CSCI-5408

DATA MANAGEMENT, WAREHOUSING, & ANALYTICS

LAB ASSIGNMENT - 5

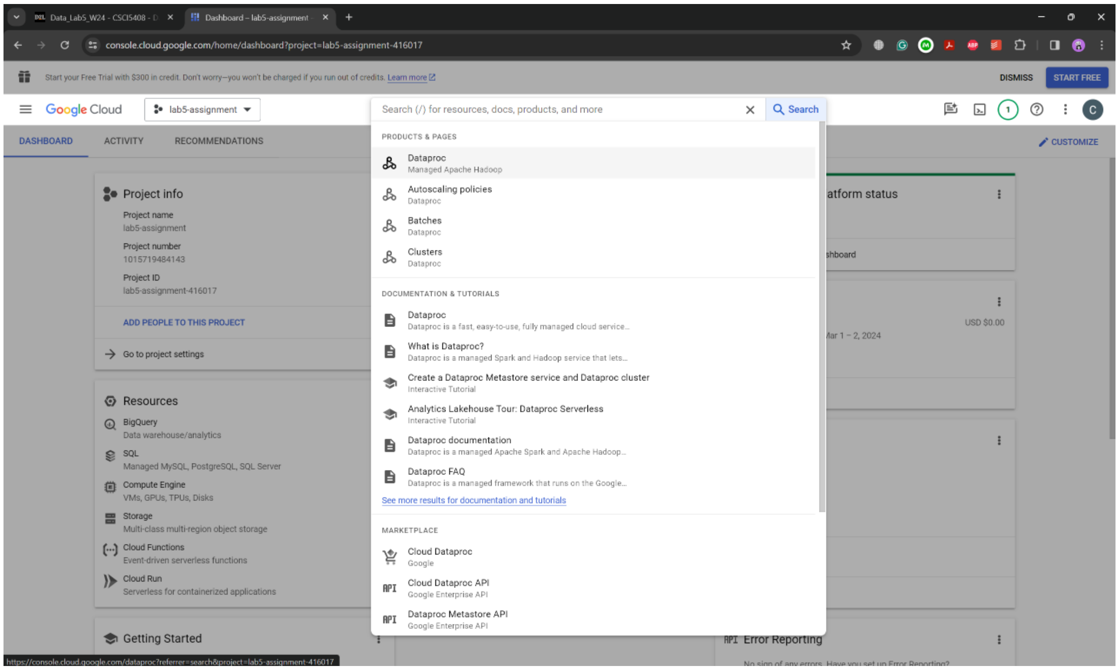
Banner ID: B00977669

GitLab Assignment Link:
https://git.cs.dal.ca/saji/csci5408_w24_b00977669_christin_saji

Table of Contents

Task 1: Screenshots of the step-by-step process followed to create the Apache Spark (GCP Dataproc) cluster and execute the job (WordCounter.jar) file on it.

First, I created a project named lab5-assignment then selected Dataproc.



console.cloud.google.com/home/dashboard?project=lab5-assignment-416017

Start your Free Trial with $300 in credit. Don't worry—you won't be charged if you run out of credits. Learn more

Project info

Project name
lab5-assignment

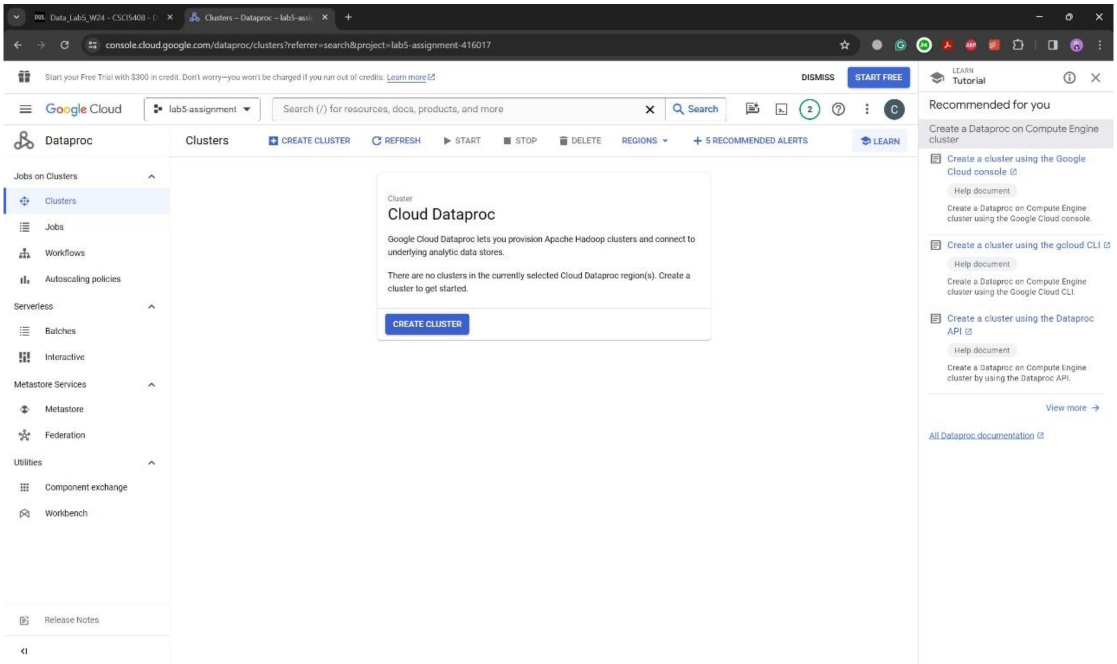
Project number
1015719484143

Project ID
lab5-assignment-416017

https://console.cloud.google.com/dataproc?referrer=search&project=lab5-assignment-416017

Data_Lab5_W24 - CSCI5408
Dashboard – lab5-assignment
DISMISS
START FREE
Google Cloud
lab5-assignment
Search (/) for resources, docs, products, and more
Search
DASHBOARD
ACTIVITY
RECOMMENDATIONS
CUSTOMIZE
ADD PEOPLE TO THIS PROJECT
Go to project settings
Resources
BigQuery
Data warehouse/analytics
SQL
Managed MySQL, PostgreSQL, SQL Server
Compute Engine
VMs, GPUs, TPUs, Disks
Storage
Multi-class multi-region object storage
Cloud Functions
Event-driven serverless functions
Cloud Run
Serverless for containerized applications
Getting Started
PRODUCTS & PAGES
Dataproc
Managed Apache Hadoop
Autoscaling policies
Dataproc
Batches
Dataproc
Clusters
Dataproc
DOCUMENTATION & TUTORIALS
Dataproc
Dataproc is a fast, easy-to-use, fully managed cloud service...
What is Dataproc?
Dataproc is a managed Spark and Hadoop service that lets...
Create a Dataproc Metastore service and Dataproc cluster
Interactive Tutorial
Analytics Lakehouse Tour: Dataproc Serverless
Interactive Tutorial
Dataproc documentation
Dataproc is a managed Apache Spark and Apache Hadoop...
Dataproc FAQ
Dataproc is a managed framework that runs on the Google...
See more results for documentation and tutorials
MARKETPLACE
Cloud Dataproc
Google
Cloud Dataproc API
Google Enterprise API
Dataproc Metastore API
Google Enterprise API
atform status
shboard
USD $0.00
lar 1 – 2, 2024
Error Reporting
No sign of any errors. Have you set up Error Reporting?

Figure 1 Selected Dataproc



console.cloud.google.com/dataproc/clusters?referrer=search&project=lab5-assignment-416017

Start your Free Trial with $300 in credit. Don't worry—you won't be charged if you run out of credits. Learn more

Cluster

Cloud Dataproc

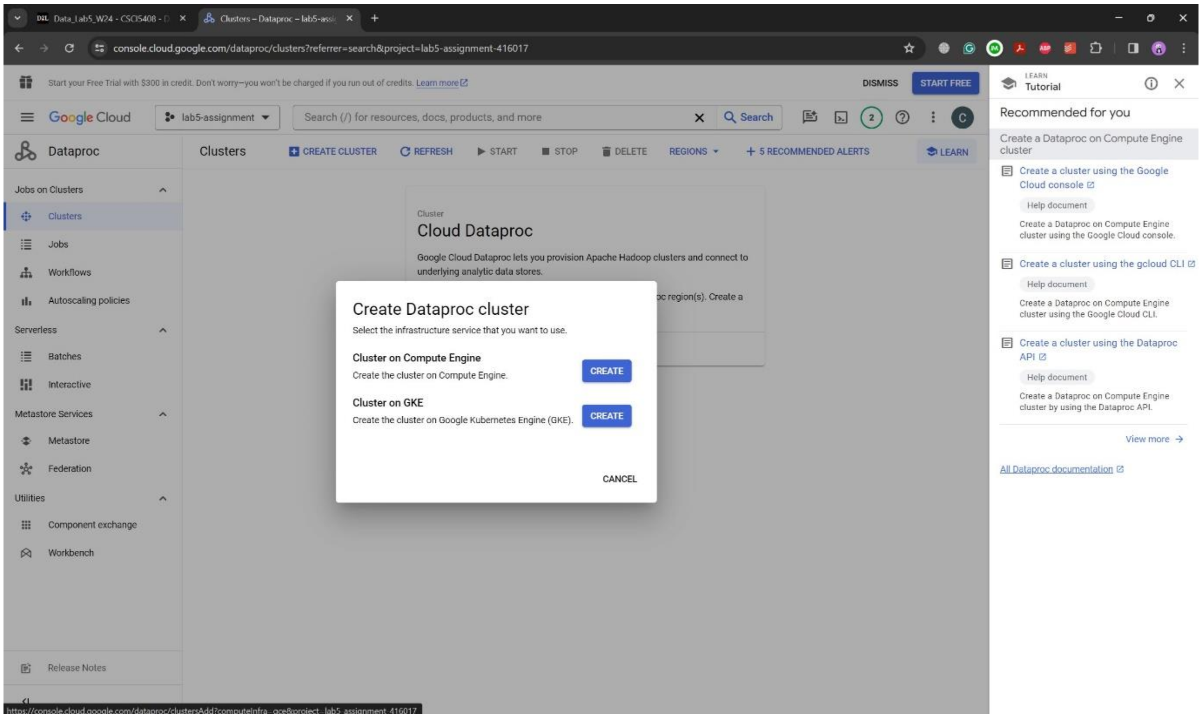
Google Cloud Dataproc lets you provision Apache Hadoop clusters and connect to underlying analytic data stores.

There are no clusters in the currently selected Cloud Dataproc region(s). Create a cluster to get started.

CREATE CLUSTER

Data_Lab5_W24 - CSCI5408
Clusters – Dataproc – lab5-assi
DISMISS
START FREE
Google Cloud
lab5-assignment
Search (/) for resources, docs, products, and more
Search
Dataproc
Clusters
CREATE CLUSTER
REFRESH
START
STOP
DELETE
REGIONS
5 RECOMMENDED ALERTS
LEARN
Jobs on Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Interactive
Metastore Services
Metastore
Federation
Utilities
Component exchange
Workbench
Release Notes
LEARN
Tutorial
Recommended for you
Create a Dataproc on Compute Engine cluster
Create a cluster using the Google Cloud console
Help document
Create a Dataproc on Compute Engine cluster using the Google Cloud console.
Create a cluster using the gcloud CLI
Help document
Create a Dataproc on Compute Engine cluster using the Google Cloud CLI.
Create a cluster using the Dataproc API
Help document
Create a Dataproc on Compute Engine cluster by using the Dataproc API.
View more
All Dataproc documentation

Figure 2 Create Cluster


Data_Lab5_W24 - CSCI5408 - D
Clusters – Dataproc – lab5-assi
console.cloud.google.com/dataproc/clusters?referrer=search&project=lab5-assignment-416017
Start your Free Trial with $300 in credit. Don't worry—you won't be charged if you run out of credits. Learn more
DISMISS
START FREE
Google Cloud
lab5-assignment
Search (/) for resources, docs, products, and more
Search
Dataproc
Clusters
CREATE CLUSTER
REFRESH
START
STOP
DELETE
REGIONS
5 RECOMMENDED ALERTS
LEARN
Jobs on Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Interactive
Metastore Services
Metastore
Federation
Utilities
Component exchange
Workbench
Release Notes
LEARN
Tutorial
Recommended for you
Create a Dataproc on Compute Engine cluster
Create a cluster using the Google Cloud console
Help document
Create a cluster using the gcloud CLI
Help document
Create a cluster using the Dataproc API
Help document
View more
All Dataproc documentation

Cluster

Cloud Dataproc

Google Cloud Dataproc lets you provision Apache Hadoop clusters and connect to underlying analytic data stores.

Create Dataproc cluster

Select the infrastructure service that you want to use.

Cluster on Compute Engine

Create the cluster on Compute Engine. CREATE

Cluster on GKE

Create the cluster on Google Kubernetes Engine (GKE). CREATE

CANCEL

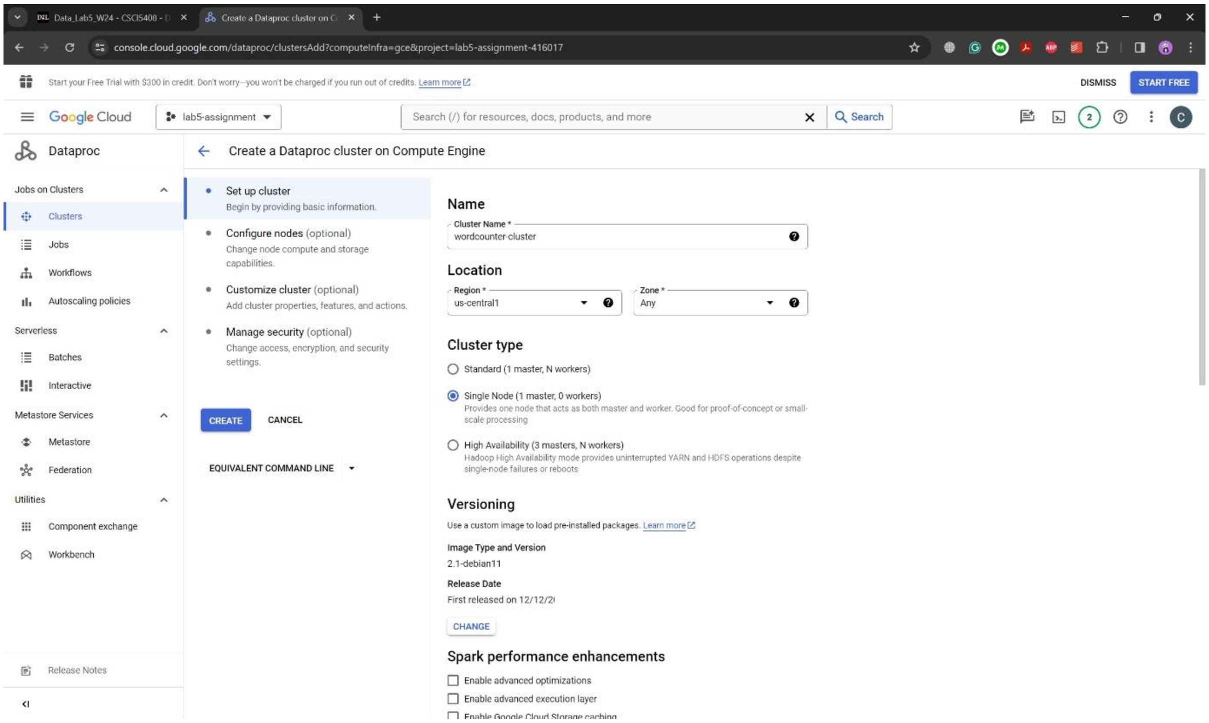
Create a Dataproc on Compute Engine cluster using the Google Cloud console.

Create a Dataproc on Compute Engine cluster using the Google Cloud CLI.

Create a Dataproc on Compute Engine cluster by using the Dataproc API.

Figure 3 Cluster on Compute Engine

Renamed the cluster name to wordcounter-cluster and selected single node option for cluster type.


Data_Lab5_W24 - CSCI5408 - D
Create a Dataproc cluster on C
console.cloud.google.com/dataproc/clustersAdd?computeInfra=gce&project=lab5-assignment-416017
Start your Free Trial with $300 in credit. Don't worry—you won't be charged if you run out of credits. Learn more
DISMISS
START FREE
Google Cloud
lab5-assignment
Search (/) for resources, docs, products, and more
Search
Dataproc
Jobs on Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Interactive
Metastore Services
Metastore
Federation
Utilities
Component exchange
Workbench
Release Notes
CREATE
CANCEL
EQUIVALENT COMMAND LINE
CHANGE

Create a Dataproc cluster on Compute Engine

- Set up cluster
  Begin by providing basic information.
- Configure nodes (optional)
  Change node compute and storage capabilities.
- Customize cluster (optional)
  Add cluster properties, features, and actions.
- Manage security (optional)
  Change access, encryption, and security settings.

Name

Cluster Name * wordcounter-cluster

Location

Region * us-central1

Zone * Any

Cluster type

○ Standard (1 master, N workers)

● Single Node (1 master, 0 workers)
Provides one node that acts as both master and worker. Good for proof-of-concept or small-scale processing

○ High Availability (3 masters, N workers)
Hadoop High Availability mode provides uninterrupted YARN and HDFS operations despite single-node failures or reboots

Versioning

Use a custom image to load pre-installed packages. Learn more

Image Type and Version
2.1-debian11

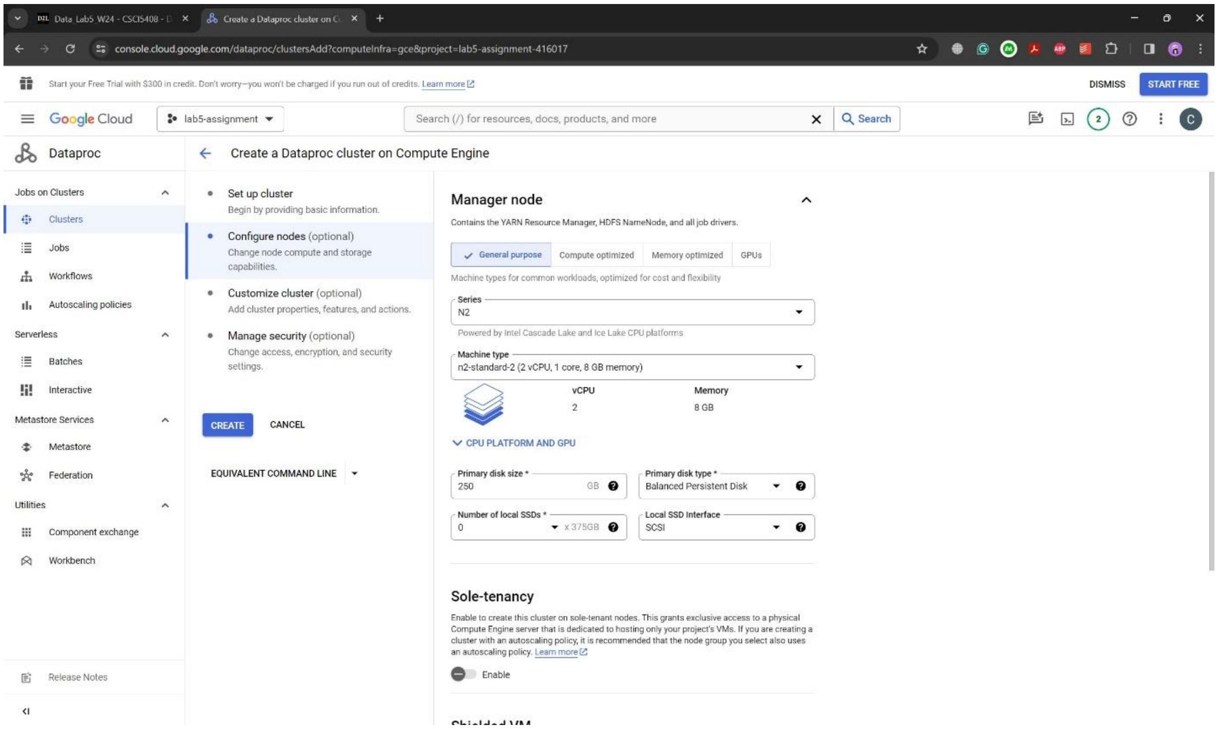
Release Date
First released on 12/12/2(

Spark performance enhancements

□ Enable advanced optimizations
□ Enable advanced execution layer
□ Enable Google Cloud Storage caching

Figure 4 Settings for set up cluster

In configure nodes, I changed machine type to 2 vCpu with 8 GB memory and primary disk size to 250 GB.



Create a Dataproc cluster on Compute Engine

- Set up cluster
  Begin by providing basic information.
- Configure nodes (optional)
  Change node compute and storage capabilities.
- Customize cluster (optional)
  Add cluster properties, features, and actions.
- Manage security (optional)
  Change access, encryption, and security settings.

Manager node

Contains the YARN Resource Manager, HDFS NameNode, and all job drivers.

Machine types for common workloads, optimized for cost and flexibility

Series: N2

Powered by Intel Cascade Lake and Ice Lake CPU platforms

Machine type: n2-standard-2 (2 vCPU, 1 core, 8 GB memory)

vCPU	Memory
2	8 GB

Primary disk size *: 250 GB

Primary disk type *: Balanced Persistent Disk

Number of local SSDs *: 0 x 375GB

Local SSD Interface: SCSI

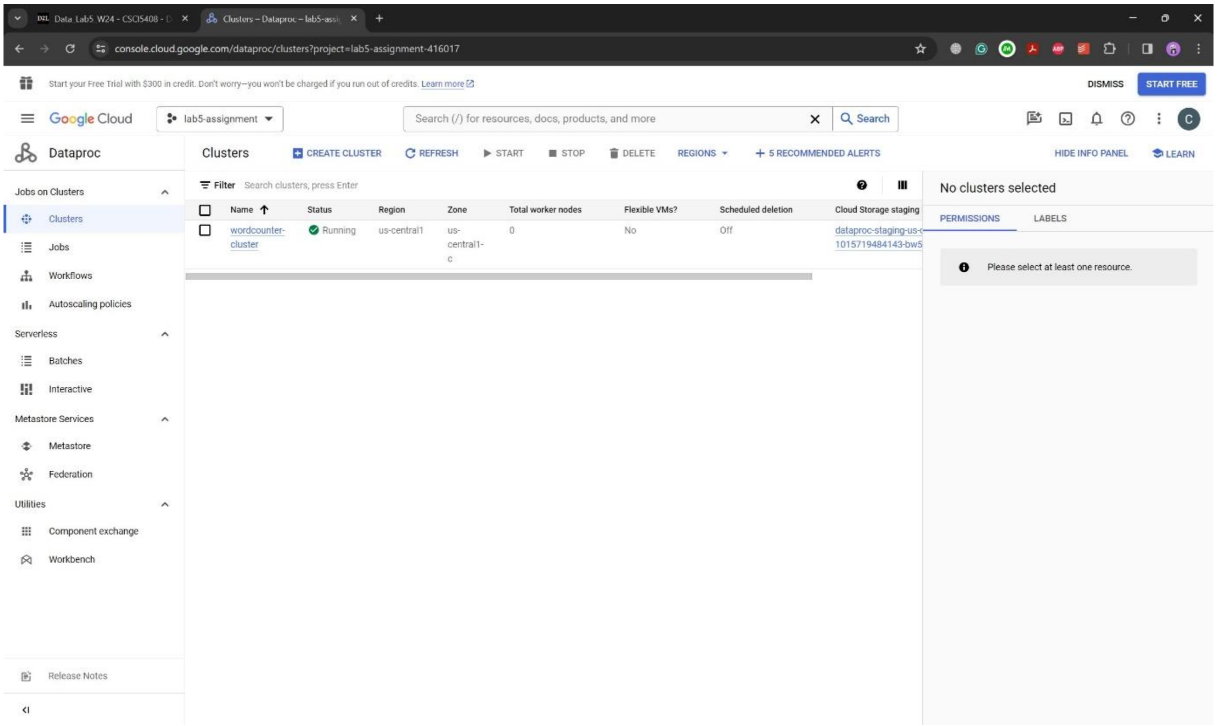
Sole-tenancy

Enable to create this cluster on sole-tenant nodes. This grants exclusive access to a physical Compute Engine server that is dedicated to hosting only your project's VMs. If you are creating a cluster with an autoscaling policy, it is recommended that the node group you select also uses an autoscaling policy. Learn more

Enable

Shielded VM

Figure 5 Settings for configure nodes



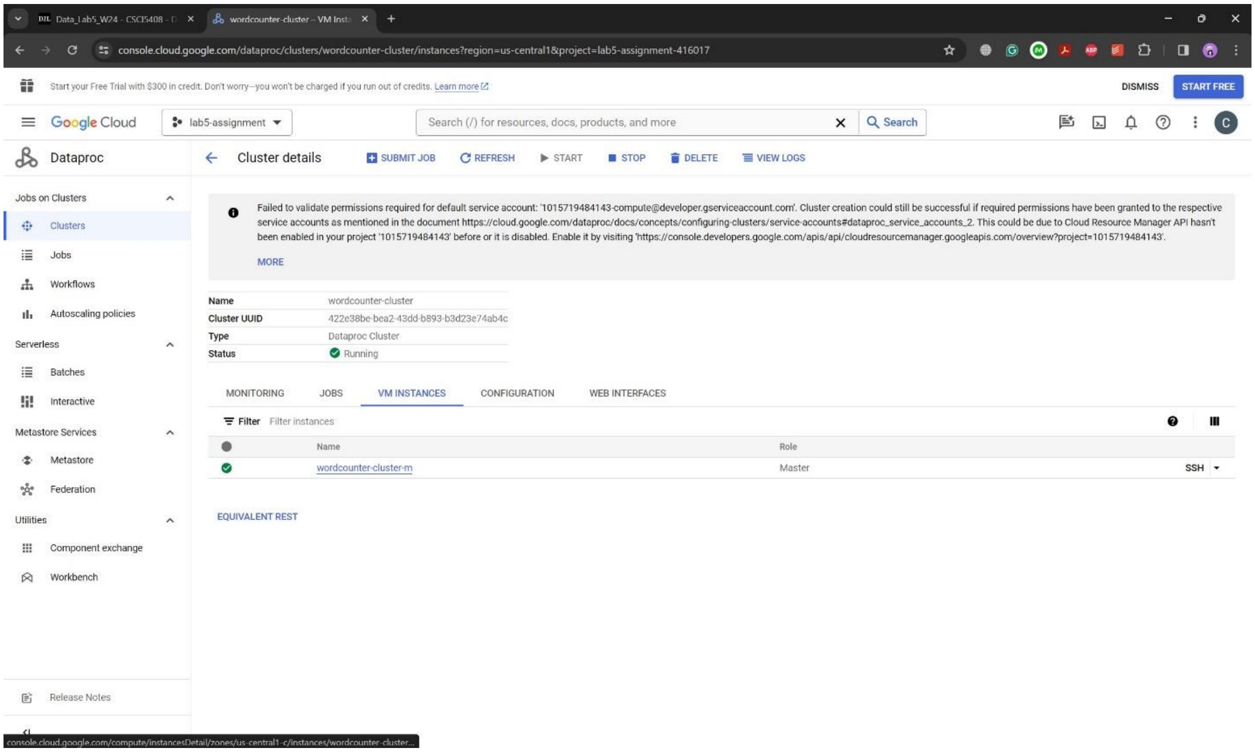
Clusters

Name	Status	Region	Zone	Total worker nodes	Flexible VMs?	Scheduled deletion	Cloud Storage staging
wordcounter-cluster	Running	us-central1	us-central1-c	0	No	Off	dataproc-staging-us-c 1015719484143-bw5

No clusters selected

Please select at least one resource.

Figure 6 Clusters menu

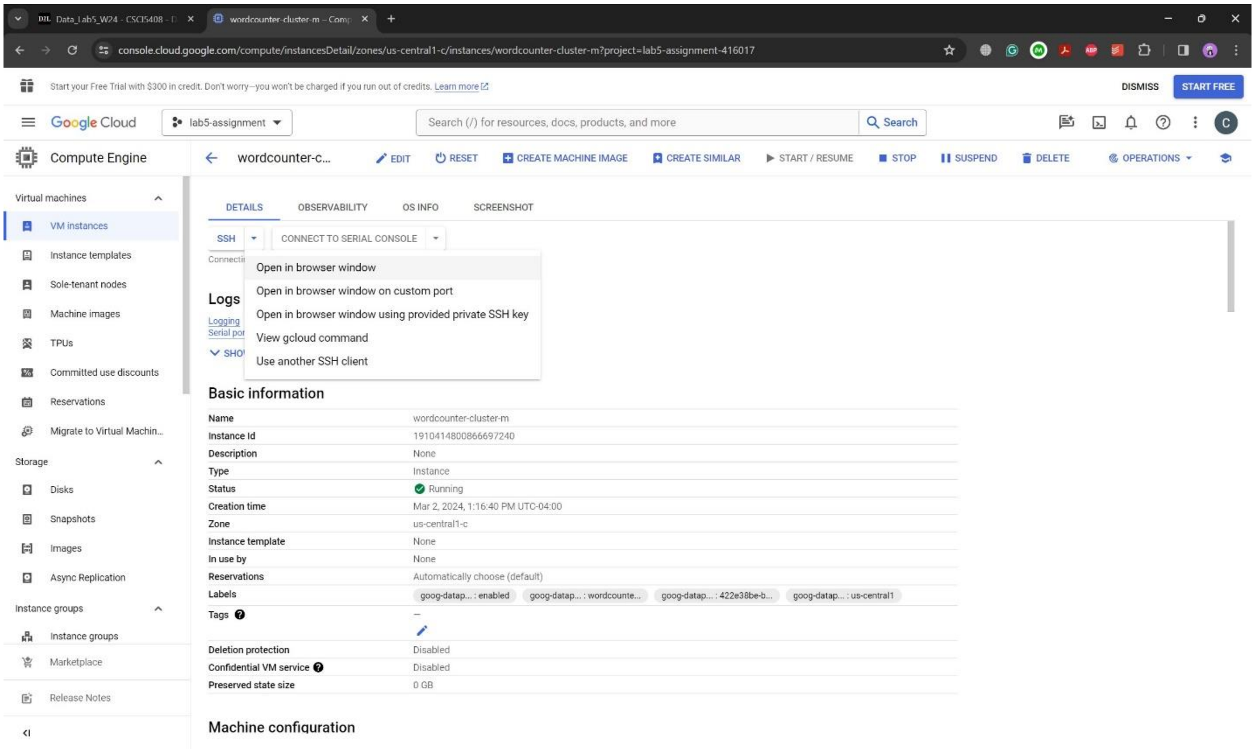

Data_Lab5_W24 - CSCI5408
wordcounter-cluster – VM Insta
console.cloud.google.com/dataproc/clusters/wordcounter-cluster/instances?region=us-central1&project=lab5-assignment-416017
Start your Free Trial with $300 in credit. Don't worry—you won't be charged if you run out of credits. Learn more
DISMISS
START FREE
Google Cloud
lab5-assignment
Search (/) for resources, docs, products, and more
Search
Dataproc
Jobs on Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Interactive
Metastore Services
Metastore
Federation
Utilities
Component exchange
Workbench
Release Notes
Cluster details
SUBMIT JOB
REFRESH
START
STOP
DELETE
VIEW LOGS
MORE
MONITORING
JOBS
VM INSTANCES
CONFIGURATION
WEB INTERFACES
Filter
Filter instances
SSH
EQUIVALENT REST
console.cloud.google.com/compute/instancesDetail/zones/us-central1-c/instances/wordcounter-cluster...

Failed to validate permissions required for default service account: '1015719484143-compute@developer.gserviceaccount.com'. Cluster creation could still be successful if required permissions have been granted to the respective service accounts as mentioned in the document https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/service-accounts#dataproc_service_accounts_2. This could be due to Cloud Resource Manager API hasn't been enabled in your project '1015719484143' before or it is disabled. Enable it by visiting 'https://console.developers.google.com/apis/api/cloudresourcemanager.googleapis.com/overview?project=1015719484143'.

Name	wordcounter-cluster
Cluster UUID	422e38be-bea2-43dd-b893-b3d23e74ab4c
Type	Dataproc Cluster
Status	✓ Running

●	Name	Role
✓	wordcounter-cluster-m	Master

Figure 7 VM instances for wordcounter-cluster

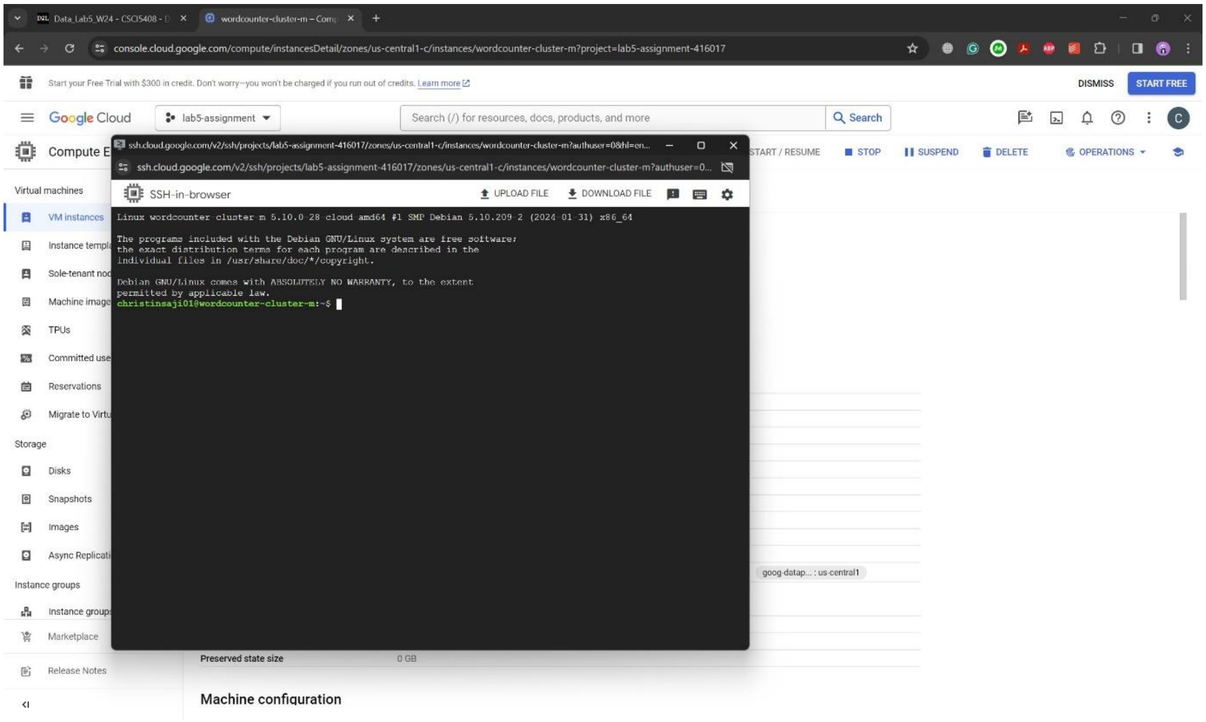


Basic information

Name	wordcounter-cluster-m
Instance Id	1910414800866697240
Description	None
Type	Instance
Status	✓ Running
Creation time	Mar 2, 2024, 1:16:40 PM UTC-04:00
Zone	us-central1-c
Instance template	None
In use by	None
Reservations	Automatically choose (default)
Labels	goog-datap... : enabled goog-datap... : wordcounte... goog-datap... : 422e38be-b... goog-datap... : us-central1
Tags	–
Deletion protection	Disabled
Confidential VM service	Disabled
Preserved state size	0 GB

Machine configuration

Figure 8 SSH in browser window


console.cloud.google.com/compute/instancesDetail/zones/us-central1-c/instances/wordcounter-cluster-m?project=lab5-assignment-416017
Start your Free Trial with $300 in credit. Don't worry—you won't be charged if you run out of credits. Learn more
DISMISS
START FREE
Google Cloud
lab5-assignment
Search (/) for resources, docs, products, and more
Search
ssh.cloud.google.com/v2/ssh/projects/lab5-assignment-416017/zones/us-central1-c/instances/wordcounter-cluster-m?authuser=0&hl=en...
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
START / RESUME
STOP
SUSPEND
DELETE
OPERATIONS
Virtual machines
VM instances
Storage
Disks
Snapshots
Images
Instance groups
Marketplace
Release Notes
Preserved state size
0 GB
Machine configuration

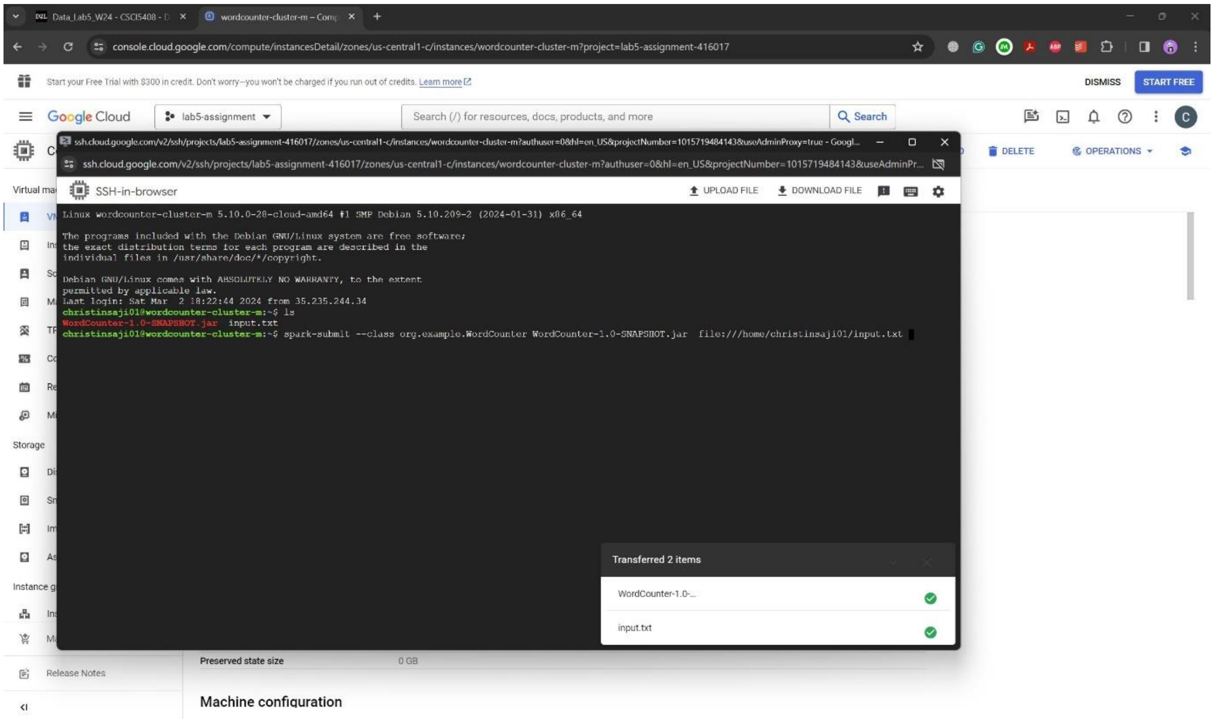
```
Linux wordcounter cluster m 5.10.0 28 cloud amd64 #1 SMP Debian 5.10.209 2 (2024 01 31) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
christinsaji01@wordcounter-cluster-m:~$
```

Figure 9 SSH-in browser terminal

I uploaded WordCounter-1.0-SNAPSHOT.jar and input.txt


console.cloud.google.com/compute/instancesDetail/zones/us-central1-c/instances/wordcounter-cluster-m?project=lab5-assignment-416017
ssh.cloud.google.com/v2/ssh/projects/lab5-assignment-416017/zones/us-central1-c/instances/wordcounter-cluster-m?authuser=0&hl=en_US&projectNumber=1015719484143&useAdminProxy=true - Googl...
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
Transferred 2 items
WordCounter-1.0-...
input.txt
Preserved state size
0 GB
Machine configuration

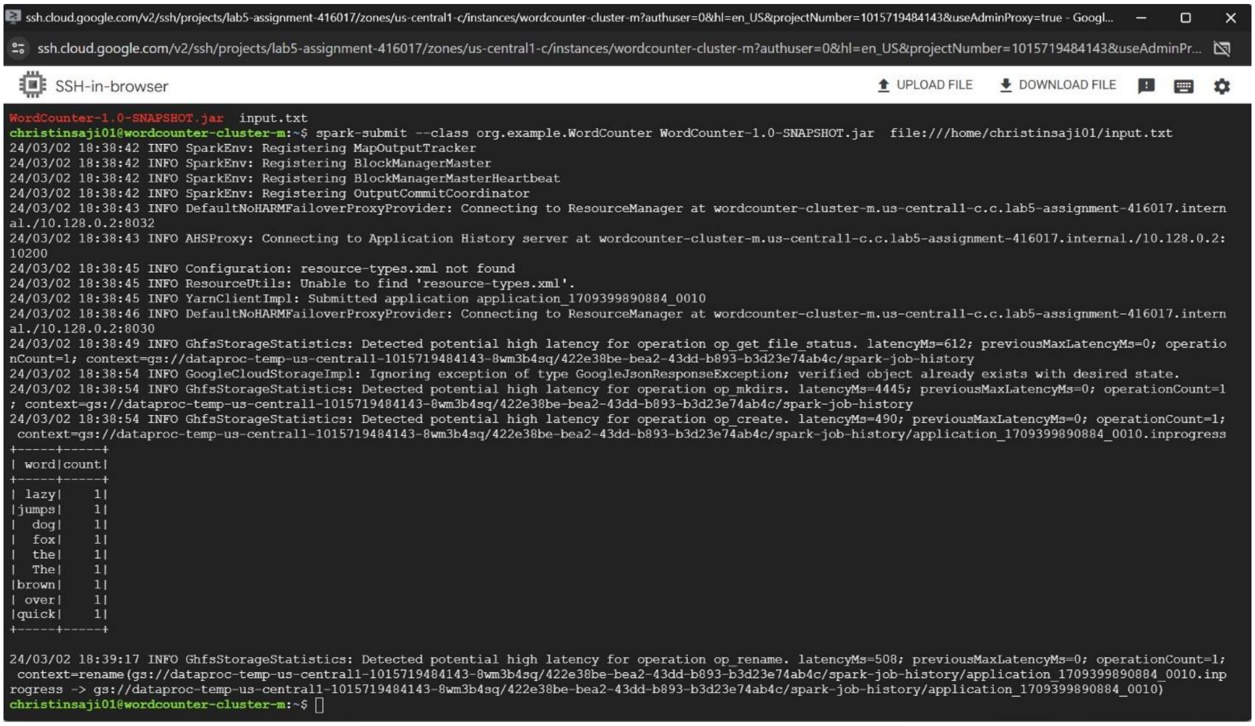
```
Linux wordcounter-cluster-m 5.10.0-28-cloud-amd64 #1 SMP Debian 5.10.209-2 (2024-01-31) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Sat Mar  2 18:22:44 2024 from 35.235.244.34
christinsaji01@wordcounter-cluster-m:~$ ls
WordCounter-1.0-SNAPSHOT.jar  input.txt
christinsaji01@wordcounter-cluster-m:~$ spark-submit --class org.example.WordCounter WordCounter-1.0-SNAPSHOT.jar  file:///home/christinsaji01/input.txt
```

Figure 10 Uploaded files in the terminal

Executed the WordCounter program using the command “spark-submit –class org.example.WordCounter WordCounter-1.0-SNAPSHOT.jar file:///home/christinsaji01/input.txt”.



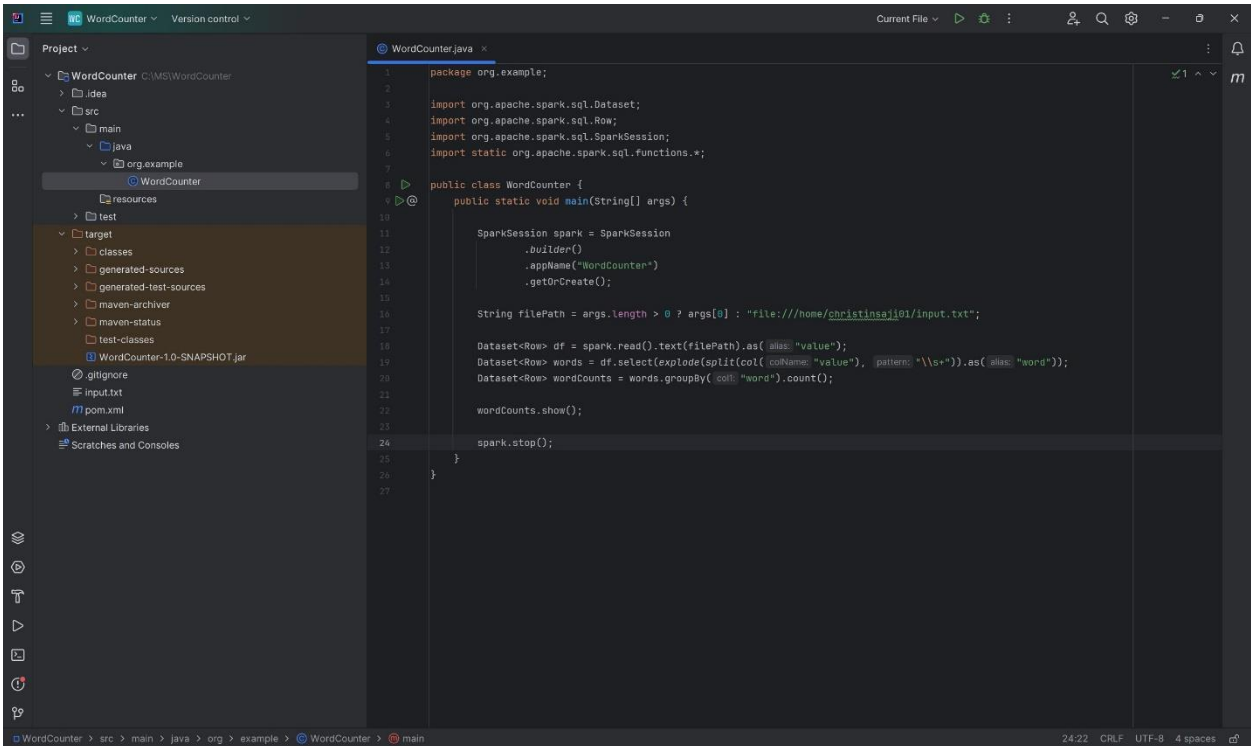
ssh.cloud.google.com/v2/ssh/projects/lab5-assignment-416017/zones/us-central1-c/instances/wordcounter-cluster-m?authuser=0&hl=en_US&projectNumber=1015719484143&useAdminProxy=true - Googl...
ssh.cloud.google.com/v2/ssh/projects/lab5-assignment-416017/zones/us-central1-c/instances/wordcounter-cluster-m?authuser=0&hl=en_US&projectNumber=1015719484143&useAdminPr...
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

```
WordCounter-1.0-SNAPSHOT.jar  input.txt
christinsaji01@wordcounter-cluster-m:~$ spark-submit --class org.example.WordCounter WordCounter-1.0-SNAPSHOT.jar  file:///home/christinsaji01/input.txt
24/03/02 18:38:42 INFO SparkEnv: Registering MapOutputTracker
24/03/02 18:38:42 INFO SparkEnv: Registering BlockManagerMaster
24/03/02 18:38:42 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/03/02 18:38:42 INFO SparkEnv: Registering OutputCommitCoordinator
24/03/02 18:38:43 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at wordcounter-cluster-m.us-central1-c.c.lab5-assignment-416017.intern
al./10.128.0.2:8032
24/03/02 18:38:43 INFO AHSProxy: Connecting to Application History server at wordcounter-cluster-m.us-central1-c.c.lab5-assignment-416017.internal./10.128.0.2:
10200
24/03/02 18:38:45 INFO Configuration: resource-types.xml not found
24/03/02 18:38:45 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/03/02 18:38:45 INFO YarnClientImpl: Submitted application application_1709399890884_0010
24/03/02 18:38:46 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at wordcounter-cluster-m.us-central1-c.c.lab5-assignment-416017.intern
al./10.128.0.2:8030
24/03/02 18:38:49 INFO GhfsStorageStatistics: Detected potential high latency for operation op_get_file_status. latencyMs=612; previousMaxLatencyMs=0; operatio
nCount=1; context=gs://dataproc-temp-us-central1-1015719484143-8wm3b4sq/422e38be-bea2-43dd-b893-b3d23e74ab4c/spark-job-history
24/03/02 18:38:54 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
24/03/02 18:38:54 INFO GhfsStorageStatistics: Detected potential high latency for operation op_mkdirs. latencyMs=4445; previousMaxLatencyMs=0; operationCount=1
; context=gs://dataproc-temp-us-central1-1015719484143-8wm3b4sq/422e38be-bea2-43dd-b893-b3d23e74ab4c/spark-job-history
24/03/02 18:38:54 INFO GhfsStorageStatistics: Detected potential high latency for operation op_create. latencyMs=490; previousMaxLatencyMs=0; operationCount=1;
 context=gs://dataproc-temp-us-central1-1015719484143-8wm3b4sq/422e38be-bea2-43dd-b893-b3d23e74ab4c/spark-job-history/application_1709399890884_0010.inprogress
+-----+-----+
| word|count|
+-----+-----+
| lazy|    1|
|jumps|    1|
|  dog|    1|
|  fox|    1|
|  the|    1|
|  The|    1|
|brown|    1|
| over|    1|
|quick|    1|
+-----+-----+

24/03/02 18:39:17 INFO GhfsStorageStatistics: Detected potential high latency for operation op_rename. latencyMs=508; previousMaxLatencyMs=0; operationCount=1;
 context=rename(gs://dataproc-temp-us-central1-1015719484143-8wm3b4sq/422e38be-bea2-43dd-b893-b3d23e74ab4c/spark-job-history/application_1709399890884_0010.inp
rogress -> gs://dataproc-temp-us-central1-1015719484143-8wm3b4sq/422e38be-bea2-43dd-b893-b3d23e74ab4c/spark-job-history/application_1709399890884_0010)
christinsaji01@wordcounter-cluster-m:~$ 
```

Figure 11 Output of the WordCounter program

Task 2: Explanation of the Java Spark program with the screenshots of the code.



```
package org.example;

import org.apache.spark.sql.Dataset;
import org.apache.spark.sql.Row;
import org.apache.spark.sql.SparkSession;
import static org.apache.spark.sql.functions.*;

public class WordCounter {
    public static void main(String[] args) {

        SparkSession spark = SparkSession
                .builder()
                .appName("WordCounter")
                .getOrCreate();

        String filePath = args.length > 0 ? args[0] : "file:///home/christinsaji01/input.txt";

        Dataset<Row> df = spark.read().text(filePath).as( alias: "value");
        Dataset<Row> words = df.select(explode(split(col( colName: "value"),  pattern: "\\s+")).as( alias: "word"));
        Dataset<Row> wordCounts = words.groupBy( col1: "word").count();

        wordCounts.show();

        spark.stop();
    }
}
```

Figure 12 Java Spark program to count words

Step 1: First, I created an instance for “SparkSession” using the Dataset and DataFrame API.

Step 2: I created a filePath to fetch the file path from the command-line argument and used it; otherwise, I set a default path.

Step 3: The input text was read into the DataFrame, with each row containing a single column called “value.”

Step 4: I split the single line using regex by whitespaces and used the “explode” function to separate each word into a separate row with a column named “words.”

Step 5: I grouped the resulting DataFrame by the word and counted the occurrence of that word.

Step 6: I displayed the result using the “show” method.

Step 7: Finally, I used “stop” to terminate the Spark session.

References

[1] Naveen (NNK), "Spark Read Text File: RDD: DataFrame," *Spark By {Examples}*, [Online], Feb 8, 2023. Available: https://sparkbyexamples.com/spark/spark-read-text-file-rdd-dataframe/ [Accessed: March 2, 2024].

[2] Singh, Chandan, "An Introduction to Apache Spark with Java," *Stack Abuse*, [Online] Aug 3, 2023. Available: https://stackabuse.com/an-introduction-to-apache-spark-with-java/ [Accessed: March 2, 2024].