

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.

Amazon EMR has launched a new console experience. [Learn more](#) or [switch to the new console](#)

Create Cluster - Quick Options

[Go to advanced options](#)

General Configuration

Cluster name:

☒ Logging ?

S3 folder:

Launch mode: ☒ Cluster ? ☐ Step execution ?

Software configuration

Release:

Applications:

Hardware configuration

Instance type: The selected instance type adds 32 GiB of GP2 EBS storage per instance by default. [Learn more](#)

Number of instances: (1 master and 2 core nodes)

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

Software Configuration

Release **emr-6.7.0** ⓘ

- | | | |
|--|---|--|
| <input checked="" type="checkbox"/> Hadoop 3.2.1 | <input type="checkbox"/> Zeppelin 0.10.0 | <input checked="" type="checkbox"/> Livy 0.7.1 |
| <input checked="" type="checkbox"/> JupyterHub 1.4.1 | <input type="checkbox"/> Tez 0.9.2 | <input type="checkbox"/> Flink 1.14.2 |
| <input type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 2.4.4 | <input type="checkbox"/> Pig 0.17.0 |
| <input type="checkbox"/> Hive 3.1.3 | <input type="checkbox"/> Presto 0.272 | <input type="checkbox"/> ZooKeeper 3.5.7 |
| <input checked="" type="checkbox"/> JupyterEnterpriseGateway 2.1.0 | <input type="checkbox"/> MXNet 1.8.0 | <input type="checkbox"/> Sqoop 1.4.7 |
| <input checked="" type="checkbox"/> Hue 4.10.0 | <input type="checkbox"/> Phoenix 5.1.2 | <input type="checkbox"/> Trino 378 |
| <input type="checkbox"/> Oozie 5.2.1 | <input checked="" type="checkbox"/> Spark 3.2.1 | <input type="checkbox"/> HCatalog 3.1.3 |
| <input type="checkbox"/> TensorFlow 2.4.1 | | |

Multiple master nodes (optional)

- ☐ Use multiple master nodes to improve cluster availability. [Learn more](#) ⓘ

AWS Glue Data Catalog settings (optional)

☐ Use Glue Data Catalog

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

Cluster Nodes and Instances

Choose the instance type, number of instances, and a purchasing option. [Learn more about instance purchasing options](#) ⓘ

Console options for automatic scaling have changed. [Learn more](#) ⓘ

Node type	Instance type	Instance count	Purchasing option
Master Master - 1 ⓘ	m4.xlarge ⓘ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB ⓘ ⓘ Add configuration settings ⓘ	1 Instances	<input checked="" type="radio"/> On-demand ⓘ <input type="radio"/> Spot ⓘ Use on-demand as max price ▾
Core Core - 2 ⓘ	m5.xlarge ⓘ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB ⓘ ⓘ Add configuration settings ⓘ	<input type="text" value="0"/> Instances	<input checked="" type="radio"/> On-demand ⓘ <input type="radio"/> Spot ⓘ Use on-demand as max price ▾
Task Task - 3 ⓘ	m5.xlarge ⓘ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB ⓘ ⓘ Add configuration settings ⓘ	<input type="text" value="0"/> Instances	<input checked="" type="radio"/> On-demand ⓘ <input type="radio"/> Spot ⓘ Use on-demand as max price ▾

Proceed to set up the EMR cluster.

General Options

Cluster name

☒ Logging ⓘ
S3 folder ⓘ

☐ Log encryption ⓘ

☐ Termination protection ⓘ

Tags ⓘ

Key	Value (optional)
<input type="text" value="Add a key to create a tag"/>	<input type="text"/>

Additional Options

☐ EMRFS consistent view ⓘ

Custom AMI ID ⓘ

▶ Bootstrap Actions

Cancel

Previous

Next

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

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Security Options

EC2 key pair ⓘ

☒ Cluster visible to all IAM users in account ⓘ

Permissions ⓘ

☐ Default ☒ Custom

Select custom roles to tailor permissions for your cluster.

EMR role ⓘ

EC2 instance profile ⓘ

Auto Scaling role ⓘ

▶ Security Configuration

▶ EC2 security groups

Cancel

Previous

Create cluster

Security Groups:

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

Cluster: Spark_Cluster Starting Configuring cluster software

[Summary](#)
[Application user interfaces](#)
[Monitoring](#)
[Hardware](#)
[Configurations](#)
[Events](#)
[Steps](#)
[Bootstrap actions](#)

Summary

ID: j-8CEM6XM2BE2G
 Creation date: 2022-12-20 10:49 (UTC+5:30)
 Elapsed time: 2 minutes
 After last step completes: Cluster waits
 Termination protection: Off [Change](#)
 Tags: -- [View All / Edit](#)
 Master public DNS: ec2-3-81-218-37.compute-1.amazonaws.com [Connect to the Master Node Using SSH](#)

Configuration details

Release label: emr-6.5.0
 Hadoop distribution: Amazon 3.2.1
 Applications: Hue 4.9.0, JupyterHub 1.4.1, JupyterEnterpriseGateway 2.1.0, Spark 3.1.2, Livy 0.7.1
 Log URI: s3://aws-logs-607926466132-us-east-1/elasticmapreduce/ [View](#)
 EMRFS consistent view: Disabled
 Custom AMI ID: --

Application user interfaces

Persistent user -- interfaces [View](#)
 On-cluster user Not Enabled [Enable an SSH Connection](#)
 interfaces [View](#)

Network and hardware

Availability zone: us-east-1d
 Subnet ID: [subnet-0169cf538af361817](#) [View](#)
 Master: Bootstrapping 1 m4.xlarge
 Core: --
 Task: --
 Cluster scaling: Not enabled
 Auto-termination: Not enabled

Security and access

Key name: XXXXXXXXXX
 EC2 instance profile: EMR_EC2_DefaultRole
 EMR role: EMR_DefaultRole
 Auto Scaling role: EMR_AutoScaling_DefaultRole
 Visible to all users: All [Change](#)
 Security groups for Master: sg-0bf8fe25e1ab068ce [View](#) (ElasticMapReduce-master)
 Security groups for Core & Task: sg-0e160b3b32c0f303f [View](#) (ElasticMapReduce-slave)

Click on the Security Group

Security Groups (2) Info

Filter security groups

search: sg-0bf8fe25e1ab068ce

Clear filters

Actions

Export security groups to CSV

Create security group

< 1 >

<input type="checkbox"/>	Name	Security group ID	Security group name	VPC ID	Description	Owner	Inbound rules count
<input type="checkbox"/>	-	sg-0ef60b3b32c0f303f	ElasticMapReduce-slave	vpc-04faa3ec703acc6e2	Slave group for Elastic ...	607926466132	6 Permission entries
<input type="checkbox"/>	-	sg-0bf8fe25e1ab068ce	ElasticMapReduce-mas...	vpc-04faa3ec703acc6e2	Master group for Elasti...	607926466132	21 Permission entries

Click on Edit Inbound Rules

sg-0bf8fe25e1ab068ce - ElasticMapReduce-master

Details

Security group name ElasticMapReduce-master	Security group ID sg-0bf8fe25e1ab068ce	Description Master group for Elastic MapReduce created on 2022-12-01T05:39:58.035Z	VPC ID vpc-04faa3ec703acc6e2
Owner 607926466132	Inbound rules count 21 Permission entries	Outbound rules count 1 Permission entry	

Inbound rules (21)

You can now check network connectivity with Reachability Analyzer [Run Reachability Analyzer](#)

[Filter security group rules](#)

[Manage tags](#) [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.

Security group ID	Type	Protocol	Port	Source	Action
sg-0f6217e0a64cf0081	Custom TCP	TCP	8443	54.240.217.64/28	Delete
sg-0977ad19c9253e731	Custom TCP	TCP	8443	207.171.172.6/32	Delete
sg-099678441d80bf25b	Custom TCP	TCP	8443	54.239.98.0/24	Delete
-	Custom TCP	TCP	22	207.171.167.101/32	Delete
-	Custom TCP	TCP	9443	0.0.0.0/0	Delete
-	Custom TCP	TCP	8888	0.0.0.0/0	Delete

[Add rule](#)

[Cancel](#) [Preview changes](#) [Save rules](#)

EC2 > Security Groups > sg-0bf8fe25e1ab068ce - ElasticMapReduce-master > Edit inbound rules: Processing

Edit inbound rules: Processing

Modifying your security group

↶ New

0%

▶ Details

Accessing Jupyter Service:

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

Cluster: Spark_Cluster **Waiting** Cluster ready to run steps.

Summary **Application user interfaces** Monitoring Hardware Configurations Events Steps Bootstrap actions

Summary

ID: j-1163PMLY9JKEU
 Creation date: 2022-12-16 11:19 (UTC+5:30)
 Elapsed time: 21 minutes
 After last step completes: Cluster waits
 Termination protection: Off [Change](#)
 Tags: -- [View All / Edit](#)
 Master public DNS: ec2-3-91-229-31.compute-1.amazonaws.com
[Connect to the Master Node Using SSH](#)

Configuration details

Release label: emr-5.36.0
 Hadoop distribution: Amazon 2.10.1
 Applications: Hive 2.3.9, Hue 4.10.0, JupyterHub 1.4.1, JupyterEnterpriseGateway 2.1.0, Spark 2.4.8, Livy 0.7.1
 Log URI: s3://aws-logs-607926466132-us-east-1/elasticmapreduce/
 EMRFS consistent view: Disabled
 Custom AMI ID: --
 Amazon Linux Release: 2.0.20221103.3 [Learn more](#)

Application user interfaces

Persistent user [Spark history server, YARN timeline server, Tez UI interfaces](#)
 On-cluster user [HDFS Name Node, Hue, Spark History Server, JupyterHub, Livy, Resource Manager](#)

Network and hardware

Availability zone: us-east-1d
 Subnet ID: [subnet-0169cf538af361817](#)
 Master: **Running** 1 m4.xlarge
 Core: --
 Task: --
 Cluster scaling: Not enabled
 Auto-termination: Not enabled

Security and access

Key name: rkey

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.

Cluster: Spark_Cluster **Waiting** Cluster ready to run steps.

Summary **Application user interfaces** Monitoring Hardware Configurations Events Steps Bootstrap actions

Persistent application user interfaces

Applications installed on the Amazon EMR cluster publish user interfaces (UI) as web sites to monitor cluster activity. Persistent UI logs are available for 30 days after an application ends. Persistent UI don't required SS hosted off of the cluster.

Application user interface

[YARN timeline server](#)
[Tez UI](#)
[Spark history server](#)

On-cluster application user interfaces

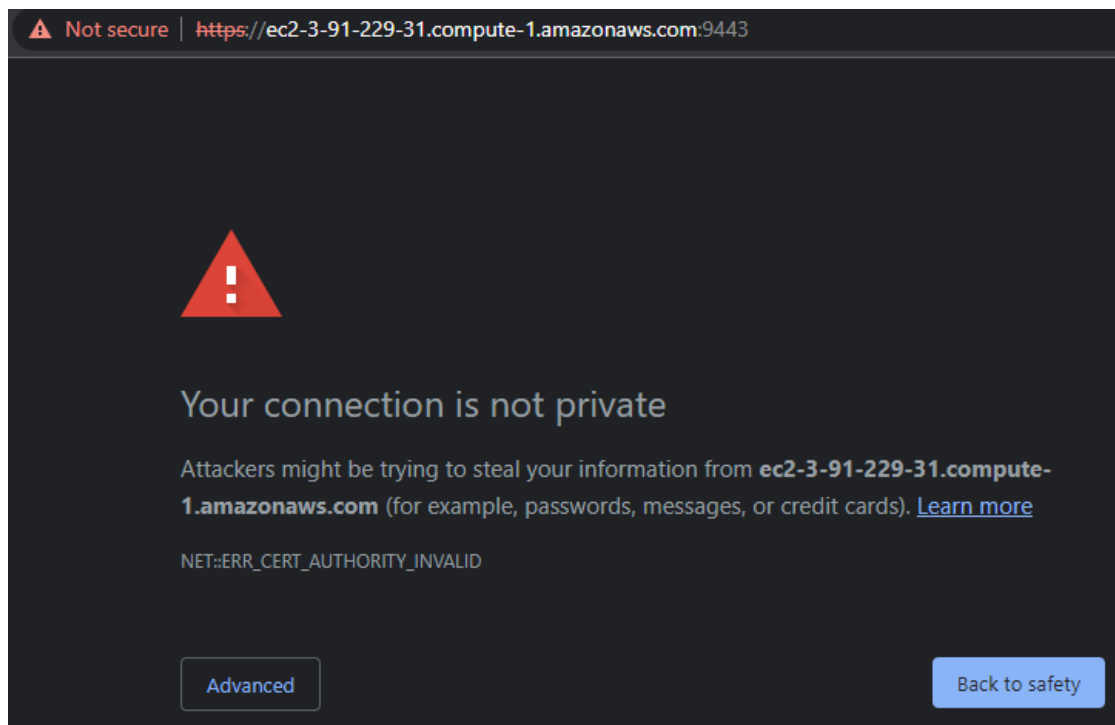
On-cluster UI are available only while clusters are running. Because they are hosted on the master node, on-cluster UI require a connection via SSH tunneling. Set up SSH tunneling before accessing these application

Application	User interface URL	Status
HDFS Name Node	http://ec2-3-91-229-31.compute-1.amazonaws.com:50070/	Available
Hue	http://ec2-3-91-229-31.compute-1.amazonaws.com:8888/	Available
JupyterHub	https://ec2-3-91-229-31.compute-1.amazonaws.com:9443/	Available
Spark History Server	http://ec2-3-91-229-31.compute-1.amazonaws.com:18080/	Available
Livy	http://ec2-3-91-229-31.compute-1.amazonaws.com:8998/	Available
Resource Manager	http://ec2-3-91-229-31.compute-1.amazonaws.com:8088/	Available

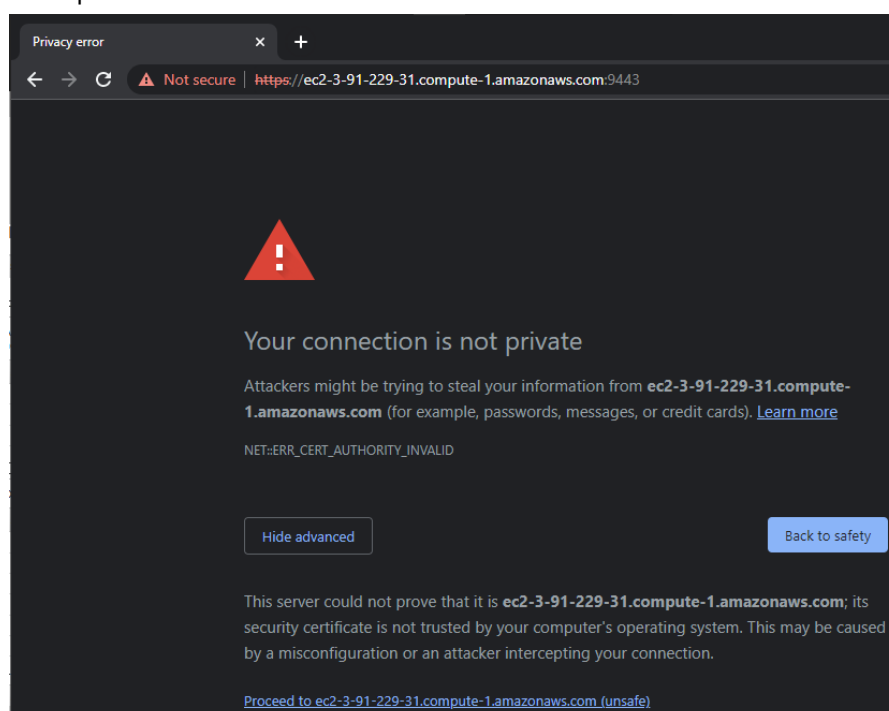
The following table lists web interfaces you can view on the task nodes:

Application	User interface URL
-------------	--------------------

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



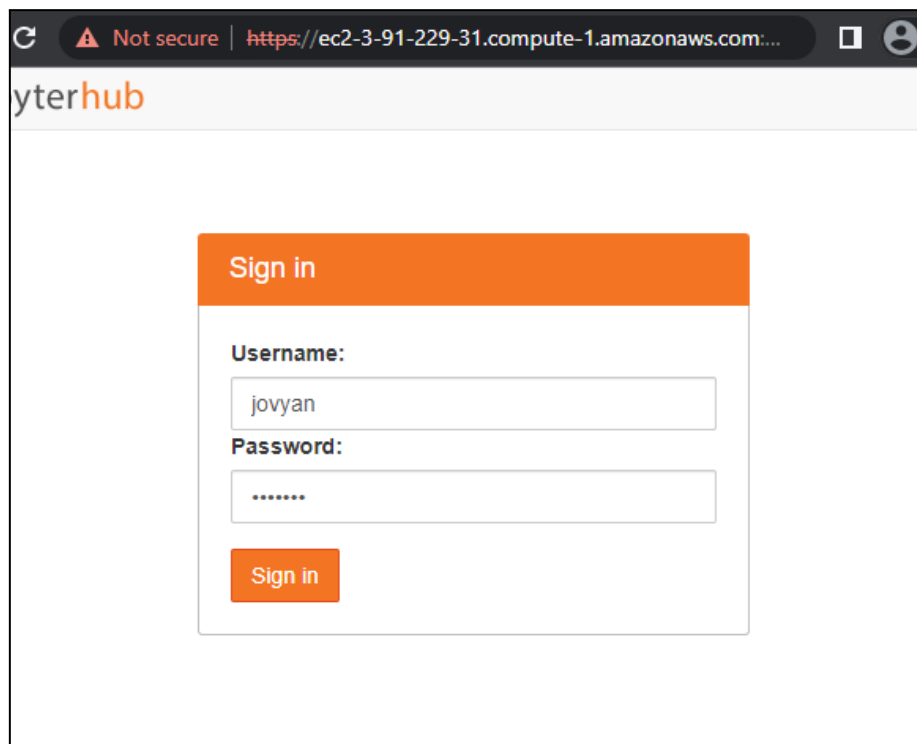
Click on '**Advanced**' and proceed to the link.



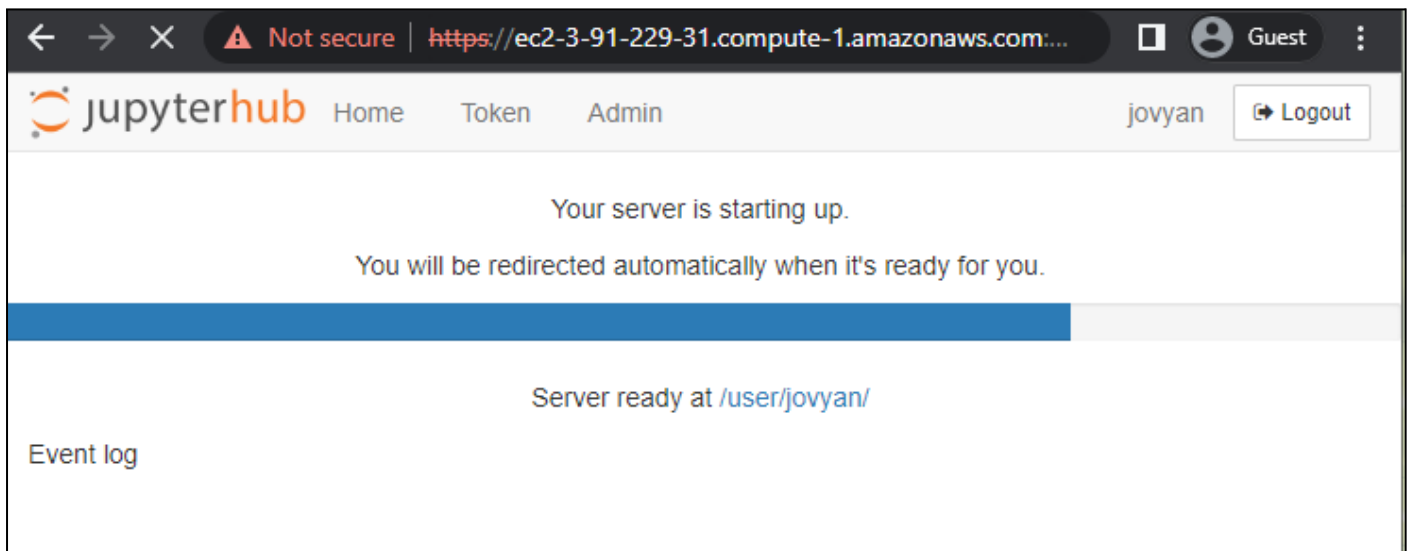
Enter the following credentials once the login page appears and click on the **Sign In** button.

Username: jovyan

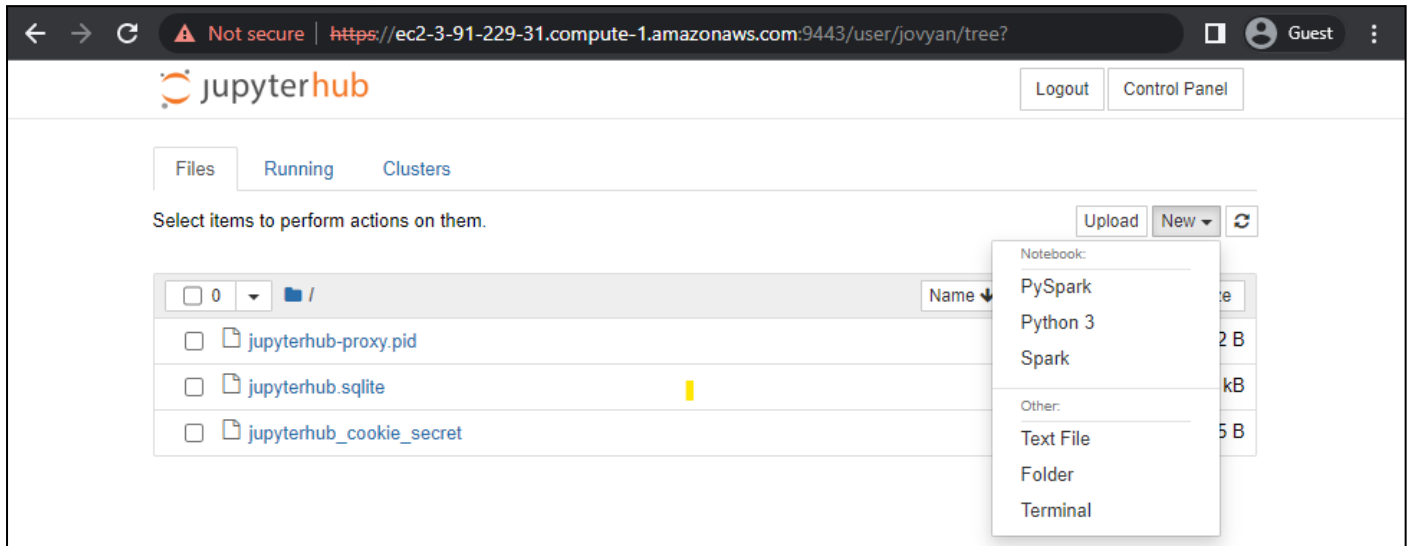
Password: jupyter



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



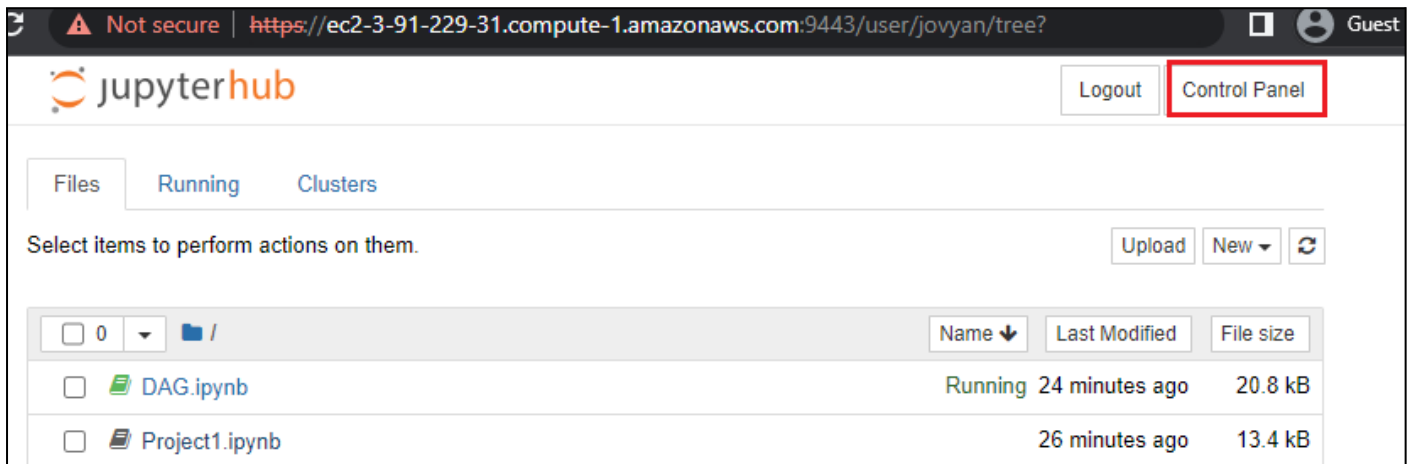
Click on the **Upload** button to upload the notebooks or the **New** button to create a new PySpark notebook.



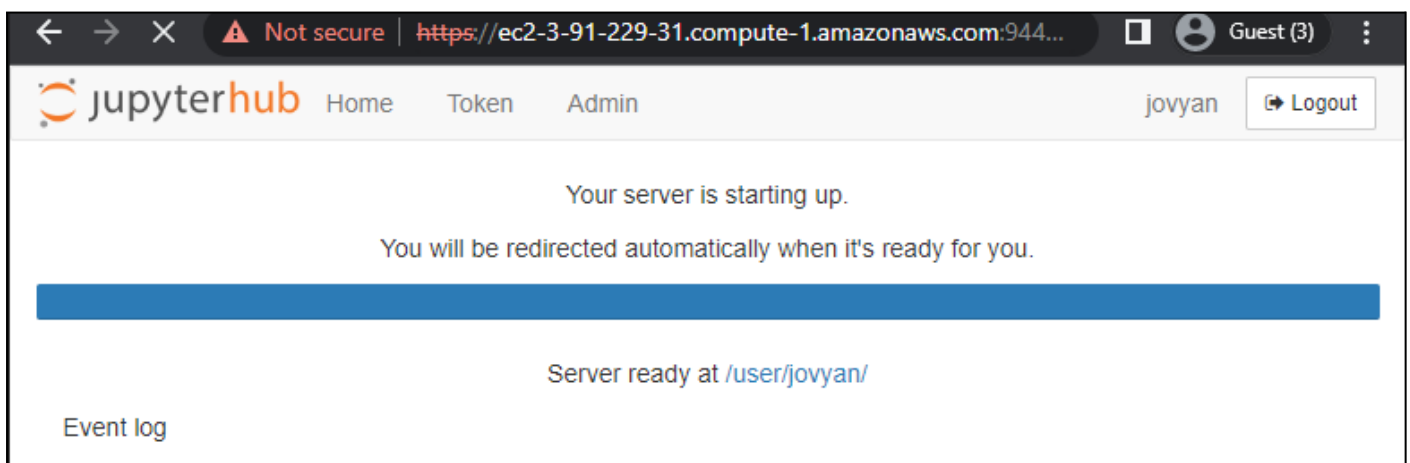
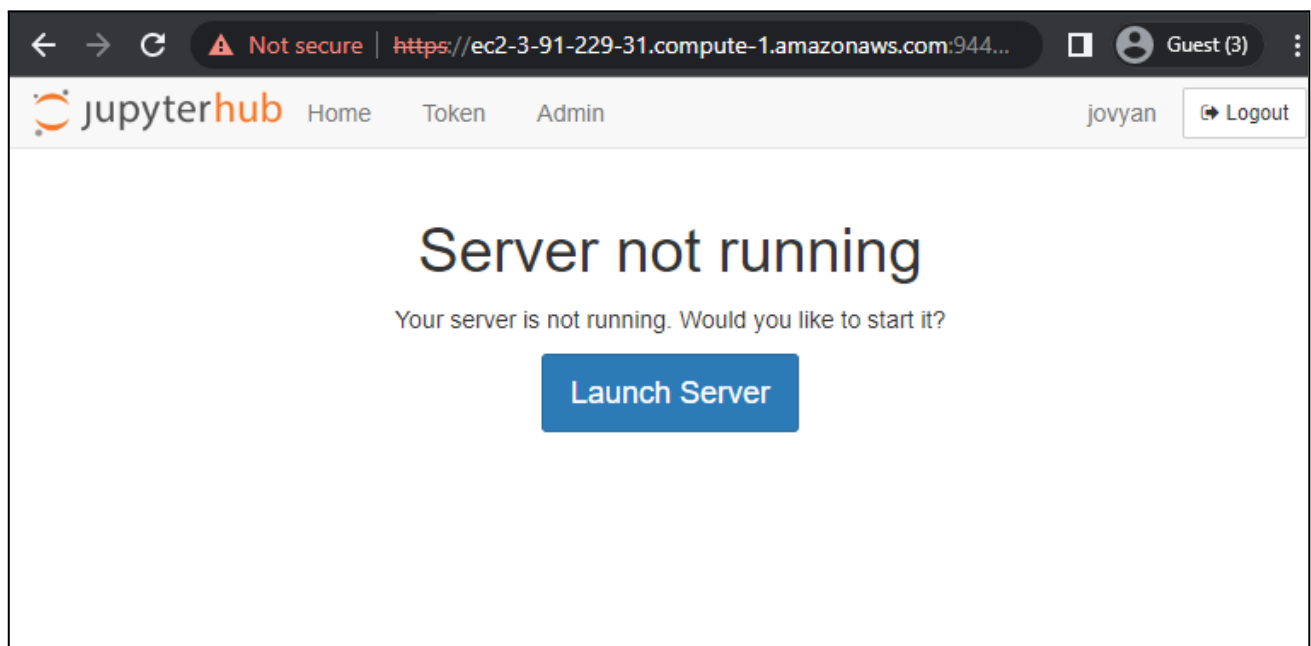
Troubleshooting:

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")
```

Disclaimer: All content and material on the upGrad website is copyrighted material, either belonging to upGrad or its bonafide contributors and is purely for the dissemination of education. You are permitted to access print and download extracts from this site purely for your own education only and on the following basis:

- You can download this document from the website for self-use only.
- Any copies of this document, in part or full, saved to disc or to any other storage medium may only be used for subsequent, self-viewing purposes or to print an individual extract or copy for non-commercial personal use only.
- Any further dissemination, distribution, reproduction, copying of the content of the document herein or the uploading thereof on other websites or use of the content for any other commercial/unauthorized purposes in any way which could infringe the intellectual property rights of upGrad or its contributors, is strictly prohibited.
- No graphics, images or photographs from any accompanying text in this document will be used separately for unauthorised purposes.
- No material in this document will be modified, adapted or altered in any way.
- No part of this document or upGrad content may be reproduced or stored in any other web site or included in any public or private electronic retrieval system or service without upGrad's prior written permission.
- Any rights not expressly granted in these terms are reserved.