**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. Madisetti, ÒCloud Computing Solutions Architect: A Hands-On ApproachÓ, 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

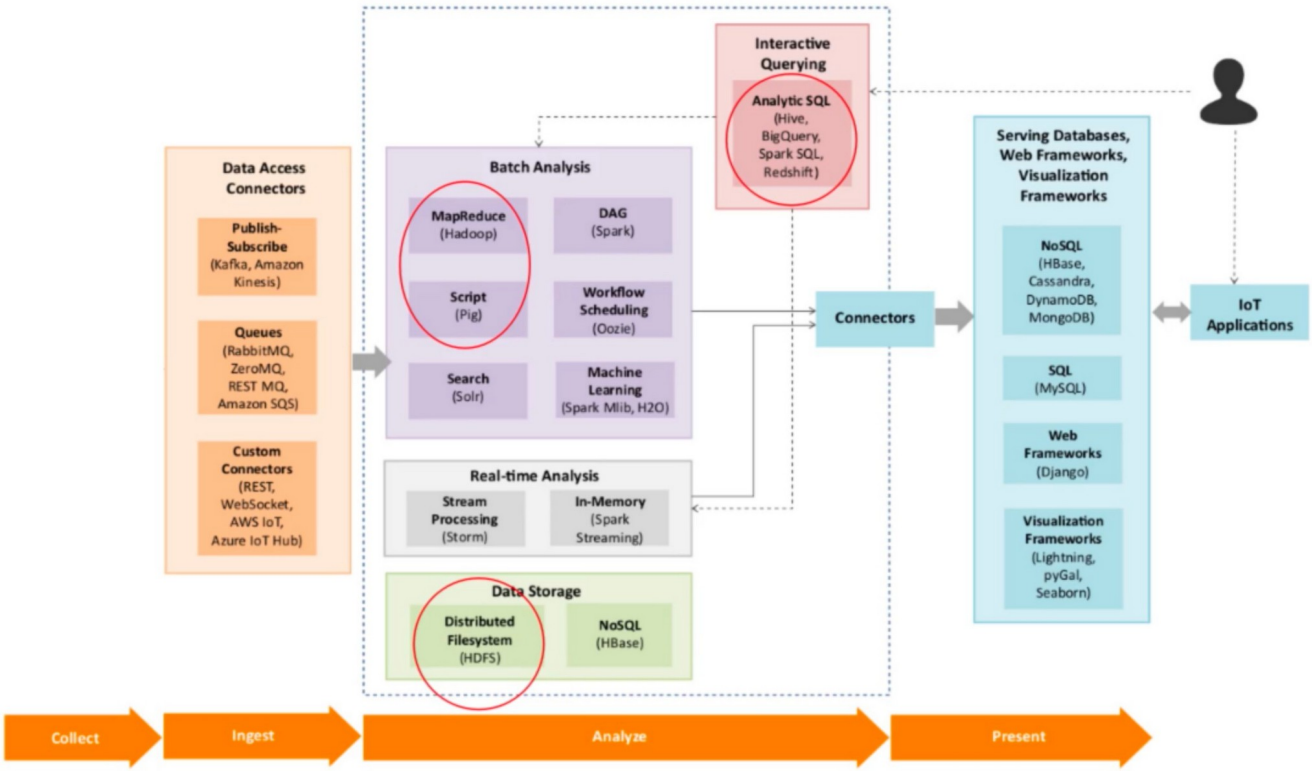

## Due Date:

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

---

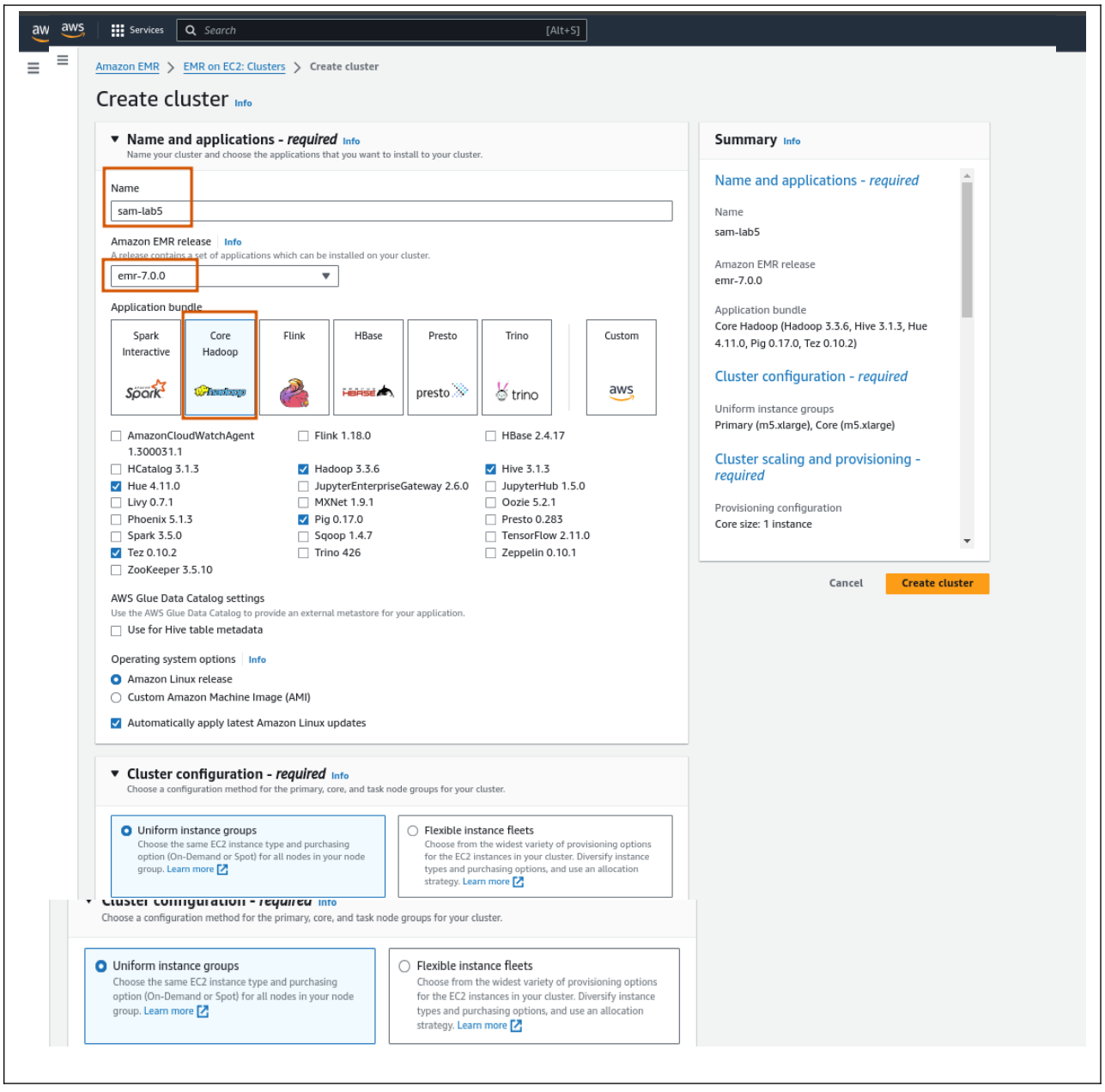
In this lab you will learn how setup a Hadoop cluster and run MapReduce, Pig and Hive job.

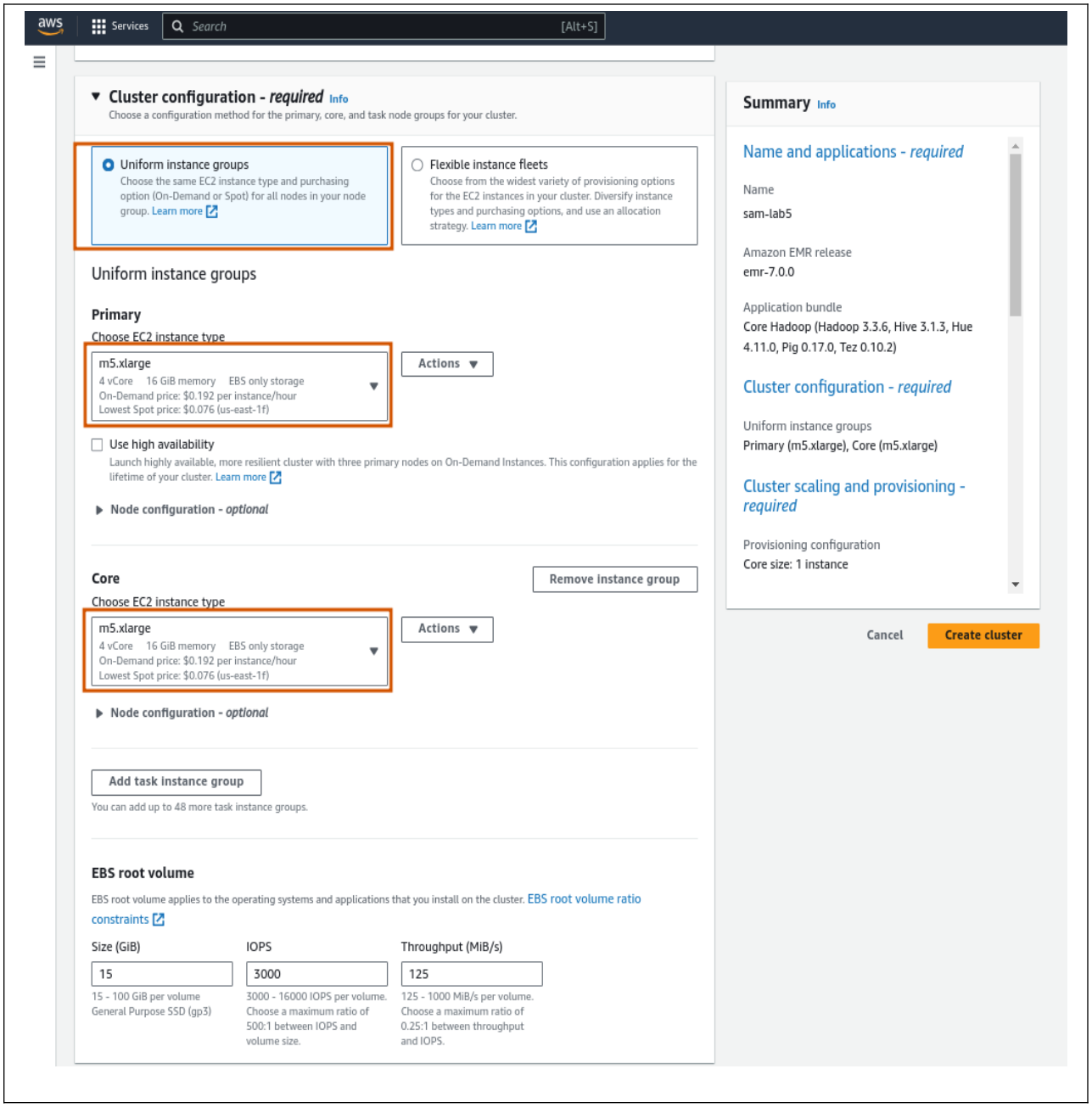
Fig.1 Architecture diagram of data processing in Hadoop



## 1. Set up a Hadoop Cluster with EMR

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




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
Remove instance group
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 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 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

## 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***

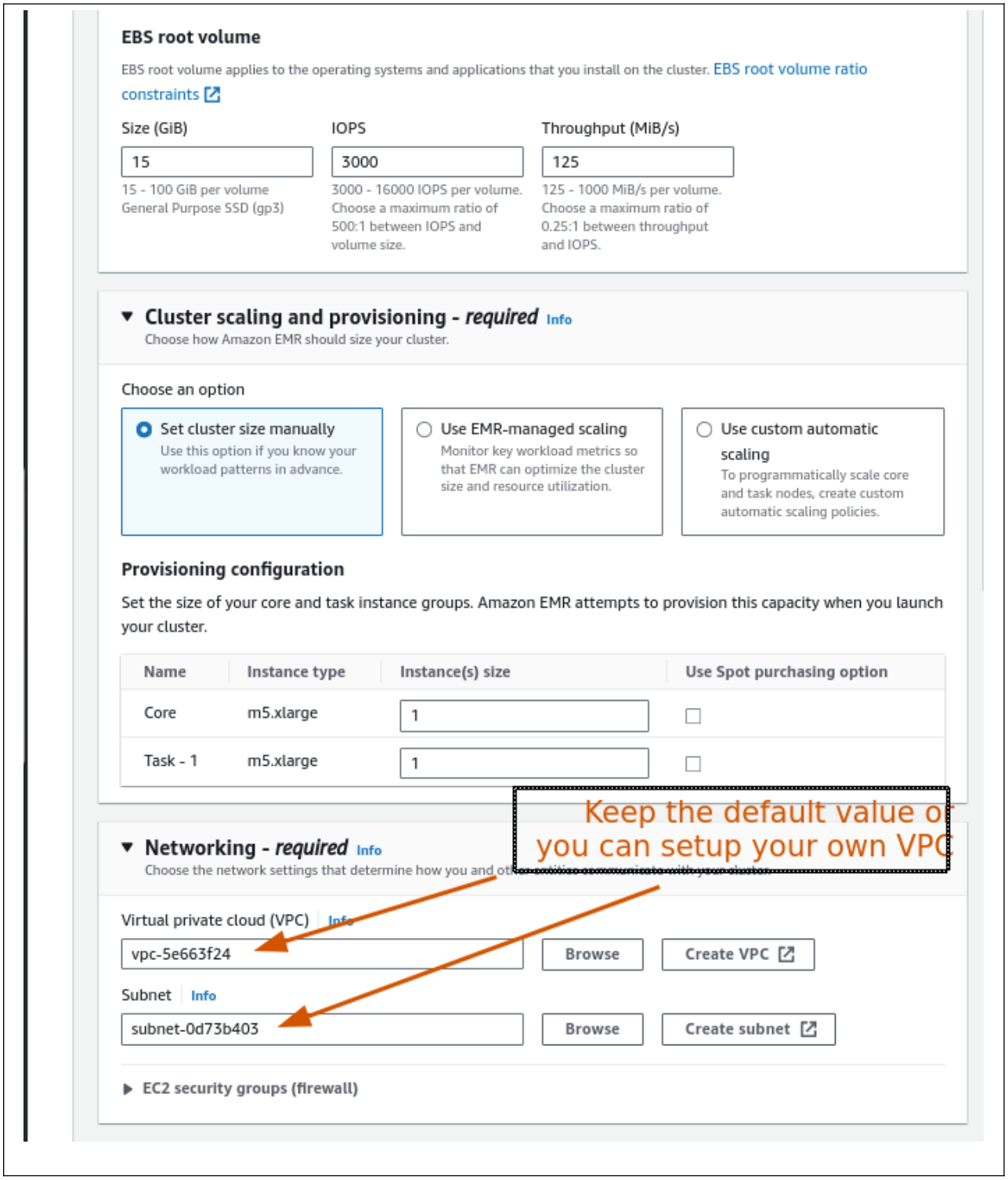
Add task instance group

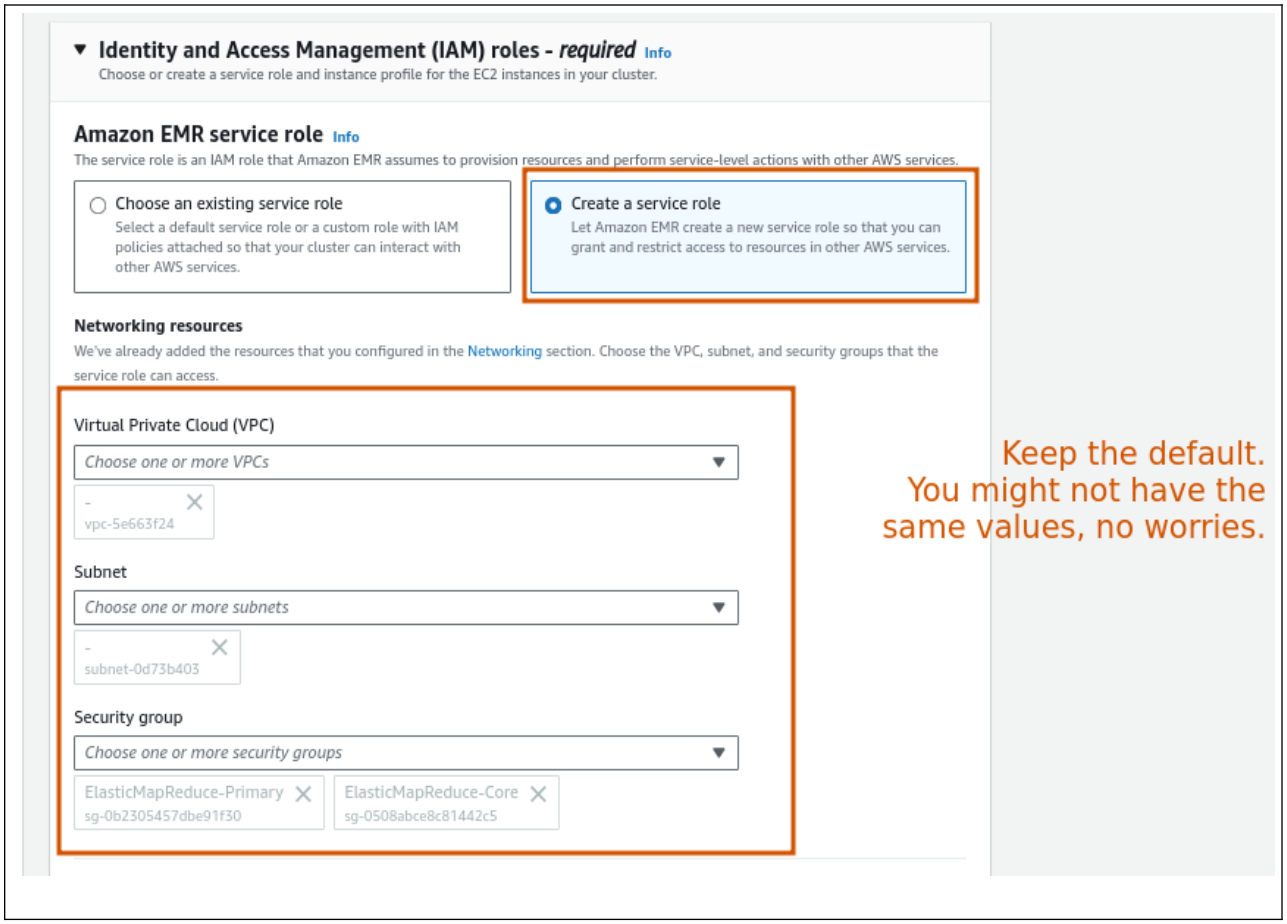
You can add up to 47 more task instance groups.

### EBS root volume

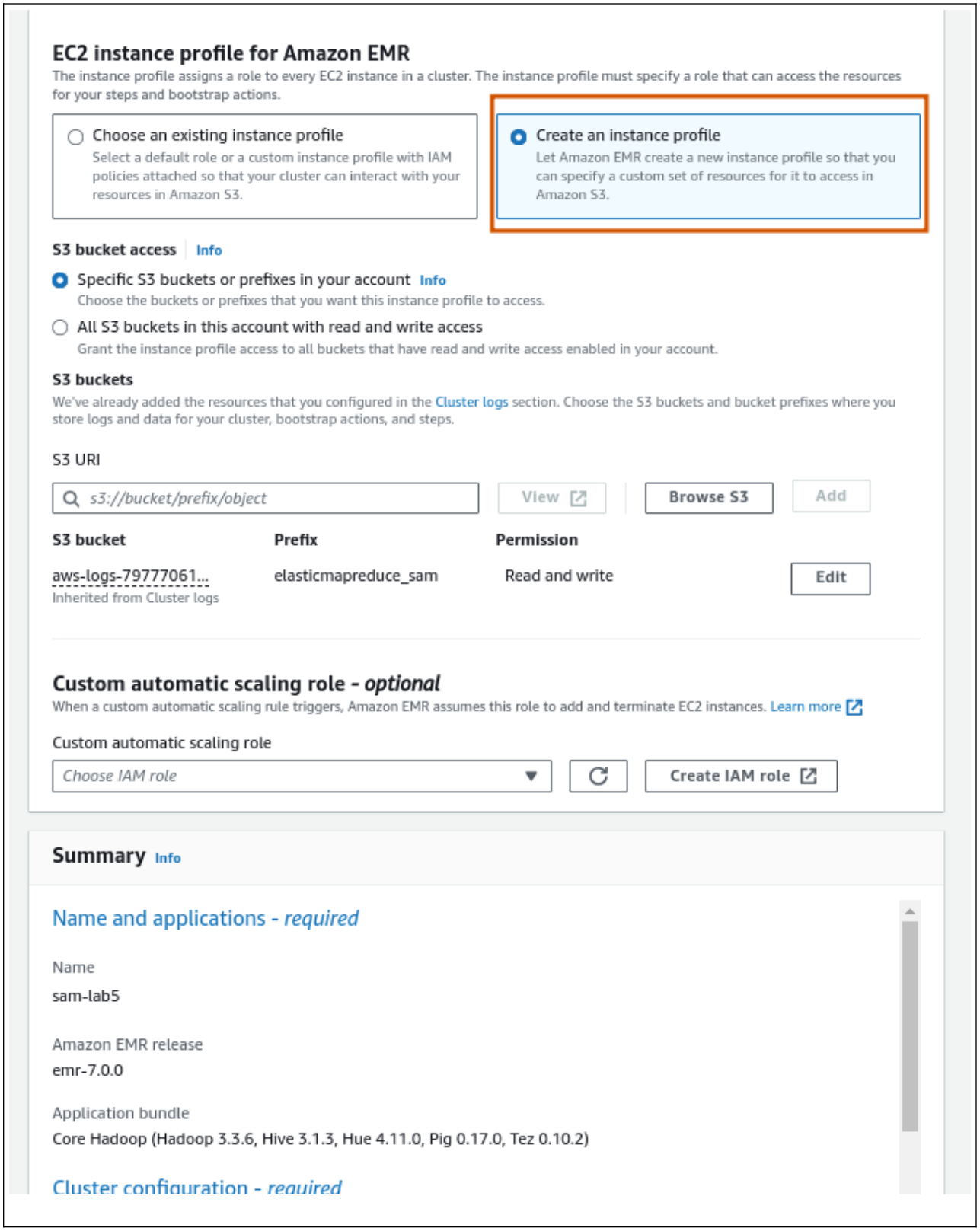
EBS root volume applies to the operating systems and applications that you install on the cluster. EBS root volume ratio constraints

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
Task - 1
m5.xlarge
1
Keep the default value o
you can setup your own VP
Networking - required Info
Virtual private cloud (VPC) Info
vpc-5e663f24
Browse
Create VPC
Subnet Info
subnet-0d73b403
Browse
Create subnet
EC2 security groups (firewall)



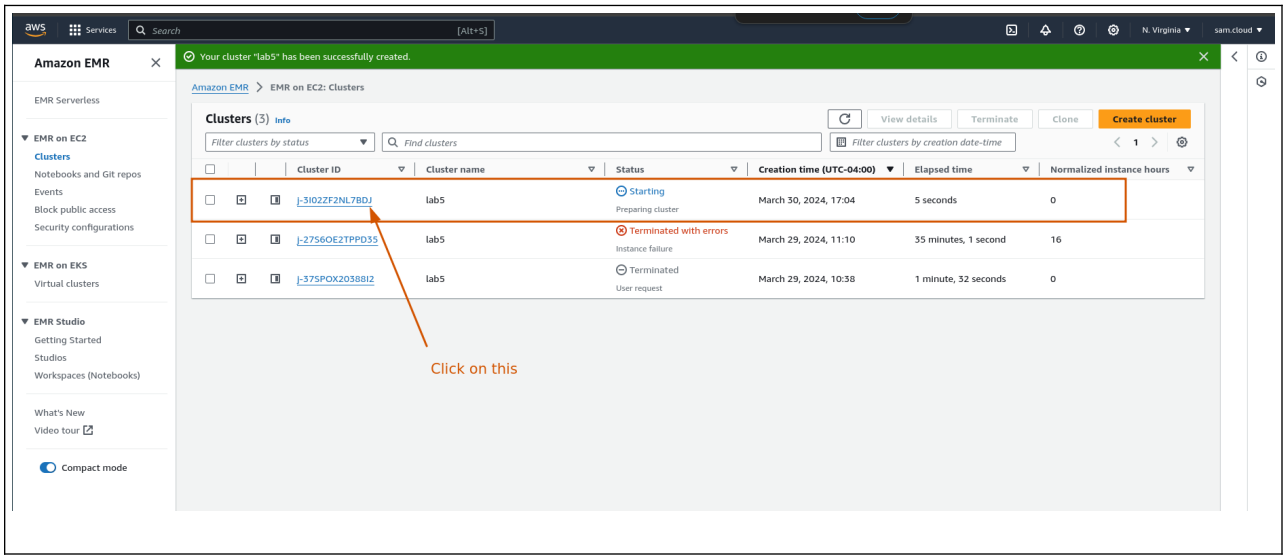

Please refer to the endnote on how to quickly and informally set up your **security group** (this method is not recommended, but you can experiment with it).



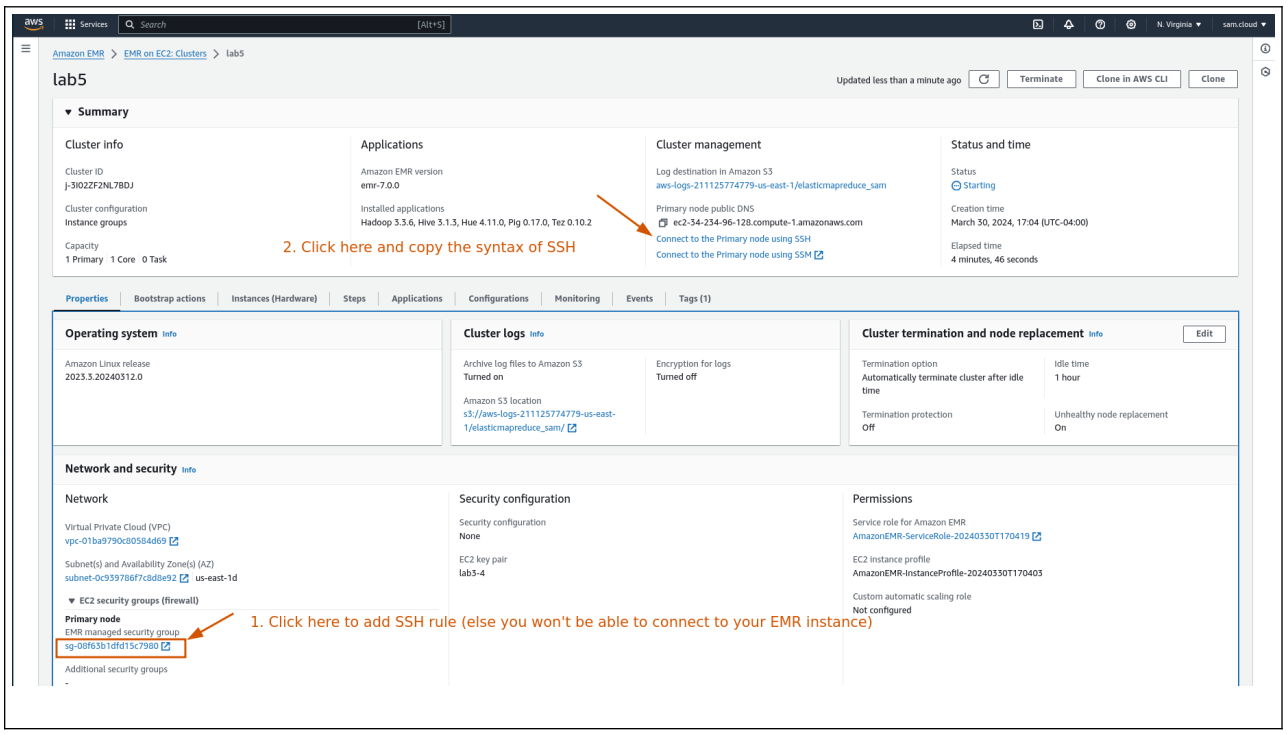

Click on **Create**

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

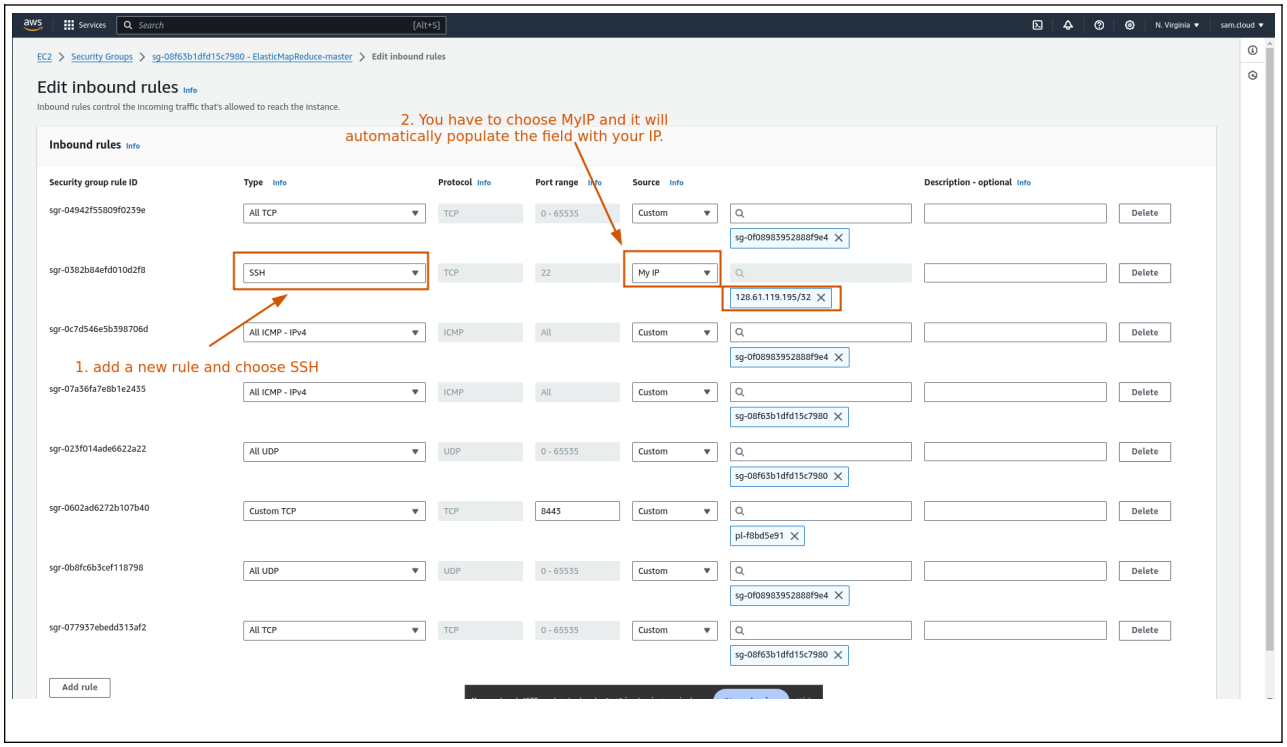
Navigate to **EC2 instance** and open the one that's running, which holds the cluster that you just created:



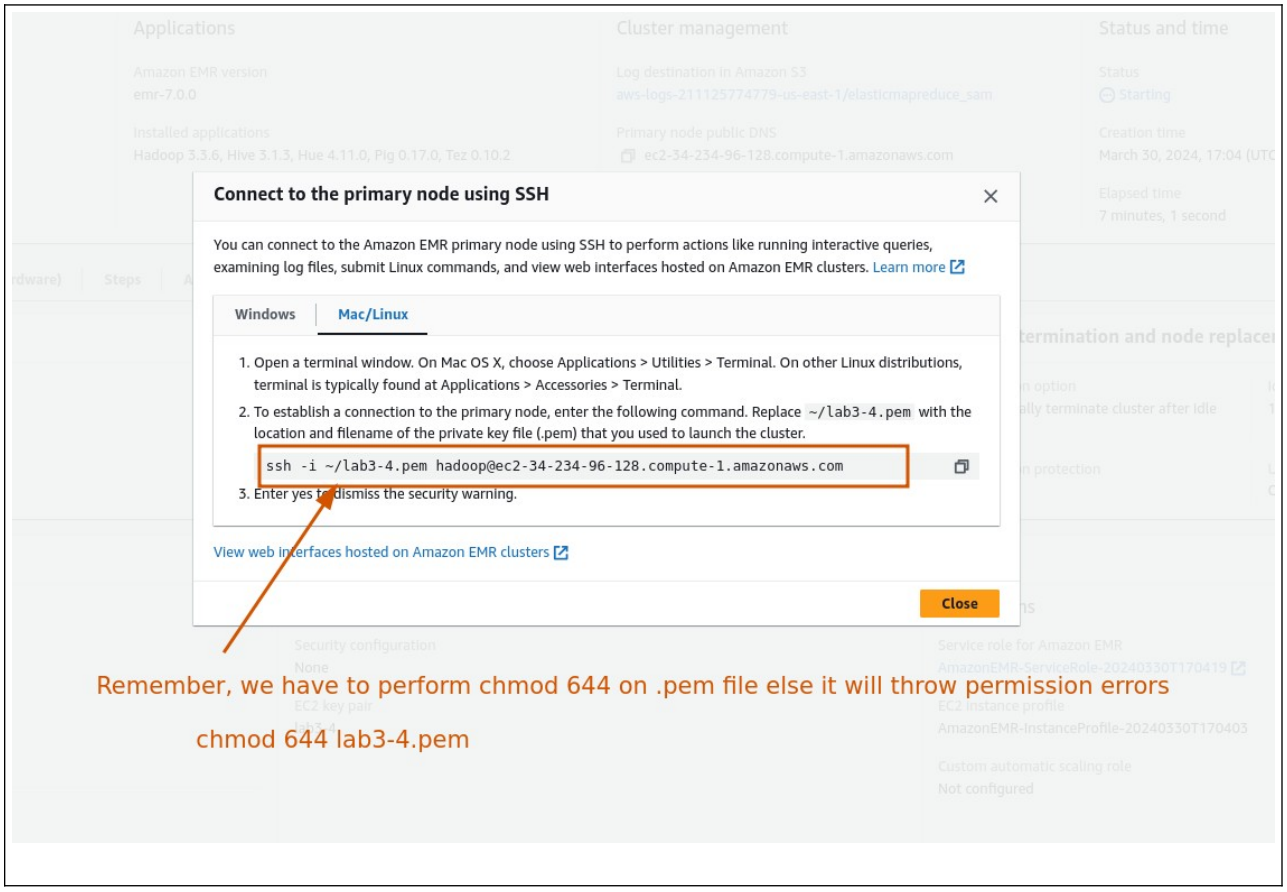

Go to **Security-Security groups** and open the security group for the master cluster, in this case **ElasticMapReduce-master**:



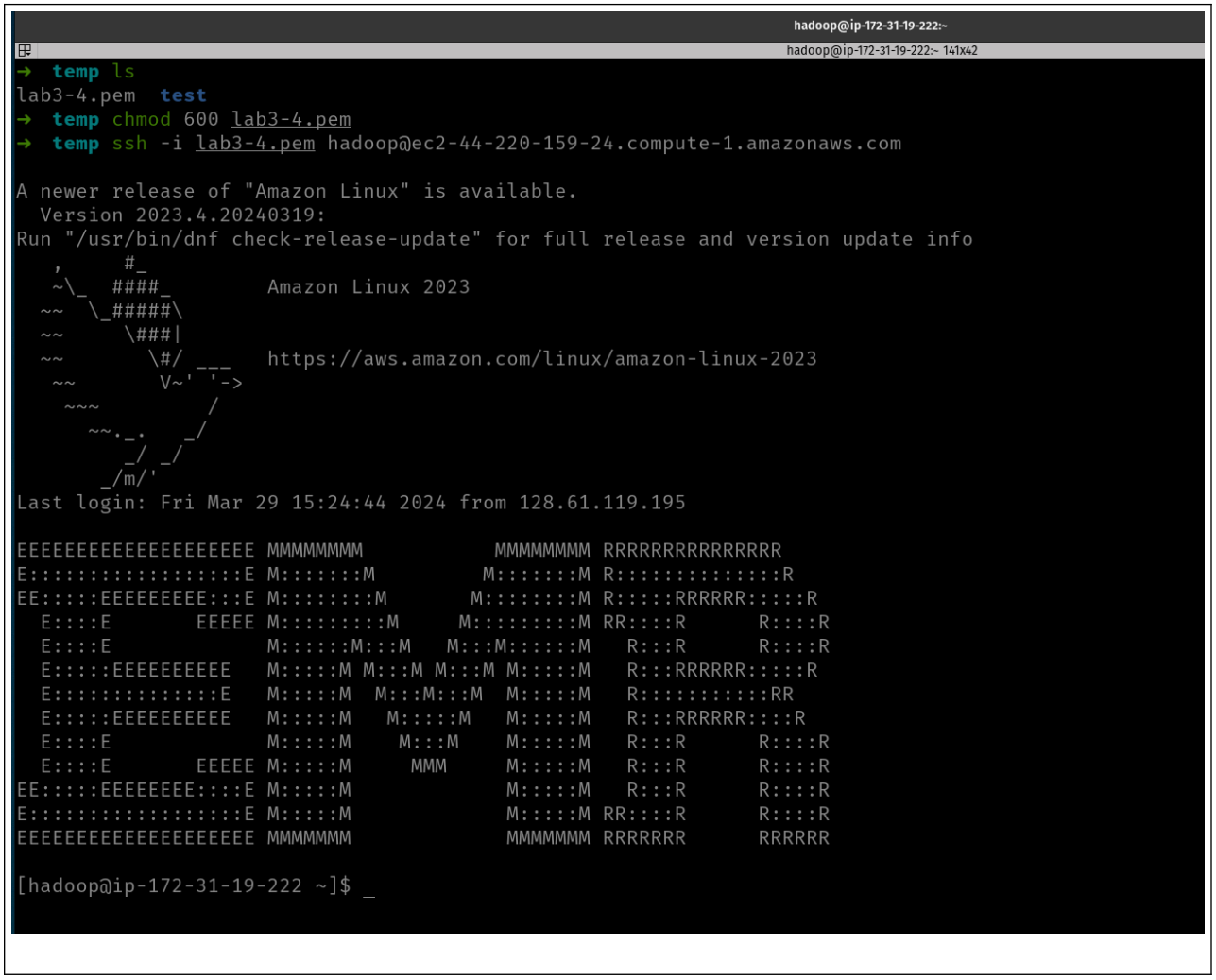

Edit the inbound rules and add a rule for SSH and save it:




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:

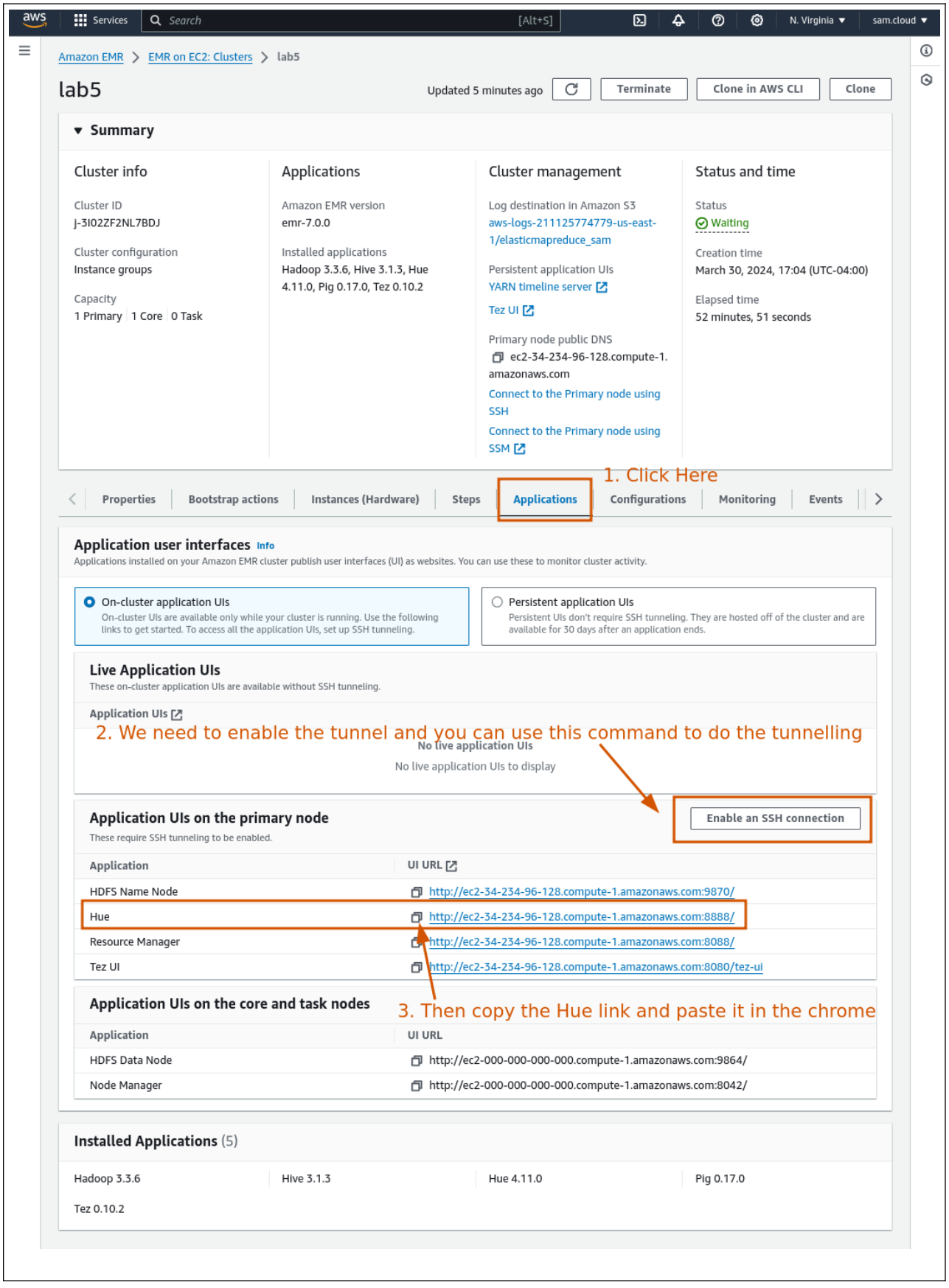


For example, a successful connection would appear like:  (You can use either 644 or 600 for the permission. Usually 600 is recommended)



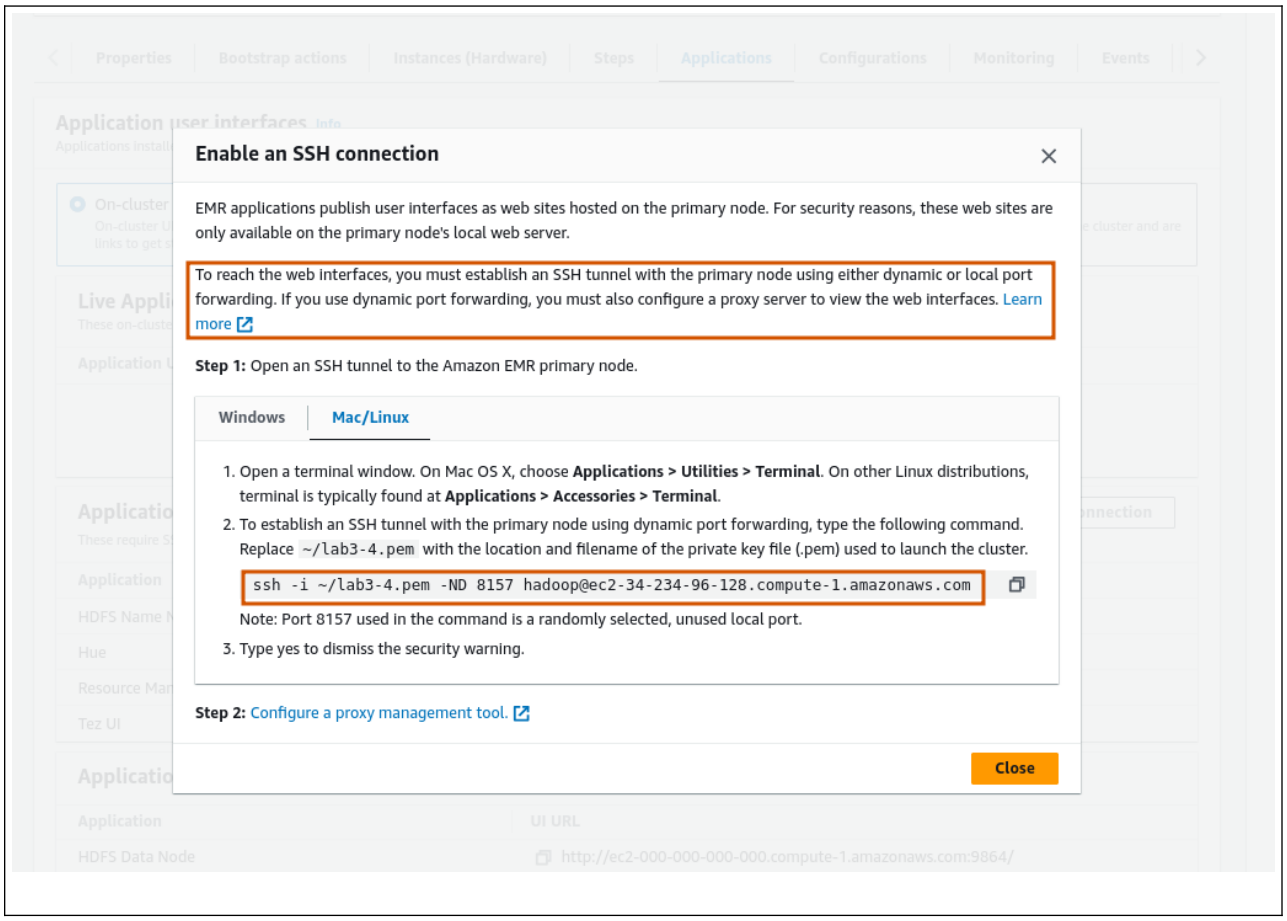

## 2. Upload Datasets to HDFS

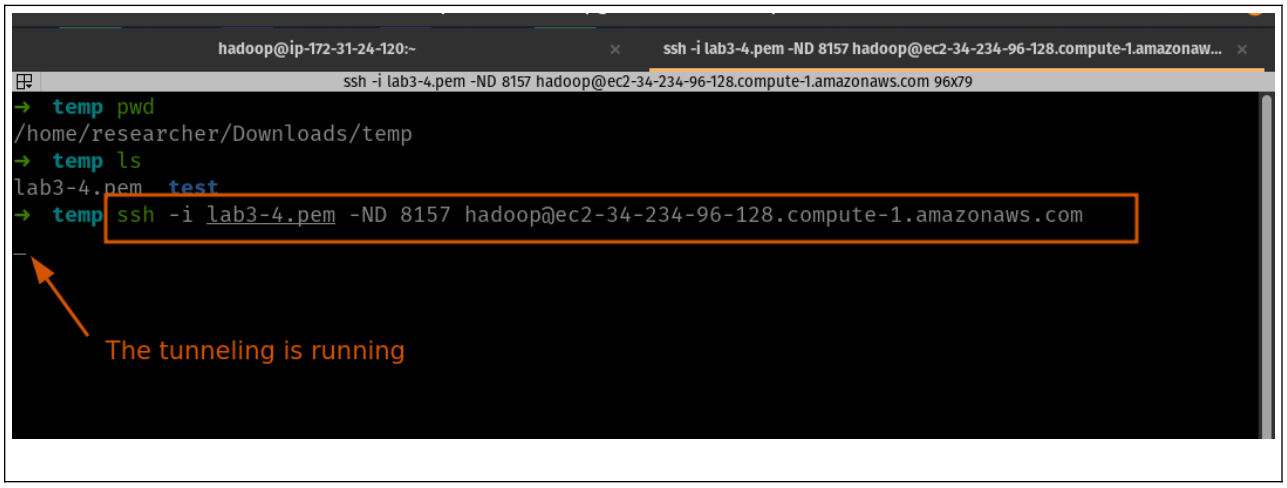
**2.1 In the AWS Image of EMR has kept the old configuration however, new version of the** hadoop runs HEAVs in a different port, so we have to manually change it. Here is the step to change that.


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
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
1. Click Here
Properties
Bootstrap actions
Instances (Hardware)
Steps
Applications
Configurations
Monitoring
Events
Application user interfaces Info
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
2. We need to enable the tunnel and you can use this command to do the tunnelling
No live application UIs
No live application UIs to display
Application UIs on the primary node
These require SSH tunneling to be enabled.
Enable an SSH connection
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
3. Then copy the Hue link and paste it in the chrome
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

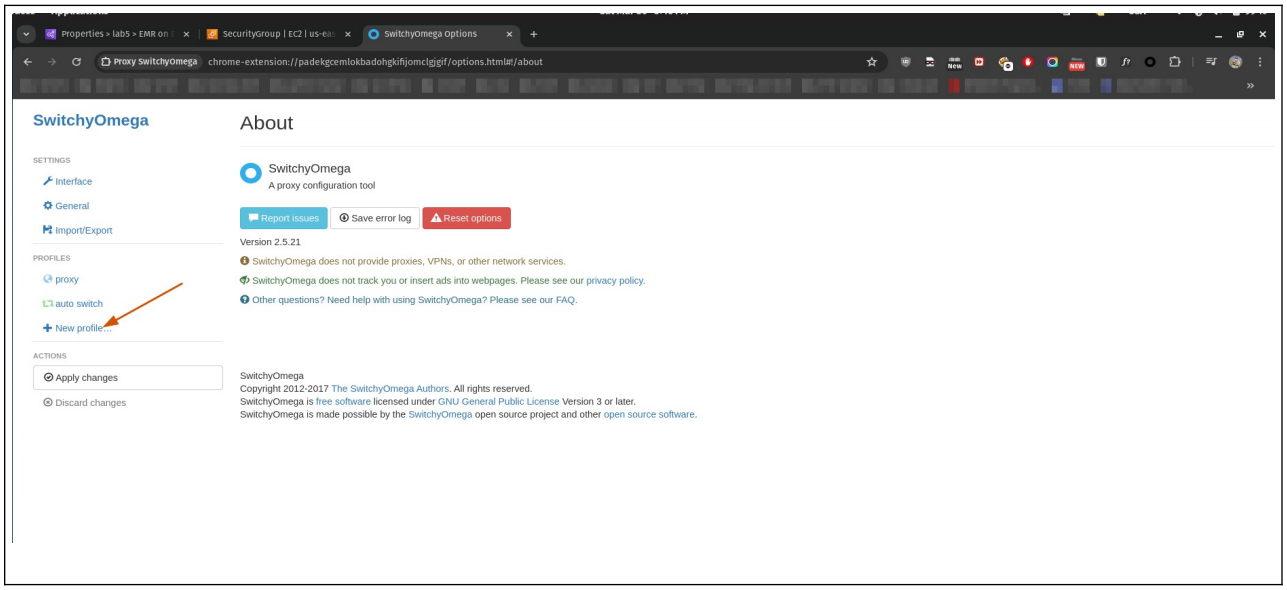
First, we need to enable SSH tunnel in the browser. Navigate to EMR cluster and click on “**Enable an SSH Connection**” in Application user interface:

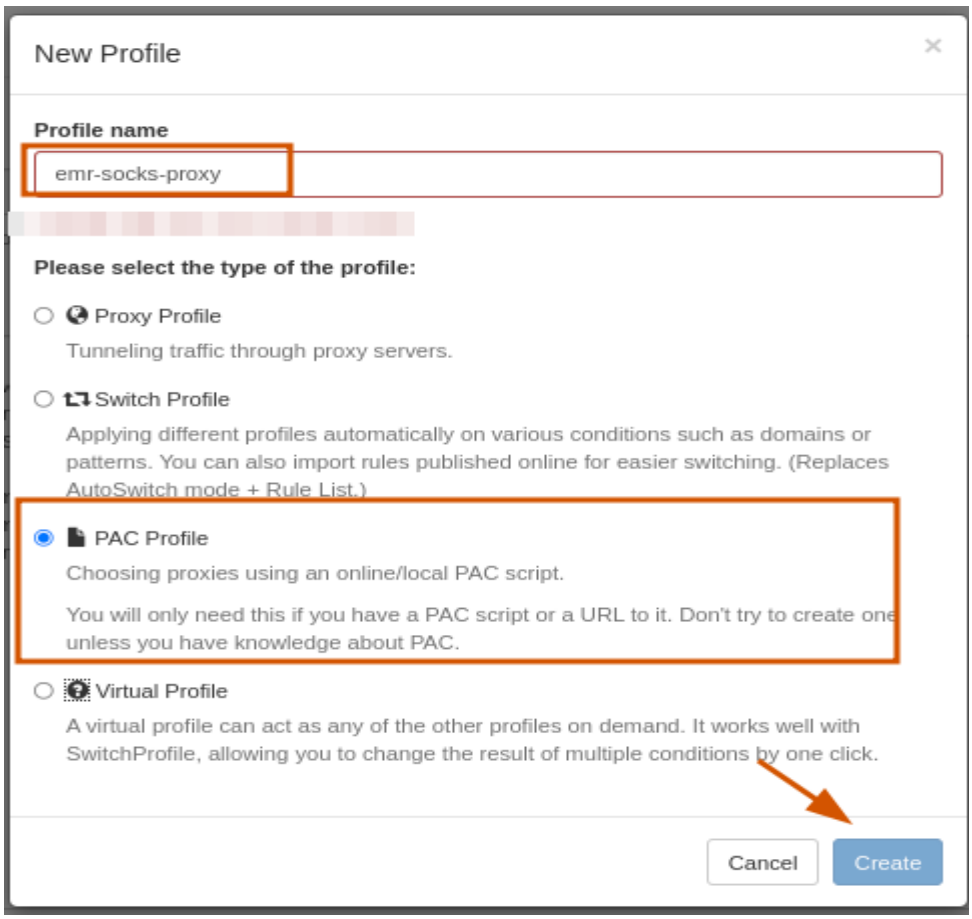
First, follow the instructions to enable an SSH tunnel to the EMR Master Node:



Then, install **SwitchyOmega** in Chrome. Here is the link to download it in chrome https://chromewebstore.google.com/detail/proxy-switchyomega/padekgcemlokbadohgkifijomclgjgif 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.

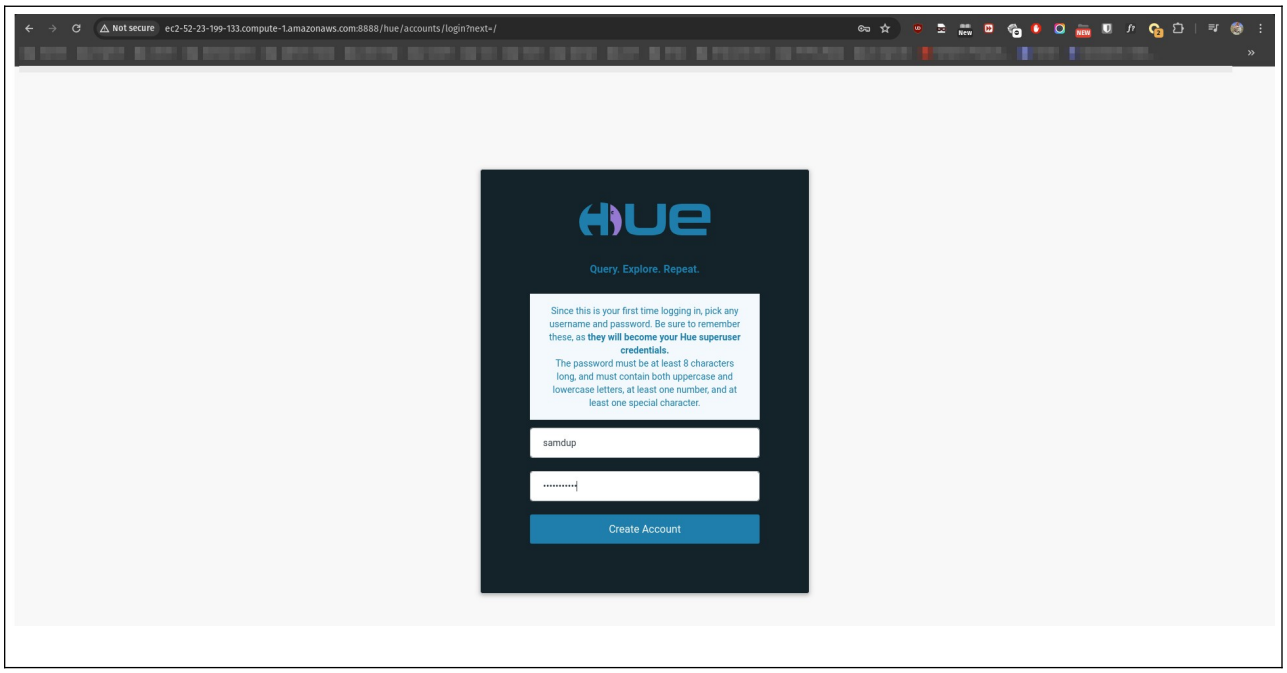


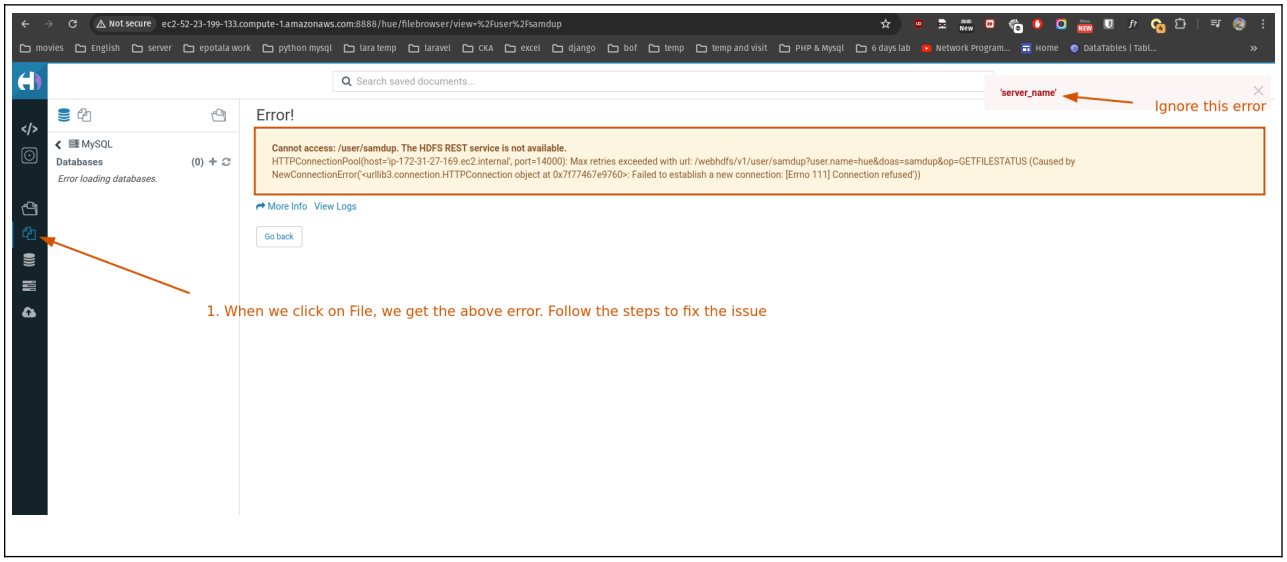




You can copy the script from this link

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**

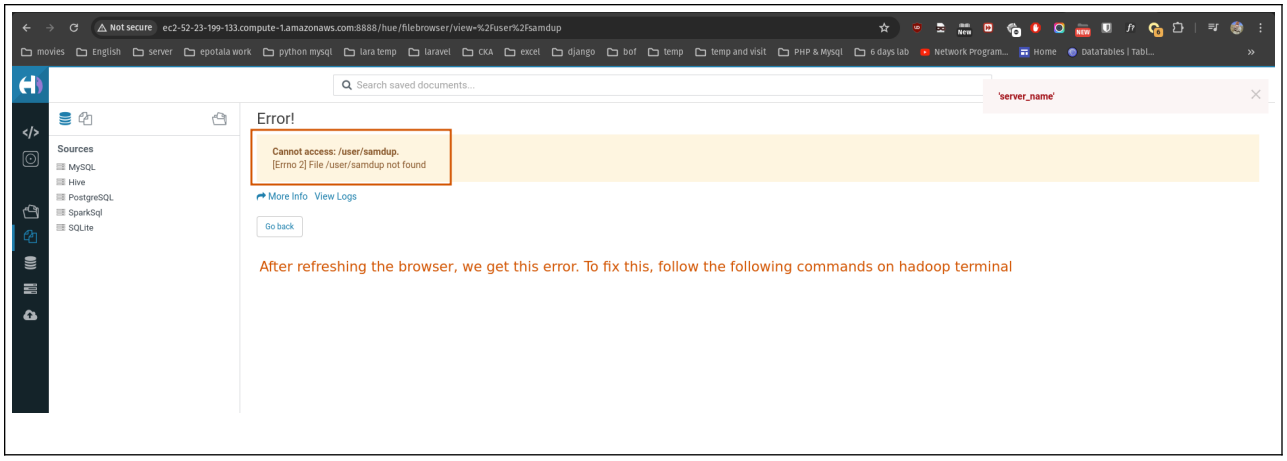
**search 14000 and replace this port number with 9870 (to search, you can use /14000 and press enter)**

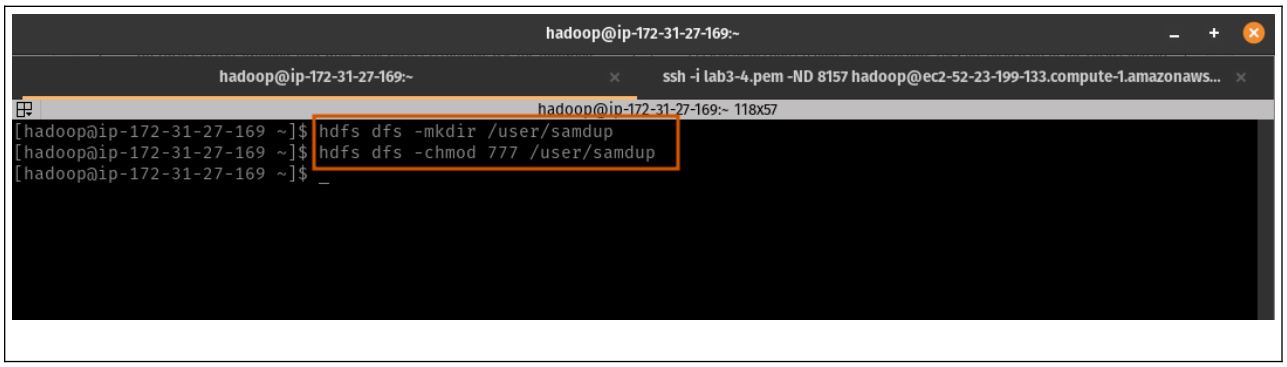
- How to search in vim (https://monovm.com/blog/how-to-search-in-vim-editor/)
- How to save file in vim (https://phoenixnap.com/kb/how-to-vim-save-quit-exit)

```
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)
```

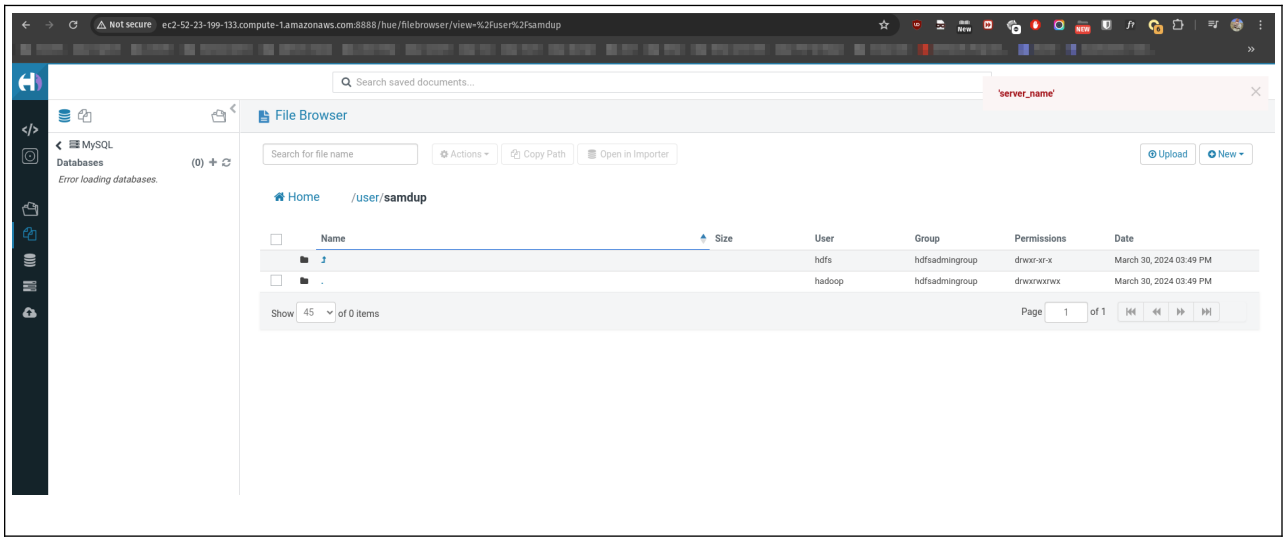
## sudo systemctl restart hue

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

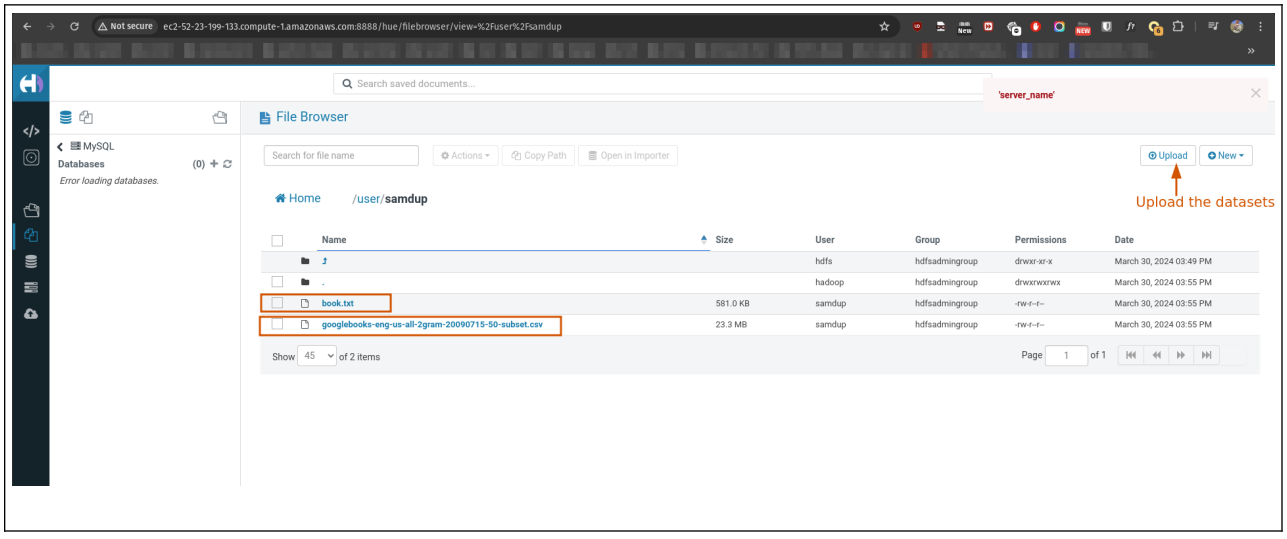





Home page of Hue:




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

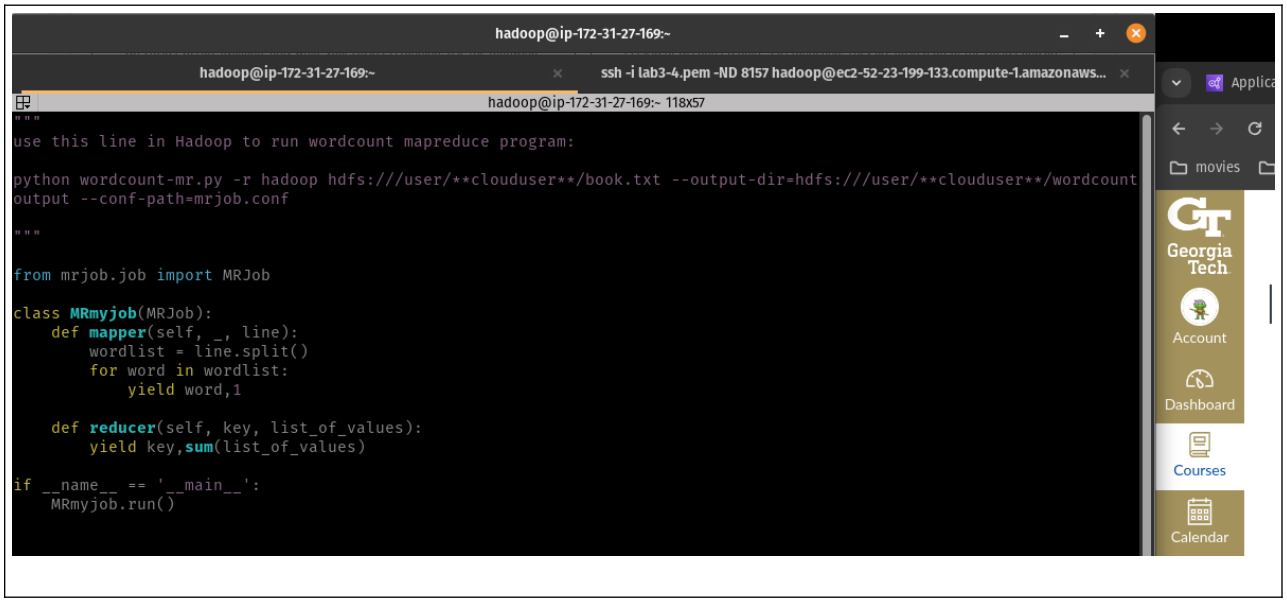


Next, create the mrjob configuration file on the Hadoop master node with vim (or nano) editor using the provided mrjob.conf, save and exit:

**vim  mrjob.conf**


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

**vim wordcount-mr.py**

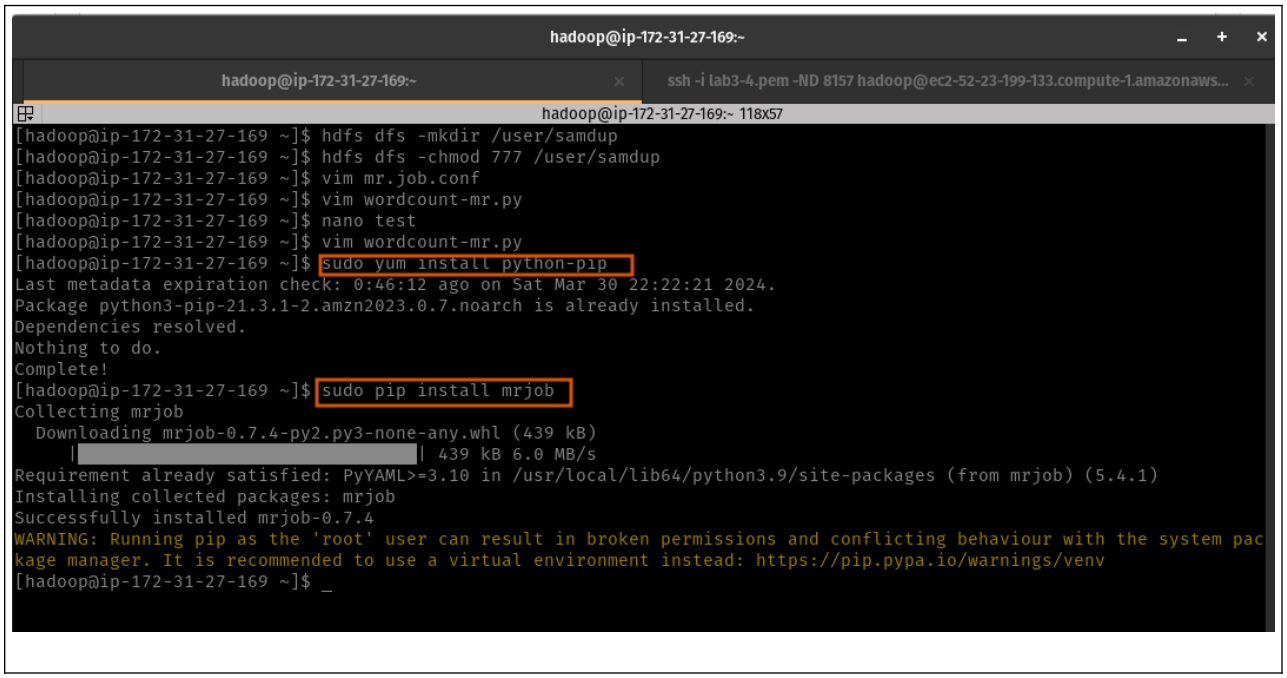


## 3. Setup mrJob on Master Node

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

sudo yum install python-pip

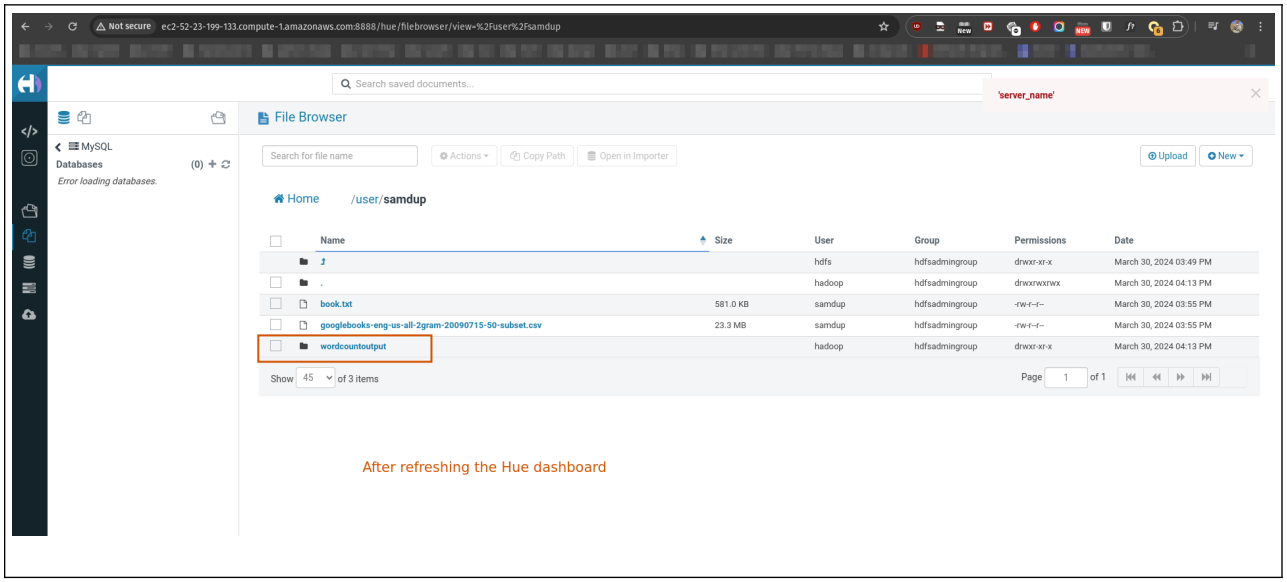
sudo pip install mrjob



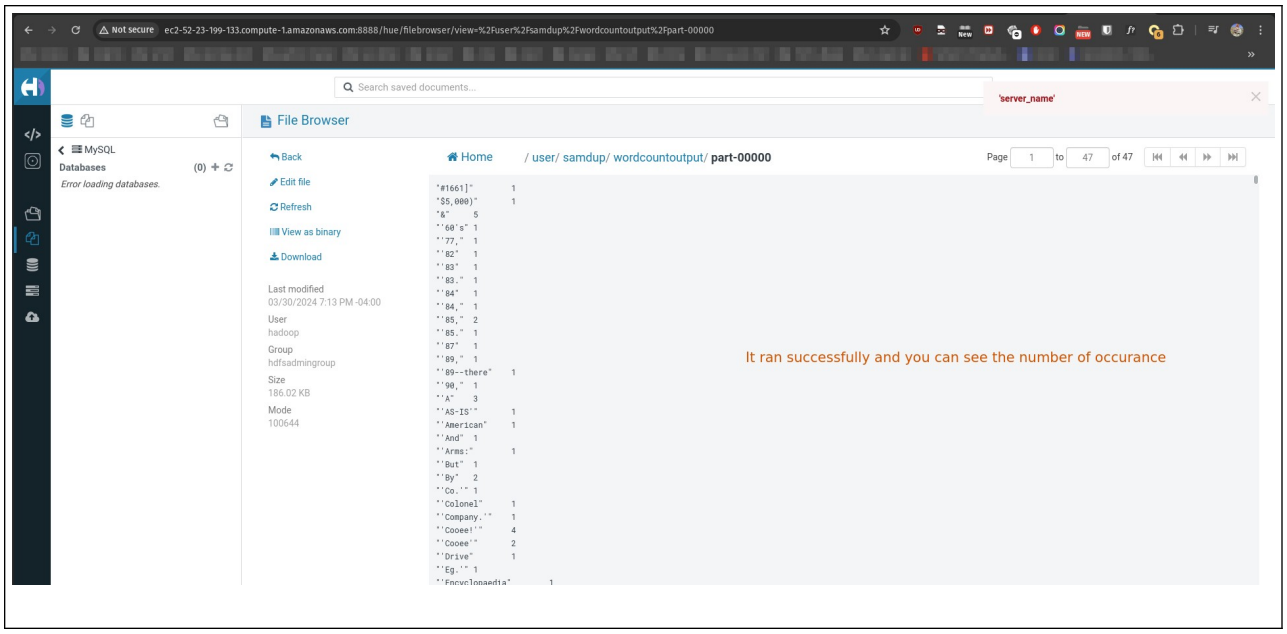

## 4. Run MapReduce Program

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

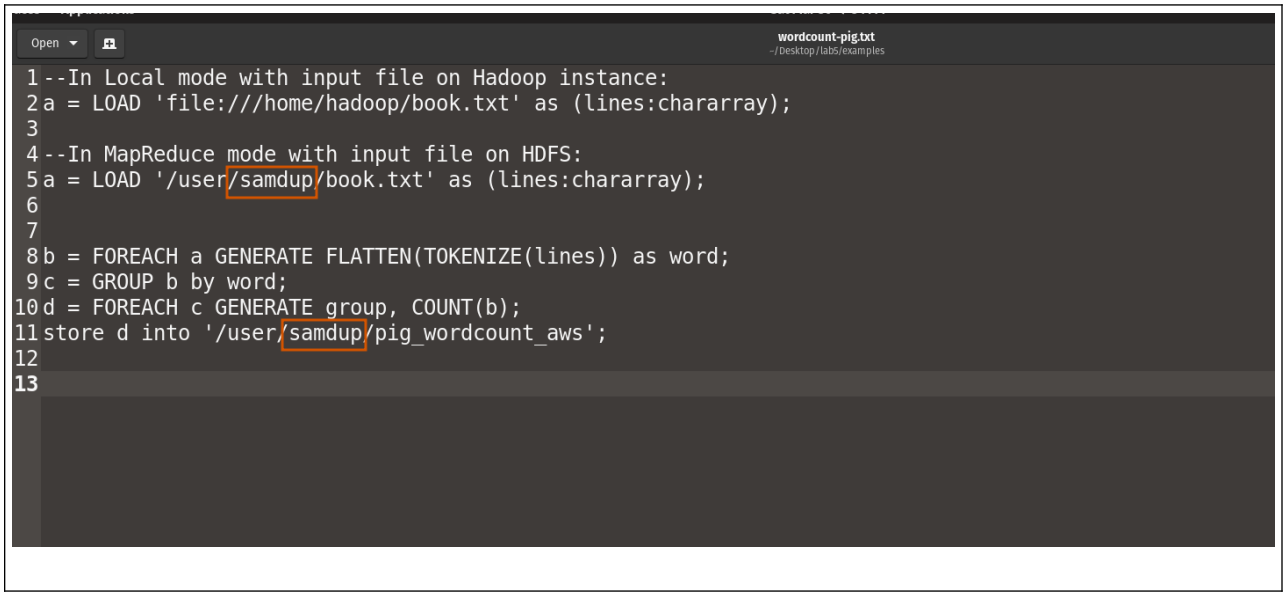



After refreshing the Hue dashboard

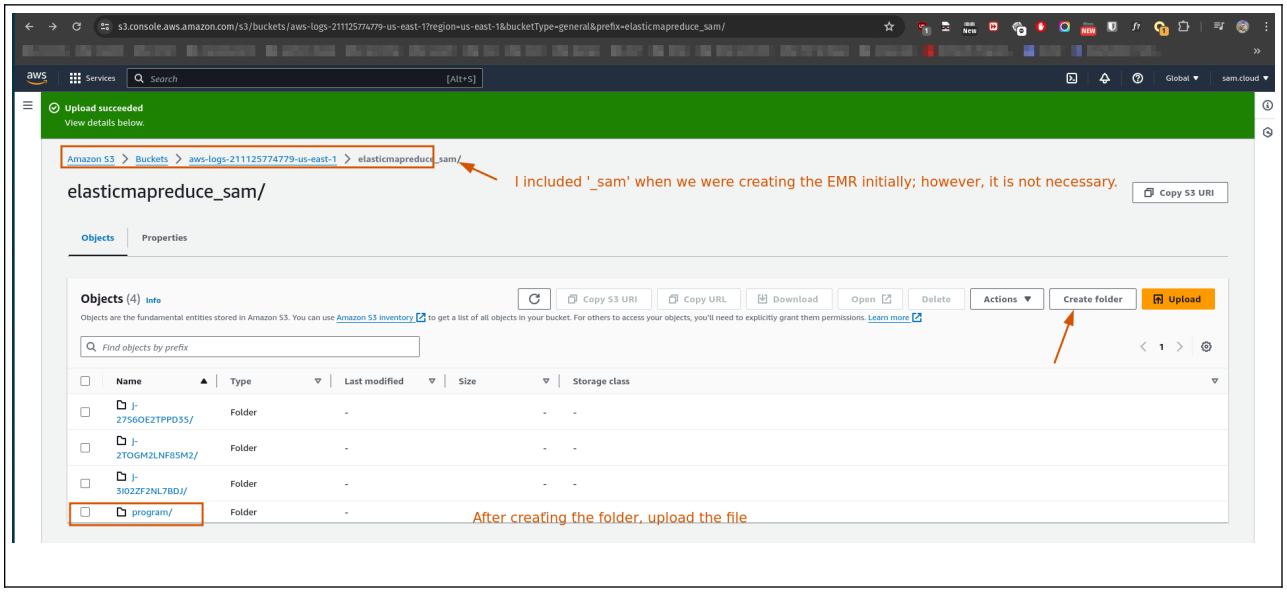


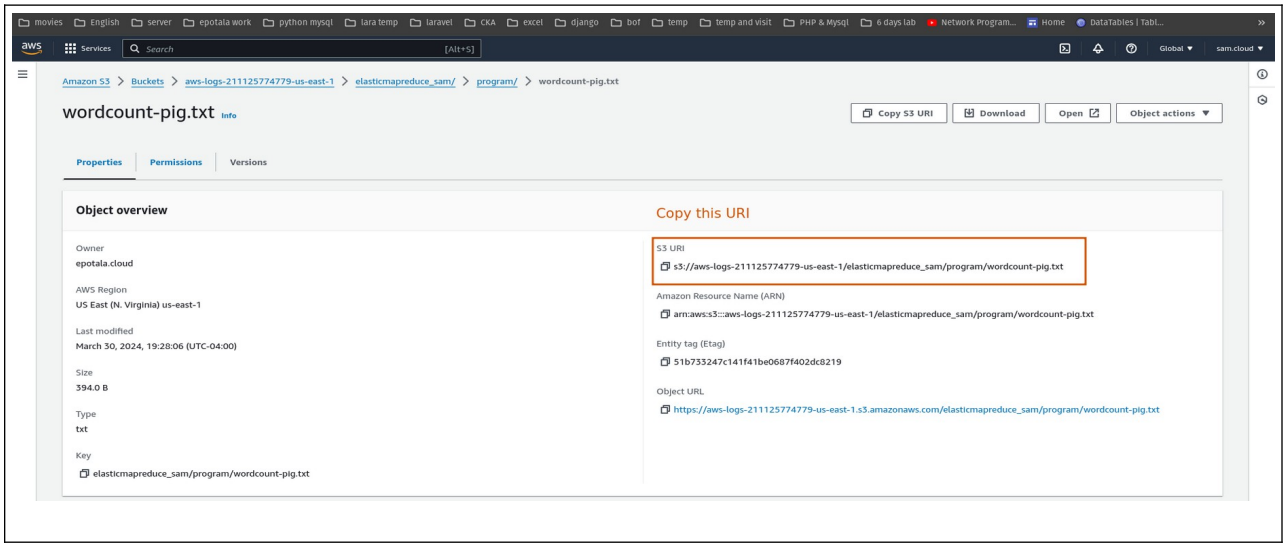

## 5. Run Pig Program (2 ways to approach this)

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



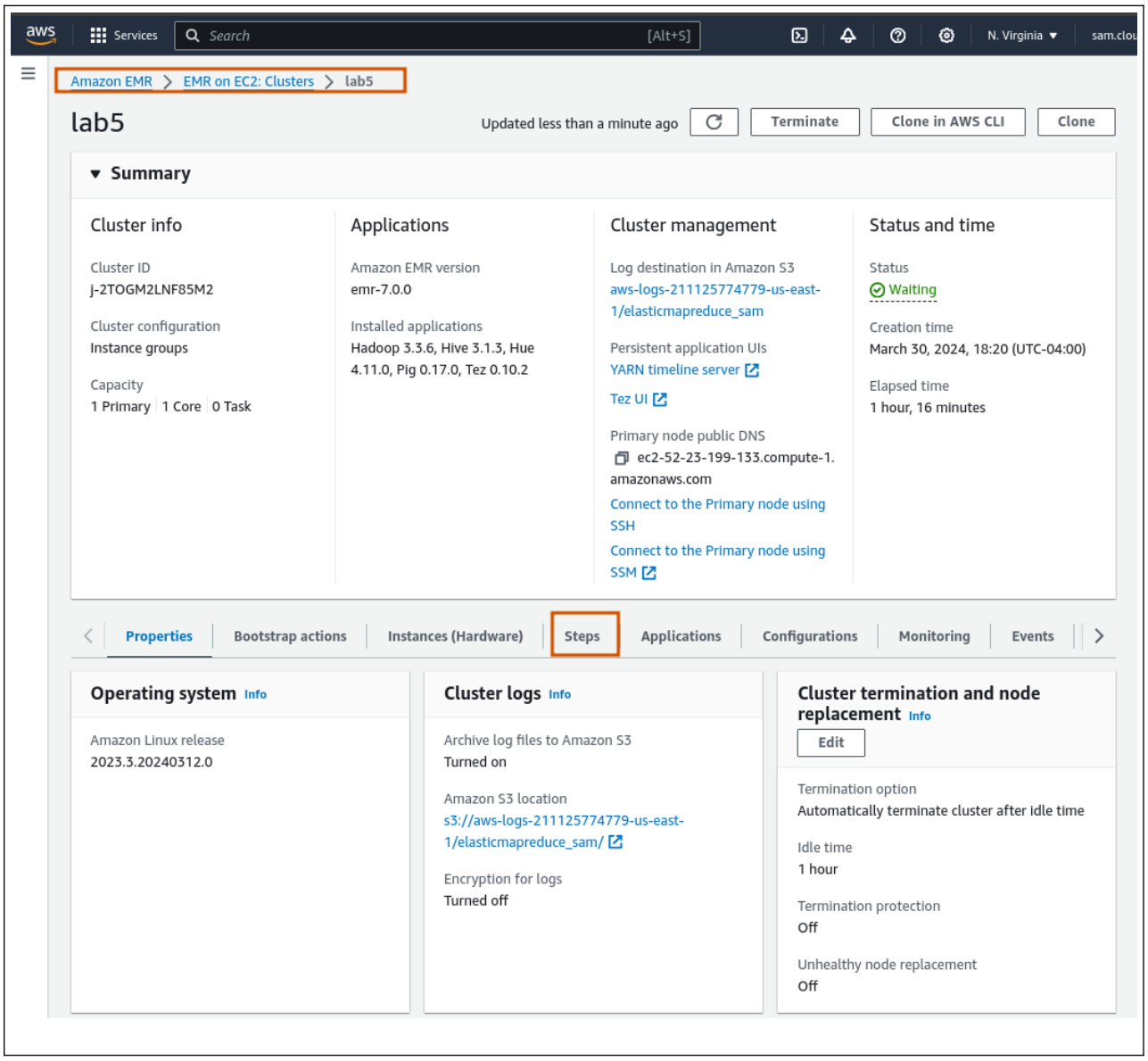
## b) Goto S3 bucket and create a folder ‘project’ and upload the wordcount-pig.txt

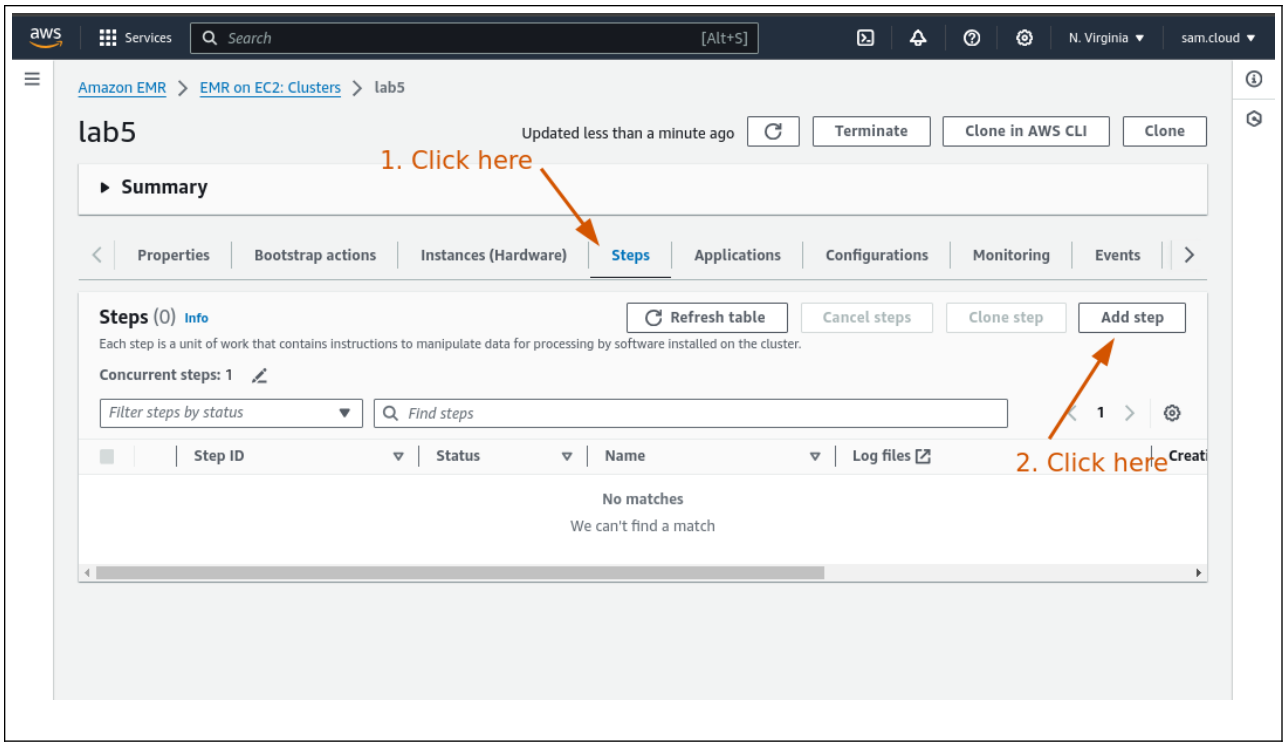


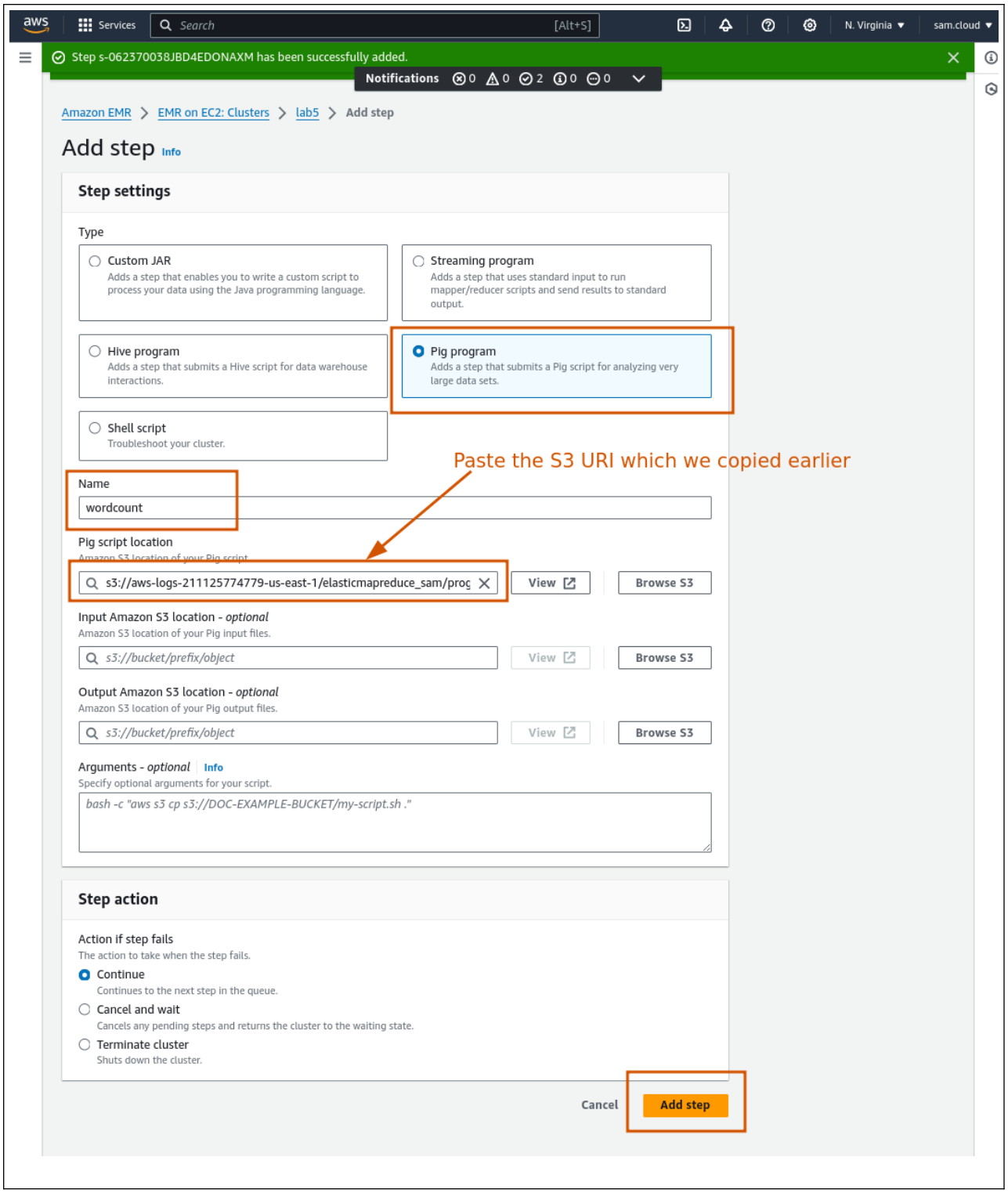



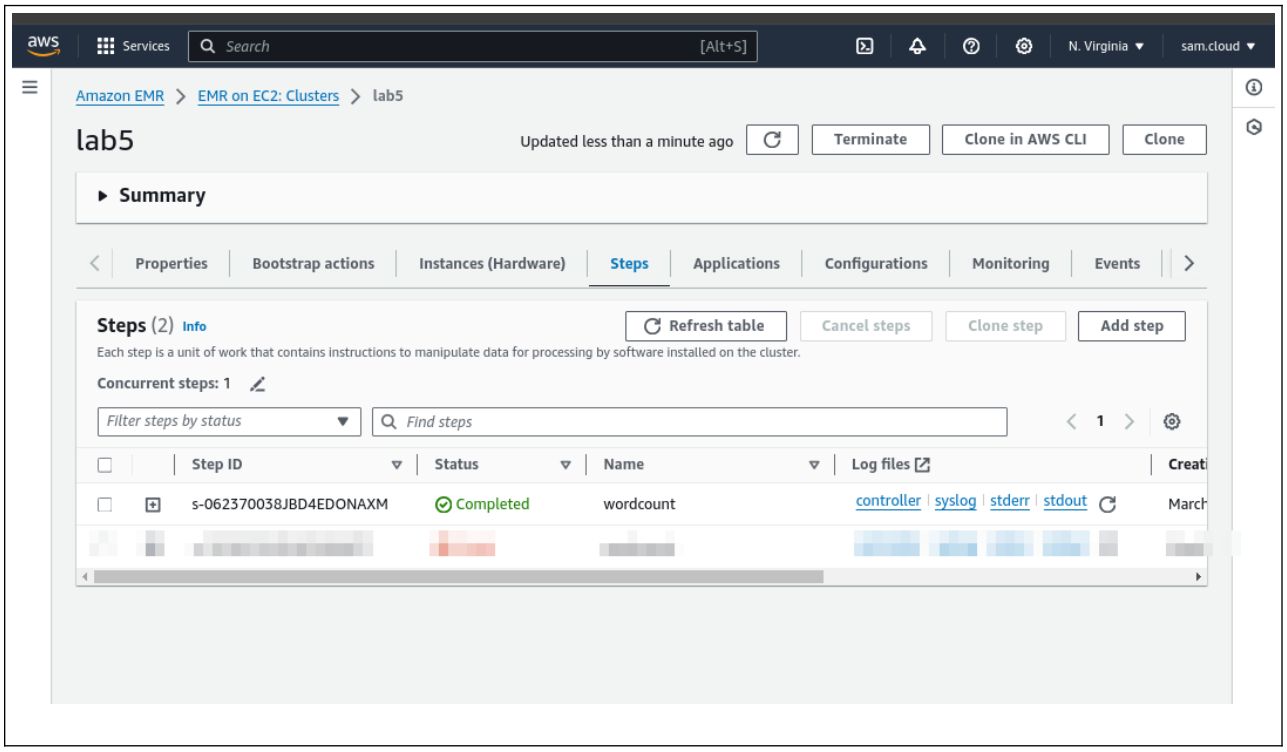

## Approach 1:

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

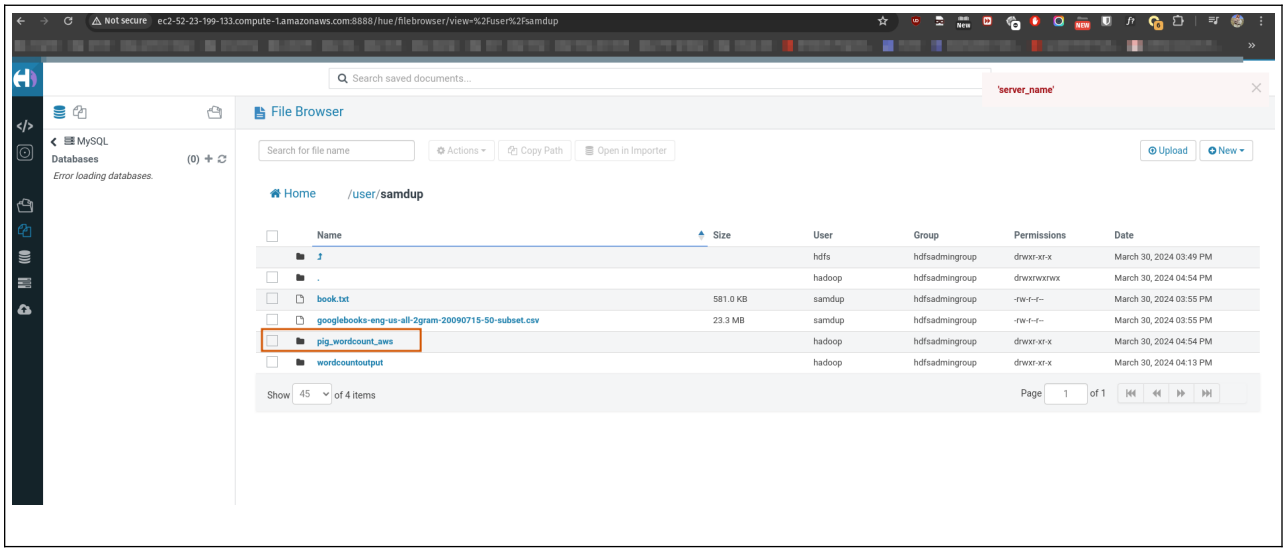


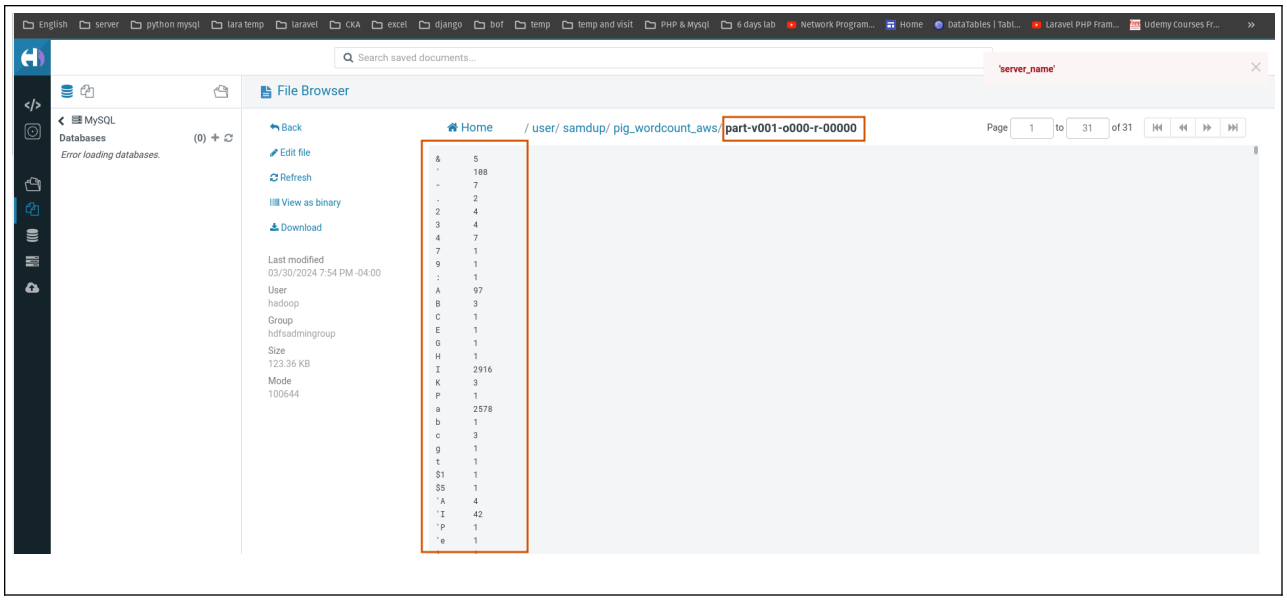

aws
Services
Search
[Alt+S]
N. Virginia
sam.cloud
Amazon EMR
EMR on EC2: Clusters
lab5
lab5
Updated less than a minute ago
Terminate
Clone in AWS CLI
Clone
1. Click here
Summary
Properties
Bootstrap actions
Instances (Hardware)
Steps
Applications
Configurations
Monitoring
Events
Steps (0) Info
Refresh table
Cancel steps
Clone step
Add step
Each step is a unit of work that contains instructions to manipulate data for processing by software installed on the cluster.
Concurrent steps: 1
Filter steps by status
Find steps
1
Step ID
Status
Name
Log files
2. Click here
No matches
We can't find a match


Step s-062370038JBD4EDONAXM has been successfully added.
Notifications
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.
Paste the S3 URI which we copied earlier
Name
wordcount
Pig script location
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
Output Amazon S3 location - optional
Amazon S3 location of your Pig output files.
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




After the job is finished, you can view the output in file browser:

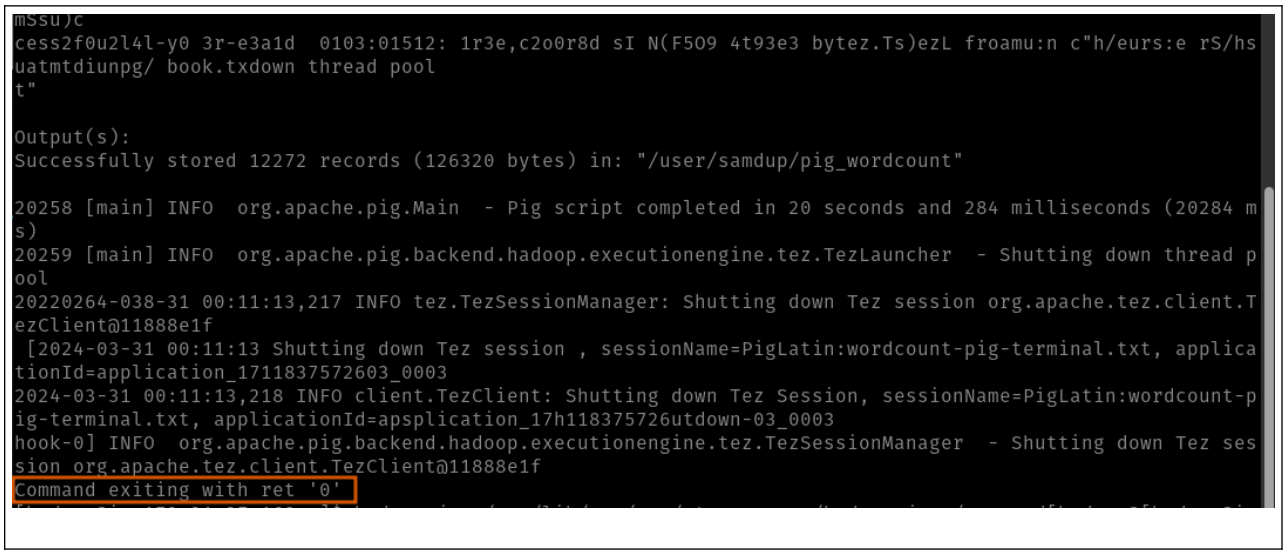


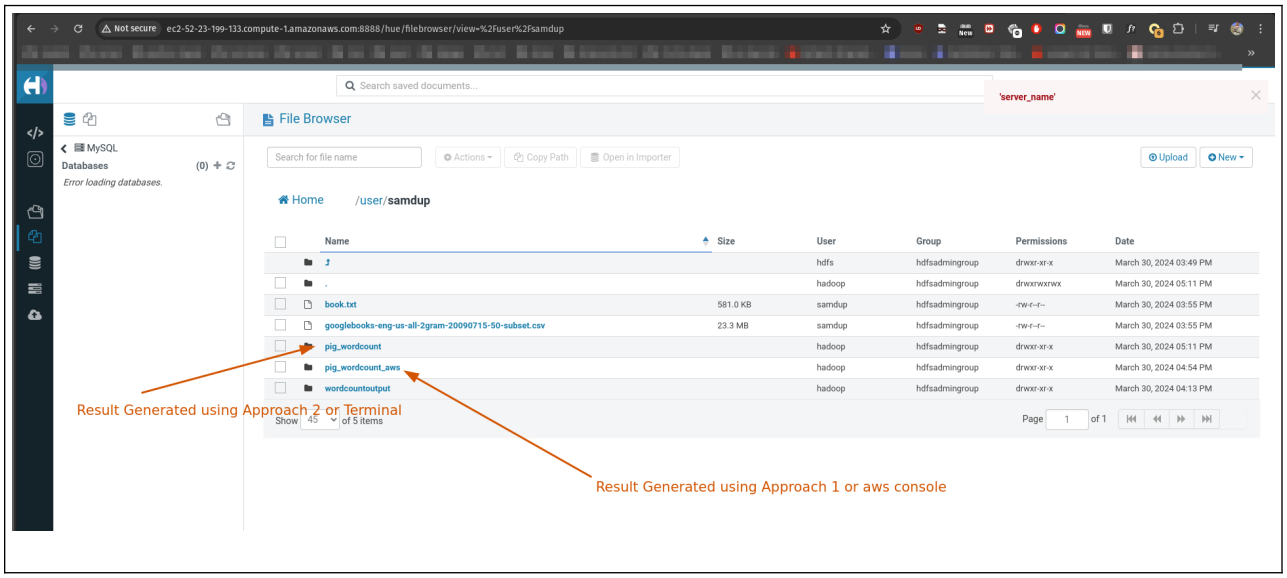


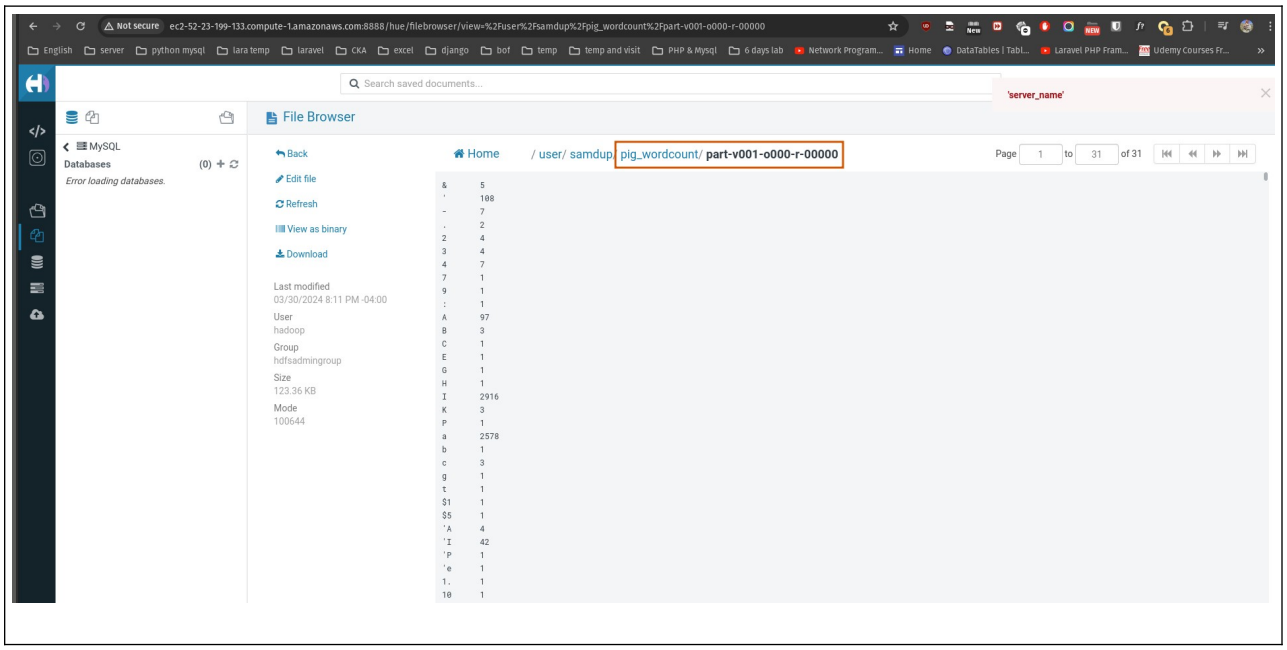

## Approach 2:

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**




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 thedataset (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), 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.