By: Zia ur rehman

Date: 15-march-2023

**Task#2**

**Data Marts:**

Data marts are mini data warehouses that are designed to serve specific audience within an organization. They are typically focused on a specific subject area, such as sales, marketing, or finance, and contain only the data that is relevant to that area. They can be physically separate databases, or virtual data marts that draw data from a centralized data warehouse.

**Data Lakehouse:**

Data Lakehouse is a modern data architecture that combines the best features of data warehouse and data lakes. It provide platform for storing, processing, and analyzing structured and unstructured data at scale. However, unlike a data lake, a data lakehouse includes a processing layer that allows users to query the data directly without having to move it to a separate data warehouse. The processing layer in a data lakehouse is based on a modern compute engine, such as Apache Spark, that can handle both batch and real-time data processing.

**Data Mesh:**

Data mesh is a concept that aims to address the challenges of centralized data management within large organizations. It treats data as a product. The data ownership is decentralized and more domain oriented. Every domain is responsible for their own data. That means each domain team is responsible for the data they produce, including its quality, delivery, and consumption.

**DWH vs Data Lake:**

A data warehouse is a centralized repository of structured data that is organized to support business intelligence and analytics applications. In contrast, a data lake is a centralized repository of raw, unstructured, and semi-structured data, often in its native format. Data warehouses are optimized for querying and reporting, while data lakes are designed for data exploration and analysis.

**OLTP vs OLAP:**

OLTP (Online Transaction Processing) is a type of database system that is designed to support transactional processing applications. It stores data like recording customer orders, invoices etc. OLTP systems are typically used by businesses that need to process large volumes of transactions in real-time, such as e-commerce.

OLAP (Online Analytical Processing) is a type of database system that is designed to support complex analytical and reporting applications. It stores data like sales data, customer demographics etc. OLAP systems are typically used by businesses that need to perform complex analysis.

**Task#3**

**- Can a database be used as DWH?**

Yes a database can be used as a data warehouse. Many data warehouses are built using relational databases such as Oracle, SQL Server, and MySQL. However u need to enhance the database with such features that enable it to be used as a data warehouse. To build a database that can be used as a data warehouse, certain design considerations should be followed, such as having a schema design, be able to integrate data from multiple sources, have transformed data to have unification, and should be optimized for complex queries and analytics.

**- Major differences between structured and Un-structured data.**

Structured data refers to data that is organized in a specific format such as a database, spreadsheet, or table. Unstructured data, on the other hand, refers to data that does not have a predefined format or structure, and it can take various forms such as text, images, audio, and video files.

The major differences between structured and unstructured data include:

* Format and structure: Structured data has a defined format and structure, while unstructured data does not have a predefined format or structure.
* Processing and analysis: Structured data is easier to process and analyze because it follows a specific format and structure, while unstructured data requires more processing and analysis to derive meaningful insights.
* Storage: Structured data is usually stored in a database or spreadsheet, while unstructured data can be stored in various formats, including text files, audio files, and video files.
* Accuracy: Structured data is more accurate because it follows a specific model and its fields are clearly defined, while unstructured data may contain errors or inconsistencies.
* Volume: Unstructured data is typically much larger in volume than structured data because it includes a variety of sources such as social media, emails, and customer reviews.

**- What are the duties of a data engineer?**

A data engineer is responsible for designing, building, and maintaining data infrastructures.

Some of the key duties of a data engineer include:

* Data Pipeline Design: Designing and building data pipelines that move data from various sources into data storage systems for further processing.
* Data Storage Design: Designing and building data storage systems, such as databases or data warehouses, to store large volumes of structured and unstructured data.
* Data Integration: Integrating data from multiple sources and ensuring data quality and consistency.
* Data Processing: Developing data processing workflows and data transformations to prepare data for analysis.
* Data Modeling: Developing and implementing data models that facilitate data analysis and reporting.
* Data Security: Ensuring data security and compliance with data protection regulations.
* Data Architecture: Designing and maintaining the overall data architecture to ensure scalability, performance, and reliability.
* Performance Optimization: Optimizing data pipelines, storage systems, and processing workflows for improved performance and efficiency.