Paper 161-28

The Value of ETL and Data Quality

Tho Nguyen, SAS Institute Inc. Cary, North Carolina

ABSTRACT

The adage “garbage in, garbage out” becomes an unfortunate reality when data quality is not addressed. This is the information age and we base our decisions on insights gained from data. If inaccurate data is entered without subsequent data quality checks, only inaccurate information will prevail. Bad data can affect businesses in varying degrees, ranging from simple misrepresentation of information to multimillion-dollar mistakes. In fact, numerous research and studies have concluded that data quality is the culprit cause of many failed data warehousing or Customer Relationship Management (CRM) projects. With the large price tags on these high-profile initiatives and the importance of accurate information to business intelligence, improving data quality has become a top management priority. By integrating ETL and data quality, a consistent, reliable view of your organization ensures that critical decisions are based only on accurate information. If you trust your data, you can trust your decisions.

INTRODUCTION

Data warehousing and CRM porjects have failed due to data quality issues. It is a documented fact that billions and billions of dollars are lost because of poor data quality. The cost accounts only for US business losses in postage, printing and staffing overhead. Frighteningly, the real cost of poor data quality is much higher. Beyond wasted resources, there are disgruntled customers, decreases in sales revenues, erosion of credibility and the inability to make sound business decisions. Sometimes the effects of bad data cause enough for complete business failure.

THE BUSINESS IMPACT

Organizations depend on data. Regardless of the industry, revenue size or the market it serves, every organization relies on its data to produce useful information for effective business decision making. Unfortunately, with so much emphasis on such information, data quality rarely gets the attention it deserves. No one wants to consciously admit that their business decisions are based on inaccurate or incomplete data. A recent survey revealed that 75% of organizations have no data quality processes in place. It appears that the majority of the businesses have taken no steps to determine the severity of data quality issues and its impact on the bottom line.

Companies invest hundreds of thousands of dollars and significantly large portions of information technology (IT) budgets on building sophisticated databases and data warehouses. In the quest for successful business intelligence, various applications and systems will be deployed and information gathering processes will be created. However, many overlook the essence that is the underlying data that matters. All of the fantastic screens and reports in the world will not make a positive difference if the data that supports the system is inconsistent, redundant and full of errors.

There are many reasons why the quality of the data that companies collect and analyze is so poor. The reasons vary everything from the very ambiguous nature of the data itself to the reliance on data entry perfection. But none is more compelling than the simple fact that companies rely on so many different data sources to obtain a holistic view of the business.

Information collection is increasing more than tenfold each year, with the Internet a major driver in this trend. As more and more data is collected, the reality of a multi-channel world that includes e-business, direct sales, call centers, and existing systems sets in. Bad data (i.e. inconsistent, incomplete, duplicate, or redundant data) is affecting companies at an alarming rate and the dilemma is to ensure that how to optimize the use of corporate data within every application, system and database throughout the enterprise.

Take into consideration the director of data warehousing at a major electronic component manufacturer who realized there was a problem linking information between an inventory database and a customer order database. The inventory database had data fields that represented product numbers differently than the customer order database. Nothing was done about it. There were hundreds of thousands of records that would have to be reviewed and changed but there were no resources available to take on this tremendous, time-consuming task. As a result, more than 500 customer orders were unfulfilled. At an average order per customer of $5000, the resulting loss of $2.5 million in revenue was incredibly significant.

PROBLEMS AND CHALLENGES

Bad data originates from a variety of sources – errors in data entry, erroneous data received from Internet forms, faulty data purchased or acquired from an outside source, or simply combining good data with outdated data and not having the ability to distinguish between the two.

One obstacle to creating good data is simply examining what is available and developing a plan for how it will be used. You will need to determine which data is good, which is bad and how bad the bad data is. Most organizations have little or no idea about the quality of the data residing in their systems and applications. In a 2002 survey conducted by The Data Warehousing Institute, almost half (44 percent) of the respondents said the quality of the data within their companies was “worse than everyone thinks”. This same report chronicles examples of costs and missed opportunities due to inaccurate or incomplete data:

• A telecommunications firm lost $8 million a month because data entry errors incorrectly coded accounts, preventing bills from being sent out.

• An insurance company lost hundreds of thousands of dollars annually in mailing costs (postage, returns, collateral and staff to process returns) due to duplicate customer and prospect records.

• A global chemical company discovered it was losing millions of dollars in volume discounts in procuring supplies because it could not correctly identify and reconcile suppliers on a global basis.

Sadly, most companies are oblivious to the true business impact of poor quality data. The simple truth is that poor data quality absolutely affects the bottom line. Accurate, complete data reduces costs in a number of ways, from the simple and obvious marketing savings (postage and production costs on a direct marketing piece, for example) to the less obvious organizational efficiency savings.

According to TDWI’s Data Quality Survey (mentioned above), almost half of the surveyed companies suffered losses, problems or costs due to poor quality data. Companies also cited extra costs due to duplicate mailings, excess inventory, inaccurate

billing and lost discounts, as well as customer dissatisfaction, delays in deploying new systems and loss of revenue.

MAKING THE MOST OF YOUR DATA

Data should be treated as a key strategic asset, so ensuring its quality is imperative. Organizations collect data from various sources: legacy, databases, external providers, the Web, and so on. Due to the tremendous amount of data variety and sources, quality is often compromised. It is a common problem that many organizations are reluctant to admit and address. The single most challenging aspect for companies is to recognize and determine the severity of their data quality issues and face the problem head-on to obtain resolution. Spending the money, time and resources to collect massive volumes of data without ensuring the quality of the data is futile and only leads to disappointment.

Cleansing data at the source is a significant way to enhance the success of a data warehouse or CRM project. Thus, it becomes a proactive rather than reactive model. Simply collecting data is no longer sufficient. It is more important to make proper sense of the data and to ensure its accuracy. As the amount of data escalates, so does the amount of inaccurate information obtained from the data. Data should be cleansed at the source in order to detect and address problems early in the process so that quality issues are prevented further down the line.

Information is all about integration and interaction of data points. Inaccuracies in a single data column can ultimately affect the results and may directly affect the cost of doing business and the quality of business decisions. Preventive measures to ensure data quality usually is more economical and less painful. Delaying the inevitable data cleansing dramatically increases the cost of doing so, as well as increases how long the cleansing process takes to complete.

Improved synergy between the extraction, loading and transformation (ETL) warehousing process and data quality offers the ability to more easily manage complex data integration. By applying data quality in the ETL process, data integrity and accuracy is assured. Much of the data warehousing effort is concentrated in the ETL process with the extraction of records and fields from various data sources, conversion of the data to new formats and the loading of data to other target destinations such as a warehouse or a data mart. The purpose of the ETL process is to load the warehouse with integrated and cleansed data. Data quality focuses on the contents of the individual records to ensure the data loaded into the target destination is accurate, reliable and consistent.

DATA QUALITY DEFINED

Data quality is often defined as a process of arranging information so that individual records are accurate, updated and consistently represented. Accurate information relies on clean and consistent data that usually includes names, addresses, e-mail addresses, phone numbers, and so on. Good data quality means that an organization's data is accurate, complete, consistent, timely, unique, and valid. The better the data, the more clearly it presents an accurate, consolidated view of the organization, across systems, departments and line of businesses.

Technological advancements that use data pattern analysis, "smart clustering," numerous data algorithms, and a host of other sophisticated capabilities ensure that data gathered throughout the organization is accurate, usable and consistent. By intelligently identifying, standardizing, correcting, matching and consolidating data, software solutions offer much needed relief to organizational data quality headaches.

Today organizations are looking for a wide range of features in data quality tools. According to the survey by TDWI, standardization and verification top the list of desired capabilities, followed by tools that define and validate business rules. Other important features include matching, consolidation and integration with other enterprise applications such as ETL tools.

SAS® provides the industry's only solution that integrates data quality into the ETL process. In addition, we provide a process and methodology to help your organization cleanse data before it is loaded into a warehouse for analysis.

THE VALUE OF ETL AND DATA QUALITY

Integrating ETL and data quality provides value to both business analysts and IT professionals. Many of the IT professionals focus heavily on the ETL process since its purpose is to load the data warehouse with integrated and cleansed data. However, data quality is a key component in the preparation for entry into the data warehouse. The best place to clean the data is in the source system so the defects cannot extend to the data warehouse and other inter-related systems. There are several methods to achieve integrated and cleansed data.

DATA AUDITING AND STANDARDIZATION

Data in a database or data store typically is inconsistent and lacks conformity. There are many ways to say the same thing. In Table 1, a title may be expressed in various ways.

Name	Title
John Doe1	VP of Sales
John Doe2	Vice President of sales
John Doe3	V.P. of Sales
John Doe4	Vice Pres. of Sales

Table 1: Variations of Title

If a direct mailing campaign plans to extract 3,000 "VP of Sales" out of a database by writing a query, and the query does not include every possible way that "vice president of sales" is represented, then the search will miss some of the target names. Inconsistently represented data is more difficult to aggregate or query. If the data is not represented consistently, it is even more difficult to perform any meaningful analysis of the data.

The first step in the cleansing process is data auditing. Data auditing provides counts and frequencies for data fields, identifies unique values and range reports with maximum and minimum values. In this phase, you should also define your business and cleansing rules.

DATA LINKING AND CONSOLIDATION

Anyone who gets two or three of the same advertisement or magazine instead of a single copy to which you subscribed is likely experiencing a perfect example of a "duplicate record" problem in the customer database. In some cases, there are probably small variations in the way the subscriber's name or address appear in the data base. Table 2 illustrates the example.

Name	Street	Country
William Smith	100 Sunset Dr.	USA
Will Smith	100 Sun Set Dr	US
Bill Smith	100 Sunset Drive	U.S.A
Billy Smith	100 Sunset	U.S.
Mr. W.H. Smith	100 SunSet Dr	United States

Table 2: Duplicate Records

William Smith may have several contact points and enter various flavors as contact information. A human being would look at these variations and instantly recognize that these names represent the same person, but a computer would store these as different records – hence multiple copies of the same advertisement or magazine. The circulation manager might think

that there are 150,000 customers but there are only 135,000 in reality. In this case, resources are allocated on inaccurate information extracted and loaded into the data warehouse with duplicate or redundant data. This problem can obviously be very costly for any organization that routinely mails against a large customer database.

Data linking and consolidation are fundamental to the ETL and data quality process. These capabilities link data from multiple sources and identify records that represent the same individual, company or entity. Consolidation, or householding, combines the elements of matching records into a single, complete record.

DATA ENHANCEMENT

Enhancement of data involves the addition of data to an existing data set, or actually changing the data in some way to make it more useful in the long run. Advanced enhancement technology enables users to easily append external data to their existing data sets without extensive programming. For example, many organizations want to profile customers so they can identify characteristics of current customers in order to conduct a sales and marketing outreach to people with similar characteristics but in a different geographic area. This company might want to enhance or increase the value of its customer database by adding data from an outside data source such as Dun & Bradstreet.

But, this process can produce problems if the data in the D&B database is represented in a different way than it is in the current customer database. In this case, the organization would need a sophisticated technology that can recognize and match the different data in order to enhance or increase the value of the primary database. Enhancement technology allows a company to get more from its data by enriching its data with additional information.

DATA CLEANSING AND VERIFICATION

SAS uses matching and standardization routines to analyze, cleanse and standardize data from a variety of platforms and sources. It includes defect analysis and corrections for invalid data and data verification processes.

Verification is the process of validating data against a known standard. For example, if a company wants to import some data from an outside vendor, they might use the U.S. Post Office database to ensure that the ZIP codes match the addresses and that the addresses are deliverable. If not, the organization potentially could incur a great deal of undeliverable mail.

CONCLUSION

It is a fact that ignoring data quality is costly and it affects every industry that relies on accurate and consistent information. Data quality is a continuous process and effort. By addressing data quality at the source, data cleansing becomes a proactive rather than a reactive process. The integration of data quality and ETL minimizes the risk of failure, cost and number of resources to manage the data. The synergy between ETL and data quality provides integrated and cleansed data for your warehouse, mart or other repository. Data is a vital resource and it is should treated as a key strategic asset in order to obtain reliable and accurate view of your organization.

REFERENCES

TDWI Report Series, Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data, by Wayne Eckerson, January 2002

W.H. Inmon, Data Quality in the Corporate Information Factory, October 2001

Larry English, Improving Data Warehouse and Business Information Quality, Wiley and Son, Copyright 2000

ACKNOWLEDGMENTS

I would like to thank the members of the SAS and DataFlux teams who provided comments and content to this paper. Special recognition goes to Katie Fabiszak and Stephanie Townsend who helped to organize this paper.

CONTACT INFORMATION

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

Tho Nguyen
SAS Institute Inc.
1 SAS Campus Dr.
Cary, North Carolina 27513
Work Phone: 919-531-4867
Fax: 919-677-4444
Email: tho.nguyen@sas.com
Web: www.sas.com