Data Mart:-

→ A data mart is a specialized subset of a data warehouse focused on a specific functional area or department within an organization.

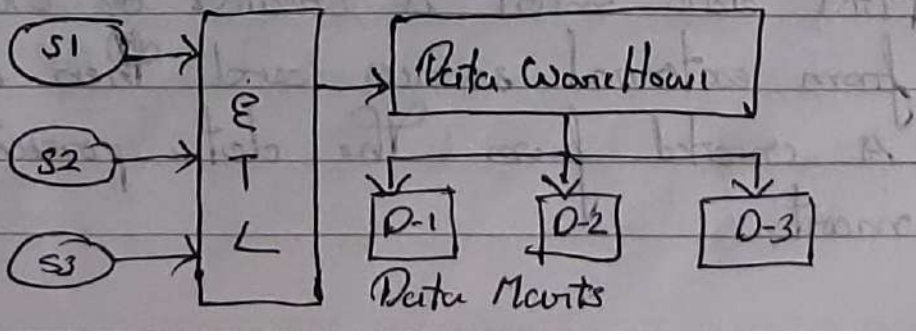
→ It provides a simplified and targeted view of data addressing specific reporting and analytical needs

→ They are organized around specific subjects, such as sales, customer data, or product information and are structured, transformed and optimized for efficient querying and analysis within the domain.

Types of Data Mart:-

i Dependent Data Mart
ii Independent Data Mart
iii Hybrid Data Mart

Dependent Data Mart:-



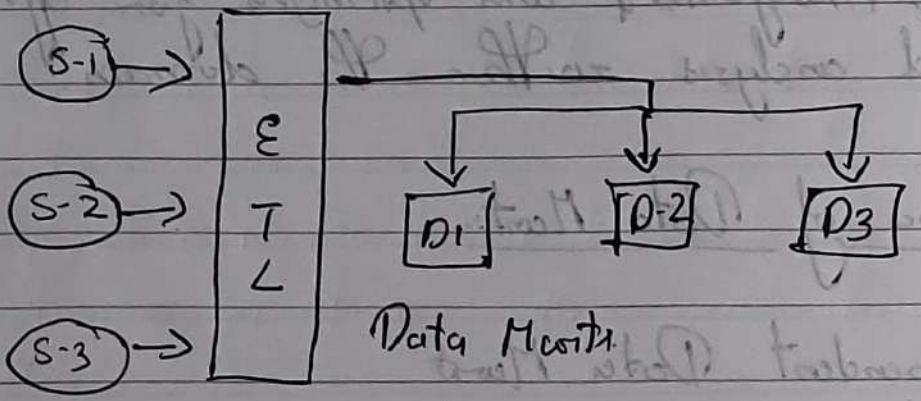
→ Created by extracting data from central repository i.e Datawarehouse

→ First data warehouse is created by extracting data (through ETL Tool) from external sources and then data mart is created from data warehouse.

⇒ It uses the top-down approach.

→ Used by big organizations

Independent Data Mart

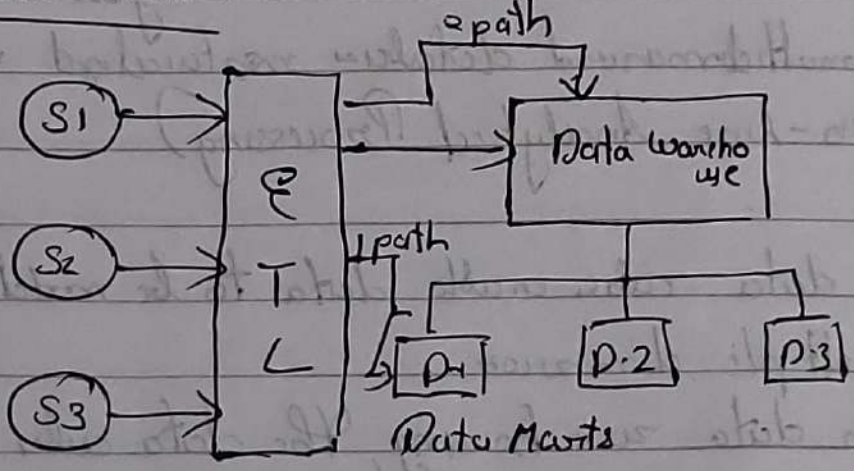


- Created directly from external sources instead of data warehouses.
- First data mart is created by extracting data from external sources and then data warehouse is created from the data present in data mart.

-> Uses Bottom up approach.

-> Used by Smaller organizations

Hybrid Data Mart:



- Created by extracting data from operational source or from data warehouse
- Path 1 reflects accessing data directly from external source and path 2 reflects dependent data model of data mart.

MultiDimensional data Modelling

It is a data modeling technique used in data warehouses to organize data in the database in an efficient manner to analyze future trends and pattern.

Data Cube:-

→ When data is grouped or combined in multidimensional matrices, it is called Data Cube.

→ The data cube method has a few alternative name like a multidimensional database, materialized view and OLAP (On-line Analytical Processing)

→ A data cube enable data to be modeled and viewed in multiple dimension.

⇒ In data warehousing, the data cubes are n-dimensional

→ The cuboid which holds the lowest level of summarization is called a base cuboid.

→ The topmost 0-D cuboid, which holds the highest level of summarization, is known as the apex cuboid.

→ In this example, this is the total sales or dollars sold, summarized over all four dimensions.

→ Data cube is represented by dimensions and facts

→ Each dimension can have a table related to it and is known as a dimension table.

→ Eg:- Dimension Table for an item can include the attributes items name, Brand and type.

-> A multidimensional data model is generally organized around a central design called facts

-> The fact table includes the names of the facts or measures and keys to each of the associated dimension table.

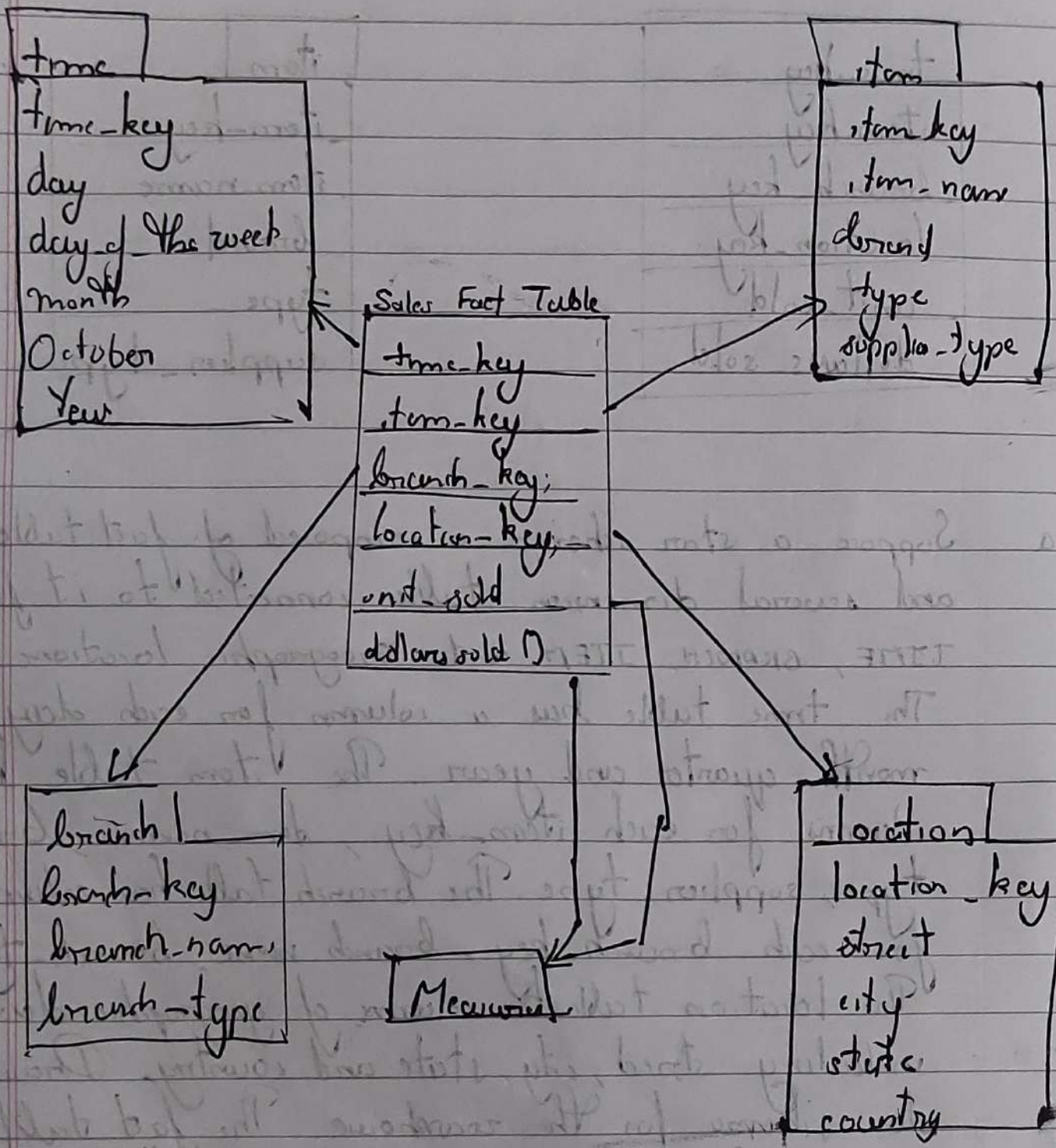
Fact Table - Sales

time_key
item_key
branch_key
location_key
unit_sold
dollars_sold

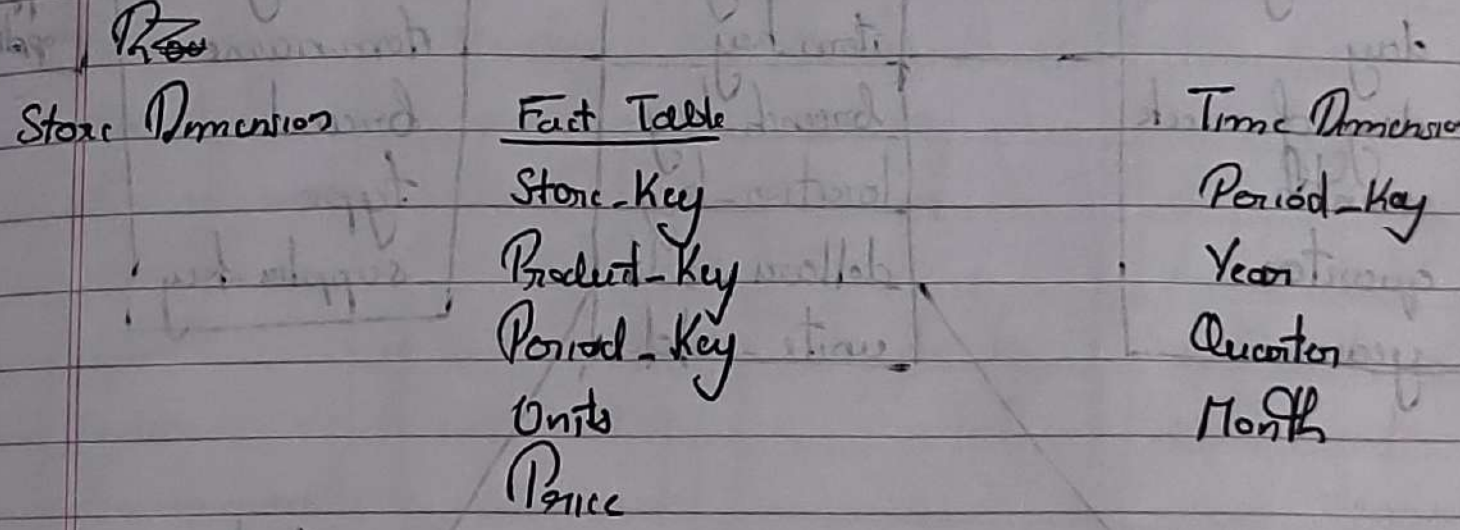
Dimension Table - Item

item
item_key
item_name
brand
type
supplier_type

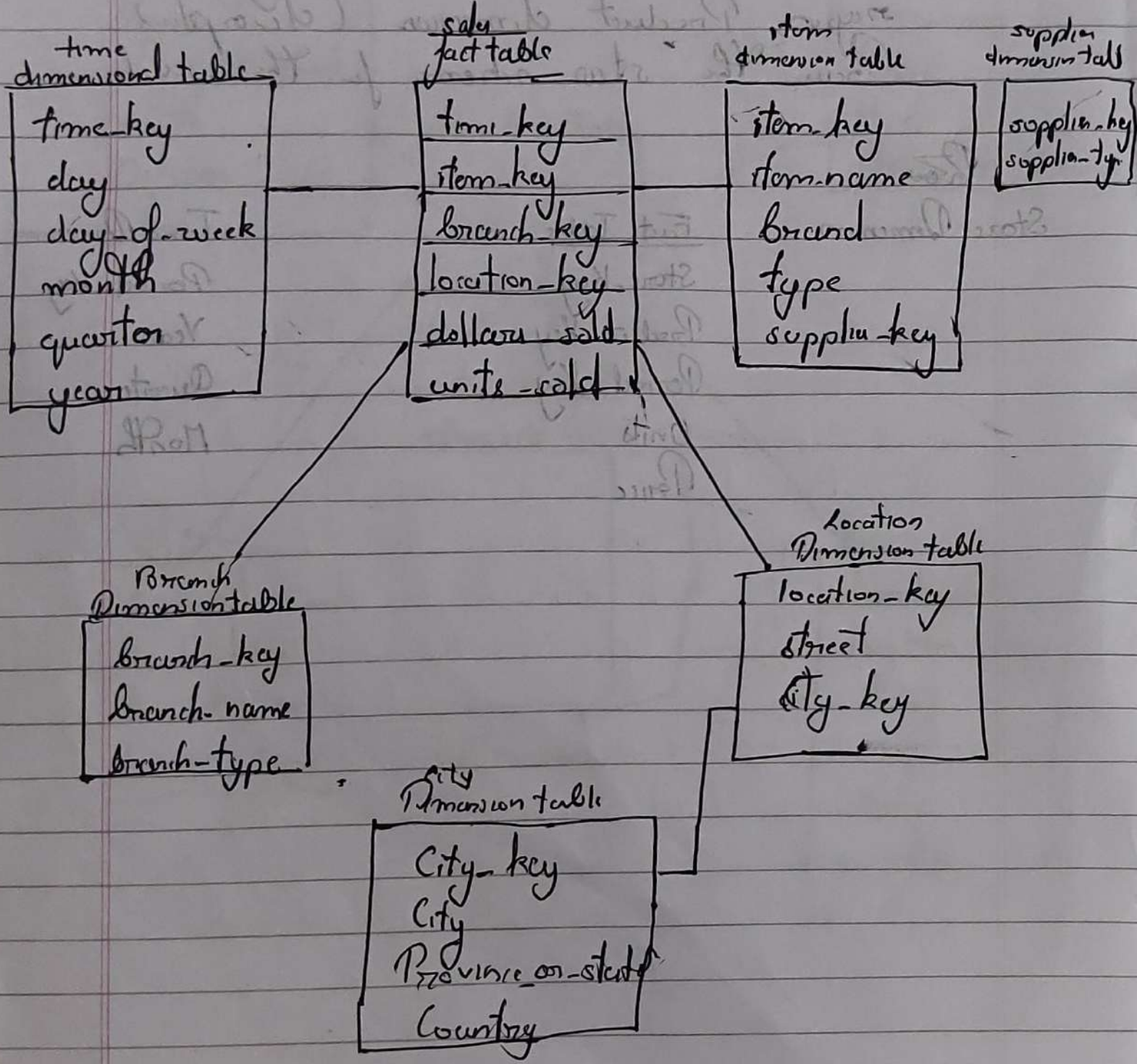
a Suppose a star schema is composed of fact table, sales and several dimension tables connected to it for TIME, BRANCH, ITEM and Geographic locations. The time table has a column for each day, month, quarter and year. The item table has columns for each item_key, item_name, brand, type, supplier_type. The branch table has columns for each branch_key, branch_name, branch_type. The location table has columns of geographic data including street, city, state and country. Draw the star schema for the warehouse. The fact table contains two measures units_sold and dollars_sold.


time
time_key
day
day_of_the_week
month
October
Year
Sales Fact Table
time_key
item_key
branch_key
location_key
unit_sold
dollars_sold
item
item_key
item_name
brand
type
supplier_type
branch
branch_key
branch_name
branch_type
location
location_key
street
city
state
country
Measures

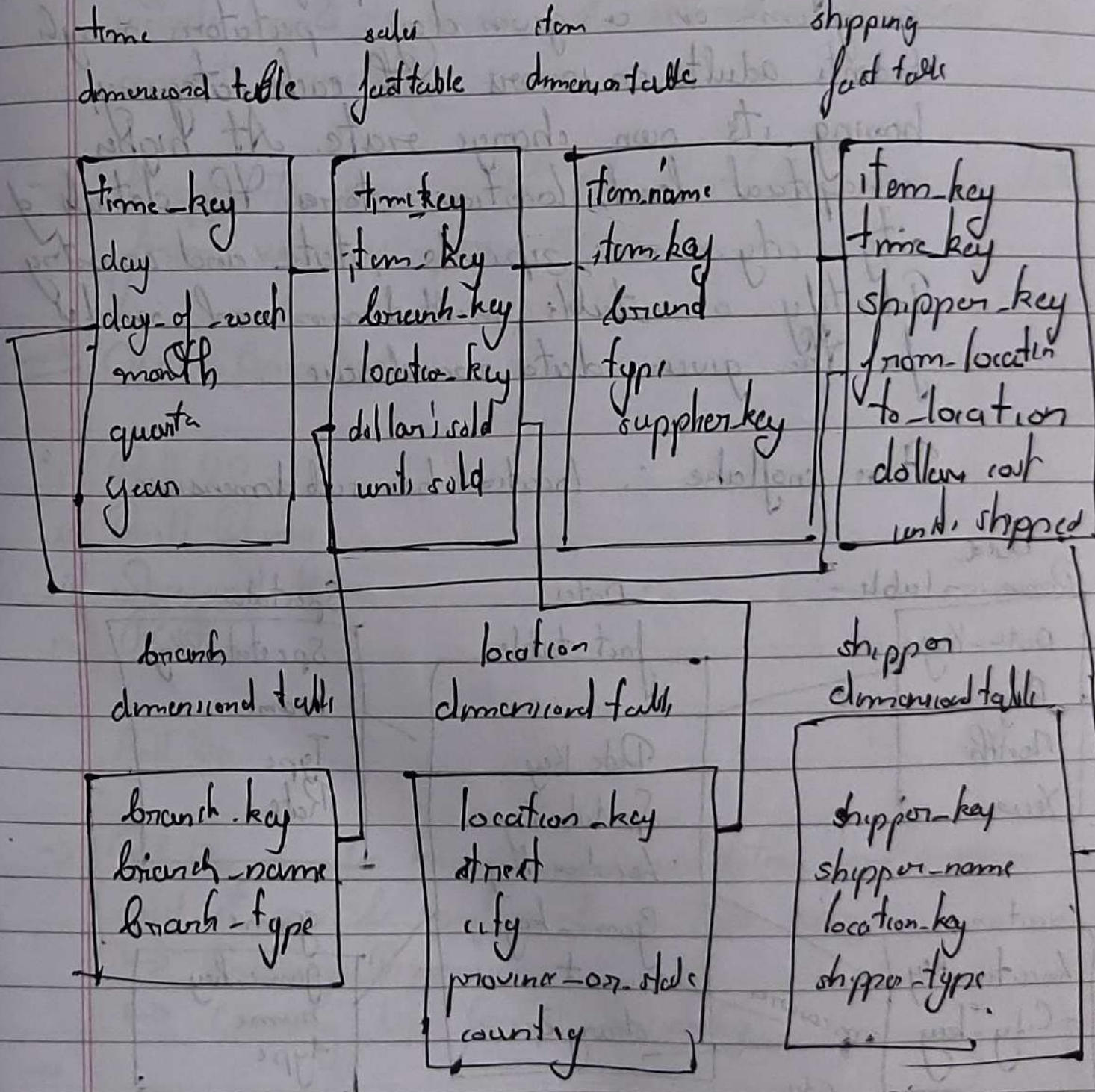
a A data warehouse for a shop consists of the following 3 dimensions. Store, time, product with 2 measures units and price. Time dimension (year, quarter, month) store dimension store name city state region, Product dimension (description) Draw the star schema for the warehouse.



* Snowflake Schema has only one fact table and multiple dimension tables. The dimension tables are connected to sub dimension tables if required.

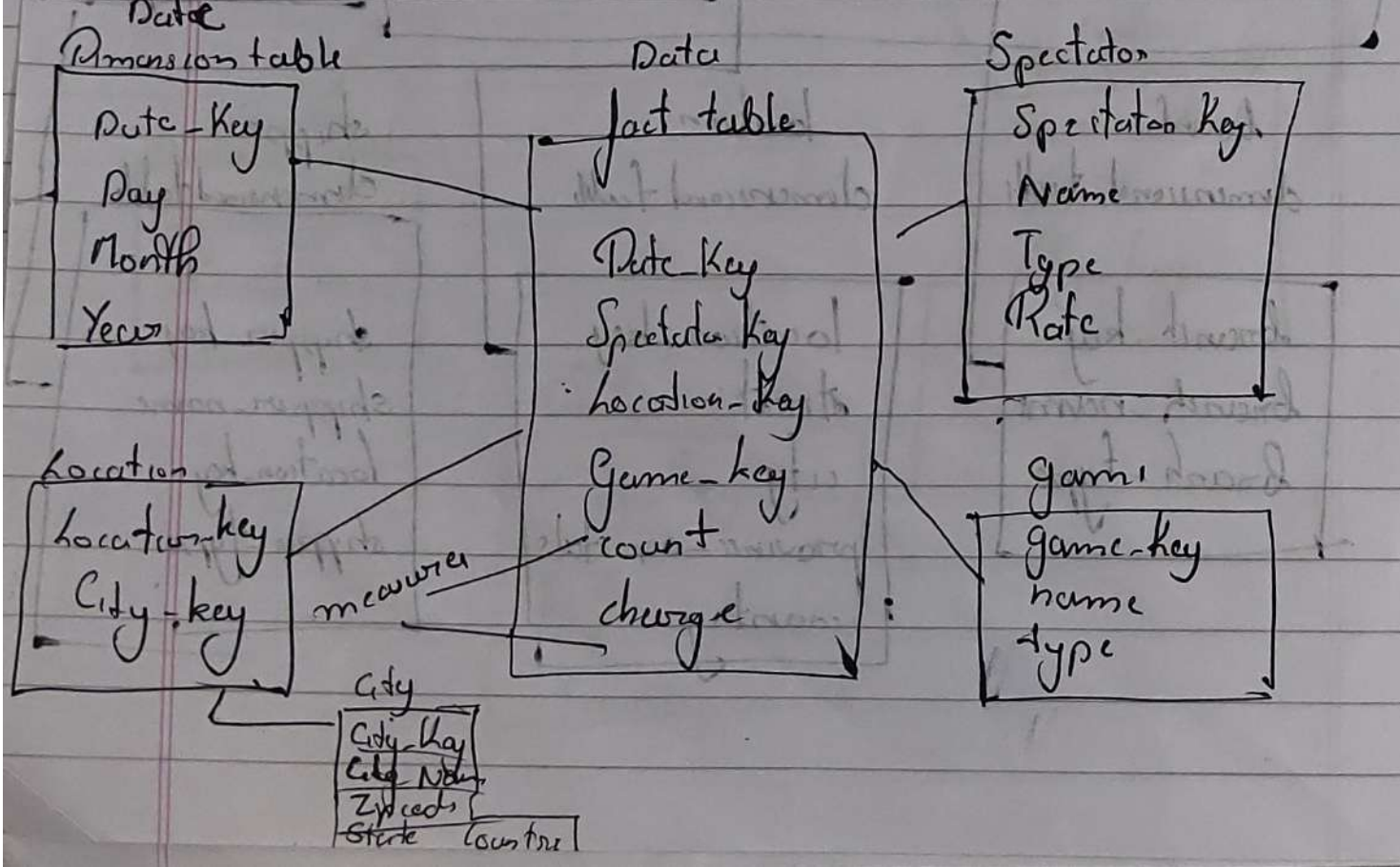


* Fact Constellation has multiple fact table with dimension table and the dimension table share the same packet.



a Suppose a data warehouse consists of 4 dimensions date, spectator, location and game with two measures count and charge where charge is the fare that a spectator pays when watching a game on a given day. Spectators may be students, adults or seniors with each category having its own charge rate. At higher conceptual level, location stores the details of city, city_name, zipcode, state and country. Identify a suitable multidimensional model for the given data warehouse

It is snowflake : location has sub dimension



OLAP

- OLAP stands for online analytical processing
- In the multidimensional model, the records are organized into various dimensions, and each dimension includes multiple levels of abstraction described by concept hierarchy.

- It supports users with the flexibility to view data from various perspectives.

marks **OLAP Operations**

i Roll Up:-
ii Drill Down
iii Slice and Dice
iv Pivot

Roll Up

- Roll Up is also known as consolidation or aggregation and it can be performed in two ways:
  - Reducing Dimensions
  - Climbing Up Concept Hierarchy

- Concept Hierarchy is a system of grouping things

Based on their order or level.

Eg:- Roll up on location from cities to countries
↳ New Jersey and Los Angeles are rolled up into country USA

Module 3

Machine learning

Machine learning is a growing technology that enables computers to learn automatically from past data.

It uses various algorithms for building mathematical models and making predictions using historical data or information

The term machine learning was introduced by Arthur Samuel in 1959.

Def → Machine learning enables a machine to automatically learn from data improve performance from experiences and predict things.

Types of ML

i Supervised learning
ii Unsupervised learning
iii Reinforcement learning

Supervised Learning:

Supervised machine learning is based on supervision.

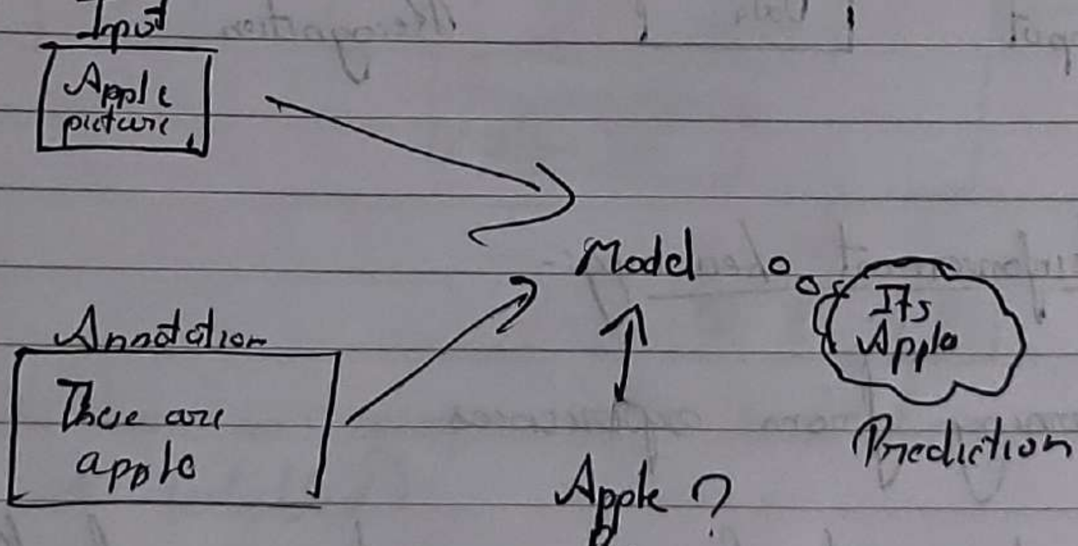
In this technique, we train the machine using the "labelled" dataset and based on the training the machine predicts the outcome.

The main goal of the supervised learning technique is to map the input variable (x) with output variable y.

Supervised learning Types:

i Classification

ii Regression



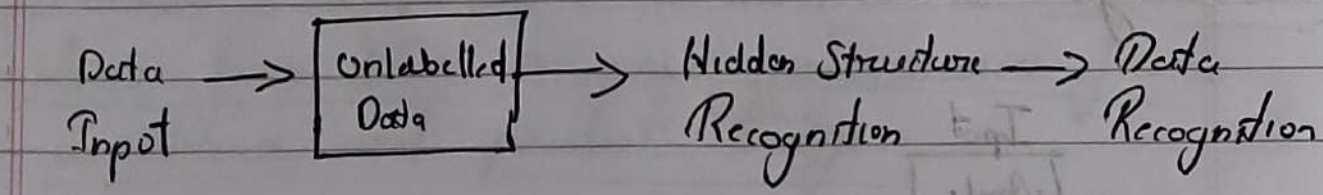
Unsupervised learning :-

→ There is no supervision

→ The machine is trained using unlabelled dataset and the machine predicts the output.

→ Aim of this algorithm is used to group or categorize the dataset according to similarities, pattern and differences.

Two Type
i Clustering
ii Association



Reinforcement learning :-

→ learning from experiences

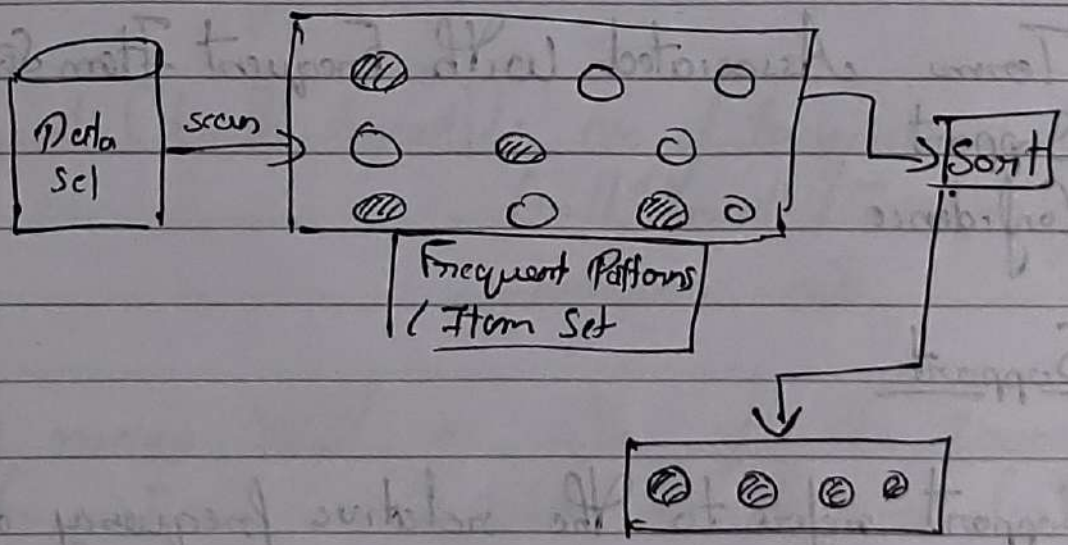
→ Reinforcement learning works on a feedback based process in which an AI agent (software

component) automatically explore its surroundings, learning from experiences and improving its performance.

→ The agent gets rewarded for each good action and get punished for each bad action, hence the goal of reinforcement learning agent is to maximize the rewards.

Frequent Pattern Mining :-

Frequent pattern refers to itemset, subsequences or substructures that appear frequently in a dataset.



Solution (This is an unsupervised dataset because the data is not labelled)

Applications

-> Market Basket Analysis:

- In DM, it is to analyze the combination of products which has been bought together
- It is used by retailers to increase sales by better understanding customer purchasing patterns.

It involves analyzing large data sets, such as purchase history to reveal product groupings, as well as products that are likely to be purchased together.

- Terms Associated with Frequent Item Set
  - i Support
  - ii Confidence

Support:

- Support refers to the relative frequency of an item in a dataset.
- The support of an itemset is the number of transaction in which the itemset appears, divided

by the total number of transactions

Support is calculated as follows

-> Support (X) = (Number of transactions containing X) / Total no of transactions.

where X is the item set for which we are calculating support.

For example, suppose we have a dataset of 1000 transactions and the itemset {milk, bread} appears in 100 of those transactions.

The support of the itemset {milk, bread} would be calculated as follows

Support ({milk, bread}) = No of transactions containing {milk, bread} / Total no of transaction

= 100/1000 = 10%

This means that in 10% of the transactions, the items milk and bread were both purchased

Algorithms that can be used for classification using frequent pattern mining

i Apriori Algorithm
ii FP Growth Algorithm

a Consider the dataset

Transaction id	items
T1	hot dogs, buns, ketchup
T2	hot dogs, buns
T3	hot dogs, coke, chips
T4	chips, coke
T5	chips, ketchup
T6	hotdogs, coke, chips

Find the frequent item set and generate association rules using apriori algorithm. Assume the minimum support threshold is 33.33% and minimum confidence is 60%.

Minimum Support Count

$$= \text{No of transactions} \times \frac{\text{support}}{100}$$

$$= 6 \times \frac{33.33}{100} = 1.99 = 2$$

Step 1

Item	Support Count
Hot Dogs	4
Buns	2
Ketchup	2
Coke	3
Chips	4

If any item having the count less than minimum support count, then we need to remove that item (Data pruning)

Step 2

Item	Support Count	
Hot Dogs, Buns	2	
Hot Dogs, Ketchup	1	x
Hot Dogs, Coke	2	
Hot Dogs, Chips	2	
Buns, Ketchup	1	x
Buns, Coke	0	x
Buns, Chips	0	x
Ketchup, Coke	0	x
Ketchup, Chips	1	x
Coke, Chips	3	

Steps Data Burning

Item	Support Count
Hot Dogs, Buns	2
Hot Dogs, coke	2
Hot Dogs, chips	2
Coke, Chips	3

Step 4 Combination of 3

Item	Support Count
Hot Dogs, Buns, Ketup	1
Hot Dogs, Buns, Coke	0
Hot Dogs, Bun, Chips	0
Hot Dogs, Ketup, Coke	0
Hot Dogs, Kitdup Chips	0
Hot Dogs, Coke, Chips	2
Buns, Ketup, Coke	0
Bun Ketchup, Chips	0
Ketchup, Coke, Chips	0
Bun, Coke, Chips	

tops Data Pruning

Items	Min Support
Hotdogs, chips, coke	2

Frequent itemset: {hotdogs, chips, coke}

Association Rule

{hotdogs, coke} ⟹ {chips}

$$\text{Confidence} = \frac{\text{Support count } (\{\text{hotdogs, chips}\})}{\text{Support count } (\{\text{hotdogs, coke}\})} \times 100$$

$$= \frac{2}{2} \times 100 \Rightarrow 100\%$$

{hotdogs, chips} ⟹ {coke}

$$\text{Confidence} = \frac{\text{Support count } (\{\text{hotdogs, coke}\})}{\text{Support count } (\{\text{hotdogs, chips}\})} \times 100$$

$$= \frac{2}{2} \times 100 = 100\%$$

{coke, chips} ⟹ {hotdogs}

$$\text{Confidence} = \frac{\text{Support count } (\{\text{coke, hotdog}\})}{\text{Support count } (\{\text{coke, chips}\})} \times 100$$

$$= \frac{2}{3} \times 100 = 66.66\%$$

{hotdogs} => {coke, chips}

Confidence : $\frac{2}{4} \times 100 = 50\%$

{coke} => {hotdogs, chips}

Confidence = $\frac{2}{3} \times 100 = 66.66\%$

{chips} => {hotdogs, coke}

Confidence = $\frac{2}{4} \times 100 = 50\%$

Strong Association Rule

1. {hotdogs, coke} => {chips}
2. {hotdogs, chips} => {coke}
3. {coke, chips} => {hotdogs}
4. {coke} => {hotdogs, chips}
5. {chips} => {hotdogs, coke}

Module IV

Classification & Prediction:

→ There are two forms of data analysis that can be used to extract models describing important classes or predict future data trends

They are:
↳ classification
↳ Prediction.

These are the two main method used to mine the data

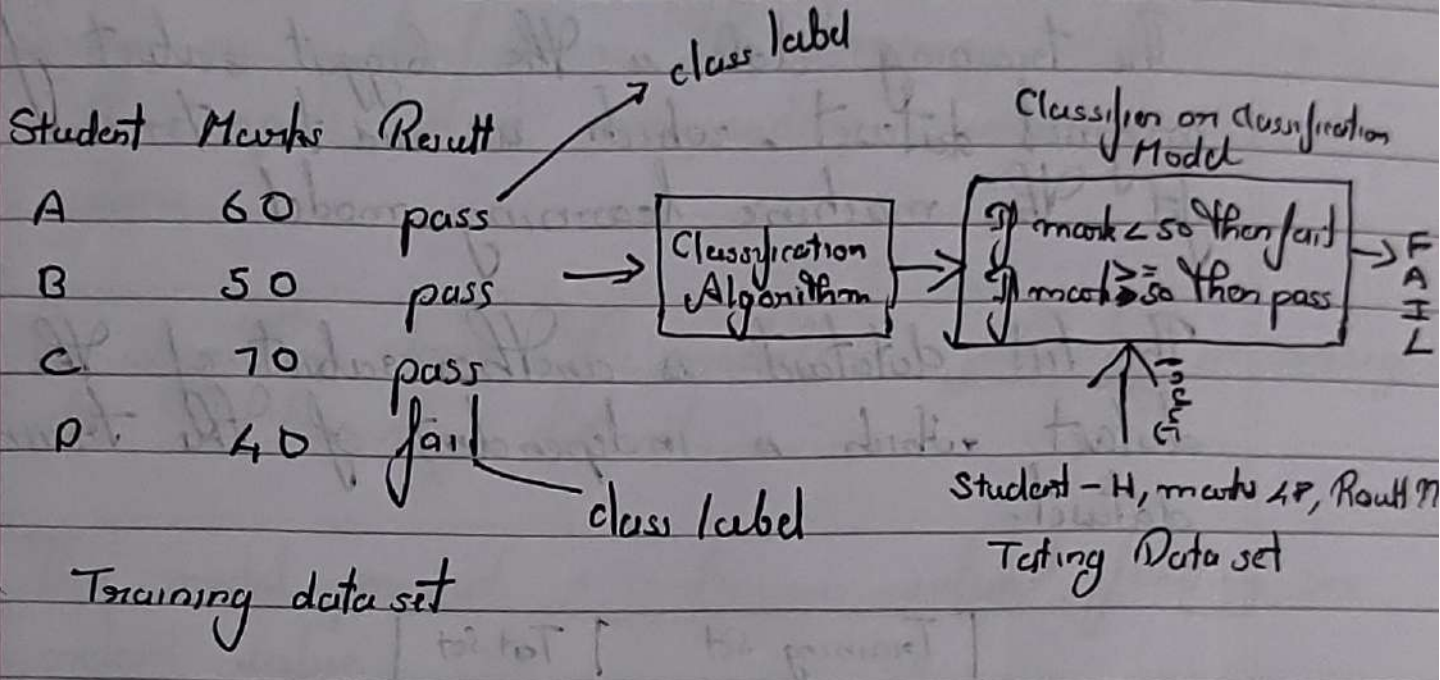
Classification:-

It is for converting large datasets to useful info.

Application → To classify people with or without a certain disease
→ To classify spam messages

Classification is the process of dividing the dataset into different categories or groups by

adding class label



→ Supervised learning always has a training set
∴ Classification is a supervised learning algorithm

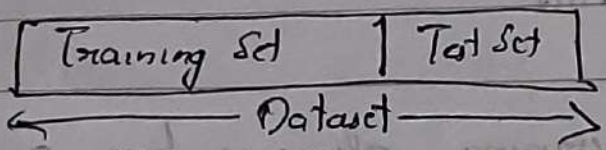
→ First a set of data is used as training data
→ The set of input data and the corresponding outputs are given to the algorithm
→ So the training dataset includes the input data and the associated class labels

→ Using the training dataset, the algorithm derive a model or classifier.

Training and Testing Data:

The training data is the biggest subset of the original dataset, which is used to train or fit the machine learning model.

The test dataset is another subset of the original dataset which is independent of the training dataset.



The training dataset is generally larger in size compared to testing dataset. General ratios of splitting train and test datasets are 80:20, 70:30, or 90:10

Prediction:

Forecast about a future event.

Eg Rainfall Prediction,

In this method, we need to predict the missing data for a new observation depending

on the previous data

Same as in classification. The training dataset contains the inputs and corresponding numerical output values.

The algorithm derives the model or a predictor according to the training dataset.

The model predicts a continuous-valued function or ordered value.

Bayesian Classification :- (Naive Bayes Classifier)

→ Naive Bayes is a probabilistic machine learning algorithm that can be used in a wide variety of classification tasks

→ Bayesian classification (classifier) can predict the probability that a given tuple belongs to a particular class

→ Based on Baye's Theorem.

→ More speed and accuracy as compared to decision Tree classifier

Bayes' Theorm

To calculate the probability of an event, based on prior knowledge related to that event.

Expressed as → P(A|B) . P(B|A) * P(A) / P(B)

P(A|B) is the posterior probability. (Probability of hypothesis A on the observed event B.

P(B|A) is likelihood probability. (Probability of the evidence given that the probability of a hypothesis is true.

P(A) and P(B) are the probabilities of event A and B

② Consider the dataset

Find out - class label on the instance outlook: sunny

	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
	Overcast	Yes
	Sunny	No