Name: Moiz Zulfiqar

Task 3

Can a database be used as DWH?

Yes, a database can be used as a Data Warehouse (DWH). In fact, a data warehouse is a type of database that is specifically designed to support business intelligence activities, such as data analysis, data mining, and reporting. A data warehouse is usually a large, centralized repository that stores data from a variety of sources, such as operational systems, external data sources, and legacy systems.

Some popular databases that are used as data warehouses include Oracle, Microsoft SQL Server, and Teradata. These databases have features that are specifically designed to support data warehousing, such as partitioning, indexing, and query optimization. However, it's important to note that not all databases are suitable for data warehousing, and it's important to choose a database that is optimized for the specific requirements of your data warehousing environment.

Major differences between structured and Un-structured data.

The following elements differentiate structured and unstructured data.

Formats

- Structured data is in the form of numbers and text, presented in standardized, readable formats like XML and CSV.
- Unstructured data comes in various shapes and sizes, does not conform to a predefined data model, and stays in the native formats, such as video and audio files.

Data Model

- Structured data follows a predefined relational data model describing the relationship of data elements.
- Unstructured data does not have a set data model but can have a hidden structure.

Storage

- Organizations store structured data in relational databases, while they store unstructured data in raw formats, not in databases.
- Data warehouses help centralize large volumes of stored structured data from different databases, while data lakes can store large amounts of unstructured data.

Database Type

- Structured data typically resides in a relational database, arranged in tables with rows and columns.

- Unstructured data often resides in a non-relational (NoSQL) database, which stores multiple data models without tables, such as document, wide-column, graph, and key-volume databases.

Searchability and Ease of Use

- Structured data is usually easier to search and use, while unstructured data involves more complex search and analysis.
- Structured data is older, so there are more analytics tools available, while standard data mining solutions cannot handle unstructured data.

Quantitative vs. Qualitative

- Structured data is quantitative, meaning that it has countable elements and is easier to analyze by classifying items based on common characteristics, investigating the relationships between variables, or clustering the data into attribute-based groups.
- Unstructured data is qualitative, meaning the information it contains is subjective, and traditional analytics tools and methods can't handle it. Techniques include splitting and stacking data volumes into logical groupings, data mining, and pattern detection.

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

The duties of a data engineer generally involve designing, building, and maintaining the infrastructure required for data storage, processing, and analysis.

High-level responsibilities of a data engineer:

1. Data Architecture Design: Develop and maintain data architectures, including data models, database schemas, and data processing workflows.
2. Data Pipeline Development: Build and maintain data pipelines to move data between systems, ensure data quality, and handle data transformations.
3. Database Management: Configure, deploy, and maintain databases that store and manage large amounts of structured and unstructured data.
4. Data Warehousing: Develop and maintain data warehouses that consolidate data from different sources and provide a single source of truth for reporting and analysis.
5. Data Quality Assurance: Ensure data is accurate, complete, and consistent across systems by developing and implementing data quality checks and monitoring processes.
6. Data Security and Privacy: Implement and maintain data security and privacy policies, including access control, encryption, and anonymization.
7. Performance Optimization: Optimize data processing and storage systems to improve query performance, reduce latency, and increase scalability.
8. Collaboration: Work closely with cross-functional teams, including data scientists, analysts, and business stakeholders, to understand their data needs and provide solutions to their data-related problems.
9. Emerging Technologies: Keep up-to-date with emerging technologies and data management tools to continuously improve data engineering practices and systems.

Overall, a data engineer plays a crucial role in ensuring that data is accessible, reliable, and usable by the organization.

References

- Structured VS Unstructured Data: https://www.imperva.com/learn/data-security/structured-and-unstructured-data/#:~:text=Structured%20data%20is%20usually%20easier,are%20more%20analytics%20tools%20available.