Instructions and Evaluation

NoSQL

A4- S8

ESILV
jihane.mali (at) devinci.fr

Chapter

Contents

Chapter 1

Instructions

1.1 Introduction

All the course content is available on DeVinci Learning: PDF, videos, external links... Please read carefully the content in order to be aware of the notions required to work properly during practice works.

Practice works will focus on NoSQL databases interactions, especially the DSL (Domain Specific Language) of each solution.

1.2 Softwares Installation

To earn a LOT of time during practice works, we will spend time during the first session to install ALL softwares required for it. So, all databases must be installed BEFORE the corresponding practice work (Cassandra, MongoDB, Elasticsearch, Neo4j). The professor won't spend time during this session for installation.

Please, refer to the installation guide online.

1.3 Evaluation and Grades

This course will be evaluated in two ways:

$\frac{8}{100}$ For each course notion will be evaluated: JSON, Cassandra, MongoDB, Elasticsearch and Neo4j. Those notions MUST be known. A single quizz will be applied during the last course.

$\frac{15}{100}$ For each report on the usage of different NoSQL databases with your chosen dataset.

$$5 \times \frac{8}{100} + 4 \times \frac{15}{100} = 100$$

Negative points will be recorded for each delay on:

- Installing softwares for the practice work (Cassandra, MongoDB/Compass, elasticsearch/curl/kibana, Neo4jDesktop). It must be install and tested BEFORE the practice work on it,
- Report submission delays,
- Lack of participation that is noted with your team.

1.4 Quizz

A Quizz will validate all the notions acquired during the previous sessions. The purpose is to evaluate the knowledge but also some skills developed during practice works. So, pay attention to course content, ask questions during the course and TD.

For example, ask yourself some questions:

- What NoSQL stands for?
- What is the CAP theorem?
- Does Cassandra is more Consistent or Available?
- MongoDB is a document store or a wide-columned-oriented store?
- What is MapReduce?

- How many mistakes this JSON document contains?
- What is the proper answer for this Neo4j query?
- …

Chapter 2

Datasets and Reports

2.1 Choose your dataset

Different datasets are available online linked to each group. First constitute groups of 4 students at most and choose your dataset online: first coming, first served.

This dataset will be used for each of the database: Cassandra, MongoDB, elasticsearch and Neo4j. You don't have to change each time. The purpose of this work is to apply what you have learned[1] during the practice work on **YOUR** own data (the dataset).

You have to import the dataset, produce queries and explain if some operations must have been made to allow importation.

Each report has to be produced once for a chosen dataset and NoSQL solution.

Evaluation Grid. A grid is available for each report. It will be used to grade you production according to some criteria:

- Description of data import, database creation/schema/index, data cleaning, data transformation, etc.
- Queries of various complexities (see below): number and quality
- New features that you propose on your queries that are not presented during the course
- Quality of your report (easy to read, English writing, arguments)
- Delay of deposit (in advance, on time, late)

Obtaining a good mark is challenging!

Datasets complexity. Every datasets are not of same complexities. We adapt the requirements of your report according to this complexity by varying the number of queries on different complexities[2]: *simple*, *complex and hard*. Your mark will rely on this complexity, so do the queries to produce for each report (database):

2.1.1 Difficulty 1: Quite simple to use.

- 5 simple queries, normal filters, projection and countings
- 2 complex queries, strong filters, aggregates, geolocalisation
- 2 hard queries, complex aggregates, combined with filters (post filters?), transformation, MapReduce, etc.

2.1.2 Difficulty 2: Not really difficult to use.

- 6 simple queries, normal filters, projection and countings
- 2 complex queries, strong filters, aggregates, geolocalisation
- 1 hard query, complex aggregates, combined with filters (post filters?), transformation, MapReduce, etc.

2.1.3 Difficulty 3: Quite difficult to use.

- 8 simple queries, normal filters, projection and countings
- 1 complex queries, strong filters, aggregates, geolocalisation
- 1 hard query, complex aggregates, combined with filters (post filters?), transformation, MapReduce, etc.

NB If a query is consider to be too simple for the required complexity, it will be avoided from the mark.

[1] The queries studied during practice works will be different from the one you will provide on your dataset (not the same schema and purpose). You have to produce YOUR queries.

[2] Simple, complex and hard queries are detailed on each practice work to help you to guide the classification

2.2 Reports on Datasets

When you have constituted a team on a dataset (at most four students per team), you will have to test it on each database. You need to import the dataset (after requirements of the database) and then query it.

To query the database, provide the query in **natural language**, and then translate it into the **DSL of the database**. The number of query to provide is dependant on your dataset complexity.

Watch the *"rubrique"* of your report in order to know on which criteria you will be evaluated (import, queries, writing, deadline, misc.).

For each report, you must give how members of the group have contributed to the report. Different marks can be given according to an unbalanced repartition.

Four reports have to be submitted online, one for each NoSQL database, the deadline is given online.