

Database Systems
SOEN 363 - Winter 2020
Project - Phase 2

Out: March 16, 2020

Due: April 06 at 3 pm, 2020

1 Project Objectives











The objectives of phase two of the project are to help students in: **(a)** practicing and applying the data systems concepts, mainly modeling, storing, and querying datasets on large datasets **(b)** using NoSQL to querying a real dataset, and **(c)** appreciating the power of NoSQL in extracting and analyzing big datasets.

2 NoSQL Databases¹

- 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.

¹<http://nosql-database.org/>

- 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. There are also other examples, such as Elasticsearch, Google BigTable, Amazon Dynamo.
- Each team has to choose only one of these NoSQL systems for phase two.

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

3 Analyzing Big Data Using NoSQL Systems

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 [data.gov](#). Some real datasets are available at the level of cities, such as criminal acts recorded by the Police of Montreal or New York city. NoSQL Systems are developed to support big data applications. Each team has to use only one of the NoSQL systems mentioned in Section 2 to:

- | | |
|-------|--|
| 15pts | (a) Download a big real dataset; it is recommended to get a dataset of at least 0.5 GBs. |
| 10pts | (b) Provide the data model for your datasets, i.e., graph, document, key-value, or column-store. |
| 15pts | (c) Create a NoSQL database for a real dataset of your choice. |
| 10pts | (d) Load the dataset into your NoSQL system. |
| 50pts | (e) Write at least 10 different queries, that show some useful information about the dataset. This should include different aspects of your NoSQL. |
| 50pts | (f) Investigate the balance between the consistency and availability in your NoSQL system. |
| 50pts | (g) Investigate the indexing techniques available in your NoSQL system. |

4 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.

5 The Deliverable

- All the scripts and code developed in Phase 2 will be submitted as a PDF file named `Project Phase2-<your_andrew_id>.pdf`. For example: `Project Phase2-Kevin Smith-99887766.pdf`
- Each team will also give a 5 minutes presentation. The presentations will take place on April 6 during the lecture. The presentation will include 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
 - **Demonstrate live the queries, i.e., using your NoSQL system**

6 Submission

Zip all your files into a single archive file and submit it to Moodle. In case of any problems, you can email your project archive to the Professors and the TA.

7 Late Policy

- If you hand in on time, there is no penalty.
- 0-24 hours late = 25% penalty.
- 24-48 hours late = 50% penalty.
- More than 48 hours late = you lose all the points for this project.