Name: Moiz Zulfiqar

Task 2

Data Mart

Data mart is a simplified form of data warehouse that is designed to focus on a particular line of business or subject area. It provides easy access to data required by specific teams or business units within an organization, enabling them to gain insights faster without having to search within a more complex data warehouse or manually aggregate data from various sources.

Data marts are created to eliminate the need for teams to rely on spreadsheets to share data, which can result in errors, confusion, complex reconciliations, and multiple sources of truth. By providing a centralized location for data collection and organization, data marts enable the creation of reports, dashboards, and visualizations that can be used to make informed decisions.

The benefits of data mart include a single source of truth, quicker access to data, faster insights leading to faster decision-making, simpler and faster implementation, and creating agile and scalable data management. It enables analysts to focus on specific challenges and opportunities, leading to better and faster decision-making.

Data mart is particularly useful for short-lived data analytics projects. It allows teams to set up a data mart rapidly and efficiently, improving productivity for both business and IT teams.

Data Lakehouse

Data lakehouse is a modern data platform that combines the features of a data warehouse and a data lake. The integration of these two tools brings together the flexible storage of unstructured data from a data lake and the management features and tools from data warehouses, resulting in a more effective solution for today's digital world. A data lakehouse offers various features, including data management features, open storage formats, flexible storage, support for streaming, and diverse workloads.

A data lakehouse can be used to streamline the overall data management process, breaking down the silo walls between multiple repositories and creating a much more efficient end-to-end process over curated data sources. This integration creates several benefits, including less administration, better data governance, simplified standards, and increased cost-effectiveness.

In conclusion, the concept of a data lakehouse is an innovative approach to data management that combines the strengths of a data lake and a data warehouse. It offers an excellent solution for organizations looking to consolidate and streamline their data management processes, and it provides several advantages, such as simplified administration, improved data governance, simplified standards, and cost-effectiveness.

Data Mesh

Data mesh is a decentralized data architecture that organizes data by specific business domains, enabling more ownership to be given to data producers of a particular dataset. Under this approach, data governance policies focus on documentation, quality, and access. Data mesh architecture shifts the use of traditional storage systems, such as data lakes or data warehouses, from a single, centralized data platform to multiple decentralized data repositories. It also promotes the adoption of cloud-native and cloud platform technologies to scale and achieve the goals of data management.

Data mesh requires a cultural shift in the way companies think about data. Instead of data being a by-product of a process, it becomes the product, where data producers act as data product owners. Under a data mesh model, domain teams become responsible for their ETL data pipelines, while a centralized data governance team enforces the standards and procedures around the data. Data mesh architecture enables data to be treated as a product that can be accessed by users across the organization, allowing for more flexible data integration and interoperable functionality.

The benefits of data mesh include data democratization, cost efficiencies, less technical debt, interoperability, and security and compliance. By making data more discoverable and accessible via this domain-driven design, it reduces data silos and operational bottlenecks, enabling faster decision-making and freeing up technical users to prioritize tasks that better utilize their skillsets. Data mesh promotes the adoption of cloud data platforms and streaming pipelines to collect data in real-time, providing cost advantages by allowing data teams to spin up large clusters as needed, paying only for the storage specified. It also reduces technical debt by distributing the data pipeline by domain ownership, reducing technical strains on the storage system.

Data mesh is suitable for larger organizations with complex enterprise data needs. It is particularly helpful in scaling data needs across an organization, enabling self-service use. While distributed data mesh architectures are still gaining adoption, they are helping teams attain their goals of scalability for common big data use cases such as business intelligence dashboards.

DWH VS Data Lake

Data lakes and data warehouses are both solutions for storing and managing data, but they have different purposes, data structures, users, costs, accessibility/agility, and security.

A data warehouse is designed to store and manage structured, refined data that is processed for a specific purpose, such as log and event management, sales reporting, or security analysis. In contrast, a data lake is a repository for any type of raw data, including unstructured, structured, and semi-structured data, which can be retained for future use. This approach has longer-term hazards in terms of cost and sustainability of storage because only a small percentage of the collected data is actually used and applied.

Data warehouses use a schema-on-write approach for structured data, while data lakes use a schema-on-read method for unprocessed data. Data warehouses are typically set up and

interpreted by data analysts or business analysts, while data lakes require the specialist knowledge of data scientists or data engineers to interpret and organize unprocessed data.

Data lakes are more cost-effective than data warehouses because they can store large amounts of any type of data without adhering to a fixed schema. However, structured data in a data warehouse can be analyzed more quickly and easily than data in a data lake.

Data lakes are more agile and flexible than data warehouses because they allow data to be added and stored more easily, and they enable tools for big data analytics. Data warehouses have a specific structure and are more difficult to alter. They typically have a read-only format which analysts can scan to garner insights from historical, clean data.

Data lakes are inherently less secure than data warehouses because they store petabytes of information and have a lack of selectivity on the data stored. However, big data security measures are rapidly evolving, and it is likely that data lakes will eventually become more secure.

In summary, data warehouses are better suited for structured data analysis, while data lakes are better suited for storing and processing large amounts of unstructured data. Ultimately, the choice between a data warehouse and a data lake depends on the specific needs and goals of an organization.

OLTP VS OLAP

OLAP and OLTP are two distinct systems used in data engineering with different objectives and processing types. OLAP (online analytical processing) is designed for complex data analysis, business intelligence, and decision-making, while OLTP (online transactional processing) is optimized for real-time execution of large numbers of database transactions.

The core of most OLAP databases is the OLAP cube, which allows you to quickly query, report on and analyze multidimensional data. OLAP systems use a multi-dimensional schema and support complex queries of multiple data facts from current and historical data. On the other hand, OLTP systems use a traditional DBMS and are designed for fast processing of simple insertions, updates, and deletions in databases.

Other key differences between OLAP and OLTP include their focus, data source, processing time, and availability. OLAP systems are ideal for making complex queries involving large numbers of records, while OLTP systems are designed for making simple read and write operations via SQL. OLAP systems require less frequent backups since they don't modify current data, while OLTP systems require frequent or concurrent backups to maintain data integrity.

Choosing the right system for a particular situation depends on the objectives of the organization. OLAP can help unlock value from vast amounts of data for business insights, while OLTP is designed for managing daily transactions. Most organizations use both OLAP and OLTP systems, and OLAP systems may be used to analyze data that leads to business process improvements in OLTP systems.

References

- Data Mart: https://www.oracle.com/pk/autonomous-database/what-is-data-mart/
- Data Lakehouse: https://www.oracle.com/data-lakehouse/what-is-data-lakehouse/
- Data Mesh: https://www.ibm.com/topics/data-mesh
- DWH VS Data Lake: https://www.splunk.com/en_us/blog/learn/data-warehouse-vs-data-lake.html
- OLTP VS OLAP: https://www.ibm.com/cloud/blog/olap-vs-oltp