Course Code : CST 406 **GTHS/RS – 19 / 7619**

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

DATA WAREHOUSING AND MINING

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

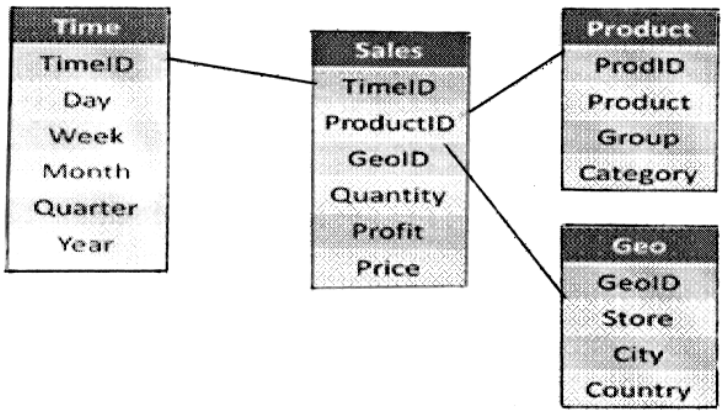
(1) All questions carry marks as indicated against them.
(2) Due credit will be given to neatness and adequate dimensions.
(3) Assume suitable data and illustrate answers with neat sketches wherever necessary.

1. (a) What is the difference between slowly changing and rapidly changing dimensions ? Give an example of each. 3 (CO 1)

(b) Given are the fact table PropertySale *(branchNo, PropertyType, YearMonth SaleAmount)* and dimension table *Branch (branchNo, city),*
Write SQL statement to answer the following query.
"Retrieve total amount of sale in January and February 2007 in Manchester, Edinburgh, and Birmingham, with subtotals for each property type, month, and city (including all cross-tabular subtotals)"
Write sample queries for : partial rollup, grouping set assuming suitable data. What is the use of grouping function ? 4 (CO 1)

(c) Why is metadata important in a data warehouse ? Explain the different components of metadata repository. 3 (CO 1)

2. (a) Consider the following diagram.



What kind of a schema is presented ? Considering that the product dimension is subject to change often, how would you transform the schema to accommodate this ? Draw the new schema. 3 (CO 2)

(b) A data warehouse of a train company contains information about train segments. It consists of six dimensions, namely, departure station, arrival station, trip, train, arrival time, and departure time and three measures, namely, number of passengers, duration and number of kilometers. Write OLAP operations to be performed in order to answer the following queries. Propose the dimension hierarchies whenever needed.

(i) Total number of kilometers made by Alstom trains during 2012 departing from French or Belgian stations.

(ii) Total duration of international trips during 2012, that is, trips departing from a station located in one country and arriving at a station located in another country. 4 (CO 2)

(c) You are data design specialist on the data warehouse project team for a manufacturing company. Design a STAR schema to track the production quantities. Production quantities are normally analyzed along the business dimensions of product, time, parts used, production facility and production run. State your assumptions and mention the concept hierarchies. 3 (CO 2)

3. (a) How are Index Oragnized Tables(IOT) different from B-Tree indexes ? Give syntax for creating a IOT with overflow area. Write SQL command to know which columns are in the overflow segment. 5 (CO 2)

OR

(b) How is data stored in a hash cluster ? How is it retrieved ? Explain the purpose of the following clauses in the create cluster command :

CLUSTER_KEY <datatype>, SIZE <size_number>, SINGLE TABLE, HASHKEYS <hash_keys_numer>, HASH IS <expr> 5 (CO 2)

(c) State what is a Bitmap Join Index. List advantages of creating a bitmap join index over normal joins. Give the command for creating a bitmap join index. 5 (CO 2)

4. (a) Write a short note on Online Analytical Mining(OLAM). 3 (CO 3)

OR

(b) Write a short note on classification of data mining systems. 3 (CO 3)

(c) Discretize the following values using Equi-width and Equi-Depth binning : 13, 15, 16, 16, 19, 20, 21, 22, 22, 25, 30, 33, 35, 35, 36, 40, 45. 3 (CO 3)

(d) A manager wants to see if geographical region is associated with ownership of a Macintosh computer. The manager surveys 100 people and the data breaks down as follows :

	Mac	No Mac	Row total
North East	12	14	26
South West	21	18	39
Mid West	17	18	35
Column Total	50	50	100

Using chi-square test determine if owning a mac and the geographical region where the owner lives are related. (From chi-square table, the chi-square value needed to reject the hypothesis at 0.05 significance level is 3.84146). 4 (CO 3)

5. (a) Generate the frequent itemsets using the FP-growth algorithm for the transaction database shown below and a minimum supports_min = 3.

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

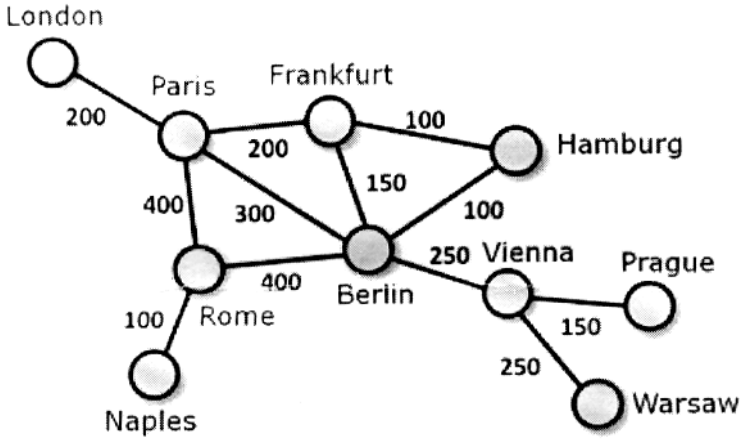
7 (CO 4)

(b) What is bagging and boosting ? State why it may improve the accuracy of decision tree induction. 3 (CO 4)

6. (a) If Epsilon is 2 and minpoint is 2, what are the clusters that DBScan would discover with the following 8 examples : A1 = (2, 10), A2 = (2, 5), A3 = (8, 4), A4 = (5, 8), A5 = (7, 5), A6 = (6, 4), A7 = (1, 2), A8 = (4, 9). Draw the 10 by 10 space and illustrate the discovered clusters. What if Epsilon is increased to 10 ? 6 (CO 4)

OR

(b) Write the MST algorithm. Using MST with maximum space clustering technique, create two clusters of following graph.



6 (CO 4)

(c) Both k-means and k-medoids algorithms can perform effective clustering. Illustrate the strength and weakness of k-means in comparison with the k-medoids algorithm. Also, illustrate the strength and weakness of these schemes in comparison with a hierarchical clustering scheme. 4 (CO 4)