* Data Warehousing and Data Mining :-

* **Data Warehousing :-** It is a coherent collection of subject oriented, integrated, time variant and non-volatile collection of data

Subject oriented :- stores data related to particular subject

eg :- Sales, products

Integrated :- It combines multiple homogeneous and heterogeneous data sources.

Time variant :- collects data related to particular period

eg :- Sales.

Non-volatile :- If you want to add some data the previous data is not changed

* **Data Mining :-**

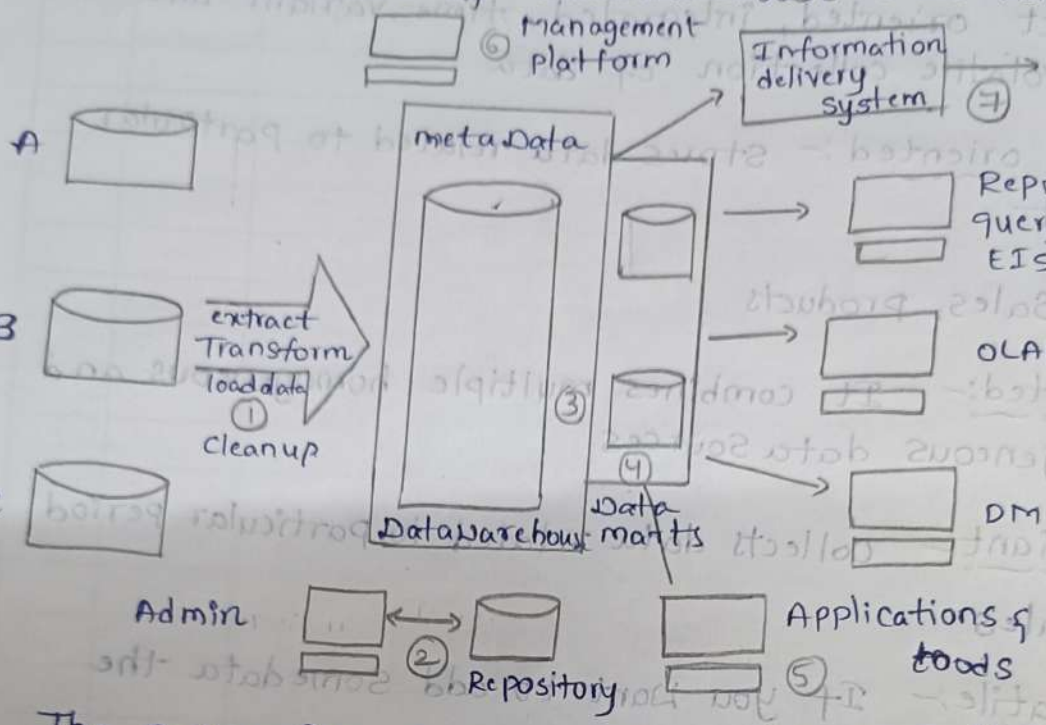
It is the process of extracting information from huge volumes of data in data warehouse is called data mining. apply some data mining techniques to extract information from data warehouse

* **Advantages of DWDM :-**

1) Taking better buisness decisions

* Architecture of DataWarehouse and its components

The Architecture of DataWarehouse is as follows:-



* The Source for DataWarehouse is operational databases (or) files.

2) The Data entered into the DataWarehouse is transformed into an integrated structure and formats

1) DataWarehouse Database:-
This is the central part of DataWarehousing environment. This is implemented based on RDBMS technology.

2) Sourcing, acquiring, cleanup, transform tools:-
1) To remove unwanted data from operational database
2) converting to common datanames & attributes
3) calculating Summaries & derived data

4) Establishing defaults for missing data

3) meta data:- It is data about data. It is used for maintaining and managing and using the data-warehouse.

4) Access tools:-

Its purpose is to provide information to the buisness users for decision making.

1) Data query and reporting tools
2) Application development tools
3) Executive information system tools
4) OLAP tools
5) Data Mining tools

} tools for accessing data

~~Dal~~

5) Data Marts:-

Departmental subsets that focuses on selected subjects.

6) Data Warehouse Admin and Management:-

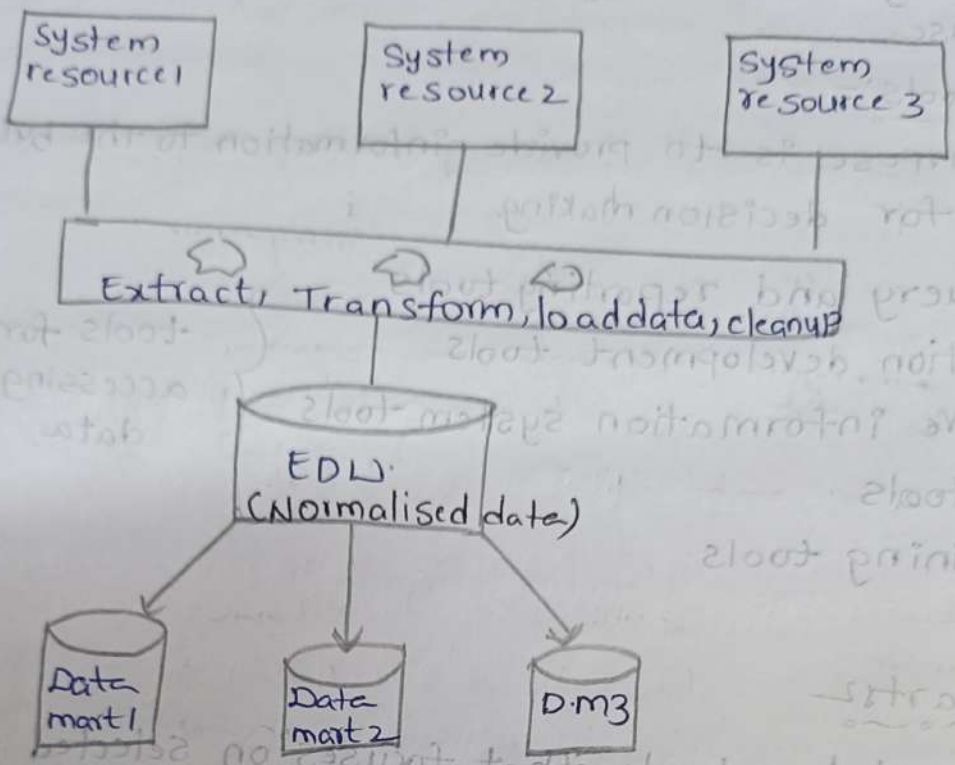
i) Security and priority management
ii) Monitoring updates from multiple sources
iii) Data quality checks
iv) Managing & updating meta data

7) Information delivery system:-

The Information delivery system delivers the information to the data warehouse user.

* Building a data warehouse :-

* Top down approach :-



* There are 2 reasons why organisations consider data warehousing a critical need :-

i) Buisness factors :-

Buisness users wants to take decisions quickly & crctly using all available data

ii) Technical factors :-

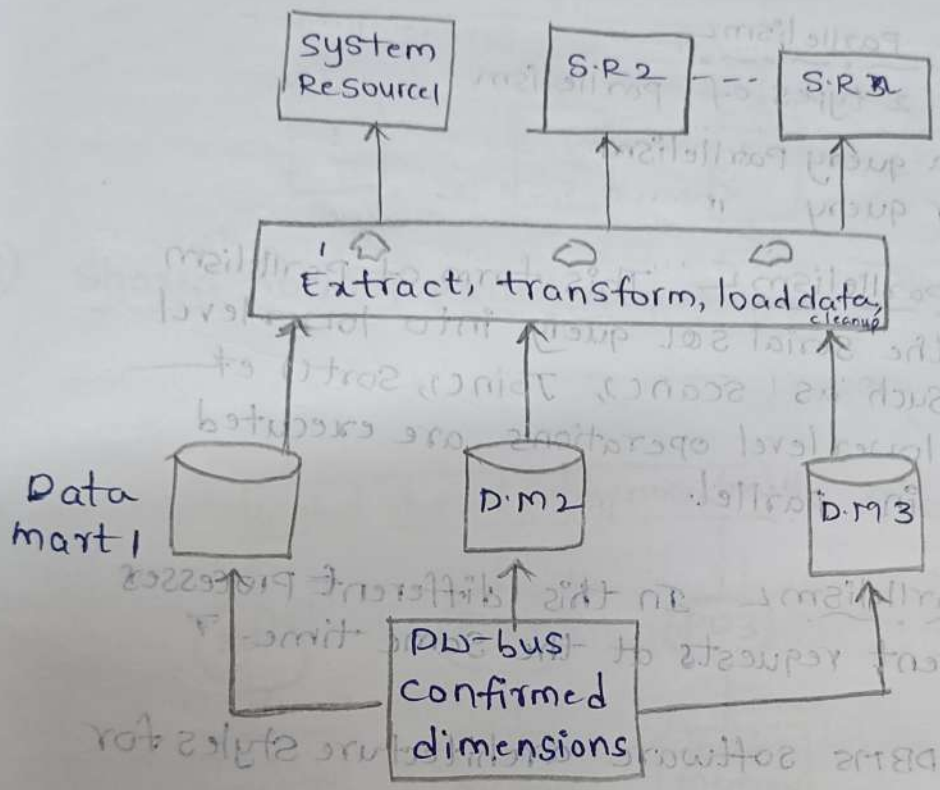
1) TO address the incompability of operational data stores

2) IT Infrastructure is changing rapidly. Its capacity is increasing & cost is decreasing. So, building a data warehouse is easy.

* There are 2 ways to design a data warehouse:-

1) Top down approach
2) Bottom up approach

2) Bottom up approach:-



* Database Architecture for Parllel Processing:-

1) The functions of datawarehouse are based on the relational database technology. The RDBMS technology is implemented in parllel manner.

2) There are 2 advantages of having parllel Relational database technology for data warehouse:-

1) Linear - speed up
2) Linear - scale up

Linear- Speed up:- It refers the ability to increase the number of processers to reduce the response time

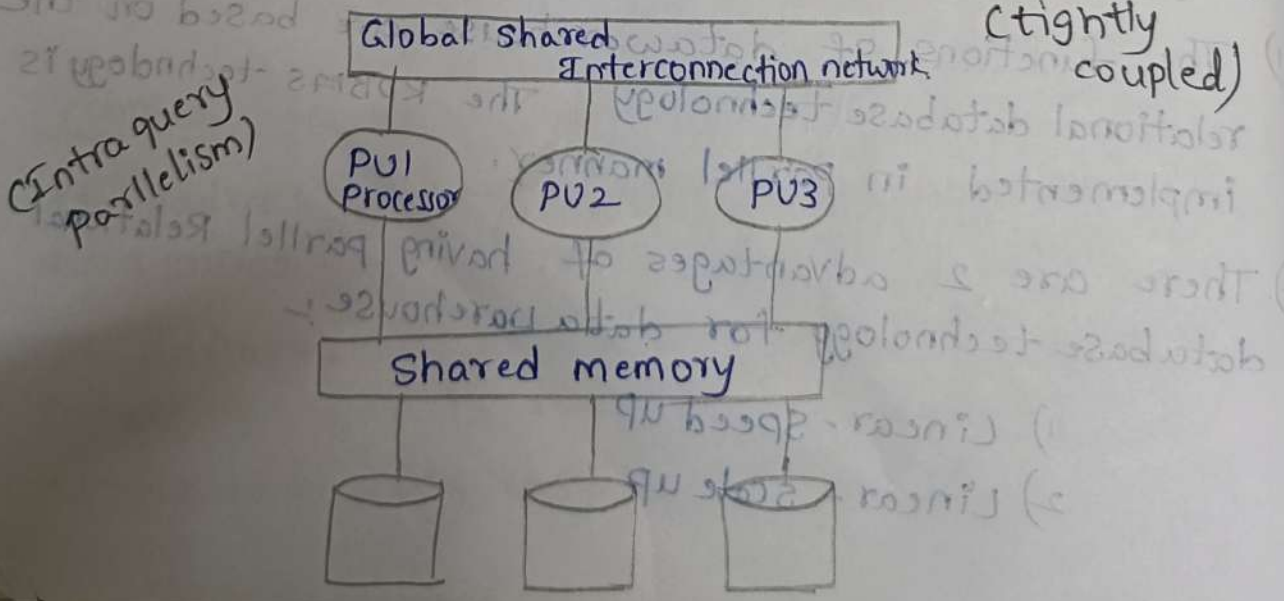
Linear- scale up:- It refers the ability to provide same performance on the same requests as the database size increases.

* Types of parllelism:-
* There are 2 types of parllelism
  1) Intra query parllelism
  2) Inter query "

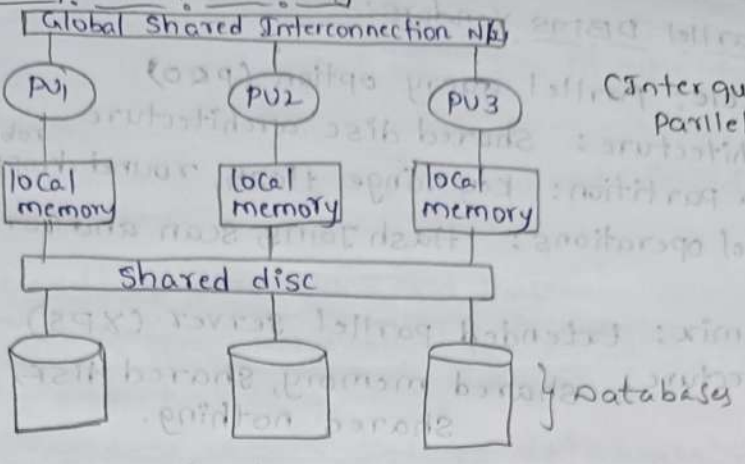
Intra query parllelism:- This form of parllelism decomposes the serial SQL query into lowerlevel operations such as scan(), Join(), Sort() etc then these lowerlevel operations are executed concurrently in parllel.

Inter query parllelism:- In this different processes handle different requests at the same time.

* There are 3 DBMS software architecture styles for parllel processing:-
  1) Shared memory / shared everything parllel processing:

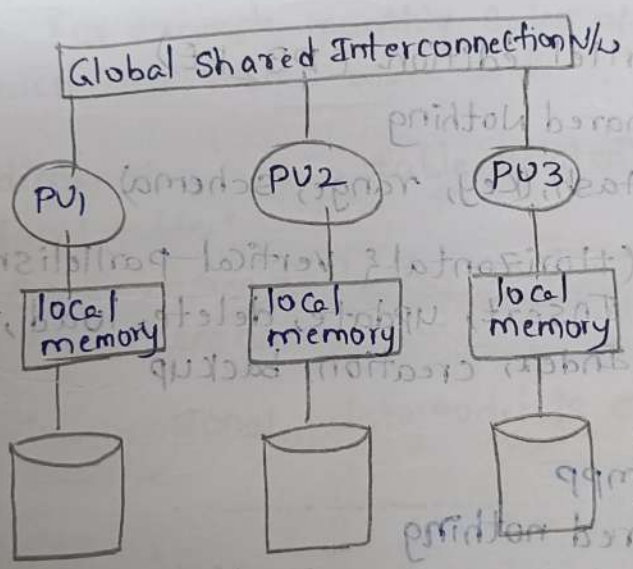


2) Shared disk parllel processing :-



(Inter query parallelism)

3) Shared nothing parllel processing :-



* Parllel DBMS Vendors:-

1) oracle: Parllel query option (PQO)
   Architecture: Shared disc architecture
   Data partition: Key range, Hash, round robin dropping
   Parllel operations: Hash Joins, Scan and Sort

2) Informix: Extended parllel server (XPS)
   Architecture: Shared memory, Shared disk, Shared nothing.
   Data partition: round robin, hash, schema, key range, User defined.
   Parllel operations: Insert, update, delete

3) IBM: DB 2 Parllel edition (DB_2PE)
   Architecture: Shared Nothing
   Data partition: Hash, (Key, range, schema)
   Parllel operations: (~~Horizontal~~ & vertical parllelism)
   Insert, update, delete, load, recovery, Index, creation, backup

4) Sybase: Sybase mpp
   Architecture: Shared nothing
   Data partition: Hash, key, range, schema
   Parllel operations: Horizontal & vertical parllelism.

* Multidimensional datamodel:

1) A multidimensional datamodel views data in the form of datacube.

2) A datacube enables data to be model and viewed in multiple dimensions.

3) It is defined by dimensions and Facts

4) The dimensions are perspectives (or) Entities concentrating which an organisation keeps records

5) ex:- A shop may create a sales datawarehouse to keep records of stores sales for the dimension time, item and location. These dimensions allow the save to keep track of things.

ex-2:- For example monthly sales of items and locations at which the items were sold.

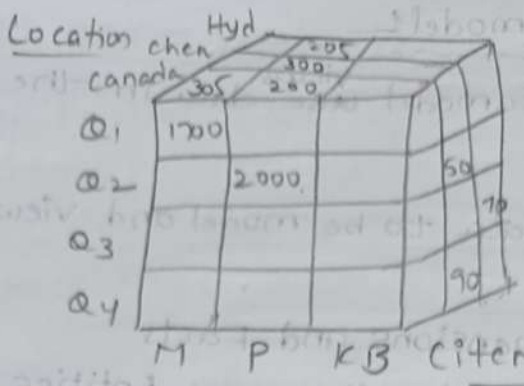
5) Each dimension has a table related to it called a "dimensional table"

ex:- A dimension table for an item may contain the attributes item name, brand and type.

6) A Multidimensional datamodel is organised around a central theme

ex:- Sales

7) This theme is represented by a "fact table". Facts are numerical measures. The fact table contains the names of facts (or) measures of related tables



8) Multidimensional datamodel is based on OLAP tool

* Database Schemas for decision support :-

1) Star schema
2) Snowflake schema
3) Fact constellation

* Concept hirearchy :-

- The mapping of low level form to higher level (or) Generalised form is called 'concept hirearchy."
(or)
- It defines the sequence of mappings from low level concepts to higher-level (or) Generalised concepts

* Types :-

1) Schema hirearchy → cuboid (collection of nodes)
2) Set Grouping hirearchy

Base value, Target value has to be identified. then high (low) level data, converted to low (high) level format

* characterstics of OLAP systems:-

Transactions/ which are performed on hyge volumes of Analysis data is OLAP.

OLTP → Performs on database
OLAP → Performs on data warehouse.

OLAP → complex Bigtasks, OLTP → simple-tasks & easy

→ The Analysis which is performed on previous data → OLAP

* characterstics:-

1) In FASMI characterstics of OLAP methods the term derivered from first letters of the characterstics

F - Fast
A - Analysis
S - Shared
M - Multidimensional
I - Information

Fast:- It defines which the system targeted to deliver the most feedback to the client within about 5 seconds. Taking no more than 1 sec & very few taking more than 20 sec

Analysis:-

Compare OLAP & OLTP Systems:-

	OLTP	OLAP
1. characteristics	Handles a large no. of small transactions	Handles large volumes of data with complex queries
2. Query types	Simple, standardised queries	Complex queries
3. operations	Based on insert, update, delete commands	Based on select comm to aggregate data for reporting
4. Responsetime	milliseconds	Based on the query typ seconds, minutes (or) hours depending on amount of data to proc
5. Design	Industry-specific such as retail, manufacturing (or) Banking	Subject-specific such as Sales., inventory (or) marketing
6. Source	Transactions	aggregated data fro Transactions.
7. purpose	Control & run essential Buisness operations in real-time	Plan, solve problem, support decisions, discover hidden insides
8. Space requirements	Generally, small until historical data is achieved	Generally large, due to aggregating large data-sets

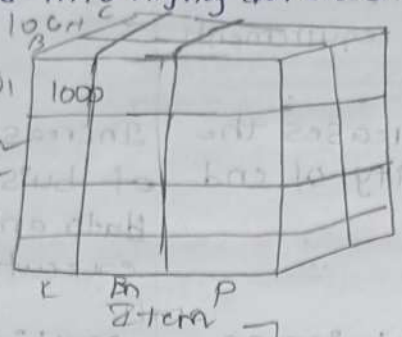
9. Backups Recovery	Regular backups required to acheive Buisness continuity & neat legal & Governance requirements	Lost data can be reloaded from OLTP Database as needed in view of regular backups
10. productivity.	OLTP increases the Productivity of end users	Increases productivity of buisness managers, data analysts & executives
11. dataview:	day-day buisness Transactions	Multidimensional view of enterprise data
12. database design:	Normalised databases for efficieney	Denormalised databases for analysis

* Typical operations of OLAP systems:-

1) OLAP stands for online analytical processing Server
2) It is a software technology that allows users to analyse information from multiple database systems at the same time
3) It is Based on multidimensional data model and allows the user to query on multidimensional data
4) OLAP databases are divided into one (or) more cubes and these cubes are known as "hypercubes"
5) There are 5 Basic Analytical operations that can be performed on OLAP cube:-

ii) Drill down

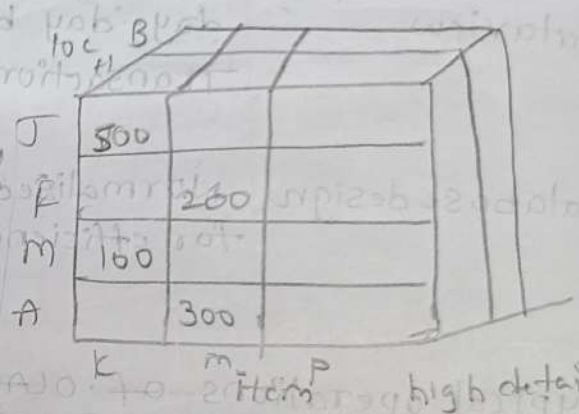
1) Drilldown: In this operation, the less detailed data is converted into highly detailed data



(no. of dimensions Increased)

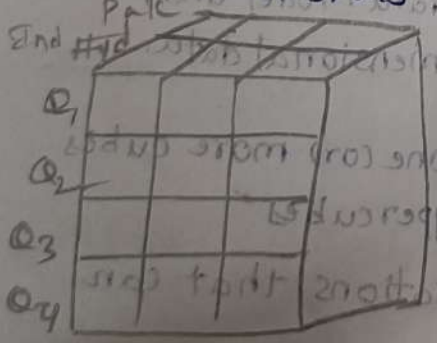
high level
less detailed data

low level
Time

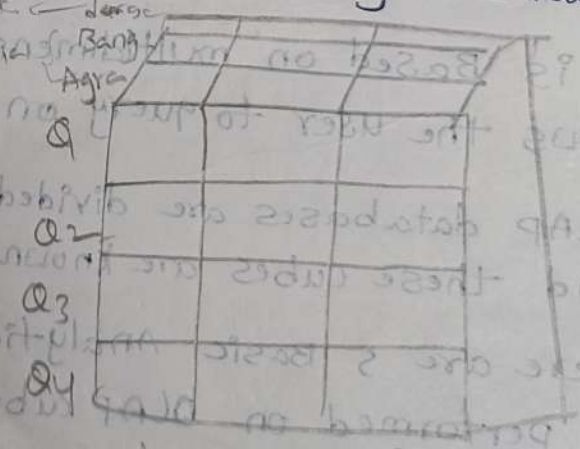


high detailed data

2) Roll up: It is just opposite of the drilldown operation. It performs aggregation on OLAP cube. It can be done by climbing up in the concept hirearchy, Reducing the dimensions.



RollUP



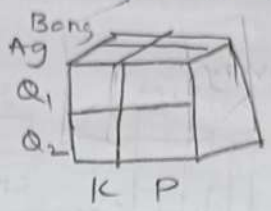
(no of dimens reduced

low level data

high level data

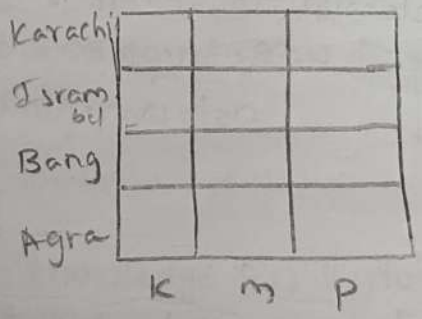
3) Dice:- It selects a subcube from the OLAP cube by selecting 2 (or) more dimensions. In the cube given in the overview section a subcube is selected by selecting following dimensions with criteria

Location = "Agra" (or) "Banglore"
Time = "Q_1" / "Q_2"
Item = "keyboard" (or) "printer"



→ Dice

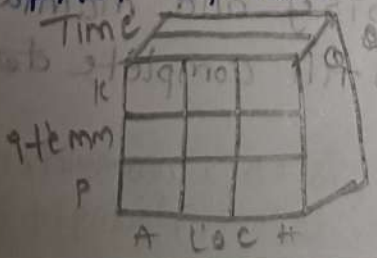
4) Slice:- Selects a single dimension from the OLAP cube which Results a new sub-cube creation, In the cube given in the overview section, slice is performed on the "dimension time = Q_1"



Slice

5) pivot:- It is also known as rotation operation as it rotates the current view to get a new view of the Representation.

*In the subcube obtain after the slice operation, performing pivot operation gives a new cube of it



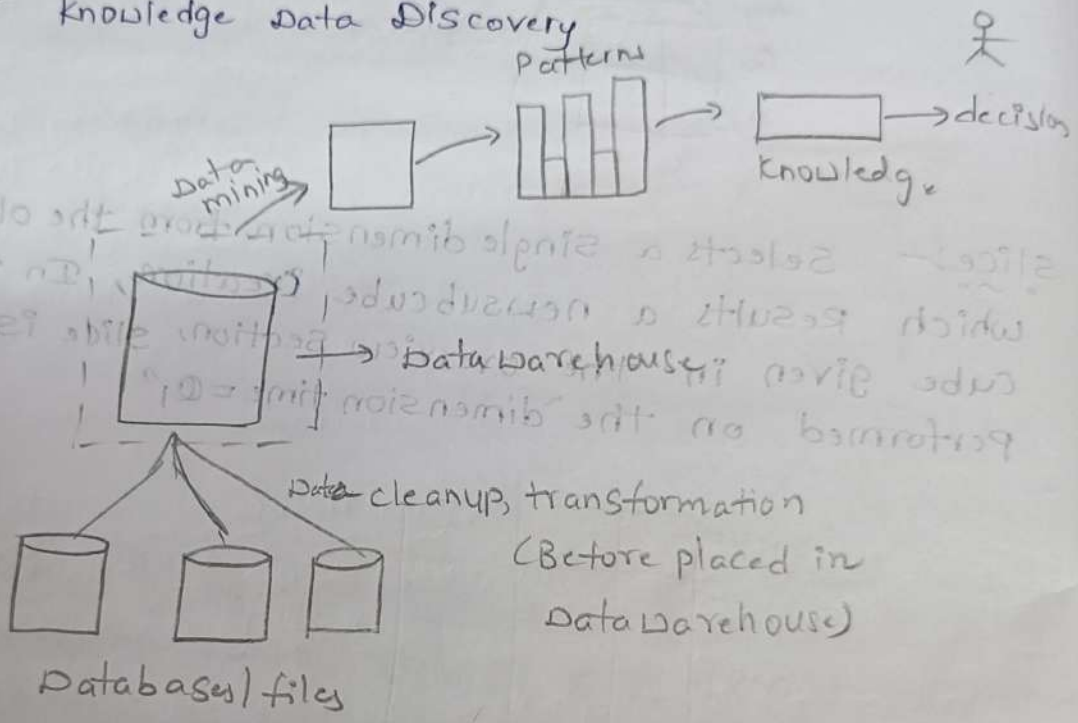
(dimensions rotates)
Or) Interchanges

2. Introduction to Datamining:-

1) Datamining is a process used by organisations to extract specific data from huge databases to solve Buisness problems.

2) The main Aim of datamining turns raw data into useful Information.

KDD:- knowledge Data Discovery



1) Data Selection:-

Select task relevant data from different operational databases (or) flat-files

2) Data cleaning:-

In order to maintain quality data in the Warehouse for the purpose of extracting correct information then by removing noisy and maintain Data consistency and to fill complete ~~details~~ data.

(Remove, noise, inconsistent, incompleteness in the data before placing it in the warehouse)

Ie, noise, inconsistency, Incomplete

3) Data transformation:-
To convert the selected data to required format related to our task.

4) Data Dare house:-
The collection of raw data from different heterogeneous (or) homogeneous databases (or) flat files related to subject

5) Datamining:-
Extraction of Information (or) knowledge from huge volumes of Raw data

6) Patterns:-
The extracted information is Reported to the data warehouse user by using intresting patterns like Graphs, charts etc

7) knowledge (or) Information:-
After processing of raw data what we can get is called 'knowledge'

8) Decision:-
Based on knowledge (or) Information the data warehouse user/ decision maker to take correct and Quality decision regarding about his organisation.

* Data mining:-

* Datamining is the process by which organisation detects patterns in data for insights relevant to their Buisness needs

* It is essential for Both Buisness intelligence, in datascience

* There are many datamining techniques:-

1) Association rules (or) mining
2) Classification
3) Prediction
4) Clustering
5) Regression
6) Artificial neutral network
7) outlayer detection
8) Genetic algorithm

* Applications of Datamining:-

1) The Amount of data collected is said to be almost double of every year.

2) An extracting data (or) seeking knowledge from massive data, datamining techniques are used.

3) In almost all datamining is used in all most all places where a large amount of data is stored, processed :-

1) Scientific Analysis
2) Inclusion detection
3) Buisness transaction
4) Market Basket Analysis
5) Education
6) Reasearch

7) Health care & Insurance
8) Transportation

* Issues (or) problems in datamining:-

1) Datamining is not a easy task as the algorithms used is very complex & it is not available at 1 place

2) It needs to be Integrated from various heterogeneous Datasources.

3) These factors also create some issues.

4) The major issues are:-

i) Mining methodology & user interaction
ii) Performance
iii) Diverse data drivers issue

1) Mining methodology & user interaction issues:-

i) Mining different kinds of knowledge from different databases
ii) Interactive Mining of knowledge at multiple levels of abstraction
iii) Incorporation of background knowledge
iv) datamining query language & adhok
v) Presentation & visualisation of datamining result.
vi) Handling noisy & incomplete data
vii) Pattern Evaluation

2) Performance issues:-

Efficiency & scalability of algorithms can be used Parllel, distributed & incremental mining algorithms

3) Diverse datatypes issue:-
Handling of relational & complex types of data
Mining information from heterogenous databases and
global information system

* Data objects & Attribute types:-

1) object → Real time thing
2) Attribute:- Property/ characterstics of an object
ex:- Attributes of student :- sid, sname, phno, add

*Types of Attributes:-

1) Nominal attribute:- Represents the names of the objects
~~like~~ It also consider as a categorical attribute

ex:-

Attribute	colour values
color	Blue, red, green
Fruit	Banana, mango

2) Binary attribute:- always represents 2 values
ie, affected / uneffected (or) True / false.

* These are further divided into
1) Symetric type :-
2) unsymetric type :-

Symetric:- Equal importance

Attribute	values
Gender	Male/female

Sametype

2) Asymetrical — Result — Pass, fail

3) ordinal attributes — Represents the attributes in a particular order i.e. either character values / numeric values etc.

eg :-

Attribute	Value
Gender	
Grade	O, S, A+, A, B, C, D, F

4) Numerical attributes — Represents the Numerical value i.e. finite no. of values.

1) Ratio scaled :- multiply to get new value
2) Interval scaled (equal interval)

eg :- calender

eg :-

Attribute	Value
Age	12, 20, 40 -

5) Discrete attribute — stores the infinite no. of values.

eg :-

Attribute	Value
Weight	60 40 35
Height	5.4 5.6 5.2

*// Statistical Description of Information

Numerical attributes — These are quantitative which can be measurable represented as integer (or) real values

- There 2 types of numeric attributes :-

1) Interval scaled attribute
2) Ratio scaled attribute

Interval scaled — These attributes will have values which follows some order. We can add / subtract to get next value in the sequence

ex :- Temp, Calender (1 - 31)

4) Ratio Scaled :- These attributes will have the values we can multiply the previous value to get next value in a sequence.

eg :- 10k is twice greater than 5k

5k 10k 20k 40k

5) Discrete attribute :- These attributes have finite (or) countably infinite set of values.

eg :-

Attribute	Values
Zipcode	522403, 522017.
Profession	Teacher, doctor

Continuous attribute :- These attributes have infinite no. of values. These values are float type

ex :- values b/w 2 & 3 are infinite

Attribute	Values
Height	5.4, 5.7, 5.8 ——

* Statistical description of Data :-

The Statistical functions are useful to identify the properties of Data and also used to find noisy & outliers.

1) Measuring central tendency :-

Mean
Median
mode

1) Mean:- Let us consider the marks of 6 students in a particular subject as $x_1, x_2, x_3, x_4, x_5, x_6$ the value will be resp 6, 7, 4, 3, 4, 7 respectively then, the mean

$$\text{mean}(\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{N} = \frac{\text{Sum of values}}{\text{no of values}}$$

Here $N = 6$ $= \frac{31.0}{6} = 5.16$

2) Median:- Values — It is the middle value among all ordered

ex:- To calculate median the values should be in arranged order (ascending order)

Consider the above example

~~6 7 4 3 4 7~~

unsorted order:- 3 4 4 6 7 7 9

odd:- $\frac{N}{2} = \frac{7}{2} = 3.5 = 4$

$N = N/2+1$

even:- 3 4 ④ ⑥ 7 9

$$\text{Median} = \frac{\frac{N}{2} + \left(\frac{N}{2}+1\right)}{2} = \left[\frac{\frac{6}{2} + \left(\frac{6}{2}+1\right)}{2} = \frac{3+(4)}{2} = \frac{7}{2} = 3.5\right]$$

$$= \frac{4+6}{2} = \frac{10}{2} = 5$$

Note:- i) for odd no. of values

Median = $\frac{N}{2}$ observation will be median

ii) For even no. of values

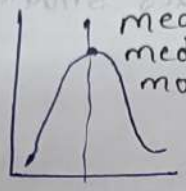
Median = $\frac{\frac{N}{2} + \left(\frac{N}{2}+1\right)}{2}$ observation will be median

3) Mode:- The repeated order in a sequence is called "mode"

1) Unimodal :- No value is Repeated
2) bimodal :- 1 value is Repeated
3) Trimodal :- 2 values are Repeated

ex:- 3 4 4 6 7 7
mode = 4, 7 (Trimodal)

unimode:- In case of symmetric data unimode data is represented as

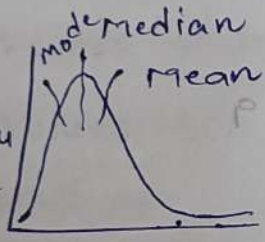


} Coincides at center of curve

← unimode

Bimode:-

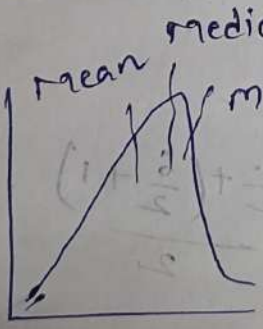
Data distribu at left side



mean > mode

← positively skewed data / asymetric data

*Trimode :-



mean < mode

← Negatively skewed data / D.D at right side

*Data in Real time applications is not in symetric format. It may be either positively skewed data (or) Negatively skewed data

* Measuring dispersion of data:-

1) **Range:-** The average of min & max value

$$= \frac{min + max}{2}$$

2) **Quantile:-** It represents half of the data distribution

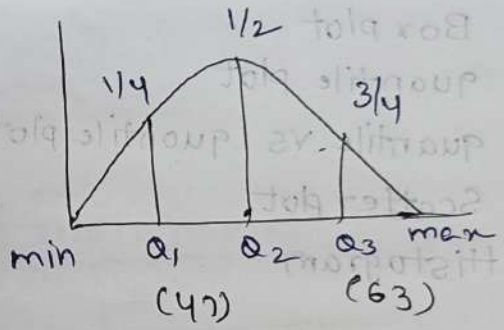
3) **Quartile:-** It represents $\frac{1}{4}$th of the Data distribution

4) **Percentile:-** It represents 100th part of the Data distribution

5) **Inter Quartile range (IQR):-** It is the difference b/w $\frac{3}{4}$th & $\frac{1}{4}$th of the Data Distribution i.e.,

$$IQR = Q_3 - Q_1$$

$$\therefore IQR = 63 - 47$$
$$= 16$$

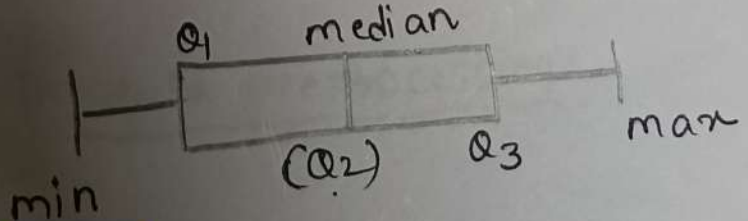


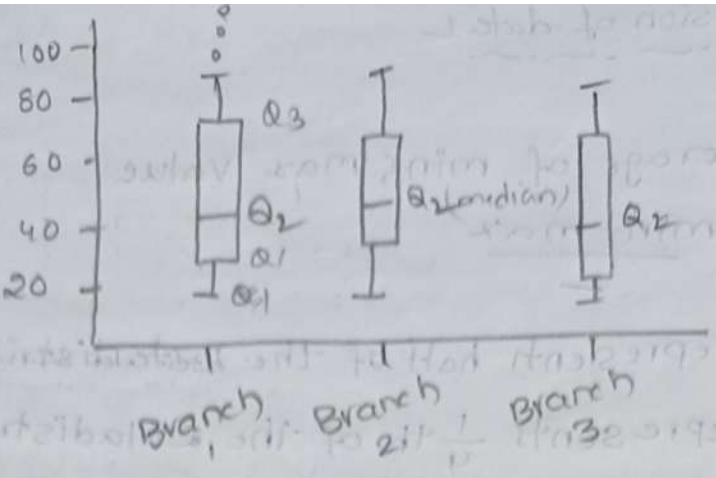
6) **5 number summary:-**

It consists of min, Q_1, median(Q_2), Q_3, max. The outlier can be identified by single out value which falls either below 1.5*IQR (or) the above

$$Q_1 < 1.5 \times IQR > Q_3 \rightarrow \text{outliers}$$

Box-plot:- The 5 number summary is represented in the following Boxplot

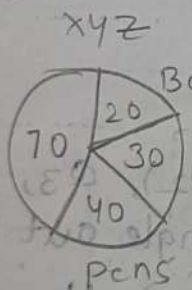




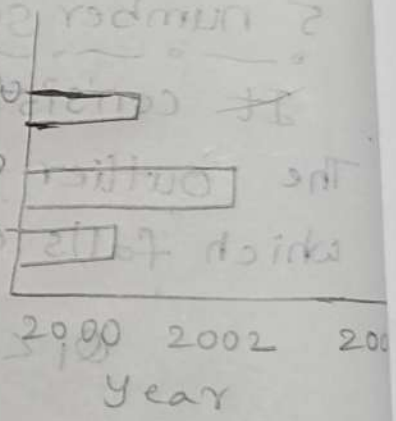
* Display of Information

1) Pie chart
2) Bar chart
3) Box plot
4) quantile plot
5) quantile vs quantile plot
6) Scatter plot.
7) Histogram

Piechart

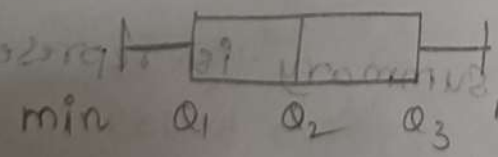


Barchart

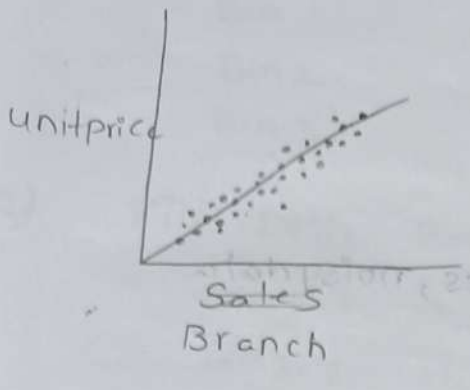


(Year Wise Sales)

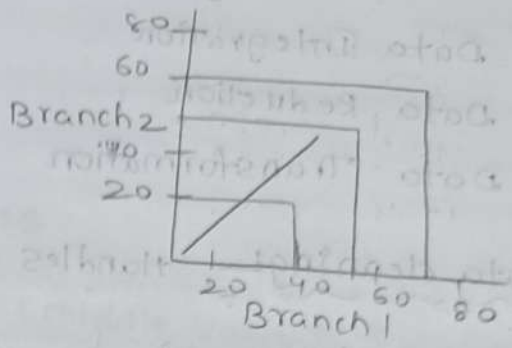
Boxplot



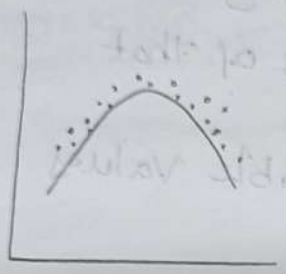
4) quantile plot



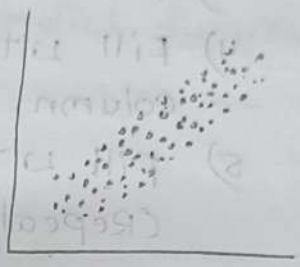
5) quantile vs quantile plot



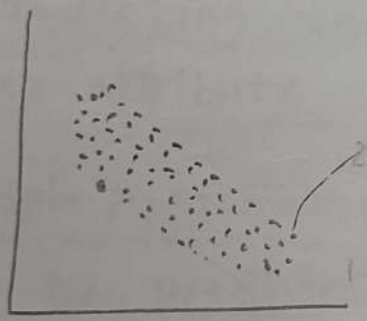
6) scatterplot



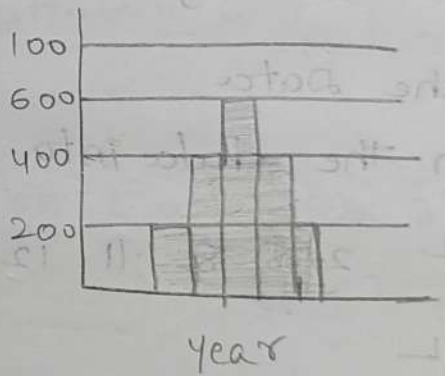
positive scatterplot



negative scatterplot



7) histogram



* Data preprocessing :-

Converting raw data into useful information is preprocessing

* steps in data preprocessing

1) Data cleaning
2) Data Integration
3) Data Reduction
4) Data Transformation

1) Data cleaning :- Handles missing values, noisy data

i) Missing values :-
1) Ignore the tuple
2) use Global constant (unknown, $-\infty$)
3) Fill the values manually
4) Fill with mean values of that column.
5) Fill with most probable values (Repeated values)

ii) handling noisy data (or) errors :-
Data Smoothing (or) Binning

1) Sort the Data
2) Partition the data into Bins/ Buckets
3) Sort :- 2 5 8 11 12 14 16 19 21

Partion :-

Bin 1 :- 2 5 8
Bin 2 :- 11 12 14
Bin 3 :- 16 19 21

} if any error is detected

) If any error is detected/ occured in bins then do the following :-

1) Fill with the mean value

Bin 1

Bin 2

Bin 3

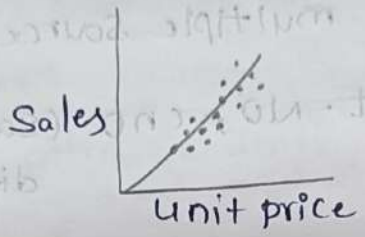
2) Fill With Boundary values

Bin 1 : 2, 2, 8 (middle value is obtained based on which boundary is it nearer to)

Bin 2 : 11, 11, 14

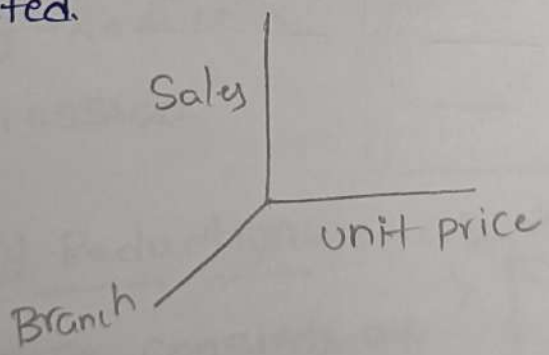
Bin 3 : 16, 21, 21

3) Regression :- We can predict future value based on regression



* Predicting the value of 1 attribute with the help of other attribute

Multiple Regression :- Based on 2 attributes 3rd value can be predicted.



* Outlier Analysis :-

1) clustering :- Grouping of objects based on simila characterstics.

With the help of clustering we can identify the outliers

2) Data Integration:-
It is the process of collecting data from multiple data sources.

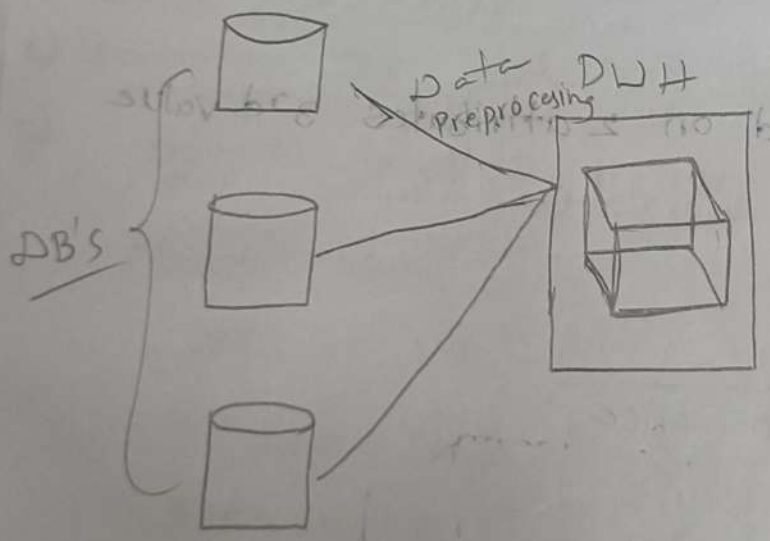
Issues:- 1) Schema Integration & object matching
2) Redundancy & corelation Analysis
3) Data value conflict and Resolution

1) Schema Integration:-
It occurs due to mismatch in attribute names which are collected from multiple sources

ex:- cust-ID, cust-NO, cno (same data but with diff names)

2) Redundancy:- Same type of data is available.
Redundancy ~~der~~ occurs due to derived attribute

ex Name, DOB, age



object matching:-

eg: Type of currency → $, rs, pounds

* When merging is done either $ changes to rs & vice versa. Mismatching of data occurs in case of object matching

Corelation analysis:- Describing Relation b/w normal attribute and the Derived attribute

ex:- Sname → stid

Sid depends on Sname

Dno " " Dname

Dname → Dno

[Relation b/w 2 attributes]

3) Data Value conflict & Resolution:-

i) converting different units into single unit

diff attribute units

circulum: CBSE, ICSE

iii) Data Reduction:-

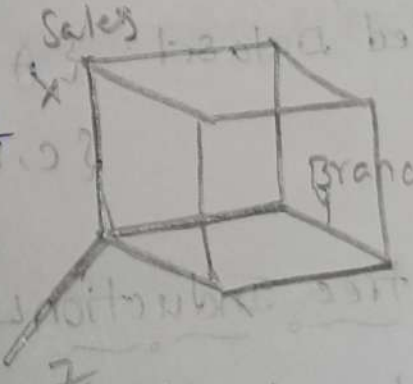
Reducing the size of data before loading into DWH

* Steps in Data Reduction:-

1) Dimensionality Reduction
2) Numerosity Reduction
3) Data compression

1) Dimensionality Reduction:-

i) Data Warehouse consists of Datacubes



ii) Reducing the Irrelavant data in DWH (or)

Removing the unused attributes from datacube

Steps in Dimensionality Reduction

1) Stepwise forward selection
2) Stepwise Backward elimination
3) Decision tree Induction

1) Stepwise forward selection

* original Dataset : {A, B, C, D, E, F}
  - E → aadhar
  - B → name
  - C → Sid, Eid

i) Identify most Relevant attributes & add it to Reduced dataset

ii) Reduced dataset : {}
{B}
{B, C}
{B, C, E}

2) Stepwise Backward elimination

1) opposite to the forward selection
2) Identify most irrelevant attributes from dataset & Remove them

Ex: original Dataset = {A, B, C, D, E}

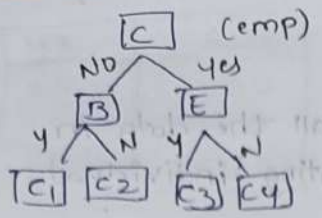
Reduced Dataset : {A, B, C, D}
{C, B, D}

3) Decisiontree Induction

Represents in tree format

ii) most relevant attribute becomes root node & 2nd, 3rd relevant attributes becomes childs to root node

eg:-



} similar to if-else ladder

2) Numerosity Reduction:-

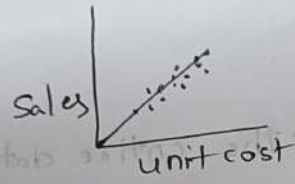
1) Parametric Reduction :- Linear Regression, multiple "

2) Non parametric Reduction → histograms, clustering, sampling, Data cube aggregation

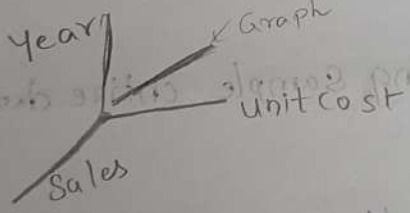
1) Linear Regression:- We can Represent entire data within a single straight line (b→ predictor variable)

[Applied on Databases ie. 2d)

(y=ax+b) Represents all data in single st. line



2) Multiple Regression:- Applied on Datacubes ie 3d



logy = ax+b

(more than 2 attributes)

* Entire data is reduced and represented in the straight line