

Tutorial – Install Spark on Windows Machine

Ryan Jones – COP 6526 – Dr. Wang – Fall 2020 MSDA Program

1. I have a laptop running Windows 10. I will be installing Spark using the WSL2 interface and Ubuntu.
2. Install Java and Set Path.

```
# sudo apt update
# sudo apt install default-jre default-jdk
# java -version
# export JAVA_HOME=/lib/jvm/java-11-openjdk-amd64
# echo $JAVA_HOME
```

```
rjones@RyanLaptop: ~
(base) rjones@RyanLaptop:~$ echo $JAVA_HOME
/lib/jvm/java-11-openjdk-amd64
(base) rjones@RyanLaptop:~$
```

3. Setup SSH Without Password.

```
# ssh-keygen -t rsa
# cd ~/.ssh/
# cat id_rsa.pub >> authorized_keys
# ssh localhost
```

```
rjones@RyanLaptop: ~
(base) rjones@RyanLaptop:~$ ssh localhost
Welcome to Ubuntu 20.04.1 LTS (GNU/Linux 4.19.128-microsoft-standard x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/advantage

System information as of Mon Sep 28 16:12:00 EDT 2020

System load:  0.04               Processes:    30
Usage of /:   3.0% of 250.98GB   Users logged in: 0
Memory usage: 12%               IPv4 address for eth0: 192.168.98.201
Swap usage:   0%

5 updates can be installed immediately.
0 of these updates are security updates.
To see these additional updates run: apt list --upgradable

*** System restart required ***
Last login: Mon Sep 21 19:43:32 2020 from 127.0.0.1
(base) rjones@RyanLaptop:~$
```

```
# service --status-all
```

```
[ + ] ssh
[ + ] udev
[ - ] ufw
[ - ] unattended-upgrades
[ - ] uuid
[ - ] x11-common
(base) rjones@RyanLaptop:~$
```

4. Install Apache Spark and Configure

```
# wget https://downloads.apache.org/spark/spark-3.0.1/spark-3.0.1-bin-hadoop2.7.tgz
# tar xf spark-3.0.1-bin-hadoop2.7.tgz
```

```
# cd ~
# vi .bashrc
```

Enter the below information into the .bashrc file and save it.

```
rjones@RyanLaptop: ~
GNU nano 4.8 .bashrc
# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# for examples

#Set Spark
export SPARK_HOME=/home/rjones/spark-3.0.1-bin-hadoop2.7
export PATH=$PATH:$SPARK_HOME/bin
```

```
# cd $SPARK_HOME/conf
# cp spark-env.sh.template spark-env.sh
# nano spark-env.sh
```

Enter the below information into the spark-env.sh file and save it.

```
rjones@RyanLaptop: ~/spark-3.0.1-bin-hadoop2.7/conf
GNU nano 4.8
export JAVA_HOME=/usr/lib/jvm/default-java
export SPARK_LOCAL_IP="hostname -i"
export SPARK_MASTER_HOST=192.168.98.201
if [ "${SPARK_LOCAL_IP}" = "${SPARK_MASTER_HOST}" ]
then
export SPARK_MASTER_PORT=7077
else
export SPARK_MASTER_PORT=6066
fi
export SPARK_WORKER_MEMORY=8g
export SPARK_WORKER_CORES=2
export SPARK_WORKER_INSTANCES=1
export MASTER=spark://${SPARK_MASTER_HOST}:7077
```

```
# cp slaves.template slaves
```

5. Run a Program on Spark

```
# cd $SPARK_HOME
# ./sbin/start-all.sh
# jps
```

```
rjones@RyanLaptop: ~/spark-3.0.1-bin-hadoop2.7
(base) rjones@RyanLaptop:~/spark-3.0.1-bin-hadoop2.7$ ./sbin/start-all.sh
starting org.apache.spark.deploy.master.Master, logging to /home/rjones/spark-3.0.1-bin-hadoop2.7/logs/spark-rjones-org.
apache.spark.deploy.master.Master-1-RyanLaptop.out
localhost: starting org.apache.spark.deploy.worker.Worker, logging to /home/rjones/spark-3.0.1-bin-hadoop2.7/logs/spark-
rjones-org.apache.spark.deploy.worker.Worker-1-RyanLaptop.out
(base) rjones@RyanLaptop:~/spark-3.0.1-bin-hadoop2.7$ jps
6138 Worker
6222 Jps
5983 Master
(base) rjones@RyanLaptop:~/spark-3.0.1-bin-hadoop2.7$
```

```
# spark-submit --class org.apache.spark.examples.SparkPi --master spark://192.168.98.201:6066 --
executor-memory 6G /home/rjones/spark-3.0.1-bin-hadoop2.7/examples/jars/spark-
examples_2.12-3.0.1.jar 10
```

```
rjones@RyanLaptop: ~  
Pi is roughly 3.141187141187141
```

```
# cd $SPARK_HOME
```

```
# ./sbin/stop-all.sh
```

```
# jps
```

```
rjones@RyanLaptop: ~/spark-3.0.1-bin-hadoop2.7
```

```
(base) rjones@RyanLaptop:~$ cd $SPARK_HOME  
(base) rjones@RyanLaptop:~/spark-3.0.1-bin-hadoop2.7$ ./sbin/stop-all.sh  
localhost: stopping org.apache.spark.deploy.worker.Worker  
stopping org.apache.spark.deploy.master.Master  
(base) rjones@RyanLaptop:~/spark-3.0.1-bin-hadoop2.7$ jps  
6683 Jps  
(base) rjones@RyanLaptop:~/spark-3.0.1-bin-hadoop2.7$
```