Guidance on a Basic Understanding of the Relational Model Theory and Data Warehouse Technologies

Tomohide Iwao*

Institute for Advancement of Clinical and Translational Science (iACT) Kyoto University Hospital. 54 Shogoin Kawahara-cho Sakyo-ku, Kyoto, 606-8507, Japan

Abstract

In recent years, many medical studies using healthcare databases have been carried out. And, in most cases, relational databases are used as database management systems. Since it is necessary to handle data by making full use of database manipulation language for analysis purposes, skills different from those for operation purposes are required. What is important here is a deep understanding of the definitions and benefits of the relational data model theory. However, some people do not understand the exact definition of relational data model theory and are confused with relational databases. This paper promotes a deeper understanding of the relational model theory by explaining in detail the relationship between the predicate logic and relational model theory using easy-to-understand examples.

Keywords

Data warehouse, Real world data, Relational data model, Big data analysis

Correspondence to:

Tomohide Iwao
Institute for Advancement of Clinical and Translational Science (IACT),
Kyoto University Hospital,
54 Shogoin Kawahara-cho Sakyo-ku,
Kyoto, 606-8507, Japan
Email: tomohide@kuhp.kyoto-u.ac.jp
Tel: +81-75-751-4879

Citation: Iwao T (2022). Guidance on a Basic Understanding of the Relational Model Theory and Data Warehouse Technologies. EJBI. 18(6):52-55.

DOI: 10.24105/ejbi.2022.18.6.52-55

Received: 10-Jun-2022, Manuscript No. ejbi-22-66339;
Editor assigned: 11-Jun-2022, Pre QC No. ejbi-22-66339(PQ);
Reviewed: 21-Jun-2022, QC No. ejbi-22-66339;
Revised: 23-Jun-2022, Manuscript No. ejbi-22-66339(R);
Published: 30-Jun-2022

1. Introduction

Most healthcare databases are operated by relational databases. For this reason, it is essential for those who challenge big data analysis today to have a correct understanding of the relational data model theory. The relational data model theory is a theory of database operations based on mathematics called predicate logic. Most introductory textbooks omit the relationship between the predicate logic and relations, so they are often confused with the relational model theory and relational databases, or with relations and tables being the same. This paper promotes a deep understanding of the relational model theory by explaining in detail the relationship between predicate logic and relational models using examples. After that, the merits of the relational model theory will be described. Finally, a brief description of data warehousing technologies based on relational model theory and its applications.

2. Relational Data Model

The Relational Data Model is a database management system theory developed in 1970 by Edgar Frank „Ted“ Codd [1]. It was devised on the basis of predicate logic, a field within mathematical logic, and set theory. The following is a simple explanation of predicate logic and set theory.

3. Predicate Logic and Set Theory

The concept of predicate logic is an extension of propositional logic. Propositional logic is the study dealing with the truth or falsehood of propositions.

> Proposition 1:
> Patient X1 is female

The “X1” part of Patient X1 in Proposition 1 can be given various different names, but in predicate logic, it is called a variable [2]. The phrase “is female” in Proposition 1 expresses a quality or relationship of this variable, and this is called a predicate. More general propositions such as the one below can be devised by increasing the number of predicates.

> Proposition 2:
> Patient X1 is a female aged 57 years’ old who was diagnosed on September 24, 2007 with stroke and died on October 15, 2007

Increasing the number of variables thus enables the handling of a large number of propositions, which can be displayed in the form of tables like Example Table 1 below. Each line in the table corresponds to a single proposition. Codd focused on the fact that if the table headings {Patient, Sex, Age, Diagnosis Date, Disease name, Death date} were considered as a set of elements, they could be handled as a data construct. As a result, each of the

records in the table below is simultaneously a proposition and a set composed of individual elements (Table 1).

4. Advantages of the Relational Data Model

As described above, Codd proposed a database management system that defined general data constructs on the basis of predicate logic and set theory. This data model has two main advantages. The first is that when populating a data model, it can be expressed in the form of a table. This enables the creation of interfaces that are highly intuitive and easily visualized. The second advantage is that it enables efficient data manipulation by means of the use of truth functions and set operations, which are reliant on predicate logic and set theory, respectively, and are suited to computer processing. Thanks to these two advantages, the Relational Data Model has been broadly accepted worldwide. Today, various vendors have released Relational Database (RDB) products supporting the Relational Data Model, and new versions are repeatedly being released [3]. Truth functions and set operations in data manipulation are implemented in the SQL database language, which is incorporated into the various RDB products. The following is a simple explanation of SQL. SQL is a database language implemented for data manipulation in the Relational Data Model. Strictly speaking, however, SQL is not a specification of the Relational Data Model, but one form of implementation that meets its specifications. It communicates with the Database Management System (DBMS) to implement mainly data definitions and manipulation. The current standard is SQL: 2016 [4], and this is extended in individually different ways in products from different vendors. Demand for analysis use-functions has risen in recent years, and Online Analytical Processing (OLAP) functions are now attracting attention [5].

5. Data Warehouse

Data Warehouse (DW) is a database structure suited to data analysis for decision-making that was suggested in 1990 by William H. Inmon [6]. DWs are data aggregates that integrate time series without assuming data deletion or updating, and this structure has now been introduced by many organizations, mainly in the form of core database systems for business use. The Dimensional Model was published several years after the Data Warehouse structure was proposed. Unlike the Relational Model and other structures based on fundamental logic, the DW concept indicates a data construct handled in relational databases.

6. Dimensional Data Model

The Dimensional Model is a data model proposed in the writings of Ralph Kimball [7]. Moreover, recently the term “data mart” has become widely used almost synonymously with Data Warehouse [8]. In recent years, in particular, data marts have often been constructed with the aim of speeding up database queries in the development of software to enable easier exploratory analysis, such as Business Intelligence (BI) tools [9]. The following explanation is based on the i2b2 Star Schema, constructed by the i2b2 tranSMART Foundation which supplies common platforms for precision medicine [10]. The Star Schema has the most typical structure of a Dimensional Model and is well known. As shown in Figure 1, the Dimensional Model is composed of two individual tables, Fact and Dimension. The Fact table is generally the largest table and includes the attributes to be analyzed. The Dimension table is a table that contains analyzable attributes by virtue of its linkage to the Fact table and corresponds to the master table of a regular database. In most cases, its size is therefore smaller than that of the Fact table. In the following example, Fact is only the observation fact table, but several such tables may exist. The attributes listed in the Fact table have already been tabulated, and if any other data are needed, these can be obtained by linking to an appropriate Dimension table as required. However, although the attributes listed in the Fact table are suitable for queries involving rapid searches and aggregations, they do not possess the means to deal with other queries with more complex conditions. Generally speaking, however, as almost all the data required or decision-making in business is patterned, most cases can be dealt with by the use of several Fact tables. Constructing a DW for analysis purposes is now almost a common-sense process, and since the late 1990s, numerous studies have been published, mainly concerning the Dimensional Model that is now the mainstream model [11,12]. Although it was originally devised for decision-making in business, today DW is widely used for decision-making in many other areas, such as government agencies. Medicine is one area in which it is also used for purposes other than decision-making. The demand for DW architecture in medical fields can be broadly divided into analysis emphasizing searches and aggregations, and analysis emphasizing statistical analyses, such as logistic analysis and hypothesis testing. Exploratory analysis is a typical example of the former [13]. Exploratory analysis is a data-driven analysis method that does not assume a hypothesis but discovers unanticipated correlations through trial and error, and in recent years, it has been attracting attention in database epidemiological studies as a useful technique for constructing hypotheses. It is also frequently used for diagnostic support in clinical settings, enabling searches, such as those for similar cases or the previous use of clinical procedure, and has also been developed as a decision-making tool [14].

Table 1: Example of propositions.

Patient	Sex	Age	Diagnosis date	Disease name	Death date
X1	Female	57	2007-09-04	Stroke	2007-10-15
X2	Male	62	1985-06-05	Liver cancer	1988-11-30
X3	Female	89	2013-02-15	Nontuberculous mycobacteria	2016-03-12
X4	Male	54	1988-02-03	Cerebral hemorrhage	1988-02-05

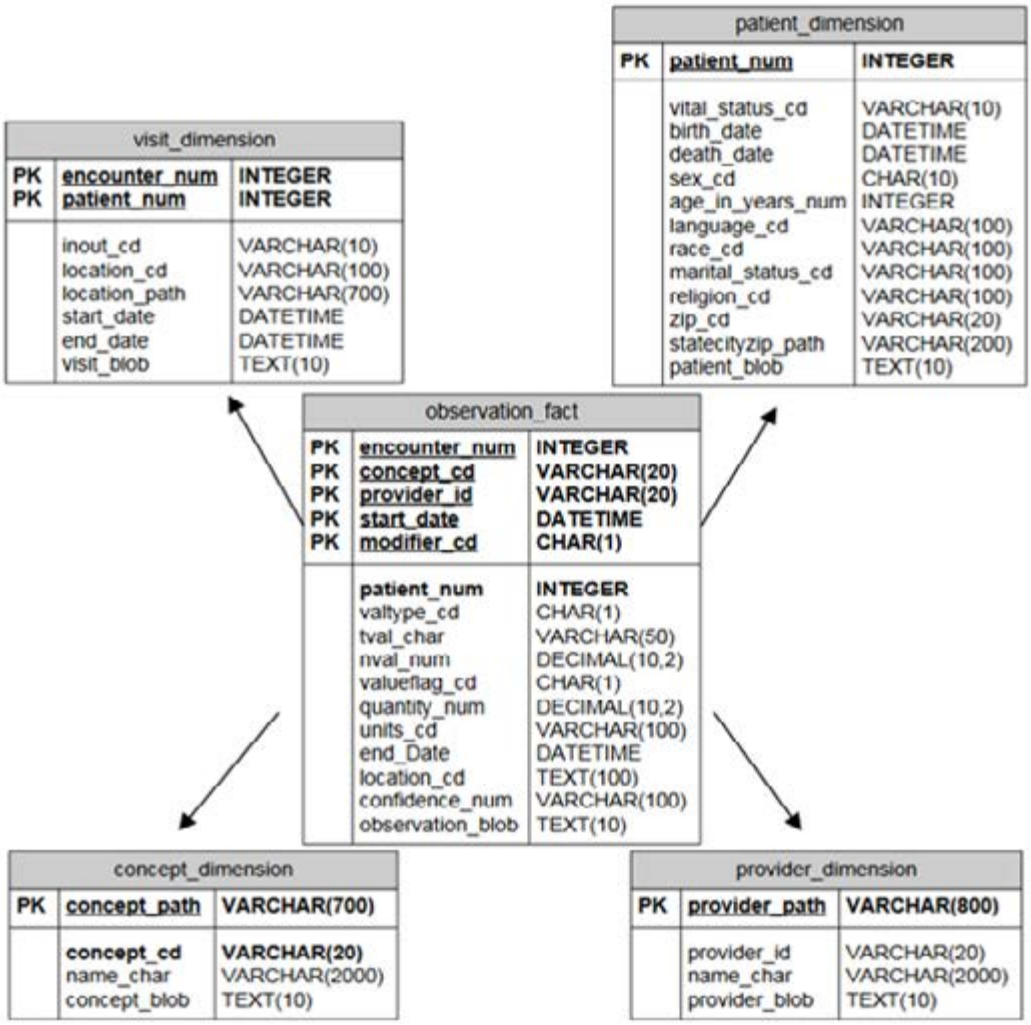

Figure 1: Star Schema of the i2b2 (taken from [10]).

7. Conclusion

In the process of data handling of medical research using a database, it is necessary to handle a large amount of data in one time, so the processing speed is extremely important. If you're dealing with the amount of data that fits in your computer's main memory, it's best to use a programming language. However, when dealing with large amounts of data that does not fit in main memory, a relational database is the best choice. Discussions based on relational models such as data warehouses will continue to be active, and many technologies suitable for analysis and data handling for medical research will be created in the coming decades.

8. Disclosure

The authors declare no conflicts of interest with respect to this research and paper.

9. Funding

This article was performed with the assistance of a research grant from the Ministry of Education, Culture, Sports, Science, and Technology (grant number: 20K23159) and AMED under Grant Number 21lk1503009j0005.

10. References

1. Codd EF. A relation model for large shared data banks. Comm ACM. 1970; 13(6):159-176.
2. Shigeki Y. Logic: An Introduction: University of Tokyo press. 1994.
3. BARC. Relational databases.
4. Standardization IOf. ISO/IEC 9075-1:2016 Information technology -- Database languages -- SQL -- Part 1: Framework (SQL/Framework). 2016: 78-79.
5. Codd EF. Providing OLAP to User Analysts: An IT Mandate. New York: Codd & Associates. 1993.
6. Inmon WH. Building the Data Warehouse. New York: Wiley. 2005.
7. Ralph K, Margy R. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. New York: Wiley. 2013.

8. Loshin D. Business intelligence: the savvy manager‘s guide. Newnes. 2012.
9. Chuck B, Daniel F, Amit G, Carlos M, Stanislav V. Dimensional Modeling: In a Business Intelligence Environment. IBM International Technical Support Organization. 2006.
10. Shawn M, Vivian G. Working assumptions of i2b2 data. 2019.
11. Tebourski W, Kara. BA, Wahiba, Ben Ghezela H. A survey on medical datawarehouse. International Conference on Control, Decision and Information Technologies (CoDIT) 2013: 933-936.
12. Romero O, Abello A. A survey of multidimensional modeling methodologies. Int J Data Warehousing Mining (IJDWM). 2009; 5(2):1-23.
13. John T. Exporative Data Analysis. New York: Addison Wesley PUB CO INC; 1977.
14. Chan CL, Van PD, Yang N-P. Building a Decision Support Tool for Taiwan’s National Health Insurance Data – An Application to Fractures Study. Intellig Dec Tech.