Mid Semester Exam
Course Id: 406035 Data Warehousing and Data Mining

Date: September 18th, 2003.
Total Time: 2 Hours   **Max. Marks: 90**
Answer all questions. Clearly state any reasonable assumptions you make.

1.a-f 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 iff all the correct options are chosen. **12 Marks**

a. In a star schema, usually:
  (i)   the fact table is normalized
  (ii)   the dimension tables are normalized
  (iii)   the fact table is denormalized
  (iv)   the dimension tables are denormalized

b. In multi-dimensional analysis, roll-up can be achieved by:
  (i)   moving up a dimension hierarchy like city -> state, etc.
  (ii)   moving down a dimension hierarchy like state -> city, etc.
  (iii)   adding a new dimension
  (iv)   removing one or more dimensions

c. In a data warehouse schema, if D1 and D2 are two conformed dimensions, then:
  (i)   D1 may be an exact replica of D2
  (ii)   D1 may be at a rolled up level of granularity compared to D2
  (iii)   Columns of D1 may be a subset of D2 and vice versa
  (iv)   Rows of D1 may be a subset of D2 and vice versa

d. A data warehouse bus matrix helps to:
  (i)   integrate data marts
  (ii)   improve speed of retrieval from the database
  (iii)   index OLAP data
  (iv)   generate aggregated information

e. 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

f. In a star schema fact table, a degenerate dimension is a column:
  (i)   that is associated with only one dimension table
  (ii)   that is associated with more than one dimension table
  (iii)   that is associated with no dimension table
  (iv)   that is a measure which is not additive across any dimension

2. Consider the following business scenario. A telecom company plans to maintain a CRM data warehouse. There are 10 million customers of the company. Besides the usual attributes, the company wants to maintain additional demographic information like literacy percentage, male/female ratio, average life expectancy and average income of the people belonging to the state to which each customer belongs. The company also wants to maintain information about the age group, income level and marital status of its customers. They also need to run queries like the number of married and unmarried customers they have at any point in time.

     a. Design an efficient data warehouse schema that satisfies the above business scenario. Clearly identify the fact table(s), dimension table(s), primary key(s) and foreign key(s).
     b. Write an SQL statement that generates the number of married and unmarried customers that the company has today.

15+3=18 Marks

3. Consider a hypothetical sales fact table that contains the columns item_code and state_code as its dimensions. The corresponding dimension tables are also shown below.

Sales fact Table

RID	Item_code	State_code
1	A	AB
2	A	CD
3	B	AB
4	C	CD
5	D	XY
6	D	AB

Item Dimension Table

Item_code	Other Columns
A	
B	
C	
D	

State Dimension Table

State_code	Other Columns
AB	
CD	
XY	

     a. Draw the bitmap index tables.
     b. Draw the join index tables.

10+10=20 Marks

4. Consider a 3-D data array consisting of 3 dimensions A, B and C. The 3-D array is partitioned into 64 memory-based chunks. Dimension A is organized into 4-equisized partitions a0, a1, a2 and a3. Similarly dimensions B and C are also organized into 4-equisized partitions each. Chunks are numbered as 1, 2, 3, …, 64 corresponding to the sub cubes a0b0c0, a1b0c0, a2b0c0, a3b0c0, a0b1c0, …, a3b3c3, respectively. Suppose the size of the array of the dimensions A, B and C are 300, 3,000 and 30,000, respectively. If we perform multi-way array aggregation, then what is the minimum memory requirement for holding all relevant 2-D partial sums in chunk memory, if the chunks are brought into memory in the order: 1, 17, 33, 49, 5, 21, …, 13, 29, 45, 61, 2, 18, …

20 Marks

5. Consider the following lattice of views along with a representation of the number of rows in each view where A is the base cuboid:

If you have to choose 3 views to materialize apart from the base cuboid, which of the views B-H would you choose and how? Assume that the cost of running a query is linearly proportional to the number of rows in the view from which it is run.

20 Marks