Architecture of Data Warehouse: A Comprehensive Study

Rashmi Bhatia

***Abstract*— The Data Warehousing in simple terms is to create a central location and enduring storage space for the numerous data sources required to support a company's analysis, reporting and other Business Intelligence functions. It provides the Business with deep insight into the existing data, which helps them in taking decisions more efficiently. Dependence of Business on Data Warehouse is tremendously increasing. As the decisions are based on the information lying in its databases, competitive advantage is gained using this technology. Thus importance of Data Warehouse can't be denied. In this paper the data warehouse is discussed in detail with thorough study of its architecture, from the acquisition of the data to its detailed design, its storage and access, and metadata management components.**

***Index Terms*— Data Cubes, Data Marts, Meta Data, Roll-Up Display.**

I. INTRODUCTION

Data Warehouse is "a subject-oriented, integrated, non-volatile, time-variant collection of data in support of management's decision" (Inmon, W.H.,1992, in Elmasri and Navathe,2000,p. 842). The data is extracted from heterogeneous operational systems and external data sources, cleansed in order to ensure validity, transformed so as to remove inconsistencies and homogenize the data, aggregated, and in the last loaded into data warehouse. The data once inserted in data warehouse cannot be changed until modifications of the source data are propagated into the warehouse, but can be deleted.

In other words, data warehousing comprises a set of decision support technologies, which support the knowledge worker (executive, analyst, manager etc.) with adequate and high quality information, so that they can make better and faster decisions.

II. FEATURES OF DATA IN DATA WAREHOUSE

The definition given specifies few of the features of data warehouse, which are as follows: -

A. Subject-Oriented

Data in data warehouse is subject-oriented, which means that the data queried is related with some specific subject area (e.g., products, customers, regions etc.)

B. Integrated

Manuscript received Feb. 20, 2014.

The data is integrated from several, possibly heterogeneous operational systems such as database systems, flat files, etc. and also from various external data sources like World Wide Web, statistical databases etc., in the data warehouse. The data from various sources is homogenized before the integration takes place i.e. theformat inconsistencies are removed. The data collected from various sources can be incomplete or erroneous. So in order to ensure validity the data is cleaned up. Then the data is installed in the data model of the warehouse.

C. Non-Volatile

This feature of data warehouse states that warehouse data is mostly non-volatile, which implies that the data is read-only. "The term non-volatile means that, once inserted, data cannot be changed, though it might be deleted."(Date, C.J. 2000). Any changes in the data take place only when modifications of the source data are propagated into the warehouse.

D. Time Variant

The time variant feature indicates the need to access historical data, which is one of the reasons for adopting data warehouse. As decision-making requires various Business Trend Analysis, for which historical data is required. Data warehouse keeps the periodical snapshots of the corresponding operational data, which is necessary in various analysis with respect to the time.

E. Different from OLTP Databases

The traditional Online Transaction Processing (OLTP) systems are not a right choice to provide support in decision-making. Data warehouse supports On-line Analytical Processing (OLAP). In OLTP even if the high speed networks are established but still the information accessibility problems persist because of the following reasons: -

- OLTP database maintain current data in great detail. Each transaction requires detailed, up-to-date data, where that part of the database is updated, which is accessed, immediately after any operation completed on the database. But in OLAP, rather than detailed data, the historical, summarized and consolidated data is of more importance because more stress is on decision making.
- Data in OLTP can be hundreds of megabytes to gigabytes in size. Whereas the size of data in data warehouse can be much larger than operational databases. It can vary from hundreds of gigabytes to terabytes in size.
- The performance key in OLTP is the maximization of the transaction throughput whereas in OLAP Query throughput and response times are more important than transaction throughput.

F. Basis for Management's Decision

The data warehouse aims at providing the management with the information in the format required by them, so as to affect their decision capability. It emphasizes on presenting the right information, in the right place at the right time, in the right format with the right cost, so as to support the right decision.

III. Architecture of Data Warehouse

The architecture of data warehouse can be divided into three components, where the first component is the Data Import and Preparation (acquisition) Component, the next is Design and Storage and the third is the Access Component which includes the applications so as to make use of the information stored in data warehouse in OLAP and data mining applications. There is one more component i.e. Meta Data Management Component which manages, defines and accesses the various types of metadata.

The overall Data Warehouse System (DWS) is implemented in two phases, Configuration Phase and Operation Phase.

- In Configuration Phase, according to the user's requirements, a conceptual view of the data warehouse is developed. The decisions regarding the various data sources from which the data will be imported and the way data will be loaded in data warehouse is determined. How the storage takes place and the data will be accessed is also decided.
- After the first load of data warehouse in configuration phase, the Operation Phase takes place, in which the warehouse data is regularly refreshed so that the modifications of the source data must be propagated into the warehouse.

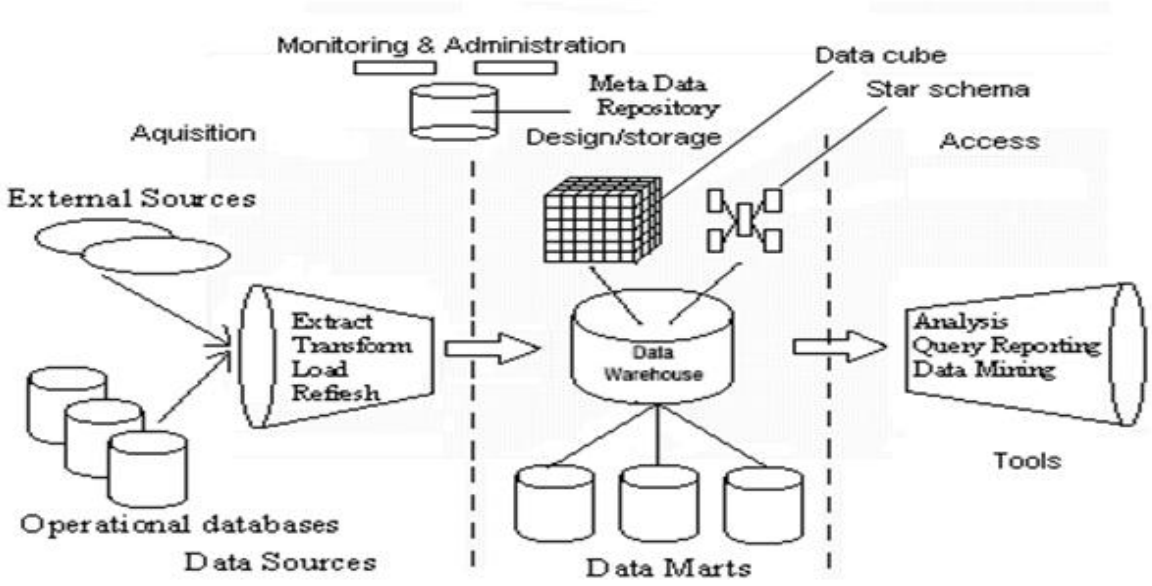

Figure 1. Architecture of Data Warehouse

A. ACQUISION OF DATA

It includes the various tools for extracting data from various tools for extracting data from various external and internal sources, cleaning, transforming and integrating this data, so as to load the data into data warehouse and also periodical refreshing of the data.

- **Extraction: -** Data is extracted from multiple operational databases and external sources which may include files acquired from independent systems and platforms.
- **Cleaning and Transforming: -** The data is collected from heterogeneous sources. That's why data values can be incorrect, inconsistent, unreadable or incomplete. This means that there can be inconsistent field lengths, inconsistent descriptions, inconsistent value assignments, missing entries and violation of integrity constraints. But as data warehousing is used to support decision making, one needs the correct and high quality data. So some data cleaning tools are used to detect the data anomalies and finally correct them before loading the data into the warehouse. Therefore these tools are as follows:
  a) **Data Migration Tools** in which simple transformation rules are specified. E.g. the string "address" can be replaced by "location".
  b) **Data Scrubbing Tools** in which the scrubbing of data is done by making use of the domain-specific knowledge. These tools make use of parsing and fuzzy matching techniques for cleaning data from multiple sources.
  c) **Data Auditing Tools** discover rules and relationships by scanning the data.

- **Load: -** After extracting, cleaning and transforming the homogenized data is required to be loaded into the data warehouse. Before or during loading data into the data warehouse, some additional tasks like
  a) Checking integrity constraints
  b) Sorting
  c) Summarization, aggregation and other computations, in order to build the derived tables stored in the warehouse
  d) Indexing

e) And portioning to multiple target storage areas are often required.

Batch Load Utilities are used for this purpose. But the process of loading takes a lot of time because the data quantity is large. Sequential loads take longer time than Pipelined and Partitioned Parallelism. Provisions are made to monitor status, to cancel, suspend and resume a load, and also to restart with no loss of data integrity, if any failure occurs. Also periodic checkpoints can be added, so that if any failure occurs during load process, the process can restart from the immediate last checkpoint, which saves time as well.

- **Refresh: -** Periodic refreshing of data is required n order to keep the database reasonably current. If any update into the source data takes place then these updates must be propagated into the data warehouse so as to update the base data and derived data stored in data warehouse. The following are some questions, the answers to which decide a good Refresh Policy. These questions are: -
  a) How up-to-date the data should be? So as to decide after how much time the data must be refreshed (e.g. daily, weekly etc.).
  b) Whether a warehouse can go off-line and for how much time. If yes then refreshing can take place at the time the warehouse goes offline.
  c) What are the various data interdependencies?
  d) How much storage is available?
  e) How much time it took to load the data completely including other processes such as sorting, indexing, partitioning etc.

Chaudhuri, S. and Dayal, U. (p.4) states that most contemporary database systems provide replication servers that support incremental techniques for propagating updates from a primary databases to one or more replicas.
The two basic replication techniques are Data Shipping and Transaction Shipping.

B. DATA WAREHOUSE DESIGN

Designing a data warehouse is very difficult as compared to designing traditional operational systems. It involves thinking in terms of much broader, and more difficult to define, business concept than designing an operational system.

Multidimensional Data and Data Cubes: - In data warehouse designing, multidimensional models are used, which make use of the inherent relationships in data to populate data in data cubes (multidimensional matrices). Let us take an example of three dimensional data cube with Product, Region and Date as each of its dimension

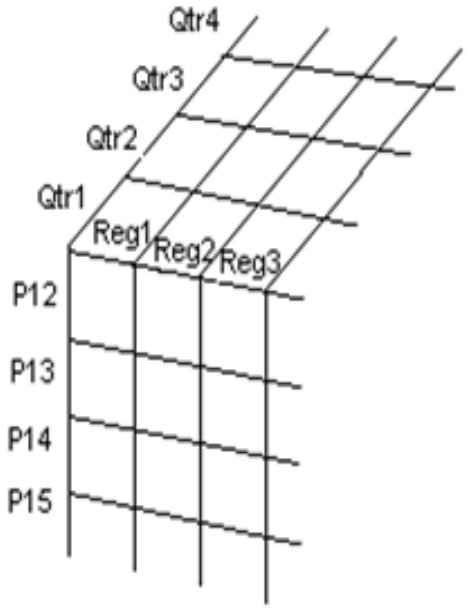

Figure 2. A Data Cube

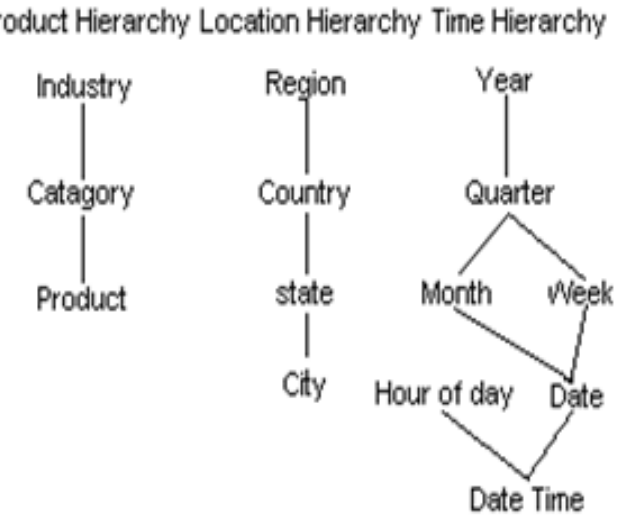

Figure 3. Hierarchical Summarization Paths

Each cell of the data cube contain data for a specific product, specific region and on specific date. If more than three dimensions are there then Data Hypercube can be produced. Each of the dimensions is described by a set of attributes and these attributes are related through a hierarchy of relationships. The product dimension may consist of three attributes: the category, industry, and year of its introduction. E.g., the colacola belongs to beverage category and food industry, year of introduction is 1990. Two popular multidimensional schemas are as follows: -

a) **Star Schema: -** In this schema there is one fact table and a single table for each dimension. All the attributes of a dimension become the columns of that particular dimension table. Each tuple in the fact table references multiple dimensional table tuples each one representing a dimension of interest like products, customers, time etc. Since dimension tables are not normalized, joining fact table with the dimension tables various views (dimensions) of the data warehouse data can be generated in an efficient way.

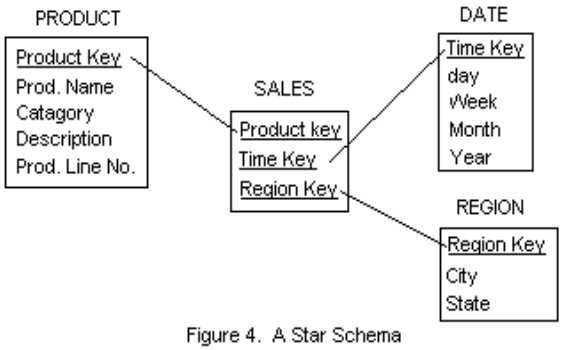

Figure 4. A Star Schema

a) **Snowflake Schema: -** Snowflake schema is a variation on the star schema, where the dimensional tables from star schema are organized into a hierarchy by normalizing them. It helps in maintaining the dimension tables, but the demoralized structure in star schema is better for browsing the dimensions.

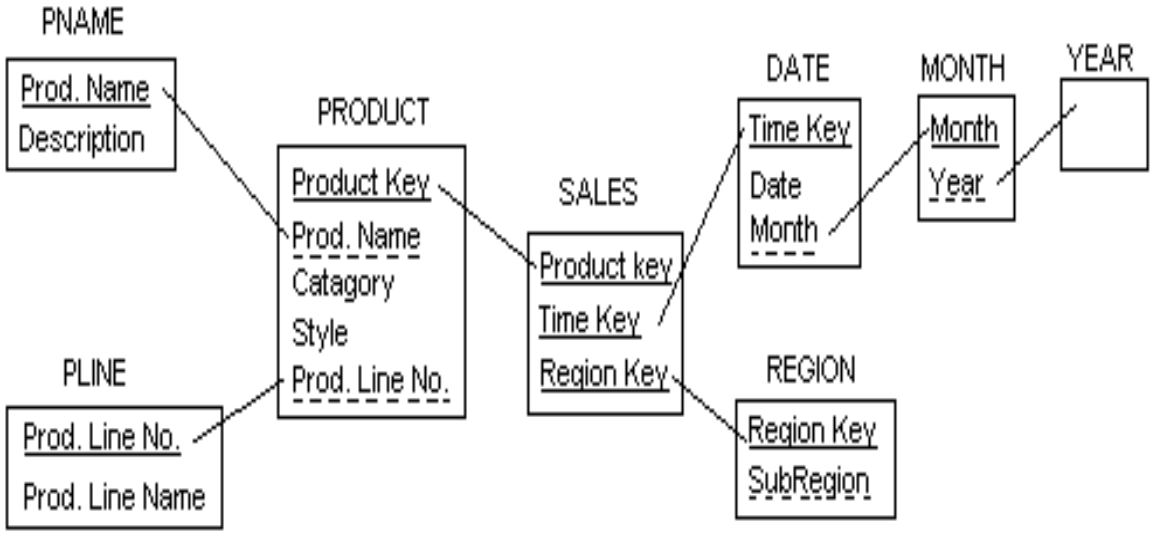

Figure 5. A SnowFlake Schema

C. DATA STORAGE AND ACCESS

The special nature of warehouse data requires adjusted mechanisms for data storage, query processing and transaction management. Query optimization is one of the main objective, where OLAP user poses many queries and expect short response time. In order to reduce the access time Bitmap Indices and various forms of join indices can be used. The warehouse data is non-volatile, that's why complex concurrency control techniques and transaction management can be adapted.

The data can be queried directly in any combination of dimensions and there are various data according to the user's choice of dimensions. These are: -

- **Pivoting: -** Pivoting means rotating the cube. In this technique the data cube can be rotated to show different orientation of the axes, thus cross tabulation is possible. E.g., one can pivot the data cube to show regional sales revenue as rows, the data dimension as column and the product as the third dimension.

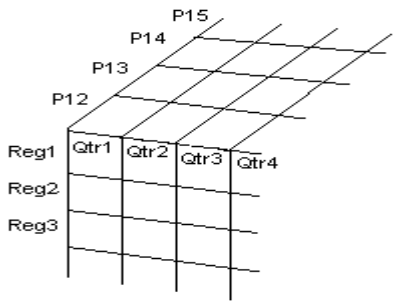

Figure 6. The Pivot Operation

- **Roll-Up Display: -** Roll up display takes the current data object and perform group-by on one of the dimensions. In other words data is summarized with increasing generalization (e.g., weekly to quarterly to annually).

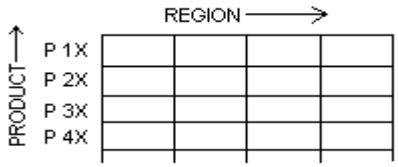

Figure 7. The Roll-Up Operation

- **Drill-Down Display: -** It is just opposite to Roll-Up display where increasing levels of details are revealed. The figure above displays the disaggregating of country sales by regions and then regional sales by sub region and also breaking up products by styles.

Figure 8. The Drill-Down Operation

Slicing-Dicing: -Which implies selecting a subset of cube by performing projection operation on dimensions.

In order to reduce the response time **Data Marts** are used. Data Mart is "a specialized, subject-oriented, integrated, volatile, time-variant data store in support of a specific subset of management's decisions." (Date, C.J. 2000, p.710). In order to achieve better performance and scalability, there can be several departmental data marts other than the main data warehouse. Warehouse server(s) manage the data stored in data mart and data warehouse.

D. D. META DATA MANAGEMENT COMPONENT

Meta data can be defined as "data about data" Gatzice, S., Vavouras, A. (1999, p.2) states that, In data warehousing, there are various types of metadata. E.g., information about the operational sources the structure and semantics of the data warehouse data, the task performed during the construction, the maintenance and access of a data warehouse etc.

A high quality Meta data results in high quality information received from data warehouse.

REFERENCES

[1] W.H. Inmon, "Building the Data Warehouse". New York: John Wiley. in *Fundamentals of Database Systems.* 3rd ed., ELMASRI, R. and NAVATHE, S.B. , Singapore: Pearson Education, 2000

[2] C.J. Date, *An Introduction to Database Systems*. 7th ed., Singapore: Pearson Education, 2000

[3] B. Mento and B. Rapple, *Data Mining and Data Warehousing* [Online]. Washington,D.C: Association of Research Libraries, 2003. Available: http://babel.hathitrust.org/cgi/pt?id=mdp.39015052882167;view=1up;seq=7

[4] E. Rich and K. Knight, *Artificial Intelligence*. 2nd ed., New Delhi: Tata McGraw-Hill Publishing Company Limited, 1991

[5] A. Silberschatz, H.F. Korth and S. Sudarshan, *Database Systems Concepts.* 4th ed., New York: McGraw-Hill., 2002

[6] A. Vavouras and S. Gatziu, *Data Warehousing: Concepts and Mechanisms.* [Online]. University of Zurich: Computer Science Department. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.1043&rep=rep1&type=pdf

[7] S. Chaudhuri and U. Dayal, *An Overview of Data Warehousing and OLAP Technologies.* [Online]. Available: http://research.microsoft.com/pubs/76058/sigrecord.pdf

[8] *Data Mining: What is Data Mining* [Online]. Available: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[9] D. Alexander, *Data Mining* [Online]., Available: http://www.eco.utexas.edu/~norman/BUS.FOR/course.mat/Alex/

[10] A. Perkins, *Developing a Data Warehouse: The Enterprise Engineering Approach.* [Online]. Available: http://www.ies.aust.com/PDF-papers/dw.pdf

111] Statistica*: Data Warehouse* [Online]. Available: http://www.statsoft.com/Portals/0/Support/Download/Brochures/Data_Warehouse.pdf