Data-Mining-Series-1-Important-Topics

ⓘ **For more notes visit**

https://rtpnotes.vercel.app

1. Data mining and application

What is Data Mining?

Data mining is the process of extracting useful information or patterns from large sets of data. Think of it like searching for valuable nuggets of gold in a huge pile of rocks.

It is also known as:

- Knowledge Discovery in Databases (KDD)
- Data analysis
- Business intelligence
- Information harvesting

Applications of Data Mining

1. **Classification** – Sorting data into categories.
    - Example: In a loan database, classifying applicants as **eligible or risky** based on their credit history, income, and other details.
2. **Estimation** – Predicting an unknown value based on existing data.
    - Example: Estimating a student's future grades based on their past performance.
3. **Prediction** – Forecasting future outcomes using data.
    - Example: Predicting **next week's stock price** based on previous trends.

4. **Market Basket Analysis (Association Rule Mining)** – Identifying patterns in shopping behavior.
   - Example: If a customer buys **a pen and a pencil**, they are likely to buy **a notebook** too.
5. **Clustering** – Grouping similar data points together when the categories are unknown.
   - Example: Grouping **customers with similar buying habits** for targeted marketing.

Where is Data Mining Used?

- Business intelligence
- Data analytics
- Bioinformatics (e.g., analyzing genetic data)
- Web mining (e.g., improving search engines)
- Text mining (e.g., summarizing large documents)
- Social network analysis (e.g., understanding user interactions)

2. OLAP and OLTP

1. **OLTP (Online Transaction Processing)**: A system that handles **daily transactions** like banking, online shopping, or airline bookings. It is designed for **fast and frequent updates** to data, ensuring accuracy and quick responses.
2. **OLAP (Online Analytical Processing)**: A system that is used for **analyzing large amounts of historical data** to help businesses make decisions. It focuses on **data trends, summaries, and complex queries** rather than quick updates. Examples include sales forecasting and business intelligence reports.

Difference Between OLTP and OLAP

Feature	OLTP (Online Transaction Processing)	OLAP (Online Analytical Processing)
1. Users & Purpose	Used by **clerks, clients, IT professionals** for day-to-day transactions and queries.	Used by **managers, executives, analysts** for decision-making and data analysis.
2. Data Type	Stores **current** and frequently updated data.	Stores **historical** data for analysis, summaries, and trends.

Feature	OLTP (Online Transaction Processing)	OLAP (Online Analytical Processing)
3. Database Design	Uses **Entity-Relationship (ER) models** with an application-oriented design.	Uses **Star or Snowflake schema** with a subject-oriented design.
4. Data Scope	Focuses on **real-time data** for an enterprise or department.	Includes **historical data** from multiple versions and organizations.
5. Storage & Access	Small-sized transactions, needs fast processing, concurrency control, and recovery.	Large, complex queries that are mostly **read-only** and analyzed in a multidimensional way.

- **OLTP** is for quick transactions (e.g., banking, order processing).
- **OLAP** is for analyzing large amounts of data (e.g., sales trends, business intelligence).

3. Data Preprocessing

What is Data Preprocessing?

Data preprocessing is the process of **preparing raw data** for analysis by cleaning, transforming, and organizing it. Since real-world data is often **incomplete, inconsistent, or contains errors**, preprocessing helps improve **data quality** and ensures better results in data mining.

Why is Data Preprocessing Important?

Good data leads to accurate insights. Poor-quality data can result in incorrect conclusions. High-quality data should be:

- **Accurate** (free from errors)
- **Complete** (no missing values)
- **Consistent** (data format is uniform)
- **Timely** (up-to-date)
- **Believable** (trustworthy)
- **Interpretable** (easy to understand)

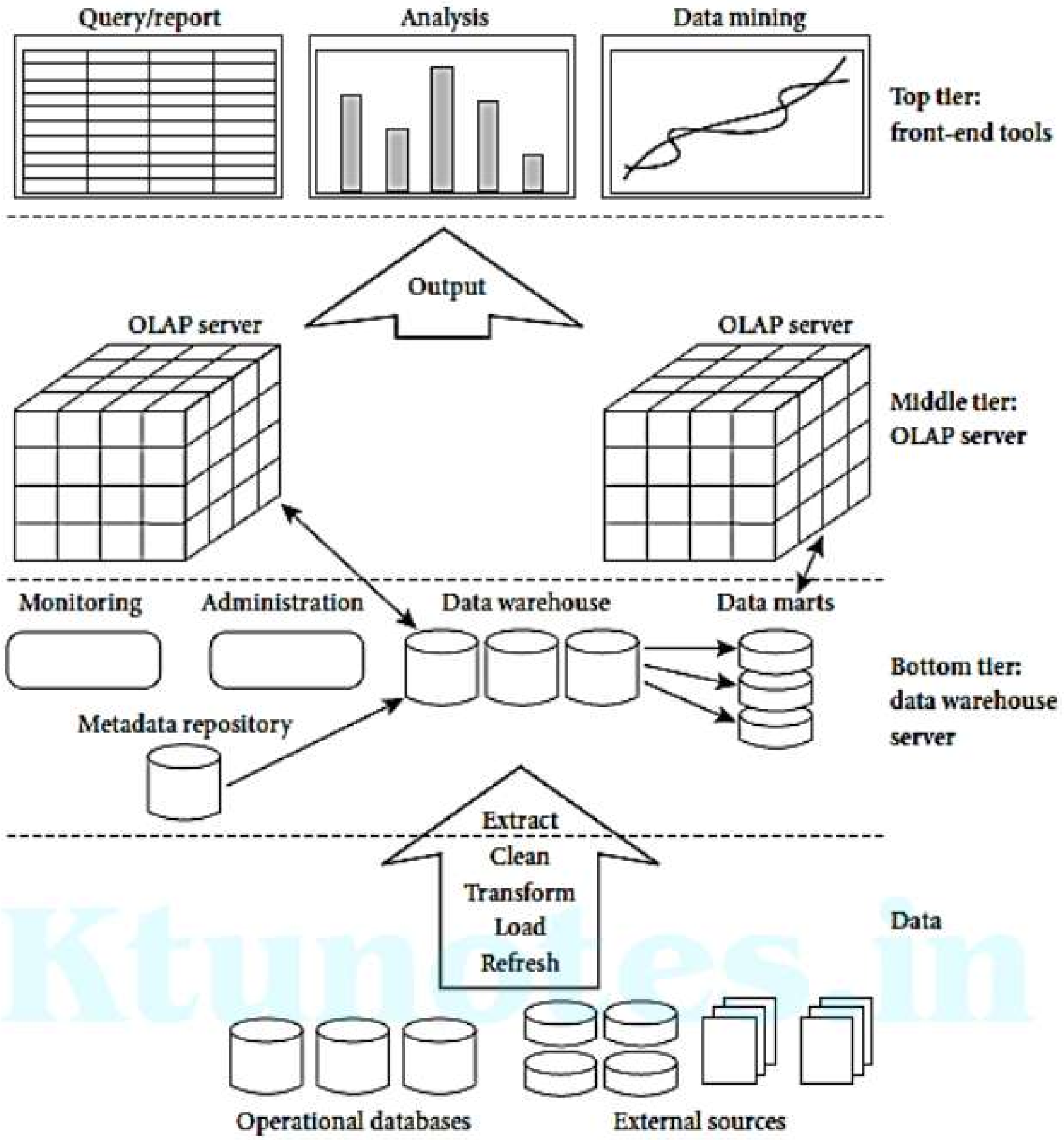
Key Steps in Data Preprocessing

1. **Data Cleaning** – Fixing errors and removing noise.
   - Example: Filling in missing values or correcting typos in names and dates.
2. **Data Integration** – Combining data from multiple sources into a single dataset.
   - Example: Merging customer data from an online store and a physical store into one system.
3. **Data Reduction** – Making data smaller without losing important information.
   - Example: Removing duplicate records or summarizing data to reduce its size.
4. **Data Transformation** – Converting data into a format suitable for analysis.
   - Example: **Normalization**, which scales data values to fit within a range (e.g., 0 to 1) to improve the performance of machine learning models.

These steps often work **together** to improve the quality and efficiency of data mining. For example, data cleaning may also involve transformation, like converting all date formats to a single standard.

4. Three tier architecture with figure

A Three-Tier Data Warehouse Architecture



A **three-tier data warehouse** organizes data processing into three levels: **Bottom Tier (Storage), Middle Tier (Processing), and Top Tier (User Interaction).**

1. Bottom Tier – Data Storage & Management

- This is the **database layer**, usually a **relational database (RDBMS)**.
- It collects data from different **sources** (e.g., operational databases, external systems).
- **Tasks performed:**
  - **Data extraction:** Pulling data from multiple sources.
  - **Data cleaning & transformation:** Merging and formatting data.
  - **Loading & refreshing:** Keeping the warehouse up-to-date.

- **Tools used:**
  - **Gateways** (connects applications to databases) – Examples:
    - **ODBC** (Open Database Connection)
    - **JDBC** (Java Database Connection)
    - **OLEDB** (Microsoft's database linking tool)
  - **Metadata repository** (stores details about data and structure).

2. Middle Tier – OLAP Server (Processing & Analysis)

- This layer processes and organizes data for fast retrieval.
- Uses **OLAP (Online Analytical Processing) Servers**, which come in two types:
  1. **ROLAP (Relational OLAP):** Uses relational databases to store and process multidimensional data.
  2. **MOLAP (Multidimensional OLAP):** Uses a specialized database that directly handles multidimensional data for faster processing.

3. Top Tier – User Interaction (Front-End Tools)

- This is the **interface layer**, where users access and analyze data.
- Includes **query tools, reporting tools, and data mining tools** to identify trends, patterns, and predictions.
- Examples of tasks:
  - Running **reports** (e.g., sales performance summaries).
  - Performing **trend analysis** (e.g., predicting next quarter's sales).

📌 **Bottom Tier** – Stores and manages raw data (RDBMS).
📌 **Middle Tier** – Processes data using OLAP (ROLAP or MOLAP).
📌 **Top Tier** – Provides user-friendly tools for data analysis and decision-making.

5. KDD

What is KDD?

KDD (Knowledge Discovery in Databases) is a **step-by-step process** of finding useful insights and patterns in large datasets. **Data Mining** is just one step in this process.

KDD vs. Data Mining

- **KDD** is the **entire process** of discovering patterns and knowledge.
- **Data Mining** is a **step within KDD** that applies algorithms to extract meaningful patterns.

Steps in KDD

1. **Data Cleaning** – Remove errors, noise, and inconsistencies.
2. **Data Integration** – Combine data from different sources.
3. **Data Selection** – Pick relevant data for analysis.
4. **Data Transformation** – Convert data into a suitable format (e.g., summarizing or normalizing values).
5. **Data Mining** – Use algorithms to find patterns and trends.
6. **Pattern Evaluation** – Identify the most useful patterns.
7. **Knowledge Presentation** – Show results using charts, graphs, and reports.

Alternative Names for KDD

- Data mining
- Knowledge extraction
- Data analysis
- Information harvesting
- Business intelligence

6. SRSWOR & SRSWR

What is Sampling in Data Reduction?

Sampling helps reduce the size of a large dataset by selecting a smaller **representative subset**. This makes data processing faster while still keeping important patterns intact.

Two Main Types of Simple Random Sampling

SRSWOR (Simple Random Sampling Without Replacement)

- Each selected data point **is NOT returned** after being picked.
- Once a data point is chosen, it **cannot appear again** in the sample.
- Example: Drawing lottery numbers where each number is unique.

SRSWR (Simple Random Sampling With Replacement)

- Each selected data point **is returned** to the dataset after being picked.
- A data point **can appear multiple times** in the sample.
- Example: Rolling a die multiple times—each roll is independent of the previous one.

7. Min-Max and Z-score normalization (problem)

Example

- Use the two methods below to normalise the following group of data
  - 1000, 2000, 3000, 5000, 9000
- Min-Max Normalization by setting Min=0 and Max=1
- Z-Score Normalization

Min-Max Normalization

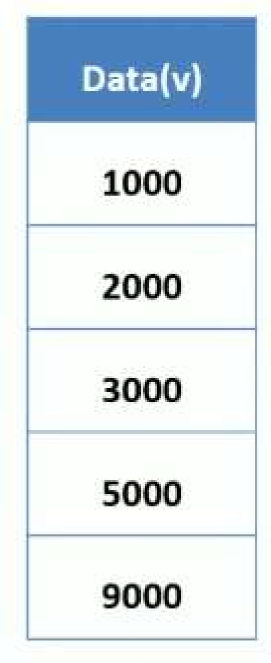
Normalization formula

$$V = \frac{x - min}{max - min}$$

- Here x is the value to be normalized
- Min is the minimum value
- Max is the maximum value

For example:

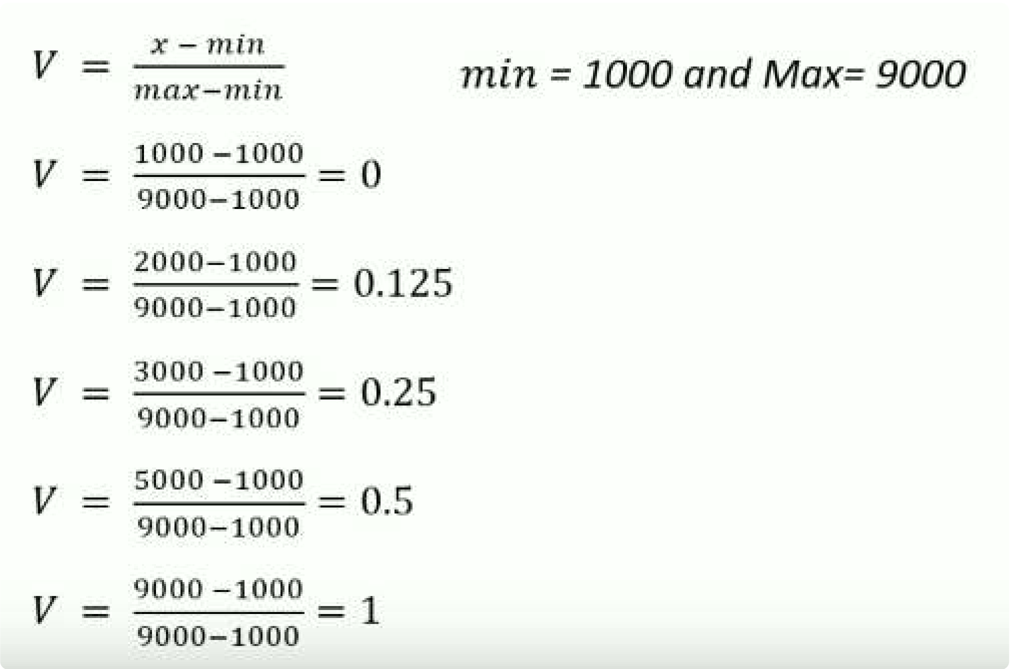
We have the following data as per the question



Data(v)
1000
2000
3000
5000
9000

For each, 1000,2000,3000,5000,9000 we will apply this formula

Let min and max be 1000 and 9000



$$V = \frac{x - min}{max - min}$$

min = 1000 and Max= 9000

$$V = \frac{1000 - 1000}{9000 - 1000} = 0$$

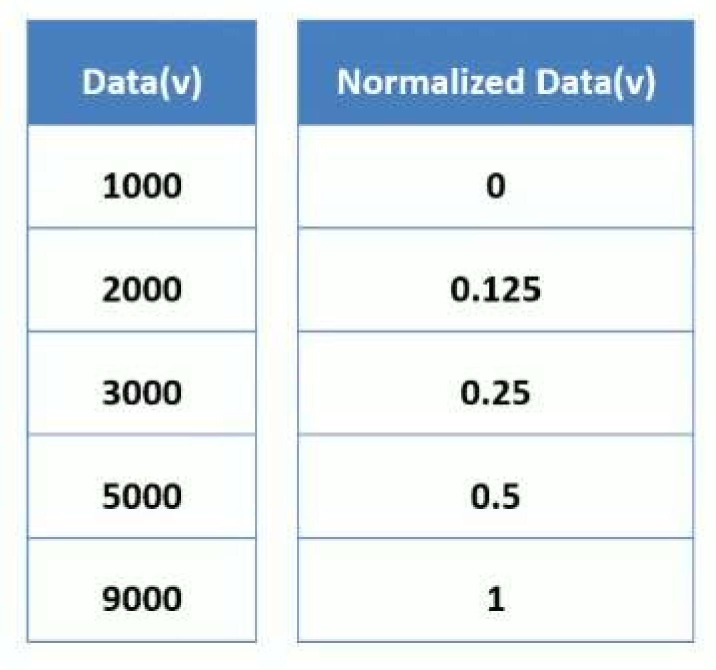
$$V = \frac{2000 - 1000}{9000 - 1000} = 0.125$$

$$V = \frac{3000 - 1000}{9000 - 1000} = 0.25$$

$$V = \frac{5000 - 1000}{9000 - 1000} = 0.5$$

$$V = \frac{9000 - 1000}{9000 - 1000} = 1$$

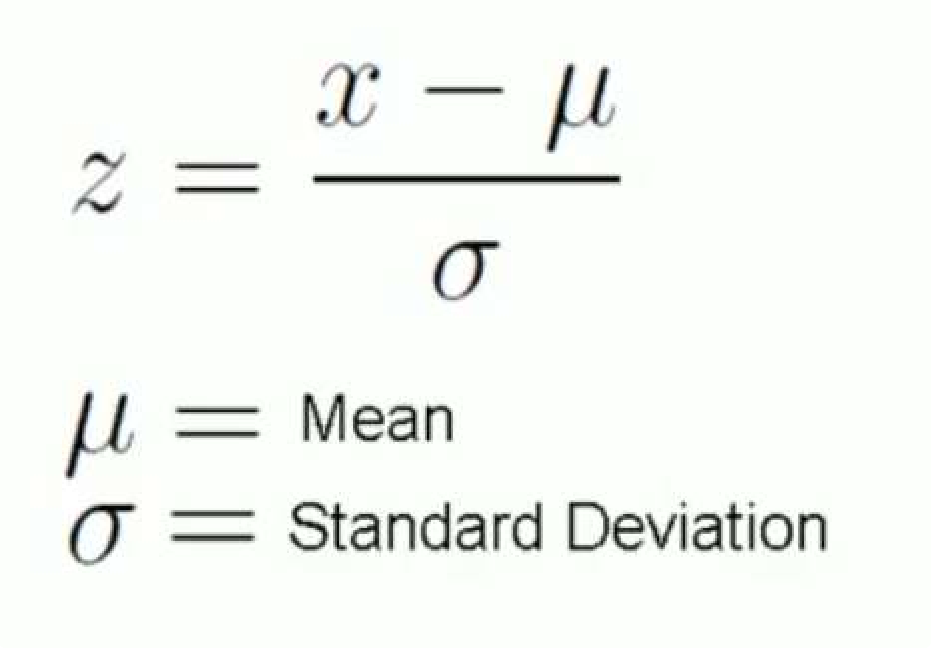
So we get our normalized data as



Data(v)	Normalized Data(v)
1000	0
2000	0.125
3000	0.25
5000	0.5
9000	1

Z-Score Normalization

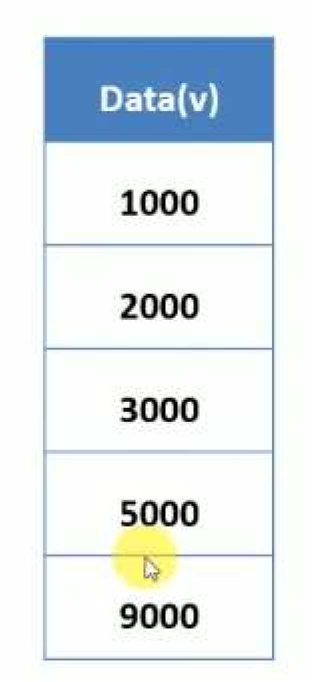
Normalization formula



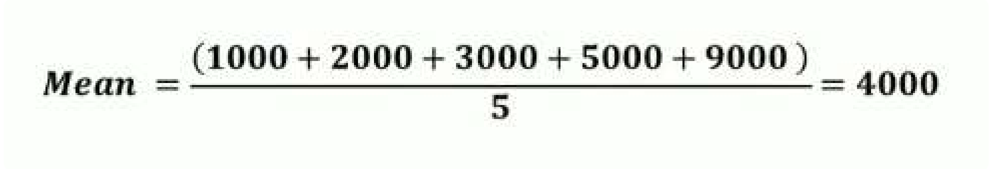
$$z = \frac{x - \mu}{\sigma}$$

$$\mu = \text{Mean}$$
$$\sigma = \text{Standard Deviation}$$

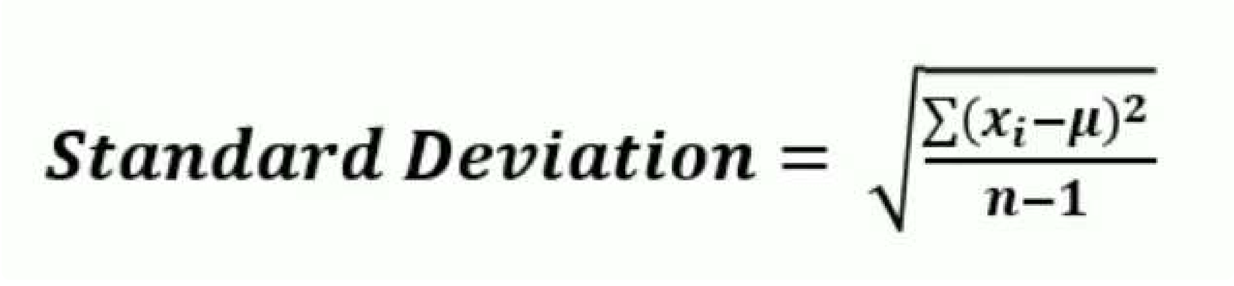
- Lets take the previous example

- 

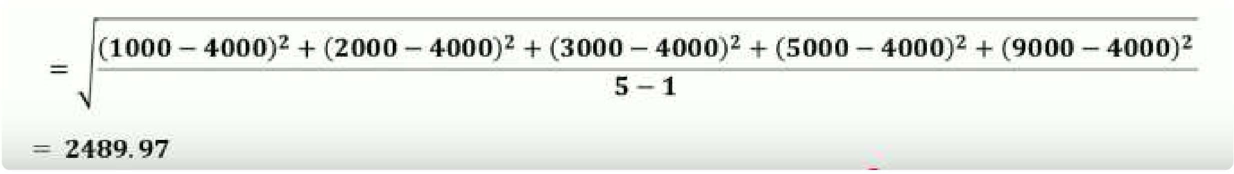
Data(v)
1000
2000
3000
5000
9000

- Calculating mean, we get
  - 

$$Mean = \frac{(1000 + 2000 + 3000 + 5000 + 9000\ )}{5} = 4000$$

- Formula for Standard Deviation
  - 

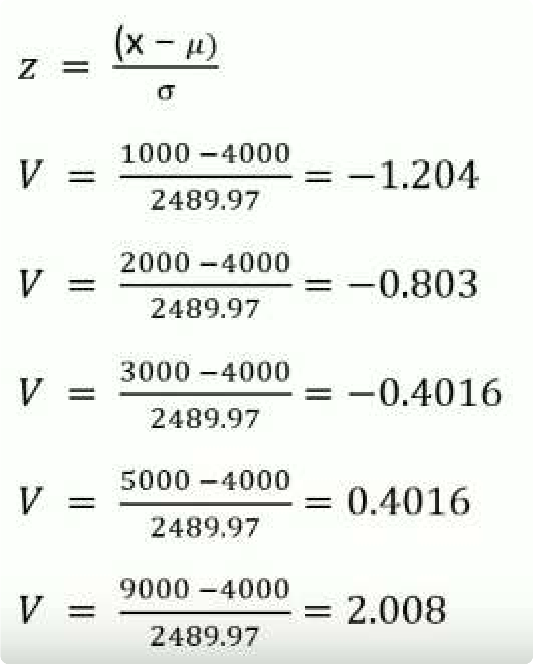
$$Standard\ Deviation = \sqrt{\frac{\sum(x_i - \mu)^2}{n-1}}$$

  - 

$$= \sqrt{\frac{(1000-4000)^2 + (2000-4000)^2 + (3000-4000)^2 + (5000-4000)^2 + (9000-4000)^2}{5-1}}$$

$$= 2489.97$$

Applying the values in the formula and getting all the normalized values



$$z = \frac{(x - \mu)}{\sigma}$$

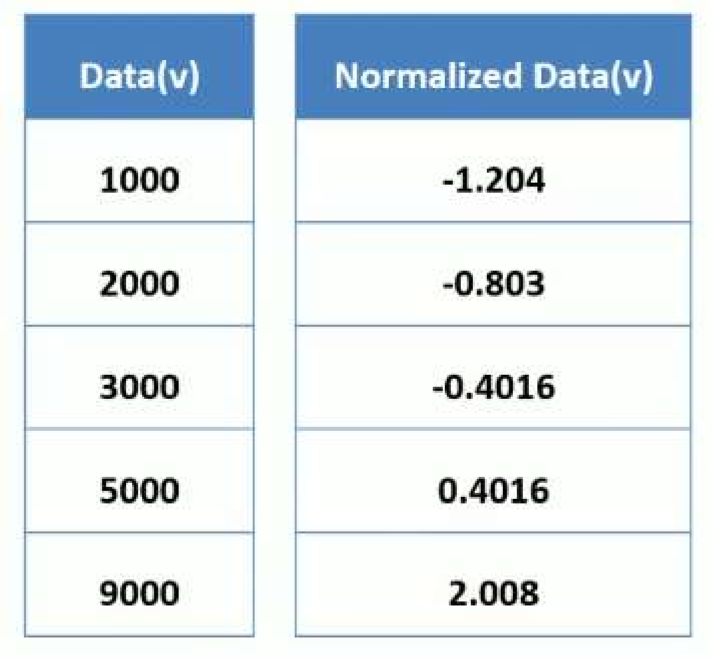
$$V = \frac{1000 - 4000}{2489.97} = -1.204$$

$$V = \frac{2000 - 4000}{2489.97} = -0.803$$

$$V = \frac{3000 - 4000}{2489.97} = -0.4016$$

$$V = \frac{5000 - 4000}{2489.97} = 0.4016$$

$$V = \frac{9000 - 4000}{2489.97} = 2.008$$



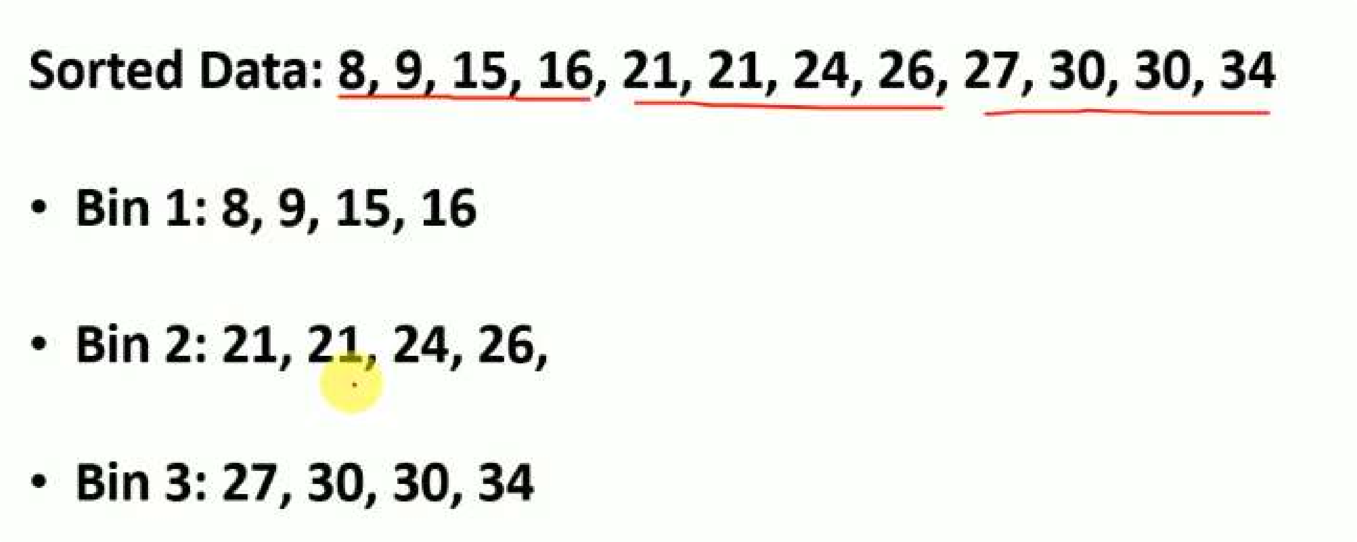
Data(v)	Normalized Data(v)
1000	-1.204
2000	-0.803
3000	-0.4016
5000	0.4016
9000	2.008

8. Smoothing by means (problem)

- The binning method can be used for smoothing the data
- Mostly data is full of noise
- Data smoothing is a data preprocessing technique used to remove the noise from the data set

Data Smoothing by Equal Frequency Bins

- Example
    - Unsorted data for price in dollars
    - Before sorting
        - 8,16,9,15,21,21,24,30,26,27,30,34
    - After Sorting
        - 8,9,15,16,21,21,24,26,27,30,30,34
- Suppose Bin size = 4
- The number of data points is 12 (there are 12 numbers)
- 12 / 4 = 3 bins
- So we divide into 3 equal pieces

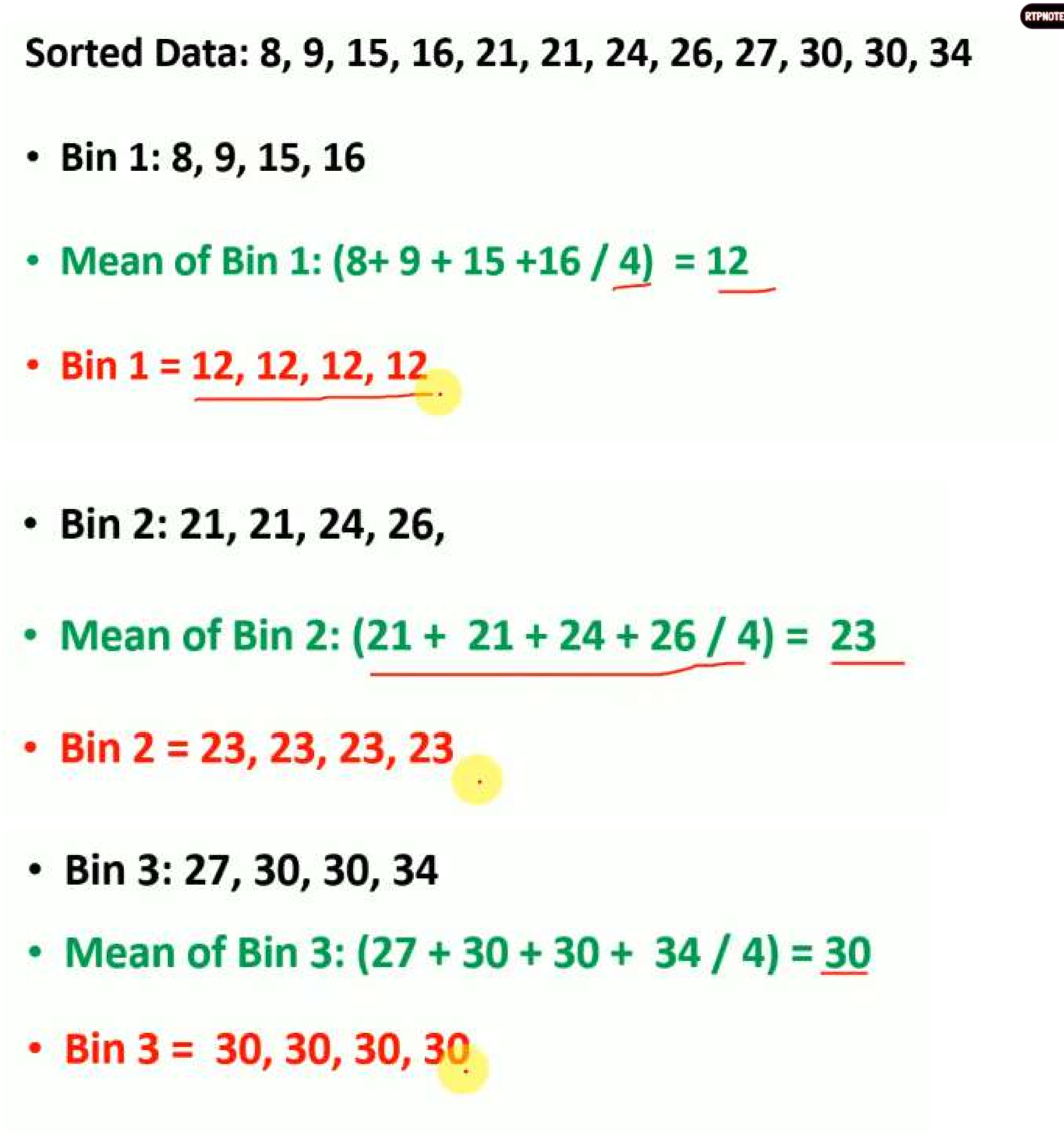


Sorted Data: 8, 9, 15, 16, 21, 21, 24, 26, 27, 30, 30, 34

- Bin 1: 8, 9, 15, 16
- Bin 2: 21, 21, 24, 26,
- Bin 3: 27, 30, 30, 34

Data Smoothing by Bin Means

- We will do the same steps as earlier
- After we get the bins, we will calculate the mean of each bin
    - Assign the mean to all the data points in the bin



Sorted Data: 8, 9, 15, 16, 21, 21, 24, 26, 27, 30, 30, 34

- Bin 1: 8, 9, 15, 16
- Mean of Bin 1: (8+ 9 + 15 +16 / 4) = 12
- Bin 1 = 12, 12, 12, 12

- Bin 2: 21, 21, 24, 26,
- Mean of Bin 2: (21 + 21 + 24 + 26 / 4) = 23
- Bin 2 = 23, 23, 23, 23

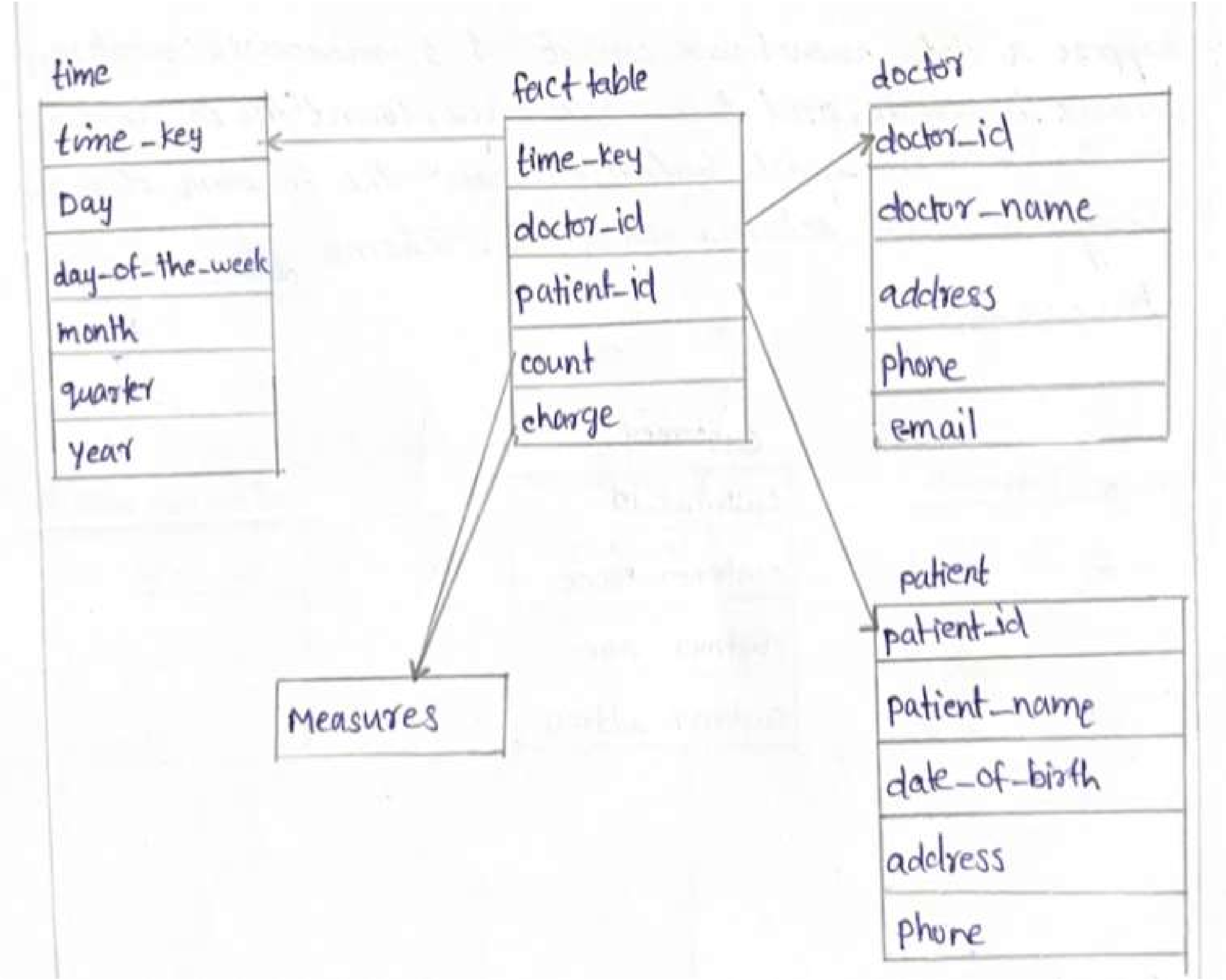
- Bin 3: 27, 30, 30, 34
- Mean of Bin 3: (27 + 30 + 30 + 34 / 4) = 30
- Bin 3 = 30, 30, 30, 30

9. Problem related to schema.

Problem 1

Suppose that a data warehouse consists of 3 dimensions, time, doctor and patient and 2 measure count and charges, where charge is the fee that a doctor charges a patient for a visit

a) Draw a schema diagram for the above data warehouse using one of the schema (Star, snowflake, constellation)



b) Starting with the base cuboid [day, doctor, patient] What specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?

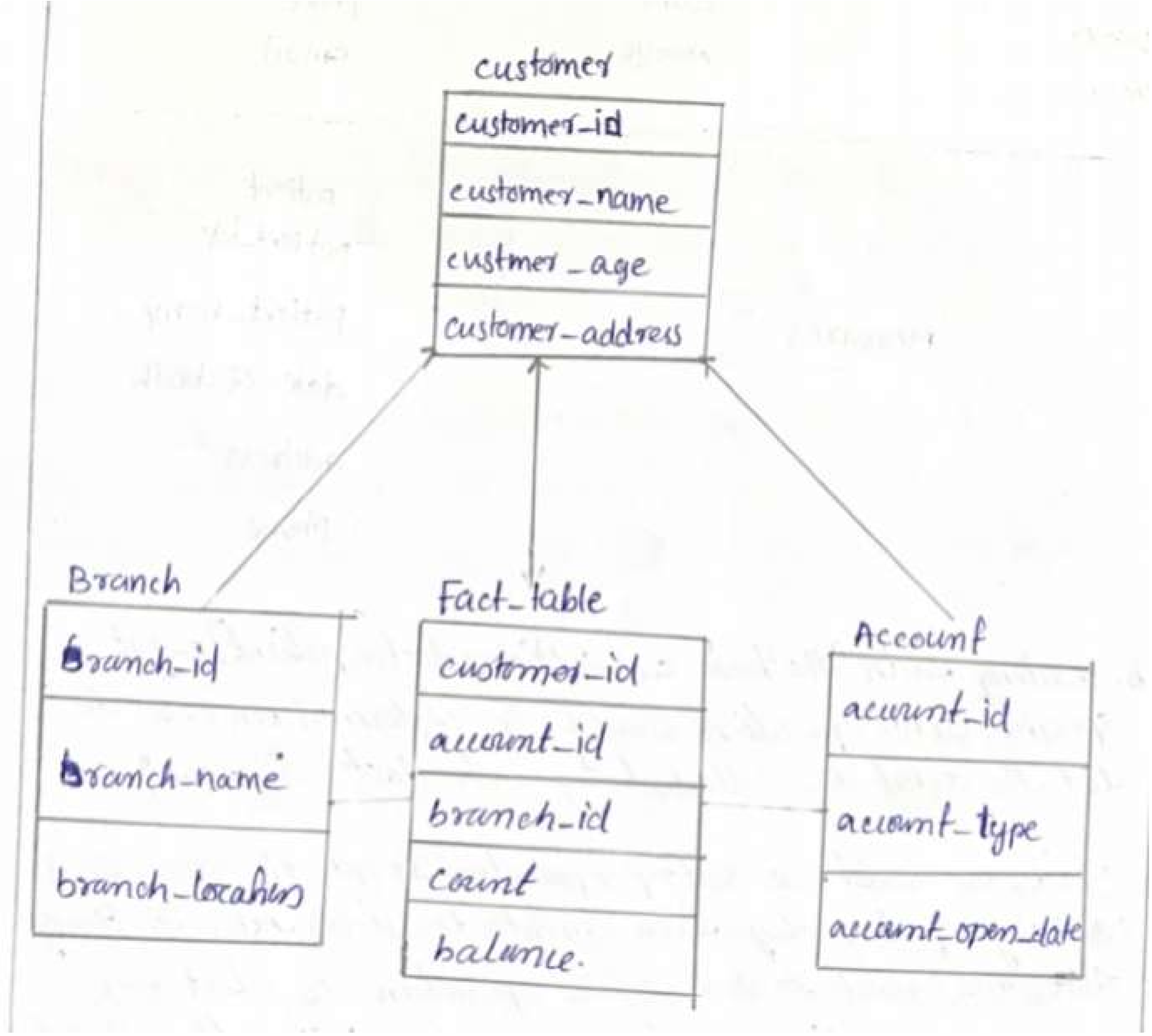
First we should use rollup operation to get the year 2004 (rolling up from day then month to year). After getting that, we need to use slice operation to select one need to use slice operation to sellect call. Finally we get list the total collected by each dcotor in 2004.

1. Roll up from day to month to year
2. Slice for year = "2004"
3. Roll upon patient from individual patient to all
4. Slice for patient = "all"
5. Get the list of total fee collected by each doctor in 2004

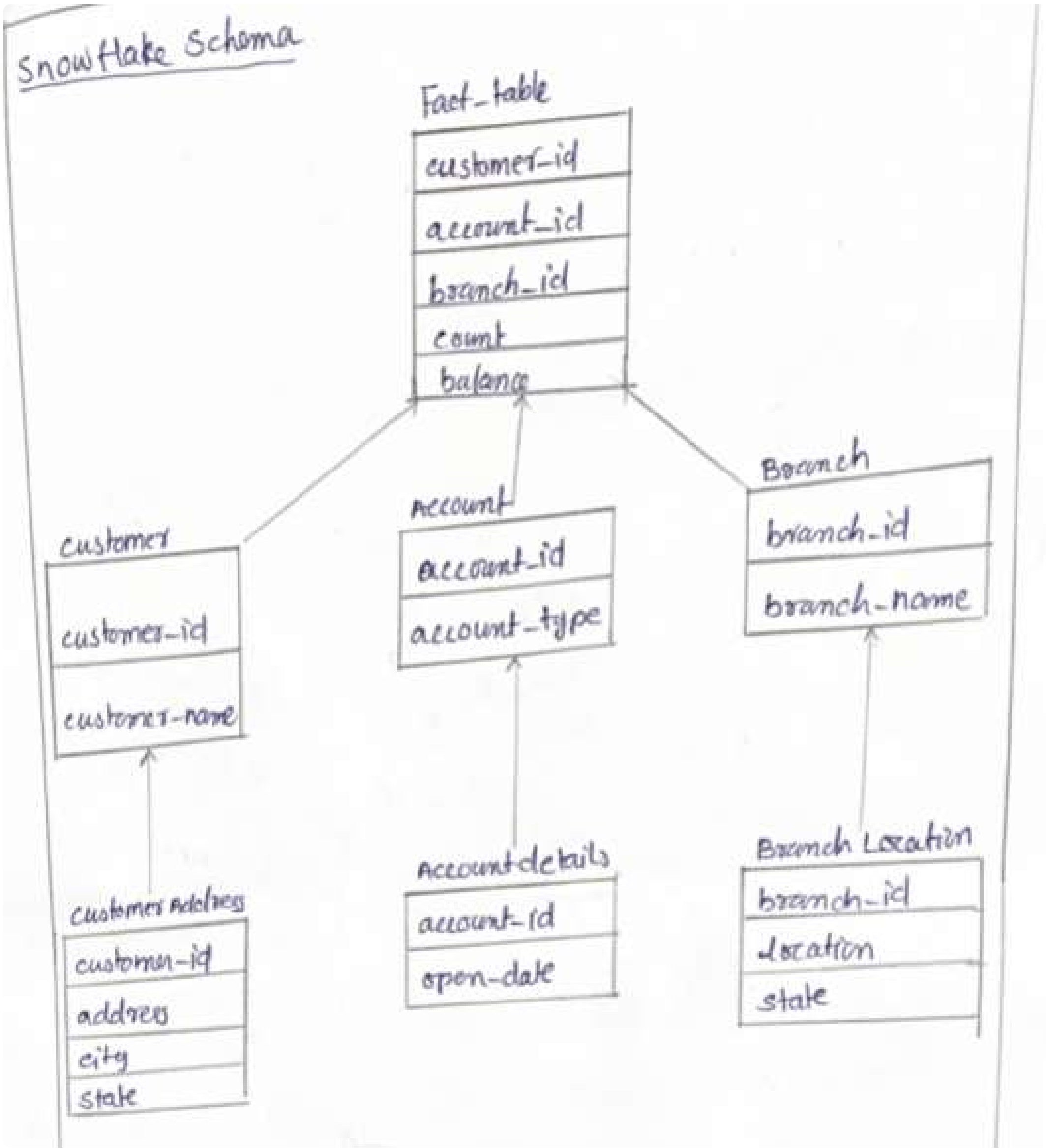
Problem 2

Suppose a data warehouse consists of 3 dimensions: Customer, account, branch, and two measures, count (no of customers on the branch and balance) draw the schema diagram using snowflake schema and star-schema

Star Schema



Snowflake Schema



10. Missing data dealing method

1 **Ignore the Tuple**

- Skip rows with missing values.
- Works **only** if the missing data is small and random.
- Not ideal if missing data is common.

2 **Fill Missing Values Manually**

- A person manually fills in missing data.
- **Very time-consuming** and not practical for large datasets.

3 Use a Global Constant

- Replace missing values with a fixed value like **"Unknown"** or **-∞**.
- **Problem:** The model may treat "Unknown" as a meaningful category.

4 Use Mean (Average) or Median

- Replace missing values with the **mean or median** of that attribute.
- **Example:** If the average income is **$56,000**, missing values for income are replaced with **$56,000**.
- **Good when data is normally distributed.**

5 Use Class-Specific Mean or Median

- Instead of the global mean, use the mean of similar groups.
- **Example:** If classifying customers by **credit risk**, use the **average income** of people in the same credit category.
- **More accurate than using the global mean.**

6 Use the Most Probable Value (Prediction-Based)

- Predict missing values using **regression, Bayesian models, or decision trees**.
- **Example:** A decision tree predicts missing **income** based on other factors like age, education, and job title.
- **Best accuracy but requires more computation.**