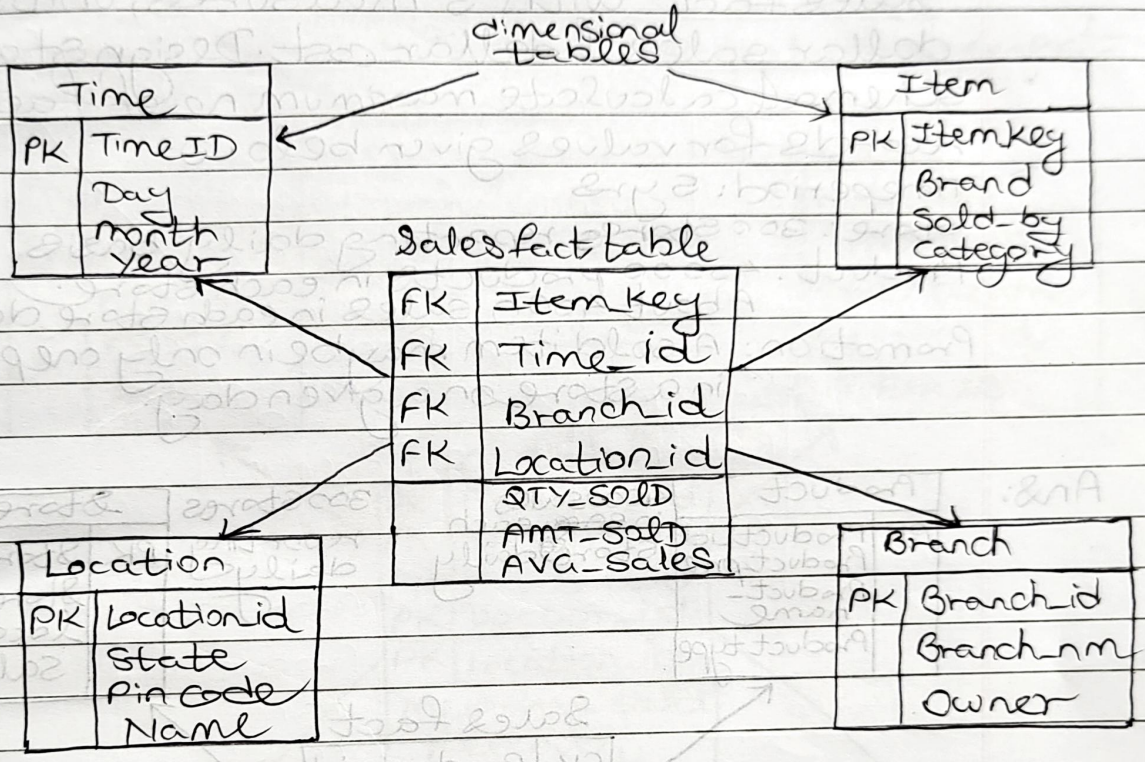
18.7.2024

* Data warehousing & Mining *

- ## Fact & dimensional table:

1) Star schema:



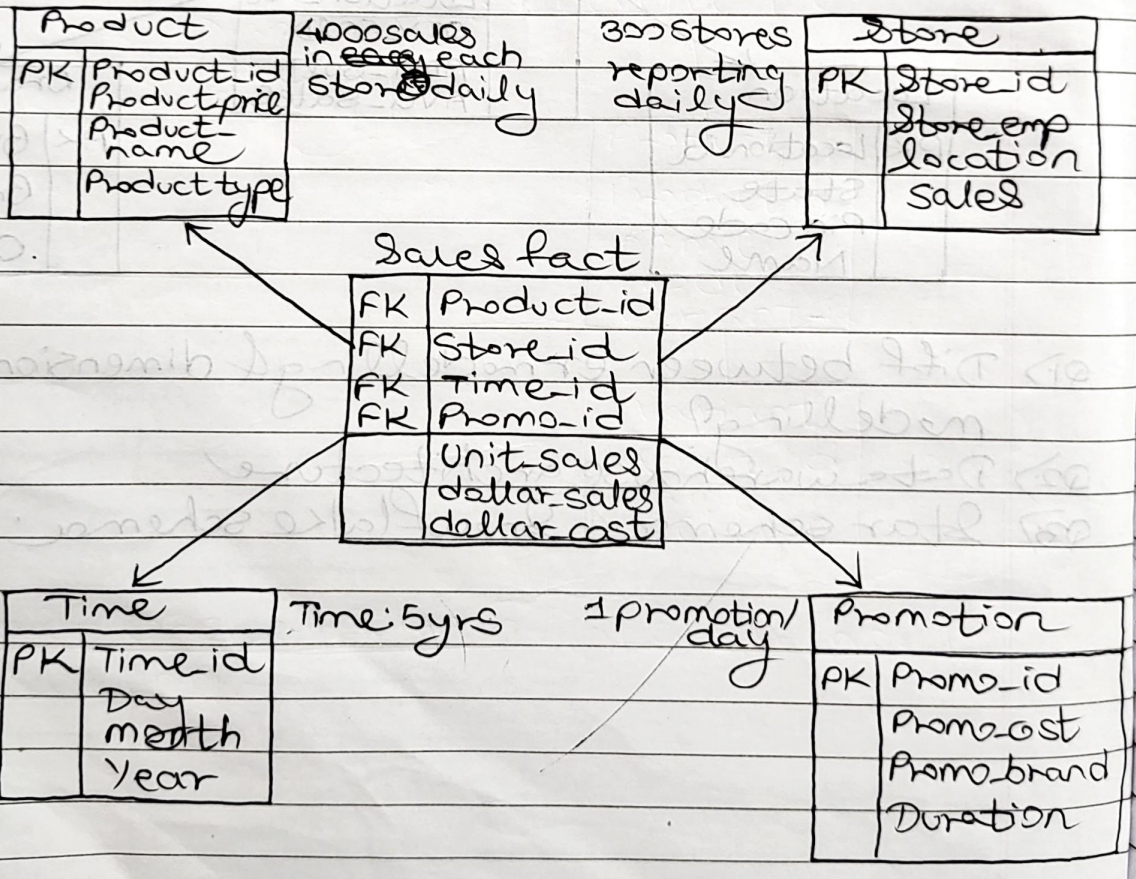
Q1) Diff between er modelling & dimensional modelling?

Q2) Data warehouse architecture

Q3) Star schema & Snowflake schema

Q) for a supermarket chain, consider the following dimensions,
i) product, ii) store, iii) time, iv) promotion.
The schema consists a central fact table "Sales fact" with 3 measures, unit sales, dollar sales & dollar cost. Design star schema & calculate maximum no. of fact table records for values given below:
Time period: 5 yrs
Store: 300 stores reporting daily basis
Product: 40000 products in each store.
About 4000 sales in each store daily.
Promotion: A sold item may be in only one promotion in a store on a given day.

Ans:

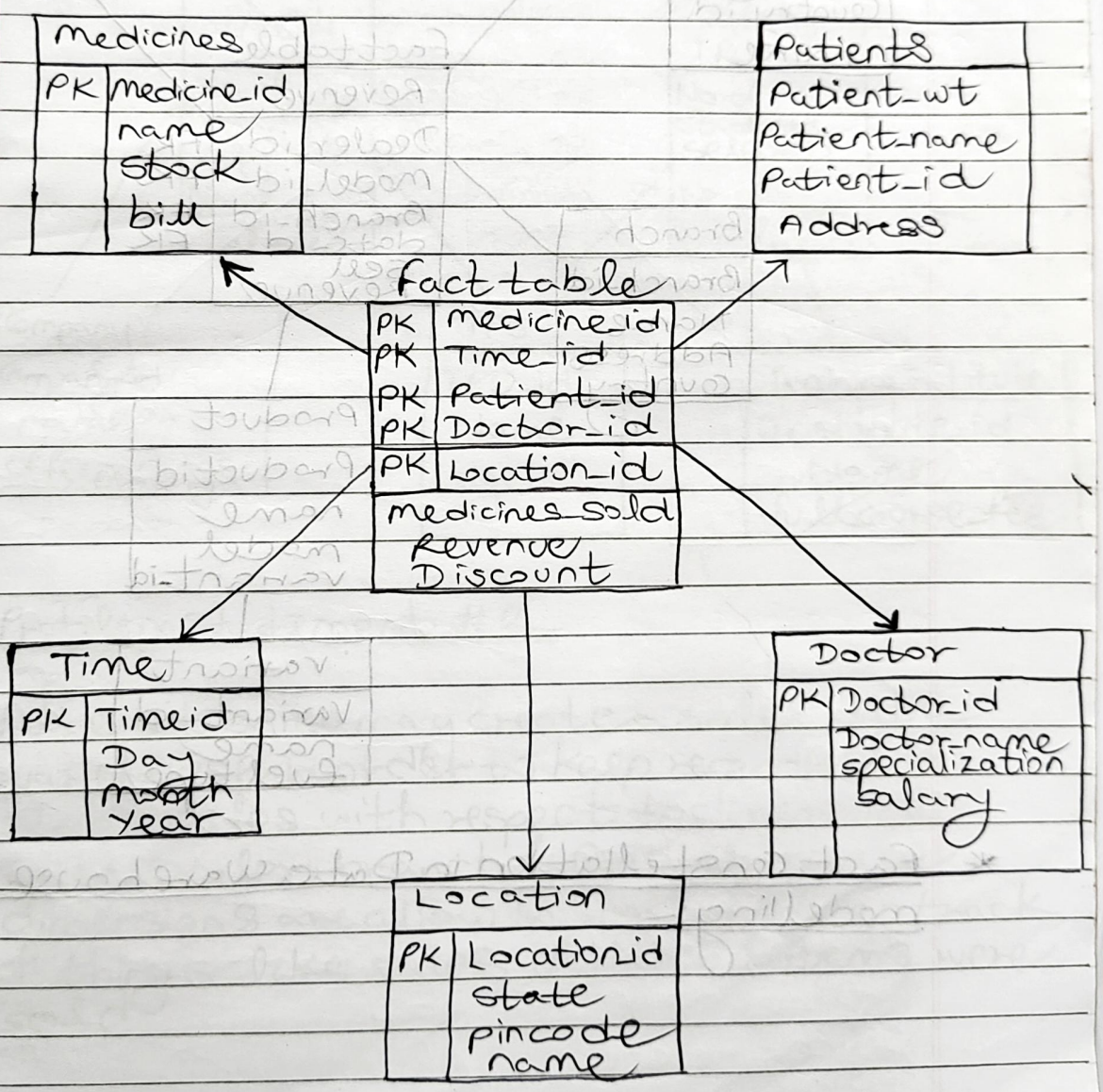


Page No.
Date

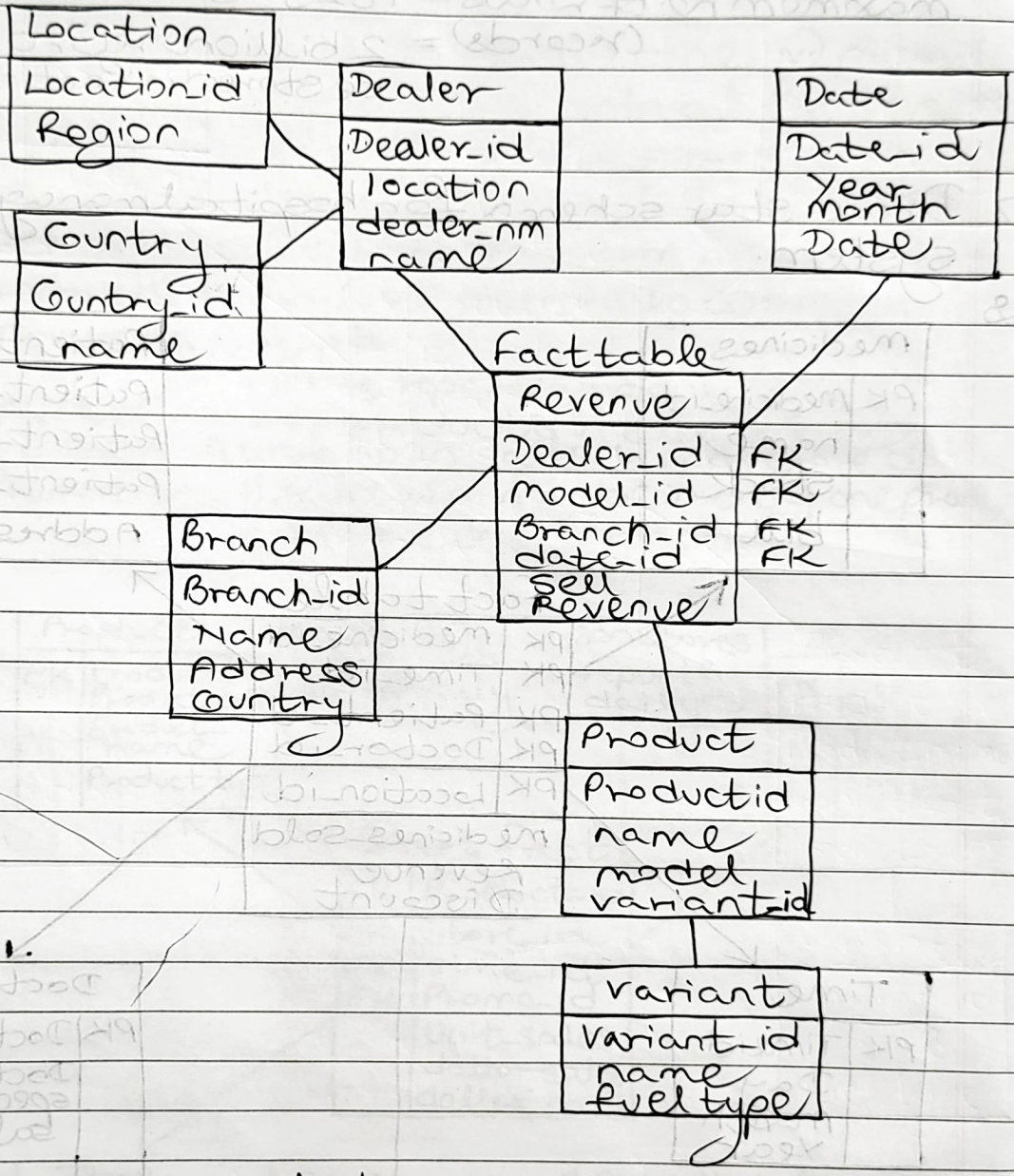
Timeperiod = 5yrs × 365 days = 1825
maximum no. of rows = 1825 × 300 × 4000 × 1
(records) = 2 billion record will be stored in fact table.

Q27 Draw star schema for hospital management system.

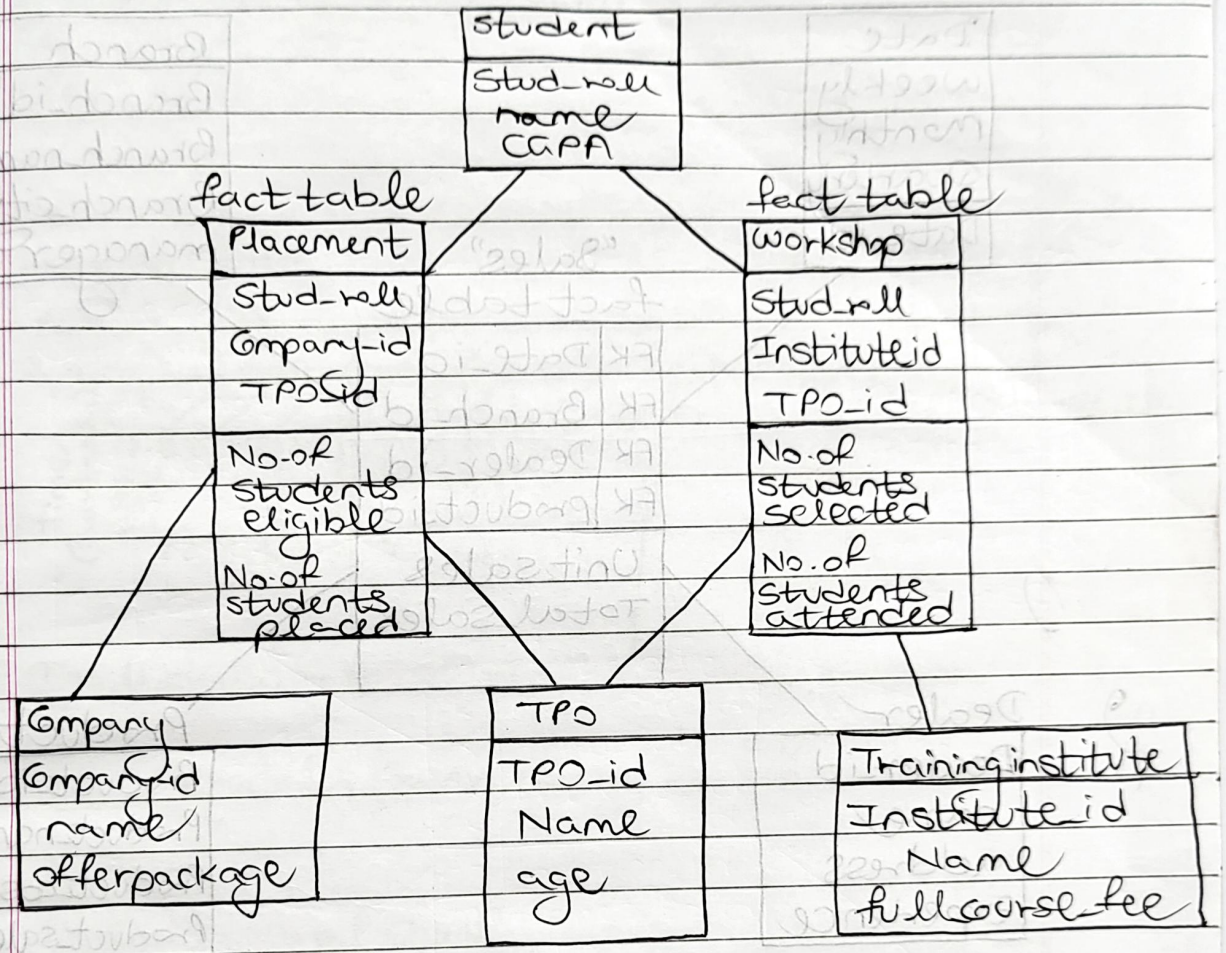
Ans



2) Snowflake schema:-

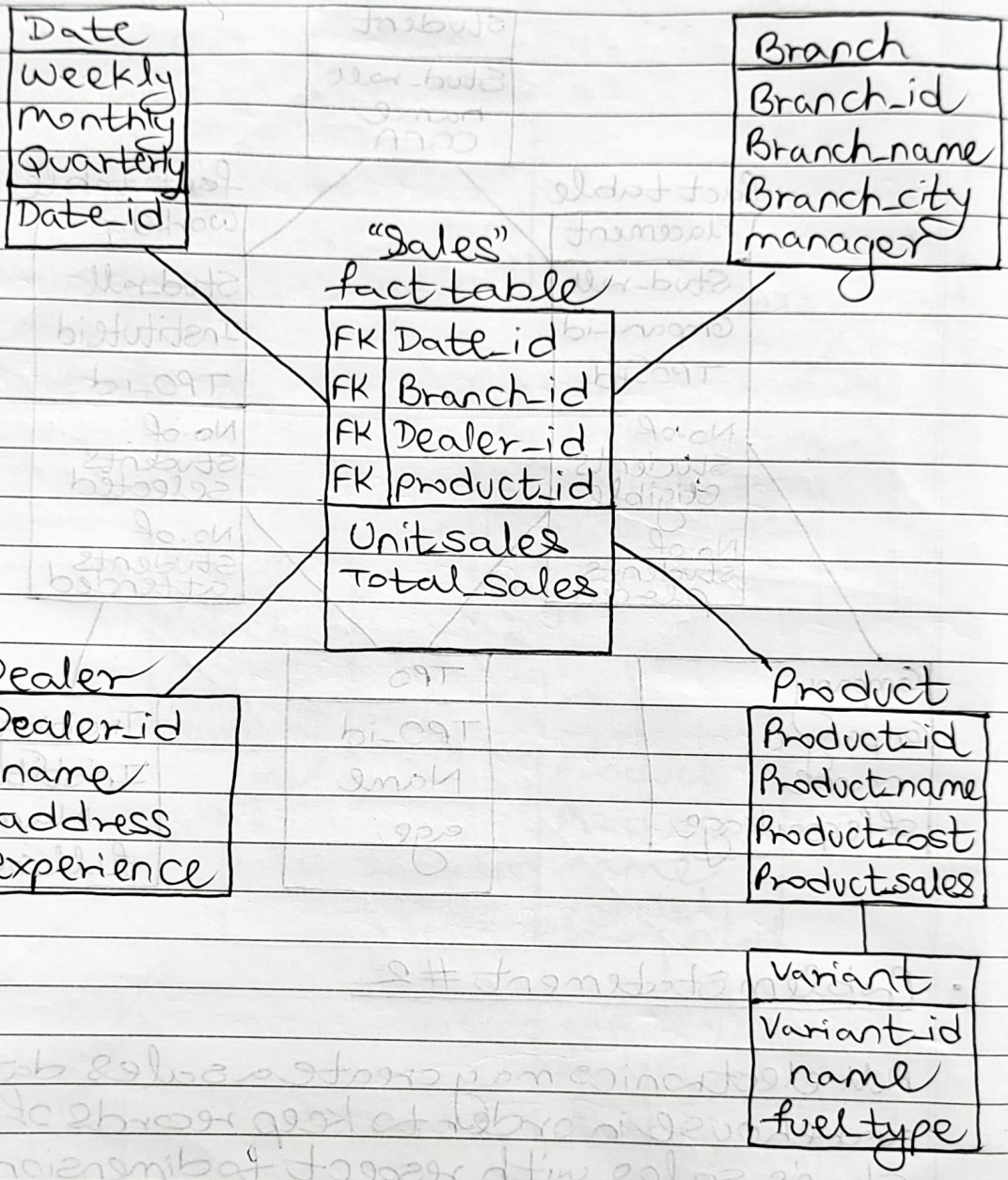


* Fact Constellation in Data Warehouse modelling:-



- Problem Statement #2

All electronics may create a sales data warehouse in order to keep records of the store's sales with respect to dimensions Date, Dealer, branch & Product. These dimensions are allow to store to keep track of things like sales of items & items were sold

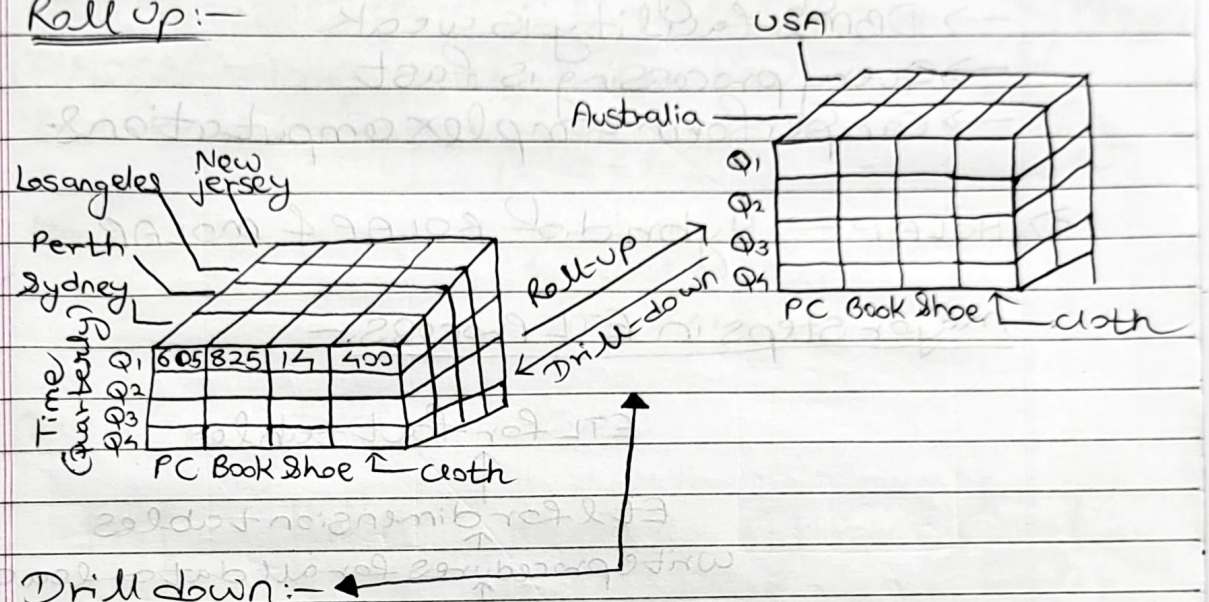


- **OLAP operations:-**

Roll up (drill-up): summarize data
Drill-~~up~~ down (roll down)
Slice: select
Dice: project
Pivot (rotate): reorient the cube

Q) Difference between OLAP & OLTP?

I. Roll up:-



II. Drill down:-

* OLAP models/Servers:

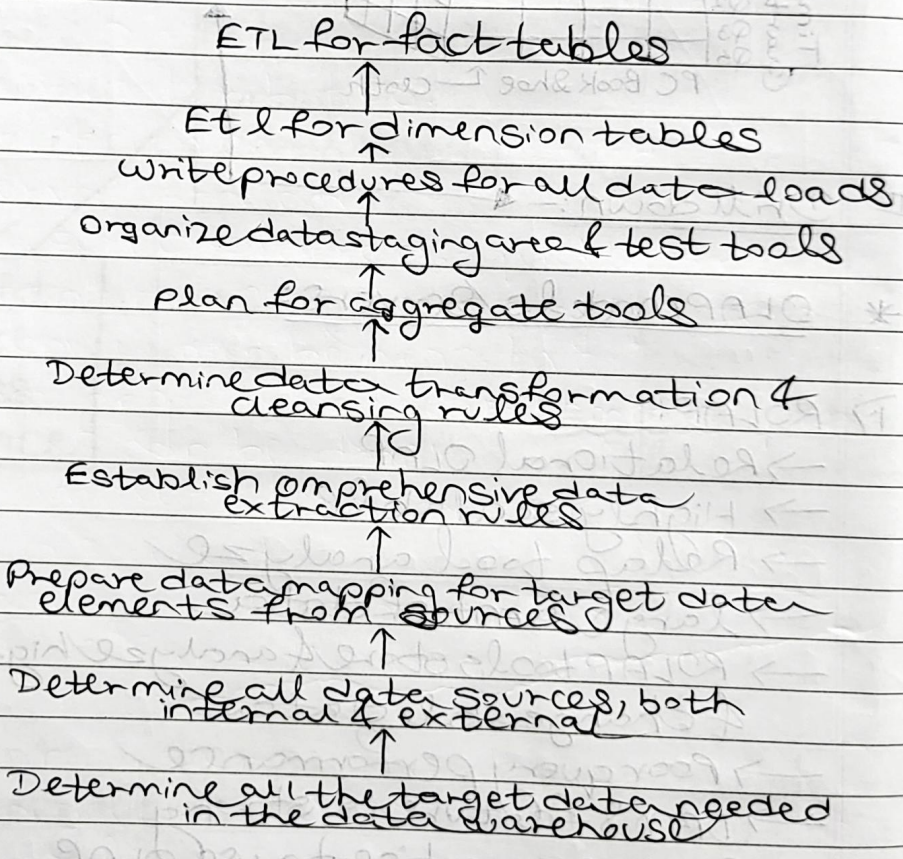
1) ROLAP:
→ Relational OLAP
→ Highly scalable
→ Rolap tool analyze
→ large amt of data
→ ROLAP tools store & analyze highly volatile & changeable data
→ Poor query performance
→ DBMS feature is strong
→ Requires expertise to use OLAP

2) MOLAP:
→ Multidimensional OLAP
→ MOLAP are not capable of containing detailed data

→ It is fast.
→ very easy to use
→ DBMS facility is weak
→ Query processing is fast
→ can perform complex computations.

3) HOLAP → Hybrid of ROLAP & MOLAP

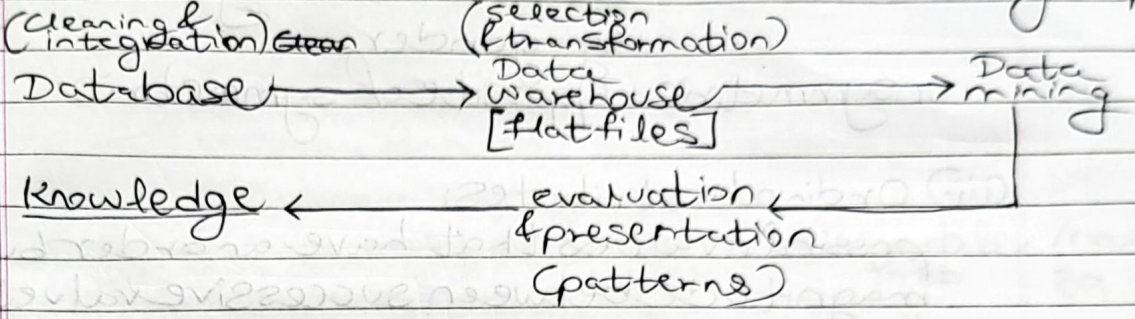
- major steps in ETL Process:-



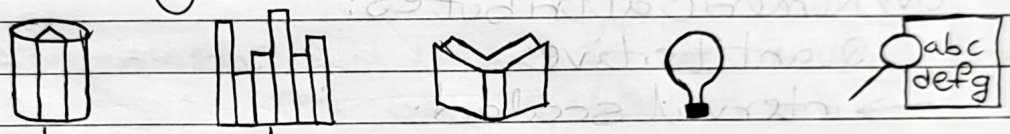
30.7.2024

* Data mining – unit 2:

Imp: KDD Process: Iterative sequence of following steps



- Data mining task primitives:



Knowledge type to be mined: characterization, Discrimination, Association,

Task-relevant
~~data~~
database or
data warehouse
name Database
tables or data
warehouse cubes
Conditions for data
~~selection~~ selection
relevant attributes
or dimensions
Data grouping criteria

- What is an attribute?

Types:-

(1) Nominal: Symbols or names of things & Each value represent some kind of category, code or state & so nominal attributes.
eg- hair color, maritical status, occupation
Non-quantitative

(iii) Binary attributes:
symmetric: both of its states are equally valuable & carry same weight
eg- gender
unsymmetric: Opposite of symmetric

(iii) Ordinal attributes:
possible values that have an order but magnitude between successive values is not known. eg- grade: A+, A, A-, B-

(iv) Numeric attributes:
Quantitative
→ interval scaled:
measured, ordered, equidistant, doesn't have any meaningful zero, interval can be negative
Grouping, sorting & arithmetic operations can be done
example- thermometer
→ ratio scaled:
measured in form of numbers.
It has rank & order
equidistant
meaning zero
no negative values
example: distance travelled, weight, age
grouping, sorting, ~~arithmedia~~
arithmetic op can be performed
mean, median, mode

Statistical description of data:

mean: Center of set of data

$$\bar{x} = \sum_{i=1}^{n} x_i / N$$

Five number summary:

Q1) minimum, 1st quantile, median, 2nd quantile, max

10 | 25th → 25 | 50th → 33 | 75th → 36 | 59

datapoints: 10 11 12 25 25 27 31 33 34 34 35 36 43 50 59

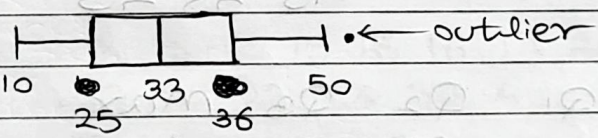
1st quantile = 25, 2nd quantile = 36

IQR = 11

outliers: Datavalue < Q1 − 1.5(IQR) < 8.5

Datavalue > Q3 + 1.5(IQR) > 52.5

= 59



Q2) Dataset: 6 7 8 9 10 11 12

$\sigma^2 = 4$, $sd = \sqrt{4} = 2$

Q3) Datapoints: 5 10 13 15 16 16 20 20 21 22 22 25 25 25 25 30 33 33 35 35 35 35 36 40 45 46 52 70 85

⇒ mean = 30.68

median = 25

$Q_1 = 25^{th}$ percentile $= \frac{25(30)}{100} = 7.50$

$= 20 + 0.50(20 - 20) = 20$

mode median = sort.

Page No.
Date

$Q3 = 75^{th}$ percentile $\Rightarrow \frac{75(30)}{100} = 22.50$

$= \frac{35+36}{2} = 35.5$

Boxplot:

Min	Q1	Q2	Q3	max
5	20	25	35	85

$IQR = Q3 - Q1 = 35 - 20 = 15$

Outliers: Datavalue $< Q1 - 1.5(IQR) < 2.5$

Datavalue $> Q3 + 1.5(IQR) > 57.5$

Q5) Find median, mean mode for data. Show boxplot.

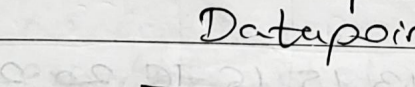
Ans Datapoints = 11 13 13 15 15 16 19 20 20 21 21 22 23 24 30 40 45 45 45

Min	Q1	Q2	Q3	max
11	15	21	30	45

$IQR = 30 - 15 = 15$

Outliers: Datapoint $< 15 - 1.5(15) = 7.5$

Datapoint $> 30 + 1.5(15) = 52.5$



Cab can cap

Page No.
Date

* Binning → dealing with noisy data

Q1) 4, 8, 15, 21, 21, 24, 25, 28, 34. make 3 equal bins.

Ans Bin 1: 4 8 15, Avg = 9
Bin 2: 15 21 21, Avg = 22
Bin 3: 24 25 28, Avg = 29

decat

(i) Smooth by Bin mean
Bin 1: 9 9 9
Bin 2: 22 22 22
Bin 3: 29 29 29

(ii) Smooth by Bin median:
Bin 1: 8 8 8
Bin 2: 21 21 21
Bin 3: 25 25 25

(iii) Smooth by Bin boundaries:
Bin 1: 4 4 15
Bin 2: 15 21 21
Bin 3: 24 24 28

Q2) Datapoints: 11 13 13 15 15 16 19 20 20 20 21 21 22 23 24 30 40 45 45 45 71 72 73 75

make 4 equal bins.

Ans Bin 1: 11 13 13 15 15 16, Avg = 13.80
Bin 2: 19 20 20 20 21 21, Avg = 20.10
Bin 3: 22 23 24 30 40 45, Avg = 30.60
Bin 4: 45 45 71 72 73 75, Avg = 63.50

(i) Smooth by Bin mean:
Bin 1: 13.8 13.8 13.8 13.8 13.8 13.8
2: 20.1 20.1 20.1 20.1 20.1 20.1
3: 30.6 30.6 30.6 30.6 30.6 30.6
4: 63.5 63.5 63.5 63.5 63.5 63.5

(ii) Smoothing by Bin median:

Bin 1: 14 14 14 14 14 14
2: 20 20 20 20 20 20
3: 27 27 27 27 27 27
4: 71.5 71.5 71.5 71.5 71.5 71.5

(iii) Smooth by Bin boundary:

Bin 1: 11 11 11 16 16 16
2: 19 19 19 21 21 21
3: 22 22 22 22 45 45
4: 45 45 75 75 75 75

Bayes Theorem:-

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

* ID3 - Desision Tree classification

p = yes = 9 } ⇒ 14
n = no = 5 }

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

$$= 0.940$$

Compute entropy - outlook

outlook	p	n	I(p,n)
Sunny	2	3	0.971
Overcast	4	0	0
Rain	3	2	0.971

$$E(outlook) = \frac{5}{14}I(2,3) + \frac{4}{14}\cdot I(4,0) + \frac{5}{14}\cdot I(3,2)$$

$$= 0.694$$

Page No.
Date

gain(outlook) = I(p,n) − E(outlook)
= 0.940 − 0.694 = 0.246

compute entropy – temperature

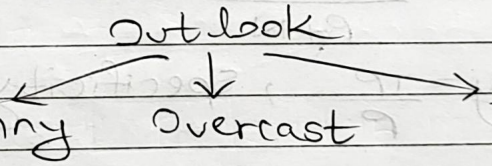
temperature	p	n	I(p,n)
hot	2	2	1
cold	3	1	0.81
mild	4	2	0.918

$E(temp) = \frac{4}{14} I(2,2) + \frac{4}{14} I(3,1) + \frac{6}{14} I(4,2)$

= 0.91

gain(temp) = I(p,n) − E(temp) = 0.940 − 0.91
= 0.03

Gain(humidity) = 0.940 − 0.786 = 0.154
Gain(wind) = 0.048



- Using If–Then Classification:

The If part of the rule is called antecedant or precondition.
Then part is called rule consequent.

Assessment of a rule:
ncovers = # of tuples covered by R
ncorrect = # of tuples correctly classified by R.
coverage(R) = ncovers / |D|, D = training data set
accuracy(R) = ncorrect / ncovers.

Page No.
Date

Q1) R1 which covers 2/14 tuples correctly classified tuple 2.

Ans Coverage = $\frac{2}{14} = 0.14$

Accuracy = $\frac{2}{2} = 1$

- Confusion matrix:

actual ↓ / predicted →	N	P
N	True Negative (TN)	False Positive (FP) ← Type 1 error
P	False Negative (FN) ← Type 2 error	True Positive (TP)

Accuracy = $\frac{TP + TN}{P + N}$

Error rate = 1 − accuracy

$= \frac{FP + FN}{P + N}$

sensitivity = $\frac{TP}{P}$ (True +ve recognition rate), specificity = $\frac{TN}{N}$

precision = $\frac{TP}{TP + FP}$, recall = $\frac{TP}{TP + FN}$

measure of exactness. what % of tuples labelled as +ve & are actually +ve.

Recall is measure of completeness. what % of positive tuples are labelled as +ve.

23-08-

Q1)

Ans

Q1)

Ans

23-08-2024

Page No.
Date

- **Imp questions:**

Qn) Total tuples in dataset is 165. Among which 60 was ~~known~~ classified as 'No' & 105 are classified as 'Yes'. After applying a model, 50 were predicted 'No' & 100 were predicted 'Yes'. Design confusion matrix, accuracy, error rate, precision & recall.

Ans

		Predicted	
		N	Y
105	N	50	10
60	Y	5	100

$$\text{Accuracy} = \frac{100+50}{165} = 0.90 = 90\%$$

$$\text{error rate} = 1 - \text{accuracy} = 1 - 0.90 = 10\%$$

$$\text{precision} = \frac{100}{110} = 0.90 = 90\%$$

$$\text{recall} = \frac{100}{105} = 95\%$$

* **Decision tree problems:**

Qn) [Gatesmashers]

Ans (i) IG of weather

Entropy of entire dataset

$$\Rightarrow S[+9,-5] = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

entropy of attributes:

Sunny: $S[+2,-3] = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

Cloudy: $S[+4,0] = 0$

Rain: $S[+3,-2] = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.97$

Page No.
Date

Information gain = Entropy (whole data) $-\frac{5}{14}$ Ent(S) $-\frac{4}{14}$ Ent(C) $-\frac{5}{14}$ Ent(R)

$= 0.246$

(ii) IG of Temperature:

Entropy of entire dataset

$S[+9,-5] = 0.94$

Entropy of all attributes:

Hot: $S[+2,-2] = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1.0$

Mild: $S[+4,-2] = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.91$

Cold: $S[+3,-1] = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.81$

Information gain $= 0.94 - \frac{4}{14}\times 1 - \frac{6}{14}\times 0.91 - \frac{4}{14}\times 0.81 = 0.029$

(iii) IG of humidity:

Entropy of entire dataset

$S[+9,-5] = 0.94$

Entropy of all attributes:

High: $S[+3,-4] = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.98$

Normal: $S[+6,-1] = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.59$

Gain $= 0.94 - \frac{7}{14}\times 0.98 - \frac{7}{14}\times 0.59 = 0.15$

(iv) IG of wind:

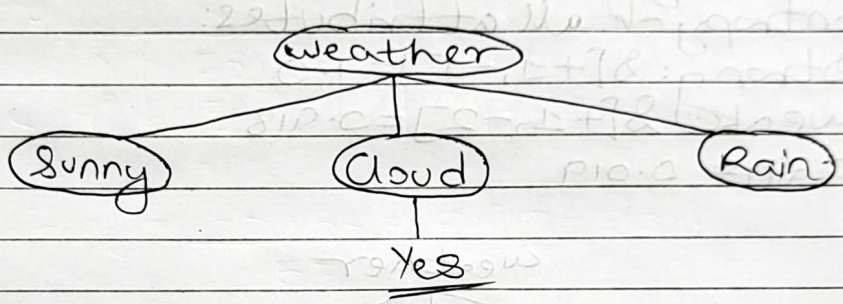
Entropy of entire dataset = 0.94

Entropy of all attributes:

Strong: $S[+3,-3] = \frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1.0$

Normal: $S[+6,-2] = \frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0.81$

$I.Gain = 0.94 - \frac{6}{14} \times 1.0 - \frac{8}{14} \times 0.81 = 0.0478$



Calculating for sunny:

(i) IG of temperature:

Entropy of Sunny $\Rightarrow S[+2,-3] = \frac{-2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

Entropy of all attributes:

Hot: $S[+0,-2] = 0$

mild: $S[+1,-1] = 1.0$

Cool: $S[+1,-0] = 0$

IG of temp $= 0.97 - \frac{2}{5} \times 0 - \frac{2}{5} \times 1 - \frac{1}{5} \times 0$

$= 0.57$

(ii) IG of humidity:

Entropy of Sunny = 0.97

Entropy of all attributes:

High: $S[+0, -3] = 0$

Normal: $S[+2, -0] = 0$

Gain = 0.97

(iii) IG of wind:

Entropy of wind = 0.97

Entropy of all attributes:

Strong: $S[+1, -1] = 1.0$

Weak: $S[+1, -2] = 0.918$

Gain = 0.019