CS315A Assignment

Dipesh Khandelwal (180249)

April 2, 2021

The nine Databases used in the assignment are as follows:

1. (A-100.csv, B-100-3-4.csv)
2. (A-100.csv, B-100-5-3.csv)
3. (A-100.csv, B-100-10-3.csv)
4. (A-1000.csv, B-1000-5-3.csv)
5. (A-1000.csv, B-1000-10-1.csv)
6. (A-1000.csv, B-1000-50-1.csv)
7. (A-10000.csv, B-10000-5-3.csv)
8. (A-10000.csv, B-10000-50-1.csv)
9. (A-10000.csv, B-10000-500-1.csv)

Exercise 1

The queries in SQL query language are as follows:

Query 1: SELECT * FROM A WHERE A1 <= 50;

Query 2: SELECT * FROM B ORDER BY B3;

Query 3: SELECT cast(COUNT(*) as real)/COUNT(DISTINCT B2) FROM B;

Query 4: SELECT * FROM A INNER JOIN B ON A.A1 = B.B2;

The queries in MongoDB query language(as used with pymongo(python3)) are as follows:

Query 1:
```
db["A"].find({"A1": {"$lte":50}})
```

Query 2:
```
db["B"].aggregate(pipeline=[{"$sort":{"B3":1}}], allowDiskUse = True )
```

Query 3:
```
db["B"].aggregate(pipeline=[
{"$group": { "_id": "$B2", "countb2": { "$sum": 1 } }},
{"$group": { "_id": None, "result": { "$avg": "$countb2" } }}
], allowDiskUse = True)
```

Query 4:
```
db["B"].aggregate(pipeline=[
{
"$lookup":{ "from": "A", "localField" : "B2", "foreignField" : "A1", "as" : "q4" }
},
{
"$replaceRoot": { "newRoot": { "$mergeObjects": [ { "$arrayElemAt": [ "$q4", 0 ] }, "$$ROOT" ] } }
},
{ "$project": { "B1":1,"B2":1,"B3":1,"A2":1 } }
], allowDiskUse = True)
```

Exercise 2

The tables below show the mean time and standard deviation of the time taken to run the queries in each of the four different database management system and each of the nine different databases mentioned on the first page of the report.

I did two different types of analysis. First, cache is enabled in mariaDB, which is the default option and second, cache is disabled in mariaDB. In first case, the queries won't be running independently as the database will make use of the cache and thus, there will be difference in time taken for running the query with cache and without cache.

The first 2 tables contains the data in which cache is disabled while the last 2 tables contains the data in which cache is enabled.

NOTE: All the data mentioned in the table is in milliseconds(ms).

	DBMS	1	2	3	4	5	6	7	8	9
query1	sqlite3	0.75	0.75	0.75	0.0	0.25	0.75	0.5	0.5	0.5
query1	mariadb(without index)	0.74	0.76	0.29	2.53	1.61	2.35	11.78	7.24	105.91
query1	mariadb(with index)	0.72	0.32	1.02	0.85	0.86	0.28	0.74	0.3	0.44
query1	mongodb	0.0	0.0	0.0	0.75	0.75	1.0	11.25	6.75	24.75
query2	sqlite3	0.5	1.0	1.75	7.5	16.25	44.0	44.75	286.25	2876.75
query2	mariadb(without index)	2.15	2.18	1.36	12.17	37.8	52.04	84.07	571.15	100226.45
query2	mariadb(with index)	1.63	1.15	3.92	32.66	23.03	77.43	141.16	881.31	626026.12
query2	mongodb	1.0	0.75	2.0	6.25	12.75	42.75	51.0	288.75	4409.5
query3	sqlite3	0.0	0.75	0.25	1.75	2.0	3.5	7.25	37.5	316.0
query3	mariadb(without index)	1.39	1.16	0.56	2.45	6.99	10.7	14.9	462.82	12654.98
query3	mariadb(with index)	0.98	0.46	0.54	3.11	3.1	8.64	13.49	82.41	17668.29
query3	mongodb	0.75	0.25	1.0	5.0	7.25	24.75	28.5	176.0	1849.25
query4	sqlite3	0.5	1.0	1.0	6.25	7.75	18.0	24.5	186.25	1836.25
query4	mariadb(without index)	16.92	8.92	9.18	309.75	518.5	2185.95	27481.62	205946.06	2101488.57
query4	mariadb(with index)	1.75	2.71	1.47	9.79	11.36	102.62	54.13	986.06	37985.11
query4	mongodb	18.5	13.5	14.0	45.75	44.0	47.0	298.75	300.25	329.0

Figure 1: Mean time(in ms) taken by each of the four queries in each of the nine databases when cache is disabled in mariaDB

	DBMS	1	2	3	4	5	6	7	8	9
query1	sqlite3	0.5	0.5	0.5	0.0	0.5	0.5	0.58	0.58	0.58
query1	mariadb(without index)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
query1	mariadb(with index)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
query1	mongodb	0.0	0.0	0.0	0.5	0.5	0.0	3.2	1.26	36.84
query2	sqlite3	0.58	0.0	0.5	2.52	2.99	4.08	7.93	14.48	84.24
query2	mariadb(without index)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
query2	mariadb(with index)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
query2	mongodb	0.0	0.5	0.82	0.96	4.86	6.9	8.41	21.33	523.0
query3	sqlite3	0.0	0.5	0.5	0.96	0.0	0.58	0.96	3.11	17.17
query3	mariadb(without index)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
query3	mariadb(with index)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
query3	mongodb	0.5	0.5	1.41	0.82	1.5	9.03	1.73	12.83	462.79
query4	sqlite3	0.58	0.0	0.0	2.75	0.96	1.83	2.65	14.66	214.2
query4	mariadb(without index)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
query4	mariadb(with index)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
query4	mongodb	9.85	7.05	3.74	2.22	2.16	3.56	3.77	5.06	28.62

Figure 2: Standard Deviation of time taken by each of the four queries in each of the nine databases when cache is disabled in mariaDB

	DBMS	1	2	3	4	5	6	7	8	9
query1	sqlite3	0.75	0.75	0.75	0.0	0.25	0.75	0.5	0.5	0.5
query1	mariadb(without index)	0.66	0.7	0.75	1.77	1.78	2.02	7.61	7.04	7.37
query1	mariadb(with index)	0.7	0.74	0.71	0.85	0.68	0.75	0.75	0.31	686.63
query1	mongodb	0.0	0.0	0.0	0.75	0.75	1.0	11.25	6.75	24.75
query2	sqlite3	0.5	1.0	1.75	7.5	16.25	44.0	44.75	286.25	2876.75
query2	mariadb(without index)	1.17	34.81	6.92	18.42	26.24	70.16	76.41	957.57	87036.79
query2	mariadb(with index)	1.32	1.7	3.2	15.97	31.18	89.7	134.78	1437.44	384389.22
query2	mongodb	1.0	0.75	2.0	6.25	12.75	42.75	51.0	288.75	4409.5
query3	sqlite3	0.0	0.75	0.25	1.75	2.0	3.5	7.25	37.5	316.0
query3	mariadb(without index)	0.73	0.6	0.94	3.92	5.04	13.66	16.97	151.33	15549.53
query3	mariadb(with index)	0.78	0.86	1.24	1.98	3.25	8.54	13.23	94.85	14498.75
query3	mongodb	0.75	0.25	1.0	5.0	7.25	24.75	28.5	176.0	1849.25
query4	sqlite3	0.5	1.0	1.0	6.25	7.75	18.0	24.5	186.25	1836.25
query4	mariadb(without index)	6.53	9.19	14.39	325.51	553.94	2407.94	29102.95	219451.14	2212212.61
query4	mariadb(with index)	1.51	2.08	4.07	250.84	172.69	39.67	170.95	4328.73	35023.61
query4	mongodb	18.5	13.5	14.0	45.75	44.0	47.0	298.75	300.25	329.0

Figure 3: Mean time(in ms) taken by each of the four queries in each of the nine databases when cache is enabled in mariaDB

	DBMS	1	2	3	4	5	6	7	8	9
query1	sqlite3	0.5	0.5	0.5	0.0	0.5	0.5	0.58	0.58	0.58
query1	mariadb(without index)	0.19	0.11	0.15	1.02	0.89	0.69	1.14	1.05	0.62
query1	mariadb(with index)	0.31	0.03	0.12	0.09	0.1	0.13	0.02	0.03	1372.66
query1	mongodb	0.0	0.0	0.0	0.5	0.5	0.0	3.2	1.26	36.84
query2	sqlite3	0.58	0.0	0.5	2.52	2.99	4.08	7.93	14.48	84.24
query2	mariadb(without index)	0.48	66.27	8.15	8.7	7.16	8.48	5.35	144.55	2029.05
query2	mariadb(with index)	0.41	0.64	0.42	3.62	8.48	7.98	6.38	341.97	130505.65
query2	mongodb	0.0	0.5	0.82	0.96	4.86	6.9	8.41	21.33	523.0
query3	sqlite3	0.0	0.5	0.5	0.96	0.0	0.58	0.96	3.11	17.17
query3	mariadb(without index)	0.28	0.3	0.56	1.72	0.95	1.13	0.51	43.13	6299.45
query3	mariadb(with index)	0.22	0.33	0.26	0.07	0.43	0.16	0.33	7.26	2026.96
query3	mongodb	0.5	0.5	1.41	0.82	1.5	9.03	1.73	12.83	462.79
query4	sqlite3	0.58	0.0	0.0	2.75	0.96	1.83	2.65	14.66	214.2
query4	mariadb(without index)	3.06	6.34	6.64	18.38	43.59	271.1	1718.87	7530.62	56110.06
query4	mariadb(with index)	0.25	0.73	0.39	453.66	279.35	9.42	122.51	884.13	4492.6
query4	mongodb	9.85	7.05	3.74	2.22	2.16	3.56	3.77	5.06	28.62

Figure 4: Standard Deviation of time taken by each of the four queries in each of the nine databases when cache is enabled in mariaDB

Exercise 3

The graphs contain the database number on x-axis and time taken(in ms) on the y-axis. The vertical lines in the graph show us the standard deviation.

The graphs when caching is disabled in mariaDB.

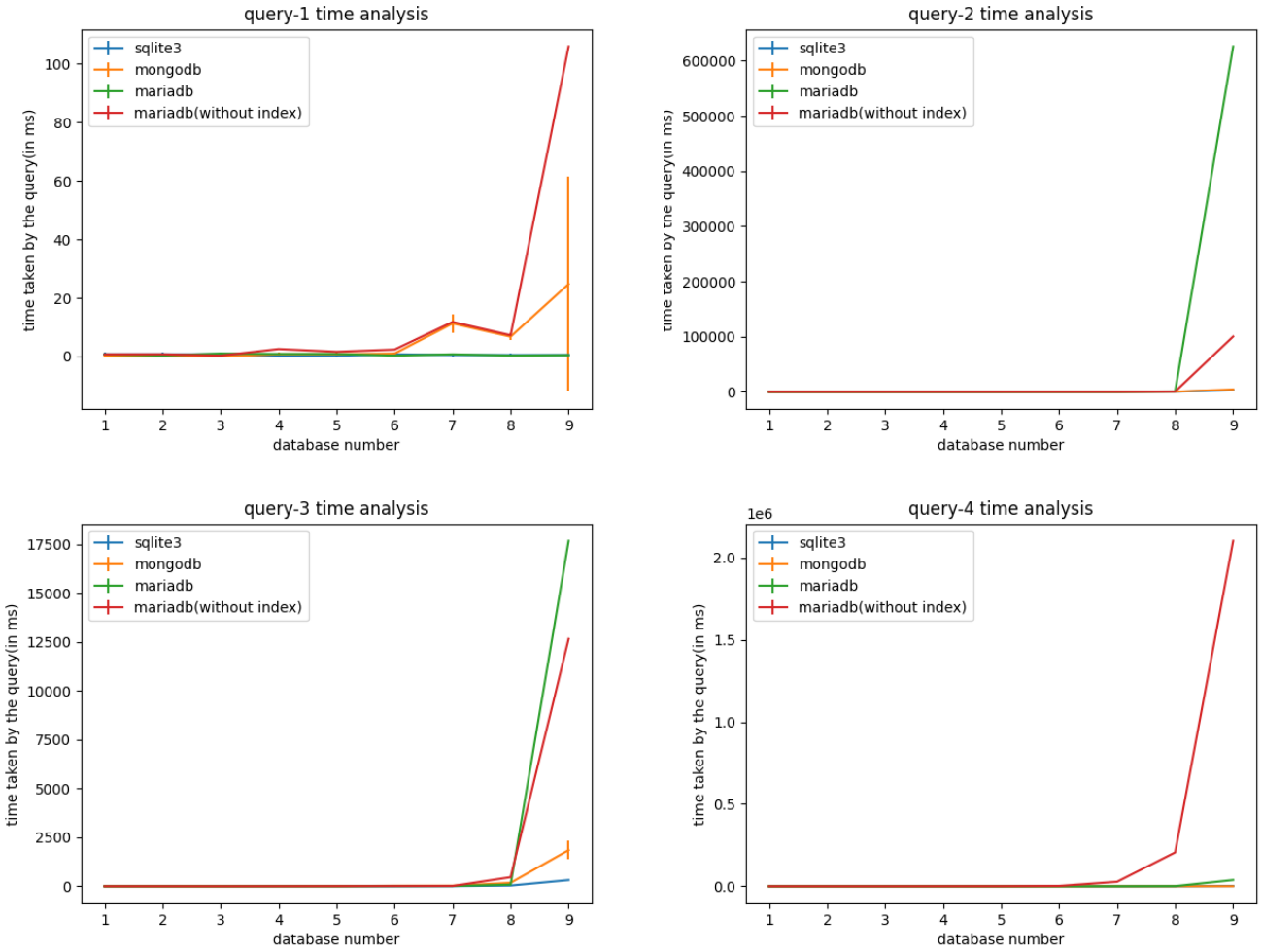

Figure 5: Graph for the time taken for executing all the 4 queries in all the 4 DBMS and all 9 databases when cache is disabled in mariaDB

As the 9th database is very large and hence, it takes a large amount of time in comparison to other databases, thus making it difficult to compare the time of the first 8 databases using graphs. Hence, I created graphs specifically for analysing the time taken by the first 8 databases and these are given below.

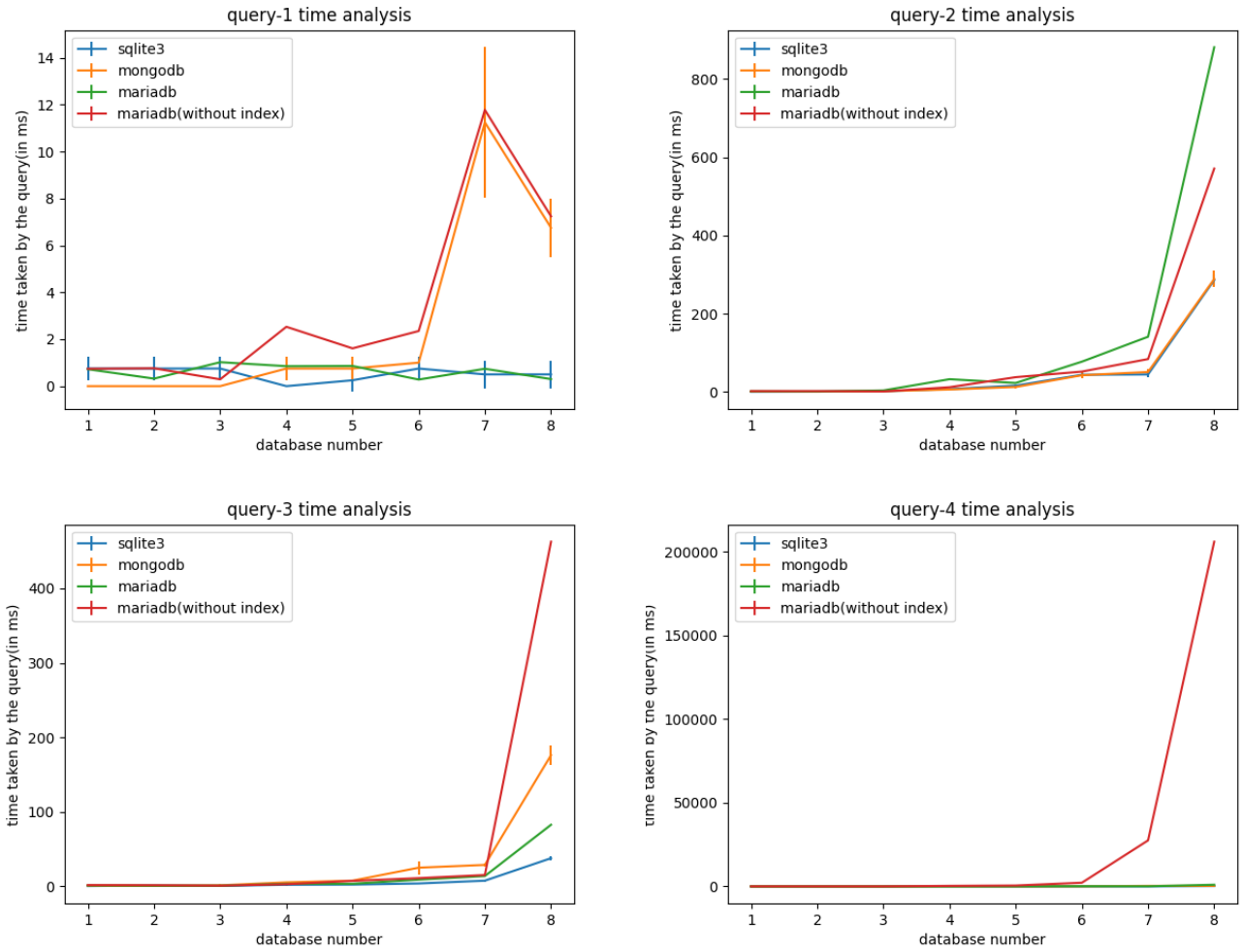

Figure 6: Graph for the time taken for executing all the 4 queries in all the 4 DBMS and first 8 databases when cache is disabled in mariaDB

The graphs when caching is enabled in mariaDB ,i.e, the default mode

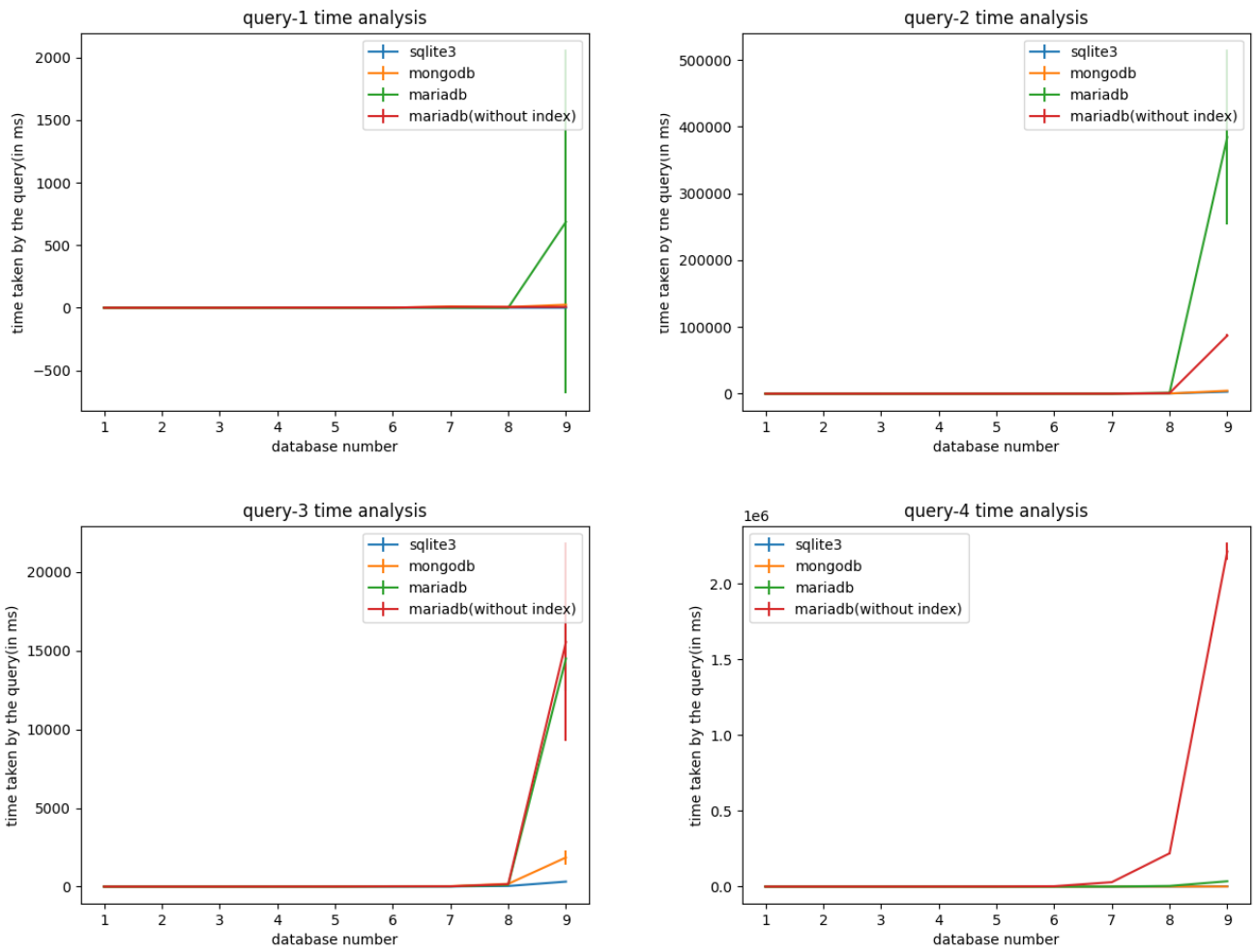

Figure 7: Graph for the time taken for executing all the 4 queries in all the 4 DBMS and all 9 databases when cache is enabled in mariaDB

As the 9th database is very large and hence, it takes a large amount of time in comparison to other databases, thus making it difficult to compare the time of the first 8 databases using graphs. Hence, I created graphs specifically for analysing the time taken by the first 8 databases and these are given below.

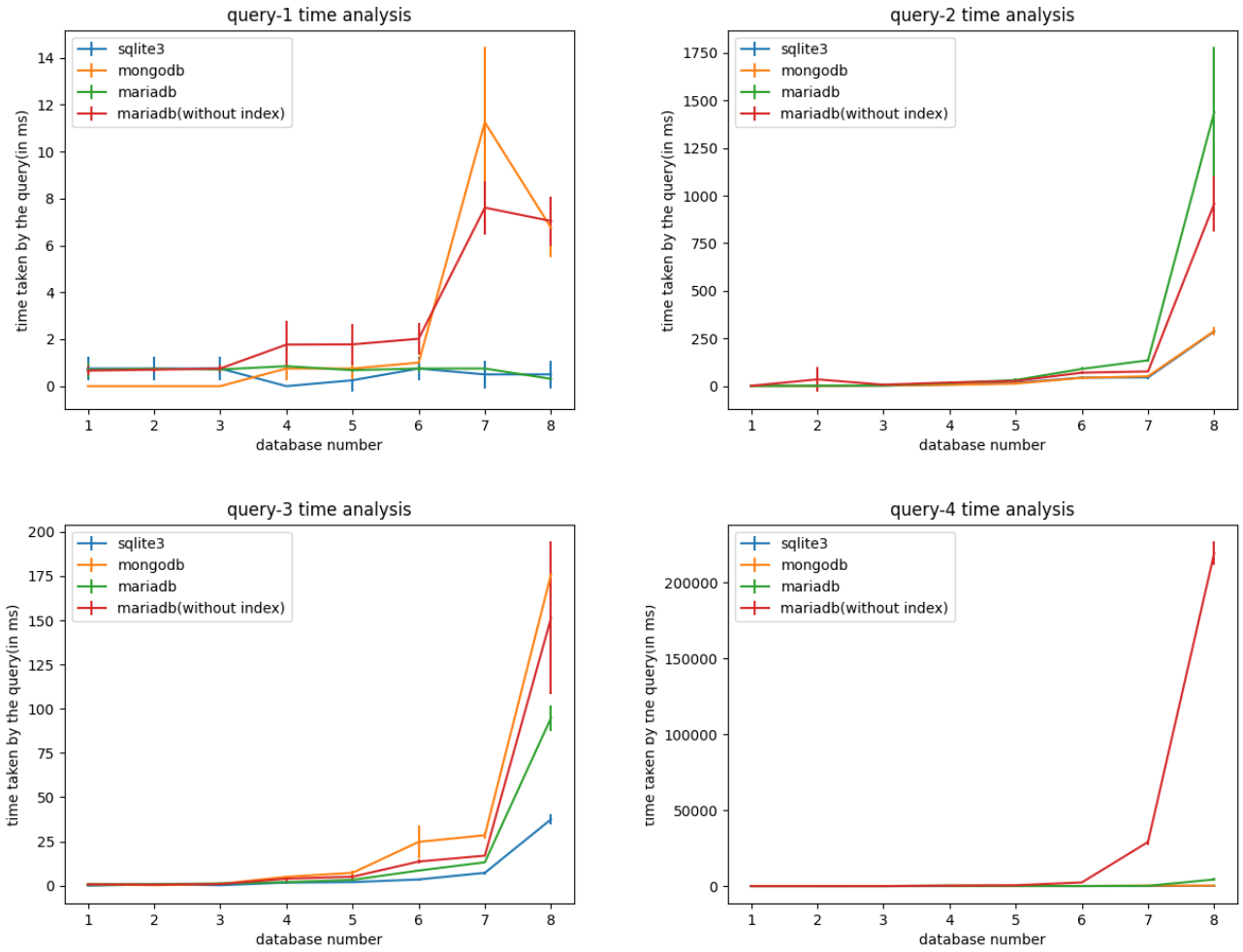

Figure 8: Graph for the time taken for executing all the 4 queries in all the 4 DBMS and first 8 databases when cache is enabled in mariaDB

Exercise 4

System Configurations:

- OS: Ubuntu 20.04.2 LTS, 64-bit
- Graphics: NV118 / Mesa Intel® UHD Graphics 620 (KBL GT2)
- Firmware : BIOS, by Dell Inc. version 1.18.0
- Memory : 8GiB SODIMM DDR4 Synchronous Unbuffered (Unregistered) 2400 MHz (0.4 ns)
- CPU: Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz
- HDD: 1 TiB
- SSD: 128 GiB

These are the system configurations extracted from the output of the command "$sudo lshw". For detailed system configurations refer to file systemConfig.txt

Observations and Conclusions:

1. The only DBMS which is portable among these four is sqlite3. You can create the database(which is created as a file) and then put it anywhere or send it anywhere to anyone and they can directly use the database. Rest all databases are server-based, i.e, not portable and have to be setup wherever you want to use them. This is a cubersome task. Also, if you want to send the database to anyone then, you need to dump these databases and all their outputs into files and then send them separately, which again needs to be imported into the database system on the other end for usage.
2. Both MariaDB and Sqlite3 are SQL based DBMS. But there is a significant difference in their speeds for the same queries which is clearly visible from the graphs and this difference in speeds is observed more clearly when the database contains large amount of data. In our case, the largest database is database 9, on which for every query there is a significant difference in time taken by mariaDB and slite3 as is visible from tables and graph above.
3. MariaDB is the slowest DBMS of all. This difference is more clearly seen in case of large databases where the difference in time is about 10-100x of the second slowest DBMS. Also, sqlite3 is the fastest DBMS of all the four DBMS.
4. MariaDB(with index) and MariaDB(without index) have significant difference in processing the queries. But, this difference is query dependent. In our case, query-1 and query-4 takes more time in processing on mariadb(without index) then mariadb(with index). On the other hand, query-2 and query-3 takes more time in processing on mariadb(with index) then mariadb(without index). These observations were made when caching was disabled so that each query is independently handled. The observations in the case when caching is enabled is quite different with query-1 and query-2 taking more time on mariadb(with index) and query-4 taking more time on mariadb(without index), while query-3 takes almost equal time on both. This leads to conclusion that queries are significantly affected by caching and indexing.
5. The most memory intensive query is query-2 as it exceeds memory limit of 100MiB in case on mongoDB. This is the reason that external sorting has to be allowed in case of mongoDB.
6. If we take the average time taken to process each query then by that data we can conclude that the most time taking query is query-4.
7. The DBMS in which implementation is most difficult is mongoDB. The fourth query which is a simple natural join query in SQL is very hard to implement in mongoDB.
8. The time taken by the query when printing the output and when the output being redirected to a file is different. The difference is significant as printing takes a lot of time as it includes I/O operations which are OS, CPU and other system configurations related.

9. The time taken by the database to process the query which is given by the DBMS is very less from the time actually taken to run the whole query. This leads to conclusion that to run a query there are a lots of overheads involved which the database ignores while taking the time.

Exercise 5

The details of the scripts and running of the scripts are as follows:

1. The csv files used contains headers.
2. Files:
   (a) main.py: This is the main script which needs be run to get the time taken for running the queries. This script creates 9*4 files(9 files for each of the four DBMS). Each file contains time taken for running all the 4 queries for a particular database in a particular DBMS. The naming system of the files will clearly tell which database and DBMS output this file contains. This scripts depends on 6 more scripts which it internally uses. Their description is given below.

   (b) chooseDatabases.py: this file contains functionality of giving you the databases when roll number is passed to it.

   (c) mariadb_utils.py: contains functions used to connect and run queries in mariadb(with index)

   (d) mariadb_without_index_utils.py: contains functions used to connect and run queries in mariadb(without index)

   (e) mongodb_utils.py: contains functions used to connect and run queries in mongoDB

   (f) sqlite3_utils.py: contains functions used to connect and run queries in sqlite3

   (g) queries.sql: contains sql queries for sqlite3

   (h) run_sqlite_queries.sh: runs queries from queries.sql into sqlite3 using terminal interface. This script is ran by sqlite3_utils.py file

   (i) createGraphs.py: this script read the output files generated by main.py and then extract the time from them and creates tables and graphs.

   (j) databasesUsed.txt: contains the names of the databases used

   (k) requirements.txt: contains all the python libraries used for running these programs

   (l) systemConfig.txt: contains output of "$sudo lshw" command. It contains system configurations in detail.

   (m) tables_cache _disabled.txt: contains table containing mean time and standard deviation of the time taken to run queries when cache is disabled in mariadb. These tables are included in the exercise 2 above.

   (n) tables.txt :contains table containing mean time and standard deviation of the time taken to run queries when cache is enabled in mariadb. These tables are included in the exercise 2 above.
3. I ran main.py 7 times at different points of time and collected the output and then ran createGraphs.py which automatically extracts time from the files created by main.py and creates tables and graphs.