Paper SAS034-2014

What's New in SAS® Data Management

Nancy Rausch, SAS Institute Inc., Cary, NC; Mike Frost, SAS Institute Inc., Cary, NC, Mike Ames, SAS Institute Inc., Cary

ABSTRACT

The latest releases of SAS® Data Integration Studio and SAS® Data Management 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 clean, standardize, transform, master, and manage your data. The latest features include:

- capabilities for building complex job processes
- web and tablet environments for managing your data
- enhanced ELT transformation capabilities
- big data transformation capabilities for Hadoop
- integration with the SAS® LASR™ platform
- enhanced features for lineage tracing and impact analysis
- 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 for leveraging 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, data integration specialists, and data scientists 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; features in support of quality, collaboration, and governance; and new monitoring features. This paper will showcase some of the newest features available in the SAS® Data Management products.

DATA MANAGEMENT CONSOLE

The SAS® Data Management Console is shown in Figure 1. The console provides at-a-glance information about your data management environment and includes links for launching other Data Management products. It is a role-based and user-customizable web client that provides information and status about data problems, workflow management, job authoring, locking and versioning, job status, job monitoring, and other data related information. There are views that provide up-to-date information on data management activities in your enterprise and alert you to problem areas. The console is fully integrated with the SAS web infrastructure platform. New components that have been added to the console include alerts about to-do workflow items from the Business Data Network, access to the new Lineage viewer, and additional Master Data Management views that show information about the state of your master data hub.

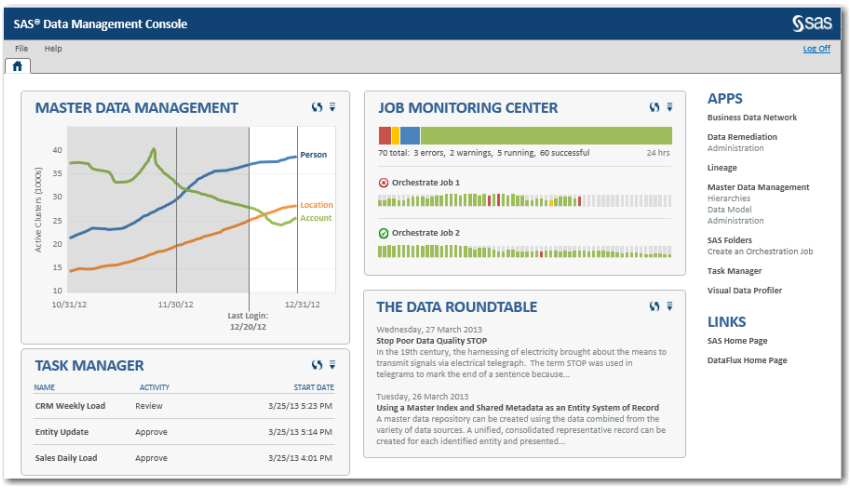

Figure 1: Data Management Console Example

BUSINESS DATA NETWORK

The Business Data Network enables collaboration of domain knowledge between business, technical, and data steward users. The 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. Technical users use the network to document information about tables and columns that implement the business terminology, as a data dictionary to describe details of data models and other data related information, to relate jobs and other information to terms, and to share knowledge about data transformations. Data stewards can view data from a business standpoint to better visualize problem areas by domain so as to identify and fix data issues more effectively. Figure 2 shows the Business Data Network main view.

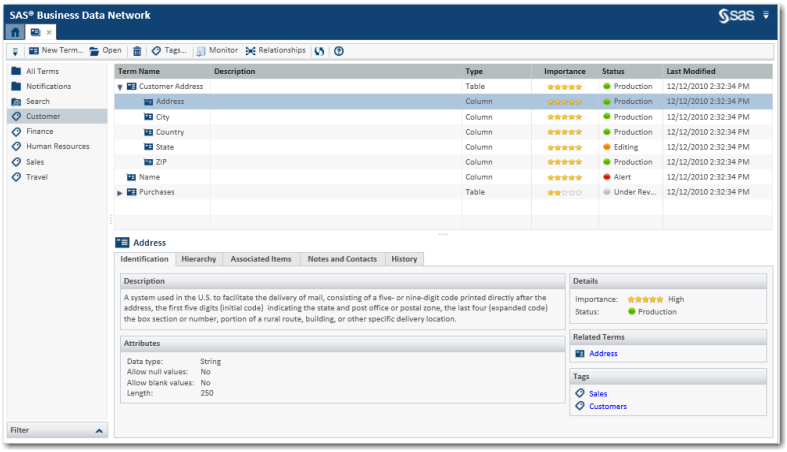

Figure 2: Business Data Network Main View

Typically a user that understands their business terminology would provide the initial information in the Business Data Network. This user may also attach documents or rules that describe each term. There are also import and export features available to support the ability to quickly populate and exchange information. The technical user adds additional information related to the term such as jobs that are used to modify the term, and data that is related to the term. The network is fully integrated with impact analysis to help you understand how your physical data and business processes interrelate. Figure 3 shows a typical diagram of relationships stored in the network.

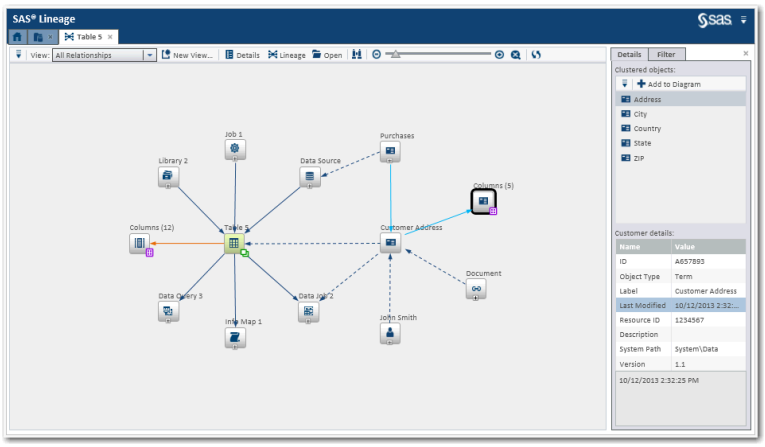

Figure 3: Example of the Relationships View Showing Business and Technical Metadata

There are a number of new features in the latest release of the Business Data Network. The user interface now supports roles, capabilities and security for terms, and term attributes. The roles and capabilities are fully customizable to match your site requirements. Figure 4 displays the many roles, capabilities, and security settings available.

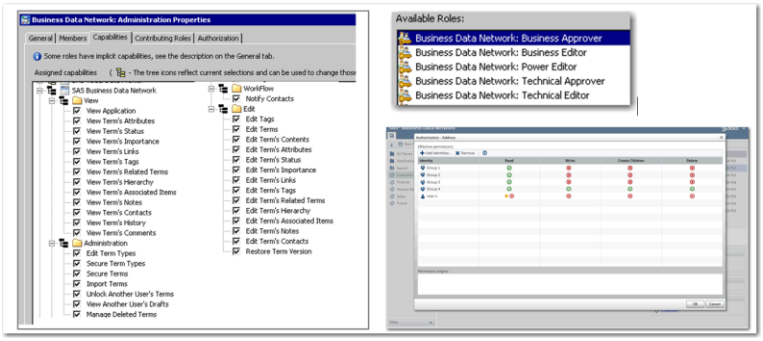

Figure 4: Examples of Security and Role Features Available in Business Data Network

Integration with SAS workflow is also now available. Users can send terms into workflow for review and approval before publishing. There are several default workflows available which you can customize, or you can create your own workflows to match your business needs. Figure 5 is an example of a workflow process applied to Business Data Network. The network reads the workflow state and customizes the UI to display buttons to help you interact with each workflow state. Status is also shown at each step in the workflow.

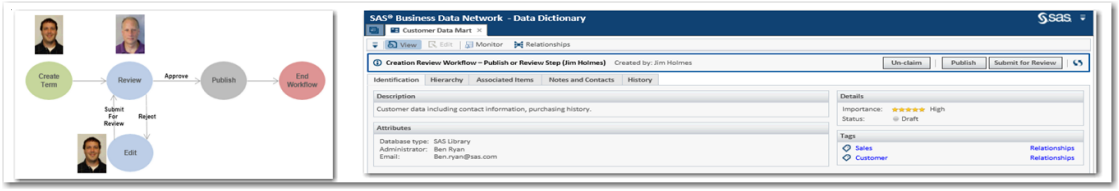

Figure 5: An Example of Using Workflow in Business Data Network

Users can quickly see workflow tasks that are waiting on their input in the task manager view in data management console and in views in the network. There are also a number of quick actions available for users such as being able to update the workflow for multiple terms together.

Different workflows can be used for different actions in the network. For example you can have one workflow that you want to use when creating terms, and another when deleting terms. You can also tie different workflows to different term groups, for example you can have one workflow when working with supplier information, and a different one for working with your data dictionary tables and columns.

Another important new feature is support for multiple, customized term templates. Administrators are now able to create templates with custom attributes for terms and term hierarchies. For example, you might have a set of terms that represent the tables and columns in your data dictionary. You can create a table template with the information

that you want to use to describe tables in your system, and a different column template with the information you want to capture about columns. You can have any number of custom templates that match the information you want to capture in your terms. You also have options to specify whether the template should be inherited in a hierarchy of terms, whether attributes are required, which can be useful if you want to enforce the collection of standard information for every term that is built from the template, and default values for attributes in a term. Most of the attributes of a term are now fully customizable via the term template. Figure 6 is an example of some of the features available for term templates:

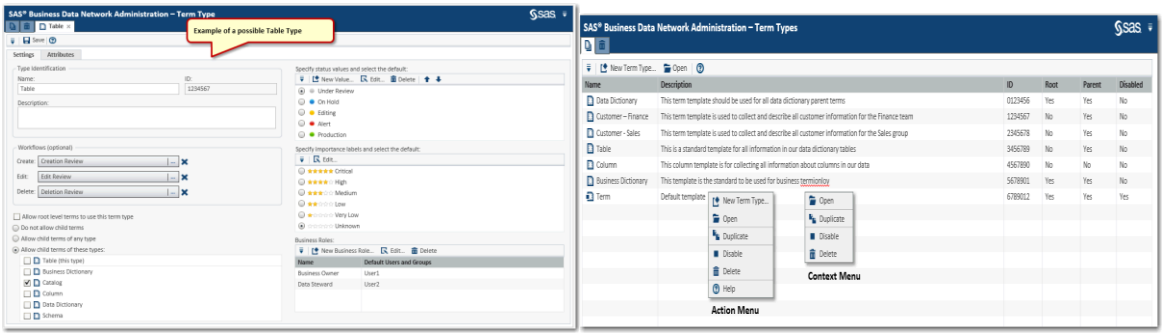

Figure 6: Customizing Term Template Examples in Business Data Network

LINEAGE AND IMPACT ANALYSIS

A number of important new features have been added to the lineage and impact analysis feature. SAS has created 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. There is also the ability to import content from third party sources.

The relationships web viewer has been enhanced to add a number of new views for displaying information stored in the service. Figure 7 is an example of the Impact Data Flow view. There are also views for all Relationships, and for Data Governance.

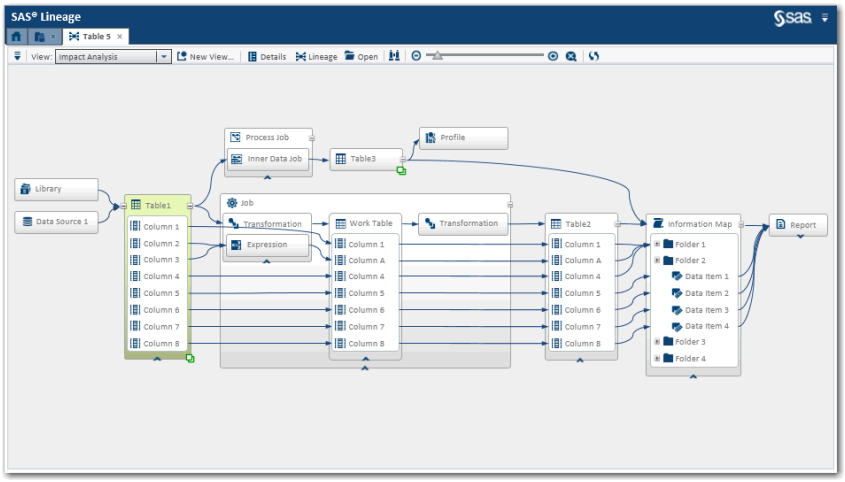

Figure 7: Lineage Viewer Showing Table, Job, and Column Relationships

You can also create your own views using the filtering capabilities of the viewer. This can help you subset the information to just the objects and relationships that you want to see. In addition there are helpful features such as grouping node sets, allowing you to expand on demand, and an overview window with details of objects. Figure 8 illustrates some of these new features.

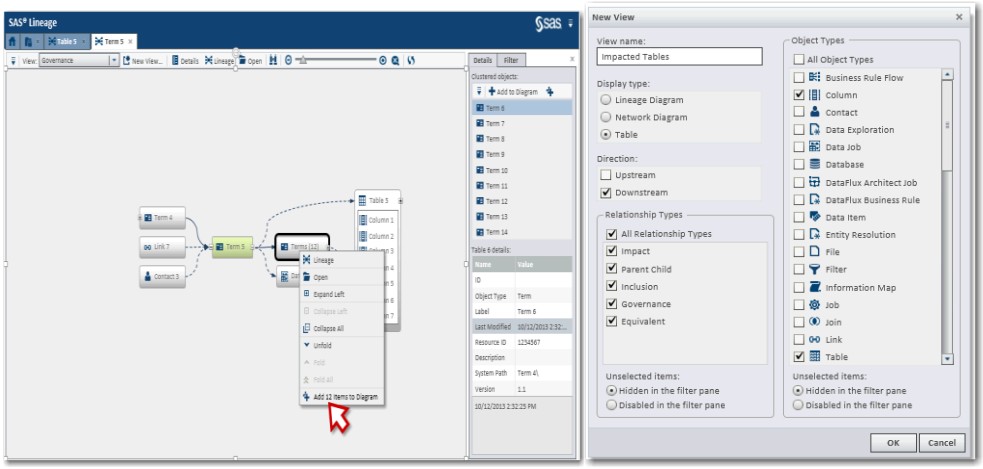

Figure 8: Custom Views, Collapsing and Expanding Multiple Nodes, and Node Details in Lineage Viewer

BIG DATA – DATA DIRECTOR FOR IN-HADOOP ETL

When traditional data storage and computational technologies struggle to provide either the storage or computation power required to work with their data, an organization is said to have a big data issue. One scenario that drives this challenge can be related to data volumes. As data volumes increase they can stretch the capacity of existing storage systems to handle. Another big data challenge is computation windows. Tasks such as simulations and risk calculations, which work on relatively small amounts of data, can still generate computations that can take days to complete, which place them outside the expected decision-making window needed. Finally, business processes may require long-running ETL-style processes or significant data manipulation.

The most significant new technology trend that has emerged for working with big data is Apache Hadoop. Hadoop is an open source set of technologies that provide a simple, distributed storage system paired with a fault tolerant parallel processing approach that is well suited to commodity hardware. Many organizations have incorporated Hadoop into their enterprise leveraging the ability for Hadoop to process and analyze large volumes of data at low cost using commodity hardware.

SAS is integrating with the Hadoop platform to bring the power of SAS to help address big data challenges in Hadoop. SAS, via the SAS/Access® technologies and code accelerator products, has been optimized to push down computation to the data stored in Hadoop. By reducing data movement, processing times decrease and users are able to more efficiently use compute resources and database systems. SAS is also supplying new user interfaces for working with Hadoop data and leveraging SAS capabilities that are installed in the Hadoop system.

The new SAS Data Director client is how you interact with SAS and Hadoop. SAS Data Director allows you to create and run Hadoop jobs, view status and results, and access and view data stored in the Hadoop system. When SAS is installed on the Hadoop cluster, SAS Data Director will also leverage SAS in-database and code accelerator capabilities. Figure 9 is an example of how Data Director interacts with the Hadoop system and can leverage SAS embedded into Hadoop.

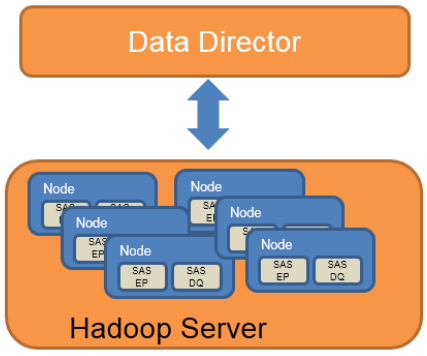

Figure 9: SAS Data Director Architecture Showing SAS Running in Hadoop

SAS Data Director offers an easy to use point and click user interface that allows you to:

- Support fast data copies, in parallel, to and from RDBMS systems and the Hadoop file system using Hadoop technologies such as SQOOP
- Create and run MapReduce jobs in Hadoop for common ETL capabilities such as sort, filter, subset, and join
- Easy movement of data from Hadoop to SAS LASR and SAS Visual Analytics servers
- Leverage the power of SAS in Hadoop by allowing you to build and run SAS programs inside of Hadoop using the SAS embedded process for Hadoop and SAS/Access for Hadoop
- Cleanse your Hadoop data via the SAS Data Quality code accelerator for Hadoop
- Profile your Hadoop data using the SAS Profile engine for Hadoop
- Build and access your Hadoop jobs and data from a web browser or from mobile devices such as tablets

Figure 10 is an example of the Data Director main page.

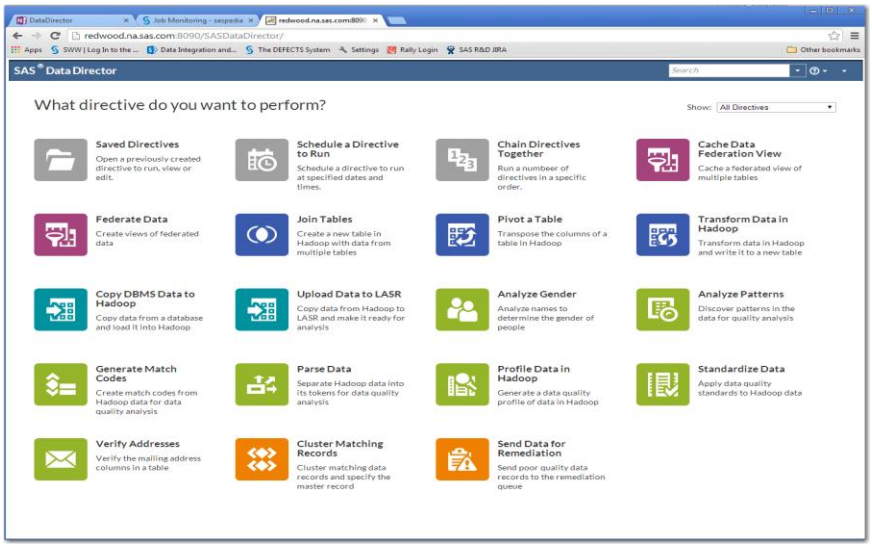

Figure 10: Data Director Main Page Example

From this page, users can select the type of job in Hadoop they want to perform. Jobs such as data consolidation and movement are simple point and click actions. There are also actions to cleanse and profile data, as well as transform the data in Hadoop.

Figure 11 is an example of one of the customization screens for the Transform Data in Hadoop job. Users customize the job by adding the source and target data, and optionally applying in Hadoop transformations. Some of the available transformations in Hadoop are shown in Figure 11.

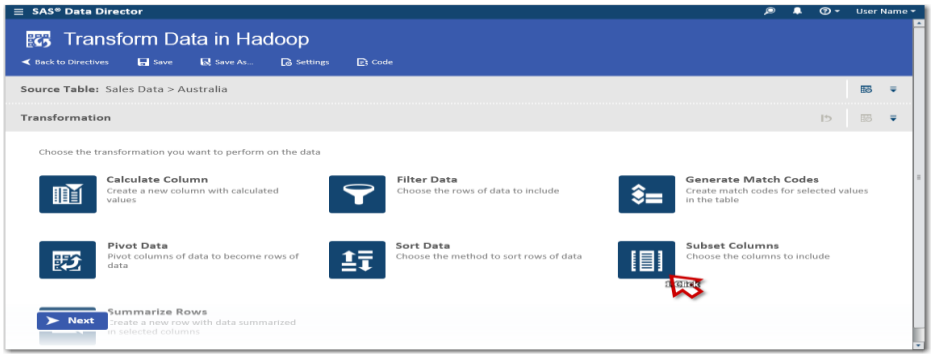

Figure 11: Examples of Some of the Hadoop ETL Transformations

Once the user has completed configuring their job, they can choose to run the jobs from the client. Interfaces are also available to schedule job runs. Run progress is available immediately or the user can choose to close and return later to see run status, results, and logs from the job status view.

BIG DATA - IN DATABASE DATA QUALITY FOR HADOOP

SAS has added the capability to apply data quality routines in-database to Hadoop data. Data Director will allow you to generate jobs that can cleanse, standardize, and apply data quality to your data in Hadoop. Figure 12 is an example of the standardize data quality feature. This example shows an example of some data in a Hadoop dataset. The original data, outlined in blue in the figure below, has state names that in some records are two letter abbreviations, and in other records are full names. Using the in-database data quality standardize technique, you can apply a state name standard to your data. The result is placed into a new column in the table. The result column outlined in red in the figure below shows the new column added to the table with state names standardized to the full state name.

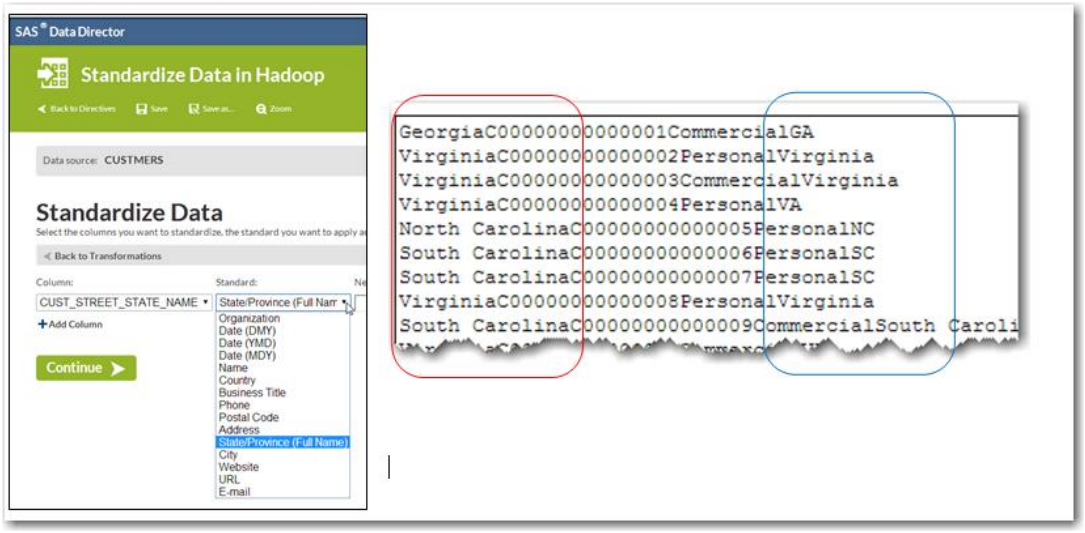

Figure 12: Hadoop Data Quality Standardize Example

Running the SAS data quality technique in Hadoop allows you to cleanse your data stored in Hadoop and significantly improves run time performance over extracting data to apply data quality. The SAS Data Quality Accelerator supports the following data quality operations in Hadoop

- Parsing
- Extraction
- Pattern Analysis
- Identification Analysis
- Gender Analysis
- Standardization
- Casing
- Matching

BIG DATA - IN DATABASE PROFILING FOR HADOOP

SAS also supports profiling Hadoop data via the new SAS in-database profile engine. Using the in-database features of profile allow you to quickly get an idea of the quality of your data stored in Hadoop. This feature is similar to the in-database data quality accelerator, in that all of the processing is done in the Hadoop server. There is no longer a need to extract data to another system to measure data quality, which will significantly improve the performance of data quality analysis when working with big data.

The web profile viewer displays the results of your data quality analysis in Hadoop. The viewer allows you to view a

number of data quality metrics associated with your data, such as count of missing values, range of values, string patterns, frequency analysis, and other data quality metrics. Figure 13 is an example of some of the available profile results views of Hadoop data.

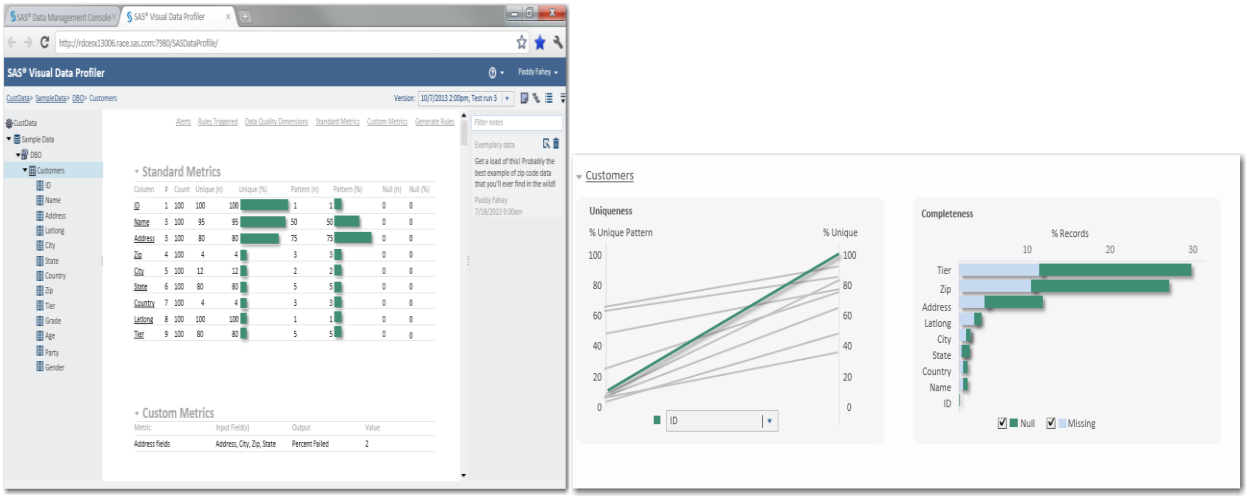

Figure 13: Example of Profile Data Results in Hadoop

JOB ORCHESTRATION

As data volumes grow, jobs that gather, transform, and manage data need to become increasingly complex to handle new performance challenges. In addition, data integrators increasingly have to manage content, including other jobs, coming from diverse systems and sources. The Job Orchestration feature in SAS Data Management is designed to help data integrators better manage their job flows. It offers a job authoring and runtime 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 14 below. 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 runtime statistics for each embedded node. The Job Orchestration feature is fully integrated with the SAS platform.

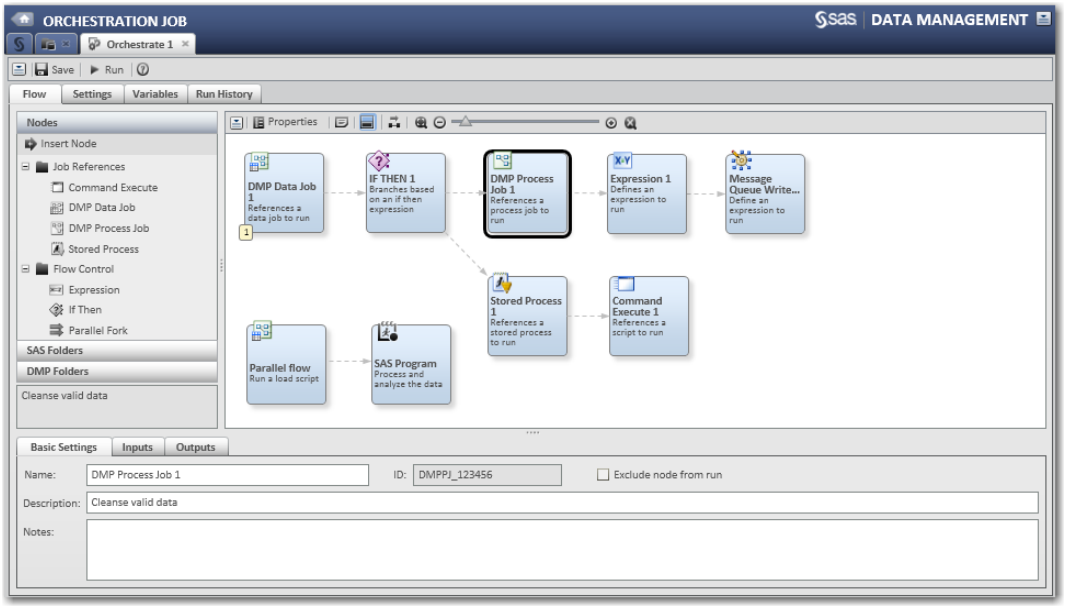

Figure 14: Job Orchestration Example

Job 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 illustrated in Figure 15.

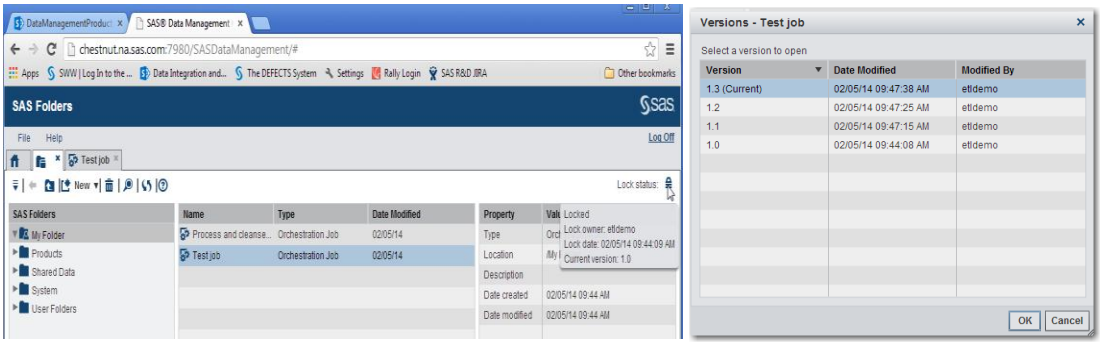

Figure 15: Examples of Locking and Versioning of Jobs

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 node types supported are:

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

JOB MONITORING

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 Job Monitor web client in the SAS Environment Manager supports this best practice. Figure 16 are some examples of the job monitor interface.

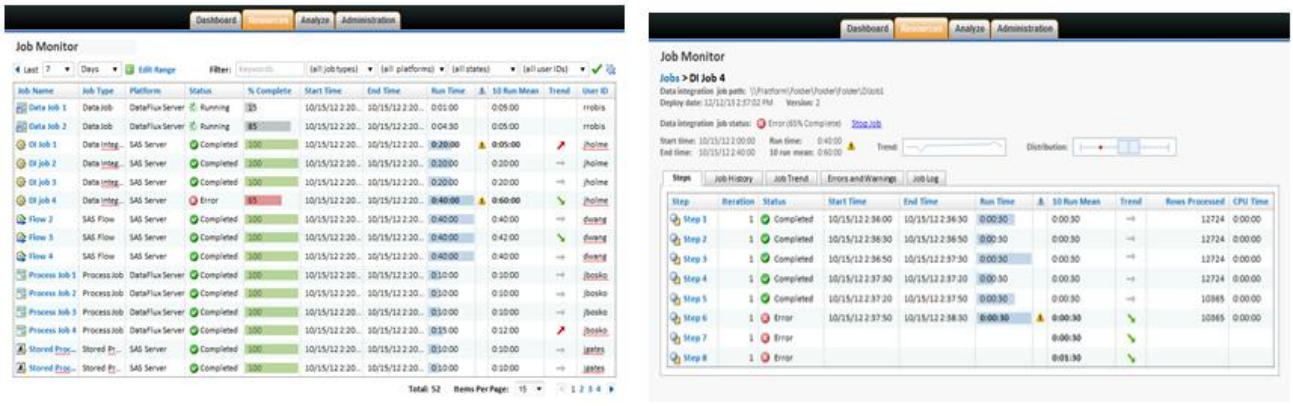

Figure 16: Job Monitor

The monitor shows a list of all job runs for all jobs being monitored over a period of time. The user time is configurable. You can also 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 compare the run times of individual runs. You can also drill in to see individual step run times inside the job, and get this information for multiple levels of nested jobs.

Figure 17 below shows the job monitor quick view on the SAS Data Management Console main page, which allows you to monitor the historical run time performance of individual jobs you are interested in. From this view you can launch Job Monitor to see more details about your jobs.

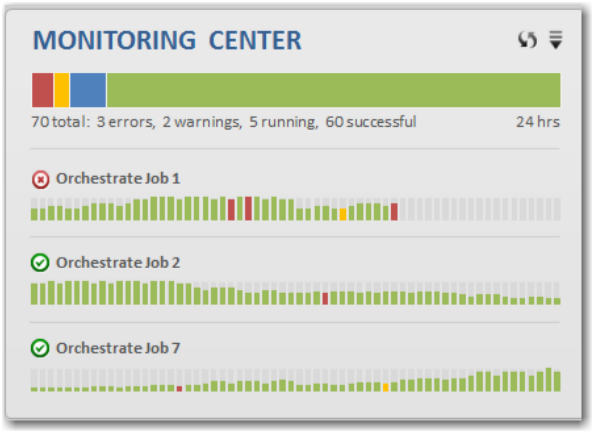

Figure 17: Job Monitor Quick View

DATA INTEGRATION STUDIO AND DATA MANAGEMENT NEW FEATURES

There are a number of new features that have been added to SAS Data Integration Studio. There is a new control flow node for performing conditional processing inside of your jobs. An example using this new node is shown in Figure 18. You can add any number of conditional start and end nodes to your job flow. Once you setup the condition, the job will read the run time value of the condition and process the steps in the job based on the value.

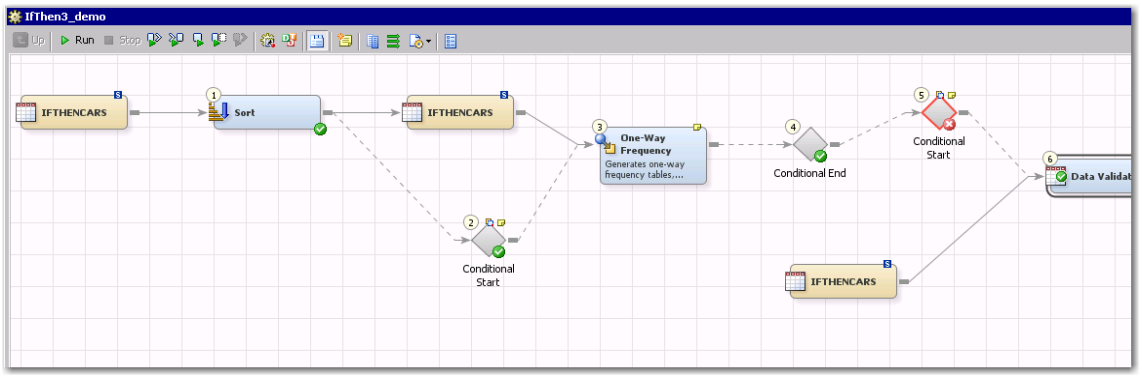

Figure 18: Data Integration Studio Conditional Node Example

Other feature enhancements in Data Integration Studio include additional source designers for supporting new SAS/Access engines and enhanced support for Decision Management. The SAP Data Surveyor has been updated to support InfoCubes and semantically partitioned objects. There are also new in-database access features to support SAS in-database capabilities for SAP HANA.

In SAS Data Management, new features have been added to enhance data quality clustering capabiltiies. The Quailty Knowledge Base has new components for customizing to meet your site needs, and for importing and exporting content between versions. Data Management Studio and Data Management Server repositories can now be configured for the SAS open source database, and there are new features incorporated into the Unstructured Data nodes for enhancing support of reading and parsing unstructured data.

MASTER DATA MANAGEMENT

When consolidating data from diverse source systems, there is often a need to select the best record out of all possible records that represent the same data, so that you can pass one version of the data to downstream jobs and reports. For example, you may have multiple diverse customer records coming in from various source systems, and

you need to be able to consolidate on a single, best record, with standardized values for fields such address and phone number.

Figure 19 is an example of a customer best record selected from three different source records.

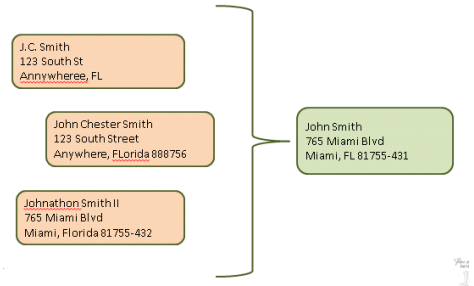

Figure 19: Customer Best Record Selection

SAS® Master Data Management (MDM) automates the process of selecting the best records from source data. MDM works through a technique called “clustering”, which is difficult to do with traditional SQL transformation logic. The technology supports sophisticated techniques such as probabilistic matching that are able to pull out the best record based on analytical processes. For example, if two records are similar to each other the technology can create a score based on 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 20 is an example of the results view in SAS MDM showing an example set of incoming records and the selected best record:

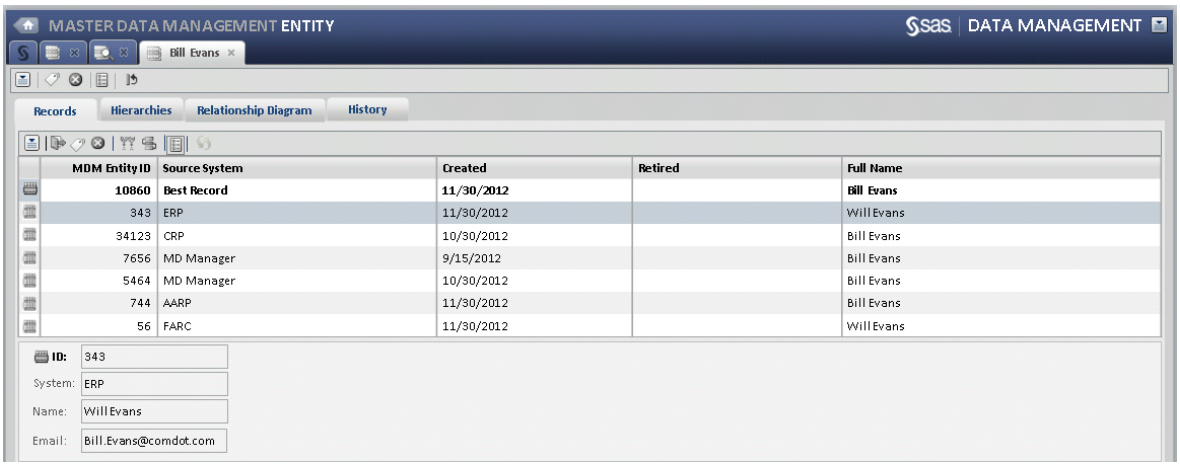

Figure 20: Master Data Management Best Record View

SAS MDM has features that allow you to interact with the data hub, which include the data repository that stores the best records. New features have been added to support workflow and remediation of data when working with the hub. With the latest release, you can now send data records into a workflow for users to review, approve or reject changes, and get notifications when you or others in your workgroup are required to take some action on the data. For example, you may want to review the best record that was selected, or adjust the data first and have it reviewed by your data quality experts before adding it to the hub. Figure 21 is an example of the Data Remediation view.

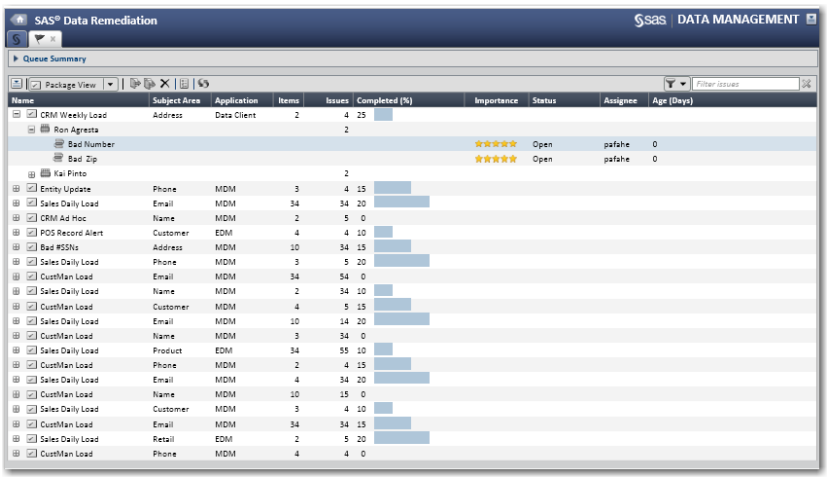

Figure 21: Data Remediation View

Another new feature is the ability to create multiple hierarchies and add records to them to match the data in your enterprise. Records can be referenced from multiple hierarchies and the relationships between the data in the hierachies are user configurable. Figure 22 is an example of hierarchy management in SAS MDM.

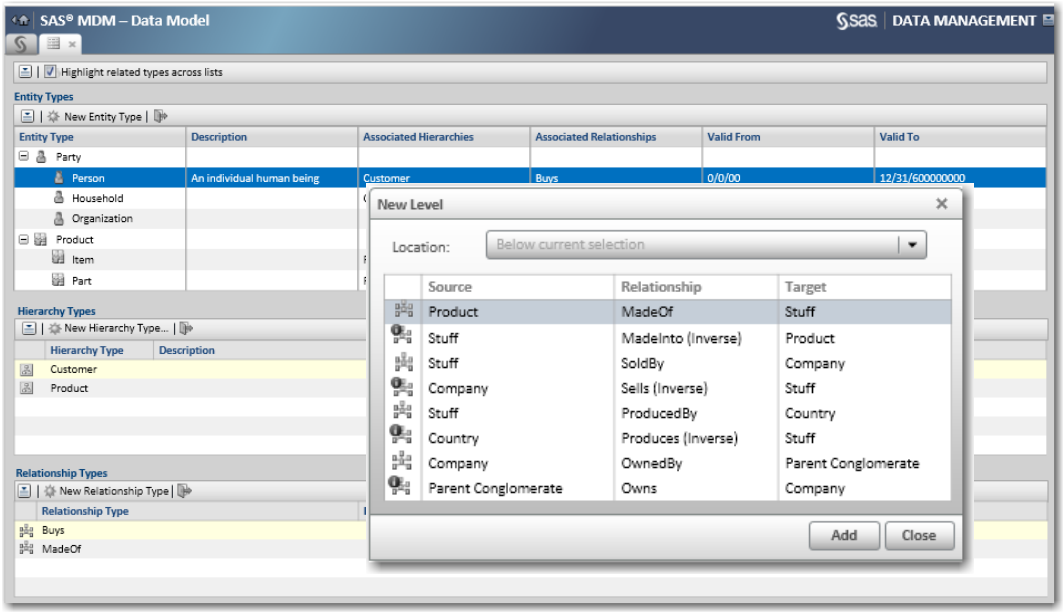

Figure 22: Master Data Management Hierarchy Examples

DATA FEDERATION

Data federation is a data integration methodology that allows a collection of data tables to be manipulated as views created from diverse source systems. It differs from traditional ETL/ELT methods because it pulls only the data needed out of the source system. Figure 23 is an illustration of the differences between traditional ETL/ELT and data federation.

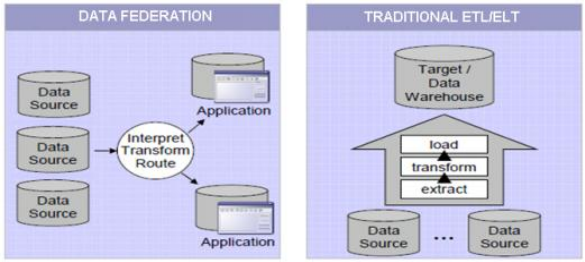

Figure 23: Illustration of the Differences Between Traditional ETL/ELT and Data Federation

Typically a data federation methodology is used when traditional data integration techniques cannot meet the data needs. One example scenario is when the data is too large to be extracted out of the source systems. Data federation solves this challenge because only the needed information is gathered from the source systems as a view that can be delivered to downstream processes. A second scenario well suited to data federation is when working with sensitive data. For example, data may be owned by organizations that do not want to grant direct access to their tables, or may charge for access to their tables. In this case, federation can play a key role because the data can be extracted and stored in a persistent cache. The cache can be updated periodically or scheduled to be refreshed during non-mission critical times. A third scenario for data federation is when data is diversified, that is spread across many diverse source systems. Managing security for all of the various source systems, and updating all applications to pull data from the various systems can become a large burden for the traditional data integration model. Data federation is well suited for this usage scenario because it allows system integrators to have a single point of control for managing security, and for updating views when source systems change.

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, an integrated scheduler for managing cache refresh, a number of data source native engines for database access, full support for SAS datasets, auditing and monitoring capabilities, many security features including table, column, and row level security, along with a number of additional key features. Figure 24 is a high-level overview of the SAS Federation Server.

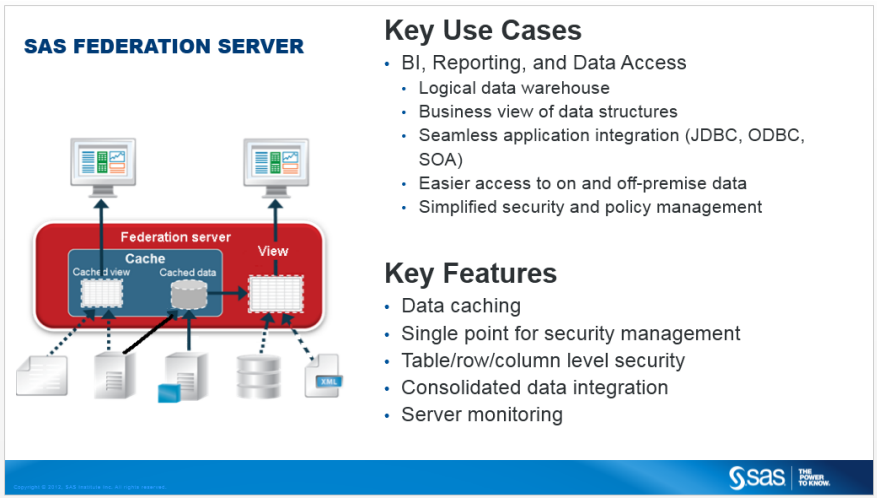

Figure 24: SAS Data Federation Server Overview

New features in the SAS Federation Server include support for data masking. Three new functions, ENCRYPT, DECRYPT, and HASH support the ability to mask sensitive information in your data tables. Below is example source code using data masking.

```
// Create table w/ encrypted NAME column:
create table "EMPLOYEES_ENCR" as
select *,
       syscat.dm.mask('ENCRYPT', "NAME",
                      'alg', 'AES',
                      'deterministic','yes',
                      'cta_values','yes',
                      'key','xyzzy') as "NAME_ENCRYPTED"
from "EMPLOYEES";
```

Figure 25 is an example of applying the above function to some data. The first column in the dataset on the left contains the original, unencrypted name. Applying the data masking function to the data results in the dataset on the right side. The result set has masked the name column.

1 Alfred	M	14	116.94168422...	135.0706559852119
2 Alice	F	13	68.906553332...	85.63003016134127
3 Barbara	F	13	105.26025321...	117.89926300818239
4 Carol	F	14	96.375045159...	107.2893197211607
5 Henry	M	14	98.982069499...	110.14093775704566

1 D022CDAB	M	14	116.94168422...	135.07065598...
2 F0282E2E	F	13	68.906553332...	85.630030161...
3 F028DDA5	F	13	105.26025321...	117.89926300...
4 F0280141	F	14	96.375045159...	107.28931972...
5 F0280EE8	M	14	98.982069499...	110.14093775...

Figure 25: Before and After Data Example Using a Federation Server Data Masking Function

Another new feature in SAS Federation Server is the addition of a number of available new source database types that you can read data from. Some of the available types are shown in Figure 26.

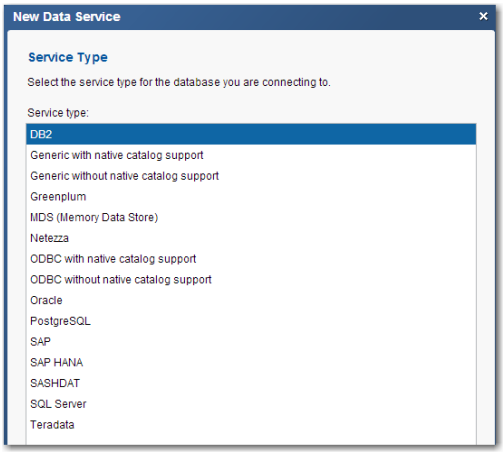

Figure 26: Federation Server Supported Data Sources

CONCLUSION

The latest releases of SAS Data Integration Studio and SAS Data Management products provide 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, enhanced metadata management capabilities, and new features in support of big data. 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
- McIntosh, Liz, et al. 2014. “Understanding Change in the Enterprise.” *Proceedings of the SAS Global Forum 2014 Conference*. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings14/SAS396-2014.pdf.
- Rausch, Nancy, et al. 2013. “What’s New in SAS Data Management.” *Proceedings of the SAS Global Forum 2013 Conference*. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings13/070-2013.pdf.
- Rausch, Nancy, et al. 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/proceedings13/077-2013.pdf.

CONTACT INFORMATION

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

Nancy Rausch
940 NW Cary Parkway, Suite 201
Cary, NC 27513
SAS Institute Inc.
Work Phone: (919) 677-8000
Fax: (919) 677-4444
E-mail: Nancy.Rausch@sas.com
Web: support.sas.com

Mike Frost
940 NW Cary Parkway, Suite 201
Cary, NC 27513
SAS Institute Inc.
Work Phone: (919) 677-8000
Fax: (919) 677-4444
E-mail: Mike.Frost@sas.com
Web: support.sas.com

Mike Ames
100 Campus Drive
Cary, NC 27513
SAS Institute Inc.
Work Phone: (919) 677-8000
Fax: (919) 677-4444
E-mail: Mike.Ames@sas.com
Web: support.sas.com