IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB7511

Chapter 9

A Methodology for Datawarehouse Design: Conceptual Modeling

José María Cavero and Esperanza Marcos
Universidad Rey Juan Carlos, Spain

Mario Piattini
Universidad de Castilla-La Mancha, Spain

Adolfo Sánchez
Cronos Ibérica, S.A., Spain

Data warehousing and online analytical processing (OLAP) technologies have become growing interest areas in latest years. Specific issues, such as conceptual modeling, schemes translation from operational systems, physical design, etc... have been widely treated, but there is not a general accepted complete methodology for datawarehouse design. In this work, we present a multidimensional datawarehouse development methodology based on and integrated with a Public software development methodology.

INTRODUCTION

The concept of datawarehouse first appeared in Inmon (1993) to describe a "subject oriented, integrated, non-volatile, and time variant collection of data in support of management's decisions". It is a concept very related to the OLAP technology, first introduced by Dr. E.F. Codd in 1993 (Codd, Codd, and Salley, 1993) to characterize the requirements of aggregation, consolidation, view production, formulae application and data synthesis in many dimensions. A datawarehouse is a repository of information mainly coming from on-line transactional processing (OLTP) systems that provides data for analytical processing and decision support.

Previously Published in *Managing Information Technology in a Global Economy*, edited by Mehdi Khosrow-Pour, Copyright © 2001, Idea Group Publishing.
This chapter appears in the book, Data Warehousing and Web Engineering by Shirley Becker Copyright © 2002, Idea Group Publishing.

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

Although many solutions have been developed "for interesting subproblems like handling multidimensional data as typical requirement for datawarehouses, view maintenance for aggregated data, data integration etc., combining these partial and often very abstract and formal solutions to an overall design methodology and warehousing strategy is still left over to the practitioners" (Gatziu, Jeusfeld, Staudt and Vassiliou, 1999).

Despite the obvious importance of having a methodological support for the development of OLAP systems, the design process has received very little attention of the scientific community and the product providers. Models usually utilized for operational data base design (like E/R model) shouldn't be used without further ado for analytical environments design. Attending only to 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, 1996). Multidimensional paradigm should be used not only in data base queries, but also during its design and maintenance. "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" (Dinter, Sapia, Blaschka and Hofling, 1999). In Wu and Buchmann (1997), authors claim for datawarehouse design methodologies and tools "with the appropriate support for aggregation hierarchies, mapping between the multidimensional and the relational models, and cost models for partitioning and aggregation that can be used from the early design stages."

The rest of the chapter is organized as follows: in the next section we briefly present the state of the art related to data warehousing methodologies. The following section outlines the IDEA conceptual multidimensional model, an overall view of the methodology, and will go more deeply into the conceptual modeling activity. Finally, in the last section, we present some conclusions and future works.