

Improving the Data Warehouse: Metadata Management, Data Cleansing and Information Stewardship

Research Paper Author: Brian Evans—2005 Graduate, IT Systems Analyst, Mentor Graphics Corporation

In Brief: Market leaders know that organizational data are strategic assets. Technology-savvy organizations are leveraging data to drive strategic decision-making and organizational knowledge, which are the core capabilities of successful organizations.

A data warehouse is intended to provide valuable information for strategic decision-making. However its Achilles heel is poor data quality, which costs American businesses more than \$600 billion annually, and consumes approximately 10 percent of a typical organization's revenue.

High profile news stories in recent years such as the 2000 U.S. Presidential election, incorrect pricing on Amazon's website, and delayed toy shipments from Toys 'R Us illustrate just a few of the crippling effects of poor data quality.

Quality Data is:	
Accurate	Data fall within an acceptable range of values
Current	Data are up-to-date
Complete	Data are populated
Consistent	Data maintain values across entities

Figure 1: Data quality indicators

Organizations also require high quality information to accommodate rapidly changing business needs. A data quality program utilizing proven data quality techniques can greatly improve the strategic value and success of a data warehouse. A brief discussion of the key quality technique follows:

Metadata Management: Data quality tools can help manage metadata quality—in particular reverse engineering tools, data profiling tools, data monitoring tools and metadata management and quality tools. The ability of an IT data warehousing team to manage all forms of corporate-wide metadata may be hindered by time and resource constraints.

Data Cleansing: Three key decisions relate to the use of data cleansing for a data warehouse. The first decision is whether data cleansing is most appropriate at the source system, during the ETL process, at the staging database, or within the data warehouse. Another is to decide what data elements to cleanse and how to cleanse them. Finally, the IT data warehousing professional must be aware that not all data cleansing steps can be automated. The role of data cleanup coordinator is an important role for data cleansing activities. While the role of tools is limited, the data cleanup coordinator



may utilize data cleansing tools, and the information quality analyst may use metadata quality reporting tools.

Information Stewardship:

Information stewardship is primarily a human resource endeavor in that an IT data warehousing professional must consider how the data quality policy and guidelines will be communicated and distributed.

Abstract: The corporate data warehouse provides strategic information to support decision-making (Kimball,

et al., 1998). High quality data may be the most important factor for data warehouse success (Loshin, 2003).

This study examines three data management techniques that improve data quality: metadata management, data cleansing, and information stewardship. Content analysis of 14 references, published between 1992 and 2004, results in lists of themes, synonyms, and defini-

tions for each technique, designed for data warehouse analysts and developers.

Although a data warehouse is intended to provide valuable information for strategic decision-making, its Achilles heel is poor data quality.

References: (Selected citations only)

- Brackett, Michael H. (1996). *The Data Warehouse Challenge: Taming Data Chaos*. New York: Wiley Computer Publishing.
- Eckerson, Wayne W. (2002). "Data Quality and the Bottom Line". TDWI [Online]. Retrieved September 13, 2005 from TDWI's Data Quality Report.
- English, Larry P. (1999). *Improving Data Warehouse and Business Information Quality*. New York: Wiley Computer Publishing.
- Kelly, Sean. (1995). *Data Warehousing: The Route to Mass Customization*. New York: John Wiley & Sons.
- Kimball, Ralph, & Caserta, Joe. (2004). *The Data Warehouse ETL Toolkit*. New York: Wiley Computer Publishing.
- Loshin, David. (2003). *Business Intelligence*. San Francisco: Morgan Kaufmann Publishers.
- Redman, Thomas C. (2001). *Data Quality: The Field Guide*. Boston: Digital Press.



UNIVERSITY OF OREGON

<http://aimdegree.com/research/ebriefings/>

(800) 824-2714 • (541) 346-4231 • aim@uoregon.edu