**HASNAIN AJMAL**

**Data Engineering**

**Week 1**

**Day 02: 15/02/23**

**TASK 02**

**Data Marts:**

A Data Mart is a subset of a data warehouse that is designed for a specific business unit or department in an organization. Data Marts are used to provide a focused view of the data that is relevant to a particular group of users. For example, a retail organization may have a data mart specifically for sales data, which can be used by the sales department to analyze sales trends, customer behavior, and other relevant information.

**Data Lakehouse:**

A Data Lakehouse is a modern data architecture that combines the best of both worlds from data warehouses and data lakes. It provides the benefits of a data lake, such as the ability to store raw and unstructured data at scale, with the structured querying and ACID-compliant transactions of a data warehouse. For example, a data lakehouse may store all of an organization's data in a raw format in a data lake, and then transform and optimize it into a structured format in a data warehouse for efficient querying and analysis.

**Data Mesh:**

Data Mesh is a decentralized approach to data architecture that emphasizes domain-driven distributed architecture. It involves breaking down monolithic data systems into smaller, domain-specific data products, each with their own teams and responsibilities. For example, a large retail organization may have a data mesh architecture where each product line has its own team responsible for the data products specific to that product line.

**DWH vs Data Lake:**

A Data Warehouse (DWH) is a centralized repository that aggregates data from various sources to provide a single source of truth for business reporting and analysis. A Data Lake, on the other hand, is a storage repository that allows organizations to store vast amounts of structured, semi-structured, and unstructured data at scale. The key difference between the two is that a data warehouse is designed for structured data and uses a schema-on-write approach, while a data lake is designed for unstructured and semi-structured data and uses a schema-on-read approach.

**OLTP vs OLAP:**

Online Transaction Processing (OLTP) is a system that supports transaction-oriented applications, such as order entry or banking systems, where the emphasis is on fast transaction processing and data integrity. On the other hand, Online Analytical Processing (OLAP) is a system that supports analytical applications, such as data mining and business intelligence, where the emphasis is on complex queries and analysis of large amounts of data. The key difference between the two is that OLTP systems are designed for transactional processing, while OLAP systems are designed for analytical processing.

**TASK 03**

**Can a database be used as Datawarehouse?**

Yes, a database can be used as a Data Warehouse, but it depends on the size and complexity of the data. Data Warehouses are designed to handle large amounts of structured and unstructured data from multiple sources, and provide fast querying and reporting capabilities. Traditional databases may not be able to handle the volume and variety of data required for a Data Warehouse, but with the right configuration, a database can be used as a smaller-scale Data Warehouse for smaller organizations or specific use cases.

**Major differences between structured and unstructured data.**

Structured data refers to data that is organized in a defined manner and follows a specific data model. This data is typically stored in a relational database and can be easily queried and analyzed using traditional methods. Examples of structured data include customer names, addresses, and transaction amounts.

Unstructured data, on the other hand, refers to data that has no predefined structure or organization. This type of data includes things like text, images, videos, and social media posts. Unstructured data is often stored in a data lake or other NoSQL databases and requires special tools and techniques to extract insights from it.

The major differences between structured and unstructured data include the organization and complexity of the data, the type of database required to store and process the data, and the tools and techniques required to analyze and extract insights from the data.

**What are the duties of a data engineer? (high-level)**

A data engineer is responsible for the design, implementation, and maintenance of an organization's data architecture. Their primary duties include:

* Designing and implementing data pipelines to move and transform data from various sources to the data warehouse or data lake.
* Building and maintaining scalable and efficient data infrastructure to store and process large volumes of data.
* Ensuring data quality and integrity through data validation, testing, and monitoring.
* Collaborating with data analysts, data scientists, and other stakeholders to ensure data is accessible and usable for business insights and decision-making.
* Staying up-to-date with the latest data technologies and tools, and making recommendations for improvements to the data architecture.