Paper SAS195-2017

What's New in SAS® Data Management

Nancy Rausch, SAS Institute Inc.

ABSTRACT

The latest releases of SAS® Data Management software provide a comprehensive and integrated set of capabilities for collecting, transforming, and managing your data. The latest features in the product suite include capabilities for working with data from a wide variety of environments and types including Hadoop, cloud data sources, RDBMS, files, unstructured data, streaming, and others, and the ability to perform ETL and ELT transformations in diverse run-time environments including SAS®, database systems, Hadoop, Spark, SAS® Analytics, cloud, and data virtualization environments. There are also new capabilities for lineage, impact analysis, clustering, and other data governance features for enhancements to master data and support metadata management. This paper provides an overview of the latest features of the SAS® Data Management product suite and includes use cases and examples for leveraging product capabilities.

INTRODUCTION

Figure 1. The Data Management Lifecycle

The latest releases of SAS® Data Integration Studio, DataFlux® Data Management Studio, and other SAS® Data Management features provide new and enhanced features to help data warehouse developers, data integration specialists, and data scientists carry out data management tasks more efficiently and with greater control and flexibility. Major focus areas for the latest releases include features to assist in managing all aspects of your data throughout the data lifecycle as illustrated in Figure 1. This paper will showcase some of the newest features available in the SAS® Data Management products.

DISCOVER AND INGEST

There are a number of new data connectivity features in support of discovering and ingesting new data content. New features include enhanced support of big data platforms such as Hadoop and the Cloud. There are also a number of enhancements to existing data connectivity features for improved performance by supporting more DBMS pushdown features. Finally, there are new options for regulatory compliance. Here is a summary of these enhancements.

CONNECTING TO SAS® VIYA™

SAS® Data Integration Studio, SAS® Enterprise Guide®, SAS® Enterprise Miner™, and other SAS® applications have added new capabilities to integrate with SAS® Viya™ and SAS® Cloud Analytic Services. These capabilities package up the SAS/CONNECT® code needed to support transferring data and programs to the SAS Viya platform.

SAS Data Integration Studio has implemented a new transform that supports this integration. In order to use the transform, you first register a SAS Cloud Analytic Server (CAS) definition and a new application

server in the SAS® Metadata Repository that contains a SAS Connect Server to the SAS Viya Platform. You can then register a library that references that application server.

Once you have set up the environment in the SAS® Metadata Server, you can use the transformation to move data and programs to and from the CAS environment. You can also submit your own code via the code tab in the transformation to the CAS server, and the code will execute in the CAS system. Figure 2 illustrates the transform and a partial example of the generated code in SAS® Data Integration Studio.

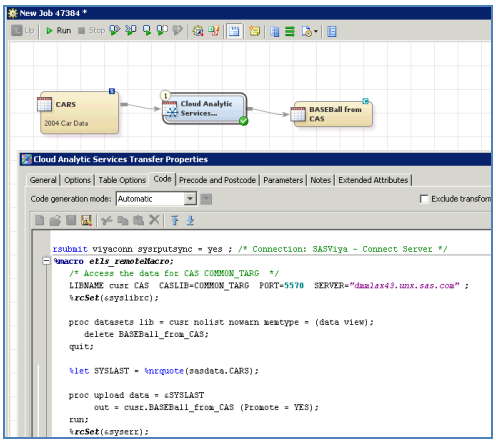

Figure 2. Cloud Analytic Server Transform in SAS® Data Integration Studio

TRANSFERRING DATA TO AND FROM THE CLOUD

There are a number of new features in SAS for supporting data access to and from cloud data sources. One new feature is integration with Amazon Redshift. Available from Amazon web services (AWS), Redshift is a highly scalable cloud data storage system. The SAS/ACCESS® to Amazon Redshift engine, introduced in the third maintenance release for SAS® 9.4, has been enhanced to support bulk data loads for increased performance. This data source type has also been added to SAS® Data Integration Studio and other SAS clients for easy connectivity of this data source to jobs and reports. Enhancements have also been made to support SQL pushdown to Amazon Redshift for SAS procedures, including FREQ, MEANS, SUMMARY, and TABULATE.

For moving data to and from various cloud sources, SAS supports secure FTP (SFTP) via the FILENAME statement. For content specific to Amazon, SAS has also released a new procedure, PROC S3. Figure 3 illustrates these two methods.

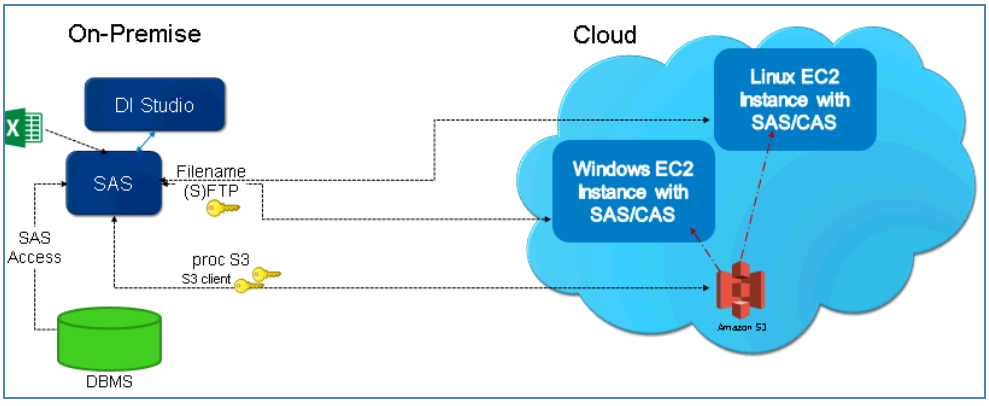

Figure 3. Overview of Some Methods for Transferring Data to and from Cloud Data Sources

SAS Data Integration Studio has added three new transforms that make it easier to generate SAS code using these methods. An example of the new transforms is shown in Figure 4.

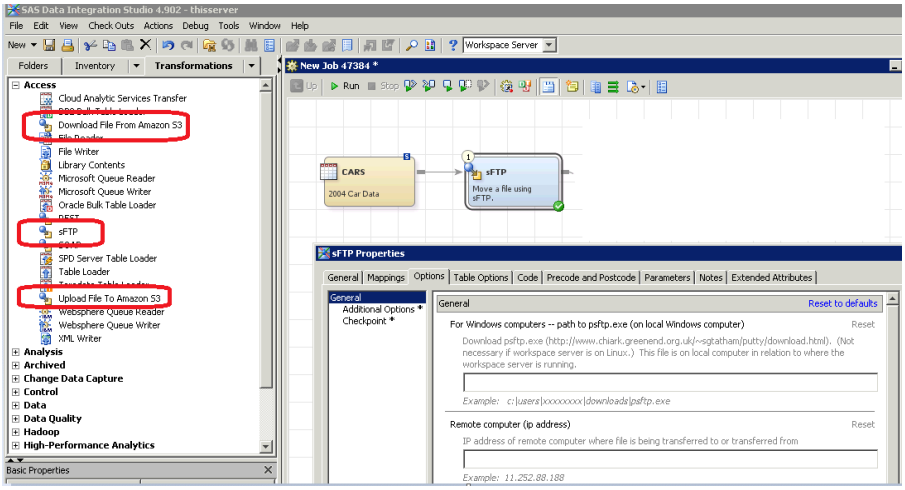

Figure 4. SAS Data Integration Studio Cloud Transforms

The SFTP transformation uses the Secure File Transfer Protocol (SFTP) access method in the SAS FILENAME statement to provide a secure connection for file transfers to and from the cloud. Both commands and data are encrypted. You can use the SFTP transformation to securely access servers, navigate to the correct directories, and transfer data securely to and from the cloud.

To connect and transfer files with SFTP, there must be an SFTP server running on the instance. In most Linux instances, an SFTP server will be running by default. For Windows instances, you must start an SFTP server on the instance and open the port that it is listening on. On the local host, you will also need the PSFTP executable to enable the transfer. PSFTP is the PuTTY SFTP client tool for transferring files securely between computers using an SSH connection. The file psftp.exe will need to be referenced either in the FILENAME statement with the PATH= parameter or you can update the PATH environment variable on your PC to include the folder in which psftp.exe is located.

If you connect to Amazon Web Services (AWS) for your Cloud Computing Service, you are also provided with a key pair. Amazon Web Services (AWS) stores the public key and you store the private key. Amazon provides a Privacy Enhanced Mail Security Certificate also known as a PEM file, which gets stored on your computer. If you are connecting to a Linux instance from Windows, you must convert this file to a PuTTY Private Key (*.ppk) file type using PuttyGen, which is a tool available in the PUTTY system.

Here is an example of the code that the SAS Data Integration Studio SFTP transform will generate for transferring data using SFTP with an Amazon key pair.

```
filename remote sftp "/home/ec2-user/class.csv" /* the file to transfer */
                     host="10.345.678.90"
                     user="ec2-user"
                     optionsx="-i C:\file_location\key_name.ppk" /*(key
from Amazon)*/
                     path="C:\file_location\psftp.exe" debug;
quit;
```

PROC S3 is another new feature in the fourth maintenance release for SAS 9.4. The two SAS Data Integration Studio transforms shown in Figure 4 support uploading and downloading data to Amazon Simple Storage Service (Amazon S3). Using the new PROC, you can create Amazon buckets and add files to S3.

Similar to the SFTP transfer mechanism when using AWS, you will need to set up a public and private key. This information can be passed in the options of the SAS® Data Integration Studio transform, or can

be set into a configuration file. The configuration file contains the details about your connection. Following illustrates a sample set of options:

```
ssl=yes     // YES if data encryption is being used
keyId=AKFKI8OMEVIM3XJHKEUQ // your access key from Amazon
secret=wb89GergI/3xejxudQugFj5Wi4iqlFJhGpLvYVv  // your secret key
region=usstd // AWS region for the connection
```

Once you have configured your settings, you can then specify the file location in Amazon that you want to use to either read or write content. The file has to be relative to the bucket being used, similar to the following.

```
my Amazon S3 bucket name/my Amazon S3 folder/my Amazon S3 file
```

Your files should now transfer correctly. Note that Amazon S3 also has a nice feature for parallel data transfers. If S3 detects that the size of the data being transferred is over some threshold, S3 will automatically open additional threads to move the data in parallel.

CONNECTIVITY AND PUSHDOWN OPTIMIZATIONS

A number of SAS/ACCESS engines have been enhanced in the latest releases for additional database SQL pushdown. SAS/ACCESS to Azure, SQL Server, Postgres, and Vertica all have been enhanced to support pushdown of FREQ, MEANS, SUMMARY, TABULATE, and other SAS procedures. These procedures will now execute in database, which can greatly improve their performance.

Other important connectivity features include:

- SAS/ACCESS® Interface to PC Files now supports writing Microsoft Excel (.xlsx) format
- Kerberos is a popular connection protocol used by various data environments, including Teradata, for secure, password-less connectivity. In addition to already available connectivity using Kerberos, SAS has added support for connecting to Teradata using Kerberos.
- SAS® Data Quality Server can now communicate with secured instances of SAS Data Management Server. The language elements in SAS Data Quality Server use HTTPS to communicate with instances of SAS Data Management Server that are protected with SSL.

HADOOP INTEGRATION AND CONNECTIVITY

There are new Hadoop related connectivity features in the latest releases of SAS. Updates include support for the latest Hadoop vendor distributions, including more Kerberos options. Hadoop deployment options for SAS content have also been simplified. You can now use the SAS Deployment Manager to pull Hadoop JAR and configuration files for your SAS deployments. Below are some additional enhancements:

- SAS/ACCESS to Hadoop has added support for Sentry and Knox, which are features in the Hadoop platforms for enhanced table security
- For the SAS® In-Database Code Accelerator for Hadoop, the SPD Engine SerDe can be used to access Hive tables.
- Partitioned Avro and Parquet files are now supported
- SAS® Data Loader for Hadoop 3.1 now supports new IT-friendly administrative deployment process using SAS® Deployment Manager, and high-performance connectivity options using SAS/ACCESS software
- SAS® Data Integration Studio now can run SAS® Data Loader jobs
- SAS® Data Integration Studio had additional in-database processing via SQL pushdown to Hadoop for PROC TRANSPOSE.

SAS FEDERATION SERVER

SAS® Federation Server is a data server that provides scalable, threaded, multi-user, and standards-based data access technology in order to process and seamlessly integrate data from multiple data services. The server acts as a data hub and supports accessing, managing, and sharing data from database systems, Hadoop, SAS, and other data sources and targets. SAS® Federation Server also supports security and data source management. You can apply security to data from many sources without moving or copying the data.

There are several new features available in SAS® Federation Server. Support has been added to read and write from SAS Scalable Performance Data Server data server. SAS Federation Server is also a key component of the SAS solution for protection of personally identifiable information (PII) as part of the European Union (EU) General Data Protection Regulation (GDPR) support. The GDPR initiative will be enforced starting in May of 2018, and will require full disclosure of how personal data is being used, kept, protected, and deleted. Companies that fail to comply will face high fines.

SAS Federation Server is an excellent solution for PII and GDPR initiatives because it supports the ability to manage secure data access with minimal disruption to existing data flows. For example, SAS® Federation Server security features include role-based data access, row-based security, and data masking such as hashing, randomization, and encryption, as data flows through the server. This means that you can implement a data protection methodology with minimal interruption to existing systems. Figure 5 is an example of how SAS Federation Server can be used to support PII and GDPR initiatives.

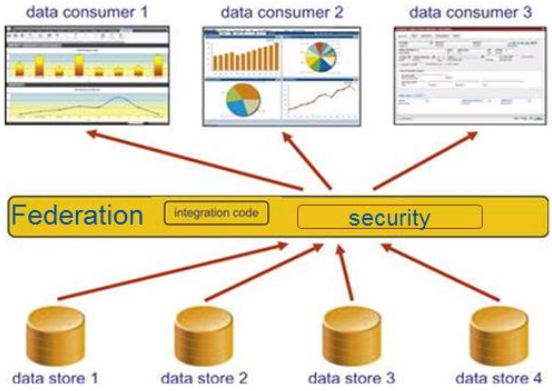

Figure 5. SAS Federation Server Security Architecture Overview

CLEANSE AND STRUCTURE

There are new and updated features in support of cleansing and structuring data. SAS® Data Remediation has a new feature to support custom user interfaces for correcting data issues records. This feature allows data stewards to fix data quality problems in the remediation interface without having to write any code.

To enable this feature, we added another set of options to the remediation administration screen. This allows you to set the data management server to use that contains two user-written jobs. The first job is the one you want to use to retrieve the record, and the other is used to fix the record in the database. Figure 6 is an example of these options.

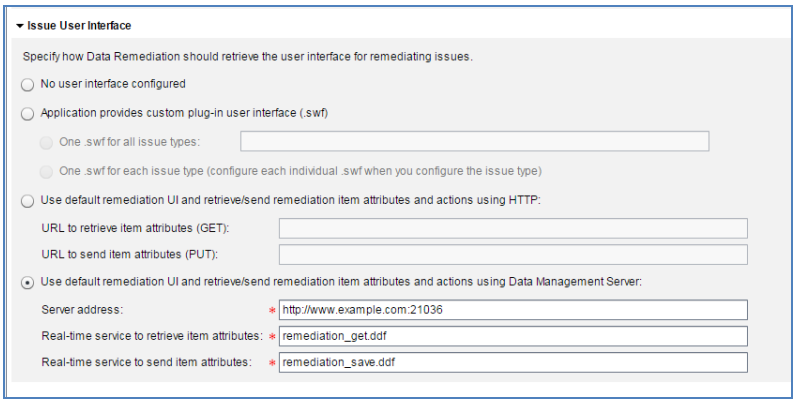

Figure 6. New SAS Data Remediation Features

After the administrative interface is configured, you can fix data issues in the remediation application. The remediation application will display a user interface that mirrors the fields defined in your jobs. Figure 7 is an example of the job and the corresponding user entry fields in the remediation interface. Remediation also has several predefined fields that can be programmatically queried in your data management jobs to identify settings such as the application URL, pass in credentials, and other integration features.

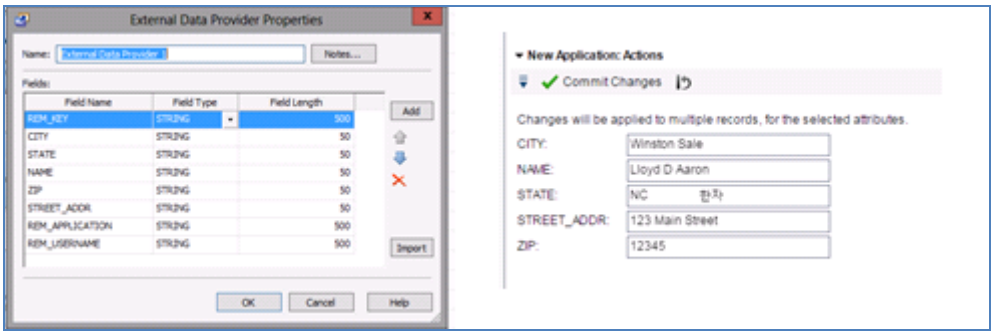

Figure 7. Example of New Remediation Features

Another new feature in remediation is the ability to invoke an external application using the action buttons on the remediation toolbar. The presence or active states of the buttons are controlled by SAS workflow status that contains an attribute with a valid URL. We also allow value substitution by allowing the value of named data objects at the status level to be used as place holders in braces. Figure 8 shows how you would configure this feature in SAS® Workflow Studio, and how the setting is presented in the remediation application. In the example, a user would be adding an e-signature to the issue, indicating that it has been resolved.

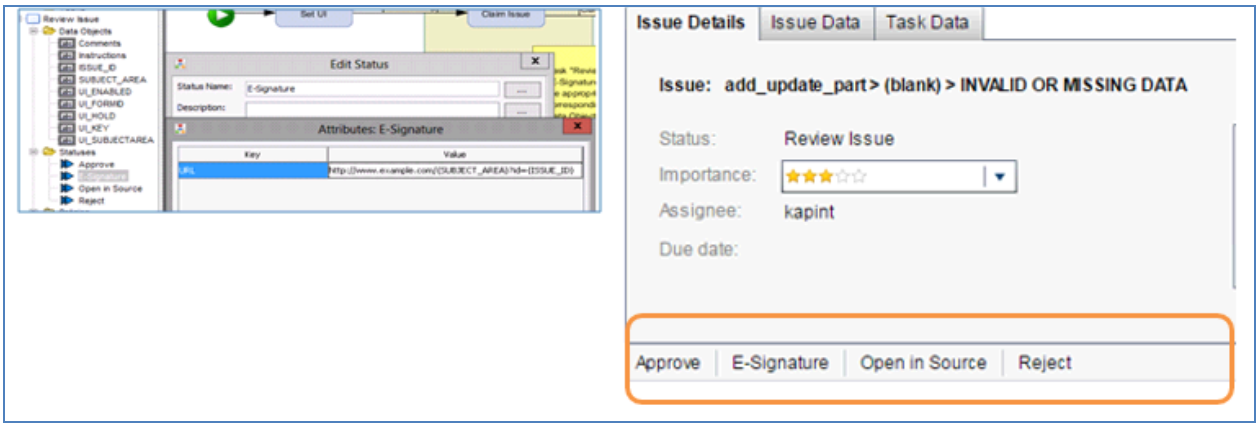

Figure 8. Example Use of New Remediation Features

TRANSFORM AND LOAD

Data integration is the process of consolidating data from a variety of sources in order to produce a unified view of the data. SAS® Data Integration Studio provides a powerful visual design tool for building, implementing and managing data integration processes for most types of data sources, applications, and platforms. An easy-to-manage, multiple-user environment enables collaboration on large enterprise projects with repeatable processes that are easily shared.

SAS Data Integration Studio has new connectivity options documented above in the data connectivity section. Another new feature in SAS Data Integration Studio is the ability to run Hadoop jobs created in SAS® Data Loader. A new transformation allows you to select which saved SAS Data Loader jobs you want to run. One advantage of this new feature is to support integrated impact analysis. You can now see impact analysis across both SAS and Hadoop environments. Figure 9 and Figure 10 illustrate some of these new features.

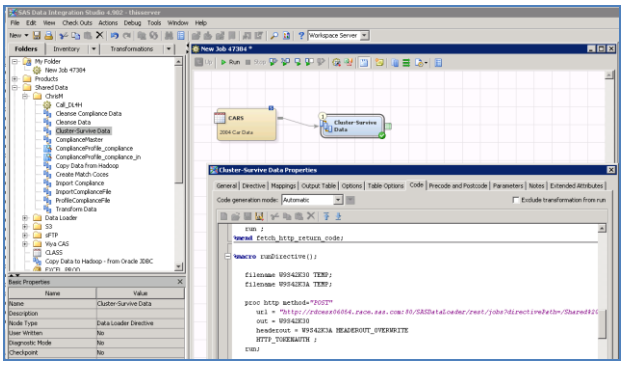

Figure 9. SAS Data Loader Integration in SAS Data Integration Studio

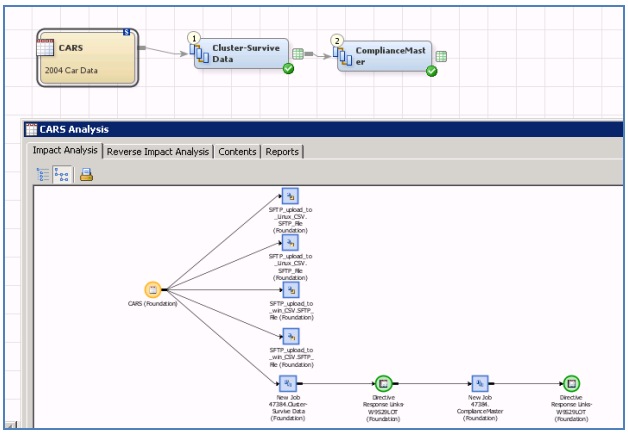

Figure 10. SAS Data Loader Jobs and Impact Analysis in SAS Data Integration Studio

Several enhancements have been made to better ensure that the code generated in SAS® Data Integration Studio will run optimally in the Hadoop database. One useful setting allows you to disable the generation of column formats. This is important because formats cannot be expressed in the Hadoop execution environment, and so data would transfer unnecessarily between the two environments. By setting this option in SAS Data Integration Studio, you can avoid this extra data transfer. Figure 11 illustrates how to set this option on new and existing jobs.

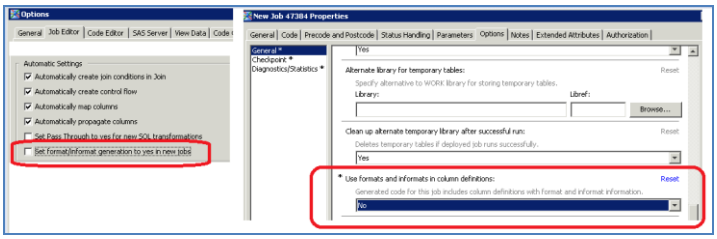

Figure 11. Disabling Column Formats Option

SAS® DATA LOADER FOR HADOOP - BIG DATA ETL

When traditional data storage or computational technologies struggle to provide either the storage or computation power required to work with large amounts of data, an organization is said to have a big data issue. Big data is frequently defined as the point at which the volume, velocity, and/or variety of data exceeds an organization's storage or computation capacity for accurate and timely decision-making.

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

SAS integrates into the Hadoop platform to bring the power of SAS to help address big data challenges. SAS, via the SAS/ACCESS® technologies and the SAS embedded code accelerator products, has been optimized to push down computation and augment native Hadoop capabilities, allowing user to use SAS to process data stored in Hadoop. By reducing data movement, processing times decrease and users are able to more efficiently use Hadoop resources.

SAS® Data Loader for Hadoop is the SAS offering that provides support for big data. Figure 12 is an overview of the architecture of SAS® Data Loader.

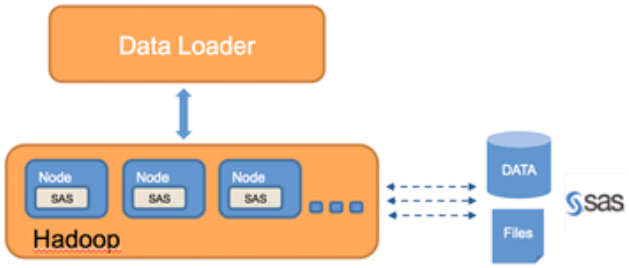

Figure 12. SAS Data Loader Architecture

SAS® Data Loader includes a client for building and running jobs in Hadoop that can leverage both Hadoop and SAS embedded code accelerator capabilities, and the components required to install and distribute SAS on the Hadoop cluster to enable Data Quality, ETL, and Analytic Data Prep features using SAS and Hadoop. Figure 13 is a screenshot of the main screen of the SAS® Data Loader client showing some of the transformations and features available to work with data in Hadoop.

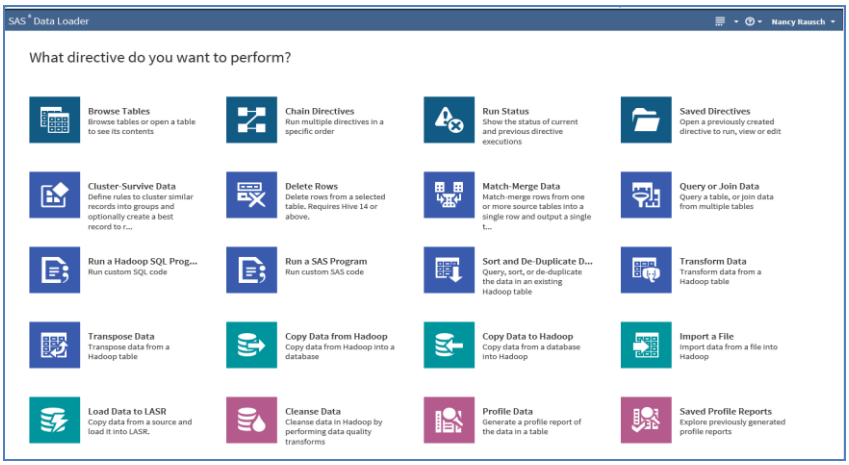

Figure 13. Data Loader Main Screen

SAS® Data Loader has been significantly enhanced in the latest release. The multi-user web application is now fully integrated into the SAS Intelligence Architecture. This enhancement enables multiuser capabilities, content sharing using SAS Folders, integration with the SAS® Metadata Repository, an easier deployment process by leveraging the standard SAS deployment tools, and other features. There is also updated support for the latest Hadoop distributions, and enhancements for Hadoop features such as better Kerberos integration and High Availability.

Using the latest release, you can now manage servers, authorizations, and directives through SAS® Management Console Figure 14 is an example of some of the new Hadoop environment and web application configuration settings in SAS Management Console.

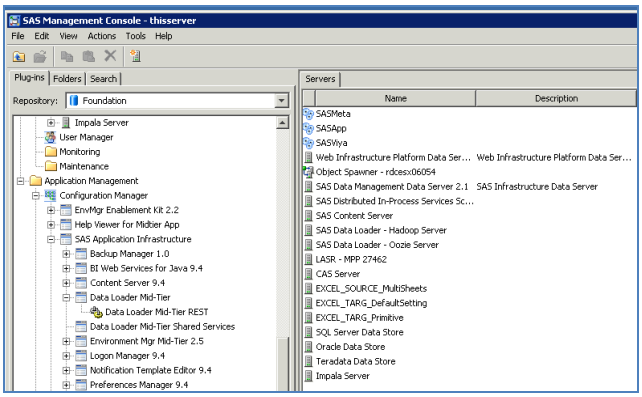

Figure 14. SAS Data Loader Configuration Settings in SAS Management Console

Figure 15 illustrates some of the new features in support of enhanced authorization and security. Administrators can now secure both directive templates and saved jobs. For example, if you want to limit which users can move data into and out of Hadoop, you can now set permissions on specific directive templates. Those users that don't have the appropriate permissions will not see those templates in their application, which gives them a customized application user experience modified for their site-specific permissions.

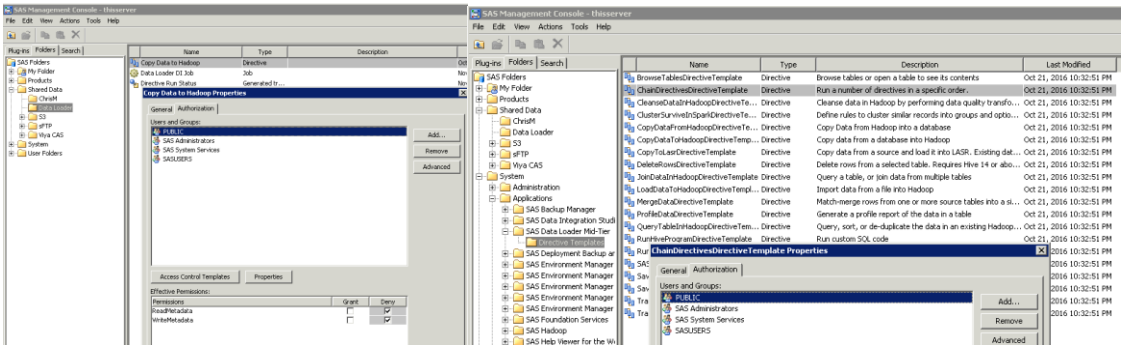

Figure 15. Permissions and Authentication Options for SAS Data Loader

You can also save and manage saved jobs in SAS folders for sharing content with other users and for integrating with other SAS applications such as SAS® Data Integration Studio. You can even leverage the permission capabilities of SAS folders to control access to saved jobs. Figure 16 illustrates this feature.

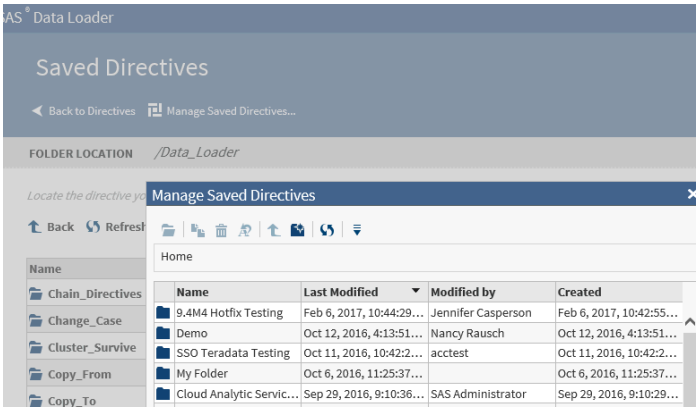

Figure 16. Data Loader Directives in SAS Folders

There are also now many more high-performance connectivity options. You are now able to use the SAS/ACCESS technologies installed in your environment to support moving data to and from your

Hadoop system using SAS Data Loader. All registered SAS libraries are accessible to you from the SAS Data Loader interface. Figure 17 illustrates this.

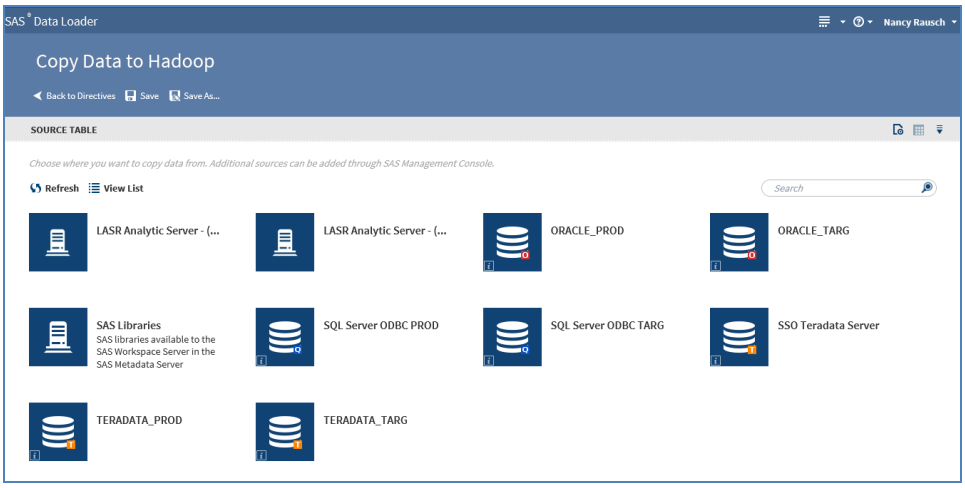

Figure 17: Many Data Connection Options in SAS Data Loader

SAS® Event Stream Processing

SAS® Event Stream Processing enables the analysis of real-time or near-real time data called event streams. These streams comprise events that occur as the result of an action or set of actions, such as a financial transaction, equipment failure, or some other trigger. These triggers can be very granular, such as something that happens within a system at a point in time—a click, a sensor reading, a tweet or some other measurable activity. The growing number of connected devices—the Internet of Things—is exponentially increasing the volume of events that surround business activity. SAS Event Stream Processing derives actionable information from these event streams in real-time using SAS for data transformation, data quality and cleansing, analytics, and pattern-of-interest detection.

There are a number of new features available in SAS Event Stream Processing. Access controls have been added, which allow you to use explicitly set read/write permissions on engine, project, query, and window objects on a per user basis.

Window objects are the transformations that you can make on the data streams as they data flows through the SAS Event Stream Processing engine. New windowing features include a new geofence window for tracking devices in and out of a specific area; and a new text topic window, which allows you to process text content. There are also enhancements to the connectors and adapters that allow the SAS Event Stream Processing technology to extract data from external systems for processing.

MANAGE AND MAINTAIN

The final phase in the data management lifecycle is to manage and maintain the data so that it continues to remain high quality and meet the needs of the data consumers. One of the key capabilities of SAS in this phase of the data lifecycle is the SAS® Business Data Network. This feature enables collaboration of domain knowledge between business, technical, and data steward users. The SAS Business Data Network can be used as a single entry point for all data consumers to better understand their data.

There are several key new features that have been added to SAS Business Data Network. One of the most useful for data governance initiatives is the ability to support snapshots. Snapshots provide the ability to back up the entire business term repository, so that users and regulators can view previous versions of the dictionary, which can help with regulatory compliance.

Snapshots are created and deleted in the manage snapshots dialog box, available from the main menu of SAS Business Data Network. In order to create, delete, or view snapshots, an administrator can grant authorization via SAS Management Console. Once authorized, users can open previous versions of the dictionary and view their content. Figure 18 is an example of a snapshot view.

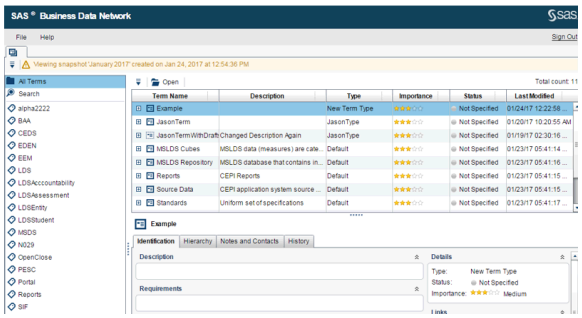

Figure 18. Snapshots in SAS Business Data Network

There are also a number of new attribute types including Date, URL, and RTF types. Figure 19 shows a usage example of the RTF type.

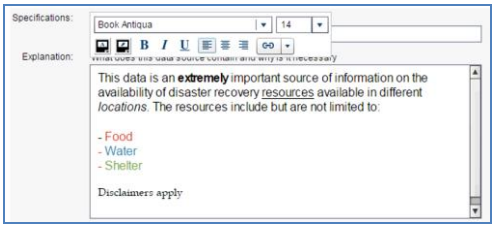

Figure 19. Example of the New RTF Data Type in SAS Business Data Network

There are also several new features that simplify populating the dictionary. You can now import of content into SAS Business Data Network from CSV files. Figure 20 is an example of the CSV import syntax.

Name	path	Type	Description	Requirements
CSVBasic1	CSVBasic1			requirements
CSVBasic2	CSVBasic1	Default		
CSVBasic3	CSVBasic1\CSVBasic2\CSVBasic3		Description, w	
CSVBasic4	CSVBasic1\CSVBasic2\CSVBasic3		Description	requirements
CSVBasic5				

Figure 20. CSV Import Example

Import performance has also been improved by allowing you to disable populating lineage until import is complete. This enhancement can significantly speed up import performance when importing a lot of content at one time.

SAS Business Data Network also has added an application programming interface (API). The API allows you to interact with the dictionary programmatically, such as from other applications. Some of the key features of the interface include:

- Request any object from the dictionary such as tags, terms, sets of terms and term types
- Create, rename, and manage terms
- Change, update, manage attributes stored in the term such as references, associated objects, contacts, and status
- Programmatically manage workflow for a term, such as in review, editing, published, and retired.

Figure 21 is an example of the API.

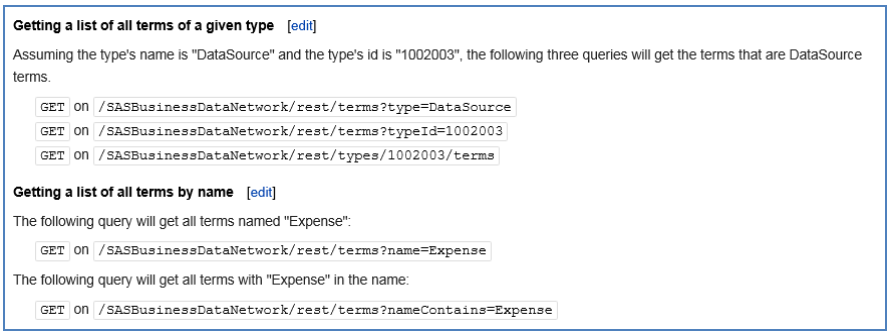

Figure 21. Externally Available Programmatic Interface for SAS Business Data Network

Lineage has also been enhanced to allow users to open objects into a preferred application. This allows users to identify and open an object directly from the lineage viewer. For example, you can link back to a business term in the BDN from the viewer. Figure 22 is an example of this new feature.

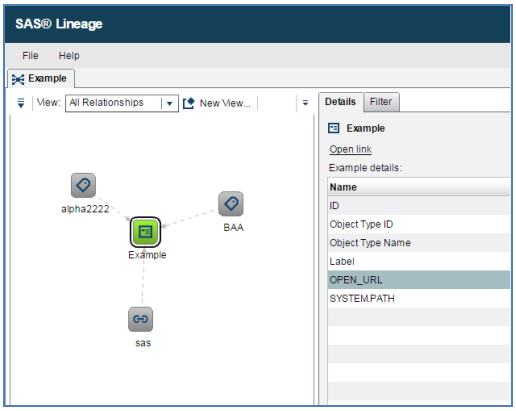

Figure 22. Lineage Integration Enhancements with SAS Business Data Network

SAS® Master Data Management

SAS® Master Data Management automates the process of selecting the best records from source data. SAS Master Data Management 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 identify a best record based on analytical techniques. 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 23 is an example of a view in SAS Master Data Management showing an example set of incoming records and the selected best record:

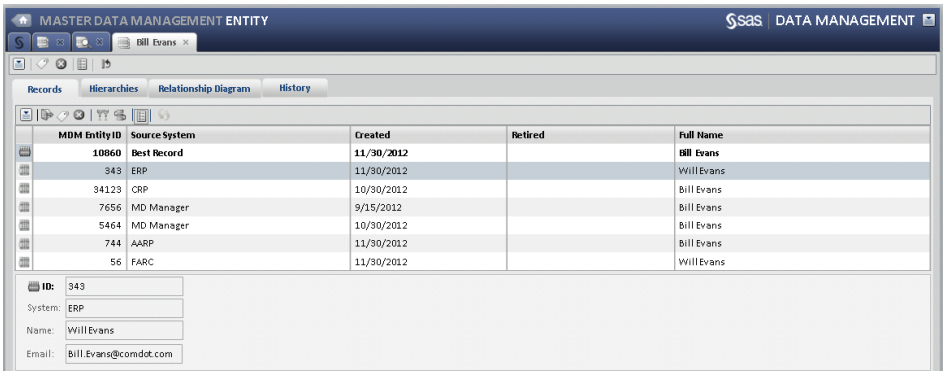

Figure 23. SAS Master Data Management New Features

SAS Master Data Management 4.3 has added the following new features:

- Added support for SQL Server 2014 as another database option for hosting the SAS Master Data Management data hub.
- Users can now set configurable limits on the number of records presented when viewing clusters. Clusters can get very large over time, and you might need to only see a sample of the cluster in order to understand the data in it. This feature will allow you to control the number of records you view.
- Automated cleanup of remediation issues. Previous versions of SAS Master Data Management required manual maintenance of the remediation database to remove data for closed issues. In this release, these issues are removed from the database automatically.
- A new feature allows administrators to control which users are able to view, create, and search for hierarchies in the user interface.

Figure 24 is an example of the cluster viewer in SAS Master Data Management, with configurable limits set. In the example, SAS Master Data Management has identified the many different spellings of a specific company name, and automatically clustered the results together.

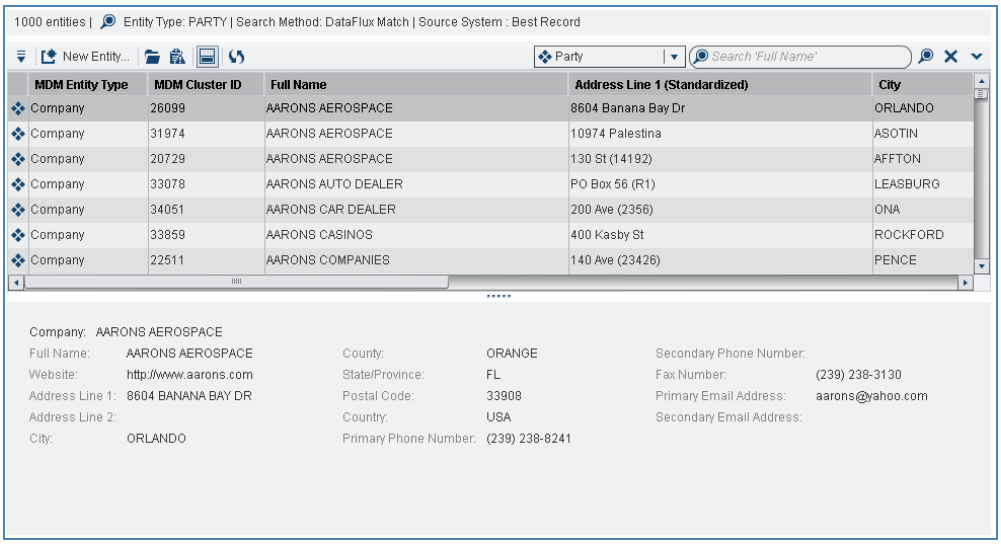

Figure 24. Enhancements to the SAS Master Data Management Cluster Viewer

CONCLUSION

The latest releases of 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. Enhancements have been made in all areas of the data management lifecycle. Customers will find many reasons to upgrade to the latest versions of SAS Data Management.

RECOMMENDED READING

- SAS® Data Management Community. Available at https://communities.sas.com/t5/Data-Management/ct-p/data_management
- Hazejager, W. & Rausch, N., 2016. “Ten Tips to Unlock the Power of Hadoop with SAS”. *Proceedings of the SAS Global Forum 2016 Conference*, Cary, NC. SAS Institute. Available at http://support.sas.com/resources/papers/proceedings16/SAS2560-2016.pdf

- Rausch, Nancy, et al. 2016. “What’s New in SAS Data Management.” *Proceedings of the SAS Global Forum 2016 Conference*. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings16/SAS2400-2016.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
Email: Nancy.Rausch@sas.com
Web: support.sas.com