CS 3380 Lab Assignment 7

1 Directions

This assignment must be completed by **Sunday, March 15 at 11:59 PM**. You must upload your SQL file to Blackboard. The uploaded file should be named `lab7.sql`. You should **not** host your code in your Babbage account for this assignment. Late submissions will not be accepted.

2 Tasks

2.1 Download

Begin by downloading an SQL dump file by executing the following commands in your terminal:

```
mkdir ~/cs3380/lab7
cd ~/cs3380/lab7
wget http://babbage.cs.missouri.edu/~klaricm/ss15/cs3380/lab7/banks.csv
```

Note that you might not be able to copy-paste the above commands. You may need to type them manually into your terminal.

Next, run the `psql` command to login to your database. Then issue the following commands to load the data into your database:

```
CREATE SCHEMA lab7;

SET search_path = lab7;

CREATE TABLE banks (
    id integer PRIMARY KEY,
    is_active boolean,
    name text,
    established date,
    insured date,
    last_updated date,
    address text,
    city varchar (50),
    state varchar(30),
    assets numeric,
    deposits numeric,
    ots_region varchar(15),
    offices integer,
    offices_domestic integer,
    officies_foreign integer,
    fed_district varchar(15)
);

\copy banks FROM ‘banks.csv’ CSV HEADER
```

Note, you only need to do this one time. After the data has been loaded once, you can return to your database at any time and it will be loaded in the `lab7.banks` table.

2.2 Inspect the Data

The data for this lab contains information about banks insured by the United States FDIC. Recall, to write SQL queries that reference tables held within schemas simply qualifying the table name with the schema name. A simple example follows.

```
SELECT * FROM lab7.banks LIMIT 1;
```

2.3 Implementation

You are responsible for answering 7 multi-part questions in this lab exercise dealing with database indexing and query plans. For all questions that require that you use the `EXPLAIN` command, you should use the `ANALYZE` option. This causes the statement in question to actually be executed leading to the most accurate statistics possible.

You'll be responsible for submitting a SQL file for this assignment, however several of the questions ask that you also provide written answers for questions or copy-pasted content of query plans. All of this "human readable" content should be contained within comments. See section 4.1.5 for information on the comment syntax: http://www.postgresql.org/docs/9.1/interactive/sql-syntax-lexical.html#SQL-SYNTAX-COMMENTS

The questions for this assignment follow:

1. Run the following EXPLAIN command and look at the query plan that is shown. Even though we haven't explicitly created an index, describe why the query plan shows than an index will be scanned. Where did this index come from?

   ```
   EXPLAIN ANALYZE SELECT id, is_active, assets, name FROM banks WHERE id = '17317';
   ```

2. First, write the query that returns all banks in the state of Missouri. Show this query and it's query plan (from the EXPLAIN command) in your answer. Then, write the command that creates an index for the "state" field in table and execute it. Next, rerun your query for all Missouri banks again. Finally, indicate how much faster this search is with the new index now in place in milliseconds and in percent form. Also, record the new query plan. In summary, your answer should contain 5 things: (1) the query for Missouri banks, (2) the query plan for this query, (3) the index creation command, (4) the query plan now that an index has been added, and (5) the speed up in ms and % form.

3. Now, write a query that returns all banks ordered by their names. Include the query plan for this command in your answer. Then, create an index on the name field in the table and show the command in your answer. Finally, re-run the statement that returns all banks sorted by their name. Include the query plan in your answer and the speedup. Your answer for this should contain the same items required for question 2.

4. Perhaps we want to be able to filter our searches based on whether or not a bank is "active". Create an index on the data in that field as well.

5. After creating the index in the previous question, which of the following two queries uses an index? Which not? Use EXPLAIN to determine the answer. Also, describe reason that an index is used when executing one of these but not the other, even though they are almost identical.

   ```
   SELECT * FROM banks  WHERE is_active = TRUE;
   SELECT * FROM banks  WHERE is_active = FALSE;
   ```

6. Write a query that returns all banks with an "insured" date on or after 2000-01-01. Add this query to your SQL file and also generate it's query plan and add that to your file. Recall from history class, that the federal government insuring bank deposits during the Great Depression. This means that our dataset will show many banks with an insured data of 1934-01-01. Create an index on the "insured" field, but make your index exclude all records with a value of 1934-01-01 in that field. Finally, re-run

your search (the one with the insured date on or after 2000-01-01) such the new index is used. Record the query plan and note the speedup. Your answer for this should contain the same items required for question 2.

7. The federal government has an interest in tracking which banks have a low asset to deposit ratio. Write a query that returns the id, name, city, state, assets, and deposits for all banks with an asset/deposit ratio less than 0.5. (Note, that you will have to exclude records that have a 0 value in the deposit field.) Include the query plan for this command in your answer. Then, create an index on this asset/deposit expression, again excluding records with a 0 value in the deposit field. Include your index creation command in your answer. Then, re-run your query that finds banks with an asset/deposit ratio of less than 0.5. Include its new query plan and speedup. Your answer for this should contain the same items required for question 2.