Paper 49-27

Data Quality – The Fuel that Drives the Business Engine

Tony Fisher, DataFlux™ Corporation, Cary, NC

ABSTRACT

In today's information age companies will live and die by information. This information is the fuel that drives the business engine. As more and more data is collected, the reality of a multi-channel world that includes e-commerce, direct sales, call centers and existing systems sets in. Bad data is affecting companies at an alarming rate and the dilemma is clear: how can a company ensure that it is getting the most out of every application, every system and every database to maximize the use of corporate data throughout the enterprise? In the end, the companies with the most complete, accurate and reliable information will gain and retain market share.

Recognizing that you have a data quality problem is just the start. The real question is how to fix the problem. In this session you will learn how to determine if incomplete, inaccurate and unreliable data is affecting your business. Once you understand the problem, you'll discover how to treat it by learning the requirements to consider when looking for a total data quality management solution.

Allowing poor data to exist within your company is inexcusable because there are quick, affordable options available that can improve the accuracy and consistency of your data. Ensure that your business engine will run smoothly by fueling it with clean and reliable data.

INTRODUCTION

Enterprises live and die by information. Companies with the most complete, accurate and reliable information will gain and retain market share. That information is used to increase revenues and to reduce expenses. In this information age, companies that know the most about their customers, partners, products, transactions, vendors and business units will be the companies best positioned for success.

Information powers the business engine of today's information centric enterprise. But the corporation is a large and complex place. Information that stands unorganized and alone is almost meaningless. In order to achieve potency, information needs to be ordered into a framework – a structure. Today's business intelligence applications build the framework that generates and harnesses the power of information and data is the fuel that feeds that engine. The business intelligence engine, like any engine, consists of a series of interacting components that work in concert with each other to provide a robust infrastructure. In order for a business to run smoothly, each component in the engine must execute at peak efficiency.

Like any engine, the intelligence engine requires clean and reliable fuel – the data that runs the business. Any failure in the data flow, any slowing, any impurities or incompleteness in the content of data used for the basis of decisions, causes the engine to run poorly and has a negative business impact on the entire enterprise. If the components of the intelligence engine cannot maintain a high degree of reliability, then the entire enterprise suffers. In contrast, an enterprise with complete and reliable data is able to generate information so rapidly that information is available before competitors have the equivalent information. The result is speed to market, which yields a powerful and distinct competitive advantage. The goal of this paper is to detail the data quality and data integration processes that need to be in place to obtain the strategic information yield. Data Quality is instrumental in ensuring that the components of the intelligence engine can be successfully integrated and can operate at optimal efficiency. Data integration provides complete information to the intelligence engine enabling you to have the most complete business intelligence about your customers, partners, vendors and business practices.

THE CASE FOR DATA QUALITY

Information intelligence is a strategic imperative that enables the corporation to respond quickly and efficiently to shifting market conditions. As business intelligence applications continue to expand, the underlying data that feeds the outlying decision support applications becomes increasingly more important. Data that provides accurate, consistent, and standardized data enables the corporation to achieve both revenue generating and cost-optimization goals. Precise and complete data, fully leveraged using rich analytics, yields the business intelligence necessary for a distinct competitive advantage.

There is a tremendous quantity of data residing in the enterprise systems that feed the corporate data warehouse. Thousands and, in some cases millions, of records continually move among the diverse systems. In addition, data travels between the corporation and vendor and partner systems and streams into the enterprise via data entry systems and unstructured web responses. The increase in data and system complexity alone can easily lead to ambiguous data representations.

As an example of ambiguous data representations, names and addresses can be depicted in various ways, depending upon the system in which the data was originally entered. For customer Robert Smith, a call center representative might enter the following: Bob Smythe, 100 E. Johnson Street. The invoicing system, however, might use a different designation: Robert Smith, 100 East Johnson Street. Mr. Smith may have entered Bob Smith with an address of Johnson Street on a customer website. Similarly, two separate systems might use different numbering schemes to encode customer information, with one system using the customer's last name and a number, and the other one using a random number. There are then different ways that represent the same customer:

Robert Smith
100 East Johnson Street
Customer id: Robert Smith

Bob Smythe
100 E. Johnson Street
Customer id: A0004972

Bob Smith
Johnson Street
Customer id: none

Robert Smiht
100 East Johnson St.
Customer id: SMI047

In the operational environment, these different representations of the same customer may be tolerable because the operational activities operate as discrete applications. But as the information is acquired by the operational applications and is transformed into applications, it is imperative that these different representations of the same customer become consolidated into the representation

of a single customer, which in fact is the case. In this example, allowing different representations of the same Robert Smith to be propagated to the Business Intelligence platform will produce data quality problems as there is no one accurate representation of the information that relates to Robert Smith.

IGNORANCE IS BLISS

Critical business decisions, allocation of resources, price changes, marketing campaigns, and daily operations revolve around key enterprise data. In no small measure a company's success or failure is based on the quality of information contained within the corporate data. Yet, despite its fundamental importance, data quality is often ignored. The perception is that implementation of data quality is typically costly, resource intensive, and time consuming.

But, perhaps, the most common reason for ignoring data quality issues is that pains and problems have not been attributed to data quality. It seems that often a company will wait until a major problem has occurred affecting revenue or impacting customers or forcing poor executive decisions before an effort is made to find, correct and prevent data quality problems.

The following are some real life examples of the pains that can happen if data quality problems are ignored:

- 390,000 income tax refund checks came back to the IRS marked as "undelivered." The USPS processes 43 million change of address requests each year.
- Barbra Streisand pulled her investment account from her investment bank because it misspelled her name as 'Barbara'
- A mail order company faced a massive lawsuit for sending out demeaning catalogue deliveries. Practical jokers took advantage of the fact that there was no quality checking on a company web page where "customers" could request a catalog. The company did not catch the bogus data and sent the catalogs to the unsuspecting recipients.
- 15-20% of voter records were marked as moved or deceased when compared with USPS.
- Three zeros added and reported as the trade volume of an inside executive of an Atlanta company caused its stock to plummet over 30% in one day.
- An acquiring company learned post acquisition that their newly purchased company had 50% less customers than thought due to duplicate data in the customer database
- The U.S. Attorney General's office has stated that "approximately $23 billion, or 14 percent of the health care dollar, is wasted in fraud or inaccurate billing."
- A corporate president and bank customer with $1M invested in a bank's mutual funds just removed all of his money. Why? The bank called to raise the rates of a checking account with a balance consistently under $100. The bank had no way to get full access to the customer's interactions with the bank.

You do not want your company to be added to this list. It is estimated that a staggering 20% of all data in an organization's databases is erroneous or unusable. And this only becomes worse as company's share data with partners, purchase external data or even allow their customers to enter information directly into the corporate databases via their website.

Of course, focusing on data quality and data integration is not just about protecting your company from embarrassment. In many instances, implementation of data quality practices can be responsible for new revenue and for savings on expenses, as in these examples.

- Problems in accounts payable and accounts receivable are common:
  - A European company discovered through a data audit that it was not invoicing 4 percent of its orders. For a company with $2 billion in revenues, this meant that $80 million in orders went unpaid.
  - It is common for large companies to pay the same invoice multiple times because an invoice gets posted multiple times with a slightly different spelling of the name of the supplier or product.
- A major holding company for a number of health insurance plans implemented a data quality initiative to find fraudulent claims that were duplicated across different plans. This initiative found that between 8-10% of all claims were duplicate claims.
- To improve customer satisfaction and reduce returned merchandise, an online retailer validated customer information entered on its website against both the USPS database and against their corporate customer database. This resulted in fulfillment savings of 22% from one holiday season to the next in addition to the added customer satisfaction.

The efficiencies provided by these data integration initiatives can be remarkable. Any initiative that can have a multi-percent increase in revenues or decrease in expenses can have a substantial impact on a company.

DEFINING DATA QUALITY AND DATA INTEGRATION PROBLEMS

Implementing data quality and data integration processes is to transform imperfect and isolated data into accurate, consolidated information assets. By implementing these processes, organizations can realize improved accuracy, more timely delivery of information and more confidence in decision making.

Thus far, we have looked at these processes in very broad strokes and we have looked at the ramifications of poor data fundamentals and the advantages of proactive data maintenance. Now, let's dig deeper into the basic data quality and data integration activities that help deliver quality, consolidated data to the right place at the right time.

There are many challenges that application developers face in trying to provide rational, consolidated data, including: data access, cleansing, data integration, data matching, data augmentation, data standardization, data consolidation and rules to maintain these processes. Although there are many aspects to a comprehensive approach, these can be grouped into four major categories:

- Data Quality
- Data Linking
- Data Enhancement
- Data Rationalization

Let's take a close look at how to effectively implement these processes.

DATA QUALITY

Successful applications begin with data quality. It does not matter how fast your application is, it does not matter how slick the interface is, it does not matter if an application is small or large, if the data that drives the application is poor, the results of the application will be poor. Data quality techniques are designed into applications to improve the accuracy of data. There are several data quality processes that are necessary:

- Data standardization
- Data deduplication
- Data validation

DATA STANDARDIZATION

Unfortunately, data can be ambiguously represented. This fact often is positioned at the very root of an organization's data quality issues. If multiple permutations of a piece of data exist within a data set, then every query or summation report generated by the data set must account for each and every instance of these multiple permutations. Otherwise, important data points can be missed and can severely impact the output of these processes.

For example, a company name can be represented a multitude of ways:

IBM, Int. Business Machines, I.B.M., ibm, Intl Bus Machines

As can a product name:

Blue Turtle Neck, Turtle Neck: Blue, B Turtle Neck, Shirt:TurNeck B

Or an address:

100 E Main Str, 100 East Main Street, 100 East Main, 100 Main St.

They all have the same meaning, but are represented very differently. It is obvious to surmise what kinds of analytical problems can and will arise if the same data is dissimilarly represented within a data set as these examples demonstrate.

Imagine a life insurance company wanting to determine the top ten companies that their policyholders work for in a given geographic region in order to tailor policies to those specific companies. Inaccurate aggregation results are likely because of all the permutations of data for a given company name will be difficult to account for.

Consider a marketing campaign that personalizes its communication based on a household profile but there are a number of profiles for customers at the same address, only the addresses are inconsistently represented. Variations in addresses can have a nightmare effect on these types of focused campaigns, and can cause improper personalization or too many generic communication pieces to be generated, wasting dollars on both material production and creative efforts of the group and alienating customers.

Picture an apparel company trying to determine what products to manufacture, where to manufacture them, how many products are in inventory and where to ship them if they cannot get a total understanding of product sales history because their systems do not standardize product description information across systems.

While these are simple data inconsistency examples, these and other similar situations are endemic to databases worldwide. Fortunately, data quality technology now exists that identifies these various permutations of data and can rectify the situation a number of ways. These include physically standardizing the data within the data set, creating synonym tables/filters, or correcting undesired permutations before they enter the data set in the first place. And, more importantly, these rules for standardization can be maintained external to an application or data set and applied to various applications to standardize across a corporation.

DATA DEDUPLICATION

Another common example of a data quality issue is duplicate data or redundant data. Again, because data can be ambiguously represented, the same customer, prospect, part, item for sale, transaction, or other important data could be occurring multiple times. In cases like these, the redundancy can only be determined by looking across multiple fields.

The following are examples of duplicate data that cannot be caught without some form of data quality technology (or else long, endless hours of human inspection, unlikely to catch as high of a percentage, and impossible with anything more than small volumes):

Robert Smith, 100 E Johnson Street
Bob Smythe, 100 East Johnson
Dr. Robert J. Smith, 100 E. Johnston St.

Ms. Kathleen Anderson, Box 12 – 9 Canary Street
Katie Andersen, 9 Canary St. #12

Large Camping Knife
Knife, Camping Lg.

The Briggs Corporation, Saint Louis
Brigs Corp, St. Louis

Problems that can arise from redundant data within a data set include inaccurate analysis, increased marketing/mailing costs, customer annoyance, and relationship breakdown across a relational system. Again, as data such as this serves as the foundation and infrastructure of our business intelligence systems, it is imperative that these situations be identified and snuffed out in order to achieve success.

DATA VALIDATION

Every company has basic business rules. These business rules cover everything from basic lookup rules:

Salary Grade	Salary Range Low	Salary Range High
20	$25,000	$52,000
21	$32,000	$60,000
22	$35,000	$75,000

To complex, very specific formulas:

Reorder_Quantity = (QuantPerUnit*EstUnit) [Unit_type] -Inventory_onHand

Many basic business rules can be checked at data entry time and, potentially, rechecked on an ad-hoc basis. Problems that arise from lack of validation can be extensive, from over-paying expenses to running out of inventory, to undercounting revenue.

Applications today need the ability to store, access and implement these basic business rules for data validation. Data validation rules should be stored external to an application so they can be shared by all applications, thereby avoiding conflicts across application data stores.

DATA LINKING

Data linking becomes a data quality and application design issue when the columns that constitute the join fields between multiple data sets contain data that is inconsistently represented. For example, trying to combine a customer table with an outside demographic data source will have undesirable results if the join column is a column commonly containing ambiguous representations of data such as company name:

Data Source A (Customer Data set)
Columns: Customer Name, Contact
Data: First Bank of Denver, Joe Snow

Data Source B (Demographics)
Columns: Company Key, Num Employees, Business Type, Annual Sales
Data: The 1st Bank of Denver, 850, Financial, $62 million

Obviously a standard SQL join statement would not recognize that these two banks are the same and therefore the demographic data would not be joined to the customer data.

One way to achieve a join that would indeed succeed in this scenario is by using a match code that *unambiguously* represents the company name. Data quality algorithms can be used to generate this unambiguous code. The code itself might be represented by something covert, such as **RX19E4**, however the same code will be generated when any permutation of the "First Bank of Denver" is passed through the match code generation algorithm. This unambiguous code then becomes the basis of the match between data sources, and can be constructed using any number and combination of columns. These codes can be stored as an extra column in each data source, stored in a temporary table or file, or generated solely at runtime.

While data integration may not be considered a "quality" problem by some, the same types of algorithms and procedures apply that can achieve much higher match rates and therefore much better success when combining data from multiple sources. Often, these integrated data sets form the basis from which many business intelligence applications thrive. Data linking in an application environment might be either explicit when multiple data sources are physically joined and a new data store is created. Or the data might just be linked implicitly using the match code data to perform a run-time join of the data for use by an application without creating a consolidated data store of the input data sources.

DATA ENHANCEMENT

Another component of data quality that will make applications more effective concerns resolving missing and/or inaccurate data by using an external data reference. This includes not only filling in missing values and replacing inaccurate values, but also adding additional data values to a record or data observation that provides a more complete picture of the entity that is being stored in the data set.

A common example of this is using the United States Postal Services' master address database to verify and/or correct existing addresses within a database. In addition, you can append other useful demographic postal data such as Zip+4, carrier routes, congressional districts, counties, delivery points, etc. This can greatly increase address integrity, as well as provide a basis for additional applications such as geocoding, mapping, and other visualization technologies that require a valid address as a starting point. Obviously, technology such as this can go a long way as an integral part of a business intelligence application.

Another typical way for companies to facilitate this is to create a master database with valid data (product data, customer data, patient data) and keep the master database clean by periodic quality, standardization, and deduplication techniques. Then, other applications use linking technology to ensure that duplicate data is not entered into other data stores and to make sure that application data is consistent with the master database.

DATA RATIONALIZATION

The final aspect of quality application development is developing an understanding of what data is in use in a company and the relationship of data in other applications to the application being developed. Having pieces of relevant data spread out across many different application data stores makes it difficult to develop a complete understanding of the data.

Looking back at the previous data linking example, there were multiple data stores that contained valid information about customers (in this example, a customer data set and a demographics data set.) A marketing group within a company may create a new initiative to do demographics based campaigns. Without a thorough knowledge of the data within the company, the marketing department would be tasked with establishing the demographics for the campaign. This represents a duplication of effort. And, more importantly, it does not take advantage of the potentially valuable, customized information in the existing demographics database – a database that might be updated depending on the behavior of current customers.

Data rationalization is more of a metadata discovery technique than it is an activity that directly manipulates the data. During application design, tools that can be used to investigate existing data and understand the relationships across data stores should be employed. Applications should be developed with the goal of providing further integration of existing environments rather than just concentrating on the independent needs of the application.

SO, LET'S GET STARTED

Data quality tools to assist in applications need to be available for the various different aspects of application development: design, implementation and post-production maintenance. The combination of the SAS® and DataFlux products provide key technologies to facilitate these aspects.

DataFlux's end-user product dfPower Studio™ brings data quality and data integration capabilities to your desktop. dfPower Studio is customizable, easy to use and can be utilized by any one in any department. The solution provides data management functionality for identifying and fixing data inconsistencies, matching and integrating items within and across data sources, and identifying and correcting duplicate data in a data source. dfPower Studio also provides data enrichment functionality to enhance your business data with valuable geographic and demographic data. dfPower Studio is a powerful tool for data rationalization, data augmentation and ad-hoc data validation/correction in a production application environment.

Blue Fusion™ SDK, a software developer kit, is a packaged set of callable libraries that easily integrate into internal operational, data warehouse, decision support, and e-commerce or web portal applications.

Blue Fusion™ CS is a software developer kit built on the Blue Fusion SDK. Blue Fusion CS provides you with the same functionality as Blue Fusion SDK, but in a client/server architecture.

SAS® Data Quality - Cleanse is a SAS language product that provides the same functionality as Blue Fusion SDK and Blue Fusion CS for building data quality, linking, augmentation and rationalization into SAS applications. All of these products use the same business rules (metadata) to ensure integrated applicability of the development environments.

Working together or separately, dfPower Studio, Blue Fusion SDK and Blue Fusion CS and SAS Data Quality - Cleanse ensure the highest data quality environment, producing better business decisions and improved data driven initiatives.

The combination of these technologies provides:

- Data quality assessment and data rationalization
- Technology for improving data quality at the application level
- Technology for applying data quality techniques when data from different sources is integrated
- Data quality tools that can be used to do ad-hoc data analysis on existing data.
- A metadata repository to store the business rules that control all of these areas.

For quality applications you need a quality development environment.

ACKNOWLEDGEMENTS

CONTACT INFORMATION

Tony Fisher
President/CEO
DataFlux Corporation
4001 Weston Parkway
Suite 300
Cary, NC 27513
919.674.2153
tony.fisher@dataflux.com