IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1117, USA ITB7520
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

Chapter 18

The Development of Ordered SQL Packages to Support Data Warehousing

Wilfred Ng
Hong Kong University of Science and Technology, Hong Kong

Mark Levene
University of London, UK

Data warehousing is a corporate strategy that needs to integrate information from several sources of separately developed Database Management Systems (DBMSs). A future DBMS of a data warehouse should provide adequate facilities to manage a wide range of information arising from such integration. We propose that the capabilities of database languages should be enhanced to manipulate user-defined data orderings, since business queries in an enterprise usually involve order. We extend the relational model to incorporate partial orderings into data domains and describe the ordered relational model. We have already defined and implemented a minimal extension of SQL, called OSQL, which allows querying over ordered relational databases. One of the important facilities provided by OSQL is that it allows users to capture the underlying semantics of the ordering of the data for a given application. Herein we demonstrate that OSQL aided with a package discipline can be an effective means to manage the inter-related operations and the underlying data domains of a wide range of advanced applications that are vital in data warehousing, such as temporal, incomplete and fuzzy information. We present the details of the generic operations arising

Previously Published in the *Journal of Database Management, vol.12, no.4,* Copyright © 2001, Idea Group Publishing.

This chapter appears in the book, Data Warehousing and Web Engineering by Shirley Becker.

from these applications in the form of three OSQL packages called: OSQL_TIME, OSQL_INCOMP and OSQL_FUZZY.

Data warehousing is a corporate strategy that addresses a broad range of decision support requirements such as querying information over its underlying databases and managing ordered data for the purpose of analysis. One of the main characteristics of data warehousing is that in order to build its foundation, it should consist of integrated data from several sources of separately developed information systems. The transmission of data relies on the network system which connects all these information systems. As a result, the integrated database has the following important features:

- **It involves huge amounts of historical data.**

Data warehouse is described as a "subject-oriented, integrated, non-volatile, time variant" collection of data which is intended to support management decisions (Inmon, 1996). It is widely recognised that the underlying database in a data warehouse should capture transactions and snapshots in time in an efficient manner in order to carry out the activities of market forecast and strategic planning (McCabe & Grossman, 1996).

- **It is usually incomplete.**

This is due to two main reasons. First, some sources of the databases may be incomplete in order to protect sensitive data or to improve the speed of the process of data downloading via a network. Second, it has been observed in Libkin (1995) that even if each source of the database is complete, the integrated database may still not be complete. Hence, incompleteness may show up in the integrated database or in the answer to users' queries.

- **It is mainly used for decision support in an enterprise.**

However, many management professionals may not necessarily have good knowledge about the technical aspects of a data warehouse. As a result, their queries over the database are sometimes fuzzy in nature due to the ambiguity of natural languages. For example, they may ask to find the "best performed" shares in the Hong Kong stock market this month in order to carry out some share trading activities.

Many database researchers have recently recognised that ordering is inherent to the underlying structure of data in many database applications (Maier & Vance, 1993; Libkin, 1995; Buneman et al., 1997) including temporal information (Tansel et al., 1993), incomplete information (Codd, 1986) and fuzzy information (Buckles & Petry, 1982). However, current relational Database Management Systems