JUNAID . GIRKAR
60004190057
TE COMPS A4
SEM : 5

11-02-22

DMW
END SEM EXAM

Q1

a] FACTS : - Number of occupied rooms
- Number of vacant rooms revenue

DIMENSIONS : - Time
- Hotel
- Room
- Customer

DIMENSION HIERARCHY :

-TIME : Data, day of week, day of month, week, month, quarter, half, year, holiday.

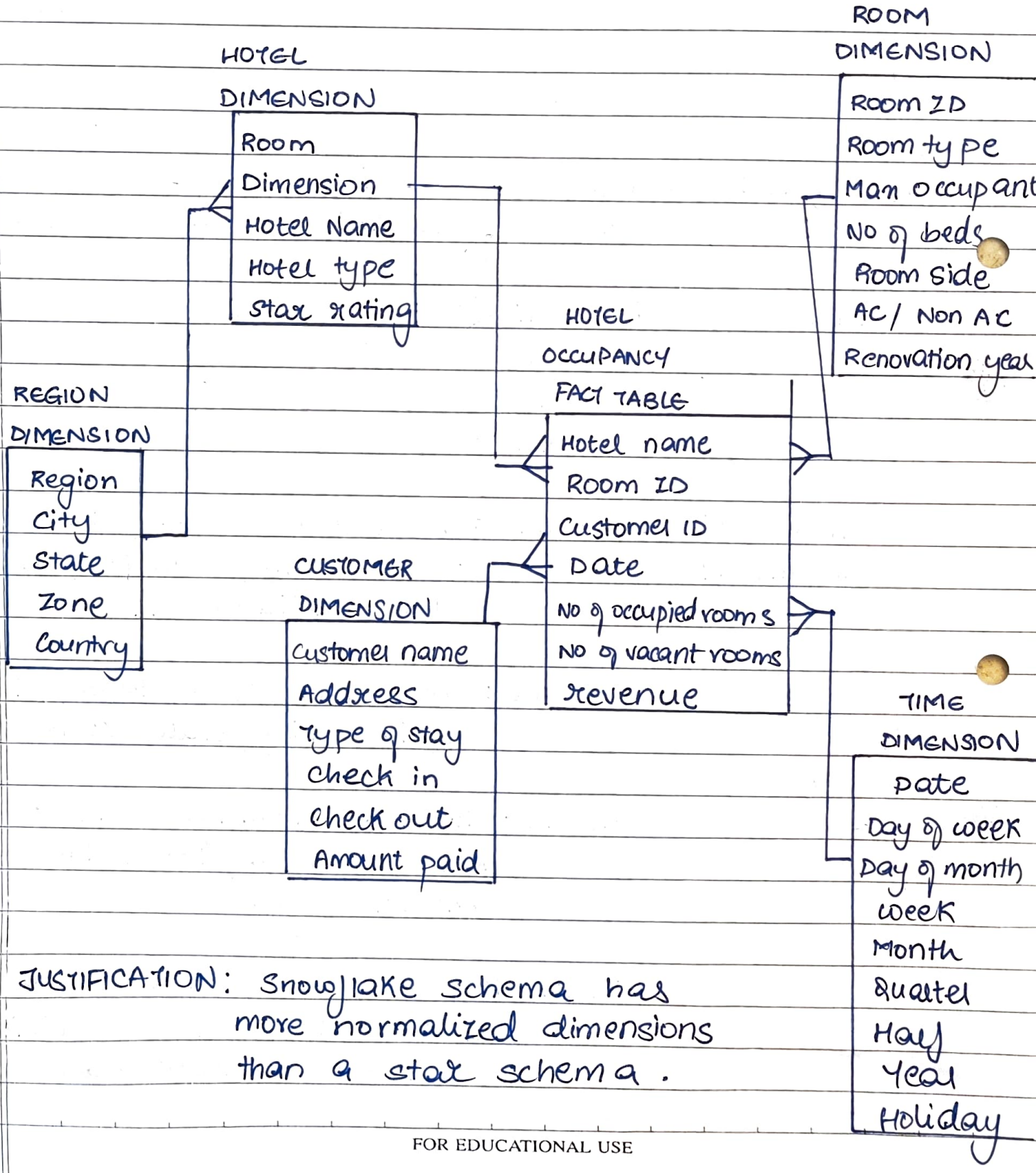
-HOTEL : Hotel name, hotel type, rating, regime, city, state, zone, country

-ROOM : Room ID, Room type, max occupants, number of beds, Room side, AC non AC, renovation year.

-CUSTOMER : Customer ID, customer name, address, type of stay, check in, check out, amount paid.

600041900 57

SCHEMA TYPE : Snowflake Schema




JUSTIFICATION: Snowflake schema has more normalized dimensions than a star schema.

6000419005-7

Q1 b] Spatial data is any type of data that directly or indirectly references a specific geographical area or location. Sometimes called geospatial data or geographic information, spatial data can also numerically represent a physical object in a geographic coordinate system.

Structures that provide information required for computer to reconstruct spatial data model in digital form are defined as spatial data structure. Many GIS software have specific capabilities for storing and manipulating attributes data in addition to spatial information. However, basic spatial data structures in GIS are mainly vector and raster.

RASTER: Raster or grid data structure refers to the storage of the raster data for data processing and analysis by the computer. There are mainly three commonly used data structures such as cell-by-cell encoding, run-length encoding and quadtree

VECTOR: Desaiption of geological phenomena explained in the form of point, line or polygons is called as vector data structure. There are commonly two data structure used in vector GIS data storage viz. topological and non-topological structures

6000419005J

IMPORYANCE :

The most common way of spatial data processing is using a GIS or geographic information system. These are programs that work together to help users make sense of their spatial data. Each spatial dataset may be referred to as a layer.

Q2 a]

Data transformation is a technique used to convert the raw data into a suitable format that eases data mining in retrieving the strategic information efficiently and quickly.

Data transformation is one of the essential data pre-processing technique that must be performed on the data before data mining in order to provide patterns that are easier to understand.

Strategies used in data transformation are :-

1) Data Smoothing : Smoothing the data means removing noise from the considered data set using techniques such as binning, regression and clustering.

2) Attribute construction : In this, new attributes are constructed consulting the existing set of attributes in order to construct a new data set that eases data mining.

3) Data Aggregation : Data aggregation transforms a large set of data to a smaller volume by implementing aggregation operation on the dataset.

4) DATA NORMALIZATION : Normalizing the data refers to scaling the data values to a much smaller range like such as [-1, 1] or [0.0, 1.0]. There are different methods to normalize the data. such as :-

→ Min-max normalization
→ Z-score normalization
→ Decimal scaling.

5) CONCEPT HIERARCHY GENERATION OF NOMINAL DATA : The nominal attribute is one which has a finite number of unique values and there is no ordering between the values. The nominal attributes form the concept hierarchy which converts the data into multiple levels.

6) DATA DISCRETIZATION : Data discretization promotes the transformation of data by replacing the values of numeric data by ~~interval~~ interval labels.

Data discretization can be classified into two types : supervised discretization where the class information is used and the other is unsupervised discretization which is based on which direction does the process proceed i.e. "top-down" or "bottom-up" strategy.

some methods to implement data discretization are :-

6000419005

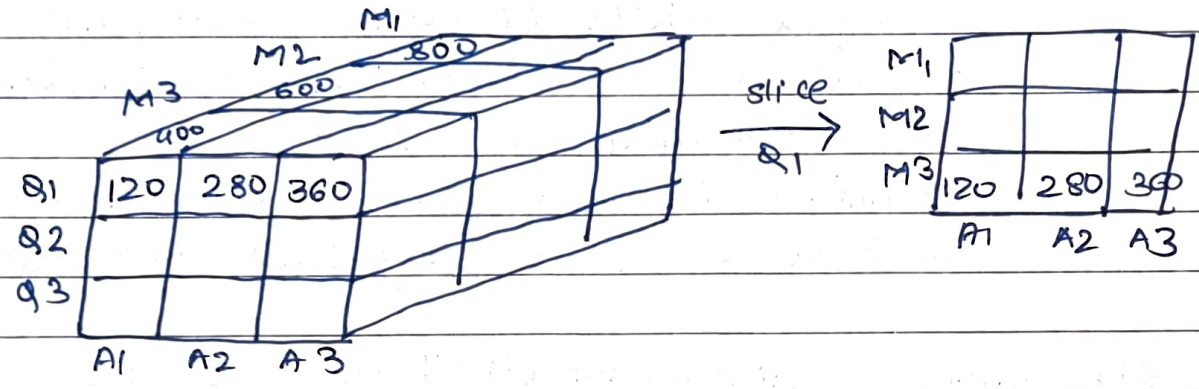
a) BINNING : This method can be used for data discretization and further - for developing concept hierarchy. The observed values or for an attribute are distributed into a number of bins of equal-width or equal-frequency. Then the values in each bin are smoothened using bin mean or bin median.
Binning is unsupervised discretization as it does not use any class information.

b) HISTOGRAM ANALYSIS : Histogram distibute the observed values of an attribute into disjoint subset which is also termed as buckets or bins. This is also unsupervised discretization

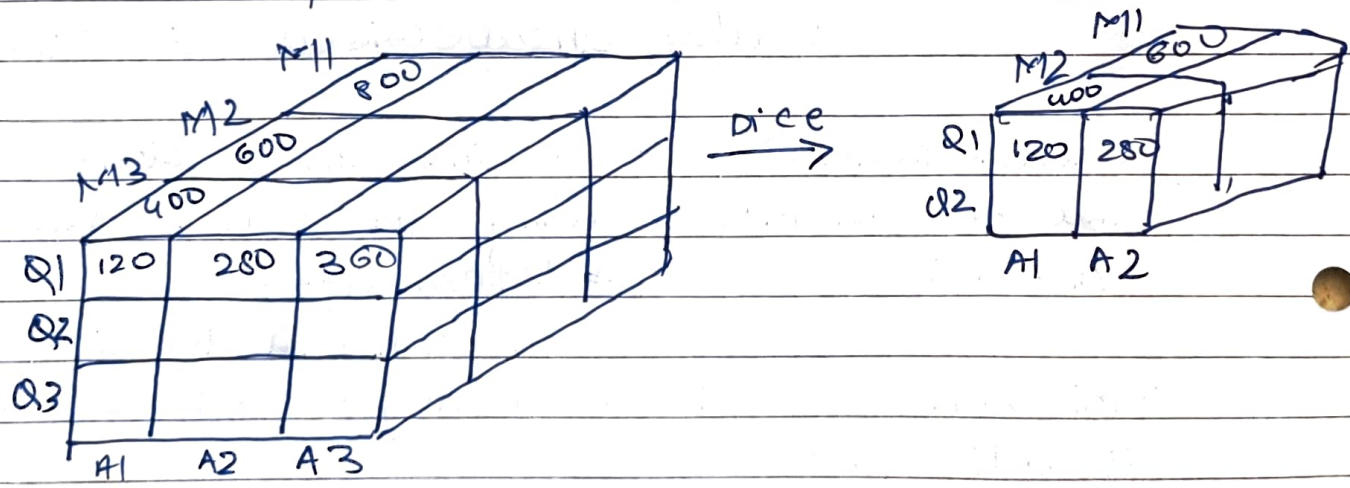
c) CLUSTER, DECISION TREE AND CORRELATION ANALYSIS : Clustering can be done either using top-down strategy or bottom-up.
Generating a decision tree implements top-down strategy and is supervised discretization.
Correlation analysis uses bottom-up merging strategy.

Q2 b] Various OLAP operations are:

(i) SLICE : In this, we select one particular dimension and provide a new sub-cube




(ii) DICE: In dice, we select two or more dimensions a provide a new sub-cube.



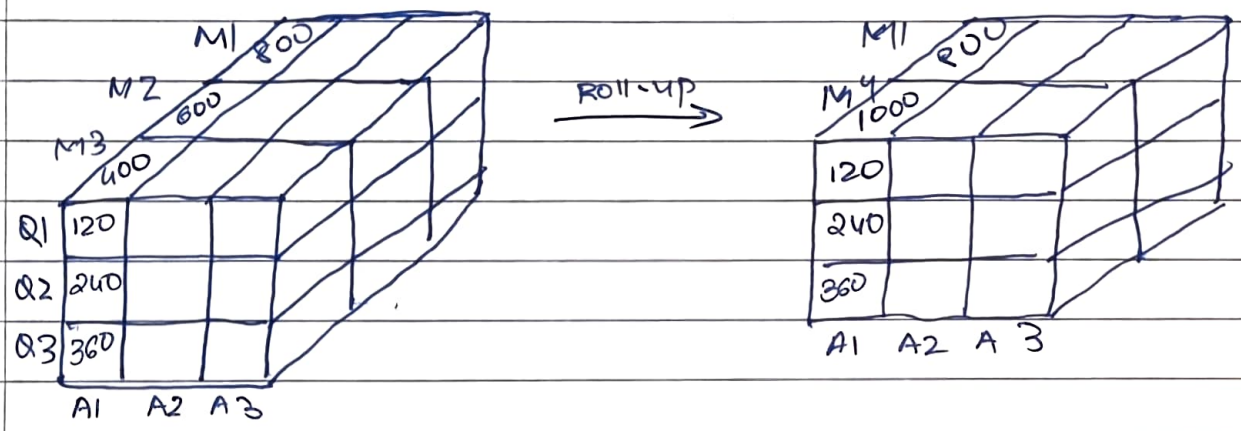

$M_i$ = Medine i
$Q_i$ = Quarter i
$A_i$ = Area i

(iii) ROLL-UP :

In roll-up, we perform data-cube aggregation by :-

→ Dimension reduction

→ Moving up the concept hierarchy

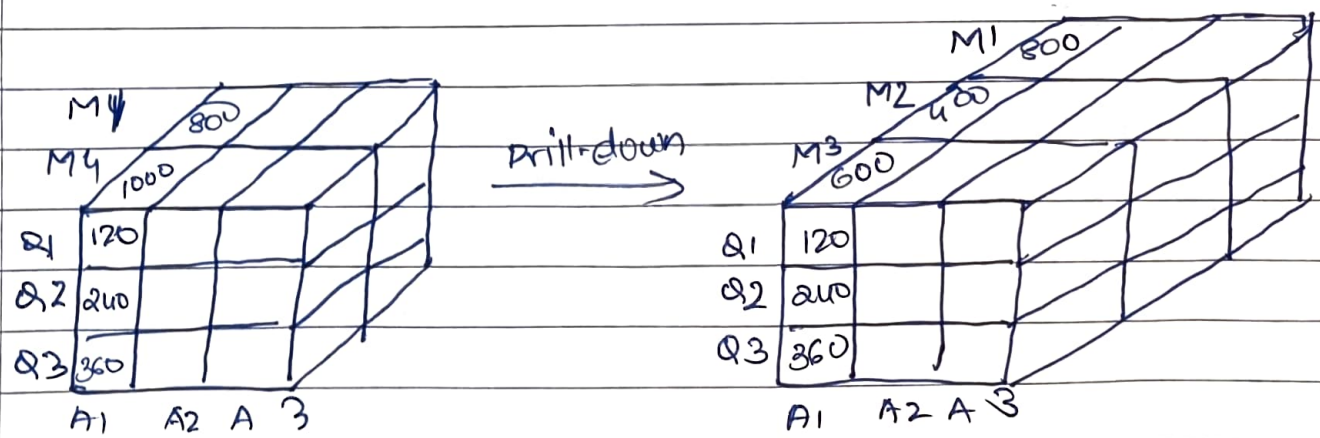



(iv) DRILL-DOWN :

In drill-down, we perform reverse of roll-up :

→ Moving down concept hierarchy

→ Introducing new dimens.



6000419005 7

Q3 a&b]

ANS Techniques to improve classification of an algorithm:

1 ADD MORE DATA :
Having more data allows the data to tell for itself instead of relying on assumptions and weak correlations. Presence of move data results in bettel and accurate models.

2 TREAT MISSING AND OUTLIER VALUES :
The unwanted presence of missing and outlier values in the training data often reduces the accuracy of a model and leads to a biased model which does inaccurate predictions.

3. FEATURE ENGINEERING :
This helps to extract more information from existing data which is extracted in terms of new features which may have a higher ability to explain the variance in the training data. thus, giving improved model accuracy

4 FEATURE SELECTION
Feature selection is a process of finding out the best subset of attributes which bettel explains the relationship of independent variables with taget variable.

5 MULTIPLE ALGORITHMS

Hitting at the right machine learning algorithm is the ideal approach to achieve higher accuracy but this is difficult. some algorithms are better suited to a particular type of data set than others. Hence we should try and compare all relevant models

6 ENSEMBLE METHODS

This is a common technique which simply combines the result of multiple weak models and produce better results. This can be achieved through two ways :-

- Bagging
- Boosting

~~Q3) Metadata can be~~

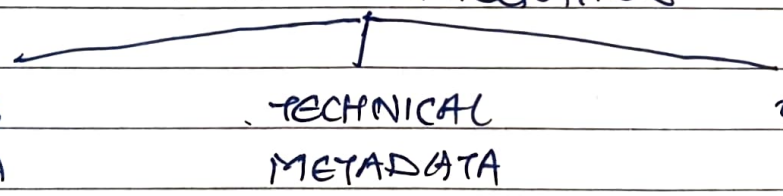
Q3 b]

ANS Metadata can be broadly categorized into three categories

- BUSINESS METADATA :
  It has the data ownership information, business defination, and changing policies

- TECHNICAL METADATA :
  It includes database system names, table and column names and sizes, data types and allowed values. It also includes structural information such as primary and foreign key attributes and indices.

- OPERATIONAL METADATA :
  It includes currency of data and data lineage. Currency of data means whether the data is active, archived or purged. Lineage of data means - the history of data migrated and transformation is applied on it.



60004190057

Qu a]

| Transaction ID | Items |
|---|---|
| 100 | B, A, T |
| 200 | A, C |
| 300 | A, S |
| 400 | B, A, C |
| 500 | B, S |
| 600 | B, S |
| 700 | B, S, A, T |
| 800 | B, S |
| 900 | B, A, S |

ANS

| ITEMS | FREQUENCY |
|---|---|
| B | 7 |
| A | 6 |
| T | 2 |
| C | 2 |
| S | 6 |

∴ Sorted manner.

∴

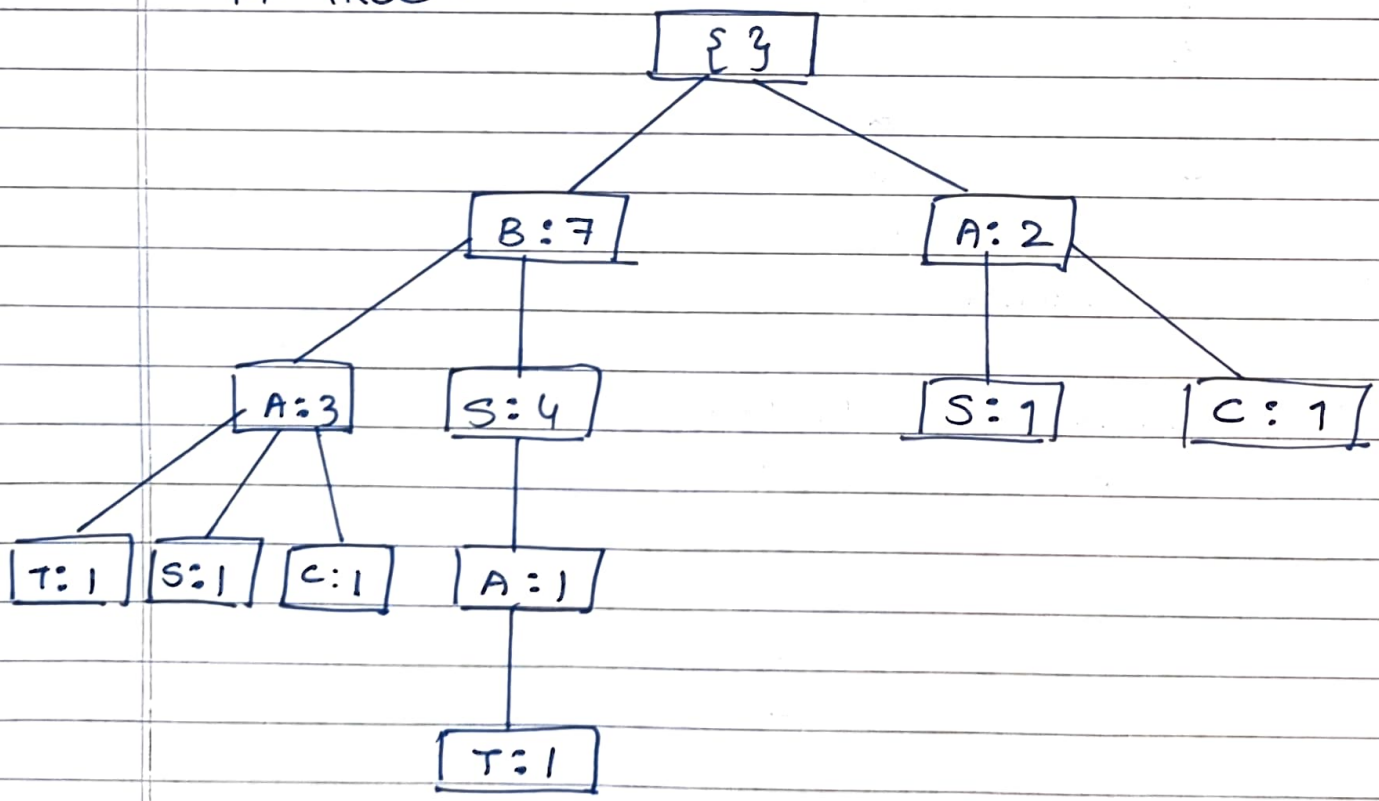
| ITEMSET | FREQUENCY |
|---|---|
| B | 7 |
| A | 6 |
| S | 6 |
| T | 2 |
| C | 2 |

6000419005 7

∴

| Transaction-ID | Itemset |
|---|---|
| 100 | B, A, T |
| 200 | A, C |
| 300 | A, S |
| 400 | B, A, C |
| 500 | B, S |
| 600 | B, S |
| 700 | B, S, A, T |
| 800 | B, S |
| 900 | B, A, S |

∴ FP TREE :



6000U1900S

Q4 b

ANS Steps involved in knowledge discovery process :-

- DATA CLEANING : In this step, the noise and inconsistent data is removed
- DATA INTEGRATION : In this step, multiple data sources are combined
- DATA SELECTION : In this step, data relevent to the analysis task are retrieved from the database
- DATA TRANSFORMATION : In this step, data is transformed or consolidated into forms appropriate for mining by peyorming summary or aggregation operations
- DATA MINING : In this step, intelligent methods are applied in order to extract data patterns
- PATTERN EVALUATION : In this step, data patterns are evaluated.
- KNOWLEDGE PRESENTATION : In this step, knowledge is represented.

8

6000419005?

Q5 a]

| Transaction ID | Items |
|---|---|
| t1 | 1, 3, 4 |
| t2 | 2, 3, 5 |
| t3 | 1, 2, 3, 5 |
| t4 | 2, 5 |
| t5 | 1, 2, 3, 5 |
| ~~t6~~ | |

APRIORI ALGORITHM:

Minimum support = 40 %

$40\% = \frac{x \times 100}{5}$

$\therefore x = 2$ = Minimum support.

Minimum confidence = 70%

C1

| ITEM | COUNT |
|---|---|
| 1 | 3 |
| 2 | 4 |
| 3 | 4 |
| 4 | 1 |
| 5 | 4 |

L1

| ITEM | COUNT |
|---|---|
| 1 | 3 |
| 2 | 4 |
| 3 | 4 |
| 5 | 4 |

C2:

| ITEM | COUNT |
|---|---|
| {1, 2} | 2 |
| {1, 3} | 3 |
| {1, 5} | 2 |
| {2, 3} | 3 |
| {2, 5} | 4 |
| {3, 5} | 3 |

L2:

| ITEM | COUNT |
|---|---|
| {1, 2} | 2 |
| {1, 3} | 3 |
| {1, 5} | 2 |
| {2, 3} | 3 |
| {2, 5} | 4 |
| {3, 5} | 3 |

G0004190057

C3:

| ITEM | COUNT |
|---|---|
| {1, 2, 3} | 2 |
| {1, 2, 5} | 2 |
| {2, 3, 5} | 3 |
| {1, 3, 5} | 2 |

L3:

| ITEM | COUNT |
|---|---|
| {1, 2, 3} | 2 |
| {1, 2, 5} | 2 |
| {2, 3, 5} | 3 |
| {1, 3, 5} | 2 |

∴ Frequent Itemsets : {1, 2, 3}
{1, 2, 5}
{2, 3, 5}
{1, 3, 5}

60004190057

Q5 b

| ANS | ROLAP | MOLAP | HOLAP |
|---|---|---|---|
| 1. | Relational database is used as storage location for summary aggregation | Multidimensional database is used as storage location for summary aggregation | Multidimentional database is used as storage location for summary aggregation |
| 2. | Processing time is very slow | Processing time is fast | Processing time is fast |
| 3. | Large storage space requirement as compared to MOLAP and HOLAP | Medium space required for storage as compared to ROLAP and HOLAP | small storage space requirement as compared to MOLAP and ROLAP |
| 4. | Low latency as compared to MOLAP and HOLAP | High latency as compared to ROLAP and HOLAP | Medium latency as compared to MOLAP and ROLAP |