CS392 Database System Concept

Assignment 9

Due May 15th, 2014

1. What factors could result in skew when a relation is partitioned on one of its attributes by:
   a. Hash partitioning
   b. Range partitioning?

   In each case, what can be done to reduce the skew?

2. Recall that histograms are used for constructing load-balanced range partitions. Suppose you have a histogram where values are between 1 and 100, and are partitioned into 10 ranges, 1-10, 11-20, …, 91-100, with frequencies 15, 5, 20, 10, 10, 5, 5, 20, 5 and 5 respectively. Give a load-balanced range partitioning function to divide the values into 5 partitions.

3. Show how to express **group by cube**(a, b, c, d) using **rollup**; your answer should have only one **group by** clause.

4. Given relation r(a, b, c), show how to use the extended SQL features to generate a histogram of c versus a, dividing a into 20 equal-sized partitions (that is, where each partition contains 5 percent of the tuples in r, sorted by a).

5. Consider that *balance* attribute of the *account* relation. Write an SQL query to compute a histogram of *balance* values, dividing the range 0 to the maximum account balance present, into three equal ranges.

6. Construct a decision-tree classifier with binary splits at each node, using tuples in relation r(A, B, C) shown below as training data; attribute C denotes the class. Show the final tree, and with each node show the best split for each attribute along with its information gain value.
   (1, 2, a), (2, 1, a), (2, 5, b), (3, 3, b), (3, 6, b), (4, 5, b), (5, 5, c), (6, 3, b), (6, 7, c)