Chapter XV
A Survey of Data Warehouse Model Evolution

Cécile Favre
University of Lyon (ERIC Lyon 2), France

Fadila Bentayeb
University of Lyon (ERIC Lyon 2), France

Omar Boussaid
University of Lyon (ERIC Lyon 2), France

INTRODUCTION

A data warehouse allows the integration of heterogeneous data sources for analysis purposes. One of the key points for the success of the data warehousing process is the design of the model according to the available data sources and the analysis needs (Nabli, Soussi, Feki, Ben-Abdallah & Gargouri, 2005).

However, as the business environment evolves, several changes in the content and structure of the underlying data sources may occur. In addition to these changes, analysis needs may also evolve, requiring an adaptation to the existing data warehouse's model.

In this chapter, we provide an overall view of the state of the art in data warehouse model evolution. We present a set of comparison criteria and compare the various works. Moreover, we discuss the future trends in data warehouse model evolution.

BACKGROUND

Schema and Data Evolution in Data Warehouses: The Coherence Problem

The main objective of a data warehouse is to provide an analysis support for decision-making.

The analysis possibilities of a data warehouse mainly depend on its schema. The analysis results depend on the data. Following the evolution of sources and analysis needs, the data warehouse can undergo evolution on the level of its schema and its data at the same time.

From the schema evolution point of view, the following evolutions can be envisaged:

- dimension (adding/deletion)
- measure (adding/deletion)
- hierarchy structural updating (level adding/deletion)

These evolutions enrich or deteriorate the analysis possibilities of data warehouses. However, they do not induce erroneous analysis as evolution of the data does.

In regard to data evolution, we have identified three operations: insertion, deletion, and updating of data in the data warehouse. These operations can be performed on either the fact table or the dimension tables, and depending on the case, do not have the same impact on analysis coherence. The insertion (in the fact table or in the dimension tables) corresponds to the usual data-loading process of the data warehouse.

However, since data warehouses contain historical and nonvolatile data, records should not be updated or deleted. However, as Rizzi and Golfarelli (2006) point out, updates in the fact table could be required in order to correct errors or to reflect the evolution of the events.

Furthermore, the usual assumption in data warehouse modeling is the independency of the dimensions. Thus, defining a dimension to characterize time induces that other dimensions are independent of the time dimension. In other words, these dimensions are supposed to be time-invariant. However, this case is extremely rare. Thus, in order to ensure correct analysis, these dimensions have to evolve in a consistent way (Letz, Henn & Vossen, 2002).

Kimball (1996) introduced three types of "slowly changing dimensions" that consist in three possible ways of handling changes in dimensions. The basic hypothesis is that an identifier cannot change, but the descriptors can. The first way consists of updating the value of the attribute. In this case, the historization of changes is not available. Thus, this solution has consequences on analysis coherence only if this updated attribute is used to carry out the analysis. The second type allows keeping all the versions of the attribute's value by creating another record valid for a time period. The drawback of this approach is the loss of comparisons throughout versions. This is due to the fact that the links between evolutions are not kept even if evolutions are preserved. The last type consists of creating another descriptor to keep track of the old value in the same record. Thus, we keep the link between the two versions. However, if there are several evolutions, there is a problem to consider the different versions with changes on several attributes that do not occur at the same time.

As Body, Miquel, Bédard, and Tchounikine (2002) summed it up; the study of Kimball takes into account most users' needs and points out the necessity of keeping track of both history and links between transitions. Indeed, the main objective of a data warehouse is to support correct analysis in the course of time and ensure good decisions.

This objective mainly depends on the capacity of the data warehouse to be a mirror of reality. From our point of view, the model evolution problem must not be separated from the problem of analysis coherence. Thus, we think we have to identify when the evolution induces incoherence of analysis. Data historization and, more precisely, dimension historization are required for descriptors that are involved in the analysis process. Note that for analysis purposes, it is necessary to be able to translate facts by getting data in a consistent time.

In order to take into account these evolutions, we can distinguish in the literature two types of approaches: model updating and temporal mod-