SELECTED TOPICS OF RECENT TRENDS IN INFORMATION TECHNOLOGY

UNIT I

Database and Data Ware Housing

DWH Constitute Entire Information Base For All Time.

- Database Constitute Real Time Information.
- DWH Supports DM and Business Intelligence.
- Database Is Used To Running The Business

DWH Is How to Run the Business

What is a Data Warehouse?

A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a way they can understand and use in a business context.

A process of **transforming data into information** and making it available to users in a timely enough manner to make a difference.

A process -

- It is a relational or multidimensional database management system designed to support management decision making.
- A data warehousing is a copy of transaction data specifically structured for querying and reporting.
- Technique for assembling and managing data from various sources for the purpose of answering business questions. Thus making decisions that were not previous possible.

- **Subject Oriented:** Data that gives information about a particular subject instead of about a company's ongoing operations.
- **Integrated:** Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.
- **Time-variant:** All data in the data warehouse is identified with a particular time period.

- **Non-volatile:** Data is stable in a data warehouse. More data is added but data is never removed. This enables management to gain a consistent picture of the business.
- Data warehousing is combining data from multiple and usually varied sources into one comprehensive and easily manipulated database.
- Common accessing systems of data warehousing include queries, analysis and reporting.
- Because data warehousing creates one database in the end, the number of sources can be anything you want it to be, provided that the system can handle the volume, of course.
- The final result, however, is homogeneous data, which can be more easily manipulated.

History of data warehousing:

- The concept of data warehousing dates back to the late 1980s when IBM researchers Barry Devlin and Paul Murphy developed the "business data warehouse".
- 1960s - General Mills and Dartmouth College, in a joint research project, develop the terms *dimensions* and *facts*.
- 1970s - ACNielsen and IRI provide dimensional data marts for retail sales.
- 1983 – Tera data introduces a database management system specifically designed for decision support.
- 1988 - Barry Devlin and Paul Murphy publish the article *An architecture for a business and information systems* in *IBM Systems Journal* where they introduce the term "business data warehouse".

A data warehouse is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use it in a business context.

Evolution in organizational use of data warehouses

Organizations generally start off with relatively simple use of data warehousing. Over time, more sophisticated use of data warehousing evolves. The following general stages of use of the data warehouse can be distinguished:

- **Off line Operational Database**
  - Data warehouses in this initial stage are developed by simply copying the data off an operational system to another server where the processing load of

reporting against the copied data does not impact the operational system's performance.

- **Off line Data Warehouse**
  - Data warehouses at this stage are updated from data in the operational systems on a regular basis and the data warehouse data is stored in a data structure designed to facilitate reporting.
- **Real Time Data Warehouse**
  - Data warehouses at this stage are updated every time an operational system performs a transaction (e.g. an order or a delivery or a booking.)
- **Integrated Data Warehouse**
  - Data warehouses at this stage are updated every time an operational system performs a transaction. The data warehouses then generate transactions that are passed back into the operational systems.

Benefits of Data Warehousing

The successful implementation of a data warehouse can bring major, benefits to an organization including:

• Potential high returns on investment

Implementation of data warehousing by an organization requires a huge investment typically from Rs 10 lack to 50 lacks. However, a study by the International Data Corporation (IDC) in 1996 reported that average three-year returns on investment (RO I) in data warehousing reached 401%.

• Competitive advantage

The huge returns on investment for those companies that have successfully implemented a data warehouse is evidence of the enormous competitive advantage that accompanies this technology. The competitive advantage is gained by allowing decision-makers access to data that can reveal previously unavailable, unknown, and untapped information on, for example, customers, trends, and demands.

• Increased productivity of corporate decision-makers

Data warehousing improves the productivity of corporate decision-makers by creating an integrated database of consistent, subject-oriented, historical data. It integrates data from multiple incompatible systems into a form that provides one consistent view of the organization. By transforming data into meaningful information, a data warehouse allows business managers to perform more substantive, accurate, and consistent analysis.

• More cost-effective decision-making

Data warehousing helps to reduce the overall cost of the· product· by reducing the number of channels.

• Better enterprise intelligence.

It helps to provide better enterprise intelligence.

• Enhanced customer service.

• It is used to enhance customer" service.

OLTP- ONLINE TRANSACTION PROCESSING

- Special data organization, access methods and implementation methods are needed to support data warehouse queries (typically multidimensional queries)
- OLTP systems are tuned for known transactions and workloads while workload is not known a priori in a data warehouse
  - e.g., *average amount spent on phone calls between 9AM-5PM in Pune during the month of December*

OLTP vs Data Warehouse

OLTP	**Data Warehouse**
• Application Oriented • Used to run business • Detailed data • Current up to date • Isolated Data • Clerical User • Few Records accessed at a time (tens) • Read/Update Access • No data redundancy • Database Size 100MB -100 GB • Transaction throughput is the performance metric • Thousands of users • Managed in entirety	– Subject Oriented – Used to analyze business – Summarized and refined – Snapshot data – Integrated Data – Knowledge User (Manager) – Large volumes accessed at a time (millions) – Mostly Read (Batch Update) – Redundancy present – Database Size 100 GB - few terabytes – Query throughput is the performance metric – Hundreds of users – Managed by subsets

Problems of Data Warehousing

The problems associated with developing and managing a data warehousing are as follows:

Underestimation of resources of data loading

Sometimes we underestimate the time required to extract, clean, and load the data into the warehouse. It may take the significant proportion of the total development time.

Hidden problems with source systems

Sometimes hidden .problems associated with the source systems feeding the data warehouse may be identified after years of being undetected. For example, when entering the details of a new property, certain fields may allow nulls which may result in staff entering incomplete property data, even when available and applicable.

Required data not captured

In some cases the required data is not captured by the source systems which may be very important for the data warehouse purpose. For example the date of registration for the property may be not used in source system but it may be very important analysis purpose.

Increased end-user demands

After satisfying some of end-users queries, requests for support from staff may increase rather than decrease. This is caused by an increasing awareness of the users on the capabilities and value of the data warehouse. Another reason for increasing demands is that once a data warehouse is online, it is often the case that the number of users and queries increase together with requests for answers to more and more complex queries.

Data homogenization

The concept of data warehouse deals with similarity of data formats between different data sources. Thus, results in to lose of some important value of the data.

High demand for resources

The data warehouse requires large amounts of data.

Data ownership

Data warehousing may change the attitude of end-users to the ownership of data. Sensitive data that owned by one department has to be loaded in data warehouse for decision making purpose. But some time it results in to reluctance of that department because it may hesitate to share it with others.

High maintenance

Data warehouses are high maintenance systems. Any reorganization· of the business processes and the source systems may affect the data warehouse and it results high maintenance cost.

Complexity of integration

The most important area for the management of a data warehouse is the integration capabilities. An organization must spend a significant amount of time determining how well the various different data warehousing tools can be integrated into the overall solution that is needed. This can be a very difficult task, as there are a number of tools for every operation of the data warehouse.

Operational Data and Data store -

An operational data store (ODS) is a type of database that's often used as an interim logical area for a data warehouse.

While in the ODS, data can be scrubbed, resolved for redundancy and checked for compliance with the corresponding business rules. An ODS can be used for integrating disparate data from multiple sources so that business operations, analysis and reporting can be carried out while business operations are occurring. This is the place where most of the data used in current operation is housed before it's transferred to the data warehouse for longer term storage or archiving.

An ODS is designed for relatively simple queries on small amounts of data (such as finding the status of a customer order), rather than the complex queries on large amounts of data typical of the data warehouse. An ODS is similar to your short term memory in that it stores only very recent information; in comparison, the data warehouse is more like long term memory in that it stores relatively permanent information.

Process managers are responsible for maintaining the flow of data both into and out of the data warehouse. There are three different types of process managers:

- Load manager
- Warehouse manager
- Query manager

Data Warehouse Load Manager

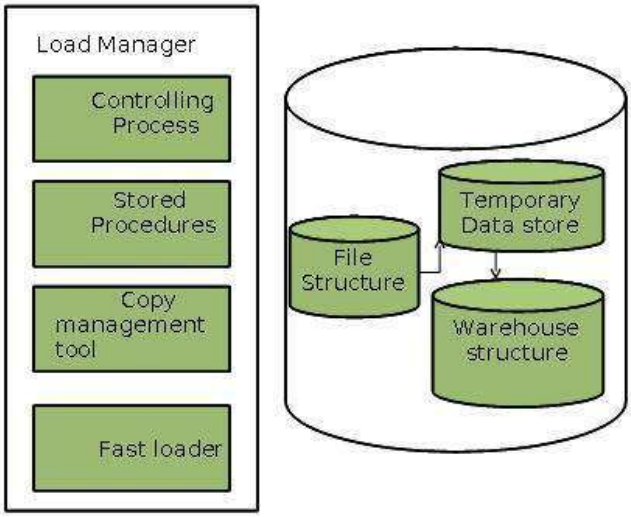
Load manager performs the operations required to extract and load the data into the database. The size and complexity of a load manager varies between specific solutions from one data warehouse to another.

Load Manager Architecture

The load manager does performs the following functions:

- Extract data from the source system.

- Fast load the extracted data into temporary data store.
- Perform simple transformations into structure similar to the one in the data warehouse.



Extract Data from Source

The data is extracted from the operational databases or the external information providers. Gateways are the application programs that are used to extract data. It is supported by underlying DBMS and allows the client program to generate SQL to be executed at a server. Open Database Connection (ODBC) and Java Database Connection (JDBC) are examples of gateway.

Fast Load

- In order to minimize the total load window, the data needs to be loaded into the warehouse in the fastest possible time.
- Transformations affect the speed of data processing.
- It is more effective to load the data into a relational database prior to applying transformations and checks.
- Gateway technology is not suitable, since they are inefficient when large data volumes are involved.

Simple Transformations

While loading, it may be required to perform simple transformations. After completing simple transformations, we can do complex checks. Suppose we are loading the EPOS sales transaction, we need to perform the following checks:

- Strip out all the columns that are not required within the warehouse.
- Convert all the values to required data types.

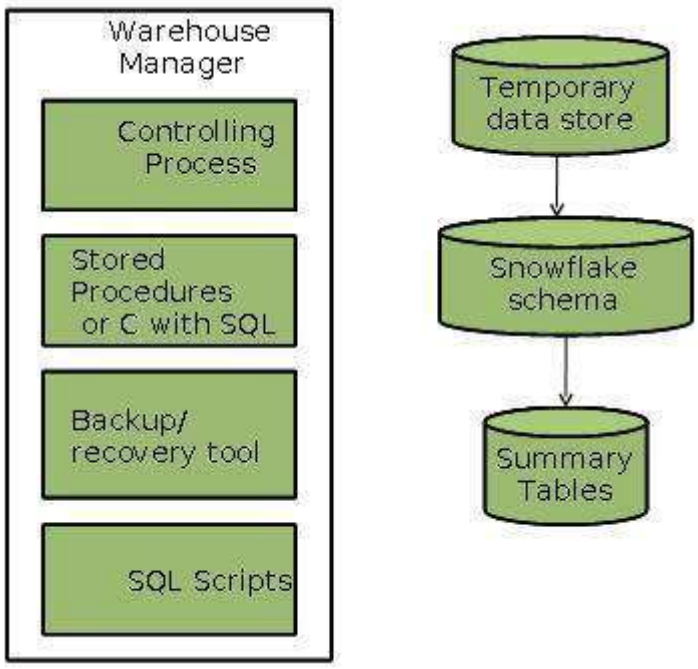
Warehouse Manager

The warehouse manager is responsible for the warehouse management process. It consists of a third-party system software, C programs, and shell scripts. The size and complexity of a warehouse manager varies between specific solutions.

Warehouse Manager Architecture

A warehouse manager includes the following:

- The controlling process
- Stored procedures or C with SQL
- Backup/Recovery tool
- SQL scripts



Functions of Warehouse Manager

A warehouse manager performs the following functions:

- Analyzes the data to perform consistency and referential integrity checks.
- Creates indexes, business views, partition views against the base data.

- Generates new aggregations and updates the existing aggregations.
- Generates normalizations.
- Transforms and merges the source data of the temporary store into the published data warehouse.
- Backs up the data in the data warehouse.
- Archives the data that has reached the end of its captured life.

Note: A warehouse Manager analyzes query profiles to determine whether the index and aggregations are appropriate.

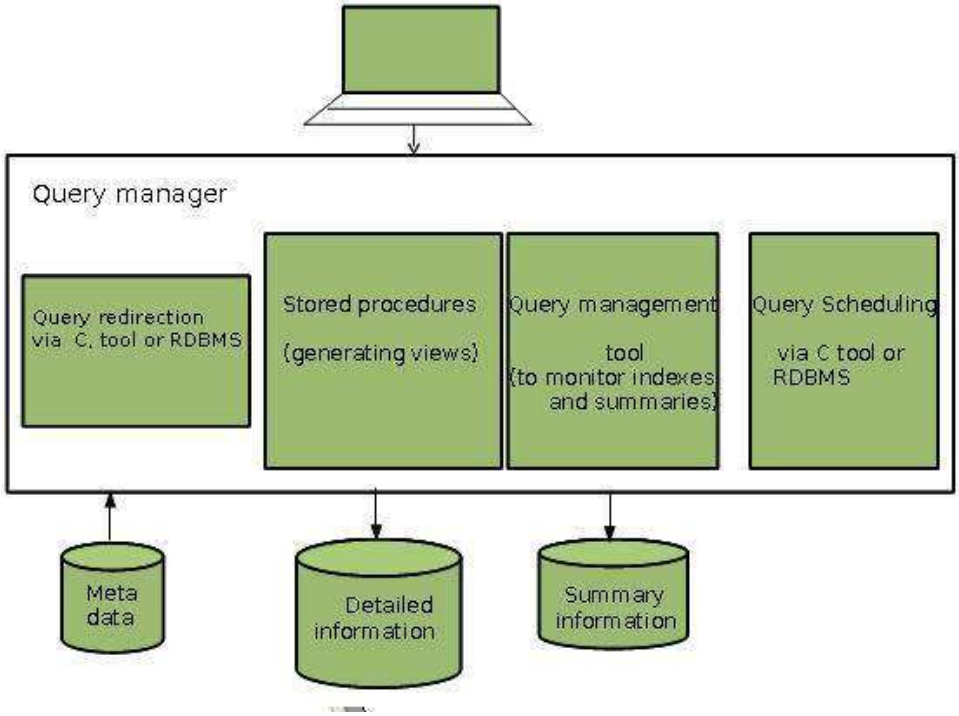
Query Manager

The query manager is responsible for directing the queries to suitable tables. By directing the queries to appropriate tables, it speeds up the query request and response process. In addition, the query manager is responsible for scheduling the execution of the queries posted by the user.

Query Manager Architecture

A query manager includes the following components:

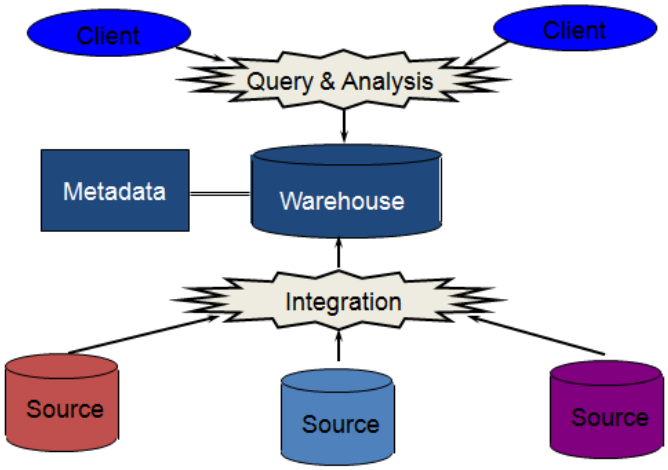
- Query redirection via C tool or RDBMS
- Stored procedures
- Query management tool
- Query scheduling via C tool or RDBMS
- Query scheduling via third-party software



Functions of Query Manager

- It presents the data to the user in a form they understand.
- It schedules the execution of the queries posted by the end-user.
- It stores query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate.

Data Warehouse Architecture



- The data has been selected from various sources and then integrate and store the data in a single and particular format.

- Data warehouses contain current detailed data, historical detailed data, lightly and highly summarized data, and metadata.
- Current and historical data are voluminous because they are stored at the highest level of detail.
- **Lightly and highly summarized data** are necessary to save processing time when users request them and are readily accessible.
- ***Metadata*** are "data about data". It is important for designing, constructing, retrieving, and controlling the warehouse data.

Technical metadata include where the data come from, how the data were changed, how the data are organized, how the data are stored, who owns the data, who is responsible for the data and how to contact them, who can access the data , and the date of last update.

Business metadata include what data are available, where the data are, what the data mean, how to access the data, predefined reports and queries, and how current the data are.

In **Lightly Summarized Data** the evaluational data is summarized by removing one, or a few, data characteristic from the primary key of the data focus.

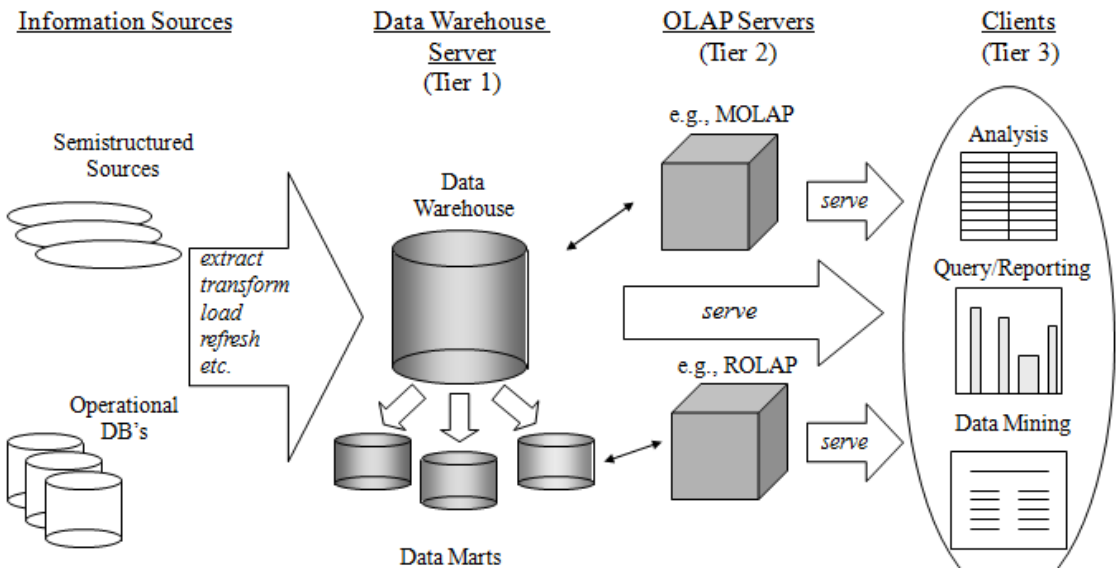
Highly Summarized Data are summarized data obtained by removing many data characteristics from the primary key of the data focus. Dealing with highly summarized data which refer to evaluational data, greatly falls under the domain of a data warehousing implementation.

DW Three-tier architecture:

a. Data acquisition software (back-end).
b. The data warehouse that contains the data & software
c. Client (front-end) software that allows users to access and analyze data from the warehouse

Two-tier architecture: First two tiers in three-tier architecture is combined into one,

The Complete Decision Support System



Data Warehouse vs. Data Marts

- *Enterprise warehouse*: collects all information about subjects (*customers,products,sales,assets, personnel*) that span the entire organization
  - Requires extensive business modeling (may take years to design and build)
- *Data Marts*: Departmental subsets that focus on selected subjects
  - Marketing data mart: customer, product, sales
  - Faster roll out, but complex integration in the long run
- *Virtual warehouse*: views over operational dbs
  - Materialize sel. summary views for efficient query processing
  - Easy to build but require excess capability on operat. db servers