**Task #2**

**Data Mart**

A data mart is a type of database (more of a subset of Data Warehouse) that focuses on meeting the specific needs of a particular business unit or department within an organization. It is designed to support the reporting and analysis needs of a specific group of users, such as a marketing department or a sales team.

Like data warehouses, data marts store structured data that has been transformed and integrated from various sources. However, unlike data warehouses, data marts are designed to support the needs of a specific set of users, such as a marketing department or a sales team, and are optimized for their specific reporting and analysis needs.

Data Marts are usually used along with data warehouse. Data Warehouse provides the view of the whole organization (stores all the data in a central repository) while data mart stores all the data of a particular unit of the organization.

**Data Lakehouse**

A data lakehouse is a hybrid data storage that combines the benefits of a data warehouse and a data lake. It stores both structured data like Data Warehouse and also stores unstructured data like Data Lake. By doing this, it combines the benefits of both of these and allows you to process both kinds of data depending upon how you want to process it.

For example, you can build your Reporting dashboards using the structured data from Data Lakehouse and at the same time, you can also query the unstructured data for your Machine Learning algorithms and etc.

So, considering WAPDA as an example. You can query the bills details in structured format to do your reporting on dashboards, but at the same time you can also query images of the “meters” to do your machine learning algorithm training (like detecting the reading of the units in the meter)

**Data Mesh**

Data Mesh is a new way of organizing data within a company. Instead of having one big team manage all the data, each department has its own team responsible for their own data. These teams treat data like a product, and are responsible for making sure the data is high quality and easy to use.

For example, if a company sells products online, they might have a team responsible for managing the product data, another team responsible for managing customer data, and so on. Each team would make sure their data is easy to access and use for the other teams. This makes it easier for everyone in the company to use the data they need to make good decisions.

Data Mesh provides each department more control over their own data and this further helps in maintaining high quality of data within the organization. This high quality of data then can be used effectively for further analytical or Machine Learning purposes.

**Data Lake vs Data Warehouse**

1. Data Warehouse has structured data but Data Lake can store any kind of data.
2. Data Warehouse has a pre-defined schema, while Data Lake don’t have any schema (more like a NoSQL Database)
3. Data Warehouse is designed to store historical data that has been cleansed, formatted and is to be used for analytical purposes, while Data Lake is designed to store vast amounts of data in its original format, without any transformation or modeling.
4. Data Warehouse typically uses Relational Databases while Data Lake usually follows NoSQL Database structure.

**OLTP vs OLAP**

1. OLTP is optimized for transactional processing, while OLAP is optimized for analytical processing.
2. OLTP supports real-time, high-concurrency read-write operations on normalized databases, while OLAP supports complex queries on denormalized databases with aggregated data (for analytical purposes).
3. OLTP typically uses a normalized database structure while OLAP usually uses snowflake or star schema (denormalized database)
4. OLTP databases mostly follow row-level storage while OLAP mostly follows columnar storage for higher processing speed.

**Task #3**

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

From what I have analyzed by reading/watching different things, it depends upon the *amount of data* that you have in the database. If the data is not that large, then yes, you can use the Database as Data Warehouse for your analytical purposes (because querying your data wouldn’t take long)

However, if the data is large then it is best to go with a specialized Data Warehouse solution where data will be stored in snowflake/star schema following columnar structure best suited for fast querying and analytical purposes. So, if the data is large then no, an OLTP database wouldn’t be a good option to serve as a Data Warehouse.

**Structured Data vs Unstructured Data**

1. Structured data has a defined schema or format and can be easily understood (for example, sales data will always be in dollars) while unstructured data don’t have any defined format. This data is also more difficult for machines to understand.
2. Structured data is usually stored in Excel sheets or Relational databases, while unstructured data is usually stored in File Systems or NoSQL Databases.
3. Structured data can easily be analyzed or queried using SQL but unstructured is not that easy to search or analyze because of having no proper format.
4. Examples of structured data include Financial Records, Sales Records etc., while unstructured data can be image files, audio files etc.

**Role of Data Engineer**

1. The key responsibility of Data Engineer is to build the “ETL Pipelines” to extract data from the source databases (which can be multiple), transform it into usable format for the analytical or machine learning teams and store it in a place accessible to these teams (often called the Consumption Layer)
2. They are also responsible for designing data models that can support the requirements of analytics team.
3. They are responsible to ensure that data pipelines (ETL pipelines) they have built are scalable, efficient and also optimized for performance.
4. They are responsible to maintain the quality of the data during the ETL pipelines.
5. They are responsible for performance tuning and resolve bottlenecks between the data pipelines.