Assignment 1

Uploading Data to the Database

Database Tuning

Start date: October 06, 2016
Due date: October 18, 23:59
Grading: 5 points

Note: This assignment involves reading the documentation of Java, the JDBC driver, and PostgreSQL. Finding the relevant sources of information is part of the challenge.

1. Download `dblp.zip` at:

   http://dbresearch.uni-salzburg.at/downloads/teaching/2016ws/dbt/dblp.zip

   This archive contains two tab separated files (`publ.tsv` and `auth.tsv`) that store authors and their publications as found in the DBLP[1] bibliography. The imported tables have the following schemas:

   - `Auth(name(49),pubID(129))`
   - `Publ(pubID(129),type(13),title(700),booktitle(132),`
     `year(4),publisher(196))`

   You can assume that all attribute values are strings; the maximum string length is shown in brackets.

2. The straightforward algorithm to load the data from the TSV file to a table issues an SQL INSERT query for each line in the TSV file.

   Task 1: Implement the straightforward approach to load `auth.tsv` to the database (PostgreSQL, Java).[2]

   Task 2: The straightforward approach is slow. There are other approaches that are significantly faster. Figure out how the efficient approaches work and implement two of them.

   Report: Describe the two efficient approaches. Give the runtime for loading `auth.tsv` with the straightforward and the efficient approaches. Why are the efficient approaches faster? Which tuning principle did you apply?

[1] http://www.informatik.uni-trier.de/~ley/db

[2] This might be very slow. Instead of loading all the data with this approach, you can also load part of the data and assume that runtime scales linearly. Mention this in the report.

Access parameters for PostgreSQL:
host: `biber.cosy.sbg.ac.at`, port: 5432
db: `dbtuning_ws2016`
user/password: you should have received them via email

The database server (`biber`) is accessible only from inside the university network. If you would like to work from home, please connect to `fanny.cosy.sbg.ac.at` via ssh. Java and the PostgreSQL client as well as Python are installed on this machine.

Please indicate the average time per group member that was spent solving this assignment. The time that you indicate will have *no* impact on your grade.

Grading scheme:

Category	**max. Points**
Description of your setup	0.5
Why is the straightforward approach slow?	1.5
Why is your efficient approach 1 fast?	1.5
Why is your efficient approach 2 fast?	1.5
Bonus: You found the most efficient approach	0.5