CSCI 5408

DATA MANAGEMENT AND WAREHOUSING

LAB - 5

Banner ID: B00984406

GitLab Assignment Link: <https://git.cs.dal.ca/jems/csci5408_s24_b00984406_jems_patel.git>

Table of Contents

Screenshots of the step-by-step process followed to create the Apache Spark (GCP Dataproc) cluster and execute the job (Addition.jar) file on it…………………………………………………….3

Report any challenges faced while executing the .jar file on the Apache spark cluster…………..8

Explanation of the Java Spark program with the screenshots of the code…………………………9

References……………………………………………..………………………………………………..10

**Screenshots of the step-by-step process followed to create the Apache Spark (GCP Dataproc) cluster and execute the job (Addition.jar) file on it.**

Here are the steps that are taken to do this:  
  
Step:1

A screenshot of a computer

Description automatically generated

Figure 1.1: Enable the API

Step:2

A screenshot of a computer

Description automatically generated

Figure 1.2: Searching the cloud Dataproc

Step:3

A screenshot of a computer

Description automatically generated

Figure 1.3: Setup the cluster

Step:4

A screenshot of a computer

Description automatically generated

Figure 1.4: Configure the manager node

Step:5

A screenshot of a computer

Description automatically generated

Figure 1.5: Configure the worker node

Step:6

A screenshot of a computer

Description automatically generated

Figure 1.6: Creating the cluster

Step:7

A screenshot of a computer

Description automatically generated

Figure 1.7: List of VM Instances

Step:8

A screenshot of a computer screen

Description automatically generated

Figure 1.8: Uploading the jar file of the java code

Step:9

A screenshot of a computer

Description automatically generated

Figure 1.9: list the uploaded files

Step:10

A screenshot of a computer

Description automatically generated

Figure 1.10: Sum of the numbers

**Report any challenges faced while executing the .jar file on the Apache spark cluster.**

While executing the .jar file in the Apache Spark cluster, I encountered two issues.

The first issue was related to creating the cluster according to the provided steps. Despite following the instructions correctly, the process often failed after making me wait up to 30 minutes. I resolved this by changing the configuration for the worker node to N1, which eventually succeeded.

The second issue involved accessing the input file in the Java code. Although I uploaded the input.txt file via the SSH terminal, the code couldn't locate it. Upon investigation, I found that the file wasn't present in Hadoop. To address this, I first created the necessary directory using the following command:

**hadoop fs -mkdir -p hdfs://lab-5-m/user/jems007patel**

After this I copy the txt file into hadoop manually by using this given command and successfully run the jar file  
  
 **hadoop fs -copyFromLocal ./input.txt user/jems007patel/input.txt**

**Explanation of the Java Spark program with the screenshots of the code.**

**A screenshot of a computer program

Description automatically generated**

Figure 1.11: Java Spark Code

Here is the step-by-step explanation for the above java code:

* To initialize this Spark application, a SparkConf object is created, where the property conf.setAppName(”SumOfNumbers”) sets the name of the application and conf.setMaster(”local”) specifies to run the application locally.
* Using this configuration, a `JavaSparkContext` is created, opening the gate to Spark functionalities.
* After that this program reads `input.txt` and convert file into an RDD, where an element in an RDD is equal to a line in the file.
* FlatMap Transformation Splits each line into numbers and then RDD of number is flattened out as an RDD of strings.
* Map Transformation Transforms all the string numbers, present in the input string, into integers.
* Reduce Action Reduces the RDD to single value by summing all the Integers.
* The sum of all these numbers is the total sum and it is printed in the console and shuts down the Spark context.

**References**

[1] "Apache Spark Documentation," Apache Spark, [Online]. Available: https://spark.apache.org/docs/latest/. [Accessed 16 June 2024].

[2] " K. Saxena, "All You Need to Know About Google Cloud Dataproc," Medium, [Online].

Available: <https://medium.com/google-cloud/all-you-need-to-know-about-google-cloud->

dataproc-23fe91369678. [Accessed 16 June 2024].