International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 3.45 (SJIF-2015), e-ISSN: 2455-2585
Volume 4, Issue 5, May-2018

DATA WAREHOUSE: DESIGN & ISSUES

SWATI
Assistant Professor, Department of Computer Science
Amity University Haryana, Gurgaon, India
swattiguptta@gmail.com

ABSTRACT

Datawarehouse is an upcoming industry field with many interesting research problem. The data warehouse is always build on specific subject,whose focus lies on a particular subject which could be sales,marketing. In this paper we are discussing about the importance of data warehouse.The data warehouse design and usage has being analyzed.The three main schemas of data warehouse has being discussed in detail.The current research area of data warehouse has being discussed.

Keywords - Data Analysis, Data Warehousing, Data Warehouse Design, Process.

1. INTRODUCTION

A data warehouse is subject oriented, integrated, time variant, & Non volatile collection of huge amount of data which is being stored to support multiple users..[2,3] The complete descriptions of the properties are as follows:

a. Subject Oriented: A data warehouse is generaaly build around major subjects such as sale,customer, product.
b. Integrated:The information in data warehouse is being gathered from multiple homogenous & heterogeneous sources.
c. Time Variant: A data warehouse is collection of historical data where data is collected from the historical perspective(generally 5-10 years)
d. Non Volatile: A data warehouse is separate physical store, which does not require any recovery control & system concurrency .

Data warehouse is a huge repository of data which collects information from various sources, the data is managed for efficient storage and retrieval, and delivers it to large group of people, usually to meet analysis support to make critical decisions. [5,14].

Need For Data Warehouse

The need for data warehouses was emerged due to following reasons:

a. In order to handle complex OLAP query a repository was required.
b. It is being used to store historical data where data is mainatained foe 5-10 years for making strategic decision.
c. It may present relevant information to provide a competitive advantage and make critical adjustments to help win over competitors.
d. Data warehouse can accurately describes the organization which increases business productivity because it is able to quickly and efficient gather large information.
e. It can facilitate by supporting customer relationship management because it provides a consistent view of customers and items across all lines of business, all departments, and all markets.
f. Finally, a data warehouse is design to achieve cost reduction by tracking trends, patterns and exceptions over long periods.

IJTIMES-2018@All rights reserved

2. DATA WAREHOUSE DESIGN PROCESS

There are four different views regarding a data warehouse design[1,8]: the top-down view, the data source view, the data warehouse view, of the information system.

- **The top - Down view** allows the selection of the relevant information necessary for the data warehouse. This information gathered must match with the current and future business needs as it provides the top down view of the data reflection.
- **The Data source view** The information gathered can be documented at various levels of detail and accuracy, from individual data source tables to integrate at various levels of detail and accuracy, form individual data source tables to integrated data source tables.Exposes the information being captured, stored, and managed by operational system.
- **The Data warehouse view** includes fact tables and dimension tables. The dimension table are the perspective with respect to which we store the information about any entity while fact are the numeric measures for measuring any entity.
- **The Business Query View** is the data perspective in the data warehouse form the end-user's view point of view.

So, building and using a data warehouse is a complex task because as it requires business skill, technology skills, and program management skills because a huge massive collection of data is required[10,12]

3. DATA WAREHOUSE:MULTI DIMENSIONAL MODEL

A data warehouse supports multi dimension model, where data is represented in the form of data cube.The data cube supports data in multiple dimension.The three schemas supported by data warehouse are as follows:

3.1 Star Schema

A Star schema[3] is the most simplest form of dimensional model, in which data is organized in the form of star comprises of one central fact table and multiple dimension table.

A fact is the perceptive with respect to which an organization wants to keep record. It describes the characteristics of an entity, we therefore specify that dimension contains reference information about the fact, such as date, product, or customer. It has become a common term used to connect any dimensional model.

Database designers used star schema to describe dimensional models because the resulting structure looks like a star[5,6].The main feature of a star schema is a fact table is 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.

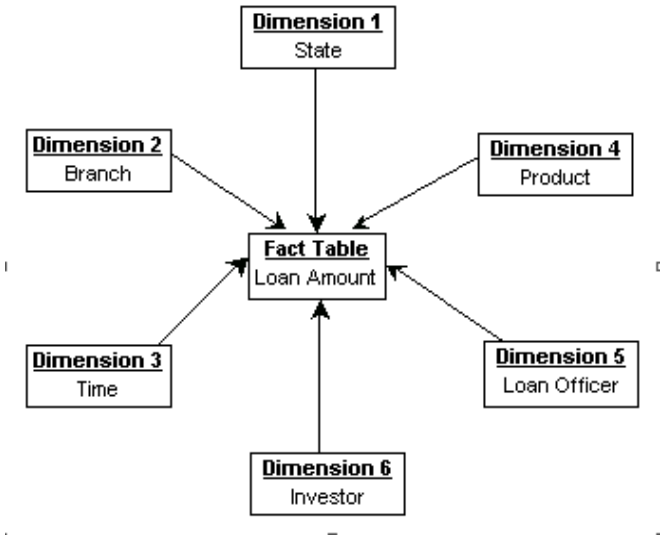

Figure 1: A Star Schema[6]

3.2 Snowflake Schema

The snowflake schema[11,9] is an extension of the star schema, where each point of the star explodes into more points. In a star schema, each dimension is being represented by a single dimensional table, whereas in a snowflake schema, that dimensional table is further normalized into multiple fact tables, each representing a level in the dimensional hierarchy.

The architecture of snowflake schema 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 modelers to understand and for database designers to use for the analysis of dimensions.

The main advantage of the snowflake schema[4] is the improvement in query performance due to minimized disk storage requirements as it helps in getting the result by joining smaller dimension tables. The main disadvantage of the snowflake schema is the additional maintenance cost needed to maintain large number of dimension tables. The diagram of snowflake schema is given below

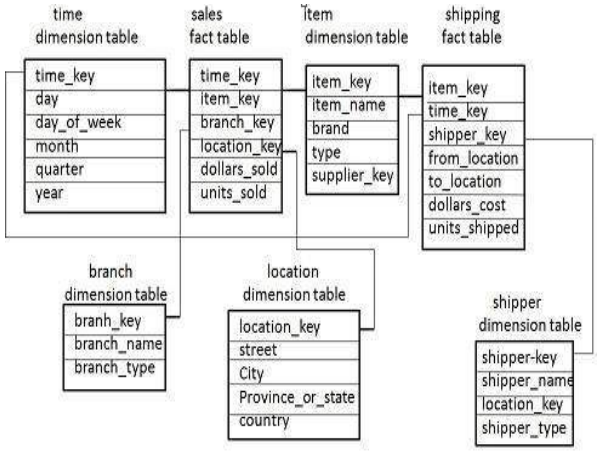

Figure 2: A Snow Flake Schema[4]

3.3 Fact Constellation Schema

The fact constellation schema is quite different from star schema or snowflake schema.This kind of schema can be viewed as a collection of different stars and hence that is why it is called Fact Constellation.However this schema is more complex than star or snowflake architecture, which is because it contains multiple fact tables.

In fact constellation schema[4,11], the different fact tables are explicitly assigned to the dimensions, which are for used for elaborating the desired facts for a given dimension.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.Basically in fact constellation schema we maintain more than one fact table thereby more joins are needed to execute the query in a given OLAP system.

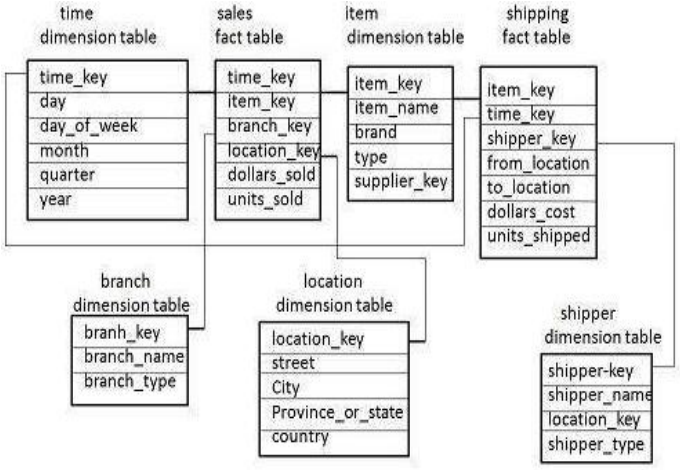

Figure 3: Fact Constellation Schema

4. MULTIDIMENSIONAL MODEL Vs RELATIONSHIP MODEL

We next present the realationship between multi dimensional model and ER model.The ER diagram is a logical design technique that seeks to remove the redundancy in data. This coupled with normalization of data enables with 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.

Basically relatinal model maintain current data but data warehouse always contain historical data.In future we can explore the different research area in data awarehouse such as :OLAP Engine,Data marts,Integration of data warehouse with data mining.

5. Conclusion

In this paper we have discuss data warehouse, need for using the data warehouse. We have also covered different design issues in data warehouse. The schemas used in data warehouse is discussed in detail 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 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. In order to make strategic decision making a separate store known as data warehouse was required for further carrying out data mining task.

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.Works as expected

[14] Swati Gupta "A Review on Data Mining Techniques" in International Journal of Software and Web Sciences (IJSWS) ISSN (Print): 2279-0063,May 2017