Subject: D.W.M Sem: V

Q.» Explain different OLAP operations, with suitable example.

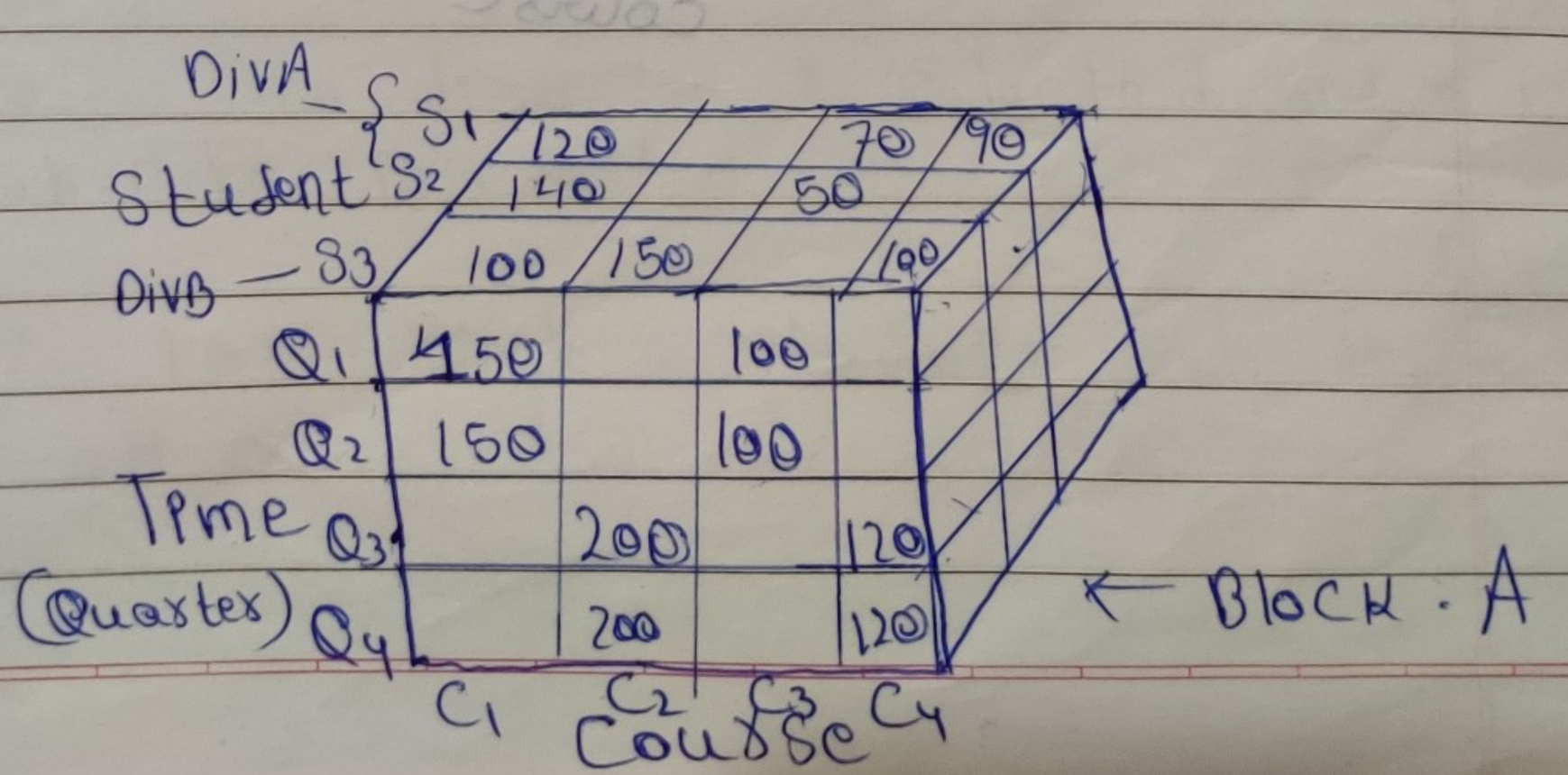
Ans) (i) Online Analytical Processing (OLAP) is a Category software that allows users to analyze information from multiple database systems at the same time. It is a technology that enables analysts to extract and view business data from different point of view.

(ii) The OLAP Cube consists of numeric facts called measures which are categorized by dimensions. OLAP cube is also called the hyper cube.

(iii) Four types of analytical OLAP operations are: 1. Roll-up 2. Drill down 3. Slice and dice 4. Pivot.

(iv) Let, us understand these operations through example of college database having dimensions 1. Course 2. Student 3. Time and fact (measure): Aggregate marks.

(v) Hyper-Cube:-

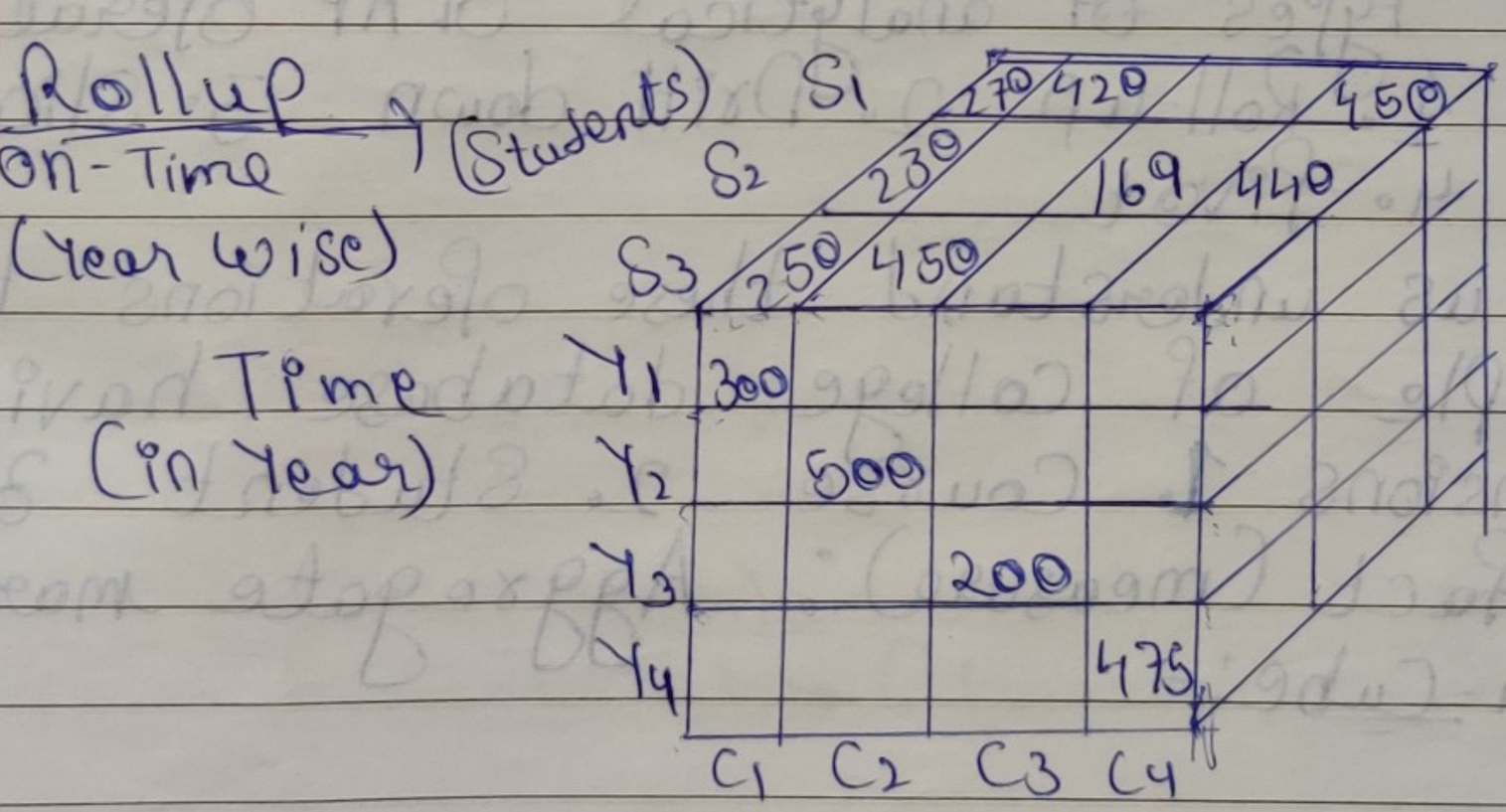


In Block A we have dimension Stundetnt at level individual Student, Time in Quarter and courses at deparment level. to columns and rows i.e each block fill with aggregate marks or NA values (Empty blocks).

1 Roll-up:- Roll-up is also known as "Consolidation" or "aggregation". The Roll-up operation can be performed in 2-ways:-

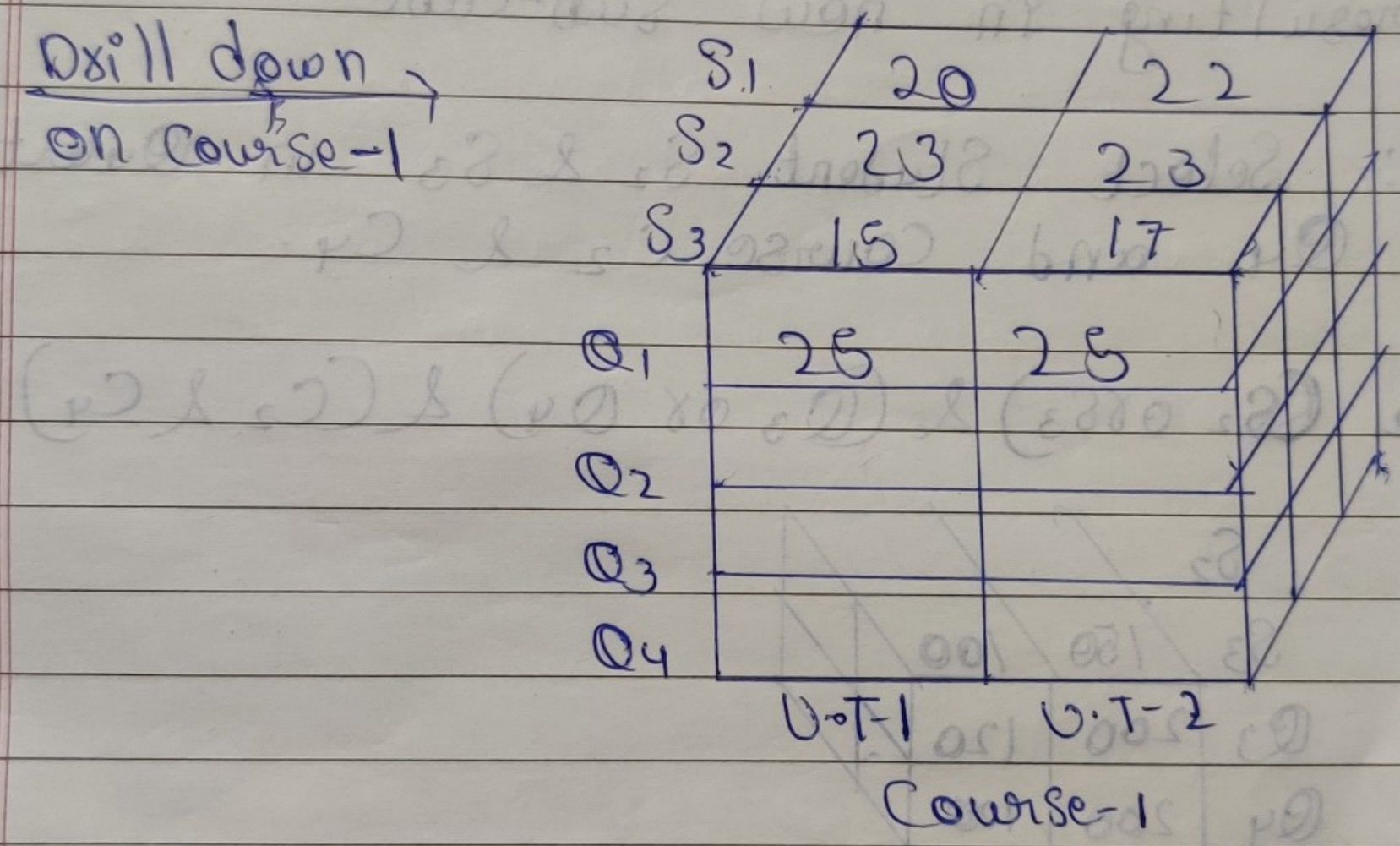
(a) Reducing dimensions (b) climbing up concept hierarchy

e.g:- Get the year wise marks for all Students is roll-up as Semester wise marks are added to get year wise marks & it's one level up hierarchy of dimension:- Time.



2) Drill-down:- In drill-down data is fragmented into smaller parts. It is the opposite of the rollup process. The drill-down operation can be performed in two ways:-
(a) Moving down the concept hierarchy.
(b) Increasing dimension.

e.g:- Drill down on courses to get unit test marks (Hatif Syllabus performance)



If we drill-down on each course, dimensions will be double of original dimension ~~of~~
So, we dice only Course-1.

3) (a) Slice:- One dimension is selected, and a new sub-cube is created.

e.g:- For Single students get the marks for all courses and all Semester.

Slice on Student S_1

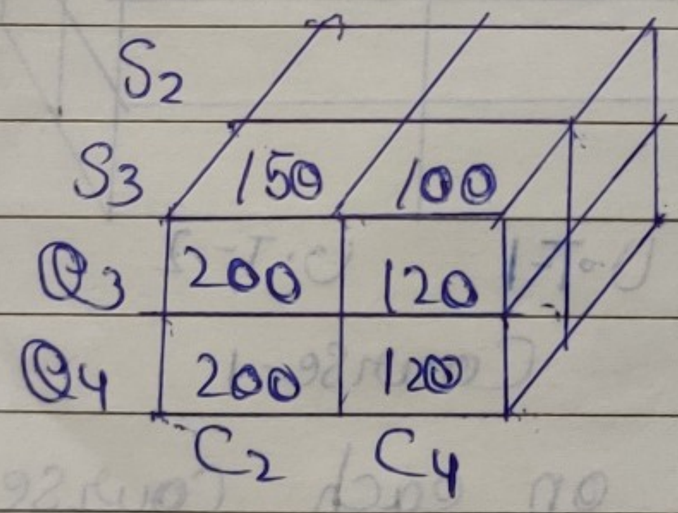
Q_1	120		70	
Q_2	120		~~70~~	
Q_3				90
Q_4				90
	C_1	C_2	C_3	C_4

C_2 is empty because S_1 didn't allote for it.

3) (b) Dice :- Two or more dimension is selected, resulting in new sub-cube.

e.g:- Select Student S_2 & S_3 in Quarter Q_3 & Q_4 and course C_2 & C_4.

Dice $(S_2$ or $S_3)$ & $(Q_3$ or $Q_4)$ & $(C_2$ & $C_4)$:-



4) Pivot :- In Pivot, you rotate the data axes to provide a substitue presentation of data. Analogous to matrix transpose.

e.g:- Pivoting result obtained in Slice.

C_1	120	120		
C_2				
C_3	70	70		
C_4			90	90
	Q_1	Q_2	Q_3	Q_4

Q.» Consider a data warehouse for a hospital where there are 3-dimensions namely
(a) Doctor (b) Patient (c) Time and
2 measures (i) Count (ii) Charge. Where,
Charge is the fee that the doctor
charges a patient for a visit.

(i) Draw Star and Snow-flake Schema.

(Ans) There are four tables:- 3-Dimensions tables
+ 1- fact table with 2 measures.

Dimension Table:-

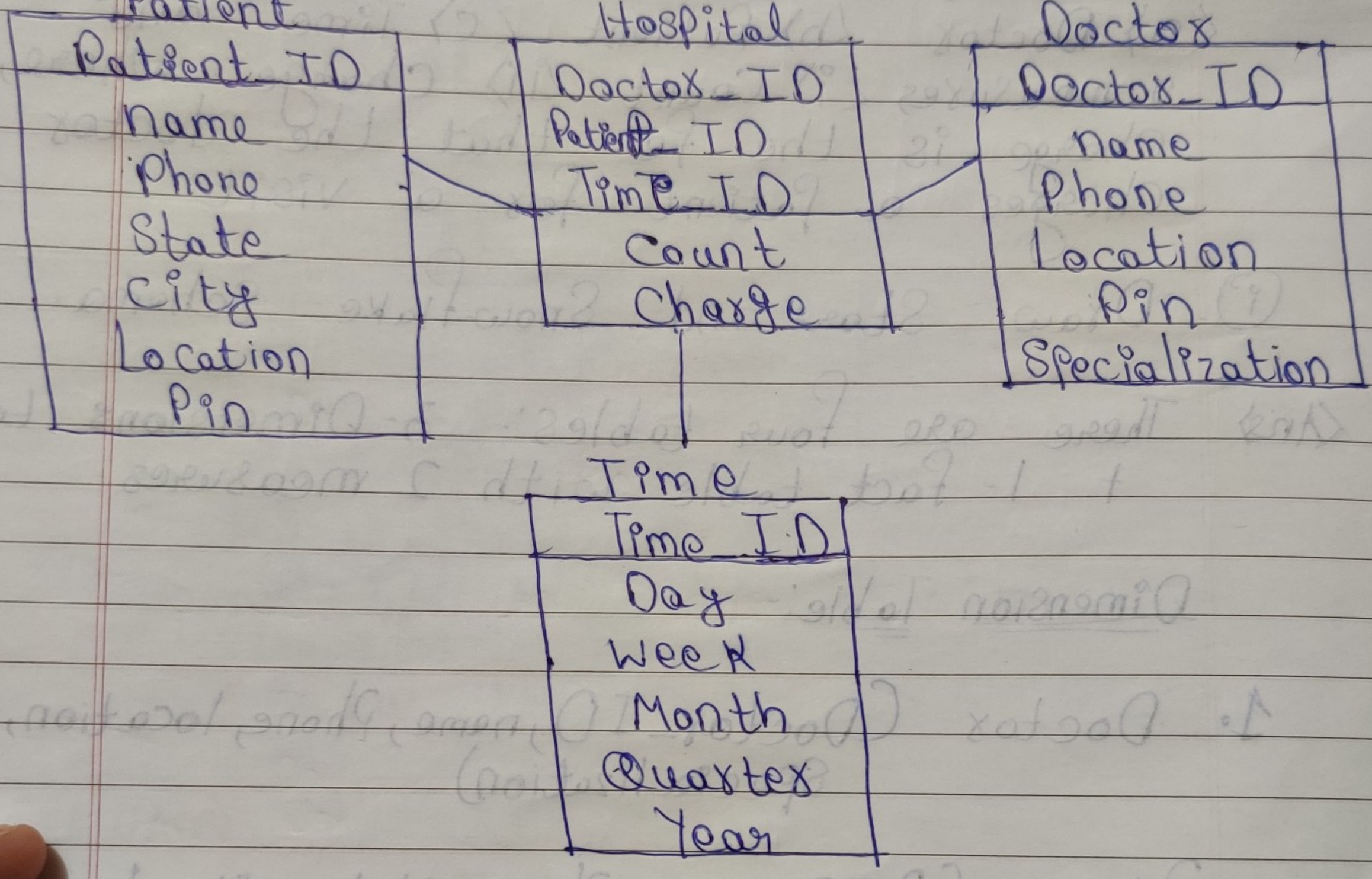
1. Doctor (Doctor_ID, name, Phone, location, Pin
Specialization)

2. Patient (P_ID, name, Phone, State, city, location,
Pin)

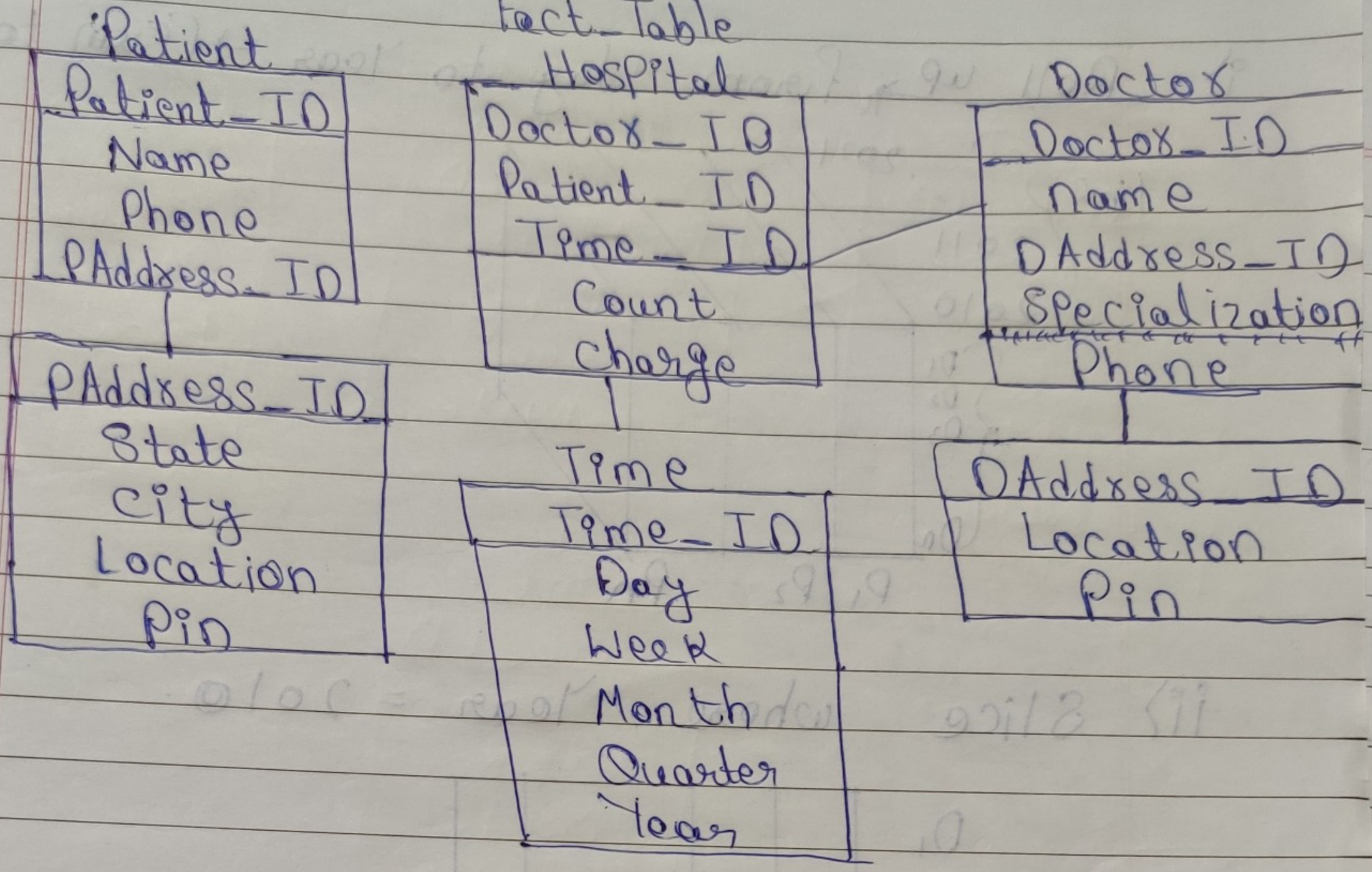
3. Time (T_ID, day, month, week, Quarter, Year)

Fact-Table:-

4. Fact_Table (Doctor_ID, P_ID, T_ID, Count, Charge)


Fact_Table
Hospital
Doctor_ID
Patient_ID
Time_ID
Count
Charge
Patient
Patient_ID
name
Phone
State
City
Location
Pin
Doctor
Doctor_ID
name
Phone
Location
Pin
Specialization
Time
Time_ID
Day
Week
Month
Quarter
Year

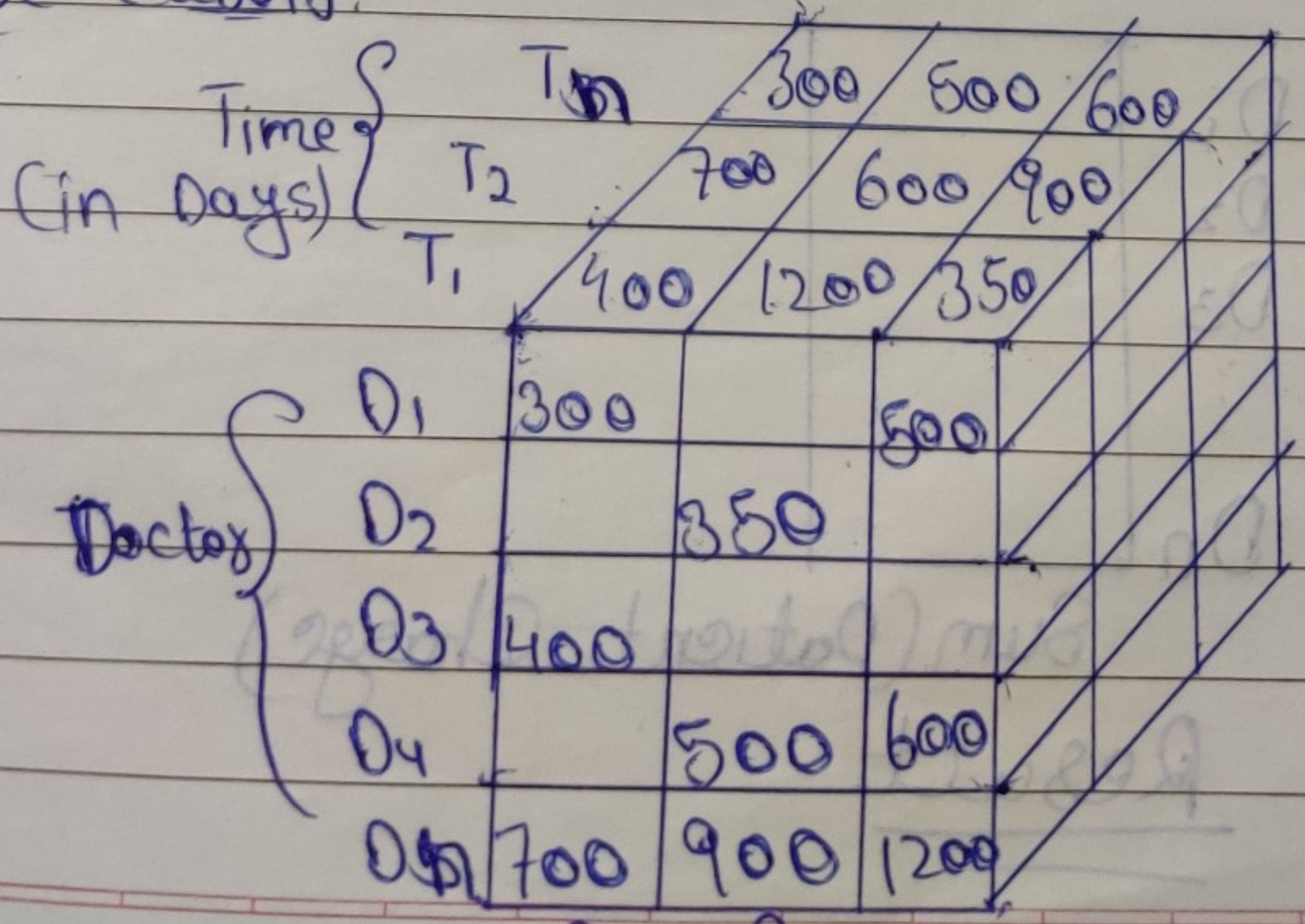
Star - Schema



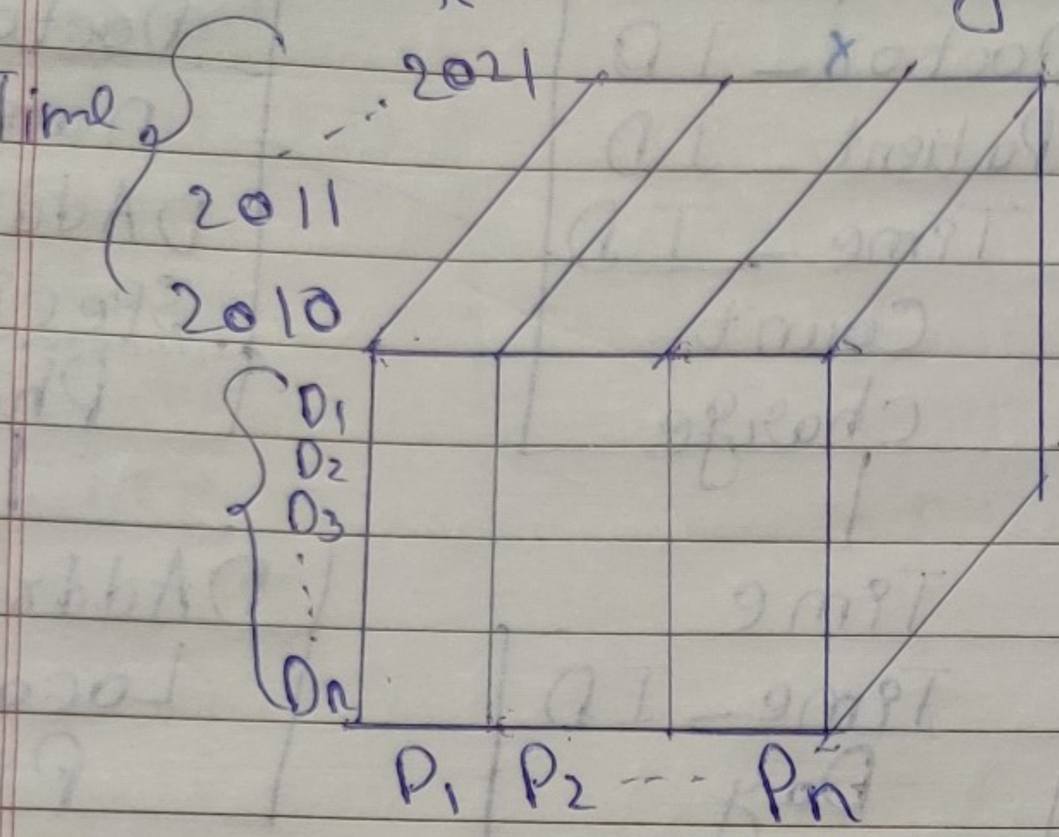
Snow-flake schema.

(ii) Starting with based cuboid [Day, Doctor, Patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2010.

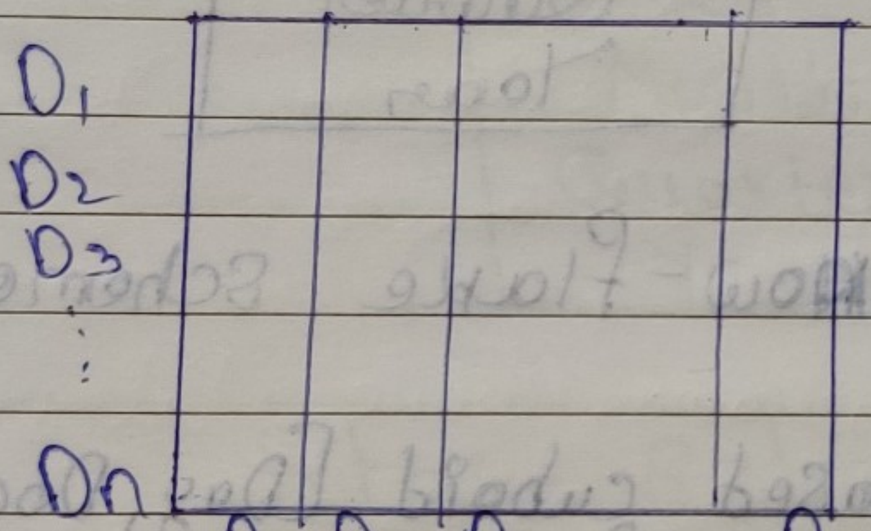
Ans) Base Cuboid:-



i) Roll up → from day to Year → (Sum to Year)

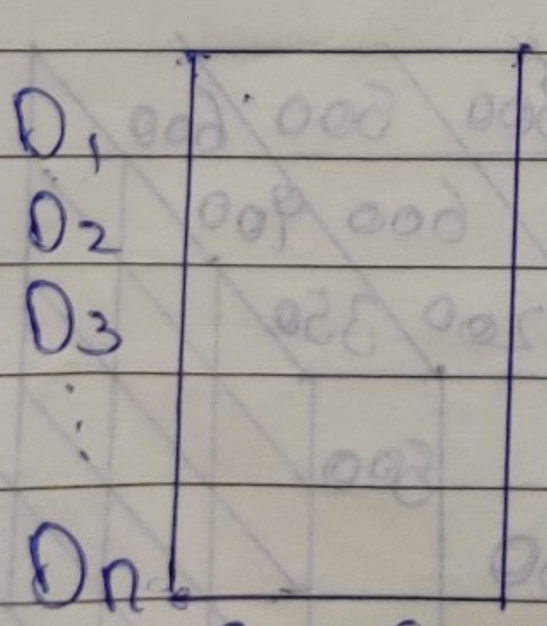


ii) Slice where Year = 2010.



Year = 2010

(iii) Roll-up on Patient group by doctor.



Sum(Patient_Charge)

Result

iii) To obtain the same list write an SQL query the data are stored in R.D.B with the schema Fee. [Day, Month, Year, Doctor, Hospital, Patient, Count, Charge]

Ans) SELECT Doctor, SUM(Charge) FROM Fee WHERE Year = 2010 GROUP BY Doctor.

Q.) Suppose that the data for analysis includes the attribute Age, the Age values for the data tuples are (in increasing order)
[13, 15, 16, 16, 19, 20, 23, 29, 33, 41, 44, 53, 62, 69, 72] use min_max normalization to transform the value 45 for ~~Charge~~ Age onto the range [0.0, 1.0]

Ans) $v' = \frac{(v - minage)}{maxage - minage} [new\_maxage - new\_min_{Age}] + new\_minage.$

$v = 45$, $minage = 13$, $maxage = 72$
$new\_maxage = 1.0$, $new\_minage = 0.0$

$\therefore v' = \frac{(45 - 13)}{72 - 13} [1.0 - 0.0] + 0.0$

$\therefore$ $v' = \frac{16}{31} = 0.51613$

Q.» Describe the steps involved in data mining when viewed as a process of Knowledge discovery.

Ans Steps involved in data mining when viewed as process of Knowledge discovery:-

1.) Data cleaning :- Data from multiple source (DataBase, flat-files, and different form) are extracted and clean to remove noise and inconsitent data.

2.) Data integration:- Since, data is extracted from multiple source they may be combined to provide useful information Domain Knowledge is necessary for proper integration. Generally, stored in R.D.B.

3.) Data Selection:- After cleaning & integration data relevant to the analysis task are retrived from the data ware-house. e.g:- Feature Sub-set Selection.

4.) Data transformation:- Data are transformed and consolidated into forms appropriate for mining by performing Data reduction, different transformation technique & aggregation.

5) Data-mining:- It is an essential process where Intelligent methods are applied, algorithms from various different fields such as Machine learning, Image processing and Natural language processing are used to extract novel data patterns.

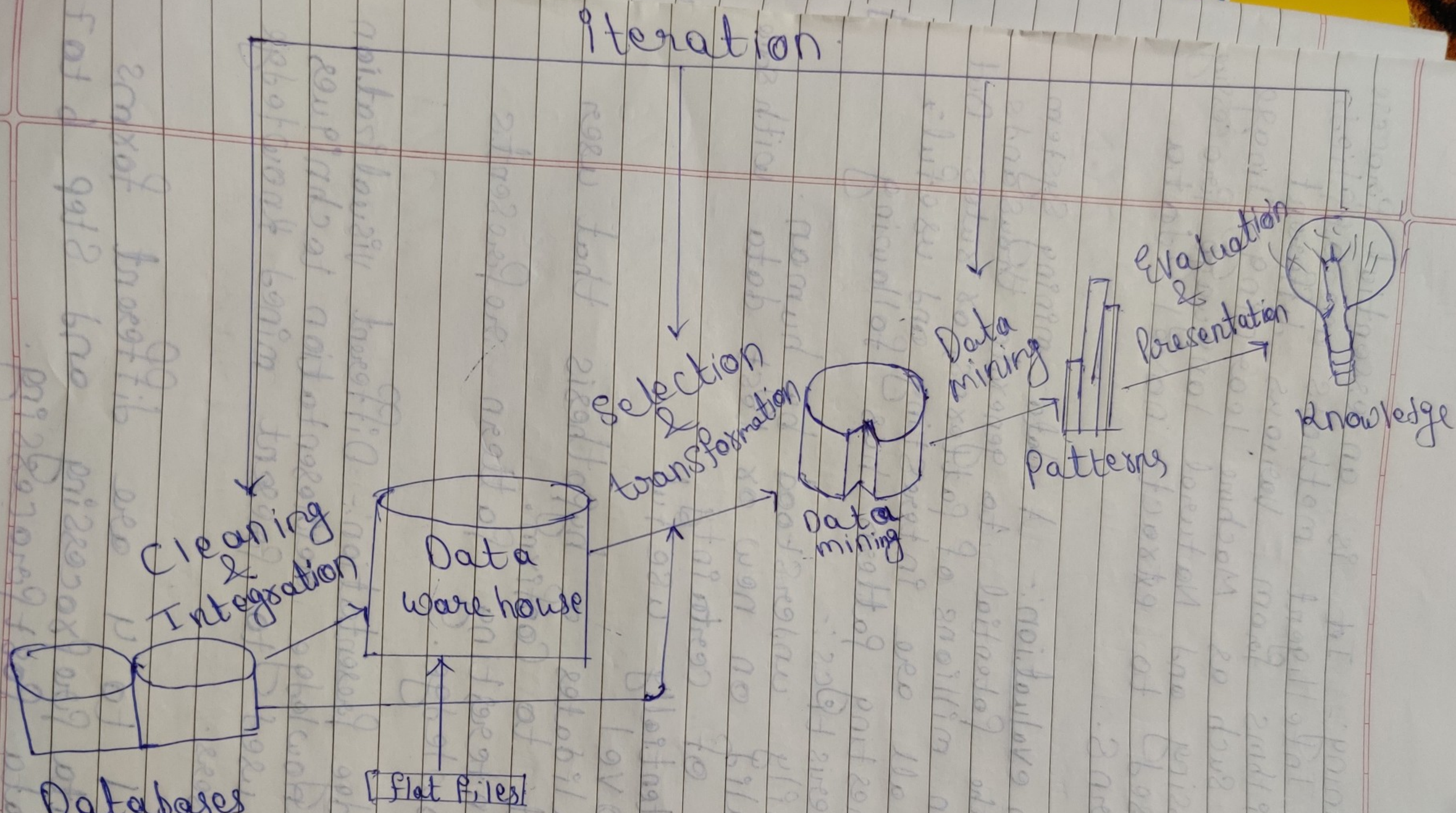
6) Pattern evaluation:- A data mining system has the potential to generate thousands or even millions of patterns, or rules. But Not all are interesting and useful; Interesting pattern has following Charateristics:-

(a) Easily understood by human.
(b) Valid on new or test data with some degree of certainty
(c) Potentially useful
(d) Novel
(e) Validates a hypothesis that user Sought to Confirm.

An interesting pattern represents Knowledge.

7) Knowledge presentation:- Different visualization and Knowledge representation techniques are used to present mined Knowledge to users.

Step 1 to 4 are different forms of data preprocessing and Step 6 to 7 is data postprocessing.



Data mining as a step in the process of knowledge discovery.

<<Q.>> (i) Why is tree pruning is useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning. (ii) Given a decision tree you have the option of (a) Converting D.T to results & then pruning the resulting rules. (b) Pruning the D.T & then converting pruned tree to rules, what advantage does (a) have over (b)

Ans (1) (I) The decision tree built may overfit the training data. There could be too many branches some of which may reflect anomalies in training data due to noise or outliers.

(2) Tree pruning addresses this issue of overfitting the data by removing the least reliable branches (using statistical measures)

(3) This generally results in more compact and reliable decision tree that is faster and more accurate in its classification of data.

(4) The drawback of using separate set of tuples to evaluate pruning is that it may not be representative of training tuples used to create the original decision tree.

(5) If the separate set of tuples are skewed then using them to evaluate the pruned tree would not be a good indicator of the pruned tree's classification accuracy.

(6) Further more, using separate set of tuples to evaluate pruning means there are less tuples to use for creation and testing of the tree.

(ii) once a decision tree has been constructed, it is a simple matter to convert it into an equivalent set of rules.
Converting a decision tree to rules before has three main advantages [advantages of (a) over (b)]:
1. Converting to rules allows distinguishing among the different contexts in which a decision node is used.
(a) Since each distinct path through the decision tree node produces a distinct rule, the pruning decision regarding that attribute test can be made differently for each path.
(b) In constrast, if the tree itself were pruned, then only two choices would be: (1) Remove the decision node completely or (2) Retain it in its original form.
2. Converting to rule removes the distinction between attribute tests that occur near the root of the tree & those that occur near the leaves. We thus avoid messy book keeping issues such as how to reorganize the tree if the root node is pruned while retaining part of subtree below this test.
3. Converting to rules improves readability. Rules are often easier for human to understand.
To generate rules, trace each path in the decision tree, from root node to leaf node, recording

the test outcomes as antecedents and the leaf-node classification as the consequent.

Q) A database has 5 transactions, let min. Support = 60% & min. confidence = 80%

TID	Item_Bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

(a) Find all frequent item set using Apriori algorithm.

(b) List all the strong association rules matching the following, where 'x' is a variable representing Customers & item i denotes variable representing items {e.

$$\forall x \in \text{transaction}, buys(x, item_i) \wedge buys(x, item_j) \Rightarrow buy(x, item_k)$$

Soln: 1 frequent item set:

A → 1 → 1/5 < 60% X
C → 2 → 2/5 < 60% X
D → 1 → 1/5 < 60% X
E → 4 → 4/5 >= 60% ✓
K → 5 → 5/5 >= 60% ✓
I → 1 → 1/5 < 60% X
M → 3 → 3/5 >= 60% ✓
N → 2 → 2/5 < 60% X
O → 4 → 4/5 >= 60% ✓
U → 1 → 1/5 < 60% X

Y → 3 → 3/5 >= 60% ✓

1st frequent item-set = {E, K, M, O, Y}

2nd frequent item-set:

{E, K} → 4 → 4/5 ✓

{E, M} → 2 → 2/5 X

{E, O} → 3 → 3/5 ✓

{E, Y} → 2 → 2/5 X

{K, M} → 3 → 3/5 ✓

{K, O} → 3 → 3/5 ✓

{K, Y} → 3 → 3/5 ✓

{M, O} → 1 → 1/5 X

{M, Y} → 2 → 2/5 X

{O, Y} → 2 → 2/5 X

2nd frequent item-set => {{E, K}, {E, O}, {K, M}, {K, O}, {K, Y}, ~~{O, Y}~~}

3rd Frequent item-set:

{E, K, O} → 3 → 3/8 ✓

{E, K, M} → 1 → 1/8 ×

{E, K, Y} → 2 → 2/8 ×

{K, M, O} → 1 → ×

{K, M, Y} → 2 → 2/8 ×

{K, O, Y} → 2 → 2/8 ×

3rd - Frequent item set = {E, K, O}

Association Rule	Support	Confidence
{E, K} → {O}	3	3/4 ⇒ 75% ×
{O} → {E, K}	3	3/4 ×
{E, O} → {K}	3	3/3 ⇒ 100% ✓
{K} → {E, O}	3	3/5 ⇒ 60% ×
{O, K} → {E}	3	3/3 ⇒ 100% ✓
{E} → {O, K}	3	3/4 ⇒ 75% ×

b) List of association rules that satisfies min. confidence ⇒

⟨E ∧ O⟩ ⟶ K

⟨O ∧ K⟩ ⟶ E