CSCI 5408

DATA MANAGEMENT AND WAREHOUSING

LAB ASSIGNMENT - 5

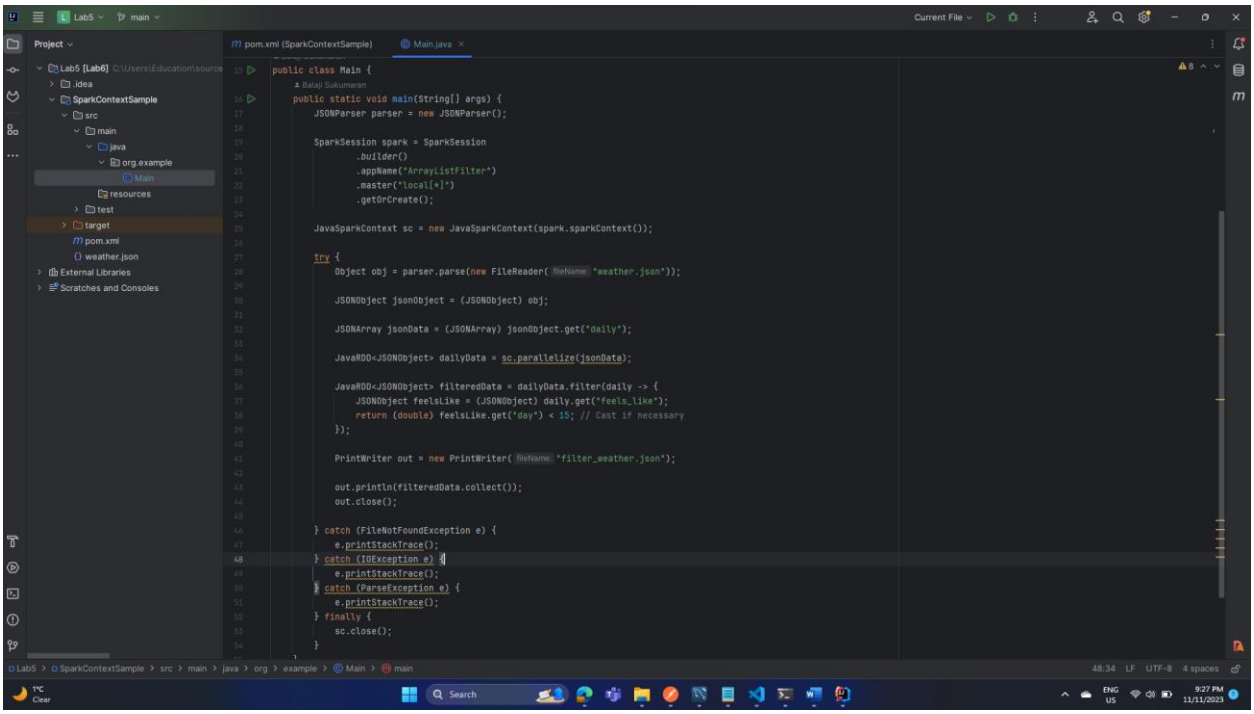
Banner ID: B00948977

Git Assignment Link :
https://git.cs.dal.ca/sukumaran/csci5408_f23_b00948977_balaji_sukumaran

Table of contents

Problem Statement 1: **Create a java program to find 5 days weather data**

The below java program will read the weather.json file and extract the data which whose feels like temperature is less than 5 C. **It uses the spark context to filter the daily weather object**.



```java
public class Main {
    public static void main(String[] args) {
        JSONParser parser = new JSONParser();

        SparkSession spark = SparkSession
                .builder()
                .appName("ArrayListFilter")
                .master("local[*]")
                .getOrCreate();

        JavaSparkContext sc = new JavaSparkContext(spark.sparkContext());

        try {
            Object obj = parser.parse(new FileReader( fileName: "weather.json"));

            JSONObject jsonObject = (JSONObject) obj;

            JSONArray jsonData = (JSONArray) jsonObject.get("daily");

            JavaRDD<JSONObject> dailyData = sc.parallelize(jsonData);

            JavaRDD<JSONObject> filteredData = dailyData.filter(daily -> {
                JSONObject feelsLike = (JSONObject) daily.get("feels_like");
                return (double) feelsLike.get("day") < 15; // Cast if necessary
            });

            PrintWriter out = new PrintWriter( fileName: "filter_weather.json");

            out.println(filteredData.collect());
            out.close();

        } catch (FileNotFoundException e) {
            e.printStackTrace();
        } catch (IOException e) {
            e.printStackTrace();
        } catch (ParseException e) {
            e.printStackTrace();
        } finally {
            sc.close();
        }
```

Figure 1: java program for parsing the weather file

Problem Statement 2: **Create an apache spark instance in GCP**

Created an apache spark instance in the Google cloud platform to run the java file.

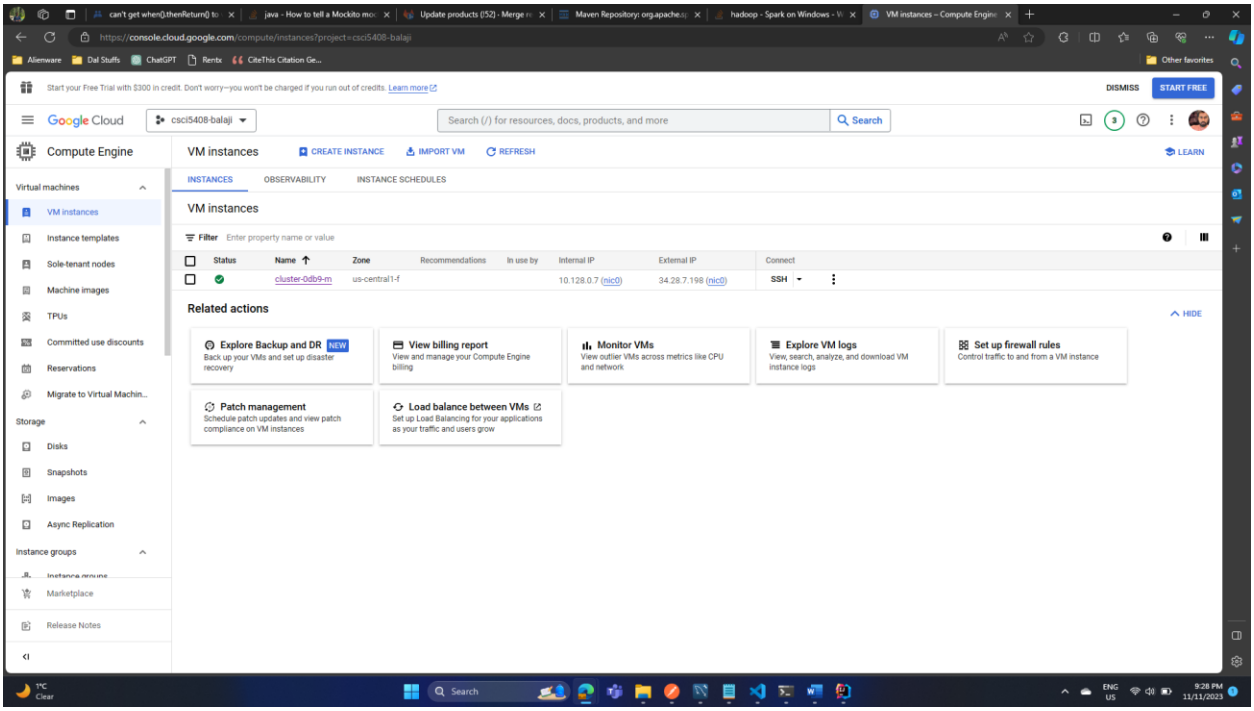


Figure 2: apache spark instance

Problem Statement 3: Execute the Java program along with its dependency JAR files in the create d apache spark cluster

Step 1: Uploaded the following java program JAR and json-simple-1.1.1.jar and weather.json

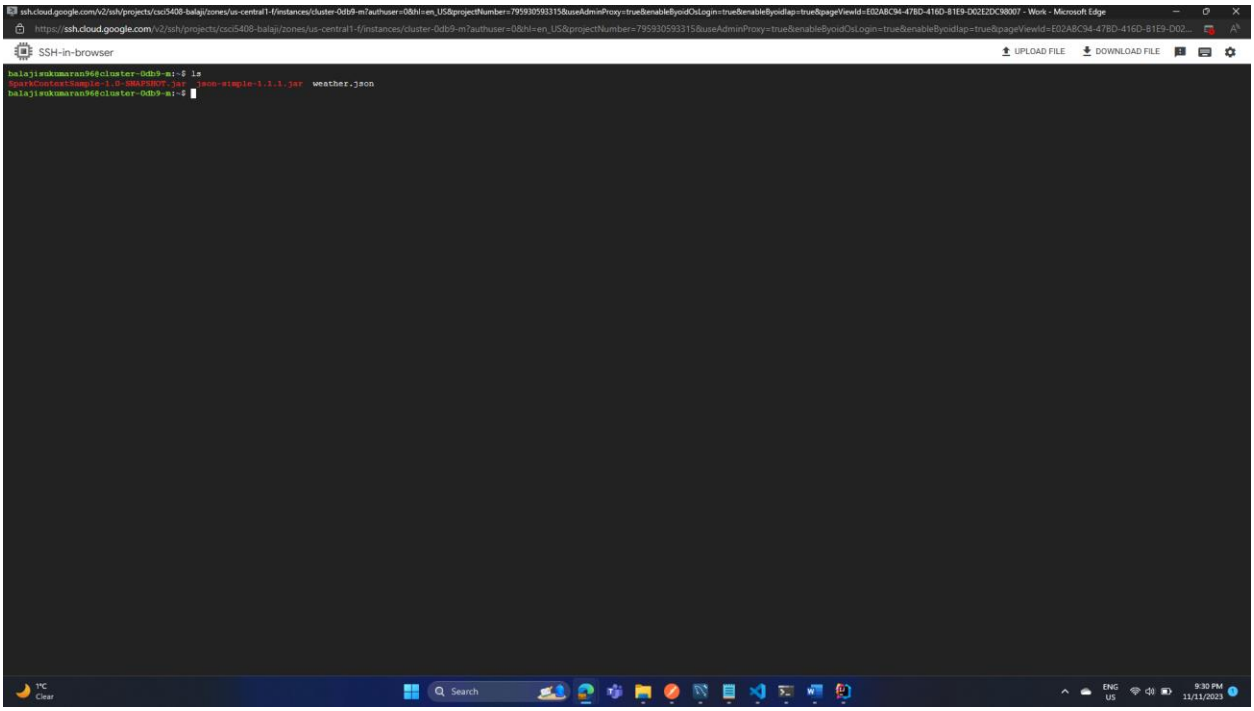


Figure 3: cluster contents

Step 2: Execute the JAR file with the dependency JARs using this commend “spark-submit --jars ./json-simple-1.1.1.jar --class org.example.Main ./SparkContextSample-1.0-SNAPSHOT.jar”

Executed successfully.

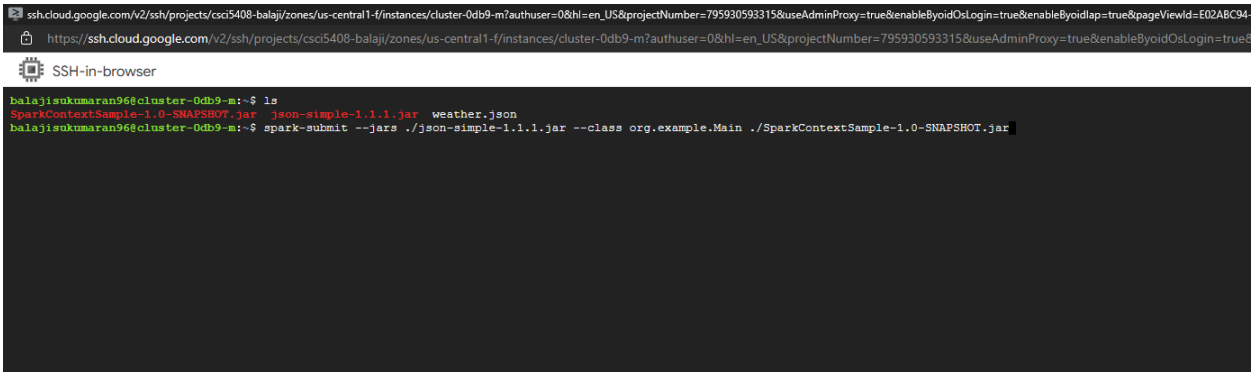


Figure 4: executed successfully

Step 3: Following is the generated output file.

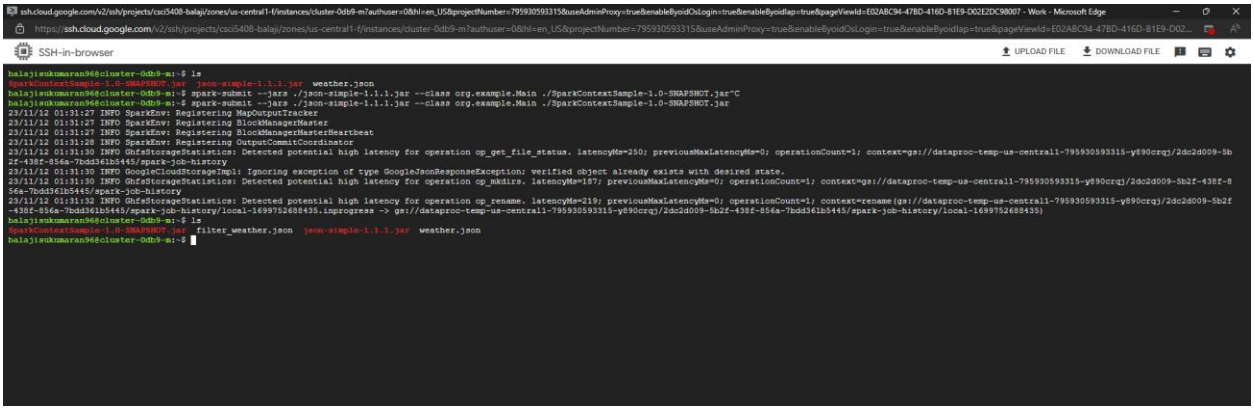


Figure 5: output file is generated

Following is the output file.

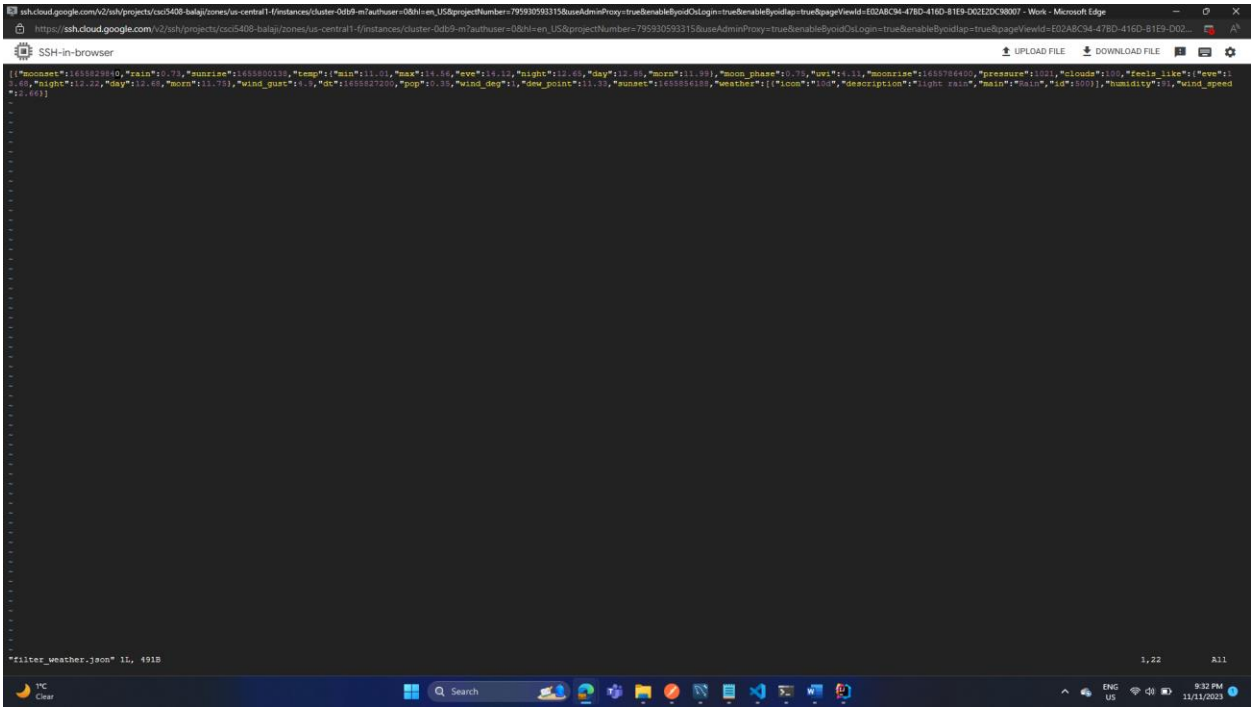


Figure 6: output file contents

REFERRRENCES:

[1] "Apache Hadoop," IBM. [Online]. Available: https://www.ibm.com/analytics/hadoop. [Accessed: 21-Oct-2023].

[2] "Spark SQL, DataFrames and datasets guide," Spark SQL and DataFrames - Spark3.0.1 Documentation. [Online]. Available: https://spark.apache.org/docs/3.0.1/sql-programming-guide.html. [Accessed: 21-Oct-2023].