School of Information Technology
IIT Kharagpur

Course Id: IT60107 Data Warehousing and Data Mining

Date: September 19, 2005 **Total Time: 2 Hours**
Max. Marks: 70

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. One 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. **[8]**

a. In a star schema, usually
   (i) the fact table is normalized
   (ii) the number of rows in dimension tables is usually much less compared to that in fact table(s)
   (iii) the fact table is de-normalized
   (iv) the 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) the number of levels in dimensional hierarchies are very large and there are too many dimensions
   (ii) the speed of retrieval is the primary objective
   (iii) the cardinality of the dimension is high
   (iv) we can implement a greedy algorithm for selecting the views to be materialized

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 two and only two dimension tables
   (iii) associated with no dimension table
   (iv) a measure which is not additive across any dimension

2. A very large tele-communications company called "Cell9", providing cellular phone services to a number of states in various regions of the country, plans to build a data warehouse for decision support. They have millions of subscribers in the country. They want to track the duration (in minutes) as well as the prevailing rate (per minute) of each phone call made by its subscribers. They also want to analyze if there is any link between the total amount of time spent in talking on cellphones by a subscriber and the number of graduates in the state or the number of married persons in the state or the male-female ratio of the state to which the subscriber belongs. Further, they want to analyse the relation between the age, salary and marital status of the customers to their total bill amount per day/month/year. One other important requirement is to make queries like determining the current total number of customers in the various age groups for each state having certain ranges of male-female ratio.

(a) Design a suitable relational database schema for such a data warehouse, clearly identifying the fact table(s), the facts in the fact table(s), the dimension table(s), their primary key(s) and foreign key(s). Your schema should at least be able to satisfy the above mentioned analysis requirements. You may consider other suitable attributes for the dimension table(s).
(b) Classify the facts in your fact table(s) as additive, non-additive and semi-additive.
(c) Draw possible concept hierarchies for each dimension that you have designed, identifying whether these are schema hierarchies or set grouping hierachies.
(d) Write an SQL query that runs on your schema and returns the region-wise yearly average bill amounts of married and unmarried customers.
(e) Draw a cuboid to represent the result of your query.
(f) From this cuboid, which sequence of OLAP operations would you perform to get the average monthly bill amounts of all the customers for the states of Bihar and West Bengal?
(g) Write an SQL query to return the current total number of customers in the various age groups for each state with male-female ratio between 0.9 and 1.1.
(h) For any one fact table (You may have only one, depending on your design), and any one attribute of any one dimension table, draw the bitmap index table(s) and join index table(s). Before drawing the index tables, first mention the representative rows in the tables.

[10+4+4+5+2+5+5+(5+5)=45]

3. 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, 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.

[17]