Data Warehouse

A data warehouse is constructed by integrating data from multiple heterogeneous source. It supports analytical reporting, structured or adhoc queries and decision making.

- A data ware is a subject oriented, integrated, time-variant and non-volatile collection of data.

* Data warehouse is kept separate from the organization's operational database.
* There is no frequent updating done in a data warehouse.
* It possesses consolidated historical data, which helps the organization to analyze its business.
* Data warehouse also provides us Online Analytical processing (OLAP) tools. These tools helps us in interactive and effective analysis of data in a multidimensional space. This analysis results in data generalization and data mining.

Operational Database

- Constructed for well-known tasks and workloads such as searching particular records, indexing, etc.
- Support concurrent processing of multiple transactions. Concurrency control and recovery mechanisms are required for operational databases to ensure robustness and consistency of the database.

- An operational database query allows to read and modify operations, while an OLAP query needs only read only access of stored data.
- An operational database maintains current data. While data warehouse maintains historical data.

Data warehouse Features.

1) Subject-oriented - provides info about the subject such as product, customers, suppliers, sales, revenue etc.; and do not focus on the ongoing operations.
2) Integrated - data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files etc.
3) Time variant - data collected is identified with a particular time.
4) Non-volatile - Previous data is not erased when new data is added to it.

Using Data warehouse Information

→ Tuning Production strategies
- Customer Analysis - analysing customer buying preferences, ~~b~~ buying time etc.
- Operations Analysis - helps in customer relationship management & making environmental corrections.

Functions of Data warehouse Tools and Utilities

- Data Extraction
- Data cleaning
- Data Transformation
- Data Loading
- Refreshing

Metadata -

- Metadata is defined as data about data. It is the summarized data that leads us to detailed data.
- Metadata is a road-map to datawarehouse
- Metadata in datawarehouse defines the warehouse objects.
- Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

Metadata Repository

- Integral part of a datawarehouse system. It contains following Metadata -
- Business metadata, operational metadata, Data for mapping from operational environment to datawarehouse, algorithms for summarization.

Data cube

- Helps us to represent data in multiple dimensions. Dimensions are the entities with respect to which an enterprise preserves the records.

Data Mart

- Data Mart contains a subset of organization-wide data that is valuable to specific groups of people in an organization.
  eg. Marketing data may contain only data related to items, customers & sales.

Virtual Warehouse

- The view over an operational data warehouse is known as virtual warehouse.

DWH Architecture

Business Analysis Framework (for the data warehouse design & architecture of DW)

- The business analyst get the information from the data warehouses. to measure the performance and make critical adjustments in order to win over other business holders in the market.
- since a data warehouse can gather information quickly & efficiently, it can enhance business productivity.
- To design an effective and efficient data warehouse we need to understand and analyze the business needs and construct a business analysis frameworks.

Different views for designing DW

- Top-down view
- Data source view
- Data Warehouse View
- Business query view

3-Tier Data Warehouse Architecture

⇒ **Bottom Tier** - is the data warehouse database server. It is a Relational database system.

⇒ **Middle Tier** - In middle tier we have OLAP server that can be implemented in following way -

1) By Relational OLAP (ROLAP), which is extended Relational database management system.
2) By Multidimensional OLAP (MOLAP) - implement multidimensional data & operations.

⇒ **Top-Tier** - is the front-end client layer. This layer holds query tools, reporting, analysis & data mining tools.

Data Warehouse models -

- Virtual Warehouse
- Data Mart
- Enterprise Warehouse

Load Manager

- This component performs the operations required to extract and load process.
- It perfoms following functions :-
  - Extract the data from source system
    - from operational databases or external information providers.

Gateways is the application programs that are used to extract data. It is supported by DBMS & allows client programs to generate SQL to be executed at a server.
ODBC & JDBC are examples of Gateways.

- Fast load the extracted data into temporary data store.
- Perform simple transformations into structure

Warehouse Manager

- Responsible for the warehouse management process. It consists of third-party system software, C Programs, shell scripts.
- Warehouse Manager includes -
  - The controlling process
  - Stored procedures or C with SQL
  - Backup / Recovery tool
  - SQL scripts

Query Manager

- It is responsible for directing queries to suitable tables.

- It is responsible for scheduling the execution of queries posed by the user.

DWH- OLAP

Online Analytical Processing Servers (OLAP) is based on the multidimensional data model. It allows managers, and analysts to get an insight of the information through fast, consistent and interactive access to information.

Types of OLAP servers

1. Relational OLAP (ROLAP) -
ROLAP servers are placed between back-end server and client front-end tools. To store and manage warehouse data, ROLAP uses relational or extended relational DBMS.

2) Multidimensional OLAP :
MOLAP uses array-based multidimensional storage engines for multidimensional views of data.

3) Hybrid OLAP -
Hybrid OLAP is a combination of both ROLAP & MOLAP. It offers high scalability of ROLAP and faster computation of MOLAP. It allows to store the large data volumes of detailed information.

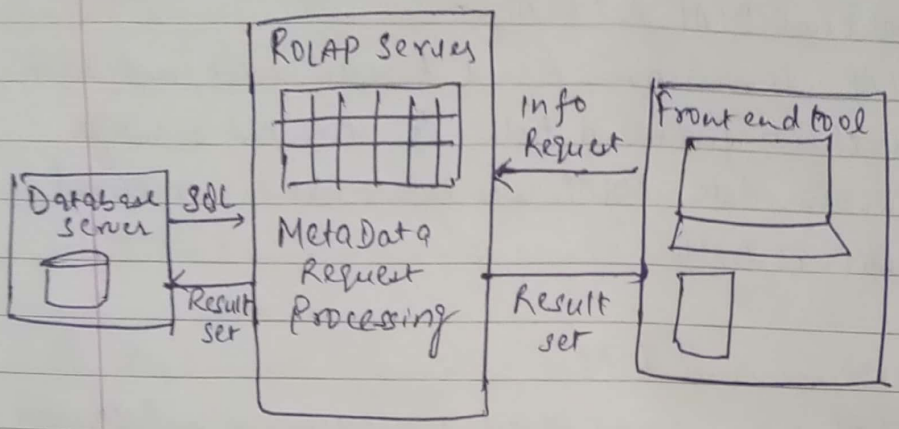
4) Specialized SQL servers -
Provides advanced query language and query processing support for SQL queries over star and snowflake schemas in a read only environment

OLAP operations

- Roll-up
  Perform aggregation of data cube by climbing up concept hierarchy for a dimension or by dimension reduction

- Drill-down
- Slice and Dice
- Pivot (rotate)

ROLAP Architecture



DWH - Schemas

Schema is a logical description of the entire database. It includes name and description of records of all record types including all associated data-items and aggregates.

- Much like database, a data warehouse also requires to maintain a schema.
- A database uses relational model, while data warehouse uses Star, Snowflake and Fact Constellation schema.

- Star Schema - Each dimension in a star schema is represented with only one-dimension table

- eg. Sales data of a company can have 4 dimensions namely time, item, branch, location

Snowflake schema
- Some dimension tables in the snowflake schema are normalized.
Normalization splits up the data into additional tables.
eg. item dimension can be split into item & supplier table

Fact constellation schema
- It has multiple fact tables. Also known as galaxy schema.

- Multidimensional schema is defined using Data Mining Query Language (DMQL)

eg define cube sales star [time, item, branch, location]:
define dimension time as (time key, day, day of week, month, quarter, year)
" " item
" " branch
" " location

Why need Data Mart -
- To partition data in order to impose access control strategies
- To speed up the queries by reducing the volume of data to be scanned.
- To segment data into diff - hardware platforms
- To structure data in a form suitable for a user access tool.