Name: Uzman Ali

Track: Data Engineering

Task: 01– by Bytewise Team

# What is a Data Warehouse?

In simple words, It's a copy of a data transaction that is specifically structured for query and analysis.

It has some characteristics like a Data Warehouse is:

* **Integrated:** It combines data coming from various streams
* **Time Variant:** Every data in a data warehouse is related to time analysis.
* **Non-Volatile (Static):** Whatever data comes into the data warehouse can’t be changed there are different mechanisms for that.
* **Subject-Oriented:** Building a Data Warehouse we need to keep in mind that it should be subject oriented. If you are building it for sales insight, then it should be designed in such a way that it is sales oriented or all data should be related to sales. The question here should be what exactly you want that for and then you start designing it.

And to end this topic it's an OLAP database different from OLTP databases.

# What is a Database?

In simple words, a Database is a collection of tables. Software systems that manage data models.

All the transactional data is stored in the system. In a corporate company suppose all customer sales transaction data is stored in a sales database and data like HR data is stored in a separate database and more. All these databases are made for faster writes because the objective here is to store data as fast as possible.

Then all the data from different operational databases will be pushed to the data warehouse for analysis as here the objective is faster reads.

It’s also called OLTP- Online transaction processes which is different from data warehouses called OLAP- Online analytical processes.

There are two types of databases: Relational and Non-Relational databases also known as SQL and NoSQL.

# What is a Data Lake?

The Data Lake philosophy is the opposite of the Data Warehouse. Mostly when we store data in a data warehouse it is structured data and as mentioned above it should be subject oriented.

It is called **Think First then Load!**

But in Data Lake we do the opposite **Load First then Think!**

We first load all the data which may be structure, un-structure, semi-structure, and logs. Then we think after that what should we do about that data, unlike Data Warehouse.

.

To make it simple I compare both Data Lake and Data Warehouse. If a data lake is a water lake that can be used for agriculture purposes or drinking etc. Then the data warehouse is the water tank of the house which is solemnly for chores of the house.

We can use both data lake and data warehouse at the same time. Firstly, corporations create data lakes and from it create different data warehouses for different purposes.

A data lake is more agile than compared to data warehouse and as well as less costly. We need to properly manage our metadata - meaning which data is coming and becoming part of the data lake. It should not be duplicated as other such issues.

# What is Big Data?

Big data refers to a large amount of data that is too complex and voluminous to be processed using traditional methods. It typically includes data from various sources, such as social media, sensors, transactions, and other sources. To understand Big Data Fundamentals you have to understand the 5 V’s of Big Data which are:

**Volume:** It’s defined as the amount of data that is processed or generated.

**Velocity**: Speed at which data is being collected/gathered and analyzed. It depends on the size of the data.

**Veracity:** Data continuously flows through multiple channels such as computer systems, social media, etc. While Volume and Velocity of data is an important factor that adds value to the business. Big Data also needs to process diverse data types collected from various sources. **Veracity or Validity** of Big Data is defined as the assurance of the quality or credibility of the data.

**Value:** Collecting data is of no use, data that ‘add’ value to the business is most important. So, Value amounts to how worthy data is to positively impact the company’s business.

**Variety:** It includes different types and sources of data, such as structured, unstructured, and streaming data.