Paper 070-2013

What's New in SAS® Data Management

Nancy Rausch, SAS Institute Inc., Cary, NC; Malcolm Alexander, SAS Institute Inc., Cary, NC

ABSTRACT

The latest releases of SAS® Data Integration Studio and DataFlux® Data Management Studio provide an integrated environment for managing and transforming your data to meet new and increasingly complex data management challenges. The enhancements help develop efficient processes that can cleanse, standardize, transform, master, and manage your data. Latest features include capabilities for building complex job processes; new web-based development and job monitoring environments; enhanced Extract/Load/Transform (ELT) transformation capabilities; big data transformation capabilities for Hadoop; integration with the SAS® LASR™ platform; enhanced features for lineage tracing and impact analysis; and new features for master data and metadata management. This paper provides an overview of the latest features of the products and includes use cases and examples of the product capabilities.

INTRODUCTION

The latest releases of SAS® Data Integration Studio, DataFlux® Data Management Studio, and other SAS Data Management features provide many new enhancements to help both data warehouse developers and data integration specialists carry out data-oriented tasks more efficiently and with greater control and flexibility. Major focus areas for the release include a new integrated console and enhanced authoring environments; features in support of big data and cloud computing; features in support of quality, collaboration, governance, and new monitoring features. This paper will showcase some of the newest features available in the SAS® Data Management products.

INTRODUCING THE DATA MANAGEMENT CONSOLE

The SAS Data Management Console shown in Figure 1 is a new web client that supports a customizable user experience for many data management activities including data remediation, workflow management, job authoring, locking and versioning, scheduling, monitoring, collaboration, and others. The user experience is based on roles and capabilities that can be customized to meet site specific needs. There are at-a-glance features such as portlets that can provide up-to-date information on data management activities in your enterprise, and alert you to problem areas quickly so that you can respond to them. The console is fully integrated with the SAS web infrastructure platform. Many data management features are already integrated and can be launched from this new interface, and additional product components will be integrated in future releases.

Figure 1. SAS® Data Management Console

BUSINESS DATA NETWORK – DATA DICTIONARY USE CASE

The Business Data Network, a data management feature that is part of DataFlux Data Management Web Studio and is also launch-able from the new SAS Data management console, supports collaboration of knowledge between those that manage data. It provides a single entry point for data producers and consumers to document their data. The Business Data Network can provide information about data for a wide range of scenarios. One common use

case for the Business Data Network is that of a data dictionary. A data dictionary captures knowledge about data in an enterprise so that consumers of the data are able to understand the data: what it means and how to use it. The Business Data Network is ideally suited to this use case because it supports documenting data details and provides useful features such as email notification about change, and information about the state of the data. Figure 2 is an example of this use case.

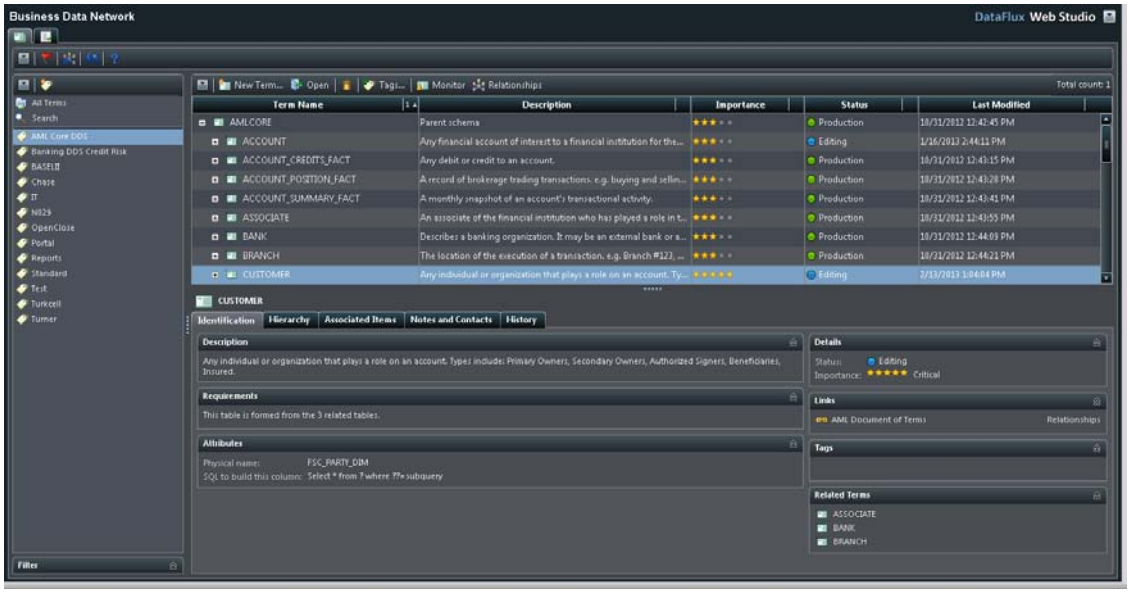

Figure 2. Business Data Network Data Dictionary

The primary data item in the Business Data Network is a Term. A Term is a descriptive piece of information that consists of a name, description, requirements, related objects such as documents, rules, jobs, applications, tables and columns, and other information. Figure 3 shows an example Data Dictionary for the SAS anti-money laundering solution, highlighting the Customer table which is one of the tables in the data model for the solution.

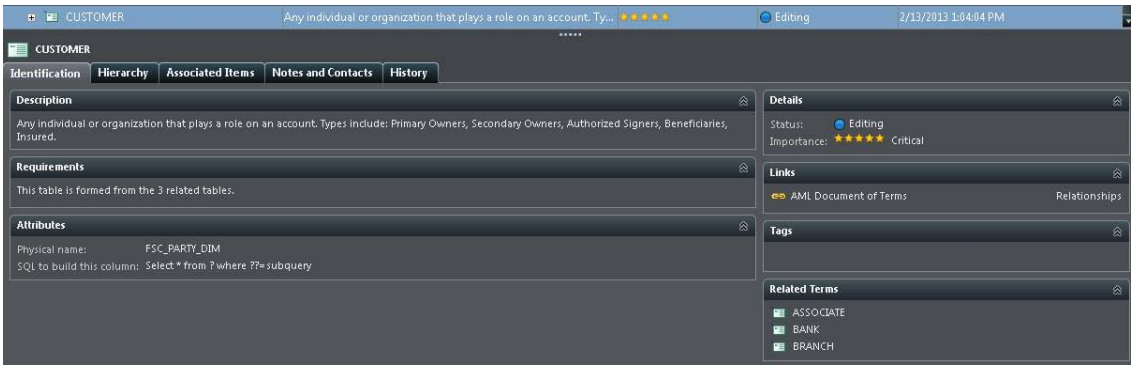

Figure 3. Customer Table in the Data Dictionary

The Customer term contains information about the importance of this table, related tables, and specific details about how to implement the table. Physical objects such as source systems that provide the data to create the table can be linked to the term. Collaboration between those that are responsible for the data contained in the table and those that consume the data is supported on the Notes and Contacts tab. Full versioning of the data model and history of changes is supported and displayed on the History tab. Some of these features are illustrated in Figure 4.

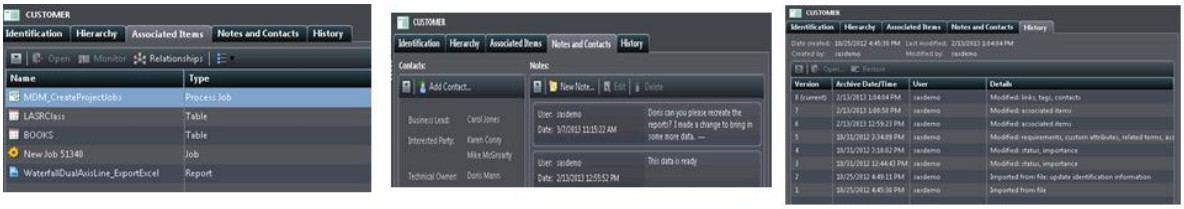

Figure 4. Details of the Customer Table are Stored in the Data Dictionary

The Reference Data Manager component of DataFlux Data Management Web Studio is integrated in with the Business Data Network. It can help you capture and record valid data values for terms. In the example data dictionary, these reference values could represent lookup codes that are valid for the data contained in the table. This is illustrated in Figure 5.

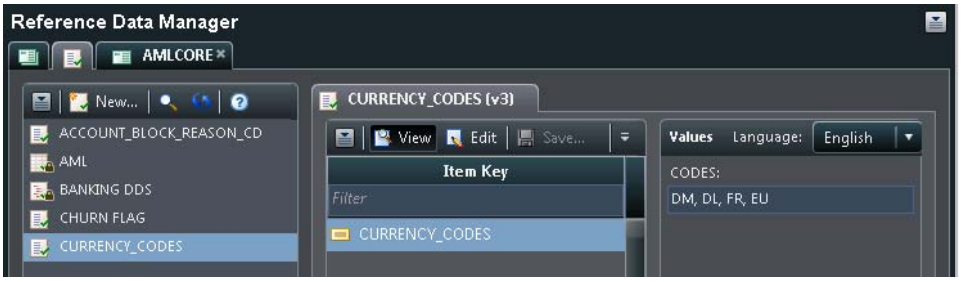

Figure 5. Reference Data Manager Can Capture Valid Values for Tables and Columns in the Data Dictionary

Content to populate the Business Data Network can be imported from a variety of sources. The import format is XML-based, and the product ships with samples to help you format and import your own content.

LINEAGE AND IMPACT ANALYSIS

A new web-based lineage and impact analysis feature is now available. In the Business Data Network example above, it can be used to show how all of the pieces in a data model fit together as shown in Figure 6. The lineage viewer supports many different types of relationships in its view including data dependencies, associated objects, synonyms, and other object relationships. The viewer can include relationships to third-party objects such as Microsoft Excel spreadsheets or documentation. It has an extensible object set so that additional objects can be incorporated into the view.

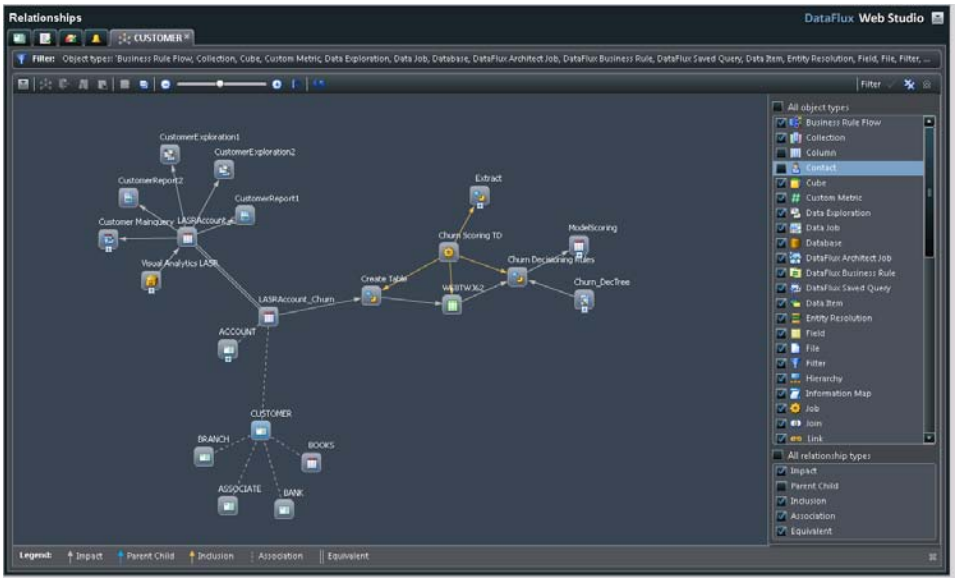

Figure 6. New Impact Analysis Viewer

Many object types are already available in the viewer and it supports filtering so you can pinpoint exactly the

information you are interested in. An extractor in SAS Data Integration Studio and DataFlux Data Management Server is available to extract lineage information into the new lineage viewer. The extractors can be scheduled to be run on a periodic basis sending data to a web service in the viewer which is listening for updated lineage information.

From SAS Data Integration Studio, you can directly launch the web lineage viewer from any table in the Folders tree.

BUSINESS RULES

Business rules de-couple business logic from application logic. Business rules are essentially a set of logic that can be stored independently of data, and reused across multiple data environments to apply the business logic to the application data. SAS Data Management offers several products tailored to specific use cases that enable you to author business rules and then apply them in your jobs to validate and cleanse your data.

SAS® Business Rules Manager, shown in Figure 7, is a web-based rule authoring user interface for authoring business rules for the SAS language. It supports development, simulation, management, and monitoring of rules; a rules engine for supporting batch, real-time, and service execution modes; support for integrating mining models for scoring; and a transformation in SAS Data Integration Studio for deploying rule packages into your jobs. The product allows you to author, validate, and test your rules, and then deploy them as packages into the SAS® Metadata Server. Rules packages are fully versioned so that you can work on future versions without affecting your production jobs.

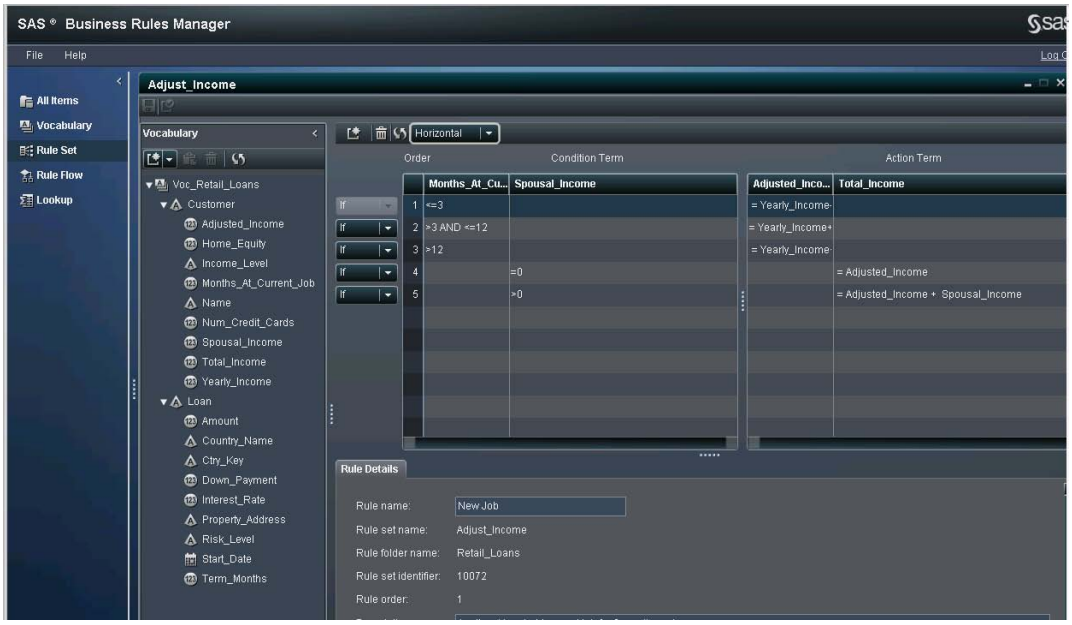

Figure 7. SAS Business Rules Manager

Figure 8 shows an example of using SAS Business Rules Manager rules in SAS Data Integration Studio. The rules are published and appear in the SAS Metadata Folders tree. SAS Data Integration Studio has a Business Rules transform that understands rules packages. The node allows you to map source data to rules and capture rules statistics in the output tables.

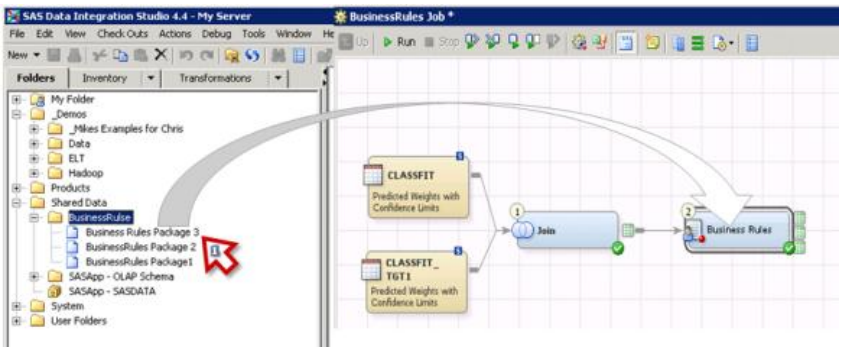

Figure 8. SAS Business Rules Manager Transform in SAS Data Integration Studio

Data Quality rules can also be built and monitored to capture data quality problems in data flows. The Business Rules editor in DataFlux Data Management Studio supports building and deploying data quality business rules. The editor is shown in Figure 9. Data Quality rules checks can be embedded in SAS Data Integration Studio flows by using the DataFlux service nodes available in the transforms library.

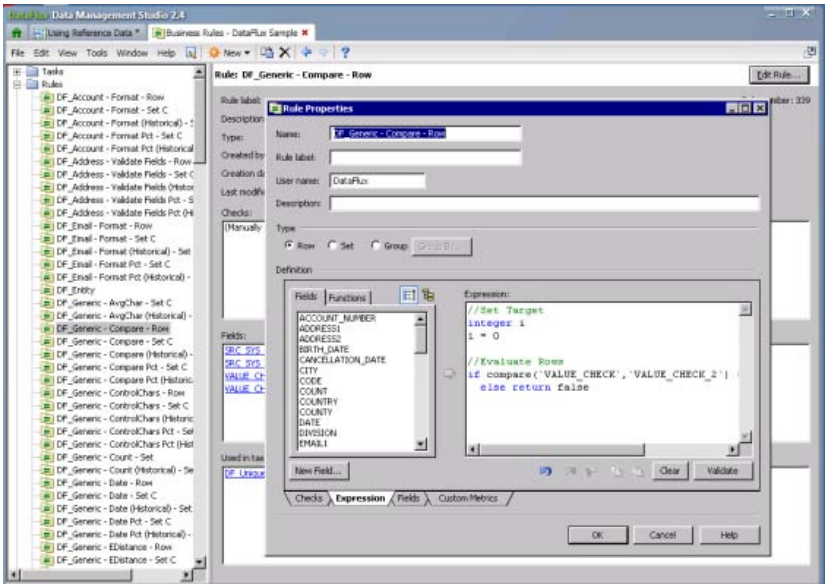

Figure 9. Data Quality Business Rules Editor

JOB ORCHESTRATION

As data volumes grow, jobs that gather, transform, and manage data need to become increasingly complex to handle the new performance challenges. In addition, data integrators increasingly have to manage content, including other jobs, coming from diverse systems and sources. The new Job Orchestration feature in SAS Data Management Console is designed to help data integrators better manage their job flows. It offers a job authoring and run time environment to create jobs that can orchestrate all sorts of other jobs from SAS code to SQL scripts to web services.

The Job Orchestration feature is shown in Figure 10 below. It provides a new web authoring environment that is launched from the SAS Data Management Console. The authoring environment supports a number of nodes that can be used to build jobs that run other jobs. It supports parallelization of nested jobs, control logic such as IF/THEN/ELSE handling and looping, event management, error checking, and run time statistics for each embedded node. The Job Orchestration feature is fully integrated with the SAS platform.

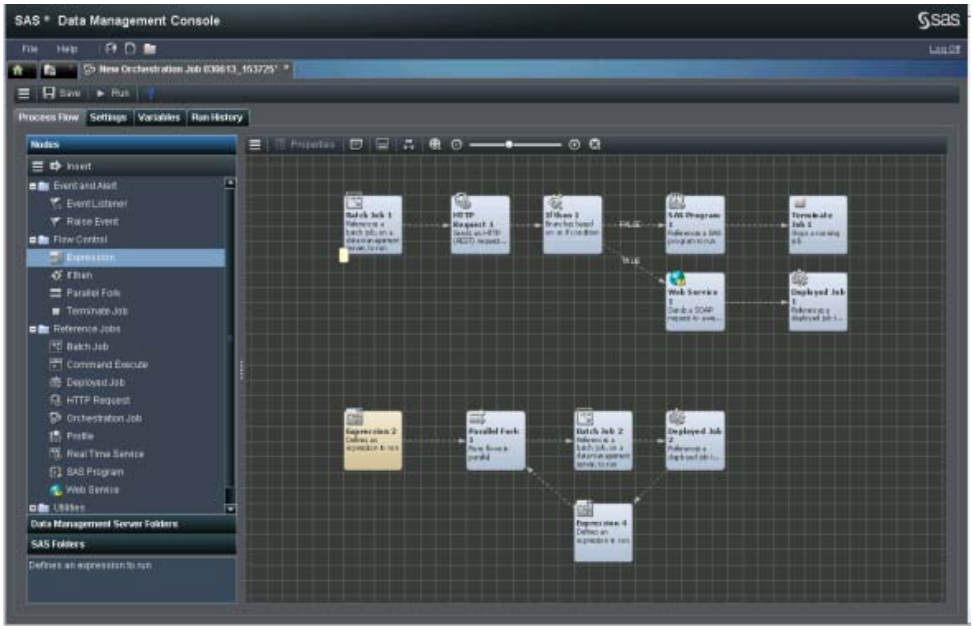

Figure 10. Job Orchestration

Orchestration jobs live in SAS folders, and are transportable between development, test, and production environments using the SAS object promotion framework in SAS Management Console. The job authoring environment also supports locking and versioning of orchestration jobs, as shown in Figure 11.

Figure 11. Locking and Versioning of Process Orchestration Jobs

Many types of objects can be orchestrated to run in serial or parallel and a number of control nodes are available to help manage job flows. Some of the nodes supported are:

- Operating system scripts
- Job nesting (jobs inside jobs inside jobs)
- DBMS scripts
- Web services
- REST services
- SAS Data Integration Studio deployed jobs
- Stored processes
- Batch jobs
- Real time services
- Event listener, event raise
- Expression logic nodes
- Process FORK and loops
- Parallel flows

Process Orchestration includes a highly optimized server that runs process orchestration jobs. The server supports many complex job management capabilities. Illustrated in Figure 12 are some of the example use cases supported by the process orchestration server.

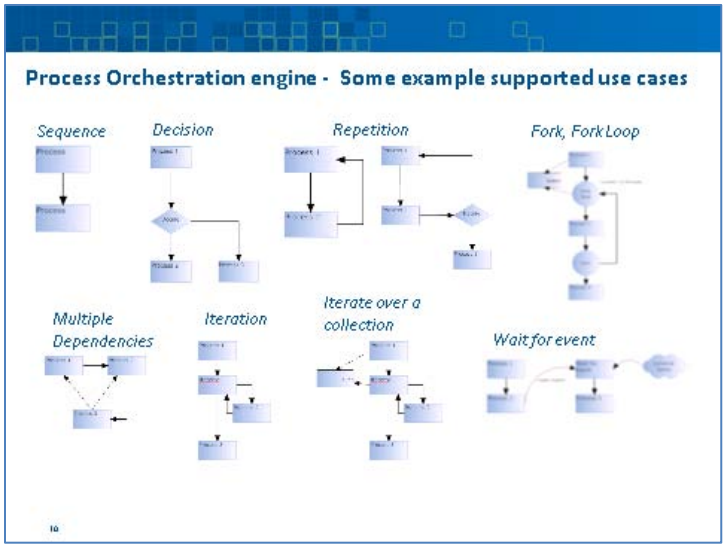

Figure 12. Use Cases Supported by the Process Orchestration Server

DATA FEDERATION

Data federation is a data integration capability that allows a collection of data tables to be manipulated as if they were a single table while retaining their existing autonomy and integrity. It differs from traditional ETL/ELT methods because it pulls only the data needed out of the source system. Figure 13 is an illustration of the differences between traditional ETL/ELT and data federation.

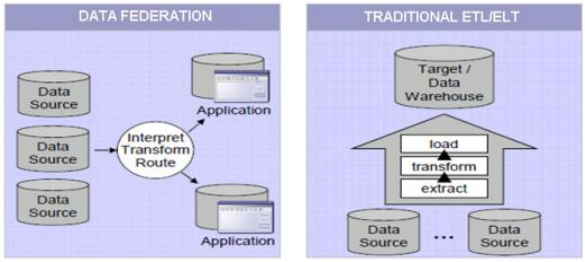

Figure 13. Illustration of the Differences Between Traditional ETL/ELT and Data Federation

The SAS Federation Server fully supports the data federation use case. It includes a data federation engine, multi-threaded I/O, pushdown optimization support, in-database caching of query results, many security features including table, column, and row level security, an integrated scheduler for managing cache refresh, a number of data source native engines for database access, full support for SAS data sets, auditing and monitoring capabilities, and a number of other key features. Figure 14 is a high-level overview of the SAS Federation Server.

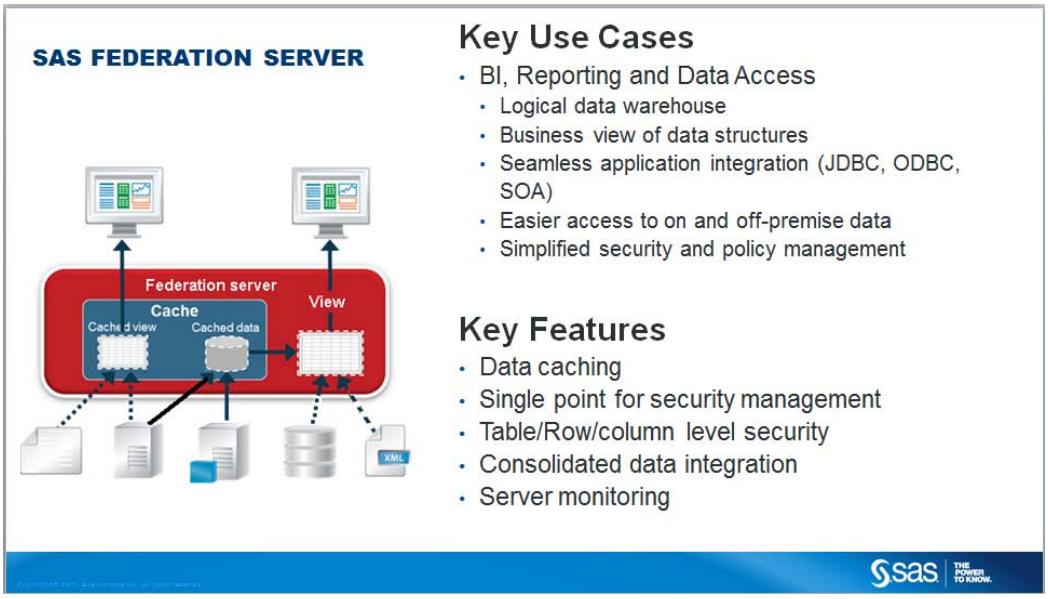

Figure 14. Overview of the SAS Federation Server Architecture

There are a number of scenarios that the SAS Federation server is ideally suited for. One key scenario is illustrated in Figure 15. The scenario illustrates data that is owned by organizations that charge for each data access such as data that resides on a mainframe, or is deemed mission critical. In this environment, users may not be allowed to go directly against the actual tables. SAS Federation Server provides an ideal answer to this problem, because it funnels the data access through the Federation Server itself, so that multiple users do not go through to the base tables. A data cache can be optionally inserted into the result stream, so that even if the underlying tables are not accessible, for example if the source system is down, data can still be provided to end users. The Federation Server also provides a single point for managing all security so that users do not have to be granted direct access to the underlying tables.

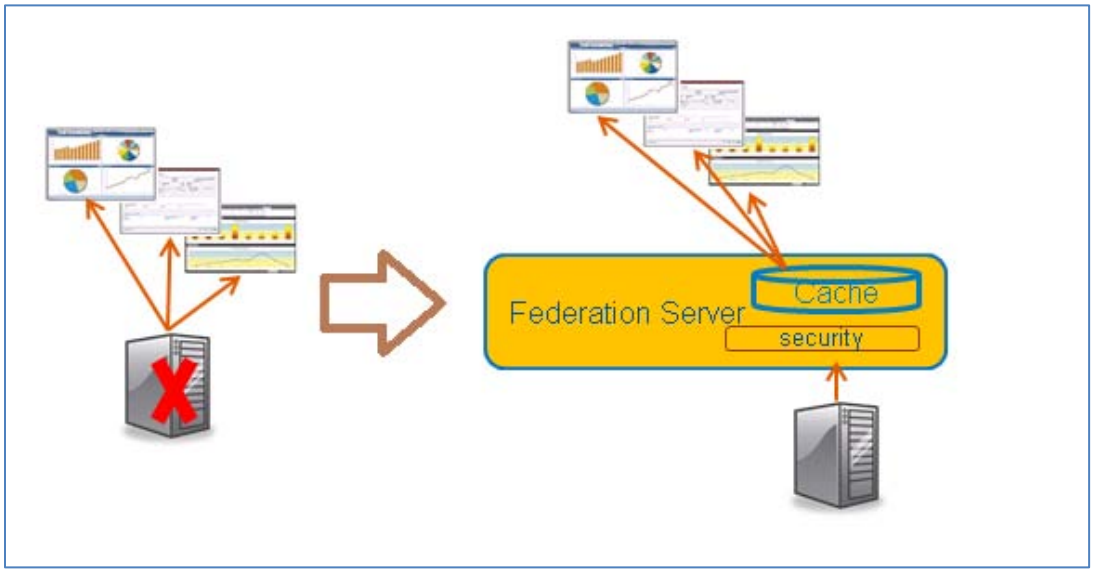

Figure 15. Scenario Highlighting the Value of Data Federation

SAS Federation Server supports access to the server using ODBC, JDBC, or through the SAS/ACCESS to SAS Federation Server LIBNAME engine. A server management web client has been developed for administering and monitoring the server. Figure 16 is an example of the manager web client. From this interface you can monitor multiple servers from the web client interface as well as administer users and permissions. You can create cached queries, which can live in a database target that you choose. You can schedule cache refreshes to occur periodically with the integrated scheduler or use your own scheduler. There are also many monitoring capabilities that let you

track server and data access, identify performance bottlenecks, and view system usage.

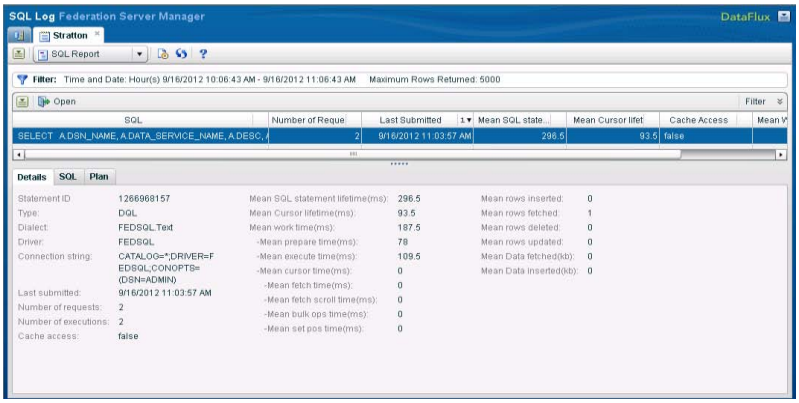

Figure 16. SAS Federation Server Manager Web Client

BIG DATA - HADOOP, LASR, AND IN-DATABASE DATA QUALITY

Apache Hadoop is an open source technology for large data volume storage and processing. Its scalability comes from the marriage of its high bandwidth, clustered storage, called the Hadoop Distributed File System (HDFS), and its fault-tolerant distributed processing algorithm (called MapReduce). SAS Data Integration Studio provides integration with Hadoop in three ways: reading and writing data to and from Hadoop HDFS; data processing for sending programs and managing execution of programs in Hadoop systems; and a data transformation library for writing Hadoop programs in Hadoop languages including Pig, Hive, and MapReduce. The basic architecture of the SAS Hadoop integration is illustrated in Figure 17.

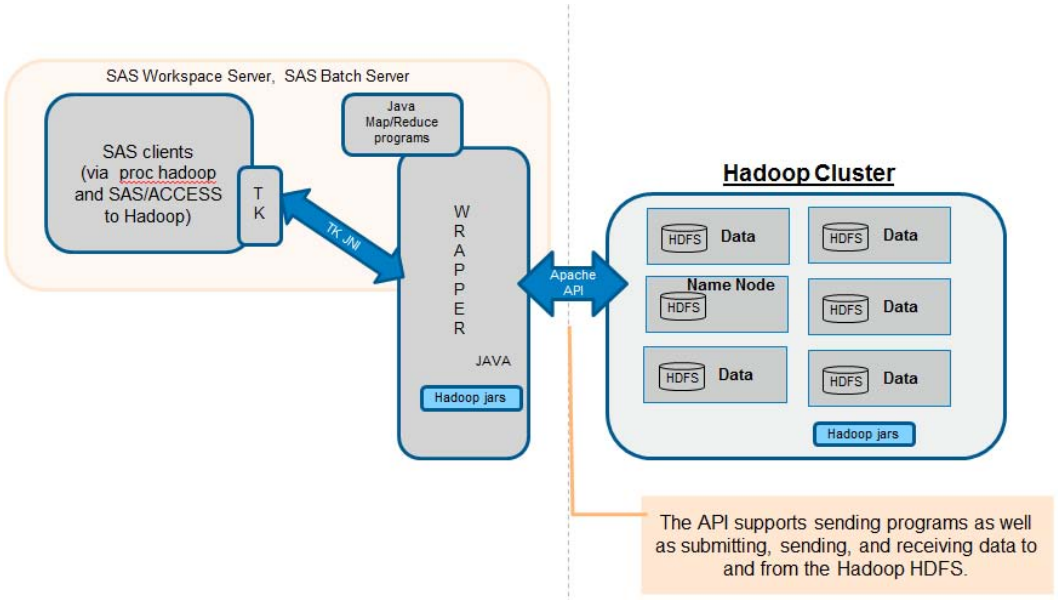

Figure 17. SAS and Hadoop Architecture

SAS can run programs on data in the Hadoop file system using a number of new transforms which are available in SAS Data Integration Studio. Figure 18 shows some of the available transformations and an example flow in SAS Data Integration Studio using one of the transforms.

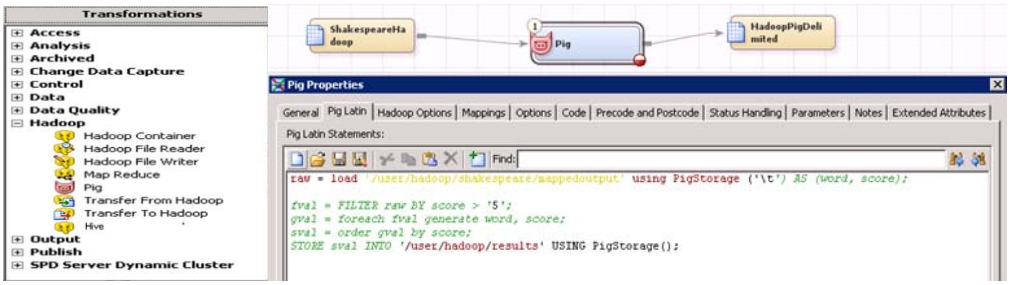

Figure 18. Using Hadoop in SAS Data Integration Studio

All programs submitted from SAS use a new SAS procedure developed to interact with Hadoop called PROC Hadoop. Figure 19 shows an example of the new PROC syntax. The procedure supports submitting and managing programs that are running in the Hadoop system

```
proc hadoop options=cfg username="" password =""
hdfs delete="/user/sasxxw/output_customer";
hdfs delete="/user/sasxxw/outputtest";

pig code=pigcode
    registerjar= "c:/hadoop/myudf.jar"
                 "c:/hadoop/myudflower.jar"
    parameters=pigparam ;
run;
```

Figure 19. PROC Hadoop Syntax Example

The SAS LASR analytic server also leverages Hadoop for data storage. The LASR server has its own HADOOP file storage format, called SAS HDAT, which is highly optimized for fast load of data into the LASR server. SAS can write data in this format from any SAS system using new SAS/ACCESS engines to LASR. If your data is already in Hadoop, SAS Data Integration Studio provides a converter in the Data Integration Studio PIG transform to convert the data from traditional HDFS format to SAS LASR format. Figure 20 is an example of how to use the converter. For data that is already in Hadoop, the converter is a good choice to get your data into the LASR format because you do not have to move it out of your existing Hadoop system.

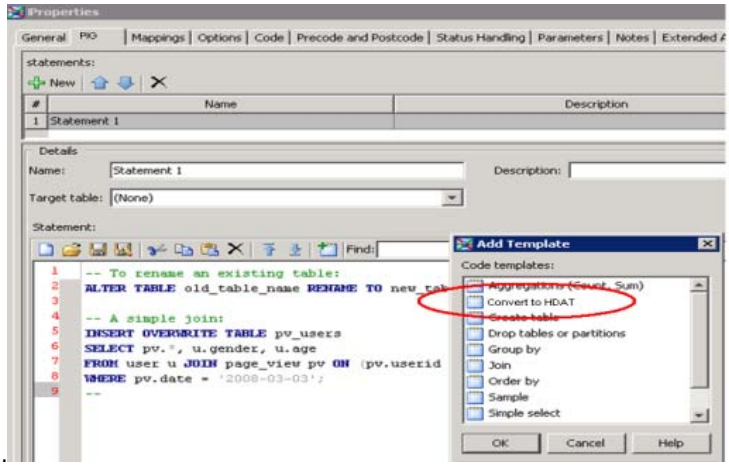

Figure 20. Converter to Convert From HDFS to LASR Format

Other SAS Data Management recent features in support of big data include the in-database data quality feature, and event stream support. In-database data quality involves pushing the data quality capabilities available in the SAS Data Quality Server into the database. The quality functions reside in the DBMS as a set of user-defined functions. This allows you to perform SAS Data Quality functions directly on the data in the database, and can provide significant performance improvements when working with database data. The SAS Event Stream Processing Engine is another key feature in support of big data. Event streaming allows for analytics to be applied to data in real time as it is streaming in from very high speed data sources such as broker trades and financial transactions. The SAS Event Stream Engine is a highly optimized engine for capturing, cleansing, and analyzing data in real time.

JOB MONITOR

Once jobs are deployed to production systems, it is a best practice to monitor the jobs to ensure that they are running as expected. The new Job Monitor web client in the SAS Environment Manager supports this best practice. Figure 21 and Figure 22 are examples of the new job monitor interface.

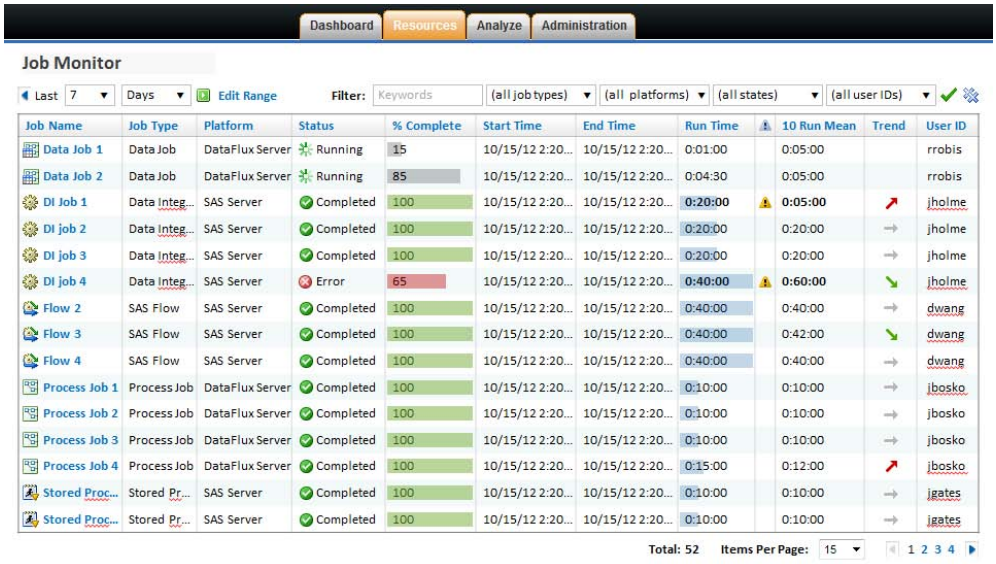

Figure 21: Job Monitor Main Page

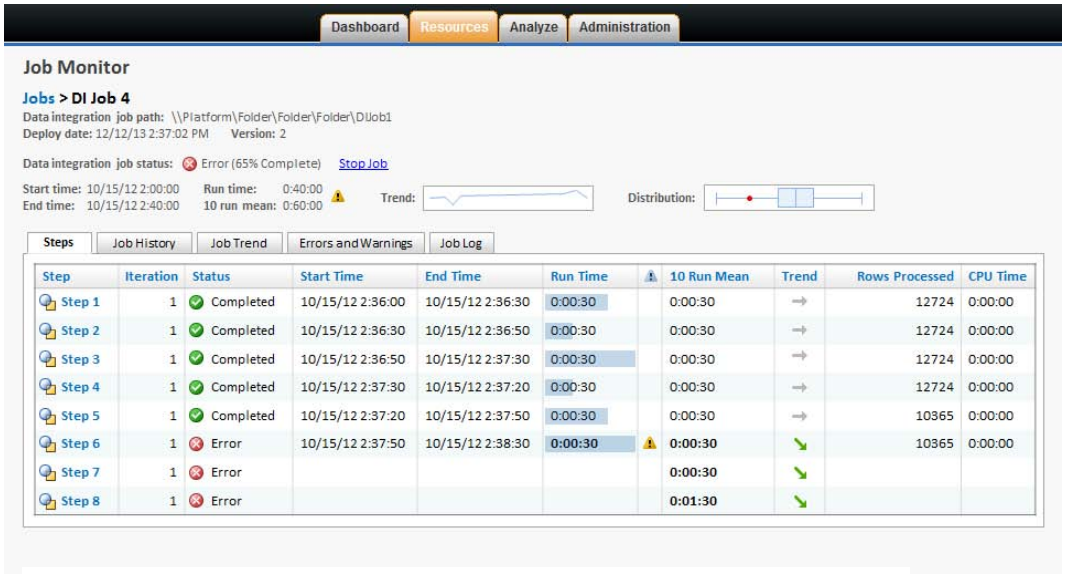

Figure 22. Drill-down in Job Monitor to See Details of a Specific Job

The Job Monitor client comes configured for monitoring Process Orchestration jobs, SAS Data Integration Studio jobs, and DataFlux Data Management Server jobs. There are no changes required to your jobs to be able to monitor them. The client shows a list of all job runs for all jobs being monitor over a period of time that is user configurable. You can filter, search, and sort this list. You can view status such as success, warnings, and errors for each job run. You can also see run-time performance compared to historical run-times for this job; any differences are noted. You can access the job log and can compare the run-times of individual runs as shown in Figure 23. You can also drill in to see individual step run-times inside the job, and you can get this information for multiple levels of nested jobs.

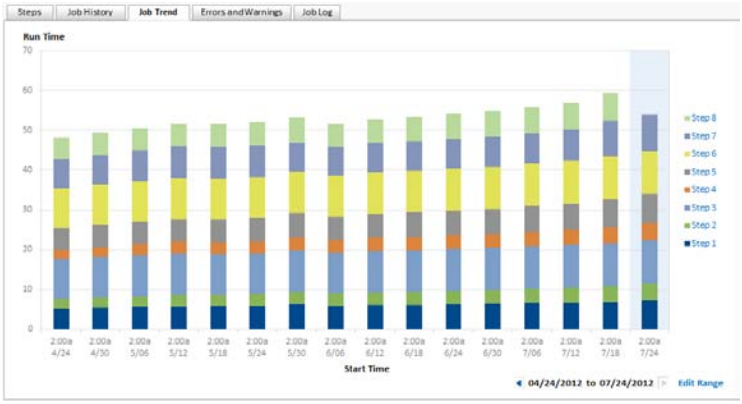

Figure 23: Performance comparison between runs of a job in Job Monitor

A portlet, show in Figure 24, available on the SAS Data Management Console main page, allows you to monitor the historical run-time performance of individual jobs you are interested in. From this view you can launch Job Monitor if you want to see more details about specific job runs.

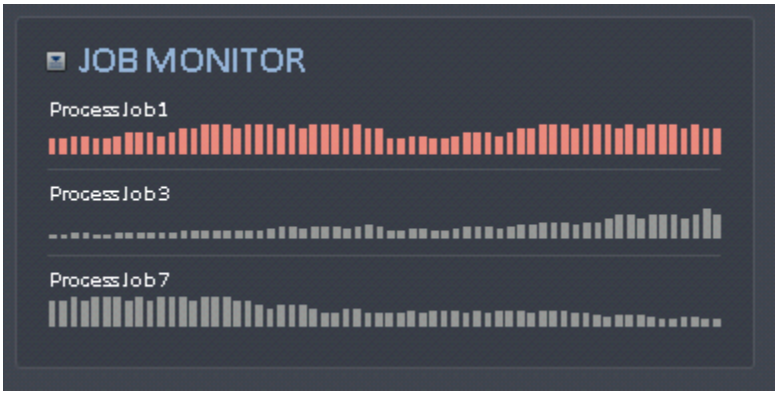

Figure 24. Portlet on the SAS Data Management Console for Monitoring Individual Job Performance

MONITORING AND CORRECTING DATA PROBLEMS

Just as it is a best practice to monitor jobs, it is also a best practice to capture data quality problems as early in your jobs as possible so that invalid data does not make its way into your analytics and reports. You will also want to fix the data so that it can be placed back into your jobs after it has been cleansed. For example, you might have data that fails your business rules that you have in your job because of missing or invalid values, or values out of range. The data is still important, but it needs to be corrected before it can be included in your downstream processes.

SAS Data Management Console has a new data remediation capability that is designed to correct invalid data. Invalid data records or sets of records can be programmatically directed to a Remediation Queue. Once in the queue, the remediation interface supports workflow, so that you will be notified when something needs your attention. The remediation interface supports user alerts and notifications, status tracking, filtering, and workflow queue management. An example of the Remediation Queue interface is shown in Figure 25.

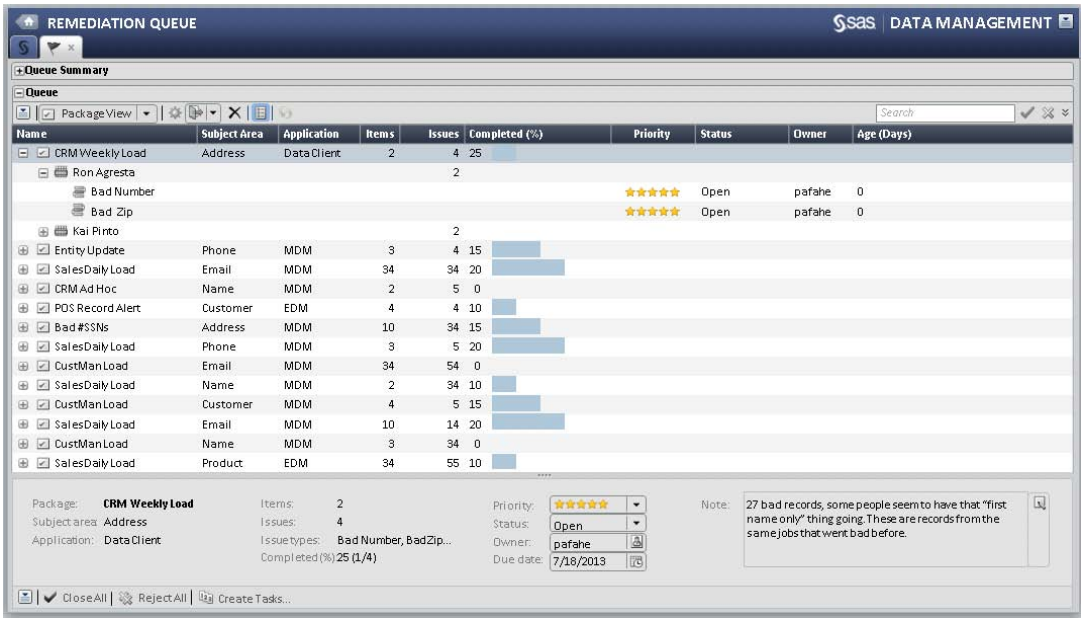

Figure 25. Remediation Queue Interface

You can open the queue to view status of individual records or sets of records. You can see what the data problem is, optionally view the data itself, assign others to work on the issue, assign priority to the issue, and optionally correct the data. All of these activities are managed based on your personal authorizations; an administrator can manage data access for users. You also have the ability to notify others, perform collaboration tasks, and optionally launch jobs to redirect the data back into your processes.

The SAS Data Management Console has a quick view that can be configured to show you the most recent state of the Remediation Queue for those tasks where you are assigned as shown in Figure 26.

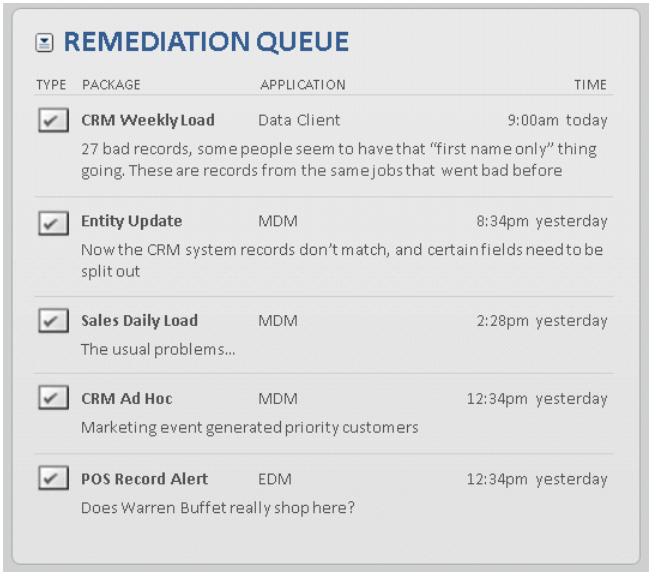

Figure 26. Example Portlet on the SAS Data Management Console Showing Remediation Tasks

You can click on any of these tasks to launch the remediation client.

DATA MONITOR AND QUALITY DASHBOARD

The Data Monitor and Dashboard clients are available to visualize quality problems and show how they are trending over time. The Dashboard allows you to drill in to view more detail about specific dimensions such as Accuracy, Integrity, and others. Dimensions and thresholds are fully user-configurable. Trending is also displayed so you can see how your data is performing over time. This enables you to better react to potential errors and fix problems more quickly. You can also drill in to view specific problem records. An example of the data quality dashboard is shown in Figure 27.

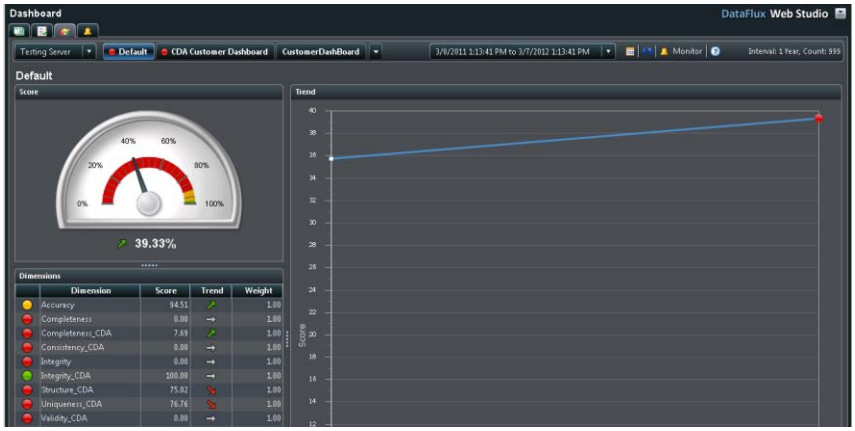

Figure 27. Example of the SAS Data Quality Dashboard Client Interface

From the Dashboard you can drill in further to the Data Monitor to see details on specific data errors. An example of the Data Monitor is shown in Figure 28.

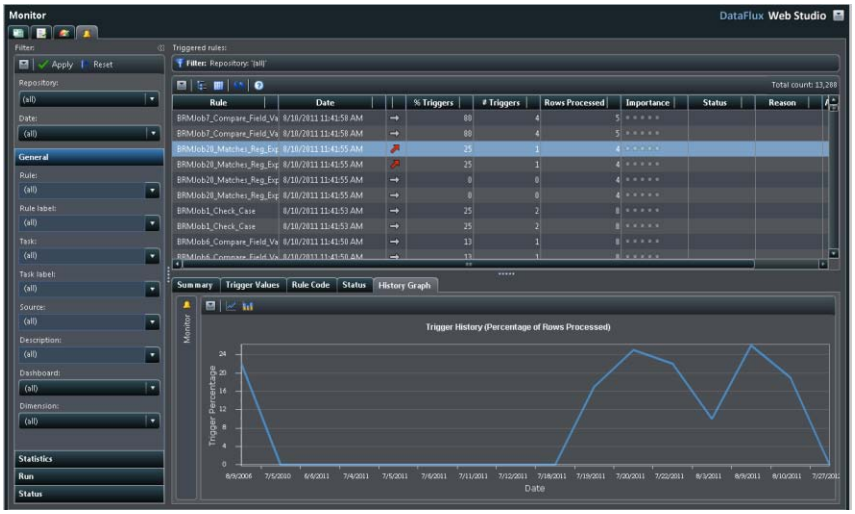

Figure 28. Data Monitor

SAS MASTER DATA MANAGEMENT

When working with certain types of data, there is frequently a need to select the best record out of all possible records prior to including the data in transformation logic. The data might include customer records, supplier information, or other information that has a high chance of having duplicate information.

Figure 29 is an example of a best record.

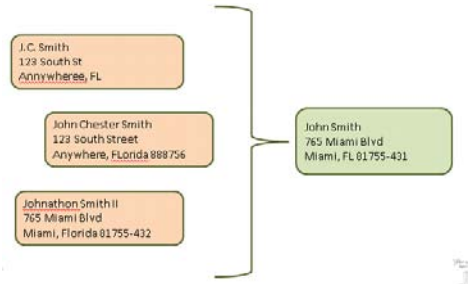

Figure 29. Example of a Best Record Selected From Multiple Records

SAS® Master Data Management (MDM) is a feature available in the SAS Data Management product suite. It allows you to automate the process of selecting the best record. It has been updated and enhanced to integrate into the SAS Data Management Console. MDM works through a technique called "clustering", which is difficult to do with traditional SQL transformation logic .The technology includes support for probabilistic matching. That is, if two records are similar to each other the technology can create a score based on configurable rules as to how likely or how probable the records match. The best record is automatically selected for you into a single, cleansed and de-duplicated data record. Figure 30 shows the new MDM cluster viewer with a selected best record.

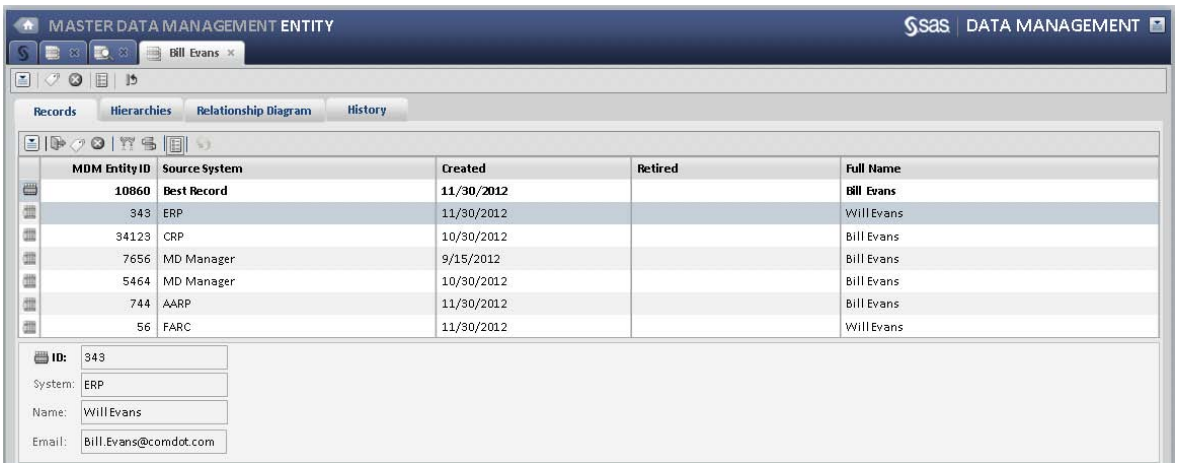

Figure 30. New MDM Cluster Viewer

The MDM client is accessible directly from the SAS Data Management console. A portlet is available to show you quick status about the health of your MDM environment. The portlet is shown in Figure 31.

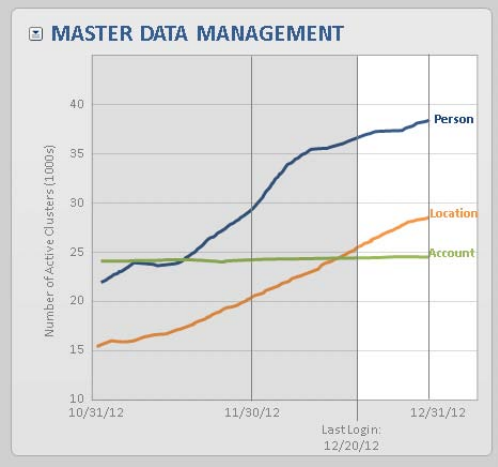

Figure 25. MDM Portlet on the SAS Data Management Console

ADDITIONAL FEATURES

There are a number of additional features in the SAS Data Management products that are worth mentioning. In SAS Data Integration Studio, a new Decision Management node has been added to allow you to map data and run SAS Enterprise Decision Management flows in your Data Integration jobs. The Data Validation transform has been updated and enhanced to support a number of new use cases. New nodes have been added to support SOAP and REST calls out to third-party web services from your jobs. Support for two new engines, an enhanced XML engine, and the SAS/ACCESS engine for POSTGRES have been added.

In DataFlux Data Management Platform, support has been added for interfacing with Java Message Service Message queues. Support for optional SAS Metadata Repository authentication has been added. New features in support of unstructured data have been added to support converting data from a variety of unstructured document types into a format that can be used to better extract data elements. There is a new data driver in support of SalesForce.com. Finally there have been a number of enhancements to the address validation nodes.

CONCLUSION

The latest releases of SAS Data Integration Studio, DataFlux Data Management Studio, and other SAS Data Management products provide many new enhancements to help both data warehouse developers and data integration specialists carry out data-oriented processes more efficiently and with greater control and flexibility. Major focus areas for the release include features for job performance and manageability, many usability enhancements, and the introduction of new transformations to assist you in optimizing your job flows for common data integration tasks. Customers will find many reasons to upgrade to the latest version of SAS Data Management.

RECOMMENDED READING

- SAS® Enterprise Data Management & Integration Discussion Forum, Available at http://communities.sas.com/community/sas_enterprise_data_management_integration
- Alexander, Malcolm, and Nancy Rausch. 2013. “Best Practices in SAS Data Management for Big Data.” Proceedings of the SAS Global Forum 2013 Conference. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings13/007-2013.pdf.
- Rausch, Nancy, et al.2012. “What’s New in SAS® Data Management.” Proceedings of the SAS Global Forum 2012 Conference. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings12/110-2012.pdf.
- Rausch, Nancy, and Stearn, Tim. 2011. “Best Practices in Data Integration: Advanced Data Management.” Proceedings of the SAS Global Forum 2011 Conference. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings11/137-2011.pdf.
- Hazejager, Wilbram, and Pat Herbert. 2011. “Innovations in Data Management – Introduction to Data Management Platform.” Proceedings of the SAS Global Forum 2011 Conference. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings11/141-2011.pdf.

- Hazejager, Wilbram, and Pat Herbert. 2011. “Master Data Management, the Third Leg of the Data Management Stool: a.k.a. the DataFlux® qMDM Solution.” Proceedings of the SAS Global Forum 2011 Conference. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings11/146-2011.pdf.

CONTACT INFORMATION

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

Nancy Rausch
SAS Institute Inc.
Cary, NC 27513
Work Phone: (919) 677-8000
Fax: (919) 677-4444
E-mail: Nancy.Rausch@sas.com
Web: support.sas.com

Malcolm Alexander
SAS Institute Inc.
Cary, NC 27513
Work Phone: (919) 677-8000
Fax: (919) 677-4444
E-mail: Malcolm.Alexander@sas.com
Web: support.sas.com