Name: Muhammad Usman

**Data Engineering Track**

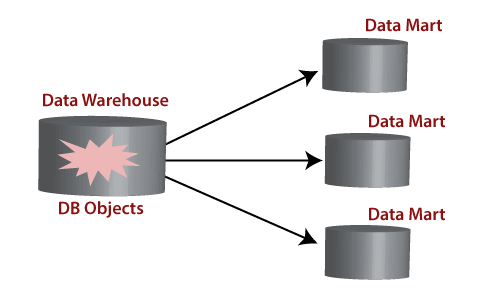
**Task 2**

**Data Marts:**

In data engineering, a data mart is a subset of a larger data warehouse that is designed to serve a particular department, team, or function within an organization. Data marts are typically smaller in scope and more focused on specific business needs than the larger data warehouse.

The purpose of a data mart is to provide easy and quick access to relevant data for the specific needs of the users. It can be thought of as a filtered or aggregated view of the data warehouse, optimized for a particular use case. Data marts may also include additional data that is not available in the data warehouse but is necessary for specific business requirements.

Data marts are often used in large organizations where there are multiple departments or teams with different data needs. By creating data marts, organizations can provide relevant data to each department while minimizing the risk of data duplication or inconsistency.



**Data Lakehouse:**

A data lakehouse is a modern data architecture that combines the benefits of both data lakes and data warehouses. It is a centralized repository that stores raw and processed data in its native format, making it more accessible for analysis and processing.

Data lakehouses are designed to handle both structured and unstructured data, allowing organizations to store and process large amounts of data from a variety of sources in a flexible and cost-effective manner.

One of the main benefits of a data lakehouse is that it enables organizations to perform real-time data processing and analytics on their data. This is achieved by leveraging technologies such as Apache Spark and Delta Lake, which allow for efficient and scalable processing of large datasets.

Another advantage of a data lakehouse is that it provides a single source of truth for all data, ensuring data consistency and reducing the risk of data duplication or inconsistencies.

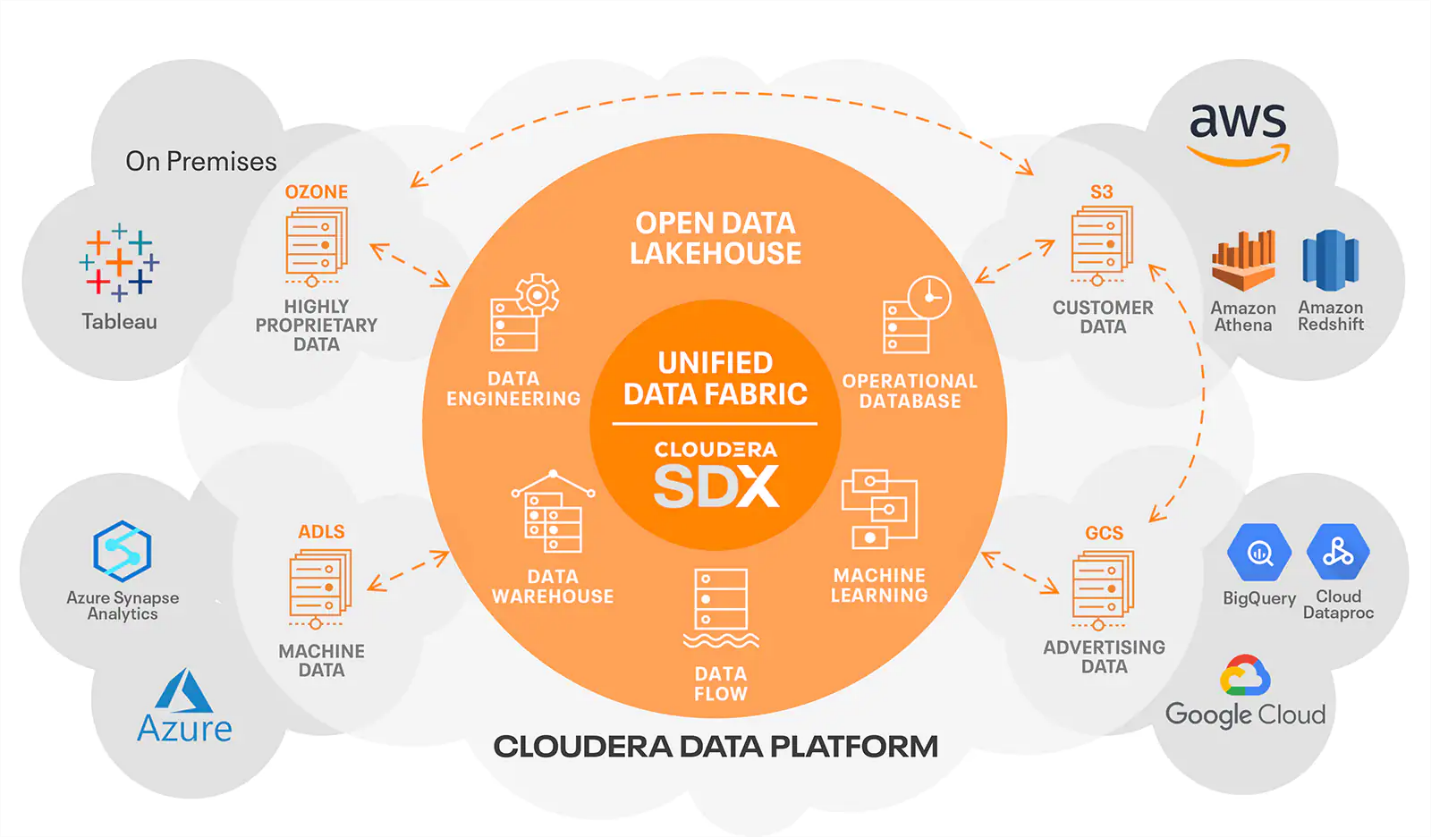
**Data Mesh:**

Data Mesh is a modern approach to managing data in large, complex organizations. It emphasizes the distribution of data ownership and management across various teams or domains, rather than centralizing data management in a single department.

In a Data Mesh architecture, each domain or team is responsible for managing their own data, including defining their data models, schemas, and APIs. This approach allows for greater autonomy and flexibility in managing data, while also ensuring that data is properly governed and can be easily integrated with other domains.

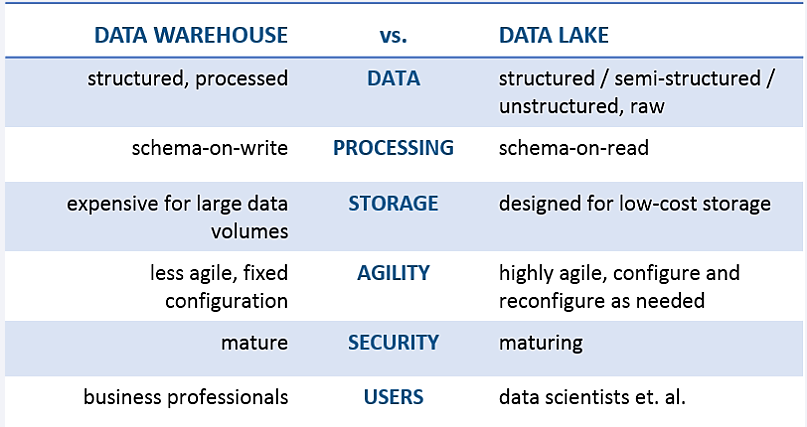
Data Mesh also emphasizes the use of data products, which are self-contained units of data that can be easily consumed by other domains. Data products are designed to be discoverable, reusable, and scalable, and can be managed using standard product management principles.

One of the main benefits of a Data Mesh architecture is that it allows for faster and more agile development of data-driven applications. By distributing data ownership and management across various teams, organizations can better align data management with business needs and reduce the time and effort required to integrate data from various sources.



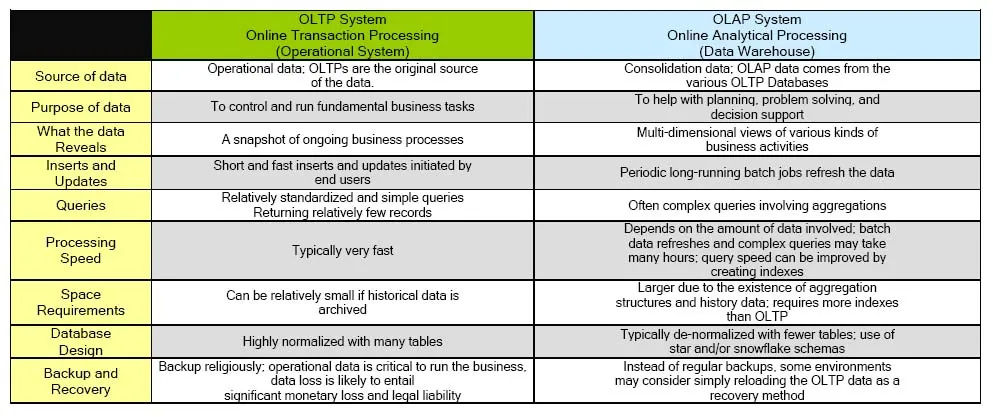
DWH VS Data Lake

While both DWHs (Data Warehouse) and data lakes are used for data storage and analysis, they differ in their architecture, purpose, and optimization. DWHs are optimized for querying structured data and supporting business intelligence, while data lakes are optimized for storing and processing raw data in any format, allowing for greater flexibility and exploration.



**OLTP VS OLAP:**

OLTP and OLAP are two different types of data processing used in data engineering. OLTP is designed for real-time transactional data processing, while OLAP is designed for data analysis and decision-making. OLTP is optimized for fast, reliable, and consistent data processing, while OLAP is optimized for complex queries and comprehensive analysis.



**Task#3**

**Questions:**

**QNO.1 Can a database be used as DWH?**

“A database can be used as a DWH if it is properly designed and optimized for DWH functions. However, a DWH requires more than just a database to be fully functional and efficient. It involves a process of ETL, schema design, and optimization for reporting and analysis.. However, a database used as a DWH will require additional design considerations, such as partitioning, indexing, and data compression, to ensure optimal query performance. The database schema will also need to be designed to support the reporting and analysis needs of the organization.”

**QNO.2 Major differences between structured and Un-structured data?**

| **Structured Data** | **Unstructured Data** |
| --- | --- |
| Organized in a predefined format with a fixed schema | Does not have a predefined format |
| Stored in a relational database | Typically stored in non-relational databases, such as document-oriented databases or object storage systems |
| Easily processed using SQL queries | Requires specialized tools and techniques, such as natural language processing (NLP) or machine learning, to extract insights and patterns |
| Represents a smaller volume of data | Can generate huge amounts of data that are difficult to store and analyze |
| Relatively simple and easy to understand | Can be complex and require more advanced tools and techniques to analyze |

As a data engineer, our main tasks involve managing and processing large volumes of data. We will:

* Sort and structure data so that it's easy to work with.
* Create efficient data pipelines that allow data to be moved between different systems.
* Understand what the business needs and work to create solutions that help achieve these goals.
* Use data analysis tools to understand trends and patterns within the data.
* Prepare data so that it can be used to create predictive models.
* Build and test different algorithms that can be used to gain insights from the data.