SAS® Text Miner 12.3

High-Performance Procedures

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2013. *SAS® Text Miner 12.3: High-Performance Procedures*. Cary, NC: SAS Institute Inc.

SAS® Text Miner 12.3: High-Performance Procedures

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

July 2013

Contents

Credits and Acknowledgments

Credits

Documentation

Editing	Anne Baxter, Ed Huddleston
Documentation Support	Tim Arnold, Melanie Gratton

Software

The SAS high-Performance text mining procedures were implemented by the following members of the development staff. Program development includes design, programming, debugging, support, and documentation. In the following list, the names of the developers who currently provide primary support are listed first; other developers and previous developers are also listed.

HPTMINE	Zheng (Alan) Zhao, Russell Albright, David Bultman, Joshua Griffin, James Cox
HPTMSCORE	Zheng (Alan) Zhao, Russell Albright, David Bultman, Joshua Griffin, James Cox
High-performance computing foundation	Steve E. Krueger
High-performance analytics foundation	Robert Cohen, Georges H. Guirguis, Trevor Kearney, Richard Knight, Gang Meng, Oliver Schabenberger, Charles Shorb, Tom P. Weber
Numerical routines	Georges H. Guirguis

The following people contribute to the SAS high-Performance text mining procedures with their leadership and support: Saratendu Sethi, Bob Johnson.

Testing

Leslie Anderson, Tim Carter, Enzo D'Andreti, Alex Fang, Angela Jones, Paul Kovach, Jim McKenzie, Jim Metcalf, Kelly Miller, Laura Rock, Kim Stott, Carol Thompson.

Internationalization Testing

Alex Chai, Mi-Na Chu, Jacky Dong, Feng Gao, Masayuki Iizuka, David Li, Lan Luan, Haiyong Rong, Bin Sun, Frank Wang, Lina Xu, Catherine Yang.

Technical Support

Craig DeVault, Ann Kou.

Acknowledgments

Many people make significant and continuing contributions to the development of SAS software products.

The final responsibility for the SAS System lies with SAS alone. We hope that you will always let us know your opinions about the SAS System and its documentation. It is through your participation that SAS software is continuously improved.

Please visit us at http://www.sas.com/software/high-performance-computing/.

Chapter 1
Introduction

Contents

Overview of the SAS Text Miner High-Performance procedures

The SAS Text Miner high-performance procedures provide text mining tools that have been specially developed to take advantage of parallel processing in both multithreaded single-machine mode and distributed multiple machine mode. The SAS Text Miner high-performance procedures provide full-spectrum support for text mining, including document parsing, term weighting and filtering, term-by-document matrix creation, dimensionality reduction via singular value de-composition (SVD), and scoring.

In addition to the procedures described in this book, SAS Text Miner includes high-performance utility procedures, which are described in *Base SAS Procedures Guide: High-Performance Procedures*. You can run all of these procedures in single-machine mode without licensing SAS High-Performance Text Miner. However, to run these procedures in distributed mode, you must license SAS High-Performance Text Miner.

Installation and Configuration

Before you can use the SAS Text Miner high-performance procedures, you need to install and configure the SAS High-Performance Analytics infrastrucure. For more information, see the *SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide*.

Additional installation and configuration is required for SAS Text Miner high-performance procedures. For more information, see the section “System Configuration” on page 61 in Chapter 3, “The HPTMINE Procedure.”

About This Book

This book assumes that you are familiar with Base SAS software and with the books *SAS Language Reference: Concepts* and the *Base SAS Procedures Guide*. It also assumes that you are familiar with basic SAS System concepts, such as using the DATA step to create SAS data sets and using Base SAS procedures (such as, the PRINT and SORT procedures) to manipulate SAS data sets.

Chapter Organization

This book is organized as follows.

Chapter 1, this chapter, provides an overview of SAS Text Miner high-performance procedures.

Chapter 2, "Shared Concepts and Topics," describes the modes in which SAS Text Miner high-performance procedures can execute.

Subsequent chapters describe the SAS Text Miner high-performance procedures. These chapters appear in alphabetical order by procedure name. Each chapter is organized as follows:

- The "Overview" section provides a brief description of the analysis provided by the procedure.
- The "Getting Started" section provides a quick introduction to the procedure through a simple example.
- The "Syntax" section describes the SAS statements and options that control the procedure.
- The "Details" section discusses methodology and miscellaneous details, such as ODS tables.
- The "Examples" section contains examples that use the procedure.
- The "References" section contains references for the methodology and for examples of the procedure.

Typographical Conventions

This book uses several type styles for presenting information. The following list explains the meaning of the typographical conventions used in this book:

roman	is the standard type style used for most text.
UPPERCASE ROMAN	is used for SAS statements, options, and other SAS language elements when they appear in the text. However, you can enter these elements in your own SAS programs in lowercase, uppercase, or a mixture of the two.
UPPERCASE BOLD	is used in the "Syntax" sections' initial lists of SAS statements and options.
oblique	is used for user-supplied values for options in the syntax definitions and in text.
VariableName	is used for the names of variables and data sets when they appear in the text.
bold	is used to refer to matrices and vectors.

italic	is used for terms that are defined in the text, for emphasis, and for references to publications.
`monospace`	is used for example code. In most cases, this book uses lowercase type for SAS code.

Options Used in Examples

Output of Examples

Most of the output shown in this book is produced with the following SAS System options:

```
options linesize=80 pagesize=500 nonumber nodate;
```

The HTMLBLUE style is used to create the HTML output and graphs that appear in the online documentation. A style template controls stylistic elements such as colors, fonts, and presentation attributes. The style template is specified in the ODS HTML statement as follows:

```
ods html style=HTMLBlue;
```

If you run the examples, you might get slightly different output. This is a function of the SAS System options used and the precision used by your computer for floating-point calculations.

Online Documentation

You can access the documentation by going to `http://support.sas.com/documentation`

SAS Technical Support Services

As with all SAS products, the SAS Technical Support staff is available to respond to problems and answer technical questions regarding the use of the SAS Text Miner high-performance procedures. Go to `http://support.sas.com/techsup` for more information.

Chapter 2
Shared Concepts and Topics

Contents

Overview

This chapter describes the modes of execution in which SAS high-performance analytical procedures can execute. If you have SAS Text Miner installed, you can run any procedure in this book on a single machine.

However, to run procedures in this book in distributed mode, you must also have SAS High-Performance Text Miner software installed. For more information about these modes, see the next section.

This chapter provides details of how you can control the modes of execution and includes the syntax for the PERFORMANCE statement, which is common to all high-performance analytical procedures.

Processing Modes

Single-Machine Mode

Single-machine mode is a computing model in which multiple processors or multiple cores are controlled by a single operating system and can access shared resources, such as disks and memory. In this book, single-machine mode refers to an application running multiple concurrent threads on a multicore machine in order to take advantage of parallel execution on multiple processing units. More simply, single-machine mode for high-performance analytical procedures means multithreading on the client machine.

All high-performance analytical procedures are capable of running in single-machine mode, and this is the default mode when a procedure runs on the client machine. The procedure uses the number of CPUs (cores) on the machine to determine the number of concurrent threads. High-performance analytical procedures use different methods to map core count to the number of concurrent threads, depending on the analytic task. Using one thread per core is not uncommon for the procedures that implement data-parallel algorithms.

Distributed Mode

Distributed mode is a computing model in which several nodes in a distributed computing environment participate in the calculations. In this book, the distributed mode of a high-performance analytical procedure refers to the procedure performing the analytics on an appliance that consists of a cluster of nodes. This appliance can be one of the following:

- a database management system (DBMS) appliance on which the SAS High-Performance Analytics infrastructure is also installed
- a cluster of nodes that have the SAS High-Performance Analytics infrastructure installed but no DBMS software installed

Distributed mode has several variations:

- Client-data (or local-data) mode: The input data for the analytic task are not stored on the appliance or cluster but are distributed to the distributed computing environment by the SAS High-Performance Analytics infrastructure when the procedure runs.
- Alongside-the-database mode: The data are stored in the distributed database and are read from the DBMS in parallel into a high-performance analytical procedure that runs on the database appliance.

- Alongside-HDFS mode: The data are stored in the Hadoop Distributed File System (HDFS) and are read in parallel from the HDFS. This mode is available if you install the SAS High-Performance Deployment of Hadoop on the appliance or when you configure a Cloudera 4 Hadoop deployment on the appliance to operate with the SAS High-Performance Analytics infrastructure. For more information about installing the SAS High-Performance Deployment of Hadoop, see the *SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide*.
- Alongside-LASR mode: The data are loaded from a SAS LASR Analytic Server that runs on the appliance.

Symmetric and Asymmetric Distributed Modes

SAS high-performance analytical procedures can run alongside the database or alongside HDFS in asymmetric mode. The primary reason for providing the asymmetric mode is to enable you to manage and house data on one appliance (the data appliance) and to run the high-performance analytical procedure on a second appliance (the computing appliance). You can also run in asymmetric mode on a single appliance that functions as both the data appliance and the computing appliance. This enables you to run alongside the database or alongside HDFS, where computations are done on a different set of nodes from the nodes that contain the data. The following subsections provide more details.

Symmetric Mode

When SAS high-performance analytical procedures run in symmetric distributed mode, the data appliance and the computing appliance must be the same appliance. Both the SAS Embedded Process and the high-performance analytical procedures execute in a SAS process that runs on the same hardware where the DBMS process executes. This is called symmetric mode because the number of nodes on which the DBMS executes is the same as the number of nodes on which the high-performance analytical procedures execute. The initial data movement from the DBMS to the high-performance analytical procedure does not cross node boundaries.

Asymmetric Mode

When SAS high-performance analytical procedures run in asymmetric distributed mode, the data appliance and computing appliance are usually distinct appliances. The high-performance analytical procedures execute in a SAS process that runs on the computing appliance. The DBMS and a SAS Embedded Process run on the data appliance. Data are requested by a SAS data feeder that runs on the computing appliance and communicates with the SAS Embedded Process on the data appliance. The SAS Embedded Process transfers the data in parallel to the SAS data feeder that runs on each of the nodes of the computing appliance. This is called asymmetric mode because the number of nodes on the data appliance does not need to be the same as the number of nodes on the computing appliance.

Controlling the Execution Mode with Environment Variables and Performance Statement Options

You control the execution mode by using environment variables or by specifying options in the PERFORMANCE statement in high-performance analytical procedures, or by a combination of these methods.

The important environment variables follow:

- *grid host* identifies the domain name system (DNS) or IP address of the appliance node to which the SAS High-Performance Text Miner software connects to run in distributed mode.
- *installation location* identifies the directory where the SAS High-Performance Text Miner software is installed on the appliance.
- *data server* identifies the database server on Teradata appliances as defined in the *hosts* file on the client. This data server is the same entry that you usually specify in the SERVER= entry of a LIBNAME statement for Teradata. For more information about specifying LIBNAME statements for Teradata and other engines, see the DBMS-specific section of *SAS/ACCESS for Relational Databases: Reference* for your engine.
- *grid mode* specifies whether the high-performance analytical procedures execute in symmetric or asymmetric mode. Valid values for this variable are `'sym'` for symmetric mode and `'asym'` for asymmetric mode. The default is symmetric mode.

You can set an environment variable directly from the SAS program by using the OPTION SET= command. For example, the following statements define three variables for a Teradata appliance (the grid mode is the default symmetric mode):

```
option set=GRIDHOST      ="hpa.sas.com";
option set=GRIDINSTALLLOC="/opt/TKGrid";
option set=GRIDDATASERVER="myserver";
```

Alternatively, you can set the parameters in the PERFORMANCE statement in high-performance analytical procedures. For example:

```
performance host       ="hpa.sas.com"
            install    ="/opt/TKGrid"
            dataserver="myserver";
```

The following statements define three variables that are needed to run asymmetrically on a computing appliance.

```
option set=GRIDHOST      ="compute_appliance.sas.com";
option set=GRIDINSTALLLOC="/opt/TKGrid";
option set=GRIDMODE      ="asym";
```

Alternatively, you can set the parameters in the PERFORMANCE statement in high-performance analytical procedures. For example:

```
performance host       ="compute_appliance.sas.com"
            install    ="/opt/TKGrid"
            gridmode   ="asym"
```

A specification in the PERFORMANCE statement overrides a specification of an environment variable without resetting its value. An environment variable that you set in the SAS session by using an OPTION SET= command remains in effect until it is modified or until the SAS session terminates.

Specifying a data server is necessary only on Teradata systems when you do not explicitly set the *gridmode* environment variable or specify the GRIDMODE= option in the PERFORMANCE statement. The data server specification depends on the entries in the (client) *hosts* file. The file specifies the server (suffixed by *cop* and a number) and an IP address. For example:

```
myservercop1  33.44.55.66
```

The key variable that determines whether a high-performance analytical procedure executes in single-machine or distributed mode is the *grid host*. The installation location and data server are needed to ensure that a connection to the grid host can be made, given that a host is specified. This book assumes that the installation location and data server (if necessary) have been set by your system administrator.

The following sets of SAS statements are functionally equivalent:

```
proc hpreduce;
   reduce unsupervised x:;
   performance host="hpa.sas.com";
run;

option set=GRIDHOST="hpa.sas.com";
proc hpreduce;
   reduce unsupervised x:;
run;
```

Determining Single-Machine Mode or Distributed Mode

High-performance analytical procedures use the following rules to determine whether they run in single-machine mode or distributed mode:

- If a grid host is not specified, the analysis is carried out in single-machine mode on the client machine that runs the SAS session.
- If a grid host is specified, the behavior depends on whether the execution is alongside the database or alongside HDFS. If the data are local to the client (that is, not stored in the distributed database or HDFS on the appliance), you need to use the NODES= option in the PERFORMANCE statement to specify the number of nodes on the appliance or cluster that you want to engage in the analysis. If the procedure executes alongside the database or alongside HDFS, you do not need to specify the NODES= option.

The following example shows single-machine and client-data distributed configurations for a data set of 100,000 observations that are simulated from a logistic regression model. The following DATA step generates the data:

```
data simData;
   array _a{8} _temporary_ (0,0,0,1,0,1,1,1);
   array _b{8} _temporary_ (0,0,1,0,1,0,1,1);
   array _c{8} _temporary_ (0,1,0,0,1,1,0,1);
```

```
      do obsno=1 to 100000;
         x  = rantbl(1,0.28,0.18,0.14,0.14,0.03,0.09,0.08,0.06);
         a  = _a{x};
         b  = _b{x};
         c  = _c{x};
         x1 = int(ranuni(1)*400);
         x2 = 52 + ranuni(1)*38;
         x3 = ranuni(1)*12;
         lp = 6. -0.015*(1-a) + 0.7*(1-b) + 0.6*(1-c) + 0.02*x1 -0.05*x2 - 0.1*x3;
         y  = ranbin(1,1,(1/(1+exp(lp))));
         output;
      end;
      drop x lp;
   run;
```

The following statements run PROC HPLOGISTIC to fit a logistic regression model:

```
   proc hplogistic data=simData;
      class a b c;
      model y = a b c x1 x2 x3;
   run;
```

Figure 2.1 shows the results from the analysis.

Figure 2.1 Results from Logistic Regression in Single-Machine Mode

```
                         The HPLOGISTIC Procedure

                         Performance Information

                   Execution Mode       Single-Machine
                   Number of Threads    4


                            Model Information

          Data Source                 WORK.SIMDATA
          Response Variable           y
          Class Parameterization      GLM
          Distribution                Binary
          Link Function               Logit
          Optimization Technique      Newton-Raphson with Ridging
```

Figure 2.1 *continued*

```
                          Parameter Estimates

                                      Standard
Parameter        Estimate          Error       DF     t Value     Pr > |t|

Intercept          5.7011         0.2539    Infty       22.45       <.0001
a 0              -0.01020        0.06627    Infty       -0.15       0.8777
a 1                     0              .        .           .            .
b 0                0.7124        0.06558    Infty       10.86       <.0001
b 1                     0              .        .           .            .
c 0                0.8036        0.06456    Infty       12.45       <.0001
c 1                     0              .        .           .            .
x1                0.01975       0.000614    Infty       32.15       <.0001
x2               -0.04728       0.003098    Infty      -15.26       <.0001
x3                -0.1017       0.009470    Infty      -10.74       <.0001
```

The entries in the "Performance Information" table show that the HPLOGISTIC procedure runs in single-machine mode and uses four threads, which are chosen according to the number of CPUs on the client machine. You can force a certain number of threads on any machine that is involved in the computations by specifying the NTHREADS option in the PERFORMANCE statement. Another indication of execution on the client is the following message, which is issued in the SAS log by all high-performance analytical procedures:

```
NOTE: The HPLOGISTIC procedure is executing on the client.
```

The following statements use 10 nodes (in distributed mode) to analyze the data on the appliance; results appear in Figure 2.2:

```
proc hplogistic data=simData;
   class a b c;
   model y = a b c x1 x2 x3;
   performance host="hpa.sas.com" nodes=10;
run;
```

Figure 2.2 Results from Logistic Regression in Distributed Mode

```
                      The HPLOGISTIC Procedure

                      Performance Information

             Host Node                        hpa.sas.com
             Execution Mode                   Distributed
             Grid Mode                        Symmetric
             Number of Compute Nodes          10
             Number of Threads per Node       24
```

Figure 2.2 *continued*

Model Information

Data Source	WORK.SIMDATA
Response Variable	y
Class Parameterization	GLM
Distribution	Binary
Link Function	Logit
Optimization Technique	Newton-Raphson with Ridging

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5.7011	0.2539	Infty	22.45	<.0001
a 0	-0.01020	0.06627	Infty	-0.15	0.8777
a 1	0	.	.	.	.
b 0	0.7124	0.06558	Infty	10.86	<.0001
b 1	0	.	.	.	.
c 0	0.8036	0.06456	Infty	12.45	<.0001
c 1	0	.	.	.	.
x1	0.01975	0.000614	Infty	32.15	<.0001
x2	-0.04728	0.003098	Infty	-15.26	<.0001
x3	-0.1017	0.009470	Infty	-10.74	<.0001

The specification of a host causes the “Performance Information” table to display the name of the host node of the appliance. The “Performance Information” table also indicates that the calculations were performed in a distributed environment on the appliance. Twenty-four threads on each of 10 nodes were used to perform the calculations—for a total of 240 threads.

Another indication of distributed execution on the appliance is the following message, which is issued in the SAS log by all high-performance analytical procedures:

```
NOTE: The HPLOGISTIC procedure is executing in the distributed
      computing environment with 10 worker nodes.
```

You can override the presence of a grid host and force the computations into single-machine mode by specifying the NODES=0 option in the PERFORMANCE statement:

```
proc hplogistic data=simData;
   class a b c;
   model y = a b c x1 x2 x3;
   performance host="hpa.sas.com" nodes=0;
run;
```

Figure 2.3 shows the “Performance Information” table. The numeric results are not reproduced here, but they agree with the previous analyses, which are shown in Figure 2.1 and Figure 2.2.

Figure 2.3 Single-Machine Mode Despite Host Specification

```
                 The HPLOGISTIC Procedure

                 Performance Information

           Execution Mode       Single-Machine
           Number of Threads    4
```

The "Performance Information" table indicates that the HPLOGISTIC procedure executes in single-machine mode on the client. This information is also reported in the following message, which is issued in the SAS log:

```
NOTE: The HPLOGISTIC procedure is executing on the client.
```

In the analysis shown previously in Figure 2.2, the data set Work.simData is local to the client, and the HPLOGISTIC procedure distributed the data to 10 nodes on the appliance. The High-Performance Analytics infrastructure does not keep these data on the appliance. When the procedure terminates, the in-memory representation of the input data on the appliance is freed.

When the input data set is large, the time that is spent sending client-side data to the appliance might dominate the execution time. In practice, transfer speeds are usually lower than the theoretical limits of the network connection or disk I/O rates. At a transfer rate of 40 megabytes per second, sending a 10-gigabyte data set to the appliance requires more than four minutes. If analytic execution time is in the range of seconds, the "performance" of the process is dominated by data movement.

The alongside-the-database execution model, unique to high-performance analytical procedures, enables you to read and write data in distributed form from the database that is installed on the appliance.

Alongside-the-Database Execution

High-performance analytical procedures interface with the distributed database management system (DBMS) on the appliance in a unique way. If the input data are stored in the DBMS and the grid host is the appliance that houses the data, high-performance analytical procedures create a distributed computing environment in which an analytic process is co-located with the nodes of the DBMS. Data then pass from the DBMS to the analytic process on each node. Instead of moving across the network and possibly back to the client machine, the data pass locally between the processes on each node of the appliance.

Because the analytic processes on the appliance are separate from the database processes, the technique is referred to as alongside-the-database execution in contrast to in-database execution, where the analytic code executes in the database process.

In general, when you have a large amount of input data, you can achieve the best performance from high-performance analytical procedures if execution is alongside the database.

Before you can run alongside the database, you must distribute the data to the appliance. The following statements use the HPDS2 procedure to distribute the data set Work.simData into the mydb database on the hpa.sas.com appliance. In this example, the appliance houses a Greenplum database.

```
option set=GRIDHOST="hpa.sas.com";
libname applianc greenplm
        server  ="hpa.sas.com"
        user    =XXXXXX
        password=YYYYY
        database=mydb;


proc datasets lib=applianc nolist; delete simData;
proc hpds2 data=simData
           out =applianc.simData(distributed_by='distributed randomly');
  performance commit=10000 nodes=all;
  data DS2GTF.out;
     method run();
        set DS2GTF.in;
     end;
  enddata;
run;
```

If the output table applianc.simData exists, the DATASETS procedure removes the table from the Greenplum database because a DBMS does not usually support replacement operations on tables.

Note that the libref for the output table points to the appliance. The data set option informs the HPDS2 procedure to distribute the records randomly among the data segments of the appliance. The statements that follow the PERFORMANCE statement are the DS2 program that copies the input data to the output data without further transformations.

Because you loaded the data into a database on the appliance, you can use the following HPLOGISTIC statements to perform the analysis on the appliance in the alongside-the-database mode. These statements are almost identical to the first PROC HPLOGISTIC example in the previous section, which executed in single-machine mode.

```
proc hplogistic data=applianc.simData;
   class a b c;
   model y = a b c x1 x2 x3;
run;
```

The subtle differences are as follows:

- The grid host environment variable that you specified in an OPTION SET= command is still in effect.
- The DATA= option in the high-performance analytical procedure uses a libref that identifies the data source as being housed on the appliance. This libref was specified in a prior LIBNAME statement.

Figure 2.4 shows the results from this analysis. The "Performance Information" table shows that the execution was in distributed mode. In this case the execution was alongside the Greenplum database. The numeric results agree with the previous analyses, which are shown in Figure 2.1 and Figure 2.2.

Figure 2.4 Alongside-the-Database Execution on Greenplum

```
                    The HPLOGISTIC Procedure

                    Performance Information

         Host Node                       hpa.sas.com
         Execution Mode                  Distributed
         Grid Mode                       Symmetric
         Number of Compute Nodes         8
         Number of Threads per Node      24


                       Model Information

   Data Source                 SIMDATA
   Response Variable           y
   Class Parameterization      GLM
   Distribution                Binary
   Link Function               Logit
   Optimization Technique      Newton-Raphson with Ridging
```

Figure 2.4 *continued*

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5.7011	0.2539	Infty	22.45	<.0001
a 0	-0.01020	0.06627	Infty	-0.15	0.8777
a 1	0	.	.	.	.
b 0	0.7124	0.06558	Infty	10.86	<.0001
b 1	0	.	.	.	.
c 0	0.8036	0.06456	Infty	12.45	<.0001
c 1	0	.	.	.	.
x1	0.01975	0.000614	Infty	32.15	<.0001
x2	-0.04728	0.003098	Infty	-15.26	<.0001
x3	-0.1017	0.009470	Infty	-10.74	<.0001

When high-performance analytical procedures execute symmetrically alongside the database, any nonzero specification of the NODES= option in the PERFORMANCE statement is ignored. If the data are read alongside the database, the number of compute nodes is determined by the layout of the database and cannot be modified. In this example, the appliance contains 16 nodes. (See the "Performance Information" table.)

However, when high-performance analytical procedures execute asymmetrically alongside the database, the number of compute nodes that you specify in the PERFORMANCE statement can differ from the number of nodes across which the data are partitioned. For an example, see the section "Running High-Performance Analytical Procedures in Asymmetric Mode" on page 19.

Alongside-LASR Distributed Execution

You can execute high-performance analytical procedures in distributed mode alongside a SAS LASR Analytic Server. When high-performance analytical procedures execute in this mode, the data are preloaded in distributed form in memory that is managed by a LASR Analytic Server. The data on the nodes of the appliance are accessed in parallel in the process that runs the LASR Analytic Server, and they are transferred to the process where the high-performance analytical procedure runs. In general, each high-performance analytical procedure copies the data to memory that persists only while that procedure executes. Hence, when a high-performance analytical procedure runs alongside a LASR Analytic Server, both the high-performance analytical procedure and the LASR Analytic Server have a copy of the subset of the data that is used by the high-performance analytical procedure. The advantage of running high-performance analytical procedures alongside a LASR Analytic Server (as opposed to running alongside a DBMS table or alongside HDFS) is that the initial transfer of data from the LASR Analytic Server to the high-performance analytical procedure is a memory-to-memory operation that is faster than the disk-to-memory operation when the procedure runs alongside a DBMS or HDFS. When the cost of preloading a table into a LASR Analytic Server is amortized by multiple uses of these data in separate runs of high-performance analytical procedures, using the LASR Analytic Server can result in improved performance.

Running High-Performance Analytical Procedures Alongside a SAS LASR Analytic Server in Distributed Mode

This section provides an example of steps that you can use to start and load data into a SAS LASR Analytic Server instance and then run high-performance analytical procedures alongside this LASR Analytic Server instance.

Starting a SAS LASR Analytic Server Instance

The following statements create a SAS LASR Analytic Server instance and load it with the simData data set that is used in the preceding examples. The data that are loaded into the LASR Analytic Server persist in memory across procedure boundaries until these data are explicitly deleted or until the server instance is terminated.

```
proc lasr port=12345
          data=simData
          path="/tmp/";
   performance host="hpa.sas.com" nodes=ALL;
run;
```

The PORT= option specifies a network port number to use. The PATH= option specifies the directory in which the server and table signature files are to be stored. The specified directory must exist on each machine in the cluster. The DATA= option specifies the name of a data set that is loaded into this LASR Analytic Server instance. (You do not need to specify the DATA= option at this time because you can add tables to the LASR Analytic Server instance at any stage of its life.) For more information about starting and using a LASR Analytic Server, see the *SAS LASR Analytic Server: Administration Guide*.

The NODES=ALL option in the PERFORMANCE statement specifies that the LASR Analytic Server run on all the nodes on the appliance. You can start a LASR Analytic Server on a subset of the nodes on an appliance, but this might affect whether high-performance analytical procedures can run alongside the LASR Analytic Server. For more information, see the section “Alongside-LASR Distributed Execution on a Subset of the Appliance Nodes” on page 19.

Figure 2.5 shows the “Performance Information” table, which shows that the LASR procedure executes in distributed mode on 16 nodes.

Figure 2.5 Performance Information

```
                    The LASR Procedure

                 Performance Information

      Host Node                    hpa.sas.com
      Execution Mode               Distributed
      Grid Mode                    Symmetric
      Number of Compute Nodes      8
```

Associating a SAS Libref with the SAS LASR Analytic Server Instance

The following statements use a LIBNAME statement that associates a SAS libref (named MyLasr) with tables on the server instance as follows:

```
libname MyLasr sasiola port=12345;
```

The SASIOLA option requests that the MyLasr libref use the SASIOLA engine, and the PORT= value associates this libref with the appropriate server instance. For more information about creating a libref that uses the SASIOLA engine, see the *SAS LASR Analytic Server: Administration Guide*.

Running a High-Performance Analytical Procedure Alongside the SAS LASR Analytic Server Instance

You can use the MyLasr libref to specify the input data for high-performance analytical procedures. You can also create output data sets in the SAS LASR Analytic Server instance by using this libref to request that the output data set be held in memory by the server instance as follows:

```
proc hplogistic data=MyLasr.simData;
   class a b c;
   model y = a b c x1 x2 x3;
   output out=MyLasr.simulateScores pred=PredictedProbabliity;
run;
```

Because you previously specified the GRIDHOST= environment variable and the input data are held in distributed form in the associated server instance, this PROC HPLOGISTIC step runs in distributed mode alongside the LASR Analytic Server, as indicated in the “Performance Information” table shown in Figure 2.6.

Figure 2.6 Performance Information

Performance Information

Host Node	hpa.sas.com
Execution Mode	Distributed
Grid Mode	Symmetric
Number of Compute Nodes	8
Number of Threads per Node	24

The preceding OUTPUT statement creates an output table that is added to the LASR Analytic Server instance. Output data sets do not have to be created in the same server instance that holds the input data. You can use a different LASR Analytic Server instance to hold the output data set. However, in order for the output data to be created in alongside mode, all the nodes that are used by the server instance that holds the input data must also be used by the server instance that holds the output data.

Terminating a SAS LASR Analytic Server Instance

You can continue to run high-performance analytical procedures and add and delete tables from the SAS LASR Analytic Server instance until you terminate the server instance as follows:

```
proc lasr term port=12345;
run;
```

Alongside-LASR Distributed Execution on a Subset of the Appliance Nodes

When you run PROC LASR to start a SAS LASR Analytic Server, you can specify the NODES= option in a PERFORMANCE statement to control how many nodes the LASR Analytic Server executes on. Similarly, a high-performance analytical procedure can execute on a subset of the nodes either because you specify the NODES= option in a PERFORMANCE statement or because you run alongside a DBMS or HDFS with an input data set that is distributed on a subset of the nodes on an appliance. In such situations, if a high-performance analytical procedure uses nodes on which the LASR Analytic Server is not running, then running alongside LASR is not supported. You can avoid this issue by specifying the NODES=ALL in the PERFORMANCE statement when you use PROC LASR to start the LASR Analytic Server.

Running High-Performance Analytical Procedures in Asymmetric Mode

This section provides examples of how you can run high-performance analytical procedures in asymmetric mode. It also includes examples that run in symmetric mode to highlight differences between the modes. For a description of asymmetric mode, see the section “Symmetric and Asymmetric Distributed Modes” on page 7.

Asymmetric mode is commonly used when the data appliance and the computing appliance are distinct appliances. In order to be able to use an appliance as a data provider for high-performance analytical procedures that run in asymmetric mode on another appliance, it is not necessary that SAS High-Performance Text Miner be installed on the data appliance. However, it is essential that a SAS Embedded Process be installed on the data appliance and that SAS High-Performance Text Miner be installed on the computing appliance.

The following examples use a 24-node data appliance named “data_appliance.sas.com,” which houses a Teradata DBMS and has a SAS Embedded Process installed. Because SAS High-Performance Text Miner is also installed on this appliance, it can be used to run high-performance analytical procedures in both symmetric and asymmetric modes.

The following statements load the simData data set of the preceding sections onto the data appliance:

```
libname dataLib teradata
        server  ="tera2650"
        user    =XXXXXX
        password=YYYYY
        database=mydb;

data dataLib.simData;
   set simData;
run;
```

NOTE: You can provision the appliance with data even if SAS High-Performance Text Miner software is not installed on the appliance.

The following subsections show how you can run the HPLOGISTIC procedure symmetrically and asymmetrically on a single data appliance and asymmetrically on distinct data and computing appliances.

Running in Symmetric Mode

The following statements run the HPLOGISTIC procedure in symmetric mode on the data appliance:

```
proc hplogistic data=dataLib.simData;
   class a b c;
   model y = a b c x1 x2 x3;
   performance host       = "data_appliance.sas.com"
               nodes      = 10
               gridmode   = sym;
run;
```

Because you explicitly specified the GRIDMODE= option, you do not need to also specify the DATASERVER= option in the PERFORMANCE statement. Figure 2.7 shows the results of this analysis.

Figure 2.7 Alongside-the-Database Execution in Symmetric Mode on Teradata

The HPLOGISTIC Procedure

Performance Information

Host Node	data_appliance.sas.com
Execution Mode	Distributed
Grid Mode	Symmetric
Number of Compute Nodes	24
Number of Threads per Node	24

Figure 2.7 *continued*

Model Information

Data Source	simData
Response Variable	y
Class Parameterization	GLM
Distribution	Binary
Link Function	Logit
Optimization Technique	Newton-Raphson with Ridging

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5.7011	0.2539	Infty	22.45	<.0001
a 0	-0.01020	0.06627	Infty	-0.15	0.8777
a 1	0	.	.	.	.
b 0	0.7124	0.06558	Infty	10.86	<.0001
b 1	0	.	.	.	.
c 0	0.8036	0.06456	Infty	12.45	<.0001
c 1	0	.	.	.	.
x1	0.01975	0.000614	Infty	32.15	<.0001
x2	-0.04728	0.003098	Infty	-15.26	<.0001
x3	-0.1017	0.009470	Infty	-10.74	<.0001

The “Performance Information” table shows that the execution occurs in symmetric mode on the 24 nodes of the data appliance. In this case, the NODES=10 option in the PERFORMANCE statement is ignored because the number of nodes that are used is determined by the number of nodes across which the data are distributed, as indicated in the following warning message in the SAS log:

```
WARNING: The NODES=10 option in the PERFORMANCE statement is ignored because
         you are running alongside the distributed data source
         DATALIB.simData.DATA. The number of compute nodes is determined by the
         configuration of the distributed DBMS.
```

Running in Asymmetric Mode on One Appliance

You can switch to running the HPLOGISTIC procedure in asymmetric mode by specifying the GRIDMODE=ASYM option in the PERFORMANCE statement as follows:

```
proc hplogistic data=dataLib.simData;
   class a b c;
   model y = a b c x1 x2 x3;
   performance host       = "data_appliance.sas.com"
               nodes      = 10
               gridmode   = asym;
run;
```

Figure 2.8 shows the “Performance Information” table.

Figure 2.8 Alongside Teradata Execution in Asymmetric Mode

The HPLOGISTIC Procedure

Performance Information

Host Node	data_appliance.sas.com
Execution Mode	Distributed
Grid Mode	Asymmetric
Number of Compute Nodes	10
Number of Threads per Node	24

You can see that now the grid mode is asymmetric. Furthermore, the NODES=10 option that you specified in the PERFORMANCE statement is honored. The data are moved in parallel from the 24 nodes on which the data are stored to the 10 nodes on which the execution occurs. The numeric results are not reproduced here, but they agree with the previous analyses.

Running in Asymmetric Mode on Distinct Appliances

Usually, there is no advantage to executing high-performance analytical procedures in asymmetric mode on one appliance, because data might have to be unnecessarily moved between nodes. The following example demonstrates the more typical use of asymmetric mode. In this example, the specified grid host “compute_appliance.sas.com” is a computing appliance that has 15 compute nodes, and it is a different appliance from the 24-node data appliance “data_appliance.sas.com,” which houses the Teradata DBMS where the data reside.

The advantage of using different computing and data appliances is that the data appliance is not affected by the execution of high-performance analytical procedures except during the initial parallel data transfer. A potential disadvantage of this asymmetric mode of execution is that the performance can be limited by the bandwidth with which data can be moved between the appliances. However, because this data movement takes place in parallel from the nodes of the data appliance to the nodes of the computing appliance, this potential performance bottleneck can be overcome with appropriately provisioned hardware. The following statements show how this is done:

```
proc hplogistic data=dataLib.simData;
   class a b c;
   model y = a b c x1 x2 x3;
   performance host        = "compute_appliance.sas.com"
               gridmode    = asym;
run;
```

Figure 2.9 shows the “Performance Information” table.

Figure 2.9 Asymmetric Mode with Distinct Data and Computing Appliances

```
                    The HPLOGISTIC Procedure

                    Performance Information

        Host Node                       compute_appliance.sas.com
        Execution Mode                  Distributed
        Grid Mode                       Asymmetric
        Number of Compute Nodes         15
        Number of Threads per Node      24
```

PROC HPLOGISTIC ran on the 15 nodes of the computing appliance, even though the data are partitioned across the 24 nodes of the data appliance. The numeric results are not reproduced here, but they agree with the previous analyses shown in Figure 2.1 and Figure 2.2.

Every time you run a high-performance analytical procedure in asymmetric mode that uses different computing and data appliances, data are transferred between these appliances. If you plan to make repeated use of the same data, then it might be advantageous to temporarily persist the data that you need on the computing appliance. One way to persist the data is to store them as a table in a SAS LASR Analytic Server that runs on the computing appliance. By running PROC LASR in asymmetric mode, you can load the data in parallel from the data appliance nodes to the nodes on which the LASR Analytic Server runs on the computing appliance. You can then use a LIBNAME statement that associates a SAS libref with tables on the LASR Analytic Server. The following statements show how you do this:

```
proc lasr port=54321
          data=dataLib.simData
          path="/tmp/";
   performance host     ="compute_appliance.sas.com"
               gridmode = asym;
run;

libname MyLasr sasiola tag="dataLib" port=54321 host="compute_appliance.sas.com" ;
```

Figure 2.10 show the “Performance Information” table.

Figure 2.10 PROC LASR Running in Asymmetric Mode

```
                       The LASR Procedure

                    Performance Information

          Host Node                    compute_appliance.sas.com
          Execution Mode               Distributed
          Grid Mode                    Asymmetric
          Number of Compute Nodes      15
```

PROC LASR ran in asymmetric mode on the computing appliance, which has 15 compute nodes. In this mode, the data are loaded in parallel from the 24 data appliance nodes to the 15 compute nodes on the computing appliance. By default, all the nodes on the computing appliance are used. You can use the NODES= option in the PERFORMANCE statement to run the LASR Analytic Server on a subset of the nodes on the computing appliance. If you omit the GRIDMODE=ASYM option from the PERFORMANCE statement, PROC LASR still runs successfully but much less efficiently. The Teradata access engine transfers the simData data set to a temporary table on the client, and the High-Performance Analytics infrastructure then transfers these data from the temporary table on the client to the grid nodes on the computing appliance.

After the data are loaded into a LASR Analytic Server that runs on the computing appliance, you can run high-performance analytical procedures alongside this LASR Analytic Server. Because these procedures run on the same computing appliance where the LASR Analytic Server is running, it is best to run these procedures in symmetric mode, which is the default or can be explicitly specified in the GRIDMODE=SYM option in the PERFORMANCE statement. The following statements provide an example. The OUTPUT statement creates an output data set that is held in memory by the LASR Analytic Server. The data appliance has no role in executing these statements.

```
proc hplogistic data=MyLasr.simData;
   class a b c;
   model y = a b c x1 x2 x3;
   output out=MyLasr.myOutputData pred=myPred;
   performance host = "compute_appliance.sas.com";
run;
```

The following note, which appears in the SAS log, confirms that the output data set is created successfully:

```
NOTE: The table DATALIB.MYOUTPUTDATA has been added to the LASR Analytic Server
      with port 54321. The Libname is MYLASR.
```

You can use the dataLib libref that you used to load the data onto the data appliance to create an output data set on the data appliance. In order for this output to be directly written in parallel from the nodes of the computing appliance to the nodes of the data appliance, you need to run the HPLOGISTIC procedure in asymmetric mode by specifying the GRIDMODE=ASYM option in the PERFORMANCE statement as follows:

```
proc hplogistic data=MyLasr.simData;
   class a b c;
   model y = a b c x1 x2 x3;
   output out=dataLib.myOutputData pred=myPred;
   performance host       = "compute_appliance.sas.com"
               gridmode   = asym;
run;
```

The following note, which appears in the SAS log, confirms that the output data set is created successfully on the data appliance:

```
NOTE: The data set DATALIB.myOutputData has 100000 observations and 1 variables.
```

When you run a high-performance analytical procedure on a computing appliance and either read data from or write data to a different data appliance, it is important to run the high-performance analytical procedures in asymmetric mode so that the Read and Write operations take place in parallel without any movement of data to and from the SAS client. If you omit running the preceding PROC HPLOGISTIC step in asymmetric mode, then the output data set would be created much less efficiently: the output data would be moved sequentially to a temporary table on the client, after which the Teradata access engine sequentially would write this table to the data appliance.

When you no longer need the data in the SAS LASR Analytic Server, you should terminate the server instance as follows:

```
proc lasr term port=54321;
    performance host="compute_appliance.sas.com";
run;
```

If you configured Hadoop on the computing appliance, then you can create output data tables that are stored in the HDFS on the computing appliance. You can do this by using the SASHDAT engine as described in the section "Alongside-HDFS Execution" on page 25.

Alongside-HDFS Execution

Running high-performance analytical procedures alongside HDFS shares many features with running alongside the database. You can execute high-performance analytical procedures alongside HDFS by using either the SASHDAT engine or the Hadoop engine.

You use the SASHDAT engine to read and write data that are stored in HDFS in a proprietary SASHDAT format. In SASHDAT format, metadata that describe the data in the Hadoop files are included with the data. This enables you to access files in SASHDAT format without supplying any additional metadata. Additionally, you can also use the SASHDAT engine to read data in CSV (comma-separated value) format, but you need supply metadata that describe the contents of the CSV data. The SASHDAT engine provides highly optimized access to data in HDFS that are stored in SASHDAT format.

The Hadoop engine reads data that are stored in various formats from HDFS and writes data to HDFS in CSV format. This engine can use metadata that are stored in Hive, which is a data warehouse that supplies metadata about data that are stored in Hadoop files. In addition, this engine can use metadata that you create by using the HDMD procedure.

The following subsections provide details about using the SASHDAT and Hadoop engines to execute high-performance analytical procedures alongside HDFS.

Alongside-HDFS Execution by Using the SASHDAT Engine

If the grid host is a cluster that houses data that have been distributed by using the SASHDAT engine, then high-performance analytical procedures can analyze those data in the alongside-HDFS mode. The procedures use the distributed computing environment in which an analytic process is co-located with the nodes of the cluster. Data then pass from HDFS to the analytic process on each node of the cluster.

Before you can run a procedure alongside HDFS, you must distribute the data to the cluster. The following statements use the SASHDAT engine to distribute to HDFS the simData data set that was used in the previous two sections:

```
option set=GRIDHOST="hpa.sas.com";

libname hdatLib sashdat
        path="/hps";
```

```
data hdatLib.simData (replace = yes) ;
    set simData;
run;
```

In this example, the GRIDHOST is a cluster where the SAS Data in HDFS Engine is installed. If a data set that is named simData already exists in the hps directory in HDFS, it is overwritten because the REPLACE=YES data set option is specified. For more information about using this LIBNAME statement, see the section "LIBNAME Statement for the SAS Data in HDFS Engine" in the *SAS LASR Analytic Server: Administration Guide*.

The following HPLOGISTIC procedure statements perform the analysis in alongside-HDFS mode. These statements are almost identical to the PROC HPLOGISTIC example in the previous two sections, which executed in single-machine mode and alongside-the-database distributed mode, respectively.

```
proc hplogistic data=hdatLib.simData;
   class a b c;
   model y = a b c x1 x2 x3;

run;
```

Figure 2.11 shows the "Performance Information" table. You see that the procedure ran in distributed mode. The numeric results shown in Figure 2.12 agree with the previous analyses shown in Figure 2.1, Figure 2.2, and Figure 2.4.

Figure 2.11 Alongside-HDFS Execution Performance Information

Performance Information

Host Node	hpa.sas.com
Execution Mode	Distributed
Grid Mode	Symmetric
Number of Compute Nodes	206
Number of Threads per Node	8

Figure 2.12 Alongside-HDFS Execution Model Information

Model Information

Data Source	HDATLIB.SIMDATA
Response Variable	y
Class Parameterization	GLM
Distribution	Binary
Link Function	Logit
Optimization Technique	Newton-Raphson with Ridging

Figure 2.12 *continued*

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5.7011	0.2539	Infty	22.45	<.0001
a 0	-0.01020	0.06627	Infty	-0.15	0.8777
a 1	0	.	.	.	.
b 0	0.7124	0.06558	Infty	10.86	<.0001
b 1	0	.	.	.	.
c 0	0.8036	0.06456	Infty	12.45	<.0001
c 1	0	.	.	.	.
x1	0.01975	0.000614	Infty	32.15	<.0001
x2	-0.04728	0.003098	Infty	-15.26	<.0001
x3	-0.1017	0.009470	Infty	-10.74	<.0001

Alongside-HDFS Execution by Using the Hadoop Engine

The following LIBNAME statement sets up a libref that you can use to access data that are stored in HDFS and have metadata in Hive:

```
libname hdoopLib hadoop
        server       = "hpa.sas.com"
        user         = XXXXX
        password     = YYYYY
        database     = myDB
        config       = "demo.xml" ;
```

For more information about LIBNAME options available for the Hadoop engine, see the LIBNAME topic in the Hadoop section of *SAS/ACCESS for Relational Databases: Reference*. The configuration file that you specify in the CONFIG= option contains information that is needed to access the Hive server. It also contains information that enables this configuration file to be used to access data in HDFS without using the Hive server. This information can also be used to specify replication factors and block sizes that are used when the engine writes data to HDFS. The following XML shows the contents of the file demo.xml that is used in this example:

```
<configuration>
<property>
  <name>fs.default.name</name>
  <value>hdfs://hpa.sas.com:8020</value>
</property>
<property>
  <name>mapred.job.tracker</name>
  <value>hpa.sas.com:8021</value>
</property>
<property>
    <name>dfs.replication</name>
```

```
    <value>1</value>
</property>
<property>
    <name>dfs.block.size</name>
    <value>33554432</value>
</property>
</configuration>
```

The following DATA step uses the Hadoop engine to distribute to HDFS the simData data set that was used in the previous sections. The engine creates metadata for the data set in Hive.

```
data hdoopLib.simData;
    set simData;
run;
```

After you have loaded data or if you are accessing preexisting data in HDFS that have metadata in Hive, you can access this data alongside HDFS by using high-performance analytics procedures. The following HPLOGISTIC procedure statements perform the analysis in alongside-HDFS mode. These statements are similar to the PROC HPLOGISTIC example in the previous sections. However, whenever you use the Hadoop engine, you must execute the analysis in asymmetric mode to cause the execution to occur alongside HDFS.

```
proc hplogistic data=hdoopLib.simData;
   class a b c;
   model y = a b c x1 x2 x3;
   performance host     = "compute_appliance.sas.com"
               gridmode = asym;
run;
```

Figure 2.13 shows the "Performance Information" table. You see that the procedure ran asymmetrically in distributed mode. The numeric results shown in Figure 2.14 agree with the previous analyses.

Figure 2.13 Alongside-HDFS Execution by Using the Hadoop Engine

The HPLOGISTIC Procedure

Performance Information

Host Node	compute_appliance.sas.com
Execution Mode	Distributed
Grid Mode	Asymmetric
Number of Compute Nodes	15
Number of Threads per Node	24

Figure 2.14 Alongside-HDFS Execution by Using the Hadoop Engine

Model Information

Data Source	HDOOPLIB.SIMDATA
Response Variable	y
Class Parameterization	GLM
Distribution	Binary
Link Function	Logit
Optimization Technique	Newton-Raphson with Ridging

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5.7011	0.2539	Infty	22.45	<.0001
a 0	-0.01020	0.06627	Infty	-0.15	0.8777
a 1	0	.	.	.	.
b 0	0.7124	0.06558	Infty	10.86	<.0001
b 1	0	.	.	.	.
c 0	0.8036	0.06456	Infty	12.45	<.0001
c 1	0	.	.	.	.
x1	0.01975	0.000614	Infty	32.15	<.0001
x2	-0.04728	0.003098	Infty	-15.26	<.0001
x3	-0.1017	0.009470	Infty	-10.74	<.0001

The Hadoop engine also enables you to access tables in HDFS that are stored in various formats and that are not registered in Hive. You can use the HDMD procedure to generate metadata for tables that are stored in the following file formats:

- delimited text
- fixed-record length binary
- JavaScript Object Notation (JSON)
- sequence files
- XML text

To read any other kind of file in Hadoop, you can write a custom file reader plug-in in Java for use with PROC HDMD. For more information about LIBNAME options available for the Hadoop engine, see the LIBNAME topic in the Hadoop section of *SAS/ACCESS for Relational Databases: Reference*.

The following example shows how you can use PROC HDMD to register metadata for CSV data independently from Hive and then analyze these data by using high-performance analytics procedures. The CSV data in the table csvExample.csv is stored in HDFS in the directory /user/demo/data. Each record in this table consists of the following fields, in the order shown and separated by commas.

1. a string of at most six characters
2. a numeric field with values of 0 or 1
3. a numeric field with real numbers

Suppose you want to fit a logistic regression model to these data, where the second field represents a target variable named Success, the third field represents a regressor named Dose, and the first field represents a classification variable named Group.

The first step is to use PROC HDMD to create metadata that are needed to interpret the table, as in the following statements:

```
libname hdoopLib hadoop
                 server       = "hpa.sas.com"
                 user         = XXXXX
                 password     = YYYYY
                 HDFS_PERMDIR = "/user/demo/data"
                 HDFS_METADIR = "/user/demo/meta"
                 config       = "demo.xml"
                 DBCREATE_TABLE_EXTERNAL=YES;

proc hdmd name=hdoopLib.csvExample data_file='csvExample.csv'
          format=delimited encoding=utf8 sep = ',';

     column Group    char(6);
     column Success  double;
     column Dose     double;
run;
```

The metadata that are created by PROC HDMD for this table are stored in the directory /user/demo/meta that you specified in the HDFS_METADIR = option in the preceding LIBNAME statement. After you create the metadata, you can execute high-performance analytics procedures with these data by using the hdoopLib libref. For example, the following statements fit a logistic regression model to the CSV data that are stored in csvExample.csv table.

```
proc hplogistic data=hdoopLib.csvExample;
    class Group;
    model Success = Dose;
    performance host     = "compute_appliance.sas.com"
                gridmode = asym;
run;
```

Figure 2.15 shows the results of this analysis. You see that the procedure ran asymmetrically in distributed mode. The metadata that you created by using the HDMD procedure have been used successfully in executing this analysis.

Figure 2.15 Alongside-HDFS Execution with CSV Data

```
                    The HPLOGISTIC Procedure

                   Performance Information

       Host Node                     compute_appliance.sas.com
       Execution Mode                Distributed
       Grid Mode                     Asymmetric
       Number of Compute Nodes       15
       Number of Threads per Node    24


                      Model Information

        Data Source                GRIDLIB.CSVEXAMPLE
        Response Variable          Success
        Class Parameterization     GLM
        Distribution               Binary
        Link Function              Logit
        Optimization Technique     Newton-Raphson with Ridging


                   Class Level Information

               Class     Levels    Values

               Group          3    group1 group2 group3


            Number of Observations Read           1000
            Number of Observations Used           1000


                     Parameter Estimates

                                 Standard
    Parameter     Estimate          Error        DF     t Value     Pr > |t|

    Intercept       0.1243         0.1295     Infty        0.96       0.3371
    Dose           -0.2674         0.2216     Infty       -1.21       0.2277
```

Output Data Sets

In the alongside-the-database mode, the data are read in distributed form, minimizing data movement for best performance. Similarly, when you write output data sets and a high-performance analytical procedure executes in distributed mode, the data can be written in parallel into the database.

For example, in the following statements, the HPLOGISTIC procedure executes in distributed mode by using eight nodes on the appliance to perform the logistic regression on work.simData:

```
proc hplogistic data=simData;
   class a b c;
   model y = a b c x1 x2 x3;
   id a;
   output out=applianc.simData_out pred=p;
   performance host="hpa.sas.com" nodes=8;
run;
```

The output data set applianc.simData_out is written in parallel into the database. Although the data are fed on eight nodes, the database might distribute the data on more nodes.

When a high-performance analytical procedure executes in single-machine mode, all output objects are created on the client. If the libref of the output data sets points to the appliance, the data are transferred to the database on the appliance. This can lead to considerable performance degradation compared to execution in distributed mode.

Many procedures in SAS software add the variables from the input data set when an observationwise output data set is created. The assumption of high-performance analytical procedures is that the input data sets can be large and contain many variables. For performance reasons, the output data set contains the following:

- variables that are explicitly created by the statement
- variables that are listed in the ID statement
- distribution keys or hash keys that are transferred from the input data set

Including this information enables you to add to the output data set information necessary for subsequent SQL joins without copying the entire input data set to the output data set.

Working with Formats

You can use SAS formats and user-defined formats with high-performance analytical procedures as you can with other procedures in the SAS System. However, because the analytic work is carried out in a distributed environment and might depend on the formatted values of variables, some special handling can improve the efficiency of work with formats.

High-performance analytical procedures examine the variables that are used in an analysis for association with user-defined formats. Any user-defined formats that are found by a procedure are transmitted automatically to the appliance. If you are running multiple high-performance analytical procedures in a SAS session and the analysis variables depend on user-defined formats, you can preprocess the formats. This step involves generating an XML stream (a file) of the formats and passing the stream to the high-performance analytical procedures.

Suppose that the following formats are defined in your SAS program:

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

The next group of SAS statements create the XML stream for the formats in the file *Myfmt.xml*, associate that file with the file reference myxml, and pass the file reference with the FMTLIBXML= option in the PROC HPLOGISTIC statement:

```
filename myxml 'Myfmt.xml';
libname  myxml XML92 xmltype=sasfmt tagset=tagsets.XMLsuv;
proc format cntlout=myxml.allfmts;
run;


proc hplogistic data=six fmtlibxml=myxml;
   class wheeze cit age;
   format wheeze best4. cit $cityChar.;
   model wheeze = cit age;
run;
```

Generation and destruction of the stream can be wrapped in convenience macros:

```
%macro Make_XMLStream(name=tempxml);
    filename &name 'fmt.xml';
    libname  &name XML92 xmltype=sasfmt tagset=tagsets.XMLsuv;
    proc format cntlout=&name..allfmts;
    run;
%mend;

%macro Delete_XMLStream(fref);
    %let rc=%sysfunc(fdelete(&fref));
%mend;
```

If you do not pass an XML stream to a high-performance analytical procedure that supports the FMTLIBXML= option, the procedure generates an XML stream as needed when it is invoked.

PERFORMANCE Statement

PERFORMANCE < *performance-options* > **;**

The PERFORMANCE statement defines performance parameters for multithreaded and distributed computing, passes variables that describe the distributed computing environment, and requests detailed results about the performance characteristics of a high-performance analytical procedure.

You can also use the PERFORMANCE statement to control whether a high-performance analytical procedure executes in single-machine or distributed mode.

You can specify the following *performance-options* in the PERFORMANCE statement:

COMMIT=*n*

requests that the high-performance analytical procedure write periodic updates to the SAS log when observations are sent from the client to the appliance for distributed processing.

High-performance analytical procedures do not have to use input data that are stored on the appliance. You can perform distributed computations regardless of the origin or format of the input data, provided that the data are in a format that can be read by the SAS System (for example, because a SAS/ACCESS engine is available).

In the following example, the HPREG procedure performs LASSO variable selection where the input data set is stored on the client:

```
proc hpreg data=work.one;
   model y = x1-x500;
   selection method=lasso;
   performance nodes=10 host='mydca' commit=10000;
run;
```

In order to perform the work as requested using 10 nodes on the appliance, the data set Work.One needs to be distributed to the appliance.

High-performance analytical procedures send the data in blocks to the appliance. Whenever the number of observations sent exceeds an integer multiple of the COMMIT= size, a SAS log message is produced. The message indicates the actual number of observations distributed, and not an integer multiple of the COMMIT= size.

DATASERVER=*"name"*

specifies the name of the server on Teradata systems as defined through the *hosts* file and as used in the LIBNAME statement for Teradata. For example, assume that the *hosts* file defines the server for Teradata as follows:

```
myservercop1  33.44.55.66
```

Then a LIBNAME specification would be as follows:

```
libname TDLib teradata server=myserver user= password= database= ;
```

A PERFORMANCE statement to induce running alongside the Teradata server would specify the following:

```
performance dataserver="myserver";
```

The DATASERVER= option is not required if you specify the GRIDMODE=option in the PERFORMANCE statement or if you set the GRIDMODE environment variable.

Specifying the DATASERVER= option overrides the GRIDDATASERVER environment variable.

DETAILS
requests a table that shows a timing breakdown of the procedure steps.

GRIDHOST=*"name"*
HOST=*"name"*
specifies the name of the appliance host in single or double quotation marks. If this option is specified, it overrides the value of the GRIDHOST environment variable.

GRIDMODE=SYM | ASYM
MODE=SYM | ASYM
specifies whether the high-performance analytical procedure runs in symmetric (SYM) mode or asymmetric (ASYM) mode. The default is GRIDMODE=SYM. For more information about these modes, see the section "Symmetric and Asymmetric Distributed Modes" on page 7.

If this option is specified, it overrides the GRIDMODE environment variable.

GRIDTIMEOUT=*s*
TIMEOUT=*s*
specifies the time-out in seconds for a high-performance analytical procedure to wait for a connection to the appliance and establish a connection back to the client. The default is 120 seconds. If jobs are submitted to the appliance through workload management tools that might suspend access to the appliance for a longer period, you might want to increase the time-out value.

INSTALL=*"name"*
INSTALLLOC=*"name"*
specifies the directory in which the shared libraries for the high-performance analytical procedure are installed on the appliance. Specifying the INSTALL= option overrides the GRIDINSTALLLOC environment variable.

LASRSERVER=*"path"*
LASR=*"path"*
specifies the fully qualified path to the description file of a SAS LASR Analytic Server instance. If the input data set is held in memory by this LASR Analytic Server instance, then the procedure runs alongside LASR. This option is not needed to run alongside LASR if the DATA= specification of the input data uses a libref that is associated with a LASR Analytic Server instance. For more information, see the section "Alongside-LASR Distributed Execution" on page 16 and the *SAS LASR Analytic Server: Administration Guide*.

NODES=ALL | *n*

NNODES=ALL | *n*

specifies the number of nodes in the distributed computing environment, provided that the data are not processed alongside the database.

Specifying NODES=0 indicates that you want to process the data in single-machine mode on the client machine. If the input data are not alongside the database, this is the default. The high-performance analytical procedures then perform the analysis on the client. For example, the following sets of statements are equivalent:

```
proc hplogistic data=one;
   model y = x;
run;
```

```
proc hplogistic data=one;
   model y = x;
   performance nodes=0;
run;
```

If the data are not read alongside the database, the NODES= option specifies the number of nodes on the appliance that are involved in the analysis. For example, the following statements perform the analysis in distributed mode by using 10 units of work on the appliance that is identified in the HOST= option:

```
proc hplogistic data=one;
   model y = x;
   performance nodes=10 host="hpa.sas.com";
run;
```

If the number of nodes can be modified by the application, you can specify a NODES=*n* option, where *n* exceeds the number of physical nodes on the appliance. The SAS High-Performance Text Miner software then *oversubscribes* the nodes and associates nodes with multiple units of work. For example, on a system that has 16 appliance nodes, the following statements oversubscribe the system by a factor of 3:

```
proc hplogistic data=one;
   model y = x;
   performance nodes=48 host="hpa.sas.com";
run;
```

Usually, it is not advisable to oversubscribe the system because the analytic code is optimized for a certain level of multithreading on the nodes that depends on the CPU count. You can specify NODES=ALL if you want to use all available nodes on the appliance without oversubscribing the system.

If the data are read alongside the distributed database on the appliance, specifying a nonzero value for the NODES= option has no effect. The number of units of work in the distributed computing environment is then determined by the distribution of the data and cannot be altered. For example, if you are running alongside an appliance with 24 nodes, the NODES= option in the following statements is ignored:

```
libname GPLib greenplm server=gpdca user=XXX password=YYY
              database=ZZZ;
proc hplogistic data=gplib.one;
   model y = x;
   performance nodes=10 host="hpa.sas.com";
run;
```

NTHREADS=*n*

THREADS=*n*

specifies the number of threads for analytic computations and overrides the SAS system option THREADS | NOTHREADS. If you do not specify the NTHREADS= option, the number of threads is determined based on the number of CPUs on the host on which the analytic computations execute. The algorithm by which a CPU count is converted to a thread count is specific to the high-performance analytical procedure. Most procedures create one thread per CPU for the analytic computations.

By default, high-performance analytical procedures execute in multiple concurrent threads unless multithreading has been turned off by the NOTHREADS system option or you force single-threaded execution by specifying NTHREADS=1. The largest number that can be specified for *n* is 256. Individual high-performance analytical procedures can impose more stringent limits if called for by algorithmic considerations.

NOTE: The SAS system options THREADS | NOTHREADS apply to the client machine on which the SAS high-performance analytical procedures execute. They do not apply to the compute nodes in a distributed environment.

Chapter 3
The HPTMINE Procedure

Contents

Overview: HPTMINE Procedure

The HPTMINE procedure is a high-performance procedure that analyzes large-scale textual data. PROC HPTMINE provides an essential capability for high-performance text mining and supports a wide range of fundamental text analysis features, which include tokenizing, stemming, part-of-speech tagging, noun group extraction, default or customized stop or start lists, entity parsing, multiword tokens, synonym lists, term weighting, term-by-document matrix creation, and dimension reduction by term filtering and singular value decomposition (SVD).

PROC HPTMINE integrates the functionalities that are provided by the TGPARSE, TMUTIL, and SPSVD procedures from SAS Text Miner, and achieves high efficiency and scalability through parallel processing. The HPTMINE procedure can also generate results that can be used to facilitate scoring by the HPTMSCORE procedure. These results include a configuration data set, a term data set, and a data set that contains the SVD projection matrix.

You can use the HPTMINE procedure to read data in distributed form and perform text analysis in parallel in single-machine mode or distributed mode. For more information about how to configure the execution mode of SAS high-performance analytical procedures, see the section “Processing Modes” on page 6 in Chapter 2, “Shared Concepts and Topics.”

NOTE: Distributed mode requires SAS High-Performance Text Miner.

PROC HPTMINE Features

The HPTMINE procedure processes large-scale textual data in parallel to achieve efficiency and scalability. The following list summarizes the basic features of PROC HPTMINE:

- Functionalities that are related to document parsing, term-by-document matrix creation, and dimension reduction are integrated into one procedure to process data more efficiently.

- Parsing supports essential natural language processing (NLP) features, which include tokenizing, stemming, part-of-speech tagging, noun group extraction, default or customized stop or start lists, entity parsing, multiword tokens, synonym lists.
- Term weighting and filtering are supported for term-by-document matrix creation.
- Parsing and term-by-document matrix creation is parallelized.
- Computation of singular value decomposition (SVD) is parallelized.
- Analysis can be performed on a massively parallel SAS high-performance appliance.
- All phases of processing make use of a high degree of multithreading.
- Input data can be read in parallel when the data source is the appliance database.

PROC HPTMINE Contrasted with Other SAS Procedures

The following remarks compare the HPTMINE procedure with the TGPARSE, TMUTIL, and SPSVD procedures in SAS Text Miner software.

PROC HPTMINE maximizes performance by combining the key functionalities that are found in the TGPARSE, TMUTIL, and SPSVD procedures. By functioning as an "all-in-one" procedure, PROC HPTMINE avoids expensive I/O operations. The PARSE option in the HPTMINE procedure effectively replaces the necessary options that are provided by specifying PROC TGPARSE followed by PROC TMUTIL in the conventional SAS environment. You can use the PARSE option to control parsing, accumulation, and ultimately the construction of the underlying weighted term-document frequency matrix. This matrix serves as input for singular value decomposition, which is controlled by the SVD statement in the HPTMINE procedure.

PROC HPTMINE's functionality also replaces many of the core functions that were available in the SPSVD procedure. Both PROC HPTMINE and PROC SPSVD produce an output that holds the reduced dimensional representation for the input documents. They also generate an output **U** matrix that can be used to project new documents in this same reduced space during scoring.

Getting Started: HPTMINE Procedure

The following DATA step contains 36 observations that have two variables. The text variable contains the input documents, and the did variable contains the ID of the documents. Each row in the data set represents a document for analysis.

```
data getstart;
    infile cards delimiter='|' missover;
    length text $150;
    input text$ did$;
    cards;
        High-performance analytics hold the key to |d01
```

```
        unlocking the unprecedented business value of big data.|d02
        Organizations looking for optimal ways to gain insights|d03
        from big data in shorter reporting windows are turning to SAS.|d04
        As the gold-standard leader in business analytics |d05
        for more than 36 years,|d06
        SAS frees enterprises from the limitations of |d07
        traditional computing and enables them |d08
        to draw instant benefits from big data.|d09
        Faster Time to Insight.|d10
        From banking to retail to health care to insurance, |d11
        SAS is helping industries glean insights from data |d12
        that once took days or weeks in just hours, minutes or seconds.|d13
        It's all about getting to and analyzing relevant data faster.|d14
        Revealing previously unseen patterns, sentiments and relationships.|d15
        Identifying unknown risks.|d16
        And speeding the time to insights.|d17
        High-Performance Analytics from SAS Combining industry-leading |d18
        analytics software with high-performance computing technologies|d19
        produces fast and precise answers to unsolvable problems|d20
        and enables our customers to gain greater competitive advantage.|d21
        SAS In-Memory Analytics eliminate the need for disk-based processing|d22
        allowing for much faster analysis.|d23
        SAS In-Database executes analytic logic into the database itself |d24
        for improved agility and governance.|d25
        SAS Grid Computing creates a centrally managed,|d26
        shared environment for processing large jobs|d27
        and supporting a growing number of users efficiently.|d28
        Together, the components of this integrated, |d29
        supercharged platform are changing the decision-making landscape|d30
        and redefining how the world solves big data business problems.|d31
        Big data is a popular term used to describe the exponential growth,|d32
        availability and use of information,|d33
        both structured and unstructured.|d34
        Much has been written on the big data trend and how it can |d35
        serve as the basis for innovation, differentiation and growth.|d36
run;
```

The following statements use singular value decomposition to parse the input text data and generate a lower-dimensional representation of the documents. The statements specify that only the terms that appear at least twice in the document collection be kept for generating the term-by-document matrix. The summary information about the terms in the document collection is stored in a SAS data set named terms. The SVD statement requests that the top 10 singular values and singular vectors be computed. The projection of the documents is stored in a SAS data set named docpro.

```
proc hptmine data=getstart;
doc_id      did;
variable    text;
parse
            outterms  = terms
            reducef   = 2;
svd
            k         = 10
            outdocpro = docpro;
performance details;
run;
```

The output from this analysis is presented in Figure 3.1 through Figure 3.5.

Figure 3.1 shows the "Performance Information" table, which indicates that PROC HPTMINE executes in single-machine mode. That is, PROC HPTMINE runs on the machine where the SAS system is running. The table also shows that two threads are used for computing.

Figure 3.1 Performance Information

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Figure 3.2 shows the "Procedure Task Timing" table, which provides details about how much time is used by each processing step.

Figure 3.2 Procedure Task Timing

Procedure Task Timing

Task	Time (sec.)	
Parse documents	0.92	90.9%
Analyze terms	0.00	0.03%
Obtain term frequency	0.03	3.20%
Filter terms	0.02	1.61%
Generate term-document matrix	0.01	1.40%
Output OUTTERMS table	0.01	1.41%
Compute SVD	0.01	1.14%
Output OUTDOCPRO table	0.00	0.26%

Figure 3.3 shows the SAS log that is generated by PROC HPTMINE; the log provides information about the default configurations used by the procedure, about where the procedure runs, and about the input and output files. The log shows that the terms data set contains 271 observations. This means that the HPTMINE procedure identified 271 individual terms in the input document collection. Because K=10 in the SVD statement, the docpro data set contains 11 variables: the first variable is the document ID, and the remaining 10 variables are obtained by projecting the original document to the 10 left singular vectors that are computed by singular value decomposition.

Figure 3.3 SAS Log

```
NOTE: Stemming will be used in parsing.
NOTE: Tagging will be used in parsing.
NOTE: Noun groups will be used in parsing.
NOTE: No TERMWGT option is specified. TERMWGT=ENTROPY will be run by default.
NOTE: No CELLWGT option is specified. CELLWGT=LOG will be run by default.
NOTE: No ENTITIES option is specified.ENTITIES=STD will be run by default.
NOTE: The HPTMINE procedure is executing on the client.
NOTE: There were 36 observations read from the data set WORK.GETSTART.
NOTE: The data set WORK.TERMS has 271 observations and 11 variables.
NOTE: The data set WORK.DOCPRO has 36 observations and 11 variables.
```

The following statements use PROC PRINT to show the contents of the first 10 rows of the docpro data set that is generated by the HPTMINE procedure:

```
proc print data = docpro (obs=10) round;
run;
```

Figure 3.4 shows the output of PROC PRINT. For information about the output of the OUTDOCPRO= option, see the section "The OUTDOCPRO= Data Set" on page 58.

Figure 3.4 The docpro Data Set

Obs	did	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8	COL9	COL10
1	d01	0.26	-0.76	0.21	-0.22	0.33	-0.18	0.25	0.08	-0.01	-0.24
2	d02	0.74	0.24	0.39	-0.04	-0.15	-0.02	-0.09	-0.09	-0.45	-0.01
3	d03	0.22	-0.23	-0.62	-0.23	-0.17	0.07	-0.48	-0.04	0.03	-0.44
4	d04	0.74	-0.01	0.13	0.35	0.08	-0.13	-0.41	0.19	0.18	0.23
5	d05	0.30	-0.23	0.49	-0.54	-0.30	-0.18	-0.19	0.38	0.06	0.14
6	d06	0.11	-0.24	-0.30	-0.23	-0.62	0.46	-0.08	-0.10	0.42	-0.11
7	d07	0.42	-0.33	0.06	0.40	-0.23	0.07	-0.18	-0.06	-0.55	0.39
8	d08	0.09	-0.16	-0.12	-0.40	0.26	-0.14	-0.24	-0.71	0.06	0.37
9	d09	0.85	0.10	-0.03	0.36	0.16	-0.23	-0.20	0.05	-0.06	-0.09
10	d10	0.23	-0.08	-0.60	-0.28	0.22	0.08	0.15	0.53	-0.16	0.34

The following statements use PROC SORT and PROC PRINT to show the contents of the terms data set that is generated by the HPTMINE procedure. In the output, only the terms that have a key value from 1 to 35 are shown. And for these terms only the values of the variables term, role, freq, numdocs, key, and parent are displayed.

```
proc sort data = terms; by key; run;
proc print data = terms (obs=35);
var term role freq numdocs key parent;
run;
```

Figure 3.5 shows the output of PROC PRINT, which provides details about the terms that are identified by the HPTMINE procedure. For example, it shows that the key of “problems” is 32 and its parent’s key is 9, which is the term “problem.” The HPTMINE procedure also identified that “sas” is a proper noun and that “big data” and “high-performance analytics” are noun groups. For information about the output of the OUTTERMS= option, see the section “The OUTTERMS= Data Set” on page 59.

Figure 3.5 The terms Data Set

Obs	Term	Role	Freq	numdocs	Key	Parent
1	high-performance	Adj	3	3	1	.
2	big data	NOUN_GROUP	5	5	2	.
3	to	Prep	13	11	3	.
4	fast	Adj	4	4	4	4
5	fast	Adj	1	1	4	.
6	data	Noun	8	8	5	.
7	gain	Verb	2	2	6	.
8	big	Adj	6	6	7	.
9	for	Prep	6	6	8	.
10	problem	Noun	2	2	9	9
11	as	Prep	2	2	10	.
12	business	Noun	3	3	11	.
13	analytics	Prop	2	2	12	.
14	the	Det	13	13	13	.
15	time	Noun	2	2	14	.
16	in	Prep	3	3	15	.
17	process	Verb	2	2	16	16
18	and	Conj	13	13	17	.
19	insight	Noun	3	3	18	18
20	a	Det	3	3	19	.
21	compute	Verb	2	2	20	20
22	be	Verb	3	3	21	21
23	high-performance analytics	NOUN_GROUP	2	2	22	.
24	be	Aux	3	3	23	23
25	from	Prep	6	6	24	.
26	how	Conj	2	2	25	.
27	growth	Noun	2	2	26	.
28	it	Pron	2	2	27	.
29	enable	Verb	2	2	28	28
30	of	Prep	5	5	29	.
31	analytics	Noun	3	3	30	.
32	sas	Prop	7	7	31	.
33	problems	Noun	2	2	32	9
34	insights	Noun	3	3	33	18
35	enables	Verb	2	2	34	28

Syntax: HPTMINE Procedure

The following statements are available in the HPTMINE procedure:

PROC HPTMINE < *options* > ;
  VARIABLE *variable* ;
  TARGET *variable* ;
  DOC_ID *variable* ;
  PARSE < *parse-options* > ;
  SVD < *svd-options* > ;
  PERFORMANCE *performance-options* ;

PROC HPTMINE Statement

PROC HPTMINE < *options* > ;

The PROC HPTMINE statement invokes the procedure. Table 3.1 summarizes the *options* in the statement by function. The *options* are then described fully in alphabetical order.

Table 3.1 PROC HPTMINE Statement Options

option	**Description**
Basic Options	
DATA | DOC=	Specifies the input document data set
Output Options	
NOPRINT	Suppresses ODS output

You can specify the following *options* in the PROC HPTMINE statement:

DATA=*SAS-data-set*
DOC=*SAS-data-set*
names the input SAS data set of documents to be used by PROC HPTMINE. The default is the most recently created data set. If PROC HPTMINE executes in distributed mode, the input data are distributed to memory on the appliance nodes and analyzed in parallel, unless the data are already distributed in the appliance database. When data are already distributed, PROC HPTMINE reads the data alongside the distributed database. For more information, see the sections "Processing Modes" on page 6 and "Alongside-the-Database Execution" on page 13 in Chapter 2, "Shared Concepts and Topics." Each row of the input data set must contain one text field and one ID field that correspond to the text and the unique ID of a document, respectively.

NOPRINT
suppresses the generation of ODS output.

DOC_ID Statement

DOC_ID *variable* < *docid-options* > **;**

This statement specifies the *variable* that contains the ID of each document. In the input data set, each row corresponds to one document. The ID of each document must be unique; it can be either a number or a string of characters.

You can specify the following *docid-options* in the DOC_ID statement:

DUPLICATIONCHECK | DUPCHK < (WARNING | STOP) >

checks whether the ID of each document is unique. When this option is not specified, no duplication check is performed. When you specify this option along with the WARNING keyword and duplicate document IDs are detected, the HPTMINE procedure outputs a warning message that shows the number of duplicate document IDs that have been detected. When you specify this option along with the STOP keyword, the HPTMINE procedure outputs an error message and terminates. The error message shows the number of duplicate document IDs that have been detected. If you specify the DUPLICATIONCHECK option without the WARNING or STOP keyword, the HPTMINE procedure takes the WARNING keyword by default.

PARSE Statement

PARSE < *parse-options* > **;**

The PARSE statement specifies the options for parsing the input documents and creating the term-by-document matrix. Table 3.2 summarizes the *parse-options* in the statement by function. The *parse-options* are then described fully in alphabetical order.

Table 3.2 PARSE Statement Options

parse-option	**Description**
Parsing Options	
NOSTEMMING	Specifies no stemming in parsing
NOTAGGING	Specifies no part-of-speech tagging in parsing
NONOUNGROUPS | NONG	Specifies no noun group extraction in parsing
ENTITIES=	Specifies whether to extract entities in parsing
STOP=	Specifies the stop list
START=	Specifies the start list
MULTITERM=	Specifies the multiword term list
SYNONYM | SYN=	Specifies the synonym list
Term-by-Document Matrix Creation Options	
TERMWGT=	Specifies how terms are weighted
CELLWGT=	Specifies how cells are weighted
REDUCEF=	Specifies the frequency for term filtering

Table 3.2 *continued*

parse-option	**Description**
Output Options	
OUTPARENT=	Specifies the data set to contain the term-by-document frequency matrix. Child terms are not represented in the data set. The frequencies of child terms are attributed to their corresponding parents.
OUTCHILD=	Specifies a secondary choice to contain the term-by-document frequency matrix. All kept terms, whether or not they are child terms, are represented in the data set with their corresponding frequency.
OUTTERMS=	Specifies the data set to contain the summary information about the terms in the document collection
OUTCONFIG=	Specifies the data set to contain the option settings that PROC HPTMINE uses in the current run

You can specify the following *options* in the PROC HPTMINE statement.

CELLWGT=LOG | NONE

specifies how the elements in the term-by-document matrix are weighted. The available cell weights are as follows:

LOG specifies that cells be weighted by using the LOG formulation.

NONE specifies that no cell weight be applied.

For information about the LOG formulation for cell weighting, see the section "Term and Cell Weighting" on page 55.

ENTITIES=STD | NONE

determines whether the entity extractor should use the standard list of entities. When ENTITIES=STD, entity extraction is turned on and standard entities are used; terms such as "George W. Bush" are 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. By default, ENTITIES=NONE.

MULTITERM=*OS-file-reference*

specifies the path to a file that contains a list of multiword terms. These terms are case-sensitive and are treated as a single entry by the HPTMINE procedure. Thus, the terms "Thank You" and "thank you" are processed differently. Consequently, you must convert all text strings to lowercase or add each of the multiterm's case variations to the list before using the HPTMINE procedure to create consistent multiword terms. The multiterm file can be any file type as long as it is formatted in the following manner:

```
multiterm: 3: pos
```

Specifically, the first item is the multiword term itself followed by a colon, the second item is a number that is specific to Teragram and represents the token type followed by a colon, and the third item is the part of speech that the multiword term represents. **NOTE:** The token type 3 is the most common token type for multiterm lists and represents compound words.

NONOUNGROUPS | NONG

determines whether the noun group extractor should be used. When this option is not specified, noun group are extracted and the HPTMINE procedure 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.

NOSTEMMING

determines whether words should be stemmed. When this option is not specified, words are stemmed and terms such as “advises” and “advising” are mapped to the parent term “advise.”

NOTAGGING

determines whether terms should be tagged. When this option is not specified, terms are tagged and the HPTMINE procedure identifies a term’s part of speech based on context clues. The identified part of speech is provided in the Role variable of the OUTTERMS= data set.

OUTCHILD=*SAS-data-set*

specifies the data set to contain a compressed representation of the sparse term-by-document frequency matrix. The term counts cannot be weighted. The data set saves only the kept, representative terms. The child frequencies are not attributed to their corresponding parent (as they are in the OUTPARENT= data set). For more information about the compressed representation of the sparse term-by-document frequency matrix, see the section “Output Data Sets” on page 57.

OUTCONFIG=*SAS-data-set*

specifies the data set to contain configuration information that is used for the current run of PROC HPTMINE. The primary purpose of this data set is to relay the configuration information from the HPTMINE procedure to the HPTMSCORE procedure. The HPTMSCORE procedure is forced to use options that are consistent with the HPTMINE procedure. Thus, the data set that is created by using the OUTCONFIG= option becomes an input data set for PROC HPTMSCORE and ensures that the parsing options are consistent between the two runs. For more information about this data set, see the section “Output Data Sets” on page 57.

OUTPARENT=*SAS-data-set*

specifies the data set to contain a compressed representation of the sparse term-by-document frequency matrix. The term counts can be weighted, if requested. The data set contains only the kept, representative terms, and the child frequencies are attributed to the corresponding parent. To obtain information about the children, use the OUTCHILD= option. For more information about the compressed representation of the sparse term-by-document frequency matrix, see the section “Output Data Sets” on page 57.

OUTTERMS=*SAS-data-set*

specifies the data set to contain the summary information about the terms in the document collection. For more information about this data set, see the section “Output Data Sets” on page 57.

REDUCEF=*n*

removes terms that are not in at least *n* documents. The value of *n* must be a positive integer. By default, REDUCEF=4.

START=*SAS-data-set*

specifies the data set that contains the terms that are to be kept for the analysis. These terms are displayed in the OUTTERMS= data set with a keep status of Y, and all other terms have a keep status of N. The START data set must have a Term variable and can also have a Role variable. You cannot specify both the START and the STOP options.

STOP=*SAS-data-set*
specifies the data set that contains the terms to exclude from the analysis. These terms are displayed in the OUTTERMS= data set with a keep status of N. They are not identified as parents or children. The STOP data set must have a Term variable and can also have a Role variable. You cannot specify both the START and the STOP options.

SYNONYM=*SAS-data-set*

SYN=*SAS-data-set*
specifies the data set that contains user-defined synonyms to be used in the analysis. The data set specifies parent-child relationships that enable you to map child terms to a representative parent. The synonym relationship is indicated in the data set that is specified in the OUTTERMS= option and is also reflected in the term-by-document data set that is specified in the OUTPARENT= option. The input synonym data set must have either the two variables Term and Parent or the four variables Term, Parent, Termrole, and Parentrole. When stemming is turned on, this list overrides any relationships that are identified when terms are stemmed.

TERMWGT=ENTROPY | MI | NONE
specifies how terms are weighted. The available term weights are as follows:

ENTROPY requests that terms be weighted using the entropy formulation.

MI requests that terms be weighted using the mutual information formulation.

NONE requests that no term weight be applied.

For more information about the entropy formulation and the mutual information formulation for term weighting, see the section “Term and Cell Weighting” on page 55.

PERFORMANCE Statement

PERFORMANCE < *performance-options*> **;**

The PERFORMANCE statement defines performance parameters for multithreaded and distributed computing, passes variables that describe the distributed computing environment, and requests detailed results about the performance characteristics of the HPTMINE procedure. You can also use the PERFORMANCE statement to control whether the HPTMINE procedure executes in single-machine or distributed mode. The PERFORMANCE statement is documented further in the section “PERFORMANCE Statement” on page 34 of Chapter 2, “Shared Concepts and Topics.”

SVD Statement

SVD <*svd-options*> **;**

The SVD statement specifies the options for calculating a truncated singular value decomposition (SVD) of the large, sparse term-by-document frequency matrix that was created during the parsing phase of PROC

HPTMINE. Table 3.3 summarizes the *svd-options* in the statement by function. The *svd-options* are then described fully in alphabetical order.

Table 3.3 SVD Statement Options

svd-option	**Description**
SVD Computation Options	
K=	Specifies the number of dimensions to be extracted by SVD
MAX_K=	Specifies the maximum number of dimensions to be extracted by SVD
TOL=	Specifies the maximum allowable tolerance for the singular value
RESOLUTION | RES=	Specifies the recommended number of dimensions (resolution) extracted by SVD, when the MAX_K= option is used
Output Options	
SVDU=	Specifies the U matrix, which contains the left singular vectors
SVDV=	Specifies the V matrix, which contains the right singular vectors
SVDS=	Specifies the S matrix, whose diagonal elements are the singular values
OUTDOCPRO=	Specifies the projections of the documents

You can specify the following *svd-options* in the SVD statement.

K=*k*

specifies the number of columns in the matrices U, V, and S. This value is the number of dimensions of the data set after the SVD is performed. If the value of *k* is too large, then the HPTMINE procedure runs for an unnecessarily long time. You cannot specify both this option and the MAX_K= option.

MAX_K=*n*

specifies the maximum value that the HPTMINE procedure should return as the recommended value of *n*. If the RES= argument is specified to recommend the value of *k*, this option limits that value to at most *n*. The HPTMINE procedure attempts to calculate (as opposed to recommends) *k* dimensions when it performs SVD. You cannot specify both this option and the K= option.

OUTDOCPRO=*SAS-data-set* <KEEPVARS | KEEPVARIABLES=*variable-list*>

specifies the data set to contain 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 that has much lower dimensionality.

The variables in the data set that is specified by the DATA= option of the PROC statement can be copied to the output of this option. When KEEPVARS | KEEPVARIABLES=*variable-list* is used. The content of the variables that is specified in the *variable-list* is attached to the output. These variables must appear in the data set that is specified by the DATA= option of the PROC statement.

RESOLUTION=LOW | MED | HIGH

RES=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 MAX_K= option. A low-resolution singular value decomposition returns fewer dimensions than a high-resolution singular value decomposition. This option recommends the value of k heuristically based on the value specified in the MAX_K= option. Assume that the MAX_K= option is set to *n* and a singular value decomposition with *n* dimensions accounts for $t\%$ of the total variance. The option HIGH always recommends the maximum number of dimensions; that is, $k = n$. The option MED recommends a k that explains $(5/6) * t\%$ of the total variance. The option LOW recommends a k that explains $(2/3) * t\%$ of the total variance.

SVDS=*SAS-data-set*
specifies the data set to contain the calculated singular values.

SVDU=*SAS-data-set*
specifies the data set to contain the calculated left singular vectors.

SVDV=*SAS-data-set*
specifies the data set to contain the calculated right singular vectors.

TOL=ϵ
specifies the maximum allowable tolerance for the singular value. Let $\mathbf{A}$ be a matrix. Suppose λ_i is the ith singular value of $\mathbf{A}$ and ξ_i is the corresponding right singular vector. The SVD computation terminates when for all $i \in \{1, \ldots, k\}$, λ_i and o_i satisfy $\|\mathbf{A}^\top \mathbf{A}\xi - \lambda\xi\|^2 \leq \epsilon$. The default value of ϵ is 10^{-6}, which is more than adequate for most text mining problems.

TARGET Statement

TARGET *variable* **;**

This statement specifies the *variable* that contains the information about the category that a document belongs to. A target variable can be any nominal or ordinal variable. It is used in calculating supervised term weighting, such as the mutual information formulation.

VARIABLE Statement

VARIABLE *variable* **;**

VAR *variable* **;**

This statement specifies the *variable* that contains the text to be processed.

Details: HPTMINE Procedure

PROC HPTMINE integrates the functionalities from both natural language processing and statistical analysis to provide essential support for high-performance text mining. This section provides details about various aspects of the HPTMINE procedure.

Natural Language Processing

Natural language processing (NLP) techniques can be used to extracting meaningful information from natural language input. The following sections describe features from SAS linguistic technologies that are supported in the HPTMINE procedure to support nature language processing.

Stemming

Stemming (a special case of morphological analysis) 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.

Part-of-Speech Tagging

Using SAS linguistic technologies, tagging identifies, or disambiguates, the grammatical category of a word by analyzing it within its context. For example:

```
I like to bank at the local branch of my bank.
```

In this case the first "bank" is tagged as a verb (V), and the second "bank" is tagged as a noun (N). Table 3.4 shows all possible part-of-speech tags.

Table 3.4 All Part-of-Speech Tags

Part-of-Speech Tag	Description
ABBR	Abbreviation
ADJ	Adjective
ADV	Adverb
AUX	Auxiliary or modal term
CONJ	Conjunction
DET	Determiner
INTERJ	Interjection
NOUN	Noun
NOUN_GROUP	Compound noun
NUM	Number or numeric expression
PART	Infinitive marker, negative participle, or possessive marker
PREF	Prefix
PREP	Preposition
PRON	Pronoun
PROP	Proper noun
PUNCT	Punctuation
VERB	Verb
VERBADJ	Verbal adjective

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 “weeklong 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. Table 3.5 shows all possible entities that are valid only for English.

Table 3.5 All Valid English Entities

Entities	Description
ADDRESS	Postal address or number and street name
COMPANY	Company name
CURRENCY	Currency or currency expression
INTERNET	E-mail address or URL
LOCATION	City, county, state, political or geographical place or region
MEASURE	Measurement or measurement expression
NOUN_GROUP	Phrases that contain multiple words
ORGANIZATION	Government, legal, or service agency
PERCENT	Percentage or percentage expression
PERSON	Person’s name
PHONE	Telephone number
PROP_MISC	Proper noun with an ambiguous classification
SSN	Social Security number
TIME	Time or time expression
TIME_PERIOD	Measure of time expressions
TITLE	Person’s title or position
VEHICLE	Motor vehicle, including color, year, make, and model

Multiword Terms Handling

By default, SAS linguistic technologies tokenize the text to individual words and operate at the word level. Multiword terms provide a control that enables you to specify sequences of words to be interpreted as individual units. For example, “greater than,” “in spite of,” and “as well as” can be defined as multiword terms.

Language Support

In the current release, English is the only supported language. By turning off some of the advanced parsing functionality, you might be able to use PROC HPTMINE effectively with other space-delimited languages.

Term and Cell Weighting

The TERMWGT= option and the CELLWGT= option control how to weight the frequencies in the compressed term-by-document matrix. The term weight is a positive number that is assigned to each term based on the distribution of that term in the document collection. This weight can be interpreted as an indication of the importance of that term to the document collection. The cell weight is a function that is applied to every entry in the term-by-document matrix; it moderates the effect of a term that is repeated within a document.

Let $f_{i,j}$ be the entry in the ith row and jth column of the term-by-document matrix, which indicates the time of appearance of term i in document j. Assuming that the term weight of term i is w_i and the cell weight function is $g(x)$, the weighted frequency of each entry in the term-by-document matrix is given by $w_i \times g(f_{i,j})$.

When the CELLWGT=LOG option is specified, the following equation is used to weight cells:

$$g(x) = \log_2(f_{i,j} + 1)$$

The equation reduces the influence of highly frequent terms by applying the log function.

When the TERMWGT=ENTROPY option is specified, the following equation is used to weight terms:

$$w_i = 1 + \sum_j \frac{p_{i,j} \log_2(p_{i,j})}{\log_2(n)}$$

In this equation, n is the number of documents, and $p_{i,j}$ is the probability that term i appears in document j, which can be estimated by $p_{i,j} = \frac{f_{i,j}}{g_i}$, where g_i is the global term frequency for term i.

When the TERMWGT=MI option is specified, the following equation is used to weight terms:

$$w_i = \max_{C_k} \left(\log \left(\frac{P(t_i, C_k)}{P(t_i)\, P(C_k)} \right) \right)$$

In this equation, C_k is the set of documents that belong to category k, $P(C_k)$ is the percentage of documents that belong to category k, and $P(t_i, C_k)$ is the percentage of documents that contain term t_i and belong to category k. Let d_i be the number of documents that term i appears in. Then $P(t_i) = \frac{d_i}{n}$.

Singular Value Decomposition

Singular value decomposition (SVD) of a matrix $\mathbf{A}$ factors $\mathbf{A}$ into three matrices such that $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$. The singular value decomposition also requires that the columns of $\mathbf{U}$ and $\mathbf{V}$ be orthogonal and that Σ be a real-valued diagonal matrix with monotonically decreasing, nonnegative entries. The entries of Σ are called singular values. The columns of $\mathbf{U}$ and $\mathbf{V}$ are called left and right singular vectors, respectively. A truncated singular value decomposition calculates only the first k singular values and their corresponding left and right singular vectors. In information retrieval, singular value decomposition is also known as latent semantic indexing (LSI).

Applications in Text Mining

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a term-by-document matrix, where m is the number of terms and n is the number of documents. The SVD statement has two main functions: to calculate a truncated singular value decomposition (SVD) of $\mathbf{A}$, and to project the columns of $\mathbf{A}$ onto the left singular vectors to generate a new representation of the documents with much lower dimensionality. The output of the SVD statement is a truncated singular value decomposition of $\mathbf{A}$, for which you specify the parameter k to define how many singular values and singular vectors to compute. A singular value decomposition reduces the dimension of the term-by-document matrix and reveals themes that are present in the document collection.

In general, the value of k must be large enough to capture the meaning of the document collection, yet small enough to ignore the noise. You can specify this value explicitly or accept a value that is recommended by the HPTMINE procedure. A value between 50 and 200 should work well for a document collection that contains thousands of documents.

An important purpose of singular value decomposition is to reduce a high-dimensional term-by-document matrix into a low-dimensional representation that reveals information about the document collection. The columns of the $\mathbf{A}$ form the coordinates of the document space, and the rows form the coordinates of the term space. Each document in the collection is represented as a vector in m-dimensional space and each term as a vector in n-dimensional space. The singular value decomposition captures this same information by using a smaller number of basis vectors than would be necessary if you analyzed $\mathbf{A}$ directly.

For example, consider the columns of $\mathbf{A}$, which represent the document space. By construction, the columns of $\mathbf{U}$ also reside in m-dimensional space. If $\mathbf{U}$ has only one column, the line between that vector and the origin would form the best fit line, in a least squares sense, to the original document space. If $\mathbf{U}$ has two columns, then these columns would form the best fit plane to the original document space. In general, the first k columns of $\mathbf{U}$ form the best fit k-dimensional subspace for the document space. Thus, you can project the columns of $\mathbf{A}$ onto the first k columns of $\mathbf{U}$ in order to optimally reduce the dimension of the document space from m to k.

The projection of a document $\mathbf{d}$ (one column of $\mathbf{A}$) onto $\mathbf{U}$ results in k real numbers that are defined by the inner product $\mathbf{d}$ with each column of $\mathbf{U}$. That is, $p_i = \mathbf{d}^\top \mathbf{u}_i$. With this representation, each document forms a k-dimensional vector that can be considered a theme in the document collection. You can then calculate the Euclidean distance between each document and each column of $\mathbf{U}$ to determine the documents that are described by this theme.

In a similar fashion, you can repeat the previous process by using the rows of $\mathbf{A}$ and the first k columns of $\mathbf{V}$. This generates a best fit k-dimensional subspace for the term space. This representation is used to group terms into similar clusters. These clusters also represent concepts that are prevalent in the document collection. Thus, singular value decomposition can be used to cluster both the terms and documents into meaningful representations of the entire document collection.

Computation

The computation of the singular vector decomposition is fully parallelized in PROC HPTMINE to support both single-machine mode via multithreading and distributed model via distributed computing. In the current release, computing singular value decomposition requires the input data to contain at least 25 documents and at least as many documents as there are nodes in the grid. If p nodes are used for computing singular value decomposition in a distributed computing environment, then the input data must contain at least $\max(p, 25)$ documents. Computing singular value decomposition is an iterative process that involves considerable communication among the computer nodes in a distributed computing environment. Therefore, adding more computer nodes for computing singular value decomposition might not always improve efficiency. Conversely, when the data size is not large enough, adding too many computer nodes for computation might lead to a noticeable increase in communication time and sometimes might even slow down the overall computation.

Output Data Sets

This section describes the output data sets that PROC HPTMINE produces when you specify the corresponding option.

The OUTCONFIG= Data Set

The OUTCONFIG= option specifies a SAS data set to contain the configuration that PROC HPTMINE uses in the current run. The primary purpose of this data set is to relay the configuration information from the HPTMINE procedure to the HPTMSCORE procedure. Thus, the HPTMSCORE procedure can use the options that are consistent with the HPTMINE procedure during scoring.

Table 3.6 shows the configuration information that is contained in this data set.

Table 3.6 Fields in the OUTCONFIG= Data Set

Field	Indicates
namedfile	Whether the parsed variable is in an external file or in the input data set (this field is not applicable in the current release)
language	The source language of the documents ("English" is the only supported language)
encoding	The encoding of the documents ("wlatin1" is the only supported encoding)
plugin	The plug-in that is used to parse the document collection ("Teragram" is the only supported plug-in)
stemming	Whether stemming is used: "Y" indicates that stemming is used, and "N" indicates that it is not used
tagging	Whether tagging is used: "Y" indicates that tagging is used, and "N" indicates that it is not used
NG	Whether noun group is used: "Y" indicates that noun group is used, and "N" indicates that it is not used
entities	Whether entities should be extracted: "STD" indicates that entities should be extracted, and "N" indicates that entities should not be extracted
entList	The entities that are kept (this field is not applicable in the current release)
attributes	Whether attributes are kept or dropped: "Y" indicates that attributes are kept, and "N" indicates that attributes are dropped
attrList	The attributes that are kept or dropped (this field is not applicable in the current release)
pos	Whether parts of speech are kept or dropped (this field is not applicable in the current release)

Table 3.6 *continued*

Field	Description
posList	The parts of speech that are kept or dropped (this field is not applicable in the current release)
multiterm	The path to the multiterm list file
litiList	The path to your custom entities (this field is not applicable in the current release)
cellwgt	How the cells of the term-by-document frequency matrix are weighted

The content of this data set is case-sensitive.

The OUTDOCPRO= Data Set

The OUTDOCPRO= option specifies a SAS data set to contain 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. If the K= option in the SVD statement is set to k and the input data set contains n documents, the output will have n rows and $k + 1$ columns. Each row of the output corresponds to a document. The first column of the output contains the ID of the documents, and the name of the column is the same as the variable that is specified in the DOC_ID statement. The remaining k columns are the projections and are named “COL1” to “COLk.”

The OUTPARENT= Data Set

The OUTPARENT= option specifies a SAS data set to contain a compressed representation of the sparse term-by-document frequency matrix. The term-by-document frequency matrix is usually very sparse.[1] To save space, the transactional style is used as a compressed representation to store the sparse matrix. In this format, the matrix is represented as a set of triples $(i,\ j,\ x)$, where x is an entry in the matrix and i and j denote its row and column indices, respectively. The transactional style can also be used when the output contains too many columns to fit in a data set. For example, in most database systems the maximum number of columns in a table is usually limited to several thousand. If an output matrix contains more columns than the limit, you must use the transactional style to avoid errors that would arise from writing too many columns to the table. When the transactional style is used, all 0 entries in the matrix are ignored in the output, thereby saving storing space when the matrix is sparse.

The data set of the OUTPARENT= option contains three columns: _TERMNUM_, _DOCUMENT_, and _COUNT_. The _TERMNUM_ column contains the ID of the terms, which corresponds to the “Key” column of the data set that is generated by the OUTTERMS= option. The _DOCUMENT_ column contains the ID of the documents, and the _COUNT_ column contains the term counts. For example, (`t1 d1 k`) means that term `t1` appears `k` times in document `d1`.

The term counts can be weighted, if requested. The data set saves only the terms that are marked as kept in the data set that is specified in the OUTTERMS= option of the PARSE statement. In the data set the child frequencies are attributed to the corresponding parent. For example, if “said” has term ID `t1` and appears eight times in the document `d1`, “say” has term ID `t2` and appears one time in the document `d1`. Assume that

[1] Many elements of the matrix are 0.

"say" is the parent of "said" and that neither cell weighting nor term weighting is applied. Then the data set that is specified in the OUTPARENT= option will have an entry

```
t2    d1    9
```

The term count of "said" in `d1` is attributed to its parent, "say."

The OUTCHILD= Data Set

The OUTCHILD= option specifies the data set to contain a compressed representation of the sparse term-by-document frequency matrix, which is usually very sparse. To save storage space, the transactional style is used as a compressed representation to store the sparse matrix. Like the data set that is specified in the OUTPARENT= option, the data set specified in the OUTCHILD= option also contains three columns: _TERMNUM_, _DOCUMENT_, and _COUNT_. The _TERMNUM_ column contains the ID of the terms, which corresponds to the "Key" column of the data set generated by the OUTTERMS= option. The _DOCUMENT_ column contains the ID of the documents, and the _COUNT_ column contains the term counts. For example, (`t1 d1 k`) means that term `t1` appears `k` times in document `d1`.

The data set saves only the terms that are marked as kept in the data set specified in the OUTTERMS= option of the PARSE statement. The child frequencies are not attributed to their corresponding parent (as they are in the data set specified in the OUTPARENT= option). Using the example in the previous section, the data set that is generated by the OUTCHILD= option will have two entries:

```
t1    d1    8
t2    d1    1
```

The term count of "said" in `d1` is not attributed to its parent, "say." The data set that is specified in the OUTCHILD= option can be combined with the data set that is specified in the OUTTERMS= option to construct the data set that is specified in the OUTPARENT= option.

The OUTTERMS= Data Set

The data set specified in the OUTTERMS= option contains the summary information about the terms in the document collection. Table 3.7 shows the fields in this data set.

Table 3.7 Fields in the OUTTERMS= Data Set

Fields	Description	
Term	A lowercase version of the term	
Role	The term's part of speech (this variable is empty if the NOTAGGING option is specified in the PARSE statement)	
Attribute	An indication of the characters that compose the term. Possible attributes are as follows:	
	Alpha	only alphabetic characters
	Mixed	a combination of attributes
	Num	only numbers
	Punct	punctuation characters
	Entity	an identified entity

Table 3.7 *continued*

Fields	Description
Freq	The frequency of a term in the entire document collection
numdocs	The number of documents that contain the term
_keep	The keep status of the term: "Y" indicates that the term is kept for analysis, and "N" indicates that the term should be dropped in later stages of analysis
Key	The assigned term number (each unique term in the parsed documents and each unique parent term has a unique Key value)
Parent	The Key value of the term's parent or a "." (period): • If a term has a parent, this variable gives the term number of that parent. • If a term does not have a parent, this value is a "." (period). • If the values of Key, Parent, and Parent_id are identical, the parent occurs as itself. • If the values of Parent and Parent_id are identical but differ from Key, the observation is a child.
Parent_id	Another description of the term's parent: Parent gives the parent's term number if a term is a child, but Parent_id gives this value for all terms.
_ispar	An indication of term's status as a parent, child, or neither: • A "+" (plus sign) indicates that the term is a parent. • A "." (period) indicates that the term is a child. • A missing value indicates that the term is neither a parent nor a child.
Weight	The weights of the terms

Displayed Output

The following sections describe the output that PROC HPTMINE produces. The output is organized into various tables, which are discussed in their order of appearance.

Performance Information

The "Performance Information" table is produced by default; it displays information about the grid host for distributed execution and information about whether the procedure executes in single-machine mode, distributed mode, or alongside-the-database mode. The numbers of computing nodes and threads are also displayed, depending on the environment.

Procedure Task Timing

When the DETAILS option is specified in the PERFORMANCE statement, PROC HPTMINE produces a "Procedure Task Timing" table, which displays the elapsed time (absolute and relative) for the main tasks.

ODS Table Names

Each table that is created by the HPTMINE procedure has a name associated with it, and you must use this name to refer to the table when you use ODS statements. These names are listed in Table 3.8.

Table 3.8 ODS Tables Produced by PROC HPTMINE

Table Name	Description	Required Statement / Option
PerformanceInfo	Information about the high-performance computing environment	Default output
Timing	Absolute and relative times for tasks performed by the procedure	DETAILS option in the PERFORMANCE statement

System Configuration

Prerequisites for Running PROC HPTMINE

To use the HPTMINE procedure, you must have a valid SAS Text Miner license, and the language binary files that are provided under that license must be available to PROC HPTMINE for parsing text.

Input Data Accessibility

When you run PROC HPTMINE in distributed mode, the following input data sets need to be accessible to the single-machine machine:

- MULTITERM= data set
- START= data set
- STOP= data set
- SYNONYM= data set

When the HPTMINE procedure runs in distributed mode, it first reads these data sets on the single-machine machine and then sends them to the grid.

Configuration for Distributed Mode

When PROC HPTMINE runs in distributed mode, it needs to locate on the grid the language binary files that are used in parsing text. These binary files must be deployed to the grid. In addition, the GRID_TEXTANALYTICS_BIN_LOC macro variable must be specified to indicate the location of the binary files on the grid.

Deploying Language Binary Files on the Grid

When SAS Text Miner is installed, it includes language binary files that are used to parse text. These files can be found in a folder under `$SASROOT`. To run PROC HPTMINE in a grid environment, the binary files must be copied from the `$SASROOT` folder to the grid.

Choose one of the following methods to copy the binary files to the grid:

- Place the files in a location that is accessible to all nodes of the grid.
- Place a copy of the files on each node of the grid.

The binary files are very large, and a shared location requires less space to store them. However, using a shared location means that the files need to be distributed to the grid when PROC HPTMINE runs, and this can be time-consuming. Consult the grid administrator for a recommended binary location on the grid.

Installation Location of Language Binary Files

The language binary files are originally installed in the following locations:

- In a Windows environment, the files are installed in the subdirectory `tktg\sasmisc` under `$SASROOT`. For example: `C:\Program Files\SASHome\SASFoundation\9.3\tktg\sasmisc`.
- In a UNIX environment, the files are installed in `$SASROOT/misc/tktg`.

English is the only language that PROC HPTMINE supports in this release, so you need to copy to the grid only the following binary files, which are used for English processing:

- *en-compounds.txt*
- *en-ne.li*
- *en-sastags.txt*
- *en-utf8.stkzo*
- *en-utf8.tkzo*
- *en-utf8-AnaInfSinFas.mdic*
- *en-utf8-std.htagger*
- *case_mapping.bin*

- *chlangid.bin*
- *chmap.bin*
- *tix.config*

If you have licensed SAS Text Miner languages other than English, you will see other files in your installation directory. You do not need to copy the files for other licensed languages to the grid.

The GRID_TEXTANALYTICS_BIN_LOC Macro

After you copy the language binary files to the grid, you can use the GRID_TEXTANALYTICS_BIN_LOC macro to tell the HPTMINE procedure where to find these binary files. The GRID_TEXTANALYTICS_BIN_LOC macro is required only when you run PROC HPTMINE on the grid. When you run PROC HPTMINE in single-machine mode, the language binary files that were installed in the `$SASROOT` location during your SAS Text Miner installation are used for parsing text.

Assume that the grid administrator has copied the language binary files to a directory, named `/global_dir/tktg/misc`, which is accessible to all nodes of the grid. To tell PROC HPTMINE the location of the language binary files, insert the following statement before the statements that call the procedure:

```
%let GRID_TEXTANALYTICS_BIN_LOC=/global_dir/tktg/misc;
```

When storage space permits, you can ensure optimal performance by placing a copy of the language binary files on each node and using a relative pathname for the GRID_TEXTANALYTICS_BIN_LOC macro variable. For example, the grid administrator can create a directory whose pathname is `/local_dir/tktg/misc` on each node and can store the language binary files in that directory. When the following statement is specified, the HPTMINE procedure goes to the directory `/local_dir/tktg/misc` on each node to load the binary files:

```
%let GRID_TEXTANALYTICS_BIN_LOC=/local_dir/tktg/misc;
```

Examples: HPTMINE Procedure

Example 3.1: Parsing with No Options Turned On

This example parses five documents, which are in a generated data set. The following DATA step generates the five documents:

```
/* 1)  create data set */

data CarNominations;
infile cards delimiter='|' missover;
length text $70 ;
input text$ i;
cards;
    The Ford Taurus is the World Car of the Year. |1
```

```
        Hyundai won the award last year. |2
        Toyota sold the Toyota Tacoma in bright green. |3
        The Ford Taurus is sold in all colors except for lime green. |4
        The Honda Insight was World Car of the Year in 2008. |5
        ;
    run;
```

The following statements run PROC HPTMINE to parse the documents. Because no host for distributed computing is specified and the NODES option is not specified in the PERFORMANCE statement, PROC HPTMINE runs automatically in soloist mode.

```
/* 2) starting code */
proc hptmine data=work.CarNominations;
doc_id i;
var text;
parse
    nostemming notagging nonoungroups
    termwgt        = none
    cellwgt        = none
    reducef        = 0
    entities       = none
    outparent      = outparent
    outterms       = outterms
    outchild       = outchild
    outconfig      = outconfig
    ;
performance details;
run;

/* 3) print outterms data set */
proc print data=outterms; title "OUTTERMS Data Set"; run;
```

Output 3.1.1 shows the content of the outterms data set. In this example, none of the following parsing options are turned on: stemming, part-of-speech tagging, noun group extraction, entity identification, term and cell weighting, and term filtering. No synonym list, multiterms word list, or stop list is specified. As a result of this configuration, there is no child term in the outterms data set. Also, the outparent data set and the outchild data set are exactly the same. The HPTMINE procedure automatically drops punctuation and numbers. For example, in the outterms data set the "." (period) term has a "keep" value of "N".

Output 3.1.1 The outterms Data Set

Obs	Term	Role	Attribute	Freq	numdocs	_keep	Key	Parent	Parent_ id	_ispar	Weight
1	.		Punct	5	1	N	27	.	27		1
2	all		Alpha	1	1	Y	1	.	1		1
3	toyota		Alpha	2	1	Y	2	.	2		1
4	ford		Alpha	2	1	Y	3	.	3		1
5	tacoma		Alpha	1	1	Y	4	.	4		1
6	year		Alpha	3	1	Y	5	.	5		1
7	taurus		Alpha	2	1	Y	6	.	6		1
8	won		Alpha	1	1	Y	7	.	7		1
9	honda		Alpha	1	1	Y	8	.	8		1
10	bright		Alpha	1	1	Y	9	.	9		1
11	sold		Alpha	2	1	Y	10	.	10		1
12	colors		Alpha	1	1	Y	11	.	11		1
13	lime		Alpha	1	1	Y	12	.	12		1
14	except		Alpha	1	1	Y	13	.	13		1
15	hyundai		Alpha	1	1	Y	14	.	14		1
16	in		Alpha	3	1	Y	15	.	15		1
17	is		Alpha	2	1	Y	16	.	16		1
18	for		Alpha	1	1	Y	17	.	17		1
19	world		Alpha	2	1	Y	18	.	18		1
20	green		Alpha	2	1	Y	19	.	19		1
21	the		Alpha	8	1	Y	20	.	20		1
22	of		Alpha	2	1	Y	21	.	21		1
23	award		Alpha	1	1	Y	22	.	22		1
24	was		Alpha	1	1	Y	23	.	23		1
25	2008		Num	1	1	N	28	.	28		1
26	car		Alpha	2	1	Y	24	.	24		1
27	insight		Alpha	1	1	Y	25	.	25		1
28	last		Alpha	1	1	Y	26	.	26		1

Example 3.2: Parsing with Stemming Turned On

This example uses the data set that is generated in Example 3.1. The following statements run PROC HPTMINE to parse the documents. Because the NOSTEMMING option is not specified in the PARSE statement, words are stemmed (the default).

```
proc hptmine data=work.CarNominations;
doc_id i;
var text;
parse
    notagging nonoungroups
    termwgt  = none
    cellwgt  = none
    reducef  = 0
    entities = none
```

```
      outparent= outparent
      outterms = outterms
      outchild = outchild
      outconfig= outconfig
      ;
   performance details;
   run;

   proc print data = outterms; title "OUTTERMS Data Set"; run;
```

Output 3.2.1 shows the content of the outterms data set. In this example, words are stemmed. You can see that the term “sold” now stems to have the parent term “sell.” Also, the outparent data set and the outchild data set are different. The parent term “sell” shows up in outparent (key=11), but not the child term “sold” (key=26). Only “sold” appears in the outchild data set, and “sell” does not appear.

Output 3.2.1 The outterms Data Set with Stemming

Obs	Term	Role	Attribute	Freq	numdocs	_keep	Key	Parent	Parent_id	_ispar	Weight
1	.		Punct	5	1	N	30	.	30		1
2	all		Alpha	1	1	Y	1	.	1		1
3	toyota		Alpha	2	1	Y	2	.	2		1
4	ford		Alpha	2	1	Y	3	.	3		1
5	tacoma		Alpha	1	1	Y	4	.	4		1
6	year		Alpha	3	1	Y	5	.	5		1
7	taurus		Alpha	2	1	Y	6	.	6		1
8	won		Alpha	1	1	Y	7	.	7		1
9	honda		Alpha	1	1	Y	8	.	8		1
10	bright		Alpha	1	1	Y	9	.	9		1
11	be		Alpha	3	1	Y	10	10	10	+	1
12	sold		Alpha	2	1	Y	26	11	11	.	1
13	sell		Alpha	2	1	Y	11	11	11	+	1
14	colors		Alpha	1	1	Y	27	23	23	.	1
15	lime		Alpha	1	1	Y	12	.	12		1
16	except		Alpha	1	1	Y	13	.	13		1
17	hyundai		Alpha	1	1	Y	14	.	14		1
18	in		Alpha	3	1	Y	15	.	15		1
19	is		Alpha	2	1	Y	28	10	10	.	1
20	for		Alpha	1	1	Y	16	.	16		1
21	world		Alpha	2	1	Y	17	.	17		1
22	green		Alpha	2	1	Y	18	.	18		1
23	the		Alpha	8	1	Y	19	.	19		1
24	of		Alpha	2	1	Y	20	.	20		1
25	award		Alpha	1	1	Y	21	.	21		1
26	was		Alpha	1	1	Y	29	10	10	.	1
27	2008		Num	1	1	N	31	.	31		1
28	car		Alpha	2	1	Y	22	.	22		1
29	color		Alpha	1	1	Y	23	23	23	+	1
30	insight		Alpha	1	1	Y	24	.	24		1
31	last		Alpha	1	1	Y	25	.	25		1

Example 3.3: Adding Entities and Noun Groups

This example uses the data set that is generated in Example 3.1. The following statements run PROC HPTMINE to parse the documents. Because the NONOUNGROUP option is not specified in the PARSE statement, noun group extraction is turned on, and because the ENTITIES option is not specified, entity identification is turned on.

```
proc hptmine data=work.CarNominations;
doc_id i;
var text;
PARSE
    notagging
    termwgt   = none
    cellwgt   = none
    reducef   = 0
    entities  = std
    outparent = outparent
    outterms  = outterms
    outchild  = outchild
    outconfig = outconfig
    ;
performance details;
run;

proc print data=outterms; title "OUTTERMS Data Set"; run;
```

Output 3.3.1 shows the content of the outterms data set. Compared to Output 3.2.1, the outterms data set is longer, because it contains entities and noun groups. For example, “honda insight” is included in the outterms data set as an entity with Role=Vehicle, and “bright green” is also included in the outterms data set as a noun group.

Output 3.3.1 The outterms Data Set with Noun Group Extraction and Entity Identification

Obs	Term	Role	Attribute	Freq	numdocs	_keep	Key	Parent	Parent_ id
1	lime green	NOUN_GROUP	Alpha	1	1	Y	1	.	1
2	.		Punct	5	1	N	37	.	37
3	all		Alpha	1	1	Y	2	.	2
4	toyota		Alpha	1	1	Y	3	.	3
5	ford		Alpha	2	1	Y	4	.	4
6	toyota tacoma	VEHICLE	Entity	1	1	Y	5	.	5
7	tacoma		Alpha	1	1	Y	6	.	6
8	year		Alpha	3	1	Y	7	.	7
9	taurus		Alpha	2	1	Y	8	.	8
10	won		Alpha	1	1	Y	9	.	9
11	honda		Alpha	1	1	Y	10	.	10
12	bright		Alpha	1	1	Y	11	.	11
13	be		Alpha	3	1	Y	12	12	12
14	sold		Alpha	2	1	Y	33	15	15
15	bright green	NOUN_GROUP	Alpha	1	1	Y	13	.	13
16	ford taurus	VEHICLE	Entity	2	1	Y	14	.	14
17	sell		Alpha	2	1	Y	15	15	15
18	hyundai	COMPANY	Entity	1	1	Y	16	.	16
19	colors		Alpha	1	1	Y	34	30	30
20	lime		Alpha	1	1	Y	17	.	17
21	except		Alpha	1	1	Y	18	.	18
22	honda insight	VEHICLE	Entity	1	1	Y	19	.	19
23	in		Alpha	3	1	Y	20	.	20
24	is		Alpha	2	1	Y	35	12	12
25	for		Alpha	1	1	Y	21	.	21
26	toyota	COMPANY	Entity	1	1	Y	22	.	22
27	world		Alpha	2	1	Y	23	.	23
28	green		Alpha	2	1	Y	24	.	24
29	the		Alpha	8	1	Y	25	.	25
30	of		Alpha	2	1	Y	26	.	26
31	world car	PROP_MISC	Entity	2	1	Y	27	.	27
32	award		Alpha	1	1	Y	28	.	28
33	was		Alpha	1	1	Y	36	12	12
34	2008		Num	1	1	N	38	.	38
35	car		Alpha	2	1	Y	29	.	29
36	color		Alpha	1	1	Y	30	30	30
37	insight		Alpha	1	1	Y	31	.	31
38	last		Alpha	1	1	Y	32	.	32

Example 3.4: Adding Part-of-Speech Tagging

This example uses the data set that is generated in Example 3.1. The following statements run PROC HPTMINE to parse the documents. Because the NOTAGGING option is not specified in the PARSE statement, PROC HPTMINE uses context clues to determine a term's part of speech.

```
proc hptmine data=work.CarNominations;
doc_id i;
var text;
PARSE
    termwgt    = none
    cellwgt    = none
    reducef    = 0
    entities   = std
    outparent = outparent
    outterms   = outterms
    outchild   = outchild
    outconfig = outconfig
    ;
performance details;
run;

proc print data=outterms; title "OUTTERMS Data Set"; run;
```

Output 3.4.1 shows the content of the outterms data set. Compared to Output 3.3.1, the outterms data set also contains the part-of-speech tag for the terms.

Output 3.4.1 The outterms Data Set with Part-of-Speech Tagging

Obs	Term	Role	Attribute	Freq	numdocs	_keep	Key	Parent	Parent_ id
1	car	Prop	Alpha	1	1	Y	1	.	1
2	lime green	NOUN_GROUP	Alpha	1	1	Y	2	.	2
3	last	Adj	Alpha	1	1	Y	3	.	3
4	lime	Adj	Alpha	1	1	Y	4	.	4
5	sell	Verb	Alpha	2	1	Y	5	5	5
6	won	Noun	Alpha	1	1	Y	6	.	6
7	toyota tacoma	VEHICLE	Entity	1	1	Y	7	.	7
8	bright	Adj	Alpha	1	1	Y	8	.	8
9	honda	Prop	Alpha	1	1	Y	9	.	9
10	sold	Verb	Alpha	2	1	Y	34	5	5
11	colors	Noun	Alpha	1	1	Y	35	12	12
12	for	Prep	Alpha	1	1	Y	10	.	10
13	award	Verb	Alpha	1	1	Y	11	.	11
14	color	Noun	Alpha	1	1	Y	12	12	12
15	2008	Num	Num	1	1	N	38	.	38
16	bright green	NOUN_GROUP	Alpha	1	1	Y	13	.	13
17	the	Det	Alpha	8	1	Y	14	.	14
18	is	Verb	Alpha	2	1	Y	36	26	26
19	ford taurus	VEHICLE	Entity	2	1	Y	15	.	15
20	hyundai	COMPANY	Entity	1	1	Y	16	.	16
21	toyota	Prop	Alpha	1	1	Y	17	.	17
22	car	Noun	Alpha	1	1	Y	18	.	18
23	except	Verb	Alpha	1	1	Y	19	.	19
24	in	Prep	Alpha	3	1	Y	20	.	20
25	was	Verb	Alpha	1	1	Y	37	26	26
26	honda insight	VEHICLE	Entity	1	1	Y	21	.	21
27	insight	Noun	Alpha	1	1	Y	22	.	22
28	ford	Prop	Alpha	2	1	Y	23	.	23
29	world	Prop	Alpha	2	1	Y	24	.	24
30	.	Punct	Punct	5	1	N	39	.	39
31	year	Noun	Alpha	3	1	Y	25	.	25
32	be	Verb	Alpha	3	1	Y	26	26	26
33	all	Adj	Alpha	1	1	Y	27	.	27
34	toyota	COMPANY	Entity	1	1	Y	28	.	28
35	taurus	Noun	Alpha	2	1	Y	29	.	29
36	world car	PROP_MISC	Entity	2	1	Y	30	.	30
37	of	Prep	Alpha	2	1	Y	31	.	31
38	tacoma	Prop	Alpha	1	1	Y	32	.	32
39	green	Noun	Alpha	2	1	Y	33	.	33

Example 3.5: Adding Synonyms

This example uses the data set that is generated in Example 3.1. So far, by looking at the outterms data sets that are generated by Example 3.1 to Example 3.4, you can see that the data are very "vehicle focused." But suppose what you really care about are the Companies. You can use a synonym list in parsing to map each vehicle to the company that produces it. The following statements show this mapping.

```
data synds;
infile cards delimiter=',';
length TERM $13;
input TERM $ TERMROLE $ PARENT $ parentrole$;
cards;
    honda insight, VEHICLE , honda,  COMPANY,
    ford taurus,   VEHICLE,  ford,   COMPANY,
    toyota tacoma, VEHICLE,  toyota, COMPANY,
;
run;

proc hptmine data=work.CarNominations;
doc_id i;
var text;
parse
    termwgt    = none
    cellwgt    = none
    reducef    = 0
    entities   = std
    synonym    = synds
    outparent  = outparent
    outterms   = outterms
    outchild   = outchild
    outconfig  = outconfig
    ;
PERFORMANCE details;
RUN;

proc print data=outterms; title "OUTTERMS Data Set"; run;
```

Output 3.5.1 shows the content of the outterms data set. You can see that the term "honda insight" (key=39) is assigned the parent term "honda" (key=5). Only the term "honda" appears in the outparent data set.

Output 3.5.1 The outterms Data Set with Synonym Mapping

Obs	Term	Role	Attribute	Freq	numdocs	_keep	Key	Parent	Parent_id
1	car	Prop	Alpha	1	1	Y	1	.	1
2	lime green	NOUN_GROUP	Alpha	1	1	Y	2	.	2
3	last	Adj	Alpha	1	1	Y	3	.	3
4	lime	Adj	Alpha	1	1	Y	4	.	4
5	honda	COMPANY	Entity	1	1	Y	5	5	5
6	sell	Verb	Alpha	2	1	Y	6	6	6
7	won	Noun	Alpha	1	1	Y	7	.	7
8	toyota tacoma	VEHICLE	Entity	1	1	Y	33	26	26
9	bright	Adj	Alpha	1	1	Y	8	.	8
10	honda	Prop	Alpha	1	1	Y	9	.	9
11	sold	Verb	Alpha	2	1	Y	34	6	6
12	colors	Noun	Alpha	1	1	Y	35	12	12
13	for	Prep	Alpha	1	1	Y	10	.	10
14	award	Verb	Alpha	1	1	Y	11	.	11
15	color	Noun	Alpha	1	1	Y	12	12	12
16	2008	Num	Num	1	1	N	40	.	40
17	bright green	NOUN_GROUP	Alpha	1	1	Y	13	.	13
18	the	Det	Alpha	8	1	Y	14	.	14
19	is	Verb	Alpha	2	1	Y	36	24	24
20	ford taurus	VEHICLE	Entity	2	1	Y	37	27	27
21	hyundai	COMPANY	Entity	1	1	Y	15	.	15
22	toyota	Prop	Alpha	1	1	Y	16	.	16
23	car	Noun	Alpha	1	1	Y	17	.	17
24	except	Verb	Alpha	1	1	Y	18	.	18
25	in	Prep	Alpha	3	1	Y	19	.	19
26	was	Verb	Alpha	1	1	Y	38	24	24
27	honda insight	VEHICLE	Entity	1	1	Y	39	5	5
28	insight	Noun	Alpha	1	1	Y	20	.	20
29	ford	Prop	Alpha	2	1	Y	21	.	21
30	world	Prop	Alpha	2	1	Y	22	.	22
31	.	Punct	Punct	5	1	N	41	.	41
32	year	Noun	Alpha	3	1	Y	23	.	23
33	be	Verb	Alpha	3	1	Y	24	24	24
34	all	Adj	Alpha	1	1	Y	25	.	25
35	toyota	COMPANY	Entity	2	1	Y	26	26	26
36	toyota	COMPANY	Entity	1	1	Y	26	.	26
37	ford	COMPANY	Entity	2	1	Y	27	27	27
38	taurus	Noun	Alpha	2	1	Y	28	.	28
39	world car	PROP_MISC	Entity	2	1	Y	29	.	29
40	of	Prep	Alpha	2	1	Y	30	.	30
41	tacoma	Prop	Alpha	1	1	Y	31	.	31
42	green	Noun	Alpha	2	1	Y	32	.	32

Example 3.6: Adding a Stop List

This example uses the data set that is generated in Example 3.1. If you want to eliminate TOYOTA from the analysis, you can enter the parent term "Toyota" with role=COMPANY in a stop list. When this stop list is an input, PROC HPTMIME drops the term "Toyota" (role=COMPANY) and all its children terms in the outterms data set by marking Keep=N for these terms.

```
data synds;
infile cards delimiter=',';
length TERM $13;
input TERM $ TERMROLE $ PARENT $ parentrole$;
cards;
    honda insight, VEHICLE , honda,   COMPANY,
    ford taurus,   VEHICLE,  ford,    COMPANY,
    toyota tacoma, VEHICLE,  toyota, COMPANY,
;
run;

/* create stop list*/
data stopList;
infile cards delimiter='|' missover;
length term $25 role $40;
input term$ role$ ;
    cards;
        toyota| COMPANY
    ;
run;

proc hptmine data=work.CarNominations;
doc_id i;
var text;
parse
    termwgt    = none
    cellwgt    = none
    reducef    = 0
    entities   = std
    synonym    = synds
    stop       = stopList
    outparent  = outparent
    outterms   = outterms
    outchild   = outchild
    outconfig  = outconfig
    ;
performance details;
run;

proc print data=outterms; title "OUTTERMS Data Set"; run;
```

Output 3.6.1 shows the content of the outterms data set. You can see that the term "Toyota" (key=40) and the term "Toyota Tacoma" (key=41) are assigned Keep=N. The outparent data set is shorter than the one generated in Example 3.5.

Output 3.6.1 The outterms Data Set Filtered Using Stop List

Obs	Term	Role	Attribute	Freq	numdocs	_keep	Key	Parent	Parent_ id
1	car	Prop	Alpha	1	1	Y	1	.	1
2	lime green	NOUN_GROUP	Alpha	1	1	Y	2	.	2
3	last	Adj	Alpha	1	1	Y	3	.	3
4	lime	Adj	Alpha	1	1	Y	4	.	4
5	honda	COMPANY	Entity	1	1	Y	5	5	5
6	sell	Verb	Alpha	2	1	Y	6	6	6
7	won	Noun	Alpha	1	1	Y	7	.	7
8	toyota tacoma	VEHICLE	Entity	1	1	N	41	40	40
9	bright	Adj	Alpha	1	1	Y	8	.	8
10	honda	Prop	Alpha	1	1	Y	9	.	9
11	sold	Verb	Alpha	2	1	Y	31	6	6
12	colors	Noun	Alpha	1	1	Y	32	12	12
13	for	Prep	Alpha	1	1	Y	10	.	10
14	award	Verb	Alpha	1	1	Y	11	.	11
15	color	Noun	Alpha	1	1	Y	12	12	12
16	2008	Num	Num	1	1	N	37	.	37
17	bright green	NOUN_GROUP	Alpha	1	1	Y	13	.	13
18	the	Det	Alpha	8	1	Y	14	.	14
19	is	Verb	Alpha	2	1	Y	33	23	23
20	ford taurus	VEHICLE	Entity	2	1	Y	34	25	25
21	hyundai	COMPANY	Entity	1	1	Y	15	.	15
22	toyota	Prop	Alpha	1	1	N	38	.	38
23	car	Noun	Alpha	1	1	Y	16	.	16
24	except	Verb	Alpha	1	1	Y	17	.	17
25	in	Prep	Alpha	3	1	Y	18	.	18
26	was	Verb	Alpha	1	1	Y	35	23	23
27	honda insight	VEHICLE	Entity	1	1	Y	36	5	5
28	insight	Noun	Alpha	1	1	Y	19	.	19
29	ford	Prop	Alpha	2	1	Y	20	.	20
30	world	Prop	Alpha	2	1	Y	21	.	21
31	.	Punct	Punct	5	1	N	39	.	39
32	year	Noun	Alpha	3	1	Y	22	.	22
33	be	Verb	Alpha	3	1	Y	23	23	23
34	all	Adj	Alpha	1	1	Y	24	.	24
35	toyota	COMPANY	Entity	2	1	N	40	40	40
36	toyota	COMPANY	Entity	1	1	N	40	.	40
37	ford	COMPANY	Entity	2	1	Y	25	25	25
38	taurus	Noun	Alpha	2	1	Y	26	.	26
39	world car	PROP_MISC	Entity	2	1	Y	27	.	27
40	of	Prep	Alpha	2	1	Y	28	.	28
41	tacoma	Prop	Alpha	1	1	Y	29	.	29
42	green	Noun	Alpha	2	1	Y	30	.	30

Example 3.7: Adding a Multiterm List

A multiterm list can be used to define terms that consist of multiple words. This example uses the data set that is generated in Example 3.1 to show how to use MULTITERM option.

```
data synds;
infile cards delimiter=',';
length TERM $13;
input TERM $ TERMROLE $ PARENT $ parentrole$;
cards;
    honda insight, VEHICLE , honda,   COMPANY,
    ford taurus,   VEHICLE,  ford,    COMPANY,
    toyota tacoma, VEHICLE,  toyota, COMPANY,
;
run;

data stopList;
infile cards delimiter='|' missover;
length term $25 role $40;
input term$ role$ ;
    cards;
        toyota| COMPANY
    ;
run;

/* create multiterms list*/
data _null_ ;
    FILE  'encomp.txt' ;
    PUT 'except for :3:Prep';
run ;

proc hptmine data=work.CarNominations;
doc_id i;
var text;
parse
    termwgt     = none
    cellwgt     = none
    reducef     = 0
    entities    = std
    synonym     = synds
    stop        = stopList
    multiterm   ="encomp.txt"
    outparent   = outparent
    outterms    = outterms
    outchild    = outchild
    outconfig   = outconfig
    ;
performance details;
run;

proc print data=outterms; title "OUTTERMS Data Set"; run;
```

Output 3.7.1 shows the content of the outterms data set. In the preceding statements, “except for” is defined as an individual term. In the outterms data set, you can see that the two terms “except” and “for” have become one term, “except for.”

Output 3.7.1 The outterms Data Set

Obs	Term	Role	Attribute	Freq	numdocs	_keep	Key	Parent	Parent_id
1	car	Prop	Alpha	1	1	Y	1	.	1
2	lime green	NOUN_GROUP	Alpha	1	1	Y	2	.	2
3	last	Adj	Alpha	1	1	Y	3	.	3
4	lime	Adj	Alpha	1	1	Y	4	.	4
5	honda	COMPANY	Entity	1	1	Y	5	5	5
6	sell	Verb	Alpha	2	1	Y	6	6	6
7	won	Noun	Alpha	1	1	Y	7	.	7
8	toyota tacoma	VEHICLE	Entity	1	1	N	40	39	39
9	bright	Adj	Alpha	1	1	Y	8	.	8
10	honda	Prop	Alpha	1	1	Y	9	.	9
11	sold	Verb	Alpha	2	1	Y	30	6	6
12	colors	Noun	Alpha	1	1	Y	31	12	12
13	award	Verb	Alpha	1	1	Y	10	.	10
14	except for	Pre	Entity	1	1	Y	11	.	11
15	color	Noun	Alpha	1	1	Y	12	12	12
16	2008	Num	Num	1	1	N	36	.	36
17	bright green	NOUN_GROUP	Alpha	1	1	Y	13	.	13
18	the	Det	Alpha	8	1	Y	14	.	14
19	is	Verb	Alpha	2	1	Y	32	22	22
20	ford taurus	VEHICLE	Entity	2	1	Y	33	24	24
21	hyundai	COMPANY	Entity	1	1	Y	15	.	15
22	toyota	Prop	Alpha	1	1	N	37	.	37
23	car	Noun	Alpha	1	1	Y	16	.	16
24	in	Prep	Alpha	3	1	Y	17	.	17
25	was	Verb	Alpha	1	1	Y	34	22	22
26	honda insight	VEHICLE	Entity	1	1	Y	35	5	5
27	insight	Noun	Alpha	1	1	Y	18	.	18
28	ford	Prop	Alpha	2	1	Y	19	.	19
29	world	Prop	Alpha	2	1	Y	20	.	20
30	.	Punct	Punct	5	1	N	38	.	38
31	year	Noun	Alpha	3	1	Y	21	.	21
32	be	Verb	Alpha	3	1	Y	22	22	22
33	all	Adj	Alpha	1	1	Y	23	.	23
34	toyota	COMPANY	Entity	2	1	N	39	39	39
35	toyota	COMPANY	Entity	1	1	N	39	.	39
36	ford	COMPANY	Entity	2	1	Y	24	24	24
37	taurus	Noun	Alpha	2	1	Y	25	.	25
38	world car	PROP_MISC	Entity	2	1	Y	26	.	26
39	of	Prep	Alpha	2	1	Y	27	.	27
40	tacoma	Prop	Alpha	1	1	Y	28	.	28
41	green	Noun	Alpha	2	1	Y	29	.	29

Example 3.8: Running in Distributed Mode

This example uses the data set that is generated in the section "Getting Started: HPTMINE Procedure" on page 41. When a host for distributed computing is specified and the NODES option in the PERFORMANCE statement is specified as in the following statements, PROC HPTMINE uses the specified host for computing and runs in distributed mode.

```
%let GRID_TEXTANALYTICS_BIN_LOC=&GRIDTXTBIN;

option set=GRIDHOST="&GRIDHOST";
option set=GRIDINSTALLLOC="&GRIDINSTALLLOC";

proc hptmine data=getstart;
doc_id      did;
variable    text;
parse
            outterms  = terms
            reducef   = 2;
svd
            k         = 10
            outdocpro = docpro;
performance nodes=2 details;
run;
```

In the preceding statements, the macro variable GRID_TEXTANALYTICS_BIN_LOC specifies the location on the grid where the language binary files are installed. These files are used by the HPTMINE procedure in parsing text.

Chapter 4
The HPTMSCORE Procedure

Contents

Overview: HPTMSCORE Procedure

For high-performance text mining, scoring is the process of applying the parsing and SVD projections to new textual data. The HPTMSCORE procedure performs this scoring of new documents, and its primary output is the document data set Outdocpro, which holds the reduced dimensional representation of the score collection. PROC HPTMSCORE uses some of the output data sets of the HPTMINE procedure as input data to ensure consistency between scoring and training. During scoring, the new textual data must be parsed using the same settings that the training data were parsed with, indexed using only the subset of terms that were used during training, and projected onto the reduced dimensional subspace of the singular value decomposition that was derived from the training data. To facilitate this process, you specify the CONFIG=, TERMS=, and SVDU= options in PROC HPTMINE to create three data sets, Outconfig, Outterms, and Svdu, respectively, and then you specify those three data sets as inputs to PROC HPTMSCORE. For more information about these data sets, see their respective options in the section "PROC HPTMSCORE Statement" on page 84.

PROC HPTMSCORE runs in either single-machine mode or distributed mode.

NOTE: Distributed mode requires SAS High-Performance Text Miner.

PROC HPTMSCORE Features

The HPTMSCORE procedure processes large-scale textual data in parallel to achieve efficiency and scalability. The following list summarizes the basic features of PROC HPTMSCORE:

- Functionalities that are related to document parsing, term-by-document matrix creation, and dimension reduction are integrated into one procedure to process data more efficiently.
- Parsing and term-by-document matrix creation are performed in parallel.
- Computation of document projection is performed in parallel.
- Analysis can be performed on a massively parallel SAS high-performance appliance.
- All phases of processing make use of a high degree of multithreading.
- Input data can be read in parallel when the data source is the appliance database.

PROC HPTMSCORE Contrasted with Other SAS Procedures

The high-performance PROC HPTMSCORE functionality replaces what has in the past been done by using the DATA step function TGSCORE followed by a call to the SPSVD procedure. The TGSCORE function was used to parse each document and to append each document's term counts to an Outparent data set. Then PROC SPSVD applied the document projections. PROC HPTMSCORE combines these tasks into a single parallel procedure. For PROC HPTMSCORE, you provide the necessary inputs to perform the parsing, term-by-document table creation, and document projections in a single call. The primary output data set of PROC HPTMSCORE contains the document projections and is specified in the SVDDOCPRO= option.

Getting Started: HPTMSCORE Procedure

The following DATA steps generate two data sets. The getstart data set contains 36 observations, and the getstart_score data set contains 31 observations. Both data sets have two variables. The text variable contains the input documents, and the did variable contains the ID of the documents. Each row in the data sets represents a "document" for analysis.

```
data getstart;
    infile cards delimiter='|' missover;
    length text $150;
    input text$ did$;
    cards;
        High-performance analytics hold the key to |d01
        unlocking the unprecedented business value of big data.|d02
        Organizations looking for optimal ways to gain insights|d03
        from big data in shorter reporting windows are turning to SAS.|d04
        As the gold-standard leader in business analytics |d05
```

```
        for more than 36 years,|d06
        SAS frees enterprises from the limitations of |d07
        traditional computing and enables them |d08
        to draw instant benefits from big data.|d09
        Faster Time to Insight.|d10
        From banking to retail to health care to insurance, |d11
        SAS is helping industries glean insights from data |d12
        that once took days or weeks in just hours, minutes or seconds.|d13
        It's all about getting to and analyzing relevant data faster.|d14
        Revealing previously unseen patterns, sentiments and relationships.|d15
        Identifying unknown risks.|d16
        And speeding the time to insights.|d17
        High-Performance Analytics from SAS Combining industry-leading |d18
        analytics software with high-performance computing technologies|d19
        produces fast and precise answers to unsolvable problems|d20
        and enables our customers to gain greater competitive advantage.|d21
        SAS In-Memory Analytics eliminate the need for disk-based processing|d22
        allowing for much faster analysis.|d23
        SAS In-Database executes analytic logic into the database itself |d24
        for improved agility and governance.|d25
        SAS Grid Computing creates a centrally managed,|d26
        shared environment for processing large jobs|d27
        and supporting a growing number of users efficiently.|d28
        Together, the components of this integrated, |d29
        supercharged platform are changing the decision-making landscape|d30
        and redefining how the world solves big data business problems.|d31
        Big data is a popular term used to describe the exponential growth,|d32
        availability and use of information,|d33
        both structured and unstructured.|d34
        Much has been written on the big data trend and how it can |d35
        serve as the basis for innovation, differentiation and growth.|d36
run;

data getstart_score;
    infile cards delimiter='|' missover;
    length text $150;
    input text$ did$;
    cards;
        Big data according to SAS|d1
        At SAS, we consider two other dimensions|d2
        when thinking about big data:|d3
        Variability. In addition to the|d4
        increasing velocities and varieties of data, data|d5
        flows can be highly inconsistent with periodic peaks.|d6
        Is something big trending in the social media?|d7
        Perhaps there is a high-profile IPO looming.|d8
        Maybe swimming with pigs in the Bahamas is suddenly|d9
        the must-do vacation activity. Daily, seasonal and|d10
        event-triggered peak data loads can be challenging|d11
        to manage - especially with social media involved.|d12
        Complexity. When you deal with huge volumes of data,|d13
        it comes from multiple sources. It is quite an|d14
        undertaking to link, match, cleanse and|d15
        transform data across systems. However,|d16
```

```
            it is necessary to connect and correlate|d17
            relationships, hierarchies and multiple data|d18
            linkages or your data can quickly spiral out of|d19
            control. Data governance can help you determine|d20
            how disparate data relates to common definitions|d21
            and how to systematically integrate structured|d22
            and unstructured data assets to produce|d23
            high-quality information that is useful,|d24
            appropriate and up-to-date.|d25
            Ultimately, regardless of the factors involved,|d26
            we believe that the term big data is relative|d27
            it applies (per Gartner's assessment)|d28
            whenever an organization's ability|d29
            to handle, store and analyze data|d30
            exceeds its current capacity.|d31
   run;
```

The following statements use PROC HPTMINE for processing the input text data set getstart and create three data sets, outconfig, outterms, and svdu, which can be used in PROC HPTMSCORE for scoring. The statements then use PROC HPTMSCORE to score the input text data set getstart_score. The statements take the three data sets that are generated by PROC HPTMINE as input and create a SAS data set named docpro, which contains the projection of the documents in the input data set getstart_score.

```
proc hptmine data      = getstart;
doc_id      did;
variable    text;
parse
            outterms  = outterms
            outconfig = outconfig
            reducef   = 2;
svd
            k         = 10
            svdu      = svdu;
performance details;
run;

proc hptmscore
            data      = getstart_score
            terms     = outterms
            config    = outconfig
            svdu      = svdu
            svddocpro = docpro;
doc_id      did;
variable    text;
performance details;
run;
```

The output from this analysis is presented in Figure 4.1 through Figure 4.3.

Figure 4.1 shows the “Performance Information” table, which indicates that PROC HPTMSCORE executes in single-machine mode. That is, PROC HPTMSCORE runs on the machine where the SAS system is running. The table also shows that two threads are used for computing.

Figure 4.1 Performance Information

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Figure 4.2 shows the "Procedure Task Timing" table, which provides details about how much time is used by each processing step.

Figure 4.2 Procedure Task Timing

Procedure Task Timing

Task	Time (sec.)	
Parse documents	0.71	96.7%
Generate term-document matrix	0.01	1.65%
Compute SVD	0.01	1.34%
Output OUTDOCPRO table	0.00	0.30%

The following statements use PROC PRINT to show the content of the first 10 rows of the docpro data set that is generated by the HPTMSCORE procedure.

```
proc print data = docpro (obs=10) round; run;
```

Figure 4.3 shows the output of PROC PRINT. For more information about the output of the OUTDOCPRO= option, see the description of the SVDDOCPRO= option.

Figure 4.3 The DOCPRO Data Set

Obs	did	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8	COL9	COL10
1	d1	0.90	0.09	-0.01	0.37	0.06	-0.13	-0.14	-0.03	-0.03	-0.04
2	d2	0.29	-0.84	0.14	0.36	0.12	0.10	0.11	-0.10	-0.09	0.08
3	d3	0.82	0.36	0.20	0.27	0.08	-0.17	-0.06	-0.12	0.05	-0.18
4	d4	0.50	-0.38	0.13	-0.21	0.00	-0.27	-0.36	0.47	0.14	0.32
5	d5	0.69	-0.12	0.11	-0.01	0.12	-0.04	0.21	-0.34	-0.54	0.19
6	d6	0.27	-0.62	0.01	0.01	0.32	-0.28	0.54	-0.10	0.07	-0.24
7	d7	0.67	-0.48	0.26	-0.02	0.14	-0.37	0.21	0.14	0.16	-0.04
8	d8	0.35	-0.10	-0.26	0.16	-0.21	-0.49	0.64	-0.25	-0.11	0.00
9	d9	0.49	-0.64	0.18	-0.07	0.13	-0.39	0.26	0.19	0.19	0.04
10	d10	0.36	-0.68	0.20	-0.32	0.27	0.05	0.33	-0.24	-0.06	-0.15

Syntax: HPTMSCORE Procedure

The following statements are available in the HPTMSCORE procedure:

PROC HPTMSCORE < *options* > ;
  VARIABLE *variable* ;
  DOC_ID *variable* ;
  PERFORMANCE *performance-options* ;

PROC HPTMSCORE Statement

PROC HPTMSCORE < *options* > ;

The PROC HPTMSCORE statement invokes the procedure. Table 4.1 summarizes the *options* in the statement by function. The *options* are then described fully in alphabetical order.

Table 4.1 PROC HPTMSCORE Statement Options

option	**Description**
Basic Options	
DATA | DOC=	Specifies the input document data set
TERMS=	Specifies the data set that contains the summary information about the terms that are to be used for scoring
CONFIG=	Specifies the data set that contains the option settings that are to be used for scoring
SVDU=	Specifies the data set that contains the U matrix whose columns are the left singular vectors
Output Options	
NOPRINT	Suppresses ODS output
OUTPARENT=	Specifies the term-by-document frequency matrix that is used to model the document collection, in which the child terms are not represented and child terms' frequencies are attributed to their corresponding parents
SVDDOCPRO=	Specifies the projections of the documents

You can specify the following *options* in the PROC HPTMSCORE statement.

CONFIG=*SAS-data-set*

specifies the input SAS data set that contains configuration information for PROC HPTMSCORE. The data set that is used for this option should be the one that is generated by the OUTCONFIG option in the PARSE statement of the HPTMINE procedure during training. For more information about this data set, see the section "The OUTCONFIG= Data Set" on page 57 of Chapter 3, "The HPTMINE Procedure."

DATA | DOC=*SAS-data-set*

specifies the input SAS data set of documents to be used by PROC HPTMSCORE. The default is the most recently created data set. If PROC HPTMSCORE executes in distributed mode, the input data are distributed to memory on the appliance nodes and analyzed in parallel, unless the data are already distributed in the appliance database. When data are already distributed, PROC HPTMSCORE reads the data alongside the distributed database. For more information, see the sections "Processing Modes" on page 6 and "Alongside-the-Database Execution" on page 13. Each row of the input data set must contain one text field and one ID field, which correspond to the text and the unique ID of a document, respectively.

NOPRINT

suppresses the generation of ODS output.

OUTPARENT=*SAS-data-set*

specifies the output data set to contain a compressed representation of the sparse term-by-document frequency matrix. The data set contains only the kept, representative terms, and the child frequencies are attributed to the corresponding parent. For more information about the compressed representation of the sparse term-by-document frequency matrix, see the section "The OUTPARENT= Data Set" on page 58 of Chapter 3, "The HPTMINE Procedure."

SVDDOCPRO=*SAS-data-set* **<KEEPVARS | KEEPVARIABLES=***variable-list***>**

specifies the output data set to contain the reduced dimensional projections for each document. The contents of this data set are formed by multiplying the term-by-document frequency matrix by the input data set that is specified in the SVDU= option and then normalizing the result.

The variables in the data set that is specified by the DATA= option can be copied to the output of this option. When KEEPVARS | KEEPVARIABLES=*variable-list* is used, the content of the variables that is specified in the *variable-list* is attached to the output. These variables must appear in the data set that is specified by the DATA= option.

SVDU=*SAS-data-set*

specifies the input data set that contains the **U** matrix, which is created during training by PROC HPTMINE. The data set contains the information that is needed to project each document into the reduced dimensional space. For more information about the contents of this data set, see the SVDU= option in Chapter 3, "The HPTMINE Procedure."

TERMS=*SAS-data-set*

specifies the input SAS data set of terms to be used by PROC HPTMSCORE. The data set that is used for this option should be the one that is generated by the OUTTERMS= option in the PARSE statement of the HPTMINE procedure during training. This data set conveys to PROC HPTMSCORE which terms should be used in the analysis and whether they should be mapped to a parent. The data set also assigns them a key that corresponds to the key that is used in the input data set that is specified by the SVDU= option. For more information about this data set, see the section "The OUTTERMS= Data Set" on page 59 of Chapter 3, "The HPTMINE Procedure."

DOC_ID Statement

DOC_ID *variable* < *docid-options* > **;**

This statement specifies the variable that contains the ID of each document. In the input data set, each row corresponds to one document. The ID of each document must be unique; it can be either a number or a string of characters.

You can specify the following *docid-option* in the DOC_ID statement:

DUPLICATIONCHECK | DUPCHK < (WARNING | STOP) >

checks whether the ID of each document is unique. When this option is not specified, no duplication check is performed. When you specify this option along with the WARNING keyword and duplicate document IDs are detected, the HPTMSCORE procedure outputs a warning message, which shows the number of duplicate document IDs that have been detected. When you specify this option along with the STOP keyword, the HPTMSCORE procedure outputs an error message and terminates. The error message shows the number of duplicate document IDs that have been detected. If you specify the DUPLICATIONCHECK option without the WARNING or STOP keyword, the HPTMSCORE procedure takes the WARNING keyword by default.

PERFORMANCE Statement

PERFORMANCE < *performance-options* > **;**

The PERFORMANCE statement defines performance parameters for multithreaded and distributed computing, passes variables that describe the distributed computing environment, and requests detailed results about the performance characteristics of the HPTMSCORE procedure.

You can also use the PERFORMANCE statement to control whether the HPTMSCORE procedure executes in single-machine or distributed mode.

The PERFORMANCE statement for is documented further in the section "PERFORMANCE Statement" on page 34 of Chapter 2, "Shared Concepts and Topics."

VARIABLE Statement

VARIABLE | VAR *variable* **;**

This statement specifies the variable that contains the text to be processed.

Details: HPTMSCORE Procedure

For information about the techniques that are used for nature language processing, term processing, and singular value decomposition, see the section "Details: HPTMINE Procedure" on page 52 of Chapter 3, "The HPTMINE Procedure."

Displayed Output

The following sections describe the output that PROC HPTMSCORE produces. The output is organized into various tables, which are discussed in their order of appearance.

Performance Information

The "Performance Information" table is produced by default. It displays information about the execution mode. For single-machine mode, the table displays the number of threads used. For distributed mode, the table displays the grid mode (symmetric or asymmetric), the number of compute nodes, and the number of threads per node. If you specify the DETAILS option in the PERFORMANCE statement, the procedure also produces a "Timing" table in which elapsed times (absolute and relative) for the main tasks of the procedure are displayed.

Procedure Task Timing

When the DETAILS option is specified in the PERFORMANCE statement, the procedure produces a "Procedure Task Timing" table, which displays the elapsed time (absolute and relative) for the main tasks.

ODS Table Names

Each table that is created by the HPTMSCORE procedure has a name associated with it, and you must use this name to refer to the table when you use ODS statements. These names are listed in Table 4.2.

Table 4.2 ODS Tables Produced by PROC HPTMINE

Table Name	Description	Required Statement / Option
PerformanceInfo	Information about the high-performance computing environment	Default output
Timing	Absolute and relative times for tasks that are performed by the procedure	DETAILS option in the PERFORMANCE statement

System Configuration

Prerequisites for Running PROC HPTMSCORE

To use the HPTMSCORE procedure, you must have a valid SAS Text Miner license, and the language binary files that are provided under that license must be available to PROC HPTMSCORE for parsing text.

Input Data Accessibility

When you run PROC HPTMSCORE in distributed mode, the following input data sets need to be accessible to the client machine:

- TERMS= data set
- CONFIG= data set
- SVDU= data set

When the HPTMSCORE procedure runs in distributed mode, it first reads these data sets on the client machine and then sends them to the grid.

Configuration for Distributed Mode

When PROC HPTMSCORE runs in distributed mode, it needs to locate on the grid the language binary files that are used in parsing text. These binary files must be deployed to the grid. In addition, the GRID_TEXTANALYTICS_BIN_LOC macro must be specified to indicate to location of the binary files on your grid. For more information, see the section "Configuration for Distributed Mode" on page 62 of Chapter 3, "The HPTMINE Procedure."

Subject Index

Syntax Index

Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to **yourturn@sas.com**. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to **suggest@sas.com**.