CSCI 5408

Data Management, Warehousing, And Analytics

Lab 5: Big Data: Hadoop and Apache Spark

Prepared By

Bhavisha Oza (B00935827)

Summary

Apache Spark is an open-source cluster computing framework that provides fast and scalable data processing and analytics. Hadoop offers distributed storage and processing, while Spark enhances it with in-memory computation and real-time analytics. Advantages of Hadoop include fault tolerance and scalability, while Spark provides faster processing and better support for iterative algorithms. Disadvantages of Hadoop are high latency and complexity, whereas Spark has a steeper learning curve and higher memory requirements.

In the given lab did some hands-on for the Apache Spark program to fetch the data from given weather.json file

Steps followed:

- Completed setup of Apache Spark on GCP as taught in the lab [2].
- Create a free account on the OpenWeather API.
- Retrieved the 5-days weather forecast data for "Halifax" using the API
- There was some issue generating json file so used the weather.json file given in the teams channel.
- Created a Java program that filters the response data where the daily "feels_like" temperature for the next 5-days is greater than 15°C during the "day" time. Exclude the current, minutely, and hourly fields.
- Save the filtered data into a new file – "summer_wealther.json"

Lab exercise:

1. Completed setup of Apache Spark on GCP

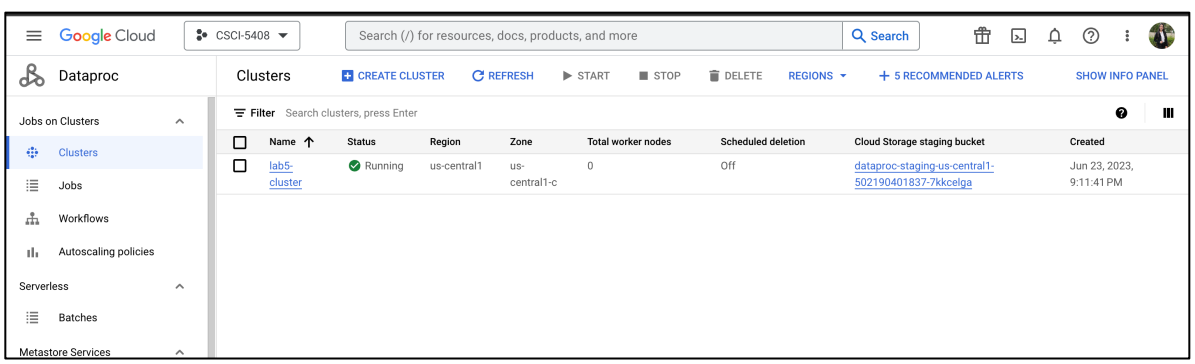

Figure 1: Apache Spark Setup on GCP

2. Created the account on OpenWeatherAPI

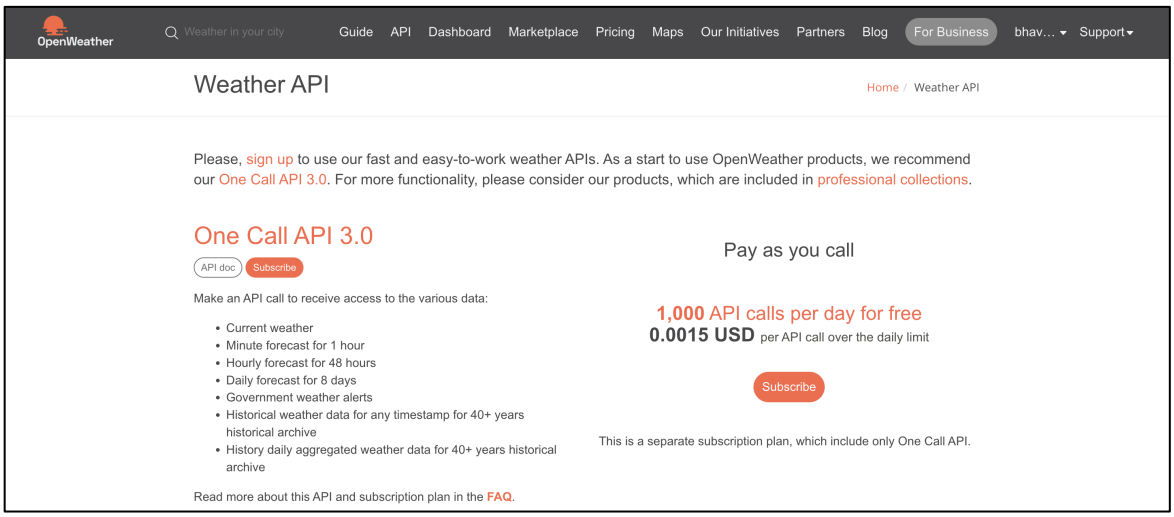

Figure 2: Account creation on OpenWeather website

3. Tried to get the API response for the 5 days data from postman.

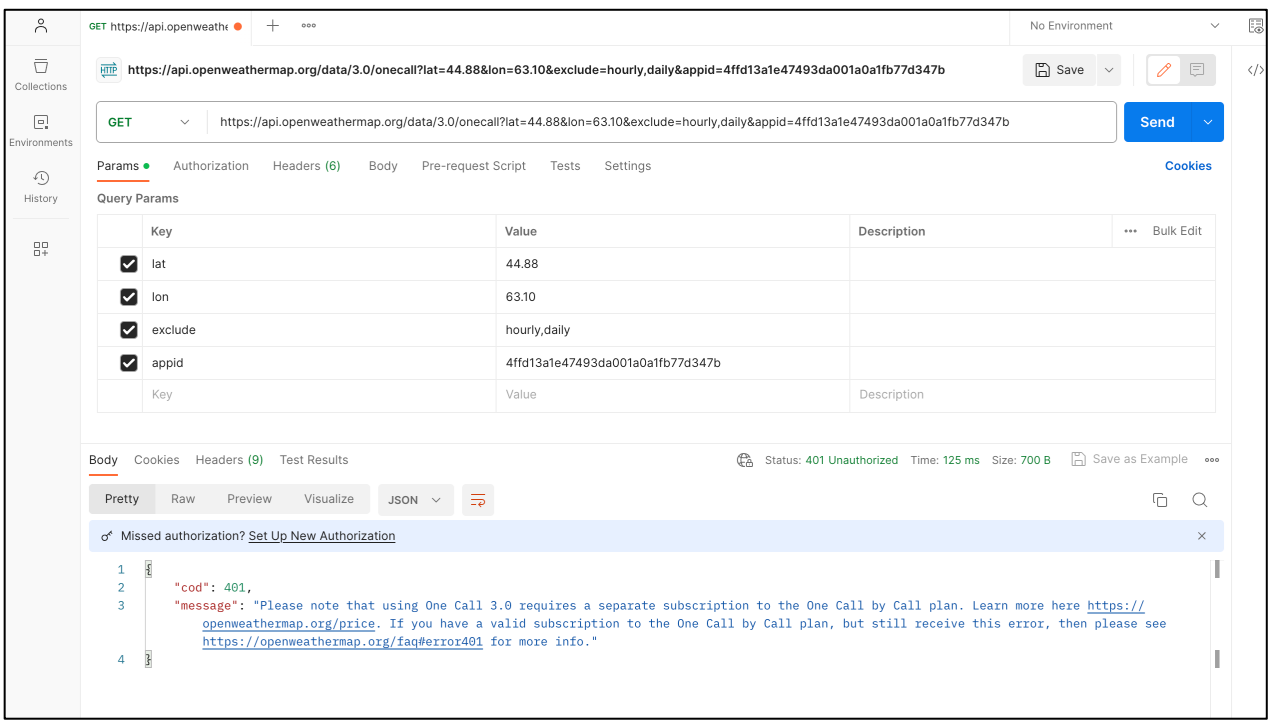

Figure 3: Postman API call

4. Write a Spark program to filter the response data where the daily "feels_like" temperature for the next 5-days is greater than 15°C during the "day" time. Exclude the current, minutely, and hourly fields.

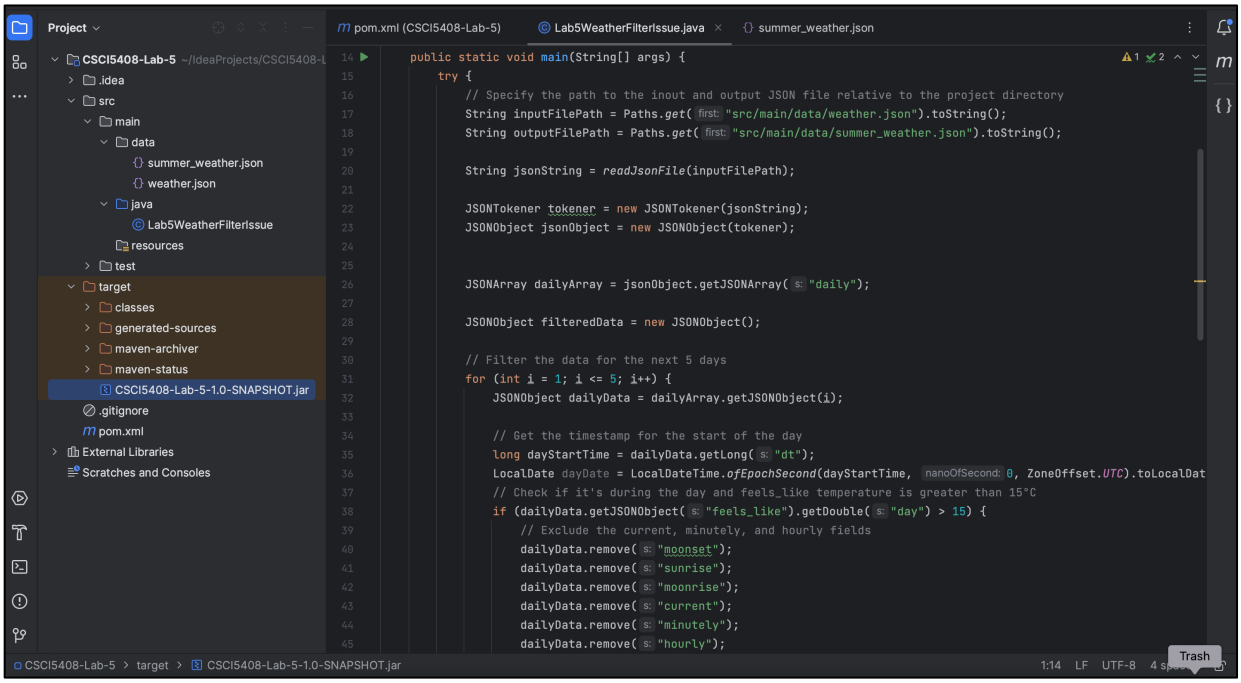

Figure 4: Java program for filtering the data.

5. Save the filtered data into a new file – "summer_wealther.json"

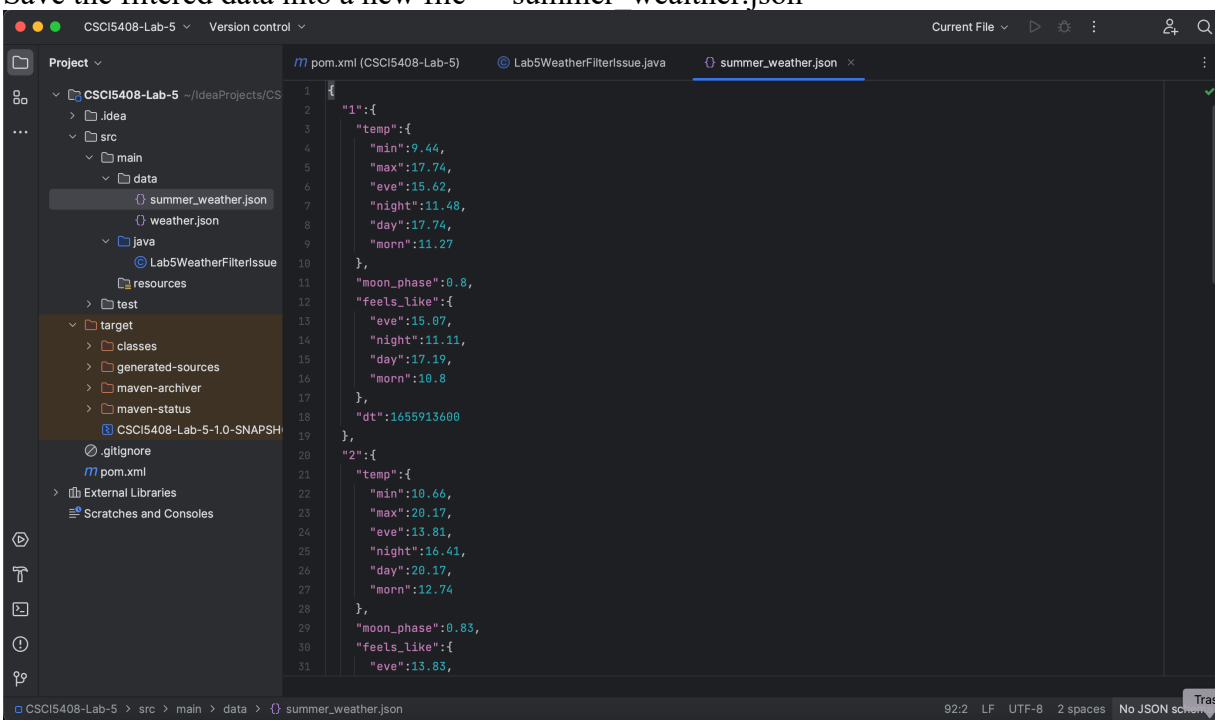

Figure 5: Filtered data stored in summer_weather.json file

6. Checking on GCP for the same

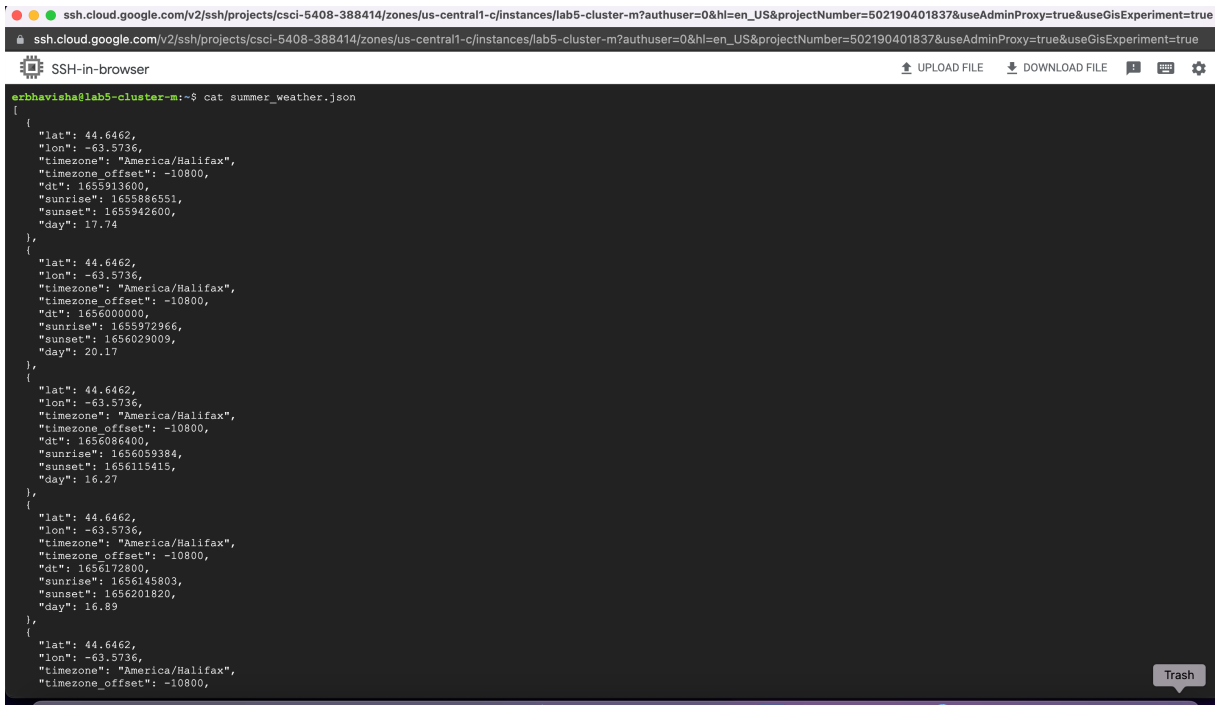

Figure 6: summer_weather.json file on GCP cluster

References:

[1] "MySQL Community Downloads," *MySQL* [Online]. Available: https://dev.mysql.com/downloads/workbench/ [Accessed: May 10, 2023].

[2] "Lab-5," *Brightspace Dalhousie University* [Online]. Available: https://dal.brightspace.com/d2l/le/content/271677/viewContent/3661458/Viewe [Accessed: June 21, 2023].