Assignment 4

Index Tuning – Selection

Database Tuning

Due date: May 8, 2020
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_ss2020`
- 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:

- Upon request via email (martin.schaeler@kit.edu)

In this assignment you will experiment with indexes using PostgreSQL 9.6.

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[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.

[1] `http://dblp.uni-trier.de/db/`

Comparison of Index Types

Compare

1. clustering B^+ tree,
2. non-clustering B^+ tree,
3. non-clustering hash index, and
4. table scan (no index)

for the following queries and measure the throughput.

Repeat the queries multiple times with different conditions for `pubID`, `booktitle`, and `year`, respectively.

```
/* Point Query */
SELECT * FROM Publ WHERE pubID = ...

/* Multipoint Query - Low Selectivity */
SELECT * FROM Publ WHERE booktitle = ...

/* Multipoint Query with IN predicate - Low Selectivity */
SELECT * FROM Publ WHERE pubID IN (List of, say, three authors)

/* Multipoint Query - High Selectivity */
SELECT * FROM Publ WHERE year = ...
```

Report

1. Explain your experimental setup, i.e., how did you send the queries to the database and how did you measure throughput?
2. Which conditions did you use for each of the query types (`pubID, booktitle, year`)? Use the same conditions for all index settings on a particular query type.
3. Give the throughput results and the query plan for each query type and each index setting.
4. Discuss your observations. Are the results expected? Why (not)?

Notes about the Experimental Setup

- To ensure statistical soundness of the throughput, compute the average runtime per query, the overall runtime should be *more than one minute*, and the client program generating the queries should run on the server to avoid network latencies.
- When you repeat a query to measure the throughput, do not use the same condition in the `WHERE` clauses of the repeated queries. Instead, use *different* conditions for each call, for example, `year = 1980`, `year = 2001`, `year = 2004`, and so on. **Briefly state in report (once) how you ensured this and why this is important.**
- Do *not* specify a primary key, a foreign key, or uniqueness constraints when you create the tables. PostgreSQL automatically creates an index to ensure uniqueness, which you want to avoid for some of the queries.
- To test the non-clustering indexes, cluster the table according to an attribute that is independent of the indexed attribute, e.g., cluster the table according to `title` for the condition on `year`.

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';
```

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 (1.)	1
Selection of conditions (2.)	0.5
Throughput results and query plan (3.)	1
Interpretation of query plans	1
Interpretation of the results	1.5

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