Course Code : CST 411 **MQNR/MW – 19 / 9659**

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

DATA WAREHOUSING AND MINING

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

(1) Number your answer properly.
(2) Assume suitable data and illustrate answers with neat sketches wherever necessary.
(3) Plot neat graphs on graph papers.

1. (a) Describe how the top-down and bottom-up approaches for building a data warehouse are different ? Compare Hub-and-Spoke with Data-mart bus type of architecture in this regard. 6(CO1)

(b) Give example of each of the following :

(i) Degenerate dimension

(ii) Junk dimension

(iii) Semi additive fact table

(iv) Additive fact table. 4(CO1)

2. (a) Consider a table Sales (brand, model, category, sales_amt)
Write the following queries in SQL :

(i) the sales amount grouped by brand and category

(ii) the sales data grouped by (brand, category), (brand), (category), ()

(iii) calculate the sales amount by brand (subtotal) and both brand and category (total)

(iv) calculate the sales amount by both brand and category (subtotal) and both brand and category (total).

(v) Write a query demonstrating partial roll-up and comment on the result obtained. 5(CO2)

(b) For the following scenario, design a star schemas to handle the heterogeneous product dimension. An insurance company offers household and automobile insurance. They would like to analyze claims by customer and policy type. Common to both types of policy is a policy number, an insured amount and a premium amount paid by the customer. Household insurance policies, however, record address details of the property, numbers, of rooms, security system and householders. Automobile insurance policies record Vehicle identification Numbers (VINs) number plate, primary driver, vehicle make and model. Assume three years of claim data, 150,000 customers with an average of one claim per customer over the three years. 5(C)2)

3. (a) Say you have a table called SALES with many columns including two special ones that are candidates for partitioning; state _code, which stores a two-digit code for the state where the sale was made, and product_code a three-digit number identifying the product sold by that sales record. Users query on the table filtering on both columns equally, and the archival requirements are also based on both these two columns. When you apply the principles of partitioning decisions, you find that both these columns are good candidates for partitioning keys. Which columns would you choose as the partitioning key ? Defend your suggestion. Also write the command to create this partitioned table. 5(CO2)

(b) Create a cluster EMP_CLUSTER with dept_no as cluster key. The cluster has two tables, employees and departments. Also write command to create an index on this cluster.

Create a single table hash cluster COSTOMERS_CLUSTER_HASH with custno as cluster key with size 512 bytes and 500 hashkeys. Allocate 500K initially and extend it with 50K next. What are the benefits of using clusters ? 5(CO2)

4. (a) Consider the dataset : 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) Draw a boxplot for this data.

(ii) Use min-max normalization to transform the value 33 for age onto the range [0.0,1.0].

(iii) Plot an equal-width histogram of width 10. 5(CO3)

(b) Prove or disprove the hypothesis "is there any correlation between attending class and passing the exam". for the following contingency table :

	Pass	Fail	Total
Attended	25	6	31
Skipped	8	15	23
Total	33	21	54

Given : For DF = 1, SL = 0.0006, chi-square = 5.23. 5(CO3)

5. (a) Assume the APRIORI algorithm identified the following 7 4-item sets that satisfy a user given support threshold: abcd, abce, abcf, acde, adef, bcde, and bcef. What initial candidate 5-itemsets are created by the APRIORI algorithm ? From your observations, infer which of those survive subset pruning? 4(CO4)

(b) Apply Naive Bayesian classifier on the following dataset and classify a Red Domestic SUV.

Example No.	Color	Type	Origin	Stolen ?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

6(CO4)

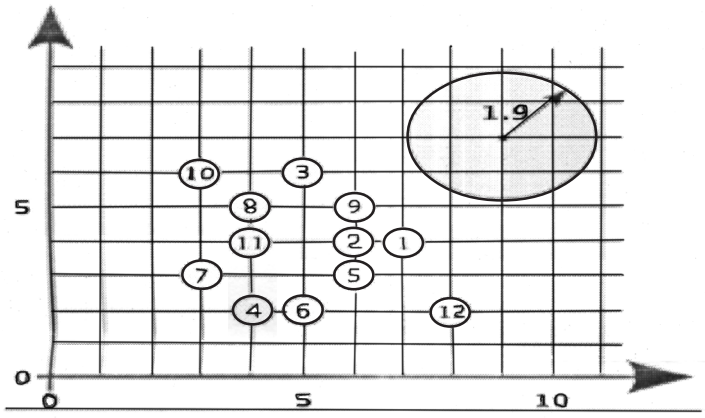
6. (a) Use single, complete, and average link agglomerative clustering to group the data describe by the following distance matrix. Compare the dendrograms and give your comments.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

10(CO4)

OR

(b) Apply DBSCAN Algorithm with radius 1.9 and MinPts = 4 (3 neighbors + the point you are considering as center for computing the density).



(i) Indicate if a point is a core, border or noise point.

(ii) Plot the clusters obtained.

(iii) Compare the results with radius 2.2. 10(CO4)