TASK # 2 & 3

Data Mart:

A data mart is a subset of a data warehouse that is designed to serve a specific business unit or department within an organisation. It contains a subset of the data from the data warehouse that is relevant to the specific business unit, and is designed to support the analytical needs of that unit. Data marts are typically smaller and easier to manage than data warehouses, and can be designed to support specific business processes or functions.

Data Lake House: A data lake house is a combination of a data lake and a data warehouse. It is a centralised repository that stores all of an organisation's raw and structured data, but also provides tools and capabilities to transform and analyse the data in a more structured way, like a data warehouse. The idea behind a data lake house is to provide the flexibility and scalability of a data lake, with the structure and governance of a data warehouse.

Data Mesh: Data mesh is a new approach to data architecture that emphasises a decentralised and domain-driven approach to data management. In a data mesh architecture, each domain or business unit within an organisation is responsible for managing their own data, and data is treated as a product that is produced and consumed by different domains. The data mesh approach emphasises data ownership, data autonomy, and self-service data access, and encourages the use of APIs and microservices to make data more discoverable and usable across the organisation. The goal of data mesh is to make data more agile, scalable, and responsive to changing business needs, while reducing the burden on centralised data teams.

Data warehouses (DWH) and **Data Lakes** are two different approaches to storing and managing data in an organisation.

DATA WAREHOUSE:
A data warehouse is a **centralised repository** that **stores structured, cleaned, and transformed data** from different sources. The data is optimised for querying and analysis, and is typically **used for business intelligence and reporting** purposes. DWHs usually follow a strict schema design and are built for known and predictable data analysis requirements. They are **often based on relational databases** and require significant upfront planning, design, and development.

DATA LAKE:
A data lake is a **decentralised repository** that **stores raw, unstructured, and heterogeneous data** from different sources. Data lakes are **optimised for data exploration, data science, and big data** processing. Data in a data lake is **not structured,** and schema on read enables the data to be flexible and scalable. Data lakes are often based on distributed storage systems and do not require upfront planning, design, and development.

DATA WAREHOUSE	DATA LAKE
Data warehouses are optimised for structured, processed, and curated data.	Data lakes are optimised for raw, unstructured, and heterogeneous data
Data warehouses require upfront planning, design, and development.	Data lakes are more flexible and scalable, and do not require as much upfront work.
Data warehouses are typically used for business intelligence and reporting.	Data lakes are used for exploratory data analysis, machine learning, and advanced analytics.
Schema-on-Write	Schema-on Read

OLTP (**O**nline **T**ransaction **P**rocessing) and **OLAP** (**O**nline **A**nalytical **P**rocessing) are two different types of systems used for processing and managing data in an organisation.

OLTP is designed for **transactional** processing, which involves **collecting, storing, and processing data related to day-to-day business transactions**. OLTP systems are optimised for handling high volumes of transactions with low latency and high throughput. Examples of OLTP systems include banking systems, e-commerce systems, and inventory management systems. The data in an OLTP system is typically normalised and organised in a **relational database** structure with a focus on maintaining data integrity and consistency.

OLAP, on the other hand, is designed for **analytical** processing, which involves **aggregating**, **summarising, and analysing** large volumes of data for d**ecision-making purposes**. OLAP systems are **optimised** for **querying and analysis**, and typically involve data that has been transformed and aggregated from various OLTP systems. Examples of **OLAP systems include data warehouses and business intelligence systems.** The data in an OLAP system is typically **denormalized** and **organised in a dimensional structure, optimised for querying and reporting,** with a focus on providing fast response times and flexible analysis capabilities.

Here are some key differences between OLTP and OLAP systems:

OLTP systems are optimised for transaction processing	OLAP systems are optimised for analytical processing
OLTP systems are typically used to manage day-to-day business operations.	OLAP systems are used to support decision-making processes.
OLTP systems are designed to maintain data integrity and consistency.	OLAP systems prioritise fast response times and flexible analysis capabilities.
OLTP systems have a normalised relational database structure.	OLAP systems have a denormalized dimensional structure.
OLTP systems process individual transactions in real-time.	OLAP systems process aggregated data over time.

Q 1- Can a database be used as a data warehouse ?
Yes, a database can be used as a data warehouse (DWH) if it is designed and optimised for analytical processing.

Q2- What is the major difference between structured and unstructured data?
The major difference between structured and unstructured data is in their format and organisation:

Structured data is data that has a predefined format and is organised in a specific way. Structured data can be easily stored, processed, and analysed using traditional data processing tools such as relational databases and spreadsheets. Structured data is typically represented in tables with rows and columns, and each column has a defined data type.

Unstructured data, on the other hand, does not have a predefined format and is not organised in a specific way. Unstructured data can be much more difficult to store, process, and analyse than structured data, because it can include a wide variety of file types, such as text, images, audio, and video.

Q3- What are the duties of a data engineer?
The duties of a data engineer can vary depending on the organisation and industry, but here are some common responsibilities:

Data Architecture: Designing, building, and maintaining the data architecture for an organisation, including data models, data pipelines, data warehouses, and data lakes.

Data Integration: Integrating data from various sources into the data architecture, including APIs, databases, third-party systems, and IoT devices.

Data Pipeline Development: Developing and maintaining data pipelines that move data from source systems to the data warehouse or data lake, ensuring that the data is clean, accurate, and up-to-date.

Data Transformation: Developing and implementing data transformation processes, such as data cleaning, data normalisation, and data enrichment, to ensure that the data is in a usable format for analysis.

Data Quality: Ensuring the quality and accuracy of the data through data profiling, data validation, and data cleansing.

Data Security: Ensuring the security and privacy of the data through data encryption, access control, and data masking.

Data Monitoring: Monitoring data pipelines, data flows, and data storage systems to ensure that data is being processed and stored correctly.