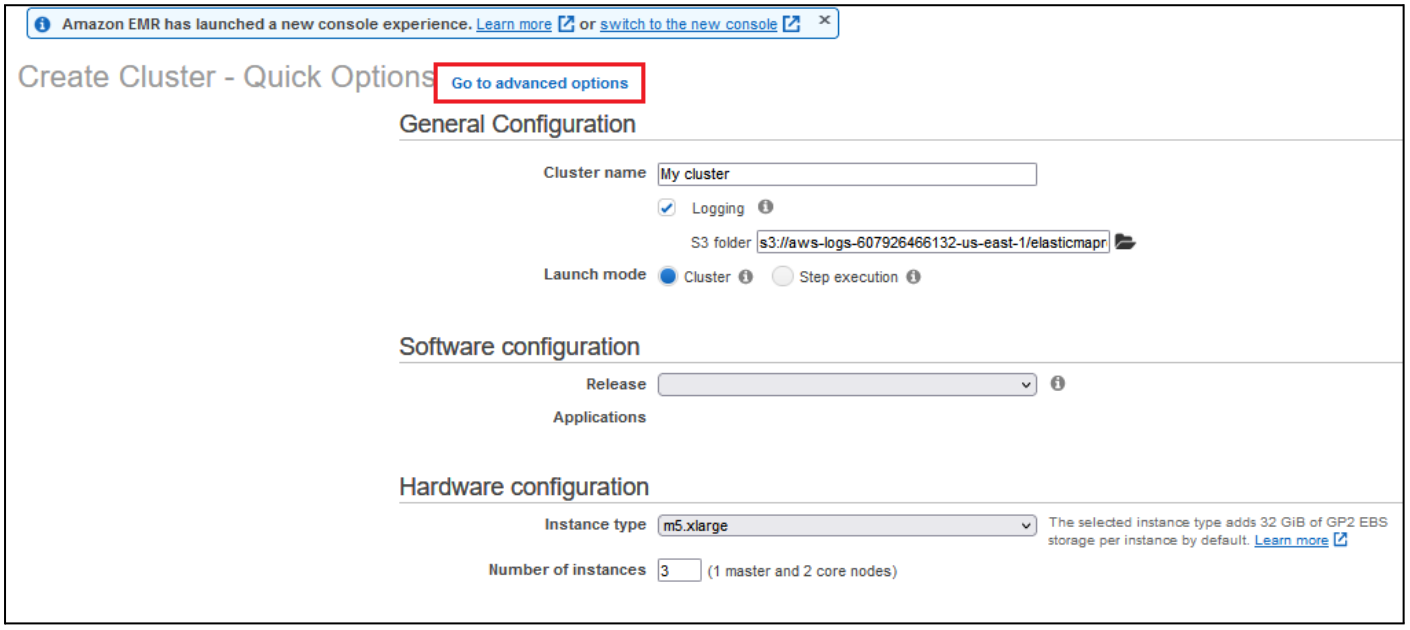
Creating a Jupyter Notebook on an EMR Cluster

This document contains the steps to work with Jupyter Notebooks and Apache Spark in EMR clusters.

EMR Cluster Creation:

Navigate to the advanced options when creating the EMR cluster.

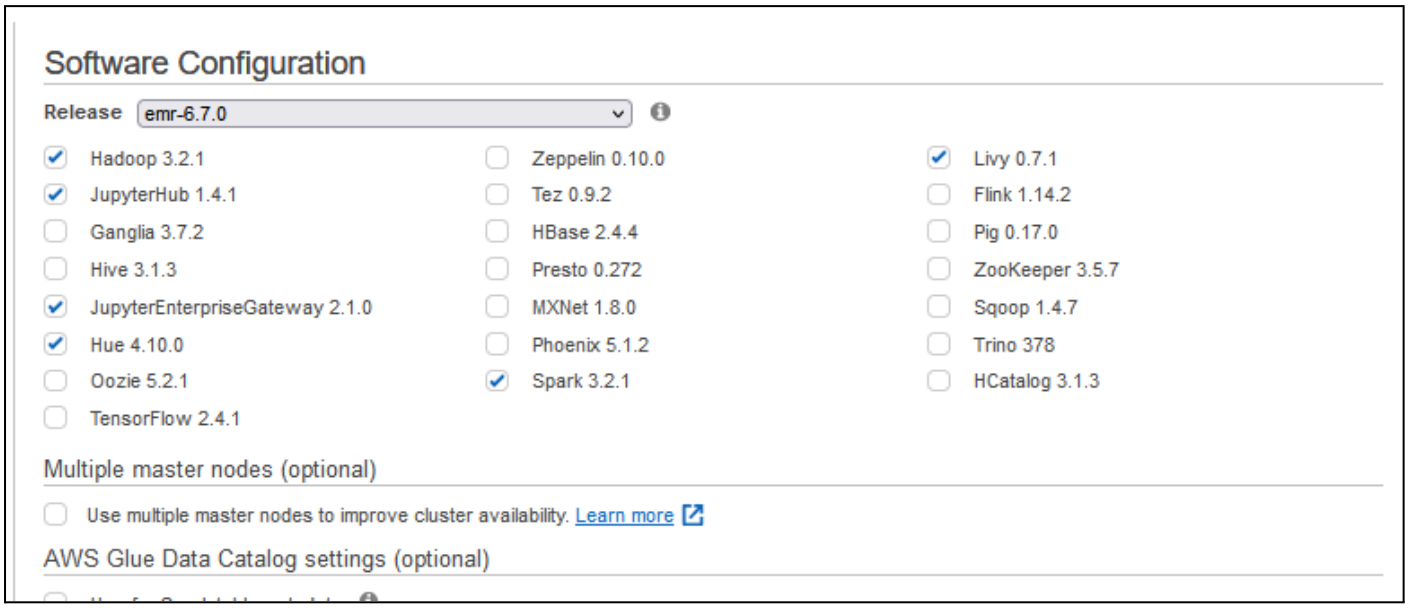


Select the following services while setting up the EMR cluster.

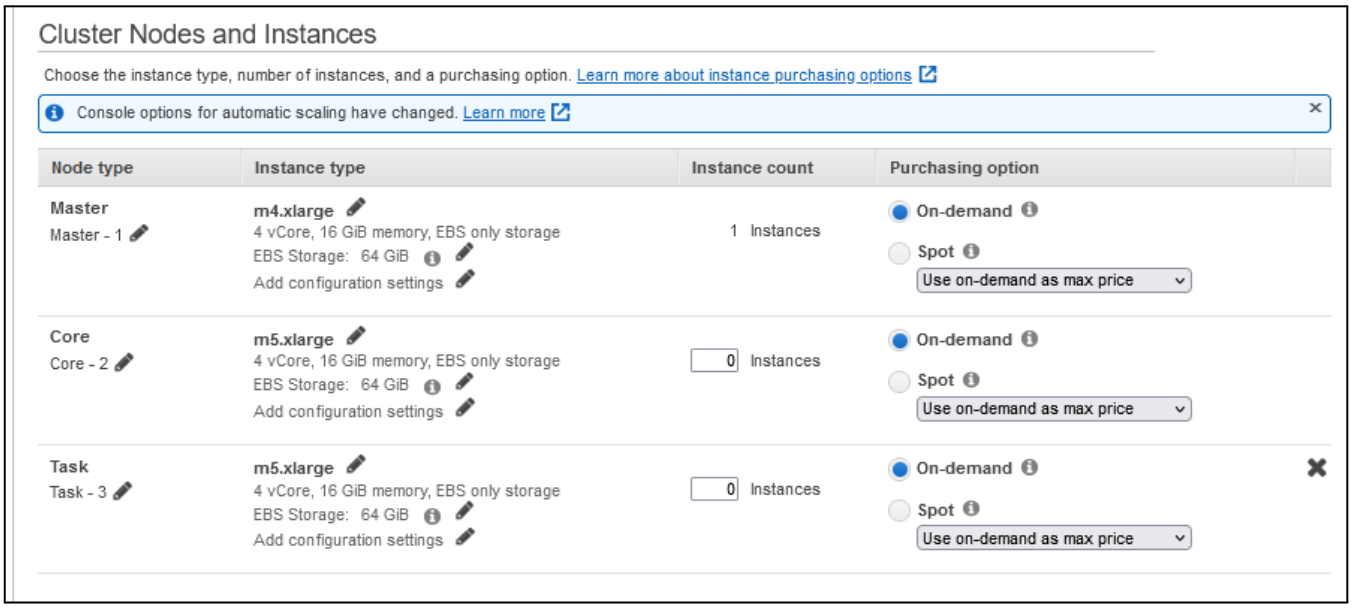
- Hadoop
- JupyterHub
- Jupyter Enterprise Gateway
- Hue
- Spark
- Livy

NOTE: In this documentation, the Jupyter service has been validated against the EMR version 6.7.0. Here, the services chosen are Hue, Jupyter, JupyterEnterpriseGateway, Livy, and Spark services for accessing Jupyter Notebooks in the EMR cluster. Depending on the EMR version, make sure to select the appropriate services.

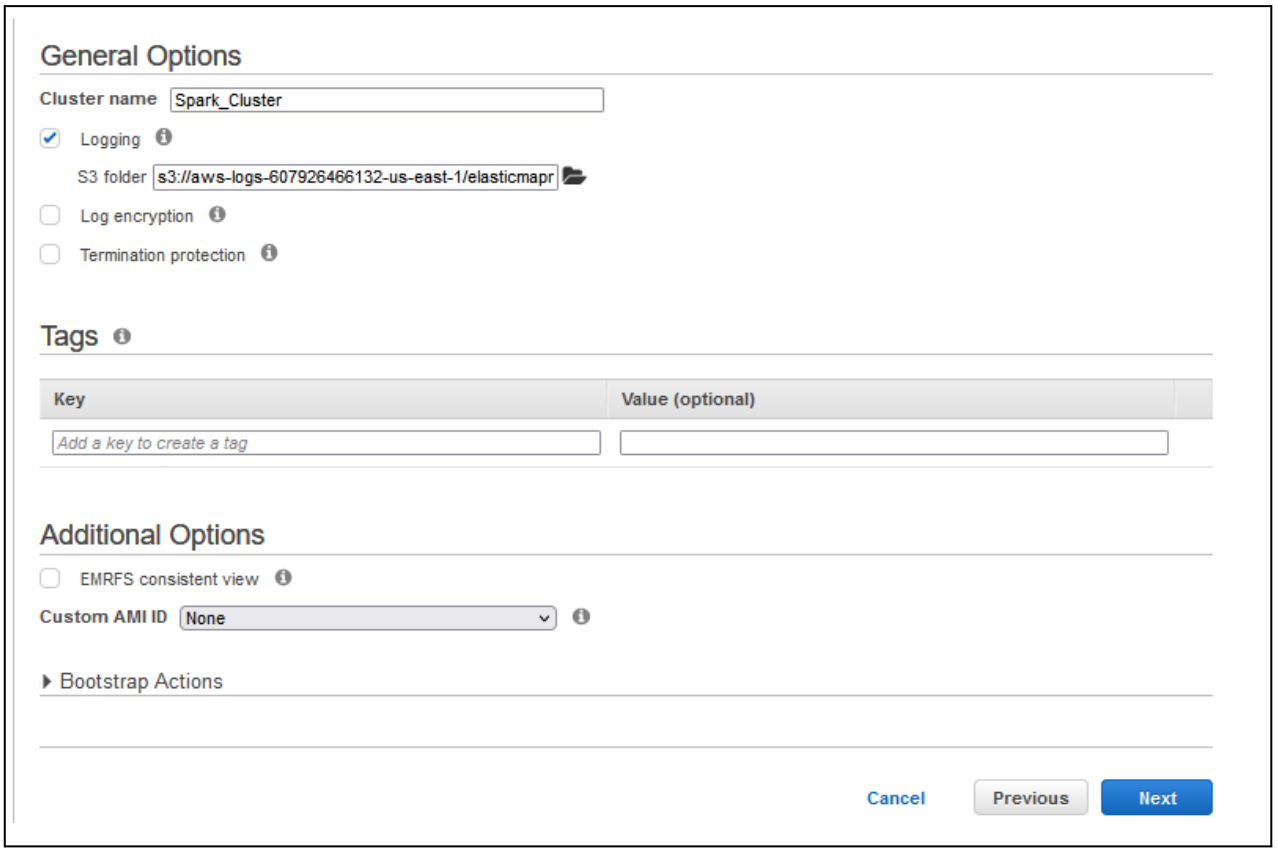
NOTE: Make sure that the Spark version you're selecting is greater than 3.2.0



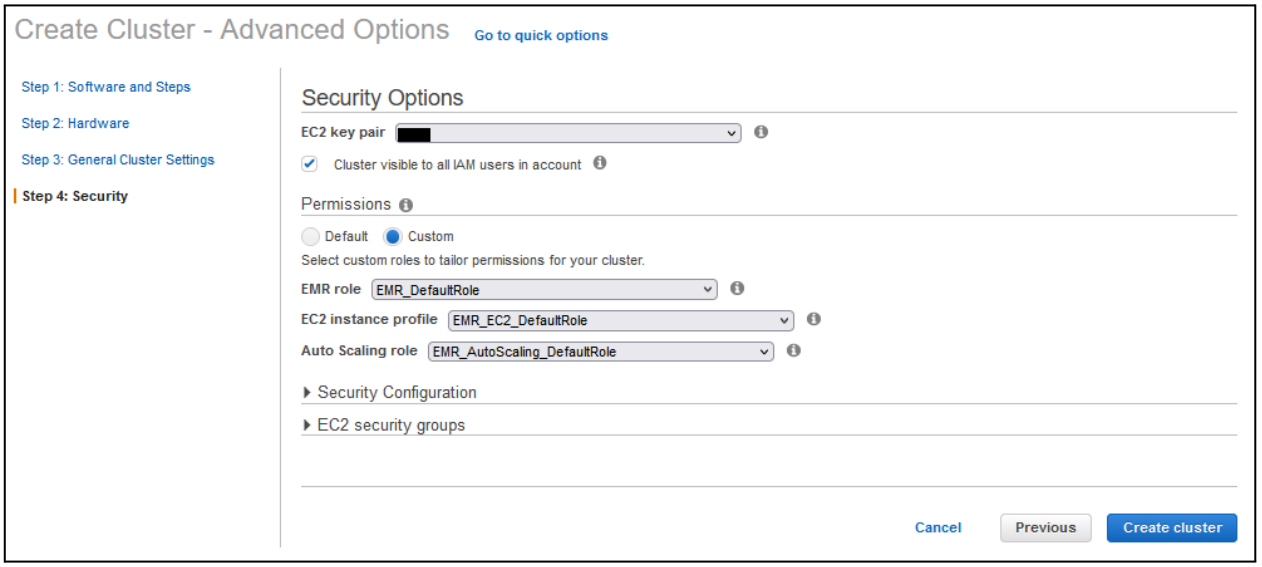
Choose the master node configuration as **m4.xlarge**.



Proceed to set up the EMR cluster.

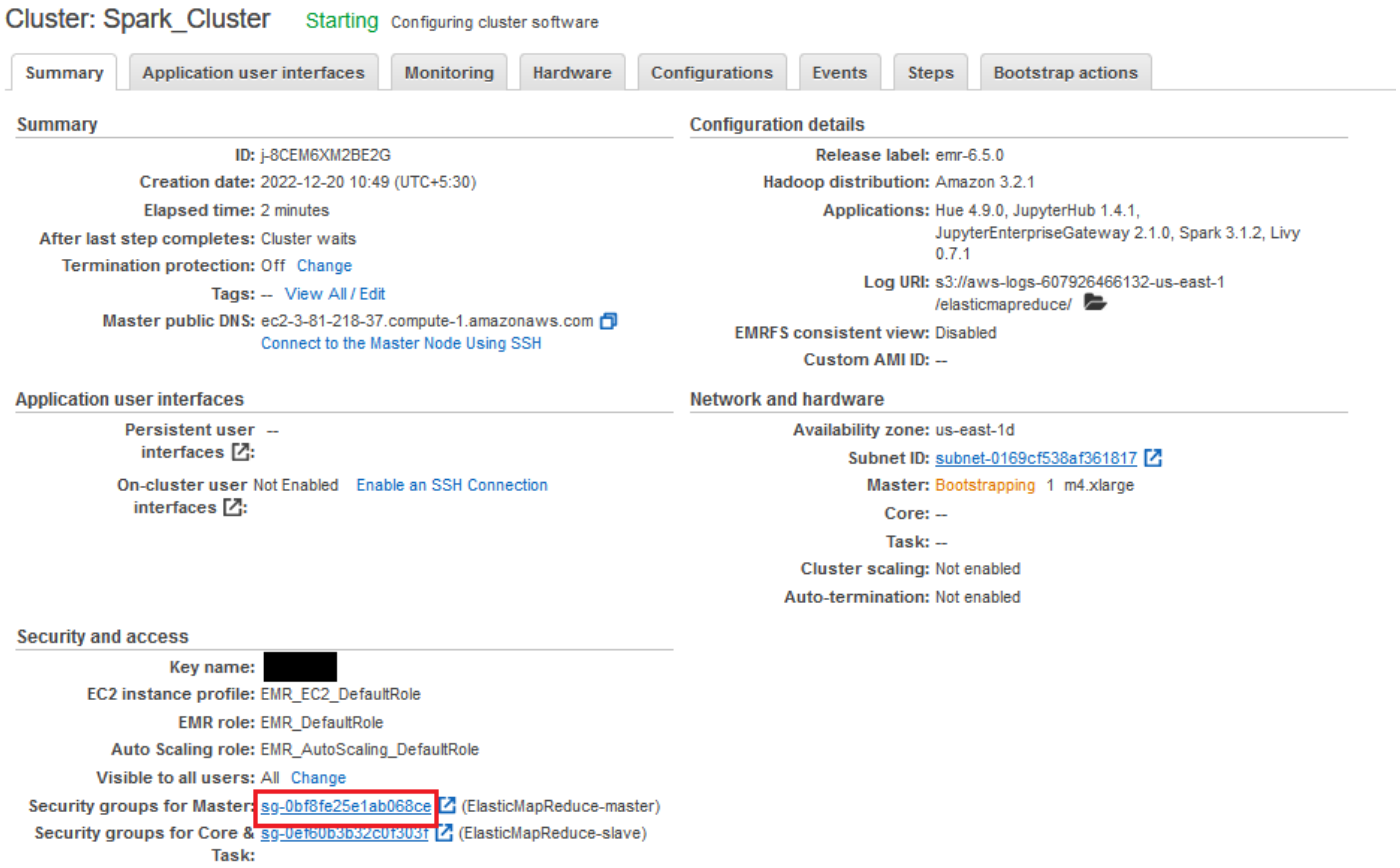


Select the EC2 key pair and click on **Create Cluster**.

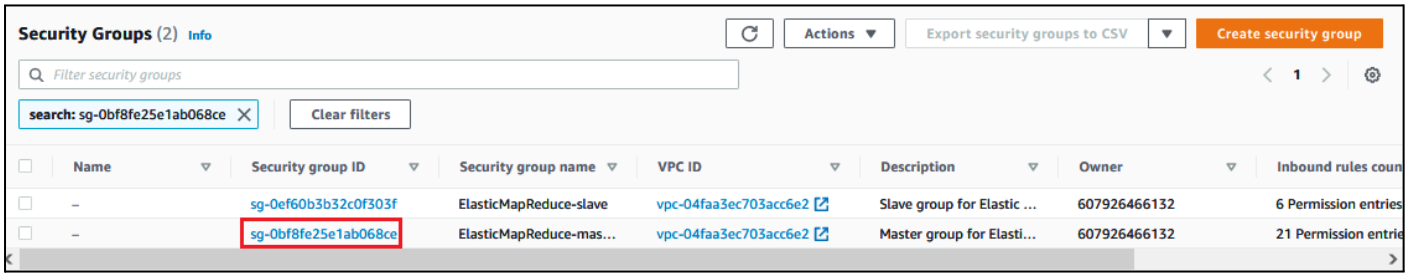


Security Groups:

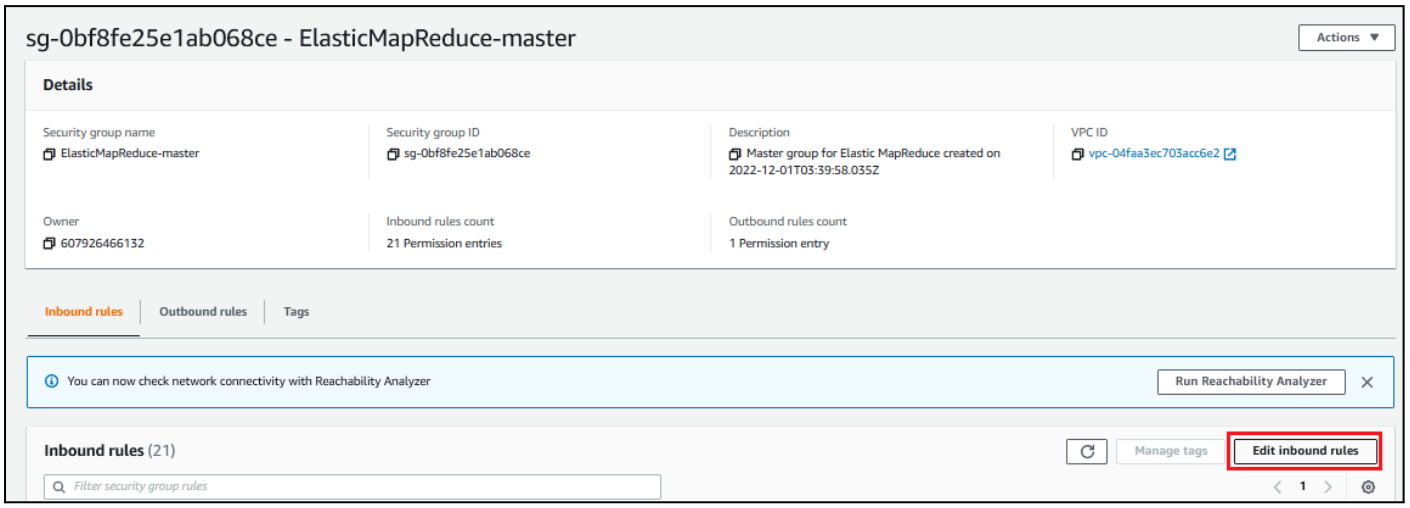
Click on the security groups of the master node in the EMR cluster page.



Click on the Security Group



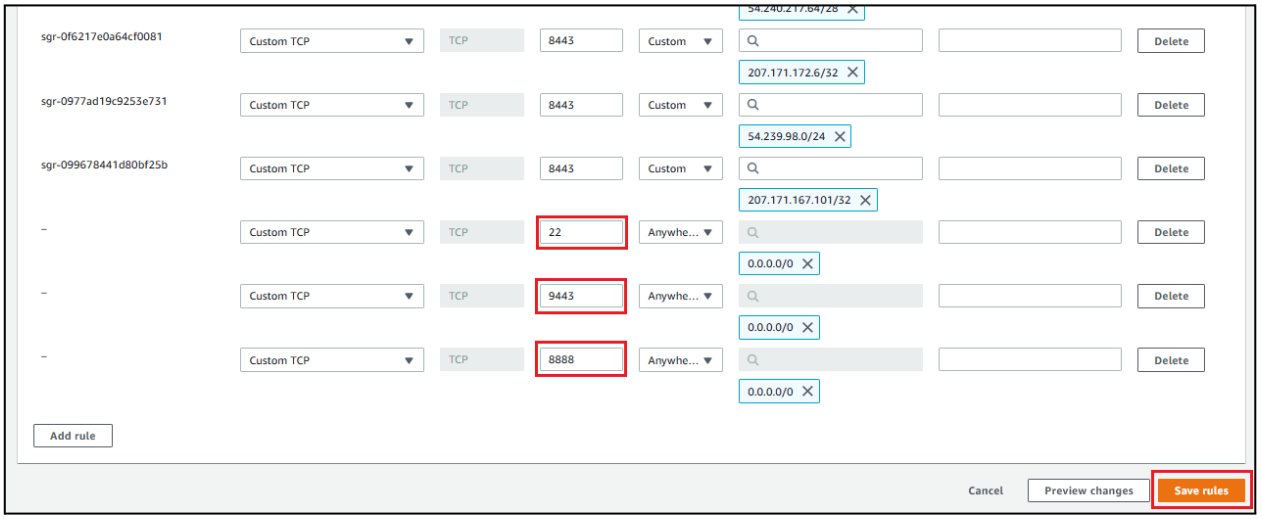
Click on Edit Inbound Rules

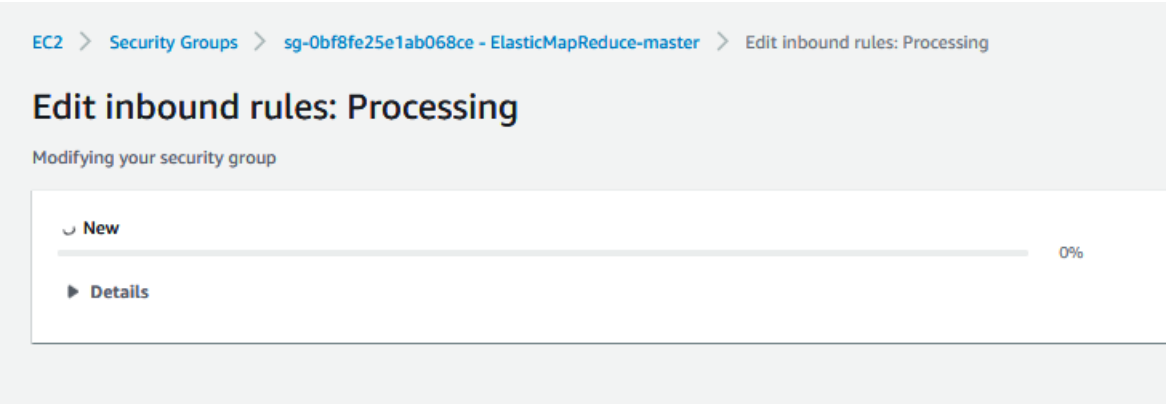


Add the following ports to the inbound rules 22, 9443, 8888.
The ports correspond to the application user interfaces of the services installed on the EMR cluster.
The values to be entered are the following

Type: Custom TCP
Protocol: TCP
Port: Enter the port number to be added
Source: Anywhere-IPv4

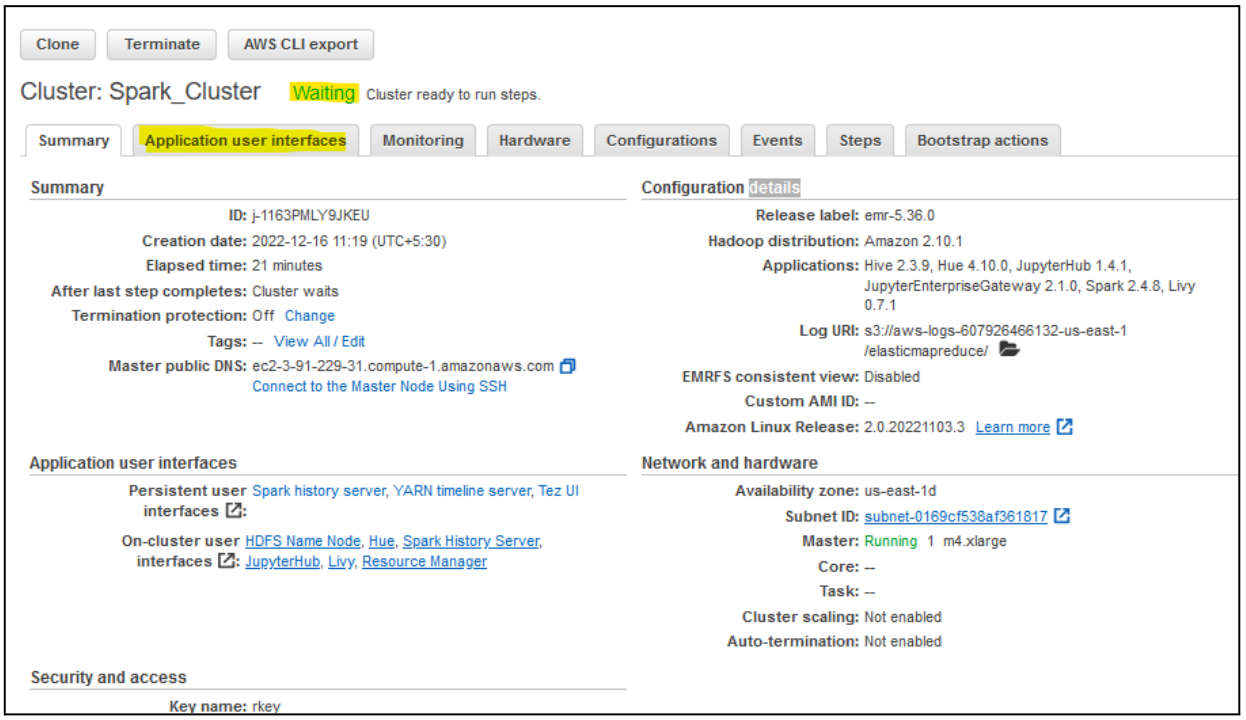
Click on **Save Rules** once done.
Note: The Port 22 is used for SSHing into the EMR cluster and may already be added in the security groups. You can then proceed to add the other ports.



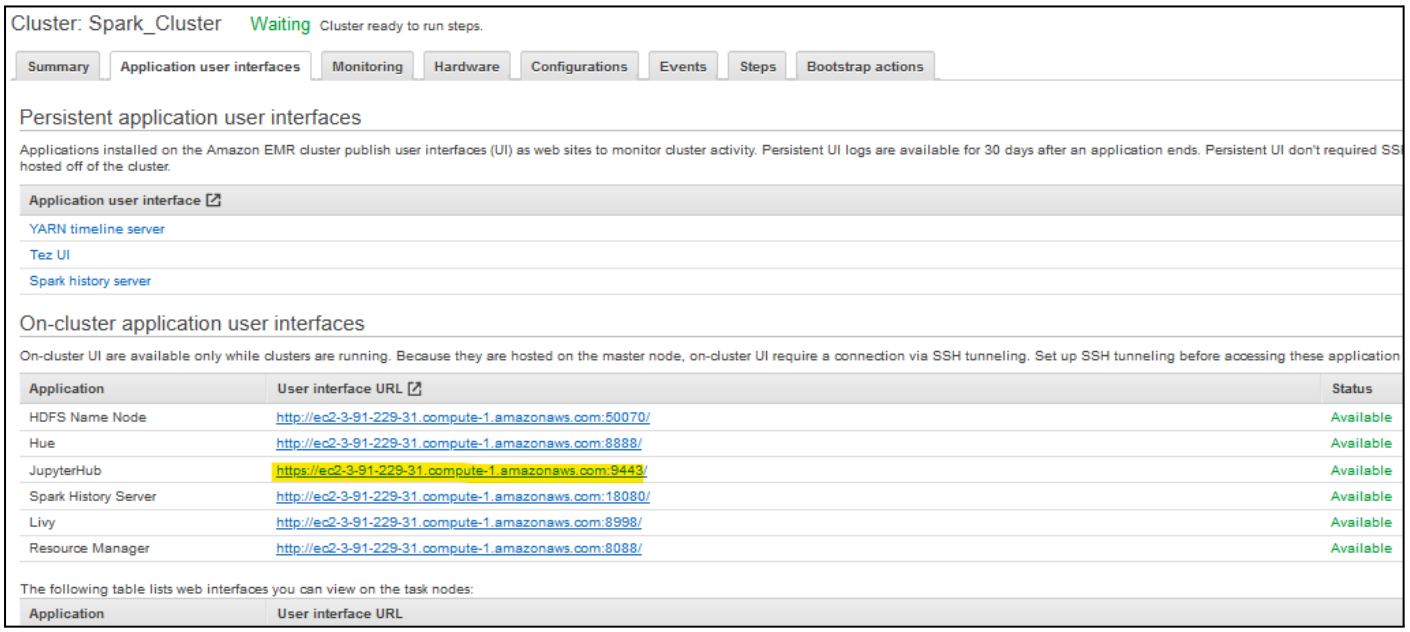


Accessing Jupyter Service:

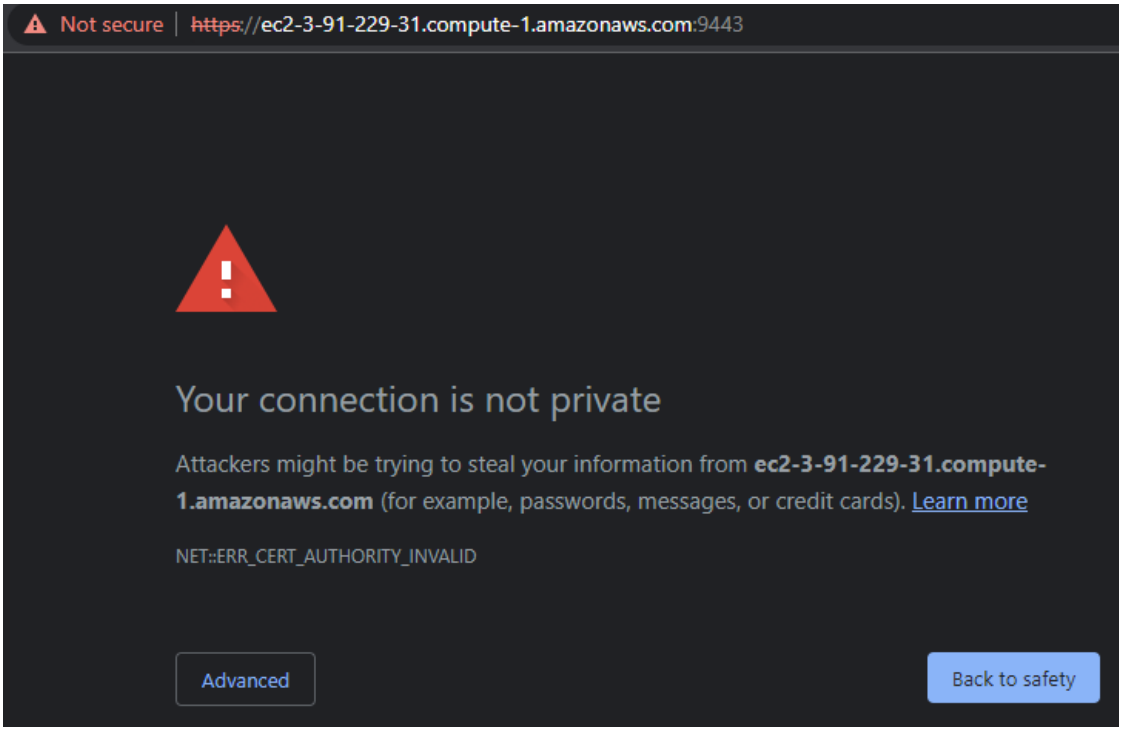
Once the cluster enters the waiting phase, navigate to the "**Application user interfaces**" tab.



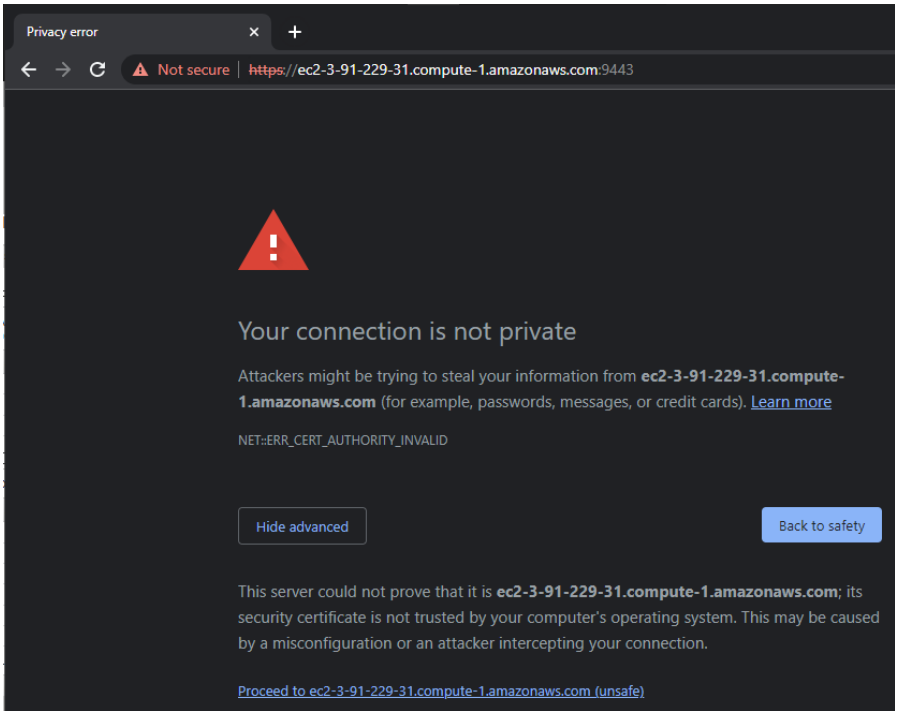
Here, you'll be able to access the application interfaces of the various services that you've installed on to your EMR cluster. We'll be using this to access the Jupyter service that was installed.
Click on the JupyterHub link.
NOTE: For this step to work, make sure that you've opened the corresponding port in the security group settings of the EMR cluster.



You may receive the following warning message when you try opening the link.



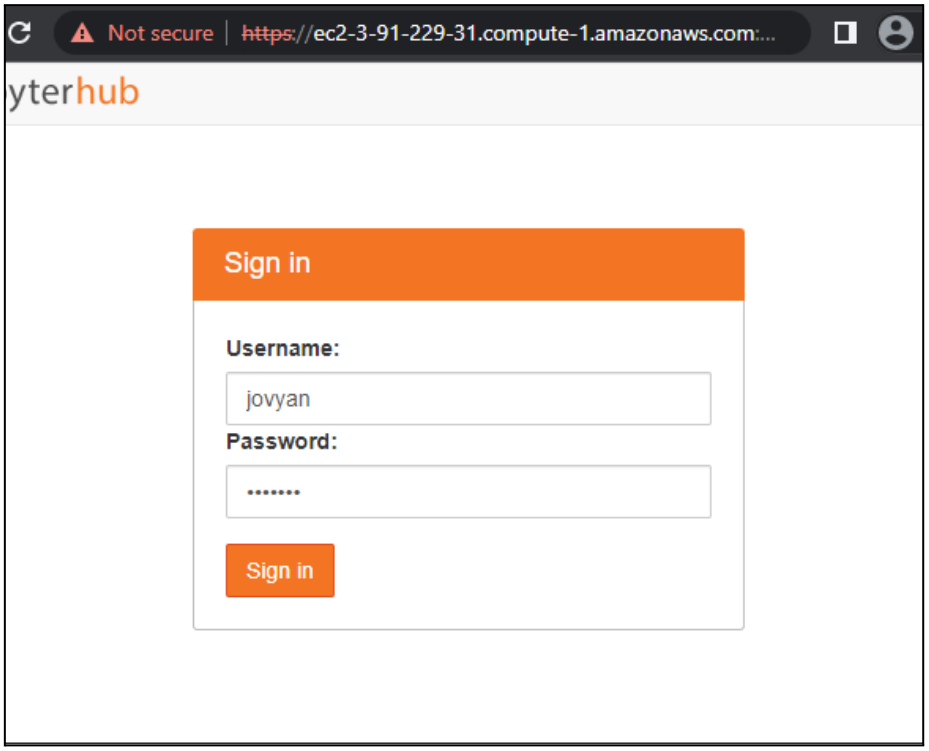
Click on **'Advanced'** and proceed to the link.



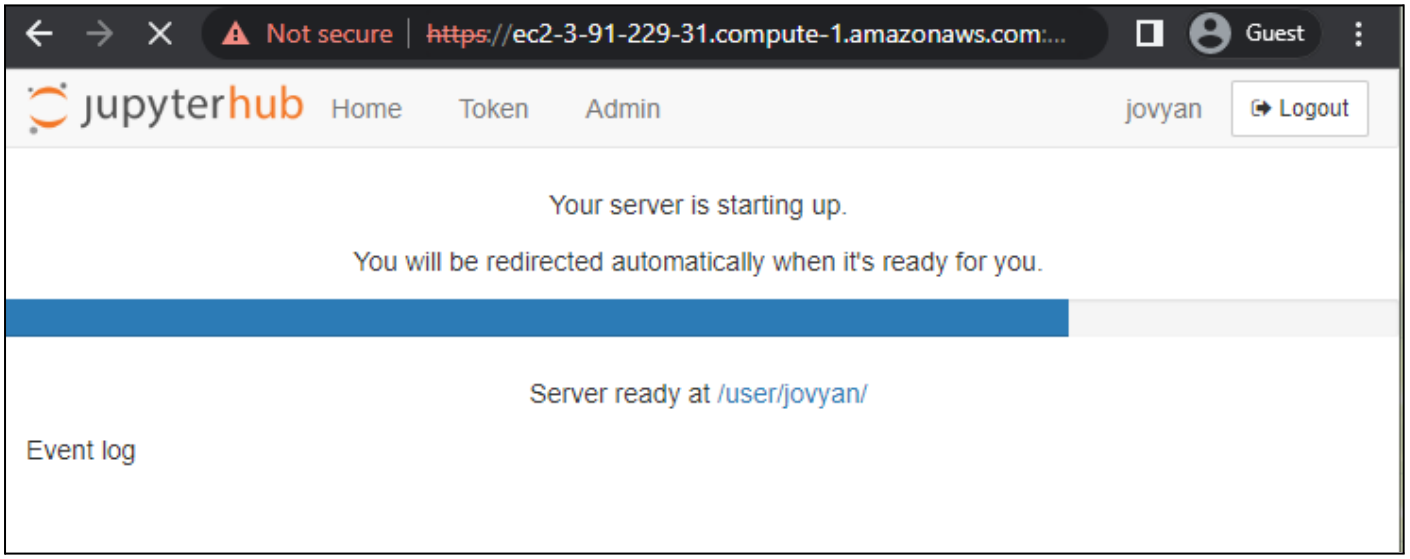
Enter the following credentials once the login page appears and click on the **Sign In** button.

Username: jovyan

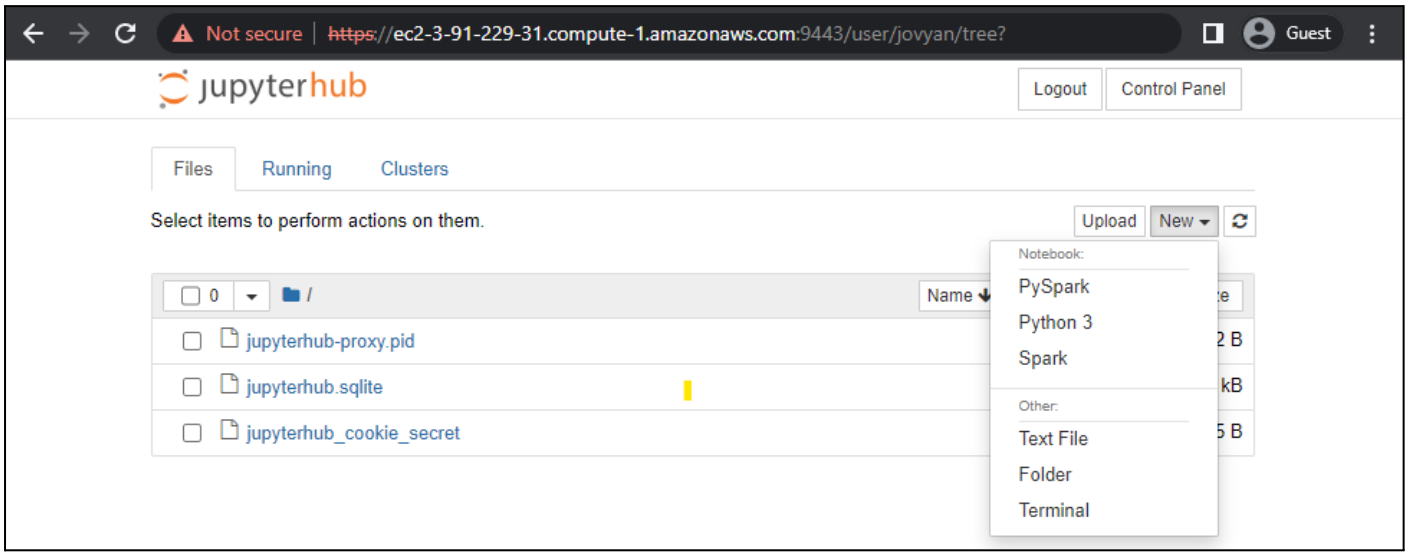
Password: jupyter



The following screen shows up. You'll now be able to access the JupyterHub service via the port on the EMR cluster.



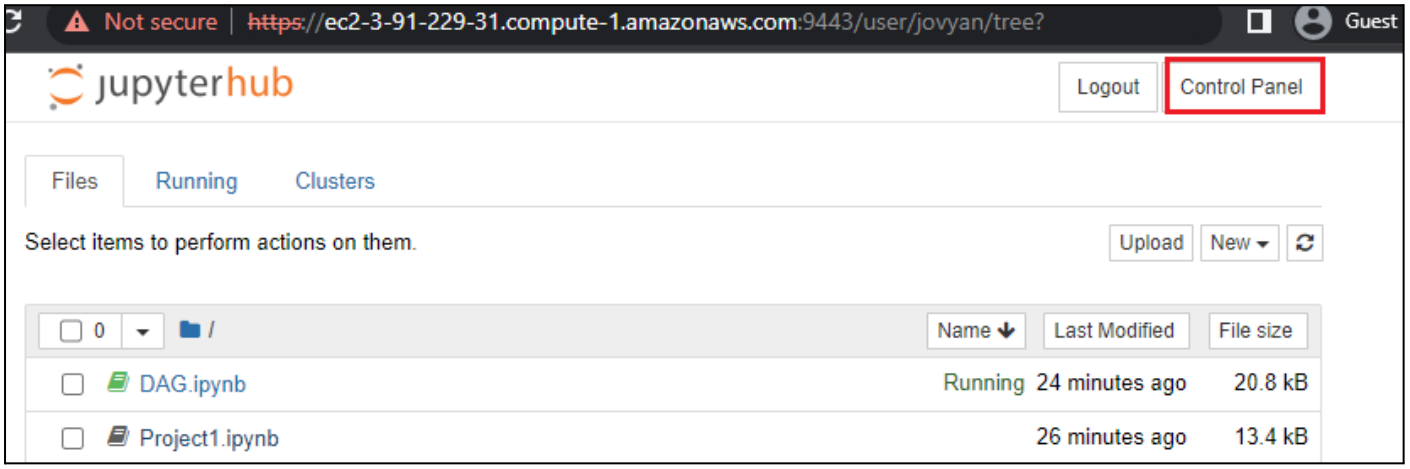
Click on the **Upload** button to upload the notebooks or the **New** button to create a new PySpark notebook.



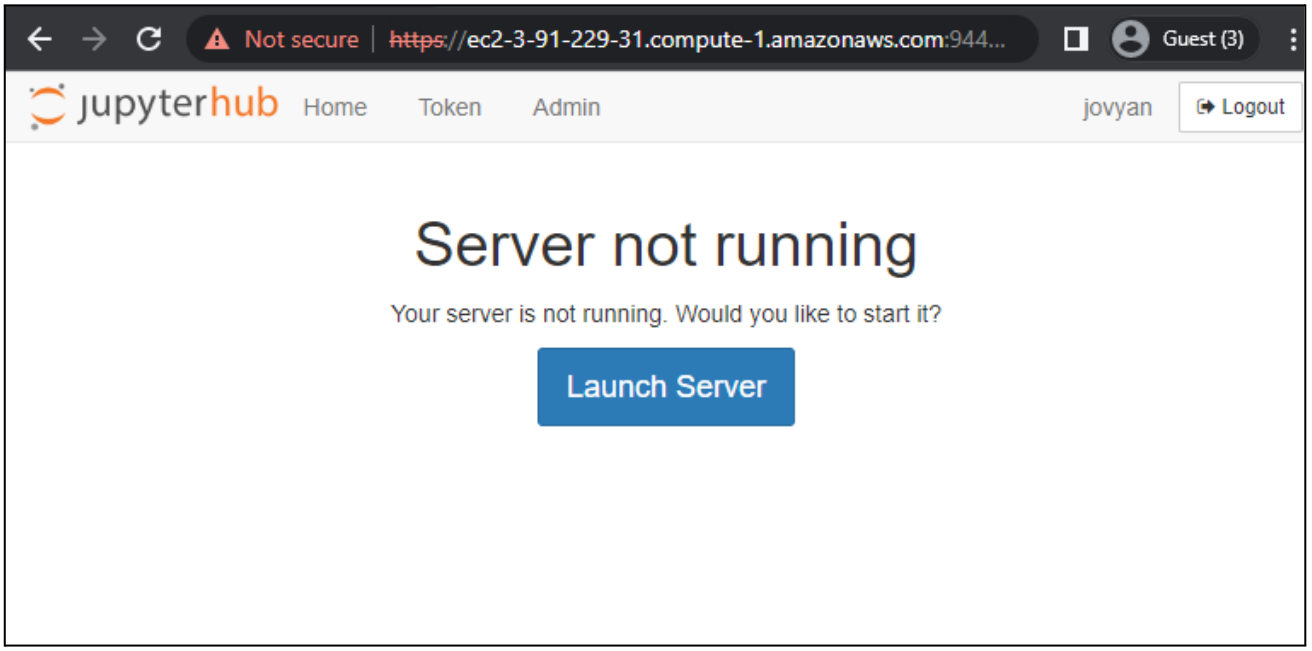
Troubleshooting:

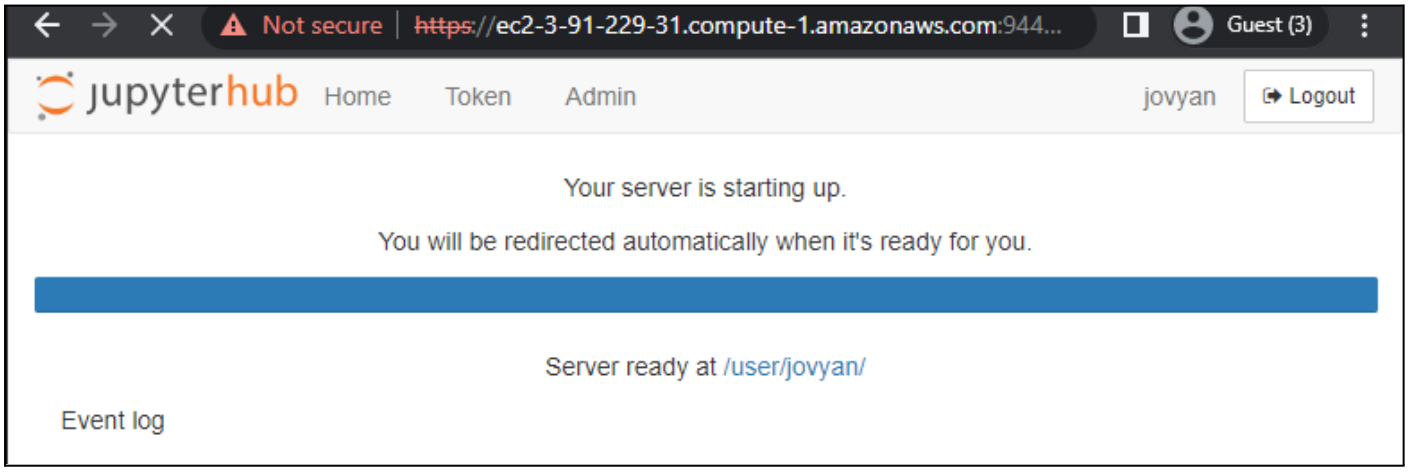
Restarting Jupyter Service

Sometimes the Jupyter service may stop working. In such situations, you may be required to restart the service. For this step, navigate to the **Control Panel** on the top right corner of the page.



Click on Launch Server to restart the service





You'll now be able to use the JupyterHub service on the EMR cluster.

Environment Variables:

The following are the environment variables for the EMR and PySpark version used in this documentation. Please use this wherever necessary.

```
import os
import sys
os.environ["PYSPARK_PYTHON"] = "/usr/bin/python3"
os.environ["JAVA_HOME"] = "/usr/java/jdk1.8.0_161/jre"
os.environ["SPARK_HOME"] = "/usr/lib/spark"
os.environ["PYLIB"] = os.environ["SPARK_HOME"] + "/python/lib"
sys.path.insert(0, os.environ["PYLIB"] +"/py4j-0.10.7-src.zip")
sys.path.insert(0, os.environ["PYLIB"] +"/pyspark.zip")
```