7 Steps for setting up Apache Spark on a free (or not free) tiered Linux EC2 instance and running PySpark on a Jupyter Notebook Server that you can access on your remote machine

Please note--- all of the resources mentioned on this document have been deleted, including the s3 bucket. Also, the keys used to log-in to my now deleted EC2 instance have been deleted. Nothing is secure; never underestimate the abilities of hackers... Finally, the cost for this project ended up being a little over a dollar. If you don't spin down your resources, etc... AWS will eventually charge you for its service.

Step 1. Spin up a (free tier) Linux EC2 instance on AWS ; ssh into and update

```
cyberpunk@Robs-Air project % pwd
/Users/cyberpunk/Projects/UMD_COURSES/msml651/project
cyberpunk@Robs-Air project %  ssh -i "umdprojectinstances.pem"
135.compute-1.amazonaws.com
```

```
ubuntu@ip-172-31-22-78:~$ sudo apt-get update
```

Step 2. Go to Apache Spark's Website and install Spark on you EC2 instance


```
ubuntu@ip-172-31-22-78:~$ wget https://dlcdn.apache.org/spark/spark-3.2.0/spa
-bin-hadoop3.2.tgz
--2021-11-15 00:50:16--  https://dlcdn.apache.org/spark/spark-3.2.0/spark-3.2
adoop3.2.tgz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::64
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... conne
HTTP request sent, awaiting response... 200 OK
Length: 300965906 (287M) [application/x-gzip]
Saving to: 'spark-3.2.0-bin-hadoop3.2.tgz'

spark-3.2.0-bin-hadoo 100%[=====================>] 287.02M  44.9MB/s    in 6

2021-11-15 00:50:22 (46.0 MB/s) - 'spark-3.2.0-bin-hadoop3.2.tgz' saved [3009
0965906]
```

STEP 3. Untar ; save the folder in /opt

```
ubuntu@ip-172-31-22-78:~$ tar -zxvf spark-3.2.0-bin-hadoop3.2.tgz
```

STEP 4. Make the following installations

A.) PIP3

```
ubuntu@ip-172-31-22-78:~$ sudo apt-get -y install python3-pip
```

B.) Py4J

```
ubuntu@ip-172-31-22-78:~$ pip3 install py4j
```

C.) Jupyter Notebook

```
ubuntu@ip-172-31-22-78:~$ pip3 install jupyter
```

D.) JAVA 11

```
ubuntu@ip-172-31-22-78:~$ java --version

Command 'java' not found, but can be installed with:

sudo apt install openjdk-11-jre-headless  # version 11.0.11+9-0ubuntu2~20.04,
sudo apt install default-jre              # version 2:1.11-72
sudo apt install openjdk-13-jre-headless  # version 13.0.7+5-0ubuntu1~20.04
sudo apt install openjdk-16-jre-headless  # version 16.0.1+9-1~20.04
sudo apt install openjdk-17-jre-headless  # version 17+35-1~20.04
sudo apt install openjdk-8-jre-headless   # version 8u292-b10-0ubuntu1~20.04

ubuntu@ip-172-31-22-78:~$ sudo apt install openjdk-11-jre-headless
```

E.) SCALA

```
ubuntu@ip-172-31-22-78:~$ sudo apt-get install scala
```

F.) AWS Command Line

```
ubuntu@ip-172-31-22-78:~$ aws

Command 'aws' not found, but can be installed with:

sudo snap install aws-cli  # version 1.15.58, or
sudo apt  install awscli   # version 1.18.69-1ubuntu0.20.04.1

See 'snap info aws-cli' for additional versions.

ubuntu@ip-172-31-22-78:~$ sudo apt  install awscli
```

STEP 5. Spark Configurations in ~/.profile; includes using Jupyter notebook when launching PySpark

```
export SPARK_HOME=/opt/spark
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd
export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbi
export PYTHONPATH=$SPARK_HOME/python:$PYTHONPATH
export PYSPARK_DRIVER_PYTHON="jupyter"
export PYSPARK_DRIVER_PYTHON_OPTS="notebook"
export PYSPARK_PYTHON=python3
```

A.) Also make sure following edits made to /opt/spark/bin/pyspark

```
# Add the PySpark classes to the Python path:
export PYTHONPATH="${SPARK_HOME}/python/:$PYTHON
export PYTHONPATH="${SPARK_HOME}/python/lib/py4j

# Load the PySpark shell.py script when ./pyspar
export OLD_PYTHONSTARTUP="$PYTHONSTARTUP"
export PYTHONSTARTUP="${SPARK_HOME}/python/pyspa
```

Note: line below is last line of /opt/spark/bin/pyspark; (I used "more pyspark" on the command line to show its contents)

```
exec "${SPARK_HOME}"/bin/spark-submit pyspark-she
ubuntu@ip-172-31-22-78:/opt/spark/bin$
```

STEP 6. Generate Jupyter config file (can be found in .jupyter folder of home directory); and make configurations as needed. Below are a few I made

```
ubuntu@ip-172-31-22-78:~$ jupyter notebook --generate-config
```

```
c.NotebookApp.allow_password_change = False
```

```
c.NotebookApp.allow_remote_access = True
```

```
#  Default: 1000000
c.NotebookApp.iopub_data_rate_limit = 10000000
```

```
ubuntu@ip-172-31-22-78:~$ ipython
Python 3.8.10 (default, Sep 28 2021, 16:10:42)
Type 'copyright', 'credits' or 'license' for more information
IPython 7.29.0 -- An enhanced Interactive Python. Type '?' for help.

In [1]: from notebook.auth import passwd

In [2]: passwd()
Enter password:
Verify password:
```

```
c.NotebookApp.password = '
```

```
#  Default: 8888
c.NotebookApp.port = 8888
```

```
.NotebookApp.ip = 'ec2-54-174-30-135.compute-1.amazonaws.com'
```

STEP 7. Set up security group; for my project only allowed my Local, remote IP address for inbound

Filter rules

Security group rule ID	Port range	Protocol	Source	Security groups
sgr-087ea929826a513f9	22	TCP	[redacted]	launch-wizard-3
sgr-0c6b308dda56de78a	8888	TCP		launch-wizard-3

▼ Outbound rules

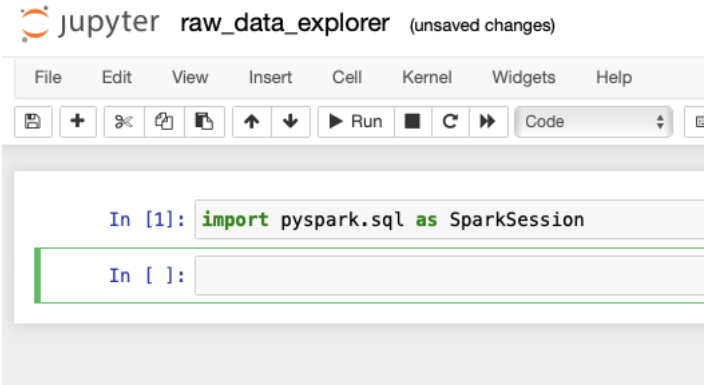
Filter rules

Security group rule ID	Port range	Protocol	Destination	Security groups
sgr-021e5249f06a2740d	All	All	0.0.0.0/0	launch-wizard-3

STEP 7: START SPARK AND LOGIN FROM LOCAL MACHINE; please note, that setting up an s3 bucket and giving the EC2 instance access rights to that s3 bucket are not covered in this write-up. Here is the link that I used to do this though, if interested: https://aws.amazon.com/premiumsupport/knowledge-center/ec2-instance-access-s3-bucket/

```
ubuntu@ip-172-31-22-78:~$ pyspark
[I 01:26:44.772 NotebookApp] Writing notebook server cookie secret to /home/u
are/jupyter/runtime/notebook_cookie_secret
[I 01:26:45.311 NotebookApp] Serving notebooks from local directory: /home/ub
[I 01:26:45.312 NotebookApp] Jupyter Notebook 6.4.5 is running at:
[I 01:26:45.312 NotebookApp] http://ec2-54-174-30-135.compute-1.amazonaws.com
[I 01:26:45.312 NotebookApp] Use Control-C to stop this server and shut down
ice to skip confirmation).
[W 01:26:45.318 NotebookApp] No web browser found: could not locate runnable
```


jupyter
Password:
Log in


jupyter
raw_data_explorer
(unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code
In [1]: import pyspark.sql as SparkSession
In [ ]:

In [ ]: