MapReduce Job for Data Processing on Google Cloud Platform

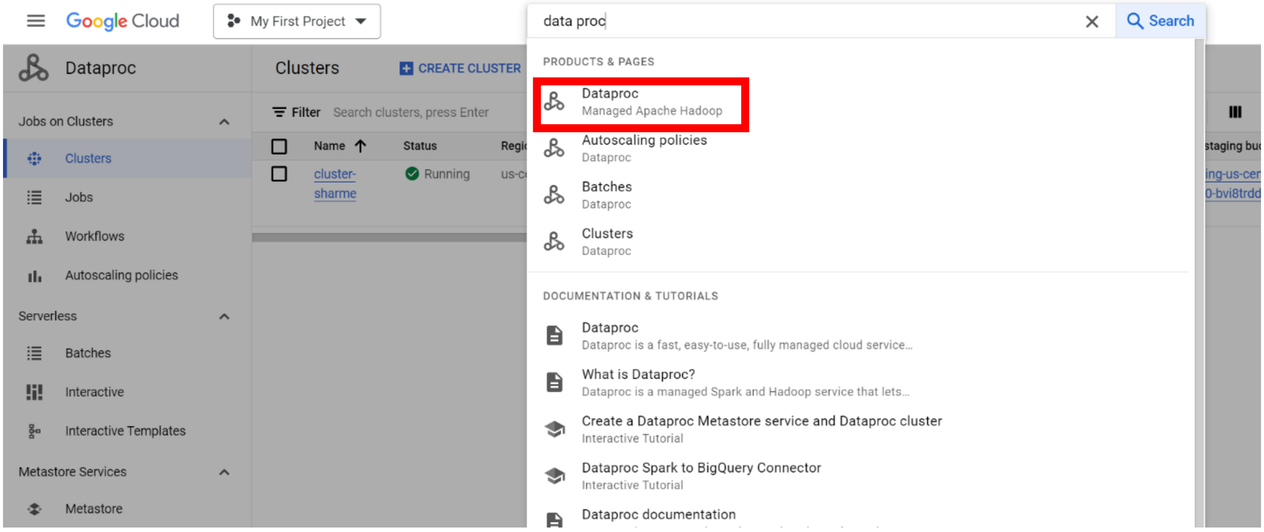
SHARMELE SOMU

Contents

CREATION OF CLUSTERS

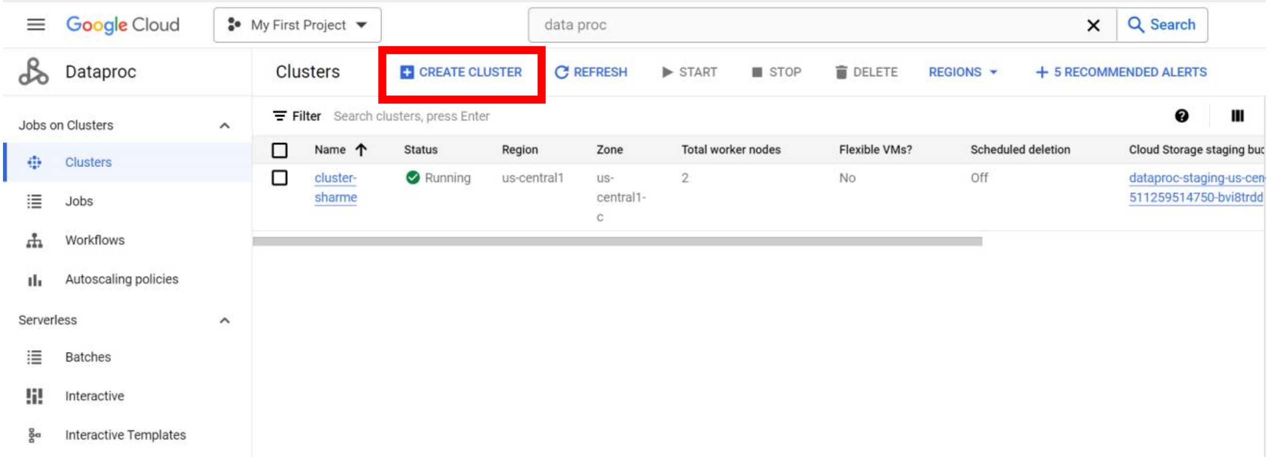
STEP 1:

To create a dataproc cluster, open Google cloud console, type ‘dataproc’ on the search bar and select Dataproc



STEP 2:

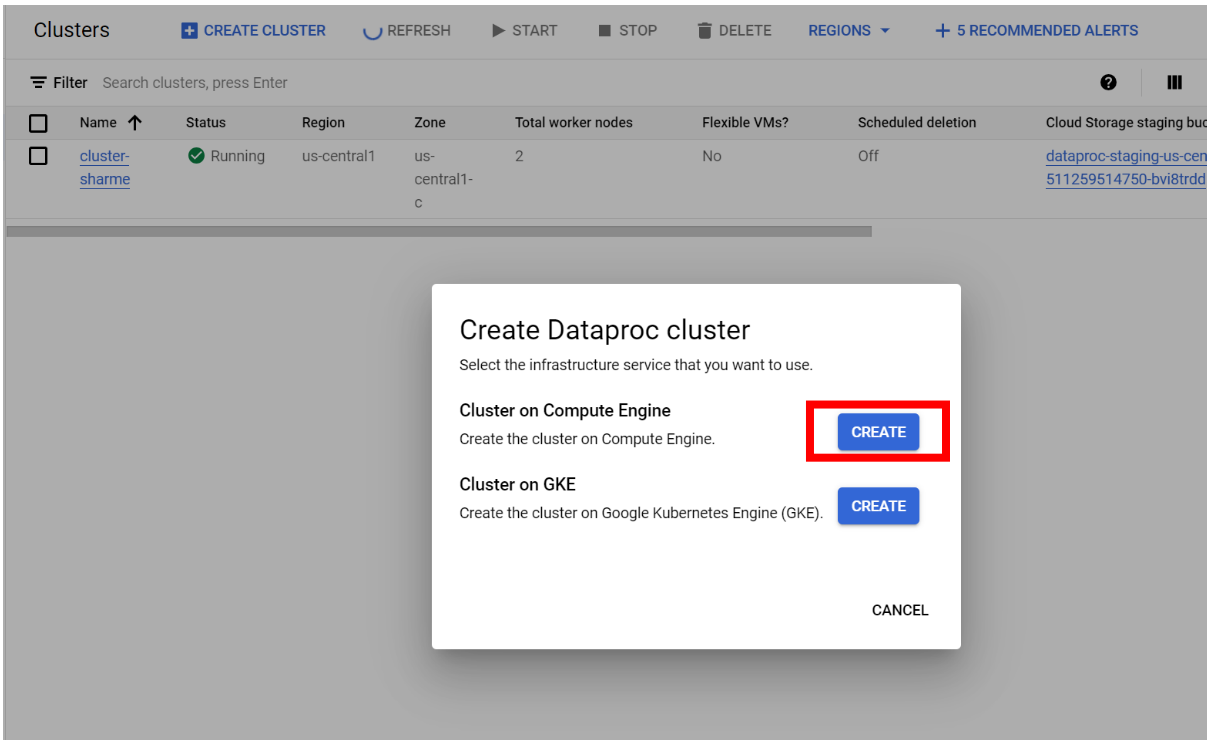
Click on create cluster to create the cluster.



☐	Name	Status	Region	Zone	Total worker nodes	Flexible VMs?	Scheduled deletion	Cloud Storage staging buc
☐	cluster-sharme	✓ Running	us-central1	us-central1-c	2	No	Off	dataproc-staging-us-cen 511259514750-bvi8trdd

STEP 3:

Click on create to create the cluster on compute engine.



Name	Status	Region	Zone	Total worker nodes	Flexible VMs?	Scheduled deletion	Cloud Storage staging buc
cluster-sharme	Running	us-central1	us-central1-c	2	No	Off	dataproc-staging-us-cen 511259514750-bvi8trdd

Create Dataproc cluster

Select the infrastructure service that you want to use.

Cluster on Compute Engine

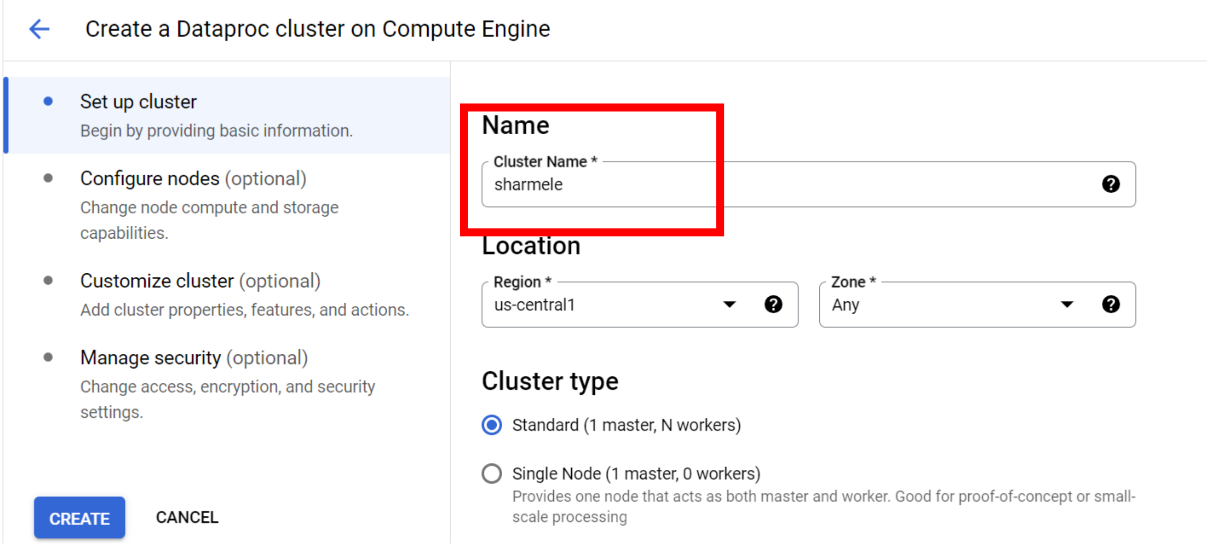
Create the cluster on Compute Engine.

Cluster on GKE

Create the cluster on Google Kubernetes Engine (GKE).

STEP 4:

Next step is to set up the cluster. Mention the name of the cluster.


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 * sharmele

Location

Region * us-central1

Zone * Any

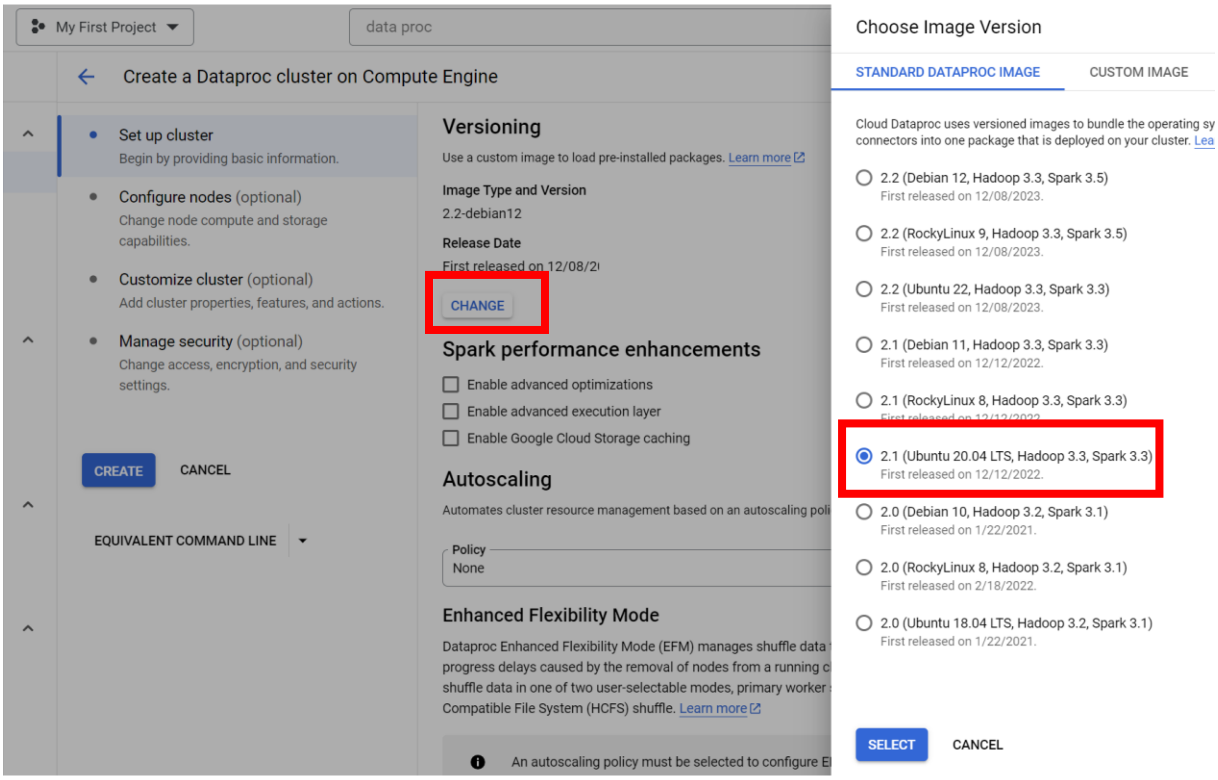
Cluster type

● 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

STEP 5:

Select the name of operating system and version of operating system, Hadoop and spark.



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.

Versioning

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

Image Type and Version
2.2-debian12

Release Date
First released on 12/08/2

Spark performance enhancements

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

Autoscaling

Automates cluster resource management based on an autoscaling poli

Policy
None

Enhanced Flexibility Mode

Dataproc Enhanced Flexibility Mode (EFM) manages shuffle data progress delays caused by the removal of nodes from a running c shuffle data in one of two user-selectable modes, primary worker Compatible File System (HCFS) shuffle. Learn more

An autoscaling policy must be selected to configure E

Choose Image Version

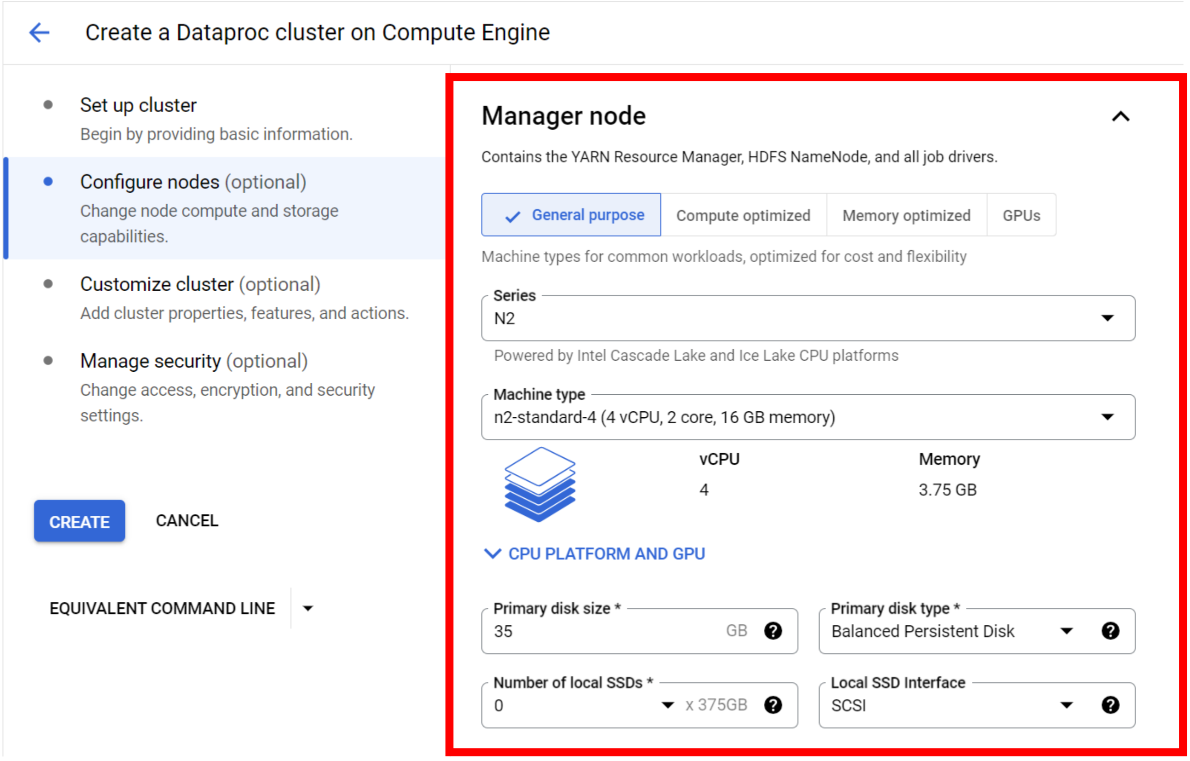
STANDARD DATAPROC IMAGE | CUSTOM IMAGE

Cloud Dataproc uses versioned images to bundle the operating sy connectors into one package that is deployed on your cluster. Lea

○ 2.2 (Debian 12, Hadoop 3.3, Spark 3.5)
First released on 12/08/2023.
○ 2.2 (RockyLinux 9, Hadoop 3.3, Spark 3.5)
First released on 12/08/2023.
○ 2.2 (Ubuntu 22, Hadoop 3.3, Spark 3.3)
First released on 12/08/2023.
○ 2.1 (Debian 11, Hadoop 3.3, Spark 3.3)
First released on 12/12/2022.
○ 2.1 (RockyLinux 8, Hadoop 3.3, Spark 3.3)
First released on 12/12/2022.
● 2.1 (Ubuntu 20.04 LTS, Hadoop 3.3, Spark 3.3)
First released on 12/12/2022.
○ 2.0 (Debian 10, Hadoop 3.2, Spark 3.1)
First released on 1/22/2021.
○ 2.0 (RockyLinux 8, Hadoop 3.2, Spark 3.1)
First released on 2/18/2022.
○ 2.0 (Ubuntu 18.04 LTS, Hadoop 3.2, Spark 3.1)
First released on 1/22/2021.

STEP 6:

In configure nodes, setup the manager and worker nodes. Select the series type, machine type and disk size for the manager node.



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.

✓ General purpose | Compute optimized | Memory optimized | GPUs

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-4 (4 vCPU, 2 core, 16 GB memory)

vCPU	Memory
4	3.75 GB

Primary disk size *
35 GB

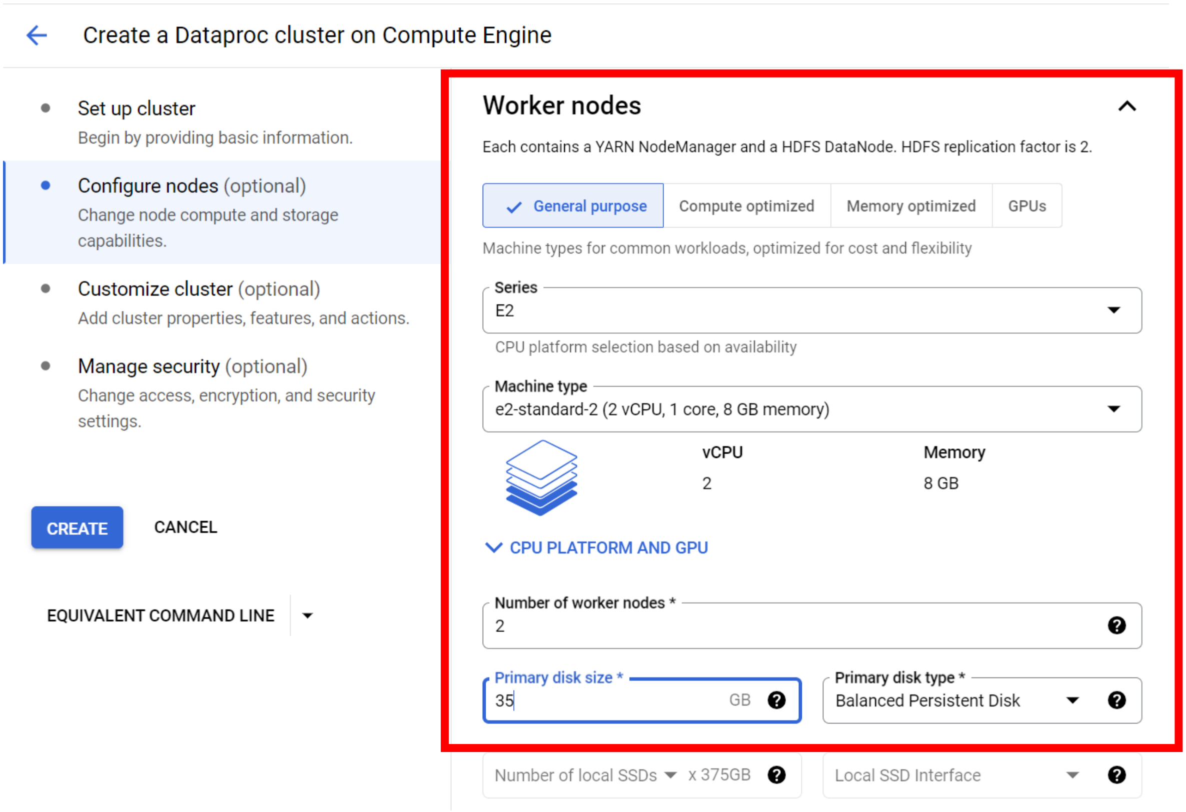
Primary disk type *
Balanced Persistent Disk

Number of local SSDs *
0 x 375GB

Local SSD Interface
SCSI

STEP 7:

Select the series type, machine type and disk size for the worker node and click on create to create the cluster with 1 manager/name node and 2 worker nodes.



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.

CREATE CANCEL

EQUIVALENT COMMAND LINE

Worker nodes

Each contains a YARN NodeManager and a HDFS DataNode. HDFS replication factor is 2.

✓ General purpose | Compute optimized | Memory optimized | GPUs

Machine types for common workloads, optimized for cost and flexibility

Series: E2

CPU platform selection based on availability

Machine type: e2-standard-2 (2 vCPU, 1 core, 8 GB memory)

vCPU	Memory
2	8 GB

CPU PLATFORM AND GPU

Number of worker nodes *: 2

Primary disk size *: 35 GB

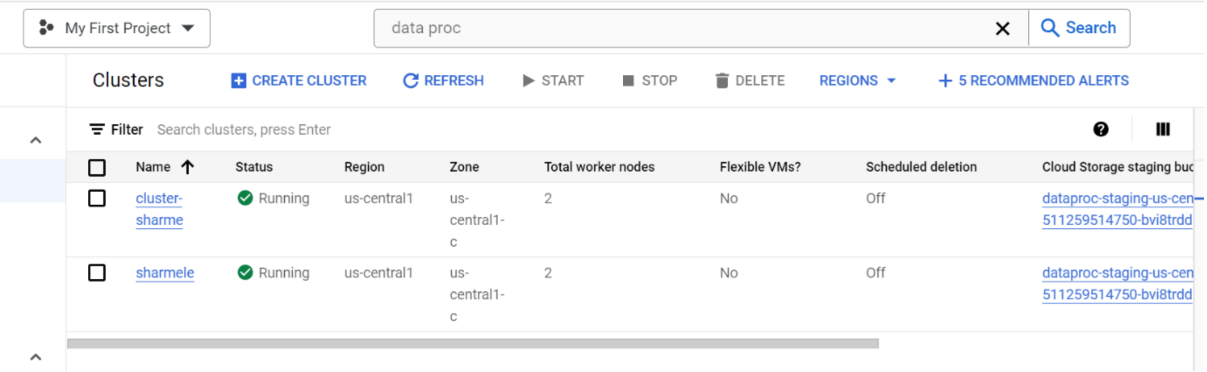
Primary disk type *: Balanced Persistent Disk

Number of local SSDs x 375GB

Local SSD Interface

STEP 8:

Click on the cluster to view the VM instances.



My First Project

data proc

Search

Clusters

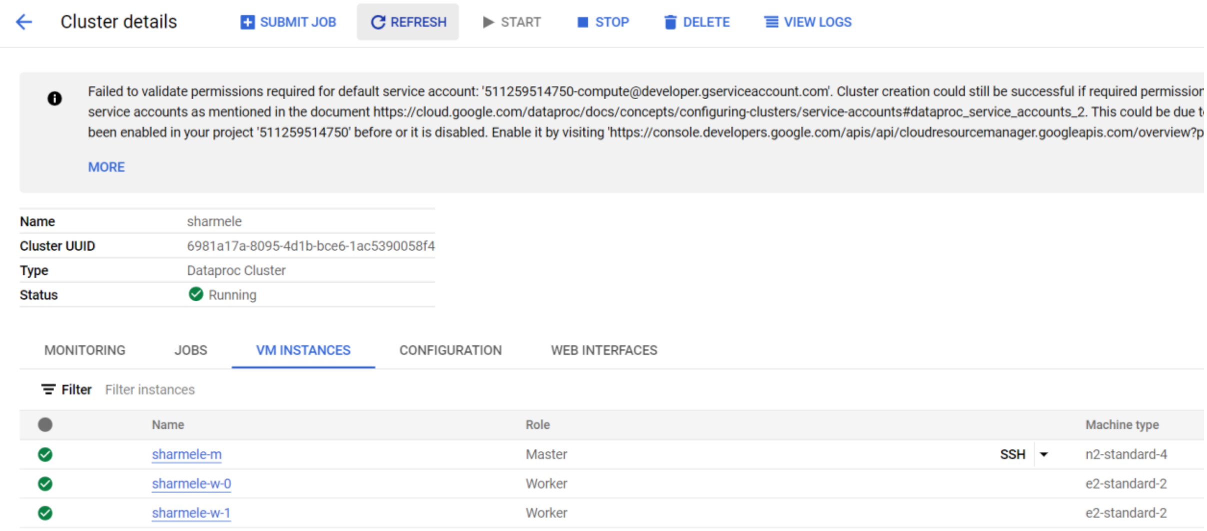
CREATE CLUSTER REFRESH START STOP DELETE REGIONS 5 RECOMMENDED ALERTS

Filter Search clusters, press Enter

□	Name ↑	Status	Region	Zone	Total worker nodes	Flexible VMs?	Scheduled deletion	Cloud Storage staging buc
□	cluster-sharme	✓ Running	us-central1	us-central1-c	2	No	Off	dataproc-staging-us-cen-511259514750-bvi8trdd
□	sharmele	✓ Running	us-central1	us-central1-c	2	No	Off	dataproc-staging-us-cen 511259514750-bvi8trdd

STEP 9:

Click on SSH to connect to the manager node. It opens the terminal window directly on the browser allowing users to interact with the VM's command line interface.

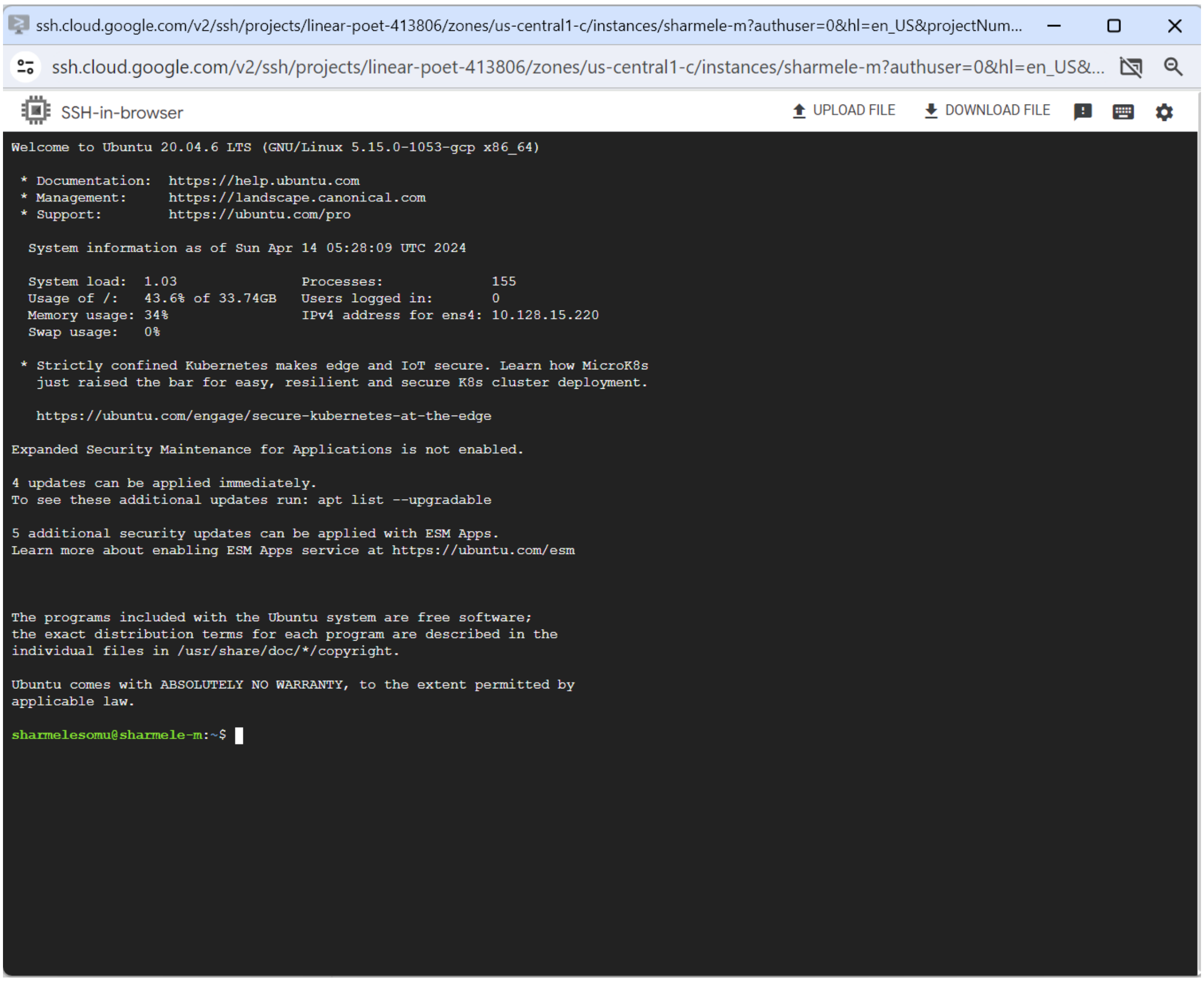


Failed to validate permissions required for default service account: '511259514750-compute@developer.gserviceaccount.com'. Cluster creation could still be successful if required permission service accounts as mentioned in the document https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/service-accounts#dataproc_service_accounts_2. This could be due t been enabled in your project '511259514750' before or it is disabled. Enable it by visiting 'https://console.developers.google.com/apis/api/cloudresourcemanager.googleapis.com/overview?p

MORE

Name	sharmele
Cluster UUID	6981a17a-8095-4d1b-bce6-1ac5390058f4
Type	Dataproc Cluster
Status	✓ Running

●	Name	Role		Machine type
✓	sharmele-m	Master	SSH ▾	n2-standard-4
✓	sharmele-w-0	Worker		e2-standard-2
✓	sharmele-w-1	Worker		e2-standard-2


ssh.cloud.google.com/v2/ssh/projects/linear-poet-413806/zones/us-central1-c/instances/sharmele-m?authuser=0&hl=en_US&projectNum...
ssh.cloud.google.com/v2/ssh/projects/linear-poet-413806/zones/us-central1-c/instances/sharmele-m?authuser=0&hl=en_US&...
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

```
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-1053-gcp x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/pro

  System information as of Sun Apr 14 05:28:09 UTC 2024

  System load:  1.03               Processes:             155
  Usage of /:   43.6% of 33.74GB   Users logged in:       0
  Memory usage: 34%                IPv4 address for ens4: 10.128.15.220
  Swap usage:   0%

 * Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
   just raised the bar for easy, resilient and secure K8s cluster deployment.

   https://ubuntu.com/engage/secure-kubernetes-at-the-edge

Expanded Security Maintenance for Applications is not enabled.

4 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

5 additional security updates can be applied with ESM Apps.
Learn more about enabling ESM Apps service at https://ubuntu.com/esm


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

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

sharmelesomu@sharmele-m:~$
```

TASK A: EXECUTING THE MAPPER AND REDUCER CODE ON THE NAME NODE IN GCP

The objective is to join two tables TA.csv and TB.csv and display the student records for those students whose date of birth is greater than ‘01/01/1997’.

MAPPER EXPLANATION

The mapper code reads the input from the standard input, extracts fields by using ‘ , ’ as the delimiter, checks if the fields are from table A or table B. If the fields are from table A, the records are assigned value 1. If the fields are from table B the records are assigned value 2. Then the output of the mapper is displayed as a key-value pair. Joining of tables is done at the reducer side.

REFERENCES

I have used the following references and ChatGPT to generate the code for the mapper and reducer.

https://www.youtube.com/watch?v=ai0E4ovoA5k&t=151s
https://www.edureka.co/blog/mapreduce-example-reduce-side-join/
https://ars.els-cdn.com/content/image/1-s2.0-S1319157820303694-gr5_lrg.jpg

MAPPER CODE – mapper.py

```
#!/usr/bin/env python
import sys
value=0
# reads input from standard input
for line in sys.stdin:
# extracts fields by using ‘,’ as a delimiter
  fields = line.strip().split(',')
# Assigns fields to student_id, name and dob
  student_id = fields[0]
  name = fields[1]
  dob = fields[2]
# In both tables, when StudentId is encounter, the header is skipped
  if fields[0] == 'StudentId':
    continue
# Assigning value 1 for records from table A to generate a composite key
  elif fields[1] in ['Alice','Tom','John']:
    value = 1
# Assigning value 2 for records from table B to generate a composite key
  else:
    value = 2
# print the mapper output - key and values
  print(student_id,value,[name,dob])
```

REDUCER EXPLANATION

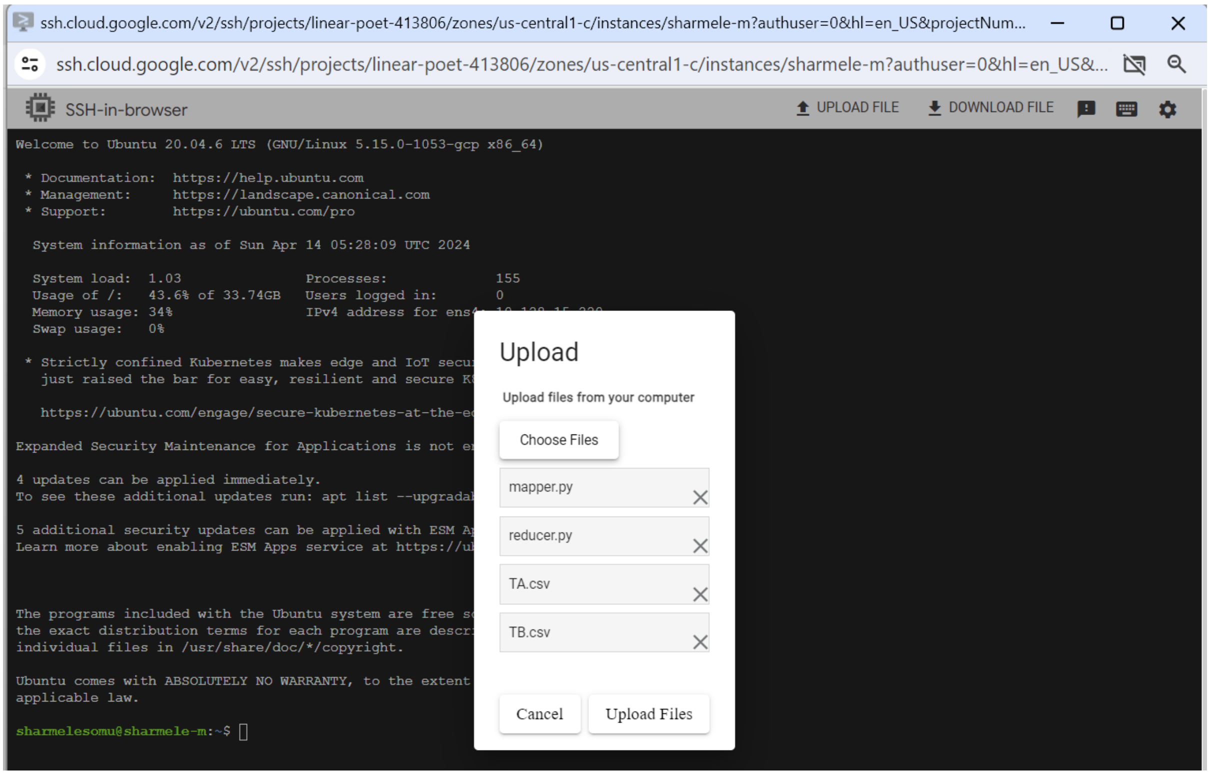
The reducer code reads the data from standard input, splits the data with the table number. If the table number is ‘1’, the code converts the date from string to date time format and stores the data in student_info_tableA. If the table number is ‘2’, the code appends the course details to student_id. Then it checks if student_id from student_info_tableA is present in student_info_tableB and joines the tables. There is a counter logic which counts the number of occurrences of each student_id. Then the student records are filtered to check if date of birth is than ‘01/01/1997’ and the resulting records are printed.

REDUCER CODE – reducer1.py

```
import sys
from datetime import datetime
# Dictionary to store information for each student ID
student_info_tableA = {}
student_info_tableB = {}
prev_student_id = None
counter = 0
x = 0
student_records = {}
# Read input lines from standard input
for line in sys.stdin:
  # Split the line into its components: student ID, table number, and the rest of the data
  student_id, table_number, rest = line.strip().split(' ', 2)
  # If the table number is '1', update student_info_tableA with the student's name and DOB
  if table_number == '1':
    name, dob_str = eval(rest)
    dob = datetime.strptime(dob_str, '%m/%d/%Y').date()
    student_info_tableA.setdefault(student_id, {'name': name, 'dob': dob})
  # If the table number is '2', update student_info_tableB with the student's course data
  elif table_number == '2':
    course_id, grade = eval(rest)
    student_info_tableB.setdefault(student_id, []).append((course_id, grade))
# Perform the join operation and filter records
for student_id in student_info_tableA:
  # Check if the student is present in both tables
  if student_id in student_info_tableB:
    # Combine the information from both tables
    for course_data in student_info_tableB[student_id]:
      name = student_info_tableA[student_id]['name']
      dob = student_info_tableA[student_id]['dob']
        # counter value to tag each student record
      if prev_student_id == student_id:
        counter = counter
        x=counter
      else:
        counter += 1
        prev_student_id = student_id
      # filters those student records for which date is above ‘01/01/1995’
      if dob > datetime(1995, 1, 1).date():  # Filter based on date of birth
        student_records.setdefault(student_id, []).append([name, dob_str, course_data[0],
course_data[1]])
    for student_id, records in student_records.items():
      counter = 0  # Initialize counter for each student ID
      for record in records:
        counter += 1  # Increment the counter for each record
        print(f'{student_id}\t{counter}\t{record}') # print the student records after filtering dob
```

EXECUTING THE MAPPER AND REDUCER ON THE NAME NODE IN GCP

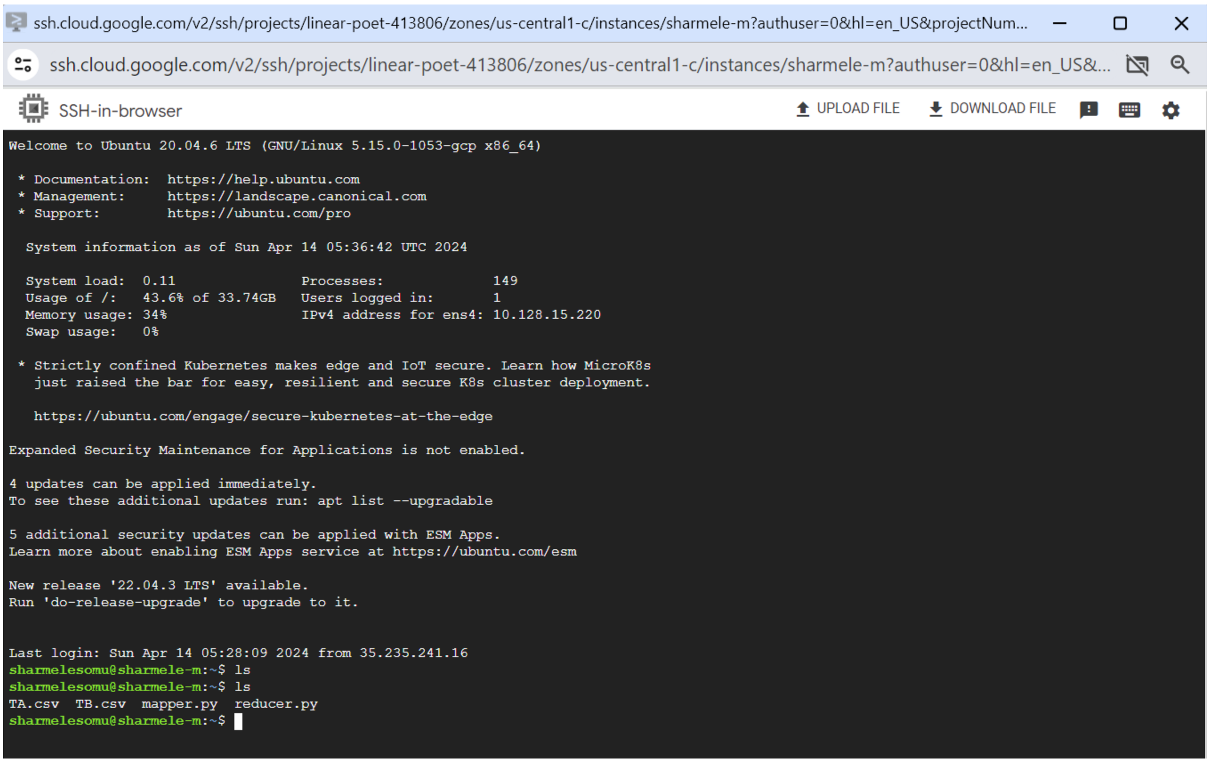
STEP 1: Upload the csv files – TA.csv, TB.csv, python files - mapper and reducer on the name node.



ssh.cloud.google.com/v2/ssh/projects/linear-poet-413806/zones/us-central1-c/instances/sharmele-m?authuser=0&hl=en_US&projectNum...
ssh.cloud.google.com/v2/ssh/projects/linear-poet-413806/zones/us-central1-c/instances/sharmele-m?authuser=0&hl=en_US&...
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
Upload
Upload files from your computer
Choose Files
mapper.py
reducer.py
TA.csv
TB.csv
Cancel
Upload Files

```
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-1053-gcp x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/pro

  System information as of Sun Apr 14 05:28:09 UTC 2024

  System load:  1.03              Processes:             155
  Usage of /:   43.6% of 33.74GB   Users logged in:       0
  Memory usage: 34%               IPv4 address for ens4
  Swap usage:   0%

 * Strictly confined Kubernetes makes edge and IoT secu
   just raised the bar for easy, resilient and secure K

   https://ubuntu.com/engage/secure-kubernetes-at-the-e

Expanded Security Maintenance for Applications is not e

4 updates can be applied immediately.
To see these additional updates run: apt list --upgrada

5 additional security updates can be applied with ESM A
Learn more about enabling ESM Apps service at https://u


The programs included with the Ubuntu system are free s
the exact distribution terms for each program are descr
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent
applicable law.

sharmelesomu@sharmele-m:~$
```

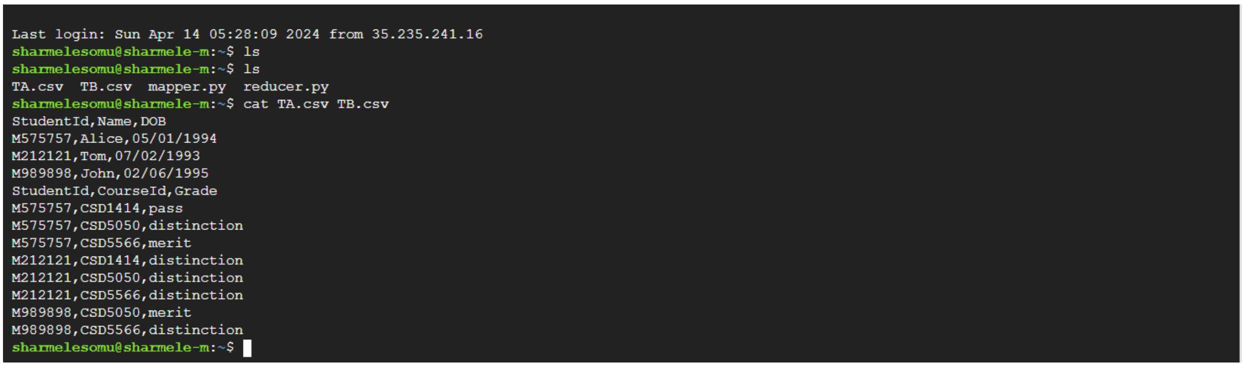
STEP 2: use the command ‘ls’ to check the list of loaded files.



ssh.cloud.google.com/v2/ssh/projects/linear-poet-413806/zones/us-central1-c/instances/sharmele-m?authuser=0&hl=en_US&projectNum...
ssh.cloud.google.com/v2/ssh/projects/linear-poet-413806/zones/us-central1-c/instances/sharmele-m?authuser=0&hl=en_US&...
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

```
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-1053-gcp x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/pro

  System information as of Sun Apr 14 05:36:42 UTC 2024

  System load:  0.11               Processes:             149
  Usage of /:   43.6% of 33.74GB   Users logged in:       1
  Memory usage: 34%                IPv4 address for ens4: 10.128.15.220
  Swap usage:   0%

 * Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
   just raised the bar for easy, resilient and secure K8s cluster deployment.

   https://ubuntu.com/engage/secure-kubernetes-at-the-edge

Expanded Security Maintenance for Applications is not enabled.

4 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

5 additional security updates can be applied with ESM Apps.
Learn more about enabling ESM Apps service at https://ubuntu.com/esm

New release '22.04.3 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Sun Apr 14 05:28:09 2024 from 35.235.241.16
sharmelesomu@sharmele-m:~$ ls
sharmelesomu@sharmele-m:~$ ls
TA.csv  TB.csv  mapper.py  reducer.py
sharmelesomu@sharmele-m:~$ 
```

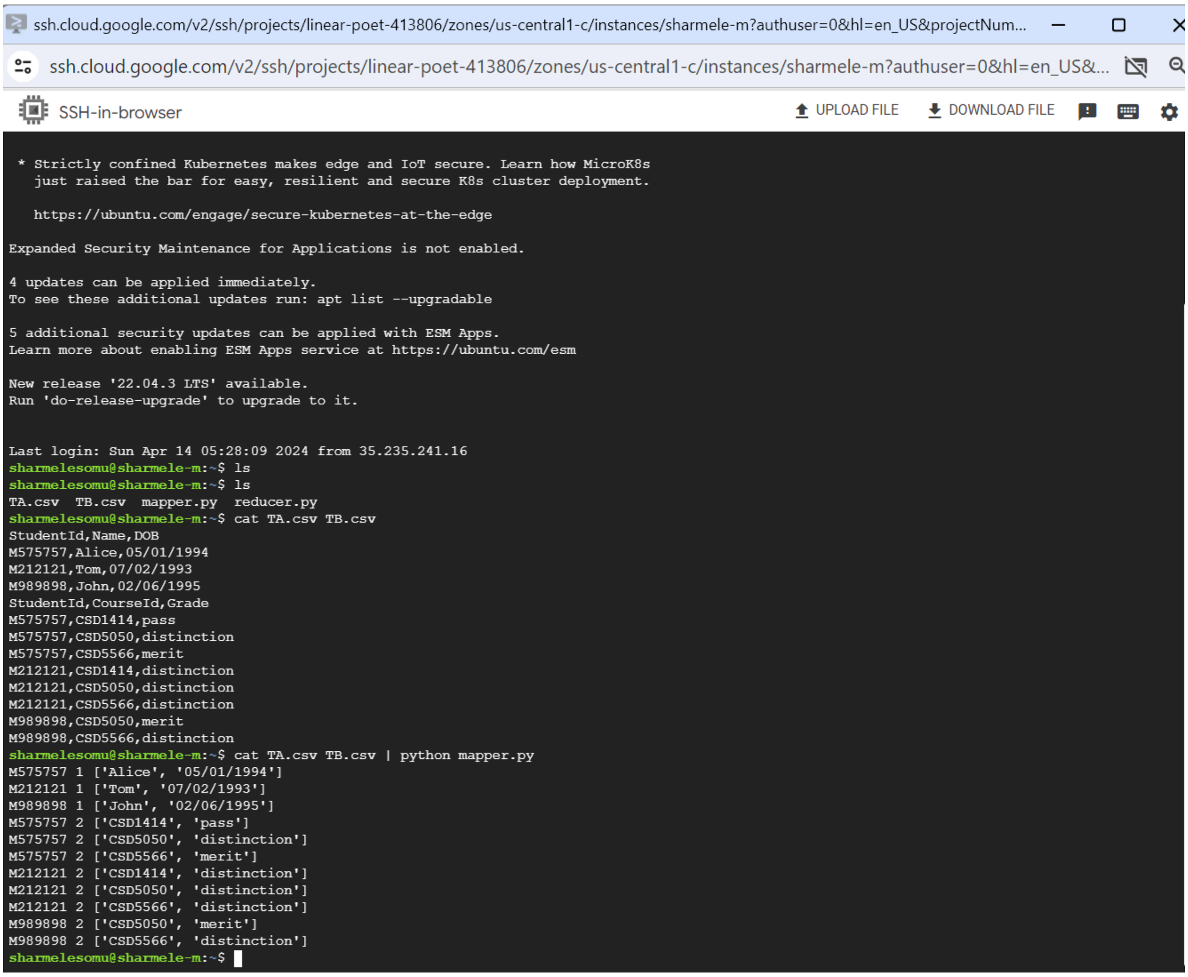
STEP 3: Display the content of Table 1 and Table 2 using **‘cat’** command.



```
Last login: Sun Apr 14 05:28:09 2024 from 35.235.241.16
sharmelesomu@sharmele-m:~$ ls
sharmelesomu@sharmele-m:~$ ls
TA.csv  TB.csv  mapper.py  reducer.py
sharmelesomu@sharmele-m:~$ cat TA.csv TB.csv
StudentId,Name,DOB
M575757,Alice,05/01/1994
M212121,Tom,07/02/1993
M989898,John,02/06/1995
StudentId,CourseId,Grade
M575757,CSD1414,pass
M575757,CSD5050,distinction
M575757,CSD5566,merit
M212121,CSD1414,distinction
M212121,CSD5050,distinction
M212121,CSD5566,distinction
M989898,CSD5050,merit
M989898,CSD5566,distinction
sharmelesomu@sharmele-m:~$ 
```

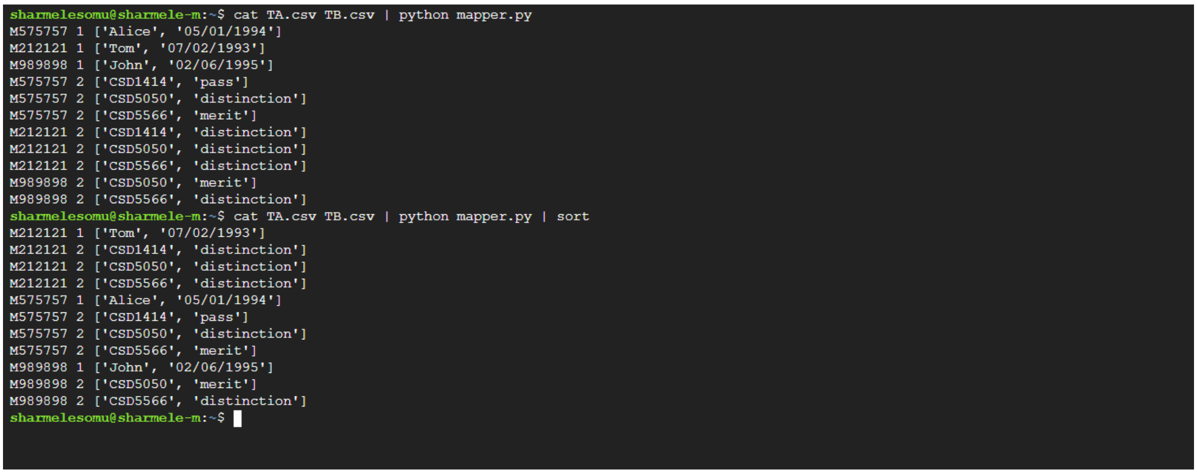
STEP 4: Execute the mapper code – mapper.py by using the command python mapper.py. The output from above cat command is the input to the mapper. We use the '|' symbol to input the output from one command to another command

cat TA.csv TB.csv | python mapper.py


ssh.cloud.google.com/v2/ssh/projects/linear-poet-413806/zones/us-central1-c/instances/sharmele-m?authuser=0&hl=en_US&projectNum...
ssh.cloud.google.com/v2/ssh/projects/linear-poet-413806/zones/us-central1-c/instances/sharmele-m?authuser=0&hl=en_US&...
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

```
 * Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
   just raised the bar for easy, resilient and secure K8s cluster deployment.

   https://ubuntu.com/engage/secure-kubernetes-at-the-edge

Expanded Security Maintenance for Applications is not enabled.

4 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

5 additional security updates can be applied with ESM Apps.
Learn more about enabling ESM Apps service at https://ubuntu.com/esm

New release '22.04.3 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Sun Apr 14 05:28:09 2024 from 35.235.241.16
sharmelesomu@sharmele-m:~$ ls
sharmelesomu@sharmele-m:~$ ls
TA.csv  TB.csv  mapper.py  reducer.py
sharmelesomu@sharmele-m:~$ cat TA.csv TB.csv
StudentId,Name,DOB
M575757,Alice,05/01/1994
M212121,Tom,07/02/1993
M989898,John,02/06/1995
StudentId,CourseId,Grade
M575757,CSD1414,pass
M575757,CSD5050,distinction
M575757,CSD5566,merit
M212121,CSD1414,distinction
M212121,CSD5050,distinction
M212121,CSD5566,distinction
M989898,CSD5050,merit
M989898,CSD5566,distinction
sharmelesomu@sharmele-m:~$ cat TA.csv TB.csv | python mapper.py
M575757 1 ['Alice', '05/01/1994']
M212121 1 ['Tom', '07/02/1993']
M989898 1 ['John', '02/06/1995']
M575757 2 ['CSD1414', 'pass']
M575757 2 ['CSD5050', 'distinction']
M575757 2 ['CSD5566', 'merit']
M212121 2 ['CSD1414', 'distinction']
M212121 2 ['CSD5050', 'distinction']
M212121 2 ['CSD5566', 'distinction']
M989898 2 ['CSD5050', 'merit']
M989898 2 ['CSD5566', 'distinction']
sharmelesomu@sharmele-m:~$ 
```

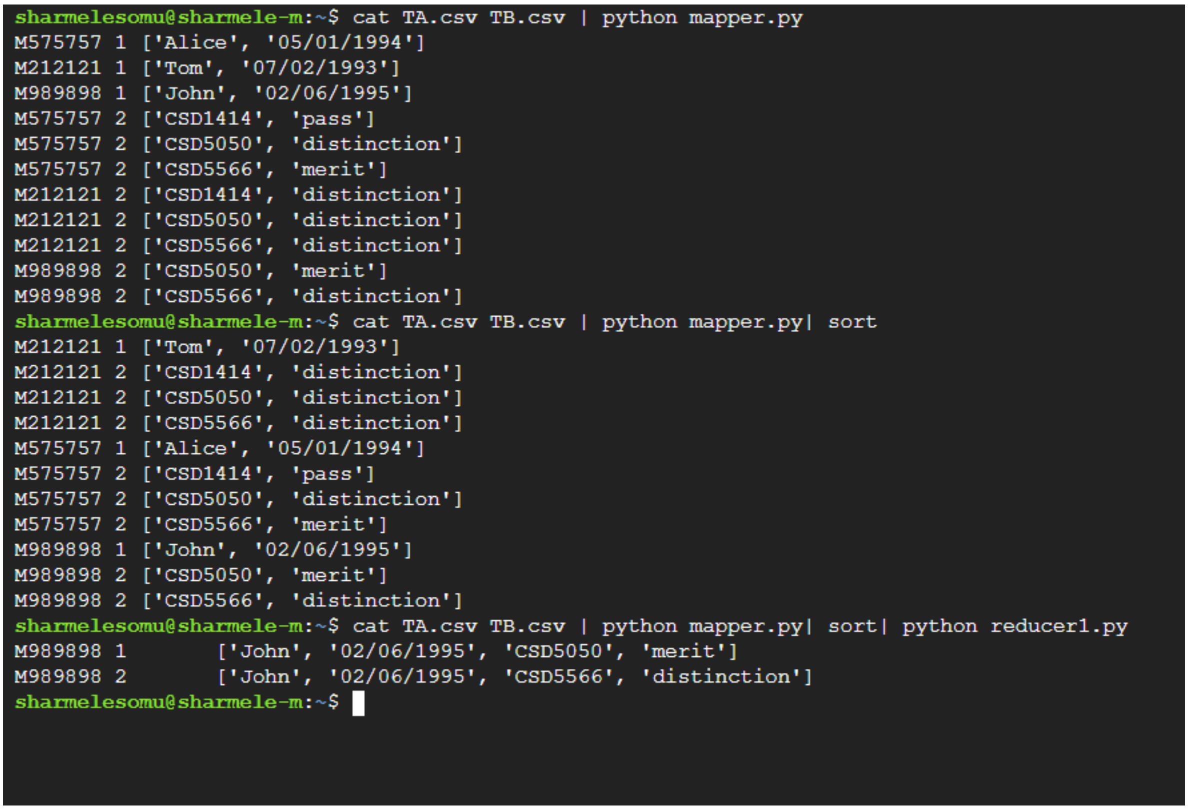
STEP 5: Sort the output from the mapper.



```
sharmelesomu@sharmele-m:~$ cat TA.csv TB.csv | python mapper.py
M575757 1 ['Alice', '05/01/1994']
M212121 1 ['Tom', '07/02/1993']
M989898 1 ['John', '02/06/1995']
M575757 2 ['CSD1414', 'pass']
M575757 2 ['CSD5050', 'distinction']
M575757 2 ['CSD5566', 'merit']
M212121 2 ['CSD1414', 'distinction']
M212121 2 ['CSD5050', 'distinction']
M212121 2 ['CSD5566', 'distinction']
M989898 2 ['CSD5050', 'merit']
M989898 2 ['CSD5566', 'distinction']
sharmelesomu@sharmele-m:~$ cat TA.csv TB.csv | python mapper.py | sort
M212121 1 ['Tom', '07/02/1993']
M212121 2 ['CSD1414', 'distinction']
M212121 2 ['CSD5050', 'distinction']
M212121 2 ['CSD5566', 'distinction']
M575757 1 ['Alice', '05/01/1994']
M575757 2 ['CSD1414', 'pass']
M575757 2 ['CSD5050', 'distinction']
M575757 2 ['CSD5566', 'merit']
M989898 1 ['John', '02/06/1995']
M989898 2 ['CSD5050', 'merit']
M989898 2 ['CSD5566', 'distinction']
sharmelesomu@sharmele-m:~$ 
```

STEP 5: Executing the reducer code – reducer1.py by using the command python reducer.py. The output from the mapper is given as an input to the reducer.

cat TA.csv TB.csv | python mapper.py |sort | python reducer1.py

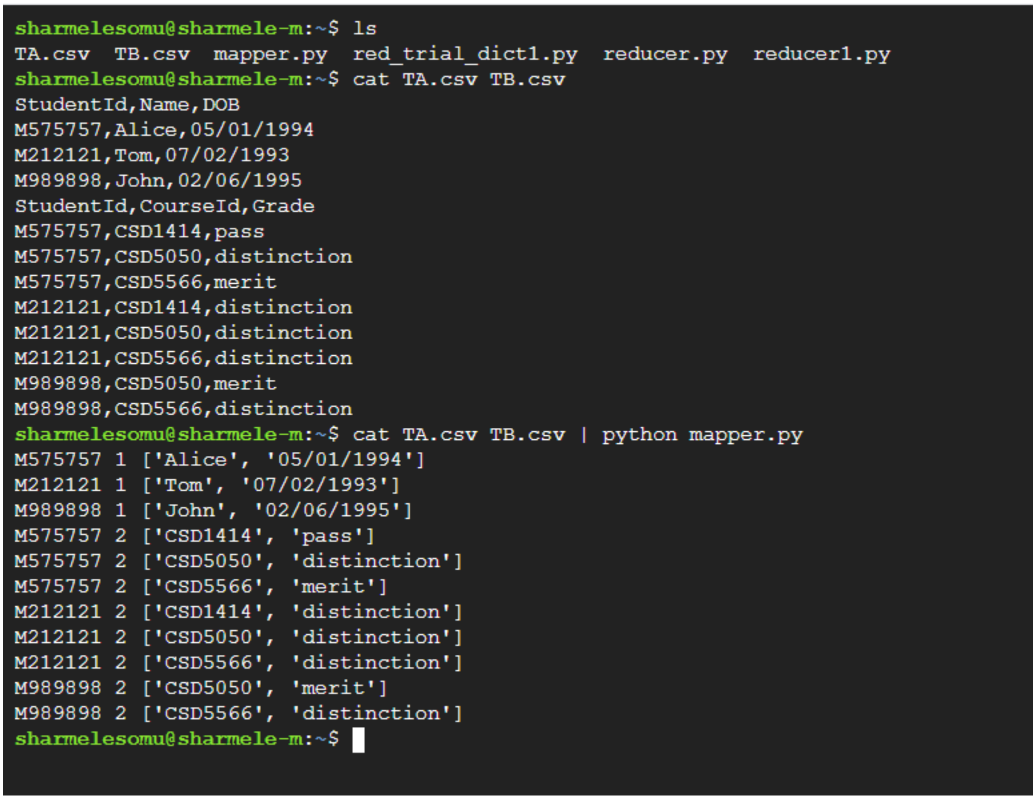


```
sharmelesomu@sharmele-m:~$ cat TA.csv TB.csv | python mapper.py
M575757 1 ['Alice', '05/01/1994']
M212121 1 ['Tom', '07/02/1993']
M989898 1 ['John', '02/06/1995']
M575757 2 ['CSD1414', 'pass']
M575757 2 ['CSD5050', 'distinction']
M575757 2 ['CSD5566', 'merit']
M212121 2 ['CSD1414', 'distinction']
M212121 2 ['CSD5050', 'distinction']
M212121 2 ['CSD5566', 'distinction']
M989898 2 ['CSD5050', 'merit']
M989898 2 ['CSD5566', 'distinction']
sharmelesomu@sharmele-m:~$ cat TA.csv TB.csv | python mapper.py| sort
M212121 1 ['Tom', '07/02/1993']
M212121 2 ['CSD1414', 'distinction']
M212121 2 ['CSD5050', 'distinction']
M212121 2 ['CSD5566', 'distinction']
M575757 1 ['Alice', '05/01/1994']
M575757 2 ['CSD1414', 'pass']
M575757 2 ['CSD5050', 'distinction']
M575757 2 ['CSD5566', 'merit']
M989898 1 ['John', '02/06/1995']
M989898 2 ['CSD5050', 'merit']
M989898 2 ['CSD5566', 'distinction']
sharmelesomu@sharmele-m:~$ cat TA.csv TB.csv | python mapper.py| sort| python reducer1.py
M989898 1       ['John', '02/06/1995', 'CSD5050', 'merit']
M989898 2       ['John', '02/06/1995', 'CSD5566', 'distinction']
sharmelesomu@sharmele-m:~$
```

TASK B: RESULTS OF THE MAPPER AND REDUCER ON NAME NODE

RESULT OF MAPPER

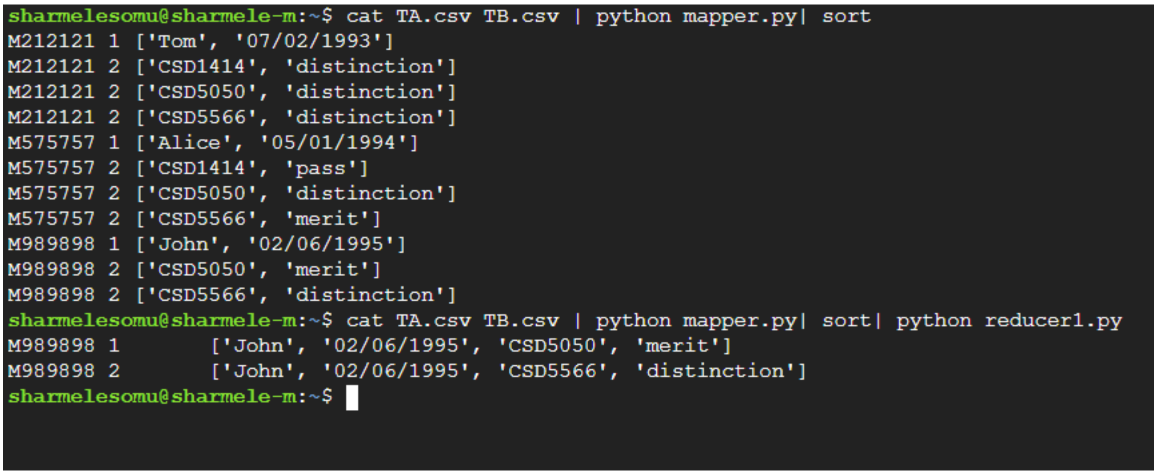
Mapper combines both tables and displays in key-value format



```
sharmelesomu@sharmele-m:~$ ls
TA.csv  TB.csv  mapper.py  red_trial_dict1.py  reducer.py  reducer1.py
sharmelesomu@sharmele-m:~$ cat TA.csv TB.csv
StudentId,Name,DOB
M575757,Alice,05/01/1994
M212121,Tom,07/02/1993
M989898,John,02/06/1995
StudentId,CourseId,Grade
M575757,CSD1414,pass
M575757,CSD5050,distinction
M575757,CSD5566,merit
M212121,CSD1414,distinction
M212121,CSD5050,distinction
M212121,CSD5566,distinction
M989898,CSD5050,merit
M989898,CSD5566,distinction
sharmelesomu@sharmele-m:~$ cat TA.csv TB.csv | python mapper.py
M575757 1 ['Alice', '05/01/1994']
M212121 1 ['Tom', '07/02/1993']
M989898 1 ['John', '02/06/1995']
M575757 2 ['CSD1414', 'pass']
M575757 2 ['CSD5050', 'distinction']
M575757 2 ['CSD5566', 'merit']
M212121 2 ['CSD1414', 'distinction']
M212121 2 ['CSD5050', 'distinction']
M212121 2 ['CSD5566', 'distinction']
M989898 2 ['CSD5050', 'merit']
M989898 2 ['CSD5566', 'distinction']
sharmelesomu@sharmele-m:~$
```

RESULT OF REDUCER

Reducer joins the tables and displays those records for which date of birth is greater than ’01/01/1995’



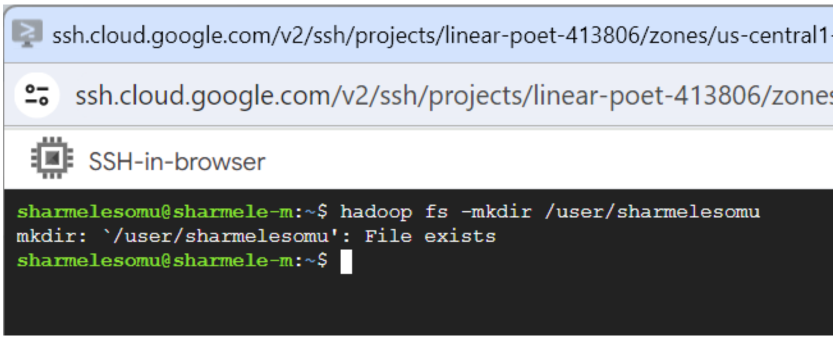
```
sharmelesomu@sharmele-m:~$ cat TA.csv TB.csv | python mapper.py| sort
M212121 1 ['Tom', '07/02/1993']
M212121 2 ['CSD1414', 'distinction']
M212121 2 ['CSD5050', 'distinction']
M212121 2 ['CSD5566', 'distinction']
M575757 1 ['Alice', '05/01/1994']
M575757 2 ['CSD1414', 'pass']
M575757 2 ['CSD5050', 'distinction']
M575757 2 ['CSD5566', 'merit']
M989898 1 ['John', '02/06/1995']
M989898 2 ['CSD5050', 'merit']
M989898 2 ['CSD5566', 'distinction']
sharmelesomu@sharmele-m:~$ cat TA.csv TB.csv | python mapper.py| sort| python reducer1.py
M989898 1       ['John', '02/06/1995', 'CSD5050', 'merit']
M989898 2       ['John', '02/06/1995', 'CSD5566', 'distinction']
sharmelesomu@sharmele-m:~$
```

TASK C: EXECUTING THE MAPPER AND REDUCER ON HADOOP STREAMING SERVICE

PREPARING THE HADOOP ENVIRONMENT

STEP 1: Create default folder in the Hadoop filesystem.

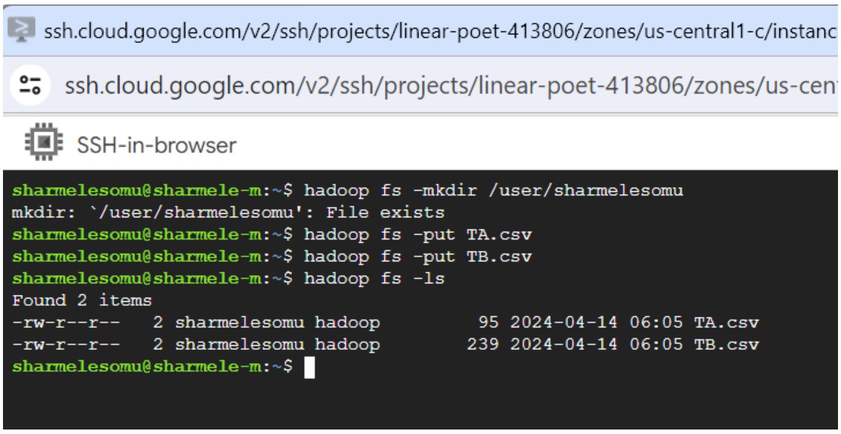
Command: Hadoop fs -mkdir /user/sharmelesomu



ssh.cloud.google.com/v2/ssh/projects/linear-poet-413806/zones/us-central1
ssh.cloud.google.com/v2/ssh/projects/linear-poet-413806/zones
SSH-in-browser

```
sharmelesomu@sharmele-m:~$ hadoop fs -mkdir /user/sharmelesomu
mkdir: `/user/sharmelesomu': File exists
sharmelesomu@sharmele-m:~$
```

STEP 2: Store the input files on the Hadoop default folder



ssh.cloud.google.com/v2/ssh/projects/linear-poet-413806/zones/us-central1-c/instanc
ssh.cloud.google.com/v2/ssh/projects/linear-poet-413806/zones/us-cen
SSH-in-browser

```
sharmelesomu@sharmele-m:~$ hadoop fs -mkdir /user/sharmelesomu
mkdir: `/user/sharmelesomu': File exists
sharmelesomu@sharmele-m:~$ hadoop fs -put TA.csv
sharmelesomu@sharmele-m:~$ hadoop fs -put TB.csv
sharmelesomu@sharmele-m:~$ hadoop fs -ls
Found 2 items
-rw-r--r--   2 sharmelesomu hadoop         95 2024-04-14 06:05 TA.csv
-rw-r--r--   2 sharmelesomu hadoop        239 2024-04-14 06:05 TB.csv
sharmelesomu@sharmele-m:~$
```

STEP 3: Execute the mapper using Hadoop streaming service.

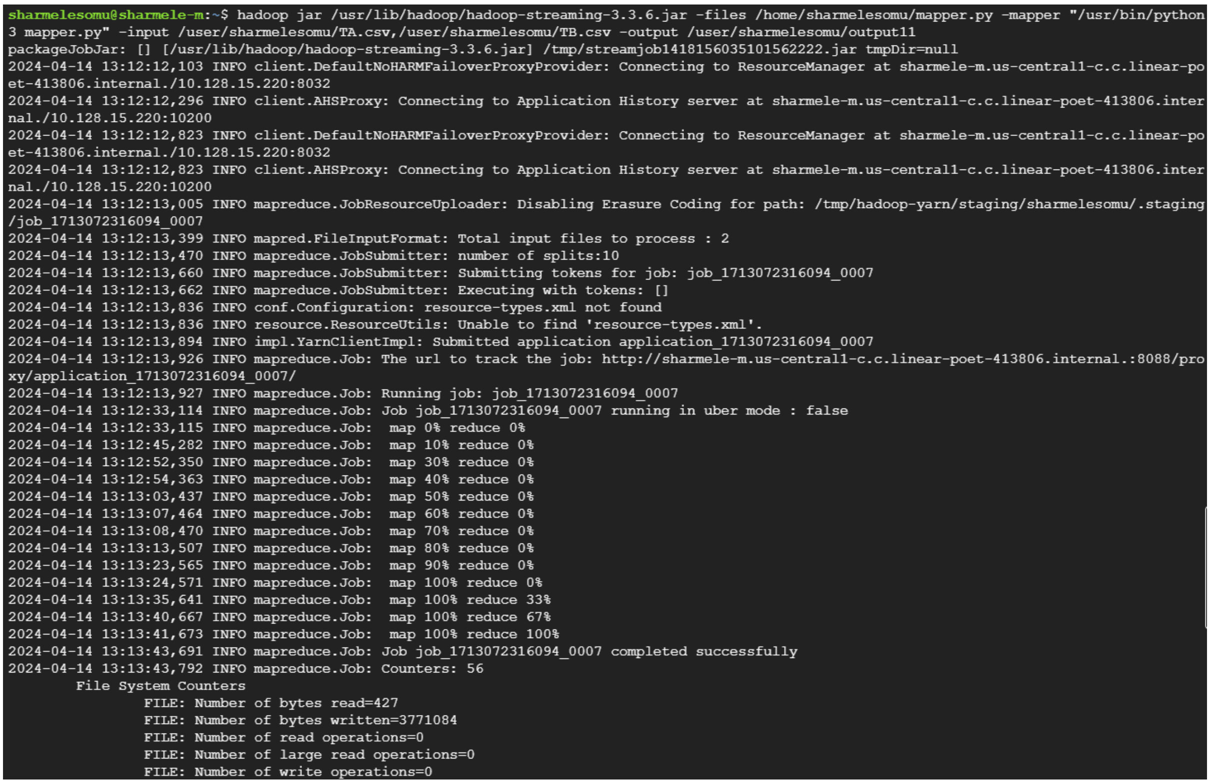
Below command is used to execute the mapper on the Hadoop environment.

**hadoop jar /usr/lib/hadoop/hadoop-streaming-3.3.6.jar **

**-files /home/sharmelesomu/mapper.py -mapper "/usr/bin/python3 mapper.py" **

**-input /user/sharmelesomu/TA.csv,/user/sharmelesomu/TB.csv **

-output /user/sharmelesomu/output11



```
sharmelesomu@sharmele-m:~$ hadoop jar /usr/lib/hadoop/hadoop-streaming-3.3.6.jar -files /home/sharmelesomu/mapper.py -mapper "/usr/bin/python
3 mapper.py" -input /user/sharmelesomu/TA.csv,/user/sharmelesomu/TB.csv -output /user/sharmelesomu/output11
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6.jar] /tmp/streamjob1418156035101562222.jar tmpDir=null
2024-04-14 13:12:12,103 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at sharmele-m.us-central1-c.c.linear-po
et-413806.internal./10.128.15.220:8032
2024-04-14 13:12:12,296 INFO client.AHSProxy: Connecting to Application History server at sharmele-m.us-central1-c.c.linear-poet-413806.inter
nal./10.128.15.220:10200
2024-04-14 13:12:12,823 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at sharmele-m.us-central1-c.c.linear-po
et-413806.internal./10.128.15.220:8032
2024-04-14 13:12:12,823 INFO client.AHSProxy: Connecting to Application History server at sharmele-m.us-central1-c.c.linear-poet-413806.inter
nal./10.128.15.220:10200
2024-04-14 13:12:13,005 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sharmelesomu/.staging
/job_1713072316094_0007
2024-04-14 13:12:13,399 INFO mapred.FileInputFormat: Total input files to process : 2
2024-04-14 13:12:13,470 INFO mapreduce.JobSubmitter: number of splits:10
2024-04-14 13:12:13,660 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1713072316094_0007
2024-04-14 13:12:13,662 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-04-14 13:12:13,836 INFO conf.Configuration: resource-types.xml not found
2024-04-14 13:12:13,836 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-04-14 13:12:13,894 INFO impl.YarnClientImpl: Submitted application application_1713072316094_0007
2024-04-14 13:12:13,926 INFO mapreduce.Job: The url to track the job: http://sharmele-m.us-central1-c.c.linear-poet-413806.internal.:8088/pro
xy/application_1713072316094_0007/
2024-04-14 13:12:13,927 INFO mapreduce.Job: Running job: job_1713072316094_0007
2024-04-14 13:12:33,114 INFO mapreduce.Job: Job job_1713072316094_0007 running in uber mode : false
2024-04-14 13:12:33,115 INFO mapreduce.Job:  map 0% reduce 0%
2024-04-14 13:12:45,282 INFO mapreduce.Job:  map 10% reduce 0%
2024-04-14 13:12:52,350 INFO mapreduce.Job:  map 30% reduce 0%
2024-04-14 13:12:54,363 INFO mapreduce.Job:  map 40% reduce 0%
2024-04-14 13:13:03,437 INFO mapreduce.Job:  map 50% reduce 0%
2024-04-14 13:13:07,464 INFO mapreduce.Job:  map 60% reduce 0%
2024-04-14 13:13:08,470 INFO mapreduce.Job:  map 70% reduce 0%
2024-04-14 13:13:13,507 INFO mapreduce.Job:  map 80% reduce 0%
2024-04-14 13:13:23,565 INFO mapreduce.Job:  map 90% reduce 0%
2024-04-14 13:13:24,571 INFO mapreduce.Job:  map 100% reduce 0%
2024-04-14 13:13:35,641 INFO mapreduce.Job:  map 100% reduce 33%
2024-04-14 13:13:40,667 INFO mapreduce.Job:  map 100% reduce 67%
2024-04-14 13:13:41,673 INFO mapreduce.Job:  map 100% reduce 100%
2024-04-14 13:13:43,691 INFO mapreduce.Job: Job job_1713072316094_0007 completed successfully
2024-04-14 13:13:43,792 INFO mapreduce.Job: Counters: 56
        File System Counters
                FILE: Number of bytes read=427
                FILE: Number of bytes written=3771084
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
```

STEP 4: Execute the reducer using Hadoop streaming service.

Below command is used for executing the reducer in the Hadoop environment

hadoop jar /usr/lib/hadoop/hadoop-streaming-3.3.6.jar

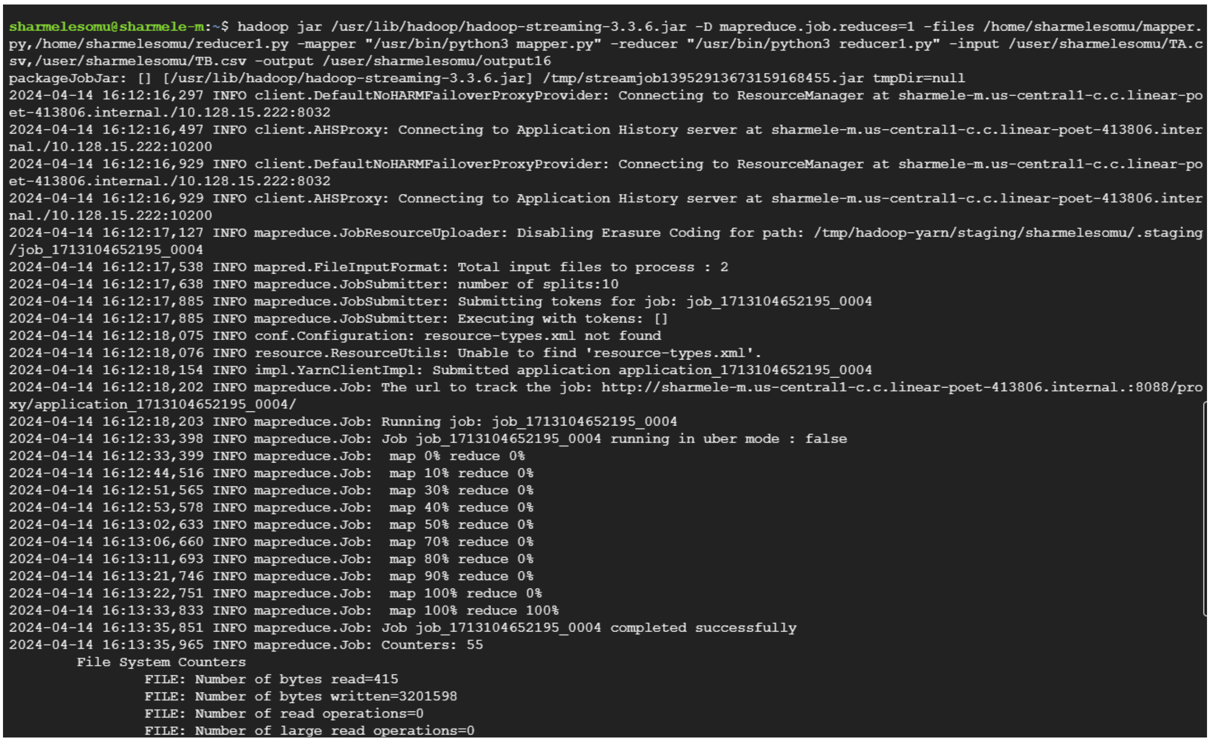
-D mapreduce.job.reduces=1

-files /home/sharmelesomu/mapper.py,/home/sharmelesomu/reducer1.py

-mapper "/usr/bin/python3 mapper.py" -reducer "/usr/bin/python3 reducer1.py"

-input /user/sharmelesomu/TA.csv,/user/sharmelesomu/TB.csv

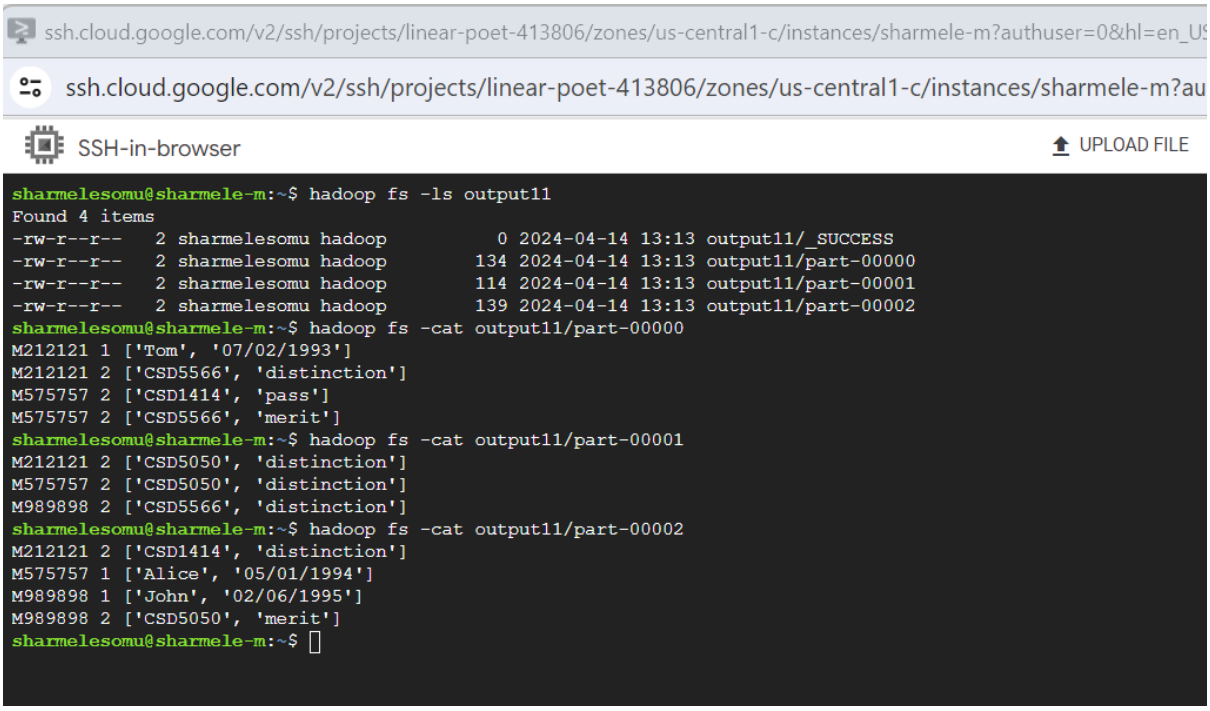
-output /user/sharmelesomu/output16



```
sharmelesomu@sharmele-m:~$ hadoop jar /usr/lib/hadoop/hadoop-streaming-3.3.6.jar -D mapreduce.job.reduces=1 -files /home/sharmelesomu/mapper.
py,/home/sharmelesomu/reducer1.py -mapper "/usr/bin/python3 mapper.py" -reducer "/usr/bin/python3 reducer1.py" -input /user/sharmelesomu/TA.c
sv,/user/sharmelesomu/TB.csv -output /user/sharmelesomu/output16
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6.jar] /tmp/streamjob13952913673159168455.jar tmpDir=null
2024-04-14 16:12:16,297 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at sharmele-m.us-central1-c.c.linear-po
et-413806.internal./10.128.15.222:8032
2024-04-14 16:12:16,497 INFO client.AHSProxy: Connecting to Application History server at sharmele-m.us-central1-c.c.linear-poet-413806.inter
nal./10.128.15.222:10200
2024-04-14 16:12:16,929 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at sharmele-m.us-central1-c.c.linear-po
et-413806.internal./10.128.15.222:8032
2024-04-14 16:12:16,929 INFO client.AHSProxy: Connecting to Application History server at sharmele-m.us-central1-c.c.linear-poet-413806.inter
nal./10.128.15.222:10200
2024-04-14 16:12:17,127 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sharmelesomu/.staging
/job_1713104652195_0004
2024-04-14 16:12:17,538 INFO mapred.FileInputFormat: Total input files to process : 2
2024-04-14 16:12:17,638 INFO mapreduce.JobSubmitter: number of splits:10
2024-04-14 16:12:17,885 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1713104652195_0004
2024-04-14 16:12:17,885 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-04-14 16:12:18,075 INFO conf.Configuration: resource-types.xml not found
2024-04-14 16:12:18,076 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-04-14 16:12:18,154 INFO impl.YarnClientImpl: Submitted application application_1713104652195_0004
2024-04-14 16:12:18,202 INFO mapreduce.Job: The url to track the job: http://sharmele-m.us-central1-c.c.linear-poet-413806.internal.:8088/pro
xy/application_1713104652195_0004/
2024-04-14 16:12:18,203 INFO mapreduce.Job: Running job: job_1713104652195_0004
2024-04-14 16:12:33,398 INFO mapreduce.Job: Job job_1713104652195_0004 running in uber mode : false
2024-04-14 16:12:33,399 INFO mapreduce.Job:  map 0% reduce 0%
2024-04-14 16:12:44,516 INFO mapreduce.Job:  map 10% reduce 0%
2024-04-14 16:12:51,565 INFO mapreduce.Job:  map 30% reduce 0%
2024-04-14 16:12:53,578 INFO mapreduce.Job:  map 40% reduce 0%
2024-04-14 16:13:02,633 INFO mapreduce.Job:  map 50% reduce 0%
2024-04-14 16:13:06,660 INFO mapreduce.Job:  map 70% reduce 0%
2024-04-14 16:13:11,693 INFO mapreduce.Job:  map 80% reduce 0%
2024-04-14 16:13:21,746 INFO mapreduce.Job:  map 90% reduce 0%
2024-04-14 16:13:22,751 INFO mapreduce.Job:  map 100% reduce 0%
2024-04-14 16:13:33,833 INFO mapreduce.Job:  map 100% reduce 100%
2024-04-14 16:13:35,851 INFO mapreduce.Job: Job job_1713104652195_0004 completed successfully
2024-04-14 16:13:35,965 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=415
                FILE: Number of bytes written=3201598
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
```

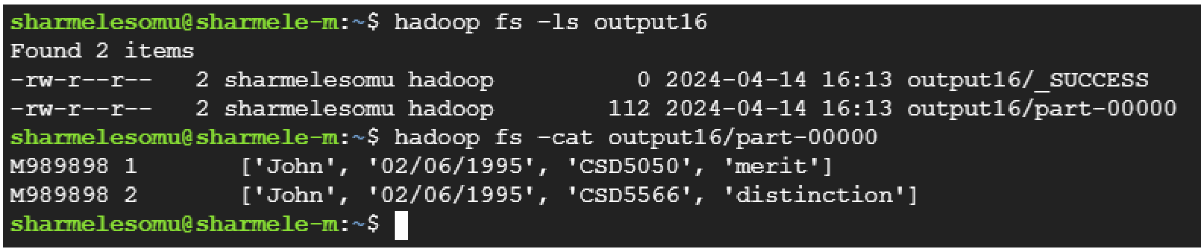
TASK D: RESULT OF THE MAPPER AND REDUCER USING HADOOP STREAMING SERVICE

RESULT OF THE MAPPER IN HADOOP ENVIRONMENT


ssh.cloud.google.com/v2/ssh/projects/linear-poet-413806/zones/us-central1-c/instances/sharmele-m?authuser=0&hl=en_US
ssh.cloud.google.com/v2/ssh/projects/linear-poet-413806/zones/us-central1-c/instances/sharmele-m?au
SSH-in-browser
UPLOAD FILE

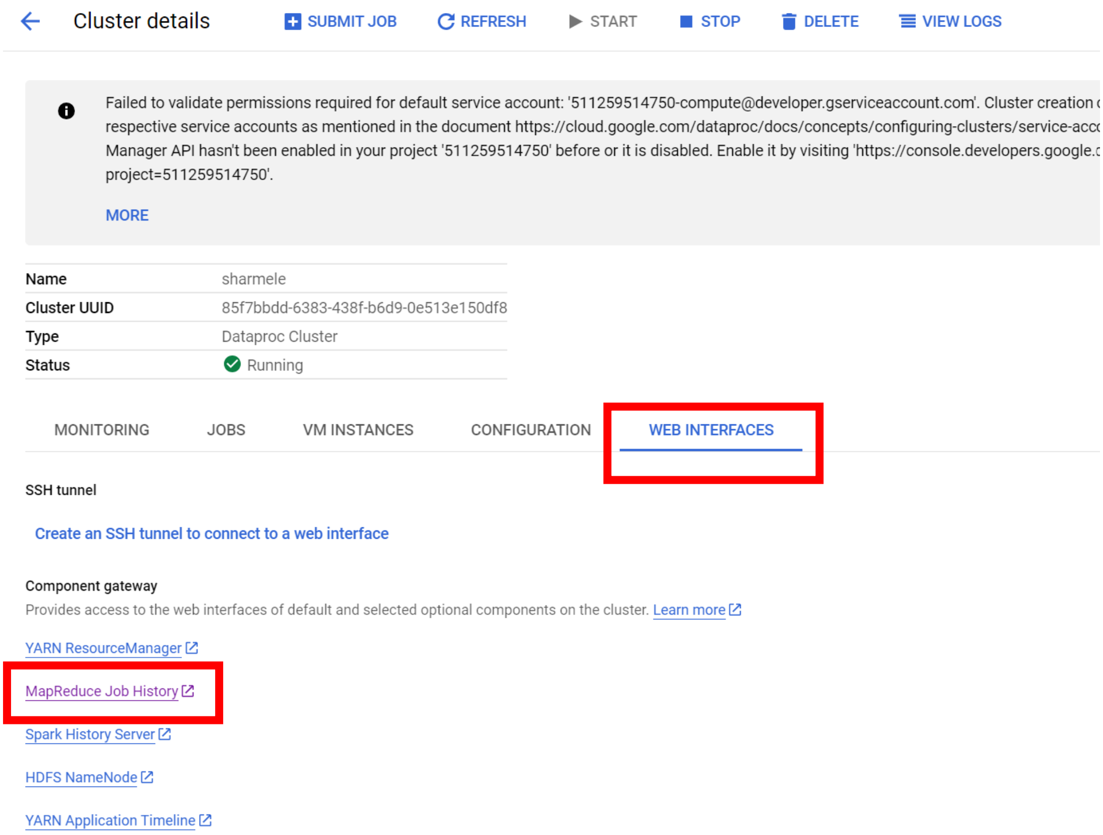
```
sharmelesomu@sharmele-m:~$ hadoop fs -ls output11
Found 4 items
-rw-r--r--   2 sharmelesomu hadoop          0 2024-04-14 13:13 output11/_SUCCESS
-rw-r--r--   2 sharmelesomu hadoop        134 2024-04-14 13:13 output11/part-00000
-rw-r--r--   2 sharmelesomu hadoop        114 2024-04-14 13:13 output11/part-00001
-rw-r--r--   2 sharmelesomu hadoop        139 2024-04-14 13:13 output11/part-00002
sharmelesomu@sharmele-m:~$ hadoop fs -cat output11/part-00000
M212121 1 ['Tom', '07/02/1993']
M212121 2 ['CSD5566', 'distinction']
M575757 2 ['CSD1414', 'pass']
M575757 2 ['CSD5566', 'merit']
sharmelesomu@sharmele-m:~$ hadoop fs -cat output11/part-00001
M212121 2 ['CSD5050', 'distinction']
M575757 2 ['CSD5050', 'distinction']
M989898 2 ['CSD5566', 'distinction']
sharmelesomu@sharmele-m:~$ hadoop fs -cat output11/part-00002
M212121 2 ['CSD1414', 'distinction']
M575757 1 ['Alice', '05/01/1994']
M989898 1 ['John', '02/06/1995']
M989898 2 ['CSD5050', 'merit']
sharmelesomu@sharmele-m:~$
```

RESULT OF THE REDUCER IN HADOOP ENVIRONMENT



```
sharmelesomu@sharmele-m:~$ hadoop fs -ls output16
Found 2 items
-rw-r--r--   2 sharmelesomu hadoop          0 2024-04-14 16:13 output16/_SUCCESS
-rw-r--r--   2 sharmelesomu hadoop        112 2024-04-14 16:13 output16/part-00000
sharmelesomu@sharmele-m:~$ hadoop fs -cat output16/part-00000
M989898 1       ['John', '02/06/1995', 'CSD5050', 'merit']
M989898 2       ['John', '02/06/1995', 'CSD5566', 'distinction']
sharmelesomu@sharmele-m:~$
```

To view the number of mappers and reducers in the Google cloud platform. Navigate to the created cluster, select web interfaces, and click on Mapreduce Job History.



Failed to validate permissions required for default service account: '511259514750-compute@developer.gserviceaccount.com'. Cluster creation
respective service accounts as mentioned in the document https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/service-acc
Manager API hasn't been enabled in your project '511259514750' before or it is disabled. Enable it by visiting 'https://console.developers.google.c
project=511259514750'.

Name	sharmele
Cluster UUID	85f7bbdd-6383-438f-b6d9-0e513e150df8
Type	Dataproc Cluster
Status	Running

SSH tunnel

Create an SSH tunnel to connect to a web interface

Component gateway

Provides access to the web interfaces of default and selected optional components on the cluster. Learn more

YARN ResourceManager

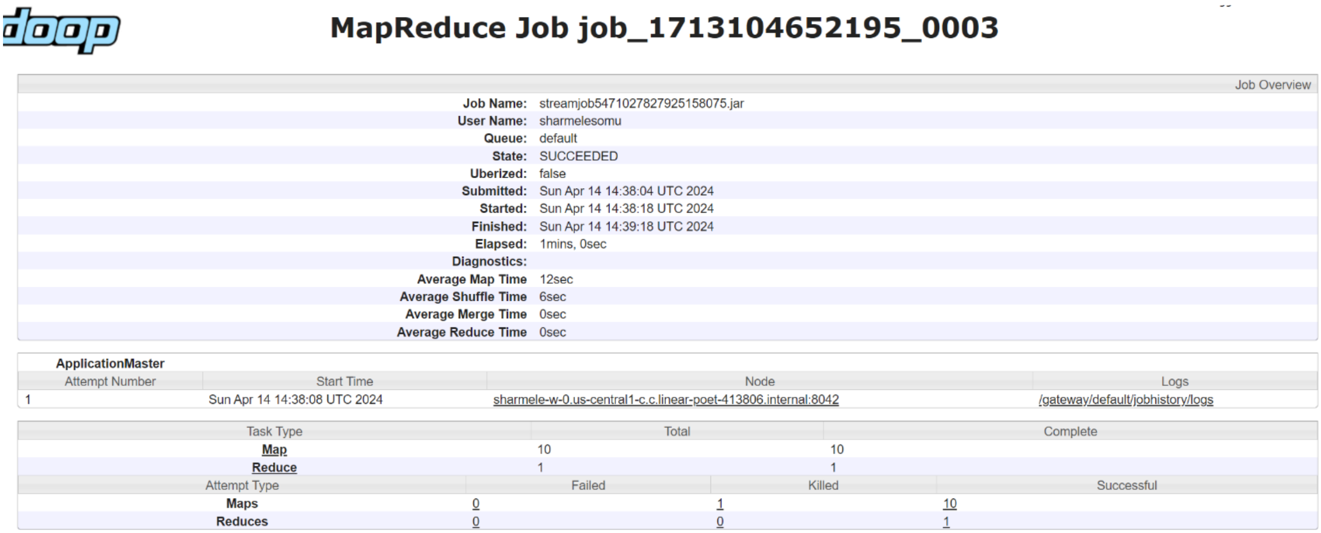
MapReduce Job History

Spark History Server

HDFS NameNode

YARN Application Timeline

We could observe that there are 10 mappers and 1 reducer created for joining two tables and fetching student records with date of birth greater than '01/01/1995'.



MapReduce Job job_1713104652195_0003

Job Overview	
Job Name:	streamjob5471027827925158075.jar
User Name:	sharmelesomu
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sun Apr 14 14:38:04 UTC 2024
Started:	Sun Apr 14 14:38:18 UTC 2024
Finished:	Sun Apr 14 14:39:18 UTC 2024
Elapsed:	1mins, 0sec
Diagnostics:	
Average Map Time	12sec
Average Shuffle Time	6sec
Average Merge Time	0sec
Average Reduce Time	0sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Sun Apr 14 14:38:08 UTC 2024	sharmele-w-0.us-central1-c.c.linear-poet-413806.internal:8042	/gateway/default/jobhistory/logs

Task Type	Total	Complete	
Map	10	10	
Reduce	1	1	
Attempt Type	Failed	Killed	Successful
Maps	0	1	10
Reduces	0	0	1