**Basics of Data Engineering 2**

**Submitted by:** Afraz Ahmed

* **Data Mart:**

A data mart is a subset of a larger data warehouse that is designed to serve a specific business unit or department. Data marts are smaller, more focused versions of data warehouses that contain data relevant to a particular business function, such as sales, marketing, or finance. They are typically designed to be easier to use and more accessible than a full data warehouse, with data that is optimized for specific types of queries. Data marts are created by selecting and filtering data from the larger data warehouse and organizing it in a way that is most useful for the specific business unit or department.

* **Data Lakehouse:**

A data lakehouse is a hybrid data architecture that combines the benefits of a data lake and a data warehouse. It is designed to provide a single platform for managing and analyzing large volumes of structured and unstructured data, and to support both traditional BI (business intelligence) and advanced analytics. Like a data lake, a data lakehouse stores data in its raw, unprocessed form, without any predefined schema.

* **Data mesh:**

Data mesh is a relatively new architectural approach to data management that is designed to improve the scalability, flexibility, and agility of data management in large organizations. The core idea behind data mesh is to shift the focus of data management from a centralized model, where a single team manages and governs all data, to a decentralized model, where individual teams are responsible for their own data domains. In this approach, each team owns and governs its own data, and data is treated as a product that is designed, built, and maintained by the team that uses it.

* **DWH vs Data Lake:**

Data warehouses and data lakes are two different approaches to storing and managing data. A data warehouse is a centralized repository of structured data that is designed to support business intelligence activities, such as reporting and analysis. Data warehouses are optimized for query processing and analysis, and they typically use a schema-on-write approach, where data is transformed and loaded into the warehouse in a structured format. Data warehouses are often designed around specific business domains, such as sales, marketing, or finance, and they are managed and governed by a centralized team.

A data lake, on the other hand, is a centralized repository of both structured and unstructured data that is designed to support a wide range of data analytics and processing activities. Data lakes are optimized for storage and processing of large volumes of data, and they typically use a schema-on-read approach, where data is stored in its raw, untransformed form, and is transformed and processed as needed. Data lakes are often used for data exploration, data science, and machine learning, and they are typically managed and governed by a decentralized team.

* **OLAP vs OLTP:**

OLTP (Online Transaction Processing) is a type of system that is designed for processing transactions in real-time. OLAP (Online Analytical Processing), on the other hand, is a type of system that is designed for performing complex analytical queries on large volumes of data. The main difference between OLTP and OLAP is the type of processing they are optimized for. OLTP systems are optimized for transaction processing, while OLAP systems are optimized for data analysis and reporting. OLTP systems typically use a row-oriented storage model, where data is stored and accessed one row at a time, while OLAP systems typically use a column-oriented storage model, where data is stored and accessed by columns.

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

Yes, a database can be used as a data warehouse, although it may require some additional design and optimization to be effective. To use a database as a data warehouse, you would typically need to design and optimize the database for analytical queries. This might involve creating specific indexes or data structures that are optimized for analytical queries and designing the schema of the database to support analytical queries.

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

Structured data is organized in a predefined format that is easy to process and analyze. Structured data is typically stored in a tabular format, such as a spreadsheet or database, where each data element is represented in a consistent format and is associated with a specific field or column. Structured data is typically easy to search, sort, and analyze.

Structured data is organized in a predefined format that is easy to process and analyze. Structured data is typically stored in a tabular format, such as a spreadsheet or database, where each data element is represented in a consistent format and is associated with a specific field or column. Structured data is typically easy to search, sort, and analyze.

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

1. Designing and building data pipelines
2. Data modeling and database design
3. Data quality management
4. Data governance and security
5. Collaborating with data analysts and scientists