Module 3 - Introduction to Data Warehousing and Multi-Dimensional Modelling.

* Operational Support System vs Decisional Support Sys-

O.S.S	D.S.S
① In terms of purpose, O.S.S automates routine tasks and processes.	① In terms of purpose, D.S.S Provides data and tools for decision making and analysis.
② In terms of data, O.S.S has Structural and transactional data.	② In terms of data, D.S.S has Historical and predictive data.
③ The concurrent transaction volume in O.S.S is very high.	③ Concurrent transaction levels in D.S.S is medium or low.
④ Regarded as update transactions.	④ Regarded as query transactions. (Read-only)
⑤ O.S.S supports transactions which are happening in real-time	⑤ D.S.S supports transactions which have happened a long time ago to analyze and make decisions on them.

* Data Warehousing

→ Data Warehousing is a process of collecting, organizing and storing large amounts of data from various sources in a centralized repository.

→ The data warehouse is designed to support buisness intelligence activities such as data analysis, reporting, and decision-making.

** Features of Data Warehouse

① Subject-oriented - A data warehouse is designed around specific subjects or topics related to an organisation's buisness, such as sales, customer data and financial data.

② Integrated - Data in a Data Warehouse is integrated from multiple sources, including internal systems and external data sources. Helps data stay accurate.

③ Time-variant - Data in a data warehouse is stored over time allowing analyst to track changes in data and identify trends over time.

④ Non-Volatile - Data in a data warehouse is read-only, so data cannot be updated or changed directly.

* The Need for Data Warehousing

→ There are many reasons why organisations need data warehousing.

① Centralised data storage - organizations generally collect data from various sources, including transactional systems, websites, social media, articles etc. making it easier to access and analyze data

② Improved data quality - Inconsistent data is a major issue, but since data comes from various sources, the accuracy and quality of data is improved significantly.

③ Analytics and reports - Data warehouses are designed for analytics and reporting; with tools and techniques optimized for data analysis and visualisation

④ Faster decision-making. - Data warehouses provide timely access to the data they need. This is imp for organisations working in competitive markets.

⑤ Historical analysis - Data warehouses store large amount of data for a long time, allowing ~~organisation~~ to analyze change in trend patterns since many years.

* Meta data

-> Meta-data refers to the info about the data stored in the warehouse.

-> It provides a description of data which is stored in the databases, such as origin, format structure.

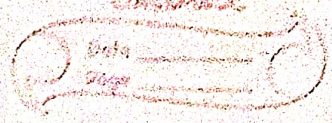
-> Metadata plays a key role in data warehousing by providing valuable information. It helps in

① understanding the data - Helps users to analyzing and report by knowing the origin, format and structure.

② Managing data quality - Metadata also includes information regarding the accuracy, precision and completeness of the information.

③ Data Integration - Meta data helps integrate data from multiple sources ensures data is consistent across the organisation.

④ Data governance - Metadata plays an imp. role in data governance by providing information regarding data ownership, data usage and data security.



④ Data Lineage - Metadata provides info. about the history of data helps users trace the lineage and understand concept.

---

* Classification of Metadata.

① Technical Metadata - Gives information regarding the technical details like format of data, storage, data model, data dictionaries etc

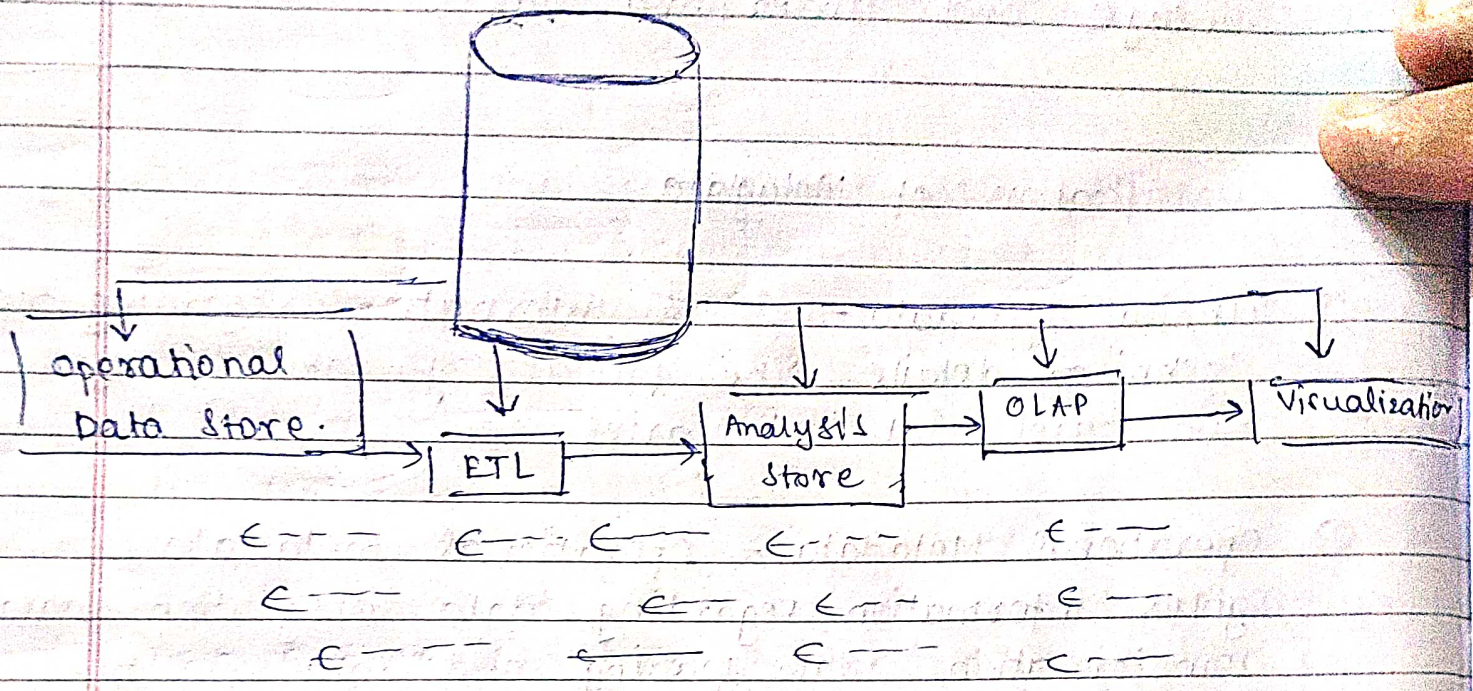
② Operational Metadata - Operational metadata gives information regarding data integration, data transformation, data loading and data quality.

③ Business Metadata - Contains business concepts represented by data.

④ Usage Metadata - Contains info regarding query execution, report generation and data visualization. used to analyse data

⑤ Data Lineage Metadata - Describes the history of the data stored. transformation, date it came into the warehouse etc.

* Information flow Mechanism and Architecture.



① Operational data store (ODS) - ODS is a database which stores the latest and real time data from various operating systems.

② ETL - contains 3 steps.

a) Extract - Data is extracted from ODS and other source systems, such as CRM, ERP etc.

b) Transform - The extracted data is then transformed into a format that is suitable for the data warehouse.

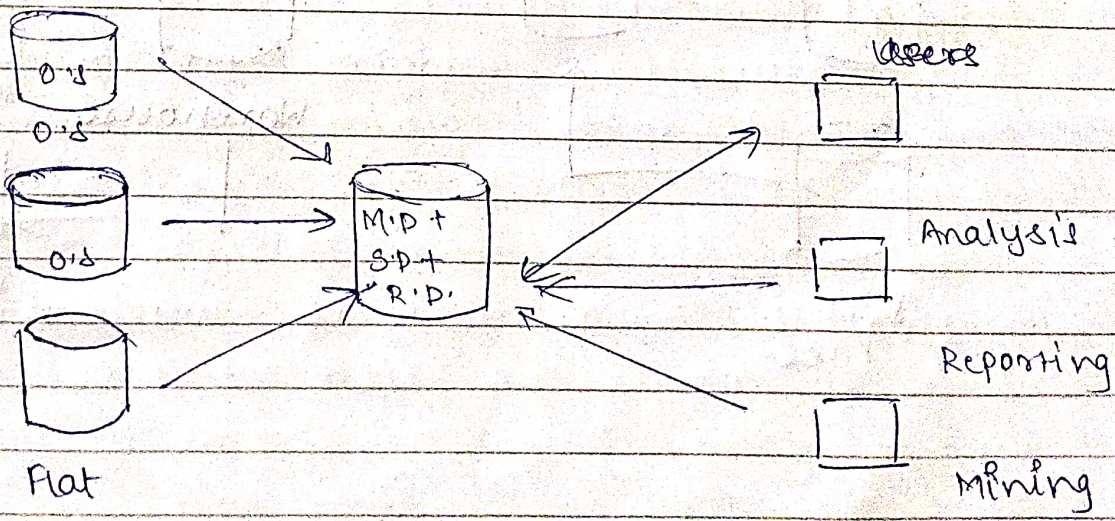
c) Load - Transformed data is then loaded onto the data warehouse. This involves putting data into appropriate tables and columns.

d) Analysis Store - Data is organised, structured and stored in a way which makes it easy to query and analyse.

e) OLAP (Online Analytical Processing) - organised data in a multi-dimensional structure. This allows users analyse data from different perspectives. OLAP provides tools to analyze trend patterns and anomalies.

f) Visualisation - Creates reports, patterns, trends, insights dashboards etc using data within the data warehouse.

* Architecture (Basic)

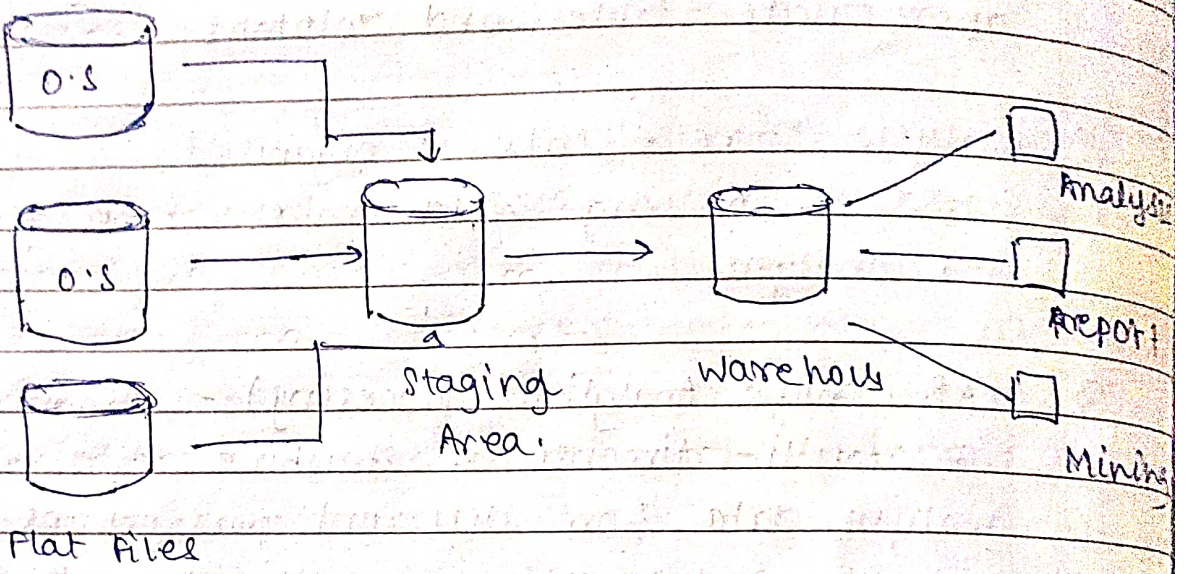


M.D → Meta Data

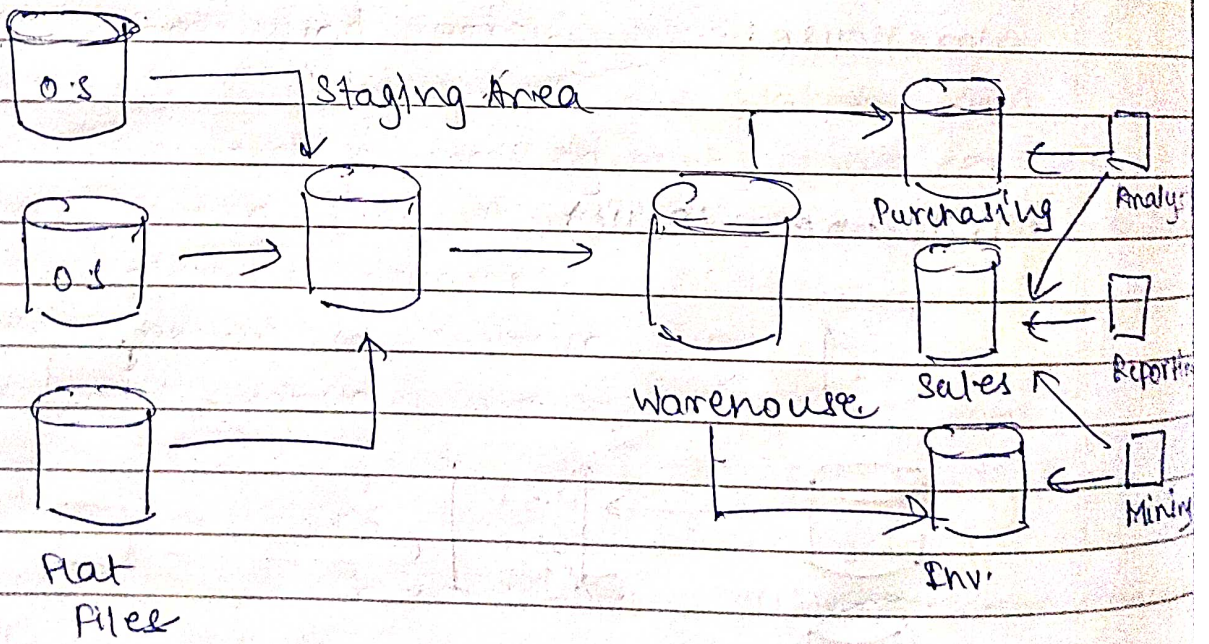
S.D → Summary data

R.D → Raw data

* Architecture (with Staging Area)



* Architecture (Staging Area and Data Marts)



Module 3 - (From 3.3 to 3.6)

* DataWarehouse VS Data Marts.

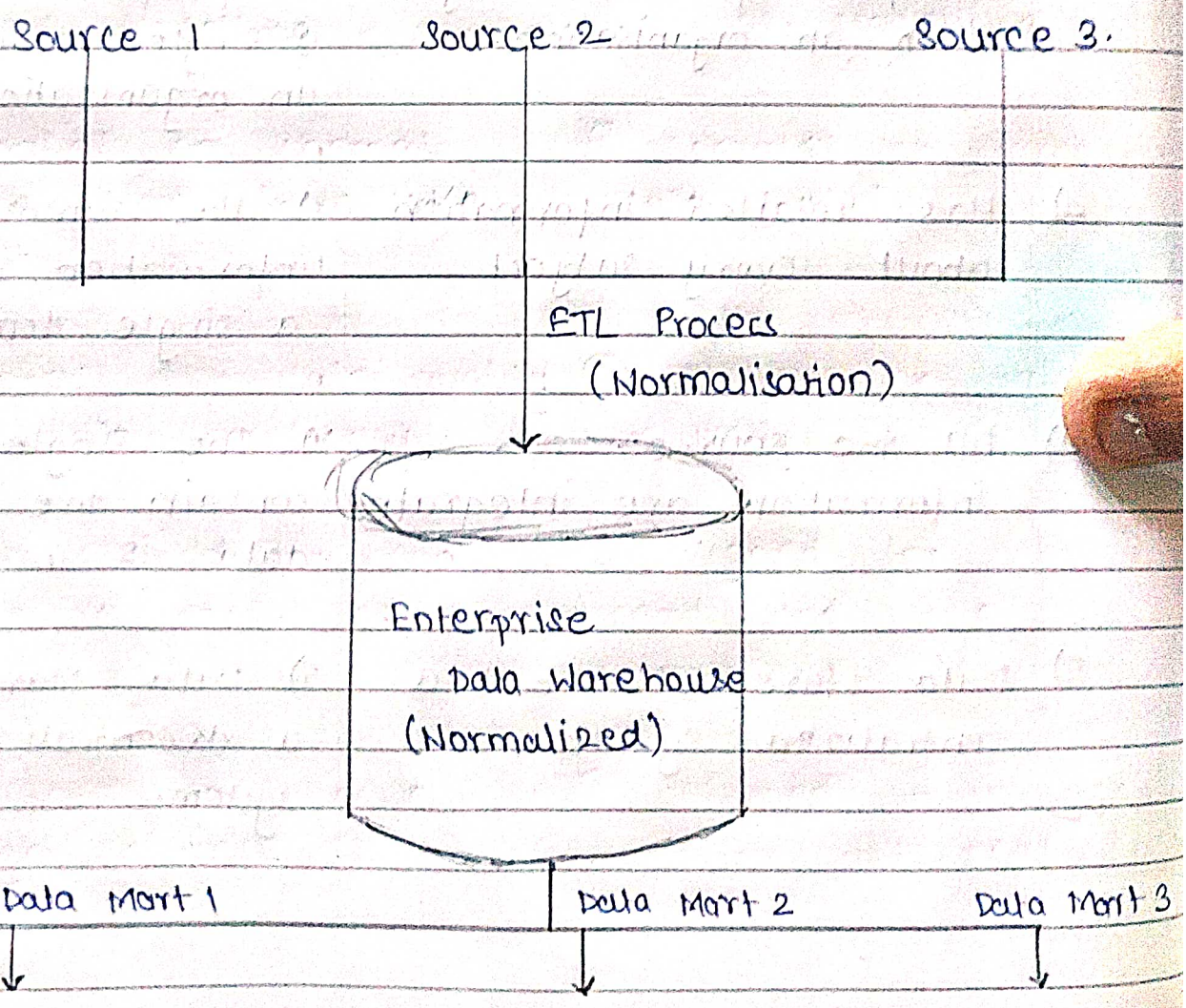
	Data Warehouse	Data Mart.
1)	Data Warehouse have a vast repository of collected information for internal as well as external sources	1) Data Mart is a sub category of data warehouse created to meet the requirements of a smaller sub-group
2)	Data Warehouse holds information related to multiple topics and subjects within an organisation.	2) Data Mart holds information related to a specific domain or a single ~~of~~ subject within an organisation.
3)	Has detailed information about every subject	3) Has summarized information regarding a single domain.
4)	All the sources of information are integrated	4) The single source contain the specific data is integrated.
5)	Data Warehouse has a centralized system.	5) Data Mart has a decentralized/distributed system.

* Data Warehouse design strategies.

→ Data Warehouse are used across multiple businesses to help with critical decision making and data analysis.

→ It has two main approaches.

① Top - Down Approach :- Also known as enterprise data warehouse (EDW) which involves creating a comprehensive, centralised data warehouse that serves an entire organisation.



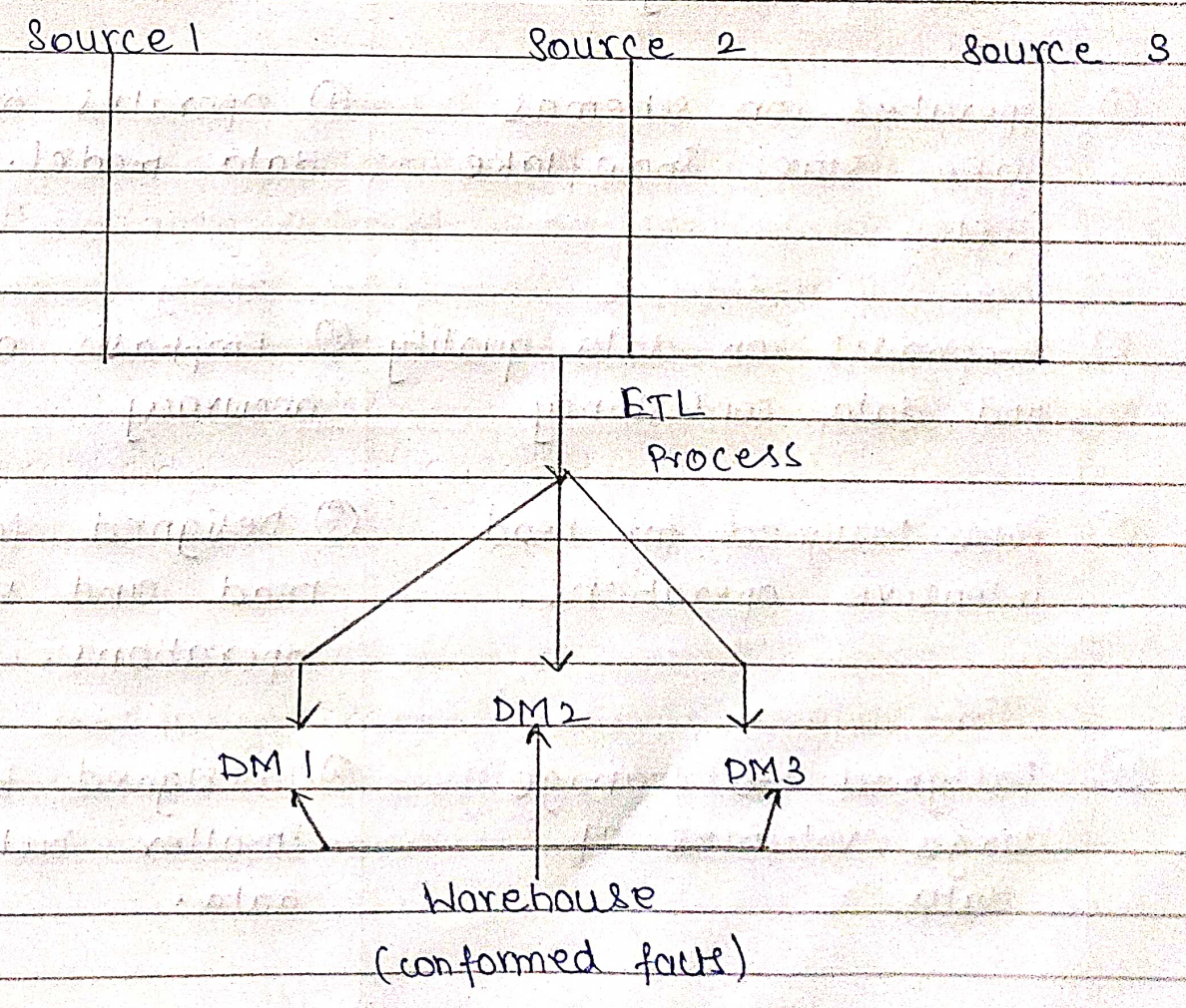
(3) Bottom Advantages of Top-Down Approach.

→ Ensures data consistency and accuracy.
→ Developing new data marts is easier.

→ Disadvantages of Top-Down Approach.

→ Inflexible to change in departmental needs
→ cost is high.

(4) Bottom-Up Approach :- Also known as the data-mart approach. This approach involves small, targetted data marts that serve specific business function and department



* Advantages of Bottom Up Approach:

→ Documents are generated quickly.
→ Cheaper than Top-Down Approach.

* Disadvantages of Bottom-Up Approach.

→ Leads to data inconsistency.
→ Difficult to manage and govern data.

* Data Warehouse Modelling vs Operational Database Modelling.

DW Modelling	OD. Modelling
① Operates on schemas like star, snowflake etc.	① Operates on normalized data model.
② Emphasis on data quality and data consistency	② Emphasis on data accuracy.
③ Designed for read intensive operations.	③ Designed for both read and write operations.
④ Designed to manage large volumes of data	④ Designed to mangge smaller but real-time data.

* Advantages of Bottom Up Approach.

→ Documents are generated quickly.
→ Cheaper than Top-Down Approach.

* Disadvantages of Bottom-Up Approach.

→ Leads to data inconsistency.
→ Difficult to manage and govern data.

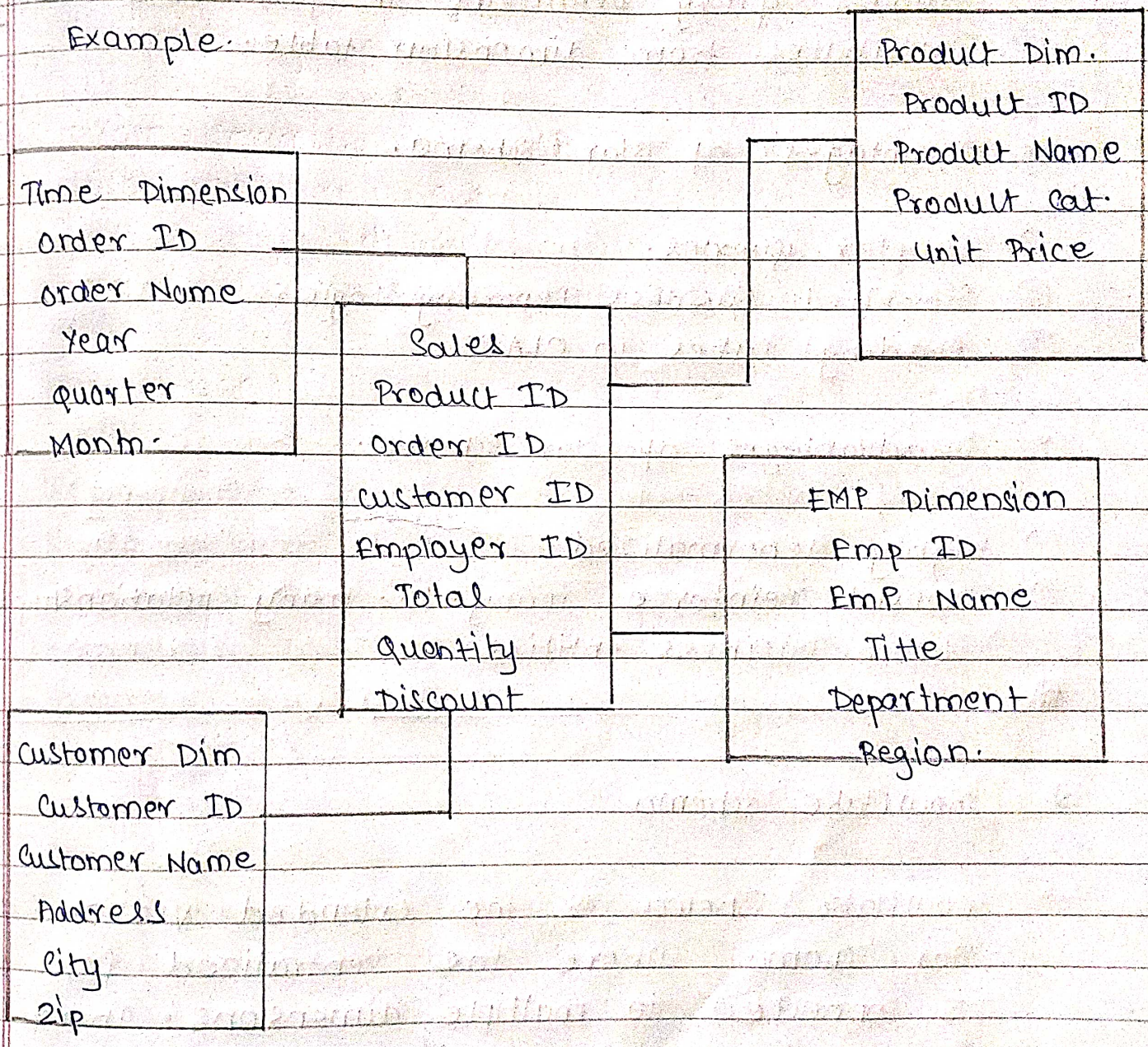
* Data Warehouse Modelling vs Operational Database Modelling.

	DW Modelling	OD. Modelling
①	Operates on schemas like star, snowflake etc.	① Operates on normalized data model.
②	Emphasis on data quality and data consistency	② Emphasis on data accuracy.
③	Designed for read intensive operations.	③ Designed for both read and write operations.
④	Designed to manage large volumes of data	④ Designed to manage smaller but real-time data.

* Star Schema

→ In Star Schema, data is organized into a central fact table that contains the measures of interest, surrounded by dimension tables that describe the attributes of the measure.

Example.



→ In Star Schema, Quantitative data is stored in fact tables and dimensions. Which are the context for the quantitative data is stored in dimension tables.

→ Each dimension table is joined to the fact table through a foreign key relationship. This allows query executions in fact table using attributes from dimension table.

* Advantages of Star Schema.

① Simpler Queries.
② Simplified Business Reporting Logic
③ Feeding Cubes in OLAP.

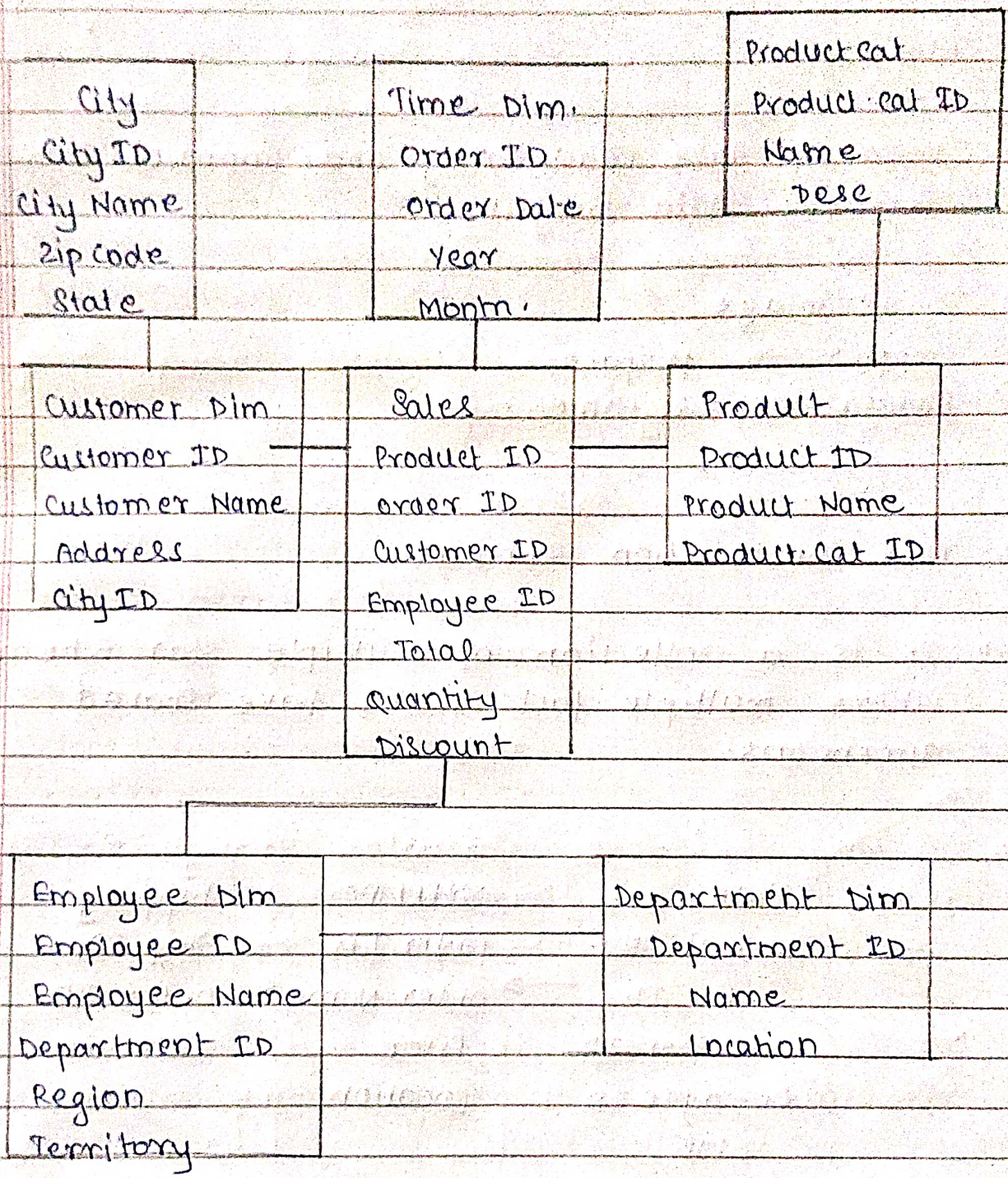
* Disadvantages of Star Schema.

① Highly denormalized.
② Doesn't reinforce many-to-many relationships within business entities.

---

* Snowflake Schema

→ Snowflake Schema is the extended version of the Star Schema where the centralized fact table is connected to multiple dimensions. There are several levels of relationship and hierarchies



→ The fact table is located at the centre of the schema, surrounded by dimension tables, Each dimension table is broken down into smaller dimensions.

* Advantages

① Reduces data redundancy and improves consistency
② Improves data integrity.

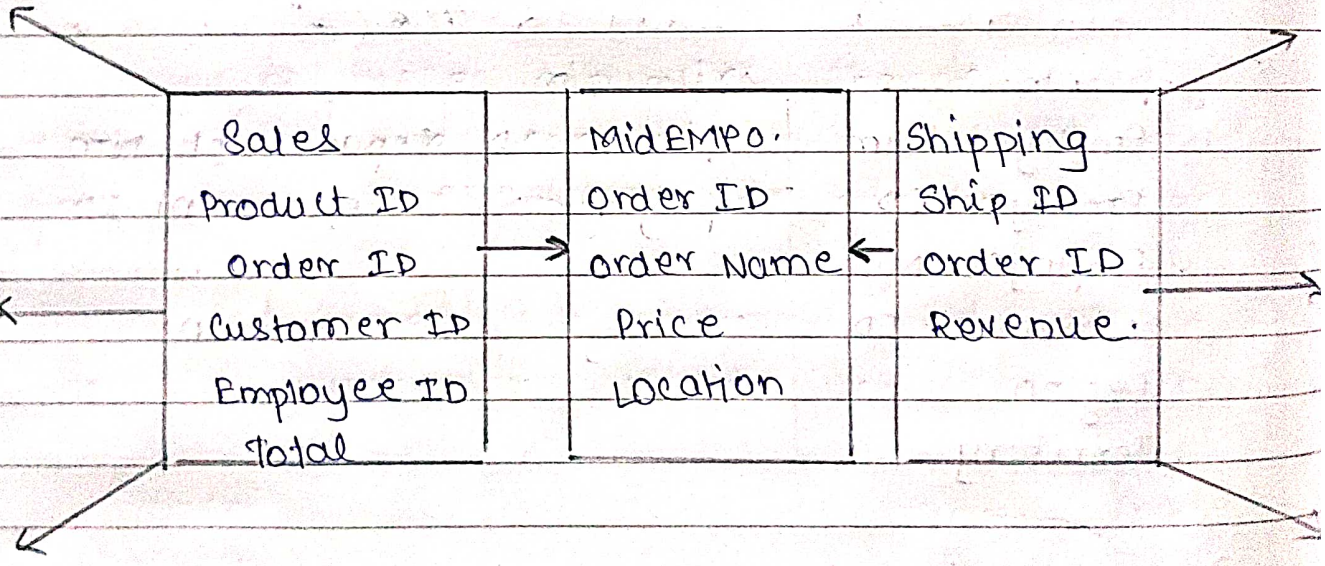
* Disadvantage

① Complex to prepare.
② Slower response time.

---

* Fact constellation Schema.

→ It is a collection of multiple star schemas where multiple fact tables have shared dimensions.



* Factless - fact table.

→ Factless fact tables are ones with only foreign key column present while facts/measures are absent. Meaning they are only used to establish a relations b/w 2 dimensions.

* Slowly changing Dimensions

→ As data constantly increases due to high transaction rates, the fact tables a tend to change frequently but over time changes can occur in dimension tables due to change in their attributes.

i) Type 1 change - overwriting existing data with new data having no record or log of the past and no changes except overwriting.

ii) Type 2 change - Tracking all the changes that are made by creating a new row for updated values.

iii) Type 3 change :- The change in creating new fields in the dimension record to capture the last value for that attribute

* Rapidly changing Dimensions

→ When one or more attributes change frequently eg. BMI, height, weight.

For eg- Patient dimension

Pat._ID
Name
Gender
Weight
BMI

→ If any of these attributes change rapidly, many rows are created and dataset would get larger and larger for minute changes

→ These attributes are removed and put in a seperate dimension called Junk Dimension.

Patient
Pat_ID
Name
Gender
↑
dimension
table

Patient_JNK
Pat_SK
BMI
index.

→ Pat_SK acts as a surrogate key in junk dimension which is the primary key.

* Data Lake.

→ A Data Lake is a large, centralised repository that allows you to store and manage vast amounts of raw data in its native format. Unlike traditional dw, Data Lake can store data in semi-structured, unstructured and structured format.

→ Architecture