Lab 03: Apache Spark with MongoDB

Instructor: Doan Dinh Toan
Department of Computer Science
Faculty of Information Technology
University of Science - VNU-HCM
toandd.i81@gmail.com

Abstract

In this lab assignment, you will be using Apache Spark to perform simple queries on a MongoDB database. By completing this lab assignment, you will gain experience using PySpark to work with real-world data and perform various data manipulation tasks.

0 Preliminary

0.1 Reminder

The main objective of this course is to truly learn. You can discuss this with your classmate, but you need to take responsibility for your submission, which actually depends on your understanding.

For any kind of cheating and plagiarism, students will be graded 0 marks for the course.

0.2 Submission guideline

Each team submits its result to a folder named `teamABC`, with `ABC` being the team's name. The folder structure is as follows:

```
.
└──── teamABC/
    ├──── src/
    │    └──── problem01
    │    └──── problem02
    ├──── docs/
    │    ├──── report.pdf
    │    └──── report.[md|tex|typ]
    │    └──── images
```

- `src` is the folder for your source code. If the lab assignment is split into multiple sections, you have to save your script in a separate folder, corresponding to the given lab assignment.
- `docs` is the folder for your documents, including the work report and images associated with your report. If the lab assignment requires screenshots as proof, the images need to be stored in this folder if you inserted them in the report.
  - `report.[md|tex|typ]` is your raw report file. The report must be written in English. The report must include the following items:
    - Information about the course, the assignment, and notes to the instructors (if any).
    - Information about your team (Student ID, full name of each member).
    - Your team's result (How much work, in percent (%), have you finished in each section?)

  - The answer to each section's tasks.
  - Reflection of your team. (Does your journey to the deadline have any bugs? How have you overcome it? What have you learned after this process? If you cannot overcome the bugs, describe where the bottlenecks are in your work.)
  - References to your work.
- `report.pdf` is the PDF file of your report. You need to check this file carefully before submit[1].

0.3 Rubrics

Students can earn 0.6 points for each problem in section 1 and 1.6 points for each problem in section 2.

0.4 Notes

- If you complete the project using Google Colab, please provide the Google Colab link(s). The data should be organized such that the grader can run your code with minor adjustments. No modification after the deadline is allowed.

- Alternatively, please submit the whole source code (in Python only) and write a careful guide of how to run the code.

- This assignment has an embedded notebook in Google Colab. In this notebook, the environment has a fixed configuration, but it is not quite usual in the industry environment. To set up your environment (in case someday you participate in a job position as Data Engineer), please read the listed docs in the references[1, 2]

- For each question, you should give notes to every important lines of code.

- Due to the fact that Colab has upgrade their environment to Ubuntu 22.04, some environment is not work at all[2]. You should using your own MongoDB server to do this lab on local if you can not resolve the environment conflict.

[1] The report must be written in one of the following formats: Markdown (.md), LaTeX (.tex), or Typst style (.typ). The report is embedded with a specific template in Typst, therefore, regardless of the tool used, it is **essential to ensure that the formatting and design layout remain consistent with the provided template**.

[2] `https://medium.com/google-colab/colab-updated-to-ubuntu-22-04-lts-709a91555b3c`

1 Introduction to PySpark

In this lab assignment, we will work with a movie dataset loaded into our MongoDB at `input_data.movies_lang`. We will use PySpark to perform the following tasks:

a Count the number of movies by country. Sort by count in decreasing order.
b Return the titles of the movies produced in France.
c Return the title of the movies of which Sofia Coppola is one of the actresses.
d Return the names and birth dates of the directors of movies produced in France.
e Return the average number of actors in a film.
f Return the name of the actor that acted in the most movies.

2 Real-world Data Manipulation

In this part of the lab, we will work with two collections in our MongoDB: `gia_ke_khai_raw` and `thuoc_raw` [3] loaded at `input_data.gia_ke_khai_raw` and `input_data.thuoc_raw` respectively. We will use PySpark to perform the following tasks:

a Read the datasets into a DataFrame and print out the schema and the number of records.

b Show all records in the `thuoc_raw` collection that have the same active pharmaceutical ingredient (API) in their `hoatChat` field as their medicine name. Note that the API not included the dosing amount like teaspoon (tps) or milliliters (ml),…

c Create a new DataFrame from the `thuoc_raw` collection that splits the API in the `hoatChat` field into multiple rows. For example, "paracetamol" is the API in "Paracetamol 500 mg," and "amoxicillin" is the API in various medications such as "Amogentine 500mg/125mg," "Augbactam 1g/200mg," and "Viamomentin." The resulting DataFrame should have two columns: `hoatChat` and `thuocTuongUng` as a list. After processing the data, write it back to our MongoDB at `output_data.thuocthaythe`.

d Create a new DataFrame from the two collections mentioned above that contains the fields `tenThuoc`, `hoatChat`, `dongGoi`, `dvt`, and `giaBan`. After processing the data, write it back to our MongoDB at `output_data.giathuoc`.

Bibliography

[1] "Maven – Maven Documentation." Accessed: Apr. 6, 2023. [Online]. Available: https://maven.apache.org/guides/

[2] "MongoDB Connector for Spark — MongoDB Spark Connector." Accessed: Apr. 6, 2023. [Online]. Available: https://www.mongodb.com/docs/spark-connector/current/

[3] C. Q. lý Dược - Bộ Y tế, "DrugBank \textbar Ngân hàng dữ liệu ngành Dược." (\url{https://drugbank.vn/})