Process-Embedded Data Integrity

Yang W. Lee, Northeastern University, USA
Leo Pipino, University of Massachusetts, Lowell, USA
Diane M. Strong, Worcester Polytechnic Institute, USA
Richard Y. Wang, Massachusetts Institute of Technology, USA

ABSTRACT

Despite the established theory and the history of the practical use of integrity rules, data quality problems, which should be solvable using data integrity rules, persist in organizations. One effective mechanism to solve this problem is to embed data integrity in a continuous data quality improvement process. The result is an iterative data quality improvement process as data integrity rules are defined, violations of these rules are measured and analyzed, and then the rules are redefined to reflect the dynamic and global context of business process changes. Using action research, we study a global manufacturing company that applied these ideas for improving data quality as it built a global data warehouse. This research merges data integrity theory with management theories about quality improvement using a data quality lens, and it demonstrates the usefulness of the combined theory for data quality improvement.

Keywords: data quality; data integrity; action research; data warehouse

INTRODUCTION

Data integrity rules are the recommended approach for enforcing data quality in operational databases. These rules are well-grounded in relational database theory (Codd, 1970, 1990) and are widely used. Despite established theory and history of the practical use of integrity rules, data quality problems persist in organizations (Becker, 1998; Brodie, 1980; Preston, 2001; Price, 1994; Tayi & Ballou, 1999; Segev, 1996; Huang et al. 1999).

The failure to link integrity rules to organizational changes is among the reasons that data quality problems plague organizations. In essence, conventional practice is to view the application of data integrity as a one-time static process applied when data enters the database. We pro-
pose that the application of data integrity be viewed as a dynamic, continuous process, embedded in an overall data quality improvement process. We develop such an embedded data integrity process and demonstrate its usefulness to a real organization.

Similar issues are apparent in data warehouses. A common difficulty with data warehousing is the poor quality of data in operational databases, the sources of data for data warehouses (Ballou & Tayi, 1989; Celko & McDonald, 1995). If poor-quality data enter the data warehouse, they can impede global access to information and knowledge (Aho, 1996). Similarly, software tools used to achieve data quality (Brown, 1997; CRG, 1997, 1998) should be applied within a coherent and systematic improvement process, but are typically applied in an ad hoc fashion.

A key challenge, then, is to understand how data integrity principles can be embedded in an ongoing continuous data-quality improvement process. While the literature provides principles of data quality improvement, data integrity, and software tools, we can only fully understand their appropriate use through the dynamic instantiation in an organization.

PROCESS-EMBEDDED DATA INTEGRITY

Total Data Quality Management (TDQM) Cycle

Total Quality Management (TQM), a practical approach for improving quality (Deming, 1986; Juran & Godfrey, 1999), explicitly links quality to a continuous improvement process. The Total Data Quality Management (TDQM) Cycle (Madnick & Wang, 1992), an adaptation of TQM principles to the context of data, consists of defining, measuring, analyzing, and improving data quality through multiple, continuous improvement cycles (Figure 1).

First, we define data quality as data that are fit for use (Strong et al., 1997), following the quality definition in the TQM literature. To data consumers, fitness for use means data that are accurate, believable, objective, relevant, timely, reputable, value-added, appropriate in amount, concisely represented, consistently represented, complete, interpretable, accessible, understandable, and secure (Wang & Strong, 1996). Typically, in a data quality improvement project, a subset of these dimensions is chosen.

Second, we measure data quality along the chosen dimensions using metrics (Pipino et al., 2002). Consider accuracy, which means that stored data values do not differ from “true” data values. Since determining true values is costly, an appropriate accuracy metric is a value in an appropriate range. Iteration between the definition and measurement steps serves to develop metrics that are measurable at a reasonable cost and useful for improvement activities.

Third, analyze is interpreting the measures and deciding whether and how to improve the quality of data. Measures of

Figure 1: The TDQM Cycle
Related Content

Efficient Placement and Processing in Shared-Nothing Data Warehouses
[www.irma-international.org/chapter/efficient-placement-processing-shared-nothing/7023/](www.irma-international.org/chapter/efficient-placement-processing-shared-nothing/7023/)

Online Analytical Mining of Path Traversal Patterns for Web Measurement
[www.irma-international.org/article/online-analytical-mining-path-traversal/3286/](www.irma-international.org/article/online-analytical-mining-path-traversal/3286/)

Integrative Document and Content Management Systems’ Architecture
[www.irma-international.org/chapter/integrative-document-content-management-systems/11162/](www.irma-international.org/chapter/integrative-document-content-management-systems/11162/)

Approximate Computation of Distance-Based Queries
[www.irma-international.org/chapter/approximate-computation-distance-based-queries/29662/](www.irma-international.org/chapter/approximate-computation-distance-based-queries/29662/)

Building an Environmental GIS Knowledge Infrastructure
[www.irma-international.org/chapter/building-environmental-gis-knowledge-infrastructure/7942/](www.irma-international.org/chapter/building-environmental-gis-knowledge-infrastructure/7942/)