**Day -1**

**Madhu Kalyani Gadi (04-12-2023)**

# Data Engineering

Data engineering is designing, constructing, testing, and maintaining architectures for collecting, storing, and retrieving data.

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

ETL involves extracting data from source systems, transforming it into a usable format, and loading it into a destination system.

**Data Classification**

Data has been classified into 3 types. They are

1. **Raw data** is the unprocessed and unorganized data that is collected directly from sources without any manipulation or analysis.
2. **Processed data** is data that has undergone some level of transformation or analysis to make it more organized, structured, and meaningful.
3. **Cooked data** describes data that has undergone extensive analysis, modeling, or transformation, often resulting in a final, refined form.

## Big Data

Big data enables organizations to analyze large and diverse datasets, extracting valuable insights for informed decision-making.

**Properties of Big Data:**

Big data is characterized by four main properties 4Vs

1. **Volume**- large amount of data
2. **Velocity-** high speed of data generation
3. **Variety-** diversity in data types and sources
4. **Veracity-** ensuring data accuracy and reliability

## Data Processing Methods:

1. **Batch Processing:**

Batch processing involves collecting, processing, and analyzing a fixed set or batch of data at once.



### Stream Processing:

It is designed for handling data in real-time as it is generated. Here the data is continuously ingesting, processing, and reacting to data as it flows through the system.

**Streaming Methods:**

* At Least Once
* At Most Once
* Exactly once

## Map Reduce:

Big data works on Map reduce which is Key-Value Pairing.

1. Organize the data into keys and values.
2. Sort by the key.
3. Combine the data with matching keys.
4. Repeat this until we find the final key-value.

Example:

* Hadoop
* Apache Spark
* Beam
* Samza
* Azure databricks

## Data Warehousing

Data Warehouse is like a collection of data which stores and manages data in large volumes for an efficient way of data analysis.

**Features of Data Warehouse**

* Subject-Oriented
* Time variant
* Non-volatile
* Integrated

## DSS (Decision Support System)

DSS helps to assess and solve everyday business questions as it uses the useful information of raw data, docs, personal knowledge or business models of an organization.

## DSS Architectural Styles

1. OLTP (Online Transaction Processing)
2. OLAP (Online Analytical Processing)

## OLTP

* It is a method which manages and records day to day transactions.
* It is used by traditional OS.

**Architecture of OLTP**

## Benefits of OLTP

## Simplicity and Efficiency

* Data Integrity
* Fast query processing

**Pitfalls of OLTP**

* It requires instant updates.
* The processed data after OLTP is not suitable for data analysis.
* To do any transaction, we need to join multiple tables by using join.

**Operational Data:**

Operational data refers to real-time and transactional information generated during day-to-day business operations.

**DSS data Vs Operational Data**

The difference between DSS and operational data lies in the following aspects

* Timespan
* Granularity
* Dimensions

## Data Stores

## Data stores contain two kinds of data. They are

1. Business data - comprises of both operational data and external data relevant to business
2. Business data model - it is a visual representation that illustrates how different data entities, their attributes, and relationships are structured and interconnected within an organization.

## Data Marts:

Data marts are customized subsets of a data warehouse, made to specific business needs or user groups within distinct organizational areas.

## Data marts Vs Data Warehouses

Data Warehouse:

* It is comprehensive and vast.
* It supports strategic decision-making with a centralized, unified view of data.

Data Mart:

* It is a subset of a data warehouse, focusing on specific business functions or user groups.
* It is to meet the needs of a particular department or business unit.