Roll No. _____Name _____ Section ___

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:	**60 Minutes**	**Total Marks:**	**25**
	Paper Date:	**2-Nov-17**	**Weight**	**12.5%**
	Section:	**CS**	**Page(s):**	**4**
	Exam:	**Midterm-2**		

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. *(3+2= 5 points)*

a) Name four types of the major transformation tasks. Give an example for any of them.

b) Describe briefly the entity identification problem in data integration and consolidation. How do you resolve this problem?

Roll No. _____Name _____ Section ___

Q2. *(10 points)*

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

Car (CarID, Model, Make, Color, …); Sale (SaleID, SalesPersonID, CarID, CustomerID, SalesDate);

Assume car and sale tables containing 20,000 and 1,000,000 rows respectively (*Car:Sale* ratio is *1:50*). Each row and each index entry takes 500 bytes and 8 bytes space respectively. Data block size is 16KB and available memory size is 100 blocks. Suppose make= 'Honda' has a selectivity of 20%, and color= ('White or 'Black') has a selectivity of (40% + 30%).

Query:

*SELECT * FROM car JOIN sale ON car.carID = sale.carID*
WHERE Make='Honda' AND (Color='White' OR Color='Black');

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

Ans:

R=500, Ri=8 B=16K, bfr=32, bfri=2048, K=100, b_{car}=625, b_{sales}=31250, car:sale ratio 1:50

Combine selectivity = 20% of (40+30)% of 20,000 = 2800 rows (88 qualifying blocks)

SMJ:

Filtering Cost + SORT car table + SORT Sales table + Merge Cost

= 625 + 88 + (31250 * log(31250/100)) + (88 + 31250)

= 625 + 88 + (31250 * 9) + (88 + 31250) = **313,301**

HJ:

Filtering Cost of car + Hashing Cost (It's best case of hash join)

625 + (88 + 31250) = **31,963**

Roll No. _____Name _____ Section ___

Q3. *(10 points)*

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

Sale (SaleID, SalesPersonID, CarID, CustomerID, SalesDate);

Block Size= 16 KB; Available Memory= 100 Blocks; Rows= 1,000,000; Row Width= 500 bytes; Index entry size (i.e. RID Width)= 8 bytes. Assume sale with 'S10' salesPersonID are 2%, with 'S12' salesPersonID are 6%, with 'S15' salesPersonID are 1%, with 'H20' carID are 4%, and with 'A30' carID are 2%.

Query: SELECT * FROM sale WHERE salesPersonID IN ('S10', 'S12', 'S15') AND carID IN ('H20', 'A30');

Calculate the I/O cost for the above query using:

a) Combining multiple indexes (Assume indexes exist on salesPersonID and carID columns separately)

b) Static bitmap index access (Assume static bitmap indexes exist on salesPersonID and carID columns)

Ans:

R=500, Ri=8 B=16K, bfr=32, bfri=2048, K=100

combine selectivity (S10, S12, S15) and (H20, A30): 9% of 6% of 1 million = 5400

a) Combining Multiple Indexes:

Index for salesperson (2%+6%+1%)= 20000/2048 + 60000/2048 + 10000/2048 = 10 + 30 +5 = **45**

Index for car (4%+2%)= 40000/2048 + 20000/2048 = 20 + 10 = **30**

Total I/Os = Index cost + Base table cost = 75 + 5400 = **5475**

b) Static Bitmap Indexes:

One bitmap access cost= 8 (i.e. 1m/(16k*8)); total 5 bitmaps are required to access.

Total I/Os = Index cost + Base table cost = 5*8 + 5400 = **5440**

Roll No. _____Name _____ Section ___