Name: Muqaddas Fatima

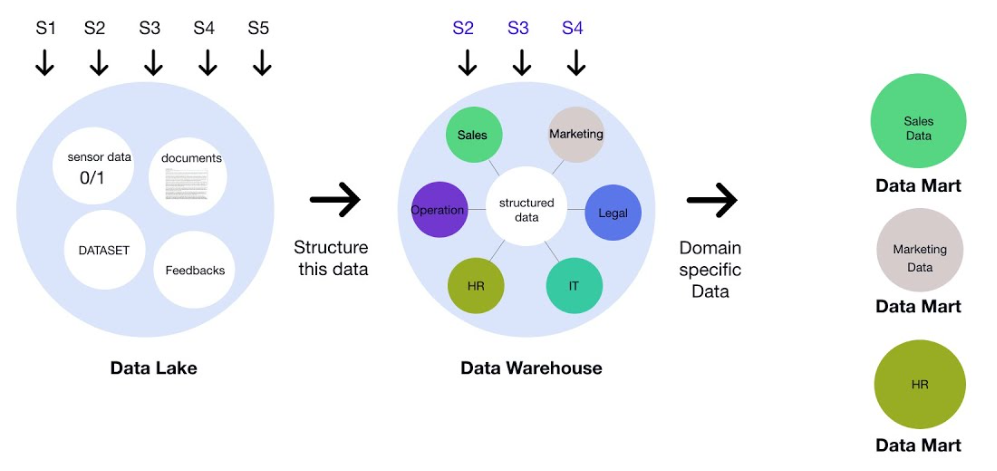
**Data Engineering Track**

**Task 2**

**Data Marts:**

A Data Mart is a subset of a Data Warehouse that is designed for a specific business unit or department. It contains a subset of the data stored in the Data Warehouse, organized in a way that is optimized for the needs of the specific unit or department.

* supplies subject-oriented data
* E.g. by limiting the data to a particular business unit (for example, the marketing department)
* Acts as a mini-data warehouse, holding aggregated data
* Data is limited in scope
* Often uses a star schema or similar structure
* Reports and dashboards use the data from the data mart



**Data Lakehouse:**

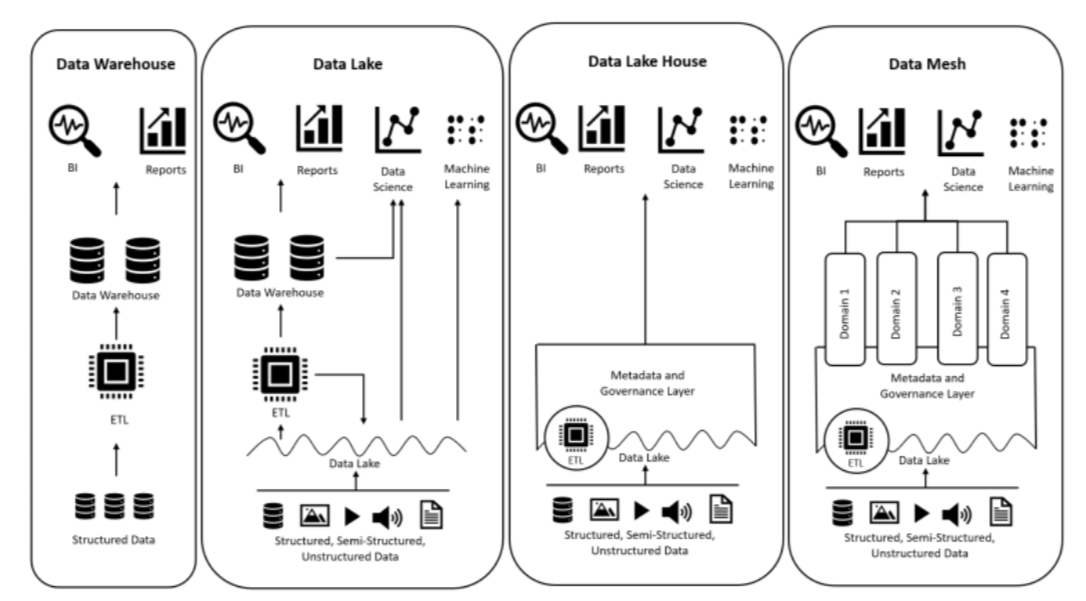
A Data Lakehouse is a modern data architecture that combines the benefits of a Data Warehouse and a Data Lake. It provides the scalability and flexibility of a Data Lake, while also offering the structured querying capabilities of a Data Warehouse.

* a single platform for storing, processing, and analyzing all types of data, including structured, semi-structured, and unstructured data.
* Data integration in a Data Lakehouse involves extracting data from various sources, transforming it into a format that is optimized for analytics, and loading it into the Data Lakehouse.
* increases scalability, flexibility, and agility in data processing and analysis. By providing a single platform for storing and processing all types of data, organizations can reduce data silos and improve the speed and accuracy of decision-making.
* particularly useful in organizations that have large amounts of data, or that require specialized data analysis for specific business functions.

**Data Mesh:**

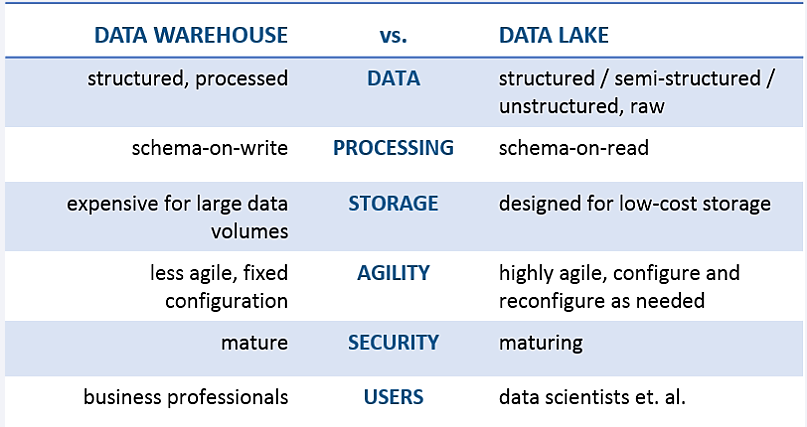
Data Mesh is a new approach to data architecture that emphasizes **decentralized data ownership** and management. It involves creating self-contained data domains, called data products, that are owned and managed by individual teams.

* encourages the creation of "data products," which are self-contained data artifacts that are designed to be consumed by other parts of the organization. By creating self-contained data products, domain teams can ensure that their data is accurate, consistent, and easily consumable.



**DWH vs Data Lake:**

The main difference between a Data Warehouse and a Data Lake is that a Data Warehouse is designed to store structured data from specific sources, while a Data Lake is designed to store a variety of structured, semi-structured, and unstructured data from various sources. A Data Warehouse typically uses a schema-on-write approach, while a Data Lake uses a schema-on-read approach.

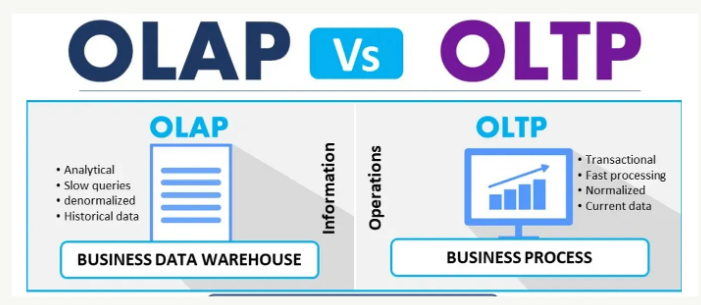


**OLTP vs OLAP**:

OLTP stands for Online Transaction Processing and refers to systems that are designed for processing high volumes of transactional data, such as sales or financial transactions. (for operational data)

OLAP stands for Online Analytical Processing and refers to systems that are designed for analyzing large volumes of data, typically for business intelligence and decision-making purposes. (for historical data)

OLAP systems often use a dimensional model, while OLTP systems typically use a normalized data model.



**TASK # 3**

Questions:

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

Yes, a database can be used as a Data Warehouse (DWH) if it is designed and optimized for efficient querying and analysis of large volumes of structured data. In fact, many Data Warehouses are built using a traditional relational database management system (RDBMS), such as Oracle, SQL Server, or MySQL

While a database can serve as a DWH, it is important to note that a DWH built on a database may not be as flexible or scalable as a Data Lake, which can store both structured and unstructured data in its native format. A database-based DWH may also require more upfront planning and design to ensure that it can handle the volume and complexity of the data being stored and analyzed.

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

Structured data is standardized, clearly defined, and searchable data, while unstructured data is usually stored in its native format.

Structured data is quantitative, while unstructured data is qualitative. Structured data is often stored in data warehouses, while unstructured data is stored in data lakes.

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

The duties of a data engineer are:

* Analyze and organize raw data.
* Build data systems and pipelines.
* Evaluate business needs and objectives.
* Interpret trends and patterns.
* Conduct complex data analysis and report on results.
* Prepare data for prescriptive and predictive modeling.
* Build algorithms and prototypes.