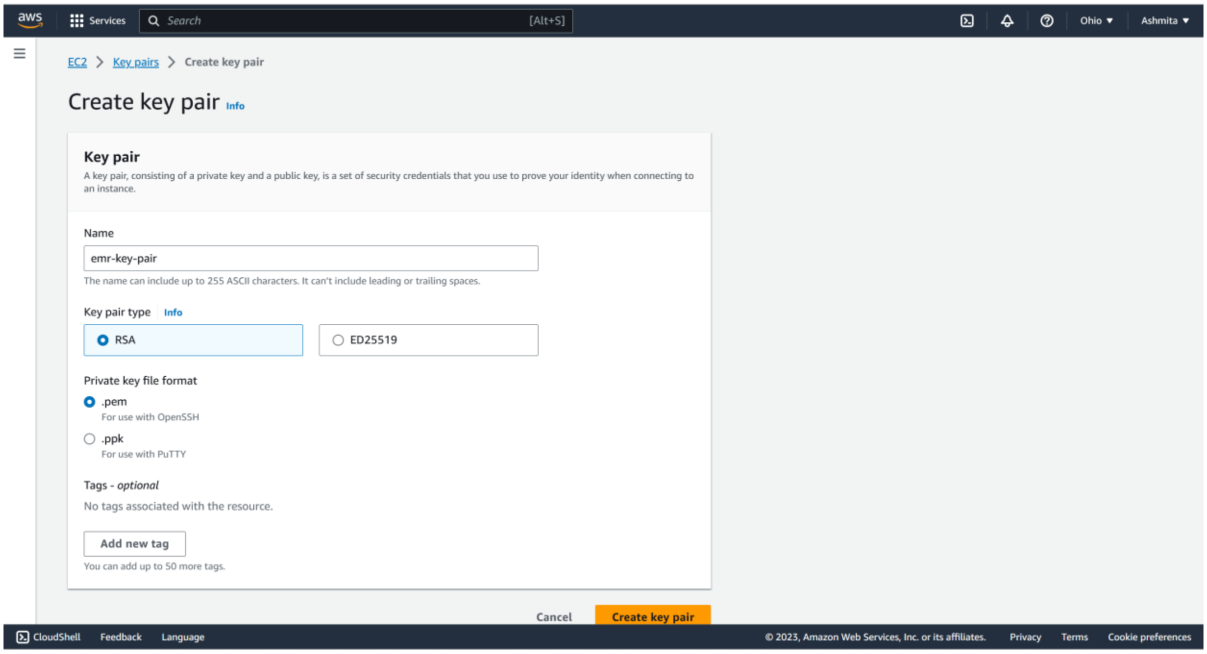
CSP 554

HOMEWORK 2

ASHMITA GUPTA (A20512498)

1. Key Pair using the EC2 service



Create key pair Info

Key pair

A key pair, consisting of a private key and a public key, is a set of security credentials that you use to prove your identity when connecting to an instance.

Name

emr-key-pair

The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.

Key pair type Info

● RSA ○ ED25519

Private key file format

● .pem
For use with OpenSSH

○ .ppk
For use with PuTTY

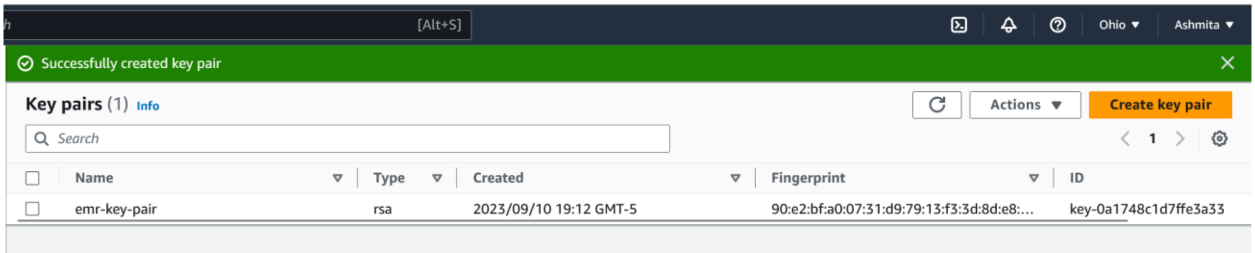
Tags - *optional*

No tags associated with the resource.

Add new tag

You can add up to 50 more tags.

2. Key pair successfully created

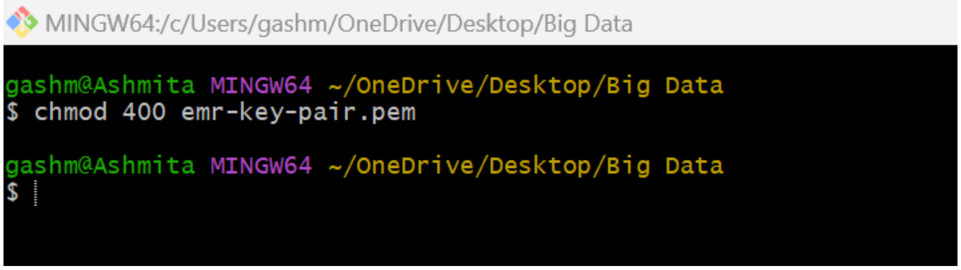


Successfully created key pair

Key pairs (1) Info

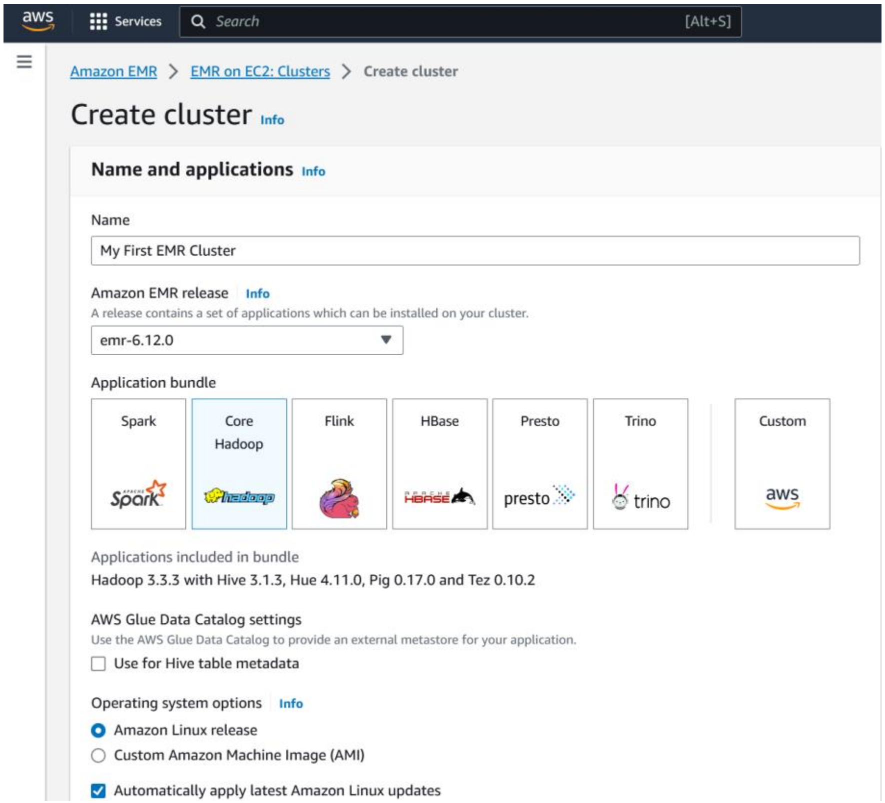
□	Name	Type	Created	Fingerprint	ID
□	emr-key-pair	rsa	2023/09/10 19:12 GMT-5	90:e2:bf:a0:07:31:d9:79:13:f3:3d:8d:e8:...	key-0a1748c1d7ffe3a33

3. Modifying permission for key pair



```
gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ chmod 400 emr-key-pair.pem

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$
```

4. Creating Amazon EMR cluster by using steps given in the instruction



Create cluster Info

Name and applications Info

Name

My First EMR Cluster

Amazon EMR release Info

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

emr-6.12.0

Application bundle

Applications included in bundle

Hadoop 3.3.3 with Hive 3.1.3, Hue 4.11.0, Pig 0.17.0 and Tez 0.10.2

AWS Glue Data Catalog settings

Use the AWS Glue Data Catalog to provide an external metastore for your application.

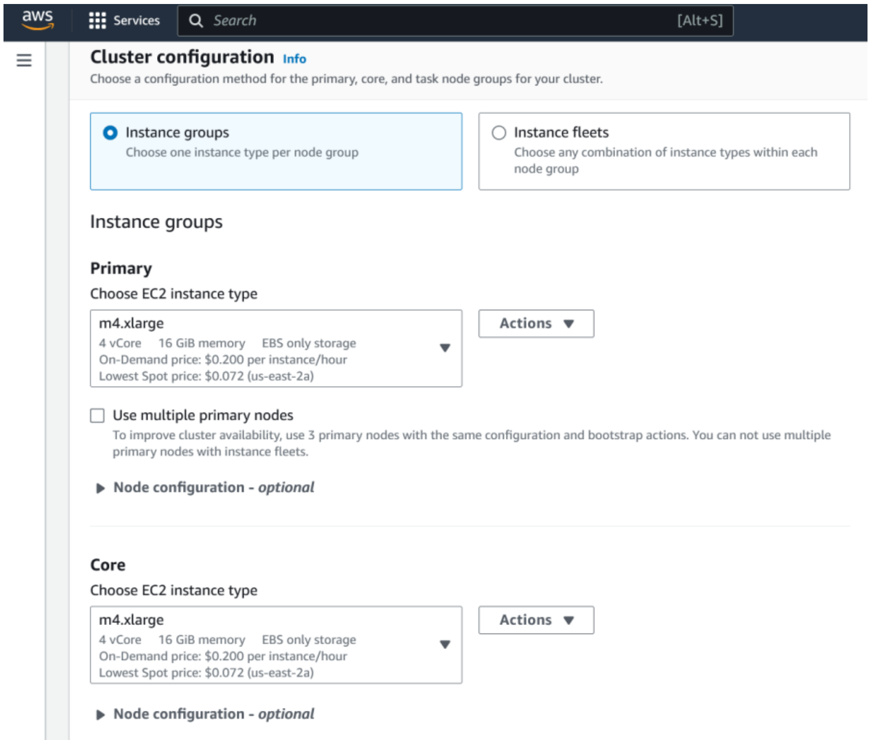
☐ Use for Hive table metadata

Operating system options Info

● Amazon Linux release

○ Custom Amazon Machine Image (AMI)

☑ Automatically apply latest Amazon Linux updates



Cluster configuration Info

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

● Instance groups
Choose one instance type per node group

○ Instance fleets
Choose any combination of instance types within each node group

Instance groups

Primary

Choose EC2 instance type

m4.xlarge
4 vCore 16 GiB memory EBS only storage
On-Demand price: $0.200 per instance/hour
Lowest Spot price: $0.072 (us-east-2a)

☐ Use multiple primary nodes
To improve cluster availability, use 3 primary nodes with the same configuration and bootstrap actions. You can not use multiple primary nodes with instance fleets.

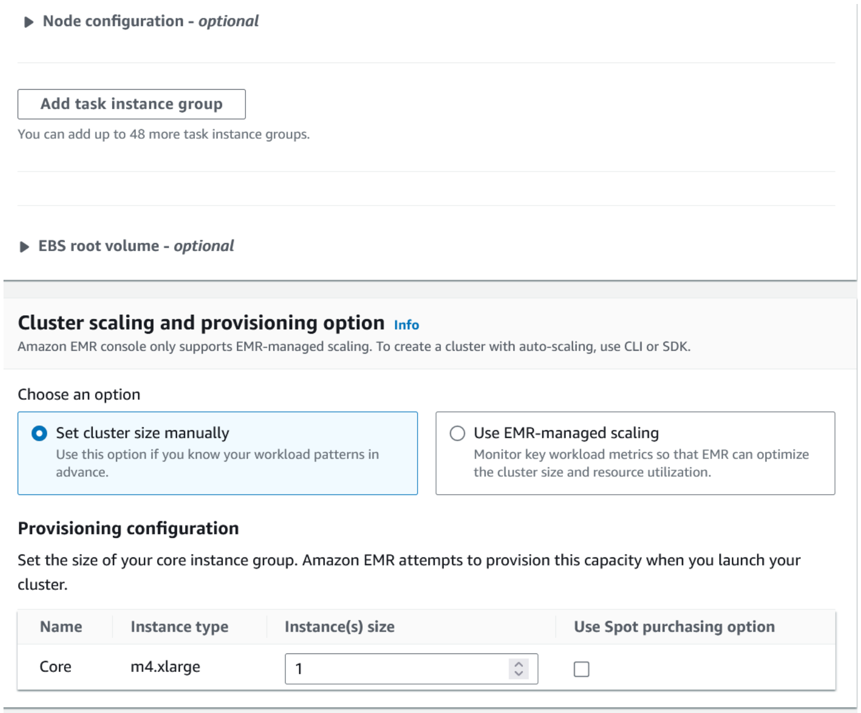
▶ Node configuration - *optional*

Core

Choose EC2 instance type

m4.xlarge
4 vCore 16 GiB memory EBS only storage
On-Demand price: $0.200 per instance/hour
Lowest Spot price: $0.072 (us-east-2a)

▶ Node configuration - *optional*



▶ Node configuration - *optional*

Add task instance group

You can add up to 48 more task instance groups.

▶ EBS root volume - *optional*

Cluster scaling and provisioning option Info

Amazon EMR console only supports EMR-managed scaling. To create a cluster with auto-scaling, use CLI or SDK.

Choose an option

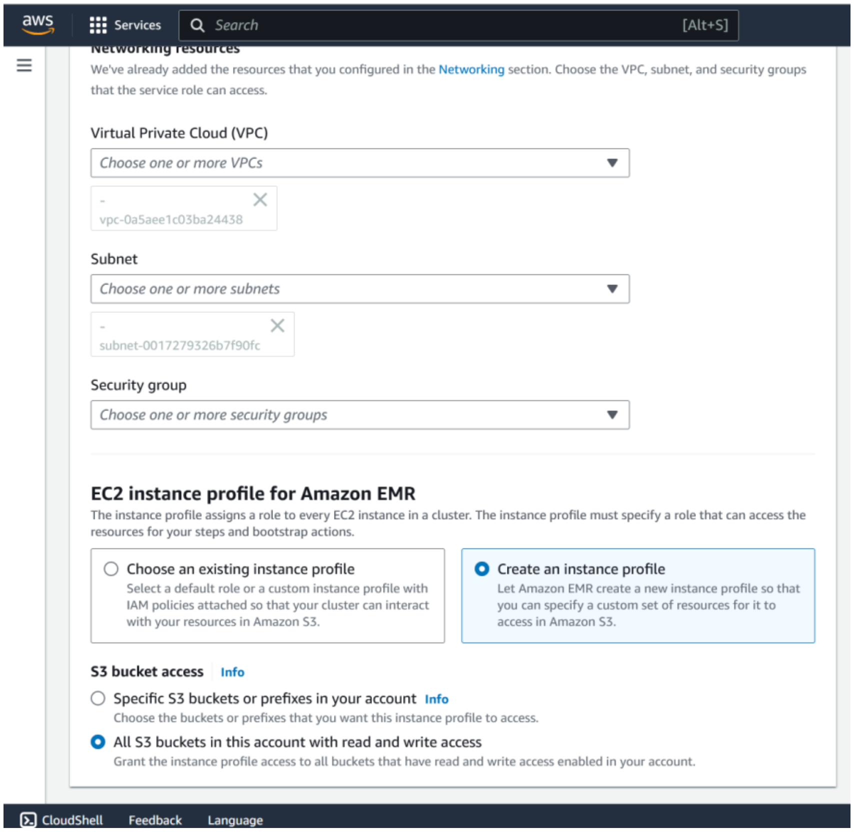
● Set cluster size manually
Use this option if you know your workload patterns in advance.

○ Use EMR-managed scaling
Monitor key workload metrics so that EMR can optimize the cluster size and resource utilization.

Provisioning configuration

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

Name	Instance type	Instance(s) size	Use Spot purchasing option
Core	m4.xlarge	1	□



Networking resources

We've already added the resources that you configured in the Networking section. Choose the VPC, subnet, and security groups that the service role can access.

Virtual Private Cloud (VPC)

Choose one or more VPCs

-
vpc-0a5aee1c03ba24438

Subnet

Choose one or more subnets

-
subnet-0017279326b7f90fc

Security group

Choose one or more security groups

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.

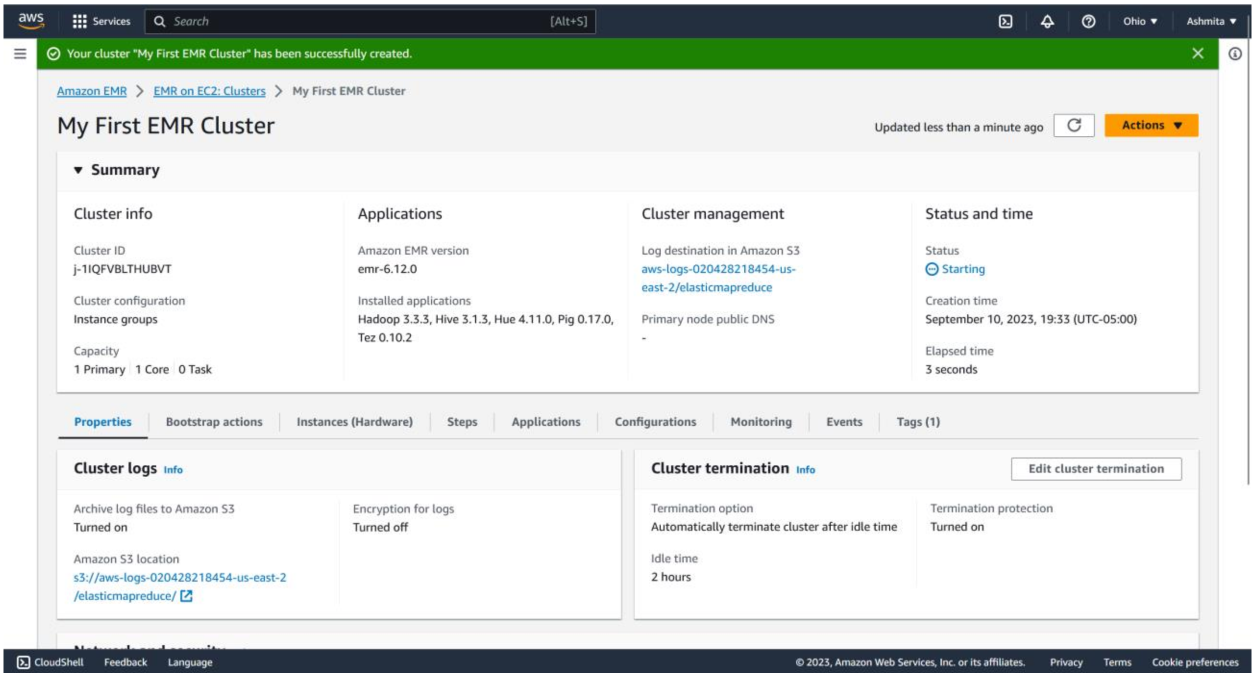
S3 bucket access Info

○ 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.

Cluster created:

Starting state:



Your cluster "My First EMR Cluster" has been successfully created.

My First EMR Cluster

Updated less than a minute ago

▾ Summary

Cluster info

Cluster ID
j-1IQFVBLTHUBVT

Cluster configuration
Instance groups

Capacity
1 Primary | 1 Core | 0 Task

Applications

Amazon EMR version
emr-6.12.0

Installed applications
Hadoop 3.3.3, 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-020428218454-us-east-2/elasticmapreduce

Primary node public DNS
-

Status and time

Status
Starting

Creation time
September 10, 2023, 19:33 (UTC-05:00)

Elapsed time
3 seconds

Cluster logs Info

Archive log files to Amazon S3
Turned on

Encryption for logs
Turned off

Amazon S3 location
s3://aws-logs-020428218454-us-east-2/elasticmapreduce/

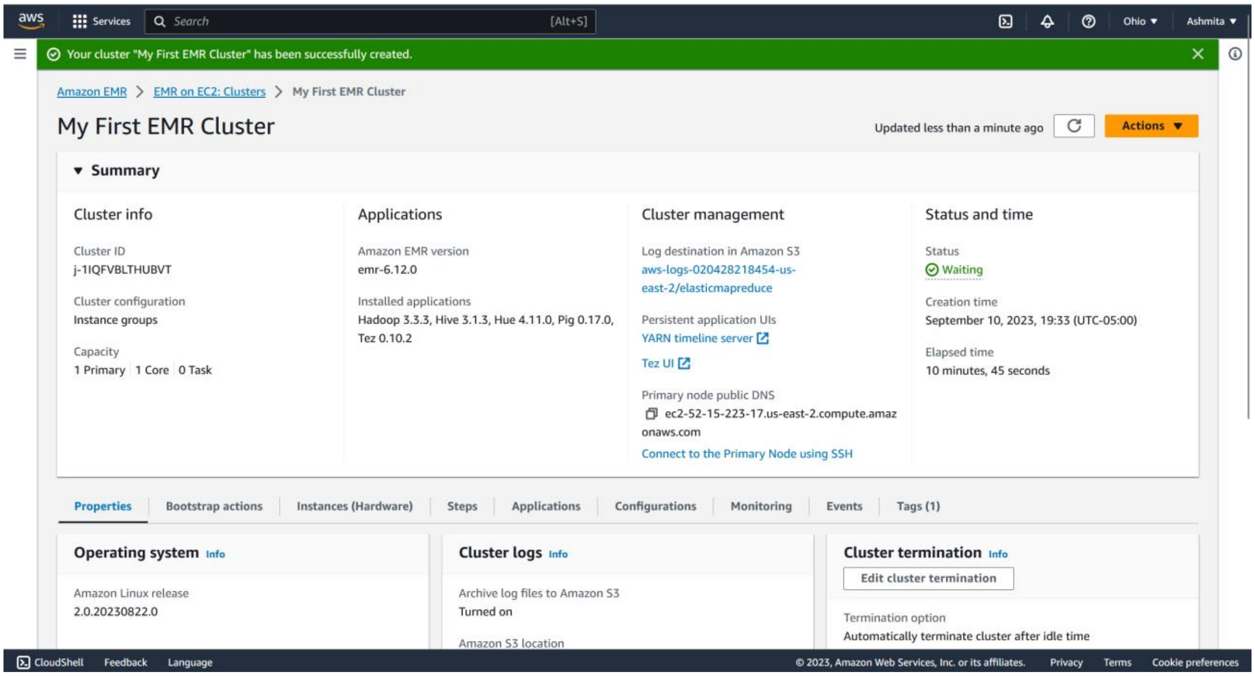
Cluster termination Info

Termination option
Automatically terminate cluster after idle time

Termination protection
Turned on

Idle time
2 hours

Waiting state:



Your cluster "My First EMR Cluster" has been successfully created.

My First EMR Cluster

Updated less than a minute ago

▾ Summary

Cluster info

Cluster ID
j-1IQFVBLTHUBVT

Cluster configuration
Instance groups

Capacity
1 Primary | 1 Core | 0 Task

Applications

Amazon EMR version
emr-6.12.0

Installed applications
Hadoop 3.3.3, 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-020428218454-us-east-2/elasticmapreduce

Persistent application UIs
YARN timeline server
Tez UI

Primary node public DNS
ec2-52-15-223-17.us-east-2.compute.amazonaws.com
Connect to the Primary Node using SSH

Status and time

Status
✓ Waiting

Creation time
September 10, 2023, 19:33 (UTC-05:00)

Elapsed time
10 minutes, 45 seconds

Operating system Info

Amazon Linux release
2.0.20230822.0

Cluster logs Info

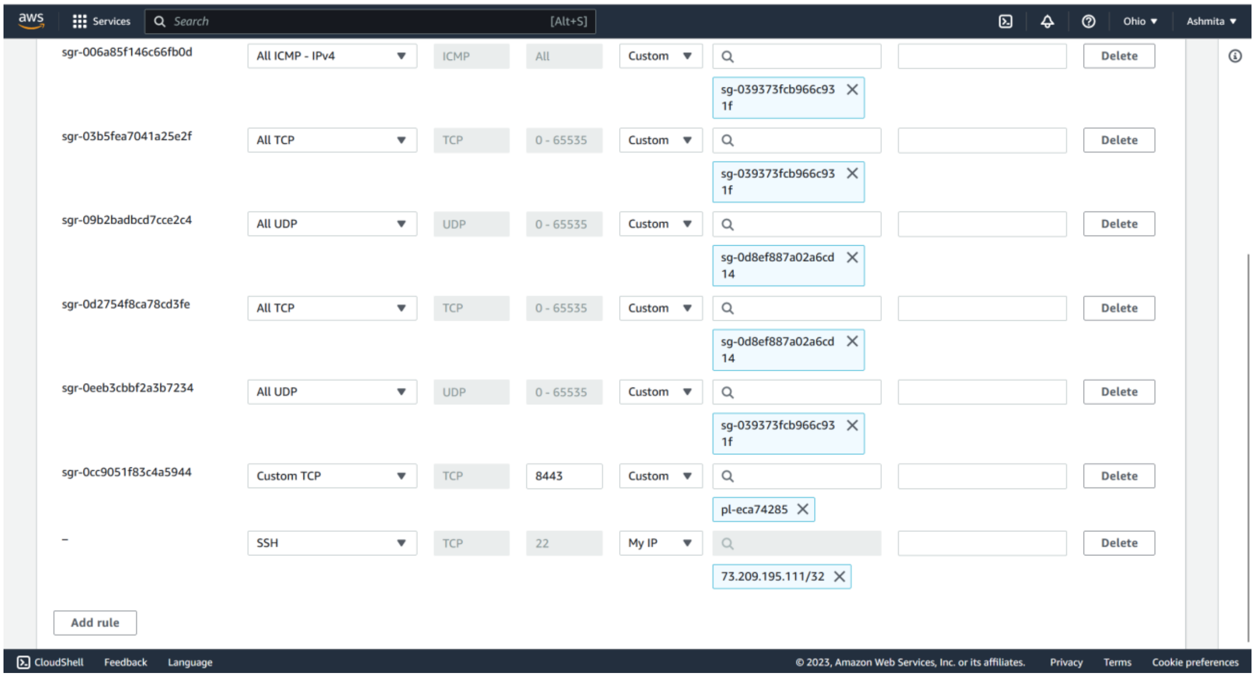
Archive log files to Amazon S3
Turned on

Amazon S3 location

Cluster termination Info

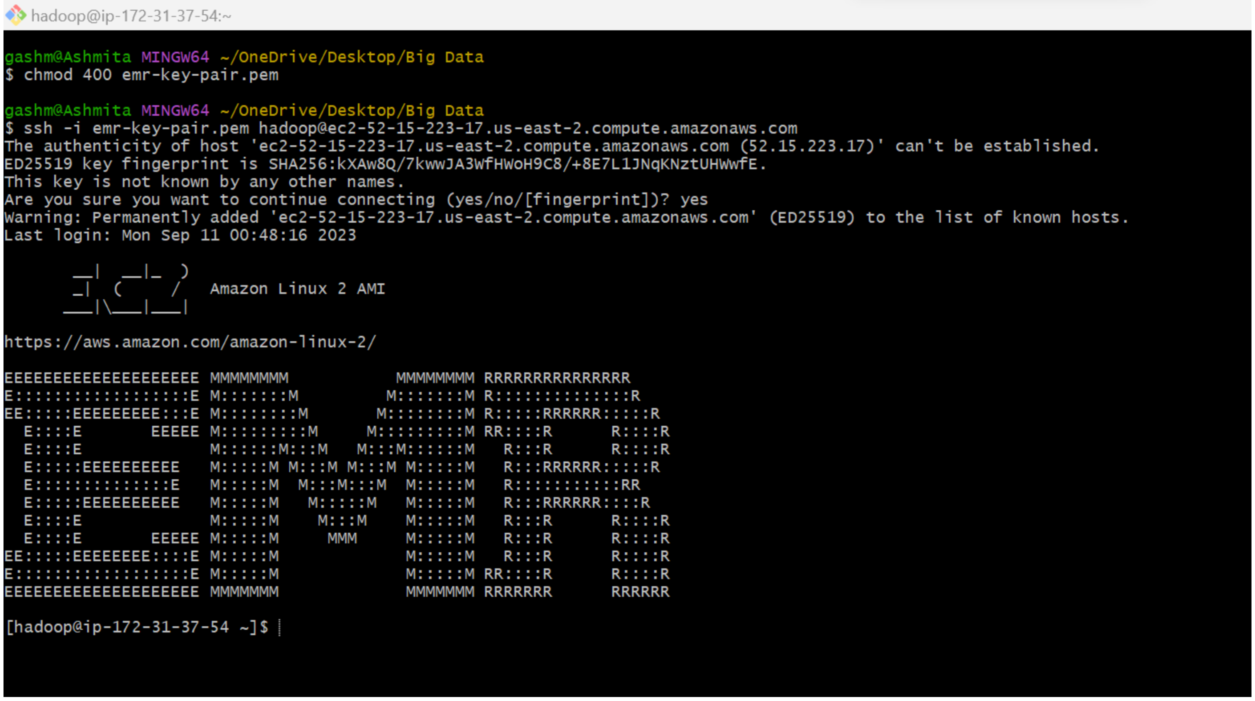
Termination option
Automatically terminate cluster after idle time

Inbound rules:



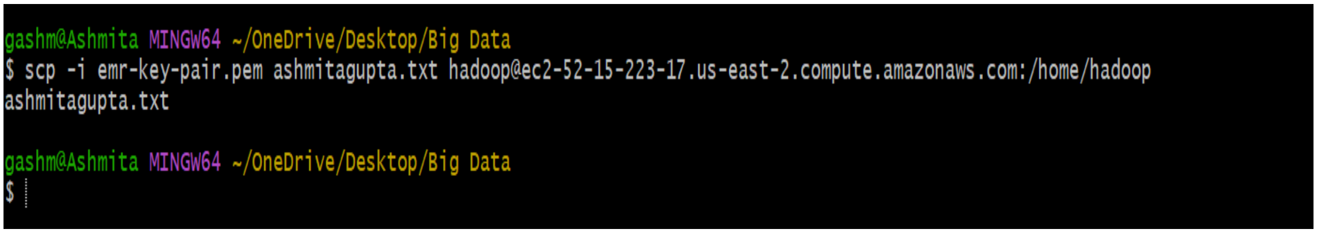
sgr-006a85f146c66fb0d	All ICMP - IPv4 ▼	ICMP	All	Custom ▼	sg-039373fcb966c931f ✕		Delete
sgr-03b5fea7041a25e2f	All TCP ▼	TCP	0 - 65535	Custom ▼	sg-039373fcb966c931f ✕		Delete
sgr-09b2badbcd7cce2c4	All UDP ▼	UDP	0 - 65535	Custom ▼	sg-0d8ef887a02a6cd14 ✕		Delete
sgr-0d2754f8ca78cd3fe	All TCP ▼	TCP	0 - 65535	Custom ▼	sg-0d8ef887a02a6cd14 ✕		Delete
sgr-0eeb3cbbf2a3b7234	All UDP ▼	UDP	0 - 65535	Custom ▼	sg-039373fcb966c931f ✕		Delete
sgr-0cc9051f83c4a5944	Custom TCP ▼	TCP	8443	Custom ▼	pl-eca74285 ✕		Delete
–	SSH ▼	TCP	22	My IP ▼	73.209.195.111/32 ✕		Delete

Connecting to the Primary Node Using SSH



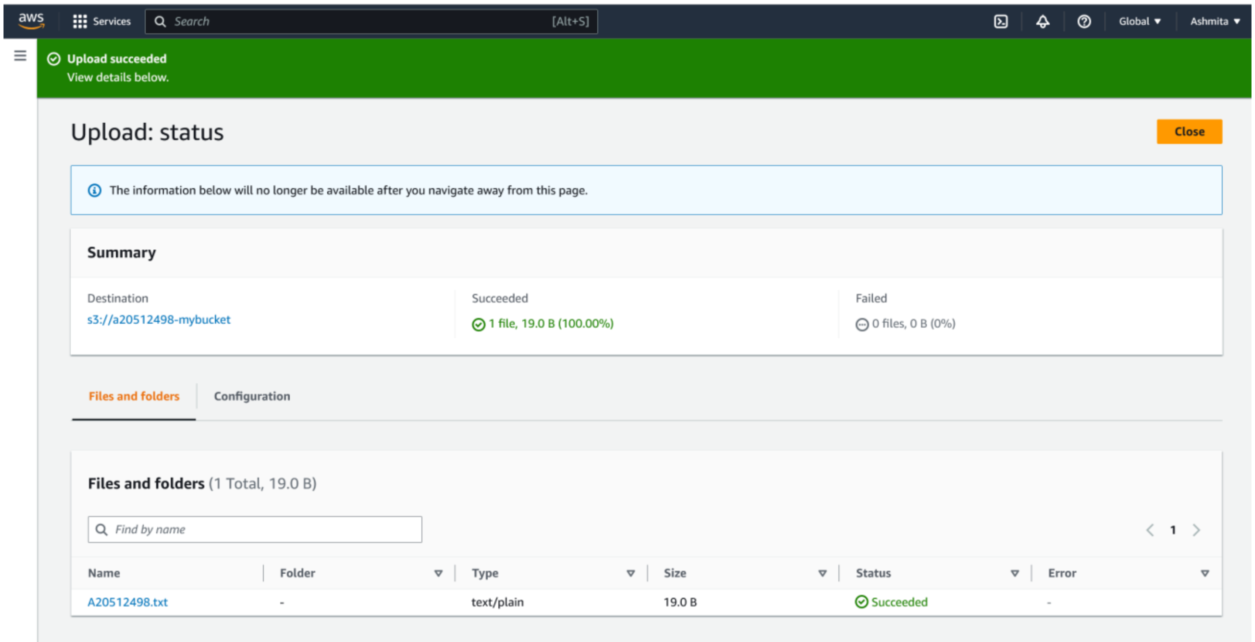
```
gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ chmod 400 emr-key-pair.pem

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ ssh -i emr-key-pair.pem hadoop@ec2-52-15-223-17.us-east-2.compute.amazonaws.com
The authenticity of host 'ec2-52-15-223-17.us-east-2.compute.amazonaws.com (52.15.223.17)' can't be established.
ED25519 key fingerprint is SHA256:kXAw8Q/7kwwJA3WfHWoH9C8/+8E7L1JNqKNztUHWwfE.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-52-15-223-17.us-east-2.compute.amazonaws.com' (ED25519) to the list of known hosts.
Last login: Mon Sep 11 00:48:16 2023

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/

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-172-31-37-54 ~]$ 
```

SCP command executed to copy file from local machine to the home directory of Hadoop master node account:



```
gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ scp -i emr-key-pair.pem ashmitagupta.txt hadoop@ec2-52-15-223-17.us-east-2.compute.amazonaws.com:/home/hadoop
ashmitagupta.txt

gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$
```

A20512498.txt file uploaded to my bucket:



Upload succeeded
View details below.

Upload: status

The information below will no longer be available after you navigate away from this page.

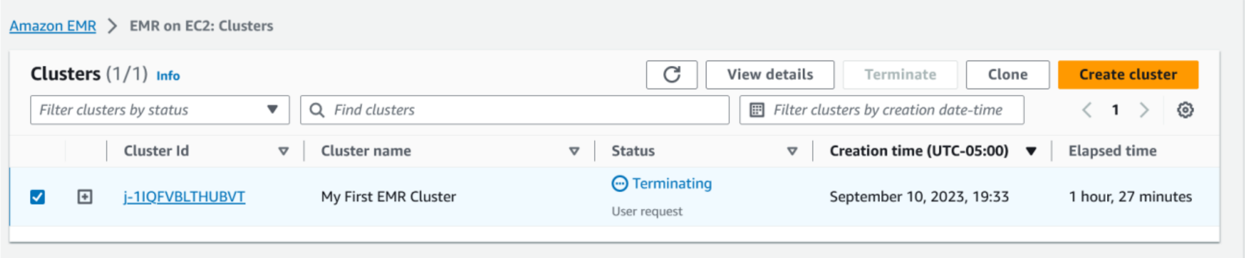
Summary

Destination	Succeeded	Failed
s3://a20512498-mybucket	✓ 1 file, 19.0 B (100.00%)	0 files, 0 B (0%)

Files and folders (1 Total, 19.0 B)

Name	Folder	Type	Size	Status	Error
A20512498.txt	-	text/plain	19.0 B	✓ Succeeded	-

Cluster created as below:



Amazon EMR > EMR on EC2: Clusters
View details
Terminate
Clone
Create cluster
Filter clusters by status
Find clusters
Filter clusters by creation date-time

Clusters (1/1) Info

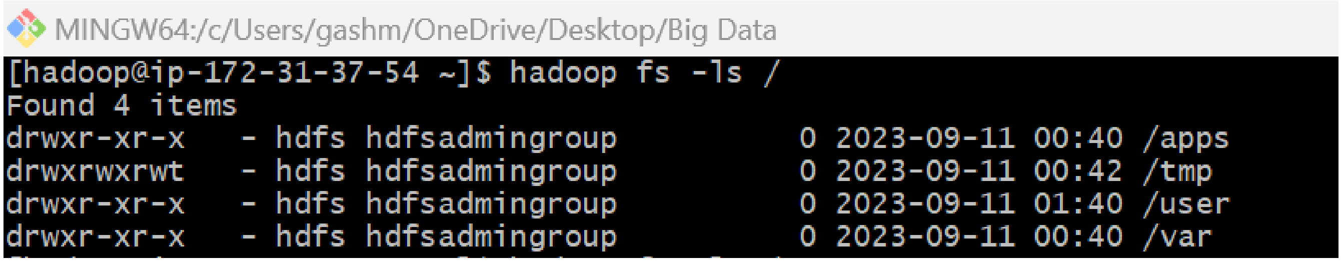
		Cluster Id	Cluster name	Status	Creation time (UTC-05:00)	Elapsed time
☑	⊞	j-1IQFVBLTHUBVT	My First EMR Cluster	Terminating User request	September 10, 2023, 19:33	1 hour, 27 minutes

9. **(2 points) Execute the following hdfs command to list the files or directories that are listed (also indicating which is a file and which a directory):**

   hadoop fs –ls /

Take a screen snapshot of names of the files or directories that are listed and include it in your assignment submission:

Ans. Command: **hadoop fs -ls /**



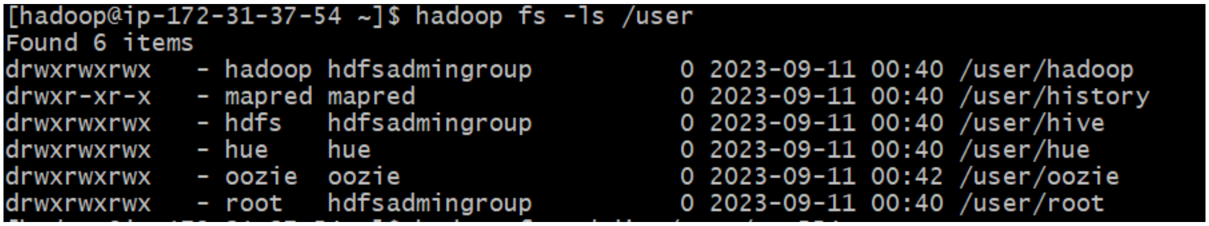
```
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hdfsadmingroup          0 2023-09-11 00:40 /apps
drwxrwxrwt   - hdfs hdfsadmingroup          0 2023-09-11 00:42 /tmp
drwxr-xr-x   - hdfs hdfsadmingroup          0 2023-09-11 01:40 /user
drwxr-xr-x   - hdfs hdfsadmingroup          0 2023-09-11 00:40 /var
```

10. (2 points) Execute a command (you needed to figure out which one) to list the files and directories under the hdfs directory listed below:

/user

Write down the command you executed and also take a screen snapshot of names of the files or directories that are listed and include it in your assignment submission.

Ans. **hadoop fs -ls /user**



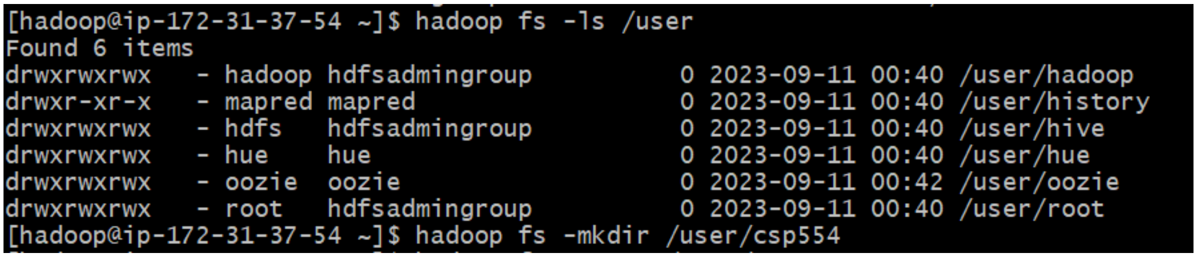
```
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -ls /user
Found 6 items
drwxrwxrwx   - hadoop hdfsadmingroup          0 2023-09-11 00:40 /user/hadoop
drwxr-xr-x   - mapred mapred                  0 2023-09-11 00:40 /user/history
drwxrwxrwx   - hdfs   hdfsadmingroup          0 2023-09-11 00:40 /user/hive
drwxrwxrwx   - hue    hue                     0 2023-09-11 00:40 /user/hue
drwxrwxrwx   - oozie  oozie                   0 2023-09-11 00:42 /user/oozie
drwxrwxrwx   - root   hdfsadmingroup          0 2023-09-11 00:40 /user/root
```

11. (2 points) Execute a command to create the following HDFS directory:

/user/csp554

Record the command you executed and include it in your assignment submission.

Ans. Command: **hadoop fs -mkdir /user/csp554**



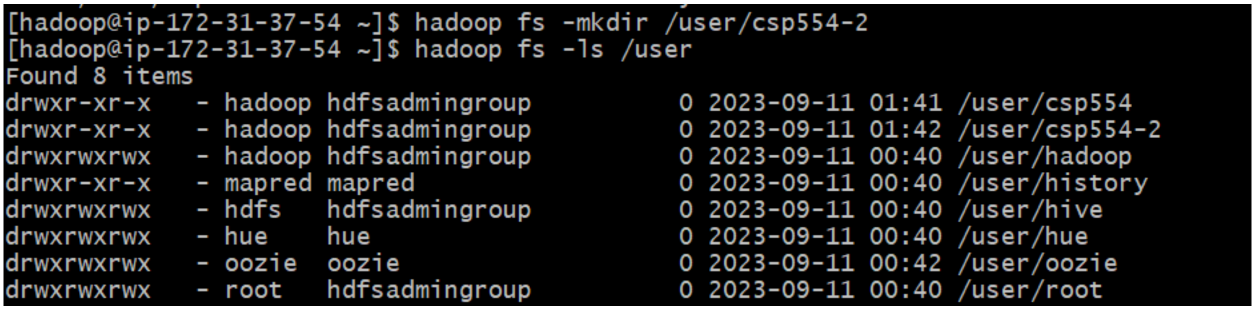
```
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -ls /user
Found 6 items
drwxrwxrwx   - hadoop hdfsadmingroup          0 2023-09-11 00:40 /user/hadoop
drwxr-xr-x   - mapred mapred                  0 2023-09-11 00:40 /user/history
drwxrwxrwx   - hdfs   hdfsadmingroup          0 2023-09-11 00:40 /user/hive
drwxrwxrwx   - hue    hue                     0 2023-09-11 00:40 /user/hue
drwxrwxrwx   - oozie  oozie                   0 2023-09-11 00:42 /user/oozie
drwxrwxrwx   - root   hdfsadmingroup          0 2023-09-11 00:40 /user/root
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -mkdir /user/csp554
```

12. (2 points) Execute a command to create the following HDFS directory:

/user/csp554-2

Ans. Command**: hadoop fs -mkdir /user/csp554-2**

Both csp554 and csp554-2 HDFS directories created as shown below:



```
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -mkdir /user/csp554-2
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -ls /user
Found 8 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-09-11 01:41 /user/csp554
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-09-11 01:42 /user/csp554-2
drwxrwxrwx   - hadoop hdfsadmingroup          0 2023-09-11 00:40 /user/hadoop
drwxr-xr-x   - mapred mapred                  0 2023-09-11 00:40 /user/history
drwxrwxrwx   - hdfs   hdfsadmingroup          0 2023-09-11 00:40 /user/hive
drwxrwxrwx   - hue    hue                     0 2023-09-11 00:40 /user/hue
drwxrwxrwx   - oozie  oozie                   0 2023-09-11 00:42 /user/oozie
drwxrwxrwx   - root   hdfsadmingroup          0 2023-09-11 00:40 /user/root
```

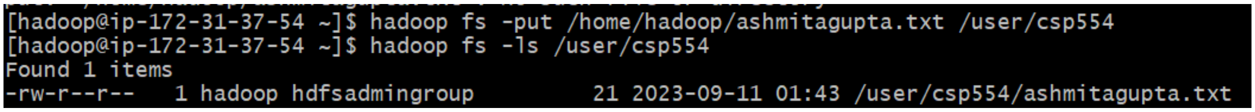
Record the command you executed and include it in your assignment submission.

13. (2 points) Execute a command that copies a given local file to the given hdfs directory :

Source local file: /home/hadoop/myname.txt (where the actual name is your name as described above)

Destination HDFS directory: /user/csp554

Ans. Command: **hadoop fs -put /home/hadoop/ashmitagupta.txt /user/csp554**



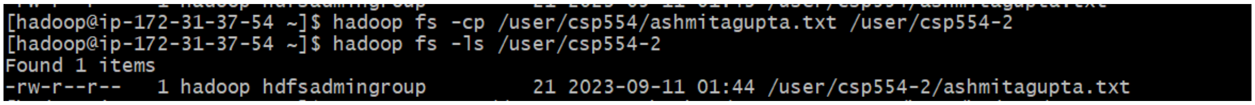
```
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -put /home/hadoop/ashmitagupta.txt /user/csp554
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -ls /user/csp554
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup         21 2023-09-11 01:43 /user/csp554/ashmitagupta.txt
```

14. (2 points) Copy a file from one hdfs directory to another hdfs directory and write down the command

Source hdfs file: /user/csp554/myname.txt (where the actual name is your name as described above)

Destination HDFS directory: /user/csp554-2

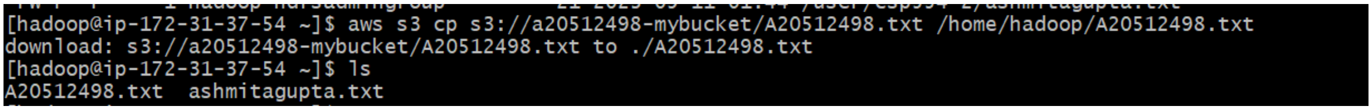
Ans. Command: **hadoop fs -cp /user/csp554/ashmitagupta.txt /user/csp554-2**



```
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -cp /user/csp554/ashmitagupta.txt /user/csp554-2
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -ls /user/csp554-2
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup         21 2023-09-11 01:44 /user/csp554-2/ashmitagupta.txt
```

15. (2 points) Copy the object myid.txt you uploaded to an S3 bucket into the Hadoop master node Linux file system. The actual object includes your student id as above.

Ans. Command: **aws s3 cp s3://a20512498-mybucket/A20512498.txt /home/hadoop/A20512498.txt**



```
[hadoop@ip-172-31-37-54 ~]$ aws s3 cp s3://a20512498-mybucket/A20512498.txt /home/hadoop/A20512498.txt
download: s3://a20512498-mybucket/A20512498.txt to ./A20512498.txt
[hadoop@ip-172-31-37-54 ~]$ ls
A20512498.txt  ashmitagupta.txt
```

16. **(2 points) Copy the same object myid.txt you created in an S3 bucket into HDFS into the directory /users/csp554**

    hadoop fs -cp s3://mybucket/myid.txt   hdfs:///user/csp554-2

    After you executed the above command, execute another command (you needed to figure out which one) to list the files and directories under the hdfs directory listed below:

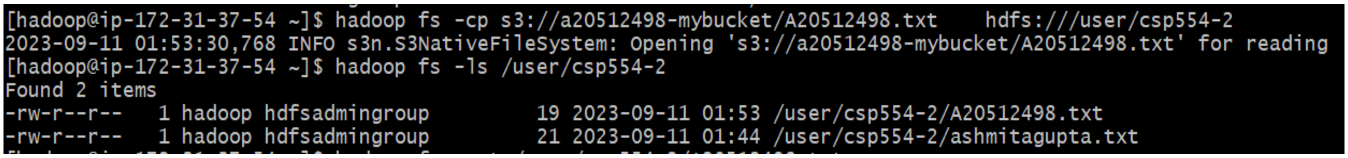
    /user/csp554-2

Write down the command you executed and also take a screen snapshot of names of the files or directories that are listed and include it in your assignment submission.

Ans. Commands:

hadoop fs -cp s3://a20512498-mybucket/A20512498.txt hdfs:///user/csp554-2

hadoop fs -ls /user/csp554-2



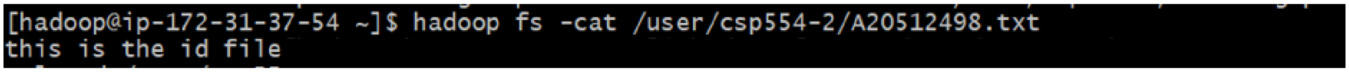
```
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -cp s3://a20512498-mybucket/A20512498.txt    hdfs:///user/csp554-2
2023-09-11 01:53:30,768 INFO s3n.S3NativeFileSystem: Opening 's3://a20512498-mybucket/A20512498.txt' for reading
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -ls /user/csp554-2
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup         19 2023-09-11 01:53 /user/csp554-2/A20512498.txt
-rw-r--r--   1 hadoop hdfsadmingroup         21 2023-09-11 01:44 /user/csp554-2/ashmitagupta.txt
```

17. **(2 points) Execute a command to show the contents of the myid.txt file in the hdfs directory /user/csp554-2**

    Clue: look up about how to use the "cat" command in the file system shell document.

Write down the command you executed and also take a screen snapshot of the listed content of the file and include it in your assignment submission.

Ans. Command: **hadoop fs -cat /user/csp554-2/A20512498.txt**



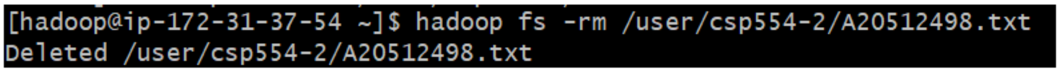
```
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -cat /user/csp554-2/A20512498.txt
this is the id file
```

18. **(2 points) Execute a command to remove the myid.txt file in the hdfs directory /user/csp554-2**

    Clue: look up about how to use the "rm" command in the file system shell document.

Write down the command you executed, then list the content of the /user/csp554-2 HDFS directory and take a screen snapshot of the listed content of the directory and include it in your assignment submission.

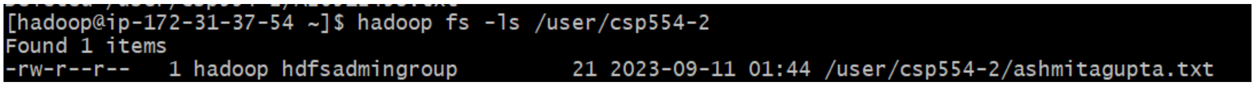
Ans. Command: **hadoop fs -rm /user/csp554-2/A20512498.txt**



```
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -rm /user/csp554-2/A20512498.txt
Deleted /user/csp554-2/A20512498.txt
```

Screenshot of contents in directory:

Command: **hadoop fs -ls /user/CS554-2**



```
[hadoop@ip-172-31-37-54 ~]$ hadoop fs -ls /user/csp554-2
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup         21 2023-09-11 01:44 /user/csp554-2/ashmitagupta.txt
```