An Object-Relational Data Warehouse Modeling for Complex Data

Wen-Yang Lin[1], Chin-Ang Wu[2,*], Chuan-Chun Wu[3]

[1]Department of Computer Science and Information Engineering, National University of Kaohsiung
700, Kaohsiung University Rd., Nan-Tzu Dist., Kaohsiung, Taiwan 811, R.O.C
[2,3]Department of Information Engineering, I-Shou University
1, Section 1, Hsueh-Cheng Rd., Ta-Hsu Hsiang, Kaohsiung County, Taiwan 840, R.O.C.

Abstract

A data warehouse is an essential component to the decision support system. The traditional data warehouse provides only numeric and character data analysis. But as information technologies progress, complex data such as semi-structured and unstructured data become vastly used. Developing new data warehouse technologies to accommodate the demand of integrating complex data is important. In this paper, we propose a data model for the complex data warehouse. This model incorporates the well-known star schema with object-relational concepts, inheriting the easy-understanding and multidimensional characteristic, and providing mechanisms to integrate complex data in a data warehouse model.

Keywords: Data warehouse, object relational, data model, star schema, complex data

1. Introduction

Data warehouse is characterized as "a subject-oriented, integrated, nonvolatile, time-variant collection of data in support of management's decisions" [7]. Data warehouse provide complex data analysis, knowledge discovery and decision making support. Most of the traditional data warehouse encompasses only structured numeric and character data types. But as information technologies progress, semi-structured and unstructured data become vastly used. Many useful reports could hence be produced and the data mining in the type of association rules, sequential patterns, classifications and predictions could possibly provide new knowledge to the users. The need of a more powerful data warehousing system that can integrate, represent and store both structured and complex data is obvious.

In this paper we attempt to propose a data model for the complex data warehouse. This model incorporates the well-known star schema with object-relational concepts, inheriting the easy-understanding and multidimensional characteristic, and possessing the mechanism to integrate complex data in a data warehouse model.

The rest of this paper is organized as the following. Section 2 is the review of the research background and section 3 is the study of related work. In section 4 the proposed data warehouse model is described and finally, the conclusions are stated in Section 5.

2. Background

2.1. Relational Database Management System and Star Schema

A relational DBMS stores data in a database system, which consists of one or many tables of rows and columns. These tables can have relationships with other tables. The data types that can be stored are confined to atomic such as integer, real, character, string, date time and currency. Any field of a row stores only a single value. It is advantageous for users to generate varieties of reports.

For a complete enterprise data warehouse system, the source databases are in the backend; data warehouse and data marts are in the core; and analysis tools are in the front end. The famous multidimensional modeling of star schema presented by Kimball [7] is the data model referenced and used by most of the OLAP users. A star schema includes a big fact table in the middle and surrounded by many usually smaller dimension tables. A fact table contains the foreign keys, which related to the primary key of

* She is also a lecturer in Cheng Shiu University, Nioa Song, Kaohsiung County, Taiwan.

each of the dimension tables. A fact table also contains the numerical measures, which are the targets for analysis [7]. A dimension table contains attributes which describes the fact and answers 'who', 'what', 'when', 'where', 'how', and 'why' about the fact [9]. It provides aggregations from different aspects and on multiple levels.

The advantages of a star schema are that it is easy for users to understand and it is suitable for query processes in relational databases [11]. Star schemas fit very well to the relational DBMS almost in every respect but there are some limitations also. First of all, it does not have the explicit hierarchy classifications [1][4]. Secondly, it does not support the handling of complex heterogeneous and unstructured data. It is deficient for modeling a data warehouse of complex data types.

2.2. Object Relational Database Systems

Complex data types including semi-structured or non-structured such as video, audio, graphic and text are best served by object-relational technologies [14]. In addition to its traditional role in the safe and efficient management of relational data, an object-relational database provides support for complex objects. In addition to multimedia data, complex objects also provide multiple base atomic types and user defined object types such as composite attributes, collection sets or variable arrays. In general, complex objects have system-assigned OIDs, while base atomic types do not. Objects have operations and user defined functions. Some object-relational DBMS support the inheritance of properties from super-type to sub-type [14].

The query language of object-relational database is SQL extensions from relational database. Specifically, object-relational database provides query language for manipulating composite data types as well as collection set types. Methods presenting behavior functions, system provides or user defined, can be shown both in selection list and predicates [8].

3. Related Work

The multidimensional data warehouse technologies are described in [6]. Complex multidimensional data issues are discussed such as non-strict, irregular dimension hierarchies and non-additive problems.

Structural level of facts and dimensions are presented using an object oriented conceptual model in [2]. The many-to-many relationships between the fact and some particular dimensions are denoted by * on the fact and 1..* on the dimension side in the graph. An object as degenerated dimension is defined in the fact table to bridge the multiple items. The following dimension hierarchies are discussed: multiple and alternative classification hierarchies, strict vs. non-strict, complete classification and categorization of dimensions.

Object oriented modeling of data warehouse supports complex data and is rich in semantics [5]. Many researches have involved in incorporating object-oriented schema with multidimensional schema. In [10], UML is used to exhibit an object-oriented multidimensional model. The concepts of dimension, measure, data cube and the concepts of operations such as rolling up, drilling down are defined and being mapped to the UML as object classes. Six desirable types of object-oriented dimensions were discussed conceptually in [12]. They are classification vs. instantiation, generalization vs. specialization, aggregation vs. decomposition, caller vs. called, derivability and dynamicity. The use of object-oriented multidimensional model does not imply that the data are stored in an object-oriented database [3].

An object-oriented database is a set of individual objects and not data set groupings. A pure object-oriented database does not run user reports as good as a relational database does. In order to meet the demand of analyzing both atomic and complex type of data for decision support in all sorts of applications, relational-based object-oriented modelings are necessary.

4. The Object Relational Data Warehouse Model

Object-relational database organizes data in the relational tabular structures, which can integrate complex objects [8].

Traditionally, data for analysis is mostly numeric; but as information technology progress, complex data such as semi-structured data and multimedia data should be able to be analyzed. Some object relational database management systems provide complex objects, which include multiple atomic types and user defined object types such as composite attributes, collection sets, variable array and multimedia objects. The reference mechanism to the complex data is also provided. This makes it conceivable that the data warehouse should be able to incorporate the complex data.

Different data warehouse models have been proposed at 'conceptual level' that are close to the user's view and independent of the implementation; at 'logical level' that are database management technology dependent; at 'physical level' specific DBMS is chosen [1][3]. The data warehouse model we attempt to propose here is of logical level, which combines object oriented functionalities into the relational multidimensional star schema.

As shown in Fig. 1, this model can be viewed as an extension of relational star schema that incorporates the features of object-oriented data model. Its components are fact tables surrounded by dimensions. In the middle of the schema stand the Payment Fact and the Image Fact. They relate with dimensions by the corresponding keys.

Usually data for analysis is numeric like SelfPay and Insurance in the Payment Fact. These by default are endowed the standard aggregation functions such as Sum, Min, Max, Count and Avg along dimensions. In addition to this, user defined functions, e.g. UDF(IncreaseRate), are provided to handle varieties of atomic data analysis such as statistical functions or mathematical calculations. When the interested value is in percentage, for example as mentioned in [13] the HbA1c% measurement for diabetes, standard deviation and other statistical functions are necessary. User defined functions therefore provide unlimited rules of calculations for measures.

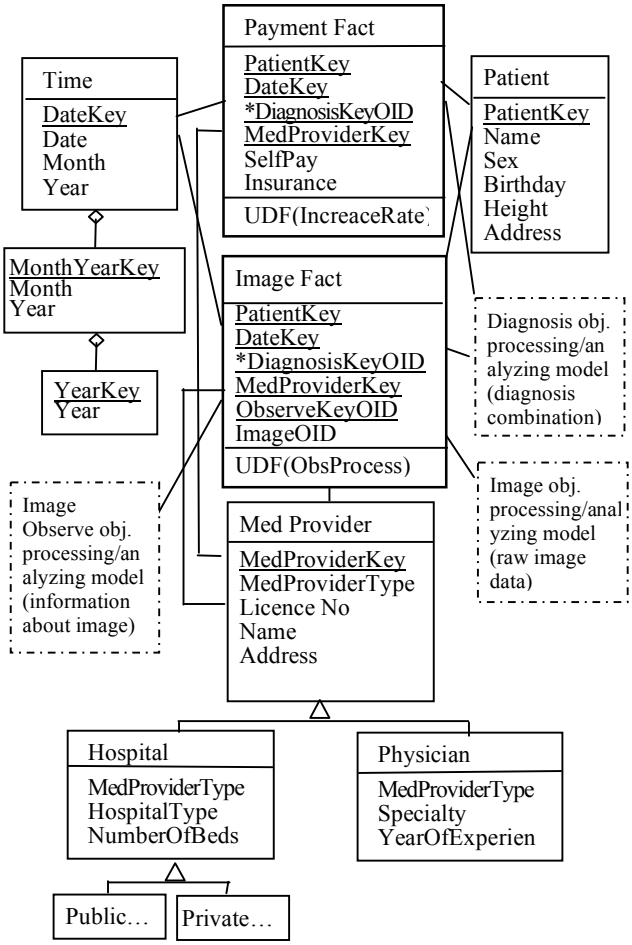

Fig. 1: Object-relational star schema for complex data.

The complex data is defined as objects and the analysis accordingly is non-additive. The operations for analysis are specific to different types of objects and thus a processing/analyzing model (the dashed boxes) for each object is displayed, which can be implemented by extra programming efforts. A typical example of object analysis to the ImageOID is to query similar images from the data warehouse.

In some data warehouse application, many-to-many relations between fact and dimensions are inevitable. For example, in a clinical application, a patient could have multiple diagnoses in one visit and the same single diagnosis could happen to many different patients. In a traditional multidimensional data warehouse, a bridge table with all possible combinations of diagnoses is often built between them to accommodate this problem [11]. But the combinations of the diagnoses can lead to an excessive amount of data, which means extra storage and sluggish retrieval speed [11]. A complex data object with an OID in the fact can be used to integrate these many items together and treated as a degenerated dimension [2] like the *DiagnosisKeyOID in both fact tables. Each *DiagnosisKeyOID object defines a different combinations of diagnoses

The ObserveKeyOID in the Image Fact is another example which includes information about the image, such as the type of modality (e.g. X-Ray), and on the site of the body (e.g. chest) where the image was taken and also the medical professional's observations (e.g. finding a mass, diagnosing a metastasis and suggesting biopsy). It is a very complex structure that involves the concerns of medical data.

The Time dimension is aggregation hierarchy or whole-part relationships. The UML whole-part notation is adopted to demonstrate the relationships between the hierarchies. Left of Fig. 1 shows an example. A complete date includes day-month-year. A month-year is part of day-month-year and a year is part of month-year.

In some other cases dimension classification hierarchies have the generalization and specialization relations. An UML generalization/specialization notation is used to denote it. In Fig. 1, both hospital and physician are medical service providers. They share some common attributes like license number, name, and address. They also possess some special attributes of their own. Hospitals have information about number of beds but physicians do not while physicians have specialty which hospitals have not. If traditional relational scheme is used, a dimension of medical service provider will have all the attributes to cover both hospital and physician, thus entries of hospital would have empty values on specialty and year of experience and vice versa. Consequently, a sparse matrix is generated. Complex object scheme of user defined type and the property of inheritance can be used to avoid sparse matrix derived from using a pure relational scheme. The following shows the

attempted implementation in object-relational like language.

```
Create Type MedProviderType(
    MedproviderType  numeric,
    LicenceNo        Char(20),
    Name             VarChar(30),
    Address          VarChar(40);

Create Type  HospitalType(
    HospitalCode     Char(10),
    NumberOfBeds     Numeric,
    Under MedProviderType);

Create Type PhysicianType(
    Specialty        VarChar(30),
    YearOfExperience Numeric,
    Under MedProviderType);
```

The model we proposed here meant to prepare the data for dealing with varieties of analytical questions like: What's the difference of amount spend on liver disease among different local areas? What's the sex difference of diabetes diseases spending? Are there pattern associations of X-Ray image from cancer patients in the same area? Get the similar stroke MRIs, et cetera. The necessary data marts can be generated from this model according to user requirements.

5. Conclusion

The data warehouse model proposed in this paper is easy for users to understand and based on the concept of object-relational database and star schema. Object-relational database is different from object-oriented database in that the former provides a data repository separated from the application programs [8].

In fact tables of the proposed model, complex data can be measures which users are interested in analyzing with. Many-to-many relationships between a fact table and some dimensions can be modeled as degenerated dimensions via complex objects, and simple aggregation of whole-part relationships are modeled as traditional relational tables while generalization-specialization are drawn with type inheritance from super-class to sub-class.

The model we proposed in this paper is an initial work. Many issues such as handling temporal change, bi-temporal problems, strict and non-strict dimension hierarchies [13][2] and many other multimedia storage and retrieval issues are to be taken care of. Further advanced researches on these related issues will be conducted in the future.

Acknowledgment

This work was partially supported by the National Science Council of ROC with grant No. NSC 91-2213-E-214-017.

6. References

[1] A. Abello, J. Samos and F. Saltor, "A data warehouse multidimensional data models classification," *Italian Association for Artificial Intelligence AI*IA Notizie*, Vol. 1, pp. 9-21, 1999.

[2] J. Trujillo, M. Palomar, J. Gomez, and I.Y. Song, "Designing data warehouses with OO conceptual models," *IEEE Computer*, Vol. 34, No. 12, pp. 66-75, 2001.

[3] C. Batini, S. Ceri, and S.B. Navathe, *Conceptual Database Design: An Entity-Relationship Approach*, Benjamin/Cummings, 1991.

[4] A. Bauer, W. Hummer, and W. Lehner, "An alternative relational OLAP modeling approach," *Proc. 2nd Int. Conf. on Data Warehouse and Knowledge Discovery, LNCS 1874*, pp. 189-198, 2000.

[5] V. Gopalkrishnan, Q. Li, and K. Karlapalem, "Star/snow-flake schema driven object-relational data warehouse design and query processing strategies," *Proc. 1st Int. Conf. on Data Warehousing and Knowledge Discovery, LNCS 1676*, pp. 11-22, 1999.

[6] T.B. Pedersen and C.S. Jensen, "Multidimensional Database Technology," *IEEE Computer*, Vol. 34, No. 12, pp. 40-46, 2001.

[7] R. Kimball, *The Data Warehouse Toolkit*, John Wiley & Sons, 1996.

[8] S. McClure, "Object database vs. object-relational databases," *IDC Bulletin* No. 14821E, 1997.

[9] D.L. Moody and M.A.R. Kortink, "From enterprise models to dimensional models: A methodology for data warehouse and data mart design," *Proc. Int. Workshop on Design and Management of Data Warehouses*, 2000.

[10] T.B. Nguyen, A.M. Tjoa, and R.R. Wagner. "An object oriented multidimensional data model for OLAP," *Proc. 1st Int. Conf. on Web-Age Information management, LNCS 1846*, pp. 69-82, 2000.

[11] P. Ponniah, *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professional*, John Wiley & Sons Inc., 2001.

[12] T.B. Pedersen and C.S. Jensen, "Clinical data warehousing–A survey," *Proc. 7th Mediterranean Conf. on Medical and Biological Engineering and Computing*, 1998.

[13] T.B. Pedersen and C.S. Jensen, "Research issues in clinical data warehousing," *Proc. 10th Int. Conf. on Scientific and Statistical Database Management*, pp. 43-52, 1998.

[14] M. Stonebraker and P. Brown, Object-Relational *DBMSs: Tracking the next Great Wave*, Morgan Kaufmann, 1999.