Assignment 5

Join Tuning

Database Tuning

Due date: May 22, 2019, 23:55
Grading: 5 points

Notes

- It is suggested that you also have a look at the report template before you start working on the assignment.

Access Parameters for PostgreSQL

- Host: `biber.cosy.sbg.ac.at`
- Port: `5432`
- Database: `dbtuning_ss2019`
- User/Password: you should have received them via email

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

Support

If there are any ambiguities or problems of understanding regarding the assignment, you have the following possibilities to clarify them:

- Slack channel **#dbt2019**[1] (preferred way of communication)
- Upon request via email (dkocher@cs.sbg.ac.at)

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

Download `https://dbresearch.uni-salzburg.at/downloads/teaching/2018ss/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[2] 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.

[1] `https://dbteaching.slack.com`
[2] `http://dblp.uni-trier.de/db/`

Join Strategies

Study indexed nested loop join, sort-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'
```

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

Report

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

   Give the response times and discuss your observations. Is the choice of the strategy expected? How does the system come to this choice?
2. Test the indexed nested loop join with a non-clustering index
   a) on `Publ.pubID`,
   b) on `Auth.pubID`, and
   c) both `Publ.pubID` and `Auth.pubID`.

   Give the response times and discuss the query plans. Are the response times expected? Why (not)?
3. Test the sort-merge join
   a) without index,
   b) with two non-clustering indexes, and
   c) with two clustering indexes.

   Give the response times and discuss the query plans. Are the response times expected? Why (not)?
4. Test the hash join without index. Give the response times and discuss the query plans. Explain the response times of the hash join in comparison to the response times of sort-merge and indexed nested loop join.

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).
  Example:

  ```
  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's 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 (1.)	0.5
Response times (2. - 4.)	0.5
Query plans discussion (2. - 4.)	1.5
Interpretation of results	2

Reminder: Additional questions about the involved topics/techniques will be asked during the meeting.

Important: If the grading scheme is unclear, ask the instructor!