**[spark-1.2.0 集群环境搭建](http://stark-summer.iteye.com/blog/2173219)**

1、下载scala2.11.4版本 下载地址为：<http://www.scala-lang.org/download/2.11.4.html> ，也可以使用wget http://downloads.typesafe.com/scala/2.11.4/scala-2.11.4.tgz?\_ga=1.248348352.61371242.1418807768

2、解压和安装： 解压 ：[spark@S1PA11 scala]$ tar -xvf scala-2.11.4.tgz  ,安装：[spark@S1PA11 scala]$ mv scala-2.11.4 ~/opt/

3、编辑 ~/.bash\_profile文件 增加SCALA\_HOME环境变量配置，

export JAVA\_HOME=/home/spark/opt/java/jdk1.6.0\_37  
export CLASSPATH=.:$JAVA\_HOME/jre/lib:$JAVA\_HOME/lib:$JAVA\_HOME/lib/tools.jar  
export SCALA\_HOME=/home/spark/opt/scala-2.11.4  
export HADOOP\_HOME=/home/spark/opt/hadoop-2.6.0  
PATH=$PATH:$HOME/bin:$JAVA\_HOME/bin:${SCALA\_HOME}/bin

立即生效 bash\_profile  ，[spark@S1PA11 scala]$ source ~/.bash\_profile

4、验证scala：[spark@S1PA11 scala]$ scala -version  
Scala code runner version 2.11.4 -- Copyright 2002-2013, LAMP/EPFL

[spark@S1PA11 scala]$ scala  
Welcome to Scala version 2.11.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.6.0\_37).  
Type in expressions to have them evaluated.  
Type :help for more information.

scala> var str = "SB is"+"SB"  
str: String = SB isSB  
  
scala>

5、copy到slave机器 ，[spark@S1PA11 scala]$ scp  ~/.bash\_profile  spark@10.126.45.56:~/.bash\_profile

6、下载spark，wget <http://d3kbcqa49mib13.cloudfront.net/spark-1.2.0-bin-hadoop2.4.tgz>

7、在master主机配置spark ：

将下载的spark-1.2.0-bin-hadoop2.4.tgz 解压到 ~/opt/目前了即 ~/opt/spark-1.2.0-bin-hadoop2.4，配置环境变量SPARK\_HOME

# set  java env  
export JAVA\_HOME=/home/spark/opt/java/jdk1.6.0\_37  
export CLASSPATH=.:$JAVA\_HOME/jre/lib:$JAVA\_HOME/lib:$JAVA\_HOME/lib/tools.jar  
export SCALA\_HOME=/home/spark/opt/scala-2.11.4  
export HADOOP\_HOME=/home/spark/opt/hadoop-2.6.0  
export SPARK\_HOME=/home/spark/opt/spark-1.2.0-bin-hadoop2.4  
PATH=$PATH:$HOME/bin:$JAVA\_HOME/bin:${SCALA\_HOME}/bin:${SPARK\_HOME}/bin:${HADOOP\_HOME}/bin

配置完成后使用source命令使配置生效

进入 spark conf目录：

[spark@S1PA11 opt]$ cd spark-1.2.0-bin-hadoop2.4/  
[spark@S1PA11 spark-1.2.0-bin-hadoop2.4]$ ls  
bin  conf  data  ec2  examples  lib  LICENSE  logs  NOTICE  python  README.md  RELEASE  sbin  work  
[spark@S1PA11 spark-1.2.0-bin-hadoop2.4]$ cd conf/  
[spark@S1PA11 conf]$ ls  
fairscheduler.xml.template  metrics.properties.template  slaves.template               spark-env.sh  
log4j.properties.template   slaves                       spark-defaults.conf.template  spark-env.sh.template

first ：修改slaves文件，增加两个slave节点S1PA11、S1PA222

[spark@S1PA11 conf]$ vi slaves  
S1PA11  
S1PA222

second：配置spark-env.sh

首先把spark-env.sh.template copy spark-env.sh

vi spark-env.sh文件 在最下面增加：

export JAVA\_HOME=/home/spark/opt/java/jdk1.6.0\_37  
export SCALA\_HOME=/home/spark/opt/scala-2.11.4  
export SPARK\_MASTER\_IP=10.58.44.47  
export SPARK\_WORKER\_MEMORY=2g  
export HADOOP\_CONF\_DIR=/home/spark/opt/hadoop-2.6.0/etc/hadoop

HADOOP\_CONF\_DIR是Hadoop配置文件目录，SPARK\_MASTER\_IP主机IP地址，SPARK\_WORKER\_MEMORY是worker使用的最大内存

完成配置后，将spark目录copy slave机器 scp -r ~/opt/spark-1.2.0-bin-hadoop2.4  spark@10.126.45.56:~/opt/

8、启动spark分布式集群并查看信息

[spark@S1PA11 sbin]$ ./start-all.sh

查看：

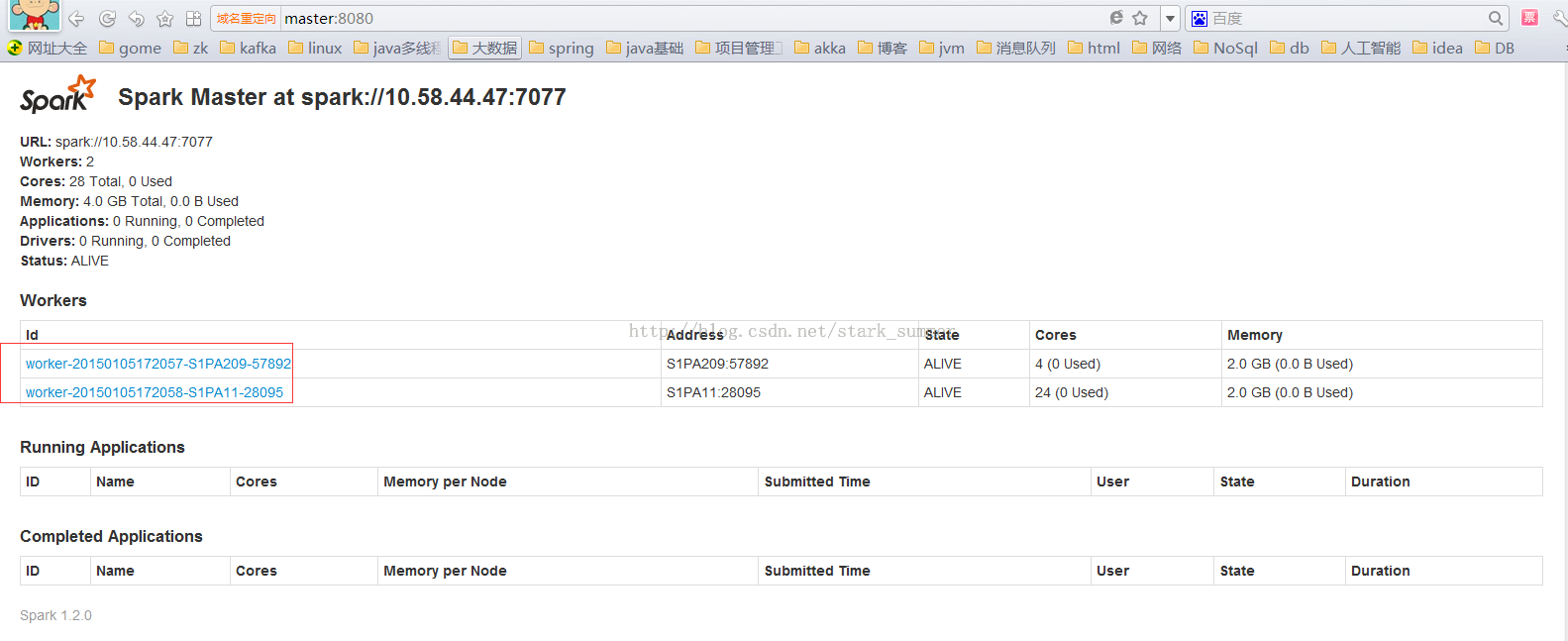
[spark@S1PA11 sbin]$ jps  
31233 ResourceManager  
27201 Jps  
30498 NameNode  
30733 SecondaryNameNode  
5648 Worker  
5399 Master  
15888 JobHistoryServer  
如果HDFS没有启动 ，请启动起来，参考hadoop集群搭建[点击打开链接](http://blog.csdn.net/stark_summer/article/details/42424279" \t "_blank)

查看slave节点：

[spark@S1PA222 scala]$ jps  
20352 Bootstrap  
30737 NodeManager  
7219 Jps  
30482 DataNode  
29500 Bootstrap  
757 Worker

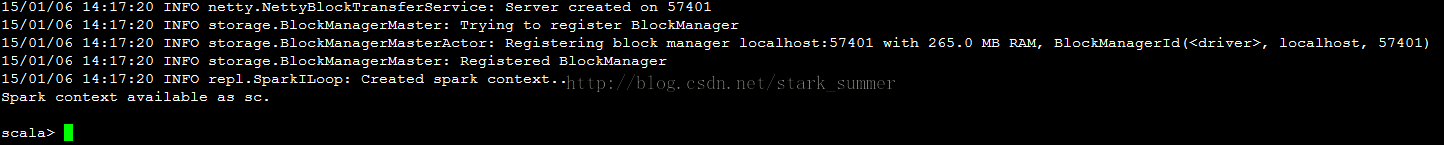
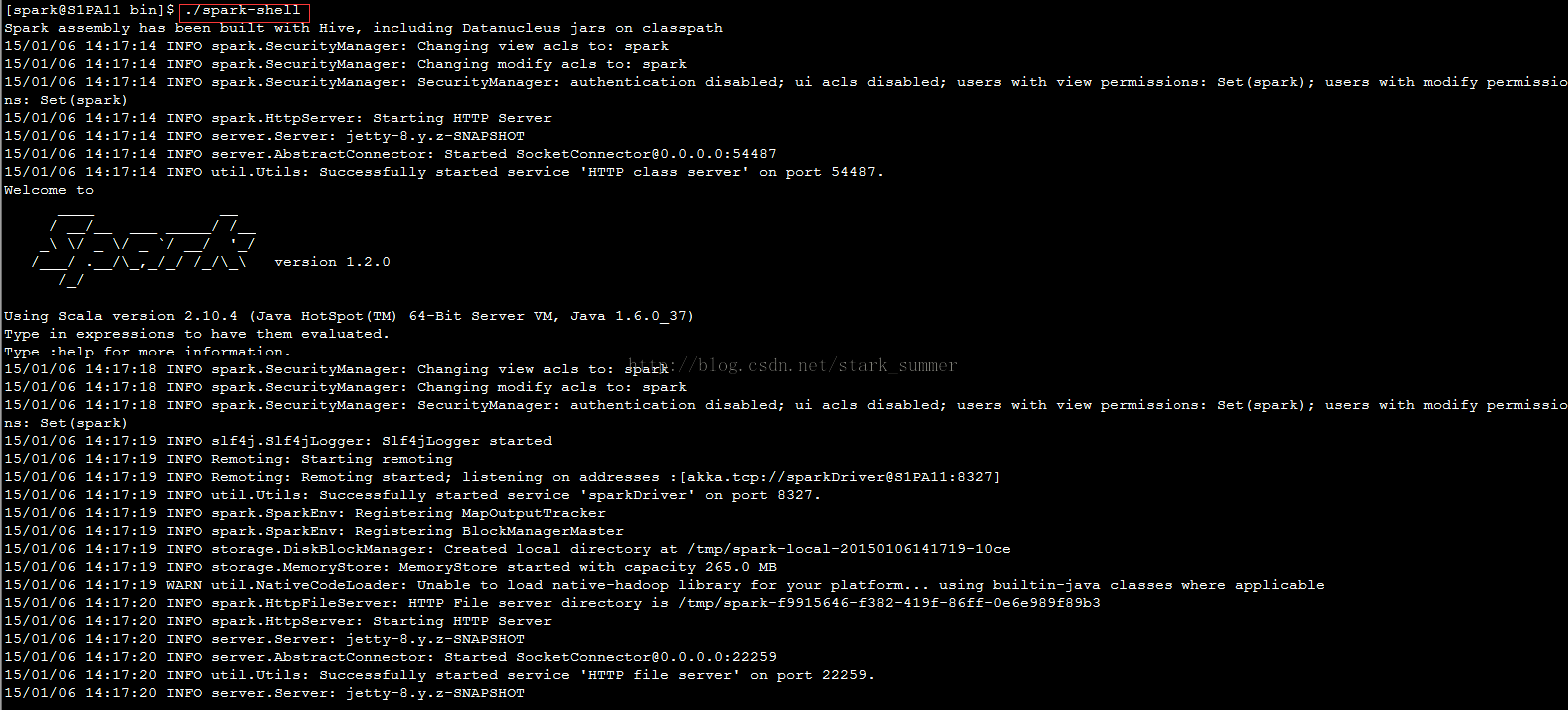
9、页面查看集群状况：

进去spark集群的web管理页面，访问



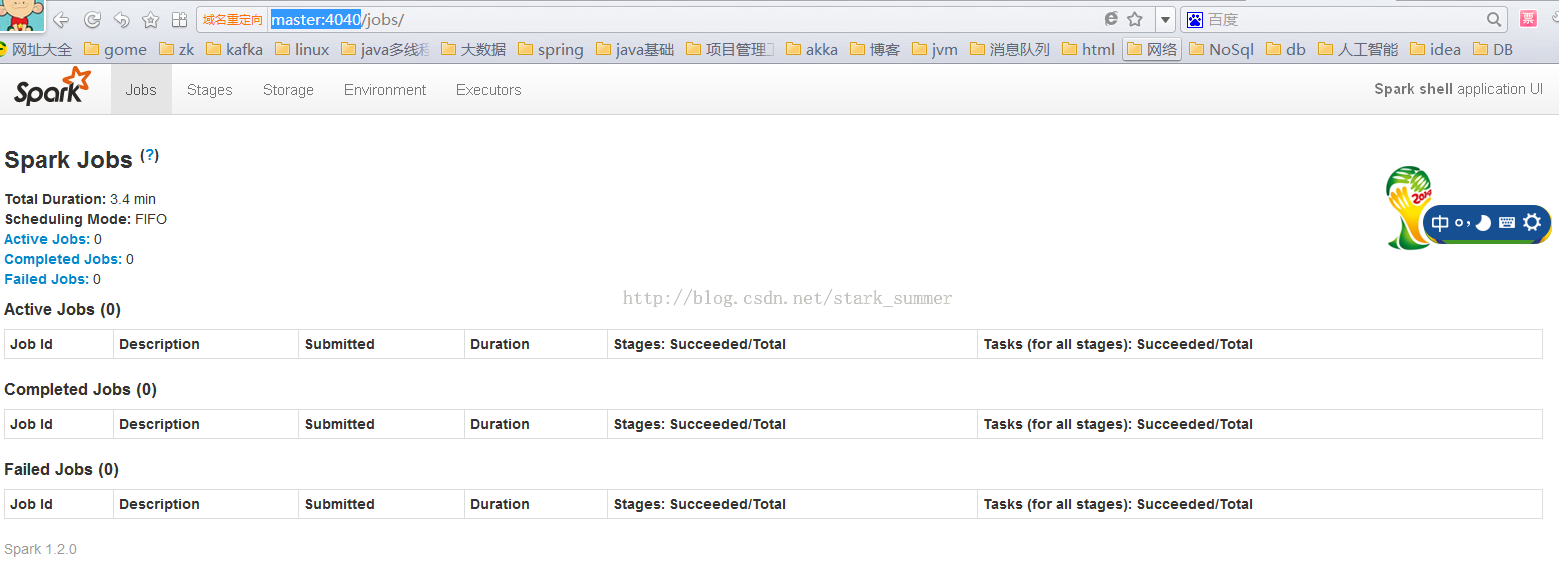
因为我们 看到两个worker节点，因为master和slave都是worker节点

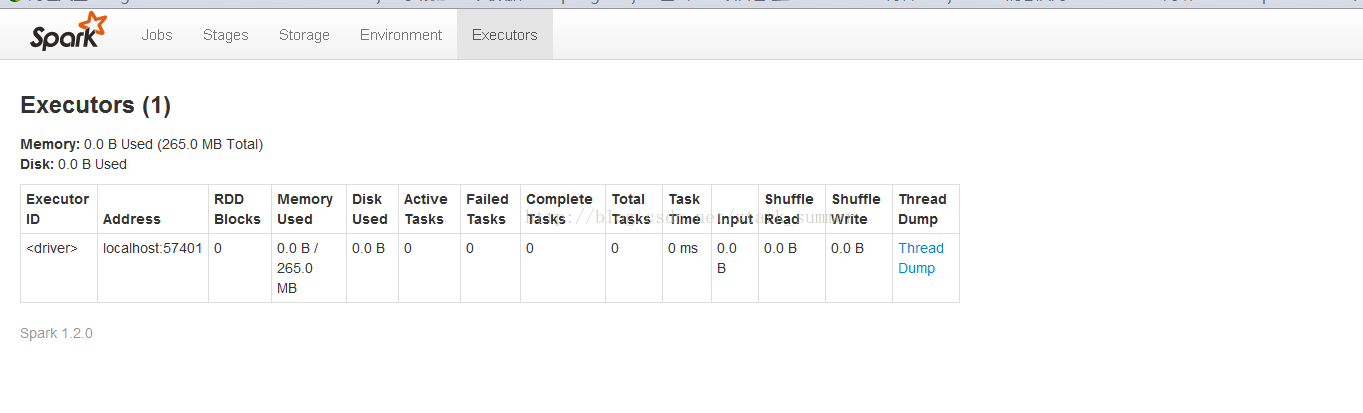
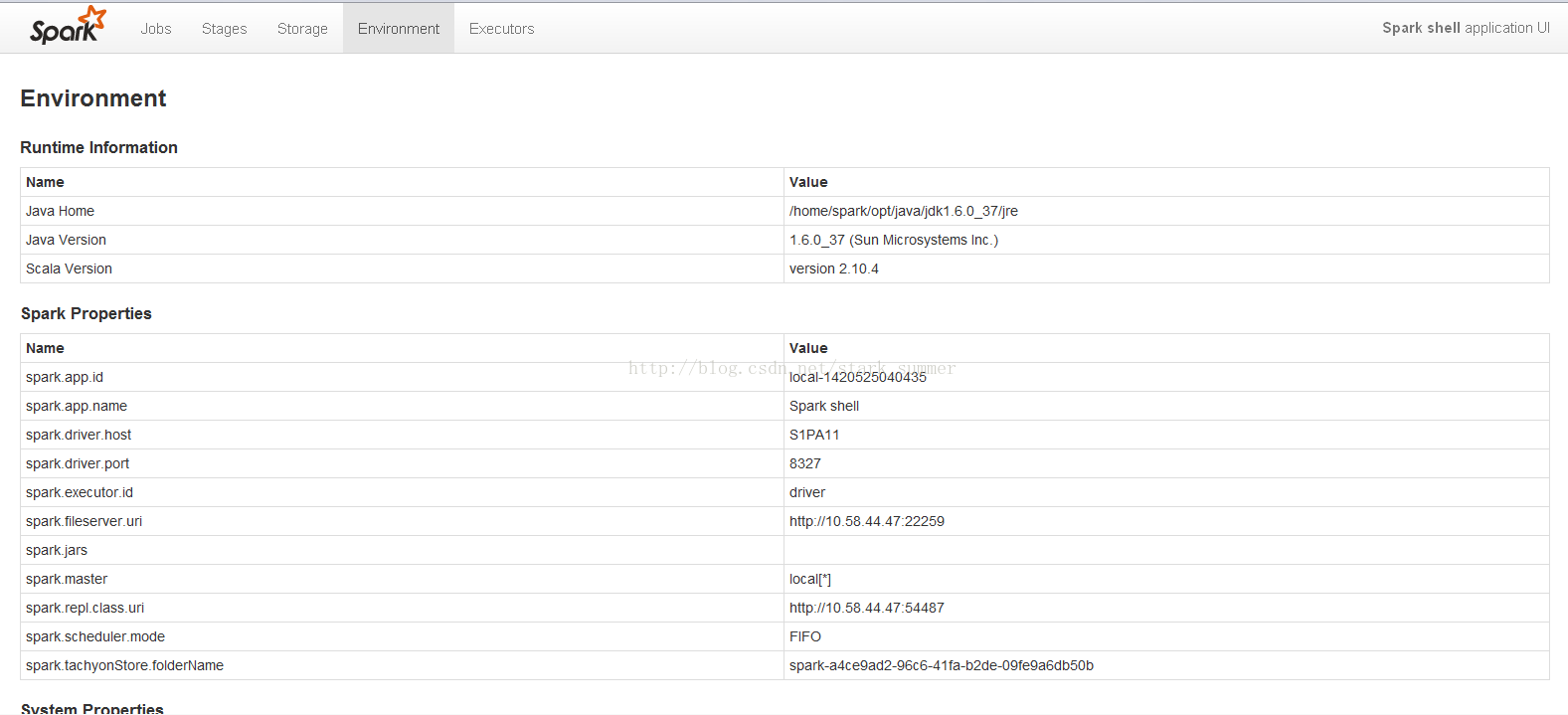
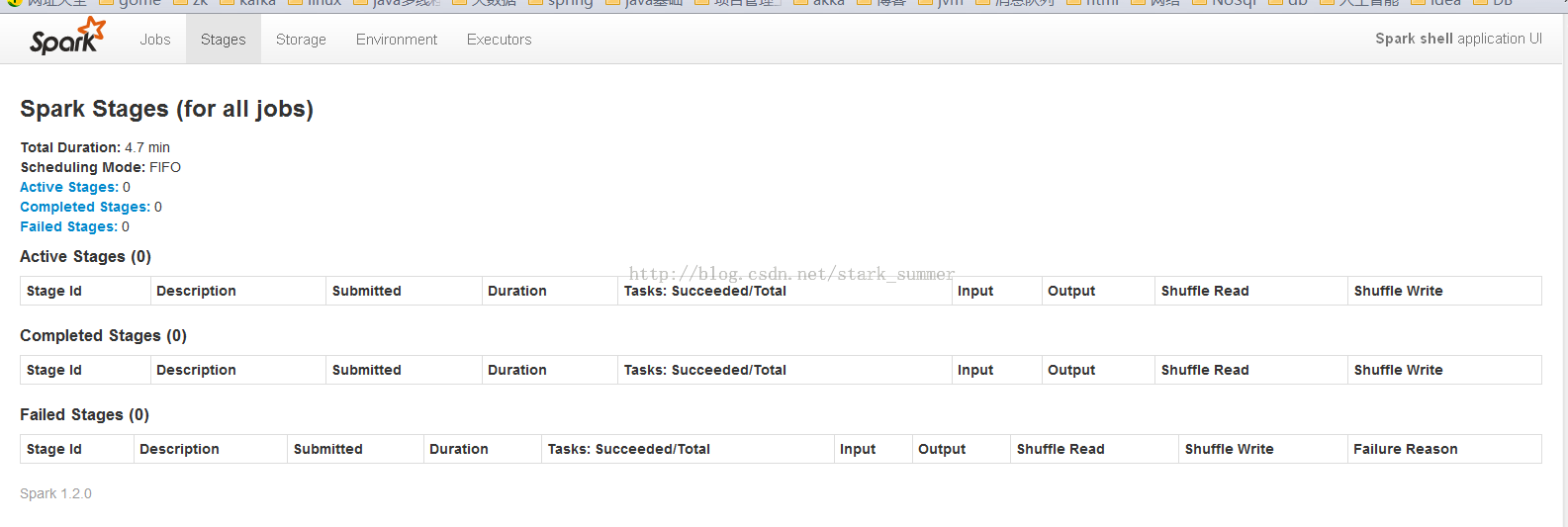
我们进入spark的bin目录，启动spark-shell控制台



现在我们已经顺利进入spark-shell的世界了 ，O(∩\_∩)O

访问http://master:4040/，我们可以看到spark WEBUI页面



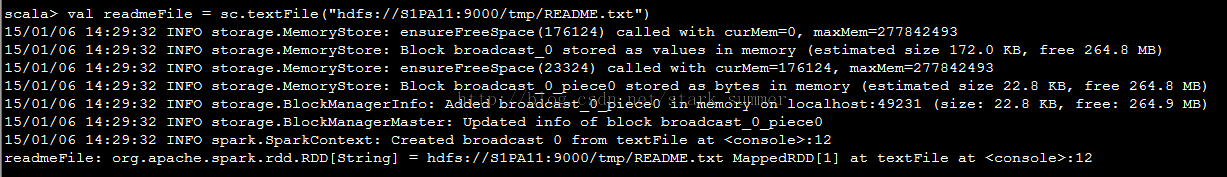


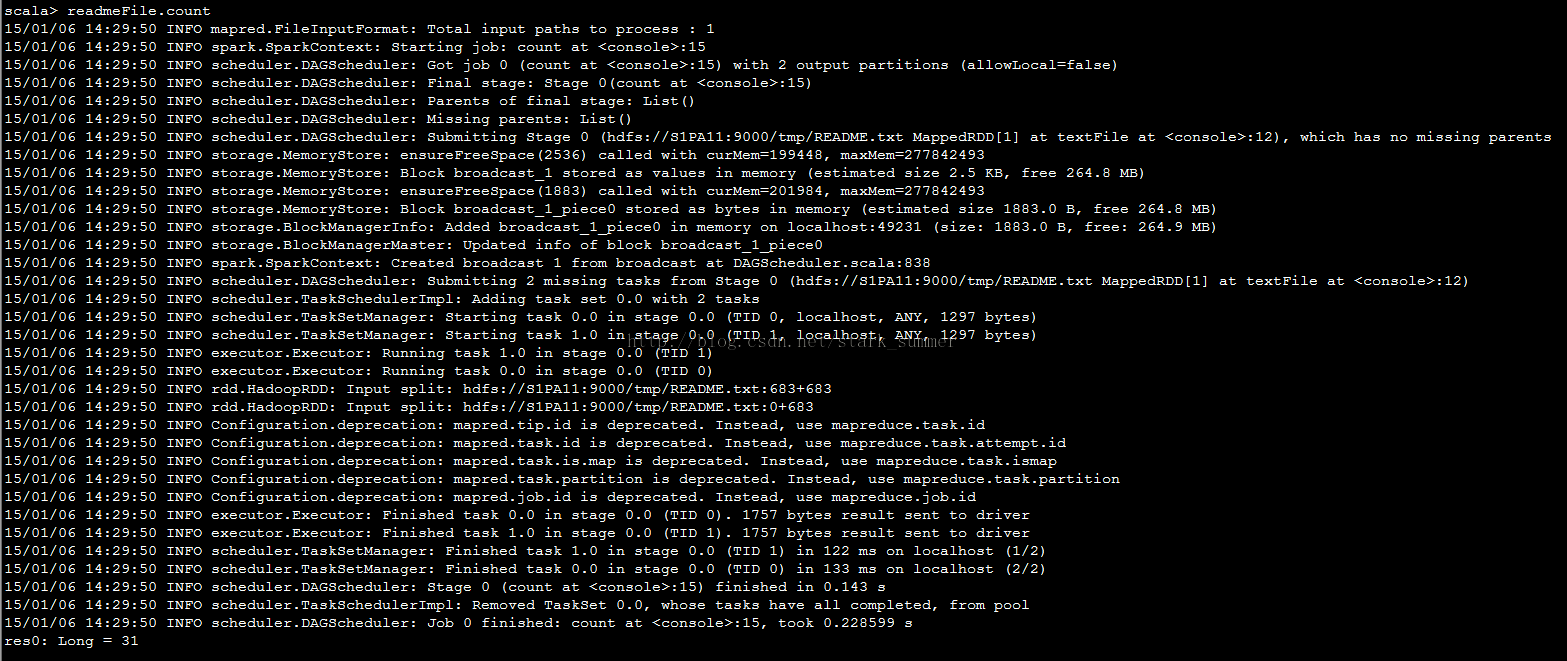
到目前为止，我们的spark集群环境搭建成功了

10、运行spark-shell 测试

之前我们在/tmp目录上传了一个README.txt文件，我们现在就用spark读取hdfs中README.txt文件

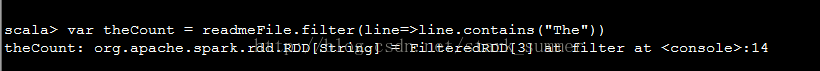


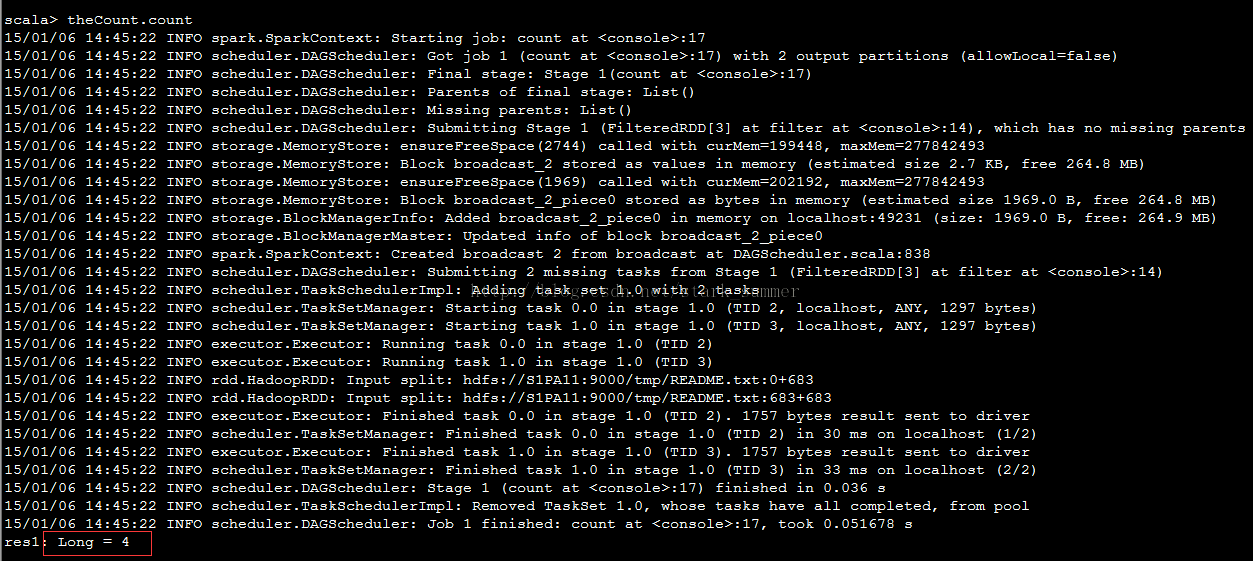
取得hdfs文件：

count下READM.txt文件中文字总数，

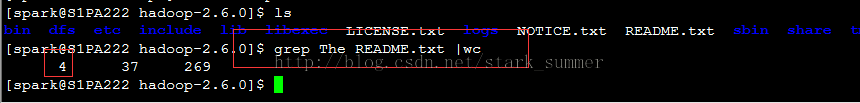
我们过滤README.txt

包括The单词有多个





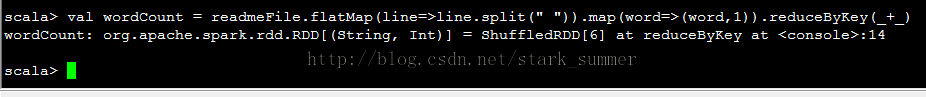
我们算出来 一共有4个The单词



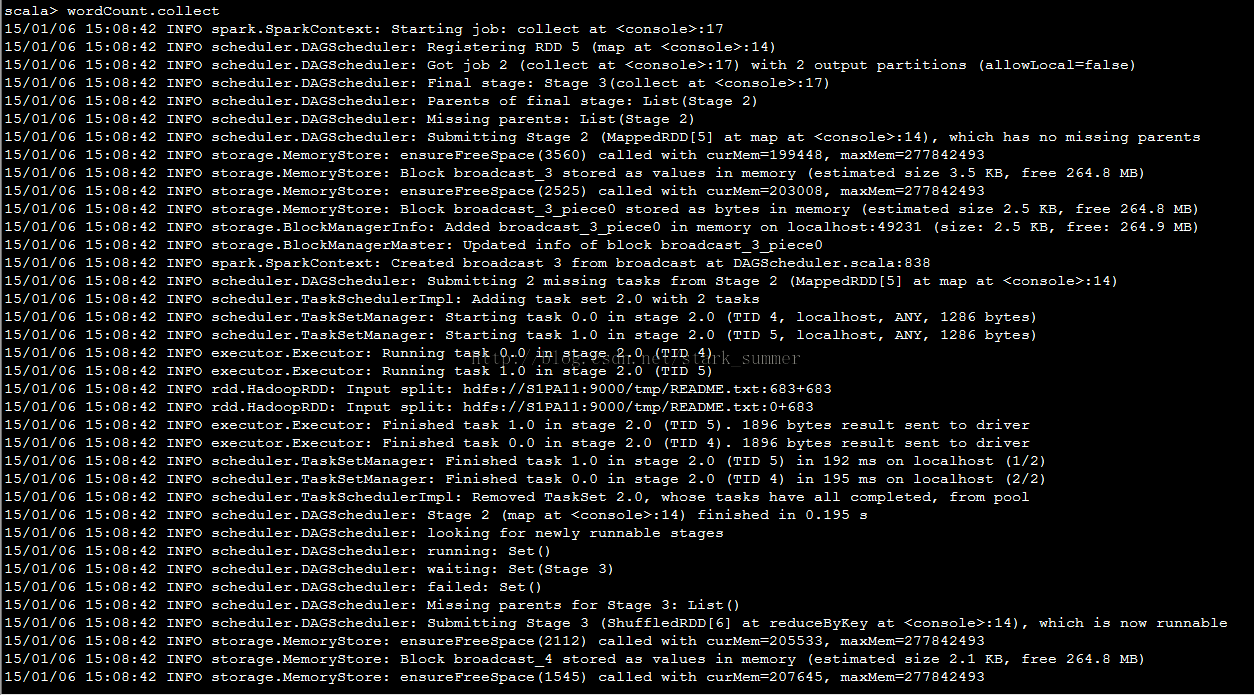
我们通过wc也算出来有4个The单词

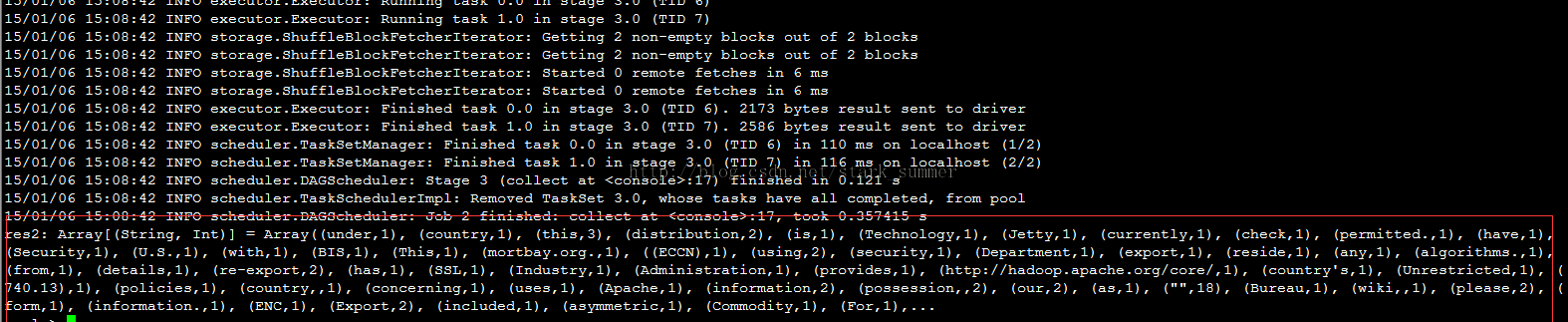
我们再实现下Hadoop wordcount功能：

首先对读取的readmeFile执行以下命令：

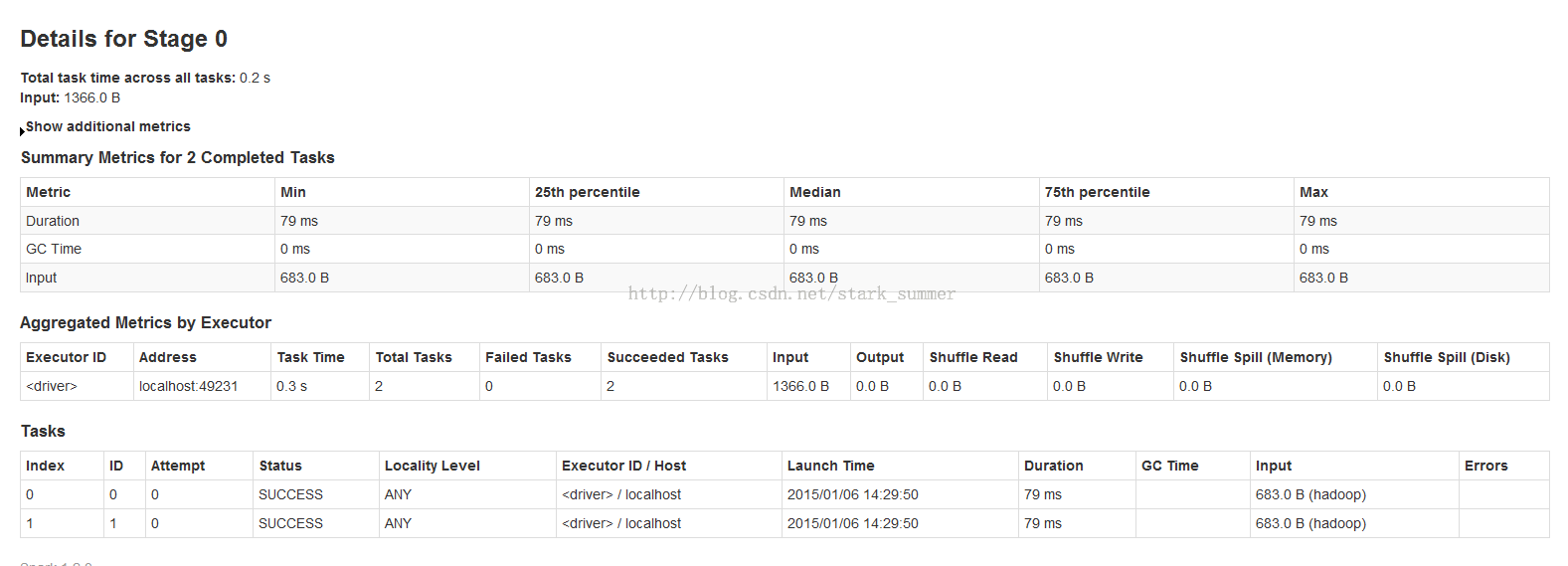


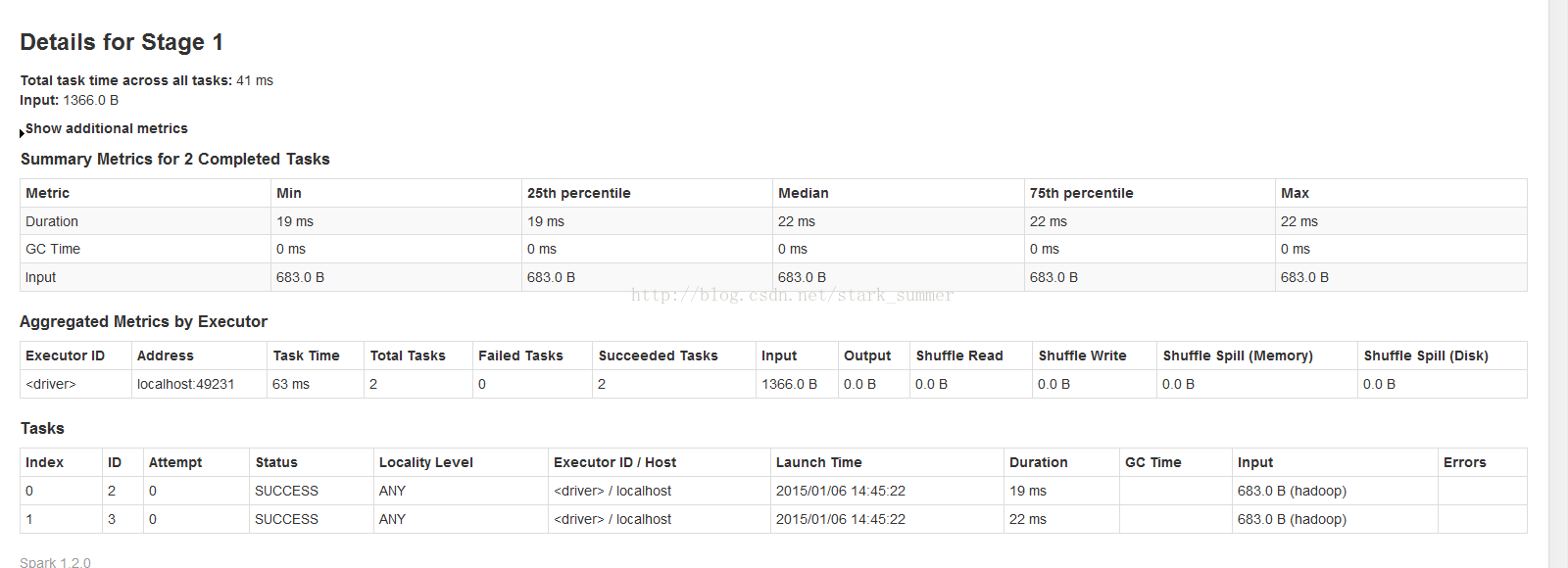
其次使用collect命令提交并执行job:

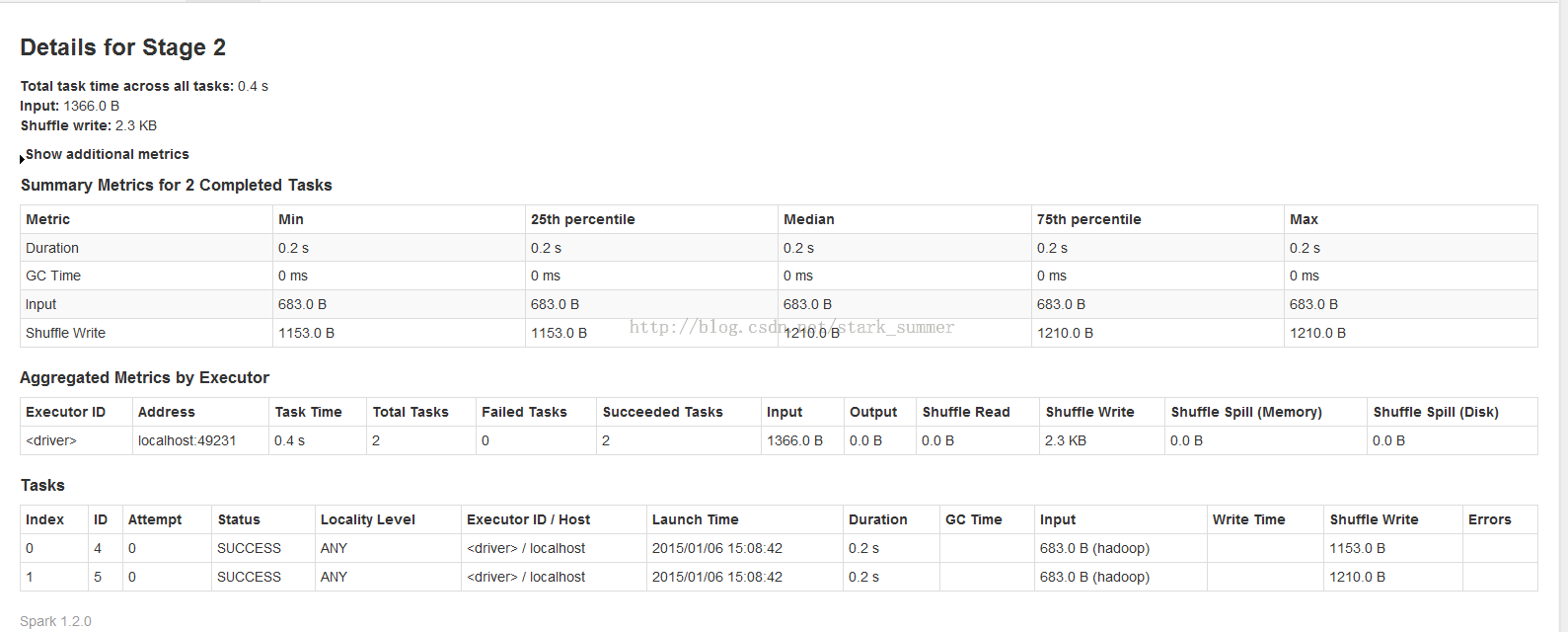


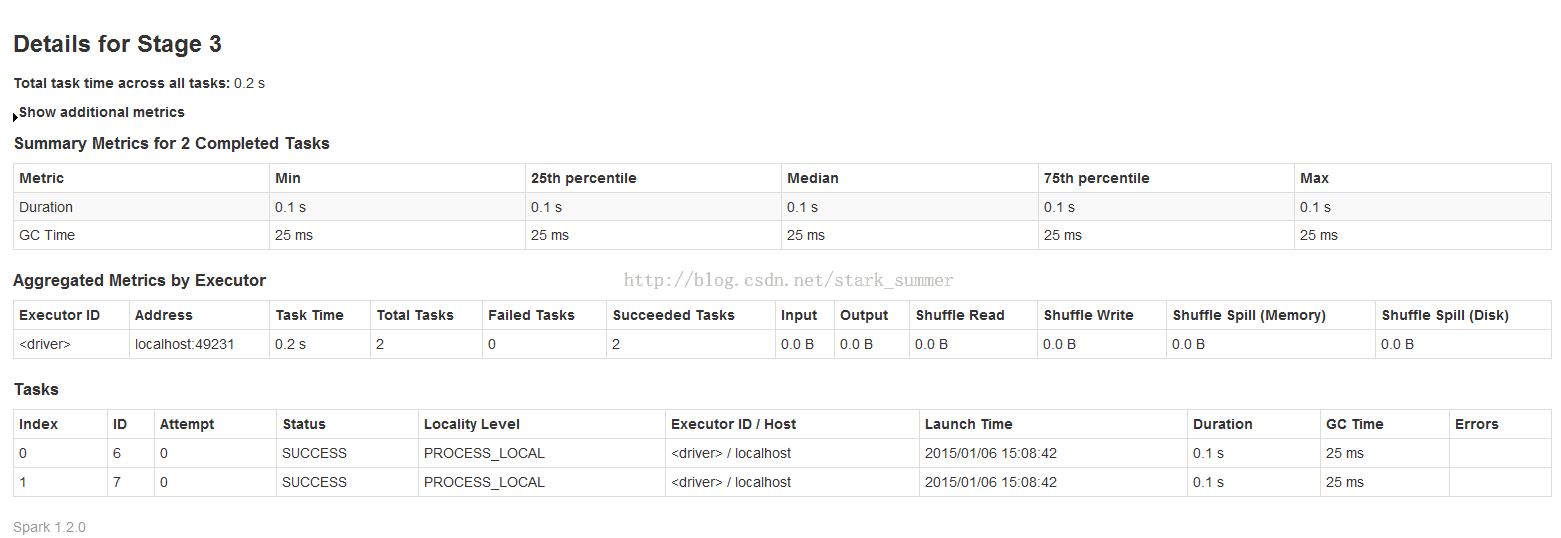


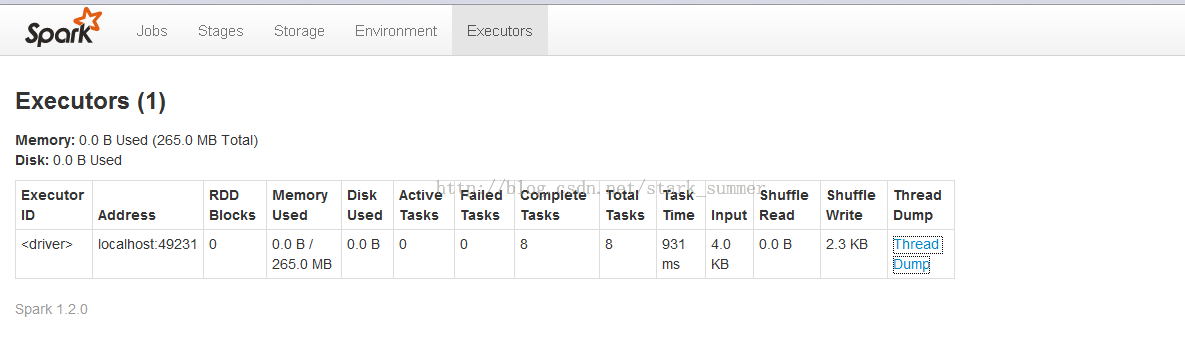
我们看下WEBUI界面执行效果：











OK，完成所有任务，(\*^\_\_^\*) spark-1.2.0 集群环境搭建