Paper 134-2011

What's New in SAS® Data Integration

Nancy Rausch and Tim Stearn, SAS Institute, Cary, NC USA

ABSTRACT

SAS® Data Integration Studio 4.3 provides 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. A major focus of the release is to deliver new features to help bring code into the managed environment using the SAS® Code Analyzer, provide performance optimizations through DBMS pushdown, and support Enterprise-level development with integrated versioning and rollback support. The introduction of the complimentary SAS® Data Management product helps leverage the capabilities provided in SAS Data Integration Studio and introduces new capabilities around data visualization, job management, parallelization, and database integration. Customers will find many reasons to upgrade to the latest version of SAS Data Integration Studio.

INTRODUCTION

The latest release of SAS Data Integration Studio and the complimentary DataFlux Data Management Platform introduce many new features to enhance your efficiency in managing your jobs, source code, runtime environment, data, and user workflow. New features include integrated versioning, enhancements to the SAS code importer, new transforms, enhanced grid support, and new features around job parallelization and management. These features enable you to build more efficient job processes and better manage your system environment.

VERSIONING

One of the major user workflow enhancements in this release is integrated versioning support. This feature enables you to archive content to a 3rd party versioning system of your choice. SAS package files can be archived, and integration includes differencing capabilities, the ability to rollback to previous versions of objects, and the ability to inspect which objects are contained within a specific version.

Versioning works by moving content such as jobs and other objects into a file and archiving that file in a versioning system. SAS Data Integration Studio creates the file and writes it into the Source Management System. To bring content back into the repository, SAS Data Integration Studio retrieves the content stored in the source management system and places it back into the SAS® Metadata Repository . In this way users can create different versions of content and restore previous versions of content when needed.

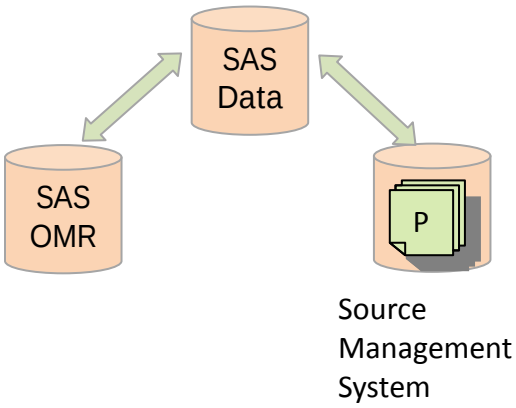

Figure 1: Overview of the Versioning Architecture

Objects can be versioned independently or with other objects to make up a package of related content. This allows you to archive sets of objects that are logically related, such as all of the content in a project. You can optionally also choose to generate source code for a job and store it along with the job as text content. This makes it easy to see the source code associated with a specific version of a job. You can view archive results of any object to see when it was last versioned. This lets you identify previous version of objects that you may want to restore and maintain a history about changes.

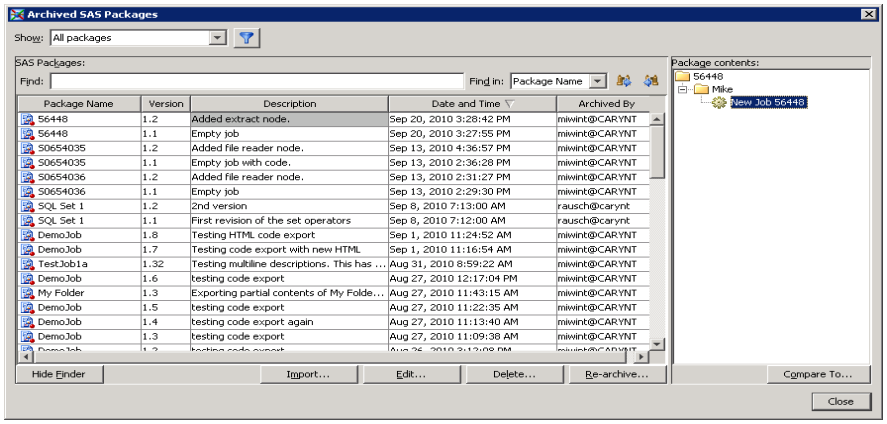

Figure 2: Dialog Showing Different Versions of Objects

There is also a differencing feature. You can select an object and view the differences between versions of the selected object, or between an archived version and the current version of that object.

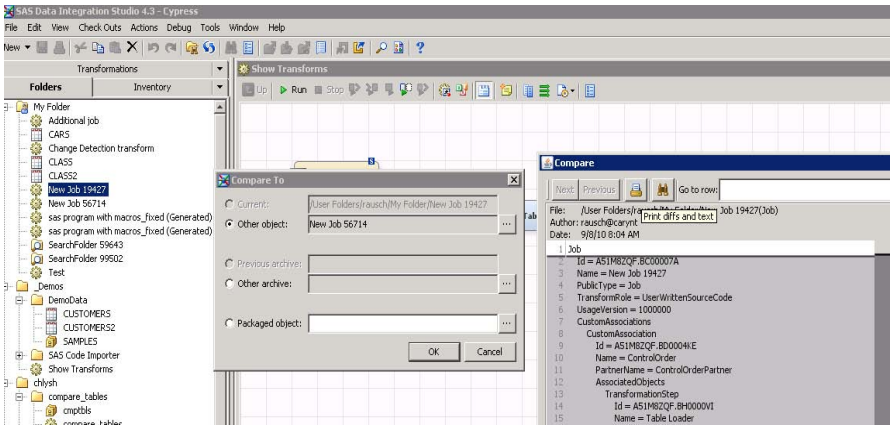

Figure 3: Differencing Features Available when versioning objects

Several out-of-the-box components are provided for integrating with two open-source versioning systems, CVS, and SubVersion. In addition, source code has been provided as an example if a different source management system is preferred. There is a documented application programming interface (API) to integrate different source management systems with SAS Data Integration Studio.

JOB MANAGEMENT REPORTS

Another key manageability feature in the new release of SAS Data Integration Studio is enhanced job performance and status reporting. There are pre-built reports for SAS® Web Report Studio that you can use to provide performance information, current and historical, and job status. You can also optionally use the SAS® Stored Process Server to generate HTML reports with similar information.

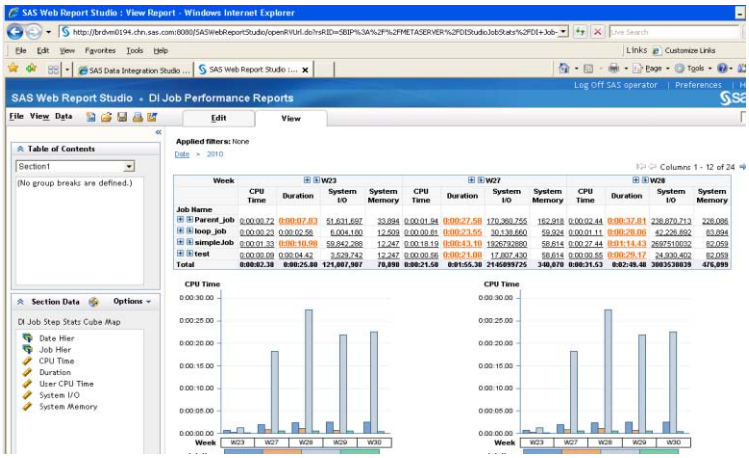

Figure 4: Example Job Management report

SAS CODE IMPORTER

There are a number of new features in the SAS Code Importer. There is an enhancements to optionally expand out SAS macros in your jobs, and create a node for each step inside of your macros. In the following example, the left figure shows what the imported job looks like without macro expansion enabled. The right figure shows the same job imported with macro expansion. As illustrated in the right figure, expanding out the macros provides additional detail about your job and how it works. When you run your job with the macros expansion option enabled you can get more performance information such as slow running steps, which steps use more memory or I/O, and CPU performance details.

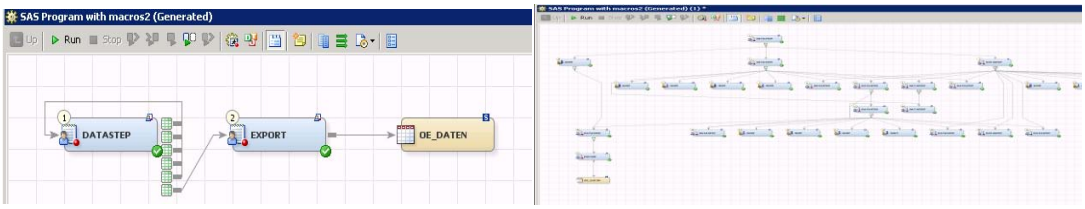

Figure 5: Examples of Output from the SAS Code Importer

Another option allows you to register all work tables as physical tables in a WORK library. This allows you to import SAS code that uses temporary tables that are both the source and target of a step. You can also analyze your job to determine the type and number of steps in your job. This information is provided in a report that you can review prior to importing the job.

INTEGRATED SEARCH AND FIND

There are now integrated search and find capabilities. You can search for objects by name including the ability to search for patterns. You can subset a search to a specific folder, search by type, by last change date, or by other user defined criteria. You can also save searches to a folder and bring them up later when needed. For example, you can use the saved search feature to maintain a "recently changed" object list

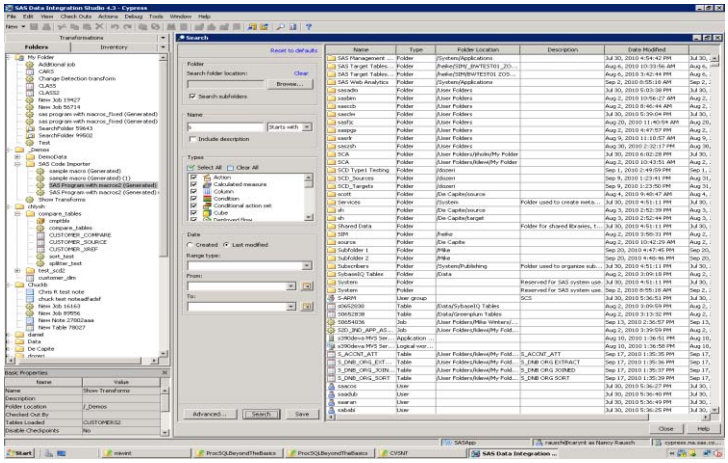

Figure 6: Find and Search Window

BATCH JOB DEPLOYMENT

Another feature that will helps with system management is the new ability to deploy jobs from a command line. An example batch file is installed with SAS Data Integration Studio that shows you how to use this new feature. You can use this feature to deploy any number of jobs without having to bring up the Data Integration Studio application.

ENHANCED SUPPORT FOR z/OS

There are additional enhancements for the z/OS operating system as well. Code generation line lengths can be limited to 80 characters or less, and deployed jobs JCL can also be restricted to fit within the z/OS 80 character line length limit. Lines that go beyond the 80 character limit will flow over onto the next line.

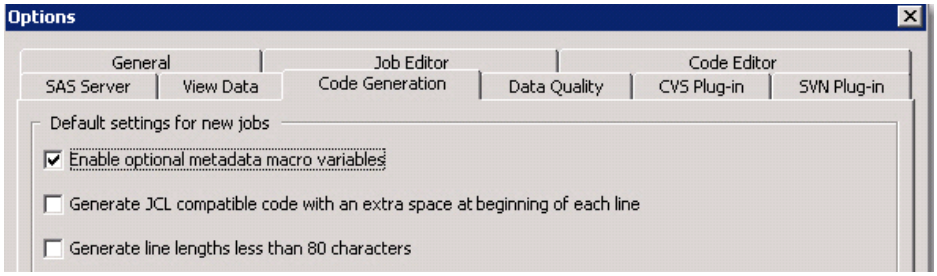

Figure 7: Option to Support the z/OS Operating System During Code Generation

COLUMN STANDARDIZATION

A new column standardization wizard is available to help you update table column metadata between tables so that they match. You can use this wizard to standardize column lengths between two or more tables, formats, and other attributes that you would like to match between the tables.

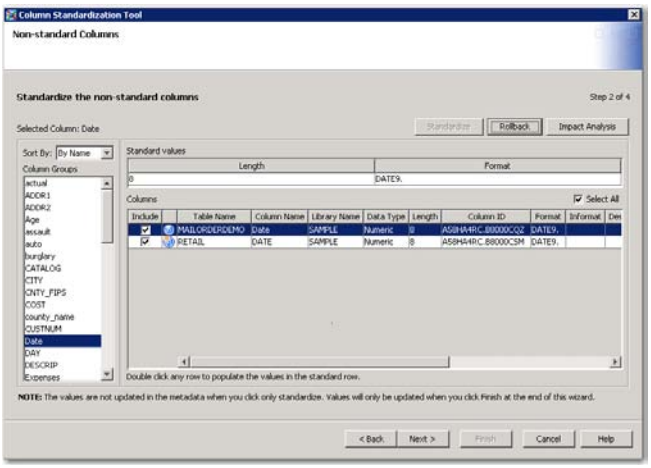

Figure 8: Column Standardization Dialog Example

You can also use this feature to generate a report about column differences, or log updates for audit purposes.

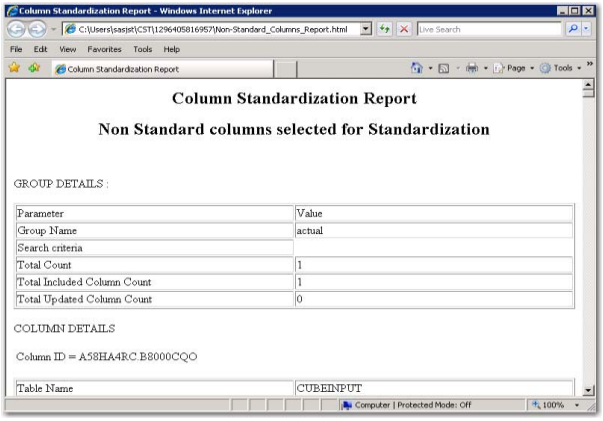

Figure 9: Column Standardization Example Report

USER DEFINED FORMATS AND ANALYTIC FORMAT INTEGRATION

The latest release of Data Integration Studio includes the ability to discover and register user defined formats and deployed analytic scoring functions discovered from a relational database. The formats and functions can be discovered and then registered so that they appear as expressions available from the Data Integration Studio expression builder.

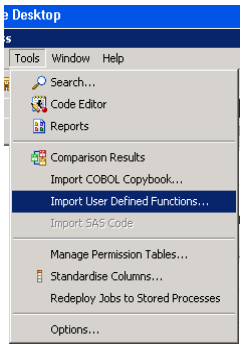

Figure 10: Import User Defined Functions and Formats is Available From the Tools Menu

The discovered functions and formats can be placed in a folder location of your choice. Parameters on the function are also discovered and registered so that you will have enough detailed information to be able to use the function or format in your jobs.

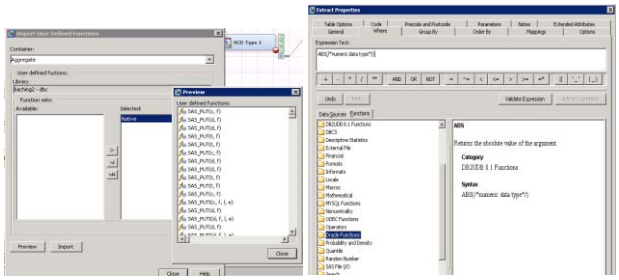

Figure 11: Import User Defined Functions and Formats Examples

ADDITIONAL GRID ENABLEMENT CAPABILITIES

The new version of SAS Data Integration Studio provides additional Grid integration to allow better workload balancing. This feature utilizes SAS® Grid Computing to achieve scalability and performance for advanced workload balancing in complex environments. Using a grid allows administrators more fine-grained control over their environment while providing developers the benefit of finding the server on which their job can run most efficiently.

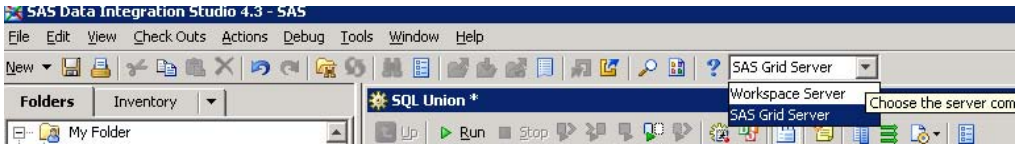

Figure 12: Interactive Submit Options in Data Integration Studio

Interactive submits to a grid give administrators the ability to configure and automate workload through prioritization, implementation of resource utilization thresholds, including suspend and resume, and providing the ability to limit the number of concurrent jobs. In addition, Grid supports the ability to implement run policies, such as a Fair Share policy, which allows prioritization of jobs based on user and workload.

When running interactively on a grid, in previous versions Data Integration Studio would create a new session for each job execution and terminate the session when the job finished. Starting with version 4.3, Data Integration Studio will keep the session open until the user closes the job. This supports incremental job development better, since intermediate work tables will remain while the session is up allowing you to inspect run results. You can also use the various debugging features available such as running specific transforms individually.

Grid must be configured correctly to optimally support the interactive use pattern described above. To improve performance, administrators should increase the number of job slots available and utilize resource thresholds to handle high concurrency. For additional details on how to best optimize a Grid deployment when doing interactive submits to a grid, see [1] in the references at the end of this paper.

TRANSFORMATIONS

There are a number of new transformations available in the latest release of SAS Data Integration Studio. These transformations help you develop higher performing, more efficient jobs.

COMPARE TABLES

The Compare Tables transformation compares two tables (an update and a master table) and classifies records as follows:

1. In Update, Not In Master: *New*
2. In Update, In Master, Changed Fields in Update: *Updated*
3. In Master, Not in Update: *Missing*
4. Unchanged

The job below shows an example use of the transformation:

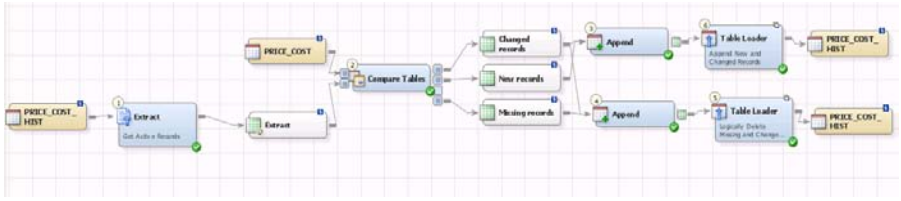

Figure 13: Example Job that Uses the Compare Tables Transform

In the example, the PRICE_COST table represents the update table and the PRICE_COST_HIST table is the master table. The PRICE_COST table could represent a marketing application table that contains the unit cost and unit price of all products currently being promoted through a channel. The PRICE_COST_HIST table keeps a historical record of all unit costs and unit prices for current and previous campaigns. The compare tables would transformation would update the tables as follows:

1. New PRICE_COST records are added to PRICE_COST_HIST
2. Any updates to UNIT_PRICE or UNIT_COST for a PRICE_COST record cause the current historical record to be logically deleted and the updated record is added.
3. Records that now longer appear on the current PRICE_COST table are logically deleted from the PRICE_COST_HIST table
4. All logically deleted records are available for query or reactivation, since they are retained in the table but are marked as deleted.

This transformation supports either a direct lookup (hash object) or disk based compare through a MERGE statement. The hash lookup will perform faster but requires that the entire table fit into memory. If this is not the case, you can choose the MERGE statement method instead. The transformation can handle New, Update Missing, and Unchanged tables as output. You can choose to retain or delete any of the possible outputs as needed to increase efficiency. The transformation also generates its results in a single pass of the data.

SLOWLY CHANGING DIMENSIONS TYPE 1

The SCD Type 1 Loader transformation is useful for tables where all column changes will be handled via overwrite (Type 1 processing). It operates by comparing an incoming table against a master table and handles the following cases:

1. On Incoming, Not on Master: Action = Insert Rows to Master Table
2. On Incoming, On Master, Changed Columns: Action = Update Matching Row In Master Table
3. On Incoming, On Master, No Changed Columns = Ignore Row

The screen shot below shows an example job which uses the SCD Type 1 transformation

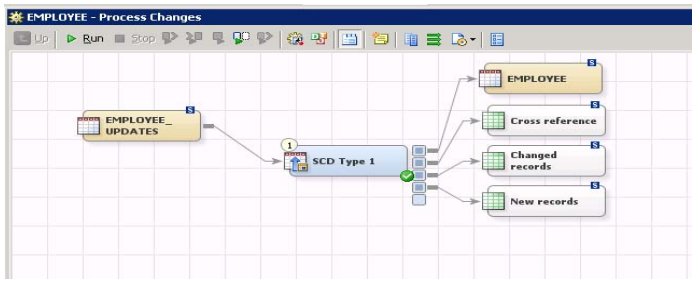

Figure 14: Example Job Using the New Slowly Changing Dimensions Transform

As shown, the incoming table (EMPLOYEE_UPDATES) is mapped as the sole input and the master table (EMPLOYEE) is mapped to output. In addition to updating the master table the transformation can optionally produce other output tables such as a cross reference table, which includes the business key, surrogate key and the change digest used for comparisons.

SQLSET OPERATORS

The latest version of SAS Data Integration Studio adds support for 3 SQL Set operators:

1. UNION
2. INTERSECT
3. EXCEPT (aka MINUS)

These operations are generally defined as depicted below:

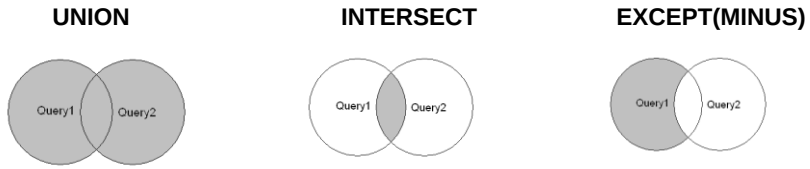

Figure 15: Join Types

Each operator also supports the "ALL" option, which removes duplicate rows from the result set. The advantage of using the SQL Set Operator transformation over the existing SQL Join transformation is primarily for performance. You can write complex queries on either side of the SQL Set Operator, combining results sets that have the same attributes but require different access paths. You can often do this with less complication by just creating two select statements and then combing them with a SQL Set operator rather than trying to integrate all logic into a single join.

Another advantage of the SET operator method is due to the limitations of a SQL join. In order to achieve the same result using a join, frequently you have to use the "NOT IN(<subquery>)" expression which is almost always slower than using a SET operator.

The job depicted below shows an example of a SQL SET transform:

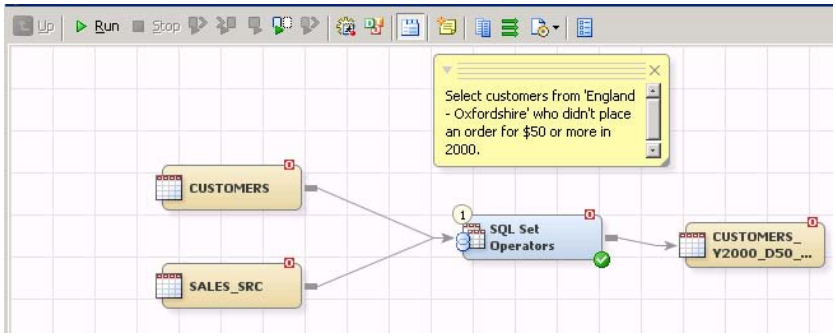

Figure 16: Example Job Using the SQL Set Operator Transformation

This transformation also supports full pushdown capability so that the entire transformation will be pushed down to the database when the source and target schemas match.

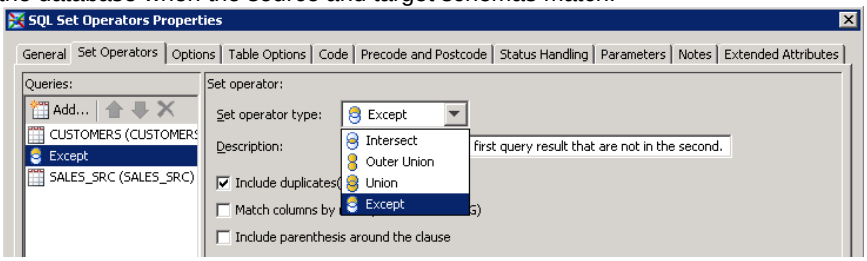

Figure 17: Example settings

OPTIMIZED ORACLE BULK LOAD SUPPORT

A new loader transformation has been added to support optimized bulk load support for the Oracle database. This transformation supports the ability to configure all options applicable to a bulk load of Oracle. You can select options affecting how indexes, constraints and table statistics are handled:, including the percentage of rows sampled when gathering statistics.

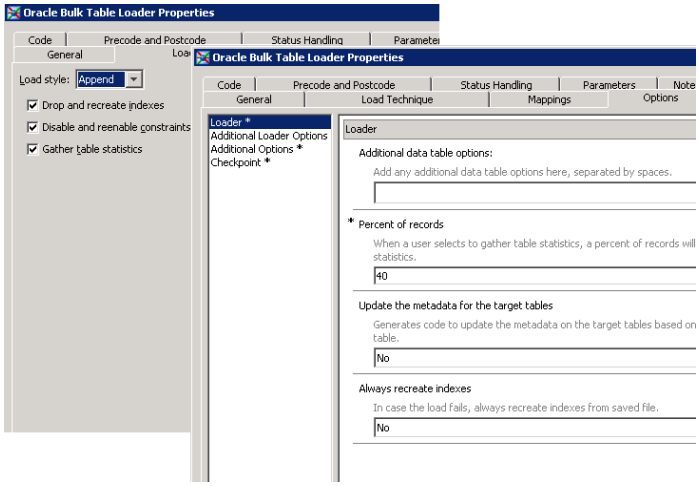

Figure 18: Oracle Bulk Loader Transformation Options Panel

Other important load options for Oracle including partition support and commit level and Direct Path load support are also configurable using this loader.

DATA MANAGEMENT PLATFORM

The latest release of SAS Data Integration includes the new DataFlux Data Management Platform which supports many new data quality, real-time data management, and parallel process management capabilities. Data Management Platform includes Data Management Studio, which leverages features and capabilities of the platform.

JOB MANAGEMENT FEATURES

DataFlux® Data Management Platform has the ability to manage independent job processes. You can implement logic between job flows to manage errors, run jobs in parallel, and pass process related data between your jobs.

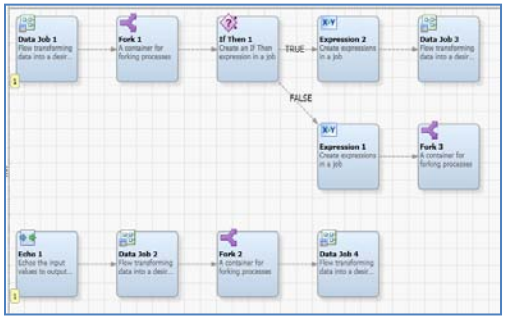

Figure 19: Example Parallel Job Flow

As shown above, you can setup parallel flows that run independently of each other with logic steps to handle job status between steps in your flow. You can check return code values or publish variables out of steps in your job and create decision logic in the flow to take different paths based on that variable value. You can also publish asynchronous events out of your job nodes that you can listen for and take some additional flow path based on the results of the event.

DATA VISUALIZATION FEATURES

There are a number of useful data visualization features available in DataFlux® Data Management Studio. You can view the structure of your data to see how the tables in your system are related. The data model view shows tables related by primary/foreign keys and indexes.

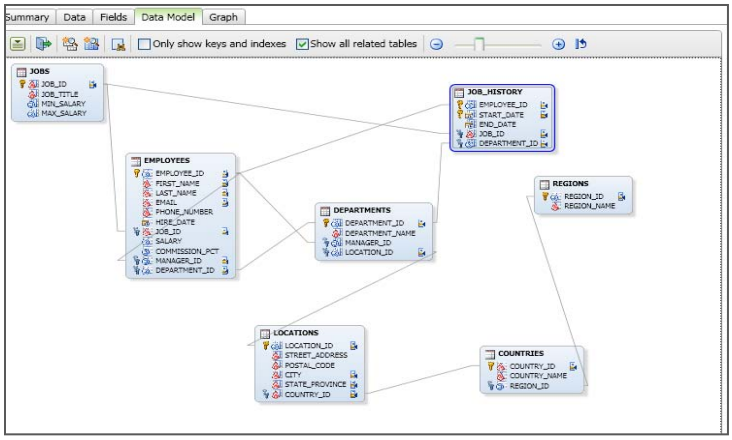

Figure 20: Data Model View

There are also some useful statistics that you can use to help you better understand the data contained in your tables. There are simple graphs that show how the data is distributed in specific columns in your table, as well as showing the length of the data and mean, and median values. You can also plot one column by another and view sum or count statistics for the data. This is done interactively on a data sample via a simple point and click interface.

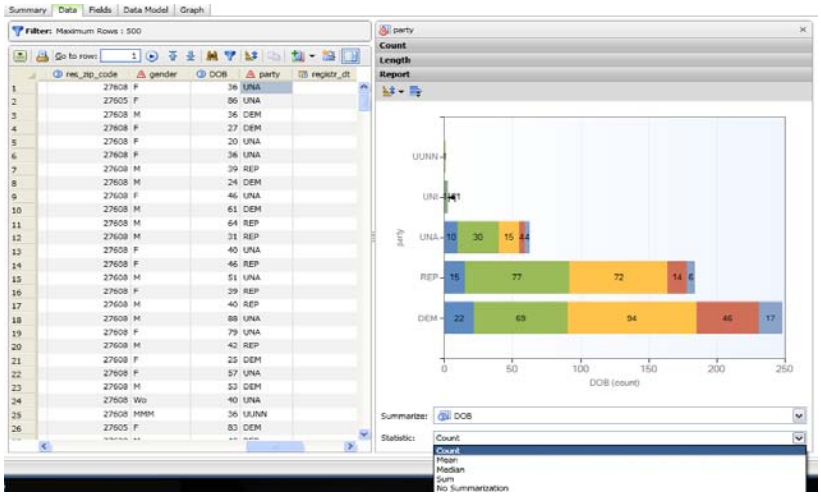

Figure 21: Data Report Statistics View

SAS INTEGRATION AND SUPPORT FOR WEB SERVICES

You can also run SAS programs, including jobs that you generated using SAS Data Integration Studio, in the Data Management Platform. These jobs can be included as a part of job process orchestration, integrated in with other data quality jobs, or as real time services. Data Management Platform now has the ability to run both data and process real time services, as well as call 3rd party Web services from its process layer.

The SAS code node in Data Management Studio uses an intelligent SAS code editor with integrated help, auto-complete capabilities, and syntax checking. You can reference a file of SAS code, such as the file generated as part of a deployed job from Data Integration Studio, and include it in your jobs, or copy/paste code directly into the node. This code will run on any SAS workspace server as a separate process and return runtime status and logs back to the calling job. You can also pass in macros from an outer process and get back run results, the SAS log, and other values from the job run.

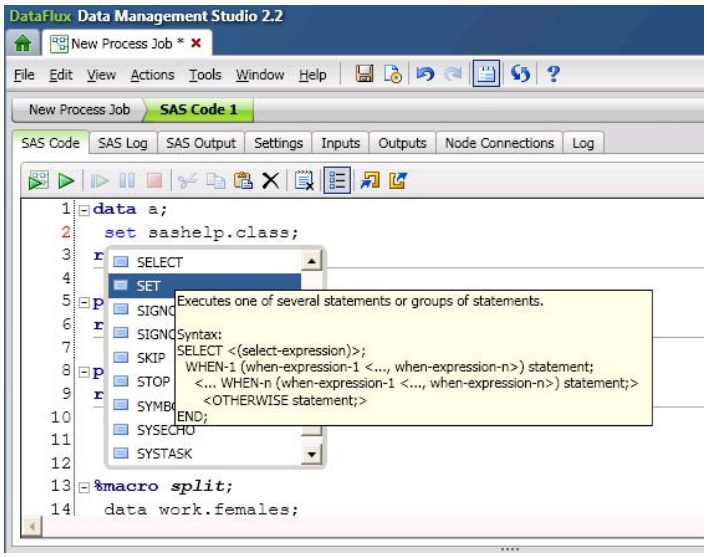

Figure 22: SAS Code Node Editor Example

ELT FEATURES

There are some new transformation nodes available in Data Management Studio that help increase the performance of your jobs when working with relational DBMS data using SQL. New nodes are available that generate SQL for creating tables, inserting data into tables, and updating tables. These nodes include the ability to set table options such as bulk load options.

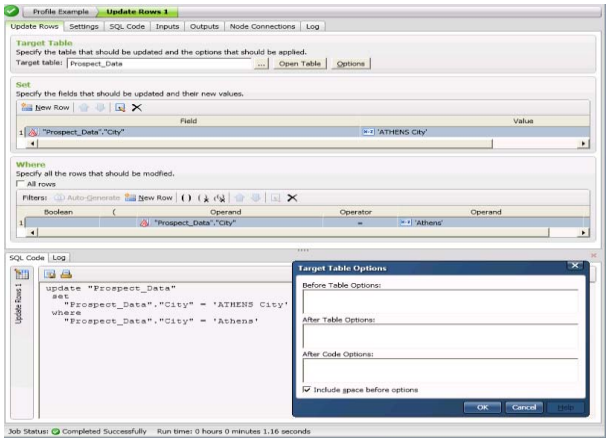

Figure 23: SQL Update Transformation Example

There is also a node that can be used to call any SQL script. This node can be useful if you have one or more SQL scripts that you want to manage from your processes. For example you might have a script that sets up some tables prior to running some other job on the data. You can use Data Management Studio to visualize and help you manage your runtime environment for these types of processes.

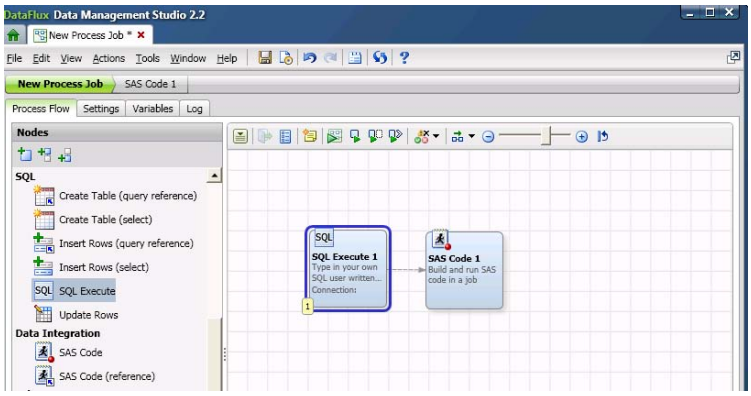

Figure 24: SQL Execute Node Example

CONCLUSION

The latest releases of SAS® Data Integration Studio and the integrated DataFlux Data Management Studio 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, the introduction of new transformations to assist you in optimizing your job flows for common data integration tasks, and enterprise features such as versioning and rollback support. The introduction of the complimentary DataFlux Data Management Studio product helps leverage the capabilities provided in SAS Data Integration Studio and introduces new capabilities around data visualization, job management, parallelization, and database integration. Customers will find many reasons to upgrade to the latest version of SAS Data Integration. .

REFERENCES

Rausch, Nancy A., and Tim Stearn, 2011, "Best Practices in Data Integration: Advanced Data Management", Proceedings of the SAS Global Forum 2011 Conference, Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings11/137-2011.pdf.

RECOMMENDED READING

- SAS Enterprise Data Management and Integration Discussion Forum, Available at http://support.sas.com/forums/forum.jspa?forumID=59
- 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.
- Stander, Jeff. 2010. " SAS® Data Integration Studio: Tips and Techniques for Implementing ELT." Proceedings of the SAS Global Forum 2010 Conference. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings10/116-2010.pdf.
- Hunley, Eric, and Nancy Rausch. 2009. "What's New in SAS Data Integration Studio 4.2." Proceedings of the SAS Global Forum 2009 Conference. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings09/093-2009.pdf.
- Doninger, Cheryl, and Nancy Rausch. 2009. "Data Integration in a Grid-Enabled Environment." Proceedings of the SAS Global Forum 2009 Conference. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings09/098-2009.pdf

Contact Information

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

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