td>
<td>1.0000</td>
<td>1000.00</td>
<td>240</td>
<td>122567</td>
<td>502.46</td>
<td>301.36</td>
<td>19.4518</td>
<td>19.4518</td>
<td>69.9757</td>
</tr>
<tr>
<td>U.S.A.</td>
<td>West</td>
<td>Actual</td>
<td>6.0000</td>
<td>984.00</td>
<td>240</td>
<td>119120</td>
<td>496.33</td>
<td>285.67</td>
<td>18.4401</td>
<td>18.4401</td>
<td>57.5567</td>
</tr>
<tr>
<td>U.S.A.</td>
<td>West</td>
<td>Predict</td>
<td>22.0000</td>
<td>999.00</td>
<td>240</td>
<td>121135</td>
<td>504.73</td>
<td>280.10</td>
<td>18.0800</td>
<td>18.0800</td>
<td>65.4544</td>
</tr>
</tbody>
</table>
```

As an alternative, you can leave the result set in an in-memory table by adding the TEMPTABLE option to the SUMMARY statement:

```sas
summary actual predict / partition temptable;
run;
```

The previous SAS statements generate the following output in the SAS session.

```
The SAS System
The IMSTAT Procedure

Temporary Table Information for Table WORK.PRDSALE
Statement SUMMARY
Temporary Table _T_960738b7_10f2b778
```

The temporary table is assigned a name by the server. When the IMSTAT procedure ends, any temporary tables created during the procedure run are removed from the server. Since the generated name is not predictable, the procedure assigns the name of the most recently generated temporary table to the _TEMPLAST_ macro variable.

You can use the TABLE statement to switch the active table to the temporary table and perform analyses. Make sure that the statement that generated the temporary table is separated from the next statement with a RUN statement. Otherwise, you receive an error that the table specified in the TABLE statement does not exist. The temporary table does not exist at parse time, it is created at run time when the statement is executed.
The following statements retrieve information about the temporary table, the formatted values for (up to) the first twenty rows, and perform a summarization:

```plaintext
table lasr.&_templatst_
    tableinfo;
    columninfo;
    fetch / from=1 to=20 format;
    summary;
quit;
```

The output for the TABLEINFO, COLUMNINFO, and FETCH statements is not shown. The results for the SUMMARY statement are as follows:

<table>
<thead>
<tr>
<th>Column</th>
<th>Min</th>
<th>Max</th>
<th>N</th>
<th>Sum</th>
<th>Mean</th>
<th>Std Dev.</th>
<th>Std Error</th>
<th>Coefficient of Variation</th>
<th>Number Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0</td>
<td>22.0000</td>
<td>12</td>
<td>67</td>
<td>5.5633</td>
<td>6.2298</td>
<td>1.7934</td>
<td>111.56</td>
<td>0</td>
</tr>
<tr>
<td>Max</td>
<td>981.00</td>
<td>1000.00</td>
<td>12</td>
<td>11922</td>
<td>593.50</td>
<td>7.5619</td>
<td>2.1829</td>
<td>0.7611</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td>240.00</td>
<td>240.00</td>
<td>12</td>
<td>2880</td>
<td>240.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NMissing</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Mean</td>
<td>498.22</td>
<td>531.19</td>
<td>12</td>
<td>5986.96967</td>
<td>498.03</td>
<td>17.0557</td>
<td>4.9247</td>
<td>3.4199</td>
<td>0</td>
</tr>
<tr>
<td>Sum</td>
<td>112373</td>
<td>127485</td>
<td>12</td>
<td>1436532</td>
<td>119219</td>
<td>4094.34</td>
<td>1181.93</td>
<td>3.4199</td>
<td>0</td>
</tr>
<tr>
<td>Std</td>
<td>275.99</td>
<td>301.36</td>
<td>12</td>
<td>3437.5553</td>
<td>298.47</td>
<td>7.3828</td>
<td>2.1313</td>
<td>2.5772</td>
<td>0</td>
</tr>
<tr>
<td>StdErr</td>
<td>17.0150</td>
<td>18.4516</td>
<td>12</td>
<td>221.95020</td>
<td>16.4513</td>
<td>0.4766</td>
<td>0.1376</td>
<td>2.5772</td>
<td>0</td>
</tr>
<tr>
<td>Var</td>
<td>76170</td>
<td>98089</td>
<td>12</td>
<td>98634.763</td>
<td>82113</td>
<td>4263.83</td>
<td>1230.86</td>
<td>6.1926</td>
<td>0</td>
</tr>
<tr>
<td>USS</td>
<td>70620.047</td>
<td>96672.749</td>
<td>12</td>
<td>96280.0466</td>
<td>79408.671</td>
<td>4419.32</td>
<td>12756.84</td>
<td>6.5651</td>
<td>0</td>
</tr>
<tr>
<td>CSS</td>
<td>18204657</td>
<td>21703393</td>
<td>12</td>
<td>236499786</td>
<td>19624982</td>
<td>10190.53</td>
<td>2941.75</td>
<td>5.1926</td>
<td>0</td>
</tr>
<tr>
<td>CV</td>
<td>53.0160</td>
<td>59.9577</td>
<td>12</td>
<td>889.72793</td>
<td>57.4773</td>
<td>2.1126</td>
<td>0.6098</td>
<td>3.6755</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>26.8303</td>
<td>27.2212</td>
<td>12</td>
<td>323.849820</td>
<td>26.9574</td>
<td>1.0171</td>
<td>0.2396</td>
<td>3.7687</td>
<td>0</td>
</tr>
<tr>
<td>PRT</td>
<td>7.49E-01</td>
<td>4.00E-01</td>
<td>12</td>
<td>7.4100E-01</td>
<td>6.11E-01</td>
<td>1.22E-01</td>
<td>3.52E-02</td>
<td>.</td>
<td>0</td>
</tr>
</tbody>
</table>

**Creating Temporary Tables with GROUPBY Variables**

Temporary tables play an important role in partitioning in the server. Temporary tables that are created by the DISTINCT, SUMMARY, CROSSTAB, or PARTITION statements are partitioned.

If the input table is not partitioned, you can still use temporary tables with the SUMMARY and CROSSTAB statements, provided that you perform a group-by analysis. The temporary table that is created by the server is automatically partitioned by the group-by variables. This potentially involves a redistribution of the groups in the result set to transfer all the result records for a particular group to the same worker node.

**Creating Temporary Variables**

You can use temporary variables in a table. For example, if a table has variables that are named x1, x2, and x3, you can calculate the summary statistics for variable d1 = x1 + x2 / 3 * x3. One way is to declare d1 as a temporary variable of the table (with data set options for the input table). You can use the temporary variables (temporary for the duration of each statement) with DATA step expression scripts.

```plaintext
class imstat data=lasr.lis.table1(array=(d,1));
    summary d1 x1 -- x3 / tempnames=d1 tempexpress="d1 = x1 + x2 / 3*x3;"
    run;
    summary d1 / tempnames=d1 tempexpress="d1 = mean(x1, x2);"
```
Because the temporary variable exists for the duration of the statement, its name can be reused in subsequent statements. The second SUMMARY statement uses the same name, d1, for the temporary variable but it has a different value.

You can also create temporary character variables. The following example creates a program that concatenates the first character of the Type variable and the first character of the Origin variable.

```sas
libname lasrlib sasiola host="hostname.example.com" port=10010 tag=hps;

data lasrlib.cars; set sashelp.cars; run;

data _null_
   file 'concat.sas';
   put "c1 = substr(type,1,1) || substr(origin,1,1);";
run;

filename fref 'concat.sas';

proc imstat;
   table lasrlib.cars(tempnames=(c1 $));  
   summary horsepower / groupby=c1 tn=(c1) te=fref;  
run;
   crosstab c1*type / tn=(c1) te=fref;
run;
```

1. The TEMPNAMES= data set option for the SAS LASR Analytic Server engine reserves the variable name, c1, for the temporary variable that does not exist. The $ indicates that it is a character variable.

2. The temporary variable name can be used in a variety of expressions. The TEMPNAMES= and TEMPEXPRESS= options must be specified in every statement that uses the temporary variable.
Syntax: IMSTAT Procedure (Data and Server Management)

PROC IMSTAT <options>;
   BALANCE <options>;
   COLUMNINFO </options>;
   COMPUTE column-name "SAS-statements" </ option>;
   DELETEROWS </ options>;
   DISTRIBUTIONINFO </ option>;
   DROPTABLE <libref.member-name>;
   FETCH <variable-list> </ options>;
   FREE resourceSpecification;
   IMPORTCUBE <options>;
   LIFETIME time-specification </ MODE= ABSOLUTE | LASTUSE >;
   NUMROWS;
   PARTITION variable-list </ options>;
   PARTITIONINFO </ options>;
   PROMOTE member-name </ options>;
   PURGETEMPTABLES <options>;
   REPLAY table-list;
   SAVE <options>;
   SCHEMA dim-specification1 <dim-specification2 ...> </ options>;
   SCORE CODE=file-reference <options>;
   SERVERINFO <option>;
   SERVERPARM <options>;
   SERVERTERM <options>;
   SERVERWAIT <options>;
   SET set-specification1 <set-specification2 ...> </ options>;
   STORE <table-name> [table-number]>
      (row-number | _ALL_ | _LAST_ | row-list | WHERE=(where-clause) <,>
         column-number | _ALL_ | COLS=column-list) = macro-variable-name </ options>;
   TABLE <libref.member-name>;
   TABLEINFO </ options>;
   UPDATE variable1=value1 <variable2=value2 ...> </options>;
QUIT;

<table>
<thead>
<tr>
<th>Statement</th>
<th>Task</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>BALANCE</td>
<td>Rebalancing a table</td>
<td>Ex. 3</td>
</tr>
<tr>
<td>DELETEROWS</td>
<td>Deleting rows and saving a table to HDFS</td>
<td>Ex. 4</td>
</tr>
<tr>
<td>PARTITION</td>
<td>Partitioning a table into a temporary table</td>
<td>Ex. 1</td>
</tr>
</tbody>
</table>
**PROC IMSTAT (Data and Server Management) Statement**
Manages in-memory tables in a SAS LASR Analytic Server instance.

**Syntax**

```
PROC IMSTAT <options>;
```

**Optional Arguments**

**BATCHMODE**
By default, the IMSTAT procedure operates in interactive mode. If your program contains errors that prevent SAS from parsing or executing statements, the errors are reported in the SAS log, but they do not stop the procedure. If the errors are fatal errors such as running out of memory on the SAS client, the procedure stops.

In contrast, when the BATCHMODE option is specified in the PROC IMSTAT statement, the procedure behaves with respect to error handling as if it were not an interactive procedure. Whenever an error occurs, the procedure terminates and sets the SYSERR macro variable.

Alias  **BATCH**

**DATA=** *libref.member-name*
specifies the table to access from memory. The libref must be assigned from a SAS LASR Analytic Server engine LIBNAME statement.

**FMTLIBXML=** *file-reference*
specifies the file reference for a format stream. For more information, see “FMTLIBXML” in the LASR procedure.

**IMMEDIATE**
specifies that the procedure executes one statement at a time rather than accumulating statements in RUN blocks.

Alias  **SINGLESTEP**

**NODATE**
specifies to suppress the display of time and date-dependent information in results from the TABLEINFO statement.

**NOPREPARSE**
prevents the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects
problems with the code, the error messages might not to be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

When you specify this option in the PROC IMSTAT statement, the option applies to all statements that can generate code. You can also exclude specific statements from pre-parsing by using the NOPREPARSE option in statements that allow temporary columns or the SCORE statement.

Alias NOPREP

NOPRINT
This option suppresses the generation of ODS tables and other printed output in the IMSTAT procedure. You can use this option to suppress printed output during execution of the actions, and then use the REPLAY statement to print the tables at a later point in the procedure execution.

NOTIMINGMSG
When an action completes successfully, the IMSTAT procedure generates a SAS log message that contains the execution time of the request. Specify this option to suppress the message.

Alias NOTIME

PGMMSG
specifies to capture messages associated with user-provided SAS statements on the server in a temporary table. Messages are produced when parsing errors occur, when code generation fails, or by PUT statements in a SAS program.

You can use this option as a debugging feature for SAS code that you submit through temporary column expressions. The macro variable _PGMMSG_ is used in the IMSTAT procedure to capture the name of the table. The _TEMPLAST_ macro variable is also updated in case this temporary table is the most recently created temporary table.

Alias PROGMST

SIGNER=authorization-web-service-uri
specifies the URI for the SAS LASR Authorization web service. For more information, see SAS Visual Analytics: Administration Guide.

Example SIGNER="https://server.example.com/SASLASRAuthorization"

TEMPTABLEINFO
specifies to add additional information for temporary tables to the ODS table that is created on the SAS client. The information includes the time at which the temporary table was created in the server, the number of rows, and the number of columns.

Alias TEMPINFO

UCA
specifies that you want to use Unicode Collation Algorithms (UCA) to determine the ordering of character variables in the GROUPBY= operations and other operations that depend on the order of formatted values.
BALANCE Statement

The BALANCE statement creates a temporary table from the active table and re-balances it so that the number of rows on the worker nodes are balanced as evenly as possible. The rows are balanced within ± 1 row of each other.

Example: “Example 3: Rebalancing a Table” on page 238

Syntax

BALANCE;

Without Arguments

The re-balancing removes any observations marked as deleted or marked for purging in the active table. A WHERE clause is observed when the data are rebalanced.

One case for re-balancing is if the data distribution for a table has become uneven due to block movement within the Hadoop Distributed File System. This can occur when nodes fail in Hadoop or Hadoop processes have exited on some nodes. Another situation where re-balancing is useful is when a partitioned table has uneven distribution across the worker nodes due to uneven sizes of the partition. This can affect the performance of all actions running in the LASR Analytic Server since typically the nodes with the most records determine the overall performance.

Rebalancing of a table removes partition and ordering information from the table.

The BALANCE statement can be used with non-distributed servers as well. However, it is less important because all records of a table reside on the same machine. It might be useful, however, to derive from a partitioned table a new table subject to a WHERE clause and that has deleted records removed and is not partitioned.

COLUMNINFO Statement

The COLUMNINFO statement is used to return information for all the columns in an in-memory table.

Syntax

COLUMNINFO <options>;

COLUMNINFO Statement Options

SAVE=table-name

saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

SETSIZE

requests that the server estimate the size of the result set. The procedure does not create a result table if the SETSIZE option is specified. Instead, the procedure reports
the number of rows that are returned by the request and the expected memory consumption for the result set (in KB). If you specify the SETSIZE option, the SAS log includes the number of observations and the estimated result set size. See the following log sample:

NOTE: The LASR Analytic Server action request for the STATEMENT statement would return 17 rows and approximately 3.641 kBytes of data.

The typical use of the SETSIZE option is to get an estimate of the size of the result set in situations where you are unsure whether the SAS session can handle a large result set. Be aware that in order to determine the size of the result set, the server has to perform the work as if you were receiving the actual result set. Requesting the estimated size of the result set does consume resources on the server. The estimated number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.

### COMPUTE Statement

The COMPUTE statement adds a permanent computed column to an in-memory table.

#### Syntax

```
COMPUTE column-name file-reference option;
COMPUTE column-name "SAS-statements" option;
```

#### Required Argument

- **column-name**
  - specifies the name to use for the computed column. The name cannot already be in use in the table.

#### COMPUTE Statement Options

**NOPREPARSE**

prevents the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects problems with the code, the error messages might not be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

When you specify this option in the PROC IMSTAT statement, the option applies to all statements that can generate code. You can also exclude specific statements from pre-parsing by using the NOPREPARSE option in statements that allow temporary columns or the SCORE statement.

Alias **NOPREP**
DELETEROWS Statement

The DELETEROWS statement is used to mark rows as deleted, to undelete rows, and to purge rows from an in-memory table. Rows that are marked for deletion or purging are not included in the calculations performed by the server.

Interaction: If a WHERE clause is active, the rows that meet the criteria are marked for deletion or purging. When no WHERE clause is active, a delete request marks all rows for deletion (they can be undeleted), but the PURGE option removes the rows that are already marked for deletion rather than removing all rows.

Example: “Example 4: Deleting Rows and Saving a Table to HDFS” on page 240

Syntax

\[ \text{DELETEROWS } <\text{options}>; \]

DELETEROWS Statement Options

PURGE

specifies to remove from memory the rows that are marked for deletion. The memory that was used by the rows is freed. The purged rows cannot be undeleted. One use case for purging rows is to remove older records from an in-memory table after new records were appended to the table.

If a WHERE clause is active, the rows that meet the criteria are purged. If no WHERE clause is active, then the PURGE request removes the rows that are already marked for deletion. It does not remove all the rows in the table. This was implemented to prevent accidentally removing all rows from a table.

SAVE=\text{table-name}

saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for \text{table-name} must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

\text{TEMPEXPRESS}="\text{SAS-expressions}"

\text{TEMPEXPRESS}=\text{file-reference}

specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias TE=

\text{TEMPNAMES}.vars\text{variable-name}

\text{TEMPNAMES}=(\text{variable-list})

specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the \text{TEMPEXPRESS} option.

Alias TN=

UNDELETE

specifies to clear the deletion mark from rows.
If a WHERE clause is active, only the rows that meet the criteria have the deletion mark cleared. A row that has been marked for purging from the table cannot be undeleted.

**DISTRIBUTIONINFO Statement**

The DISTRIBUTIONINFO statement returns the number of partitions and the number of records on each worker node in the SAS LASR Analytic Server instance. This information provides the approximate distribution characteristics of the data across the worker nodes. If you want more details about the data distribution, then use the PARTITIONINFO statement.

**Syntax**

```
DISTRIBUTIONINFO <option>;
```

**DISTRIBUTIONINFO Statement Options**

- **SAVE=table-name**
  saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for `table-name` must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

**DROPTABLE Statement**

The DROPTABLE statement is used to remove an in-memory table from a server. You must specify a table as an active table with the DATA= procedure option or in a TABLE statement before you can use the DROPTABLE statement. Once a table is active, you can specify that table, another table, or a temporary table.

**Syntax**

```
DROPTABLE <libref.member-name>;
```

**Optional Argument**

- **libref.member-name**
  specifies the name of the in-memory table to remove from the server. If you do not specify the table, then the active table is dropped.

**FETCH Statement**

The FETCH statement is used to retrieve rows from an in-memory table. You can use the FETCH statement to retrieve calculated columns that are calculated according to a script as part of the request. The columns that are calculated this way do not persist beyond the time it takes to execute in the server.
Syntax

FETCH <variable-list> </ options>;

Optional Argument

variable-list
specifies the numeric and character variables to retrieve.

FETCH Statement Options

ARRAYSTART=n
specifies the starting element of an array when the record of an in-memory table represents a variable array. This is the case, for example, when a pricing cube from SAS High-Performance Risk is loaded into a server. There might then be 10,000 columns for a variable. Specifying the ARRAYSTART= and ARRAYLENGTH= options enables you to page through the data more conveniently.

ARRAYLENGTH=k
specifies the length of the array to fetched when the record of an in-memory table represents a variable array. Use this option with the ARRAYSTART= option.

DESCENDING=variable-name
DESCENDING=(variable-list)
specifies which variables of the ORDERBY= list are used with descending sort order. Specifying the DESCENDING= option by itself has no effect. The option is specified in addition to the ORDERBY= option. The following example specifies to fetch data on columns A and B of the active table ordered by ascending unformatted values of B and descending unformatted values of C.

Alias DESC=

Example fetch a b / orderby=(b c) descending=c;

FORMAT <=("format-specification", ... )>
specifies the formats to use for the variables. By default, the FETCH statement retrieves the unformatted values. If you specify the FORMAT option without a list of format names, then the server applies the default format that is associated with each variable.

Be aware that when you retrieve unformatted values and you create an output data set with the OUT= option, the variable information such as formats and labels are transferred to the output data set.

If you want to use the default format for a variable, specify an empty string. The following example uses the default format for column A and the $10 format for column B.

Alias FORMATS=

Example fetch a b / format=("", "$10");

FROM=first-index
specifies the index of the first row to retrieve (inclusive). The value for first-index is 1-based.

Default FROM=1
The value for FROM= is applied after the evaluation of a WHERE clause.

**NOPREPARSE**

prevents the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects problems with the code, the error messages might not be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

When you specify this option in the PROC IMSTAT statement, the option applies to all statements that can generate code. You can also exclude specific statements from pre-parsing by using the NOPREPARSE option in statements that allow temporary columns or the SCORE statement.

Alias **NOPREP**

**ORDERBY=variable-name**

**ORDERBY=(variable-list)**

specifies one or more variables by which to order the results. The default sort order of the ORDERBY= variables is ascending in unformatted values and follows location and collation rules. If you want to arrange some ORDERBY= variables in descending sort order, then list the variable names in the DESCENDING= option (in addition to listing them in the ORDERBY= option).

If you want to assign a format to ORDERBY= variables, you can use the ORDERBYFORMAT= option. That option can also be used to specify which variables are sorted by formatted values and which variables are sorted by unformatted values.

**ORDERBYFORMAT=("format-specification", ...)**

specifies the formats to use for sorting of the ORDERBY= variables. By default, if you specify an ORDERBY= variable or variable list, the server sorts by the ascending unformatted values. If you want to apply unformatted value ordering for some ORDERBY= variables, and formatted value ordering for other ORDERBY= variables, you can specify the keyword _RAW_ for the variables to sort by unformatted value.

The following example specifies to retrieve unformatted values of columns Make, Model, Type, Invoice, and Mpg_City. The rows are retrieved in the order of ascending formatted value of Type, using the $ format, and descending unformatted values of Invoice.

```
Example fetch make model type invoice mpg_city /
   orderby=(type invoice)
   desc invoice
   orderbyformat=($", 
      _RAW_"));
```

**OUT=libref.member-name**

specifies the name of the data set to store the result set of the FETCH statement.
SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

SETSIZE
requests that the server estimate the size of the result set. The procedure does not create a result table if the SETSIZE option is specified. Instead, the procedure reports the number of rows that are returned by the request and the expected memory consumption for the result set (in KB). If you specify the SETSIZE option, the SAS log includes the number of observations and the estimated result set size. See the following log sample:

NOTE: The LASR Analytic Server action request for the STATEMENT statement would return 17 rows and approximately 3.641 kBytes of data.

The typical use of the SETSIZE option is to get an estimate of the size of the result set in situations where you are unsure whether the SAS session can handle a large result set. Be aware that in order to determine the size of the result set, the server has to perform the work as if you were receiving the actual result set. Requesting the estimated size of the result set does consume resources on the server. The estimated number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.

TEMPEXPRESS="SAS-expressions"

TEMPEXPRESS=file-reference
specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias

TE=

TEMPNAMES=variable-name
TEMPNAMES=(variable-list)
specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

Alias

TN=

TO=last-index
specifies the index of the last row to retrieve (inclusive). The value for last-index is 1-based.

Default
The default value is FROM=first-index + 19.

Interaction
The value for TO= is applied after the evaluation of a WHERE clause.

FREE Statement
The FREE statement is used to release resources from STORE statements or SAVE= options. If you save a result table (one table or sets of tables) with the SAVE= option in any of the analytic statements of the IMSTAT procedure, then you can release the resources with the FREE statement. Once a table has been
freed, you can reuse the table name. While a saved table exists, you cannot create a table by the same name.

---

**Syntax**

```plaintext
FREE table-list;
FREE _ALL_;
FREE TABLE=one-table;
FREE MACRO=macro-variable-name;
```

**FREE Statement Options**

- **table-list**: specifies a list of tables to release.
- **_ALL_**: specifies to release the resources for all the tables that were saved throughout the procedure execution.
- **TABLE=one-table**: specifies the name of the table to release.
- **MACRO=macro-variable-name**: specifies the name of a macro variable to release.

---

**IMPORTCUBE Statement**

The IMPORTCUBE statement is used to read a cube descriptor file for a SAS High-Performance Risk pricing cube. The server converts the pricing information that is in the cube into an in-memory table.

---

**Syntax**

```plaintext
IMPORTCUBE <options>;
```

**IMPORTCUBE Statement Options**

- **CUBE="path-to-descriptor-file"**: specifies the SAS High-Performance Risk descriptor file for the pricing cube. The procedure opens the descriptor file and examines the contents. If the file is not a cube descriptor file for a SAS High-Performance Risk cube or the version of the file does not match the IMSTAT procedure version, the procedure writes an error to the SAS log and stops processing statements.
  
  **Alias**  
  RISKCUBE=

- **HOST="host-name"**: specifies the machine to use for exporting the SAS High-Performance Risk cube. If you do not specify the host, the procedure examines the GRIDHOST environment variable. You can use the PORT= and HOST= options to specify the server to use. The procedure does not determine the host and port information for the IMPORTCUBE statement from the currently active table.
PORT=number
specifies the port on which the server listens for connections. You can combine the
PORT= and HOST= options to specify the server to use. If you do not specify the
PORT= option the procedure uses the port number that is stored in the LASRPORT
macro variable. The procedure does not determine the host and port information for
the IMPORTCUBE statement from the currently active table.

TABLENAME="table-name"
 specifies the name of the in-memory table to use for the imported cube. A table by
the same name must not already exist in the server.

Alias NAME=

LIFETIME Statement
The LIFETIME statement enables you to associate an expiration date with an in-memory table. You must
have sufficient authorization, which is equivalent to the authorization to drop the table. The server checks
periodically if any tables have expired and drops the expired ones. This frees all resources associated with
those tables.

Syntax
LIFETIME time-specification <| MODE= ABSOLUTE | LASTUSE >;

Required Argument
time-specification
specifies the duration (in seconds) that the table is to remain in memory. The
minimum value is 1 second.

LIFETIME Statement Options
MODE=ABSOLUTE | LASTUSE
specifies how to use the time-specification. If MODE=ABSOLUTE is specified, then
the server drops the table after the specified number of seconds. If
MODE=LASTUSE is specified, then the server drops the table the specified number
of seconds after the last successful access to the table.

Default ABSOLUTE

NUMROWS Statement
The NUMROWS statement identifies how many rows satisfy a selection criterion. The selection observes
the WHERE clause and records marked for deletion or purging.

Syntax
NUMROWS;
PARTITION Statement

The PARTITION statement is used to generate a temporary table from the active table and partition it according to the statement options. This re-partitioning of tables is supported for distributed and non-distributed SAS LASR Analytic Server.

**Examples:**
- “Example 1: Partitioning a Table into a Temporary Table” on page 235
- “Example 7: Appending a Non-Partitioned Table to a Partitioned Table” on page 244

**Syntax**

`PARTITION variable-list <options>;`

**PARTITION Statement Options**

**DESCENDING=(variable-list)**

`DESCENDING=variable-name`

specifies the variables in the ORDERBY= list to use with a descending sort order. Specifying the DESCENDING= option by itself has no effect on ordering within a partition. The option is specified in addition to the ORDERBY= option.

**Alias**

DESC=

**Example**

The following statement requests partitioning of the active table by variables A and B, and ordering within the partition by ascending value of column C and descending value of column D:

`partition a b / orderby=(c d) descending=d;`

**FORMATS=("format-specification"<,...>)**

specifies the formats for the PARTITIONBY= variables. If you do not specify the FORMATS= option, the default format is applied for that variable. The format of a partitioning variable is important because the equality of the partition keys is determined from the formatted values.

Enclose each format specification in quotation marks and separate each format specification with a comma.

**NOMISSING**

specifies that missing values are excluded in the determination of partition keys. By default, observations with missing values are included.

**Alias**

NOMISS

**ORDERBY=(variable-list)**

`ORDERBY=variable-name`

specifies the variable or variables to use for ordering the observations within a partition. By default, the sort order for the ORDERBY= variables is ascending in raw values and follows location and collation rules. If you want to order some ORDERBY= variables in descending sort order, then specify the variable names in the DESCENDING= option (in addition to listing them in the ORDERBY= option).

The ORDERBY= variables are transferred automatically to the partitioned temporary table, whether you list them in the VARS= option or not.
SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

TEMPEXPRESS="SAS-expressions"
TEMPEXPRESS=file-reference
specifies either a quoted string that contains the SAS expression that defines the temporary variables or a file reference to an external file with the SAS statements.

Alias TE=

TEMPNAMES=variable-name
TEMPNAMES=(variable-list)
specifies the list of temporary variables for the request. Each temporary variable must be defined through SAS statements that you supply with the TEMPEXPRESS= option.

Alias TN=

VARIABLES=(variable-list)
VARIABLES=variable-name
specifies the variable or variables to include in the temporary table in addition to the partitioning variables. If you do not specify the VARS= option, then all the variables are transferred from the active table. Temporary calculated columns are also transferred to the temporary table.

Alias VARS=

PARTITIONINFO Statement
The PARTITIONINFO statement produces very detailed information about the partitions of the data for each node of the server. Be aware that depending on the number of partitions, the result table can be very large. If you use the statement on non-partitioned data, it reports the number of bytes and records for each node. These results are similar to the result set you get from a DISTRIBUTIONINFO statement.

Syntax
PARTITIONINFO <options>;

PARTITIONINFO Statement Options
SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

SETSIZE
requests that the server estimate the size of the result set. The procedure does not create a result table if the SETSIZE option is specified. Instead, the procedure reports the number of rows that are returned by the request and the expected memory consumption for the result set (in KB). If you specify the SETSIZE option, the SAS
The log includes the number of observations and the estimated result set size. See the following log sample:

NOTE: The LASR Analytic Server action request for the STATEMENT statement would return 17 rows and approximately 3,641 kBytes of data.

The typical use of the SETSIZE option is to get an estimate of the size of the result set in situations where you are unsure whether the SAS session can handle a large result set. Be aware that in order to determine the size of the result set, the server has to perform the work as if you were receiving the actual result set. Requesting the estimated size of the result set does consume resources on the server. The estimated number of KB is very close to the actual memory consumption of the result set. It might not be immediately obvious how this size relates to the displayed table, since many tables contain hidden columns. In addition, some elements of the result set might not be converted to tabular output by the procedure.

**PROMOTE Statement**

The PROMOTE statement is used to change a temporary table to a regular in-memory table. The currently active table must be a temporary table and the table identified with the `member-name` parameter must not already exist in the server. Promoting a temporary table requires authorization to add a table to the server. You can also specify a tag for the in-memory table with the `TAG=` option.

**Example:**  
“Example 5: Creating a Star Schema” on page 241

**Syntax**

```plaintext
PROMOTE member-name <options>;
```

**Required Argument**

`member-name`

specifies the name to use for the table that is promoted.

**PROMOTE Statement Options**

`PERM=mode`

specifies the access permission for the newly created table as an integer. The mode value is specified as an integer value such as 755. The mode corresponds to the mode values that are used for UNIX file access permissions. You must specify a value that preserves Read and Write permission for your user ID.

- **Alias** `PERMISSION=`
- **Range** 600 to 777

`TAG='server-tag'`

specifies the tag to use for naming the table. If no `TAG=` option is specified, then the `TAG=` option from the LIBNAME statement is used. If the LIBNAME statement does not specify the `TAG=` option, then the name of the libref is used as the server tag.
**PURGETEMP TABLES Statement**

The PURGETEMP TABLES removes all temporary tables from a server. The action requires server-level authorization because it removes temporary tables created by all users. To execute this command successfully, you must have the same authorization that is required to terminate the server.

**Syntax**

```
PURGETEMP TABLES <options>;
```

**Without Arguments**

The server to use is identified from the active table. If you do not have an active table, then you can connect to a specific server with the HOST= and PORT= options.

**PURGETEMP TABLES Statement Options**

- **HOST=** "host-name"
  - specifies the host name for the server. Use this option with the PORT= option.
- **PORT=** number
  - specifies the port number for the server.

---

**REPLAY Statement**

The REPLAY statement enables you to display a previously saved result table or set of tables. The REPLAY statement displays the saved result tables regardless of the NOPRINT option. This enables you to suppress output generation with the NOPRINT option and to then display the tables that you want in a different order.

**Syntax**

```
REPLAY table-list;
```

**Optional Argument**

```
table-list
```

- specifies the saved result tables to display.

**Example: Display Result Tables in a Different Order**

The following SAS statements suppress the display of output with the NOPRINT option. Then, the tables are displayed in the reverse order.

```
proc imstat data=sales.prdsale noprint;
  fetch country region actual / save=salestab from=1 to=5;
  fetch predict actual / save=predicttab from=1 to=10;
  replay predicttab salestab;
quit;
```
SAVE Statement

The SAVE statement enables you to save an in-memory table to HDFS. The table must be the active table. You specify the active table with the DATA= option in the IMSTAT procedure or with the TABLE statement.

Example:  “Example 4: Deleting Rows and Saving a Table to HDFS” on page 240

Syntax

SAVE <options>;

Optional Arguments

BLOCKSIZE
specifies the block size to use for distributing the data set. Suffix values are B (bytes), K (kilobytes), M (megabytes), and G (gigabytes).

Interaction  If the in-memory table is partitioned, the BLOCKSIZE= specification is ignored. The server determines the block size based on the size of the partitions.

COPIES=n
specifies the number of replications to make for the data set (beyond the original blocks). The default value is 1. You can specify COPIES=0 if you do not need replications for redundancy.

FULLPATH
specifies that the value for the PATH= option specifies the full path for the file, including the filename (without the SASHDAT extension). The filename portion of the quoted string is expected to be in lowercase characters.

PATH='HDFS-path'
specifies the directory in HDFS in which to store the table as a SASHDAT file. The value is case sensitive. The filename for the SASHDAT file that is stored in the path is always lowercase.

Note:  If the PATH= option is not specified, the server attempts to save the table in the /user/userid directory. The userid is the user ID that started the server instance.

REPLACE
specifies that the SASHDAT file should be overwritten if it already exists.

SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

SCHEMA Statement

The SCHEMA statement is used to define a simple star schema in the server from a single fact table and one or more dimension tables. The result of the SCHEMA statement is a temporary table that you can use as it is, or with the PROMOTE statement.
Syntax

SCHEMA dim-specification1 <dim-specification2 ...> </options>;

Required Argument

dim-specification

specifies how to use the dimension table with the fact table. You must specify the variables to use as keys for the fact table (fact-key) and the dimension table (dim-key). The variables do not need to have the same name, but they do need to have the same type.

dim-table-name (fact-key = dim-key < PREFIX = dim-prefix >
< TAG = 'server-tag' >, variable-list >)

dim-table-name

specifies the name of the dimension table.

fact-key

specifies the variable name in the fact table to use.

dim-key

specifies the variable name from the dimension table to use.

PREFIX=dim-prefix

specifies a prefix to use for naming variables in the schema. If you do not specify PREFIX=, then up to the first sixteen characters of the dim-table-name are used as the dimension prefix for naming the variables in the schema.

Alias

NAME=

TAG='server-tag'

specifies the server tag to use for identifying the dimension table.

variable-list

specifies the variables from the dimension table to join with the fact table. By default, all variables except the dimension key are transferred from the dimension table. The dimension key is never transferred because a corresponding value is available through the fact-key.

SCHEMA Statement Options

MODE=VIEW | TABLE

specifies whether the rows of the schema are materialized when the statement executes in the server. The default is MODE=VIEW and implies that the server resolves the relations in the tables but defers the resolution (formation) of the rows until the view is accessed. If you specify MODE=TABLE, then the table is resolved (flattened) when the statement executes. A view consumes much fewer resources (almost none), but data access is slower compared to a flattened table.

Default VIEW

SAVE=table-name

saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.
TEMPEXPRESS="SAS-expressions"
TEMPEXPRESS=file-reference

specifies either a quoted string that contains the SAS expression that defines the
temporary variables or a file reference to an external file with the SAS statements.

Alias TE=

TEMPNAMES=variable-name
TEMPNAMES=variable-list

specifies the list of temporary variables for the request. Each temporary variable
must be defined through SAS statements that you supply with the TEMPEXPRESS=
option.

Alias TN=

Details

Assumptions

The server makes the following assumptions with regard to fact and dimension tables:

- Dimension tables do not have repeat values for the dimension keys. If a key value
  appears multiple times, the first record that matches the key is used.
- The relation between the fact table and a dimension table is expressed by one pair of
  keys. That is, one variable in the fact table defines the relation to one variable in the
dimension table.
- The variable names of keys in the fact table and dimension tables do not have to be
  the same.
- The lookups for the keys are performed based on raw values with fuzzing for
doubles.
- If the fact table or a dimension table contains pre-levelized CLASS variables, the
  class-variable signature is removed when the schema is created.
- Partitioning and order-by information is preserved when the schema is created.
  However, only partitioning of the fact table is taken into consideration and the
  resulting table or view is partitioned by the same variables as the fact table.
- The relations are resolved when the schema is created. This strategy makes passes
  through the data more efficient.

About Views and Tables

When the SCHEMA statement executes, a temporary table is generated and the name of
that temporary table is returned to the client as the result set. You use the
&_TEMPLAST_ macro variable to refer to the star schema.

By default, the server creates a view from the schema definition. The temporary table
then has columns for all variables in the schema. The relations have been resolved, but
the rows of the view have not been formed.

You can request that row resolution takes place when the temporary table is formed. The
result is a flattened temporary table where the rows of the schema are materialized.

There are advantages and disadvantages to using views and flattened tables. The primary
consideration is whether there is enough memory for the data volume. The following list
identifies some of the considerations:
• A flattened table consumes memory when the statement executes. If the memory requirement of the fully joined schema exceeds the capacity of the machine, the statement fails. For example, if you intend to work with relations that expand to seven terabytes of memory, then you cannot flatten the table unless you have that much memory on your system.

• If a flattened table can be held in memory, data access is faster because it is a regular in-memory table.

• A view does not consume memory until it is accessed. At that time, the table is never materialized fully in memory. The joined rows are formed only when a buffer is needed. This enables you to work with views that exceed the memory capacity of the system.

• The performance difference between resolving a view at run time and accessing a flattened table is difficult to quantify. It depends, for example, on the number of columns to resolve and the data access pattern. A request that passes through the data multiple times suffers a greater performance hit (compared to a flat table) than a single-pass request.

Some operations are not supported with views (but are supported with materialized schemas):

• You cannot append tables or rows to a view.

• You cannot perform row updates of the view.

• You cannot re-partition a view.

• You cannot use a view in another schema.

If a view is based on a partitioned fact table and you want to change the partition key, then re-partition the fact table and re-create the view with another SCHEMA statement.

A view is static. For example, if you append rows to the fact table, the append operation succeeds and every new access to the fact table can use the appended rows. However, the view is not affected by the addition of rows to the fact table. The view resolves to the state of the fact table when the view was formed.

If you want a schema to change with appends and updates, then you can materialize it and then append or update the flattened table. Likewise, you can append or update the fact table and dimension tables, drop the view, and re-create it.

Using a view as the fact table or as a dimension table in a SCHEMA statement is not supported.

---

**SCORE Statement**

The SCORE statement applies scoring rules to an in-memory table. The results can take different forms. They can be displayed as tables in the SAS session, output data sets on the client, or temporary tables in the server. The most common use of the SCORE statement is to execute DATA step code on some or all rows of an in-memory table and to produce corresponding output records.

**Syntax**

```sas
SCORE CODE=file-reference <options>;
```
**Required Argument**

**CODE=**file-reference

specifies a file reference to the SAS program that performs the scoring.

Alias  PGM=

**SCORE Statement Options**

**DROP=**(variable-list)

**DROP=**variable-name

specifies one or more variables that you want to drop from the input data when transferring results to the scoring results. By default, the SCORE statement does not transfer any variables from the input table to the scoring results. If you specify the **KEEP=** option, only the specified variables are moved to the result. If you specify the **DROP=** option, all variables except the specified variables are moved to the result set.

Interaction  Do not use the **DROP=** and **KEEP=** options together. The **KEEP=** option takes precedence.

**DSRETAIN**

requests that the scoring code implements DATA step retention behavior for output symbols. By default, the server automatically retains output variables, whereas the DATA step sets variables to missing unless they are retained explicitly with a RETAIN statement. Specifying this option means that the automatic retention behavior is disabled in the server and you must specify RETAIN statements in your DATA step program to retain values.

Alias  NOAUTORETAIN

**HASHDATA**(table-name1 <, table-name2...>)

specifies the names of one or more in-memory tables in server that are used as input tables for DATA step hash objects. The names are the actual table names in the server, not names where tags are masked by the libref. The names must match the names that you use in the DECLARE HASH statements in the scoring program.

The server checks your authorization to access the secondary tables. The scoring action fails with a permission error if you are not authorized to use any one of the tables. The scoring action also fails if the names that you specify in the **HASHDATA** option do not match the names that are used in the scoring program.

Alias  HASH

Example  “Example: Joining Data during Scoring with a DATA Step Hash Object” on page 218

**KEEP=**(variable-list)

**KEEP=**variable-name

specifies one or more variables that you want to transfer from the input data to the scoring results. You can use _ALL_ for all variables, _NUMERIC_ for all numeric variables, and other valid variable list names. If this option is not specified, then no variables are transferred from the input data (the table that is being specified), unless they are assigned values in the scoring code.

Alias  TABLEVARS=
NOPREPARSE
specifies to prevent pre-parsing and pre-generating the program code that is referenced in the CODE= option. If you know the code is correct, you can specify this option to save resources. The code is always parsed by the server, but you might get more detailed error messages when the procedure parses the code rather than the server. The server assumes that the code is correct. If the code fails to compile, the server indicates that it could not parse the code, but not where the error occurred.

Alias  NOPREP

OUT=libref.member-name
specifies the name of an output data set in which to store the scoring results. If the result set contains variables that match those in the input data set, then format information is transferred to the output data set. The OUT= option and the TEMPTABLE option are mutually exclusive. If you specify the OUT= option, a temporary table is not created in the server.

Alias  DATA=

PARTITION <=partition-key>
specifies to take advantage of partitioning for tables that are partitioned. When this option is specified, the scoring code is executed in the order of the partitions. If the data are also ordered within the partition, the observations are processed in that order. If the scoring code uses the reserved symbols __first_in_partition or __last_in_partition, then the data are also processed in partitioned order. Although the observations are processed in a specific order, the execution occurs in concurrent threads (in parallel). Different threads are assigned to work on different partitions.

If you do not specify the optional partition-key, then the analysis is performed for all partitions. If you do specify a partition-key, then the analysis is performed for the partitions that match the specified key value only. You can use the PARTITIONINFO statement to retrieve the valid partition-key values for a table.

You can specify a partition-key in two ways. You can supply a single quoted string that is passed to the server, or you can specify the elements of a composite key separated by commas. For example, if you partition a table by variables GENDER and AGE, with formats $1 and BEST12, respectively, then the composite partition key has a length of 13. You can specify the partition for the 11 year-old females as follows:

score / partition="F 11"; /* passed directly to the server */
score / partition="F","11";      /* composed by the procedure */

If you choose the second format, the procedure composes a key based on formatting information from the server.

Alias  PART=

Interaction  This option is effective when used with partitioned in-memory tables only.

SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.
SETSIZE
requests that the server estimate the size of the result set. The procedure does not
create a result table if the SETSIZE option is specified. Instead, the procedure reports
the number of rows that are returned by the request and the expected memory
consumption for the result set (in KB). If you specify the SETSIZE option, the SAS
log includes the number of observations and the estimated result set size. See the
following log sample:

NOTE: The LASR Analytic Server action request for the STATEMENT
statement would return 17 rows and approximately
3.641 kBytes of data.

The typical use of the SETSIZE option is to get an estimate of the size of the result
set in situations where you are unsure whether the SAS session can handle a large
result set. Be aware that in order to determine the size of the result set, the server has
to perform the work as if you were receiving the actual result set. Requesting the
estimated size of the result set does consume resources on the server. The estimated
number of KB is very close to the actual memory consumption of the result set. It
might not be immediately obvious how this size relates to the displayed table, since
many tables contain hidden columns. In addition, some elements of the result set
might not be converted to tabular output by the procedure.

SYMBOLS=(symbol-list)
specifies one or more symbols that are calculated in the scoring code that you want
to transfer as columns to the scoring results. If the SYMBOLS= option is not
specified, then all symbols that are assigned values in the program—and that are not
just placeholders for intermediate calculations—are transferred to the results. If you
use a large program with many assignments, you might want to use the SYMBOLS=
option to limit the columns in the results.

Alias SYM=

TEMPTABLE
generates an in-memory temporary table from the result set. The IMSTAT procedure
displays the name of the table and stores it in the _TEMPLAST_ macro variable,
provided that the statement executed successfully.

When the IMSTAT procedure exits, all temporary tables created during the IMSTAT
session are removed. Temporary tables are not displayed on a TABLEINFO request,
unless the temporary table is the active table for the request.

Details
To help manage how output is generated, options in the SCORE statement can be
brought to bear together with special syntax elements in the scoring code. For example,
the PARTITION<> option can be used to specify that the scoring code is executed
separately for each partition or for a specific partition of the data only. If you want to
tcontrol precisely which observations are used to generate output records, you can use the
__lasr_output symbol in your SAS program. When this symbol is set to 1, the row
is output. You can also use the __first_in_partition and
__last_in_partition variables to programmatically determine the first and last
observation in a partition.

The following SAS code is an example:

__lasr_output = 0;
if __first_in_partition then do;
  totalmsrp = msrp;
  minmsrp   = msrp;

For the first observation within a partition, three variables are initialized. The minimum MSRP, the total MSRP, and the number of records in the partition are then computed.

The variable ORGDRIVE is obtained by concatenating the strings of the ORIGIN and DRIVETRAIN variables.

When the last record within the partition is reached, the ___lasr_output automatic variable is set to 1, this is used to add of the current record to the result set.

The execution of the SCORE code observes the active WHERE clause in the IMSTAT run block—in other words, the scoring code is executed only for those observations that meet the WHERE condition if a WHERE clause is active.

The following example loads the SASHELP.CARS data set partitioned by the TYPE variable, and executes the previous code sample.

```sas
data lasrlib.cars(partition=(type));
  set sashelp.cars;
run;

filename fref '/path/to/scorepgm.sas';

proc imstat;
  table lasrlib.cars;
  score pgm=fref / partition;
run;
```

The PARTITION option in the SCORE statement requests the server to execute the code separately for each partition of the data. Because the code outputs one observation per partition and there are six unique values of the TYPE variable in the SASHELP.CARS data set, the scoring results show six rows:

<table>
<thead>
<tr>
<th>totalmsrp</th>
<th>minmsrp</th>
<th>numCars</th>
<th>orgdrive</th>
<th>mpgdiff</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>59760</td>
<td>19110</td>
<td>3.000000</td>
<td>Asia Front</td>
<td>-8.000000</td>
<td>Hybrid</td>
</tr>
<tr>
<td>2087415</td>
<td>17163</td>
<td>60.000000</td>
<td>EuropeAll</td>
<td>5.000000</td>
<td>SUV</td>
</tr>
<tr>
<td>7800688</td>
<td>10280</td>
<td>262.000000</td>
<td>EuropeFront</td>
<td>7.000000</td>
<td>Sedan</td>
</tr>
<tr>
<td>2615966</td>
<td>18345</td>
<td>49.000000</td>
<td>Asia Rear</td>
<td>6.000000</td>
<td>Sports</td>
</tr>
<tr>
<td>598593</td>
<td>12800</td>
<td>24.000000</td>
<td>Asia All</td>
<td>3.000000</td>
<td>Truck</td>
</tr>
<tr>
<td>865216</td>
<td>11905</td>
<td>30.000000</td>
<td>EuropeAll</td>
<td>7.000000</td>
<td>Wagon</td>
</tr>
</tbody>
</table>

The SCORE statement does not support the temporary expressions that are available in other IMSTAT statements. This is because you can compute all necessary temporary variables in the scoring code.
Example: Joining Data during Scoring with a DATA Step Hash Object

You can use many SAS functions in the scoring code that is processed by the server, including some functionality from the DATA step hash object. You can specify an in-memory table in the DECLARE statement to populate a hash object from a table.

This example shows how to use a fact table and dimension tables to instantiate one hash object for each dimension table. This effectively produces a join of the fact and dimension tables that is resolved at run time of the scoring program. The following display shows the relations:

The details of the relationships are as follows:

- MailOrder is the fact table.
- Customers joins to MailOrder on the CustNum column.
- Products joins to MailOrder on the PCode column.
- Catalog joins to MailOrder on the CatCode column.

The following code shows how to create a table with scoring code that combines the columns Date and Year from the fact table with the columns Catalog, Cost, Price, Type, Name, City, and State from the dimension tables. The code limits the results to customers from the state of New Jersey.

**File 5.1 Scoring Program**

```sas
length catcode $6;  
length pcode $6;  
length catalog $20;  
length type $15;  
length name $32;  
length city $20;  
length state $3;  
length cost 8;  
length price 8;  

declare hash cst(dataset:"star.customers");  
rc = cst.defineKey ('custnum');  
rc = cst.defineData('name');  
rc = cst.defineData('city');
```
rc = cst.defineData('state');  
rc = cst.defineDone();  
rc_cst = cst.find();  
if ((rc_cst=0) and (state = 'NJ ')) then do;  
   declare hash cat(dataset:"star.catalog");  
   rc = cat.defineKey ('catcode');  
   rc = cat.defineData('catalog');  
   rc = cat.defineDone();  
   declare hash prd(dataset:"star.products");  
   rc = prd.defineKey ('pcode');  
   rc = prd.defineData('type');  
   rc = prd.defineData('price');  
   rc = prd.defineData('cost');  
   rc = prd.defineDone();  
   rc_cat = cat.find();  
   rc_prd = prd.find();  
   if (rc_cat=0 and rc_prd=0) then output;  
end;  

1 The LENGTH statements define the variables that are accessed from the dimension tables through the hash objects. The specified types and lengths must match the data types and lengths in the dimension tables.

2 Three hash objects are declared in the program, one for each dimension table. The DECLARE HASH statement defines and names the object. The DATASET option specifies the in-memory table to use for populating the hash object. The names specified in the program code are the actual table names in the server, and not the names where a libref masks the tag.

3 The DEFINEKEY function defines the key for the hash object. The keys in the example are the same that would be used in a star schema to join the dimension tables to the fact table.

4 The DEFINEDATA function lists the columns from the dimension tables that you want to make available to the scoring step through the hash object.

5 Because this example uses the STATE variable as a filter, the program determines whether a record matches New Jersey first. Only then does the program look for matching records in the other hash objects. The FIND function is used to locate a matching record for the particular hash object. The code compares the State variable against the literal 'NJ ' after the CST.FIND() call. Otherwise, the State variable would not have been updated with the value that corresponds to the current record in the fact table.

6 A record that passes the filter and can be successfully looked up in the hash objects is output to the scoring result set.


The following example shows how the scoring program is used with the HASHDATA option:

filename fref './hash_schema.txt';
proc imstat;
  table lasr.mailorder;
  score pgm=fref hashdata(star.catalog, 2
    star.products,
    star.customers)
    symbols(catalog cost price type name city state) 3
  keep   (date year)
  temptable; 4
run;

1 The scoring program is saved in a file that is named hash_schema.txt. The file is referenced with a FILENAME statement.

2 The HASHDATA option specifies the names of the dimension tables. In this example, the table names do not match the libref that is used for accessing the fact table. Specify the tables in the HASHDATA option as they are shown in the results of the TABLEINFO / HOST="hostname.example.com" PORT=number statement. The names in the DATASET options in the DECLARE HASH statements of the scoring program must match.

3 The SYMBOLS option specifies the columns to transfer from the dimension tables to the result set. Although these are columns in an in-memory table, they are not columns in the active table, and therefore must be transferred explicitly to the result set. The KEEP option specifies the columns from the active table, MailOrder, to transfer to the result set.

4 The TEMPTABLE option specifies to store the result set as an in-memory table. Otherwise, the result set is transferred to your SAS session.

SERVERINFO Statement

The SERVERINFO statement returns information about the SAS LASR Analytic Server.

Syntax

SERVERINFO <option>;

SERVERINFO Statement Options

HOST="host-name"
  specifies the host name for the SAS LASR Analytic Server. Use this option with the PORT= option.

NORANKS
  specifies to omit the list of host names for the worker nodes. This option reduces the output of the SERVERINFO option considerably for large environments.

PORT=number
  specifies the port number for the SAS LASR Analytic Server. If you do not specify a PORT= value, then behavior of the SERVERINFO statement depends on whether an in-memory table is active. If there is no active table, then the procedure attempts to connect to the server using the LASRPORT macro variable. If a table is active, the information is gathered for the server that is implied by the libref of the active table.
SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

SERVERPARM Statement
The SERVERPARM statement enables you to change some global settings for the server if you have sufficient authorization. The user account that starts the server has privileges to modify server parameters.

Syntax
SERVERPARM <options>;

SERVERPARM Statement Options

CONCURRENT=number
specifies the number of concurrent requests that can execute in the server. Once the threshold is met, the requests are queued and then executed as the currently running requests complete.

Alias  NACTIONS=
Default  20

EXTERNALMEM=pct
specifies the percentage of memory that can be allocated before the server stops transferring data to external processes such as external actions and the SAS High-Performance Analytics procedures. If the percentage is exceeded, the server stops transferring data.

Default  75

HADOOPHOME="path"
specifies the path for the HADOOP_HOME environment variable. Changing this variable is useful for migrating SASHDAT files from one Hadoop installation to another.

Setting the HADOOP_HOME environment variable is a server-wide change. All requests, by all users, for reading files from HDFS and saving files, use the specified HADOOP_HOME. This can cause unexpected results if users are not aware of the change.

Note: If you are using this option to migrate SASHDAT files, then consider starting a server for that exclusive purpose.

Alias  HADOOP=

HOST="host-name"
specifies the host name for the SAS LASR Analytic Server. Use this option with the PORT= option.
PORT=number
specifies the port number for the SAS LASR Analytic Server. If you do not specify a PORT= value, then behavior of the SERVERPARM statement depends on whether an in-memory table is active. If there is no active table, then the procedure attempts to connect to the server using the LASRPORT macro variable. If a table is active, the information is gathered for the server that is implied by the libref of the active table.

TABLEMEM=pct
specifies the percentage of memory that can be allocated before the server rejects requests to add tables or append data. If the percentage is exceeded, adding a table or appending rows to tables fails. These operations continue to fail until the percentage is reset or the memory usage on the server drops below the threshold.

This option has no effect for non-distributed servers. For non-distributed servers, the memory limits can be controlled with the MEMSIZE system option.

Note: The specified pct value does not specify the percentage of memory allocated to in-memory tables. It is the percentage of all memory used by the entire machine that—if exceeded—prevents further addition of data to the server. The memory used is not measured at the process or user level, it is computed for the entire machine. In other words, if operating system processes allocate a lot of memory, then loading tables into the server might fail. The threshold is not affected by memory that is associated with SASHDAT tables that are loaded from HDFS.

Alias MEMLOAD=
Default 75

TEMPNAMES=YES | NO
specifies whether the server writes the full name of temporary tables to the server log file (TEMPNAMES=YES) or whether the names are masked (TEMPNAMES=NO). Because the name of the temporary table provides access to the table, the server does not display the full names of temporary tables, by default.

This option can be useful if you need to debug a series of requests that use temporary tables as input. By changing the handling of temporary names with TEMPNAMES=YES, you can see the full temporary table names in the log file.

SERVERTERM Statement
The SERVERTERM statement sends a termination request to the server that is identified through the statement options. You must have sufficient authorization for this request to succeed.

Syntax
SERVERTERM <options>;

SERVERTERM Statement Options
HOST="host-name"
specifies the host name for the SAS LASR Analytic Server. Use this option with the PORT= option.
PORT=number
specifies the port number for the SAS LASR Analytic Server.
SERVERWAIT Statement

The SERVERWAIT statement suspends execution of the IMSTAT procedure until the server that it uses receives a termination request.

Syntax

SERVERWAIT <options>;

SERVERWAIT Statement Options

HOST="host-name"
  specifies the host name for the SAS LASR Analytic Server. Use this option with the PORT= option.

PORT=number
  specifies the port number for the SAS LASR Analytic Server.

SET Statement

The SET statement is used to append in-memory tables to each other. The result of the operation is not a temporary table, but the appending of rows from the secondary tables to the active table.

Examples:
  "Example 6: Appending Tables" on page 243
  "Example 7: Appending a Non-Partitioned Table to a Partitioned Table" on page 244

Syntax

SET set-specification1 <set-specification2 ...> </options>;

Required Argument

set-specification
  specifies the table to append to the active table and options. You can list multiple set-specifications. A table can be used in more than one set-specification, and you can specify the active table in a set-specification.

  table-name <(TAG='server-tag')>
  table-name
    specifies the table to append to the active table.

  TAG='server-tag'
    specifies the server tag to use for identifying the table to append.

SET Statement Options

DROP
  specifies that the secondary tables (the tables specified in the set-specifications) are dropped from the server after the statement executes successfully. If the active table is listed in a set-specification, it is not dropped.
NOPARTITION
specifies to append the secondary tables and undo the partitioning of the active table. If the active table is partitioned, and you append partitioned tables to it, then the server rejects the request unless all the tables have the same partitioning variables, in the same order, and have the same key length. When this option is specified, the active table is no longer partitioned if the SET statement succeeds.

Alias  NOPART

WHEREWITHALL
specifies to apply the WHERE clause to the active table, in addition to the secondary tables to append. By default, the rows of the secondary tables that are appended to the active table are filtered according to the WHERE clause. Rows marked for deletion or purging are not appended to the main table. By default, the WHERE clause does not filter rows of the active table. If you want the WHERE clause to apply to the active table, specify this option.

Alias  ALLWHERE

STORE Statement
The STORE statement enables you to assign the contents of previously saved tables to macro variables. You can reuse the results from one statement as input for subsequent statements in the same IMSTAT procedure.

Syntax

STORE table-name <[table-number]>
  (row-number | _ALL_ | _LAST_ | row-list | WHERE=(where-clause) <,>
  column-number | _ALL_ | COLS=column-list) = macro-variable-name </options>;

Required Arguments

column-number
 specifies the column number to access as it appears in the default output or with the REPLAY statement. Be aware that hidden columns that might appear in an output data set when an ODS table is converted to a SAS data set are not counted. The first column is numbered 1. You can specify _ALL_ as an alternative, or specify the column names in the COLS= option.

macro-variable-name
 specifies the name of a macro variable to use for storing the value.

table-name
 specifies the saved result table to use.

row-number
 specifies the row number to access as it appears in the default output or with the REPLAY statement. The first row is numbered 1. You can specify _ALL_, _LAST_, a numeric list of rows (row-list), or a WHERE clause as alternatives. When you specify a row-list, any row number less than 1, greater than the number of rows, and duplicate row numbers, are ignored.
If you specify a WHERE= clause, you can use Boolean expressions like NOT, BETWEEN, CONTAINS, LIKE, and mathematical expressions. The following operators are available:

- prefix +, prefix -
- <>, ><, ** (max, min, and power)
- *, /
- infix +, infix -
- || (concatenation)
- =, ^=, <, <=, >, >=, IN(set), IS NULL, IS MISSING
- AND, &
- OR, |

**Examples**

The following row-list accesses rows 0, 1, and 2 from the results.

\( (0, 1, 2) \)

The following row-list accesses rows 0, 1, 2, 4, 8, 12, 16, and 20.

\( (0 \text{ to } 2 \text{ by } 1, 4, 8 \text{ to } 20 \text{ by } 4) \)

**Optional Arguments**

[**table-number**]

Specifies the table to use for accessing multi-part tables. In the following example, the HISTOGRAM statement generates one histogram table for the Cylinders variable and a second for the EngineSize variable. The two histogram tables are stored in the temporary HistTab table. In order to access the second histogram table, the \( [2] \) is used. If you do not specify a table-number, the first table is used.

**Example**

```bash
proc imstat data=mylasr.cars(tag=sashelp);
    histogram Cylinders EngineSize / save=HistTab;
    store HistTab[2] (2,6) = Engsz_Pct;
quit;
%
```

**CONTROL=**"**control-string**"

Specifies the string that used to format cell values before they are stored in the macro variable. The control string must include the same number of placeholders as the number of columns. The default placeholder is %. There are two more placeholders besides %, # and ^.

If the \( i \)-th placeholder in CONTROL= is:

- #, then it is a placeholder for the value of the \( i \)-th relevant column in the first relevant row
- ^, then it is a placeholder for the value of the \( i \)-th relevant column in the last relevant row

In the preceding statements, first means the relevant row with the smallest row number and last means the relevant row that with the largest row number. See Example 4 on page 230.

If you read character values that you want to use in a WHERE clause, you might need to enclose the placeholders in quotation marks.

**LEFT=**"**left-side-string**"

Specifies the string to assign as a prefix to the macro variable.
NODUPS
specifies to ignore duplicate formatted cell values.

Restriction
This option applies to numeric values only and when only one column is accessed from the results.

RIGHT="right-side-string"
specifies the string to assign as a suffix to macro variable.

SEPARATOR="separator-string"
specifies the string to use for separating the formatted cell values.

Default " " (the space character)

Details
The simplest use of the STORE statement is to read a value from a cell in the results, assign it to a macro variable, and use it in subsequent statements. The following statement is copied from part of Example 1 and demonstrates reading the value of a single cell into a macro variable.

store mpgtab (_last_, cols=Mean) = avgmpgcity;

More sophisticated uses of the STORE statement are possible. Example 2 on page 227 shows how variable names are read from a results table and used in a subsequent programming statement.

Perhaps the most sophisticated use of the STORE statement is to construct a string that can be used in a WHERE clause. Such a use typically requires use of the LEFT=, CONTROL=, SEPARATOR=, and RIGHT= options. The following steps describe the concept and flow:

1. Specify the cells from the results to use by specifying the rows and columns.
2. Use options to control how to construct the macro value, made up of cell values and the following:
   a. A string to prefix to the left side of the macro variable. Typically, "(" is common.
   b. A control string to use for formatting cell values for each row.
   c. A string to use for separating the formatted control strings. The control string and the separator are used.
   d. A string to use as a suffix for the right side of the macro variable. Typically, ")" is common.

For examples that demonstrate using these options, see Example 3 and Example 4.

Examples

Example 1: Accessing Multi-Part Table Output and Storing a Single Value
The following STORE statement reads the value from the second row of the Percent column into the Bin2Pct macro variable.

```plaintext
proc imstat data=example.cars(tag=sashelp);
    histogram Cylinders EngineSize / save=HistTab;
    store HistTab[2] [2,cols= percent) = Bin2Pct;
quit;
```
The PUT statement shows the value from the Bin2Pct macro variable in the SAS log.
The macro variable shows a different number of significant digits than the table output.

The PUT statement shows the value from the Bin2Pct macro variable in the SAS log.
The macro variable shows a different number of significant digits than the table output.

Example 2: Storing Variable Names
The following DISTINCT statement calculates the number of unique values for the numeric variables and stores the results in DistinctTab.

```sas
proc imstat data=example.cars(tag=sashelp);
  distinct _numeric_ / save=distincttab;
run;
```
The following display shows the results.

| Number of Distinct Values in Table SASHELP.CARS |
|-------------------------|-------------------------|------------------------|
| **Column**              | **Number of Distinct Values** | **Number of Missing Values** |
| MSRP                   | 410                      | 0                      |
| Invoice                | 425                      | 0                      |
| Engine Size            | 43                       | 0                      |
| Cylinders              | 8                        | 2                      |
| Horsepower             | 110                      | 0                      |
| MPG_City               | 23                       | 0                      |
| MPG_Highway            | 33                       | 0                      |
| Weight                 | 348                      | 0                      |
| Wheelbase              | 40                       | 0                      |
| Length                 | 67                       | 0                      |

The following STORE statement reads the results and filters for rows with a number of distinct values that is greater than 100. The values from the first column (the variable names) are stored in the macro variable VarList1.

```sas
store distincttab (where="NDistinct > 100", 1) = varlist1;
run;
```

The SAS log shows the constructed string:

```
NOTE: The string MSRP Invoice Horsepower Weight related to table distincttab has been stored in the macro variable varlist1.
```

The following CORR statement uses the variable names that were stored in the VarList1 macro variable.

```sas
corr &varlist1;
run;
```

The following display shows the results of the CORR statement.

| Pairwise Correlations for Table SASHELP.CARS |
|-------------------------|-------------------------|------------------------|
| **Column**              | **Row** | **MSRP** | **Invoice** | **Horsepower** | **Weight** |
| MSRP                   | 1       | 1.0000   | 0.9991      | 0.8269        | 0.4484     |
| Invoice                | 2       | 0.9991   | 1.0000      | 0.8237        | 0.4423     |
| Horsepower             | 3       | 0.8269   | 0.8237      | 1.0000        | 0.5308     |
| Weight                 | 4       | 0.4484   | 0.4423      | 0.6308        | 1.0000     |
Example 3: Storing Values from Multiple Cells

```sas
proc imstat data=example.cars(tag=sashelp);
  summary mpg_city / save=mpgtab;
run;
```

The following display shows the results. The overall average Mpg_City for all makes and models is highlighted.

```sas
store mpgtab (_last_, cols=Mean) = avgmpgcity;
run;
```

```sas
summary mpg_city / groupby=make save=summarytab;
run;
```

The following display shows the part of the results.

```sas
store summarytab (where="Mean > &avgmpgcity",cols=Make) = highmpgmakes /
  left="Make in (" control="'%' separator=", " right=")""); run;
```

The SAS log shows the constructed string. Notice how the LEFT= value is used to begin the string, and the CONTROL= and SEPARATOR= values are used for each row in the results. The RIGHT= value is used to end the string.

```
NOTE: The string Make in ('Honda', 'Hyundai', 'Kia', 'MINI', 'Mazda', 'Mitsubishi', 'Oldsmobile', 'Pontiac', 'Saab', 'Saturn', 'Scion', 'Subaru', 'Suzuki', 'Toyota', 'Volkswagen') related to table summarytab has been stored in the macro variable highmpgmakes.
```
Finally, the constructed string is used in a WHERE clause. The CROSSTAB statement shows the frequency of each MPG_City value for each Make that is specified in the WHERE clause.

```
where &highmpgmakes;
   crosstab mpg_city*make;
run;
```

The following display shows part of the crosstabulation results.

<p>| Cross-tabulation of MPG_City by Make for Table SASHELP.CARS with Aggregator N |
|-----------------------------------------------|-------------|-------------|-----------|---|</p>
<table>
<thead>
<tr>
<th>MPG_City</th>
<th>Honda</th>
<th>Hyundai</th>
<th>Kia</th>
<th>MINI</th>
<th>Mazda</th>
<th>Mitsubishi</th>
<th>Oldsmobile</th>
<th>Pontiac</th>
<th>Saab</th>
<th>Saturn</th>
<th>Scion</th>
<th>Subaru</th>
<th>Suzuki</th>
<th>Toyota</th>
<th>Volkswagen</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Example 4: Storing Values from Vertically Stacked Cells**

The following PRECENTILE statement calculates the quantiles for the Mpg_City variable and saves the results in MpgTab.

```
proc imstat data=example.cars(tag=sashelp);
   percentile mpg_city / save=mpgtab;
run;
```

The following display shows the results.

<p>| Percentiles and Quantiles for Table SASHELP.CARS |
|-----------------------------------------------|-------------|-------------|-----------|</p>
<table>
<thead>
<tr>
<th>Column</th>
<th>Percentile</th>
<th>Value</th>
<th>Converged</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP_City</td>
<td>25</td>
<td>17.00000000</td>
<td>Yes</td>
</tr>
<tr>
<td>MP_City</td>
<td>50</td>
<td>19.00000000</td>
<td>Yes</td>
</tr>
<tr>
<td>MP_City</td>
<td>75</td>
<td>21.50000000</td>
<td>Yes</td>
</tr>
</tbody>
</table>

To create a WHERE clause that uses the values for the third quartile of the Mpg_City variable (between 19 and 21.5), the following STORE statement is used.

```
store mpgtab (where="Pctl >= 50", cols=Value Value) = q3 /
   control="(mpg_city between # and ^)"; run;
```

The STORE statement performs the following:

- Filters the results to the rows where Percentile is greater than or equal to 50 (two rows).
- Reads the Value column (twice).
- Substitutes the first value in the first row for the # placeholder. Substitutes the second value from last row in the ^ placeholder, because it is the second placeholder.
- Stores the resulting string (mpg_city between 19 and 21.5) in the Q3 macro variable.
The SAS log shows the constructed string:

```
NOTE: The string (mpg_city between 19 and 21.5) related to table mpgtab has been
stored in the macro variable q3.
```

The following statements then use a constructed string in a WHERE clause and request a
summary of the Mpg_City variable that is grouped by Make.

```sas
where &q3.;
  summary mpg_city / groupby=make;
run;
```

The following display shows part of the results.

<table>
<thead>
<tr>
<th>Make</th>
<th>Column</th>
<th>Min</th>
<th>Max</th>
<th>N</th>
<th>Sum</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Std Error</th>
<th>Coefficient of Variation</th>
<th>Number Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acura</td>
<td>MPG_City</td>
<td>20.0000</td>
<td>20.0000</td>
<td>1</td>
<td>20</td>
<td>20.0000</td>
<td>0.4082</td>
<td>0.1667</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Audi</td>
<td>MPG_City</td>
<td>20.0000</td>
<td>21.0000</td>
<td>6</td>
<td>121</td>
<td>20.1967</td>
<td>0.6304</td>
<td>0.1746</td>
<td>2.0244</td>
<td>0</td>
</tr>
<tr>
<td>BMW</td>
<td>MPG_City</td>
<td>19.0000</td>
<td>21.0000</td>
<td>13</td>
<td>256</td>
<td>19.6323</td>
<td>0.4082</td>
<td>0.1667</td>
<td>3.2014</td>
<td>0</td>
</tr>
<tr>
<td>Buick</td>
<td>MPG_City</td>
<td>19.0000</td>
<td>20.0000</td>
<td>6</td>
<td>119</td>
<td>19.8333</td>
<td>0.4082</td>
<td>0.1667</td>
<td>2.0564</td>
<td>0</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>MPG_City</td>
<td>19.0000</td>
<td>21.0000</td>
<td>5</td>
<td>100</td>
<td>20.0000</td>
<td>1.0000</td>
<td>0.4472</td>
<td>5.0000</td>
<td>0</td>
</tr>
</tbody>
</table>

### TABLE Statement

The TABLE statement is used to specify the in-memory table to use for subsequent IMSTAT procedure
statements. You can use this statement to switch between different in-memory tables in a single run of the
IMSTAT procedure.

**Example:**  “Example 5: Creating a Star Schema” on page 241

### Syntax

```
TABLE <libref.member-name>;
```

### Optional Argument

`libref.member-name`

specifies the libref for the SAS LASR Analytic Server and the table name.

If you do not specify the libref and member-name, the procedure closes the table that
is currently open.

### Example

A common use for the TABLE statement is to reference a temporary table with the
`libref.&_TEMPLAST_` macro variable. Temporary tables are in-memory tables that are
created with the results of a statement that supports the TEMPTABLE option.

```sas
proc imstat data=lasrlib.sales2012;
  partition customerid;
```
run;

table lasrlib.&_templast_;  
run;

/*
* More statements for the partitioned table.
* The PROMOTE statement can be used to convert the
* temporary table to a regular table.
*/
quit;

---

**TABLEINFO Statement**

The TABLEINFO statement is used to return information about an in-memory table. This information includes the table name, label, number of rows and column, owner, encoding, and the time of table creation. If no table is in use, then information is returned for the in-memory tables for the server specified in the HOST= and PORT= options.

### Syntax

```
TABLEINFO <options>;
```

### TABLEINFO Statement Options

- **HOST=host-name**  
  specifies the host name for the SAS LASR Analytic Server. Use this option with the PORT= option.

- **PARTVARS**  
  specifies to include information about partition and orderby variables in the output of the TABLEINFO statement. This enables you to retrieve the names of those variables. If a table is not partitioned or ordered, "N/A" is displayed.

- **PORT=number**  
  specifies the port number for the SAS LASR Analytic Server. If you do not specify a PORT= value, then behavior of the TABLEINFO statement depends on whether an in-memory table is active. If there is no active table, then the procedure attempts to connect to the server using the LASRPORT macro variable. If a table is active, the information is gathered for the server that is implied by the libref of the active table.

- **SAVE=table-name**  
  saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.

---

**UPDATE Statement**

The UPDATE statement performs rowwise updates of the data in an in-memory table.
Syntax

UPDATE variable1=value1 <variable2=value2 ...> </options>;
UPDATE DATA=libref.member-name </options>;

Required Arguments

variable
specifies the name of the variable to update.

type
specifies the value to assign to the variable.

libref.member-name
specifies the libref and table name of a SAS data set to use for updating the in-memory table. The data set must contain the variables and values that you want to update. You can specify a WHERE variable in the data set to apply as a filter to the particular set of update values. This clause in the data set augments the overall WHERE clause, if one is specified.

UPDATE Statement Options

CODE=file-reference
specifies a file reference to a SAS program to use for the row update (an update script). You can combine the specification of a SAS program through the CODE= option with the name-value pair specification or the DATA= specification for bulk updates. The updates that are specified in the name-value pair and DATE= specifications are performed first and then the update script executes on the modified row to produce the update.

Alias PGM=

NOPREPARSE
prevents the procedure from pre-parsing and pre-generating code for temporary expressions, scoring programs, and other user-written SAS statements.

When this option is specified, the user-written statements are sent to the server "as-is" and then the server attempts to generate code from it. If the server detects problems with the code, the error messages might not be as detailed as the messages that are generated by SAS client. If you are debugging your user-written program, then you might want to pre-parse and pre-generate code in the procedure. However, if your SAS statements compile and run as you want them to, then you can specify this option to avoid the work of parsing and generating code on the SAS client.

When you specify this option in the PROC IMSTAT statement, the option applies to all statements that can generate code. You can also exclude specific statements from pre-parsing by using the NOPREPARSE option in statements that allow temporary columns or the SCORE statement.

Alias NOPREP

SAVE=table-name
saves the result table so that you can use it in other IMSTAT procedure statements like STORE, REPLAY, and FREE. The value for table-name must be unique within the scope of the procedure execution. The name of a table that has been freed with the FREE statement can be used again in subsequent SAVE= options.
TEMPEXPRESS="SAS-expressions"
TEMPEXPRESS=file-reference
    specifies either a quoted string that contains the SAS expression that defines the
temporary variables or a file reference to an external file with the SAS statements.

    Alias TE=

TEMPNAMES=variable-name
TEMPNAMES=(variable-list)
    specifies the list of temporary variables for the request. Each temporary variable
must be defined through SAS statements that you supply with the TEMPEXPRESS=
option.

    Alias TN=

Details
It is common to use the UPDATE statement with a WHERE clause. The clause filters
the rows to which the updates are applied. If you are unsure about the number of rows
that can be updated, use the NUMROWS statement to determine how many rows would
be affected by the rowwise update.

You can update the values of ORDERBY variables, but you cannot update the value of
variables that are used for constructing partition keys.

You cannot update the values of permanent computed variables. Their values are
determined by the SAS program that originally defined them.

QUIT Statement
The QUIT statement is used to end the procedure execution. When the procedure reaches the QUIT
statement, all resources allocated by the procedure are released. You can no longer execute procedure
statements without invoking the procedure again. However, the connection to the server is not lost,
because that connection was made through the SAS LASR Analytic Server engine. As a result, any
subsequent invocation of the procedure that uses the same libref executes almost instantaneously
because the engine is already connected to the server.

Interaction: Using a DATA step or another procedure step is equivalent to issuing a QUIT
statement. If there is an error during the procedure execution, it is also equivalent to
issuing a QUIT statement.

Syntax
QUIT;

Examples: IMSTAT Procedure (Data and Server Management)

Example 1: Partitioning a Table into a Temporary Table

Details
This PROC IMSTAT example demonstrates partitioning a table as it is loaded to memory and then saving it to a temporary table with different partitioning variables.

Program

libname example sasiola host="grid001.example.com" port=10010 tag='hps';

data example.prdsale(partition=(country region)); 1
set sashelp.prdsale;
run;

proc imstat data=example.prdsale;
    partitioninfo;
    summary actual predict / partition; 2
run;

    /* partition the active table, example.prdsale, by region and prodtype */
    partition region prodtype; 3
run;
    table example.&_templast_; 4
run;
    partitioninfo;
    summary actual predict / partition; 3
quit;

Program Description
1. The Prdsale data set is loaded into memory and partitioned by the unique combinations of the formatted values for the Country and Region variables.

2. The procedure examines the partitioning of the table and requests a summarization of the Actual and Predict variables by the partition values (unique combinations of Country and Region).

3. In order to accommodate a different data access pattern, the table is partitioned by unique combinations of the Region and Prodtype variables. The table is stored in a temporary table and the name is assigned to the _TEMPLAST_ macro variable.

4. The TABLE statement references the _TEMPLAST_ macro variable and sets the temporary table as the active table. All statements that follow use the temporary table.
5. As with the previous SUMMARY statement, the partitioning is examined and the summary is requested for the Actual and Predict variables by the unique combinations of the Region and Prodtype variables.

Output

Output 5.1  Partitions for Prdsale When Partitioned by Country and Region

Output 5.2  Summary Statistics for Prdsale When Partitioned by Country and Region

Output 5.3  Partitions for Prdsale When Partitioned by Region and Prodtype
Summary Statistics for Prdsale When Partitioned by Region and Prodtype

<table>
<thead>
<tr>
<th>REGION</th>
<th>PRODTYPE</th>
<th>Column</th>
<th>Min</th>
<th>Max</th>
<th>N</th>
<th>Sum</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Std Error</th>
<th>Coefficient of Variation</th>
<th>Number Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAST</td>
<td>FURNITURE</td>
<td>ACTUAL</td>
<td>4.000</td>
<td>999.0</td>
<td>288</td>
<td>146471</td>
<td>590.56</td>
<td>283.91</td>
<td>16.7296</td>
<td>55.6241</td>
<td>0</td>
</tr>
<tr>
<td>EAST</td>
<td>FURNITURE</td>
<td>PREDICT</td>
<td>0.000</td>
<td>1000.0</td>
<td>288</td>
<td>141116</td>
<td>489.99</td>
<td>299.16</td>
<td>17.6200</td>
<td>61.0542</td>
<td>0</td>
</tr>
<tr>
<td>EAST</td>
<td>OFFICE</td>
<td>ACTUAL</td>
<td>13.000</td>
<td>1000.0</td>
<td>432</td>
<td>223390</td>
<td>515.03</td>
<td>284.22</td>
<td>13.6745</td>
<td>54.9653</td>
<td>0</td>
</tr>
<tr>
<td>EAST</td>
<td>OFFICE</td>
<td>PREDICT</td>
<td>12.000</td>
<td>994.00</td>
<td>432</td>
<td>217896</td>
<td>503.99</td>
<td>288.08</td>
<td>13.8996</td>
<td>57.1640</td>
<td>0</td>
</tr>
<tr>
<td>WEST</td>
<td>FURNITURE</td>
<td>ACTUAL</td>
<td>3.000</td>
<td>996.00</td>
<td>288</td>
<td>140154</td>
<td>500.53</td>
<td>297.06</td>
<td>17.5046</td>
<td>59.3491</td>
<td>0</td>
</tr>
<tr>
<td>WEST</td>
<td>FURNITURE</td>
<td>PREDICT</td>
<td>0.000</td>
<td>1000.0</td>
<td>288</td>
<td>137202</td>
<td>476.40</td>
<td>265.60</td>
<td>16.0320</td>
<td>59.9604</td>
<td>0</td>
</tr>
<tr>
<td>WEST</td>
<td>OFFICE</td>
<td>ACTUAL</td>
<td>6.000</td>
<td>1000.00</td>
<td>432</td>
<td>215522</td>
<td>499.82</td>
<td>285.74</td>
<td>13.7477</td>
<td>57.1689</td>
<td>0</td>
</tr>
<tr>
<td>WEST</td>
<td>OFFICE</td>
<td>PREDICT</td>
<td>6.000</td>
<td>999.00</td>
<td>432</td>
<td>210281</td>
<td>488.76</td>
<td>274.62</td>
<td>13.2125</td>
<td>56.4170</td>
<td>0</td>
</tr>
</tbody>
</table>

**Example 2: Promoting Temporary Tables to Regular Tables**

**Details**

The SUMMARY, SCORE, CROSSTAB, PERCENTILE, and DISTINCT statements offer a TEMPTABLE option. When you specify the TEMPTABLE option, the results of the statement are written to a temporary table. The PARTITION statement also results in a temporary table. If you want to keep the table, you can use the PROMOTE statement to convert the table from being a temporary table to a regular table. Once you do this, other users can access the data.

**Program**

```sas
libname example sasiola host="grid001.unx.sas.com" port=10010 tag='hps';

data example.prdsale; set sashelp.prdsale; run;

proc imstat data=example.prdsale;
  summary / groupby=(country) temptable;
run;
  table example.&_templast_; run;
  promote sum_by_country;
run;
  table example.sum_by_country; run;
  fetch / format to=10; quit;
```

**Program Description**

1. The TEMPTABLE option stores the results of the SUMMARY statement to a temporary table. If this table is not promoted before the QUIT statement, it is removed from memory.

2. The TABLE statement references the _TEMPLAST_ macro variable and sets the temporary table as the active table. All statements that follow use the temporary table.
3. The PROMOTE statement converts the temporary table to a regular table with the name Sum_By_Country. The table is associated with the current library through the libref, Example. The SAS log also includes a note that indicates how to specify the libref and table name.

4. The TABLE statement makes the table the active table explicitly by specifying the libref and table name. The Sum_By_Country table is not removed from memory when the IMSTAT procedure terminates.

5. All the subsequent statements that follow the TABLE statement use the newly promoted table.

The example does not show the use of SAS LASR Analytic Server engine server tags. You can use server tags with the PROMOTE statement as shown in the following code sample.

```sas
proc imstat data=example.prdsale;
   summary / groupby=(region) temptable;
   run;
   table example.&_templast_;
   run;
   promote sum_by_region / tag="sales";
   run;
   table example.sum_by_country (tag="sales");
   run;
quit;
```

As shown in the previous example, the TAG= option is used in the PROMOTE statement. To access the table, the TABLE statement uses the TAG= data set option.

As shown in the following sample, the SAS log indicates the libref, table name, and server tag to use for accessing the table.

**Log 5.1  SAS Log for the PROMOTE Statement with the TAG= Option**

```
NOTE: The temporary table _T_BE5C2602_45A0DCB8 was successfully promoted to the LASR Analytic Server table WORK.SUM_BY_COUNTRY. You can access this table with the TABLE statement as table EXAMPLE.sum_by_country(tag='sales').
```

---

**Example 3: Rebalancing a Table**

**Details**

It might be beneficial to rebalance the rows of a table if the data access patterns do not take advantage of partitioning or if the HDFS block distribution becomes uneven.

**Program**

```sas
libname example sasiola host="grid001.example.com" port=10010 tag='hps';

proc imstat immediate;
   table example.table1;
   distributioninfo;
```
balance;
droptable; 2
table example.&_templast_; 3
promote table1; 4
table example.table1;
distributioninfo; 5
/* save path="/hps" replace; */ 6
quit;

Program Description

1. The DISTRIBUTIONINFO statement displays the number of rows from Table1 on each machine in the cluster.

2. The DROPTABLE statement is used to drop the active table, Table1.

3. The BALANCE statement rebalanced Table1 into a temporary table. The TABLE statement is used with the &_TEMPLAST_ macro variable to access the temporary table.

4. The PROMOTE statement changes the temporary table into a regular in-memory table with the original table name, Table1.

5. After setting the Table1 as the active table with the TABLE statement, the DISTRIBUTIONINFO statement displays the nearly homogenous distribution of rows.

6. The SAVE statement can be used to save the table back to HDFS with the homogeneous block distribution.

Output

The following output shows the partial display for the first DISTRIBUTIONINFO statement. One machine has zero rows and another machine has approximately twice the number of rows.

Output 5.5  Uneven Row Distribution

<table>
<thead>
<tr>
<th>LASR Node</th>
<th>Number of Partitions</th>
<th>Number of Records</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>165971</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>165964</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>165964</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>165964</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>165964</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>331955</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>165964</td>
</tr>
</tbody>
</table>
The following output shows the homogenous distribution of rows after the BALANCE statement is used.

**Output 5.6  Homogenous Row Distribution**

<table>
<thead>
<tr>
<th>LASR Node</th>
<th>Number of Partitions</th>
<th>Number of Records</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>165977</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>165977</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>165977</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>165977</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>165977</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>165977</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>165977</td>
</tr>
</tbody>
</table>

---

**Example 4: Deleting Rows and Saving a Table to HDFS**

**Details**

The server can delete rows from in-memory tables and also save tables to HDFS. The following example demonstrates using WHERE clause processing across RUN-group boundaries to copy a subset of an in-memory table to HDFS and then delete the subset from memory.

**Program**

```sas
libname example sasiola host="grid001.example.com" port=10010 tag='hps';

data example.prdsale; set sashelp.prdsale; run;

proc imstat data=example.prdsale;
   where year=1994 and quarter=1; 1
   save path="/dept/sales/y1994q1" copies=1 fullpath; 2
   run;
   deleterows / purge; 3
   run;
   where; 4
   summary actual;
run;
```
Program Description

1. Once the WHERE clause is specified, it applies to the statements that follow it. It also crosses RUN boundaries.

2. The SAVE statement is subject to the WHERE clause. As a result, the records from the Prdsale table that meet the WHERE clause are saved to /dept/sales/y1994q1.sashdat. The FULLPATH option is used to specify the table name instead of using the name of the active table. This is particularly useful when saving temporary tables.

3. The DELETEROWS statement is also subject to the WHERE clause. The records that were just saved to HDFS are now deleted and purged from memory. (The DELETEROWS statement without the PURGE option would mark the records for deletion and exclude them from being used in calculations, but it does not free the memory resources.)

4. The WHERE clause is cleared and the SUMMARY statement that follows is performed against all the remaining records in the Prdsale table.

This pattern of using a WHERE clause to subset an in-memory table, save the records to HDFS, and then delete them can be combined with the APPEND data set option of the SAS LASR Analytic Server engine. You can create a sliding window for keeping months or years of data in memory for analysis, yet keeping it up-to-date by appending the most recent records.

Output

<table>
<thead>
<tr>
<th>Information from Saving Table SALES.PRDSALE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Path</td>
</tr>
<tr>
<td>Number of Records</td>
</tr>
<tr>
<td>Block Size (kBytes)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Status of Records in Table SALES.PRDSALE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Records</td>
</tr>
<tr>
<td>----------------------------</td>
</tr>
<tr>
<td>1440</td>
</tr>
</tbody>
</table>

Example 5: Creating a Star Schema

Details

The following example demonstrates using the SCHEMA statement to join dimension tables with a fact table.

Program

```
libname example sasiola host="grid001.example.com" port=10010 tag='hps';
```
proc imstat;
  table example.mailorder; 1
    schema catalog  (catCode=CatCode)
      products (pcodes=pcode )
      customers(custnum=custnum); 2
  run;

  table example.&_templast_ 3
  run;
  columninfo;
quit;

**Program Description**

1. Table Example.MailOrder is set as the active table. This table is the fact table for the star schema.

2. The SCHEMA statement joins the tables Catalog, Products, and Customers to the active table, MailOrder. The columns to use as keys for joining each table are enclosed in parenthesis.

3. The result of the SCHEMA statement is a temporary table or view. Use the &_TEMPLAST_ macro variable to refer to the star schema. If you want to persist the star schema, use the PROMOTE statement.
The following output shows the temporary table name and how the dimension table names are used as prefixes for the column names.

### Example 6: Appending Tables

#### Details

The following example demonstrates using the `SET` statement to append tables with the active table.

#### Program

```plaintext
libname example sasiola host="grid001.example.com" port=10010 tag='hps';
libname hdfs sashdat host="grid001.example.com" install="/opt/TKGrid" path="/hps";
```
proc lasr add data=hdfs.january port=10010; 2
   performance host="grid001.example.com" nodes=all;
run;

proc lasr add data=hdfs.february port=10010; 3
   performance host="grid001.example.com" nodes=all;
run;

data example.march; 4
   set otherlib.march;
run;

proc imstat;
   table example.january; 5
   set february / drop; 6
   set march; 7
run;

   save path="/hps/qtr1" copies=1 replace fullpath; 8
quit;

Program Description
1. The value for the TAG= option in the SAS LASR Analytic Server LIBNAME statement matches the PATH= value for the SAS Data in HDFS engine LIBNAME statement.
2. The first table, January, is loaded to memory from HDFS.
3. The second table, February, is loaded to memory from HDFS. The tables are still independent in-memory tables.
4. The third table, March, is loaded from another library into the server with the SAS LASR Analytic Server engine.
5. The first table, January, is set as the active table.
6. The second table, February, is appended to the active table. The DROP option specifies to remove the February table from memory as soon as the SET statement completes.
7. The third table, March, is appended to the active table. This table remains in memory.
8. The February and March tables are now appended to the active table, January. The SAVE statement saves the table to HDFS with the name Qtr1.

Example 7: Appending a Non-Partitioned Table to a Partitioned Table

Details
The following example demonstrates how to append a table that is not partitioned to an in-memory table that is partitioned. The SET statement is used to append the table.
As an alternative, if the table to append is not already in memory, you can append the rows to the partitioned in-memory table with the SAS LASR Analytic Server engine. For more information, see “APPEND= Data Set Option” on page 309.

Program

```sas
libname example sasiola host="grid001.example.com" port=10010 tag='hps';
libname hdfs sashdat host="grid001.example.com" install="/opt/TKGrid" path="/hps";

proc lasr add data=hdfs.transactions(partition=(customerid)) port=10010;
    performance host="grid001.example.com" nodes=all;
run;

proc lasr add data=hdfs.recenttrans(partition=(dateid)) port=10010;
    performance host="grid001.example.com" nodes=all;
run;

proc imstat;
    table example.recenttrans;
    partition customerid;
run;
    table example.transactions;
    set &_templast_ / drop;
quit;
```

Program Description

1. The value for the TAG= option in the SAS LASR Analytic Server LIBNAME statement matches the PATH= value for the SAS Data in HDFS engine LIBNAME statement.
2. The first table, Transactions, is loaded to memory from HDFS. The table is partitioned by values of the CustomerId variable.
3. The second table, RecentTrans, is loaded to memory from HDFS. The table is partitioned by values of the DateId variable.
4. The second table, RecentTrans, is set as the active table and then partitioned into a temporary table with the PARTITION statement. The temporary table is partitioned by values of the CustomerId variable.
5. The first table, Transactions, is set as the active table.
6. The temporary table is appended to the active table. The DROP option specifies to remove the temporary table from memory as soon as the SET statement completes.

Example 8: Storing Temporary Variables

Details

Many statements offer a SAVE= option that is used to save the result table of the statement for use in other IMSTAT procedure statements. You can use the STORE statement to assign a value from the saved result table to a macro variable.
Program

libname example sasiola host="grid001.unx.sas.com" port=10010 tag='hps';

data example.prdsale(partition=(country region)); set sashelp.prdsale; run;

proc imstat data=example.prdsale immediate noprint;
    percentile actual / partition="U.S.A.", "EAST" save=tab1;
run;
    percentile actual / partition="CANADA", "EAST" save=tab2;
run;
    store tab1(3,5) = us_SeventyFivePct;
    run;
    store tab2(3,5) = ca_SeventyFivePct;
    run;
%put %sysevalf(&us_SeventyFivePct - &ca_SeventyFivePct);
replay tab2;
run;

free tab1 tab2;
free macro=us_SeventyFivePct;
free macro=ca_SeventyFivePct;
quit;

Program Description
1. The NOPRINT option suppresses displaying the results tables.
2. The results tables for the PERCENTILE statements are saved to temporary tables.
3. The STORE statements access the results tables and store the value from the fifth column in the third row (the 75th percentile) to macro variables.
4. The REPLAY statement displays the results table for the second PERCENTILE statement.
5. The FREE statements releases the memory used by results tables and the macro variables.

Log Output
The SAS log describes how the values are stored to the macro variables.

store tab1(3,5) = us_SeventyFivePct;
NOTE: The numeric value 746.5 from row 3, column 5 of table tab1 has been stored in the macro variable us_SeventyFivePct.
run;

store tab2(3,5) = ca_SeventyFivePct;
NOTE: The numeric value 759.5 from row 3, column 5 of table tab2 has been stored in the macro variable ca_SeventyFivePct.
Overview: IMXFER Procedure

What Does the IMXFER Procedure Do?

The IMXFER procedure is used to transfer in-memory tables between two distributed SAS LASR Analytic Server instances. The procedure takes advantage of network topology and parallelism as much as possible.

The IMXFER procedure cannot be used with a non-distributed SAS LASR Analytic Server.

Syntax: IMXFER Procedure

PROC IMXFER <options>;
   SERVER server-name <HOST=host-name> <PORT=number>;
   TABLE export-server-name export-table-name
      import-server-name <import-table-name> </options>;
QUIT;

Examples: IMXFER Procedure

Example 1: Copying Tables from One Server to Another
Example 2: Copying Tables from One Cluster to Another
PROC IMXFER Statement
Transfers an in-memory table.

Syntax

PROC IMXFER <options>;

Optional Arguments

HOSTONLY
specifies to transfer the tables through the root nodes on the two clusters. With this
option, the data are collected by the exporting root node before sending them to the
importing root node. The importing root node then distributes the data to its worker
nodes before passing the data to the importing server instance.

Specify this option if you know in advance that the worker nodes of the server
instances do not have network communication with each other. (Even if you do not
specify this option when this network topology exists, the procedure detects the lack
of communication, and routes the data this way automatically.) Specify the option so
that time is not lost trying to establish network connections between the clusters.

Alias NOWORKER

IMMEDIATE
specifies that the procedure executes one statement at a time rather than
accumulating statements in RUN blocks.

Alias SINGLESTEP

LASRERROR
specifies that the procedure terminate when an error message is received from one of
the servers.

If you do not specify this option, the IMXFER procedure attempts to continue
interactive processing of programming statements. For example, if you receive an
error that a table with the same name already exists in the importing server instance,
you might prefer to change the name and continue rather than end the procedure.

NOPRINT
This option suppresses the generation of ODS tables and other printed output in the
IMXFER procedure.

NOTIMINGMSG
When an action completes successfully, the IMXFER procedure generates a SAS log
message that contains the execution time of the request. Specify this option to
suppress the message.

Alias NOTIME

TIMEOUT=n
specifies the time in seconds that the worker nodes of the exporting server waits for
network connections. When this interval of time has passed, the data transfer occurs
through the root nodes only.
SERVER Statement

The SERVER statement is used to specify a server instance to use in a transfer. In the statement, you assign a logical name to the server and you use that name subsequently to refer to the particular server instance. There is no limit to the number of SERVER statements. You can establish connections to more than two servers with the IMXFER procedure.

Syntax

SERVER server-name <HOST=host-name> <PORT=number>;

Required Argument

server-name
specifies the name to use for referring to the server instance. The name is used in the TABLE statement to identify the exporting server and the importing server.

SERVER Statement Options

HOST="host-name"
specifies the host name for the SAS LASR Analytic Server. If this option is not specified, then the GRIDHOST environment variable is used.

PORT=number
specifies the port number for the SAS LASR Analytic Server.

Default 10616

TABLE Statement

The TABLE statement is used to specify the table to export from one server and import to another server.

Syntax

TABLE export-server-name export-table-name import-server-name <import-table-name> </options>;

Required Arguments

export-server-name
specifies the name to use for the server instance that is exporting the table.

export-table-name
specifies the in-memory table to export. The name is specified as server-tag.member-name.

import-server-name
specifies the name to use for the server instance that is importing the table.
**Optional Argument**

`import-table-name`

specifies the name to use for the imported table.

If you do not specify a name, then the IMXFER procedure attempts to create a table with the same name as the exported table. If a table with the same name already exists in the importing server, then the transfer fails.

If you specify a table name, prefix the table name with the tag that you want to use for accessing the table. For example, `HPS` and `USER.SASDEMO` can be used as tags for a table name. For more information, see “Understanding Server Tags” on page 296.

**TABLE Statement Options**

`DELETED= INCLUDE | INC | EXCLUDE`

specifies how rows that are marked for deletion are handled in the transfer. By default, `DELETED=EXCLUDE`, which implies that any row that has a deletion mark is not transferred.

If you specify `DELETED=INCLUDE`, the IMXFER procedure instructs the server to ignore the deletion marks. Any rows that are marked for purging are not transferred, regardless of the `DELETED=` option.

Default `EXCLUDE`

`WHERE=where-expression`

specifies the WHERE clause to apply to the exported table. Only rows that meet the conditions of the WHERE expression are transferred.

Alias `FILTER=`

`PARTITION= NO | REMOVE | YES`

specifies how to handle partitioning (and ordering within the partitions) when a partitioned table is transferred. By default, `PARTITION=YES`, and implies that a partitioned table is transferred to the importing server and remains partitioned and ordered by the same variables. When the servers have different numbers of worker nodes, there is no guarantee that partitions end up on the same nodes. However, it is guaranteed that partitions appear together on a node in the importing server.

Partitioning incurs some overhead and if you transfer a table from a smaller to a larger number of nodes, you might not want to apply partitioning. (Removing the partitioning spreads the data out more evenly in the importing server.) Or, you might not want to maintain partitioning on transfer if the transfer is for archival purposes. In that case, specify `PARTITION=NO` or `PARTITION=REMOVE`. This transfers the table to the importing server without the partitioning information.

Default `YES`

`PERMISSION=mode`

specifies the permission setting for accessing the imported table. The mode value is specified as an integer value such as 755. The mode corresponds to the mode values that are used for UNIX file access permissions.

Alias `PERM=`

Range 600 to 777
QUIT Statement

This statement terminates the IMXFER procedure. When the QUIT statement is reached, all resources allocated by the procedure are released and all connections to servers are terminated.

Interaction: Using a DATA step or another procedure step is equivalent to issuing a QUIT statement. If there is an error during the procedure execution, it is also equivalent to issuing a QUIT statement.

Syntax

QUIT;

Examples: IMXFER Procedure

Example 1: Copying Tables from One Server to Another

Details

It might be necessary to copy tables from one SAS LASR Analytic Server instance to another.

Program

```sas
proc imxfer;
   server s1 host="grid001.example.com" port=10031;
   server s2 host="grid001.example.com" port=10010;
   table s1 public.fact_table s2;  
quit;
```

Program Description

1. The first SERVER statement creates a reference to the server that is listening on port 10031. The second SERVER statement creates a reference to a server on the same host, but listening on port 10010.

2. The TABLE statement transfers the Hps.Fact_Table table from the server that is listening on port 10031 to the server that is listening on port 10010. Because no import-table-name is specified, the table uses the name Hps.Fact_Table on the importing server.
Example 2: Copying Tables from One Cluster to Another

Details

It might be necessary to copy tables from one SAS LASR Analytic Server instance on one cluster to a server that is running on a different cluster. The clusters can have different numbers of machines.

When the number of machines is not the same, the IMXFER procedure automatically redistributes the rows of the table to provide the most even distribution possible. In most cases, equalizing the data distribution equalizes the work load and provides the best performance. By default, partitioned tables remain partitioned on the importing server. For more information, see the PARTITION= on page 250 option for the TABLE statement.

Program

```
proc imxfer;
  server s1 host="grid001.example.com" port=10031;
  server s2 host="cluster2.example.com" port=10010;
  table s1 public.inventory s2 hps.inventory;
quit;

/* access the transferred table */
libname cluster2 sasiola host="cluster2.example.com" port=10010 tag="hps";
proc imstat;
  table cluster2.inventory;
run;

distributioninfo;
quit;
```

Program Description

1. The first SERVER statement creates a reference to the server that is listening on port 10031. The second SERVER statement creates a reference to a server on a different cluster that is listening on port 10010.

2. The TABLE statement transfers the Public.Inventory table from the server that is listening on port 10031 to the server on the other cluster. The table is renamed to Hps.Inventory on the importing server.

3. To access the transferred table, the LIBNAME statement must use the value "hps" as the server tag.

4. The DISTRIBUTIONINFO statement for the IMSTAT procedure displays the number of rows that are used on each machine in the second cluster.
Chapter 7
OLIPHANT Procedure

Overview: OLIPHANT Procedure

What about the SAS Data in HDFS Engine?

The SAS Data in HDFS engine replaces the functionality provided by the OLIPHANT procedure. For more information, see “Using the SAS Data in HDFS Engine” on page 315.

What Does the OLIPHANT Procedure Do?

The OLIPHANT procedure is used to add, delete, and manage SASHDAT files that are stored in the Hadoop Distributed File System (HDFS). The procedure is used to add data sets from SAS libraries into HDFS. Once the data is in HDFS, it is stored as a SASHDAT file. The filename for the SASHDAT file is always lowercase. The procedure is also used to remove SASHDAT files from HDFS. For the data in SASHDAT files, the procedure can provide information about the data such as file size, block size, column count, row count, and so on.
Understanding How SAS LASR Analytic Server Uses HDFS

The SAS LASR Analytic Server reads data in parallel from the SASDAT files that are added to HDFS.

Concepts: OLIPHANT Procedure

Adding Big Data

The best performance for reading data into memory on the SAS LASR Analytic Server occurs when the server is co-located with the distributed data and the data is distributed evenly. The OLIPHANT procedure distributes the data such that parallel read performance by the SAS LASR Analytic Server is maximized. In addition, the distribution also ensures an even workload for query activity performed by the SAS LASR Analytic Server.

In order to produce an even distribution of data, it is important to understand that Hadoop stores data in blocks and that any block that contains data occupies the full size of the block on disk. The default block size is 32 megabytes and blocks are padded to reach the block size after the data is written. The data is distributed among the machines in the cluster in round-robin fashion. In order to maximize disk space, you can specify a block size that minimizes the padding.

It is important to know the size of the input data set such as the row count and the length of a row. This information, along with the number of machines in the cluster, can be used to set a block size that distributes the blocks evenly on the machines in the cluster and uses the space in the blocks efficiently.

For example, if the input data set is approximately 25 million rows with a row length of 1300 bytes, then the data set is approximately 30 gigabytes. If the hardware is a cluster of 16 machines, with 15 used to provide HDFS storage, then storing 2 gigabytes on each machine is optimal. In this case, a BLOCKSIZE= setting of 32 megabytes or 64 megabytes would fill the overwhelming majority of blocks with data and reduce the space that is wasted by padding.

Adding Small Data

If the amount of data to add is not very large, then distributing it evenly can lead to poor block space utilization because at least one block is used on each machine in the cluster. However, the blocks might be mostly padding and contain little data. In these cases, the INNAMEONLY option can be used. This option sends the data to the Hadoop NameNode only. The blocks are distributed according to the default strategy used by Hadoop. The distribution is likely to be unbalanced, but the performance is not reduced because the data set is not large.
PROC OLIPHANT Statement

Enables adding, removing, and managing SASHDAT files in Hadoop Distributed File System (HDFS).

Syntax

PROC OLIPHANT HOST=root-node INSTALL="grid-install-path"
</PATH='HDFS-path'> </LOGUPDATE> <INNAMEONLY>;

<table>
<thead>
<tr>
<th>Statement</th>
<th>Task</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADD</td>
<td>Add a data set.</td>
<td>Ex. 1</td>
</tr>
<tr>
<td>REMOVE</td>
<td>Remove a data set.</td>
<td>Ex. 1</td>
</tr>
<tr>
<td>DETAILS</td>
<td>Query data set metadata.</td>
<td>Ex. 2</td>
</tr>
</tbody>
</table>

Required Arguments

HOST=

specifies the host name or IP address of the grid host. This is the machine that is running the Hadoop NameNode that is provided by SAS High-Performance Deployment of Hadoop. If you do not specify the HOST= option, it is determined from the GRIDHOST= environment variable.

Alias NAMENODE=

INSTALL=

specifies the path to the TKGrid software on the grid host. If you do not specify this option, it is determined from the GRIDINSTALLLOC= environment variable.

Alias INSTALLLOC=
**Oliphant Options**

**PATH=**
specifies the directory in HDFS to use. This value can be overridden with a PATH= option on an ADD, REMOVE, or DETAILS statement.

Alias **OUTDIR=**

**LOGUPDATE**
provides progress messages in the SAS log about the data transfer to the grid host. The data transfer size is not necessarily the same as the block size that is used to form blocks in HDFS. The data transfer size is selected to optimize network throughput.

Alias **LOGNOTE**

**INNAMEONLY**
specifies that data identified in an ADD statement should be sent as a single block to the Hadoop NameNode for distribution. This option is appropriate for smaller data sets.

Restriction The BLOCKSIZE= option is ignored.

**ADD Statement**

Adds a data set to HDFS as a SASHDAT file.

**Example:**  "Example 1: Adding and Removing Files in HDFS" on page 258

**Syntax**

ADD libref.member-name <add-statement-options>;

**Add Statement Options**

**BLOCKSIZE=**
specifies the block size to use for distributing the data set. Suffix values are B (bytes), K (kilobytes), M (megabytes), and G (gigabytes). The default block size is 32M.

Alias **BLOCK=**

**COPIES=**
specifies the number of replications to make for the data set (beyond the original blocks). The default value is 2 when the INNAMEONLY option is specified and otherwise is 1. Replicated blocks are used to provide fault tolerance for HDFS. If a machine in the cluster becomes unavailable, then the blocks needed for the SASHDAT file can be retrieved from replications on other machines.

Alias **COPY=**

**(input-data-set-options)**
specifies any data set options to apply to the input data set.
Typically, you specify a description for the data set with the LABEL= option. The LABEL= option assigns the description to the SASHDAT file when the data set is stored in HDFS. The LABEL= option is used to override the label that is associated with the data set. Enclose the options in parentheses.

**PATH='HDFS-path'**
specifies the directory in HDFS in which to store the SASHDAT file. The value is case sensitive. The filename for the SASHDAT file that is stored in the path is always lowercase.

Alias OUTDIR=

**REPLACE**
specifies that the SASHDAT file should be overwritten if it already exists.

Alias OVERWRITE

**VARS=(<variables>)**
specifies the variables from the input data set to include in the SASHDAT file that is stored to HDFS. The default action is to include all the variables from the input data set.

---

**REMOVE Statement**

Removes a SASHDAT file from HDFS.

**Example:**

“Example 1: Adding and Removing Files in HDFS” on page 258

---

**Syntax**

REMOVE SASHDAT-file PATH='HDFS-path';

**Required Arguments**

**SASHDAT-file**
specifies the name of the file to remove. Do not specify a fully qualified HDFS path. Do not enclose the value in quotation marks. Do not include the SASHDAT filename suffix. The name is converted to lowercase and the filename of the SASHDAT file in HDFS must also be in lowercase.

**PATH='HDFS-path'**
specifies the HDFS directory.

Alias OUTDIR=

---

**DETAILS Statement**

Queries information about the data in a SASHDAT file.

**Example:**

“Example 2: Querying File Details from HDFS” on page 259
Syntax

DETAILS <details-statement-options>;

Details Statement Options

ALL
includes the number of rows for each SASHDAT file in the SAS output.

Alias ALLFILES

COLUMN
includes the column attributes for the specified SASHDAT file in the SAS output.

Alias COLUMNINFO

FILE=SASHDAT-file
specifies the name of the SASHDAT file to use. Do not specify a fully qualified HDFS path. Do not enclose the value in quotation marks. Do not include the SASHDAT filename suffix. The name is converted to lowercase and the filename of the SASHDAT file in HDFS must also be in lowercase.

Alias TABLE=

PATH='HDFS-path'
specify the fully qualified HDFS directory name.

Alias OUTDIR=

RECURSIVE
when FILE= is not specified, the details are reported for all SASHDAT files that are found in the path and child directories.

ROWCOUNT
includes the number of observations in the specified SASHDAT file.

Examples: OLIPHANT Procedure

Example 1: Adding and Removing Files in HDFS

Details
This PROC OLIPHANT example demonstrates adding and removing data sets to HDFS. One data set is added and a different SASHDAT file is removed.

Program

libname hrdata */data/hr/2011*;

proc oliphant host="grid001.example.com" install="/opt/TKGrid";
Program Description

1. The PROC OLIPHANT statement uses the HOST= and INSTALL= options to identify the SAS High-Performance Deployment of Hadoop cluster to use.
2. The ADD statement copies the EMPS data set to the HDFS path. The data set is distributed in blocks of 16 megabytes each. If an emps.sashdat file for the EMPS data set already exists, it is replaced.
3. This ADD statement includes a LABEL= option for the input data set.
4. The REMOVE statement deletes the salary.sashdat file from the HDFS path.

Example 2: Querying File Details from HDFS

Details

This PROC OLIPHANT example demonstrates how to query the details of SASHDAT files.

Program

```
proc oliphant host="grid001.example.com" install="/opt/TKGrid";
   details path="/sasdata/2011/" recursive;
   details file=emps path="/sasdata/2011/" column;
run;
```

Program Description

1. The PROC OLIPHANT statement uses the HOST= and INSTALL= options to identify the SAS High-Performance Deployment of Hadoop to use.
2. The table information details for all SASHDAT files in the /sasdata/2011 directory and any subdirectories are displayed.
3. The column information for the emps.sashdat file is displayed.
Chapter 8
RECOMMEND Procedure

Overview: RECOMMEND Procedure

Purpose of the RECOMMEND Procedure

PROC RECOMMEND is an interactive procedure that executes the statements within RUN blocks, similar to the IMSTAT procedure. All tasks in a SAS LASR Analytic Server-based recommender system can be performed from the RECOMMEND procedure.

You use the RECOMMEND procedure to perform the following tasks:

• start a recommender system in a SAS LASR Analytic Server
• connect to an existing recommender system
• interact with one or more recommender systems in the same SAS LASR Analytic Server
• filter data using a WHERE clause
• remove a recommender system from a SAS LASR Analytic Server
• populate a recommender system with content-based data for use with the cluster-based method or with rating data for collaborative filtering

Working with Recommender Systems
• add methods to a recommender system
• remove methods from a recommender system
• define and optimize ensemble methods
• hold a subset of the data to use for model validation
• obtain recommendations for users

Working with Recommender Systems

Identify a recommender system in the SAS LASR Analytic Server by a two-level name, similar to a LIBNAME.MEMBER construct. The name, which is case-insensitive, must be unique among the recommender systems in the SAS LASR Analytic Server. It is possible for an in-memory table and a recommender system to have the same name, because the table and application name spaces are separate. However, as a best practice, avoid using the same names to prevent confusion for users.

You can specify the recommender system that you want to work with through options in the PROC RECOMMEND statement or through options on other statements. When you specify the recommender system in the PROC RECOMMEND statement, the procedure uses that system as the default during execution. You can explicitly call other recommender systems in statements within the RECOMMEND procedure. If no application name is specified, the default recommender system is RECOM.SYSTEM.

You add tables to a recommender system to identify the user content, item content, and ratings. Typically, you load a user-item-ratings table into the recommender system. The SAS LASR Analytic Server derives a number of internal tables from the tables that you provide. The subsequent operations of the recommender system use the internal tables only. Make sure the SAS LASR Analytic Server has sufficient resources to build the recommender system and its associated tables. You can drop the user-item-ratings table after the recommender system has been set up.

Access to a recommender system is governed by signature files in the same way that those files are used to control access to SAS LASR Analytic Server tables. However, a recommender system contains multiple derived tables. For example, granting Read access to a recommender system implies Read access to all associated tables for that system.

Syntax: RECOMMEND Procedure

PROC RECOMMEND <option(s)>;
   ADD recommender-system <\ option(s)>;
   ADDTABLE table <\ option(s)>;
   INFO <\ option(s)>;
   METHOD method-name <\ option(s)>;
   PREDICT <\ option(s)>;
   REMOVE recommender-system <\ option(s)>;

<table>
<thead>
<tr>
<th>Statement</th>
<th>Task</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADD</td>
<td>Add a recommender system.</td>
<td>Ex. 1, Ex. 2</td>
</tr>
</tbody>
</table>
PROC RECOMMEND Statement

invokes the RECOMMEND procedure.

Syntax

PROC RECOMMEND <option(s)>;

Optional Arguments

**SYSTEM=**recommender-system

specifies the name of the recommender system in the SAS LASR Analytic Server that the procedure works with. Specify a two-level name, similar to a LIBNAME.MEMBER construct.

To work with an existing recommender system, the name that you specify identifies the application in the SAS LASR Analytic Server. When you create a new recommender system, the specified value becomes the name of that new system.

**Alias**

**RECOM=**

**Default**

**RECOM.SYSTEM**

**HOST=**host-name

specifies the host name of the SAS LASR Analytic Server.

**Default**

GRIDHOST value or the machine name of the SAS session if GRIDHOST is not specified.

**IMMEDIATE**

specifies that the RECOMMEND procedure executes each statement individually rather than running all the procedure statements in a single block at the end of the procedure code.

**NOPRINT**

suppresses printed output and SAS ODS tables.

**PORT=**port-number

specifies the number of the port that the SAS LASR Analytic Server uses to listen for requests.

**Default**

LASRPORT macro variable if that variable is set.
ADD Statement

The ADD statement is used to add a recommender system to the SAS LASR Analytic Server. If you do not provide a system name, the recommender system for the PROC RECOMMEND statement is used.

**Requirement:** The recommender system that you specify must not already exist in the SAS LASR Analytic Server.

**Syntax**

ADD <recommender-system> / <option(s)>;

**Optional Argument**

*recommender-system*

specifies the name of the recommender system to create.

**Default** Value provided for the SYSTEM= or RECOM= option in the PROC RECOMMEND statement. If no value is specified, the default value RECOM.SYSTEM is used.

**ADD Statement Options**

You can specify the following optional arguments after the slash (/) in the ADD statement. These options apply to the recommender system that is used by the ADD statement.

**DATAFILTER=**"expression"

specifies an optional WHERE clause for the recommender system. All of the data is filtered by this WHERE clause.

**DESCENDING=**variable-name

specifies which variables of the ORDERBY= list are used with descending sort order. Specifying the DESCENDING= option by itself has no effect. The option is specified in addition to the ORDERBY= option.

Alias **DESC=**

**ITEM=**column-name

specifies the name of the column that contains item identification in the in-memory tables.

**LABEL=**'string'

specifies a label that you can use to identify the recommender system. The label is returned in output from the SAS LASR Analytic Server.

**ORDERBY=**(variable-list)

specifies one or more variables to use for sorting ratings for a user or for an item in the derived tables. For example, you can specify a timestamp variable to arrange ratings in chronological order for each user or for each item. Separate multiple variables with a space.
PERMISSION=mode
specifies the permission setting for accessing the recommender system. The mode 
value is expressed as an integer, such as 755. The mode corresponds to the mode 
values that are used for UNIX file access permissions.

Alias PERM=

Default Permissions are set according to the UNIX file access permissions for the 
SAS LASR Analytic Server process.

RATING=column-name
specifies the name of the column that contains ratings in the in-memory tables.

Alias RATE=column-name

USER=column-name
specifies the name of the column that contains user identification in the in-memory 
tables.

ADDTABLE Statement
The ADDTABLE statement specifies a table for a recommender system that the SAS LASR Analytic Server 
uses to derive internal tables. The ADDTABLE statement is used most often to identify a table that 
contains user-item-ratings information.

Syntax
ADDTABLE libref.member-name<option(s)>;

Required Argument
table
specifies a source table that the SAS LASR Analytic Server uses to generate internal 
tables. Specify the table name in Lasrlib.Name format, where the Lasrlib value is the 
libref of the SAS LASR Analytic Server on which the recommender system is 
defined.

Requirement The table that you specify must already be in memory.

ADDTABLE Statement Options
You can specify the following optional arguments after the slash (/) in the ADDTABLE 
statement. These options apply to the recommender system that is used by the 
ADDTABLE statement.

SYSTEM=recommender-system
specifies the name of the recommender system in the SAS LASR Analytic Server 
that the procedure works with. Specify a two-level name, similar to a 
LIBNAME.MEMBER construct.

To work with an existing recommender system, the name that you specify identifies 
the application in the SAS LASR Analytic Server. When you create a new 
recommender system, the specified value becomes the name of that new system.

Alias RECOM=
specifies what type of information the table contains. A table of TYPE=ITEM contains the item column that was specified in the ADD statement and content information for the items. A table of TYPE=USER contains the user column that was specified in the ADD statement and content information for the users. A table of TYPE=RATING contains user-item-ratings information.

Default  
RATING

VARS=(variable-list)  
specifies a list of one or more variables to transfer to the internal tables that the SAS LASR Analytic Server derives from the in-memory table. For example, if you are adding a table of item content, then only a subset of the variables are useful to the recommender system. List the useful variables in the VARS= option.

Default  
Transfer all variables

**INFO Statement**

The INFO statement requests information about one or more recommender applications in the SAS LASR Analytic Server.

**Syntax**

INFO < / option(s)>;

**Optional Arguments**

You can supply the following options in the INFO statement after the slash (/):

**ALL**  
requests a list of all recommender systems in the SAS LASR Analytic Server. When you use the ALL option, an additional RECOM= or SYSTEM= option is ignored.

**METHODS=("method-name1" <, "method-name2" ...>)**  
requests information about one or more methods that are registered with the recommender system. Specify each method with a quoted string, and separate multiple methods with commas.

**SYSTEM= recommender-system**  
specifies the name of the recommender system in the SAS LASR Analytic Server that the procedure works with. Specify a two-level name, similar to a LIBNAME.MEMBER construct.

To work with an existing recommender system, the name that you specify identifies the application in the SAS LASR Analytic Server. When you create a new recommender system, the specified value becomes the name of that new system.

**Alias**  
RECOM=

**Default**  
RECOM.SYSTEM
METHOD Statement

The METHOD statement adds a method for computing recommendations for a recommender system. You can use the METHOD statement to specify details for a method definition, rather than using the default settings for that method.

Syntax

METHOD method-name <\ option(s)>;

Required Argument

method-name

specifies the name of the method to add to a recommender system. The method name can be one of the following values:

- **AVERAGE | AVE | AVG**: a default method that is used to produce recommendations for users that have insufficient information in the recommender system. For example, a method might require that at least two ratings are on record for a user. If that is not the case, a request for a recommendation is provided with the AVG method.
- **SLOPEONE | SLOPE1**: a simple regression-based method.
- **NEAREST | KNN**: a \( k \)-nearest-neighbor method that is based on measures of association between items or users. This method is also called a collaborative filter.
- **SVD**: a recommender method that is based on a singular-value decomposition of a user-item-ratings matrix.
- **ENSEMBLE**: a collection of other methods that you specify.
- **ARM**: a method that performs associative rule mining (ARM).
- **CLUSTER**: a cluster-based method that uses item or user profiles. Items or users are clustered. Then the similarity information between items or users for each cluster is computed to make recommendations.

Any method that has not already been defined for the SAS LASR Analytic Server, except for the CLUSTER and ARM methods, is created when a prediction request is processed. For example, suppose that you request a prediction with the \( k \)-nearest-neighbor method, and that method has not been added to the recommender system. The SAS LASR Analytic Server creates that method with default parameters and adds the method to the recommender system at that time.

Use the METHOD statement to add a method with explicitly defined parameters if you do not want to use the defaults. You must use the METHOD statement to add either the ARM or CLUSTER method, because those methods cannot be created using default values. For more information, see “ARM Statement” on page 45 or “CLUSTER Statement” on page 63.
Optional Arguments

**DATAFILTER=expression**
specifies an optional WHERE clause for each method. All of the data is filtered by this WHERE clause.

**DETAILS**
requests that additional details are provided for the numerically intensive SVD and ensemble methods.

**LABEL='string'**
specifies a label by which the method can be identified. A label is important if you have multiple instances of a method definition (with different parameter values) in the recommender system.

**FCONV=r**
specifies a relative function convergence criterion for the numerical optimization in SVD and ensemble methods.

**GCONV=r**
specifies a relative gradient convergence criterion for the numerical optimization in SVD and ensemble methods.

**MAXITER=n**
specifies the maximum number of iterations for the numerical optimization in SVD and ensemble methods.

Default 1 (a one-step update)

**MAXFEVAL=n**
specifies the maximum number of function evaluations for the numerical optimization in SVD and ensemble methods.

**SEED=n**
specifies the seed for random number generation in SVD and ensemble methods.

**SYSTEM=recommender-system**
specifies the name of the recommender system in the SAS LASR Analytic Server that the procedure works with. Specify a two-level name, similar to a LIBNAME.MEMBER construct.

To work with an existing recommender system, the name that you specify identifies the application in the SAS LASR Analytic Server. When you create a new recommender system, the specified value becomes the name of that new system.

Alias **RECOM=**

Default **RECOM.SYSTEM**

Options for the SLOPEONE Method

**HOLD=n**
specifies the number of ratings to hold for users that are selected by the WITHHOLD= option. The specified number of ratings are selected at random to be held in a validation data set, which is a subset of the original data set.

Typically, you specify a positive number for the HOLD= option. However, you can specify a negative number, which indicates that all ratings should be held in the validation data set except for the specified number of ratings. For example, HOLD=–2 means that all ratings but two should be held in the validation data set.
WITHHOLD=\(r\)  
specifies a relative percentage of users whose ratings are included in a validation data set, which is a subset of the original data set. For example, WITHHOLD=0.1 indicates that 10% of users should be selected at random. A portion of the selected users’ ratings are held in the validation data set. The number of ratings to select is specified by the HOLD= option.

Range 0–1, exclusive

Options for the KNN Method

**HOLD=\(n\)**  
specifies the number of ratings to hold for users that are selected by the WITHHOLD= option. The specified number of ratings are selected at random to be held in a validation data set, which is a subset of the original data set.

Typically, you specify a positive number for the HOLD= option. However, you can specify a negative number, which indicates that all ratings should be held in the validation data set except for the specified number of ratings. For example, HOLD=–2 means that all ratings but two should be held in the validation data set.

Default 1

**NEAREST=\(k\)**  
specifies the parameter \(k\) for a \(k\)-nearest-neighbor method. Only the \(k\) nearest neighbors are considered in deriving a recommendation for a particular user.

Alias **K=**

**NONNEGATIVE**  
requests that only positive associations are used when computing a neighborhood in a \(k\)-nearest-neighbor method.

Alias **POSITIVE**

**PREFILTER=NONE**  
**PREFILTER=TOP(\(n\))**  
**PREFILTER=THRESHOLD(\(r\))**  
specifies the type of prefiltering to apply when computing a neighborhood. If you specify PREFILTER=TOP(\(n\)), then a list of only the \(n\) nearest neighbors and their similarities are kept. If you specify PREFILTER=THRESHOLD(\(r\)), then the list of nearest neighbors includes items or users with similarities that exceed the threshold value \(r\). If you specify PREFILTER=NONE, then neighborhoods are formed based on all similarities.

Default TOP(10)
SIMILARITY=COSINE | COS | CV
SIMILARITY=CORR | PEARSON | PC
SIMILARITY=ADJCOS | AC

specifies the similarity measure that is used in $k$-nearest-neighbor collaborative filtering. If you specify SIMILARITY=COSINE (or COS or CV), then the cosine measure is the similarity measure. If you specify SIMILARITY=CORR (or PEARSON or PC), then the Pearson’s correlation coefficient, or product-moment correlation, is the similarity measure. If you specify SIMILARITY=ADJCOS (or AC), then the adjusted cosine measure is the similarity measure. For more information, see “How Similarity Measures Are Calculated” on page 273.

WITHHOLD=$r$

specifies a relative percentage of users whose ratings are included in a validation data set, which is a subset of the original data set. For example, WITHHOLD=0.1 indicates that 10% of users should be selected at random. A portion of the selected users’ ratings are held in the validation data set. The number of ratings to select is specified by the HOLD= option.

Range 0–1, exclusive

Options for the SVD Method

BINARY=$n$

specifies a rule to generate a binary rating. If a numeric rating exceeds $n$, then the binary rating is set to 1. Otherwise, the binary rating is set to 0.

BINALPHA=$m$

specifies a weighting factor for the squared errors in the loss function of the matrix factorization.

HOLD=$n$

specifies the number of ratings to hold for users that are selected by the WITHHOLD= option. The specified number of ratings are selected at random to be held in a validation data set, which is a subset of the original data set.

Typically, you specify a positive number for the HOLD= option. However, you can specify a negative number, which indicates that all ratings should be held in the validation data set except for the specified number of ratings. For example, HOLD=–2 means that all ratings but two should be held in the validation data set.

Default 1

Interaction The HOLD= option is ignored if the WITHHOLD= option is not also specified.

LOSS=SE
LOSS=SEREG
LOSS=SEWREG
LOSS=KL | ENTROPY

specifies the loss function for the matrix factorization. The LOSS=SE option indicates that the squared-error function is the loss function. The LOSS=SEREG and LOSS=SEWREG options are modifications of the squared-error loss function that include regularization terms in matrix norms or weighted matrix norms, respectively. Weighted regularization terms are weighted by $\lambda$, and you set the value of this parameter with the LAMBDA= option. The LOSS=KL (or ENTROPY) option indicates that the Kullback-Leibler divergence, or relative entropy, is the loss function.
LAMBDA=\lambda
    specifies the regularization factor for the loss functions.

Applies to   LOSS=SEREG or LOSS=SEWREG

TECHNIQUE=LBFGS

TECHNIQUE=ALS
    specifies the optimization method for the singular-value decomposition. The
    TECHNIQUE=LBFGS option indicates a limited-memory Broyden-Fletcher-Goldfarb-Shanno
    (BFGS) optimization method. This method is often used for
    solving neural network problems. The TECHNIQUE=ALS option indicates an
    alternating least squares optimization method.

WITHHOLD=r
    specifies a relative percentage of users whose ratings are included in a validation
    data set, which is a subset of the original data set. For example, WITHHOLD=0.1
    indicates that 10% of users should be selected at random. A portion of the selected
    users’ ratings are held in the validation data set. The number of ratings to select is
    specified by the HOLD= option.

Range   0–1, exclusive

Options for Ensemble Method

CONSTRAINT
    restricts the weights in the ensemble to lie between 0 and 1.

HOLD=n
    specifies the number of ratings to hold for users that are selected by the
    WITHHOLD= option. The specified number of ratings are selected at random to be
    held in a validation data set, which is a subset of the original data set.

Typically, you specify a positive number for the HOLD= option. However, you can
    specify a negative number, which indicates that all ratings should be held in the
    validation data set except for the specified number of ratings. For example,
    HOLD=-2 means that all ratings but two should be held in the validation data set.

Default   1

Interaction  The HOLD= option is ignored if the WITHHOLD= option is not also
    specified.

METHODS=("method1", "method2" <,"method3" ...>)
    specifies the methods that participate in the ensemble. Enclose each method in
    quotation marks, and separate multiple values with a comma.

Default   All methods except the AVERAGE method.

Restriction  The AVERAGE method is not part of any ensemble.

WITHHOLD=r
    specifies a relative percentage of users whose ratings are included in a validation
    data set, which is a subset of the original data set. For example, WITHHOLD=0.1
    indicates that 10% of users should be selected at random. A portion of the selected
    users’ ratings are held in the validation data set. The number of ratings to select is
    specified by the HOLD= option.

Range   0–1, exclusive
Options for the Cluster Method

**BUBMAXPTS=**\(n\)

specifies the maximum number of points in each bubble. This number must exceed the value of the BUBMINPTS= option.

**BUBMINPTS=**\(n\)

specifies the minimum number of points in each bubble.

Default 1

**CLUSTINFO**

generates the temporary table that contains the cluster results for each user or item.

**CLUSTVARS=(variable-list)**

lists the variables to use with the CLUSTER method.

**CLUSTERTECH=KMEANS**

**CLUSTERTECH=DBSCAN**

specifies the clustering technique.

Default KMEANS

**CONV=**\(c\)

specifies the convergence criterion \(c\) for the \(k\)-means analysis. When the relative change in WCSS between successive iterations is less than \(c\), the analysis is presumed to have converged.

Default 0.00001

**DIST=EUC | SQUAREDEUC | MANHATTAN | MAXIMUM | COSINE | JACCARD | HAMMING**

specifies the distance measure that is used in the clustering method. The \(k\)-means method uses DIST=EUC.

Applies to CLUSTERTECH=DBSCAN

**DMAX=**\(v\)

specifies the maximum diameter of bubbles with the given distance measure.

Default 0

**EPS=**\(r\)

specifies the distance value for neighborhood querying. For more information, see “CLUSTER Statement” on page 63.

Applies to CLUSTERTECH=DBSCAN

**INITMETHOD=FORGY | RAND | AVG**

specifies the method for obtaining the initial estimate of cluster assignment. For more information, see “CLUSTER Statement” on page 63.

Alias **INIT=**

**MINPTS=**\(n\)

specifies the minimum number of points that are required in one cluster.

Applies to CLUSTERTECH=DBSCAN
NOCASE
specifies that the comparisons between terms and the values of character variables are case insensitive. By default, comparisons are case-sensitive.

NOIDF
specifies that only the term frequency is used to construct the vectors and that inverse document frequency is not used.

NONORM
specifies that the TF-IDF vectors are not normalized.

NREP=\(k\)
specifies the number of representative points for each bubble.

Default 1

NUMCLUSTERS=\(k\)
specifies the number of clusters for the \(k\)-means analysis.

Alias NUMCLUS=

Default 2

SAVETERMS
saves the TF-IDF vectors in the temporary table when the CLUSTINFO option is enabled.

TERMS=("term1" <, "term2"...>)
specifies terms that are used to compute term frequency. Each string represents one term. For more information, see “CLUSTER Statement” on page 63.

TERMDATA=table-name
specifies an in-memory table in the server that contains the term list. For more information, see “CLUSTER Statement” on page 63.

TOKENS=("token1" <, "token2"...>)
specifies the tokens that separate terms when scanning character variables. For more information, see “CLUSTER Statement” on page 63.

TOKENDATA=table-name
specifies an in-memory table in the server that contains the tokens list.

TYPE=ITEM | USER
specifies which type of profile is used for the CLUSTER method. The CLUSTER method that uses a user profile table cannot be used in the ensemble model with other methods.

Requirement The user or item table must be added into the recommender system.

Details

How Similarity Measures Are Calculated
Similarity measures are used to determine the \( k \) nearest neighbors of an item or a user for the KNN method. You can select cosine, adjusted cosine, or Pearson’s correlation coefficient to measure the similarity between items or users.

Suppose that \( r_{ui} \) is the rating of user \( u \) for item \( i \). Then the user-based similarity measures between users \( u \) and \( v \) are computed as follows:
Cosine \((u, v)\) = \[ \frac{\sum_{i \in I_{uv}} u_i v_i}{\sqrt{\sum_{i \in I_u} u_i^2 \sum_{i \in I_v} v_i^2}} \]

Corr \((u, v)\) = \[ \frac{\sum_{i \in I_{uv}} (u_i - \bar{r}_u)(v_i - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (u_i - \bar{r}_u)^2 \sum_{i \in I_{uv}} (v_i - \bar{r}_v)^2}} \]

In these expressions, \(I_{uv}\) denotes the set of items that have been rated by user \(u\) and user \(v\). The value \(\bar{r}_u\) is the average rating by user \(u\) across all items that she rated.

In an item-based recommender system, the similarity measures associate ratings for items across the set of users who rated items \(i\) and \(j\). Denote this set as \(U_{ij}\). The Pearson correlation measure between items \(i\) and \(j\) are calculated as follows:

Corr \((i, j)\) = \[ \frac{\sum_{u \in U_{ij}} (u_i - \bar{r}_u)(u_j - \bar{r}_j)}{\sqrt{\sum_{u \in U_{ij}} (u_i - \bar{r}_u)^2 \sum_{u \in U_{ij}} (u_j - \bar{r}_j)^2}} \]

AC \((i, j)\) = \[ \frac{\sum_{u \in U_{ij}} (u_i - \bar{r}_u)(u_j - \bar{r}_u)}{\sqrt{\sum_{u \in U_{ij}} (u_i - \bar{r}_u)^2 \sum_{u \in U_{ij}} (u_j - \bar{r}_u)^2}} \]

The equations for item similarity measures reflect that the differences in rating scales among users are typically more pronounced than the differences in ratings for individual items. That is, an item might get low scores from most users, but the ranges of scores among users often vary widely.

---

**PREDICT Statement**

The PREDICT statement generates recommendations (predictions) for one or more users.

**Syntax**

```
PREDICT <option(s)>;
```

**Optional Arguments**

You can specify the following options in the PREDICT statement after the slash (/):

- `LABEL='string'`
  
  specifies the label that is assigned to the desired method of computing recommendations.
METHOD=AVERAGE | AVE | AVG
METHOD=NEAREST | KNN
METHOD=SLOPEONE | SLOPE1
METHOD=SVD
METHOD=ENSEMBLE
METHOD=ARM
METHOD=CLUSTER

specifies the method to use for computing recommendations. If the requested method
is not yet defined for the recommender system, then the SAS LASR Analytic Server
adds the method with default parameters and computes the recommendation. For
more information, see “METHOD Statement” on page 267.

NRECOMM=n
specifies the upper limit for the number of recommendations that are returned per
user.

Alias NUM=
Default 10

SYSTEM=recommender-system
specifies the name of the recommender system in the SAS LASR Analytic Server
that the procedure works with. Specify a two-level name, similar to a LIBNAME.MEMBER construct.

To work with an existing recommender system, the name that you specify identifies
the application in the SAS LASR Analytic Server. When you create a new
recommender system, the specified value becomes the name of that new system.

Alias RECOM=
Default RECOM.SYSTEM

OUT=SAS-data-set
specifies a SAS data set that stores recommendations.

Alias OUTDATA=

USERDATA=table-name
specifies an in-memory table that contains the user IDs for which you want
recommendations.

USERLIST=("userID1" <"userID2" ...>
specifies the user IDs for which you want recommendations. This is a convenient
format if you want recommendations for only a small number of users.

Alias USERS=

Interaction If a value is also specified for USERDATA, then only the USERDATA
table is used.

REMOVE Statement

Removes a recommender system from the SAS LASR Analytic Server, or removes a method from a
recommender system.
Syntax

REMOVE <recommender-system> </opti(s)>;

Optional Argument

recommender-system

specifies the name of the recommender system to remove from the SAS LASR Analytic Server. To remove the currently active recommender system, issue the REMOVE statement with no additional arguments or options.

Default Value provided for RECOM= or SYSTEM= in the PROC RECOMMEND statement. If no value was specified, the default value RECOM.SYSTEM is used.

Options for the REMOVE Statement

METHOD=NEAREST | KNN
METHOD=SLOPEONE | SLOPE1
METHOD=SVD
METHOD=ENSEMBLE
METHOD=ARM
METHOD=CLUSTER

specifies the method that you want to remove from a recommender system. If the method is included in an ensemble, then the ensemble is removed also.

LABEL='string'

specifies the label of the method to remove.

Examples: RECOMMEND Procedure

Example 1: Recommendations from Explicit Ratings

Details

Problem Description

This example draws on data that is derived from online movie viewing companies. A company wants to offer its customers recommendations of movies that they might like. These recommendations are based on ratings that are provided by users. The following table contains an example of a user-item-ratings matrix that online movie viewing companies might use.

Table 8.1 Sample Movie Ratings from Customers

<table>
<thead>
<tr>
<th>Customers</th>
<th>Movie 1</th>
<th>Movie 2</th>
<th>Movie 3</th>
<th>Movie 4</th>
<th>Movie 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>4</td>
<td></td>
<td>4</td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>
In the sample data, customers rate the movies that they have seen using a 1–5 scale, where 1 is the lowest rating and 5 is the highest rating. This table represents a user-item-ratings matrix. Blank cells correspond to movies that a customer has not rated.

In practice, the matrix would be even more sparse than the sample data, as the typical customer rates only a very small fraction of all available movies. The goal of the recommender system is to predict ratings for all of the blank cells. Would User 1 like Movie 4? From the sample above, there is very little data available. However, with much more data, we might observe that User 1 and User 3 have similar taste in movies. Therefore, we could conclude that User 1 would also give Item 4 a high rating, based on the information that we already have from User 3. In a real-world situation, it is not necessary to predict every blank entry in a utility matrix. The system is required to supply only a few suggestions that a customer would rate highly.

### Example Data

This example uses the MovieLens data set (1 MB) that was developed by the GroupLens project at the University of Minnesota. The data that is displayed was downloaded in February 2014 from the GroupLens website.

**Note:** Per University of Minnesota guidelines, you cannot use this data for any commercial or revenue-bearing purpose without first obtaining permission from a faculty member of the GroupLens Research Project.

Before invoking the RECOMMEND procedure, we can print a part of the tables that are included in the recommender system by invoking the IMSTAT procedure.

```sas
proc imstat;
  table MYlasr.movierating;
  fetch/format;
  run;

  table MYlasr.movieprofile;
  fetch/format;
  run;

  table MYlasr.userprofile;
  fetch/format;
  run;
quit;
```

The first FETCH statement prints a portion of the MovieRating table. The table contains rating information made by users (customers) about items (movies).
Output 8.1  Sample Data from the MovieRatings Table

The second and third FETCH statements print a portion of the MovieProfile and UserProfile tables. These tables contain information about each movie and user (customer), respectively.

Output 8.2  Sample Profile Data
Program

```
proc recommend port=&portNumber recom = rs.movielens;
  add rs movimiento / item = itemid user = userid rating = rating;
  addtable MYlasr.movierating / recom = rs.movielens type = rating
          vars=(itemid userid rating);
  addtable MYlasr.movieprofile / recom = rs.movielens type = item;
  addtable MYlasr.userprofile / recom = rs.movielens type = user;
run;

method knn / label = "knn" k = 20 positive similarity = pc seed = 1234;
run;

method slope1 / label = "slope1";
run;

method svd / factors = 20 label = "svd" fconv = 1e-3 gconv = 1e-3
          maxiter = 100 seed = 1234 MAXFEVAL = 5000 function=L2
          lamda = 0.2 technique = lbfgs;
run;

method ensemble / methods = ("svd","knn") label = "ensemble" details
          MAXFEVAL=5000 maxiter=100 seed=1234 hold=2
          withhold=0.1;
run;

predict / method = knn  label="knn" Num = 5
          users = ("1","33","478","2035");
run;

info;
run;

remove rs.movielens;
run;
```
Program Description

1. The ADD statement adds a recommender system, MovieLens, to the SAS LASR Analytic Server.

2. Three ADDTABLE statements add the MovieRating, MovieProfile, and UserProfile tables to the recommender system.

3. Each METHOD statement adds a method for computing recommendations to the recommender system. The methods KNN, SLOPE1, SVD, and ENSEMBLE were added with options specified for each method. For more information, see “METHOD Statement” on page 267.

4. The PREDICT statement generates five predictions for each specified user (1, 33, 478, and 2035).

5. The INFO statement requests information about all recommender systems on the SAS LASR Analytic Server. You can also specify a specific recommender system in the INFO statement to filter the results.

6. The REMOVE statement removes a recommender system from the server or removes a method.

METHOD Statement Output

Output 8.3  Output from the METHOD Statement Using the ENSEMBLE Option

![Output 8.3](image1)

Output 8.4  Output from the METHOD Statement Using the Details Option

![Output 8.4](image2)
PREDICT Statement Output

Output 8.5 Output from the PREDICT Statement

<table>
<thead>
<tr>
<th>User</th>
<th>Rank</th>
<th>Rating</th>
<th>itemID</th>
<th>year</th>
<th>title</th>
<th>category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5.3542</td>
<td>557.000000</td>
<td>1962.000000</td>
<td>Mamma Roma</td>
<td>Drama</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>5.0897</td>
<td>2503.000000</td>
<td>1998.000000</td>
<td>Apple, The (Sib)</td>
<td>Drama</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>5.0719</td>
<td>1178.000000</td>
<td>1957.000000</td>
<td>Paths of Glory</td>
<td>Drama</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>5.0651</td>
<td>2360.000000</td>
<td>1998.000000</td>
<td>Celebration, The (Festen)</td>
<td>Drama</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>5.0172</td>
<td>3245.000000</td>
<td>1964.000000</td>
<td>I Am Cuba (Soy Cuba/Ya Kuba)</td>
<td>Drama</td>
</tr>
<tr>
<td>33</td>
<td>1</td>
<td>4.6101</td>
<td>2905.000000</td>
<td>1962.000000</td>
<td>Sanjuro</td>
<td>Action</td>
</tr>
<tr>
<td>33</td>
<td>2</td>
<td>4.5831</td>
<td>3897.000000</td>
<td>2000.000000</td>
<td>Almost Famous</td>
<td>Comedy</td>
</tr>
<tr>
<td>33</td>
<td>3</td>
<td>4.5613</td>
<td>2503.000000</td>
<td>1998.000000</td>
<td>Apple, The (Sib)</td>
<td>Drama</td>
</tr>
<tr>
<td>33</td>
<td>4</td>
<td>4.5612</td>
<td>53.000000</td>
<td>1994.000000</td>
<td>Lamerica</td>
<td>Drama</td>
</tr>
<tr>
<td>33</td>
<td>5</td>
<td>4.5317</td>
<td>457.000000</td>
<td>1993.000000</td>
<td>Fugitive, The</td>
<td>Action</td>
</tr>
</tbody>
</table>
Example 2: Recommendations from Implicit Information

Details

In many situations, it is difficult to obtain explicit ratings that can be directly used to infer user preferences. Instead, we are provided with feedback such as purchase history,
browsing history, search patterns, or time spent on a website. This documentation presents two ways to infer user preferences from abundant implicit feedback to build a recommender system.

Let $U$ denote the number of users, and let $I$ denote the number of items. Let $f_u$ denote the frequency of user $u$ purchasing any item, and let $f_i$ denote the frequency of any user purchasing item $i$. Let $f_{ui}$ denote the frequency of item $i$ being purchased by user $u$.

Similar to the term frequency-inverse document frequency weight ($tf-idf$) in text mining, there are two methods that we can use to convert frequency $f_{ui}$ to ratings $r_{ui}$:

$$r_{ui} = \log (f_{ui} + 1) \log \left( \frac{U}{f_{i} + 1} \right)$$  \hspace{1cm} (1)

$$r_{ui} = \log (f_{ui} + 1) \log \left( \frac{I}{f_{u} + 1} \right)$$  \hspace{1cm} (2)

Equation 1 places larger weight on items that are purchased less frequently. Equation 2 places larger weight on users who purchase items less frequently.

**Program: Transform a Transaction Table into a Rating Table**

```plaintext
proc imstat;
  table MYlasr.salesfact; 1
    tableinfo / save = tabinf;
    store tabinf(1,3) = nObs;
  run;
  distinct / save = dtab;
  store dtab(1,2) = numItem;
  store dtab(2,2) = numUser;
  run;
  table MYlasr.salesfact; 2
    compute joinkey "joinkey = &userID || &itemID;";
  run;
  table MYlasr.salesfact(tempnames=(t1));
    summary t1 / groupby=(&userID &itemID) temptable tn=t1 3
      te="t1=1;" save=tab1;
    summary t1 / groupby=(&userID) temptable tn=t1
      te="t1=1;" save=tab2;
    summary t1 / groupby=(&itemID) temptable tn=t1
      te="t1=1;" save=tab3;
  run;
  store tab1(2,2) = freq_user_item; 4
  store tab2(2,2) = freq_user;
  store tab3(2,2) = freq_item;
  run;
  table MYlasr.&freq_user_item;
    schema &freq_user(&userID=&userID / prefix=UserTotal,_n_)
      &freq_item(&itemID=&itemID / prefix=ItemTotal,_n_) / mode=table;
  run;
  table MYlasr.&_templast_; 5
```

**Example 2: Recommendations from Implicit Information**
compute joinkey "joinkey = &userID || &itemID;";
run;
table MYlasr.salesfact;
  schema &_templast_(joinkey=joinkey / prefix=r,_n_
    UserTotal__N_ ItemTotal__N_);
run;
table MYlasr.&_templast_; 
  compute Rating_iuf "Rating_iuf = 
    log10(r__N_+1)*log10(&nObs/(r_ItemTotal__n_+1));";
run;
  compute Rating_iif "Rating_iif = 
    log10(r__N+1)*log10(&nObs/(r_UserTotal__n_+1));";
run;
  compute Rating_simple "Rating_simple = 
    Round((r__N_/r_UserTotal__n_)*10+1,1);";
run;
table MYlasr.&_templast_; 
  save path="/hps/rating_tfidf" copies=1 replace fullpath;
run;

Program Description
1. The STORE statements assign the number of observations, the number of distinct users, and the number of distinct items to macro variables.
2. The COMPUTE statement adds a permanent column by concatenating the user ID and the item ID values.
3. The SUMMARY statements with the GROUPBY= option produce descriptive statistics in the temporary column (t1). The results of the SUMMARY statements are saved to temporary tables.
4. The STORE statements assign the names of the three temporary tables to three macro variables, Freq_User_Item, Freq_User, and Freq_Item.
5. The SCHEMA statement joins the star tables Freq_User_Item, Freq_User, and Freq_Item and creates a new temporary table.
6. Generate the ratings table. By using equations 1 and 2, as described in “Details” on page 282, convert the frequency counts into ratings.
7. The SAVE statement saves the rating table directly into HDFS as a SASHDAT table for future use.

Program: Create a Recommender System with the Rating Table

proc recommend port=&lasrport recom = rs.DEPTSTORE;
  add rs.DEPTSTORE /item = item_sk user = household_sk rating = &rating;

  addtable MYlasr.MBA_rating_tfidf / recom = rs.DEPTSTORE 
    type = rating 
    vars=(item_sk household_sk &rating);

  addtable MYlasr.household / recom = rs.DEPTSTORE type = user;
  addtable MYlasr.item / recom = rs.DEPTSTORE type = item;
run;
method cluster / clusttech=kmeans numclus=50 dist=euc type=user 3 maxiter=10 label="clust" details
   terms=("Convenience", "Occasional", "Unk")
   tokens=" *") noidf seed=1234 clustinfo
   clustvars=(LIFESTYLE SEGMENT FAMILY PET);
run;

predict / users="11815911" method=cluster label="clust" num = 5;
run;

method svd / factors = 20 label = "svd_1" fconv = 1e-3 gconv = 1e-3 5 maxiter = 100 seed = 12314 MAXFEVAL = 5000
   function=L2 lambda = 0.2 technique = als;
run;

predict / method = svd label = "svd_1" Num = 5 6 users = ("11815911");
run;

quit;

Program Description
1. The ADD statement adds the Rs.DEPTSTORE system to the server.
2. The ADDTABLE statements add the tables to be analyzed to the system.
3. The METHOD statement using the CLUSTER method applies a model that first clusters users into several groups according to user profiles. Then, the method uses the nearest neighbor method to predict unknown ratings.
4. The PREDICT statement generates the top 5 ranked products for user 11815911 using the cluster method.
5. The next METHOD statement using the SVD method applies a model that is based on a singular-value decomposition of a user-item-ratings matrix.
6. The next PREDICT statement generates the top 5 ranked products for user 11815911 using the SVD method.

PREDICT Statement Output

Output 8.7 Output from the PREDICT Statement (Cluster Method)

<table>
<thead>
<tr>
<th>User</th>
<th>Rank</th>
<th>Rating</th>
<th>Item_sk</th>
<th>Category</th>
<th>Desc</th>
</tr>
</thead>
<tbody>
<tr>
<td>11815911</td>
<td>1</td>
<td>4.96</td>
<td>1278</td>
<td>produce</td>
<td>apples</td>
</tr>
<tr>
<td>11815911</td>
<td>2</td>
<td>4.87</td>
<td>1189</td>
<td>lifestyle</td>
<td>ice_cream</td>
</tr>
<tr>
<td>11815911</td>
<td>3</td>
<td>4.77</td>
<td>1342</td>
<td>produce</td>
<td>peppers</td>
</tr>
<tr>
<td>11815911</td>
<td>4</td>
<td>4.72</td>
<td>5592</td>
<td>butcher</td>
<td>steak</td>
</tr>
<tr>
<td>11815911</td>
<td>5</td>
<td>4.64</td>
<td>1359</td>
<td>canned</td>
<td>olives</td>
</tr>
</tbody>
</table>
### Output 8.8  Output from the PREDICT Statement (SVD Method)

<table>
<thead>
<tr>
<th>User</th>
<th>Rank</th>
<th>Rating</th>
<th>Item_sk</th>
<th>Category</th>
<th>Desc</th>
</tr>
</thead>
<tbody>
<tr>
<td>11815911</td>
<td>1</td>
<td>1.87</td>
<td>1342</td>
<td>produce</td>
<td>peppers</td>
</tr>
<tr>
<td>11815911</td>
<td>2</td>
<td>1.84</td>
<td>1278</td>
<td>produce</td>
<td>apples</td>
</tr>
<tr>
<td>11815911</td>
<td>3</td>
<td>1.79</td>
<td>1359</td>
<td>canned</td>
<td>olives</td>
</tr>
<tr>
<td>11815911</td>
<td>4</td>
<td>1.74</td>
<td>5592</td>
<td>butcher</td>
<td>steak</td>
</tr>
<tr>
<td>11815911</td>
<td>5</td>
<td>1.69</td>
<td>1189</td>
<td>lifestyl</td>
<td>ice_crea</td>
</tr>
</tbody>
</table>
Overview: VASMP Procedure

What Does the VASMP Procedure Do?

The VASMP procedure is used to list in-memory tables and perform administration of Non-distributed SAS LASR Analytic Server instances.

Syntax: VASMP Procedure

PROC VASMP <options>;
  SERVERINFO <option>;
  SERVERPARM <option>;
  SERVERTERM <options>;
  SERVERWAIT <options>;
  TABLEINFO <options>;
QUIT;

Example: Copying Tables from One Hadoop Installation to Another
PROC VASMP Statement

in a SAS LASR Analytic Server instance.

Syntax

PROC VASMP <options>;

Optional Arguments

DATA=libref.member-name
specifies the table to access from memory. The libref must be assigned from a SAS LASR Analytic Server engine LIBNAME statement.

IMMEDIATE
specifies that the procedure executes one statement at a time rather than accumulating statements in RUN blocks.

Alias SINGLESTEP

NOPRINT
This option suppresses the generation of ODS tables and other printed output in the VASMP procedure.

NOTIMINGMSG
When an action completes successfully, the VASMP procedure generates a SAS log message that contains the execution time of the request. Specify this option to suppress the message.

Alias NOTIME

SERVERINFO Statement

The SERVERINFO statement returns information about the SAS LASR Analytic Server.

Syntax

SERVERINFO </option>;

SERVERINFO Statement Options

HOST="host-name"
specifies the host name for the SAS LASR Analytic Server. Use this option with the PORT= option.

NORANKS
specifies to omit the list of host names for the worker nodes. This option reduces the output of the SERVERINFO option considerably for large environments.

PORT=number
specifies the port number for the SAS LASR Analytic Server. If you do not specify a PORT= value, then behavior of the SERVERINFO statement depends on whether an
in-memory table is active. If there is no active table, then the procedure attempts to connect to the server using the LASRPORT macro variable. If a table is active, the information is gathered for the server that is implied by the libref of the active table.

**SERVERPARM Statement**

The SERVERPARM statement enables you to change some global settings for the server if you have sufficient authorization. The user account that starts the server has privileges to modify server parameters.

**Syntax**

```
SERVERPARM <options>;
```

**SERVERPARM Statement Options**

**CONCURRENT=number**

specifies the number of concurrent requests that can execute in the server. Once the threshold is met, the requests are queued and then executed as the currently running requests complete.

Alias  NACCTIONS=

Default  20

**EXTERNALMEM=pct**

specifies the percentage of memory that can be allocated before the server stops transferring data to external processes such as external actions and the SAS High-Performance Analytics procedures. If the percentage is exceeded, the server stops transferring data.

Default  75

**HADOOPHOME="path"**

specifies the path for the HADOOP_HOME environment variable. Changing this variable is useful for migrating SASHDAT files from one Hadoop installation to another.

Setting the HADOOP_HOME environment variable is a server-wide change. All requests, by all users, for reading files from HDFS and saving files, use the specified HADOOP_HOME. This can cause unexpected results if users are not aware of the change.

*Note:* If you are using this option to migrate SASHDAT files, then consider starting a server for that exclusive purpose.

Alias  HADOOP=

**HOST="host-name"**

specifies the host name for the SAS LASR Analytic Server. Use this option with the PORT= option.

**PORT=number**

specifies the port number for the SAS LASR Analytic Server. If you do not specify a PORT= value, then behavior of the SERVERPARM statement depends on whether an in-memory table is active. If there is no active table, then the procedure attempts
to connect to the server using the LASRPORT macro variable. If a table is active, the information is gathered for the server that is implied by the libref of the active table.

**TABLEMEM=pct**

specifies the percentage of memory that can be allocated before the server rejects requests to add tables or append data. If the percentage is exceeded, adding a table or appending rows to tables fails. These operations continue to fail until the percentage is reset or the memory usage on the server drops below the threshold.

This option has no effect for non-distributed servers. For non-distributed servers, the memory limits can be controlled with the MEMSIZE system option.

*Note:* The specified *pct* value does not specify the percentage of memory allocated to in-memory tables. It is the percentage of all memory used by the entire machine that—if exceeded—prevents further addition of data to the server. The memory used is not measured at the process or user level, it is computed for the entire machine. In other words, if operating system processes allocate a lot of memory, then loading tables into the server might fail. The threshold is not affected by memory that is associated with SASHDAT tables that are loaded from HDFS.

```plaintext
Alias MEMLOAD=
Default 75
```

**SERVERTERM Statement**

The SERVERTERM statement sends a termination request to the server that is identified through the statement options. You must have sufficient authorization for this request to succeed.

---

**Syntax**

```plaintext
SERVERTERM <options>;
```

**SERVERTERM Statement Options**

- **HOST="host-name"**
  - specifies the host name for the SAS LASR Analytic Server. Use this option with the PORT= option.

- **PORT=number**
  - specifies the port number for the SAS LASR Analytic Server.

---

**SERVERWAIT Statement**

The SERVERWAIT statement suspends execution of the VASMP procedure until the server that it uses receives a termination request. This is useful for starting a non-distributed server from a batch program. This statement suspends the SAS session in which it is executed until the server stops or until an interrupt signal is received.
Syntax

SERVERWAIT <options>;

**SERVERWAIT Statement Options**

HOST="host-name"
specifies the host name for the SAS LASR Analytic Server. Use this option with the PORT= option.

PORT=number
specifies the port number for the SAS LASR Analytic Server.

---

**TABLEINFO Statement**

The TABLEINFO statement is used to return information about an in-memory table. This information includes the table name, label, number of rows and column, owner, encoding, and the time of table creation. If no table is in use, then information is returned for the in-memory tables for the server specified in the HOST= and PORT= options.

Syntax

TABLEINFO <options>;

**TABLEINFO Statement Options**

HOST="host-name"
specifies the host name for the SAS LASR Analytic Server. Use this option with the PORT= option.

PARTVARS
specifies to include information about partition and orderby variables in the output of the TABLEINFO statement. This enables you to retrieve the names of those variables. If a table is not partitioned or ordered, "N/A" is displayed.

PORT=number
specifies the port number for the SAS LASR Analytic Server. If you do not specify a PORT= value, then behavior of the TABLEINFO statement depends on whether an in-memory table is active. If there is no active table, then the procedure attempts to connect to the server using the LASRPORT macro variable. If a table is active, the information is gathered for the server that is implied by the libref of the active table.

---

**QUIT Statement**

The QUIT statement is used to end the procedure execution. When the procedure reaches the QUIT statement, all resources allocated by the procedure are released. You can no longer execute procedure statements without invoking the procedure again. However, the connection to the server is not lost, because that connection was made through the SAS LASR Analytic Server engine. As a result, any subsequent invocation of the procedure that uses the same libref executes almost instantaneously because the engine is already connected to the server.
Interaction: Using a DATA step or another procedure step is equivalent to issuing a QUIT statement. If there is an error during the procedure execution, it is also equivalent to issuing a QUIT statement.

Syntax
QUIT;

Example: Copying Tables from One Hadoop Installation to Another

Details
This example does not apply to a non-distributed SAS LASR Analytic Server. It might be necessary to work with more than one Hadoop installation so that you can copy SASHDAT files from one Hadoop installation to a newer version. The SAS LASR Analytic Server must be co-located with both Hadoop installations and both versions of Hadoop must be running.

Note: Using the HADOOPHOME= option to switch between Hadoop installations is a server-wide change. If users access the server while the setting is being switched, they might accidentally access the older Hadoop installation. Consider starting a server for the exclusive use of copying files.

Program

```sas
proc lasr create port=12636 serverpermissions=700;  
   performance host="grid001.example.com" install="/opt/TKGrid" nodes=all;  
run;

libname private sasiola host="grid001.example.com" port=12636 tag='hps';

data private.iris; set sashelp.iris; run; /* a table must be active */

proc VASMP data=private.iris;  
   serverparm hadoophome="/olderhadoop/path";  
quit;

proc lasr add hdfs(path="/dept/sales/y2011" direct) port=12636;  
   performance host="grid001.example.com";
run;

proc VASMP data=private.y2011(tag="dept.sales");  
   serverparm hadoophome="/newerhadoop/path";
run;
   save path="/dept/sales/";  
quit;
```

Program Description

1. Starting a server with SERVERPERMISSIONS=700 creates a single-user server. This is not required but can be used to prevent users from accessing the server while
the HADOOP_HOME value is changed and accidentally accessing older or incorrect data.

2. You must have an active table. You can specify an active table with the DATA= option. Any table, such as the Iris data set can be used.

3. Use the SERVERPARM statement to specify the path to the older Hadoop installation with the HADOOPHOME= option. Specify the same path that is returned for the HADOOP_HOME environment variable for the older installation. Example: /hadoop/hadoop-0.21.

4. You must specify the DIRECT option. This statement loads table y2011 into memory from the /dept/sales directory in HDFS.

5. The TAG= option must be used to specify the in-memory table. The server tag matches the HDFS path to the table, but the slashes are replaced with periods (.). If the table was loaded from /, then specify TAG=HADOOP.

6. Use the SERVERPARM statement to specify the path to the newer Hadoop installation. Example: /hadoop-0.23/hadoop-0.23.1.

7. The SAVE statement writes the y2011 table to HDFS in the /dept/sales directory. The HDFS directory is in the newer Hadoop installation.
Chapter 10
Using the SAS LASR Analytic Server Engine

What Does the SAS LASR Analytic Server Engine Do? .................................. 295
Understanding How the SAS LASR Analytic Server Engine Works .............. 295
Understanding Server Tags ................................................................. 296
  What is a Server Tag? ................................................................... 296
  Why Use a Server Tag? ................................................................. 296
Comparing the SAS LASR Analytic Server Engine with the LASR Procedure 296
What is Required to Use the SAS LASR Analytic Server Engine? .............. 297
What is Supported? ........................................................................... 297

What Does the SAS LASR Analytic Server Engine Do?

The SAS LASR Analytic Server engine is used to add, remove, and access tables in a SAS LASR Analytic Server instance.

Typically, the tables that are loaded in memory are very large on a SAS LASR Analytic Server instance. The engine makes it possible to access a table and use procedures like the UNIVARIATE procedure. However, in this case, the entire table is transferred from the server instance to the SAS session and then the procedure is executed on the data. If the table is large, the data volume can overwhelm the SAS session.

The best performance for accessing the data through the engine is with a SAS High-Performance Analytics procedure. These procedures are designed to operate in a distributed computing environment and can read data in parallel from a SAS LASR Analytic Server instance.

Understanding How the SAS LASR Analytic Server Engine Works

An engine is a component of SAS software that reads from or writes to a file. The SAS LASR Analytic Server engine provides Read and Write access for data and metadata information such as variable attributes. Each engine enables SAS to access files that are in a particular format. There are several types of SAS engines.
You use the SAS LASR Analytic Server engine like other SAS data access engines. That is, you execute a LIBNAME statement to assign a libref and to specify the engine. You then use that libref throughout the SAS session where a libref is valid to access a SAS LASR Analytic Server instance.

Understanding Server Tags

What is a Server Tag?

A server tag is a text string that is associated with a table that is loaded into memory on a SAS LASR Analytic Server instance. The server tag is specified in the LIBNAME statement or as a data set option. The server tag and the table name are used together to match the name used for tables in the SAS LASR Analytic Server.

Why Use a Server Tag?

The following list identifies some reasons for specifying a server tag:

- You must use a server tag in a LIBNAME statement or as a data set option to access tables that are loaded from HDFS.
- Different users can load tables with the same name, such as Forecast, into a server instance. You use a server tag and the Forecast table name to specify which table to access.
- Tables that are loaded into memory with the LASR procedure (but not from HDFS) use the libref as the server tag. In order to access these tables, you must specify the server tag.
- When you load a table into memory from HDFS with the LASR procedure, the table is assigned a server tag. The server tag represents the directory path from which the SASHDAT file was loaded. You need to use that server tag to access the table.

See Also

- “Example 4: Accessing Tables Loaded with a DATA Step” on page 304
- “Example 5: Accessing Tables Loaded with the LASR Procedure” on page 305
- “Example 6: Accessing Tables That Are Loaded from HDFS” on page 305

Comparing the SAS LASR Analytic Server Engine with the LASR Procedure

The engine and the LASR procedure are similar in that you can use them to load tables to memory in a SAS LASR Analytic Server instance. You can also use the engine and the procedure to unload tables from memory.

You can use the engine with the APPEND= data set option to add data to an existing table. The procedure cannot modify the data.
You cannot use the engine to load tables into memory from HDFS. Only the LASR procedure can be used to load tables into memory from HDFS.

You can use the LASR procedure to save in-memory tables to HDFS. The procedure writes the data in parallel because the server instance uses SAS High-Performance Deployment of Hadoop as a co-located data provider.

You can use the engine to supply a libref to SAS procedures or DATA steps. However, be aware that if you use the engine as an input data source, the data volume can be large. Large data volumes can overwhelm the SAS session.

What is Required to Use the SAS LASR Analytic Server Engine?

To use the SAS LASR Analytic Server engine, the following are required:

• access to the machines in the cluster where a SAS LASR Analytic Server is running. A server instance is started with the LASR procedure.

• an operating system user ID that is configured for passwordless secure shell (SSH) on the machines in the cluster

The requirement for passwordless SSH is not unique to using the engine. Passwordless SSH is used throughout SAS High-Performance Analytics. The SAS High-Performance Computing Management Console can be used to simplify configuring users for passwordless SSH.

What is Supported?

The following list identifies some usage notes:

• The engine does not support views or BY-group processing.

• You cannot replace or overwrite tables in memory. You must unload the table and then load the new table.

• You cannot use the APPEND procedure. However, you can use an APPEND= data set option to achieve the same result.

• Loading tables into memory from HDFS is performed with the LASR procedure. You cannot load tables into memory from HDFS with the engine.

• The engine guarantees the data order for a particular configuration of worker nodes. If you load a table into a server and you retrieve the data three times, the order of the data is the same. However, if you start another server and load the same table into a different number of worker nodes, then the order in which you retrieve the data is different. However, it is reproducible within fetches from a single server.

• Any order-dependent operation, such as the LAG or DIF functions, cannot rely on stability of results beyond that which can be guaranteed by the distribution model of the data.
Chapter 11
LIBNAME Statement for the SAS LASR Analytic Server Engine

Dictionary

LIBNAME Statement Syntax

associates a SAS libref with tables on a SAS LASR Analytic Server.

Valid in: Anywhere
Category: Data Access

Syntax

LIBNAME libref SASIOLA <LASR= "server-description-file">
   <HOST= "grid-host"> <PORT= number>
   <TAG= server-tag> <FORMATEXPORT= DATA | NONE | ALL>
   <NODEFAULTFORMAT= YES | NO>
   <STARTSERVER <=(non-distributed-server-options)>>
   <SIGNER= "authorization-web-service-uri">
   <VERBOSE= YES | NO>;

Required Arguments

libref
   is a valid SAS name that serves as a shortcut name to associate with the tables on the SAS LASR Analytic Server. The name must conform to the rules for SAS names. A libref cannot exceed eight characters.

SASIOLA
   is the engine name for the SAS LASR Analytic Server engine.
Optional Arguments

FORMATEXPORT= DATA | NONE | ALL
specifies how the engine interacts with user-defined formats when tables are added
to the server instance. The default value is FORMATEXPORT=DATA. This option
can be overridden in a data set option. This option has no effect for input data sets
(data sets that are transferred from the server instance to the SAS client).

DATA
specifies that the definition of all user-defined formats associated with variables
written to the server instance are transferred to the server. You can then use those
formats when you access the table (from a client such as SAS Visual Analytics).
The user-defined formats are transferred to the server only once. The formats are
not transferred as XML streams on subsequent requests to the server.

NONE
specifies that user-defined formats are not transferred to the server.

ALL
specifies that all formats in the format catalog search path are converted and
transferred to the server with the table. This option is useful if the catalog search
path contains user-defined formats that are not associated with variables in the
table, but you might want to use later. Considerable resources can be required to
generate the XML representation of the formats for deployments that have large
catalogs or a deep search path.

HOST="grid-host"
specifies the grid host that has a running server instance. Enclose the host name in
quotation marks. If you do not specify the HOST= option, it is determined from the
GRIDHOST= environment variable.

Alias | SERVER=

Interaction | If the LASR= option is specified, then the host name specified in the
HOST= option is ignored.

LASR="server-description-file"
specifies the server to use. Provide the fully qualified path to the server description
file.

Interaction | If you specify the server description file to use, then you do not need to
specify the HOST= or PORT= options.

NODEFAULTFORMAT= YES | NO
specifies whether a default format that is applied to a variable is reported by the
engine.

If you do not specify a format for a variable when you add a table to the server, the
engine adds a default format. The server applies BEST. for numeric variables and $.
for character variables.

The engine displays this "forced" format in procedures that list variable attributes,
such as the CONTENTS and DATASETS procedures. If you specify
NODEFAULTFORMAT=YES, then the display of the "forced" format is
suppressed.

Note: This setting does not control whether formats are applied to a variable.
PORT=number
specifies the port number to use for connecting to the running server instance. If you use the PORT= option when you start a non-distributed server instance, then use this option to specify the network port number for the server.

Interaction
The LASR procedure stores the port number of the last server instance that is started in the LASRPORT macro variable. You can specify PORT=&LASRPORT to use the macro variable.

SIGNER="authorization-web-service-uri"
specifies the URI for the SAS LASR Authorization web service. The web service is provided by the SAS Visual Analytics software. For more information, see SAS Visual Analytics: Administration Guide.

Example
SIGNER="https://server.example.com/SASLASRAuthorization"

STARTSERVER= YES | NO
STARTSERVER <=(non-distributed-server-options)>
specifies to start a non-distributed server instance. Options are specified as name and value pairs. Separate each option with a space. The following options are available:

AFFINITY= YES | NO
requests that the concurrently executing threads of the server are associated with specific CPUs. When thread affinity is set to YES, a thread does not bounce between CPUs.

Default NO

CLF= YES | NO
specifies to use the common log format for log files. This format is a standardized text file format that is frequently analyzed by web analysis software. Specifying this option implies the LOGGING option.

KEEPLOG= YES | NO
specifies to keep the log files when the server exits instead of deleting them. By default, the log files are removed when the server exits. Specifying this option implies the LOGGING option.

LOGGING= YES | NO
specifies to enabling logging of server actions. The log file is stored with the signature files in the directory that is specified in the PATH= option. The log file is named in the pattern LASR.timestamp.0.saslasr.log.

MAXLOGSIZE=n
specifies the maximum log file size, in megabytes, for a log file. When the log file reaches the specified size, the log file is rolled over and renamed with a sequentially assigned index number (for example, .log.1). The default value is 100 megabytes. Specifying this option implies the LOGGING option.

TIP Do not include an MB or M suffix when you specify the size.

MAXLOGROLL=n
specifies the maximum number of log files to create. When the maximum has been reached, the server begins to overwrite existing log files. The oldest log file is overwritten first. The default value is 10. Specifying this option implies the LOGGING option.
MERGELIMIT=\(n\)
specifies the limit for merging large result sets into smaller groups. The
MERGEBINS= option specifies the size of the group. If MERGEBINS= is not
specified, then \(n\) is the bin limit.

MERGEBINS=\(b\)
specifies the number of bins that numeric variables are binned into when
MERGELIMIT=\(n\) is reached.

NOHOSTCHECK = YES | NO
specifies that the server does not check that the host name specified in the
HOST= option is the local host. This option can be useful with unusual network
configurations.

Interaction
When the SIGNER= option is also specified, the host name that is
specified in the HOST= option is sent to the SAS LASR
Authorization Service.

NTHREADS=\(n\)
specifies the number of threads to use for the server. By default, \(n\) equals the
number of CPU cores on the machine.

PATH="signature-file-path"
specifies the directory to use for storing the server and table signature files. The
specified directory must already exist.

If you do not specify a value for PATH=, the signature files are stored in the
default utility file directory of the SAS session.

PERMISSION=mode
specifies the permission setting for accessing the server instance. The mode value
is specified as an integer value such as 755. The mode corresponds to the mode
values that are used for UNIX file access permissions.

Alias
PERM=

Range
600 to 777

Alias
START=

TAG=server-tag
specifies the tag to use for identifying the tables in the server instance. The value for
server-tag cannot exceed 128 characters in length.

VERBOSE= YES | NO
specifies whether the engine accepts and reports extra messages from TKGrid.
Specifying VERBOSE=YES can help diagnose problems with passwordless SSH
setups, grid install locations, and so on.

The following message in the SAS log shows an example of a problem with
passwordless SSH configuration.

ERROR: Failed to load the SAS LASR Analytic Server access extension
in the distributed computing environment.

Server refused our key from:
/home/sasdemo/.ssh/id_rsa
Timeout waiting for Grid connection.
Examples

Example 1: Submitting a LIBNAME Statement Using the Defaults Program

The following example shows the code for starting a server with the LASR procedure and then connecting to the same server with a LIBNAME statement:

```r
option set=GRIDHOST="grid001.example.com";
option set GRIDINSTALLLOC="/opt/TKGrid";

proc lasr
   create port=10010
   path="/tmp" noclass;
   performance nodes=all;
run;

libname salesvr sasiola;
```

NOTE: No tag was specified in the LIBNAME statement. The default tag (WORK) is used to name and identify tables in the LASR Analytic Server. You can specify a tag as a data set option.

NOTE: Libref SALESSVR was successfully assigned as follows:
Engine:        SASIOLA
Physical Name: SAS LASR Analytic Server engine on host 'grid001.example.com', port 10010

Program Description

1. The grid host is specified in the GRIDHOST environment variable.
2. The default LIBNAME statement does not include the LASR=, HOST=, or PORT= options. The LIBNAME statement uses host name from the GRIDHOST environment variable and the LASRPORT macro variable and connect to server instance.

Example 2: Submitting a LIBNAME Statement Using the LASR= Option

The following example shows a LIBNAME statement that uses the LASR= option to specify the server instance to use:

```r
proc lasr
   create="/tmp/hrsvr" path="/opt/VADP/var/hr"
   noclass;
   performance host="grid001.example.com" install="/opt/TKGrid" nodes=all;
run;

libname hrsvr sasiola lasr="/tmp/hrsvr";
```

Program Description

1. A server instance is started with the CREATE= option. The server description file is /tmp/hrsvr.
2. The HOST= option is specified in the PERFORMANCE statement rather than specifying the GRIDHOST environment variable.

3. The LASR= option specifies the server description file that was created when the server instance started.

**Example 3: Submitting a LIBNAME Statement Using the HOST= and PORT= Options**

The following example shows the code for starting a server with the LASR procedure and then submitting a LIBNAME statement to use the same server by specifying the HOST= and PORT= options.

```sas
proc lasr
   create port=10010
   path="/tmp"
   noclass;

   performance host="grid001.example.com" install="/opt/TKGrid" nodes=all;
run;
```

NOTE: The LASR procedure is executing in the distributed computing environment with 7 worker nodes.

NOTE: The server started on 'grid001.example.com' port 10010.

NOTE: The LASR Analytic Server port '12637' has been assigned to the macro variable "LASRPORT".

```sas
libname hrdata sasiola host="grid001.example.com" port=10010 tag='hr';
```

NOTE: Libref hrdata was successfully assigned as follows:

- Engine: SASIOLA
- Physical Name: SAS LASR Analytic Server engine on host 'grid001.example.com', port 10010

**Program Description**

1. When a server instance is started, the SAS log indicates the port number for the server instance.

2. The PORT= option in the LIBNAME statement references the port number. The value for the PORT= option can also be specified as PORT=&LASRPORT to use the port number for the most recently started server instance.

**Example 4: Accessing Tables Loaded with a DATA Step**

The following example shows how to use the engine without a server tag in a DATA step.

```sas
libname sales sasiola port=10010;

data sales.prdsale;
   set sashelp.prdsale;
run;

proc datasets lib=sales;
quit;
```

* a server tag is not needed to access the data ;
proc print data=sales.prdsale(obs=5);
run;

When no server tag is specified, a default server tag that is named WORK is used.

**Output 11.1** DATASETS Procedure Output Showing the WORK Server Tag

<table>
<thead>
<tr>
<th>Directory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Libref</td>
</tr>
<tr>
<td>Engine</td>
</tr>
<tr>
<td>Physical Name</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Member Type</th>
<th>Number of Rows</th>
<th>Number of Columns</th>
<th>Last Modified</th>
<th>Server Tag</th>
<th>Data Encoding</th>
<th>Owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PRDSALE</td>
<td>DATA</td>
<td>1440</td>
<td>10</td>
<td>29Nov12:20:53:10</td>
<td>WORK</td>
<td>latin1</td>
<td>sasdemo</td>
</tr>
</tbody>
</table>

**Example 5: Accessing Tables Loaded with the LASR Procedure**

When tables are loaded to memory on a server instance with the LASR procedure, the libref that is used with the procedure is set as the server tag. The following example shows how to add a table to a server instance and then access the table with a LIBNAME statement that includes a server tag.

```
proc lasr port=10010 add data=sashelp.prdsale noclass;
run;
libname lasr2 sasiola tag=sashelp;
proc datasets lib=lasr2;
run;
* a server tag is not needed to access the data             ;
* because a server tag is specified in the LIBNAME statement   ;
proc print data=lasr2.prdsale(obs=5);
run;
```

By default, the libref is used as the server tag. The following display shows sashelp used as the server tag.

**Output 11.2** DATASETS Procedure Output Showing the SASHELP Server Tag

<table>
<thead>
<tr>
<th>Directory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Libref</td>
</tr>
<tr>
<td>Engine</td>
</tr>
<tr>
<td>Physical Name</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Member Type</th>
<th>Number of Rows</th>
<th>Number of Columns</th>
<th>Last Modified</th>
<th>Server Tag</th>
<th>Data Encoding</th>
<th>Owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PRDSALE</td>
<td>DATA</td>
<td>1440</td>
<td>10</td>
<td>29Nov12:21:34:08</td>
<td>SASHELP</td>
<td>latin1</td>
<td>sasdemo</td>
</tr>
</tbody>
</table>

**Example 6: Accessing Tables That Are Loaded from HDFS**

When tables are loaded into memory on the server instance with the LASR procedure and the SAS Data in HDFS engine, the server tag is related to the HDFS directory name. The server tag is the same as the HDFS path to the SASHDAT file, but is delimited with periods (.) instead of slashes (/).
The following example shows how to add a table to a server instance from HDFS and then access the table with a LIBNAME statement that includes a server tag.

```
libname sales sashdat path="/dept/sales";
proc lasr port=10010 add data=sales.sales2012 noclass;
run;
libname lasr3 sasiola tag="dept.sales";
proc datasets lib=lasr3;
run;
* access the data with the "dept.sales" server tag;
proc print data=lasr3.sales2012(obs=5);
run;
```

**Output 11.3** DATASETS Procedure Output Showing the DEPT.SALES Server Tag

### Example 7: Loading a Table and Partitioning

Partitioning a table as it is loaded to memory can be a powerful feature for reducing processing times. For more information, see “Data Partitioning and Ordering” on page 14.

```
libname lasrlib sasiola host="grid001.example.com" port=10010 tag="sales";
data lasrlib.prdsale(partition=(country region) orderby=(descending year));
  set sashelp.prdsale;
run;
```

**Program Description**

The Prdsale table is distributed to the machines in the cluster according to the PARTITION= data set option. The rows are distributed according to the unique combinations of the formatted values for the variables Country and Region. In addition, the ORDERBY= option is used to sort the rows in each partition by Year, in descending order.

### Example 8: Creating an Empty Table

Creating an empty table can be useful to seed the column information for a table that you later append to. The following statements create an empty table with two numeric and two character variables:

```
libname lasrlib sasiola host="grid001.example.com" port=10010 tag="sales";
data lasrlib.empty;
  length c1 $15;
```
Example 8: Creating an Empty Table

Program Description

The Empty table is added to the server.

```plaintext
length c2 $12;
x=1;
y=1;
c1="";
c2="";
delete;
run;
```
Chapter 12
Data Set Options for the SAS LASR Analytic Server Engine

Dictionary

APPEND= Data Set Option
specs to append the data to an existing table in the server instance.

Interaction: You must use the NOCLASS option if you load the initial table with the LASR procedure.

Syntax

APPEND=YES | NO

Details

By default, the SAS LASR Analytic Server engine does not permit appending observations to tables. The APPEND= data set option can be used to permit adding observations to an existing table.

Example Code 12.1 Using the APPEND= Data Set Option

    proc lasr add data=grp1.sales noclass port=10010;
    run;
libname grp1lasr host="grid001.example.com" port=10010 tag=grp1;

data grp1lasr.sales (append=yes);
  set yr2012.sales (keep=date location amount);
run;

As shown in the preceding example, the APPEND= data set option can be used to add observations to an existing table. The KEEP= option on the input data set specifies the variables from the input data to append. Any variables for which the input data set does not append data are set to missing. You cannot add new variables to the table.

The example also shows how to load the initial table to memory with the LASR procedure. The NOCLASS option must be specified if you use the LASR procedure. As an alternative, you can load the initial table to memory with the SAS LASR Analytic Server engine.

---

**ARRAY= Data Set Option**

requests that the name space for the table on the SAS session is extended with names that are derived from a temporary numeric array.

**Syntax**

```
ARRAY=(array-name, n)
```

**Details**

The following example shows how to specify a temporary numeric array with variables named Temp1–Temp4:

```
proc imstat;
  table lasrlib.sales (array=(temp,4));
```

The variables in the temporary numeric array do not exist in the in-memory table, but the SAS session assumes that they are there. Using temporary names this way can be useful when your SAS program refers to calculated temporary columns that do not exist when table is opened for input. For example, this option can enable you to retrieve the results for calculated columns with the FETCH statement of the IMSTAT procedure.

This option is used for numeric variables only. If you want to refer to specific temporary variable names, you can also use the TEMPNAMES= option. The TEMPNAMES= option enables you to specify the variable type, and in the case of character variables, the length of the variable.

---

**AUTOCOMMIT= Data Set Option**

specifies how rows are committed to an in-memory table during append operations.

**Syntax**

```
AUTOCOMMIT= nR
AUTOCOMMIT= kS
```
Details
By default, rows are not committed to an in-memory table until the DATA step completes. That is, the rows are held in intermediate storage areas in the server and are not included in requests for data or computational results until the DATA step completes. If you specify AUTOCOMMIT=n or AUTOCOMMIT=nR, then the server commits the rows when at least n rows have been received.

If you specify AUTOCOMMIT=kS, the server commits any rows received within k seconds of the start of the append operation or within k seconds of the previous commit.

FORMATEXPORT= Data Set Option
specifies how the engine interacts with user-defined formats when tables are added to the server instance.

Syntax
FORMATEXPORT=DATA | NONE | ALL

Details
This option is used to override the FORMATEXPORT= option for the LIBNAME statement.

See Also
FORMATEXPORT= option in the LIBNAME statement

HASH= Data Set Option
specifies that when partitioning data, the distribution of partitions is not determined by a tree, but by a hashing algorithm. As a result, the distribution of the partitions is not as evenly balanced, but it is effective when working with high-cardinality partition keys (in the order of millions of partitions).

Syntax
PARTITION=(variable-list) HASH=YES | NO

Example
```bash
data lasrlib.transactions(partition=(cust_id year) hash=yes);
set somelib.sometable;
run;
```

NODEFAULTFORMAT= Data Set Option
specifies whether a default format that is applied to a variable is reported by the server.
Syntax
NODEFAULTFORMAT=YES | NO

Details
This option is used to override the NODEFAULTFORMAT= option for the LIBNAME statement.

ORDERBY= Data Set Option
specifies the variables by which to order the data within a partition.

Example: “Example 7: Loading a Table and Partitioning” on page 306

Syntax
ORDERBY=(variable-list)
ORDERBY=(variable-name <DESCENDING> variable-name)

Details
The variable names in the variable-list are separated by spaces.
The ordering is hierarchical. For example, ORDERBY=(A B) specifies ordering by the values of variable B within the ordered values of variable A. The specified variables must exist and cannot be specified as partitioning variables. The order is determined based on the raw value of the variables and uses locale-sensitive collation for character variables. By default, values are arranged in ascending order. You can specify descending order by preceding the variable name in the variable-list with the keyword DESCENDING.

Example
The following code sample orders the data in the partitions by Year in ascending order and then by Quarter in descending order.

```sas
data lasr1ib.prdsale (partition=(country region)
   orderby=(year descending quarter));
set sashelp.prdsale;
run;
```

PARTITION= Data Set Option
specifies the list of partitioning variables to use for partitioning the table.

Example: “Example 7: Loading a Table and Partitioning” on page 306

Syntax
PARTITION=(variable-list)
Details

Partitioning is available only when you create tables. User-defined format definitions for partitioning variables are always transferred to the server, regardless of the FORMATEXPORT= option.

Partitioning by a variable that does not exist in the output table is an error. Partitioning by a variable listed in the ORDERBY= option is also an error. Partition keys are derived based on the formatted values in the order of the variable names in the variable-list.

Be aware that the key construction is not hierarchical. That is, PARTITION=(A B) specifies that any unique combination of formatted values for variables A and B defines a partition.

PERM= Data Set Option

Specify the permission setting for the table in the server.

Alias: PERMISSION=

Syntax

PERM=mode

Details

The mode is specified as an integer (for example, PERM=755). The value is converted by the engine to a umask. If no permission is specified, the access permissions for the table are set according to the umask of user that loads the table.

SIGNER= Data Set Option

Specifies the URI of the SAS LASR Authorization web service.

Details

This option is used to override the SIGNER= option for the LIBNAME statement.

TAG= Data Set Option

Specifies the tag to use for identifying the tables in the server instance.

Syntax

TAG='server-tag'

Details

If no TAG= option is specified as a data set option, then the TAG= option from the LIBNAME statement is used. If the LIBNAME statement does not specify the TAG= option, then the name of the libref is used as the server tag.
**TEMPNAMES= Data Set Option**

requests that the name space for the table on the SAS session is extended with the specified names.

**Syntax**

\[
\text{TEMPNAMES} = (\text{variable-name}_1, \text{variable-name}_2, \ldots)
\]

**Details**

The following example shows how to specify two temporary numeric variables named Difference and Ratio:

```sas
proc imstat;
   table lasrlib.prdsale (tempnames=(difference ratio));
run;
fetch actual -- month predict ratio / tempnames=(difference ratio)
   tempexpr='difference = actual - predict; ratio = actual / predict';
run;
```

By default, temporary variables that are specified through the TEMPNAMES= option are of numeric type. To define a temporary character variable, follow the variable name with a dollar sign ($) and an optional length.

The following example shows the variable Cust_Name as a character variable with a length of 20:

```sas
table lasrlib.accounts(tempnames=(total_deposits
   cust_name $ 20
   deposit_count
   branch_count));
```

**UCA= Data Set Option**

specifies that you want to use Unicode Collation Algorithms (UCA) to determine the ordering of character variables in the ORDERBY= option.

**Syntax**

\[
\text{PARTITION} = (\text{key}) \quad \text{ORDERBY} = (\text{variable-list}) \quad \text{UCA} = \text{YES} \mid \text{NO}
\]
What Does the SAS Data in HDFS Engine Do?

The SAS Data in HDFS engine is used to distribute data in the Hadoop Distributed File System (HDFS) that is provided by SAS High-Performance Deployment of Hadoop. The engine enables you to distribute the data in a format that is designed for high-performance analytics. The block redundancy and distributed computing provided by SAS High-Performance Deployment of Hadoop is complemented by the block structure that is created with the engine.

The engine is designed to distribute data in HDFS only. Because the data volumes in HDFS are typically very large, the engine is not designed to read from HDFS and transfer data back to the SAS client. For example, consider the case of reading several terabytes of data from a distributed computing environment, transferring that data back to a SAS session, and then using the UNIVARIATE or REG procedures on such a large volume of data. In contrast, the SAS High-Performance Analytics procedures are designed to operate in a distributed computing environment and to read data in parallel from a co-located data provider like SAS High-Performance Deployment of Hadoop.

Understanding How the SAS Data in HDFS Engine Works

An engine is a component of SAS software that reads from or writes to a file. The SAS Data in HDFS engine is write-only for data and read-write for metadata information such as variable attributes. Each engine enables SAS to access files that are in a particular format. There are several types of SAS engines.
You use the SAS Data in HDFS engine like other SAS data access engines. That is, you execute a LIBNAME statement to assign a libref and to specify the engine. You then use that libref throughout the SAS session where a libref is valid to transfer data to the Hadoop Distributed File System (HDFS) or to retrieve information about a table in HDFS.

**What is Required to Use the SAS Data in HDFS Engine?**

To use the SAS Data in HDFS engine, the following are required:

- access to the machines in the cluster where SAS High-Performance Deployment of Hadoop is installed and running
- an operating system user ID that is configured for passwordless secure shell (SSH) on the machines in the cluster

The requirement for passwordless SSH is not unique to using the engine. Passwordless SSH is used throughout SAS High-Performance Analytics. The SAS High-Performance Computing Management Console can be used to simplify configuring users for passwordless SSH.

**What is Supported?**

The SAS Data in HDFS engine is used with SAS High-Performance Deployment of Hadoop only.

The engine is designed as a write-only engine for transferring data to HDFS. However, SAS High-Performance Analytics procedures are designed to read data in parallel from a co-located data provider. The LASR procedure, and other procedures such as HPREG and HPLOGISTIC, can read data from HDFS with the engine. The HPDS2 procedure is designed to read data and write data in parallel. The HPDS2 procedure can be used with the engine to read data from HDFS and create new tables in HDFS.

Whenever a SAS High-Performance Analytics procedure is used to create data in HDFS, the procedure creates the data with a default block size of 8 megabytes. This size can be overridden with the BLOCKSIZE= data set option.

The engine does not support views.

**Common HDFS Commands**

**Start or Stop Hadoop**

To start or stop SAS High-Performance Deployment of Hadoop, log on to the machine that is used as the NameNode. Log on with the user ID that was selected as the service account (that account is often named hadoop).

- The `start` command is as follows:

  ```bash
  /hadoop-installation-directory/sbin/start-dfs.sh
  ```
• The stop command is as follows:

```
/hadoop-installation-directory/sbin/stop-dfs.sh
```

Note: A typical installation directory for the SAS High-Performance Deployment of Hadoop is /hadoop/hadoop-version.

Create and Protect Directories

To create and manage access to directories in HDFS, log on to the machine that hosts the NameNode, and use the hadoop command.

Note: To get started, use the hadoop user account to create and manage directories. Once the beginning of a directory structure is created and permissions are changed, then other user accounts can be used to manage access to directories.

To create a general purpose directory:

1. As the hadoop user account, create a directory named /shared:

```
./hadoop fs -mkdir /shared
```

2. Open up access permissions on the directory:

```
./hadoop fs -chmod 1777 /shared
```

Note: This permissions mode enables only the superuser, directory owner, and file owner to delete or move files within the directory.

3. Confirm that the commands succeeded:

```
./hadoop fs -ls /
```

Found 3 items

- drwxr-xr-x - hadoop supergroup 0 2014-02-03 21:38 /data
- drwxrwxrwt - hadoop supergroup 0 2014-02-14 21:23 /shared
- drwxrwxrwt - hadoop supergroup 0 2014-01-17 11:07 /tmp
- drwxr-xr-x - hadoop supergroup 0 2014-02-13 08:45 /user

To set up a directory for members of the sales group:

1. Create a directory named /dept/sales:

```
./hadoop fs -mkdir -p /dept/sales
```

2. Change the group ID:

```
./hadoop fs -chgrp sales /dept/sales
```

Note: The preceding command assumes that an operating system group that is named sales exists. You can use the SAS High-Performance Computing Management Console to create the group on the machines in the cluster. After you create the group, stop and then start Hadoop (so that the group is recognized).

3. Provide access to only the hadoop user account and members of the sales group:

```
./hadoop fs -chmod 770 /dept/sales
```

4. Confirm that the commands succeeded:

```
./hadoop fs -ls /dept
```

Found 1 items

- drwxrwx----- - hadoop sales 0 2014-02-14 21:29 /dept/sales
Note: The HDFS directory structure is similar to a UNIX file system. Directories have a user ID, group ID, and associated access permissions. More information about the hadoop command is available from http://hadoop.apache.org.
Chapter 14

LIBNAME Statement for the SAS Data in HDFS Engine

Dictionary

LIBNAME Statement Syntax

Associates a SAS libref with SASHDAT tables stored in HDFS.

**Valid in:** Anywhere

**Category:** Data Access

**Syntax**

LIBNAME *libref* SASHDAT

*<HOST="grid-host">* *<INSTALL="grid-install-location">*

*<PATH="HDFS-path">* *<COPIES=n> *<INNAMEONLY=YES | NO>*

*<NODEFAULTFORMAT=YES | NO>*

*<VERBOSE=YES | NO>*;

**Required Arguments**

*libref*

is a valid SAS name that serves as a shortcut name to associate with the SASHDAT tables that are stored in the Hadoop Distributed File System (HDFS). The name must conform to the rules for SAS names. A libref cannot exceed eight characters.

*SASHDAT*

is the engine name for the SAS Data in HDFS engine.

**Optional Arguments**

**COPIES=n**

specifies the number of replications to make for the data set (beyond the original blocks). The default value is 2 when the INNAMEONLY option is specified and otherwise is 1. Replicated blocks are used to provide fault tolerance for HDFS. If a
machine in the cluster becomes unavailable, then the blocks needed for the
SASHDAT file can be retrieved from replications on other machines. If you specify
COPIES=0, then the original blocks are distributed, but no replications are made and
there is no fault tolerance for the data.

HOST="grid-host"
specifies the grid host that has a running Hadoop NameNode. Enclose the host name
in quotation marks. If you do not specify the HOST= option, it is determined from
the GRIDHOST= environment variable.

INNAMEONLY=YES | NO
specifies that when data is added to HDFS, that it should be sent as a single block to
the Hadoop NameNode for distribution. This option is appropriate for smaller data
sets.

Alias NODIST

INSTALL="grid-install-location"
specifies the path to the TKGrid software on the grid host. If you do not specify this
option, it is determined from the GRIDINSTALLLOC= environment variable.

NODEFAULTFORMAT=YES | NO
specifies whether a default format that is applied to a variable is reported by the
engine.

If you do not specify a format for a variable when you add a table to HDFS, the
engine adds a default format. The server applies BEST. for numeric variables and $. for character variables.

The engine displays this "forced" format in procedures that list variable attributes,
such as the CONTENTS and DATASETS procedures. If you specify
NODEFAULTFORMAT=YES, then the display of the "forced" format is
suppressed.

Note: This setting does not control whether formats are applied to a variable.

PATH="HDFS-path"
specifies the fully qualified path to the HDFS directory to use for SASHDAT files.
You do not need to specify this option in the LIBNAME statement because it can be
specified as a data set option.

VERBOSE=YES | NO
specifies whether the engine accepts and reports extra messages from TKGrid. For
more information, see the VERBOSE= option on page 302 for the SAS LASR
Analytic Server engine.

Examples

Example 1: Submitting a LIBNAME Statement Using the Defaults Program

The following example shows the code for connecting to a Hadoop NameNode with a
LIBNAME statement:

```
option set=GRIDHOST="grid001.example.com";
option set GRIDINSTALLLOC="/opt/TKGrid";
libname hdfs sashdat;
```
NOTE: Libref HDFS was successfully assigned as follows:
Engine:        SASHDAT
Physical Name: grid001.example.com

Program Description

1. The host name for the Hadoop NameNode is specified in the GRIDHOST environment variable.

2. The LIBNAME statement uses host name from the GRIDHOST environment variable and the path to TKGrid from the GRIDINSTALLLOC environment variable. The PATH= and COPIES= options can be specified as data set options.

Example 2: Submitting a LIBNAME Statement Using the HOST=, INSTALL=, and PATH= Options

The following example shows the code for submitting a LIBNAME statement with the HOST=, INSTALL=, and PATH= options.

```sas
libname hdfs sashdat host=grid001.example.com install=/opt/TKGrid path=/user/sasdemo;
```

NOTE: Libref HDFS was successfully assigned as follows:
Engine:        SASHDAT
Physical Name: Directory '/user/sasdemo' of HDFS cluster on host grid001.example.com

Example 3: Adding Tables to HDFS

The following code sample demonstrates the LIBNAME statement and the REPLACE= and BLOCKSIZE= data set options. The LABEL= data set option is common to many engines.

```sas
libname arch /data/archive;
libname hdfs sashdat host=grid001.example.com install=/opt/TKGrid path=/dept;

data hdfs.allyears(label="Sales records for previous years" replace=yes blocksize=32m);
  set arch.sales2012
    arch.sales2011
    ...
  ;
run;
```

Example 4: Adding a Table to HDFS with Partitioning

The following code sample demonstrates the PARTITION= and ORDERBY= data set options. The rows are partitioned according to the unique combinations of the formatted values for the Year and Month variables. Within each partition, the rows are sorted by descending values of the Prodtype variable. For more information, see “Data Partitioning and Ordering” on page 14.

```sas
libname hdfs sashdat host=grid001.example.com install=/opt/TKGrid path=/dept;

data hdfs.prdsale(partition=(year month) orderby=(descending prodtype));
  set sashelp.prdsale;
run;
```
Example 5: Removing Tables from HDFS
Removing tables from HDFS can be performed with the DATASETS procedure.

```
libname hdfs sashdat host="grid001.example.com" install="/opt/TKGrid"
    path="/dept";
proc datasets lib=hdfs;
    delete allyears;
run;
```

**NOTE:** Deleting HDFS.ALLYEARS (memtype=DATA).

Example 6: Creating a SASHDAT File from Another SASHDAT File
The following example shows copying a data set from HDFS, adding a calculated variable, and then writing the data to HDFS in the same library. The BLOCKSIZE= data set option is used to override the default 8-megabyte block size that is created by SAS High-Performance Analytics procedures. The COPIES=0 data set option is used to specify that no redundant blocks are created for the output SASHDAT file.

```
libname hdfs sashdat host="grid001.example.com" install="/opt/TKGrid"
    path="/dept";
proc hpds2
    in = hdfs.allyears(where=(region=212))
    out = hdfs.avgsales(blocksize=32m copies=0);
    data DS2GTF.out;
        dcl double avgsales;
        method run();
        set DS2GTF.in;
        avgsales = avg(month1-month12);
        end;
    enddata;
run;
```

1 The WHERE clause is used to subset the data in the input SASHDAT file.
2 The BLOCKSIZE= and COPIES= options are used to override the default values.

Example 7: Working with CSV Files
The comma-separated value (CSV) file format is a popular format for files stored in HDFS. The SAS Data in HDFS engine can read these files in parallel. The engine does not write CSV files.

List the Variables in a CSV File
The following example shows how to access a CSV file in HDFS and use the CONTENTS procedure to list the variables in the file. For this example, the first line in the CSV file lists the variables names. The GETNAMES data set option is used to read them from the first line in the file.

```
libname csvfiles sashdat host="grid001.example.com" install="/opt/TKGrid"
    path="/user/sasdemo/csv";
```
proc contents data=csvfiles.rep(filetype=csv getnames=yes);
run;

**Output 14.1** List the Variables in a CSV File with the CONTENTS Procedure

<table>
<thead>
<tr>
<th>Data Set Name</th>
<th>user\sasdemocsv\rep.csv</th>
<th>Observations</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Member Type</td>
<td>DATA</td>
<td>Variables</td>
<td>6</td>
</tr>
<tr>
<td>Engine</td>
<td>SASHDAT</td>
<td>Indexes</td>
<td>0</td>
</tr>
<tr>
<td>Created</td>
<td>Tuesday, July 03, 2012 08:53:38 AM</td>
<td>Observation Length</td>
<td>208</td>
</tr>
<tr>
<td>Last Modified</td>
<td>Thursday, June 20, 2012 02:46:41 PM</td>
<td>Deleted Observations</td>
<td>0</td>
</tr>
<tr>
<td>Protection</td>
<td>Compressed</td>
<td></td>
<td>NO</td>
</tr>
<tr>
<td>Data Set Type</td>
<td>Sorted</td>
<td></td>
<td>NO</td>
</tr>
<tr>
<td>Label</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Representation</td>
<td>SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encoding</td>
<td>Default</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Convert a CSV File to SASHDAT

The engine is not designed to transfer data from HDFS to a SAS client. As a consequence, the contents of a CSV file can be accessed only by a SAS High-Performance Analytics procedure that runs on the same cluster that is used for HDFS. The SAS High-Performance Analytics procedures can read the data because the procedures are designed to read data in parallel from a co-located data provider.

The following code sample shows how to convert a CSV file to a SASHDAT file with the HPDS2 procedure.

```sas
option set=GRIDHOST="grid001.example.com";
option set=GRIDINSTALLLOC="/opt/TKGrid";
libname csvfiles sashdat path="/user/sasdemo/csv";
proc hpds2 in=csvfiles.rep(filetype=csv getnames=yes) out=csvfiles.rephdat(path="/user/sasdemo" copies=0 blocksize=32m);
data DS2GTF.out;
method run();
set DS2GTF.in;
end;
enddata;
run;
```
The values for the GRIDHOST and GRIDINSTALLLOC environment variables are read by the SAS Data in HDFS engine in the LIBNAME statement and by the HPDS2 procedure.

The FILETYPE=CSV data set option enables the engine to read the CSV file. The GETNAMES= data set option is used to read the variable names from the first line in the CSV file.

The PATH= data set option is used to store the output as /user/sasdemo/rephdat.sashdat. The COPIES=0 data set option is used to specify that no redundant blocks are created for the rephdat.sashdat file.
Chapter 15
Data Set Options for the SAS Data in HDFS Engine

Dictionary

<table>
<thead>
<tr>
<th>Blocksize= Data Set Option</th>
<th>325</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOCKSIZE= Data Set Option</td>
<td>325</td>
</tr>
<tr>
<td>COLUMNS= Data Set Option</td>
<td>326</td>
</tr>
<tr>
<td>COPIES= Data Set Option</td>
<td>327</td>
</tr>
<tr>
<td>FILETYPE= Data Set Option</td>
<td>327</td>
</tr>
<tr>
<td>GETNAME= Data Set Option</td>
<td>328</td>
</tr>
<tr>
<td>GETOBS= Data Set Option</td>
<td>328</td>
</tr>
<tr>
<td>GUESSROWS= Data Set Option</td>
<td>329</td>
</tr>
<tr>
<td>HASH= Data Set Option</td>
<td>329</td>
</tr>
<tr>
<td>LOGUPDATE= Data Set Option</td>
<td>330</td>
</tr>
<tr>
<td>ORDERBY= Data Set Option</td>
<td>330</td>
</tr>
<tr>
<td>PARTITION= Data Set Option</td>
<td>331</td>
</tr>
<tr>
<td>PATH= Data Set Option</td>
<td>332</td>
</tr>
<tr>
<td>PERM= Data Set Option</td>
<td>332</td>
</tr>
<tr>
<td>REPLACE= Data Set Option</td>
<td>332</td>
</tr>
<tr>
<td>UCA= Data Set Option</td>
<td>333</td>
</tr>
</tbody>
</table>

Dictionary

**BLOCKSIZE= Data Set Option**

specifies the block size to use for distributing the data set.

**Valid in:** DATA Step  
**Default:** 2 megabytes  
**Example:** “Example 6: Creating a SASHDAT File from Another SASHDAT File” on page 322

**Syntax**

```
BLOCKSIZE=
```

**Details**

By default, the SAS Data in HDFS engine distributes data in 2-megabyte blocks or the length of a record, which ever is greater. You can override this value by specifying the block size to use. Suffix values are B (bytes), K (kilobytes), M (megabytes), and G.
The actual block size is slightly larger than the value that you specify. This occurs for any of the following reasons:

- to reach the record length. This occurs if the specified size is less than the record length.
- to align on a 512-byte boundary.
- to include a metadata header in HDFS for the SASHDAT file.

The following code shows an example of specifying the BLOCKSIZE= option.

**Example Code 15.1 Using the BLOCKSIZE= Data Set Option**

```sas
data hdfs.sales (blocksize=48M);
  set yr2012.sales;
run;
```

### COLUMNS= Data Set Option

specifies the variable names and types for a CSV file.

**Alias:** COLS=

**Applies to:** Reading CSV files

**Syntax**

```
COLUMNS=(column-specification < …column-specification>);
```

**Required Argument**

column-specification

is a name-value pair that specifies the column name and data type. For numeric data, specify `double` as the data type. For character data, specify `'char(length)'`.

**Default**

Any variables that are not named are assigned the name `VARn`.

**Example**

```
columns=(station='char(4)' obsdate='char(18)' tempf=double precip=double)
```

**Details**

Numeric variables use eight bytes. For character variables, if the byte length is not specified, then the default action is to use eight bytes. If the variable in the CSV file uses fewer bytes than the specified length, then the variable is padded with spaces up to the specified length. If the variable in the CSV file uses more bytes than the specified length, then the variable is truncated to the specified length.

If the variable name is not specified, then the variable is named automatically. Automatically named variables are named `VARn`, starting at 1. If the data type is not specified and cannot be determined, the variable is assigned as `char(8)`.

**TIP**

Do not use a comma between each column specification. Enclose `'char(n)'` in quotation marks.
COPIES= Data Set Option

specifies the number of replications to make for the data set (beyond the original blocks).

Default: 1

Syntax

COPIES=n

Details

The default value is 1. This default value creates one copy of each block, in addition to the original block. When the INNAMEONLY option is specified, the default is 2. Replicated blocks are used to provide fault tolerance for HDFS. If a machine in the cluster becomes unavailable, then the blocks needed for the SASHDAT file can be retrieved from replications on other machines.

You can specify COPIES=0 to avoid creating redundant blocks for the SASHDAT file. This option can be useful to preserve storage space when you have redundancy for the source data.

FILETYPE= Data Set Option

specifies whether to access a comma-separated value (CSV) file instead of a SASHDAT file.

Applies to: Reading CSV files

Syntax

FILETYPE=CSV

Details

The SAS Data in HDFS engine can be used to read CSV files. The engine does not write CSV files. Specify this option to use the file as input for a SAS High-Performance Analytics procedure or the SAS LASR Analytic Server.

The filename for CSV files in HDFS can be upper, mixed, or lower case. If more than one file in the directory has the same name (but with different casing), the engine does not read the file because the file reference is ambiguous.

See Also

- COLUMNS= data set option
- GETNAMES data set option
- GUESSROWS= data set option
**GETNAMES= Data Set Option**

specifies to read variable names from the first line in the CSV file.

**Applies to:** Reading CSV files

**Syntax**

```
GETNAMES= YES | NO
```

**Details**

Specify GETNAMES=YES if the first line of a CSV file contains the variable names for the file. Alternatively, you can specify the variable names in the `COLUMNS=` data set option, or you can use the default names that are provided by the SAS Data in HDFS engine.

**GETOBS= Data Set Option**

specifies to retrieve the number of observations in SASHDAT files.

**Syntax**

```
GETOBS= YES | NO
```

**Details**

By default, the SAS Data in HDFS engine does not compute the number of observations in a SASHDAT file. This improves performance for SASHDAT files that are distributed among a large number of blocks, or for HDFS directories that have a large number of SASHDAT files. When you specify GETOBS=YES, the engine calculates the number of observations in a SASHDAT file.

```plaintext
ods select attributes;

proc datasets library=hdfs;
   contents data=sales2012(getobs=yes);
run;
```
**GUESSROWS= Data Set Option**

specifies the number of lines in CSV file to scan for determining variable types and lengths.

**Default:** 20

**Applies to:** Reading CSV files

**Syntax**

\[ \text{GUESSROWS} = n \]

**Details**

The SAS Data in HDFS engine scans the specified number of lines from the CSV file to determine the variable types and lengths. If the GETNAMES data set option is specified, then the engine begins scanning lines from the second line in the file.

---

**HASH= Data Set Option**

specifies that when partitioning data, the distribution of partitions is not determined by a tree, but by a hashing algorithm. As a result, the distribution of the partitions is not as evenly balanced, but it is effective when working with high-cardinality partition keys (in the order of millions of partitions).

**Syntax**

\[ \text{PARTITION} = (\text{variable-list}) \ \text{HASH=} \ \text{YES} | \ \text{NO} \]

**Example**

```sas
data hdfs.transactions(partition=(cust_id year) hash=yes);
  set somelib.sometable;
run;
```
LOGUPDATE= Data Set Option

specifies to provide progress messages in the SAS log about the data transfer to the grid host.

Syntax

```
LOGUPDATE= YES | NO
```

Details

The data transfer size is not necessarily the same as the block size that is used to form blocks in HDFS. The data transfer size is selected to optimize network throughput. A message in the SAS log does not mean that a block was written to HDFS. The message indicates the transfer progress only.

```
data hdfs.sales2012(logupdate=yes);
    set saleslib.sales2012;
run;
```

```
NOTE: 4096 kBytes (5191 records) have been transmitted (1.91 MB/sec).
NOTE: 8192 kBytes (10382 records) have been transmitted (3.65 MB/sec).
NOTE: 12288 kBytes (15573 records) have been transmitted (5.19 MB/sec).
NOTE: 16384 kBytes (20764 records) have been transmitted (6.15 MB/sec).
NOTE: 20480 kBytes (25955 records) have been transmitted (7.3 MB/sec).
NOTE: 24576 kBytes (31146 records) have been transmitted (8.16 MB/sec).
NOTE: 28672 kBytes (36337 records) have been transmitted (8.83 MB/sec).
NOTE: 32768 kBytes (41528 records) have been transmitted (9.73 MB/sec).
NOTE: 36864 kBytes (46719 records) have been transmitted (10.3 MB/sec).
NOTE: 40960 kBytes (51910 records) have been transmitted (10.8 MB/sec).
NOTE: 45056 kBytes (57101 records) have been transmitted (11.6 MB/sec).
NOTE: 49152 kBytes (62292 records) have been transmitted (12 MB/sec).
NOTE: 53248 kBytes (67483 records) have been transmitted (12.4 MB/sec).
NOTE: 57344 kBytes (72674 records) have been transmitted (12.9 MB/sec).
NOTE: 61440 kBytes (77865 records) have been transmitted (13.2 MB/sec).
NOTE: 65536 kBytes (83056 records) have been transmitted (13.5 MB/sec).
NOTE: 69632 kBytes (88247 records) have been transmitted (13.9 MB/sec).
NOTE: 73728 kBytes (93438 records) have been transmitted (14.1 MB/sec).
NOTE: 77824 kBytes (98629 records) have been transmitted (14.3 MB/sec).
NOTE: There were 100000 observations read from the data set SALESLIB.YEAR2012.
NOTE: The data set /user/sasdemo/sales2012 has 100000 observations and 86 variables.
NOTE: 78906 kBytes (100000 records) have been transmitted (14.3 MB/sec).
```

ORDERBY= Data Set Option

specifies the variables by which to order the data within a partition.

Example:  "Example 4: Adding a Table to HDFS with Partitioning" on page 321

Syntax

```
ORDERBY=(variable-list)
ORDERBY=(variable-name <DESCENDING> variable-name)
```
Details

The variable names in the variable-list are separated by spaces.

The ordering is hierarchical. For example, ORDERBY=(A B) specifies ordering by the values of variable B within the ordered values of variable A. The specified variables must exist and cannot be specified as partitioning variables. The order is determined based on the raw value of the variables and uses locale-sensitive collation for character variables. By default, values are arranged in ascending order. You can specify descending order by preceding the variable name in the variable-list with the keyword DESCENDING.

Example

The following code sample orders the data in the partitions by Year in ascending order and then by Quarter in descending order.

```sas
data hdfs.prdsale (partition=(country region)
   orderby=(year descending quarter));
set sashelp.prdsale;
run;
```

PARTITION= Data Set Option

specifies the list of partitioning variables to use for partitioning the table.

Interaction: If you specify the PARTITION= option and the BLOCKSIZE= option, but the block size is less than the calculated size that is needed for a block, the operation fails and the table is not added to HDFS. If you do not specify a block size, the size is calculated to accommodate the largest partition.

Example: “Example 4: Adding a Table to HDFS with Partitioning” on page 321

Syntax

```
PARTITION=(variable-list)
```

Details

Partitioning is available only when you add tables to HDFS. If you partition the table when you add it to HDFS, it becomes a partitioned in-memory table when you load it to SAS LASR Analytic Server. If you also specify the ORDERBY= option, then the ordering is preserved when the table is loaded to memory too.

Partition keys are derived based on the formatted values in the order of the variable names in the variable-list. All of the rows with the same partition key are stored in a single block. This ensures that all the data for a partition is loaded into memory on a single machine in the cluster. The blocks are replicated according to the default replication factor or the value that you specify for the COPIES= option.

If user-defined formats are used, then the format name is stored with the table, but not the format. The format for the variable must be available to the SAS LASR Analytic Server when the table is loaded into memory. This can be done by having the format in the format catalog search path for the SAS session.
Be aware that the key construction is not hierarchical. That is, `PARTITION=(A B)` specifies that any unique combination of formatted values for variables A and B defines a partition.

Partitioning by a variable that does not exist in the output table is an error. Partitioning by a variable listed in the `ORDERBY=` option is also an error.

**PATH= Data Set Option**

specifies the fully qualified path to the HDFS directory to use for SASHDAT files.

**Syntax**

```plaintext
PATH='HDFS-path'
```

**Details**

This option overrides the `PATH=` option specified in the `LIBNAME` statement.

**PERM= Data Set Option**

specifies how the engine sets the file access permissions on the SASHDAT file.

**Alias:** `PERMISSION=`

**Syntax**

```plaintext
PERM=mode
```

**Details**

The `mode` value is specified as an integer value such as 755. The mode corresponds to the mode values that are used for UNIX file access permissions.

**REPLACE= Data Set Option**

specifies whether to overwrite an existing SASHDAT file.

**Syntax**

```plaintext
REPLACE=YES | NO
```

**Details**

By default, the SAS Data in HDFS engine does not replace SASHDAT files. Specify `REPLACE=YES` as a data set option to replace a SASHDAT file by overwriting it.
UCA= Data Set Option

specifies that you want to use Unicode Collation Algorithms (UCA) to determine the ordering of character variables in the ORDERBY= option.

**Syntax**

```
PARTITION=(key) ORDERBY=(variable-list) UCA= YES | NO
```
Chapter 16
Programming with SAS LASR Analytic Server

About Programming

When programming with SAS LASR Analytic Server, it is important to understand where the computation occurs and memory utilization.

- The IMSTAT procedure always performs the computation in the server, and the analysis is performed against the original in-memory table.
- Other procedures (for example, FREQ, UNIVARIATE, and RANK) transfer the in-memory table to the client machine. After the transfer, the session on the client machine performs the analysis on the copy of the data.
- Most DATA step programs operate by transferring the in-memory table to the client and then performing the computation. However, if a DATA step is written to use in-memory tables for input and output, the DATA step can run in-memory, with restrictions. The next section describes how to use this feature.

DATA Step Programming for Scoring In SAS LASR Analytic Server

Scoring In-Memory Tables Using DATA Step Processing

The DATA step can process in-memory tables for scoring with limitations:

- Only one input file and one output file is allowed.
- Only functions and formats that are supported by the DS2 language compile successfully.
Some DATA step statements are not allowed, such as those pertaining to input and output. For more information, see “Requirements for LASR Score Mode DATA Step Processing” on page 337 and “Restrictions in DATA Step Processing” on page 338.

To enable the DATA step to score in-memory SAS tables, set the system option DSACCEL=ANY.

If a SAS program does not meet the requirements for running in the SAS LASR Analytic Server, SAS writes informational or error messages to the log and executes the code in your Base SAS session. In this case, SAS reads and writes large tables over the network. You can determine whether your code is compliant with the SAS LASR Analytic Server compiler by setting the system option MSGLEVEL= to I. When MSGLEVEL=I, SAS writes log messages that identify the non-compliant code.

**Example 1: A DATA Step Program For SAS LASR Analytic Server**

This example demonstrates executing a DATA step program in the SAS LASR Analytic Server:

```sas
/* Enable DATA step parallel processing using the system option   */ /* and enable messages to view non-compliant code in the SAS log. */ options dsaccel=any msglevel=i;

/* Create a libref for in-memory tables. */
libname lasr sasiola host="grid001.example.com" port=10010 tag='hps';

/* Create a libref for the input data that is stored on disk. */
libname score '/myScoreData/';

/* Load the input table into memory */
data lasr.intr;
set score.intrid;
run;

/* Execute the score code using the in-memory tables. */ /* Both tables must use the same libref. */
data lasr.sumnormtable;
set lasr.intr;

/* Execute the score code. */
if sum > 1000
then score=1;
run;
```

**Example 2: Using User-Defined Formats with In-Memory Tables**

You can use user-defined formats in a DATA step by using the CATALOG procedure to copy the format catalog to a library. This example copies the format library to Work:

```sas
/* Enable DATA step parallel processing using the system option   */ /* and enable messages to view non-compliant code in the SAS log. */
/* Create a libref for in-memory tables. */
libname lasr sasiola host="grid001.example.com" port=10010 tag='hps';

/* Create a libref for the input data that is stored on disk. */
libname score '/myScoreData/';

/* Load the input table into memory */
data lasr.intr;
set score.intrid;
run;
```

```sas
/* Execute the score code using the in-memory tables. */ /* Both tables must use the same libref. */
data lasr.sumnormtable;
set lasr.intr;

/* Execute the score code. */
if sum > 1000
then score=1;
run;
```
options dsaccel=any msglevel=i;

/* Create a libref for the in-memory tables. */
libname lasr sasiola host="grid001.example.com" port=10010 tag='hps';

/* Create a libref for the input data and format catalog that is stored on disk. */
libname score '/myScoreData/';

/* Copy the demx format catalog to the Work library */
proc catalog catalog=score.dmex;
    copy out=work.formats;
quit;

/* Enable in-memory processing (dsaccel) and load the input table into memory. */
data lasr.dmex;
    set score.dmex;
run;

/* Enable in-memory processing (dsaccel) and execute the score code using the in-memory tables. */
data lasr.dmexout;
    set lasr.dmex;
    %inc "dmex.sas";
run;

SAS automatically searches the Work and Library libraries for a format catalog. If you copy the format library to a library other than Work or Library, then you must use the FMTSEARCH= system option to let SAS know the location of the format library.

options fmtsearch=(myFmtLib);

You must also specify the FMTSEARCH= system option if the format catalog name is not format:

options fmtsearch=(myFmtLib.myFmts);

---

**Requirements for LASR Score Mode DATA Step Processing**

In order to score in-memory tables in SAS LASR Analytic Server, the following is required:

- The DSACCEL=ANY system option is set.
- The code must contain a LIBNAME statement using the SASIOLA engine.
- The input and output tables must use the same libref for the SASIOLA engine.
- The DATA statement must be followed immediately by the SET statement.

This example demonstrates these requirements:

libname lasr sasiola;
data lasr.out;
    set lasr.in;
    /* DATA step code */
run;
Restrictions in DATA Step Processing

Here are the restrictions for using the DATA step in SAS LASR Analytic Server:

- More than one SET statement is not supported. SET statement options are not allowed.
- These statements are not supported:
  - BY (or FIRST. and LAST. variables)
  - CONTINUE
  - DISPLAY
  - FILE
  - Sub-setting IF
  - INFILE
  - INPUT
  - LEAVE
  - MERGE
  - MODIFY
  - OUTPUT
  - PUT
  - REMOVE
  - RENAME
  - REPLACE
  - RETAIN
  - UPDATE
  - WHERE
  - WINDOW
- The ABORT statement has these restrictions:
  - The ABORT statement does not accept arguments.
  - The ABORT statement is not supported within functions. It is valid only in the main program.
- These functions are not supported:
  - DIF
  - LAG
- The INPUT function does not support the question mark (?) and double question mark (??) modifiers.
- Some CALL routines are not supported. Routines are supported if there is an equivalent function.
- You can use only SAS formats and functions that are supported by the DS2 language. These formats and functions are documented in SAS DS2 Language Reference.
• Component objects are not supported.
• Scoring input variables cannot be modified.
Chapter 17
Text Analytics in SAS LASR
Analytic Server

SAS Linguistic Files .......................................................... 341

Language Processing Concepts .......................................... 342
Stemming ........................................................................... 342
Tagging .............................................................................. 342
Noun Group Extraction ....................................................... 343
Entity Identification .......................................................... 343

Data Preparation ............................................................... 344
Width Limitation for Text Fields ......................................... 344
Encoding ............................................................................ 344
A Document ID Variable Is Required ................................. 344

Output Tables for the TEXTPARSE Statement ...................... 344
Sample Data and Program ................................................ 344
Terms Table ....................................................................... 346
Parent Table ....................................................................... 347
SVD V Table ..................................................................... 349
SVD U Table ..................................................................... 349
Topics Table ...................................................................... 350
Terms by Topic Table ........................................................ 351
Document Projection Table ................................................. 352

SAS Linguistic Files

The text analytic capabilities of the server depend on accessing SAS linguistic files. The server uses these files to perform the parsing, term extraction, stemming, tagging of terms, and so on. The files must be made available to the server and the path must be specified in either the TKTXTANIO_BINDAT_DIR or TKPARSE_BINDAT_DIR environment variables. (See the examples.)

Information about installing the linguistic files for distributed deployments is provided in SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide. A common location is /opt/TKGData, but the location can be customized at your site.

For non-distributed deployments, the files are installed with SAS as follows:

Windows Specifics
\SASFoundation\9.4\tktg\sasmisc

UNIX Specifics
/SASFoundation/9.4/misc/tkg
Language Processing Concepts

**Stemming**

Stemming identifies the possible root form of an inflected word. For example, the word talk is the stem of the words talk, talks, talking, and talked. In this case talk is the parent, and talk, talks, talking, and talked are its children.

**Tagging**

Tagging disambiguates the grammatical category of a word by analyzing it in context. For example, consider the following sentence:

> I like to bank at the local branch of my bank.

The first bank is tagged as a verb and the second bank is tagged as a noun. The possible speech tags that you might see are as follows:

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABBR</td>
<td>Abbreviation</td>
</tr>
<tr>
<td>ADJ</td>
<td>Adjective</td>
</tr>
<tr>
<td>ADV</td>
<td>Adverb</td>
</tr>
<tr>
<td>AUX</td>
<td>Auxiliary or modal term</td>
</tr>
<tr>
<td>CONJ</td>
<td>Conjunction</td>
</tr>
<tr>
<td>DET</td>
<td>Determiner</td>
</tr>
<tr>
<td>INTERJ</td>
<td>Interjection</td>
</tr>
<tr>
<td>NOUN</td>
<td>Noun</td>
</tr>
<tr>
<td>NOUN_GROUP</td>
<td>Compound noun</td>
</tr>
<tr>
<td>NUM</td>
<td>Number or numeric expression</td>
</tr>
<tr>
<td>PART</td>
<td>Infinitive marker, negative participle, or possessive marker</td>
</tr>
<tr>
<td>PREF</td>
<td>Prefix</td>
</tr>
<tr>
<td>PREP</td>
<td>Preposition</td>
</tr>
<tr>
<td>PROP</td>
<td>Proper noun</td>
</tr>
<tr>
<td>PUNCT</td>
<td>Punctuation</td>
</tr>
</tbody>
</table>
Noun Group Extraction

Noun groups provide more relevant information than simple nouns. A noun group is defined as a sequence of nouns and their modifiers. Noun group extraction uses part-of-speech tagging to identify nouns and their related words that together form a noun group. Examples of noun groups are "week-long cruises" and "Middle Eastern languages."

Entity Identification

Entity identification uses SAS linguistic technologies to classify sequences of words into predefined classes. These classes are assigned as roles for the corresponding sequences. For example, "Person," "Location," "Company," and "Measurement" are identified as classes for "George W. Bush," "Boston," "SAS Institute," "2.5 inches," respectively. The following table lists the possible entities for English.

<table>
<thead>
<tr>
<th>Entity</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADDRESS</td>
<td>Postal address or number and street name</td>
</tr>
<tr>
<td>COMPANY</td>
<td>Company name</td>
</tr>
<tr>
<td>CURRENCY</td>
<td>Currency or currency expression</td>
</tr>
<tr>
<td>INTERNET</td>
<td>E-mail address or URL</td>
</tr>
<tr>
<td>LOCATION</td>
<td>City, county, state, political or geographical place or region</td>
</tr>
<tr>
<td>MEASURE</td>
<td>Measurement or measurement expression</td>
</tr>
<tr>
<td>NOUN_GROUP</td>
<td>Phrases that contain multiple words</td>
</tr>
<tr>
<td>ORGANIZATION</td>
<td>Government, legal, or service agency</td>
</tr>
<tr>
<td>PERCENT</td>
<td>Percentage or percentage expression</td>
</tr>
<tr>
<td>PERSON</td>
<td>Person’s name</td>
</tr>
<tr>
<td>PHONE</td>
<td>Telephone number</td>
</tr>
<tr>
<td>PROP_MISC</td>
<td>Proper noun with an ambiguous classification</td>
</tr>
<tr>
<td>SSN</td>
<td>Social Security number</td>
</tr>
<tr>
<td>TIME</td>
<td>Time or time expression</td>
</tr>
<tr>
<td>Entity</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>TIME_PERIOD</td>
<td>Measure of time expressions</td>
</tr>
<tr>
<td>TITLE</td>
<td>Person’s title or position</td>
</tr>
<tr>
<td>VEHICLE</td>
<td>Motor vehicle, including color, year, make, and model</td>
</tr>
</tbody>
</table>

Data Preparation

**Width Limitation for Text Fields**

Character variables in SAS data sets cannot exceed 32K. So, when using the TEXTPARSE statement in a SAS program, you are limited to 32K for a text fields.

**Encoding**

The linguistic data files and libraries are sensitive to the NLS encoding. The encoding used in the TEXTPARSE statement is derived from the encoding of the active table that you parse.

**A Document ID Variable Is Required**

The TEXTPARSE statement requires that each document is uniquely identified by a column in the table. This column essentially holds a record ID in the table, as each row is considered a separate document.

Output Tables for the TEXTPARSE Statement

**Sample Data and Program**

The TEXTPARSE statement generates a summary table and up to seven temporary tables. The following program provides sample data and statements for getting started and then showing the basic layout for each of the temporary tables.

```sas
data getstart;
infile cards delimiter='|' missover;
length text $150;
input text$ docid$;
cards;
High-performance analytics hold the key to unlocking the unprecedented business value of big data. Organizations looking for optimal ways to gain insights from big data in shorter reporting windows are turning to SAS. As the gold-standard leader in business analytics for more than 36 years,
```
SAS frees enterprises from the limitations of traditional computing and enables them to draw instant benefits from big data. Faster Time to Insight.

From banking to retail to health care to insurance, SAS is helping industries glean insights from data that once took days or weeks in just hours, minutes, or seconds. It's all about getting to and analyzing relevant data faster. Revealing previously unseen patterns, sentiments, and relationships. Identifying unknown risks. And speeding the time to insights.

High-Performance Analytics from SAS Combining industry-leading analytics software with high-performance computing technologies produces fast and precise answers to unsolvable problems and enables our customers to gain greater competitive advantage. SAS In-Memory Analytics eliminate the need for disk-based processing allowing for much faster analysis. SAS In-Database executes analytic logic into the database itself for improved agility and governance. SAS Grid Computing creates a centrally managed shared environment for processing large jobs and supporting a growing number of users efficiently. Together, the components of this integrated, supercharged platform are changing the decision-making landscape and redefining how the world solves big data business problems.

Big data is a popular term used to describe the exponential growth, availability and use of information, both structured and unstructured. Much has been written on the big data trend and how it can serve as the basis for innovation, differentiation and growth.

Computing singular value decomposition requires the input data to contain at least 25 documents and at least as many documents as there are machines in the cluster. By default, REDUCEF=4 but in this example is set to 2 to specify that a word only needs to appear twice to be kept for generating the term-by-document matrix. The default dimension for the singular-value decomposition is $k=10$ and the server generates ten topics.
The `TEXTPARSE` statement produces the following output. The names of the temporary tables are reported and begin with `_T_`. These names (and the rest of the output in the table) is stored in a temporary buffer that is named `TXTSUMMARY`.

<table>
<thead>
<tr>
<th>Data Source</th>
<th>HPS.GETSTART</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document Variable</td>
<td>docid</td>
</tr>
<tr>
<td>Text Variable</td>
<td>text</td>
</tr>
<tr>
<td>Number of Documents</td>
<td>36</td>
</tr>
<tr>
<td>Number of Terms</td>
<td>31</td>
</tr>
<tr>
<td>Term Information</td>
<td>_T_DBDFC26_7F21541ABD38</td>
</tr>
<tr>
<td>Transactional Terms (Bag of Words)</td>
<td>_T_DBDFC2B_7F21541ABE58</td>
</tr>
<tr>
<td>V Matrix from SVD</td>
<td>_T_DBDFC78_7F21541ABBF8</td>
</tr>
<tr>
<td>U Matrix from SVD</td>
<td>_T_DBDFC68_7F21541ABC08</td>
</tr>
<tr>
<td>Topics</td>
<td>_T_DBDFC80_7F21541ABDC8</td>
</tr>
<tr>
<td>Terms and Topics</td>
<td>_T_DBDFC88_7F21541ABDC8</td>
</tr>
<tr>
<td>Document Projection</td>
<td>_T_DBDFC92_7F21541ABC98</td>
</tr>
</tbody>
</table>

Because the `IMSTAT` procedure was used in interactive mode with a `RUN` statement instead of `QUIT`, the `STORE` statement can be used to create macro variables for the temporary table names as follows:

```sas
store txtsummary(6,2) as Terms;
store txtsummary(7,2) as Parent;
store txtsummary(8,2) as V;
store txtsummary(9,2) as U;
store txtsummary(10,2) as Topics;
store txtsummary(11,2) as TermsByTopics;
store txtsummary(12,2) as DocPro;
run;
```

Each of the tables can then be accessed with a `libref` and the macro variable, such as `example.&Terms.;`. The following sections show how to access these tables.

### Terms Table

Based on the statements in the previous section, you can access the terms table as follows:

```sas
table example.&Terms.;
umrows;
columninfo;
where _ispar ne ".";
fetch / format to=10;
run;
```
Filtering out the observations where _Ispar is missing results in showing only the terms that are used in the subsequent singular-value decomposition and topic generation.

The following output shows the number of observations in the Terms table, the data structure, and details about the terms that are identified by the TEXTPARSE statement.

<table>
<thead>
<tr>
<th>Number of Rows Action for Table _T_DBDFC26_7F21541ABD38</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Records</td>
</tr>
<tr>
<td>43</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column Information for Table _T_DBDFC26_7F21541ABD38</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Selected Records from Table _T_DBDFC26_7F21541ABD38</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term</td>
</tr>
<tr>
<td>sas</td>
</tr>
<tr>
<td>high-performance</td>
</tr>
<tr>
<td>big data</td>
</tr>
<tr>
<td>to</td>
</tr>
<tr>
<td>fast</td>
</tr>
<tr>
<td>data</td>
</tr>
<tr>
<td>gain</td>
</tr>
<tr>
<td>big</td>
</tr>
<tr>
<td>for</td>
</tr>
<tr>
<td>problem</td>
</tr>
</tbody>
</table>

**Parent Table**

The Parent table is also known as the bag of words or as the term document. It is a sparse representation of the term by document weights. It is the input to the singular-value decomposition (SVD). The SVD then reduces the dimensionality of the problem.
by focusing on the $k$ dimensions with the largest singular values. It is essentially a dimension reduction technique.

```plaintext
where; /* clear the _ispar filter that was in use */
table example.&Parent.;
umrows;
columninfo;
fetch / format to=10;
run;
```

There are 133 nonzero entries in the term × document matrix. The full matrix would be a $43 \times 36$ matrix.

| Column Information for Table _T_DBDFC2B_7F21541ABE58 |
|---|---|---|---|---|
| Id | Column | Type | Length | Format |
| 1 | _termnum_ | Num | 8 | BEST12. |
| 2 | _id_ | Num | 8 | BEST12. |
| 3 | _count_ | Num | 8 | BEST12. |
| 4 | docid | Char | 8 | $8$. |

| Selected Records from Table _T_DBDFC2B_7F21541ABE58 |
|---|---|---|---|
| _termnum_ | _id_ | _count_ | docid |
| 2 | 1 | 0.6934264036 | d01 |
| 4 | 1 | 0.3549850065 | d01 |
| 14 | 1 | 0.2842372535 | d01 |
| 23 | 1 | 0.8065735964 | d01 |
| 31 | 1 | 0.6934264036 | d01 |
| 3 | 2 | 0.5508777991 | d32 |
| 4 | 2 | 0.3549850065 | d32 |
| 6 | 2 | 0.4197207991 | d32 |
| 8 | 2 | 0.5 | d32 |
| 14 | 2 | 0.2842372535 | d32 |

There are 133 nonzero entries in the term × document matrix. The full matrix would be a $43 \times 36$ matrix.

There are four columns in the parent table. The _termnum_, and _id_ column are the term number and an internal identifier. Values in the _termnum_ column correspond to values in the Key column in the Terms table. The _count_ column represents the term
weight. The last column is the document ID that corresponds to the DOCID= variable specified in the TEXTPARSE statement. In this example, it is named Docid.

**SVD V Table**

```sas
table example.&V.;
numrows;
columninfo;
fetch / format to=10;
runch;
```

This table is one of the results of the singular-value decomposition of the sparse parent table (bag of words). The number of rows equals the number of documents in the input table. The number of columns equals the number of topics (the K= value of the SVD option) plus one for an ID variable.

### Output Tables for the TEXTPARSE Statement

**SVD U Table**

```sas
table example.&U.;
numrows;
columninfo;
fetch / format to=10;
```
This table contains the U matrix from the singular-value decomposition. The number of rows equals the number of terms in the SVD (31). The number of columns equals the number of topics (the K= value of the SVD option) plus an _ID_ column.

### Note:
The U matrix produced by the TEXTPARSE statement in the server does not match the U matrix that is generated by the HPTMINE procedure for SAS Text Miner. The SAS LASR Analytic Server performs a varimax rotation on the matrix. This step is done by SAS Text Miner on the HPTMINE output with the FACTOR procedure.

### Topics Table

```sql
table example.&Topics;
columninfo;
fetch / format to=10;
run;
```

This table displays the topics identified by the server, the term weight cutoff for the topic, and a descriptive label formed from the terms in the topic with the highest weights. By default, five terms are used to label the topic. You can specify a different number with the NUMLABELS= option.
### Terms by Topic Table

```r
table example.TermsByTopic;
columninfo;
fetch / format to=10;
run;
```

The terms-by-topic table is a sparse representation of the term-by-topic matrix.
The projected document contains at a minimum the document ID variable from the input table and the projection onto the topic space. The number of rows equals the number of documents in the input table. The number of columns equals the number of topics (the K= value of the SVD option) plus the document ID column. Each of the Col1 to ColK columns are the projections.
You can request that other variables, beside the document ID variable, are transferred to the projected document table. If you do not transfer variables, then the document projection table is a candidate for a join with the input table on the document ID in order to associate the documents with their topic weights.
SAS In-Memory Statistics for Hadoop is an offering that provides the data scientist or analytical expert with interactive programming access to in-memory data and integrates seamlessly with Hadoop.

In order to use the offering, the following must be true:

- You are using a distributed SAS LASR Analytic Server only.
- The SAS LASR Analytic Server is co-located with SAS High-Performance Deployment of Hadoop or a commercial Hadoop distribution that has been configured with the services from SAS High-Performance Deployment of Hadoop. The services enable you to use the SASHDAT file format for storing tables in HDFS.
- SAS/ACCESS Interface to Hadoop is configured on a client machine that you use for submitting SAS programs. Be sure to install the SAS Embedded Process on the machines in the Hadoop cluster. The SAS/ACCESS engine, the embedded process, and the HDMDS procedure enable you to describe your data that is in Hadoop and access it directly without an intermediate metadata repository such as Hive.
- SAS Studio provides an interactive web-based development application that enables you to write and submit SAS programs. Make sure that your user ID is configured for passwordless SSH to the machines in the cluster. Also make sure that you have passwordless SSH access from the machine that hosts SAS Studio to the machines in
Writing and Running SAS Programs

Log On to SAS Studio

Start a web browser and direct it to a URL that is similar to the following example:

http://hostname.example.com/SASStudio

In addition to the host name being different, the protocol might be HTTPS and you might need to specify a port number.

TIP After you log on, click in the toolbar to enable interactive mode. This mode is needed for working interactively with the IMSTAT procedure.

Start a SAS LASR Analytic Server

You start a distributed server with the LASR procedure. See the following example:

options set=GRIDHOST="grid001.example.com";
options set=GRIDINSTALLLOC="/opt/TKGrid_REP";
options set=GRIDMODE=asym;

proc lasr create port=10011;
  performance nodes=all;
run;

1 Using a TKGrid_REP installation location and the GRIDMODE=ASYM options enable reading data in parallel with SAS/ACCESS engines from distributed databases and Hadoop clusters.

Load a Table to Memory

Use the SAS LASR Analytic Server Engine

A common way to load tables into memory is to transfer them to the server using the SAS LASR Analytic Server engine. The following example transfers a table that is named Webscore to the server.

libname lasrlib sasiola host="grid001.example.com" port=10011;

data lasrlib.webscore;
  set somelib.webscore;
run;

Use the Hadoop Engine

Your Hadoop cluster might already have data in HDFS that you want to analyze, such as DBMS tables that were imported with Sqoop or log files that were imported with Flume. In this case, you can use the SAS/ACCESS Interface to Hadoop to access that data.
The LIBNAME statement with the Hadoop engine uses the HDFS_METADIR= option. This option enables working with XML-based table definitions called SASHDMD descriptors.

The HDMD procedure is used to create the SASHDMD descriptors from existing data in HDFS, which in this case is a delimited file.

For information about the Hadoop engine and the HDMD procedure, see SAS/ACCESS for Relational Databases: Reference.

Sample the Data

The following statements show one way to sample data. A calculated column is created with a number that is uniformly distributed between 0 and 1. The sampling is done based on the value of the new column.

```sas
%let seed = 12345;

proc imstat;
  table lasrlib.webdata;  
  tableinfo;
  columninfo;
  frequency goalVar;
run;

  compute sampkey "sampkey = ranuni(&seed.*)";  
run;

  table lasrlib.webdata;  
deleterows;
  where sampkey ge 0.31 and goalVar = 0;
run;
```
1 The TABLE statement specifies the Webdata table that was loaded to memory as the active table. The following three statements, TABLEINFO, COLUMNINFO, and FREQUENCY provide information about the table and a variable that is named Goalvar.

2 The COMPUTE statement creates a column that is named Sampkey. The column is permanent and is added to the table.

3 The TABLE statement is used again to reopen the table. This enables SAS to access the newly created column, Sampkey. The DELETEROWS statement is subject to the WHERE clause and marks 70% of the table for deletion where the goal was not met. Because the PURGE option is not used, the rows are not actually deleted. Instead, the rows are just disregarded in subsequent analyses that use the table.

Create a Forest of Decision Trees, Assess, and Score

The following code sample demonstrates using the RANDOMWOODS statement with training and validation data.

```
table lasrlib.webdata;
  where sampkey ge 0.3; /* training set */
  randomwoods goalVar /
    input  = ( browser var1-var8 )
    nominal = ( browser )
    nbins=100 maxlevel=10 maxbranches=2 /* tree specs */
    greedy gain leafsize=50
    ntree=100 seed=1314 m=5 /* forest spec */
    temptable
  ;
run;

table lasrlib.&_templast_; /* promote the model into */
  promote RF; /* a permanent table */
run;

table lasrlib.webdata;
  where sampkey lt 0.3; /* validation set */
  randomwoods /
    lasrtree  = lasrlib.RF
    nominal   = ( browser )
    temptable
    assess
    vars      = ( userid goalVar )
  ;
run;

table lasrlib.&_templast_; /* assess */
  where strip(_RF_Level_) eq '1';
  assess _RF_P_/ y = goalVar event = '1'
    nbins = 10 step = 0.001;
run;

table lasrlib.webscore; /* score */
  compute goalVar "goalVar = 2";
```
randomwoods /
lasrtree = lasrlib.RF
nominal = { browser }
temptable
assess
vars = { userid }
;
run;
   table lasrlib.&_templast_; 
promote scorerresult;
quit;

**About Passwordless SSH and SAS Studio**

When working with distributed SAS LASR Analytic Server, you must be configured for passwordless SSH. If SAS Studio is configured and running on a Windows host, you might need to perform the following steps to copy your SSH keys from a UNIX host to the Windows machine.

To copy your SSH keys to a Windows machine:

1. Determine your Windows home directory. Enter the following command in a command window:
   ```
echo %HOMEDRIVE%%HOMEPATH%
   ```
   The results are typically something like `C:\Users\sasdemo`.

2. You can use Windows Explorer to drag-and-drop the `.ssh` directory from your UNIX home directory, or you can use a command like the following to copy it:
   ```
   xcopy driverLetter:\.ssh\* "%HOMEDRIVE%%HOMEPATH%\.ssh" /s /i
   ```

---

**Deploying SAS In-Memory Statistics for Hadoop**

**Installation Sequence**

If SAS In-Memory Statistics for Hadoop is installed along with a SAS solution such as SAS Visual Analytics, then follow the steps that are provided in the installation guide for the solution. The software is automatically installed when it is delivered with a SAS solution.

If you are not installing the software as part of a solution, then you are performing a "Basic" installation instead of a "Planned" installation. Use the documents in the following sections to install the software.

**Software for Your Hadoop Cluster**

**Basic Steps**

Information about installing and configuring SAS High-Performance Analytics Environment, SAS High-Performance Deployment for Hadoop, and SAS High-Performance Computing Management Console is available in the *SAS High-*
Performance Analytics Infrastructure: Installation and Configuration Guide. This book is available at the following URL:

http://support.sas.com/documentation/solutions/hpainfrastructure/

Note: SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide directs you to install the SAS Embedded Process. You can install it only after SAS Foundation and SAS/ACCESS are installed.

If you are using SAS High-Performance Deployment of Hadoop, keep in mind the following items as you follow the procedures in SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide:

• When you follow the instructions in section "Preparing Your Data Provider for a Parallel Connection with SAS," the list of JAR files is not supplied for SAS High-Performance Deployment of Hadoop. There are no MapReduce 1 JAR files. The files for Common, HDFS, and MapReduce 2 are in the following directories:

  Common
  $HADOOP_HOME/share/hadoop/common/lib/*.jar
  $HADOOP_HOME/share/hadoop/common/*.jar

  HDFS
  $HADOOP_HOME/share/hadoop/hdfs/lib/*.jar
  $HADOOP_HOME/share/hadoop/hdfs/*.jar

  MapReduce 2
  $HADOOP_HOME/share/hadoop/mapreduce/lib/*.jar
  $HADOOP_HOME/share/hadoop/mapreduce/*.jar

  Copy the JAR files from those directories into a single directory on the machine that is used as the Hadoop NameNode.

• You need to perform the steps in section "Configuring for a Remote Data Store." Those steps install TKGrid_REP. When the installation script prompts you to specify the location of the Hadoop and client JAR files, enter the path to the single directory that has the collection of JAR files.

Additional Steps for SAS High-Performance Deployment of Hadoop

SAS High-Performance Deployment of Hadoop provides Apache Hadoop, SAS services that run inside Hadoop to enable working with the SASHDAT format, and an installation program that helps you to install and configure the software.

SAS High-Performance Deployment of Hadoop is initially configured for using HDFS and SAS LASR Analytic Server together. To enable use with the SAS Embedded Process that is needed to access data in formats other than SASHDAT in parallel, you need to configure Hadoop for MapReduce and Yarn services.

To configure the MapReduce and Yarn services:

1. Log on to the host for the Hadoop NameNode as root or with an account that has sudo privileges.

2. Edit the $HADOOP_HOME/etc/hadoop/yarn-site.xml file. Include the following properties. Substitute the host name of your NameNode for grid001.example.com:

   <?xml version="1.0"?>
   <configuration>

   ...
3. Edit the $HADOOP_HOME/etc/hadoop/mapred-site.xml file. Add the following property above the closing </configuration> tag:

```xml
<property>
  <name>mapreduce.framework.name</name>
  <value>yarn</value>
</property>
```

4. Copy the two files to the $HADOOP_HOME/etc/hadoop directory on the other machines in the cluster.

   **TIP** You can use /opt/TKGrid/bin/simcp to copy the files.

5. Edit /etc/profile.d/java.sh and include the following lines. Adjust the values to match your environment:

   ```bash
   export JAVA_HOME=/usr/lib/jvm/jre
   export PATH=$PATH:$JAVA_HOME/bin
   export YARN_HOME=/hadoop/hadoop-0.23.1
   ```

6. Copy the /etc/profile.d/java.sh file to the /etc/profile.d directory on the other machines in the cluster.

7. Using the account that you use to run Hadoop, start Yarn.

   ```bash
   su - hadoop  
   $HADOOP_HOME/sbin/start-yarn.sh  
   ```
SAS Foundation and Related Software

Install SAS Foundation
On the machine that you will use as the SAS client for writing and submitting SAS programs, run the SAS Deployment Wizard to install SAS Foundation. The type of SAS Studio that is installed depends on your host operating system. See the following sections for details.

About SAS Studio Basic and SAS Studio - Single User
SAS Studio is a development application for writing SAS programs and submitting them. You can access SAS Studio through your web browser.

- SAS Studio Basic is included with an order for SAS In-Memory Statistics for Hadoop on Linux for x64.
- SAS Studio - Single User is included with an order for SAS In-Memory Statistics for Hadoop on Windows.


UNIX Hosts
Use the SAS Deployment Wizard to install SAS. Refer to the documentation for UNIX hosts at the following URL:

http://support.sas.com/documentation/installcenter/94/unx/index.html

When you run the SAS Deployment Wizard, you can specify to install SAS Studio Basic. Refer to "SAS Studio Basic" in the SAS Studio: Administrator's Guide.

After the SAS Deployment Wizard installs the software, be sure to follow the instructions for configuring Hadoop JAR files. Refer to Configuration Guide for SAS 9.4 Foundation for UNIX Environments available at the preceding URL.

Windows Hosts
Use the SAS Deployment Wizard to install SAS. Refer to the documentation for Windows hosts at the following URL:

http://support.sas.com/documentation/installcenter/94/win/index.html

After the SAS Deployment Wizard installs the software, be sure to follow the instructions for configuring Hadoop JAR files. Refer to Configuration Guide for SAS 9.4 Foundation for Microsoft Windows for x64 available at the preceding URL.

SAS In-Database Products and SAS Embedded Process
After SAS Foundation and SAS/ACCESS are installed, follow the instructions in the SAS In-Database Products: Administrator's Guide to install the deployment package for Hadoop on the cluster.

Note: SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide that is mentioned in a preceding section directs you to install the SAS Embedded Process, too. You do not need to install the SAS Embedded Process...
twice, it is mentioned at this point because it is possible to install it only after SAS
Foundation and SAS/ACCESS are installed.

To perform the procedure for installing the deployment package, you need a list of JAR
files. If you are using SAS High-Performance Deployment of Hadoop, then use the same
directory with the JAR files that you collected earlier. You need to copy the contents of
the directory to the client machine that you will use to run SAS programs. As described
in SAS In-Database Products: Administrator's Guide, you also need to copy the
sas.hadoop.ep.apache023.jar and sas.hadoop.ep.apache023.nls.jar
files to the same directory on the client machine. This is the directory that you will
specify as the SAS_HADOOP_JAR_PATH on the client machine.
Appendix 2
Removing a Machine from the Cluster

About Removing a Machine from the Cluster

This information applies to deployments that use a distributed SAS LASR Analytic Server and SAS High-Performance Deployment of Hadoop.

This information does not apply to non-distributed deployments.

For deployments that use Teradata EDW, Greenplum DCA, or a commercial distribution of Hadoop, such as Cloudera or Hortonworks, follow the product documentation for the appliance or cluster. If it is possible to run the appliance or cluster in a degraded state, you can follow the steps to remove the host name from the grid.hosts file. This can make it possible to start and run SAS LASR Analytic Server instances.

Which Servers Can I Leave Running When I Remove a Machine?

In most cases, all SAS LASR Analytic Server instances stop automatically if communication with any machine in the cluster fails. However, if any servers are still running, stop them.

Stop SAS High-Performance Deployment of Hadoop to avoid logging messages about the machine being unavailable. For SAS Visual Analytics deployments, stop the SAS LASR Analytic Server Monitor, too.
Remove the Host Name from the grid.hosts File

The host name for the machine to remove must be removed from the `/opt/TKGrid/grid.hosts` file. (The path might be different for your deployment.)

To remove the host name:

1. Log on to the root node for the deployment as the root user or the user ID that installed the High-Performance Analytics Environment.
2. Edit the `/opt/TKGrid/grid.hosts` file and delete the line that includes the host name to remove. Save and close the file.
3. Copy the updated `grid.hosts` file to the same location on all the machines in the cluster. You can use the `simcp` command to perform this task:

   ```bash
   /opt/TKGrid/bin/simcp /opt/TKGrid/grid.hosts /opt/TKGrid
   ```

   If your deployment uses the SAS High-Performance Computing Management Console, remove the host name from the `/etc/gridhosts` file, too.

Remove the Host Name from the Hadoop Slaves File

If the machine is still running, stop any Hadoop processes that are running on it before you continue.

To remove the host name from the Hadoop slaves file:

1. Log on to the root node for the deployment as the root user or the user ID that installed the SAS High-Performance Deployment of Hadoop.
2. Edit the `/hadoop/hadoop-0.23.1/etc/hadoop/slaves` file and delete the line that includes the host name to remove. Save and close the file.

   The path might be different for your deployment.

Restart Servers

You can restart SAS LASR Analytic Server instances at this point. When you load tables from HDFS, data are loaded from redundant blocks on the remaining machines.

When the machine becomes available again (such as replacement hardware or new hardware), follow the steps in “Adding Machines to the Cluster”.

---

Appendix 2 • Removing a Machine from the Cluster
Appendix 3
Adding Machines to the Cluster

About Adding Machines to the Cluster

This information applies to deployments that use a distributed SAS LASR Analytic Server and SAS High-Performance Deployment of Hadoop. These steps do not include information about licensing SAS software for additional CPUs. Contact your SAS
representative for information about getting additional licensing and applying the license.

This information does not apply to non-distributed deployments.

For deployments that use Teradata EDW, Greenplum DCA, or a commercial distribution of Hadoop, follow the product documentation for information about expanding the appliance or cluster. After following those procedures, install the SAS High-Performance Analytics environment as described in *SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide*.

### Which Servers Can I Leave Running When I Add a Machine?

It is best to stop all SAS LASR Analytic Server instances and SAS High-Performance Deployment of Hadoop. For SAS Visual Analytics deployments, stop the SAS LASR Analytic Server Monitor, too.

If you prefer to deny access to the environment while performing this procedure, you can use the SAS High-Performance Computing Management Console to perform an SSH lockout. Be sure to permit access for root, the SAS installer account, the account that is used to run HDFS, and at least one administrator that can start and stop server instances.

You can ensure that saving files to HDFS is denied by using safe mode. The following command is an example:

```
$HADOOP_HOME/bin/hdfs dfsadmin -safemode enter
```

### Configure System Settings

Each of the additional machines must be configured identically to the existing machines with regard to operating system, drivers, and tuning settings. The high-level tasks are as follows:

- Configure passwordless SSH for the root user ID.
- Disable SELinux if it is disabled on the existing machines.
- Modify `/etc/ssh/sshd_config` with the following setting:
  ```
  MaxStartups 1000
  ```
- Modify `/etc/security/limits.conf` with the following settings:
  ```
  soft nproc 65536
  hard nproc 65536
  soft nofile 350000
  hard nofile 350000
  ```
- Modify `/etc/security/limits.d/90-nproc.conf` with the following setting:
  ```
  soft nproc 65536
  ```
- Modify `/etc/sysconfig/cpuspeed` with the following setting:
Add Host Names to Gridhosts

If you use SAS High-Performance Computing Management Console to manage your cluster, you must add the host names for the additional machines to the /etc/gridhosts file. You can do this with a text editor, or you can use the Gridhosts File Management feature in the console. If you do not use the console for managing the cluster, then you do not need to perform this task.

If you use the console to add the hosts, make sure that the SSH Connectivity column indicates Yes for all the machines. The following display shows an example.

A No in the SSH Connectivity column indicates that passwordless SSH for the root user ID is not configured properly for the machine identified in the Node Name column.
Propagate Operating System User IDs

Which User IDs Must Be Propagated?

You must propagate the operating system user IDs that are used for Hadoop and for managing SAS LASR Analytic Server instances (starting, stopping, and loading tables).

About Passwords

In most deployments, passwords are not used to log on to the machines in the cluster. Therefore, in most cases, it is not necessary to propagate passwords to the additional machines in the cluster.

However, if you want to preserve passwords and you use the SAS High-Performance Computing Management Console, you can view the value from the `/etc/shadow` file or within the console. When you add the user with the console, you can paste the value in the Pre-encrypted password field.

Option 1: User Management Software

If your site uses account management software applications such as LDAP, NIS, or Active Directory on Linux for managing user accounts, then use that software to make the user accounts available on the additional machines.

Option 2: Delete and Re-Add Users

If you use SAS High-Performance Computing Management Console to manage the machines in the cluster, then you can delete each user and then add the user back to the system. This is an option when the number of operating system accounts is fairly low. When you add the user, the account is re-created on the original machines, and it is also added to the new machines.

Note: When using this method, be sure to note the UID and primary GID of the user before it is deleted and to reuse the same values when re-creating the user account.

Note: If you choose this option, be aware that using the Generate and Propagate SSH Keys option when you create the user account removes existing SSH keys. If you do not delete the home directory when you delete the user and you do not generate new SSH keys, then the existing keys can be reused.

Option 3: Use Operating System Commands to Add Users

About Using the Simultaneous Shell Command

You can view the existing user IDs and groups with the SAS High-Performance Computing Management Console, or from the `/etc/passwd` and `/etc/group` files. You can use operating system commands to add the groups and users to the addition machines.

The following sample commands use the `simsh` utility that is included with SAS High-Performance Computing Management Console. This utility attempts to configure users
and groups on the machines that already have the user accounts and groups. This results in an error message from those machines. The error message is harmless because the commands do not modify the existing configuration, but you might find them distracting. As an alternative, you can use the console to create a simultaneous utilities machine group that contains the host names for the new machines only. You can then specify the group name with the `simsh` command so that only the new machines are affected.

This document demonstrates using the `simsh` command with a machine group that is named `newnodes` for simplicity.

### Add Groups

Identify the groups to add to the new machine by viewing the console or looking at the `/etc/group` file. Make sure that you identify each group ID.

The following example shows how to add the group that is named `sasdemo` with a group number of 102 to the `newnodes` machine group:

```
/opt/webmin/utilbin/simsh -g newnodes "groupadd -g 102 sasdemo"
```

### Add Users

Identify the user IDs to add to the new machine by viewing the console or looking at the `/etc/passwd` file. Make sure that you identify each user ID.

The following example shows how to add a user:

```
/opt/webmin/utilbin/simsh -g newnodes "useradd -u 503 -g 102 -d /home/sasdemo -s /bin/bash sasdemo"
```

*Note:* The command must be entered on a single line.

### Propagating Secure Shell Keys

One way to propagate existing SSH keys to the new machines is to copy them to all the new machines in the cluster. The following example shows one way to perform this operation:

```
simcp /home/user/.ssh /home/user/
simsh chown -R user:group /home/user/.ssh
```

You can use scripting to simplify this task and all the previous operating system commands, too. You also do not need to follow this strategy. Any method that is able to propagate the groups, user IDs, and SSH keys is acceptable.

---

## Configure SAS High-Performance Deployment of Hadoop

### Install a Java Runtime Environment

A Java Runtime Environment or Java Development Kit is necessary to run SAS High-Performance Deployment of Hadoop on the additional machines. Be sure to install it in
the same path as the existing machines. One way to perform this task is to use the
\texttt{simcp} command as shown in the following example:

\begin{verbatim}
/opt/webmin/utilbin/simsh -g newnodes mkdir -p /data/java/

/opt/webmin/utilbin/simcp -g newnodes /data/java/jdk_1.6.0_29
/data/java/
\end{verbatim}

\section*{Install Hadoop on the Additional Machines}

Install Hadoop on the additional machines with the same package, \texttt{sashadoop.tar.gz}, that is installed on the existing machines. Install the software with the same user account and use the same installation path as the existing machines.

Typically, you create a text file with the host names for all the machines in the cluster and supply it to the installation program. In this case, create a text file with the host names for the new machines only. Use a filename such as \texttt{-/addhosts.txt}. When you run the installation program, \texttt{hadoopInstall}, supply the fully qualified path to the \texttt{addhosts.txt} file.

The previous tasks result in a new cluster that is independent of the existing cluster. When the configuration files are overwritten in the next step, the additional machines no longer belong to their own cluster. They become part of the existing cluster.

When you run the installation program on the new machines, if you are unsure of the directory paths to specify, you can view the following files on an existing machine:

\begin{verbatim}
$HADOOP_HOME/etc/hadoop/hdfs-site.xml
\end{verbatim}

Look for the values of the \texttt{dfs.name.dir} and \texttt{dfs.data.dir} properties.

\begin{verbatim}
$HADOOP_HOME/etc/hadoop/mapred-site.xml
\end{verbatim}

Look for the values of the \texttt{mapred.system.dir} and \texttt{mapred.local.dir} properties.

\section*{Update the Hadoop Configuration Files}

As the user ID that runs HDFS, modify the \texttt{$HADOOP_HOME/etc/hadoop/slaves} file on the existing machine that is used for the NameNode. Add the host names of the additional machines to the file.

You can use the \texttt{simcp} command to copy the file and other configuration files to the new machines:

\begin{verbatim}
/opt/webmin/utilbin/simcp -g newnodes $HADOOP_HOME/etc/hadoop/slaves $HADOOP_HOME/etc/hadoop/

/opt/webmin/utilbin/simcp -g newnodes $HADOOP_HOME/etc/hadoop/master $HADOOP_HOME/etc/hadoop/

/opt/webmin/utilbin/simcp -g newnodes $HADOOP_HOME/etc/hadoop/core-site.xml $HADOOP_HOME/etc/hadoop/

/opt/webmin/utilbin/simcp -g newnodes $HADOOP_HOME/etc/hadoop/hdfs-site.xml $HADOOP_HOME/etc/hadoop/
\end{verbatim}
Start the DataNodes on the Additional Machines

For each of the new machines, run a command that is similar to the following example:

```
ssh hostname /data/hadoop/hadoop-0.23.1/sbin/hadoop-daemon.sh start datanode
```

**Note:** Run the command as the user account that is used for HDFS.

Make sure that you specify the actual path to `hadoop-daemon.sh`. Once you have started the DataNode process on each new machine, view the [http://namenode-machine:50070/dfshealth.jsp](http://namenode-machine:50070/dfshealth.jsp) page to view the number of live nodes.

Run `$HADOOP_HOME/bin/hdfs dfsadmin -printTopology` to confirm that the new machines are part of the cluster. The following listing shows a sample of the command output:

```
Rack: /default-rack
  192.168.8.148:50010 (grid103.example.com)
  192.168.8.153:50010 (grid104.example.com)
  192.168.8.217:50010 (grid106.example.com)
  192.168.8.230:50010 (grid105.example.com)
  192.168.9.158:50010 (grid109.example.com)
  192.168.9.159:50010 (grid100.example.com)
  192.168.9.160:50010 (grid101.example.com)
```

Copy a File to HDFS

If you put HDFS in safe mode at the beginning of this procedure, leave that state with a command that is similar to the following:

```
$HADOOP_HOME/bin/hdfs dfsadmin -safemode leave
```

To confirm that the additional machines are used, you can copy a file to HDFS and then list the locations of the blocks. Use a command that is similar to the following:

```
$HADOOP_HOME/bin/hadoop fs -D dfs.blocksize=512 -put /etc/fstab /hps
```

**Note:** The very small block size shown in the example is used to increase the number of blocks written and increase the likelihood that the new machines are used.

You can list the block locations with a command that is similar to the following:

```
$HADOOP_HOME/bin/hdfs fsck /hps/fstab -files -locations -blocks
```

Review the output to check for IP addresses for the new machines.
Configure SAS High-Performance Analytics Infrastructure

Strategies

The SAS High-Performance Analytics Infrastructure software must be installed on the new machines in the cluster. In addition, the existing installations must be updated so that the grid.hosts file includes the new host names.

The first option is to re-install the software. This adds the software to the new machines and updates the grid.hosts file. This option has the advantage of being very simple.

The second option is to copy the files to the new machines and then copy an updated grid.hosts file to all the machines.

**T I P** Installing the software to a new directory is not suggested because the path to the software might be specified in numerous server definitions that are registered in SAS metadata.

Option 1: Re-Install the Software

The software is installed by running the TKGrid/Linux_x86_64.sh executable. For details about installing the software, see SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide.

If you choose this option, stop all SAS LASR Analytic Server instances. Stopping the servers avoids the possibility of errors related to overwriting executables and libraries.

Option 2: Copy the Software

Use a command that is similar to the following for copying the software to the new machines:

```bash
/opt/webmin/utilbin/simcp -g newnodes /opt/TKGrid/ /opt/TKGrid/
```

Modify the `/opt/TKGrid/grid.hosts` file and add the host names for the new machines. Then, copy the file to all machines, even the existing machines:

```bash
/opt/webmin/utilbin/simcp /opt/TKGrid/grid.hosts /opt/TKGrid/
```

Validate the Change

Use the mpirun command to confirm that the new machines are accessible.

```bash
cd /opt/TKGrid/mpich2-install

./bin/mpirun -f ../grid.hosts hostname
```
The **hostname** command is run on each machine and the results are returned. Make sure that the response includes all the host names in the cluster.

```
grid098.example.com
grid103.example.com
grid106.example.com
grid100.example.com
grid105.example.com
grid101.example.com
grid104.example.com
grid099.example.com
```

---

**Restart SAS LASR Analytic Server Monitor**

For SAS Visual Analytics deployments, restart the monitor:

```
    cd SAS-config-dir/Levn/Applications/SASVisualAnalyticsX.X/
    HighPerformanceConfiguration/LASRMonitor.sh restart
```

Restarting the monitor causes an error for any users logged on to SAS Visual Analytics Administrator. Those users need to log off and log on again.

---

**Restart Servers and Redistribute HDFS Blocks**

Log on to SAS Visual Analytics Administrator and perform the following steps:

1. Select **LASR ⇨ Monitor Resources** and sure that the additional machines appear in the Real-Time View.
2. Select **LASR ⇨ Manage Servers** and start each of the servers.
3. Select **Tools ⇨ Explore HDFS** and review the block distribution.

Any tables that are loaded from HDFS cannot initially use the additional hardware because the blocks have not been replicated to the additional machines. Within SAS Visual Analytic Administrator, you can view the block distribution from the **HDFS** tab to confirm that blocks are on the original machines only.

Hadoop includes a balancer process that can move blocks to under-utilized machines. The process is intentionally slow so that resource consumption is low and so that it does not interfere with other tasks on the cluster. To begin the balancing process:

```
./hadoop-daemon.sh start balancer
```

As an alternative, if adding the data to HDFS again is possible, then you can delete files with SAS Visual Analytics Administrator and then add them back.

---

**View Explorations and Reports**

To confirm that the additional machines are working as intended, view existing explorations and reports.
If you did not delete and then add the data back to HDFS, then make sure that you view explorations and reports that use data that is streamed to the server instance (instead of being loaded from HDFS). Or, make sure that new data sets are added to HDFS and create explorations and reports from the new data.
Glossary

Apache Hadoop
a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model.

BY-group processing
the process of using the BY statement to process observations that are ordered, grouped, or indexed according to the values of one or more variables. Many SAS procedures and the DATA step support BY-group processing. For example, you can use BY-group processing with the PRINT procedure to print separate reports for different groups of observations in a single SAS data set.

colocated data provider
a distributed data source, such as SAS Visual Analytics Hadoop or a third-party vendor database, that has SAS High-Performance Analytics software installed on the same machines. The SAS software on each machine processes the data that is local to the machine or that the data source makes available as the result of a query.

grid host
the machine to which the SAS client makes an initial connection in a SAS High-Performance Analytics application.

Hadoop Distributed File System
a framework for managing files as blocks of equal size, which are replicated across the machines in a Hadoop cluster to provide fault tolerance.

HDFS
See Hadoop Distributed File System.

Message Passing Interface
is a message-passing library interface specification. SAS High-Performance Analytics applications implement MPI for use in high-performance computing environments.

MPI
See Message Passing Interface.

root node
in a SAS High-Performance Analytics application, the role of the software that distributes and coordinates the workload of the worker nodes. In most deployments
the root node runs on the machine that is identified as the grid host. SAS High-Performance Analytics applications assign the highest MPI rank to the root node.

**SASHDAT file**
the data format used for tables that are added to HDFS by SAS. SASHDAT files are read in parallel by the server.

**server description file**
a file that is created by a SAS client when the LASR procedure executes to create a server. The file contains information about the machines that are used by the server. It also contains the name of the server signature file that controls access to the server.

**signature file**
small files that are created by the server to control access to the server and to the tables loaded in the server. There is one server signature file for each server instance. There is one table signature file for each table that is loaded into memory on a server instance.

**worker node**
in a SAS High-Performance Analytics application, the role of the software that receives the workload from the root node.
Index

A
ADD statement
  RECOMMEND procedure 264, 265
appending data 306
ARM statement
  IMSTAT procedure 45
ASSESS statement
  IMSTAT procedure 57

B
BALANCE statement
  IMSTAT procedure 197
BOXPLOT statement
  IMSTAT procedure 60

C
CLUSTER statement
  IMSTAT procedure 63
COLUMNINFO statement
  IMSTAT procedure 197
COMPUTE statement
  IMSTAT procedure 198
connecting to a Hadoop NameNode
  LIBNAME statement, SAS Data in HDFS engine 320
connecting to a server
  LIBNAME statement, SAS LASR Analytic Server engine 303
CORR statement
  IMSTAT procedure 69
CROSSTAB statement
  IMSTAT procedure 69

D
DECISIONTREE statement
  IMSTAT procedure 74
DELETEROWS statement
  IMSTAT procedure 199
DISTINCT statement
  IMSTAT procedure 80
DISTRIBUTIONINFO statement
  IMSTAT procedure 200
DROPTABLE statement
  IMSTAT procedure 200

E
eengine 295, 315

F
FETCH statement
  IMSTAT procedure 200
FORECAST statement
  IMSTAT procedure 87
formats
  LASR procedure 35
FREE statement
  IMSTAT procedure 203
FREQUENCY statement
  IMSTAT procedure 90

G
GENMODEL statement
  IMSTAT procedure 92
GLM statement
  IMSTAT procedure 103
gpfdist
  distributing data 9
Greenplum
  distributing data 7
GROUPBY statement
  IMSTAT procedure 118

H
HDFS
accessing with the SAS LASR Analytic Server engine 305
HISTOGRAM statement
IMSTAT procedure 122
HPDS2 procedure
SAS Data in HDFS engine 322

I
IMPORTCUBE statement
IMSTAT procedure 204
IMSTAT procedure
concepts 42, 188
temporary variables 192
WHERE clause processing 189
IMSTAT procedure examples
appending tables 243
appending tables to partitioned tables 244
boxplot with NOUTLIERLIMIT= 175
cluster analysis 176
correlation 177
crosstabulation 178
decisiontree 180, 182
multi-dimensional summary 184
partitioning 235
percentiles and quartiles 172
promote temporary tables 237
rebalancing a table 238
regression models 185
retrieving box plot values 174
saving tables 240
star schema 241
storing temporary variables 245
IMXFER procedure 247
IMXFER procedure examples
copying a table 251, 252
INFO statement
RECOMMEND procedure 266

K
KDE statement
IMSTAT procedure 124

L
LABEL= option 257
LASR procedure
accessing tables with the SAS LASR Analytic Server engine 305
compared with the SAS LASR Analytic Server engine 296
concepts 23
FMTLIBXML= option 37
LASR procedure examples
load a table from Greenplum 35
load a table from Teradata 34
Loading a table from HDFS 33
logging 33
saving tables 38
starting a server 32, 356
stopping a server 36
unload a table from memory 36
user-defined formats 37
working with formats 37
LIBNAME statement, SAS Data in HDFS engine
syntax 319
LIBNAME statement, SAS LASR Analytic Server engine
syntax 299
LIFETIME statement
IMSTAT procedure 205
logging
default log file location 15
insufficient authorization 16
LOGISTIC statement
IMSTAT procedure 125

M
MDSUMMARY statement
IMSTAT procedure 140
METHOD statement
RECOMMEND procedure 267

N
NUMROWS statement
IMSTAT procedure 205

O
OLIPHANT procedure
concepts 253
syntax 255
OLIPHANT procedure examples
adding files to HDFS 258
querying file details from HDFS 259
OPTIMIZE statement
IMSTAT procedure 142

P
PARTITION statement
IMSTAT procedure 206
PARTITIONINFO statement
IMSTAT procedure 207
PERCENTILE statement
IMSTAT procedure 147
PERFORMANCE statement
LSR procedure 29
PREDICT statement
   RECOMMEND procedure 274
PROC IMSTAT statement 43, 195
PROC IMXFER statement 248
PROC LASR statement 24
PROC OLIPHANT statement 255
PROC RECOMMEND statement 263
PROC VASMP statement 288
programming 356
PROMOTE statement
   IMSTAT procedure 208
PURGETEMPTABLES statement
   VASMP procedure 209

Q
QUIT statement
   VASMP procedure 171, 234, 251, 291

R
RANDOMWOODS statement
   IMSTAT procedure 151
RECCORR statement
   IMSTAT procedure 156
REMOVE statement
   LASR procedure 31
   RECOMMEND procedure 275
REPLAY statement
   IMSTAT procedure 209

S
sampling data 357
   IMSTAT procedure 357
SAS Data in HDFS engine
   BLOCKSIZE= data set option 325
   COPIES= data set option 327, 328
   FILETYPE= data set option 327
   GUESSROWS= data set option 329
   HASH data set option 329
   how it works 315
   LOGUPDATE data set option 330
   ORDERBY= data set option 330
   PARTITION= data set option 331
   PATH= data set option 332
   PERM= data set option 332
   requirements for using 316
   what is supported 316
SAS In-Memory Statistics for Hadoop 355
SAS LASR Analytic Server engine
   accessing tables loaded from HDFS 305
   accessing tables loaded with a DATA step 304
   accessing tables loaded with the LASR procedure 305
   APPEND= data set option 309
   ARRAY= data set option 310
   AUTOCOMMIT= data set option 310
   compared with the LASR procedure 296
   FORMATEXPORT= data set option 311
   HASH= data set option 311
   ORDERBY= data set option 312
   PARTITION= data set option 312
   PERMISSION= data set option 312
   SIGNER= data set option 313
   TAG= data set option 313
   TEMPNAMES= data set option 314
   UCA= data set option 314
   understanding server tags 296
SAS LASR Analytic Server engine SAS Data in HDFS engine
   how it works 295
   requirements for using 297
   what is supported 297
SAS/ACCESS engines 9
SAVE statement
   IMSTAT procedure 210
   LASR procedure 31
SCHEMA statement
   IMSTAT procedure 210
SCORE statement
   IMSTAT procedure 213
security
   hadoop command 317
   server run time 8
   SERVER statement 249
   server tag 296
   accessing a table loaded with a DATA step 304
   accessing tables loaded from HDFS 305
   accessing tables loaded with the LASR procedure 305
   SERVERINFO statement 288
   IMSTAT procedure 220
   SERVERPARAM statement 289
   IMSTAT procedure 221
   SERVERTERM statement 290
   IMSTAT procedure 222
   SERVERWAIT statement
   IMSTAT procedure 223
   VASMP procedure 290
SET statement
   IMSTAT procedure 223
specifying host and port
LIBNAME statement, SAS LASR Analytic Server engine 304
specifying host and software installation location
LIBNAME statement, SAS Data in HDFS engine 321
starting a server 356
STORE statement
IMSTAT procedure 224
SUMMARY statement
IMSTAT procedure 158

T
TABLE statement 249
IMSTAT procedure 231
TABLEINFO statement 291
IMSTAT procedure 232
temporary tables 13
ARM statement 53
association rules 55
BALANCE statement 197
frequent itemsets 53
managing memory 13
partitioned 190
SCHEMA statement 210
sequences 56
temporary variables 192
example 193
TEMPNAMES= data set option 314
Teradata
distributing data 10
TEXTPARSE statement
IMSTAT procedure 164
TOPK statement
IMSTAT procedure 169

U
UPDATE statement
IMSTAT procedure 232

V
VASMP procedure 287
VASMP procedure examples
more than one Hadoop installation 292

W
WHERE clause
processing in the IMSTAT procedure 189