I have checked with various versions of python and Spark compatibility and the one which worked for me is 2020.02 (Feb 2020) version (Else I was getting Numpy errors and few other errors while running code in Jupyter Notebook)

First, you needed to clean everything from your virtual machine, by running this command:

$ rm -rfv / home/ubuntu /{\*,.\*}

Getting Anaconda:

$ wget http://repo.continuum.io/archive/Anaconda3-2020.02-Linux-x86\_64.sh

$ bash Anaconda3-2020.02-Linux-x86\_64.sh

then run:

$ source ~/.bashrc

This will change your environment to Conda's (Base)

(base) **ubuntu@ip-172-31-16-113**

now we will install everything in this environment only:

let's check the version of Python:

$ which python

/home/ubuntu/anaconda3/bin/python

$ python --version

it will give you -- Python 3.7.6

Now run below codes for configuration of Jupyter Notebook and connection with EC2

$ jupyter notebook --generate-config

$ mkdir certs

$ cd certs

$ sudo openssl req -x509 -nodes -days 365 -newkey rsa:1024 -keyout mycert.pem -out mycert.pem

$ sudo chmod 777 mycert.pem

$ cd ~/.jupyter/

$ vi [jupyter\_notebook\_config.py](https://jupyter_notebook_config.py/)

press "i" and the paste this in file:

**# Configuration file for jupyter-notebook.**

**c = get\_config()**

**# Notebook config this is where you saved your pem cert**

**c.NotebookApp.certfile = u'/home/ubuntu/certs/mycert.pem'**

**# Listen on all IPs**

**c.NotebookApp.ip = '0.0.0.0'**

**# Allow all origins**

**c.NotebookApp.allow\_origin = '\*'**

**# Do not open browser by default**

**c.NotebookApp.open\_browser = False**

**# Fix port to 8888**

**c.NotebookApp.port = 8888**

--  now press escape and then wq!

Now come out to main home directory and run jupyter Notebook command:

$sudo chown ubuntu:ubuntu /home/ubuntu/certs/mycert.pem

$ cd

$sudo chown ubuntu:ubuntu /home/ubuntu/certs/mycert.pem

$ jupyter notebook

Now go to the browser and for me, Chrome didn't work and didn't let me pass through the security warning so I use Safari which gives me a lame warning before letting me in:

copy your DNS from Ec2 and paste it as mentioned below:

https://**DNS**:8888

DNS  = [ec2-52-15-252-200.us](https://ec2-52-15-252-200.us/)-east-2.compute.amazonaws.com

so actually you will be pasting in browser(Safari):

**https://ec2-52-15-252-200.us-east-2.compute.amazonaws.com:8888**

Due to security reasons jupyter notebook will ask you token, which you can get easily from your terminal:

[I 21:52:47.706 NotebookApp] Serving notebooks from local directory: /home/ubuntu

[I 21:52:47.707 NotebookApp] The Jupyter Notebook is running at:

[I 21:52:47.707 NotebookApp] https://ip-172-31-16-113:8888/?token=**675a256b95cd7ae813217424243deeed4908f8b46ea1a252**

[I 21:52:47.707 NotebookApp] or https://127.0.0.1:8888/?token=**675a256b95cd7ae813217424243deeed4908f8b46ea1a252**

(Don't worry I change some of the values so you won't be able to access my virtual machine ;)

so copy this token and paste it in Jupyter Notebook and it will let you in!

Now press CTRL+C to end your session for further installations:

Installing Java and Scala:

$ cd ~/.jupyter/

$ sudo apt-get update

$ cd

$ sudo apt-get install default-jre

$ java -version

$ sudo apt-get install scala

$ scala –version

Now install PIP:

export PATH=$PATH:$HOME/anaconda3/bin

conda install pip

**IMP: Execution will ask you to upgrade Conda but please don't upgrade by simply typing No when prompter will ask you.**

now install:

$ pip install py4j

$ now install Hadoop and Spark:

$ wget https://ftp.nluug.nl/internet/apache/spark/spark-3.0.1/spark-3.0.1-bin-hadoop2.7.tgz

$ sudo tar -zxvf spark-3.0.1-bin-hadoop2.7.tgz

$ export SPARK\_HOME='/home/ubuntu/spark-3.0.1-bin-hadoop2.7'

$ export PATH=$SPARK\_HOME:$PATH

$ export PYTHONPATH=$SPARK\_HOME/python:$PYTHONPATH

$ jupyter notebook

Now again open Jupyter Notebook:

$ Jupyter Notebook

**from pyspark import SparkContext**

**sc = SparkContext()**

I hope this will work smoothly for you !! all the best!

Future readers, you may not need to delete everything and start fresh. If your problem is exclusive to spark/hadoop version, simply run the following two commands:

$ sudo rm -r spark-2.0.0-bin-hadoop2.7

$ sudo rm -r spark-2.0.0-bin-hadoop2.7.tgz

Note: safest bet is to type **$ sudo rm -r s**and then hit tab to autocomplete the spark/hadoop packages

then, install spark and hadoop as described above.