



THE
POWER
TO KNOW.

SAS[®] 9.4 Data Management

Overview

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2016. *SAS® 9.4 Data Management: Overview*. Cary, NC: SAS Institute Inc.

SAS® 9.4 Data Management: Overview

Copyright © 2016, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

For a hard-copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a) and DFAR 227.7202-4 and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513-2414.

April 2016

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

For additional information, see the Legal Notices appendix.

Contents

<i>Accessibility Notice</i>	v
Chapter 1 • Introduction	1
A Data Management Scenario	1
Challenges	2
Opportunities	2
Plan	4
Act	5
Monitor	8
Chapter 2 • Data Access	11
Overview	11
SAS/ACCESS Interfaces	12
SAS/ACCESS Interface to Hadoop and SAS/ ACCESS Interface to Impala	13
SAS Federation Server	14
SAS Event Stream Processing	16
Other Applications That Connect to Data	16
Chapter 3 • Data Quality	23
Overview	23
SAS Quality Knowledge Base	24
DataFlux Data Management Studio	26
SAS Master Data Management	45
Other Applications with Data Quality Functions	48
Chapter 4 • Data Integration	53
Overview	53
SAS Data Loader for Hadoop	54
SAS Data Integration Studio	59
SAS Visual Process Orchestration	62
DataFlux Data Management Studio	63

DataFlux Data Management Server	64
Chapter 5 • Data Governance	65
Overview	65
Shared Metadata	66
DataFlux Web Studio	66
SAS Business Data Network	69
SAS Lineage	71
SAS Data Remediation	74
Chapter 6 • Architecture	77
Overview	78
Data Access	79
Desktop Clients	80
Web Tier	80
Server Tier	81
Appendix 1 • Legal Notices	85
Recommended Reading	95

Accessibility

Accessibility Notice

For information about the accessibility of any of the products mentioned in this document, see the usage documentation for that product.

1

Introduction

<i>A Data Management Scenario</i>	1
<i>Challenges</i>	2
<i>Opportunities</i>	2
<i>Plan</i>	4
<i>Act</i>	5
<i>Monitor</i>	8

A Data Management Scenario

SAS data management applications work together to solve complicated business problems. For example, suppose that a national grocery store chain in the United States named Groceryrama buys a smaller regional grocery store chain named GreenVillage. What can SAS software do to help the data systems of Groceryrama and GreenVillage work well together?

Challenges

Such a large purchase brings problems along with its opportunities. For example, Groceryrama and GreenVillage could store their data in different data management systems. Their data warehouses could be organized in incompatible ways. Key data fields, such as address, pricing, and personnel data could be stored in formats that conflict. The two chains could also use different systems for reviewing and reporting on their data.

Opportunities

What can SAS data management offerings do to help bring order to this chaos?

Software from SAS comes in offerings, which provide specific functions to help specific groups of users. These offerings generally include SAS/ACCESS interfaces that enable data architects and administrators to efficiently load data from a variety of sources including relational databases, personal computer files, and Hadoop.

The Data Management Advanced offering also includes integration and data quality tools such SAS Data Integration Studio and DataFlux Data Management Studio that aid administrators and architects in cleansing data and organizing it for better performance.

SAS data governance applications such as SAS Lineage and SAS Business Data Network helps business users and managers see and understand data more clearly.

A good way to understand how to use the SAS data management applications is to look at them as tools using the SAS Data Management methodology. This methodology divides your data management engagement into planning, action, and monitoring stages. The following stages are included:

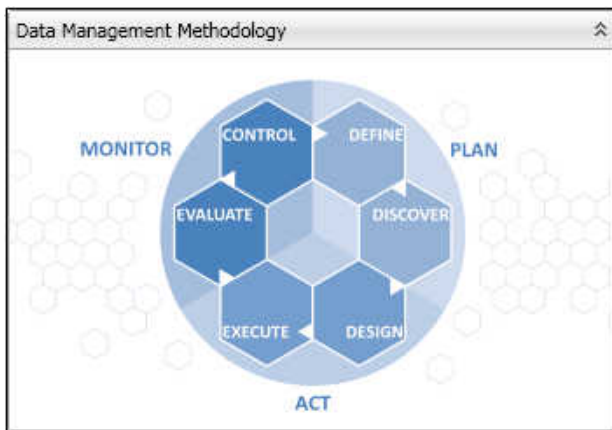
- [“Plan” on page 4](#)
- [“Act” on page 5](#)

- “Monitor” on page 8

The SAS Data Management methodology is a step-by-step process for performing data management tasks such as data quality, data integration, data migrations, and master data management. Within this methodology, the Plan stage contains Define and Discover phases. Similarly, the Act stage includes Design and Execute phases, and the Monitor stage contains Evaluate and Control phases.

The following display shows the parts of the SAS Data Management Methodology:

Figure 1.1 SAS Data Management Methodology



Your organization benefits when it plans, takes action on, and monitors data management projects. This methodology helps you build the foundation to optimize revenue, control costs, and mitigate risks. Moreover, all of these stages interact. For example, you can revisit your initial plans and operational designs when you evaluate your progress in the Monitoring stage.

The remainder of this introduction shows how the Groceryrama and GreenVillage staff use the SAS data management applications to work together through a data management project.

Plan

The Plan phase gives you the chance to define of the people, processes, and technologies that are used for your data management project. It also gives you time to discover and categorize your data assets. Groceryrama and GreenVillage staff include data scientists, data architects, business users, system administrators, and managers. They need to meet, discuss their data needs, and discover solutions to their problems. These meetings can begin addressing a series of questions that help define the parameters of the project. These questions include the following:

- **People:** Who is involved? And for what purpose?
- **Roadmap:** Where are we now? Where do we want to go? What obstacles are in our way?
- **Source systems:** What data do we need? Where is that data coming from?
- **Business processes:** Which business processes are affected? How can better data enhance how the organization operates?
- **Business rules and data definitions:** How do we define “customer?” How do we want to optimize procurement and spend?

The answers to these questions guide the collection, organization, enhancement, monitoring, and retirement of your data assets throughout the process. While you do not need all the answers at the beginning, you need a solid plan about how to proceed and what the ultimate success indicators will be. These discussions help identify the business rules and data definitions that guide the data management project. For example, you need clear guidance on the reason for the project (such as to cut costs, mitigate risks, and enhance revenue).

During the discovery portion of planning, Groceryrama and GreenVillage business analysts and data architects might run the following processes:

- **Data exploration:** This diagnostic phase is concerned with documenting the data in your organization and the characteristics of that data.

- **Data profiling and auditing:** Data profiling alerts you to data that does not match the characteristics defined in the metadata compiled during data exploration.
- **Data cataloging and business vocabulary:** You need a development environment where data sources can be combined and rationalized.

Most of the processes in the planning stage are supported by tools in DataFlux Data Management Studio. Data explorations enable you and your organization to identify data redundancies and extract and organize metadata from multiple sources. Then, the Groceryrama and GreenVillage team can use the profiling tools to dig deeper and identify data management issues and plan and scope data quality processes appropriately. Profiling is one of the SAS applications that uses the SAS Quality Knowledge Base (QKB). The QKB provides a set of files that contains rules, expressions, and reference data that are combined to analyze and transform text data in various SAS products. Finally, data collections are provided as a means to select data fields in different tables of different data connections. A collection provides a convenient way for you to build up a data set using those fields.

Data cataloging lays the groundwork for all data management tasks to follow. Data catalogs must be augmented with business definitions and vocabularies, allowing the business user to comfortably navigate the landscape. SAS Business Data Network enables you to manage these business terms. You can set up workflows and establish relationships between terms and processes. These tasks promote a common understanding of the key concepts and practices used in an enterprise.

Planning comes first. However, you might want to refer back to your plan as you move forward and adjust it as you learn more about the data needs of Groceryrama and GreenVillage.

Act

The team at Groceryrama and GreenVillage can begin to act by designing a system for dealing with their data needs and executing the processes defined in that system. Start the design phase by taking stock of all of these different structures, formats, data sources, and data feeds in the define and discover phases of the data management methodology. Then you can create an environment that accommodates the needs of

your business. In the design phase, you consolidate and coordinate your data management activities by concentrating on the following imperatives:

- **Consistency of rules:** Ultimately, an organization needs one set of business rules that can be stored centrally but deployed across all data sources, applications, and lines of business.
- **Consistency of the data model:** The data model is the single, definitive source for how your data maps to your business. Through the process of creating a well-structured data model, you identify the appropriate source systems and begin to reconcile multiple views, if required.
- **Consistency of business processes:** During the planning stage, you identify processes that are potentially impacted. Now, the task is to provide consistency across these processes.

This is the time to gather teams of business analysts, data architects, and IT specialists and begin to make practical decisions about how the data will be organized and regulated.

For example, you need to make sure that the content of the data files works together. Do your customer names have same format? Or do some sources put the surname first and others put it last? Are all of dates in the same format? Do your product lists contain duplicated records? These questions and other like them can be addressed with the data quality and master data management functions in applications such as DataFlux Data Management Studio and SAS Master Data Management.

Another common data problem is choosing among similar, but not identical, rows of data. For example, you might have records in the supplier information tables for both of the merging entities that appear to refer to the same supplier. The supplier records for Acme Pork Products for one company spell out the name of the state where the supplier is located and include a postal code. However, the supplier records for the other company use a two-letter state code and omit the postal code. Which supplier records should the merged Groceryrama and GreenVillage enterprise use?

You can use the entity resolution tools in SAS Master Data Management to diagnose the extent of the problem. Then you can designate one survivor record for suppliers that you can use throughout the enterprise. Business users have established how the data

and rules should be defined. The IT staff can now ensure that databases and applications comply with the definitions.

SAS Data Integration Studio is an Enterprise extract, transform, load (ETL) application that gives a big productivity boost to SAS coders doing data preparation and data management. It contains several features that help you get your data working together. First, you can register your data as metadata and group it into libraries of source and target tables. Then, you add the items in the libraries to job flows that enable you to perform the extract, transform, and load tasks at the core of data integration. Finally, you can deploy these jobs and schedule them for execution in batches. The application also supports related processes such as SQL queries, table loading, analytics, and reporting.

DataFlux Data Management Studio support data job flows and process job flows to improve data quality. DataFlux Data Management Studio is designed to work with DataFlux Data Management Server. Any authorized user can review and work with the jobs on the server. SAS Visual Process Orchestration adds the ability to integrate executable files from various systems into a single process flow. It enables you to build orchestration jobs, which are process jobs that run other jobs.

SAS Data Integration Studio and DataFlux Data Management Studio are often used by data management specialists. SAS Data Loader brings data management within reach of the general business user. SAS Data Loader simplifies the process of working with large distributed Hadoop data sources. Then it provides a series of wizard-based directives that help you perform tasks like transforming, profiling, and querying data in Hadoop

You can use SAS software to quickly and efficiently acquire data from a wide variety of data sources. For example, you can use SAS/ACCESS interfaces for critical sources such as the following:

- Oracle, Sybase, DB2, and Microsoft
- SQL Server, and Teradata databases
- Hadoop and Impala data
- data from enterprise resource planning applications such as SAS

Then, you can use the external file wizards in SAS Data Integration Studio to acquire data from fixed-width, delimited, and user-written external files. You can also use SAS Data Integration Studio to register metadata from all of your data sources into libraries that are used in jobs. You can work with all of this data in your SAS applications.

Monitor

A healthy data lifecycle for Groceryrama and GreenVillage requires a robust monitoring and reporting system. The data needs to be consistently monitored so that it remains fit-for-purpose for your organization. Why is this so critically important? You just spent time, energy, and resources to get your systems to a point where the business users have a consistent and validated view of your organization. Is it not time to just enjoy the success of all this effort?

Actually, the opposite is true. Very few organizations are static. They are forever growing and evolving. For example, you add new partners that bring new data to the table. Your business changes, sales regions are created or modified, you take on new initiatives, and you develop new products. All of these changes must be reflected in your data, which makes the evaluate phase so important.

Your mantra for success at this point needs to be as follows:

- **Monitor:** Data should be monitored and validated as it enters your organization to verify it is meeting your rules. Those rules need to be constantly monitored to ensure they are still meeting the needs of your business.
- **Review:** Efforts in discovery, design, and execution enable you to consolidate the rules and requirements into a single environment.
- **Optimize:** With the ability to centralize the required data management rules, the changes can be immediately propagated across the organization, without duplication of effort.

DataFlux Web Studio, SAS Business Data Network, SAS Visual Analytics and Reporting, and SAS Lineage are key applications in this stage. Broadly speaking, these applications help you keep track of your data management architecture and ensure that you and your coworkers share a common understanding of its component parts.

DataFlux Web Studio includes the DataFlux Monitor Viewer and Dashboard Viewer, which provide web interfaces for viewing exceptions to monitored business rules. It also contains the Reference Data Manager, which provides a web interface for creating and managing reference data. Some examples of reference data included are a list of valid values for a Gender field and a list of valid ZIP codes with their associated cities and states.

SAS Business Data Network enables you to manage business terms. You can set up workflows and establish relationships between terms and processes. These tasks promote a common understanding of the key concepts and practices used in an enterprise.

SAS Visual Analytics and Reporting provides visualization and reporting capabilities to help you visualize your data, level of data health, and remediation issues. You can also share information using the in-memory capabilities of the SAS LASR Analytic Server.

SAS Lineage lets you view your data and relationships between your data objects. It can use a dedicated SAS relationship service to view a wide variety of data objects including most types of SAS metadata and data taken from third-party sources. It can generate a network diagram that displays all available relationships or specialized impact analysis diagrams. It includes powerful search and filtering tools so that you can see exactly the data that you need.

You can use the processes in the monitor stage to test the efficacy of the work that you have done in the earlier stage. Then you can prepare for the additional work that you need to keep your data current and relevant. This stage requires an ongoing commitment to the quality and utility of the data management practices of your organization.

2

Data Access

<i>Overview</i>	11
<i>SAS/ACCESS Interfaces</i>	12
<i>SAS/ACCESS Interface to Hadoop and SAS/ACCESS Interface to Impala</i>	13
<i>SAS Federation Server</i>	14
<i>SAS Event Stream Processing</i>	16
<i>Other Applications That Connect to Data</i>	16
SAS Data Loader for Hadoop	16
SAS Data Integration Studio	17
DataFlux Data Management Studio	18

Overview

Data access in SAS data management includes the following technologies:

- “SAS/ACCESS Interfaces” on page 12
- “SAS/ACCESS Interface to Hadoop and SAS/ACCESS Interface to Impala” on page 13
- “SAS Federation Server” on page 14

- [“SAS Event Stream Processing” on page 16](#)

These access technologies help organizations that struggle with accessing and integrating diverse, large, or unstructured data. Data access includes the ability to read, write, and update tables or individual records in relational databases. Native text file importers are also provided for delimited, fixed width, and multi-length text files. Finally, the capacity to work with non-structured and semi-structured data such as XML, message queues, web services, and extracted data is included in the appropriate SAS data management applications. These data access capabilities work to speed the creation of reports and the preparation of data to run analytics for making better decisions.

Other applications such as SAS Data Loader for Hadoop, SAS Data Integration Studio, and DataFlux Data Management Studio (part of the SAS Data Quality bundle) use these technologies to support their data connections. See [“Other Applications That Connect to Data” on page 16](#) for more information.

SAS/ACCESS Interfaces

Data can be stored in a wide range of third-party databases, including the following:

- relational databases such as Oracle, Sybase, DB2, Microsoft SQL Server, and Teradata
- hierarchical databases such as IBM Information Management System (IMS)
- Computer Associates Integrated Database Management System (CA-IDMS), a network model database system
- Enterprise resource planning applications such as SAP

SAS/ACCESS interfaces provide fast, efficient loading of data to and from these facilities. With these interfaces, SAS software can work directly from the data sources without making a copy. Several SAS/ACCESS engines use an input/output (I/O) subsystem so that applications can read entire blocks of data instead of reading only one record at a time. This feature reduces I/O bottlenecks so that procedures can read data as quickly as they can process it. SAS/ACCESS engines for Oracle, Sybase, DB2

(on UNIX and PC), ODBC, Microsoft SQL Server, and Teradata support this functionality.

These engines, as well as the DB2 engine on z/OS, can also access database management system (DBMS) data in parallel by using multiple threads to the parallel DBMS server. You can get even greater performance gains by using threaded SAS procedures with these SAS/ACCESS ESS (Enterprise Systems Support) engines.

Some ESS engines also provide database-specific performance-tuning options and support features like bulk loading. Selected ESS engines include database pushdown capabilities such as Code, Scoring, and Data Quality accelerators. These pushdown features take advantage of database processing power by processing the data in place instead of moving it to the SAS environment.

SAS/ACCESS Interface to Hadoop and SAS/ACCESS Interface to Impala

Hadoop is general-purpose data storage and computing platform that includes database-like tools, such as Hive and HiveServer2. SAS/ACCESS Interface to Hadoop lets you work with your data using SQL constructs through Hive and HiveServer2. It also lets you access data directly from the underlying data storage layer, the Hadoop Distributed File System (HDFS). This differs from the traditional SAS/ACCESS engine behavior, which exclusively uses database SQL to read and write data.

Cloudera Impala is an open-source, massively parallel processing (MPP) query engine that runs natively on Apache Hadoop. You can use it to issue SQL queries to data stored in HDFS and Apache Hbase without moving or transforming data. Similar to other SAS/ACCESS engines, SAS/ACCESS Interface to Impala lets you run SAS procedures against data that is stored in Impala and returns the results to SAS.

You can read and write data to and from Hadoop as if they were any other relational data source to which SAS can connect with SAS/ACCESS Interface to Hadoop and SAS/ACCESS Interface to Impala . SAS/ACCESS Interface to Hadoop for Hive and Hive Server2 provides fast, efficient access to data stored in Hadoop through HiveQL. You can access Hive tables as if they were native SAS data sets and then analyze them using SAS.

You can also get virtual views of Hadoop data without physically moving it. SAS Federation Server offers simplified data access, administration, security, and performance by creating a virtual data layer without physically moving data. This frees business users from the complexities of the Hadoop environment. They can view data in Hadoop and virtually blend it with other database systems such as SAP HANA, IBM DB2, Oracle, or Teradata. Improved security and governance features, such as dynamic data masking, ensure that the right users have access to the right data.

The HADOOP procedure enables SAS to interact with Hadoop data by running Apache Hadoop code. Apache Hadoop is an open-source framework that is written in Java and provides distributed data storage and processing of large amounts of data.

The HADOOP procedure interfaces with the Hadoop JobTracker. This is the service within Hadoop that controls tasks to specific nodes in the cluster.

PROC HADOOP enables you to submit the following:

- Hadoop Distributed File System (HDFS) commands
- MapReduce programs
- Pig language code

SAS Federation Server

SAS Federation Server is a data server that provides scalable, threaded, multi-user, and standards-based data access technology in order to process and seamlessly integrate data from multiple data sources. The server acts as a hub that provides clients with data by accessing, managing, and sharing SAS data. It also supports several popular relational databases.

SAS Federation Server provides a secure, blended, and virtual view of your data without having to move it. Dynamic data masking, on-demand data quality, and in-memory caching help to make information technology staff more agile in responding to data provisioning requests from business users.

SAS Federation Server supports native access to Oracle, Teradata, SQLServer, or DB2. The Federation Server also supports the use of ODBC drivers not delivered by SAS. Federation Server administrators create DSNs on the Federation Server and provide access to users. Users can set up a Federation Server DSN in the Data Management Platform that gives them access to data without ever having to set up ODBC on their specific box.

SAS Federation Server enables powerful querying capabilities and improved data source management. With SAS Federation Server, you can efficiently unite data from many sources, without moving or copying the data.

SAS Federation Server provides the following data access capabilities:

- a central location for setup and maintenance of database connections.
- access to popular database systems, including DB2, Netezza, Oracle, SAP, SQL Server, PostgreSQL, Teradata, and Greenplum.
- ODBC and native drivers to connect to select data sources.
- threaded data access technology that enhances enterprise intelligence and analytical processes.
- multi-server services that enable multiple clients to access the same data concurrently.
- ability to reference data from disparate data sources with a single query, known as data federation. Also included is its own SQL syntax, Federated Query Language (FedSQL), to provide consistent functionality that is independent of the underlying data source.
- a data abstraction layer, providing the ability to present a consistent data model throughout the organization. This abstraction layer is created with FedSQL views.
- data access control with user permissions and data source security.
- federated view building application that creates dynamic views of heterogeneous data and is made available to other systems through ODBC, JDBC, or web services.
- support for data masking, caching, and in-view data quality transformations.
- table-level, row-level, and column-level data access controls.

SAS Event Stream Processing

SAS Event Stream Processing enables you to access and process structured and unstructured streaming data from sources such as sensors, operational systems, and devices. You can perform the following data management and data quality related operations while in stream:

- Ingest a wide variety of streaming inputs.
- Cleanse and transform flowing data.
- Apply business rules.
- Decide to discard or take action before data is even stored.
- Integrate with YARN on Hadoop.

Then you can perform the following analytical tasks while in stream:

- Analyze for patterns of interest.
- Score with advanced analytic models.
- Inject new analytic insights into streams.
- Perform machine learning in-stream.

Other Applications That Connect to Data

SAS Data Loader for Hadoop

SAS Data Loader for Hadoop uses the Copy Data to Hadoop, Import a File, and Copy Data from Hadoop directives. These directives enable you to move data from your files system or database management systems into and out of Hadoop. The Import a File

directive helps you import miscellaneous files from your file system into Hadoop as columnar tables. SAS Data Loader for Hadoop provides a point-and-click interface for moving, cleansing, and analyzing data in Hadoop. It enables business users and data scientists to do self-service data preparation on a Hadoop cluster.

SAS Data Integration Studio

SAS Data Integration Studio supports data access with the following components:

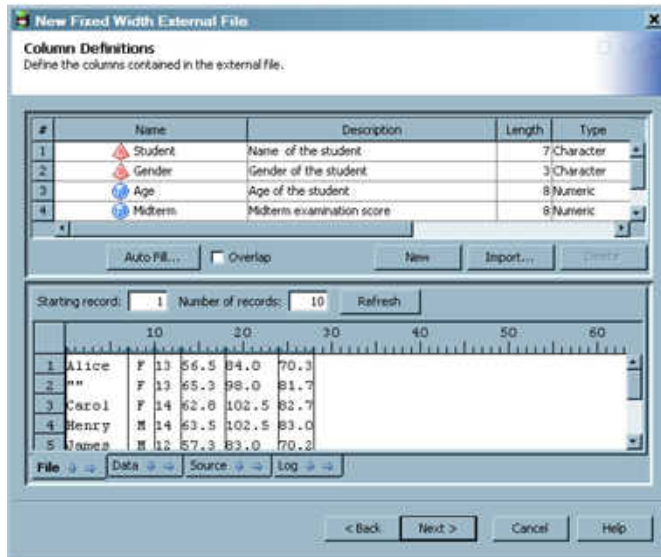
- external file tools
- tools for accessing web services, message queues, and XML data
- SAS table and relational database table tools
- high-performance data tools
- Hadoop and SAS LASR Analytic Server tools

SAS Data Integration Studio enables you to import and manage external files. An external file, sometimes called a flat file or a raw data file, is a plain text file that often contains one record per line. Within each record, the fields can have a fixed length or they can be separated by delimiters, such as commas. Like SAS or DBMS tables, external files can be used as inputs and outputs in SAS Data Integration Studio jobs.

Unlike SAS or DBMS tables, which are accessed with SAS LIBNAME engines, external files are accessed with SAS INFILE and FILE statements. Accordingly, external files have their own registration wizards for processing delimited external files, fixed-width external files, and external files with user-written code. These registration wizards enable you to create metadata for external files. The metadata is saved to a SAS Metadata Server, where SAS Data Integration Studio can access it.

The following display shows the Column Definitions page from the fixed-width external file wizard:

Figure 2.1 Fixed-Width Column Definitions



You can use the REST and SOAP transformations to access third-party web services. You can use the queue writer and queue reader transformations for both Microsoft and WebSphere message queues. Finally, you can use the XML Reader and XML Writer transformations to work with XML data. All of these transformations enable you to work with these data types in SAS Data Integration Studio jobs.

DataFlux Data Management Studio

DataFlux Data Management Studio provides connection definitions that you can use to connect to a wide variety of data sources.

The following supported connection types are supported:

- ODBC connections
- domain-enabled ODBC connections
- ODBC connections for Excel
- ODBC connections for Hadoop

- SAS data set connections
- Federation Server connections
- Custom SAP connections
- Custom SQLite connections
- SQL Queries
- XML files
- Java Message Queues
- documents from third-party applications such as Adobe Acrobat, Microsoft PowerPoint, and Microsoft Visio

Most types of data connections in DataFlux Data Management Studio are set with the ODBC Data Source Administrator window. You can scroll the list of data types to select the type of connection that you need. Then, you can use the tabs in the window to set options and parameters for the connection. You can see the current list in of the available databases in the “Supported Databases for Data Storage” section of the “Working with Databases” topic. This topic can be found in the *DataFlux Data Management Studio: User’s Guide* for your version of DataFlux Data Management Studio. The “Working with Databases” topic is located in the “Data Riser Bar” chapter.

You can create domain-enabled ODBC connections that reference the ODBC connection and an appropriate authentication server domain, which can prevent users from needing to constantly authenticate to the connection.

Domain-enabled ODBC connections are based on the following prerequisites:

- a standard ODBC DSN for the data source that you want to access
- an authentication domain, user, and login for this ODBC DSN

Note: Domain-enabled connections cannot use shared logins.

Specialized ODBC connections for Excel and Hadoop simplify the process of accessing these data types. The Excel process enables you to select the appropriate driver for your version of Excel and create an ODBC DSN to read named ranges in an Excel spreadsheet. The Hadoop process enables you to select the appropriate DataFlux

Apache Hive Wire Protocol driver or the DataFlux Impala Wire Protocol driver for your site. Then, you can click **System DSN** to create a connection to a data source that all users on the machine can access.

SAS data sets can be accessed through the SAS Data Set Connection window, which connects to a folder that contains one or more SAS data sets. The data is accessed directly on disk without mediation by a SAS Application Server. The SAS connection points to a folder on the file system that contains SAS data sets.

The host that executes the connection must be able to access the folder that contains the SAS data. For example, the DataFlux Data Management Studio host is a Windows host. If the SAS data sets are on a UNIX host, you need a networking protocol like SAMBA (SMB/CIFS). You also could use a network file system (NFS) that exposes the UNIX file system as a Windows directory.

These SAS data set connections can be configured in the SAS Data Set Connection window. For example, you can specify an access level and specify whether the data should be compressed. You can also specify options for features such as table locking and encryption. Finally, you can check the connection string to see whether the appropriate options encoding has been selected for a given connection.

If a SAS Federation Server is available on your site, you can use the Data riser to connect to that server and access the DSN connections that are managed by that server. The Federation Server Connection window enables you to specify a server and port for the connection. It also supports compression and credentials settings. You can also test the connection to the server.

You can add a user-defined connection to an SAP system. This connection could be used as the data source for the **SAP Remote Function Call** node, a data job node. SAP libraries (DLLs) must be installed on all computers where this custom connection is used.

You can also add a user-defined connection for an SQLite database file. For example, an SQLite connection can be used in the definition for an Address Update repository.

DataFlux Data Management Studio contains a set of data job nodes that enable you to use XML data and XML column data as inputs and outputs in data jobs. Similarly, it supports the Java Message Service (JMS). This Java API enables applications to create, send, receive, and read messages with data and process job reader and writer

nodes. Web services are supported with the **Web Service** and **HTTP Request** data job nodes. You can use the **Document Extraction** data job node to find information that you need to process that is not always found in traditional databases. For example, you might need to take data from a Microsoft Word file or an HTML file. Then you can convert it into a format that you can process in a DataFlux Data Management Studio job.

3

Data Quality

Overview	23
SAS Quality Knowledge Base	24
DataFlux Data Management Studio	26
Quality Nodes	26
Data Enrichment	32
Data Profiling	36
Data Exploration	40
Data Matching and Entity Resolution	41
SAS Master Data Management	45
Other Applications with Data Quality Functions	48
SAS Data Loader for Hadoop	48
SAS Data Quality Accelerator for Teradata	48
SAS Data Integration Studio	49

Overview

Data quality provides the processes and tools to assess, correct, monitor, and maintain data health in an organization. The data quality features included in SAS applications address data quality issues quickly and easily. These features perform functions that include profiling of records, entity resolution, data lineage, text extraction, data remediation, process orchestration, and job monitoring.

Data quality and enrichment in SAS data management includes the following applications:

- [“SAS Quality Knowledge Base” on page 24](#)
- [“DataFlux Data Management Studio” on page 26](#)
- [“SAS Master Data Management” on page 45](#)

SAS offers additional quality-focused applications. SAS In-Database Technologies for Teradata includes a Data Quality Accelerator that enables you to run Data Quality functions inside Teradata. SAS Data Loader for Hadoop enables business users and data scientists to profile and run data quality functions inside Hadoop for improved performance. SAS Data Integration Studio contains transformations that bring data quality functions to the application. For more information, see [“Other Applications with Data Quality Functions” on page 48](#).

SAS Quality Knowledge Base

SAS Quality Knowledge Base (QKB) is a collection of files that store data and logic that define data management operations such as parsing, standardization, and matching. SAS software products refer to the QKB when you perform data management operations such as data cleansing or address parsing.

There are two QKB products, QKB for Contact Information and QKB for Product Data. QKB for Contact Information supports the management of commonly used contact information for individuals and organizations, such as names, addresses, company names, and phone numbers. QKB for Product Data supports the management of common attributes related to products and services, such as dimensions, color, materials, packaging terms, and part numbers.

A QKB contains hundreds of thousands data points and rules that enable the computer to analyze and correct data like a human. This collection of data quality rules is shared across the entire SAS suite, supporting a “write once, use anywhere” data quality rules strategy. This same set of rules can be used in-stream, in-database, or in-memory.

You can modify and extend a QKB to cleanse literally any type of data simply by modifying or creating new pattern libraries. It can also be used to manage entities such as word vocabularies, phonetic match rules, and standardization rules. When you upgrade the QKB, the installation automatically identifies user-defined modifications and merges the modifications into the new release.

Each QKB supports and is licensed by a locale. The locale is organized by language and country (for example, English, United States; English, Canada; and French, Canada). SAS supports QKB locales for more than 40 language regions, including French, German, Italian, Russian, Chinese, and Polish. You can license support for one or more locales for each QKB for your enterprise. To process data that originates in specific locales, license those locales for the QKB that handles that type of data.

The data quality algorithms used in the QKB are completely tunable through an application called Customize that enables you to modify the QKBs used in your DataFlux Data Management Studio data flow jobs. Customize enables you to add or modify the following components:

- pattern rules
- word vocabularies
- standardization or recode rules
- phonetic match rules
- regular expression rules

You can use the customization interface to teach the SAS engine how to parse, match, and standardize content that includes product names, descriptions, numeric information, and more. The ability to customize enables you to create completely new data quality definitions to better meet your projects' needs. These new definitions that are added to the Quality Knowledge Base are instantaneously available across the entire SAS suite.

DataFlux Data Management Studio

Quality Nodes

Data jobs are the main way to process data in DataFlux Data Management Studio, which is available as a part of the SAS Data Quality bundle.

Overview

You can use the data quality job nodes in the application and the Quality Knowledge Base to analyze data that is specified in these data jobs.

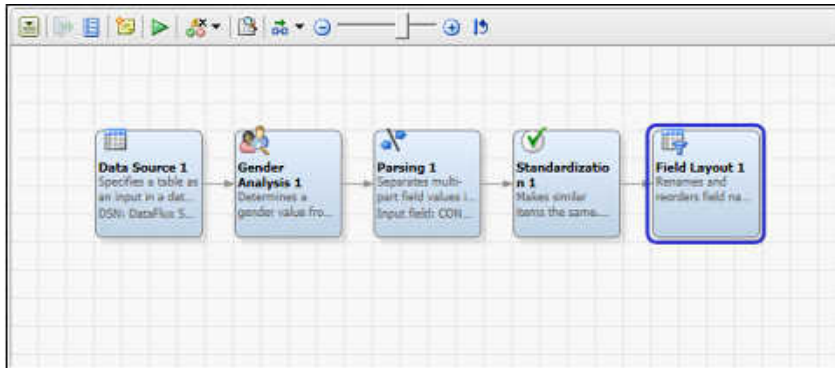
Some of the key DataFlux Data Management Studio data quality processes are addressed with the following data job nodes:

- [“Gender Analysis” on page 28](#)
- [“Identification Analysis” on page 28](#)
- [“Parsing” on page 28](#)
- [“Standardization” on page 31](#)
- [“Change Case” on page 31](#)
- [“Locale Guessing” on page 32](#)

DataFlux Data Management Studio also contains data job nodes for **Right Fielding**, **Create Scheme**, **Dynamic Scheme Application**, **Field Extraction**, and **Document Extraction**. Other data job and process job nodes support functions such as data flow into and out of jobs, data integration, and data monitoring. These nodes can be use to enhance jobs that include data quality nodes. For more information about these nodes, see the *DataFlux Data Management Studio User’s Guide* for your version of the application. You can also review the Data Access, Data Integration, and Data Governance sections of this overview.

The following display shows a data cleansing job in DataFlux Data Management Studio:

Figure 3.1 Data Cleansing Job



The output of the data cleansing job is shown in the following display:

Figure 3.2 Data Cleansing Job Output

	COMPANY	CONTACT	CONTACT_Gender	ADDRESS	STATE
1	First Merit Bank	James E. Briggs	M	19 East Broad Street	Missouri
2	DataFlux Corporation	Bob Brauer	M	6512 Six Forks Road - 404B	North Carolina
3	KAISER HOSPITAL	LUTHER BAKER	M	3560 E 116TH ST	CA
4	Transamerica Life Insurance	Irene Greaves	F	555 W Fifth St	OH
5	Transamerica Life Ins & Annuity Co	Rob Drain	M	7718 Elder Way	OH
6	Transamerica Occidental	Tonia Gerstner	F	1018 N Hayworth Ave	OH
7	Transamerica Life Insurance	Joshua Hodgekin	M	244 Evans	OH
8	Transamerica Occidental Life	Nancy Weinstock	F	2134 Estrado Cir	(null)
9	Transamerica Financial Services	Mary Little	F	13845 N 56th Pl	OH
10	Transamerica Occidental Life	Kate Lindemood	F	33032 Lighthouse Ct	OH
11	Transamerica Financial Group	Justin Echavarria	M	2662 E 2nd St	OH
12	Transamerica Financial Service	Olivia Bach	F	PO Box 8179	OH
13	Huntington Beach Union HS Dist	G. Weston	U	17272 Chapparral Ln	OH

Many of these nodes access information from an appropriate SAS Quality Knowledge Base (QKB). A QKB is a set of files that contain rules, expressions, and reference data that are combined to analyze and transform text data in various SAS products. You can modify these QKBs to create custom data quality logic for operations such as custom parsing or matching rules.

Gender Analysis

These nodes are designed to process data in ways that generate very specific pieces of information. For example, the **Gender Analysis** node determines a gender value from a list of names by using advanced word vocabularies and associated gender rules. The node parses names into name prefix, given name, middle name, and last name.

Then it scores the gender of each word to determine the actual gender, as shown in the following display:

Figure 3.3 Gender Analysis Output

Name	Pattern	Gender
Mr.·Mike·J·Smith	Male·Male·Unknown·Last·Name	Male
Chris·Smith·····	Unknown·Last·Name →	Unknown
Chris·Angela·Smith	Unknown·Female·Last·Name	Female

Identification Analysis

Identification analysis is the process of accurately flagging the type of data found within a field and bucketing the identified content into the appropriate field. Contact-related information is often used across contact fields such as name, address1, address2, and so on. You cannot assume that only names are found in a field simply because the field is named “CONTACT NAME”. This same example can be extended to address information, which is often used across many different address lines. SAS uses advanced word vocabularies and pattern libraries to correctly identify the type of data found in each field. After identifying the content, it moves the data to a new uniform field.

Parsing

SAS uses extensive pattern libraries and categorized lookup tables to intelligently break multi-value fields into parsed elements.

The SAS solution includes the following pre-defined parsing definitions:

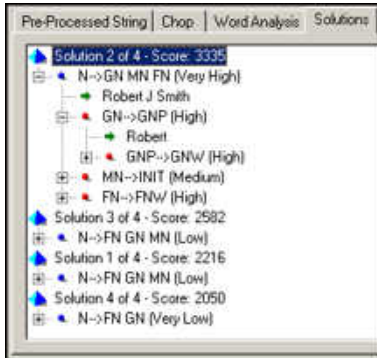
- Address
- City

- City, State/Province, Post Code
- Email
- Phone
- Website URLs
- Date
- Name
- Two names
- Account number
- Date time stamps
- Company names
- Phone numbers
- State/Province
- Post codes

For a given data type, the SAS engine uses customized pattern libraries that are tuned for the specific type of content being processed. These definitions include pattern rules, word vocabularies, automatic categorization, and other advanced processes that allow the engine to intelligently break data into pre-defined groups. After categorizing and identifying every word in a field, the engine uses a proprietary, patent-pending natural language processing engine to identify the best pattern match based on heuristic rules.

The following display represents one of the processes used to parse a name into multiple patterns:

Figure 3.4 Parsing a Name



In this case, the parsing engine identifies four potential solutions for breaking the name into first name, middle name, and last name. These solutions are (indicated by the “x of 4” text next to the score. The engine then selects the parse solution with the highest score and outputs the name tokens according to the defined pattern.

The example illustrates the complex logic applied during the parsing process. The first pattern, solution 2, matches the name to a known pattern [Given Name] [Middle Name] [Last Name]. Solution 3 matches the name to a different pattern [Family Name] [Given Name] [Middle Name]. In this case, the solution is given a lower score by the engine because the pattern has a lower priority than the previous pattern. Solution 1 and solution 4 are also scored as low-priority patterns. Therefore, the natural language engine decides that solution 2 is the best way to break apart the data (the “best solution”). Word categorization and pattern scores are completely user-tunable. If the selected pattern is not the best pattern, you can simply increase the pattern likelihood in order to modify the output.

This output is shown in the following display:

Figure 3.5 Parsing Output

Name	Name Prefix	First Name	Middle Name	Last Name
Mr Mike J Smith	Mr.	Mike	J	Smith
Chris Smith		Chris		Smith
Smith, Angela B and Chris		Angela	B	Smith
		Chris		Smith

Standardization

SAS supports advanced standardization routines that include element standardization, phrase standardization, and pattern standardization. Each standardization approach supports the ability to eliminate semantic differences found in source data. These differences include multiple spellings of the same information, multiple patterns of the same data, or the translation of inventory codes to product descriptions. For example, phrase standardization describes the process for modifying original source data from some value to a common value.

The company name that is rendered as “GM” or “Gen’l Motors” can be translated, based on standardization rules, from the original value to General Motors. The output is shown in the following display:

Figure 3.6 Standardization Output

Organization Name	Standardized Value
G.M. Auto	General Motors
Gen’l Motors	General Motors

Change Case

SAS change case functionality includes standard uppercase, lowercase, and proper case of words, as well as intelligent casing. In addition to normal proper casing rules,

intelligent casing includes the ability to recognize casing exceptions such as eBay, SAS, and others. Intelligent casing definitions exist for the following types of data:

- addresses
- city, state, or province
- name
- company name
- business title
- text

Locale Guessing

The **Locale Guessing** node compares with your data to guess the country (locale) to which your data applies. Locale identification supports the ability to automatically identify the locale associated with a name, address, and other types of information in order to optimize the matching, parsing, and standardization process. Name, address, and other types of data vary greatly based on the associated locale or country. Locale guessing ensures that the best country rules are used.

Data Enrichment

You can also use the nodes in the data jobs enrichment category to enrich, standardize, and augment the data that is specified in a data job.

Overview

The data jobs enrichment nodes support processes such as the following:

- [“Address Update” on page 34](#)
- [“Geocoding” on page 34](#)
- [“County Processing” on page 34](#)
- [“Phone Analysis and Enrichment” on page 35](#)
- [“Area Code Processing” on page 35](#)

Address Verification

The US address verification engine in DataFlux Data Management Studio has been certified by the United States Postal Service (USPS) and conforms to all of the CASS certification rules.

US address verification performs the following corrections in order to standardize and enrich the quality of any US address:

- spelling correction (street names and street elements)
- street type and directional append
- format standardization according to USPS rules

Additional address verification functions include the following:

Delivery Point Validation (DPV)

Delivery Point Validation validates the actual existence of a location that can receive mail. Although an address might be CASS certifiable, the actual physical address might not even exist. This can occur because the USPS reserves the address for a future building (apartment building, office building, or consumer residence).

Locatable Address Conversion Service (LACS)

Locatable Address Conversion Service (LACS) is a product that allows mailers to identify and convert a rural route address to a “city-style” address.

Residential Delivery Indicator (RDI)

The Residential Delivery Indicator engine can be used in conjunction with RDI data provided by the USPS. This indicator validates the type of address as a residential address or a non-residential address. RDI data is not distributed by SAS, but is available from the USPS for a modest annual fee. The RDI engine returns an RDI result code.

eLOT™

eLOT™ gives mailers the ability to sort their mailings in approximate carrier-casing sequence. eLOT™ processing can be used by mailers to qualify for enhanced carrier route presort discounts.

An address verification node is also available for Canada, and two nodes cover the rest of the world.

Address Update

DataFlux Data Management Studio also contains **Address Update** nodes in the job and process node **Enrichment** category. You can use these **Address Update** nodes to apply the functions included in the National Change of Address (NCOA) service to your data. This service makes address update information available to mailers, which helps reduce undeliverable mail.

US city, state, and ZIP code lookup enables lookup of either the city and state by the ZIP code or the ZIP code by city and state. If you look up the city and state by the ZIP code, you can return more than one city name for the supplied ZIP code. You can configure the number of possibilities that are returned.

Geocoding

SAS offers a **Geocode** node to support geocoding for any address within the US and Canada. Geocoding support includes the ability to identify the longitude and latitude of a specific address. It also enables the identification of census code information.

Address verification and geocoding support is available for most countries in the world. However, no census information is associated with this support.

For a given address, the geocoding engine returns the following:

- latitude
- longitude
- state/county/tract/block

There are two options available for geocoding: centroid of the postal code or rooftop-level geocoding.

County Processing

The SAS solution supports the ability to identify US county information from a Federal Information Processing Standard (FIPS) code. County enrichment is licensed along with the geocoding engine, and the reference file is updated quarterly.

The FIPS code can be generated by the geocode or phone analysis and can be used to identify the following information:

- US county name

- county seat (government capital)
- time zone offset from GMT (Greenwich Mean Time)
- type of county
- population
- area in square miles
- result

Phone Analysis and Enrichment

Phone analysis can be used to identify discriminatory attributes associated with a phone number, including items such as city, state, and MSA code. The phone enrichment library is updated quarterly and supports area code updates, splits, and much more.

Phone enrichment includes the following information (for US and Canadian phone numbers):

- country ISO name
- state
- city
- FIPS code assigned by US Census Bureau
- Metropolitan Statistical Area (MSA): closest major city
- primary and largest MSA (closest, largest major city)
- phone type (such as standard, cell, and beeper)
- area code
- overlay (additional valid area codes)

Area Code Processing

Area code enhancement supports the ability to determine the valid area code associated with a ZIP code. This enrichment, coupled with address validation works to ensure that the contact information is always current. The area codes reference library is updated quarterly along with the geocode library.

Data Profiling

You can also use DataFlux Data Management Studio to create and analyze data profiles. SAS data profiling provides a more granular, table-level view of the data's strengths and weaknesses. It supports the ability to connect to virtually any data source, including flat files, relational databases, and mainframe systems. Through an intuitive user interface, analysts can simply point the engine at a one or more sources and initiate the analysis. SAS profiling extracts the data from the source system and performs a three-phase profiling analysis composed of Data Discovery, Data Completeness, and Relationship Analysis. Profiling then presents the discovered statistics and potential anomalies to the analyst for rapid data analysis.

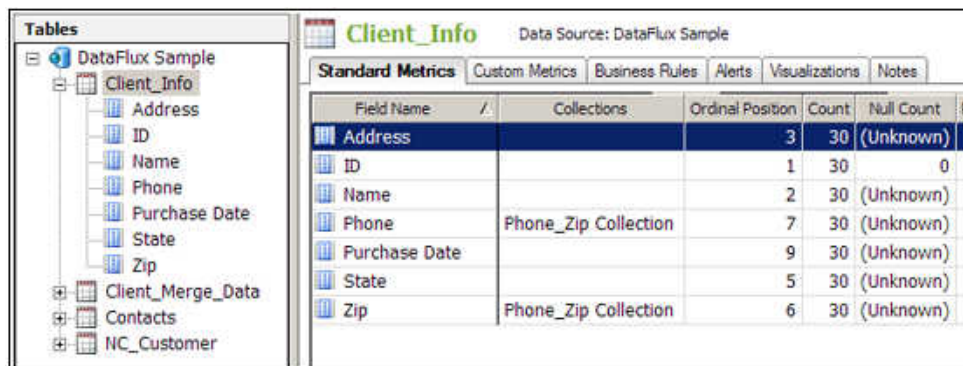
When you profile your data, you can perform the following tasks more efficiently:

- identifying issues early in the data management process, when they are easier and less expensive to manage
- obtaining more accurate project scopes and resource estimates
- better understanding existing databases and evaluating how well they might support potential marketing activities
- determining which steps you should take to address data problems
- making better business decisions about your data
- running periodic profiles to monitor your data and ensure that your data management processes are working well

Data profiling encompasses discovery and audit activities that help you assess the composition, organization, and quality of databases. Thus, a typical data profiling process helps you recognize patterns, identify scarcity in the data, and calculate frequency and basic statistics. Data profiling can also aid in identifying redundant data across tables and cross-column dependencies. All of these tasks are critical to optimal planning and monitoring.

Data profiling displays tabular output, as shown in the following display:

Figure 3.7 Tabular Output from Profiling



Field Name	Collections	Ordinal Position	Count	Null Count
Address		3	30	(Unknown)
ID		1	30	0
Name		2	30	(Unknown)
Phone	Phone_Zip Collection	7	30	(Unknown)
Purchase Date		9	30	(Unknown)
State		5	30	(Unknown)
Zip	Phone_Zip Collection	6	30	(Unknown)

The data profiling engine supports the following statistical analysis of source data:

Rule Validation

Apply advanced business rules to ensure information quality. Rules can include the following elements within rows, across rows, or across data sources: field comparison, mathematical calculations, Boolean tests, and data analysis.

Sampling

Perform full data source analysis or sample analysis based on a percentage of the data. The user has the ability to specify the sample interval.

Redundant Data Analysis (join tests)

Identify redundant information across tables or across data sources. SAS data profiling supports both exact field redundant analysis and fuzzy matching analysis. Fuzzy matching analysis applies domain-specific data quality match rules to the field before testing for joins. For example, if you need to identify potential overlap between two systems by company name, the SAS profiling engine can identify the overlap between *5th 3rd Bank Copr.* and *The Fifth Third Bank Corporation*. Redundant data analysis supports drill-down directly to the source record.

Key Relationship Analysis (primary and foreign key analysis)

Perform primary and foreign key analysis that ensures 100% correspondence across primary and foreign tables. You can also automatically identify orphaned records within a database or across databases.

SQL Query Virtual Table Creation

Profile the results of SQL queries and multiple table joins with full drill-down to the source record. You can either enter your own SQL queries or use the included SQL query builder.

Business Rules

Use profiling to take advantage of the business rules that have been defined by enabling users to apply them to a profile. This function enables you to identify issues early in the process. When you view the profile report, the violations are called out as alerts. These alerts identify areas that should get your immediate attention.

Extended Reports Options (such as text file, HTML, by column, and by metric)

Data profiling analysis can be exported to HTML, Microsoft Excel, or text files. Organizations can also write custom reports against the SAS profiling repository because it is not a closed repository.

Batch Scheduling

You can use third-party schedulers such as cron and Windows Scheduler to schedule profile jobs. These jobs can be run at a specified interval, such as hourly, daily, weekly, and monthly.

Open Repository

Statistics and information derived during the profiling phase are not stored in a closed repository. They can be stored in any ODBC-compliant database.

Data Trending

You can generate GUI-based control charts that graph the change in profiling statistics and custom metrics over time. Chart types include pie charts, line graphs, bar charts, and more. You can automatically invoke events when data violating user-defined rules is identified. These events include items such as a log exception record, an email user, write data to database table, and a log record to a text file. Time series profile metric collection enables you to view profiling statistics from any point in time during the monitoring lifecycle.

Metric calculations

SAS data profiling capabilities display standard statistical calculations, including those listed in the following list:

- Data Type

- Count
- Unique Count
- Null Count
- Non-Null Count
- Blank Count
- Pattern Count
- Minimum Value
- Maximum Value
- Maximum Length
- Mean
- Median
- Standard Error
- Mode
- Standard Deviation
- Custom Metrics (which create company-specific profiling statistics)
- Data Length
- Ordinal Position
- Primary Key Candidate
- Nullable
- Decimal Places
- Actual Type
- Uniqueness
- Minimum Length
- Outliers

- Percent Null
- Percentiles
- Pattern Distribution (which consists of a list of values with drill-down capability, filtering, and graphics)
- Frequency Distribution (which consists of a list of values with drill-down capability, filtering, and graphs)

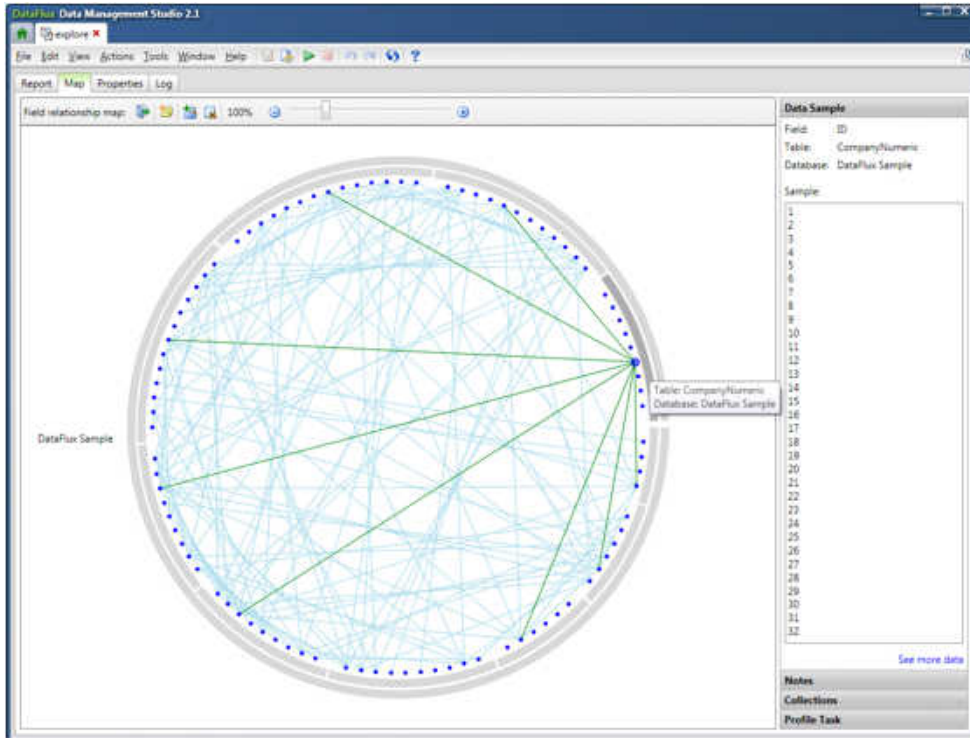
Data Exploration

You can use data explorations to identify data redundancies and extract and organize metadata from multiple sources. Relationships between metadata can be identified and cataloged into types of data by specified business data types and processes. A data exploration reads data from databases and categorizes the fields in the selected tables into categories. These categories have been predefined in the QKB. Data explorations perform this categorization by matching column names. You also have the option of sampling the data in the table to determine whether the data is one of the specific types of categories in the QKB.

For example, your customer metadata might be grouped into one catalog and your address metadata might be grouped in another catalog. Once you have organized your metadata into manageable chunks, you can identify relationships between the metadata by table-level profiling. Analyzing this data, relationships between tables can be identified and cataloged to help users identify areas that need further investigation. Data Exploration is a tool that helps you understand your data and your data problems. Creating a data exploration enables you to analyze tables within databases to locate potential matches and plan for the profiles that you need to run.

The following display shows a data exploration map:

Figure 3.8 Data Exploration Map



Each point contained in the circle created by the outer database and the inner table rings is a field in the data. Move the mouse over the point for a field to see the name of the field and the names of the table and database that contain it. You can also find the selected field in a report.

Once you have identified possible matches, you can plan the best way to handle your data and create a profile job for any database, table, or field. Thus, you can use a data exploration of your metadata to decide on the most efficient and profitable way to profile your physical data.

Data Matching and Entity Resolution

Entity resolution is the process of merging multiple files (or duplicate records within a single file). Once merged, the records referring to the same physical object are treated

as a single record. Records are matched based on the information that they have in common. The records that you can merge appear to be different but can actually refer to the same person or thing. Entity resolution, record matching, and surviving records are supported in DataFlux Data Management Studio. “[SAS Master Data Management](#)” on page 45 is a more powerful and specialized application that performs similar functions.

The SAS match engine has been designed to enable both the identification of duplicate records within a single data source. It also works across multiple sources. The rules-based matching engine uses a combination of parsing rules, standardization rules, phonetic matching, and token-based weighting to strip the ambiguity out of source information. After applying hundreds of thousands of rules to each and every field, the engine outputs a *match key*. This match key is an accurate representation of all versions of the same data generated at any point in time. A *sensitivity* level enables you to define the closeness of the match in order to support both high-confidence and low-confidence match sets.

As well as offering pre-defined match rules, SAS has designed an extremely customizable match engine that allows your organization to define custom matching rules. These match rules can include any number of fields, as well as any number of match conditions coupled together using Boolean rules (AND/OR). Matching records are then assigned a single group ID that can be persisted and maintained over time. Duplicate records are grouped based on linkage rules or automatically consolidated into a single *best record*.

Note that the SAS engine always generates the exact same match key for similar data generated at any point in time across the enterprise. SAS is the only data quality vendor that generates a single match key for an entity. These keys can be generated in real time. They can also be persisted to a data source to facilitate cross-system matching in batch or real-time environments.

The following display illustrates how the SAS matching engine is able to strip ambiguity from an address and generate a key for three distinct representations of the same address:

Figure 3.9 Address Matching

Address
100 N Mane St, Floor 12
100-12 North Main
#12, 100 No. Main Street

After eliminating pattern differences, the match engine applies a series of string manipulations to each token. Then, after applying hundreds of thousands of rules, it outputs a 15-character match key. The match key can be used to identify the address relationship across any number of disparate data sources. When other fields such as business name and postal code are combined in the match process, the records can be rationalized as a single entity.

Following this stage, the engine can be configured to perform one or all of the following tasks:

- displaying the matches in report format
- automatically consolidating the records into one best record
- appending grouping keys to each record and persist the keys over time
- writing match keys to an index or a cross-references table

The SAS engine does not require a change in the source or *presentation* data to match the records. The SAS match algorithm applies data parsing, standardization, and other algorithms during the match process. Some data quality vendors require adherence to a strict methodology: parse first, standardize second, and match third. The SAS engine applies all of these processes during the actual match process.

The SAS engine supports matching the types of information listed in the following list:

- Address

- City
- City, State/Province, Post Code
- Email
- Phone
- Website URLs
- Date
- Name
- Two names
- Account number
- Date time stamps
- Company names
- Business titles
- Phone numbers
- State/Province
- Postal codes
- Countries
- Text

Record consolidation, or duplicate elimination, merges multiple records into a single best record. The match engine supports user-defined *record-level* and *field-level* rules. These rules enable you to use the engine to pick and choose information from multiple records to compile a single version of the entity.

This record consolidation includes both basic and advanced rules such as the following:

- Field is not null or field is null
- Field is not equal to X or is equal to X
- Field has the highest occurring value within the duplicate record set

- Field contains the highest value within the duplicate record set
- Most or least recent create date
- Source is equal to *specified field* or string

The SAS engine supports persisted key clustering (or house holding). This type of clustering enables the assignment of a single unique key to any record that conforms to user-defined match rules. The engine uses sophisticated SAS match keys to group records and assign an integer-based unique identifier. If new records enter the system, the SAS clustering engine assigns the existing integer ID to the new record and then logs the record into the cluster table.

Sometimes, a new record enters the system that causes two or more unique households to collapse into one household. In that case, the SAS engine assigns the existing cluster ID to the records. Then it logs the old household ID in the grouping table. This engine supports the business need to track the lifetime activity of any entity from supplier to end consumer and can be run in both batch and real time.

For example, a record-level rule might call for the preservation of a record with the most recent edit or create date. However, this record might not include accurate address information. If the address exists in another record, field-level rules can be used to extract the address from the secondary record. Then the address in the primary record can be replaced with this trusted content.

SAS Master Data Management

Master data management (MDM) is the creation of a single, accurate and unified view of corporate data, integrating information from various data sources into one master record. This master data is then used to feed information back to the applications, creating a consistent view of data across the enterprise. MDM allows for consistent, reliable data to feed operational applications, helping end users access customer, patient, citizen, employee, product, asset, or location data.

SAS Master Data Management performs the following functions:

- extracts business information from your data sources

- validates and standardizes the data by using data quality functions described in this chapter
- captures data errors through user-defined business rules and sequesters the data for review and correction
- consolidates the information into a single view of the information available from all the data sources

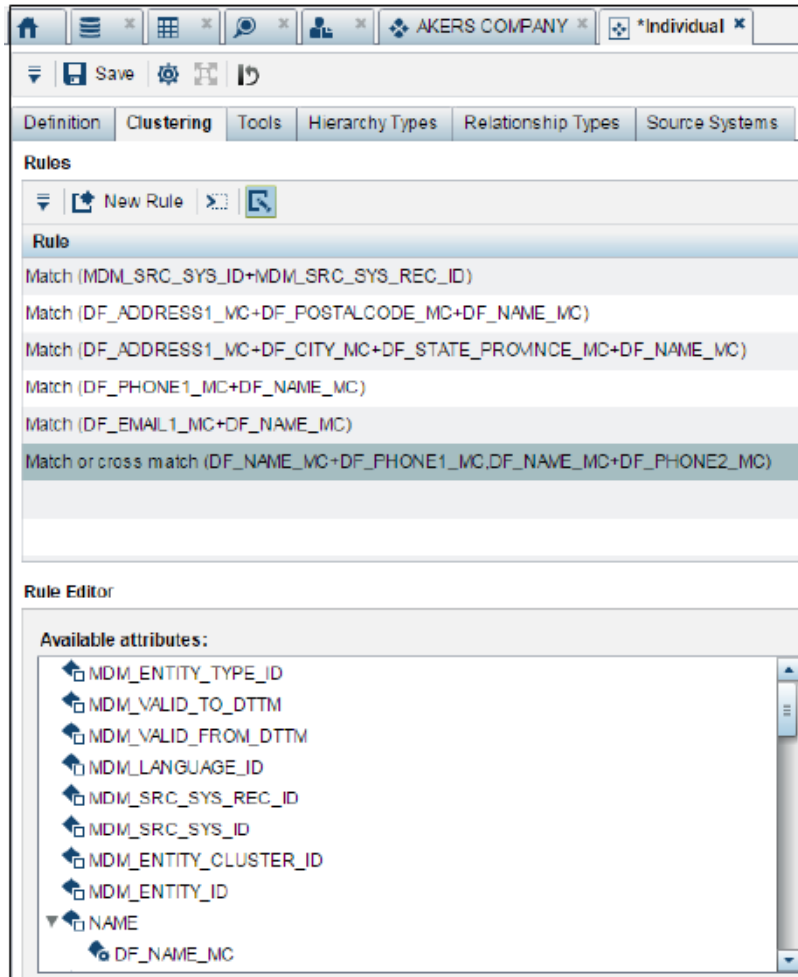
The organizational data that you provide can be customer data, product data, services data, or data for other entity types. SAS MDM applies a rigorous methodology to the problem of integrating disparate enterprise data.

SAS MDM is a combination of software, templates, documentation, data models, and services that provide the functionality and processes necessary to build and maintain a master entity database. An entity can be a customer, product, patient, site, or any business data object that you define. You can also define attributes for each entity. Also, you can use transformations that support data quality and identity management functionality.

The SAS MDM hub is a set of tables and table relationships that contain entity information and required entity keys and linking information. This hub provides end users or applications with a complete view of an entity and its relationships with other entities. In addition, the hub can be used as the single source for creating and maintaining survivor records that can be used in other applications or processes. The original source system IDs are important to other enterprise applications or data sources. They are also maintained in the hub to facilitate linking activities with other data providers or consumers.

The following display shows the interface for creating and editing rules on the **Clusters** tab for an entity type:

Figure 3.10 Entity Type Clustering Tab



Note: SAS MDM includes and is integrated with DataFlux Data Management Platform. Therefore, to use SAS MDM successfully, you must be familiar with DataFlux Data Management Platform.

Other Applications with Data Quality Functions

SAS Data Loader for Hadoop

SAS Data Loader for Hadoop addresses data quality concerns with the Cleanse Data, Cluster-Survive Data, and Match-Merge Data directives. Directives are wizards that help business users and data scientists perform tasks with their data without using specialized coding skills.

SAS Data Quality Accelerator for Hadoop and SAS Code Accelerator for Hadoop are bundled with and sold only as part of the SAS Data Loader for Hadoop. The data quality functionality includes match code generation, gender analysis, identification analysis, casing, and standardization.

The SAS Data Quality Accelerator for Hadoop and the SAS In-Database Code Accelerator for Hadoop process SAS DS2 code inside Hadoop. They use the MapReduce SAS® data-to-decision life cycle framework and the SAS Embedded Process. This approach minimizes the time spent moving and manipulating data for analytic processing and cleansing. It also eliminates the need to know how to code this capability in MapReduce, which improves performance and execution response times by harnessing the power of the Hadoop cluster.

SAS Code Accelerator for Hadoop and SAS Data Quality Accelerator for Hadoop provide the ability to push processing to the cluster. This ability enables the SAS programmer to apply data quality routines to data stored in Hadoop. This includes match coding, standardization routines, and parsing.

SAS Data Quality Accelerator for Teradata

SAS applications are often built to work with large volumes of data in environments that demand rigorous IT security and management. When the data is stored in an external database such as Teradata, the transfer of large data sets to computers running the

SAS System can cause a performance bottleneck. There are also possible unwanted security and resource management consequences for local data storage.

SAS Data Quality Accelerator for Teradata, which is included in SAS In-Database Technologies for Teradata, addresses these challenges. It moves computational tasks closer to the data and improves the integration between the SAS System and the database management system (DBMS).

SAS Data Quality Accelerator for Teradata provides in-database data quality operations as Teradata stored procedures. A stored procedure is a subroutine that is stored in the database and is available to applications that access a relational database.

The stored procedures perform the following data quality operations:

- changing to the proper case
- attribute extraction
- gender analysis
- identification analysis
- match code generation
- parsing
- pattern analysis
- standardization

SAS Data Integration Studio

SAS Data Integration Studio supports data quality improvement with the following transformations listed under the Data Quality category in the Transformations tree:

- Apply Lookup Standardization
- Create Match Code
- Standardize with Definition
- DataFlux Batch Job

■ DataFlux Data Service

Like DataFlux Data Management Studio, SAS Data Integration Studio is included in the SAS Data Management bundle.

You can use the DataFlux schemes in the Apply Lookup Standardization transformation to standardize the format, casing, and spelling of character columns in a source table. Similarly, you can select and apply DataFlux standardization definitions in the Standardize with Definition transformation to elements within a text string. For example, you might want to change all instances of “Mister” to “Mr.” but only when “Mister” is used as a salutation. However, this approach requires SAS Data Quality Server.

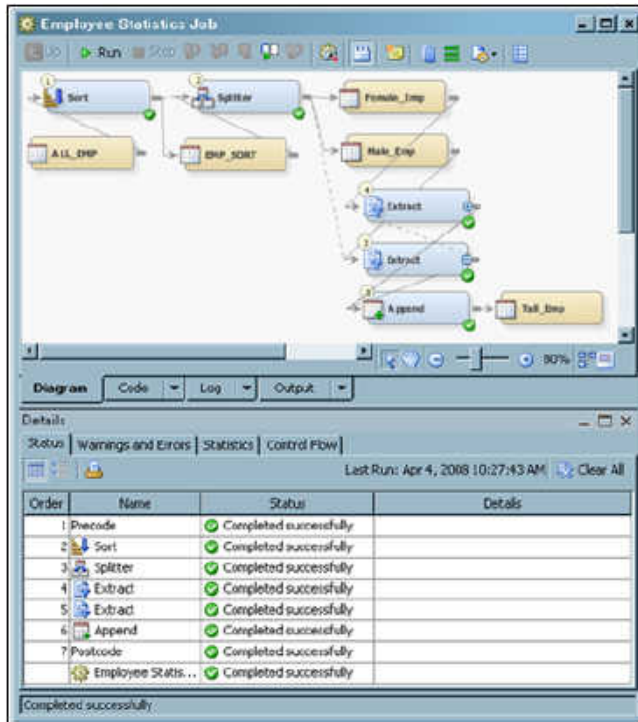
The Create Match Code transformation enables you to analyze source data and generate match codes based on common information shared by clusters of records. Comparing match codes instead of actual data enables you to identify records that are in fact the same entity, despite minor variations in the data.

The DataFlux Batch Job and DataFlux Data Service transformations enable you to select and execute DataFlux jobs and jobs configured as real time from a DataFlux Data Management Server. Then, you can perform DataFlux quality activities such as data jobs, process jobs, and profiles.

Many of the features in SAS Data Quality Server and the DataFlux Data Management Platform can be used in SAS Data Integration Studio jobs. For example, you can use DataFlux standardization schemes and definitions in SAS Data Integration Studio jobs. You can also execute DataFlux jobs, profiles, and services from SAS Data Integration Studio.

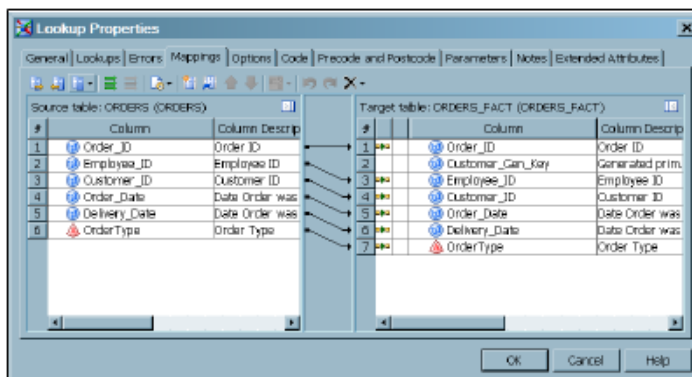
The following display shows a SAS Data Integration Studio job that sorts, splits, extracts, and appends data:

Figure 3.11 Data Job



The **Mapping** tab for a fact table used by the Lookup transformation in SAS Data Integration Studio is shown in the following display:

Figure 3.12 Fact Table Mapping Tab



4

Data Integration

<i>Overview</i>	53
<i>SAS Data Loader for Hadoop</i>	54
<i>SAS Data Integration Studio</i>	59
<i>SAS Visual Process Orchestration</i>	62
<i>DataFlux Data Management Studio</i>	63
<i>DataFlux Data Management Server</i>	64

Overview

The key applications for data integration in SAS data management include:

- “SAS Data Loader for Hadoop” on page 54
- “SAS Data Integration Studio” on page 59
- “DataFlux Data Management Studio” on page 63
- “SAS Visual Process Orchestration” on page 62

These approaches of these applications are summarized in the following table:

Application	Approach to Data Integration
SAS Data Loader for Hadoop	Provides directive-based data processing in Hadoop
SAS Data Integration Studio	Provides better data access capabilities, can run code in-database, is fully integrated into the SAS metadata layer, and has more options for interacting with Hadoop.
DataFlux Data Management Studio	Provides user interfaces for performing data exploration and data profiling. Includes a business rules component.
SAS Visual Process Orchestration	Provides a web-based design environment that pulls together SAS, DataFlux, and external jobs in logical flows.

DataFlux Data Management Server also performs data integration functions. For more information, see [“DataFlux Data Management Server”](#) on page 64.

SAS Data Loader for Hadoop

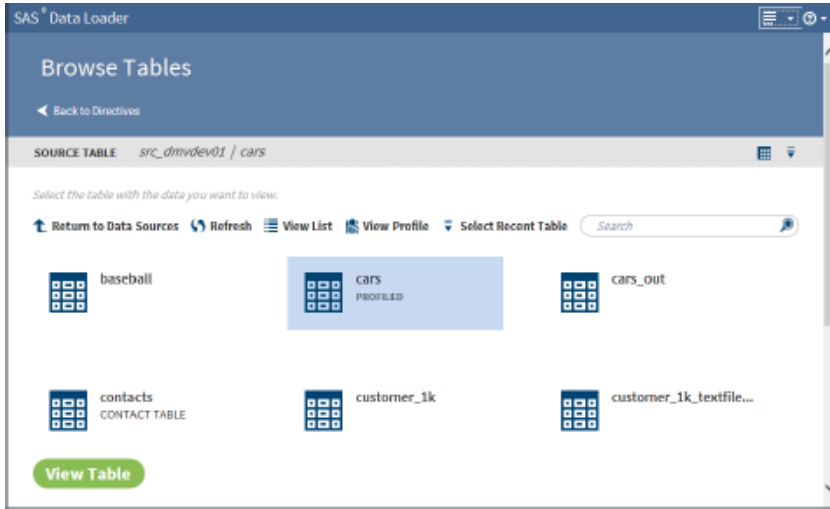
SAS Data Loader for Hadoop provides data integration, data quality, and data preparation capabilities without requiring you to write code. Your advanced users can edit and run HiveQL, Impala, or SAS DS2 code. Code runs inside Hadoop for improved performance.

You can also use SAS Data Loader for Hadoop to lift data into memory on the SAS LASR Analytic Server. Then you can use SAS Visual Analytics to do further visualization or analysis of that data.

You can use Hadoop regardless of your technical background. If you are a business analyst with little or no experience with Hadoop, you can use the wizard-based directives to merge, filter, and sort large distributed data sources.

For example, you can use the **Browse Tables** directive to see the available tables, as shown in the following display:

Figure 4.1 Browse Tables Directive



Then you can click **View Table** to see the table in the Table Viewer interface, which is shown in the following display:

Figure 4.2 Table Viewer

The screenshot shows the SAS Data Loader Table Viewer interface. The browser address bar displays the URL: `http://192.168.25.131/SASDataLoader/view/hadoopserver/path/HadoopServer/sas/DataLoader/ctrl-ctrl-ctrl-as74-4`. The interface includes a menu bar (File, Edit, View, Favorites, Tools, Help) and a title bar (SAS Data Loader - Table V...). The main area shows the table viewer for schema `name:tc_binadavet3` and table `name:car`. The row limit is set to 100. On the left, a 'Columns' list shows various attributes checked, including `make`, `model`, `type`, `origin`, `drive_train`, `msrp`, `invoice`, `engine_size`, `cylinders`, `horsepower`, `mpg_city`, `mpg_highway`, and `weight`. Below the columns list is a 'Property Value' table. The main table displays 23 rows of car data with columns: `index`, `make`, `model`, `type`, `origin`, `drive_train`, `msrp`, and `invoice`.

index	make	model	type	origin	drive_train	msrp	invoice
1	Acura	MDX	SUV	Asia	All	30945	333
2	Acura	RSX Type S	Sedan	Asia	Front	23820	217
3	Acura	TSX 4dr	Sedan	Asia	Front	20990	240
4	Acura	TL 4dr	Sedan	Asia	Front	33195	300
5	Acura	3.5 RL 4dr	Sedan	Asia	Front	43755	390
6	Acura	3.5 RL w/Na	Sedan	Asia	Front	46100	411
7	Acura	NSX coupe	Sports	Asia	Rear	80765	790
8	Audi	A4 1.8T 4dr	Sedan	Europe	Front	25940	235
9	Audi	A4 1.8T conv	Sedan	Europe	Front	30940	323
10	Audi	A4 3.0 4dr	Sedan	Europe	Front	31840	288
11	Audi	A4 3.0 Quatt	Sedan	Europe	All	33430	303
12	Audi	A4 3.0 Quatt	Sedan	Europe	All	34480	313
13	Audi	A6 3.0 4dr	Sedan	Europe	Front	36640	331
14	Audi	A6 3.0 Quatt	Sedan	Europe	All	36640	331
15	Audi	A4 3.0 conv	Sedan	Europe	Front	42490	383
16	Audi	A4 3.0 Quatt	Sedan	Europe	All	44240	400
17	Audi	A6 2.7 Turbi	Sedan	Europe	All	42840	388
18	Audi	A6 4.2 Quatt	Sedan	Europe	All	40690	449
19	Audi	A8 L Quatt	Sedan	Europe	All	69190	647
20	Audi	S4 Quattro	Sedan	Europe	All	48040	435
21	Audi	RS 6 4dr	Sports	Europe	Front	64600	764
22	Audi	TT 1.8 conv	Sports	Europe	Front	35940	325

Directives to perform the following types of functions are provided:

Acquire and Discover

Browse Tables, Import a File, Profile Data, Save Profile Reports, Copy Data to Hadoop

Transform and Integrate

Chain Directives, Query a Table in Hadoop, Query or Join Data, Run a Hadoop SQL Program, Run a SAS Program, Run Status, Transpose Data, Transform Data in Hadoop, Delete Rows, Save Directives

Cleanse and Deliver

Cleanse Data in Hadoop, Load Data to LASR, Match-Merge Data, Cluster-Survive Data, Sort and De-Duplicate Data

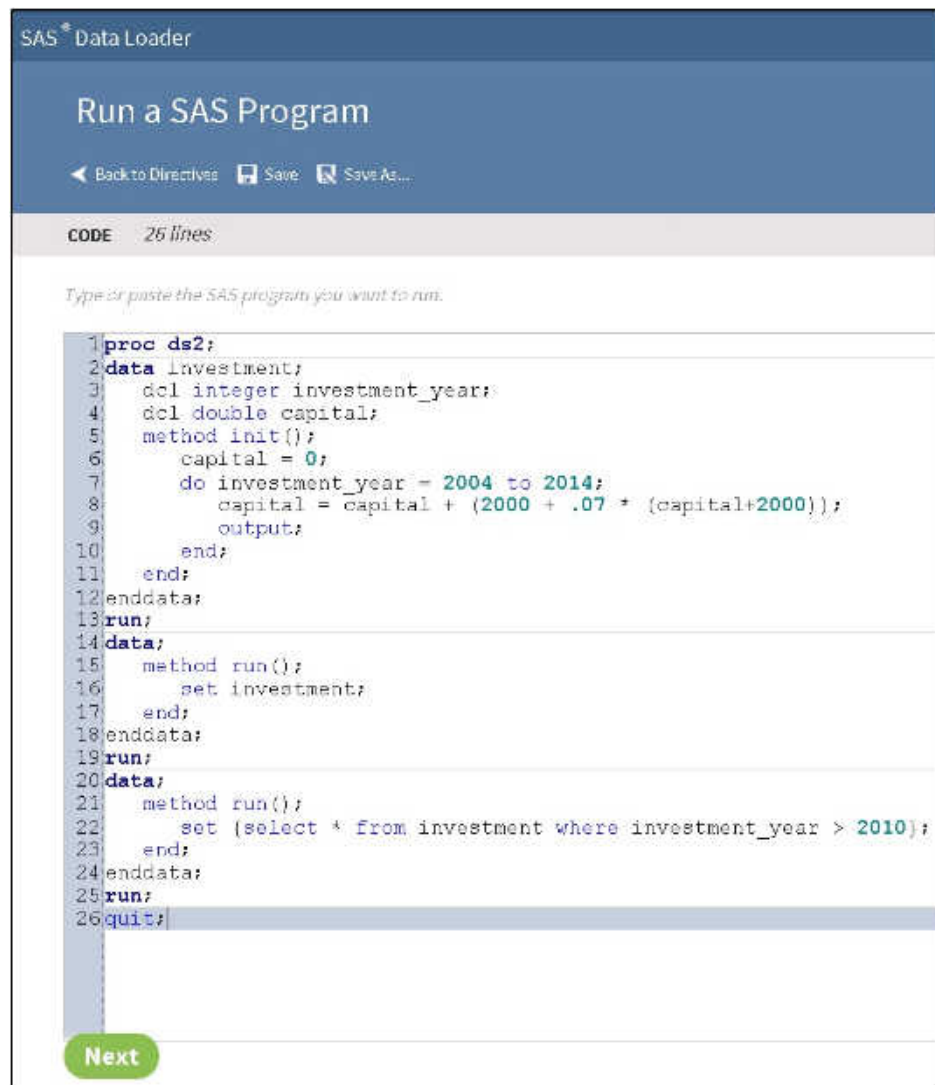
SAS Data Loader for Hadoop supports the data query and manipulation features in HiveQL. It also supports the data transformation features offered with SAS DS2, Quality Knowledge Bases, and SAS LASR Analytic Server.

You can use the Transform Data in Hadoop directive, for example, to select a source table, select one or more transformations, and select a target. Then you can apply transformations to filter, manage columns, and create summarized (aggregate) columns. The Profile Data directive enables you to select source columns from one or more tables to report uniqueness, incompleteness (null or blank), and patterns.

Programmers with experience in Hadoop and SAS also benefit from the simplicity of SAS Data Loader for Hadoop. Existing SAS DS2 programs and Hive SQL queries can be dropped into directives for repeat execution in Hadoop, with status monitoring in SAS Data Loader. HiveQL and DS2 code is generated as a result of certain directives. This code can be edited.

The following display shows SAS code running in SAS Data Loader for Hadoop:

Figure 4.3 Running Code in SAS Data Loader for Hadoop



The screenshot shows the SAS Data Loader interface. At the top, it says "SAS Data Loader" and "Run a SAS Program". Below that are navigation buttons: "Back to Directives", "Save", and "Save As...". A section labeled "CODE" indicates "26 lines". A prompt says "Type or paste the SAS program you want to run." Below this is a text area containing the following SAS code:

```
1 proc ds2;
2 data investment;
3   dcl integer investment_year;
4   dcl double capital;
5   method init();
6     capital = 0;
7     do investment_year = 2004 to 2014;
8       capital = capital + (2000 + .07 * (capital+2000));
9       output;
10    end;
11  end;
12 enddata;
13 run;
14 data;
15   method run();
16     set investment;
17   end;
18 enddata;
19 run;
20 data;
21   method run();
22     set (select * from investment where investment_year > 2010);
23   end;
24 enddata;
25 run;
26 quit;
```

At the bottom left, there is a green "Next" button.

SAS Data Integration Studio

SAS Data Integration Studio provides a graphical user interface and process flow to ease the creation and deployment of jobs that manage data.

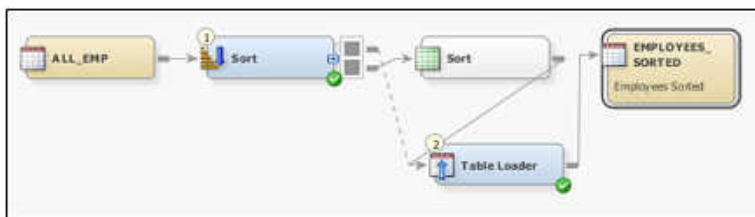
It supports the following goals:

- reducing development time by enabling the rapid generation of data warehouses, data marts, and data streams
- controlling the costs of data integration by supporting collaboration, code reuse, and common metadata
- increasing returns on existing IT investments by providing multi-platform scalability and interoperability
- creating process flows that are reusable, easily modified, and support embedded data quality processing. The flows are self-documenting and support data lineage analysis.

Data integration in SAS Data Integration Studio uses job flows. You can pull data into these jobs and modify it with transformations, user-written code, and wizards. Then you can store it in output tables that you can use for tasks such as analysis and reporting.

The following display shows a sample SAS Data Integration Studio job flow:

Figure 4.4 SAS Data Integration Studio Job Flow



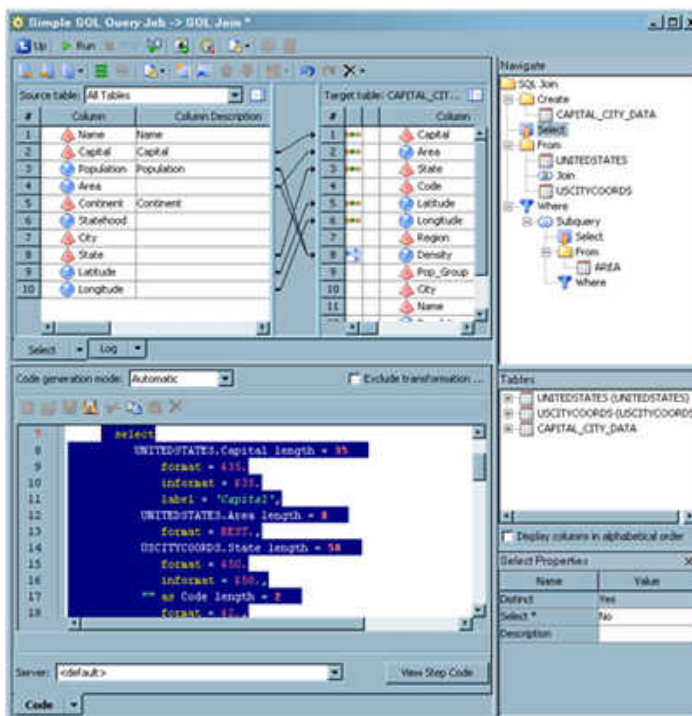
Note that this job flow contains a default temporary output table and a table loader. Job flows are created on the **Diagram** tab. This tab features tools that enable you to start

and restart jobs. You can also resize job flows, save images of flows, and perform other tasks that help you build and maintain the flows.

The Transformations tree groups the transformations available in SAS Data Integration Studio into folders. Several of these folders contain transformations that are useful for data integration tasks. For example, the Control folder contains transformations that help you work with conditional and loop processing. The SQL folder contains the Join transformation, which can run queries from a full-featured SQL wizard. It also contains specialized interfaces that work with specific SQL statements such as the Delete and Execute transformations.

The following display shows the **Designer** tab of the SQL Join transformation:

Figure 4.5 SQL Designer Tab



You can use the sections in the **Designer** tab to map source columns to target columns. You can also navigate within the query, set code generation options, review tables and columns, and review SQL Join properties.

Most data integration transformations are found in the Data folder.

Data transformations include the following:

- Append
- Compare Tables
- Data Validation
- Extract
- Rank
- Sort
- Transpose
- User-Written Code

SAS Data Integration Studio supports the following data management tasks:

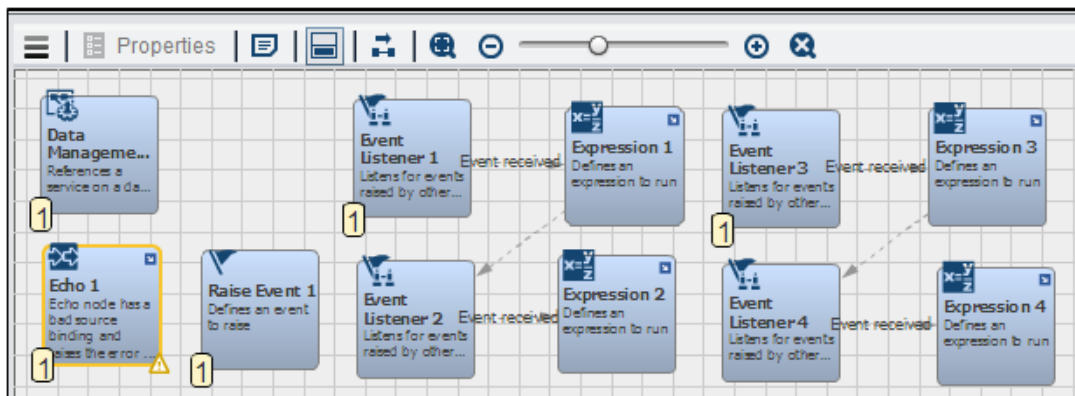
- job flow processing with features that include checkpoints, restarting jobs, batch deployment, and status handling
- job versioning, data lineage, and impact analysis
- importing and exporting metadata for individual objects or sets of related objects. You can work with two types of metadata. The first is SAS metadata in SAS Package format. The second is relational metadata such as metadata for libraries, tables, columns, indexes, and keys. This metadata must be in formats that can be accessed with a SAS Metadata Bridge.
- job management tasks such as submission, reviewing, and debugging
- job deployment and redeployment tasks. You can deploy jobs from a command line, using a scheduler, and as a stored process.
- running job documentation reports
- running DataFlux Data Management Platform jobs and processes such as profiles and standardization schemes

SAS Visual Process Orchestration

SAS Visual Process Orchestration is a web-based process flow authoring environment that is launched from SAS Data Management Console. The authoring environment provides nodes that can be used to build orchestration jobs, which are process jobs that run other jobs. When you create data integration and data quality jobs, you can use SAS Visual Process Orchestration to ensure the coordinated execution of these jobs. This coordinated execution is based on control logic.

The following display shows a SAS Visual Process Orchestration job flow:

Figure 4.6 SAS Visual Process Orchestration Job Flow



SAS Visual Process Orchestration provides the following benefits:

- It enables you to integrate jobs (executable files) from various systems into a single orchestration job. Referenced jobs can include SAS Data Integration Studio jobs, DataFlux Data Management Studio jobs, SAS code, third-party programs, scripts, and web services.
- Orchestration jobs can execute referenced jobs in parallel.
- Orchestration jobs can apply control logic such as looping and IF/THEN/ELSE handling.

- Orchestration jobs can handle events, error-checking, and run-time statistics for each node in the job.
- Orchestration jobs can be versioned and locked.
- SAS Visual Process Orchestration is fully integrated with the SAS platform. For example, orchestration jobs are stored in SAS folders. You can use the SAS object promotion framework in SAS Management Console to move orchestration jobs between test, development, and production environments.

DataFlux Data Management Studio

You can use data jobs and process jobs to perform data integration tasks in DataFlux Data Management Studio. For example, the data inputs data job node category contains nodes that tasks such as running SQL queries, extracting table metadata, and processing XML input. Similarly, the data output nodes support tasks such as deleting records and generating an HTML-formatted report from the results of a data job. They can also produce a report that lists the duplicate records identified with match criteria that you have specified.

The nodes in the data integration category support a range of tasks. These tasks include sorting and joining your data, combining the data from two data sets, and SQL lookup and execution. You can also use data integration nodes to issue SOAP and HTTP requests.

You can use the process job SQL nodes to perform tasks such as running SQL in parallel and managing custom SQL scripts. You can also write your own SQL and create or insert content into tables. The process job data integration nodes are useful when you need to write some code into the node or point to a file that contains some SAS code. For example, the **SAS Code Reference** process job node enables you to point to a SAS code file on the file system or in a DataFlux repository. You can then execute that code as part of a process job. A Data Integration license is required to get these nodes.

DataFlux Data Management Server

The DataFlux Data Management Server provides consistent, accurate, and reliable access to data across a network by integrating real-time data quality, data integration, and data governance routines. Jobs created with DataFlux Data Management Studio are deployed to the DataFlux Data Management Server for faster execution in both real-time and batch modes. With DataFlux Data Management Server, you can replicate your business logic for acceptable data across applications and systems, enabling you to build a single, unified view of your enterprise. The server implements business rules that you create in DataFlux Data Management Studio, in both batch and real-time environments. DataFlux Data Management Server enables pervasive data quality, data integration, process orchestration, and master data management (MDM) throughout your enterprise.

The Data Management Server provides a service-oriented architecture (SOA) application server that enables you to execute batch or profile jobs on a server-based platform, in Windows, Linux, or UNIX. By processing batch and profile jobs where the data resides, you avoid network bottlenecks and take advantage of performance features available with higher-performance computers. DataFlux Data Management Server has both a SOAP based and a REST-based web services API.

In addition, the DataFlux Data Management Server executes real-time data services and real-time process services. These services can be invoked by any web service application, such as SAP, Siebel, Tibco, or Oracle. You can convert your existing batch jobs to real-time services to reuse the business logic that you developed for data migration or to load a data warehouse. You can apply your real-time services at the point of data entry to ensure consistent, accurate, and reliable data across your enterprise.

5

Data Governance

<i>Overview</i>	65
<i>Shared Metadata</i>	66
<i>DataFlux Web Studio</i>	66
<i>SAS Business Data Network</i>	69
<i>SAS Lineage</i>	71
<i>SAS Data Remediation</i>	74

Overview

Data Governance encompasses the definition and implementation of policies and procedures for managing your data as a strategic asset. For example, SAS Business Data Network can be used to store definitions and owners for business terms like *profit* or *customer*. Then these business terms can be linked to the technical metadata that defines where customer data is stored in different databases, tables, and fields across the organization. Business Rules can be defined and implemented to improve and monitor the health of your data.

The key technologies for data governance in SAS data management include:

- [“Shared Metadata” on page 66](#)
- [“DataFlux Web Studio” on page 66](#)

- [“SAS Business Data Network” on page 69](#)
- [“SAS Lineage” on page 71](#)
- [“SAS Data Remediation” on page 74](#)

Shared Metadata

SAS Data Management provides the tightest integration in the industry that spans the entire data management lifecycle. Metadata is shared between the data management and analytics domains. For example, during the profiling phase, data correction strategies are identified and documented within the SAS repository. After documenting these rules, the profiling engine can be prompted, with a single click of the mouse, to automatically build the data correction workflow. The profiling engine shares all metadata with the data quality engine. This shared metadata includes items such as data source connection information, data quality rules defined during the profiling phase, and field names.

This automatically generated workflow can be invoked through Service-Oriented Architectures. For example, users, groups, and logins can be shared with data quality jobs to streamline integration with SAS analytical solutions. This integration layer can include complex business logic. It can also contain core data quality algorithms such parsing, standardization, and matching.

Shared metadata in SAS is common throughout the platform. It uses metadata bridges that are available to integrated metadata across applications. SAS applications use a relationship importer that enables metadata to flow across the metadata bridges.

DataFlux Web Studio

DataFlux Web Studio supports data management functions with three components: Dashboard, Data Monitor, Reference Data Manager.

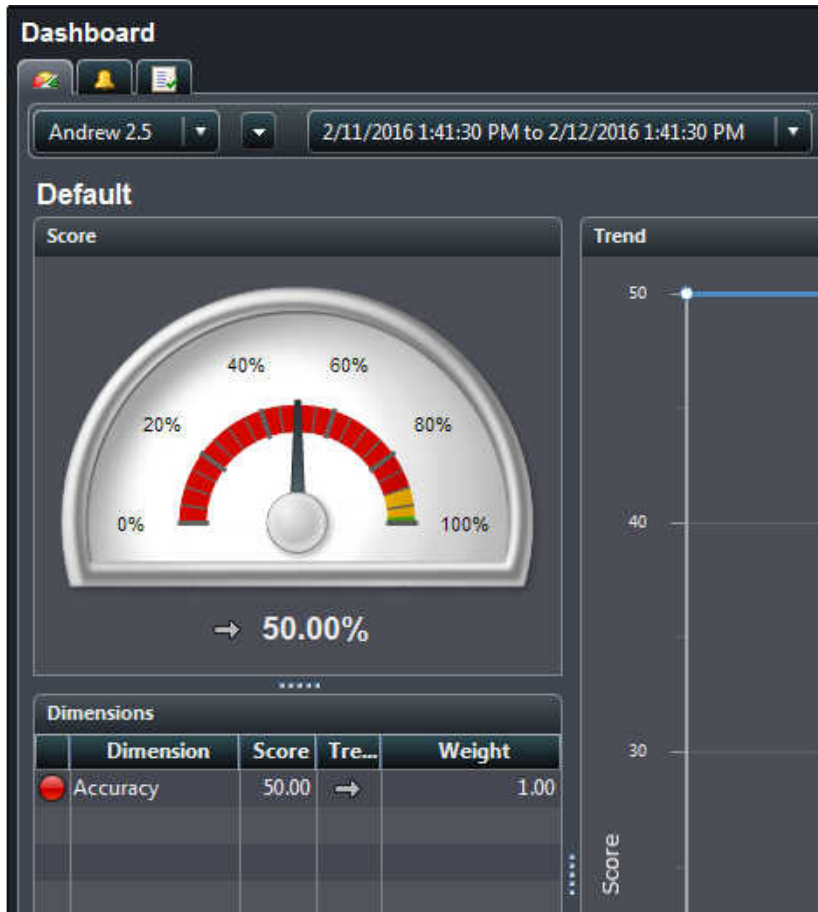
The Monitor Viewer and Dashboard Viewer components in DataFlux Web Studio provide web interfaces for viewing exceptions to monitored business rules. The exceptions are generated by data monitoring jobs that are created in DataFlux Data Management Studio and are deployed to a DataFlux Data Management Server. These web interfaces are similar to the Monitor Viewer and Dashboard tab in Data Management Studio. However, the web interfaces are available to those who do not have access to DataFlux Data Management Studio.

The SAS monitoring capability supports the ability to detect data anomalies. These anomalies include variances, values inside or outside of ranges, values that violate mathematical calculations, values that vary from historical values, and much more. Once identified, these anomalies are presented to the business user through an intuitive graphical dashboard for further analysis. The dashboard not only displays data violations, but also supports the ability to trend values over time as well as compare current values to historical values.

Graphical displays aid in visualizing and trending issues that have been uncovered during the monitoring of specified business rules. These data quality dashboards provide high-level access and understanding to all data quality issues discovered by the SAS solution.

The following display shows the Dashboard interface, which displays the status of parameters that are determined by business rules that are set up in DataFlux Data Management Studio.

Figure 5.1 Dashboard Viewer



The Monitor Viewer provides a detailed, interactive view of the exceptions to monitored business rules. The exceptions are generated by data monitoring jobs that are created in DataFlux Data Management Studio. The output of these jobs can be monitored in DataFlux Web Studio, using Monitor Viewer.

The Reference Data Manager is a component of DataFlux Web Studio. It provides a web interface for creating and managing reference data. Examples of reference data include a list of valid values for a product code and a complete list of countries with their

associated cities and states. The values are organized into domains. Reference Data Manager domains can be used in Data Management Studio data jobs and in business rules.

Reference Data Manager enables you to build and manage relationships and hierarchies across data elements. Reference Data Manager offers a web-based interface that can be used across departments and lines of business to store and manage different types of hierarchies.

Reference Data Manager allows data governance and data management teams to perform the following tasks:

- maintain reference data entities consistently across the organization, with definitions and relationships clearly defined for each entity and the entity values
- validate data in any source system using the reference data managed in this application
- create reference data required for effective master data management efforts

Using the Reference Data Manager component, you can access a single repository to manage important reference data that is accessed by technical, business, and IT users. This repository is centrally managed, versioned (so that the history and progression of changes to the data can be maintained and managed), and exported to external systems.

SAS Business Data Network

SAS Business Data Network is an application that enables you to manage business terms. It supports a collaborative approach to managing the following information:

- descriptions of business terms, including their requirements and attributes
- related source data and reference data
- contacts (such as technical owners, business owners, and interested parties)
- relationships between terms and processes (such as SAS Data Management Studio jobs, services, and business rules)

By linking terms to business rules and data monitoring processes, SAS Business Data Network provides a single entry point for all data consumers to better understand their data. Data stewards, IT staff, and enterprise architects can use the terms to promote a common vocabulary across projects and business units. Permissions can be set to allow only specific users to access or control the data in SAS Business Data Network.

SAS Business Data Network enables collaboration of domain knowledge between business users, technical users, and data stewards. SAS Business Data Network can be used as a single entry point for all data consumers to better understand their data. It consists of a web user interface that documents business terms and their associated rules, jobs, applications, data, documentation, and other information.

The following display shows a segment of the **All Terms** list:

Figure 5.2 All Terms List

Term Name	Description	Type	Import...	Status	Workflow
Logistics	Logistics Division	create	★★★★☆	● Producti...	
Loading...	test	Create Extend...	★★★★☆	● Under R...	Extended Cre...
Wareho...	Warehouse organization	Default	★★★★☆	● Not Spe...	
Wareho...	Storage facility for goods...	Create Extend...	★★★★☆	● Under R...	

The Type, Status, and Workflow columns provide workflow information for the terms.

Technical users use the network to document information about tables and columns that implement the business terminology. This information can be used to relate jobs and other information to terms and to share knowledge about data transformations. It serves as a data dictionary to describe details of data models and other data-related information. Data stewards can view data from a business standpoint to better visualize problem areas by domain in order to identify and fix data issues more effectively.

SAS Business Data Network provides the following benefits:

- ability to enter non-collaborative terms, which do not undergo a review and approval process, or collaborative terms, which are reviewed and approved.

The following display shows a Notifications list, which is used in the workflow for collaborative term approval:

Figure 5.3 Notifications List

Term Name	Description	Type	Importan...	Status
Shelving Rack	Shelving Rack in a warehous...	Create Extended	★★★★☆	Editing

- integration with SAS Workflow Studio. This integration enables you to quickly see collaborative review tasks that are waiting on their input in the Task Manager View in SAS Data Management Console and in views in the network.
- support for multiple, customized term templates.
- ability to customize the attributes of a term through the term type.

SAS Lineage

SAS Lineage is a web-based diagram component for visualizing relationships between objects. It is used as a stand-alone lineage and relationship viewer that can be accessed by SAS database management and business intelligence applications. The component has two modes:

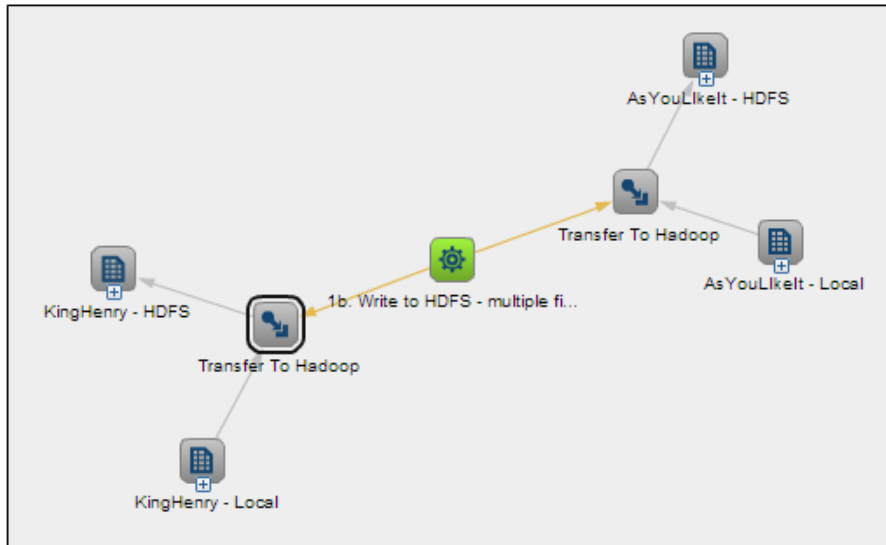
- A network diagram displays all relationships.
- Two left-to-right dependency diagrams are available: one that displays data governance information (governance) and another that displays parent and child relationships only (impact analysis).

The relationship information displayed in these diagrams is drawn from the Relationship database that is a part of the Web Infrastructure Platform Data Server.

SAS Lineage can display most types of SAS metadata. This includes data objects, including columns, tables, external files, information maps, reports, stored processes, SAS Enterprise Guide projects and associated objects, and the levels and measures in OLAP cubes. You can also display objects created in SAS Business Data Network, such as terms, tags, and associated items.

The following display shows a network diagram in SAS Lineage:

Figure 5.4 Network Diagram



Lineage and impact analysis include the following features:

- a shared store for all relationship information, called the SAS relationship service. Most SAS products and object types are now integrated into the SAS relationship service.
- the ability to import content from third-party sources.
- enhancements to the relationships web viewer to add a number of new views for displaying information stored in the service.
- views for Relationships and Data Governance. These views are named All Relationships, Governance, and Impact Analysis.

You can also create your own views using the filtering capabilities of the viewer. This can help you subset the information to only the objects and relationships that you want to see.

The following display shows a user-defined view definition:

Figure 5.5 User-Defined View Definition

In addition, there are helpful features such as grouping node sets, which enable you to expand on demand, and an overview window with details of objects. Metadata from external systems can be imported with a batch API and metadata bridges to third-party systems. This capacity enables you to see lineage of objects that go beyond the SAS environment.

SAS Data Remediation

For many data-intensive IT projects, anomalies or inconsistencies in the data prevent systems involved from operating optimally and providing clean and timely data to each other and end users. Data remediation provides a means to identify, review, and correct the problem data before it reaches the downstream systems.

SAS Data Remediation makes it easy to capture and review problems found in enterprise data. SAS Data Remediation has a web-based interface for data administrators and a representational state transfer (REST) web service API for system integration. Both of these interfaces interact with a remediation database that contains information about where the problem data is located. You can use this information to review which system generated the data, who should see the data, and how the data might be corrected.

Data remediation allows user- or role-based access to data exceptions, which are categorized by application and subject area. Once data remediation issues have been reviewed, they can be corrected through the same application, eliminating the need for another user to complete the correction process. All data remediation issues can also be associated with workflow definitions that route the issues to the correct decision maker for approval or instructions for additional action.

The following display shows the Data Remediation Summary portlet in the SAS Data Management Console:

Figure 5.6 Data Remediation Summary Portlet



A portion of the SAS Data Remediation tab is shown in the following display:

Figure 5.7 SAS Data Remediation Tab



6

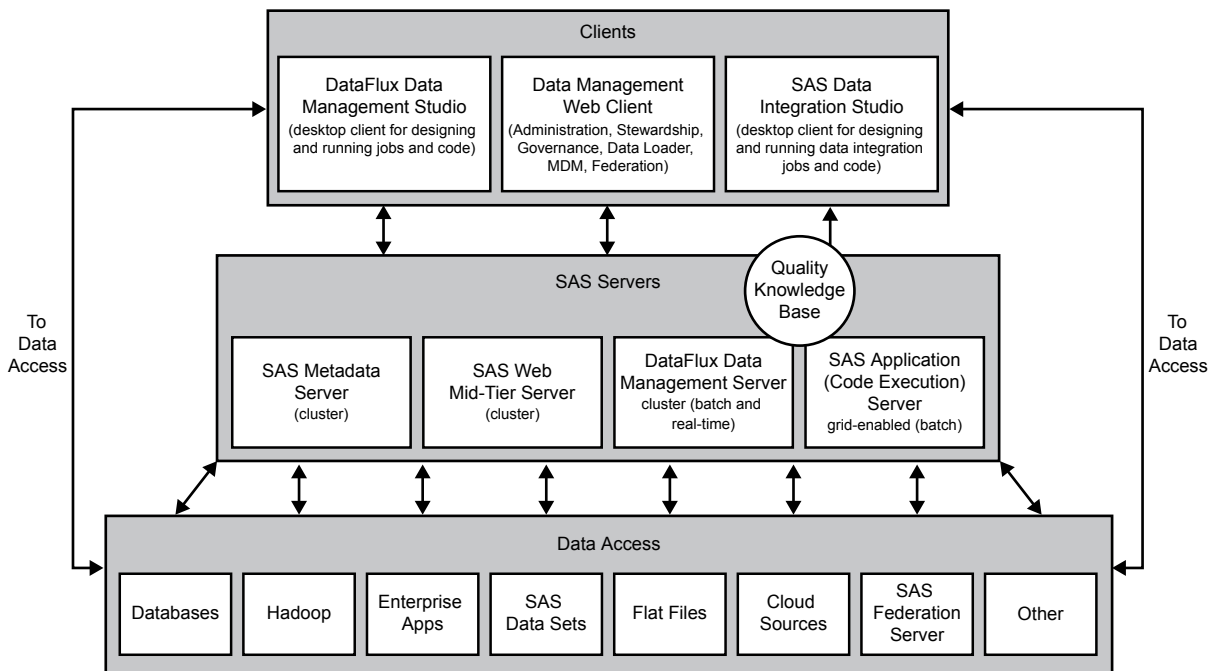
Architecture

Overview	78
Data Access	79
Desktop Clients	80
Web Tier	80
Server Tier	81
SAS Metadata Server	81
SAS Application Servers	82
DataFlux Data Management Server	83

Overview

The following display presents a high-level overview of the SAS data management architecture:

Figure 6.1 SAS Data Management Architecture Overview



Data moves between the parts of the SAS data architecture through the use of the following methods:

- JDBC
- ODBC
- native access
- SAS/ACCESS
- accelerators

- REST API
- Web Services

The SAS data management architecture consists of the following components:

- [“Data Access” on page 79](#)
- [“Desktop Clients” on page 80](#)
- [“Web Tier” on page 80](#)
- [“Server Tier” on page 81](#)

Data Access

SAS Data Management offerings can access most data types. The accessible data sources include the following:

- relational and non-relational databases
- web services
- Hadoop
- Enterprise applications
- SAS data sets
- flat files
- cloud data sources

SAS data management applications include data loaders, ODBC interfaces, SAS/ACCESS engines, Hadoop transformations, and external file wizards to facilitate access to these types of data.

Desktop Clients

DataFlux Data Management Studio is the primary desktop client for performing data quality tasks in SAS data management. It is closely integrated with DataFlux Data Management Server and DataFlux Web Studio. SAS Data Integration Studio is the primary desktop client for data integration and data management tasks. Unlike DataFlux Data Management Studio, SAS Data Integration Studio generates SAS code.

Web Tier

The web tier provides an environment in which the data management web applications, such as SAS Business Data Network, SAS Visual Process Orchestration, and SAS Lineage, can execute. These products run in a web application server and communicate with the user by sending data to and receiving data from the user's web browser. The applications rely on servers on the SAS server tier to perform SAS processing.

The SAS Web Server and SAS Web Application Server provide a highly scalable, easy-to-manage environment that is dedicated to running SAS web applications. The SAS Web Server provides static HTTP content to users. It forwards requests for dynamic content to the SAS Web Application Server, which provides an execution environment for applications and services.

The SAS Web Infrastructure Platform is a collection of services and applications that provide common infrastructure and integration features to be used by SAS web applications.

These services and applications provide the following benefits:

- consistency in installation, configuration, and administration tasks for web applications
- greater consistency in users' interactions with web applications

- integration among web applications as a result of the ability to share common resources

DataFlux Web Studio is a DataFlux application that works in conjunction with the DataFlux Web Studio Server and DataFlux Data Management Studio.

Server Tier

The server tier includes [“SAS Metadata Server” on page 81](#), [“SAS Application Servers” on page 82](#), and [“DataFlux Data Management Server” on page 83](#).

SAS Metadata Server

The SAS Metadata Server is a critical software component in SAS data management. All of the client applications and the other SAS servers depend on the SAS Metadata Server and cannot operate without it.

The SAS Metadata Server is a multi-user server that serves metadata from one or more SAS Metadata Repositories to all of the client applications in your environment. The SAS Metadata Server enables centralized control so that all users access consistent and accurate data.

The functionality of the SAS Metadata Server is provided through the SAS Open Metadata Architecture, which is a metadata management facility that provides common metadata services to applications. One metadata server supports all of the SAS applications in your environment and can support hundreds of concurrent users.

This architecture enables the following:

- exchanging metadata between applications. so that applications can work together more easily.
- centrally managing metadata resources. Because there is a common framework for creating, accessing, and updating metadata, it is easier to manage the applications that rely on this metadata.

The SAS Metadata Server stores information about the following:

- the enterprise data sources and data structures that are accessed by SAS applications
- resources that are created and used by SAS applications, including information maps, OLAP cubes, report definitions, stored process definitions, and scheduled jobs
- the servers that run SAS processes
- the users and groups of users that use the system, and the levels of access that users and groups have to resources

SAS Application Servers

When SAS data management is installed at your site, a metadata object that represents the SAS server tier in your environment was defined. In the SAS Management Console interface, this type of object is called a SAS Application Server. By default, this application server is named SASApp.

A SAS Application Server is not an actual server that can execute SAS code submitted by clients. Rather, it is a logical container for a set of application server components, which do execute code. This code is typically SAS code, although some components can execute Java code or MDX queries. For example, a SAS Application Server might contain a workspace server, which can execute SAS code that is generated by clients such as SAS Data Integration Studio. A SAS Application Server might also contain a stored process server, which executes SAS Stored Processes. It also might contain a SAS/CONNECT Server, which can upload or download data and execute SAS code that is submitted from a remote machine.

The following table lists the main SAS Application Server components and describes how each one is used:

Table 6.1 SAS Application Servers

Server	How the Server Is Used
SAS Workspace Server	Executes SAS code; reads and writes data.

Server	How the Server Is Used
SAS/ CONNECT Server	Submits generated SAS code to machines that are remote from the default SAS Application Server; can also be used for interactive access to remote libraries.
Stored Process Server	Submits stored processes for execution by a SAS session. Stored processes are SAS programs that are stored and can be executed by client applications.
SAS Grid Server	Supports a compute grid that can execute grid-enabled jobs that are created in SAS Data Integration Studio.

All of the SAS Application Servers are specified as a component in a SAS Application Server object.

DataFlux Data Management Server

The DataFlux Data Management Server provides consistent, accurate, and reliable access to data across a network by integrating real-time data quality, data integration, and data governance routines. With DataFlux Data Management Server, you can replicate your business rules for acceptable data across applications and systems, enabling you to build a single, unified view of your enterprise. The server implements business rules that you create in DataFlux Data Management Studio, in both batch and real-time environments. DataFlux Data Management Server enables pervasive data quality, data integration, process orchestration, and master data management (MDM) throughout your enterprise.

The DataFlux Data Management Server provides a service-oriented architecture (SOA) application server that enables you to execute batch or profile jobs on a server-based platform in Windows, Linux, or UNIX. By processing batch and profile jobs where the data resides, you avoid network bottlenecks and take advantage of performance features that are available with higher-performance computers.

In addition, the DataFlux Data Management Server executes real-time data services and real-time process services. These services can be invoked by any web service application, such as SAP, Siebel, Tibco, or Oracle. You can convert your existing batch jobs to real-time services to reuse the business logic that you developed for data

migration or to load a data warehouse. You can apply your real-time services at the point of data entry to ensure consistent, accurate, and reliable data across your enterprise.

Appendix 1

Legal Notices

Apache Portable Runtime License Disclosure

Copyright © 2008 DataFlux Corporation LLC, Cary, NC USA.

Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at <http://www.apache.org/licenses/LICENSE-2.0>

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

Apache/Xerces Copyright Disclosure

The Apache Software License, Version 3.1

Copyright © 1999-2003 The Apache Software Foundation. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- 1 Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

- 2 Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- 3 The end-user documentation included with the redistribution, if any, must include the following acknowledgment: "This product includes software developed by the Apache Software Foundation (<http://www.apache.org>).". Alternately, this acknowledgment may appear in the software itself, if and wherever such third-party acknowledgments normally appear.
- 4 The names "Xerces" and "Apache Software Foundation" must not be used to endorse or promote products derived from this software without prior written permission. For written permission, please contact apache@apache.org.
- 5 Products derived from this software may not be called "Apache", nor may "Apache" appear in their name, without prior written permission of the Apache Software Foundation.

THIS SOFTWARE IS PROVIDED "AS IS" AND ANY EXPRESSED OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE APACHE SOFTWARE FOUNDATION OR ITS CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

This software consists of voluntary contributions made by many individuals on behalf of the Apache Software Foundation and was originally based on software copyright (c) 1999, International Business Machines, Inc., <http://www.ibm.com>. For more information on the Apache Software Foundation, please see <http://www.apache.org>.

Boost Software License Disclosure

Boost Software License - Version 1.0 - August 17, 2003

Permission is hereby granted, free of charge, to any person or organization obtaining a copy of the software and accompanying documentation covered by this license (the "Software") to use, reproduce, display, distribute, execute, and transmit the Software, and to prepare derivative works of the Software, and to permit third-parties to whom the Software is furnished to do so, all subject to the following:

The copyright notices in the Software and this entire statement, including the above license grant, this restriction and the following disclaimer, must be included in all copies of the Software, in whole or in part, and all derivative works of the Software, unless such copies or derivative works are solely in the form of machine-executable object code generated by a source language processor.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE AND NON-INFRINGEMENT. IN NO EVENT SHALL THE COPYRIGHT HOLDERS OR ANYONE DISTRIBUTING THE SOFTWARE BE LIABLE FOR ANY DAMAGES OR OTHER LIABILITY, WHETHER IN CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Canada Post Copyright Disclosure

The Data for areas of Canada includes information taken with permission from Canadian authorities, including: © Her Majesty the Queen in Right of Canada, © Queen's Printer for Ontario, © Canada Post Corporation, GeoBase®, © Department of Natural Resources Canada. All rights reserved.

DataDirect Copyright Disclosure

Portions of this software are copyrighted by DataDirect Technologies Corp., 1991 - 2008.

Expat Copyright Disclosure

Part of the software embedded in this product is Expat software.

Copyright © 1998, 1999, 2000 Thai Open Source Software Center Ltd.

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

gSOAP Copyright Disclosure

Part of the software embedded in this product is gSOAP software.

Portions created by gSOAP are Copyright © 2001-2004 Robert A. van Engelen, Genivia inc. All Rights Reserved.

THE SOFTWARE IN THIS PRODUCT WAS IN PART PROVIDED BY GENIVIA INC AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO,

PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

IBM Copyright Disclosure

ICU License - ICU 1.8.1 and later [as used in DataFlux clients and servers.]

COPYRIGHT AND PERMISSION NOTICE

Copyright © 1995-2005 International Business Machines Corporation and others. All Rights Reserved.

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, provided that the above copyright notice(s) and this permission notice appear in all copies of the Software and that both the above copyright notice(s) and this permission notice appear in supporting documentation.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR HOLDERS INCLUDED IN THIS NOTICE BE LIABLE FOR ANY CLAIM, OR ANY SPECIAL INDIRECT OR CONSEQUENTIAL DAMAGES, OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.

Except as contained in this notice, the name of a copyright holder shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization of the copyright holder.

Informatica Address Doctor Copyright Disclosure

AddressDoctor® Software, © 1994-2015 Platon Data Technology GmbH

Loqate Copyright Disclosure

The Customer hereby acknowledges the following Copyright notices may apply to reference data.

Australia: Copyright. Based on data provided under license from PSMA Australia Limited (www.pasma.com.au)

Austria: © Bundesamt für Eich- und Vermessungswesen

Brazil: Conteúdo fornecido por MapLink. Brazil POIs may not be used in publically accessible, internet-based web sites whereby consumers obtain POI data for their personal use.

Canada:

Copyright Notice: This data includes information taken with permission from Canadian authorities, including © Her Majesty, © Queen's Printer for Ontario, © Canada Post, GeoBase ®.

End User Terms: The Data may include or reflect data of licensors including Her Majesty and Canada Post. Such data is licensed on an "as is" basis. The licensors, including Her Majesty and Canada Post, make no guarantees, representation, or warranties respecting such data, either express or implied, arising by law or otherwise, including but not limited to, effectiveness, completeness, accuracy, or fitness for a purpose. The licensors, including Her Majesty and Canada Post, shall not be liable in respect of any claim, demand or action, irrespective of the nature of the cause of the claim, demand or action alleging any loss, injury or damages, direct or indirect, which may result from the use or possession of the data or the Data.

The licensors, including Her Majesty and Canada Post, shall not be liable in any way for loss of revenues or contracts, or any other consequential loss of any kind resulting from any defect in the data or in the Data.

End User shall indemnify and save harmless the licensors, including Her Majesty the Queen, the Minister of Natural Resources of Canada and Canada Post, and their officers, employees and agents from and against any claim, demand or action, irrespective of the nature of the cause of the claim, demand or action, alleging loss, costs, expenses, damages, or injuries (including injuries resulting in death) arising out of the use of possession of the data or the Data.

Croatia, Cyprus, Estonia, Latvia, Lithuania, Moldova, Poland, Slovenia, and/or Ukraine:
© EuroGeographics

France: source: G oroute® IGN France & BD Carto® IGN France

Germany: Die Grundlagendaten wurden mit Genehmigung der zust ndigen Beh rden entnommen

Great Britain: Based upon Crown Copyright material.

Greece: Copyright Geomatics Ltd. Hungary: Copyright   2003; Top-Map Ltd.

Italy: La Banca Dati Italiana   stata prodotta usando quale riferimento anche cartografia numerica ed al tratto prodotta e fornita dalla Regione Toscana.

Norway: Copyright   2000; Norwegian Mapping Authority

Portugal: Source: IgeoE – Portugal

Spain: Informaci n geogr fica propiedad del CNIG

Sweden: Based upon electronic data   National Land Survey Sweden.

Switzerland: Topografische Grundlage   Bundesamt f r Landestopographie.

Microsoft Copyright Disclosure

Microsoft®, Windows, NT, SQL Server, and Access, are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.

Oracle Copyright Disclosure

Oracle, JD Edwards, PeopleSoft, and Siebel are registered trademarks of Oracle Corporation and/or its affiliates.

PCRE Copyright Disclosure

A modified version of the open source software PCRE library package, written by Philip Hazel and copyrighted by the University of Cambridge, England, has been used by DataFlux for regular expression support. More information on this library can be found at: <ftp://ftp.csx.cam.ac.uk/pub/software/programming/pcre/>.

Copyright © 1997-2005 University of Cambridge. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of the University of Cambridge nor the name of Google Inc. nor the names of their contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Red Hat Copyright Disclosure

Red Hat® Enterprise Linux®, and Red Hat Fedora™ are registered trademarks of Red Hat, Inc. in the United States and other countries.

SAS Copyright Disclosure

Portions of this software and documentation are copyrighted by SAS® Institute Inc., Cary, NC, USA, 2009. All Rights Reserved.

SQLite Copyright Disclosure

The original author of SQLite has dedicated the code to the public domain. Anyone is free to copy, modify, publish, use, compile, sell, or distribute the original SQLite code, either in source code form or as a compiled binary, for any purpose, commercial or non-commercial, and by any means.

Sun Microsystems Copyright Disclosure

Java™ is a trademark of Sun Microsystems, Inc. in the U.S. or other countries.

TomTom Copyright Disclosure

© 2006-2015 TomTom. All rights reserved. This material is proprietary and the subject of copyright protection, database right protection, and other intellectual property rights owned by TomTom or its suppliers. The use of this material is subject to the terms of a license agreement. Any unauthorized copying or disclosure of this material will lead to criminal and civil liabilities.

USPS Copyright Disclosure

National ZIP®, ZIP+4®, Delivery Point Barcode Information, DPV, RDI, and NCOALink®. © United States Postal Service 2005. ZIP Code® and ZIP+4® are registered trademarks of the U.S. Postal Service.

DataFlux is a non-exclusive interface distributor of the United States Postal Service and holds a non-exclusive license from the United States Postal Service to publish and sell USPS CASS, DPV, and RDI information. This information is confidential and proprietary to the United States Postal Service. The price of these products is neither established, controlled, or approved by the United States Postal Service.

VMware Copyright Disclosure

VMware® virtual environment provided those products faithfully replicate the native hardware and provided the native hardware is one supported in the applicable DataFlux product documentation. All DataFlux technical support is provided under the terms of a written license agreement signed by the DataFlux customer.

The VMware virtual environment may affect certain functions in DataFlux products (for example, sizing and recommendations), and it may not be possible to fix all problems.

If DataFlux believes the virtualization layer is the root cause of an incident; the customer will be directed to contact the appropriate VMware support provider to resolve the VMware issue and DataFlux shall have no further obligation for the issue.

Recommended Reading

- *SAS/ACCESS for Relational Databases: Reference*
- *SAS Federation Server: Administrator's Guide*
- *SAS Event Stream Processing: User's Guide*
- *SAS Data Loader for Hadoop: User's Guide*
- *SAS Data Integration Studio: User's Guide*
- *DataFlux Data Management Studio User Guide*
- *SAS Quality Knowledge Base for Contact Information: Help*
- *DataFlux® Quality Knowledge Base for Product Data 2013A - User's Online Help*
- *SAS Data Remediation: User's Guide*
- *SAS Data Quality Accelerator for Teradata: User's Guide*
- *SAS Visual Process Orchestration: User's Guide*
- *DataFlux Web Studio: User's Guide*
- *SAS Business Data Network: User's Guide*
- *SAS Lineage: User's Guide*
- *SAS Data Remediation: User's Guide*
- *Cody's Data Cleaning Techniques Using SAS*
- *The Data Asset: How Smart Companies Govern Their Data for Business Success*
- *Data Quality for Analytics Using SAS*

For a complete list of SAS publications, go to sas.com/store/books. If you have questions about which titles you need, please contact a SAS Representative:

SAS Books

SAS Campus Drive

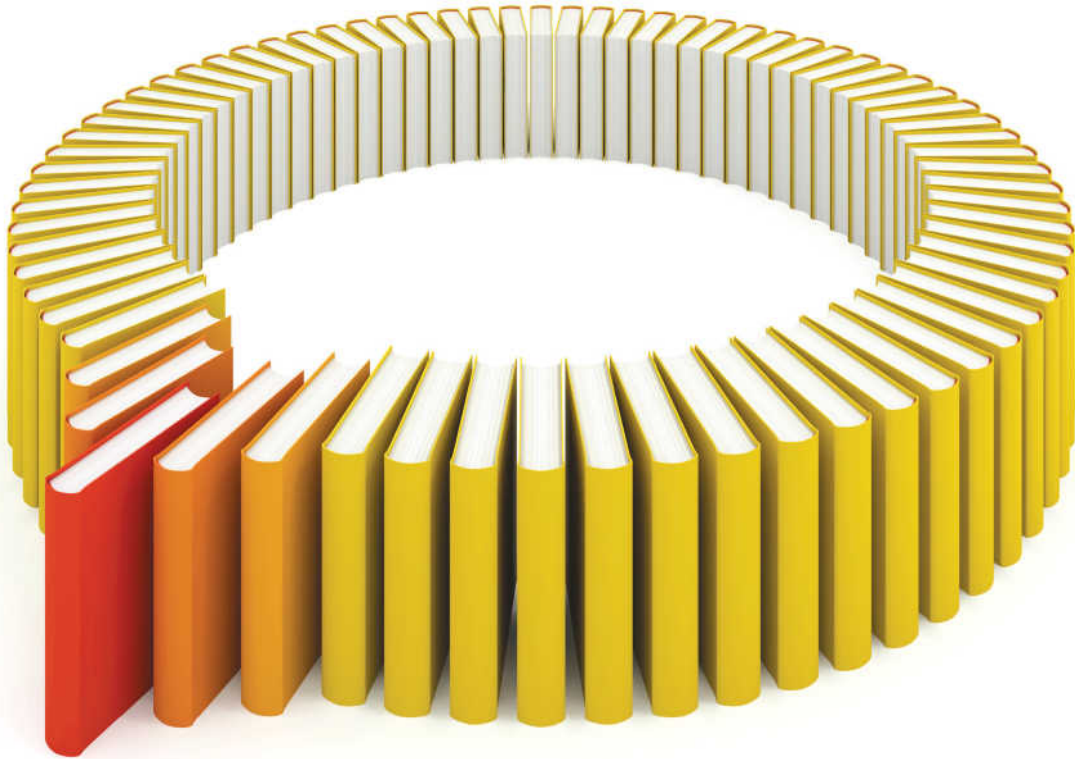
Cary, NC 27513-2414

Phone: 1-800-727-0025

Fax: 1-919-677-4444

Email: sasbook@sas.com

Web address: sas.com/store/books



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

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

