CHAPTER

1

Introduction to SAS Data Integration

About SAS Data Integration

Data integration is the process of consolidating data from a variety of sources in order to produce a unified view of the data. SAS supports data integration in the following ways:

□ *Connectivity and metadata.* A shared metadata environment provides consistent data definition across all data sources. SAS software enables you to connect to, acquire, store, and write data back to a variety of data stores, streams, applications, and systems on a variety of platforms and in many different environments. For example, you can manage information in Enterprise Resource

Planning (ERP) systems; relational database management systems (RDBMS), flat files, legacy systems, message queues, and XML.

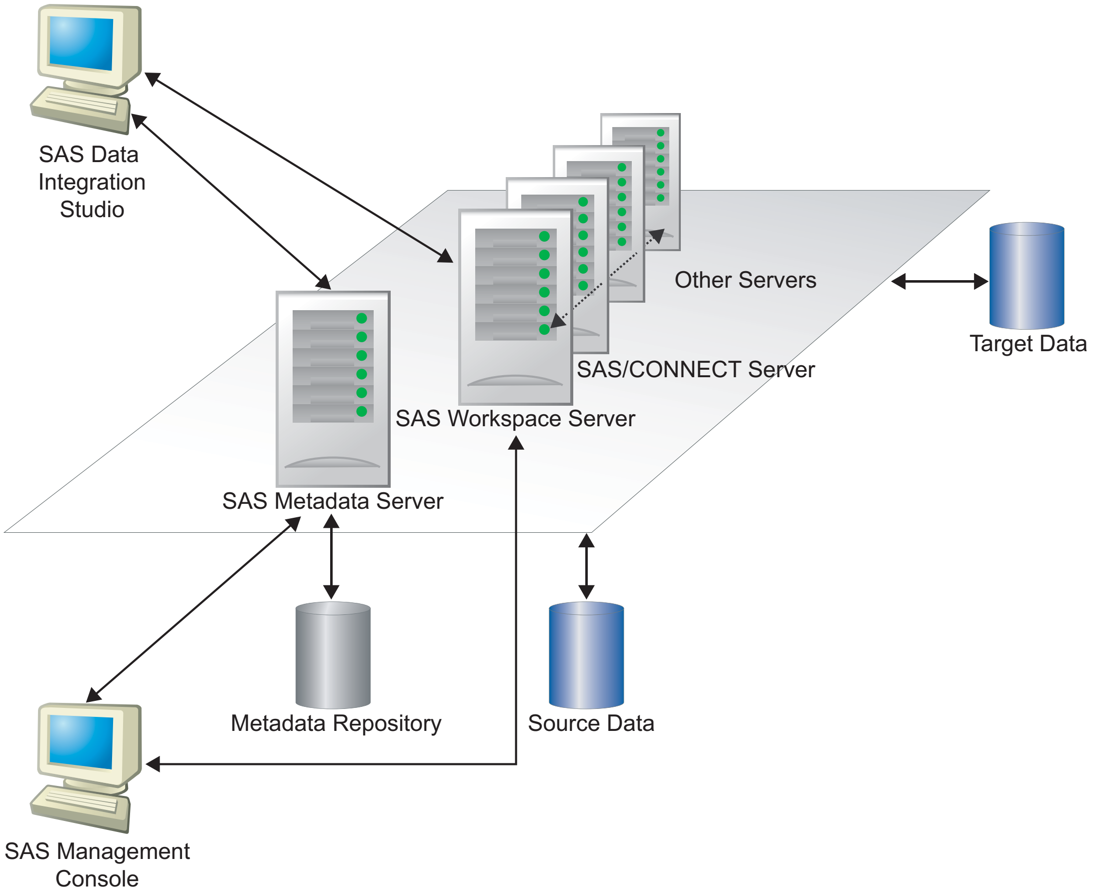
- □ *Data cleansing and enrichment*. Integrated SAS Data Quality software enables you to profile, cleanse, augment, and monitor data to create consistent, reliable information. SAS Data Integration Studio provides a number of transformations and functions that can improve the quality of your data.
- □ *Extraction, transformation, and loading (ETL)*. SAS Data Integration Studio enables you to extract, transform, and load data from across the enterprise to create consistent, accurate information. It provides a point-and-click interface that enables designers to build process flows, quickly identify inputs and outputs, and create business rules in metadata, all of which enable the rapid generation of data warehouses, data marts, and data streams.
- □ *Migration and synchronization*. SAS Data Integration Studio enables you to migrate, synchronize, and replicate data among different operational systems and data sources. Data transformations are available for altering, reformatting, and consolidating information. Real-time data quality integration allows data to be cleansed as it is being moved, replicated, or synchronized, and you can easily build a library of reusable business rules.
- □ *Data federation*. SAS Data Integration Studio enables you to query and use data across multiple systems without the physical movement of source data. It provides virtual access to database structures, ERP applications, legacy files, text, XML, message queues, and a host of other sources. It enables you to join data across these virtual data sources for real-time access and analysis. The semantic business metadata layer shields business staff from underlying data complexity.
- □ *Master data management*. SAS Data Integration Studio enables you to create a unified view of enterprise data from multiple sources. Semantic data descriptions of input and output data sources uniquely identify each instance of a business element (such as customer, product, and account) and standardize the master data model to provide a single source of truth. Transformations and embedded data quality processes ensure that master data is correct.

A Basic Data Integration Environment

Overview of a Data Integration Environment

The following figure shows the main clients and servers in a SAS data integration environment.

Figure 1.1 SAS Data Integration Studio Environment



Administrators use SAS Management Console to connect to a SAS Metadata Server. They enter metadata about servers, libraries, and other resources on your network and save this metadata to a repository. SAS Data Integration Studio users connect to the same metadata server and register any additional libraries and tables that they need. Then, they create process flows that read source tables and create target tables in physical storage.

SAS Management Console

SAS Management Console provides a single interface through which administrators can explore and manage metadata repositories. With this interface, administrators can efficiently set up system resources, manage user and group accounts, and administer security.

SAS Data Integration Studio

SAS Data Integration Studio is a visual design tool that enables you to consolidate and manage enterprise data from a variety of source systems, applications, and technologies. This software enables you to create process flows that accomplish the following tasks:

□ extract, transform, and load data for use in data warehouses and data marts

- □ cleanse, migrate, synchronize, replicate, and promote data for applications and business services

SAS Data Integration Studio enables you to create metadata that defines sources, targets, and the processes that connect them. This metadata is stored in one or more shareable repositories. SAS Data Integration Studio uses the metadata to generate or retrieve SAS code that reads sources and creates targets in physical storage. Other applications that share the same repositories can use the metadata to access the targets and use them as the basis for reports, queries, or analyses.

Through its metadata, SAS Data Integration Studio provides a single point of control for managing the following resources:

- □ data sources (from any platform that is accessible to SAS and from any format that is accessible to SAS)
- □ data targets (to any platform that is accessible to SAS, and to any format that is supported by SAS)
- □ processes that specify how data is extracted, transformed, and loaded from a source to a target
- □ jobs that organize a set of sources, targets, and processes (transformations)
- □ source code generated by SAS Data Integration Studio
- □ user-written source code

Note: SAS Data Integration Studio was formerly named SAS ETL Studio. △

Servers

SAS Application Servers

When the SAS Intelligence Platform was 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. If you have a SAS server, such as a SAS Workspace Server, on the same machine as your SAS Metadata Server, the application server object is named **`SASMain`**; otherwise, it 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—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, and a SAS/CONNECT Server, which can upload or download data and execute SAS code submitted from a remote machine.

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

Table 1.1 SAS Application Servers

Server	How Used	How Specified
SAS Metadata Server	Reads and writes metadata in a SAS Metadata Repository.	In each user's metadata profile.
SAS Workspace Server	Executes SAS code; reads and writes data.	As a component in a SAS Application Server object.
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.	As a component in a SAS Application Server object.
SAS OLAP Server	Creates cubes and processes queries against cubes.	As a component in a SAS Application Server object.
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.	As a component in a SAS Application Server object.
SAS Grid Server	Supports a compute grid that can execute grid-enabled jobs created in SAS Data Integration Studio.	As a component in a SAS Application Server object.

Typically, administrators install, start, and register SAS Application Server components. SAS Data Integration Studio users are told which SAS Application Server object to use.

SAS Data Servers

The following table lists two special-purpose servers for managing SAS data.

Table 1.2 SAS Data Servers

Server	How Used	How Specified
SAS/SHARE Server	Enables concurrent access of server libraries from multiple users.	In a SAS/SHARE library.
SAS Scalable Performance Data (SPD) Server	Provides parallel processing for large SAS data stores; provides a comprehensive security infrastructure, backup and restore utilities, and sophisticated administrative and tuning options.	In an SPD Server library.

Typically, administrators install, start, and register these servers and register the SAS/SHARE library or the SPD Server library. SAS Data Integration Studio users are told which library to use.

Database Management System (DBMS) Servers

SAS Data Integration Studio uses a SAS Application Server and a database server to access tables in database management systems such as Oracle and DB2.

When you start a source designer or a target designer, the wizard tries to connect to a SAS Application Server. You are then prompted to select an appropriate database library. SAS Data Integration Studio uses the metadata for the database library to generate a SAS/ACCESS LIBNAME statement, and the statement is submitted to the SAS Application Server for execution.

The SAS/ACCESS LIBNAME statement specifies options that are required to communicate with the relevant database server. The options are specific to the DBMS to which you are connecting. For example, here is a SAS/ACCESS LIBNAME statement that could be used to access an Oracle database:

```
libname mydb oracle user=admin1 pass=ad1min path='V2o7223.world'
```

Typically, administrators install, start, and register DBMS servers and register the DBMS libraries. SAS Data Integration Studio users are told which library to use.

Enterprise Resource Management Servers

Optional data surveyor wizards can be installed that provide access to the metadata and data from enterprise applications. Applications from vendors such as SAP, Oracle, PeopleSoft, and Siebel are supported. Typically, administrators install, start, and register ERP servers. SAS Data Integration Studio users are told which server metadata to use.

Libraries

In SAS software, a library is a collection of one or more files that are recognized by SAS and that are referenced and stored as a unit. Libraries are critical to SAS Data Integration Studio. You cannot begin to enter metadata for sources, targets, or jobs until the appropriate libraries have been registered in a metadata repository.

Accordingly, one of the first tasks in a SAS Data Integration Studio project is to specify metadata for the libraries that contain sources, targets, or other resources. At some sites, an administrator adds and maintains most of the libraries that are needed, and the administrator tells SAS Data Integration Studio users which libraries to use. The steps for specifying metadata about a Base SAS library are described in “Registering Any Libraries That You Need” on page 55.

Additional Information

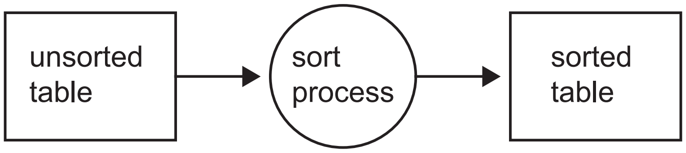
For more information about setting up a data integration environment, administrators should see “Administrative Documentation for SAS Data Integration Studio” on page 12.

Overview of Building a Process Flow

Problem

You want to become familiar with SAS Data Integration Studio, so you decide to create a simple process flow that reads data from a source table, sorts the data, and then writes the sorted data to a target table, as shown in the following figure.

Figure 1.2 Simple Process Flow



Solution

Create a job in SAS Data Integration Studio that specifies the desired process flow. Perform the following tasks:

□ Connect to a metadata server.

□ Register the source table.

□ Register the target table.

□ Create an empty job.

□ Drag and drop the SAS Sort transformation on the job.

□ Drag and drop the source table metadata and target table metadata on the job.

□ Update the metadata for the tables and the SAS Sort transformation as needed for your environment.

□ Execute the job.

It is assumed that administrators have installed, configured, and registered the relevant servers, libraries, and other resources that are required to support SAS Data Integration Studio in your environment.

Tasks

Connect to the Metadata Server

Most servers, data, and other resources on your network are not available to SAS Data Integration Studio until they are registered in a repository on a SAS Metadata Server. Accordingly, when you start SAS Data Integration Studio, you are prompted to select a metadata profile which specifies a connection to a metadata server. You might have a number of different profiles that connect to different metadata servers at your site. Select the profile that will connect to the metadata server with the metadata that you will need during the current session.

For details about creating a metadata profile, see “Connecting to a Metadata Server” on page 51.

Register Source Tables

Suppose that the source table in the example process flow is an existing SAS table, but the table is not currently registered; that is, metadata about this table has not been saved to the current metadata server. One way to register a table that exists in physical storage is to use a source designer wizard. To display the source designer wizard for a SAS table, select **Tools ▶ Source Designer** from the menu bar. A selection window displays. Click **SAS**, and then click **OK**. The SAS source designer displays. A source designer wizard enables you to:

- □ specify the library that contains the table to be registered (typically, this library has been registered ahead of time)
- □ display a list of tables contained in the selected library
- □ select one or more tables in that library
- □ generate and save metadata for the selected tables

For details about using source designers, see "Registering Tables with a Source Designer" on page 85.

Register Target Tables

Suppose that the target table in the example process flow is a new table, one that does not yet exist in physical storage. You could use the Target Table wizard to specify metadata for the table. Later, you can drag and drop this metadata on the target position in a process flow. When the process flow is executed, SAS Data Integration Studio will use the metadata for the target table to create a physical instance of that table.

One way to register a table that does not exist in physical storage is to use the Target Table wizard. To display the Target Table wizard, select **Tools ▶ Target Designer** from the menu bar. A selection window displays. Click **Target Table**, and then click **OK**. The Target Table wizard displays.

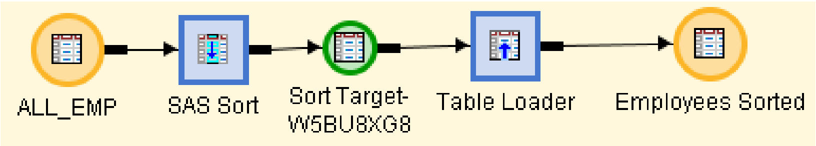
The Target Table wizard enables you to specify the physical location, column structure, and other attributes of the target table and save that metadata to the current repository.

For details about using the Target Table wizard, see "Registering Tables with the Target Table Wizard" on page 87.

Create a Job That Specifies the Desired Process Flow

In SAS Data Integration Studio, a process flow is contained in a job. One way to create a job is to use the New Job wizard to create an empty job, then drag and drop metadata for the source tables, the target tables, and the desired transformations onto the empty job, and build the desired process flow. For details about this method, see "Creating an Empty Job" on page 143. For now, assume that you have used this method to create the process flow shown in the following display.

Display 1.1 Process Flow Diagram for a Job That Sorts Data



Given the direction of the arrows in the previous display:

- □ ALL_EMP specifies metadata for the source table.
- □ SAS Sort specifies metadata for the sort process, which writes its output to a temporary output table, Sort Target-W5BU8XGB. (For more information about temporary output tables, see "Manage Temporary and Permanent Tables for Transformations" on page 239.)
- □ Table Loader specifies metadata for a process that reads the output from the previous step and loads this data into a target table.
- □ Employees Sorted specifies metadata for the target table.

SAS Data Integration Studio uses the preceding process flow diagram to generate SAS code that reads ALL_EMP, sorts this information, and writes the sorted information to a temporary output table. Then, the information is written to the Employees Sorted table.

Run the Job

One way to execute a SAS Data Integration Studio job is to select **Process ▶ Submit** from the menu bar. The code is then submitted to a SAS Application Server, which executes the code. If the job is successful, the output of the job is created or updated.

Next Tasks

The output from a job can become the source for another job in SAS Data Integration Studio, or it can be the source for a report or query in another SAS application. Any tables, libraries, or other resources that were registered in order to create the job are also available to other SAS applications that connected to the same metadata repository.

Impact of Change Management

The change management feature adds a few steps to some of the previous tasks. For more information, see "Working with Change Management" on page 59.

Advantages of SAS Data Integration

SAS data integration projects have a number of advantages over projects that rely heavily on custom code and multiple tools that are not well integrated.

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

Online Help for SAS Data Integration Studio

The online Help describes all windows in SAS Data Integration Studio, and it summarizes the main tasks that you can perform with the software. The Help includes examples for all source designer wizards, all target designer wizards, and all transformations in the Process Library. The Help also includes a What's New topic and a set of Usage Note topics for the current version of the software.

Perform the following steps to display the main Help window for SAS Data Integration Studio.

1 Start SAS Data Integration Studio as described in "Starting SAS Data Integration Studio" on page 50.

2 From the menu bar, select **Help ▶ Contents**. The main Help window displays.

To display the Help for an active window or tab, click its **`Help`** button. If the window or tab does not have a **`Help`** button, press the **`F1`** key.

To search for topics about concepts or features that are identified by specific words, such as "application server," display the main Help window. Then, click the **`Search`** tab (magnifying glass icon). Enter the text to be found and press the **`Enter`** key.

Administrative Documentation for SAS Data Integration Studio

Many administrative tasks, such as setting up the servers that are used to execute jobs, are performed outside of the SAS Data Integration Studio interface. Such tasks are described in SAS Intelligence Platform documentation, which can be found at the following location: **`http://support.sas.com/913administration`**.

The following table identifies the main SAS Intelligence Platform documentation for SAS Data Integration Studio.

Table 1.3 SAS Intelligence Platform Documentation for SAS Data Integration Studio

Administrative Task	Related Documentation
□ Set up metadata servers and metadata repositories.	*SAS Intelligence Platform: System Administration Guide*
□ Set up data servers and libraries for common data sources.	*SAS Intelligence Platform: Data Administration Guide*

Administrative Task	Related Documentation
□ Set up SAS Application Servers. □ Set up grid computing (so that jobs can execute on a grid).	*SAS Intelligence Platform: Application Server Administration Guide*
□ Set up change management. □ Manage operating system privileges on target tables (job outputs). □ Set up servers and libraries for remote data (multi-tier environments). □ Set up security for Custom tree folders. □ Set up a central repository for importing and exporting generated transformations. □ Set up support for message queue jobs. □ Set up support for Web service jobs and other stored process jobs. □ Enable the bulk-loading of data into target tables in a DBMS. □ Set up SAS Data Quality software. □ Set up support for job status handling. □ Set up support for FTP and HTTP access to external files.	*SAS Intelligence Platform: Desktop Application Administration Guide*

Accessibility Features in SAS Data Integration Studio

Accessibility Standards

SAS Data Integration Studio includes features that improve usability of the product for users with disabilities. These features are related to accessibility standards for electronic information technology that were adopted by the U.S. Government under Section 508 of the U.S. Rehabilitation Act of 1973, as amended. SAS Data Integration Studio supports Section 508 standards except as noted in the following table.

Table 1.4 Accessibility Exceptions

Section 508 Accessibility Criteria	Support Status	Explanation
(a) When software is designed to run on a system that has a keyboard, product functions shall be executable from a keyboard where the function itself or the result of performing a function can be discerned textually.	Supported with exceptions	The software supports keyboard equivalents for all user actions. Tree controls in the user interface can be individually managed and navigated through using the keyboard. However, some exceptions exist. Some ALT key shortcuts are not functional. Also, some more advanced manipulations require a mouse. Still, the basic functionality for displaying trees in the product is accessible from the keyboard. Based on guidance from the Access Board, keyboard access to drawing tasks does not appear to be required for compliance with Section 508 standards. Accordingly, keyboard access does not appear to be required for the **Process Editor** tab in the Process Designer window, or the **Designer** tab in the SQL Join properties window. Specifically, use of the **Process Editor** tab in the Process Flow Diagram and the **Designer** tab in the SQL Join Properties window are functions that cannot be discerned textually. Both involve choosing a drawing piece, dragging it into the workspace, and designing a flow. These tasks required a level of control that is provided by a pointing device. Moreover, the same result can be achieved by editing the source code for flows. **Example:** Use of the **Process Editor** tab in the Process Flow Diagram is designed for visual rather than textual manipulation. Therefore, it cannot be operated via keyboard. If you have difficulty using a mouse, then you can create process flows with user-written source code. See Chapter 12, "Working with User-Written Code," on page 215.
(c) A well-defined on-screen indication of the current focus shall be provided that moves among interactive interface elements as the input focus changes. The focus shall be programmatically exposed so that Assistive Technology can track focus and focus changes.	Supported with exceptions	In some wizards, when focus is on an element in a wizard pane, rather than a button, focus is not apparent. If an element in the pane is highlighted and focus is moved to a button, the element appearance is unchanged, so the user might not be certain when focus is on such an item. **Example:** When you launch the Target Designer and press the down arrow, you can traverse the Targets tree to select the type of target that you want to design even though no object has visual focus.

Section 508 Accessibility Criteria	**Support Status**	**Explanation**
(d) Sufficient information about a user interface element including the identity, operation, and state of the element shall be available to Assistive Technology. When an image represents a program element, the information conveyed by the image must also be available in text.	Supported with exceptions	In some wizards, identity, operation, and state of some interface elements is ambiguous. SAS currently plans to address this in a future release. **`Example:`** When you select a library in the Source Designer wizard, you must use the SAS Library combo box. If you are using the JAWS screen reader, the reader immediately reads not only the library name but also all of its details. If you want to know the libref, you must know that the label exists and that its shortcut is Alt+F. Then, you must press Alt+F so that the JAWS screen reader will read the label and its read-only text. You can move among the items in Library Details only after you use a shortcut to get to one of them.
(g) Applications shall not override user selected contrast and color selections and other individual display attributes.	Supported with exceptions	When the user sets the operating system settings to high contrast, some attributes of that setting are not inherited. **`Example:`** As with most other Java applications, system font settings are not inherited in the main application window. If you need larger fonts, consider using a screen magnifier.
(l) When electronic forms are used, the form shall allow people using Assistive Technology to access the information, field elements, and functionality required for completion and submission of the form, including all directions and cues.	Supported with exceptions	When navigating with a keyboard to choose a path in the Browse dialog box, the focus disappears. To work around the problem, either (1) count the number of times you press the TAB key and listen closely to the items, or (2) type the path explicitly. **`Example:`** In some wizards such as the Source Designer, the visual focus can disappear sometimes when you operate the software with only a keyboard. If so, continue to press the TAB key until an interface element regains focus.

If you have questions or concerns about the accessibility of SAS products, send e-mail to **`accessibility@sas.com`**.

Enabling Assistive Technologies

For instructions on how to configure SAS Data Integration Studio software so that assistive technologies will work with the application, see the information about downloading the Java Access Bridge in the section about accessibility features in the *SAS Intelligence Platform: Desktop Application Administration Guide*.