Chapter 04:-
Data warehouse

Data warehousing:-

Data warehousing is a process for collecting and managing data from varied sources to provide meaningful business insights. A data warehouse is typically

Mon Tue Wed Thu Fri Sat

used to connect and analyze business data from heterogenous sources. The datawarehouse is the core of the BI system which is built for data analysis and reporting.

According to definition, data warehouses are:

Subject-oriented:-

They can analyze data about a particular subject or functional area. (such as sales).

Integrated:-

Data warehouses create consistency among different data types from disperate sources.

Non-volatile:-

One data is in a data warehouse, it's stable and doesn't change.

Mon Tue Wed Thu Fri Sat

تاریخ:

Time - variant:-

Data warehouse analysis looks at change over time.

OLTP Vs. OLAP

OLAP stands for (Online analytical processing) and OLTP stands for (Online transaction processing).

Category	OLAP	OLTP
Defination	It is well-known as an online database queury managment system.	It is well-known as an online database modifying system.
Data Source	Consist of historical data from various databases.	Consist of operational current data.
Method used	It makes use of a data warehouse.	It makes use of a standarad (DBMS).
Application	It is subject-oriented. Used for	It is application-oriented. Used

Category	OLAP	OLTP
	Data Mining Analytics, Decisions making, etc.	for business tasks.
Normalized	In OLAP database, tables are not normalized.	In OLTP database, tables are normalized (3NF).
Usage of data	The data is used in planning, problem-solving, and decision making.	The data is used to perform day-by-day fundamental operations.
Task	It reveals a snapshot of present business tasks.	It provides a multidimensional view of different business tasks.
Productivity	Improves the efficiency of business analysts.	Enhances the user's productivity.
Nature of audience	Process that is focused on the	Process that is focused on the

Category	OLAP	OLTP
	customer.	market.
Types of users	The data is generally managed by CEO, MD, GM.	The data is managed by clerks, managers.
Volume of data	A large amount of data is stored typically in TB, PB.	The size of data is relatively small as historical data as in MB, GB.

Why Separate Data Warehouse?

High performance for both systems:-

→ DBMS turned for OLTP: access methods, indexing, concurrency control, recovery.

→ Warehouse turned for OLAP: complex OLAP queuries, multidimensional view, consolidation.

Different functions and data:-

→ Missing data: Decision support requires historical data which operational DBs do not typically maintain.

→ Data consolidation:- Ds requires consolidation (aggregation, summarization) of data from heterogeneous sources.

→ Data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled.

Multi-tier architecture of Data warehouse:-

A data warehouse is representable by data integration from multiple heterogenous sources.

Data warehouse is refferred to data repository that is maintained separately from the organization's

operational data.

Multi-tier data warehouse architecture consist of following components.

→ Bottom Tier

→ Middle Tier

→ Top Tier

Bottom Tier (Data Sources and Data Storage):-

→ The bottom tier usually consist of Data sources and Data storage.

→ It is a warehouse database server. For example RDBMS.

→ In bottom tier, using the application program interface (called gateways), data is extracted from operational and external sources.

→ Application program interface likes ODBC (Open Database Connection), OLEDB (Open-Linking and Embedding for Database), JDBC (Java Database Connection) is supported.

Mon Tue Wed Thu Fri Sat

Middle Tier:-

The middle tier is an OLAP server that is typically implemented using either: A relation OLAP (ROLAP) model (i.e an extended relational DBMS that maps operation from standarad data to standarad data) or a multidimensional OLAP (MOLAP) model (i.e a special purpose server that indirectly implements multidimensional data and operations).

Top Tier:-

Top tier is a front-end client layer, which includes queury and reporting tools, analysis tools, and/or data mining tools (e.g. trend analysis, prediction, etc).

Data Warehouse Models:-

From the architecture point of view, there are three warehouse models.

→ Enterprise Warehouse

→ Data Mart

→ Virtual Warehouse

Enterprise Warehouse:-

→ An enterprise warehouse collects all information topics spread throughout the organization.

→ It provides corporate-wide data integration, typically from one or several operational systems or external information providers, and is cross-functional in scope.

→ It usually contains detailed data as well as summarized data and can range in size from a few gigabytes to hundreds of gigabytes,

terabytes or beyond. Can be an enterprise data warehouse.

→ The traditional mainframe, computer super server, or parallel architecture has been implemented on platforms. This require extensive commercial modeling and may take years to design and manufacture.

Data Mart:-

→ A data mart contains a subset of corporate-wide data that is important to a specific group of users.

→ The scope is limited to specific selected subjects.

→ For example, a marketing data mart may limit its topics to customers, goods and sales.

→ The data contained in the data marts are summarized. Data marts are typically applied

to low-cast departmental servers that are linux / unix or windows based.

Virtual Warehouse:-

→ A virtual warehouse is a group of views on a operational database.

→ For efficient queury processing, only a few possible summary views can be physical.

→ Creating a virtual warehouse is easy, but requires additional capacity on operational database servers.

Extraction, Transformattion and Loading (ETL).

Extract:-

The first step of this process

is extracting data from the target sources that are usually heterogeneous.

Three Data Extraction methods:

→ **Partial Extraction**:-

The easiest way to obtain the data is if the source system notifies you when a record has been changed.

→ **Partial Extraction** (with **update notification**):-

Not all systems can provide a notification in case an update has taken place; however they can point to those records that have been changed and provide an extract of such records.

→ **Full Extract**:-

There are certain systems that cannot identify which data has

been changed it all. In this case, a full extract is the only possibility to extract the data out of the system. This method requires having a copy of the last extract in the same format so you can identify the changes that have been made.

Transform:-

The second step consists of transforming the raw data that has been extracted from the sources into a format that can be used by different applications. In this stage, data gets cleansed, mapped and transformed, often to a specific schema, so it meets operational needs. During this stage, you have the possibility to generate audit

reports for regulatory compilance, or diagnose and repair any data issues.

Load:-

Finally, the load function is the process of writing converted data from a staging area to a target database, which may or may not have previously existed. Depending on the requirements of the application, this process may be either quite simple or intricate. Each of these steps can be done with ETL tools or custom code.

Metadata Repository:-

Metadata is the data defining warehouse objects. It stores:

- Description of structure of datawarehouse
- Operational meta-data
- The algorithms used for summarization

→ The mapping from operational environment to the data warehouse
→ Data related to system performance
→ Business Data

Data Cube:-

A data cube is a multidimensional data model that store the optimized, summarized or aggregated data which ease the OLAP tools for the quick and easy analysis. Data cube stores the precomputed data and eases online analytical processing.

Data stored in data cube is represented in term of dimensions and facts. Each dimension has a dimension table which contains a further description of that dimension. Such as branch dimension may have branch-name, branch-code, branch-address, etc.