Paper 043-2008

ETL and Data Quality: Which Comes First?

Emilio Power, ThotWave Technolgies, LLC, Chapel Hill, NC
Greg Nelson, ThotWave Technolgies, LLC, Chapel Hill, NC

ABSTRACT

Usually, an early task in any data warehousing project is a detailed examination of the source systems, including an audit of data quality. Data quality issues could include inconsistent data representation, missing data and difficulty around understanding relationships between the various source systems.

As ETL and Data Quality technologies converge, it's important to use the right tools at the right time to fully take advantage of the strength of each individual tool. Within SAS ® Data Integration Server, there are several opportunities to address data quality issues – this paper will go over the development of an ETL process with an emphasis on data quality. A course of action will be established with suggested roles of stakeholders that can have an input on the ETL process outside of the direct development team. The paper will then cover the topics of discovering data issues, how to address them and which tool to use to achieve the ultimate goal of having a clean output suitable for a data warehouse.

INTRODUCTION

Typically, whenever we build a data warehouse we have certain expectations about the level of data quality we expect in the data. We know that most data warehouses have lots of raw data source systems and we have an idea of our ideal data structure for the warehouse, but what is often overlooked is a central foundation for data quality. In a data warehouse fact tables are very sensitive to data 'impurities' in the sense that foreign key relations to dimensions and levels of granularity can be seriously impacted. This has a cascading effect to the extent that these errors will degrade the functionality and level of value provided by the warehouse. Thus developing a plan for building in data quality right from the start becomes extremely important. In a previous paper (Nelson, 2007), we outlined the process for planning for and designing a data warehouse. In that papers, we suggested that you first identify your business goals, then actually verify the contents of your source systems through data profiling and then, and only then, continue with your ETL development (assuming you have properly designed your data model and the source to target mappings).

In this paper we will use a fictitious case study based on building a data warehouse in higher education. We will go through the major steps of developing the ETL and planning for the data. At the end we will discuss the sources of error, data quality methods and the implications of decisions made around these techniques.

BUSINESS GOALS

In the development of a data warehouse, the identification of business goals to develop a model will often directly or indirectly identify data quality issues. Stakeholders involved in all levels of the current business processes have an opportunity to bring attention to potential data quality issues before actually looking at any of the source systems. Some common goals and questions:

1. The data warehouse should allow easy tracking of entities* across transactions.
2. Develop a single identification system for entities
3. Identification of a "person" which often introduces the requirement house-holding or grouping to a logical entity
4. Track history of an entity

* Examples of an entity are a person, product, individual sample, inventory item.

BUSINESS PROBLEM

So first, let's describe our business problem. In our case study, we have an educational institution that would like a data warehouse built that will allow them to track information about their students over time. The institution has been trying to aggregate records from various systems which were created for different programs to deal with students from high school to graduate school. In the past, the various systems were treated independently but these valuable information sources need to be brought together in order to better understand student patterns and history as they moved from one program to another. By better understanding students the institution can do a better job of serving students and understanding the "person" versus just records in separate systems.

ANALYSIS

There are always a number of ETL issues that arise when we bring data from multiple systems on multiple platforms. One of the most pressing of these issues is how to standardize (conform) data elements that may or may not have been recognized as the same in the separate source system. For example, a person in one program may or may not be the same person in another system. Hopefully, we have a good key like social security number to tie them, but as you might suspect, this is seldom the case when we are dealing with this type of disparate data. The problem is the lack of a standard, centralized method for identifying students across all systems. The solution to the student identification problem will have to be able to discern students using a range of fields such as student identification numbers, addresses, dates and city of birth – all to get to a single, consistent version of the truth about who a person is and what interactions we may have had with that person over time.

DATA PROFILING

TACKLING DATA QUALITY WHEN THERE IS NO DATA

Identifying data sources and their associated origins allows developers to predict specific sources of error and emphasize certain areas that should have greater scrutiny. Sources such as in store registers, user-filled surveys and customer driven data collection can be potential focus points for predicting what type of data issues will be present. This will assist the developer when creating a data quality management plan for the ETL process. Communication, including interviews with stakeholders involved in the process of data collection, allows for a level of profiling without ever seeing data. Such a conversation might include taking time to speak with a clerk involved in the process to solicit an honest level of communication. This can lead to identification of data quality issues such as when data entry is done with bogus numbers such as a client's phone number in order to get a transaction completed. Overloaded fields due to a static process are also areas where identification of data quality issues persist and will need attention but may not be clearly visible even with data profiling tools.

BUSINESS PROBLEM

In our case study, we interviewed various stakeholders that were responsible for the operational systems in which the data was originally collected, including managers and users of their central enrollment system. In these interviews it became known that different programs have different systems and that while they all have identification numbers that share the same logical name, the data was not shared across systems and thus not useful for matching students across the systems. In addition when a student needed to update their mailing or billing address, the record is simply over written manually by a phone operator. It was also identified that for a period of time, when a piece of mail was returned, the intended recipient's address was changed to the school's mailing address in hopes of making a list to identify what students had a bad address.

ANALYSIS

There are three data quality issues that were identified before we looked at any data. The first was that an identification number may not have a 1 to 1 relationship. In other words a single person has multiple identification numbers across systems. The second was the need to standardize addresses for individual students. These items would probably be found through standard data profiling but the third piece of having a non-standardized entry that identifies issues around a bad address may have been missed. Having data with multiple meanings in a single column are referred to as overloaded fields which is much more insidious because the data is technically correct, but logically, it means something very different than what we expect.

Identified Requirements

- Clustering of Records
- Lack of Standardization
- Overloaded fields

TACKLING DATA QUALITY WHEN THERE IS DATA

After some work has been completed to understand how people "think" the data is organized, the next step is to actually get your hands dirty by digging into the data. SAS Data Integration Studio ® offers a couple of tools to productively accomplish this task. Using the information gathered in interviews with process owners or stakeholders you can develop target profile jobs looking for previously identified problems. This is where it is beneficial to involve someone with a strong familiarity of the business process to help identify issues that were not identified in the previous steps.

One of the tools in *SASData Integration Solution* available for profiling data can be found within *DataFlux® dfPower® Studio*. The *SAS® Quality Knowledge Base (QKB)* serves as the central repository for the *SAS Data Quality Solution*. *DataFlux® dfPower® Studio's* primary role (when bundled with the *SAS Data Integration Solution*) is its process for customizing the business rules found in the QKB that will be used for data cleansing. Another significant benefit from *DataFlux® dfPower® Studio* is the *Profile* tool which allows ETL developers to peer into data and asses the quality of its source data. Discovery of unknown problems and assessing known issues is facilitated with a series of metrics and tools. One of the strongest tools found in the *Profile* module is the ability to assess data patterns found in variables.

A second often over looked tool in the *Data Quality Solution* are the identification or "guessing" functions that can be found in both *DataFlux® dfPower® Studio* and *Data Quality Server*. These are great tools for identifying common overloaded fields in order to qualify data types. In *Data Quality Server* the DQIDENTIFY process performs this function by leveraging identification definitions found in the installed *Quality Knowledge Base*. One of the interesting predefined definitions in the QKB 2007a ENUSA Locale is the 'Contact Info' which allows DQIDENTIFY to determine a contact element such as street address, city street zip code, and organization name.

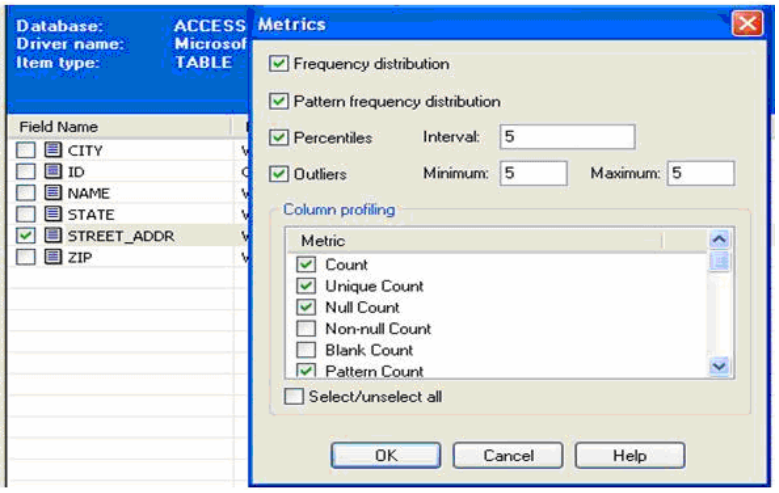
BUSINESS PROBLEM

So continuing with our case study, we know through the interviews and provided documentation, that there is a source data system for each program the student attends or participates in and the general design has to account for information available across all systems – even though those operational systems were not designed to really aggregate the data across systems. The educational institution has also been able to provide a mapping of the various systems and has given the developers access to the data.

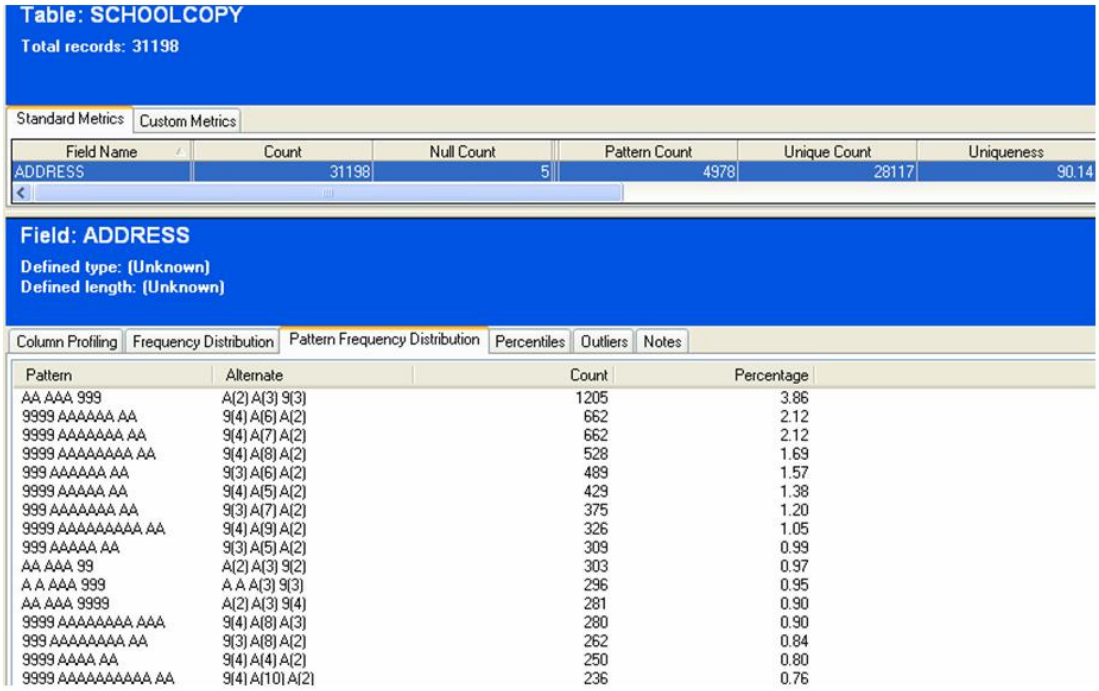
ANALYSIS

In this specific case the address variables are profiled through *DataFlux® dfPower® Studio* and *SAS ® Data Quality Solution*. Frequency distribution of patterns and values will probably be of the greatest interest along with count, unique count, null count, pattern count, maximum length, and percent null.

In the screenshot below, we make our selections to include gathering some basic diagnostic information about the field called STREET_ADDR.



The original suspicion of the need to standardize data is confirmed by reviewing the outcomes of the profiling methods as there are mixed versions of standard names such as ST and STREET and there are hints of various data types.



Using the Identity function we scanned the data set for various data types mixed into the address field. The following code produces a frequency count on a dataset using the DQIDENTIFY function to provide a result on what kind of data is found in the address data set.

```
%dqload(DQLOCALE=(ENUSA), DQSETUPLOC='C:\Program Files\SAS\SAS
9.1\dquality\sasmisc\dqsetup.txt', DQINFO=1);

proc sql;

create table work.addresstype as
    select distinct(dqIdentify(ADDRESS, 'Contact Info', 'ENUSA')) as Address_ID,
    count(*) as Typecount
    from address.schoolcopy
    group by Address_ID;

quit;
```

VIEWTABLE: Work.Addresstype

	Address_ID	Typecount
1	ADDR	26980
2	ADDR2	3
3	CSZ	18
4	IND	29
5	ORG	44
6	PHONE	2
7	POB	3997
8	UNK	125

These profiling results have identified various steps that need to be taken in the design of the ETL to address data quality. Data will not only need a level of standardization but there is a need to address mixed data types. The results show that the ETL will require two main steps:

1. Identify and isolate data types
2. Apply a standardization customized for each data type

DEVELOPMENT OF ETL

SAS ® DATA QUALITY SOLUTION

With the business information gathered and the source data profiled, an ETL design can now be developed. The seamless ability to leverage *SAS ® Data Quality Server* through *SAS Data Integration Studio* allows developers to use the data quality functions in mappings, SQL transformations, and procedures. The process of profiling decisions around how data will be standardized should be reflected in the *SAS Quality Knowledge Base Locale*, which will serve as the cornerstone to all data quality functions.

SAS ® Data Quality and *DI Studio* are the main tools in *SAS ® Data Integration solution* in which ETL development can directly address data quality and leverage specific functions and procedures. The most valuable functionality brought by *SAS ® Data Quality Solution* revolve around data parsing, match code generation and a clustering algorithm that uses straight matching and can accept multiple conditions. These core functions are what make it possible to easily handle various issues that can range from the creation of standardization schemes to solving householding questions.

BUSINESS PROBLEM

Our goals were identified in the previous steps which suggested that we need to standardize data and identify students across systems. Data standards have been set and the Quality Knowledge Base has been optimized to offer standardization schemes that match the definitions by data stewards. The educational institution would like to identify students through the following conditions: Name, City of Birth, Date of Birth, Last 4 digits of Social Security Number or Name, Address, Sex, and Date of Birth. Some systems have tables with historical information and they would like to leverage that data.

ANALYSIS

This is where many would consider, 'where the rubber meets the road', actually takes place. A Master Data Management approach will usually lean towards the control of data quality early in the process and ideally at the time of collection. This requires coordination across many systems, which in most data warehousing projects is an afterthought. The most common situation found in a data warehousing project with some sort of Data Management is that it will usually reside at a specific system level and will not be spread through-out the enterprise.

This case study follows the most common situation where data standards have been applied to different systems but they were done so independently. Without an enterprise wide effort for data management it is even likely that, in the future, data from the Data Warehouse may be brought into another system and the cycle will continue. This is why a Master Data Management approach helps minimize the difficulties encountered with data quality.

STANDARDIZATION

We have identified that addresses need to be standardized which can be done through the DQSTANDARDIZE function. The DQSTANDARDIZE function as all the other DQ functions is reliant on a QKB definition. The QKB should be constantly monitored to ensure maximum accuracy.

The following code is an example used in SAS ® Data Quality Solution to standardize data.

```
%dqload(DQLOCALE=(ENUSA), DQSETUPLOC='C:\Program Files\SAS\SAS
9.1\dquality\sasmisc\dqsetup.txt', DQINFO=1);

data Standardized_Addresses;
 set school.addresses;
 StandardizedAddress=DQSTANDARDIZE(Address, 'Address' , 'ENUSA') ;
run;
```

The following screenshot is of the standardized data created by the DQStandardize function.

	ADDRESS	StandardizedAddress
444	4861 ROSALIA DRIVE	4861 Rosalia Dr
445	2001 CONSTANCE STREET	2001 Constance St
446	300 COLONIAL CLUB DRIVE	300 Colonial Club Dr
447	2323 S. GALVEZ ST	2323 S Galvez St
448	1020 N PRIEUR STREET	1020 N Prieur St
449	2035 TOLEDANO STREET	2035 Toledano St
450	2437 JENA STREET	2437 Jena St
451	1415 TECHE STREET	1415 Teche St
452	315 CIVIC DRIVE	315 Civic Dr
453	215 BETZ PLACE	215 Betz Pl
454	75 E CHALMETTE CIRCLE	75 E Chalmette Cir
455	5701 VETERAN'S MEMORIAL BLVD.	5701 Veteran'S Memorial Blvd
456	726 JOHN HILL TAYLOR DR	726 John Hl Taylor Dr
457	1207 EAST BROADWAY	1207 E Broadway
458	170 E GAP HILL RD	170 E Gap Hl Rd
459	200 ENTERPRISE DRIVE	200 Enterprise Dr
460	RT 4 BOX 90	RR 4 PO Box 90
461	170 W A JENKINS ROAD	170 W A Jenkins Rd
462	357 WEST ARCH ST	357 W Arch St
463	10362 ST RTE 138	10362 St RR 138
464	MAIN CROSS ST BOX 69	Main Cross St PO Box 69

CLUSTERING

The next requirement was to match student records across systems. In previous ETL steps the data has been gathered to meet the requirements that the business identified for clustering records. Proc DQMatch in *SAS Data Quality* is a procedure that can tackle these two functions. The first is the creation of match codes and the second is clustering through a series of identical matching of conditions.

```
proc dqmatch
    data=ClusterData
    out=ClusteredData  cluster=Cluster_Code
 NO_CLUSTER_BLANKS NODELIMITER ;
 /* Criteria One */
    criteria condition=1 EXACT  var=DOB;
    criteria condition=1 matchdef="Name" var=Full_Name sensitivity=95
       matchcode=Match_Code1;
    criteria condition=1 EXACT  var=BIRTHCITY;
    criteria condition=1 EXACT  var=FOURDIGSSN;
 /* Criteria Two */
    criteria condition=2 matchdef="Name" var=Full_Name sensitivity=95
       matchcode=Match_Code1;
    criteria condition=2 EXACT  var=SEX;
    criteria condition=2 EXACT  var=ADDRESS;
    criteria condition=2 EXACT  var=DOB;
 /* Criteria Three */
    criteria condition=3 EXACT  var=SYSTEMID;
run;
```

SOURCES OF ERROR

The business case here offers a very interesting question on how the organization goes about identifying students. There is an apparent detail that needs attention which, if ignored, can cause serious implications for data quality in the data warehouse. The ETL development team has two ways to solve the issue around the identification of students. The first is to standardize data and then cluster records based on the standardized data. The second option is to cluster data based on raw records. Most people would quickly jump to the conclusion that you must first standardize data and then do the clustering. While this has its obvious arguments the question really comes down to a decision of what level of error you are willing to accept. Data Quality is not a fool proof way that takes incorrect data and corrects and reformats data.

In the business case we are addressing it is important to remember there are three main potential sources for error which include Data Entry or Collection Error, Data Standardization Error, and Clustering Logic Error. The decision behind a method has to be made based on the best process set forth that will minimize the sources of error while keeping in mind that processes themselves can impact the relative levels of risk. Every case should have an analysis to identify potential sources of error.

Identified Sources of Error

- Data Entry or Collection Error
- Data Standardization Error
- Clustering Logic Error

In this case we will be looking at the way that each source of error can impact the clustering results and give us a Type I or Type II Error (the probability of accepting or rejecting a match when we are, in fact ,either right or wrong).

DATA ENTRY

Data entry is the main source of variation in data and this can lead to both types of clustering errors. A student could have a “date of birth” mistyped on his second application which would lead to a clustering result of a Type I error. This would mean that both of his records were not united but should have been with correct data. The second error could occur where an identification number was mistyped matching a second student thus causing a Type II error of matching two records of two different people.

DATA STANDARDIZATION

The level of data standardization has a great impact on the source of error. In this case standardization is not a format change such as changing text entries to “proper case” it is where actual changes are made to the data such as PO 456 and PO Box 456 changed to P.O. BOX 456. By reducing noise in data you are increasing your chances of making a Type I error or False Positive when clustering. Constant updating of a QKB definition and a level of confidence in results are key in understanding the level of impact on clustering results. You must ask if the business is sensitive to having these types of mistakes. A Type II error due to standardization in the clustering of medical records could lead to serious implications.

LOGICAL ERROR

A logical error stems around having requirements that infer a relationship that is based on probability. It is unlikely that you find two students with the same name, date of birth, street address, mother’s maiden name, social security number and favorite ice cream flavor. This model however is also impractical that a balance is made to maximize efficiency by getting the greatest number of correct clusters with the minimal amount of data. This is probably the most variable source error in terms of Type I and Type II. In the straight matching algorithm that’s found in Proc DQMATCH the introduction of more requirements in a single condition reduces the Type I error yet it increases the Type II error. It is important to develop a strong model and make sound business decisions when deciding on clustering requirements.

RISK CONSIDERATION

In the final decision of developing data integration process it is important to consider the types of risks we introduce and the value gained. In the case of standardizing data before clustering begins, the business assumes that they are potentially increasing the risk for a Type I error. So while hopefully the actual percentage of errors is small the type of error made is important when making decisions on whether or not to include standardization. This will vary depending on the use of the warehouse and may often actually imply limits of use. For example, it would not be wise to do confidential mailings based on a system with a high Type I error percentage.

CONCLUSION

It is important in the development of a data warehouse ETL process to pay special attention to data quality as the nature of your results will impact directly the quality and value of the data warehouse. Business knowledge acquired from all levels of the business makes involvement of stake holders at every step of the process invaluable. Besides the technical workings of creating an ETL process, business knowledge identified including the underlying business model and goals will dictate important steps that the ETL in regards to data quality will require. The technical workings of ETL including the seamless use of *Data Quality* functions and procedures along with the attention to a business process makes development in SAS a true data integration process.

REFERENCES

Nelson, Gregory S. “Planning for and Designing a Data Warehouse: A Hands on Workshop”. Hand on Workshop SAS Global Forum 2007. Orlando, Florida.

SAS Institute Inc. 2004. SAS® 9.1.2 Data Quality Server: Reference. Cary, NC: SAS Institute Inc

ACKNOWLEDGMENTS

A big thanks goes out to everyone at ThotWave Technolgies, LLC.
A very special thank goes to Shawn Edney and Richard Phillips.

CONTACT INFORMATION

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

Emilio Power
Greg Nelson
ThotWave Technologies, LLC
510 Meadowmont Village Circle #192
Chapel Hill, NC 27517
(800) 584 2819
http://www.thotwave.com