Paper 166-28

Better Decisions Through Better Data

Tony Fisher, President and General Manager, DataFlux
George Marinos, National Data Quality Partner, PricewaterhouseCoopers LLP

ABSTRACT

In today's market, everyone is seeking a way to gain competitive advantage. Most companies are focused on building new systems, implementing new strategies, or identifying new markets. What is often ignored, or assumed, is management of the quality of the data that supports the existing decisions and how it can be improved to help gain a competitive analysis

It is critical to understand that the quality of the results from any analysis is only as good as the quality of the inputs (the data) that feed that analysis. Given today's accessibility and quantity of data, people often try to perform new types of analysis to gain a more competitive edge. Many times they are trying to answer the same question; however due to poor data management they seek alternate analysis methods instead of trying to improve their existing analysis. Many business decisions are based upon a few key analyses; therefore it is our recommendation to improve the quality of these analyses through improved quality of data before developing new analysis.

INTRODUCTION

Organizations depend on data. Regardless of industry, revenue size or the market it serves, every company relies on its data to produce information for business decision-making. With so much riding on such information, does data quality get the attention it deserves in your organization?

While no one wants to consciously admit that their business decisions are based on inaccurate or incomplete data, 75 percent of organizations have no data quality management processes in place. Backing up this figure are recent surveys by PricewaterhouseCoopers and a 2002 survey conducted by The Data Warehousing Institute (Data Quality and the Bottom Line). It appears that the majority of organizations have not taken the necessary steps to determine the severity of data quality issues and its impact on the bottom line.

Companies spend 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. Unfortunately, many overlook the fact that it is the underlying data that matters. All of the fantastic screens and reports in the world won't make a positive difference if the data that supports the system is inconsistent, redundant and full of errors.

Many individuals believe that improving the quality of an organization's data assets is easier said than done, often choosing to ignore data quality rather than address it. Everyone knows data quality is a problem, but most everyone in denial as to the extent of the problem and its impact on the overall organization. Most often, data quality problems are not given the attention they deserve until an underlying application or system fails, such as CRM or data warehousing. Not until an initiative is deemed a failure, or the return on investment (ROI) is not achieved, does data quality come to the forefront.

Effective data quality management relies on the combination of people, process and technology. This paper will explore the problems of data quality management and then look more closely at how to utilize the people, process and technology available to your organization to achieve better data quality. Two case studies of successful dataquality management implementation will also be presented.

WHAT CONSTITUTES DATA QUALITY?

The question "what is data quality?" begs to be answered. The industry analyst group Current Analysis says, "Data quality is simply a reflection of the accuracy of an organization's data. 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 business lines." (Current Analysis, Data Quality Product Assessment, October 19, 2001).

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; as well as non-name and address data like part numbers, product names, sales figures and so on.

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 and integration is impossible. Spending the money, time and resources to collect massive volumes of data without ensuring proper management of the data is futile and only leads to disappointment.

PROBLEMS AND CHALLENGES IN CREATING TRUSTWORTHY DATA

Bad data originates from a variety of sources – errors in data entry, erroneous data received from internal forms or Internet forms, system discrepancies among different parts of an organization, 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.

To create good data, you must examine what is available and develop a data management plan for how it will be used. You will need to determine which data is good (accurate, complete, timely, unique and valid), which is bad and how bad it is.

Most organizations have little or no idea about the quality of the data residing in their systems and applications. In a 2001 survey conducted by PricewaterhouseCoopers, only one-third of the companies were very confident in the quality of their data. This lack of confidence is for a good reason. There are many data anomalies that plague these various systems:

- No standardization:
  - Differences in the spelling of common words, such as Street, st, St, and so forth.
  - States may be shortened or spelled out, such as TX or Texas.
  - Salutation can differ from one application to another, such as Mr., Mister, no salutation, etc.
  - Name fields are particularly troublesome fields. Some names have many forms, such as Robert, Bob, Rob, Bobbie, Bobby, etc.
  - Formats are usually inconsistent across different applications. For example, different systems use different date formats: dd/mm/yy or yyddmm, etc.
  - The same data is expressed in different ways. Gender may be M and F, 0 and 1, etc. Many systems overload fields like this and allow entries like "Unknown" or "Joint Account" or "Corporation."

  Companies may be represented differently. GM might

mean General Mills or General Motors.

- Incorrect data. Often times, special codes are utilized to signify unknown or default data. For example, entering a social security number of 111-11-1111 or a birth date of 01/01/01 may have been permitted by systems in order to complete a transaction when data was unknown.
- Data does not adhere to corporate business rules. For example, a salary may be out of range for a salary grade.
- Data is stale. Data that has decayed or changed over time is now of little value.
- Data may be ambiguous across systems with no consistent key to allow the joining of the information across systems. For example, 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	BOB SMYTHE
100 East Johnson Street	100 E. Johnson Street
CUSTOMER ID: ROBERT SMITH	CUSTOMER ID: A0004972

Bob Smith	Robert Smiht
Johnson Street	100 East Johnson St.
Customer id: none	Customer id: SMI047

As the information is acquired by the systems and is transformed into analytic 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.

HOW DO YOU ACHIEVE EFFECTIVE DATA MANAGEMENT?

Today, organizations are looking for solutions to improve data quality. These solutions must include process, people and technology. Let's look more closely at the people, process, and technology for effective data quality management.

People

Data quality has traditionally been considered a data standardization activity best left to the database management team. The tendency has been to let the IT departments not only physically change and correct the data in corporate databases, but to also create the rules and routines by which the data is transformed. The effect has been that what is considered good clean usable data from a technical standpoint by the database management team may not be the complete, accurate, and timely information the business user (the ultimate "consumer" of the data) actually needs to perform a given set of tasks.

Effective data quality requires not only the IT user to efficiently perform his or her job of physically transforming the data, but also the business user to define what actually constitutes "good" data for a particular task, process, or project. Data quality then moves from being about the technical "correctness" of the data to a concept of data integration, where a company's accumulation of digital information is transformed into a strategic asset by creating a consistent, timely, reliable view of enterprise data.

Process and Technology

Let's dig deeper into the basic data quality management activities that help deliver quality, consolidated data to the right place at the right time.

Because the needs and goals of organization can vary so widely, and because the data and the problems related to data quality depend so heavily on how the data is collected, stored, and used, a general process is required to tie all aspects of data quality management projects together. And, technology is needed to drive the components of the data quality management process.

The four basic steps of the process are:

- Data Profiling
- Data Cleansing
- Data Integration
- Data Enhancement

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

Data Profiling

The profiling is the phase of discovery – understanding and documenting where the problems with your data supposedly reside. Listing databases and tables, or even departments from where problem data gets generated will help you establish where to begin drilling down into the data for more tell-tale signs of data quality problem. The drill-down phase could mean profiling the suspect data for ranges, or frequency counts or could mean checking for valid primary key relationships. This discovery phase could also be where one searches for patterns or duplicates records or even valid value analyses.

The point is that the discovery phase, for whatever your data management initiative, helps you generate the rules or procedures from which all operations will flow. Remember, the profiling phase is about looking at your *own* data and data collection or data transformation processes, and then generating or applying techniques to fix problems found there. Resist the temptation to dive in and transform the data without first exploring potential sources of trouble. Data that on the surface looks to be in error could in fact be the correct representation after further examining the data or process from which it was generated.

Data Cleansing

Data cleansing techniques are designed into applications to improve the accuracy of data. There are several data quality processes that are necessary:

Data standardization

Data validation (domain, range and missing 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 dataset, then every query or summation report generated by the dataset 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 dataset 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.

Picture a marketing campaign that personalizes its communication based on a household profiles but there are a number of profiles for customers at the same address, only the address 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.

Envision an apparel company trying to determine what products to manufacture, where to manufacture them, how many products are in inventory and where ship them if they can't get a total understanding of product sales history because their systems don't 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 cleansing technology now exists that identifies these various permutations of data and can rectify the situation a number of ways, including physically standardizing the data within the dataset, creating synonym tables/filters, or correcting undesired permutations before they enter the dataset in the first place. And, more importantly, these rules for standardization can be maintain external to an application or dataset and applied to various applications to standardize across a corporation.

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 Integration

Data integration is a combination of:

Linking: finding common fields between databases or within a single database
Consolidation: once you have found common fields how do you join or merge the data.

Data Linking

Data linking is an issue when the columns that constitute the join fields between multiple datasets may 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 Dataset)
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 datasets 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 Consolidation

Once you have determined that multiple records represent the same data element, you must determine what process to follow to consolidate the duplicate/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 dataset include inaccurate analysis, increased marketing/mailing costs, customer annoyance, and relationship breakdown across a relational system. Again, data such as this serves as the foundation and infrastructure of our business intelligence systems, it is imperative that these situations be identified and eliminated in order to achieve success.

Your consolidation process might consist of deduplication (removing certain records), merging (choosing the best information across multiple records), or keeping the information from all data sources.

Data Enhancement

Another aspect of data quality that will make applications more effective concerns the area of 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 dataset.

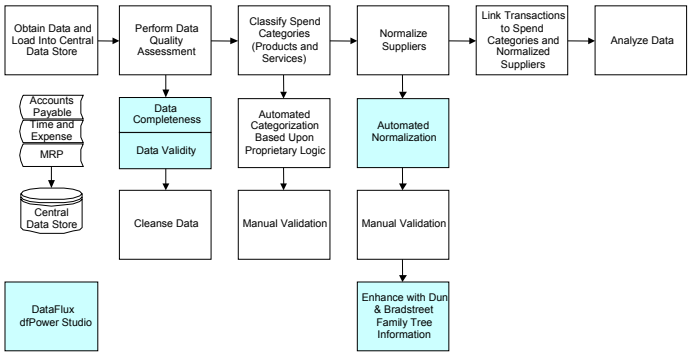
A common example of this is using the United States Postal Services' master address database to verify and/or correct existing addresses within a database, as well as append other useful demographic postal data such as Zip+4, carrier routes, congressional districts, counties, delivery points, etc. This 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.

CASE STUDY: PROCUREMENT ANALYSIS

In order to achieve the cost savings' goals defined by senior management, the Procurement Officer of a Global Insurance organization, initiated a project to analyze how and where the organization spends its monies. The Procurement Officer was interested in understanding: 1) how much was the total spend; 2) what type of products or services were being purchased; and 3)who was the organization purchasing from. The answers to these questions would assist with the ability to: negotiate more favorable contracts based upon the amount purchased from a vendor; consolidate the purchasing of specific types of products or services; and determine the level of compliance to existing vendor contracts.

The questions the Procurement Officer was asking were very rudimentary, however the reason why these questions had gone unanswered were that: the data required to support the answers and provide this analysis was stored in multiple systems; no link existed to combine the data from these systems; and some of the key data were not captured by the systems or were captured inconsistently. Therefore, the Procurement Officer knew that the success of this analysis was dependent upon integrating, standardizing, and enhancing the quality of this data.

This analysis hinged upon the quality of the source data and the ability to link and standardize the suppliers and products/services. The following process was implemented to address this issue:



This solution included: people, process, and technology. The

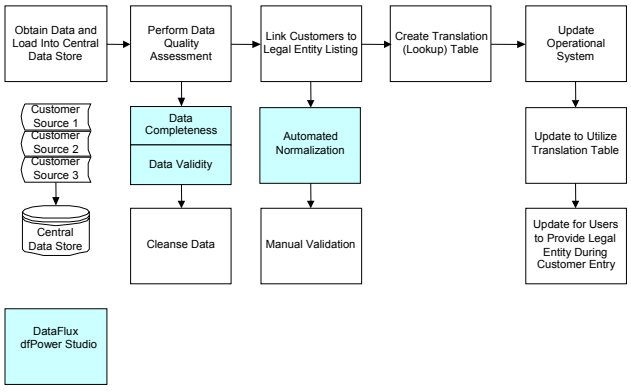
people consisted of the analysis team and the end users that understood the data. The process consisted of piecing together the various steps, and the technology consisted of dfPower Studio (DataFlux) and and the Procurement database.

If the above process was not implemented, the analysis would not have been successful. The users would not have been able to obtain the full picture of the suppliers (e.g., there were over 20 accounts in the various source systems that represented one normalized account, however without the above process the analysis would have shown these as separate organizations and leverage in negotiation would have been lost). In addition, without standardizing the categories of spend, the analysis would not have revealed the true amount of spending on various products and services – the information would have been either too broad (e.g. information technology) or too detailed (e.g. a specific type of computer).

This is a good representation of how the quality of the source data and how it is improved and enhanced will determine the success of your business decisions. The questions being asked were not complex, however being able to provide high quality, integrated data was critical to the success of this project.

CASE STUDY: ANTI-MONEY LAUNDERING

The USA Patriot has caused compliance departments in a number of industries to take a hard look at the quality of their data. For example, the Compliance Officer of a Global Banking Institution needed to ensure that the organization was in compliance with the act. Specifically, the organization needed to gather additional information on their customers and clean existing information they had. They also had to put the appropriate policies in place to ensure that their data would be maintained at the quality that was needed to appropriately risk rate their clients. They needed to implement a solution that would take their existing customer data and properly link those customers to additional information available from third parties. After this initial linking, the existing operational sytems would be updated to capture the additional information. This analysis hinged upon the quality of the source data and the ability to link the customer base to a standardized listing of the required fields (e.g. country of jurisdiction, address, name contained in the data obtained by third parties.The following process was implemented to address this issue:



This solution included: people, process, and technology. The people consisted of the project team and the end users that understood the data. The process consisted of piecing together the various steps, and the technology consisted of dfPower Studio (DataFlux), the database of the current customer information and another database with the new gathered information.

The above process was chosen due to the large volume of data and the short timeframe required to implement this solution. In order to ensure the highest quality of matching, a separate manual validation process was implemented. The manual validation process provided the end users with the ability to either accept the automated link or override. In this case, a separate project was implemented to provide an additional data element that had not been captured but was critical to this organization's compliance with the Patriot Act.

CONCLUSION

To recap, the four cornerstones of Data Management are Data Profiling, Data Cleansing, Data Integration and Data Enhancement. Execution of these steps might seem overwhelming, but it is not as expensive, difficult and time consuming as it once was. Data Management tools are easy-to-use and are also customizable, allowing implementation to be within your reach.

It is through the application of people, process, and technology to these cornerstones that data quality can be significantly improved , resulting in better decisions. As was demonstrated in the case studies, the quality of decisions are dependent upon addressing all four of these cornerstones to ensure that accurate, integrated, and complete data are available. Today, the processes and tools are available to improve the quality of your data that will give you the ability to make better decisions and gain a competitive advantage.

You need a solution and a methodology that easily integrates every aspect of a fully functional CRM system: operational applications and analytical applications. When you can ensure the data management environment, you will be enabling better business decisions and improved data-driven initiatives.

Every puzzle begins with the first piece and builds from there. The first piece of the Data Management puzzle is to take stock of your systems. Only then can you begin to uncover the integration issues within these systems. Understanding the problem is a large part of the solution. Analysis and data discovery techniques allow you to investigate your current Enterprise systems.

Begin with discovery. End with enlightenment.

CONTACT INFORMATION

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

Author Name	Tony Fisher
Company	DataFlux
Address	4001 Weston Parkway, Ste300
City state ZIP	Cary, NC 27513
Work Phone:	919-674-2153
Fax:	919-678-8330
Email:	tony.fisher@dataflux.com
Web:	www.dataflux.com

Author Name	George Marinos
Company	PricewaterhouseCoopers LLP
Address	
City state ZIP	
Work Phone:	
Fax:	
Email:	
Web:	