Université Paris-Saclay
Master 2 Data and Knowledge

Professor: Ioana Manolescu
Teaching assistant: Paweł Guzewicz[1]

Big Data Architectures Lab 2

September 30, 2019

Submission deadline: Sunday, October 13, 2019 at 23:59:59.

1 Lab organization[2]

1.1 MongoDB setup

MongoDB is supported under several Linux distributions, macOS and Windows platforms.

1. Download and install MongoDB following instructions at `https://docs.mongodb.com/manual/administration/install-community`.
2. Launch MongoDB and check if it works.

1.2 MongoDB tutorials

Please start by reading an introduction to data management in MongoDB here: `https://docs.mongodb.com/manual`. You can find many tutorials at official MongoDB website: `https://docs.mongodb.com/manual/tutorial`. Also at `https://www.tutorialspoint.com/mongodb` or `https://www.tutorialkart.com/mongodb/mongodb-tutorial`.

1.3 Assignment and questions

Please read the assignment description in the next section. If you have any questions about the lab or on the course material, don't hesitate to ask them during the lab session or via email (don't wait until the last moment).

2 Assignment

Download this archive file: `https://www.lix.polytechnique.fr/Labo/Pawel.Guzewicz/teaching/2019_2020_Big_Data_Architectures/lab_2_datasets.zip`. It contains four JSONL files (JSONL stands for JSON Lines, a file format in which each line is written in JSON data format). Using these datasets, you need to solve the following tasks.

Task 0: setup

1. Create a directory for the MongoDB server. Launch the server (mongod) and write down the commands you used. On which port does it run?
2. Import the document `moviepeople-10.jsonl` into the server.
3. Launch a client (mongo shell), retrieve all the documents by asking a query in the client.

Task 1: import and querying

1. Import the documents `moviepeople-3000.jsonl` and `cities.jsonl` into the server (the one set up in Task 0).
2. In the mongo shell client, write queries for finding:
   (a) the person named Anabela Teixeira

[1] pawel.guzewicz@lix.polytechnique.fr; `https://www.lix.polytechnique.fr/Labo/Pawel.Guzewicz/teaching`
[2] All lab materials based on previous course editions courtesy of Ioana Manolescu and Silviu Maniu (with some changes)

(b) the birthplace of Steven Spielberg

(c) the number of people born in Lisbon

(d) the people taller than 170 cm

(e) the names of people whose information contains "Opera"

(f) for each movie person whose birth place is known, the latitude, longitude and population of that city (if that information exists in the city document).

You may use functions, several commands, aggregates, lookups etc.

Task 2: replication

1. Create working directories for 3 MongoDB servers.
2. Create a replica set for a collection called `small-movie`.
3. Launch the three MongoDB servers (in different shells) and let them run.
4. Connect a client (mongo) to one server. Through the client, initialize the replication: add one other server as secondary, and add the third one as arbiter.
5. Identify the master from the outputs of the servers' and by requesting replica set information from the servers.
6. Import `moviepeople-1000.jsonl` through the master. Observe the output of the two other servers.
7. Once the synchronization is finished, stop (ctrl-c) the master. Observe the output of the two other servers.

Task 3: sharding

1. Start two shard servers.
2. Shard the cities by the country.

2.1 Report

Write a report on your solutions for the tasks. It should include the following elements.

1. Setup information: for each task write down **all the commands** you used to launch the MongoDB server(s)/client(s).
2. Answers: for each solution write down **the complete list of commands or queries** you used, as well as **all the results/output**. Comment on them.

You can separate the output of commands and queries, and/or screenshots of terminal(s) into some files and then refer to those files in the report. (Please include any external files in the submission archive.)

2.2 Submission guidelines

Please follow submission rules and guidelines: `https://www.lix.polytechnique.fr/Labo/Pawel.Guzewicz/teaching/2019_2020_Big_Data_Architectures/submission_rules_and_guidelines.pdf`.

Moreover, I encourage you to read my advice on lab sessions and submissions: `https://www.lix.polytechnique.fr/Labo/Pawel.Guzewicz/teaching/2019_2020_Big_Data_Architectures/advice_on_lab_sessions_and_submissions.pdf`