International Refereed Journal of Reviews and Research

International Refereed and Indexed Journal for Research Scholars and Practitioners

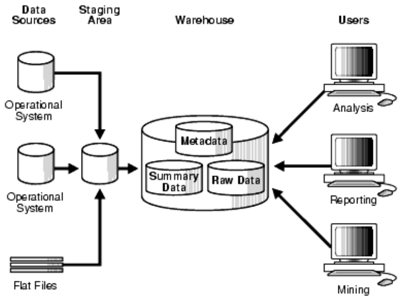
DATA WAREHOUSING, BUSINESS INTELLIGENCE AND MACHINE LEARNING

M. Srivastava

Freelance Writer

Mumbai, India

Data warehouse is a repository or storage house of an organization's electronic data. Data warehouses are used to facilitate reporting and analysis. As compared to data warehouses, operational databases support day-to-day transaction processing.



ADVANTAGES OF DATA WAREHOUSING

- It provides a common data model for all data of interest regardless of the data's source.
- Development of report and analysis of information is easy than using multiple data models to retrieve information such as sales invoices, order receipts, ledger charges.
- In data warehouse, inconsistencies can be easily identified and resolved.
- Simplification in reporting and analysis.
- Information is under the control of data warehouse administrator so that the information in the warehouse can be stored safely for extended span of time.
- Data Warehouses provides retrieval of data and reporting without slowing down the operational systems.
- Data warehouses also work with decision support system applications such as market reports, exception reports, and reports that show actual performance versus goals.

DISADVANTAGES OF DATA WAREHOUSING

- Not suitable and efficient for unstructured data.
- Maintaining the data warehouses are generally very costly.
- The data warehouse is usually not static in nature.
- Maintenance and upgradation costs are high.

APPLICATIONS OF DATA WAREHOUSING

- Credit card analysis
- Insurance Fraud analysis
- Internet Fraud Analysis
- Cyber Forensic
- E-mail Sender Investigation
- Call Tracing Analysis
- Call record analysis
- Logistics management

BUSINESS INTELLIGENCE

Business intelligence is closely related to data warehousing. This section discusses business intelligence, as well as the relationship between business intelligence and data warehousing.

FACT TABLE

Fact table is a relation of database which contains the measures of interest. For example, profit amount would be such a measure. This fact or measure is stored in the fact table with the appropriate granularity. In this example, the fact table will contain three columns: date, storename, and a profit amount.

LOOKUP TABLE

Lookup table gives the detailed information about the attributes. For example, the lookup table for the Month attribute would include a list of all of the Months available in the data warehouse. Each row (each quarter) may have several fields, one for the unique ID that identifies the quarter, and one or more additional fields that specifies how that particular quarter is represented on a report (for example, first quarter of 2001 may be represented as "M1 2001" or "2001 M1").

COMPONENTS OF DATA WAREHOUSE

The data residing in the data warehouse arrives from operational systems of the organization as well as from other external sources. It is collectively known as ***Source Systems***. The data which is ***Extracted*** from source systems is stored in a area called ***Data Staging Area***, where the data is cleaned, ***Transformed***, combined, unduplicated to prepare the data for us in the data warehouse.

The data staging area is a collection of machines where simple activities including sorting and sequential processing. The data staging area do not provide any query or presentation services. When a system provides query or presentation services, it is categorized as a ***Presentation Server***.

A presentation server is the key machine on which the data is ***Loaded*** from the data staging area organized. It is stored for direct querying by end users, report writers and other applications.

Three different kinds of systems that are required for a data warehouse are:

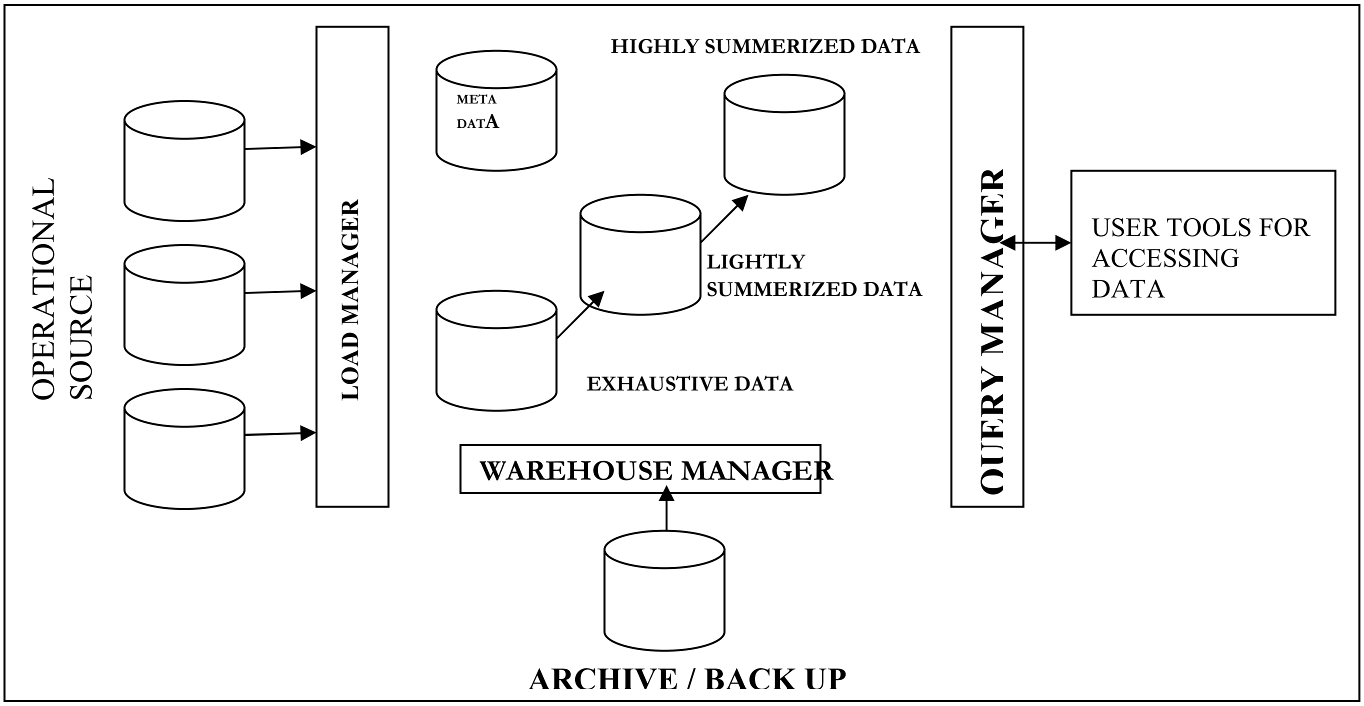
1. Source Systems
2. Data Staging Area
3. Presentation servers

The data gets transmitted from source systems to presentation servers via the data staging area.

This entire process is popularly known as ETL (**Extract, Transform, And Load**) or ETT (**Extract, Transform, And Transfer**).

The ETL tool of **Oracle** is called **Oracle Warehouse Builder (OWB)** and **MS SQL Server's ETL** tool is called **Data Transformation Services (DTS).**

ARCHITECTURE OF A DATA WAREHOUSE



OPERATIONAL DATA

The key sources of data for the data warehouse are supplied from the data from the main systems in the traditional network and hierarchical format. The data can also arrive from the relational DBMS like Oracle, Informix, MySQL. Moreover, operational data also includes external data obtained from commercial databases and databases associated with supplier and customers.

LOAD MANAGER

This component is responsible for performing all the operations associated with extraction and loading data into the data warehouse. Such operations include simple transformations of the data to the format of warehouse. The size and complexity factor of this component may vary between data warehouses and constructed using a combination of vendor data loading tools and custom built programs.

WAREHOUSE MANAGER

This component performs all the operations related with the management of data in the warehouse. It is built using vendor data management tools and custom built programs.

The operations performed by warehouse manager are

(i) Analysis of consistency

(ii) Transformation and Joining the source data from temporary storage into data warehouse tables

(iii) Creating indexes and views on the base table.

(iv) Denormalization

(v) Production of aggregation

(vi) Backing up, Storage and Archiving of data

QUERY MANAGER

Query manager performs all the functions associated with management of user queries. It is usually constructed using vendor end-user access tools, data warehousing monitoring tools, database facilities and custom built programs. The complexity of this component is determined by facilities provided by the end-user access tools and database.

DETAILED DATA

It is the area which stores all the detailed data in the database schema. Generally, detailed data is not stored online but aggregated to the next level of details. Still, the detailed data is updated and inserted regularly to the warehouse to supplement the aggregated data.

LIGHTLY AND HIGHLY SUMMERIZED DATA

It is the area which stores all the predefined lightly and highly summarized (aggregated) data produced by the warehouse manager. This area of the warehouse is temporary as it will be subject to changes on an ongoing basis to respond to the changing query profiles. The main purpose of the summarized information is to speed up the query performance. The summarized data is updated continuously as soon as new data is loaded into the warehouse.

ARCHIVE AND BACK UP DATA

This area stores detailed and summarized data for the purpose of archiving and backup. The data is sent to storage archives such as magnetic tapes or optical disks.

META DATA

All the Meta data or data about data definitions are stored by the warehouse.

It is used for

(i) The extraction and loading process – Meta data maps data sources to a common view of information within the warehouse.

(ii) The warehouse management process – Meta data automate the production of summary tables.

(iii) Meta data directs a query to the most appropriate data source.

END-USER ACCESS TOOLS

The main purpose of data warehouse is to give information to the business managers for strategic decision-making. The users interact with the warehouse using end user access tools.

Examples of some of the end user access tools are

(iv) Reporting and Query Tools

(v) Application Development Tools

(vi) Executive Information Systems Tools

(vii) Online Analytical Processing Tools

(viii) Data Mining Tools

THE E T L (EXTRACT TRANSFORMATION LOAD) PROCESS

These are used to extract (data from the operational systems and pull it to the data warehouse), transform (the data into internal format and structure of the data warehouse), cleanse (to make sure it is of sufficient quality to be used for decision making) and load (cleanse data is put into the data warehouse).

There are four processes from extraction through loading and referred collectively as Data Staging.

EXTRACT

Some data elements in the operational database may reasonably be useful in the decision making, but others may be of less value for that purpose. For this purpose, it is necessary to extract the relevant data from the operational database before pushing into the data warehouse.

Many commercial tools are available which helps in the extraction process. Data Junction is one of the available commercial products. The user of these tools has an easy-to-use windowed interface by which to specify the following:

(i) Which files and tables should be accessed in the source database?

(ii) Which fields should be extracted from them?

(iii) What are required in the resulting database?

(iv) What is the database format of the output?

(v) What is the schedule for the extraction process?

TRANSFORM

The operational databases can be stored on any set of priorities, which changes time to time with the requirements. Therefore the developers of the data warehouse face different inconsistency problems among their data sources. Transformation process deals with removing any inconsistency.

CLEANSING

Information quality is the main consideration in determining the value of the information. It is necessary to go through the data entered into the data warehouse and make it as error free. This process is known as Data Cleansing. Data Cleansing deal with many types of errors. These may include missing data and incorrect data at one source; inconsistent data and conflicting data when two or more sources are involved. Several algorithms are followed to clean the data

LOADING

Loading means the movement of the data from the source database(s) to that which will store the data warehouse database. This takes place after the extraction phase. The very common channel for data movement is a high-speed communication link. Its examples are Oracle Warehouse Builder is the API from Oracle, that provides the features to perform the ETL task on Oracle Data Warehouse.