BASICS OF DATA ENGINEERING [2] + QA

SUBMITTED BY: MUHAMMAD FAHAD

TASKS DAY 2 [15/03/2023]

Task # 2: Familiarize yourself with the following topics:-

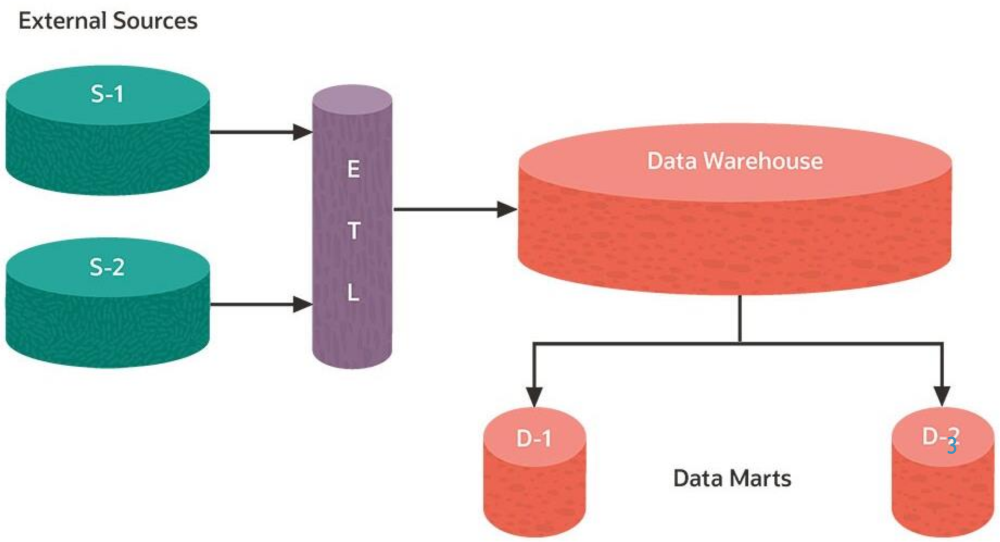
- Data Marts
- Data Lakehouse
- Data Mesh
- DWH vs Data Lake
- OLTP vs OLAP

Task # 3: After you complete these topics, please answer the following questions in your document:-

- Can a database be used as DWH?
- Major differences between structured and Un-structured data.
- What are the duties of a data engineer? (high-level)

TASK#02
DATA MART

A data mart is a data storage system that contains information specific to an organization's business unit. It contains a small and selected part of the data that the company stores in a larger storage system. Companies use a data mart to analyze department specific information more efficiently. It provides summarized data that key stakeholders can use to quickly make informed decisions.



TYPES OF DATA MART

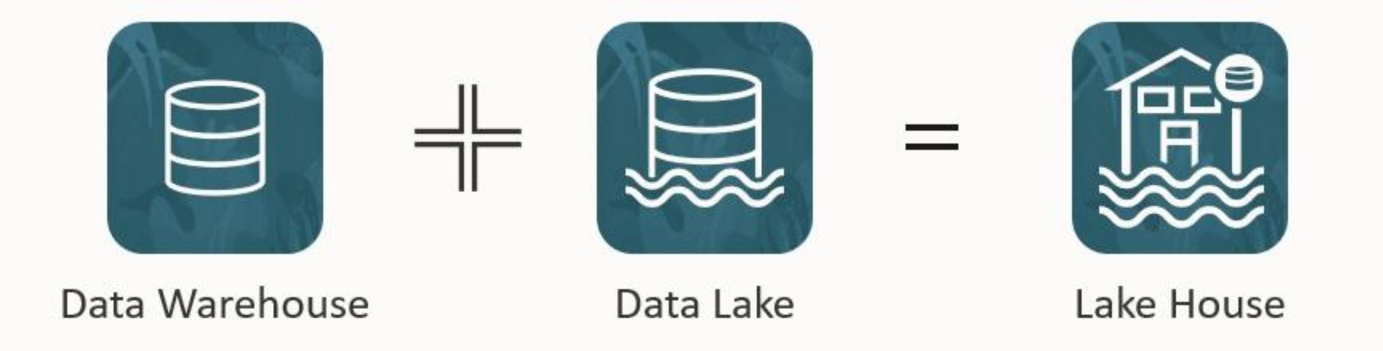
- Dependent Data Mart: Dependent data marts are created by drawing data directly from operational, external or both sources.

- Independent Data Mart: Independent data mart is created without the use of a central data warehouse.

- Hybrid Data Mart: This type of data marts can take data from data warehouses or operational systems.

DATA LAKE HOUSE

A data lake house is a data management architecture that combines the benefits of data lakes and data warehouses. It is a hybrid approach that allows organizations to store, manage, and analyze both structured and unstructured data in a single system. Data lakes traditionally store raw data in its native format, which allows for fast and flexible data processing. However, this approach lacks the structure needed for analytical queries, making it difficult to derive meaningful insights. On the other hand, data warehouses provide a more structured approach to data management and are optimized for analytical queries. However, they are not designed to handle large amounts of unstructured data.

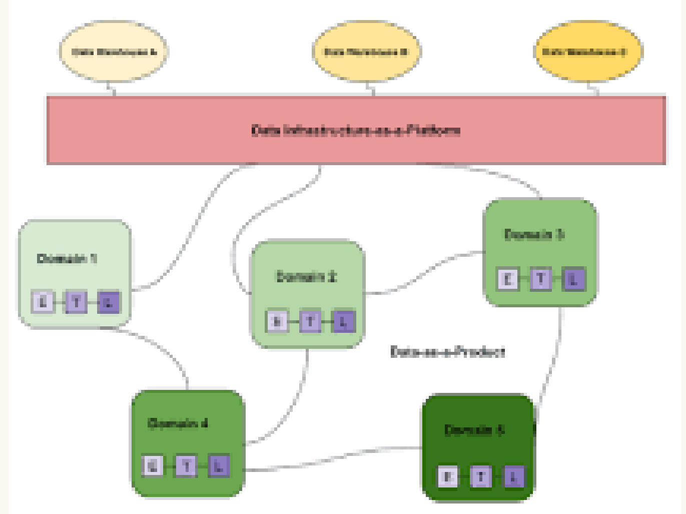


FEATURES OF DATA LAKE HOUSE

- A data lake house is a hybrid data management architecture that combines the benefits of data lakes and data warehouses
- Data lake houses are scalable and flexible, making it easier for organizations to manage and analyze large and diverse data sets as they grow.
- It provides a centralized repository for both structured and unstructured data, allowing organizations to maintain the flexibility of data lakes while also providing the structure needed for analytical queries.
- Data lake houses support all types of data in all file formats, making it easier for organizations to store and manage diverse data sets.
- Schema support and mechanisms for data governance ensure that data is organized and managed effectively, making it easier to derive meaningful insights.
- Real-time capabilities enable data science, machine learning, and data analytics projects to be performed in real-time, providing faster and more accurate results.

DATA MESH

Data Mesh is a data platform architecture that promotes easy access to important data by end-users without the need to transfer it to a data warehouse or data lake, or to involve expert data teams. It achieves this through decentralization, whereby data ownership is distributed among teams who can manage it as a product independently and securely. This reduces bottlenecks and silos in data management and allows for scalability without compromising data governance.



BENEFITS OF DATA MESH

- Data Mesh enables decentralized data operations, improving agility, scalability and business domain agility.
- Enterprises using Data Mesh benefit from increased flexibility and independence, avoiding being locked into a single data platform or product.
- Data Mesh offers faster access to critical data through a self-service model and a centralized infrastructure, making SQL queries faster and easier.
- Decentralized data ownership in Data Mesh promotes transparency and cross-functional use across teams, reducing dependence on expert data teams and increasing collaboration in data management.

DATA WAREHOUSE VS DATA LAKE

Data Warehouse	Data Lake
Structured Data	Unstructured or Semi Unstructured Data
Processed Data	Raw Data
Batch Processing	Real Time / Batch Processing
Analyzing Historical Data	Exploring data and discovering insights
SQL Queries	Various tools and languages
Fixed Schema	Dynamic Schema

OLTP VS OLAP

OLTP	OLAP
Online Transaction Processing	Online Analytical Processing
Supports Transactional Processing	Supports Analytical Processing
Current / Real Time Data	Historical Data
Simple Transactional Queries	Complex Analytical Queries
Fixed Schema, changes are difficult to implement	Flexible Schema, changes are easier to implement
Operational Staff	Business Analyst / Data Scientist

TASK # 03
QA

Can a database be used as DWH?

- Although it is possible to use a database as a data warehouse, there are some key differences to consider. Databases are designed for transactional processing with a focus on fast read and write operations and low-latency access to data. In contrast, data warehouses are designed for analytical processing, with a focus on heavy read operations and querying for reporting. If using a database as a data warehouse, additional design considerations and optimizations such as denormalization and indexing may be required to support complex analytical queries and reporting. Additionally, databases may not be as efficient as dedicated data warehouse platforms when handling large volumes of data. Therefore, while it is possible to use a database as a data warehouse, it may not be the most optimal solution in every scenario.

STRUCTURED VS UNSTRUCTURED DATA

Structured Data	Unstructured Data
Several Formats	Huge variety of formats
Organized information	Diverse structure for information
Data Warehouses	Data Lakes
Easy to Search	Difficult to Search
Relational Database (SQL)	Non Relational Databases (No SQL)
Requires less storage	Requires more storage
Customers Data, Financial Transactions	Images, Videos, Audio, Text files

WHAT ARE THE DUTIES OF DATA ENGINEER?

A Data Engineer is responsible for:

- Design and Develop Data Architectures: Data Engineers design and develop the architecture for data systems including Data pipelines, data warehouse and data lakes. They ensure that the architecture is scalable, reliable and optimized
- Developing Data Pipelines: Data engineers build and maintain data pipelines that transport data from source to target systems. This includes creating connectors to different data sources performing data transformation and ensuring data quality and consistency
- Ensure Data Quality: Data engineers are responsible for ensuring that data is consistent, accurate and trustworthy. They implement data validation and quality checks. They work with data scientists and analyst to resolve any issues arise.
- Implement Data Governance Policies: Data engineers implement data governance policies to ensure that data is managed in compliance with legal and regulatory requirements and ensure the security of sensitive data by implementing access controls, encryption, and other security measures
- Data Integration: Collect, clean and transform data from different sources so that it can be used for analysis. This may involve designing and implementing ETL (Extract, Transform, Load) processes that move data from one system to another.
- Performance Optimization: Optimize data processing and query performance by tuning databases, data pipelines, and other systems. This includes monitoring system performance and implementing optimizations to improve system efficiency