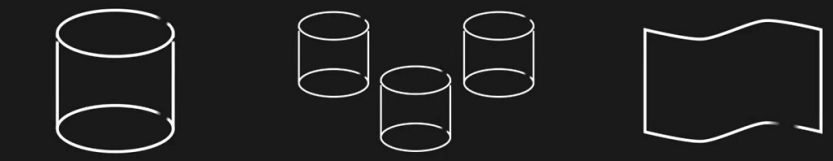
Data Warehouses, Data Marts, and Data Lakes

Introduction

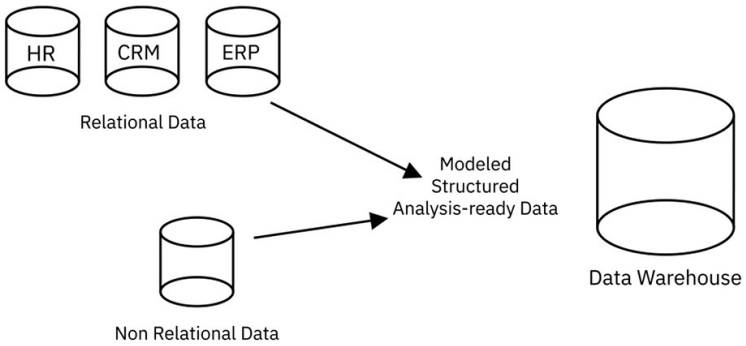
Data Mining Repositories store data for:

- Reporting
- Analysis
- Deriving insights



Data Warehouses | Data Marts | Data Lakes

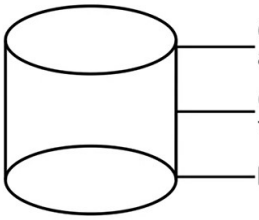
Data Warehouses



- **Relational data from transactional systems and operational databases**
- **Non-relational data**

Data Warehouses

A Data Warehouse has a 3-tier architecture:



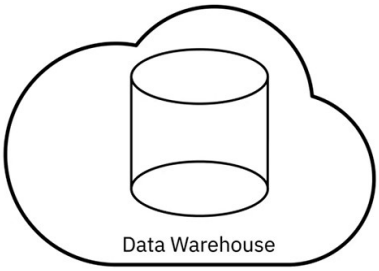
Client Front-End Layer (querying, reporting, and analyzing data)

OLAP Server (process and analyze information coming from database servers)

Database Servers (extracting data from different sources)

Data Warehouse

Data Warehouses



Benefits of cloud-based data warehouses:

- Lower costs
- Limitless storage and compute capabilities
- Scale on a pay-as-you-go basis
- Faster disaster recovery

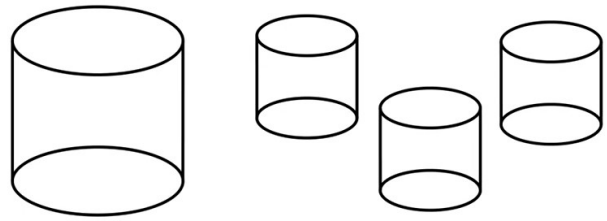
Data Warehouses

teradata. ORACLE EXADATA IBM Db2 NETEZZA

amazon REDSHIFT Google BigQuery cloudera snowflake

Data Marts

A data mart is a sub-section of the data warehouse, built specifically for a particular business function, purpose, or community of users.

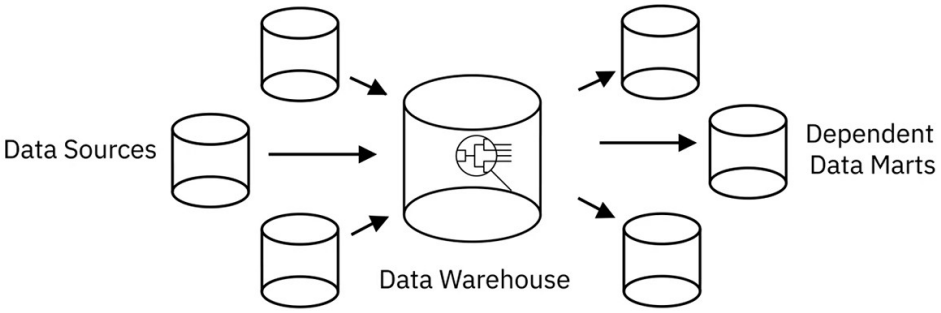


Data Warehouse Data Marts

Three types of data marts:

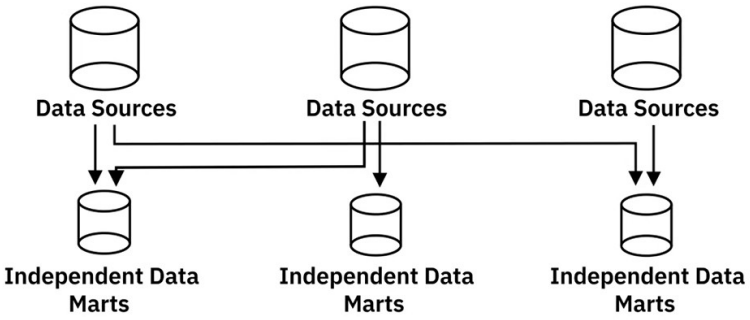
- Dependent
- Independent
- Hybrid

Data Marts



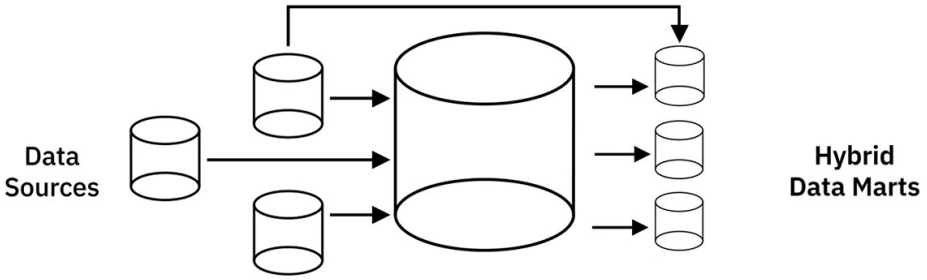
Dependent Data Marts offer analytical capabilities for a restricted area of a Data Warehouse.

Data Marts



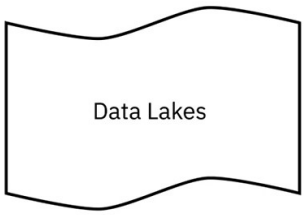
Independent Data Marts are created from sources other than an Enterprise Data Warehouse, such as Internal Operational Systems or External Data.

Data Marts



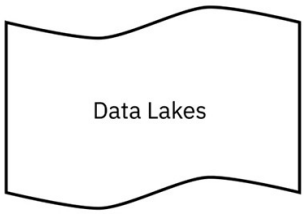
Hybrid Data Marts combine inputs from Data Warehouses, Operational Systems, and External Systems.

Data Lakes



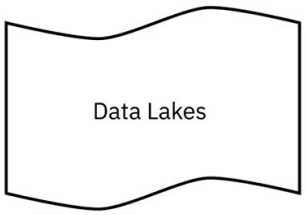
- Store large amounts of structured, semi-structured, and unstructured data in their native format
- Data can be loaded without defining the structure and schema of data
- Exist as a repository of raw data straight from the source, to be transformed based on the use case
- Data is classified, protected, and governed

Data Lakes



- A reference architecture that combines multiple technologies
- Can be deployed using
  - > Cloud Object Storage, such as Amazon S3
  - > Large-scale distributed systems such as Apache Hadoop
  - > Relational Database Management Systems, as well as NoSQL data repositories

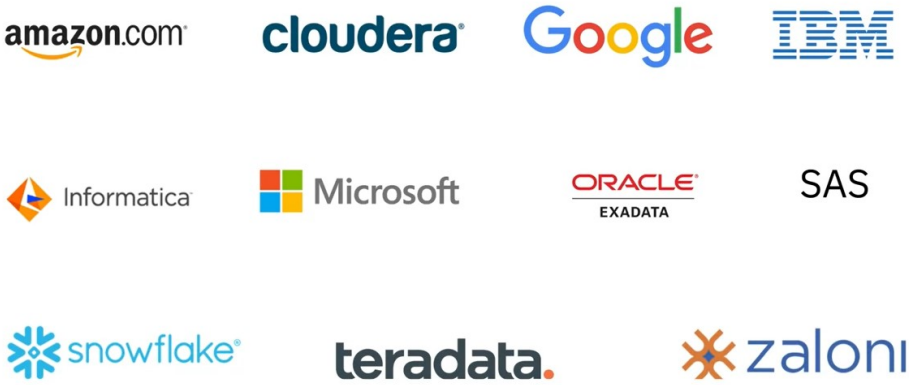
Data Lakes



Benefits:

- Ability to store all types of data (unstructured, semi-structured and structured data)
- Agility to scale based on storage capacity (growing from terabytes to petabytes)
- Saving time in defining structures, schemas, and transformations (data is imported in its original format)
- Ability to repurpose data in several different ways and wide-ranging use cases

Data Lakes


amazon.com
cloudera
Google
IBM
Informatica
Microsoft
ORACLE
EXADATA
SAS
snowflake
teradata.
zaloni