Week1_DataMarts_and_Lake_Overview

What is a data mart

Data mart is an isolated part of a larger enterprise Data Warehouse specifically built to serve a particular business function, purpose or community of users.

Examples:

- Sales Data Mart
- Finance Data Mart
- Marketing Data Mart

Data marts are used to:

- Provide support for tactical decision-making
- Help end users focus only on relevant data
- Save time otherwise spent searching the data warehouse for answers

Typical structure of a data mart:

- Relational database
- Star or snowflake schema
- Central fact table of business metrics
- Surrounded by associated dimension tables

It is a smalll data warehouse with tactical scope, and is lean an fast

Types of data marts

Dependent, independent and hybrid.

Dependent: Draw data from the enterprise data warehouse.

> Inherit security from the EDW
> Use cleaned and transformed data
> Have simpler data pipelines

Independent: Created directed from sources, internal data or outside sources

> Require custom ETL data pipelines
> May require additional security measures

Hybrid: Depend partially on the enterprise data warehouse, combine data from the warehouse and external data

The purpose is to provide timely and relevant data, rapid query responses, cost efficiency and secure access.

What is a Data Lake

It is a storage repository that can store a large amount of structured, unstructured and semi structure data in their native format, classified and tagged with metadata.

- Sometimes used as a staging area prior to load into a warehouse or mart
- Data can be loaded without defining the structure or schema of data
- Use cases do not need to be known in advance
- Exist as a repository of raw data straight from the source
- A reference architecture that combines multiple technologies
- Can be deployed using:
    - Cloud object storage
    - Large-scale distributed systems
    - Relational database management systems
    - NoSQL data repositories

Benefits

- Handles all types of data
- Scalable storage capacity
- Saves time that would have been used to define structures, create schemas and transform data
- Can quickly repurpose data for a wide range of use cases
