Spark on hadoop 集群搭建教程

环境：

安装 spark 之前需要先安装 hadoop 集群。

ubuntu14.04

Hadoop 2.6.0

scala 2.11.7

spark 1.6.0

集群：

1 个 master　nathychen@Master　192.168.10.45

1 个 slave　nathychen@Slave　192.168.10.46

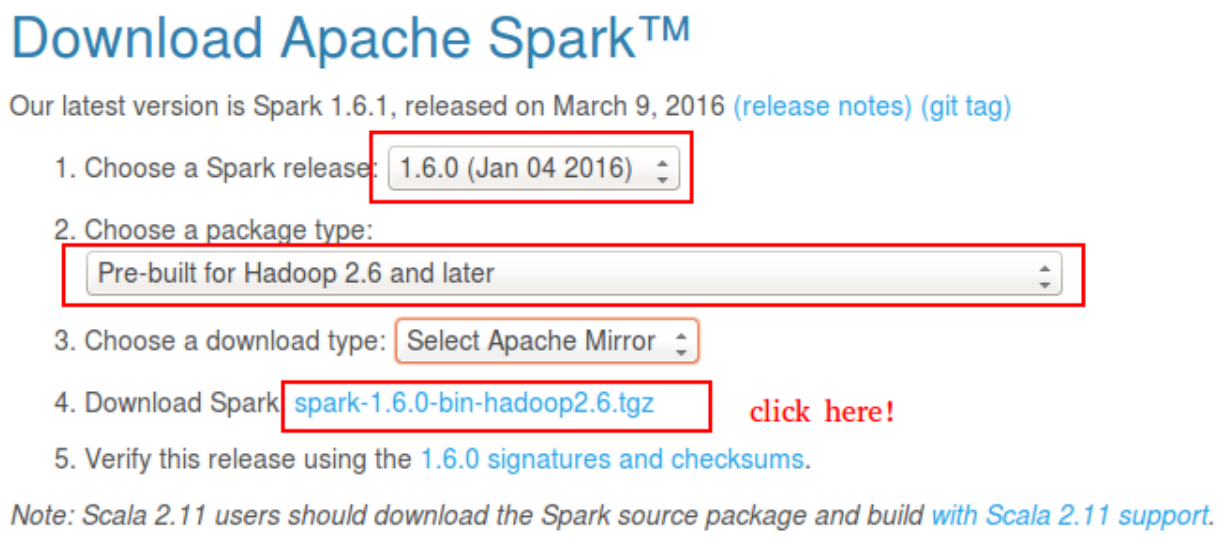
1、下载 scala 2.11.7

下载 scala 包：http://www.scala-lang.org/download/2.11.7.html

2、下载 spark 1.6.0

下载 spark 包：http://spark.apache.org/downloads.html

选择如下：



注：通过浏览器下载的安装包默认保存在./下载里

3、安装 scala（在 Master 上）

- mkdir /usr/local/scala　　#在/usr/local 下新建文件夹 scala

- cp ~/下载/scala-2.11.7.tgz /usr/local/scala/　　#将./下载里的 scala 包拷贝到 scala 文件夹
- cd /usr/local/scala
- tar -zxvf scala-2.11.7.tgz　　#解压
- ln -s /usr/local/scala/scala-2.11.7 /usr/local/scala/scala　　#创建软链接

接下来，修改环境变量，添加 SCALA_HOME，并修改 PATH，若提示没有权限，则在命令前使用 sudo：

- vim /etc/profile.d/java.sh

在新建的 java.sh 文件中写入：

```
export SCALA_HOME=/usr/local/scala/scala-2.11.7
export PATH=$JAVA_HOME/bin:$HADOOP_HOME/bin:$SCALA_HOME/bin:$PATH
```

保存退出后执行如下命令，使环境配置立即生效：

- source /etc/profile

通过以下命令验证是否安装成功

- scala -version

显示如下：

```
nathychen@Slave:/usr/local$ scala -version
Scala code runner version 2.11.7 -- Copyright 2002-2013, LAMP/EPFL
nathychen@Slave:/usr/local$
```

4、安装 spark（在 Master 上）

- mkdir /usr/local/spark　　#新建 spark 文件夹
- cp ~/下载/spark-1.6.0-bin-hadoop2.6.tgz /usr/local/spark/
- tar zxvf spark-16.0-bin-hadoop2.6.tgz

接下来，修改环境变量，将 SPARK_HOME 添加进去，并修改 PATH

- vim /etc/profile.d/java.sh

在 java.sh 添加：

export SPARK_HOME=/usr/local/spark/spark-1.6.0-bin-hadoop2.6

export PATH=$JAVA_HOME/bin:$HADOOP_HOME/bin:$SCALA_HOME/bin:$SPARK_HOME/bin:$PATH

保存退出后执行如下命令，使环境配置立即生效：

- source /etc/profile

接下来，修改 spark 的配置文件

4.1、修改 spark-env.sh 文件：

- cd /usr/local/spark/spark-1.6.0-bin-hadoop2.6/conf
- cp spark-env.sh.template spark-env.sh
- vim spark-env.sh

在 spark-env.sh 最底端追加：

export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64

export SCALA_HOME=/usr/local/scala/scala

export SPARK_MASTER_IP=192.168.10.45

export SPARK_WORKER_MEMORY=512m

export HADOOP_CONF_DIR=/usr/local/hadoop/etc/hadoop

4.2、修改 slaves 文件

- cp slaves.template slaves
- vim slaves

添加如下（可能有默认 localhost，将其修改成 Master）：

Master

Slave

5、将上述所有在 Master 上的安装拷贝到 Slave 上

直接复制到 Slave 的/usr/local 中容易出现权限问题，建议先复制到/tmp 文件夹中，再从 Slave 上将其拷贝过去：

5.1、scala

在 Master 上

- scp -r /usr/local/scala nathychen@Slave:/tmp

在 Slave 上

- mv ./tmp/scala /usr/local/　　#将 scala 文件夹移动到/usr/local/下

5.2、spark

在 Master 上

- scp -r /usr/local/spark nathychen@Slave:/tmp

在 Slave 上

- mv ./tmp/spark /usr/local/　　#将 spark 文件夹移动到/usr/local 下

5.3、环境变量

在 Master 上

- scp -r /etc/profile.d/java.sh nathychen@Slave:/tmp

在 Slave 上

- mv ./tmp/java.sh /etc/profile.d/
- source /etc/profile

6、在 Master 上修改 spark 文件夹的权限

- sudo chown -R -v nathychen:users /usr/local/spark

可进入/usr/local 文件夹通过如下命令查看 spark 文件夹的权限：

- ls -l

```
nathychen@Master:/usr/local/spark/spark-1.6.0-bin-hadoop2.6$ cd /usr/local
nathychen@Master:/usr/local$ ls -l
总用量 44
drwxr-xr-x  2 root      root     4096  8月  5  2015 bin
drwxr-xr-x  2 root      root     4096  8月  5  2015 etc
drwxr-xr-x  2 root      root     4096  8月  5  2015 games
drwxr-xr-x 13 nathychen hadoopNC 4096  3月 31 16:00 hadoop
drwxr-xr-x  2 root      root     4096  8月  5  2015 include
drwxr-xr-x  4 root      root     4096  8月  5  2015 lib
lrwxrwxrwx  1 root      root        9 10月 20 16:42 man -> share/man
drwxr-xr-x  2 root      root     4096  8月  5  2015 sbin
drwxr-xr-x  3 root      root     4096  4月  6 11:03 scala
drwxr-xr-x  7 root      root     4096  8月  5  2015 share
drwxr-xr-x  3 nathychen users    4096  4月  6 15:59 spark
drwxr-xr-x  2 root      root     4096  8月  5  2015 src
nathychen@Master:/usr/local$
```

可看到，spark 文件夹的所属者已被该为 nathychen。

7、启动 spark 集群

- cd /usr/local/spark/spark-1.6.0-bin-hadoop2.6/sbin
- ./start-all.sh
- jps

```
nathychen@Master:/usr/local$ jps
15037 NameNode
18097 Jps
16485 Master
16654 Worker
15877 NodeManager
15564 ResourceManager
15397 SecondaryNameNode
15197 DataNode
nathychen@Master:/usr/local$
```

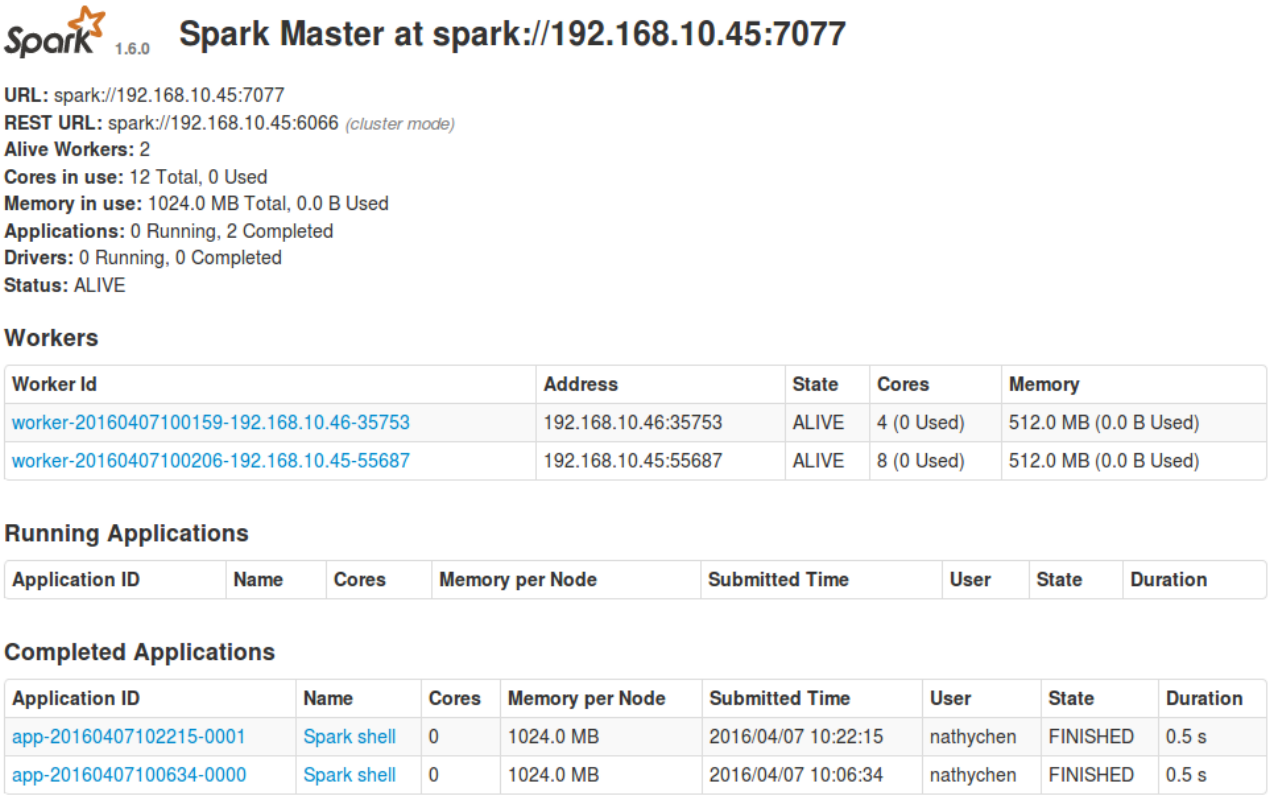
比 hadoop 集群启动时多了 Master 和 Worker。

输入如下命令

- /usr/local/spark/spark-1.6.0-bin-hadoop2.6/bin/spark-shell

出现 scala>则说明成功。

在浏览器中输入 192.168.10.45：8080 时，会看到如下图，有两个 worker。



Spark 1.6.0 **Spark Master at spark://192.168.10.45:7077**

URL: spark://192.168.10.45:7077
REST URL: spark://192.168.10.45:6066 *(cluster mode)*
Alive Workers: 2
Cores in use: 12 Total, 0 Used
Memory in use: 1024.0 MB Total, 0.0 B Used
Applications: 0 Running, 2 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers

Worker Id	Address	State	Cores	Memory
worker-20160407100159-192.168.10.46-35753	192.168.10.46:35753	ALIVE	4 (0 Used)	512.0 MB (0.0 B Used)
worker-20160407100206-192.168.10.45-55687	192.168.10.45:55687	ALIVE	8 (0 Used)	512.0 MB (0.0 B Used)

Running Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration

Completed Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
app-20160407102215-0001	Spark shell	0	1024.0 MB	2016/04/07 10:22:15	nathychen	FINISHED	0.5 s
app-20160407100634-0000	Spark shell	0	1024.0 MB	2016/04/07 10:06:34	nathychen	FINISHED	0.5 s