ABSTRACT

Statistics Canada is Canada’s Governmental agency that collects and processes data for statistical purposes via about 350 surveys in order to disseminate the nation’s statistics. We are the largest user of SAS® in Canada and are researching SAS Data Integration Studio as a tool for integrating our vast stores of data. This paper is meant as an overview of SAS Data Integration Studio. It will present some of the many, many advantages of using SAS Data Integration Studio.

INTRODUCTION

Data Integration Studio is a natural fit for our environment at Statistics Canada because we do a lot of data transformation. We are constantly reformatting, renaming, calculating, converting, concatenating, subdividing and anything else you can imagine you can do with data.

Many months of research have been spent exploring the advantages of SAS Data Integration Studio at Statistics Canada. In this paper, a definition of Data Integration is presented. Then, through a series of screen shots and text, this paper summarizes what I see as some of the principal advantages and some of the disadvantages of SAS Data Integration Studio.

WHAT IS DATA INTEGRATION?

Imagine each source of data represents a different, small, assembled puzzle. Each individual, assembled puzzle represent different pictures of the same man. Puzzle number one might represent a picture of this man in a suit. Puzzle number two might represent a picture of this man dressed like he is on vacation and finally, puzzle number three might represent a picture of this man dressed up in a police uniform. Each puzzle represents a different version of reality concerning this man – just like each source of data in our organization represents a different view of our respondents to surveys conducted by Statistics Canada.

Now, take a part all three puzzles and put all of these puzzle pieces which represent different pictures into one big bag and shake the bag around to mix them all up. Try and put these pieces of all these smaller puzzles together to make one big, huge puzzle and one big, huge picture. This would be a seemingly impossible task because the puzzle pieces could be made of different materials, have different shapes, and represent different pictures.

Similarly, all kinds of problems are encountered when making one master data source or one version of reality from different data sources. Just like the pieces of the puzzles, the pieces of data are all unique in their own way. Let’s use as an example as a data item called marital status. If there are no standards that exist, the same data item from different sources could have different names, lengths, code sets, and formats. Maybe in one data source, the marital status item could have a length of 25 characters, have values such as “Married”, “Single”, “Separated” and have five possible values. While in another source, the marital status could be coded with a length of 1 character, have values such as “1”, “2”, “3” and can consist of 9 possible different values.

If you would like to combine these two data items from different sources into one data item, a process of data harmonization or integration has to be done. The field would have to have a standard length, a standard value and a standard code set and so on. The process of fitting these data items together to form one source of data or one picture of reality is the process called data Integration. Without the process of data integration, it would be impossible to fit the different sources of data together to create a master data source (just like it is impossible to fit pieces of the three different puzzles together to form one big picture).
FEATURES OF SAS DATA INTEGRATION STUDIO

According to my SAS course notes, “it is a visual design tool for:

1. building processes,
2. implementing processes, AND
3. managing processes.”

Regardless Of: Data sources, Applications, OR Platforms

There are too many SAS Data Integration Studio features to go through in this paper. Therefore, this paper will focus on the features that this author finds to be the most useful:

1. Meta Data Rich
2. Visual
3. Integration with SAS® Dataflux
4. Create stored process and SOA modules
5. Schedule jobs via the platform scheduler
6. Import existing SAS Code
7. Push the processing to the database itself.
8. Integration with Third Party Source Control
9. Security
10. Can import column definitions from other tables
11. Impact Analysis/Reverse Impact Analysis

META DATA RICH ($)

As a client to the SAS Platform, SAS Data Integration Studio has access to and can be used to generate a large volume and variety of metadata. Metadata, in this context, does not only refer to the classical definition of “Data about Data”. It is data about almost any element in the computing environment such as servers, users, schedules, processes, data mappings, data sources, and data columns.

In fact, SAS Data Integration Studio is nothing without metadata. If you click cancel when asked to log onto the metadata server, there is absolutely nothing displayed. SAS Data Integration Studio saves you time and money by allowing the user access to this tremendous amount of metadata in developing SAS Data Integration Studio jobs. It allows the user to develop modules that can be run anytime, on any server with any type of data.

VISUAL

One of the best features of SAS Data Integration Studio is that it is very, very visual. It reads like a flowchart so much so that there is one SAS Global Forum paper that suggests it can be used as a Computer Aided Software Engineering (CASE) tool to communicate systems design between the systems designer and the client(s) for which the system is being designed. This visual feature of SAS Data Integration Studio is possible due to the vast array of transformations made available by SAS Data Integration Studio. The flowchart produced by Data Integration Studio becomes part of your system documentation which definitely makes the system easier to maintain.

A transformation is a pre-defined process that takes an input and turns it into something else in the form of an output. The popular data processing model of input-process-output is at the core of the transformation. The input of the transformation is called the source, the output is called the target and the process is the transformation itself. There are seventy three transformations in Data Integration Studio 4.4 that can be used to make a job that is easy to understand. For example, there are transformations for sort, load and transpose.
For example, once the user is comfortable with the transformation concept of Data Integration Studio, he can quickly grasp the meaning of the job stream below (see following screen capture.):

1. There exists a table.
2. There is some sort of extraction that takes place and
3. The result is sorted.

Display 1: Example Data Flow

If there is no transformation available to accomplish the desired task, a new transformation can be created to perform the specific task. This ability to create custom transforms greatly adds to the flexibility of SAS Data Integration Studio.

INTEGRATION WITH DATAFLUX

SAS Data Integration Studio has transformations that integrate some of the functionality of the SAS Dataflux product. The SAS Dataflux product is used to cleanse and standardize data elements such as names and addresses. Records from different data sources can be linked together via these new, standardized data elements along with any other data elements on the files being linked together. For example, you might want to link two data sources based on date of birth, address and gender.

This function will immensely increase the value of SAS Data Integration Studio at Statistics Canada as we deal with data quality issues all of the time and perform record linkage on a regular basis.

CREATE STORED PROCESSES AND WEB SERVICES

SAS’s devotion to allowing the user to be able to run SAS anywhere really shines with the ability to create stored processes and web services.

As in SAS® Enterprise Guide, users are able to convert their job flow into a stored process. By using the SAS® add-in for Microsoft® Office, stored processes can be run from any popular Microsoft package such as Outlook, PowerPoint and Word. Stored Processes can also be run on SAS clients such as SAS® Enterprise Guide and SAS® Web Report Studio. Once the job is deployed as a stored process, the client can deploy the stored process as a service which would allow the service to be integrated into a system which uses the Service Oriented Architecture (SOA) standard.

Simply by right-clicking on a job and picking the appropriate menu item, the job can be converted into a stored process via an easy to use wizard. Similarly, by right-clicking on a stored process, the stored process can be converted into a web service via an easy to use wizard.

SCHEDULE JOBS VIA THE SAS® PLATFORM SCHEDULER

The platform scheduler is an industrial-strength scheduler and is a complement to SAS Data Integration Studio. Scheduling a DI Studio job is a two step process. First, the job is deployed from SAS Data Integration Studio and then it is scheduled via the platform scheduler. A job cannot be scheduled directly from SAS Data Integration Studio. In order to activate the platform scheduler, it must be activated via a plug-in in SAS® Management Console.

It works with job flows just like SAS Data Integration Studio (see following screen capture). Deployed jobs can be chained together and you can have one job within another job. Jobs can be activated according to events such as the occurrence of a date and time or if a certain file exists or upon completion of another job. The platform scheduler can even be upgraded to be used to run jobs other than SAS jobs.
Deployed jobs can be run on different servers and on a SAS grid.

Display 2. Picture of a Scheduled Flow from within the Platform Scheduler

**IMPORT EXISTING CODE**

Base SAS users are not left out in the SAS Data Integration Studio world. SAS Data Integration Studio will import existing Base SAS code. It will do its best to try and import your Base SAS code and turn it into a SAS Data Integration Studio Job. This import is not perfect – so, the imported job may demand some manual intervention to make the job work in the SAS Data Integration environment.

**ABILITY TO PUSH THE PROCESSING TO THE DATABASE ITSELF**

In certain situations, it is optimal to push the SQL processing to the database instead of running it directly in SAS. For example, a SAS Data Integration Studio job could be set-up to extract data from a database such as Microsoft® SQL Server, derive some variables and push that data along with the derived variables back to the same Microsoft database. In this instance, it would probably be more efficient to push the processing to the database.

In order to explicitly indicate that you wish to push the SQL processing to the database, set the “optimized pass-through facility for SQL statements” option to “Yes”. This option can be found on the options tab of the transformation properties window (see following screen capture):

Display 3. Additional Options screen

Traditional processing performed by products such as SAS Data Integration Studio is usually thought about in terms of extracting data from their sources, transforming or processing the extracted data and lastly, load the transformed data into their target database (ETL - Extract, Transform and Load). SAS Data Integration Studio also allows the user to submit non-SAS code to the database and have it processed from within the database instead of by SAS. With this method SAS Data Integration Studio allows the user to extract the data first, load the data into its target table and transform or process this data in the actual target database (ELT - Extract, Load and Transform).

Giving the user of SAS Data Integration Studio a choice between ETL processing and ELT processing is a feature that gives the user tremendous flexibility in terms of optimizing their jobs.
INTEGRATION OF THIRD PARTY SOURCE CONTROL

In SAS Data Integration Studio 4.3, one of the major improvements is support for integrating SAS Data Integration Studio with two open-source versioning systems: SubVersion and CVS. Also, there is an API for other source versioning systems.

This feature allows the user to enter the world of “managed” code. The user can specify a version for individual objects or for groups of related objects. Also, the user can choose to store the code that the job generates along with the source code. As well, the system can compare versions of objects and display the differences between the different versions.

With managed code, changes can be documented and the history of these changes can be tracked. The ability to rollback to previous versions of objects is one of the most powerful advantages of this feature.

SECURITY

Since SAS Data Integration Studio is one of the many clients of the SAS Platform, it can leverage the security features of the SAS Platform via the SAS Metadata Server. There is group-based security and role-based security. Group-based security is used to manage access to application features; whereas, role-based security tells what a user can do after they have access to a specific application (what data they have access to). Given the importance of security these days, this is an important advantage of being able to leverage the SAS Platform.

IMPORT METADATA FROM OTHER TABLES

The metadata of any columns can be imported from other tables. This feature leads to more consistent definitions of columns within SAS Data Integration Studio jobs. Instead of going to a data dictionary or other tool to look up a data definition, a data column can be imported from an existing table. Not only does the name, length and data type of the column get imported, but the format, description, and label also do (see following screen capture.).
IMPACT ANALYSIS/REVERSE IMPACT ANALYSIS

The user can right click on any object within a job and ask for an “analysis”. This “analysis” of the object tells the user which jobs and which objects will be affected by a change to the object being analyzed. This feature facilitates the inevitable changes that occur in a job as a result of changes in the environmental, specifications, records layouts, etc....

Display 5. Impact Analysis
PERFORMANCE STATISTICS

Performance statistics are collected for each job and displayed in a tabular format or graph in the Statistics tab below the job. The name, status, records, start time, end time, duration, CPU time, memory usage, I/O, server and the number of threads that are collected for the processes in the job (see following screen capture). This is a tremendous help for the job designers wishing to optimize their SAS Data Integration Studio jobs.

Using these statistics, the job designers can identify any bottlenecks in their jobs and take appropriate actions. For example, depending on the problem, the solution could be to push the processing to the database as we discussed earlier, push the process to different servers, add new hardware or re-design the SAS Data Integration Studio job altogether.

Display 5. Statistics Tab

ABILITY TO RUN THE SAS DATA INTEGRATION STUDIO JOB ON A SAS GRID

SAS Data Integration Studio gives the user the ability to run the Data Integration job on a grid. Instead of rolling the job on one server, the job can be run on several servers. This ability to run jobs on several servers can greatly speed up the job execution time for certain jobs. For example, a job that uses a lot of by statements that takes several hours to run can take advantage of the SAS grid to run in a lot less time.

DISADVANTAGES OF SAS DATA INTEGRATION STUDIO

This paper would not be complete without mentioning some of the disadvantages of SAS Data Integration Studio. The advantages greatly outweigh the disadvantages of SAS Data Integration Studio, but the following disadvantages should be known and understood before adopting the tool:

- Installation of SAS Data Integration is complex because it exists as part of the SAS Platform and it is a client-server application.
- There is definitely a big learning curve with SAS Data Integration Studio. If you have never worked with the Platform metadata before, it is a very different paradigm. It takes a while before the user can be comfortable working with the different types of metadata available in the SAS platform.
- Also, there is the overhead of a metadata administrator required. With SAS Data Integration Studio, for example, you cannot simply create a table. This table must be registered in metadata before it can be accessed effectively. A metadata administrator would take care of the metadata permissions, server definitions, and library definitions, etc…

The same reasons why these disadvantages exist; however, is why a lot of the advantages exist in SAS Data Integration Studio exists. As was stated before, the advantages of SAS Data Integration Studio greatly outweigh the disadvantages. For example, it is expected that the introduction of any sophisticated tool implies some kind of a learning curve. Passed this learning curve however, it is expected that the user will become more productive.
CONCLUSION

In conclusion, the many features of SAS Data Integration Studio make it an invaluable addition to the Statistics Canada toolbox where processing data is a daily ritual. SAS Data Integration Studio allows the user to do the Extract, Transform and Load (ETL) process or Extract, Load and Transform (ELT) process in a more transparent way than using Base SAS. With these many features, people should not be asking “Why use SAS Data Integration Studio”, they should be asking themselves “Why not use SAS Data Integration Studio”.

With this paper, I hope I have inspired you to Just Go with the Flow and give Data Integration Studio a try!
REFERENCES


RECOMMENDED READING


CONTACT INFORMATION

David Dean
Statistics Canada
R.H. Coats 14O
100 Tunney’s Pasture Driveway
Ottawa, Ontario
Canada
K0A 0T6
E-mail: David.Dean@statcan.gc.ca

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.