Course Code : CST 406 **IRJQ/MS – 19 /8619**

Eighth Semester B. E. (Computer Science and Engineering) Examination

DATA WAREHOUSING AND MINING

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

(1) Number your answers properly.

(2) Assume suitable data and illustrate answers with neat sketches wherever necessary.

1. (a) Write note on : metadata repository and data cube materialization. 5(CO1)

(b) Write a short note on data warehouse development life cycle. 5(CO1)

2. (a) What is the purpose of 'refresh' in ETL process ? When should we refresh ? What are the different refresh techniques ? 5(CO2)

(b) Suppose that a data warehouse consists of the following four dimensions : student, course, semester, and instructor, and two measures count and avg grade. When at the lowest conceptual level (e. g. for a given student, course, semester, and 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 (e. g., roll – up from semester to year) should one perform in order to list the average grade of CS courses for each Big University student. 5(CO2)

3. (a) Create a Index Organized Table with overflow area. Give command to insert two sample rows in this table. What are the pros and cons of an index organized table over a heap organized table. 6(CO1)

(b) Explain the need to create function based indexes. Write a command to create a function based index on emp_name column of EMPLOYEE table. 4(CO1)

4. (a) 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 , 20 , 21 , 22 , 22 , 25 , 25 , 25 , 25 , 30 , 33 , 33 , 35 , 35 , 35 , 35 , 36 , 40 , 45 , 46 , 52 , 70.

(i) What is the mean of the data ? What is the median ?

(ii) What is the mode of the data ? Comment on the data's modality.

(iii) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data ?

(iv) Give the five – number summary of the data.

(v) Show a boxplot of the data. 5(CO3)

(b) Consider the following snapshot of SALES table :—
Extract of Sales Data

Address or Rowid	**Date**	**Product**	**Color**	**Region**	**Sale ($)**
00001 BFE. 0012.0111	15 – Nov – 00	Dishwasher	White	East	300
00001 BFE. 0013.0114	15 – Nov – 00	Dryer	Almond	West	450
00001 BFF. 0012.0115	16 – Nov – 00	Dishwasher	Almond	West	350
00001 BFF. 0012.0138	16 – Nov – 00	Washer	Black	North	550
00001 BFF. 0012.0145	17 – Nov – 00	Washer	White	South	500
00001 BFF. 00.12.0.157	17 – Nov – 00	Dryer	White	East	400
00001 BFF. 0014.0165	17 – Nov – 00	Washer	Almond	South	575

Explain how the query : Select the rows from the Sales table where product is "Washer" and color is "Almond" and division is "East" or "South" will be executed if bitmap indexes are created on Product, Color, and Region columns. Show the intermediate steps. 5(CO3)

5. (a) Explain with the help of a neat diagram knowledge discovery process in data mining. 3(CO4)

(b) A database has five transactions. Let min_sup = 60% and min_conf = 80%.

TID	*items_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 }

(i) Find all frequent item sets using Apriori and FP – growth, respectively. Compare the efficiency of the two mining processes.

(ii) List all of the strong association rules (with support s and confidence c) matching the following metarule, where X is a variable representing customers, and item_i denotes variables representing items (e. g. "A" , "B" etc.) :

$\forall_x \in transaction\ ,\ buys\ (X\ ,\ item_1)\ \wedge\ buys\ (X\ ,\ item_2) \Longrightarrow buys\ (X\ ,\ item_3)\ [s,c]$

7(CO4)

6. (a) Describe each of the clustering algorithm in terms of the following criterion :—

(1) Shape of the cluster that can be determined

(2) Input parameter that must be specified

(3) Limitations

(4) Time complexity of the algorithm

(i) CLARA and (ii) BIRCH 4(CO4)

(b) Suppose that the data mining task is to cluster points (with (x , y) representing location) into three clusters, where the points are A1 (2 , 10) , A2 (2 , 5) , A3 (8 , 4) , B1 (5 , 8) , B2 (7 , 5) , B3 (6 , 4) , C1 (1 , 2) , C2 (4 , 0). The distance function is Euclidean distance. Suppose initially we assign A1 , B1 and C1 as the center of cluster, respectively. Use the K – means algorithm to show only the three culster center after the second round of execution.

6(CO4)