National University of Computer and Emerging Sciences, Lahore Campus

Course:	Data Warehousing & Data Mining	Course Code:	CS409
Program:	BS(Computer Science)	Semester:	Fall 2016
Duration:	60 min	Total Marks:	30
Paper Date:	14-Nov-2016	Weight	12.5%
Section:	All	Page(s):	4
Exam:	Mid II	Reg. No. (Section)	---------- ( )

Instruction/Notes: **All the questions are to be solved on question paper, write your Roll no on every sheet**

Question1: (5 Points)

What are the two general categories of data stored in source operational systems? Give an example for each.

Question 2: (5 Points)

Assume that 50,000 rows out of 10 million rows in Account dimension table changes on each data refresh. Which loading strategy should you follow? Explain the reasons for your selection. Also suggest some practical steps that expedite the data loading process.

Question 3 (10 Points)

Consider the following tables and statistics which are part of a student system:

Student (RollNo, Name, gpa, DeptID, BatchID, DegreeID, …..); Attendance (RollNo, CourseCode, Semester, AttFlag, …..);

Assume student and attendance tables containing 128,000 and 2,560,000 rows respectively (*Student:Attendance* ratio is *1:20*). Each row and each index entry takes 256 bytes and 16 bytes space respectively. Data block size is 16KB and available memory size is 100 blocks. Suppose degree= 'MS' has a selectivity of 3%, batch= ('2015' or '2005') has a selectivity of (4% + 2%), and dept= ('CS or 'EE') has a selectivity of (10% + 5%).

Query:

```
SELECT AVG(gpa) FROM student JOIN attendance ON student.rollno=attendance.rollno
WHERE DegreeID='MS' AND (BatchID='2015' OR BatchID='2005') AND (DeptID='CS' OR DeptID='EE');
```

Calculate the total I/O cost (including the I/O cost to filter the condition on student table) for the above Query using hash join and block nested loop join techniques. You are supposed to filter the condition first and then join. Show all steps clearly.

Question 4 (3+3+4= 10 Points)

Consider the following tables and statistics which are part of a bank system:

ACCOUNT (accId, title, accType, rating, openingDate, …);

Block Size= 4 KB; Available Memory= 100 Blocks; Rows= 250,000; Row Width= 500 bytes; Index entry size (i.e. RID Width)= 8 bytes. Assume accounts with 'SAVING' accType are 4%, accounts with 'CHECKING' accType are 10%, and accounts with '1' rating are 6%.

Query: SELECT COUNT(*) FROM account WHERE (accType= 'SAVING' OR accType= 'CHECKING') AND Rating= 1

Calculate the I/O cost for the above query using

a) Composite index access (Assume a composite index exist on accType and rating columns)
b) Dynamic Bitmap index access (Assume indexes exist on accType and rating columns separately)
c) Clustered index access (Assume only clustered index exist on accType column)