Institut Polytechnique de Paris
DataAI

Professor: Ioana Manolescu
Teaching assistant: Madhulika Mohanty[1]

Architectures for Big Data: Lab 2 - MongoDB

October 18, 2022

Submission deadline: Monday, October 31, 2022 at 23:55:00 (Paris time).

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://www.mongodb.com/docs/manual`. In particular, to get familiar with MongoDB client syntax, please read: `https://www.mongodb.com/docs/manual/tutorial/getting-started/`. You can find many tutorials at the official MongoDB website: `https://www.mongodb.com/docs/manual/tutorial/` and 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, do not hesitate to ask them during the lab session, via Moodle or over email.

2 Assignment

Please download the `Lab 2 datasets` zip file from Moodle. The archive contains four JSONL files[3]. Using these datasets, you need to solve the following tasks.

Task 0: setup

1. Create a data store 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 file `moviepeople-10.jsonl` into the server.
3. Launch a mongo shell client, and ask a query to retrieve all the data.

Task 1: import and querying

1. Import the files `moviepeople-3000.jsonl` and `cities.jsonl` into the server (the one set up in Task 0).
2. In the mongo client shell, write queries to find:
   (a) information about the person called Bárbara Norton de Matos

[1] madhulika.mohanty@inria.fr

[2] All lab materials based on previous editions of the course (with changes) courtesy of Ioana Manolescu, Silviu Maniu and Pawel Guzewicz.

[3] JSONL stands for JSON Lines, a file format, where each line is a valid JSON value; see also: `https://jsonlines.org`.

(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) the last spouse of each person who have ever had one

(g) for each movie person whose birth place is known, the latitude, longitude and population of the corresponding city (if such information exist for the city).

You may use various features of the MongoDB query language such as aggregates, lookups etc.

Task 2: replication

1. Create data store directories for three MongoDB servers.
2. Create a replica set for a collection called `small-movie`. *Hint*: to do that you will need to launch the three MongoDB servers (in different shells) and let them run.
3. Connect a mongo client to one of the servers. Through the client, initialize the replication: add one other server as secondary, and add the third one as arbiter.
4. Identify the master from the outputs of the servers, as well as by requesting replica set information from the servers.
5. Import `moviepeople-1000.jsonl` through the master. Observe the output of the two other servers.
6. Once the synchronization is finished, stop the master. Observe and report the output of the two other servers.

Task 3: sharding

1. Create data store directories and start two shard servers.
2. Shard the cities from `cities.jsonl` 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** (you may provide an excerpt if too big). Also, elaborate on them discussing what do the outputs indicate.

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 the submission rules and guidelines available at Moodle for Lab-1.