Group Project

Michail Kazimianec
Werner Nutt
Damiano Somenzi

Coursework Instructions Milestone 5: Indexing and Query Execution

The goal of the fifth (and last) part of your IDA project is to

- speed up query execution by adding indexes to your database;
- create a test database to investigate whether the indexes are being used and which is their effect on performance;
- find out which collection of indexes is the best one for your application.

For this milestone, you will have an occasion to apply the material of the lectures on file organisation and indexing, query execution, and on indexes and query plans in PostgreSQL.

Work on this milestone is not needed for the normal 30% weighting of the coursework for the final mark. It will be considered an extra and increases the weight by which the coursework will be considered by another 8%.

Selection of the Query to Optimise

Identify a query in your application that is likely to be run frequently and that is costly to execute. For the analysis to be worthwhile, the query should have at least two joins and one selection or one join and two selections. This may be a query that you have formulated for the last two milestones, it may as well be a new query. In reality, one wouldn't probably consider a single query, but rather a query pattern (e.g., a prepared statement in JDBC).

Creation of a Test Database

Create a test version of your database for which you generate random data. You may want to use the Java programs prepared for the labs as a starting point for your own data generation programs. Load the data into tables with the PostgreSQL bulk loader, calling the `\copy` command of `psql`. command.

Identification of Possible Indexes

Identify different options to support the execution of the query by indexes. For each option,

- specify which indexes need to be created for which attributes,
- describe how you expect the indexes to be used by the query engine, that is, write up an execution plan, either textually or graphically, that would benefit from those indexes.

For your query, you should explore at least two design options.

Testing the Design Options

First, run your query on the test database without indexes and measure the time (`EXPLAIN ANALYZE` may be useful for doing this).

Then create the indexes for the first option. Use `EXPLAIN`, possibly in pgAdmin, to check whether the indexes are employed in the way you have foreseen. If this is the case, take measurements. If not, modify your index design.

Then drop the indexes and test the second option in a similar way.

If you have chosen a parameterized query, make sure that you test the design for a sufficient number of paramenters. It may be the case that the optimiser chooses different plans, depending on the parameter.

Based on your findings, explain which is the best choice of indexes for your query.

It is not a problem if your test results are different from what you expected. The point of the exercise is to explore how the query optimiser works.

Deliverable

The deliverable will be a document consisting of four parts:

1. a section introducing the query you want to optimise and a short explanation why it is important;
2. a section describing the test database: size of relations, number of different values for an attribute, the choice of the range of values, etc.;
3. a section presenting the design options you explore;
4. a section reporting on the test results.

Deadline and Submission

The work is to be submitted by publishing it on the web site of your IDA group. The deadline is

Monday, 1 February 2009, 7pm.