COMPUTING & DATA ANALYTICS 5540 – Assignment 3

This is a TEAM assignment, where teams are expected to be the same as the PROJECT teams.

This assignment is accompanied by data files and scripts, which can be copied from

```
/home/course/u00/sk_public/as3_data+scripts.zip
```

and for which we have the following.

1. The script "make_person_table.sql" creates the MySQL table "person" and inserts tuples using the 20,000 person lines from the data file "persons.tsv".

2. The script "make_history_table.sql" creates the MySQL table "history" and inserts tuples using the event lines from the file "phistory.tsv".

3. The python script "make_person_history.py" can be used to generate any number of events about the 20,000 persons (technically about persons with id's between 1 and 20,000). If **N** is the number of events given as input, the script generates N event lines and saves them in the data file "phistory.tsv". Note that the data files "cities.txt" and "countries.txt" are used by the script to generate city and country data.

Your first task is to test the effect of an index on the performance of the query:

```
select lname, fname from person, history as H
where _id=pid and eyear=2000 and H.city='Las Vegas';
```

Test the query without an index on history(city) and then with an index on history(city). Start with the given file "phistory.tsv" (made with **N** = 300,000) and then repeat for several increasing values of N. Make a table in which you record, for each N, the milliseconds for the search with and without an index. Do this for at least 10 values of N, where the increase should be as large as possible.

Your second task is to explore the effect of an index when inserting new data in a table. Make a second history file "phistory2.tsv" of 1,000 event lines. Make a transaction that inserts the 1,000 event lines when the index exists, and a transaction when it does not exist, and record the times required for these insertions. Do that for all the values of N you have chosen above. In all cases rollback the transactions so that, in the end, the insertions have no effect.

Your third task is to explore the effect of an index when deleting existing rows in a table. This time you are supposed to devise the experiment yourselves. Repeat it for all the values of N you have chosen above. In all cases rollback the transactions so that, in the end, the deletions have no effect. You should consider two types of deletions: one irrelevant to the index field and one based on the index field.

Of course, it makes sense to **do** the three **tasks above together** for each N. As the amount of data and computation time might be large, you are **not** supposed to do the above tasks on the Linux server dbcourse.cs-smu.ca. Do them **on your laptops**.

YOUR REPORT: write a report that contains (i) your team number, the team's names and IDs, title and a brief introduction of what the report is about; (ii) your measurements in table form; (iii) a few paragraphs explaining your findings about the performance of the index and a clear description of the method you used to test the effect of index on deletions; (iv) snapshots of the execution of scripts/MySQL-statements used in the above tasks. (v) It is absolutely fine to use AI systems, in which case you are supposed to have a paragraph stating which AI systems you used (including version) and how you used them.

IMPORTANT COMMENTS

1. Late submissions will not be accepted.
2. Submission procedure: **One** submission per team. Submit in the Brightspace dropbox one pdf file and state in your submission message your TEAM NUMBER.
3. Teams are not allowed to share solutions, but they are allowed to exchange ideas verbally without taking notes during these exchanges.
4. While you are not required to submit code, you might be asked to demonstrate that your code actually works, if we feel this is necessary.