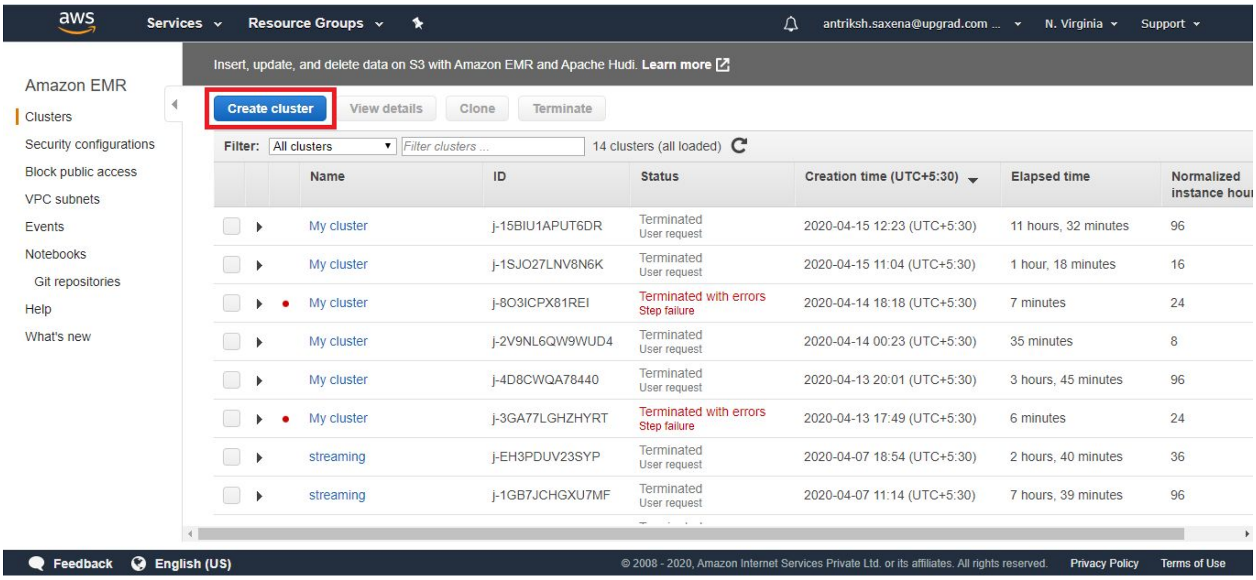
upGrad

Spark Setup on EMR Cluster

In this document, you will learn how to launch an EMR cluster with Spark installed on it. You will also see how to run a Jypyter notebook on the EMR cluster.

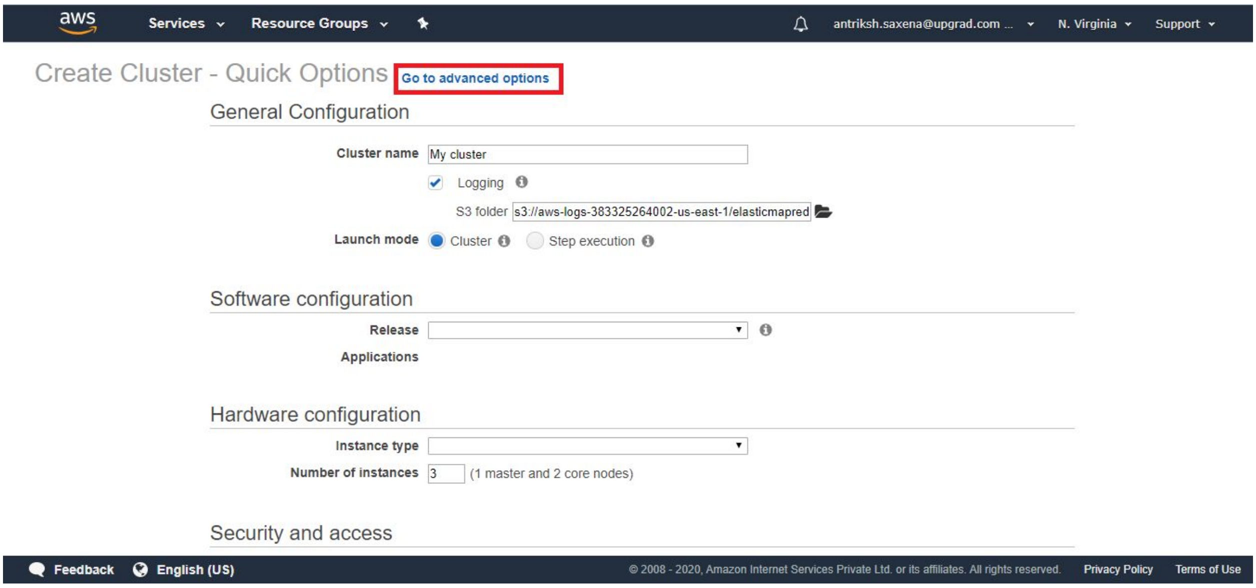
Steps to launch an EMR Cluster with spark installed on it

1. First, go to the EMR services section and click on the “**Create Cluster”** button.



Name	ID	Status	Creation time (UTC+5:30)	Elapsed time	Normalized instance hou
My cluster	j-15BIU1APUT6DR	Terminated User request	2020-04-15 12:23 (UTC+5:30)	11 hours, 32 minutes	96
My cluster	j-1SJO27LNV8N6K	Terminated User request	2020-04-15 11:04 (UTC+5:30)	1 hour, 18 minutes	16
My cluster	j-8O3ICPX81REI	Terminated with errors Step failure	2020-04-14 18:18 (UTC+5:30)	7 minutes	24
My cluster	j-2V9NL6QW9WUD4	Terminated User request	2020-04-14 00:23 (UTC+5:30)	35 minutes	8
My cluster	j-4D8CWQA78440	Terminated User request	2020-04-13 20:01 (UTC+5:30)	3 hours, 45 minutes	96
My cluster	j-3GA77LGHZHYRT	Terminated with errors Step failure	2020-04-13 17:49 (UTC+5:30)	6 minutes	24
streaming	j-EH3PDUV23SYP	Terminated User request	2020-04-07 18:54 (UTC+5:30)	2 hours, 40 minutes	36
streaming	j-1GB7JCHGXU7MF	Terminated User request	2020-04-07 11:14 (UTC+5:30)	7 hours, 39 minutes	96

2. You will be redirected to the below window, here click on the “**Go to advanced options”** button.


aws
Services
Resource Groups
antriksh.saxena@upgrad.com ...
N. Virginia
Support
Feedback
English (US)
© 2008 - 2020, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Create Cluster - Quick Options Go to advanced options

General Configuration

Cluster name: My cluster

☑ Logging

S3 folder: s3://aws-logs-383325264002-us-east-1/elasticmapred

Launch mode: ● Cluster ○ Step execution

Software configuration

Release:

Applications

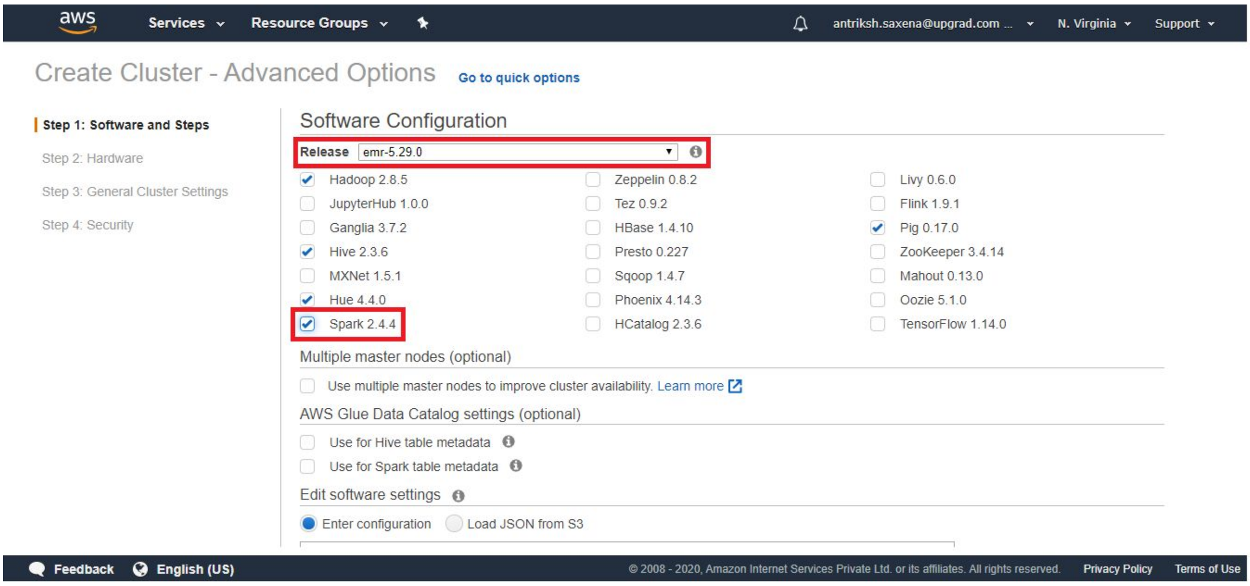
Hardware configuration

Instance type:

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

Security and access

3. Under the **Release** section and select the checkbox **Spark.**


aws
Services
Resource Groups
antriksh.saxena@upgrad.com ...
N. Virginia
Support
Feedback
English (US)
© 2008 - 2020, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Create Cluster - Advanced Options Go to quick options

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release: emr-5.29.0

☑ Hadoop 2.8.5
□ JupyterHub 1.0.0
□ Ganglia 3.7.2
☑ Hive 2.3.6
□ MXNet 1.5.1
☑ Hue 4.4.0
☑ Spark 2.4.4
□ Zeppelin 0.8.2
□ Tez 0.9.2
□ HBase 1.4.10
□ Presto 0.227
□ Sqoop 1.4.7
□ Phoenix 4.14.3
□ HCatalog 2.3.6
□ Livy 0.6.0
□ Flink 1.9.1
☑ Pig 0.17.0
□ ZooKeeper 3.4.14
□ Mahout 0.13.0
□ Oozie 5.1.0
□ TensorFlow 1.14.0

Multiple master nodes (optional)

□ Use multiple master nodes to improve cluster availability. Learn more

AWS Glue Data Catalog settings (optional)

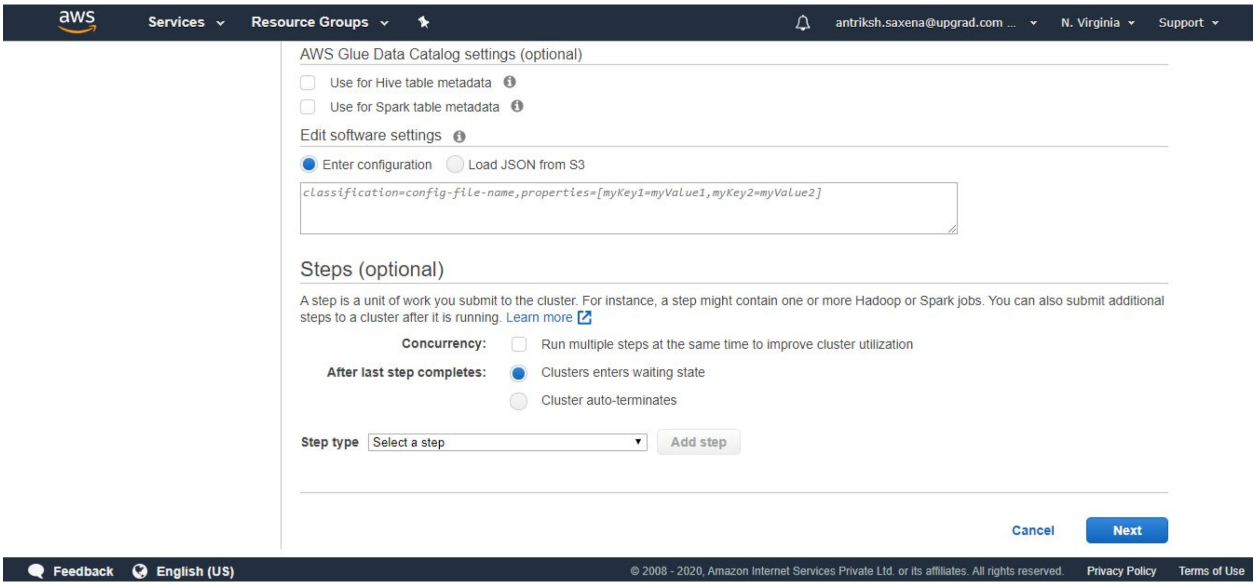
□ Use for Hive table metadata

□ Use for Spark table metadata

Edit software settings

● Enter configuration ○ Load JSON from S3

4. After selecting Spark, you need to scroll down and press “**Next”.** Here you don’t need to add any step in **“Add step”** section.


aws
Services
Resource Groups
antriksh.saxena@upgrad.com ...
N. Virginia
Support
Feedback
English (US)
© 2008 - 2020, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

AWS Glue Data Catalog settings (optional)

□ Use for Hive 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]
```

Steps (optional)

A step is a unit of work you submit to the cluster. For instance, a step might contain one or more Hadoop or Spark jobs. You can also submit additional steps to a cluster after it is running. Learn more

Concurrency: □ Run multiple steps at the same time to improve cluster utilization

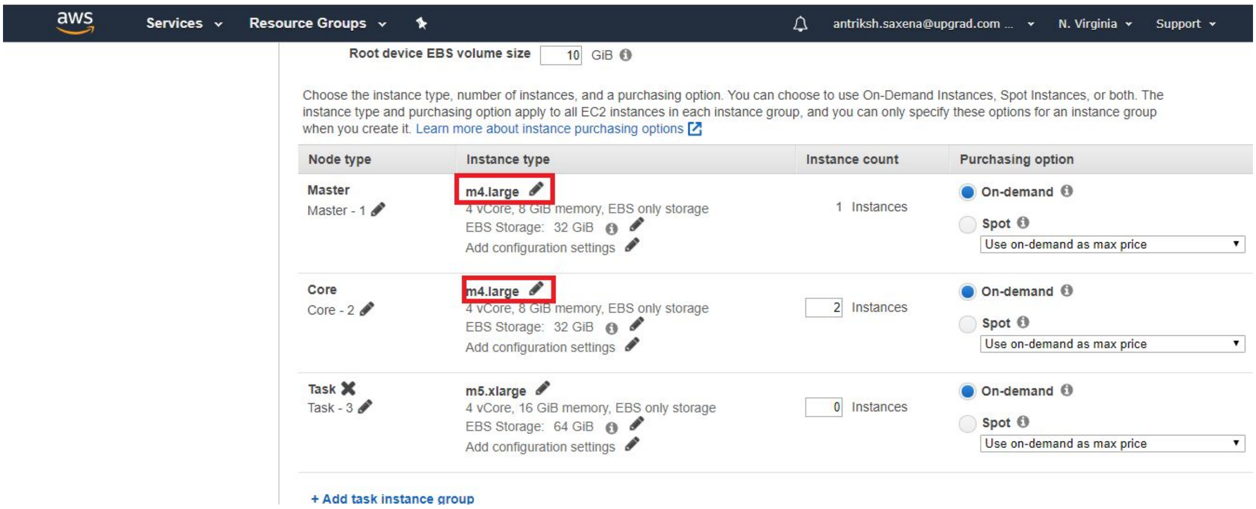
After last step completes: ● Clusters enters waiting state

○ Cluster auto-terminates

Step type Select a step ▾ Add step

Cancel Next

5. **Important -** Now change the **Master** and **Core** to **m4.large** and press **Next**. Ensure you are doing this else you would incur huge cost with the default settings.



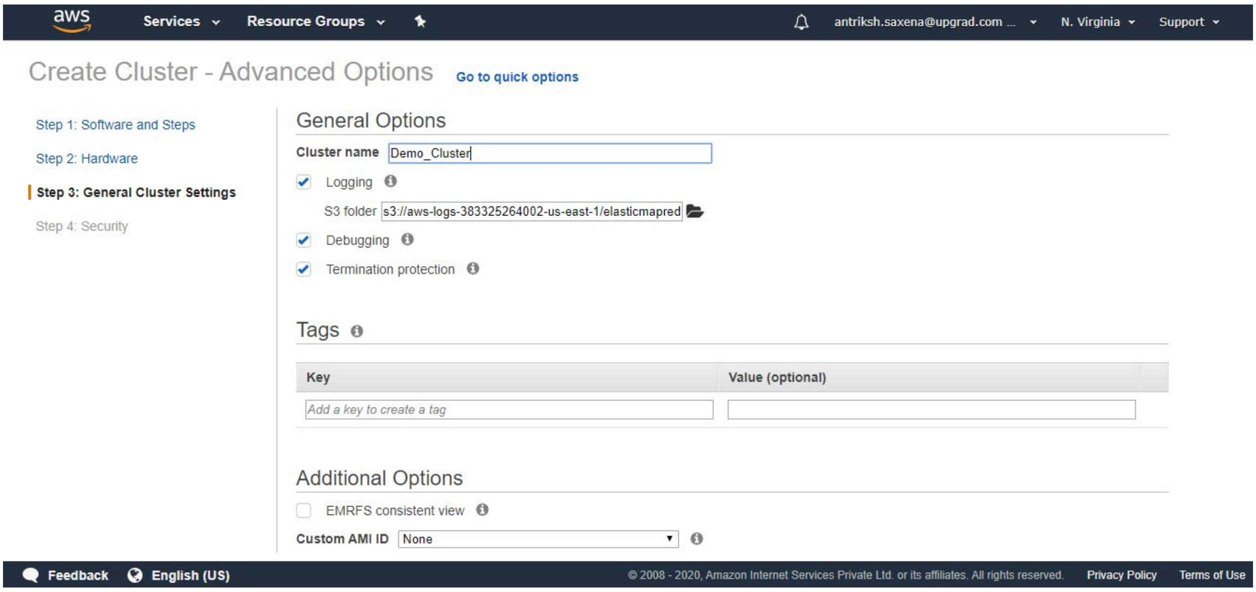
Root device EBS volume size 10 GiB

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	m4.large 4 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	● On-demand ○ Spot Use on-demand as max price ▾
Core Core - 2	m4.large 4 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	2 Instances	● On-demand ○ Spot Use on-demand as max price ▾
Task ✖ Task - 3	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	0 Instances	● On-demand ○ Spot Use on-demand as max price ▾

+ Add task instance group

6. Name your cluster as you can see we have named it “**Demo_Cluster**”. After naming the cluster, press **Next**.


aws
Services
Resource Groups
antriksh.saxena@upgrad.com ...
N. Virginia
Support
Feedback
English (US)
© 2008 - 2020, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Create Cluster - Advanced Options Go to quick options

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

General Options

Cluster name Demo_Cluster

☑ Logging

S3 folder s3://aws-logs-383325264002-us-east-1/elasticmapred

☑ Debugging

☑ Termination protection

Tags

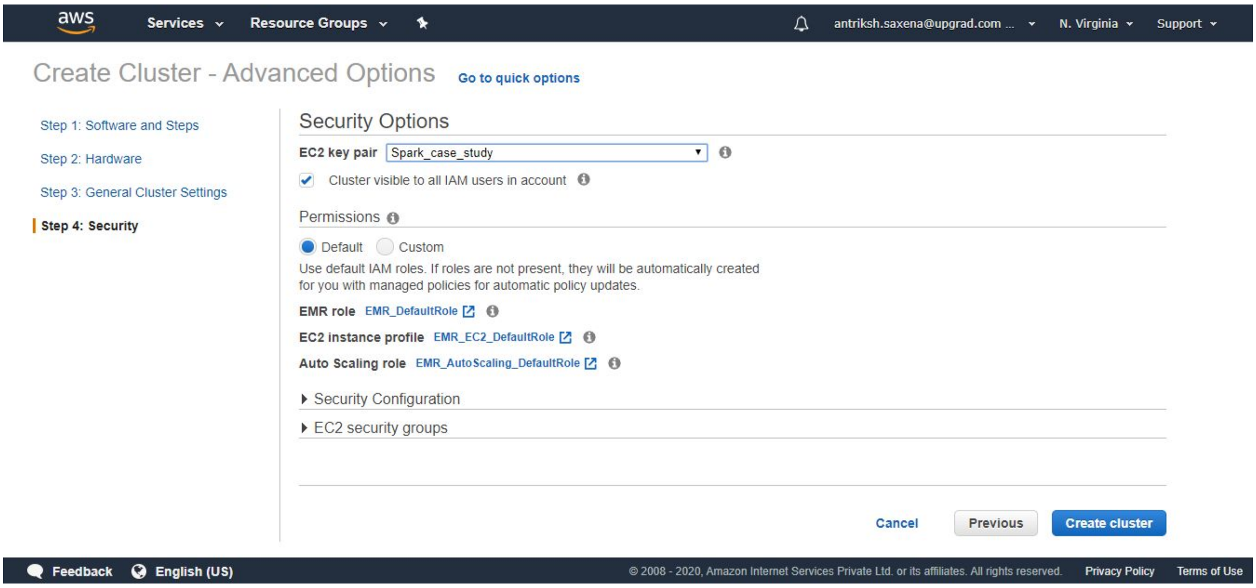
Key	Value (optional)
Add a key to create a tag	

Additional Options

□ EMRFS consistent view

Custom AMI ID None

7. In this step select an existing key pair and click on **Create cluster** button.You can proceed further with the existing EC2 key pair or any other keypair for which the corresponding PPK file is available. In case you don’t have a PPK file for any of the existing key pairs then you need to create a new key pair, you can refer to previous modules where creating a new pair is demonstrated to you Finally hit the **“Create cluster”** button.


aws
Services
Resource Groups
antriksh.saxena@upgrad.com ...
N. Virginia
Support
Cancel
Previous
Create cluster
Feedback
English (US)
© 2008 - 2020, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Create Cluster - Advanced Options Go to quick options

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

Security Options

EC2 key pair Spark_case_study

☑ 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

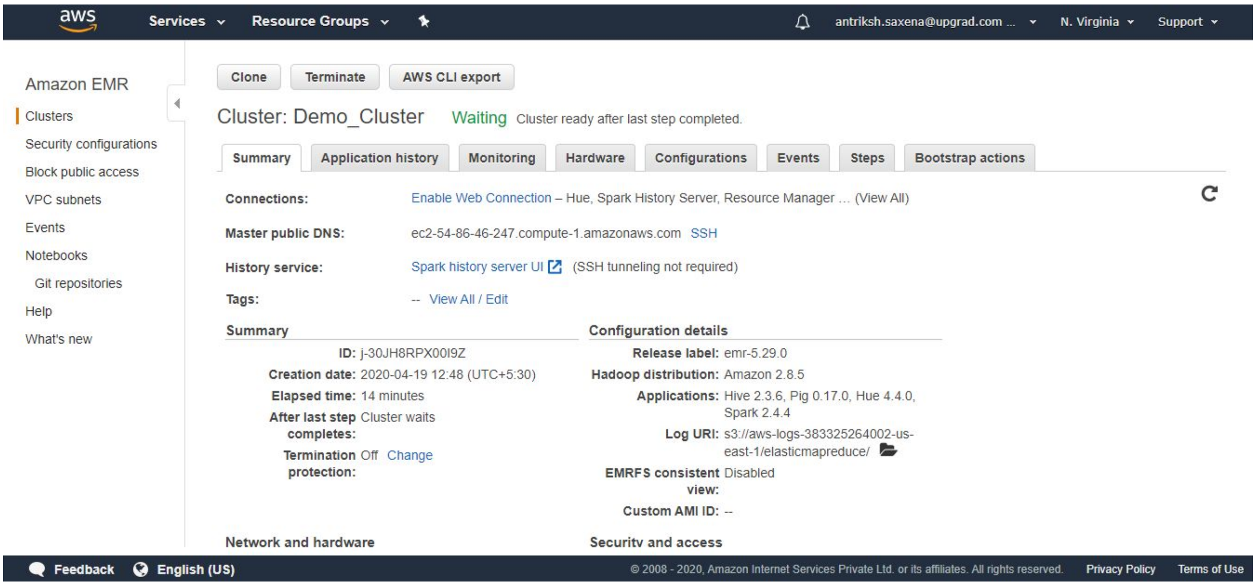
EC2 instance profile EMR_EC2_DefaultRole

Auto Scaling role EMR_AutoScaling_DefaultRole

▸ Security Configuration

▸ EC2 security groups

8. This will take some time to create the cluster and then you will find cluster status as **Waiting**.



Cluster: Demo_Cluster Waiting Cluster ready after last step completed.

Connections:	Enable Web Connection – Hue, Spark History Server, Resource Manager ... (View All)
Master public DNS:	ec2-54-86-46-247.compute-1.amazonaws.com SSH
History service:	Spark history server UI (SSH tunneling not required)
Tags:	-- View All / Edit

Summary

ID:	j-30JH8RPX00I9Z
Creation date:	2020-04-19 12:48 (UTC+5:30)
Elapsed time:	14 minutes
After last step completes:	Cluster waits
Termination protection:	Off Change

Configuration details

Release label:	emr-5.29.0
Hadoop distribution:	Amazon 2.8.5
Applications:	Hive 2.3.6, Pig 0.17.0, Hue 4.4.0, Spark 2.4.4
Log URI:	s3://aws-logs-383325264002-us-east-1/elasticmapreduce/
EMRFS consistent view:	Disabled
Custom AMI ID:	--

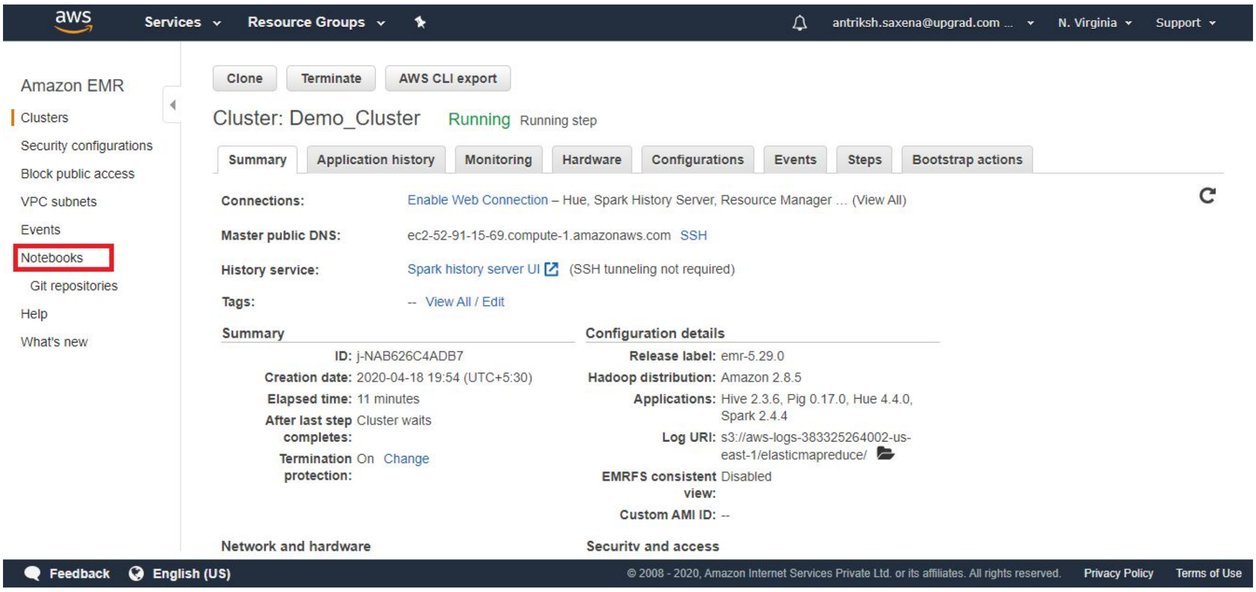
Network and hardware

Securitv and access

© 2008 - 2020, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.

Running Jupyter notebook on EMR cluster

9. Now you need to go to the **“Notebook”** section as shown in the image below.



Cluster: Demo_Cluster Running Running step

Connections: Enable Web Connection – Hue, Spark History Server, Resource Manager ... (View All)

Master public DNS: ec2-52-91-15-69.compute-1.amazonaws.com SSH

History service: Spark history server UI (SSH tunneling not required)

Tags: -- View All / Edit

Summary

ID:	j-NAB626C4ADB7
Creation date:	2020-04-18 19:54 (UTC+5:30)
Elapsed time:	11 minutes
After last step completes:	Cluster waits
Termination protection:	On Change

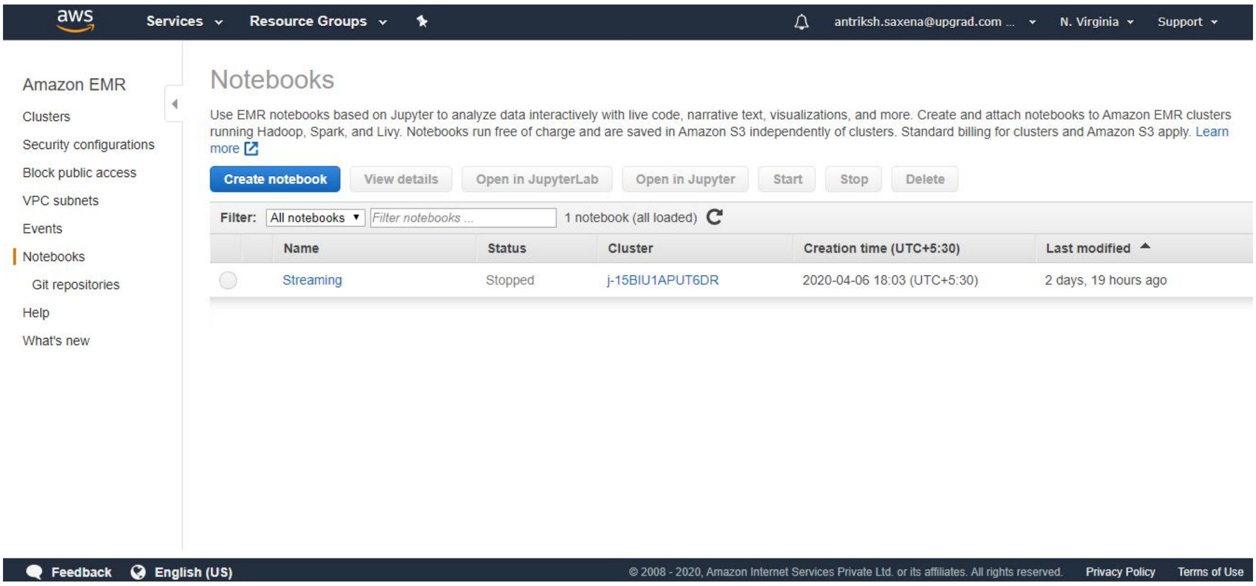
Configuration details

Release label:	emr-5.29.0
Hadoop distribution:	Amazon 2.8.5
Applications:	Hive 2.3.6, Pig 0.17.0, Hue 4.4.0, Spark 2.4.4
Log URI:	s3://aws-logs-383325264002-us-east-1/elasticmapreduce/
EMRFS consistent view:	Disabled
Custom AMI ID:	--

Network and hardware

Security and access

10. Hit on the **Create notebook** button.



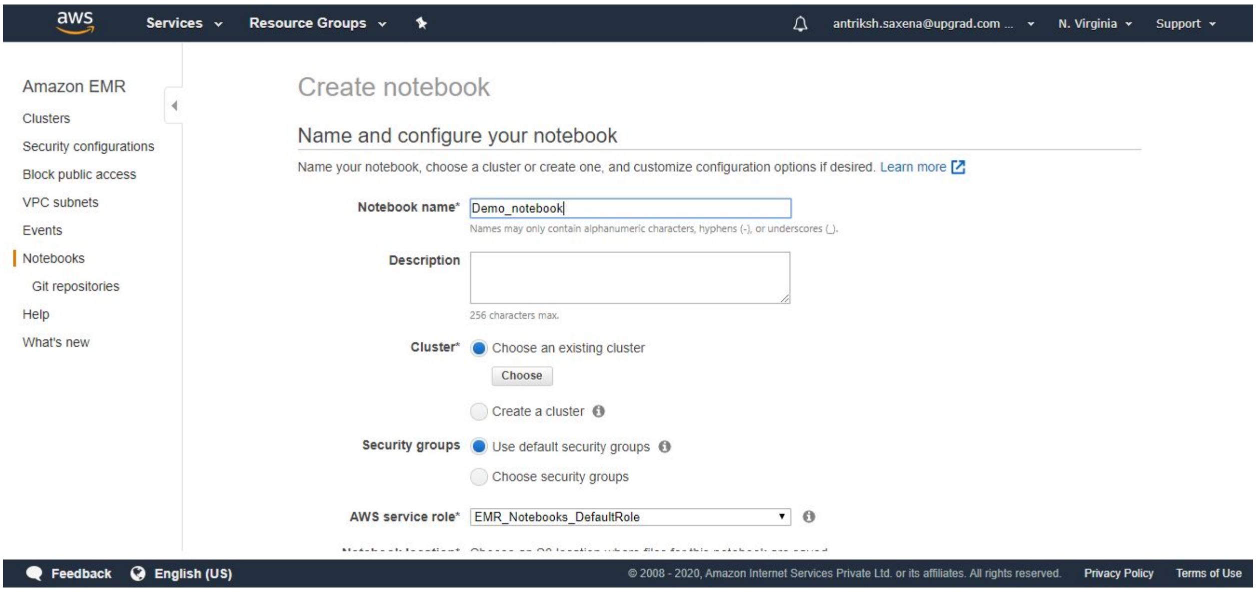
Notebooks

Use EMR notebooks based on Jupyter to analyze data interactively with live code, narrative text, visualizations, and more. Create and attach notebooks to Amazon EMR clusters running Hadoop, Spark, and Livy. Notebooks run free of charge and are saved in Amazon S3 independently of clusters. Standard billing for clusters and Amazon S3 apply. Learn more

Filter: All notebooks | Filter notebooks ... | 1 notebook (all loaded)

	Name	Status	Cluster	Creation time (UTC+5:30)	Last modified
○	Streaming	Stopped	j-15BIU1APUT6DR	2020-04-06 18:03 (UTC+5:30)	2 days, 19 hours ago

11. Name the notebook as per your desire. As you can see in our case the name is “**Demo_notebook**”



Create notebook

Name and configure your notebook

Name your notebook, choose a cluster or create one, and customize configuration options if desired. Learn more

Notebook name* Demo_notebook
Names may only contain alphanumeric characters, hyphens (-), or underscores (_).

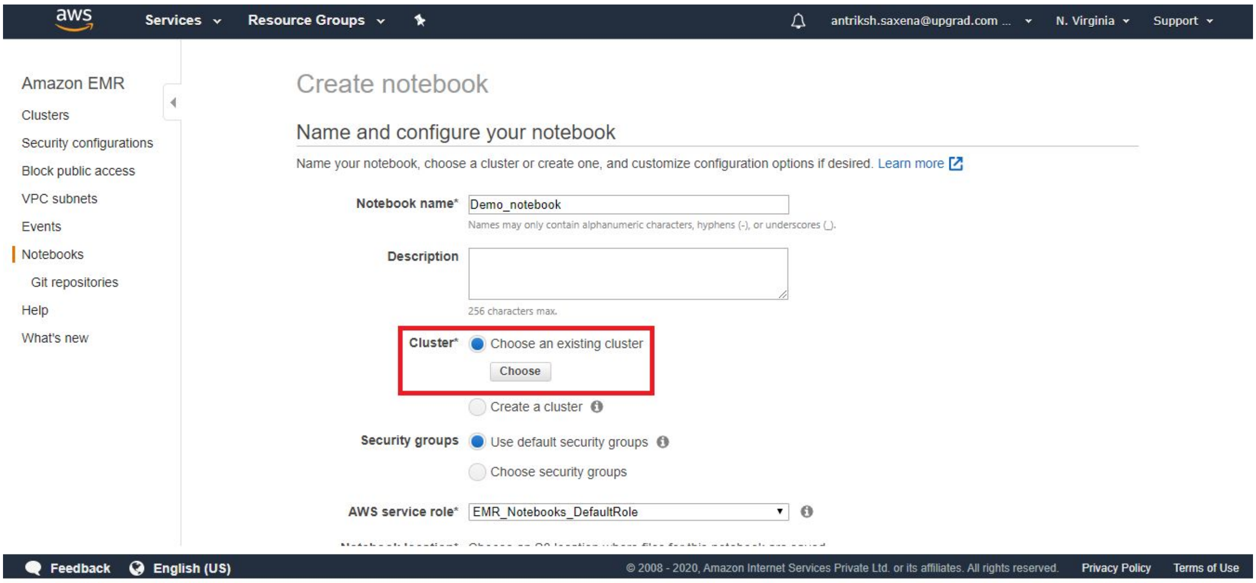
Description
256 characters max.

Cluster* ● Choose an existing cluster
Choose
○ Create a cluster

Security groups ● Use default security groups
○ Choose security groups

AWS service role* EMR_Notebooks_DefaultRole

12. Then click on **choose** in the cluster section.



Create notebook

Name and configure your notebook

Name your notebook, choose a cluster or create one, and customize configuration options if desired. Learn more

Notebook name* Demo_notebook
Names may only contain alphanumeric characters, hyphens (-), or underscores (_).

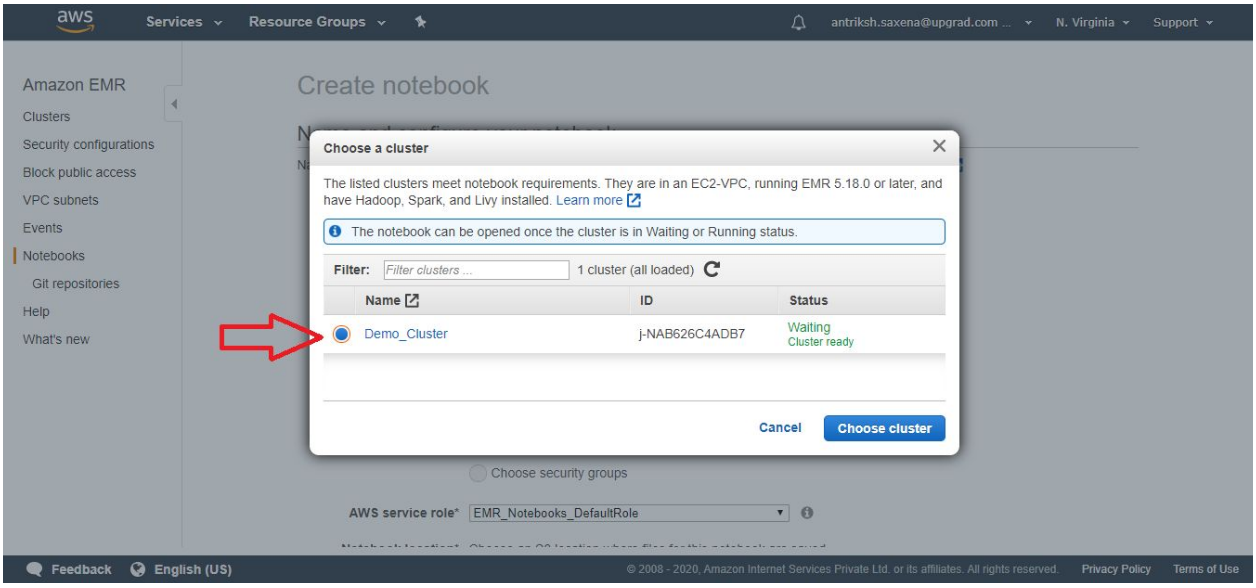
Description
256 characters max.

Cluster* ● Choose an existing cluster
Choose
○ Create a cluster

Security groups ● Use default security groups
○ Choose security groups

AWS service role* EMR_Notebooks_DefaultRole

13. Now, select the existing cluster as shown in the image below.



Choose a cluster

The listed clusters meet notebook requirements. They are in an EC2-VPC, running EMR 5.18.0 or later, and have Hadoop, Spark, and Livy installed. Learn more

The notebook can be opened once the cluster is in Waiting or Running status.

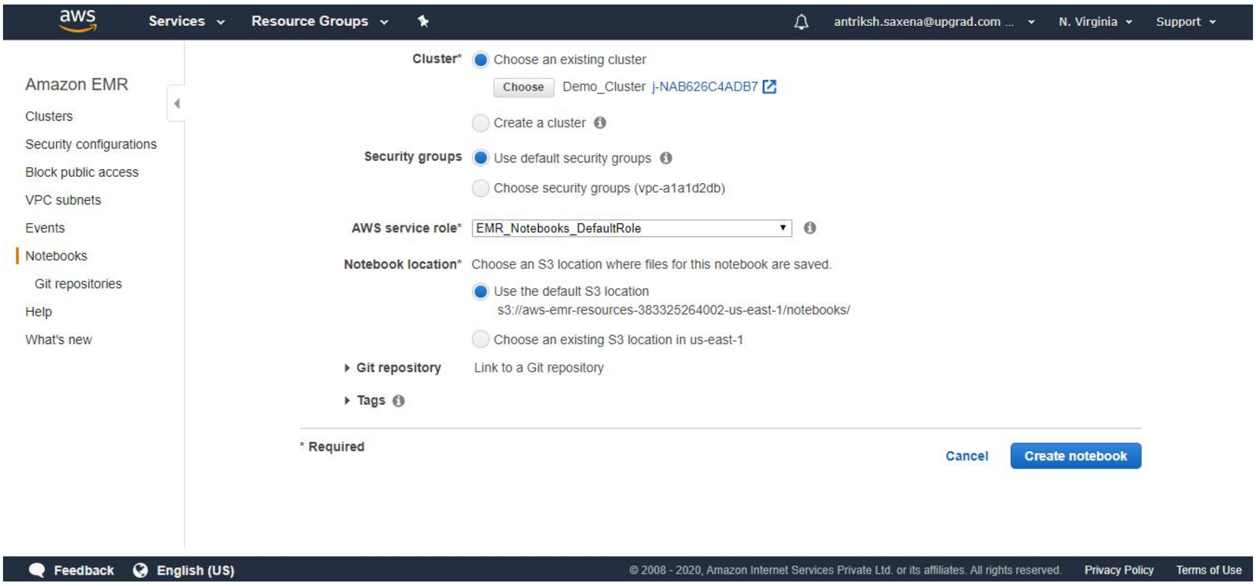
Filter: Filter clusters ... 1 cluster (all loaded)

	Name	ID	Status
●	Demo_Cluster	j-NAB626C4ADB7	Waiting Cluster ready

○ Choose security groups

AWS service role* EMR_Notebooks_DefaultRole

14. Once you choose the cluster, then you will have to hit the “**Create notebook**” button at the bottom of the window. It will take some time to create a notebook.



Cluster* ● Choose an existing cluster
Choose Demo_Cluster j-NAB626C4ADB7
○ Create a cluster

Security groups ● Use default security groups
○ Choose security groups (vpc-a1a1d2db)

AWS service role* EMR_Notebooks_DefaultRole

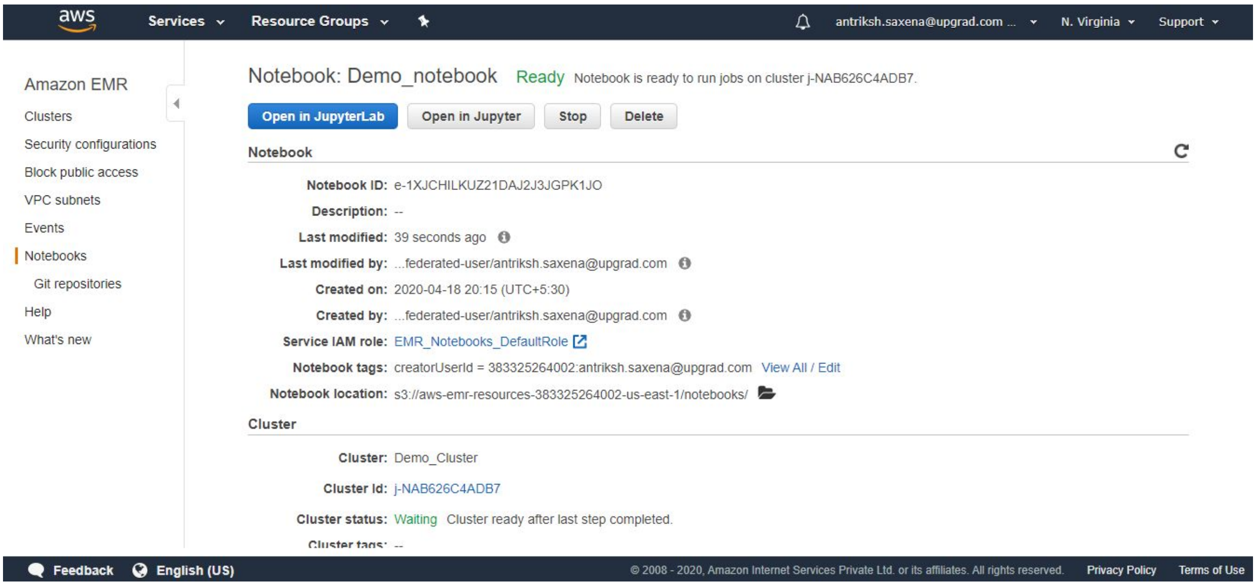
Notebook location* Choose an S3 location where files for this notebook are saved.
● Use the default S3 location
s3://aws-emr-resources-383325264002-us-east-1/notebooks/
○ Choose an existing S3 location in us-east-1

Git repository Link to a Git repository

Tags

* Required

15. Now click on the “**Open in Jupyter**” button. This will land you to the jupyter notebook tab.



Notebook: Demo_notebook Ready Notebook is ready to run jobs on cluster j-NAB626C4ADB7.

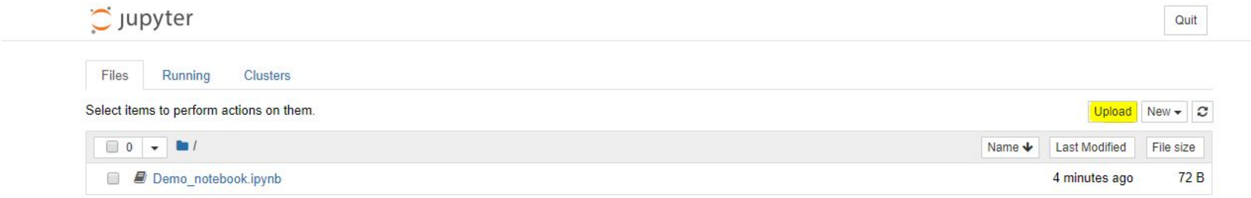
Notebook

Notebook ID: e-1XJCHILKUZ21DAJ2J3JGPK1JO
Description: --
Last modified: 39 seconds ago
Last modified by: ...federated-user/antriksh.saxena@upgrad.com
Created on: 2020-04-18 20:15 (UTC+5:30)
Created by: ...federated-user/antriksh.saxena@upgrad.com
Service IAM role: EMR_Notebooks_DefaultRole
Notebook tags: creatorUserId = 383325264002:antriksh.saxena@upgrad.com View All / Edit
Notebook location: s3://aws-emr-resources-383325264002-us-east-1/notebooks/

Cluster

Cluster: Demo_Cluster
Cluster Id: j-NAB626C4ADB7
Cluster status: Waiting Cluster ready after last step completed.
Cluster tags: --

16. You can upload your own notebook from the “**upload**” button on the top right part of the window.

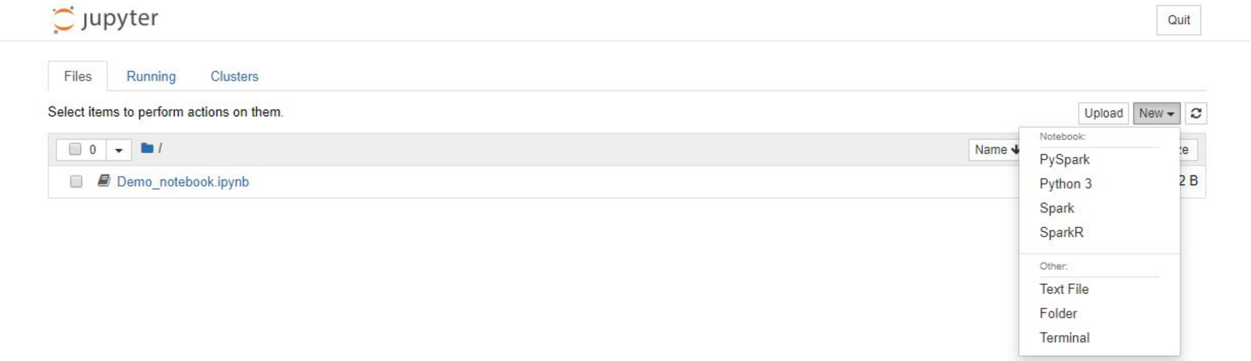


Select items to perform actions on them.

Demo_notebook.ipynb 4 minutes ago 72 B

.

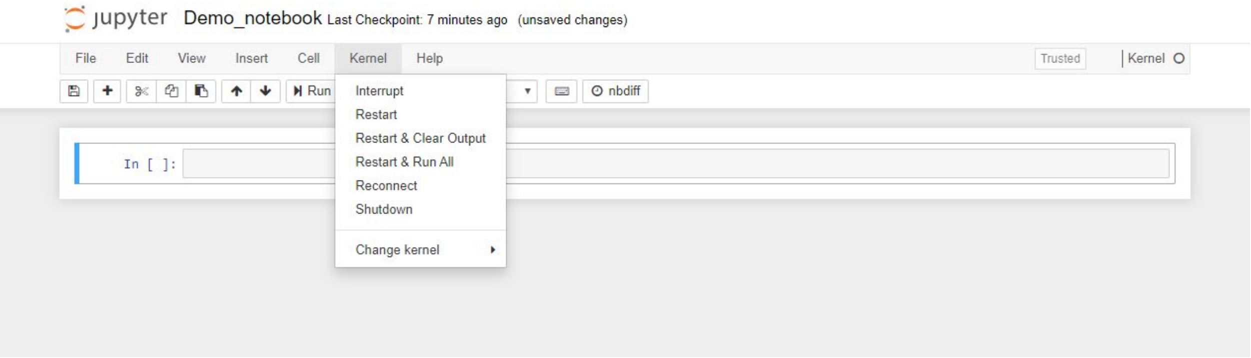
17. There is also an option to create a new notebook with the desired kernel .



Select items to perform actions on them.

Demo_notebook.ipynb

18. There is an option in the notebook to change the kernel also. From which you can select the desired kernel.



In this way you can run the Spark application on the EMR cluster.