LearnKartS

A Training Services Company

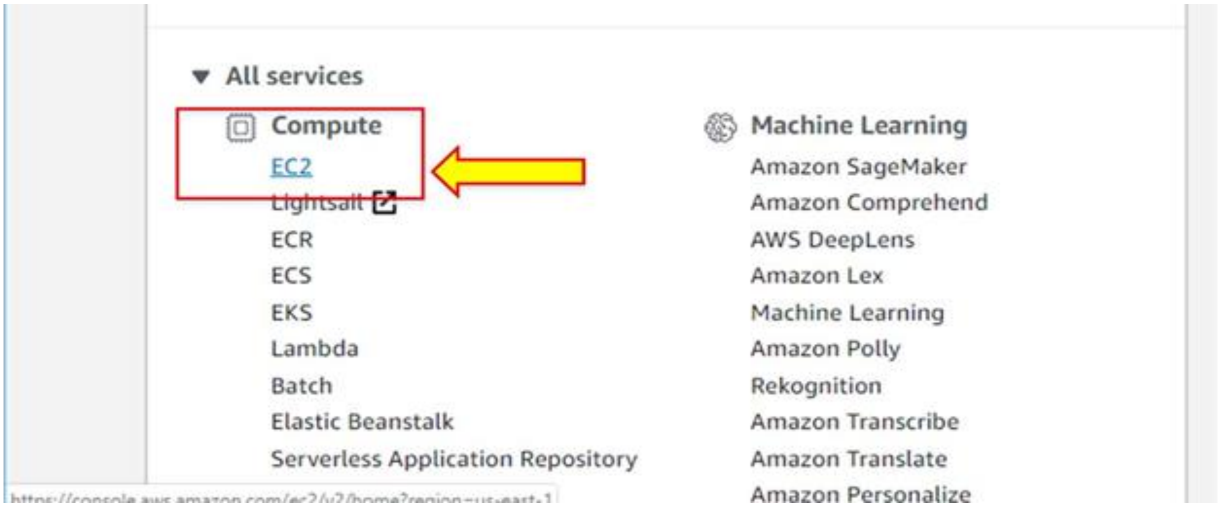
SOLUTION:

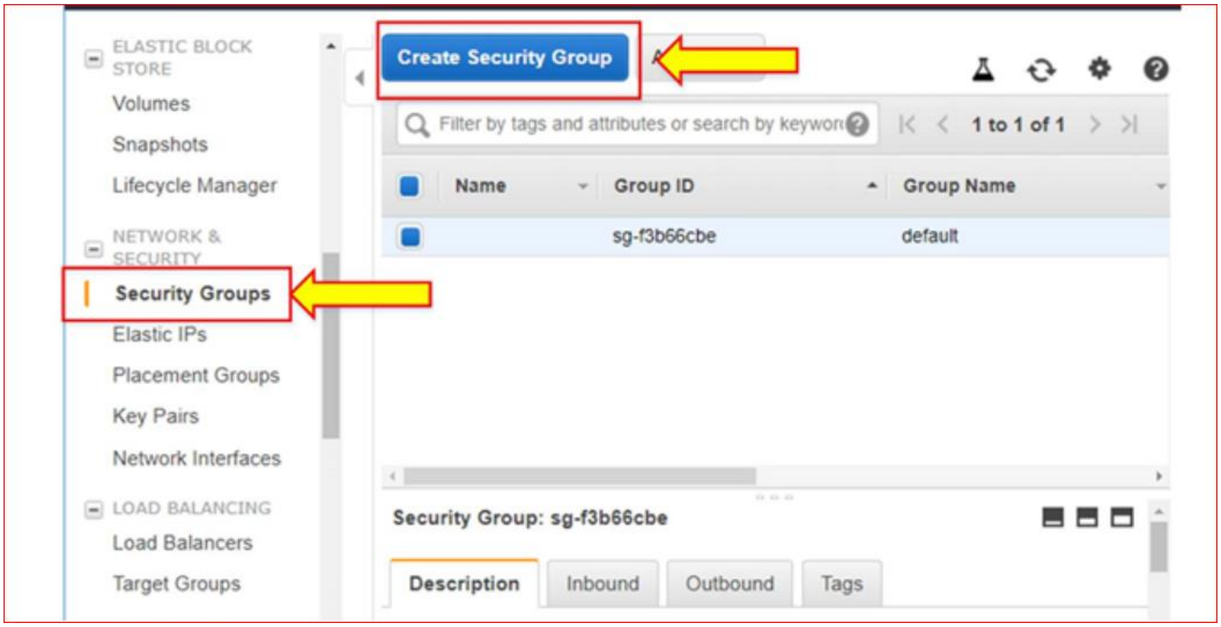
BIG DATA ANALYTICS FOR ANALYZING LARGE DATA SETS

Steps:

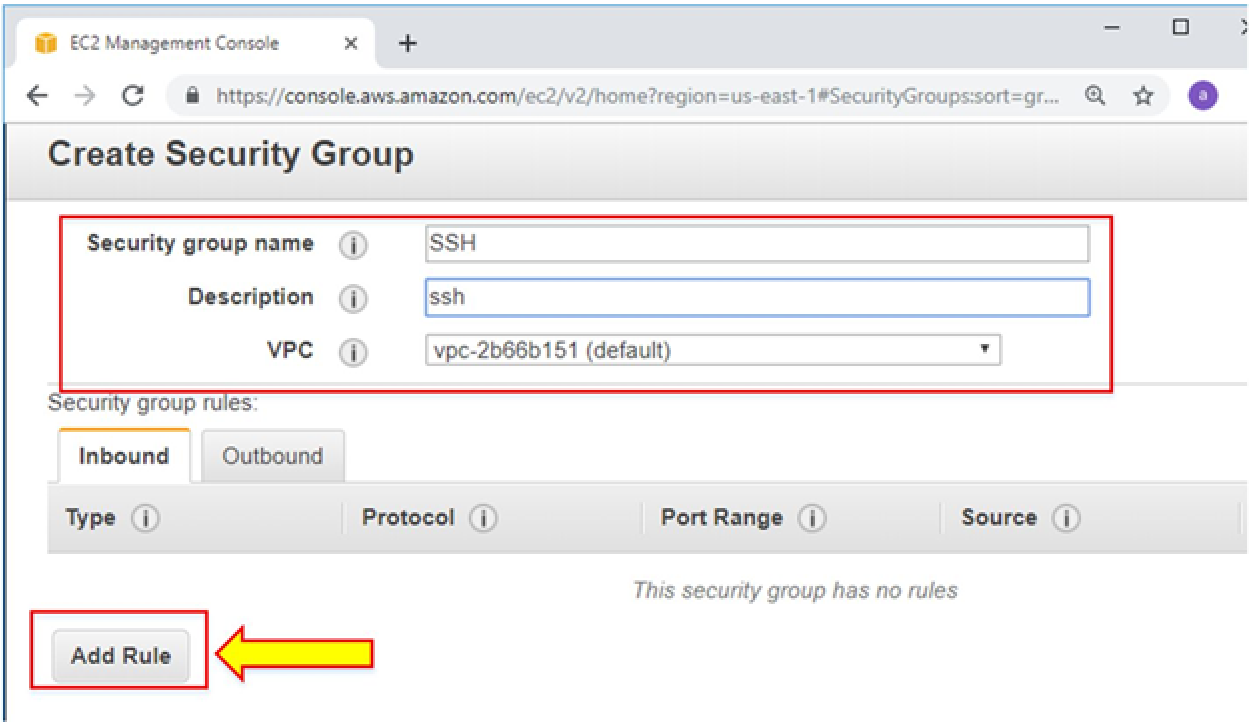
For the requirement, they use Elastic MapReduce tool for analyzing the process and in that use Hive for querying.

Step 1: First, create a Security group with appropriate SSH rule for EMR master node.





Name	Group ID	Group Name
	sg-f3b66cbe	default



Create Security Group

Security group name: SSH

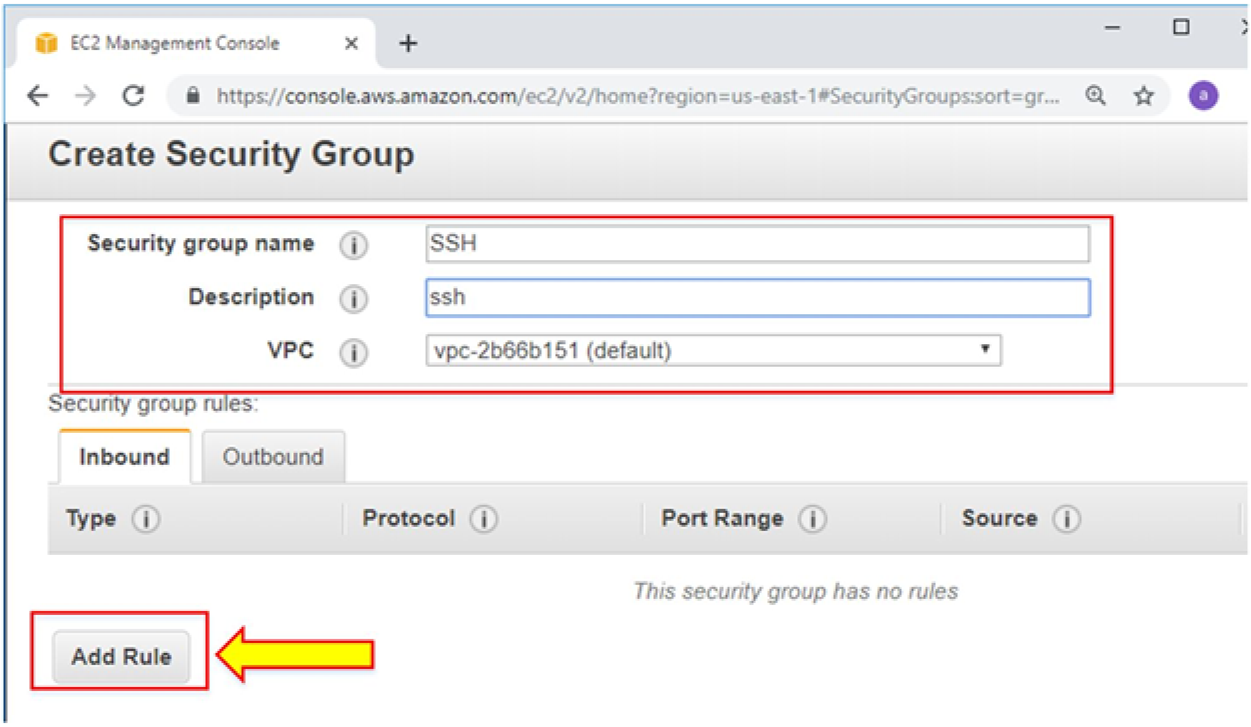
Description: ssh

VPC: vpc-2b66b151 (default)

Security group rules:

Type	Protocol	Port Range	Source

This security group has no rules


EC2 Management Console
https://console.aws.amazon.com/ec2/v2/home?region=us-east-1#SecurityGroups:sort=gr...

Create Security Group

Security group name: SSH

Description: ssh

VPC: vpc-2b66b151 (default)

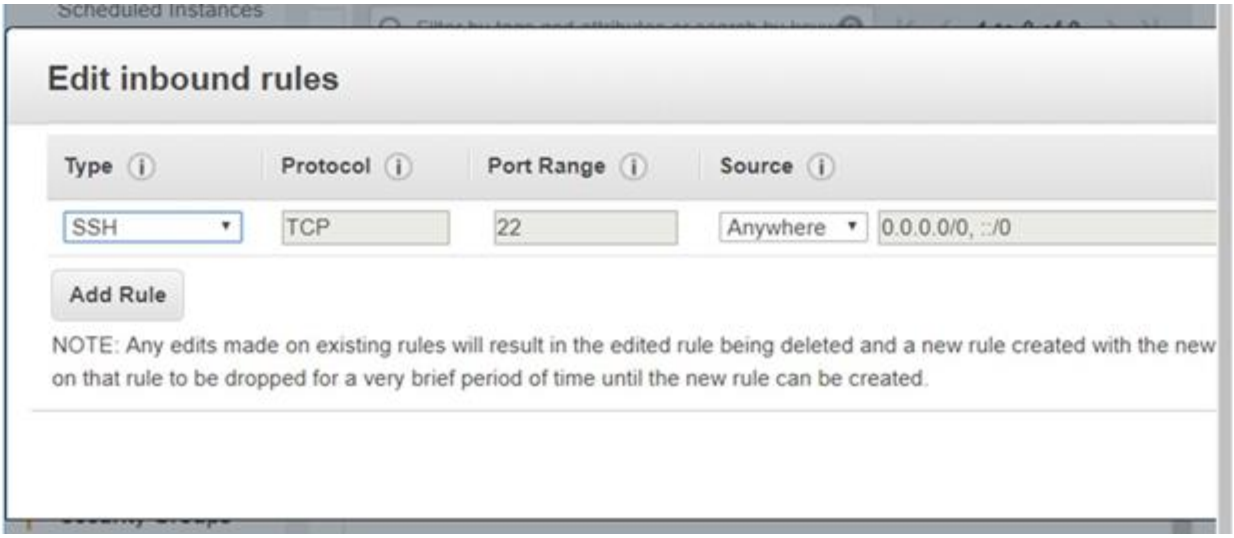
Security group rules:

Inbound
Outbound

Type	Protocol	Port Range	Source

This security group has no rules

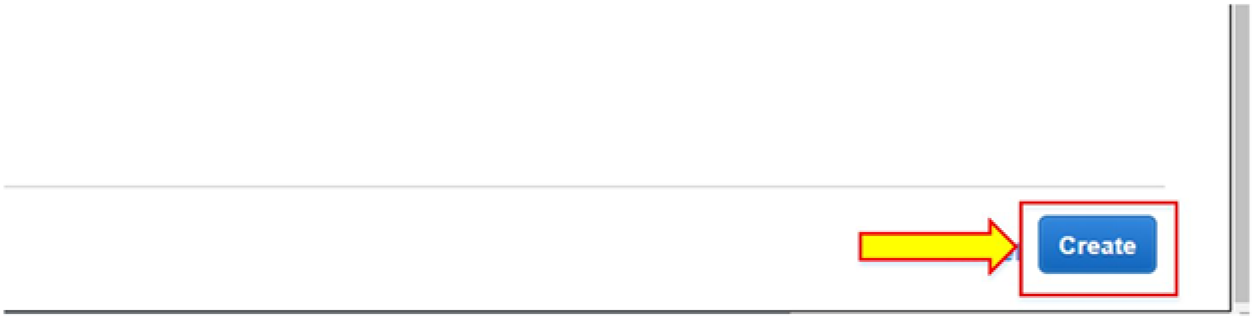
Add Rule

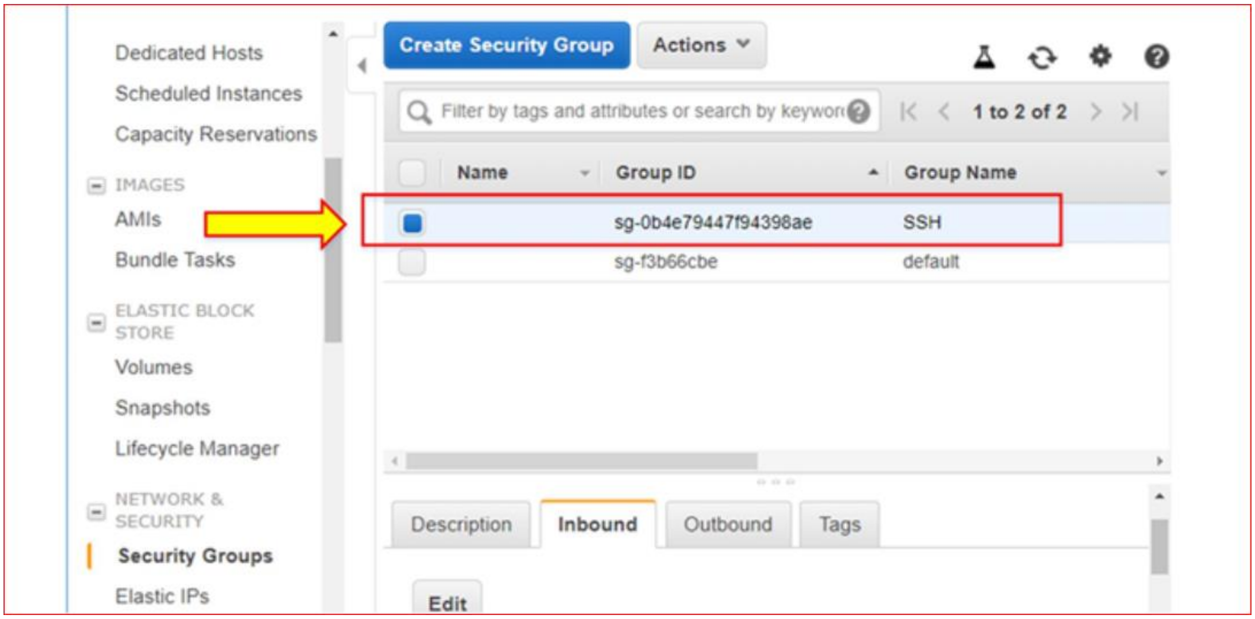


Edit inbound rules

Type	Protocol	Port Range	Source	
SSH	TCP	22	Anywhere	0.0.0.0/0, ::/0

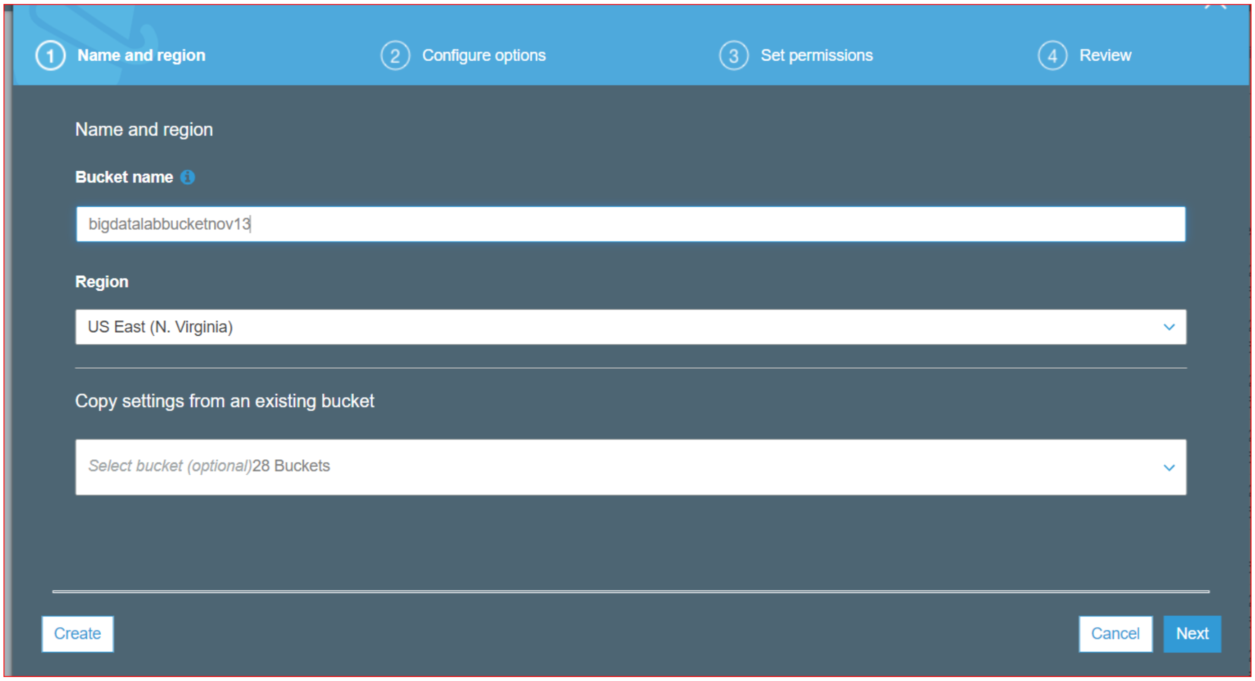
NOTE: Any edits made on existing rules will result in the edited rule being deleted and a new rule created with the new on that rule to be dropped for a very brief period of time until the new rule can be created.





	Name	Group ID	Group Name
■		sg-0b4e79447f94398ae	SSH
□		sg-f3b66cbe	default

Step 2: Create an S3 bucket which is public to store the table.

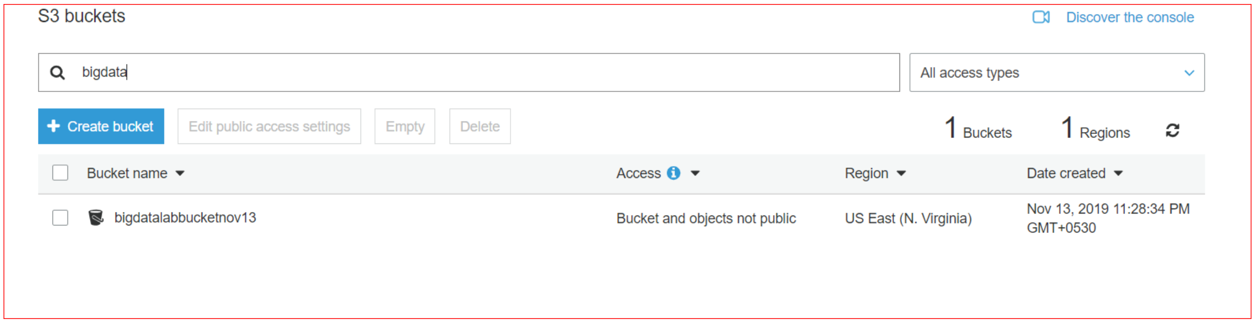


Name and region

Bucket name: bigdatalabbucketnov13

Region: US East (N. Virginia)

Copy settings from an existing bucket: Select bucket (optional)28 Buckets

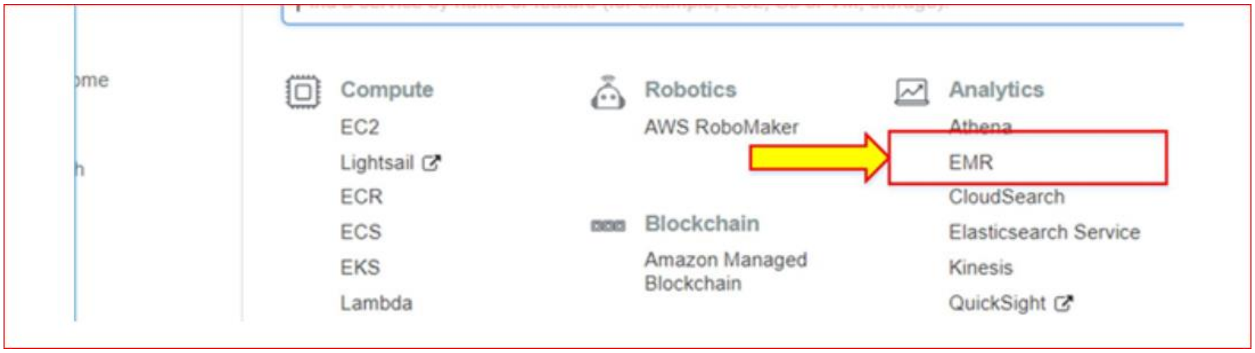


bigdata

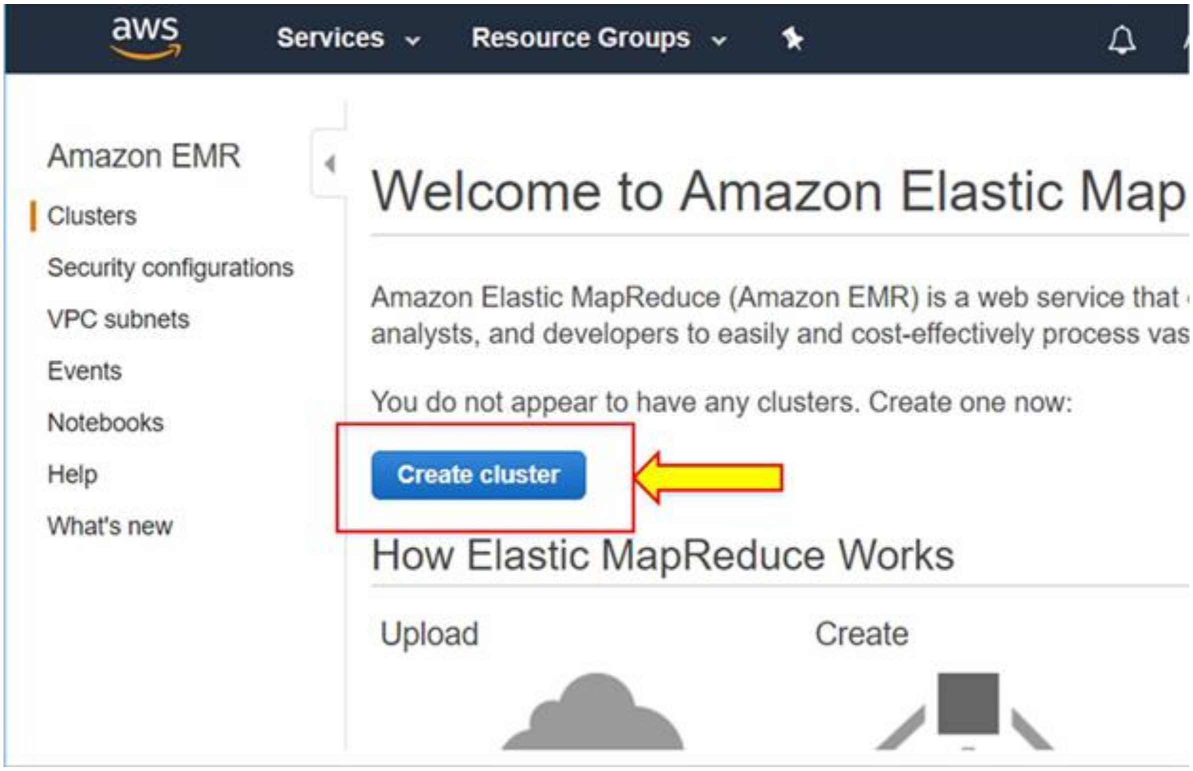
1 Buckets 1 Regions

Bucket name	Access	Region	Date created
bigdatalabbucketnov13	Bucket and objects not public	US East (N. Virginia)	Nov 13, 2019 11:28:34 PM GMT+0530

Step 3: Go to EMR under Analytics on AWS Console



Step 4: Click on Create Cluster and go to advanced options.

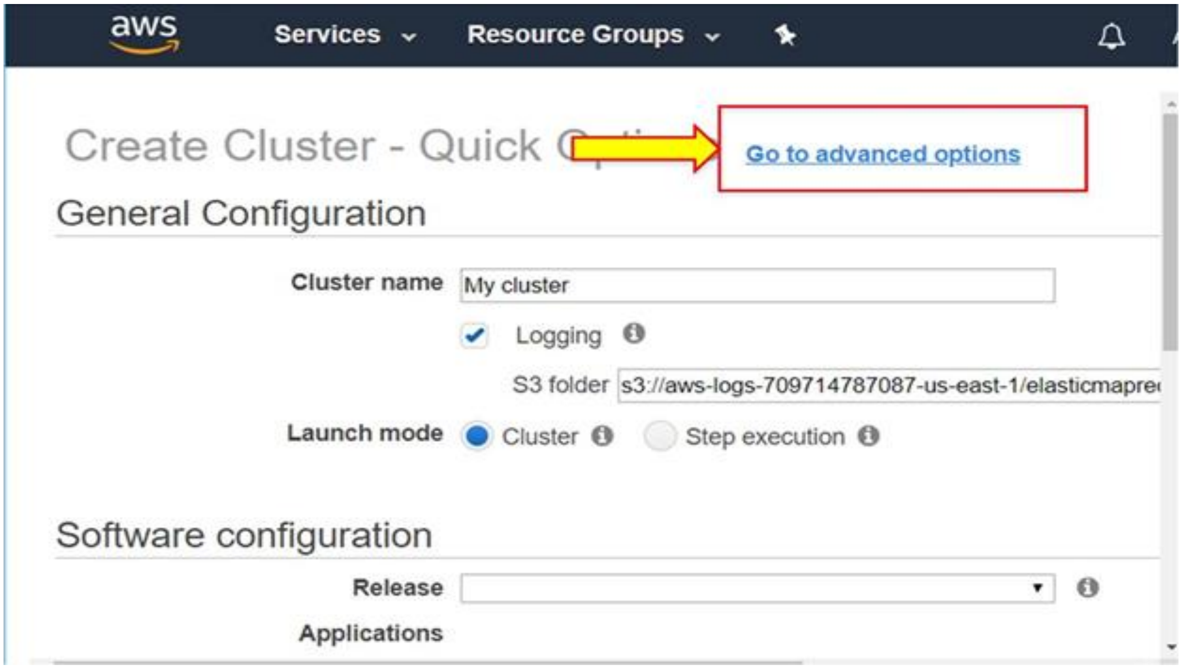


Welcome to Amazon Elastic Map

Amazon Elastic MapReduce (Amazon EMR) is a web service that
analysts, and developers to easily and cost-effectively process vas

You do not appear to have any clusters. Create one now:

How Elastic MapReduce Works



Create Cluster - Quick Options

General Configuration

Cluster name: My cluster

☑ Logging

S3 folder: s3://aws-logs-709714787087-us-east-1/elasticmapre

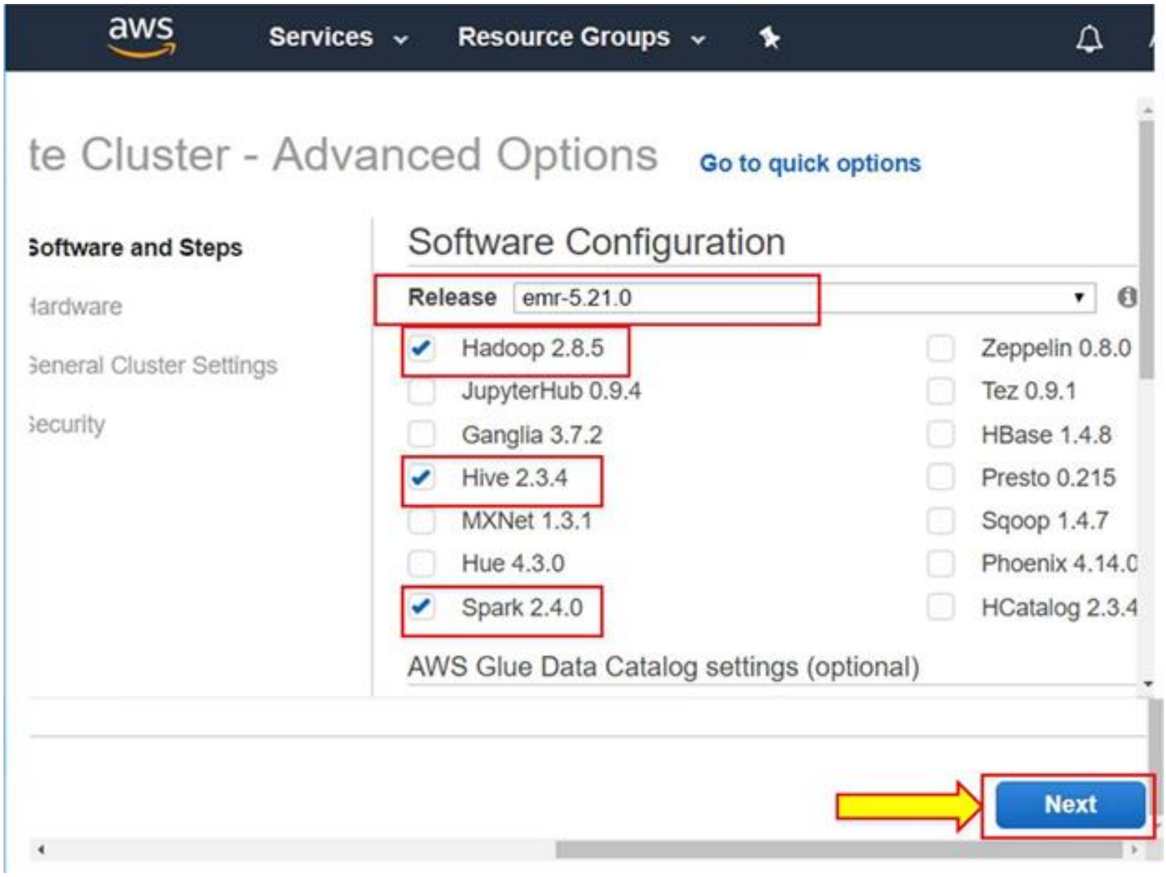
Launch mode: ● Cluster ○ Step execution

Software configuration

Release:

Applications

Step 5: Go to the latest EMR release. Check Hadoop, Hive and Spark as we are going to use these tools. Click Next.



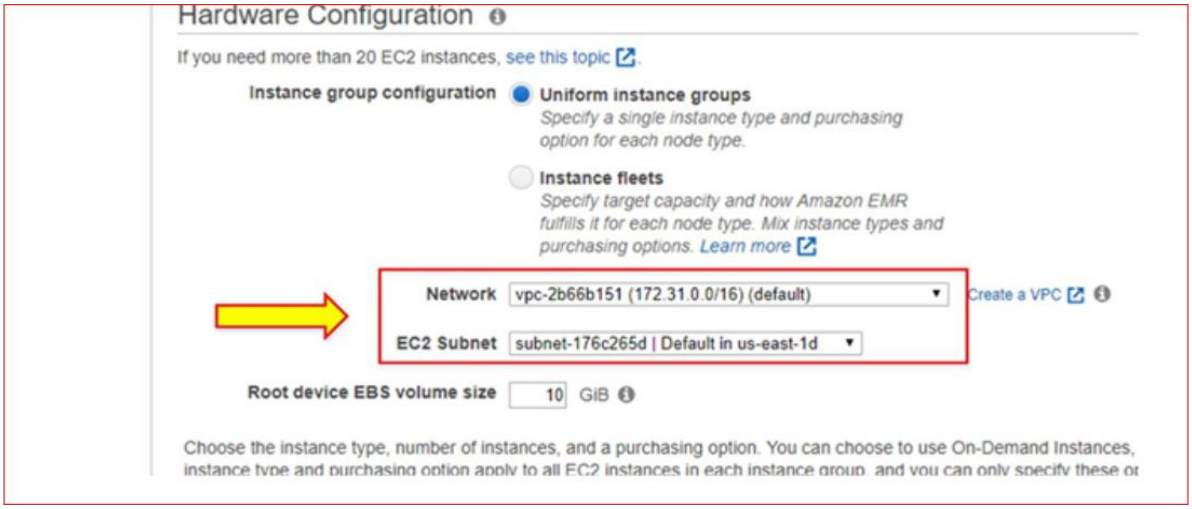
Software Configuration

Release emr-5.21.0

☑ Hadoop 2.8.5	□ Zeppelin 0.8.0
□ JupyterHub 0.9.4	□ Tez 0.9.1
□ Ganglia 3.7.2	□ HBase 1.4.8
☑ Hive 2.3.4	□ Presto 0.215
□ MXNet 1.3.1	□ Sqoop 1.4.7
□ Hue 4.3.0	□ Phoenix 4.14.0
☑ Spark 2.4.0	□ HCatalog 2.3.4

AWS Glue Data Catalog settings (optional)

Step 6: In Hardware configuration, select the VPC which has public subnet. Here we are using default VPC.



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

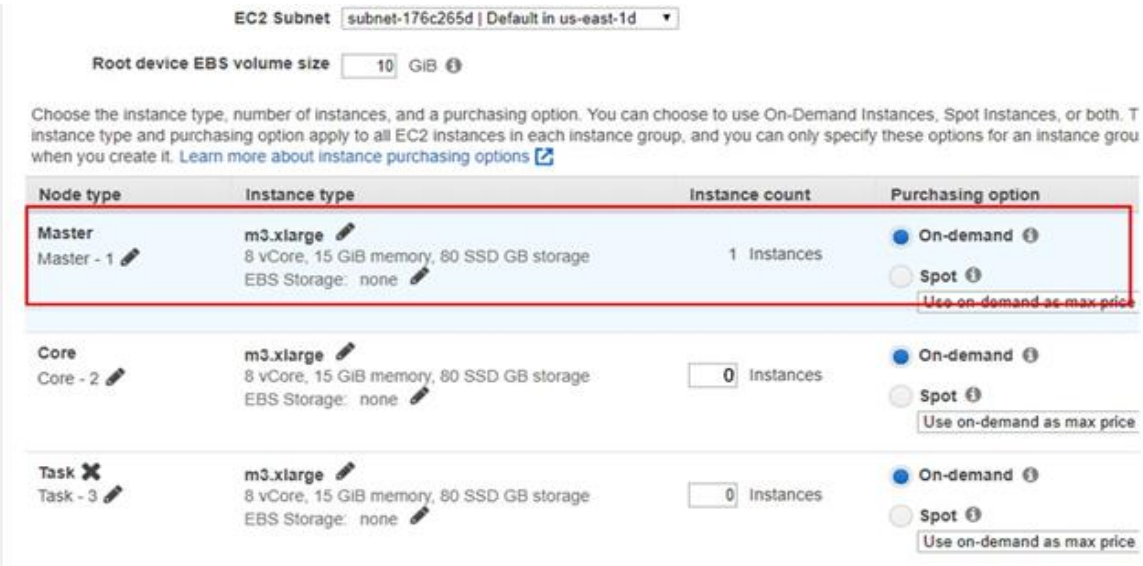
Network vpc-2b66b151 (172.31.0.0/16) (default)

EC2 Subnet subnet-176c265d | Default in us-east-1d

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, instance type and purchasing option apply to all EC2 instances in each instance group, and you can only specify these or

Step 7: Now go to the node and select the type for master node and its purchasing option. There is no need for a core node and task node.



EC2 Subnet: subnet-176c265d | Default in us-east-1d

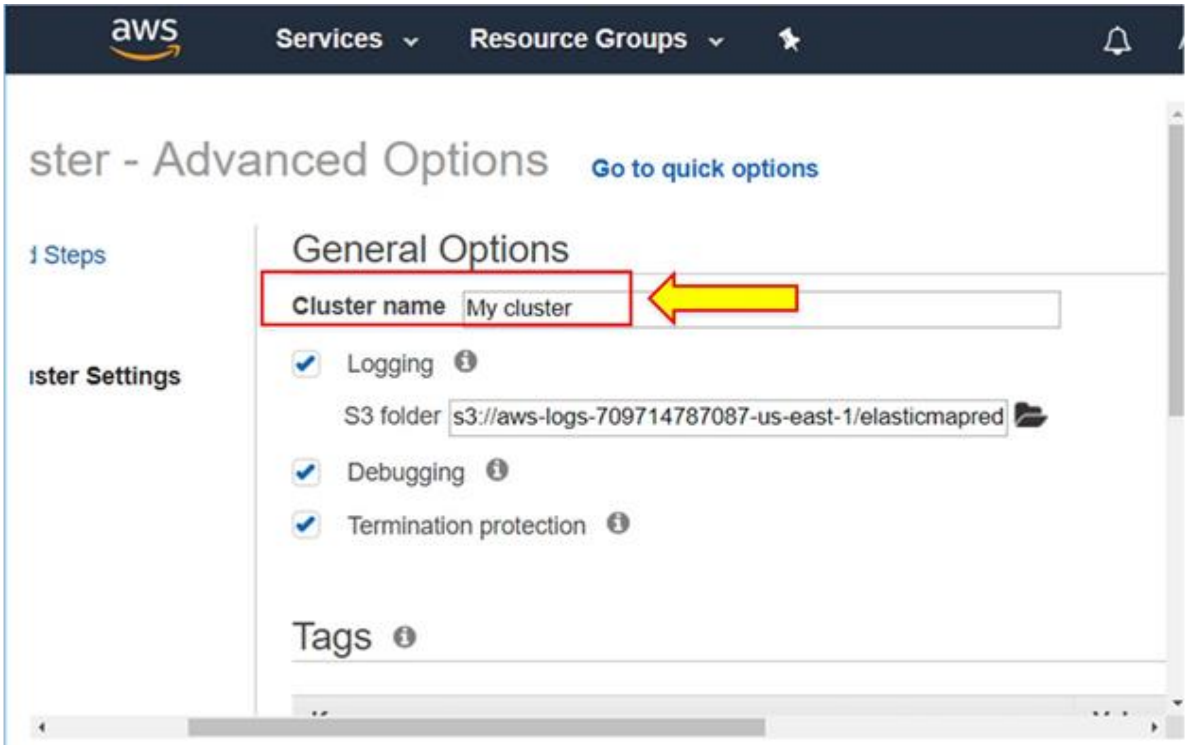
Root device EBS volume size: 10 GiB

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

Node type	Instance type	Instance count	Purchasing option
Master Master - 1	m3.xlarge 8 vCore, 15 GiB memory, 80 SSD GB storage EBS Storage: none	1 Instances	● On-demand ○ Spot Use on-demand as max price
Core Core - 2	m3.xlarge 8 vCore, 15 GiB memory, 80 SSD GB storage EBS Storage: none	0 Instances	● On-demand ○ Spot Use on-demand as max price
Task Task - 3	m3.xlarge 8 vCore, 15 GiB memory, 80 SSD GB storage EBS Storage: none	0 Instances	● On-demand ○ Spot Use on-demand as max price

Step 8: Click Next

Step 9: Enter the cluster name and leave everything as it is. Click Next.



ster - Advanced Options Go to quick options

General Options

Cluster name: My cluster

☑ Logging

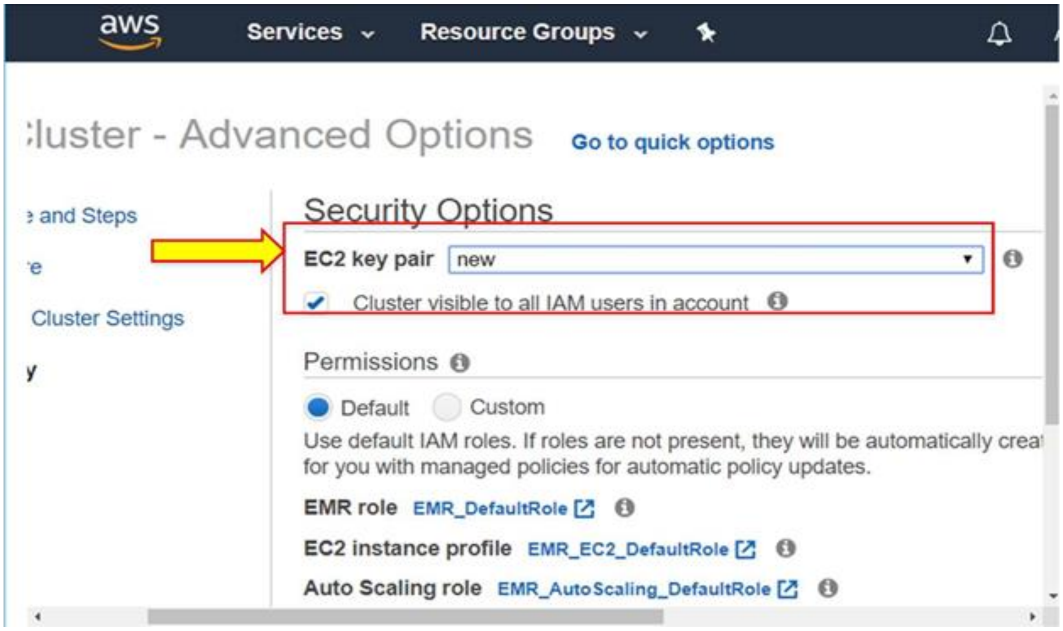
S3 folder: s3://aws-logs-709714787087-us-east-1/elasticmapred

☑ Debugging

☑ Termination protection

Tags

Step 10: In security option, select EC2 key pair as we want that to SSH the master node. Here we have used the key which already created in EC2 service.



Security Options

EC2 key pair: new

☑ Cluster visible to all IAM users in account

Permissions

● Default ○ Custom

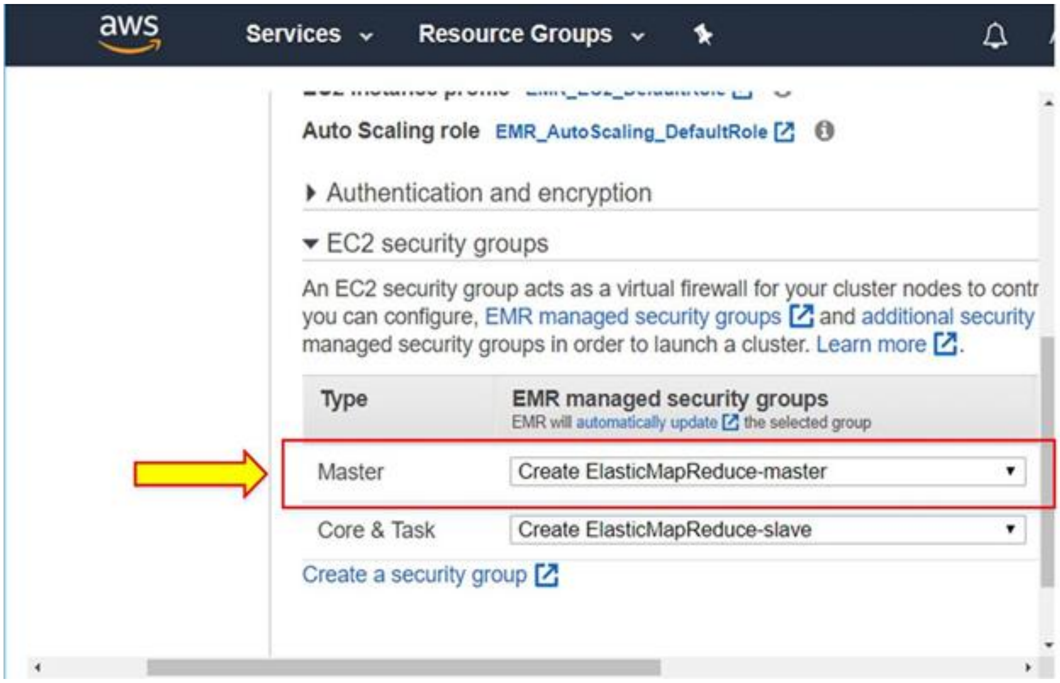
Use default IAM roles. If roles are not present, they will be automatically creat for you with managed policies for automatic policy updates.

EMR role EMR_DefaultRole

EC2 instance profile EMR_EC2_DefaultRole

Auto Scaling role EMR_AutoScaling_DefaultRole

Step 11: Leave everything as it is and in the security group of master node add additional security group. Here we add that Security group that we created in the start for SSH. Click on Assign Security group.



Auto Scaling role EMR_AutoScaling_DefaultRole

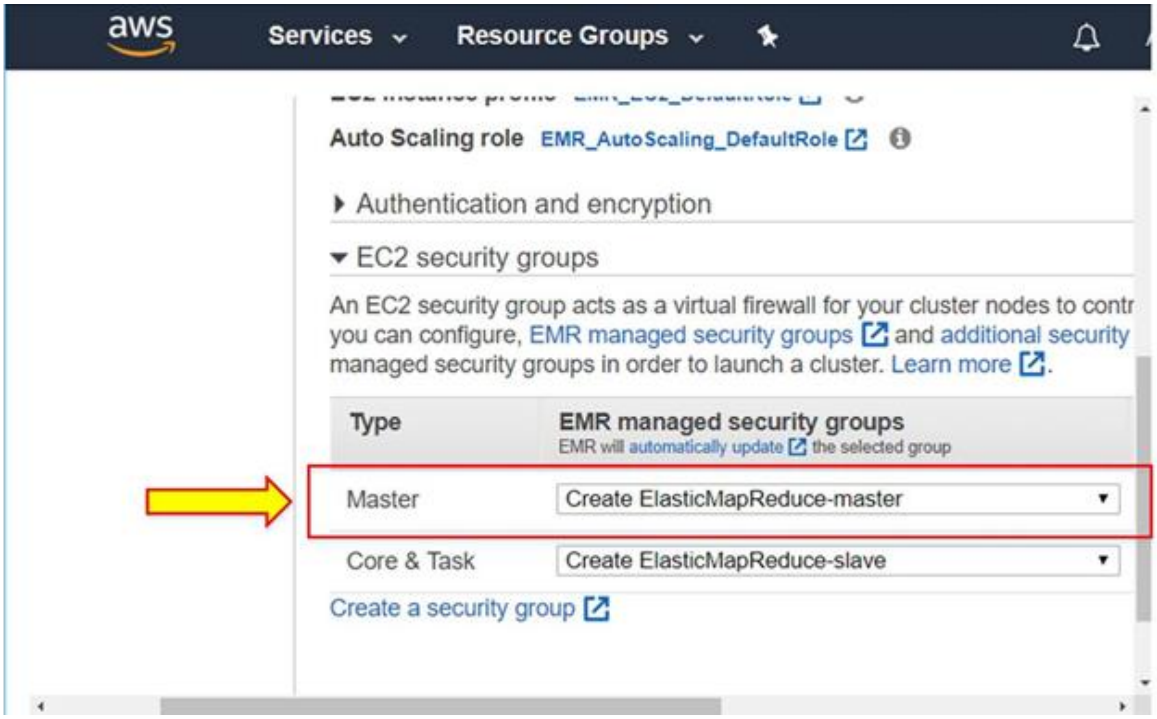
▸ Authentication and encryption

▾ EC2 security groups

An EC2 security group acts as a virtual firewall for your cluster nodes to contr you can configure, EMR managed security groups and additional security managed security groups in order to launch a cluster. Learn more.

Type	EMR managed security groups EMR will automatically update the selected group
Master	Create ElasticMapReduce-master
Core & Task	Create ElasticMapReduce-slave

Create a security group



Auto Scaling role EMR_AutoScaling_DefaultRole

Authentication and encryption

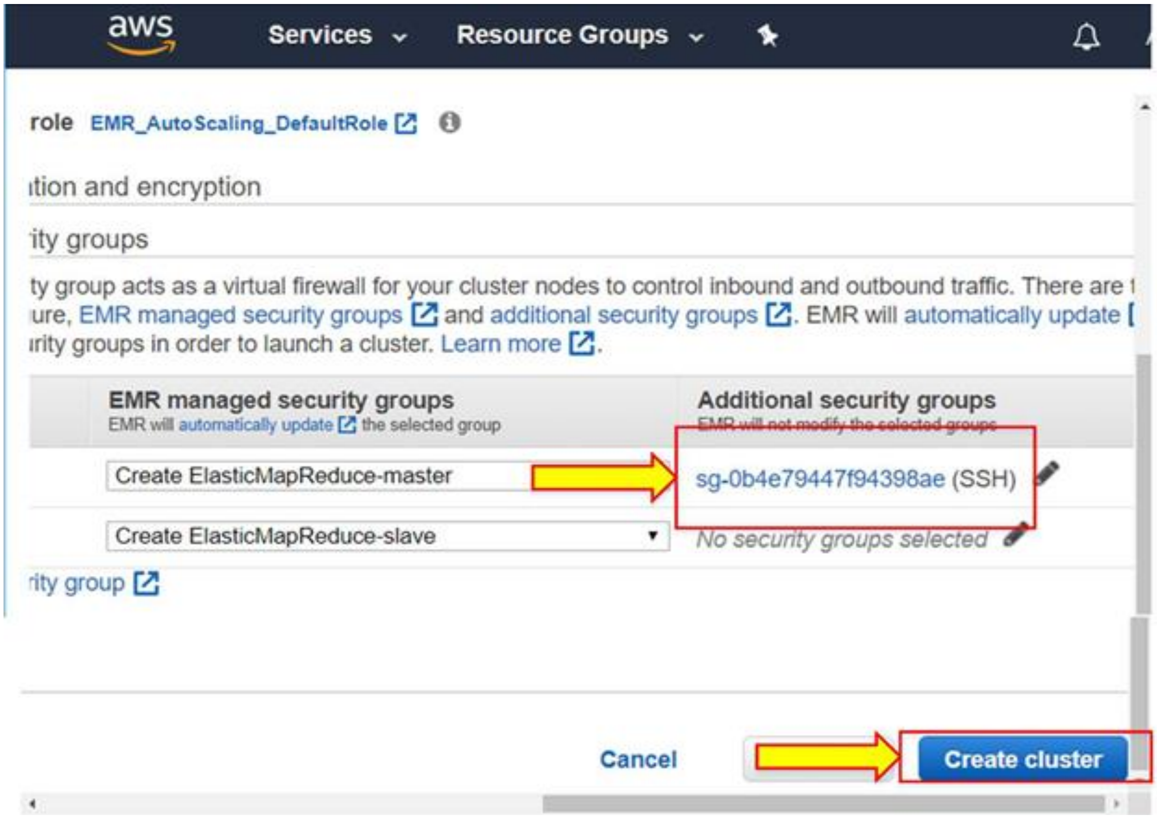
EC2 security groups

An EC2 security group acts as a virtual firewall for your cluster nodes to contr
you can configure, EMR managed security groups and additional security
managed security groups in order to launch a cluster. Learn more.

Type	EMR managed security groups EMR will automatically update the selected group
Master	Create ElasticMapReduce-master
Core & Task	Create ElasticMapReduce-slave

Create a security group

Step 12: Click on Create Cluster.



role EMR_AutoScaling_DefaultRole

ıtion and encryption

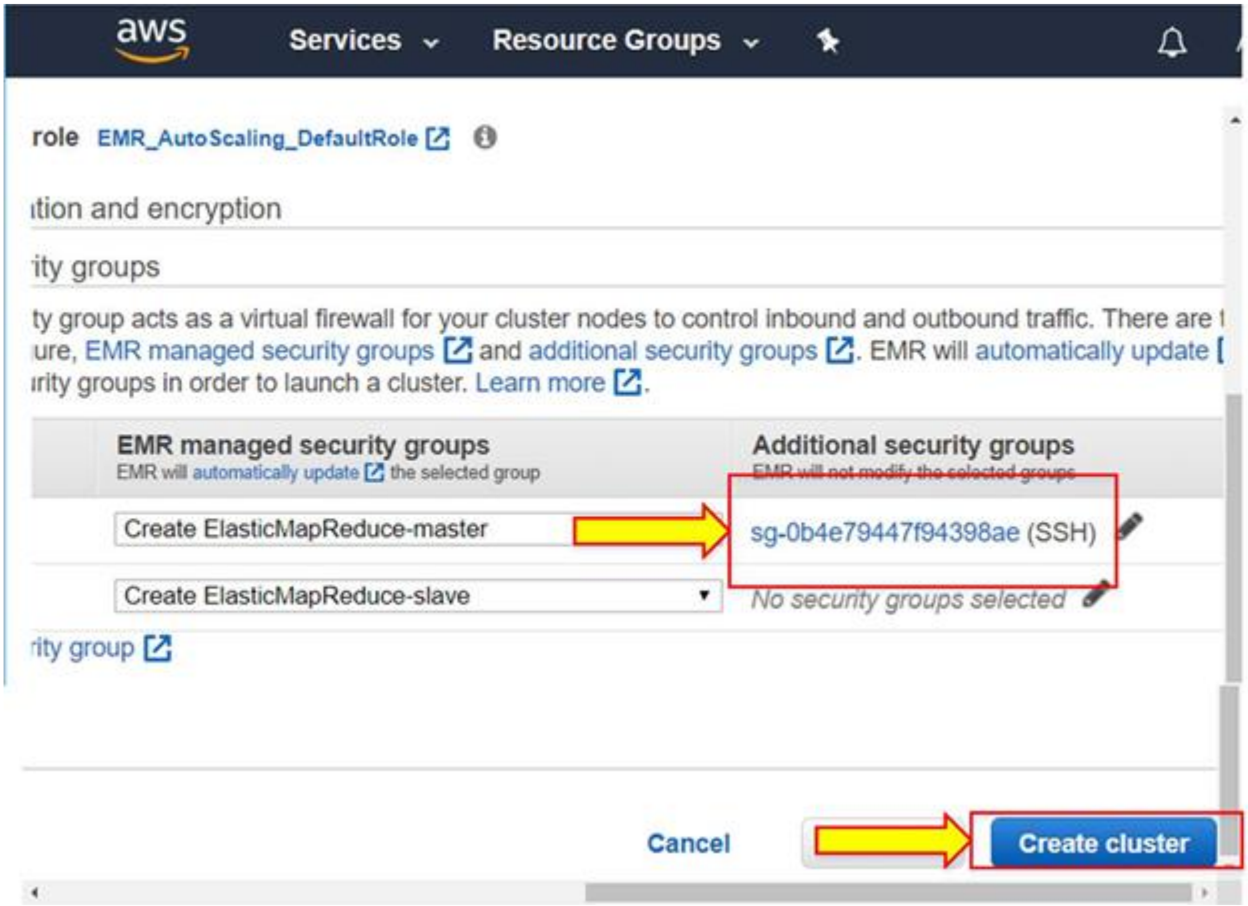
ity groups

ty group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are t
ure, EMR managed security groups and additional security groups. EMR will automatically update
ırity groups in order to launch a cluster. Learn more.

EMR managed security groups EMR will automatically update the selected group	Additional security groups EMR will not modify the selected groups
Create ElasticMapReduce-master	sg-0b4e79447f94398ae (SSH)
Create ElasticMapReduce-slave	No security groups selected

rity group

Step 13: Now cluster has started creating. Once it is created, take public DNS name of the cluster for SSH.



role EMR_AutoScaling_DefaultRole

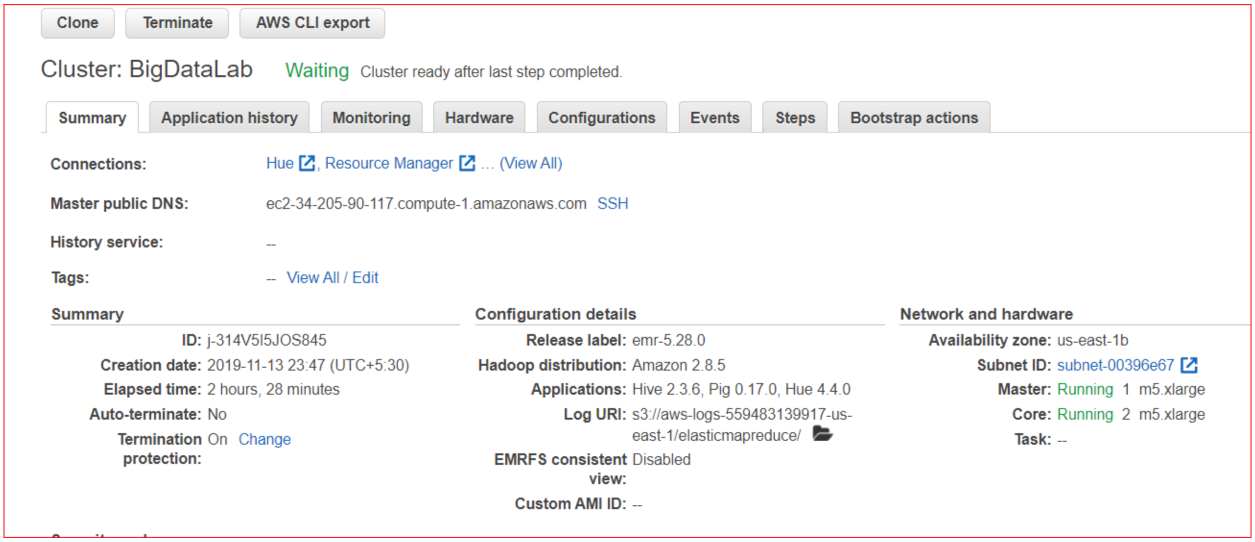
ition and encryption

ity groups

ty group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are
ure, EMR managed security groups and additional security groups. EMR will automatically update
irity groups in order to launch a cluster. Learn more.

EMR managed security groups EMR will automatically update the selected group	Additional security groups EMR will not modify the selected groups
Create ElasticMapReduce-master	sg-0b4e79447f94398ae (SSH)
Create ElasticMapReduce-slave	No security groups selected

rity group



Cluster: BigDataLab Waiting Cluster ready after last step completed.

Connections: Hue, Resource Manager ... (View All)

Master public DNS: ec2-34-205-90-117.compute-1.amazonaws.com SSH

History service: --

Tags: -- View All / Edit

Summary

ID: j-314V5I5JOS845
Creation date: 2019-11-13 23:47 (UTC+5:30)
Elapsed time: 2 hours, 28 minutes
Auto-terminate: No
Termination protection: On Change

Configuration details

Release label: emr-5.28.0
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, Pig 0.17.0, Hue 4.4.0
Log URI: s3://aws-logs-559483139917-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

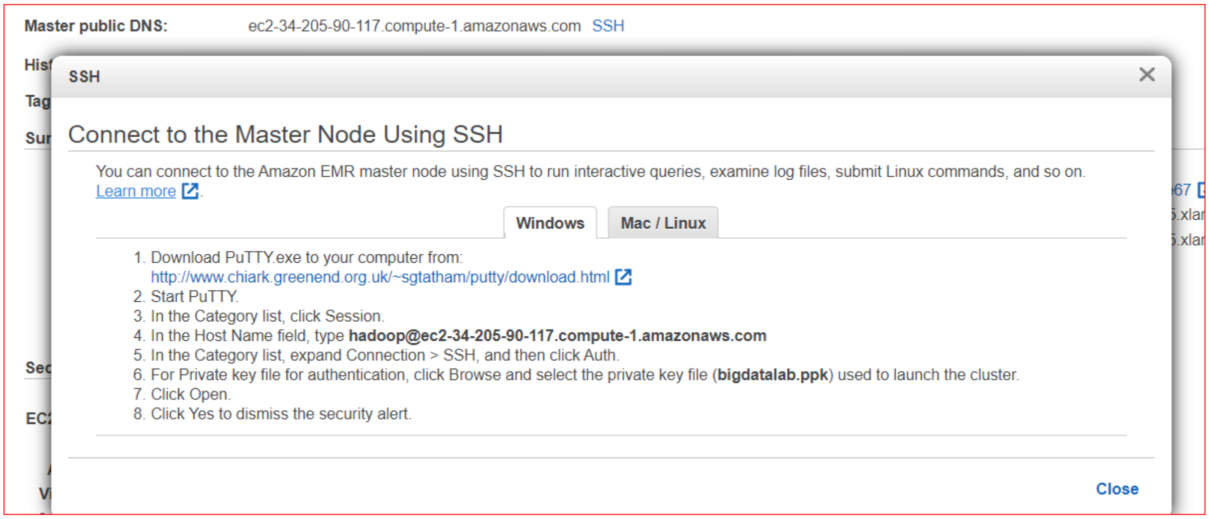
Network and hardware

Availability zone: us-east-1b
Subnet ID: subnet-00396e67
Master: Running 1 m5.xlarge
Core: Running 2 m5.xlarge
Task: --

The cluster status will turn to waiting.

Step 14: For Windows Users use PuttyKeyGen to generate a PPK file out of the PEM file and login using putty.

Click on the SSH link as shown in the picture:



Master public DNS: ec2-34-205-90-117.compute-1.amazonaws.com SSH

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.

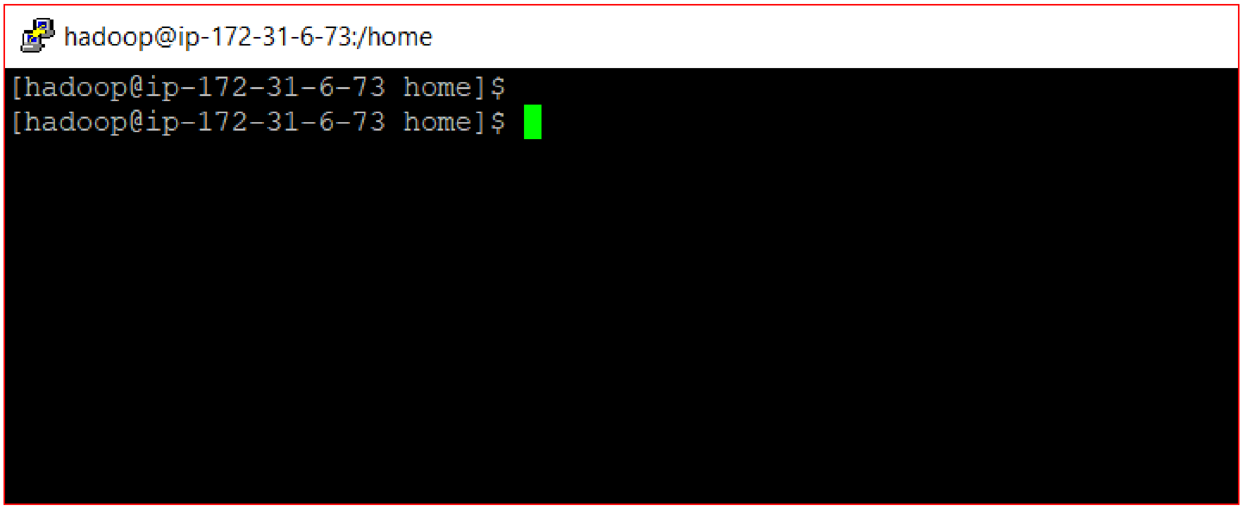
1. Download PuTTY.exe to your computer from: http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html
2. Start PuTTY.
3. In the Category list, click Session.
4. In the Host Name field, type **hadoop@ec2-34-205-90-117.compute-1.amazonaws.com**
5. In the Category list, expand Connection > SSH, and then click Auth.
6. For Private key file for authentication, click Browse and select the private key file (**bigdatalab.ppk**) used to launch the cluster.
7. Click Open.
8. Click Yes to dismiss the security alert.

Go to terminal here using Linux. Go to SSH via pem key and DNS name to connect to master node.

<u>For MAC and LInux users:</u>

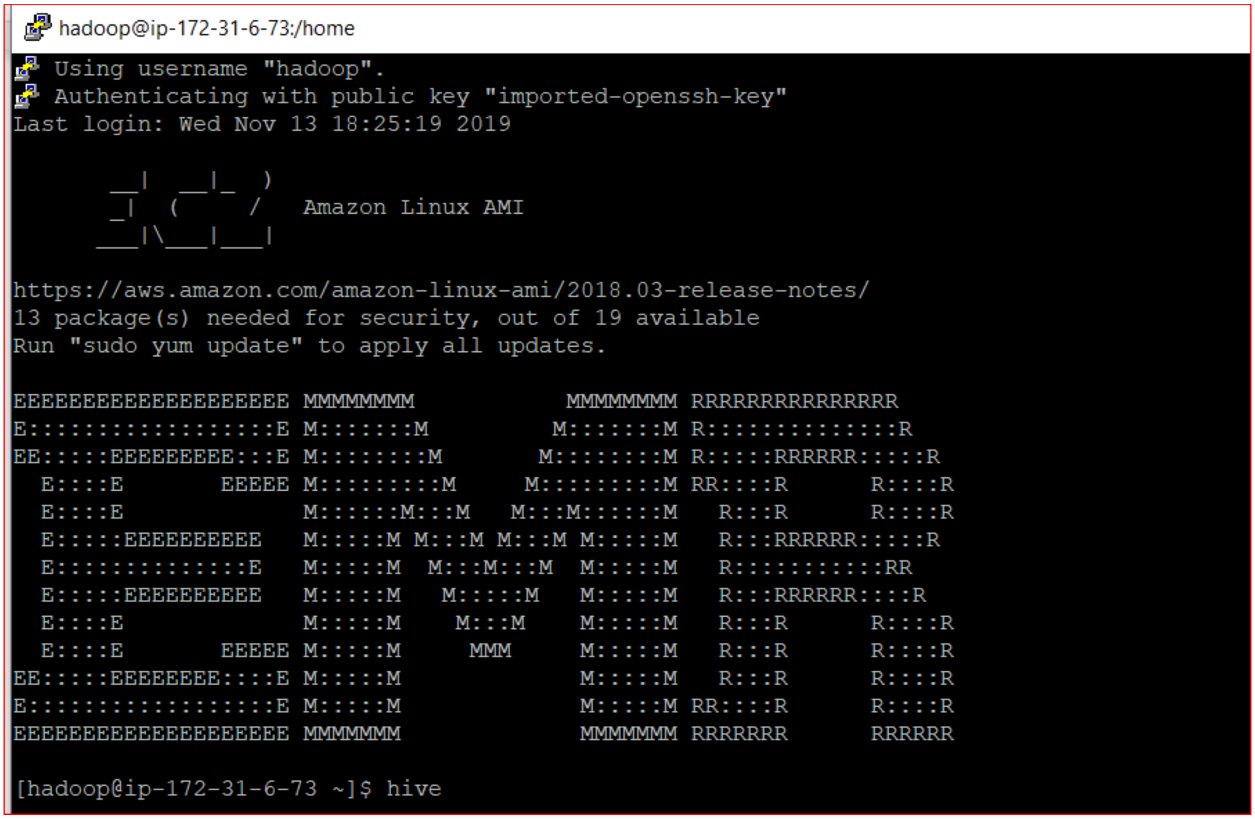
“chmod 400 new.pem “

“ssh -i new.pem hadoop@<DNS name>”



```
[hadoop@ip-172-31-6-73 home]$
[hadoop@ip-172-31-6-73 home]$
```

Step 15: Now you are connected to master node. For using Hive, type ‘hive’. Now in Hive, create a table which for which the storage is S3. Create a folder called “bigdatatest” in the S3 bucket.



hadoop@ip-172-31-6-73:/home

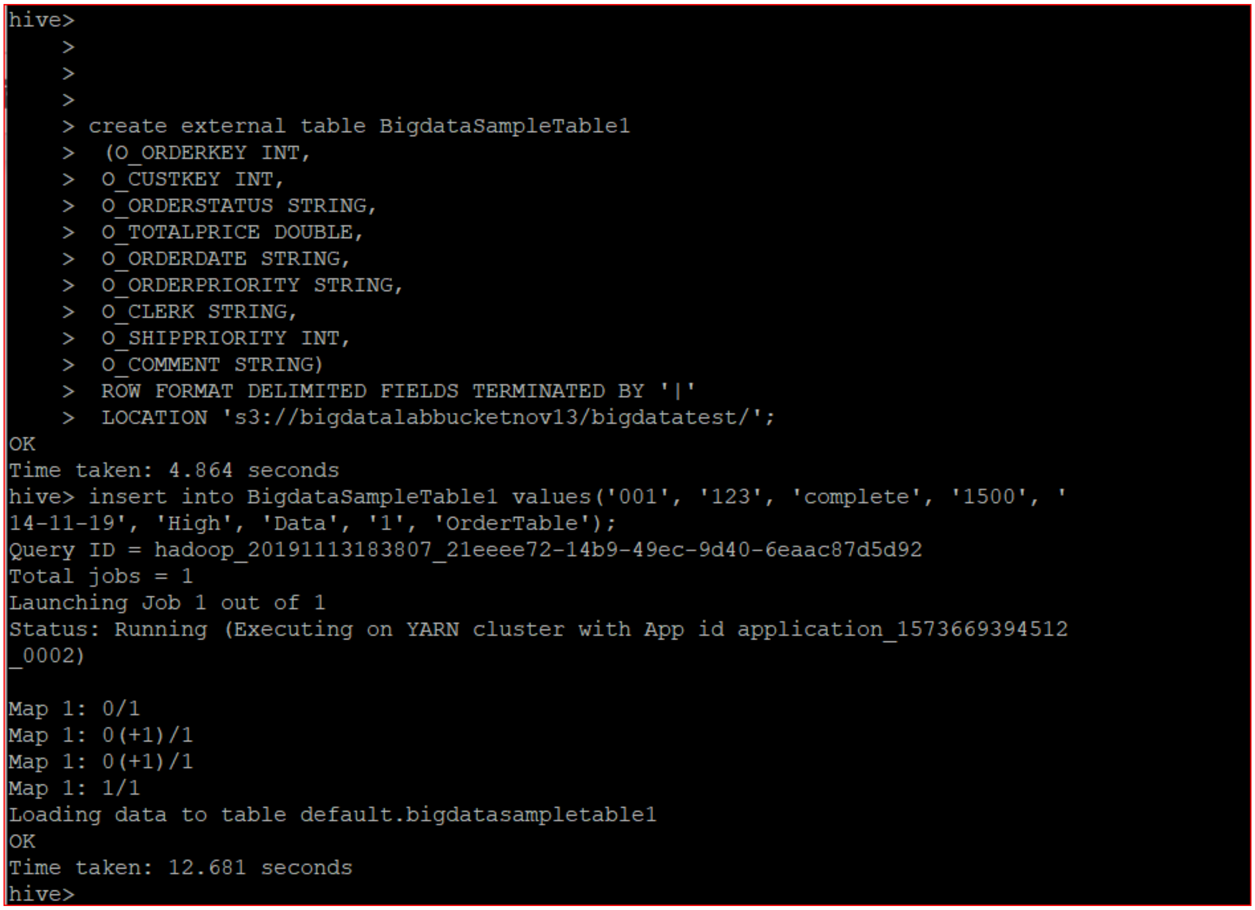
```
 Using username "hadoop".
 Authenticating with public key "imported-openssh-key"
Last login: Wed Nov 13 18:25:19 2019

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
13 package(s) needed for security, out of 19 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   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-172-31-6-73 ~]$ hive
```



```
hive>
    >
    >
    >
    > create external table BigdataSampleTable1
    >  (O_ORDERKEY INT,
    >  O_CUSTKEY INT,
    >  O_ORDERSTATUS STRING,
    >  O_TOTALPRICE DOUBLE,
    >  O_ORDERDATE STRING,
    >  O_ORDERPRIORITY STRING,
    >  O_CLERK STRING,
    >  O_SHIPPRIORITY INT,
    >  O_COMMENT STRING)
    >  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'
    >  LOCATION 's3://bigdatalabbucketnov13/bigdatatest/';
OK
Time taken: 4.864 seconds
hive> insert into BigdataSampleTable1 values('001', '123', 'complete', '1500', '
14-11-19', 'High', 'Data', '1', 'OrderTable');
Query ID = hadoop_20191113183807_21eeee72-14b9-49ec-9d40-6eaac87d5d92
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1573669394512
_0002)

Map 1: 0/1
Map 1: 0(+1)/1
Map 1: 0(+1)/1
Map 1: 1/1
Loading data to table default.bigdatasampletable1
OK
Time taken: 12.681 seconds
hive>
```

*****

create external table BigdataSampleTable1

 (O_ORDERKEY INT,

 O_CUSTKEY INT,

 O_ORDERSTATUS STRING,

 O_TOTALPRICE DOUBLE,

 O_ORDERDATE STRING,

 O_ORDERPRIORITY STRING,

 O_CLERK STRING,

 O_SHIPPRIORITY INT,

 O_COMMENT STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'

LOCATION 's3://bigdatalabbucketnov13/bigdatatest/';

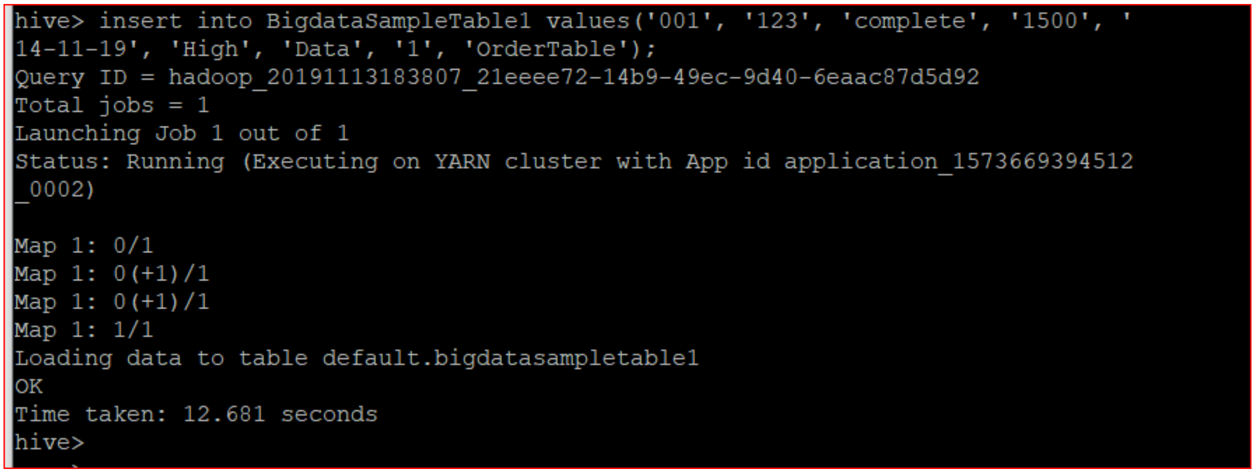
****

Load the Data to the first table : BigdataSampleTable1

****

insert into BigdataSampleTable1 values('001', '123', 'complete', '1500', '14-11-19', 'High', 'Data', '1', 'OrderTable');

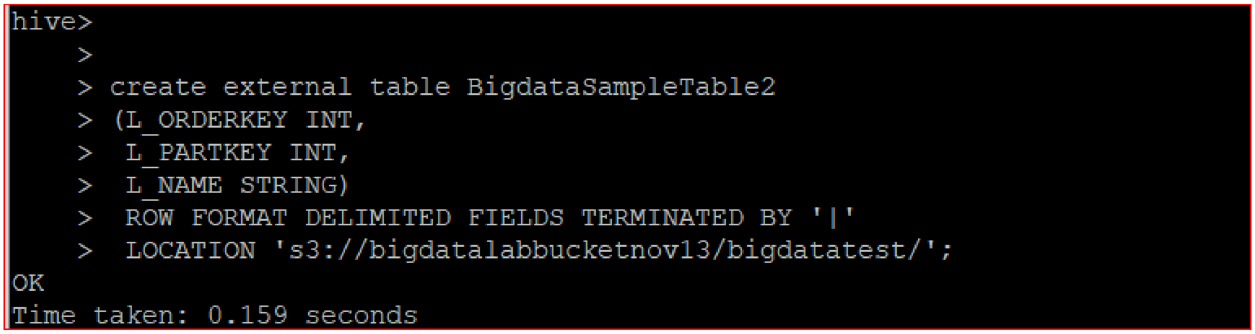
*****



```
hive> insert into BigdataSampleTable1 values('001', '123', 'complete', '1500', '
14-11-19', 'High', 'Data', '1', 'OrderTable');
Query ID = hadoop_20191113183807_21eeee72-14b9-49ec-9d40-6eaac87d5d92
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1573669394512
_0002)

Map 1: 0/1
Map 1: 0(+1)/1
Map 1: 0(+1)/1
Map 1: 1/1
Loading data to table default.bigdatasampletable1
OK
Time taken: 12.681 seconds
hive>
```

Step 16: Create another table in the same way.



```
hive>
    >
    > create external table BigdataSampleTable2
    > (L_ORDERKEY INT,
    >  L_PARTKEY INT,
    >  L_NAME STRING)
    >  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'
    >  LOCATION 's3://bigdatalabbucketnov13/bigdatatest/';
OK
Time taken: 0.159 seconds
```

*****

create external table BigdataSampleTable2

(L_ORDERKEY INT,

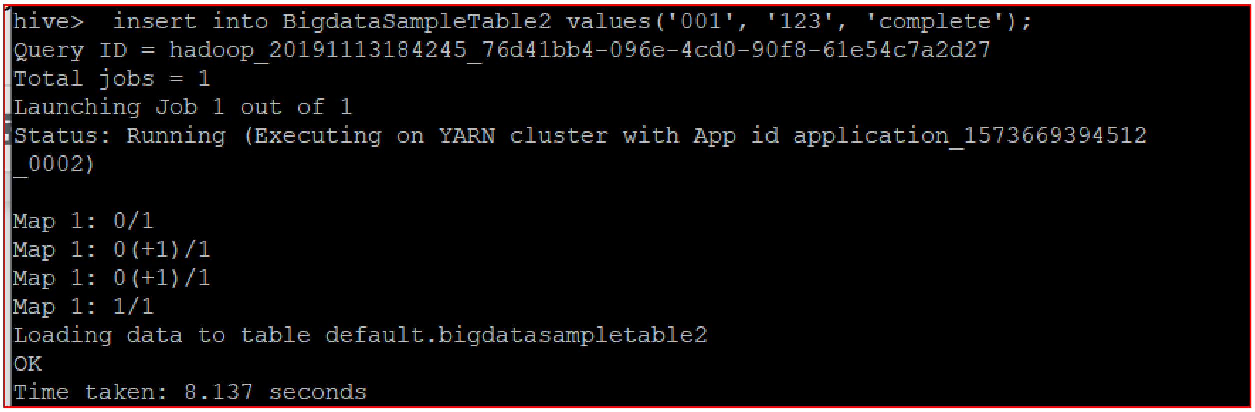
L_PARTKEY INT,

L_NAME STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'

LOCATION 's3://bigdatalabbucketnov13/bigdatatest/';

*****



```
hive>  insert into BigdataSampleTable2 values('001', '123', 'complete');
Query ID = hadoop_20191113184245_76d41bb4-096e-4cd0-90f8-61e54c7a2d27
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1573669394512
_0002)

Map 1: 0/1
Map 1: 0(+1)/1
Map 1: 0(+1)/1
Map 1: 1/1
Loading data to table default.bigdatasampletable2
OK
Time taken: 8.137 seconds
```

Load the Data to the first table : BigdataSampleTable2

****

insert into BigdataSampleTable2 values('001', '890', 'complete');

****

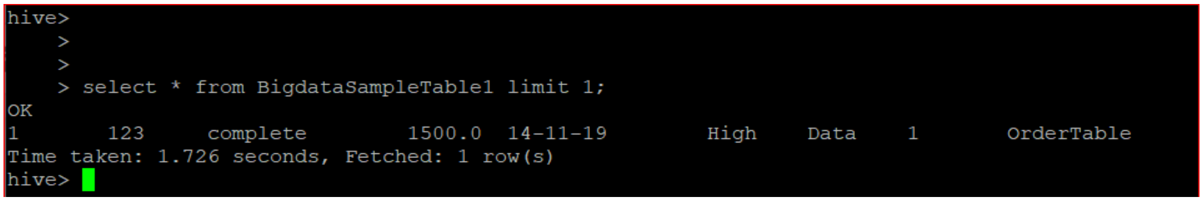
Step 17: Now both tables have been created. Run query on tables

***

select * from BigdataSampleTable1 limit 1;

****

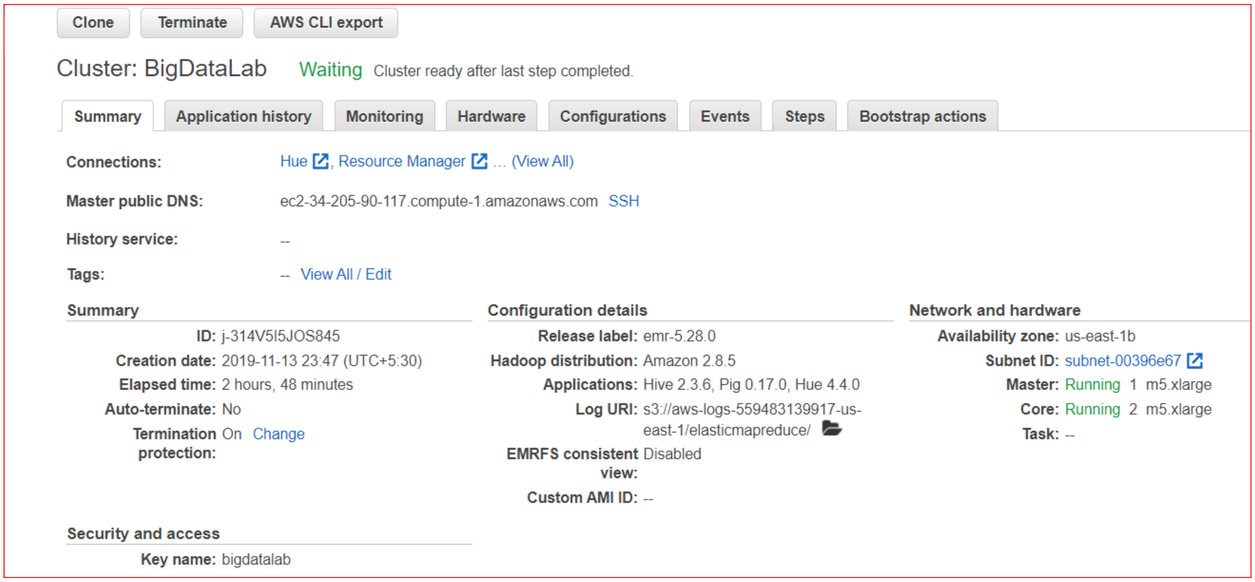
Now you will get the output with the time it takes to run the query.



```
hive>
    >
    >
    > select * from BigdataSampleTable1 limit 1;
OK
1       123     complete        1500.0  14-11-19        High    Data    1       OrderTable
Time taken: 1.726 seconds, Fetched: 1 row(s)
hive> 
```

Additional tasks:

1. You can check the EMR logs in the s3 bucket and path mentioned in the LogURI.



Cluster: BigDataLab Waiting Cluster ready after last step completed.

Connections: Hue, Resource Manager ... (View All)

Master public DNS: ec2-34-205-90-117.compute-1.amazonaws.com SSH

History service: --

Tags: -- View All / Edit

Summary

ID: j-314V5I5JOS845
Creation date: 2019-11-13 23:47 (UTC+5:30)
Elapsed time: 2 hours, 48 minutes
Auto-terminate: No
Termination protection: On Change

Configuration details

Release label: emr-5.28.0
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, Pig 0.17.0, Hue 4.4.0
Log URI: s3://aws-logs-559483139917-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

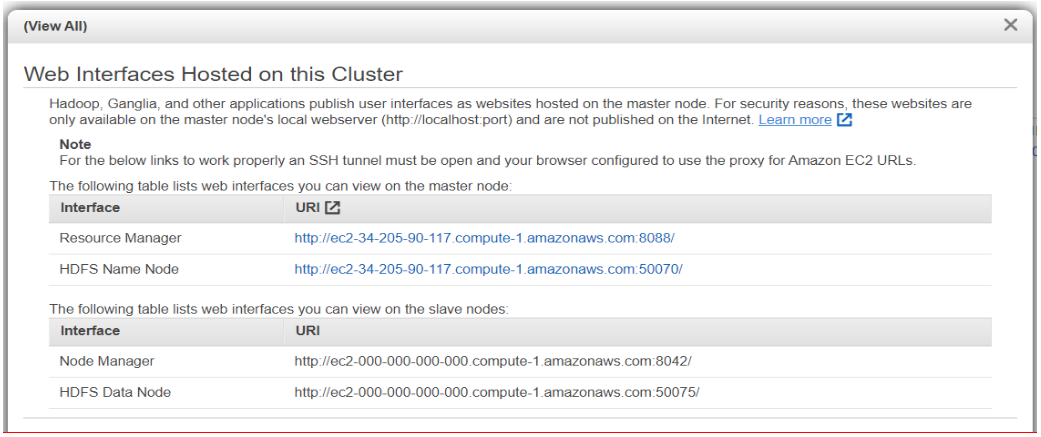
Network and hardware

Availability zone: us-east-1b
Subnet ID: subnet-00396e67
Master: Running 1 m5.xlarge
Core: Running 2 m5.xlarge
Task: --

Security and access

Key name: bigdatalab

2. Check the Web Interfaces Hosted on this Cluster
   a. Hadoop, Ganglia, and other applications publish user interfaces as websites hosted on the master node



(View All)

Web Interfaces Hosted on this Cluster

Hadoop, Ganglia, and other applications publish user interfaces as websites hosted on the master node. For security reasons, these websites are only available on the master node's local webserver (http://localhost:port) and are not published on the Internet. Learn more

Note
For the below links to work properly an SSH tunnel must be open and your browser configured to use the proxy for Amazon EC2 URLs.

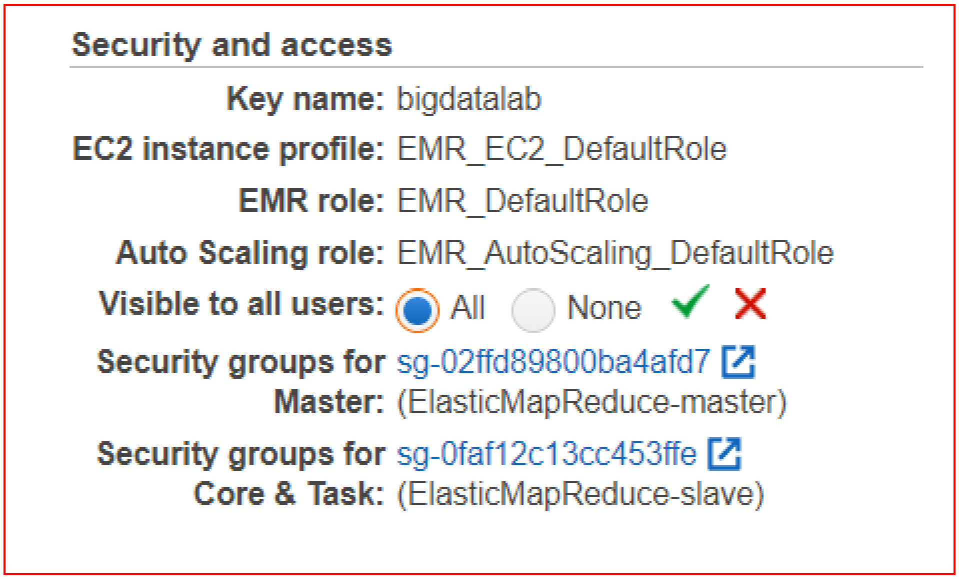
The following table lists web interfaces you can view on the master node:

Interface	URI
Resource Manager	http://ec2-34-205-90-117.compute-1.amazonaws.com:8088/
HDFS Name Node	http://ec2-34-205-90-117.compute-1.amazonaws.com:50070/

The following table lists web interfaces you can view on the slave nodes:

Interface	URI
Node Manager	http://ec2-000-000-000-000.compute-1.amazonaws.com:8042/
HDFS Data Node	http://ec2-000-000-000-000.compute-1.amazonaws.com:50075/

3. You can protect view access to all users by changing the below settings.



Security and access

Key name: bigdatalab

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

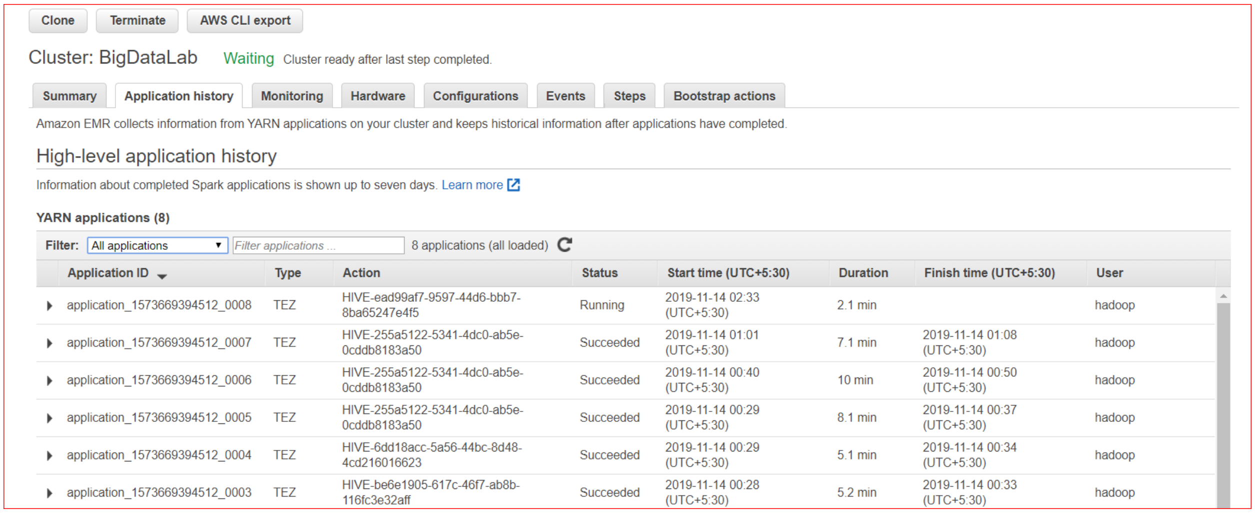
Auto Scaling role: EMR_AutoScaling_DefaultRole

Visible to all users: ● All ○ None ✓ ✗

Security groups for Master: sg-02ffd89800ba4afd7 (ElasticMapReduce-master)

Security groups for Core & Task: sg-0faf12c13cc453ffe (ElasticMapReduce-slave)

4. Any time you login to the applications installed in the EMR, it can be monitored under "Application history".



Cluster: BigDataLab Waiting Cluster ready after last step completed.

Amazon EMR collects information from YARN applications on your cluster and keeps historical information after applications have completed.

High-level application history

Information about completed Spark applications is shown up to seven days. Learn more

YARN applications (8)

Filter: All applications ▾ Filter applications ... 8 applications (all loaded)

Application ID	Type	Action	Status	Start time (UTC+5:30)	Duration	Finish time (UTC+5:30)	User
application_1573669394512_0008	TEZ	HIVE-ead99af7-9597-44d6-bbb7-8ba65247e4f5	Running	2019-11-14 02:33 (UTC+5:30)	2.1 min		hadoop
application_1573669394512_0007	TEZ	HIVE-255a5122-5341-4dc0-ab5e-0cddb8183a50	Succeeded	2019-11-14 01:01 (UTC+5:30)	7.1 min	2019-11-14 01:08 (UTC+5:30)	hadoop
application_1573669394512_0006	TEZ	HIVE-255a5122-5341-4dc0-ab5e-0cddb8183a50	Succeeded	2019-11-14 00:40 (UTC+5:30)	10 min	2019-11-14 00:50 (UTC+5:30)	hadoop
application_1573669394512_0005	TEZ	HIVE-255a5122-5341-4dc0-ab5e-0cddb8183a50	Succeeded	2019-11-14 00:29 (UTC+5:30)	8.1 min	2019-11-14 00:37 (UTC+5:30)	hadoop
application_1573669394512_0004	TEZ	HIVE-6dd18acc-5a56-44bc-8d48-4cd216016623	Succeeded	2019-11-14 00:29 (UTC+5:30)	5.1 min	2019-11-14 00:34 (UTC+5:30)	hadoop
application_1573669394512_0003	TEZ	HIVE-be6e1905-617c-46f7-ab8b-116fc3e32aff	Succeeded	2019-11-14 00:28 (UTC+5:30)	5.2 min	2019-11-14 00:33 (UTC+5:30)	hadoop