EMR Workshop Lab 1 - Cluster Creation

(Updated 19-July-18)

This lab demonstrates the steps involved in cluster creation.

1. Create VPC

a) In AWS Management Console
   - Click on VPC

b) In VPC Dashboard
   - Choose Start VPC Wizard

c) In Step 1: Select a VPC Configuration
   - Choose VPC with a Single Public Subnet

d) In Step 2: VPC with a Single Public Subnet
   - Enter a VPC name.
   - Keep the defaults on everything else.
   - Click Create VPC

2. EC2 key pair

Make sure you have an EC2 key pair in the region you are using.

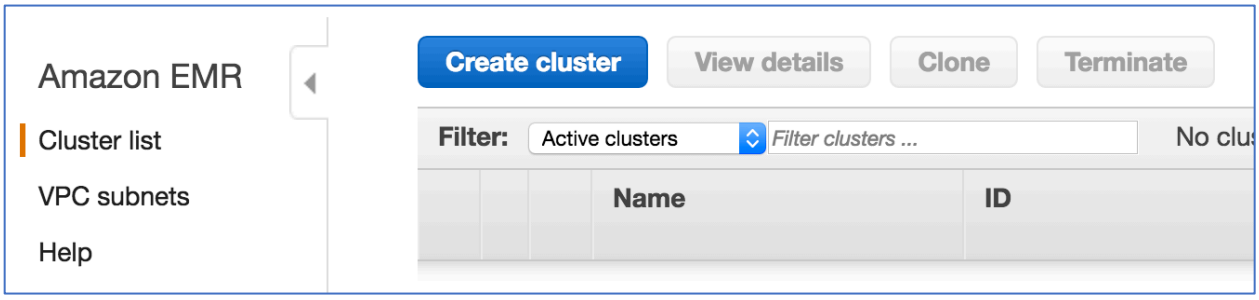
http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-key- pairs.html#having-ec2-create-your-key-pair

3. Launch EMR Cluster

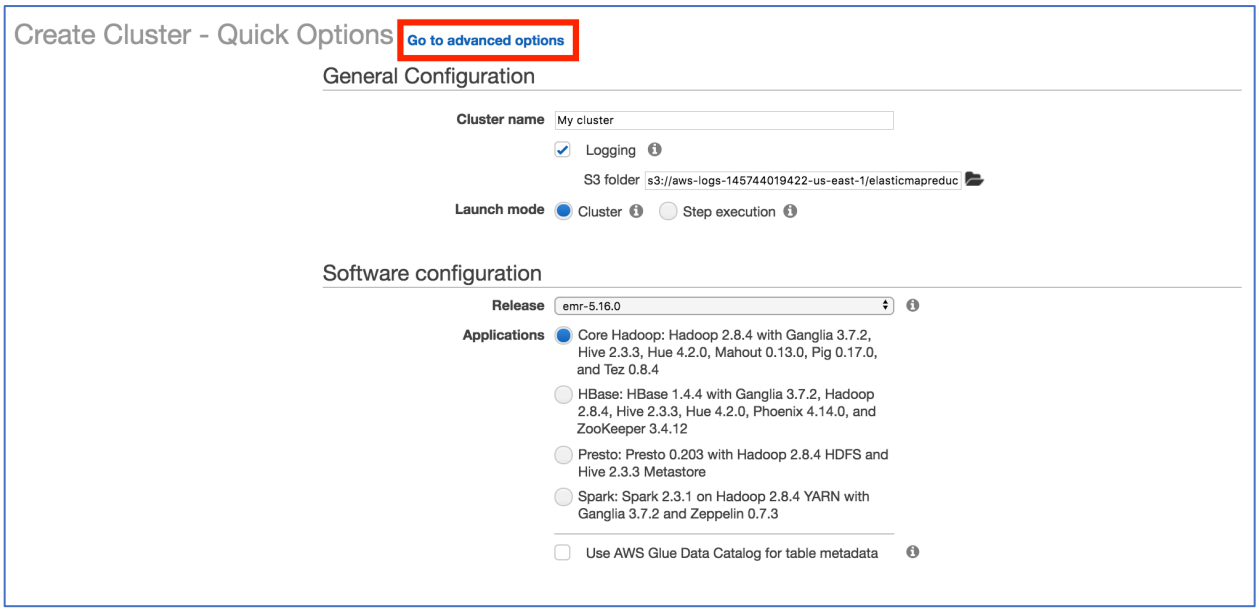
Open the Amazon EMR console at

https://console.aws.amazon.com/elasticmapreduce/

a) Click Create cluster.

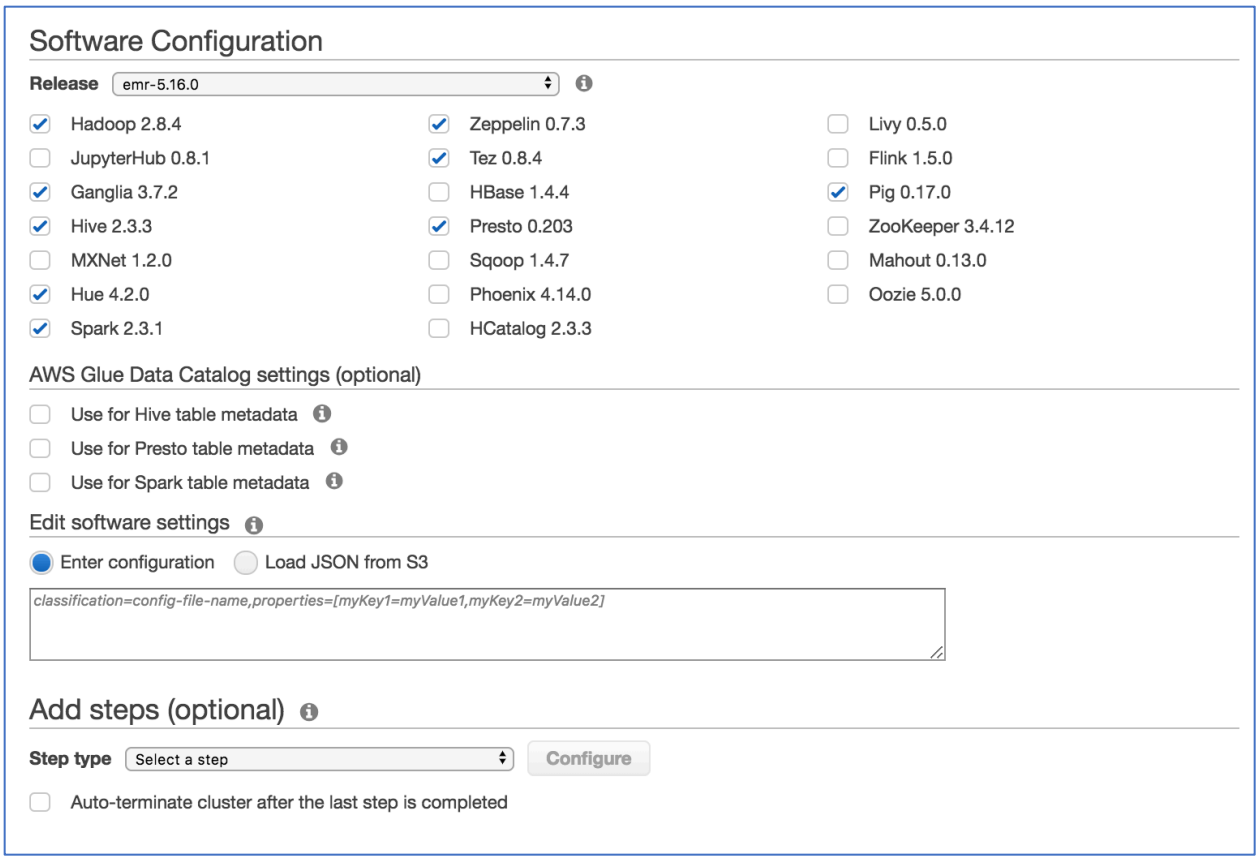


b) Click ‘Go to advanced options’



Step 1: Software and Steps

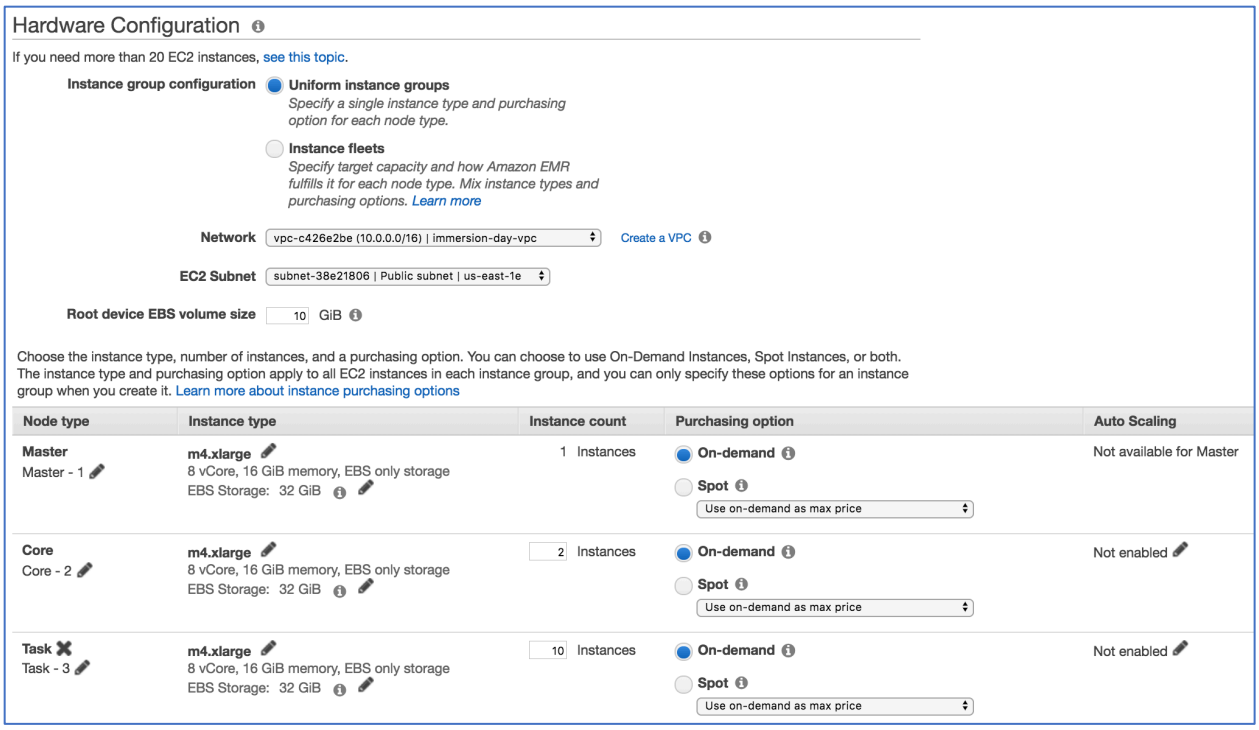
Release	Leave as default
Software Configuration	Ensure that the following are checked: • Hadoop • Ganglia • Hive • Zeppelin • Presto • Tez • Pig • Hue • Spark
AWS Glue Data Catalog settings	Leave as default
Edit Software Settings	Leave as default
Add Steps	Leave as default



c) Click 'Next'

Step 2: Hardware Configuration

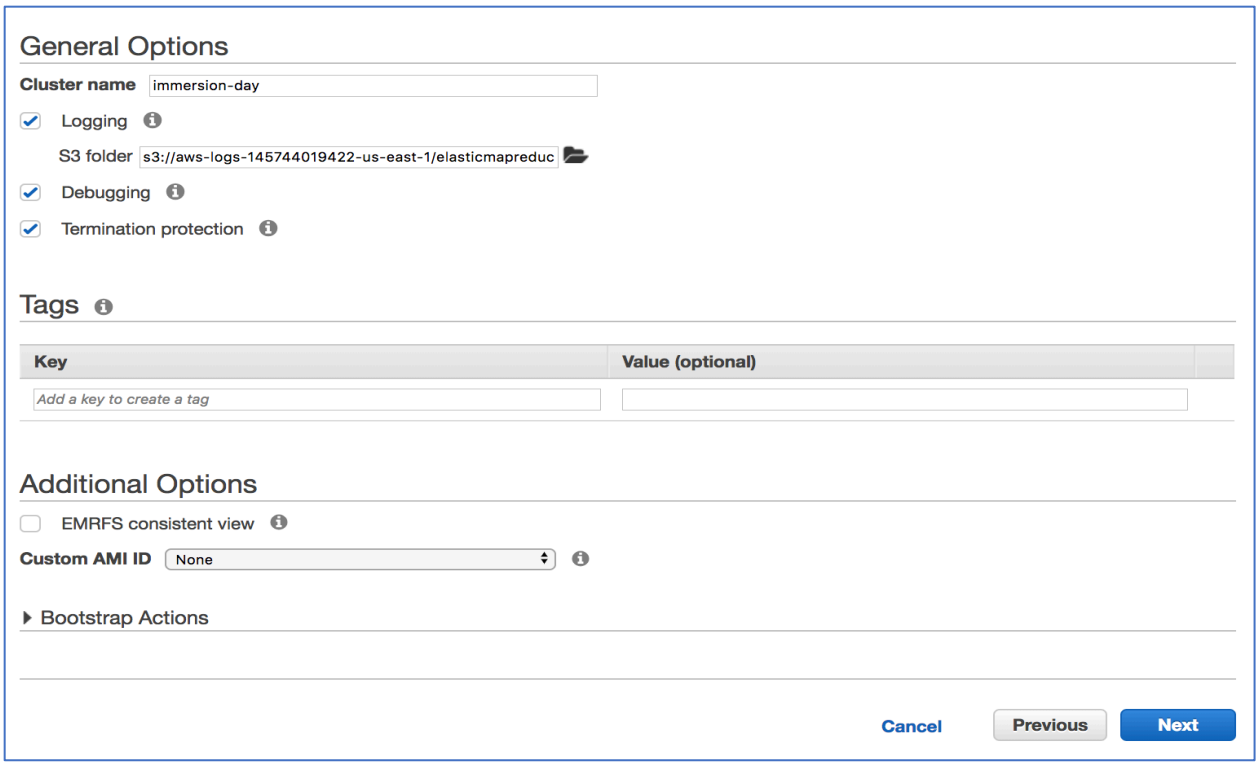
Instance group configuration	Leave as default
Network	Choose previously created VPC
EC2 Subnet	Choose the public subnet
Instances	Set the cluster instances and counts as follows: • Master: m4.xlarge, count = 1 • Core: m4.xlarge, count = 2 • Task: m4.xlarge, count = 10



d) Click ‘Next’

Step 3: General Cluster Settings

Cluster Name	Name your cluster.
Logging	Leave checked Choose a bucket in this region
Debugging	Leave checked.
Termination Protection	Leave checked
Tags	Leave Blank
EMRFS Consistent View	Leave unchecked
Bootstrap actions	Leave alone

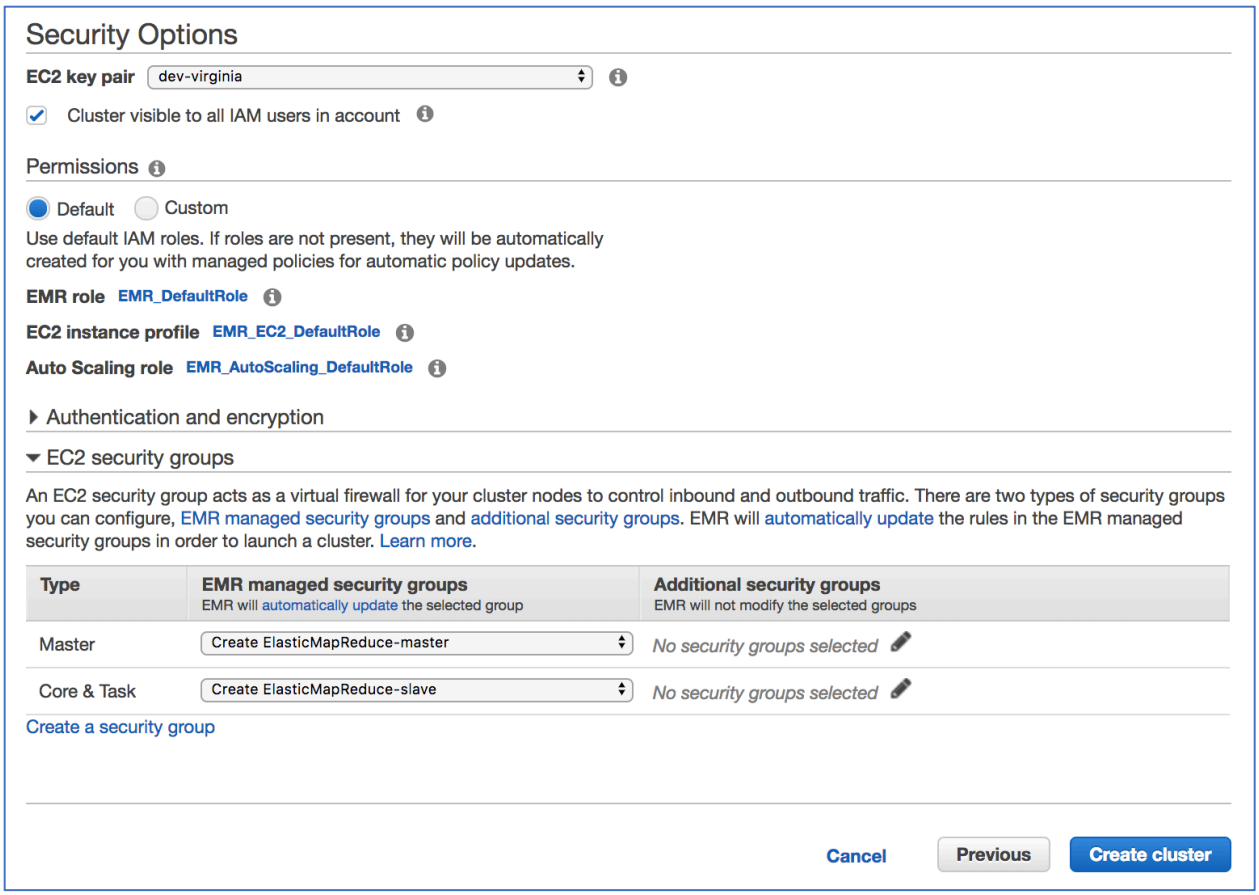


e) Click 'Next'

Step 4: Security

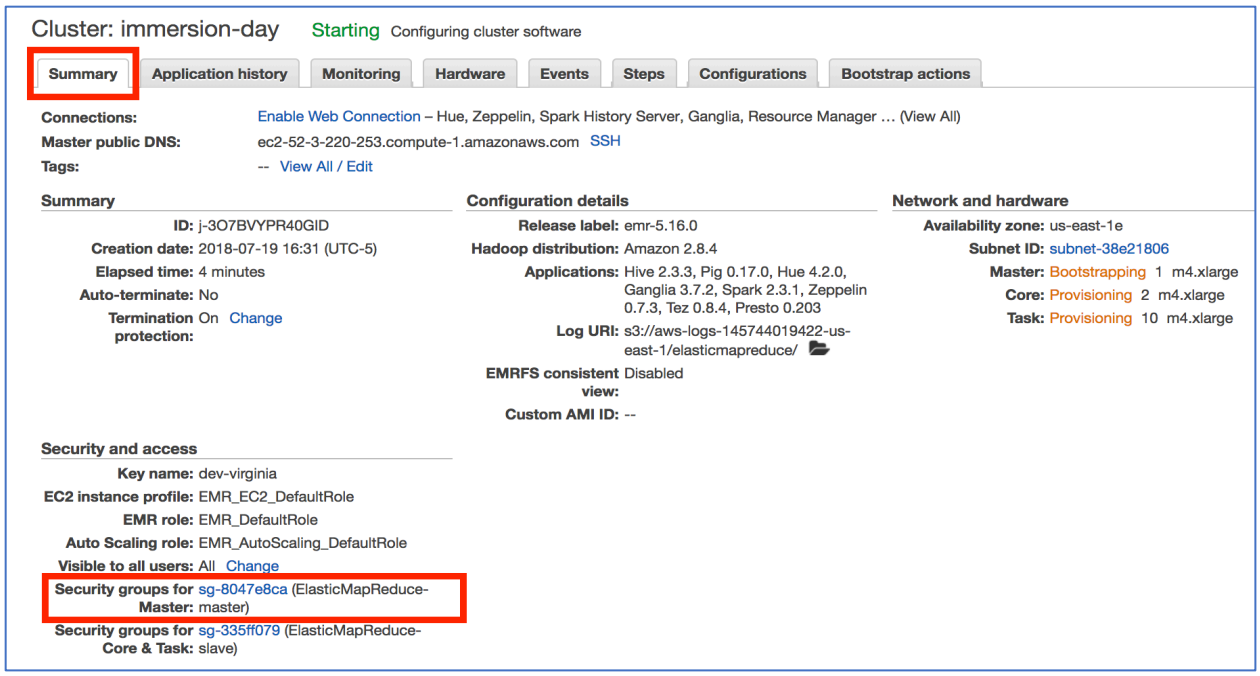
Click 'Create Cluster'.

EC2 Key Pair	Choose a key pair in the region
Cluster visible	Leave checked
Permissions	Choose Default
Authentication and encryption	Leave as default
EC2 Security Groups	Leave as default



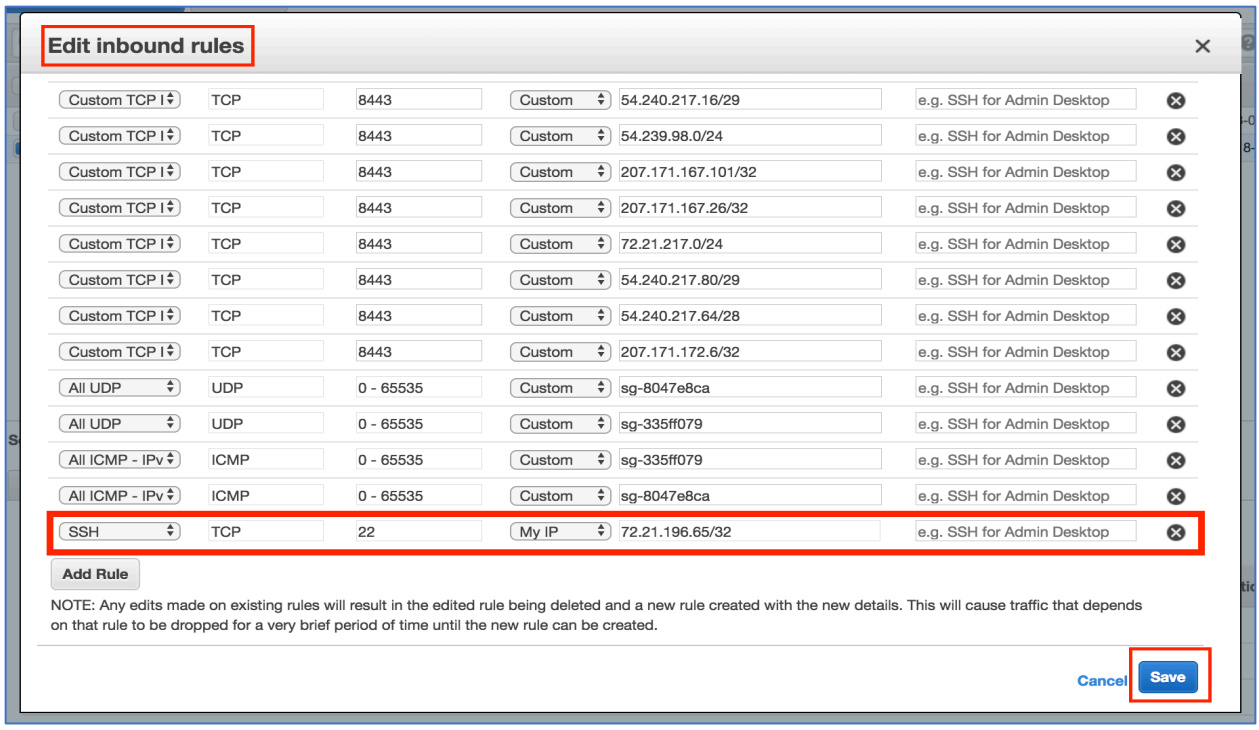
Update Security Group

- In the “Summary” tab for your cluster, scroll down to the “Security and access” section and click on the security group shown for ‘Security Group for Master’



- Click on the security group for ‘ElasticMapReduce-master’

- Click on the ‘Inbound’ tab.
- Click the ‘Edit’ button.
- Click the Add Rule button.
- Add a rule that allows SSH from your IP Address.



- Click Save.

  This will you to SSH into the cluster when it comes up in about 10-15 mins.

➢ ssh -I <<key-pair>> hadoop@<<emr-master-public-dns-address>>

```
9801a7add675:Keys tanzir$ ssh -i dev-virginia.pem hadoop@ec2-52-3-220-253.compute-1.amazonaws.com
Last login: Thu Jul 19 21:46:13 2018 from 72-21-196-65.amazon.com

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
3 package(s) needed for security, out of 4 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEE MMMMMMM           MMMMMMM 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
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR
```