Course Code : CST 411 **CXDW/RW – 18 / 5107**

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

DATA WAREHOUSING AND MINING

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

(1) Assume suitable data wherever necessary and clearly state your assumptions.
(2) Give examples wherever necessary.

1. (a) We want to store a multidimensional structure containing the following information about sales :

- quantity (number of items sold)
- customer (name of the customer)

over the following dimensions :

- Time (day, week, month, quarter, year)
- Product (type, band, category, group)
- Location (city, region, country, continent)

(i) Define a star schema to represent the above multidimensional structure. Show implicit and explicit hierarchy in each dimension.

(ii) Define a snowflake schema that reduces (at least on one dimension) the redundancy of the star schema defined at the previous point.

(iii) Write an SQL query over the star schema defined at point

(i) That returns the names of the customers who bought a product from category "Car" in 2015 in Italy.

(iv) Write the SQL query over the snowflake schema defined at point

(ii) That returns the names of the customers who bought a product from category "Car" in 2015 in Italy.

8 (CO 1)

(b) Explain the importance of aggregates in data warehousing.
A company is using OLAP to provide monthly summary information about its products and branch sales to the company managers. How many different aggregates would be required to fill a data cube on product, branches and dates if there are 20 products, 10 branches and five years of monthly data ? 2 (CO 1)

2. (a) A cuboid (day, pname, cname) of 100 GB is already materialized. Remaining choices are :
(day, pname) - 60 GB
(day, cname) - 20 GB
(pname, cname) - 1 GB
(day) - 10 GB
(pname) - 200 MB
(cname) - 30 MB
(ALL) - 8 bytes

The following queries are fired with equal probability :
Q1: total sales per (pname,cname)
Q2: total sales per (pname)
Q3: total sales per (cname)

Which cuboids should we materialize if available space is and why :

(1) 10 GB (2) 1 GB (3) 100 MB. 4 (CO 2)

(b) Assume you are in the insurance business. Find two example of Type 2 slowly changing dimensions in that business. As an analyst on the project, write the specifications for applying the Type 2 changes to the data warehouse with regard to the two examples. 3 (CO 2)

(c) What is a cuboid ? Explain OLAP operations on data cube with example. 3 (CO 2)

3. (a) Consider a table Orders (order_mode, order_status, order_date) where order_mode can be online or offline and order status can be 0 or 1 or 2 or 3. Assuming suitable data in the table, illustrate the concept of key compression. What is the effect of changing prefix length in key compression? 5 (CO 2)

(b) What is clustering factor ? Can rebuilding the index improve the clustering factor ? Justify your answer. Consider a table ORDERS which grows every day. Index is created on the order date and another one on the customer id. Because orders don't get deleted there are no holes in the table so that each new order is added to the end. The table grows chronologically. Which index will have a good clustering factor and which will have a bad clustering factor ? Why ? 5 (CO 2)

4. (a) Assume an integer - valued attribute A whose values are distributed as follows : 0, 0.0, 1, 1 ,1 ,2, 2, 5, 6, 7, 17, 18, 19, 20, 25, 28, 29, 33, 39, 43, 44, 44, 46, 51, 58, 59, 60, 61, 65, 77 , 78, 81, 99, 120.

Apply min-max normalization and z-score normalization for attribute A for value 25. If the value of the attribute A that has been normalized by min-max normalization is 0.25 what does this value tell you about the location of the attribute value 0.25 relative to the other values for attribute A ? If the value of the attribute A that has been normalized using z-score is-2 what does this value tell you about the location of the attribute value-2 relative to the other values for attribute A ? 5 (CO 3)

(b) Imagine that a local clothing manufacturer has 2,700 employees. The personnel manager decides to ask the employees for suggestions on how to improve their workplace. It would take too long to survey everyone, so the manager chooses to systematically sample 300 of the employees.

(a) What would be the sampling interval ?

(b) If the number 8 was your first randomly drawn number, what would be the first 5 numbers of your sample ? 2 (CO 3)

(c) Suppose two stock A and B have the following values in one week : (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).

If the stocks are affected by the same industry trends, will their prices rise or fall together ? 3 (CO 3)

5. (a) Assume we want to mine rules of the form $X \longrightarrow Y$ where X and Y are sets of items that maximize a measure of interestingness called Lift:

Lift (X, Y) = confidence $(X \longrightarrow Y)$/support (Y)

Assume we mine rules using this measure for the transaction set given below:

T1 : a, b, c, d
T2 : b, c, d
T3 : a, b, d, e
T4 : a, c, d, e
T5 : b, c, d, e
T6 : b, d, e
T7 : c, d
T8 : a, b, c
T9 : a, d, e
T10 : b, d

We only consider rules that have a single item on the left hand and right hand side; e. g. $X \longrightarrow Y$ and we additionally request $X < Y$. Compute the 3 rules that have the highest lift for the given set of transactions. Compute the 3 rules that have the highest confidence for the given data set. How is using lift as the measure of interestingness different from support and confidence? What kind of relationships does lift association rule mining reveal? Give reasons for your answer. 10 (CO 4)

OR

(b) The following table summarizes a data set with three attributes *A, B, C* and two class labels +, –. Build a two - level decision tree.

A	B	C	Numbers of Instances	
			+	–
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

According to the classification error rate, which attribute would be chosen

as the first splitting attribute ? For each attribute, show the contingency table and the gains in classification error rate. 10 (CO 4)

6. (a) Compare the clusters that can be found with K - means with clusters that can be found with grid - based clustering algorithms (e. g. the basic grid - based algorithm). Give examples that illustrate limitations of both approaches. Given are the points A = (1, 2), B = (2, 2), C = (2, 1), D = (-1, 4), E = (-2, -1), F = (-1, -1)

(i) Starting from initial clusters Cluster1 = {A} which contains only the point A and Cluster 2 = {D} which contains only the point D, run the K - means clustering algorithm and report the final clusters.
Use L1 distance as the distance between points which is given by :
$d((x1, y1), (x2, y2)) = |x1 - x2| + |y1 - y2|$

(ii) Draw the points on a 2 – D grid and check if the clusters make sense. 10 (CO 4)

OR

(b) Assume the following points are given : (2, 2), (3, 3), (7, 6), (6, 7), (7, 7), (1, 2), (5, 3), (9, 9). Moreover Manhattan distance is used as the distance functions and min - distance is used as the cluster distance function.

(i) Assume that agglomerative average linkage hierarchical clustering is applied to the problem. What will be the result of the clustering process ? Provide computations you performed to reach the result.

(ii) Now assume DBSCAN is applied to the same problem with MINPOINTS = 2 and $\varepsilon = 3$. What will be the result of applying DBSCAN ; which points in the dataset of core, border or outliers (noise points) ? Does the result change, if we set MINPOINTS to 3 ? 10 (CO 4)