DOC – 4 | 21/05/21 – BY ABHINAV SHRESHTH

Work on making a CI-CD (Continuous Integration and Continuous Delivery) the task to submit a spark word count program from S3 to the EMR cluster using AWS code pipeline.

DevOps

A set of practices intended to reduce the time between committing a change to a system and the change being placed into normal production, while ensuring high quality. is the combination of practices and tools designed to increase an organization's ability to deliver applications and services faster than traditional software development processes.

Continuous Integration and Delivery

Step 1 – Split the entire chunk of codes into segments

Step 2 – keep small segments of manageable code

Step 3 – integrate the segmented code, multiple times a day

Step 4 – Adopt a continuous integration methodology to coordinate with your team

we have a source code repository where the developers work continuously submit their pieces of the code repository. Such that a central place where the changes are constantly committed then we have a belt server where everything is compiled reviewed tested integrated and then packets as well finally started test final test goes to the minorities and then it goes to the production environment, where this process the building the Staging takes palace and the committee process it

automating this process becomes very important and if we will do it manually we will suffer a lot

AWS Codepipeline

AWS CodePipeline is a continuous delivery service you can use to model, visualize, and automate the steps required to release your software.

Its features -

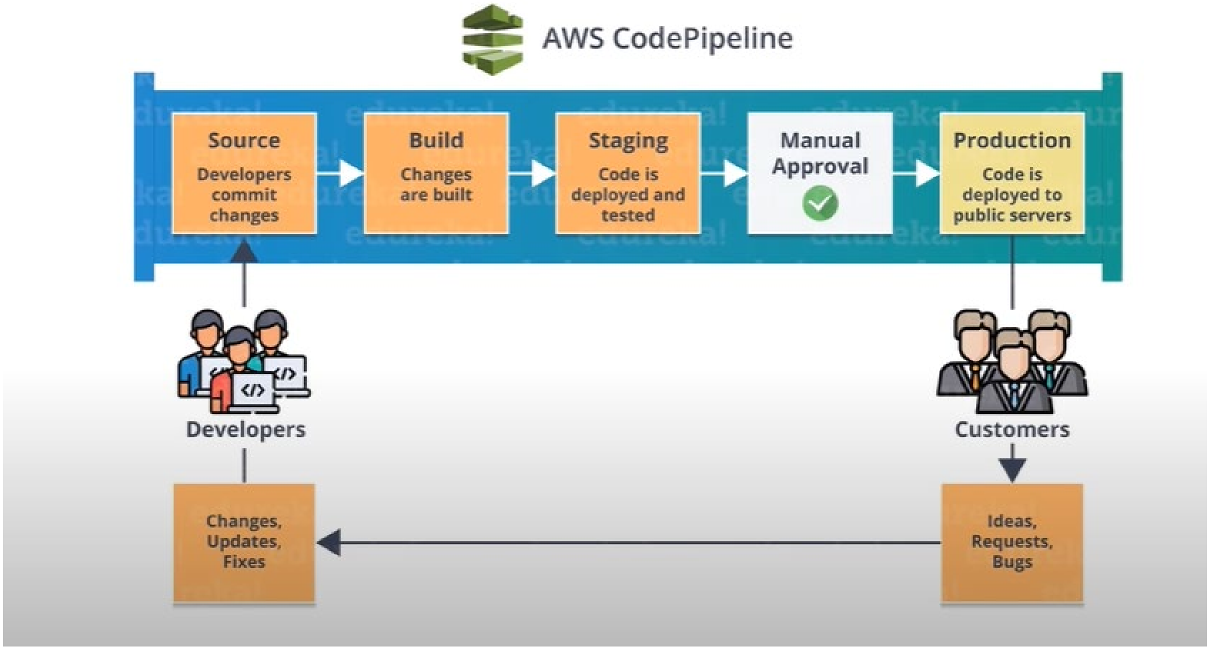
Monitor your processes in real-time

Ensure Consistent Release Process

Speed up delivery while improving quality

View pipeline history details

The architecture of the code pipeline looks like this



AWS CodePipeline

Source: Developers commit changes

Build: Changes are built

Staging: Code is deployed and tested

Manual Approval

Production: Code is deployed to public servers

Developers

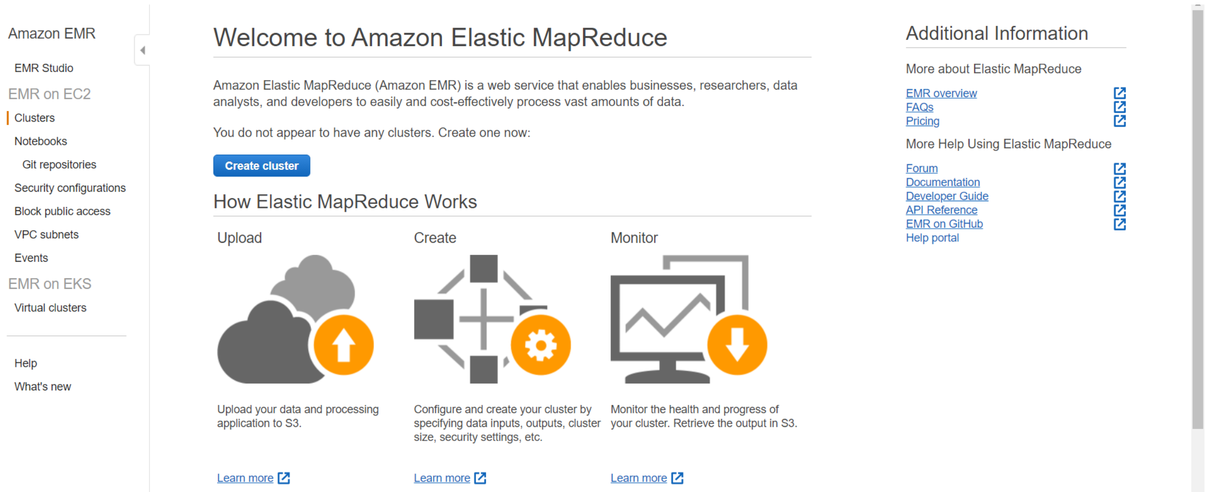
Customers

Changes, Updates, Fixes

Ideas, Requests, Bugs

Further, I have signed up for the Amazon AWS console.

Then I opened AWS EMR and then as I don't know how to proceed further then I YouTube it to know how to proceed



Welcome to Amazon Elastic MapReduce

Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.

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

How Elastic MapReduce Works

Upload: Upload your data and processing application to S3.

Create: Configure and create your cluster by specifying data inputs, outputs, cluster size, security settings, etc.

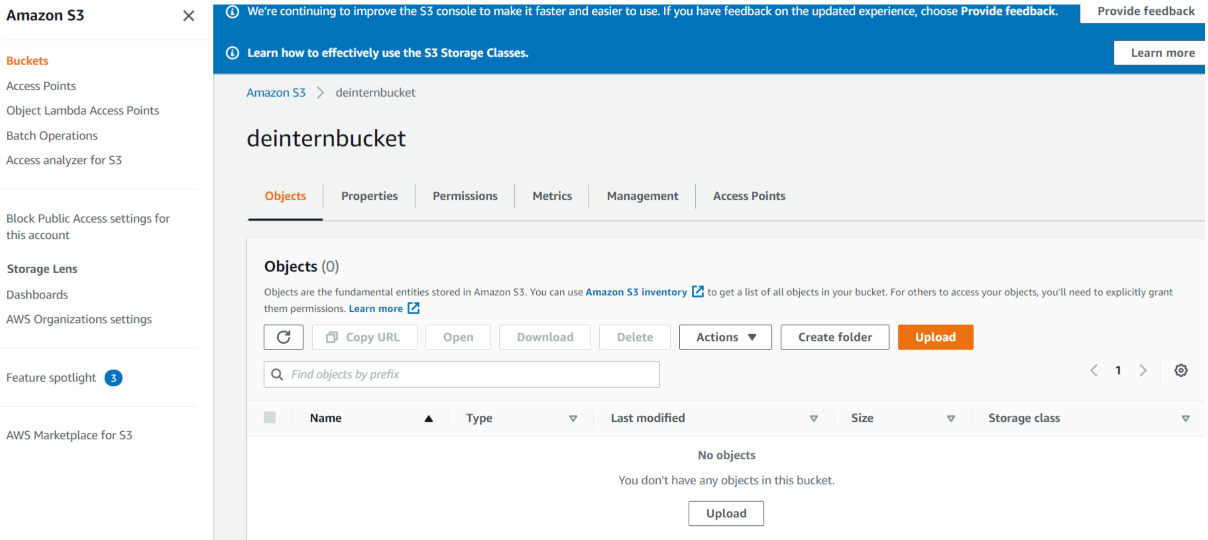
Monitor: Monitor the health and progress of your cluster. Retrieve the output in S3.

Additional Information

More about Elastic MapReduce

More Help Using Elastic MapReduce

On AWS s3 I created bucket there



We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback.**

Learn how to effectively use the S3 Storage Classes.

deinternbucket

Objects (0)

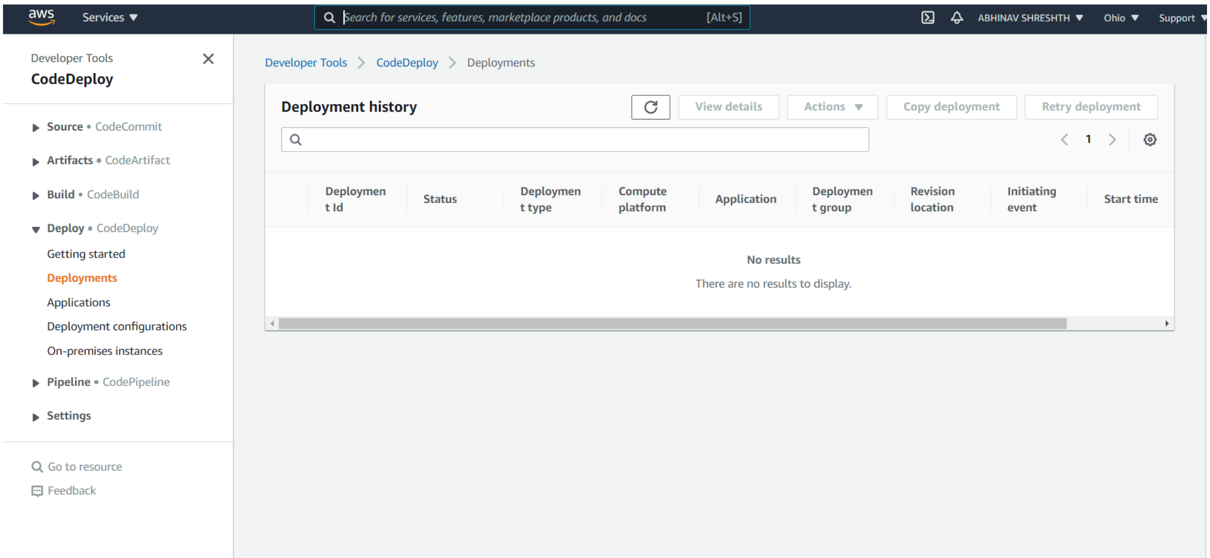
Objects are the fundamental entities stored in Amazon S3. You can use **Amazon S3 inventory** to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. **Learn more**

Name	Type	Last modified	Size	Storage class

No objects

You don't have any objects in this bucket.

further



Deployment history

Deployment Id	Status	Deployment type	Compute platform	Application	Deployment group	Revision location	Initiating event	Start time

No results

There are no results to display.

Amazon EMR

Amazon EMR cluster provides a managed Hadoop framework that makes it easy, fast, and cost-effective to process vast amounts of data across dynamically scalable Amazon EC2 instances.

We can also run other popular distributed frameworks such as Apache **Spark** and HBase in Amazon EMR, and interact with data in other AWS data stores such as Amazon S3 and Amazon DynamoDB.

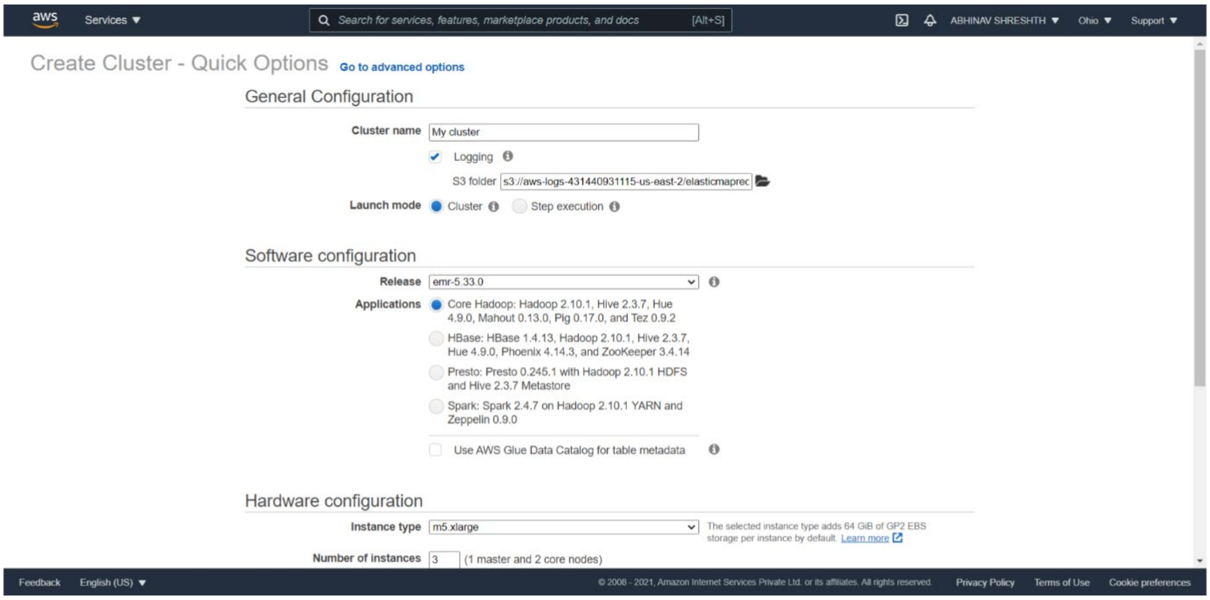
S3(Amazon Simple Storage Service)

We are going to run our Spark application on top of the Hadoop cluster, and we will put the input data source into the S3.

S3 is a distributed storage system and AWS's equivalent to HDFS. We because we want to make sure that

- Our data is coming from some distributed file system that can be accessed by every node on our Spark cluster.

- Our Spark application doesn't assume that our input data sits somewhere on our local disk because that will not scale. By saving our input data source into S3, each spark node deployed on the EMR cluster can read the input data source from S3.


aws
Services ▼
Search for services, features, marketplace products, and docs [Alt+S]
ABHINAV SHRESHTH ▼
Ohio ▼
Support ▼
Go to advanced options
Learn more
Feedback
English (US) ▼
© 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Cookie preferences

Create Cluster - Quick Options

General Configuration

Cluster name: My cluster

☑ Logging

S3 folder: s3://aws-logs-431440931115-us-east-2/elasticmaprec

Launch mode: ● Cluster ○ Step execution

Software configuration

Release: emr-5.33.0

Applications:
- ● Core Hadoop: Hadoop 2.10.1, Hive 2.3.7, Hue 4.9.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
- ○ HBase: HBase 1.4.13, Hadoop 2.10.1, Hive 2.3.7, Hue 4.9.0, Phoenix 4.14.3, and ZooKeeper 3.4.14
- ○ Presto: Presto 0.245.1 with Hadoop 2.10.1 HDFS and Hive 2.3.7 Metastore
- ○ Spark: Spark 2.4.7 on Hadoop 2.10.1 YARN and Zeppelin 0.9.0

□ Use AWS Glue Data Catalog for table metadata

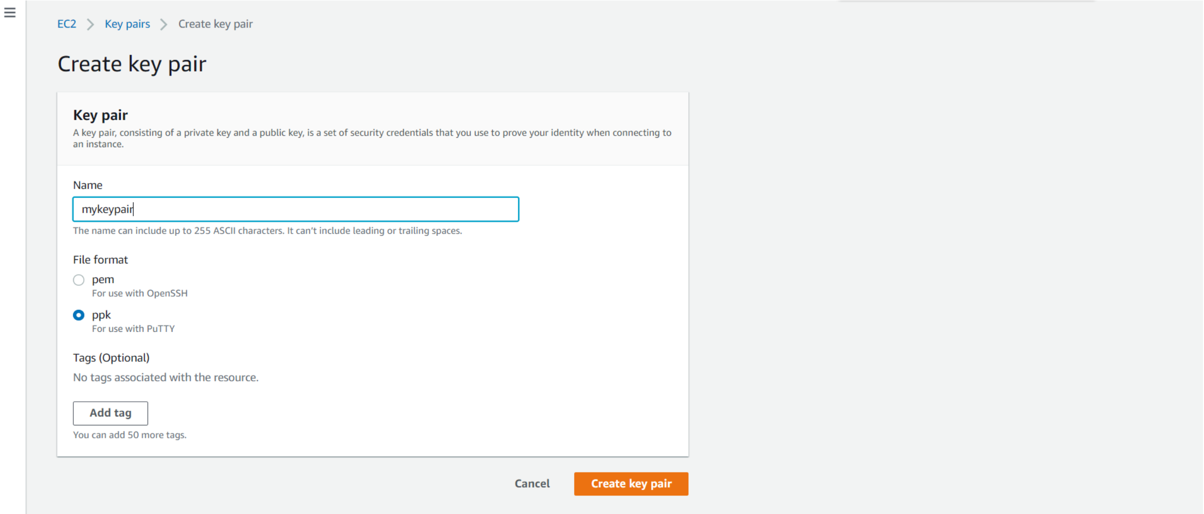
Hardware configuration

Instance type: m5.xlarge — The selected instance type adds 64 GiB of GP2 EBS storage per instance by default.

Number of instances: 3 (1 master and 2 core nodes)

To create an Amazon EC2 key pair:

1. Go to the Amazon EC2 console
2. In the Navigation pane, click Key Pairs



Create key pair

Key pair

A key pair, consisting of a private key and a public key, is a set of security credentials that you use to prove your identity when connecting to an instance.

Name: mykeypair

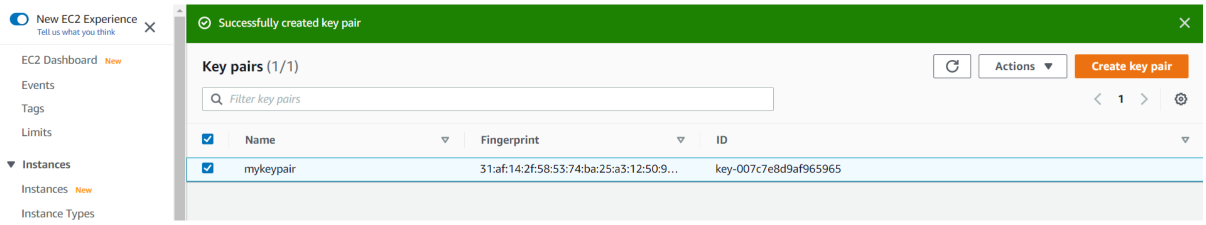
The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.

File format
- ○ pem — For use with OpenSSH
- ● ppk — For use with PuTTY

Tags (Optional)

No tags associated with the resource.

You can add 50 more tags.

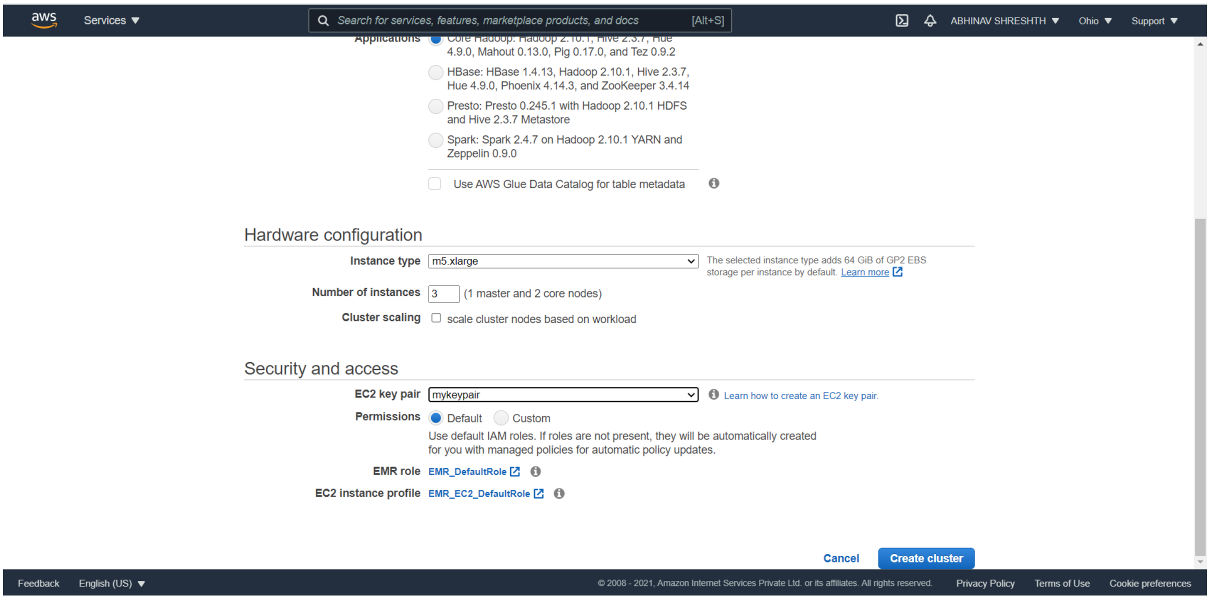


Successfully created key pair

Key pairs (1/1)

☑	Name	Fingerprint	ID
☑	mykeypair	31:af:14:2f:58:53:74:ba:25:a3:12:50:9...	key-007c7e8d9af965965

3. On the Key Pairs page, click Create Key Pair
4. In the Create Key Pair dialogue box, enter a name for your key pair, such as mykeypair
5. Click Create
6. Save the resulting PEM file in a safe location


aws
Services ▼
Search for services, features, marketplace products, and docs [Alt+S]
ABHINAV SHRESHTH ▼
Ohio ▼
Support ▼

Applications ● Core Hadoop: Hadoop 2.10.1, Hive 2.3.7, Hue 4.9.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2

○ HBase: HBase 1.4.13, Hadoop 2.10.1, Hive 2.3.7, Hue 4.9.0, Phoenix 4.14.3, and ZooKeeper 3.4.14

○ Presto: Presto 0.245.1 with Hadoop 2.10.1 HDFS and Hive 2.3.7 Metastore

○ Spark: Spark 2.4.7 on Hadoop 2.10.1 YARN and Zeppelin 0.9.0

□ Use AWS Glue Data Catalog for table metadata

Hardware configuration

Instance type: m5.xlarge — The selected instance type adds 64 GiB of GP2 EBS storage per instance by default. Learn more

Number of instances: 3 (1 master and 2 core nodes)

Cluster scaling: □ scale cluster nodes based on workload

Security and access

EC2 key pair: mykeypair — Learn how to create an EC2 key pair.

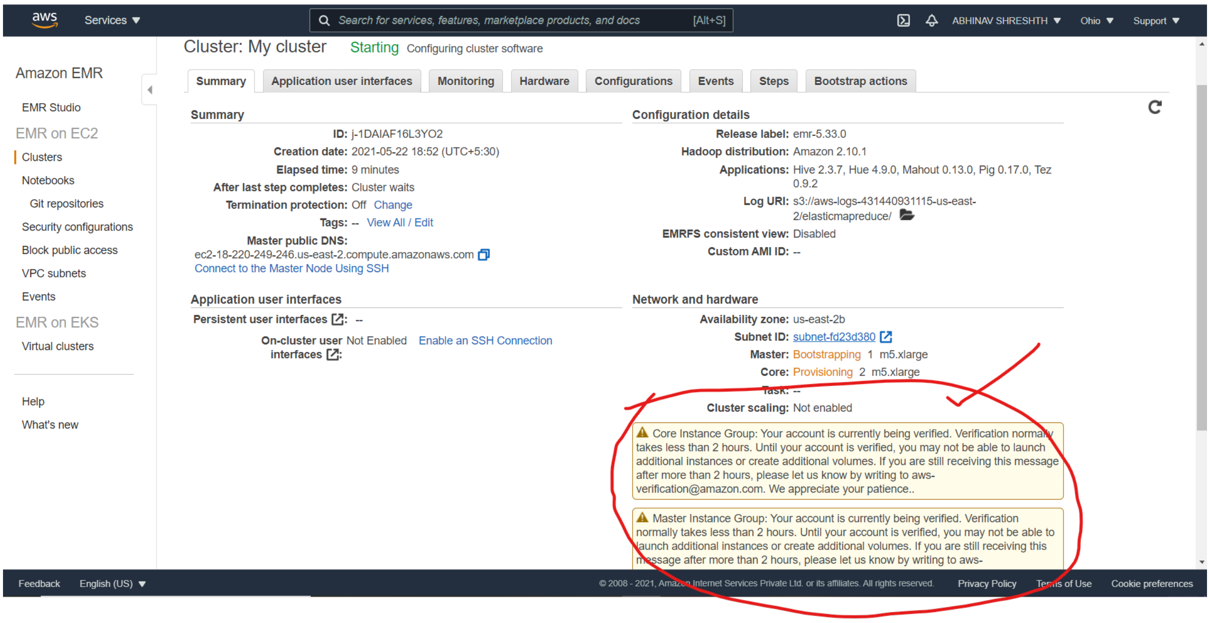
Permissions: ● Default ○ Custom

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

Cancel
Create cluster
Feedback
English (US) ▼
© 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Cookie preferences



Cluster: My cluster Starting Configuring cluster software

Summary

ID: j-1DAIAF16L3YO2
Creation date: 2021-05-22 18:52 (UTC+5:30)
Elapsed time: 9 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: ec2-18-220-249-246.us-east-2.compute.amazonaws.com
Connect to the Master Node Using SSH

Application user interfaces

Persistent user interfaces: --
On-cluster user interfaces: Not Enabled Enable an SSH Connection

Configuration details

Release label: emr-5.33.0
Hadoop distribution: Amazon 2.10.1
Applications: Hive 2.3.7, Hue 4.9.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2
Log URI: s3://aws-logs-431440931115-us-east-2/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

Availability zone: us-east-2b
Subnet ID: subnet-fd23d380
Master: Bootstrapping 1 m5.xlarge
Core: Provisioning 2 m5.xlarge
Task: --
Cluster scaling: Not enabled

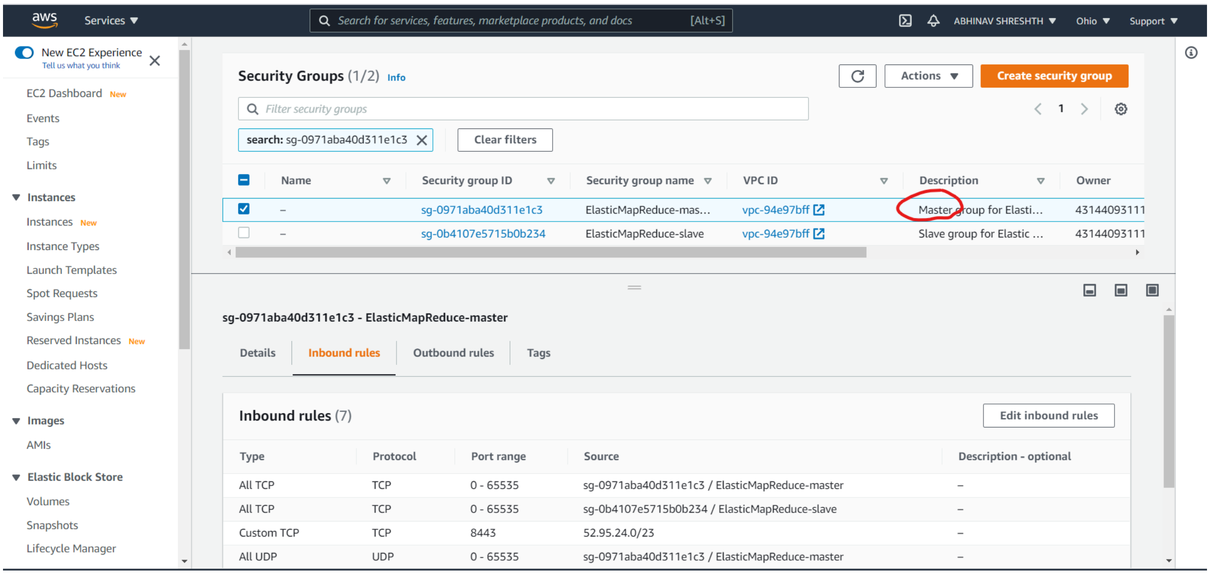
⚠ Core Instance Group: Your account is currently being verified. Verification normally takes less than 2 hours. Until your account is verified, you may not be able to launch additional instances or create additional volumes. If you are still receiving this message after more than 2 hours, please let us know by writing to aws-verification@amazon.com. We appreciate your patience..

⚠ Master Instance Group: Your account is currently being verified. Verification normally takes less than 2 hours. Until your account is verified, you may not be able to launch additional instances or create additional volumes. If you are still receiving this message after more than 2 hours, please let us know by writing to aws-

Feedback
English (US) ▼
© 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Cookie preferences

After the activation of the account under **Security and access -> Security groups for Master**

Click on it then this window will open



Security Groups (1/2) Info

search: sg-0971aba40d311e1c3

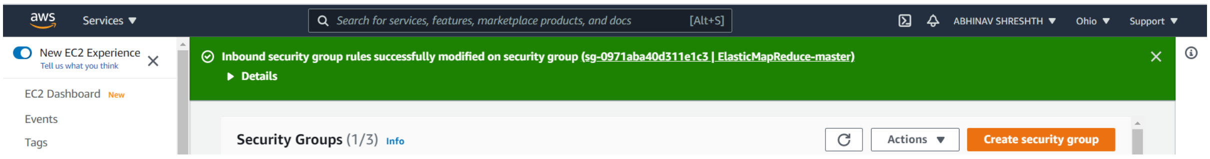
	Name	Security group ID	Security group name	VPC ID	Description	Owner
☑	–	sg-0971aba40d311e1c3	ElasticMapReduce-mas…	vpc-94e97bff	Master group for Elasti…	43144093111
□	–	sg-0b4107e5715b0b234	ElasticMapReduce-slave	vpc-94e97bff	Slave group for Elastic …	43144093111

sg-0971aba40d311e1c3 - ElasticMapReduce-master

Inbound rules (7)

Type	Protocol	Port range	Source	Description - optional
All TCP	TCP	0 - 65535	sg-0971aba40d311e1c3 / ElasticMapReduce-master	–
All TCP	TCP	0 - 65535	sg-0b4107e5715b0b234 / ElasticMapReduce-slave	–
Custom TCP	TCP	8443	52.95.24.0/23	–
All UDP	UDP	0 - 65535	sg-0971aba40d311e1c3 / ElasticMapReduce-master	–

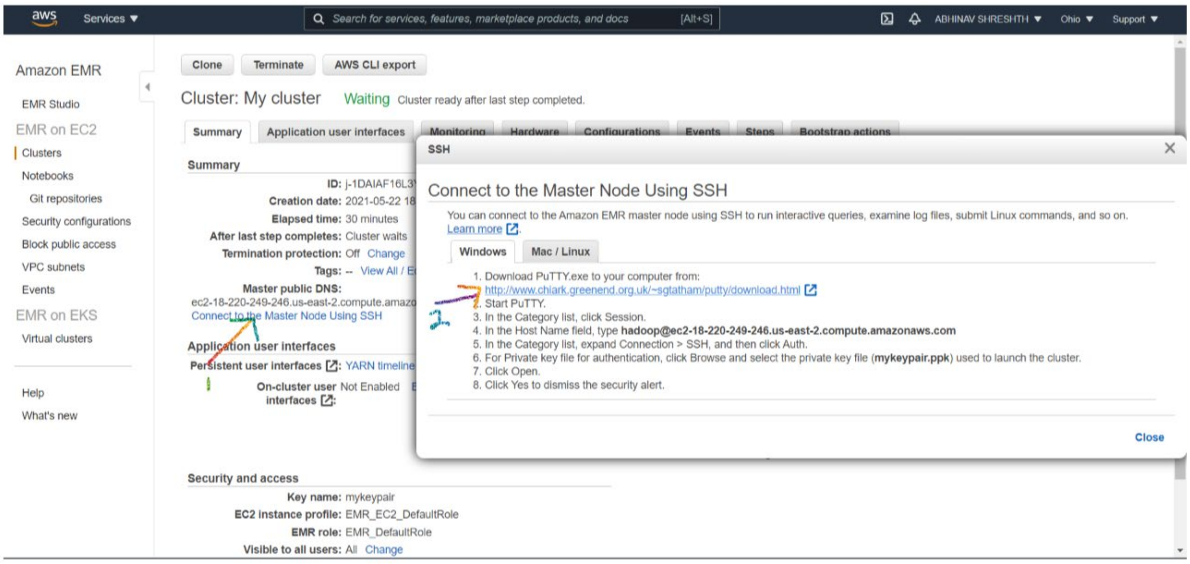
There create a new rule of an inbound connection



Inbound security group rules successfully modified on security group (sg-0971aba40d311e1c3 | ElasticMapReduce-master)

Security Groups (1/3) Info

Next step – download putty application



Cluster: My cluster Waiting Cluster ready after last step completed.

Summary

ID: j-1DAIAF16L3
Creation date: 2021-05-22 18
Elapsed time: 30 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / E
Master public DNS:
ec2-18-220-249-246.us-east-2.compute.amazo
Connect to the Master Node Using SSH

Application user interfaces

Persistent user interfaces: YARN timeline
On-cluster user interfaces: Not Enabled

SSH

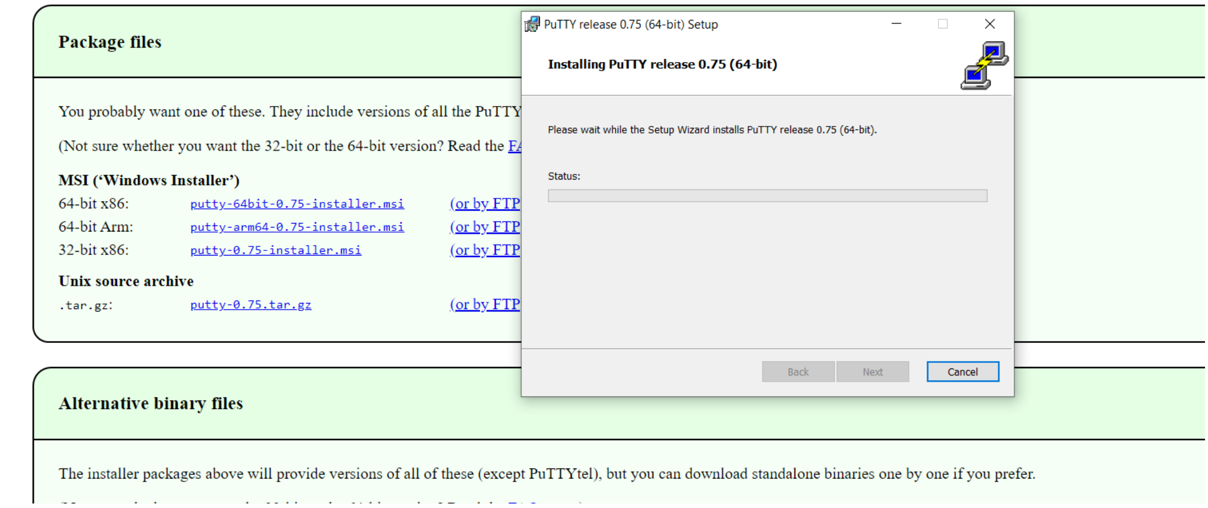
Connect to the Master Node Using SSH

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@ec2-18-220-249-246.us-east-2.compute.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 (**mykeypair.ppk**) used to launch the cluster.
7. Click Open.
8. Click Yes to dismiss the security alert.

Security and access

Key name: mykeypair
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change



Package files

You probably want one of these. They include versions of all the PuTTY

(Not sure whether you want the 32-bit or the 64-bit version? Read the FA

MSI ('Windows Installer')

64-bit x86:	putty-64bit-0.75-installer.msi	(or by FTP
64-bit Arm:	putty-arm64-0.75-installer.msi	(or by FTP
32-bit x86:	putty-0.75-installer.msi	(or by FTP

Unix source archive

.tar.gz:	putty-0.75.tar.gz	(or by FTP

Alternative binary files

The installer packages above will provide versions of all of these (except PuTTYtel), but you can download standalone binaries one by one if you prefer.

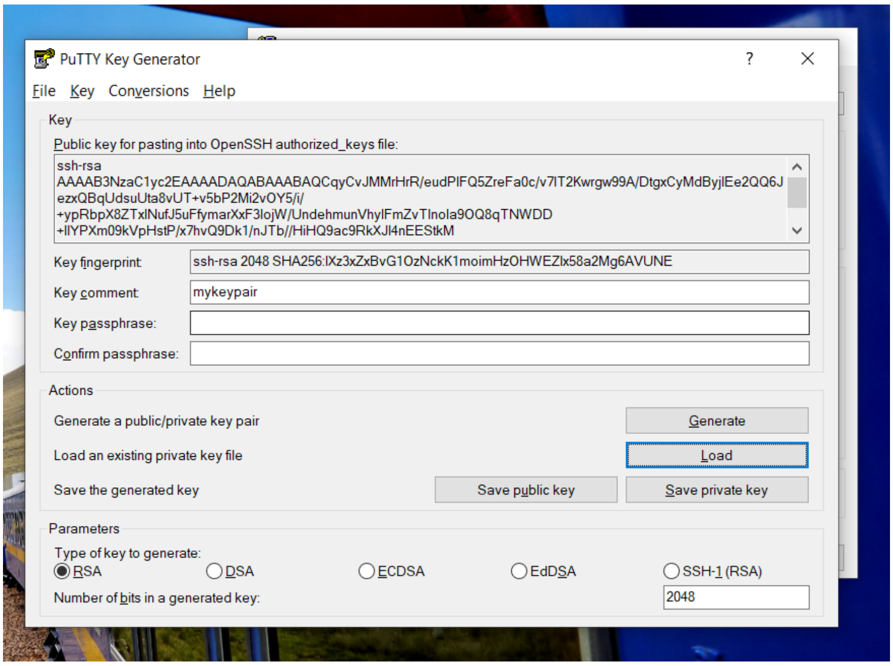
PuTTY release 0.75 (64-bit) Setup

Installing PuTTY release 0.75 (64-bit)

Please wait while the Setup Wizard installs PuTTY release 0.75 (64-bit).

Status:

Open the Putty key generator then load the key you have downloaded before from aws



PuTTY Key Generator

Key

Public key for pasting into OpenSSH authorized_keys file:

```
ssh-rsa
AAAAB3NzaC1yc2EAAAADAQABAAABAQCqyCvJMMrHrR/eudPIFQ5ZreFa0c/v7IT2Kwrgw99A/DtgxCyMdByjIEe2QQ6J
ezxQBqUdsuUta8vUT+v5bP2Mi2vOY5/i/
+ypRbpX8ZTxlNufJ5uFfymarXxF3lojW/UndehmunVhylFmZvTlnola9OQ8qTNWDD
+lIYPXm09kVpHstP/x7hvQ9Dk1/nJTb//HiHQ9ac9RkXJl4nEEStkM
```

Key fingerprint: ssh-rsa 2048 SHA256:lXz3xZxBvG1OzNckK1moimHzOHWEZlx58a2Mg6AVUNE

Key comment: mykeypair

Key passphrase:

Confirm passphrase:

Actions

Generate a public/private key pair

Load an existing private key file

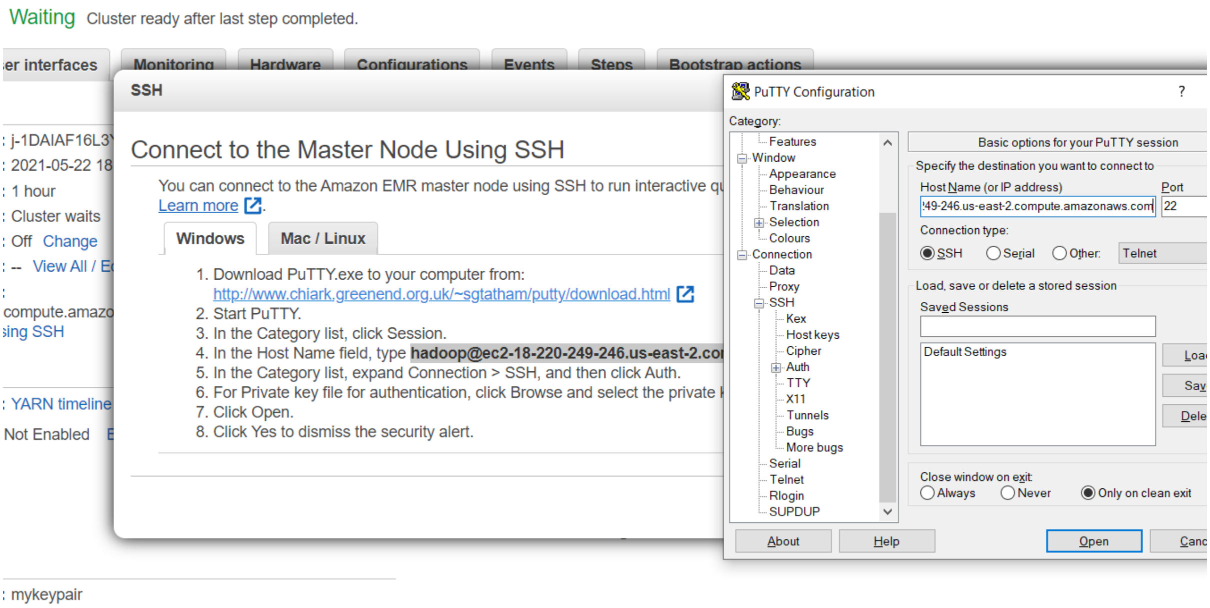
Save the generated key

Parameters

Type of key to generate:

● RSA ○ DSA ○ ECDSA ○ EdDSA ○ SSH-1 (RSA)

Number of bits in a generated key: 2048



Waiting Cluster ready after last step completed.

: j-1DAIAF16L3
: 2021-05-22 18
: 1 hour
: Cluster waits
: Off Change
: -- View All / E
:
compute.amazo
sing SSH
: YARN timeline
Not Enabled E

SSH

Connect to the Master Node Using SSH

You can connect to the Amazon EMR master node using SSH to run interactive qu
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@ec2-18-220-249-246.us-east-2.co**
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 k
7. Click Open.
8. Click Yes to dismiss the security alert.

: mykeypair

PuTTY Configuration ?

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address): '49-246.us-east-2.compute.amazonaws.com

Port: 22

Connection type: ● SSH ○ Serial ○ Other: Telnet

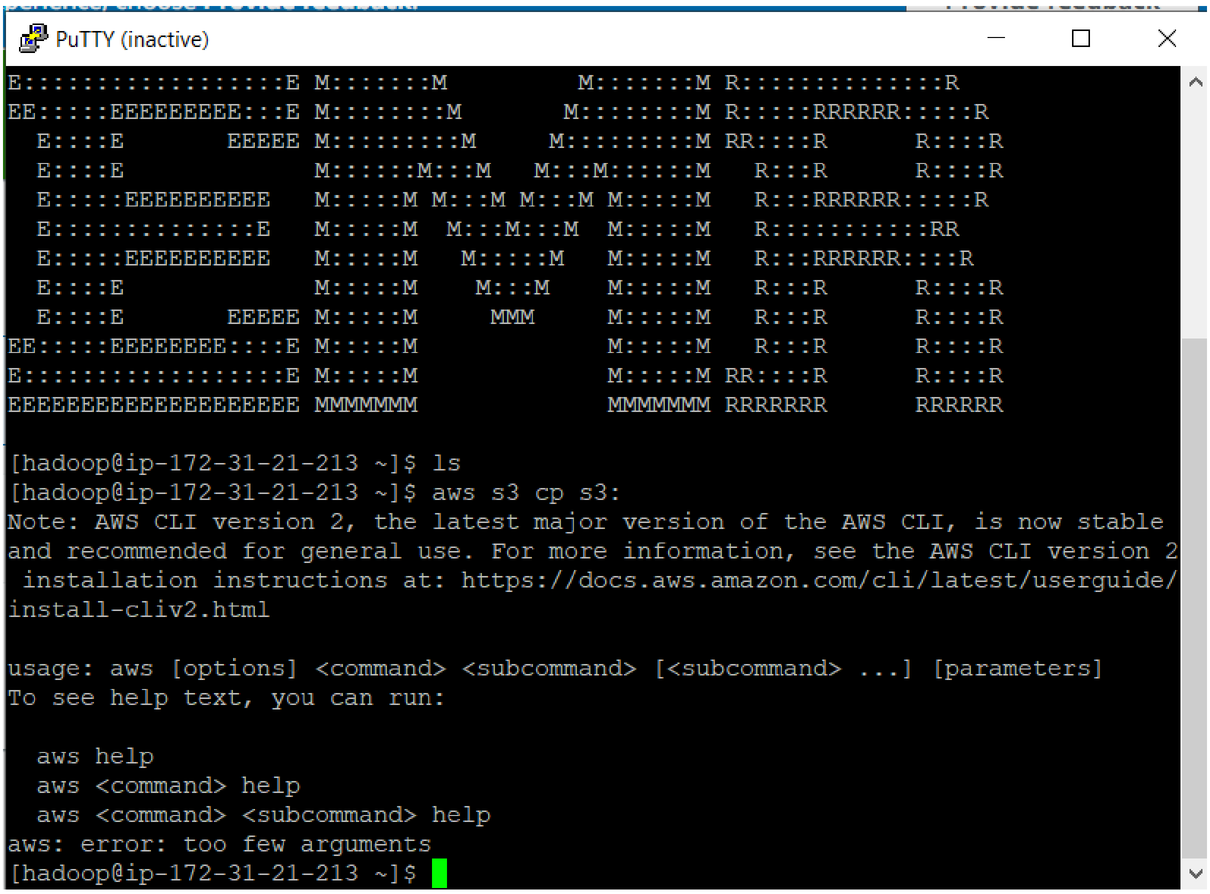
Load, save or delete a stored session

Saved Sessions

Default Settings

Close window on exit: ○ Always ○ Never ● Only on clean exit

Now I am inside EMR cluster



PuTTY (inactive)

```
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-172-31-21-213 ~]$ ls
[hadoop@ip-172-31-21-213 ~]$ aws s3 cp s3:
Note: AWS CLI version 2, the latest major version of the AWS CLI, is now stable
and recommended for general use. For more information, see the AWS CLI version 2
 installation instructions at: https://docs.aws.amazon.com/cli/latest/userguide/
install-cliv2.html

usage: aws [options] <command> <subcommand> [<subcommand> ...] [parameters]
To see help text, you can run:

  aws help
  aws <command> help
  aws <command> <subcommand> help
aws: error: too few arguments
[hadoop@ip-172-31-21-213 ~]$
```