Data Warehousing

Neha Pandit, Priyanka Tiwari
Students
Dronacharya College of Engineering, Greater Noida

Abstract: **A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process. Data Warehouse provides an effective way for the analysis and statistic to the mass data and helps to do the decision making. Many commercial products and services are now available and all of the principal database management system vendors now have offerings in these areas. The paper introduces the data warehouse and the online analysis process with an accent on their new requirements.**

Index Terms**: Data Warehouse, Subject Oriented, Integrated, Time Variant, Non Volatile.**

I. Introduction:

Data Warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions. It provides architecture and tools for business executives to systematically organize, understand and use their data to make strategic decisions. Data Warehouse is a database used for reporting and analysis. It refers to the database that is maintained separately from an organization's operational databases. The data stored in the data warehouse is uploaded from the operational systems. Data Warehouse systems allow for the integration of a variety of application systems. They support information processing by providing a solid platform of consolidated historical data for analysis. Data warehousing technologies have been successfully deployed in many industries: manufacturing (for order shipment and customer support), retail (for inventory management), financial services (for credit card analysis, risk analysis, and fraud detection), utilities (for power usage analysis), and healthcare (for outcomes analysis). This paper presents an overview of data warehousing technologies, focusing on the special requirements that data warehouses place on Database Management Systems (The business organizations and companies usually need to store huge amount of data both historical and current for which they need a well-organized database that can not only facilitate the transaction processing but also a help in decision support and evaluating future strategy. A data warehouse is such a collection of integrated databases that has met all the demands of different business organizations. It is a subject oriented database since data stored within a data warehouse gives information about a particular subject rather than the companies' ongoing process. Data warehouses are also time stamped in nature that means all data stored within the data warehouse is identified with a particular time period. Data once stored in the data warehouse is never removed, only transaction and addition of data is possible and this enables the management to get a consistent picture of the business. Data warehouses are prevalently characterized by OLAP workload and it provides the users with data access tools and applications that are appropriate for their needs.

Subject-oriented: A data warehouse is organized around major subjects such as customer, supplier, product and sales. Rather than concentrating on the day to day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers. Data warehouses typically provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.
Integrated: A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on line transaction records. Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures and so on.
Time-variant: Data are stored to provide information from a historical perspective. Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.
Nonvolatile: A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrence control mechanisms. It usually requires only two operations in data accessing: initial loading of data and access of data.

II. Types of systems

Data mart: A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), such as sales, finance or marketing. Data marts are often built and controlled by a single department within an organization. Given their single-subject focus, data marts usually draw data from only a few sources. The sources could be internal operational systems, a central data warehouse, or external data.

Online analytical processing (OLAP) is characterized by a relatively low volume of transactions. Queries are often very complex and involve aggregations. For OLAP systems, response time is an effectiveness measure. OLAP applications are widely used by Data Mining techniques. OLAP databases store aggregated historical data in multi-dimensional schemas (usually star schemas).

OLAP systems typically have data latency of a few hours, as opposed to data marts, where latency is expected to be closer to one day.

Online Transaction Processing (OLTP) is characterized by a large number of short on-line transactions (INSERT, UPDATE, and DELETE). OLTP systems emphasize very fast query processing and maintaining data integrity in multi-access environments. For OLTP systems, effectiveness is measured by the number of transactions per second. OLTP databases contain detailed and current data. The schema used to store transactional databases is the entity model (usually 3NF).

Predictive analysis: Predictive analysis is about finding and quantifying hidden patterns in the data using complex mathematical models that can be used to predict future outcomes. Predictive analysis is different from OLAP in that OLAP focuses on historical data analysis and is reactive in nature, while predictive analysis focuses on the future. These systems are also used for CRM (Customer Relationship Management).

III. Data Warehouse Architecture

Different data warehousing systems have different structures. Some may have an ODS (operational data store), while some may have multiple data marts. Some may have a small number of data sources, while some may have dozens of data sources. In view of this, it is far more reasonable to present the different layers of data warehouse architecture rather than discussing the specifics of any one system.

In general, all data warehouse systems have the following layers:

- Data Source Layer
- Data Extraction Layer
- Staging Area
- ETL Layer
- Data Storage Layer
- Data Logic Layer
- Data Presentation Layer
- Metadata Layer
- System Operations Layer

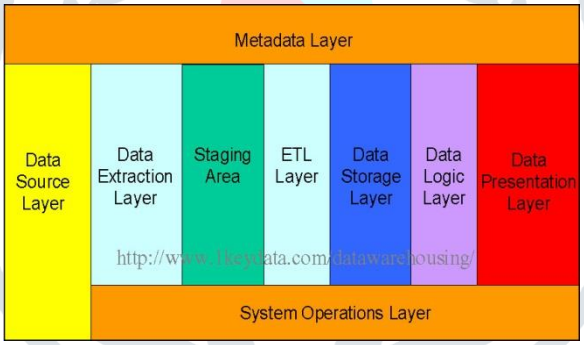

Figure 1: Relationships among the different components of the Data Warehouse Architecture

Data Source Layer

This represents the different data sources that feed data into the data warehouse. The data source can be of any format -- plain text file, relational database, other types of database, Excel file, etc., can all act as a data source. Many different types of data can be a data source:

- Operations -- such as sales data, HR data, product data, inventory data, marketing data, systems data.
- Web server logs with user browsing data.
- Internal market research data.
- Third-party data, such as census data, demographics data, or survey data.

All these data sources together form the Data Source Layer.

Data Extraction Layer

Data gets pulled from the data source into the data warehouse system. There is likely some minimal data cleansing, but there is unlikely any major data transformation.

Staging Area

This is where data sits prior to being scrubbed and transformed into a data warehouse / data mart. Having one common area makes it easier for subsequent data processing / integration.

ETL Layer

This is where data gains its "intelligence", as logic is applied to transform the data from a transactional nature to an analytical nature. This layer is also where data cleansing happens. The ETL design phase is often the most time-consuming phase in a data warehousing project, and an ETL tool is often used in this layer.

Data Storage Layer
This is where the transformed and cleansed data sit. Based on scope and functionality, 3 types of entities can be found here: data warehouse, data mart, and operational data store (ODS). In any given system, you may have just one of the three, two of the three, or all three types.

Data Logic Layer
This is where business rules are stored. Business rules stored here do not affect the underlying data transformation rules, but do affect what the report looks like.

Data Presentation Layer
This refers to the information that reaches the users. This can be in a form of a tabular / graphical report in a browser, an emailed report that gets automatically generated and sent every day, or an alert that warns users of exceptions, among others. Usually a tool and/or a reporting tool are used in this layer.

Metadata Layer
This is where information about the data stored in the data warehouse system is stored. A logical data model would be an example of something that's in the metadata layer. A metadata is often used to manage metadata.

System Operations Layer
This layer includes information on how the data warehouse system operates, such as ETL job status, system performance, and user access history.

IV. Back-End Tools and Utilities

Data warehouse systems use back-end tools and utilities to populate and refresh their data.

Data Cleaning:
Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data Since a data warehouse is used for decision making, it is important that the data in the warehouse must be correct. Some examples where data cleaning becomes necessary are: inconsistent field length, inconsistent descriptions, inconsistent value assignments, missing entries and violation of integrity constraints.

Load
After extracting, cleaning and transforming, data must be loaded into the warehouse. Additional preprocessing may still be required: checking integrity constraints; sorting; summarization; aggregation; and other computations to build the derived tables stored in the warehouse. In addition, load utility also allows the system administrator to monitor status, to cancel, to suspend and resume a load, and to restart after failure with no loss of data.

Integrity.
The load utilities for data warehouses have to deal with much larger data volumes than for operational databases.

Refresh
Refreshing a warehouse consists in propagating updates on source data to correspondingly update the base data and derived data stored in the warehouse. There are two sets of issues to consider: when to refresh and how to refresh. Usually, the warehouse is refreshed periodically. The refresh policy is set by the warehouse administrator, depending on user needs and traffic, and may be different for different sources. Refresh techniques also depends on the characteristics of the source and capabilities of the database servers. Replication servers can be used to refresh a warehouse when the sources change.

V. Multidimensional Data Model

Data warehouse and OLAP tools are based on a multidimensional data model. This model views data in the form of a data cube. A data cube allows data to be modeled and viewed in multiple dimensions. Dimensions are perspectives or entities with respect to high an organization wants to keep records. Each dimension has a table associated with it, called a dimension table, which further describes the dimension. A multidimensional data model is typically organized around a central theme. This theme is represented by a fact table. The fact table contains the names of the facts, or measures, as well as keys to each of the related dimension tables.

VI. Front-End Tools

The multidimensional data model grew out of the view of business data popularized by spreadsheet programs that are extensively used by business analysts. One of the popular operations that are supported by the multidimensional spreadsheet is pivoting. Pivot also called rotate, is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data. Other operations are roll-up, drill-down, slice and dice. The roll-up operation performs the aggregation on a data cube, either by climbing up the concept hierarchy for a dimension or by dimension reduction. Drill-down is the reverse of the roll-up. It

navigates from less detailed data to more detailed data. The slice operation performs a selection on one dimension of the cube. The dice operation performs a selection on two or more dimensions.

VII. Database Design

Most data warehouse use a **star schema** to represent the multidimensional data model. The database consists of a single fact table and a single table for each dimension. Each tuple in the fact table consists of a pointer to each of the dimension that provides its multidimensional coordinates and stores the numeric measures for that coordinates. Each dimension table consists of columns that correspond to attributes of the dimension.

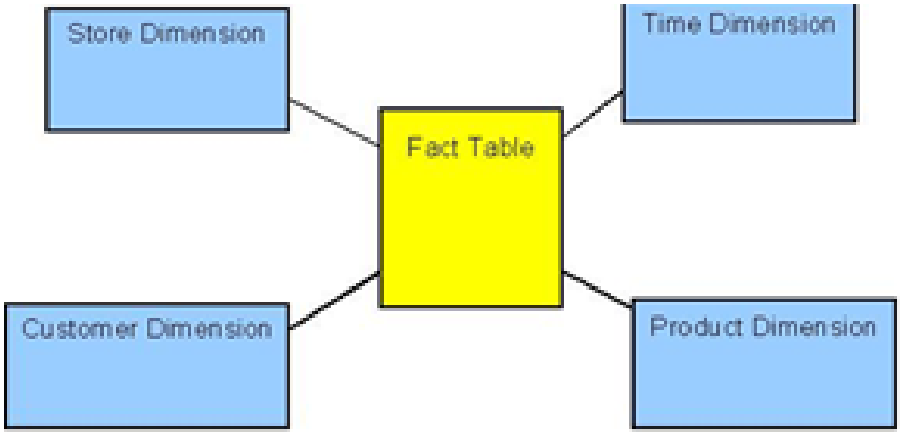

Figure 2: Star Schema

The **Snowflake schema** is the variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms the shape similar to a snowflake. The major difference between the snowflake and star schema models is that the dimension table of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space.

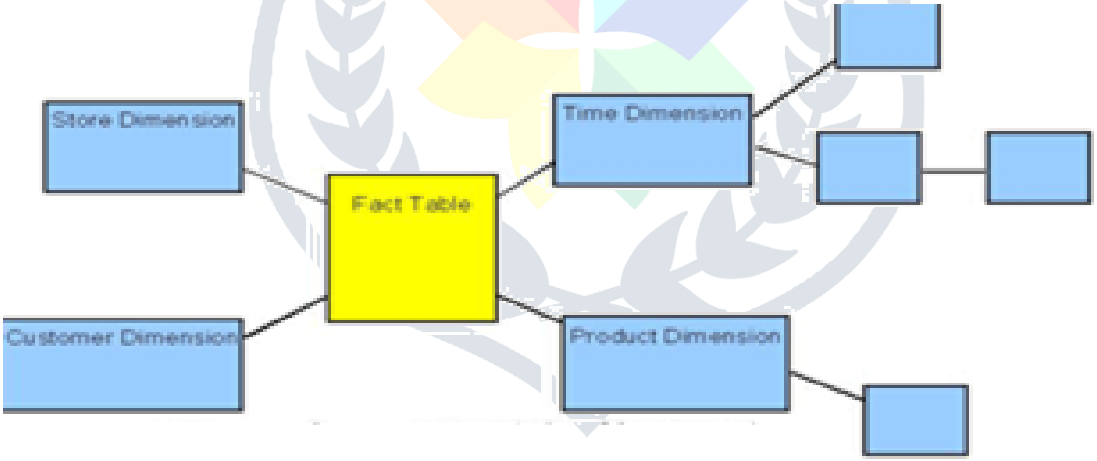

Figure 3: Snowflake Schema

VIII. CONCLUSION:

In this paper, I have discussed about the construction of data warehouses, designing of data warehouses. The construction of data warehouses involves data cleaning and data integration.Datacleaning attempt to fill in the missing values, Smooth out the noise, while identifying the outliers and remove the inconsistencies in the data. Many people feel that with competition mounting in every industry, data warehousing is the latest must-have marketing weapon-a way to keep customers by learning more about their needs.

A data warehouse is a subject oriented, integrated, time variant and non-volatile collection of data in support of management's decision making process. Data warehousing is the process of constructing and using data warehousing is very useful from the point of view of heterogeneous database integration. It provides an interesting alternative approach to the traditional approach of heterogeneous database integration. It employs an update-driven approach in which information from multiple, heterogeneous source is integrated in advance and stored in a warehouse for direct querying and analysis. Data warehouse do not contain the current information. However, data warehouse brings high performance to the integrated heterogeneous database system. It can store and integrate historical information and support complex multidimensional queries. As a result, data warehousing has become very popular in industry.

IX.ACKNOWLEDGEMENT:

This paper would not have been possible without the guidance of teachers and seniors who have encouraged us for writing the paper. We would also like to thank our parents and our friends for their unconditional support.

References:

[1] Data Warehousing- Wikipedia
[2] Jiawei Han and Micheline Kamber : Data Mining Concepts and Techniques
[3] Surajit Choudhary: Data Warehousing and OLAP technology.
[4] Umeshwar Dayal: An overview of data warehousing and technology.
[5] http://www.dwinfocenter.org
[6] http://www.carolla.com/wp-dw.htm
[7] http://system-services.com/dwintro.asp
[8] http://ciemcal.org/data-warehousing/