Advanced Database Systems - G22.2434 - Fall 2008

Professor: Dennis Shasha

Homework 2 - Due: Tuesday, November 18, 2008

Each question is worth 20 points. You may work with one partner and sign both of your names to your paper.

1. Design a program in your favorite programming language that can generate tuples where field values can be unique, have a fixed set of distinct values that are allocated to tuples according to a uniform probability distribution, and a fixed set allocated to the tuples according to a fractal probability distribution (70-30 rule). Here is a method to do this:

\[
\text{gen}(\text{frac}, N)
\]
\[
\text{begin}
\]
\[
p := \text{random permutation of numbers from 1 to } N
\]
\[
\text{outvec} := p \quad // \text{so outvec is of length } N
\]
\[
\text{while}(|p| > 1)
\]
\[
p := \text{first } \text{frac} \times |p| \text{ elements of } p
\]
\[
\text{concatenate } p \text{ to outvec}
\]
\[
\text{end while}
\]
\[
\text{return random permutation of outvec}
\]

Call it with gen(0.3, x) where x is 70,000 or so. The idea is to get to 100,000.

Populate table sales(id, itemid, quantity)
such that there are 100,000 ids, 20,000 itemids distributed with uniform probability among the sale ids. Then quantities are distributed in a fractal manner.

Then write one or more queries that find the ten items having the greatest total quantities (sum of the quantities grouped by item) and the ten items having the least total quantities. Use any relational database on any operating system you like. Show us the queries and then a terminal history session showing execution and the results. Submission by email or on hard copy.

2. Choose one or two rules of thumb from the book (e.g. having to do with chopping, indexes, commits, checkpoints) and find data distributions where they are satisfied and ones where they are not. Comment on whether the case where the rule of thumb is satisfied is more likely than when it isn’t. Also see how you would modify the rule of thumb to make it more precise.