A Practical Perspective on Data Quality Issues

Shirley Becker, Associate Editor-Industry & Practice
American University

Why Be Concerned About Data Quality?

Approximately 90% of the respondents of an informal survey on data warehousing opportunities readily identified data quality problems in their database systems. The respondents included administrators, software engineers, sales personnel, documentation and technical support, among others. The data quality issues ranged from inadequate search capabilities to major integration problems. The survey results are supported by the increasing popularity of data warehousing techniques that may be attributed at least in part to the ongoing search for higher-quality data.

One of the respondents, a former US government employee, succinctly stated that data problems may be viewed as “noise” making it difficult to obtain necessary and sufficient data. He pointed out that, “A database contains an important message which is sometimes hard to decipher from the background noise.” Other respondents described their quality problems as the result of missing, incorrect, or incomplete data thus, impacting their decision-making capabilities. Format incompatibility, missing data/documentation, and vague or unknown data definitions were identified as factors impacting timely and meaningful data retrieval.

What is the impact of poor data quality on an organization? Obviously it depends on the type of data quality issues that an organization faces and the cost associated with them. Though most of the respondents could identify data quality problems, less than 15% indicated that some type of cost/benefit analysis had been done to resolve these problems. Perhaps this disparity is best signified by one respondent’s statement, “No costs have been associated with the quality of data. (But) it is imperative to us that we have all the correct data.”

The identification of costs associated with data quality requires an understanding of the factors that can impact data design and implementation. For example, a respondent identified a potential cost associated with “sifting” through data that contains “irrelevant” or “difficult to interpret” components. This cost, though undocumented, may have resulted from poor design work such as ill-defined data and supporting documentation and/or incorrect or missing relationships and constraints. It may also be the result of a poor implementation strategy that lacked support for data integrity. Without proper data analysis, it is difficult to determine the cost implications and potential benefits of quality improvements.

The following scenario presented by a respondent demonstrates a hidden cost of data duplication and maintenance that may not be visible at higher levels in an organization. “I find myself duplicating data that is already in the (main) database because there is no easy way to retrieve the information in a form that is meaningful for my department. I can’t create (the reports) that I need ... Someone may be able to access this information, but it isn’t me ... So, I gather it all by hand, put it on my computer and manipulate the data from there.”

What Types of Problems Impact Data Quality?

One respondent stated, “I can get data but it isn’t exactly what I requested!” This and other types of problems identified in the survey are categorized as follows:

- **Data corruption due to incorrect conversion.** Several respondents identified data problems when spreadsheet, text files, and old file structures are converted to a relational format. During the conversion process, data may be corrupted or lost due to a lack of understanding of the old data and its relationships.
- **Historical and current data have different meanings.** Supporting user and system documentation need to be updated to reflect changing user requirements. Yet, this is often not the case. A reengineered database system was found to contain data values with different meanings stored in the same column. There was no supporting documentation to interpret the change in data definitions.
- **The same data has more than one data definition.** Data may appear in more than one table but with different representations making it difficult if not impossible to manipulate the data correctly. For example, columns containing social security numbers may be various lengths (9 no dashes, 11 with dashes) with different data types (numeric, character, or varchar). One respondent had identified seventeen different data definitions for a customer name.

Twenty-two survey respondents completed an informal questionnaire on data quality issues and data warehouse concepts.
Related Content

Situational Method Engineering to Support Process-Oriented Information Logistics: Identification of Development Situations
Tobias Bucher and Barbara Dinter (2012). *Journal of Database Management* (pp. 31-48).
[www.igi-global.com/article/situational-method-engineering-support-process/62031?camid=4v1a](www.igi-global.com/article/situational-method-engineering-support-process/62031?camid=4v1a)

Towards Real-Time Multi-Sensor Golf Swing Classification Using Deep CNNs
[www.igi-global.com/article/towards-real-time-multi-sensor-golf-swing-classification-using-deep-cnns/218925?camid=4v1a](www.igi-global.com/article/towards-real-time-multi-sensor-golf-swing-classification-using-deep-cnns/218925?camid=4v1a)

Cache Management for Web-Powered Databases
Dimitrios Katsaros and Yannis Manolopoulos (2003). *Web-Powered Databases* (pp. 203-244).
[www.igi-global.com/chapter/cache-management-web-powered-databases/31429?camid=4v1a](www.igi-global.com/chapter/cache-management-web-powered-databases/31429?camid=4v1a)

Agile Software Development in Practice
[www.igi-global.com/chapter/agile-software-development-practice/28431?camid=4v1a](www.igi-global.com/chapter/agile-software-development-practice/28431?camid=4v1a)