Paper SAS1907-2015

SAS® Data Management: Technology Options for Ensuring a Quality Journey Through the Data Management Process

Mark Craver, SAS Institute Inc., Cary, NC

ABSTRACT

When planning for a journey, one of the main goals is to get the best value possible. The same can be said for your corporate data as it journeys through the data management process. It is your goal to get the best data in the hands of decision makers in a timely fashion, with the lowest cost of ownership and the minimum number of obstacles. The SAS® Data Management suite of products provides you with many options for increasing the value of your data throughout the data management process. The purpose of this session is to focus on how the SAS® Data Management solution can be used to ensure the delivery of quality data, in the right format, to the right people, at the right time. The journey is yours, the technology is ours—together we can make it a fulfilling and rewarding experience.

INTRODUCTION

The SAS® Quality Knowledge Base is a robust collection of data cleansing algorithms that are applied during the data management process to control the quality of data throughout the process. These algorithms can be applied in many places in the data management process. Knowing the best place to introduce data quality into the data management process is key to successfully passing data through to the end business users.

The purpose of this paper is to discuss the SAS Data Management technology components that are available and the options (and benefits) for using them at different steps in the data management process. Through lecture and live demonstration, you will see a variety of options for applying the data quality and data cleansing algorithms to your data, and discuss the benefits and shortcomings of each.

This session includes discussions of interacting with the Quality Knowledge Base from within the following technology components:

- SAS® Data Management Studio batch jobs
- DataFlux® Data Management Server batch jobs
- DataFlux® Data Management Server real-time services
- SAS® Data Quality Server code
- SAS® Data Quality Accelerator for Hadoop

QUALITY KNOWLEDGE BASE OVERVIEW

The Quality Knowledge Base is a collection of files and definitions that are used in a variety of ways to ensure the cleanliness of data. The Quality Knowledge Base consists of a collection of files – schemes, chop tables, vocabularies, grammars, regular expression libraries, and phonetics libraries that are designed for a specific purpose and type of data.

The data cleansing algorithms in the Quality Knowledge Base is provided to the end user through what are known as “definitions”. Definitions are surfaced to the end user in a variety of ways, including through drop-down lists in the interfaces, SAS procedures, SAS functions (from within Data Step and SQL code), Data Integration Studio transformations, and more.

A definition is a set of steps for processing a piece of data, using a variety of the file components of the Quality Knowledge Base. The Quality Knowledge Base contains the definitions for the following:

- Data casing – formats data based on a given set of casing/capitalization rules.
- Gender analysis – determines the gender of a data string containing a person’s name.

- Identification analysis – determines the categorization of a data string.
- Language and locale guessing – guesses a language and/or locale from which a string might originate.
- Match coding and entity resolution – a context-specific matching algorithm that combines parsing, standardization, regular expression processing, and phonetics to identify potential duplicate records in and across database tables. Match codes can be used to household data, achieve uniqueness across records (survivorship), and in “fuzzy-logic” joins.
- Data parsing and extraction – identifies and categorizes words, or groups of words, within a text string.
- Pattern analysis - looks at input values not from an explicit examination of each character or word, but rather looks at those values with an eye to identifying patterns in the data.
- Data standardization – applies transformations to your data to ensure a standard representation of data values.

SAS QUALITY KNOWLEDGE BASE CONFIGURATION

In order to take advantage of the functionality in the Quality Knowledge Base, you need to configure the different technology components to the location of the Quality Knowledge Base. Figure 1 shows the possible configurations for the Data Management technology components to the Quality Knowledge Base.

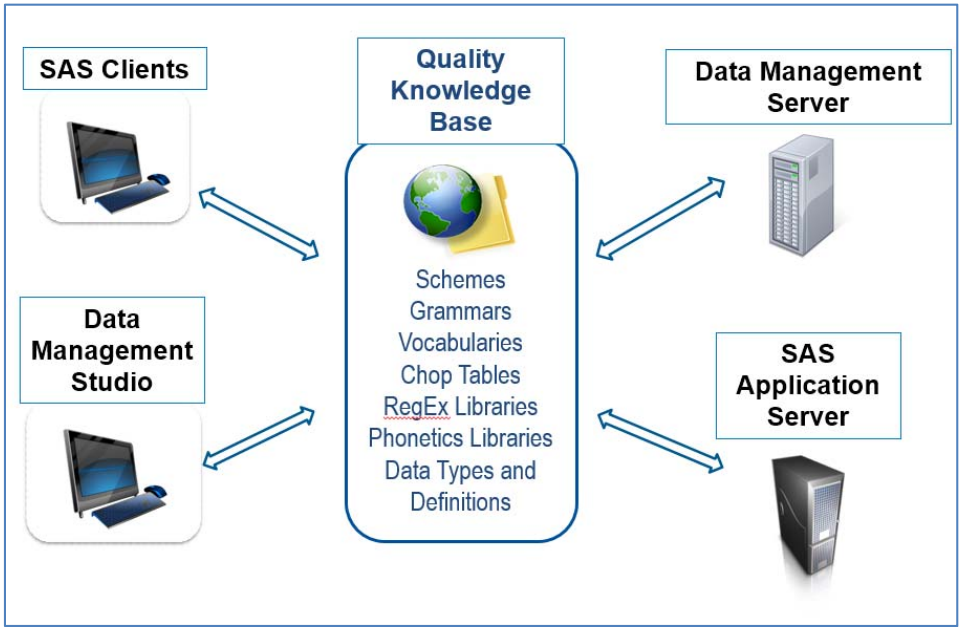

Figure 1. Quality Knowledge Base Configuration

Configuring SAS client components to the Quality Knowledge Base allows you to take advantage of the data cleansing definitions from within applications such as SAS® Data Integration Studio, SAS® Enterprise Guide®, SAS® Data Quality Accelerator for Hadoop (through SAS® Data Loader for Hadoop), and others.

For SAS code that access the components of the Quality Knowledge Base, and executing on the SAS Workspace Server, it is necessary for the server to be configured to the location of the Quality Knowledge Base.

Configuring DataFlux® Data Management Studio to the Quality Knowledge Base allows you to take advantage of the data cleansing definitions in reusable jobs and services stored in the DataFlux repository. Data Management Studio can be configured to interact with more than one Quality Knowledge

Base, although only one is registered as “active.” It is also the technology component used to make updates to the Quality Knowledge Base components, through the provided editors.

The jobs and services created in Data Management Studio can be uploaded to the repository on the DataFlux® Data Management Server, providing scalability, as well as real-time access to the jobs and services. In order to use this execution method, the Data Management Server also needs to be configured to the location of the Quality Knowledge Base.

DATA MANAGEMENT ADVANCED USE CASES

If you consider the fact that so many components in the SAS Data Management technology stack can be configured to the Quality Knowledge Base, this means you have a variety of options for applying the data cleansing algorithms to the data. In this section, we will explore several use cases for using the technology components in conjunction with the Quality Knowledge Base, as well as the benefits and shortcomings of each scenario.

SCENARIO 1: ALL PROCESSING TAKES PLACE IN A DATA MANAGEMENT STUDIO DATA JOB

One option you have for applying Quality Knowledge Base components to your source data is to use SAS Data Management Studio to do the processing. In this scenario, you would create a data job that accesses the source data, performs the necessary management/cleansing tasks, and writes the resulting data to one or more target tables, as shown in Figure 2.

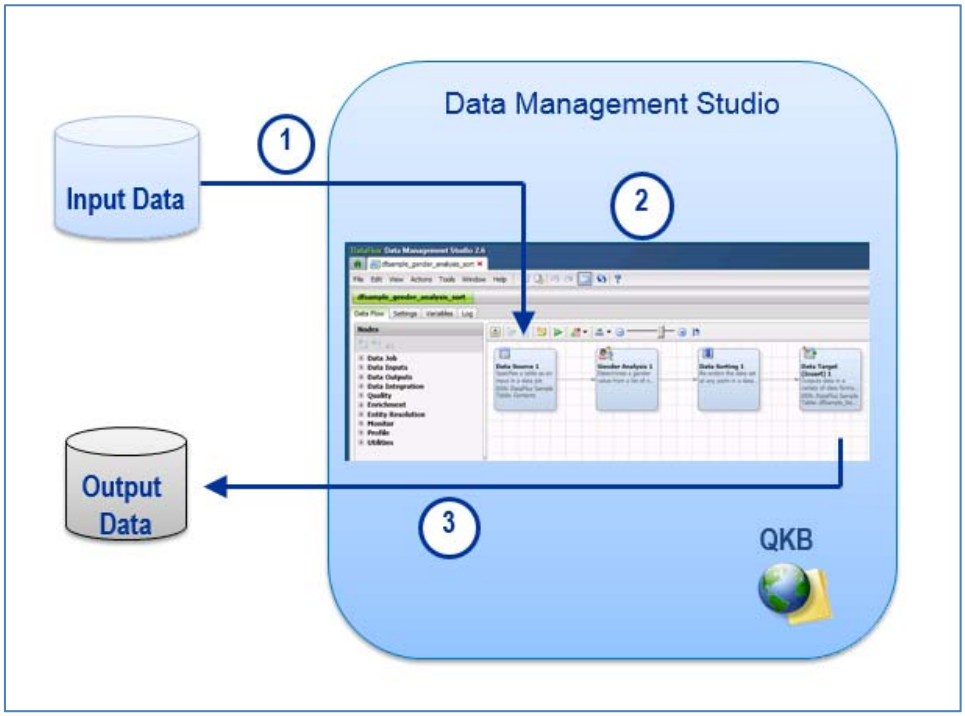

Figure 2. Data Access, Processing, and Data Output Executed in Data Management Studio

In this scenario, it is critical for Data Management Studio to be configured to the location of the Quality Knowledge Base as well as the input and output data storage locations.

There are many benefits to processing data in this manner, including the following:

- The interface is very easy to use, and suitable for a non-technical user.
- The user does not have to be a SAS programmer, and does not see any code.
- The user has complete ownership of the process, and that frees them from their dependence on IT.

- IT does not have to get involved with “data stewardship” tasks.
- The user has the option of previewing the data at each step along the way, and adjust the process if necessary.
- The jobs and services created by the user are reusable assets that can be deployed to the Data Management Server for scalability and real-time access.

There are also some shortcomings to this scenario, including the following:

- Data Management Studio is a “fat client” application, and needs to be installed locally on the user’s machine.
- The data must be brought down to the client machine for processing, making it impractical for larger volumes of data.
- Processing large volumes of data may overwhelm the client machine.
- Users must learn a new interface and understand the components of the Quality Knowledge Base.
- There is a limit to the number of data sources that can be read in to Data Management Studio.
- This process is not documented in SAS metadata.

SCENARIO 2: ALL PROCESSING IS HANDLED PROGRAMATICALLY WITH SAS DATA QUALITY SERVER CODE

One option for applying the data cleansing processes to your data is to use SAS® Data Quality Server code in SAS programs. This option allows you to execute SAS code on a SAS server process to apply data cleansing algorithms from the Quality Knowledge Base to your data.

Figure 3 shows an example of SAS code that takes advantage of the %DQLOAD macro to access the components of the Quality Knowledge Base:

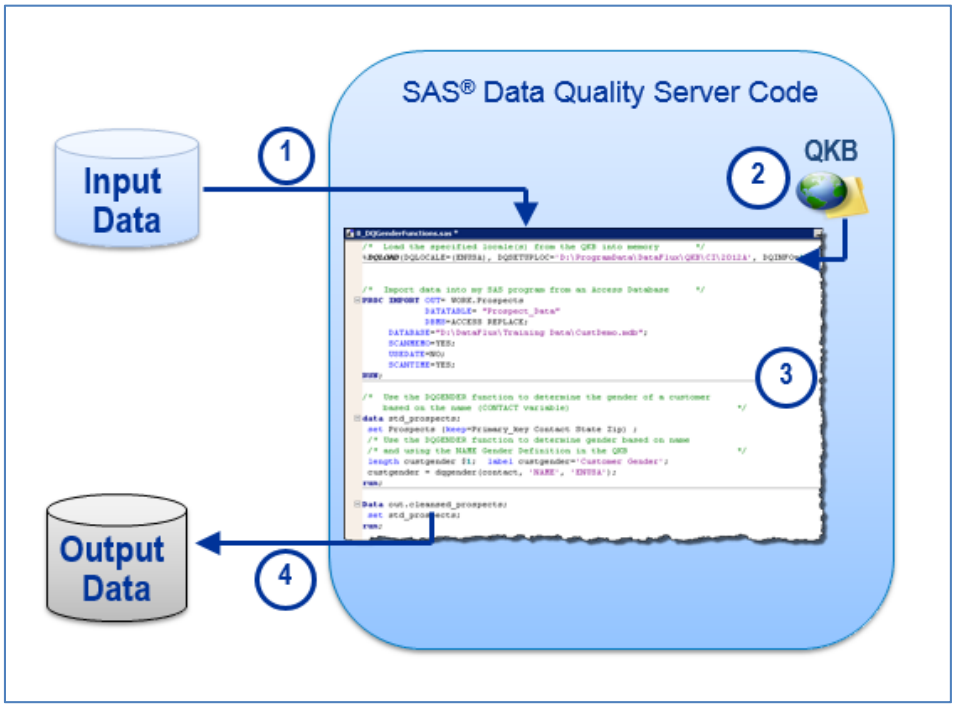

Figure 3. Data Access, Processing, and Data Output Executed in SAS Data Quality Server Code

This scenario requires that the SAS server process executing the code be configured to the location of the Quality Knowledge Base. This task is accomplished in one of the following three ways:

- Use options in the SAS configuration files for the SAS Workspace Server to reference the root installation path of the Quality Knowledge Base. This option makes the Quality Knowledge Base components available to any SAS client or web applications that are interacting with the SAS Workspace Server.
- Use the %DQLOAD macro, which is part of the Data Quality Server product, to load the Quality Knowledge Base into memory in the SAS session. This option makes the data cleansing algorithms available through the use of SAS procedures, functions, and call routines.
- Use the options in the configuration files for the local SAS session (for example, SAS Display Manager) to reference the root installation path of the SAS Quality Knowledge Base. This option makes the Quality Knowledge Base components available anytime you run a batch SAS program on the machine, or anytime you execute a batch in SAS Display Manager.

There are many benefits to accessing the Quality Knowledge Base in from within SAS code, including the following:

- The user does not have to learn to use the Data Management Studio interface.
- The user does not need Data Management Studio on the client machine in order to use the components of the Quality Knowledge Base.
- There is no separate repository of jobs and services to maintain.
- SAS code can access the production Quality Knowledge Base as other applications are accessing.
- It takes advantage of performance/scalability of the SAS Workspace Server (if it is configured as part of the SAS platform).
- It integrates easily with existing Data Integration Studio jobs, SAS Enterprise Guide jobs, as well as new and existing SAS batch programs.
- It allows you to move the process closer to where the data is, without having to move the data down to the client machine.

There are also some shortcomings to this scenario, including the following:

- The user must be familiar with how the Quality Knowledge Base components are accessed in SAS Data Quality Server code.
- The user must be a SAS programmer.
- The user must know the exact name of the components being invoked from the Quality Knowledge Base, because there are very few places where there are drop-down lists to select from.

  Note: There is a %DQPUTLOC macro that users can use in SAS code to write the Quality Knowledge Base definition names out to the SAS log. This macro shows you the exact definition name to be used in the SAS code.
- The user has no easy-to-use editors for the components of the Quality Knowledge Base if changes are needed.

SCENARIO 3: SAS CODE CALLS A REAL-TIME DATA SERVICE AND STREAMS DATA TO THE SERVICE FOR PROCESSING

Another option for improving the consistency and cleanliness of your data is to use SAS Data Quality Server code in SAS programs to stream data to real-time services on the Data Management Server. There are SAS procedures and functions available for interacting with the Data Management Server, allowing you to perform tasks such as:

- Execute batch jobs on the server
- Call real-time services on the server
- Check the status of jobs executing on the server
- Kill jobs that are running on the server
- Copy process logs from the server back to the local machine

In this scenario, SAS code is used to access and preprocess the data, then SAS Data Quality Server code is used to pass data to the Data Management Server for processing using algorithms in the Quality Knowledge Base. After completing the process, Data Management Server sends the data back to the SAS session for further processing. Figure 4 shows an example of SAS Data Quality Server code passing data into a real-time service on the Data Management Server.

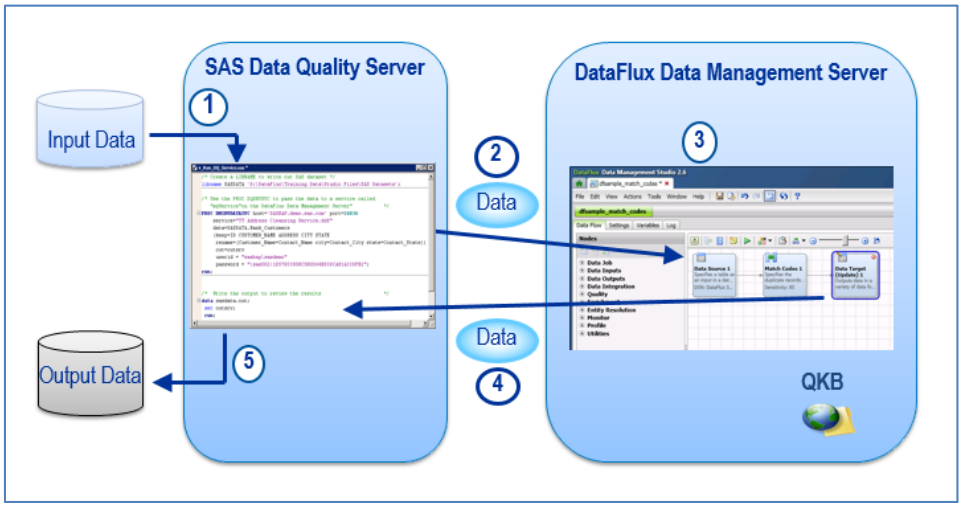

Figure 4. Data Access and Data Output Handled Programmatically in SAS, Data Passed to DataFlux Data Management Server for Processing, and Then Streamed Back to the SAS Program

This option requires that the Data Management Server is configured to the Quality Knowledge Base. It is not necessary for the SAS components to be configured to the Quality Knowledge Base, since all of the Quality Knowledge Base interaction takes place on the Data Management Server.

The benefits to this configuration include the following:

- The data cleansing service can be built and maintained by a data quality steward in the business unit who knows the data and knows what improvements need to be made to the data. All the SAS programmer needs to know is the name of the real-time service to pass the data to.
- The real-time service is a reusable asset that can be called in a variety of ways and used in conjunction with a variety of source data tables.
- You can insert a real-time service call into existing SAS jobs to send the data out of process and improve the consistency and cleanliness of the data.

- You can improve the quality and consistency of your data as early as possible in the process because you can insert real-time service calls anywhere it is appropriate in your existing SAS code.
- There are transformations in SAS Data Integration Studio that are designed and configured to interact with the Data Management Server and communicate via SOAP commands. These transforms connect to the Data Management Server, access the data services from the repository on the server, return a list of service names, expected input fields, output fields, and other items required for interacting with the service.

Shortcomings to this approach include the following:

- With large volumes of data, the network connection can prove to be a bottleneck.
- If the Data Management Server is down for some reason, then you cannot connect to the server for processing.
- For a secure Data Management Server, you must be sure that the user who is trying to access the server, has the appropriate level of permissions to pass to the server.

SCENARIO 4: SAS CODE KICKS OFF A DATA JOB ON THE DATA MANAGEMENT SERVER, AND RETURNS THE JOB STATUS TO SAS PROCESS

SAS Data Quality Server code can also be used to kick off a data job on the Data Management Server. The data job is responsible for accessing the source data, processing the data using algorithms from the Quality Knowledge Base, and then writing the output data to the target table. Figure 5 shows an example of SAS Data Quality Server code kicking off a data job and returning the status of the job back to SAS.

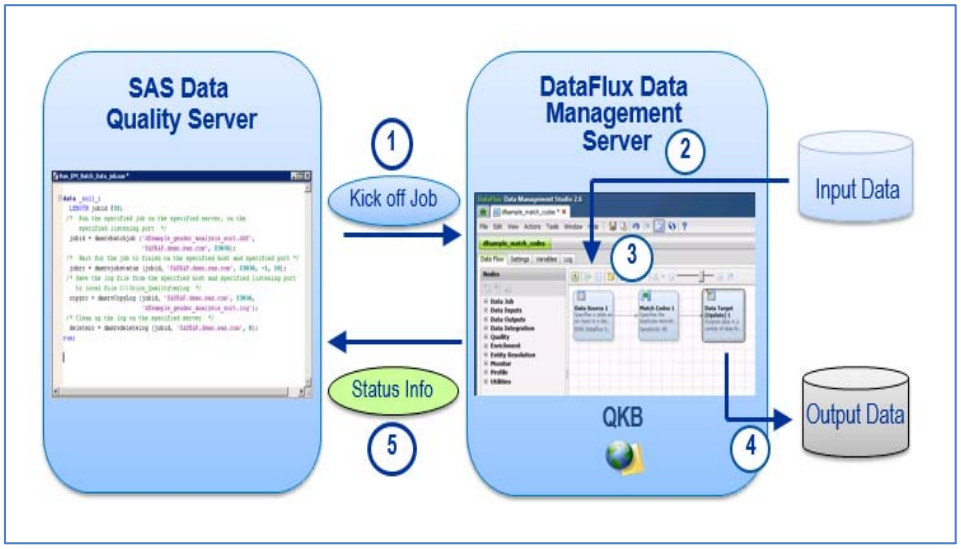

Figure 5. SAS Data Quality Server Code Kicks Off a Job, DataFlux Data Management Server Accesses the Data, Processes It, and Writes It to the Output Location, Sending the Status Information Back to the SAS Program

Since this option takes advantage of a data job on the Data Management Server, there is no need to configure SAS to the Quality Knowledge Base. In addition, since a data job has a specific input data source and also creates its own target data tables, all data access is handled on the Data Management Server. The purpose of the SAS code is to orchestrate the execution of the job and check the status before continuing with other processes.

The benefits of this scenario are the following:

- The data quality stewards in the business unit can create and maintain the data job that gets executed on the Data Management Server and not have to rely on IT to build the job for them.
- The SAS programmer can take advantage of the job to improve the quality of the data, without having to know how to create the job itself.
- The SAS programmer can monitor the status of the job as it runs and take appropriate action downstream based on the status returned for the job.
- You have the flexibility of separating out the data cleansing process into an external job and only executing that job if indicated by some trigger mechanism.

Potential shortcomings to this approach include:

- The SAS program needs to include credentials to pass to the Data Management Server.
- The programmer needs the appropriate permissions to perform the requested tasks on the server.
- The SAS programmer has no control over what takes place in the data job on the Data Management Server, and will not be able to debug the process if there is an issue in the execution of the data job.

SCENARIO 5: SAS CODE CREATES STAGING TABLES, THEN KICKS OFF A DATA JOB THAT PROCESSES AND CREATES SECOND SET OF STAGING TABLES, AND RETURNS STATUS TO SAS PROGRAM

One option for applying data cleansing processes to your data is to use a combination of SAS Data Quality Server code in conjunction with data jobs on the Data Management Server to ensure the consistency of your data. This scenario uses staging tables to provide the added benefit of tracking the progress after each step. It also gives you more flexibility if you need to restart a portion of the process somewhere in the middle. Figure 6 shows an example of SAS code that accesses data from an input data source, processes it, and creates a staging data table. That staging table is then read into a data job on Data Management Server, kicked off by the SAS process. The data job reads in the staging table, processes the data, writes it out to another staging table, and then sends a status back to the SAS program. The SAS program then reads in the second staging table, and continues processing, ultimately writing data to an output data source.

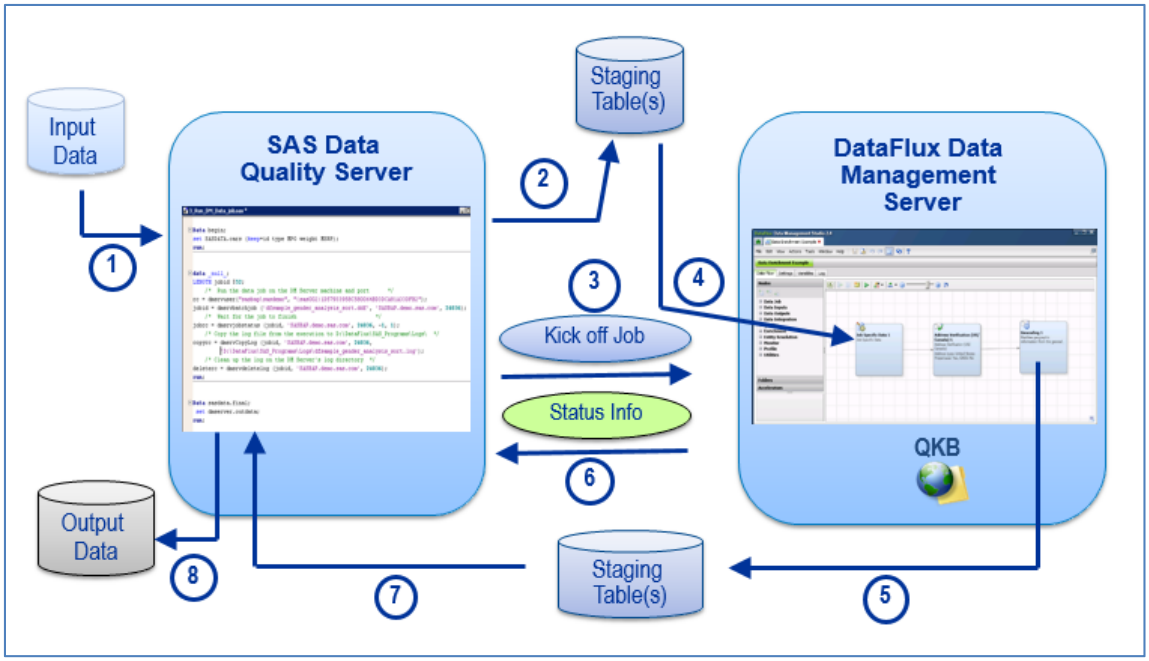

Figure 6. A SAS Program Creates a Staging Table and SAS Data Quality Server Code Kicks Off a Job. Then, DataFlux Data Management Server Accesses the Data, Processes It, and Writes It to a Staging Table, Sending the Status Information Back to the SAS Program. The SAS program Then Accesses the Staging Table and Continues Processing and Writes Data to the Output Table.

In this scenario, SAS is used for data access and orchestrating the execution of the data management process. Staging tables are written at different stages in the process, and the data is read in to the next step in the process. Since the data cleansing process takes place via a data job on the Data Management Server, the server is the only thing that needs to be configured to the Quality Knowledge Base.

Benefits to this approach include the following:

- You are not passing data to a service, so the network will not be a bottleneck.
- You can create staging tables at different points in the process. This option gives you the ability to identify and resolve issues earlier and in a more timely fashion.
- You can perform quality checks on the data in the staging tables to see if any additional processes need to be put in place.
- You can restart the staging table's job, if needed.
- You can use schedulers to check for the existence of the staging table before proceeding with the next step in the process.
- You can use the status that is returned to the SAS job to ensure things are in a good state before proceeding.

Potential drawbacks to this approach include the following:

- You are storing multiple copies of the data in the staging tables, so disk space could become an issue.
- Processes are slowed down by the disk I/O.
- If the Data Management Server is down, then this could keep the remainder of your SAS code from executing.

SCENARIO 6: SQL CODE IS USED TO INVOKE DATA QUALITY FUNCTIONALITY WITHIN THE DATBASE

A relatively new option for applying data cleansing processes to your data is to use the SAS Data Quality Accelerators to cleanse the data inside the database. This provides you the flexibility and scalability of not having to pull the data out of the database and move it to SAS for processing. The Data Quality Accelerators take advantage of the Quality Knowledge Base that has been deployed inside the database, allowing you to programmatically invoke the algorithms inside the database. Figure 7 shows an example of the architecture of the SAS Data Quality Accelerator for Hadoop.

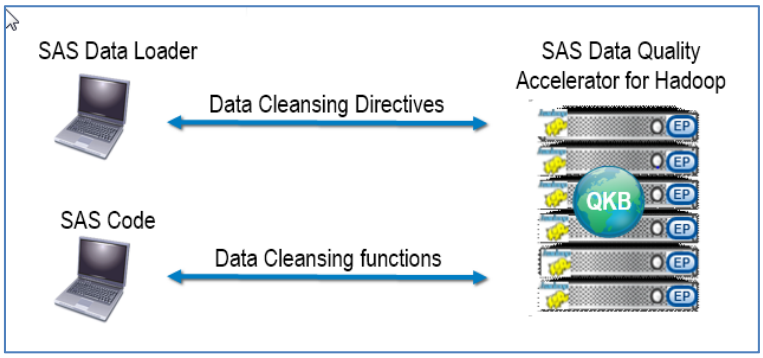

Figure 7. SAS Data Quality Accelerator for Hadoop Deployed into the Hadoop Cluster, Accessed By Data Cleansing Directives in SAS Data Loader, or from Within SAS Programs

In order to use this option for exploiting the functionality of the Quality Knowledge Base, the Quality Knowledge Base must first be deployed into the database (in this case, to the Hadoop cluster). Once the Quality Knowledge Base is deployed as a schema inside the database, you can access the data quality functions from within SAS code.

The code snippet below shows an example of SAS code is used to access the data cleansing functions inside a Hadoop cluster:

```
/* Create a libname to the Hadoop cluster with the appropriate options  */
/* for interacting with the Hadoop cluster                              */

libname WSTFD7 HADOOP port=10000 schema=default subprotocol=Hive2
 transcode_fail=warning host="192.168.74.134" user=sasdemo dbmax_text=50;

/* Use PROC DS2 code to execute functions in the Hadoop cluster to    */
/* standardize country and state columns using standardization        */
/* definitions from the QKB                                           */

proc ds2 bypartition=yes ds2accel=yes;
 thread t_pgm / overwrite=yes;
    dcl package dq dq();
    dcl varchar(3) country_std;
    dcl varchar(15) state_std;
    method init();
       dq.loadlocale('ENUSA');
    end;
   method run();

 set WSTFD7.furniture_sales;
  country_std = dq.standardize('Country (ISO 3 Char)', country);
  state_std = dq.standardize('State/Province (Abbreviation)', state);
 output;
 end;
endthread;
```

```
/* Use a DATA step to write the data to a table in the Hadoop cluster */

data WSTHVJ.furniture_sales_std (overwrite=yes);
 declare thread t_pgm t;
 method run();
  set from t;
 end;
enddata;

run;
quit;
```

The benefits of using the SAS Data Quality Accelerator to cleanse your data include the following:

- No movement of data from the database to SAS for processing.
- Much more efficient and scalable for data cleansing processes.
- Uses the database investment and infrastructure already in place.
- The Quality Knowledge Base is installed in a schema inside the Hadoop cluster, you do not have to pass files into the cluster from within your code.
- The functions to access the components of the Quality Knowledge Base are already deployed into the Hadoop cluster as part of the SAS Data Quality Accelerator deployment process.

Possible shortcomings of this approach include the following:

- Extra deployment steps are required to get the Quality Knowledge Base installed inside the database.
- Customizations to the Quality Knowledge Base require it to be redeployed into the database.
- Each time a new Quality Knowledge Base release is available, it needs to be redeployed into the database.
- The programmer must know how to write SAS DS2 code to access the functions in the Hadoop cluster.

  Note: The SAS® Data Loader for Hadoop product provides a point-and-click directive that builds the PROC DS2 code for you.
- All of the functionality in the Quality Knowledge Base is not available via functions “out-of-the-box.”

CONCLUSION

The SAS Quality Knowledge Base contains a robust set of functionality for ensuring the cleanliness and consistency of your data as it journeys through the data management process. With the option of applying that functionality in a variety of ways, and at various points in the data management process, you have an incredible amount of flexibility in how you use the Quality Knowledge Base components to improve your data. This enables you to greatly increase the business value to your data as it makes its journey from source systems and into the hands of decision makers.

In today’s competitive market, your ability to creatively use the components of the Quality Knowledge Base could give you the competitive edge that you seek. SAS gives you many options for ensuring the quality of your corporate data asset as it makes its way into the hands of decision makers. The journey is yours… the technology is ours… together, we can make it a fulfilling and rewarding experience!

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Mark Craver
100 SAS Campus Drive
Cary, NC 27513
SAS Institute Inc.
919-531-6702
Mark.Craver@sas.com
http://www.sas.com