Hall Ticket Number: **18CSD12**

III/IV B.Tech (Regular\Supplementary) DEGREE EXAMINATION

February, 2021 **Computer Science and Engineering**

Fifth Semester **Data ware Housing and Data Mining**

Time: Three Hours **Maximum:** 50 Marks

Answer Question No.1 compulsorily. (1X10 = 10 Marks)

Answer any one question from each unit. (4X10=40 Marks)

1. Answer all questions (1X10=10 Marks)
   a) Define Data Warehouse.
   b) How to deal with missing values in an attribute?
   c) What is the importance of a fact table?
   d) What is Data Mining?
   e) How to compute confidence measure for an association rule?
   f) What is Data Mart?
   g) Differentiate qualitative and quantitative attributes.
   h) Define cluster analysis?
   i) What are properties of good Clustering?
   j) Define outlier?

UNIT I

2. a) What are Data Mining functionalities? Explain briefly. 5M
   b) What is Data Cleaning? Explain various data cleaning tasks. 5M

(OR)

3. a) Discuss various issues in Data Mining. 5M
   b) Illustrate the Data Transformation by Normalization. 5M

UNIT II

4. a) Explain Data Warehouse architecture. 5M
   b) What is the difference between operational DBMS and Data Warehouse 5M

(OR)

5. a) Discuss the star and snowflake schema in detail with suitable example. 5M
   b) Explain various OLAP operations. 5M

UNIT III

6. a) A database has six transactions. Let min-sup = 50% and min-conf = 75%. 5M

TID	List of items
001	Pencil, sharpener, eraser, color papers
002	Color papers, charts, glue sticks
003	Pencil, glue stick, eraser, pen
004	Oil pastels, poster colours, correction tape
005	Whitener, pen , pencil, charts, glue stick
006	Colour pencils, crayons, eraser, pen

Find all frequent item sets using Apriori algorithm. List all the strong association rules.

   b) Write the advantages and disadvantages of Apriori and FP-growth Algorithm 5M

(OR)

7. a) What are different methods to improve Apriori algorithm's efficiency? 5M
   b) Explain constraint based rule mining. 5M

UNIT IV

8. a) What is the goal of clustering? How does partitioning around medoids algorithm achieve this goal? 5M
   b) Write K-means clustering algorithm. 5M

(OR)

9. Explain Hierarchical clustering algorithm. 10M

Hall Ticket Number: **18CSD12**

III/IV B.Tech (Regular\Supplementary) DEGREE EXAMINATION

February, 2021 **Computer Science and Engineering**

Fifth Semester **Data ware Housing and Data Mining**

Time: Three Hours **Maximum:** 50 Marks

Answer Question No.1 compulsorily. (1X10 = 10 Marks)

Answer any one question from each unit. (4X10=40 Marks)

1.	Answer all questions		(1X10=10 Marks)
	a)	Define Data Warehouse. A data warehouse is a subject-oriented, integrated, time-variant, and Non-volatile collection of data in support of management's decision making process.	
	b)	How to deal with missing values in an attribute? If written any 3 of the below, award 1M 1) Ignore the tuple. 2) Fill in the missing value manually. 3) Use a global constant to fill in the missing value. 4) Use the attribute mean to fill in the missing value. 5) Use the attribute mean for all samples belonging to the same class as the given tuple. 6) Use the most probable value to fill in the missing value	
	c)	What is the importance of a fact table? A fact table consists of two types of columns. The foreign keys column which allows joining with dimension tables and the measure columns contain the data that is being analyzed.	
	d)	What is Data Mining? Data mining is a process of discovering patterns in large data sets	
	e)	How to compute confidence measure for an association rule? Confidence (A=>B) = (Number of transactions includes both A and B)/ (Number of transactions includes only product A)	
	f)	What is Data Mart? A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), such as Sales or Finance or Marketing.	
	g)	Differentiate qualitative and quantitative attributes. **Quantitative Attributes** - Attributes whose values result from counting or measuring something. Examples: height, weight, time in the 100 yard dash, number of items sold to a shopper **Qualitative Attributes** - Attributes that are not measurement variables. Their values do not result from measuring or counting. Examples: hair color, religion, political party, profession	
	h)	Define cluster analysis? **Cluster Analysis** means that to find out the group of objects which are similar to each other in the group but are different from the object in other groups.	
	i)	What are properties of good Clustering? A good clustering method will produce high quality clusters with • high intra-class similarity • low inter-class similarity	
	j)	Define outlier? An outlier is a data point that differs significantly from other observations.	
		UNIT I	
2.	a)	What are Data Mining functionalities? Explain briefly. **Data Mining Functionalities** Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.Data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make	5M

predictions.

Concept/Class Description: Characterization and Discrimination

Data can be associated with classes or concepts. For example, in the Electronics store, classes of items for sale include computers and printers, and concepts of customers include bigSpenders and budgetSpenders.

Data characterization

Data characterization is a summarization of the general characteristics or features of a target class of data.

Data discrimination

Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.

Mining Frequent Patterns, Associations, and Correlations

Frequent patterns, are patterns that occur frequently in data. There are many kinds of frequent patterns, including itemsets, subsequences, and substructures.

Association analysis

Suppose, as a marketing manager, you would like to determine which items are frequently purchased together within the same transactions.

buys(X,"computer")=buys(X,"software") [support=1%,confidence=50%]

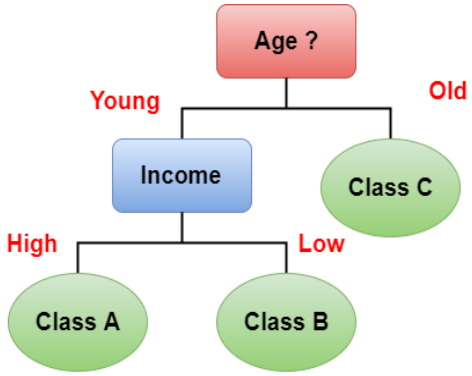
where X is a variable representing a customer.Confidence=50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. Support=1% means that 1% of all of the transactions under analysis showed that computer and software were purchased together.

Classification and Prediction

Classification is the process of finding a model that describes and distinguishes data classes for the purpose of being able to use the model to predict the class of objects whose class label is unknown.

"How is the derived model presented?" The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks.

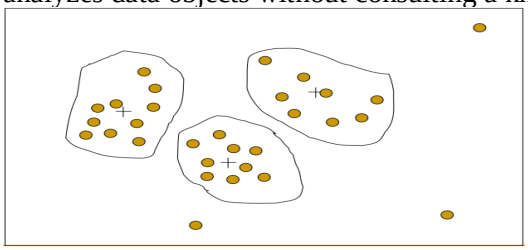
A **decision tree** is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions.



Decision tree

Cluster Analysis

In classification and prediction analyze class-labeled data objects, where as clustering analyzes data objects without consulting a known class label.



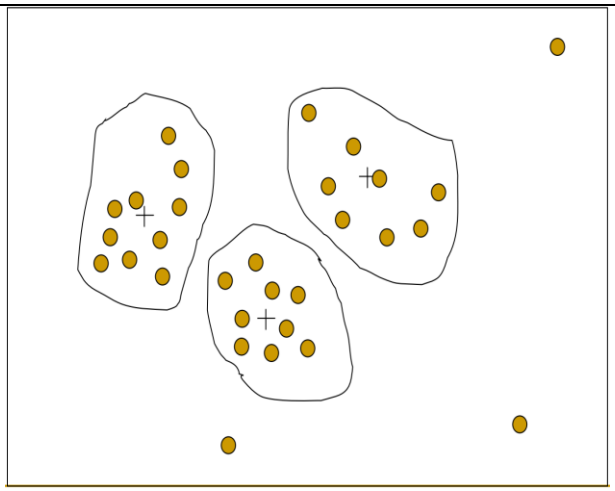
Cluster Analysis

The objects are grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

Outlier Analysis

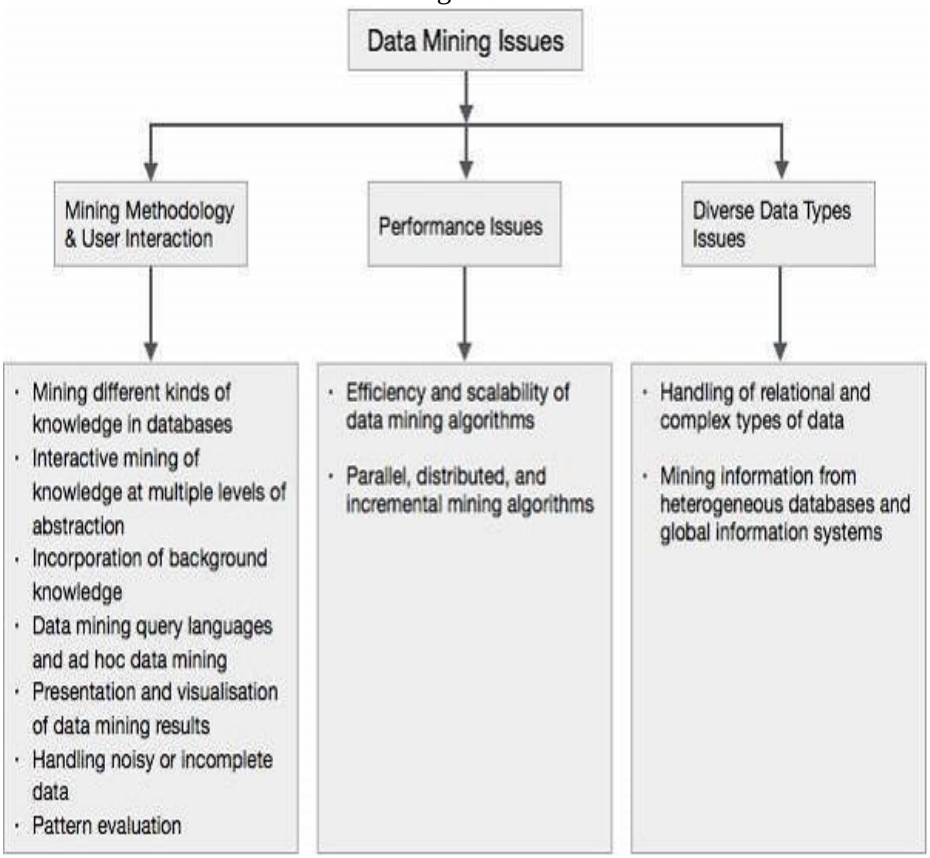
A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard

		outliers as noise or exceptions. The analysis of outlier data is referred to as outlier mining.	
	b)	What is Data Cleaning? Explain various data cleaning tasks. **Data Cleaning in Data Mining** Quality of your data is critical in getting to final analysis.Any data which tend to be incomplete, noisy and inconsistent can affect your result. Data cleaning in data mining is the process of detecting and removing corrupt or inaccurate records from a record set, table or database. Some data cleaning methods :- 1 You can ignore the tuple.This is done when class label is missing.This method is not very effective , unless the tuple contains several attributes with missing values. 2 You can fill in the missing value manually.This approach is effective on small data set with some missing values. 3 You can replace all missing attribute values with global constant, such as a label like "Unknown" or minus infinity. 4 You can use the attribute mean to fill in the missing value.For example customer average income is 25000 then you can use this value to replace missing value for income. 5 Use the most probable value to fill in the missing value. **Noisy Data** Noise is a random error or variance in a measured variable. Noisy Data may be due to faulty data collection instruments, data entry problems and technology limitation. How to Handle Noisy Data? **Binning:** Binning methods sorted data value by consulting its "neighbor- hood," that is, the values around it.The sorted values are distributed into a number of "buckets," or bins. For example Price = 4, 8, 15, 21, 21, 24, 25, 28, 34 **Partition into (equal-frequency) bins:** Bin a: 4, 8, 15 Bin b: 21, 21, 24 Bin c: 25, 28, 34 In this example, the data for price are first sorted and then partitioned into equal-frequency bins of size 3. **Smoothing by bin means:** Bin a: 9, 9, 9 Bin b: 22, 22, 22 Bin c: 29, 29, 29 In smoothing by bin means, each value in a bin is replaced by the mean value of the bin. **Smoothing by bin boundaries:** Bin a: 4, 4, 15 Bin b: 21, 21, 24 Bin c: 25, 25, 34 In smoothing by bin boundaries, each bin value is replaced by the closest boundary value. **Regression** Data can be smoothed by fitting the data into a regression functions. **Clustering:** Outliers may be detected by clustering, where similar values are organized into groups, or "clusters. Values that fall outside of the set of clusters may be considered outliers.	5M



(OR)

3. a) Discuss various issues in Data Mining. 5M



Mining Methodology and User Interaction Issues

It refers to the following kinds of issues −

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.

- **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

There can be performance-related issues such as follows −

- **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data** − The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

b) Illustrate the Data Transformation by Normalization. 5M

The data transformation is a basic element of data mining. It means transforming the data, namely converting the source data in to another format that allows processing data effectively. The main purpose of data normalization is to minimize or even exclude duplicated data. This is a very essential and important issue because it is increasingly problematic to keep in data in relational databases, which store identical data in more than one place.

Min-Max normalization

The first technique we will cover is min-max normalization. It is the linear transformation of the original unstructured data. It scales the data from 0 to 1. It is calculated by the following formula:

$$v' = \frac{v - \text{minF}}{\text{maxF} - \text{minF}}(new\_max_F - new\_min_F) + new\_min_F ,$$

where is the current value of feature *F*.

Let us consider one example to make the calculation method clear. Assume that the minimum and maximum values for the feature F are \$50,000 and \$100,000 correspondingly. It needs to range *F* from 0 to 1. In accordance with min-max normalization, *v* = \$80,000 is transformed to:

$$v' = \frac{80{,}000 - 50{,}000}{100{,}000 - 50{,}000} + (1 - 0) + 0 = \frac{3}{5} = 0{,}6$$

As you can see this technique enables to interpret the data easily. There are no large numbers, only concise data that do not require further transformation and can be used in decision-making process immediately.

Z-score normalization

The next technique is z-score normalization. It is also called zero-mean normalization. The essence of this technique is the data transformation by the values conversation to a common scale where an average number equals zero and a standard deviation is one. A value is normalized to ′ under the formula:

$$v' = \frac{v - \bar{F}}{\sigma_F},$$

Here is the mean and is the standard deviation of feature *F*.

Here is an example of the calculation of a value.

On the supposition that the mean of feature is \$65,000 and its standard deviation is \$ 18,000. Applying the z-score normalization we get the following mean of the value equals to \$85,800:

$$\frac{85{,}800-65{,}000}{18{,}000} = 1{,}156.$$

Therefore, z-score normalization is an approach to the data processing that takes into account outliers; however, it does not provide data normalization with an identical scale.

Data normalization by decimal scaling

And now we finally will move on to the decimal scaling normalization technique. It involves the data transformation by dragging the decimal points of values of feature *F*. The movement of decimals is very dependent on the absolute value of the maximum. A value of feature *F* is transformed to by calculating:

$$v' = \frac{v}{10^j}.$$

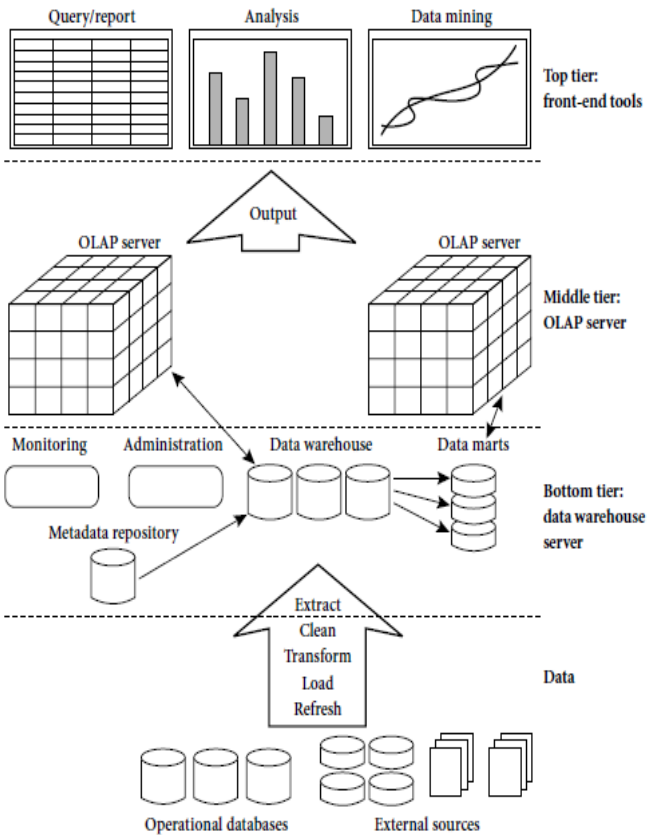
In this formula, *j* is the lowest integer while *Max* (|) < 1.

For instance, values for feature *F* scale from 850 to 825. Suppose *j* equals to three. In this case, the maximum absolute value of feature *F* equals 850. Applying the normalization with decimal scaling, we need to divide all values by 1,000. Therefore, we get 850 normalized to 0,850 as well as 825 transformed to 0,825.

This technique entails the transformation of the decimal points of the values according to the absolute value of the maximum. It follows that the means of the normalized data will always be between 0 and 1.

UNIT II

4. a) Explain Data Warehouse architecture. 5M



Data warehouses often adopt three-tier architecture**.**

1.The bottom tier is a warehouse database server that is almost always a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (such as customer profile information provided by external consultants). These tools and utilities perform data extraction, cleaning, and transformation (e.g., to merge similar data from different sources into a unified format), as well as load and refresh functions to update the data warehouse. The data are extracted using application program interfaces known as gateways. A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server. Examples of gateways include ODBC (Open Database Connection) and OLEDB (Open Linking and Embedding for Databases) by Microsoft and JDBC (Java Database Connection). This tier also contains a metadata

repository, which stores information about the data warehouse and its contents.
2. The middle tier is an OLAP server that is typically implemented using either (1) a relational OLAP (ROLAP) model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations; or (2) a multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations
3. The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).

b) What is the difference between operational DBMS and Data Warehouse — 5M

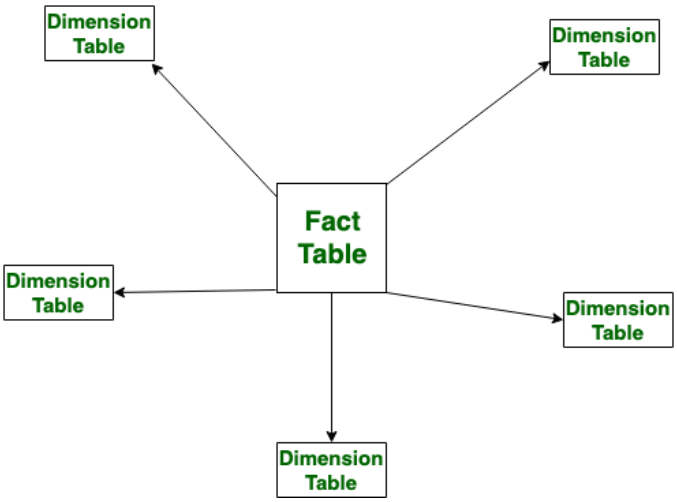
Operational Database Systems	Data Warehouses
Operational systems are generally designed to support high-volume transaction processing.	Data warehousing systems are generally designed to support high-volume analytical processing. (i.e. OLAP).
Operational systems focuses on Data in.	Data warehousing systems focuses on Information out.
In Operational systems data is stored with a functional or process orientation.	In Data warehousing systems data is stored with a subject orientation.
Performance is low for analysis queries.	Performance is high for analysis queries.
It is used for Online Transactional Processing (OLTP)	It is used for Online Analytical Processing (OLAP).
Operational systems represent current transactions.	Data warehousing systems reads the historical data.
Data within operational systems are generally updated regularly.	Data within a data warehouse is non-volatile, meaning when new data is added old data is not erased so rarely updates.
Complex data structures.	Multi dimensional data structures.

(OR)

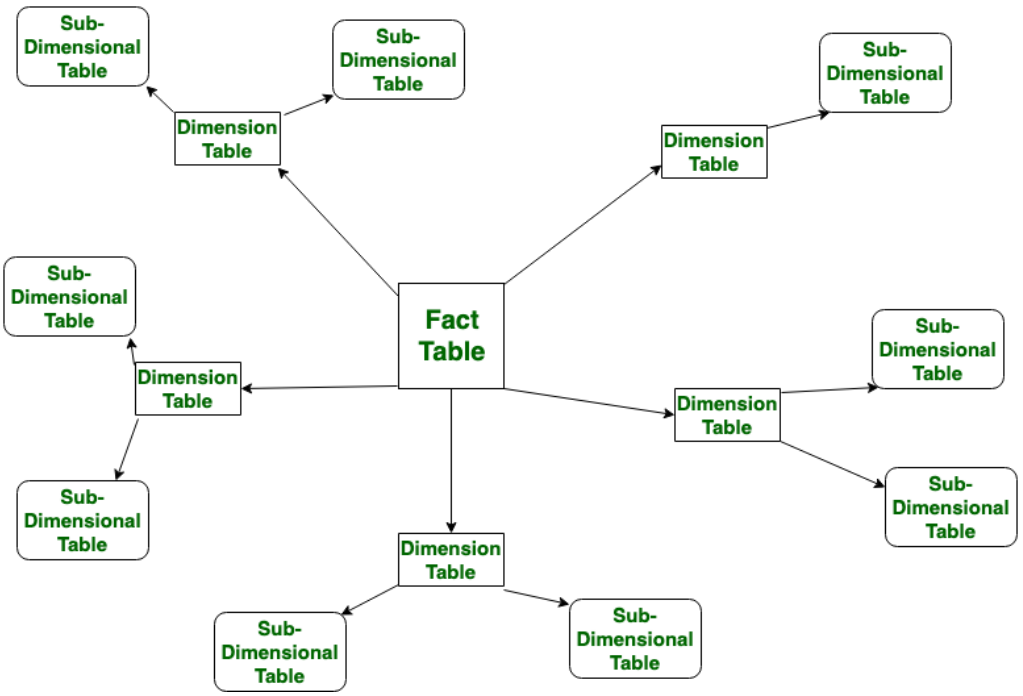
5. a) Discuss the star and snowflake schema in detail with suitable example. — 5M

Star Schema:

Star schema is the type of multidimensional model which is used for data warehouse. In star schema, The fact tables and the dimension tables are contained. In this schema fewer foreign-key join is used. This schema forms a star with fact table and dimension tables.

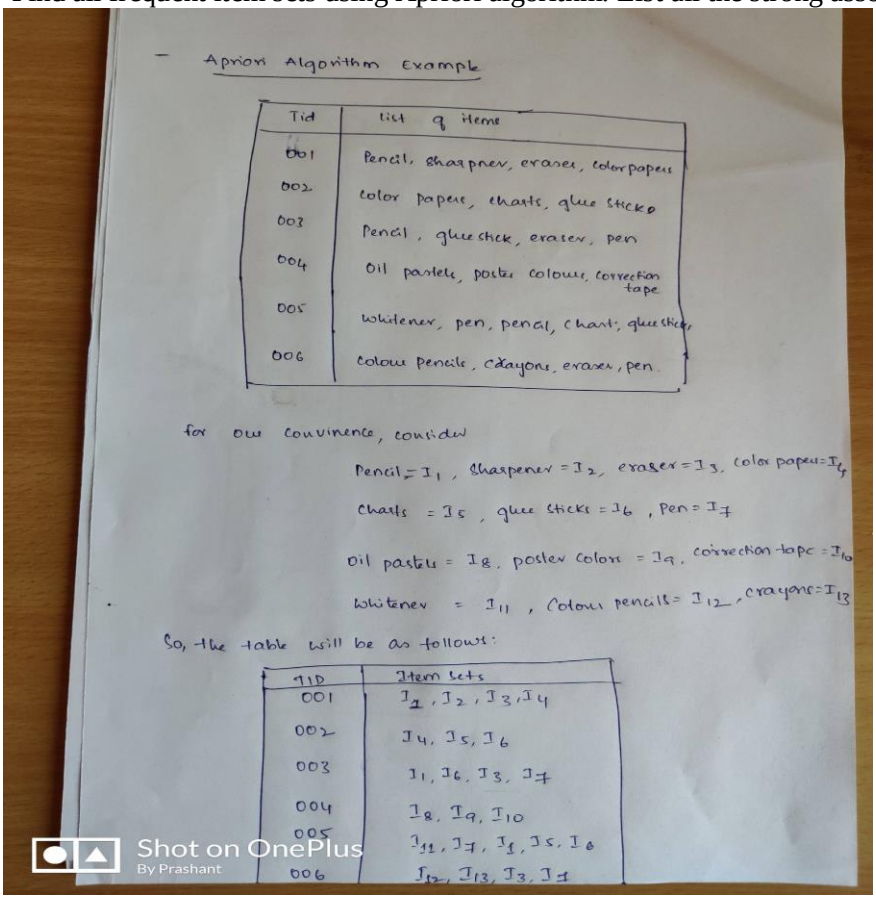


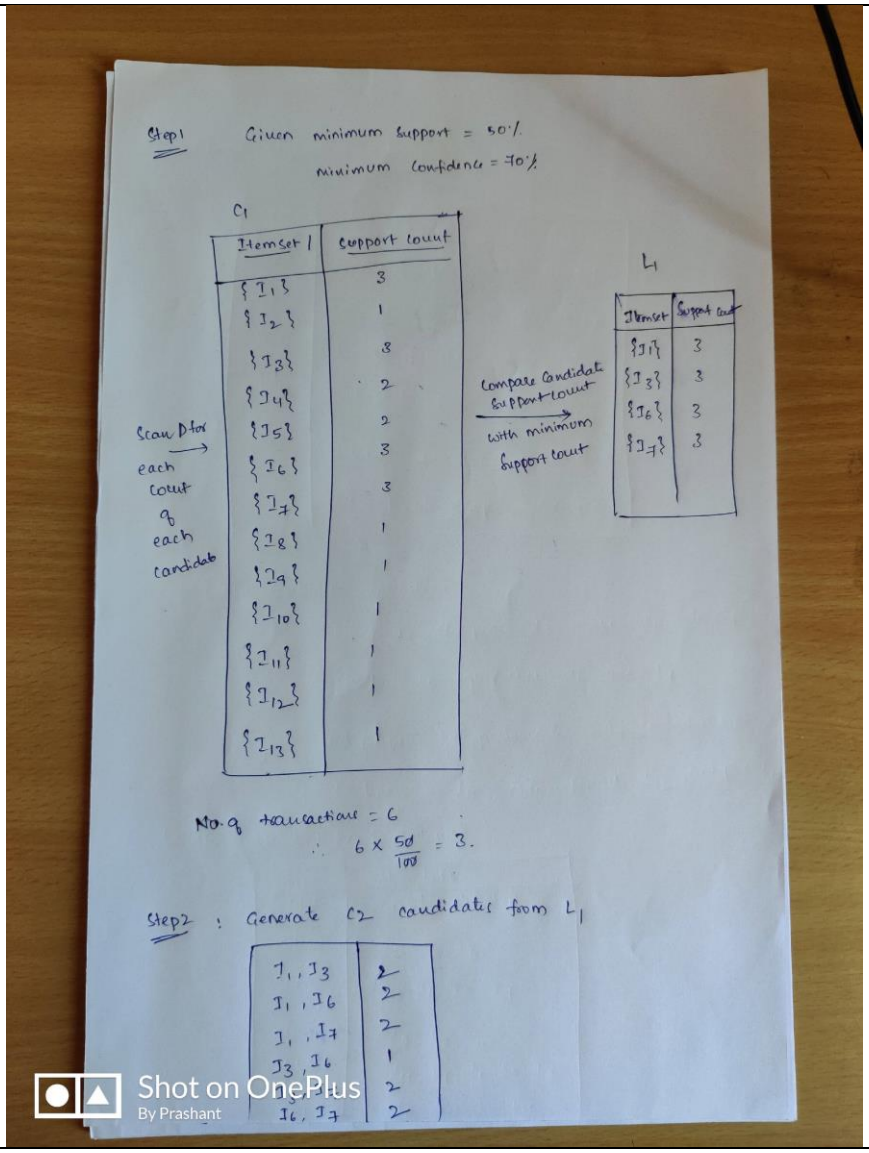
Snowflake Schema:

		Snowflake Schema is also the type of multidimensional model which is used for data warehouse. In snowflake schema, The fact tables, dimension tables as well as sub dimension tables are contained. This schema forms a snowflake with fact tables, dimension tables as well as sub-dimension tables.  Any related examples:	
	b)	Explain various OLAP operations. Four types of analytical operations in OLAP are: 1. Roll-up 2. Drill-down 3. Slice and dice 4. Pivot (rotate) **1) Roll-up:** Roll-up is also known as "consolidation" or "aggregation." The Roll-up operation can be performed in 2 ways 1. Reducing dimensions 2. Climbing up concept hierarchy. Concept hierarchy is a system of grouping things based on their order or level. **2) Drill-down** In drill-down data is fragmented into smaller parts. It is the opposite of the rollup process. It can be done via • Moving down the concept hierarchy • Increasing a dimension **3) Slice:** Here, one dimension is selected, and a new sub-cube is created. **Dice:** This operation is similar to a slice. The difference in dice is you select 2 or more dimensions that result in the creation of a sub-cube. Explanation with relevant example required.	5M
		UNIT III	
6.	a)	A database has six transactions. Let min-sup = 50% and min-conf = 75%.	5M

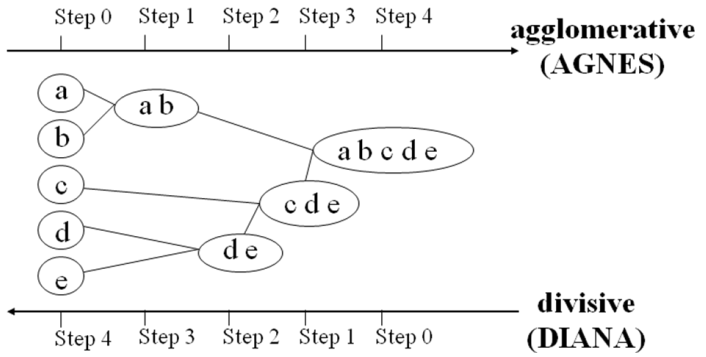
TID	List of items
001	Pencil, sharpener, eraser, color papers
002	Color papers, charts, glue sticks
003	Pencil, glue stick, eraser, pen
004	Oil pastels, poster colours, correction tape
005	Whitener, pen , pencil, charts, glue stick
006	Colour pencils, crayons, eraser, pen

Find all frequent item sets using Apriori algorithm. List all the strong association rules.



			
	b)	Write the advantages and disadvantages of Apriori and FP-growth Algorithm Apriori Algorithm Advantages: • Easy to understand algorithm • Join and Prune steps are easy to implement on large itemsets in large databases Disadvantages: • It requires high computation if the itemsets are very large and the minimum support is kept very low. • The entire database needs to be scanned. FP-Growth Algorithm Advantages: • Faster than Apriori algorithm. • No candidate generation. • Only two passes over dataset. Disadvantages of FP growth algorithm: • FP tree may not fit in memory. • FP tree is expensive to build.	5M
		(OR)	
7.	a)	What are different methods to improve Apriori algorithm's efficiency? • Hash-based itemset counting: A *k*-itemset whose corresponding hashing bucket count is below the threshold cannot be frequent • Transaction reduction: A transaction that does not contain any frequent k-itemset is useless in subsequent scans • Partitioning: Any itemset that is potentially frequent in DB must be frequent in at least one of the partitions of DB • Sampling: mining on a subset of given data, lower support threshold + a method to	5M

		determine the completeness • Dynamic itemset counting: add new candidate itemsets only when all of their subsets are estimated to be frequent	
	b)	Explain constraint based rule mining. • A data mining process may uncover thousands of rules from a given set of data, most of which end up being unrelated or uninteresting to the users. • Often, users have a good sense of which "direction" of mining may lead to interesting patterns and the "form" of the patterns or rules they would like to find. • Thus, a good heuristic is to have the users specify such intuition or expectations as constraints to confine the search space. • This strategy is known as constraint-based mining. • Constraint based mining provides o **User Flexibility:** provides constraints on what to be mined. o **System Optimization:** explores constraints to help efficient mining. • The constraints can include the following: • **Knowledge type constraints:** These specify the type of knowledge to be mined, such as association or correlation. • **Data constraints**:These specify the set of task-relevant data. o **Dimension/level constraints:** These specify the desired dimensions (or attributes) of the data, or levels of the concept hierarchies, to be used in mining. • **Interestingness constraints:** These specify thresholds on statistical measures of rule interestingness, such as support, confidence, and correlation. o **Rule constraints:** These specify the form of rules to be mined. Such constraints may be expressed as rule templates, as the maximum or minimum number of predicates that can occur in the rule antecedent or consequent, or as relationships among attributes, attribute values, and/or aggregates. The above constraints can be specified using a high-level declarative data mining query language and user interface. **Constraint based association rules:** - In order to make the mining process more efficient rule based constraint mining allows users to describe the rules that they would like to uncover. - provides a sophisticated mining query optimizer that can be used to exploit the constraints specified by the user. - encourages interactive exploratory mining and analysis.	5M
		UNIT IV	
8.	a)	What is the goal of clustering? How does partitioning around medoids algorithm achieve this goal? The goal of clustering is to identify distinct groups in a dataset. PAM algorithm identifies distinct groups in a dataset using mediods. • It is used Find *representative* objects, called medoids, in clusters • *PAM* (Partitioning Around Medoids) – starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering – *PAM* works effectively for small data sets, but does not scale well for large data sets • *CLARA* • *CLARANS* :Randomized sampling The above said are the three algorithms which are used in PAM. Focusing + spatial data structure (Ester et al., 1995)	5M
	b)	Write K-means clustering algorithm. Given *k*, the *k-means* algorithm is implemented in 4 steps: – Partition objects into *k* nonempty subsets – Compute seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster. – Assign each object to the cluster with the nearest seed point. – Go back to Step 2, stop when no more new assignment. With relevant example	5M
		(OR)	
9.		Explain Hierarchical clustering algorithm.	10M



- Use distance matrix as clustering criteria. This method does not require the number of clusters ***k*** as an input, but needs a termination condition

AGNES:

- Implemented in statistical analysis packages, e.g., Splus
- Use the Single-Link method and the dissimilarity matrix.
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster

DIANA:

- Implemented in statistical analysis packages, e.g., Splus
- Inverse order of AGNES
- Eventually each node forms a cluster on its own

With relevant example