Université Paris-Saclay
Master 2 Data and Knowledge

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

Big Data Architectures Lab 4

October 21, 2019

Submission deadline: Sunday, November 3, 2019 at 23:59:59.

1 Lab organization

1.1 Polystores recall

Please find the presentation at `https://www.lix.polytechnique.fr/Labo/Pawel.Guzewicz/teaching/2019_2020_Big_Data_Architectures/lab_4_slides.pdf`.

1.2 Apache Drill setup

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

1. Download and install Apache Drill (Embedded Mode) following instructions at `https://drill.apache.org/docs/installing-drill-on-linux-and-mac-os-x/` or `https://drill.apache.org/docs/installing-drill-on-windows/`, depending on your operating system.
2. Launch Apache Drill and check if it works by following instructions at `https://drill.apache.org/docs/starting-drill-on-linux-and-mac-os-x/` or `https://drill.apache.org/docs/starting-drill-on-windows/`, depending on your operating system.

1.3 Apache Drill tutorials

Go through a basic tutorial at `https://drill.apache.org/docs/drill-in-10-minutes/`. Having launched Apache Drill, check the Web user interface at `https://drill.apache.org/docs/starting-the-web-ui/`.

1.4 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_3_dataset.zip` (the same one as for Lab 3). Also, download "Structures egalité femmes-hommes" dataset in JSON format from `https://www.data.gouv.fr/fr/datasets/structures-egalite-femmes-hommes/`. Using these datasets, you need to solve the following tasks.

Task 1: querying MongoDB and saving results in Apache Parquet file format

1. Set up MongoDB plugin following instructions at `https://drill.apache.org/docs/mongodb-storage-plugin/`. Check if you are able to query the example zip code Mongo database using example queries from the tutorial.
2. Import `structures-egalite-femmeshommes.json` dataset into MongoDB.

[1] pawel.guzewicz@lix.polytechnique.fr; `https://www.lix.polytechnique.fr/Labo/Pawel.Guzewicz/teaching`

3. In Apache Drill, using MongoDB plugin, find in the imported dataset the number of organizations working for gender equality in Toulouse by their zip code in the descending order of size.

4. Analyze the result of the aggregation query, Is the organizations' zip codes data complete?

5. Save the result of the query into a Parquet file in `tmp` workspace using a default `dfs` plugin.

6. Run a query to display the content of the Parquet file.

Task 2: importing data in CSV and joining with data in Postgres

1. Import `boston-crime-incident-reports-10k.csv` dataset into Postgres.

2. Set up Postgres plugin following instructions at `https://drill.apache.org/docs/rdbms-storage-plugin/`. In particular, you need to create a plugin in Apache Drill Web UI, and specify database connection parameters.

3. Run a query to display the content of the dataset loaded to Postgres in Apache Drill.

4. Run a query to display the content of `boston-offense-codes-lookup.csv` file in Apache Drill (without loading it to Postgres).

5. Find all the distinct street names mentioned in reports such that their code name in a lookup CSV file contains "FIRE" and they refer to Monday.

2.1 Report

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

1. Answers: for each solution write down **the complete list of commands or queries** you used, as well as **all the results/output**.

2. Explanation of the commands or queries, i.e., what is a logical operation that you perform.

The report should not include the dataset nor any output that comprises the whole dataset.
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`