Task 2

Q1) What is data mart?

A data mart is a subset of a larger data warehouse that is designed to serve a specific business function or user group. It contains a focused set of data that is relevant to the needs of a particular department or business unit, and is optimized for querying and reporting on that specific data. Data marts are typically created by selecting a subset of data from the larger data warehouse and transforming it into a format that is optimized for the specific business needs of the user group. This may involve aggregating, summarizing, or filtering the data to provide a more streamlined and targeted view of the data.

Q2) What is data lakehouse?

A data lakehouse is a newer approach to data management that combines the best aspects of data lakes and data warehouses. It is designed to provide the scalability and flexibility of data lakes, while also providing the structured and governed approach of data warehouses. A data lakehouse typically uses a cloud-based architecture and allows for the storage of both structured and unstructured data in its native format, just like a data lake. This makes it easy to store and process large volumes of data, without the need for extensive data modeling or ETL processes. However, unlike a data lake, a data lakehouse also provides structured and governed access to the data, just like a data warehouse. This means that the data can be queried using SQL or other standard tools, and that data governance policies can be applied to ensure data quality, security, and compliance.

Q3) What is data mesh?

Data Mesh is a relatively new approach to data management that seeks to address some of the challenges and limitations of traditional centralized data management architectures. It is based on the idea of treating data as a product, and using a decentralized approach to data management that enables teams to take ownership of their own data domains. The basic idea behind Data Mesh is to break down large, monolithic data systems into smaller, more manageable pieces, or "domains". Each domain is owned and managed by a team of domain experts, who are responsible for defining the data schema, data quality standards, and access policies for their domain. To enable effective collaboration and communication between different domains, Data Mesh also emphasizes the use of standard interfaces and protocols, such as APIs, event streams, and metadata catalogs. This helps to ensure that data can be easily shared and integrated across different domains, while still allowing each domain to maintain autonomy and control over their own data.

Q4) Difference between datawarehouse and datalakes?

A data warehouse is a centralized repository of structured data that has been cleaned, transformed, and organized for efficient querying and analysis. It is designed to support business intelligence and reporting, and typically includes a schema that defines the relationships between different tables of data. Data warehouses often require extensive data modeling and ETL processes to ensure data quality and consistency, and they are optimized for performance and scalability.

A data lake, on the other hand, is a centralized repository of raw data that is stored in its native format, without any prior transformation or modeling. It is designed to store massive amounts of data, including structured, semi-structured, and unstructured data, and can be used to support a wide range of data analysis and processing tasks. Data lakes often require less upfront data modeling and ETL processes, and are optimized for flexibility and scalability.

Q5) Difference between OLTP and OLAP?

OLTP (Online Transaction Processing) systems are used for real-time transaction processing, such as recording sales or processing customer orders. They are designed to handle high volumes of simple, small transactions, and are optimized for fast data retrieval and update. OLTP systems typically have a normalized data schema that minimizes data redundancy and ensures data consistency and accuracy. They are often used in transactional systems, such as point-of-sale systems, online banking, and airline reservations.

OLAP (Online Analytical Processing) systems, on the other hand, are used for complex data analysis and reporting, such as sales trend analysis or customer segmentation. They are designed to handle large volumes of complex queries and aggregate data across multiple dimensions, such as time, geography, and product categories. OLAP systems typically have a denormalized data schema that allows for faster querying and analysis. They are often used in business intelligence and reporting systems, such as dashboards and scorecards.