Metrics for Data Warehouse Quality

Manuel Serrano
University of Castilla-La Mancha, Spain

Coral Calero
University of Castilla-La Mancha, Spain

Mario Piattini
University of Castilla-La Mancha, Spain

INTRODUCTION AND BACKGROUND

It is known that organizations are very rich in data but poor in information. Today, technology has made it possible for organizations to store vast amounts of data obtained at a relatively low cost, however these data fail to provide information (Gardner, 1998). Data warehouses have appeared as a solution to this problem supporting decision-making processes and new kinds of applications as marketing.

A data warehouse is defined as a "collection of subject-oriented, integrated, non-volatile data that supports the management decision process" (Inmon, 1997). Data warehouses have become the key trend in corporate computing in the last years, since they provide managers with the most accurate and relevant information to improve strategic decisions. Also, the future for data warehouse is promising. Jarke, Lenzerin, Vassilou, and Vassiliadis (2000) forecast a market of 12 million U.S. dollars for data warehouse markets for the next few years. However, the development of a data warehouse is a difficult and very risky task. It is essential that we can assure the information quality of the data warehouse as it becomes the main tool for strategic decisions (English, 1999).

Information quality of a data warehouse comprises data warehouse system quality and presentation quality (see Figure 1). In fact, it is very important that data in a data warehouse reflect correctly the real world, but it is also very important that data can be easily understood. In data warehouse system quality, as in an operational database (Piattini, Genero, Calero, Polo, & Ruiz, 2000), three different aspects could be considered: DBMSs quality, data model quality, and data quality.

In order to assess DBMS quality, we can use an international standard like ISO 9126 (ISO, 1999), or some of the existing product comparative studies. This type of quality should be addressed in the product selection stage of the data warehouse life cycle.

Data quality must address mostly in the extraction, filtering, cleaning and cleansing, synchronization, aggregation, loading, and so forth, activities of the life cycle. In the last few years, very interesting techniques have been proposed to assure data quality (Bouzeghoub & Kedad, 2002).

Last but not least, data warehouse model quality has a great influence in the overall information quality. The designer must choose the tables, processes, indexes and data partitions representing the logical data warehouse and facilitating its functionality (Jarke et al., 2000).

Figure 1. Information and data warehouse quality

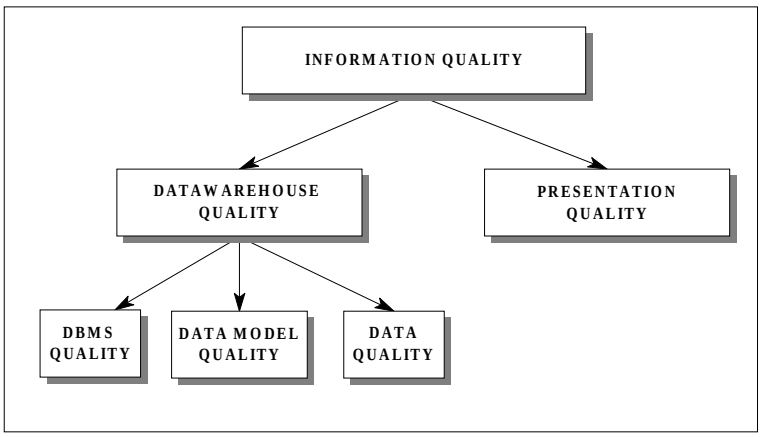
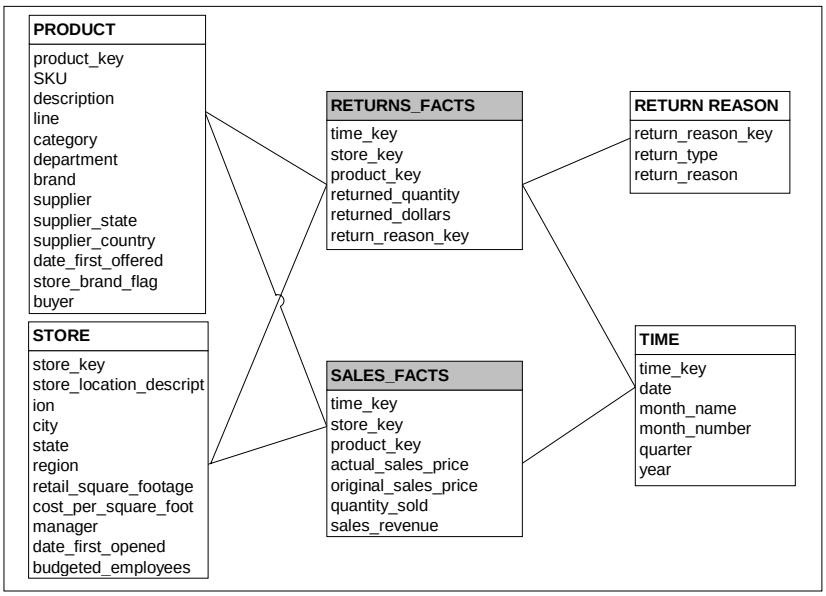


Figure 2. Example of a multidimensional data model design



Multidimensional data models are used to design data warehouses (Petersen & Jensen, 2001). A multidimensional data model is a direct reflection of the manner in which a business process is viewed. It captures the measurements of importance to a business, and the parameters by which the measurements are broken out. The measurements are referred to as *fact* or *measures*. The parameters by which a fact can be viewed are referred to as *dimensions* (Adamson & Venerable, 1998).

Usually multidimensional data models are represented as star schemas, which consist of one central table and several dimensional tables. The measures of interest are stored in the fact table (e.g., sales, inventory). For each dimension of the multidimensional model, there exists a dimensional table (e.g., product, time) that stores the information about the dimension (Jarke et al., 2000).

In Figure 2, we present an example of multidimensional data model design, with two fact tables (Returns_Facts and Sales Facts) and four dimensional tables (Product, Store, Return_Reason and Time).

In recent years, different authors have proposed some useful guidelines for designing multidimensional data models (Bouzeghoub & Kedad, 2002; Jarke et al., 2000; Vassiliadis, 2000). However, more objective indicators are needed to help designers and managers to develop quality multidimensional data models (Hammergren, 1996; Kelly, 1997; Kimball, Reeves, Ross, & Thornthwaite, 1998). Also, interesting recommendations for achieving a "good" multidimensional data model have been suggested (Adamson & Venerable, 1998; Inmon, 1997; Kimball et al., 1998), but quality criteria are not enough on their own to ensure quality in practice, as different people will generally have different interpretations of the same criteria. The goal should be to replace intuitive notions of design "quality" with formal, quantitative, objective metrics in order to reduce subjectivity and bias in the evaluation process.

The definition of a set of objective metrics for assuring multidimensional data model quality is the final aim of our work. As we know, quality depends on several factors and characteristics such as functionality, reliability, usability, understandability... (external attributes) (ISO, 1999). Several of these characteristics are influenced by the complexity (internal attribute) of the multidimensional data model. We tried to obtain a set of metrics for measuring the complexity of datawarehouse models that help designers to improve the quality of their datawarehouses.

However, it is not enough with proposing metrics, and it is fundamental to be sure that these metrics are really useful for the goal they were conceived through different kinds of validations.

In this article, we will propose metrics for multidimensional data models quality, which can characterize their complexity and the different validations we have made with them.

In the next section, we will present the framework followed to define and validate metrics. The third section summarizes the proposed metrics, and in the fourth section, the formal validation of these metrics is described. Part of the empirical validations are presented in the fifth section and conclusions and future work will be presented in the final section.