Paper SAS2400-2016

What's New in SAS® Data Management

Nancy Rausch and Ron Agresta, SAS Institute Inc.

ABSTRACT

The latest releases of SAS® Data Integration Studio, SAS® Data Management Studio and SAS® Data Integration Server, SAS® Data Governance, and SAS/ACCESS® 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, 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

The latest releases of SAS® Data Integration Studio, DataFlux® Data Management Studio, and other SAS data management products provide new and enhanced features to help 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 latest releases include a number of new features in support of big data, data management, data governance, and data federation. This paper showcases some of these latest features.

SAS AND 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 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.

SAS has extensive integration options with Hadoop to bring the power of SAS to help address big data challenges. SAS, via the SAS/ACCESS® technologies and SAS® In-Database Code Accelerator products, has been optimized to push down computation and augment native Hadoop capabilities to bring the power of SAS 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.

Hadoop support is available in SAS Data Integration Studio, the SAS/ACCESS products, the SAS® In-Database Code Accelerator for Hadoop, and the SAS® Data Quality Accelerator for Hadoop. SAS Data Loader for Hadoop (SAS Data Loader) is the primary SAS offering that combines all of these components into a single offering to provide support for big data with Hadoop.

Figure 1 is an overview of the architecture of SAS Data Loader.

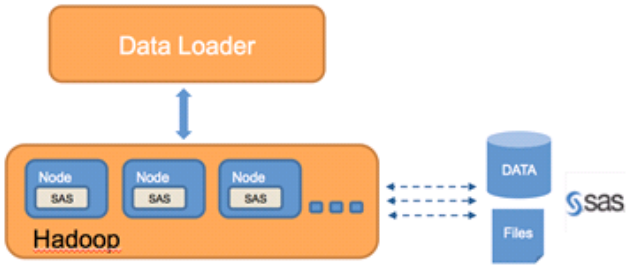

Figure 1: SAS Data Loader for Hadoop Architecture

The SAS Data Loader offering includes a client for building and running jobs in Hadoop that can leverage both Hadoop and SAS embedded process 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 Hadoop. Figure 2 is a screenshot of the main screen of the SAS Data Loader client showing some of the main transformations and features available to work with data in Hadoop.

Figure 2: SAS Data Loader Main Screen

SAS Data Loader integrates with native Hadoop capabilities such as Oozie, Sqoop, and Hive for reading, transforming, and writing data in parallel using Hadoop. It does this by generating multiple Hadoop languages, integrating with native Hadoop application programming interfaces, and generating SAS DATA step code to run in Hadoop. SAS Data Loader manages all of these different languages and the complexities of working with Hadoop for you, so that you can focus on working with your data.

SAS Data Loader includes the SAS Embedded Process technology and generates SAS code to perform various transformations on data in Hadoop. Also included is the SAS Data Quality Accelerator for Hadoop and a SAS Quality Knowledge Base offered in a number of different languages to support data quality actions such as standardizing addresses, state codes, phone numbers, parsing data into standard or customizable tokens, and identifying data into known types such as names and addresses.

SAS Data Loader capabilities include the ability to parallel load data into or out of Hadoop from RDBMS systems, SAS Data sets, and delimited files. Data can also be loaded into a SAS LASR Analytic Server. Figure 3 is an example of the Copy Data from Hadoop directive.

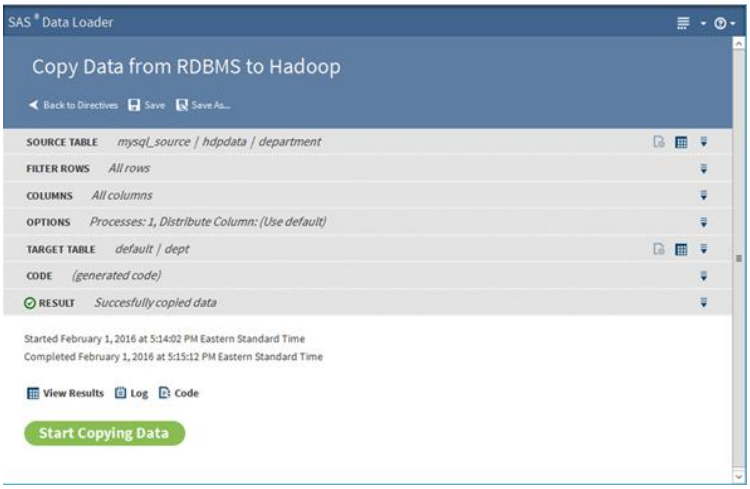

Figure 3: Copy Data Example

Figure 4 is an example of one of the languages that SAS Data Loader generates, in this case Hadoop Oozie and Sqoop code, which are languages native to the Hadoop cluster.

```
<?xml version="1.0" encoding="UTF-8"?>
<workflow-app xmlns="uri:oozie:workflow:0.4" name="CopyDatafromRDBMStoHadoop11dc5e584bfa4ce1a1d5a995aeda79d0">
 <global>
  <job-tracker>${jobTracker}</job-tracker>
  <name-node>${nameNode}</name-node>
  <job-xml>${wf:appPath()}/hive-site.xml</job-xml>
 </global>
 <credentials/>
 <start to="action-1"/>
 <action name="action-1">
  <sqoop xmlns="uri:oozie:sqoop-action:0.4">
   <prepare>
    <delete path="${nameNode}${wf:appPath()}/department"/>
   </prepare>
   <arg>import</arg>
   <arg>--connect</arg>
   <arg>jdbc:mysql://[redacted]/hdpdata</arg>
.....
.....
.....
 <kill name="kill">
  <message>Script failed, error message[${wf:errorMessage(wf:lastErrorNode())}]</message>
 </kill>
 <end name="end"/>
</workflow-app>
```

Figure 4: Example Data Loader Generated Oozie and Sqoop Code

You can also build complex Hadoop Hive and Impala SQL queries and joins of 1-N tables. In this case, SAS Data Loader will generate Hadoop Hive SQL, and Hadoop Impala SQL. The example in **Error! Reference source not found.** shows a three table join, illustrating the types of joins available.

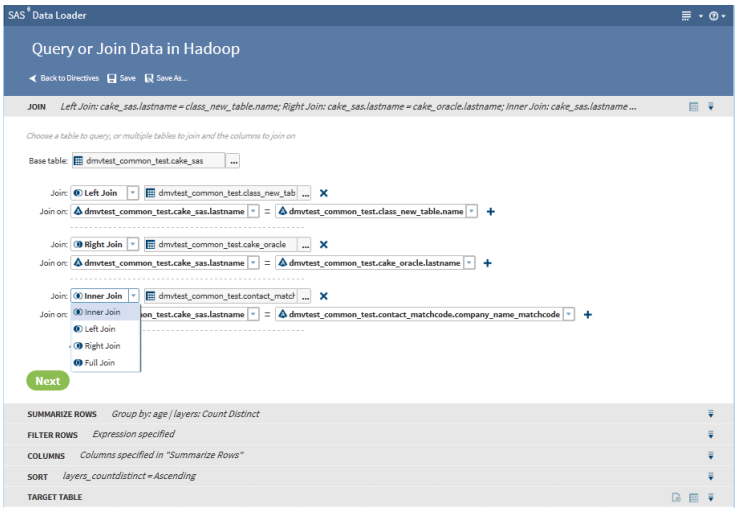

Figure 5: Example Query and Join Directive

You can perform data cleansing operations such as parsing, standardization, filtering, casing, pattern analysis, and others. Figure 6**Error! Reference source not found.** shows some of the data quality transforms available.

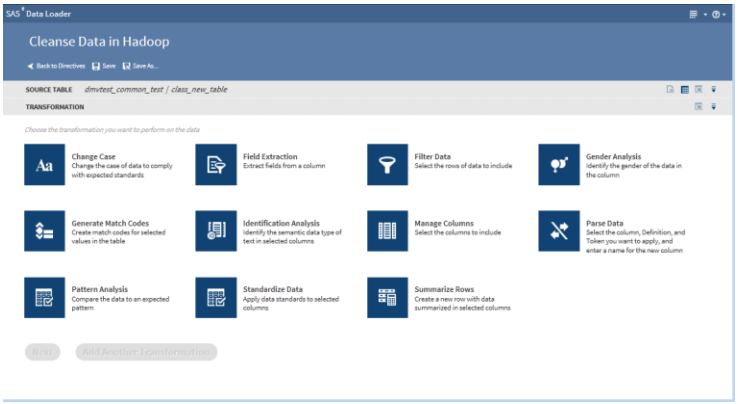

Figure 6: Data Quality Transforms for Hadoop and Spark

You can standardize columns to a variety of pre-shipped standards, or you can add your own standards. Figure 7 shows some of the available standards that are delivered with the SAS Quality Knowledge Base for Contact Information used by SAS Data Quality Accelerator. The SAS Quality Knowledge Base delivers data quality definitions for more than 35 locales.

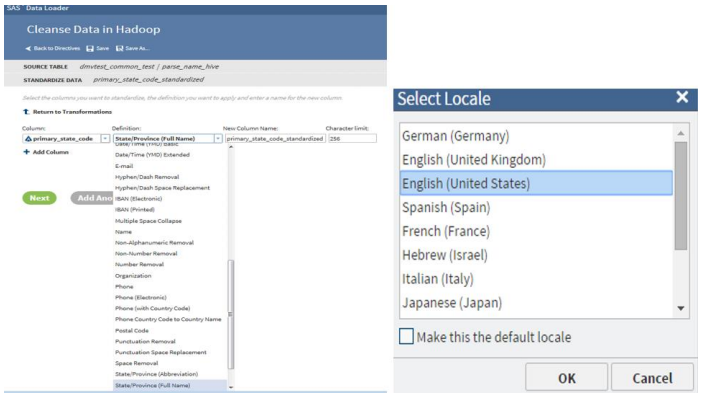

Figure 7: Data Quality Standardization across Multiple Locales

Figure 8 shows the results of applying standardization to data in a Hadoop data set. The original data, outlined in blue (right) 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 (left) in the figure below shows the new column added to the table with state names standardized to the full state name.

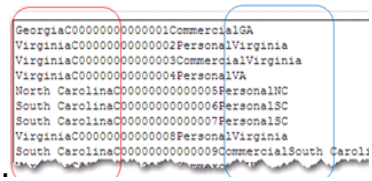

Figure 8: Applying Standard State Codes to Data in Hadoop

One of the key new capabilities available in the latest release includes support for SAS running in Apache Spark. Apache Spark is an optional in-memory feature of Hadoop that can significantly speed up the performance of complex transformations. SAS Data Loader has the ability to automatically detect if Spark is enabled on your Hadoop cluster, and use it when it is available. SAS Data Loader will also automatically fall back to using other native Hadoop processing methods when Spark is not available. This allows you to have the freedom to write your code once and have SAS Data Loader optimize it for performance, based on the capabilities of your system. Figure 9 is an example of the conjuration dialog box in SAS Data Loader that determines which method will get first preference when running your jobs in Hadoop. You can control this automatic optimization feature yourself via the settings on this panel.

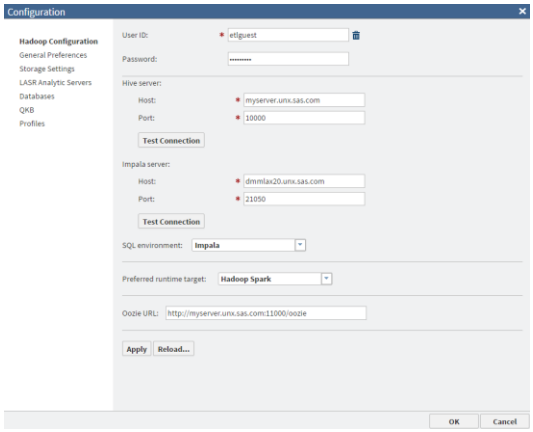

Figure 9: Hadoop Configuration Options in Data Loader

SAS Data Loader can automatically detect when the Hadoop Impala SQL server is installed on your system. Apache Impala has an in-memory SQL capability that can provide significant performance improvements when working with large data and it is fully interchangeable with Hive data sets. Figure 10 is an example of the performance improvements when using Impala to perform SQL transformations.

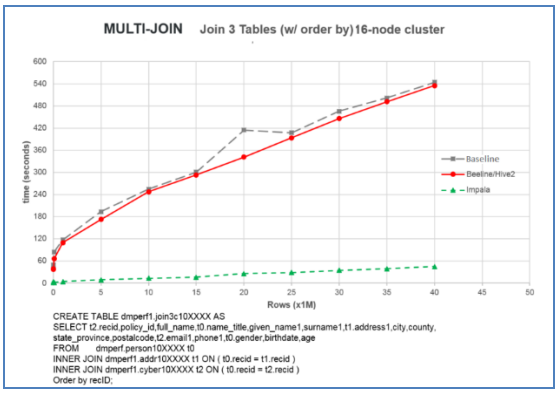

Figure 10: Hive and Impala Performance Graphs

Another new capability in the latest release is the ability to perform SAS DATA step merge running in parallel in Hadoop. DATA Step Merge is a powerful SAS capability that allows you to combine rows from two or more source tables into a single row in a target table. Rows are combined according to the values of one or more matched columns. Figure 11 is an example that illustrates how merge works.

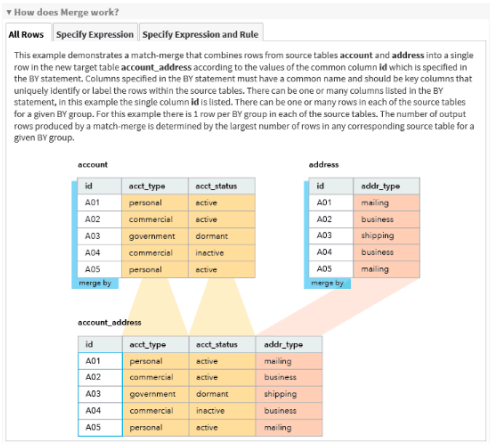

Figure 11: DATA Step Merge Example

The merge has been integrated into the SAS Embedded Process in Hadoop, so the merge will run in parallel inside the Hadoop cluster. This allows many of the familiar features of merge to be performed within the database for example, first. last. processing, row-based matching, and other useful features of the merge statement. There are a number of useful helper features integrated into the user interface to assist in building the SAS merge code, such as automatic rename support, complex row-based expression handling, target column selection, and multiple input files support via the SET statement.

Figure 12 illustrates some of the features available in the merge action, illustrating the ability to automatically identify and rename match columns, and define complex filter conditions to apply to the incoming data.

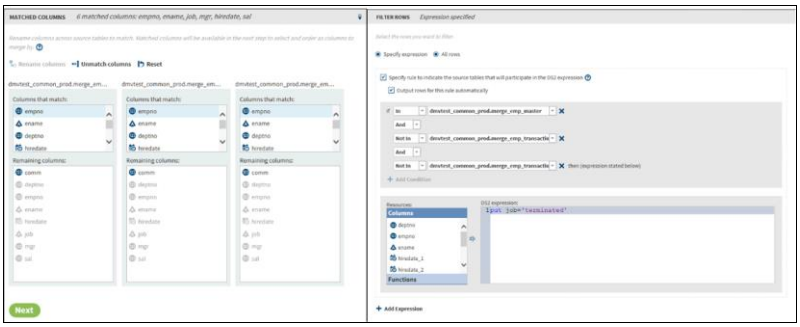

Figure 12: DATA Step Merge Example Features

SAS Data Loader generates a number of languages to optimize in-hadoop performance, including SAS, Hive SQL, Oozie, Sqoop, and other Hadoop specific languages. The merge directive generates SAS DATA step 2 code, which is very similar to SAS DATA step code, except that it supports partitioned data sets and threaded programming. Below is an example snippet of the merge code that SAS Data Loader will generate to perform the merge in hadoop:

```
/* setup some basic syntax ------------------------------------------*/
proc ds2 bypartition=yes ds2accel=yes;
thread t_pgm / overwrite=yes;

/* define what variables to keep-------------------------------------*/
merge mylib.customer_cleansed_nc (keep = (customer_id address_type ..) IN =
inTable1) mylib.order_product_total_nc (keep = (customer_rk bebop
iris_explorer jumping_sumo rolling_spider) rename = (customer_rk AS
customer_id) IN = inTable2)
;
   by customer_id;
   order_total = 0;
   order_volume_ct = 'low';

/* Perform the merge ------------------------------------------------*/
if inTable1 = 1  AND inTable2 = 1  then do;
order_total = SUM(bebop, iris_explorer, jumping_sumo, rolling_spider);
if (order_total <= 200000) then
 order_volume_ct = 'low';
 else
 order_volume_ct = 'high'
output;
end; /* end of if condition */
end;
endthread;

/* Run the program --------------------------------------------------*/
data W0W7FYMA.customer_nc_order_total (overwrite=yes);
   declare thread t_pgm t;
   method run();
      set from t;
   end;
enddata;
run;
quit;
```

SAS Data Loader has a new cluster and survivorship directive that runs in Hadoop Spark that can help you when consolidating data from diverse source systems. Frequently, in this type of scenario, there is often a need to select the best record out of all possible records that represent the same data, so that you can create a clean data set for downstream jobs and reports. For example, you might 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 13**Error! Reference source not found.** is an example of this type of data.

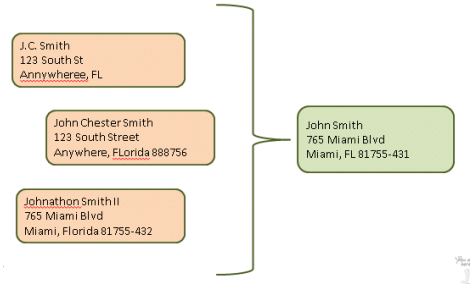

Figure 13: Example of Selecting a Best Record

Using the cluster and survivorship directive, you can match similar rows of data into a cluster, and then select the best record automatically into a single, cleansed, and de-duplicated data record. You can build simple or complex rules to define how you want to cluster the data, and then manage what rules you want to apply to select the best record. Figure 14 is an example of some of the capabilities of the cluster-survivorship directive.

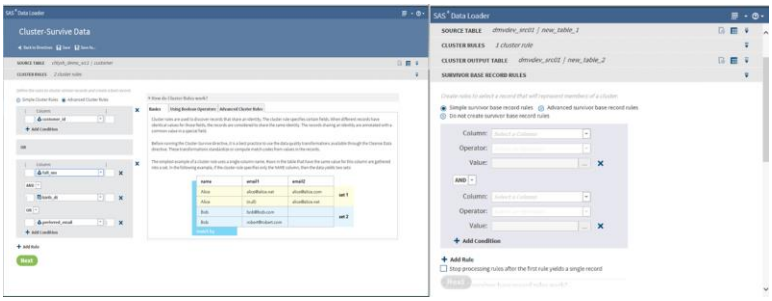

Figure 14: Example of Cluster-Survivorship Directive in Data Loader

The new chained directives feature allows you to run a sequence of jobs in serial or in parallel on the cluster. You can also build up complexity by embedding serial or parallel jobs inside of other jobs. Chained directives also let you define how to handle errors if a job in the sequence fails. Figure 15 is an example of the chained directives view.

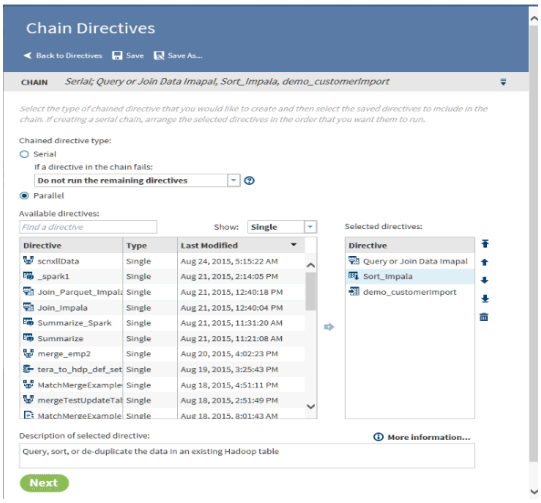

Figure 15: Chained Directives Example

A REST interface is now available for interacting with SAS in the Hadoop cluster. For example, you can query and run jobs in the cluster, query run status, get run-time performance metrics, and other features. The REST interface allows you to use these capabilities from other third-party applications, such as schedulers.

The following example illustrates the execution of a saved SAS Data Loader directive using a curl application interface. Curl is a simple open-source command line tool for calling applications using REST. The figures below are simply examples, you can use any application that can make REST calls to do the same thing as shown below.

To call the SAS Data Loader directive from an external application using REST API, first determine the ID of the directive and locate the URL to run the directive of the directive that you are interested in calling. The following curl statement displays information about the supplied directive name, ProfileData, which includes the directive ID and the URL to run.

```
$ curl -H "[LL1] Accept: application/json"
http://192.168.180.132/SASDataLoader/rest/directives?name=ProfileData
output:
{"links":{"version":1,"links":[{"method":"GET","rel":"self","href":"http://
192.168.180.132:80/SASDataLoader/rest/directives?name=ProfileData&start=0&l
imit=10","uri":"/directives?name=ProfileData&start=0&limit=10"}]},"name":"i
tems","accept":"application/vnd.sas.dataloader.directive.summary+json…other
information…}
```

Once the directive ID and the URL is identified, execute the saved directive using http method POST with parameters as URL to run. Once the POST statement is submitted, it displays JSON text with job ID.

```
$ curl -H "Accept: application/json" -X POST
http://192.168.180.132/SASDataLoader/rest/jobs?directive=7ab97872-6537-
469e-8e0b-ecce7b05c2c5
```

By using the job ID from the previous statement output, you can view the status of submitted job. The following statement shows the status as ”job running” and “completed” when it completes.

```
$ curl -H "Accept: text/plain"
http://192.168.180.132/SASDataLoader/rest/jobs/22/state
output:
running

$ curl -H "Accept: text/plain"
http://192.168.180.132/SASDataLoader/rest/jobs/22/state
output:
completed
```

The status of the SAS Data Loader directive execution, which has been called and executed using REST API can also be viewed and monitored in the “Run Status” SAS Data Loader interface window. Figure 16 is an example of the Run Status view when the ProfileData job was started via REST API.

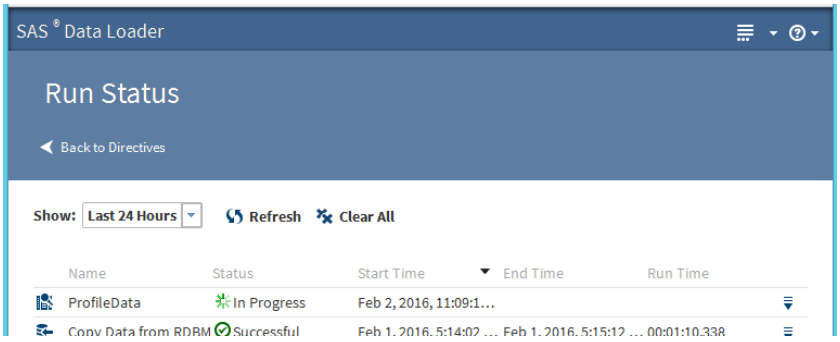

Figure 16: Example of Job Submitted via the REST API

There are a number of new distributions and deployment options enhancements to enable you to more easily evaluate data loader on your own Hadoop distributions. There is a new trial download that supports integration with cluster management tools such as Cloudera Manager and Hortonworks Ambari

for ease of deployment of the SAS components into the Hadoop cluster. There are also additional distributions supported including MapR, IBM BigInsights, and Pivotal, and enhancements to support the latest versions of Cloudera and Hortonworks. An example of some of the new deployment options is shown in Figure 17.

Figure 17: Example of New Deployment Options

SAS DATA INTEGRATION STUDIO

SAS Data Integration Studio supports traditional ETL and ELT capabilities using SAS, SQL, and pushdown SQL. The latest release of SAS Data Integration Studio has added a number of new features. One of the key enhancements is the addition of a new process Fork node that allows you to run multiple flows in parallel in a job. The fork node will spawn a new parallel SAS process when it is run inside of a job. All of the nodes between the Fork and the Fork End transform will run in a parallel process. The Wait For Completion transformation uses the output from multiple Fork transformations, allows the job to wait for all or any of the processes to complete, and then creates a single output. The Fork also supports Grid processing when Grid is available, and works similar to the existing Loop transform. Figure 18 is an example of the new Fork and Fork End transforms, showing two parallel processes.

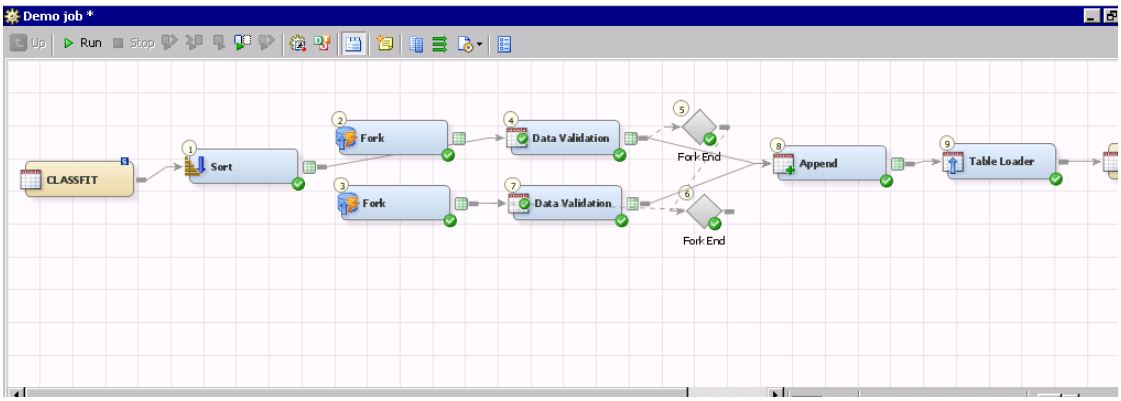

Figure 18: SAS Data Integration Studio Fork Transform Example

The command-line batch deployment tool has been updated as well. The tool enables users to batch deploy many jobs at once using a simple command-line interface. The user invokes an executable named "DeployJobs.exe" and supplies parameters to control its behavior. All options are specified as arguments to the "DeployJobs" executable now, so a manifest file is no longer required, which should simplify the use of the tool. The tool can also be used on platforms previously unsupported such as z/OS.

Below is an example of the syntax of the tool.

```
DeployJobs
connection-options
```

```
-deploytype DEPLOY | REDEPLOY
-objects source-location-1 source-location-2 …
-sourcedir
-deploymentdir
-metarepository
-metaserverid
-appservername
-servermachine
-serverport
-serverusername
-serverpassword
-batchserver
-folder
-log LOG PATH| LOG PATH AND FILENAME
-recursive
-since FROM ABSOLUTE DATE | FROM RELATIVE DATE
```

Here is a sample command-line batch deployment tool command.

```
DeployJobs –profile “My Profile” –deploytype deploy –objects “/Shared Data/My
Jobs/TransformJob” –sourcedir “c:\Source Data\Jobs” –deploymentdir
“C:\SAS\Config\Lev1\SASApp\SASEnvironment\SASCode\Jobs” –metarepository
Foundation –metaserverid A57CMFYM.AS000002 –servermachine “appserver machine
name” –serverport 8591 –serverusername “user-id” –serverpassword “password” –
batchserver “SASApp – SAS DATA Step Batch Server” –folder “Jobs/Deployed
Jobs”
```

This command does the following.

- Deploys the job TransformJob from the folder /Shared Data/My Jobs.
- Deployed job code files are written to c:\SAS\Config\Lev1\SASApp\SASEnvironment\SASCode\Jobs.
- Deployed job objects are created in the folder location Jobs/Deployed Jobs.

Several new data sources and targets have also been added. Support for Pi data, and Hawq SQL data sources and targets are now supported via two new SAS Access engines. Integration with the SAS LASR Analytic Server has also been enhanced to support the SASIOLA engine for loading tables into the SAS LASR Analytic Server.

DATAFLUX DATA MANAGEMENT STUDIO AND SERVER

DataFlux Data Management Studio and DataFlux Data Management Server (both part of all SAS Data Management offerings) have a number of new features available as well. These products have added support for integration with the SAS Metadata Server. This allows both Data Management Studio and Data Management Server to now support Integrated Windows Authentication and single sign on, as well as other authentication modes supported by the SAS Metadata Server. Shared users and groups are now also supported.

A migration tool has been developed to help migrate users and groups from a DataFlux Authentication Server to a SAS Metadata Server. To use the tool, run PROC ASExport on the DataFlux Authentication Server, and then run the dftool utility to handle the migration. The migration process and dftool run during installation. You can also choose to run migration and dftool after installation as needed to complete the migration of users and groups.

DataFlux Data Management Server can now be configured to manage HTTPS connections in a manner that complies with the Federal Information Processing Standard 140-2. FIPS compliance is required by various industries and government agencies.

DataFlux Data Management Server has added a new public REST API. The REST interfaces support access to most server features including job generation, execution, and status. Figure 19 is an example of the new server REST interface base URL.

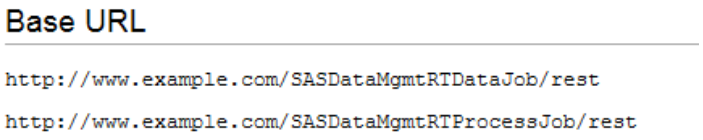

Figure 19: Example of the Data Management Server REST Interfaces

There are three categories for the public REST APIs for DataFlux Data Management Server.

- Batch Jobs: allows for management and execution of Data Management batch jobs
- Real-Time Data Jobs: allows for management and execution of Data Management real-time data jobs
- Real-Time Process Jobs: allows for management and execution of Data Management process jobs

SAS QUALITY KNOWLEDGE BASE

The SAS Quality Knowledge Base (QKB) is a collection of files that store data and logic that define data quality operations such as parsing, standardization, and matching across multiple languages and locales. SAS software products use the QKB when performing data quality operations, including the data quality operations that are available in Hadoop. A copy of the QKB can be deployed in Hadoop as part of the SAS Data Loader for Hadoop offering across all the data nodes in your Hadoop deployment.

SAS delivers many out of the box QKB definitions for use with common types of data such as names, addresses, and phone numbers. One of the more powerful features of the data quality features that SAS supports however is the ability to customize these QKB definitions to meet your organizations needs. For example, you can create definitions to help you standardize addresses in your company preferred format, or handle your own custom product codes. You can use DataFlux Data Management Studio to customize the QKB by modifying definitions or creating new definitions for use with your own business data. Since a QKB is used by all SAS products, QKB customizations are automatically available to your entire enterprise.

A number of enhancements to the DataFlux Data Management Studio Customize component have been made to assist in customizing a SAS Quality Knowledge Base. You can now copy and paste expressions, import word, category, and likelihood values from external files, update grammar, leverage string and substring values found in phrases, update diacritics and punctuation, and string together many of these features in order to handle complex string manipulating, matching, and parsing.

SAS LINEAGE AND IMPACT ANALYSIS

A number of important new features have been added to support lineage and impact analysis. 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. A relationships web viewer called SAS Lineage supports different views for displaying information stored in the service. Figure 20 is an example of the Impact Data Flow view. There are also views for all Relationships, and for Data Governance.

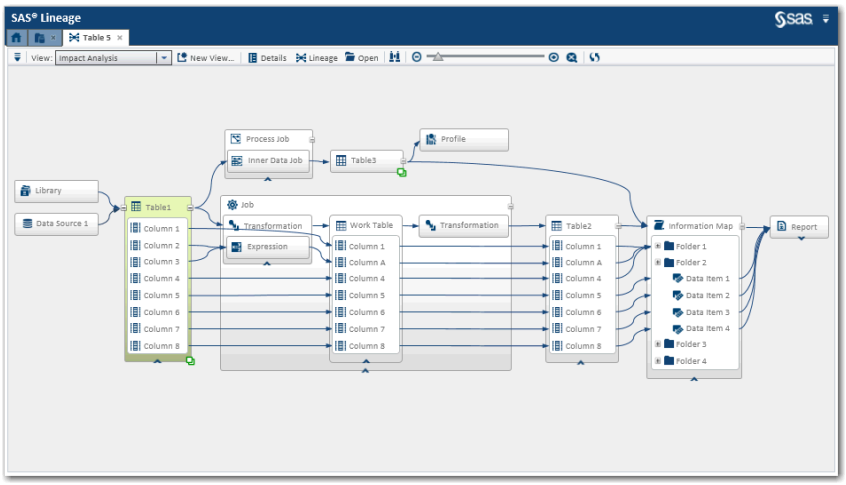

Figure 20: Lineage Viewer Showing Table, Job, and Column Relationships

There are a number of enhancements to SAS Lineage and the underlying relationship service that supports the lineage content. A key enhancement is the ability to import content from third-party metadata sources using the MetaIntegration bridge technology. The import exchange is available for hundreds of third-party sources including vendor sources such as SAP Business Objects, ERwin modeler, and many other tools. In past releases, import was limited to relational types of metadata, but this restriction has now been lifted. The content types are unlimited, in that all object types from all models can be imported.

Metadata exchange with third-party sources is available via a new command line utility that comes with a SAS installation. Below is an example of the launcher program and its install location.

```
Launcher name: sas-metabridge-relationship-loader
Install location: !SASHOME\SASMetadataBridges\4.1\tools
```

The user supplies login information to the relationship service, and an administrative user ID and password to perform the import. Other options available during the import include the ability to mark 1-N objects as equivalent to each other, so that the viewer shows the object as a single group object instead of separate objects; the ability to specify vendor options when using a particular vendor bridge; and the ability to schedule the import to occur on some predetermined schedule, for better support of synchronizing content. Below is a partial list of the options available in the utility.

```
usage: sas-metabridge-relationship-loader [options...]

Example options:
 -?,--help              Print help information.
 -bridgeDirectory       The location of the SAS Metadata Bridges if different
than default location.
 -bridgeList            Request the list of available licensed bridges.
 -bridgeOptions         Customize the import
 -clean                 Clean relationships from the third party source.
 -loadRelationships     Load relationships from a third party source

…and others
```

Figure 21 and Figure 22 are some examples of imported content from external metadata sources using the bridges.

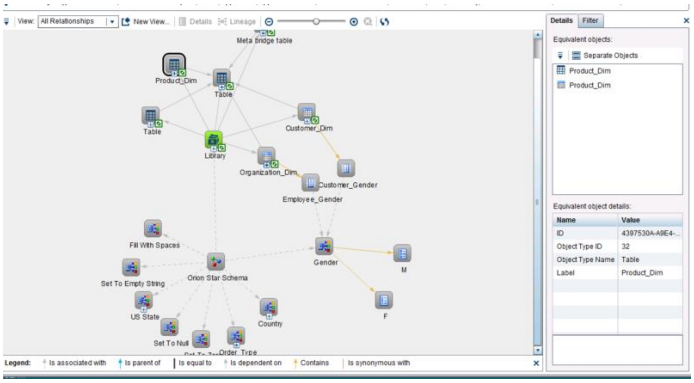

Figure 21: Example of Content Imported from an External Metadata Source

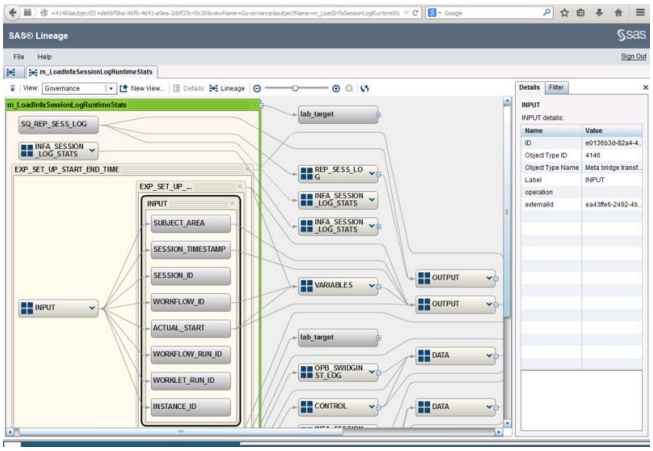

Figure 22: Example of the Governance View from an Import

SAS FEDERATION SERVER

SAS Federation Server is the SAS offering that supports 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 a conceptual diagram of data federation.

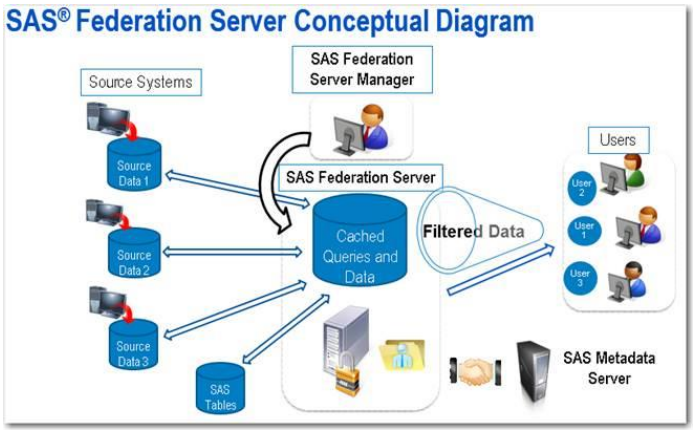

Figure 23: Federation Server Conceptual Diagram

Typically, a data federation methodology is used when traditional data integration techniques cannot meet the data needs, such as when the data is too large, or too proprietary, or too mission critical 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 views that can be delivered to downstream processes. Federation allows the data to be extracted and stored in a persistent cache, which can be updated periodically or scheduled to be refreshed during non-mission critical times. Federation is also a good choice for systems where data is diversified across source systems. Managing security, user IDs, authorizations, and so on, on all of the various source systems can be a huge burden for a 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 diverse system security environments, and for updating views when source systems change.

SAS Federation Server 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, integrated data quality functions (using the SAS Quality Knowledge Base), a number of data source native engines for database access, full support for SAS data sets, auditing and monitoring capabilities, many security features including table, column, and row-level security, and a number of other federation features. Access to SAS Federation Server is available via a number of available interfaces including a REST API, JDBC, ODBC, and via a SAS Access engine.

Figure 24Figure 24 is a high-level overview of the SAS Federation Server.

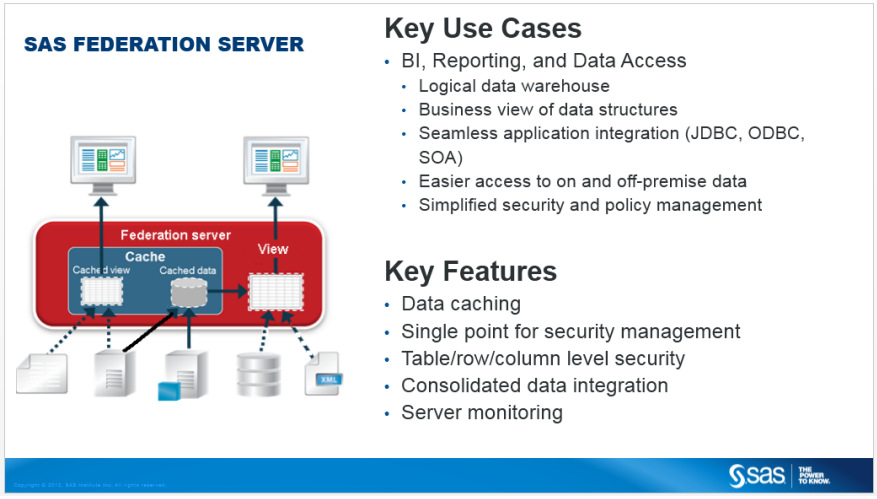

Figure 24: SAS Data Federation Server Overview

SAS Federation Server has been updated in the latest release to integrate fully with SAS Enterprise Architecture that includes the SAS Metadata Server and the SAS Web Infrastructure Platform. These features enable SAS Federation Server to support enterprise class features such as Integrated Windows Authentication, common authentication and authorization features including shared user management, consistency in installation, configuration, and administrative tasks, and standard SAS web infrastructure capabilities such as custom theming support.

SAS Federation Server includes support for data masking, and the support has been enhanced in the latest release. Data masking is a method of hiding sensitive data, or personally identifiable information (PII), within data sources. The purpose of data masking is to protect the original data by using a functional substitute in situations where the audience is not privileged to access the original data. You can use data masking to protect sensitive data while maintaining integrity of the data so that it is still usable in your applications.

Enhancements made to the SAS Federation Server for data masking include some of following functions.

- ENCRYPT and DECRYPT – mask or unmask the values in a column using symmetric key encryption

- HASH – hash the value
- TRANC (Transliterated Value) - mask values transliterating characters
- RANDOM - mask numeric values by generating a uniformly distributed pseudo-random number
- RANDATE (Random Date) - mask values in a date column by replacing them with pseudo-random date values.
- RANSTR, RANDIG – mask strings or digits with randomly generated values

Below is an example of using the ENCRYPT function in the SAS Federation Server to mask a NAME field:

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

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

There are a number of available new source database types, including Apache Hadoop Hive data, Postgres SQL, SAP HANA, and SASHDAT format for writing data files to a SAS Analytics Server.

Another key feature is the ability to persist content in memory. This can be very useful if there are critical resources that need to be available on demand. You can persist views, tables, and data caches in the in-memory storage facility and cycle it out of memory when it is no longer needed.

SAS Federation Servers can also be chained, so that one federation server can be used as a source, or target of another federation server. This can be useful if for example, you need to synchronize between different sites such as headquarters and regional offices. You can use data federation to help manage resources between the various locations using features that the federation server provides such as data caching, security, and fast table access using the in-memory store.

SAS Federation Server also supports the ability to run SAS DS2 programs and SAS Scoring programs on the data flowing through the federation server to support a variety of useful functions such as data cleansing, data consolidation, and ETL features such as joins, updates, and queries. Figure 26Figure 26 is an example of a federation server DS2 program.

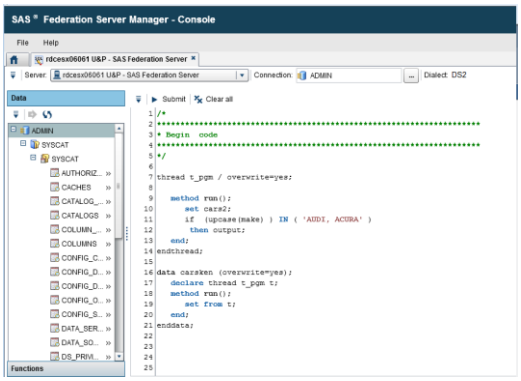

Figure 26: Example Federation Server DS2 Program

CONCLUSION

The latest releases of SAS® Data Integration Studio, DataFlux® Data Management Studio, and other SAS data management products provide new and enhanced features to help 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 latest releases include a number of new features in support of big data, management, governance, and federation.

RECOMMENDED READING

- SAS® Data Management Discussion Forum, Available at https://communities.sas.com/t5/SAS-Data-Management/bd-p/data_management
- Hazejager, W. Rausch, N. 2016, Ten Tips to Unlock the Power of Hadoop with SAS®, Available http://support.sas.com/resources/papers/proceedings16/SAS2560-2016.pdf.
- Ghazaleh, David. 2016. “Exploring SAS® Embedded Process Technologies on Hadoop®.” Proceedings of the SAS Global Forum 2016 Conference. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings16/SAS5060-2016.pdf.
- Ray, Robert. Eason, William. 2016. “Data Analysis with User-Written DS2 Packages.” Proceedings of the SAS Global Forum 2016 Conference. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings16/SAS6462-2016.pdf.
- Rausch, N. 2015. “What’s New in SAS Data Management.” *Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings15/SAS1390-2015.pdf.
- Rineer, B., 2015 “Garbage In, Gourmet Out: How to Leverage the Power of the SAS® Quality Knowledge Base”, *Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings15/SAS1852-2015.pdf.
- Agresta, R., 2015 “Master Data and Command Results: Combine MDM with SAS Analytics for Improved Insights”, *Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings15/SAS1822-2015.pdf.

CONTACT INFORMATION

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

Nancy Rausch
100 SAS Campus Drive
Cary, NC 27513
SAS Institute Inc.
Nancy.Rausch@sas.com

http://www.sas.com

Ron Agresta
100 SAS Campus Drive
Cary, NC 27513
SAS Institute Inc.
Ron.Agresta@sas.com
http://www.sas.com