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...…………………………………………………. | **1** |
| **Problem Statement 2:** Create an apache spark instance in GCP.……………………………………………………………… | **2** |
| **Problem Statement 3:** Execute the Java program along with its dependency………...……………………………………... | **3** |

**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**.

A screen shot of a computer

Description automatically generated

Figure : 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.

A screenshot of a computer

Description automatically generated

Figure : 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

A computer screen with a black screen

Description automatically generated

Figure : 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.

A screenshot of a computer

Description automatically generated

Figure : executed successfully

**Step 3:** Following is the generated output file.

A screen shot of a computer

Description automatically generated

Figure : output file is generated

Following is the output file.

A screenshot of a computer

Description automatically generated

Figure : 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].