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.

Ans: OLAP is used to analyze the past; data mining is used to predict the future.

OLAP is able to give you answers to questions on past performance. Of course, from these answers you can gain a good understanding of what happened in the past. You may make guesses about the future from these answers about past performance. In contrast, data mining can uncover specific patterns and relationships to predict the future.

OLAP Questions:
1- Who are our top 100 best customers for the last three years?
2- Which customers defaulted on their mortgages last two years?

Data Mining Questions:
1- Which 100 customers offer the best profit potential?
2- Which customers are likely to be bad credit risks?

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?

Ans: Confidence= support of {a,b}/support of {a} = 20%/25% = 80%
Rule is also interesting because confidence is greater than 60%.

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%.

Ans: First scan (1-itemsets)

ItemSet	Sup Count
A	4
B	4
~~C~~	~~2~~
D	3
~~E~~	~~2~~
~~K~~	~~1~~

L1

ItemSet	Sup Count
A	4
B	4
D	3

L2 (second scan)

ItemSet	Sup
{A,B}	4
{A,D}	3
{B,D}	3

L3 (third scan)

ItemSet	Sup
{A,B,D}	3

F= {A→B, B→A, D→A, D→B, AD→B, BD→A, D→AB}

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

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

Ans:

System Conversions, Data Aging, Heterogeneous System Integration, Poor Database Design, Incomplete Information at Data Entry, etc.

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

Ans: MDMis an umbrella approach to provide consistent and comprehensive core information across the organization. Master data generally refers to data describing core business objects such as customers, products, locations, and financials. Sometimes data about other entities such as business partners, employees, sales contacts, and physical assets are also included as master data for an organization. These may be considered as nontransactional data entities or reference data.

Benefits:

- Reduction in cost and complexity of processes that use master data and provide internal efficiencies.
- Improvement in the ability to consolidate, share, and analyze business information in a timely manner, regionally and even globally.
- Possibility to rapidly assemble new, composite applications with accurate master data and reusable business processes.
- Reduction in time to market by having a single system for creating and maintaining product information, promotions, and consumer communications.
- Improvements to the supply chain with single, accurate, well-defined definitions of products and suppliers, eliminating duplications.
- Enhanced customer service, with a complete view of each customer designed to better anticipate customer needs and provide targeted offers.

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

Ans:

Centralized Data Warehouse, Independent Data Marts, Federated, Hub-and-Spoke, Data-Mart Bus.

d) Name any three advantages of using materialized views.

Ans:

- Performance improvement
- Automatic consistency between base table contents and mv contents
- Efficient maintenance when rows are inserted/deleted/updated in base table
- Optimizer automatically recognizes opportunities for exe. plan that take advantage of the mv using cost based evaluation methods.
- Transparent from user

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

Ans:

Transaction log, db triggers, source application, date/time stamp, file comparision
Transaction log is easy inexpensive to implement due to no additional development cost and no impact on existing source system.

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

Ans:

Snowflake schema is a normalized structure and star schema is de-normalized structure.

Advantages:

- Small savings in storage space
- Normalized structures are easier to update and maintain

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.

Ans:

R=500, Ri=8 B=4K, K=100, b_{car}=2500, b_{sales}=125000, car:sale ratio 1:50

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

SMJ:

Filtering Cost + Sort car table + Sort Sales table + Merge Cost

2500 + (350 * log(350/100)) + (125000 * log(125000/100)) + (350 + 125000) = **1,503,550**

Indexed NLJ:

Filtering Cost of car + Read Cost of qualifying blocks + (Qualifying rows of car * (Sales index cost + average rows of sales per car))

2500 +350 + (2800 * (3 + 50)) = **151,250**

Question 4 (10 Points)

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

Sale (<u>SaleID</u>, 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)

Ans:

combine selectivity (10, 12, 15) and (H20, A30): 9% of 6% of 1 million = 5400

a) Index for salesperson (2%+6%+1%)= 20000/512 + 60000/512 + 10000/512 = 40 + 118 +20 = **178**

Index for car (4%+2%)= 40000/512 + 20000/512 = 79 + 40 = **119**

Total I/Os = Index cost + Base table cost = (178 + 119) + 5400 = **5697**

b) I/O cost for combination

c1: 2% of 4% of 1 million = 800/512 = 2

c2: 2% of 2% of 1 million = 400/512 = 1

c3: 6% of 4% of 1 million = 2400/512 = 5

c4: 6% of 2% of 1 million = 1200/512 = 3

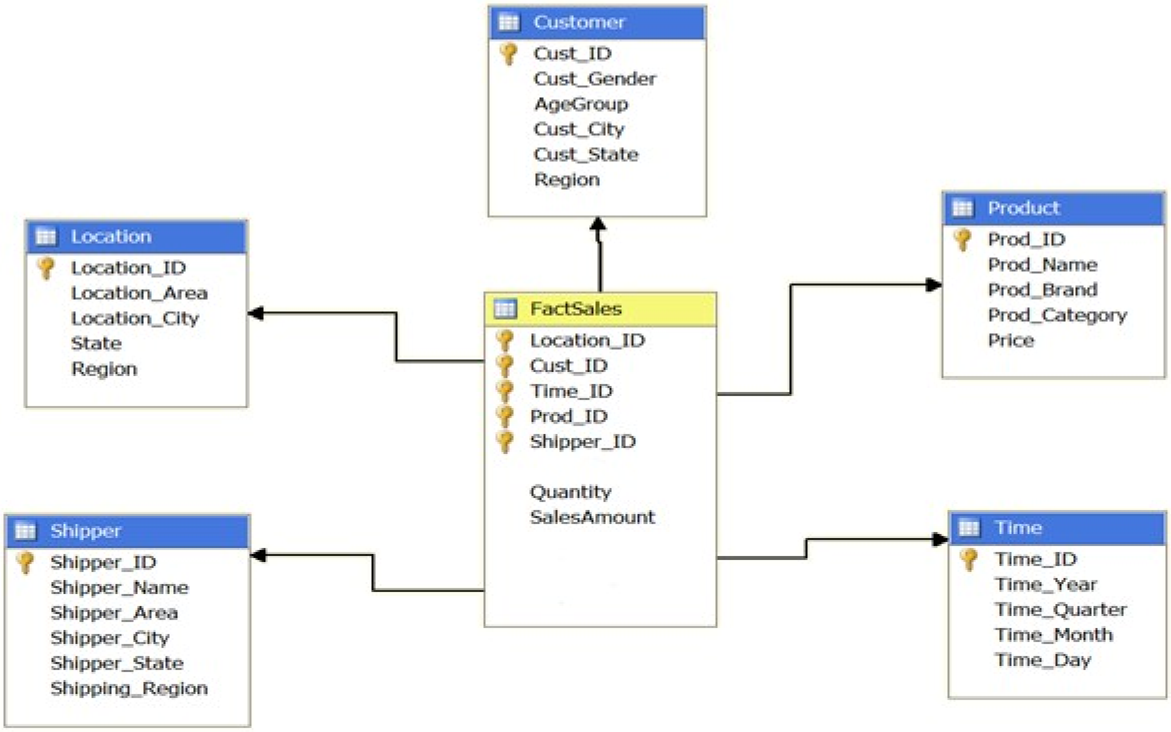
c5: 1% of 4% of 1 million = 400/512 = 1

c6: 1% of 1% of 1 million = 200/512 = 1

Total I/Os = Index cost + Base table cost = 13 + 5400 = **5413**

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.

Ans: b) Size of
1-way aggregate fact table: 150 *150 * 150 * 150 * 5 = 2,531,250,000
2-way aggregate fact table: 150 *150 * 150 * 5 * 5 = 84,375,000
3-way aggregate fact table: 150 *150 * 5 * 5 * 5 = 2,812,500