Apache Spark - Installation

# JAVA Installation

Type the following command in terminal: sudo apt install openjdk-8-jdk openjdk-8-jre Update and install the installer package

sudo apt update

Java Path Checking:

1. to check the path, type:

sudo update-alternatives --config java

1. to open environment file

sudo nano /etc/environment

1. once the file is open, type the following

JAVA\_HOME="path" here path output get in step 1

1. finally, reconfirm your java environment variable by this command

echo $JAVA\_HOME

1. Check the version

$java -version

SCALA Installation:

1.Run the command

sudo apt install scala

2​.​ check the version, by the following command

scala-version

# Apache Spark Installation

Make a folder and type the following within this folder

Download:

Wget https://archive.apache.org/dist/spark/spark-2.2.1/spark-2.2.1-bin-hadoop2.6.tgz Install:

sudo tar -xvf /home/sehrish/spark-2.2.1-bin-hadoop2.6.tgz

Type the following in the main directory

Nano .bashrc

Set Environment variables export SPARK\_PATH=/home/haris/spark/spark-2.2.1-bin-hadoop2.6 path where spark is downloaded and installed

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

Save and exit, then type the following command source ~/.bashrc Then spark-shell

# Pyspark Installation

Open linux terminal and create directory to save installer file:

mkdir anacondainstaller Change directory:

cd anacondainstaller

Now download the installer package:

wget [https://repo.anaconda.com/archive/Anaconda3-2019.10-Linux-x86\_64.s](https://repo.anaconda.com/archive/Anaconda3-2019.10-Linux-x86_64.sh)​ [h](https://repo.anaconda.com/archive/Anaconda3-2019.10-Linux-x86_64.sh) Now make it executable:

chmod +x Anaconda3-5.1.0-Linux-x86\_64.sh

Run the installer:

./Anaconda3-5.0.1-Linux-x86\_64.sh Set path variables:

export SPARK\_PATH=/home/haris/spark/spark-2.2.1-bin-hadoop2.6 export PATH=$PATH:~/anaconda3/bin export PYTHONPATH=$SPARK\_PATH/python:$PYTHONPATH export PYSPARK\_DRIVER\_PYTHON="jupyter" export PYSPARK\_DRIVER\_PYTHON\_OPTS="notebook" export PYSPARK\_PYTHON=python3

# Configure the jupyter notebook server for remote access

**Set-up:** Here, let’s define the local user and host as localuser​ and localhost​ respectively. Similarly, let’s define the remote user and remote host as remoteuser​ and remotehost.​

Run jupyter nnotebook with no browser and specify port XXXX

$ jupyter notebook --no\_browser --port=X

Forward via ssh in your local machine

$ ssh \_N \_f \_L localhost:YYYY localhost:XXXX remoteuser@remotehost

Step 1: Run Jupyter Notebook from remote machine​

Log-in to your remote machine

**remoteuser@remotehost: jupyter notebook --no-browser --port=XXXX**

***# Note: Change XXXX to the port of your choice. Usually, the default is 8888.***

***# You can try 8889 or 8890 as well.***

**Step 2: Forward port XXXX to YYYY and listen to it**

**localuser@localhost:**

**ssh**

**-**

**N**

**f**

**-**

**-**

**L**

**localhost:YYYY:localhost:XXXX**

**remoteuser@remotehost**

**Step 3: Fire-up Jupyter Notebook**

**localhost:YYYY**

**Closing all connections**

To close connection

s

**localuser@localhost: sudo netstat -lpn |grep :YYYY**