Course Code : CST 406 **CXDW/RW – 18 / 5582**

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

DATA WAREHOUSING AND MINING

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

(1) All questions carry marks as indicated against them.
(2) Number your answers properly.
(3) Assume suitable data and illustrate answers with neat sketches wherever necessary.

1. (a) Consider a data warehouse for a hospital where there are three dimensions – Doctor, Patient and Time. Consider two measures Count and Charge where Charge is the fee that the doctor charges a patient for a visit. For the above example create a Cube and illustrate the following operations.

   (i) Rollup

   (ii) Drill Down

   (iii) Slice

   (iv) Dice

   (v) Pivot 6(CO1)

   (b) Bring out the differences between

   (i) OLTP and OLAP

   (ii) ROLAP and MOLAP 4(CO1)

2. (a) Compare and contrast data warehouse and data mart. Also specify reasons for creating data mart. 4(CO2)

(b) For a Supermarket chain, consider the following dimensions namely product, store, time and promotion. The schema contains central fact table for Sales

(i) Design a STAR schema for the above application.

(ii) Calculate the maximum number of base fact table records for warehouse with the following values given below

- Time period : 5 years
- Store : 300 stores reporting daily sales
- Product : 40,000 products in each store 6(CO2)

3. (a) Write SQL query to create an IOT look_ups with the attributes (lookup_code, lookup_value, lookup_description) in tablespace ts–lookup with following specifications :

(i) Constraint : lookup_code should be primary key

(ii) PCTTHRESHOLD is 20

(iii) lookup_description should be in overflow area and Overflow should be in ts_overflow tablespace

OR

(b) Write SQL query to create a table with list partition as follows :—

(i) Table having columns deptno, deptname, quarterly_sales and state.

(ii) Create partition on state :

- Northwest on OR and WA
- Southwest on AZ , UT and NM
- Northeast on NY , VM and NJ
- Southeast on FL and GA
- Northcentral on SD and WI
- Southcentral on OK and TX

(iii) Write SQL to alter already created partitioned table to add new partition “Unknown” which will accept any default state values. 5(CO2)

(c) Describe query optimization technique with materialized views in data warehouse. Take suitable example for illustration. 3(CO2)

(d) Bring out essential difference between Cost based and Rule based optimizer. 2(CO2)

4. (a) Illustrate with example, each of the following data mining functionalities :—

(i) Association and Correlation Analysis

(ii) Classification and Prediction

(iii) Clustering and Evolution Analysis 7(CO3)

OR

(b) Using Equi-Depth binning method, partition the data given below into 4 bins and perform smoothing according to the following methods.

(i) Smoothing by bin means

(ii) Smoothing by bin median

(iii) Smoothing by bin boundaries

(iv) How outliers will be determined in the given data ?
24 , 25 , 26 , 27 , 28 , 56 , 67 , 70 , 70 , 75 , 78 , 79 , 89 , 90 , 91 , 94 , 95 , 96 , 100 , 102 , 103 , 107 , 109 , 112 7(CO3)

(c) What are the major challenges of mining a huge amount of data (such as billion of tuples) in comparison with mining a small amount of data (few hundred tuple data set) ? 3(CO3)

5. (a) A database consists of nine transactions taken from grocery store. Enumerate all the frequent itemset using, Apriori algorithm with minimum support threshold S = 3 and minimum confidence threshold C = 50%. Show the candidate and frequent itemset for each database scan. List all the association rules that are generated and highlight the strong one, sort them by confidence.

Tid	Item
1	Milk, Bread, Biscuits
2	Bread, Sugar
3	Bread, Cereal

4	Milk, Bread, Sugar
5	Milk, Cereal
6	Bread, Cereal
7	Milk, Cereal
8	Milk, Bread, Cereal, Biscuits
9	Milk, Bread, Cereal

7(CO4)

OR

(b) Apply FP – growth algorithm on following example with min_sup = 20% Construct FP – Tree and show frequent pattern generated at the end of the algorithm.

Tid	Item
1	a, b, e
2	b, d
3	b, c
4	a, b, d
5	a, c
6	b, c
7	a, c
8	a, b, c, e
9	a, b, c

7(CO4)

(c) Calculate the distance between two objects, A = (22 , 2 , 45 , 18) and B = (21 , 0 , 34 , 9) using the Manhattan distance and Minkowski distance for p = 4. 3(CO4)

6. (a) Consider five points {x1 , x2 , x3 , x4 , x5 } with the following coordinate as two demension samples for clustering :
X1 = (0 , 2) , x2 = (1 , 0) , x3 = (2 , 1), x4 = (4, 1), x5 (5, 3).

Apply K – Means algorithm on the above data set. The required number of clusters is two, initially clustered are formed from random distribution of samples : C1 (x1 , x2 , x4) and C2 (x3 , x5) 6(CO4)

OR

(b) The distance between five pair of cases given below:

	A	B	C	D	E
A	0				
B	8	0			
C	2	6	0		
D	5	4	8	0	
E	10	9	3	7	0

Cluster the five cases using below procedure and draw the Dendograms structure.

(a) Single linkage hierarchical procedure.

(b) Complete linkage hierarchical procedure.

(c) Average linkage hierarchical procedure. 6(CO4)

(c) For the following algorithms, write

(i) Shape of the cluster that can be determined and

(ii) Time complexity.

(a) CLARA

(b) CLARANS

(c) BIRCH

(d) DBSCAN 4(CO4)