DATA ENGINEERING BASICS & QA

MARCH 15, 2023

BYTEWISE LIMITED

Contents

Task 2:

1. Data Marts

- Data marts make specific data available to a defined group of users, which allows those users to quickly access critical insights without wasting time searching through an entire data warehouse.
- A data mart (as noted above) is a focused version of a data warehouse that contains a smaller subset of data important to and needed by a single team or a select group of users within an organization.
- A data mart is built from an existing data warehouse (or other data sources) through a complex procedure that involves multiple technologies and tools to design and construct a physical database, populate it with data, and set up intricate access and management protocols.
- A data mart is a simple form of data warehouse focused on a single subject or line of business.

2. Data Lakehouse

- A data lakehouse is a new, open data management architecture that combines the flexibility,
- A data lakehouse starts with a data lake architecture, and attempts to add data warehouse capabilities to it, generally with modifications to the query engine and the addition of a predefined file format.
- Separation of storage and compute
- Unlimited scale data repository
- Mixed data types: structured, semi-structured and unstructured
- Choice of languages for processing (but not always SQL)
- No need to inventory or ingest data
- Direct access to source data

3. DWH vs. Data Lake

Data Lake	DWH
The reason for storing data is undefined	There is predefined reason for storing data
Data is left raw until needed	Data is processed and ready to be queried
Used by data scientists	Used by data professionals
There is no predefined schema so it stores data in native format	Schema of data warehouse is structured and defined before storage
Schema on read .only data read during processing is parsed and adapted into schema as needed.	Schema is applied while writing data
Stores structured and unstructured data	Stores structured data

4. Data Mesh

- **A** data mesh is a type of data platform architecture that embraces the ubiquity of data in the enterprise by leveraging a domain-oriented, self-serve design.
- A data mesh architecture is a decentralized approach that enables domain teams to perform cross-domain data analysis on their own.
- A data mesh is a decentralized data architecture that organizes data by a specific business domain—for example, marketing, sales, customer service, and more—providing more ownership to the producers of a given dataset.

5. OLTP vs OLAP

OLTP	OLAP
Handles a large number of small transactions	Handles large volumes of data with complex queries
Simple standardized queries	Complex queries
Based on INSERT, UPDATE, DELETE commands	Based on SELECT commands to aggregate data for reporting
Industry-specific, such as retail, manufacturing, or banking	Subject-specific, such as sales, inventory, or marketing
Short, fast updates initiated by user	Data periodically refreshed with scheduled, long-running batch jobs

Task 3:

1. Can a database be used as DWH?

Yes,

- A database can be used as a data warehouse (DWH) with the proper design and configuration.

- A data warehouse is a large-scale data repository that stores data from different sources and allows for complex data analysis and reporting.

- A database can be used as a DWH by following certain design principles and implementing features such as data partitioning, indexing, and optimization for reporting and analysis queries.

- However, it's important to note that a database used as a DWH may require different configuration and management strategies than a transactional database, and may also require specific data integration and transformation processes to prepare the data for reporting and analysis.

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

Structured	Un-structured data
Structured data is organized in a predefined manner with a specific data model.	Unstructured data lacks a predefined structure or data model.

Structured data is usually stored in a tabular format with defined fields, rows, and columns,	While unstructured data can take many different formats, such as text, images, videos, and audio.
Structured data is usually generated in smaller volumes.	Unstructured data is often generated in large volumes.
Structured data can be easily processed and analyzed using traditional database technologies,	unstructured data requires advanced technologies such as natural language processing and machine learning to extract insights
Structured data can be analyzed using SQL queries or other structured query languages,	Unstructured data requires specialized tools and techniques to analyze.

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

- Data Modeling: Designing and implementing data models that support the organization's data storage, retrieval, and analysis needs.

- Data Pipeline Development: Building data pipelines that move and transform data from various sources to the data storage system, such as a data warehouse or data lake.

- Data Integration: Integrating data from various sources, including databases, APIs, and third-party applications, into the organization's data ecosystem.

- Data Quality Assurance: Ensuring that the data collected is accurate, consistent, and reliable by performing data cleaning, validation, and auditing.

- Data Security: Ensuring that the organization's data is secure and protected from unauthorized access or data breaches.

- Performance Optimization: Optimizing the performance of data systems, such as databases and data pipelines, to ensure that they can handle large volumes of data and queries.

- Collaboration: Collaborating with other teams, such as data scientists and analysts, to ensure that their data needs are met and that they have access to the necessary data.