Data Warehouse Development

José María Cavero Barca
Universidad Rey Juan Carlos, Spain

Esperanza Marcos Martínez
Universidad Rey Juan Carlos, Spain

Mario G. Piattini
Universidad de Castilla-La Mancha, Spain

Adolfo Sánchez de Miguel
Cronos Ibérica, S.A., Spain

INTRODUCTION

The concept of data warehouse first appeared in Inmon (1993) to describe a "subject oriented, integrated, nonvolatile, and time variant collection of data in support of management's decisions" (31). It is a concept related to the OLAP (online analytical processing) technology, first introduced by Codd et al. (1993) to characterize the requirements of aggregation, consolidation, view production, formulae application, and data synthesis in many dimensions. A data warehouse is a repository of information that mainly comes from online transactional processing (OLTP) systems that provide data for analytical processing and decision support.

The development of a data warehouse needs the integration of data that come from different sources, mainly legacy systems. The development of a data warehouse is, like any other task that implies some kind of integration of preexisting resources, complex. This process, according to Srivastava and Chen (1999), is "labor-intensive, error-prone, and generally frustrating, leading a number of warehousing projects to be abandoned midway through development" (118). OLTP and OLAP environments are profoundly different. Therefore, the techniques used for operational database design are inappropriate for data warehouse design (Kimball & Ross, 2002; Kimball et al., 1998).

Despite the obvious importance of having a methodological support for the development of OLAP systems, the scientific community and product providers have paid very little attention to the design process. Models usually utilized for operational database design (like the Entity/Relationship-E/R model-should not be used without further ado for analytical environments design. Bearing in mind just technical reasons, databases obtained from E/R models are inappropriate for decision support systems, in which query performance and data loading (including incremental loading) are important (Kimball & Ross, 2002). Multidimensional paradigm should be used not only in database queries but also during its design and maintenance. As stated in Dinter et al. (1999): "To use the multidimensional paradigm during all development phases it is necessary to define dedicated conceptual, logical and physical data models for the paradigm and to develop a sound methodology which gives guidelines how to create and transform these models during the development process." Wu & Buchmann (1997) claimed for data warehouse design methodologies and tools "with the appropriate support for aggregation hierarchies" and "mappings between the multidimensional and the relational models,"(79).

The next section summarizes existing approaches in data warehouse design. Then, our approach for the development of data warehouses is briefly described. Finally, conclusions are presented.

SUMMARY OF EXISTING APPROACHES

There are several proposals for data warehouse design; in this section, we summarize the most relevant ones.

In Kimball and Ross (2002) and Kimball et al. (1998), an approach based on two points is proposed: the data warehouse bus architecture that shows how to construct a series of data marts that, finally, will allow for the creation of a corporate data warehouse, and the business dimensional life cycle (BDL) with the purpose of development of data marts based on dimensional star schemas starting from the business requirements. It is an iterative methodology in which, after a project planning and a business requirements definition task, different activities are developed. These activities can be categorized into three groups: technology activities, data design activi-

ties, and specification and development of final user applications activities.

Last, there are two activities related to data warehouse deployment, maintenance, and growth. It is a detailed methodology and, according to the authors, is widely tested. However, in our opinion, it is focused on the relational model from its initial phases.

In Debevoise (1999), an object-oriented methodological approach is proposed, using Unified Modeling Language (UML) to detail the methodology steps. Use case diagrams are used to describe the tasks that the team has to carry out to complete each phase. Use cases will specify what every team member has to do to complete each project cycle part. This methodology is less detailed than the previous one and is a bit difficult to follow.

Cabibbo and Torlone (1998) presented a logical model for multidimensional (MD) database design, and a design methodology to obtain a MD schema from operational databases. As the starting point, they use an ER schema that describes an integrated view of the operational databases. This schema may contain all information valuable for the data warehouse, but the information is in an inappropriate format for this kind of system. The methodology consists of a series of steps for the MD model schema construction and its transformation into relational models and multidimensional matrices. The methodology is incomplete and starts from an ideal assumption; that is, all information is contained in the ER schema. In our opinion, operational schemas should be simply a support, giving more importance to analytical users' requirements.

Golfarelli and colleagues (Golfarelli & Rizzi, 1999; Golfarelli, Maio, & Rizzi, 1998) outlined a methodological framework for data warehouse design based on a conceptual multidimensional model of the same authors, called dimensional fact model (DFM). The methodology is mainly focused on a relational implementation.

Abelló et al. (2001, 2002) reviewed multidimensional data models and proposed a new one, as an extension of UML. Luján-Mora et al. (2002) also extended UML for multidimensional modeling and proposed a methodology also based on UML for the development of data warehouses (Trujillo & Luján-Mora, 2003; Trujillo et al., 2001).

There are many other partial proposals, focused on issues such as model transformation, view materialization, index, etc. For example, Sapia et al. (1999) proposed using data mining techniques in data warehouse design phases (for example, using data mining algorithms for discovering implicit information on data, for conflict resolution in schema integration for recovering lost values and incorrect data, etc.).

The problem with all these works is that they propose to use a new different methodology for data warehouse design, so organizations must use at least two different methodologies: one for OLTP environments and one for OLAP environments. We think that it is better to integrate data warehouse design in the existing methodologies, modifying and adding new activities, so that the training and learning curve for data warehouse design is less difficult.

OUR APPROACH

Our approach is based on applying the experience and knowledge obtained in relational database system development in the last decade (Structed-Query Language or SQL, ER modeling, Computer Aided Software Enginerring or CASE tools, methodologies...) to multidimensional database (MDDB) design. We propose a MDDB development methodology analogous to the traditional ones used in the relational database systems development. Instead of defining a new methodology, we adapt METRICA, an existing traditional methodology (de Miguel et al., 1998), to the development of data warehouses.

Our methodology (MIDEA) (Cavero et al., 2003) uses as reference framework the Spanish Public Methodology METRICA version 3 proposal (MV3), which is similar to British Structured Systems Analysis Design Method (SSADM) or French Merise. The considered MV3 processes are those that have more influence on the data warehouse development, that is, information system analysis, design, and construction (ASI, DSI, and CSI). The new processes, modified from the original MV3 proposal, have been named as ASI-MD (multidimensional), DSI-MD, and CSI-MD, respectively. Of course, it does not mean that the rest of the processes should not be taken into account on a data warehouse development, but we have considered that the differences should not be significant with respect to any other information system development.

MIDEA uses IDEA, Integrating Data: Elementary-Aggregated, (Sánchez et al., 1999) as a conceptual model. IDEA is a multidimensional conceptual model used to understand and represent analytical users' requirements in a similar manner as the ER model is used to interact with microdata users. Preexisting OLTP system data schema and requirements obtained from analytical data users are the main inputs to the construction of IDEA multidimensional conceptual schema.

This methodology is supported by a CASE tool that incorporates a graphical interface (de Miguel et al., 2000). This tool allows the transformation of a conceptual IDEA schema into a logical schema based on a model supported by some multidimensional or relational products.

Figure 1 presents an overview of the methodology, showing the scope of its three processes: ASI-MD, DSI-MD, and CSI-MD.