School of Information Technology  
IIT Kharagpur

Course Id: IT60107 Data Warehousing and Data Mining (Mid Semester Examination)

Date: September 27, 2011
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. For Question No. 2, if you feel, you can show sample tuples in the relations to explain your assumptions. But the assumptions must be valid.

1. a-e 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. [2x5=10]

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

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

   c. It is not beneficial or 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 not the primary objective
      (iii) number of levels in dimensional hierarchies is not very large and there are only a few dimensions
      (iv) storage space is very expensive

   d. In star schema, a non-additive fact
      (i) is not additive across any dimension
      (ii) is additive only across a subset of the dimensions
      (iii) is additive across all the dimensions
      (iv) is the same as a degenerate dimension

   e. In a star schema, using Type 2 slowly changing dimension handling technique
      (i) increases the number of rows in the dimension table
      (ii) increases the number of columns in the dimension table
      (iii) increases the number of rows in the fact table
      (iv) increases the number of columns in the fact table
2. Consider a chain of retail stores having business in India. Their initial analysis requirements include getting to know which products are purchased together in each transaction. They also want to know the sales figures (both in terms of sales amount in Rupees as well as quantity) of the individual stores and also for the city, state and region to which they belong. Further, they want to know how sales vary over different months, quarters and years. Some of the items are sold under promotion. It is important for them to keep track of which products are sold under which promotion scheme and which did not sell even after being put up on promotion.

(i) Design and draw a star/fact constellation schema for such a data warehouse clearly identifying the fact table(s) and dimension table(s), their primary keys and foreign keys. The fact table(s) must be normalized (in 3NF) and it is ok to have de-normalized dimension table(s). Clearly state which columns in the fact table(s) represent dimensions and which of them represent fact(s). Also identify whether the fact(s) is/are additive, semi-additive or non-additive. Your schema should at least be able to satisfy the analysis requirements mentioned in (a) above and also the three queries appearing below.

(ii) Write an SQL statement that runs on your schema and returns the total sales of the western region for each quarter under each promotion scheme.

(iii) Write an SQL statement that runs on your schema and returns the products that were on promotion (any form of promotion) during the month of January 2010 (it is ok to consider a product to be on promotion if it was on promotion at least on one day during this period) but did not sale at all in the city of Mumbai during that entire month.

(iv) Write an SQL statement that runs on your schema and returns the products that were not on any promotion on January 20, 2010 but were sold at least once in the store named “ABC Retail” on that day.

(b) The company later decided that other than the above-mentioned requirements, they need to keep track of the individual customer’s buying habit. They also must know how sales figures change with the hour of the day – e.g., how morning hours sales is different from evening hours sales, etc.; how buying habits of male customers are different from that of female customers; how buying habits of married customers are different from that of unmarried customers; how buying habits of customers vary with their academic degrees (e.g., B.Tech., M.Tech., Ph.D., etc.).

(i) Design and draw the updated star/fact constellation schema for such a data warehouse clearly identifying the fact table(s) and dimension table(s), their primary keys and foreign keys. The fact table(s) must be normalized (in 3NF) and it is ok to have de-normalized dimension table(s). Clearly state which columns in the fact table(s) represent dimensions and which of them represent fact(s). Also identify whether the fact(s) is/are additive, semi-additive or non-additive. Your schema should at least be able to satisfy the analysis requirements of both (a) and (b) above and also the two queries appearing below.

(ii) Write an SQL statement that runs on your schema and returns the total sales of the western region of unmarried male customers carrying out transactions between 10:00 AM-11:00 AM.

(iii) Write an SQL statement that runs on your schema and returns the current number of married female customers who have a Ph.D. degree.