Data Warehousing Special Report: Data quality and the bottom line

5/1/2002

By Wayne W. Eckerson

During the past 50 years, the developed world has moved from an industrial economy to an information economy. Companies now compete on their ability to absorb and respond to information, not just manufacture and distribute products. Intellectual capital and know-how are more important assets than physical infrastructure and equipment.

If information is the currency of the new economy, then data is a critical raw material needed for success. Just as a refinery takes crude oil and transforms it into numerous petroleum products, companies use data to generate a multiplicity of information assets. These assets form the basis of the strategic plans and actions that determine a firm's success.

Consequently, poor quality data can have a negative impact on the health of a company. If not identified and corrected early on, defective data can contaminate all downstream systems and information assets.

The problem with data is that its quality quickly degenerates over time. Experts say 2% of records in a customer file become obsolete in a month because customers die, divorce, marry and move. In addition, data-entry errors, systems migrations and changes to source systems, among other things, generate bucketloads of errors. As well, as organizations fragment into different divisions and units, interpretations of data elements mutate to meet local business needs. A data element that one individual finds valuable may be nonsense to an individual in a different group.

The Data Warehousing Institute (TDWI) estimates that poor quality customer data costs U.S. businesses a staggering $611 billion a year in postage, printing and staff overhead (TDWI estimates based on cost-savings cited by survey respondents and others who have cleaned up name and address data, combined with Dun & Bradstreet counts of U.S. businesses by number of employees.). Frighteningly, the real cost of poor quality data is much higher. Organizations can frustrate and alienate loyal customers by incorrectly addressing letters or failing to recognize them when they call, or visit a store or Web site. Once a company loses

its loyal customers, it loses its base of sales and referrals, as well as future revenue potential.

Given the business impact of poor quality data, it is bewildering to see the casual way in which most companies manage this critical resource. Most companies do not fund programs designed to build quality into their data in a proactive, systematic and sustained manner. According to TDWI's Data Quality Survey, almost half of all firms have no plan for managing data quality.

Part of the problem is that most organizations overestimate the quality of their data and underestimate the impact errors and inconsistencies can have on their bottom line. On one hand, almost half of the companies who responded to our survey believe the quality of their data is "excellent" or "good." Yet more than one-third of the respondent companies think the quality of their data is "worse than the organization thinks."

Although some firms understand the importance of high-quality data, most are oblivious to the true business impact of defective or substandard data. Thanks to a raft of new information-intensive strategic business initiatives, executives are beginning to wake up to the real cost of poor quality data. Many have bankrolled high-profile IT projects in recent years -- data warehousing, CRM and e-business projects -- that have failed or been delayed due to unanticipated data-quality problems.

For example, in 1996, FleetBoston Financial Corp. (then Fleet Bank) in New England undertook a much publicized $38 million CRM project to pull together customer information from 66 source systems. Within three years, the project was drastically downsized and the lead sponsors and technical staff were let go. A major reason the project came unraveled was the team's failure to anticipate how difficult and time consuming it would be to understand, reconcile and integrate data from 66 different systems.

According to TDWI's Industry Study 2000 survey, the top two technical challenges firms face when implementing CRM solutions are "managing data quality and consistency" (46%) and "reconciling customer records" (40%). Considering that 41% of CRM projects were "experiencing difficulties" or "a potential flop," according to the same study, it is clear that the impact of poor data quality in CRM is far reaching ("Harnessing Customer Information for Strategic Advantage: Technical Challenges

and Business Solutions." A summary can be found at http://www.dw-institute.com/ download/2000_Industry_Study.pdf.).

Data warehousing, CRM and e-business projects often expose poor quality data because they require companies to extract and integrate data from multiple operational systems. Data that is sufficient to run payroll, shipping or accounts receivable is often peppered with errors, missing values and integrity problems that do not show up until someone tries to summarize or aggregate the data.

Also, since operating groups often use different rules to define and calculate identical elements, reconciling data from diverse systems can be a huge, and sometimes insurmountable, obstacle. Sometimes the direct intervention of the CEO is the only way to resolve conflicting business practices, or political and cultural differences.

Every firm, if it looks hard enough, can uncover a host of costs and missed opportunities caused by inaccurate or incomplete data. Consider the following:

* 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 due to duplicate customer records.

* An information services firm lost $500,000 annually and alienated customers because it repeatedly recalled reports sent to subscribers due to inaccurate data.

* A large bank discovered that 62% of its home-equity loans were being calculated incorrectly, with the principal getting larger each month.

* A health insurance company in the Midwest delayed a decision support system for two years because the quality of its data was "suspect."

* 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.

* A regional bank was unable to calculate customer and product profitability due to missing and inaccurate cost data.

In addition, new industry and government regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) and Bank Secrecy Act, are upping the ante. Organizations are now required to carefully

manage customer data and privacy or face penalties, unfavorable publicity and loss of credibility.

What can go wrong?
The sources of poor quality data are myriad. Leading the pack are data-entry processes, which produce the most frequent data quality problems, and systems interfaces.

Not surprisingly, survey respondents cite data-entry errors by employees as the most common source of data defects. Examples of errors include misspellings, transposition of numerals, incorrect or missing codes, data placed in the wrong fields and unrecognizable names, nicknames, abbreviations or acronyms. These types of errors are increasing as companies move their businesses to the Web and allow customers and suppliers to enter data about themselves directly into operational systems.

Lack of validation routines. Interestingly, many data-entry errors can be prevented through the use of validation routines that check data as it is entered into Web, client/server or terminal-host systems. Respondents to the TDWI survey mentioned a "lack of adequate validation" as a source of data defects, noting this grievance in the "Other" category.

Valid, but not correct. But even validation routines cannot catch typos where the data represents a valid value. Although a person may mistype a telephone number, the number recorded is still valid -- it is just not the right one. The same holds true for social security numbers, vehicle identification numbers, part numbers and last names. Database integrity rules can catch some of these errors, but firms need to create complex business rules to catch the rest.

Mismatched syntax, formats and structures. Data-entry errors are compounded when organizations try to integrate data from multiple systems. For example, corresponding fields in each system may use different syntax (first-middle-last name vs. last-first-middle name), data formats (6 byte date field vs. 4 byte date field), or code structures (male-female vs. m-f vs. 1-2). In these cases, either a data cleansing or ETL tool needs to map these differences to a standard format before serious data cleanup can begin.

Unexpected changes in source systems. Perhaps a more pernicious problem is structural changes that occur in source systems. Sometimes these changes are deliberate, such as when an administrator adds a new field or code value and then neglects to notify the managers of

connecting systems about the changes. In other cases, front-line people reuse existing fields to capture new types of information that were not anticipated by the application designers.

Spiderweb of interfaces. Because of the complexity of systems architectures today, changes to source systems are easily and quickly replicated to many other systems, both internal and external. Most systems are connected through a spiderweb of interfaces to other systems. Updating these interfaces is time-consuming and expensive, and many changes slip through the cracks and "infect" other systems. Thus, changes in source systems can wreak havoc on downstream systems if adequate change management processes are not in place.

Lack of referential integrity checks. It is also true that target systems do not adequately check the integrity of the data they load. For example, data warehouse administrators often turn off referential integrity when loading the data warehouse for performance reasons. If source administrators change or update tables, this can create integrity problems that are not detected.

Poor system design. Source or target systems that are poorly designed can create data errors. As companies rush to deploy new systems, developers often skirt fundamental design and modeling principles, which leads to data integrity problems down the road.

Data conversion errors. In the same vein, data migration or conversion projects can generate defects, as well as ETL tools that pull data from one system and load it into another. Although systems integrators may convert databases, they often fail to migrate business processes that govern the use of data. In addition, programmers may not take the time to understand source or target data models, and may therefore write code that introduces errors. One change in a data migration program or system interface can generate errors in tens of thousands of records.

The fragmentation of definitions and rules. A much bigger problem comes from the fragmentation of our organizations into a multitude of departments, divisions and operating groups, each with its own business processes supported by distinct data management systems. Slowly and inexorably, each group begins to use slightly different definitions for common data entities -- such as "customer" or "supplier" -- and apply different rules for calculating values, such as "net sales" and "gross profits." Add mergers, acquisitions and global expansion into countries

with different languages and customs, and you have a recipe for a data-quality nightmare.

The problems that occur in this scenario have less to do with accuracy, completeness, validity or consistency, than with interpretation and protecting one's "turf." That is, people or groups often have vested interests in preserving data in a certain way even though it is inconsistent with the way the rest of the company defines data.

For example, many global companies squabble over a standard for currency conversions. Each division in a different part of the world wants the best conversion rate possible. And even when a standard is established, many groups will skirt the spirit of the standard by converting their currencies at the most opportune times, such as when a sale was posted vs. when the money was received. This type of maneuvering wreaks havoc on a data warehouse that tries to accurately measure values over time.

Slowly changing dimensions. Similarly, slowly changing dimensions can result in data-quality issues depending on the expectations of the user viewing the data. For example, an analyst at a chemical company wants to calculate the total value of goods purchased from Dow Chemical for the past year. But Dow recently merged with Union Carbide, which the chemical company also purchases materials from.

In this situation, the data warehousing manager needs to decide whether to roll up purchases made to Dow and Union Carbide separately, combine the purchases from both firms throughout the entire database, or combine them only after the date the two companies merged. Whatever approach the manager takes, it will work for some business analysts and alienate others.

In these cases, data quality is a subjective issue. Users' perception of data quality is often colored by the range of available data resources they can access. Where there is "competition" -- another data warehouse or data mart that covers the same subject area -- knowledge workers tend to be pickier about data quality, said Michael Masciandaro, director of decision support at Rohm & Haas.

Delivering high-quality data

Given the ease with which data defects can creep into systems, especially data warehouses, maintaining data quality at acceptable levels takes considerable effort and coordination throughout an organization. "Data quality is not a project, it's a lifestyle," said David Wells, enterprise

systems manager at the University of Washington and the developer of TDWI's full-day course on data cleansing ("TDWI Data Cleansing: Delivering High Quality Warehouse Data").

And progress is not always steady or easy. Improving data quality often involves exposing shoddy processes, changing business practices, gaining support for common data definitions and business rules, and delivering education and training. In short, fixing data quality often touches a tender nerve on the underbelly of an organization.

One top executive leading a data-quality initiative said, "Improving data quality and consistency involves change, pain and compromise. The key is to be persistent and get buy-in from the top. Tackle high ROI projects first, and use them as leverage to bring along other groups that may be resistant to change."

The University of Washington's Wells emphasizes that managing data quality is a never-ending process. Even if a company gets all the pieces in place to handle today's data-quality problems, there will be new challenges tomorrow. That is because business processes, customer expectations, source systems and business rules all change continuously.

To ensure high-quality data, firms need to gain broad commitment to data-quality management principles and develop processes and programs that reduce data defects over time. To lay the foundation for high-quality data, firms need to adhere to the methodology outlined below.

Step 1. Launch a data quality program. The first step to delivering high-quality data is to get top managers to admit there is a problem and take responsibility for it.

The best way to kickstart a data-quality initiative is to fold it into a corporate data stewardship or data administration program. These programs are typically chartered to establish and maintain consistent data definitions and business rules so the firm can achieve a "single version of the truth" and save time on developing new apps and looking for data.

Step 2. Develop a project plan. The next step is to develop a data-quality project plan or series of plans. A project plan should define the scope of activity, set goals, estimate ROI, perform a gap analysis, identify actions, and measure and monitor success. To perform these tasks, the team will need to dig into the data to assess its current state, define corrective actions and establish metrics for monitoring conformance to goals.

Step 3. Build a data-quality team. Organizations must assign or hire individuals to create the plan, perform initial assessment, scrub the data and set up monitoring systems to maintain adequate levels of data quality.

Step 4 and Step 5. Review business processes and data architecture. Once there is corporate backing for a data-quality plan, the stewardship committee -- or a representative group of senior managers throughout the organization -- needs to review the company's business processes for collecting, recording and using data in the subject areas defined by the scope document. With help from outside consultants, the team also needs to evaluate the underlying systems architecture that supports the business practices and information flows.

Step 6. Assess data quality. After reviewing information processes and architectures, an organization needs to undertake a thorough assessment of data quality in key subject areas. The purpose of the assessment is to identify common data defects; create metrics to detect defects as they enter the data warehouse or other systems; and create rules or recommend actions for fixing the data. This can be long, arduous and labor-intensive work, depending on the scale and scope of the project, as well as the age and cleanliness of the source files.

Step 7. Clean the data. Once the audit is complete, the job of cleaning the data begins. A fundamental principle of quality management is to detect and fix defects as close as possible to the source to minimize costs.

Prevention is the least costly response to defects, followed by correction and repair. Correction involves fixing defects in-house, while repair involves fixing defects that affect customers directly. Examples of repair are direct mail pieces that are delivered to a deceased spouse, or software bugs in a commercially available product.

Step 8. Improve business practices. As mentioned earlier, preventing data defects involves changing attitudes and optimizing business processes. "A data quality problem is a symptom of the need for change in the current process," said Brad Bergh, a veteran database designer with Double Star Inc. Improving established processes often stokes political and cultural fires, but the payoff for overcoming these challenges is great.

Having a corporate data stewardship program and an enterprise-wide commitment to data quality is critical to making progress. Under the

auspices of the CEO and the direction of corporate data stewards, a company can begin to make fundamental changes in the way it does business to improve data quality.

Step 9. Monitor data continuously. Organizations can quickly lose the benefits of their data preparation efforts if they fail to monitor data quality continuously. To do this, companies need to build a program that audits data at regular intervals, or just before or after data is loaded into another system such as a data warehouse. Companies then use the audit reports to measure their progress in achieving data-quality goals and complying with service-level agreements negotiated with business groups.

Service-level agreements should specify tolerances for critical data elements and penalties for exceeding those tolerances.

The above techniques, although they are not easy to implement in all cases, can help bring a company closer to achieving a strong foundation on which to build an information-based business. The key is to recognize that managing data quality is a perpetual endeavor. Companies must make a commitment to build data quality into all information management processes if they are going to reap the rewards of high-quality data -- and avoid the pitfalls caused by data defects.

Wayne W. Eckerson is director of education and research for The Data Warehousing Institute, where he oversees TDWI's educational curriculum, member publications, and various research and consulting services. He has published and spoken extensively on data warehousing and business intelligence subjects since 1994.

---

back to previous page