Using EMR

Kookmin University BigData Lab.

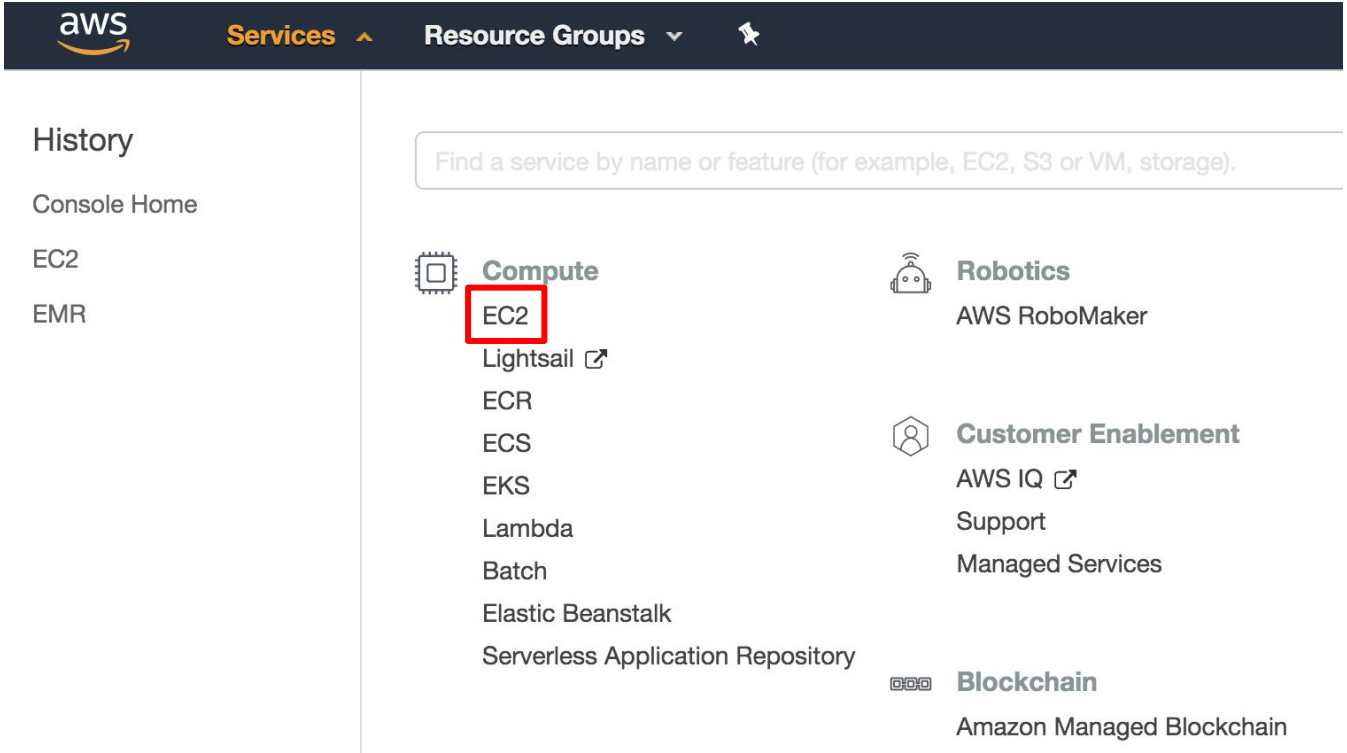
bigDATA Lab.
KOOKMIN UNIVERSITY

Before Use EMR Service

We should create key pair for EMR service(SSH)

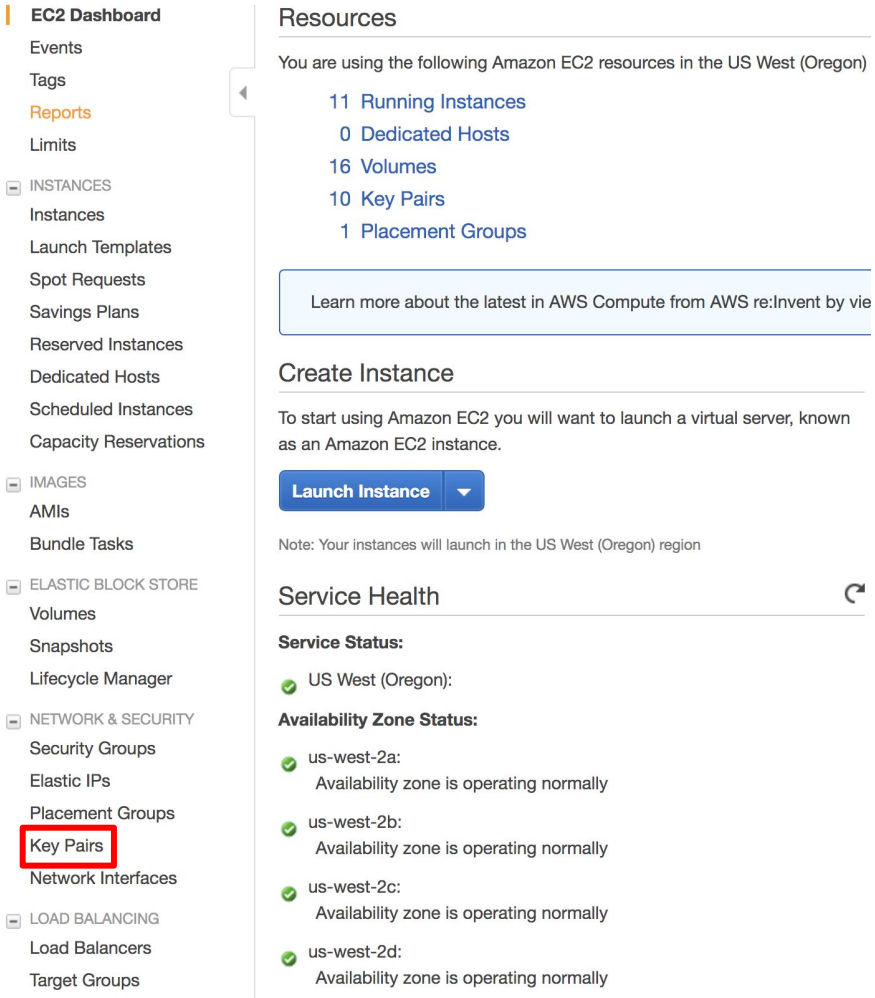
Create Key pair

Services - EC2



Create Key pair

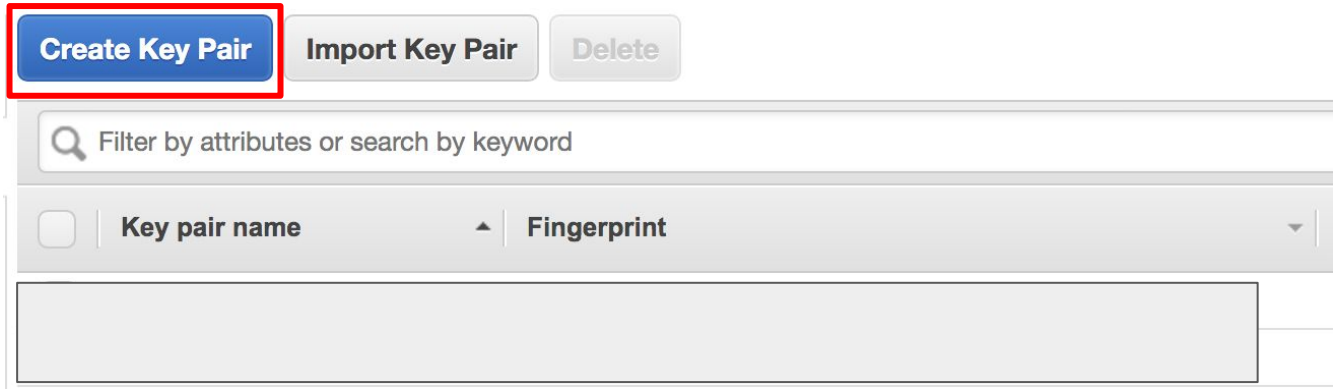
Find ‘Key Pairs’ on the left side of page



bigDATA Lab.
KOOKMIN UNIVERSITY

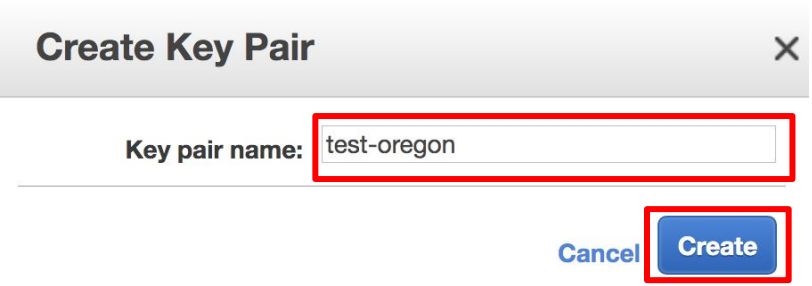
Create Key pair

Click ‘Create Key Pair’



Create Key pair

Write key pair name you want and click ‘Create’, then key pair file will be downloaded on your computer




```
-----BEGIN RSA PRIVATE KEY-----
MIIEowIBAAKCAQEAr741STu57q/4MvwUU/HHWwbayjET3lXRPJa/n0deQEes6I0gUFyxvGlHFsWr
piQ9s0KcgDDzdJB+x3mosSN5upR6r+iZsWtRXXT74UqAaCM6gtZkvWsjSeIv4T3z2S8eI+ZrG2ne
UGy9uQF+enu50AVap9vSm3IjoOpvamiQqpKCwF3VS7fStoXhbbsTPclu1tNBmxA3d/iqig6gOfBb
pbL7O75s4Jcb+6X6l+8TqJ6QQzDvGgI8akoGEsbqGBnnB5qCSPgGlbdLr/YvfjkD7ApsCJQHOZZi
woudCF0Uf2U+UVbVu/5mxoZeTsVrLfjkaLYfmZ1/oQCi8JwJ3kF2bmdLfMWv/lK9yuiBg7PGYUmV
RyOG/gM5OPYT1QtfDFlmgEVrkPCAYD5AT921HJMXNOzSnXN6WtZCzoxX+Q24kBN/MQ72R3nYs8cC
gYEAtzhiU7ay0C1tsUhF2rANZApP/Z7/DO+3jbN5LorFtIa3o8lE0jsbj1bwEkXUOb6tV4UkgxE0
RVzKTEZBMxkbWQ5PiU7vMVJFR1B+u8QF0wVCY+sqB3N733MYqpJf1KCQ5yu0eWlfmaYmloTsaLSj
4hdbleZ3lBxCpXOYLBoeuj0CgYB08lzhBDq2Ia1/j2ybW3qOPCloUsVoI5FGiax5uerLhDm8M8Vm
YqolAsgCDii6CK6KbVKW93Cr4kN18jXoe3oA5snHuhEWV17hkoAhhw0x+qf/jziglrmBqBUpFmdm
GHSLcMSSQ7lstBBHcb+b6upsAluMl/H2NMBJNvXzuWmaVQKBgD8P7bwhssPZ6ez/B/sjayehbuo/
e9tNVjU04Y4Awbxy/3/M7ubOEOwhb3n8z/4jhkcsD2yUcXwCaWNsZq0rwwPMgiADl5MLVoT2J93x
9Irln1LnNRWtHD7rahI/2CGfUflZb+a7MSLV93kOuMQnHVSd/zqmS0PIz9326AqGnp/S
-----END RSA PRIVATE KEY-----
```


bigDATA Lab.
KOOKMIN UNIVERSITY

Create Key pair

Give 600 permission on key pair file and move key pair file to user's home folder

- **Keep this file securely!**

$ chmod 600 ~/Downloads/test-oregon.pem # Give 600 permission

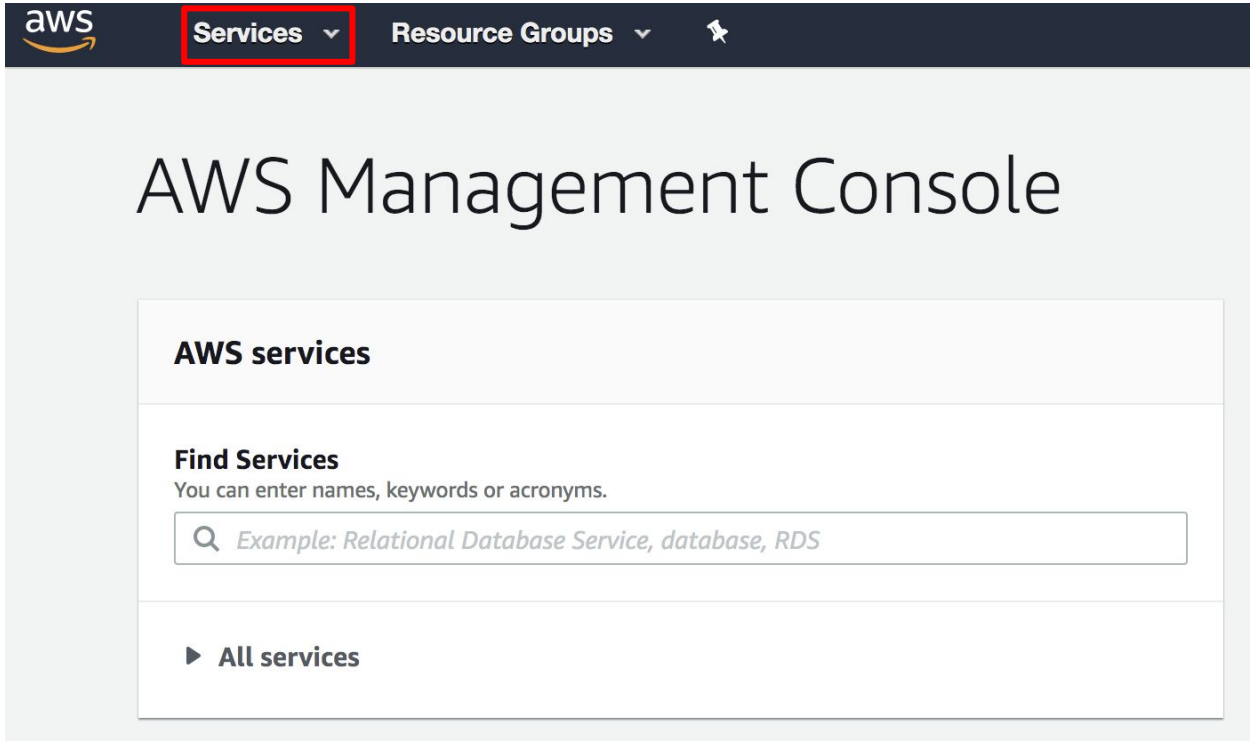
$ mv ~/Downloads/test-oregon.pem ~/ # move .pem file to home folder (downloaded path can be different)

$ ls -l ~/ | grep test-oregon # check .pem file is moved to home folder properly

```
jueon-MacBook-Pro:~ jueon$ chmod 600 ~/Downloads/test-oregon.pem
jueon-MacBook-Pro:~ jueon$ mv ~/Downloads/test-oregon.pem ~/
jueon-MacBook-Pro:~ jueon$ ls -l ~/ | grep test-oregon
-rw-------@  1 jueon  staff  1692 11  7 13:22 test-oregon.pem
jueon-MacBook-Pro:~ jueon$
```

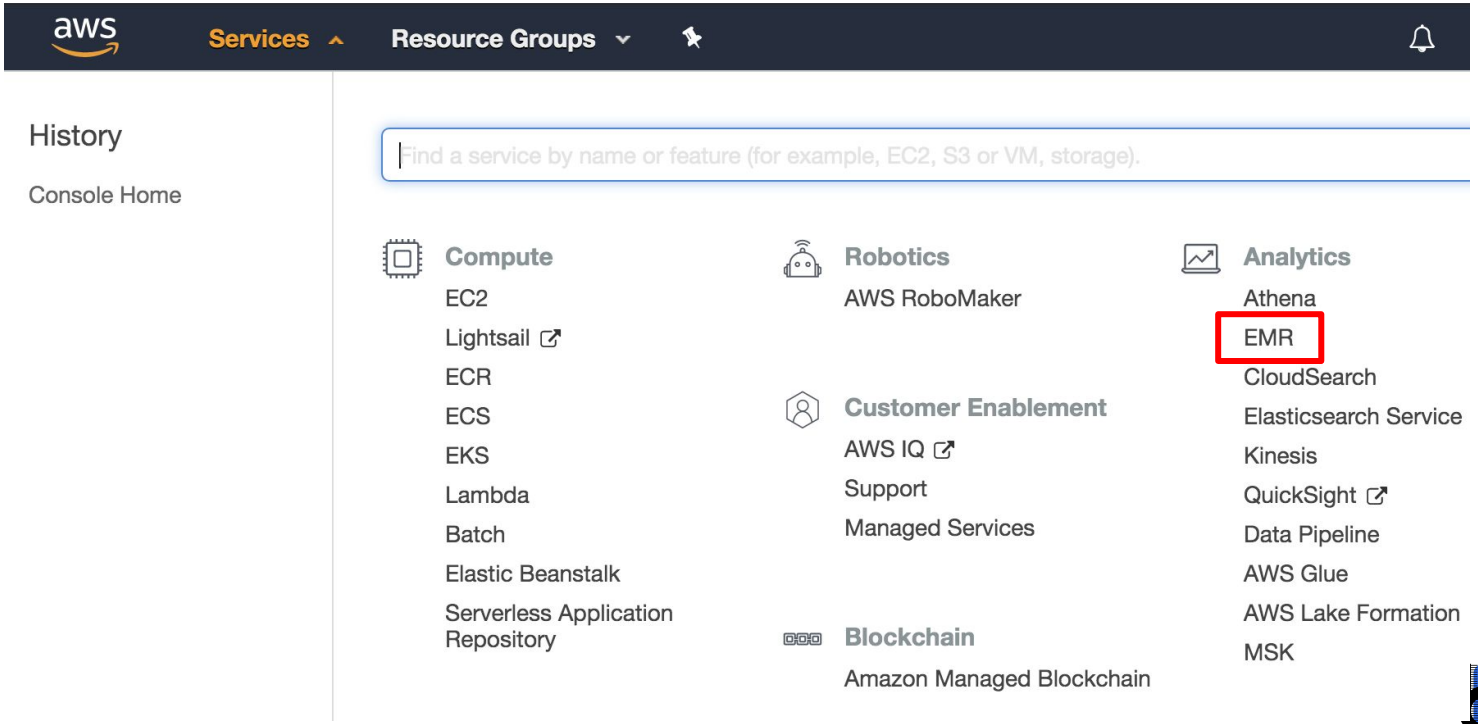
Access to EMR Service

Services - EMR



Access to EMR Service

Type EMR or find EMR on menu



bigDATA Lab.
KOOKMIN UNIVERSITY

Create Cluster

Click ‘Create cluster’ in red box

aws Services Resource Groups Support

Amazon EMR
- Clusters
- Security configurations
- Block public access
- VPC subnets
- Events
- Notebooks
- Git repositories
- Help
- What's new

Save up to 90% on compute
Lower costs and reduce time-to-insight by including Spot Instances in your EMR clusters. Learn more.

Create cluster | View details | Clone | Terminate

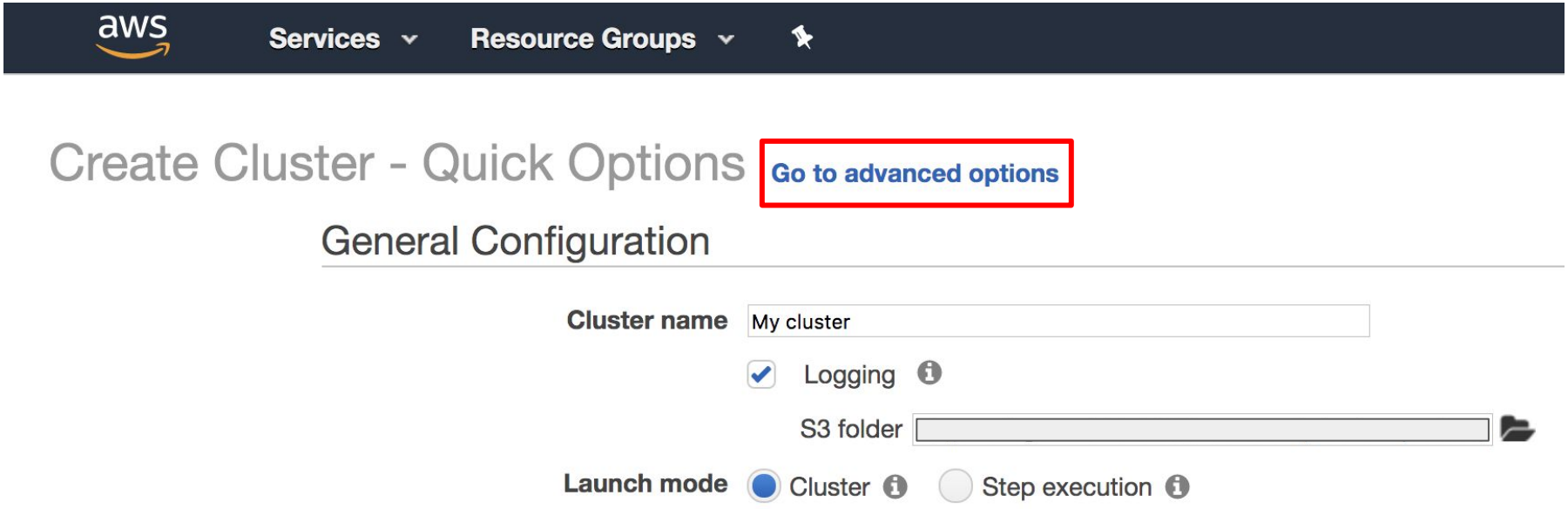
Filter: All clusters | Filter clusters ... | 27 clusters (all loaded)

Name	ID	Status	Creation time (UTC+9)	Elapsed time

bigDATA Lab. KOOKMIN UNIVERSITY

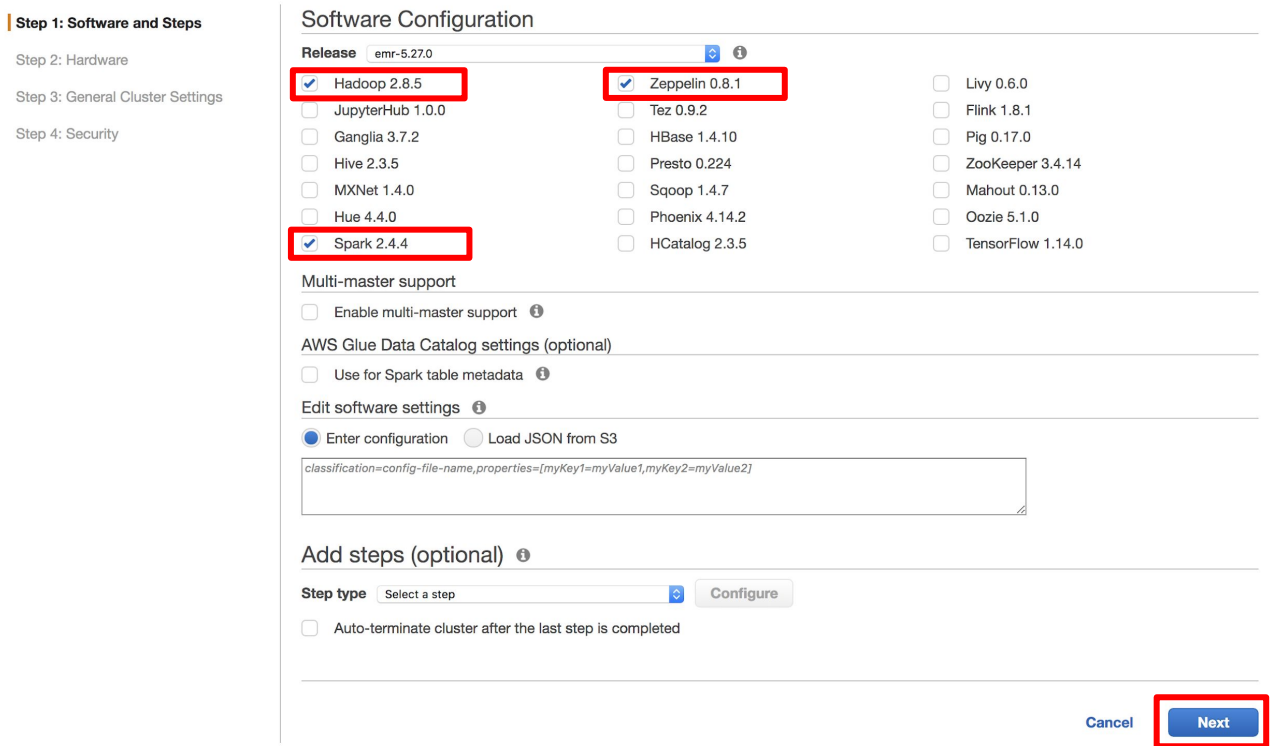
Create Cluster

Click ‘Go to advanced options’ in red box



Software and Steps

Check what you need on cluster(In our case, we need Hadoop, Spark and Zeppelin), and click ‘Next’ in red box



bigDATA Lab.
KOOKMIN UNIVERSITY

Hardware Configuration

Click red box to choose instance type you want (In our case, we will use **m1.large** for instances)

Cluster Nodes and Instances

Choose the instance type, number of instances, and a purchasing option. Learn more about instance purchasing options

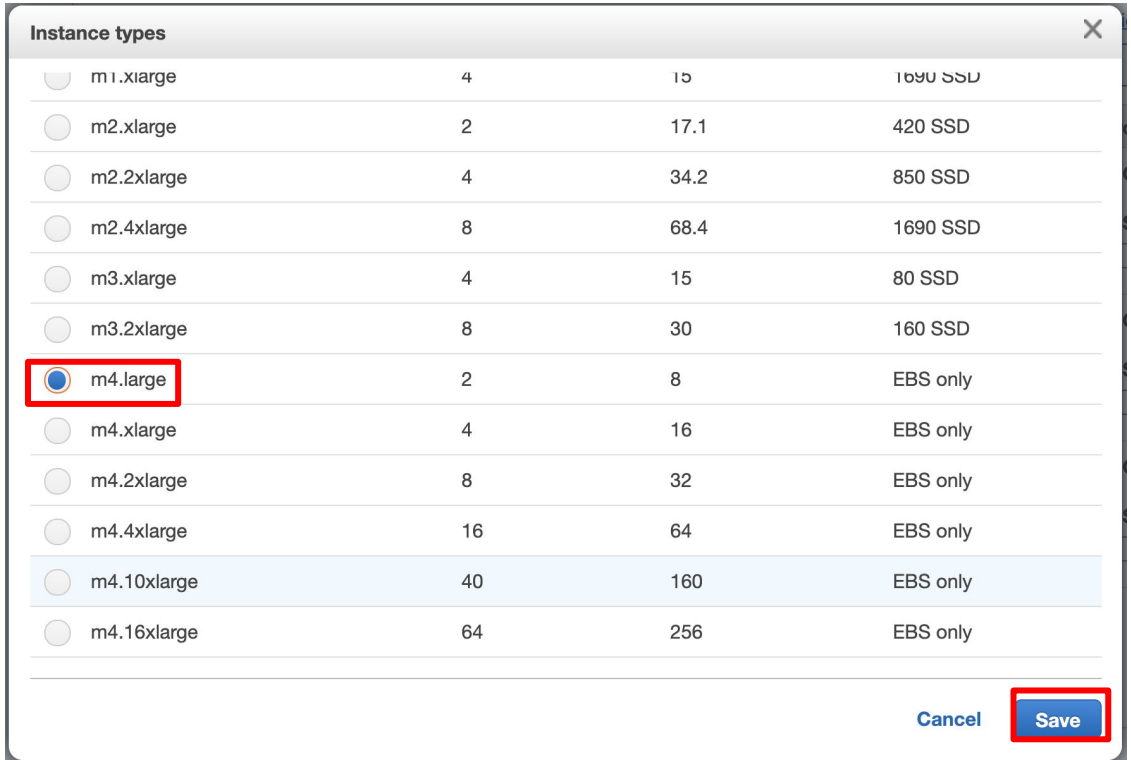
Console options for automatic scaling have changed. Learn more

Node type	Instance type	Instance count	Purchasing option
Master Master - 1	**m5.xlarge** 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	1 Instances	On-demand Spot Use on-demand as max price
Core Core - 2	**m5.xlarge** 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	2 Instances	On-demand Spot Use on-demand as max price
Task Task - 3	**m5.xlarge** 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	0 Instances	On-demand Spot Use on-demand as max price

ATA Lab.
KOOKMIN UNIVERSITY

Hardware Configuration

Find **m1.large** in Instance types



Instance types

m1.xlarge	4	15	1690 SSD
m2.xlarge	2	17.1	420 SSD
m2.2xlarge	4	34.2	850 SSD
m2.4xlarge	8	68.4	1690 SSD
m3.xlarge	4	15	80 SSD
m3.2xlarge	8	30	160 SSD
m4.large	2	8	EBS only
m4.xlarge	4	16	EBS only
m4.2xlarge	8	32	EBS only
m4.4xlarge	16	64	EBS only
m4.10xlarge	40	160	EBS only
m4.16xlarge	64	256	EBS only

Cancel Save

Hardware Configuration

Click red box to choose instance type you want (In our case, we will use **m1.large** for instances)

Cluster Nodes and Instances

Choose the instance type, number of instances, and a purchasing option. Learn more about instance purchasing options

Console options for automatic scaling have changed. Learn more

Node type	Instance type	Instance count	Purchasing option
Master Master - 1	**m4.large** 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	On-demand Spot Use on-demand as max price
Core Core - 2	**m4.large** 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	2 Instances	On-demand Spot Use on-demand as max price
Task Task - 3	**m5.xlarge** 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	0 Instances	On-demand Spot Use on-demand as max price

Hardware Configuration

Set ‘Root device EBS volume size’ to **32** GiB

EBS Root Volume

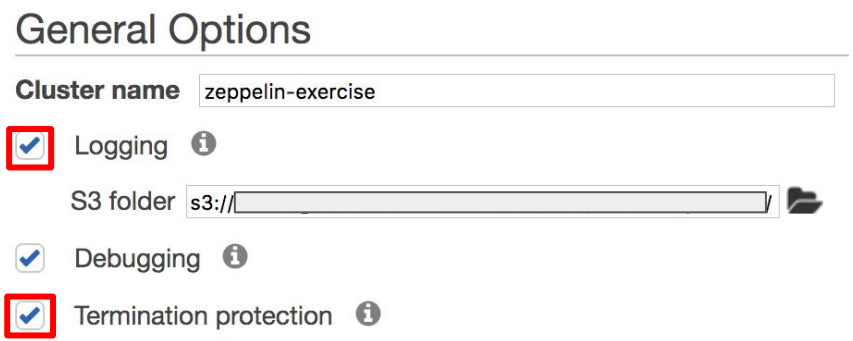
Specify the root device volume size up to 100 GiB. This sizing applies to all instances in the cluster. Learn more.

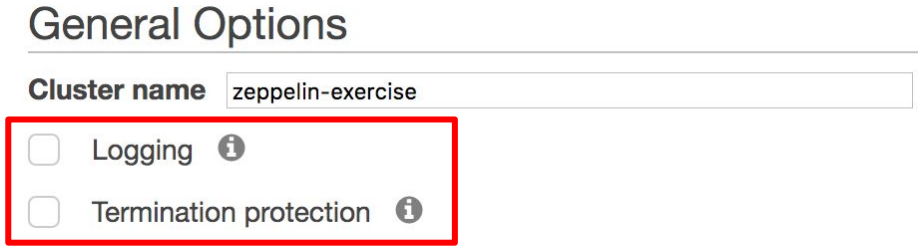
Root device EBS volume size 32 GiB

Cancel Previous Next

General Cluster Settings

Uncheck ‘Logging’ and ‘Termination protection’





Security Options

Set EC2 key pair

Security Options

EC2 key pair test-oregon

Cluster visible to all IAM users in account

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

EC2 instance profile EMR_EC2_DefaultRole

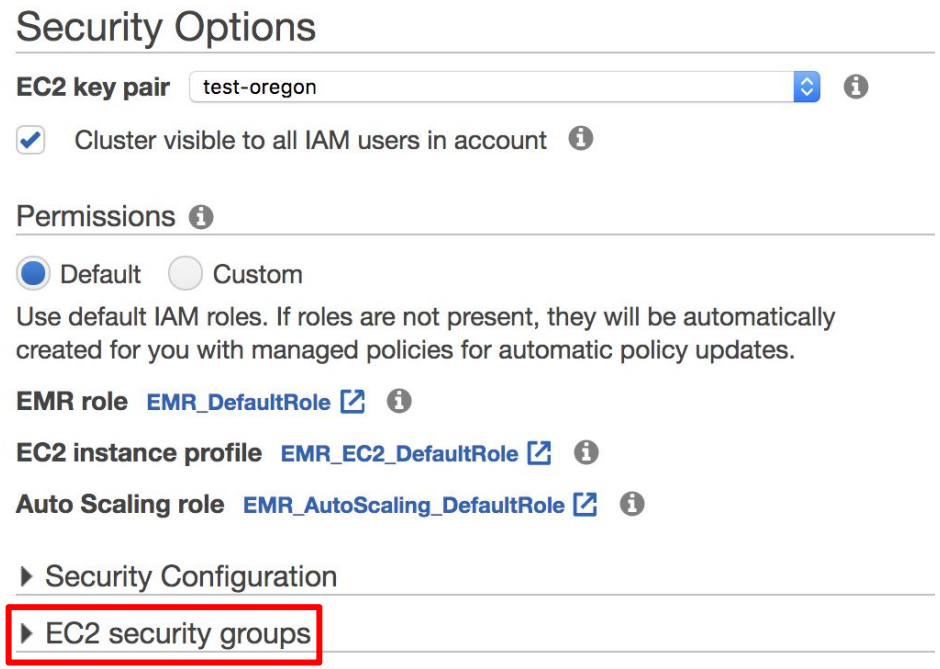
Auto Scaling role EMR_AutoScaling_DefaultRole

Security Configuration

EC2 security groups

Security Options

Click EC2 security groups to set the security group



Security Options

Setting the security groups (If you don't have, just create one like below)

▸ Security Configuration

▾ EC2 security groups

An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are two types of security groups you can configure, EMR managed security groups and additional security groups. EMR will automatically update the rules in the EMR managed security groups in order to launch a cluster. Learn more.

Type	EMR managed security groups EMR will automatically update the selected group	Additional security groups EMR will not modify the selected groups
Master	Create ElasticMapReduce-master	*No security groups selected*
Core & Task	Create ElasticMapReduce-slave	*No security groups selected*

Create a security group

Security Options

Click ‘Create cluster’ to finish

▸ Security Configuration

▾ EC2 security groups

An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are two types of security groups you can configure, EMR managed security groups and additional security groups. EMR will automatically update the rules in the EMR managed security groups in order to launch a cluster. Learn more.

Type	EMR managed security groups EMR will automatically update the selected group	Additional security groups EMR will not modify the selected groups
Master	sg-02b667fdbe6457975 (default)	*No security groups selected*
Core & Task	sg-02b667fdbe6457975 (default)	*No security groups selected*

EMR will automatically update the rules in the custom EMR managed security groups selected above to launch a cluster

Create a security group

Cancel | Previous | Create cluster

Connecting to Master Node Using SSH

Click SSH on Summary tab and copy the ssh command below

Cluster: zeppelin-exercise Waiting Cluster ready to run steps.

Summary | Application history | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Connections: Zeppelin, Spark History Server, Resource Manager ... (View All)

Master public DNS: ec2-34-221-239-37.us-west-2.compute.amazonaws.com SSH

History service: Spark history server UI (SSH tunneling not required)

Tags: -- View All / Edit

Connect to the Master Node Using SSH

You can connect to the Amazon EMR master node using SSH to run interactive queries, examine log files, submit Linux commands, and so on. Learn more.

Windows | Mac / Linux

1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found at Applications > Accessories > Terminal.
2. To establish a connection to the master node, type the following command. Replace ~/test-oregon.pem with the location and filename of the private key file (.pem) used to launch the cluster.

```
ssh -i ~/test-oregon.pem hadoop@ec2-34-221-239-37.us-west-2.compute.amazonaws.com
```

3. Type yes to dismiss the security warning.

bigDATA Lab.
KOOKMIN UNIVERSITY

Access to a master node

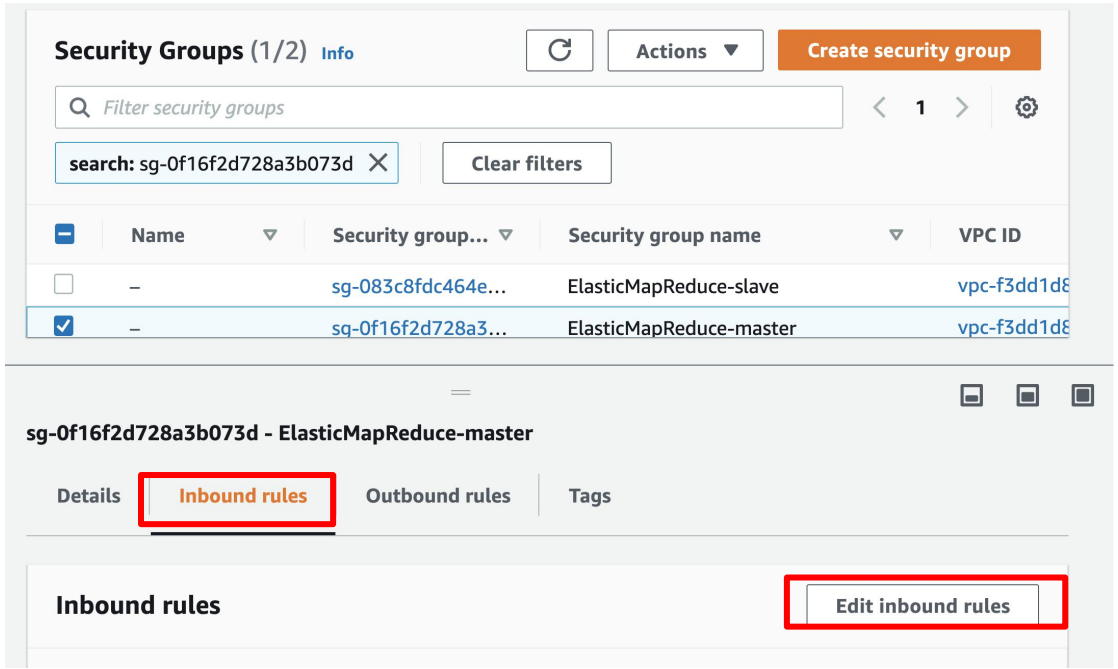
At the end of the EMR cluster page, click a master node's security group

Security and access

Key name:	test-oregon
EC2 instance profile:	EMR_EC2_DefaultRole
EMR role:	EMR_DefaultRole
Auto Scaling role:	EMR_AutoScaling_DefaultRole
Visible to all users:	All Change
Security groups for Master:	sg-0f16f2d728a3b073d (ElasticMapReduce-master)
Security groups for Core & Task:	sg-083c8fdc464e49dc6 (ElasticMapReduce-slave)

Open SSH port

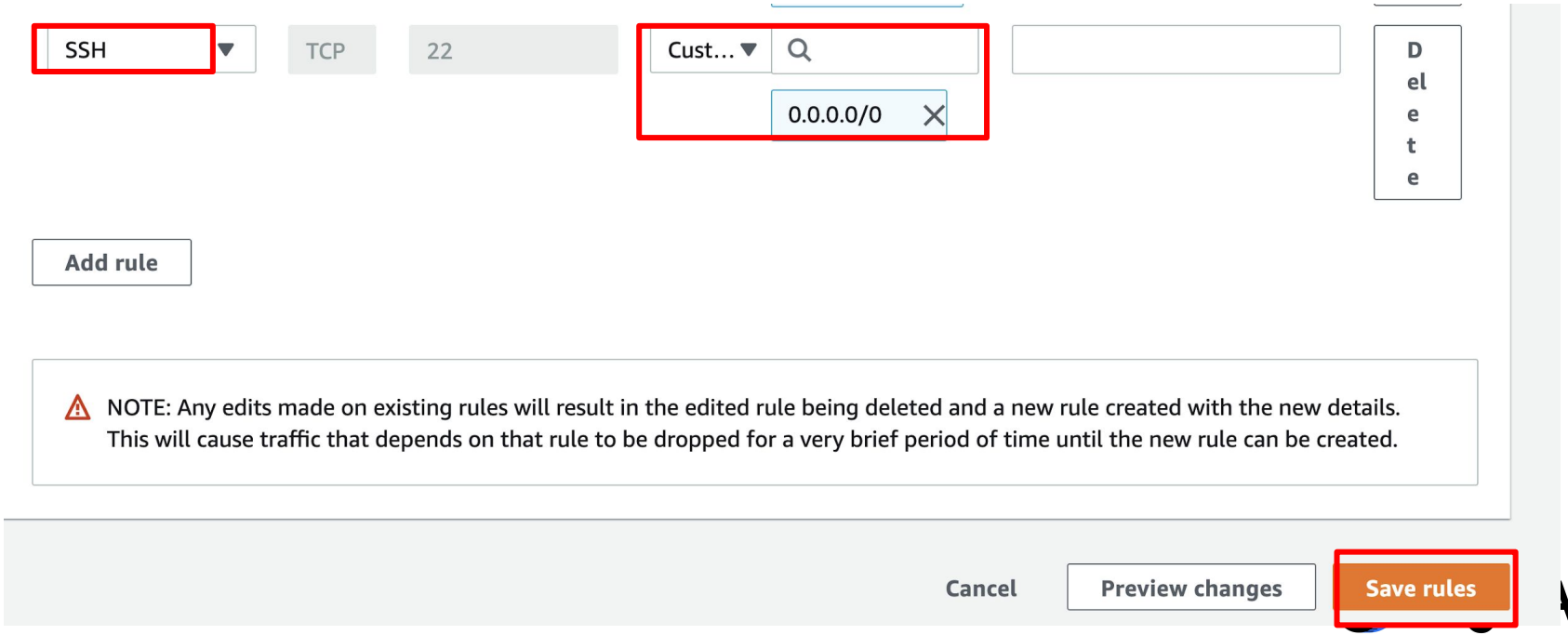
Click Inbound rules → Edit Inbound rules



Open SSH port

Click Add rule → Choose SSH in the protocol and type 0.0.0.0/0 in the destination

Click Save rules



Connecting to Master Node Using SSH

Open terminal and paste

```
jueon-MacBook-Pro:~ jueon$ ssh -i ~/test-oregon.pem hadoop@ec2-34-221-239-37.us-west-2.compute.amazonaws.com
The authenticity of host 'ec2-34-221-239-37.us-west-2.compute.amazonaws.com (34.221.239.37)' can't be established.
ECDSA key fingerprint is SHA256:EX1Xg6rIpB+bnaEp9yFUbeNcM9LG+Um1jiaVKsMG+no.
Are you sure you want to continue connecting (yes/no)? yes
```

```
Warning: Permanently added 'ec2-34-221-239-37.us-west-2.compute.amazonaws.com,34.221.239.37'
f known hosts.
Last login: Thu Nov  7 06:17:20 2019

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
19 package(s) needed for security, out of 29 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::::::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-172-31-8-185 ~]$
```

bigDATA Lab.
KOOKMIN UNIVERSITY

Shutdown a cluster after the exercise is over

In the EMR page, click a cluster and Terminate