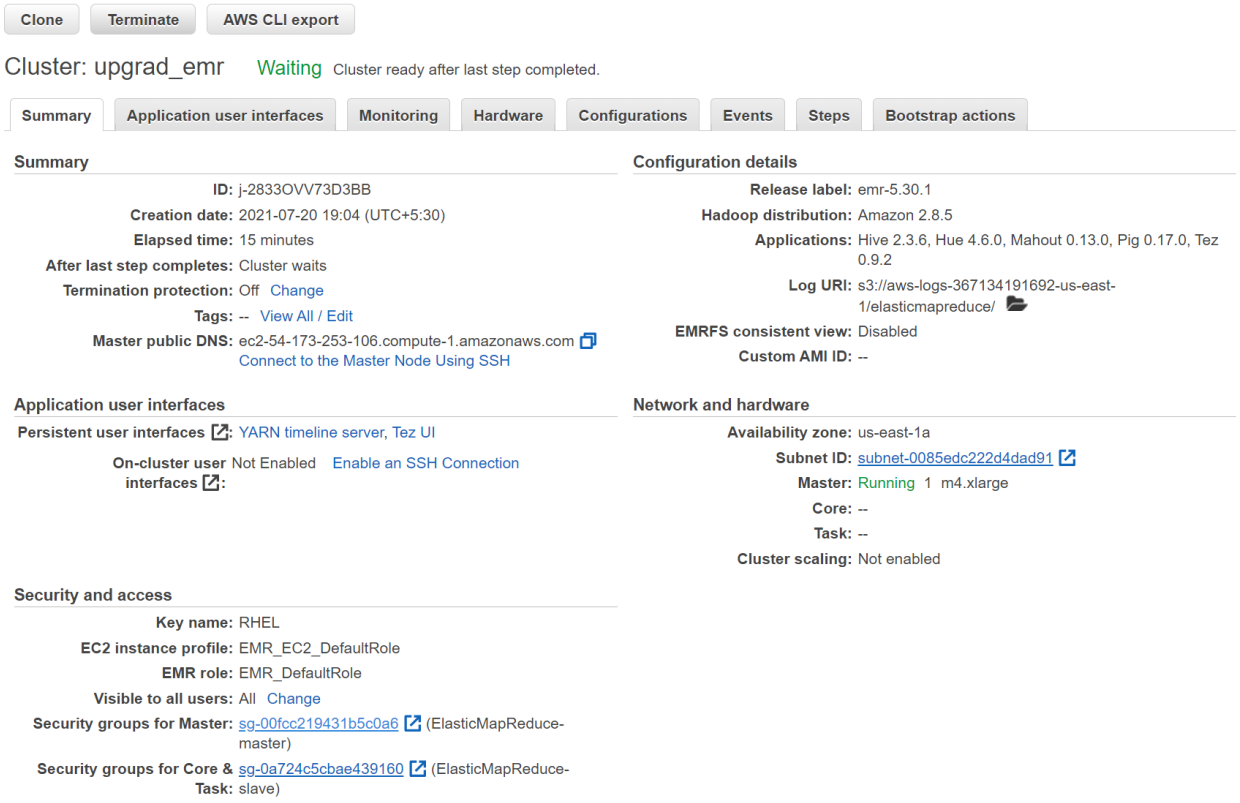
HDFS SHELL COMMANDS ON AMAZON EMR

Step 1: Start your AWS EMR instance by logging in to your AWS Management Console.



Step 2: Make sure that your instance is up and running fully. Connect to AWS EMR (via PuTTY, etc.). You learnt how to connect to AWS EMR in the previous modules.

The File System (FS) shell includes various shell-like commands that directly interact with the Hadoop Distributed File System (HDFS) as well as other file systems that Hadoop supports.

IMPORTANT INSTRUCTIONS

- The following notations have been used throughout the file:

```
[hadoop@ip-10-0-0-14 ~]$ hadoop command
Output of the command
```

As shown above, the command to be run is written in **bold.** The output of the command is written in *italics*. The [**hadoop**@ip-10-0-0-14 ~] tells us the user through which the command is to be executed.

- Please be careful with the spaces in the commands.
- If a series of commands is given in a particular order, make sure that you run them in the same order.

BASIC COMMANDS

- To check the commands that are available in the HDFS, run any of the following commands.

hadoop fs -help or **hadoop dfs -help**

- To read the list of files in the HDFS, use the 'ls' command.

```
[hadoop@ip-10-0-0-14 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hadoop          0 2021-07-20 13:41 /apps
drwxrwxrwt   - hdfs hadoop          0 2021-07-20 13:42 /tmp
drwxr-xr-x   - hdfs hadoop          0 2021-07-20 13:41 /user
drwxr-xr-x   - hdfs hadoop          0 2021-07-20 13:41 /var
```

- The 'sudo -i' command is used to switch from the hadoop to the root user. Also the superuser while using root is hdfs. The 'su -hdfs' helps us switch from the root user to the hdfs user when in the root user. To switch from the hdfs user to the root user, type 'exit'.

```
[hadoop@ip-10-0-0-14 ~]$ sudo -i
[root@ip-10-0-0-14 ~]# su - hdfs
-bash-4.2$ exit
```

```
[root@ip-10-0-0-14 ~]#
```

- **df**: This is a command to check the available space in the HDFS.

  du: This will help you check the space usage of the HDFS.

  Please note that both the commands can be run only from the hdfs user.

```
-bash-4.2$ hadoop fs -df -h
Filesystem     Size      Used  Available  Use%
hdfs://ip-172-31-32-207.ec2.internal:8020  34.5 G  810.8 M  33.4 G 2%
-bash-4.2$ hadoop fs -du -s -h /
799.8 M  /
```

```
-bash-4.2$ hadoop fs -df -h
Filesystem                                    Size     Used  Available  Use%
hdfs://ip-172-31-32-207.ec2.internal:8020  34.5 G  810.8 M     33.4 G    2%
-bash-4.2$ hadoop fs -du -s -h /
799.8 M  /
-bash-4.2$
```

Create a directory inside the HDFS

- The commands used below demonstrate how to create a directory in the HDFS.

```
[hadoop@ip-172-31-32-207 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hadoop          0 2021-07-20 13:41 /apps
drwxrwxrwt   - hdfs hadoop          0 2021-07-20 13:42 /tmp
drwxr-xr-x   - hdfs hadoop          0 2021-07-20 13:41 /user
drwxr-xr-x   - hdfs hadoop          0 2021-07-20 13:41 /var
[hadoop@ip-172-31-32-207 ~]$ hadoop fs -mkdir /user/hadoop/testdir_1
```

- You can verify the same using the command shown below.

```
[hadoop@ip-172-31-32-207 ~]$ hadoop fs -ls /user/hadoop
Found 1 items
drwxr-xr-x   - hadoop hadoop          0 2021-07-20 15:08 /user/hadoop/testdir_1
```

- Note that whenever you are performing a job using **root** user, then make sure that you are using the root directory i.e., '**/user/root**' in HDFS and similarly if you are performing a job using **hadoop** then make sure that you are using the hadoop directory, i.e. '**/user/hadoop**' in HDFS, for your operations.

Creating a file using a root user

- First, we create a file using the 'cat' command as shown below. After entering the contents of the file, we use 'Ctrl+Z' to save and exit the file.

```
[hadoop@ip-172-31-32-207 ~]$ cat > test.txt
```

- You can also use the "**vi test.txt**" command to create a text file using vi if you prefer. Keep in mind that you will have to go to the Input mode by pressing I and then write into the file and then later on press "Esc" to go back to the command mode and then type "wq!" to save and exit vi.
- Now verify whether the file has been created or not using the 'ls' command.

```
[hadoop@ip-172-31-32-207 ~]$ ls
test.txt
```

Copy a file in the local file system to the HDFS

- Now, we will use the 'put' command to copy the file created above from the local file system to the HDFS. The syntax for the put command is:

  hadoop fs -put <src> <destination>

```
[hadoop@ip-172-31-32-207 ~]$ hadoop fs -put test.txt /user/hadoop/
```

- We can verify whether the file has been copied as shown below:

```
[hadoop@ip-172-31-32-207 ~]$ hadoop fs -ls /user/hadoop/
Found 2 items
-rw-r--r--   1 hadoop hadoop         21 2021-07-20 15:24 /user/hadoop/test.txt
drwxr-xr-x   - hadoop hadoop          0 2021-07-20 15:08 /user/hadoop/testdir_1
```


```
[hadoop@ip-172-31-32-207 ~]$ hadoop fs -ls /user/hadoop/
Found 2 items
-rw-r--r--   1 hadoop hadoop         21 2021-07-20 15:24 /user/hadoop/test.txt
drwxr-xr-x   - hadoop hadoop          0 2021-07-20 15:08 /user/hadoop/testdir_1
```

- Now, check the content of the file, using the ‘cat’ command.

```
[hadoop@ip-172-31-32-207 ~]$ hadoop fs -cat /user/hadoop/test.txt
```

```
[hadoop@ip-172-31-32-207 ~]$ hadoop fs -cat /user/hadoop/test.txt
This is a test file.
[hadoop@ip-172-31-32-207 ~]$
```

Copy a file to the local file system from the HDFS

- We will now create a new directory in our local file system. Then we will copy a file from the HDFS to this local file system using the get command. Please note that in our case we are using the same file we had copied from the local file system to the HDFS. However, in this case it is copied to a new directory.

  Syntax: **hadoop fs -get <src> <destination>**

- First, we create a new directory named testing using the 'mkdir' command.

```
[hadoop@ip-172-31-32-207 ~]$ mkdir testing
```

- Now, we will copy the file from the HDFS to the local system using the 'get' command.

```
[hadoop@ip-172-31-32-207 ~]$ hadoop fs -get /user/hadoop/test.txt ./testing
```

- Now, let us verify the same by navigating to the new directory using the 'cd' command. Then use the 'ls' command and verify whether your file is present or not.

```
[hadoop@ip-172-31-32-207 ~]$ cd testing/
[hadoop@ip-172-31-32-207 testing]$ ls
test.txt
```

Change the replication factor of a particular file

- Please note that these steps are just for practise and should not be done while doing regular practise
- As we know, the default replication factor in the HDFS is 3. Now, we will use the 'setrep' command to change it to any value desired by us. In this case, we are setting the replication factor of the file test.txt to 6.

```
[hadoop@ip-172-31-32-207 testing]$ hadoop fs -setrep 6 /user/hadoop/test.txt
Replication 6 set: /user/hadoop/test.txt
```