17/1/23

Data Mining & Warehousing :-

Data Mining :- extracting knowledge from data.

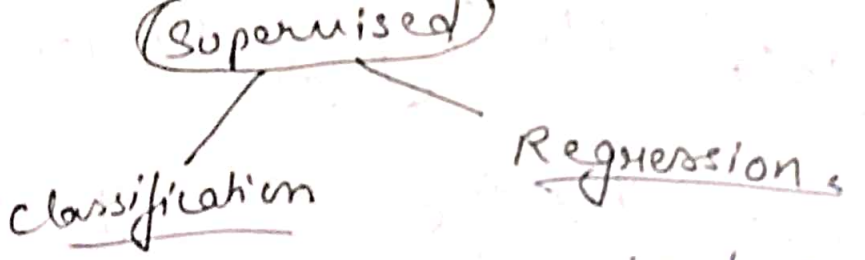
- ↳ supervised ~~Mining~~ learning.
- ↳ unsupervised ~~Mining~~ learning.
- ↳ Market Basket ~~Mining~~ Analysis - based on the products in the shopping cart of customers so that the brands can analyse.

→ Data Mining is the process of analysing data from different perspective and summarising it into useful information - info. that can be used to increase revenue, cut cost or both.

Data Mining Strategies :-

- Supervised learning.
- unsupervised clustering.
- Market Basket Analysis.
- Semi Supervised learning

1) Supervised learning :- (when we are young we use induction to form basic concept definition, we see instances of concepts, representing animals, plants, buildings, etc.) - (Here, labels given to individual instances and choose what we believe to be the defining concept feature (attribute) from our own classification model ← later we use the model, we have developed to help us Identify objects of similar structure. The name of this type of learning is supervised concept of learning or just supervised learning.



(i) Classification :- It is a supervised learning task where output is having defined labels
Types — Binary, Multi class, Multi labelled

Binary :- In binary classification the model predicts either 0 or 1, yes or no but in case of multi class classification the model predicts any one class from more than 2 classes.
In multi label classi. any one instance can have more than one label or classes.

(ii) Regression :- It is supervised learning task where output is having 'continuous value.
It is generally divided into two parts - linear regression and logistic regression.

---

(2) Semi-Supervised :- It is a supervised learning where the training data contains the very few labelled examples and a large no. of unlabelled examples.
In this initially similar data is clustered along with an unsupervised learning algo. and further it helps to label the unlabelled data into labelled data.
Semi supervised lies b/w sup. & unsup. learn^n.

(3) Unsupervised clustering :- unlike supervised learning, unsuper. learning builts model from data without pre defined classes. Data instances are grouped together based on the similarity scheme defined by the clustering system.

Clustering is the process of grouping the data into clusters, so that objects within a cluster have high similarity in compare -on to one another but are very dissimi -lar to objects in other clusters.

18/1/23

Temporal database :- column of any table has attrib -ute of time.

Time Series Databases :- there is always a fixed set

→ Data Mining can be applied on many kinds of data :-

(i) Transactional Databases :- These consist of a file where each record represent a trans -action. A transaction typically includes a unique trans? Id number & a list of item making up the transaction.

(ii) Temporal Databases :- It stores and manages time varying data, eg. financial, medical, government etc; It maintains historical info.; incorporate notion of time in the system.

(iii) Time Series Databases:- a time series is a sequence of data points, measured typically at successive points in time spaced at uniform time intervals.

→ A time series database stores sequence of values or events obtained over repeated measurement of time. Eg - hourly, daily, weekly.

(iv) Spatial Databases:- These databases contain spatial related info. Eg - geographical map databases, satellite image databases.

→ A spatial database that stores objects that change with time is called a spatio-temporal database.

Data Cleaning:-

↳ Missing Values.

↳ Noisy Data

(a) Missing Values:- (i) Ignore the tuple.

when few values are missing then don't ignore. Else ignore the table.

↳ This is usually done when several attributes with missing values are present or the class label is missing.

→

(ii) Fill in the missing values manually.

↳ In general this approach is time consuming & may not be feasible given a large data-set with many missing values.

(iii) Use a global constant to fill in the missing values.

↳ Replace all missing values by the same constant such as label like unknown, ∞, 0, etc. ~~the~~

↳ The problem with this is that the mining program may mistakenly conclude that they form a interesting concept since they all have a value in common.

(iv) Use attribute mean to fill missing values.

↳ Missing values are ~~placed~~ replaced by mean value of column.

(v) use the attribute mean for all samples belonging to the same ~~class~~ as the given tuple.

eg

A_1	A_2	A_3	A_4	Income	class
				100	C_1
				200	C_1
				?	C_1
				300	C_2
				400	C_2

mean value of C_1 only.

(vi) use the most probable value to fill in the missing values.

↳ other methods to fill values. (like regression)

(b) Noisy Data :-

(i) Binning :- ~~[illegible] method~~ ↑ smooth the sorted data values by consulting it's neighbourhood, that is the value around it.

The two common techniques that are used are as follow —

(i).(a) — Data Smoothing By Bin means

(ii).(b) — Smoothing By Bin Boundaries.

(i).(a). 4, 8, 15, 21, 21, 24, 25, 28, 34.

Bin 1: 4, 8, 15 → 9, 9, 9 → equal size bins

Bin 2: 21, 21, 24 → 22, 22, 22 partitioning.

Bin 3: 25, 28, 34 → 29, 29, 29 → data should be so ideal.

↳ replace all values by bin mean.

(ii).(b)

Bin 1: → 4, 4, 15

Bin 2: → 21, 21, 24

Bin 3: → 25, 25, 34.

replace all non boundary numbers with minimum difference boundary number

(ii) Regression:-
Data can be smooth by fitting the data to a a func^{n} such as with regression

(iii) Clustering:-
Outliers may be detected by clustering, where similar values are organised into groups or cluster.

---

1 | 2 | 23

Mean, Median, Mode:-

Q. 13, 18, 13, 14, 13, 16, 14, 21, 13.

$$\text{Mean} = \frac{135}{9} = 15.$$

Mode = 13.

Median:-
ascending - 13, 13, 13, 13, (14), 14, 16, 18, 21.

Median = 14.

Q 10, 12, 13, (16), (17), 18, 19, 21.

$$\text{Mean} = \frac{15.75}{}$$

$$\text{Median} = \left(\frac{n}{2} + 1\right)\text{th} = 4.5\text{th} = \left(\frac{5}{}\right)\text{th element, } 4 \text{ & } 6^{\text{th}} \text{ elm}$$

$$\frac{16+17}{2} = 16.5.$$

Mode = 3 Median - 2 Mean
= 3 (16.5) - 2 (15.75)
= 18.

Q:

age	frequency
1-5	200
5-15	450
15-20	300
20-50	1500
50-80	700
80-110	44

Find the median of age.

Sol^n

$$\text{Median} = L_1 + \left(\frac{N/2 - (\Sigma \text{freq})_l}{\text{freq}_{\text{median}}} \right) \text{width}$$

where L_1 = lower boundary of median interval.

N = No. of values in the entire data set.

$(\Sigma \text{freq})_l$ = sum of frequencies of all the intervals that are lower than the median interval.

$\text{freq}_{\text{median}}$ = frequency of median interval.

width = width of the median interval.

N = Sum of all frequencies.

= 3194.

$N/2$ = 1597 lies in (20-50).

∴ Median interval = (20-50).

$L_1 = 20$.

~~Median~~ $\Rightarrow (\Sigma f)_1 = 950$.

width = 30.

$(f)_{median} = 1500$.

$$\text{Median} = 20 + \left(\frac{1597 - 950}{1500}\right) \times 30.$$

$$= 20 + \frac{647 \times 30}{1500}$$

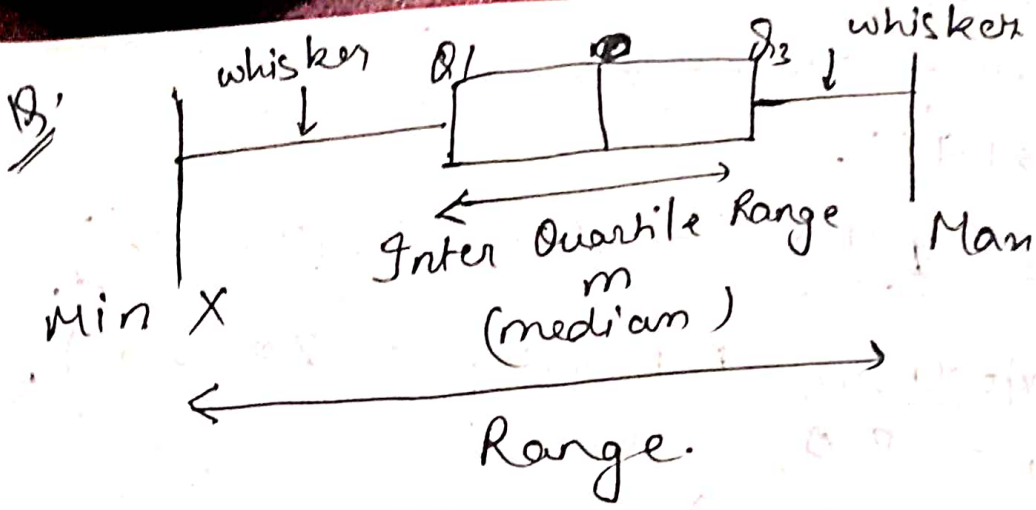
$$= 20 + \frac{19410}{1500}.$$

$$= 20 + 12.94$$

$$\boxed{\text{Median} = 32.94}$$

Box Plot

B)



Data, 76, 79, 76, 74, 75, 71, 85, 82, 82, 79, 81.

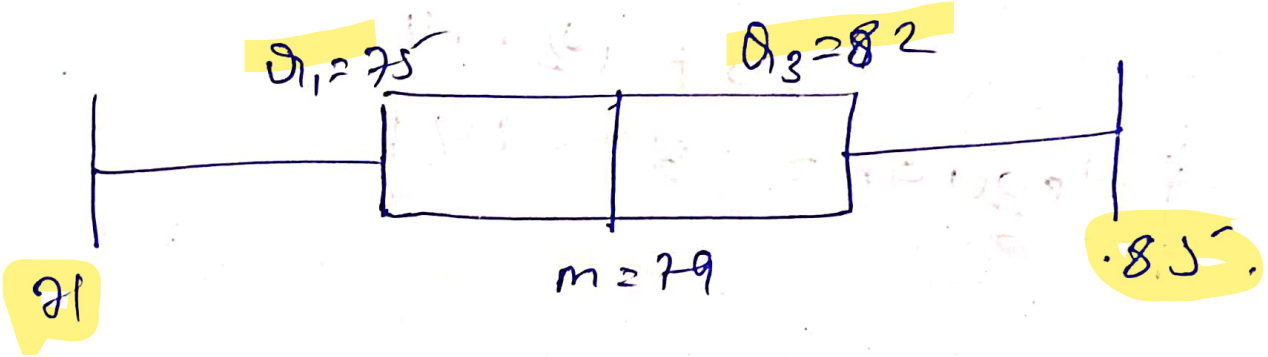
→ arrange in increasing order.

→ median find.

71, 74, 75, 76, 76, 79, 79, 81, 82, 82, 85.

left half median — right half median

Median = 79. ⇒ m = 79



Q. Calculate the standard deviation of 4, 2, 5, 8, 6.

Mean $\bar{x} = \frac{25}{5} = 5$

$\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{N}}$

$= \sqrt{\frac{1 + 9 + 0 + 9 + 1}{5}}$

$= \sqrt{\frac{20}{5}} = \sqrt{4} = 2.$

Q. Correlation Analysis?

Pearson Correlation:

→ Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data.

→ The value of pearson correlation lies b/w [-1, 1] i.e. Pearson $(-1 \leq r_{A,B} \leq +1)$

→ if $(r_{A,B} > 0)$, then A & B are (+vely) correlated. i.e. when the value of A ↑ then the value of B also ↑.

↳ Higher the value, stronger the correlation.

→ if $(r_{A,B} = 0)$, the A & B are independent and there is no correlation b/w them.

→ if $(r_{A,B} < 0)$, then A & B are (-vely) correlated.

$$r_{A,B} = \frac{\sum_{i=1}^{N} (a_i - \bar{A})(b_i - \bar{B})}{\sqrt{\sum_{i=1}^{N} (a_i - \bar{A})^2 \sum_{i=1}^{N} (b_i - \bar{B})^2}}$$

Q.

Tree Height(y)	Trunk Diameter (x)
35	8
49	9
27	7
33	6
60	13
21	7
45	11
51	12

$r = 0.886$.

Soln Mean:-

χ^2 (chi-square) Test

→ For categorical data a correlation relationship b/w two attributes, A & B, can be discovered by χ^2 test. Suppose A has 'c' distinct values i.e. ~~A_1, A_2~~ $(a_1, a_2 \ldots a_c)$ & B has 'r' distinct values i.e. $(b_1, b_2 \ldots b_r)$. Let

→ Let (A_i, B_j) denotes the event that attribute A takes the value A_i and B takes the value B_j. This is represented by a distinct slot in the table. This table is known as Contingency Table. The χ^2 value is

→ computed as

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

where, O_{ij} = observed frequency (i.e. actual count of the joint event).

e_{ij} = expected frequency of (A_i, B_j).

→ e_{ij} is computed as

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{N}$$

where, N = no. of data tuples

→ Degree of freedom = $(r-1) \times (c-1)$.

Q.	male (1)	female (2)	Total.
fiction (1)	250 (90)	200 (360)	450.
Non-fiction (2)	50. (310)	1000 (840)	1050.
Total	300	1200	1500.

e_{11} = $\frac{\text{count (male)} \times \text{count (fiction)}}{N}$

$= \frac{300 \times 450}{1500} = 90.$

e_{12} = $\frac{\text{count (female)} \times \text{count (fiction)}}{1500.}$

$= \frac{1200 \times 200}{1500} = 360.$

$\chi^2 = \frac{(250-90)^2}{90.} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} +$

$\frac{(1000-840)^2}{840.} = 507.$

significance level — 5%.] = 3.84

(table will be given):

DOF = 1.

if $x^2 < 3.84$ then these values are not correlated.

Data Transformation :—

→ Smoothing
→ Aggregation
Generalization

Attribute Construction
(New attributes are constructed from given set of attributes)

1) First step is smoothing which works to remove noise from the data e.g., such techniques include binning, regression, clustering.

2) Aggregation :- Where summary or aggregation operations are applied to the data. Eg. Daily sales may be aggregated to compute monthly or annual sales.

3) Generalisation :— Where low level primitive data are replaced by higher level concept through the use of concept hierarchies.
Eg. City can be generalized to a higher level concept like State or country

4) Normalization - where the attribute data are scaled so as to fall within specified range
Eg. -1 to 1 or 0 to 1.

5) Attribute Construction:- where new attributes are constructed from a given set of attributes.

Min-Max Normalization

$$V' = \frac{V - min_A}{max_A - min_A}\left(new\_max_A - new\_min_A\right) + new\_min_A$$

Where max_A and min_A are maximum and minimum values of an attribute A
Min-Max Algo map a value V of attribute A to a value V' in the range new_min_A to new_max_A

Suppose the minimum and maximum value of attribute income are 12000 and 98000 map the income in range 0.0 to 1.0 using min_max Norm. find the value of income 73600.

$$v' = \left(\frac{73600 - 12000}{98000 - 12000}\right)(1 - 0) + 0$$

$$= \frac{61600}{76000}$$

$$= 0.716$$

Z-score Normalization

The values of an attribute A are normalized based on the mean and standard deviation of A

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

The mean and standard deviation of the values for the attribute income are ~~surpeesly~~ 54000 and 16000. ~~tobe~~ with Z-score normalization transform the value of income to 73600.

$$v' = \frac{73600 - 54000}{16000}$$

$$= \frac{19600}{16000} = 0.6$$

Data Warehouses — OLAP
Online Analytical
by Managerial level (Decision Makers)

Databases — OLTP
Online Transactions
(Lower Level Staffs)

Data Warehouse :-
~~A Datawase~~
According to William H Inmm. A Data Warehouse is a subject Oriented, Integrated, Time variant and Non - Volatile Collection of Data in Support of Management decision making process.

SINT

1) Subject Oriented — A Data Warehouse is organised around major subject areas like customer, supplier, etc.

2) Integrated — A Data Warehouse is usually constructed by integrating multiple heterogeneous sources.

3) Time Variant → Data are stored to provide information from a historical perspective. Every key structure in the datawarehouse contains explicitly or implicitly an element of time.

It usually requires only two operation in data accessing. Initial loading of data and accessing of data.

Location – Belrar

item (type)

time (Quarter)	Home Entertainment	Computer	Phone
Q_1	605	365	12
Q_2	625	205	13
Q_3	425	195	19
Q_4	365	325	29

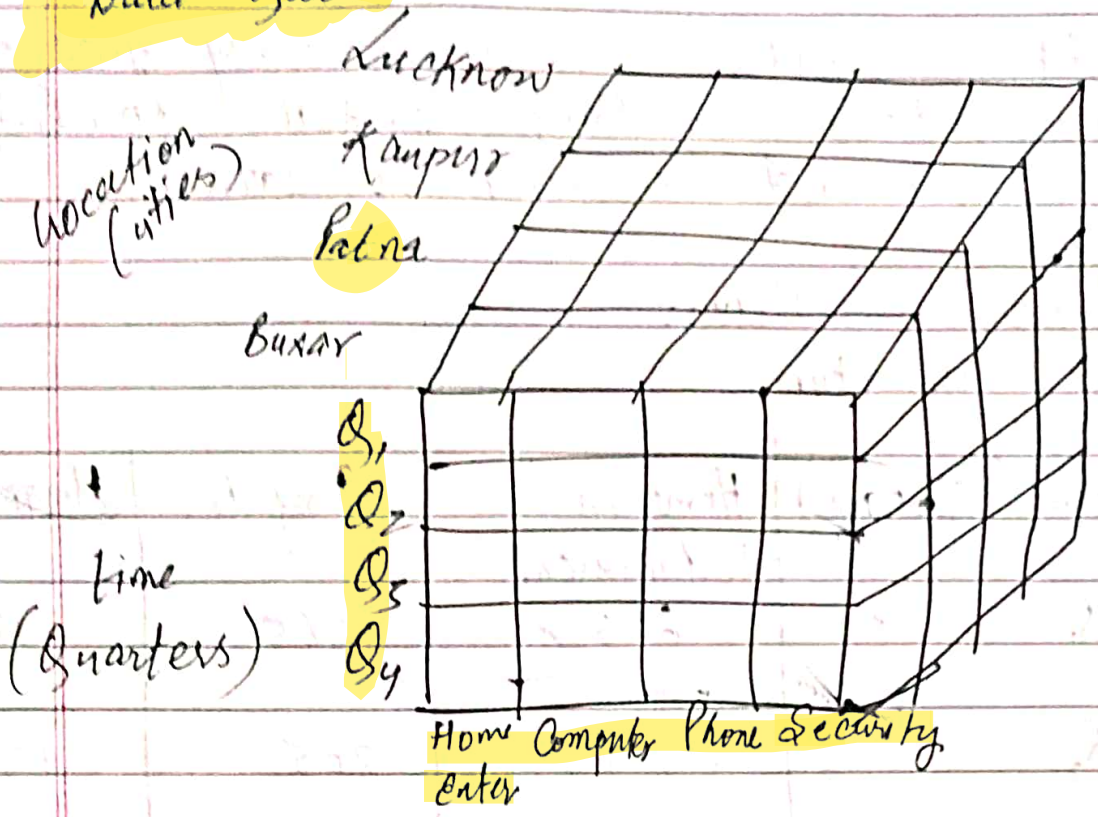
Location – "Patna"

item (type)

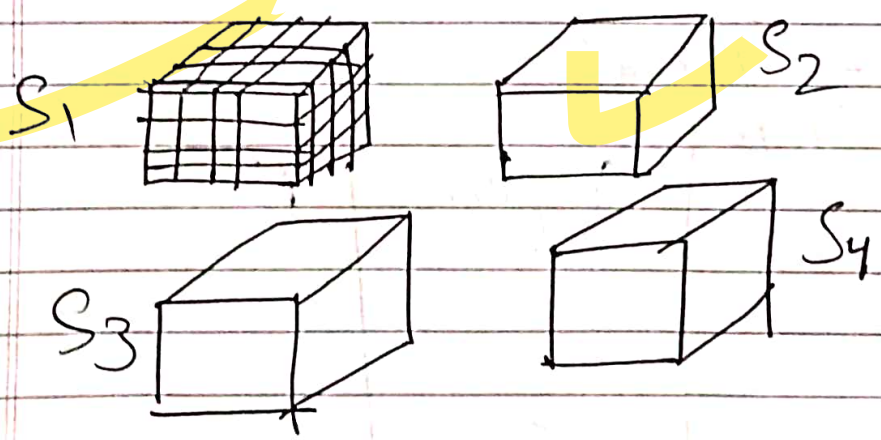
time	Home Entertainment	Computer	Phone
Q_1	395	1600	
Q_2	700	300	
Q_3	900	200	
Q_4	400	400	

Location – "Kanpur"

Data Cube



If we want to increase dimensions for eg. by adding the supplier attribute make one cube for each supplier.

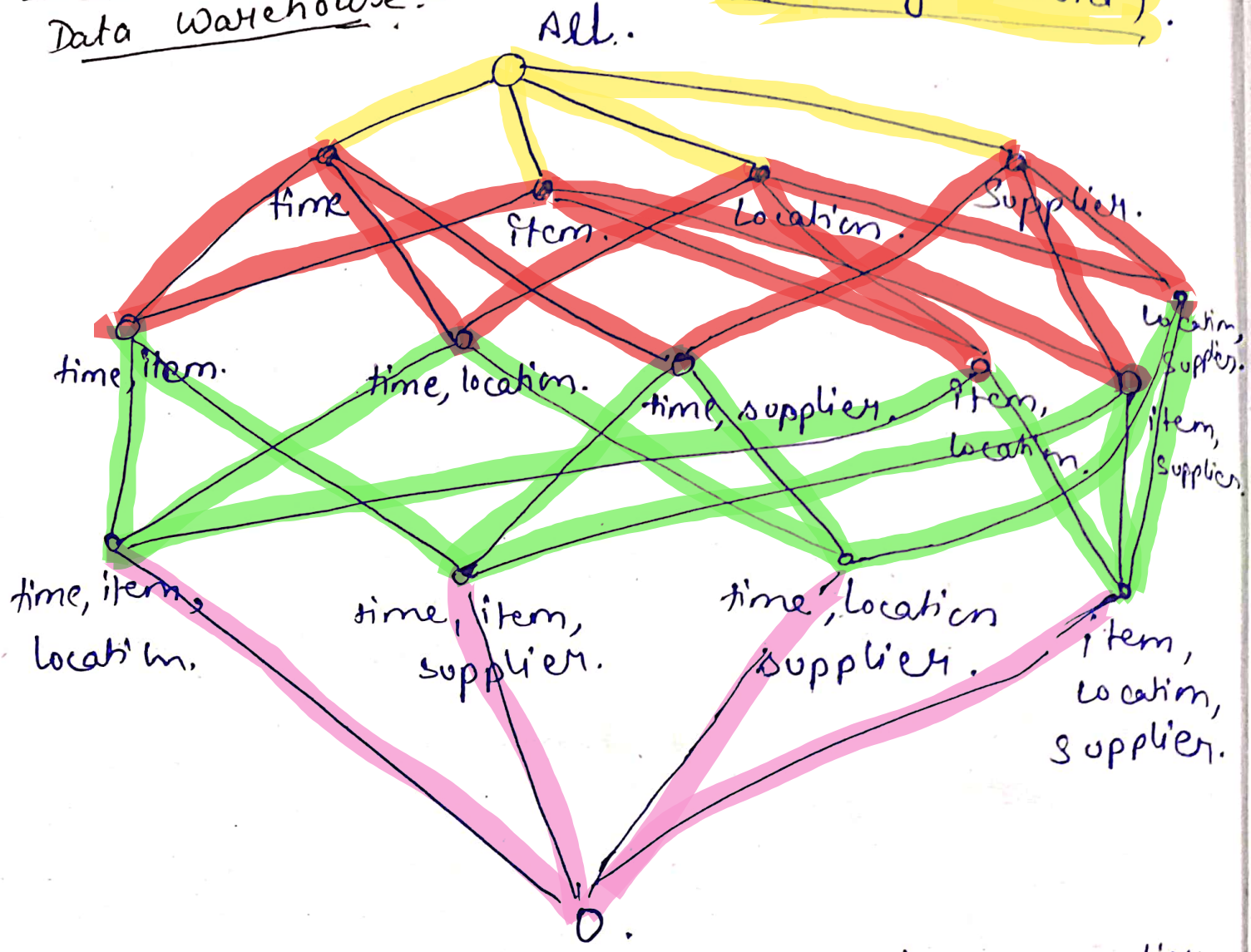


If we continue in this way the n dimensional data can be viewed as n-1 dimensional cubes

8/2/23 :-

Data Warehouse :-

(lattice of cuboid).



→ Given a set of dimensions, we can generate a cuboid, for each of the possible subsets of given dimensions. The result would form a lattice of cuboids each showing a data at different level of summarisation. The figure shows the lattice of cuboids forming a data cube for the dimension — time, item, location & supplier.

→ The cuboid that holds the lowest level of summarisation is called the base cuboid.

→ The zero dim^n cuboid which holds the highest level of summarisation is called the apex cuboid.

Schemas for Multidimensional Databases :-

→ Star
→ Snowflake
→ Fact - Constellation.

3 schemas for Multi-D. Databases.

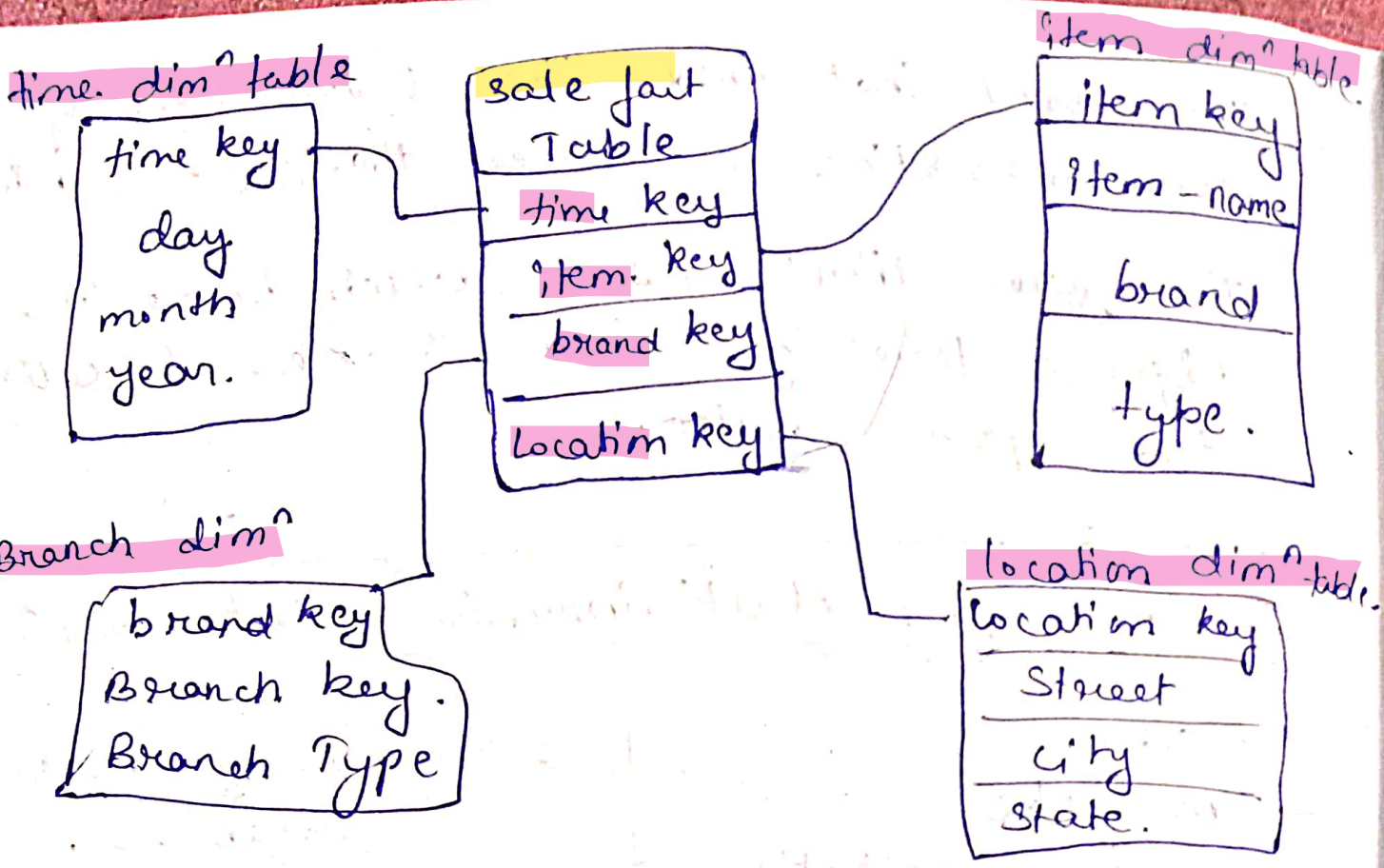
→ The E-R model is commonly used in the design of relational databases, where a database schema consist of set of entities and the relationship blw them.

→ The most common data models for data warehouse are —
- Star
- Snowflake
- Fact - Constellation.

Star :-

Sale Fact Tube
time key
item key
brand key
location key

P.T.O. →



→ The most common model is star schema in which data warehouse contain —

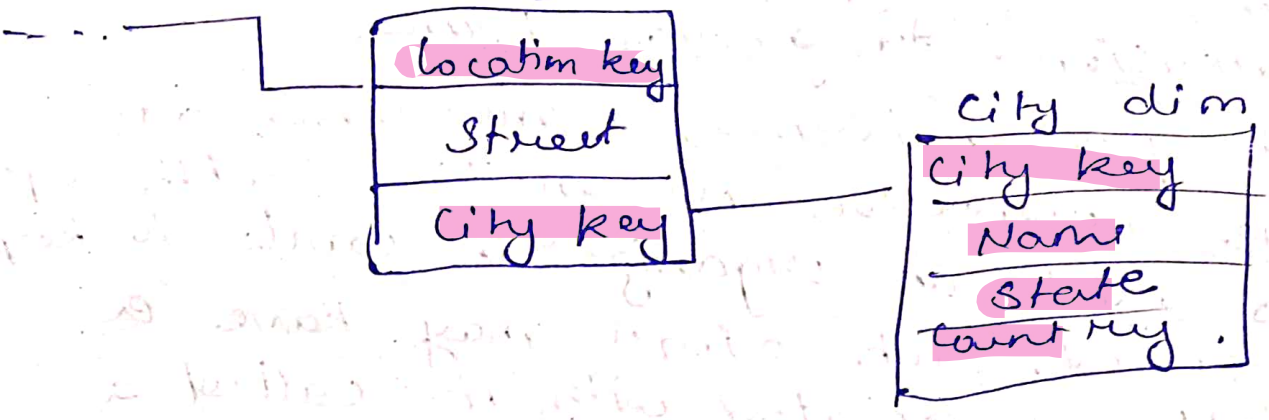
(i) a large central fact table containing a bulk of data with usually no redundancy.

(ii) A set of smaller attendant tables (dim^ tables), one for each dimension.

→ The schema graph resembles star.

Snowflake :-

→ Same diagram as star but —.



→ The snowflake schema is a variant of star schema where some dimⁿ tables are normalized, thereby further splitting of data into further tables.

Fact - Constellation :-

→ It is collection of star schema by help. of dimⁿ tables.

→ Sophisticated applications may require multiple fact tables to share dimension table. This kind of schema can be viewed as a collection of stars and hence is called as galaxy schema or a fact constellation schema.

P. T. O. →

14/2/23.

Data Cube :- A data cube allows data to be modelled & viewed in multiple dimension. ~~In general terms~~

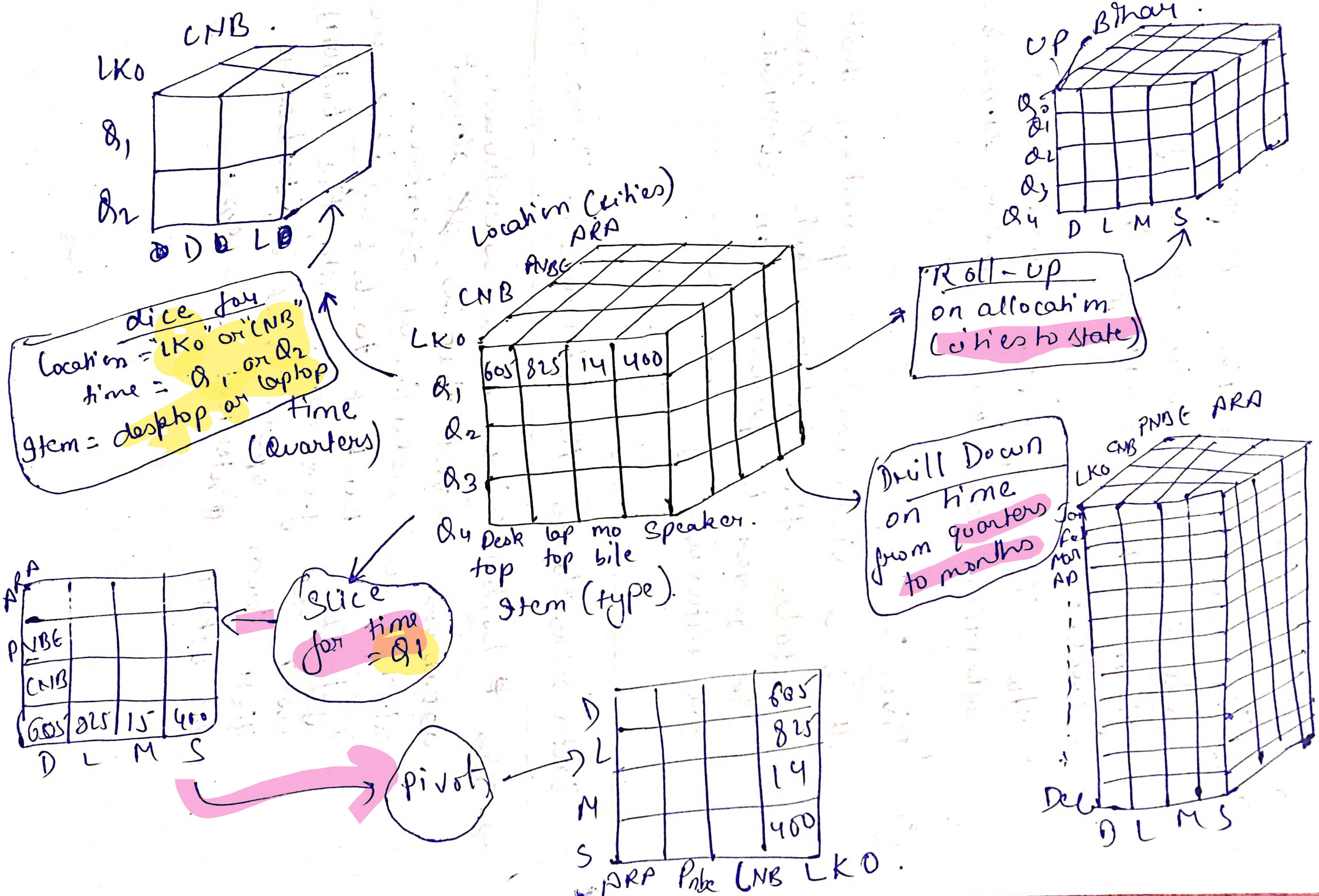
→ In general terms, dimensions are the perspective or entities with respect to which an organization wants to keep record. Each dimn may have a table associated with it called a dimension table.

→ Fact Table :- Contains the name of the fact or keys of each related dimension table.

OLAP operations :-

① Roll up - The roll up operation performs the aggregation on a data cube either by climbing up or dimensionality reduction.

② Drill Down - The drill down is reverse of roll up. It navigates from less detailed data to more detailed data. It can be realized either by stepping down a concept hierarchy for a dimension or introducing additional dimensions.


CNB
LKO
Q1
Q2
D D L D
dice for
Location = "LKO" or "CNB"
time = Q1 or Q2
Item = desktop or laptop
Location (cities)
ARA
PNBE
CNB
LKO
605
825
14
400
Q1
Q2
Q3
Q4
Desk top
lap top
mobile
Speaker.
Item (type).
time (Quarters)
UP
Bihar.
Q1
Q2
Q3
Q4
D L M S
Roll-up
on allocation
(cities to state)
Drill Down
on time
from quarters
to months
LKO
CNB
PNBE
ARA
Jan
Feb
Mar
Apr
Dec
D L M S
Slice
for time = Q1
ARA
PNBE
CNB
605
825
15
400
D L M S
pivot
D
L
M
S
605
825
14
400
ARA
Pnbe
CNB
LKO.

(3) Slice & Dice :-

→ The slice operation performs a selection on 1 dimension of the given cube, resulting in a sub cube.

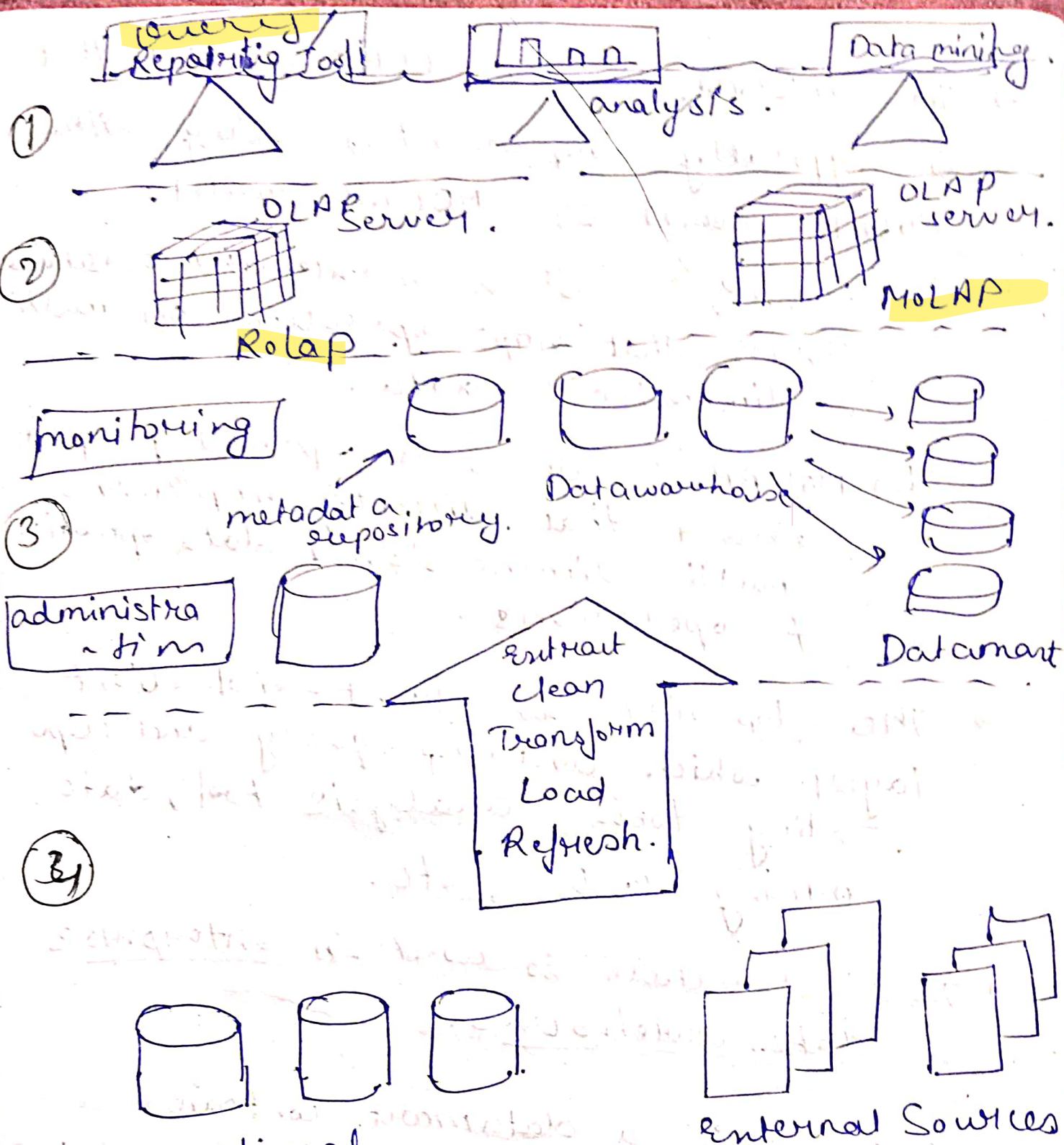
→ The dice operation performs a selection on two or more dimension.

(4) Pivot :

It is a visualisation operation that rotates the data axis in view in order to provide an alternate presentation of the data.

Datawarehouse Architecture :- (Three Tier)

→ The Bottom Tier is a warehouse database server that is almost always a relational database system. Backend tools & utilities are used to feed data into the bottom tier from operational databases or other external sources.



- (1) — Top Tier front end tools.
- (2) Middle Tier OLAP servers.
- (3) Bottom Tier Datawarehouse Server.
- (4) DataMART

→ The middle tier is OLAP server that is typically implemented using either ROLAP model or MOLAP model.

↳ ROLAP :- It is an extended relational DBMS that maps operations on multi-dimensional data.

↳ MOLAP :- It is a special purpose server that directly implements multi-dimensional data operations.

→ The top tier is front-end client layer which contains query and reporting tools, analysis tool, data mining tools, etc.

→ This structure is used in enterprise data warehouse.

→ Datamart :- a datamart contains a subset of cooperate by data, that is of value to a specific group of users. It's scope is confined to specific selected subject.

→ Depending on the source of data, datamarts can be categorised as - Independent or dependent.

→ Independent - datamarts are sourced from data captured from one or more operational system or external info providers.

→ Dependent - datamarts are sourced directly from enterprise data ware-house.

•) Independent data Marts are source from the data captured from one or more operational System or external information providers.

•) Dependent Data Mart are source directly from Enterprised datawore house.

Date - 15/02/2023 :-

Curse of dimensionality :-

Datamare house

*) ~~OLTP~~ ~~Source~~ OLAP System primary objective ?

OLAP — Analysis.

→ It is used to produce the information required by top level in less time.

OLTP system

single Group By operation- producing single 1-D Cuboid.

multiple Group By oper. — multi - Cuboid.

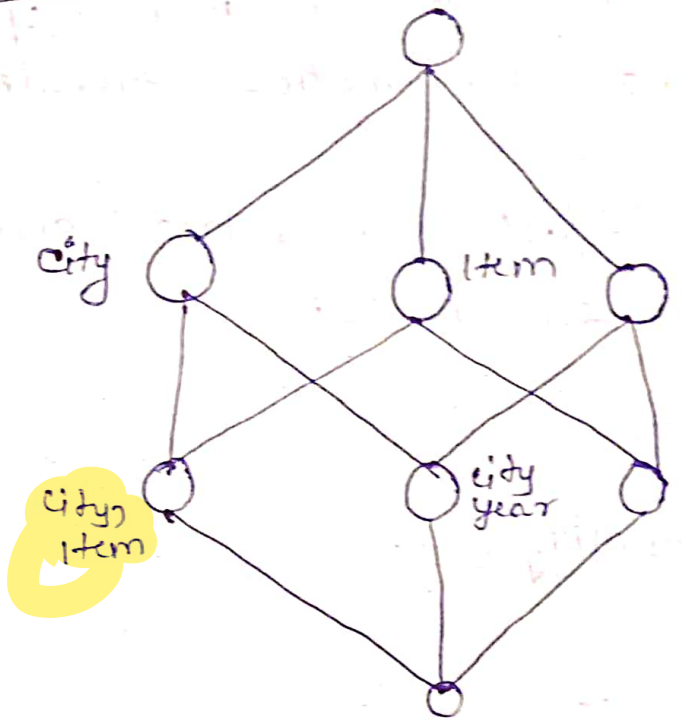
multi - Cuboid — join — latent of Cuboid.

~~#~~ . Curse of dimensionality →

no. of dim → ↑, no. of Cuboid ↑

It is very difficult to store those.

Sales Data :-



0-D Apex cuboid

1-D cuboid

2-D cuboid

3-D base cuboid.

lattice of cuboid

*) The total no. of cuboid or group-by that can be computed for this data cube is 2^3. (2^n)

*) These group by forms a lattice of cuboid for the data cube.

*) The base cuboid contains all 3-dimension (city, item & year). It can return the total set for any combination of these 3-dimensions.

*) The Apex cuboid / 0-D cuboid refers to the case where group-by is empty. It contains the total sum of all sales.

c.) This pre-computation requires large amount of storage and it may explode when many of the dimensions have associated concept hierarchy.

⟹ For 'n' dimensional data cube, the total no. of cuboids that can be generated if the dimensions have hierarchies associated with it.

$$\text{Total no. of cuboid} = \prod_{i=1}^{n} (L_i + 1)$$

where L_i is the no. of level associated with dimension 'i'.

Ques.) If the data cube has 10 dimension and each dimensions have 4 levels associated with it, then find the total no. of cuboids generated.

Solution.

$$\prod_{i=1}^{n} L_i + \prod_{i=1}^{n} 1$$

$$10 \times 4 + 1 \times 10$$

$$= 40 + 55 = 95.$$

$$= 40 + 10$$

$$= 50.$$

$$\text{no. of Cuboid} = \prod_{i=1}^{10} (4+1)$$

$$= \prod_{i=1}^{10} 5$$

$$= 5 \times 5 \text{ ---- } 5$$

$$= 5^{10}.$$

Materialization of cuboids :-

→ No materialization
→ Full ,,
→ Partial ,,

* There are 3 choices for data cube materialization given a base cuboid.

1.) No Materialization

→ Don't precompute any of the non-base cuboid but this may lead to extremely slow operations.

2) Full Materialization

→ Precompute all of the cuboids.
→ It typically requires huge amount of memory space in order to store all of the precomputed cuboids.

3) Partial Materialization

→ Selectively compute a proper subset of the whole set of possible cuboids.
→ It consider 3 factors :-
  ⊙ Identify the subset of cuboids or sub-cubes to materialise.
  ⊙ Use the materialised cuboids during query processing.
  ⊙ Efficiently update the materialized cuboid during load & refresh.

Association Rule :- Frequent Pattern.

↳ Analyse market basket of customer.

↳ Association rule is usually written in form of
$A \rightarrow B$

↳ support → eg- 2%. (Mobile & tempered glass purchased together)

↳ confidence → eg- 60%. (if 10 Mobile = 6 tempered.).