# COSC2637/2633 Big Data Processing
# Lab 1: Tutorial/Lab class (Windows User)

**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.

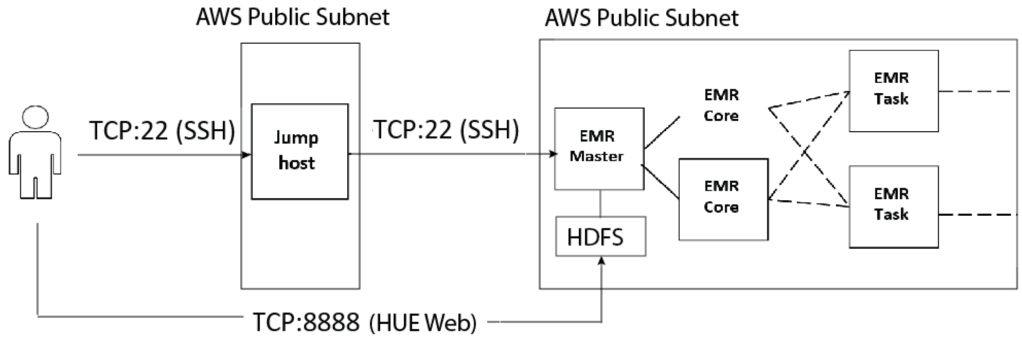
The jumphost and any EMR clusters launched will all be in the AWS US East Region (N.Virginia) (Also known as "us-east-1" or "Standard"). This means any clusters you create will have access to the Public Data Sets hosted in S3 that are provided by AWS in that region.

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 (Windows)

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

1. Save the key location on your device
2. Convert the key using PuTTYGen in Windows OS

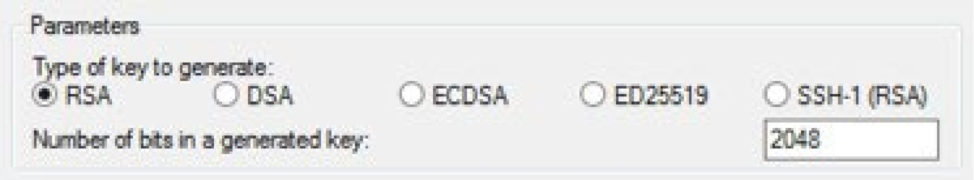
**Converting Your Private Key Using PuTTYgen**

PuTTY does not natively support the private key format (.pem) generated by Amazon EC2. PuTTY has a tool named PuTTYgen, which can convert keys to the required PuTTY format (.ppk). You must convert your private key into this format (.ppk) before attempting to connect to your instance using PuTTY.

**To convert your private key**

a. Start PuTTYgen (for example, from the **Start** menu, choose **All Programs > PuTTY > PuTTYgen**).

b. Under **Type of key to generate**, choose **RSA**.

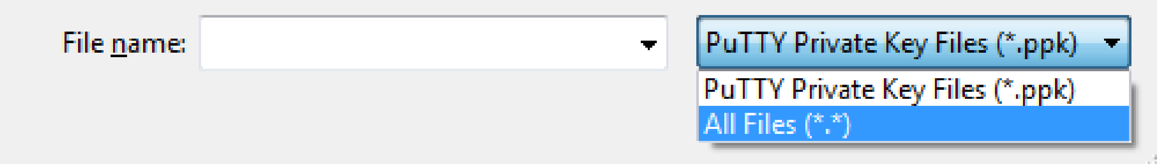


Parameters

Type of key to generate:
● RSA ○ DSA ○ ECDSA ○ ED25519 ○ SSH-1 (RSA)

Number of bits in a generated key: 2048

c. Choose **Load**. By default, PuTTYgen displays only files with the extension .ppk. To locate your .pem file, select the option to display files of all types.



File name:

PuTTY Private Key Files (*.ppk)

PuTTY Private Key Files (*.ppk)
All Files (*.*)

d. Select your .pem file for the key pair that you specified when you launched your instance, and then choose **Open**. Choose **OK** to dismiss the confirmation dialog box.

e. Choose **Save private key** to save the key in the format that PuTTY can use. PuTTYgen displays a warning about saving the key without a passphrase. Choose **Yes**.

Note: A passphrase on a private key is an extra layer of protection, so even if your private key is discovered, it can't be used without the passphrase. The downside to using a passphrase is that it makes automation harder because human intervention is needed to log on to an instance or copy files to an instance.

f. Specify the same name for the key (.pem). PuTTYgen automatically adds the .ppk file extension.

Your private key is now in the correct format for use with PuTTY. You can now connect to your instance using PuTTY's SSH client.

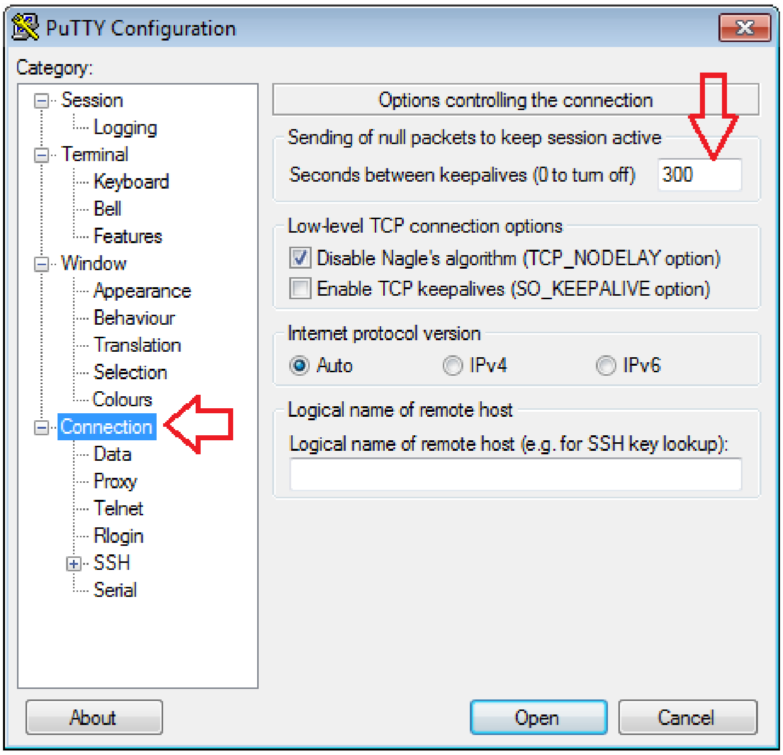
Links for more information:
*http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AccessingInstances.html*

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

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

## 3. Set up PuTTY connection

a. Configuration required to prevent the inactive shell phenomenon



Options controlling the connection

Sending of null packets to keep session active

Seconds between keepalives (0 to turn off) 300

Low-level TCP connection options

☑ Disable Nagle's algorithm (TCP_NODELAY option)

☐ Enable TCP keepalives (SO_KEEPALIVE option)

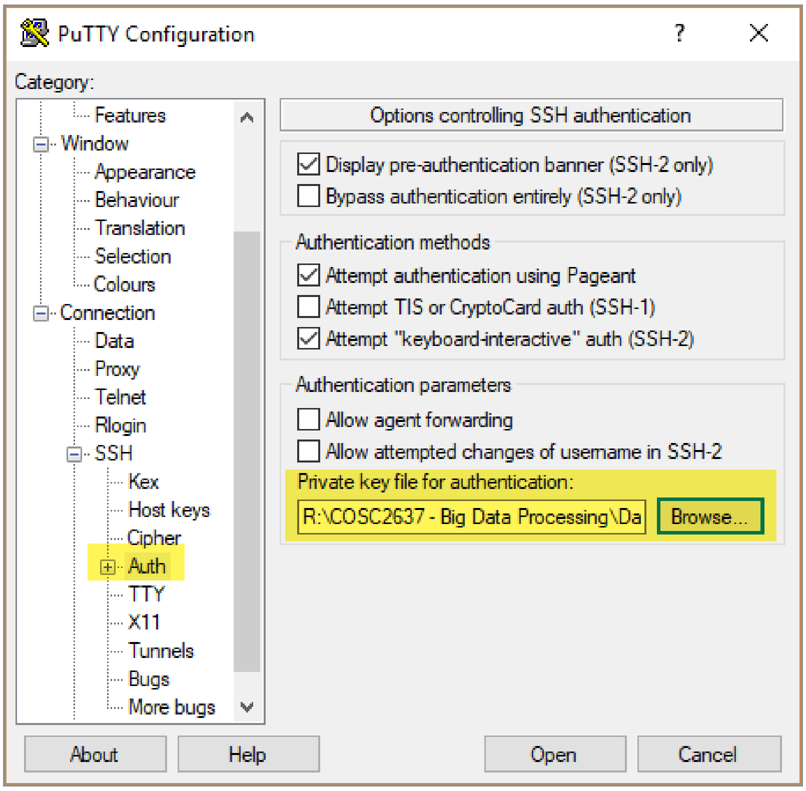
Internet protocol version

● Auto ○ IPv4 ○ IPv6

Logical name of remote host

Logical name of remote host (e.g. for SSH key lookup):

b. Add private key you saved earlier (extracted from the key pair file emailed to you)



Options controlling SSH authentication

☑ Display pre-authentication banner (SSH-2 only)

☐ Bypass authentication entirely (SSH-2 only)

Authentication methods

☑ Attempt authentication using Pageant

☐ Attempt TIS or CryptoCard auth (SSH-1)

☑ Attempt "keyboard-interactive" auth (SSH-2)

Authentication parameters

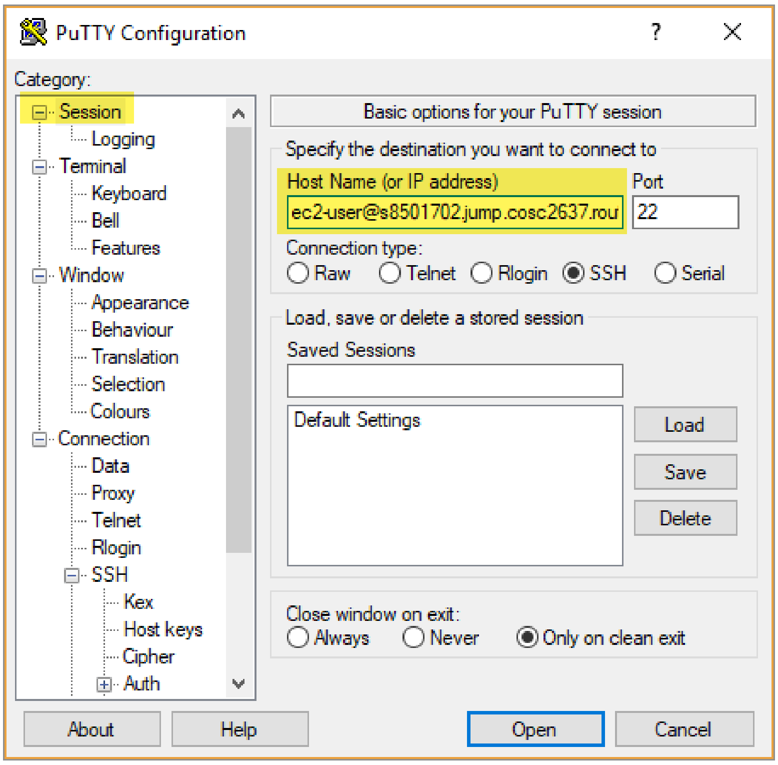
☐ Allow agent forwarding

☐ Allow attempted changes of username in SSH-2

Private key file for authentication:

R:\COSC2637 - Big Data Processing\Da Browse...

**c.** Enter jump host name:
`ec2-user@sXXXXXXX.jump.cosc2637.route53.aws.rmit.edu.au`
Where XXXXXXX is your student number



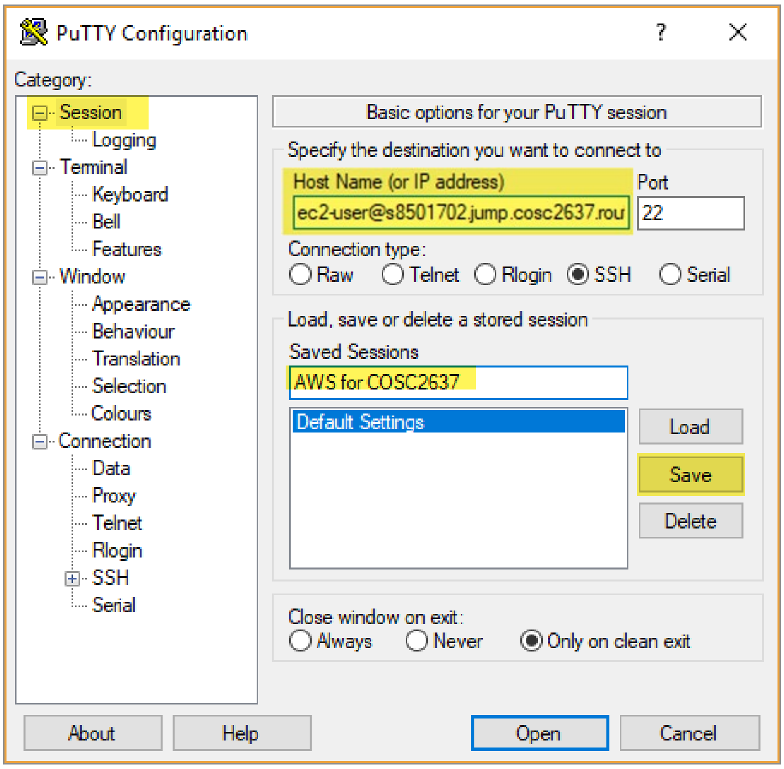
Host Name (or IP address): ec2-user@s8501702.jump.cosc2637.rou

Port: 22

Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Close window on exit: ○ Always ○ Never ● Only on clean exit

d. Save settings



Host Name (or IP address): ec2-user@s8501702.jump.cosc2637.rou

Port: 22

Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

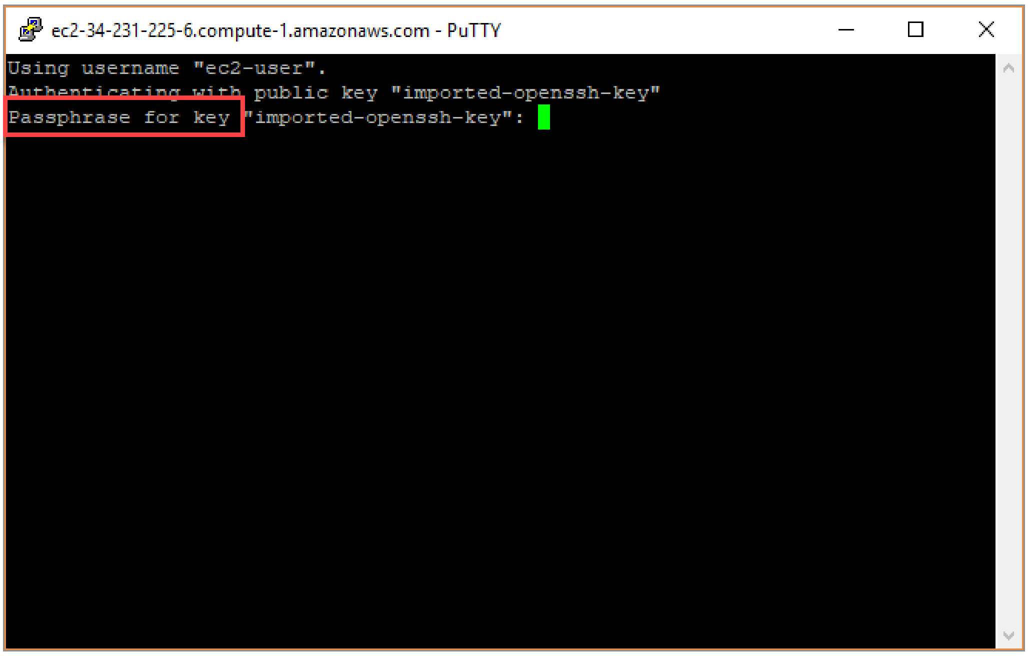
Saved Sessions: AWS for COSC2637

Close window on exit: ○ Always ○ Never ● Only on clean exit

4. Open PuTTY connection

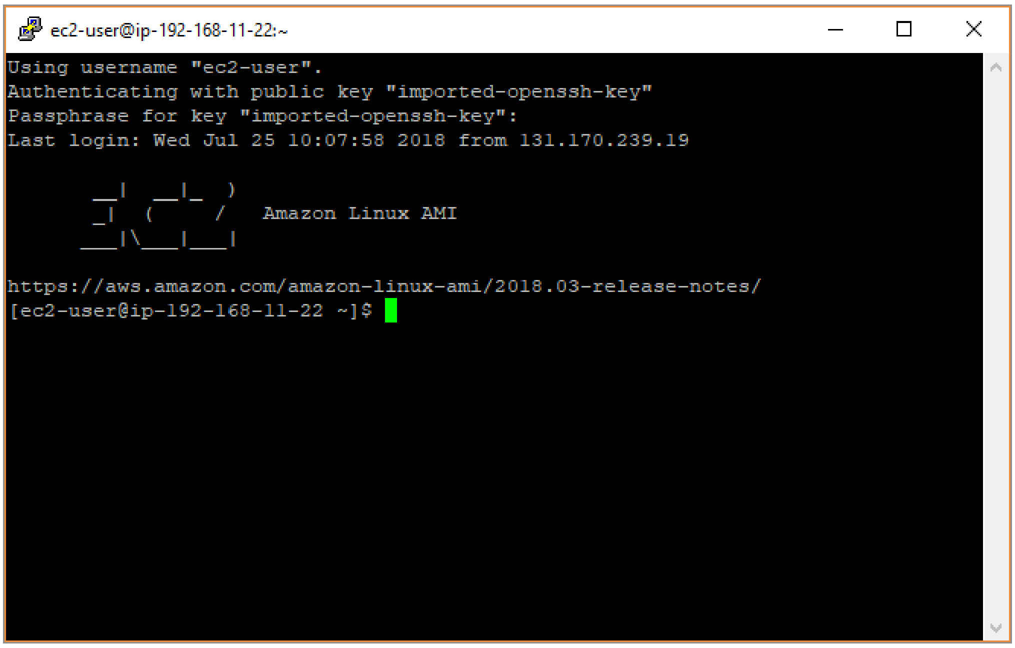
   Open button

5. If you entered a passphrase when saving the private key, you will be prompted for it now.



```
Using username "ec2-user".
Authenticating with public key "imported-openssh-key"
Passphrase for key "imported-openssh-key":
```

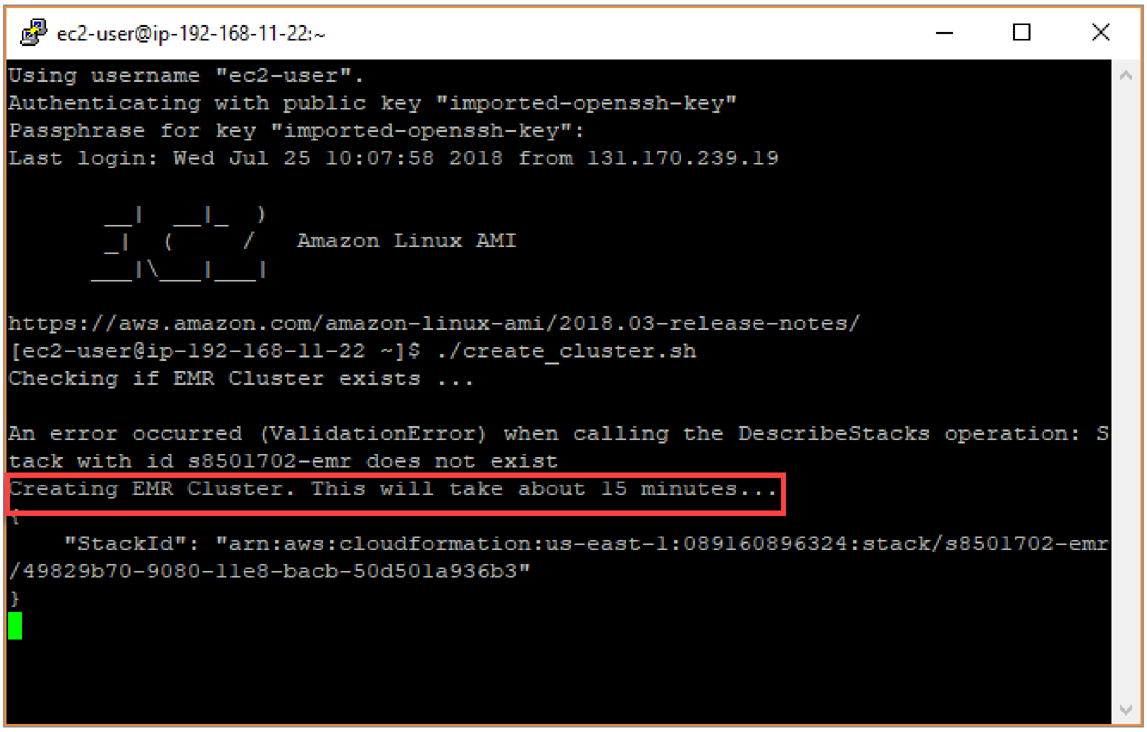
6. You should now have the AWS $ prompt



```
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 ~]$
```

7. Enter `./create_cluster.sh`

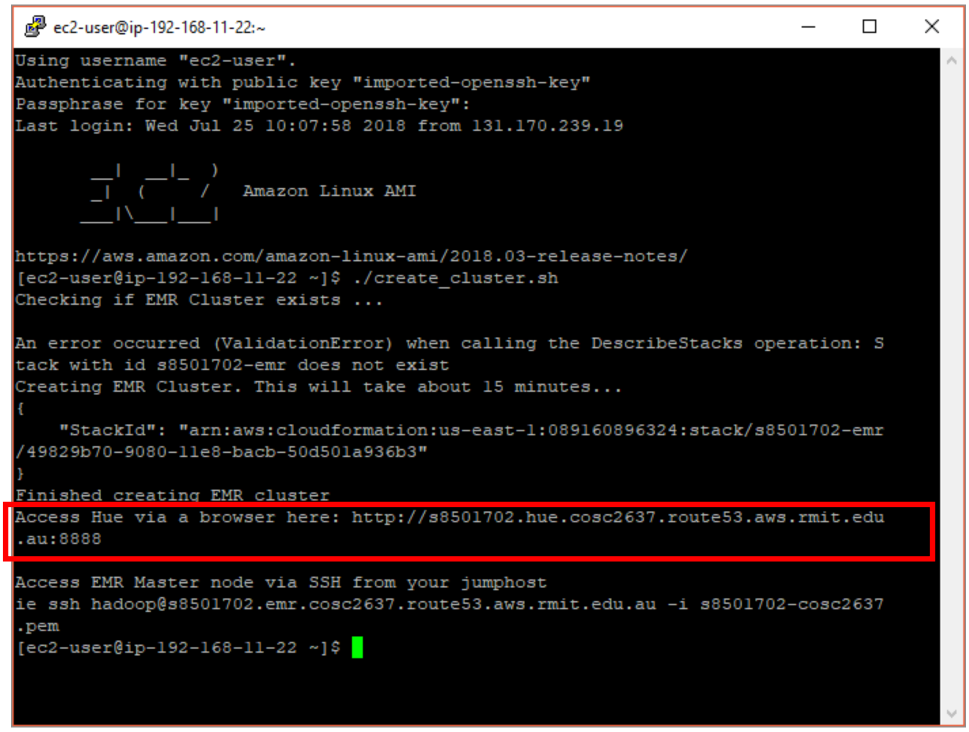
   Note: `sh create_cluster.sh` also works



```
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"
}
```

The EMR cluster will now be created. **It can take 15 minutes or more.** The script begins by checking whether a cluster already exists, so the ValidationError is expected as it is simply saying that the cluster doesn't already exist.

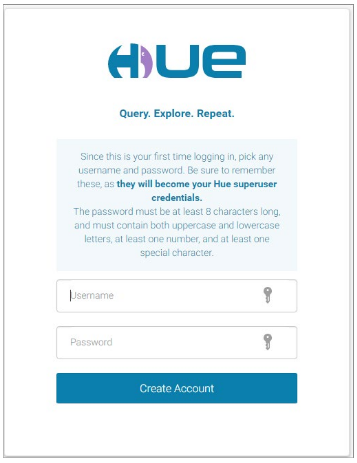
8. Once the cluster is created, it will show the URL you need use to connect to your cluster



```
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 ~]$
```

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

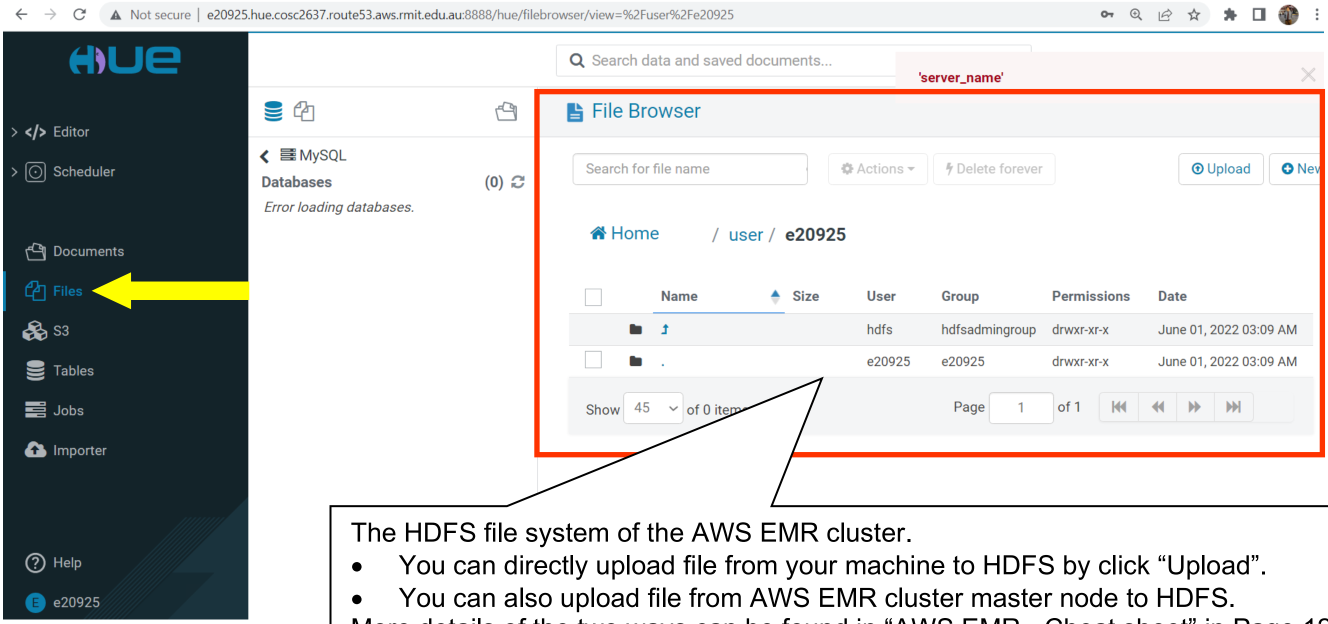
http://sXXXXXXX.hue.cosc2637.route53.aws.rmit.edu.au:8888



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

*Error loading databases.*

File Browser

| | 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

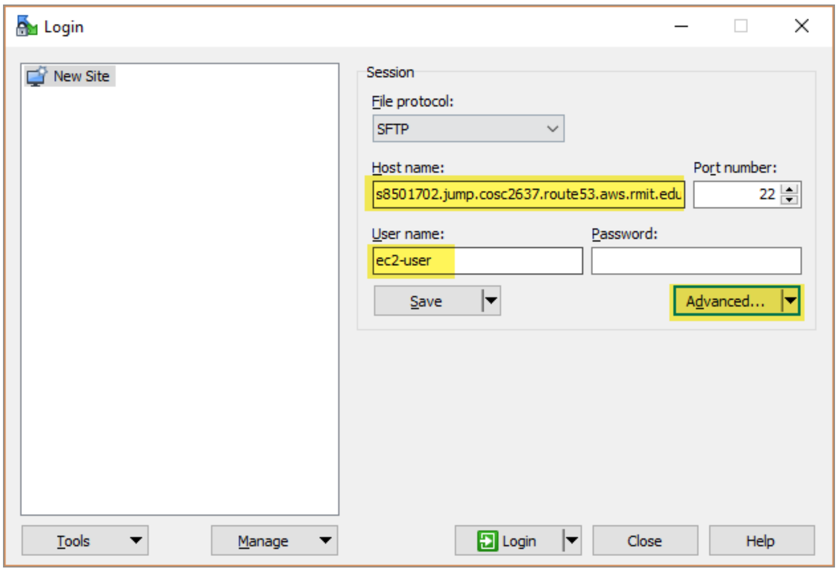
The HDFS file system of the AWS EMR cluster.

- 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 12.

**Using WinSCP for transferring key to JumpHost home (the key should be the xxx-xxx.pem emailed to you at or prior to your practical class)**

1. Enter jump host name, user and Click Advanced, then select Advanced



File protocol: SFTP

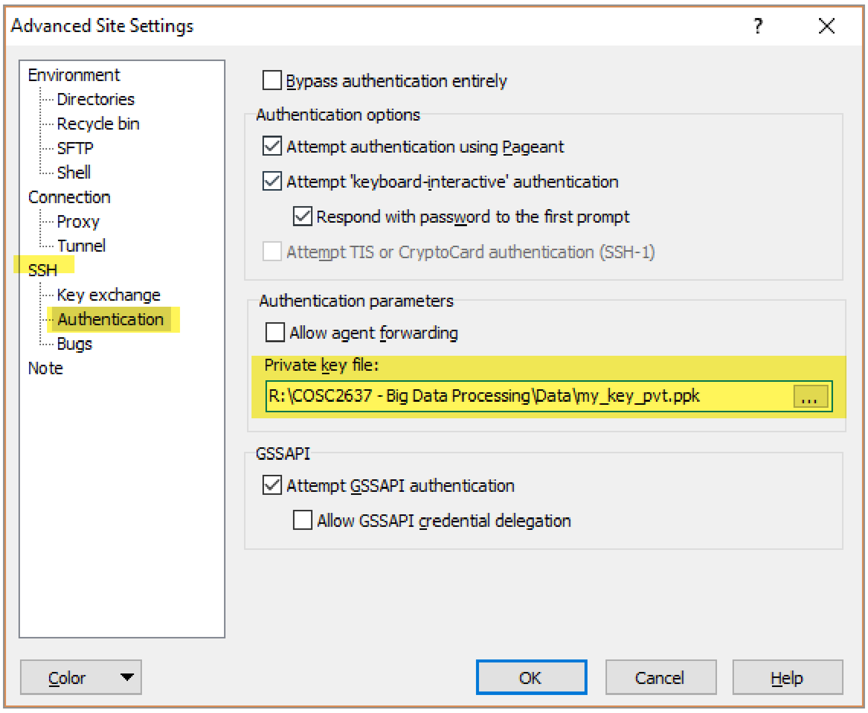
Host name: s8501702.jump.cosc2637.route53.aws.rmit.edu

Port number: 22

User name: ec2-user

Password:

2. Under SSH |Authentication, browse to the private key you extracted from the emailed key pair



□ Bypass authentication entirely

Authentication options

☑ Attempt authentication using Pageant

☑ Attempt 'keyboard-interactive' authentication

☑ Respond with password to the first prompt

□ Attempt TIS or CryptoCard authentication (SSH-1)

Authentication parameters

□ Allow agent forwarding

Private key file: R:\COSC2637 - Big Data Processing\Data\my_key_pvt.ppk

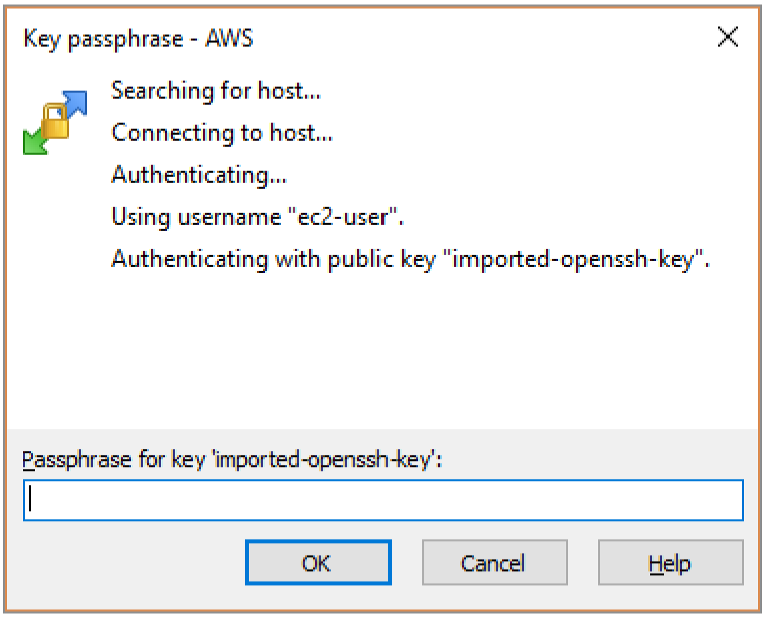
GSSAPI

☑ Attempt GSSAPI authentication

□ Allow GSSAPI credential delegation

3. Click OK, and when you're back to the Login screen, save the connection if desired.

4. Login. If you entered a passphrase when saving the private key, you will be prompted for it now



Searching for host...

Connecting to host...

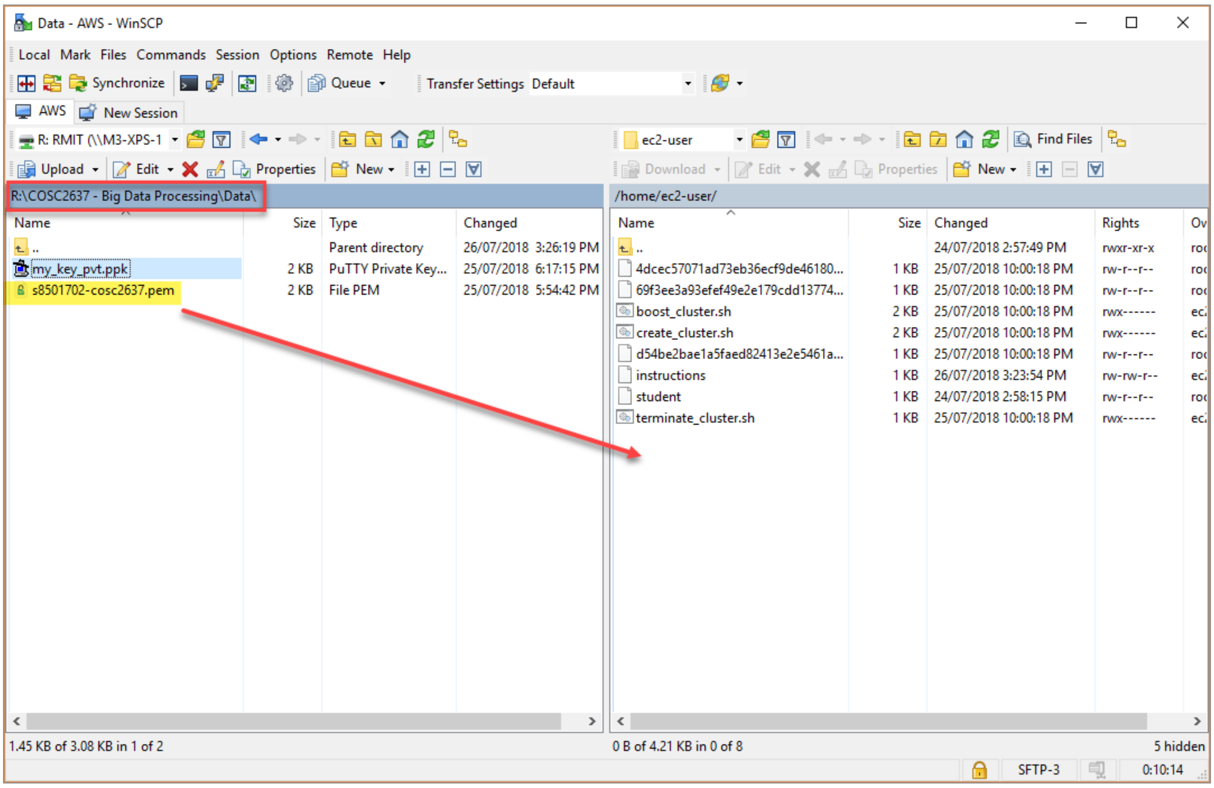
Authenticating...

Using username "ec2-user".

Authenticating with public key "imported-openssh-key".

Passphrase for key 'imported-openssh-key':

5. In the left panel, browse to the location of the key pair file you were emailed (*.pwm), then drag this to the right panel (the jump host), to copy the key pair file.



R:\COSC2637 - Big Data Processing\Data\

| Name | Size | Type | Changed |
|---|---|---|---|
| .. | | Parent directory | 26/07/2018 3:26:19 PM |
| my_key_pvt.ppk | 2 KB | PuTTY Private Key... | 25/07/2018 6:17:15 PM |
| s8501702-cosc2637.pem | 2 KB | File PEM | 25/07/2018 5:54:42 PM |

/home/ec2-user/

| Name | Size | Changed | Rights | Ov |
|---|---|---|---|---|
| .. | | 24/07/2018 2:57:49 PM | rwxr-xr-x | ro |
| 4dcec57071ad73eb36ecf9de46180... | 1 KB | 25/07/2018 10:00:18 PM | rw-r--r-- | ro |
| 69f3ee3a93efef49e2e179cdd13774... | 1 KB | 25/07/2018 10:00:18 PM | rw-r--r-- | ro |
| boost_cluster.sh | 2 KB | 25/07/2018 10:00:18 PM | rwx------ | ec |
| create_cluster.sh | 2 KB | 25/07/2018 10:00:18 PM | rwx------ | ec |
| d54be2bae1a5faed82413e2e5461a... | 1 KB | 25/07/2018 10:00:18 PM | rw-r--r-- | ro |
| instructions | 1 KB | 26/07/2018 3:23:54 PM | rw-rw-r-- | ec |
| student | 1 KB | 24/07/2018 2:58:15 PM | rw-r--r-- | ro |
| terminate_cluster.sh | 1 KB | 25/07/2018 10:00:18 PM | rwx------ | ec |

1.45 KB of 3.08 KB in 1 of 2

0 B of 4.21 KB in 0 of 8

5 hidden

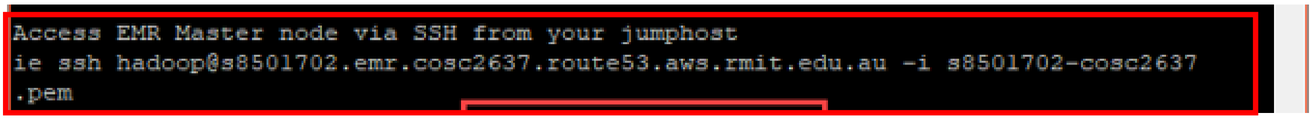
SFTP-3

0:10:14

Then, go to JumpHost home (as shown in the last picture in page 5 of this document) and you should find xxx-xxx.pem. Next

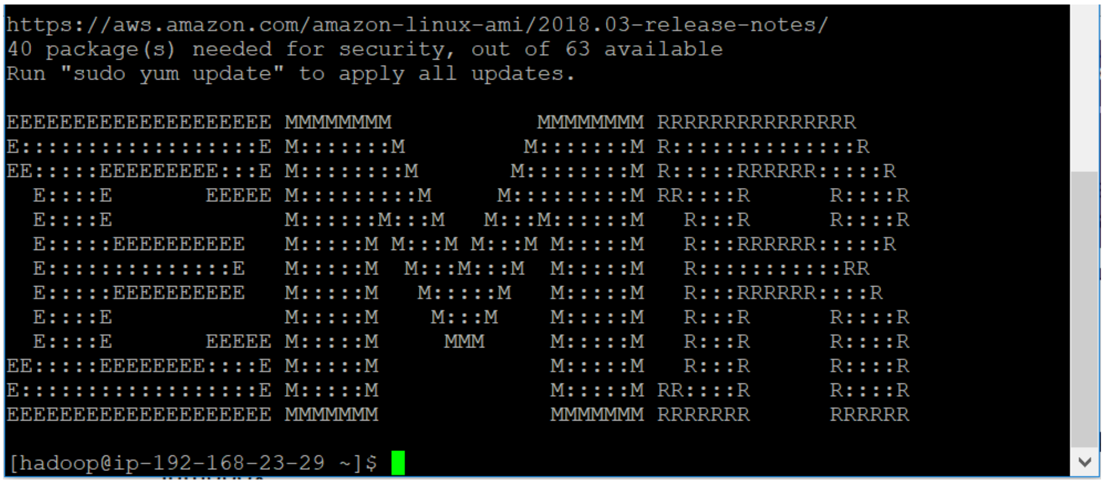
**$ chmod 400 xxx-xxx.pem**

And login the newly created cluster by ssh hadoop@sxxxx.emr.... as shown below



```
Access EMR Master node via SSH from your jumphost
ie ssh hadoop@s8501702.emr.cosc2637.route53.aws.rmit.edu.au -i s8501702-cosc2637
.pem
```

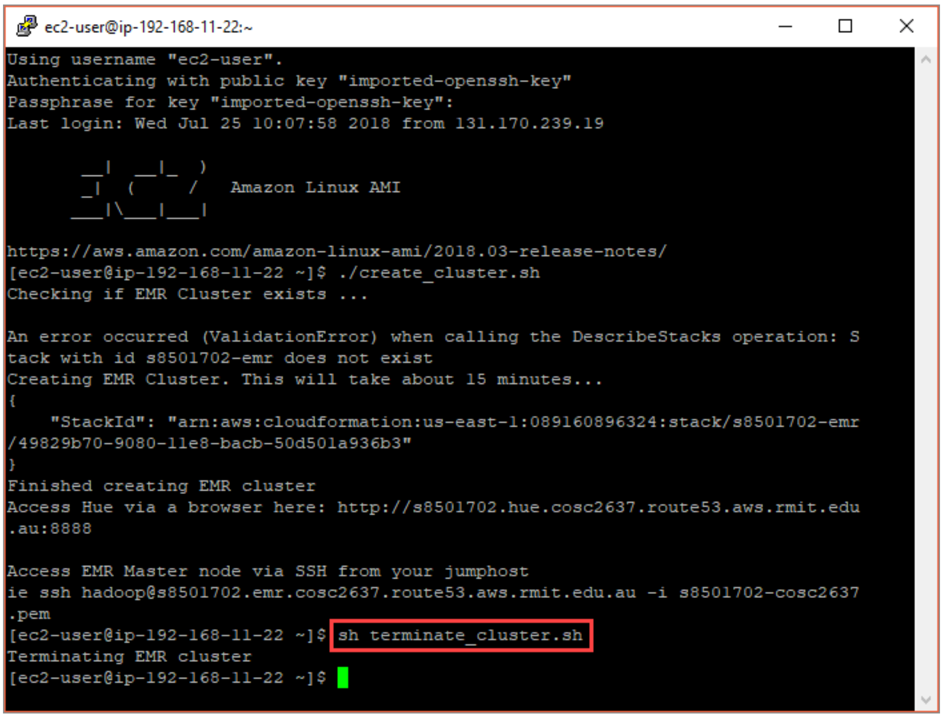
If logging in successfully, you will see



```
https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
40 package(s) needed for security, out of 63 available
Run "sudo yum update" to apply all updates.

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-192-168-23-29 ~]$
```

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.



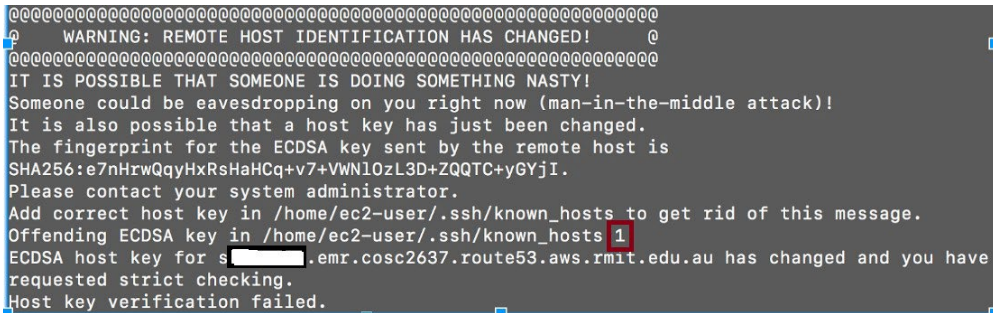
```
Using username "ec2-user".
Authenticating with public key "imported-openssh-key"
Passphrase for key "imported-openssh-key":
Last login: Wed Jul 25 10:07:58 2018 from 131.170.239.19

       __|  __|_  )
       _|  (     /   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.

### A common issue of cluster login

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

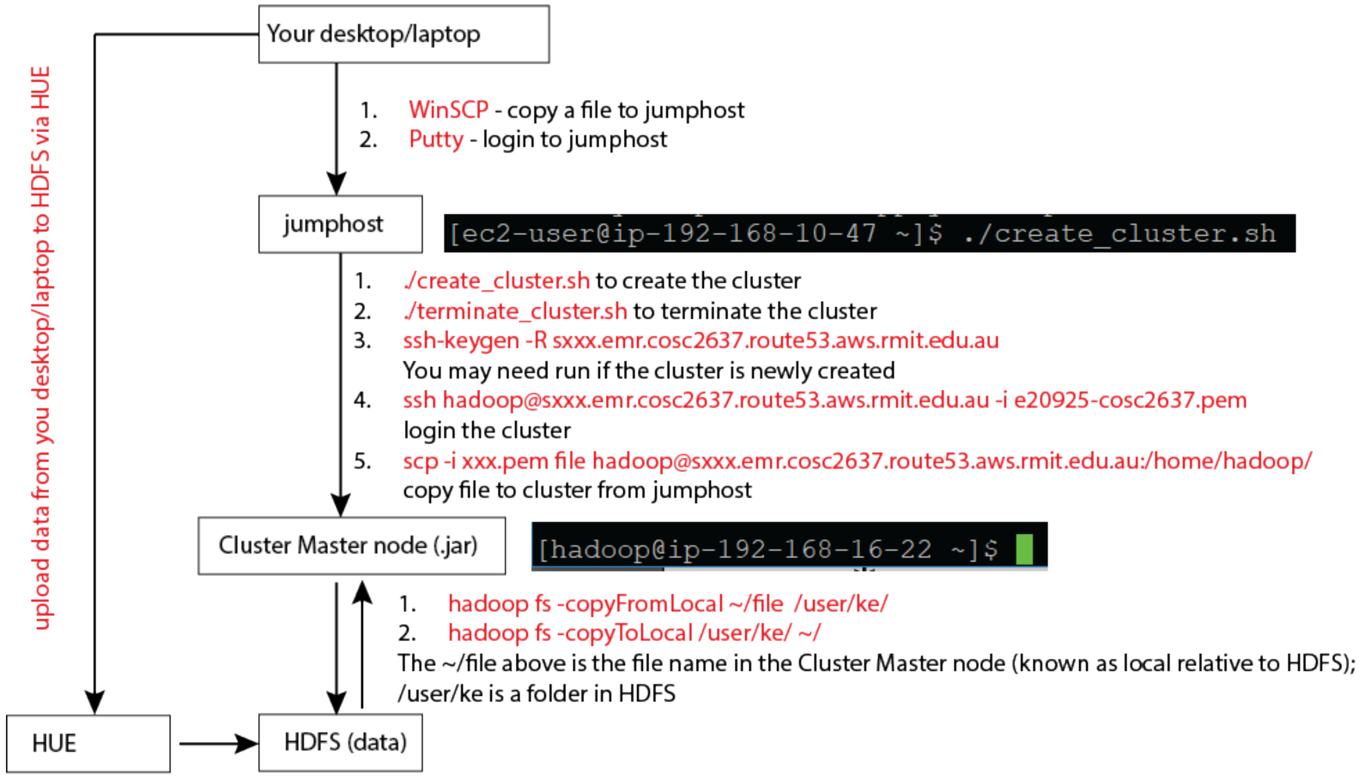


```
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@    WARNING: REMOTE HOST IDENTIFICATION HAS CHANGED!     @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
IT IS POSSIBLE THAT SOMEONE IS DOING SOMETHING NASTY!
Someone could be eavesdropping on you right now (man-in-the-middle 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
```

### AWS EMR - Cheat sheet



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

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

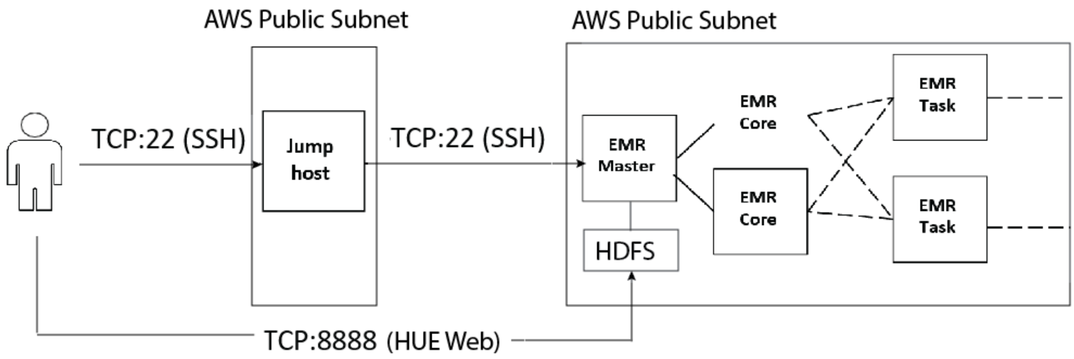
```
[ec2-user@ip-192-168-10-47 ~]$ ./create_cluster.sh
```

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



AWS Public Subnet
AWS Public Subnet
TCP:22 (SSH)
Jump host
TCP:22 (SSH)
EMR Master
HDFS
EMR Core
EMR Core
EMR Task
EMR Task
TCP:8888 (HUE Web)