Course Code : CST 411 **KOLP/RW – 19 / 9159**

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

1. (a) Given three dimension tables :
   - time : day, week, month, quarter, year.
   - part : pname, type, color, brand, manufacturer.
   - customer : cname, type, city, state, country.

   (i) List the hierarchies in each dimension.

   (ii) Starting with the base cuboid [day, pname, cname] how many cuboids would be generated ? 3 (CO 1)

   (b) Suppose that a data warehouse contains 20 dimensions, each with about five levels of granularity :

   (a) Users are mainly interested in four particular dimensions, each having three frequently accessed levels for rolling up and drilling down. How would you design a data cube structure to support this preference efficiently ?

   (b) At times, a user may want to drill through the cube, down to the raw data for one or two particular dimensions. How would you support this feature ? 4 (CO 1)

   (c) What is the purpose of 'refresh' in ETL process ? When should we refresh ? What are the different refresh techniques ? 3 (CO 1)

2. (a) Consider a dimension CUSTOMER (cust_key, cust_name, cust_code, acc_status, marital_status, address, state, zip).

   Using your knowledge of types of dimension tables, how will you

store / implement the following types of changes ? Give detailed explanation for each :

(i) Correction in name.

(ii) Address of customer changes and the application needs to keep track of current and previous address.

(iii) Acc_status values can be good, late, very late, in arrears, suspended. The history of account status of each customer is to be maintained. The account status of a customer gets changed frequently. 6 (CO 2)

(b) Design a star schema according to the following scenario. The Restaurants 'R Us wholesale restaurant company supplies equipment to 55 different restaurants in Melbourne, such as tables, chairs, table cloths, napkin holders, cutlery and so on, as well as kitchen equipment such as saucepans, knives and chef clothing, They wish to analyze their daily sales in terms of revenue, unit sales, costs and profit for each product and customer. They also would like to know this information by product line and product group (front of house, kitchen). 4 (CO 2)

3. (a) Consider a table my_table(id number(5), name varchar2(10)) is created with 100 rows. Give commands to partition this existing table into two partitions p1 and p2 containing first 50 and next 50 records respectively. 5 (CO 2)

Solve **3(b)** or **3(c)** :—

(b) Find all instructors in Computer Science department with salary of 74,000 or more. Outline the steps in answering the query, and demonstrate the final and intermediate bitmaps constructed to answer the query :

ID	name	dept_name	salary
10101	Srinivasan	Comp. Sci.	65000
12121	Wu	Finance	90000
15151	Mozart	Music	40000
22222	Einstein	Physics	95000
32343	El Said	History	60000
33456	Gold	Physics	87000
45565	Katz	Comp. Sci.	75000
58583	Califieri	History	62000
76543	Singh	Finance	80000
76766	Crick	Biology	72000
83821	Brandt	Comp. Sci.	92000
98345	Kim	Elec. Eng.	80000

5 (CO 2)

OR

(c) Consider a table sales_figures with columns : store_id NUMBER, Quarter INTEGER, month INTEGER, amount NUMBER, remarks VARCHAR2, and primary key store_id, quarter, month)).

Give command to create an index organized table with column remarks in the overflow area.

What is the difference between a btree index and IOT. 5 (CO 2)

4. (a) Find Q1, Q2 and Q3 for the following data set. Identify any mild and extreme outliers if any, and draw a box – and – whisker plot.

{5, 40, 42, 46, 48, 49, 50, 50, 52, 53, 55, 56, 58, 75, 102} 3 (CO 3)

(b) Consider the first and second exam scores of 35 students :

Student	First	Second	Student	First	Second
1	21	22	19	25	22
2	23	23	20	13	19
3	16	19	21	17	22
4	23	19	22	23	18
5	23	24	23	11	21
6	17	21	24	17	14
7	12	18	25	18	11
8	15	16	26	13	16
9	20	20	27	18	11
10	8	10	28	16	15
11	22	24	29	21	17
12	22	22	30	15	9
13	23	22	31	16	22
14	18	19	32	22	16
15	22	23	33	18	16
16	20	20	34	21	13
17	20	20	35	19	24
18	20	20			

(i) Draw a scatterplot of the data. How are the two exam scores related based on the plot ? Comment on the type of relationship.

(ii) Compute the Pearson's correlation coefficient between the first and second exam scores. Does this value support your judgment in the previous question ?

(iii) The teacher decided to curve the first exam scores by giving away 5 points. Obtain a new scatterplot and compare this with the old plot. 7 (CO 3)

Solve **Q. 5** or **Q. 6** :

5. (a) Consider the training examples shown in the following table for a binary classification problem :

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

(i) Compute the Gini index for the overall collection of training examples.

(ii) Compute the Gini index for the Customer ID attribute.

(iii) Compute the Gini index for the Gender attribute.

(iv) Compute the Gini index for the Car Type attribute using multiway split.

(v) Compute the Gini index for the Shirt Size attribute using multiway split.

(vi) Which attribute is better Gender, Car Type, or Shirt Size ? Why ?

(vii) Why Customer ID should be not used as the attribute test condition even though it has the lowest Gini ? 10 (CO 4)

OR

6. (a) The following contingency matrix shows a breakdown of transactions for coffee and tea drinkers (assume 1000 transactions) :

	Coffee	Not coffee	Total
Tea	150	50	200
Not tea	650	150	800
Total	800	200	1000

(i) Calculate support, confidence and lift for the association rule $\{tea\} \rightarrow \{coffee\}$.

(ii) Analyze why lift better represents the relationship between tea and coffee drinkers than using support and confidence. 6 (CO 4)

(b) Some classification algorithms run out of memory in trying to fit all data in memory to create the classification model. Analyze ways in which you might address the issue of memory capacity. 4 (CO 4)

7. (a) Assume the following dataset is given : (2, 2), (4, 4), (5, 5), (6, 6), (8, 8), (9, 9), (0, 4), (4, 0). K – Means is used with k = 4 to cluster

the dataset. Moreover, Manhattan distance is used as the distance function to compute distances between centroids and objects in the dataset. The initial clusters C1, C2, C3 and C4 are as follows :

C1 : {(2, 2), (4, 4), (6, 6)}

C2 : {(0, 4), (4, 0)}

C3 : {(5, 5), (9, 9)}

C4 : {(8, 8)}

Give a run of K – means for a single iteration. Evaluate the new clusters and their centroids. 5 (CO 4)

(b) Assume DBSCAN is run with MinPoints = 6 and epsilon = 0·1 for a dataset and 4 clusters are obtained and 5% of the objects in the dataset are classified as outliers. Now if DBSCAN is run with MinPoints = 8 and epsilon = 0·1, how do you expect the clustering results to change ? Justify you answer. 5 (CO 4)