Unit-4:

* Multidimensional data models and different OLAP operations:

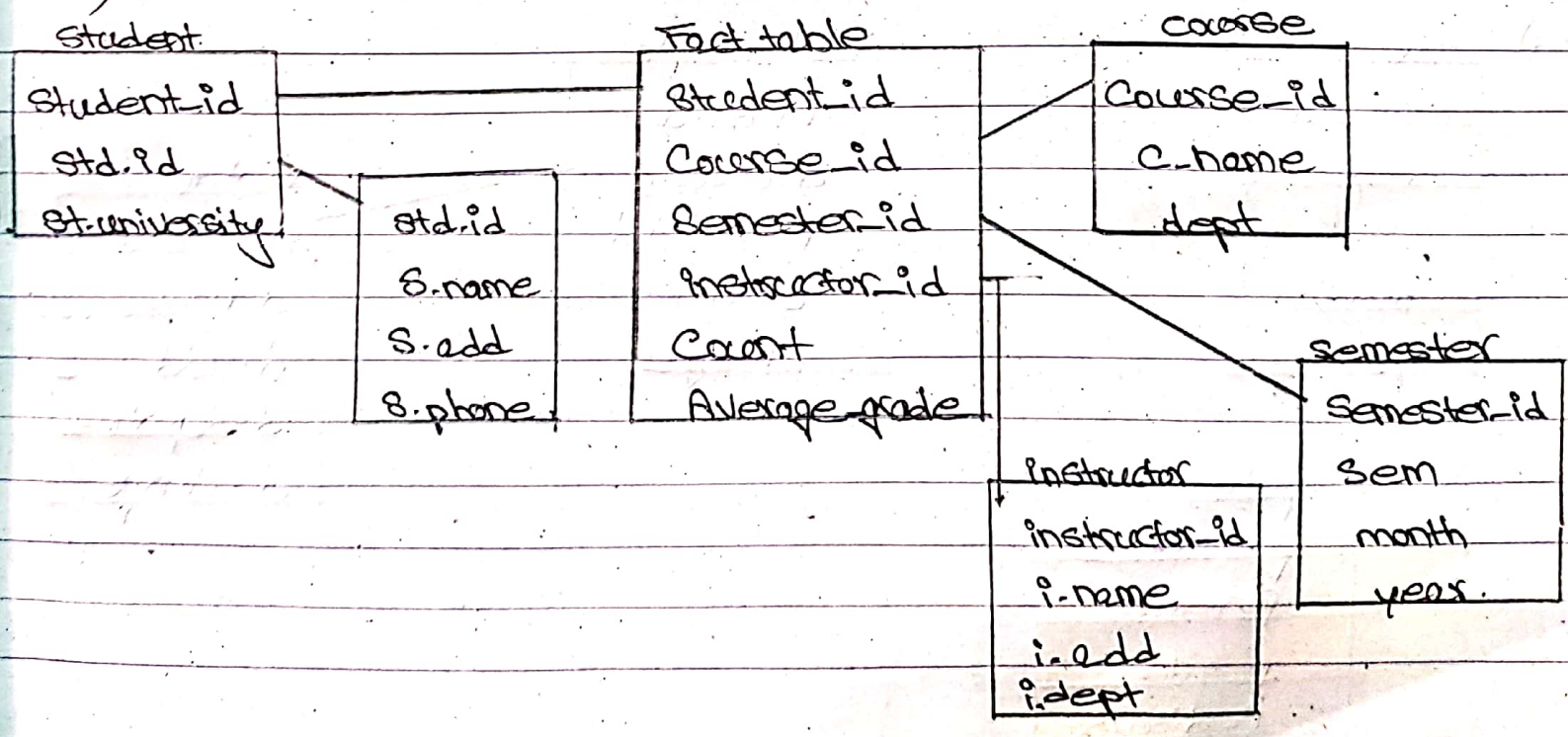
a) Suppose that a data warehouse for big university consists of the following four dimensions:

Student, Course, Semester, an instructor and two measures: Count & average grade.

When at the lowest conceptual level (e.g. for a given student, course, semester & instructor combination), the avg-grade measure stores the actual course grade of the student. At higher conceptual levels, avg-grade stores the average grade for the given combination.

i) Draw a snowflake schema diagram for the data warehouse.

ii) Starting with the base cuboid (student, course, semester, instructor) what specific OLAP operations (eg:- rollup from semester to year) should one perform in order to list the average grade of CS courses for each Big university Std = ?) =



b) 1) Roll up on course from course_id to dept
2) Roll up on student from student_id to university.
3) Dice on course, student.
(course = "CS" AND University = "BU university")

b) Suppose that a data warehouse consists of the three dimensions: time, doctor & patient and the two measures count & charge where charge is the fee that a doctor charges a patient for visit.

i) Draw a schema diagram for the above data warehouse using star, snowflake & fact constellation schemas.

ii) Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?

i)

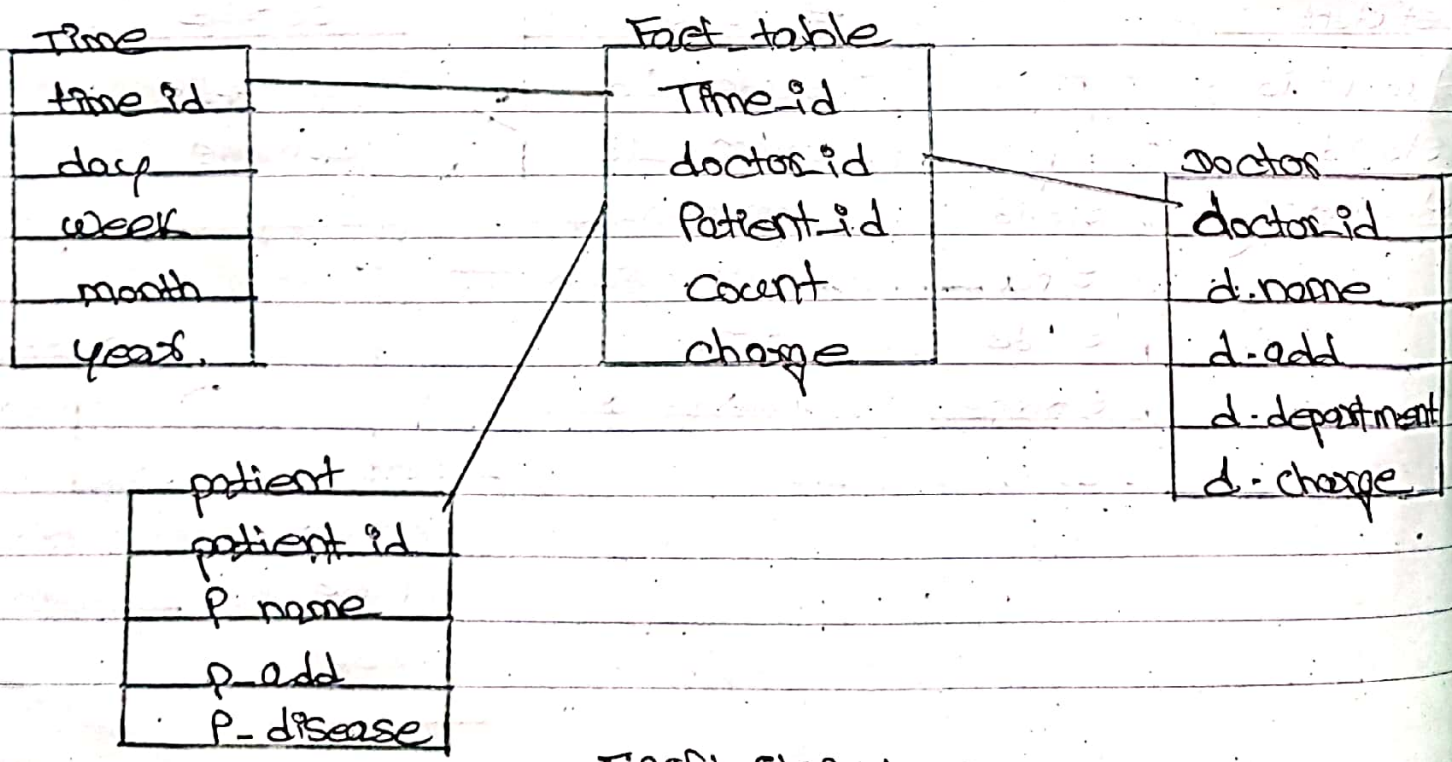


Fig of:- Star schema.

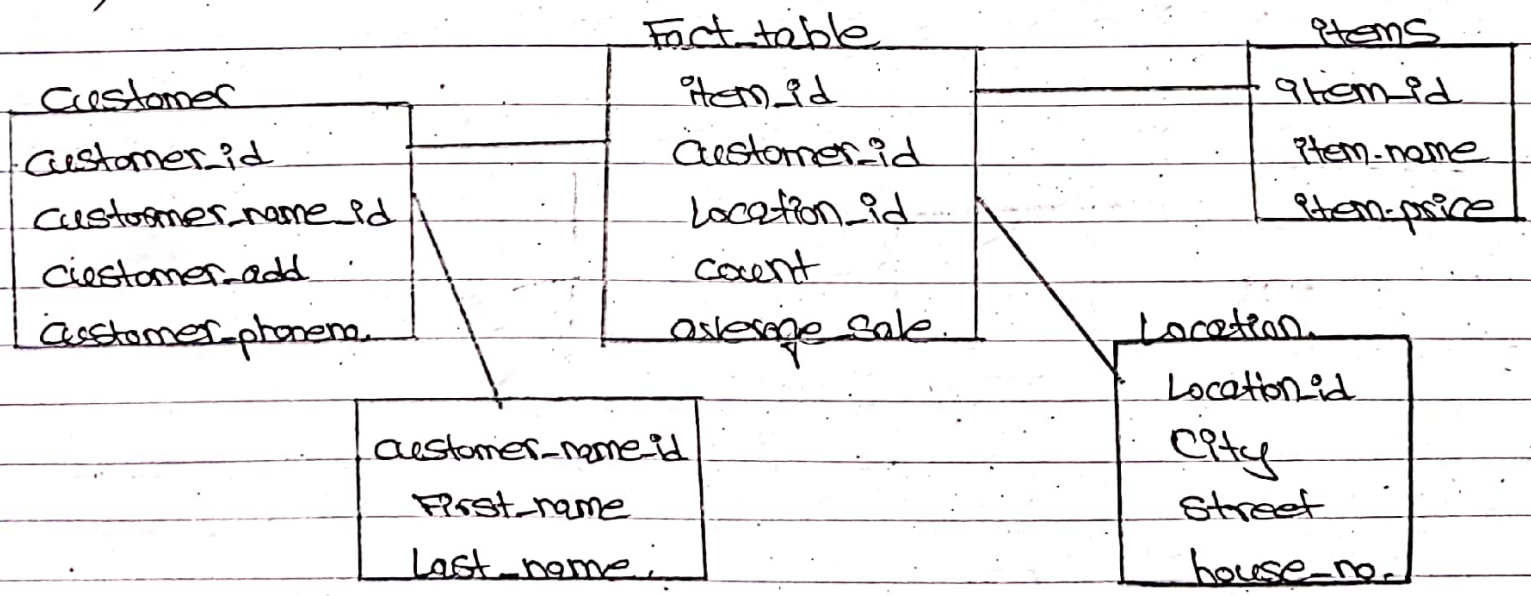
b) = i) Roll up on time from time_id to year.
ii) Roll up on doctor from doctor_id to doctor_name.
iii) Roll up on patient from patient_id to patient_name
iv) Dice on doctor, patient, year.

c) Suppose that a datawarehouse for one of the shopping center consists of the following three dimensions: item, customer, location & two measures count and average_sale.

a) Draw a snowflake schema diagram for the datawarehouse

b) Starting with the base cuboid [item, customer, location], what specific OLAP operations should one perform in order to list the average sales of items for each customer of the shopping center with reference to particular location.

= a) =



b) = Rollup on location from location_id to house_no.
Rollup on customer from customer_id to customer_name_id.
Dice on customer, items, house_no.

* Data Cube Computation:

- It is a process of calculating the no. of cuboids provided by no. of dimensions.

Two approaches:

i) Dimension without hierarchy.

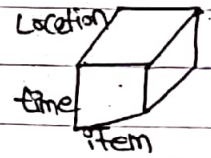
ii) Dimension with hierarchy.

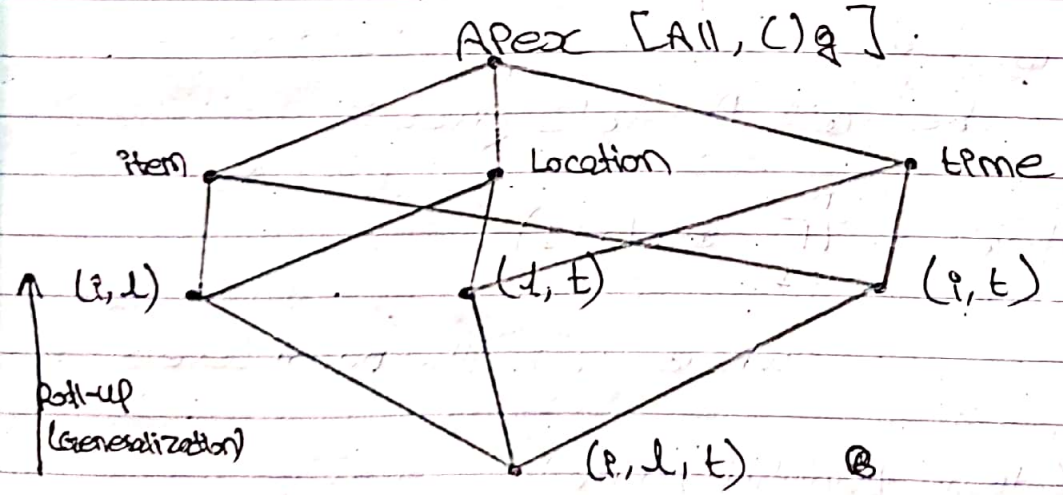
i) Dimension without hierarchy:

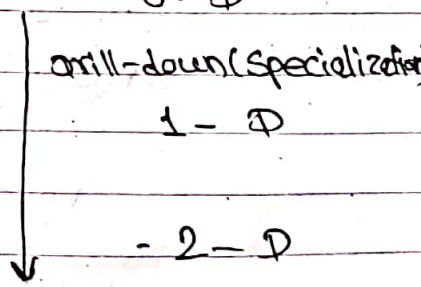
- If there were no hierarchies associated with each dimension then total number of cuboids for n-dimensional data cube, is 2^n.
- However, many dimensions do have hierarchies. for example, the dimension time is usually not explored at only one conceptual level, such as year but rather at multiple conceptual level. such as in hierarchy "day < month < quarter < year".
- For eg:-

$n = 3$ (item, location, time)

cuboids $= 2^n = 8$ cuboids.





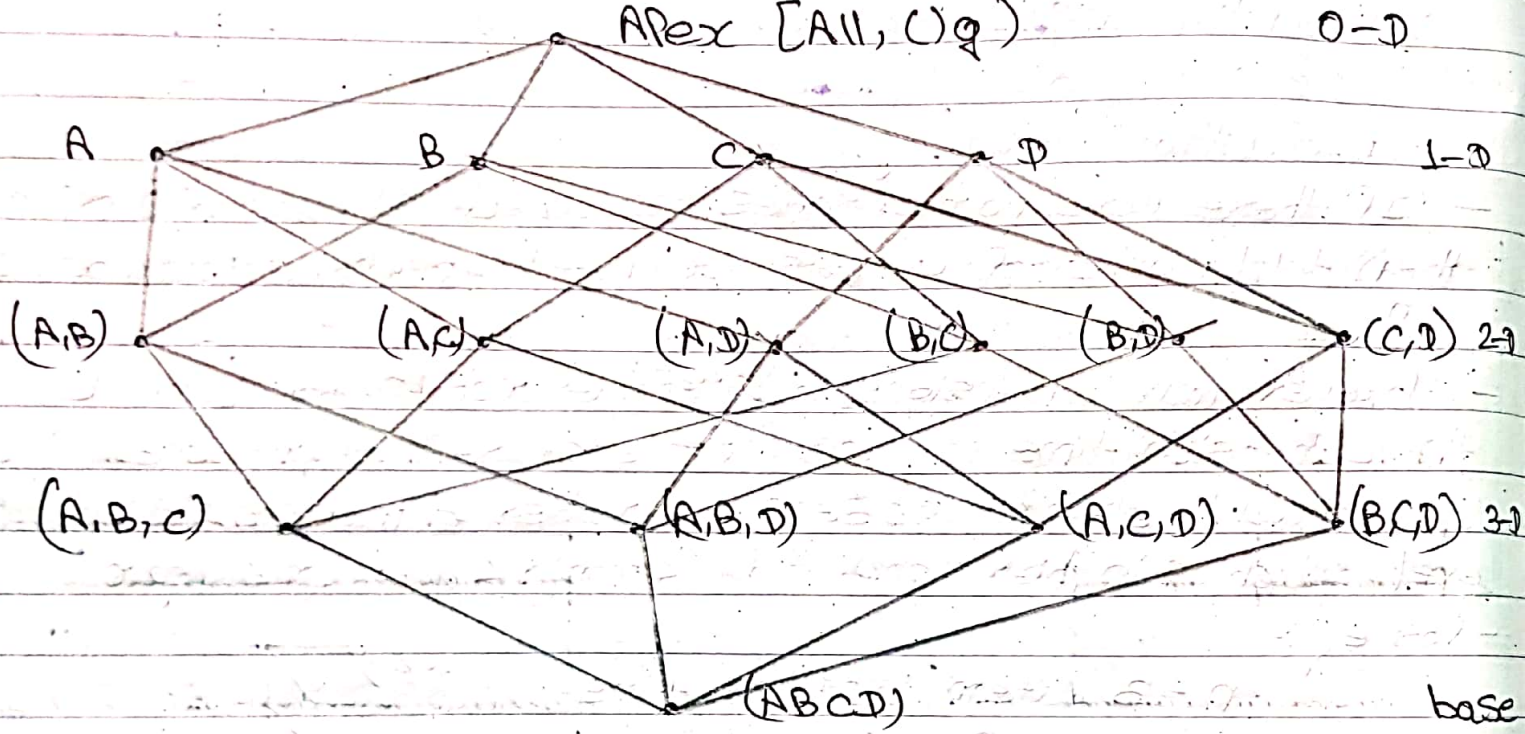


Base:

*Draw a cuboids for 4-dimension A, B, C, d. (Dimension without hierarchy)

= n = 4

cuboids = $2^n = 2^4 = 16$



ii) Dimension with hierarchy:

- n-dimensional cube with L-levels gives:

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

where, L_i is the no. of levels.

For eg:- n = 3 (Item, Location, time)

where, Item & Locations has Level 1 & Time has Level -3 i.e days → months → years

(L1) (L2) (L3)

No. of cuboids = 2 * 2 * 4

no. of cuboids = 16

* Curse of dimensionality:

- space complexity ↑ with respect to dimension ↑
- Dealing with the problem (partial materialization)

* Bitmap Index:

- There are major two categories of indexing:-
  i) Dense
  ii) Sparse.

Keys	B1
Keys	B2
⋮	⋮
Keys	Bn

Fig:- Dense-indexing.

Keys	B1 – B4
Keys	B5 – B8
⋮	⋮
Keys	Bn-4 Bn

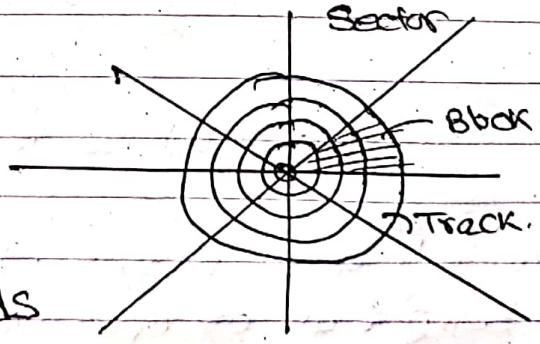
Fig:- Multi-level indexing
Sparse-Indexing.

- It is used to reduce the no. of disk access.

For eg:-

i) Block = 512 B

ii) no. of records that can be kept in a block = $\frac{512}{128} = 4$



iii) no. of blocks required for 100 records
$= \frac{1}{4} \times 100 = 25$

i.e, If no indexing is carried out the total block access is 25.

eg:-

Table 1:

ID	Name	Sex
1	A	M
2	B	F
3	A	M
4	C	F

bit-map →

ID	Name	M	F
1	A	1	0
2	B	0	1
3	A	1	0
4	C	0	1

Table 2:

T.ID	I.D	Item bought
1	1	Y
2	2	Y
3	1	M
4	3	Y

bit-map →

TID	IID	Item bought	
		Y	M
1	1	1	0
2	2	1	0
3	1	0	1
4	3	1	0

① Publish those names who is female & have bought an item?

= We need to use ~~Bitm~~ Bitmap index joint.

So,

Female = (B, C) having id (2, 4)

bought_item = (1, 2, 3)

Bitmap index joint.

So, Those whose is female & have bought an item is B.

② Publish those names who is female or have bought an item?

= being & an or operation include all names who are female or have bought an item.

i.e, (A, B, C, D).

Unit 6:

* Association Rule Mining:

① Frequent pattern: (Apriori Algorithm)

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

for e.g:

Transaction_id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F

i) Find all the rules $X \rightarrow Y$ with minimum support & confidence.

ii) min_sup = 50% and min_conf = 50%

=Sol

Support count = min_sup of total transaction

= 50% of 5

= $\frac{50}{100} \times 5$

= 2.5 Let it be (3)

now,

C1 =

Item_set	Support
A	3 ✓
B	3 ✓
C	2 ✗
D	4 ✓
E	3 ✓
F	2 ✗

L1 =

Item_set	Support
A	3
B	3
D	4
E	3

C_2 =

Item_Set	Support
AB	1 ✗
AD	3 ✓
AE	1 ✗
BD	2 ✗
BE	2 ✗
DE	2 ✗

L_2 =

Item_Set	Support
AD	3

So, The Association rules we get are,

$A \rightarrow D$

$D \rightarrow A$

i) $A \rightarrow D$

Support = probability that a transaction contain AUB (for $A \Rightarrow B$)

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

$$\text{confidence} = \frac{AUB}{\text{occurance of A}} \times 100\%$$

$= \frac{3}{3} \times 100$

$= 100\%$

ii) $D \rightarrow A$

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

confidence $= \frac{3}{4} \times 100\% = 75\%$

* Why pruning:

= Combination of not frequent itemset will never generate frequent itemset. So neglect itemset having support less than the min-support.

Also, Combination of frequent itemset may or may not generate another frequent itemset. So, we need to check every possible rules.

for eg:-

TID	items
1	A, B, C, D, E, F
2	B, C, D, E, F, G
3	A, D, E, H
4	A, D, F, I, J
5	B, D, E, K

Assume support = 60%. Generate frequent item sets using Apriori algorithm. Also generate the Association rules with confidence = 80%.

= Sol

$$\text{Support\_count} = \frac{60}{100} \times 5$$

$$= 3$$

Also, we have confidence = 80%.

C_1 =

Itemset	Support	
A	3	
B	3	
C	2	X
D	5	
E	4	
F	3	
G	1	X
H	1	X
I	1	X
J	1	X
K	1	X

L_1 =

Itemset	Support
A	3
B	3
D	5
E	4
F	3

C_2 =

Itemset	Support	
{A, B}	1	X
{A, D}	3	
{A, E}	2	X
{A, F}	2	X
{B, D}	3	
{B, E}	3	
{B, F}	2	X
{D, E}	4	
{D, F}	3	
{E, F}	2	X

L_2 =

Itemset	Support
{A, D}	3
{B, D}	3
{B, E}	3
{D, E}	4
{D, F}	3

C_3 =

Itemset	Support
{B, D, E}	3

L_3 =

Itemset	Support
{B, D, E}	3

Association rules:

A → D	confidence = $\frac{3}{3}$ = 100%	✓
D → A	confidence = $\frac{3}{5}$ = 60%	✗
B → D	confidence = $\frac{3}{3}$ = 100%	✓
D → B	confidence = $\frac{3}{5}$ = 60%	✗
B → E	confidence = $\frac{3}{3}$ = 100%	✓
E → B	confidence = $\frac{3}{4}$ = 75%	✗
D → E	confidence = $\frac{4}{5}$ = 80%	✓
E → D	confidence = $\frac{4}{4}$ = 100%	✓
D → F	confidence = $\frac{3}{5}$ = 60%	✗
F → D	confidence = $\frac{3}{3}$ = 100%	✓
B → DE	confidence = $\frac{3}{5}$ = 60%	✗
D → BE	confidence = $\frac{3}{5}$ = 60%	✗
E → BD	confidence = $\frac{3}{4}$ = 75%	✗
DE → B	confidence = $\frac{3}{4}$ = 75%	✗
BE → D	confidence = $\frac{3}{3}$ = 100%	✓
BD → E	confidence = $\frac{3}{3}$ = 100%	✓

* F-P Growth (Frequent Pattern):

Q no. 1.

TID	items
1	A, B, C, D, E, F
2	B, C, D, E, F, G
3	A, D, E, H
4	A, D, F, I, J
5	B, D, E, K

Support_Count = 3

=Sol

C_1:

Itemset	Support	
A	3	✓
B	3	✓
C	2	✗
D	5	✓
E	4	✓
F	3	✓
G	1	✗
H	1	✗
I	1	✗
J	1	✗
K	1	✗

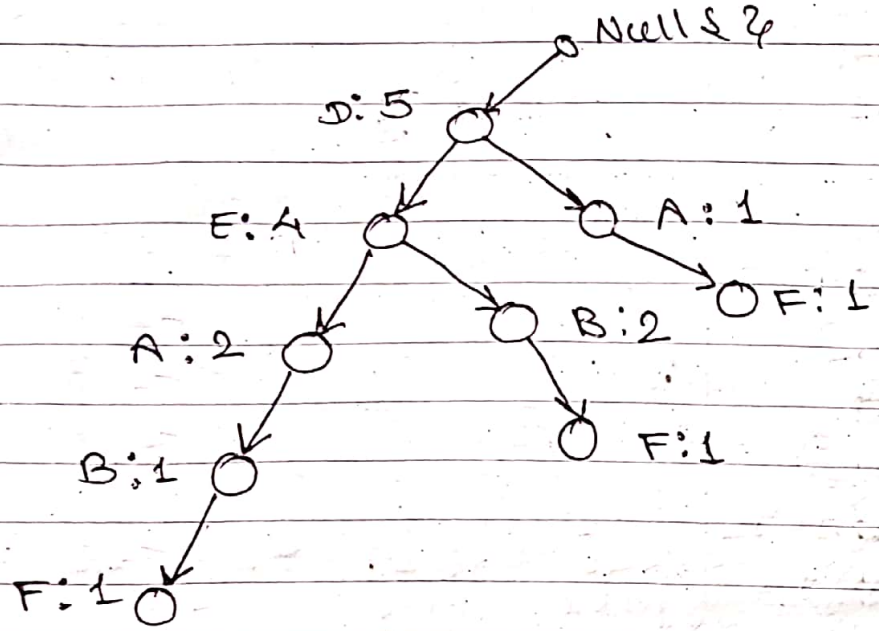
→ sorting in descending order.

L_1:

Itemset	Support
D	5
E	4
A	3
B	3
F	3

now, for each transaction, the respective ordered-itemset is:

TID	items	ordered-itemset
1	A, B, C, D, E, F	D, E, A, B, F
2	B, C, D, E, F, G	D, E, B, F
3	A, D, E, H	D, E, A
4	A, D, F, I, J	D, A, F
5	B, D, E, K	D, E, B



Items	Conditional Pattern Base	Conditional FP-tree	Frequent Patterns
F	{(D,E,A,B:1),(D,E,B:1), (D,A,F:1)}	{D:3}	{DF:3}
B	{(D,E,A:1),(D,E:2)}	{(DE:3)}	{DB:3},{E,B:3} {D,E,B:3}
A	{(D,E:2),(D:1)}	{(D:3}	{D,A:3}
E	{(D:4)}	{(D:4}	{D,E:4}
D	Null		

Q.No.2.:

TID	items
T1	I1, I2, I5
T2	I2, I4
T3	I2, I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I3
T8	I1, I2, I3, I5
T9	I1, I2, I3

↓ L1

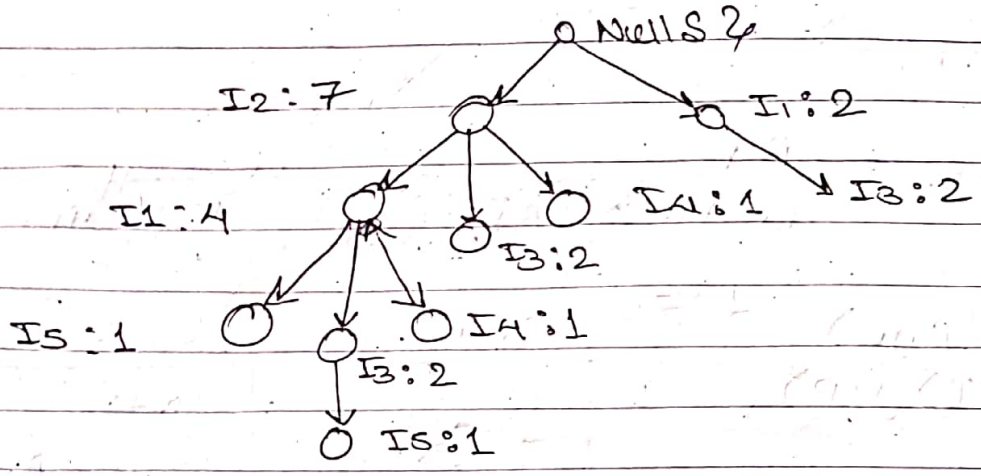
Itemset	Support
I1	6
I2	7
I3	6
I4	2
I5	2

sorting in descending order →

Itemset	Support
I2	7
I1	6
I3	6
I4	2
I5	2

now, for each transaction the respective item_set is:

TID	items	ordered item set
T1	I1, I2, I5	I2, I1, I5
T2	I2, I4	I2, I4
T3	I2, I3	I2, I3
T4	I1, I2, I4	I2, I1, I4
T5	I1, I3	I1, I3
T6	I2, I3	I2, I3
T7	I1, I3	I1, I3
T8	I1, I2, I3, I5	I2, I1, I3, I5
T9	I1, I2, I3	I2, I1, I3



Items	Conditional pattern	Conditional FP-tree	Frequent pattern
I5	℘(I2, I1 : 1), (I2, I1, I3 : 1),	℘(I2, I1 : 2)℘	(I2, I5 : 2) (I1, I5 : 2) (I1, I2, I5 : 2)
I4	℘(I2, I1, ~~I3~~ : 1), (I2 : 1)℘	℘(I2 : 2)℘	℘(I2, I4 : 2℘
I3	℘(I2, I1 : 2), (I2 : 2), (I1 : 2)℘	℘(I2 : 4, I1 : 2), (I1 : 2) ##2 >4 branch ##2	℘(I2, I3 : 4) ## (I1, I3 : 4) ## (I2, I1, I3 : 2
I1	℘(I2 : 4)℘	℘(I2 : 4)℘	℘I1, I2 : 4℘

* Correlation Analysis :

① All-confidence $= \frac{AB}{AB + \max(A\bar{B}, \bar{A}B)}$ or $\frac{\text{Support}(A \cup B)}{\max(\text{Support}(A), \text{Support}(B))}$

② Lift $= \frac{P(A \cup B)}{P(A) \cdot P(B)}$

③ Cosine $= \frac{P(A \cup B)}{\sqrt{P(A) \cdot P(B)}}$

Note:

χ^2		Lift, Cosine & All-Confidence	
+ve	-ve	+ve	-ve
<1	>1	>1,	<1
A↑B↑	A↑B↓		
correlation			

④ $\chi^2 = \sum \frac{(O-E)^2}{E}$ where, O = observed value,
E = Expected value.

eg:- Suppose we are interested in Analysing transactions at XYZ shopping center w.r.t the purchase of computer games and videos. Let game refer to the transaction containing computer games and video refers to the transaction containing videos of the 10,000 transaction analyzed, the data shows that 6,000 of the transactions included computer games, 7500 included videos & 4000 included both computer games & videos.

Let $\text{support}_{min} = 30\%$, $\text{confidence}_{min} = 60\%$

buys (x, games) ⇒ buys (x, videos) predicate?

=Sol Total transactions = 10,000
computer game Transaction = 6,000
Videos Transaction = 7,500
Both computer games & videos = 4,000

① Support of buys (X, games) = $\frac{6000}{10000} \times 100\% = 60\%$

② Support of buys (X, videos) = $\frac{7500}{10000} \times 100\% = 75\%$.

③ Support of buys (X, games & videos) = $\frac{4000}{10000} \times 100\% = 40\%$.

now, confidence of buys (X, games) → buys (X, videos)
A B

$= \frac{A \cup B}{\text{occurence of A}} \times 100\%$

$= \frac{4000}{6000} \times 100\%$

= 66.67% which is greater than min confidence.

So, This is strong rule i.e it is signifying positive correlation

Confusion Matrix for X^2-test:

	Game	Game	Σ rows
video	4000 (E = 4500)	(7500 − 4000) = 3500 (3000−E)	7500
video	(6000 − 4000) = 2000 (E = 1500)	(10000 − 4000 − 3500 − 2000) = 500 (E = 1000)	2500
Σ columns	6000	4000	10,000

now, expected values :

① For Video-game = $\frac{6000 \times 7500}{10000}$

= 4500

② For Video-$\overline{\text{game}}$ = $\frac{4000 \times 7500}{10,000}$

= 3000

③ For $\overline{\text{Video}}$-game = $\frac{6000 \times 2500}{10,000}$

= 1500

④ For $\overline{\text{Video}}$-$\overline{\text{game}}$ = $\frac{2500 \times 4000}{10,000}$

= 1000

now,

① $\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$

$= \frac{(4000-4500)^2}{4500} + \frac{(3500-3000)^2}{3000} + \frac{(2000-1500)^2}{1500}$

$+ \frac{(500-1000)^2}{1000}$

$= 55.56 + 83.33 + 166.67 + 250$

$= 555.56 > 1$

This signifies that there is a negative correlation.

② Lift (Game, Video) $= \frac{P(A \cup B)}{P(A) \cdot P(B)}$

i.e, $P(A \cup B) = \frac{4000}{10000} = 0.4$

$P(A) = \frac{6000}{10,000} = 0.6$

$P(B) = \frac{7500}{10,000} = 0.75$

i.e, $\text{Lift} = \frac{0.4}{0.6 \times 0.75}$

$= 0.89 < 1$

which signifies negative correlation.

③ Cosine $(A, B) = \frac{P(A \cup B)}{\sqrt{P(A) \cdot P(B)}}$

$= \frac{0.4}{\sqrt{0.6 \times 0.75}}$

$= 0.596 < 1$

which signifies negative correlation.

④ All confidence $= \frac{AB}{AB + \max(A\bar{B}, \bar{A}B)}$ or

$= \frac{\text{support}(A \cup B)}{\max(\text{support}(A), \text{support}(B))}$

$= \frac{40\%}{75\%}$

$= 0.53 < 1$

which signifies negative correlation.

* Consider the following Transaction with respective data for the videos & games?

Dataset	GV	$\overline{G}V$	$G\overline{V}$	$\overline{GV}$	All-confidence	Cosine	Lift	χ^2
D_0	4000	3500	2000	0	0.53	0.6	0.84	14778
D_1	4000	3500	2000	500	0.53	0.6	0.89	555.6
D_2	4000	3500	2000	10000	0.53	0.60	1.73	2913

= This are already calculated values for:
All confidence, Cosine, Lift & χ^2 but we need to calculate them using formula in initial page.

Also for the data in D_1 we have calculated the values of all 4-attributes in previous example.

For D_2 let:

Confusion matrix:

	G	$\overline{G}$	Σ row
V	4000	3500	7500
$\overline{V}$	2000	10000	12000
Σ column	6000	13500	19500

① $Lift = \frac{P(GUV)}{P(G).P(V)}$

So, $P(GUV) = \frac{4000}{19500} = 0.205$

$P(G) = \frac{6000}{19500} = 0.31$

$P(V) = \frac{7500}{19500} = 0.38$

$Lift = \frac{0.205}{0.31 * 0.38}$

$= 1.73 > 1$ which implies positive correlation.

② Cosine $= \frac{P(A \cup B)}{\sqrt{P(A) \cdot P(B)}}$

$= 0.60 < 1$ (Negative correlation)

③ All-Confidence $= \frac{A \cdot B}{AB + \max(A\bar{B}, \bar{A}B)}$

$= \frac{4000}{4000 + \max(2000, 3500)}$

$= \frac{4000}{4000 + 3500}$

$= 0.53 < 1$ (Negative correlation)

Note:

i) If these exists Null value in given data set, Lift & χ^2 is not efficient measurement Attribute (result varies largely).

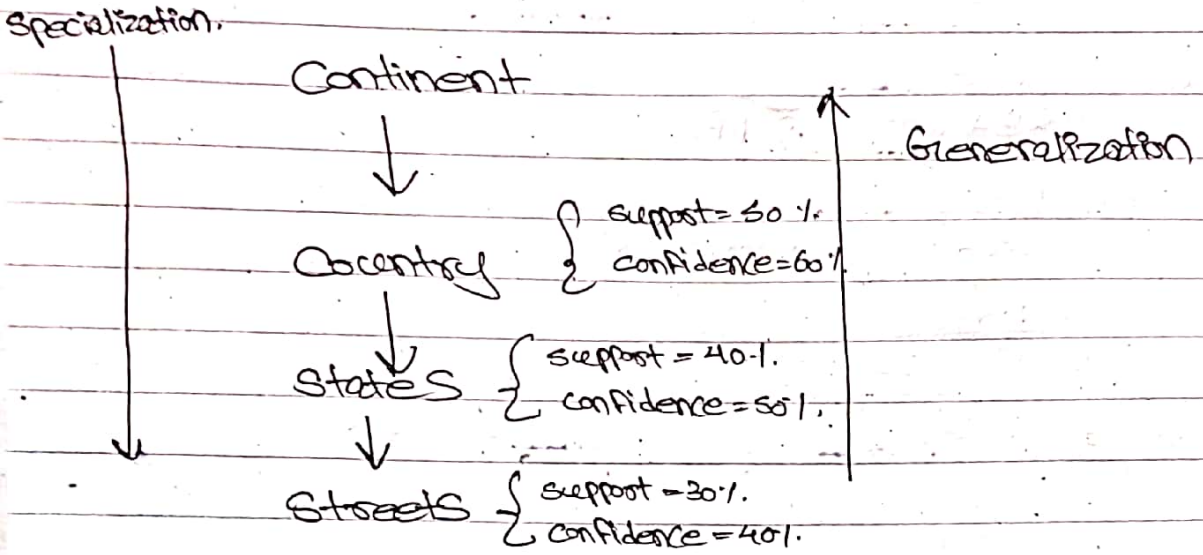
ii) If the given data set is unbalance (not evenly distributed)
- All-confidence is good measurement Attributes,
- Cosine is not good measurement Attributes.

① Multi-Level mining
② Multi-dimensional mining
③ Constrained-based mining
- Monotonic
- Antimonotonic
- Succint
- Convertible

① Multi-level Mining:

① Uniform & non-uniform

② Concept hierarchy.



- Uniform mean equal distribution of support & confidence in all levels.
- non-uniform means unequal distribution of support & confidence in all levels

e.g:- Avg-sale of Sony TV in Asia (Generalized)

let x $

Avg sale of Sony TV in Nepal (Specialized)

② Multi-dimensional:

Dimension (Attributes) - Two or more than two

eg: Age (x, 30 ... 40) ∧ Salary (x, 50K ... 80K) ⇒ buys (x, car) ∨ sales (x, bike)

constrained — Constrained

Implementing constraints as:- 30 ... 40

avg (E-Age) ≤ 25

min()

max() etc.