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.

Amazon EMR

Clusters

Notebooks

Git repositories

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.

Notebooks

Use EMR notebooks based on Jupyter to analyze data interactively with live code, narrative text, visualizations, and more. Create independently of clusters. Standard billing for clusters and Amazon S3 apply. Learn more

| Create notebook | View details | Open in JupyterLab | Open in Jupyter | Start | Stop | Delete |

Filter: All notebooks | Filter notebooks ... | 2 notebooks (all loaded) ↻

Name
○ SparkNotebook
○ MyNewNoteBook

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

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* [   ]
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   ▼] ⓘ

Notebook location* Choose an S3 location where files for this notebook are saved.

          ⦿ Use the default S3 location
            s3://aws-emr-resources-864328032829-us-east-1/notebooks/

          ○ Choose an existing S3 location in us-east-1

▸ Git repository Link to a Git repository

▸ Tags ⓘ

* Required   Cancel [Create notebook]

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.
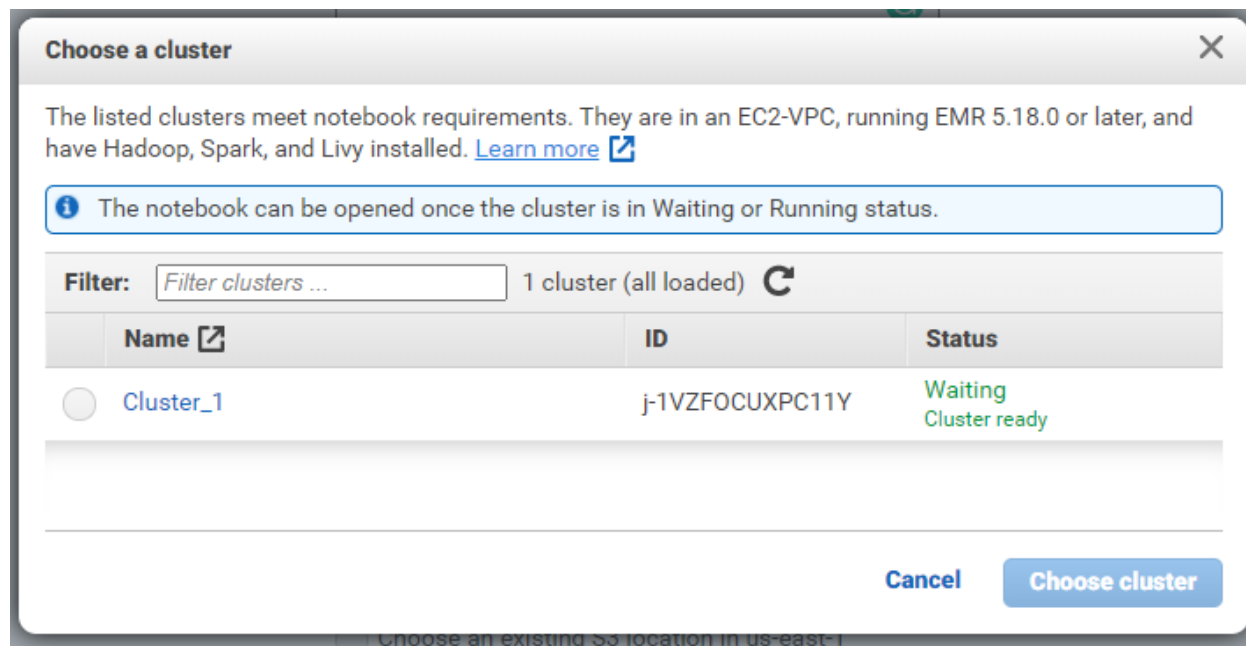
Notebook name* [Notebook1_test   ]
Names may only contain alphanumeric characters, hyphens (-), or underscores (_).

Description [This is a test notebook.   ]

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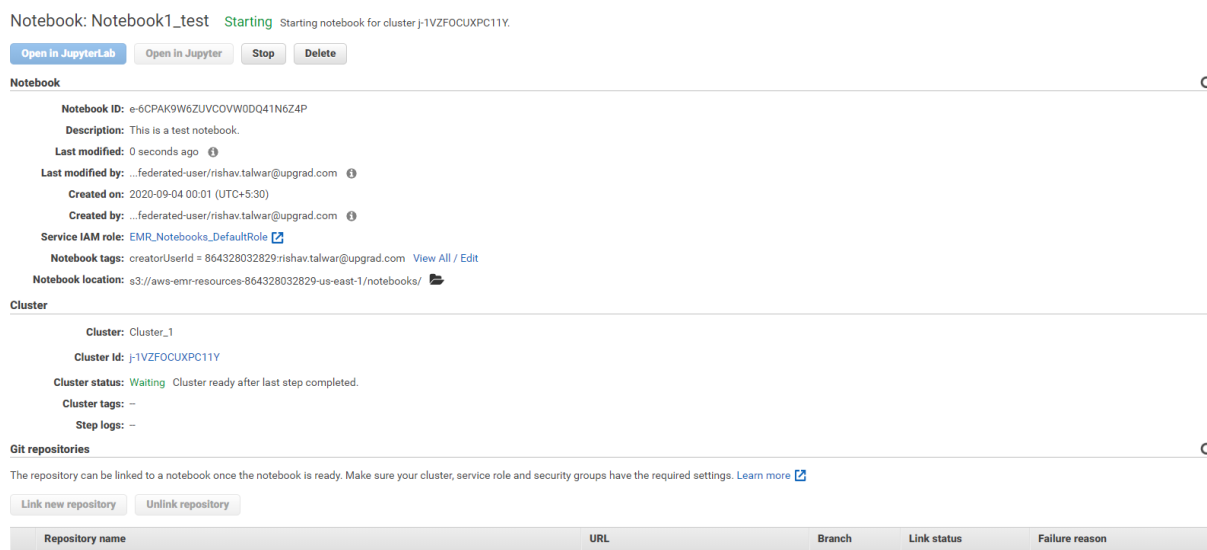
This means the Notebook has been created and is now starting. Shortly after, the notebook will show at the top with status 'Ready'.

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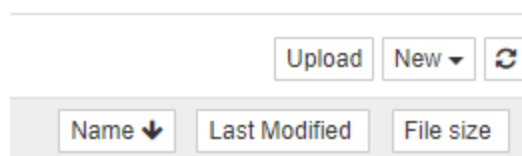
A Windows 'Open' dialogue box will appear. From here, you can simply find the location of your Jupyter Notebook and then click on the 'Open' button. After this, you will see that the name has been appended to the list of notebooks. You now need to click on the 'Upload' button next to your notebook file.
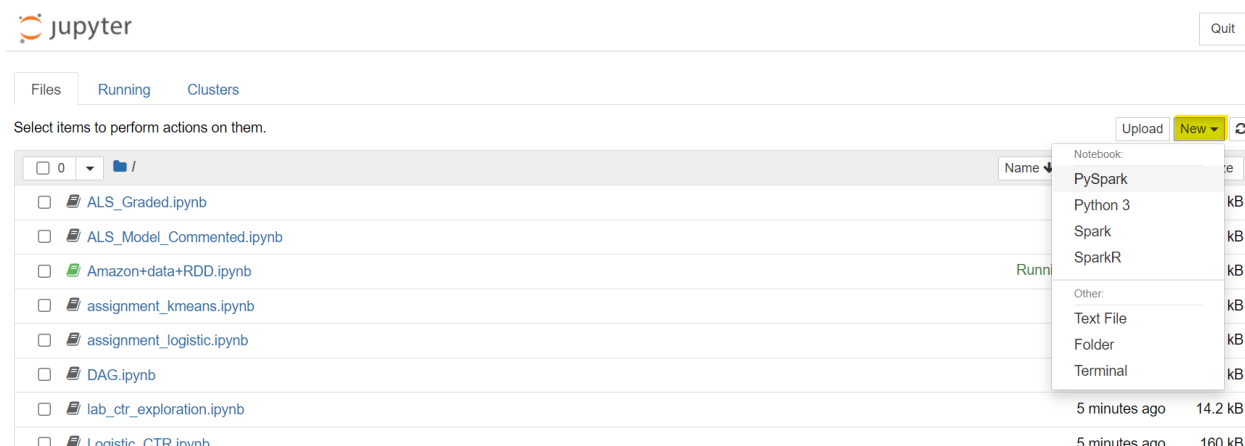

This will upload the Jupyter Notebook file to your EMR notebooks folder.

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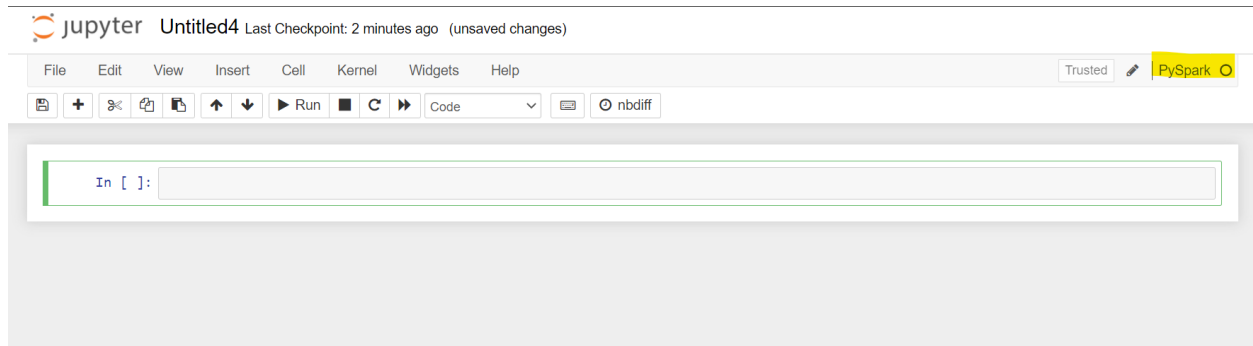
- Click on the **new button** to the top right of the Jupyter homepage.

- You will then have to select **PySpark** as your kernel in the drop-down menu. Click on PySpark
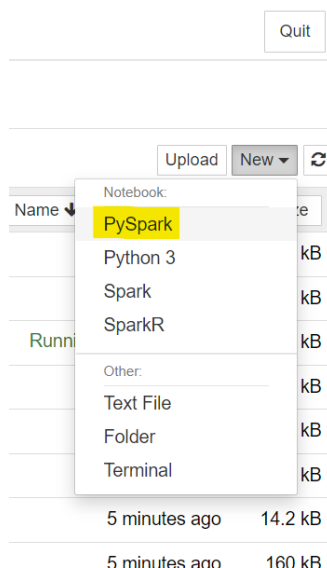
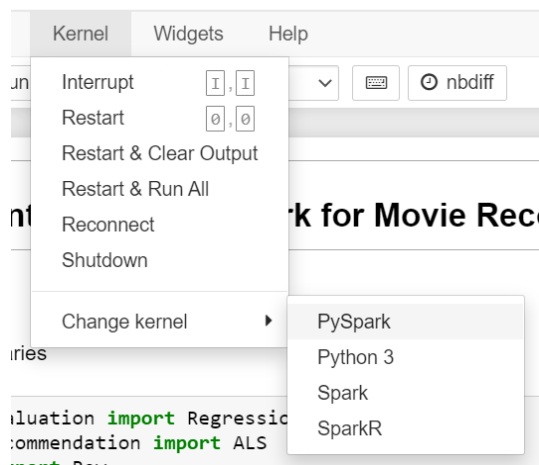● This will then open a new window where your PySpark notebook will open up.

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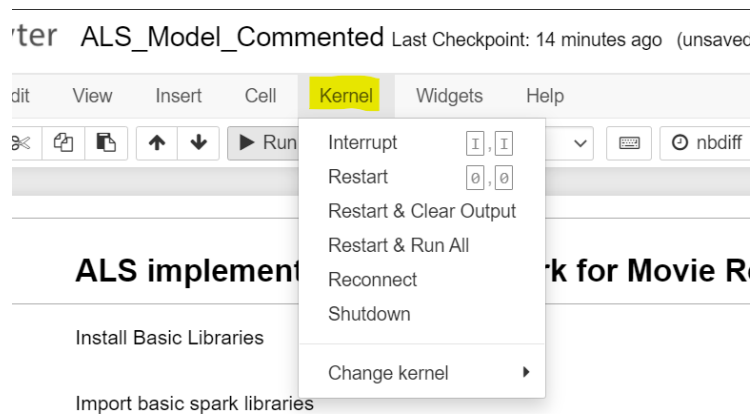
- Go to Change kernel at the end of this menu and then click on PySpark. This will change your kernel to PySpark in a few seconds. You can then start working on your PySpark Jupyter Notebook

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.