Paper 3348-2015

From ETL to ETL VCM: Ensure the Success of a Data Migration Project through a Flexible Data Quality Framework Using SAS® DataFlux Data Management

Yves Wouters, Senior Consultant, Deloitte Belgium

ABSTRACT

Data Quality is now more important than ever before. According to Gartner (2011), poor data quality is the primary reason why 40% of all business initiatives fail to achieve their targeted benefits.

To anticipate the ever growing importance of Data Quality and ensure the success of a business initiative, Deloitte Belgium has created an end-to-end Data Quality framework using SAS DataFlux Data Management to rapidly identify and resolve root causes of data quality issues to jumpstart business initiatives such as a data migration.

Moreover, the framework uses both standard SAS DataFlux Data Management functionalities (such as standardization, parsing, etc.) as well as advanced features (such as using macro's, dynamic profiling in deployment mode, extracting the profiling results, etc.) allowing the framework to be agile and flexible as well as maximize the re-usability of specific components built in SAS DataFlux Data Management.

INTRODUCTION

With the rapid evolution of our society, organizations are more and more embarking on business transformation initiatives to better understand and respond to current market needs. This usually requires organizations to replace their legacy system(s) and/or migrate/integrate data from the (different) source(s) into a new target system.

This migration/integration of data, typically, requires ETL (Extract – Transform – Load) processes that enables the availability of data into the target system. Although the phrase "garbage in… garbage out" is well-known in the information management world, it is often not acted upon due to budget issues and/or time constraints.

This paper focuses on the importance of data quality in such a context by adding three new steps in a traditional ETL process: Validate, Cleanse and Monitor or VCM. These three steps, applied continuously, significantly increases the success as well as the impact and value of a business transformation initiative if focus is put on business priorities. To ensure you understand the framework and the importance of the VCM steps an analogy is described with a car manufacturing process in the next section.

SAS DataFlux Data Management contains several components (such as standardization, deduplication, enriching, profiling, etc.) that support these steps. This paper does not focus on the standard SAS DataFlux Data Management functionalities, but rather wants to demonstrate the importance of applying continuous VCM activities and provides techniques that should stimulate you to integrate SAS DataFlux Data Management in an automated ETL process rather than using it on an ad-hoc basis.

ETL-VCM FRAMEWORK OVERVIEW

The ETL-VCM framework, shown below, focuses on building continuous VCM activities in an ETL context by integrating SAS DataFlux Data Management at each level (Source, ETL, Target) in an automated way. To realize this, key techniques will be described in the next section that enables such a set-up. T

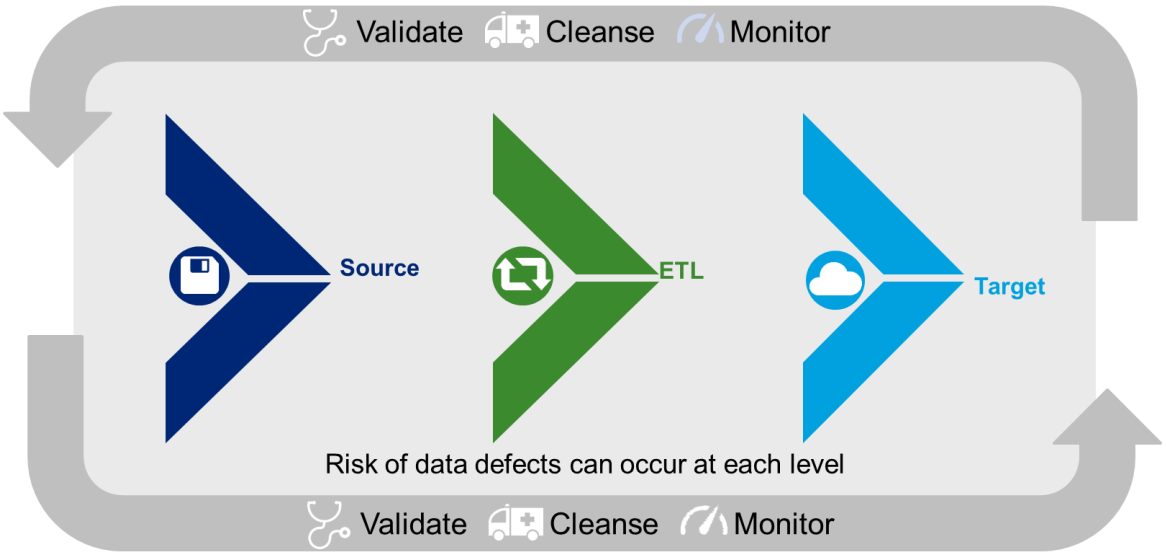


Figure 1: The ETL-VCM Framework

The framework can be best compared with a standard supply chain process. Let's take a car manufacturing process as an example. In a typical car manufacturing process, parts of a car are delivered from different suppliers, they enter the assembly line, and are assembled accordingly in order to produce a car according to the client needs. Although the process, shown in the below figure, seems simple, having the right parts at the right time in the right place is another story.

Figure 2: Simplified supply chain process of a car

Even if the right parts are at the right time in the right place, it does not mean the assembly will go smoothly. Defects within the parts (such as wrong dimensions, wrong colors, etc.) or during the assembly itself can still occur. This in turn impacts the manufacturing process which might lead to a complete stop of the process and results in huge operational costs which could have been avoided.

Today's car manufacturing processes are tremendously evolved compared to Henry Ford's automation of the process about 100 years ago. This is also due to the fact that car manufacturers realize the impact if the manufacturing process is stopped due to defects or if the car is not delivered according to the client needs. Nevertheless, car manufacturers these days are still looking to have a full end-to-end view/control on the supply chain process by following the process as of the moment parts are being constructed at the suppliers' side until the moment the car is delivered to the client.

Typical business transformation initiatives requiring data migration, integration or data warehousing activities are not that different from the process described above:

- Suppliers are the producers of data in operational/legacy sources
- The parts entering the assembly line are the data itself being shipped to the staging area
- The assembly line is the ETL process of transforming the data into the requirements of the customer(s)
- The produced car is the transformed data ready for shipment towards the new target system
- The customers are the consumers of the “new” data in the target system.

Just as in the car manufacturing example, having defects in data could potentially impact the ETL process and forces it to stop or it can lead to wrongly transformed data which in turn leads to unsatisfied consumers of data that do not use the new system or, even worse, lead to wrong decision making. The ETL-VCM framework, conceptually represented in the figure 1, follows the principle that applying data quality throughout the process by using validate, cleanse and monitor activities highly reduces the chances of defects and increases intensely the success rate of a business transformation initiative by applying an end-to-end view on your data supply chain process.

There are 3 basic principles within this framework:

1. Involve the business as much as possible (mandatory) and prioritize their needs based on defects in data with the biggest impact.
2. Apply VCM as early as possible in the process (similar to the car manufacturing process, the more effort at the source level, the less chance on defects later in the process).
3. Identify and resolve the root-causes behind poor data quality (Continuous improvement actions ensures adoption in the operations and avoids future defects).

As indicated in the framework, data defects can occur anywhere in the process. Continuous validation, cleansing and monitoring across the process is a key within the framework.

The rest of the paper will focus on key features using SAS DataFlux Data Management to help integrate it in an automated ETL process and ensure an end-to-end automated data supply chain process.

ADVANCED SAS DATAFLUX DATA MANAGEMENT TECHNIQUES SUPPORTING THE ETL-VCM FRAMEWORK

This section of the paper focuses on providing 5 key techniques that should help you understand how to use SAS DataFlux Data Management so that it can be integrated and automated in any (existing) ETL process.

1. Using Macro variables
2. Dynamic Profiling
3. Extracting profiling results
4. Running jobs from the command line
5. Using Kerberos Authentication to process batch runs

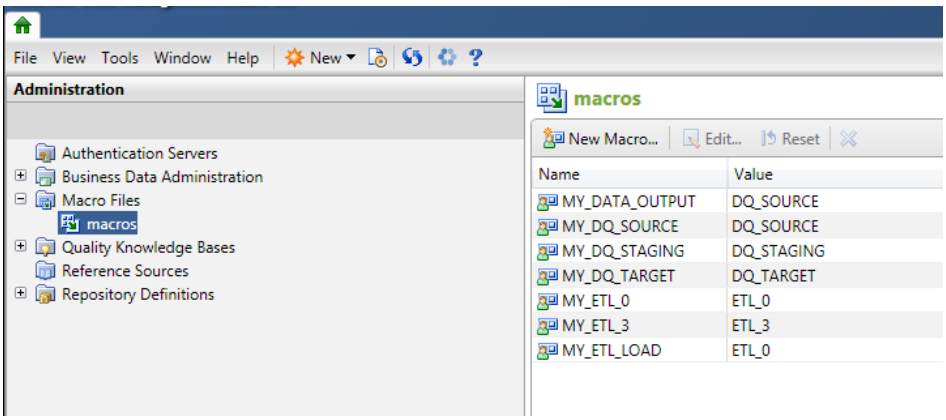
You should realize that the techniques described are very useful if SAS DataFlux Data Management is the only SAS tool used at your project. If other SAS tools are available, other techniques could be applied as well.

Moreover, the techniques help to optimize the usage of the tool when you are working on a project with different environments (Development, Test, UAT and Production) and/or if you have multiple versions of databases related to the target system. It can be for example that database v1 of the target system is set into UAT and Production, but that v2 is currently being used in the Development and Test environment. Additionally, you can also imagine a situation where you only have one version of the target system, but the structure of the source data changes over time. When the different pieces of the puzzle are combined correctly, the ETL-VCM framework can be processed in an automated way be simply specifying the version of the database and the environment you want to process in a scheduler.

TECHNIQUE 1: USING MACRO VARIABLES

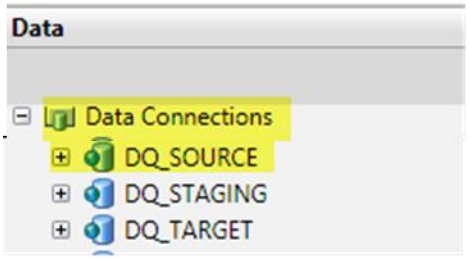
Within SAS DataFlux Data Management, you can define global parameters called macros. By defining a macro with a corresponding value, you reference this value anywhere within SAS DataFlux Data Management. If you want to use a macro variable whether it is in a data job, process job, SQL statement etc., you must use the following syntax: “%%<macro_variable_name>%%”. SAS DataFlux Data Management will consequently fill the parameter with the corresponding value.

Below displays provide an overview of where to find macros in SAS DataFlux Data Management, how to reference them and what the result looks like.



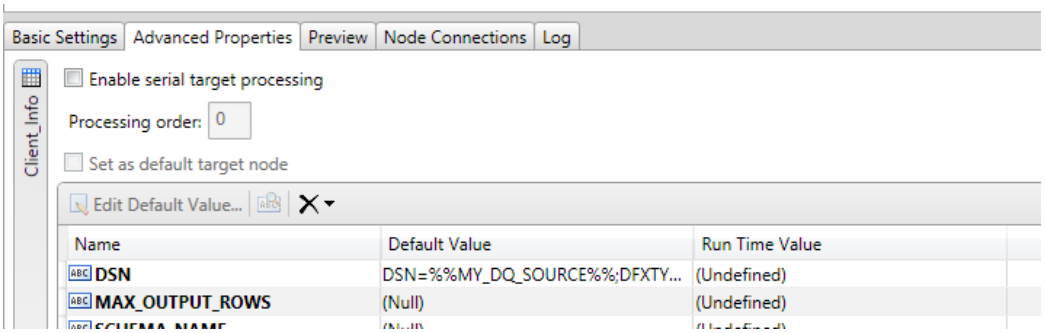
Display 1: Macros in SAS DataFlux Data Management

Macros can be found under the Administration tab by simply clicking on “New Macro” to add a new macro. Provide the macro with a name and a corresponding value. In the screenshot below, the value of the macro is referring to a DSN connection.

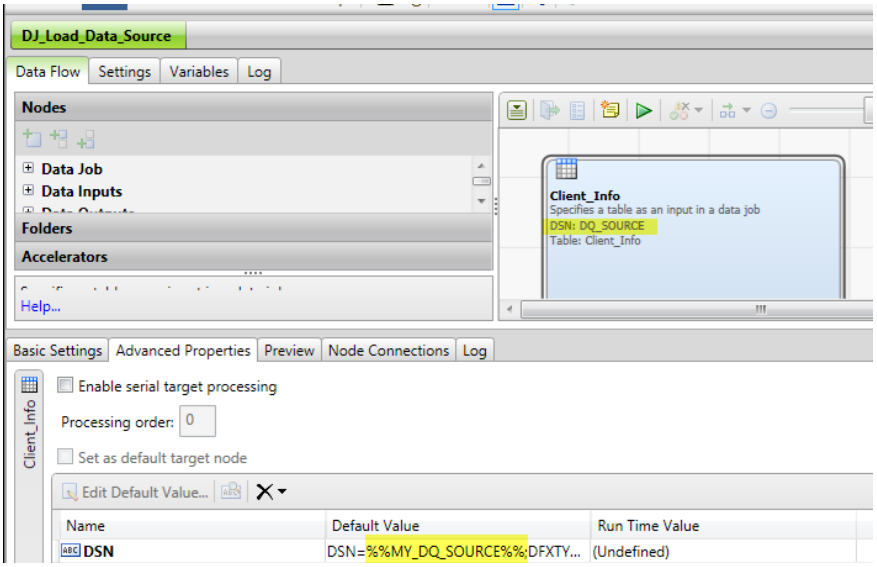


Display 2: DSN definition referred to by the macro

By following the syntax described earlier, you can then use the macro as shown in the picture below where the DSN of a data source is replaced by the macro name.



Display 3: Referencing macros in SAS DataFlux Data Management



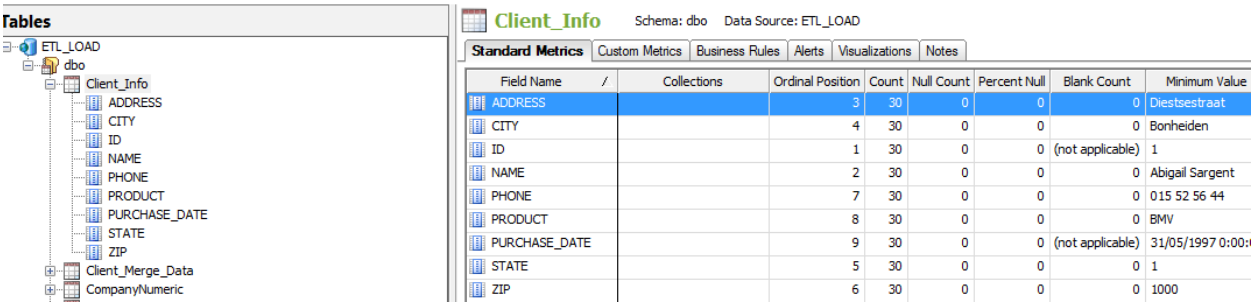
Display 4: Result of using a macro in SAS DataFlux Data Management

As shown in display 4, the DSN is replaced by the corresponding macro variable. The example shown in above pictures shows that using macro variables to call a specific data source allows you to develop a data quality validation and/or cleansing job once, but re-use it many times on different environments.

TECHNIQUE 2: DYNAMIC PROFILING

Static Profiling

A profiling job in SAS DataFlux Data Management scans your data on numerous elements to help you better understand your data and its quality (e.g. mandatory validation, numeric validation, date validation, but also pattern frequency, min and max ranges, etc.). Below example shows the results of such a profiling job.



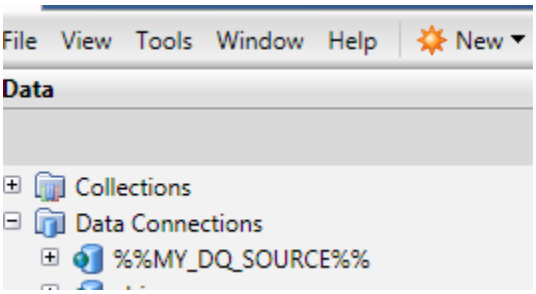
Display 5: Example of a profile result

Typically, this profiling exercise is done in a rather static way. The data source is selected, the profiling is run and the results are reviewed. As you can see from the display above, the profiling references to one specific environment and related database. However, taking the migration example, it means that for each environment and related database the profiling has to be recreated, even if the structure of the database is exactly the same.

Creating a dynamic profiling job

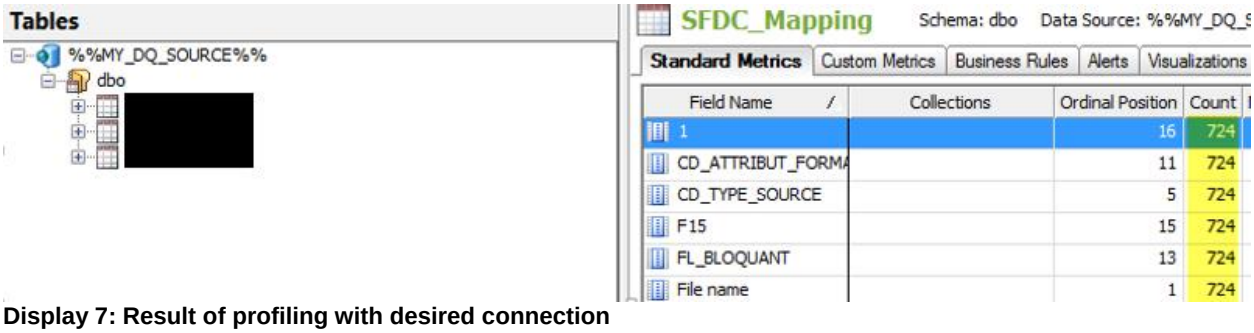
To avoid this, profiling in SAS DataFlux Data Management can be created dynamically using the macro variable principle described earlier. The below steps explain how this can be achieved.

1. Create a new DSN connection, but label the connection with the macro variable created as described earlier.



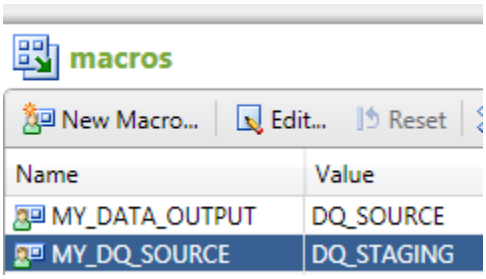
Display 6: Creating a dynamic DSN using the macro variable

2. Create and run the profiling. The profiling will take the value of the macro into account.

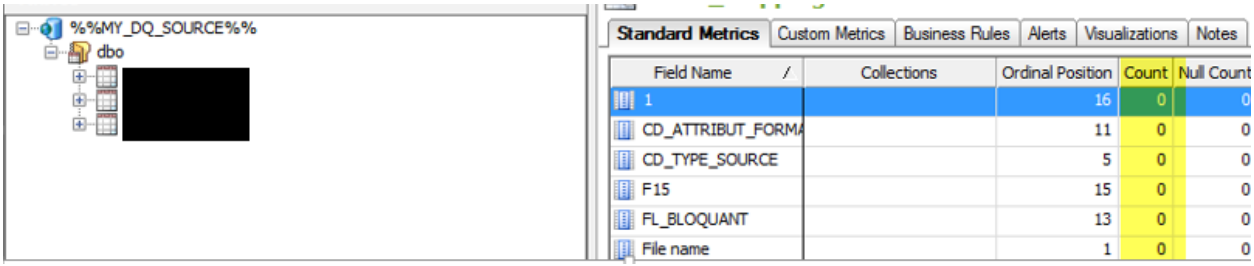


Display 7: Result of profiling with desired connection

3. Change the value of the macro to the new environment you want to profile (please note that the structure of the database tables being profiled need to match the structure defined in the initial profiling) and re-run your profile job.



Display 8: Change the desired connection via the macro



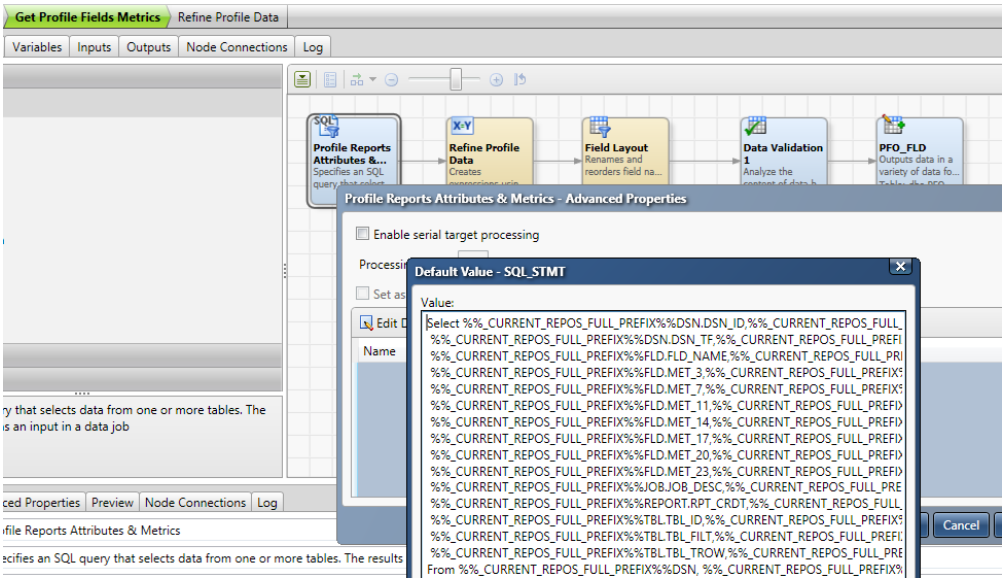
Display 9: Results of profiling with the new connection

As shown in the displays above, the same profiling is used, but the results are different. Our staging environment does not yet contain any data while the source did.

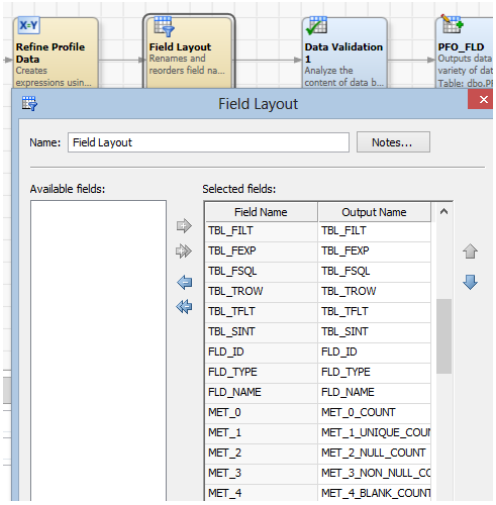
TECHNIQUE 3: EXTRACTING PROFILING RESULTS

Profiling results of SAS DataFlux Data Management are stored in its related repository. To support the ETL-VCM framework, extracting the profiling results and loading them into a database allows the framework to re-use these results. Moreover, these profiling results can then be automatically compared with the expectations from the business without having someone to review it manually.

In order to achieve this, it is important to know how the profiling results are stored in SAS DataFlux Data Management and what the meaning is behind the different fields. The displays below provide an impression on how to extract profiling results.



Display 10: Extracting the profiling results



Display 11: Providing meaning to the profiling column names stored in the repository

TECHNIQUE 4: RUNNING JOBS FROM THE COMMAND LINE

SAS DataFlux Data Management contains an executable command file called “dmpexec”. The storage location can vary, but usually can be found under: *Computer drive:\Program Files (x86)\DataFlux\DMStudio\studio1\bin\dmpexec.cmd*

This command file allows you to execute different data, profiles and/or process jobs at once. Moreover, it can be used to create log files (using “-l”) of each job executed so that it can be easily traced afterwards which jobs failed. The below code is a sample of how this command script could be used.

```
@echo off

set d=%date:~11,4%%date:~8,2%%date:~5,2%

set directory=Computer_drive:\SAS\SASHome\SASDataFluxDataManagementServer\bin\dmpexec -b
"REPOSITORY=Server Repository" -l
```

set log_file= Computer_drive:\any_location_desired

set run_dir=-j Computer_drive:\SAS\SASHome\SASDataFluxDataManagementServer\var\batch_jobs

%directory% %log_file%\%d%\.log %run_dir%

In a migration context, the above can be taken a step further. Migration code normally has to go through different environments (e.g. Development, Test, UAT, Production) before it is released in production. Different database versions can be implemented at each environment. By incorporating additional fields in the script above, you are then able to run versioned DQ jobs related to a specific database version on any environment (Dev, Test, UAT, etc.) and trace loggings of each specific run.

TECHNIQUE 5: USING KERBEROS AUTHENTICATION TO PROCESS BATCH RUNS

SAS DataFlux Data Management can be used together with “Kerberos Authentication protocol” In order to remotely execute data quality jobs from the DQ server onto data existing on another server (e.g. Microsoft® SQL Server) using a scheduler tool.

The below representation is a visual overview of such a set-up with a link from SAS DataFlux Data Management Server towards SQL Server as an example.

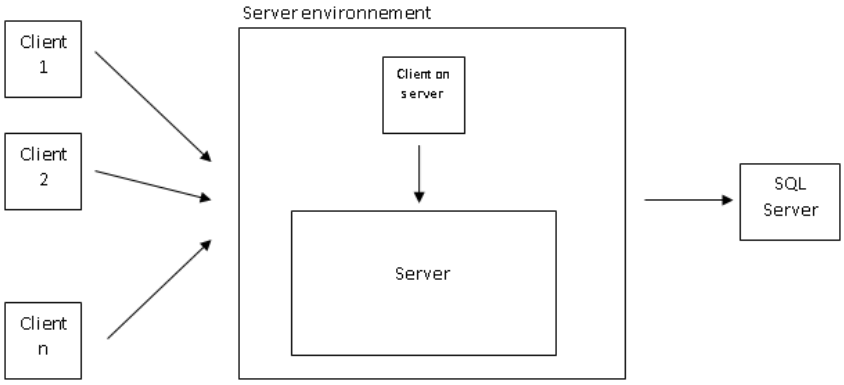


Figure 3: A Data Management server connecting to a SQL Server

Overview of the set-up:

- Local client: the machine(s) on which the DQ team works daily to make sure the correct DQ controls are performed and the data quality is analyzed.
- Data Management server: A central location to save work that needs to be executed by the server. From this central location, all the jobs can be called upon to verify the current data quality when performing an ETL run.
- SQL server: A server on which the data to check is located and the results of the DQ batch run are saved

Data Quality jobs and other Data Quality files can be created daily on local machines depicted as Client 1, Client 2, etc.. At a certain point, these items are copied to the Data Management server in order to have a centralized location to run everything from one directory during a batch run.

The server environment can be accessed by a "Remote desktop control" which can be seen as a user of the server on the actual server environment itself. This user is portrayed as a "Client on server" on the figure above and allows to create a Data Management client on the actual server.

During the batch run, a scheduler is used to call and run the centralized directory located on the SAS DataFlux Data Management server. The latter is accessed using known credentials. However, depending

on the security protocols within an organization, it can be prohibited to access a server from another server. As a result the credentials used to authenticate to the Data Management server are not retained towards the other server where the data is stored (in this case SQL Server).

This can be avoided by using the Kerberos authentication protocol which is supported by SAS DataFlux Data Management. By using the protocol, a new session is created on the SAS DataFlux Data Management server where a connection is made with the Key Distribution Center that acts as a trusted third party. This third party, in return, hands over a ticket containing the authentication necessary to connect to the SQL server. The received ticket remains valid until the end of the session.

CONCLUSION

Just as in the car manufacturing process example, identifying and resolving defects in your data as early as possible in the process is key to ensure the success of your business transformation program. Moreover, you need to ensure that no other defects occur during the next steps of your data supply chain process such as transforming the data and loading it into the target system for use by your customers.

The ETL-VCM framework helps you integrate SAS DataFlux Data Management to ensure this end-to-end data supply chain process by combining the following 5 techniques:

1. Using Macro variables
2. Dynamic Profiling
3. Extracting profiling results
4. Running jobs from the command line
5. Using Kerberos Authentication to process batch runs

RECOMMENDED READING

- *DataFlux Data Management Studio 2.5: User Guide*
- *DataFlux Expression Language 2.5: Reference Guide*

CONTACT INFORMATION

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

Yves Wouters
Deloitte
+32476578171
ywouters@deloitte.com
http://www2.deloitte.com/be/en.html