Paper 110-2012

What's New in SAS® Data Management

Nancy Rausch, Michael Ames, Wilbram Hazejager, SAS Institute, Cary, NC, USA

ABSTRACT

The latest releases of SAS® Data Integration Studio and DataFlux® Data Management Platform 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. Latest features include capabilities for building complex control processes, additional in-database ELT transformation capabilities, big data capabilities, enhanced features for monitoring data and processes, and new features for unstructured data access, 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 their combined capabilities.

INTRODUCTION

The data management lifecycle is a best practice approach for managing the flow of data in your enterprise. There are three main phases of the data management lifecycle: plan, act, and monitor. The three phases and the actions which typically take place in these phases, are illustrated in Figure 1 below. In the plan phase, you determine what data is accessible and where it is located. In the action phase, you design processes that transform the data into a form that is usable for analysis and reporting. During the monitor phase you monitor your job processes so that you can react to changes in your enterprise when needed. In this way you ensure that your data is managed to the highest efficiency and quality. SAS® Data Integration Studio and DataFlux® Data Management Platform comprise the SAS Data Management suite of products. These products include many new features that enable you to better manage your data across this lifecycle.

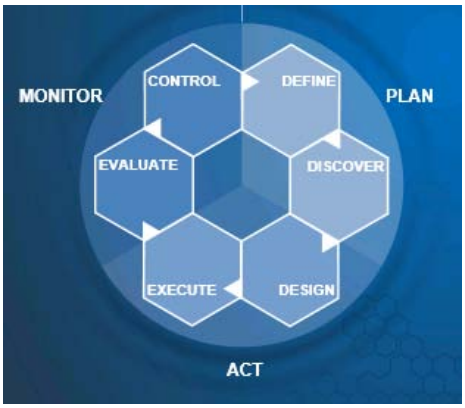

Figure 1: The Three Phases of the Data Management Lifecycle and the Tasks Found in Each Phase

DATA DISCOVERY

The first phase of data management involves data discovery and defining the data in your system. There are a number of new features in the SAS Data Management products that can help you in this phase.

BUSINESS DATA NETWORK

The Business Data Network supports 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 contains a web user interface that documents business terms and their associated rules, jobs, applications, data, documentation, and other information. Technical users can use the network to collaborate on rules used to validate data, and share knowledge about data transformations. Data stewards can view the data from a business standpoint and visualize problem areas by domain so as to identify and fix issues more effectively. Figure 2 shows the Business Data Network main page.

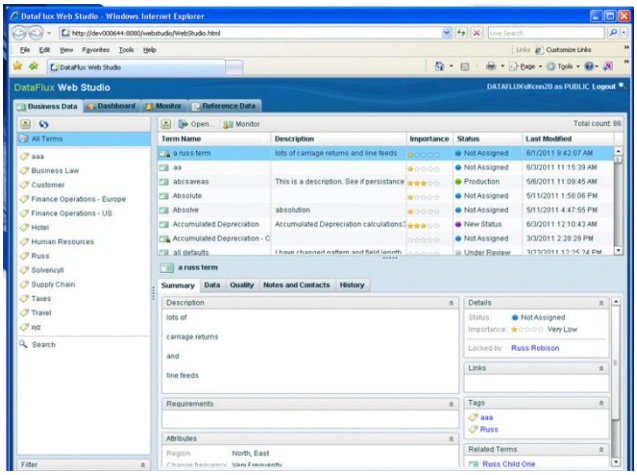

Figure 2: Business Data Network

Typically a user that understands their business terminology would provide the initial information in the Business Data Network. This user would also attach documents or rules that describe each term. 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 contains diagrams that allow you to understand how your physical data and business processes interrelate. Figure 3 shows a typical diagram of relationships stored in the network.

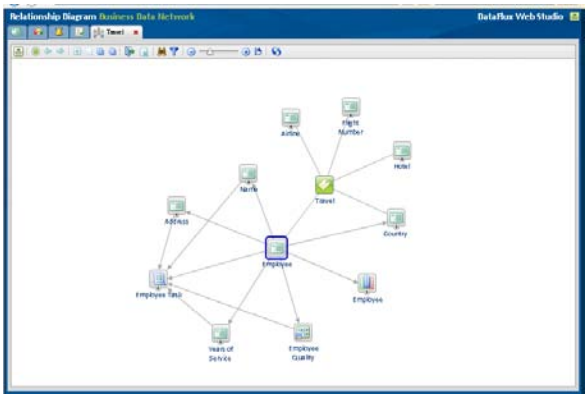

Figure 3: Business Data Network Relationship Diagram

DATA VISUALIZATION

SAS Data Management also provides views that allow you to see the structure of your data tables and how they interrelate. The data model view illustrated in Figure 4 shows the data model viewer in Data Management Studio. It visualizes table structures, primary/foreign key relationships, data types, and table indexes.

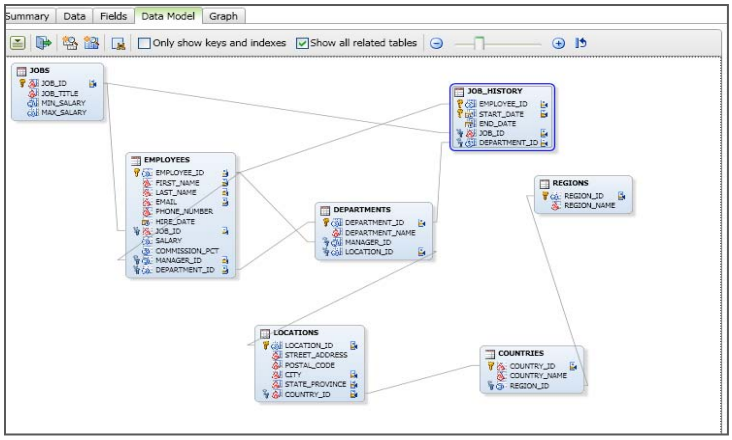

Figure 4: Data Model View

There are also some useful statistics that you can use to help you better understand the data contained in your tables. Data Management Studio includes some simple graphs such as those shown in Figure 5 below, that show how the data is distributed in your table, as well as showing data lengths, mean, and median values. You can also plot one column by another and view sum or count statistics. You can view this information via a simple point and click interface.

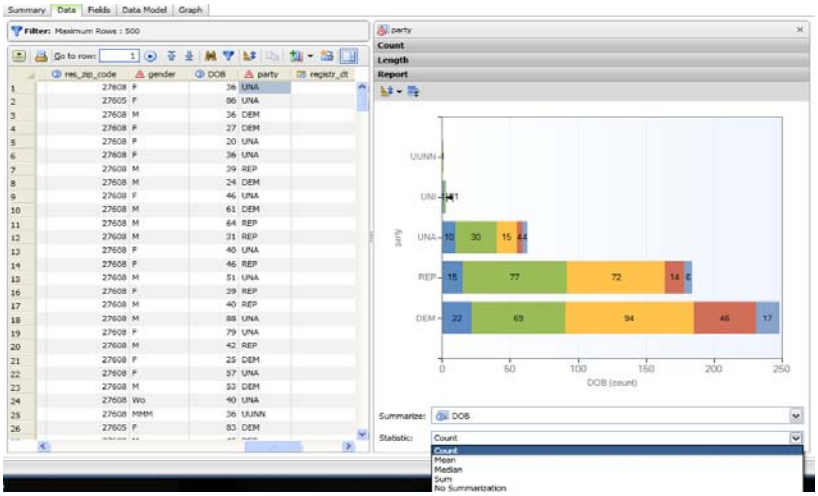

Figure 5: Data Report Statistics View

To gain further understanding of your data, you can leverage the data profiling capabilities provided in SAS Data Management. Data profiling provides reports on frequency counts, value ranges, pattern types, and more. Profiling can be done on an ad-hoc basis, or, if you need to profile large tables on a periodic basis, a job node is available as shown in Figure 6. Using the node in your jobs ensures that your profile reports are always up to date.

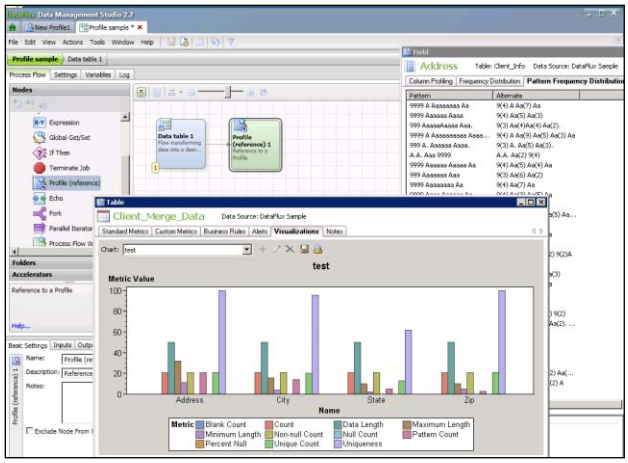

Figure 6: Profile Job Node and Profile Reports

MASTER DATA FOUNDATIONS

Frequently there is a need when working with certain kinds of data, such as customer records, supplier information, or other information that has a high chance of having duplicate related information, to select the best record out of all of the possible records before including the data in transformation logic. Figure 7 is an example showing this type of data.

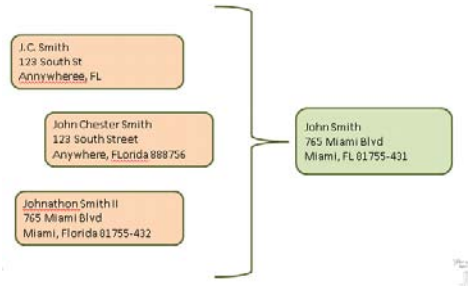

Figure 7: Example Data Cluster

Master Data Foundations is a feature available in the SAS data management software to assist you in making this selection across thousands of records. It accomplishes this through a technique called "clustering", which is difficult to do with traditional SQL transformation logic .The technology includes support for probabilistic matching. That is, if two records are similar to each other the technology can create a score based on configurable rules as to how likely or how probable the records match. Figure 8 is an example of the cluster viewer which allows you to see the cluster of records that matched, and change or modify them as needed. The best record is automatically selected for you into a single, cleansed and de-duplicated data source.

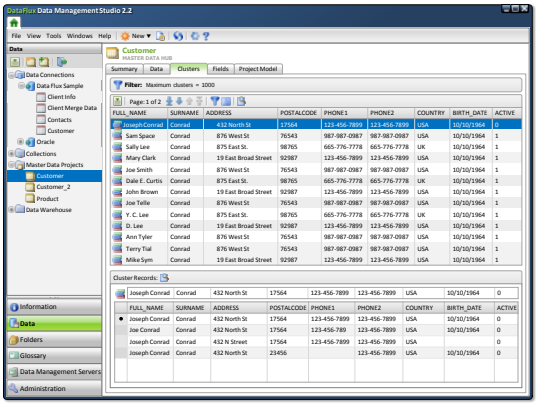

Figure 8: Cluster Viewer

DATA FEDERATION

When data needs to be consolidated from many different formats, types and databases, as illustrated in Figure 9, it becomes very inefficient to try to move the data around so that it can be appropriately joined. The SAS/DataFlux Federation Server was developed to help efficiently solve this problem. The server supports parallel, threaded, in-database processing. A data cache layer is also available. In the data cache, data can be retrieved from different data sources and persisted directly in the database as a cached view. An integrated scheduler supports periodic cache refreshes.

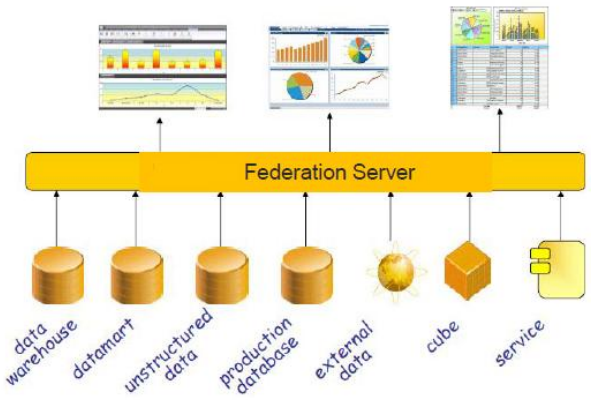

Figure 9: Data Federation Example

The federation server has a web administrative console for server monitoring and management. Figure 10 is an example of the console.

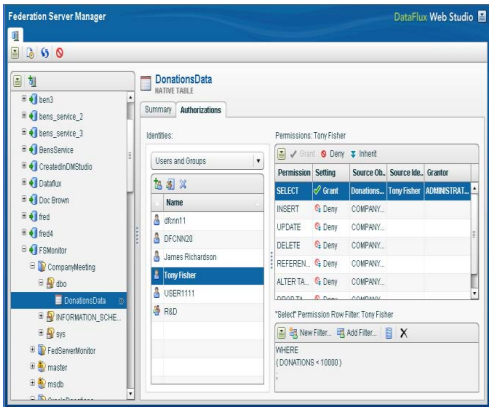

Figure 10: Federation Server Administrative Console

The federation server also supports column and row level security. This allows an administrator to configure views so that different users gain appropriate access to the data based on their permissions.

TRANSFORMING DATA

Once data has been cleansed and de-duplicated, it is ready to be transformed into structures appropriate for downstream reporting and analysis. There are a number of new features available in SAS Data Management for optimizing the transformation process. Data nodes have been added for in-database processing for optimal performance.

The following SQL nodes have been added for in-database processing:

- SQL Merge which supports updating existing records and inserting new records
- SQL Delete with an optional select statement
- SQL Update with an optional select statement
- Simplified interfaces for Create Table as Select, Insert Table with Select, and Execute SQL syntax with

example templates

- DB2 Optimized Table Loader

SQL MERGE

The SQL Merge transform supports inserting new rows and updating existing rows using the SQL merge DML command which was introduced in the SQL 2008 standard. This node generates the DBMS specific passthru syntax shown below:

```
MERGE INTO table_name
    USING (table_reference or query)
    ON (condition)
    WHEN MATCHED THEN
          UPDATE SET column1 = value1 [, column2 = value2 ...]
    WHEN NOT MATCHED THEN
          INSERT (column1 [, column2 ...]) VALUES (value1 [, value2 ...])
    LOG ERRORS INTO <table> (values) REJECT LIMIT 10;
```

Merge supports updating matching records and inserting new records when no match is found. Match can consist of a simple match value or a complex select and subquery to generate the match criteria. Merge is illustrated in Figure 11.

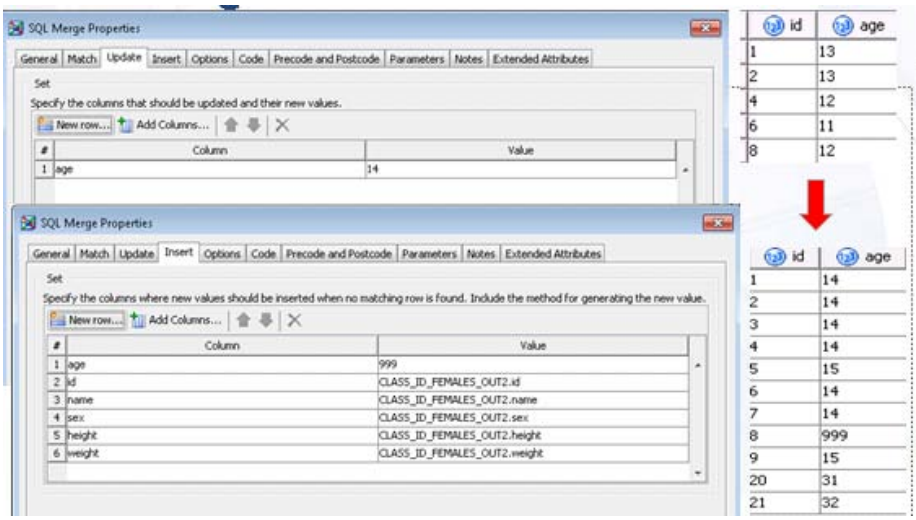

Figure 11: SQL Merge Transform Example

SQL DELETE

The SQL Delete transform supports deleting records in a database table that match the specified where clause. The where clause can be a simple value or a complex select and sub-query to generate the match criteria. A SQL delete example is illustrated in Figure 12.

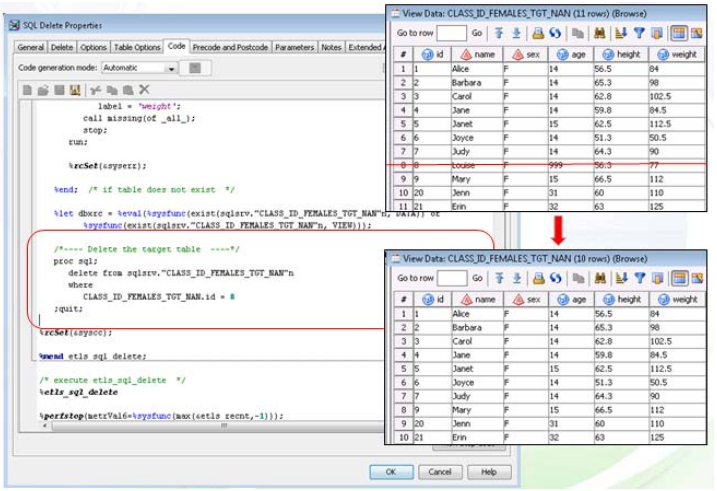

Figure 12: SQL Delete Transform Example

SQL UPDATE

The SQL Update transform updates rows in a table that match the specified where clause. The where clause can be a simple value or a complex select and sub-query. The syntax for a SQL Update node is shown below:

```
proc sql;
UPDATE Table
SET  <columns>
WHERE <complex clause>
quit;
```

A SQL Update example is illustrated in Figure 13.

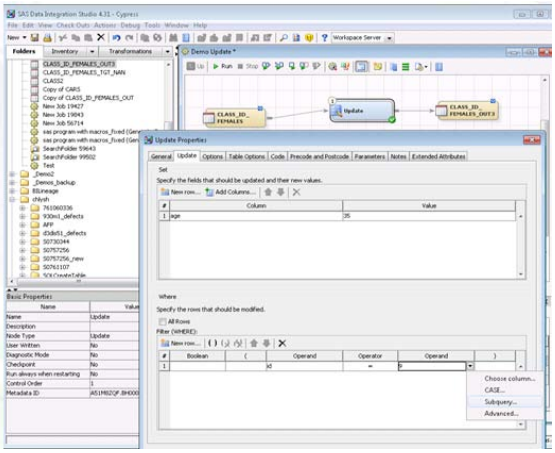

Figure 13: SQL Update Transform Example

NEW SQL INTERFACES

In addition to new functionality, some transforms have been updated to include simplified interfaces that suit a variety of user scenarios, and SQL templates have been added to help users get started. The SQL Create Table as Select and Insert Table as Select transforms have a simplified query designer interface. A user written SQL Execute transform is also available with added templates to help you get started. The SQL Execute transform and templates are illustrated in Figure 14.

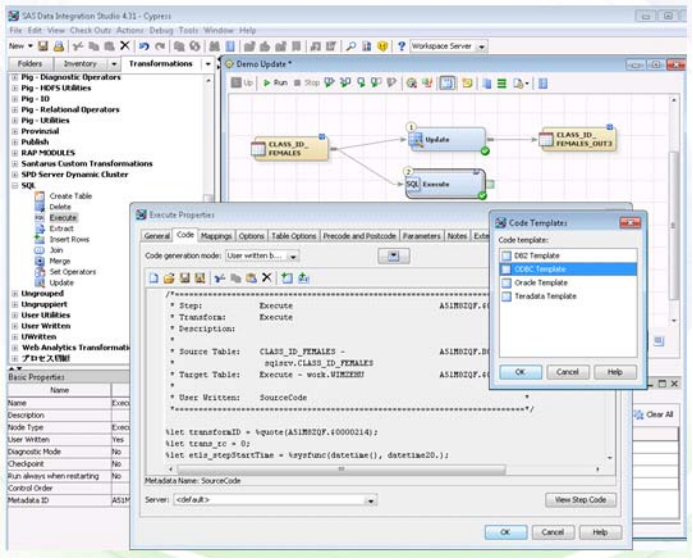

Figure 14: SQL Execute Transform Example

BUSINESS RULES

Business rules de-couple business logic from application logic. Business rules are essentially a set of logic that can be stored independently of data, and reused across multiple data environments to apply the business logic to the application data. SAS Data Management offers a new set of products to enable you to author business rules and then apply them in your jobs to standardize and cleanse your data. The components of the new Business Rules product suite include a web-based rule authoring user interface for development, simulation, management and monitoring of your rules; a rules engine for supporting batch, real-time, and service execution modes; support for integrating mining models for scoring; and a transformation in SAS Data Integration Studio for deploying rule packages into your jobs.

Figure 15 is an example of the authoring user interface, SAS Rules Studio. Rules Studio allows you to author, validate and test your rules, and then deploy them as packages into the SAS Metadata server. Rules packages are fully versioned so that you can work on future versions without affecting your production jobs.

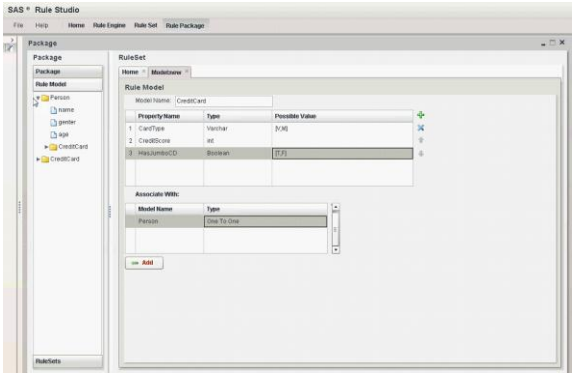

Figure 15: SAS Rules Studio

Figure 16 shows an example of using rules in SAS Data Integration Studio. The rules are published and appear in the SAS Metadata Folders tree. SAS Data Integration Studio introduces a new Business Rules node that understands rules packages. The node allows you to map your source data, and output data into and out of the rules package. The job will then apply the rules to your data when it is run.

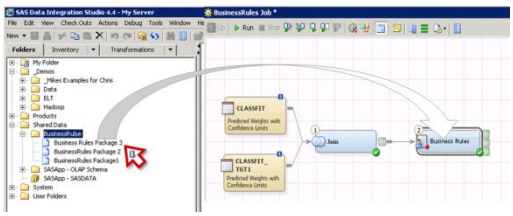

Figure 16: Rules Studio Node in Data Integration Studio

When a job is run that contains a rules package, statistics are collected about how many rules fired, invalid and valid data record values, and other details. You can use this information to further refine your data as it flows through your transformation logic.

DB2® BULK LOADER

An optimized bulk loader transformation that supports the DB2 database has been added. The loader supports multiple DB2 load techniques including import, CLILoad, and truncate. Table statistic generation is optionally available, and many performance tuning options are surfaced in the user interface.

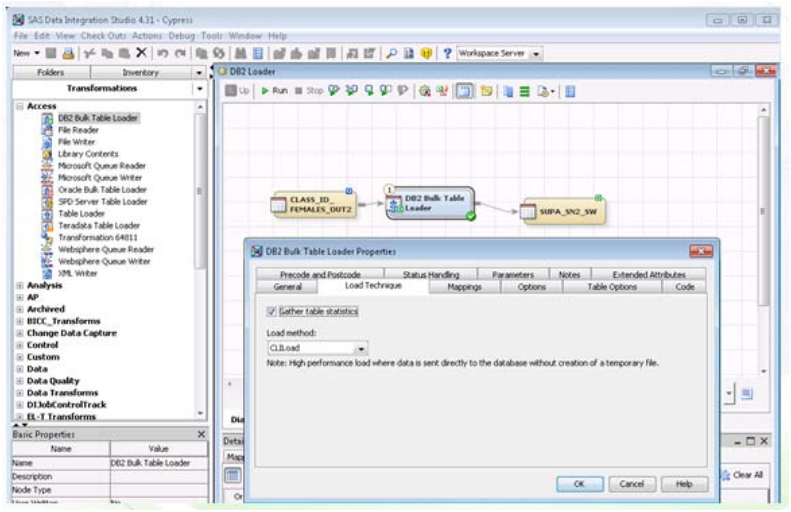

Figure 17: DB2 Bulk Loader Transformation Example

HADOOP

Apache™ Hadoop™is an open source technology for large data volume storage and processing. Its scalability comes from the marriage of its high bandwidth, clustered storage, called the HDFS, and its fault-tolerant distributed processing algorithm, called MapReduce. SAS Data Management provides integration with Hadoop in three ways: reading and writing data to and from Hadoop HDFS, data processing for sending programs and managing execution of programs in Hadoop systems, and a data transformation library for writing Hadoop programs in Hadoop languages including Pig, Hive, and MapReduce.

The basic architecture of the SAS Hadoop integration is illustrated in Figure 18.

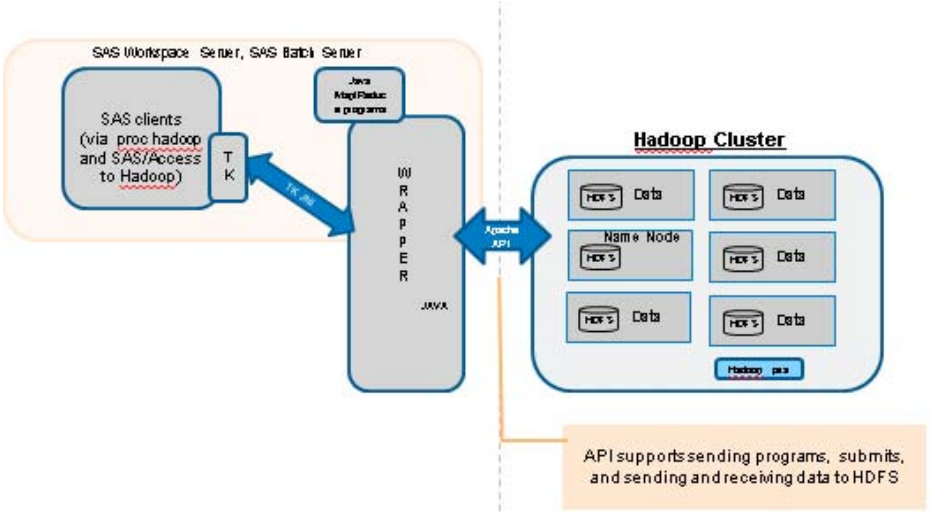

Figure 18: SAS Hadoop Architecture

SAS sends data to and from the Hadoop HDFS using either a newly available SAS file access method, or via the new SAS/Access engine to Hadoop. Once data is in the Hadoop file system a number of new transforms are available in SAS Data Management that you can use to write programs and submit them to the Hadoop system. Figure 19 shows some of the available transformations.

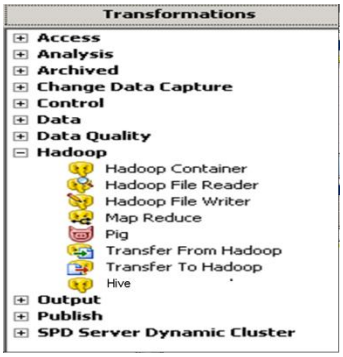

Figure 19: SAS Data Integration Hadoop Transforms

Figure 20 shows an example program written in the Hadoop PIG language. The transform has an enhanced, color coded editor specific to the language to make it easier for you to write programs for using the various languages of Hadoop.

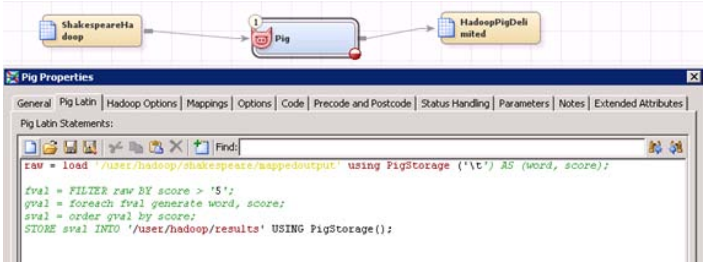

Figure 20: Example Hadoop Transform for the PIG language

All programs submitted from SAS use a new SAS procedure developed to interact with Hadoop called Proc Hadoop. Figure 21 shows an example of the new Proc syntax. The Proc supports submitting and managing programs running in the Hadoop system.

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

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

Figure 21: Example Syntax of Proc Hadoop

Once jobs have been submitted to Hadoop, you can monitor them from SAS Data Integration Studio as illustrated in Figure 22. You can monitor the status of your jobs, data loads, file system, and cluster usage.

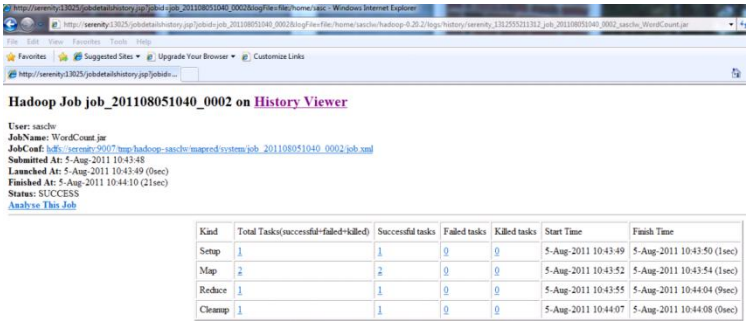

Figure 22: Example of the Hadoop Job Monitoring Features

ADDITIONAL TRANSFORMATION FEATURES

Other enhancements to the transformation capabilities of the SAS Data Management platform include nodes to support provisioning data for the SAS High Performance Analytics platform. The Data Validation transform has had a number of updates based on customer requested features such as error and exception work tables and improved performance using hashing as an optional technique for comparison matching. XML integration has been enhanced with new XML nodes that can support reading columns of XML data coming from a database and transforming it into tabular structures using XML maps or XSL translations. These nodes are illustrated in Figure 23.

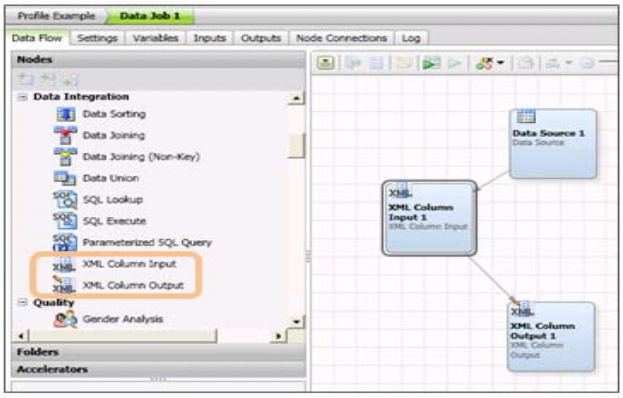

Figure 23: XML Column Nodes for Transforming Database XML Columns

There are new unstructured data nodes that can read a variety of text formats such as pdf, txt, doc, etc. and convert them into strings that can then be parsed into tokens using user definable dictionaries. A new process FORK node has been added to support parallel processing of differing job flows. Finally the slowly changing dimensions transform has been updated to support SPDS optimized loading, and additional performance enhancements.

DATA MONITORING

Once your jobs are in place to transform data, you need to monitor the results, watch for potential errors or modifications to the incoming data that you may need to react to, and manage the overall quality and performance of your system. Figure 24 shows the Data Quality dashboard available in SAS Data Management.

Figure 24: Data Quality Dashboard

The dashboard is a web client that visualizes quality problems and how they are trending over time. You can drill in to get more detail about specific dimensions such as Accuracy, Integrity, and others. Dimensions and thresholds are fully user-configurable. Trending is also visualized so you can see how your data is performing over time. This enables you to better react to potential errors and fix problems more quickly. You can also drill in to so see specific problem records.

CONCLUSION

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

RECOMMENDED READING

- SAS® Enterprise Data Management and Integration Discussion Forum, Available at http://communities.sas.com/community/sas_enterprise_data_management_integration
- Rausch, Nancy and Stearn, Tim, "What's New in SAS® Data integration", Proceedings of the SAS Global Forum 2011 Conference, Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings11/134-2011.pdf.
- Rausch, Nancy and Stearn, Tim, "Best Practices in Data Integration: Advanced Data Management", Proceedings of the SAS Global Forum 2011 Conference, Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings11/137-2011.pdf.
- Ames, Michael and Steve Sparano, " On the Horizon: Streaming Integration and Analytics", Proceedings of the SAS Global Forum 2011 Conference, Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings11/404-2011.pdf
- Hazejager, Wilbram and Pat Herbert, "Innovations in Data Management – Introduction to Data Management Platform", Proceedings of the SAS Global Forum 2011 Conference, Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings11/141-2011.pdf.
- Hazejager, Wilbram and Pat Herbert, "Master Data Management, the Third Leg of the Data Management Stool: a.k.a. the DataFlux® qMDM Solution", Proceedings of the SAS Global Forum 2011 Conference, Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings11/146-2011.pdf.

CONTACT INFORMATION

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

Nancy Rausch

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

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

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