

Assignment 5

Join Tuning

Database Tuning

Start date: Dec 11, 2014

Due date: Jan 06, 2015, 23:59

Grading: 1 point

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

1. Download <http://www.cosy.sbg.ac.at/~augsten/teaching/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 DBLP¹ 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 `dumbo.cosy.sbg.ac.at`. 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.