UNIT 4: DATA WAREHOUSING

What is data warehousing?
It is a type of computer database; where large amounts of data from different sources can be stored and managed.
The goal is to have an efficient way of managing info & analyzing data.

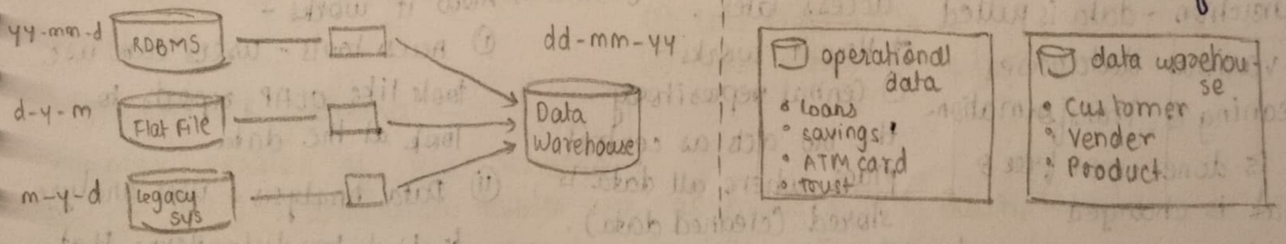
Need of data warehousing :-

① Data consolidation - homogeneous sources data is stored into one place.
② Better Reporting & Analysis - helps to analyse quickly.
③ Historical Analysis - look at past data to make future predictions.
④ Improved performance - separate analytic tasks from daily operations, that speeds up report generation.

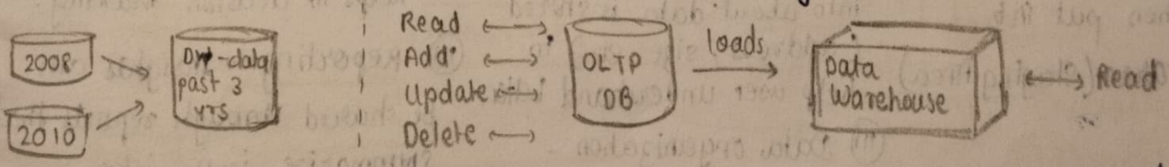
According to Bill Inmon - "A data Warehousing is a subject oriented, integrated, non-volatile, & time varient collection of data in support of management's decisions."

Characteristics of data Warehouse :-

① subject - Oriented data - focus on key areas (like sales or customer) rather than daily operations.
② Integrated - combines data from various sources into consistent format.



③ Time-varient - keeps historical data, allowing users to look at trends overtime.



④ Non-volatile - once data is stored, it does'nt change, making it reliable for reporting.

Advantages of Data Warehouse :-

① centralized Data repository - single source of truth for decision makers.
② Uniformity - end users can use a single data model & query language.
③ Stores historic data
④ Faster Query Performance
⑤ Secure information
⑥ Scalable
⑦ Make better decisions
⑧ Enhance company's performance.

Limitations of Data Warehouse :-

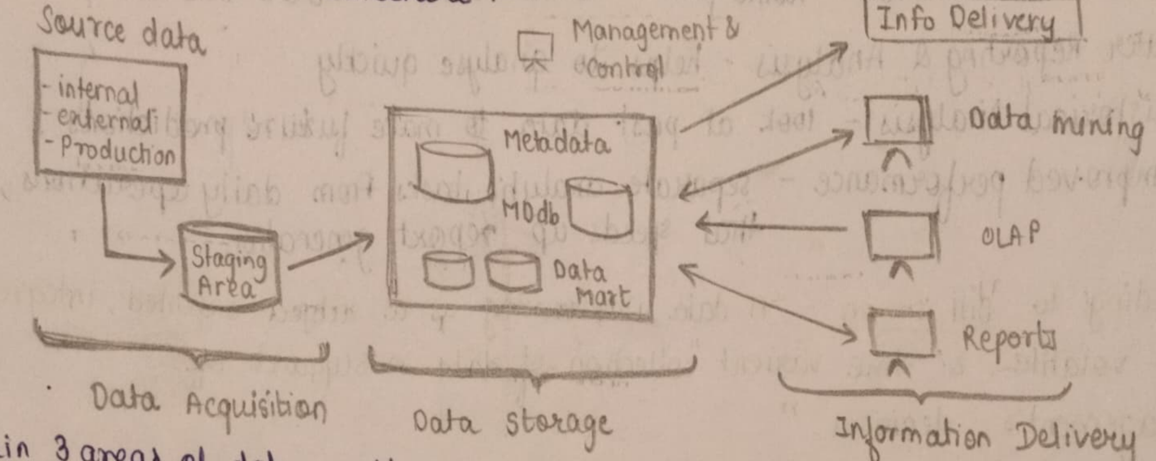
① Complexity of implementation/integration - ~~requires investment in infrastructure & skilled personnel~~ Process of extracting, transforming, loading data is time consuming & complex.

② Time consuming process. ③ High intial cost - requires investments.

④ High maintainance cost. ⑤ Data security risks

⑥ Changing requirements of end user - end user is always demanding in nature.

Data Warhouse Architecture :-



Main 3 areas of data warehouse :-

① Data Acquisiton

→ data is collected from diff. sources.

→ How it works -

(i) Extraction - data is pulled from various databases

(ii) Cleaning & Transformation- data is done error free & format is changed

(iii) Loading - this cleaned data is then put into temporary area (Staging Area)

② Data storage

→ After data is acquired, it needs to be stored properly for easy access later.

→ How it works -

(i) Central repository - the DW acts as central place where all data is stored (cleaned data).

(ii) Metadata management - info about data is stored (address, size, etc) to help user understand better.

(iii) Data organization - data is arranged in structured (tables) way for easy retrival.

③ Information delivery

→ This is final step, provides users with access to stored data so they can analyze it.

→ How it works -

(i) Access tools - users can use tools like OLAP, reports to look at the data.

(ii) Data Analysis - user analyze & find trends/patterns that helps in decision making.

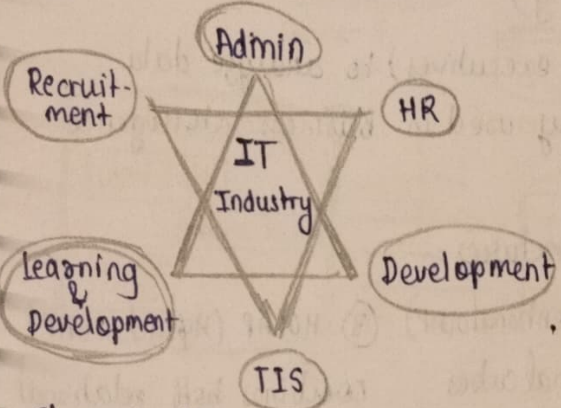
(iii) Reporting - Insights can be shared through reports that summarize imp. info.

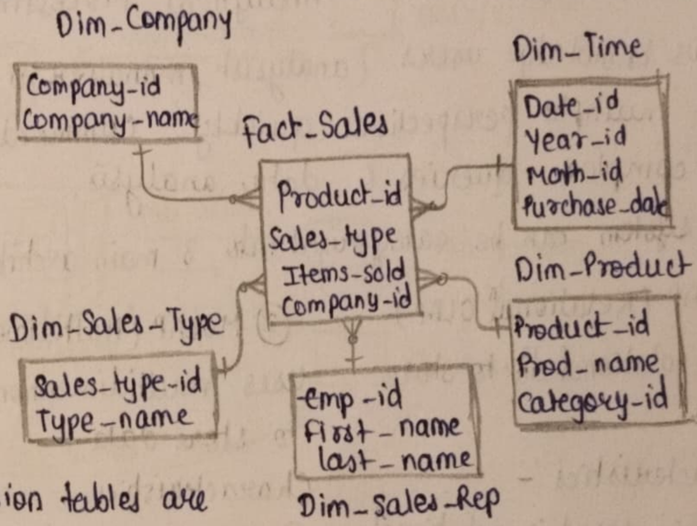
Datawarehouse schema

Schemas are used to define how data is organized and how different data elements relate to each other.

① STAR Schema -

Simplest & straightforward design, consist of entity - relationship diagram in star shape and the centre of star consist of fact table & points of the stars are dimension tables.





Characteristics :-

① Denormalized structure - Dimension tables are denormalized ie, contain all attributes so no need of joins.

② Single fact table - 1 fact table connects multiple dimension tables.

③ Easy to understand.

Advantages :-

① fast query optimization & performance.

② Simplicity.

Disadvantages :-

① Data Redundancy due to denormalization.

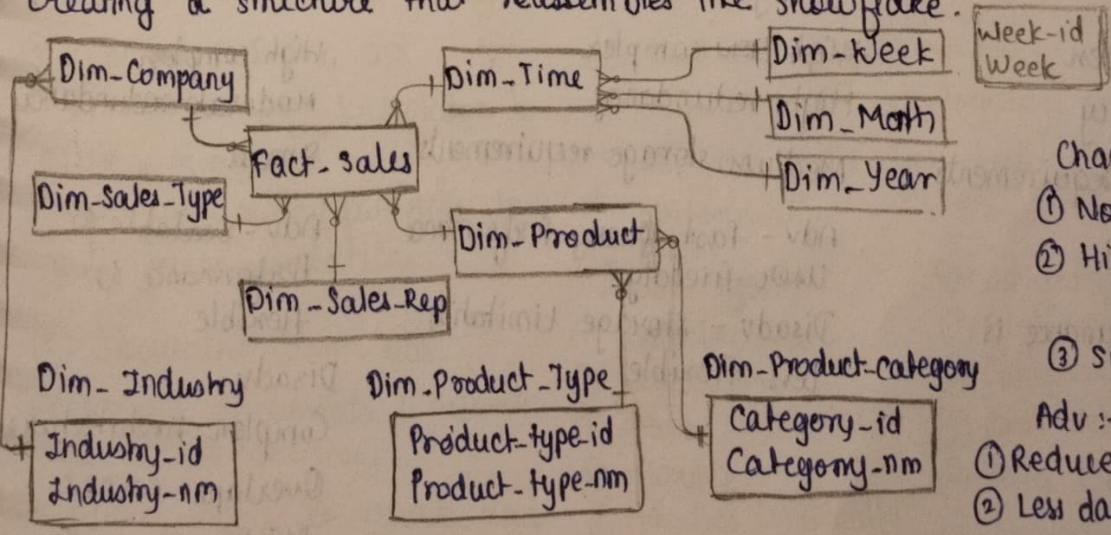
② Less flexibility.

③ Poor performance for dynamic (keeps changing) data.

② SNOWFLAKE Schema —

Complex version of Star Schema.

In this design, dimension tables are normalized into multiple related tables, creating a structure that reassembles like snowflake.



Characteristics :-

① Normalized structure

② Hierarchical Relationships

③ Single fact table

Adv :-

① Reduce Data Redundancy

② Less data inconsistency

Disadv :-

① Increase complexity, complex structure ② Complex Query, slow performance

Feature	Star schema	Snowflake schema
Structure	Denormalized.	Normalized.
No. of Tables	fewer tables.	More tables (sub-dimensions).
Query Performance	faster due to fewer joins	Slower due to more joins
Data Redundancy	High	Low
Complexity	Simple & easy to understand	More complex & harder to maintain.

OLAP Architecture (Online Analytical Processing)

→ Allows knowledge users (analysist, manager & executives) to analyse data from multiple perspectives quickly. Commonly used in Business intelligence for complex queries & data analysis.

OLAP system can be categorized into 3 main architectures:-

① ROLAP (Relational OLAP)

Uses relational db to store data.

Characteristics -

① Data is stored in relational tables.

② Uses SQL for querying, which allows detailed & complex queries.

③ Support large volm of detailed data.

② MOLAP (Multidimentional OLAP)

Uses multidimentional cubes to store data.

Characteristics -

① Data is stored in cubes

② Pre-calculation aggregation of query improves performance

③ Dimensions are organized hierarchically.

③ HOLAP (Hybrid OLAP)

Combines both relational database & cubes to store data.

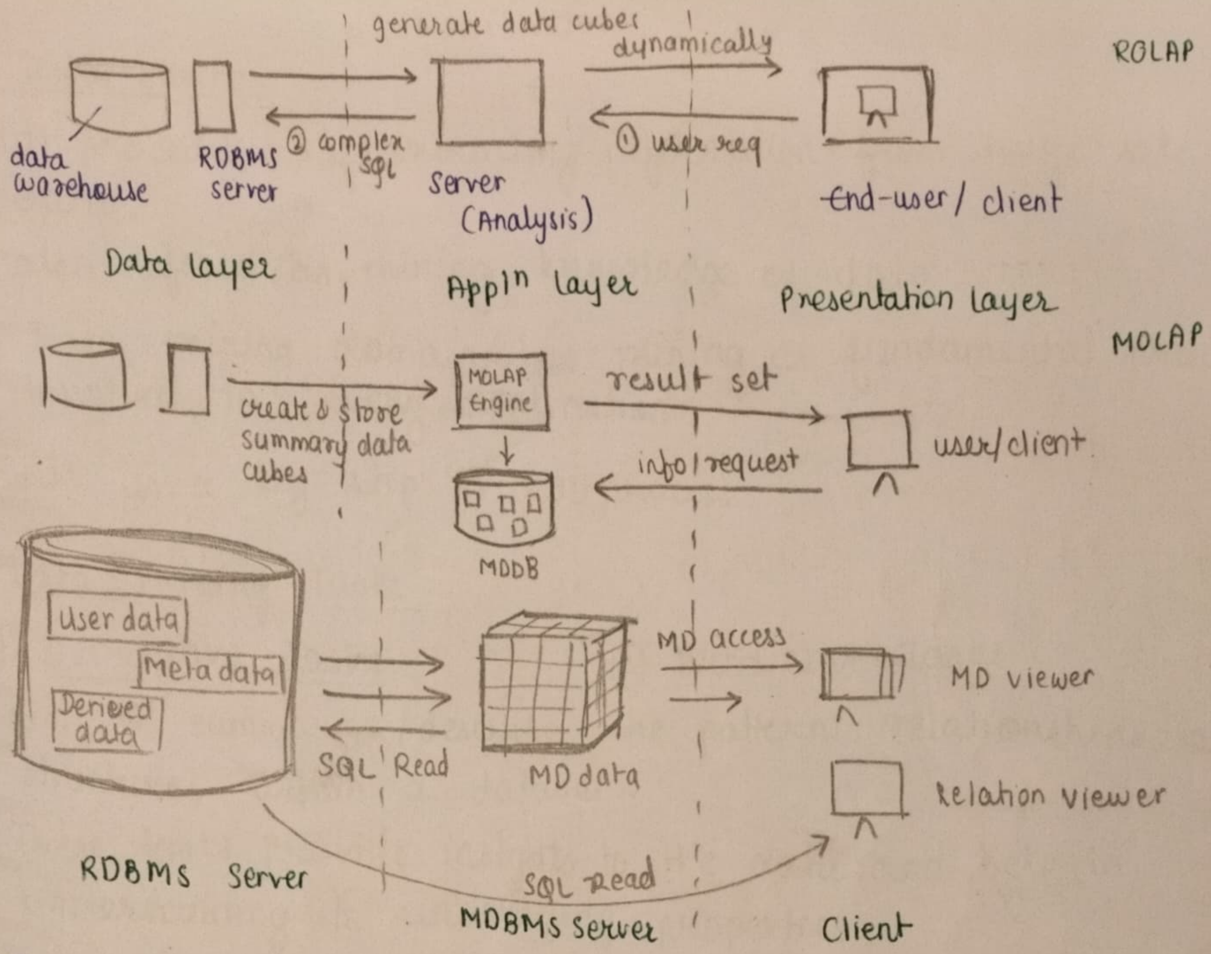
Characteristic

① of ROLAP

① of MOLAP

③ Provides balance betn detail & performance.

SQL for querying	MDX (multidimentional expression) for querying	SQL & MDX
Slow due to multiple joins	fast due to pre-calculated aggregations	Balanced performance
Moderate complex	~~High~~ Low complex	High complex
Low redundancy	High redundancy	Moderate redundancy
Large storage requirements	Medium storage requirements	Small
Adv - scalability flexibility	Adv - fast Query Performance User friendly	Adv - Scalable & Performance & flexible
Disadv - Performance is slow. Complexity.	Disadv - storage Limitation Less flexible	Disadv - Complex Architecture Overlaps ROLAP & MOLAP can create redundancy.



Decision Support System. (DSS)

It is a real time decision making tool, which assist managers & decision-makers. It integrates data from various sources & uses analytics model to support complex decision-making processes.

Types of DSS

① Data Driven DSS
- focus on analysing large amount of data.
- Useful for tasks like sales forecasting & inventory management.

② Model Driven DSS
- Uses mathematical models to help analyze data.
- Good for things like financial forecasting & optimization.

③ Knowledge Driven DSS
- Provides expert knowledge to assist in decision making.
- Often uses rules/guidelines to give recommendations.

④ Communication Driven DSS
- Helps teams work together on decision
- Includes tools for chatting & sharing information.

⑤ Document Driven DSS
- Manages documents & reports related to decisions.
- Allows users to create and view imp documents easily

Advantages:-
① Better Decisions
② Saves Time
③ Team Collaboration
④ Flexible
⑤ Handles Complexity

Disadvantages:-
① High Cost
② Complex to use
③ Over-Reliance on technology
④ Data Quality Issues
⑤ Security Risks