

Dimensional Modeling for Data Warehouse

Umashanker Sharma, Anjana Gosain
GGS, Indraprastha University, Delhi

Abstract

Many surveys indicate that a significant percentage of DWs fail to meet business objectives or are outright failures. In this paper, we are describing about multidimensional data model for the DWs, which will help in surround facts with as much relevant context dimensions. Multidimensionality is just a design technique that separates the information into facts and dimensions. It gives rise to schema with star shape, Consists of a fact table with a single table for each dimension. Snowflake Schema is a variation of star schema, in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them. Multidimensional model present information to the end-user in a way that corresponds to his normal understanding of his business, key figures or facts from the different perspectives that influence them .We are surveyed on present different multidimensional data cubes, which are the basic logical model for OLAP applications.

Key Words: Multidimensional Modeling, data warehouse, OLAP, Fact Schema,

1. Introduction

DW¹ technology is in full swing in industry as major business intelligence technology. Many business intelligence applications currently run at companies not only demand more capacity, but also new methods, models, techniques or architectures to satisfy these new needs some of the hot topics in DW. Traditional databases are not optimized for data access only they have to balance the requirement of data access with the need to ensure integrity of data. Most of the times the DW users need only read access but, need the access to be fast over a large volume of data. Most of the data required for DW analysis comes from multiple databases. The concept of DWs emerged during the nineties as an integrated data collection system for companies oriented to decision making. A data warehouse is a set of data and technologies aimed at enabling the executives, managers and analysts to make better and faster decisions. DWs to manage information efficiently as the main organizational asset. The principal role of DW in taking strategic decisions, quality is fundamental. DWs are databases consisting of cleansed, reconciled, and enhanced data integrated into logical business subject areas for improving decision making. A DW is a database that

stores information in order to satisfy decision-making requests. This kind of DB² has the following particular features. It contains data that is the result of transformations, quality improvement, and integration of data that comes from operational bases, also including indicators that give it additional value. The DWs have to support complex queries however its maintenance does not suppose transactional load. These features cause the design techniques and the used strategies to be different from the traditional ones.

- **Enterprise DW** -An enterprise DW provides a central database for decision support throughout the enterprise.
- **Operational Data Store**-This has a broad enterprise wide scope, but unlike the real. Enterprises DW, data is refreshed in near real time and used for routine business activity.
- **Data Mart** -Data mart is a subset of data warehouse and it supports a particular region, business unit or business function.

1.1- Characteristics of DW- The main characteristics of data warehouse are:

Separate, The DW is separate from the operational systems in the company. It gets its data out of these legacy systems. Available. The task of a DW is to make data accessible for the user. Integrated. The basis of this integration is the standard company model

A lot of history. Questions have to be answered; trends and correlation's have to be discovered. They are time stamped and associated with defined periods of time.

Subject oriented. Most of the time oriented on the subject 'customer'.

Not dynamic. When the data is updated, it is done only periodical, but not as on individual basis.

Aggregation performance. The data which is requested by the user has to perform well on all scales of aggregation.

Consistency. Structural and contents of the data is very important and can only be guaranteed by the use of metadata: this is independent from the source and collection date of the data.

1.2- The design process of data warehouse consists of following steps, Apply one of the existing methodologies for designing the DW schema starting from the source conceptual schema. After obtaining the DW logical schema, build it through application of transformations to the source logical schema, and apply

other necessary transformations so that the DW schema is refined according to the requirements.

2. Dimensional modeling

Dimensional modeling is a technique for conceptualizing and visualizing data models as a set of measures that are described by common aspects of the business. Dimensional modeling has two basic concepts.

Facts:

- A fact is a collection of related data items, consisting of measures.
- A fact is a focus of interest for the decision making process.
- Measures are continuously valued attributes that describe facts.
- A fact is a business measure.

Dimension:

- The parameter over which we want to perform analysis of facts
- The parameter that gives meaning to a measure number of customers is a fact, perform analysis over time.

Dimensional modeling also has emerged as the only coherent architecture for building distributed DW systems. If we come up with more complex questions for our warehouse which involves three or more dimensions. This is where the multi-dimensional database plays a significant role analysis. Dimensions are categories by which summarized data can be viewed. Cubes are data processing units composed of fact tables and dimensions from the data warehouse. Dimensional modeling also has emerged as the only coherent architecture for building distributed data warehouse systems.

3. Multi-Dimensional Modeling

Multidimensional database technology has come a long way since its inception more than 30 years ago. It has recently begun to reach the mass market, with major vendors now delivering multidimensional engines along with their relational database offerings, often at no extra cost. Multi-dimensional technology has also made significant gains in scalability and maturity. Multidimensional data model emerged for use when the objective is to analyze rather than to perform on-line transactions.

Multidimensional model is based on three key concepts:

- Modeling business rules
- Cube and measures
- Dimensions

Multidimensional data-base technology is a key factor in the interactive analysis of large amounts of data for decision-making purposes. Multidimensional data model is introduced based on relational elements. Dimensions are modeled as *dimension relations*.

Changing from one dimensional hierarchy to another is easily accomplished in a data cube by a technique called pivoting also called rotation. In this technique the data cube can be thought of as rotating to show different orientation of the axes. The multidimensional model transforms the visualization of a schema into a more business-focused environment. All these structures cubes, measures and dimensions interact with each other to provide an extremely powerful reporting environment. Each object adds new levels of interactivity when it is fully exploited. A multi-dimensional model is extensible. Users need to be able to add additional calculations quickly and easily. The creation of calculations needs to insulate the user from the structure of the associated cubes. There are no restrictions on the usage of these calculated measures. Dimensions can also be filtered based on a condition that is driven by a calculated measure. Most of the multidimensional database systems used in business analysis and decision support applications is special. Generally, they can be classified into two categories: 1st is the special relational database systems which create multi-dimensional schemas such as star schema and snowflake schema by applying the mature theory of relational database systems, the 2nd is the multi-dimensional database systems which are designed specially for online. All dimensional tables are directly connected with the factual table and do not generate connections with other dimensional tables. However, it will need to divide one dimension into many dimensions sometimes. Such structure is called the snowflake mode, which arises from some dimensions standardized. Relational database systems are suitable for applications of OLTP⁴, but it does not meet the requirements of online analytical processing applications. Relational OLAP³ systems which originally are ORDBMS can only be classified as relational database systems, because after changing into systems supporting OLTP applications, only the relational functions are used, the object features are disappeared. A multidimensional database is a type of database that is optimized for DW and OLTP applications. Multidimensional databases are frequently created using input from existing RDs, a multidimensional database allows a user to ask questions and questions related to summarizing business operations and trends. An OLTP application that accesses data from a multidimensional database is known as a multidimensional OLTP application. A multidimensional database or a multidimensional database management system implies the ability to rapidly process the data in the database so that answers can be generated quickly. A number of vendors provide products that use multidimensional databases. Approaches to how data is stored and the user interface vary. To multidimensional database systems, applications are hampered because uniform standard does not exist. They are special database systems which do not support comprehensive query

languages similar to structured query language. They can not treat all dimensions and measures symmetrically the definition of multidimensional schema describes multiple levels along a dimension, and there is at least one key attribute in each level that is included in the keys of the star schema in RD systems. Multidimensional database enable end-users to model data in a multidimensional environment. This is real product strength, as it provides for the fastest, most flexible method to process multidimensional requests.

- Data mining applications seek to discover knowledge by searching semi-automatically for previously unknown patterns and relationships in multidimensional databases. OLAP software enables analysts, managers, and executives to gain insight into the performance of an enterprise through fast access to a wide variety of views of data organized to reflect the multidimensional nature of the enterprise data.

3.1:-The Goals of Multi-Dimensional Data Models

- To present information to the end-user in a way that corresponds to his normal understanding of his business, key figures or facts from the different perspectives that influence them.
- To offer the basis for a physical implementation that the software recognizes (the OLAP engine), thus allowing a program to easily access the data required. Modeling principles of multidimensional DB are, the factual table will be analyzed firstly which include some primary dimensional yards and facts. The primary yards are the primary code to link with the dimensional tables. And facts are monitoring data representing actual measured values to be used in monitoring data modeling.

3.2:- Three important application areas.

- DWs are large repositories that integrate data from several sources in an enterprise for analysis.
- OLAP systems provide fast answers for queries that aggregate large amounts of detail data to find overall trends, efficiency to the system. So build the factual table according to the minimum detail level.

In some cases dimension table can reduce the size of factual table and the amount of repeated data in the factual table. If there is many to much relationship between some dimension table and factual table, it will need to standardize dimension through snowflake mode. When a snowflake mode is used, it needs to minimize the number of dimensional tables linking with dimensional tables which links with the factual table. Second, monitoring data should be original in the factual table. They should not be summarized. And these data should be at the same level and own a necessary size. Third, the reference integrity of monitoring data should be ensured. All monitoring information exists in the factual table should emerge in dimensional tables when designing. Multidimensional data models have three important

application areas within data analysis. Modeling process includes functional requirement should be selected first when planning a multidimensional DB system. Then it needs to determine the monitoring numerical values to identify the functions. Analysis dimension should be constructed to determine the perspectives to analyze the monitoring data and values etc. After the dimensions have been managed it is necessary to determine the size. Finally the logical model of the database should be defined when analysis objects, perspectives and detailed level has been determined.

3.3:- Logical Multidimensional Model

The multidimensional data model is important because it enforces simplicity. As Ralph Kimball states in his landmark book, *The DW Toolkit*:” The central attraction of the dimensional model of a business is its simplicity that simplicity is the fundamental key that allows users to understand DBs, and allows software to navigate databases efficiently.” The multidimensional data model is composed of logical cubes, measures, dimensions, hierarchies, levels, and attributes. The simplicity of the model is inherent because it defines objects that represent real-world business entities. Analysts know which business measures they are interested in examining, which dimensions and attributes make the data meaningful, and how the dimensions of their business are organized into levels and hierarchies. Multidimensional data cubes, are the basic *logical* model for OLAP applications. The focus of OLAP tools is to provide multidimensional analysis to the underlying information. To achieve this goal, these tools employ multidimensional models for the storage and presentation of data.

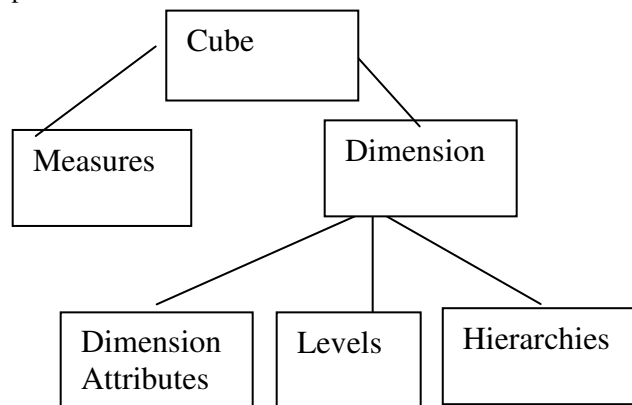


Figure1: Diagram of logical Multi dimensional model

A logical model (figure1) for cubes based on the key observation that a cube is not a self-existing entity, but rather a view over an underlying data set. Logical cubes provide a means of organizing measures that have the same shape, that is, they have the exact same dimensions. Measures in the same cube have the same relationships to

other logical objects and can easily be analyzed and displayed together. Measures populate the cells of a logical cube with the facts collected about business operations. Measures are organized by dimensions, which typically include a Time dimension. Multidimensional analysis allows users to interact with each dimension in isolation to determine which dimension values should be displayed in the target presentation. Dimensional queries are also rarely based on a single step, but consist of a number of steps. The relational model forces users to manipulate all the elements as a whole, which tends to lead to confusion and unexpected result sets. In contrast, the multi-dimensional model allows users to filter each dimension in isolation and uses more friendly terms such as Add, Keep and Remove. Users can quickly and easily create multi step queries. The multi-dimensional query model has one significant advantage over the relational query model. Each dimension can be queried separately. This allows users to breakdown what would be a very complex query into simple manageable steps. The multidimensional model also provides powerful filtering. Additionally, it is possible to create conditions based on measures that are not part of the final report. Because the dimensional query is independent of the filters, it allows complete flexibility in determining the structure of the condition. The relational implementation of the multidimensional data model is typically a star schema, or a snowflake schema.

3.4:- Star schema

Star schema consists of a fact table with a single table for each dimension. Star Schema is the special design technique for multidimensional data representations. It Optimizes data query operations instead of data update operations. Star Schema is a relational database schema for representing multidimensional data. It is the simplest form of data warehouse schema that contains one or more dimensions and fact tables. It is called a star schema because the entity-relationship diagram between dimensions and fact tables resembles a star where one fact table is connected to multiple dimensions. The center of the star schema consists of a large fact table and it points towards the dimension tables. The advantage of star schema is slicing down, performance increase and easy understanding of data.

Steps in designing star schema

- Identify a business process for analysis.
- Identify measures or facts.
- Identify the dimensions for facts.
- List the columns that describe the each dimension.
- Determine the lowest level of summary in a fact table.

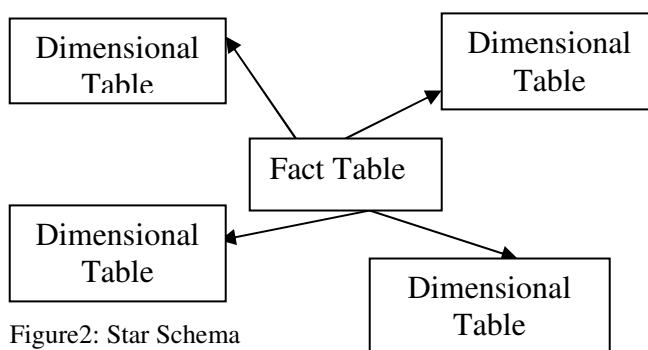


Figure2: Star Schema

3.5:- Snowflake schema

Snowflake schema is a variation on the star schema in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them. Some installations are normalizing DW up to third normal form; so that they can access the DW efficiently .i.e. dimension table hierarchies are broken into simpler tables.

Important aspects of Star Schema & Snow Flake Schema

In a star schema every dimension will have a primary key and also a dimension table will not have any parent table. Whereas in a snow flake schema, a dimension table will have one or more parent tables. Hierarchies for the dimensions are stored in the dimensional table itself in star schema. Whereas hierarchies are broken into separate tables in snow flake schema. These hierarchies help to drill down the data from topmost hierarchies to the lowermost hierarchies. Snowflake schema is the normalized form of star schema.

4. Conclusion

This paper helps us to understanding the concept of multidimensional model. Multi-dimensional data which will help in surround facts with as much relevant context dimensions. Finally this paper helps to compare the various multi-dimensional modeling according to the multi-dimensional space, language aspects and physical representation.

References

- [1] INMON, W.H.: 'Building the data warehouse' (John Wiley and Sons, September 1998, 41, (9), pp. 52-60 New York, 1997)
- [2] S.Kelly.Data Warehousing in Action.John Wiley & Sons (1997).
- [3] Kimball, R. "The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses". John Wiley and Sons 1996. ISBN 0-471-15337-0
- [4] S.Chaudhuri,U.Dayal.An overview of data warehousing and OLAP technology. SIGMOD Record 26,1 (1997).
- [5] G.Colliat.OLAP, relational and multi-dimensional database systems.SIGMOD Record 25, 3 (1996)

- [6] M. Golfarelli, D. Maio, and S. Rizzi. The dimensional fact model: a conceptual model for data warehouses. *IJCIS*, 7(2-3):215–247, 1998.
- [7] M. Jarke, M. Lenzerini, Y. Vassilious, and P. Vassiliadis, editors. *Fundamentals of Data Warehousing*. Springer-Verlag, 1999.
- [8] L. Cabibbo and R. Torlone. A logical approach to multidimensional databases. In *Proc. of EDBT-98*, 1998.
- [9] E. Franconi and U. Sattler. A data warehouse conceptual data model for multidimensional aggregation. In *Proc. of the Workshop on Design and Management of Data Warehouses (DMDW-99)*, 1999.
- [10] McGuff, F. “Data Modeling for Data Warehouses” October, 1996 from <http://members.aol.com/fmcguff7dwmodel/dwmodel.html>
- [11] Gyssens, M. and Lakshmanan, L.V.S. “A foundation for multi-dimensional databases,” Technical Report, Concordia University and University of Limburg, February 1997.
- [12] M.Blaschka,C.Sapia,G.H’ ofling, and B. Dinter. Finding Your Way through Multidimensional Data Models. In *DEXA ’98*, pages 198–203, 1998. <http://www.pentaho.org> (16.06.2006), 2006.
- [13] Antoaneta Ivanova, Boris Rachev Multidimensional models - Constructing DATA CUBE International Conference on Computer Systems and Technologies - CompSysTech’2004
- [14] Multidimensional Database Technology by *Torben Bach, Pedersen, Christian S. Jensen*, Aalborg University
- [15] Rakesh Agrawal, Ashish Gupta, and Sunita Sarawagi. Modeling multidimensional databases. Research Report, IBM Almaden Research Center, San Jose, California, 1996.
- [16] P. Vassiliadis and T.K. Sellis, “A Survey of Logical Models for OLAP Databases,” *ACM SIGMOD Record*, vol. 28, no. 4, 1999.
- [17] 21. L. Cabibbo and R. Torlone. A logical approach to multidimensional databases. In *Proc. of EDBT-98*, 1998