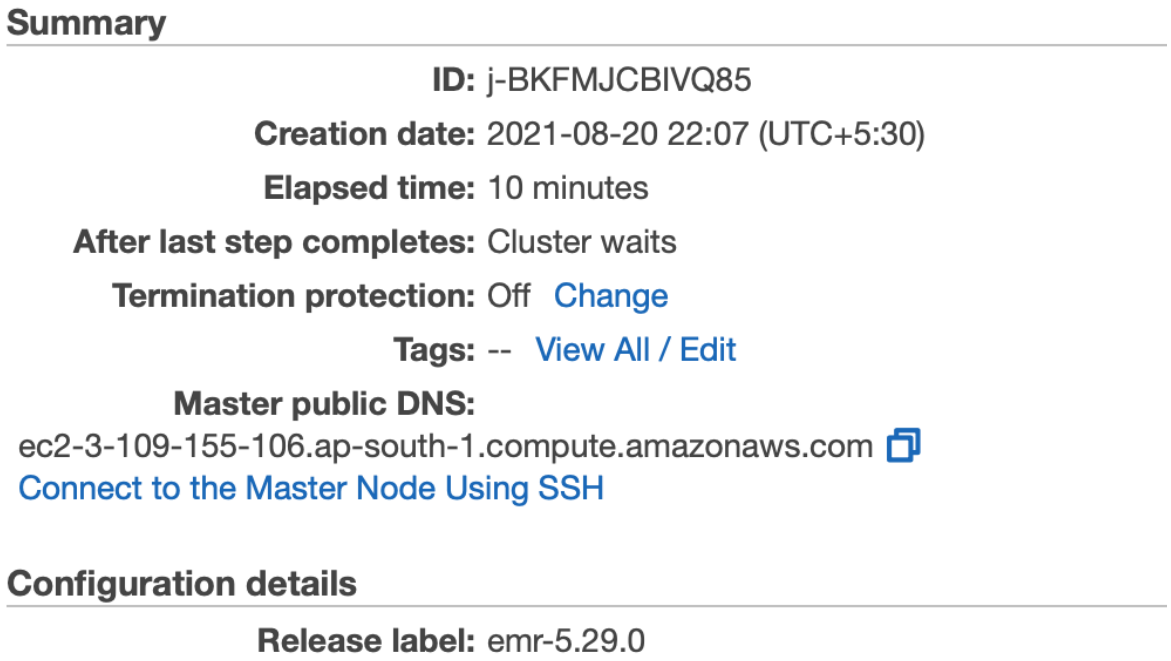
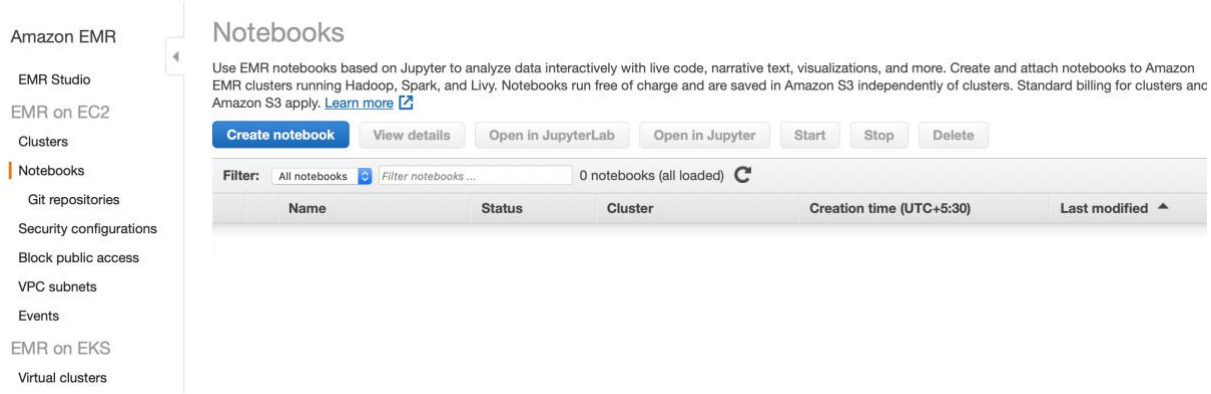
Lab – Data Processing using EMR Jupyter Notebooks

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**

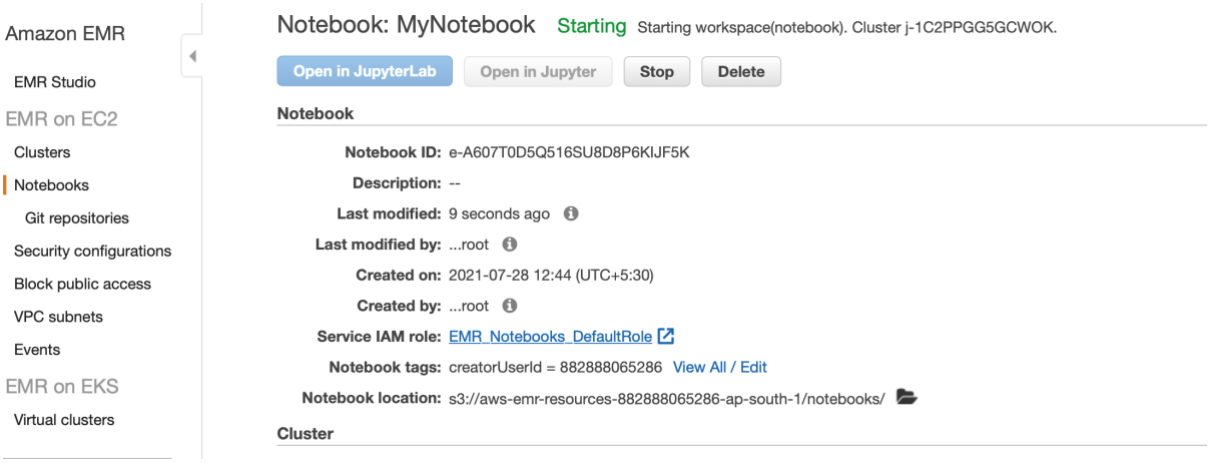


Step 3: Once the cluster is created Click on the Notebooks on the left side

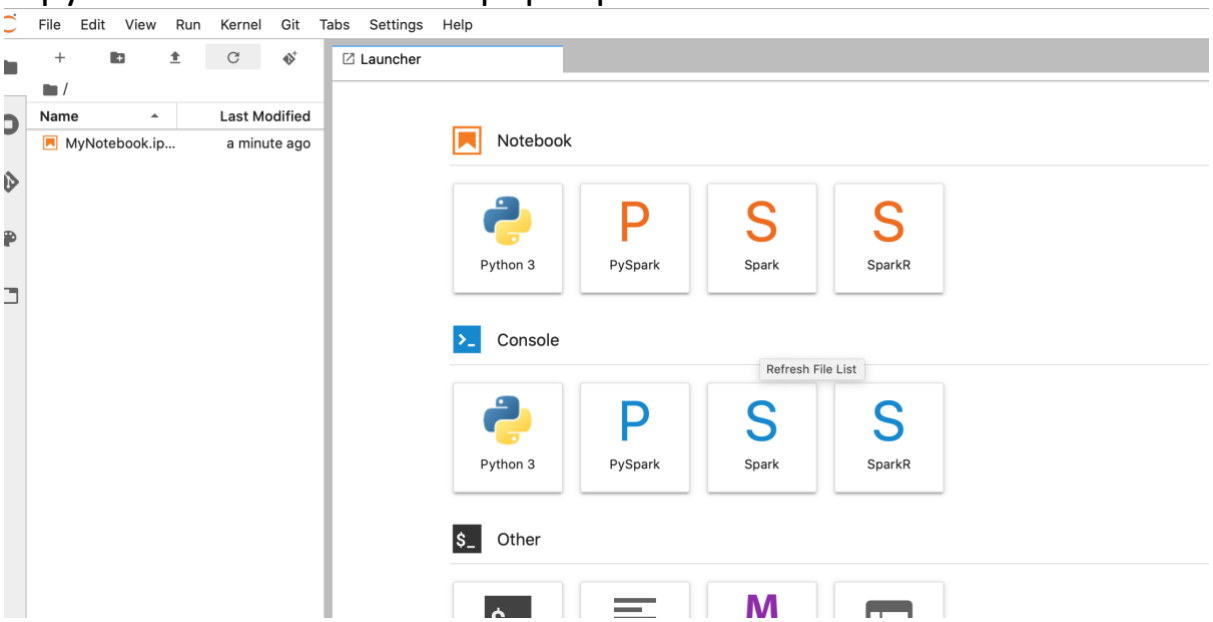


Step 4: Click Create Notebook

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



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



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))
```