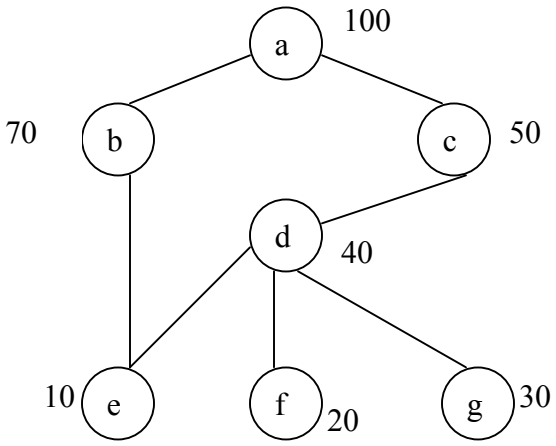
School of Information Technology
IIT Kharagpur

Course Id: IT60107 Data Warehousing and Data Mining

Date: August 30, 2006 **Total Time: 60 minutes**
Max. Marks: 40

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. Assume that the following figure represents a lattice of cuboids defined wrt a binary relation ≤ where x ≤ y denotes view x can be derived from view y. The size (in no. of rows) is marked against each cuboid. Determine a set of three cuboids other than the base cuboid (a is the base cuboid) that you would materialize to get maximum improvement in query response time with respect to no materialization. It is given that the cost of running a query is proportional to the number of rows in the view on which the query is run. **[10]**



2. Consider a chain of retail stores having business only in India. Their analysis requirements include getting to know which products are purchased together by each individual customer. They 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. They also want to know how sales varies over different months, quarters and years; 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 unmaried customers; how buying habits of customers vary with their native languages (e.g., Hindi, Bengali, Tamil, etc.).

(a) Design a star schema for such a data warehouse clearly identifying the fact table and dimension tables, their primary keys and foreign keys. Clearly state which columns in the fact table represent dimensions and which ones represent fact. Your schema should at least be able to satisfy the above mentioned analysis requirements and also the queries appearing below.

(b) Write <u>one</u> SQL statement that runs on your schema and returns the amount of purchases made during the evening hours by the married customers and the unmarried customers in the month of May 2005.

Note: your query must return two rows – one for married customers' amount of purchases during all the evening hours in the month of May 2005 added over all the stores for all the products and the other is for unmarried customers. The two rows should be distinguishable as to which one is for married customers and which one is for unmarried customers.

(c) Represent the result of the query in (b) in the form of an OLAP cube.

(d) From the OLAP cube of (c) above, which sequence of primitive OLAP operations is required to get the OLAP cube that represents (customer native language)-wise, (store region)-wise sales during each quarter of the year 2005?

[15+8+2+5=30]