

# Banking Data Analysis with Hadoop, MRJob, and Apache Hive on GCP Dataproc

## Project Overview

This project demonstrates how to process and analyze large-scale banking data using distributed computing tools such as **Hadoop**, **MapReduce** (via Python's `mrjob`), and **Apache Hive**, all deployed on **Google Cloud Platform (GCP) Dataproc**.

We explored various tasks including:

- Hadoop HDFS operations
- MapReduce jobs using `mrjob` in Python
- Apache Hive for SQL-based analysis

## Step 1: GCP Dataproc Cluster Creation

### ♦ Navigate to Dataproc

1. Go to GCP Console → Dataproc → Clusters
2. Click on “Create Cluster”
3. Choose the following main settings:

Setting	Value
Cluster Name	bank-cluster
Region	us-central1
Zone	us-central1-a
Image Version	2.3.6-debian12
Master Node Type	n4-standard-2
Worker Nodes	2
Worker Type	n4-standard-2
Optional Components	JUPYTER, HIVE_WEBHCAT

Google Cloud | My First Project | data

Dataproc / Clusters / Cluster: cluster-bank / VM Instances

Overview

Notebooks/IDE

- BigQuery Studio
- Workbench

Clusters

- Clusters
- Jobs
- Workflows
- Autoscaling policies

Serverless

- Batches
- Interactive Sessions
- Session Templates

Metastore Services

- Metastore
- Federation

Utilities

- Release Notes

Cluster details

SUBMIT JOB REFRESH START STOP DELETE VIEW LOGS

Failed to validate permissions required for default service account: '8262952441-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](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 '8262952441' before or it is disabled. Enable it by visiting <https://console.developers.google.com/apis/api/cloudresourcemanager.googleapis.com/overview?project=8262952441>.

MORE

Name cluster-bank

Cluster UUID 85a4e95e-3b9f-49d5-bdab-135be4de4c48

Type Dataproc Cluster

Status Running

MONITORING JOBS VM INSTANCES CONFIGURATION WEB INTERFACES

Filter Filter instances

Name	Role	Machine type
cluster-bank-m	Master	n4-standard-2
cluster-bank-w-0	Worker	n4-standard-2
cluster-bank-w-1	Worker	n4-standard-2

EQUIVALENT REST

## ◆ SSH into the Master Node

```
$ gcloud dataproc clusters list
```

```
$ gcloud compute ssh bank-cluster-m --zone=us-central1-a
```

SSH-in-browser

UPLOAD FILE DOWNLOAD FILE

```
pramod@cluster-bank-m:~$ hadoop version
Hadoop 3.3.6
Source code repository https://bigdataoss-internal.googleusercontent.com/third_party/apache/hadoop -r 756bf396804b712d2cb7f748cb0b3c30e70d78c9
Compiled by bigtop on 2025-07-09T21:44Z
Compiled on platform linux-x86_64
Compiled with protoc 3.25.5
From source with checksum afca53245f7160b91433a087d476293f
This command was run using /usr/lib/hadoop/hadoop-common-3.3.6.jar
pramod@cluster-bank-m:~$ hadoop fs -ls
pramod@cluster-bank-m:~$ pwd
/home/pramod
pramod@cluster-bank-m:~$ whoai
-bash: whoai: command not found
pramod@cluster-bank-m:~$ whoami
pramod
pramod@cluster-bank-m:~$ df -h
Filesystem      Size  Used Avail Use% Mounted on
udev            3.9G     0  3.9G   0% /dev
tmpfs           795M  672K  795M   1% /run
/dev/nvme0n1p1  32G   15G   15G  51% /
tmpfs           3.9G     0  3.9G   0% /dev/shm
tmpfs           5.0M     0  5.0M   0% /run/lock
tmpfs           4.0M     0  4.0M   0% /sys/fs/cgroup
/dev/nvme0n1p15 124M  12M  113M  10% /boot/efi
tmpfs           795M     0  795M   0% /run/user/108
tmpfs           795M     0  795M   0% /run/user/106
tmpfs           795M     0  795M   0% /run/user/110
tmpfs           795M     0  795M   0% /run/user/107
tmpfs           795M     0  795M   0% /run/user/111
tmpfs           795M     0  795M   0% /run/user/1002
pramod@cluster-bank-m:~$
```

## Step 2: Upload Dataset to HDFS

- ◆ Place dataset on the master node

```
$ scp bank.csv pramod@<master-node-ip>:/home/pramod/
```

- ◆ Put dataset in HDFS

```
$ hadoop fs -put bank.csv /user/pramod/
```

```
$ hadoop fs -ls /user/pramod
```



The image shows a terminal window titled "SSH-in-browser" connected to a Linux cluster. The terminal output shows the user 'pramod' logging in and running 'ls' commands to verify the file 'bank.csv' is present in the directory '/user/pramod/'. A small overlay at the bottom right of the terminal window indicates "Transferred 1 item" and "bank.csv" with a green checkmark, confirming the successful upload to HDFS.

```
Linux cluster-bank-m 6.1.0-37-cloud-amd64 #1 SMP PREEMPT_DYNAMIC Debian 6.1.140-1 (2025-05-22) 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: Fri Jul 25 06:29:10 2025 from 35.235.241.18
pramod@cluster-bank-m:~$ ls
pramod@cluster-bank-m:~$ ls
bank.csv
pramod@cluster-bank-m:~$
```

Transferred 1 item  
bank.csv


## Step 3: Run MapReduce Jobs using `mrjob`



Job	Directory	Python File	Output Dir	Command
1 Subscriptions by Job	job1_subscription_by_job	job1_subscription_by_job.py	/user/pramod/output/job1	python3 job1_subscription_by_job.py bank.csv -r hadoop --output-dir hdfs:///user/pramod/output/job1

Job	Directory	Python File	Output Dir	Command
2 Avg Balance by Education	job2_avg_balance_by_education	job2_avg_balance_by_education.py	/user/pramod/output/job2	python3 job2_avg_balance_by_education.py bank.csv -r hadoop --output-dir hdfs:///user/pramod/output/job2
3 Conversion Rate by Month	job3_conversion_rate_by_month	job3_conversion_rate_by_month.py	/user/pramod/output/job3	python3 job3_conversion_rate_by_month.py bank.csv -r hadoop --output-dir hdfs:///user/pramod/output/job3

### ♦ View Output Example

```
$ hadoop fs -cat /user/pramod/output/job1/part-*
```

 SSH-in-browser
 

UPLOAD FILE
 DOWNLOAD FILE
 



```

pramod@cluster-bank-m:~$ hadoop fs -cat /user/pramod/output/job1/part-*
"blue-collar"      69
"services"         38
"student"          19
"admin."           58
"self-employed"    20
"technician"       83
"unemployed"       13
"entrepreneur"     15
"housemaid"        14
"management"      131
"retired"          54
"unknown"          7
pramod@cluster-bank-m:~$

```

```
$ hadoop fs -cat /user/pramod/output/job2/part-*
```

```
SSH-in-browser
pramod@cluster-bank-m:~$ hadoop fs -cat /user/pramod/output/job2/part-*
"secondary"      1196.81
"tertiary"       1775.42
"primary"        1411.54
"unknown"        1701.25
pramod@cluster-bank-m:~$
```

```
$ hadoop fs -cat /user/pramod/output/job3/part-*
```

```
SSH-in-browser
pramod@cluster-bank-m:~$ hadoop fs -cat /user/pramod/output/job3/part-*
"dec" "45.00%"
"jun" "10.36%"
"may" "6.65%"
"nov" "10.03%"
"feb" "17.12%"
"jan" "10.81%"
"jul" "8.64%"
"sep" "32.69%"
"apr" "19.11%"
"aug" "12.48%"
"mar" "42.86%"
"oct" "46.25%"
pramod@cluster-bank-m:~$
```

---

## Step 4: Apache Hive Setup and Queries

Apache Hive is a distributed, fault-tolerant data warehouse system that enables analytics at a massive scale. Hive Metastore(HMS) provides a central repository of metadata that can easily be analyzed to make informed, data driven decisions, and therefore it is a critical component of many data lake architectures. Hive is built on top of Apache Hadoop and

supports storage on S3, adls, gs etc though hdfs. Hive allows users to read, write, and manage petabytes of data using SQL.

#### ♦ Launch HiveShell or Beeline

```
$ hive
```

OR

```
$ beeline
!connect jdbc:hive2://localhost:10000
```

#### ♦ Create Database and Table

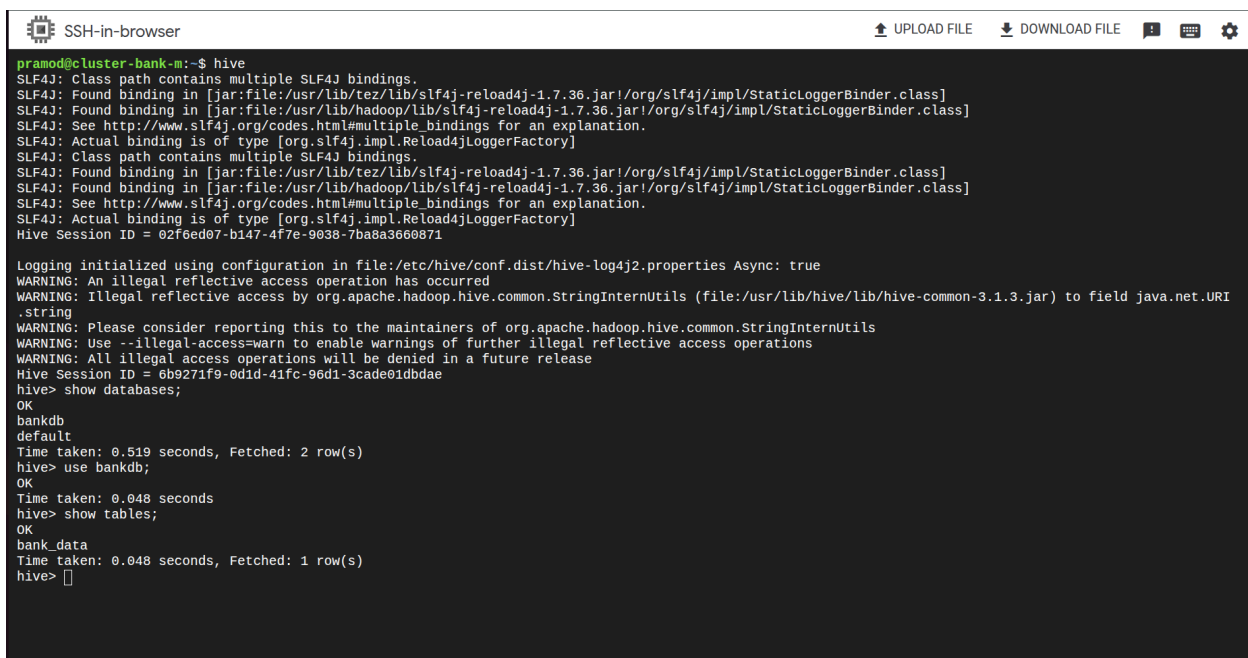
```
CREATE DATABASE IF NOT EXISTS bankdb;
USE bankdb;
```

```
CREATE EXTERNAL TABLE bank_data (
  age INT,
  job STRING,
  marital STRING,
  education STRING,
  default STRING,
  balance INT,
  housing STRING,
  loan STRING,
  contact STRING,
  day INT,
  month STRING,
  duration INT,
  campaign INT,
  pdays INT,
  previous INT,
  poutcome STRING,
  y STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/pramod/';
```

## ◆ Create External Table

Defines a table named `bank_data` that reads CSV data from HDFS without owning the file.

- **External Table** → Hive won't delete data if table is dropped
- **ROW FORMAT DELIMITED** → Data is plain text
- **FIELDS TERMINATED BY ','** → Columns separated by commas (CSV)
- **STORED AS TEXTFILE** → Stored as simple text
- **LOCATION '/user/pramod/'** → Data resides in HDFS at this path



```
pramod@cluster-bank-m:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/tez/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/tez/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
Hive Session ID = 02f6ed07-b147-4f7e-9038-7ba8a3660871

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.hive.common.StringInternUtils (file:/usr/lib/hive/lib/hive-common-3.1.3.jar) to field java.net.URI
.string
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.hive.common.StringInternUtils
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Hive Session ID = 6b9271f9-0d1d-41fc-96d1-3cade01dbdae
hive> show databases;
OK
bankdb
default
Time taken: 0.519 seconds, Fetched: 2 row(s)
hive> use bankdb;
OK
Time taken: 0.048 seconds
hive> show tables;
OK
bank_data
Time taken: 0.048 seconds, Fetched: 1 row(s)
hive>
```

## ◆ Hive Queries

### ◆ Query 1: Average balance grouped by job

```
SELECT job, AVG(balance) FROM bank_data GROUP BY job;
```

```
m7jqvof6wzcdplhrsrj4ldwm-dot-us-central1.dataproc.googleusercontent.com/gateway/default/jupyter/lab/
iconic-apricot-466612-i6 > cluster-bank Sign out
File Edit View Run Kernel Git Tabs Settings Help
Terminal 1 Terminal 2
Map 1 ..... container SUCCEEDED 3 3 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 11.34 s
INFO : Completed executing command(queryId=hive_20250728170945_ba2a3b8d-4157-4f98-b072-72c98b5d6186); Time taken: 11.543 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| job | average_age |
+-----+
| NULL | NULL |
| admin. | 1226.73640167364 |
| blue-collar | 1085.161733615222 |
| entrepreneur | 1645.125 |
| housemaid | 2083.8035714285716 |
| management | 1766.9287925696594 |
| retired | 2319.191304347826 |
| self-employed | 1392.4098360655737 |
| services | 1103.9568345323742 |
| student | 1543.8214285714287 |
| technician | 1330.99609375 |
| unemployed | 1089.421875 |
| unknown | 1501.7105263157894 |
+-----+
13 rows selected (11.92 seconds)
```

♦ Query 2: Count of term deposit subscriptions by education

**SELECT** education, **COUNT**(\*) **FROM** bank\_data **WHERE** y = 'yes' **GROUP BY** education;

```
m7jqvof6wzcdplhrsrj4ldwm-dot-us-central1.dataproc.googleusercontent.com/gateway/default/jupyter/lab/
iconic-apricot-466612-i6 > cluster-bank Sign out
File Edit View Run Kernel Git Tabs Settings Help
Terminal 1 Terminal 2
INFO : Subscribed to counters: [] for queryId: hive_20250728171454_fd109bcf-06e5-4ef7-b4cd-779e1358bb95
INFO : Session is already open
INFO : Dag name: SELECT education, COUNT(*) as su...education (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1753721570210_0001)
-----
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED 3 3 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 12.43 s
INFO : Completed executing command(queryId=hive_20250728171454_fd109bcf-06e5-4ef7-b4cd-779e1358bb95); Time taken: 12.62 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| education | subscription_count |
+-----+
| primary | 64 |
| secondary | 245 |
| tertiary | 193 |
| unknown | 19 |
+-----+
4 rows selected (12.917 seconds)
0: jdbc:hive2://localhost:10000>
```



♦ *Query 3: Month-wise campaign success*

```
SELECT month, COUNT(*) FROM bank_data WHERE y = 'yes' GROUP BY month;
```

← → ↺ m7jqvofwzcdplhrsrjr4ldwm-dot-us-central1.dataproc.googleusercontent.com/gateway/default/jupyter/lab/ 🔍 ☆ 🌐 ⋮

iconic-apricot-466612-i6 > cluster-bank Sign out

File Edit View Run Kernel Git Tabs Settings Help

Terminal 1 Terminal 2

```
Map 1 ..... container SUCCEEDED 3 3 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 12.10 s
INFO : Completed executing command(queryId=hive_20250728171814_9bf4b6f2-b6e9-48f3-9af5-8d19ccb30fde); Time taken: 12.28 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| month | successful_campaigns |
+-----+
| apr   | 56                   |
| aug   | 79                   |
| dec   | 9                    |
| feb   | 38                   |
| jan   | 16                   |
| jul   | 61                   |
| jun   | 55                   |
| mar   | 21                   |
| may   | 93                   |
| nov   | 39                   |
| oct   | 37                   |
| sep   | 17                   |
+-----+
12 rows selected (12.538 seconds)
```