Course Code : CST 406 **KOLP/RW – 19 /9620**

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.
(3) Plot neat graphs on graph papers.

1. (a) Explain the life cycle of data warehouse development. Compare top – down approach with bottom – up approach of data warehouse designing. 5(CO1)

   (b) Explain different types of hierarchies with examples. 5(CO1)

2. (a) Suppose that a data warehouse for University consists of the following four dimensions : student, course, semester and instructor and two measures count and avg_grade. When at the lowest conceptual levlel (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.

   (iii) If each dimension has five levels (including all), such as "student < major < status < university < all", how many cuboids will this cube contain (including the base and apex cuboids) ? Explain your answer. 5(CO2)

(b) Consider the following schema :
Student (studID, name, major)
Instructor (instID, dept) ;
Class (classID, univ, region, country)
Took (studID, instID, classID, score

Write OLAP queries for :—

(i) Find average scores grouped by student and instructor for courses taught in Vidarbha region.

(ii) "Roll up" your result from problem 1 so it's grouping by country only.

(iii) Find average scores grouped by student major.

(iv) "Drill down" on your result from problem 3 so it's grouping by instructor's department as well as student's major.

(v) Use "WITH ROLLUP" on attributes of table Class to get average scores for all geographical granularities : by country, region and university, as well as the overall average. 5(CO2)

3. (a) Consider the following queries. Assume a btree index exists on column employee_id, last_name, department_id, composite index on (cust_gender, cust_email), composite index on (department_id, last_name, salary).
State which type of index scan will the CBO use in each case. Clearly state your assumptions if any.

(i) SELECT *FROM employeesWHERE employee_id = 5 ;

(ii) SELECT * FROM sh.customers WHERE cust_email = 'Abbey@company.com' ;

(iii) SELECT department_id , last_name , salaryFROM employees ;

(iv) SELECT department_idFROM departments where department_id between 10 and 40 ;

(v) SELECT department_id , last_name , salaryFROM employees WHERE salary > 5000 ORDER BY department_id , last_name ; 5(CO2)

(b) Write commands to :

(i) create a cluster named PERSONNEL containing two tables EMP and DEPT.

(ii) drop the cluster.

(iii) to create hash cluster named TRIAL_CLUSTER over the same tables. What is the advantages of using TRIAL_CLUSTER over PERSONNEL ? 5(CO2)

4. (a) Consider the following confusion matrix :

		Predicted		
		Yes	**No**	**Neutral**
Actual	**Yes**	15	10	100
	No	10	15	10
	Netural	10	100	1000

Calculate Accuracy, Precision, Recall and F measure. Write your observations based on these values. 5(CO3)

(b) Consider the following data : 5 , 10 , 11 , 13 , 15 , 35 , 50 , 55 , 72 , 92 , 204 , 215

Partition this data into three bins by each of the following methods :

(a) equal – frequency (equi – depth) partitioning

(b) equal – width partitioning

And apply following techniques for smoothing : by means by boundaries. 5(CO3)

5. (a) Apply apriori algorithm and generate strong rules from the following data set. Assume min_sup of 20% and min_conf of 80%.

Trans ID	Items Purchased
101	Apple , Orange , Litchi , Grapes
102	Apple , Mango
103	Mango , Grapes , Apple
104	Apple , Orange , Litchi , Grapes
105	Pears , Litchi

5(CO4)

(b) Consider the following training set with 3 features 2 classes :

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II

Build the first level of decision tree using information gain. 5(CO4)

6. (a) Consider the following dataset : a (4 , 4) , b (8 , 4) , c (15 , 8) and d (24 , 4). Draw dendograms using simple link, average link and complete link. (Hint : use Euclidean distance to calculate distance matrix) 10(CO4)