

Lab – Data Processing using EMR Jupyter Notebooks

Step 1: Login to AWS Console and go to EMR

Step 2: Click on create a cluster and give a name or use default.
Select Spark under Applications and select m4.large machines(3).
Select an EMR version less than 30. Click create cluster and wait till the status changes to **Waiting**

Summary

ID: j-BKFMJCBIVQ85

Creation date: 2021-08-20 22:07 (UTC+5:30)

Elapsed time: 10 minutes

After last step completes: Cluster waits

Termination protection: Off [Change](#)

Tags: -- [View All / Edit](#)

Master public DNS:

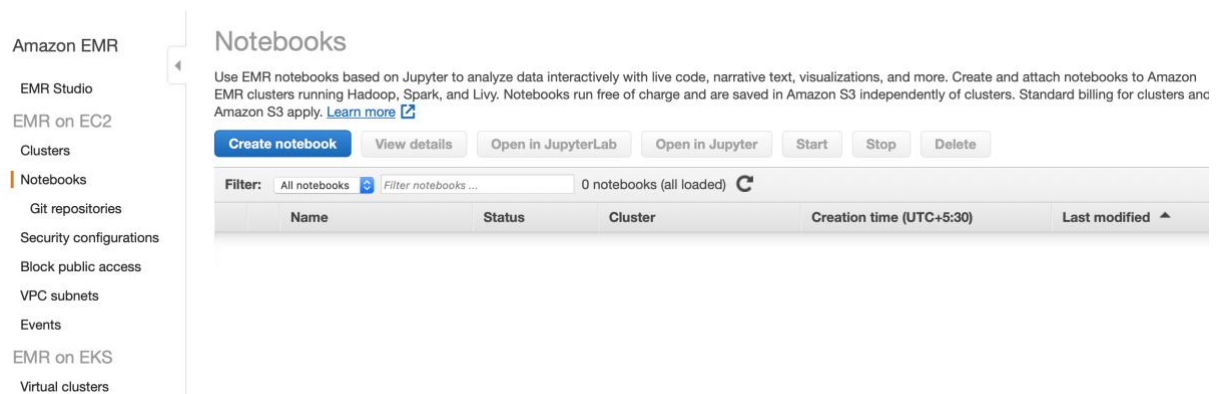
ec2-3-109-155-106.ap-south-1.compute.amazonaws.com 

[Connect to the Master Node Using SSH](#)

Configuration details

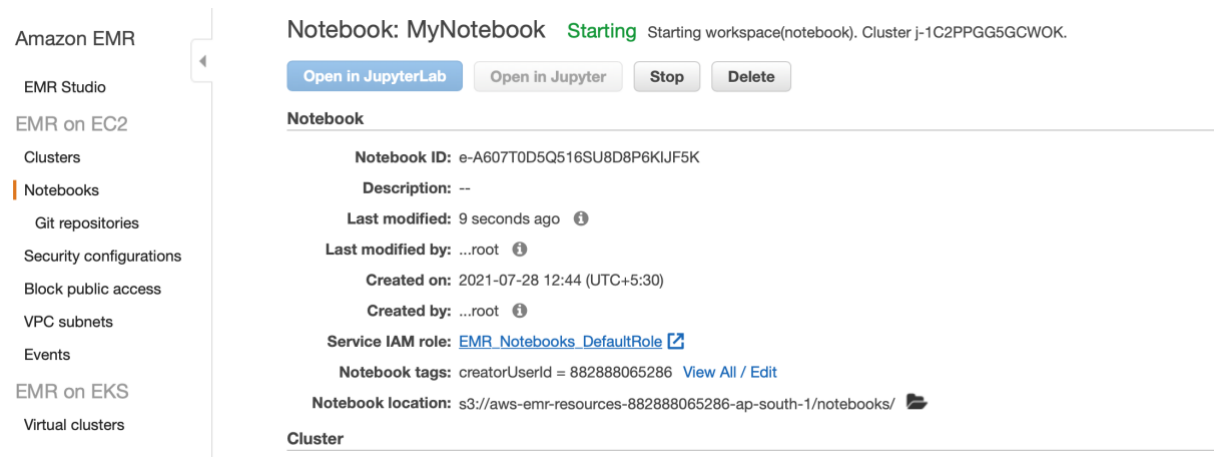
Release label: emr-5.29.0

Step 3: Once the cluster is created Click on the Notebooks on the left side



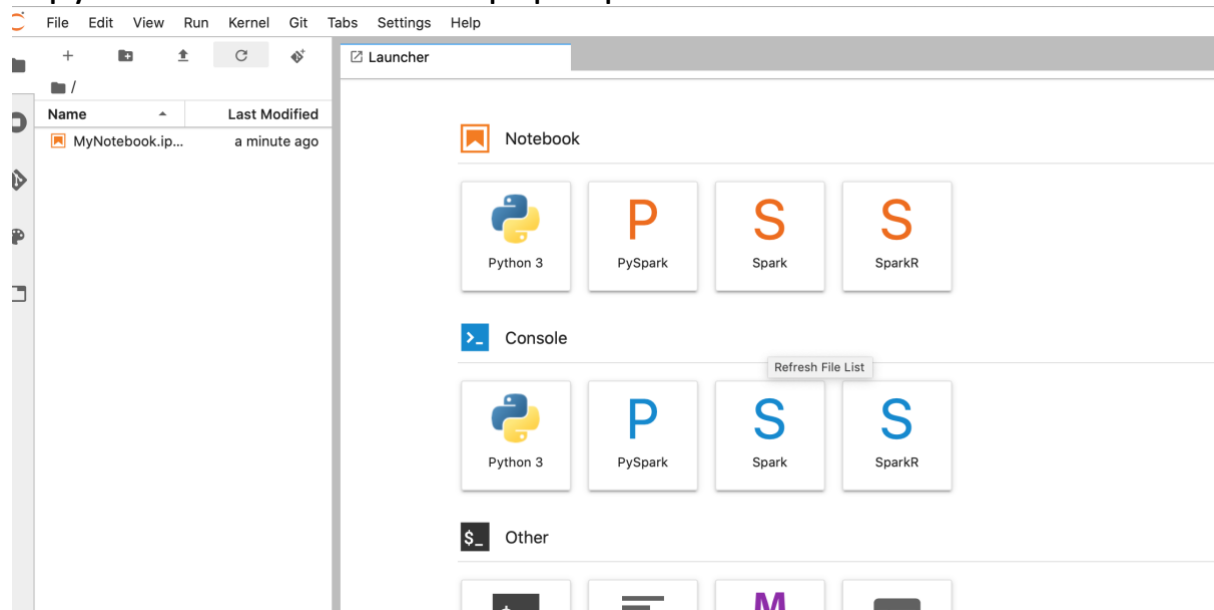
Step 4: Click Create Notebook

Step 5: Give a name to notebook and select the created cluster. Click Create Notebook



The screenshot shows the Amazon EMR console interface. On the left is a navigation menu with options like 'Amazon EMR', 'EMR Studio', 'EMR on EC2', 'Clusters', 'Notebooks', 'Git repositories', 'Security configurations', 'Block public access', 'VPC subnets', 'Events', 'EMR on EKS', and 'Virtual clusters'. The 'Notebooks' section is selected. The main area displays details for a notebook named 'MyNotebook'. At the top, it says 'Notebook: MyNotebook' with a green 'Starting' status and a sub-note 'Starting workspace(notebook). Cluster j-1C2PPGG5GCWOK.' Below this are four buttons: 'Open in JupyterLab' (highlighted in blue), 'Open in Jupyter', 'Stop', and 'Delete'. A section titled 'Notebook' contains the following information: 'Notebook ID: e-A607T0D5Q516SU8D8P6KIJF5K', 'Description: --', 'Last modified: 9 seconds ago', 'Last modified by: ...root', 'Created on: 2021-07-28 12:44 (UTC+5:30)', 'Created by: ...root', 'Service IAM role: EMR_Notebooks_DefaultRole', 'Notebook tags: creatorUserId = 882888065286', and 'Notebook location: s3://aws-emr-resources-882888065286-ap-south-1/notebooks/'. A 'Cluster' section is partially visible at the bottom.

Step 6: Once the status changes to Ready, Click on Open in JupyterLab. A new window pops up



The screenshot shows the JupyterLab interface. The top menu bar includes 'File', 'Edit', 'View', 'Run', 'Kernel', 'Git', 'Tabs', 'Settings', and 'Help'. Below the menu is a toolbar with icons for creating new notebooks, opening recent notebooks, and refreshing. The left sidebar shows a file explorer with a table of notebooks. The main area is titled 'Launcher' and displays a grid of notebook icons for 'Python 3', 'PySpark', 'Spark', and 'SparkR'. Below this is a 'Console' section with similar icons. At the bottom, there is an 'Other' section with icons for a terminal, a file explorer, and a markdown editor.

Step 7: Select PySpark and you should see an editor. Enter the following code and hit run (Shift+Enter)

```
lines =  
spark.read.text("s3://mys3kinesisfirehosebucket/2021/07/20/04/*").rdd  
print(lines.take(10))
```

