Database vs Data Warehouse vs Data lake

Database: An organized collection of structured data with defined schema → Create table

↑ intuition

Purpose: Handles transaction data and day-to-day operational activities

Characteristics:

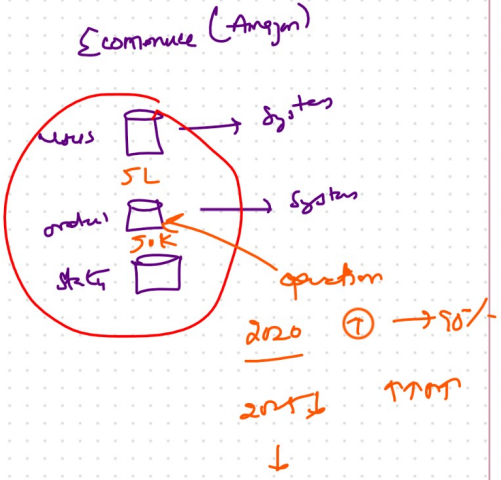
- Write heavy (optimized for writing operations)
- Structured data only
- Schema validation required [errors | mismatch]
- Stores recent data only [not historical data]
- High storage cost
- Best for OLTP [Online transaction Processing]

Examples: MySQL, PostgreSQL

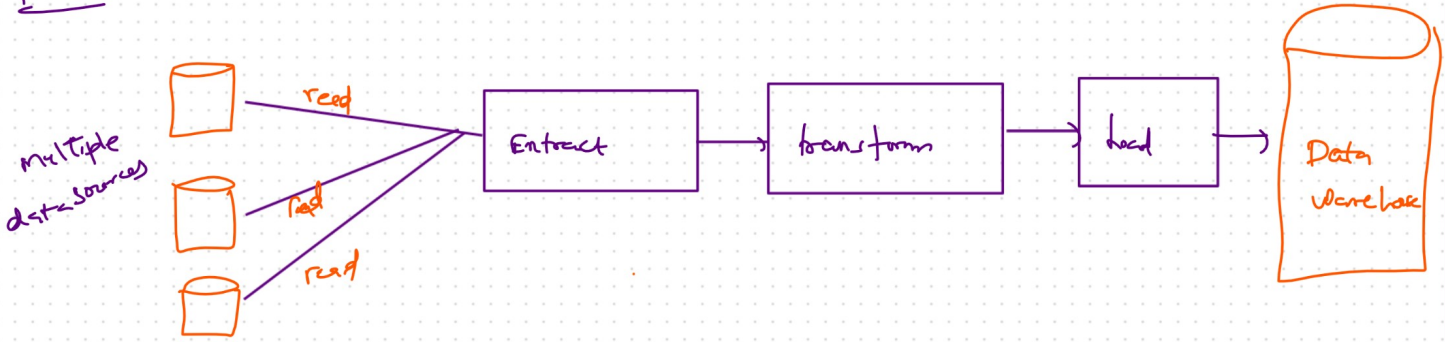
Use Case: A single grocery store tracking daily sale, inventory, employee hour

Data Warehouse: A centralized repository for structured data from multiple sources

Purpose: Used for historical data analysis & business intelligence



process: Uses ETL (Extract, Transform — load)



Characteristics :

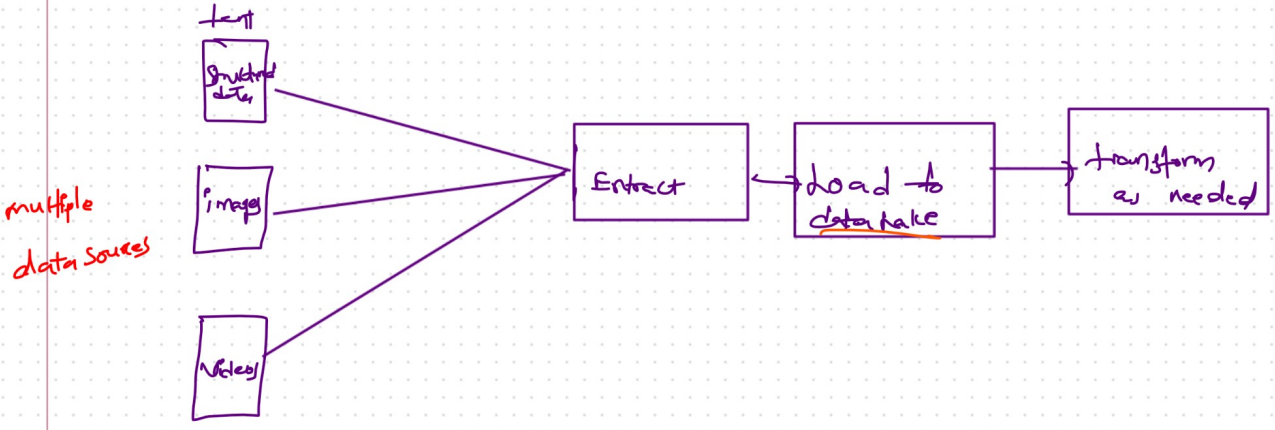
* Read-heavy (optimized for analytical queries)
* Structured data only
* Stores large amounts of historical data
* Moderate cost (lower than databases)
* Best for OLAP (Online Analytical Processing)
* Moderately scalable & flexible

example : tata data

Use Case example : Head office of grocery chain collecting data from all branches to analyse trends over time

Datalake

definition : A centralized repository that stores all kinds of raw data with no restriction



Characteristics :

* Stores all data types (structured, semi-structured, unstructured)
* No schema validation required at ingestion time
* High scalable
* Cost-effective (uses cheap storage)

- Very flexible
- transform data as when required

Example: HDFS (Hadoop file System), S3

Use case example: Grocery chain storing structured data along the side semi-structured data like reviews & unstructured data like product images & CCTV footage