This document is to install apache spark on single node aws Ubuntu instance.

Step1 :Update OS

sudo apt-get update

```
ubuntu@ip-172-31-62-27:~$ sudo apt-get update
```

Step2:Install Java 8

sudo add-apt-repository ppa:webupd8team/java –y

```
ubuntu@ip-172-31-62-27:~$ sudo add-apt-repository ppa:webupd8team/java -y
```

sudo apt-get update

sudo apt-get install oracle-java8-installer

sudo apt-get install oracle-java8-set-default

Step3 :Check java version.

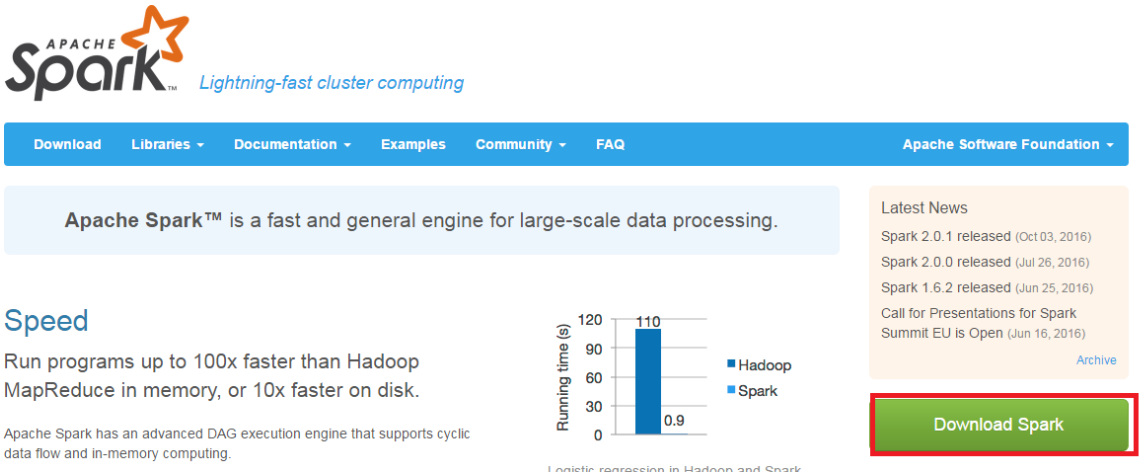
java -version

```
ubuntu@ip-172-31-62-27:~$ java -version
java version "1.8.0_111"
Java(TM) SE Runtime Environment (build 1.8.0_111-b14)
Java HotSpot(TM) 64-Bit Server VM (build 25.111-b14, mixed mode)
```

Step4 :Apache Spark Setup

a) Go to official URL of Apache spark http://spark.apache.org/

b) Click on **“Download Spark”**



c) Select **“Select Apache Mirror”** option from dropdown .(If you have are working on windows/Linux OS with GUI ,Kindly select **“Directly Download ”** option)

d) Click on **"spark-2.0.1-bin-hadoop2.7.tgz"** file.

f) Copy below mentioned URL.

g) Using wget command download spark tar file.

wget http://mirror.fibergrid.in/apache/spark/spark-2.0.1/spark-2.0.1-bin-hadoop2.7.tgz

```
ubuntu@ip-172-31-62-27:~$ wget http://mirror.fibergrid.in/apache/spark/spark-2.0.1/spark-2.0.1-bin-hadoop2.7.tgz
--2016-11-01 14:28:09--  http://mirror.fibergrid.in/apache/spark/spark-2.0.1/spark-2.0.1-bin-hadoop2.7.tgz
Resolving mirror.fibergrid.in (mirror.fibergrid.in)... 103.194.116.38, 2400:4a80::57
Connecting to mirror.fibergrid.in (mirror.fibergrid.in)|103.194.116.38|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 187275308 (179M) [application/x-gzip]
Saving to: 'spark-2.0.1-bin-hadoop2.7.tgz'

spark-2.0.1-bin-hadoop2.7  17%[=====>                              ]  30.93M  3.34MB/s    eta 57s
```

h) **Untar** spark tar file using **tar** command and rename file to **spark** using mv command.

tar -xzvf spark-2.0.1-bin-hadoop2.7.tgz

```
ubuntu@ip-172-31-62-27:~$ tar -xzvf spark-2.0.1-bin-hadoop2.7.tgz
```

mv spark-2.0.1-bin-hadoop2.7 spark

```
ubuntu@ip-172-31-62-27:~$ mv spark-2.0.1-bin-hadoop2.7 spark
```

i)Setup SPARK_HOME and PATH in bashrc file and then compile bashrc file.

vi ~/.bashrc

export SPARK_HOME=/home/ubuntu/spark

export PATH=$PATH:$SPARK_HOME/bin

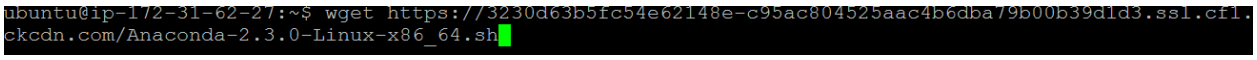
```
# ~/.bashrc: executed by bash(1) for no
# see /usr/share/doc/bash/examples/star
# for examples
export SPARK_HOME=/home/ubuntu/spark
export PATH=$PATH:$SPARK_HOME/bin
```

source ~/.bashrc

Step4 :Download and install Anaconda for pyspark

a) Download sh file .

wget https://3230d63b5fc54e62148e-c95ac804525aac4b6dba79b00b39d1d3.ssl.cf1.rackcdn.com/Anaconda-2.3.0-Linux-x86_64.sh



b) Execute below command and accept the conditions.

bash Anaconda-2.3.0-Linux-x86_64.sh