International Journal of Data Warehousing and Mining
Volume 15 • Issue 3 • July-September 2019

Maintaining Dimension's History in Data Warehouses Effectively

Canan Eren Atay, Dokuz Eylül University, Izmir, Turkey
Georgia Garani, University of Thessaly, Volos, Greece
https://orcid.org/0000-0003-1892-4183

ABSTRACT

A data warehouse is considered a key aspect of success for any decision support system. Research on temporal databases have produced important results in this field, and data warehouses, which store historical data, can clearly benefit from such studies. A slowly changing dimension is a dimension in which any of its attributes in a data warehouse can change infrequently over time. Although different solutions have been proposed, each has its own particular disadvantages. The authors propose the Object-Relational Temporal Data Warehouse (O-RTDW) model for the slowly changing dimensions in this research work. Using this approach, it is possible to keep track of the whole history of an object in a data warehouse efficiently. The proposed model has been implemented on a real data set and tested successfully. Several limitations implied in other solutions, such as redundancy, surrogate keys, incomplete historical data, and creation of additional tables are not present in our solution.

KEYWORDS

Data Warehouse, Object-Relational, Slowly Changing Dimension, Temporal Data Warehouse

INTRODUCTION

The remarkable increase in the amount of data collected by organizations these days for the purpose of extracting business information requires special consideration by the database research community. Companies collect data to generate decision-making strategies. However, multiple decision support environments, which operate independently, require data to be heavily redundant. Because of this issue, two IBM researchers, Barry Devlin and Paul Murphy, introduced the concept of a "business data warehouse" in the late 1980s. Since then, data warehousing has been an active research field. The main purpose of data warehousing is to access historical data that cannot be altered. According to Bill Inmon (2002), a data warehouse (DW) is "a collection of subject-oriented, integrated, non-volatile and time-variant data to support management's decisions." The non-volatile and time-variant data features of data warehousing suggest that it should allow changes to the data values without overwriting the existing values.

The types of data maintained by a DW are used in management reports, various business queries, decision support systems, management information systems, and data mining applications by linking the various data that are distributed throughout an organization. The increasing demands of financial analyses in healthcare (Silver et al., 2001), clinical data mining (Lyman et al., 2008), laboratory test data analyses (Allard, 2003), disease control (Wisniewski et al., 2003), adverse drug event control (Einbinder & Scully, 2002), the clinical decision process (Banek et al., 2006), and information feedback

DOI: 10.4018/IJDWM.2019070103

for hospital practice management (Grant et al., 2006) in the field of healthcare have given rise to the research and development of clinical DWs. Undoubtedly, the user of a DW needs to be assured that the data being stored are timely, accurate, and complete.

The problem of time support in a database management system has also been an active field of research for many years. Temporal databases can capture either the history of objects and their attributes (known as valid time), or the history of the database activities (known as transaction time). A valid time is necessary to model the time-varying states of an object and its attributes. These states may be in the past, present, or even future. A transaction time denotes the timestamp of any change as it is recorded in a database. This time is system generated and may not extend beyond the current time.

DWs store historical data and can therefore clearly benefit from research on temporal databases. To manage time-varying multidimensional data, temporal data warehouses (TDWs) have been proposed by combining the research fields of temporal databases and DWs. DWs commonly include a time dimension indicating the timeframe for measures. However, the time dimension cannot be used to keep track of changes in other dimensions. In many cases, changes in the dimensional data and the time at which they occur are important requirements for analysis. Such dimensions, in which attribute values may change are called slowly changing dimensions (SCDs). Kimball (2002a) proposed several implementation solutions for this problem; Type 1 overwrites a value, Type 2 adds a new row in the dimension with the updated attribute values and Type 3 adds a new attribute in the dimension to preserve the old attribute value. He also proposed other techniques where he combined hybrid approaches. Nonetheless, these solutions are not adequate because they do not preserve the entire data history or are not easy to implement. Furthermore, such solutions do not consider research related to managing the time-varying information in a temporal database.

SCDs in the world of data management can pose significant challenges to data quality if a suitable strategy is not adopted to mitigate the impact they may have on businesses. In this paper, a new Object-Relational Temporal Data Warehouse (O-RTDW) model is proposed that provides a temporal extension for each time-related attribute for TDWs. In the O-RTDW model, the valid time is attached to the attributes in the dimension tables. Most things in real life change over time and a database system should capture such changes, which resembles one of the main characteristics of a DW, i.e., the data stored in a data warehouse are historical and time-dependent (Inmon, 2002). The O-RTDW is particularly aimed at representing such types of data. Therefore, the present research aims to apply an object-relational solution, provided with logical model and formal definition, to encounter the problem of SCDs in data warehousing. A case study with performance results using an object-relational structure for SCDs which real lung cancer data made available by the American National Cancer Institute (NCI, 2017) is also presented.

The rest of the paper is organized as follows. Section 2 discusses related research work. The research objective is presented in section 3. SCDs are introduced in section 4. In section 5, the O-RTDW model is presented and explained. Implementation is presented in section 6 where different types of queries for the model are demonstrated together with experimental results. Theoretical and practical implications of the proposed research are included in Section 7. Section 8 concludes the paper and presents future work.

RELATED WORK

The necessity of managing time-varying data in a database has been acknowledged over the past several decades and has been revealed through numerous earlier researchers on temporal databases. There are two common approaches for extending a relational data model: tuple timestamping and attribute timestamping. The distinction between the two is in where the timestamps are attached. Whereas tuple timestamping uses 1NF relations, attribute timestamping requires N1NF relations (Atay & Tansel, 2009). Another taxonomy for classifying temporal databases is in terms of the supported time dimension, i.e., valid time concerns the time when an event is true in the real world, transaction