Imp Topics

13/02/2023

What is data warehouse :-

-> Data warehousing is a process of collecting and organizing large and complex data from various sources into a central repository, know as data warehouse, for reporting and analysis.

-> The primary goal of data warehousing is to support decision making by Providing business users with single, unified and integrated view of the data

-> A data warehouse is a subject-oriented, integrated, time variant and non-volatile collection of data.

Subject-oriented :-

data warehouse is organized around major subjects such as customer, sales products; supplier, etc.. Rather than concentrating on day to day transactions of an organization. data warehouse focuses on modelling and analysis of data for decision makers.

Integrated :-

A data warehouse is usually constructed by integrating various multiple heterogenous sources such as relational database, flat files etc... data cleaning and integration ensure consistancy in data warehousing.

Time variant :-

Data is stored to provide information from an historic perspective (Past 5-10 years). every key structure in data warehouse contains, either implicitly or explicitly a time element.

Non-Volatile :-

operational updates of data does not occure in data warehouse env. It does not req transaction recovery processing and concurrency control mechanisms. It only requires inital loading of data and access of data.

Difference between DBMS and data warehouse:-

-> A database management system (DBMS) and a Data warehouse (DW) are both systems used to store and manage data, but they serve different Roposes and have different characteristics.

-> A DBMS is a software system that is designed to manage a database and provides users with the ability to interact with the data base through various interfaces, such as SQL. The primary purpose of DBMS is to provide a centralized location for data storage and retrival, and to enforce rules for data integrity, consistency and security.

-> A data warehouse is a large, centralized repository of data that is specifically designed for use by decision makers and business analysts. The data in a data warehouse is usually integrated from multiple sources, and it is transformed and summarized to support business intelligence activits such as data analysis and reporting. Unlike a DBMS a data warehouse is optimized for querying and analysis, rather than transactional processing.

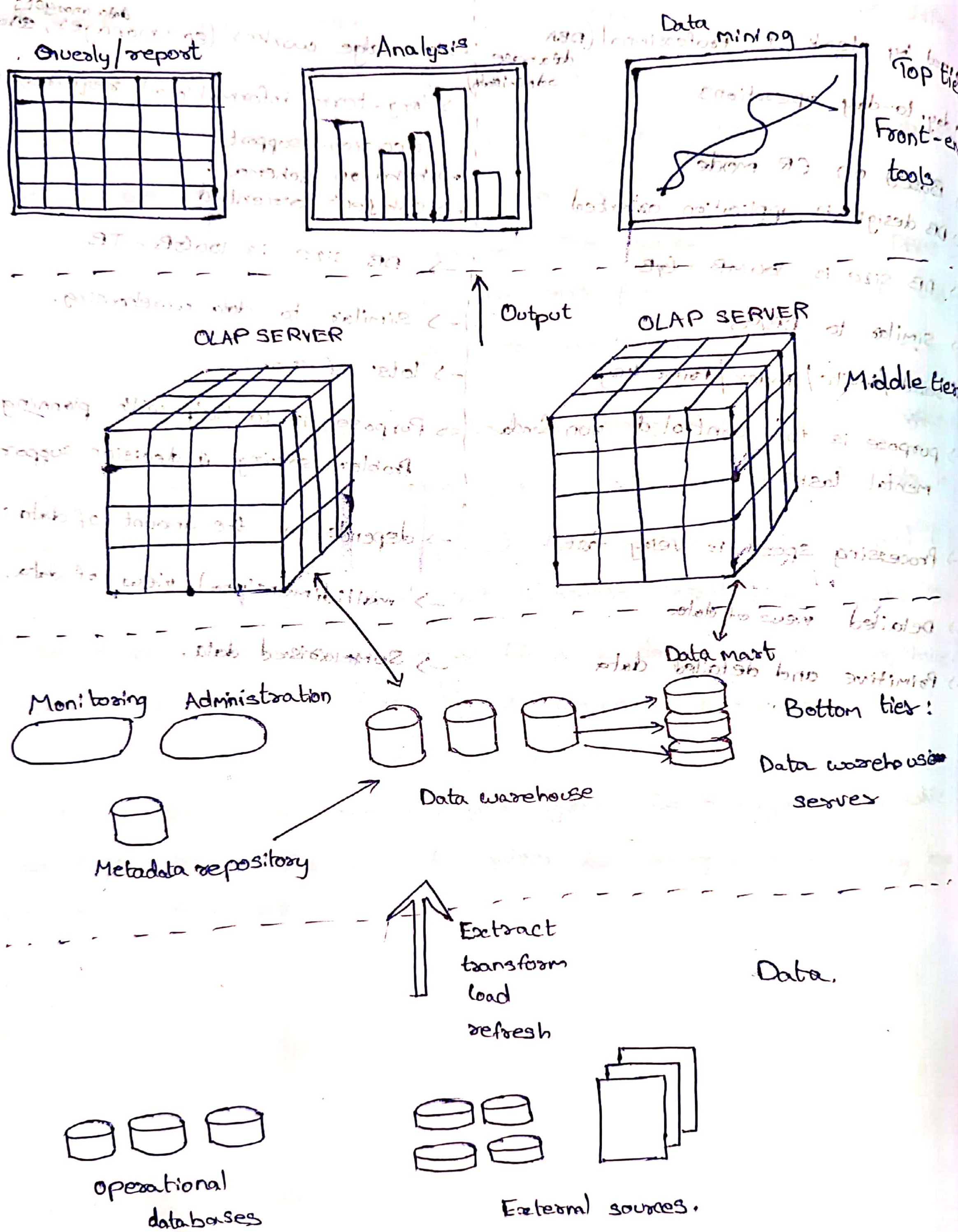
-> In sommary, a DBMS is a general-purpose system for managing data, while a data warehouse is a specialized system for storing and analyzing large amount of historical data.

Difference b/w operational database system and Data warehousing.

OLTP (online Transaction Processing)	OLAP (online Analytical Processing)
-> used by clerk, IT Professional (DBA database administrator)	-> knowledge worker (eg: manager, data analysts, analyst)
-> day-to-day operations	-> long-term informational requirements decision support.
-> Based on ER model	-> based on schema.
-> DB design is application oriented	-> subject oriented.
-> DB size is 100MB - GB	-> DB size is 100GB - TB
-> similar to DBMS	-> similar to data warehousing.
-> read/write/index/Prime key.	-> lots of scans
-> purpose is to control & run fundamental tasks	-> Purpose is to help with planning, Problem solving & decision support
-> Processing speed is very fast	-> depends on the amount of data involved
-> Detailed view of data	-> multidimensional view of data.
-> Primitive and detailed data	-> Summarised data.

3-Tier Architecture :-

→ Data warehouses often adopt a three-tier architecture.



Bottom tier:- This is also known as Database Tier. This is bottom-most layer and is responsible for storing the data. It communicates with the application tier to provide data when requested and to update the data when changes are made. This layer is also known as the data storage layer or the data layer.

Middle tier:- This is also known as Application Tier. This layer is responsible for processing the data and handling all of the business logic. It communicates with the database tier to retrieve and update the data and with the presentation tier to display the data. This layer is also known as the logic layer or the middleware layer.

Top tier:- This is also known as presentation Tier. This is the top-most layer and is responsible for presenting data to the user. It usually contains the user interface and any client-side logic that is required to interact with the user. This layer is also known as the Client layer or the user interface layer.

The three-tier architecture allows for separation of concerns, scalability, and improved security, as each tier can be managed and updated independently of the others. The 3-tier architecture is commonly used in web applications, where the presentation tier runs in a web browser, the application tier is hosted on a web server, and the database tier is hosted on a separate database server.

* Data modeling:-

Process of designing schema of detailed and summarized info of datawarehouse. The main objective of the data modeling is to improve efficiency of the system and to support complex queries.

-> 3 levels :-

1. conceptual : mainly explains the semantics of data i.e simply the meaning of the data.

2. Logical : defines all the information (or) data in a structural format using data structures ; rules ; processes ; relationships ; etc...

3. Physical : defines how the data is represented (like in form the data is represented like tables (or) views etc...

* Data warehouse models:- (Types)

1. Enterprise Warehouse:-

it will Contains all the information about subjects related to a "entire" organisation both detailed and summarized . This is implemented on mainframes / Superservers . Range from (gb to Tb)

2. Data mart :-

it will have data specific to group of users (not entire organisation) and only have summarized data.

3. Virtual warehouse :

Contains the data copied from multiple source during the process

* Fact Table:-

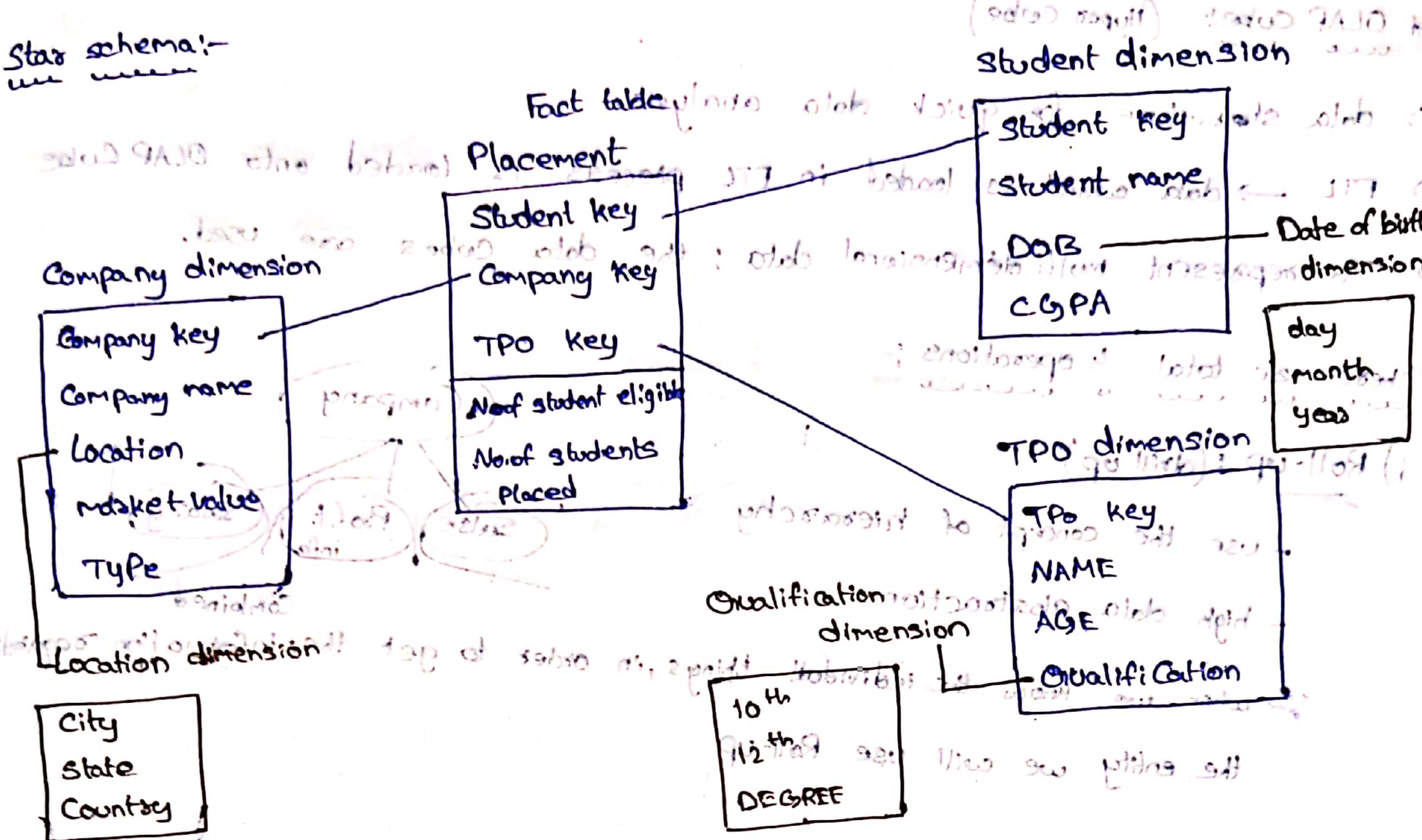
-> Relations between multidimensional data

* Dimension Table :-

-> Tables related to each dimension and helps in describing dimension further.

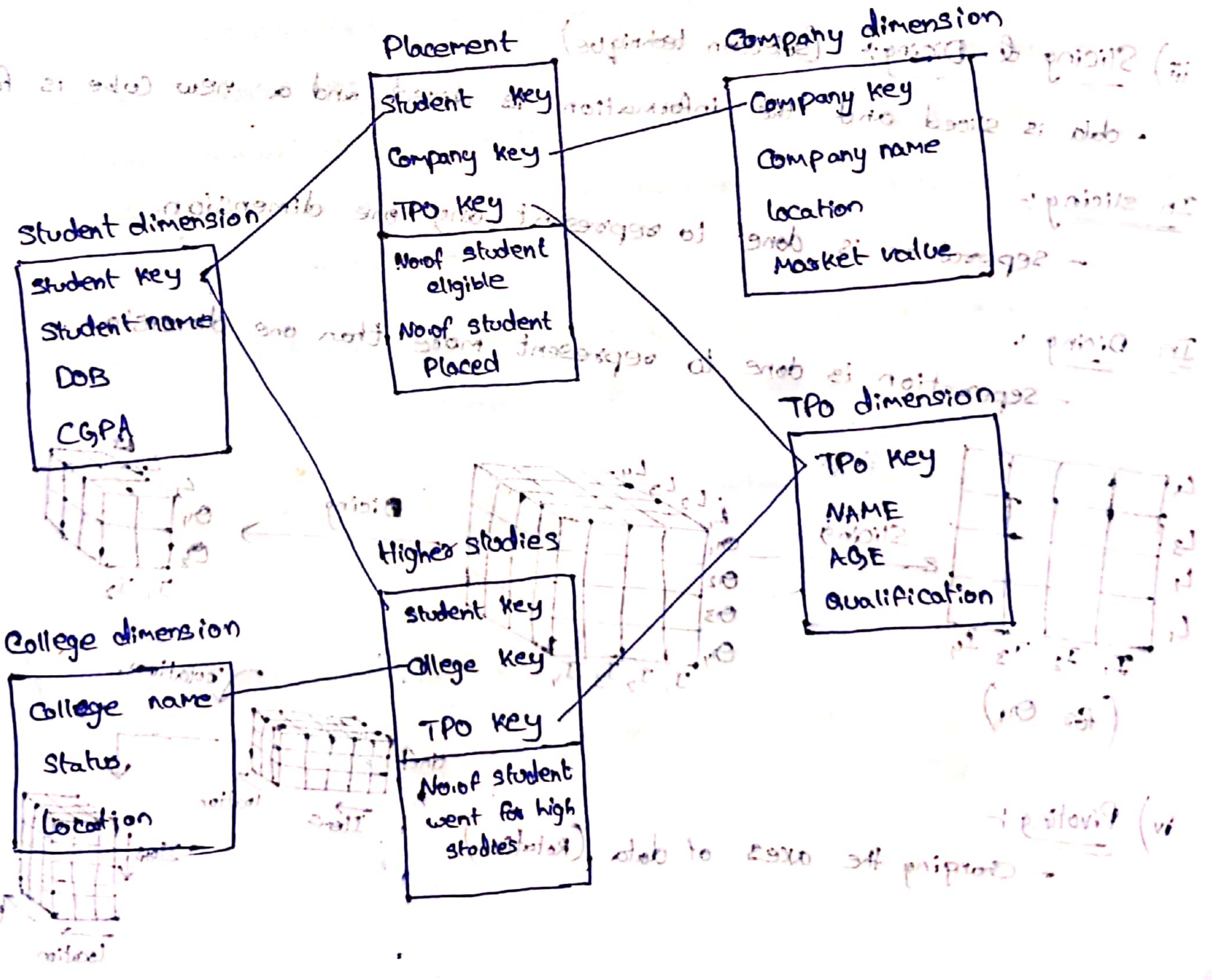
* Measures in a schema refer to numerical or quantitative data that is stored in schema

* Star schema:-



* Snow flake schema

* Fact constellation schema:-



* OLAP Cube:- (Hyper Cube)

-> data structure for quick data analysis.

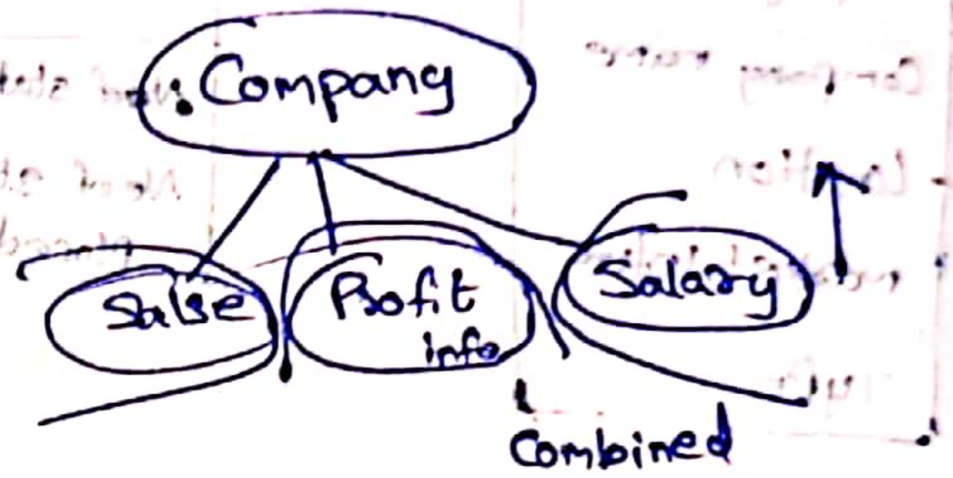
-> ETL → data which is loaded in ETL process is loaded onto OLAP Cube

-> To represent multi dimensional data ; the data cubes are used.

These are total 4 operations :-

i) Roll-up : (drill up)

- use the concept of hierarchy
- high data, abstraction
- when we know the individual things, in order to get the information regarding the entity we will use Roll-up



ii) Drill-down :

- Revesse of roll-up operation (data is drilled down)
- low data abstraction.

iii) Slicing & Dicing :- (selection technique)

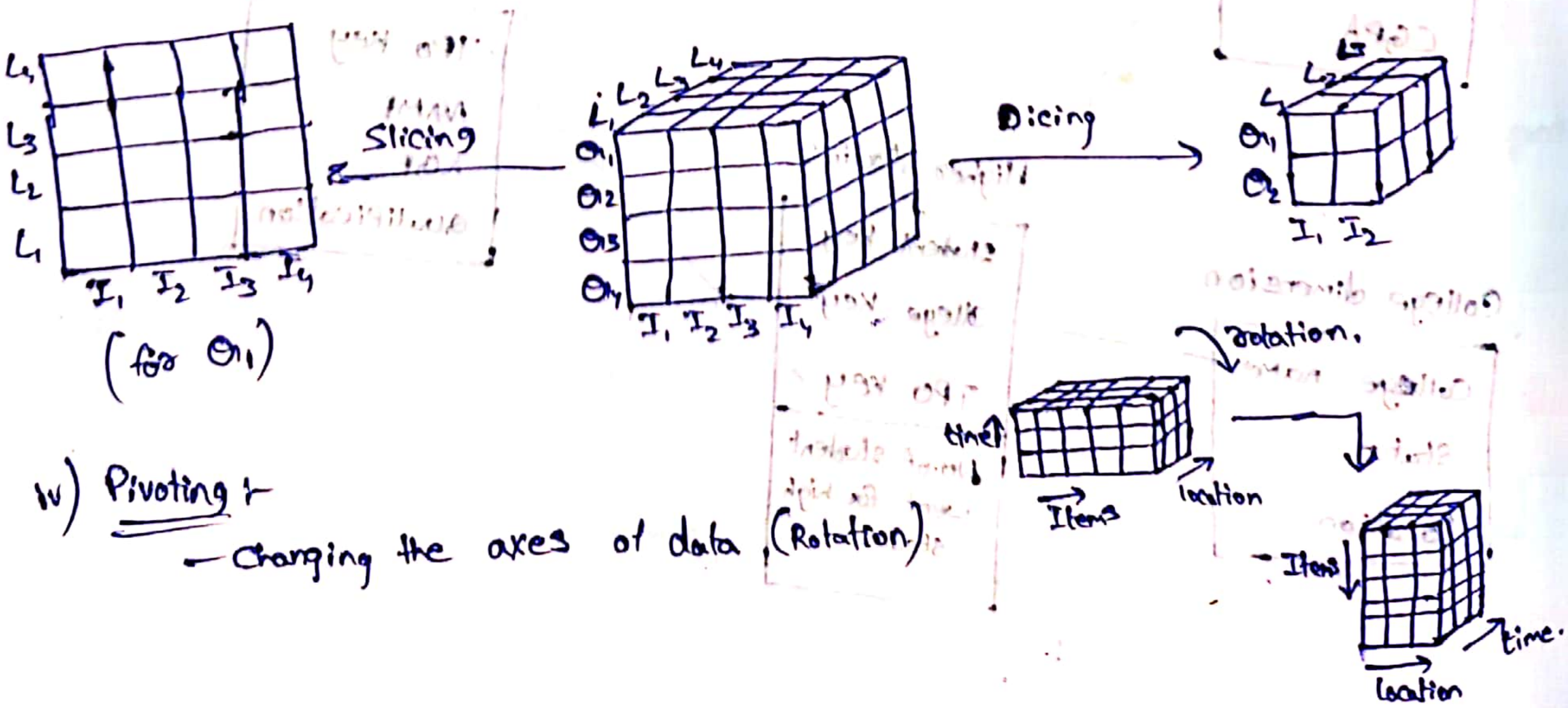
- data is sliced and the information is dived and a new cube is formed

In slicing :-

- seperation is done to represent any one dimension.

In Dicing :-

- seperation is done to represent more than one dimension.

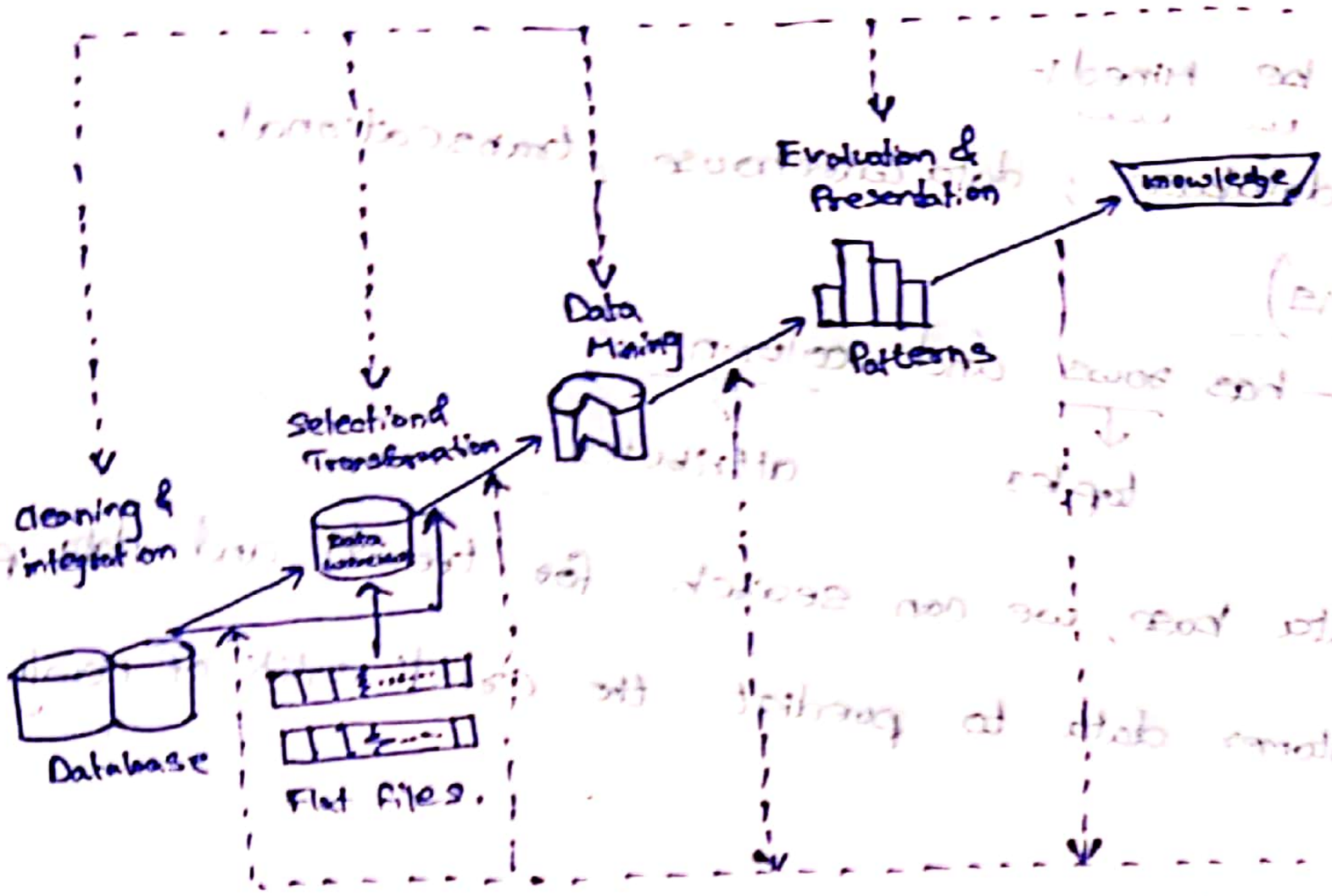


iv) Pivoting :-

- Changing the axes of data (Rotation)

KDD:-

The knowledge discovery in database process is a widely used methodology for extracting useful knowledge from the data. The KDD process typically involves the following steps:-



1. Data cleaning:- In this the raw data is cleaned and pre-processed to remove the inconsistencies, errors and missing values.
2. Data Integration:- Data from multiple sources are combined into a single dataset.
3. Data selection:- Relavent data is seleted from the integrated dataset for analysis.
4. Data Transformation:- The selected data is transformed into form by reducing that can be analyzed.
5. Data Mining:- The transformed data is analyzed by using data mining algorithms to identify patterns, trends and relationships.
6. Pattern evaluation:- The identified patterns are evaluated to determined their significance and usefulness.
7. knowledge presentation:- The evaluated patterns are presented in a form that can be easily understood and interpeted.

* DATA MINING:-

- Data mining is defined as procedure of extracting info from huge sets of data.
- also defined as mining knowledge from data.

What type of data can be mined:-

Mainly 3 - Types : database ; data warehouse ; transcational.

i) Database data (RDBMS)

- set of tabels - has rows and columns
  - rows → tuples
  - columns → attributes

- while mining data base, we can search for trends and data patterns.

example:- i) Analysing customer data to predict the credit risk of newly coming customer.

ii) Analysing sales data.

ii) Datawarehouse data:-

- Datawarehouse is a process of collection and orginizing large and complex data from various sources into a central repository for reporting & analysis.
- In datawarehouse, data is stored in multidimensional structure (data cube), where each dimension will represent each attribute.

iii) Transcational data:-

- Each record is called as transaction.
- Sales, flightbooking, user clicks on web page.
- From this we can mine frequent patterns.

Other-types of data:

- Sequence data, data streams, spatial data, engineering design data, hypertext, multimedia, web data, etc...

* Interestness of patterns :-

- This rise three questions :

i) What makes pattern interesting?

- easily understood by humans
- valid on new/test data
- potentially useful. (fufill all needs)

ii) Can data mining system generate all of the interesting patterns?

- refers to completeness of a dm system.
- In reality it is impossible to generate all interesting patterns.

iii) Can data mining system generate only interesting patterns?

- refers to optimization of a dm system.
- generating only interesting patters → challenging.
- if it is done, then it become easy and efficient for user.

* Outlier analysis is a technique used in data analysis to identify and examine data points that are significantly different from majority of data points. these points lie far from center of distribution and may effect skewing results; decreasing accuracy or creating bias.

→ Outlier analysis involve :

- Identify the outliers (examin the data to find the outliers)
- analyzing the outliers (examin the outlies is it true or it is usuale data point)
- Remove (or) handling outliers (remove or transform or use robust stastical methods) → like, Z score; boxplot; scatter plot, etc.

→ outlier analysis ensure accuracy and reliability of results obtained from the statistical analysis.

-> Proximity measure similarity (or) dissimilarity.

-> Can take two or more states

$$d(i,j) = \frac{p-m}{p}$$; where p refers to total number of attributes & m refers to total matchs.

Eg:-

1. Red A P_1
2. Green B P_2
3. Blue B P_3
4. Red A P_4

$d(1,2) \Rightarrow \frac{3-2}{3} = 1/3 = 0.3$

$d(1,3) \Rightarrow \frac{3-0}{3} = 1$

$d(1,4) \Rightarrow \frac{3-2}{3} = 1/3 = 0.3$

$d(2,3) = \frac{3-1}{3} = 2/3 = 0.6$

$d(2,4) = \frac{3-0}{3} = 1$

$d(3,4) = \frac{3-0}{3} = 1$

-> Proximity measure for ordinal attributes.

-> Proximity measure for binary attributes :-

for similarity :-

$$d(i,j) = \frac{r+s}{q+r+s+t}$$ where, q = no of attributes i.e equal to 1 for both i & j

r = no of attributes i.e equal to "1" for i & "0" for j

s = no of attributes i.e equal to "0" for i & "1" for j

t = no of attributes i.e equal to "0" for both i & j.

distance measure for asymmetric binary attributes :-

$$d(i,j) = \frac{r+s}{q+r+s}$$

Eg:-

Name	Gender	Fever	Cough	Test 1	Test 2	Test 3	Test 4
Jack	M	Y_1	N_0	P_1	N_0	N_0	N_0
Mary	F	Y_1	N_0	P_1	N_0	P_1	N_0
Jim	M	Y_1	P_1	N_0	N_0	N_0	N_0

Geneder is symmetric attributes & remaining are asymmetric attributes

~~for geneder = $\frac{r+s}{q+r+s+t}$ =~~

$q = q_{(Jack, Mary)} = 2$; $r_{(Jack, Mary)} = 0$; $s_{(Jack, Mary)} = 1$;

$t_{(Jack, Mary)} = 3$

for Jack & Mary = $\frac{0+1}{2+0+1} = \frac{1}{3} = 0.\overline{3}$

Jack & Jim :- $q = 1$; $r = 1$; $s = 1$; $t = 3$.

for Jack & Jim :- $\frac{1+1}{1+1+1} = \frac{2}{3} = 0.\overline{6}$

Mary & Jim :- $q = 1$; $r = 2$; $s = 1$; $t = 2$

for Mary & Jim :- $\frac{2+1}{1+2+1} = \frac{3}{4} = 0.75$

→ Proximity measure similarity (or) dissimilarity:

→ Can take two or more states

$d(i,j) = \frac{P-m}{P}$: where P refers to total number of attributes & m refers to total matchs.

Eg:-

1. Red A P_1
2. Green B P_2
3. Blue B P_3
4. Red A P_4

$d(1,2) \Rightarrow \frac{3-2}{3} = 1/3 = 0.3$

$d(1,3) \Rightarrow \frac{3-0}{3} = 1$

$d(1,4) \Rightarrow \frac{3-2}{3} = 1/3 = 0.3$

$d(2,3) = \frac{3-1}{3} = 2/3 = 0.6$

$d(2,4) = \frac{3-0}{3} = 1$

$d(3,4) = \frac{3-0}{3} = 1$

→ Proximity measure for ordinal attributes.

→ Proximity measure for binary attributes :-

for symmetric :-

$d(i,j) = \frac{r+s}{q+r+s+t}$ where, q = no of attributes i.e equal to 1 for both i & j

r = no of attributes i.e equal to "1" for i & "0" for j

s = no of attributes i.e equal to "0" for i & "1" for j

t = no of attributes i.e equal to "0" for both i & j

distance measure for asymmetric binary attributes :-

$d(i,j) = \frac{r+s}{q+r+s}$

Eg:-

Name	Gender	Fever	Cough	Test 1	Test 2	Test 3	Test 4
Jack	M	Y_1	N_0	P_1	N_0	N_0	N_0
Mary	F	Y_1	N_0	P_1	N_0	P_1	N_0
Jim	M	Y_1	P_1	N_0	N_0	N_0	N_0

Geneder is symmetric attributes & remaining are asymmetric attributes

~~for geneder = $\frac{r+s}{q+r+s+t}$ =~~

$q_{(Jack, Mary)} = 2$; $r_{(Jack, Mary)} = 0$; $s_{(Jack, Mary)} = 1$;

$t_{(Jack, Mary)} = 3$

for Jack & Mary = $\frac{0+1}{2+0+1} = \frac{1}{6} = 0.16$ $\frac{1}{3} = 0.\overline{3}$

Jack & Jim :- $q = 1$; $r = 1$; $s = 1$; $t = 3$.

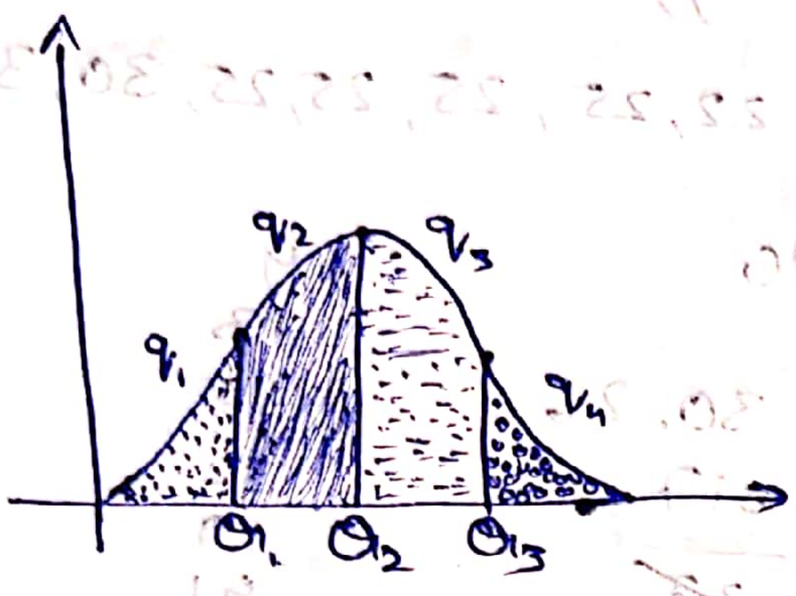
for Jack & Jim :- $\frac{1+1}{1+1+1} = \frac{2}{5} = 0.4$ $= \frac{2}{3} = 0.\overline{6}$

Mary & Jim :- $q = 1$; $r = 2$; $s = 1$; $t = 2$

for Mary & Jim :- $\frac{2+1}{1+2+1} = \frac{3}{4} = \frac{1}{2} = 0.5$ $= 0.75$

01/02/2023

* Quantiles :- are lines which divide the graph into 4 equal quarters.



here Q_1 ; Q_2 ; Q_3 are Quantiles.

q_1 ; q_2 ; q_3 ; q_4 are quarters. (each of 25%)

Q) 13, 15, 16, 16, 19, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

A) as we know that :-

$N = 26$

Mean = 30

Median = 27.5

Mode = 25, 35

q_1 : 13, 15, 16, 16, 19, 20

q_2 : 21, 22, 22, 25, 25, 25

q_3 : 25, 30, 33, 33, 35, 35, 35

q_4 : 35, 36, 40, 45, 46, 52, 70

$Q_1 = 20$; $Q_2 = 25$.

& $Q_3 = 35$

* Box Plot :-

-> Ends of the box plot represent the quantiles (Q_1 & Q_3)

-> The division in the center of the box plot is "median".

-> The two extrem ends outside of the box plot are smallest value & largest value.

-> These lines are known as whiskers.

-> The five point summary of a distribution consists of Q_1; Q_2; Q_3; (here Q_2 = median) smallest value; largest value.

Q) Suppose, that a hospital tested the age and body fat data for 18 randomlly selected adults with the following results. first Calculate the mean, median of age & fat. Draw the box Plot for age & fat.

Age	23	23	27	27	39	41	47	49	50	52	54	54	56	57	58	58	60	61
fat%	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

$N = 18$

Ans) Age :-

mean :- 23, 23, 27, 27, 39, 41, 47, 49, 50, 52, 54, 54, 56, 57, 58, 58, 60, 61

$$\text{mean} = \frac{836}{18} = 46.4$$

7.8, 9.5, 17.8, 25.9, 26.5

median = $\frac{50+52}{2} = 51$

mode = 23, 27, 54, 58.

$Q_1 = 39$; $Q_2 = 51$; $Q_3 = 57$.

Fat% :- 9.5, 26.5, 7.8, 17.8, 31.4, 25.9, 27.4, 27.2, 31.2, 34.6, 42.5, 28.8, 33.4, 30.2, 34.1, 32.9, 41.2, 35.7

518.1

$$\text{mean} = \frac{518.1}{18} = 28.783$$

$$\text{median} = \frac{31.2+34.6}{2} = \frac{65.8}{2} = 32.9$$

mode =

$Q_1 = 31.4$; $Q_2 = 32.9$; $Q_3 = 30.2$