Task: 02– by Bytewise Team

# **Data Warehouse vs Data Lake**

I will explain the difference between the above two simply in 3  steps:

1. Data Warehouse follows the philosophy of **Thinking first Loading later**. It means we need first to answer some questions like what purpose we need a data warehouse and what kind of data we need to store in it. As the data warehouse is Subject oriented. It is built for specific purposes. Mostly structured data is stored in a data warehouse using the ETL process.

When we are done with the thinking part then the loading part comes where we use the stored data in a data warehouse for multiple purposes like query or analysis.

1. Data lake on the other side follows the opposite philosophy **Loading first Thinking later**. Here we load data into a data lake that may be structured, unstructured, or semi-structured from multiple streams. After that, we start thinking about what we need to do with this metadata we have.

1. There are undoubtedly many more differences like data lake is more agile and less costly than a data warehouse and many more. Also, we can use both at the same time, like creating multiple data warehouses from a data lake.

For ease, you can think that a **Data Lake** is like an actual water lake, where water can be used for multiple purposes for drinking, agriculture, etc. and a **Data Warehouse** is like a house water tank, where its water can be used only for house chores. So you can also supply water from the lake to the water tank of a house if you want. These are just some differences to give an idea of both terms. I will end this here.

I will end the difference here!

# **What are Data Marts?**

They are a subset of data warehouses. If the data warehouse is the parent, then the data marts are the child of it.

**E.g**

Let's say you own a grocery store and you want to know which fruits are selling the most. You could go through all of your sales data, but that would be a lot of work! Instead, you could create a data mart that only contains sales data for fruits.

That way, you can easily see which fruits are selling the most without having to sort through all of your sales data.

In the same way, a data mart helps you analyze data by creating a smaller, more focused version of your larger data set. This makes it easier to find the information you need without having to search through a lot of irrelevant data.

# **What is Data Lakehouse?**

A data lakehouse is a relatively new concept in the world of data storage and analytics.It combines the features of a data warehouse and a data lake into a single platform, allowing for both structured and unstructured data to be stored, processed, and analyzed together.

In short, It enables companies to gain deeper insights into their data, leading to better business decisions and improved customer experiences.

**For Example —**

imagine you are a company that sells products online.

You have a website where customers can purchase products, and you also have social media accounts where you engage with customers. You have data on your sales, inventory, etc. stored in different systems and formats.

**In a traditional data warehouse…**

You would have to convert all of this data into a standardized format before it can be analyzed. This can be time-consuming and challenging, especially if the data is unstructured, such as social media comments.

**However, in a data lakehouse…**

You can store all of this data in its original format and still be able to analyze it. You can use tools like Spark to access and query the data, regardless of its format or source.

This allows you to gain deeper insights into customer behavior and preferences, and improve your business processes and customer experience.

# **OLTP vs OLAP**

OLTP stands for online transactional processing. It’s one of the two approaches to handling data in a business or organization

* OLTP is designed for handling day-to-day business transactions, such as processing orders, making payments, and managing inventory.
* OLTP databases are optimized for rapid updates, inserts, and deletes of individual records.
* OLTP systems focus on ensuring data integrity and maintaining consistency.
* OLTP databases are usually normalized, meaning that data is organized into multiple tables to avoid data duplication.

OLAP stands for online analytical processing.

* OLAP is designed for analyzing large amounts of data to identify trends and make strategic decisions.
* OLAP databases are optimized for fast querying and aggregation of data, typically through the use of multidimensional data cubes.
* OLAP systems allow for complex data analysis, including drill-down and roll-up functionality to explore data at different levels of detail.
* OLAP databases are usually denormalized, meaning that data is organized into fewer tables to improve query performance.

OLTP and OLAP serve different purposes, with OLTP handling day-to-day transaction processing and OLAP providing the tools for analyzing and understanding business data.

# **What is Data Mesh?**

Data Mesh is a way of organizing data so that it's easier for people in a company to find and use. It does this by treating data as a product. Each piece of data has an owner who is responsible for making sure it's accurate and useful. Also, each piece of data has a label that tells you who owns it and how you can use it.

**E.g** Imagine a company that sells shoes. In the past, if someone in the company wanted to know how many pairs of shoes were sold last month, they would have to ask someone in the IT department to run a report for them.

But with Data Mesh, the shoe sales data would be owned by the sales team, and they would make it available to everyone in the company through a data portal.

So if someone in marketing wanted to know how many pairs of blue shoes were sold last month, they could go to the data portal, find the shoe sales data, and run the report themselves without having to ask anyone in IT.