Roll No. ___________Name _________________ Section _CS_

National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Data Warehousing & Data Mining**	**Course Code:**	**CS409**
Program:	**BS(Computer Science)**	**Semester:**	**Fall 2017**
Duration:	**3 Hours**	**Total Marks:**	**50**
Paper Date:	**12-Dec-17**	**Weight**	**40%**
Section:	**CS**	**Page(s):**	**7**
Exam:	**Final Exam**		

Instruction/Notes: Scratch sheet can be used for rough work however, all the questions and steps are to be shown on question paper. No extra/rough sheets should be submitted with question paper.
You will not get any credit if you do not show proper working, reasoning and steps as asked in question statements. CALCULATORS are ALLOWED.

Q1. *(2+2+3+3= 10 points)*

Give the appropriate answers of the following questions very briefly:

a. What are the different types of OLAP? What are the operations of OLAP?

b. What is the difference between ELT and ETL?

c. When are materialized views useful? What is the use of query rewrite in materialized view?

d. What kind of situations are there where you might want to use degenerated dimensions? Give an example of degenerated dimension.

Q2. *(3+3+4= 10 points)*

a. Give an example of each of the following data mining functionalities, using a real life database that you are familiar with: classification, clustering, and association.

b. Suppose you have market basket data consisting of *1000* transactions and *30* items. If the support for *item a* is *70%*, the support for *item b* is *40%* and the support for *itemset {a, b}* is *30%*. Let the support and confidence thresholds be *25%* and *50%*, respectively. Compute the confidence of the association rule *{a}→{b}*. Is the rule interesting according to the confidence measure?

c. A database has four transactions.

TID	Items-Bought
10	{A, C, D}
20	{B, C, E}
30	{A, B, C, E}
40	{B, E}

Find all frequent itemsets using Aprori algorithm with min_sup=2, i.e., any itemset occurring in less than 2 transactions is considered to be infrequent. Also list all of the strong association rules with min_sup=2 and min_conf=100%.

Roll No. ___________Name _________________ Section _CS_

Consider the following description for next Questions# 3 and Question# 4:

Consider the following tables and statistics which are part of a student registration system:
Student (RollNo, Name, gpa, DeptID, BatchID, DegreeID, …..); Attendance (RollNo, CourseCode, Semester, AttFlag, …..);

Assume student and attendance tables containing 64,000 and 640,000 rows respectively (Student:Attendance ratio is 1:10). Each table row and each index entry takes 128 bytes and 8 bytes space respectively. Data block size is 8KB and available memory size is 10 blocks. Suppose degree= 'MS' has a selectivity of 3%, batch= ('2015' or '2014') has a selectivity of (5% + 2%), and dept= ('CS or 'EE') has a selectivity of (40% + 20%).

Q3. *(10 points)*

How many blocks of data need to be accessed to answer the query:

```
SELECT  AVG(gpa)   FROM student  JOIN attendance ON  student.rollno=attendance.rollno
WHERE  DegreeID='MS'  AND  (BatchID='2015' OR BatchID='2014') AND  (DeptID='CS' OR DeptID='EE');
```

Assume cluster indexes exist on RollNo column of student table and also on RollNo column of attendance table. You are supposed to filter the condition first and then join. Examine and use the best possible joining technique. Justify your selection and show all steps clearly.

Q4. *(10 points)*

How many blocks of data need to be accessed to answer the query:

```
SELECT  COUNT(*)   FROM student
WHERE  DegreeID=’MS’  AND  (BatchID=‘2015’ OR BatchID=’2014’) AND  (DeptID=‘CS’ OR DeptID=’EE’);
```

Suppose three secondary indexes are created on student’s attributes *deptID*, *BatchID*, and *DegreeID*. Examine and use the best possible access path. Justify your selection and show all steps clearly.

Q5. *(6+2+2= 10 points)*
Consider the following three dimensions and a fact table:
Customer: customer-ID, Name, gender, city, country, …
Account: account-ID, account-Number, open-Date, account-Type-Code, …
Month: month-End-date-ID, month-Name, calendar-Month, …
Monthly_Account: month-End-date-ID, account-ID, customer-ID, previous-Balance, total-Deposits, total-Withdrawal, available-Balance.

a. Draw the appropriate star schema that includes a base fact table, a 1-way aggregate fact table (along customer dimension), and a 2-way aggregate fact table (along customer and account dimensions). Show the primary keys, foreign keys and all the relationships between the dimensions and fact tables.

b. Identify the full-additive, semi-additive, and non-additive facts, if any, in the above base fact table.

c. Refer to the customer dimension of above star schema. Show the revised customer dimension schema that also preserves the history of changes to the customer.