BYTEWISE FELLOWSHIP PROGRAM

DATA ENGINEERING

Tehreem Fatima

**WEEK 1**

**14 – March -2023**

***Task No 1***

***Big Data***

Big data means large amounts of data, that is in the form of digital information which is too complex or massive for traditional data processing techniques to handle efficiently. It is generally described by its large volume and variety of different data types. This data can come from different sources, such as social media posts, and transaction records, etc, and more. The challenges of big data include storing, processing, and analyzing it to find patterns, insights, and other useful information.

***Data Lake***

A data lake is a centralized repository that allows the storage of all structured and unstructured data at any scale in its raw form. It is designed to store and manage huge amounts of data including structured, semi-structured, and unstructured data from diverse sources, such as databases, social media, IoT devices, etc. The purpose of a data lake is to enable organizations to shore up data that traditional data sources cannot efficiently handle.

***Database***

A database is an organized collection of data stored electronically in a computer system. It is designed to provide efficient storage, retrieval, and manipulation of data. Generally, a database consists of tables, which store data in rows and columns, and a schema, which specifies the structure of the tables, the types of data that can be stored in each column, and the relationships between the tables. Databases can be used for a wide variety of applications, including managing customer data, inventory management, financial transactions, and much more.

***Data Warehouse***

A data warehouse is a type of data management system that is designed to enable and support Business Intelligence activities, especially analytics. It is a central repository of data that has been collected from various sources within an organization and is used for analysis and decision-making purposes. Data warehouses are typically used in enterprises and organizations that generate large amounts of data, such as financial institutions, healthcare organizations, and e-commerce companies.

**15 – March -2023**

***Task No 02***

***Data Marts***

Data marts are databases designed to serve the needs of a specific group of users within an organization. They are a type of data storage system used to accelerate business processes and provide quick access to relevant information. Data marts are a subset of a data warehouse that contains data that is specific to a particular line of business, department, or subject area.

**Data Lakehouse**

A data lake house is a modern data management architecture that combines the features and benefits of both data warehouses and data lakes. It kind of merges the ease of use and flexibility of a data lake with the data management and governance provided by a data warehouse. A data lake house can be defined as a big data storage platform built from a combination of a data lake and a data warehouse.

***Data Mesh***

Data Mesh is a decentralized data architecture that focuses on self-serve design to empower domain teams to have responsibility for their data products. The Data Mesh approach involves people, processes, and technology to organize data by specific business domains, such as marketing, sales, or customer service, rather than organizing data by technical architecture.

***Data Warehouse VS Data Lake***

Both data warehouses and data lakes are used for storing large amounts of data, but they differ in their structure, purpose, and usage.

Data warehouses are used to store structured and processed data that has been filtered and transformed for analysis and reporting purposes. Data in a warehouse is usually organized and optimized for quick analysis. Data warehouses often use a star schema for data modeling and may use SQL-based tools for querying and reporting.

Data lakes, on the other hand, are used to store large amounts of raw, unstructured, and semi-structured data. Data in a data lake is often stored in its native format, without any filtration or transformation. Data lakes are used to perform data discovery and other big data operations that require a lot of flexibility and scalability. Unlike data warehouses, data lakes do not enforce any structure or schema on the data and may use a variety of tools and technologies for querying and analysis.

In summary, the main difference between a data warehouse and a data lake is in their purpose and structure. Data warehouses are designed for structured reporting and analysis, while data lakes are designed for exploratory and big data operations. Data warehouses require structured data and a schema, while data lakes can accommodate a variety of unstructured and semi-structured data formats.

***OLTP VS OLAP***

OLTP - Online Transaction Processing and OLAP- Online Analytical Processing are two different approaches to managing and analyzing data.

OLTP is used for managing day-to-day business operations, such as recording transactions, order processing, or inventory management. It is a database system that supports high transaction output with many small transactions over small volumes of data.

OLAP is a type of database or processing system that is used for the complex analysis of large volumes of data. It is used to support decision-making processes by providing access to more detailed data. It is designed to handle many complex queries over large volumes of data.

In short, the key difference is that OLTP systems are optimized for quick, transactional processing, while OLAP systems are designed for complex analysis of large volumes of data.

***Task No 03***

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

Yes, a database can be used as a Data Warehouse DWH, but there are some differences between a database and a DWH.

A database is typically used for transactional processing, where each query is focused on specific data elements. On the other hand, a DWH is used for analysis and reporting, where queries are focused on the entire dataset and not just specific data elements.

DWH is designed to handle large volumes of data from multiple sources and combine them into a single repository for analysis and reporting. While a database can be used to store and access data for analysis, it may not be optimized for this purpose, making the queries slow. Therefore, it is better to use a dedicated DWH solution rather than a traditional database for data analysis and reporting needs.

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

Structured data is highly specific and is stored in a proper format whereas unstructured data is a collection of many types of data that are not organized in a specific way.

* Structured data is easier to search and analyze. It is organized in a specific way, whereas unstructured data is more difficult to search and analyze because it does not have a predefined structure.
* Structured data is typically stored in relational databases or other structured storage systems, while unstructured data is often stored in flat files, NoSQL databases, or other diverse forms.
* Structured data is generally easier to secure than unstructured data since it is stored in robust systems and is easy to keep track of and often in encryption format.

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

The duties of a data engineer at a high level include:

* Developing and maintaining data architectures
* Converting raw data into usable information for analysis and business decisions
* Aligning data architecture with business requirements
* Having a high-level understanding of data management technology and techniques
* Working with various systems that can handle large volumes of data
* Using languages like C/C# and Golang
* Training and implementing machine learning models

Overall data engineers are responsible for managing large volumes of data and making it accessible and usable for analysis and business decision-making. They require a combination of technical expertise in data management and software development as well as an understanding of business needs.