Course Code : CST 406 **KRSJ/RS – 18 / 3612**

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) What is cube materialization ? Discuss its different types. 2(CO1)

(b) Explain the difference between the following two SQL queries :

```
SELECT gender, semester_year AS year, semester_month AS month,
SUM(num_of_student)AS total
FROM instructor_summary
GROUP BY (gender,semester_year, semester_month);
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);
```

What changes when you use CUBE operator instead of ROLLUP ?

Rewrite the query to compute the following three result groups: (gender, Semester year), (semester month) and ().

What is the result of the following query ?

```
SELECT semester_year AS year, campus,
SUM(num_of_classes)AS num_of_classes
FROM instructor_summary
GROUP BY CUBE (semester_year, campus)
ORDER BY 1;
```

5(CO1)

(c) What are dimension hierarchies ? Give three examples. 3(CO1)

2. (a) Consider a data warehouse with three dimensions date, vehicle, road. There is one measure, toll, which is the money that the driver has to pay for using a particular road.

(i) Draw a simple star schema assuming some concept hierarchy for each dimension.

(ii) Starting with the base cube and finest granularity [date, vehicle, road] which sequence of OLAP operations do you need to list the total toll collected on each road in the year 2011? 5(CO2)

(b) A data cube C, has n dimensions and each dimension has exactly p distinct values in the base cuboid. Assume that there are no concept hierarchies associated with the dimensions.

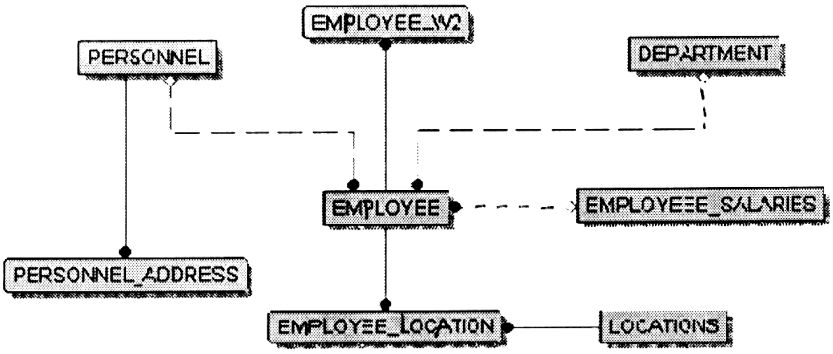
(i) What is the maximum number of cells possible in the base cuboid ?

(ii) What is the minimum number of cells possible in the base cuboid ?

(iii) What is the maximum number of cells possible (including both base cells and aggregate cells) in the data cube, C ?

(iv) What is the minimum number of cells possible in the data cube C ? 5(CO2)

3. (a) Consider the following model :



If PERSONNEL and PERSONNEL_ADDRESS are to be placed in a cluster and EMPLOYEE and EMPLOYEE_W2 are to be placed in another cluster, write commands to create these cluster and tables. 4(CO2)

(b) Explain the reason for error after the following SQL staement are executed:

SQL>create table test (coll number, col2 varchar2(20));

SQL>create index idx1 on test(coll);

SQL>drop table test;

SQL>drop index idx1;

Update col2 such that it stores values in allcaps. Write command to create a function based index on col2. 2(CO2)

(c) Consider a table called YEARLY_SALES with attributes (sales_month INTEGER, state VARCHAR2(2), sales_amount NUMBER).

Write the command to partition this table using range partitioning based on sales_month. Each partition is placed in a different tablespace.

Write the command to partition this table using list partitioning based on state. Each partition is placed in a different tablespace. 4(CO2)

4. (a) Give five–point summary of the following data set and draw the box–plot. Identify any mild and extreme outliers if any :

{5, 40, 42, 46, 48, 49, 50, 50, 52, 53, 55, 56, 58, 75, 102}.

Draw a q–plot for the above data. What can you conclude from this? 5(CO3)

OR

(b) The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(i) Plot an equal–width histogram of width 10.

(ii) Sketch examples of each of the following sampling techniques: SRSWOR, SRSWR, cluster sampling, stratified sampling. Use samples of size 5 and the strata "youth", "middle–aged", and "senior". 5(CO3)

(c) What is entity identification problem ? Consider the following two tables. Table R and Table S present at two different sources.

Table R:

name	street	cuisine
Village Work	Wash. Ave.	Chinese
Ching	Co.B Rd.	Chinese
Old Country	Co.B2 Rd	American

Table S:

name	city	manager
Village Work	Mpls	Hwang
Old Country	Roseville	Libby
Express Cafe	Burnsville	Tom

They are to be merged into a single table to be stored in the data warehouse. Why will the use of common candidate key not work in this case ? Suggest a candidate key for the merged table. 5(CO3)

5. (a) Given a transaction database: {bread milk butter beer} {bread butter water jam beer} [beer diapers bread butter jam} {butter milk juice} {diapers beer juice water}

(i) For minimal support 0.6 mine for all frequent itemsets using vertical data format.

(ii) For minimal confidence 0.7 find association rules of the form item1→item2, item3]

(iii) How many certain association rules can you find ? (with confidence 1.0) 5(CO4)

OR

(b) Given the following training dataset. (Buy Computer data), Identify the root node using information gain and draw the initial decision tree.

R/D	**age**	**income**	**student**	**credit_rating**	**Class:bus_computer**
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31...40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31...40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31...40	medium	no	excellent	yes
13	31...40	high	yes	fair	yes
14	>40	medium	no	excellent	no

5(CO4)

(c) Suppose that we would like to select between two prediction models, M1 and M2. We have performed 10 rounds of 10–fold cross–validation on each model. where the same data partitioning in round i is used for both M1 and M2. The error rates obtained for M1 are 30.5, 32.2, 20.7, 20.6, 31,0, 41.0, 27.7, 26.0, 21.5, 26.0. The error rates for M2 are 22.4, 14.5, 22.4, 19.6, 20.7, 20.4, 22.1, 19.4, 16.2, 35.0. Comment on whether one model is significantly better than the other considering a significance level of 1%. 5(CO4)

6. (a) Given the training data in Q5(b) (Buy Computer data), build an associative classifier model by generating all relevant association rules with support and confidence thresholds 10% and 60% respectively. Classify using this model the new example: age<=30, income=medium, student=yes, credit–rating=fair, selecting the rule with the highest confidence. What would be the classification if we chose to vote the class among all rules that apply ? 7(CO4)

OR

(b) Suppose that the data mining task is to cluster the following eight points (with (x, y) representing, location) into three clusters :

A1(3, 11), A2(3, 6), A3(9, 5), B1(6, 9), B2(8, 6), B3(7, 5) C1(2, 3), C2(5, 10)

The distance function is Manhattan distance. Suppose initially we assign A1 as the center of each cluster, respectively. Use the k–means algorithm to show.

(a) The three cluster centers after the first round of execution.

(b) The final three clusters. 7(CO4)

(c) Present conditions under which density–based clustering is more suitable than partitioning–based clustering and hierarchical clustering. Given some application examples to support your argument. 3(CO4)