Data Warehouse and Data mining

LAST MOMENT
TUITIONS

For Full Course

www.Lastmomenttuitions.com

Data Warehouse

Q1) What is Data Warehouse?

Ans: Bill Inmon Consider to be father of Data Warehousing Provides the following definition

A data warehouse is a

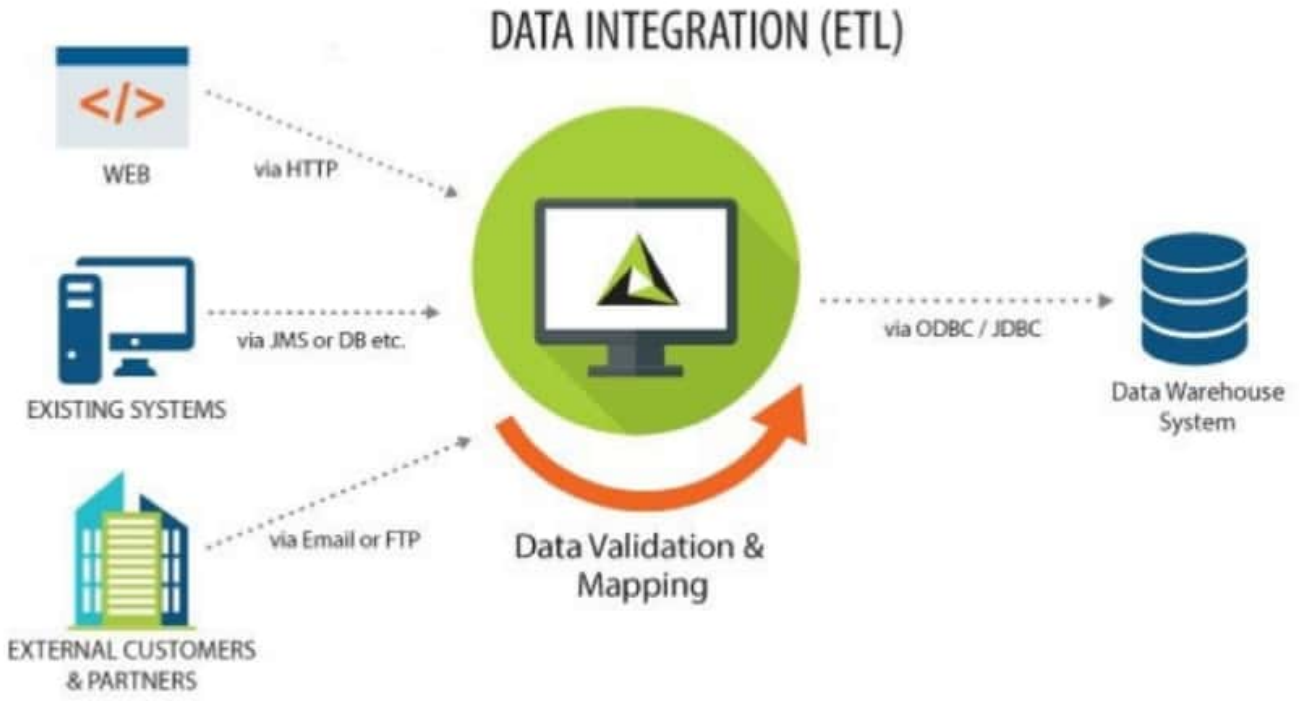
1. Subject Oriented
2. Integrated
3. Non Volatile
4. Time Variant

Collection of data in support of management's Decision

(yeh ek bada datebase hai jaha static information store hotha hai jo business ko improve karne ki decision me kaam aati hai)

1. **Subject Oriented**
    - A data warehouse is divided into major subjects (parts), such as customers, suppliers,products and sales.
    - Rather than concentrating on the day to day operation and transaction processing of an organisation a data warehouse focuses on the modeling and analysis of data for decision makers.
    - Hence data warehouse typically provides a simple and concise (kaam ki cheez) view around particular subject done by removing or eliminating data that are not useful in the decision support process

2. **Integrated Data**
    - For Proper decision making you need to pull together (nikaalna) all the relevant data (kaam ka data) from the various application.
    - A data warehouse is usually Constructed by integrating multiple heterogeneous sources, such as relational databases, Flat files and online transaction records.
    - Data Cleaning and data Integration techniques are applied to eliminate unwanted data and then bring together relevant information.



3. **Time Variant**
    - A data warehouse, because of its very nature of its purpose, has to contains historical data, not just current values .
    - For an operating system. The stored data contains the current values , On the other hand the data in the data warehouse is meant for analysis and decision making.
    - If the user is looking at the buying patterns of a specific customers the user needs data not only about the current purchase but on the past purchases as well
    - The time variant nature of the data in the data warehouse
        1. Allows for analysis of the post
        2. Relates information for the present
        3. Enables forecast for the future.

4. **Non Volatile Data**
    - The business transaction updates the operational system database in real time, we add,change, or delete data from operational system as each transaction happens but do not usually update the data in DW
    - You do not delete the data in the data warehouse in real time. Once the data is captured in the data warehouse , you do not run individual transaction to change the data there

(Data edit aaplog operational database me kar sakte ho par dataware house me nahi ek baar DW me Video dala toh aap edit yaa update nahi kar sakthe)

Data mart

Datamart is a subset of the data source usually oriented to specific purpose or major data warehouse which is divided to support business needs.
(Datamart ek part hai dataware house ka chotasa jo sirf ek particular cheez pe focus karte hai)

Eg: Agar engineering college ka data warehouse hai toh

1. Computer Department
2. IT Department
3. Civil Department
   Yeh sab department hai

Datamart Contains programs, data, software and hardware of a specific department of a company there can be separate data marts for finance sales, production or marketing

- All these data marts are different, but they can be coordinated
- Data mart of one department is different from data mart of another department
- A data mart is a small warehouse which is designed at departmental level

Q) Difference Between Data warehouse and Data Mart

Data Warehouse	Data Mart
It gives enterprise wide view of data	It gives departmental view of data
Union of all marts	Subset of Data Warehouse Or Single business process
Takes Longer time to implement (months to years)	Takes less time to implement (weeks to month)
It size is more than 100 TB	It size is less than 10 TB
Slower Response	Faster Response

(Dekho difference toh bahot easy hai par isko aur easy kar lete taki app ko exm tak yaad rahe , Ab agar aapko 5 points yaad karna hai toh aapko sirf alphabet yaad rakhna hai flow me (rstuv)

R- response
S- Size
T-Time
U-union
V-View

Meta Data

Meta data serves as a directory of the contents of your data warehouse
(yani metadata ek index ki tarah kaam karta hai data warehouse ke liye)

Meta data Types

Meta data in a data warehouse fall into three major categories

1. Operational metadata
2. Extraction and transformation metadata
3. End user meta data

1. Operational Metadata

Operational metadata contains all of the information about the operational data sources means data for the data warehouse comes from several operational system of the enterprise
(History of Migrated data and transformation Path)
(Matlab hum datawarehouse me data bahoot alag alag Branch se lake share karte hai aur user ko represent karte hai toh agar user ko janna hai ki data actual me konse gaav ka hai toh voh information operational metadata me rehta hai)

2. Extraction and Transformation Metadata

It Contains information about all the data transformation that takes place in data (staging area : Agar voh jagah data transform hota hai and load ke liye taiyar hota hai)

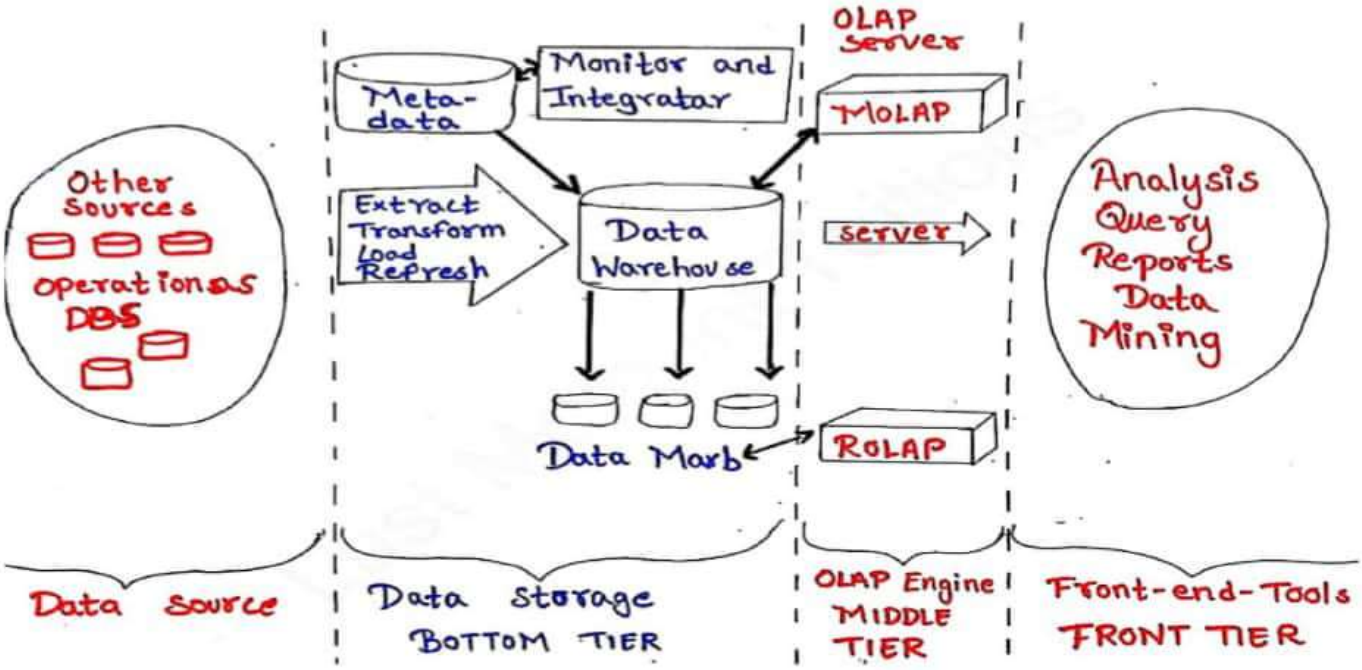
3. End user Meta data (Index)

- The End user Meta data is navigational map of data warehouse.
- It enables end users to find information from data warehouse.

Special Significance of Metadata

- First it acts as the glue/clue that connects all the parts of the data warehouse
- Next it provides information about the content of structure to the developers
- Finally it opens the door to the end user and makes the content recognizable
- Meta data in a data warehouse contains the answers to questions about the data in the data warehouse

Data warehouse Architecture



Back end Tools And Utilities: They are used to feed data into data warehouse (Bottom Tier) from operational databases or the other external sources.

- These tools and utilities perform data extraction, cleaning and transformation (eg to merge similar data from different sources into unifies format) as well as load and refresh functions to update the data warehouse
  (Matlab alag alag branches se data laake data warehouse me backend tools and utilities store kartha hai)

Bottom Tier: It is the warehouse database server that is almost always a relational database system
(relational database voh system hotha hai jisme data table from me store hota hai)

- Various Data Mart are connected to form a data warehouse
- This tier also contains a metadata repository which stores information about data warehouse and its contents.
- This tier also contains monitoring and integrator which always integrate data.
  (Yeh Main part hai datawarehouse ke architure ka jisme Data Mart Bahot saare milke data warehouse banta hai aur isme metadata block bhi hota hai)

The Middle Tier: The middle tier is an OLAP server

(OLAP server ek server hota hai jaise Waiter .Waiter Kaise order lete hai fir Khana Laake deta hai similarly server bhi query yaa request leta hai data laake deta hai)

It is typically implemented either using ROLAP or MOLAP

- **ROLAP:** is a server which performs operation of relational databases
- **MOLAP:** is a server which is special type and directly implemented on multi dimensional data and operation

TOP Tier: It is a front end client layer which contains

- query and reporting tools (matlab konsa data mangwana hai)
- Analysis tools (data aane ke baad uspe decision lena)
- Data Mining Tools (data bahar nikalna kaise hai)

(Do Not write in Hindi in exam it is for your understanding purpose)

Dimensional Modeling

STAR SCHEMA, SNOWFLAKE SCHEMA AND FACT CONSTELLATION (GALAXY SCHEMA)

(Dekho yaar star schema ka sidha sum aayega theory rarely poochthe hai so aap agar mere video deekhoge toh appko star,snowflaake, fact constellation samjh jayega apko practice karna padega. Filal chalo kuch basics bata deta hu iska)

(Jaise database me hum ER Diagram bana rahate usi tarah isme hum star schema and baki schema (jo ki map hai) banane se phele.)

Star schema banane ke liye 3 cheeze pata hona chahiye

1. Fact
2. Dimension
3. Measure

1. **Fact**: A fact is a collection of related data items, consisting of measures (matlab woh cheeze jispe DW (data warehouse) banana hai)

   Eg: Placement mere fact hai toh usme kon kon honga
   1. Student
   2. Company
   3. Tpo

   (inhe dimensions kehte hai)

   And value numeric kya hongi ?
   - No of students placed
   - No of student not placed
   - No of student eligible
   - No of student not eligible

     (jo number ki value hothi hai unhe measure kehte hai iska shape star type hota hai thoda bahot isilye star schema kehte hai)

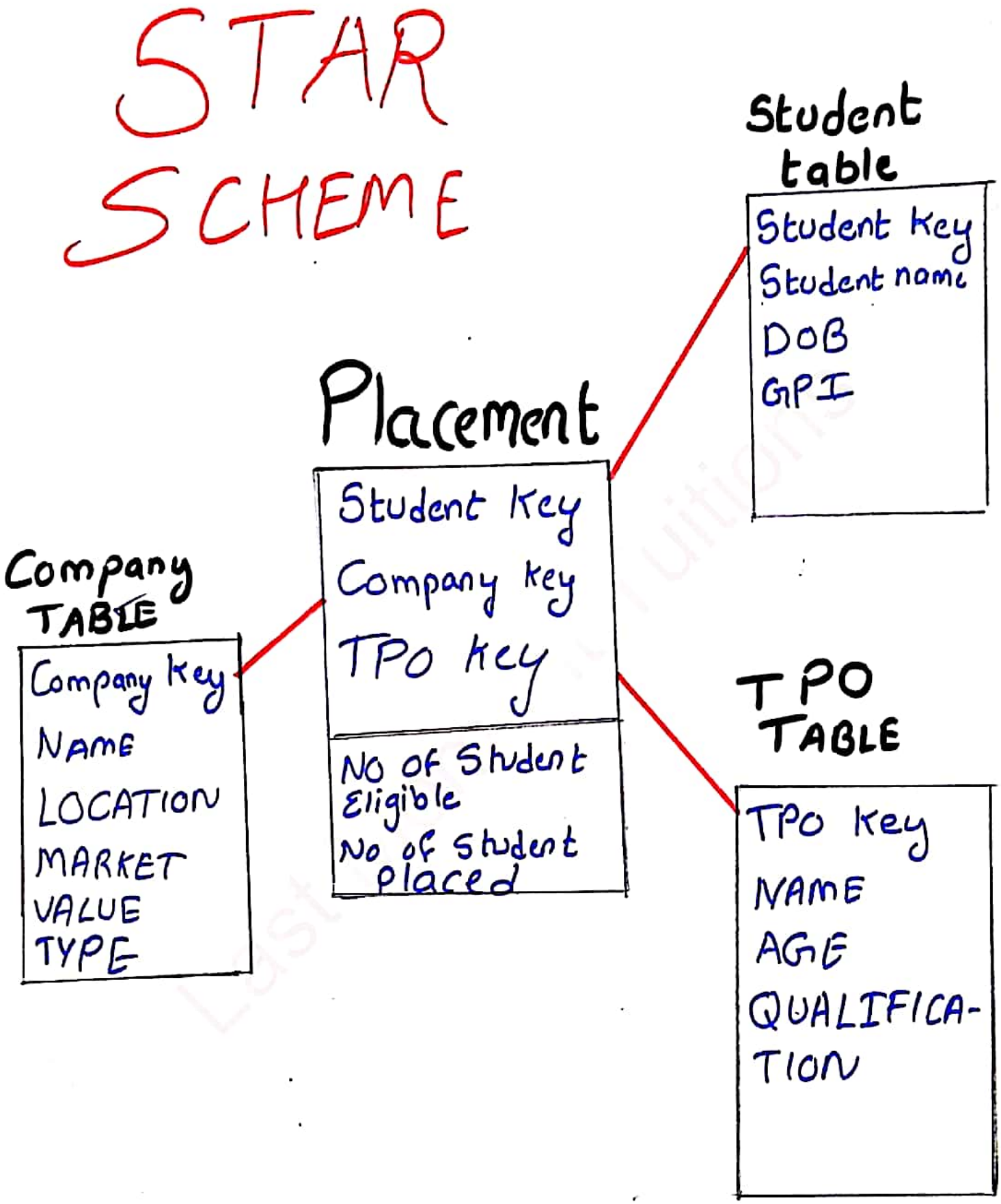
Snowflake Schema

Snowflake schema ko thoda aur detail me kholo toh woh snow flake banata hai matlab jo dimension hote hai usske bhi subparts karo)

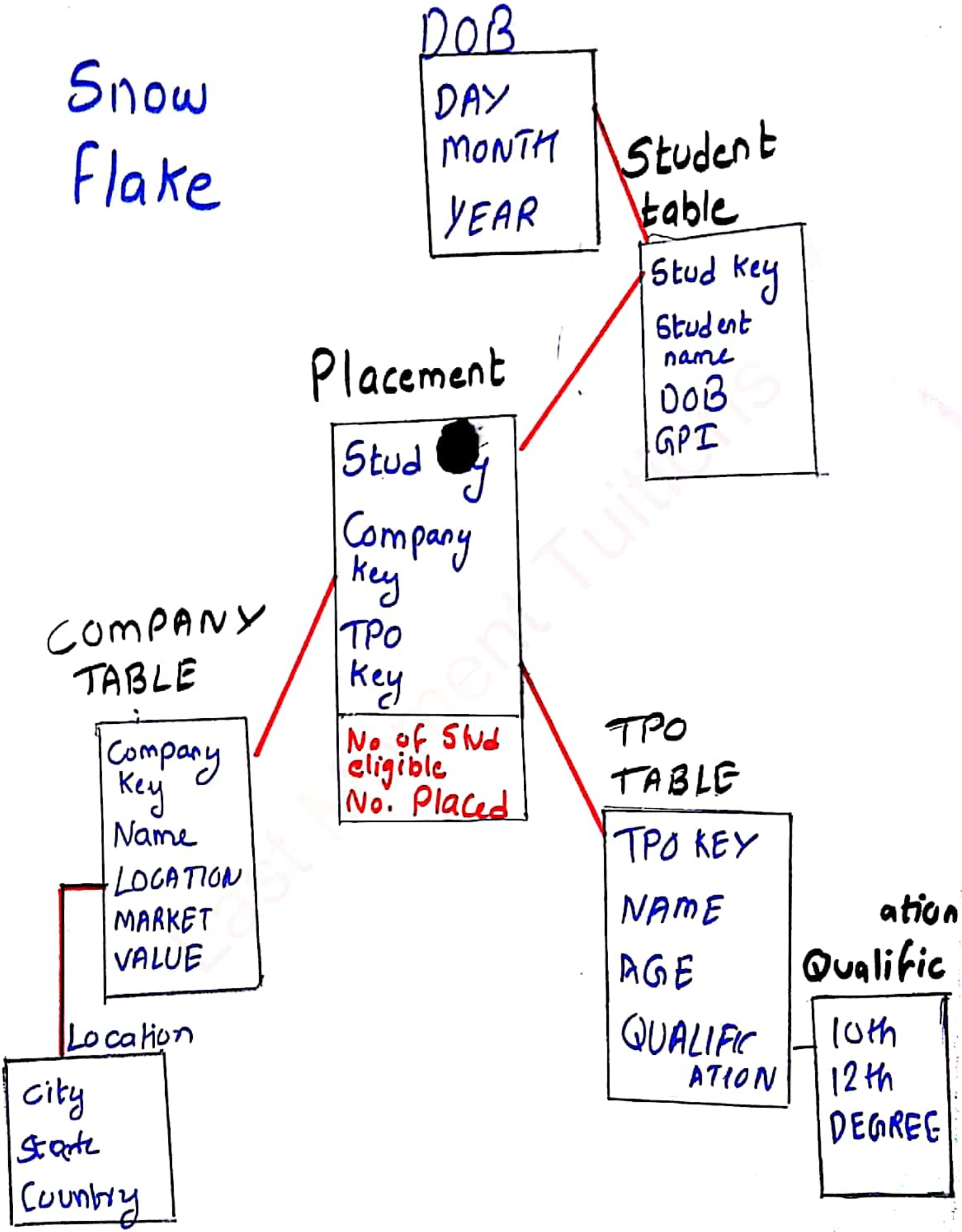
Fact Constellation

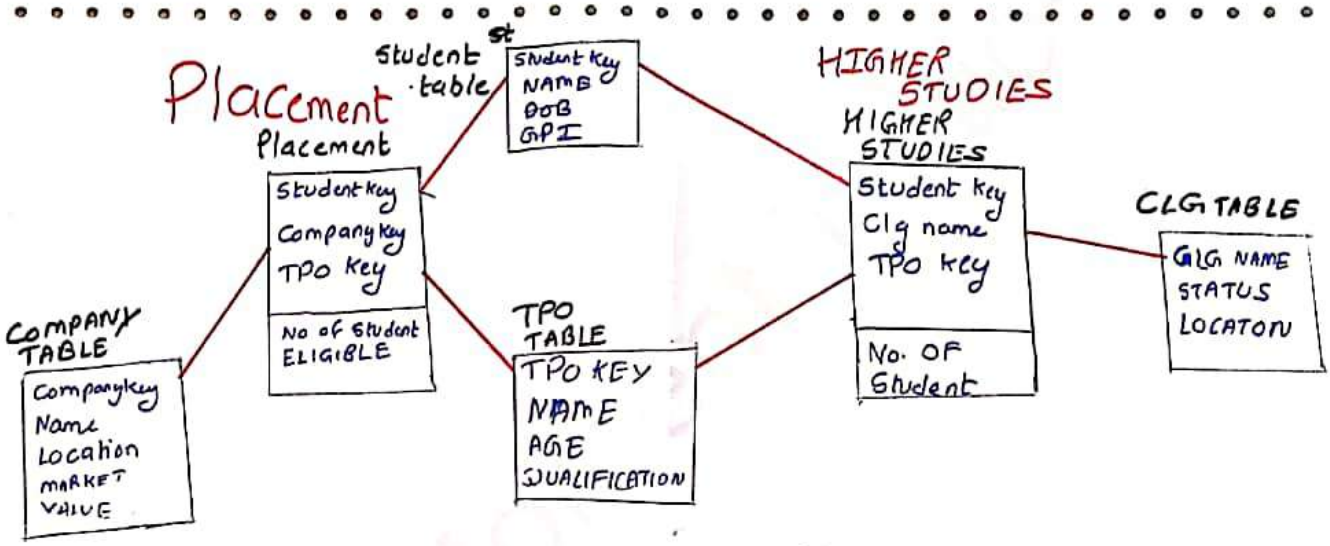
Voh schema jisme ek se jyada fact table hota hai dur jiske dimension common hote hai.

STAR SCHEME


Student table
Student key
Student name
DOB
GPI
Placement
Student key
Company key
TPO key
No of Student Eligible
No of Student placed
Company TABLE
Company key
NAME
LOCATION
MARKET VALUE
TYPE
TPO TABLE
TPO key
NAME
AGE
QUALIFICA-TION

Snow Flake


DOB
DAY
MONTH
YEAR
Student table
Stud key
Student name
DOB
GPI
Placement
Stud key
Company key
TPO key
No of Stud eligible
No. Placed
COMPANY TABLE
Company Key
Name
LOCATION
MARKET VALUE
Location
City
State
Country
TPO TABLE
TPO KEY
NAME
AGE
QUALIFICATION
Qualification
10th
12th
DEGREE



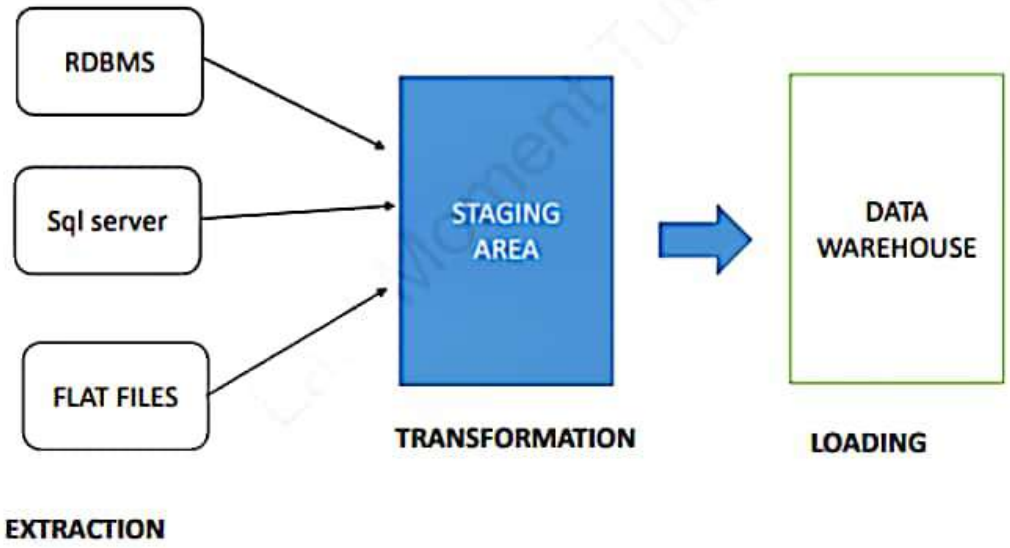
FACT CONSTELLATION

ETL

(EXTRACT, TRANSFORM , LOAD) PROCESS

Defination:

- It is a process in data warehousing of pulling data out of the source system and putting it in data warehouse



Data extraction from various source system

- Source system can RDBMS and files(json,xml)
- In this data is extracted from source system
- The main objective of this step is retrieve all required data from source system
- The extraction step should be design in such a way that it should not have negative affect on the source system

Data tranformation

- This step include cleaning ,filtering , validating and applying rules to extracted data
- The main objective of this step is to load the extracted data into target database with clean and general format
- This is because we extract data from various sources and each have their own format

- For example there are two sources A and B
- A date format is dd/mm/yyyy
- B date format is yyyy/mm/dd
- In transformation these date are bring into general format

Other thing that are carried out in this steps are:

- Cleaning(male to 'M' and female to 'F')
- Filtering(selecting only certain column to load)
- Enrichment(full name to 'first name' ,'middle name' , 'last name')
- Splitting(splitting one column into multiple column)
- Joining (together data from multiple sources)

In some cases data does not need transformation and this type of data is called rich data

LOADING

- Data extracted and transform is of no use until it is loaded in target data base
- In this step the extracted data and transform data is loaded to target database
- In order to make data load efficiently it is necessary to index the database

ETL process can be run parallel

- Data extraction take time so the second step of transformation take place simultaneously
- This prepare data for third step of loading
- As soon as some data is ready it is loaded without complete of previous steps

OLAP

(Online Analytical Processing)

Definition:

Online Analytical Processing (OLAP) is a category of software technology that enables analysts , manager and executive to gain insight into data through fast , consistent , interactive access in a wide variety of possible views of information that has been transform from raw data to reflect the real dimensionality of the enterprise as understood by the user.

Types of OLAP Servers

We have four types of OLAP servers –

- Relational OLAP (ROLAP)
- Multidimensional OLAP (MOLAP)
- Hybrid OLAP (HOLAP)
- Specialized SQL Servers

Relational OLAP

ROLAP servers are placed between relational back-end server and client front-end tools. To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.

ROLAP includes the following –

- Implementation of aggregation navigation logic.
- Optimization for each DBMS back end.
- Additional tools and services.

Multidimensional OLAP

MOLAP uses array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the data set is sparse. Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

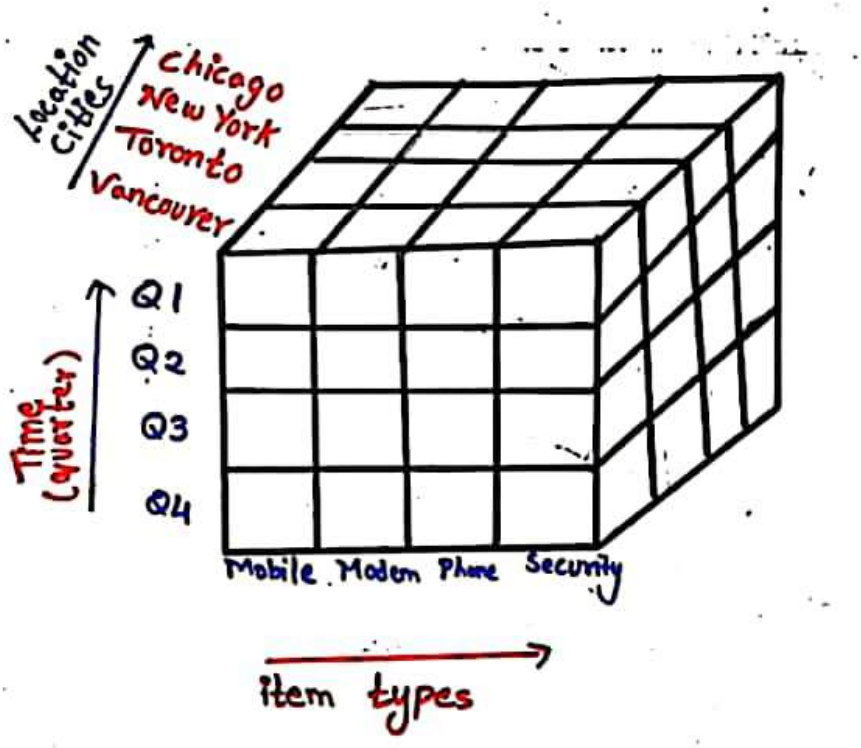
Hybrid OLAP

Hybrid OLAP is a combination of both ROLAP and MOLAP. It offers higher scalability of ROLAP and faster computation of MOLAP. HOLAP servers allows to store the large data volumes of detailed information. The aggregations are stored separately in MOLAP store.

Specialized SQL Servers

Specialized SQL servers provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

OLAP cube:



At the core of the OLAP, concept is an OLAP Cube. The OLAP cube is a data structure optimized for very quick data analysis.

- The OLAP Cube consists of numeric facts called measures which are categorized by dimensions. OLAP Cube is also called the **hypercube**.
- Usually, data operations and analysis are performed using the simple spreadsheet, where data values are arranged in row and column format.

- This is ideal for two-dimensional data. However, OLAP contains multidimensional data, with data usually obtained from a different and unrelated source.
- Using a spreadsheet is not an optimal option. The cube can store and analyze multidimensional data in a logical and orderly manner.

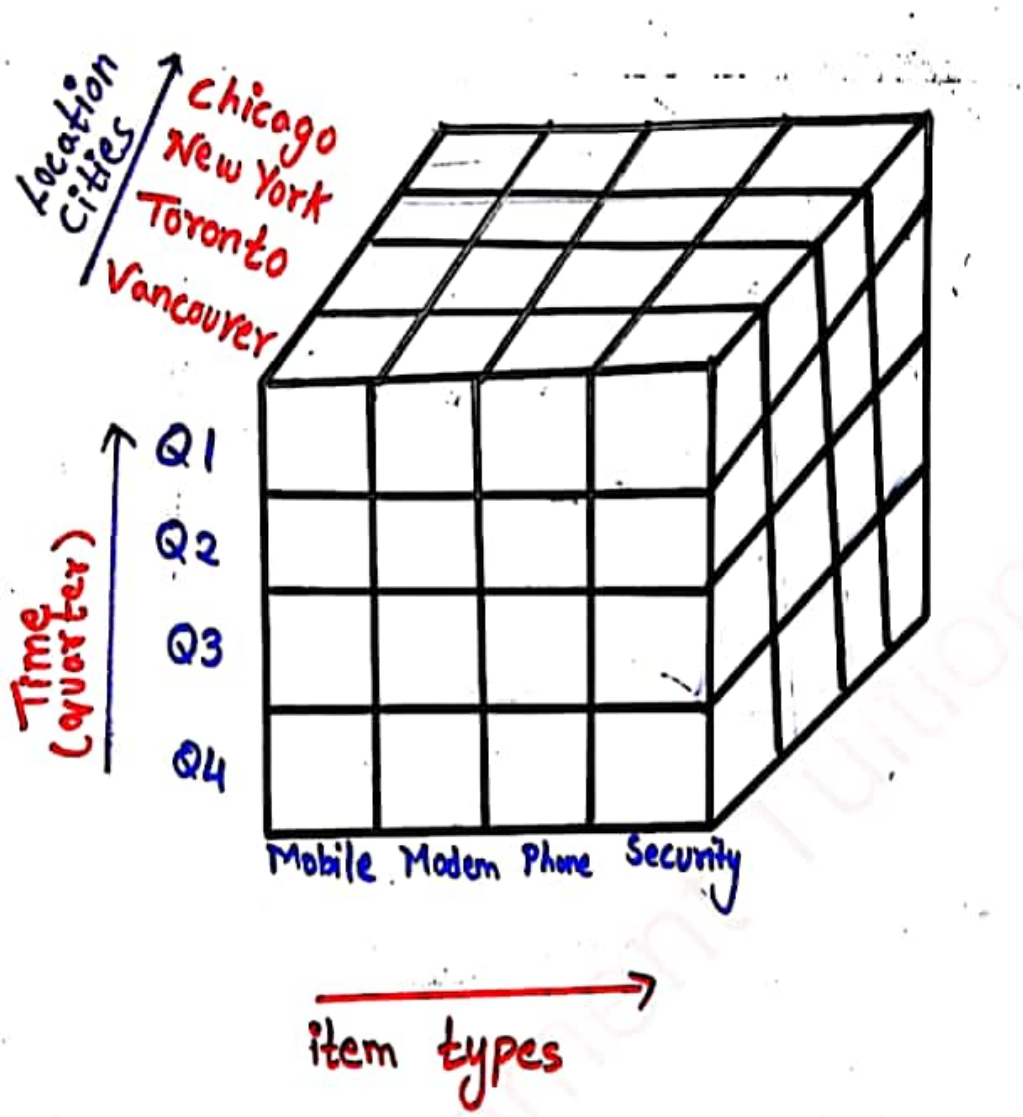
How does it work?

A Data warehouse would extract information from multiple data sources and formats like text files, excel sheet, multimedia files, etc.
The extracted data is cleaned and transformed. Data is loaded into an OLAP server (or OLAP cube) where information is pre-calculated in advance for further analysis.

Basic analytical operations of OLAP

Four types of analytical operations in OLAP are:

1. Roll-up
2. Drill-down
3. Slice and dice
4. Pivot (rotate)



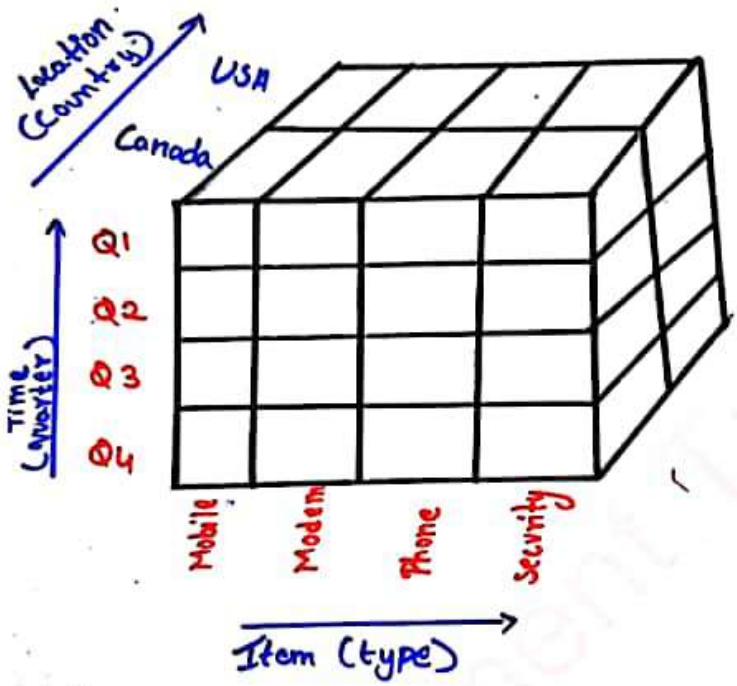
Operations on Data Warehouse

Roll-up:

This display moves up the hierarchy, grouping into larger units along a dimension (e.g., summing weekly data by quarter, or by year).

(yeha hum upper upper se cheez ko dekhte hai yani agar hum city dekh rahe the ab state dekhte hai dekhte hai dekhte hai)

ROLL UP



Example
Chicago
Newyork
Toranto
vancourer

Can be roll up to

USA
Chicago

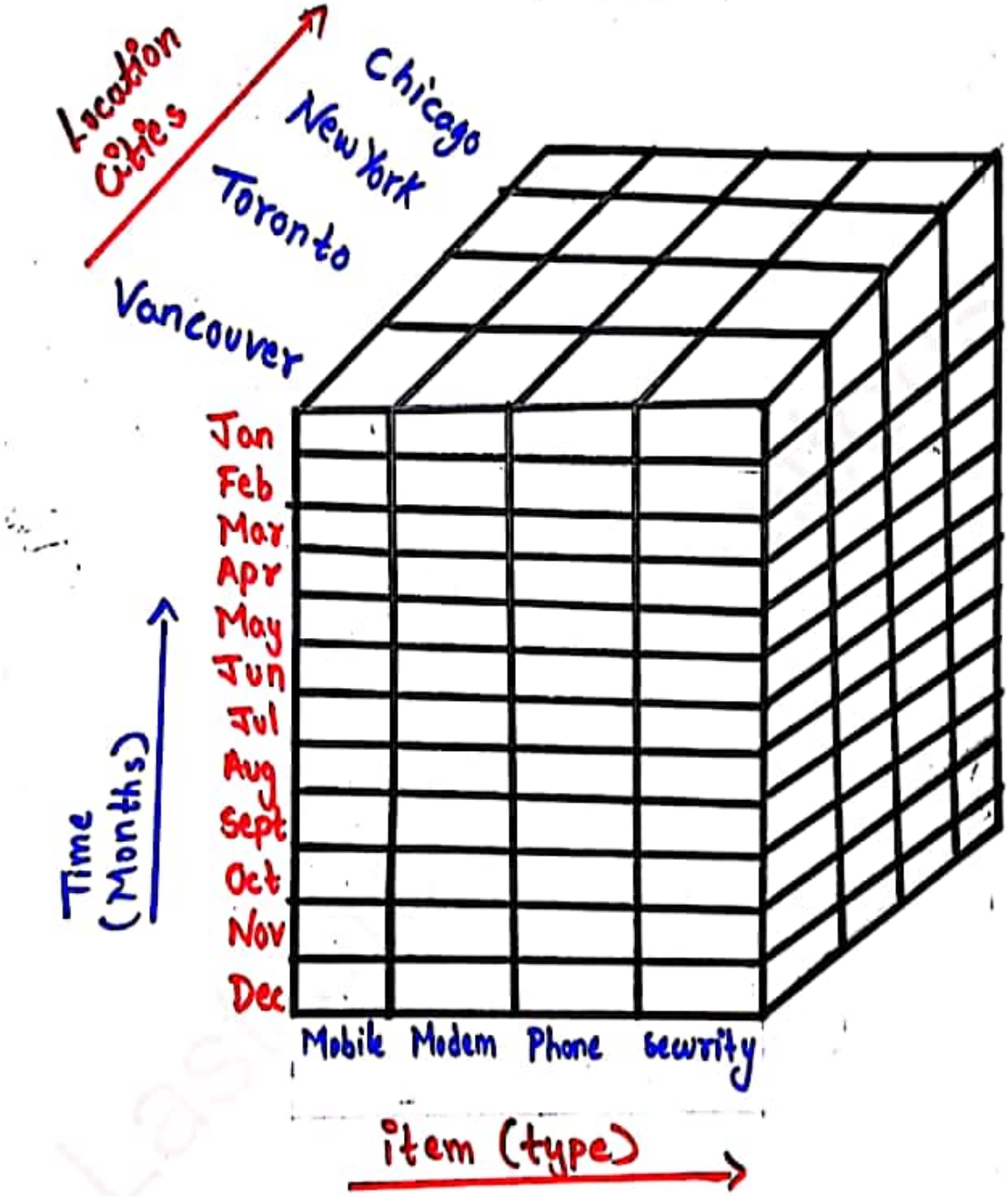
Drill-down:

This display provides the opposite capability, furnishing a finer-grained view, perhaps disaggregating country sales by region and then regional sales by subregion and also breaking up products by styles.

(yeha hum cheezo ko detail me dekhte hai

Yeni agar hum quarter year(3 month) dekh rahe the ab har mahine dekhenge)

DRILL DOWN


Location Cities
Chicago
New York
Toronto
Vancouver
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sept
Oct
Nov
Dec
Time (Months)
Mobile
Modem
Phone
Security
item (type)

Slice:

This display provides output by performing a selection on one dimension of given multidimensional cube results into sub cube.

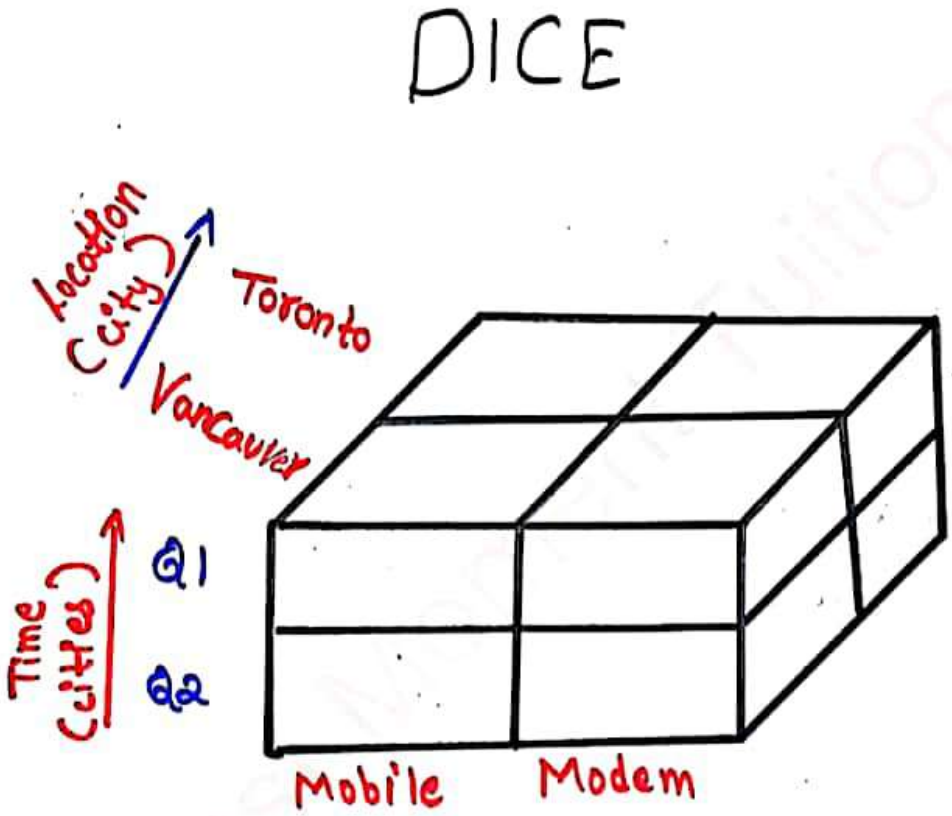
(hum mango ki kaise slice kat te hai toh kaise sirf ek hi side dikhta hai usi tarah hum cube(multidimension) ki slice kat te hai aur one dimension banate hai)

Pivot:Bus slice ko ghuma do voh pivot ban gaya.

Dice:

This display gives a sub cube as an output by performing selection operation on two or more dimensions.

(cube(multidimension) ka ek chota sa tukda dice hota hai)



Feature	Operational System (OLTP)	Information System (OLAP)
• Characteristic	operational processing	Informational processing
• Orientation	transaction	Analysis
• User	Clerk, DBA, Database professionals	Knowledge Workers (eg manager, executive)
• Focus	data in	information out
• DB design	ER based, application oriented	star/snowflake, subject-oriented
• Data	Current-guranted up-to-date	historical, consolidated
• Unit of work	short, simple transaction	Complex query
• Access	read/write	mostly read
• DB size	100MB to GB	100 GB To TB
• Priority	high performance high availability	high flexibility & end-user autonomy

TRICK to REMEMBER

FADU PCO

F - Focus

A - Access

D - Database Design

D - Database Size

D - Data

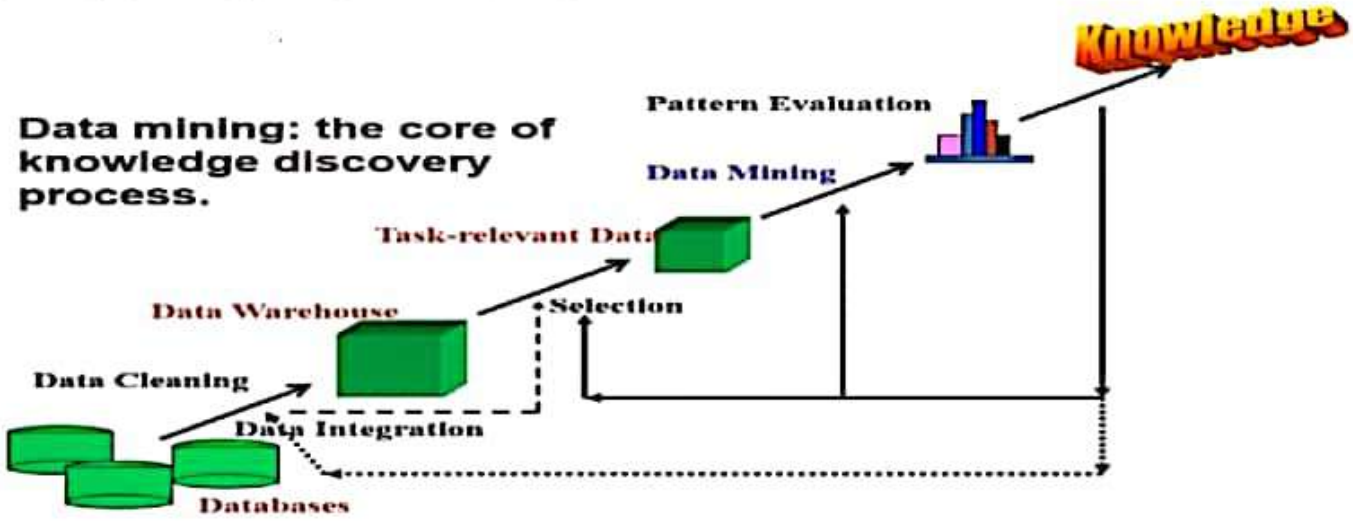
P - Priority

C - Characteristic

O - Orientation

Knowledge Discovery in Databases or Data warehouse

KDD



Data Cleaning

- To remove noise and inconsistent data example parsing the data.
- Cleaning is performed for detection of Syntax error.
- Parser decide whether the given string of data is acceptable within data specification.

Data Integration

- Where multiple data sources are combined.

Data Selection

- Where data relevant to the analysis task are retrieved from the database.

Data Transformation

- Where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance.

Data Mining

- An essential process where intelligent methods are applied in order to extract data patterns.

Pattern Evaluation

- To identify the truly interesting patterns representing knowledge based on some interesting new measures.

Knowledge Representation

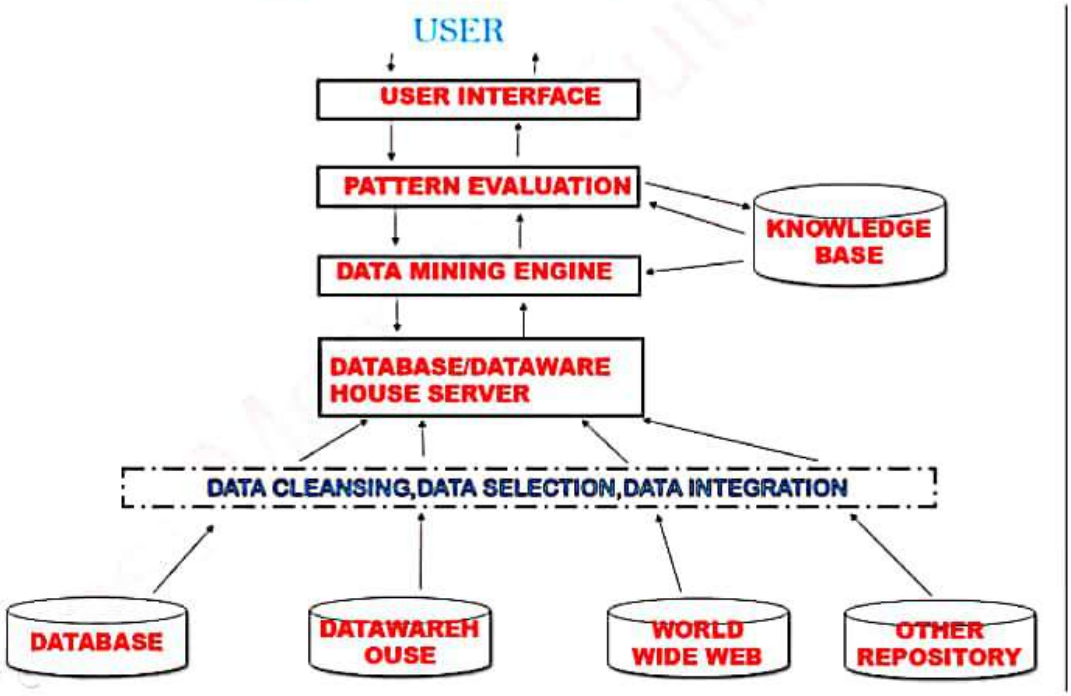
- Where visualization and knowledge representation technique are used to present the mined knowledge to the user.

DATA MINING NOTES

DEFINATION

Data mining refers to extracting knowledge from large amount of data

(hazaro GB ke data me se apne kaam ya interest ki cheez nikalne ki technique ko data mining kehte hai)



Data mining architecture

We will go bottom to top

1)Database, datawarehouse, www, other repository: This are single database or group of database or spreadsheet etc are main sources of data from where data is to be fetched for data mining

2) Data cleansing, selection, integration:

In this the data which is fetching is process for cleansing i.e removing noise from data and removing unwanted data ,data selection means selection interested data only and avoid unwanted data then the data is been integrated i.e combined

3)Database/dataware house server: The database or dataware house server are responsible for fetching relevant data based on the user data mining request

4)Data mining engine: This is essential(important) module for data mining system ideally consists of set of functional module for task such as Association and co-relation analysis, classification

, prediction , clustering analysis ,Evolution analysis etc.

5)Pattern evaluation:This component typically employs interestingness measures & interacts with data mining modules,so as to focus the search toward interesting pattern.

For efficient data mining it is highly recommended to push the evaluation of pattern as deep as possible into the mining process so as to confine the search to only the interesting pattern

6)User interface:This module communicate with user and data mining system the user interact with data mining system by specifying a data mining query or task ,providing the information that help to focus on search

Knowledge base : This is the domain knowledge use to guide the search or evaluate the interestingness of resulting patterns

No Question were Asked from this chapter yet

Steps in Data Preprocessing

- Data Preprocessing is one of the most critical steps in data mining.
- Data Preprocessing involves transforming raw data into understandable format

Data goes through series of steps during preprocessing :

1. DATA CLEANING
2. DATA INTEGRATION
3. DATA TRANSFORMATION
4. DATA REDUCTION
5. Data Discretization

DATA CLEANING

- Data cleaning is the process of removing noise and correcting inconsistencies in the data.
- Dirty data can cause confusion and result in unreliable and poor output. Hence, the first step in data preprocessing is data cleaning.
- The cleaning of data is done by :

- Parsing:
  Parsing locates and identifies individual data elements in the source files and then isolates these data elements in the target files.
- Correcting:
  Correct parsed individual data components using sophisticated data algorithms and secondary data sources.
- Standardizing:
  Standardizing applies conversion routines to transform data into its preferred and consistent format using both standard and custom business rules.
- Matching:
  Searching and matching records within and across the parsed, corrected and standardized data based on predefined business rules to eliminate duplications.

- **Consolidating:**
  Analyzing and identifying relationships between matched records and consolidating/merging them into one representation.

DATA INTEGRATION

- Data integration involves combining the data from different sources using different technologies and different databases.
- The outcome of integration is a single unified view of the data.
- The biggest challenge in data integration is the technical implementation of integrating data from different and incompatible data sources.
- This phase of preprocessing is done in the design phase and in the implementation phase.

DATA TRANSFORMATION

- Data transformation is the process of converting data from one format into another.
- It involves transforming the actual values from one representation to the target representation.
- Examples of data transformation tasks include mapping stock symbols to company names, changing date format from MM-DD to DD-MM, and replacing cities by their countries.
- There are some transformations, such as litres to gallons, which can be performed by applying a formula or a program on the input values. Such types of transformations are referred to as syntactic transformations.
- Other types of transformations, such as company name to stock symbol and event to date, require finding the mappings between the input and output values in a repository of reference data.

DATA REDUCTION

- A database or data warehouse may store terabytes of data.So it may take very long to perform data analysis and mining on such huge amounts of data.
- Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume but still contain critical information.

- Need for data reduction:
    a. Reducing the number of attributes
    b. Reducing the number of attribute values
    c. Reducing the number of tuples
- It enables us to combine features and identify useful data.

Data Reduction Strategies:-

1. **Data Cube Aggregation** : Aggregation operations are applied to the data in the construction of a data cube.
2. **Dimensionality Reduction** : In dimensionality reduction redundant attributes are detected and removed which reduce the data set size.
3. **Data Compression** : Encoding mechanisms are used to reduce the data set size.
4. **Numerosity Reduction**: In numerosity reduction where the data are replaced or estimated by alternative.
5. **Discretisation and concept hierarchy generation** :Where raw data values for attributes are replaced by ranges or higher conceptual levels.

Data Discretization

- In Data Discretization the range of continuous attribute is divided into intervals
- Data Discretization convert a large number of data value into smaller once , so that data evaluation and data management becomes easy.

Chapter 8 : Classification

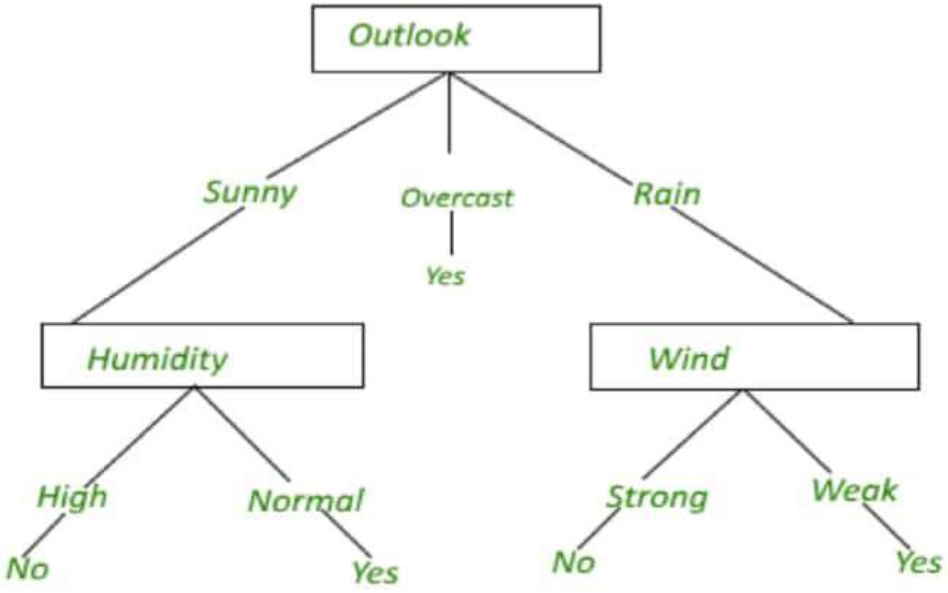
Decision Tree

- A decision tree is a graphical representation of possible solutions to a decision based on certain conditions.
- It's called a decision tree because it starts with a single box (or root), which then branches off into a number of solutions, just like a tree.
- We get a Pre-dataset based on which we create a decision tree and then when new data sample comes we can classify and Predict them.
- A decision tree represent rules and it is very popular tool for classification and prediction.
- Rules are Easy to understand and can be directly used in SQL to retrieve record from database.
- Decision tree algorithm falls under the category of the supervised learning.
- They can be used to solve both regression and classification problems.

There are many algorithm to build decision tree

1. ID3 (Iterative Dechotomiser 3)
2. C4.5 (Successor of ID3)
3. CART (Classification and Regression Tree)
4. CHAID (CHi-squared automatic interaction detector)

Example : Decision Tree for Playing Tennis



Decision Tree Representation

- Decision tree classifier has a tree type structure.
- It has a root node , leaf node and decision node.
- Root node is from where the tree starts and its important to find out which attribute must become root node in the given data set.
- A leaf node is the last node of each branch and indicate class label or value of target attribute.
- A decision node is node of the tree which has leaf node or sub-tree. (example : humidity and wind are decision node)
- A test is carried out on each value of decision node to get the decision of class label or to get next sub tree. (har Decision node pe ek test hota hai jo ya toh final class value dikhata hai ya next sub tree)

Advantages

- Decision trees generate understandable rules.
- Decision trees perform classification without requiring much computation.
- Decision trees are capable of handling both continuous and categorical variables.
- Decision trees provide a clear indication of which fields are most important for prediction or classification.

Disadvantages

- Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
- Decision trees are prone to errors in classification problems with many class and a relatively small number of training examples.

NOTES FROM Date

Last Moment tuitions

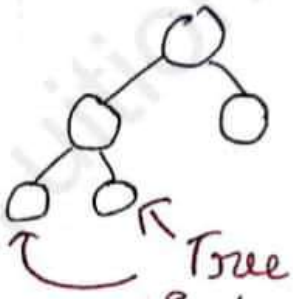
Decision tree notes

Age	Competition	Type	Profit
old	Yes	Software	Down
old	NO	Software	Down
old	NO	Hardware	Down
mid	Yes	Software	Down
mid	yes	Hardware	Down
mid	NO	Hardware	UP
mid	NO	Software	UP
new	Yes	Software	UP
new	NO	Hardware	UP
new	NO	Software	UP

Draw a decision tree for given dataset ?
yeh question us university exam me 10 marks ke liye Pucha gaya tha

Let see how to solve this

- Dekh bhai koi bhi dataset ya Table diya ho uska jo last column hota hai na voh hota hai class ATTRIBUTE
- jaise apne table me PROFIT hai aur uski jo value rahengi voh honge mere leaf node



Tree ke jo last wale node ko leaf node kehta hai

Given question me Decision tree isliye Pucha hai meko aise decision tree banake jo jisse Pata chale mera Profit UP and down kaise honge.

$$I(P_i, N_i) = \frac{-P}{P+N} \log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N} \log_2\left(\frac{N}{P+N}\right)$$

$P = 0, \quad N = 3.$

$$I(P_i, N_i)_{old} = \frac{-0}{0+3} \log_2\left(\frac{0}{0+3}\right) - \frac{3}{0+3} \log_2\left(\frac{3}{0+3}\right)$$

(Dekh bhai jab P ya N dono me se ek bhi valu 0 hai toh answer bhi 0 hi hoga)

$I(P_i, N_i)_{old} = 0$

similarly we find for

$P = 2, \quad N = 2$

$$I(P_i, N_i)_{mid} = \frac{-2}{4} \log_2\left(\frac{2}{4}\right) - \frac{2}{4} \log_2\left(\frac{2}{4}\right)$$

(When P=N then entrophy will be 1)

$I(P_i, N_i)_{mid} = 1$

P=3, N=0

$I(P_i, N_i) = 0$ //

New

∴ Entrophy of Age

$$= \sum \frac{P_i + N_i}{P+N} I(P_i, N_i)$$

↖ yeh P and N class attribute ki value hai where P=5, N=5

$$= \frac{0+3}{10}(0) + \frac{2+2}{10}(1) + \frac{3+0}{10}(0)$$

$$= \frac{4}{10} = 0.4$$

∴ Gain = Entrophy (class) − Entrophy of Age

= 1 − 0.4 ↖ (yeh jo humne starting Pe Profit ki entrophy nikali thi voh hai)

= 0.6 //

Step 1:-

In Profit

Ut
Class P = Profit (UP) = 5
Class N = Profit (down) = 5

(Toh hame class attribute hai Profit
usko entrophy nikalna hai)

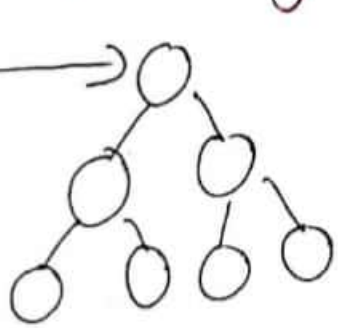
$$\text{Entrophy}(P, N) = \frac{-P}{P+N} \log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N} \log_2\left(\frac{N}{P+N}\right)$$

after putting value of P=5, N=5 ae
get

$$= \frac{-5}{10} \log_2\left(\frac{5}{10}\right) - \frac{5}{10} \log_2\left(\frac{5}{10}\right)$$

Entrophy (Profit) = 1

chalo itna Pata chal gaya ki up and down mere leaf node honge so questions comes root node kon honga



To find this - hum saare column ka gain nikalenge aur compare karenge jiska gain sabse zyaada bad voh root node

For Age	Pi	Ni	I(Pi, Ni)
old	0	3	0
mid	2	2	1
new	3	0	0

Now for Competition

	Pi	Ni	I(Pi, Ni)
Yes	1	3	0.81127
NO	4	2	

$$I(P_i, N_i)_{Yes} = \frac{-1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)$$

$$= 0.81127$$

$$I(P_i, N_i)_{NO} = \frac{-4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)$$

$$= 0.918295$$

Entrophy (competition)

$$= \sum \frac{P_i + N_i}{P+N} (I(P_i, N_i))$$

$$= \frac{(1+3)}{10}(0.81127) + \frac{(4+2)}{5+5}(0.918295)$$

$$= 0.8754$$

Gain = Class Entrophy − Entrophy (competition)

= 1 − 0.8754

= 0.124515.

for Type

	P_i	N_i	$I(P_i, N_i)$
software	3	3	1
hardware	2	2	1

$I(P_i, N_i)$ = 1 ∵because $P = N$
software

$I(P_i, N_i)$ = 1
Hardware

Entrophy (Type)

$$= \frac{3+3}{5+5}(1) + \frac{(2+2)}{5+5}(1)$$

$$= \frac{6}{10} + \frac{4}{10} = \frac{10}{10} = 1$$

Gain = Class entrophy − Entrophy (Type)

$$= 1 - 1$$

$$= 0$$

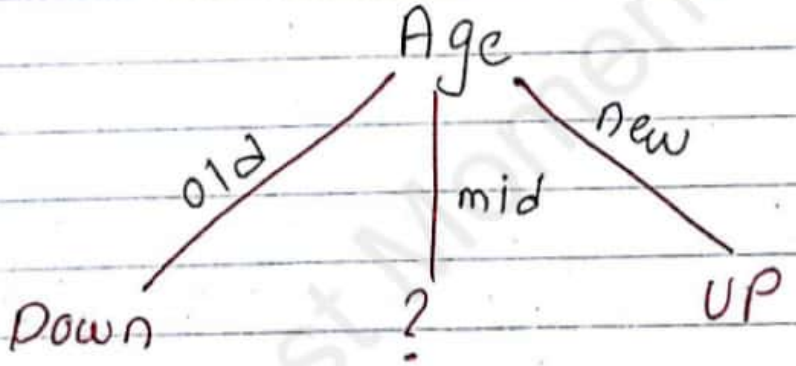
Information Gain

Age — 0.6 ← root node

Competition — 0.124515

Type — 0

- Toh gain sabse zyaada age ka hai root node will be age
- Ab AGE me jitni bhi value hai us utne root node se branches niklenge



Agar app Age old ki values and Profit ki values ko compare kare toh sab value down hai direct down likh do same for new

Age	Profit	Age	Profit
old →	Down	new →	UP
old →	Down	new →	UP
old →	Down	new →	UP

But MID me Up and down dono value ...
isliye Hum direct kuch Put nahi kar sakte
so hame firse sirf mid ka table banana
Padenga

Age	Competition	Type	Profit
mid	Yes	Software	Down
mid	Yes	Hardware	Down
mid	NO	Hardware	UP
mid	NO	Software	UP

Ab same Process firse karna hai
class antribute ki entrophy nikalni hai
and Competition and Type ka gain nikalke
jo bada honga usko mid ka node
banana hai

class P = Profit(UP) = 2
class N = Profit (down) = 2.

Entrophy (Profit) = 1 ∵(P=N).

Now lets find gain for Competition

Competition

	Pi	Ni	I(Pi, Ni)
Yes	0	2	0
NO	0	2	0

I(Pi, Ni) = 0 (Yes)

I(Pi, Ni) = 0 (NO)

(Koi bhi ek value P or N me 0 hai toh ast answer will be zero)

Entrophy (competition)

$$= \frac{2}{4}(0) + \frac{2}{4}(0) = 0$$

Gain = Entrophy class − Entrophy (competition)

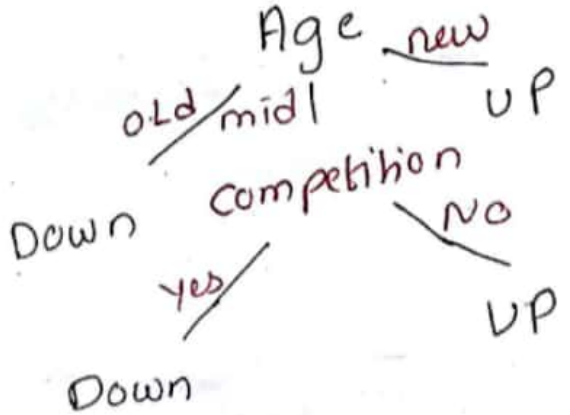
$$= 1 - 0$$

$$= 1$$

Gain → 1

Competition → 0

Type



obviously apko doubt aya hoga ki yaar fir Type kaha gaya (yes) ki and dekh bhai jab mai competition Profit ki value compare kar raha hu toh I m Getting direct answer down toh Type ka zhanjat kya Palneko

similarly for NO = UP

competition	Profit
Yes	Down
Yes	Down

Competition	Profit
NO	UP
NO	UP

Toh agar app decision tree ko analyse karo toh we can see ki agar hum new person rakhte hai and voh age me jawan hai toh Profit up and budha hai toh Profit down

Agar age mid hai toh dekhne ka competition kitna hai if Yes toh Profit → down if NO toh Profit → up

Toh iss tarah se real life me decision tree kaam karta hai.

Note: Exams me sab likhne ki jarroorat nahi hai sirf jo ka calculation voh likhna hai baki apka understanding purpose ke liye hai.

Thank you so much to study from last moment tuitions.

and please let us know if we can add some help to you from last moment tuitions aur ha

Why naive bayes is called naive?

(Naive bayes ek classifier hai jo classify karta hai cheezo ko naive ka matlab hota hai a Person/action showing a lack of experience toh isko is classifier ko naive kyu kehte hai voh iss answer me bataya hai)

- Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.
- The Naive Bayes Classifier assumes that all the features of a class are independent of each other.
- This means that anyone of the feature is missing then classification may not be perfect.
- For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter, or a ball if we don't give the red colour classification.
- So even if the features are inter-dependent on each other, the Naive Bayes Classifier will consider them independently.

It's called naive also because it makes the assumption that all attributes are independent of each other. This assumption is why it's called naive as in lots of real world situations this does not fit. Despite this the classifier works extremely well in lots of real world situations and has comparable performance to neural networks and SVM's in certain cases (though not all).

Naive Bayes Classifier

S.No.	Age	Income	Student	Credit	Buy
1	<30	High	No	Fair	No
2	<30	High	No	Excellent	No
3	31-40	High	No	Fair	Yes
4	>40	Medium	No	Fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	low	Yes	Excellent	No
7	31-40	Low	Yes	Excellent	Yes
8	<30	Medium	No	Fair	No
9	<30	low	Yes	Fair	Yes
10	>40	Medium	Yes	Fair	Yes
11	<30	Medium	Yes	Excellent	Yes
12	31-40	Medium	No	Excellent	Yes
13	31-40	High	Yes	Fair	Yes
14	>40	Medium	No	Excellent	No

X = (age = <30, income = medium, student = yes, credit = fair)

$P(buy = Yes) = 9/14 = 0.643$

$P(buy = No) = 5/14 = 0.357$

- Age.
  $P(age = <30 \mid buy = Yes) = 2/9 = 0.222$
  $P(age = <30 \mid buy = No) = 3/5 = 0.600$
- Income
  $P(Income = med \mid buy = Yes) = 4/9 = 0.44$
  $P(income = med \mid buy = No) = 2/5 = 0.40$
- Student
  $P(stud = Yes \mid buy = Yes) = 6/9 = 0.667$
  $P(stud = Yes \mid Buy = No) = 1/5 = 0.20$
- Credit
  $P(credit = fair \mid buy = Yes) = 6/9 = 0.667.$
  $P(credit = fair \mid buy = No) = 2/5 =$

$P(X/Yes) \cdot P(Yes) = P(age/Yes) \cdot P(med/Yes) \cdot P(Yes/buy) \cdot P(fair/Yes) \cdot P(Yes)$
$= 0.222 \times 0.44 \times 0.667 \times 0.667$
$= 0.0435$

$P(X/No) \cdot P(No) = P(age/No) \cdot P(med/No) \cdot P(Yes/No) \cdot P(fair/No) \cdot P(No)$
$= 0.60 \times 0.40 \times 0.20 \times 0.40$
$= 0.0192$

$P(X/Yes)$ is greater than $P(X/No)$,
$\therefore$ X will buy the product.

Chapter 9 : Clustering

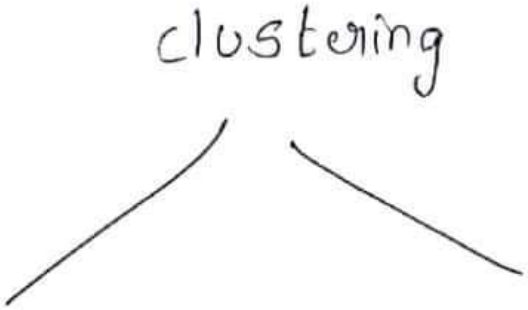
What is Clustering?

Clustering is the process of making a group of abstract object into classes of similar object.

(matlab abstract object bole toh bahut saare object me se similar object ka group banana isko kehte hai clustering.)

Application of clustering

- Clustering analysis is broadly used in many application such as market research, pattern recognization, data analysis and image Processing



K-mean clustering

- This is used in ~~creative~~ creating central cluster

- K mean clustering aims to partition n observation into K cluster in which each observation belongs to the cluster with nearest mean, serving as a prototype of the cluster. This result in partition of the data space into cell

(K mean ka maksat hota hai partition karna (n = jitna data hai) in k cluster)

- ~~[illegible]~~ steps to solve the ~~[illegible]~~ sums

step 1) Take mean value

step 2) find nearest number of mean and Pot in cluster

step 3) Repeat Process one and two untill we get same mean

Tension mat lo jab sum solve karenge toh zyaada achse samjhega.

Given Problem:-

Page No.

$K = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$ ← given data set

$K = 2$ ← jitna K diya hai utna cluster banane ke.

Step 1) Take mean value

K=2 hai dekho abhi hamara start hai so pehle se agar 2 mean value nahi di rahengi toh koi bhi 2 random value ko le lene ka 2 kya because K=2 hai isliye.

$m_1 = 4$ $m_2 = 12$

Step 2) Find nearest number of mean and Put in cluster

$K_1 = \{2, 3, 4\}$

jo 4 ke zyaada kareeb hai usko K_1 me daalo

$K_2 = \{10, 11, 12, 20, 25, 30\}$

jo 12 ke zyaad kareeb hai usko K_2 me daalo

ab hrsme mean nikalo Par asli wala K_1 ka mean and K 2 ka mean

$$m_1 = \frac{2+3+4}{3} = 3$$

$$m_2 = \frac{10+11+12+20+25+30}{6}$$

$$m_2 = \frac{108}{6} = 18$$

$\therefore m_1 = 3$ $\quad$ $m_2 = 18$

same Process cluster banao

$K_1 = \{2, 3, 4, 10\}$ $\quad$ $K_2 = \{11, 12, 20, 25, 30\}$

$m_1 = 4.75$ $\quad$ $m_2 = 19.6$

$K_1 = \{2, 3, 4, 10, 11, 12\}$ $\quad$ $K_2 = \{20, 25, 30\}$

$m_1 = 7$ $\quad$ $m_2 = 25$

$K_1 = \{2, 3, 4, 10, 11, 12\}$ $\quad$ $K_2 = \{20, 25, 30\}$

$m_1 = 7$ $\quad$ $m_2 = 25$

Thus we are getting same mean we have to stop K_1 and K_2 are new cluster.

Agglomerative with Dendrogram.

Item	E	A	C	B	D
E	0	1	2	2	3
A	1	0	2	5	3
C	2	2	0	1	6
B	2	5	1	0	3
D	3	3	6	3	0

Consider only one part (lower trangular part)

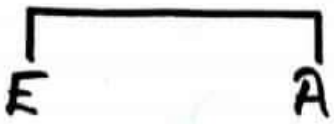
Item	E	A	C	B	D
E	0				
A	1	0			
C	2	2	0		
B	2	5	1	0	
D	3	3	6	3	0

Now find minimum distance

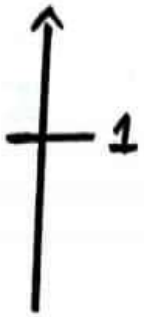
minimum distance = 1 for E, A

Hence merge EA

Dendrogram



Distance



Find min distance from EA to other

$$\min\left[\text{dist}\{(E,A),C\}\right]$$

$$= \min\left[\text{dist}(E,C), \text{dist}(A,C)\right]$$

$$= \min[2,2] = 2$$

$$\min\left[\text{dist}\{(E,A),D\}\right]$$

$$= \min[3,3] = 3$$

dist (ED, AD)
= 3

$$\min\left[\text{dist}\{(E,A),B\}\right]$$

$$= \min[2,5]$$

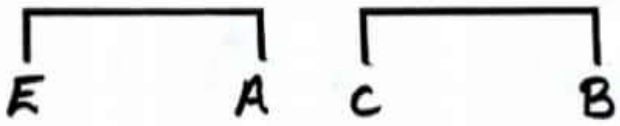
$$= 2$$

dist(EB, AB)

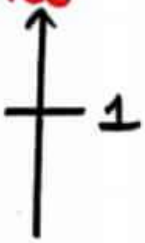
Item	EA	C	B	D
EA	0			
C	2	0		
B	2	1	0	
D	3	6	3	0

In this minimum Distance = 1 for C,B.

Dendrogram



Distance



Now again Find min distance From CB to other point.

min dist $[(CE, A), (C, B)]$

$=$ min dist $[(E, C), (E, B), (A, C), (A, B)]$

$=$ min dist $[2, 2, 2, 5] = 2$.

min dist $[(C, B), D]$

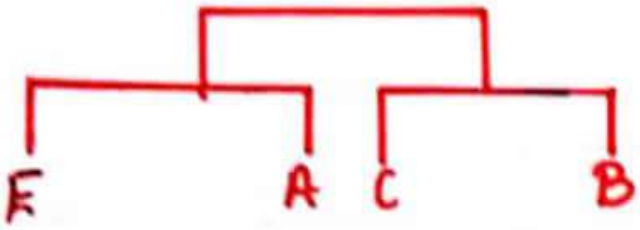
$=$ min. dist $[(C, B), D]$

$=$ min dist $[(C, D), (B, D)]$

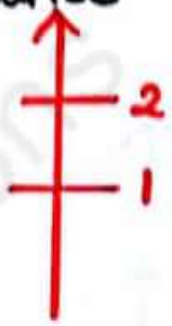
$=$ min. dist $[6, 3] = 3$

Item	EA	CB	D
EA	0		
CB	2	0	
D	3	3	0

Dendrogram



Distance



Now minimum distance = 2 for EA, CB

min. dist [(E,A,C,B), D]

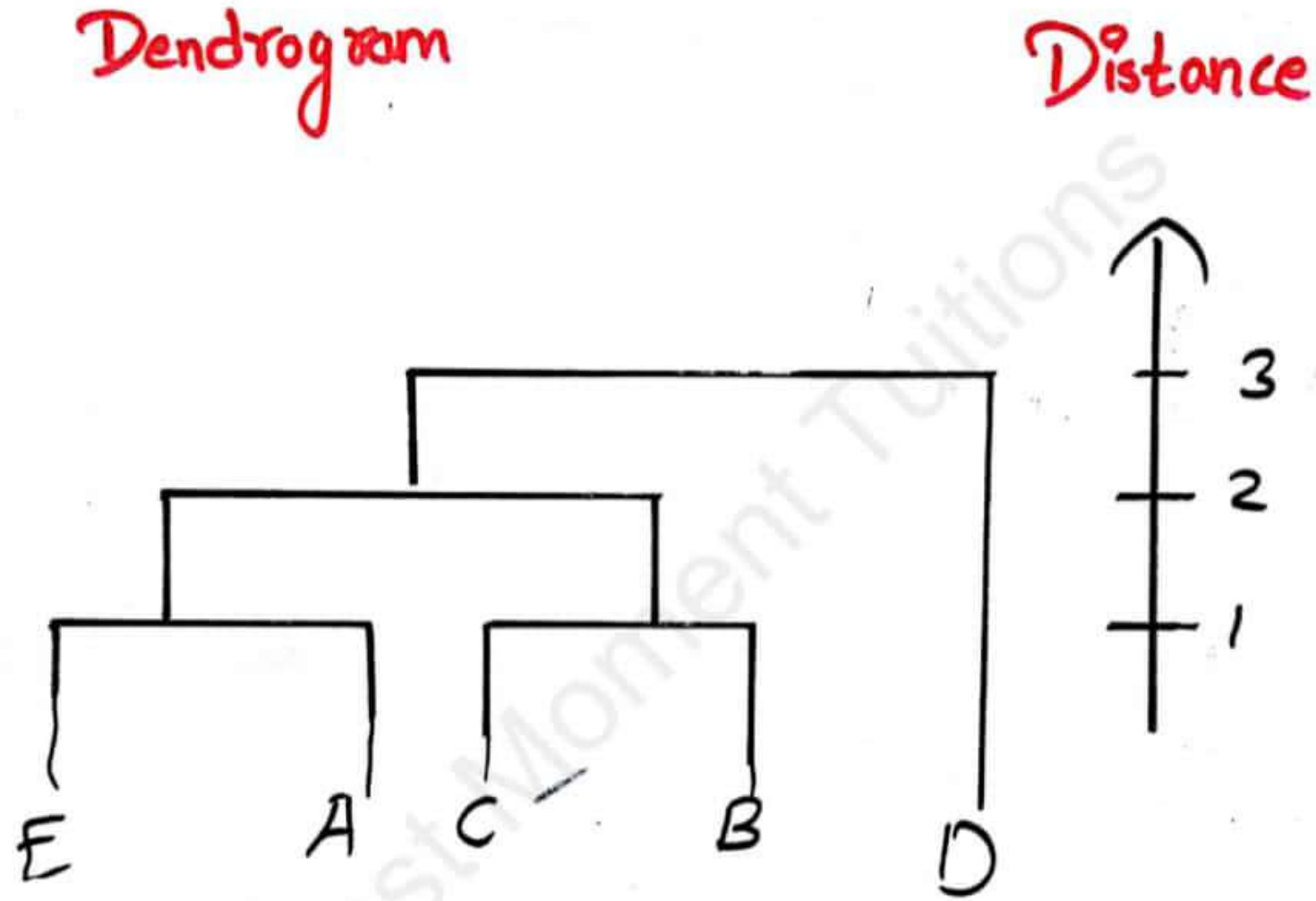
= min. dist [(E,D), (A,D), (C,D), (B,D)]

= min dist [3, 3, 6, 3]

= 3

	EA CB	D
EA CB	0	
D	3	0

Dendrogram



Chapter 10

Mining Frequent Pattern and Association rule

A Frequent pattern is a pattern (a set of items, subsequences, subgraphs, etc.) that occurs frequently in a data set.

Need of Association Mining:

- Frequent mining is generation of association rules from a Transactional Dataset.
- If there are 2 items X and Y purchased frequently then it's good to put them together in stores or provide some discount offer on one item on purchase of other item.
- This can really increase the sales. For example it is likely to find that if a customer buys **Milk** and **bread** he/she also buys **Butter**.
- So the association rule is **['milk]^['bread']=>['butter']**. So seller can suggest the customer to buy butter if he/she buys Milk and Bread.

Apriori Algorithm

- Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules.
- Apriori Uses "Bottom up" Approach where frequent subset are extended one at a time.
- Apriori is design to operate on database containing transaction (for Example Collection of item bought by customer)

Example : Find Association Rule

Apriori

min - supprot = 60%
min - confidence = 80%

Support = $\frac{60}{100} \times 5 = 3$

T- ID	Itemset
T- 1000	M, O, N, K, E, Y
T- 1001	D, O, N, K, E, Y
T- 1002	M, A, K, E
T- 1003	M, U, C, K, Y
T- 1004	C, O, O, K, E

Now find support count of each Itemset.

C_1 =

Itemset	Support Count
M	3
O	4
N	2
E	4
Y	3
D	1
A	1
U	1
C	2
K	5

Compare min support with each Itemset support count.

L_1 =

Itemset	Supp - Count
M	3
O	4
K	5
E	4
Y	3

generate pair to generate C_2

C_2 =

Itemset	Supp-Count
M,O	1
M,K	3
M,E	2
M,Y	2
O,K	3
O,E	3
O,Y	2
K,E	4
K,Y	3
E,Y	2

Now again compare C_2 with min-Support

L_2 =

Itemset	Supp-Count
M,K	3
O,K	3
O,E	3
K,E	4
K,Y	3

make pair to generate C_3

C_3 =

Itemset	Supp - Count
M, K, O	1
M, K, E	2
M, K, Y	2
O, K, E	3
O, K, Y	2

Now again compare the itemset with min-support

L_3 =

Itemset	Supp- Count
O, K, E	3

Now create association rules with support and Confidence

For O,K,E

Association rule	Support	confidence	confidency
O∧K ⇒ E	3	3/3 = 1	100
O∧E ⇒ K	3	3/3 = 1	100
K∧E ⇒ O	3	3/4 = 0.75	75
E ⇒ O∧K	3	3/4 = 0.75	75
K ⇒ O∧E	3	3/5 = 0.6	60
O ⇒ K∧E	3	3/4 = 0.75	75

Compare this with min - confidence = 80%

rules	Support	confidence
O∧K ⇒ E	3	100
O∧E ⇒ K	3	100

Hence Final association rules are

O∧K => E
O∧E => K

} market - basket analysis

FP Growth Algorithm

- Fp Growth Algorithm stand for Frequent pattern growth Algorithm.
- FP growth algorithm is an improvement of apriori algorithm
- FP growth algorithm used for finding frequent itemset in a transaction database without candidate generation.

Advantages of FP growth algorithm:-

1. Faster than apriori algorithm
2. No candidate generation
3. Only two passes over dataset

Disadvantages of FP growth algorithm:-

1. FP tree may not fit in memory
2. FP tree is expensive to build

FP TREE

TID	Item bought
1	f, a, c, d, g, i, m, p
2	a, b, c, f, L, m, o
3	b, f, h, j, o
4	b, c, K, s, p
5	a, f, c, e, l, p, m, n

min- support = 3

To Solve this sum we use following 4 step.

Step 1.] Find minimun support of each item.

Step 2.] Order frequent itemset in desending order. (Consider only items with high or equal to min support)

Step 3.] Draw FP tree.

Step 4.] Minning Fuequent Pattern from FP tree.

Step 1] Find minimum support of each item.

Item	Support
a	3
b	3
c	4
d	1
e	1
f	4
g	1
h	1
i	1
j	1
k	1
l	2
m	3
n	1
o	2
p	3
s	1

item	Support
a	3
b	3
c	4
f	4
m	3
p	3

Step 2] Order frequent item in desending order.

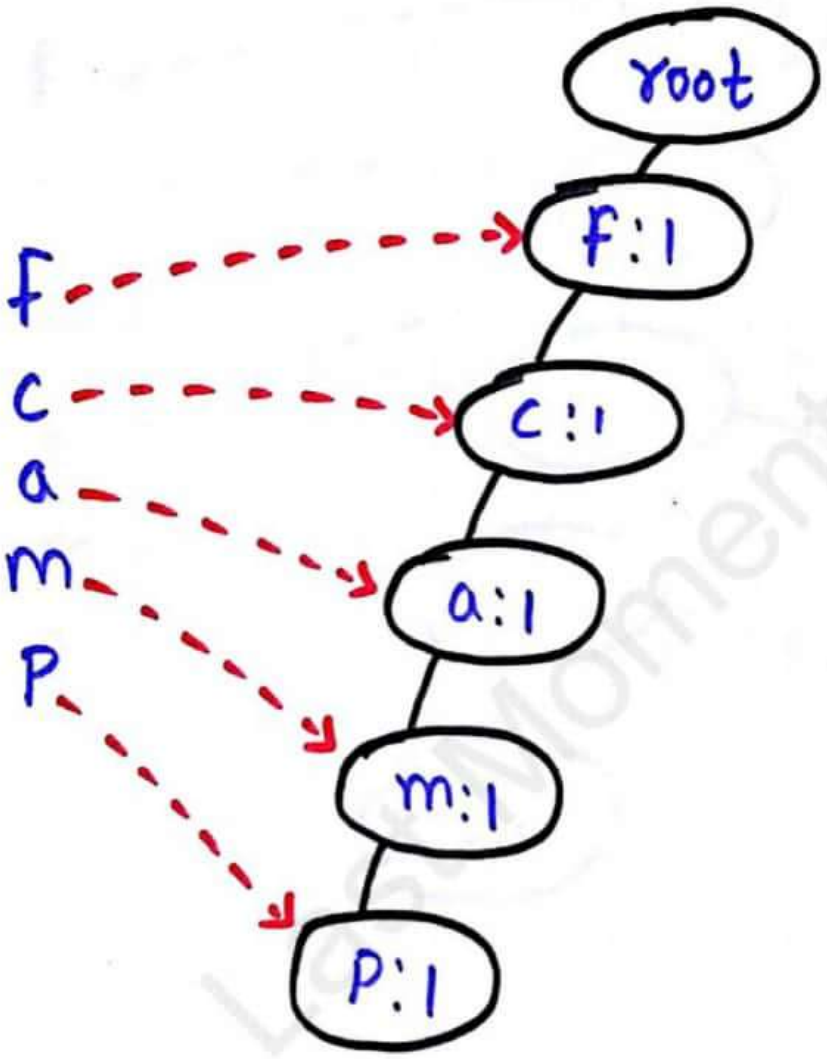
TID	Item bought	Cordered frequent items
1	f, a, c, d, g, i, m, p	f, c, a, m, p
2	a, b, c, f, L, m, o	f, c, a, b, m
3	b, f, h, j, o	f, b
4	b, c, K, s, p	c, b, p
5	a, f, c, e, p, m, n	f, c, a, m, p

f = 4, c = 4, a = 3, b = 3, m = 3, p = 3

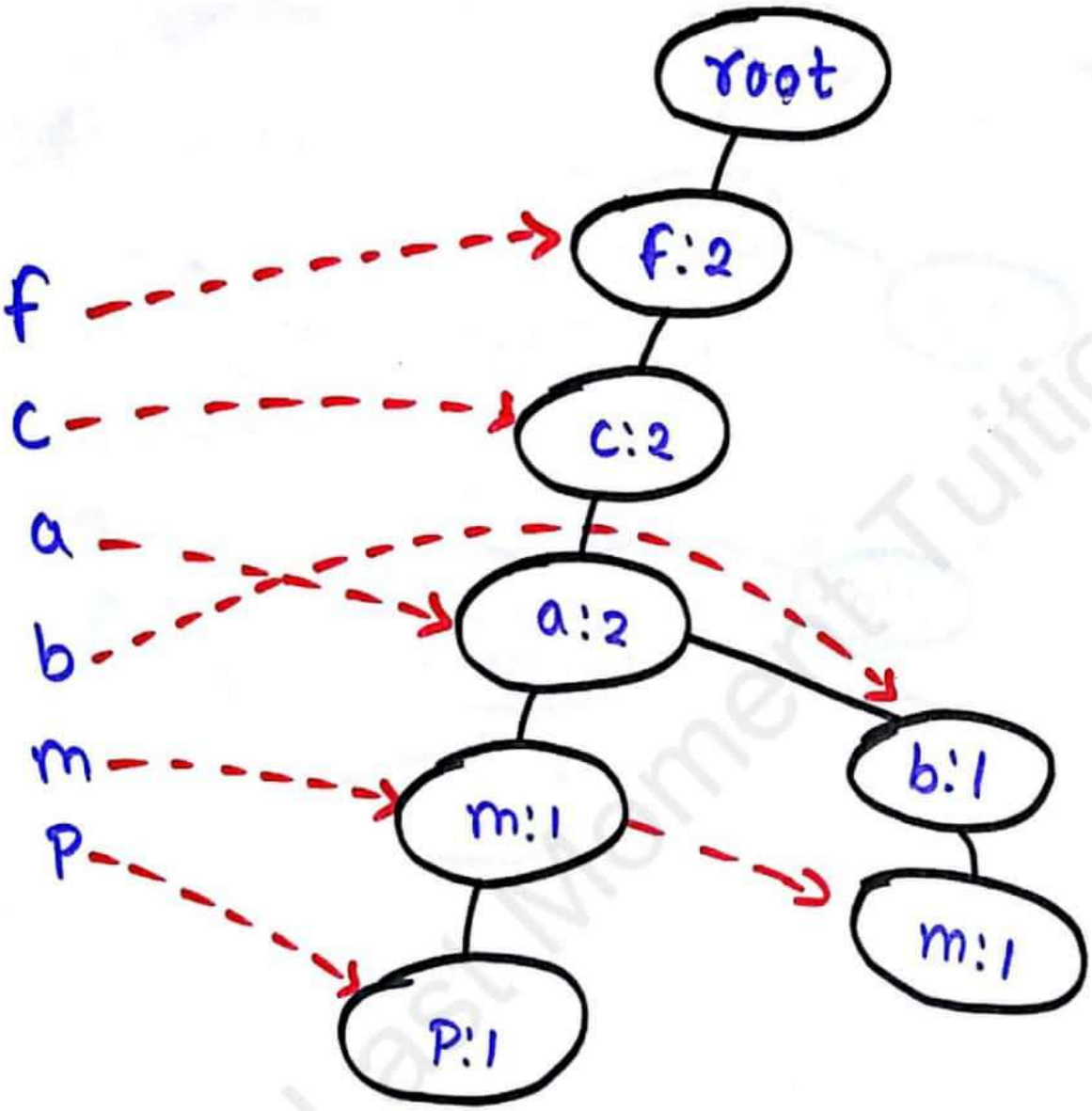
Step 3] Draw FP tree

●] root

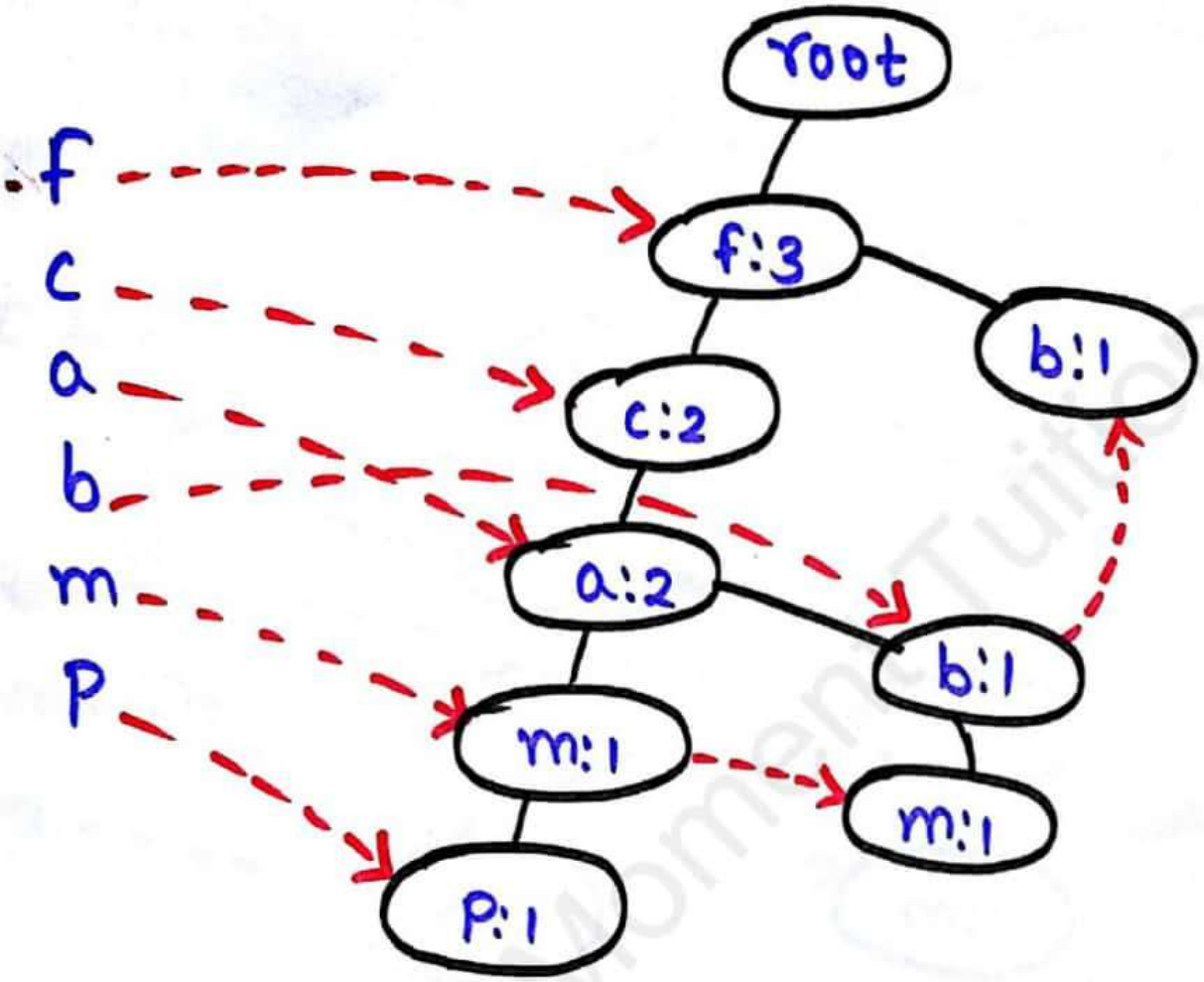
1] insert first transaction



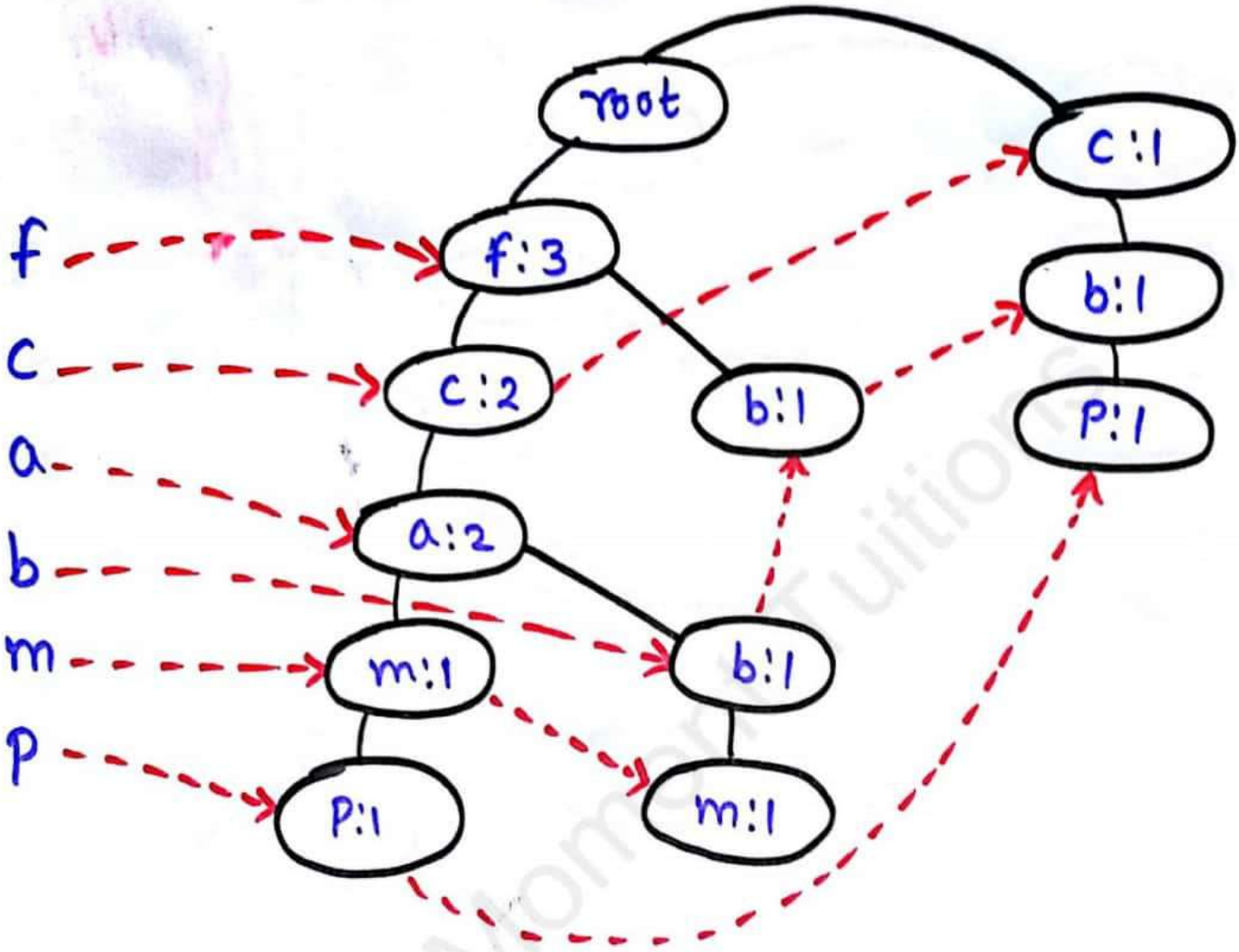
2] Insert Second transaction.



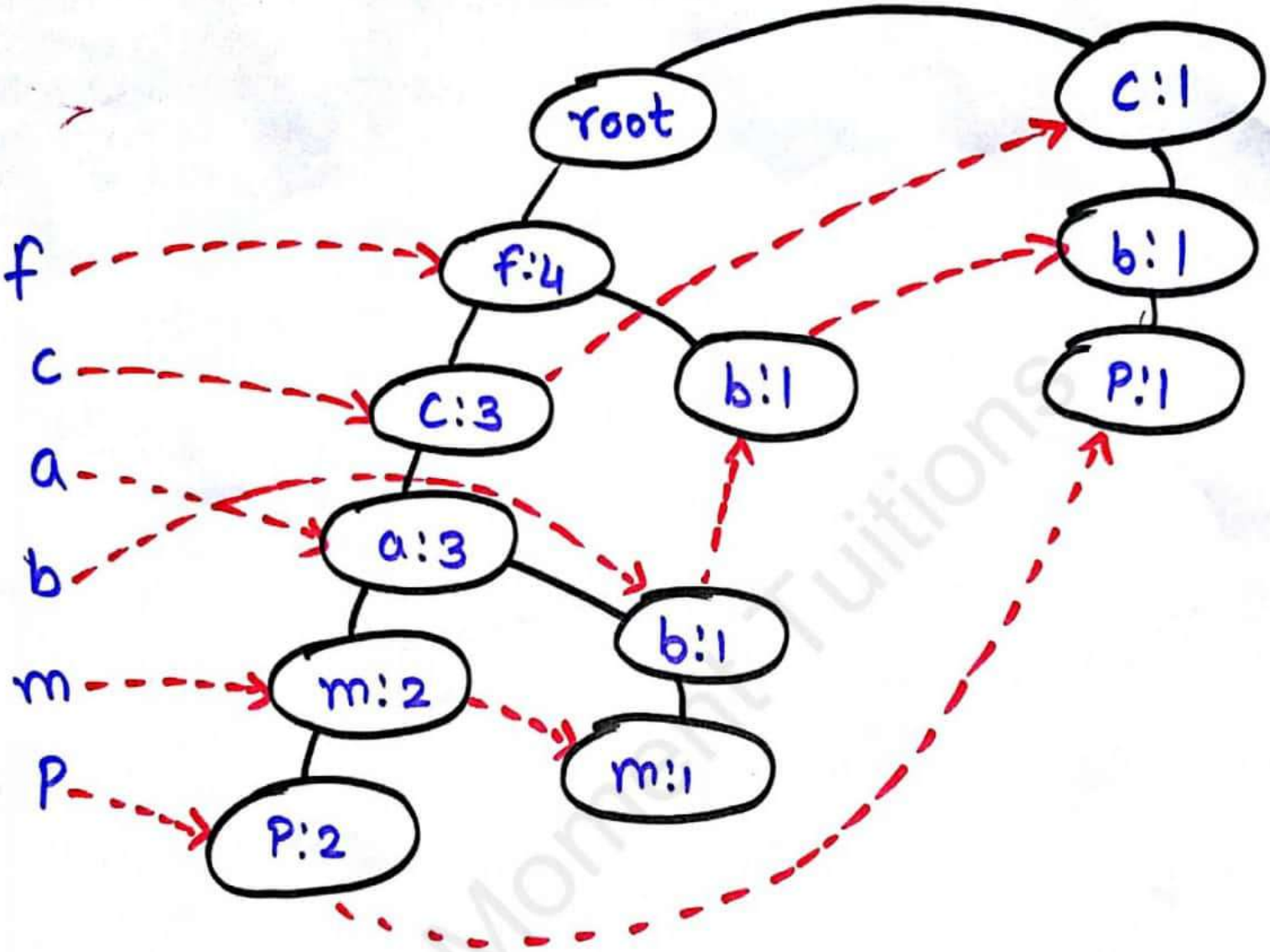
Insert third transaction



Insert fourth transaction



Insert fifth transaction



item	Conditional Pattern base	Conditional FP tree
P	(fcam:2) (cb:1)	(c:3) | P
M	(fca:2), (fcab:1)	(f:3, c:3, a:3) | m
B	(fca:1) (f:1) (c:1)	Empty
A	(fc:3)	(f:3, c:3) | a
C	(f:3)	(f:3) | c
F	Empty	Empty

Strongly suggest to see

Sandeep maheshwari

videos

To change your life