Course Code : CST 411 **EVFU/MW – 18 / 6108**

Seventh 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) A manufacturing company has huge sales network. To control the sale, it is divided into regions. Each region has multiple zones. Each zone has different cities. Each sales person is allotted different cities. The objective is to track the sales figure at different granularity levels of region and to count number of products sold. Design Star schema by considering granularity level for region, sales person and time. Convert Star schema to Snowflake Schema. 6 (CO 2)

(b) What is meant by metadata in the context of data warehuses ? Explain different types of metadata stored in data warehouse. 4 (CO 1)

2. (a) Discuss the different types of measures stored in the fact table in a Data Warehouse with suitable examples. 4 (CO 2)

OR

(b) What is factless fact table ? Design Star schema with factless fact table to track employees leave management system. 4 (CO 2)

(c) Explain the following types of Dimension with suitable examples

- Rapidly changing
- Junk dimension
- Degenerate dimensions 6 (CO 2)

3. (a) Write a query to create composite Range – List partitioning for the following scenario :—

- Employee table having attributes First_Name, Middle_Name, Last_Name, Birth_Date, State.
- Perform Range partitioning on Birth_Date attributes and List partitioning on State
- Partition definitions for range are as below :
  - Partition P1 should accept values less than 01 – Jan – 2018
  - Partition P2 should accept values less than 01 – April – 2018
  - Partition P3 should accept values less than 01 – July – 2018
  - Partition P4 should accept values greater than 01 – July – 2019
- Partition definitions for list are as below :
  - Partition East should accept values ('AK')
  - Partition South should accept values ('NY', 'NJ', 'VA', 'CT')
  - Partition North should accept values ('TX', 'MS', 'GA', 'KY')
  - Partition West should accept values ('CA', 'AZ', 'OR', 'NV')
  - Partition No_State should accept any values.

5 (CO 2)

OR

(b) Differentiate between View and Materialized views.
Write SQL query to create Materialized view for the following query with build immediate option.
SELECT gender, semester_year AS year, semester_month AS month, SUM (num_of_students) AS total
FROM instructor_summary
GROUP BY ROLLUP (gender, semester_year, semester_month) ; 5 (CO 2)

(c) "Update to the bitmap index takes bit longer time than B tree indexes". Justify with suitable example. 3 (CO 2)

(d) Bring out the difference between clustered and non – clustered index with suitable example. 2 (CO 2)

4. (a) Given is the data for age in particular region after survey
14 , 16 , 17 , 17 , 20 , 21 , 21 , 22 , 23 , 23 , 26 , 26 , 26 , 26 , 31 , 34 , 34 , 36 , 36 , 36 , 36 , 37 , 41 , 46 , 47 , 53 , 71.

Apply the following methods and show the results :—

(i) Use smoothing by bin means with a depth of 3.

(ii) Use Min – Max normalization to transform the value 36 into the reange 0.0 to 1.0

(iii) Use z – score normalization to transform the value 36 where the standard deviation of the above data is 13.94

(iv) Use normalization by decimal scaling to transform the value 36.

(v) Plot an equi – width histogram of width 10.

(vi) Sketch examples of different sampling techniques using sample of size 5 and the strata low, medium and high. 6 (CO 3)

(b) Suppose that 1000 people attended a disease prediction test. Among 300 patients having heart related disorders, 280 of them tested positive, 20 tested negative. Among the 700 people, without having any heart diseases, 685 tested negative and 15 tested positive. Find accuracy, precision, recall and specificity. 4 (CO 3)

5. (a) Consider the following datasets having five transactions, assuming minimum support as 2. Find all frequent itemsets generated using FP tree algorithm. Draw FP tree.

TID	Items Brought
T1	(a , c , d , f , g , i , m , p)
T2	(a , b , c , f , l , m , o)
T3	(b , f , h , j , o , w)
T4	(b , c , k , s , p)
T5	(a , f , c , e , 1 , p , m , n)

5 (CO 4)

(b) Find the root node of the decision tree for the following data set using information gain.

Gender	Car Ownership	Travel Cost	Income Level	Transport Mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car
Female	1	Cheap	Medium	Train
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train

5 (CO 4)

6. (a) Use the k – means algorithm and Eucliean distance to cluster the following 10 examples into 3 clusters X_1 (3 , 11) ; X_2 (3 , 6) ; X_3 (9 , 5) ; X_4 (10 , 5) ; Y_1 (6 , 9) ; Y_2 (8 , 6) ; Y_3 (7 , 5) ; Z_1 (2 , 3) ; Z_2 (5 , 10) ; Z_3 (7 , 11).
Suppose that the initial seeds (centers of each cluster) are X_1 , X_4 and Z_2
Run the k – means algorithm for 3 iteration. At the end of each iteration, show :—

(i) The new clusters. (i. e. the examples belonging to each cluster)

(ii) The centers of the new clusters.

(iii) Draw a 10 by 10 space with all the 10 points and show the clusters after each iteration. 6 (CO 4)

(b) The table below comprises sample data items including the distance between the elements.

	E	A	B	C	D
E	0	1	2	2	3
A	1	0	2	5	3
B	2	2	0	1	6
C	2	5	1	0	3
D	3	3	6	3	0

Use single link agglomerative clustering to group the data described by the following distance matrix. Produce the dendograms. 4 (CO 4)