

Database Systems
SOEN 363 - Fall 2022
Project - Phase 2

Out: Nov 10, 2022

Due: Dec 14, 2022

1 Project Objectives

The objectives of phase two of the project are to help students in: (a) appreciating the power of SQL in extracting and analyzing useful information from real datasets, (b) practicing and applying the data systems concepts, mainly modeling, storing, and querying datasets on large datasets, (c) appreciating the power of NoSQL in extracting and analyzing big datasets, and (d) providing a comparison between SQL and NoSQL systems.

2 Real Datasets from Open Data Portal or Social Apps

Governments, social applications, and public bodies produce huge quantities of real datasets. For example, Twitter provides a public API to download real tweets. Canada and USA make the governmental data more accessible to everyone, see [Canada Open Data](#) and [USA Open Data](#). Quebec data portal has more than 1200 datasets. Some real datasets are available at the level of cities, such as criminal acts recorded by the Police of Montreal or New York city. [The Linked Open Data Cloud](#) has more than 1400 datasets.

3 Analyzing Big Data Using SQL and RDBMS Systems











Each team has to choose between Oracle, Postgresql, MySQL, or Amazon Aurora. A team may use another RDBMS if the team gets approval from the professor.

- 5pts (a) Find a big real dataset; it is recommended to get a dataset of at least 5 different files of total size 0.5GB.
- 5pts (b) Create a relational database for a real dataset of your choice. Provide the ER model and the DDL statements.
- 5pts (c) Load the dataset into your database.
- 20pts (d) Write at least ten reports, i.e. write SQL queries, that show some useful information about the dataset. This should include aggregate functions, Group By, and Order By, queries.
- 5pts (e) Explore indexes to enhance the performance of these queries

4 NoSQL Databases

- NoSQL is not a single product or even a single technology. It represents a **class of products** and a collection of diverse, and sometimes related, concepts about data storage and manipulation.
- NoSQL database systems represent a new generation of low-cost, high performance database software which is increasingly gaining more and more popularity.
- The below figure illustrates some examples of NoSQL, and more examples are available at <https://dbdb.io/>.

- We highly recommend this list of NoSQL systems: Elasticsearch, Couchbase, NeDB, Apache Flink, Google Cloud Bigtable, Google BigQuery, Neo4J, Virtuoso, Apache Jena, Blazegraph, BoltDB, QuestDB, and HBase.
- It is NOT allowed to use MongoDB

Document Databases	Graph Databases
 Couchbase  MarkLogic  mongoDB	 neo4j  InfiniteGraph <small>The Distributed Graph Database</small>
Column Databases	Key-Value Databases
 redis  APACHE HBASE  riak	 HYPERTABLE <small>INC</small>  cassandra Amazon SimpleDB

5 Analyzing Big Data Using NoSQL Systems

NoSQL systems are developed to support big data applications. Each team has to use only one of the NoSQL systems mentioned above.

- | | |
|-------|--|
| 5pts | (a) Find a big real dataset; it is recommended to get a dataset of at least 1 GBs. This dataset could be the same one use with Section 1. |
| 5pts | (b) Provide the data model for your datasets, i.e., graph, document, key-value, or column-store. |
| 5pts | (c) Create a NoSQL database for a real dataset of your choice. |
| 5pts | (d) Load the dataset into your NoSQL system. |
| 20pts | (e) Write at least ten different queries that show some useful information about the dataset. This should include different aspects of your NoSQL. |
| 10pts | (f) Investigate the balance between the consistency and availability in your NoSQL system. |
| 10pts | (g) Investigate the indexing techniques available in your NoSQL system. |

6 Q&A

We use Moodle Forum as a platform for asking questions and receiving answers. Posting your questions on Moodle Forum will help the whole class benefit and will certainly avoid redundancy.

7 The Deliverable

The deliverables of this phase are 1) a 10 minutes presentation and a live demo for 10 minutes, where the team will be asked to write new queries or adjust existing ones. The presentation should include for the SQL part slides mainly about:

- What is the dataset? how big is it? the original format? how many files?
- The ER diagrams of the dataset
- Loading the dataset into the database
- **Live demonstration of the queries using your RDB system.**

and for the NoSQL part slides mainly about:

- What is the dataset? how big is it? the original format?
- A data model for the dataset
- Discuss the consistency and availability of the NoSQL used in the project
- Discuss the indexing techniques available in the NoSQL used in the project
- **Live demonstration of the queries using your NoSQL system.**

8 Submission

The full presentation should be submitted on April 12 at 11.59 pm.