Howto deploy Lyftron spark driver to AWS EMR cluster.

Requirements:

- AWS account
- AWS web console with permissions enough to create AWS EMR cluster and EC2 key pair
- scp/sftp and ssh client

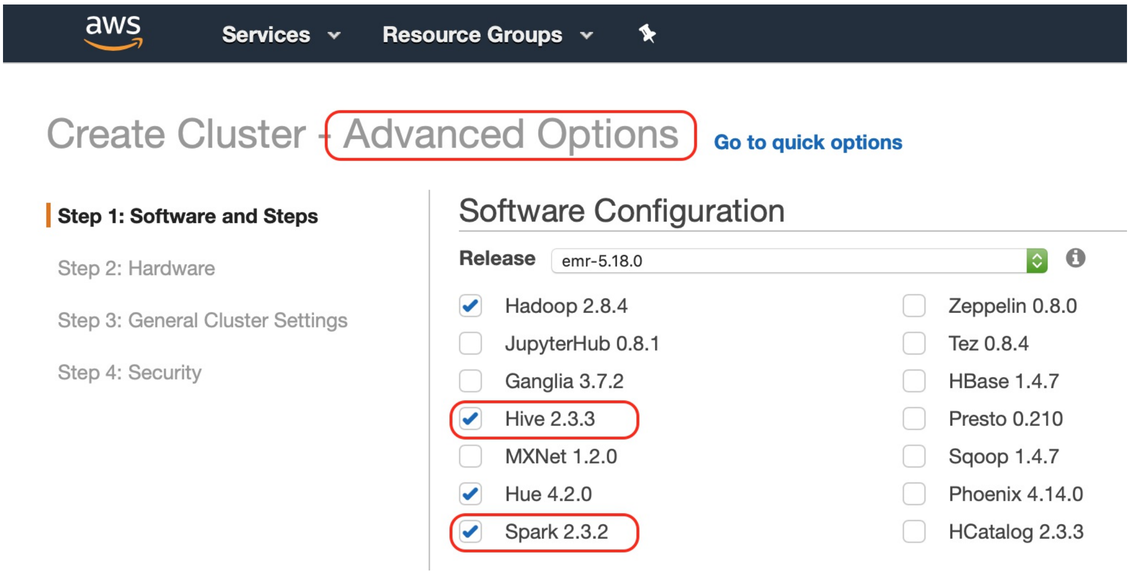
Short roadmap:

- Create and configure AWS EMR cluster
- Upload driver to the cluster
- Start driver

Create AWS EMR cluster

You have to create AWS EMR cluster to proceed. Use following details while creating

- Software configuration:
  - use `emr-5.19.0` release, it's latest version with spark 2.3 release
  - ensure both spark and hive are enabled



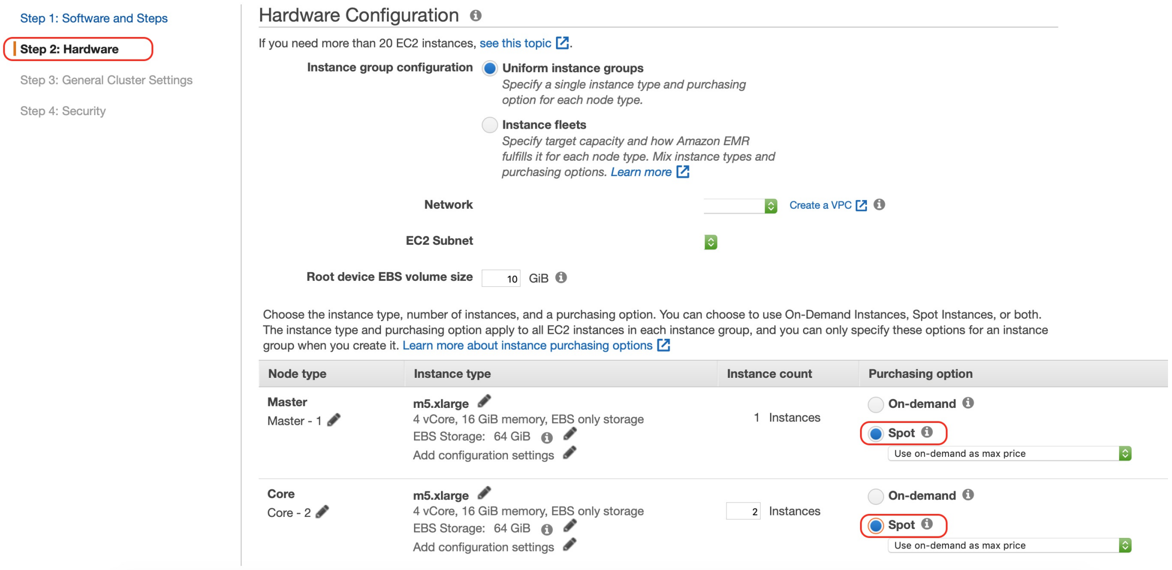
Create Cluster - Advanced Options

Software Configuration

Release emr-5.18.0

☑ 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 2.3.2
☐ Zeppelin 0.8.0
☐ Tez 0.8.4
☐ HBase 1.4.7
☐ Presto 0.210
☐ Sqoop 1.4.7
☐ Phoenix 4.14.0
☐ HCatalog 2.3.3

- Hardware configuration: pick instance type and number of instances you need. Minimal instance types are m4.large/c4.large.
  Use "Advanced options" during EMR creation to choose spot instances



Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

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

Create a VPC

EC2 Subnet

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

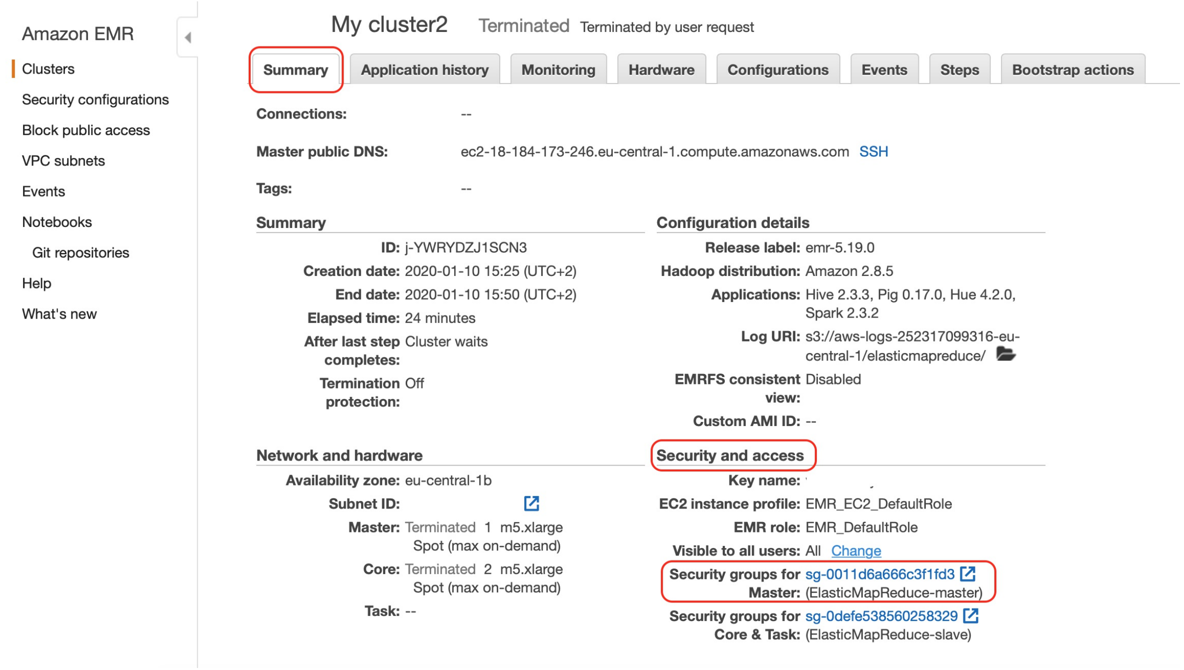
Node type	Instance type	Instance count	Purchasing option
Master Master - 1	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	1 Instances	○ On-demand ● Spot Use on-demand as max price
Core Core - 2	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: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

- it may save funds.

- Security and access:
  select existing key pair or create new EC2 key pair and select it. It allows cluster access with SSH client.

After cluster creation we need to configure Security Groups to permit us to connect to SSH and Lyftron connection:

- click to the created cluster, "Summary" tab



My cluster2 Terminated Terminated by user request

Connections: --

Master public DNS: ec2-18-184-173-246.eu-central-1.compute.amazonaws.com SSH

Tags: --

Summary

ID: j-YWRYDZJ1SCN3
Creation date: 2020-01-10 15:25 (UTC+2)
End date: 2020-01-10 15:50 (UTC+2)
Elapsed time: 24 minutes
After last step completes: Cluster waits
Termination protection: Off

Configuration details

Release label: emr-5.19.0
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.3, Pig 0.17.0, Hue 4.2.0, Spark 2.3.2
Log URI: s3://aws-logs-252317099316-eu-central-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

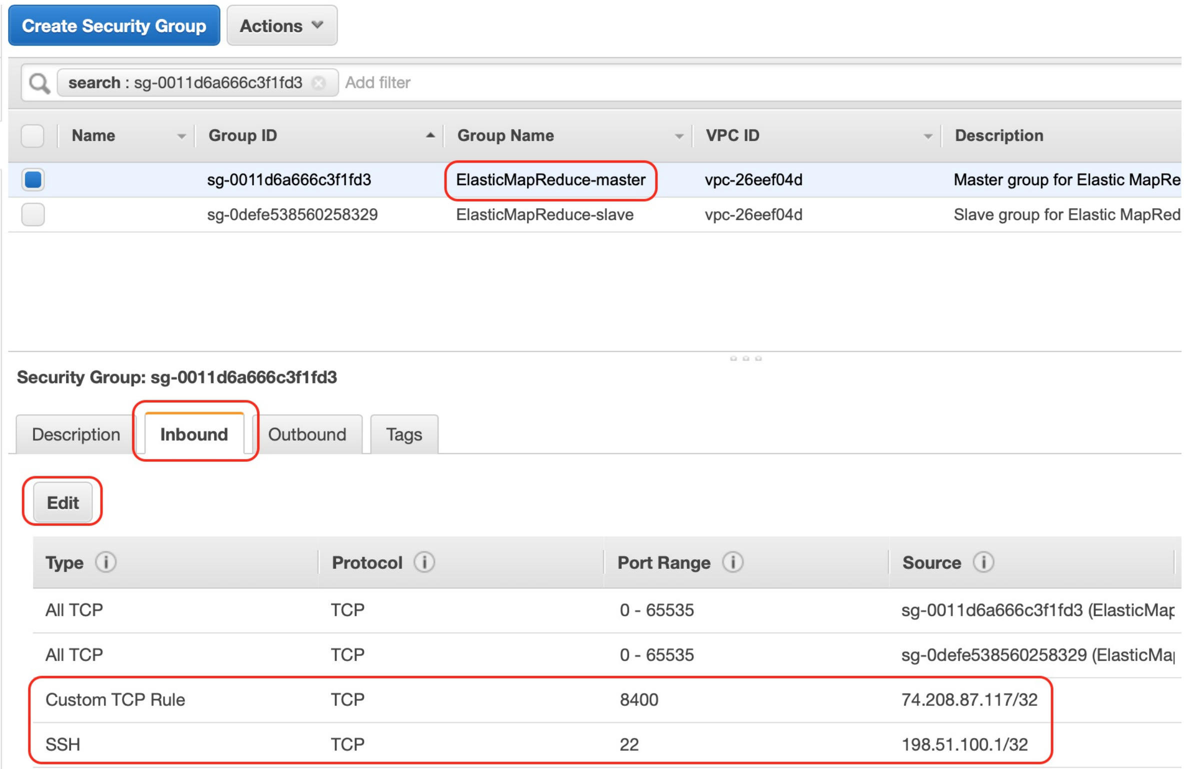
Network and hardware

Availability zone: eu-central-1b
Subnet ID:
Master: Terminated 1 m5.xlarge Spot (max on-demand)
Core: Terminated 2 m5.xlarge Spot (max on-demand)
Task: --

Security and access

Key name:
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-0011d6a666c3f1fd3 (ElasticMapReduce-master)
Security groups for Core & Task: sg-0defe538560258329 (ElasticMapReduce-slave)

- under "Security and access" click to the link value of "Security groups for Master"
- in the opened "Security groups" window choose "ElasticMapReduce-master"
- choose "Inbound" tab below and click "Edit"



search : sg-0011d6a666c3f1fd3

Name	Group ID	Group Name	VPC ID	Description
	sg-0011d6a666c3f1fd3	ElasticMapReduce-master	vpc-26eef04d	Master group for Elastic MapRe
	sg-0defe538560258329	ElasticMapReduce-slave	vpc-26eef04d	Slave group for Elastic MapRed

Security Group: sg-0011d6a666c3f1fd3

Type	Protocol	Port Range	Source
All TCP	TCP	0 - 65535	sg-0011d6a666c3f1fd3 (ElasticMap
All TCP	TCP	0 - 65535	sg-0defe538560258329 (ElasticMa
Custom TCP Rule	TCP	8400	74.208.87.117/32
SSH	TCP	22	198.51.100.1/32

- click "Add Rule", select "SSH" as type and "My IP" as Source. Don't forget to use some comment.
- click "Add Rule", select "Custom TCP rule" as type, 8400 as Port Range and custom source `74.208.87.117/32`
- click "Save"

Now we're ready to deploy driver to EMR cluster

Upload driver to the cluster

- click to the created cluster, "Summary" tab, and pick value of "Master public DNS". That's the endpoint we'll need to connect to via SSH port 22.
- upload content of this directory(not the directory itself) via scp/sftp to remote `/home/hadoop` directory(default directory),
  use `hadoop` as username and endpoint from previous line. You need to specify EC2 key you'd used to create a EMR cluster. Here is `scp` example:
  `bash scp -i emr.key -r spark_driver/* hadoop@ec2-3-120-231-87.eu-central-1.compute.amazonaws.com:`

Start driver

- connect to "Master public DNS" cluster endpoint via ssh client, use port 22, user `hadoop` and EC2 key used during EMR cluster creation.

Here is `ssh` example:

```
bash ssh -i emr.key hadoop@ec2-3-120-231-87.eu-central-1.compute.amazonaws.com
```

- run `screen` command on the remote terminal to keep driver up and running after ssh disconnect
- check what files do we have in default directory, it should be `/home/hadoop` directory. Here are commands with output example:

```
bash [hadoop@ip-172-31-5-142 ~]$ pwd /home/hadoop [hadoop@ip-172-31-5-142 ~]$ ls -la total 28580 drwxr-xr-x 6 hadoop hadoop 4096 дек 13 22:47 . drwxr-xr-x 4 root root 4096 дек 13 18:03 .. drwxr-xr-x 2 hadoop hadoop 4096 окт 23 2018 .aws -rw-r--r-- 1 hadoop hadoop 86 окт 19 2018 .bash_profile -rw-r--r-- 1 hadoop hadoop 357 окт 19 2018 .bashrc drwxr-xr-x 2 hadoop hadoop 4096 дек 13 18:35 conf -rw-rw-r-- 1 hadoop hadoop 193831 дек 5 03:13 hadoop-lzo-0.4.20.jar -rwxr-xr-x 1 hadoop hadoop 1523 дек 13 18:35 log4j.properties -rwxr-xr-x 1 hadoop hadoop 29009093 дек 13 18:35 lyftron.spark.driver-1.0-spark-2.1.0.jar drwxr-xr-x 3 hadoop hadoop 4096 дек 13 18:36 sqljdbc drwx------ 2 hadoop hadoop 4096 дек 13 18:04 .ssh -rwxr--r-- 1 hadoop hadoop 2183 дек 13 19:33 start-driver.sh [hadoop@ip-172-31-5-142 ~]$
```

- start the driver via following command:

```
bash bash ./start-driver.sh
```

You should get output like following

```
Starting Lyftron driver, cmd arguments: lyftron.driver.port=8400 lyftron.driver.p
Creating app context
Initializing configuration
Starting conductor
```

- That's it, you're done.