An Experimental Replication With Data Warehouse Metrics

Manuel Serrano, University of Castilla-La Mancha, Spain
Coral Calero, University of Castilla-La Mancha, Spain
Mario Piattini, University of Castilla-La Mancha, Spain

ABSTRACT

Data warehouses are large repositories that integrate data from several sources for analysis and decision support. Data warehouse quality is crucial, because a bad data warehouse design may lead to the rejection of the decision support system or may result in non-productive decisions. In the last years, we have been working on the definition and validation of software metrics in order to assure data warehouse quality. Some of the metrics are adapted directly from previous ones defined for relational databases, and others are specific for data warehouses. In this paper, we present part of the empirical work we have developed in order to know if the proposed metrics can be used as indicators of data warehouse quality. Previously, we have developed an experiment and its replication, and in this paper, we present the second replication we have made with the purpose of assessing data warehouse maintainability. As a result of the whole empirical work, we have obtained a subset of the proposed metrics that seem to be good indicators of data warehouse quality.

Keywords: data warehouse quality; design metric; empirical validation

INTRODUCTION

At the present time, most of the organizations face a serious problem of data pollution (Kelly, 1997), because they have great amounts of data collected at a relatively low cost that do not provide information. Nevertheless, companies must manage the information like a product of primary importance; they must capitalize on the knowledge like a main asset; in this way, they will be able to survive and to prosper in the digital economy (Huang et al., 1999). With this aim, data warehouses arose a few years ago.
Data warehouses were created to hold data drawn from several data sources and maintained by different operating units, together with historical and summary transformations. A data warehouse is a collection of technologies aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions.

Due to the increasing complexity of the data warehouses (Inmon, 1997), it is necessary to pay continuous attention to the evaluation of its quality throughout the development process. As Bouzeghoub and Kedad (2002) remark, quality in data warehouses is crucial.

A first step to obtain data warehouses with quality was the appearance of development methodologies like the proposals in Anahory and Murray (1992), Debevoise (1999), and Kimball et al. (1998).

But using development methodologies is not sufficient to assure the quality of data warehouses. Unfortunately, most of the works related to quality are focused on software quality (Arthur, 1992; Gilles, 1992; Ginac, 1998; ISO, 2001; Jones, 1997; Oskarsson & Glass, 1996), and the aspect most studied has been the program quality, disregarding the database quality (Sneed & Foshag, 1998). Even for the traditional design of databases, aspects regarding quality are not incorporated explicitly (Wang et al., 1993). All these reasons make it necessary to complement the specific methodologies with techniques, procedures, and specific metrics.

With this goal in mind, we have defined a set of metrics for measuring data warehouse star schemas in order to control its quality. Although the quality of a data warehouse depends on several factors (Jarke et al., 2000), we present in this paper the work we are developing for the logical (star) schema level.

Once the metrics have been proved as useful metrics, the data warehouse designer will be able to use them, for example, for selecting among different alternative schemas that are semantically equivalents. So, metrics could be used as design guidelines, not in the sense that they tell the designer the next step but in the sense that they can give the designer very useful information for making the best design decisions.

However, it is evident that when giving this powerful tool to designers, it is fundamental to ensure that metrics are really useful for the goal they are supposed to achieve. In this way, a methodological definition of the metrics is necessary to assure that metrics are useful for the goal for which they are intended. We have defined a method for defining valid and useful metrics that involves five main phases (identification, creation, acceptance, application, and accreditation). In this paper, we have focused only on the first two steps of the method (identification of requisites and metrics creation).

In the next section, we briefly present the method used to define metrics; in the third, fourth, and fifth sections, we will explain how we have used the method for the definition of data warehouse metrics. The conclusion is in the last section.

METHOD FOR DEFINING METRICS

Metric definition should be based on clear measurement goals. Metrics should be defined according to an organization’s needs that are related to external quality attributes. Figure 1 presents the method we apply for obtaining correct metrics. This method has been developed using the method proposed by Calero et al. (2001b) and the MMLC (Measure Model Life Cycle) method.
Related Content

User Segmentation Based on Twitter Data Using Fuzzy Clustering
Basar Öztaysi and Sezi Çevik Onar (2013). Data Mining in Dynamic Social Networks and Fuzzy Systems (pp. 316-333).
www.irma-international.org/chapter/user-segmentation-based-twitter-data/77533/

A Lattice-Based Framework for Interactively and Incrementally Mining Web Traversal Patterns
Yue-Shi Lee and Show-Jane Yen (2008). Data Mining and Knowledge Discovery Technologies (pp. 72-96).
www.irma-international.org/chapter/lattice-based-framework-interactively-incrementally/7514/

Managing Late Measurements in Data Warehouses
Matteo Golfarelli and Stefano Rizzi (2007). International Journal of Data Warehousing and Mining (pp. 51-67).
www.irma-international.org/article/managing-late-measurements-data-warehouses/1793/

Activity-Based Travel Demand Forecasting Using Micro-Simulation:
Stochastic Error Investigation of FEATHERS Framework
Qiong Bao, Bruno Kochan, Tom Bellemans, Davy Janssens and Geert Wets (2014). Data Science and Simulation in Transportation Research (pp. 167-181).
www.irma-international.org/chapter/activity-based-travel-demand-forecasting-using-micro-simulation/90071/

Mining Customer Knowledge for a Recommendation System in Convenience Stores
Shu-Hsien Liao, Chih-Hao Wen, Pei-Yuan Hsian, Chien-Wen Li and Che-Wei Hsu (2014). International Journal of Data Warehousing and Mining (pp. 55-86).
www.irma-international.org/article/mining-customer-knowledge-for-a-recommendation-system-in-convenience-stores/110386/