Chapter 3

X-WACoDa
An XML-Based Approach for Warehousing and Analyzing Complex Data

Hadj Mahboubi
Université de Lyon (ERIC Lyon 2), France

Jean-Christian Ralaivao
Université de Lyon (ERIC Lyon 2), France

Sabine Loudcher
Université de Lyon (ERIC Lyon 2), France

Omar Boussaïd
Université de Lyon (ERIC Lyon 2), France

Fadila Bentayeb
Université de Lyon (ERIC Lyon 2), France

Jérôme Darmont
Université de Lyon (ERIC Lyon 2), France

ABSTRACT

Data warehousing and OLAP applications must nowadays handle complex data that are not only numerical or symbolic. The XML language is well-suited to logically and physically represent complex data. However, its usage induces new theoretical and practical challenges at the modeling, storage and analysis levels; and a new trend toward XML warehousing has been emerging for a couple of years. Unfortunately, no standard XML data warehouse architecture emerges. In this chapter, the authors propose a unified XML warehouse reference model that synthesizes and enhances related work, and fits into a global XML warehousing and analysis approach we have developed. They also present a software platform that is based on this model, as well as a case study that illustrates its usage.

DOI: 10.4018/978-1-60566-756-0.ch003
INTRODUCTION

Data warehouses form the basis of decision-support systems (DSSs). They help integrate production data and support On-Line Analytical Processing (OLAP) or data mining. These technologies are nowadays mature. However, in most cases, the studied activity is materialized by numeric and symbolic data, whereas data exploited in decision processes are more and more diverse and heterogeneous. The development of the Web and the proliferation of multimedia documents have indeed greatly contributed to the emergence of data that can:

- Be represented in various formats (databases, texts, images, sounds, videos...);
- Be diversely structured (relational databases, XML documents...);
- Originate from several different sources;
- Be described through several channels or points of view (a video and a text that describe the same meteorological phenomenon, data expressed in different scales or languages...);
- Change in terms of definition or value over time (temporal databases, periodical surveys...).

We term data that fall in several of the above categories complex data (Darmont et al., 2005). For example, analyzing medical data regarding high-level athletes has lead us to jointly exploit information under various forms: patient records (classical database), medical history (text), radiographies and echographies (multimedia documents), physician diagnoses (texts or audio recordings), etc. (Darmont & Olivier, 2006; Darmont & Olivier, 2008)

Managing such data involves lots of different issues regarding their structure, storage and processing (Darmont & Boussaïd, 2006); and classical data warehouse architectures must be reconsidered to handle them. The XML language (Bray et al., 2006) bears many interesting features for representing complex data (Boussaïd et al., 2007; Boussaïd et al., 2008; Darmont et al., 2003; Darmont et al., 2005). First, it allows embedding data and their schema, either implicitly, or explicitly through schema definition. This type of metadata representation suits data warehouses very well. Furthermore, we can benefit from the semi-structured data model’s flexibility, extensibility and richness. XML document storage may be achieved either in relational, XML-compatible Database Management Systems (DBMSs) or in XML-native DBMSs. Finally, XML query languages such as XQuery (Boag et al., 2007) help formulate analytical queries that would be difficult to express in a relational system (Beyer et al. 2004; Beyer et al., 2005). In consequence, there has been a clear trend toward XML warehousing for a couple of years (Baril & Bellahsène, 2003; Hümmer et al., 2003; Nassis et al., 2005; Park et al., 2005; Pkorný, 2002; Vrdoljak et al., 2003; Zhang et al., 2005).

Our own motivation is to handle complex data into a complete decision-support process, which requires their integration and representation under a form processable by on-line analysis and/or data mining techniques (Darmont et al., 2003). We have already proposed a full, generic data warehousing and on-line analysis process that includes two broad axes (Boussaïd et al., 2008):

- Data warehousing, including complex data integration and modeling;
- Complex data analysis.

More precisely, the approach we propose consists in representing complex data as XML documents. Then, we recommend an additional layer to prepare them for analysis. Complex data under the form of XML documents are thus multidimensionally modeled to obtain an XML data warehouse. Finally, complex data analysis can take place from this warehouse, with on-line analysis, data mining or a combination of the two approaches.