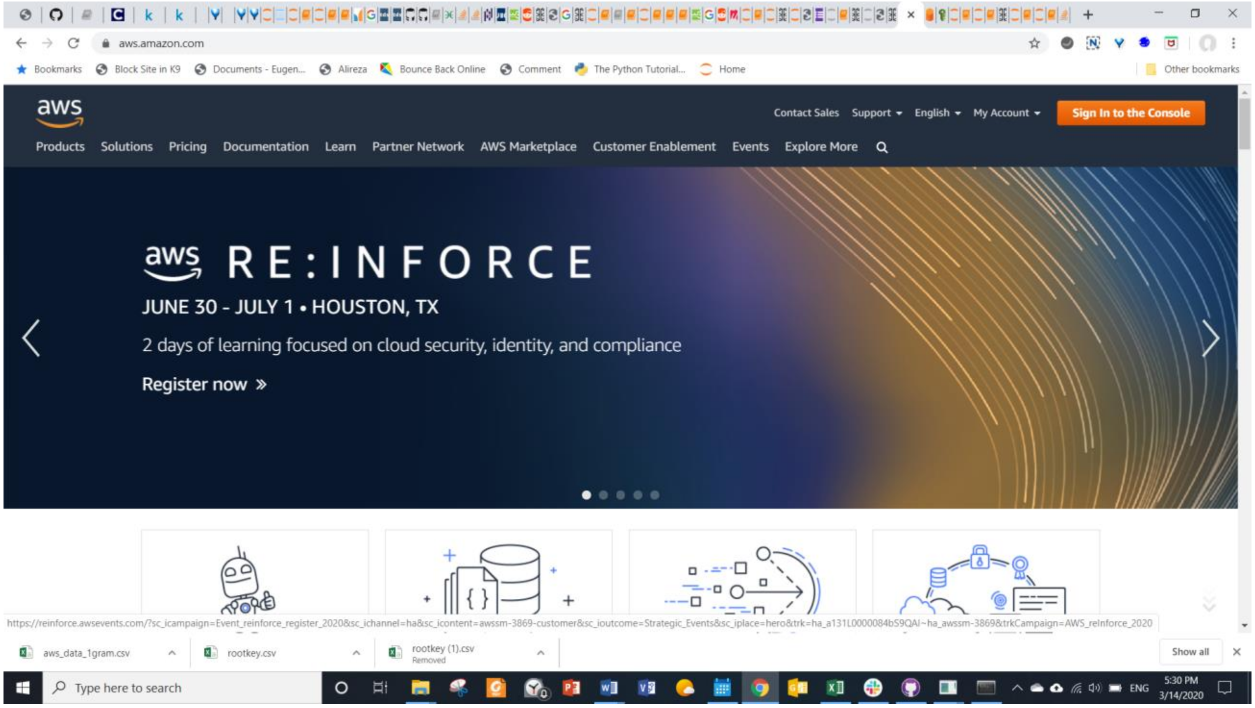
Contents

1 pin up a new EMR cluster

Go to aws.amazon.com and sign in to the console



aws RE:INFORCE

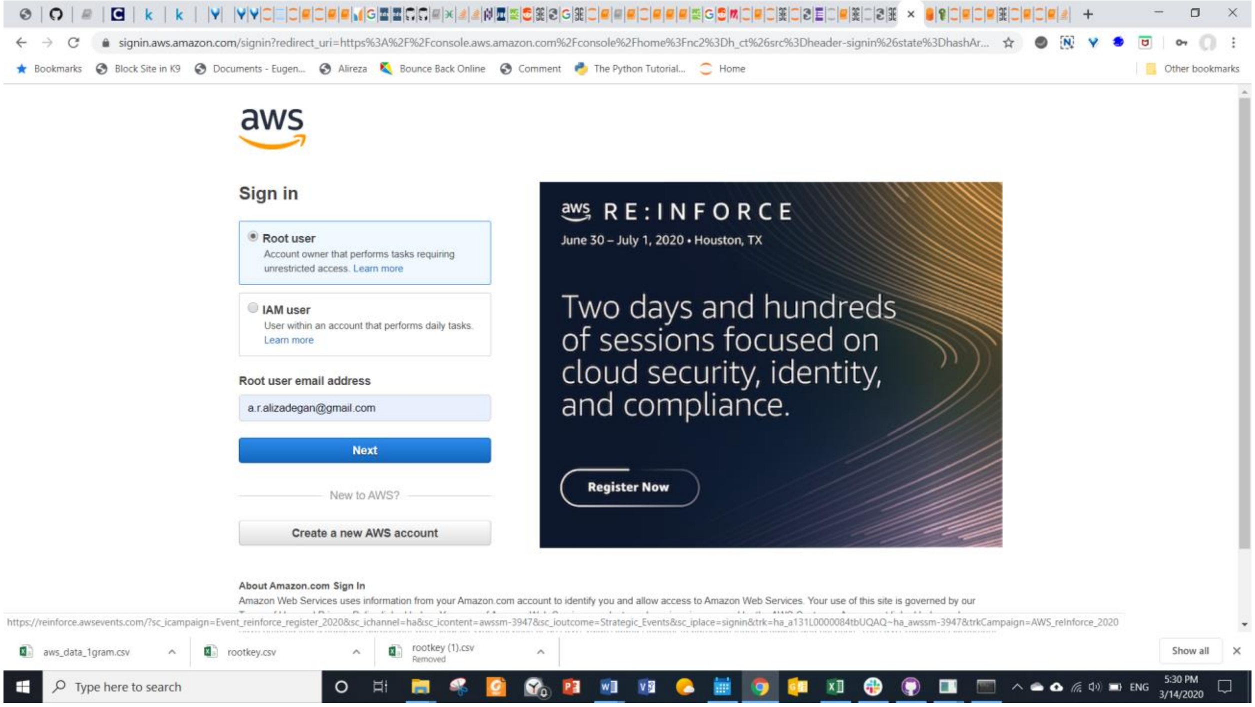
JUNE 30 - JULY 1 • HOUSTON, TX

2 days of learning focused on cloud security, identity, and compliance

Register now »

https://reinforce.awsevents.com/?sc_icampaign=Event_reinforce_register_2020&sc_ichannel=ha&sc_icontent=awssm-3869-customer&sc_ioutcome=Strategic_Events&sc_iplace=hero&trk=ha_a131L0000084bS9QAI~ha_awssm-3869&trkCampaign=AWS_reInforce_2020

Enter root user email address and click next



signin.aws.amazon.com/signin?redirect_uri=https%3A%2F%2Fconsole.aws.amazon.com%2Fconsole%2Fhome%3Fnc2%3Dh_ct%26src%3Dheader-signin%26state%3DhashAr...

Sign in

● Root user
Account owner that performs tasks requiring unrestricted access. Learn more

○ IAM user
User within an account that performs daily tasks. Learn more

Root user email address: a.r.alizadegan@gmail.com

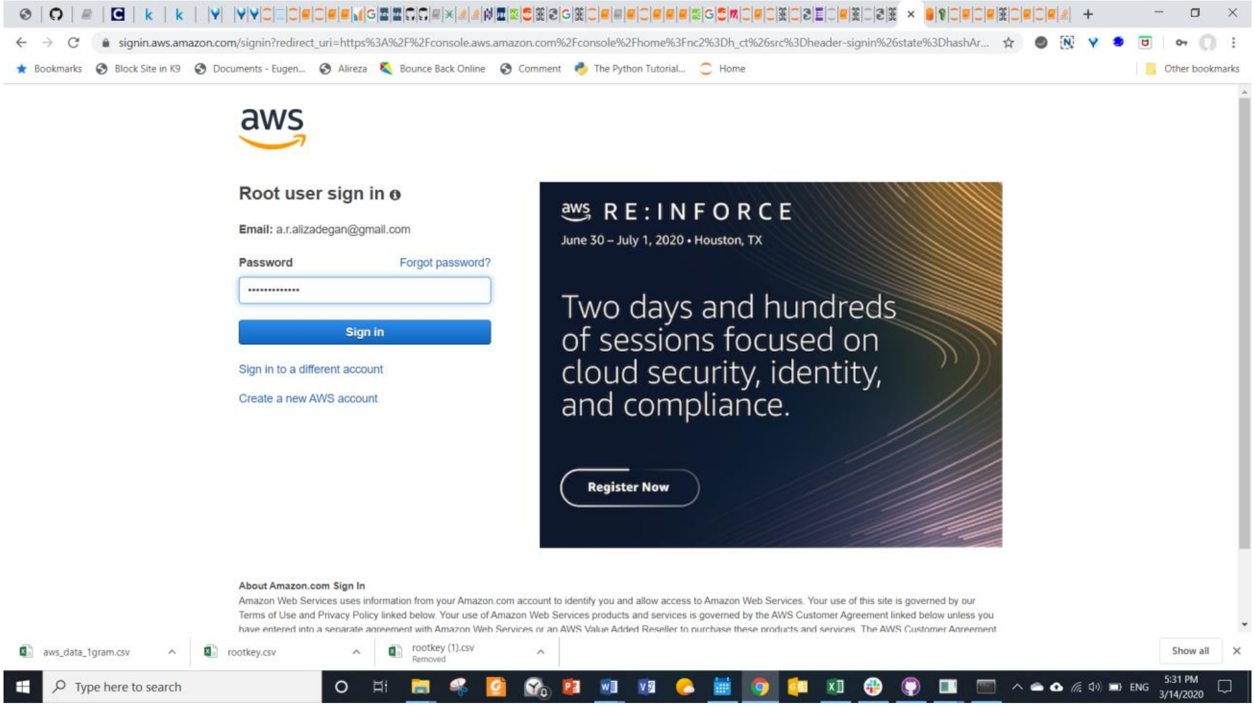
Next
New to AWS?
Create a new AWS account
aws RE:INFORCE
June 30 – July 1, 2020 • Houston, TX
Two days and hundreds of sessions focused on cloud security, identity, and compliance.
Register Now

About Amazon.com Sign In

Amazon Web Services uses information from your Amazon.com account to identify you and allow access to Amazon Web Services. Your use of this site is governed by our

https://reinforce.awsevents.com/?sc_icampaign=Event_reinforce_register_2020&sc_ichannel=ha&sc_icontent=awssm-3947&sc_ioutcome=Strategic_Events&sc_iplace=signin&trk=ha_a131L0000084tbUQAQ~ha_awssm-3947&trkCampaign=AWS_reInforce_2020

Enter password and sign in



signin.aws.amazon.com/signin?redirect_uri=https%3A%2F%2Fconsole.aws.amazon.com%2Fconsole%2Fhome%3Fnc2%3Dh_ct%26src%3Dheader-signin%26state%3DhashAr...

Root user sign in

Email: a.r.alizadegan@gmail.com

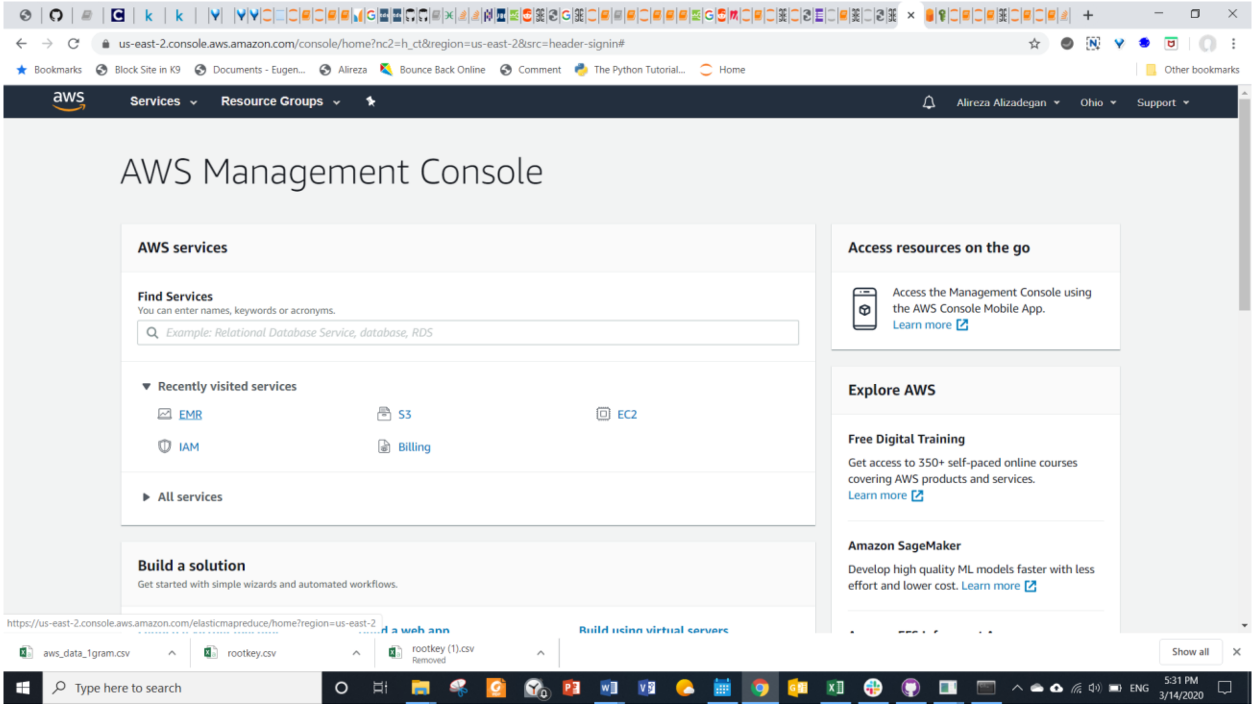
Password: ••••••••••••••

Forgot password?
Sign in
Sign in to a different account
Create a new AWS account
aws RE:INFORCE
June 30 – July 1, 2020 • Houston, TX
Two days and hundreds of sessions focused on cloud security, identity, and compliance.
Register Now

About Amazon.com Sign In

Amazon Web Services uses information from your Amazon.com account to identify you and allow access to Amazon Web Services. Your use of this site is governed by our Terms of Use and Privacy Policy linked below. Your use of Amazon Web Services products and services is governed by the AWS Customer Agreement linked below unless you have entered into a separate agreement with Amazon Web Services or an AWS Value Added Reseller to purchase these products and services. The AWS Customer Agreement

Choose EMR from AWS services in their management console



AWS Management Console

AWS services

Find Services
You can enter names, keywords or acronyms.

Example: Relational Database Service, database, RDS

▼ Recently visited services

- EMR
- S3
- EC2
- IAM
- Billing

▶ All services

Build a solution
Get started with simple wizards and automated workflows.

Access resources on the go

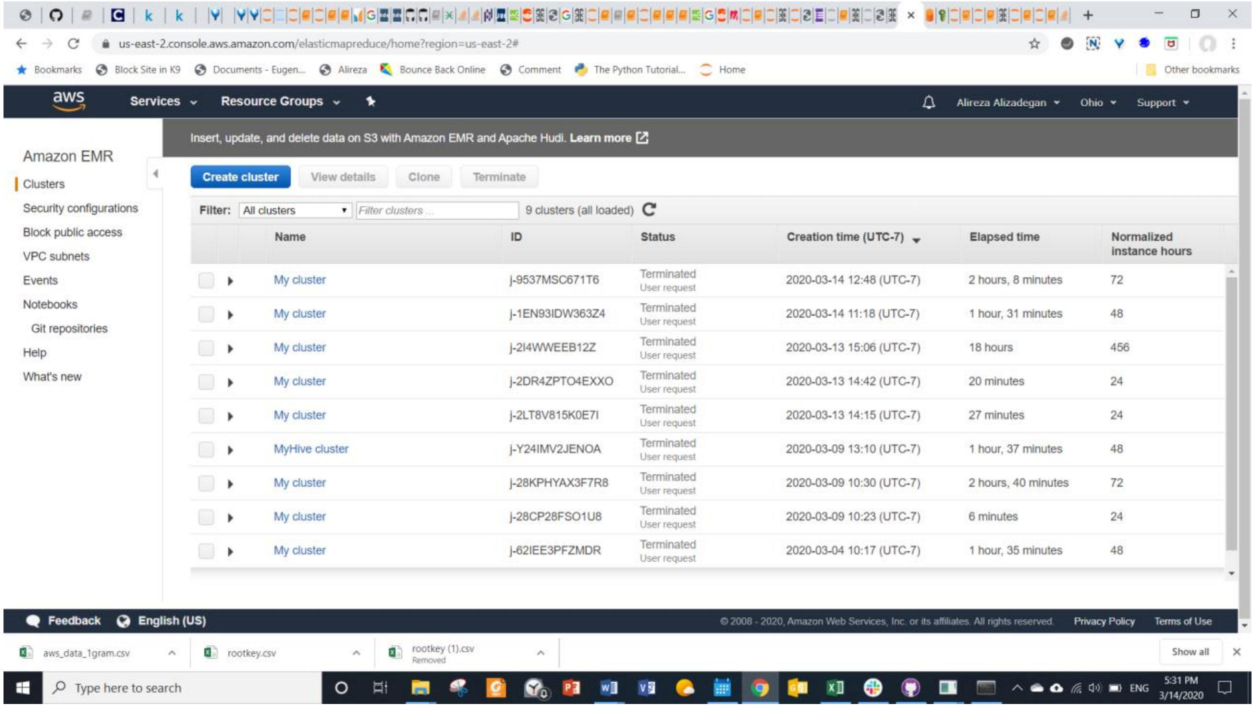
Access the Management Console using the AWS Console Mobile App. Learn more

Explore AWS

Free Digital Training
Get access to 350+ self-paced online courses covering AWS products and services. Learn more

Amazon SageMaker
Develop high quality ML models faster with less effort and lower cost. Learn more

Click create cluster in EMR clusters page

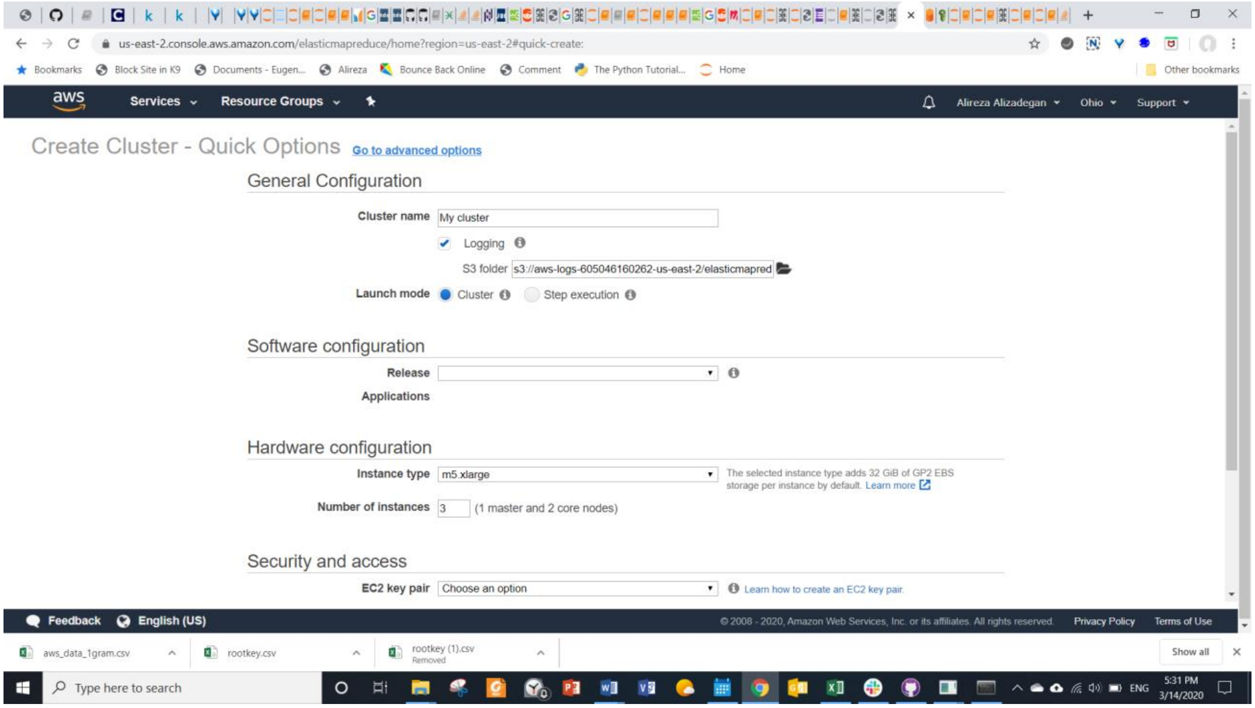


Insert, update, and delete data on S3 with Amazon EMR and Apache Hudi. **Learn more**

Filter: All clusters ▾ Filter clusters ... 9 clusters (all loaded)

		Name	ID	Status	Creation time (UTC-7) ▾	Elapsed time	Normalized instance hours
□	▶	My cluster	j-9537MSC671T6	Terminated User request	2020-03-14 12:48 (UTC-7)	2 hours, 8 minutes	72
□	▶	My cluster	j-1EN93IDW363Z4	Terminated User request	2020-03-14 11:18 (UTC-7)	1 hour, 31 minutes	48
□	▶	My cluster	j-2I4WWEEB12Z	Terminated User request	2020-03-13 15:06 (UTC-7)	18 hours	456
□	▶	My cluster	j-2DR4ZPTO4EXXO	Terminated User request	2020-03-13 14:42 (UTC-7)	20 minutes	24
□	▶	My cluster	j-2LT8V815K0E7I	Terminated User request	2020-03-13 14:15 (UTC-7)	27 minutes	24
□	▶	MyHive cluster	j-Y24IMV2JENOA	Terminated User request	2020-03-09 13:10 (UTC-7)	1 hour, 37 minutes	48
□	▶	My cluster	j-28KPHYAX3F7R8	Terminated User request	2020-03-09 10:30 (UTC-7)	2 hours, 40 minutes	72
□	▶	My cluster	j-28CP28FSO1U8	Terminated User request	2020-03-09 10:23 (UTC-7)	6 minutes	24
□	▶	My cluster	j-62IEE3PFZMDR	Terminated User request	2020-03-04 10:17 (UTC-7)	1 hour, 35 minutes	48

Go to advanced options in the options page



Create Cluster - Quick Options Go to advanced options

General Configuration

Cluster name: My cluster

☑ Logging

S3 folder: s3://aws-logs-605046160262-us-east-2/elasticmapred

Launch mode: ● Cluster ○ Step execution

Software configuration

Release:

Applications

Hardware configuration

Instance type: m5.xlarge

The selected instance type adds 32 GiB of GP2 EBS storage per instance by default. Learn more

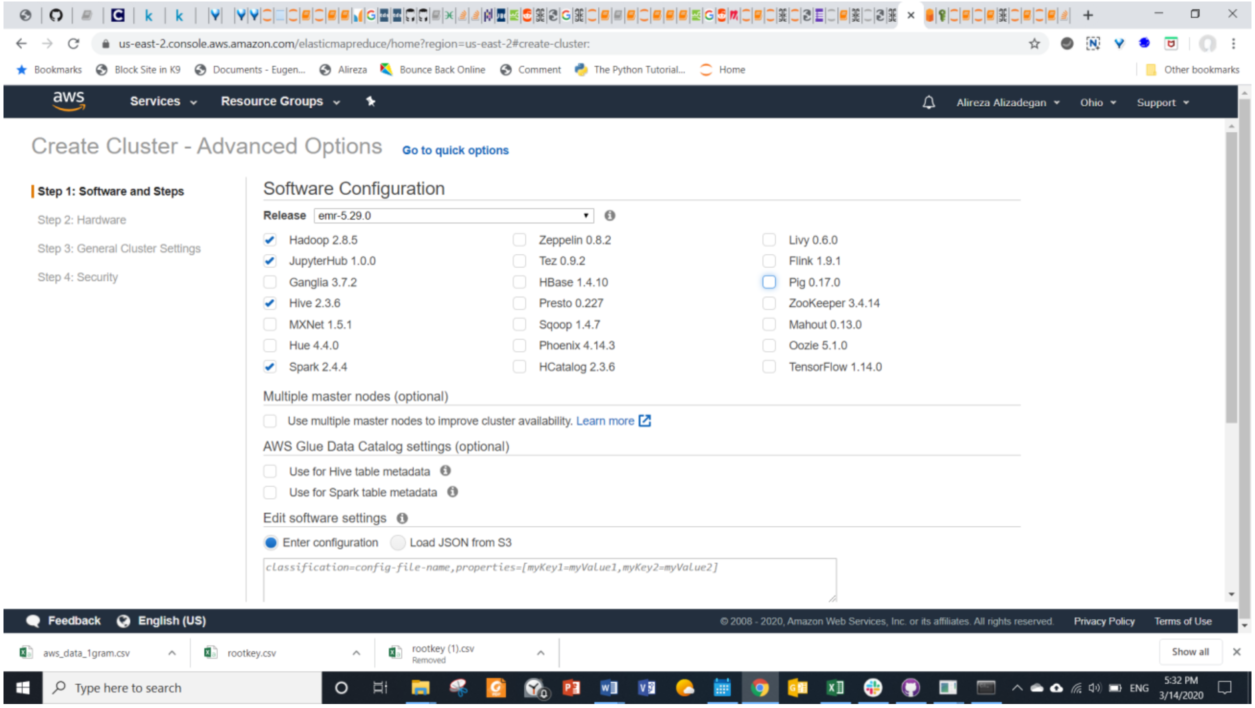
Number of instances: 3 (1 master and 2 core nodes)

Security and access

EC2 key pair: Choose an option

Learn how to create an EC2 key pair.

Select Hadoop, jupyterhub, hive, and spark in software & steps step



Create Cluster - Advanced Options Go to quick options

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release: emr-5.29.0

☑ Hadoop 2.8.5
☑ JupyterHub 1.0.0
□ Ganglia 3.7.2
☑ Hive 2.3.6
□ MXNet 1.5.1
□ Hue 4.4.0
☑ Spark 2.4.4

□ Zeppelin 0.8.2
□ Tez 0.9.2
□ HBase 1.4.10
□ Presto 0.227
□ Sqoop 1.4.7
□ Phoenix 4.14.3
□ HCatalog 2.3.6

□ Livy 0.6.0
□ Flink 1.9.1
□ Pig 0.17.0
□ ZooKeeper 3.4.14
□ Mahout 0.13.0
□ Oozie 5.1.0
□ TensorFlow 1.14.0

Multiple master nodes (optional)

□ Use multiple master nodes to improve cluster availability. Learn more

AWS Glue Data Catalog settings (optional)

□ Use for Hive table metadata

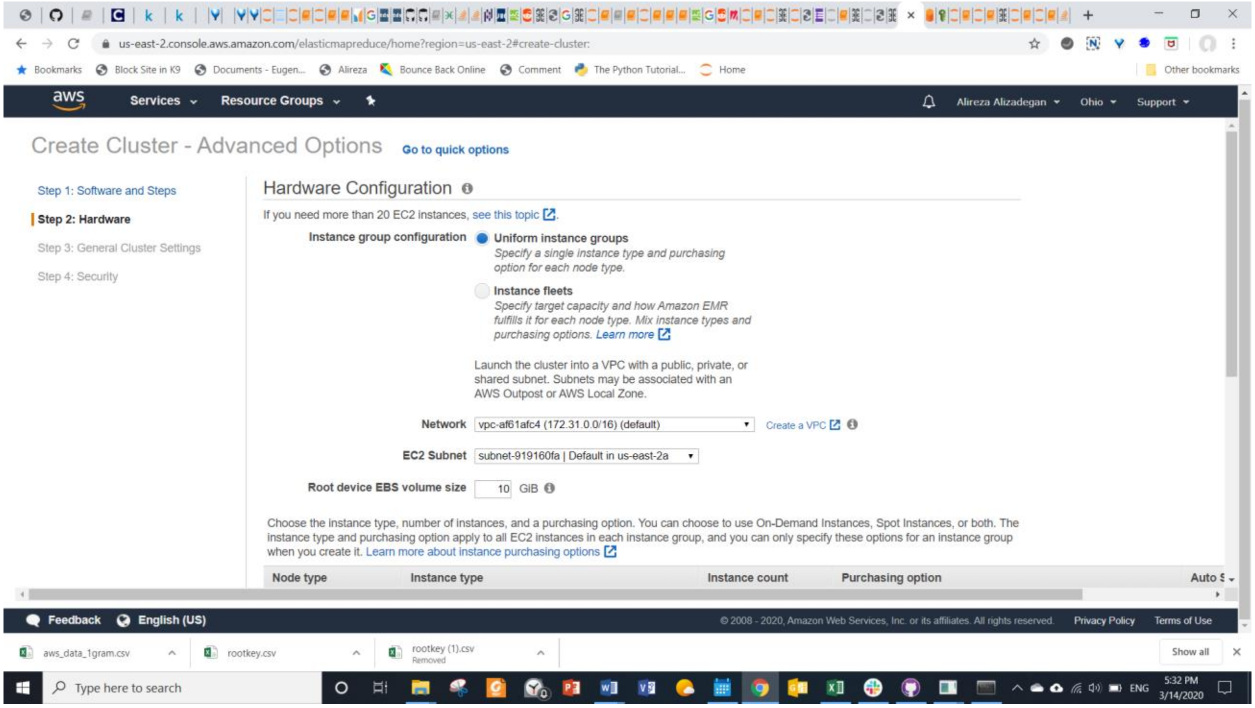
□ Use for Spark table metadata

Edit software settings

● Enter configuration ○ Load JSON from S3

```
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
```

Do not make any changes in hardware step



Create Cluster - Advanced Options Go to quick options

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

Hardware Configuration

If you need more than 20 EC2 instances, see this topic.

Instance group configuration ● **Uniform instance groups** *Specify a single instance type and purchasing option for each node type.*

○ **Instance fleets** *Specify target capacity and how Amazon EMR fulfills it for each node type. Mix instance types and purchasing options. Learn more*

Launch the cluster into a VPC with a public, private, or shared subnet. Subnets may be associated with an AWS Outpost or AWS Local Zone.

Network vpc-af61afc4 (172.31.0.0/16) (default) Create a VPC

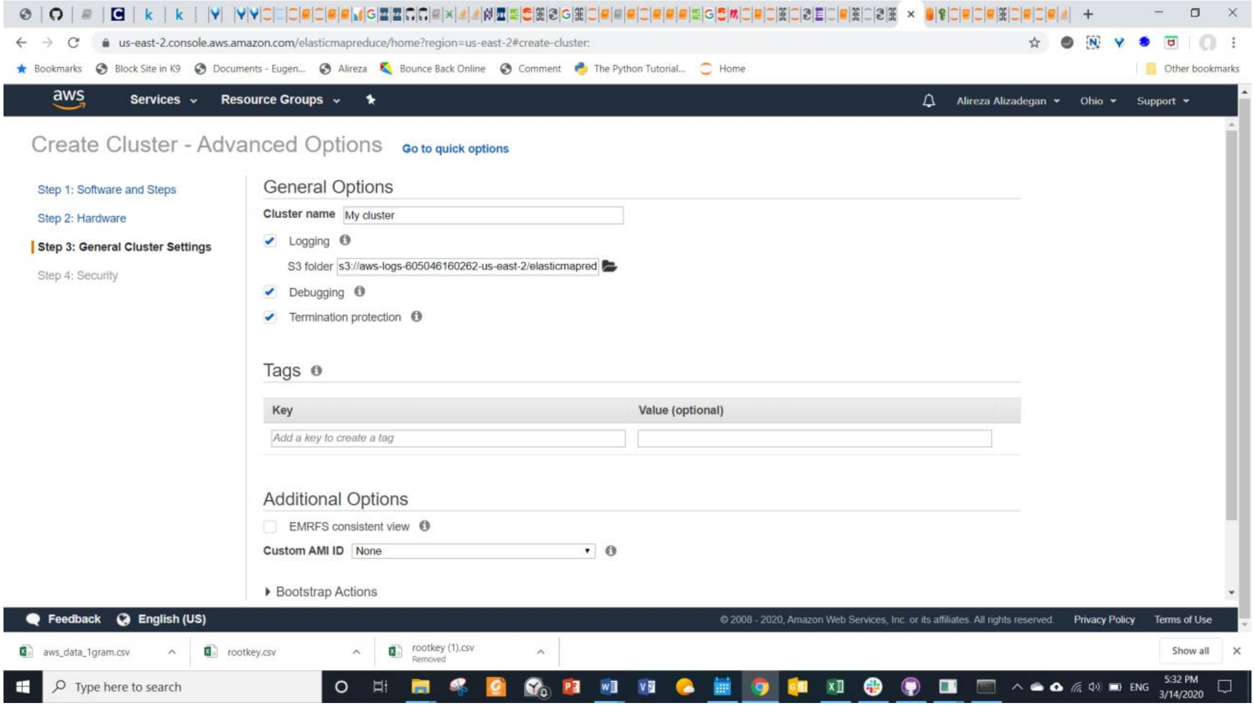
EC2 Subnet subnet-919160fa | Default in us-east-2a

Root device EBS volume size 10 GiB

Choose the instance type, number of instances, and a purchasing option. You can choose to use On-Demand Instances, Spot Instances, or both. The instance type and purchasing option apply to all EC2 instances in each instance group, and you can only specify these options for an instance group when you create it. Learn more about instance purchasing options

Node type	Instance type	Instance count	Purchasing option	Auto S

Name the cluster in the general cluster settings step



Create Cluster - Advanced Options Go to quick options

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

General Options

Cluster name My cluster

☑ Logging

S3 folder s3://aws-logs-605046160262-us-east-2/elasticmapred

☑ Debugging

☑ Termination protection

Tags

Key	Value (optional)
Add a key to create a tag	

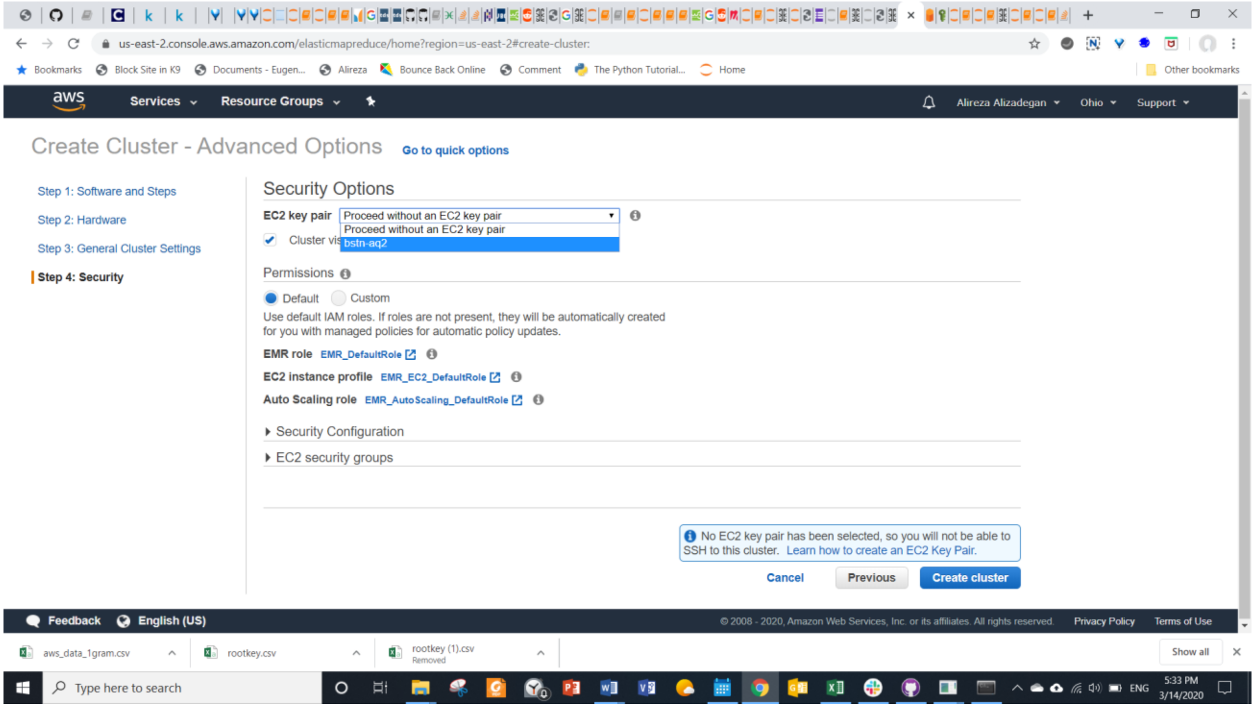
Additional Options

□ EMRFS consistent view

Custom AMI ID None

▸ Bootstrap Actions

Select the EC2 key pair in security options section of security step and click create cluster



Create Cluster - Advanced Options

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

Security Options

EC2 key pair Proceed without an EC2 key pair
Proceed without an EC2 key pair
bstn-aq2

☑ Cluster vis

Permissions

● Default ○ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role EMR_DefaultRole

EC2 instance profile EMR_EC2_DefaultRole

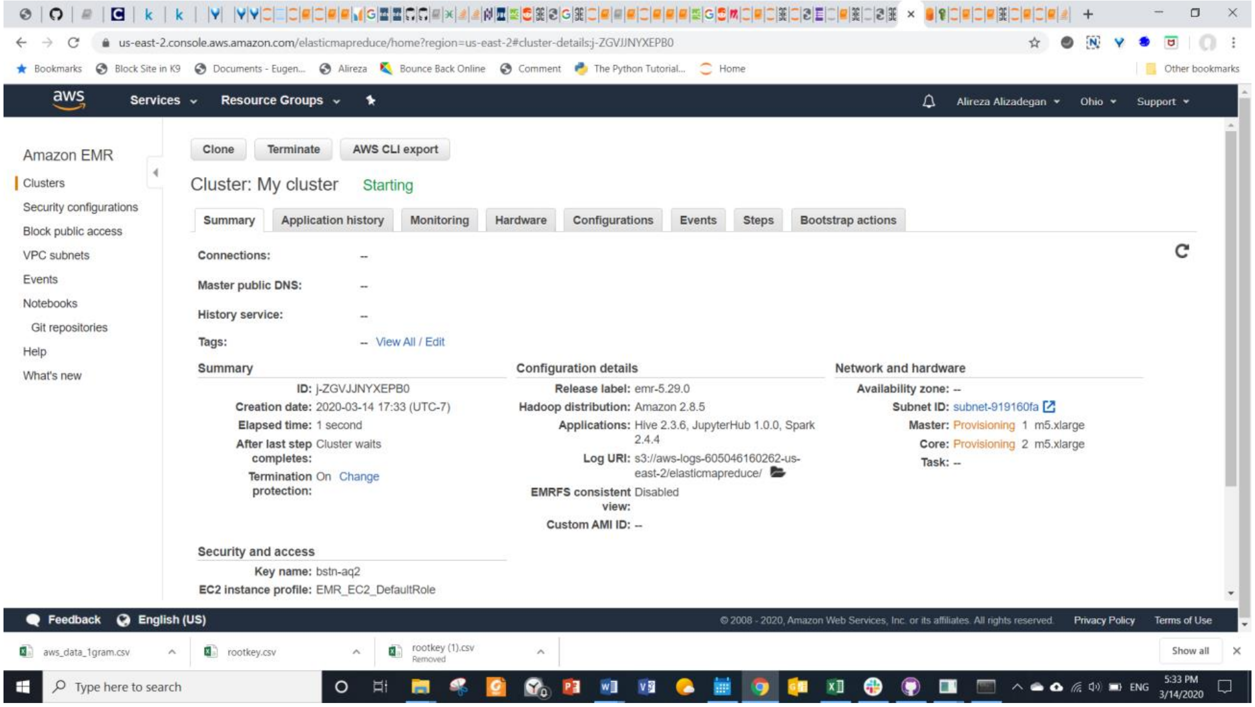
Auto Scaling role EMR_AutoScaling_DefaultRole

▸ Security Configuration

▸ EC2 security groups

No EC2 key pair has been selected, so you will not be able to SSH to this cluster. Learn how to create an EC2 Key Pair.

Starting status shows up in green font in front of the cluster name as below



Cluster: My cluster Starting

Connections: --
Master public DNS: --
History service: --
Tags: -- View All / Edit

Summary

ID: j-ZGVJJNYXEPB0
Creation date: 2020-03-14 17:33 (UTC-7)
Elapsed time: 1 second
After last step completes: Cluster waits
Termination protection: On Change

Configuration details

Release label: emr-5.29.0
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, JupyterHub 1.0.0, Spark 2.4.4
Log URI: s3://aws-logs-605046160262-us-east-2/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

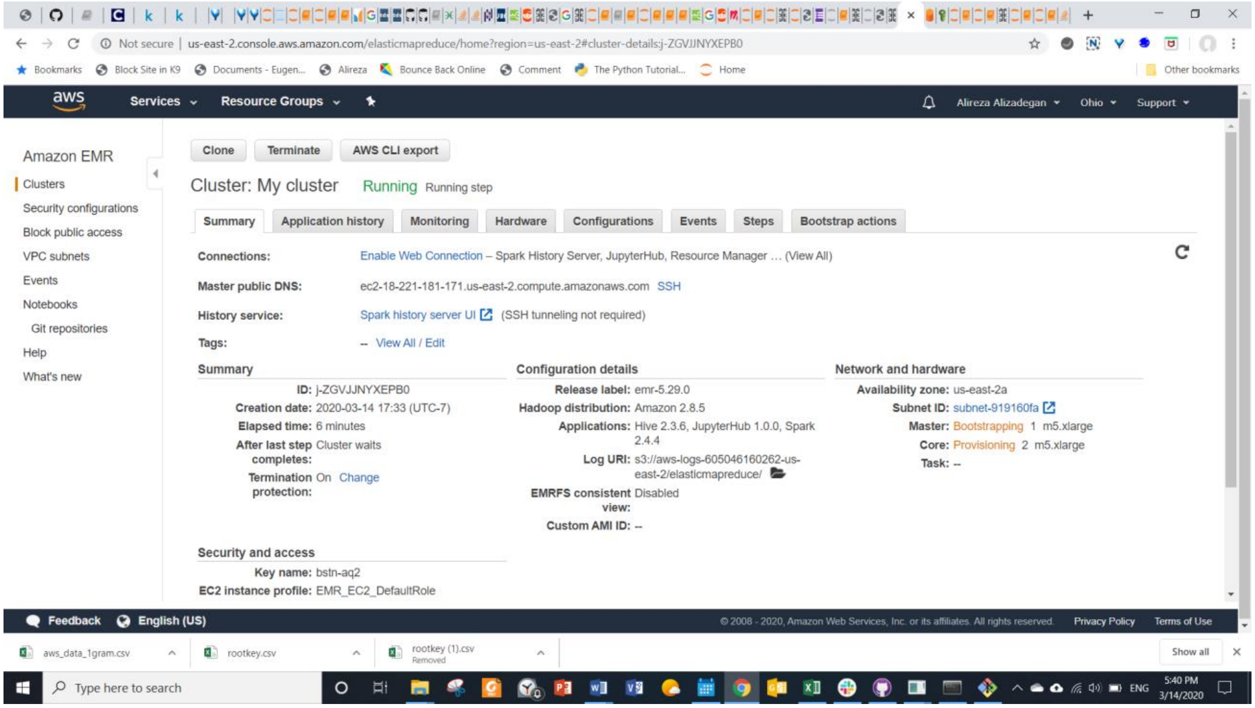
Network and hardware

Availability zone: --
Subnet ID: subnet-919160fa
Master: Provisioning 1 m5.xlarge
Core: Provisioning 2 m5.xlarge
Task: --

Security and access

Key name: bstn-aq2
EC2 instance profile: EMR_EC2_DefaultRole

Wait until status changes to running as below then click on SSH link in master public DNS



Cluster: My cluster Running Running step

Connections: Enable Web Connection – Spark History Server, JupyterHub, Resource Manager … (View All)

Master public DNS: ec2-18-221-181-171.us-east-2.compute.amazonaws.com SSH

History service: Spark history server UI (SSH tunneling not required)

Tags: -- View All / Edit

Summary

ID: j-ZGVJJNYXEPB0
Creation date: 2020-03-14 17:33 (UTC-7)
Elapsed time: 6 minutes
After last step completes: Cluster waits
Termination protection: On Change

Configuration details

Release label: emr-5.29.0
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, JupyterHub 1.0.0, Spark 2.4.4
Log URI: s3://aws-logs-605046160262-us-east-2/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

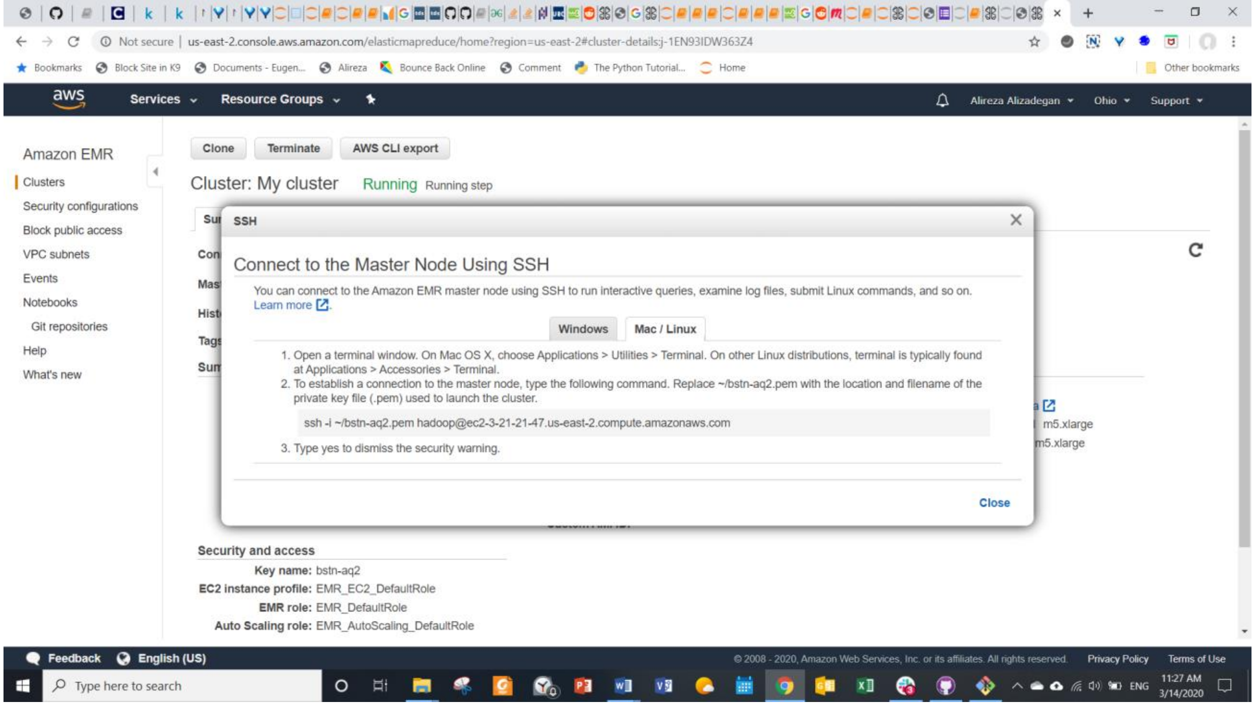
Availability zone: us-east-2a
Subnet ID: subnet-919160fa
Master: Bootstrapping 1 m5.xlarge
Core: Provisioning 2 m5.xlarge
Task: --

Security and access

Key name: bstn-aq2
EC2 instance profile: EMR_EC2_DefaultRole

2 Connect to master node of cluster

Add location of the private key to provided command in step 2 of instructions



Cluster: My cluster Running Running step

Connect to the Master Node Using SSH

You can connect to the Amazon EMR master node using SSH to run interactive queries, examine log files, submit Linux commands, and so on. Learn more.

1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found at Applications > Accessories > Terminal.
2. To establish a connection to the master node, type the following command. Replace ~/bstn-aq2.pem with the location and filename of the private key file (.pem) used to launch the cluster.

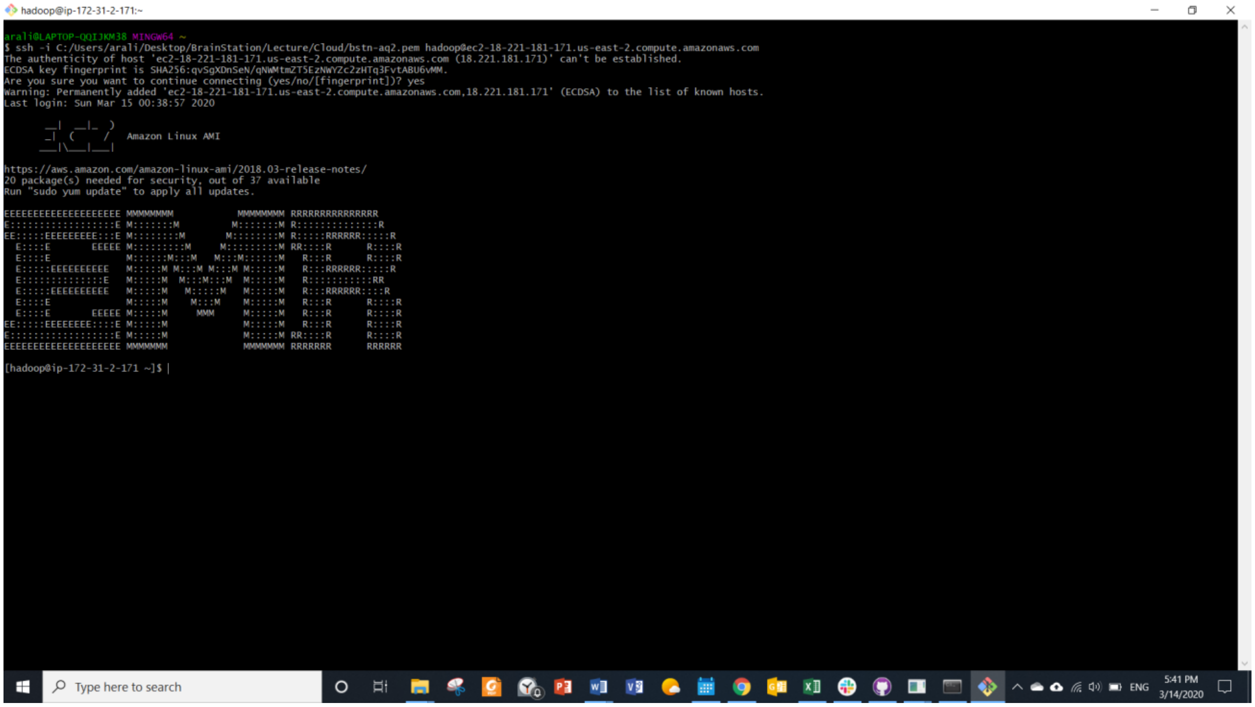
```
ssh -i ~/bstn-aq2.pem hadoop@ec2-3-21-21-47.us-east-2.compute.amazonaws.com
```

3. Type yes to dismiss the security warning.

Security and access

Key name: bstn-aq2
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole

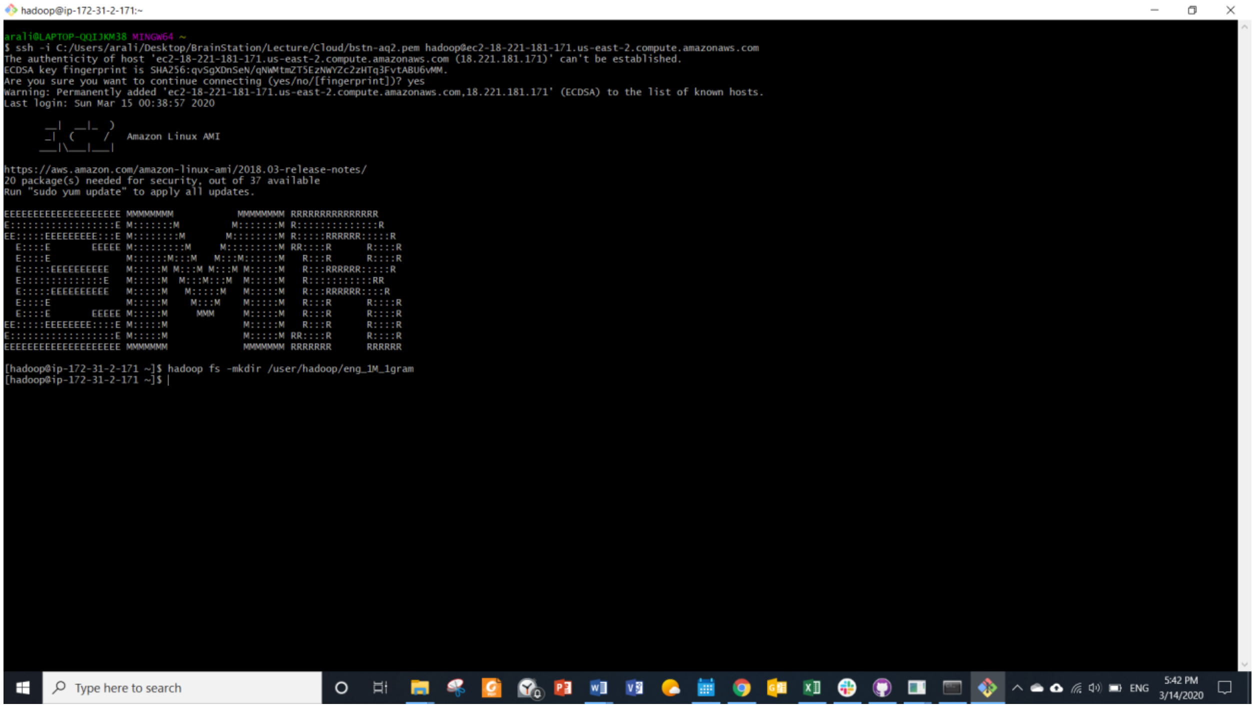
Paste command in git bash and enter Hadoop environment as below



```
arali@LAPTOP-QQIJKM38 MINGW64 ~
$ ssh -i C:/Users/arali/Desktop/BrainStation/Lecture/Cloud/bstn-aq2.pem hadoop@ec2-18-221-181-171.us-east-2.compute.amazonaws.com
The authenticity of host 'ec2-18-221-181-171.us-east-2.compute.amazonaws.com (18.221.181.171)' can't be established.
ECDSA key fingerprint is SHA256:qvSgXDnSeN/qNWMtmZT5EzNWYZc2zHTq3FvtABU6vMM.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-18-221-181-171.us-east-2.compute.amazonaws.com,18.221.181.171' (ECDSA) to the list of known hosts.
Last login: Sun Mar 15 00:38:57 2020

       __|  __|_  )
       _|  (     /   Amazon Linux AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
20 package(s) needed for security, out of 37 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-2-171 ~]$ |
```

3 Copy data from S3 bucket to Hadoop directory

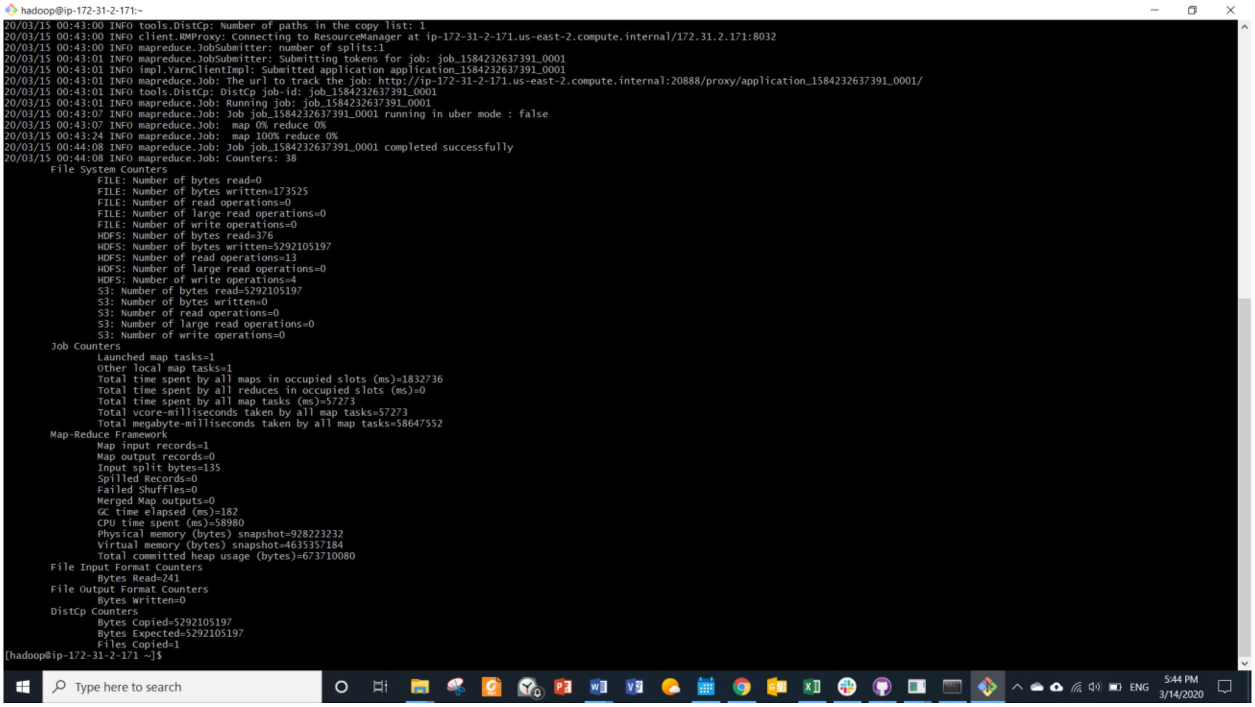
Create a directory in Hadoop environment



```
arali@LAPTOP-QQIJKM38 MINGW64 ~
$ ssh -i C:/Users/arali/Desktop/BrainStation/Lecture/Cloud/bstn-aq2.pem hadoop@ec2-18-221-181-171.us-east-2.compute.amazonaws.com
The authenticity of host 'ec2-18-221-181-171.us-east-2.compute.amazonaws.com (18.221.181.171)' can't be established.
ECDSA key fingerprint is SHA256:qvSgXDnSeN/qNWMtmZT5EzNWYZc2zHTq3FvtABU6vMM.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-18-221-181-171.us-east-2.compute.amazonaws.com,18.221.181.171' (ECDSA) to the list of known hosts.
Last login: Sun Mar 15 00:38:57 2020

       __|  __|_  )
       _|  (     /   Amazon Linux AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
20 package(s) needed for security, out of 37 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-2-171 ~]$ hadoop fs -mkdir /user/hadoop/eng_1M_1gram
[hadoop@ip-172-31-2-171 ~]$ |
```

Copy data from brainstation public repository to directory by pasting following command in git bash

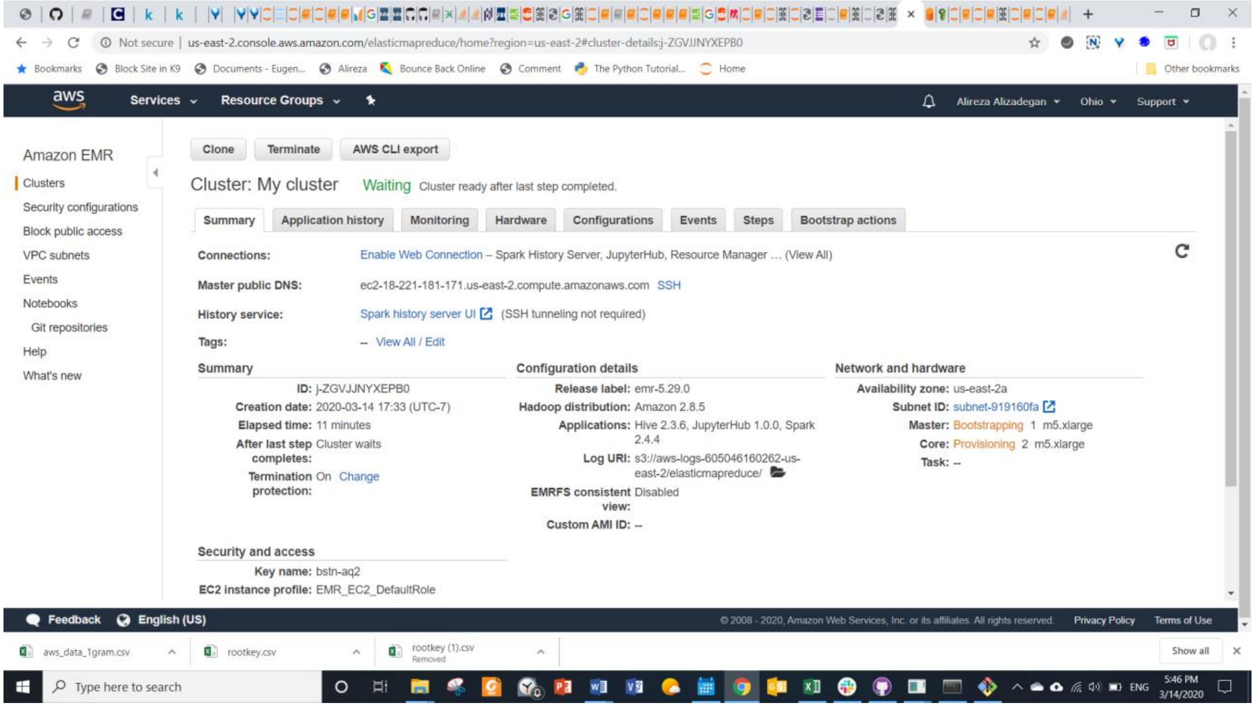
```
hadoop distcp s3://brainstation-dsft/eng_1M_1gram.csv /user/hadoop/eng_1M_1gram
```



```
20/03/15 00:43:00 INFO tools.DistCp: Number of paths in the copy list: 1
20/03/15 00:43:00 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-2-171.us-east-2.compute.internal/172.31.2.171:8032
20/03/15 00:43:00 INFO mapreduce.JobSubmitter: number of splits:1
20/03/15 00:43:01 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1584232637391_0001
20/03/15 00:43:01 INFO impl.YarnClientImpl: Submitted application application_1584232637391_0001
20/03/15 00:43:01 INFO mapreduce.Job: The url to track the job: http://ip-172-31-2-171.us-east-2.compute.internal:20888/proxy/application_1584232637391_0001/
20/03/15 00:43:01 INFO tools.DistCp: DistCp job-id: job_1584232637391_0001
20/03/15 00:43:01 INFO mapreduce.Job: Running job: job_1584232637391_0001
20/03/15 00:43:07 INFO mapreduce.Job: Job job_1584232637391_0001 running in uber mode : false
20/03/15 00:43:07 INFO mapreduce.Job:  map 0% reduce 0%
20/03/15 00:43:24 INFO mapreduce.Job:  map 100% reduce 0%
20/03/15 00:44:08 INFO mapreduce.Job: Job job_1584232637391_0001 completed successfully
20/03/15 00:44:08 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=173525
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=376
                HDFS: Number of bytes written=5292105197
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                S3: Number of bytes read=5292105197
                S3: Number of bytes written=0
                S3: Number of read operations=0
                S3: Number of large read operations=0
                S3: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=1832736
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=57273
                Total vcore-milliseconds taken by all map tasks=57273
                Total megabyte-milliseconds taken by all map tasks=58647552
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=135
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=182
                CPU time spent (ms)=58980
                Physical memory (bytes) snapshot=928223232
                Virtual memory (bytes) snapshot=4635357184
                Total committed heap usage (bytes)=673710080
        File Input Format Counters
                Bytes Read=241
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=5292105197
                Bytes Expected=5292105197
                Files Copied=1
[hadoop@ip-172-31-2-171 ~]$
```

4 Analyze the data with spark

Click enable web connections link in front of connections



Cluster: My cluster Waiting Cluster ready after last step completed.

Connections: Enable Web Connection – Spark History Server, JupyterHub, Resource Manager ... (View All)

Master public DNS: ec2-18-221-181-171.us-east-2.compute.amazonaws.com SSH

History service: Spark history server UI (SSH tunneling not required)

Tags: -- View All / Edit

Summary

ID: j-ZGVJJNYXEPB0
Creation date: 2020-03-14 17:33 (UTC-7)
Elapsed time: 11 minutes
After last step completes: Cluster waits
Termination protection: On Change

Configuration details

Release label: emr-5.29.0
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, JupyterHub 1.0.0, Spark 2.4.4
Log URI: s3://aws-logs-605046160262-us-east-2/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

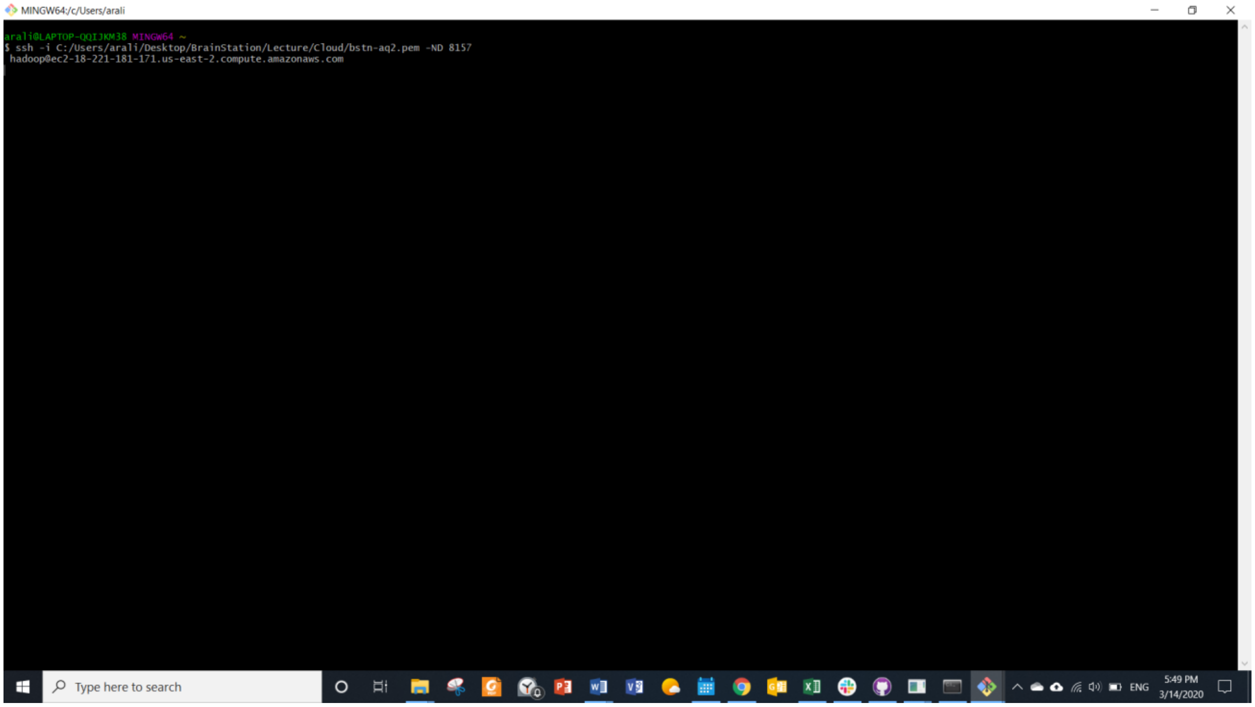
Network and hardware

Availability zone: us-east-2a
Subnet ID: subnet-919160fa
Master: Bootstrapping 1 m5.xlarge
Core: Provisioning 2 m5.xlarge
Task: --

Security and access

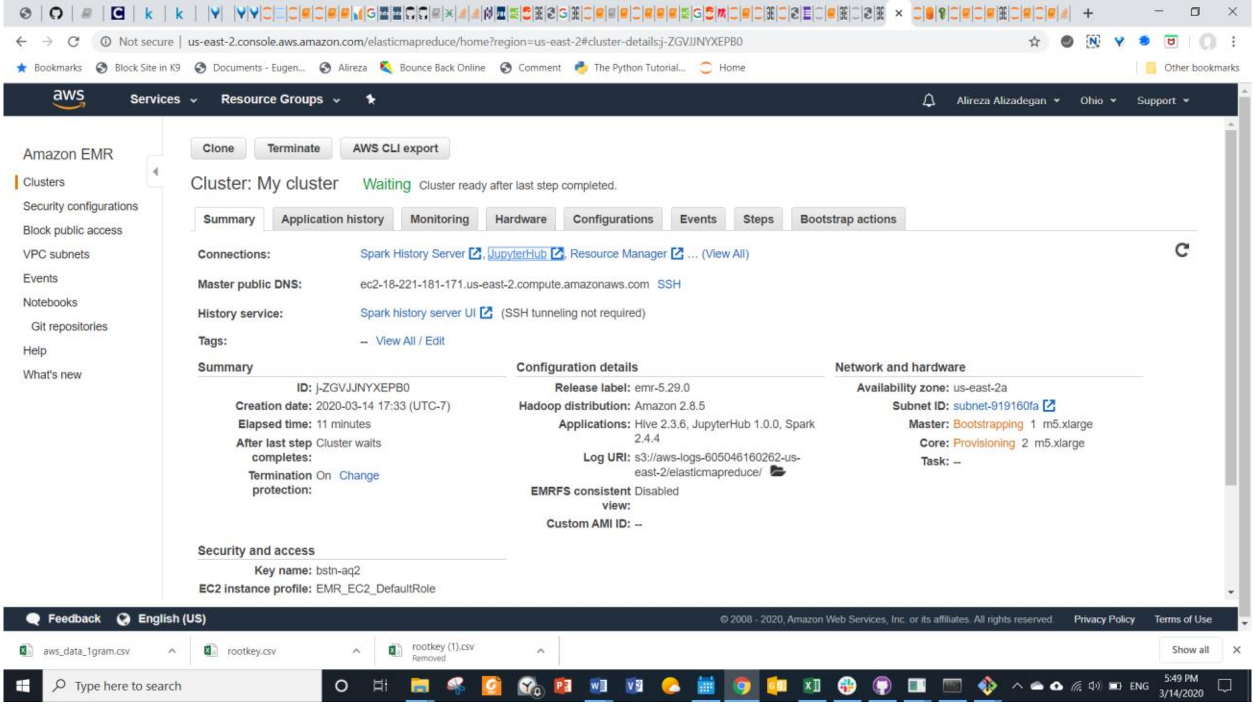
Key name: bstn-aq2
EC2 instance profile: EMR_EC2_DefaultRole

Add location of the private key to provided command in step 2 of instructions and paste it in git bash as below



```
arali@LAPTOP-QQIJKM38 MINGW64 ~
$ ssh -i C:/Users/arali/Desktop/BrainStation/Lecture/Cloud/bstn-aq2.pem -ND 8157
 hadoop@ec2-18-221-181-171.us-east-2.compute.amazonaws.com
```

Click the activated JupyterHub link in connections as below



Cluster: My cluster Waiting Cluster ready after last step completed.

Connections: Spark History Server, JupyterHub, Resource Manager ... (View All)

Master public DNS: ec2-18-221-181-171.us-east-2.compute.amazonaws.com SSH

History service: Spark history server UI (SSH tunneling not required)

Tags: -- View All / Edit

Summary

ID: j-ZGVJJNYXEPB0
Creation date: 2020-03-14 17:33 (UTC-7)
Elapsed time: 11 minutes
After last step completes: Cluster waits
Termination protection: On Change

Configuration details

Release label: emr-5.29.0
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, JupyterHub 1.0.0, Spark 2.4.4
Log URI: s3://aws-logs-605046160262-us-east-2/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

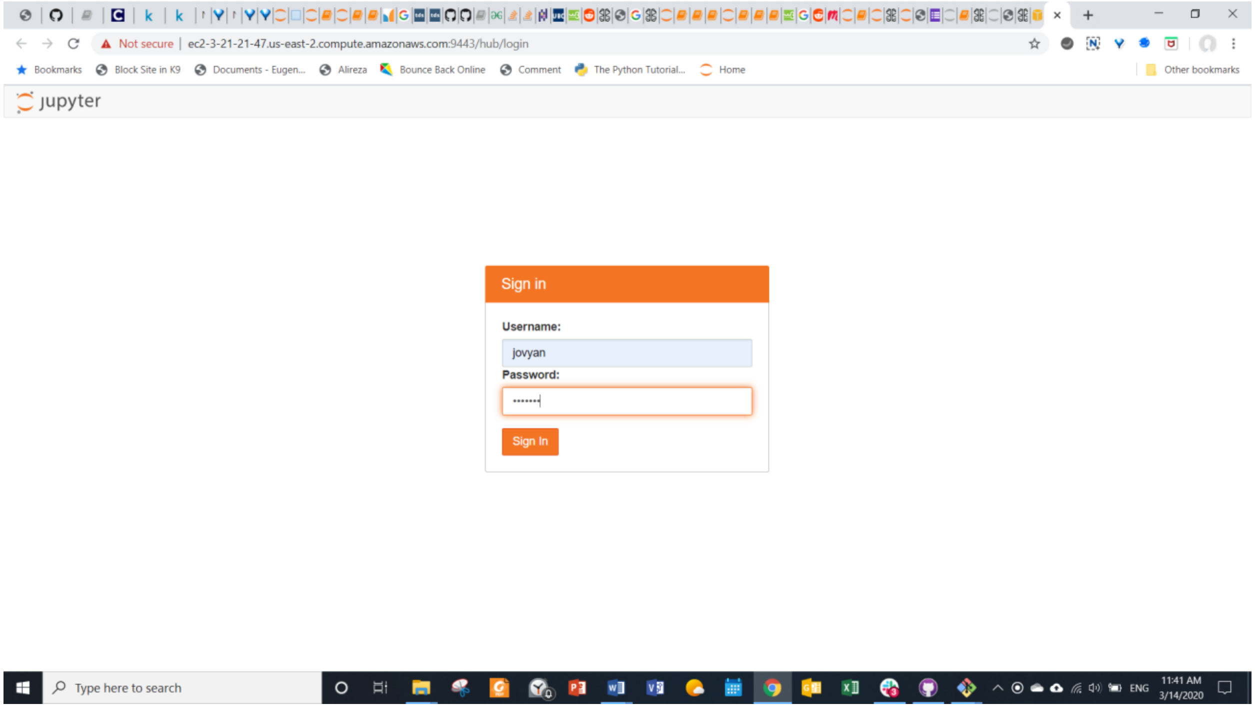
Network and hardware

Availability zone: us-east-2a
Subnet ID: subnet-919160fa
Master: Bootstrapping 1 m5.xlarge
Core: Provisioning 2 m5.xlarge
Task: --

Security and access

Key name: bstn-aq2
EC2 instance profile: EMR_EC2_DefaultRole

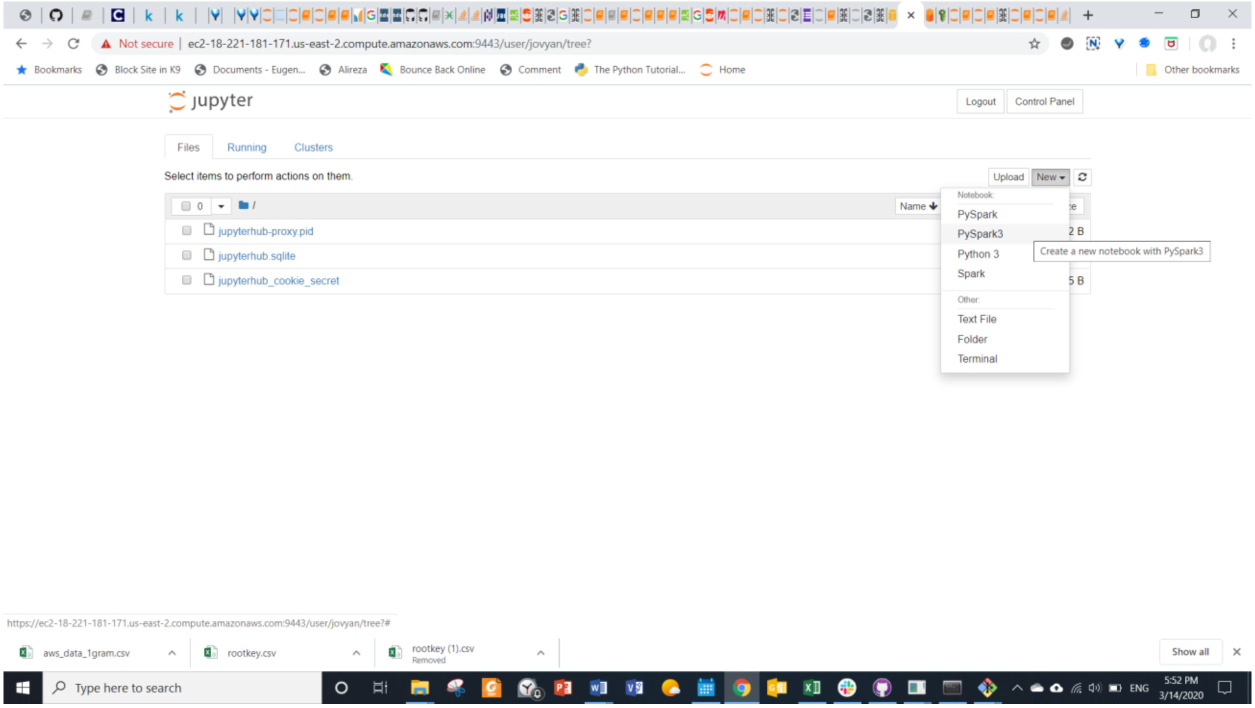
Enter ‘jovyan’ for username and ‘jupyter’ for password and sign in



Username: jovyan

Password: •••••••

Create new pyspark3 notebook and run commands as attached jupyter notebook

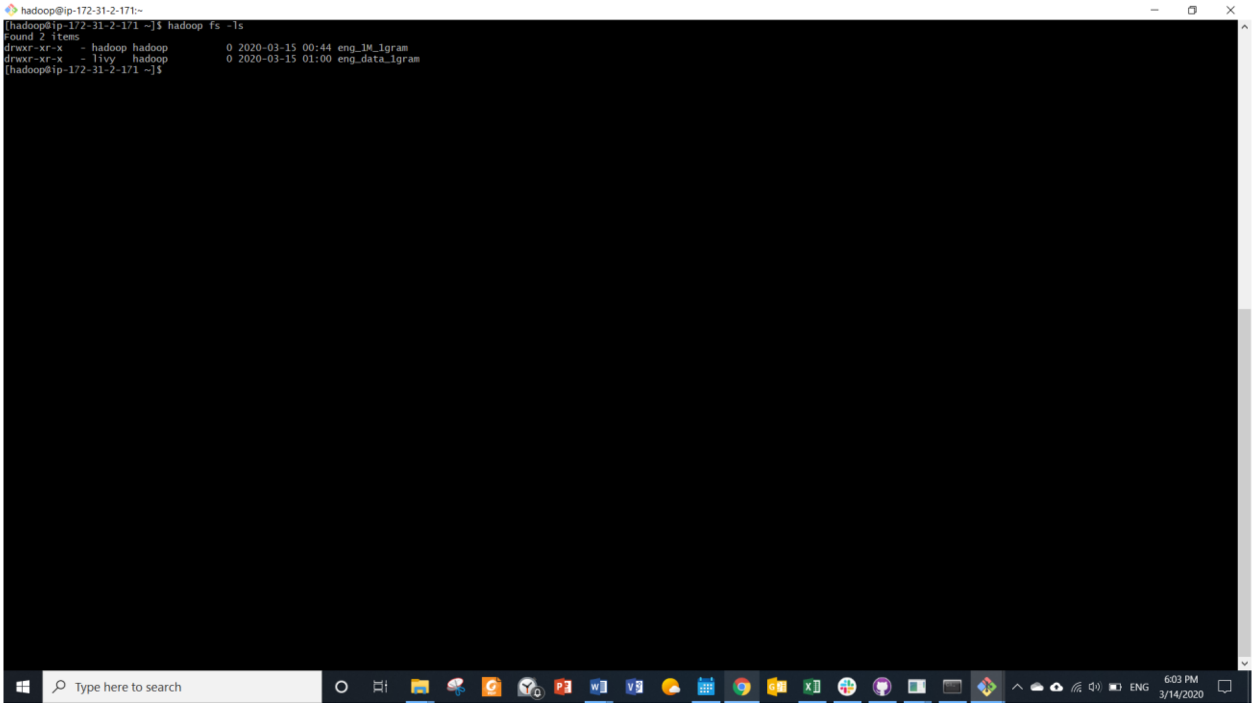


jupyterhub-proxy.pid

jupyterhub.sqlite

jupyterhub_cookie_secret

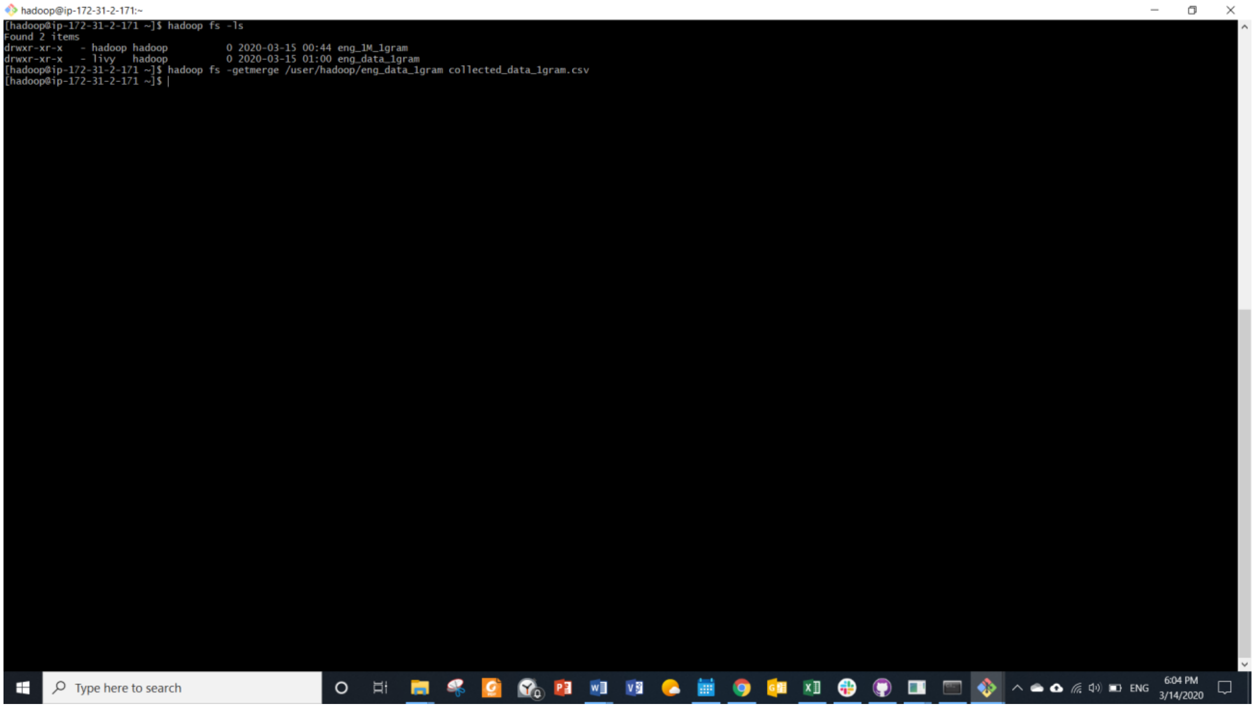
Verify that results of analytics in spark are saved to Hadoop by listing content



```
[hadoop@ip-172-31-2-171 ~]$ hadoop fs -ls
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2020-03-15 00:44 eng_1M_1gram
drwxr-xr-x   - livy   hadoop          0 2020-03-15 01:00 eng_data_1gram
[hadoop@ip-172-31-2-171 ~]$
```

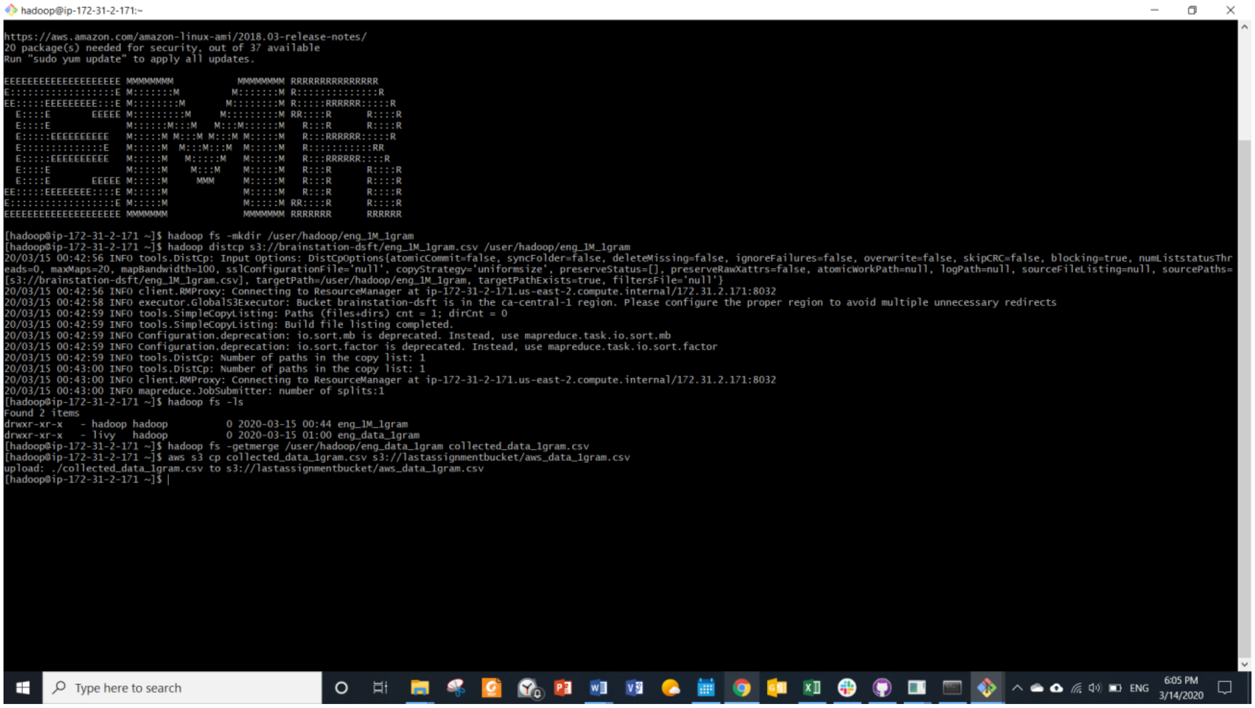
5 Merge data into master node and copy to S3 bucket

Collect the data to master node as a CSV file



```
[hadoop@ip-172-31-2-171 ~]$ hadoop fs -ls
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2020-03-15 00:44 eng_1M_1gram
drwxr-xr-x   - livy   hadoop          0 2020-03-15 01:00 eng_data_1gram
[hadoop@ip-172-31-2-171 ~]$ hadoop fs -getmerge /user/hadoop/eng_data_1gram collected_data_1gram.csv
[hadoop@ip-172-31-2-171 ~]$ |
```

Copy the file from master node to personal AWS S3 bucket ‘lastassignmentbucket’



```
https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
20 package(s) needed for security, out of 37 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-2-171 ~]$ hadoop fs -mkdir /user/hadoop/eng_1M_1gram
[hadoop@ip-172-31-2-171 ~]$ hadoop distcp s3://brainstation-dsft/eng_1M_1gram.csv /user/hadoop/eng_1M_1gram
20/03/15 00:42:56 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numListstatusThr
eads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=
[s3://brainstation-dsft/eng_1M_1gram.csv], targetPath=/user/hadoop/eng_1M_1gram, targetPathExists=true, filtersFile='null'}
20/03/15 00:42:56 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-2-171.us-east-2.compute.internal/172.31.2.171:8032
20/03/15 00:42:58 INFO executor.GlobalS3Executor: Bucket brainstation-dsft is in the ca-central-1 region. Please configure the proper region to avoid multiple unnecessary redirects
20/03/15 00:42:59 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
20/03/15 00:42:59 INFO tools.SimpleCopyListing: Build file listing completed.
20/03/15 00:42:59 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
20/03/15 00:42:59 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
20/03/15 00:42:59 INFO tools.DistCp: Number of paths in the copy list: 1
20/03/15 00:43:00 INFO tools.DistCp: Number of paths in the copy list: 1
20/03/15 00:43:00 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-2-171.us-east-2.compute.internal/172.31.2.171:8032
20/03/15 00:43:00 INFO mapreduce.JobSubmitter: number of splits:1
[hadoop@ip-172-31-2-171 ~]$ hadoop fs -ls
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2020-03-15 00:44 eng_1M_1gram
drwxr-xr-x   - livy   hadoop          0 2020-03-15 01:00 eng_data_1gram
[hadoop@ip-172-31-2-171 ~]$ hadoop fs -getmerge /user/hadoop/eng_data_1gram collected_data_1gram.csv
[hadoop@ip-172-31-2-171 ~]$ aws s3 cp collected_data_1gram.csv s3://lastassignmentbucket/aws_data_1gram.csv
upload: ./collected_data_1gram.csv to s3://lastassignmentbucket/aws_data_1gram.csv
[hadoop@ip-172-31-2-171 ~]$ |
```