Paper 117- 26

Data Quality – Spinning Straw Into Gold

Bob Brauer, DataFlux Corporation, Cary, NC

DATA QUALITY TODAY

The issue of data quality is a simple one. As IT spending soars and organizations continue to spend a large portion of their IT budgets on Business Intelligence applications such as data warehousing, customer-relationship management, data mining, marketing automation, and sales force automation, the importance of the underlying source of data that feeds these applications has become increasingly higher. After all, well into the Information Age, organizations have become entirely data-driven. Rare is the employee who does not come into daily contact with an element of information originating from a company's various databases. Critical business decisions and allocation of resources are made based upon what is found in the data. Prices are changed, marketing campaigns created, customers are communicated with, and daily operations evolve around whatever data points are churned out by an organization's various systems. In other words, companies are living and dying as a result of the information contained within their data.

THE ISSUE OF DATA QUALITY

This is hardly a distressing affair. After all, what we yearn for from our information systems is more efficiency and increased effectiveness in every facet of the organization, and that is precisely what these business intelligence systems can deliver, with one small catch. The data that serves as the foundation of these systems *must be good data*. Otherwise, we fail before we ever begin. It doesn't matter how pretty the screens are, how intuitive the interfaces are, how high the performance rockets, how automated the processes are, how innovative the methodology is and how far-reaching the access to the system is, *if the data is bad - the systems fail*. Period. And if the systems fail, or at the very least provide inaccurate information, every process, decision, resource allocation, communication, or interaction with the system will have a damaging, if not disastrous, impact on the business itself. Worst of all, all the money and resources spent within IT on the development and deployment of these Business Intelligence systems will go down the proverbial toilet, and can negatively effect, sometimes severely, the operations and success of the company we all dedicate our lives to on a daily basis.

WHY ARE DATA QUALITY ISSUES SO RAMPANT?

Nothing discussed up to this point falls short of obvious. While it is certainly true that data quality is often overlooked in the race for glory and on-time delivery during the development of information systems, rare is the individual who claims that data quality is not an important issue and *should* be addressed. However, rarer still seems to be the individual who strives to actually make it a high priority issue during systems development, and something that is part of the agenda of every status meeting that discusses the success of a particular information system. This can be quantified by a report from the META Group indicating that 75% of companies in the United States have yet to implement *any kind of data quality* procedures to their systems development lifecycles. Of those that have, the majority report that the current implementations fall far short of what the organization requires. This is a phenomenon worth considering closely.

There are several reasons why the issue of data quality often gets swept under the carpet during the design, development, and deployment of information systems. While there are certainly others, three primary reasons will be discussed here. They are *attractiveness of the subject matter, the cost and difficulty in implementing data quality, and the inability to demonstrate ROI.*

The first reason to tackle is *attractiveness of the subject matter*. Data Quality is right up there in frequency with Cobol performance-tuning and fax-machine maintenance as a career objective among today's IT professionals. Most engineers are interested in the latest and greatest hot technologies and areas such as graphical interface and screen development and all the aesthetic arts involved. They also tend to have an interest in applications programming and all of the latest technologies surrounding it such as Java, Visual Basic, Cold Fusion, and all of the algorithmic challenges that solutions development offers. And of course, the Internet and its many flavors of technological know-how also tend to seduce the more ambitious among us engineers from what we consider to be the mundane tasks of data quality. After all, it is a safe bet that most IT professionals can count the number of people who consider themselves to be a "data quality engineer" on one hand. This is an unfortunate circumstance as some of the most challenging algorithms that the art of computer science has to offer can be found in the practice and implementation of data quality to a problem set that is more often than not unique from organization to organization. Also, as more and more exciting technologies begin to emerge in this area as we are seeing now, the momentum of the science should also increase.

Perhaps it is the word "quality" that causes many to cringe when the term is discussed, and another term ought to be interchanged with it to broaden its appeal. "Director of Data Management", "Data Effectiveness Researcher", and "Data Infrastructure Engineer" would be among the list of candidates that executive management may want to consider when assigning titles to what we all agree is a very important issue within the organization.

A second major reason that Data Quality is often ignored is because of the general consensus that the *implementation of data quality procedures and tools is a costly and resource-intensive undertaking*. While it is certainly not disputed that the science of data quality can and should extend beyond software technology, there are a great deal of software solutions that are available in the industry that will take an organization a long way in overcoming their data quality challenges, and enable it to enjoy the fruits and savor the success that these applications can deliver. Unfortunately, these solutions have traditionally been very expensive (in the six and seven figure ranges), very difficult to work with (engineers from the vendor as customization consultants long-term are the norm), encumbering to implement (data typically has been required to be exported off-line before even simple quality analysis and other various data quality functions can occur), and therefore have had very long implementation cycles (often 18 months). Doing the math clearly demonstrates that these kinds of solutions have traditionally been well out of reach financially from all but a very select few, and even more so for those of us who do not have the luxury of multi-year development schedules.

Fortunately, this is dramatically changing, and data quality technologies are emerging that eliminate practically all of the aforementioned barriers to introducing data quality into the development life cycle, as well as practically anywhere else along the information technology landscape. It is important to point out that since barriers such as high cost of entry, usability, lack of piece-wise availability, and time of implementation are being eliminated, the idea of high Data Quality is permeating out of the enterprise IT departments. It is now found departmentally, on a project-by-project basis, and even in non-IT departments such as marketing, sales, and operations. This has allowed the technology to be utilized not only on the largest DB2 driven mainframe-oriented systems, but all the way down to Microsoft Access databases and at every platform step along the way. In fact, there are now free data quality analysis tools available to any organization that can be implemented in minutes on practically every platform. This can help an organization ascertain to what extent they may be facing data quality issues. Giving these free-of-charge analysis tools an opportunity to provide an information system a thorough data quality checkup is a no-brainer. Vendors have discovered that demonstrating data quality issues easily and quickly via technology are a pre-cursor to a customer acquiring further technology from the vendor that can resolve many of these same issues. SAS and DataFlux have pioneered a great deal in this area and these innovations are now currently available.

This is a good lead-in to the third primary reason that data quality is often overlooked, and that is the *inability to demonstrate clearly the return-on-investment of data quality technologies*. Since there are many different flavors and aspects of data quality, there are differing degrees to the extent of this situation. For example, undoubtedly one can easily calculate postage saved in a marketing database that has a duplicate record level of 30% of the records (the same individual or organization appearing multiple times within a dataset under different variations of their name and address) if those duplicates can be discovered and removed. However, it is often not so clear what dollar amount can be placed on data that is inaccurate or otherwise missing from a dataset, or to what degree data inconsistency within a dataset can affect the results of queries and reports that are used as the basis of decision-making, and what the financial impact is of poor decisions that are made as a result of faulty or inaccurate information. Since this is often the case and ROI analysis of data quality procedures are far from a science, persuasion of management and spending authorities can be a difficult challenge when resources are doled out from the top of the mountain.

Interestingly enough, data quality lends credence to the phenomenon that the most obvious and blatant solutions are often those most overlooked. Sometimes exercises in data quality impact analysis produce ROI figures that are so astronomical that they are simply dismissed as absurd.

To demonstrate this, let’s consider the example of a typical banking institution.

Research has shown that the average banking customer contributes $3000 per year to his or her bank (the culmination of mortgage interest, auto loans, banking fees, securities management, etc.), or $15,000 over a five-year period. Suppose a bank with a million customers implements data quality technology across the enterprise enabling a greater success in all data-driven activities such as customer interaction, marketing personalization, cross-selling, and accurate business analysis results in an increase of a mere 2% of new business. The resulting impact numbers can be astounding. This would be the equivalent of adding 20,000 new customers, multiplied by $15,000 over a five-year period, or $300 million dollars. And that is at a paltry 2%. It should be clear that while this is a simplified analysis, similar more-detailed analysis numbers that produce similar results often leave executives scratching their heads and thinking, “What have I missed here?” Usually, the answer is nothing.

THE RISKS AND COSTS OF NON-QUALITY DATA

The previous example is indicative of some of the processes that can reap huge benefits to an organization via data quality technology. However, inattention to data quality does not just result in missed opportunities, it can leave broad risk exposure and countless shortfalls from a data quality perspective, including lost dollars, lost customers, and even legal ramifications. The following are some examples that demonstrate the great pains that can be suffered if we continue to avoid data quality as a management priority.

- In a conservative estimate, more than 175,000 IRS and state tax refund checks were marked as ‘undeliverable’ by the post office last year.
- Three zeros inadvertently added and reported as a trade volume amount of an inside executive of a public company in Atlanta caused its stock to plummet over 30% in one day.

- After an audit, it was estimated that 15-20% of voters on voter registration lists have either moved or are deceased when compared to data gathered from post office relocation data.
- An acquiring company learned long after the deal was closed that their new consumer business only had ½ the customers as they thought because of the large presence of duplicate data in the customer database.
- A fiber-optics company lost $500,000 after a mislabeled shipment caused the wrong cable to be laid at the bottom of a lake.
- A mailing order company faced massive litigation because it was unable to catch bogus, insulting names from being entered into the catalog request section of their website that were ultimately mailed to unsuspecting and then humiliated receivers of the catalog.
- The US government estimates that billions of dollars are lost annually due to poor data quality.

It is staggering. Estimates have shown that 15-20% of data within an organization's databases can be erroneous or otherwise unusable, leading to an enormous effect on the bottom line.

The issue is compounding to a greater extent in the age of the Internet. We now have less control over the data collection mechanisms that feed our information systems such as website forms or B2B XML channels. These external data suppliers are growing more common by the day. Also, we have a much higher level of access to corporate data via the Internet by our customers, demanding a greater degree of accuracy still.

A CLOSER LOOK AT ACTUAL DATA QUALITY ISSUES

So far, we have discussed data quality and its ramifications in broad strokes. Now, the discussion turns to the kinds of basic data quality issues that IT professionals run into regularly, and how these can affect the results of the business intelligence applications that we rely on daily.

INCONSISTENTLY REPRESENTED DATA

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 business title:

VP Sales, Vice President Sales, V.P. Sales, Director of Sales, VP SALES

Or an operating system:

Windows NT, WINDOWS, WIN NT, Windows NT 5, Win NT 5.0

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.

Imagine a marketing campaign that personalizes its communication based upon the business title of the prospect. Variations in business titles 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.

User base platform analysis by a software company could produce improper results if the data looks like it does in the platform example cited.

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, 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.

DUPLICATE DATA

Another common example of a data quality issue is 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, 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 INTEGRATION

One of the close sisters of duplicate data is the issue of data integration. This becomes a data quality 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 VERIFICATION AND AUGMENTATION

One final area of data quality that will be discussed here 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 Service's 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.

OTHER DATA QUALITY ISSUES

In addition to the data quality issues discussed here, there are many others that should be given careful consideration by any organization who has invested in business intelligence applications. These include but are not limited to data structural incompatibilities, missing values, numerical data quality issues (can be identified using techniques of statistical analysis), cross-column-based correctness analysis (such as determination of correct gender from name data), and a whole lot more. In the aggregate, all of these issues if overlooked can become extremely hazardous to the health of a data-driven information system.

THE IMPERATIVE OF ASSESSING DATA QUALITY

Now that we have established a firm understanding of what Data Quality is, as well as its impact, it is now time to turn our thoughts to what can be done about it. As the old adage declares, every great journey begins with the first step. In the case of data quality, this first step is *data quality assessment*. The fact that a data quality assessment step is missing from the development plans of many project leaders is not only dumbfounding, it truly ought to be grounds for dismissal. Every system that has data relies on that data heavily (or else the data wouldn't be there) for its success, and anybody who eliminates data quality from the development process is acting irresponsibly. This may have been understandable in the past as analysis tools were not readily available to assess data quality, but now not only are they available, they are often available for free and require very little time to utilize. Now that we understand the impact of poor data quality, inattention to it is inexcusable.

In addition to the analysis and exception-finding tools available for this purpose, data quality assessment should at the very least consist of the following questions:

- How do we know our data isn't bad?
- What would constitute "bad data" from our perspective?
- If it were bad, what kind of effects might it have on our various systems?
- What points of data collection might cause data quality issues to occur?
- If we had a more accurate data foundation, what could we then accomplish?
- Where specifically are our costs of data imperfections high?
- What non-technology processes could be introduced to increase our level of data quality?
- What technology is available to assist us in our efforts?
- What quantifiable ways can we demonstrate ROI to executive management?

Attention to these questions as well as ones that pertain specifically to a certain type of system or business intelligence application will shed much light on how these issues can be contained if not resolved entirely. A meticulous data quality assessment will undoubtedly provide the blueprint for a solid data quality solution.

WHAT SAS AND DATAFLUX ARE DOING ABOUT DATA QUALITY

On June 9, 2000 SAS acquired the DataFlux Corporation in an effort to enhance the data infrastructure solutions currently offered to its customers, and underscore the importance of data quality as a cornerstone of any data-driven initiative.

DataFlux has developed many award-winning data quality technologies that are used throughout multiple industries today. These technologies tackle many of the issues that were discussed throughout this paper and many of those not discussed here, as well as providing easy to implement, cost-free data quality assessment as also emphasized herein. SAS has integrated some of these technologies currently, and will continue to integrate many more into SAS products such as the Data Warehouse Administrator, as well continuing to offer the DataFlux product line to the market.

More specifically, the DataFlux offering includes a suite of data quality and data integration tools that can assist significantly in the development of a bulletproof data foundation integral to any data-driven business intelligence endeavor. dfPower^{TM} Studio 4.0 is a comprehensive data quality and data integration solution that focuses on many data quality issues such as standardization, matching, data verification, de-duplication, data integration, accuracy, and data quality business rule management. These technologies are delivered via an intuitive interface, and are packaged with many other enabling technologies such as database access, providing an easy-to-use and easy-to-implement multi-faceted data quality solution.

For developers, DataFlux also provides its Blue Fusion^{TM} SDK, a series of functional libraries that contain many of the algorithms that comprise the core of dfPower Studio 4.0. The covers have been taken off and the algorithms delivered as components, allowing developers to build data quality directly into any application.

The two offerings also work very well together as a completely integrated solution, providing the ability to include data quality and better data integration at practically every point within the organization.

A GOLDEN OPPORTUNITY

In conclusion, it is very clear that data quality processes can have tremendous impact on the bottom line of an organization. Now that the available technologies in this area are gaining momentum and improving immensely with every new release date, the effects of poor data quality are no longer inescapable. Well thought-out data quality assessment combined with state-of-the-art data quality technology will take an organization a long way in achieving its goals of accuracy, consistency, usability, and completeness of all of its electronic data assets. An understanding and facilitation of these concepts will continue to be a bellwether for business intelligence success.

REFERENCES

English, Larry P. 1999. Improving Data Warehouse and Business Information Quality – Methods for reducing costs and increasing profits. Jon Wiley and Sons, Inc. New York, NY.

META Group, Inc. Stamford, CT

The author may be contacted at:

Bob Brauer
DataFlux Corporation
4001 Weston Parkway, Suite 300
Cary, North Carolina 27513
(919) 674-2153
Fax: (919) 678-8330
Email: bobb@dataflux.com
Web: www.dataflux.com