Assignment 5
Join Tuning
Database Tuning

Start date: Dec 1, 2016
Due date: December 13, 23:59
Grading: 5 points

In this assignment you will experiment with different join algorithms in PostgreSQL.

1. Download `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. `Publ.pubID` is a key.

2. Study index nested loop join, merge join, and hash join for the following queries:

```
SELECT name,title
FROM Auth, Publ
WHERE Auth.pubID=Publ.pubID;

SELECT title
FROM Auth, Publ
WHERE Auth.pubID=Publ.pubID AND Auth.name='Divesh Srivastava'
```

   (a) What join strategies does the system propose (i) without use of an index, (ii) with a unique non-clustering index on `Publ.pubID`, and (iii) with two clustering indexes, one on `Publ.pubID` and the other one on `Auth.pubID`?

   (b) Test the index nested loop join with a non-clustering index (i) on `Publ.pubID`, (ii) on `Auth.pubID`, (iii) and both `Publ.pubID` and `Auth.pubID`. Give the response times and discuss the query plans.

   (c) Test the merge join (i) without index, (ii) with two non-clustering indexes, and (iii) with two clustering indexes. Give response times and discuss the query plans.

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

(d) Test the hash join without index and give the response time.

(e) Are the results (query plan and throughput) expected? Why (not)?

Note: You can stop queries that run for more than 10 minutes on `biber`. Check the query plan to avoid queries with excessive runtime.

Notes about PostgreSQL

- *Clustering indexes*: You first create an index, then you use the index to cluster the table (i.e., physically sort the table by the index attribute):

  ```
  CREATE INDEX year_idx ON publ(year);
  CLUSTER publ USING year_idx;
  ```

- *Query plan*: The command `EXPLAIN` shows the query plan without executing the query. The command `EXPLAIN ANALYZE` also executes the query. Example:

  ```
  EXPLAIN ANALYZE SELECT * FROM publ WHERE year='2006';
  ```

- *Join strategy*: You can influence the optimizer choice with the switches `enable_hashjoin`, `enable_mergejoin`, and `enable_nestloop`. Example:

  ```
  SET enable_hashjoin TO true;
  SHOW enable_hashjoin;
  ```

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
Join strategies (2a)	0.5
Response times (2b-2d)	0.5
Query plans discussion (2b-2d)	1.5
Interpretation of results	2