**Setting up the Kafka-Spark and Jupyter lab in Unbuntu environment with Azure Virtual Machine**

**Step1:**

**Installing Java.**

java –version

sudo apt install default-jdk

**Step2:**

**Install Kafka.**

wget <https://downloads.apache.org/kafka/3.6.1/kafka-3.6.1-src.tgz>

tar xzf kafka-3.6.1-src.tgz

**Setting environment variables for kafka.**

cd kafka-3.6.1-src

pwd - **To see the path of kafka**

/home/sandeep/kafka-3.6.1-src -**copy this path and add the path in below file**

sudo nano /etc/environment  
  
add this at the end in the file **:path/bin"**

**Run the Zoopkeeper server:**

bin/zookeeper-server-start.sh config/zookeeper.properties

bin/zookeeper-server-stop.sh config/zookeeper.properties

**If zookeeper got JVM error then kill the java process and run the zookeeper again.**

Killall -9 java

**Run the Kafka server:**

bin/kafka-server-start.sh config/server.properties

bin/kafka-server-stop.sh config/server.properties

**Step3:**

**Install Spark.**

wget <https://archive.apache.org/dist/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3-scala2.13.tgz>

tar xzf spark-3.5.0-bin-hadoop3-scala2.13.tgz

cd spark-3.5.0-bin-hadoop3-scala2.13

**Set the environment variables for Spark same like kafka setup.**

pwd

**Step4:**

**Install Jupyter**

sudo apt install jupyter-core

sudo pip install jupyter

jupyter-notebook list

jupyter notebook --generate-config **– This will generate the path of the config file**

sudo nano /home/sandy/.jupyter/jupyter\_notebook\_config.py **– Opening the config File**

**Copy paste the below code in config file.**

c.NotebookApp.notebook\_dir = '/home/sandy/.jupyter/'

c.NotebookApp.open\_browser = False

c.NotebookApp.port = 5000

c.NotebookApp.ip = '0.0.0.0'

c.NotebookApp.allow\_remote\_access = True

c.ServerApp.allow\_remote\_access = True

**Install Apache**

sudo apt-get install apache2

**It is now time to configure the web server to handle requests to your subdomain and point them to your notebook server. We can do this easily by letting Apache reverse proxy our Jupyter notebook server. Execute the following to move to your Apache sites directory and begin editing your default virtualhosts file (or your config file of choice):**

cd /etc/apache2/sites-available/

sudo nano 000-default.conf

Add the following virtualhost config to the file:

<VirtualHost \*:80>

ServerName sub.domain.com

ProxyPreserveHost On

ProxyPass /api/kernels/ ws://localhost:5000/api/kernels/

ProxyPassReverse /api/kernels/ http://localhost:5000/api/kernels/

ProxyPass / http://localhost:5000/

ProxyPassReverse / http://localhost:5000/

</VirtualHost>

**Be sure to change the port from 5000 to the port you specified in your Jupyter config file. If you did not specify a port, the Jupyter server defaults to port 8888. Also don’t forget to replace sub.domain.com with your actual subdomain.**

**We now need to enable a few Apache modules to allow proxy functionality. Type the following in your terminal:**

sudo a2enmod proxy sudo a2enmod proxy\_http sudo a2enmod proxy\_wstunnel

**We can now restart Apache using**

sudo service apache2 restart

**Start Jupyter Notebook in srever.**

jupyter notebook &

**Now the jupyter server will run in VM localhost 5000 port**

**To Access Jupyter Notebook in your local desktop browser.**

**Create an inbound rule in nsg(Network Security Group) in AzureVM**

sudo ufw allow 5000 **-This command will allow Unbutu firewall to receive http: request from Port 5000 of AzureVm.**

**Now check the connection by pasting the public Ip address of your AzureVm in your local desktop browser. The Jupyter notebook will run in your browser.**

**Enabling Spark and Java environment in Jupyter-Notebook(pyspark)**

sudo nano ~/.zshrc

**Paste the below java path in above command.**

export JAVA\_HOME=/usr/lib/jvm/java-11-openjdk-amd64

source ~/.zshrc **-This command will save the java home (Run this command before starting jupyter server).**

sudo mv spark-3.5.0-bin-hadoop3-scala2.13 /opt/spark-3.5.0 -**This command move the spark directory to /opt/spark-3.5.0.**

sudo ln -s /opt/spark-3.5.0 /opt/spark. **-This will create the soft link for Spark**

sudo nano ~/.bashrc **- This will open the bash file**

**Copy-Paste the below in this above file:**

export SPARK\_HOME=/opt/spark

export PATH=$PATH:$SPARK\_HOME/bin:$SPARK\_HOME/sbin

export PYSPARK\_PYTHON=/usr/bin/python3

export PYSPARK\_DRIVER\_PYTHON=jupyter

export PYSPARK\_DRIVER\_PYTHON\_OPTS='notebook'

alias pyspark='pyspark --packages org.apache.spark:spark-sql-kafka-0-10\_2.13:3.5.0'

source ~/.bashrc. **This command will save the Spark home (Run this command before starting jupyter server).**

pyspark. -**This command will run the jupyter server.**

kill $(lsof -ti :5000) kill port -**If jupyter server is not stopped properly than kill the port using this command.**