# 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: | 3 Hours | Total Marks: | 60 |
| Paper Date: | 26-Dec-2016 | Weight | 40% |
| Section: | All | Page(s): | 8 |
| Exam: | Final | Reg. No. (Section) | ---------- (  ) |

**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.**
**Write your Roll no on every sheet.**
**You will not get any credit if you do not show proper working, reasoning and steps as asked in question statements. Unreadable answers will NOT be graded.**

**Question 1 (2+3+5= 10 Points)**

**a)** How is data mining different from OLAP? Explain briefly.

**b)** Suppose you have market basket data consisting of 100 transactions and 20 items. If the support for item a is 25%, the support for item b is 90% and the support for itemset {a, b} is 20%. Let the support and confidence thresholds be 10% and 60%, 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 |
|---|---|
| T100 | {A, B, D, K} |
| T200 | {A, B, C, D, E} |
| T300 | {A, B, C, E} |
| T400 | {A, B, D} |

Find all frequent itemsets using Aprori algorithm with min_sup=3, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent. Also list all of the strong association rules with min_sup=3 and min_conf=80%.

**Question2: (3+3+3+3+4+4= 20 Points)**

**a)** Discuss the three common sources of data pollution and provide examples.

**b)** What is master data management (MDM) approach? Also list two benefits of MDM.

**c)** List the three common and major types of architectures for building a data warehouse.

**d)** Name any three advantages of using materialized views.

**e)** Name any three data extraction techniques. Which of these are easy and inexpensive to implement? Explain briefly why.

**f)** How does a snowflake schema differ from a STAR schema? Name two advantages of the snowflake schema.

**Question 3 (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 4KB 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 index nested loop join (Assume there is an index on carID column of sale table and three $I/O_s$ are required to read index for each qualifying car). You are supposed to filter the condition first and then join. Show all steps clearly.

**Question 4 (10 Points)**

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

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

Block Size= 4 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 '10' salesPersonID are 2%, with '12' salesPersonID are 6%, with '15' salesPersonID are 1%, with 'H20' carID are 4%, and with 'A30' carID are 2%.

**Query:** SELECT * FROM sale WHERE salesPersonID IN (10, 12, 15) AND carID IN ('H20', 'A30');

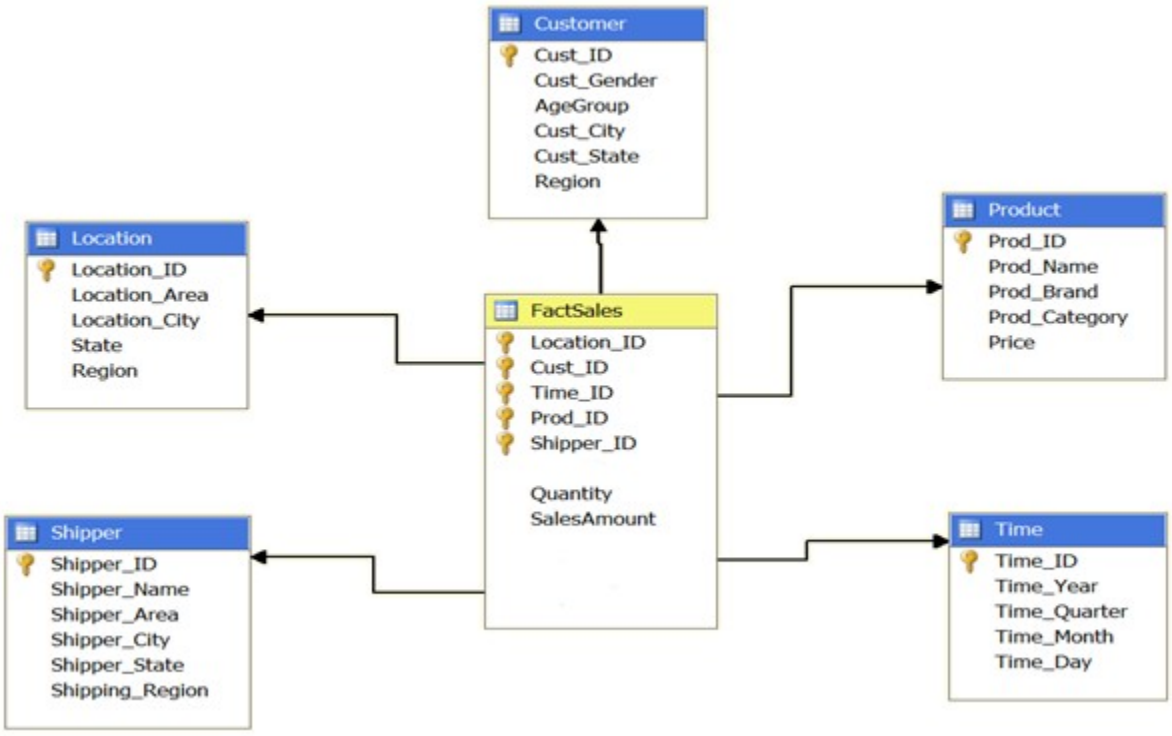
Calculate the I/O cost for the above query using:

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

**b)** Composite index access (Assume a composite index exist on salesPersonID and carID columns)

**Question 5 (7+3= 10 Points)**

Consider the following star schema:




**a)** Create a new star schema that includes a 1-way aggregate fact table (along time_month), a 2-way aggregate fact table (along time_month and cust_city), and a 3-way aggregate fact table (along time_month, cust_city, and prod_category).

**b)** Estimate the size (in rows) of all the above aggregate fact tables. Assuming that each dimension has 150 rows and the fact table records allowable events (i.e. it has a row for every combination of all dimensions). There are 5 different months, customer cities and product categories with uniform distribution among the 150 rows.