**Name -Monika Gariya**

**Email-** [**monikagariya2023@gmail.com**](mailto:monikagariya2023@gmail.com)

**Data Engineering Batch 1**

**Date – 17-01-2024**

**Topic – Data Engineering and Data Warehousing**

**Data Engineering** - Data engineering is a field of study and practice that involves designing, developing, and managing the systems and architecture for collecting, storing, processing, and analysing large volumes of data.

**Types of data**

1)Raw data **-** Raw data is an unprocessed data where no schema is applied

2)Processed data- Processed Data is raw data in which schema is applied

3)Cooked data- cooked data is a processed data that have been summarized.

**Big Data -** Big data describes large and diverse datasets that are huge in volume and also rapidly grow in size over time.

**Big data properties**

1) Volume - How much data you have

2) Velocity - How fast data is getting

3) Variety - How different your data format

4) Veracity - How reliable your data(imp)

**Data Processing Methods**

1. **Batch processing -** Batch processing refers to the processing of blocks of data that have already been stored over a period of time. For example, processing transactions that have been performed by a financial firm in a week.

**Example –** Billing and Invoicing, Scientific Research, image and video processing

1. **Stream processing -** Stream processing is a big data technology that allows us to process data in real-time as they arrive and detect conditions within a small period of time from the point of receiving the data. It allows us to feed data into analytics tools as soon as they get generated and get instant analytics results.

**Example** – Fraud Detection, Social media Monitoring, Healthcare Monitoring

**Data Warehousing**

Data Warehousing is a process of collecting, storing and managing large volume of data from different sources to support business intelligence and reporting activities.

**Features of Data Warehousing**

**1)Subject-oriented -** It focuses on providing a consistent view of information related to specific subjects, such as customers, products, or sales.

**2)Integrated** - Integration means combines data from several sources into a single, unified warehouse

**3)Time-variant** - It includes historical information and maintains a record of changes over time. The system captures and stores data at different points in time, allowing for trend analysis and historical comparisons.

**4)Non-volatile** - Once data is loaded into the data warehouse, it becomes non-volatile, meaning it is not changed or updated in response to every transaction

**Need for DSS (Decision Support System) for business**

Decision Support Systems helps us to assess and resolve everyday business questions.

It works by compiling useful information from a combination of raw data, documents, personal knowledge, or business models.

**Structured and Unstructured components of DSS**

1. **Structured Component -**This refers to well-organized, tabular data with a clear and fixed structure, often stored in relational databases. Structured data includes numerical values, dates, and categorical information organized in rows and columns.

1. **Unstructured Component -** This type of data lacks a predefined data model or does not fit neatly into traditional relational databases. Unstructured data includes text documents, emails, images, videos, social media posts, and other forms of non-tabular information.

**OLTP (Online transaction Processing)**

It is a class of systems that facilitate and manage transaction-oriented applications, typically for data entry and retrieval in real-time. OLTP systems are designed to handle a large volume of transactions, and they are optimized for efficient and quick processing of individual transactions.

**Example** – ATM (Automated Teller Machine), online booking, ticket and reservation system.

**Benefits of OLTP**

1) Simple and efficient

2) Maintain data integrity

**Pitfalls of OLTP**

1) Require instant updates

2) Not suitable for data analysis