Depicting Data Quality Issues in Business Intelligence Environment through a Metadata Framework

Te-Wei Wang, University of Illinois at Springfield, Springfield, IL, USA
Yuriy Verbitskiy, University of South Australia, Adelaide, Australia
William Yeoh, University of South Australia, Mawson Lakes, Australia and Deakin University, Burwood, Australia

ABSTRACT

Modern business intelligence systems depend highly on high quality data. The core of data quality management is to identify all possible sources of data quality problems. To achieve this goal, an extensive metadata infrastructure is the most promising solution. Through theoretical metadata model investigation, the authors identified a set of data quality dimensions by carefully examining the data quality management principles and applied those principles to current BI environment. They summarize their analysis by proposing a BI data quality framework.

KEYWORDS

Business Intelligence, Data Quality, Framework, Metadata

INTRODUCTION

Recently Business Intelligence (BI) market has experienced high growth and BI technologies have consistently received attention by many Chief Information Officers (Gartner, 2015). BI is “a broad category of technologies, applications, and processes used for gathering, storing, accessing, and analysing data to help its users make better decisions.” In a typical BI environment, data from different sources are extracted, transformed and loaded (ETL) into an Enterprise Data Warehouse (EDW) and, from there, are used for reporting across the organization (see Figure 1). During this process, data quality plays a critical role in BI success since poor data quality can hinder business decisions at all levels of the organization (Daniel et al. 2008; Khatri and Brown, 2010). Being very complex at the same time, data quality is an issue that costs billions of dollars (Khatri and Brown, 2010; Vassiliadis, 2000) because data used in the BI environment are used for decision making at every organizational levels and various business processes (Ballou and Tayi, 1999).

In general, there are two main sources of data quality issues involved in BI projects. The first is rooted in the source systems. There are always data quality issues where data do not conform to business requirements even after extensive testing by business users and IT developers. The second
source of data quality problem resides on the entire BI processes. By integrating data from the source systems BI projects create new requirements for the existing data. Now the data should conform not only to the original source system requirements but also to new business requirements gathered for the BI project (Ballou and Tayi, 1999). It is a major task of the BI processes to drive more data exploration and analysis, thereby exposing all existing data quality issues, even those which were not considered before. BI processes may themselves become the source of data quality issues because they involve very complex operations with the data during all BI stages.

In a typical BI process such as the one showing in Figure 1, the ETL represents the first key challenge for any BI project, “70% of the risk and effort in the DW/BI project comes from this step” (Kimball et al., 2008). The main reason for this is because during this step data with a different structure and designed for a different operational purpose are brought into one place and transformed in such a way as to enable them to be integrated and used together (Ballou and Tayi, 1999). It is challenging to address varying levels of data quality across source systems or even within one source system (Daniel, 2008).

The second risk for data quality is the database designed for the EDW. Database design for an EDW mainly depends on the business requirements for decision making, the structure of the available data in the source systems and quality levels of the source data. The database design process is extremely complex in its own right involving understanding of the source data and related business processes, analysing business requirements, and ensuring naming conventions and data integration through using conforming dimensions and facts.

The next step is report building. Production of reports may not seem risky in terms of data quality if the data in the EDW conform to business requirements, but this BI process has its own reasons for concern. Full understanding of original business requirements for the EDW data set, EDW data set structure, related business processes and specific report requirements are the main factors for a successful report development process. Merely to understand the data structure in the EDW can be a challenge since it usually stores multidimensional, integrated data that were designed to be used for different purposes. Presenting the data in the simplest way is another challenge. It requires solid knowledge of the available BI Reporting tools.
Intellectual Capital Management in Long-Lasting Family Firms: The DuPont Case
[www.igi-global.com/chapter/intellectual-capital-management-long-lasting/36464?camid=4v1a](www.igi-global.com/chapter/intellectual-capital-management-long-lasting/36464?camid=4v1a)

Uncovering Actionable Knowledge in Corporate Data with Qualified Association Rules
Nenad Jukic, Svetlozar Nestorov, Miguel Velasco and Jami Eddington (2013). *Principles and Applications of Business Intelligence Research* (pp. 210-229).
[www.igi-global.com/chapter/uncovering-actionable-knowledge-corporate-data/72572?camid=4v1a](www.igi-global.com/chapter/uncovering-actionable-knowledge-corporate-data/72572?camid=4v1a)

Clinical Business Intelligence to Prevent Stroke Accidents
[www.igi-global.com/chapter/clinical-business-intelligence-to-prevent-stroke-acidents/146071?camid=4v1a](www.igi-global.com/chapter/clinical-business-intelligence-to-prevent-stroke-acidents/146071?camid=4v1a)