

CSP 554

HOMEWORK 2

ASHMITA GUPTA (A20512498)

1. Key Pair using the EC2 service

The screenshot shows the AWS Management Console interface for creating a new key pair. The breadcrumb navigation indicates the path: EC2 > Key pairs > Create key pair. The main heading is 'Create key pair' with an 'Info' link. Below this, a 'Key pair' section explains that a key pair consists of a private key and a public key used for authentication. The form includes a 'Name' field with the value 'emr-key-pair' and a note that the name can include up to 255 ASCII characters. The 'Key pair type' section has two radio buttons: 'RSA' (selected) and 'ED25519'. The 'Private key file format' section has two radio buttons: '.pem' (selected, for use with OpenSSH) and '.ppk' (for use with PuTTY). There is an 'Add new tag' button and a note that no tags are currently associated with the resource. At the bottom, there are 'Cancel' and 'Create key pair' buttons.

2. Key pair successfully created

The screenshot shows the AWS Management Console 'Key pairs (1)' page. A green banner at the top indicates 'Successfully created key pair'. Below this, there is a search bar and a table listing the key pairs. The table has columns for Name, Type, Created, Fingerprint, and ID. One key pair is listed: 'emr-key-pair' of type 'rsa', created on '2023/09/10 19:12 GMT-5', with a fingerprint of '90:e2:bfa0:07:31:d9:79:13:f3:3d:8d:e8:...' and an ID of 'key-0a1748c1d7ffe3a33'. There are 'Actions' and 'Create key pair' buttons at the top right of the table.

Name	Type	Created	Fingerprint	ID
emr-key-pair	rsa	2023/09/10 19:12 GMT-5	90:e2:bfa0:07:31:d9:79:13:f3:3d:8d:e8:...	key-0a1748c1d7ffe3a33

3. Modifying permission for key pair

```
MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data
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

The screenshot shows the 'Create cluster' page in the AWS Management Console. The breadcrumb trail is 'Amazon EMR > EMR on EC2: Clusters > Create cluster'. The page title is 'Create cluster' with an 'Info' link. The section 'Name and applications' contains a 'Name' field with the value 'My First EMR Cluster'. Below it, the 'Amazon EMR release' is set to 'emr-6.12.0'. The 'Application bundle' section shows several options: Spark, Core Hadoop (selected), Flink, HBase, Presto, Trino, and Custom. Below the bundles, it lists 'Applications included in bundle' as 'Hadoop 3.3.3 with Hive 3.1.3, Hue 4.11.0, Pig 0.17.0 and Tez 0.10.2'. The 'AWS Glue Data Catalog settings' section has a checkbox for 'Use for Hive table metadata' which is unchecked. The 'Operating system options' section has two radio buttons: 'Amazon Linux release' (selected) and 'Custom Amazon Machine Image (AMI)'. At the bottom, there is a checked checkbox for 'Automatically apply latest Amazon Linux updates'.

aws Services Search [Alt+S]

Amazon EMR > EMR on EC2: Clusters > Create cluster

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

Spark Core Hadoop Flink HBase Presto Trino Custom

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

The screenshot shows the 'Cluster configuration' page in the AWS Management Console. The breadcrumb trail is 'aws Services Search [Alt+S]'. The page title is 'Cluster configuration' with an 'Info' link. Below the title, it says 'Choose a configuration method for the primary, core, and task node groups for your cluster.' There are two radio buttons: 'Instance groups' (selected) and 'Instance fleets'. The 'Instance groups' section has a sub-section 'Primary' with a dropdown menu for 'Choose EC2 instance type' set to 'm4.xlarge'. Below this, there is a checkbox for 'Use multiple primary nodes' which is unchecked. The 'Core' section also has a dropdown menu for 'Choose EC2 instance type' set to 'm4.xlarge'. Below the dropdowns, there are links for 'Node configuration - optional'.

aws Services Search [Alt+S]

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)

Actions

☐ 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)

Actions

► 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	<input type="checkbox"/>

aws

Services

Search

[Alt+S]

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.

CloudShell Feedback Language

Cluster created:

Starting state:

The screenshot displays the AWS Management Console interface for an Amazon EMR cluster. A green notification bar at the top states: "Your cluster 'My First EMR Cluster' has been successfully created." The breadcrumb navigation shows: Amazon EMR > EMR on EC2: Clusters > My First EMR Cluster. The cluster title "My First EMR Cluster" is prominently displayed, with a timestamp "Updated less than a minute ago" and an "Actions" dropdown menu.

The **Summary** section is expanded, showing four columns of information:

- Cluster info:** Cluster ID is j-1IQFVBLTHUBVT. Cluster configuration includes Instance groups and Capacity (1 Primary, 1 Core, 0 Task).
- Applications:** Amazon EMR version is emr-6.12.0. Installed applications include Hadoop 3.3.3, Hive 3.1.3, Hue 4.11.0, Pig 0.17.0, and Tez 0.10.2.
- Cluster management:** Log destination in Amazon S3 is aws-logs-020428218454-us-east-2/elasticmapreduce. Primary node public DNS is -.
- Status and time:** Status is **Starting** (indicated by a blue circle with a play icon). Creation time is September 10, 2023, 19:33 (UTC-05:00). Elapsed time is 3 seconds.

Below the summary, a horizontal tab bar includes: Properties, Bootstrap actions, Instances (Hardware), Steps, Applications, Configurations, Monitoring, Events, and Tags (1). The **Properties** tab is active, showing two sub-sections:

- Cluster logs Info:** Archive log files to Amazon S3 is Turned on. Amazon S3 location is s3://aws-logs-020428218454-us-east-2/elasticmapreduce/. Encryption for logs is Turned off.
- Cluster termination Info:** Termination option is Automatically terminate cluster after idle time. Idle time is 2 hours. Termination protection is Turned on. An "Edit cluster termination" button is present.

The footer of the console shows "CloudShell", "Feedback", "Language", and copyright information for Amazon Web Services, Inc. or its affiliates.

Waiting state:

This screenshot shows the same AWS Management Console page, but the cluster status has changed to **Waiting**, indicated by a green circle with a checkmark. The "Elapsed time" has increased to 10 minutes, 45 seconds.

The **Summary** section now includes additional details in the **Cluster management** column:

- Persistent application Uls: YARN timeline server and Tez UI (both with external links).
- Primary node public DNS: ec2-52-15-223-17.us-east-2.compute.amazonaws.com.
- A link to "Connect to the Primary Node using SSH".

The **Properties** tab remains active, but the sub-sections have updated:

- Operating system Info:** Amazon Linux release 2.0.20230822.0.
- Cluster logs Info:** Archive log files to Amazon S3 is Turned on. Amazon S3 location is -.
- Cluster termination Info:** The "Edit cluster termination" button is still present.

The footer remains consistent with the previous screenshot, showing "CloudShell", "Feedback", "Language", and copyright information.

Inbound rules:

aws

Services

Search

[Alt+]5

Ohio

Ashmita

sg-006a85f146c66fb0d

All ICMP - IPv4

ICMP

All

Custom

sg-039373fcb966c931f

X

Delete

sg-03b5fea7041a25e2f

All TCP

TCP

0 - 65535

Custom

sg-039373fcb966c931f

X

Delete

sg-09b2badbcd7c2e2c4

All UDP

UDP

0 - 65535

Custom

sg-0d8ef887a02a6cd14

X

Delete

sg-0d2754f8ca78cd3fe

All TCP

TCP

0 - 65535

Custom

sg-0d8ef887a02a6cd14

X

Delete

sg-0eeb3cbbf2a3b7234

All UDP

UDP

0 - 65535

Custom

sg-039373fcb966c931f

X

Delete

sg-0cc9051f83c4a5944

Custom TCP

TCP

8443

Custom

pl-eca74285

X

Delete

-

SSH

TCP

22

My IP

73.209.195.111/32

X

Delete

Add rule

Connecting to the Primary Node Using SSH

```

hadoop@ip-172-31-37-54:~
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/+8E7L1JNqKzntUHwWfE.
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 MMMMMMMMM MMMMMMMMM RRRRRRRRRRRRRR
E:::EEEEEEEEEEEEEEEE M:::M M:::M R:::R
EE:::EEEEEEEEEEEEEEEE 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:::EEEEEEEEEEEE M:::M M:::M M:::M R:::RRRRRR:::R
E:::EEEEEEEEEEEE M:::M M:::M M:::M R:::RRRRRR:::R
E:::EEEEEEEEEEEE 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:::EEEEEEEEEEEE M:::M M:::M R:::R R:::R
E:::EEEEEEEEEEEE M:::M M:::M RR:::R R:::R
EEEEEEEEEEEEEEEEEEEE MMMMMMMMM RRRRRRR RRRRRR

[hadoop@ip-172-31-37-54 ~]$

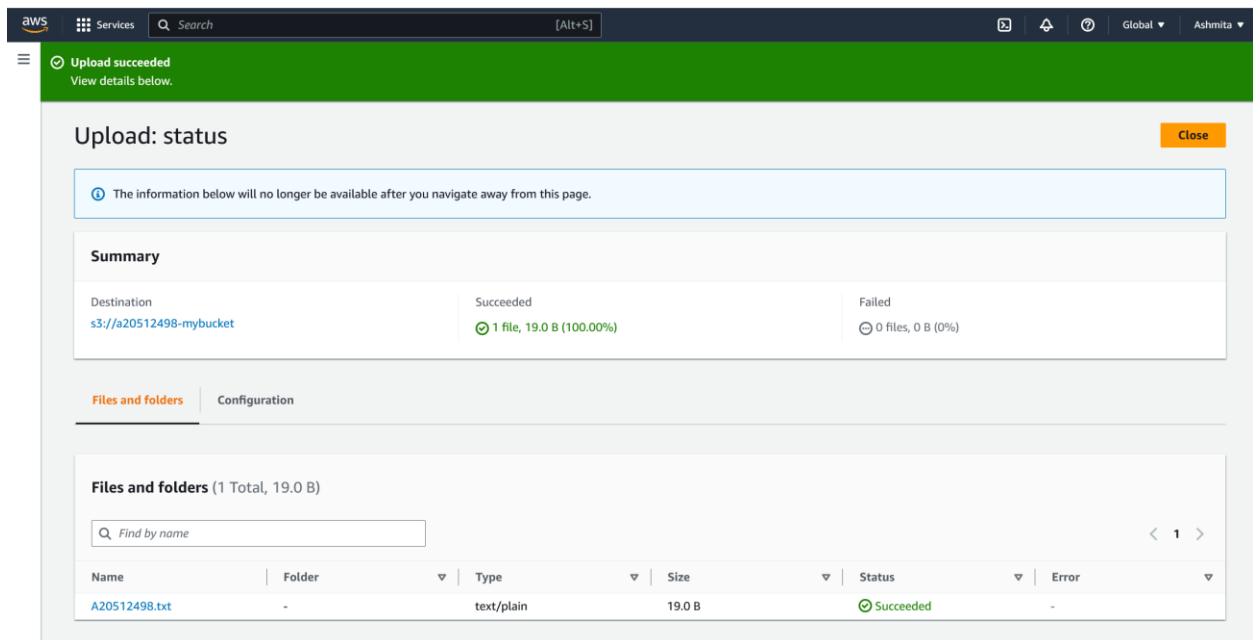
```

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
$
```

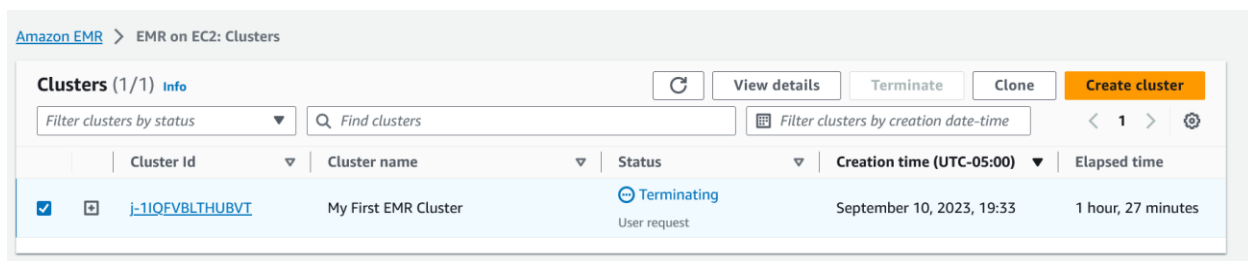
A20512498.txt file uploaded to my bucket:



The screenshot shows the AWS S3 console interface. At the top, a green banner indicates "Upload succeeded". Below this, the "Upload: status" section provides a summary of the upload. The destination is "s3://a20512498-mybucket". The summary shows "Succeeded" with "1 file, 19.0 B (100.00%)" and "Failed" with "0 files, 0 B (0%)". The "Files and folders" tab is selected, showing a table with one file: "A20512498.txt", which is "text/plain", "19.0 B", and "Succeeded".

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

Cluster created as below:



The screenshot shows the Amazon EMR console. The "Clusters (1/1)" section is active. A table lists the cluster "j-1IQFVBLTHUBVT" with the name "My First EMR Cluster". The status is "Terminating" (User request). The creation time is "September 10, 2023, 19:33" and the elapsed time is "1 hour, 27 minutes".

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 /**

```
MINGW64:/c/Users/gashm/OneDrive/Desktop/Big Data
[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 hdfsadmin group 19 2023-09-11 01:53 /user/csp554-2/A20512498.txt
-rw-r--r-- 1 hadoop hdfsadmin group 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
```