Paper 328-2010

dATa qWaliti 4 Analytics

David Barkaway, SAS Institute Inc., United Kingdom

ABSTRACT

Data quality is one of the largest challenges facing business reporting and analytical projects today. Poor data quality can undermine the success of what would otherwise be well-developed applications by stripping the business value from the information.

This paper will show how an analytical data mart for predictive modeling is created using modern software architecture and a market leading data quality toolset. The process of data gathering, data aggregation, data enrichment and joining data from various entities to deliver a final one-row-per-subject data mart will be demonstrated, accompanied by the relevant techniques of data quality control in order to achieve a trusted basis for statistical analysis and data mining .

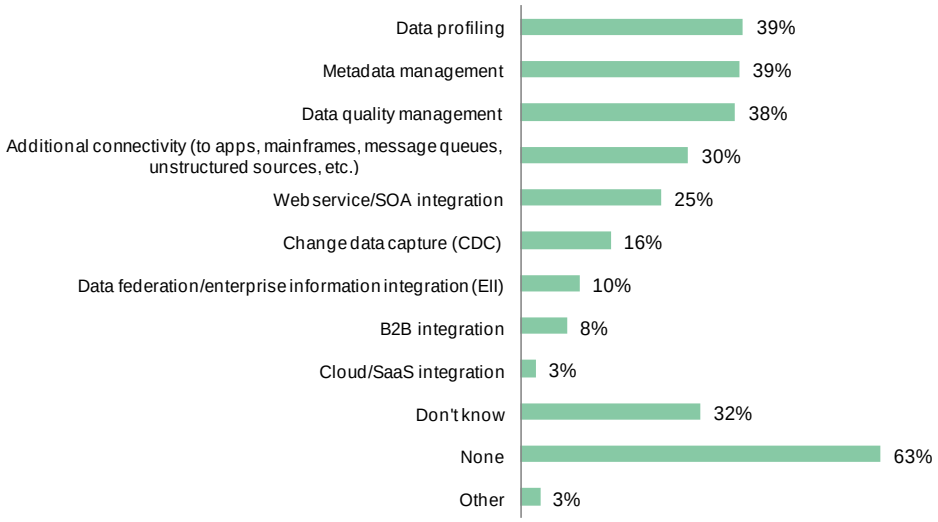
This paper is applicable to data integration, data or business analysts, and investigates the tools and technologies associated with data quality and how these tools can be applied correctly to create a foundation for accurate and trusted analytical solutions.

The paper will include practical demonstrations through the use of SAS data quality and analytical techniques.

INTRODUCTION

In a global survey conducted by Forrester in November, 2009, on the usage of data integration technologies, the results showed that nearly 40% of those surveyed had purchased a data profiling tool, metadata management tool, or data quality technology from their ETL vendor. (See the figure below.)

“Have you purchased any complimentary data management solutions through your ETL vendor?”



Source: Forrester survey November, 2009, Global ETL Online Survey,” Trends In Enterprise ETL Usage and Adoption.”

All of the technologies, data profiling, metadata management, and data quality management tools are fundamental to providing a trusted data platform for delivering an analytics or business intelligence initiative. The above survey results suggest that data integration and data quality are intrinsically linked. Delivering a business intelligence or analytics solution without considering data quality leaves you prone to issues with data definition, data content, data preparation, and information presentation, which can cause the analytics or business intelligence initiative to fail.

CAUSES OF DATA QUALITY PROBLEMS

Problems that hamper effective usage of analytics stem from many sources. Data might not be clearly defined, causing a mismatch in definition and in the facts collected. Data can be captured inaccurately, or samples in record selection can be biased. Time impacts data. A change in the real world might not be reflected in the current data. For example, if the value of a client’s account status changes significantly, then this new value needs to be reflected in the analysis of whether the client is eligible for a loan.

ANALYZING THE SEVEN INTEGRATION STEPS

If we think of data as measurements that we want to analyze there are several steps that we undertake before we can analyze the data:

1. Data capture
2. Data delivery
3. Staging
4. Integration
5. Filtering
6. Data transformation or analytical data preparation
7. Analysis

DATA CAPTURE

If data is considered measurements, then we can conclude that in the majority of cases the data in our organizations, although valuable for analysis, is primarily captured for the day to day execution of the business, and is transactional data in nature. The data is collected for the sole purpose of core business execution such as making a product, selling a product, delivering a product, and billing customers.

There are several areas we have to consider when looking at the data that we are capturing.

AN ACCURATE AND AGREED UPON DEFINITION OF THE MEASUREMENT THAT WE ARE CAPTURING

We can opportunistically use the existing data and derive value from it in a secondary usage. Ideally, we work with the business subject matter experts to develop a consensus that identifies a standard for values that we want to capture.

It is important to define an accurate, agreed upon value that we are measuring and want to capture. For example, let us assume that we are analyzing customer churn for a broadband network supplier. One of the areas that influences churn might be customer satisfaction. Customer satisfaction is influenced by many areas such as speed, downtime, and so forth. One of the key areas might be the ability of the organization to set up the client's connection efficiently, order a connection, install a router, and configure the system in the customer's premises. One of the data values we want to capture is delivery date. Delivery date might be captured in many systems but the data value will be different. It is important to identify and agree the correct measurement we want to use. If we capture delivery date from the sales system, it might record the date and time for expected delivery of the service, or it might show when the previous supplier's contract ended and the new contract began. The warehousing system might record the delivery date as the time the router was picked off the shelf, when the router was delivered to the loading door, or when the router was picked up by a 3rd party courier. The courier probably records the delivery date and time indicating when the router was signed for at the client's premises. The provisioning team might record when the router was installed and set up on the client site. The account team might record delivery date as the date that the customer made their first payment for the service. All of these parts of the business might be capturing delivery date independently. The data values might vary tremendously and it is important to agree on which measurement is required for analysis.

CAPTURING THE DATA ACCURATELY

Once we have defined and agreed the measurement, we need to ensure that it is captured accurately either manually or in an automated fashion. If the capture method is automated, it is important to ensure that the machine is appropriately situated, calibrated correctly, and taking measurements at the correct time.

In a recent project I was working on, a client had invested heavily in installing wind speed equipment on multiple locations in the United Kingdom, and was selling this information to sailors and flying enthusiasts. The data could be accessed in real-time or trended over a day, week, or months. After a period of time, however, the data showed inconsistencies for one measurement site. The site was showing lower wind speeds than other wind indicators in the area, especially when the wind was in a westerly direction. An on-site investigation identified that the local farmer had erected a 20-foot high barn within three feet of the measuring station, thus rendering the wind speed figures useless.

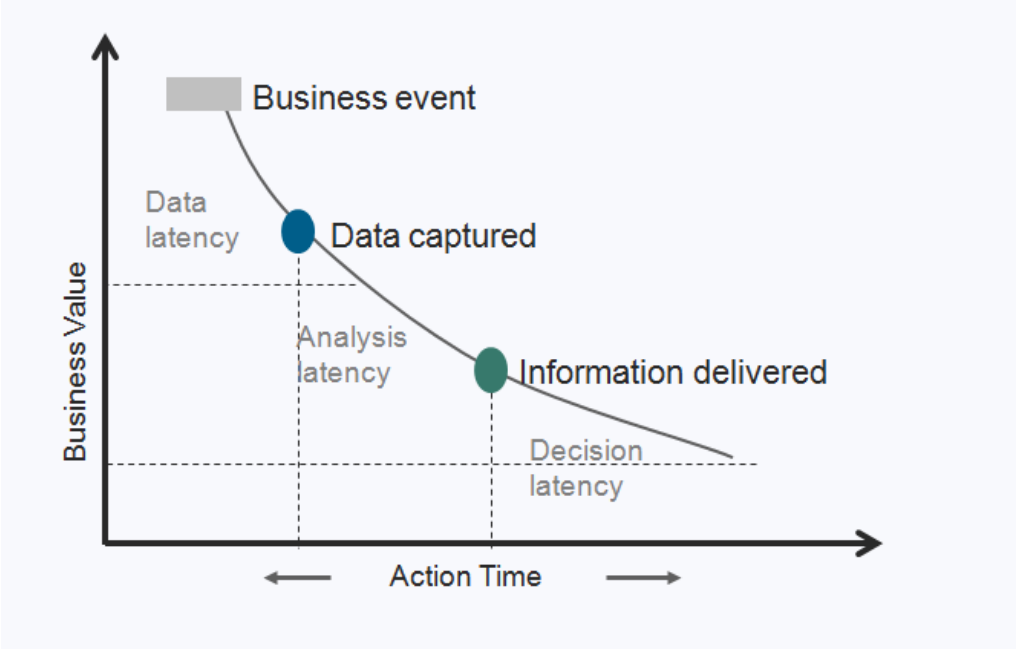
When capturing data manually it is important that the person recording the data not only understands the process of capture but also the importance of accuracy. Misspellings, inaccurate input, and misused fields, are a common problem.

A legal company specializing in auto insurance claims was analyzing the types of claims that it was litigating for a large insurance company. The system captured information entered by the legal clerks in their interviews with the injured parties. The initial analysis surprisingly showed a very large number of claims for accelerated arthritis, when in fact it was well known in the minor injury department that whiplash injuries were the main cause of injury and claim. Investigations quickly identified that legal clerks, when interviewing clients, were taking the default injury selection on the alphabetical list of their input screen to expedite form entry, rather than correctly scrolling through the list and identifying the injury in the list.

DATA LATENCY

It is important to capture changes to the source data in an efficient timescale. Ideally, when a change occurs in the source data, the change needs to be captured. There are two implications for not capturing the change quickly. One implication is that you lose the value because it might be overwritten. The other implication is that the data item can lose its value over time. The longer it takes for the data item to become information to someone for the purpose of making a decision, the less likely that information might be of value to the business.

The following figure shows that the value of data decays over time. Therefore, it is important to capture any changes to the source system within a valid time frame.



From an analytical perspective, the latency of the information might not be so important. But it is important if you are trending data over time to identify meaningful patterns. It is important that you use data of a common time period, understand the currency of all data required for a given model, and assure that the selection fits the age requirements.

SAMPLING AT CAPTURE

If sampling is required, then data must be representative of the population being studied. If there is undiscovered sampling bias and the population is not proportionally represented, the discovered trends will not be representative.

If sampling is made at information collection, assure that items or objects are selected by statistical sampling techniques so that each object has an equal likelihood of being selected. If data is being sampled, the data set must have representative samples for the real world collection of objects or the events it represents. Record samples must be made using the same statistical sampling. If there are different strata in the population, such as different classifications of customers, you need to assure a proportionate representation of each stratum. In some cases of analysis, however, you might want to develop a biased sample that includes a higher ranking of outliers in order to predict rare events, such as fraudulent transactions.

DATA DELIVERY

It has been over forty years since the concept of processing data for analysis from the source system was first implemented by using a data warehouse. We now have multiple sources of information, multiple warehouses, and multiple marts. Data must be captured and then delivered to a central system for analytical data preparation. There is an inherent risk of loss of data quality as part of the delivery process. Network connections can fail, records can be rejected, and data can be corrupted.

Most data transport technologies will have a level of functionality ensuring safe delivery of data over the network by using checksum calculations, data duplication, and recovery mechanisms. It is important that there be a level of auditing in your integration processes. Auditing can ensure that if 100,000 customers were extracted from the source system, maybe 10,000 were rejected due to validation rules, and 90,000 were loaded into the target system with no data loss, that the numbers add up, for example, Extracted = Rejected + Loaded.

Most delivery errors can occur when a delivery mechanism fails part way through. Data duplication is the usual problem when the system is restarted without addressing the partial load in the target system. If an automated mechanism such as an ETL batch job is used for data delivery, then you need to ensure that the data integration process adequately manages system restarts and partial data delivery occurrences so that data is not corrupted.

STAGING

Staging data has been necessary ever since the introduction of multiple data source feeds. Staging provides a storage space to hold data before processing and loading into the target analytical environment. One risk for data quality in the staging environment is data type mismatching. If you extract data from a source system and stage data to a different data storage platform, you need to ensure that the data is not transformed or truncated inadvertently

due to data type translation. Using the correct date formats and decimal formats are important. It is also important to ensure that text fields are not truncated.

There is also a risk of concurrency, which is the timing difference or equivalence of data in one data silo matched to another data silo based on movement of data from one store to another. Data that is extracted from different data sources might reach a given data set at different times. Sales for Today, for example, might not be in the staging area until tomorrow, because the data is batch loaded nightly. Deliveries are loaded weekly rather than daily, and returns are extracted on a monthly basis. It is important that batch extract schedules from the data sources assure that transactions represent events or objects at a single point in time or time period, and ensure that appropriate date and time stamps exist to enable the correlation of records across systems.

INTEGRATION

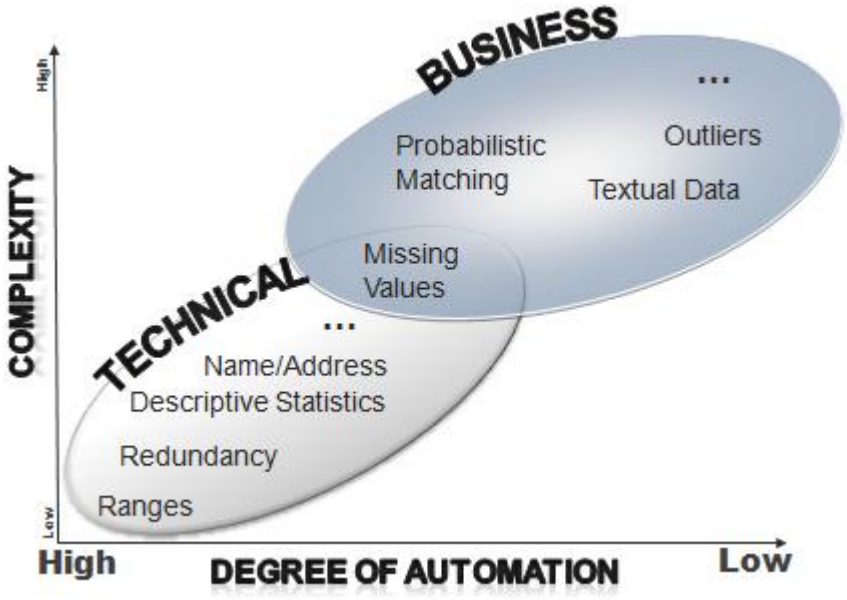
Data Integration involves combining data that resides in different sources and providing users with a unified view of this data. The difficulty and risk of combining data occurs when there is no common key to combine the data. It is highly likely that data that has been generated within a single silo will not have a common key with another silo of information. This makes it difficult to join data across silos, and you have to rely on textual data matching for names, addresses, descriptions, and so forth. This is where the data quality tool is invaluable because of its ability to use fuzzy matching, probabilistic matching techniques, and algorithms to join data.

FILTERING

Filtering involves the selection and validation of data to ensure that we have a data set that we can analyze and trust. There is a level of selection when we capture the data, but once the data is in the operational data store or staging area, there is a process of filtering that removes those records in which the data is invalid or cannot be repaired.

A data quality 'firewall' can be defined that checks all values against a set of rules and can either reject, repair, or pass the record, and provide alerts as part of the process. This data quality firewall that filters the data can be automated to a degree but requires regular monitoring. As the complexity moves away from the technical and more towards the business aspect, we can have missing values and outliers. Probabilistic matching, and then more intervention or review by the business on the data, might be required to confirm the automated filtering decisions.

As the following figure shows, the complexity moves away from the technical aspect towards the business aspect.



The most common tests for filtering data are listed below:

Ranges can be easily tested in a standard data validation node. The ranges are set, and if the value falls outside the valid range the record is rejected and logged. The data validation transform needs to record the reason the record failed and provide an alert and dashboard to show the trend of failed records.

Redundancy is part of the integration process. Redundancy should have been identified and any duplicate records or fields removed from the data set.

Individual and Organization Names, Addresses, and Phone data can be verified using an enterprise data quality tool that uses reference data to ensure consistency and accuracy in names, and also verifies the accuracy of address information. If required, it can enhance the data with geocoding information.

Missing values can result from data collection errors, incomplete customer responses, actual system and measurement failures, or from a revision of the data collection scope over time, such as tracking new variables that were not included in the previous data collection schema. If an observation contains a missing value, then by default that observation is not used for modelling methods such as a neural network or linear regression. However, there is a risk of rejecting all incomplete observations because useful or important information that is still contained in the missing variables might be ignored. Rejecting the record might also bias the sample, as records with missing values might have other things in common as well.

There is no single solution to missing values. You can obviously reject all missing data or repair the data value. Ideally, you can go back to the original source to recapture this information. You might also be able to access and look up the data against a secondary system that might contain the missing information, if the value is a simple piece of data such as age or date of birth. If the distribution of data values follows a normal population response, you might be able to estimate the data values with a mean of the variable, but this might affect the sample distribution. Another technique to consider is to replace the missing value with the mean of all of the values from the data source. This assumes that the input from the specific data source conforms to a normal distribution.

It is important to not only validate each individual record for missing values, but also to validate the complete data set. If a few records exist with missing values, these might be rejected or repaired. Beyond a certain threshold, the validity of the whole data set is invalid.

A statistician's view of an outlier is an observation that is numerically distant from the rest of the data (Wikipedia). Grubbs , in "Procedures for Detecting Outlying Observations in Samples" (F. E. Grubbs 1969), defined an outlier as 'one that appears to deviate markedly from other members of the sample in which it occurs.'

Outliers can occur by chance in any distribution, but they are often indicative either of measurement error, or a population that has a heavy tailed distribution. There is no rigid mathematical definition of what constitutes an outlier. Determining whether an observation is an outlier is ultimately a subjective exercise. Outlier detection can be automated to remove anomalous observations from the data. Most integration tools' data validation methods on outliers assume a normal distribution and identify observations deemed unlikely based on the mean and standard deviation.

Chauvenet's criterion could be used in a data validation node. The criterion uses the mean and standard deviation of the input data set values. Based on how much the suspect data differs from the mean, the transform could use the normal distribution function to determine the probability that a given record value will be at the value of the suspect record value. If you multiply this probability by the number of records in the data set, and if the result is less that 0.5, the record value is suspicious and could be invalid and rejected.

Rejecting outliers automatically based on an algorithm above has its risks. Many scientists and mathematicians think outlier rejection is inappropriate unless there is a good understanding of the underlying model and of the process being measured. Outlier rejection is also inappropriate if the usual distribution is known to confidently identify outlier records. Outliers can be rejected to remove anomalous observations from data, but they can also identify system faults and fraud before they escalate. It might be more appropriate to flag potential outliers for manual review and pass the data, rather than reject the record.

ANALYTICAL DATA PREPARATION

The final step is analytical data preparation. This is the mapping of data from operational sources to a more optimal format for analysis with the ability to transform existing values and derive new values.

The transformation of existing values from a data quality perspective could mean that the data values are standardized. It could also mean that categorical data is mapped to a numeric value so that the data can be easily processed in a mining tool. Most modelling techniques have difficulty processing alphabetic codes, and it is therefore important to transform this data into ordered numeric values that can be interpreted for correlation.

A usual requirement for part of the data preparation process is to enhance the data. Having the latitude and longitude coordinates for an address might be required, and identifying gender based on a person's name might also be required. Supplementing existing data with external sources such as interest rate fluctuations, weather information, or census information, might also be required. It is important that you take the same level of data quality rigor with external feeds as with your internal feeds.

THE APPLICATION OF DATA INTEGRATION AND DATA QUALITY TECHNOLOGIES

Data integration and data quality technologies can be applied to the seven integration steps. SAS provides a comprehensive suite of tools that address all aspects of data integration and data quality to provide a foundation for a trusted data platform. The seven integration steps are listed below:

1. Data capture
2. Data delivery
3. Staging
4. Integration
5. Filtering
6. Data transformation or analytical data preparation
7. Analysis

SOURCE SYSTEM EXPLORATION AND ANALYSIS WITH SAS TECHNOLOGY

The seven steps listed above provide a high level description of the technical steps needed to move data from the source system to a target system for analysis. The steps do not include the initial phases of an analytical project, that is, defining an analysis process and then selecting the data sources. The selection of the data sources will be dependent on the availability of the data, whether the data fits the analytical requirement, and on an assessment of the quality of the data.

DATA EXPLORATION AND DATA PROFILING

SAS® Enterprise Data Integration provides enterprise data integration and data quality functionality. It includes a product called DatFlux® dfPower Explorer. This technology provides the ability to analyze your metadata within your existing enterprise, providing a foundation to identify appropriate data for analysis. In addition, the tool is linked to data profiling. After you identify your candidate source data you can set tasks to profile the data for the purpose of identifying potential data quality issues before moving the data from the source system.

After we identify where the data that we want to use within our analytics platform resides, we can use dfPower Profile to complete the following tasks:

- Statistical analysis: Frequencies, ranges, outliers, numeric range analysis
- Data validation: Data patterns, data formats
- Pattern analysis: Redundant data, duplicate information, variant spellings, duplicate spellings
- Relationship discovery: Primary or foreign key relationships, cross-table relationships, cross-database relationships
- Business rule validation: Test the data against an organizational standard for data quality and for business processes and practices.

With data profiling we can identify data quality issues with the candidate data and define data quality rules to filter out or correct the data as required.

DATA CAPTURE WITH SAS TECHNOLOGY

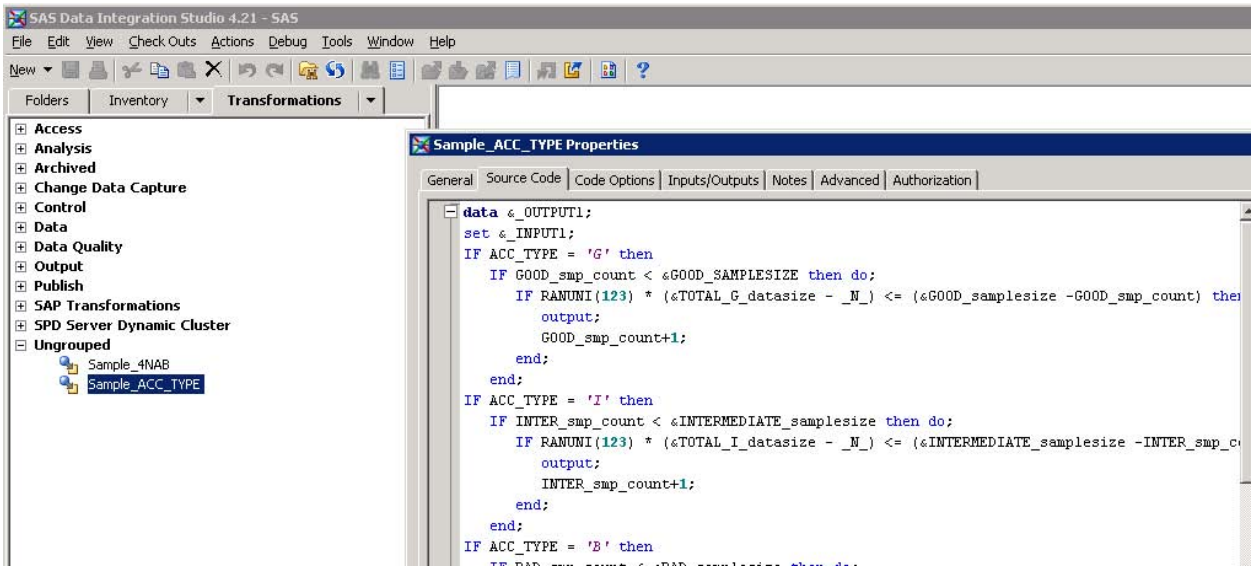
SAS® Data Integration Studio provides comprehensive capabilities to capture data in batch mode as well as real-time. The data could be provided by a Web service or a message queue. In the latest release of the product, there are new nodes that take data feeds from change data capture technology, either provided by the database or from a third party vendor, such as Attunity. The benefit of change data capture technologies is that they are low latency, low impact methods for capturing data.

In a recent project, a large telecommunications company used SAS Data Integration Studio linked to a change data capture technology to capture mobile text message information for customer data analysis. The main concern for the telecommunications company was that although they wanted to complete analysis of the data, they could not impact the operational system which was creating in excess of 10,000 new records per second. Using SAS Data Integration Studio to design a capture process in conjunction with a change data capture technology, provided the solution with zero impact on the operational system through the processing of archive data logs.

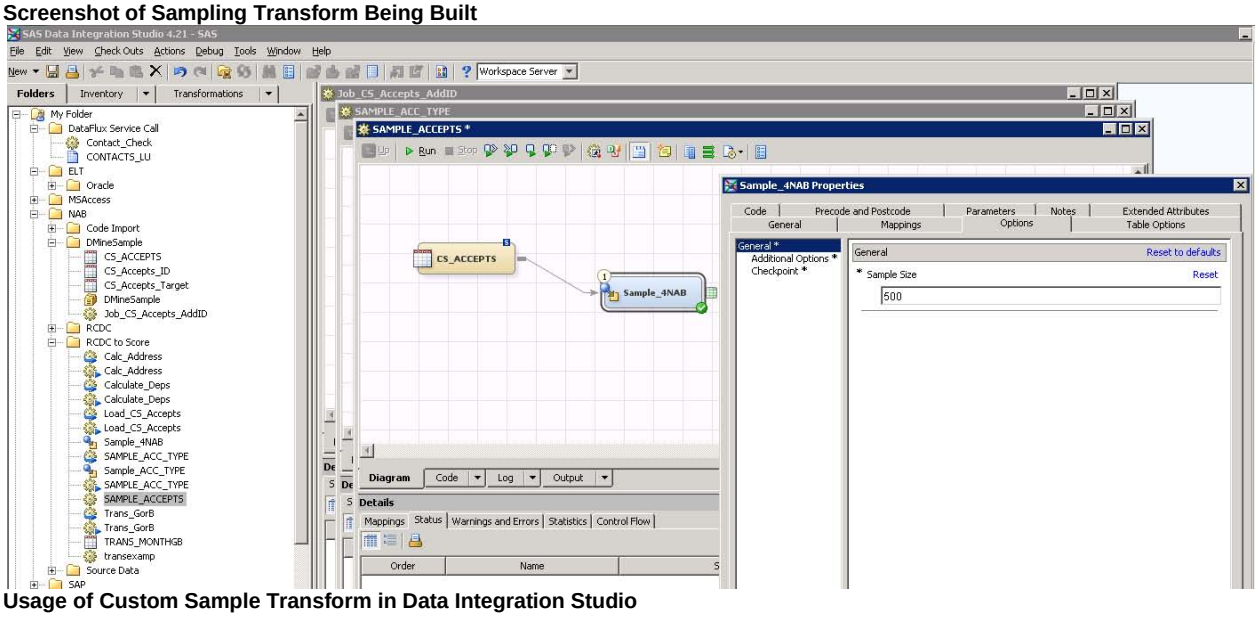
DATA SAMPLING

Data sampling within the capture part of the data integration process is something you do not normally see provided by the majority of ETL technologies. If it is provided, the functionality is not nearly as comprehensive as SAS Data Integration Studio. Sampling can be achieved very easily within SAS Data Integration Studio because it is underpinned by the power of the SAS language. This ensures that the sampling delivered provides an equal likelihood of the record being selected. If there are different strata in the population, a proportionate representation of each stratum is provided to ensure the quality of the sample data. Within SAS Data Integration Studio, it takes only a matter of minutes to define a new transform that provides comprehensive sampling capabilities. The transform then becomes a reusable graphical object that can be shared among users and across projects.

The following figures show examples of the windows in SAS Data Integration Studio:



Screenshot of Sampling Transform Being Built



Usage of Custom Sample Transform in Data Integration Studio

STAGING DATA WITH SAS TECHNOLOGY

The loading of data into an Operational Data Store or staging area can be achieved efficiently through the use of native support of the database bulk loaders. Data type mismatching can be avoided through checking the mapping of SAS data formats to the native data types of the target database. SAS Data Integration Studio, through the use of the underlying SAS® Access technology, automatically applies the correct data type if creating a new target table. If there is a data type mismatch then a warning or alert will be visible with SAS Data Integration Studio. This functionality reduces the risk of truncating textual strings or losing precision in numerical data.

INTEGRATING DATA WITH SAS TECHNOLOGY

Data integration in this context is the joining of data. As identified earlier, integration can prove difficult when there is no obvious key to link the data and we have to rely on textual fields for the data links. For example, an issue arises about whether Mr Jon Smith of SAS Institute in one system is the same as Mr John Smyth of SAS Institute in another system. SAS provides a solution to this dilemma with the market leading data quality solution from DataFlux®, which is incorporated in the SAS Enterprise Data Integration product suite.

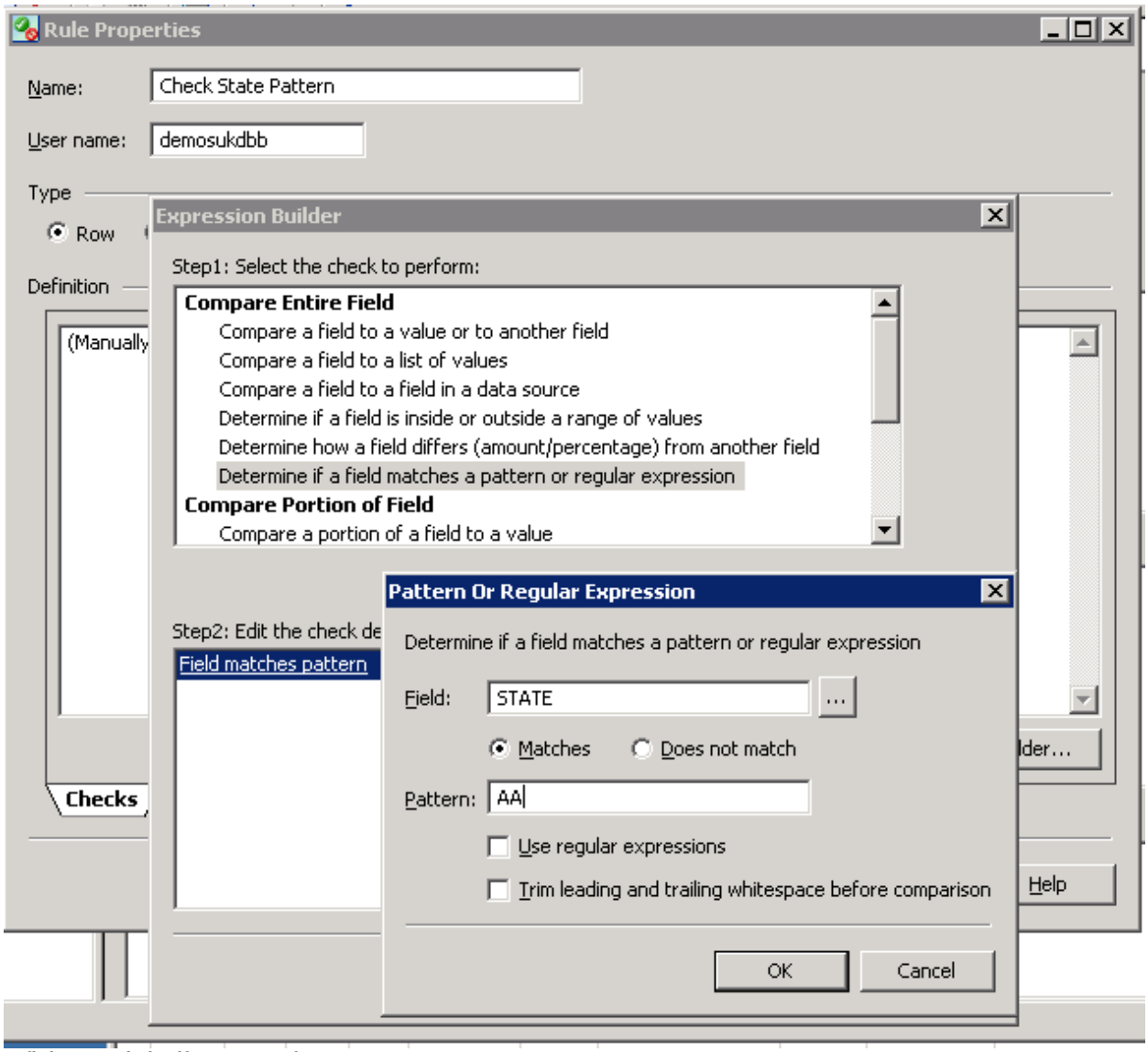
The matching capabilities in dfPower Quality are extremely comprehensive. The data can be parsed to identify the tokens for the data type (name, address, account), standardized, and then matched (See the figure below). A match code can be defined for a specific sensitivity and this match code is then used to join or not join the data.

Field	Record 1	Record 2	Record 3
Name	Robert Smith	Bob Smith	Rob Smith
Address	100 Main St	100 Main	100 Main St.
City	Phoenix	Phoenix	Raleigh
Match Code	**GHWS$$EWT$**	**GHWS$$EWT$**	GHWS$$WWI$

FILTERING

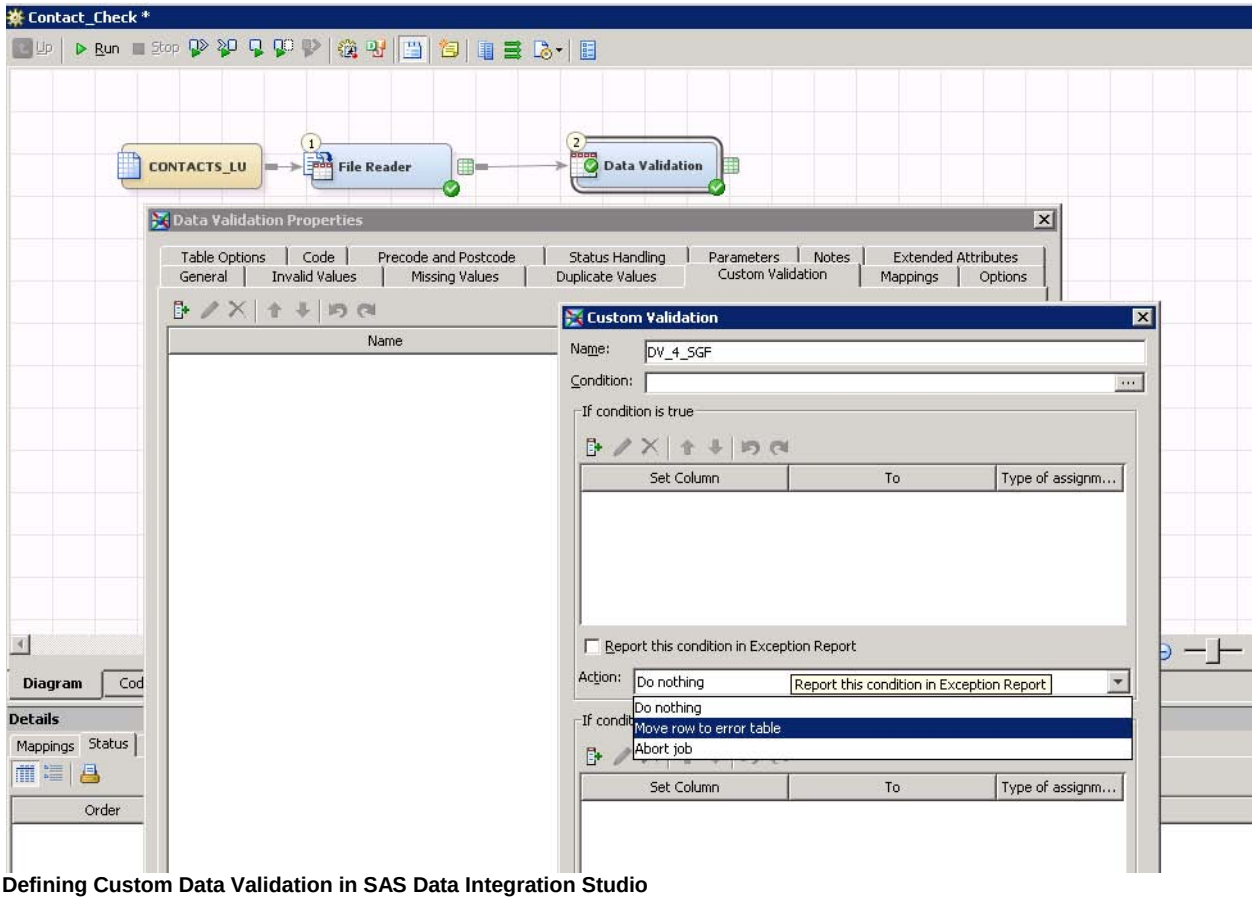
Multiple layers of filtering are provided by the SAS technology. Defining filters during data capture is an option if you use simple WHERE clauses for the selection of data. More comprehensive is the data validation transforms provided within SAS Data Integration Studio, where comprehensive tests for data validity can be defined.

Data validation can further be enhanced with the use of dfPower Monitor which is part of the DataFlux technologies, incorporated with SAS Enterprise Data Integration. dfPower Monitor enables you to define data quality rules in a central repository. These rules can be allocated to a task. Tasks are then used by a data quality job that can be exposed as a service, run as a scheduled batch job, or incorporated within a SAS Data Integration job. The data is checked against the data quality rules. If the record fails, the reason for failure can be recorded, the data can be viewed by means of an interface, and alerts can be triggered. The benefit of this approach is that all validation rules are centralized for maintenance and re-use. The following figure shows several windows in dfPower Monitor:



Defining a Rule in dfPower Monitor

The SAS Data Integration Studio technology provides the benefit that within the data validation transformation you can define data validation rules that can incorporate specific SAS analytical functions, which are unavailable in almost any other data integration technology. This can be combined with the data quality rules capability of DataFlux technology using the data quality rules and repository to define and monitor data quality in batch or real-time. The following figure shows several windows in SAS Data Integration Studio:

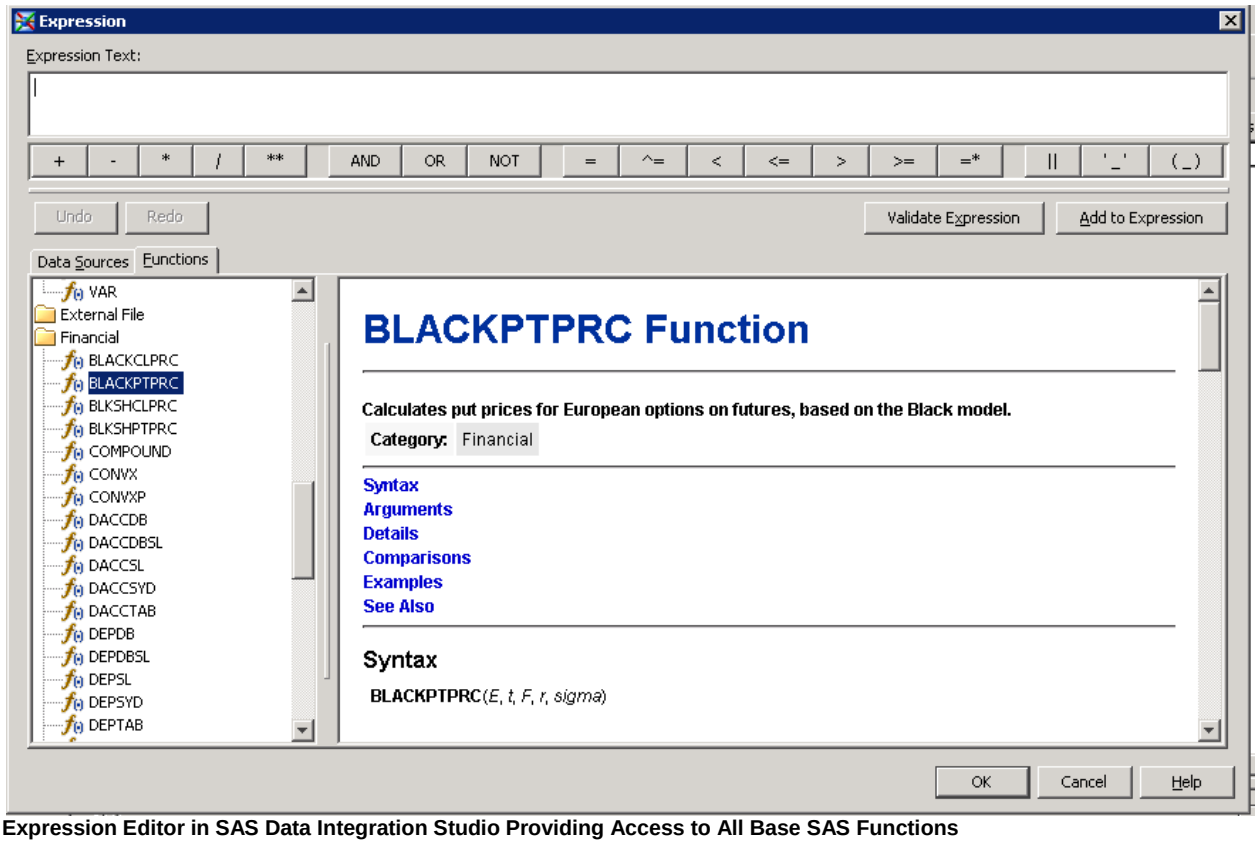


Defining Custom Data Validation in SAS Data Integration Studio

DATA TRANSFORMATION OR ANALYTICAL DATA PREPARATION WITH SAS

The power of SAS Data Integration Studio is that it provides a graphical environment to access all of the capabilities of the SAS language. There is no other data integration tool that provides such a comprehensive library of transformation functions within the data integration tool without calling out to external libraries or resources to transform the data. There are data quality functions to standardize data and create match codes as well as to call data quality services from DataFlux. You also have access to all of the financial and statistical functions available with Base SAS. The benefit of having this library of functions is the incorporation of these functions within the data validation rules, using filtering as part of your data integration process, and transforming and preparing data for analysis.

The following figure shows the Expression editor in SAS Data Integration Studio:



Expression Editor in SAS Data Integration Studio Providing Access to All Base SAS Functions

CONCLUSION

Ensuring that you have the right level of data quality for analytics and business intelligence is not just a matter of applying data validation and data quality services to your data integration processes. There must be a clear definition of what data is to be analyzed. Once the data to be analyzed has been defined, the data integration and data quality processes can be defined and developed. This will involve a definition of where the data is to be sourced, the interval for capturing the data, where to deliver the data, the data validation rules, integration rules, and ultimately how the data is to be prepared and loaded and ready for analysis.

The vast majority of the integration process can be automated with data integration and data quality tools, unlike other integration projects. The level of business involvement with the validation of the data to ensure it is 'fit for purpose' in an analytical data preparation project is higher than in most data integration projects. You are much more likely to flag data for review, or alert business users in the data quality or validation step, than in other projects because of the validation rules.

The SAS suite of data integration and data quality products provides comprehensive capabilities to automate the data integration process and reduce the burden as much as possible in providing a trusted data platform for analytics. SAS provides the technology to access and capture all enterprise data in batch and real-time. Having access to the comprehensive library of SAS functions in SAS Data Integration Studio provides unmatched capabilities when defining data validation and integration rules as well as for preparing the data for analytics. Combining the SAS Data Integration technology with the market leading data quality technology from DataFlux, provides the ultimate technical solution to address any enterprise's data integration and data quality requirements for analytics. It is just a matter of ensuring that the correct processes and implementation methods are used to ensure that the business and technology organizations can implement a solution that can identify the data and prepare the data to be loaded into the tuned model.

REFERENCES

- Chauvenet : Chauvenet's Criterion: http://en.wikipedia.org/wiki/Chauvenet's_criterion
- Forrester Research : 'Trends In Enterprise ETL Usage And Adoption', Robe Karel March 2010
- Grubbs: 'Procedures for Detecting Outlying Observations in Samples', F. E. Grubbs 1969

CONTACT INFORMATION

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

Name: David Barkaway
Enterprise: SAS Institute Inc

Address:
City, State ZIP: United Kingdom
Work Phone: +44 1628 4-86933
E-mail: David.Barkaway@suk.sas.com
Web: support.sas.com