Chapter 10

A Multidimensional Data Warehouse Development Methodology

José María Cavero
Universidad Rey Juan Carlos, Spain

Carmen Costilla
Universidad Politécnica de Madrid, Spain

Esperanza Marcos
Universidad Rey Juan Carlos, Spain
Mario G. Piattini
Universidad de Castilla-La Mancha, Spain

Adolfo Sánchez
Cronos Ibérica, Spain

ABSTRACT

Data warehousing and online analytical processing (OLAP) technologies have become growing interest areas in recent years. Specific issues such as conceptual modeling, schemes translation from operational systems, physical design, etc. have been widely treated. A few methodologies covering the entire development cycle have also been proposed, but there is still not a general, accepted, complete methodology for data warehouse design. In this work we present a multidimensional data warehouse development methodology integrated within a traditional software development methodology.
INTRODUCTION

The term data warehouse first appeared in (Inmon, 1993), defined as a “subject oriented, integrated, nonvolatile, and time variant collection of data in support of management’s decisions.” Data warehouses are closely related to OLAP technology, first introduced by Dr. E.F. Codd in 1993 (Codd, Codd, & Salley, 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 mainly coming from online transactional processing (OLTP) systems that provides data for analytical processing and decision support.

A multidimensional view of data is not anything new (in fact, it is a very old concept): managers observe the evolution of interesting data organized in dimensions, such as products, clients, promotions, sell points, and, of course, time. The need of having simply and rapidly every historical information of operational systems has pushed companies to look for new ways of structuring and accessing their data to have advantage over their competitors. There is an agreement in that traditional database systems are not appropriate for multidimensional data analysis. Traditional OLTP systems are optimized for providing high performance in processing a lot of concurrent transactions. These transactions usually affect very few records. Meanwhile, multidimensional systems have to answer to complex queries (sometimes unpredictable) that need a huge number of records (Cabibbo & Torlone, 1998). In fact, as Ralph Kimball points out (Kimball, 1996), OLTP is profoundly different from dimensional data warehousing in their users, their data content and structures, their hardware and software, their administration and management, and their daily rhythms.

Because OLTP and OLAP environments are profoundly different, the techniques used for operational database design are inappropriate for data warehouse design (Kimball, 1996; Kimball, Reeves, & Thornthwaite, 1998).

The development of a data warehouse needs the integration of data mainly from legacy systems. The process of developing a data warehouse is, like any other task that implies some kind of preexisting resources integration, profoundly complex. This process is “labor-intensive, error-prone, and generally frustrating, leading a number of warehousing projects to be abandoned midway through development” (Srivastava & Chen, 1999).

In this respect, in recent years, there have been many proposals for some particular aspects involved in the data warehouse design process. However, although many solutions have been developed “for interesting subproblems like handling multidimensional data as typical requirement for data warehouses, view maintenance for aggregated data, data integration etc., combining these partial and often very abstract and formal solutions to an overall design
Related Content

Incremental Algorithm for Discovering Frequent Subsequences in Multiple Data Streams
[www.igi-global.com/article/incremental-algorithm-discovering-frequent-subsequences/58635?camid=4v1a](www.igi-global.com/article/incremental-algorithm-discovering-frequent-subsequences/58635?camid=4v1a)

A Method of Sanitizing Privacy-Sensitive Sequence Pattern Networks Mined From Trajectories Released
U.S. Federal Data Mining Programs in the Context of the War on Terror: The Congress, Courts, and Concerns for Privacy Protections
www.igi-global.com/chapter/federal-data-mining-programs-context/29149?camid=4v1a

Technique for Transformation of Data From RDB to XML Then to RDF
www.igi-global.com/chapter/technique-for-transformation-of-data-from-rdb-to-xml-then-to-rdf/178367?camid=4v1a