

**GEORGIA INSTITUTE OF TECHNOLOGY**  
**SCHOOL of ELECTRICAL and COMPUTER ENGINEERING**

**ECE 4150-A Spring 2024**

**Lab: Batch Data Analysis using Hadoop, MapReduce, Pig & Hive**

**References:**

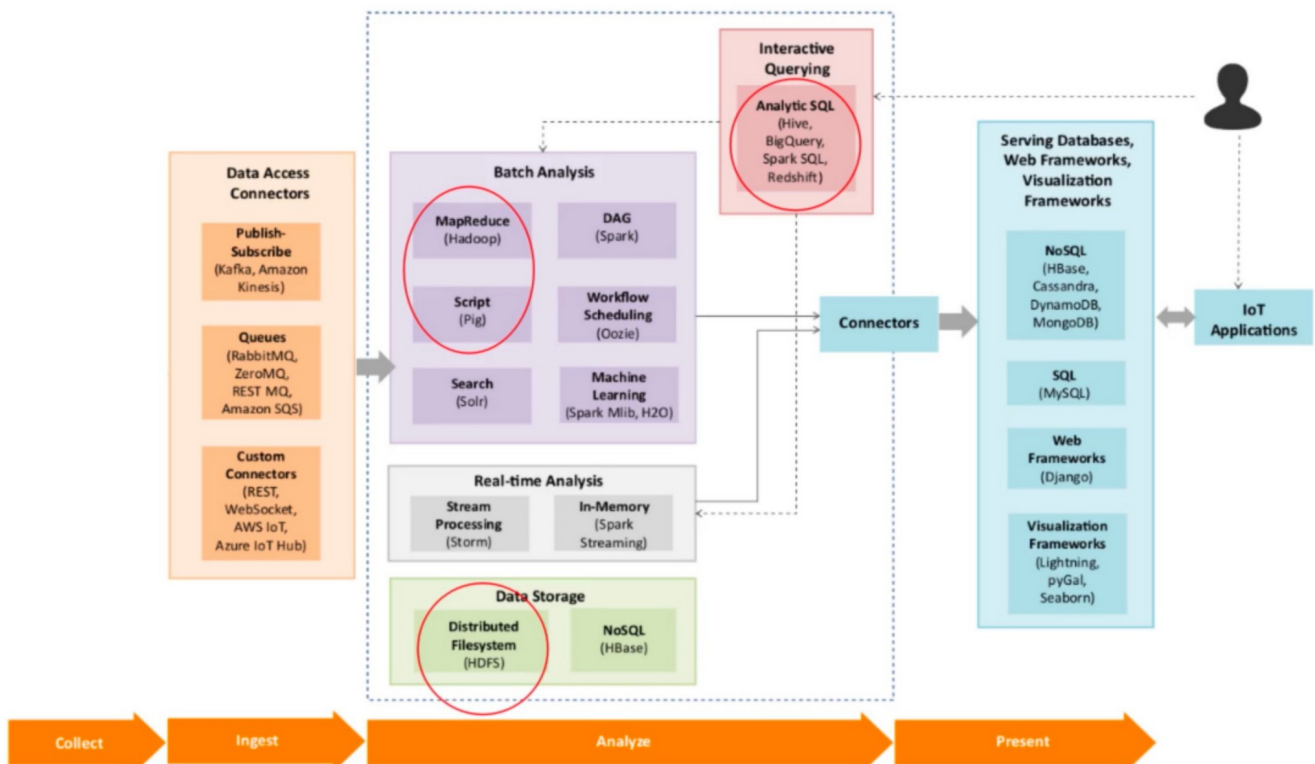
- [1] A. Bahga, V. Madiseti, *Cloud Computing Solutions Architect: A Hands-On Approach*, ISBN: 978-0996025591
- [2] <https://pythonhosted.org/mrjob/>
- [3] <http://hadoop.apache.org/>
- [4] <http://storage.googleapis.com/books/ngrams/books/datasetsv2.html>
- [5] <http://pig.apache.org/docs/r0.15.0/basic.html>
- [6] <https://cwiki.apache.org/confluence/display/Hive/LanguageManual>

**Due Date:**

The lab report will be due on **April 14th, 2024 at 11:59 PM.**

In this lab you will learn how setup a Hadoop cluster and run MapReduce, Pig and Hive job.

Fig.1 Architecture diagram of data processing in Hadoop



## 1. Set up a Hadoop Cluster with EMR

Navigate to Amazon EMR console and create a new cluster with the following configurations:

**Create cluster** [Info](#)

**Name and applications - required** [Info](#)  
Name your cluster and choose the applications that you want to install on your cluster.

Name  
sam-lab5

Amazon EMR release [Info](#)  
A release contains a set of applications which can be installed on your cluster.  
emr-7.0.0

Application bundle

Spark Interactive	<b>Core Hadoop</b>	Flink	HBase	Presto	Trino	Custom
-------------------	--------------------	-------	-------	--------	-------	--------

☐ AmazonCloudWatchAgent 1.300031.1  
☐ HCatalog 3.1.3  
☒ Hue 4.11.0  
☐ Livy 0.7.1  
☐ Phoenix 5.1.3  
☐ Spark 3.5.0  
☒ Tez 0.10.2  
☐ ZooKeeper 3.5.10

☐ Flink 1.18.0  
☒ Hadoop 3.3.6  
☐ JupyterEnterpriseGateway 2.6.0  
☐ MXNet 1.9.1  
☒ Pig 0.17.0  
☐ Sqoop 1.4.7  
☐ Trino 426

☐ HBase 2.4.17  
☒ Hive 3.1.3  
☐ JupyterHub 1.5.0  
☐ Oozie 5.2.1  
☐ Presto 0.283  
☐ TensorFlow 2.11.0  
☐ Zeppelin 0.10.1

**AWS Glue Data Catalog settings**  
Use the AWS Glue Data Catalog to provide an external metastore for your application.  
☐ Use for Hive table metadata

**Operating system options** [Info](#)  
☒ Amazon Linux release  
☐ Custom Amazon Machine Image (AMI)  
☒ Automatically apply latest Amazon Linux updates

**Cluster configuration - required** [Info](#)  
Choose a configuration method for the primary, core, and task node groups for your cluster.

☒ **Uniform instance groups**  
Choose the same EC2 instance type and purchasing option (On-Demand or Spot) for all nodes in your node group. [Learn more](#)

☐ **Flexible instance fleets**  
Choose from the widest variety of provisioning options for the EC2 instances in your cluster. Diversify instance types and purchasing options, and use an allocation strategy. [Learn more](#)

**Summary** [Info](#)

**Name and applications - required**

Name  
sam-lab5

Amazon EMR release  
emr-7.0.0

Application bundle  
Core Hadoop (Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, Pig 0.17.0, Tez 0.10.2)

**Cluster configuration - required**

Uniform instance groups  
Primary (m5.xlarge), Core (m5.xlarge)

**Cluster scaling and provisioning - required**

Provisioning configuration  
Core size: 1 instance

[Cancel](#) [Create cluster](#)

aws

Services

Search

[Alt+S]

▼ Cluster configuration - required

Info

Choose a configuration method for the primary, core, and task node groups for your cluster.

☒ Uniform instance groups

Choose the same EC2 instance type and purchasing option (On-Demand or Spot) for all nodes in your node group. [Learn more](#)

☐ Flexible instance fleets

Choose from the widest variety of provisioning options for the EC2 instances in your cluster. Diversify instance types and purchasing options, and use an allocation strategy. [Learn more](#)

Uniform instance groups

Primary

Choose EC2 instance type

m5.xlarge

4 vCore 16 GiB memory EBS only storage

On-Demand price: \$0.192 per instance/hour

Lowest Spot price: \$0.076 (us-east-1f)

Actions ▼

☐ Use high availability

Launch highly available, more resilient cluster with three primary nodes on On-Demand Instances. This configuration applies for the lifetime of your cluster. [Learn more](#)

► Node configuration - optional

Core

Choose EC2 instance type

m5.xlarge

4 vCore 16 GiB memory EBS only storage

On-Demand price: \$0.192 per instance/hour

Lowest Spot price: \$0.076 (us-east-1f)

Actions ▼

Remove instance group

► Node configuration - optional

Add task instance group

You can add up to 48 more task instance groups.

EBS root volume

EBS root volume applies to the operating systems and applications that you install on the cluster. [EBS root volume ratio constraints](#)

Size (GiB)

15

15 - 100 GiB per volume  
General Purpose SSD (gp3)

IOPS

3000

3000 - 16000 IOPS per volume.  
Choose a maximum ratio of 500:1 between IOPS and volume size.

Throughput (MiB/s)

125

125 - 1000 MiB/s per volume.  
Choose a maximum ratio of 0.25:1 between throughput and IOPS.

Summary

Info

Name

sam-lab5

Amazon EMR release

emr-7.0.0

Application bundle

Core Hadoop (Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, Pig 0.17.0, Tez 0.10.2)

Cluster configuration - required

Uniform instance groups  
Primary (m5.xlarge), Core (m5.xlarge)

Cluster scaling and provisioning - required

Provisioning configuration  
Core size: 1 instance

Cancel

Create cluster

## Uniform instance groups

### Primary


Choose EC2 instance type

m5.xlarge

4 vCore 16 GiB memory EBS only storage  
On-Demand price: \$0.192 per instance/hour  
Lowest Spot price: \$0.076 (us-east-1f)

Actions ▼

☐ Use high availability

Launch highly available, more resilient cluster with three primary nodes on On-Demand Instances. This configuration applies for the lifetime of your cluster. [Learn more](#) 

► Node configuration - optional

### Core

Choose EC2 instance type

m5.xlarge

4 vCore 16 GiB memory EBS only storage  
On-Demand price: \$0.192 per instance/hour  
Lowest Spot price: \$0.076 (us-east-1f)

Actions ▼

► Node configuration - optional

Remove Task



Remove instance group

### Task 1 of 1

Name

Task - 1

Choose EC2 instance type

m5.xlarge

4 vCore 16 GiB memory EBS only storage  
On-Demand price: \$0.192 per instance/hour  
Lowest Spot price: \$0.076 (us-east-1f)

Actions ▼

► Node configuration - optional

Add task instance group

You can add up to 47 more task instance groups.

### EBS root volume

EBS root volume applies to the operating systems and applications that you install on the cluster. [EBS root volume ratio constraints](#) 

Size (GiB)

IOPS

Throughput (MiB/s)

### EBS root volume

EBS root volume applies to the operating systems and applications that you install on the cluster. [EBS root volume ratio constraints](#)

Size (GiB)

15

15 - 100 GiB per volume  
General Purpose SSD (gp3)

IOPS

3000

3000 - 16000 IOPS per volume.  
Choose a maximum ratio of  
500:1 between IOPS and  
volume size.

Throughput (MiB/s)

125

125 - 1000 MiB/s per volume.  
Choose a maximum ratio of  
0.25:1 between throughput  
and IOPS.

### ▼ Cluster scaling and provisioning - required [Info](#)

Choose how Amazon EMR should size your cluster.

Choose an option

☒ Set cluster size manually

Use this option if you know your  
workload patterns in advance.

☐ Use EMR-managed scaling

Monitor key workload metrics so  
that EMR can optimize the cluster  
size and resource utilization.

☐ Use custom automatic  
scaling

To programmatically scale core  
and task nodes, create custom  
automatic scaling policies.

### Provisioning configuration

Set the size of your core and task instance groups. Amazon EMR attempts to provision this capacity when you launch your cluster.

Name	Instance type	Instance(s) size	Use Spot purchasing option
Core	m5.xlarge	1	<input type="checkbox"/>
Task - 1	m5.xlarge	1	<input type="checkbox"/>

### ▼ Networking - required [Info](#)

Choose the network settings that determine how you and other users can connect to your cluster.

Virtual private cloud (VPC) [Info](#)

vpc-5e663f24

[Browse](#)

[Create VPC](#)

Subnet [Info](#)

subnet-0d73b403

[Browse](#)

[Create subnet](#)

► EC2 security groups (firewall)

Keep the default value of  
you can setup your own VPC



## EC2 instance profile for Amazon EMR

The instance profile assigns a role to every EC2 instance in a cluster. The instance profile must specify a role that can access the resources for your steps and bootstrap actions.

- ☐ Choose an existing instance profile
- Select a default role or a custom instance profile with IAM policies attached so that your cluster can interact with your resources in Amazon S3.

- ☒ Create an instance profile
- Let Amazon EMR create a new instance profile so that you can specify a custom set of resources for it to access in Amazon S3.

### S3 bucket access [Info](#)

- ☒ Specific S3 buckets or prefixes in your account [Info](#)
- Choose the buckets or prefixes that you want this instance profile to access.
- ☐ All S3 buckets in this account with read and write access
- Grant the instance profile access to all buckets that have read and write access enabled in your account.

### S3 buckets

We've already added the resources that you configured in the [Cluster logs](#) section. Choose the S3 buckets and bucket prefixes where you store logs and data for your cluster, bootstrap actions, and steps.

#### S3 URI

[View](#) [Browse S3](#) [Add](#)

S3 bucket	Prefix	Permission	
aws-logs-79777061...	elasticmapreduce_sam	Read and write	<a href="#">Edit</a>
Inherited from Cluster logs			

### Custom automatic scaling role - *optional*

When a custom automatic scaling rule triggers, Amazon EMR assumes this role to add and terminate EC2 instances. [Learn more](#)

#### Custom automatic scaling role

[Create IAM role](#)

## Summary [Info](#)

### Name and applications - *required*

Name  
sam-lab5

Amazon EMR release  
emr-7.0.0

Application bundle  
Core Hadoop (Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, Pig 0.17.0, Tez 0.10.2)

### Cluster configuration - *required*

Click on **Create**

Wait for the cluster to be created and enter the state of **Waiting**, which usually takes 5 minutes to finish.

Navigate to **EC2 instance** and open the one that's running, which holds the cluster that you just created:

Amazon EMR console showing a list of clusters. The cluster 'lab5' with ID 'j-3102ZF2NL7BDJ' is highlighted in a red box. An orange arrow points to this cluster with the text 'Click on this'.

Cluster ID	Cluster name	Status	Creation time (UTC-04:00)	Elapsed time	Normalized instance hours
j-3102ZF2NL7BDJ	lab5	Starting Preparing cluster	March 30, 2024, 17:04	5 seconds	0
j-2756OE2TPPD35	lab5	Terminated with errors Instance failure	March 29, 2024, 11:10	35 minutes, 1 second	16
j-375POX2038812	lab5	Terminated User request	March 29, 2024, 10:38	1 minute, 32 seconds	0

Go to **Security-Security groups** and open the security group for the master cluster, in this case **ElasticMapReduce-master**:

Amazon EMR console showing the details of cluster 'lab5'. An orange arrow points to the 'Primary node public DNS' field in the 'Cluster management' section with the text '2. Click here and copy the syntax of SSH'. Another orange arrow points to the 'Primary node EMR managed security group' field in the 'Network and security' section with the text '1. Click here to add SSH rule (else you won't be able to connect to your EMR instance)'.

**Cluster info**

- Cluster ID: j-3102ZF2NL7BDJ
- Cluster configuration: emr-7.0.0
- Installed applications: Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, Pig 0.17.0, Tez 0.10.2
- Capacity: 1 Primary 1 Core 0 Task

**Cluster management**

- Log destination in Amazon S3: aws-logs-211125774779-us-east-1/elasticmapreduce\_sam
- Primary node public DNS: ec2-54-234-96-128.compute-1.amazonaws.com
- Connect to the Primary node using SSH
- Connect to the Primary node using SSM

**Status and time**

- Status: Starting
- Creation time: March 30, 2024, 17:04 (UTC-04:00)
- Elapsed time: 4 minutes, 46 seconds

**Operating system**

- Amazon Linux release: 2023.3.20240512.0

**Cluster logs**

- Archive log files to Amazon S3: Turned on
- Encryption for logs: Turned off

**Cluster termination and node replacement**

- Termination option: Automatically terminate cluster after idle time
- Idle time: 1 hour
- Termination protection: Off
- Unhealthy node replacement: On

**Network and security**

- Virtual Private Cloud (VPC): vpc-01ba9790c80584d69
- Subnet(s) and Availability Zone(s) (AZ): subnet-0c939786f7c8d8e92 us-east-1d
- EC2 security groups (firewall): sg-08f63b1dfd15c7980

**Security configuration**

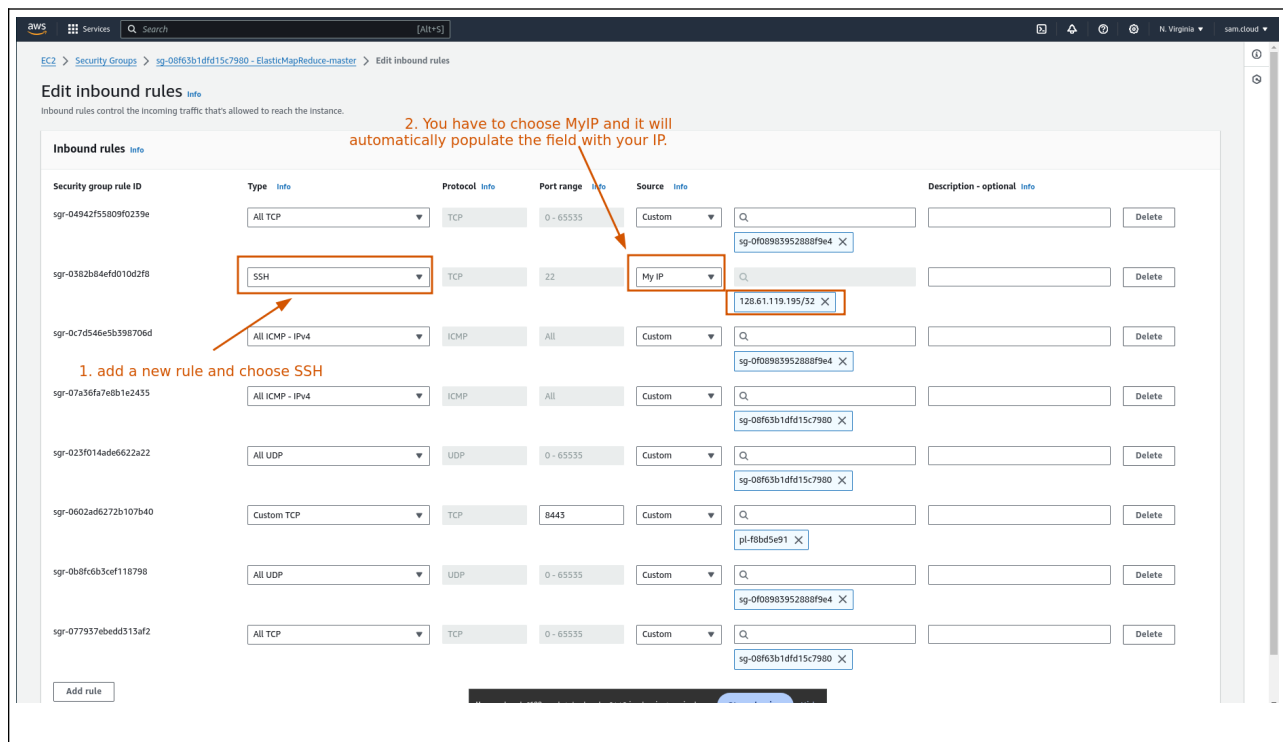
- Security configuration: None
- EC2 key pair: lab3-4

**Permissions**

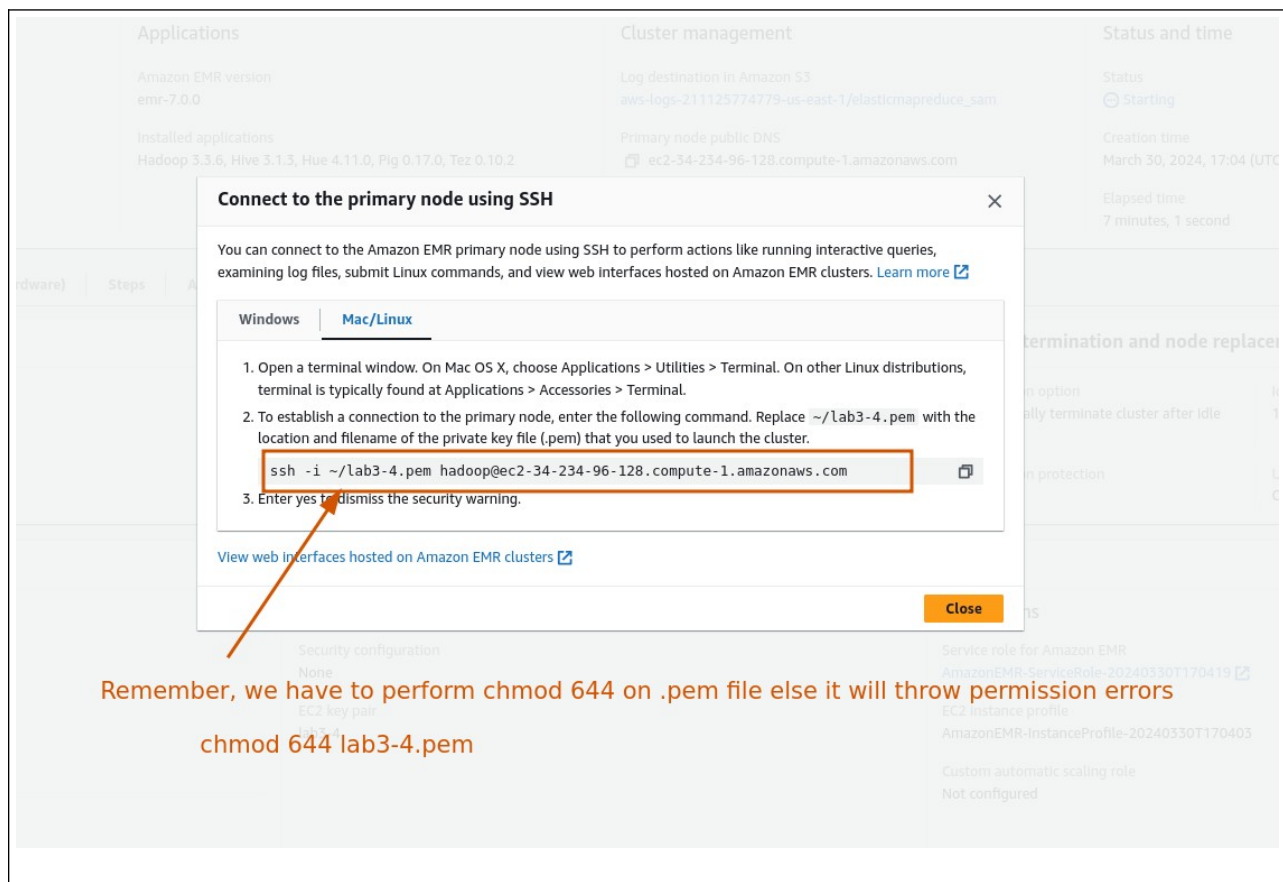
- Service role for Amazon EMR: AmazonEMR-ServiceRole-20240330T170419
- EC2 instance profile: AmazonEMR-InstanceProfile-20240330T170403
- Custom automatic scaling role: Not configured

Edit the inbound rules and add a rule for SSH and save it:





Now go back to your **EMR cluster**, click on “**connect to the master node using SSH**” and follow the instruction to connect to the master node:



For example, a successful connection would appear like: (You can use either 644 or 600 for the permission. Usually 600 is recommended)

[illegible]

## 2. Upload Datasets to HDFS

**2.1 In the AWS Image of EMR has kept the old configuration however, new version of the hadoop runs HEAVs in a different port, so we have to manually change it. Here is the step to change that.**

aws

Services

Search

[Alt+S]

N. Virginia

sam.cloud

Amazon EMR > EMR on EC2: Clusters > lab5

lab5

Updated 5 minutes ago

Terminate

Clone in AWS CLI

Clone

▼ Summary

Cluster info

Cluster ID  
j-3I02ZF2NL7BDJ

Cluster configuration

Instance groups  
1 Primary | 1 Core | 0 Task

Capacity  
1 Primary | 1 Core | 0 Task

Applications

Amazon EMR version  
emr-7.0.0

Installed applications  
Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, Pig 0.17.0, Tez 0.10.2

Cluster management

Log destination in Amazon S3  
aws-logs-211125774779-us-east-1/elasticmapreduce\_sam

Persistent application Uis  
YARN timeline server  
Tez UI

Primary node public DNS  
ec2-34-234-96-128.compute-1.amazonaws.com

Connect to the Primary node using SSH

Connect to the Primary node using SSM

Status and time

Status  
Waiting

Creation time  
March 30, 2024, 17:04 (UTC-04:00)

Elapsed time  
52 minutes, 51 seconds

Properties

Bootstrap actions

Instances (Hardware)

Steps

Applications

Configurations

Monitoring

Events

Application user interfaces

Applications installed on your Amazon EMR cluster publish user interfaces (UI) as websites. You can use these to monitor cluster activity.

On-cluster application Uis

On-cluster Uis are available only while your cluster is running. Use the following links to get started. To access all the application Uis, set up SSH tunneling.

Persistent application Uis

Persistent Uis don't require SSH tunneling. They are hosted off of the cluster and are available for 30 days after an application ends.

Live Application Uis

These on-cluster application Uis are available without SSH tunneling.

Application Uis

No live application Uis

No live application Uis to display

Application Uis on the primary node

These require SSH tunneling to be enabled.

Application	UI URL
HDFS Name Node	http://ec2-34-234-96-128.compute-1.amazonaws.com:9870/
Hue	http://ec2-34-234-96-128.compute-1.amazonaws.com:8888/
Resource Manager	http://ec2-34-234-96-128.compute-1.amazonaws.com:8088/
Tez UI	http://ec2-34-234-96-128.compute-1.amazonaws.com:8080/tez-ui

Application Uis on the core and task nodes

Application	UI URL
HDFS Data Node	http://ec2-000-000-000-000.compute-1.amazonaws.com:9864/
Node Manager	http://ec2-000-000-000-000.compute-1.amazonaws.com:8042/

Installed Applications (5)

Hadoop 3.3.6

Hive 3.1.3

Hue 4.11.0

Pig 0.17.0

Tez 0.10.2

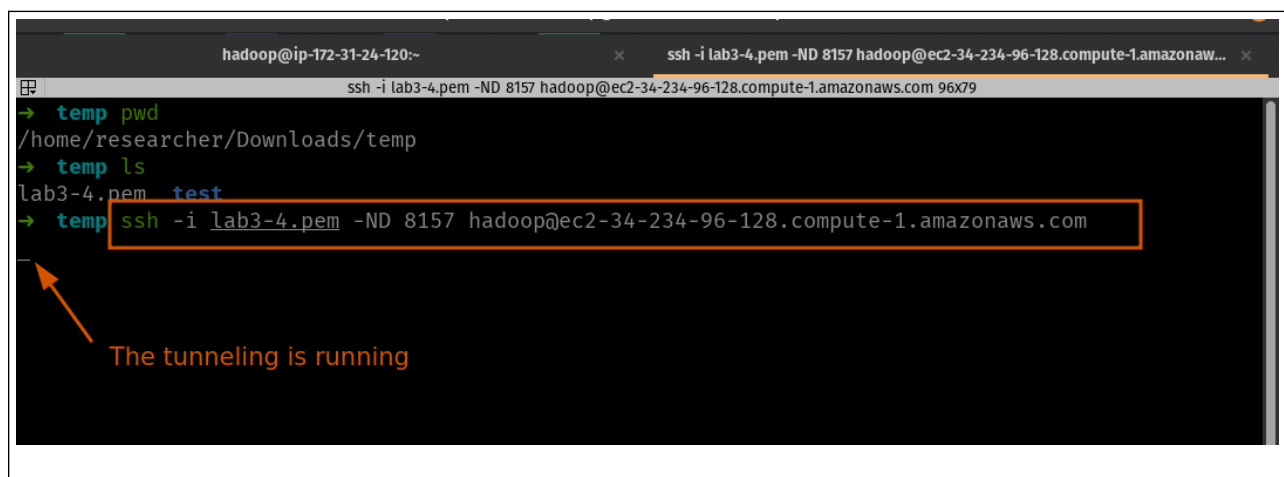
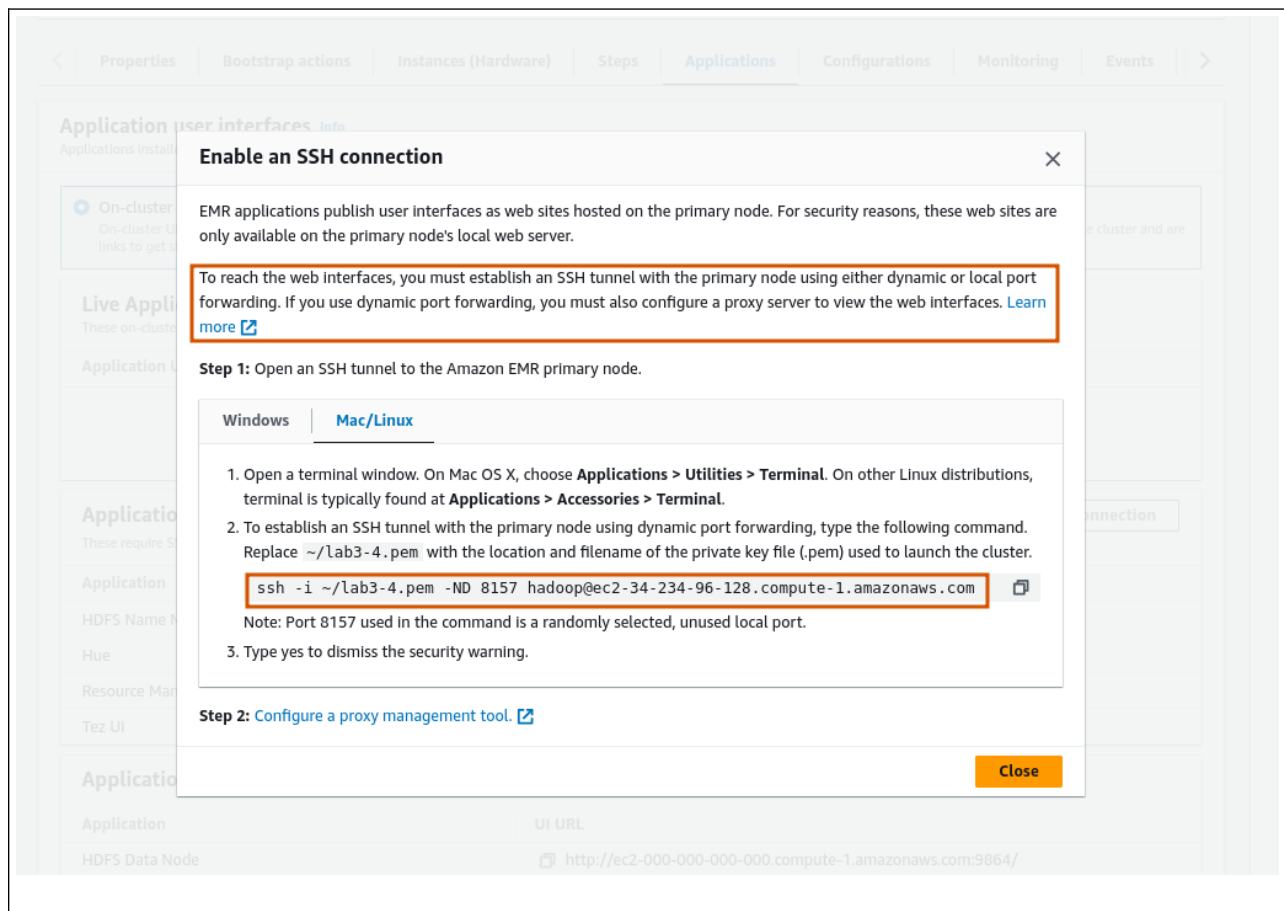
1. Click Here

2. We need to enable the tunnel and you can use this command to do the tunnelling

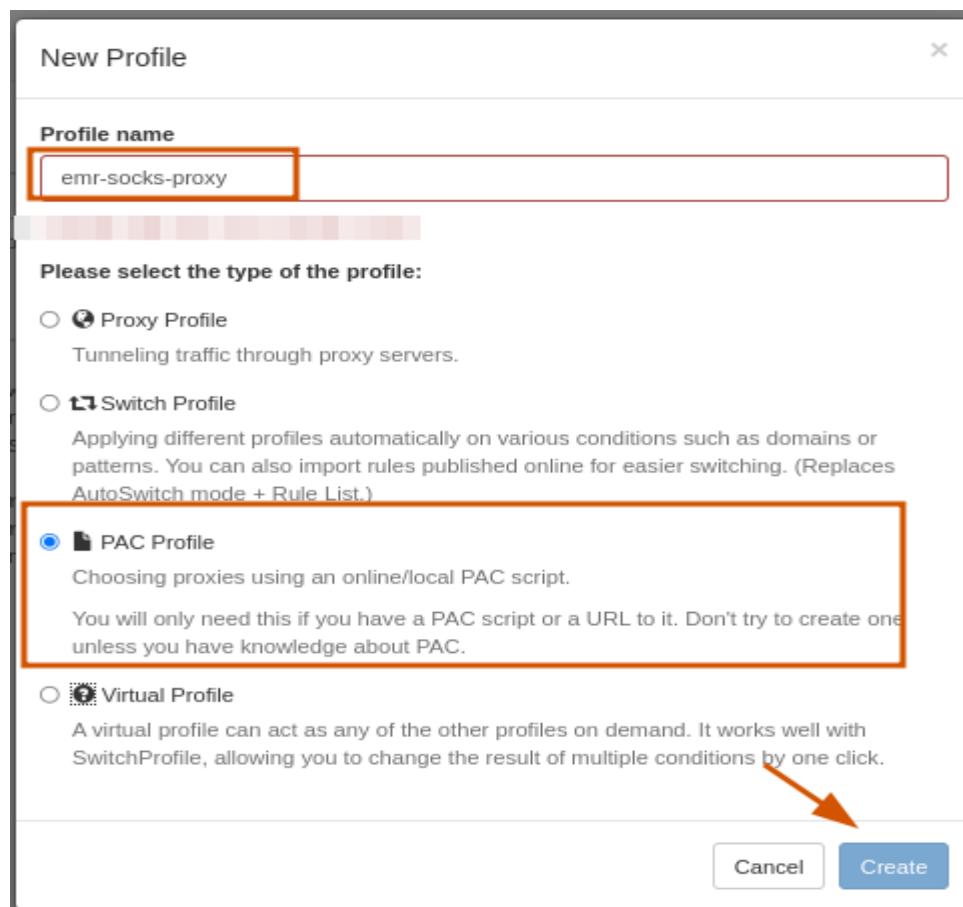
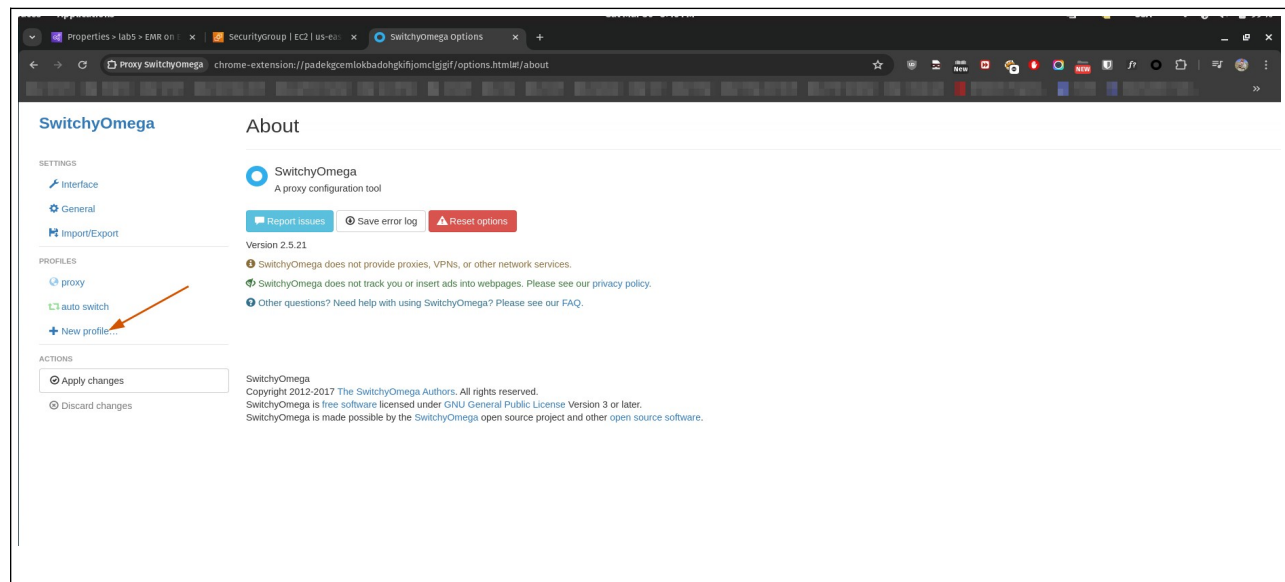
3. Then copy the Hue link and paste it in the chrome

First, we need to enable SSH tunnel in the browser. Navigate to EMR cluster and click on “**Enable an SSH Connection**” in Application user interface:

First, follow the instructions to enable an SSH tunnel to the EMR Master Node:



Then, install **SwitchyOmega** in Chrome. Here is the link to download it in chrome <https://chromewebstore.google.com/detail/proxy-switchyomega/padekgcemlokbadohgkifijomclgjif> If the url provided in the instruction doesn't work, so please manually install it as an extension on you browser. Take Chrome as an example: go to **chrome store**, search for "**SwitchyOmega**", install and add to chrome, **restart** chrome after installing.

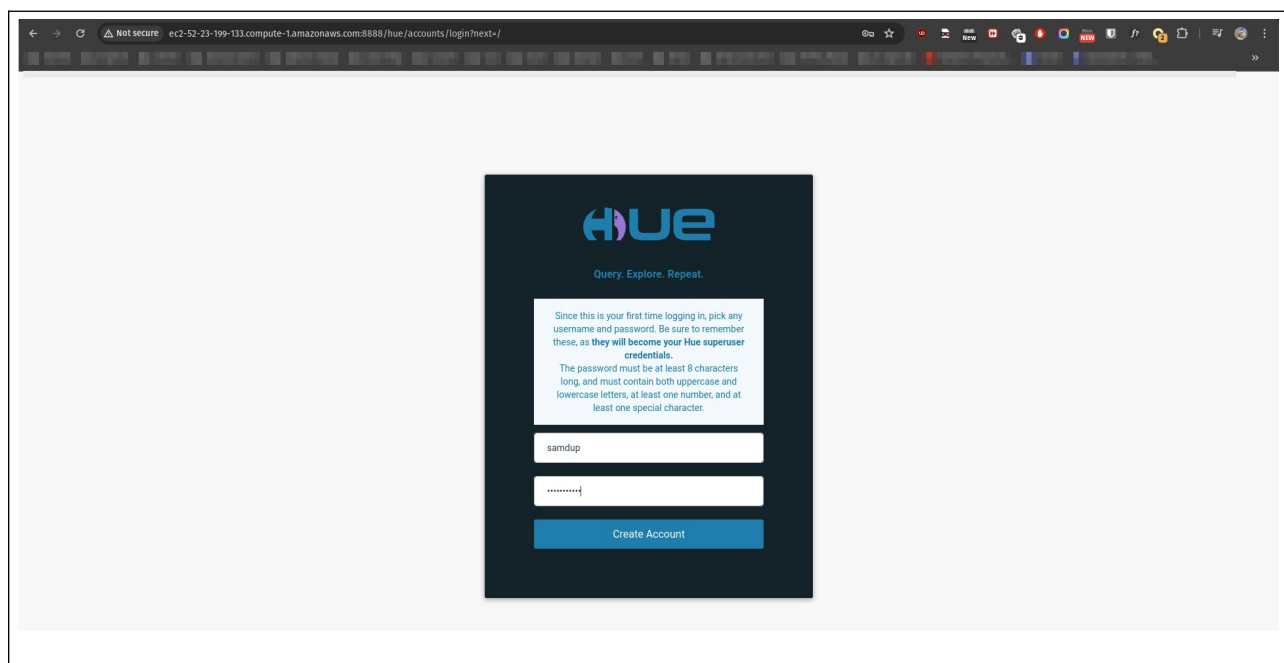


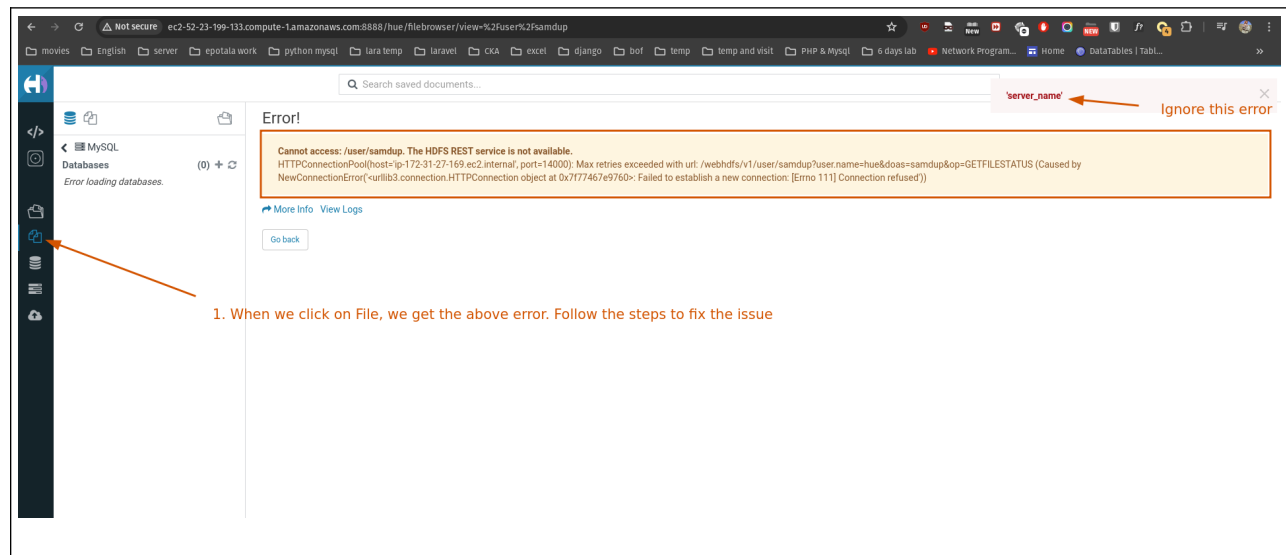


You can copy the script from this link

<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-connect-master-node-proxy.html>

Go back to the EMR cluster and navigate to **Application user interfaces**, copy the url of Hue and open it in the browser in which you just installed the **SwitchyOmega**, create a new account in Hue and **save the username**.





On the hadoop Terminal:

```
sudo chmod 777 /usr/lib/hue/desktop/conf/hue.ini
```

```
vim /usr/lib/hue/desktop/conf/hue.ini
```

**search 14000 and replace this port number with 9870 (to search, you can use /14000 and press enter)**

- How to search in vim (<https://monovm.com/blog/how-to-search-in-vim-editor/>)
- How to save file in vim (<https://phoenixnap.com/kb/how-to-vim-save-quit-exit>)

```

1011
1012 [[default]]
1013 # Enter the filesystem uri
1014 fs_defaultfs = hdfs://ip-172-31-27-169.ec2.internal:8020
1015
1016 # NameNode logical name.
1017 ## logical_name=
1018
1019 # Use WebHdfs/HttpFs as the communication mechanism.
1020 # Domain should be the NameNode or HttpFs host.
1021 # Default port is 14000 for HttpFs.
1022 webhdfs_url = http://ip-172-31-27-169.ec2.internal:9870/webhdfs/v1
1023
1024 # Change this if your HDFS cluster is Kerberos-secured
1025 security_enabled = false
1026
1027 # In secure mode (HTTPS), if SSL certificates from YARN Rest APIs
1028 # have to be verified against certificate authority
1029 ## ssl_cert_ca_verify=True
1030
1031 # Directory of the Hadoop configuration
1032 ## hadoop_conf_dir=$HADOOP_CONF_DIR when set or '/etc/hadoop/conf'
1033
1034 # Configuration for YARN (MR2)
1035 # -----
1036 [[yarn_clusters]]
1037 [[[ip-172-31-27-169.ec2.internal]]]
1038 # Enter the host on which you are running the ResourceManager
1039 resourcemanager_host = ip-172-31-27-169.ec2.internal
1040
1041 # The port where the ResourceManager IPC listens on
1042 # resourcemanager_port=
1043
1044 # Whether to submit jobs to this cluster
1045 submit_to = True
1046
1047 # Resource Manager logical name (required for HA)

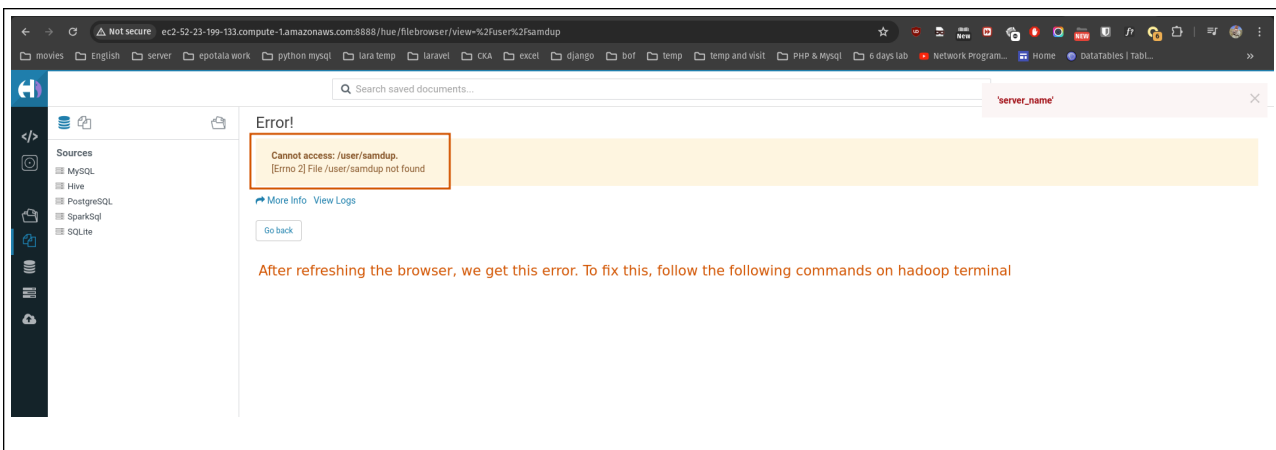
```

**sudo systemctl restart hue**

```

hadoop@ip-172-31-27-169:~ 118x57
[hadoop@ip-172-31-27-169 ~]$ sudo systemctl restart hue
[hadoop@ip-172-31-27-169 ~]$ _

```



Not secure ec2-52-23-199-133.compute-1.amazonaws.com:8888/hue/filebrowser/view-%2Fuser%2Fsamdup

Search saved documents...

server\_name

Sources

- MySQL
- Hive
- PostgreSQL
- SparkSQL
- SQLite

Error!

Cannot access: /user/samdup.  
[Errno 2] File /user/samdup not found

More info View Logs

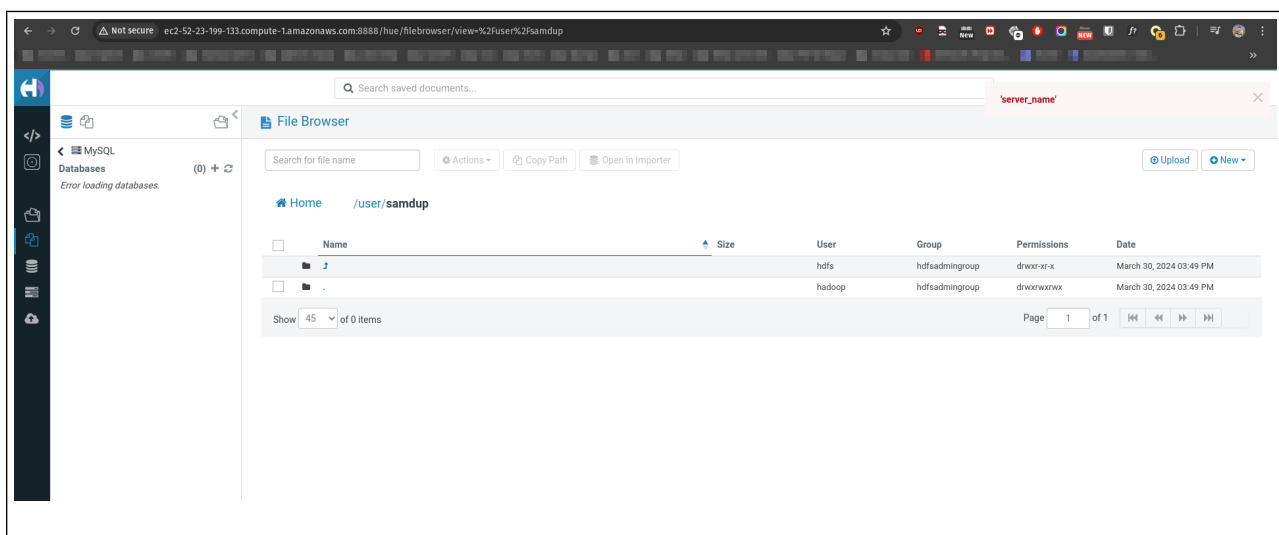
Go back

After refreshing the browser, we get this error. To fix this, follow the following commands on hadoop terminal

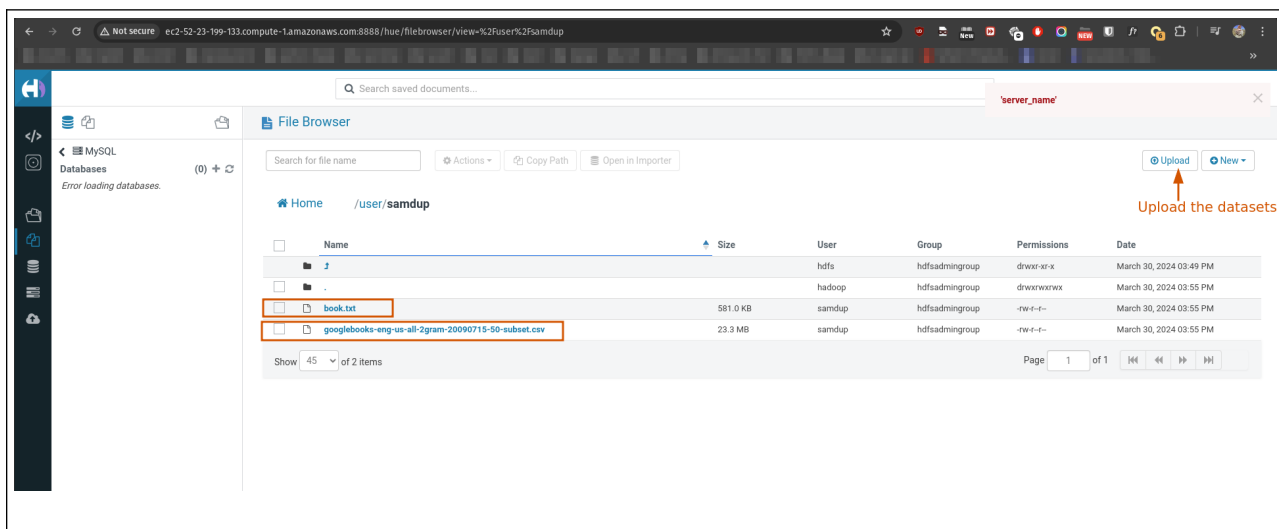


```
hadoop@ip-172-31-27-169:~  
hadoop@ip-172-31-27-169:~ ssh -i lab3-4.pem -ND 8157 hadoop@ec2-52-23-199-133.compute-1.amazonaws.com  
hadoop@ip-172-31-27-169:~ 118x57  
[hadoop@ip-172-31-27-169 ~]$ hdfs dfs -mkdir /user/samdup  
[hadoop@ip-172-31-27-169 ~]$ hdfs dfs -chmod 777 /user/samdup  
[hadoop@ip-172-31-27-169 ~]$
```

Home page of Hue:

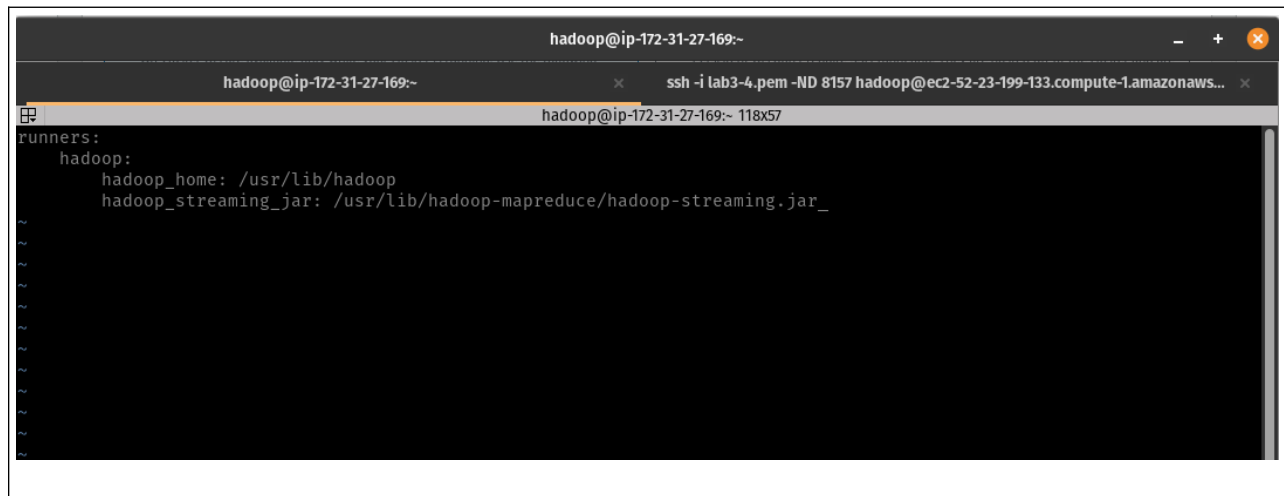


Upload the provided dataset to File Browser in the following structure (don't upload the zip file or folder, use only separate files)



Next, create the mrjob configuration file on the Hadoop master node with vim (or nano) editor using the provided mrjob.conf, save and exit:

### vim mrjob.conf

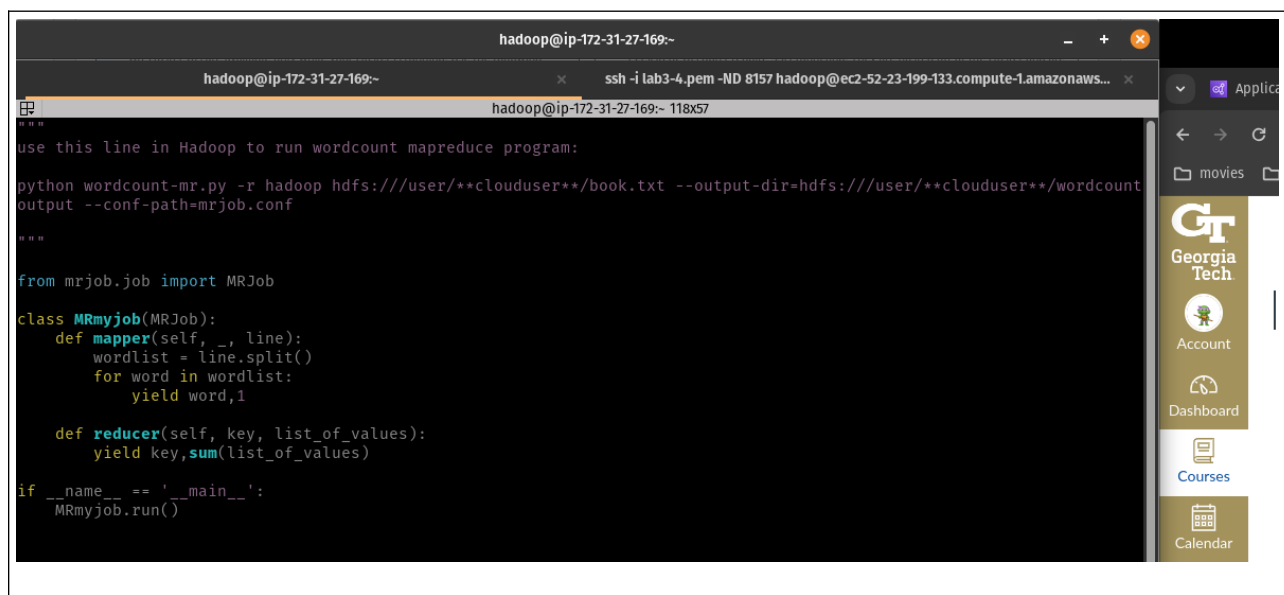


The screenshot shows a terminal window with a dark background. At the top, the title bar reads 'hadoop@ip-172-31-27-169:~'. Below it, the terminal shows the vim editor in normal mode. The file being edited is 'mrjob.conf'. The content of the file is as follows:

```
runners:
  hadoop:
    hadoop_home: /usr/lib/hadoop
    hadoop_streaming_jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar_
~
~
~
~
~
~
~
~
~
~
```

Create wordcount-mr.py on Hadoop master node with vim using the provided wordcount-mr.py, save and exit:

### vim wordcount-mr.py



The screenshot shows a terminal window with a dark background. At the top, the title bar reads 'hadoop@ip-172-31-27-169:~'. Below it, the terminal shows the vim editor in normal mode. The file being edited is 'wordcount-mr.py'. The content of the file is as follows:

```
"""
use this line in Hadoop to run wordcount mapreduce program:
python wordcount-mr.py -r hadoop hdfs:///user/**clouduser**/book.txt --output-dir=hdfs:///user/**clouduser**/wordcount
output --conf-path=mrjob.conf
"""

from mrjob.job import MRJob

class MRmyjob(MRJob):
    def mapper(self, _, line):
        wordlist = line.split()
        for word in wordlist:
            yield word,1

    def reducer(self, key, list_of_values):
        yield key,sum(list_of_values)

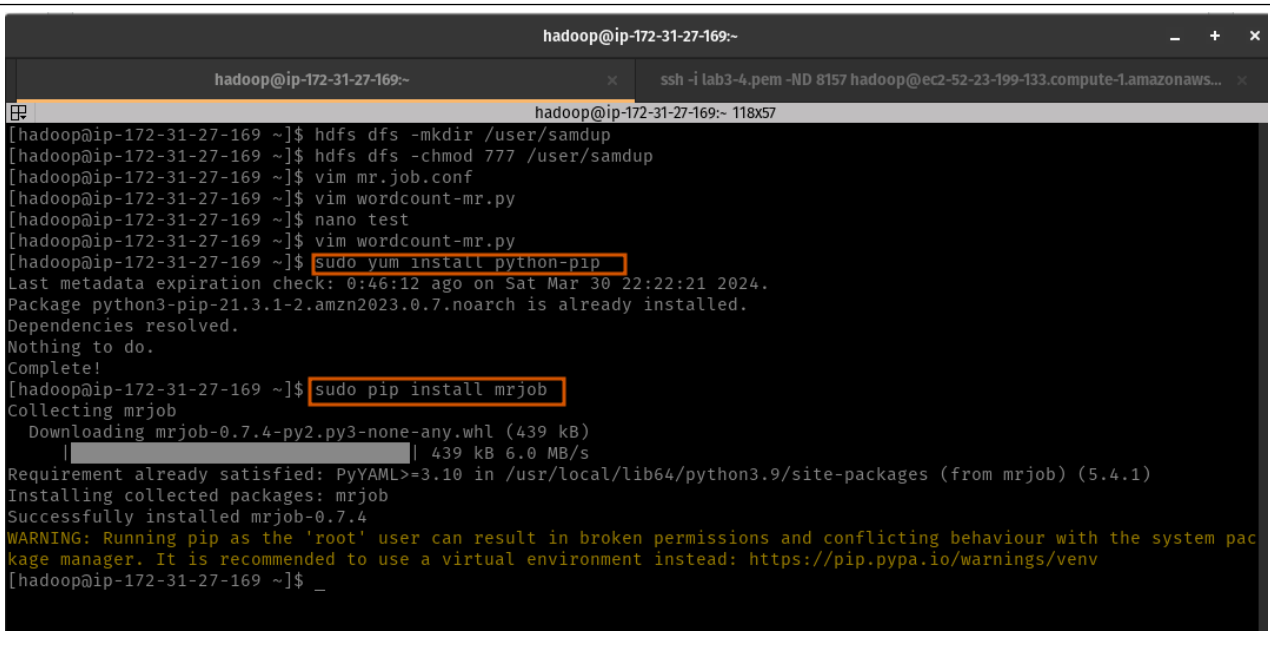
if __name__ == '__main__':
    MRmyjob.run()
```

### 3. Setup mrJob on Master Node

In the Hadoop instance, run the following commands to set up mrJob:

```
sudo yum install python-pip
```

```
sudo pip install mrjob
```

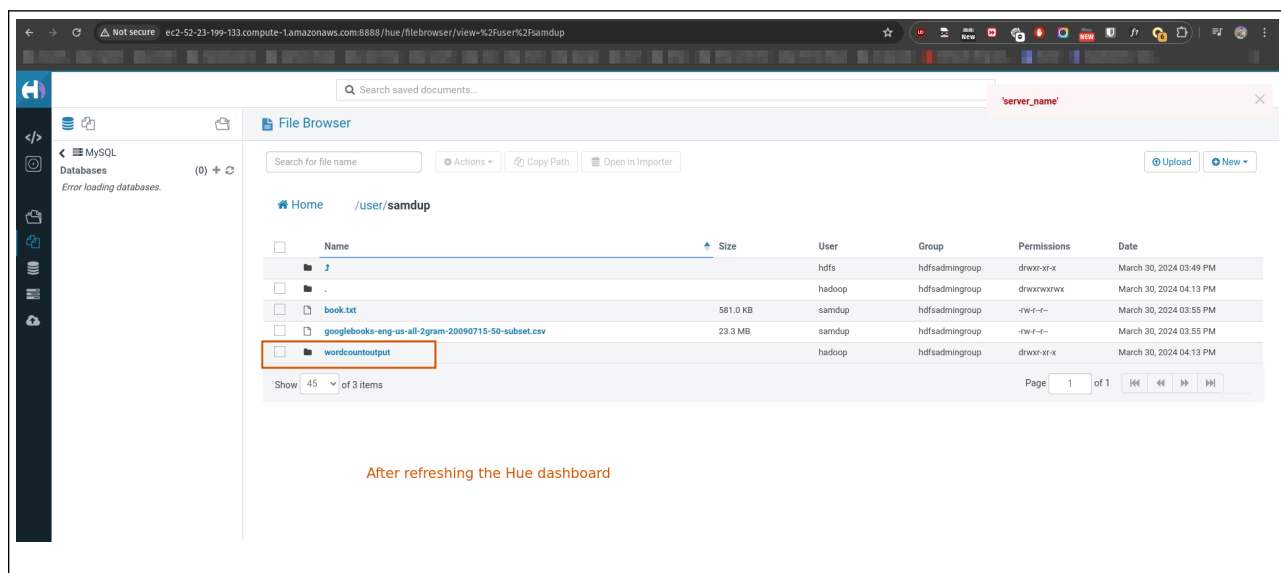


```
hadoop@ip-172-31-27-169:~  
hadoop@ip-172-31-27-169:~  
hadoop@ip-172-31-27-169:~$ hdfs dfs -mkdir /user/samdup  
hadoop@ip-172-31-27-169:~$ hdfs dfs -chmod 777 /user/samdup  
hadoop@ip-172-31-27-169:~$ vim mr.job.conf  
hadoop@ip-172-31-27-169:~$ vim wordcount-mr.py  
hadoop@ip-172-31-27-169:~$ nano test  
hadoop@ip-172-31-27-169:~$ vim wordcount-mr.py  
hadoop@ip-172-31-27-169:~$ sudo yum install python-pip  
Last metadata expiration check: 0:46:12 ago on Sat Mar 30 22:22:21 2024.  
Package python3-pip-21.3.1-2.amzn2023.0.7.noarch is already installed.  
Dependencies resolved.  
Nothing to do.  
Complete!  
hadoop@ip-172-31-27-169:~$ sudo pip install mrjob  
Collecting mrjob  
  Downloading mrjob-0.7.4-py2.py3-none-any.whl (439 kB)  
    |#####| 439 kB 6.0 MB/s  
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib64/python3.9/site-packages (from mrjob) (5.4.1)  
Installing collected packages: mrjob  
Successfully installed mrjob-0.7.4  
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv  
hadoop@ip-172-31-27-169:~$ _
```

### 4. Run MapReduce Program

In the Hadoop instance, paste the command line from wordcount-mr.py and change the clouduser to your username. After the program finishes execution, you can view the results in the file browser:

```
python wordcount-mr.py -r hadoop hdfs:///user/samdup/book.txt --output-dir=hdfs:///user/samdup/wordcountoutput --conf-path=mrjob.conf
```



The screenshot shows the Hue File Browser interface. The left sidebar contains icons for MySQL Databases, File Browser, and other tools. The main panel displays the file browser for the path `/user/samdup/wordcountoutput/part-00000`. The file list shows a single file `part-00000` with a size of 186.02 KB. The file content is displayed in a table format, showing word counts for various words. A message in the center of the file content area states: "It ran successfully and you can see the number of occurrence".

Word	Count
"#1661]"	1
"\$5,000)"	1
"\$"	5
"'60's"	1
"'77,"	1
"'82"	1
"'83"	1
"'83,"	1
"'84"	1
"'84,"	1
"'85,"	2
"'85,"	1
"'87"	1
"'89,"	1
"'89--there"	1
"'90,"	1
"'A"	3
"'A5-15'"	1
"'American"	1
"'and"	1
"'Aras:"	1
"'But"	1
"'By"	2
"'Co."	1
"'Colonel"	1
"'company,"	1
"'Coolest"	4
"'Coolest"	2
"'Drive"	1
"'Eg."	1
"'FictionTosadita"	1

## 5. Run Pig Program (2 ways to approach this)

### a) Edit the wordcount-pig.txt (change to your username)

```
wordcount-pig.txt
~/Desktop/labs/examples

1--In Local mode with input file on Hadoop instance:
2 a = LOAD 'file:///home/hadoop/book.txt' as (lines:chararray);
3
4--In MapReduce mode with input file on HDFS:
5 a = LOAD '/user/samdup/book.txt' as (lines:chararray);
6
7
8 b = FOREACH a GENERATE FLATTEN(TOKENIZE(lines)) as word;
9 c = GROUP b by word;
10 d = FOREACH c GENERATE group, COUNT(b);
11 store d into '/user/samdup/pig_wordcount_aws';
12
13
```

## b) Goto S3 bucket and create a folder 'project' and upload the wordcount-pig.txt

Upload succeeded  
View details below.

Amazon S3 > Buckets > aws-logs-211125774779-us-east-1 > elasticmapreduce\_sam/

I included '\_sam' when we were creating the EMR initially; however, it is not necessary.

elasticmapreduce\_sam/

Objects (4) info

Find objects by prefix

Name	Type	Last modified	Size	Storage class
27560E27PPD35/	Folder	-	-	-
2TOGM2LNF85M2/	Folder	-	-	-
3IO2ZF2NL7BDJ/	Folder	-	-	-
program/	Folder	-	-	-

After creating the folder, upload the file.

wordcount-pig.txt info

Copy this URI

S3 URI  
s3://aws-logs-211125774779-us-east-1/elasticmapreduce\_sam/program/wordcount-pig.txt

Amazon Resource Name (ARN)  
arn:aws:s3::aws-logs-211125774779-us-east-1/elasticmapreduce\_sam/program/wordcount-pig.txt

Entity tag (ETag)  
51b733247c141f41be0687f402dc8219

Object URL  
https://aws-logs-211125774779-us-east-1.s3.amazonaws.com/elasticmapreduce\_sam/program/wordcount-pig.txt

## Approach 1:

Goto **EMR** and Click on **Steps**

The screenshot displays the AWS Management Console interface for an Amazon EMR cluster named 'lab5'. The breadcrumb navigation at the top indicates the path: Amazon EMR > EMR on EC2: Clusters > lab5. The cluster's status is 'Waiting', and it was created on March 30, 2024, at 18:20 UTC-04:00. The 'Steps' tab is selected in the navigation bar, showing a single step: 'Archive log files to Amazon S3', which is 'Turned on'. The step's configuration includes an Amazon S3 location and encryption for logs turned off. The 'Cluster termination and node replacement' section shows the termination option set to 'Automatically terminate cluster after idle time', with an idle time of 1 hour, and both termination protection and unhealthy node replacement are turned off.

**lab5** Updated less than a minute ago

**Summary**

Cluster info	Applications	Cluster management	Status and time
<b>Cluster ID</b> J-2TOGM2LNF85M2	<b>Amazon EMR version</b> emr-7.0.0	<b>Log destination in Amazon S3</b> <a href="#">aws-logs-211125774779-us-east-1/elasticmapreduce_sam</a>	<b>Status</b> Waiting
<b>Cluster configuration</b> <b>Instance groups</b>	<b>Installed applications</b> Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, Pig 0.17.0, Tez 0.10.2	<b>Persistent application UIs</b> <a href="#">YARN timeline server</a> <a href="#">Tez UI</a>	<b>Creation time</b> March 30, 2024, 18:20 (UTC-04:00)
<b>Capacity</b> 1 Primary   1 Core   0 Task		<b>Primary node public DNS</b> ec2-52-23-199-133.compute-1.amazonaws.com <a href="#">Connect to the Primary node using SSH</a> <a href="#">Connect to the Primary node using SSM</a>	<b>Elapsed time</b> 1 hour, 16 minutes

**Steps**

Operating system	Cluster logs	Cluster termination and node replacement
<b>Amazon Linux release</b> 2023.3.20240312.0	<b>Archive log files to Amazon S3</b> Turned on  <b>Amazon S3 location</b> <a href="#">s3://aws-logs-211125774779-us-east-1/elasticmapreduce_sam/</a>  <b>Encryption for logs</b> Turned off	<b>Termination option</b> Automatically terminate cluster after idle time  <b>Idle time</b> 1 hour  <b>Termination protection</b> Off  <b>Unhealthy node replacement</b> Off

aws

Services

Search

[Alt+S]

N. Virginia

sam.cloud

Amazon EMR > EMR on EC2: Clusters > lab5

lab5

Updated less than a minute ago

Terminate

Clone in AWS CLI

Clone

1. Click here

Summary

Properties Bootstrap actions Instances (Hardware) Steps Applications Configurations Monitoring Events

Steps (0) Info

Refresh table

Cancel steps

Clone step

Add step

Concurrent steps: 1

Filter steps by status

Find steps

1

2. Click here

Step ID	Status	Name	Log files	Create
No matches				
We can't find a match				

aws

Services

Search

[Alt+S]

N. Virginia

sam.cloud

Step s-062370038JBD4EDONAXM has been successfully added.

Notifications 0 0 2 0 0

Amazon EMR > EMR on EC2: Clusters > lab5 > Add step

## Add step Info

### Step settings

Type

☐ Custom JAR  
Adds a step that enables you to write a custom script to process your data using the Java programming language.

☐ Streaming program  
Adds a step that uses standard input to run mapper/reducer scripts and send results to standard output.

☐ Hive program  
Adds a step that submits a Hive script for data warehouse interactions.

☒ Pig program  
Adds a step that submits a Pig script for analyzing very large data sets.

☐ Shell script  
Troubleshoot your cluster.

Name

wordcount

Pig script location

Amazon S3 location of your Pig script

s3://aws-logs-211125774779-us-east-1/elasticmapreduce\_sam/prog

View

Browse S3

Input Amazon S3 location - optional

Amazon S3 location of your Pig input files.

s3://bucket/prefix/object

View

Browse S3

Output Amazon S3 location - optional

Amazon S3 location of your Pig output files.

s3://bucket/prefix/object

View

Browse S3

Arguments - optional Info

Specify optional arguments for your script.

bash -c "aws s3 cp s3://DOC-EXAMPLE-BUCKET/my-script.sh."

### Step action

Action if step fails

The action to take when the step fails.

☒ Continue  
Continues to the next step in the queue.

☐ Cancel and wait  
Cancels any pending steps and returns the cluster to the waiting state.

☐ Terminate cluster  
Shuts down the cluster.

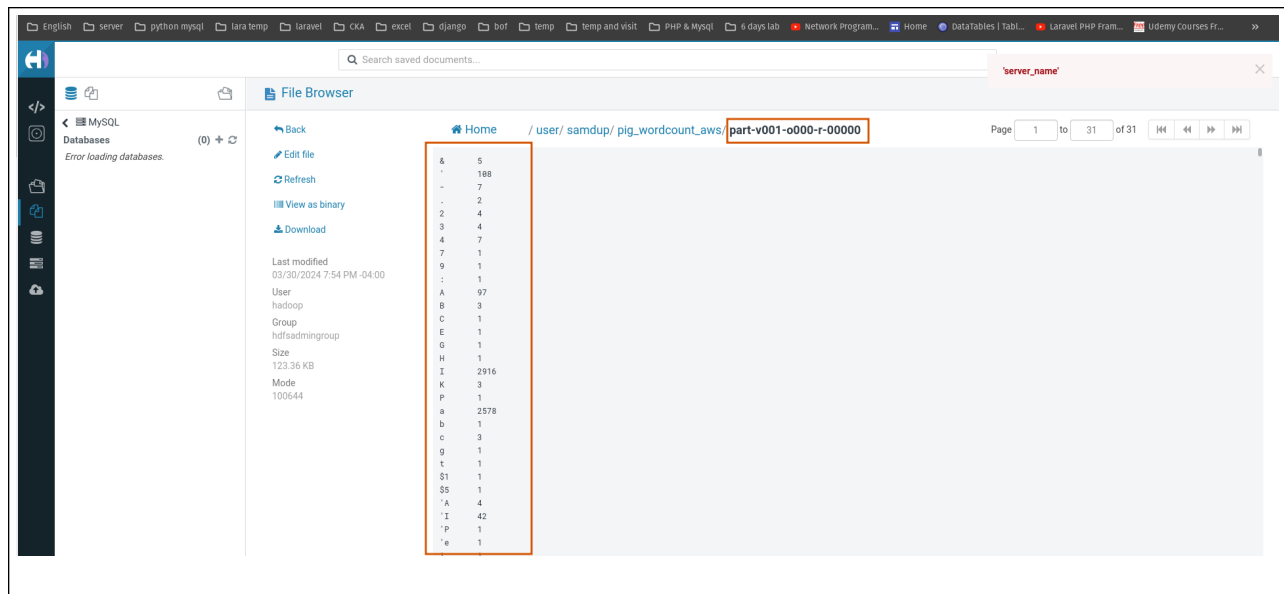
Cancel

Add step

Paste the S3 URI which we copied earlier



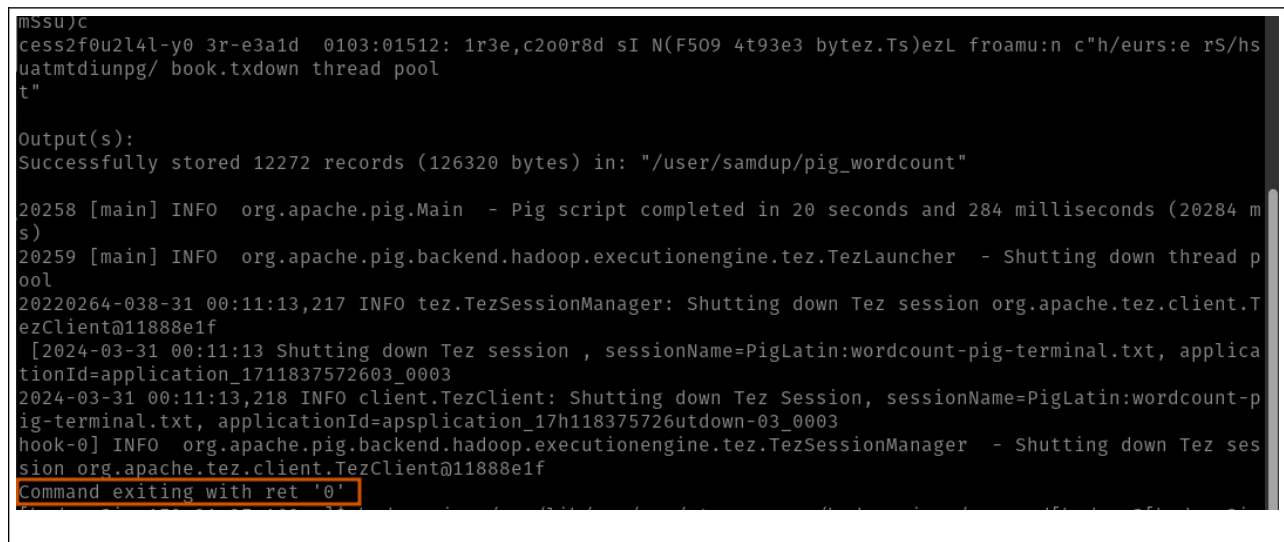
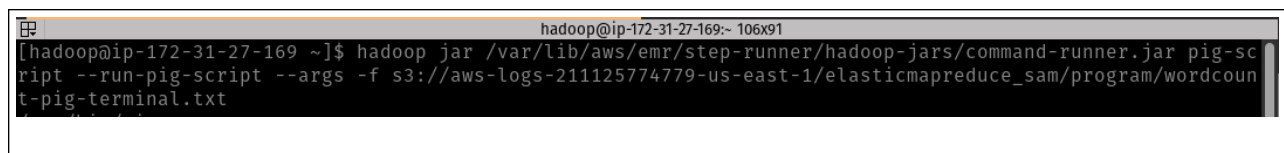




## Approach 2:

Using terminal / On your hadoop Terminal:

```
hadoop jar /var/lib/aws/emr/step-runner/hadoop-jars/command-runner.jar pig-script --run-pig-script --args -f s3://aws-logs-211125774779-us-east-1/elasticmapreduce_sam/program/wordcount-pig-terminal.txt
```



Search saved documents...

File Browser

Search for file name

Actions Copy Path Open in Importer

Upload New

Home /user/samdup

Name	Size	User	Group	Permissions	Date
.		hdfs	hdfsadmin	drwxr-xr-x	March 30, 2024 03:49 PM
..		hadoop	hdfsadmin	drwxrwxrwx	March 30, 2024 05:11 PM
book.txt	581.0 KB	samdup	hdfsadmin	-rw-r--r--	March 30, 2024 03:55 PM
googlebooks-eng-us-all-2gram-20090715-50-subset.csv	23.3 MB	samdup	hdfsadmin	-rw-r--r--	March 30, 2024 05:11 PM
pig_wordcount		hadoop	hdfsadmin	drwxr-xr-x	March 30, 2024 04:54 PM
pig_wordcount_aws		hadoop	hdfsadmin	drwxr-xr-x	March 30, 2024 04:13 PM
wordcountoutput		hadoop	hdfsadmin	drwxr-xr-x	March 30, 2024 04:13 PM

Show 45 of 5 items

Page 1 of 1

Result Generated using Approach 2 or Terminal

Result Generated using Approach 1 or aws console

Search saved documents...

File Browser

Search for file name

Actions Copy Path Open in Importer

Upload New

Home /user/samdup/pig\_wordcount/part-v001-o000-r-00000

Back

Edit file

Refresh

View as binary

Download

Last modified: 03/30/2024 8:11 PM -04:00

User: hadoop

Group: hdfsadmin

Size: 123.36 KB

Mode: 100644

```

& 5
' 198
- 7
. 2
2 4
3 4
4 7
7 1
9 1
: 1
A 97
B 3
C 1
E 1
G 1
H 1
I 2916
K 3
P 1
a 2578
b 1
c 3
g 1
t 1
$1 1
$5 1
'A 4
'I 42
'P 1
'e 1
1 1
10 1

```

Page 1 to 31 of 31

**Note: Regardless of whether you follow Approach 1 or Approach 2, the output will be the same.**

## 6. Challenges (75%)

1. Implement a MapReduce program that emits the bigrams which were coined after year 1992 (or which started appearing after the year 1992).  
Output of the program should include: (bigram, year)  
**Example output:** (mobile phone, 1996) means that the bigram 'mobile phone' first appeared in the dataset in the year 1996.
2. Implement a MapReduce program that emits the average number of times each bigram appears in a book (over all the years). [Average for a particular n-gram is the total count of n-gram (over all the years) divided by the total number of books in which the n-gram appeared (over all the years)]  
Output of the program should include: (bigram, average)  
**Example output:** (how are, 6) means that the bigram 'how are' appears on average 6 times in a book (over all the years).
3. Implement a Pig program that computes the most common bigram in the year 2003 in the dataset (as determined by the count field). Output of the program should include: (bigram, count)  
**Example output:** (how many, 5001) means that the bigram 'how many' was the most popular bigram in the year 2002 and it appeared a total of 5001 times in all the books in that year.
4. Implement a Pig program that computes the most common bigram in each year in the dataset (as determined by the count field). Output of the program should include: (year, bigram, count)  
**Example output:** (2003, mobile phone, 3012) means that in the year 2003 the most popular bigram was 'mobile phone' and it appeared 3012 times in all the books in that year. Emit such tuples for each year in the dataset.
5. Create a Hive meta-store table from the N-Gram dataset (CSV) file from the Hue web interface. Implement a Hive query (in the SQL-like Hive Query Language) to find the most popular bigram (over all the years).

## Deliverables

1. The complete code with the modifications needed to complete each exercise, including the new lambda function.
2. Output files (.txt or .csv) that contain results for each exercise program.

**Troubleshooting Section:**

1. To replicate Dr. Joel's EMR setup as demonstrated in the video lecture ([https://gatech.instructure.com/courses/377392/discussion\\_topics/1770987](https://gatech.instructure.com/courses/377392/discussion_topics/1770987)), follow strictly to his instructions. However, when configuring the EMR resource, omit the selection of any security group, leaving it unassigned. Proceed with the creation process as instructed.
2. Upon completion, terminate this EMR instance.
3. Subsequently, recreate the EMR resource, ensuring that all options are configured as per Dr. Joel's guidance. Once the EMR instance is ready, navigate to the security group settings and add the SSH rule to the inbound rule configuration.