# Concepts of Data Warehouse

## What is Data Warehousing?

The process of creating data warehouses to store a large amount of data is named Data Warehousing. Data Warehousing helps to improve the speed and efficiency of accessing different data sets and makes it easier for company decision-makers to obtain insights that will help the business and promoting marketing tactics that set them aside from their competitors. We can say that it is a blend of technologies and components which aids the strategic use of data and information. The main goal of data warehousing is to create a hoarded wealth of historical data that can be retrieved and analyzed to supply helpful insight into the organization's operations.

## Need of Data Warehousing

Data Warehousing is a progressively essential tool for business intelligence. It allows organizations to make quality business decisions. The data warehouse benefits by improving data analytics, it also helps to gain considerable revenue and the strength to compete more strategically in the market. By efficiently providing systematic, contextual data to the business intelligence tool of an organization, the data warehouses can find out more practical business strategies.

## Characteristics of Data warehouse:

1. **Subject Oriented:** A data warehouse is often subject-oriented because it delivers may be achieved on a particular theme which means the data warehousing process is proposed to handle a particular theme that is more defined. These themes are often sales, distribution, selling. etc.

2. **Time-Variant:** When the data is maintained via totally different intervals of time like weekly, monthly or annually, etc. It founds numerous time limits that are unit structured between the big datasets and are command within the online transaction method (OLTP). The time limits for the data warehouse are extended than that of operational systems. The data resided within the data warehouse is predetermined with a particular interval of time and delivers information from the historical perspective. It contains parts of time directly or indirectly.

3. **Non-volatile:** The data residing in the data warehouse is permanent and defined by its names. It additionally means that the data in the data warehouse is cannot be erased or deleted or also when new data is inserted into it. In the data warehouse, data is read-only and can only be refreshed at a particular interval of time. Operations such as delete, update and insert that is done in a software application over data is lost in the data warehouse environment. There are only two types of data operations that can be done in the data warehouse:
    1. Data Loading
    2. Data Access

4. **Integrated:** A data warehouse is created by integrating data from numerous different sources such that from mainframe computers and a relational database. Additionally, it should also have reliable naming conventions, formats, and codes. Integration of data warehouse benefits in the successful analysis of data.Dependability in naming conventions, column scaling, encoding structure, etc. needs to be confirmed. Integration of data warehouse handles numerous subject-oriented warehouses.

# OLAP vs OLTP

### Architecture & Components of Data Warehouse:

Data warehouse architecture defines the comprehensive architecture of data processing and presentation that will be useful for data analysis and decision making within the enterprise and organization. Each organization has different data warehouses depending upon their need, but all of them are characterized by some standard components.

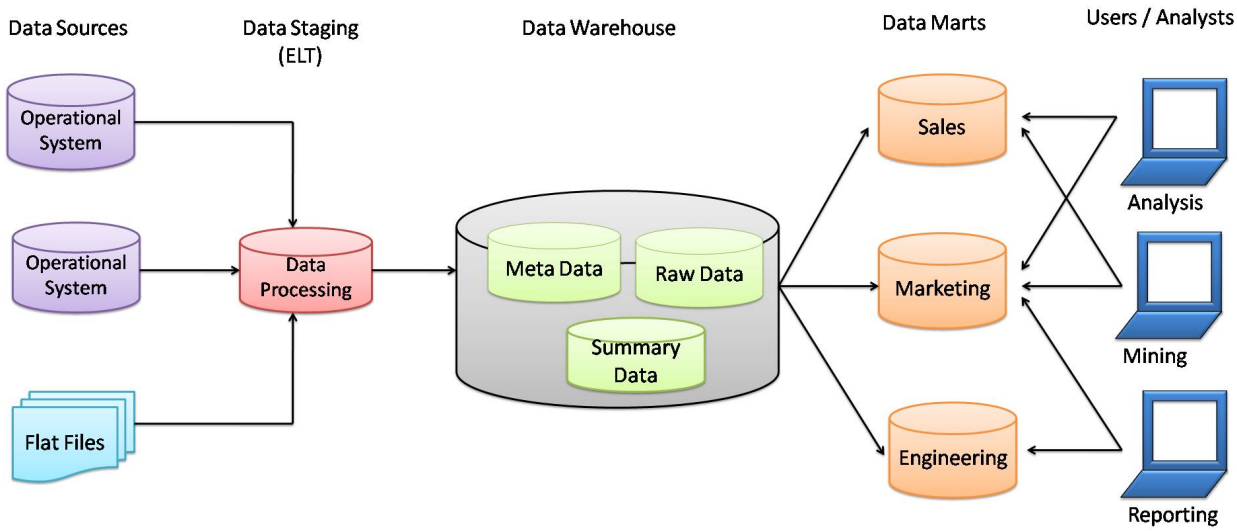
Data Warehouse applications are designed to support the user's data requirements, an example of this is online analytical processing (OLAP). These include functions such as forecasting, profiling, summary reporting, and trend analysis.

The architecture of the data warehouse mainly consists of the proper arrangement of its elements, to build an efficient data warehouse with software and hardware components.

The elements and components may vary based on the requirement of organizations. All of these depend on the organization's circumstances.

| OLAP | OLTP |
|---|---|
| ✓ Gives a multi-dimensional view of business activities. | ✓ Enables a snapshot of ongoing business processes. |
| ✓ Helps with planning, problem solving, and decision support. | ✓ Useful for controlling and running fundamental business tasks. |
| ✓ Data source is consolidated data | ✓ Data source is the operational data. |
| ✓ Includes Periodic long-running batch jobs that refresh the data. | ✓ Has short and fast inserts and updates which are initiated by end users. |
| ✓ OLAP applications are widely used by Data Mining techniques. | ✓ Large number of short on-line transactions |
| ✓ Database design is typically de-normalized and contains fewer tables. | ✓ Database design in OLTP is highly normalized. |
| ✓ Often involves complex queries along with aggregations, which in turn compels processing speed to be dependent on the amount of data involved; batch data refreshes, etc. | ✓ Involves standardized and simple queries that return relatively few records hence is faster. |

## OLAP vs OLTP




DATA WAREHOUSE ARCHITECTURE

## 1. Source Data Component:

In the Data Warehouse, the source data comes from different places. They are group into four categories:

- **External Data:** For data gathering, most of the executives and data analysts rely on information coming from external sources for a numerous amount of the information they use. They use statistical features associated with their organization that is brought out by some external sources and department.

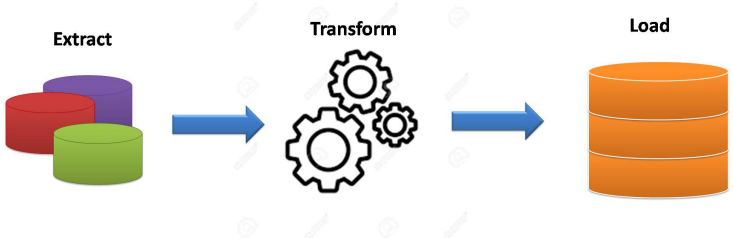
- **Internal Data:** In every organization, the consumer keeps their "private" spreadsheets, reports, client profiles, and generally even department databases. This is often the interior information, a part that might be helpful in every data warehouse.

- **Operational System data:** Operational systems are principally meant to run the business. In each operation system, we periodically take the old data and store it in achieved files.

- **Flat files:** A flat file is nothing but a text database that stores data in a plain text format. Flat files generally are text files that have all data processing and structure markup removed. A flat file contains a table with a single record per line.

## 2. Data Staging:

After the data is extracted from various sources, now it's time to prepare the data files for storing in the data warehouse. The extracted data collected from various sources must be transformed and made ready in a format that is suitable to be saved in the data warehouse for querying and analysis. The data staging contains three primary functions that take place in this part:



- **Data Extraction:** This stage handles various data sources. Data analysts should employ suitable techniques for every data source.

- **Data Transformation:** As we all know, information for a knowledge warehouse comes from many alternative sources. If information extraction for a data warehouse posture huge challenges, information transformation gifts even important challenges. We tend to perform many individual tasks as a part of information transformation. First, we tend to clean the info extracted from every source of data. Standardization of information elements forms an outsized part of data transformation. Data transformation contains several kinds of combining items of information from totally different sources. Information transformation additionally contains purging supply information that's not helpful and separating outsourced records into new mixtures. Once the data transformation performs ends, we've got a set of integrated information that's clean, standardized, and summarized.

- **Data Loading:** When we complete the structure and construction of the data warehouse and go live for the first time, we do the initial loading of the data into the data warehouse storage. The initial load moves high volumes of data consuming a considerable amount of time.

### 3. Data Storage in Warehouse:

Data storage for data warehousing is split into multiple repositories. These data repositories contain structured data in a very highly normalized form for fast and efficient processing.

- **Metadata:** Metadata means data about data i.e. it summarizes basic details regarding data, creating findings & operating with explicit instances of data. Metadata is generated by an additional correction or automatically and can contain basic information about data.

- **Raw Data:** Raw data is a set of data and information that has not yet been processed and was delivered from a particular data entity to the data supplier and hasn't been processed nonetheless by machine or human. This data is gathered out from online sources to deliver deep insight into users' online behavior.

**Summary Data or Data summary:** Data summary is an easy term for a brief conclusion of an enormous theory or a paragraph. This is often one thing where analysts write the code and in the end, they declare the ultimate end in the form of summarizing data. Data summary is the most essential thing in data mining and processing.

### 4. Data Marts:

Data marts are also the part of storage component in a data warehouse. It can store the information of a specific function of an organization that is handled by a single authority. There may be any number of data marts in a particular organization depending upon the functions. In short, data marts contain subsets of the data stored in data warehouses. Now, the users and analysts can use data for various applications like reporting, analyzing, mining, etc. The data is made available to them whenever required.

# How to Choose the right data warehouse for a data Analysis project?

The most important criteria to choose data warehouse are as follows:

• Type of data
• Volume of data
• Real-time analysis / after-fact-analysis

# Type of data :

- **Structured data:**

If you are using Structured data , a relational database like Postgres, MySQL,Amazon Redshift or BigQuery will fit your needs. These structured, relational databases are great when you know exactly what kind of data you're going to receive and how it links together, basically how rows and columns relate.

- **Unstructured data:**

If your data is of unstructured data type , you should look into a non-relational (NoSQL) database like Hadoop or Mongo.] If we are dealing with large amount of data, then non-relational database best suits because it won't impose restraints on incoming data, allowing for faster queries with scalability

- **Semi-structured data:**

Non-relational databases excel with extremely large amounts of data points of semi-structured data.

# Volume of data:

| DATABASE OPTIONS BY SCALE | | | | |
|---|---|---|---|---|
| DATA SIZE | < 1TB | 2TB-64TB | 64TB-2PB | #ALLOFTHEDATA |
| DATABASE THAT'S A GOOD FIT | Postgres MySQL | Amazon Aurora | Amazon Redshift Google BigQuery | Hadoop |

Database options by Scale

If you're under 1 TB of data, Postgres will give you a good price to performance ratio. But, it slows down around 6 TB. If you like MySQL but need a little more scale, Amazon Aurora can go up to 64 TB. For petabyte scale, Amazon Redshift is usually a good choice since it's optimized for running analytics up to 2PB.

# Real-Time analysis/ after-fact-anlaysis

If you absolutely need real-time data, you should look at an unstructured database like Hadoop. You can design your Hadoop database to load very quickly, though queries may take longer at scale depending on RAM usage, available disk space, and how you structure the data.

If you're mostly working on after-fact analysis, you should go for a database that is optimized for analytics like Redshift or BigQuery. These kind of databases are designed

under the hood to accommodate a large amount of data and to quickly read and join data, making queries fast. They can also load data reasonably fast (hourly) as long as you have someone vacuuming, resizing, and monitoring the cluster.

# Other factors considered for choosing Datawarehouse:

### Scale vs. Speed

When you need speed, consider Postgres: Under 1TB, Postgres is quite fast for loading and querying. Plus, it's affordable. As you get closer to their limit of 6TB (inherited by Amazon RDS), your queries will slow down. That's why when you need scale, Redshift is recommended. Redshift have the best cost to value ratio.

### Third-party Ecosystem

Redshift has a very large ecosystem of third-party tools. AWS has options like Segment Data Warehouses to load data into Redshift from an analytics API, and they also work with nearly every data visualization tool on the market. Fewer third-party services connect with Google, so pushing the same data into BigQuery may require more engineering time, and you won't have as many options for BI software.

If you already use Google Cloud Storage instead of Amazon S3, you may benefit from staying in the Google ecosystem.