ASSESSMENT 1

*“INTRODUCTION TO DATA ENGINEERING & DATA WAREHOUSING”*



Submitted By

VAIBHAV PATIDAR

IPS ACADEMY, INDORE (M.P.)

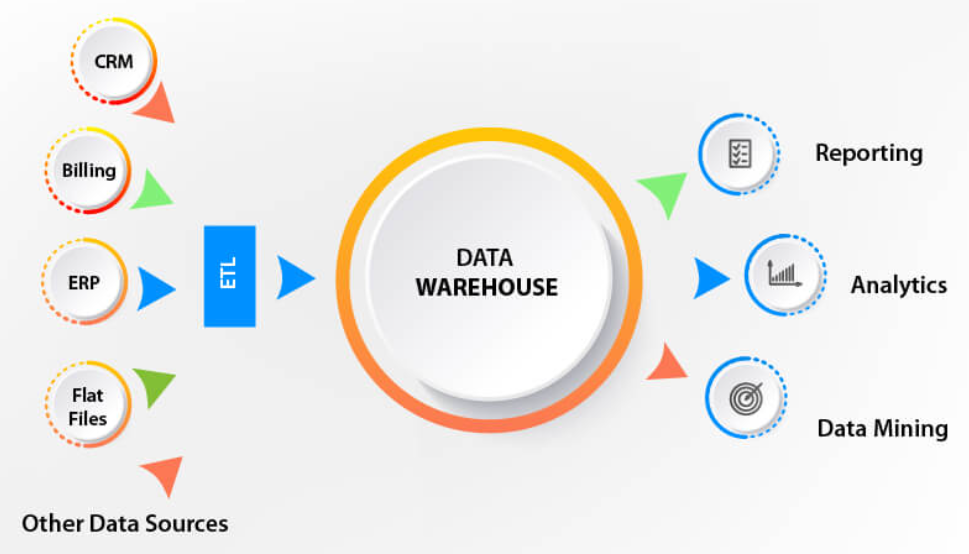
→ **Introduction to “Data Warehousing”**

Warehouse is a term refers to “repository” or “storage space” and data warehouse means - *“A centralized space or repository where data from multiple sources is collected and stored and can be used further to generate useful insights to skyrocket the business growth”.*

It is similar to physical warehouses which are used to store products and other stuffs. For example, flipkart, Amazon, Blinkit has their own warehouses where they store all the products delivered by them in an efficient well-maintained manner. So, that whenever a product is ordered , it can be easily dispatched.

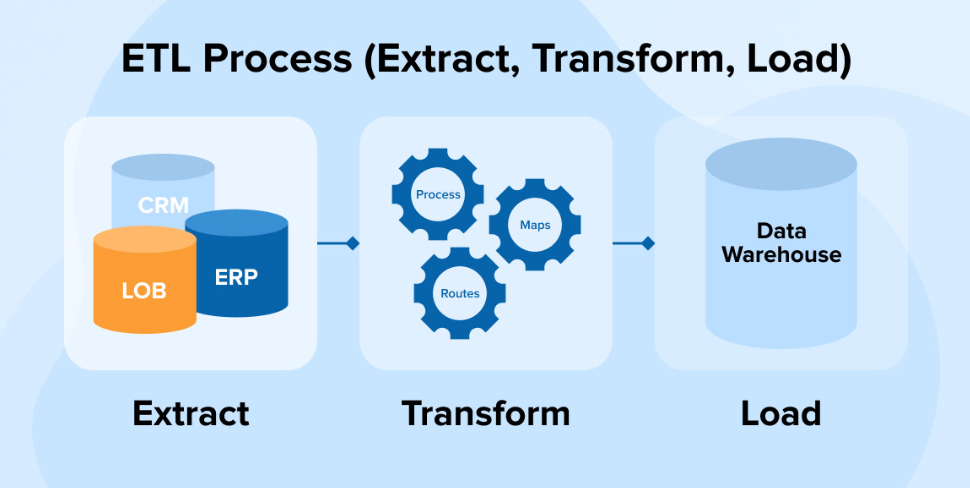
Just like this in data warehouses, first the data is gathered and stored in a such a way so that it can be easily retrieved and processed as and when required.

**Data Warehouse Architecture :**

****

**Extract, Transform & Load (ETL):**

* As we discussed, the data is collected from different sources, for this a process is followed, known as ETL (Extract, Transformation, Loading). In this multiple data from different sources is integrated in the data warehouse.
* **Some popular example of ETLs are :** Oracle Data Integrator, Microsoft SSIS (SQL Server Integration Service)
* Then once the data is stored, then multiple processing is done on data to remove the unwanted data and restored the useful data in the warehouse in the same form it was gathered.
* Then, this optimized data is further used for mining and analysis and used to create informational dashboards and predictions to study the business scenario and examine the future growth.



→ **Purpose of “Data Warehouse”**

* Data Warehouse is must needed for taking “quick & effective” decisions.
* It helps users in providing useful information.
* Data warehouse creates and maintain efficient data repository.
* It is also useful in providing historical business information to the data or business analyst.
* Chances of database failure decreases.
* It is also useful in data mining process.

**Benefits :**

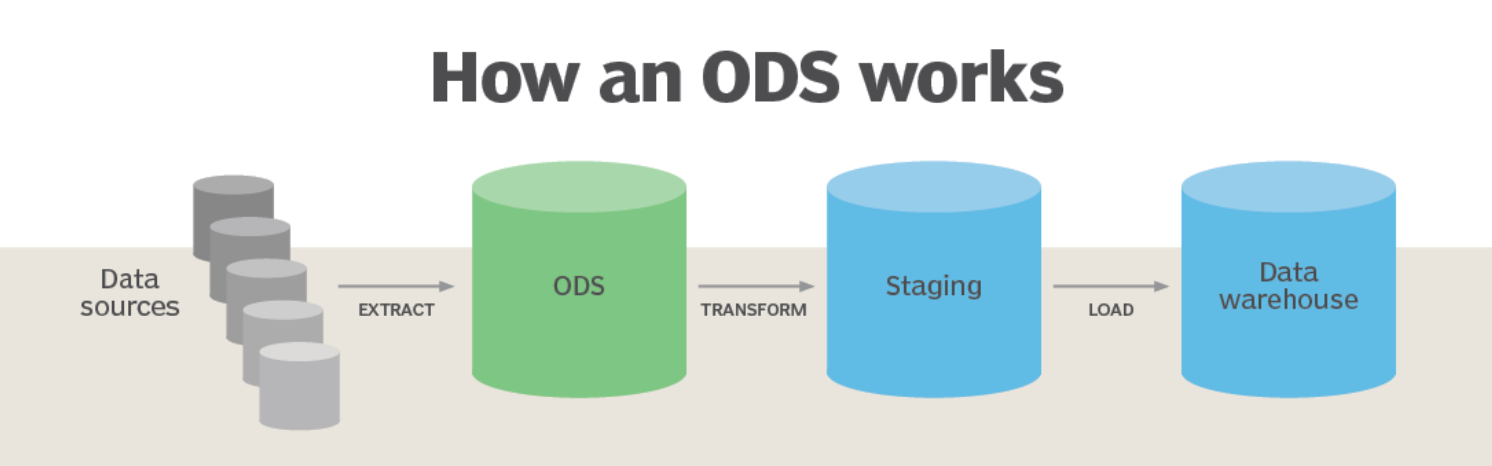
* It increases the product inventory turnover.
* Decreases the cost of production.
* Inhance the targeted market selection process.
* It develops better business intelligence.

**Goals of Data Warehousing :**

* Secured and easy access of information to the user.
* Consistent information must be provided by the user to the data warehouse.
* Clean and authentic data for analysis.
* Data collected must be accurate and verified.
* Data must be adaptive in nature.

→ **What is an “Operational Data Source”?**

* “Operational Data Source (ODS)” is the place where the data comes first during the current operations and from here it is send to the data warehouse for further storing or processing.
* ODS acts as a intermediate logical area for data warehouse, means it interact with the data prior to the warehouse.
* ODS are commonly used in applications processing transactions and generates transactional data.
* Operational Data Source, process the data in real time without any delay.
* It is connected to multiple data sources located at different targets, and collects data from all sources at the same point of time.
* ODS generally stores the data in its original form.
* Unlike ETL, where the data is fetched, processed and then send for further analysis, in ODS data is fetched and sirectly transferred to BI tools for analysis.

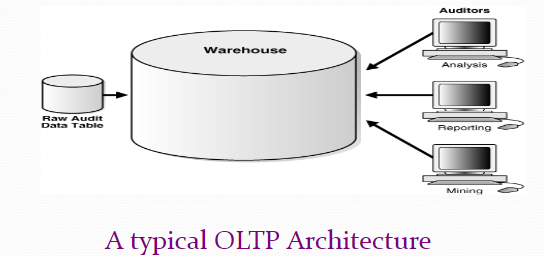


→ **Difference between OLTP and Data Warehouse Applications**

Data Warehouse is a Relational Database (RDBMS) system, which is used by application that process transaction. While, if wwe talk about OLTP (Online Transaction Processing) is a process-oriented system that processes transactions directly through connected computers over a network.

There are many differences between OLTP and Data warehouse, but the top 5 difference between OLTP and data warehouse are as follows :

1. **Updated Data :** As we know, in data warehouse a huge amount of data is stored which is not at all easy task to update the whole data. So, it is updated after a fixed period of time. While in OLTP , the data updated continuously or in real time.
2. **Completeness :** A data warehouse contains the historic to latest data of a business, so there is a chance it might have some incomplete information. While the OLTP can easily gather the incomplete information at real time.
3. **Structure :** A data warehouse generally divides a large table into smaller ones and then process them. While the OLTP can handle the data in an optimized way.
4. **Working Pressure :** Data warehouse generally supports and work on different types of queries. But, OLTP is programmed with fixed number of queries and it work around them only.



→ **Explain Data Marts**

As we know Data Warehouse is a Relational Database (RDBMS) system, which stores all type of data of an organization.

But, Data Marts is a data store which is designed specifically for a particular department of an organization.

We can also say, Data Mart is a sub set of data warehouse that is usually oriented to a specific purpose.

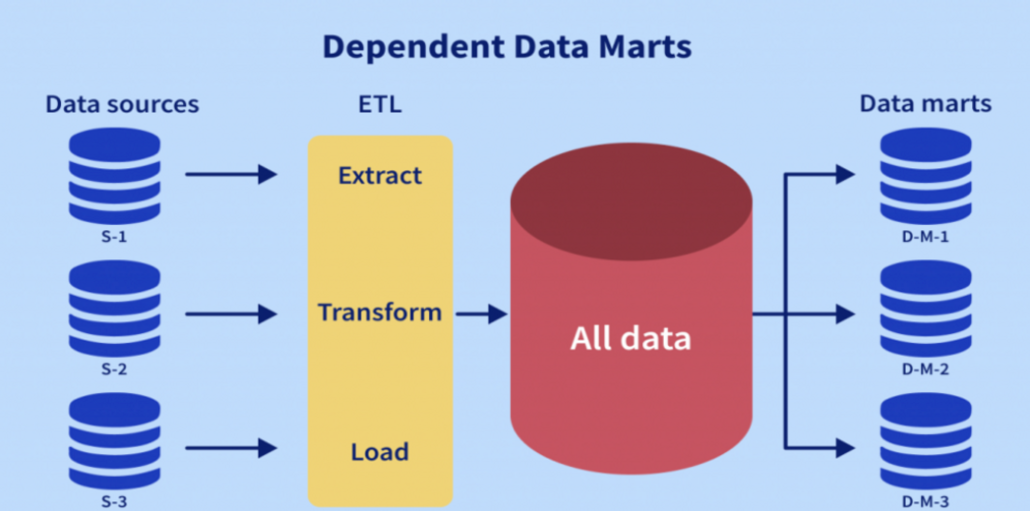
**Reasons for creating Data Marts :**

1. By data mart, easy access of frequent data is possible.
2. It also improves the response time at the user’s end.
3. Creating a data mart is easy.
4. The cost of creating a data mart is cost efficient.

**Different types pf Data Mart :**

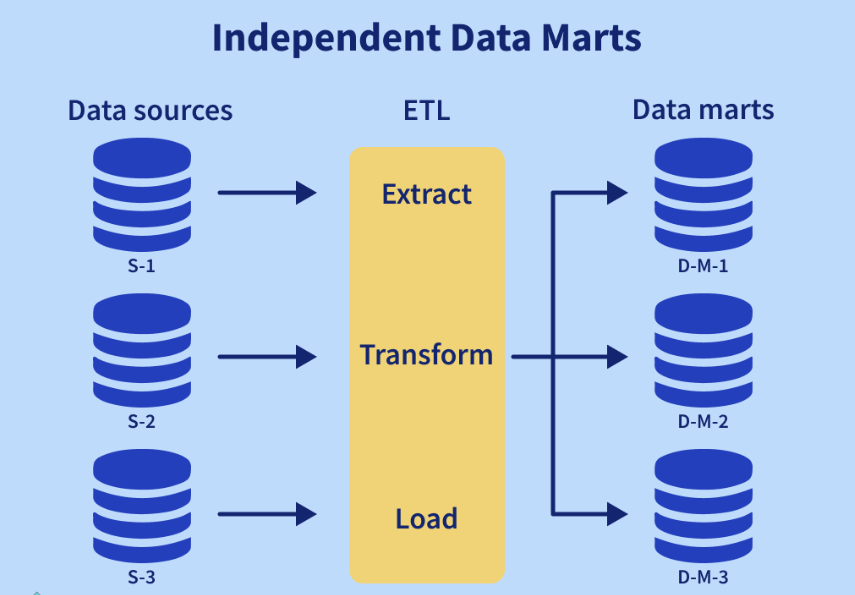
1. **Dependent Data Mart :**

* In this the Data Mart is built by fetching data from the central data warehouse that already exists.
* Means, it is depending on the data warehouse for data, that’s why it is called dependent data mart.



1. **Independent Data Mart :**

* In this the data mart is built by gathering data from operational or external sources of data or both.
* It is not dependent on any data warehouse, that’s why it is called independent data mart.



→ **Data Warehouse vs Data Marts**

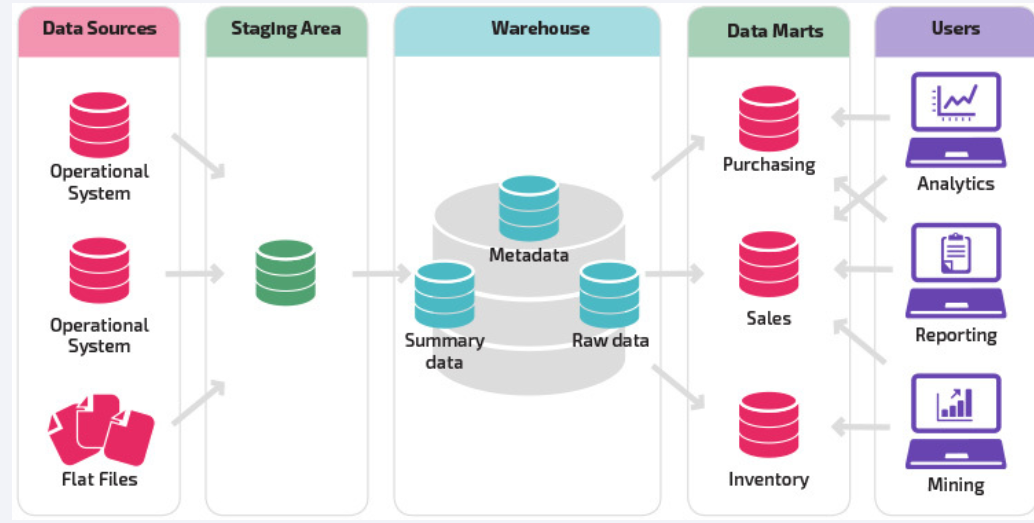
Both the terms sound similar, but they have a lot of differencesbetween them and different responsibilities and functionalities. So, here are some key differences between data warehouse and data marts.

**Data Warehouse :**

1. A data warehouse is built for an entire corporation or enterprise.
2. It is generally a union of all the data marts.
3. Normally, a data warehouse receives data from staging area.
4. In data warehouse, there is a centralized control and management of data.
5. A data warehouse keeps all the historic data even if it is not needed.

**Data Mart :**

1. A data mart is built for a particular depart of an organization or enterprise.
2. Unlike data warehouse, data mart is a single business process.
3. Whereas, a data mart receives data from **Star-join (facts & dimensions).**
4. A data mart is a optimal technology for data access and analysis of it.
5. Data mart is designed in such a way, that it suits the departmental view of data.



→ **Data Warehouse Development Life Cycle**

* A data warehouse is a place where the organizations can store information and data for a very long period of time. In this way, it become a library of historical data that can used for analysis purpose.
* Oftenly, new data is updated by the organization in different departments, such as marketing and sales.

**Components of Data Warehouse :**

1. Data Sources
2. Data Staging and ETL
3. Data Warehouse
4. Data Marts

**Data warehouse development life cycle :**

Data Warehousing is a flow process used to gather and handle structured and unstructured data from multiple sources into a centralized repository to operate actionable business decisions. With all of your data in one place, it becomes easier to perform analysis, reporting and discover meaningful insights at completely different combination levels.

**It consists of following steps :**

1. Requirement Specifications
2. Data Modelling
3. ELT Design & Development
4. OLAP Cubes
5. UI Development
6. Maintenance
7. Test and Deplyment
8. Back to 1

