Chapter 2
Dynamic Workload for Schema Evolution in Data Warehouses
A Performance Issue

Fadila Bentayeb
University of Lyon – ERIC Laboratory, France

Cécile Favre
University of Lyon – ERIC Laboratory, France

Omar Boussaid
University of Lyon – ERIC Laboratory, France

ABSTRACT
A data warehouse allows the integration of heterogeneous data sources for identified analysis purposes. The data warehouse schema is designed according to the available data sources and the users’ analysis requirements. In order to provide an answer to new individual analysis needs, the authors previously proposed, in recent work, a solution for on-line analysis personalization. They based their solution on a user-driven approach for data warehouse schema evolution which consists in creating new hierarchy levels in OLAP (on-line analytical processing) dimensions. One of the main objectives of OLAP, as the meaning of the acronym refers, is the performance during the analysis process. Since data warehouses contain a large volume of data, answering decision queries efficiently requires particular access methods. The main issue is to use redundant optimization structures such as views and indices. This implies to select an appropriate set of materialized views and indices, which minimizes total query response time, given a limited storage space. A judicious choice in this selection must be cost-driven and based on a workload which represents a set of users’ queries on the data warehouse. In this chapter, the authors address the issues related to the workload’s evolution and maintenance in data warehouse systems in response to new requirements modeling resulting from users’ personalized analysis needs. The main issue is to avoid the workload generation from scratch. Hence, they propose a workload management system which helps the administrator to maintain and adapt dynamically the workload according to changes arising on the data warehouse schema. To achieve this maintenance, the authors propose two types of workload updates: (1) maintaining existing queries consistent with respect to the new data warehouse

DOI: 10.4018/978-1-60566-748-5.ch002
schema and (2) creating new queries based on the new dimension hierarchy levels. Their system helps the administrator in adopting a pro-active behaviour in the management of the data warehouse performance. In order to validate their workload management system, the authors address the implementation issues of their proposed prototype. This latter has been developed within client/server architecture with a Web client interfaced with the Oracle 10g DataBase Management System.

INTRODUCTION

Research in data warehousing and OLAP (On-Line Analytical Processing) has produced important technologies for the design, management and use of information systems for decision support. Nevertheless, despite the maturity of these technologies, there are new data and analysis needs in companies. These needs not only demand more storage capacity, but also new methods, models, techniques or architectures. Some of the hot topics in data warehouses include schema evolution, versioning and OLAP operators for dimension updates.

To be an effective support of OLAP analysis, data warehouses store subject-oriented, integrated, time-variant and non-volatile collection of data coming from heterogeneous data sources in response to users’ analysis needs. In a data warehouse, data are organized in a multidimensional way. The objective is then to analyse facts through measures, according to dimensions which can be divided into hierarchies, representing different granularity levels of information. The granularity levels of each dimension are fixed during the design step of the data warehouse system. After deployment, these dimensions remain static because schema evolution is poorly supported in current OLAP models.

To design a data warehouse, several approaches exist in the literature. Inmon, W.H. (2002) argues that the data warehouse environment is data-driven, in comparison to classical systems which are requirements-driven. Anahory, S., & Murray, D. (1997) propose a catalogue for conducting users interviews in order to collect end-user requirements while Kimball, R. (1996) and Kimball, R. et al. (1998) state that the main step of the design process in data warehousing is based on a business process to model. User requirements describe the tasks that the users must be able to accomplish with the help of the data warehouse system. They are often collected in the design step of the data warehouse. Thus some new requirements are not satisfied and some trends are not explored. Indeed, data sources and requirements are often changing so that it is very important that the data warehouse schema evolves according to these changes. In business environment, several changes in the content and structure of the underlying data sources may occur and individual analysts’ needs can emerge and grow in time. Thus, a data warehouse schema cannot be designed in one step since it evolves over the time.

To consider this problem, two categories of research emerged: (1) temporal multidimensional data models that manage and keep the evolutions history by time-stamping relations over hierarchy levels proposed by Morzy, T., & Wrembel, R. (2004), Morzy, T., & Wrembel, R. (2003), Mendelzon, A.O., & Vaisman, A.A. (2000) and Bläujute, R. et al. (1998); and (2) extending the multidimensional algebra with a set of schema evolution operators proposed by Pourabbas, E., & Rafanelli, M. (2000), Hurtado, C.A. et al. (1999) and Blaschka, M. et al. (1999).

In the context of data warehouse evolution, we proposed in a previous work a new data warehouse architecture which takes into account personalized user’s analyses independently of the data sources. More precisely, in Bentayeb, F. et al. (2008) and Favre, C. et al. (2007a) we designed a user-driven data warehouse model in which schema evolves according to new user’s analysis needs. In this case, new dimension hierarchies are then created based on personalized user’s analyses in an