Module 1

Q1) ER modelling and dimensional modelling

=>

	ER modelling	Dimensional modelling
(1)	Transaction oriented	(1) Subject oriented
(2)	Its core components are entities and relationships	(2) Its core components are fact tables and dimension tables
(3)	It eliminates redundancy	(3) It plans for redundancy
(4)	Highly volatile data	(4) Non-volatile data
(5)	Physical and logical model	(5) Physical model
(6)	Normalization is suggested	(6) De-normalization is suggested.
(7)	OLTP application	(7) OLAP application
(8)	Type of information used is real-time information	(8) Type of information used is historical information
(9)	High transaction volume	(9) Low transaction volume.
(10)	Application is used for buying products	(10) Application is used to analyze buying patterns

Page No.
Date

Q2) OLTP and OLAP

=>

OLTP	OLAP
(1) Online transaction processing	(1) Online analytical processing
(2) It makes use of standard DBMS.	(2) It makes use of data warehouse
(3) Tables are normalized	(3) Tables are not normalized.
(4) The data is used to perform day-to-day operations.	(4) The data is used in planning, problem-solving and decision making.
(5) The size of the data is relatively small	(5) The size of the data is relatively large.
(6) Queries are very fast	(6) Queries are relatively slow
(7) Both read and write operations	(7) Only read and rarely write operations
(8) It serves the purpose to Insert, update, and delete the information from the database	(8) It extracts the information for analysis and decision-making.

Page No.
Date

(9) Star Schema & Snowflake schema

=>

	Star Schema	Snowflake schema
(1)	Top-down model	Bottom up model
(2)	Uses more space	Uses less space
(3)	Takes less time	Takes more time
(4)	Normalization is not used	Normalization and De-normalization are us
(5)	Design is very simple	Design is complex
(6)	Query complexity is low	Query complexity is higher
(7)	Less number of foreign keys	More number of foreign keys
(8)	High data redundancy	Low data redundancy
(9)	The fact tables and dimension table are contained	The fact tables, dimension tables as well as sub dimension tables are contained.

Page No.
Date

(Q4) Top down approach & bottom up approach

=> Top down approach

(1) Centralized data warehouse is built first.

(2) Development time is longer

(3) High complexity

(4) High scalability

(5) High risk

(6) Ensures consistency.

(7) High investment required.

Bottom up approach

(1) Data marts are built first.

(2) Development time is shorter.

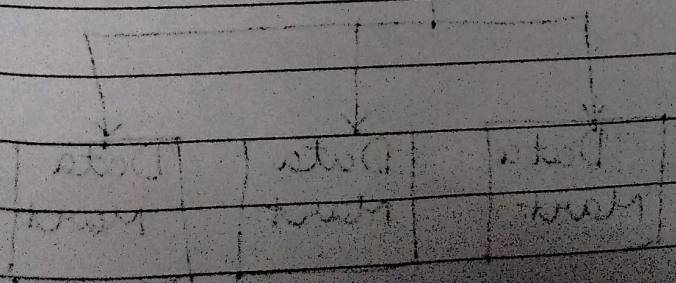
(3) Initially low complexity

(4) Moderate scalability

(5) Low risk

(6) Does not ensure consistency.

(7) Incremental investment required as data marts grow (low)



Page No.
Date

(Q4) Data warehouse design strategies

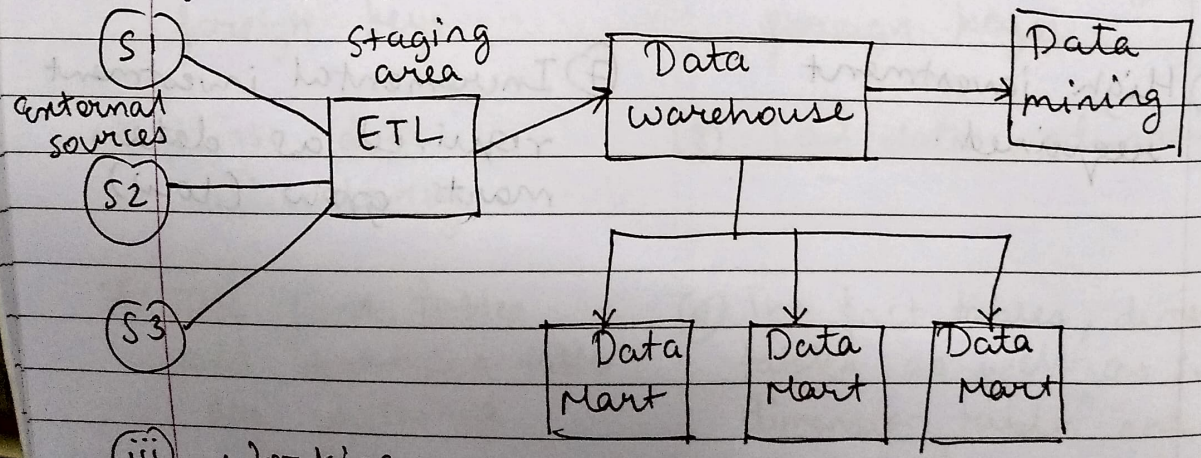
=> (1) A data warehouse is a single data repository where a record from multiple data sources is integrated for OLAP.

(2) It collects, organizes and stores data from different places to help you make better decisions.

(3) When building a data warehouse, there are two main ways to do it:
(i) Top down approach
(ii) Bottom up approach

(4) Top down approach
(i) This method starts by building a big, central data warehouse for the entire organization.

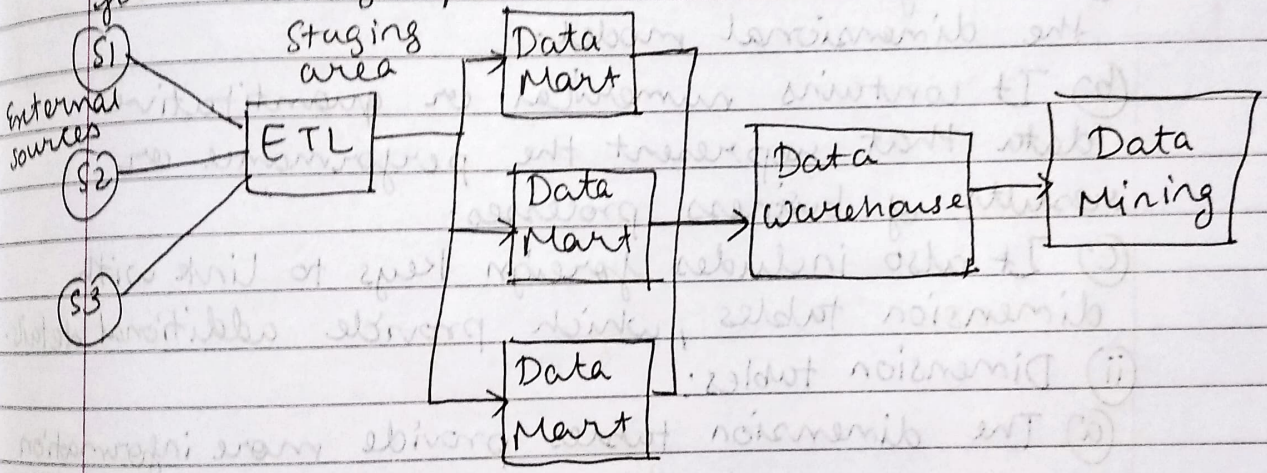
(ii) Then, smalled, focused sections called data marts are created from the main warehouse for specific departments.



(iii) Working
(a) Data is collected from various sources
(b) Data is processed & validated using ETL tools
(1) Extract : Data is collected from sources
(2) Transform: Data is standardized & cleansed
(3) Load: Data is loaded in the data warehouse

(c) After cleaning, the data is stored in single, organized location, data warehouse.
(d) Smaller data marts are created for each department to focus only on their needs.
(e) The company uses reporting tools to get insights, trends and predictions from the data.

(5) Bottom up approach
(i) The method focuses on creating small, seperate data marts first for specific departments.
(ii) These data marts are then connected later to form a larger, centralized data warehouse.



(iii) Working:
(a) Data is collected from various external sources.
(b) Data is processed using ETL tools, similar to the top-down approach.
(c) Individual data marts are built for specific needs
(d) These data marts are then combined to form a bigger, central data warehouse.
(e) Each department can use the data warehouse for analyzing the data.

Page No.
Date

Q5) Dimensional modelling.

=> (1) Dimensional modelling is a way to organize data in a database to make it easy to analyze and understand.

(2) It is mostly used in data warehouses for decision-making and organizing data into a cube for better viewing and analysis.

(3) Dimensional modelling involves two types of tables:

(i) Fact table

(a) The fact table is a central table in the dimensional model.

(b) It contains numerical or quantitative data that represent the performance or results of business processes.

(c) It also includes foreign keys to link with dimension tables, which provide additional details.

(ii) Dimension tables:

(a) The dimension tables provide more information for the facts in the fact table:

(b) It contains descriptive or categorical attributes that helps to explain the measures in the fact table.

(4) Example:

A city and state can view a store summary in a fact table. Item summary, customer information can be viewed in dimension table

Sales (storeID, ItemID, CustID, qty, price)

StoreID (storeID, city, state)

ItemID (ItemID, category, brand, color, size)

CustID (CustID, name, address)

Page No.
Date

Fact table

StoreID	ItemID	CustID	qty	price

StoreID	....

ItemID	....

CustID	....

Dimension tables

Page No.
Date

Q6) Basic building blocks of data warehouse
OR
Components of data warehouse
OR
Data warehouse architecture.

=> (1) A data warehouse is a system that stores data from various sources in one place, making it easier for businesses to analyze and make decisions.
(2) Its components or building blocks work together to collect, organize, store and present data.
(3) The components are:
(i) Data sources
(a) These are the initial points from where data is collected and generated.
(b) Its types:
(1) Operational data: Data from day-to-day business systems.
(2) Internal data: Private data such as reports
(3) Archived data: Older data stored for historical analysis.
(4) External data: Data from outside sources
(ii) ETL Process
(a) Extract: Extracts raw data from various sources using queries.
(b) Transform: Data is standardized and cleaned to remove errors
(c) Load: Save the prepared data in the data warehouse.
(iii) Data storage
(a) The central repository where the cleaned,

processed data is stored.

(b) Components :

(i) Staging area : Temporary storage where raw data is held before processing.

(2) Warehouse storage : The main storage area.

(3) Data marts : Subsections of warehouse.

(iv) Metadata

(a) Metadata describes the structure, content and the meaning of the data stored in warehouse.

(v) Information delivery

(a) This component ensures data is accessible and usable for decision-making.

(b) Eg: A dashboard showing monthly revenue trends

(vi) Data modelling

(a) The process of organizing the data in the warehouse to support efficient querying and analysis.

(b) Common models :

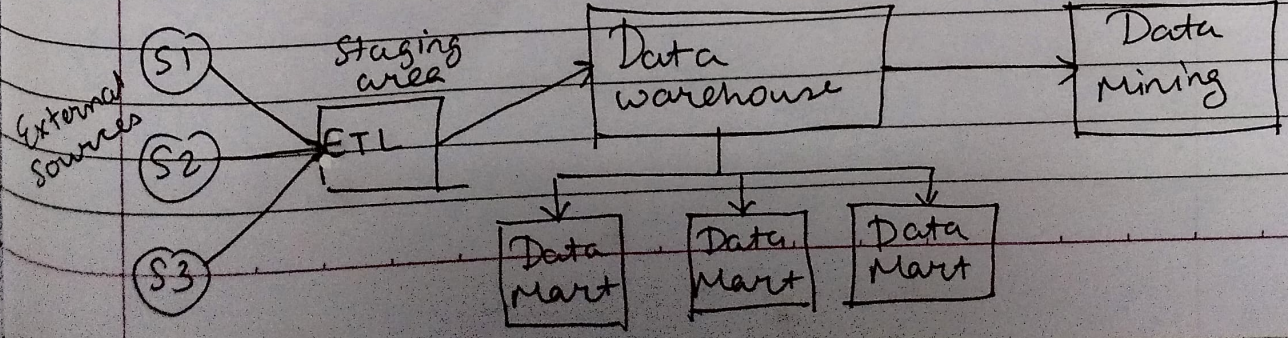
(1) Star schema

(2) Snowflake schema

(vii) Management and control

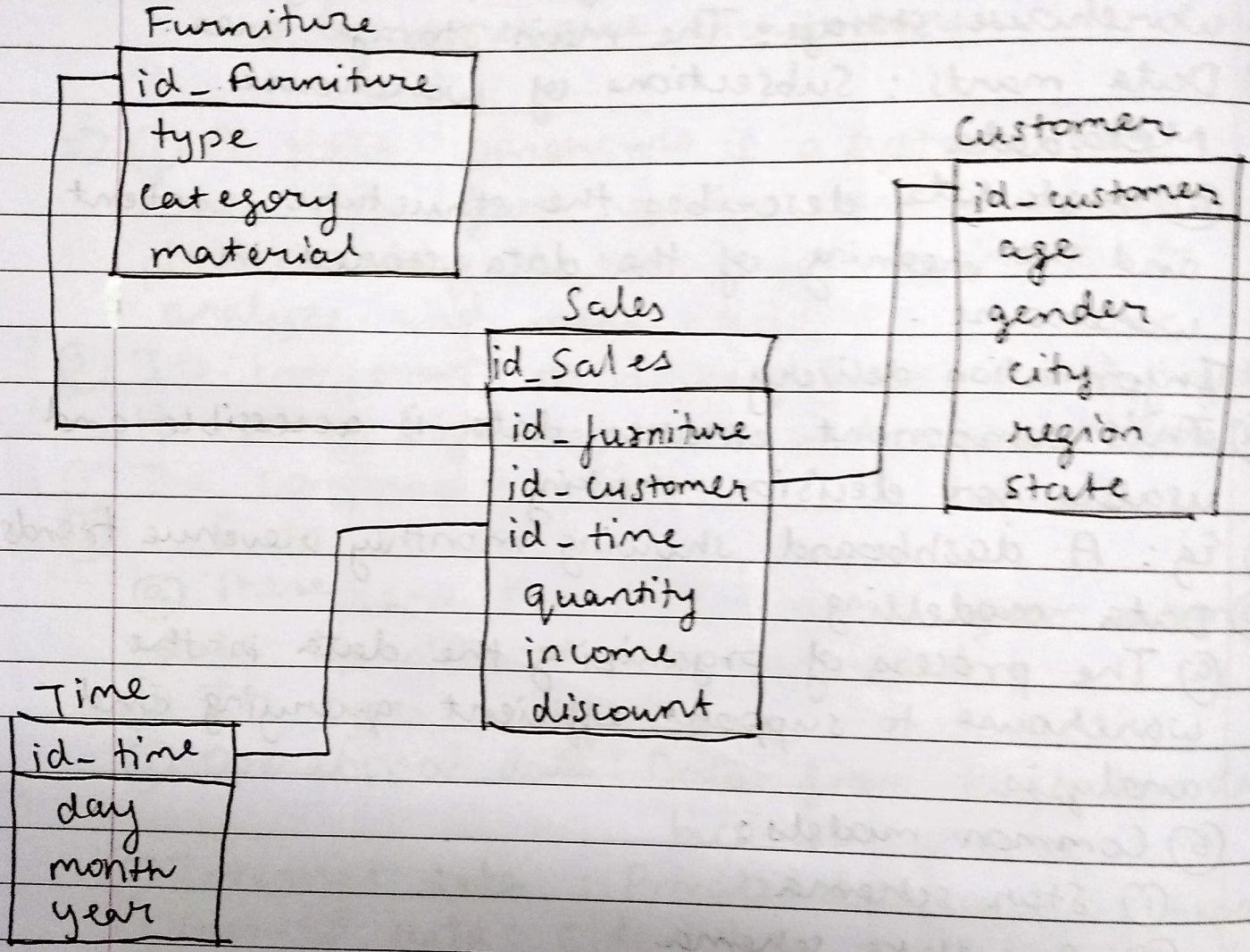
(a) This includes the systems and processes that maintain the warehouse's functionality and security

(b) It schedules ETL tasks, monitors performance, ensures data integrity, security and backups



Q7) Design the data warehouse dimensional model for a wholesale furniture company.

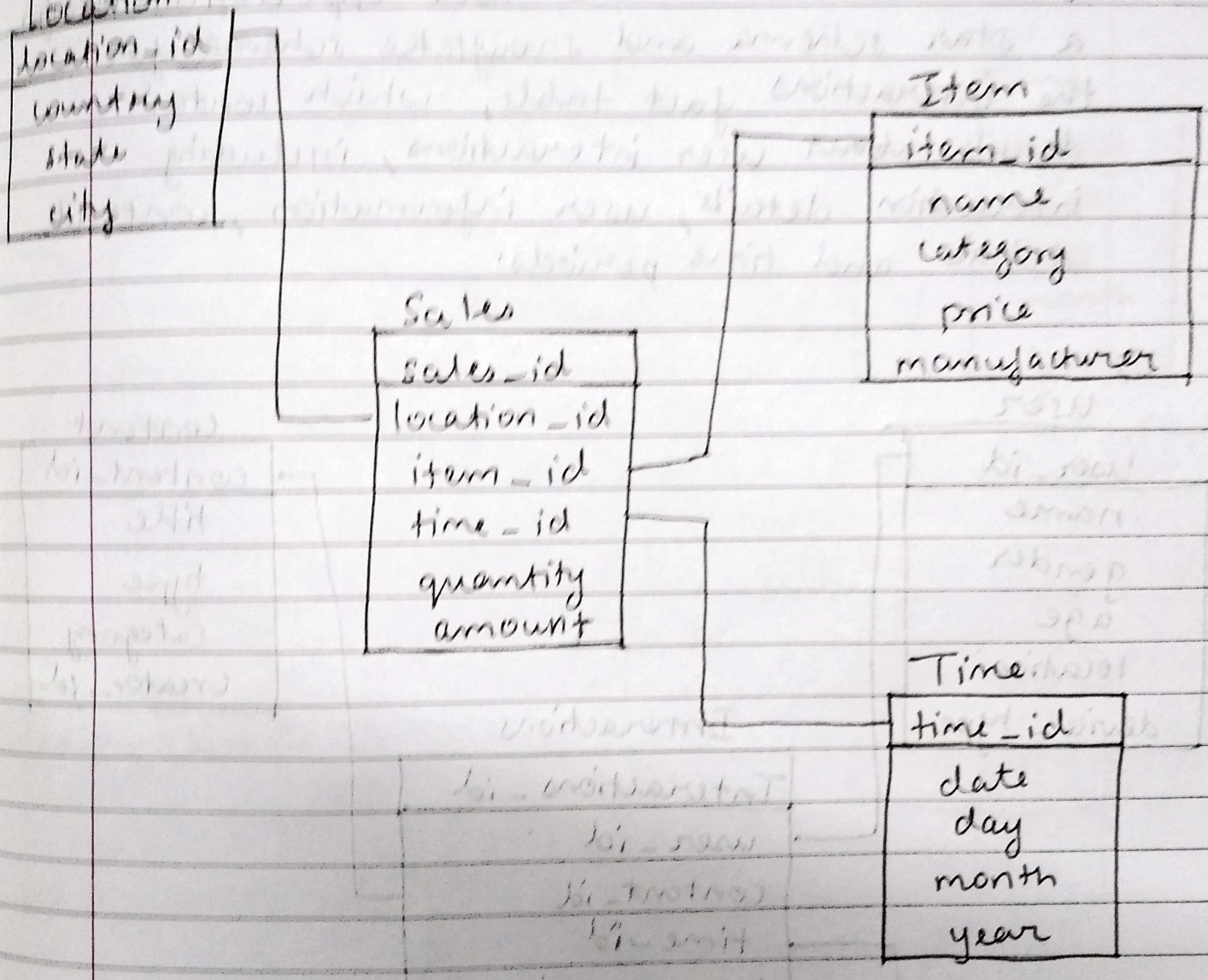
=>



Star schema
for
wholesale furniture company

Q8) Design a star schema for company sales with three dimensions: location, time & item.

=>

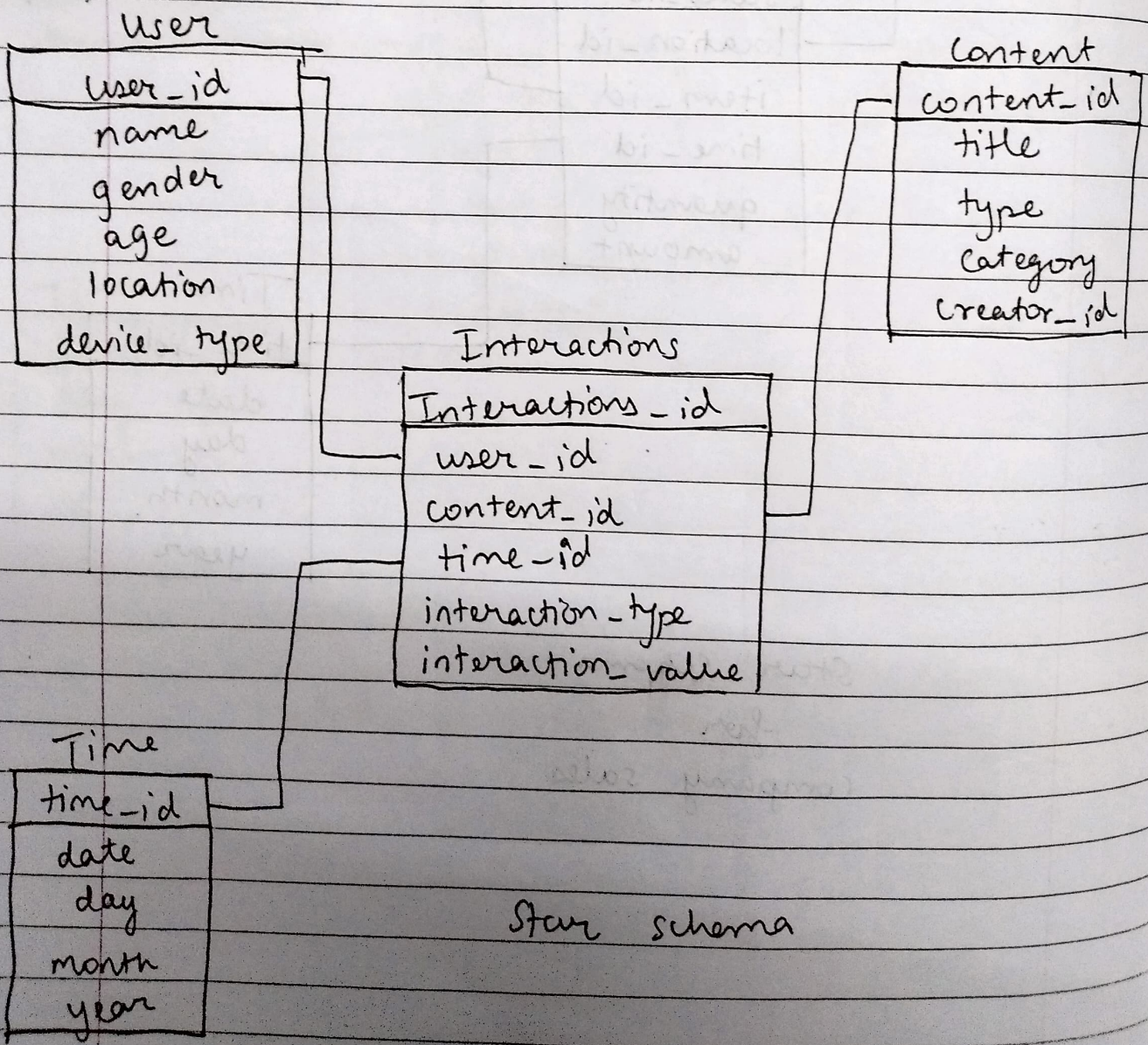


Star schema for company sales

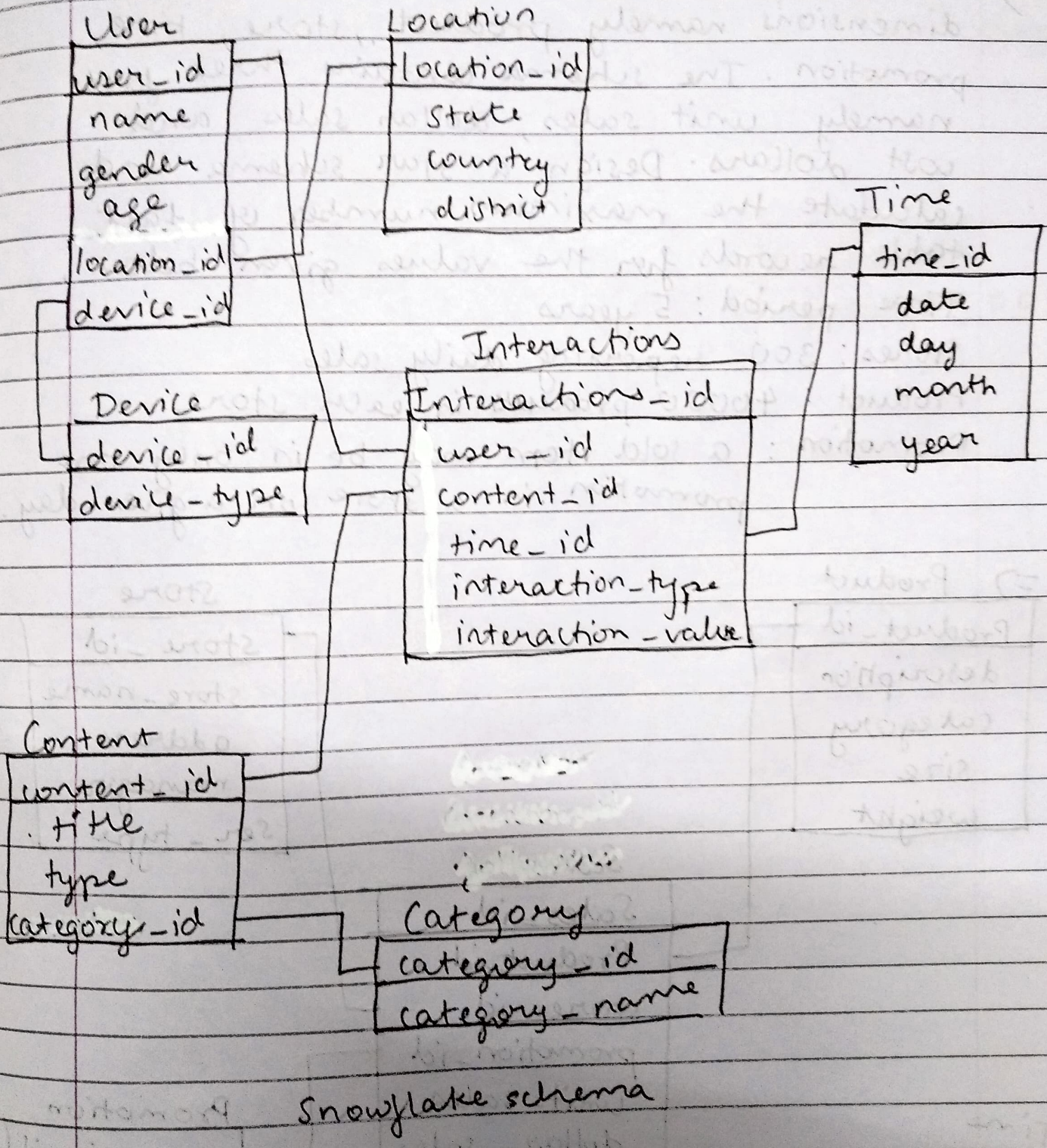
Page No.
Date

Q9) A social media platform wants to analyze user engagement data to improve content recommendations and user experience. Design a star schema and snowflake schema for the interactions fact table, which contain details about user interactions, including interaction details, user information, content details and time periods.

⇒

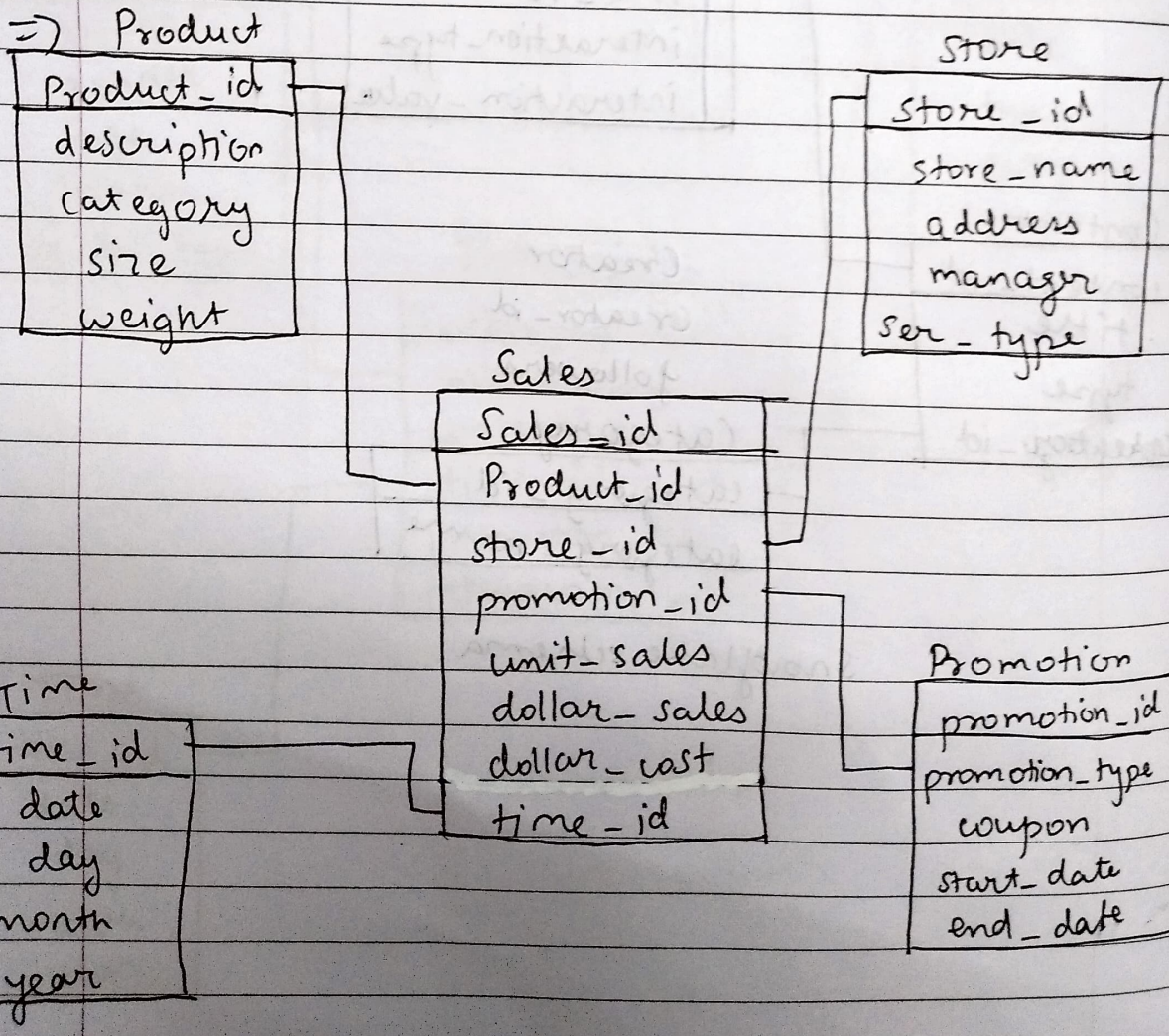


Star schema



Snowflake schema

Q10) For a supermarket chain, consider the dimensions namely product, store, time, promotion. The schema contains three facts namely unit sales, dollar sales and cost dollars. Design a star schema and calculate the maximum number of fact table records for the values given below:
Time period: 5 years
Stores: 300 reporting daily sales
Product: 40000 products in each store
Promotion: a sold item may be in only one promotion in a store on a given day.

=> 

Page No.
Date

Time period = 5 years × 365 days = 1825 days

There are 300 stores
Each store daily sale 4000
Promotion = 1
Maximum number of fact table records
= 1825 × 300 × 4000 × 1 = 2,19,00,00,000