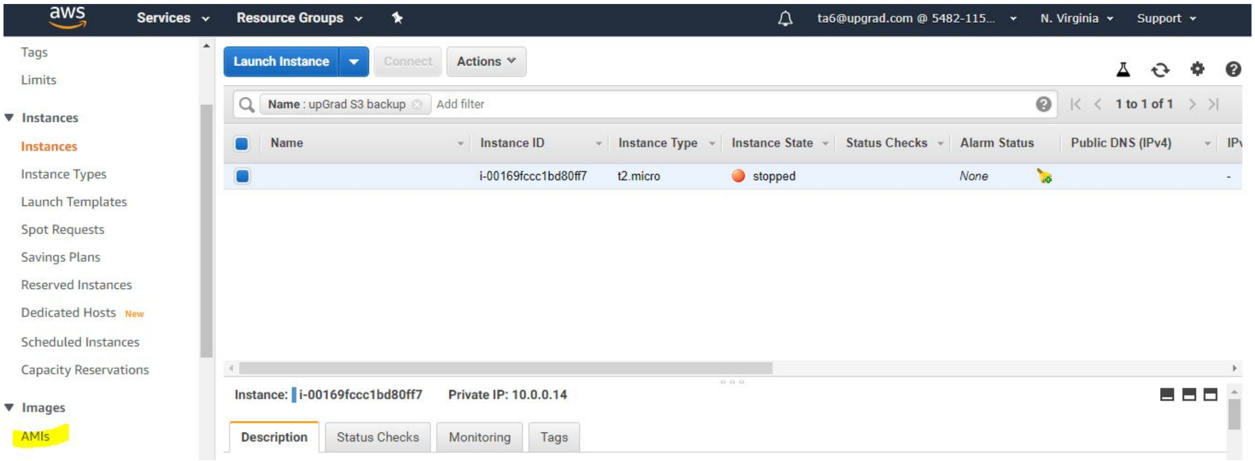
Instance Configuration

Go to the EC2 page on the AWS Management console to launch the instance. As part of this assignment, you are expected to work with the following AMI:

AMI Name - spark-jupyter
AMI ID - ami-0437a7d30763297af

To do so, follow the steps given below:

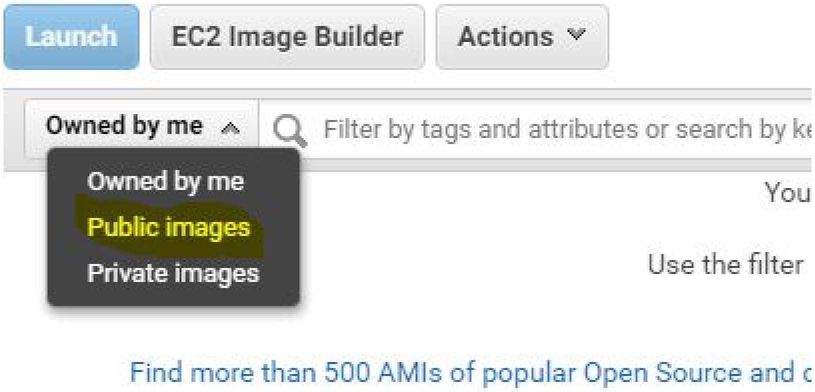
1. Go to the EC2 dashboard and open the **AMI** section.

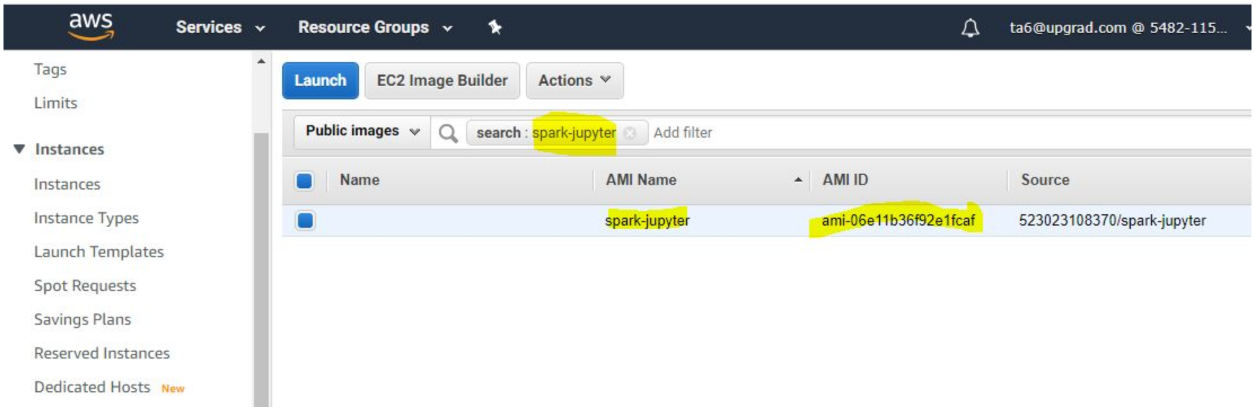


Name	Instance ID	Instance Type	Instance State	Status Checks	Alarm Status	Public DNS (IPv4)	IPv
	i-00169fccc1bd80ff7	t2.micro	stopped		None		-

Instance: i-00169fccc1bd80ff7 Private IP: 10.0.0.14

2. Under the AMI section, select the option **Public images** and search for the ami name mentioned above: **spark-jupyter**

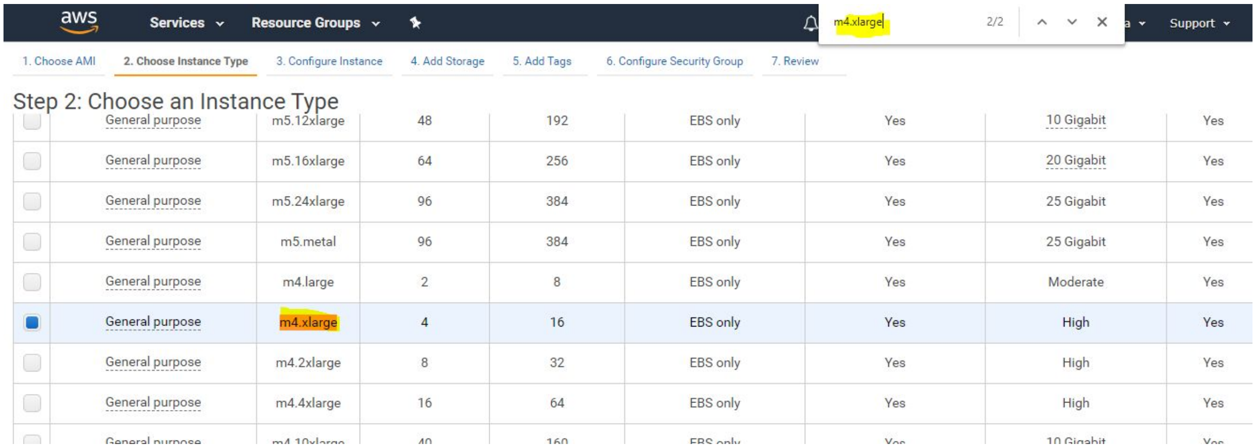

Launch
EC2 Image Builder
Actions
Owned by me
Filter by tags and attributes or search by ke
Owned by me
Public images
Private images
You
Use the filter
Find more than 500 AMIs of popular Open Source and c



Name	AMI Name	AMI ID	Source
	spark-jupyter	ami-06e11b36f92e1fcaf	523023108370/spark-jupyter

3. Select the AMI image and press the **Launch** button. You will move to the next step towards launching the EC2 instance.

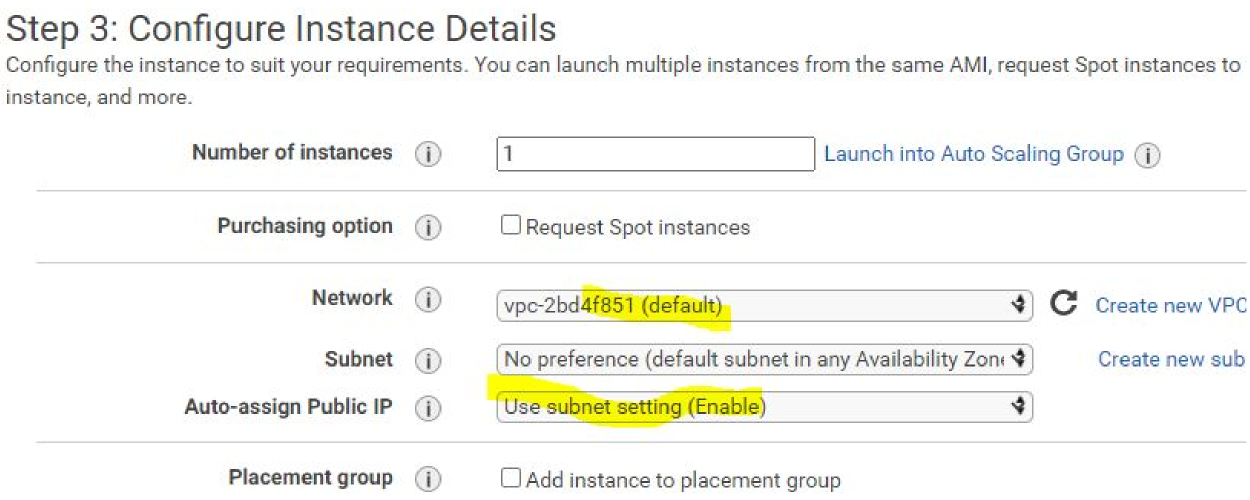
4. You are expected to work with the following instance type: **m4.xlarge**. Select the instance-type from the provided list and click on Next.



Step 2: Choose an Instance Type

□	General purpose	m5.12xlarge	48	192	EBS only	Yes	10 Gigabit	Yes
□	General purpose	m5.16xlarge	64	256	EBS only	Yes	20 Gigabit	Yes
□	General purpose	m5.24xlarge	96	384	EBS only	Yes	25 Gigabit	Yes
□	General purpose	m5.metal	96	384	EBS only	Yes	25 Gigabit	Yes
□	General purpose	m4.large	2	8	EBS only	Yes	Moderate	Yes
■	General purpose	m4.xlarge	4	16	EBS only	Yes	High	Yes
□	General purpose	m4.2xlarge	8	32	EBS only	Yes	High	Yes
□	General purpose	m4.4xlarge	16	64	EBS only	Yes	High	Yes
□	General purpose	m4.10xlarge	40	160	EBS only	Yes	10 Gigabit	Yes

5. Keep the default following VPC settings under the next section as shown in the image below:



Step 3: Configure Instance Details

Configure the instance to suit your requirements. You can launch multiple instances from the same AMI, request Spot instances to instance, and more.

Number of instances: 1 — Launch into Auto Scaling Group

Purchasing option: ☐ Request Spot instances

Network: vpc-2bd4f851 (default) — Create new VPC

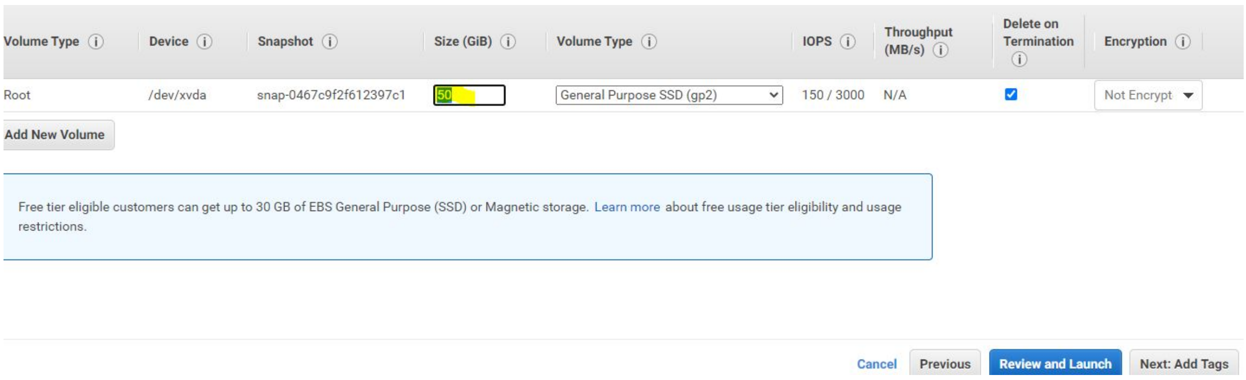
Subnet: No preference (default subnet in any Availability Zone — Create new sub

Auto-assign Public IP: Use subnet setting (Enable)

Placement group: ☐ Add instance to placement group

Click on Next once done.

6. Next, you are expected to increase the volume size of the instance to **50 GBs**. Keep the Volume Type as a **General Purpose SSD**.



Volume Type	Device	Snapshot	Size (GiB)	Volume Type	IOPS	Throughput (MB/s)	Delete on Termination	Encryption
Root	/dev/xvda	snap-0467c9f2f612397c1	50	General Purpose SSD (gp2)	150 / 3000	N/A	☑	Not Encrypt

Add New Volume

Free tier eligible customers can get up to 30 GB of EBS General Purpose (SSD) or Magnetic storage. Learn more about free usage tier eligibility and usage restrictions.

7. Under the next section, you can add the tags to your instance. This is not a compulsory step and can be ignored.

8. Under the Configure Security Group section, you must create a new security group that allows the access of the following ports to your IP address: **22** (SSH), **8888** (Custom TCP - Jupyter) and **4040** (Custom TCP - Spark UI).
Note: Remember that this IP must be checked every time you launch the instance.

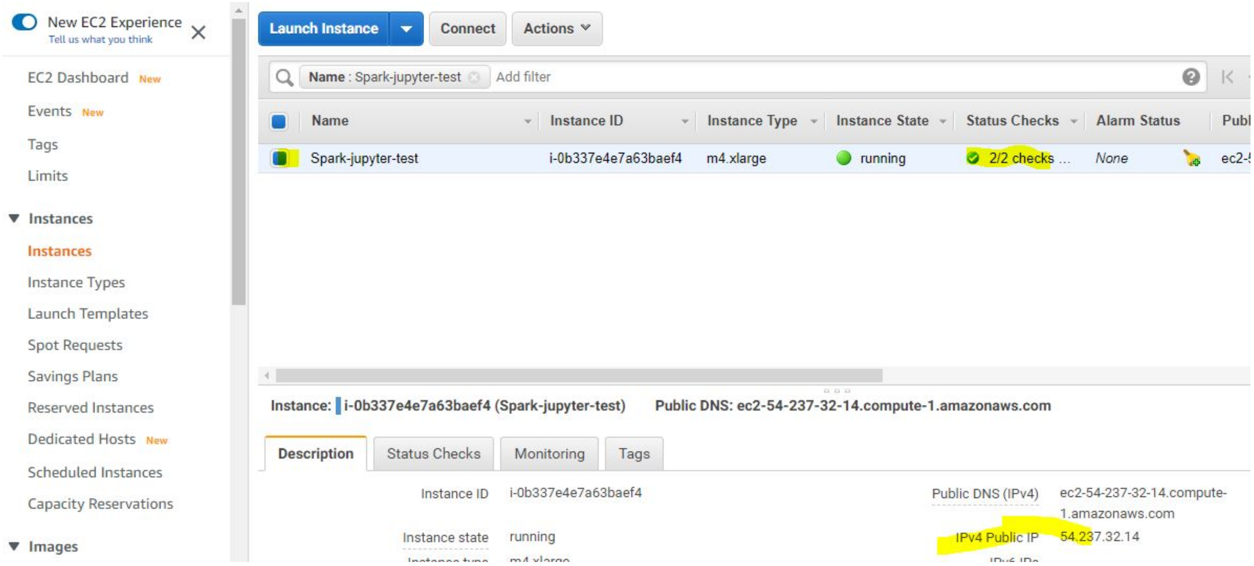
   After the group is created, you can click on **Review and Launch**.

9. Next, you will be asked for the key-pair for the instance. You can select an existing key-pair or create a new one if required. Always keep this key-pair safe as it will provide a security layer to the instance.

Launching the Jupyter Notebook

1. You can use the PuTTYgen software to convert the PEM file into a PPK file if you are using a Windows machine. Ignore this if you are working on a Linux machine or MAC.

2. To connect to the instance:
   a. Windows: Use the PuTTY software to connect to the instance.



Name : Spark-jupyter-test

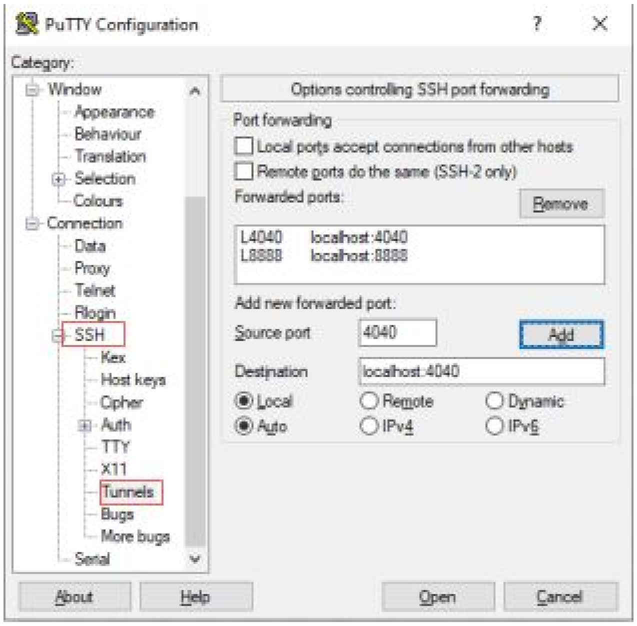
Name	Instance ID	Instance Type	Instance State	Status Checks	Alarm Status	Publ
Spark-jupyter-test	i-0b337e4e7a63baef4	m4.xlarge	running	2/2 checks ...	None	ec2-!

Instance: i-0b337e4e7a63baef4 (Spark-jupyter-test)   Public DNS: ec2-54-237-32-14.compute-1.amazonaws.com

Instance ID i-0b337e4e7a63baef4   Public DNS (IPv4) ec2-54-237-32-14.compute-1.amazonaws.com

Instance state running   IPv4 Public IP 54.237.32.14

You must provide the correct path for the PPK file under the **Auth** section of the **SSH** tab. Moreover, you must also add the required ports under the **Tunnel** section as shown below.



Options controlling SSH port forwarding

Port forwarding

☐ Local ports accept connections from other hosts

☐ Remote ports do the same (SSH-2 only)

Forwarded ports:

```
L4040    localhost:4040
L8888    localhost:8888
```

Add new forwarded port:

Source port 4040

Destination localhost:4040

● Local ○ Remote ○ Dynamic

● Auto ○ IPv4 ○ IPv6

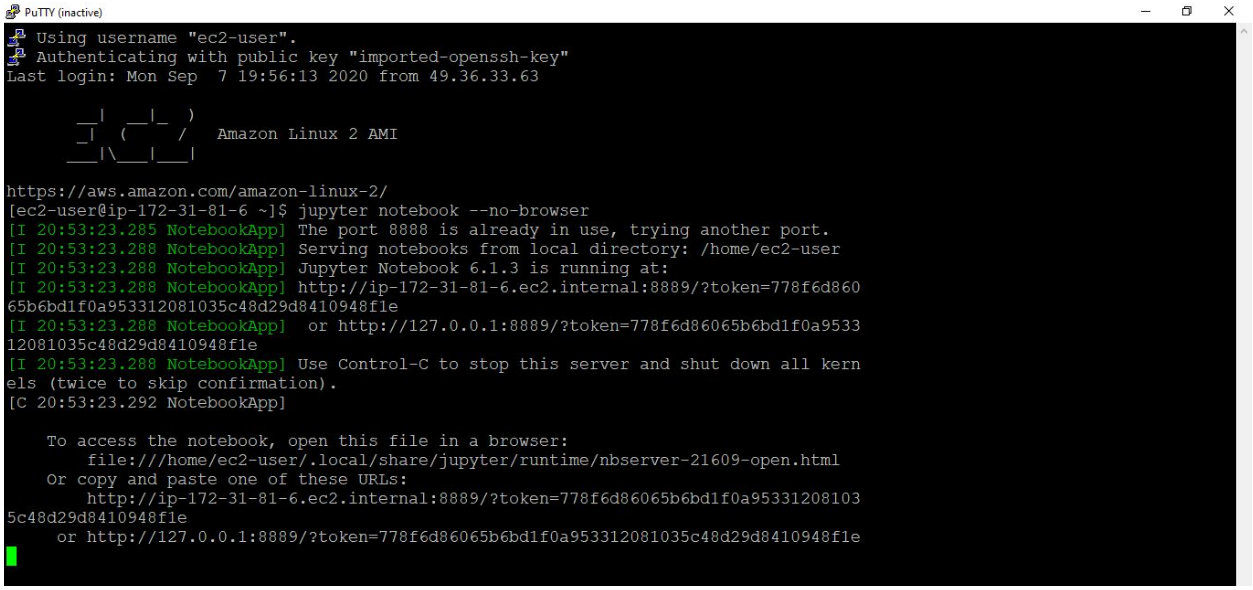
    b. Linux/MAC: To connect using your instance's public DNS name, enter the following command:

       ssh -i /path/my-key-pair.pem ec2-user@instance-public-dns-name

3. The instance is pre-loaded with all the required tools:
    a. Python
    b. Jupyter
    c. Spark

   Hence, you are expected to launch the instance directly using the following command:

jupyter notebook --no-browser



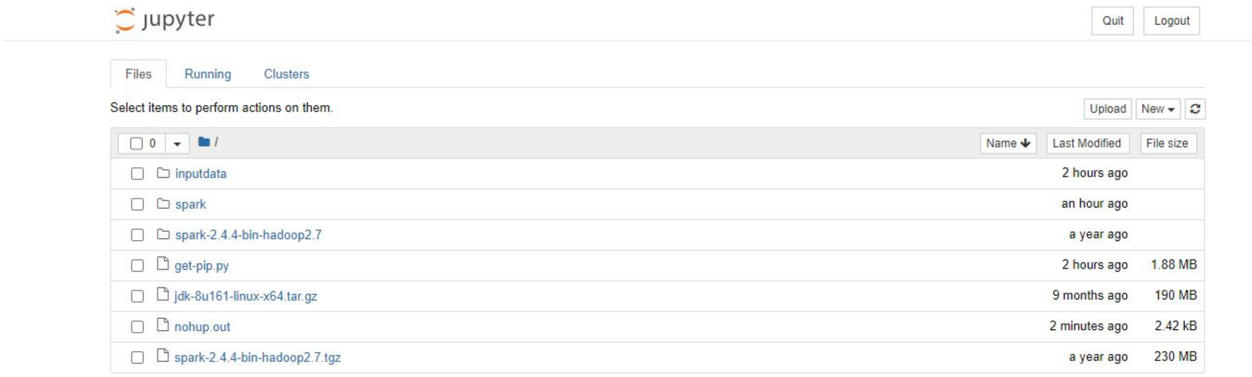
```
  Using username "ec2-user".
  Authenticating with public key "imported-openssh-key"
Last login: Mon Sep  7 19:56:13 2020 from 49.36.33.63

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
[ec2-user@ip-172-31-81-6 ~]$ jupyter notebook --no-browser
[I 20:53:23.285 NotebookApp] The port 8888 is already in use, trying another port.
[I 20:53:23.288 NotebookApp] Serving notebooks from local directory: /home/ec2-user
[I 20:53:23.288 NotebookApp] Jupyter Notebook 6.1.3 is running at:
[I 20:53:23.288 NotebookApp] http://ip-172-31-81-6.ec2.internal:8889/?token=778f6d860
65b6bd1f0a953312081035c48d29d8410948f1e
[I 20:53:23.288 NotebookApp]  or http://127.0.0.1:8889/?token=778f6d86065b6bd1f0a9533
12081035c48d29d8410948f1e
[I 20:53:23.288 NotebookApp] Use Control-C to stop this server and shut down all kern
els (twice to skip confirmation).
[C 20:53:23.292 NotebookApp]

    To access the notebook, open this file in a browser:
        file:///home/ec2-user/.local/share/jupyter/runtime/nbserver-21609-open.html
    Or copy and paste one of these URLs:
        http://ip-172-31-81-6.ec2.internal:8889/?token=778f6d86065b6bd1f0a95331208103
5c48d29d8410948f1e
     or http://127.0.0.1:8889/?token=778f6d86065b6bd1f0a953312081035c48d29d8410948f1e
```

   You can copy the provided IP address for launching the Jupyter notebook in your local browser. (Here: http://127.0.0.1:8888/?token=778f6d86065b6bd1f0a953312081035c48d29d8410948f1e)

   Remember that if you are hosting multiple sessions (8889 in the above image), you must provide access to all the ports under security groups and the tunnel section in PuTTY. You can check the sessions using the following command:

jupyter notebook list

4. Once the Jupyter notebook is live, you can use it to write your Python queries directly.



Select items to perform actions on them.

Name	Last Modified	File size
inputdata	2 hours ago	
spark	an hour ago	
spark-2.4.4-bin-hadoop2.7	a year ago	
get-pip.py	2 hours ago	1.88 MB
jdk-8u161-linux-x64.tar.gz	9 months ago	190 MB
nohup.out	2 minutes ago	2.42 kB
spark-2.4.4-bin-hadoop2.7.tgz	a year ago	230 MB

However, to use Spark, you must run the following commands in the notebook to set the environment variables:

```
import os
import sys
os.environ["PYSPARK_PYTHON"]="/usr/bin/python3"
os.environ["PYSPARK_DRIVER_PYTHON"]="/usr/bin/python3"
os.environ["PYSPARK_DRIVER_PYTHON_OPTS"]="notebook --no-browser"
os.environ["JAVA_HOME"] = "/usr/java/jdk1.8.0_161/jre"
os.environ["SPARK_HOME"] = "/home/ec2-user/spark-2.4.4-bin-hadoop2.7"
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")
```

Now, your instance is ready to perform all the expected tasks.