

---

*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.

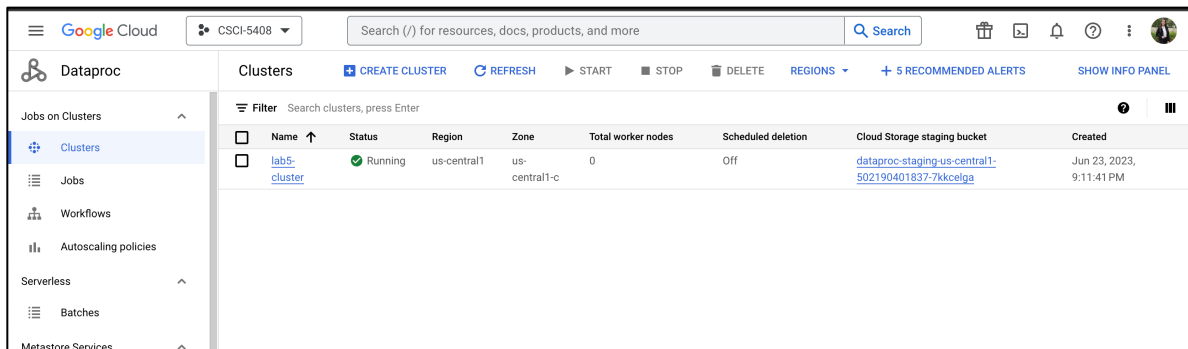
In the given lab did some hands-on for the Apache Spark program to fetch the data from given weather.json file

## 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\_weather.json”

## Lab exercise:

### 1. Completed setup of Apache Spark on GCP

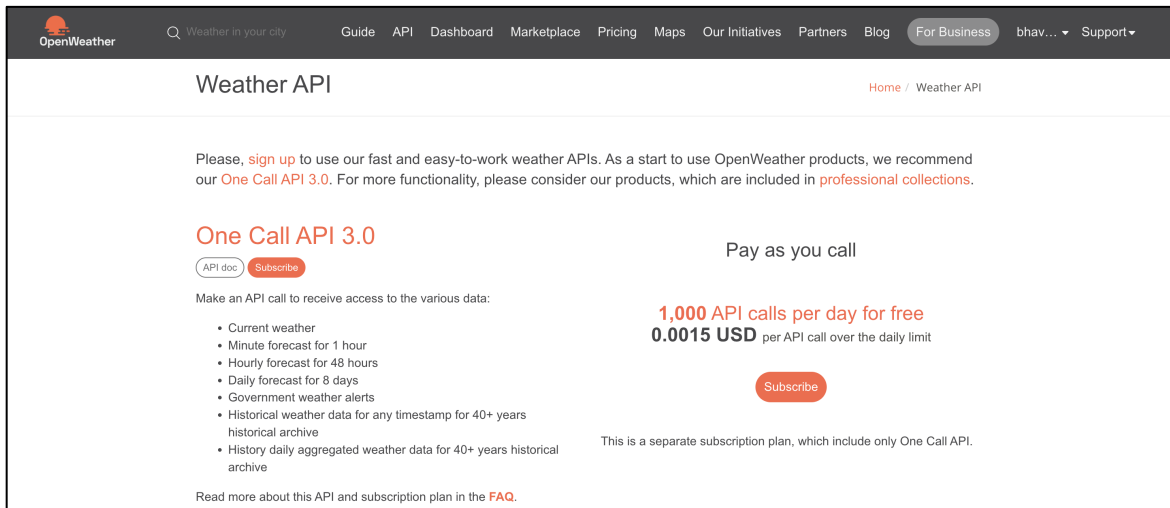


The screenshot shows the Google Cloud Dataproc Clusters page. The left sidebar contains navigation links for Clusters, Jobs, Workflows, Autoscaling policies, Serverless, Batches, and Metastore Services. The main panel displays a table of clusters. One cluster, 'lab5-cluster', is listed with a status of 'Running'. The table includes columns for Name, Status, Region, Zone, Total worker nodes, Scheduled deletion, Cloud Storage staging bucket, and Created.

Name	Status	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created
lab5-cluster	Running	us-central1	us-central1-c	0	Off	<a href="#">dataproc-staging-us-central1-502190401837-7kkcelga</a>	Jun 23, 2023, 9:11:41 PM

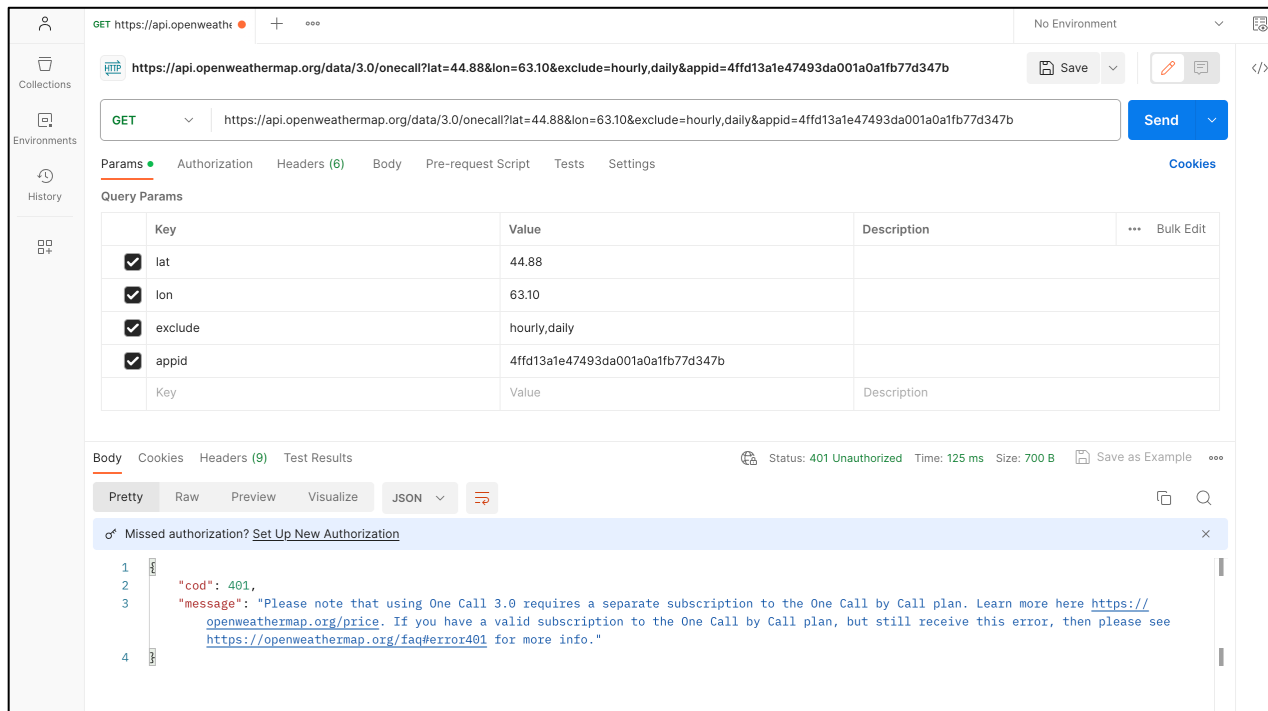
**Figure 1: Apache Spark Setup on GCP**

## 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.

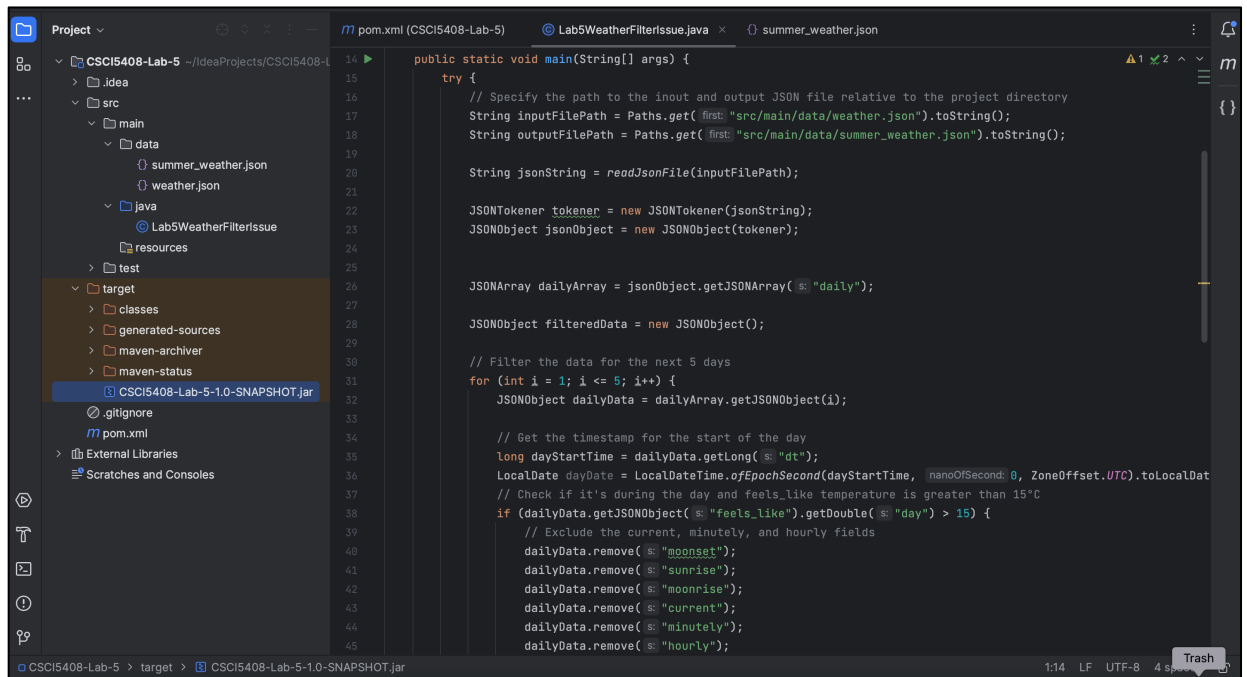


Figure 4: Java program for filtering the data.

##### 5. Save the filtered data into a new file – “summer wealther.json”

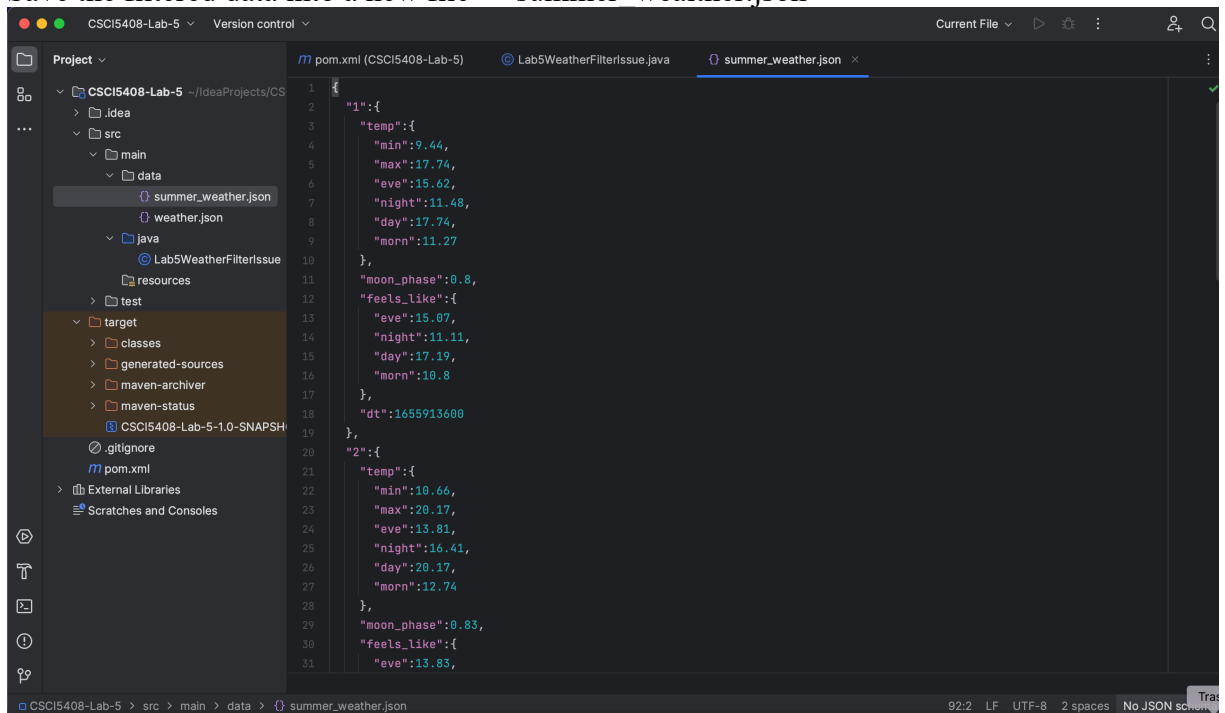
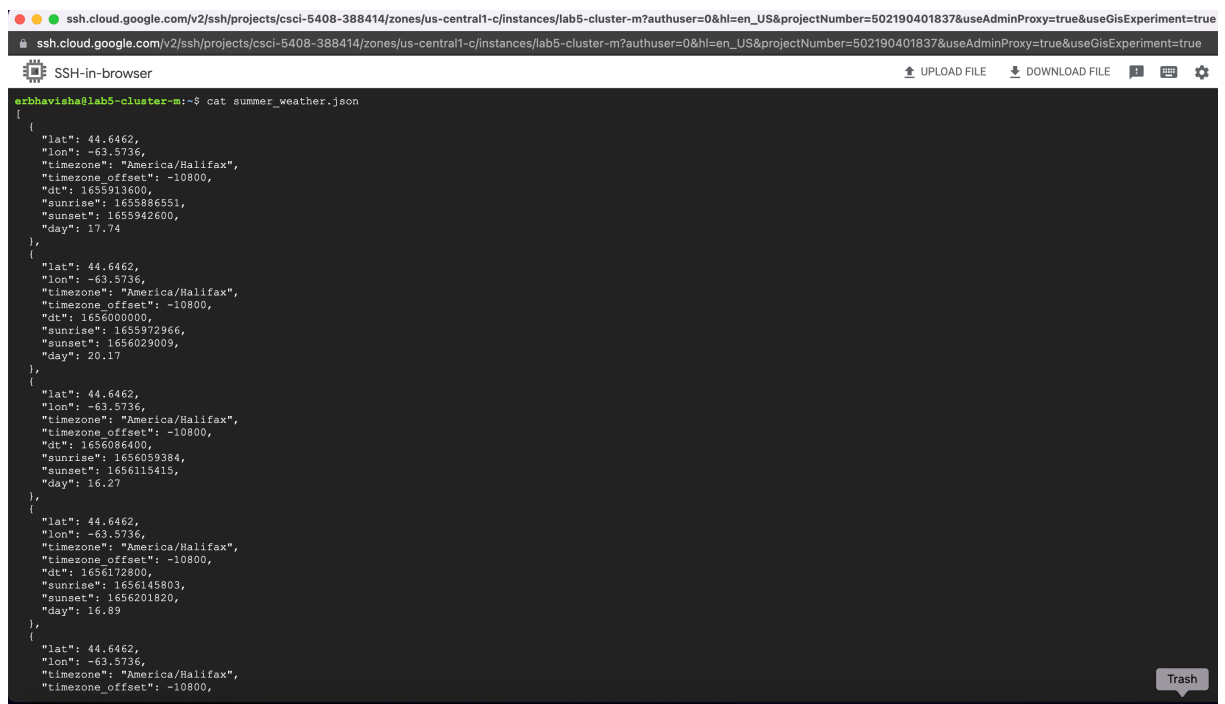


Figure 5: Filtered data stored in summer\_weather.json file

## 6. Checking on GCP for the same



```
ssh.cloud.google.com/v2/ssh/projects/csci-5408-388414/zones/us-central1-c/instances/lab5-cluster-m?authuser=0&hl=en_US&projectNumber=502190401837&useAdminProxy=true&useGisExperiment=true
ssh.cloud.google.com/v2/ssh/projects/csci-5408-388414/zones/us-central1-c/instances/lab5-cluster-m?authuser=0&hl=en_US&projectNumber=502190401837&useAdminProxy=true&useGisExperiment=true
SSH-in-browser
erbhavisha@lab5-cluster-m:~$ cat summer_weather.json
[
  {
    "lat": 44.6462,
    "lon": -63.5736,
    "timezone": "America/Halifax",
    "timezone_offset": -10800,
    "dt": 1655913600,
    "sunrise": 1655886551,
    "sunset": 1655942600,
    "day": 17.74
  },
  {
    "lat": 44.6462,
    "lon": -63.5736,
    "timezone": "America/Halifax",
    "timezone_offset": -10800,
    "dt": 1656000000,
    "sunrise": 1655972966,
    "sunset": 1656029009,
    "day": 20.17
  },
  {
    "lat": 44.6462,
    "lon": -63.5736,
    "timezone": "America/Halifax",
    "timezone_offset": -10800,
    "dt": 1656086400,
    "sunrise": 1656059384,
    "sunset": 1656115415,
    "day": 16.27
  },
  {
    "lat": 44.6462,
    "lon": -63.5736,
    "timezone": "America/Halifax",
    "timezone_offset": -10800,
    "dt": 1656172800,
    "sunrise": 1656145803,
    "sunset": 1656201820,
    "day": 16.89
  },
  {
    "lat": 44.6462,
    "lon": -63.5736,
    "timezone": "America/Halifax",
    "timezone_offset": -10800,
    "dt": 1656259200,
    "sunrise": 1656248151,
    "sunset": 1656320151,
    "day": 17.41
  }
]
```

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].