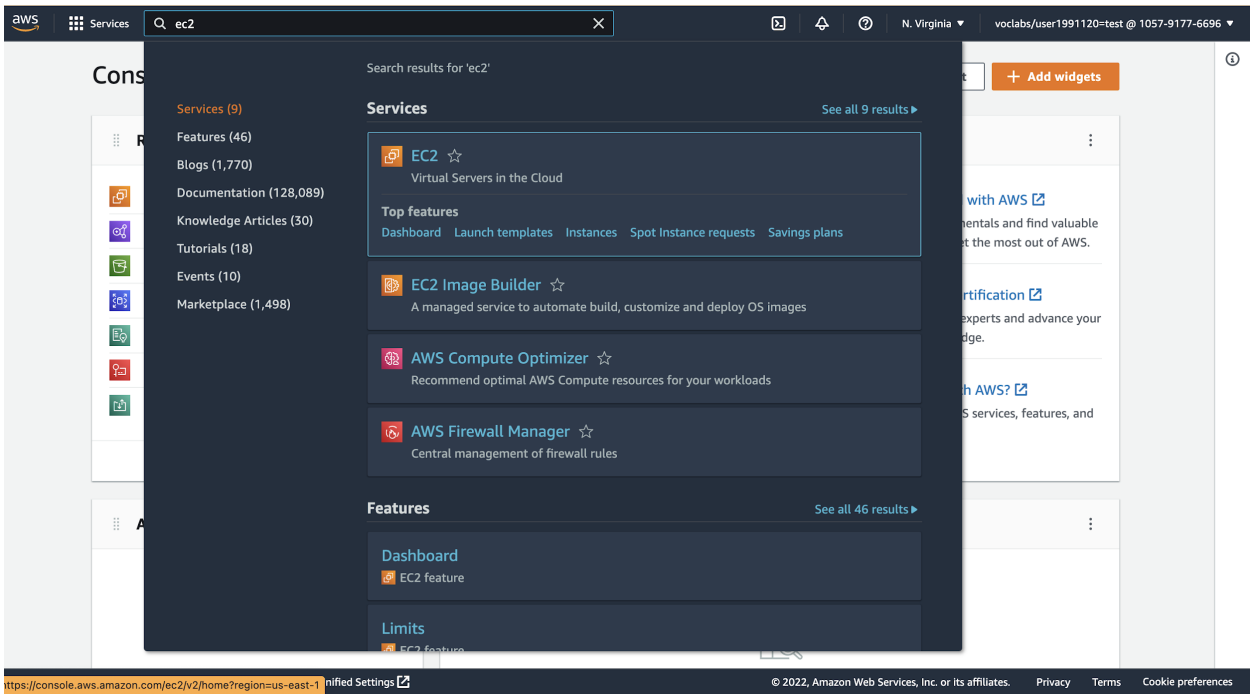
Setting up Apache Spark instance and Jupyter Notebook

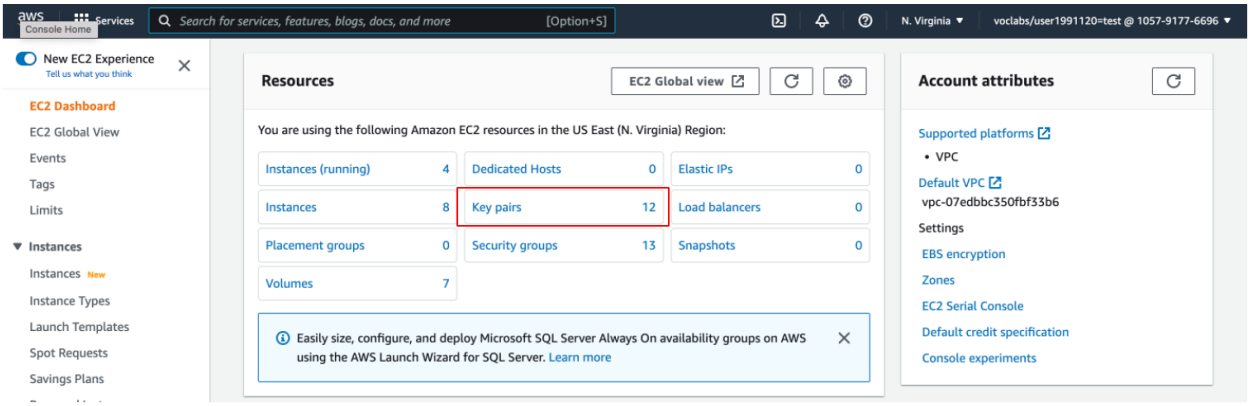
This document contains the steps to create an EMR cluster for using Apache Spark.

1. Before you create your EMR cluster, you will need to create a key pair. This is because your EMR cluster will be running on EC2 instances and you will require a key pair to connect with your instance. Let us quickly revise how you can create your key pair using the console.

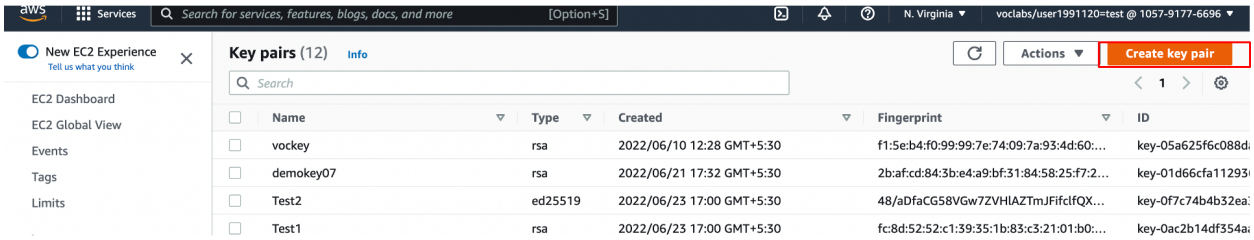
   a. Once you have logged in to your AWS account, search for ‘EC2’ under ‘Find Services’ and click on it.



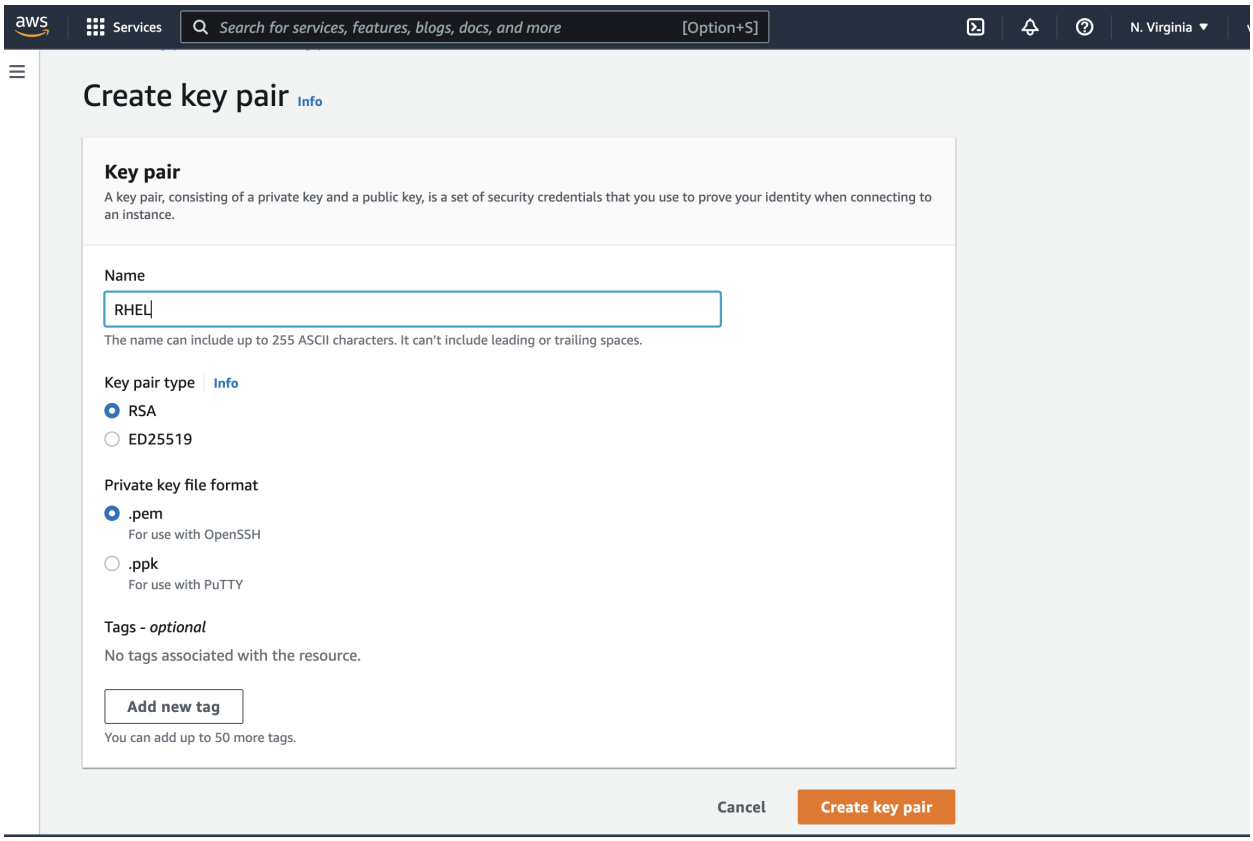
   b. Under ‘Resources’, click on ‘Key Pairs’.



c. Now click on ‘Create key pair’

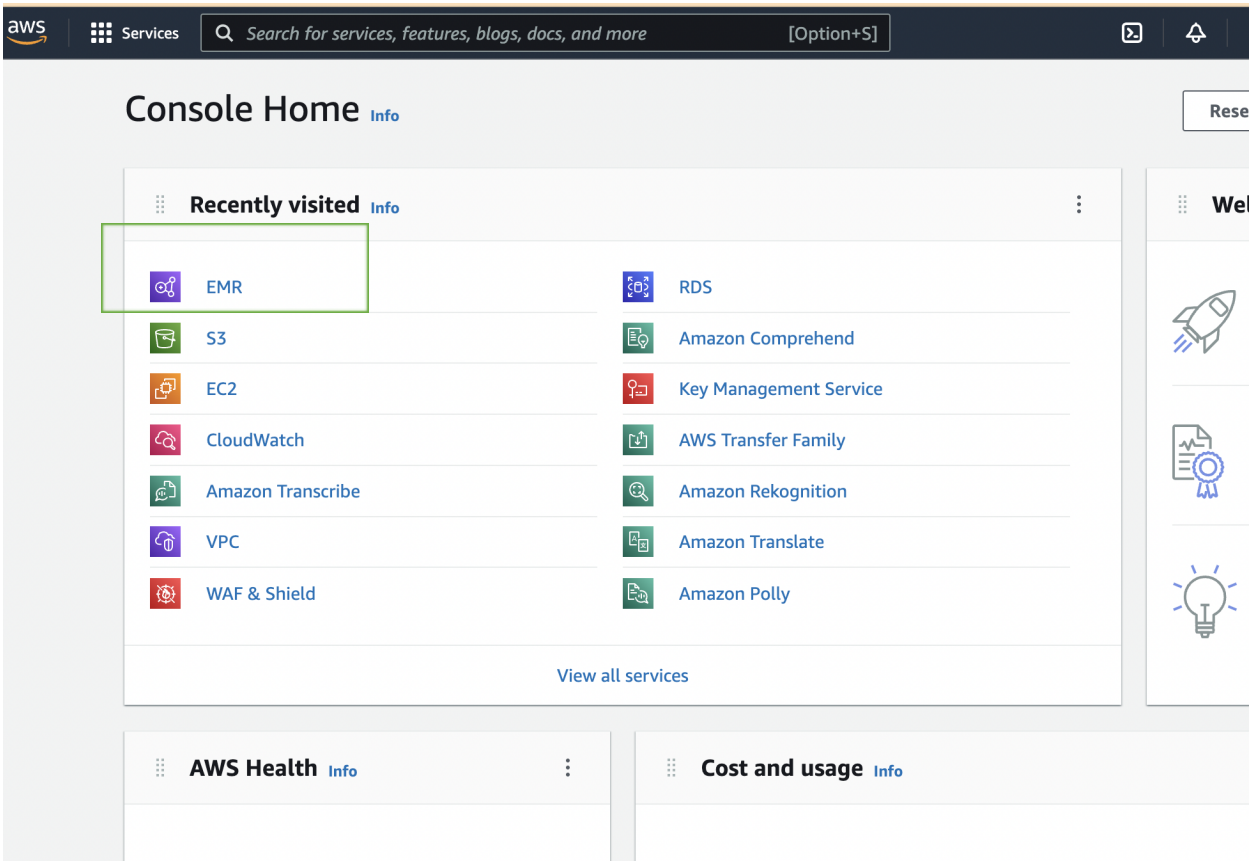


d. Give a name to your key pair. In our case, we have named it as RHEL and used the pem File format. Now click on ‘Create key pair’.

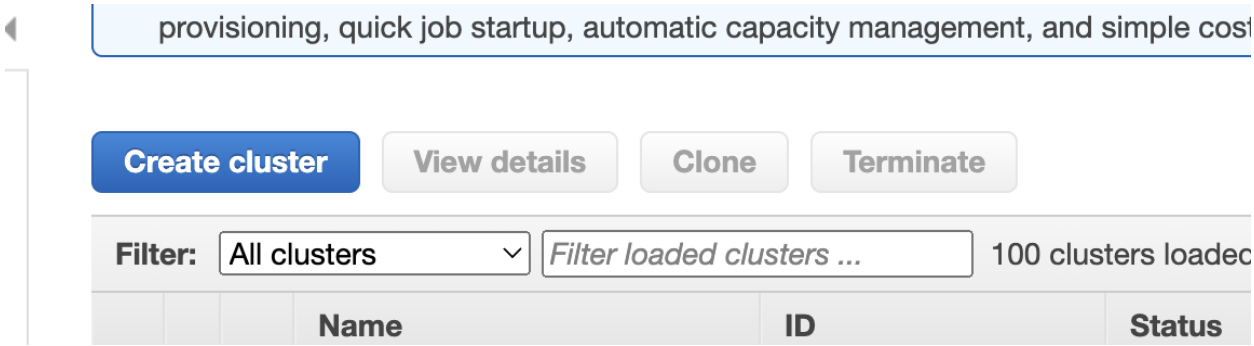


e. Great! You now have your key pair and you can proceed with launching your EMR cluster

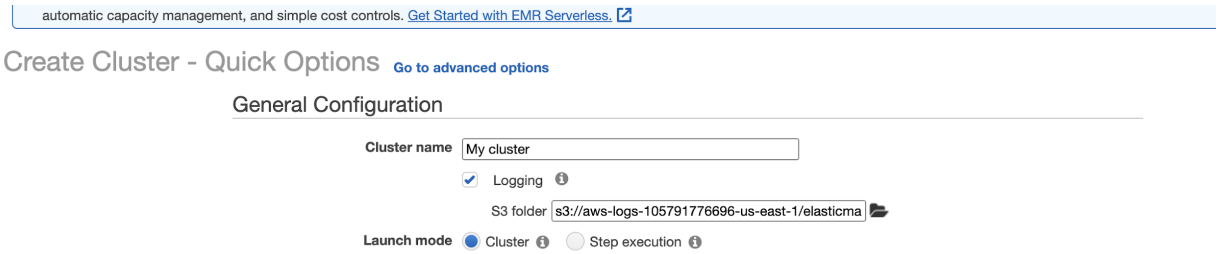
2. Click on the **Services** at the top of the AWS console and then click on the **EMR** service.



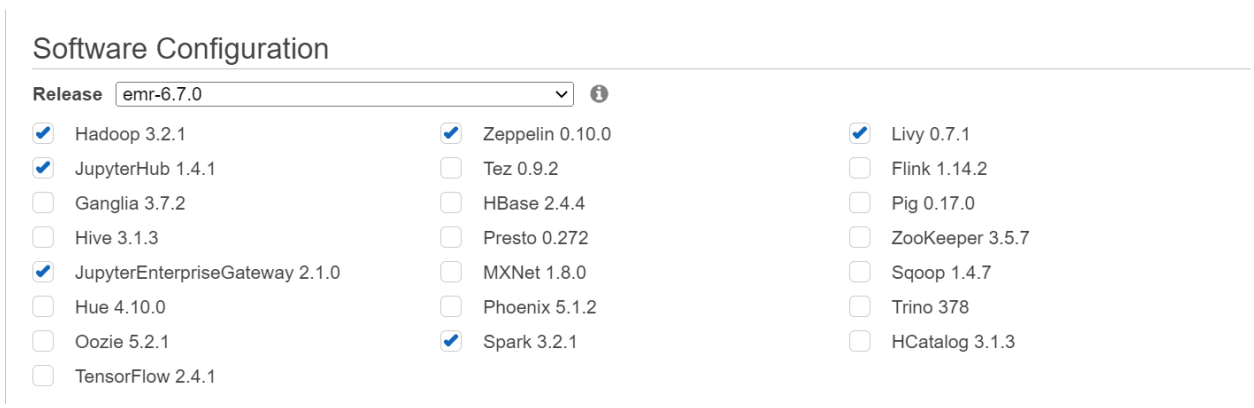
3. Click on the **Create cluster** button and you will go to the cluster creation page.



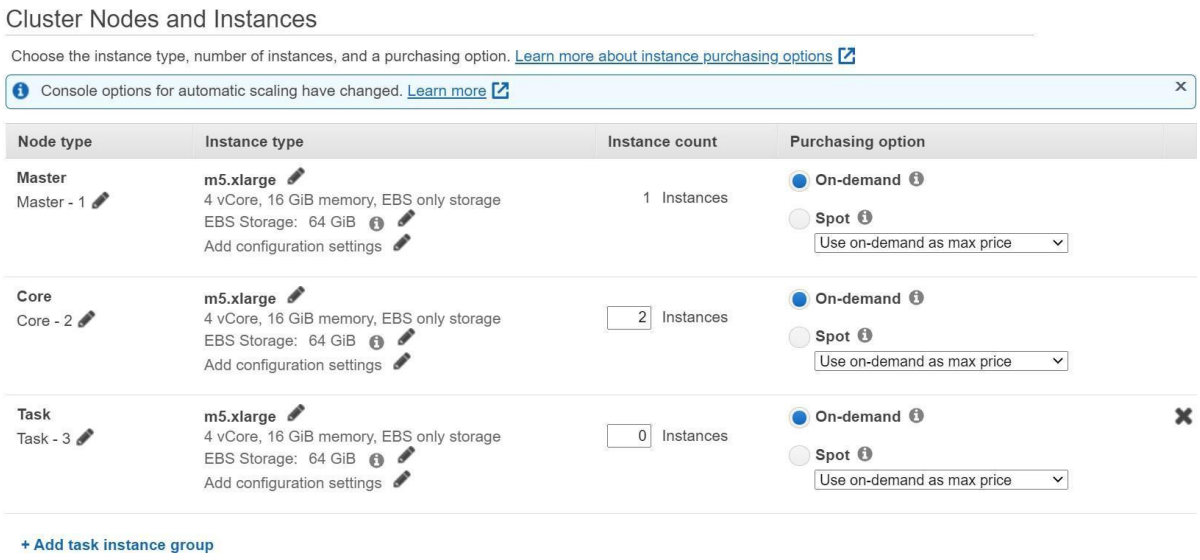
4. Once you click on the 'Create cluster' button, you need to click on the **Go to advanced options** link.



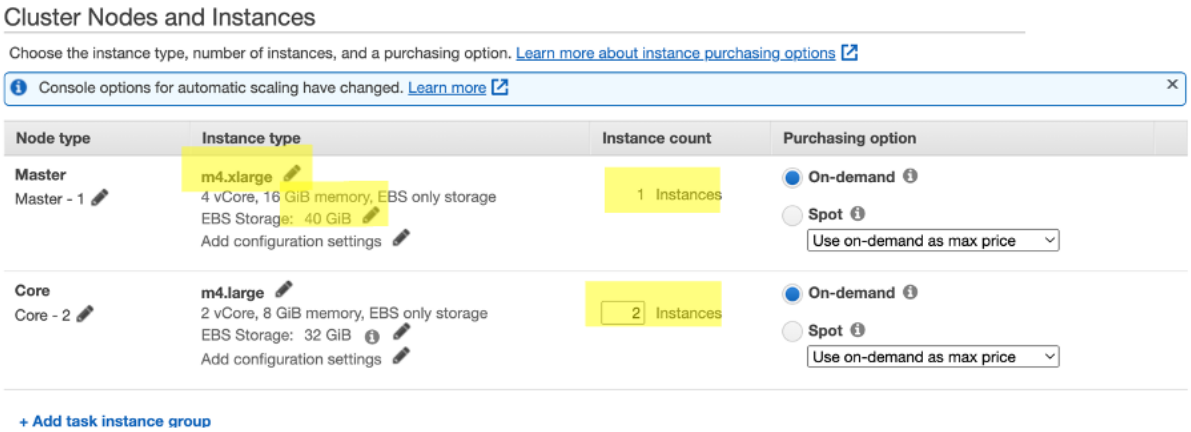
5. It will take you to the following page. In the 'Release' column, you will be choosing the **emr-6.7.0** version for your EMR cluster. You will now configure the software applications that you need for your EMR cluster. For HBase, you need to install the following services -
    - **Hadoop**
    - **Spark**
    - **Zeppelin**
    - **JupyterHub**
    - **JupyterEnterpriseGateway**
    - **Livy**



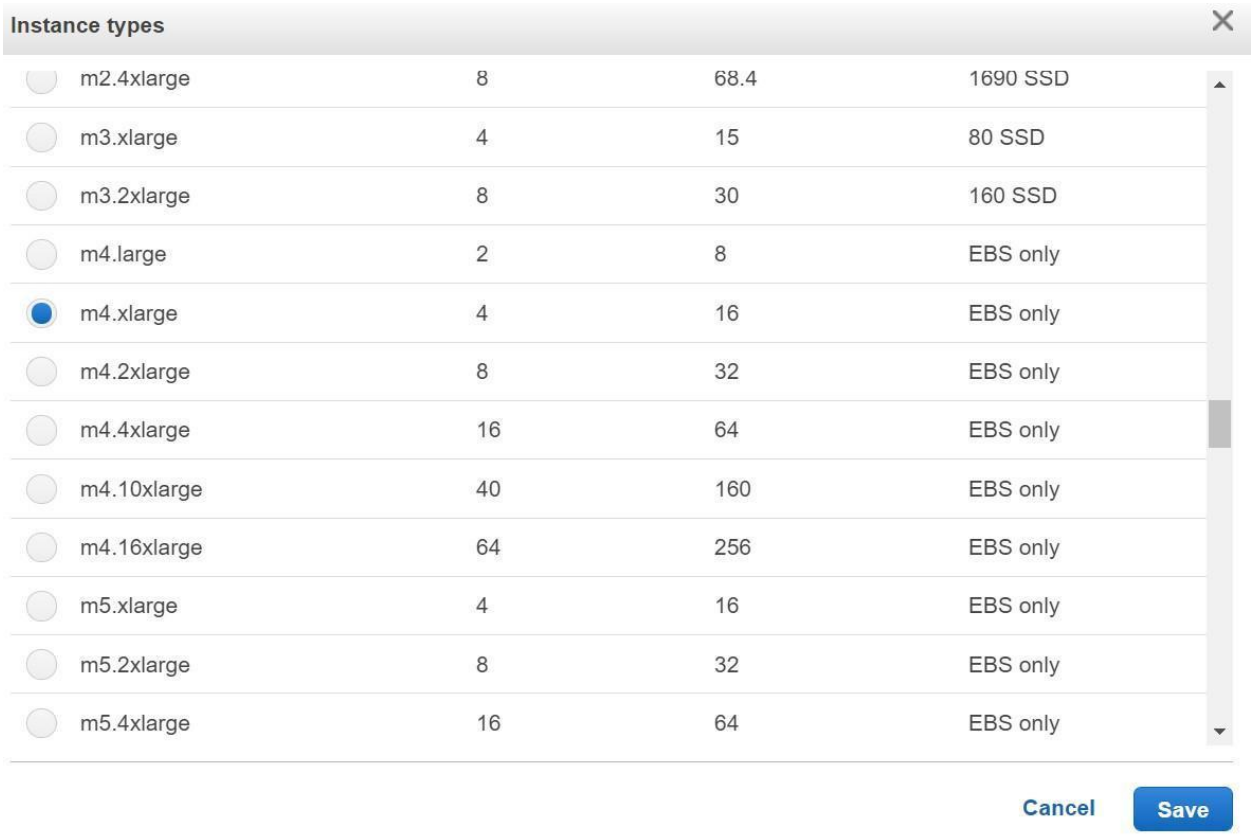
6. Click on the **Next** button at the end of the page.
7. In this page, scroll down to the 'Cluster Nodes' and 'Instances' section. You will now have to click on the **Cross** button to the right of the **Task Node**, and then under **Core Node**, you will need to type **0** under Instances. Do not create a multiple node cluster, otherwise, you might consume the entire budget in a single day.



8. Now, you need to click on the pencil button to the right of **m5.xlarge**.

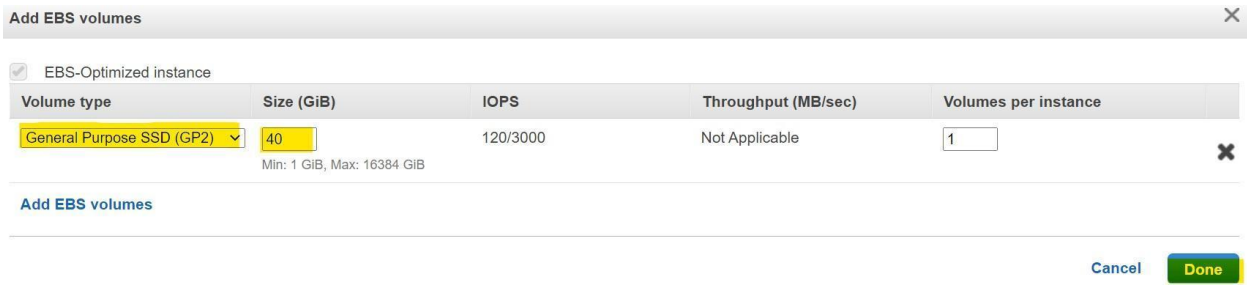


9. Now, you need to select the **m4.xlarge** instance type under the window that appears, and then click on the **Save** button. You can use 'Control + F' keys and search for m4.xlarge.



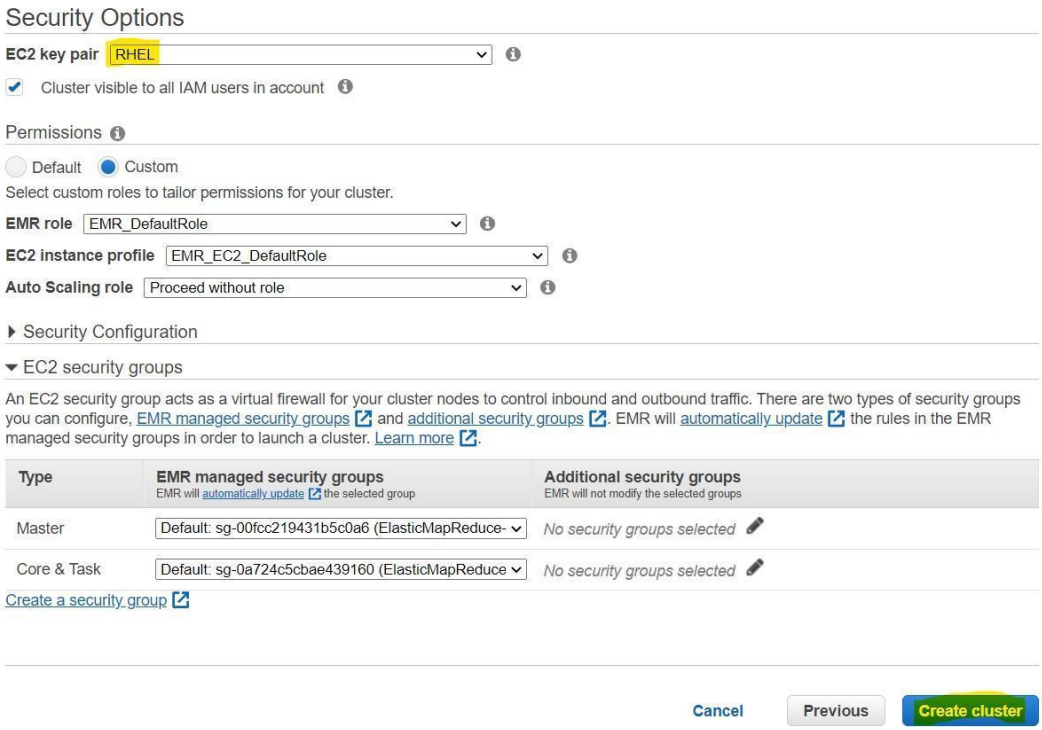
10. Now, you need to click on the pencil button to the right of the EBS storage.

11. Next, you need to configure the EBS volume for your EMR cluster. Click on the 'Volume type' and select the **General Purpose SSD (GP2)** option, and then under the 'Size' column, type **40**. Remove any other EBS volumes if present. After this, you can now click on the **Done** button. Thereafter, you can click on the **Next button** for this step of the advanced options as well.

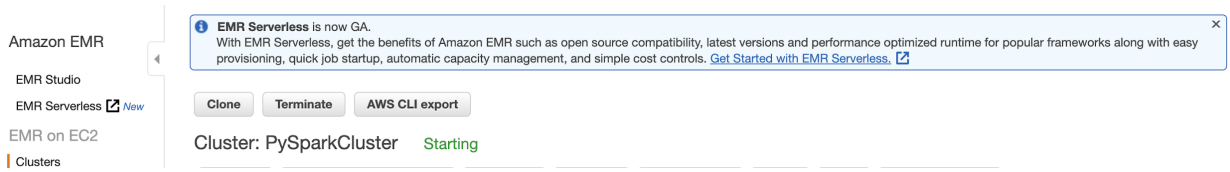


12. Now, in this step like the previous method, type the Cluster name that you want for your EMR cluster. Also, uncheck the 'Termination protection' option as this is not needed for this EMR instance. You can now click on the **Next** button.

13. Now in this step, you just need to select the EC2 key pair that you had created previously, and after that, you can click on the **Create cluster** button.



14. Thereafter, the cluster will start setting up.

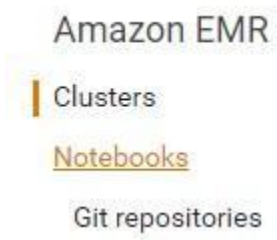


Once you have the EMR running, follow the steps to launch jupyter notebook:

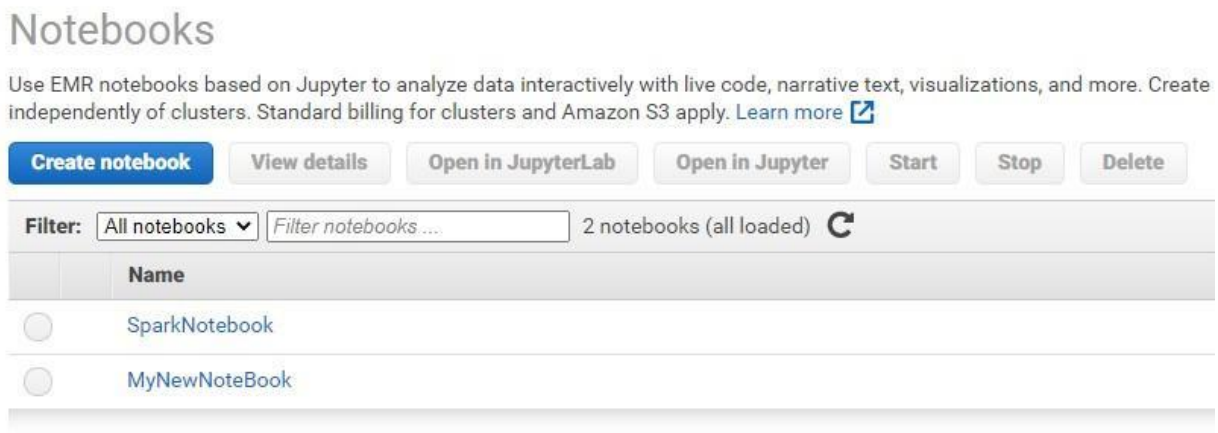
Creating a Jupyter Notebook on an EMR Cluster

Once you have created and started your EMR cluster, you can then create a new Jupyter Notebook wherein you can write your Spark jobs. Note that these notebooks are persisted on S3 even after you terminate the EMR cluster, so you don't have to worry about creating a Jupyter Notebook again from scratch. You can follow the steps below to create a Jupyter Notebook:

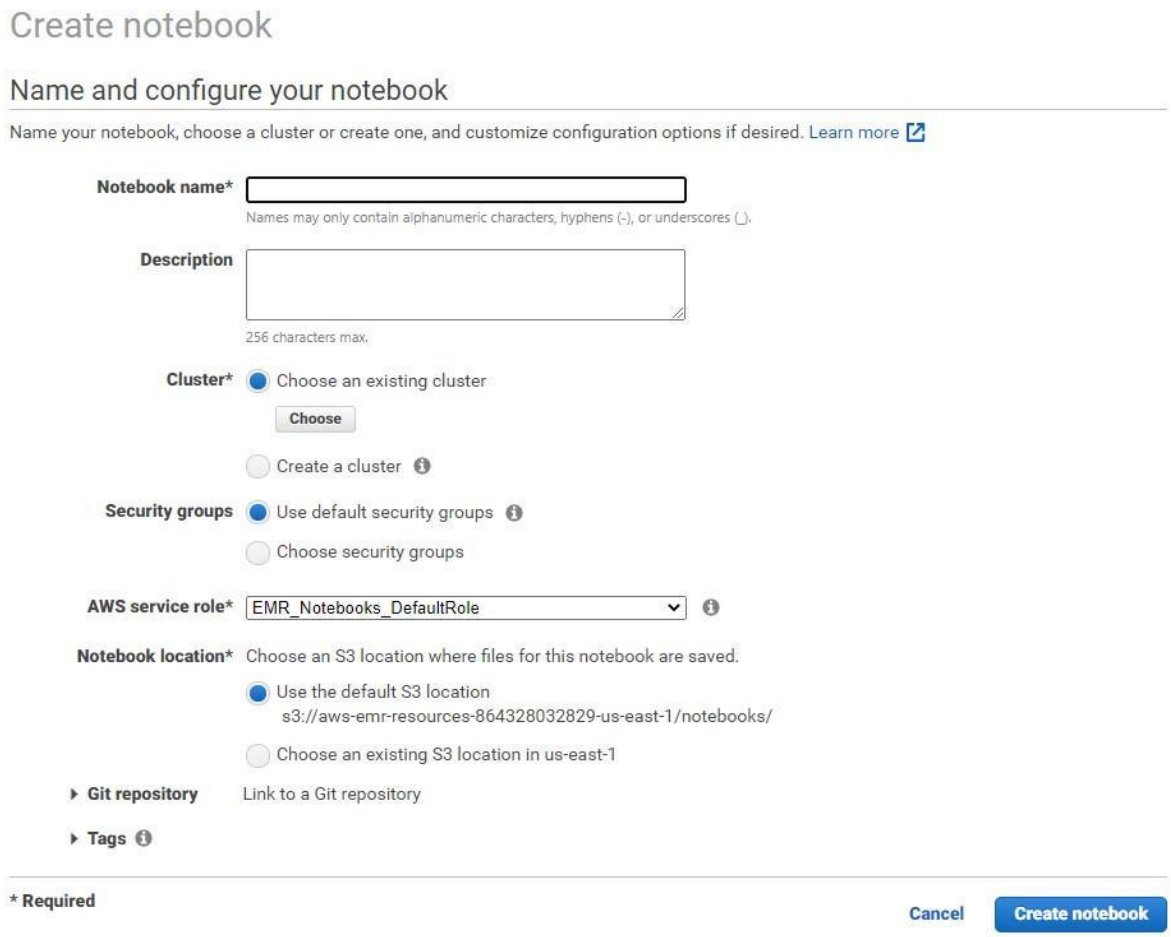
Step 1: First, you need to go to the 'Notebooks' link in the left navigation pane under Amazon EMR.



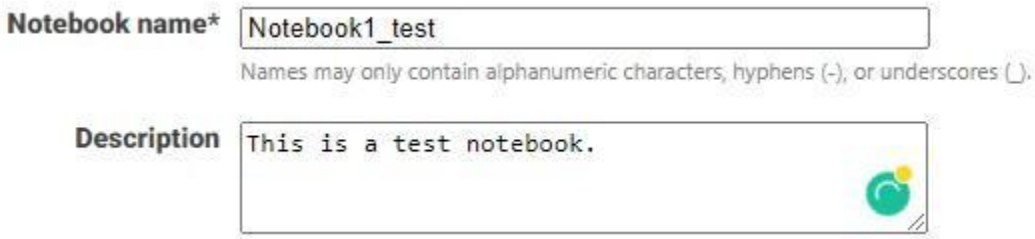
Step 2: On clicking the link, your screen will appear as shown below. Now, to create a new Jupyter Notebook, simply click on the 'Create notebook' button.



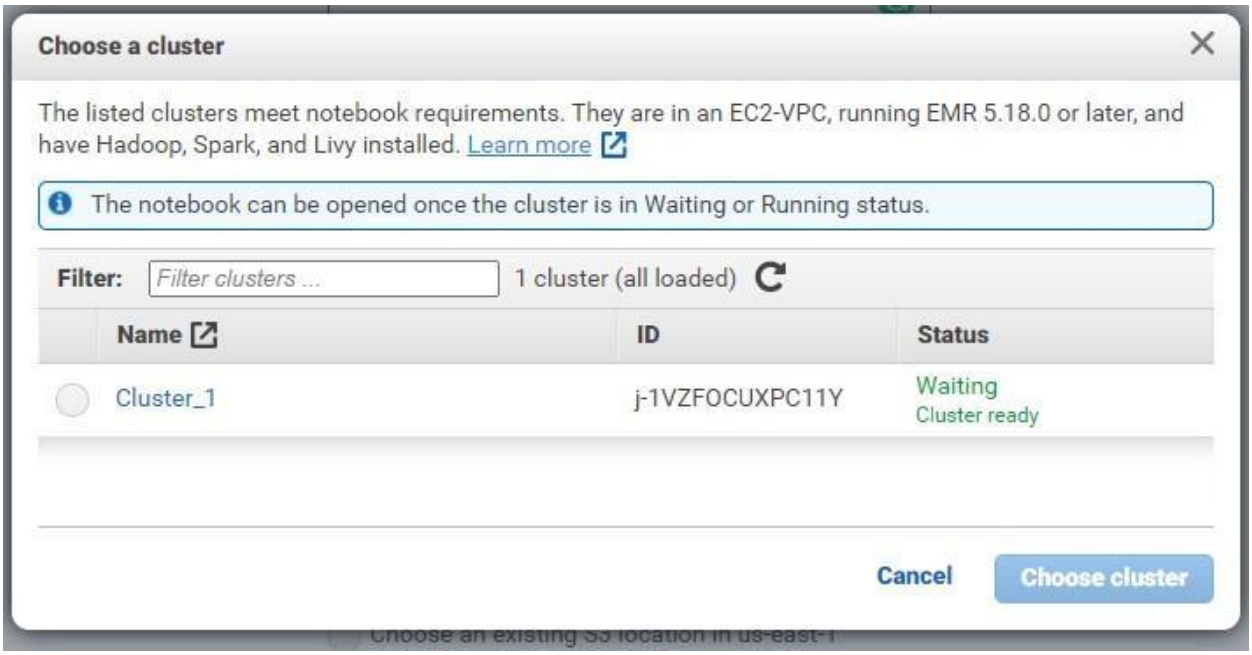
Once you click on the button, the following page will open on your screen.



Step 3: Here, you can write a 'Notebook name' for your Jupyter Notebook. Under 'Description', you can write a few lines to describe the notebook that you are creating.

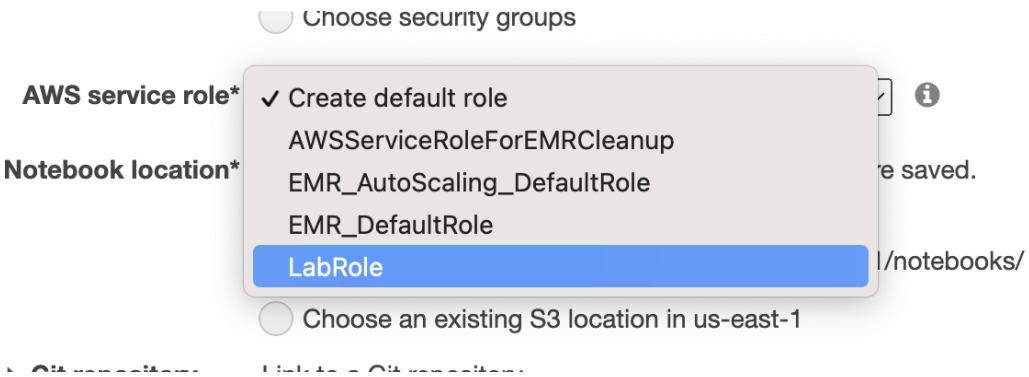


Step 4: Now, under 'Cluster', click the radio button for 'Choose an existing cluster' and then click on the 'Choose' button. As soon as you do that, a pop-up will appear, showing you the list of all currently running EMR clusters.

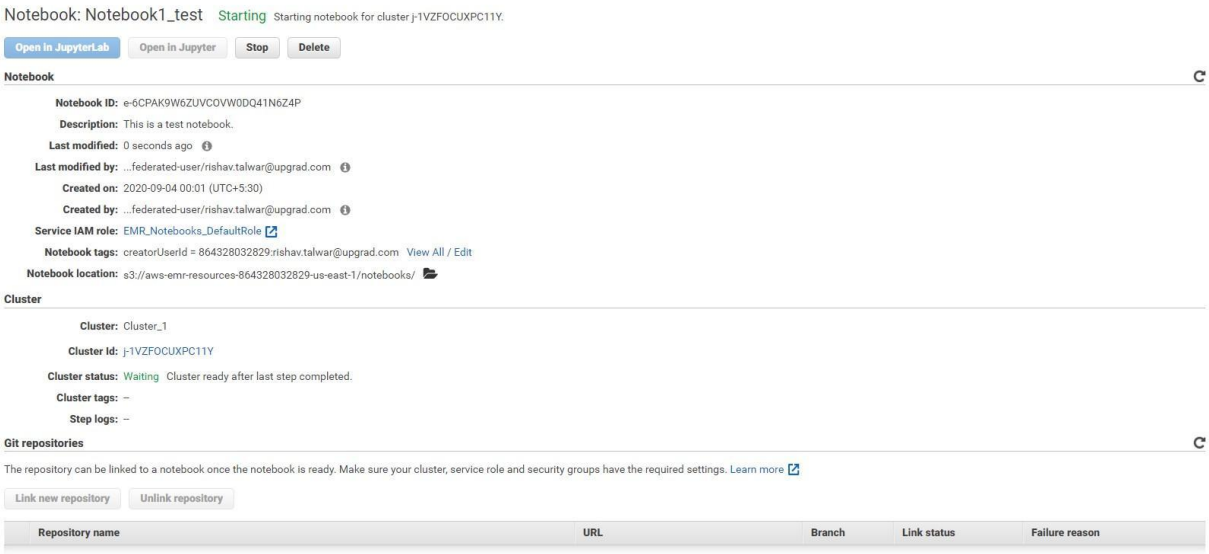


Step 5: In this step, just select the EMR cluster that you have created and then click on the 'Choose cluster' button at the bottom right corner of the page.

Step 6: Next, in AWS service role select '**LabRole**'



Step 6: After this step, you can keep the other settings as default and click on the 'Create notebook' button at the bottom right of the page. As soon as you do this, the following page will open on your screen.



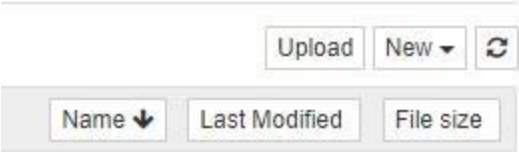
This means the Notebook has been created and is now starting. Shortly after, the notebook will show at the top with status 'Ready'.

Step 7: Finally, after the status of the Jupyter Notebook shows 'Ready', you can launch the notebook. To do this, simply click on the '**Open in Jupyter**' button.

This will open the familiar Jupyter UI. Here, you can start creating your own Jupyter Notebooks as you have done in the previous module on Spark.

Note: Please note that if you want to work with Apache Spark, you will need to set the kernel of your notebook to ""

You can also upload any Jupyter Notebook that you want easily. For this, simply click on the 'Upload' button to the top right on the Jupyter UI.



A Windows ‘Open’ dialogue box will appear. From here, you can simply find the location of your Jupyter Notebook and then click on the ‘Open’ button. After this, you will see that the name has been appended to the list of notebooks. You now need to click on the ‘Upload’ button next to your notebook file.

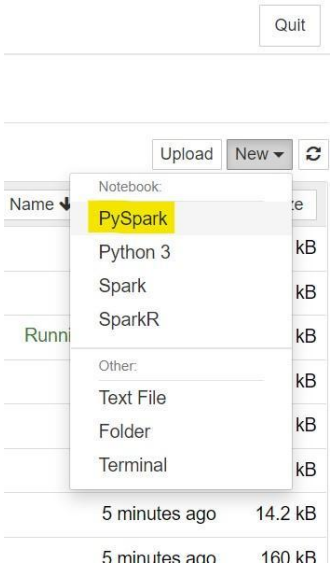
This will upload the Jupyter Notebook file to your EMR notebooks folder.

Using Jupyter Notebook with Apache Spark

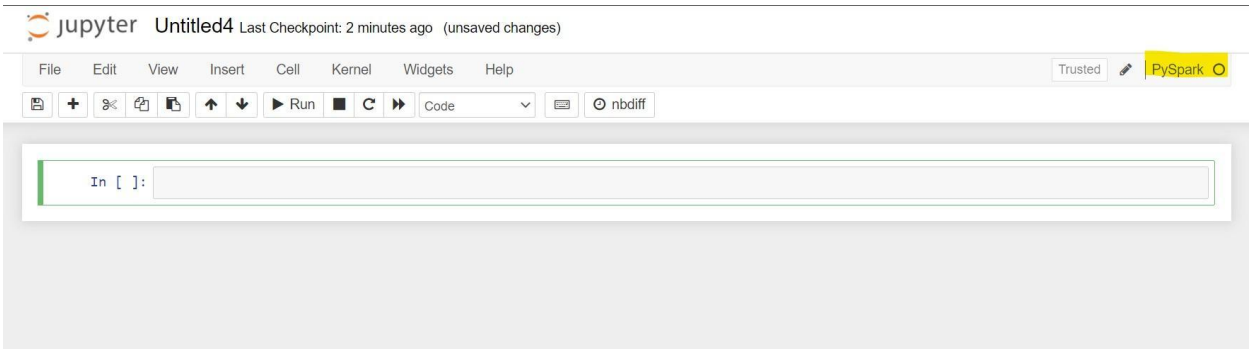
If you want to create a new Jupyter notebook to be used with Apache Spark then you need to follow these steps:

- Click on the **new button** to the top right of the Jupyter homepage.

- You will then have to select **PySpark** as your kernel in the drop-down menu. Click on PySpark

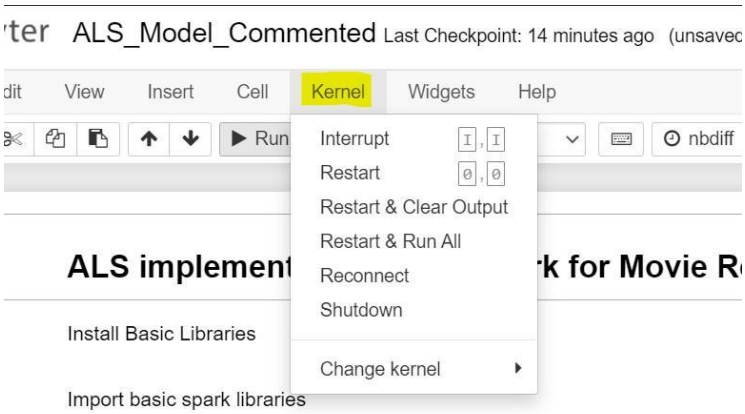


- This will then open a new window where your PySpark notebook will open up.

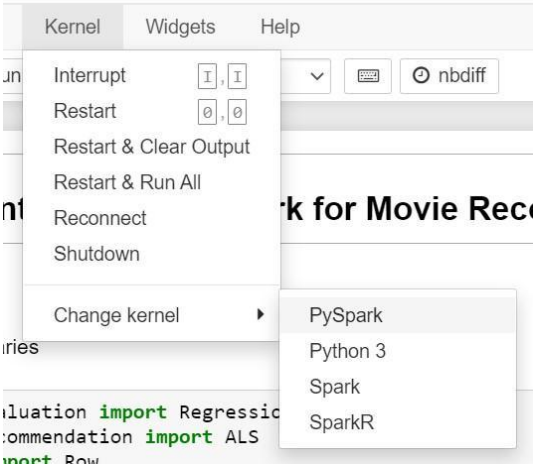


Note: Please note that if you want to work with Apache Spark on an older notebook or a notebook that you just uploaded to Jupyter, you might need to set the kernel of your notebook to "PySpark". You can do this by following these steps:

- Click on **Kernel** on the top menu of your notebook as shown in the image below.

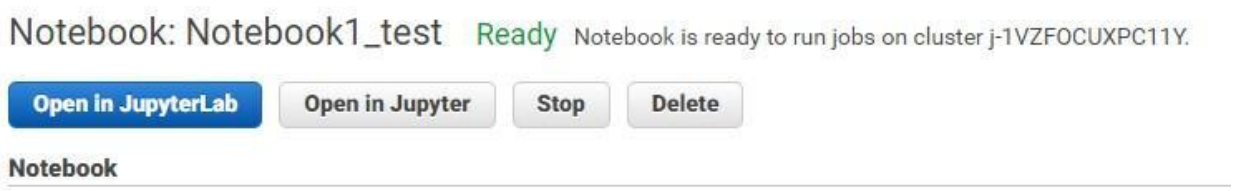


- Go to Change kernel at the end of this menu and then click on PySpark. This will change your kernel to PySpark in a few seconds. You can then start working on your PySpark Jupyter Notebook



Stopping a Jupyter Notebook

You can also stop the Notebook whenever you want by simply clicking on the **Stop** button in the Notebook UI, as shown below:



And, if you need to resume your Jupyter Notebook, then you can do so by going to the notebooks list, selecting your notebook, and clicking on the **Start** button.