Techniques for Data Warehouse and Data Modeling

[1]YERRA SANKAR RAO,
Gandhi Institute of Excellent Technocrats, Bhubaneswar, India
[2]SMITA SINGH MAHARATHA,
Samanta Chandra Sekhar Institute of Technology and Management, Koraput, Odisha, India

Abstract
The conceptual Entity-Relationship (ER) is extensively used for database design in relational database environment, which emphasized on day-to-day operations. Multidimensional (MD) data modeling, on the other hand, is crucial in data warehouse design, which targeted for managerial decision support. It supports decision making by allowing users to drill-down for a more detailed information, roll-up to view summarized information, slice and dice a dimension for a selection of a specific item of interest and pivot to re-orientate the view of MD data. When designing a MD model regardless whether it is a star or snowflake schema, it involves the identification of a fact, dimensions and measureattributes. This paper will explore on how the Multidimensional model can be used as the yardstick of data warehouse design instead of ER Model.
Keywords: Entity-Relationship Model, Multidimensional Model, Fact, Dimensions, Attributes.

I. INTRODUCTION

Data ModelingTechniques
Two data modeling techniques that are relevant in a data warehousing environment are ER modeling and Multidimensional modeling.
ER modeling produces a data model of the specific area of interest, using two basic concepts: entities and the relationships between those entities. The ER model is an abstraction tool because it can be used to understand and simplify the ambiguous data relationships in the business world and complex systems environments.
Multidimensional modeling uses three basic concepts: measures, facts, and dimensions. Multidimensional modeling is powerful in representing the requirements of the business user in the context of database tables.
Both ER and Multidimensional modeling can be used to create an abstract model of aspecific subject.

1.1 ERModeling

An ER model is represented by an ER diagram, which uses three basic graphic symbols to conceptualize the data: entity, relationship, and attribute.

1.1.1 ENTITY

An entity is defined to be a person, place, thing, or event of interest to the business or the organization. An entity represents a class of objects, which are things in the real world that can be observed and classified by their properties andcharacteristics.

1.1.2 RELATIONSHIP

A relationship is represented with lines drawn between entities. It depicts the structural interaction and association among the entities in a model. A relationship is designated grammatically by a verb, such as owns, belongs, and has. The relationship between two entities can be defined in terms of the cardinality. This is the maximum number of instances of one entity that are related to a single instance in another table and vice versa. The possible cardinalities are: one-to-one (1:1), one-to-many (1:M), and many-to-many (M:M).

1.1.3 ATTRIBUTES

Attributes describe the characteristics of properties of the entities. For clarification, attribute naming conventions are very important. An attribute name should be unique in an entity and should be self-explanatory.
When an instance has no value for an attribute, the minimum cardinality of the attribute is zero, which means either nullable or optional.
In ER modeling, if the maximum cardinality of an attribute is more than 1, the modeler will try to normalize the entity and finally elevate the attribute to another entity. Therefore, normally the maximum cardinality of an attribute is 1.

1.2 MultidimensionalModeling

In some respects, Multidimensional modeling is simpler, more expressive, and easier to understand than ER modeling. But, Multidimensional modeling is a relatively new concept and not firmly defined yet in details, especially when compared to ER modeling techniques.

1.2.1 BASICCONCEPTS

Multidimensional modeling is a technique for conceptualizing and visualizing data models as a set of measures that are described by common aspects of the business. It is especially useful for summarizing and rearranging the data and presenting views of the data to support data analysis. Multidimensional modeling focuses on numeric data, such as values, counts, weights, balances, and occurrences
Multidimensional modeling has several basic concepts:

Dimensions

- Facts
- Measures(variables)

1.2.2 FACT

A fact is a collection of related data items, consisting of measures and context data. Each fact typically represents a business item, a business transaction, or an event that can be used in analyzing the business or business processes.
In a data warehouse, facts are implemented in the core tables in which all of the numeric data is stored.

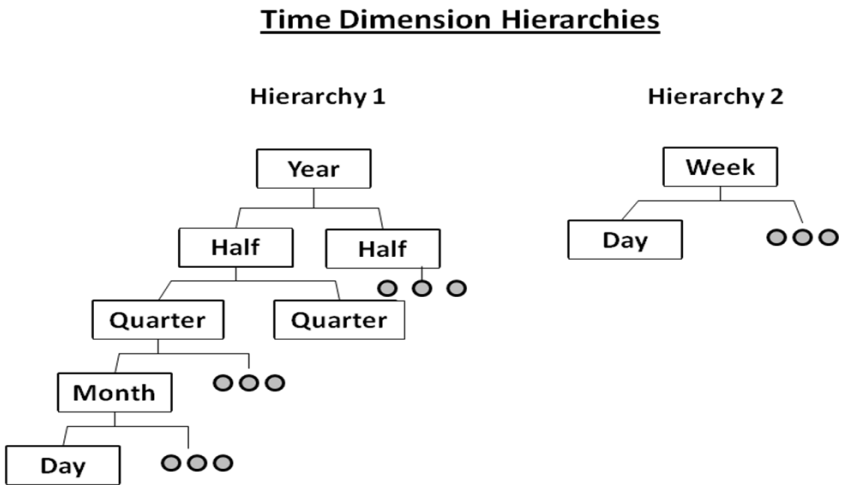
1.2.3 Dimension

A dimension is a collection of members or units of the same type of views. In a diagram, a dimension is usually represented by an axis. In a Multidimensional model, every data point in the fact table is associated with one and only one member from each of the multiple dimensions. That is, dimensions determine the contextual background for the facts. Many analytical processes are used to quantify the impact of dimensions on the facts.
Dimensions are the parameters over which we want to perform Online Analytical Processing (OLAP). For example, in a database for analyzing all sales of products, common dimensions could be:

- Time
- Location/region
- Customers
- Salesperson
- Scenarios such as actual, budgeted, or estimatednumbers

DIMENSION MEMBERS: A dimension contains many dimension members. A dimension member is a distinct name or identifier used to determine a data item's position. For example, all months, quarters, and years make up a time dimension, and all cities, regions, and countries make up a geographydimension.
DIMENSION HIERARCHIES: We can arrange the members of a dimension into one or more hierarchies. Each hierarchy can also have multiple hierarchy levels. Every member of a dimension does not locate on one hierarchy structure. A good example to consider is the time dimension hierarchy as shown in Figure1.

FIGURE 1: Multiple Hierarchies In A Time Dimension



1.2.4 Measure

A measure is a numeric attribute of a fact, representing the performance or behavior of the business relative to the dimensions. The actual numbers are called as variables. For example, measures are the sales in money, the sales volume, the quantity supplied, the supply cost, the transaction amount, and so forth. A measure is determined by combinations of the members of the dimensions and is located onfacts.
Considering Relational context, there are two basic models that are used in dimensional modeling: (i) star model and (ii) snowflake model.
The star model is the basic structure for a dimensional model. It has one large central table (fact table) and a set of smaller tables (dimensions) arranged in a radial pattern around the central table. (We show an example in Figure 2). The snowflake model is the result of decomposing one or more of the dimensions. The many-to-one relationships among sets of attributes of a dimension can separate new dimension tables, forming a hierarchy. (Figure 3 shows an example). The decomposed snowflake structure visualizes the hierarchical structure of dimensions very well.

Figure 2: Star Model

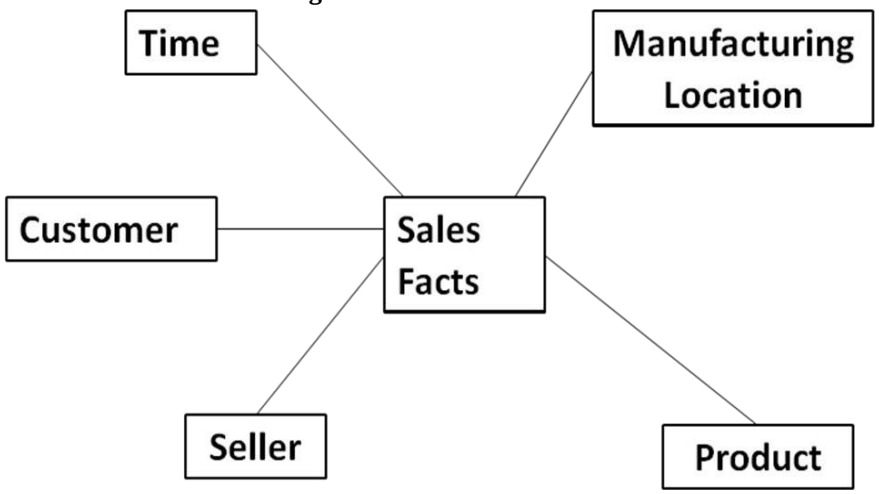
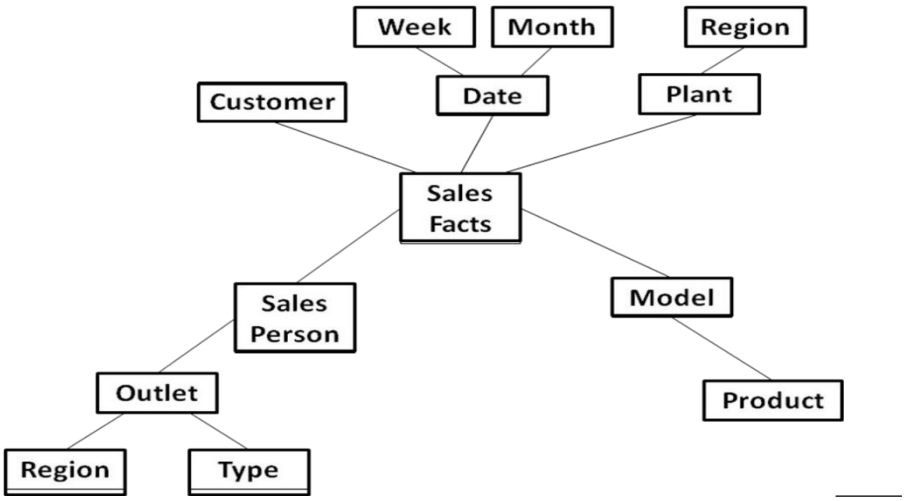


Figure 3: Snow Flake Model



II. Multidimensional Model Vs RelationshipModel

ER is a logical design technique that seeks to remove the redundancy in data. This coupled with normalization of data enables easy maintainability and improves data integrity which is a necessity for transaction processing applications. End user comprehension and the data retrieval are major show stoppers; as such a database is proliferated with dozens of tables that are linked together by a bewildering spider web of joins. Use of the ER modeling technique defeats the basic allure of data warehousing, namely intuitive and high- performance retrieval ofdata.

MD is a logical design technique that seeks to present the data in a standard, intuitive framework that allows for high-performance access. Every Multidimensional model is composed of one table with a multipart key, called the fact table, and a set of smaller tables called dimension tables. Each dimension table has a single-part primary key that corresponds exactly to one of the components of the multipart key in the fact table. This characteristic "star-like" structure is often called a star join.

Each dimensional table is logical and user identifiable and serves a business purpose by serving as an object of interest to the user. It is also maintained by the ETL process of the data ware housing application .Hence it is considered as an internal Logical file and included in the data function count.

III. WHY ER IS NOT SUITABLE FOR DATAWAREHOUSES?

- End user cannot understand or remember an ER Model. End User cannot navigate an ER Model. There is no graphical user interface or GUI that takes a general ER diagram and makes it usable by endusers.
- ER modeling is not optimized for complex, ad-hoc queries. They are optimizedfor repetitive narrowqueries
- Use of ER modeling technique defeats this basic allure of data warehousing, namely intuitive and high performance retrieval of data because it leads to highly normalized relational tables.

IV. CONCLUSION

In this paper, we examine that an E-R structured data warehouse, absent associative entities, i.e. fact tables, is a not viable concept given recent developments in data warehousing. A number of conclusions are supported by the arguments.

- Not every E-R model can be represented as a set of star schemas containing equivalentinformation
- But every properly constructed E-R data warehousing model can be sorepresented
- Many E-R data warehouse models are not properly constructed in that they don't explicitly recognize many-many relations and the need to resolve them with associative entities, i.e. facttables.
- To use data warehousing E-R models specifying atomic data dependency relationships without fact tables is to ensure poor query response performance in large databases, and therefore discourage, and often prevent, execution of a multi-stage analysis process. In effect, it is to make the data warehouse no more than a big staging area for data marts, with no independent analytical function of itsown.
- Given the development of ODSs and non-queryable centralized staging areas for storing, extracted, cleansed, and transformed data and for gathering centralized metadata.
- We don't need another non-queryable staging area called a data warehouse. What we do need, instead, is a dimensionally modeled data warehouse for enterprisewide DSS, prepared to provide the best in query response performance and to support the most advanced OLAP functionality we candevise.

V. REFERENCES

[1] A Conceptual Model for Multidimensional Data By Anand S. Kamble, Department of Information Technology, Government of India, New Delhi,India

[2] An Overview of Data Warehouse Design Approaches and Techniques, Alejandro Gutiérrez, Adriana MarottaInstituto de Computación, Facultad de Ingenieria, Universidad de la República, Montevideo, Uruguay October2000

[3] C. Li and X. S. Wang, "A Data Model for Supporting On-Line Analytical Processing" Fifth International Conference on Information and Knowledge Management, 1996, pp.81–88.

[4] Conceptual Multidimensional Model, By Manpreet Singh, Parvinder Singh, and Suman, World Academy of Science, Engineering and Technology 362007

[5] Data Modeling Techniques for Data Warehousing By Chuck Ballard, Dirk Herreman, Don Schau, Rhonda Bell, Eunsaeng Kim, AnnValencic

[6] Dimensional Modeling and E-R Modeling In The Data Warehouse By Joseph M. Firestone, Ph.D. White Paper No. Eight, June 22,1998

[7] Dimensional Model Data Warehouse: An Overview, Dan Kirpes, Fireman's Fund Insurance Company, Novato,CA

[8] Estimation Model for Data Warehousing Projects Presented In 2nd Annual International Estimation Colloquium 2007,By KarthikeyanSankaran&SujathaSivaraman ,HexawareTechnologies

[9] Kimball, Ralph, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, New York, NY: John Wiley and Sons, Inc., 2002.436pp.

[10] Len Silverston, W. H. Inmon, and Kent Graziano, The Data Model Resource Book (New York, NY: John Wiley & Sons, Inc.,1997).