

Using EMR



Kookmin University BigData Lab.

Before Use EMR Service

We should create key pair for EMR service(SSH)

Create Key pair

Services - EC2

 **Services** ^ **Resource Groups** v 



History


Console Home



EC2


EMR

Find a service by name or feature (for example, EC2, S3 or VM, storage).

 **Compute**
EC2
Lightsail 
ECR
ECS
EKS
Lambda
Batch
Elastic Beanstalk
Serverless Application Repository

 **Robotics**
AWS RoboMaker

 **Customer Enablement**
AWS IQ 
Support
Managed Services

 **Blockchain**
Amazon Managed Blockchain

Create Key pair

Find 'Key Pairs' on the left side of page

EC2 Dashboard

Events

Tags

Reports

Limits

INSTANCES

Instances

Launch Templates

Spot Requests

Savings Plans

Reserved Instances

Dedicated Hosts

Scheduled Instances

Capacity Reservations

IMAGES

AMIs

Bundle Tasks

ELASTIC BLOCK STORE

Volumes

Snapshots

Lifecycle Manager

NETWORK & SECURITY

Security Groups

Elastic IPs

Placement Groups

Key Pairs

Network Interfaces

LOAD BALANCING

Load Balancers

Target Groups

Resources

You are using the following Amazon EC2 resources in the US West (Oregon)

11 Running Instances

0 Dedicated Hosts

16 Volumes

10 Key Pairs

1 Placement Groups

Learn more about the latest in AWS Compute from AWS re:Invent by vie

Create Instance

To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.

Launch Instance

Note: Your instances will launch in the US West (Oregon) region

Service Health

Service Status:

US West (Oregon):

Availability Zone Status:

us-west-2a:
Availability zone is operating normally

us-west-2b:
Availability zone is operating normally

us-west-2c:
Availability zone is operating normally

us-west-2d:
Availability zone is operating normally

Create Key pair

Click 'Create Key Pair'

Create Key Pair

Import Key Pair

Delete

Filter by attributes or search by keyword

Key pair name

▲

Fingerprint

▼

Create Key pair

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

Create Key Pair

×

Key pair name:

Cancel Create

-----BEGIN RSA PRIVATE KEY-----

MIIeEwIBAAKCAQEAR741STu57q/4MvwUU/HHWwbayjET3lXRPJa/n0deQEes6I0gUFyxvGLHFSwr
piQ9s0KcgDDzdJB+x3mosSN5upR6r+iZsWtRXXT74UqAaCM6gtZCvWsjSeIv4T3z2S8eI+ZrG2ne
UGy9uQF+enu50AVap9vSm3Ijo0pvamiQqpKCwF3V57fStoXhbbstPclu1tNBmxA3d/iqig6g0fBb
pbL7075s4Jcb+6X6l+8TqJ6QQzDvGgI8akoGESbqGBnnB5qCSPqGlbdlr/YvfjkD7ApsCJQH0ZZi

woudCF0Uf2U+UVbVu/5mxoZeTsVrLfjkaLYfmZ1/oQC18JwJ3kF2bmdLfMWv/LK9yuiBg7PGYUmV
Ry0G/gM50PYT1QtFDfLmgEVrkPCAYD5AT921HJMXN0zSnX6WtZCzoX+Q24kBN/MQ72R3nYs8cC
gYEAtzhiU7ay0C1tsUhF2rANZApP/Z7/D0+3jbN5LorFtIa3o8lE0jsbjlbwEkXU0b6tV4UkgxE0
RVzKTEZBMxkbWQ5PiU7vMVJFR1B+u8QF0wVCY+sqB3N733MYqpJf1KCQ5yu0eWlfmaYmloTsaLSj
4hdbLeZ3lBxCpX0YLB0euj0CgYB08lzhBDq2Ia1/j2ybW3q0PCloUsVoI5FGiax5uerLhDm8M8Vm
YqolAsgCDi6CK6KbVKW93Cr4kN18jXoe3oA5snHuhEWV17hkoAhhw0x+qf/jziglrmbqBUfFmdm
GHSLcMSSQ7lstBBHcb+b6upsAluMl/H2NMBJNvXzuWmaVQKBgD8P7bwhssPZ6ez/B/sjayeHbuo/
e9tNVjU04Y4Awbxy/3/M7ub0E0whb3n8z/4jhkcsD2yUcXwCawNsZq0rwwPMgiADl5MLVoT2J93x
9Irln1LnNRwHD7rahI/2CGfUflZb+a7MSLV93k0uMQnHVSd/zqmS0PIz9326AqGnp/S
-----END RSA PRIVATE KEY-----

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

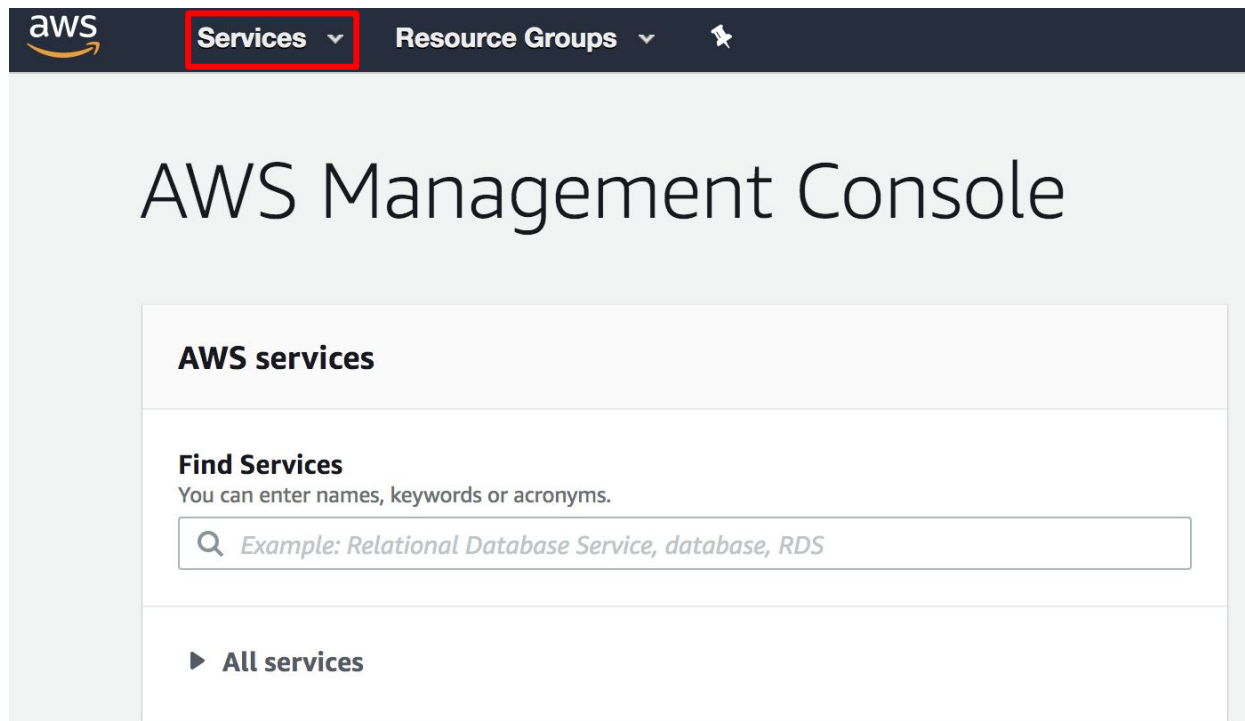
\$ `mv ~/Downloads/test-oregon.pem ~/` # move .pem file to home folder (downloaded path can be different)

\$ `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

The screenshot shows the AWS Management Console interface. At the top, there is a dark navigation bar with the AWS logo, 'Services' with a dropdown arrow, 'Resource Groups' with a dropdown arrow, and a notification bell icon. On the left side, there is a sidebar with 'History' and 'Console Home' links. The main area features a search bar with the placeholder text 'Find a service by name or feature (for example, EC2, S3 or VM, storage)'. Below the search bar, services are categorized into three columns: Compute, Robotics, and Analytics. The 'Compute' column lists EC2, Lightsail, ECR, ECS, EKS, Lambda, Batch, Elastic Beanstalk, Serverless Application Repository, and others. The 'Robotics' column lists AWS RoboMaker. The 'Analytics' column lists Athena, EMR (highlighted with a red box), CloudSearch, Elasticsearch Service, Kinesis, QuickSight, Data Pipeline, AWS Glue, AWS Lake Formation, and MSK. Below the Analytics column, there are sections for 'Customer Enablement' (AWS IQ, Support, Managed Services) and 'Blockchain' (Amazon Managed Blockchain).

aws Services Resource Groups

History
Console Home

Find a service by name or feature (for example, EC2, S3 or VM, storage).

Compute

- EC2
- Lightsail
- ECR
- ECS
- EKS
- Lambda
- Batch
- Elastic Beanstalk
- Serverless Application Repository

Robotics

- AWS RoboMaker

Analytics

- Athena
- EMR**
- CloudSearch
- Elasticsearch Service
- Kinesis
- QuickSight
- Data Pipeline
- AWS Glue
- AWS Lake Formation
- MSK

Customer Enablement

- AWS IQ
- Support
- Managed Services

Blockchain

- Amazon Managed Blockchain

Create Cluster

Click 'Create cluster' in red box

The screenshot shows the Amazon EMR console interface. At the top, there's a navigation bar with the AWS logo, 'Services', 'Resource Groups', a star icon, a search bar, and 'Support'. On the left, a sidebar lists navigation options: 'Amazon EMR', 'Clusters', 'Security configurations', 'Block public access', 'VPC subnets', 'Events', 'Notebooks', 'Git repositories', 'Help', and 'What's new'. The main content area features a blue banner at the top with an information icon, the text 'Save up to 90% on compute', and a description about Spot Instances with a 'Learn more' link. Below the banner, there are four buttons: 'Create cluster' (highlighted with a red box), 'View details', 'Clone', and 'Terminate'. Under these buttons is a filter section showing 'Filter: All clusters' and '27 clusters (all loaded)'. Below the filter is a table with columns: 'Name', 'ID', 'Status', 'Creation time (UTC+9)', and 'Elapsed time'. The table currently shows four rows, each with a checkbox and a right-pointing arrow in the first column, followed by a large grey rectangular area representing the cluster details.

aws Services ▾ Resource Groups ▾ ☆

Amazon EMR

Clusters

Security configurations

Block public access

VPC subnets

Events

Notebooks

Git repositories

Help

What's new

Create cluster View details Clone Terminate

Filter: All clusters ▾ Filter clusters ... 27 clusters (all loaded) ↻

	Name	ID	Status	Creation time (UTC+9) ▾	Elapsed time
<input type="checkbox"/> ▶					
<input type="checkbox"/> ▶					
<input type="checkbox"/> ▶					
<input type="checkbox"/> ▶					

Create Cluster

Click 'Go to advanced options' in red box



Create Cluster - Quick Options

[Go to advanced options](#)

General Configuration

Cluster name

My cluster



Logging



S3 folder



Launch mode



Cluster



Step execution



Software and Steps

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

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release

☒ Hadoop 2.8.5

☐ JupyterHub 1.0.0

☐ Ganglia 3.7.2

☐ Hive 2.3.5

☐ MXNet 1.4.0

☐ Hue 4.4.0

☒ Spark 2.4.4

☒ Zeppelin 0.8.1

☐ Tez 0.9.2

☐ HBase 1.4.10

☐ Presto 0.224

☐ Sqoop 1.4.7

☐ Phoenix 4.14.2

☐ HCatalog 2.3.5

☐ Livy 0.6.0

☐ Flink 1.8.1

☐ Pig 0.17.0

☐ ZooKeeper 3.4.14

☐ Mahout 0.13.0

☐ Oozie 5.1.0

☐ TensorFlow 1.14.0

Multi-master support

☐ Enable multi-master support ⓘ

AWS Glue Data Catalog settings (optional)

☐ Use for Spark table metadata ⓘ

Edit software settings ⓘ

☒ Enter configuration ☐ Load JSON from S3

`classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]`

Add steps (optional) ⓘ

Step type

☐ Auto-terminate cluster after the last step is completed

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	<input checked="" type="radio"/> On-demand <input type="radio"/> 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	<input checked="" type="radio"/> On-demand <input type="radio"/> 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	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price

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	<input checked="" type="radio"/> On-demand <input type="radio"/> 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	<input checked="" type="radio"/> On-demand <input type="radio"/> 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	<input checked="" type="radio"/> On-demand <input type="radio"/> 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

GiB

[Cancel](#)

[Previous](#)

[Next](#)

General Cluster Settings

Uncheck 'Logging' and 'Termination protection'

General Options

Cluster name

☒ Logging 

S3 folder 

☒ Debugging 

☒ Termination protection 

General Options

Cluster name

☐ Logging 

☐ 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

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

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	<div>Create ElasticMapReduce-master</div>	No security groups selected 
Core & Task	<div>Create ElasticMapReduce-slave</div>	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	<input type="text" value="sg-02b667fdb6457975 (default)"/>	No security groups selected
Core & Task	<input type="text" value="sg-02b667fdb6457975 (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.

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

The screenshot displays the AWS Management Console interface for Security Groups. At the top, there's a header 'Security Groups (1/2)' with an 'Info' link, a refresh button, an 'Actions' dropdown, and a 'Create security group' button. Below this is a search bar with the text 'Filter security groups' and a filter input showing 'search: sg-0f16f2d728a3b073d' with a 'Clear filters' button. A table lists security groups with columns for Name, Security group..., Security group name, and VPC ID. The second row is selected, showing 'sg-0f16f2d728a3b073d' with the name 'ElasticMapReduce-master' and VPC ID 'vpc-f3dd1d8'. Below the table, the details for 'sg-0f16f2d728a3b073d - ElasticMapReduce-master' are shown. The 'Inbound rules' tab is selected and highlighted with a red box. At the bottom, the 'Inbound rules' section is visible, with an 'Edit inbound rules' button highlighted by a red box.

Name	Security group...	Security group name	VPC ID
sg-083c8fdc464e...	ElasticMapReduce-slave	vpc-f3dd1d8	
sg-0f16f2d728a3...	ElasticMapReduce-master	vpc-f3dd1d8	

sg-0f16f2d728a3b073d - ElasticMapReduce-master

Details **Inbound rules** Outbound rules Tags

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

SSH ▼

TCP

22

Cust... ▼

Q

0.0.0.0/0 X

D
e
l
e
t
e

Add rule



NOTE: Any edits made on existing rules will result in the edited rule being deleted and a new rule created with the new details. This will cause traffic that depends on that rule to be dropped for a very brief period of time until the new rule can be created.

Cancel

Preview changes

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+bnEp9yFUbeNcM9lG+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'
to 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 MMMMMMM      MMMMMMM RRRRRRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M      M::::::::M R::::::::::::R
EE::::::::EEEEEEEE::::E M::::::::M      M::::::::M R::::::::RRRRR::::R
E::::E      EEEEE M::::::::M      M::::::::M RR::::R      R::::R
E::::E      M::::::::M      M::::::::M R::::R      R::::R
E::::EEEEEEEEEE M::::M M::::M M::::M M::::M R::::RRRRR::::R
E::::::::::::E M::::M M::::M M::::M M::::M R::::::::RRR
E::::EEEEEEEEEE M::::M M::::M M::::M R::::RRRRR::::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 ~]$
```

Shutdown a cluster after the exercise is over

In the EMR page, click a cluster and Terminate

Amazon EMR

Clusters

Notebooks

Git repositories

Security configurations

[Create cluster](#) [View details](#) [Clone](#) [Terminate](#)

Filter: All clusters 2 clusters (all loaded)

			Name	ID	Status
			My_cluster	j-1GLMATIHTFOVK	Waiting Cluster ready