COSC2637/2633 Big Data Processing
Lab 1: Tutorial/Lab class

Objectives

- Access your individual jumphost that has already been setup for each student
- Create, access and then terminate an AWS EMR cluster
- Responsible use of cloud resources

Introduction

In order to use AWS EMR, RMIT ITS has created a relatively inexpensive jumphost on AWS for each student enrolled in the course. From this jumphost you will be to create, access and manage an EMR cluster. Creating a cluster requires allocation of additional physical machines in the AWS datacentre (which takes minutes to happen, depending on the size of cluster). However, as the EMR clusters are more expensive machines (and cost by time that the cluster is left set up), it is important that clusters are not left running after you have finished using them (and saved any output). So, at the end of a lab class, it is **very important** you **terminate** your cluster.

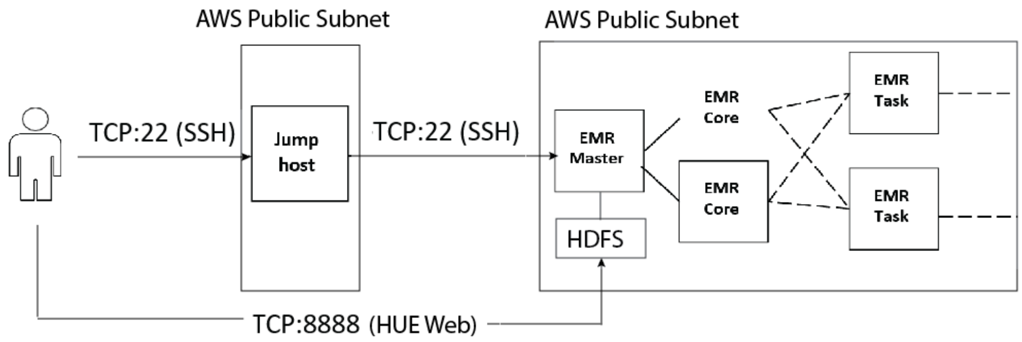
The jumphost and any EMR clusters launched will all be in the AWS US East Region (N.Virginia) (Also known as "us-east-1" or "Standard"). This means any clusters you create will have access to the Public Data Sets hosted in S3 that are provided by AWS in that region.

Example of student jumphost DNS: `sXXXXXXX.jump.cosc2637.route53.aws.rmit.edu.au`

Each student will have a personal SSH key and it will provide access to your jumphost and EMR cluster.

You can run `./create_cluster.sh` from your jumphost to launch your EMR cluster. After the script finishes running, there will be further instructions on how to access that specific cluster (Hue and Hadoop Master node).

Please **do not forget** to run `./terminate_cluster.sh` (from your jumphost) each time you have finished using your cluster (e.g., at the end of the lab class).



The steps to get you on to the AWS environment (For Windows OS users, please see another document)

Any feedback or suggestions about these instructions should be posted on the Course Canvas Discussion Board “Technical questions concerning AMS EMR” via myRMIT.

Keys (these should be emailed to you at or prior to your practical class)

Save the key location on your device

- For MacOSx/Linux, change permissions:

```
$chmod 600 sxxxxxx-cosc2637.pem
```

(replace *sXXXXXXX* with your student number, also shown in the key sent to you from lecturer)

Links for more information:

http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AccessingInstances.html

http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AccessingInstancesLinux.html

http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/putty.html

Configuration required to prevent the inactive shell phenomenon

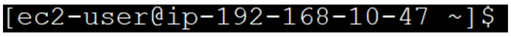
- **For MacOSx/Linux:** add the following settings into *~/.ssh/config*

```
ServerAliveInterval 15
TCPKeepAlive yes
```

SSH into your jumpbox

```
ssh ec2-user@sXXXXXXX.jump.cosc2637.route53.aws.rmit.edu.au -i sxxxxxx-cosc2637.pem
```

Accept the host, now you are in the jumpbox if you see something like:



```
[ec2-user@ip-192-168-10-47 ~]$
```

From here you can run the start and terminate script as needed. There are several scripts and other files in the directory on the jumphost that should not be changed, otherwise you may not be able to create and manage your cluster.

Start a cluster

Start the cluster using either of the following commands:

```
$ ./create_cluster.sh
```

or

```
$ sh create_cluster.sh
```

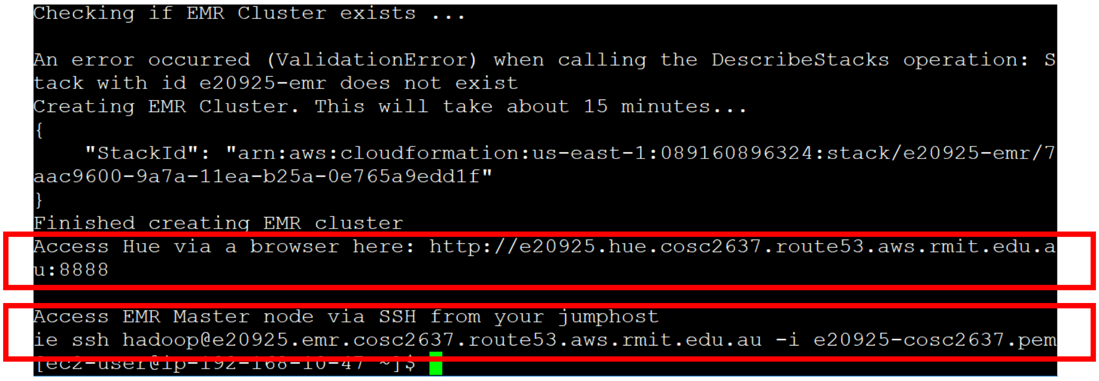
The script begins by checking whether a cluster already exists, the following message indicates it does not:

```
An error occurred (ValidationError) when calling the DescribeStacks
operation: Stack with id e23270-emr does not exist
```

Then, wait 15 min (it can be longer) for the cluster to start up, you will see when the cluster is ready

Access the cluster

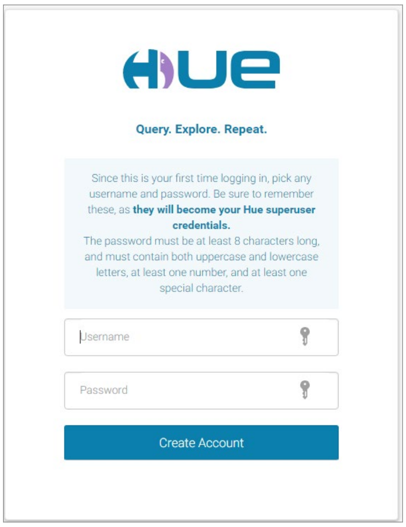
You will be shown the DNS for the cluster and the web link now.



```
Checking if EMR Cluster exists ...

An error occurred (ValidationError) when calling the DescribeStacks operation: S
tack with id e20925-emr does not exist
Creating EMR Cluster. This will take about 15 minutes...
{
    "StackId": "arn:aws:cloudformation:us-east-1:089160896324:stack/e20925-emr/7
aac9600-9a7a-11ea-b25a-0e765a9edd1f"
}
Finished creating EMR cluster
Access Hue via a browser here: http://e20925.hue.cosc2637.route53.aws.rmit.edu.a
u:8888

Access EMR Master node via SSH from your jumphost
ie ssh hadoop@e20925.emr.cosc2637.route53.aws.rmit.edu.au -i e20925-cosc2637.pem
[ec2-user@ip-192-168-10-47 ~]$
```

- Use your web browser to use the Hadoop environment via Hue. You will need to setup an account each time as the cluster is new.

http://sXXXXXXX.hue.cosc2637.route53.aws.rmit.edu.au:8888

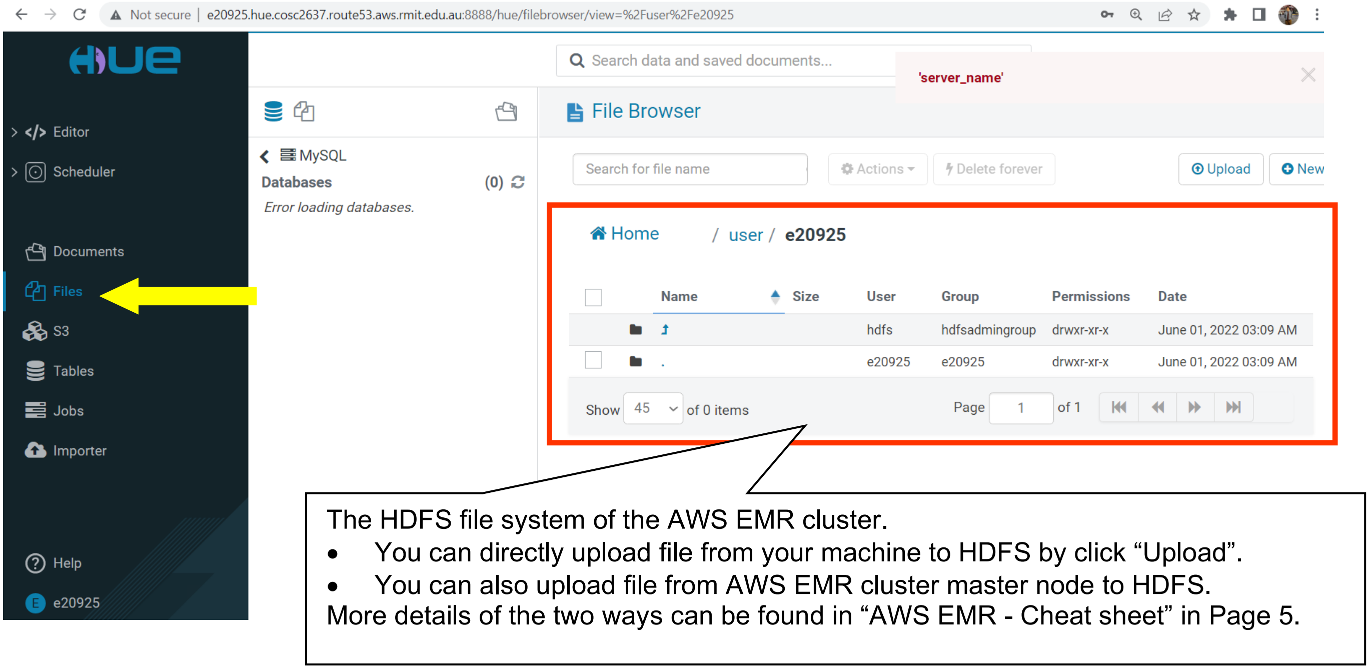


Query. Explore. Repeat.

Since this is your first time logging in, pick any username and password. Be sure to remember these, as **they will become your Hue superuser credentials.**

The password must be at least 8 characters long, and must contain both uppercase and lowercase letters, at least one number, and at least one special character.

For example, my Hue account is “e20925”. After logging in HUE, click Files, I will see the HDFS file system like below:



Not secure | e20925.hue.cosc2637.route53.aws.rmit.edu.au:8888/hue/filebrowser/view=%2Fuser%2Fe20925

'server_name'

Databases (0)

Error loading databases.

Home / user / e20925

	Name	Size	User	Group	Permissions	Date
	↥		hdfs	hdfsadmingroup	drwxr-xr-x	June 01, 2022 03:09 AM
□	.		e20925	e20925	drwxr-xr-x	June 01, 2022 03:09 AM

Show 45 of 0 items Page 1 of 1

The HDFS file system of the AWS EMR cluster.

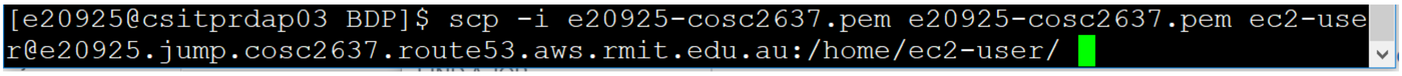
- You can directly upload file from your machine to HDFS by click “Upload”.
- You can also upload file from AWS EMR cluster master node to HDFS.

More details of the two ways can be found in “AWS EMR - Cheat sheet” in Page 5.

- Connect to the cluster via SSH

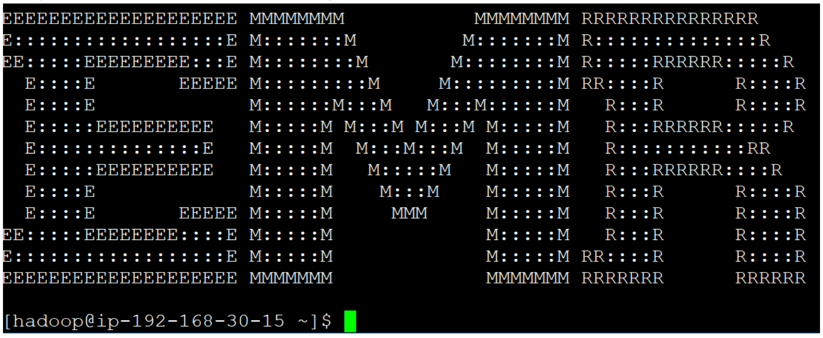
`ssh hadoop@e20925.emr.cosc2637.route53.aws.rmit.edu.au -i sxxxxxxx-cosc2637.pem`

Note before running above ssh, please make sure `sxxxxxxx-cosc2637.pem` in the jumphost. If it does not, you need run the following line on your own machine (replace e20925 by your student id sxxxxxxx).



```
[e20925@csitprdap03 BDP]$ scp -i e20925-cosc2637.pem e20925-cosc2637.pem ec2-user@e20925.jump.cosc2637.route53.aws.rmit.edu.au:/home/ec2-user/
```

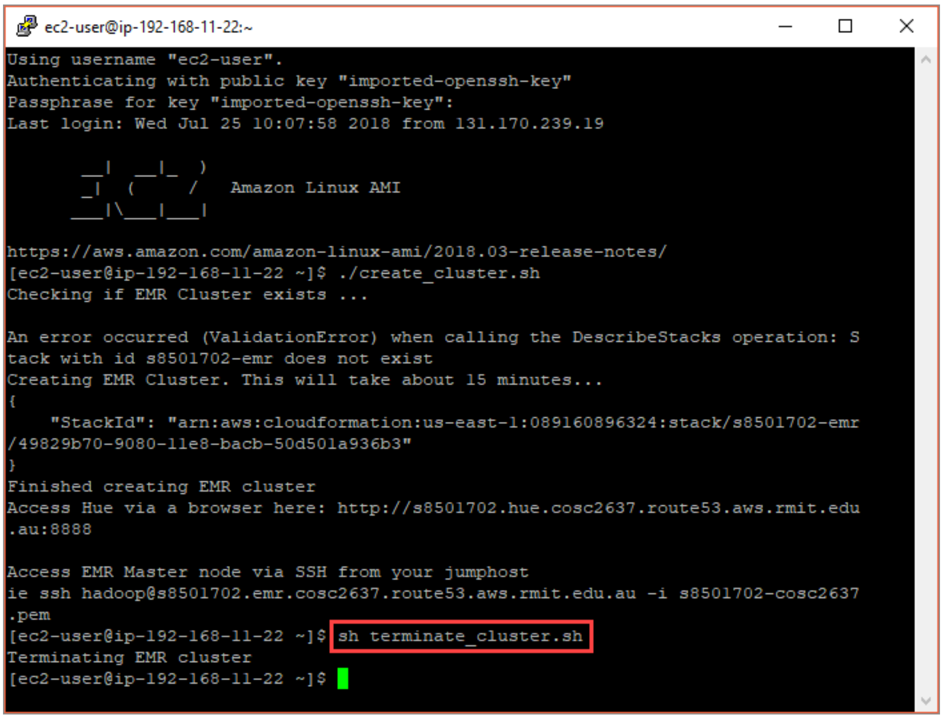
You are in the cluster, if you can see something like



```
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-192-168-30-15 ~]$
```

Leave the cluster by inputting “exit”, then you will return back to Jumphost.

Don’t forget to **shutdown the cluster using terminate_cluster.sh**. The shutting down will take about 5 minutes during which you’ll still be able to access it via the browser.

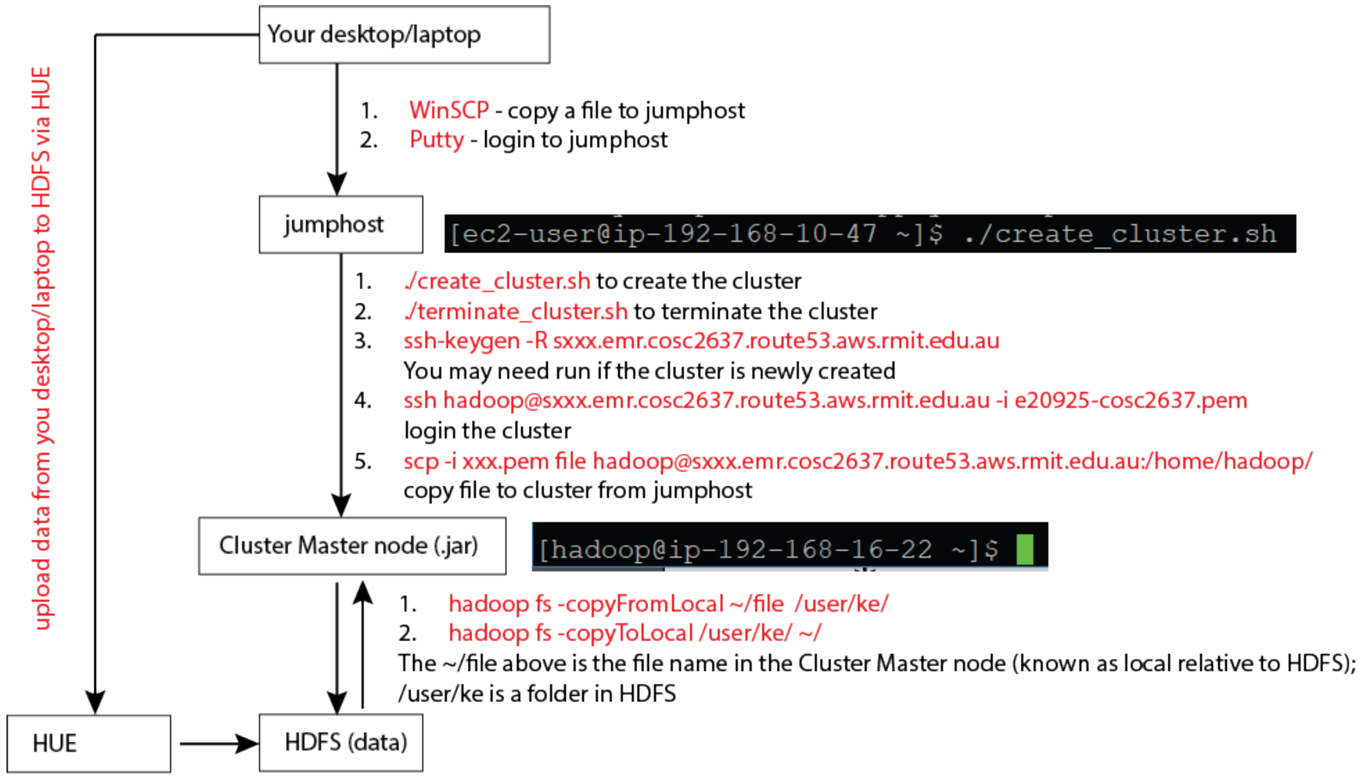


```
Using username "ec2-user".
Authenticating with public key "imported-openssh-key"
Passphrase for key "imported-openssh-key":
Last login: Wed Jul 25 10:07:58 2018 from 131.170.239.19

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
[ec2-user@ip-192-168-11-22 ~]$ ./create_cluster.sh
Checking if EMR Cluster exists ...

An error occurred (ValidationError) when calling the DescribeStacks operation: S
tack with id s8501702-emr does not exist
Creating EMR Cluster. This will take about 15 minutes...
{
    "StackId": "arn:aws:cloudformation:us-east-1:089160896324:stack/s8501702-emr
/49829b70-9080-11e8-bacb-50d501a936b3"
}
Finished creating EMR cluster
Access Hue via a browser here: http://s8501702.hue.cosc2637.route53.aws.rmit.edu
.au:8888

Access EMR Master node via SSH from your jumphost
ie ssh hadoop@s8501702.emr.cosc2637.route53.aws.rmit.edu.au -i s8501702-cosc2637
.pem
[ec2-user@ip-192-168-11-22 ~]$ sh terminate_cluster.sh
Terminating EMR cluster
[ec2-user@ip-192-168-11-22 ~]$
```

*********Useful Information************

It is a good idea to do just prior to practical class so that it is up and running. However, do NOT leave it running for days at a time as it is expensive and a waste of resources to have idle clusters that are not being used. When finished, don't forget to **shutdown the cluster** to avoid charges. Note that there is no indication when the cluster is terminated.

AWS EMR - Cheat sheet



Your desktop/laptop
jumphost
Cluster Master node (.jar)
HUE
HDFS (data)
upload data from you desktop/laptop to HDFS via HUE

1. WinSCP - copy a file to jumphost
2. Putty - login to jumphost

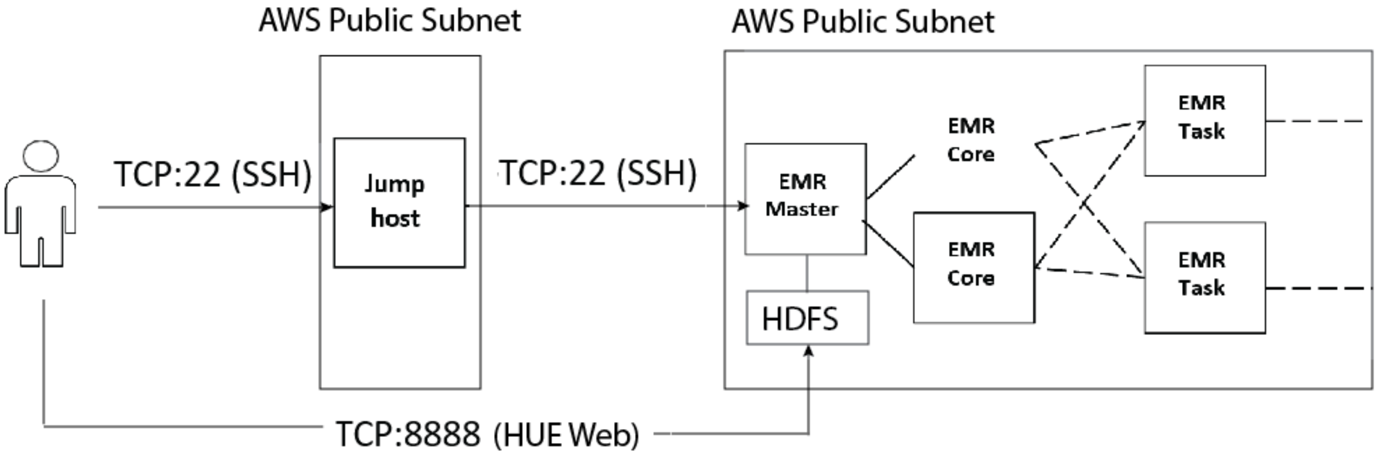
```
[ec2-user@ip-192-168-10-47 ~]$ ./create_cluster.sh
```

1. ./create_cluster.sh to create the cluster
2. ./terminate_cluster.sh to terminate the cluster
3. ssh-keygen -R sxxx.emr.cosc2637.route53.aws.rmit.edu.au
   You may need run if the cluster is newly created
4. ssh hadoop@sxxx.emr.cosc2637.route53.aws.rmit.edu.au -i e20925-cosc2637.pem
   login the cluster
5. scp -i xxx.pem file hadoop@sxxx.emr.cosc2637.route53.aws.rmit.edu.au:/home/hadoop/
   copy file to cluster from jumphost

```
[hadoop@ip-192-168-16-22 ~]$
```

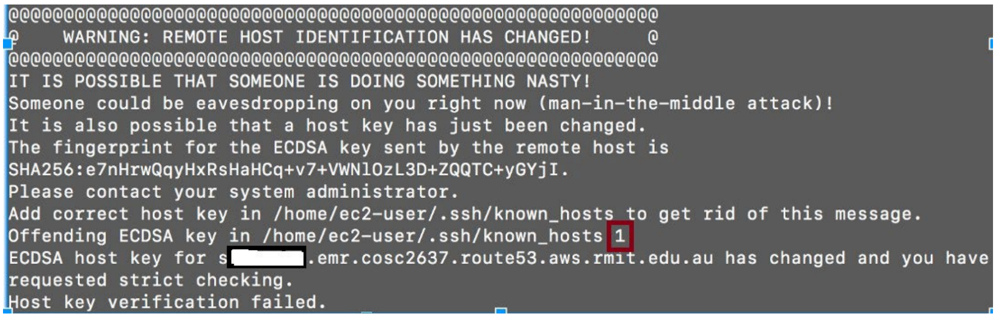
1. hadoop fs -copyFromLocal ~/file /user/ke/
2. hadoop fs -copyToLocal /user/ke/ ~/

The ~/file above is the file name in the Cluster Master node (known as local relative to HDFS); /user/ke is a folder in HDFS



A common issue of cluster login

You may see the following error when ssh from your jumphost to the hadoop master node,



```
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@    WARNING: REMOTE HOST IDENTIFICATION HAS CHANGED!     @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
IT IS POSSIBLE THAT SOMEONE IS DOING SOMETHING NASTY!
Someone could be eavesdropping on you right now (man-in-the-middle attack)!
It is also possible that a host key has just been changed.
The fingerprint for the ECDSA key sent by the remote host is
SHA256:e7nHrwQqyHxRsHaHCq+v7+VWNlOzL3D+ZQQTC+yGYjI.
Please contact your system administrator.
Add correct host key in /home/ec2-user/.ssh/known_hosts to get rid of this message.
Offending ECDSA key in /home/ec2-user/.ssh/known_hosts 1
ECDSA host key for s[redacted].emr.cosc2637.route53.aws.rmit.edu.au has changed and you have
requested strict checking.
Host key verification failed.
```

Every time a new cluster is created it is a new set of hosts so the ssh host key changes, and ssh gives you a warning that it has been changed. To fix the problem, please run the following:

$ssh-keygen -R sxxxxxxx.emr.cosc2637.route53.aws.rmit.edu.au