ByteWise Fellowship – Week 1

**Module: DE Basics [2]** Date: 15 March-2023

Question # 1

**What is Data Mart?**

A data mart is a subset of a larger data warehouse that is designed to serve the needs of a specific department or business unit within an organization.

**Characteristics:**

It is relevant to a particular group of users such as sales, marketing, or finance.

Data marts are typically designed to provide users with quick and easy access to the data they need for reporting, analysis, and decision-making.

**Advantage:**

Advantage of DM is that it can provide quick and relevant retrieval of data that is specifically targeted for that organization

Question # 2

**What is Data lakehouse?**

A data lakehouse is a unified and integrated data platform that allows organizations to store and process large volumes of data in a flexible, scalable, and cost-effective manner.

**Advantage of DataLake House:**

Data lakes are typically used for storing raw data in its native format, while data warehouses are used for storing structured, processed data that has been transformed and optimized for querying and analysis. The data lakehouse architecture seeks to bring the best of both worlds by providing a platform for storing raw data while also enabling users to perform real-time analysis and querying on that data without the need for additional processing or transformation.

**Tools:**

* Apache Hudi and Spark
* AWS Lake Formation

Question # 3

**What is DataMesh?**

Data Mesh is a relatively new concept in the field of data science that proposes a decentralized approach to data architecture and management within an organization.

It suggests that instead of having a centralized data platform and a team responsible for managing all data-related tasks, data should be treated as a product that is owned and managed by the domain teams who use it.

The main idea behind Data Mesh is to break down the silos of data ownership and management and shift the focus to empowering domain teams to take ownership of their data.

This means that each team is responsible for the quality, availability, and accessibility of their own data. They are also responsible for making sure that their data meets the needs of other teams within the organization who may want to use it.

Question # 4

**Difference between DWH and Data Lake?**

**1. Data Structure and Organization**

Data warehouses are designed to store structured data that is organized into predefined schemas. This data is typically sourced from transactional systems and transformed into a specific format that is optimized for analysis and reporting.

Data lakes, on the other hand, are designed to store both structured and unstructured data in its raw form. They do not require a predefined schema, allowing users to store all types of data in their original format.

**2. Data Storage and Processing**

Data warehouses store data in a highly structured and organized manner. They typically use a relational database management system (RDBMS) and follow a specific schema. The data is cleaned, transformed, and aggregated before being loaded into the warehouse, making it easier to query and analyze.

Data lakes, on the other hand, store data in a less structured manner. They use distributed file systems such as Hadoop or Amazon S3, and data can be stored in any format such as CSV, JSON, or Parquet. Data lakes can also leverage big data processing frameworks such as Apache Spark to perform data processing and analysis at scale.

**3. Data Usage and Accessibility**

Data warehouses are optimized for querying and analysis, with data structured and organized to support specific business use cases. Data stored in data warehouses is highly accessible to business intelligence tools, reporting tools, and other analytics applications.

Data lakes, on the other hand, are more flexible in terms of data usage and accessibility. Data can be accessed and analyzed by a wide range of tools, from traditional business intelligence tools to more advanced data science and machine learning platforms.

Question # 5

**Difference between OLTP and OLAP?**

OLTP (Online Transaction Processing) and OLAP (Online Analytical Processing) are two distinct categories of database processing systems with different objectives and characteristics.

**OLTP:**

OLTP is a type of system designed to handle real-time transactional data processing, such as recording sales, processing payments, and updating inventory levels. The primary goal of OLTP is to efficiently and accurately manage day-to-day business operations. OLTP systems are

optimized for processing a large volume of small transactions that are critical for supporting operational activities.

**OLAP**

On the other hand, OLAP is a type of system designed to support complex analysis of large volumes of historical data. The primary goal of OLAP is to provide fast access to summarized, multidimensional views of data for decision-making and strategic planning. OLAP systems are optimized for analyzing large amounts of data from multiple sources to identify trends, patterns, and relationships.

**Conclusion:**

To summarize, OLTP systems focus on transactional processing, while OLAP systems focus on analytical processing. OLTP systems are optimized for managing operational activities, while OLAP systems are optimized for supporting strategic decision-making