Course Code : CST 406 **ITSJ/RW – 17 / 1398**

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) A data warehouse for Hotel consists of three dimensions–hotel, customer, reservation and two measures amount_paid and customer_count
Analyze these dimensions and list the possible attribute for each dimension tables. Also designate a primary key for each table.
Construct snow flake schema for the above scenario. Make suitable assumptions. List the concept hierarchies. 6(CO 2)

(b) What is factless fact table ? Design Star schema with factless fact table to track a patient by diagnostic procedure and time. 4(CO 2)

OR

(c) A data warehouse is subject oriented. Identify major critical business subjects for the following companies ?

(i) An international manufacturing company.

(ii) A Hospital.

(iii) A domestic hotel chain. 4(CO 1)

2. (a) Explain the computation of measures in a data cube :

(a) Enumerate three categories of measures, based on the kind of aggregate functions used in computing a data cube.

(b) For a data cube with the three dimensions time, location and item, which category does the function variance belong to ? Describe how to compute it if the cube is partitioned into many chunks.

(c) Suppose the function "top 10 sales". Discuss how to efficiently compute this measure in data cube. 7(CO 1)

(b) In data warehouse technology a multiple dimensional cube can be implemented either by a multi–dimensional database technique (MOLAP) or by a relational database technique (ROLAP). Briefly describe each implementation technique. 3(CO 1)

3. (a) State the advantage of storing partitions in different tablespaces ? Write a query to create range partitioned table for the following scenario:

- Create a table named–Purchase consisting of four partitions, one for each quarter of Purchase for the year 2016. The column Purchase_Year is the partitioning column, while its values constitute the partitioning key of a specific row.
- The other columns for table must be prod_id, cust_id, promo_id, quantity_purchased, amount_Purchased–all in number format and Year.
- Store each partition in different tablespaces.
- Write a query to insert row in partition 3.
- Write a query to merge partition 3 and 4. 5(CO 2)

OR

(b) Illustrate Index Organized Table and function based indexes with suitable examples. 5(CO 2)

(c) Constrast beween Cost based and Rule based optimizer. Describe query optimization technique with materialized views in data warehouse. Take suitable example for illustration. 5(CO 2)

4. (a) The following data set represents measurements student made when pitching pennies against a wall. The measurements, in centimeters, recorded how far the penny was from the wall when it landed. 3, 10, 5, 6, 20, 15, 99, 64, 23, 2, 5, 6, 10, 11, 7, 8, 9, 6, 12, 15, 8, 22, 18, 11, 8, 7, 9, 11, 12, 8.

(i) What is mean and median of data ?

(ii) What is the mode of data ? Comment on the data modality (i,e. bimodal, trimodal, etc).

(iii) What is the standard deviation and midrange of the data ?

(iv) Find first quartile (Q1), third quartile (Q3), IQR of the data?

(v) Give the five–number summary of the data.

(vi) Show a boxplot of the data.

(vii) How is quantile–quantile plot different from a quantile plot ? 7(CO 3)

(b) Describe the steps involved in data mining when viewed as a process of knowledge discovery. 3(CO3)

OR

(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)

Transaction ID	**Basket Content**
1234	(Aspirin, Panadol}
4234	{Aspirin, Sudafed}
9373	{Tvlenol, Cepacol}
9843	{Aspirin, Vitamin C,Sudafed}
2941	{Tvlenol, Cepacol}
2753	{Aspirin,Cepacol}

Contd..

Transaction ID	**Basket Content**
9643	{Aspirin,Vitamin C}
9691	{Aspirin,Ibuprofen, Panadol)
5313	{Panadol, Vitamin C}
1003	{Tvlenol, Cepacol, Ibuprofen}
5636	{Tvlenol, Panadol, Cepacol}
3478	{Panadol, Sudafed, Ibuprofen}

Apply Apriori algorithm on the above dataset with support count threshold $s_{\mathbf{min}}$ = 4. Show the candidate and frequent itemsets for each iteration. Enumerate all the final frequent itemsets. Also indicate the association rules that could be generated from these itemsets and highlight the strongest ones. 7(CO 4)

(b) Compute accuracy, error rate, sensitivity, specificity, precision, recall and F measure for the following confusion matrix :—

Classes	**Buys_smartphone=yes**	**Buys_smartphone=no**
Buys_smartphone=yes	100	40
Buys_smartphone=no	60	300

3(CO3)

6. (a) Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,0). The distance function is Manhattan distance. Suppose initially we assign A1, B1 and C1 as the center of cluster, respectively. Use the K–means algorithm to show only :—

(i) The three cluster center after the first round of execution.

(ii) The final three clusters. 5(CO 4)

OR

(b) Describe each of the clustering algorithm in terms of the following criterion:

(1) Shape of the cluster that can be determine.

(2) Input parameter that must be specified.

(3) Limitations.

(4) Time complexity of the algorithm.

(a) CLARA.

(b) BIRCH. 5(CO 4)

(c) Given two objects represented by the tuples (24,3,44,12) and (22,2,38,10):

(a) Compute Euclidean distance between two objects.

(b) Compute Manhattan distance between two objects.

(c) Compute Minkowski distance between two objects, uning q=3 3(CO 4)

(d) Compare K–means and K–medoids/PAM. What are the main differences between the two algorithms ? 2(CO 4)