

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.

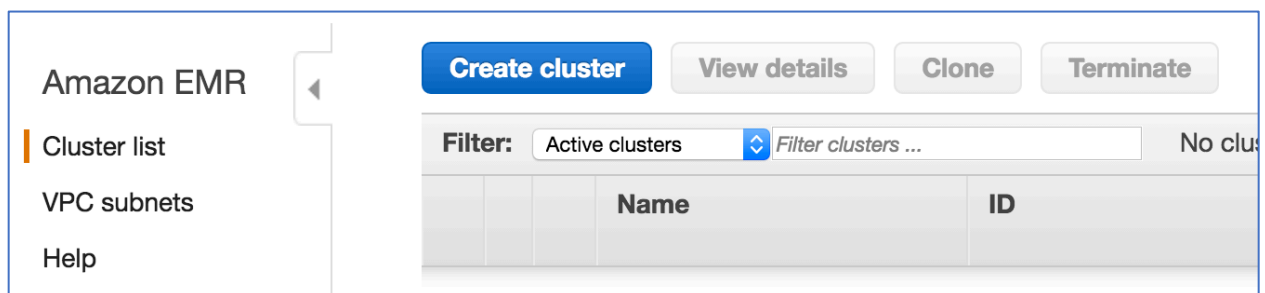
<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-key-pairs.html#having-ec2-create-your-key-pair>

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'

Create Cluster - Quick Options

Go to advanced options

General Configuration

Cluster name

My cluster

☒ Logging

S3 folder

s3://aws-logs-145744019422-us-east-1/elasticmapreduc

Launch mode

☒ Cluster

☐ Step execution

Software configuration

Release

emr-5.16.0

Applications

☒ Core Hadoop: Hadoop 2.8.4 with Ganglia 3.7.2, Hive 2.3.3, Hue 4.2.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.8.4

☐ HBase: HBase 1.4.4 with Ganglia 3.7.2, Hadoop 2.8.4, Hive 2.3.3, Hue 4.2.0, Phoenix 4.14.0, and ZooKeeper 3.4.12

☐ Presto: Presto 0.203 with Hadoop 2.8.4 HDFS and Hive 2.3.3 Metastore

☐ Spark: Spark 2.3.1 on Hadoop 2.8.4 YARN with Ganglia 3.7.2 and Zeppelin 0.7.3

☐ Use AWS Glue Data Catalog for table metadata

Step 1: Software and Steps

Release	Leave as default
Software Configuration	<div>Ensure that the following are checked:</div> <ul style="list-style-type: none">HadoopGangliaHiveZeppelinPrestoTezPigHueSpark
AWS Glue Data Catalog settings	Leave as default
Edit Software Settings	Leave as default
Add Steps	Leave as default

Software Configuration

Release

emr-5.16.0

☒ Hadoop 2.8.4
 ☐ JupyterHub 0.8.1
 ☒ Ganglia 3.7.2
 ☒ Hive 2.3.3
 ☐ MXNet 1.2.0
 ☒ Hue 4.2.0
 ☒ Spark 2.3.1

☒ Zeppelin 0.7.3
 ☒ Tez 0.8.4
 ☐ HBase 1.4.4
 ☒ Presto 0.203
 ☐ Sqoop 1.4.7
 ☐ Phoenix 4.14.0
 ☐ HCatalog 2.3.3

☐ Livy 0.5.0
 ☐ Flink 1.5.0
 ☒ Pig 0.17.0
 ☐ ZooKeeper 3.4.12
 ☐ Mahout 0.13.0
 ☐ Oozie 5.0.0

AWS Glue Data Catalog settings (optional)

☐ Use for Hive table metadata
 ☐ Use for Presto table metadata
 ☐ 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

Select a step

Configure

☐ Auto-terminate cluster after the last step is completed

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: <ul style="list-style-type: none"> Master: m4.xlarge, count = 1 Core: m4.xlarge, count = 2 Task: m4.xlarge, count = 10

Hardware Configuration

If you need more than 20 EC2 instances, [see this topic](#).

Instance group configuration
☒ **Uniform instance groups**
Specify a single instance type and purchasing option for each node type.

☐ **Instance fleets**
Specify target capacity and how Amazon EMR fulfills it for each node type. Mix instance types and purchasing options. [Learn more](#)

Network vpc-c426e2be (10.0.0.0/16) | Immersion-day-vpc Create a VPC

EC2 Subnet subnet-38e21806 | Public subnet | us-east-1e

Root device EBS volume size 10 GiB

Choose the instance type, number of instances, and a purchasing option. You can choose to use On-Demand Instances, Spot Instances, or both. The instance type and purchasing option apply to all EC2 instances in each instance group, and you can only specify these options for an instance group when you create it. [Learn more about instance purchasing options](#)

Node type	Instance type	Instance count	Purchasing option	Auto Scaling
Master Master - 1 	m4.xlarge 8 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB 	1 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price 	Not available for Master
Core Core - 2 	m4.xlarge 8 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB 	<input type="text" value="2"/> Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price 	Not enabled
Task Task - 3 	m4.xlarge 8 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB 	<input type="text" value="10"/> Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price 	Not enabled

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

General Options

Cluster name immersion-day

☒ Logging ⓘ
 S3 folder s3://aws-logs-145744019422-us-east-1/elasticmapreduc

☒ Debugging ⓘ

☒ Termination protection ⓘ

Tags ⓘ

Key	Value (optional)
Add a key to create a tag	

Additional Options

☐ EMRFS consistent view ⓘ

Custom AMI ID None ⓘ

▶ Bootstrap Actions

[Cancel](#)
[Previous](#)
[Next](#)

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

Security Options

EC2 key pair

dev-virginia

☒ 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](#)

Authentication and encryption

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](#)

Cancel

Previous

Create cluster

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’

Cluster: immersion-day **Starting** Configuring cluster software

Summary Application history Monitoring Hardware Events Steps Configurations Bootstrap actions

Connections: [Enable Web Connection](#) – Hue, Zeppelin, Spark History Server, Ganglia, Resource Manager ... (View All)

Master public DNS: ec2-52-3-220-253.compute-1.amazonaws.com [SSH](#)

Tags: -- [View All / Edit](#)

Summary	Configuration details	Network and hardware
ID: j-3O7BVYPR40GID Creation date: 2018-07-19 16:31 (UTC-5) Elapsed time: 4 minutes Auto-terminate: No Termination On protection: Change	Release label: emr-5.16.0 Hadoop distribution: Amazon 2.8.4 Applications: Hive 2.3.3, Pig 0.17.0, Hue 4.2.0, Ganglia 3.7.2, Spark 2.3.1, Zeppelin 0.7.3, Tez 0.8.4, Presto 0.203 Log URI: s3://aws-logs-145744019422-us-east-1/elasticmapreduce/ EMRFS consistent view: Disabled Custom AMI ID: --	Availability zone: us-east-1e Subnet ID: subnet-38e21806 Master: Bootstrapping 1 m4.xlarge Core: Provisioning 2 m4.xlarge Task: Provisioning 10 m4.xlarge

Security and access

Key name: dev-virginia

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 sg-8047e8ca (ElasticMapReduce-Master): [Change](#)

Security groups for sg-335ff079 (ElasticMapReduce-Core & Task): [Change](#)

- Click on the security group for 'ElasticMapReduce-master'

Create Security Group Actions

search : sg-8047e8ca Add filter

Name	Group ID	Group Name	VPC ID	Description
<input type="checkbox"/>	sg-335ff079	ElasticMapReduce-slave	vpc-c426e2be	Slave group for Elastic MapReduce created on 2018-07-19T19:02:36.904Z
<input checked="" type="checkbox"/>	sg-8047e8ca	<u>ElasticMapReduce-master</u>	vpc-c426e2be	Master group for Elastic MapReduce created on 2018-07-19T19:02:36.904Z

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

Edit inbound rules

Custom TCP I	TCP	8443	Custom	54.240.217.16/29	e.g. SSH for Admin Desktop	X
Custom TCP I	TCP	8443	Custom	54.239.98.0/24	e.g. SSH for Admin Desktop	X
Custom TCP I	TCP	8443	Custom	207.171.167.101/32	e.g. SSH for Admin Desktop	X
Custom TCP I	TCP	8443	Custom	207.171.167.26/32	e.g. SSH for Admin Desktop	X
Custom TCP I	TCP	8443	Custom	72.21.217.0/24	e.g. SSH for Admin Desktop	X
Custom TCP I	TCP	8443	Custom	54.240.217.80/29	e.g. SSH for Admin Desktop	X
Custom TCP I	TCP	8443	Custom	54.240.217.64/28	e.g. SSH for Admin Desktop	X
Custom TCP I	TCP	8443	Custom	207.171.172.6/32	e.g. SSH for Admin Desktop	X
All UDP	UDP	0 - 65535	Custom	sg-8047e8ca	e.g. SSH for Admin Desktop	X
All UDP	UDP	0 - 65535	Custom	sg-335ff079	e.g. SSH for Admin Desktop	X
All ICMP - IPv	ICMP	0 - 65535	Custom	sg-335ff079	e.g. SSH for Admin Desktop	X
All ICMP - IPv	ICMP	0 - 65535	Custom	sg-8047e8ca	e.g. SSH for Admin Desktop	X
SSH	TCP	22	My IP	72.21.196.65/32	e.g. SSH for Admin Desktop	X

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.

CancelSave

- 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

  _I_ _I_ )
 _I (  /  Amazon Linux AMI
 ___I\___I___I

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.

EEEEEEEEEEEEEEEEEEEE MMMMMMM      MMMMMMM RRRRRRRRRRRRRR
E:::~::~~::~~::~~::E M:::~::~~::M      M:::~::~~::M R:::~::~~::~~::R
EE:::~::~~::~~::~~::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 R:::~::~~::R      R:::~::~~::R
 E:::~::~~::EEEEEEEEEE M:::~::~~::M M:::~::~~::M M:::~::~~::M R:::~::~~::RRRRRR:::R
 E:::~::~~::~~::~~::E 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

```