

“Spark Installation”

Submitted By

Varsha Varadarajan

[142051002]

M. Tech. (Computer Engineering)

Under the guidance of

Dr. V. B. Nikam



Department of Computer Engineering and Information
Technology

Veermata Jijabai Technological Institute

Mumbai-400019

(An Autonomous Institute affiliated to University of
Mumbai)

2015-16

Contents

1	Spark Single Node Installation	1
2	Spark Multi-Node Cluster Configuration	4
2.1	Creating the cluster	5

List of Figures

1	spark-env.sh File	5
2	bashrc File	6
3	Web UI	6
4	slaves File	7
5	Web UI	8

1 Spark Single Node Installation

The following steps show how to install Apache Spark.

1. Verifying Java Installation Java installation is one of the mandatory things in installing Spark. Try the following command to verify the JAVA version.

```
$java -version
```

If Java is already installed on your system, you get to see the following response -

```
java version "1.7.0_71"  
Java(TM) SE Runtime Environment (build 1.7.0_71-b13)  
Java HotSpot(TM) Client VM (build 25.0-b02, mixed mode)
```

In case you do not have Java installed on your system, then Install Java before proceeding to next step.

2. Verifying Scala installation You should Scala language to implement Spark. So let us verify Scala installation using following command.

```
$scala -version
```

If Scala is already installed on your system, you get to see the following response -

```
Scala code runner version 2.11.6 — Copyright 2002–2013,  
LAMP/EPFL
```

In case you don't have Scala installed on your system, then proceed to next step for Scala installation.

3. Downloading Scala Download the latest version of Scala by visit the following link [Download Scala](#). For this tutorial, we are using scala-2.11.6 version. After downloading, you will find the Scala tar file in the download folder.

4. Installing Scala Follow the below given steps for installing Scala.

- (a) Extract the Scala tar file Type the following command for extracting the Scala tar file.

```
\$ tar xvf scala-2.11.6.tgz
```

- (b) Move Scala software files Use the following commands for moving the Scala software files, to respective directory (/usr/local/scala).

```
\$ su -  
Password :  
# cd /home/Hadoop/Downloads/  
# mv scala-2.11.6 /usr/local/scala  
# exit
```

- (c) Set PATH for Scala Use the following command for setting PATH for Scala.

```
\$ export PATH = $PATH:/usr/local/scala/bin
```

5. Downloading Apache Spark Download the latest version of Spark by visiting the following link [Download Spark](#). For this tutorial, we are using spark-1.3.1-bin-hadoop2.6 version. After downloading it, you will find the Spark tar file in the download folder.

6. Installing Spark Follow the steps given below for installing Spark.

- (a) Extracting Spark tar The following command for extracting the spark tar file.

```
\$ tar xvf spark-1.3.1-bin-hadoop2.6.tgz
```

- (b) Moving Spark software files The following commands for moving the Spark software files to respective directory (/usr/local/spark).

```
\$ su -  
Password :  
# cd /home/Hadoop/Downloads/  
# mv spark-1.3.1-bin-hadoop2.6 /usr/local/spark  
# exit
```

- (c) Setting up the environment for Spark Add the line to `/.bashrc` file. It means adding the location, where the spark software file are located to the `PATH` variable.

```
export PATH = $PATH:/usr/local/spark/bin
```

- (d) Use the following command to source the `/.bashrc` file.
-

```
\$ source ~/.bashrc
```

7. Verifying the Spark Installation Write the following command for opening Spark shell.
-

```
\$spark-shell
```

If spark is installed successfully then you will find the following output.

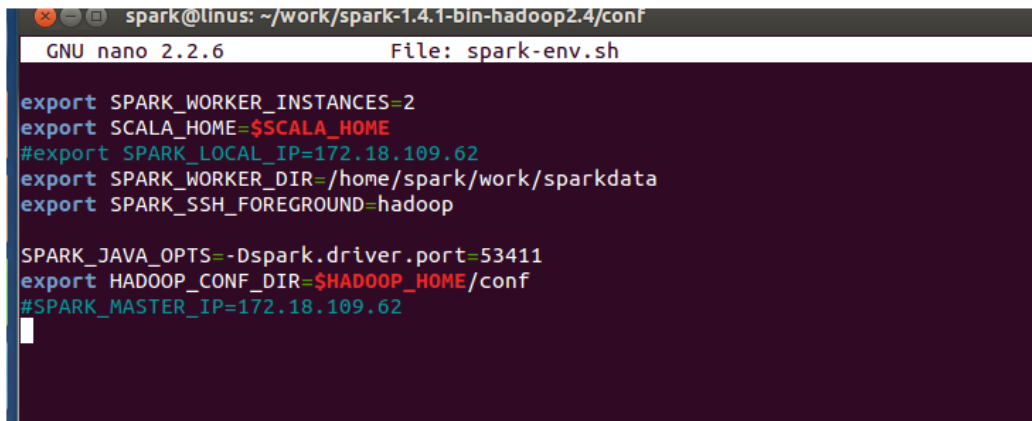
```
Spark assembly has been built with Hive, including  
Datanucleus jars on classpath  
Using Spark's default log4j profile:  
org/apache/spark/log4j-defaults.properties  
15/06/04 15:25:22 INFO SecurityManager: Changing view  
acls to: hadoop  
15/06/04 15:25:22 INFO SecurityManager: Changing modify  
acls to: hadoop  
15/06/04 15:25:22 INFO SecurityManager: SecurityManager:  
authentication disabled;  
ui acls disabled;  
users with view permissions: Set(hadoop);  
users with modify permissions: Set(hadoop)  
15/06/04 15:25:22 INFO HttpServer: Starting HTTP Server
```

```
15/06/04 15:25:23 INFO Utils: Successfully started
service 'HTTP class server' on port 43292.
Welcome to SPARK
Using Scala version 2.10.4 (Java HotSpot(TM)
64-Bit Server VM, Java 1.7.0_71)
Type in expressions to have them evaluated.
Spark context available as sc
scala>
```

2 Spark Multi-Node Cluster Configuration

1. Download apache spark:
`http://spark.apache.org/downloads.html`
2. Download compatible version of scala from
`http://www.scala-lang.org/download/`
3. Double click the scala.deb file to install scala using Package Installer in Ubuntu.
4. Unzip the Spark and Scala zip files in a location of your choice.
5. Locate conf folder in Spark folder/
6. Create a copy of spark-env.sh.template and rename to spark-env.sh
7. Add below lines to the file `/home/spark-1.0.1-bin-hadoop2/conf/spark-env.sh`. Refer Figure 1.

```
export SPARK_LOCALIP=(ip address of the local machine)
export SPARK_MASTER_IP=(ip address of the master machine)
export SPARK_WORKER_CORE=1
export SPARK_WORKER_INSTANCE=4
export SPARK_WORKER_MEMORY=2g
```

A screenshot of a terminal window with a dark background. The title bar shows 'spark@linus: ~/work/spark-1.4.1-bin-hadoop2.4/conf'. The editor is GNU nano 2.2.6, editing the file 'spark-env.sh'. The content of the file is as follows:

```
export SPARK_WORKER_INSTANCES=2
export SCALA_HOME=$SCALA_HOME
#export SPARK_LOCAL_IP=172.18.109.62
export SPARK_WORKER_DIR=/home/spark/work/sparkdata
export SPARK_SSH_FOREGROUND=hadoop

SPARK_JAVA_OPTS=-Dspark.driver.port=53411
export HADOOP_CONF_DIR=$HADOOP_HOME/conf
#SPARK_MASTER_IP=172.18.109.62
```

Figure 1: spark-env.sh File

8. Add following lines to the file `/.bashrc`. Refer Figure 2.

```
export SPARK_HOME=/home/spark/work/spark-1.4.1-bin-hadoop
export SCALA_HOME=/home/spark/work/scala-2.10.4
export PATH=$PATH:$SPARK_HOME/bin
export PATH=$PATH:$SCALA_HOME/bin
```

9. Test whether scala is working:
Type `scala` on terminal : Output should be version of scala.
10. Test whether spark is working:
Type `./bin/spark-shell`
11. If spark shell powers up, then spark has been installed on the machine.
Refer Figure 3.
12. Follow the above 8 steps on all the machines.

2.1 Creating the cluster

1. In the master machine, locate the `conf` folder in `spark` folder.
2. Locate `slaves.sh` file and make a copy of the file and rename it `slaves`.
3. Edit `slaves` file and add the ip addresses of the slave machines to it.
Refer Figure 4.


```
GNU nano 2.2.6      File: slaves
# A Spark Worker will be started on each of the machines listed below.
linus
mark
kevin
aaron
```

Figure 4: slaves File

6. In the browser type “ip address of master machine:8080”. This will show all the worker instances working. Refer 5.

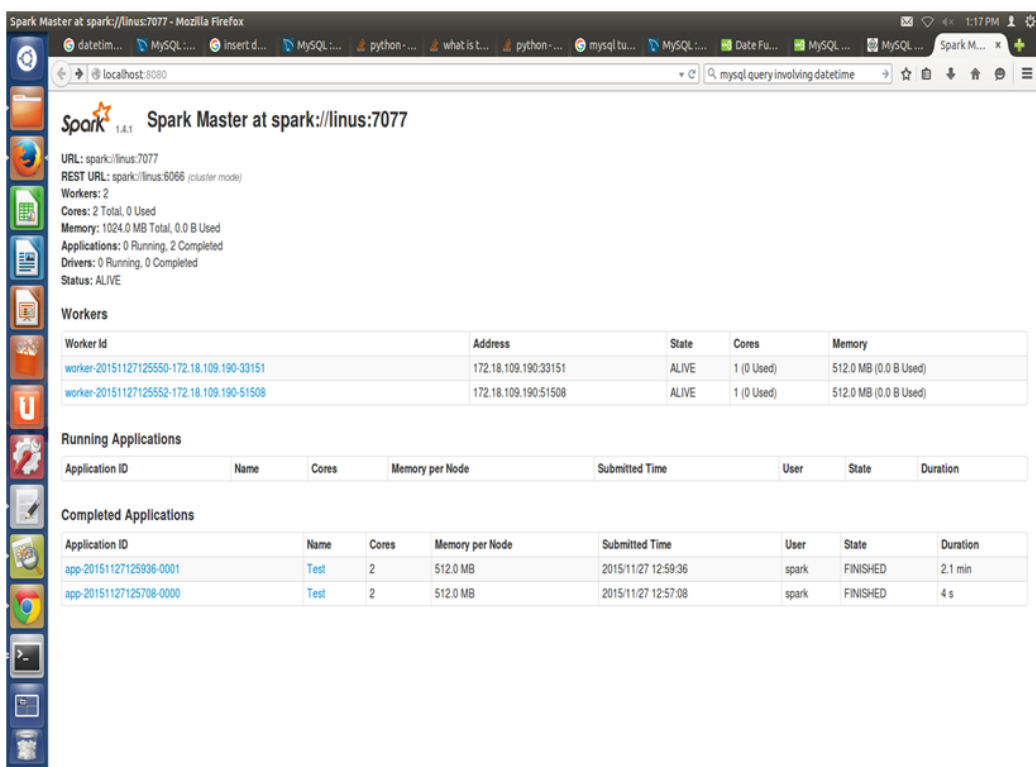


Figure 5: Web UI