

Big Data Technologies (CSP554)

Assignment-2

1. Creating new 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

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.

[Cancel](#) [Create key pair](#)

Successfully created key pair

Key pairs (1) [Info](#)

<input type="checkbox"/>	Name	Type	Created	Fingerprint	ID
<input type="checkbox"/>	emr-key-pair	rsa	2022/09/05 20:10 GMT-5	11:1a36b56a2d6594aeaac19c9f...	key-02b124735c6dcdf8

2. Giving permission to key pair

```
aasth@LAPTOP-HJTR6HMR MINGW64 ~  
$ cd Downloads/  
  
aasth@LAPTOP-HJTR6HMR MINGW64 ~/Downloads  
$ chmod 400 emr-key-pair.pem  
  
aasth@LAPTOP-HJTR6HMR MINGW64 ~/Downloads  
$ |
```

3. Creation of Amazon EMR cluster by using steps given in the document (Steps 1-5 screenshots)

Create Cluster - Quick Options [Go to advanced options](#)

General Configuration

Cluster name

☒ Logging [Info](#)

S3 folder

Launch mode ☒ Cluster [Info](#) ☐ Step execution [Info](#)

Software configuration

Release label [Info](#)

Applications

☒ Core Hadoop: Hadoop 2.10.1, Hive 2.3.9, Hue 4.10.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2

☐ HBase: HBase 1.4.13, Hadoop 2.10.1, Hive 2.3.9, Hue 4.10.0, Phoenix 4.14.3, and ZooKeeper 3.4.14

☐ Presto: Presto 0.267 with Hadoop 2.10.1 HDFS and Hive 2.3.9 Metastore

☐ Spark: Spark 2.4.8 on Hadoop 2.10.1 YARN and Zeppelin 0.10.0

☐ Use AWS Glue Data Catalog for table metadata [Info](#)

Hardware configuration

Instance type The selected instance type adds 32 GiB of GP2 EBS storage per instance by default. [Learn more](#)

Number of instances (1 master and 1 core nodes)

Cluster scaling ☐ scale cluster nodes based on workload

Auto-termination ☒ Enable auto-termination [Learn more](#)

Terminate cluster when it is idle after hours minutes

Security and access

EC2 key pair emr-key-pair Learn how to create an EC2 key pair.

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 ☐ Use EMR_DefaultRole_V2

EC2 instance profile EMR_EC2_DefaultRole

Cancel

Create cluster

4. Screenshot of the cluster when starting

Clone Terminate AWS CLI export

Cluster: My First EMR Cluster Starting

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

Summary

ID: j-3BG6RB0MQELX2
Creation date: 2022-09-05 20:35 (UTC-5)
Elapsed time: 1 second
After last step completes: Cluster waits
Termination protection: Off [Change](#)
Tags: -- [View All / Edit](#)
Master public DNS: --

Configuration details

Release label: emr-5.36.0
Hadoop distribution: Amazon 2.10.1
Applications: Hive 2.3.9, Hue 4.10.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2
Log URI: s3://aws-logs-984516399758-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --
Amazon Linux Release: 2.0.20220426.0 [Learn more](#)

Application user interfaces

Persistent user interfaces: --
On-cluster user interfaces: --

Network and hardware

Availability zone: --
Subnet ID: [subnet-066617fbce9fa0f4d](#)
Master: Provisioning 1 m4.large
Core: Provisioning 1 m4.large
Task: --
Cluster scaling: Not enabled
Auto-termination: Terminate if idle for 1 hour

Security and access

Key name: emr-key-pair
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All [Change](#)
Security groups for Master:
Security groups for Core & Task:

5. Screenshot of Security groups

AWS Services Search for services, features, blogs, docs, and more [Alt+S]

New EC2 Experience Tell us what you think

EC2 Dashboard EC2 Global View Events Tags Limits

▼ Instances Instances New

Security Groups (2) Info

Filter security groups

search: sg-04f2794e15946cc39 Clear filters

	Name	Security group ID	Security group name	VPC ID	Description	Owner	Inbound rules count	Outbound rules count
<input type="checkbox"/>	sg-04f2794e15946cc39	ElasticMapReduce-master	vpc-0900712dee503bfee	Master group for ElasticMapReduce	984516399758	18 Permission entries	1 Permission entry	
<input type="checkbox"/>	sg-0f264ee72c7d32497	ElasticMapReduce-slave	vpc-0900712dee503bfee	Slave group for ElasticMapReduce	984516399758	6 Permission entries	1 Permission entry	

6. Screenshot of Security group inbound rules

Inbound rules (18)

Filter security group rules

Manage tags Edit inbound rules

	Name	Security group rule...	IP version	Type	Protocol	Port range	Source	Description
<input type="checkbox"/>	sg-04ee86f854468a32b		--	All TCP	TCP	0 - 65535	sg-0f264ee72c7d3249...	--
<input type="checkbox"/>	sg-0ba16a876c54aace4		--	All UDP	UDP	0 - 65535	sg-0f264ee72c7d3249...	--
<input type="checkbox"/>	sg-0dc775daede40919a		IPv4	Custom TCP	TCP	8443	54.240.217.80/29	--
<input type="checkbox"/>	sg-0fd7ed3a356fa5324		IPv4	Custom TCP	TCP	8443	207.171.172.6/32	--
<input type="checkbox"/>	sg-087595561ff4fd831		IPv4	Custom TCP	TCP	8443	72.21.198.64/29	--
<input type="checkbox"/>	sg-0f9c9c9e31885aa47		--	All ICMP - IPv4	ICMP	All	sg-0f264ee72c7d3249...	--

Inbound security group rules successfully modified on security group (sg-04f2794e15946cc39 | ElasticMapReduce-master)

Details

Security Groups (1/3) Info

Filter security groups

	Name	Security group ID	Security group name	VPC ID	Description	Owner	Inbound rules count	Outbound rules count
<input checked="" type="checkbox"/>	-	sg-04f2794e15946cc39	ElasticMapReduce-master	vpc-0900712dee503bfee	Master group for ElasticMapReduce	984516399758	19 Permission entries	1 Permission entry
<input type="checkbox"/>	-	sg-0f264ee72c7d32497	ElasticMapReduce-slave	vpc-0900712dee503bfee	Slave group for ElasticMapReduce	984516399758	6 Permission entries	1 Permission entry
<input type="checkbox"/>	-	sg-0ffed0f8157515f21	default	vpc-0900712dee503bfee	default VPC security group	984516399758	1 Permission entry	1 Permission entry

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 /`

Command executed- `hadoop fs -ls /`

```
[hadoop@ip-172-31-77-128 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x - hdfs hdfsadmingroup 0 2022-09-06 01:43 /apps
drwxrwxrwt - hdfs hdfsadmingroup 0 2022-09-06 01:45 /tmp
drwxr-xr-x - hdfs hdfsadmingroup 0 2022-09-06 01:43 /user
drwxr-xr-x - hdfs hdfsadmingroup 0 2022-09-06 01:43 /var
[hadoop@ip-172-31-77-128 ~]$
```

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`

Command executed- `hadoop fs -ls /user`

```
[hadoop@ip-172-31-77-128 ~]$ hadoop fs -ls /user
Found 6 items
drwxrwxrwx - hadoop hdfsadmingroup 0 2022-09-06 01:43 /user/hadoop
drwxr-xr-x - mapred mapred 0 2022-09-06 01:43 /user/history
drwxrwxrwx - hdfs hdfsadmingroup 0 2022-09-06 01:43 /user/hive
drwxrwxrwx - hue hue 0 2022-09-06 01:43 /user/hue
drwxrwxrwx - oozie oozie 0 2022-09-06 01:45 /user/oozie
drwxrwxrwx - root hdfsadmingroup 0 2022-09-06 01:43 /user/root
[hadoop@ip-172-31-77-128 ~]$
```

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

`/user/csp554`

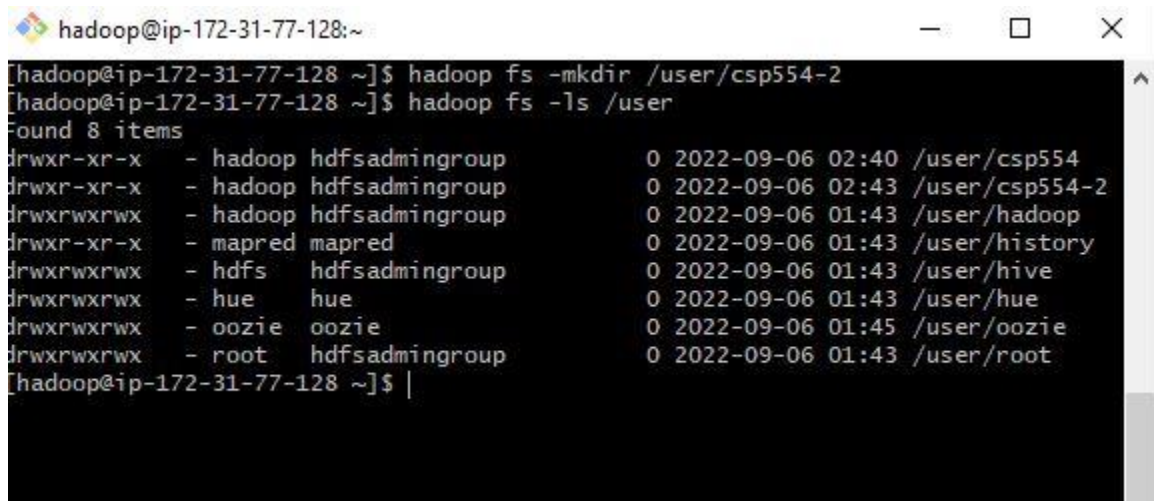
Command executed- `hadoop fs -mkdir /user/csp554`

```
[hadoop@ip-172-31-77-128 ~]$ hadoop fs -mkdir /user/csp554
[hadoop@ip-172-31-77-128 ~]$ fs -ls /user
-bash: fs: command not found
[hadoop@ip-172-31-77-128 ~]$ hadoop fs -ls /user
Found 7 items
drwxr-xr-x - hadoop hdfsadmingroup 0 2022-09-06 02:40 /user/csp554
drwxrwxrwx - hadoop hdfsadmingroup 0 2022-09-06 01:43 /user/hadoop
drwxr-xr-x - mapred mapred 0 2022-09-06 01:43 /user/history
drwxrwxrwx - hdfs hdfsadmingroup 0 2022-09-06 01:43 /user/hive
drwxrwxrwx - hue hue 0 2022-09-06 01:43 /user/hue
drwxrwxrwx - oozie oozie 0 2022-09-06 01:45 /user/oozie
drwxrwxrwx - root hdfsadmingroup 0 2022-09-06 01:43 /user/root
[hadoop@ip-172-31-77-128 ~]$
```

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

/user/csp554-2

Command executed- `hadoop fs -mkdir /user/csp554-2`



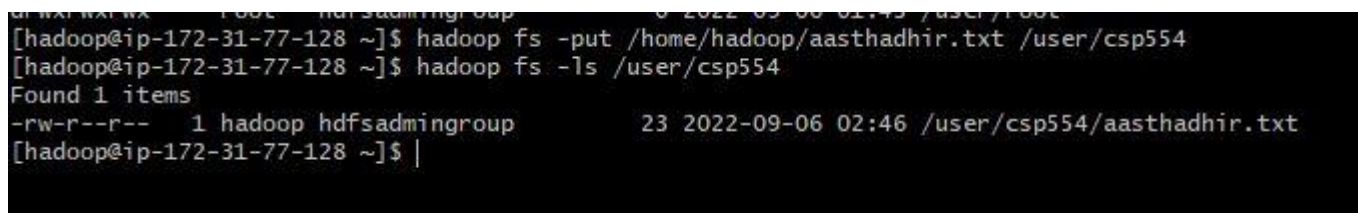
```
hadoop@ip-172-31-77-128:~  
[hadoop@ip-172-31-77-128 ~]$ hadoop fs -mkdir /user/csp554-2  
[hadoop@ip-172-31-77-128 ~]$ hadoop fs -ls /user  
Found 8 items  
drwxr-xr-x - hadoop hdfsadmingroup 0 2022-09-06 02:40 /user/csp554  
drwxr-xr-x - hadoop hdfsadmingroup 0 2022-09-06 02:43 /user/csp554-2  
drwxrwxrwx - hadoop hdfsadmingroup 0 2022-09-06 01:43 /user/hadoop  
drwxr-xr-x - mapred mapred 0 2022-09-06 01:43 /user/history  
drwxrwxrwx - hdfs hdfsadmingroup 0 2022-09-06 01:43 /user/hive  
drwxrwxrwx - hue hue 0 2022-09-06 01:43 /user/hue  
drwxrwxrwx - oozie oozie 0 2022-09-06 01:45 /user/oozie  
drwxrwxrwx - root hdfsadmingroup 0 2022-09-06 01:43 /user/root  
[hadoop@ip-172-31-77-128 ~]$ |
```

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`

Command executed- `hadoop fs -put /home/Hadoop/aasthadhir.txt /user/csp554`



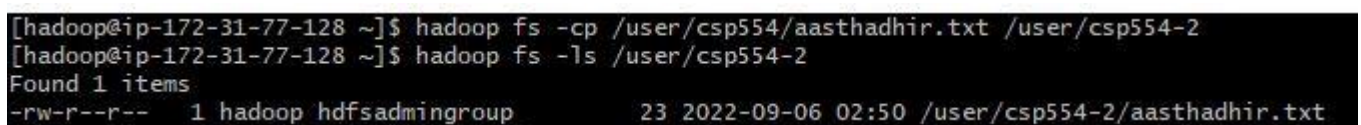
```
[hadoop@ip-172-31-77-128 ~]$ hadoop fs -put /home/hadoop/aasthadhir.txt /user/csp554  
[hadoop@ip-172-31-77-128 ~]$ hadoop fs -ls /user/csp554  
Found 1 items  
-rw-r--r-- 1 hadoop hdfsadmingroup 23 2022-09-06 02:46 /user/csp554/aasthadhir.txt  
[hadoop@ip-172-31-77-128 ~]$ |
```

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`

Command executed- `hadoop fs -cp /user/csp554/aasthadhir.txt /user/csp554-2`



```
[hadoop@ip-172-31-77-128 ~]$ hadoop fs -cp /user/csp554/aasthadhir.txt /user/csp554-2  
[hadoop@ip-172-31-77-128 ~]$ hadoop fs -ls /user/csp554-2  
Found 1 items  
-rw-r--r-- 1 hadoop hdfsadmingroup 23 2022-09-06 02:50 /user/csp554-2/aasthadhir.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.

Command executed- `aws s3 cp s3://a20468022-csp554/A20468022.txt /home/hadoop/A20468022.txt`


```
hadoop@ip-172-31-39-119:~  
[hadoop@ip-172-31-39-119 ~]$ aws s3 cp s3://a20468022-csp554/A20468022.txt /home  
/hadoop/A20468022.txt  
download: s3://a20468022-csp554/A20468022.txt to ./A20468022.txt  
[hadoop@ip-172-31-39-119 ~]$ ls  
A20468022.txt aasthadhir.txt  
[hadoop@ip-172-31-39-119 ~]$ cat A20468022.txt  
this is the id file  
[hadoop@ip-172-31-39-119 ~]$ |
```

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`

Note, the three slashes after the “hdfs:”

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

Command executed- `hadoop fs -cp s3://a20468022-csp554/A20468022.txt hdfs:///user/csp554-2`

```
[hadoop@ip-172-31-39-119 ~]$ hadoop fs -cp s3://a20468022-csp554/A20468022.txt h  
dfs:///user/csp554-2  
22/09/06 04:29:35 INFO s3n.S3NativeFileSystem: Opening 's3://a20468022-csp554/A2  
0468022.txt' for reading  
[hadoop@ip-172-31-39-119 ~]$ hadoop fs -ls /user/csp554-2  
Found 2 items  
-rw-r--r-- 1 hadoop hdfsadmin 21 2022-09-06 04:29 /user/csp554-2/  
A20468022.txt  
-rw-r--r-- 1 hadoop hdfsadmin 23 2022-09-06 04:24 /user/csp554-2/  
aasthadhir.txt  
[hadoop@ip-172-31-39-119 ~]$ |
```

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.

Command executed- `hadoop fs -cat /user/csp554-2/A20468022.txt`

```
hadoop@ip-172-31-39-119:~  
[hadoop@ip-172-31-39-119 ~]$ hadoop fs -cat /user/csp554-2/A20468022.txt  
this is the id file  
[hadoop@ip-172-31-39-119 ~]$ |
```

18. 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.

Command executed- `hadoop fs -rm /user/csp554-2/A20468022.txt`

```
[hadoop@ip-172-31-39-119 ~]$ hadoop fs -rm /user/csp554-2/A20468022.txt
Deleted /user/csp554-2/A20468022.txt
[hadoop@ip-172-31-39-119 ~]$ hadoop fs -ls /user/csp554-2
Found 1 items
-rw-r--r--  1 hadoop hdfsadmin  group      23 2022-09-06 04:24 /user/csp554-2/
aasthadhir.txt
[hadoop@ip-172-31-39-119 ~]$ |
```

Submitted By: -

Aastha Dhir

CWID- A20468022

adhir2@hawk.iit.edu