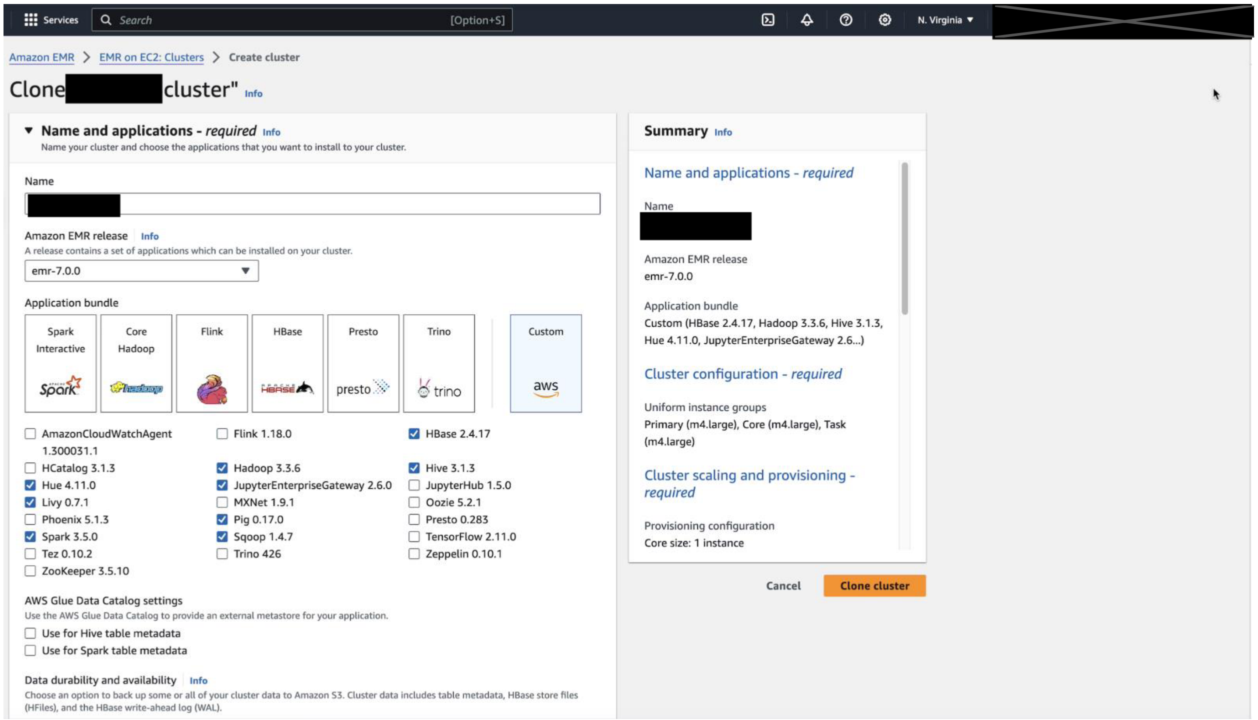
CLONING THE CLUSTER:



[redacted]

Amazon EMR > EMR on EC2: Clusters > Create cluster

Clone [redacted] cluster" Info

Name and applications - *required* Info

Name your cluster and choose the applications that you want to install to your cluster.

Name

[redacted]

Amazon EMR release Info

A release contains a set of applications which can be installed on your cluster.

emr-7.0.0

Application bundle

- ☐ 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
- ☐ Use for Spark table metadata

Data durability and availability Info

Choose an option to back up some or all of your cluster data to Amazon S3. Cluster data includes table metadata, HBase store files (HFiles), and the HBase write-ahead log (WAL).

Summary Info

Name and applications - required

Name

[redacted]

Amazon EMR release

emr-7.0.0

Application bundle

Custom (HBase 2.4.17, Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2.6...)

Cluster configuration - required

Uniform instance groups

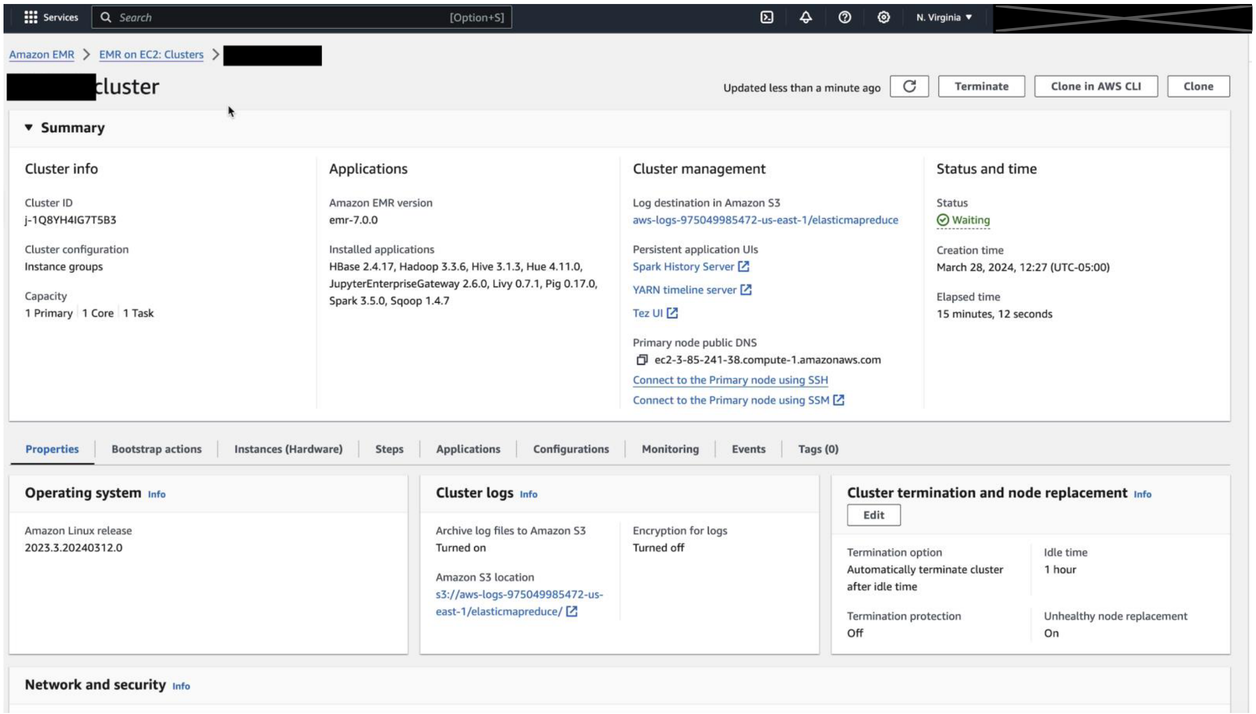
Primary (m4.large), Core (m4.large), Task (m4.large)

Cluster scaling and provisioning - required

Provisioning configuration

Core size: 1 instance

STARTED & WAITED:



[redacted]

Amazon EMR > EMR on EC2: Clusters > [redacted]

[redacted] cluster

Updated less than a minute ago

Summary

Cluster info

Cluster ID

j-1Q8YH4IG7T5B3

Cluster configuration

Instance groups

Capacity

1 Primary | 1 Core | 1 Task

Applications

Amazon EMR version

emr-7.0.0

Installed applications

HBase 2.4.17, Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2.6.0, Livy 0.7.1, Pig 0.17.0, Spark 3.5.0, Sqoop 1.4.7

Cluster management

Log destination in Amazon S3

aws-logs-975049985472-us-east-1/elasticmapreduce

Persistent application UIs

Spark History Server

YARN timeline server

Tez UI

Primary node public DNS

ec2-3-85-241-38.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 28, 2024, 12:27 (UTC-05:00)

Elapsed time

15 minutes, 12 seconds

Operating system Info

Amazon Linux release

2023.3.20240312.0

Cluster logs Info

Archive log files to Amazon S3

Turned on

Encryption for logs

Turned off

Amazon S3 location

s3://aws-logs-975049985472-us-east-1/elasticmapreduce/

Cluster termination and node replacement Info

Termination option

Automatically terminate cluster after idle time

Idle time

1 hour

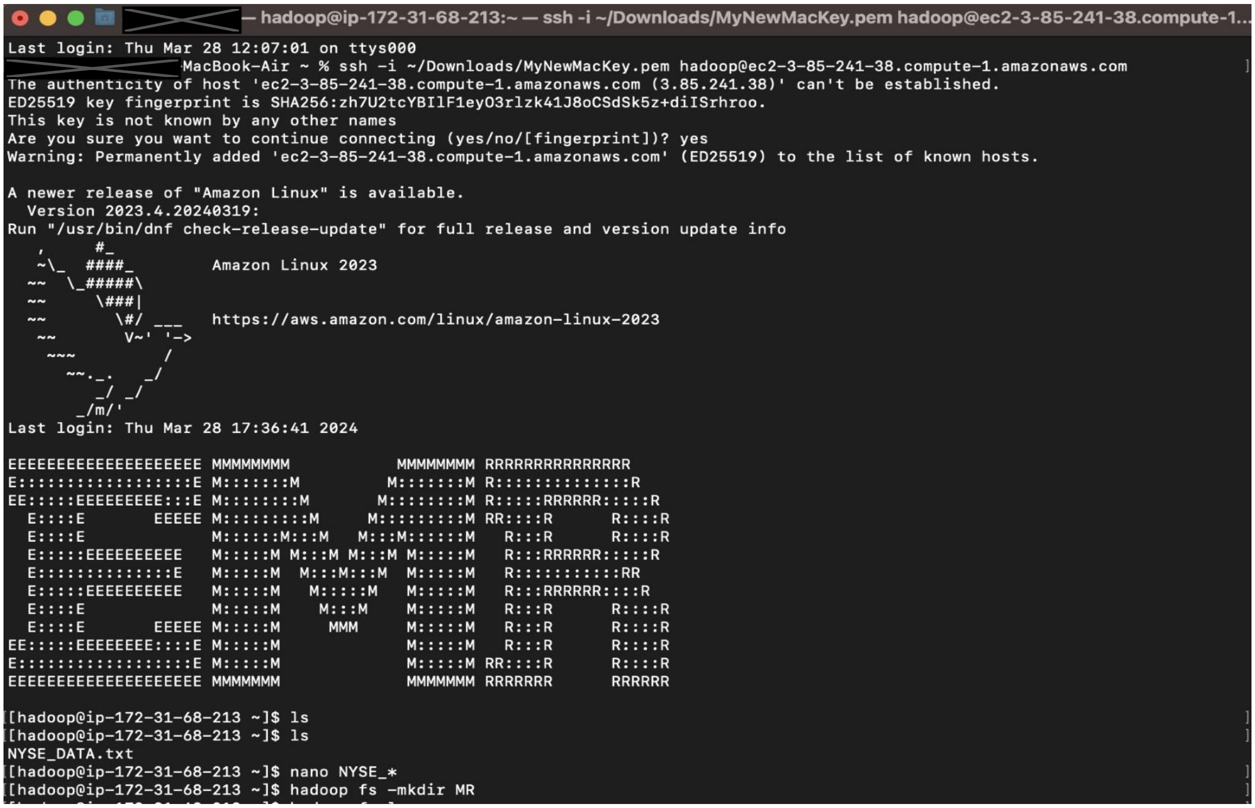
Termination protection

Off

Unhealthy node replacement

On

Network and security Info



[redacted] — hadoop@ip-172-31-68-213:~ — ssh -i ~/Downloads/MyNewMacKey.pem hadoop@ec2-3-85-241-38.compute-1...

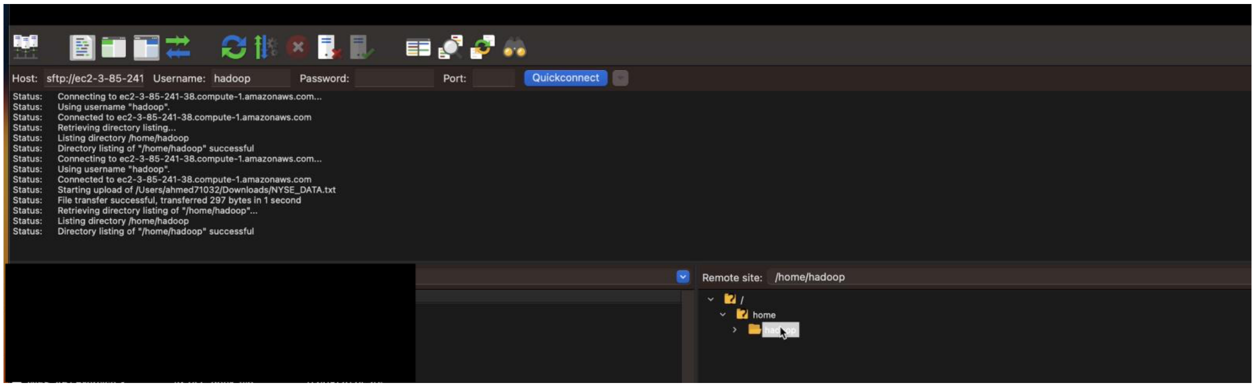
```
Last login: Thu Mar 28 12:07:01 on ttys000
[redacted]MacBook-Air ~ % ssh -i ~/Downloads/MyNewMacKey.pem hadoop@ec2-3-85-241-38.compute-1.amazonaws.com
The authenticity of host 'ec2-3-85-241-38.compute-1.amazonaws.com (3.85.241.38)' can't be established.
ED25519 key fingerprint is SHA256:zh7U2tcYBIlF1eyO3rlzk41J8oCSdSk5z+diISrhroo.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-3-85-241-38.compute-1.amazonaws.com' (ED25519) to the list of known hosts.

A newer release of "Amazon Linux" is available.
  Version 2023.4.20240319:
Run "/usr/bin/dnf check-release-update" for full release and version update info
   ,     #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'
Last login: Thu Mar 28 17:36:41 2024

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-172-31-68-213 ~]$ ls
[hadoop@ip-172-31-68-213 ~]$ ls
NYSE_DATA.txt
[hadoop@ip-172-31-68-213 ~]$ nano NYSE_*
[hadoop@ip-172-31-68-213 ~]$ hadoop fs -mkdir MR
```

FILE ZILLA STEPS:

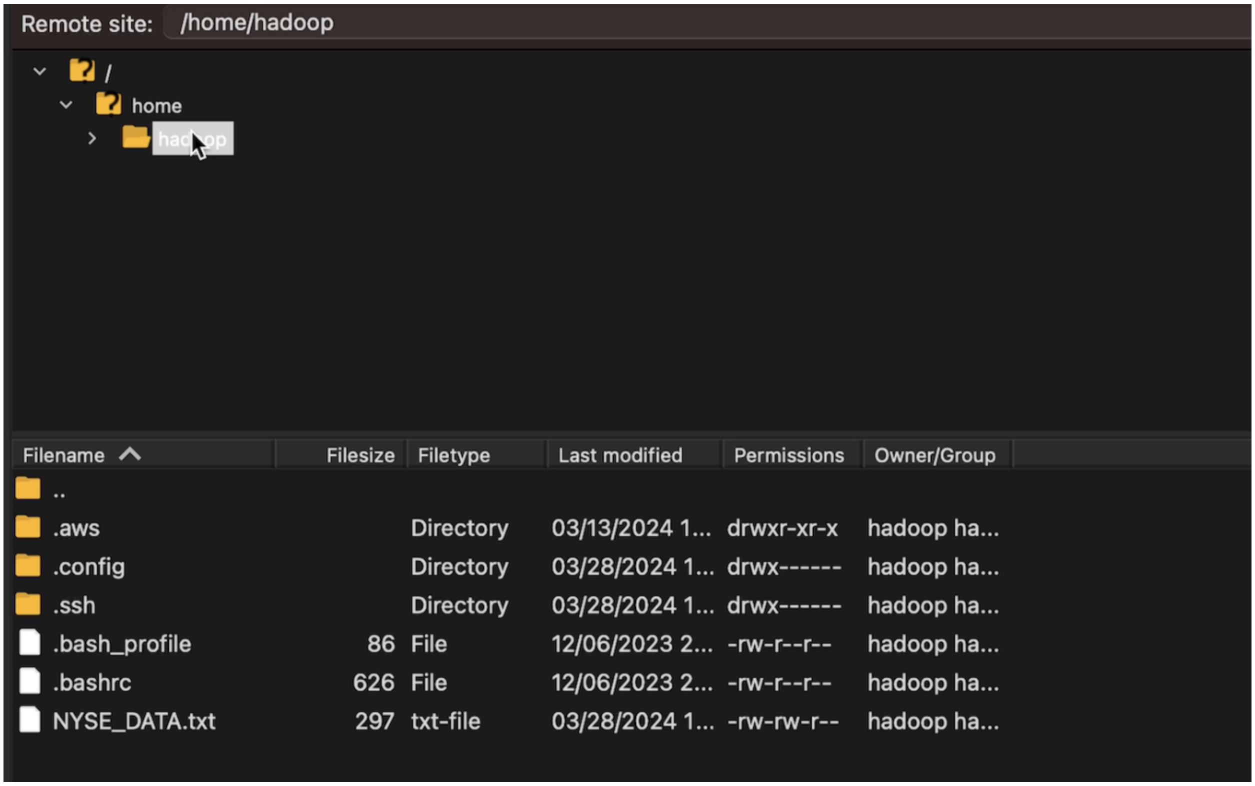


Host: sftp://ec2-3-85-241 Username: hadoop Password: Port: Quickconnect

```
Status:    Connecting to ec2-3-85-241-38.compute-1.amazonaws.com...
Status:    Using username "hadoop".
Status:    Connected to ec2-3-85-241-38.compute-1.amazonaws.com
Status:    Retrieving directory listing...
Status:    Listing directory /home/hadoop
Status:    Directory listing of "/home/hadoop" successful
Status:    Connecting to ec2-3-85-241-38.compute-1.amazonaws.com...
Status:    Using username "hadoop".
Status:    Connected to ec2-3-85-241-38.compute-1.amazonaws.com
Status:    Starting upload of /Users/ahmed71032/Downloads/NYSE_DATA.txt
Status:    File transfer successful, transferred 297 bytes in 1 second
Status:    Retrieving directory listing of "/home/hadoop"...
Status:    Listing directory /home/hadoop
Status:    Directory listing of "/home/hadoop" successful
```

[redacted]

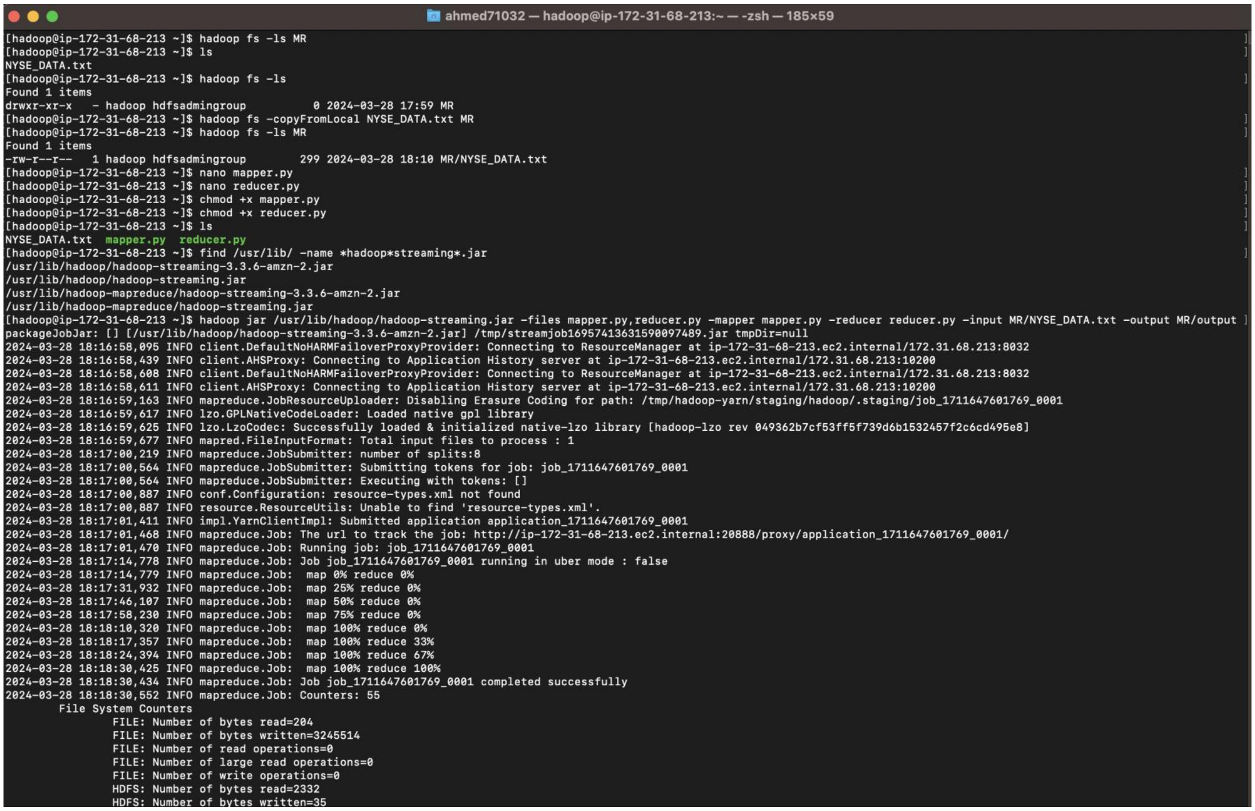
Remote site: /home/hadoop



Remote site: /home/hadoop

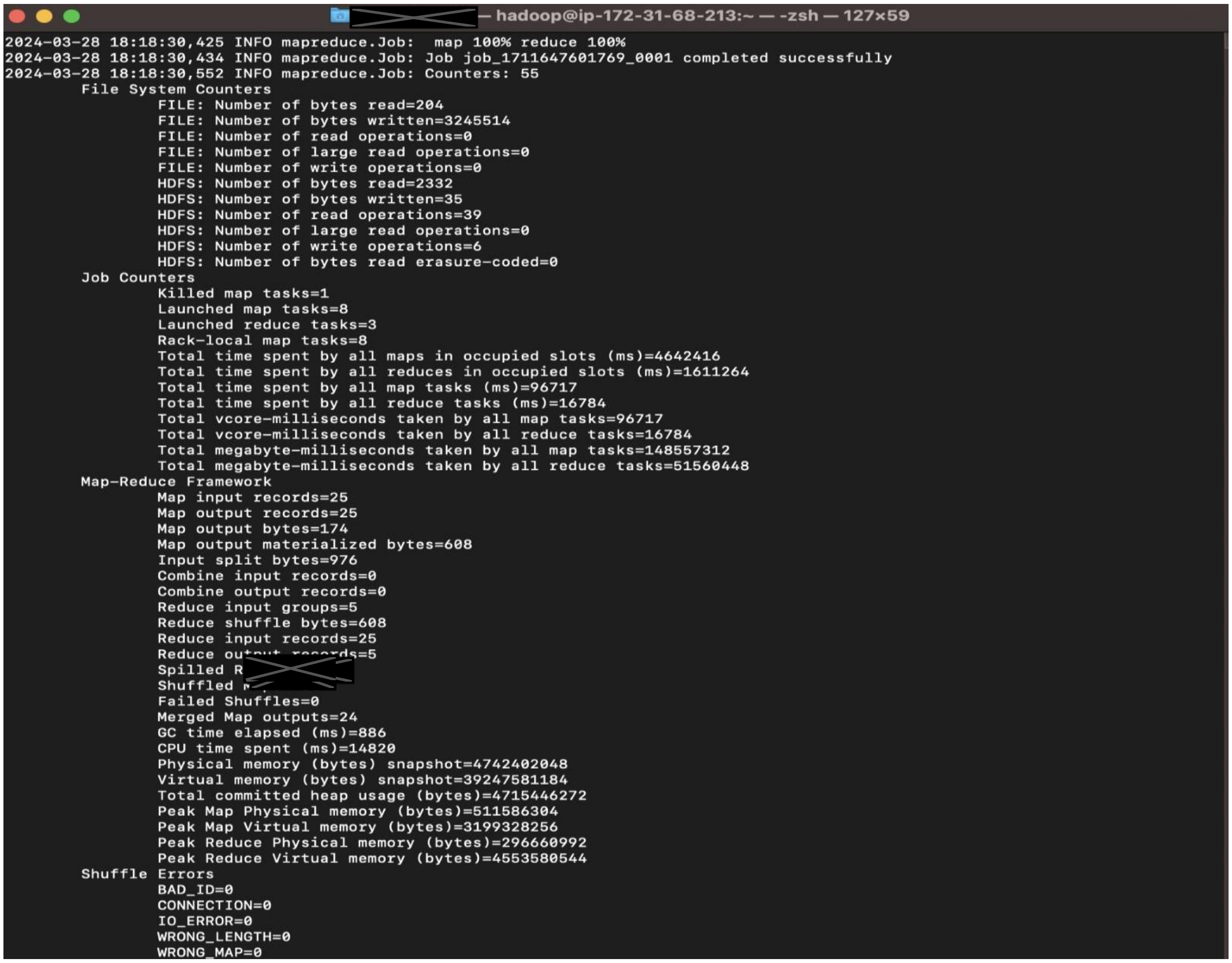
Filename	Filesize	Filetype	Last modified	Permissions	Owner/Group
..					
.aws		Directory	03/13/2024 1...	drwxr-xr-x	hadoop ha...
.config		Directory	03/28/2024 1...	drwx------	hadoop ha...
.ssh		Directory	03/28/2024 1...	drwx------	hadoop ha...
.bash_profile	86	File	12/06/2023 2...	-rw-r--r--	hadoop ha...
.bashrc	626	File	12/06/2023 2...	-rw-r--r--	hadoop ha...
NYSE_DATA.txt	297	txt-file	03/28/2024 1...	-rw-rw-r--	hadoop ha...

CREATING DIRECTORY AND MAKING EDITS TO THE MAPPER & REDUCER PYTHON FILES AND EXECUTING IT:



ahmed71032 — hadoop@ip-172-31-68-213:~ — -zsh — 185×59

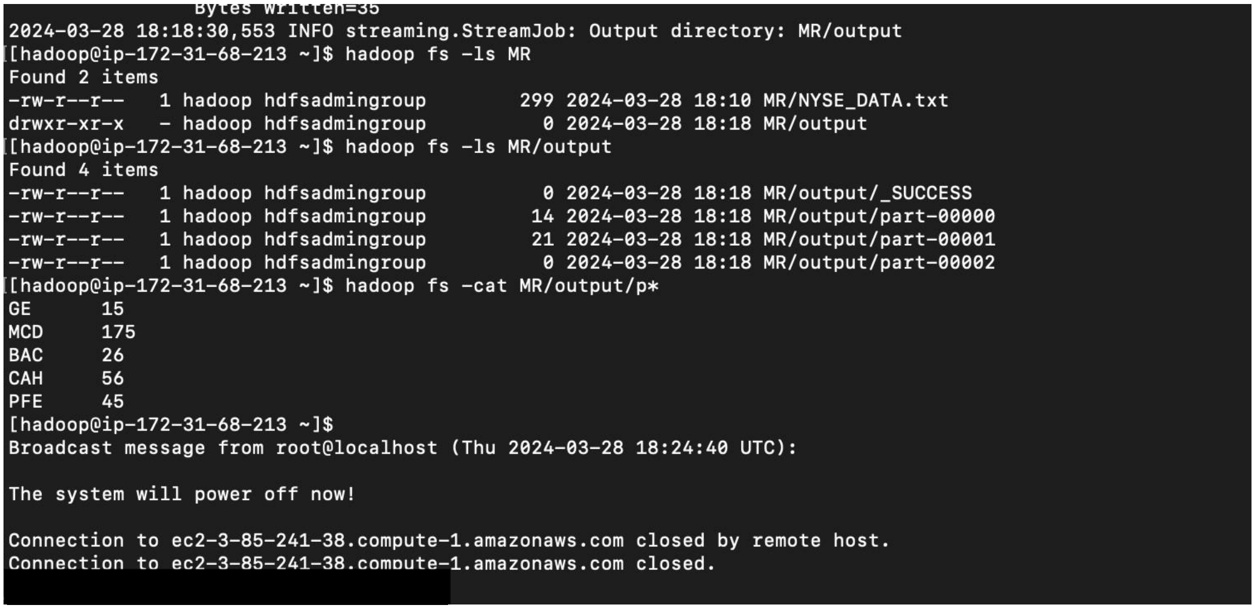
```
[hadoop@ip-172-31-68-213 ~]$ hadoop fs -ls MR
[hadoop@ip-172-31-68-213 ~]$ ls
NYSE_DATA.txt
[hadoop@ip-172-31-68-213 ~]$ hadoop fs -ls
Found 1 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2024-03-28 17:59 MR
[hadoop@ip-172-31-68-213 ~]$ hadoop fs -copyFromLocal NYSE_DATA.txt MR
[hadoop@ip-172-31-68-213 ~]$ hadoop fs -ls MR
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup        299 2024-03-28 18:10 MR/NYSE_DATA.txt
[hadoop@ip-172-31-68-213 ~]$ nano mapper.py
[hadoop@ip-172-31-68-213 ~]$ nano reducer.py
[hadoop@ip-172-31-68-213 ~]$ chmod +x mapper.py
[hadoop@ip-172-31-68-213 ~]$ chmod +x reducer.py
[hadoop@ip-172-31-68-213 ~]$ ls
NYSE_DATA.txt  mapper.py  reducer.py
[hadoop@ip-172-31-68-213 ~]$ find /usr/lib/ -name *hadoop*streaming*.jar
/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-2.jar
/usr/lib/hadoop/hadoop-streaming.jar
/usr/lib/hadoop-mapreduce/hadoop-streaming-3.3.6-amzn-2.jar
/usr/lib/hadoop-mapreduce/hadoop-streaming.jar
[hadoop@ip-172-31-68-213 ~]$ hadoop jar /usr/lib/hadoop/hadoop-streaming.jar -files mapper.py,reducer.py -mapper mapper.py -reducer reducer.py -input MR/NYSE_DATA.txt -output MR/output
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6-amzn-2.jar] /tmp/streamjob16957413631590097489.jar tmpDir=null
2024-03-28 18:16:58,095 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at ip-172-31-68-213.ec2.internal/172.31.68.213:8032
2024-03-28 18:16:58,439 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-68-213.ec2.internal/172.31.68.213:10200
2024-03-28 18:16:58,608 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at ip-172-31-68-213.ec2.internal/172.31.68.213:8032
2024-03-28 18:16:58,611 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-68-213.ec2.internal/172.31.68.213:10200
2024-03-28 18:16:59,163 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1711647601769_0001
2024-03-28 18:16:59,617 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
2024-03-28 18:16:59,625 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
2024-03-28 18:16:59,677 INFO mapred.FileInputFormat: Total input files to process : 1
2024-03-28 18:17:00,219 INFO mapreduce.JobSubmitter: number of splits:8
2024-03-28 18:17:00,564 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1711647601769_0001
2024-03-28 18:17:00,564 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-03-28 18:17:00,887 INFO conf.Configuration: resource-types.xml not found
2024-03-28 18:17:00,887 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-03-28 18:17:01,411 INFO impl.YarnClientImpl: Submitted application application_1711647601769_0001
2024-03-28 18:17:01,468 INFO mapreduce.Job: The url to track the job: http://ip-172-31-68-213.ec2.internal:20888/proxy/application_1711647601769_0001/
2024-03-28 18:17:01,470 INFO mapreduce.Job: Running job: job_1711647601769_0001
2024-03-28 18:17:14,778 INFO mapreduce.Job: Job job_1711647601769_0001 running in uber mode : false
2024-03-28 18:17:14,779 INFO mapreduce.Job:  map 0% reduce 0%
2024-03-28 18:17:31,932 INFO mapreduce.Job:  map 25% reduce 0%
2024-03-28 18:17:46,107 INFO mapreduce.Job:  map 50% reduce 0%
2024-03-28 18:17:58,230 INFO mapreduce.Job:  map 75% reduce 0%
2024-03-28 18:18:10,320 INFO mapreduce.Job:  map 100% reduce 0%
2024-03-28 18:18:17,357 INFO mapreduce.Job:  map 100% reduce 33%
2024-03-28 18:18:24,394 INFO mapreduce.Job:  map 100% reduce 67%
2024-03-28 18:18:30,425 INFO mapreduce.Job:  map 100% reduce 100%
2024-03-28 18:18:30,434 INFO mapreduce.Job: Job job_1711647601769_0001 completed successfully
2024-03-28 18:18:30,552 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=204
                FILE: Number of bytes written=3245514
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2332
                HDFS: Number of bytes written=35
```



[redacted] — hadoop@ip-172-31-68-213:~ — -zsh — 127×59

```
2024-03-28 18:18:30,425 INFO mapreduce.Job:  map 100% reduce 100%
2024-03-28 18:18:30,434 INFO mapreduce.Job: Job job_1711647601769_0001 completed successfully
2024-03-28 18:18:30,552 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=204
                FILE: Number of bytes written=3245514
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2332
                HDFS: Number of bytes written=35
                HDFS: Number of read operations=39
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=6
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=8
                Launched reduce tasks=3
                Rack-local map tasks=8
                Total time spent by all maps in occupied slots (ms)=4642416
                Total time spent by all reduces in occupied slots (ms)=1611264
                Total time spent by all map tasks (ms)=96717
                Total time spent by all reduce tasks (ms)=16784
                Total vcore-milliseconds taken by all map tasks=96717
                Total vcore-milliseconds taken by all reduce tasks=16784
                Total megabyte-milliseconds taken by all map tasks=148557312
                Total megabyte-milliseconds taken by all reduce tasks=51560448
        Map-Reduce Framework
                Map input records=25
                Map output records=25
                Map output bytes=174
                Map output materialized bytes=608
                Input split bytes=976
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=608
                Reduce input records=25
                Reduce output records=5
                Spilled R[redacted]
                Shuffled
                Failed Shuffles=0
                Merged Map outputs=24
                GC time elapsed (ms)=886
                CPU time spent (ms)=14820
                Physical memory (bytes) snapshot=4742402048
                Virtual memory (bytes) snapshot=39247581184
                Total committed heap usage (bytes)=4715446272
                Peak Map Physical memory (bytes)=511586304
                Peak Map Virtual memory (bytes)=3199328256
                Peak Reduce Physical memory (bytes)=296660992
                Peak Reduce Virtual memory (bytes)=4553580544
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
```

FINAL RESULT:

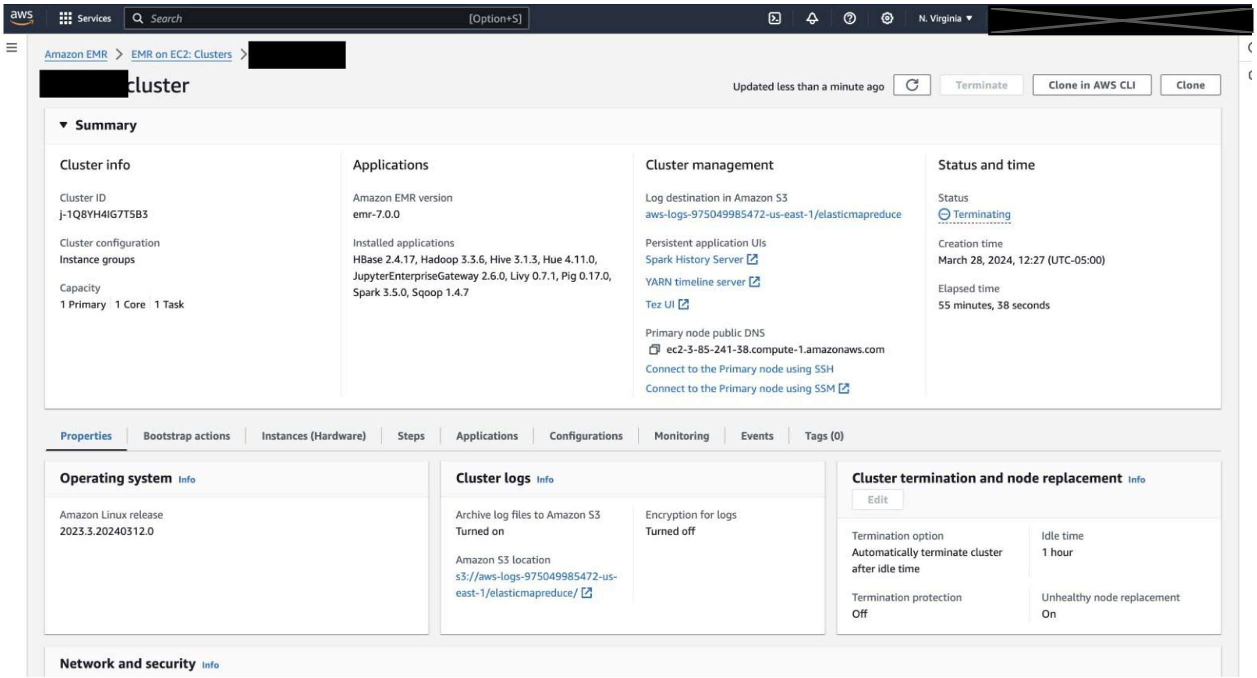


```
                Bytes Written=35
2024-03-28 18:18:30,553 INFO streaming.StreamJob: Output directory: MR/output
[hadoop@ip-172-31-68-213 ~]$ hadoop fs -ls MR
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup        299 2024-03-28 18:10 MR/NYSE_DATA.txt
drwxr-xr-x   - hadoop hdfsadmingroup          0 2024-03-28 18:18 MR/output
[hadoop@ip-172-31-68-213 ~]$ hadoop fs -ls MR/output
Found 4 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2024-03-28 18:18 MR/output/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup         14 2024-03-28 18:18 MR/output/part-00000
-rw-r--r--   1 hadoop hdfsadmingroup         21 2024-03-28 18:18 MR/output/part-00001
-rw-r--r--   1 hadoop hdfsadmingroup          0 2024-03-28 18:18 MR/output/part-00002
[hadoop@ip-172-31-68-213 ~]$ hadoop fs -cat MR/output/p*
GE      15
MCD     175
BAC     26
CAH     56
PFE     45
[hadoop@ip-172-31-68-213 ~]$
Broadcast message from root@localhost (Thu 2024-03-28 18:24:40 UTC):

The system will power off now!

Connection to ec2-3-85-241-38.compute-1.amazonaws.com closed by remote host.
Connection to ec2-3-85-241-38.compute-1.amazonaws.com closed.
[redacted]
```

TERMINATING THE CLUSTER & ENDING THE LAB:



[redacted] [redacted]

[redacted]cluster

▼ Summary

Cluster info

Cluster ID
j-1Q8YH4IG7T5B3

Cluster configuration
Instance groups

Capacity
1 Primary | 1 Core | 1 Task

Applications

Amazon EMR version
emr-7.0.0

Installed applications
HBase 2.4.17, Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2.6.0, Livy 0.7.1, Pig 0.17.0, Spark 3.5.0, Sqoop 1.4.7

Cluster management

Log destination in Amazon S3
aws-logs-975049985472-us-east-1/elasticmapreduce

Persistent application UIs
Spark History Server
YARN timeline server
Tez UI

Primary node public DNS
ec2-3-85-241-38.compute-1.amazonaws.com
Connect to the Primary node using SSH
Connect to the Primary node using SSM

Status and time

Status
⊖ Terminating

Creation time
March 28, 2024, 12:27 (UTC-05:00)

Elapsed time
55 minutes, 38 seconds

Operating system Info

Amazon Linux release
2023.3.20240312.0

Cluster logs Info

Archive log files to Amazon S3
Turned on

Amazon S3 location
s3://aws-logs-975049985472-us-east-1/elasticmapreduce/

Encryption for logs
Turned off

Cluster termination and node replacement Info

Termination option
Automatically terminate cluster after idle time

Idle time
1 hour

Termination protection
Off

Unhealthy node replacement
On

Network and security Info