Immersion Day

Launching EMR Interactively

September 2016

Rev 2016-01-13

Table of Contents

Overview

EMR allows you to launch clusters that are either long-lived/interactive clusters or transient clusters. Interactive clusters are useful for analysts and engineers to explore datasets as well as other uses that requires a cluster to be running all the time (for example – hadoop/spark based streaming analytics and for realtime access to big data stores that run on top of hdfs such as HBase and Accumulo).

The first set of labs will introduce how to run various analytics in an interactive mode. The first step we'll do for this is to create a cluster which we'll use for many of the following labs.

Part of the cluster configuration is to specify which hadoop applications that EMR should automatically load for you onto the cluster.

We'll be using the following applications. More detail on each tool will be presented in each lab:

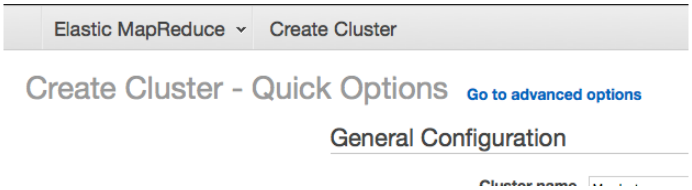
Application	Use
Pig	Script-style analytic that is often used for ETL
Hive	SQL-style analytic to query data on S3 and HDFS
Hue	Web front-end to easily run Pig/Hive jobs
Spark	In-memory analytical framework that provides multiple language support
Zeppelin	Web front-end to run various analytics, including spark

These steps assume you already have a EC2 KeyPair created and the private key accessible. If you don't, please create one that you can use during these labs.

http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-key-pairs.html

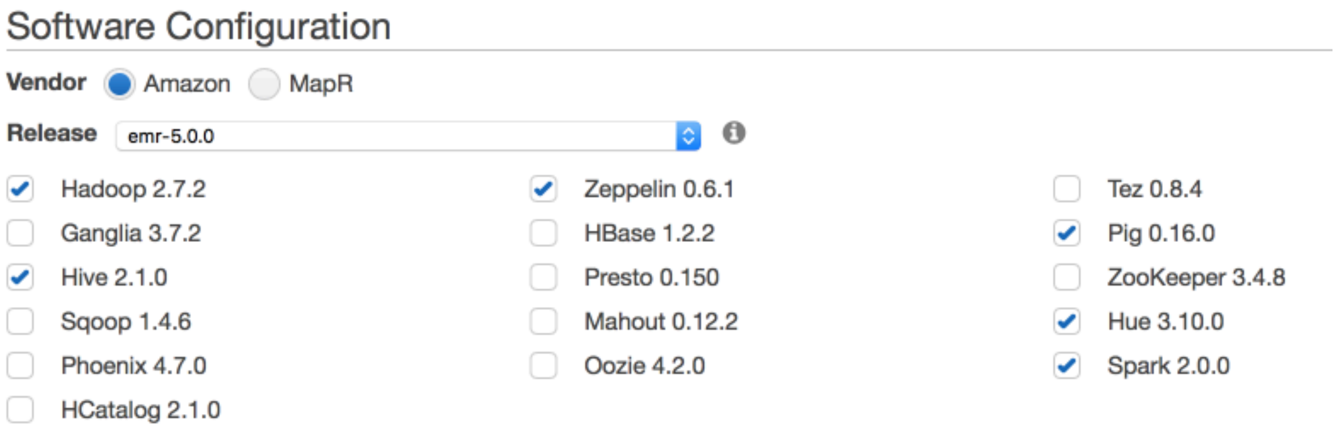
Launching an EMR cluster interactively (advanced mode)

1. Log into the console: https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#
2. NOTE: SWITCH TO ***N. Virginia (US EAST)*** if you aren't already in that region.
3. Select 'Create Cluster'
4. Select the "Go to advanced options"



5. Select the following applications:
    - Pig #.#.# (likely selected by default)
    - Hive #.#.# (likely selected by default)
    - Hue (likely selected by default)

- Spark #.#.#
- Zeppelin #.#.#



Software Configuration

Vendor ● Amazon ○ MapR

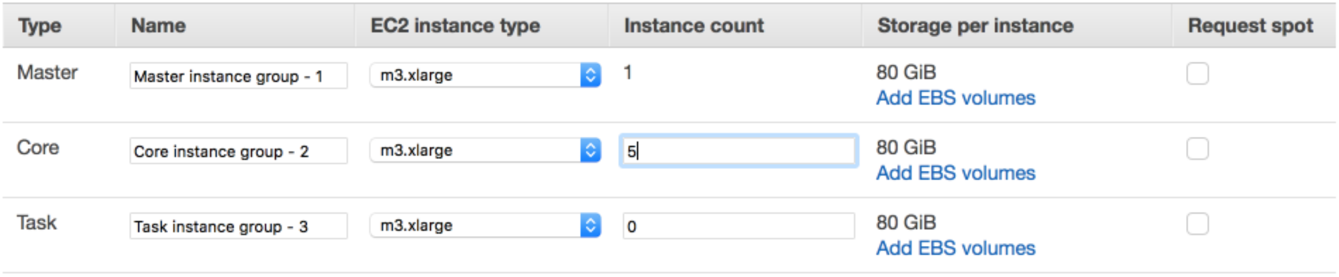
Release emr-5.0.0

☑ Hadoop 2.7.2
□ Ganglia 3.7.2
☑ Hive 2.1.0
□ Sqoop 1.4.6
□ Phoenix 4.7.0
□ HCatalog 2.1.0

☑ Zeppelin 0.6.1
□ HBase 1.2.2
□ Presto 0.150
□ Mahout 0.12.2
□ Oozie 4.2.0

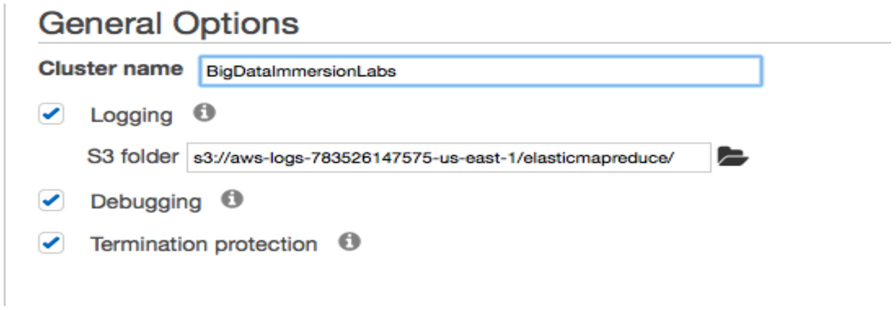
□ Tez 0.8.4
☑ Pig 0.16.0
□ ZooKeeper 3.4.8
☑ Hue 3.10.0
☑ Spark 2.0.0

6. Select Next
7. Keep the default VPC selected
8. Select m3.xlarge for both the core/master instance type and update the number of core nodes to be 5. Leave the number of Task nodes set to 0



Type	Name	EC2 instance type	Instance count	Storage per instance	Request spot
Master	Master instance group - 1	m3.xlarge	1	80 GiB Add EBS volumes	□
Core	Core instance group - 2	m3.xlarge	5	80 GiB Add EBS volumes	□
Task	Task instance group - 3	m3.xlarge	0	80 GiB Add EBS volumes	□

9. Select Next
10. Enter name for cluster: '<YourInitials>-BigDataImmersionLabs'



General Options

Cluster name BigDataImmersionLabs

☑ Logging

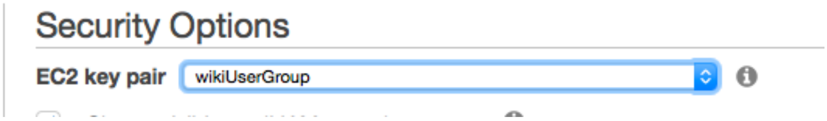
S3 folder s3://aws-logs-783526147575-us-east-1/elasticmapreduce/

☑ Debugging

☑ Termination protection

- (Specify a S3 logging location if one isn't already)

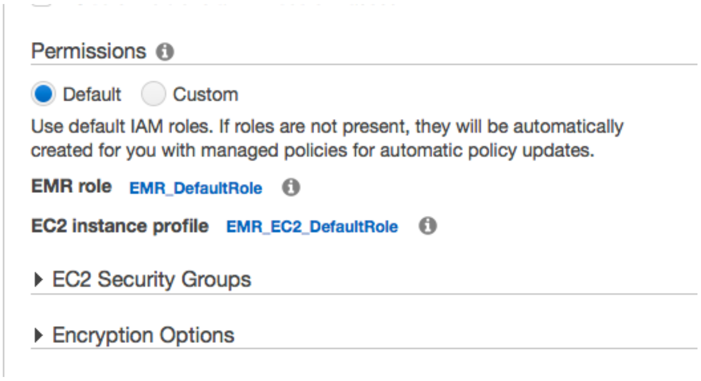
11. Select Next
12. Select your keypair



Security Options

EC2 key pair: wikiUserGroup

13. Keep the default permissions



Permissions

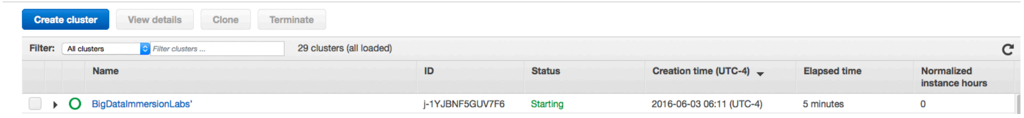
● Default ○ Custom

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

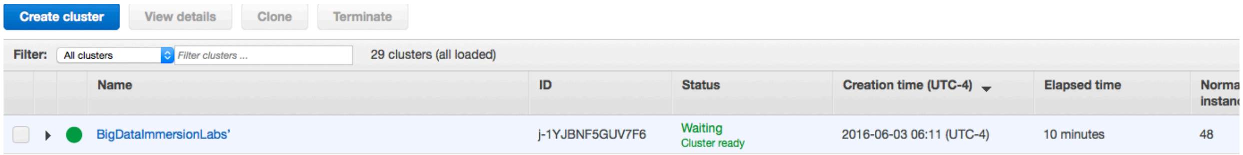
14. Select “Create cluster”
15. Go back to the EMR Cluster list.
    - You’ll notice it starting:



Filter: All clusters | Filter clusters ... | 29 clusters (all loaded)

Name	ID	Status	Creation time (UTC-4)	Elapsed time	Normalized instance hours
BigDataImmersionLabs'	j-1YJBNF5GUV7F6	Starting	2016-06-03 06:11 (UTC-4)	5 minutes	0

16. Wait until it’s in a Waiting state. Press the refresh icon in the top right of the table ↻ every couple minutes to refresh.
    - This is what it will look like when it’s ready:



Filter: All clusters | Filter clusters ... | 29 clusters (all loaded)

Name	ID	Status	Creation time (UTC-4)	Elapsed time	Norma instan
BigDataImmersionLabs'	j-1YJBNF5GUV7F6	Waiting Cluster ready	2016-06-03 06:11 (UTC-4)	10 minutes	48

Copyright 2016, Amazon Web Services, All Rights Reserved