School of Information Technology
IIT Kharagpur

Course Id: IT60107 Data Warehousing and Data Mining (Mid Sem)

Date: September 18, 2006 **Total Time: 2 Hours**
Max. Marks: 60

Instructions: Answer all questions. You may answer the questions in any order. However, all parts of the same question must be answered together. Clearly state any reasonable assumption you make.

1. a-d are multiple choice type questions. Zero or more options may be correct. 2 marks will be awarded for each correct answer, 1 mark will be deducted for each wrong answer. An answer will be considered correct if all the correct and only the correct options are chosen. If no option is correct, write "None". If you do not want to attempt a question, leave it blank. **[8]**

a. In a star schema, usually
   (i) fact table is de-normalized
   (ii) number of rows in dimension tables is much less compared to that in fact table(s)
   (iii) dimension table is de-normalized
   (iv) dimension tables contain more number of columns compared to fact tables(s)

b. In OLAP, dimension reduction can occur due to
   (i) Roll-up
   (ii) Drill-down
   (iii) Slicing
   (iv) Dicing

c. It is beneficial and practical to materialize all the views in a data cube when
   (i) number of levels in dimensional hierarchies is very large and there are too many dimensions
   (ii) speed of retrieval is the primary objective
   (iii) effective join indexes can be defined between fact table and dimension tables.
   (iv) we can implement a greedy algorithm for solving the view materialization problem

d. In a star schema fact table, a degenerate dimension is a column that is
   (i) a measure which is additive across only some of the dimensions
   (ii) associated with one and only one dimension table
   (iii) not associated with any dimension table
   (iv) a measure which is not additive across any dimension

2. Consider a 3-D data array consisting of the dimensions A, B and C. The 3-D array is partitioned into 72 memory-based chunks. Dimension A is organized into 4-equisized partitions a0, a1, a2 and a3. Dimension B is organized into 3-equisized partitions b0, b1 and b2. Dimension C is organized into 6-equisized partitions c0 … c5. Chunks are numbered as 1, 2, 3, …, 72 corresponding to the sub cubes a0b0c0, a1b0c0, a2b0c0, a3b0c0, a0b1c0, …, a3b2c5, respectively. Assume that the size of the array dimensions A, B and C are 300, 30 and 3000, respectively. If we perform multi-way array aggregation using chunking, then calculate the minimum memory requirement for holding all relevant 2-D partial sums in chunk memory when the chunks are brought into memory in the order: 1, 13, 25, …, 61, 5, 17, …, 65, 9, 21, …, 69, 2,14, …, 62,……., 12, 24,…,72. **[12]**

3. Consider the following business scenario.

A credit card issuing bank (for example, Citibank) wants to build a data warehouse for analyzing purchase behavior of its customers. It may issue one or more cards to the same customer. They would like to track how a customer is behaving as well as how he is using each of the cards issued to him.

Cards issued by the bank can be of several "partnership" types. For example, Citibank issues Indian Oil card which gives some discount on purchases of petrol from Indian Oil petrol pumps. In this case, Indian Oil is a partner of Citibank. Similarly, Citibank also issues Big Bazar Card which gives some discount when used for purchases from Big Bazar. In this case, Big Bazar is a partner of Citibank. It is important to see the performance of these partnerships. A given type of partnership card may, however, be used for purchases from any other shops or from another partnership shop as well. We will use the terms "type" and "partnership type" interchangeably.

For any on-line purchase transaction made on the card, the issuing bank gets to know only the credit card number, the name of the merchant (i.e., the shop/establishment where it was used), the partnership type of merchant (i.e., which partnership type it is and "not applicable" if it is not of any partnership type), the date of purchase, the time of purchase and the total rupee amount of purchase.

One of the goals of the issuing bank data warehouse is to find the comparative amount of purchases made on the different types of cards during different periods of time. The time period could be day, week, month, quarter and year. They would also like to see how the cards are being used during different hours of the day. It is also important to analyze the behavior of customers, both individual as well as based on their age groups, marital status, income group, % literacy of the states to which they belong, male/female ratio of the states to which they belong, etc.

a) Design an efficient data warehouse schema that satisfies the above business scenario. Clearly identify the fact table(s), fact(s), dimension table(s), primary key(s) and foreign key(s). Classify the fact(s) as additive/non-additive/semi-additive.
b) Write an SQL statement that generates the number of married and unmarried customers that the issuing bank has today for each type of card.
c) Write an SQL statement that generates the quarterly amounts of purchase made on each type of card.
d) Draw an OLAP cube to represent the result of your query of Question (c) above.
e) Write an SQL statement to view the total amount of purchase (over Rs. 1 crore) made by married customers during 5:00 PM – 7:00 PM in states where male/female ratio is less than 1.0.
f) Write an SQL statement to return the total purchases made by the customer who owns both card numbers ABC1234 and DEF2345. If the same customer does not own the two cards, it should return zero.
g) Write an SQL statement to find the amounts of purchases made by the customer who owns card no. DEF4567 from merchants of the same partnership type as that of the card and also from merchants not of the same partnership type as the card. [There should be two rows – one for the same partnership type as the card and one for all purchases made from merchants not of the same partnership type as the card. There will be two columns – "partnership type" and "amount". One row will have a value "All others" for the first column to denote purchases made from all other types of merchants including not applicable types. For the other row, the value of this column will be the name of the partnership type of the card DEF4567.]

[16+4+4+4+4+4+4=40]