CS315A: Principles of Database Systems

Assignment: Marks = 100

Due Date: 1 April, 2021, 6:00pm

Study the following database schema:

```
A (A1 integer, A2 string, primary-key(A1))
B (B1 integer, B2 integer (foreign-key = A1), B3 string, primary-key(B1))
```

Use the data given as part of this assignment to construct four database implementations using this schema:

1. SQlite3
2. MariaDB (without index)
3. MariaDB (with index)
4. MongoDB

Use the 9 databases according to the following scheme.

Take the *last* 3 digits of your roll number. Let it be a, b, c where c is the least significant digit or the last digit.

Create the following set of 9 integers:

$(a \times a)/5$; $(a \times b)/5$; $\ldots$ $(c \times c)/5$;

Use these numbers to pick up databases according to `B-100-3-?.csv`, `B-100-5-?.csv`, `B-100-10-?.csv`, `B-1000-5-?.csv`, `B-1000-10-?.csv`, `B-1000-50-?.csv`, `B-10000-5-?.csv`, `B-10000-50-?.csv`, `B-10000-500-?.csv` where ? are the corresponding numbers, and the appropriate `A-x.csv` files.

Pick up in order the files as shown in the example below.

For example, if roll number is `*123`, the student should use the following data files:

(The 9 numbers are $(1 \times 1)/5 = 1$, $(1 \times 2)/5 = 2$, $\ldots$, $(3 \times 3)/5 = 4$.)

```
(A-100.csv,B-100-3-1.csv),
(A-100.csv,B-100-5-2.csv),
(A-100.csv,B-100-10-3.csv),
(A-1000.csv,B-1000-5-2.csv),
(A-1000.csv,B-1000-10-4.csv),
(A-1000.csv,B-1000-50-1.csv),
(A-10000.csv,B-10000-5-3.csv),
(A-10000.csv,B-10000-50-1.csv),
(A-10000.csv,B-10000-500-4.csv).
```

Test the following queries:

(a) Q1: Find all A with $A1 \leq 50$.

(b) Q2: Find all B in sorted order of $B3$.

(c) Q3: Find average number of values per $A1$ by using only B table.

(d) Q4: Find all $A2$ that corresponds to B by using $B2$ (output the fields of B and $A2$).

Do the following:

(1) [20 marks] Write the equivalent queries in SQL and MongoDB query languages.

(2) [35 marks] Mention the times taken for each of the queries and each of the database implementations (4 * 4 * 9).

You should output a table with 9 columns as the 9 databases and 4 rows with 4 sub-rows each for the 4 queries and the 4 databases.

Take at least 7 runs per query at *different* times (not consecutively). Get rid of the most and least time consuming runs, and report the average and standard deviation of the remaining 5 runs.

For the largest database files, you may use only 5 runs.

(3) [15 marks] Draw *appropriate* graphs per query and per size and factor of the database implementations.

(4) [20 marks] What do you conclude? Include conclusions about scalability of queries, databases, implementations as well as system issues. Write a report. Clearly mention the O/S (with details) you have used, and include the machine configuration (which CPU, what speed, how much RAM, how much HDD/SSD storage, etc.).

(5) [10 marks] Include all the scripts necessary to run the programs. Include the data loading scripts, the queries, the tables, the graph output generators, the graphs, etc. Please do *not* include output tuples. Zip all of these in a file named `rollno.zip` with your roll number. Include the report as well in the zip. Do *not* include the data files. However, clearly mention which data files you have used according to your roll number.

Please include all the details of your running and scripts in the report.