International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Issue 4, Volume 2 (April 2015) www.ijirae.com

Data Warehouse Schemas

Mandeep Kaur Sandhu
Computer Science Deptt.
SDSPM College for Women, Rayya

Amanjot Kaur
Computer Science Deptt.
SDSPM College for Women,Rayya

Ramandeep Kaur
Computer Science Deptt.
SDSPM College for Women,Rayya

Abstract-- A data warehouse is an integrated set of data, derived basically from operational data to use in decision making strategy and business intelligence using (OLAP) techniques. The words On-Line Analytical Processing (OLAP) bring together a set of tools that use multidimensional modelling in the extraction of information from the Data Warehouse. The multidimensional data model is an integral part of On-Line Analytical Processing, or OLAP. Because OLAP is on-line, it must provide answers quickly; analysts pose iterative queries during interactive sessions, not in batch jobs that run overnight. And because OLAP is also analytic, the queries are complex. The multidimensional data model is designed to solve complex queries in real time. Most of the creation of data multidimensional data warehouses is done manually, but it is a very complex and takes a long time. Despite this, there is no noticeable efforts has been done in order to find a practical solution structured to resolve the issue. As a result, the user has to choose the candidate schema which meets the system requirements.

Keywords—Introduction, Dimensional modeling, Schemas, Star, Snowflake, Fact constellation

I. INTRODUCTION

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

A. 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 and characteristics.

B. 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).

C. 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 null able 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.

A Data Warehouse (DW) is defined as "a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process". Data warehousing is now playing a significant role in strategic decision making. It is gaining importance day by day. It provides a multidimensional view of huge amounts of historical data from operational sources, thus supplying useful information for decision makers to improve their business intelligence which has become an integral part of decision making strategy. It is a collection of integrated, subject-oriented databases designed to support the DSS function, where each unit of data is non-volatile and relevant to some moment in time. There are two forms for data management i.e. operational databases and data warehouse. The operational databases are where the data is put in. Users of this type almost deal with one record at a time and they usually perform the same tasks. The data warehouse is where we get the data out. Users of this type almost deal with set of row at a time and their questions require that thousands of rows be fetched into an answer set. The data stored in the warehouse is uploaded from the operational systems such as marketing, sales, etc. The data may pass through an operational data store for additional operations before it is used in the DW for reporting. The main **objectives** of data warehouse are:

A. Data availability: Data is structured to be ready and available for analytical processing activities such as OLAP, data mining, querying, reporting and any other decision supporting applications.

B.Easily accessible: Data warehouse content must be understandable and labelled meaningfully. It must also return query results with minimal wait times.

C.Consistently: Warehouse must be reliable and its data must be collected from a variety of sources in a relation, cleansed and quality checked.

D.Adaptive to change: The data warehouse must be able to handle changes. The existing data and applications should not be affected with changes, asking new questions and adding new data to the warehouse.

E.Security and protection: The data warehouse must control access to the confidential information.

F.Improve decision making: The data warehouse must have a trusted data to support decision making.

G.Subject-Oriented: A data warehouse is organised around major subjects such as customer, supplier, product and sales.Hence, data warehouse typically provide a simple and brief view around particular subject issues by excluding data that are not useful in the decision support process.

H.Time Variant: Historical data is kept in data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months or even older data from a data warehouse.

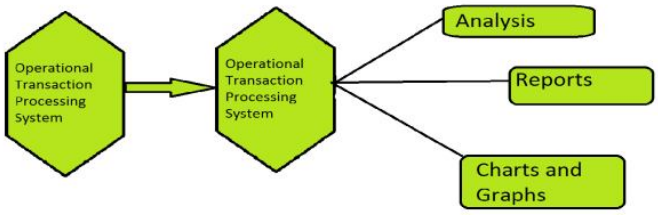

Fig 1. Data warehouse Design

I. Data Warehouse Dimensional Modelling

Dimensional modelling is the design concept used by many data warehouse designers to build their data warehouse. Dimensional model is the basic data model used by many of the commercial OLAP products available today in the market. Dimensional modelling (DM) names a set of techniques and concepts used in data warehouse design. It is considered to be different from entity-relationship modelling (ER). It is simpler, more expressive and easier to understand than ER modelling. It 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.

Dimensional modelling focuses on numeric data such as values, counts, weights, balances, and occurrences.It does not necessarily involve a relational database. The same modelling approach, at the logical level, can be used for any physical form, such as multidimensional database or even flat files. Dimensional modelling always uses the concepts of facts (measures) and dimensions (context). Facts are typically (but not always) numeric values that can be aggregated and dimensions are groups of hierarchies and descriptors that define the facts. For example, sales amount is a fact; timestamp, product, register, store, etc. are elements of dimensions. Dimensional models are built by business process area, e.g. store sales, inventory, claims, etc. Because the different business process areas share some but not all dimensions, efficiency in design, operation, and consistency, is achieved using conformed dimensions, i.e. using one copy of the shared dimension across subject areas. In this model, all data is contained in two types of tables called **Fact** Table and**Dimension** Table.

Fact Table:

In a Dimensional Model, Fact table contains the measurements or metrics or facts of business processes. If your business process is **Sales**, then a measurement of this business process such as "monthly sales number" is captured in the fact table. In addition to the measurements, the only other things a fact table contains are foreign keys for the dimension tables. The fact is a set of related data, contains analytical context data and measures. It used to represents business items or business transactions. A fact table is the central table in a star schema of a data warehouse. A fact table stores quantitative information for analysis and is often denormalized. Suppose an electronic shop sells its product. Thus, every sale is a fact that happens and the fact table is used to record these facts. For example:

Item_no.	Branch_code	Location	Unit_Sold
410	1239	Amritsar	10
472	4568	Patiala	40
235	4893	Ludhiana	87
389	3297	Barnala	58

Fig 2. Fact Table

Dimension Table:
In a Dimensional Model, frameworks of the measurements are represented in dimension tables. A dimension table is a table in a star schema of a data warehouse. A dimension table stores attributes or dimensions that describe the objects in a fact table. From the above example Item_no dimension, the attributes can beItem_name, supplier, etc. Generally the Dimension Attributes are used in report labels, and query constraints such as *where Supplier='Amit'*.

Item_no	Item_name	Supplier
410	LED	Amit
472	Refrigerator	Suma
235	A.C	Mahesh
389	Washing Machine	Zuni

Fig 3. Dimension Table

The dimension attributes also contain one or more hierarchical relationships. Before designing your data warehouse, you need to decide what this data warehouse contains. The measure is a numeric attribute in the fact table which illustrates the behaviour of the business of the dimensions.Say if you want to build a data warehouse containing monthly sales numbers across multiple store locations, across time and across products then your dimensions are:

1. *Location*
2. *Time*
3. *Product*

Each dimension table contains data for one dimension. In the above example you get all your store location information and put that into one single table called**Location**. Your store location data may be spanned across multiple tables in your OLTP system (unlike OLAP), but you need to de-normalize all that data into one single table. A data warehouse requires concise, subject oriented schemas that facilitate on-line analysis. The most popular data model for data warehouse is multidimensional model. The common models used are:

II. Star Schema

In data warehousing, a star schema is the simplest form of dimensional model, in which data is organized into facts and dimensions. A star schema is the simplest form of a dimensional model, in which data is organized into *facts* and *dimensions*. A fact is an event that is counted or measured, such as a sale or login. A dimension contains reference information about the fact, such as date, product, or customer. Star schema has become a common term used to connote a dimensional model. Database designers have long used the term star schema to describe dimensional models because the resulting structure looks like a star. A star schema is characterized by one or more very large fact tables that contain the primary information in the data warehouse, and a number of much smaller dimension tables (or lookup tables), each of which contains information about the entries for a particular attribute in the fact table . The main feature of a star schema is a fact table at the centre surrounded by dimensional tables; each one contains information about the entries for a particular attribute in the fact table. The following diagram illustrates the star schema of a company which sells various products:

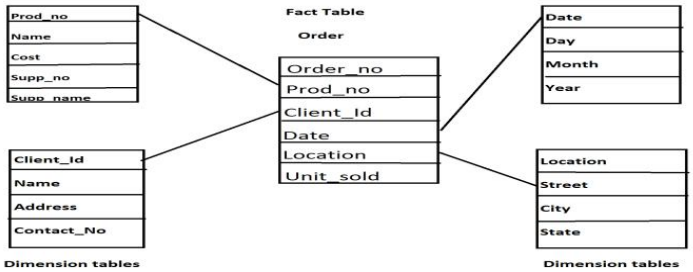

Fig 4. Star Schema

III. Snowflake Schema

The snowflake schema is an extension of the star schema, where each point of the star explodes into more points. In a star schema, each dimension is represented by a single dimensional table, whereas in a snowflake schema, that dimensional table is normalized into multiple lookup tables, each representing a level in the dimensional hierarchy. The snowflake schema architecture is a more complex because the dimensional tables are normalized. It is an enhancement of star schema. It normalizes dimensions to eliminate redundancy. The decomposed snowflake structure visualizes the hierarchical structure of dimensions very well. The snowflake model is easy for data modellers to understand and for database designers to use for the analysis of dimensions. The main advantage of the snowflake schema is the improvement in query performance due to minimized disk storage requirements and joining smaller lookup tables. The main disadvantage of the snowflake schema is the additional maintenance efforts needed due to the increase number of lookup tables. The diagram of snowflake schema is as follows:-

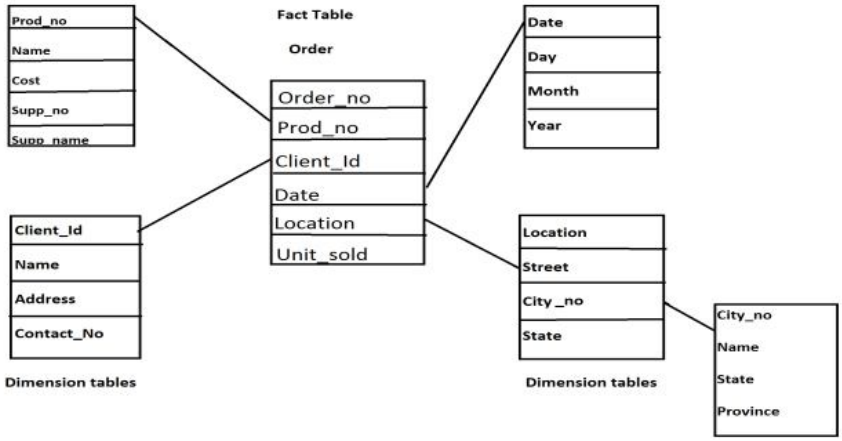

Fig 5.Snowflake Schema Diagram

IV. Fact Constellation Schema

For each star schema or snowflake schema it is possible to construct a fact **constellation schema**. This kind of schema can be viewed as a collection of stars and hence it is called Fact Constellation. This schema is more complex than star or snowflake architecture, which is because it contains multiple fact tables. This allows dimension tables to be shared amongst many fact tables. That solution is very flexible, however it may be hard to manage and support. In a fact constellation schema, different fact tables are explicitly assigned to the dimensions, which are for given facts relevant. This may be useful in cases when some facts are associated with a given dimension level and other facts with a deeper dimension level. The fact constellation architecture contains multiple fact tables that share many dimension tables. It is possible to construct fact constellation schema by splitting the original star schema into more star schemes each of them describes facts on another level of dimension hierarchies. The dimensions in this schema are large. They must be split into independent dimensions based on the levels of hierarchy. It is used mainly for the aggregate fact tables and for better understanding.The main disadvantage of the fact constellation schema is a more complicated design because many variants of aggregation must be considered.

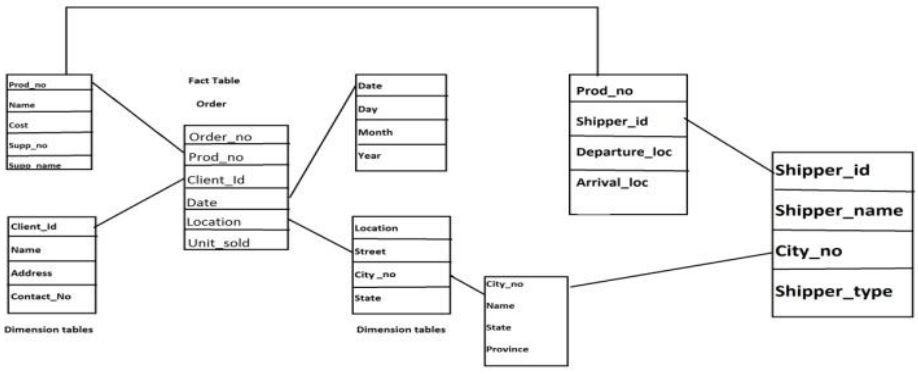

Fig 6. Fact constellation

V. MULTIDIMENSIONAL MODEL Vs RELATIONSHIP MODEL

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 of data. 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.

Conclusion

At last in this paper discuss data warehouse schemas and different types of multidimensional schemas such as star schema, snowflake schema and fact constellation. The advantage of using these schemas is that they are simpler and communicative, easy to read then E-R models. It is useful for terse and rearranging the data and presenting views of the data to support data analysis. Data warehouses separate analysis workload from transaction workload and enable an organization to consolidate data from several sources. Dimensional modeling is a logical design technique for structuring data so that it's intuitive to business users and delivers fast query performance. Data presented to the business intelligence tools must be grounded in simplicity to stand any chance of success. Simplicity is a fundamental requirement because it ensures that users can easily understand databases, as well as allows software to efficiently navigate databases.

References

[1] Soumya Sen, Ranak Ghosh, Debanjali Paul, Nabendu Chaki "Integrating related XML data into multiple data warehouse schemas" research paper 2012-2013.
[2] MS.Alpa R. Patel, "Data Modeling techniques for data warehouse" International Journal of Multidisciplinary Research, Vol.2 Issue 2, February 2012, ISSN 2231 5780.
[3] Anirban Sarkar "Data Warehouse Requirements Analysis Framework: Business-Object Based Approach" International Journal of Advanced Computer Science and Applications, Vol. 3, No. 1, 2012.
[4] Keshav Dev Gupta, Jyoti Gupta, 3Prakati Prasoon "Novel Architecture with Dimensional Approach of Data Warehouse" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.
[5] Frank S.C. Tseng, Chia Wei Chen: Integrating heterogeneous data warehouses using XML technologies, Journal of Information Science Volume-31, Issue:3 (June 2005) Page-209-229
[6] Boris Vrdoljak, Marko Banek, and Stefano Rizzi: Designing Web Warehouses from XML Schemas Y. Kambayashi, M. Mohania, W. Wöß (Eds.): DaWaK 2003, LNCS 2737, pp. 89-98, 2003. SpringerVerlag Berlin Heidelberg 2003
[7] Wolfgang Hummer, Andreas Bauer, Gunnar Harde: XCube – XML For Data Warehouses, DOLAP'03, November 7, 003, USA.
[8] M. Golfarelli, S. Rizzi, and B. Vrdoljak, .Data warehouse design from XML sources., Proc. DOLAP'01, Atlanta, pp. 40-47, 2001.
[9] Data Mining Concepts and Technique,2nd Edition, Jiawei Han and Micheline Kamber, Morgan Kaufmann Publisher.
[10] Daneva, M., Wieringa R "Requirements engineering for cross-organizational ERP implementation undocumented assumptions and potential mismatches", 13th IEEE International Conference on Requirements Engineering, 2005.
[11] The Data Warehouse Toolkit: The Complete Guide to Dimensional Modelling 2nd Edition, Ralph Kimball and Margy Ross, John Willy & Sons.
[12] The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data 1st Edition, Ralph Kimball and Joe Caserta, John Willy & Sons.
[13] Li Jian; Xu Bihua; "ETL Tool Research and Implementation Based on Drilling Data Warehouse" 7th Int'l Conference on Fuzzy Systems and Knowledge Discovery, Chnegdu, China.

