Program for Generating Skewed Data Distributions for TPC-D

Surajit Chaudhuri, Vivek Narasayya

The schema and queries of the TPC-D benchmark are widely used by people in the database community. One of the requirements of the benchmark is that data for columns in the database are generated from a *uniform* distribution. However, this requirement makes it hard for users to conclude about the robustness/effectiveness of their system using only TPC-D. We have therefore created a new data loading program for TPC-D that is capable of generating a database where the columns have non-uniform (skewed) data distributions. In particular, the loading program can generate data from a Zipfian distribution, where the Zipf value (**z**), which controls the degree of skew in the data, is a parameter that can be specified to the program. This program supports any real value of z >= 0. For example, the parameter z=0 generates a uniform distribution for each column in the database, whereas z=4 generates a highly-skewed distribution (i.e., a few values occur very frequently) for each column. In addition, the program allows the generation of a database with "mixed" data distribution where the skew of a column in the database is randomly chosen from the Zipfian values {0,1,2,3,4}. Note that the total number of rows in the tables and the total database size are not affected by our changes.

Usage

The program *dbgen.exe* in this directory generates TPC-D data with skewed distributions. We refer the user to the Readme file in this directory which explains each parameter to *dbgen* in detail. Here we describe the additional parameter that we have introduced:

dbgen.exe <other_parameters> **–z <zipfValue >**

zipfValue is any real value >=0. The program generates data for *each column* in the database from the Zipfian distribution with the specified z value. If zipfValue is –1, the program assigns a random skew value from the set {0, 1, 2, 3, 4} to each column in the database; and generates data for that column with the assigned skew value. Therefore a parameter value of –1 option allows the user the ability to create a database where different columns have different degree of skew. If the –z option is not used, data is generated for all columns from a uniform distribution (i.e., it is equivalent to –z 0).

Examples

1. dbgen.exe –s 0.1 –z 2

The above command creates a TPC-D database with scaling factor 0.1 where data in each column is generated from a Zipfian distribution where the degree of skew (z) is 2.

2. dbgen.exe –s 0.1 –z 1.6

This example shows that z need not be an integer. In this example, each column has a skew of 1.6.

3. dbgen.exe –s 1.0 –z -1

The above command creates a TPC-D database with scaling factor 1.0 where data in each column is generated from a Zipfian distribution where the degree of skew is picked uniformly at randomly from the set {0,1,2,3,4}.

Source files

No new source files have been added for the above modification. The files that have been modified from the original version are: dss.h, rnd.h, rnd.c, driver.c, build.c, and print.c.

Building dbgen.exe

A copy of **dbgen.exe** already exists in this directory. If you wish to re-build the executable or make other changes to the code, note that no additional compile time flags are required due to this change.

Other Notes

1. These changes affect each column in the database where the data was previously being generated from a uniform data distribution.
2. Because of the way rows in orders and lineitem are generated, the number of rows in the lineitem table can be slightly different, (a few rows) from that using the old generation program.
3. The file DistrTest.sql contains queries that output the distribution (i.e. count of each distinct value) of various columns in the database.