



个人资料



Hspark

+ 加关注

发私信

访问：996次

积分：283

等级：BLOG > 2

排名：千里之外

原创：1篇

转载：0篇

译文：0篇

评论：1条

文章搜索

Q

- 文章分类
- spark (1)

- 文章存档
- 2015年05月 (1)

- 阅读排行
- Hadoop-2.6.0+Zookeep (998)

- 评论排行
- Hadoop-2.6.0+Zookeep (1)

- 推荐文章
- \*HTML抽屉效果的实现与展示
- \*2015年校招求职之旅
- \*iOS9使用提示框的正确实现方式

💡

专家和你聊Web，速来报名

微信开发学习路线高级篇上线

免费公开课平台正式上线啦

恭喜July新书上市

原

Hadoop-2.6.0+Zookeeper-3.4.6+Spark-1.5.0+Hbase-1.1.2+Hive-1.2.0集群搭建

分类：spark

2015-05-12 22:09

🔍 998人阅读

💬 评论(1)

🔖 收藏

🚩 举报

- hive
- hadoop
- hbase
- zookeeper
- spark

前言

本博客目的在于跟大家分享大数据平台搭建过程，是笔者半年的结晶。在大数据搭建过程中,希望能给大家提过一些帮助，这也是本博客的价值所在。

博客内容分为将五个部分，具体如下：

- 第一部分 集群平台搭建
- 第二部分 SSH登录和JDK安装
- 第三部分 Hadoop-2.6.0+zookeeper-3.4.6集群搭建
- 第四部分 Spark集群搭建
- 第五部分 Hive集群搭建
- 第六部分 Hbase 集群搭建

以下是本集群所需要的软件，大家可按需要进行下载：

- Ubuntu14.04
- VMware Workstation 12 Pro for Linux
- Hadoop-2.6.0
- jdk-7u80-linux-x64.gz
- Zookeeper-3.4.6
- Spark-1.5.0
- Hbase-1.1.2
- Hive-1.2.0

第一部分 Ubuntu物理机和虚拟机搭建

集群框架，采用一台物理机和两台虚拟机。在PC机上用easyBCD方式安装Win和Ubuntu双系统。

一.安装easyBCD



\* 你不知道的JavaScript--Item18  
JScript的Bug与内存管理

Android Studio中配置及使用  
OpenCV示例

\* 最老程序员创业开发实训4--IOS  
平台下MVC架构

最新评论

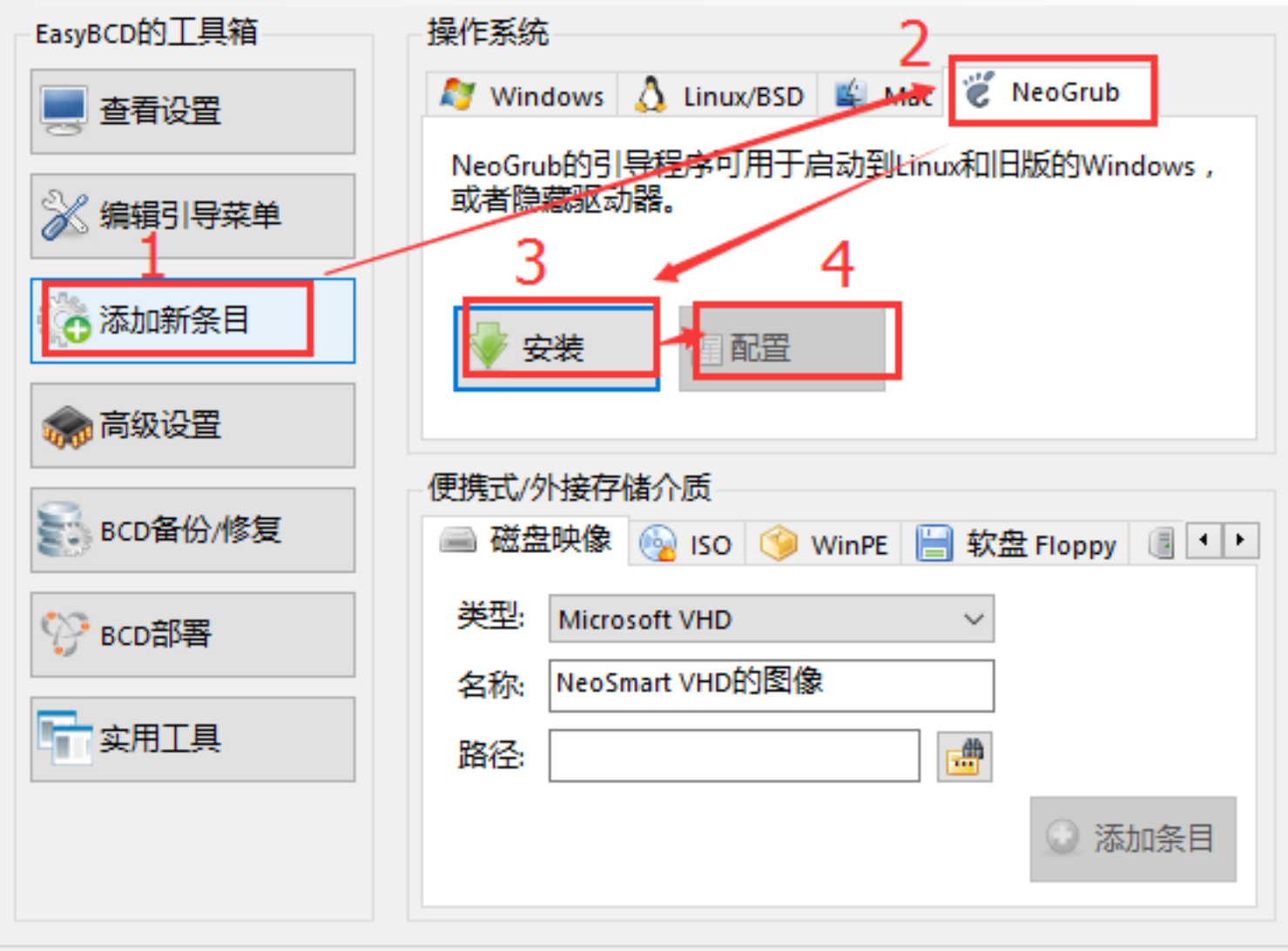
Hadoop-2.6.0+Zookeeper-3.4.6+  
caojunxu: export  
SPARK\_HOME=/spark/spark-  
1.3.1export PATH=PA...

1.下载

网址：<http://172.28.0.26/cache/4/02/crsky.com/f5eedf182f9ae67bdd575d0feaf80406/EasyBCD-v2.2.zip>

2.设置引导

安装好easyBCD，打开并点击添加新条目->安装->NeoGrub->配置，如图所示：



点击配置后，会跳出menu.lst文件，删除文件里面的所有内容，并复制以下内容到文件内：

```
1 title Install Ubuntu
2
3 root (hd0,1)
4
5 kernel (hd0,1)/vmlinuz.efi boot=casper iso-scan/filename=/ubuntu-14.04-desl
6
7 initrd (hd0,1)/initrd.lz
```

注：如果Win操作系统有系统保留分区，则root处为1，即root(hd0,1)；如果没有则为0,即root(hd0,0)。

### 二.安装Ubuntu

将buntu-14.04-desktop-amd64.iso文件拷贝到C盘。用解压软件打开，在casper文件下有两个文件：initrd.lz、vmlinuz.efi，将它们复制到C盘。

接下来重启PC,在引导界面选择NeoGrub，进入安装界面后，ctrl+alt+t打开终端命令，输入以下命令需要使用的命令取消光盘驱动器挂载：

```
1 sudo umount -l /isodevice
```

### 三.Ubuntu相关配置

1.打开软件中心更改软件更新源，一般选择国内的网易。

2.列举本地更新

```
1 sudo apt-get update
```

3.安装可用更新

```
1 sudo apt-get upgrade
```

4.修改root用户登录

```
1 sudo gedit /usr/share/lightdm/lightdm.conf.d/50-ubuntu.conf
```



添加

```
1  [SeatDefaults]
2  greeter-session=unity-greeter
3  user-session=ubuntu
4  greeter-show-manual-login=true
5  allow-guest=false
```

5.设置 root用户密码，,并输入root用户的密码

```
1  sudo passwd root
```

6.卸载防火墙

```
1  ufw disable    //关闭
2  sudo apt-get remove ufw    //卸载
3  sudo ufw status    //查看
```

7.安装ssh

安装ssh-server

```
1  sudo apt-get install openssh-server
```

安装ssh-client

```
1  sudo apt-get install openssh-client
```

```
1  验证：
2  ps -e | grep ssh
3
4  450 ??00:00:00 sshd    //如果出现这个,则安装成功
5
6  重启ssh：
7  /etc/init.d/ssh restart
```

四.安装Ubuntu虚拟机

1.下载VMware Workstation 12 Pro for Linux

网址：<http://www.vmware.com/cn/products/workstation/workstation-evaluation>

密钥：

```
1  5A02H-AU243-TZJ49-GTC7K-3C61N
```

2.安装VMware

使用以下命令安装VMware，”/home/huai/桌面/“为我在Ubuntu下文件路径名，大家可跟据自己的做更改。

```
1  sudo chmod +x /home/huai/桌面/VMware-Workstation-Full-12.0.0-2985596.x86_64
2
3  sudo /home/huai/桌面/VMware-Workstation-Full-12.0.0-2985596.x86_64.bundle
```

3.安装虚拟机

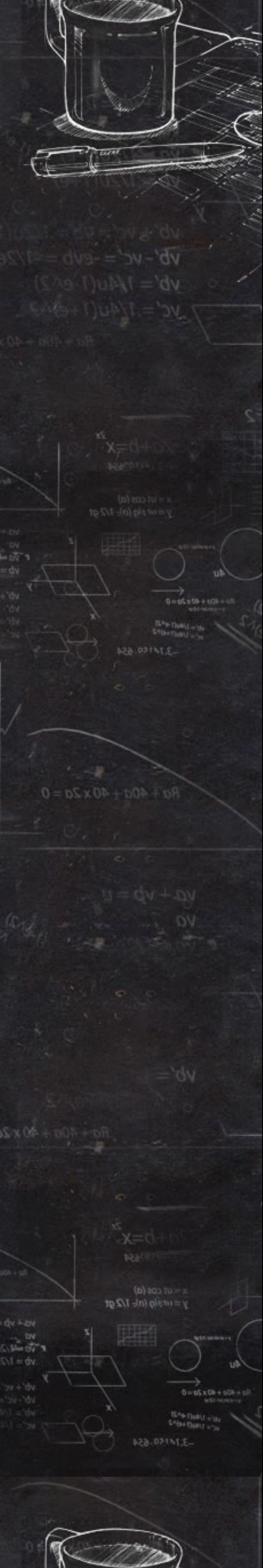
注：虚拟机用户名要与物理机的用户名保持一致，如笔者的三台集群的用户名都是“huai”。

用VMware安装虚拟机，采用典型方式安装，操作过程很简单，就不在这叙述了，大家可另行百度。对于刚安装好Ubuntu虚拟机，是可以联网的。按照上面Ubuntu物理机的相关配置对虚拟机Ubuntu进行同样操作。

配置好虚拟机,关闭虚拟机和VMware。复制安装的虚拟机文件，构成两台虚拟机。

4.修改主机名

因为笔者搭建此集群是为了学习spark,故将三台集群主机名命为spark01、spark02和spark03。其中，park01为物理



机，充当主节点。spark02和spark03为两台虚拟机的主机名。

修改物理机主机名,使用以下命令将主机名改为spark01:

```
1  sudo gedit /etc/hostname
```

修改虚拟机主机名：同上,分别修改为spark02和spark03。

五.修改IP

集群IP分别设为

```
1  192.168.8.101, 192.168.8.102, 192.168.8.103
```

1.设置虚拟机IP模式

```
虚拟机网卡设置为host-only模式( 具体操作，请百度 )
```

2.设置spark02的IP

通过界面设置IP

```
1  系统设置-->网络-->选项  -->IPv4设置-->  method选择为manual-->  点击add按钮--添加
2  IP: 192.168.8.102  子网掩码：255.255.255.0  网关：192.168.1.1-->  apply
```

3.设置spark03的IP

通过同样的方式把spark03的IP设置为：192.168.8.103

4.设置物理机spark01的IP

给物理机设置静态IP有点麻烦，因为在学校用的是内网，根据网上很多方法设置后，都无法再上网了。以下这种方式是网上比较常见的修改方式，但是不好使，只作参考，修改方式如下：

1) 一种令人不愉快的方式

```
1  打开：
2  sudo gedit /etc/network/interfaces
3
4  添加：
5  auto eth0
6  iface eth0 inet static
7  address 192.168.8.101
8  netmask 255.255.255.0
9  gateway 192.168.8.1
```

设置DNS域名解析：

```
1  打开：
2  sudo gedit /etc/resolv.conf
3
4  添加：
5  search nwu.edu.cn
6  nameserver 202.117.96.5
7  nameserver 202.117.96.10
8  nameserver 127.0.1.1
9
10  打开：
11  sudo gedit /etc/resolvconf/resolv.conf.d/tail
12
13  添加：
14  search nwu.edu.cn
15  nameserver 202.117.96.5
16  nameserver 202.117.96.10
17  nameserver 127.0.1.1
```



```
18
19 重启networking命令:
20
21 sudo /etc/init.d/networking restart
22 sudo /etc/init.d/resolvconf restart
```

笔者自己找到了两种方式设置一个固定的IP，简单且容易实现，关键在百度上都找不到。一种是采用桥接,第二种是绑定

2) 采用桥接

```
1 1.点击桌面右上角网络图标栏中的"编辑连接",网络连接中选中"桥接",点击"添加";
2 2.对跳转出来选择连接类型窗口,选择"虚拟--桥接",点击"新建",
3 3.在桥接--桥接连接(c)处,点击"添加",选择"以太网",点击新建,在设备MAC地址(D)处.选择"eth
4 4.点击IPv4设置,方法处选择为"手动",>点击"添加"设置IP为: 192.168.8.101 子网掩码: 255.
```

3) 采用绑定

```
1 采用绑定的方式与桥接方式操作一样，就是在操作2中，选择“虚拟--绑定”，其他操作一样。
```

六.修改主机名和IP的映射关系

整个集群的主机名和IP的映射关系的内容设置是一样的， 分别在三台机子上操作以下命令:

```
1 sudo gedit /etc/hosts
2
3 添加:
4
5 192.168.8.101 spark01
6 192.168.8.102 spark02
7 192.168.8.103 spark03
```

七.重启

```
1 关闭两台虚拟机，并重启Ubuntu物理机
```

第二部分 SSH登录和JDK安装

一.SSH登录配置

本集群的用户名为”huai”,集群尽量用统一的用户名,用不同的ssh登录时,还要进行一些其他操作，很麻烦。

1.在物理机spark01上.使用以下命令,四个回车，并cd到相关目录，操作如下:

```
1 huai@spark01:~$ ssh-keygen -t rsa
2 huai@spark01:~$ cd /home/huai/.ssh
3 huai@spark01:~/.ssh$ ls
4 id_rsa id_rsa.pub
```

生成两个文件：id\_rsa（私钥）、id\_rsa.pub（公钥）

2.对两台虚拟机进行同样操作(如上)

3.将虚拟机的id\_rsa.pub（公钥）传到spark01,”huai”是我集群的用户名,”huai”应改为你自己设定的用户名。分别在spark02和spark03进行以下操作:

```
1 scp id_rsa.pub huai@spark01:/home/huai/.ssh/id_rsa.pub.spark02
2 scp id_rsa.pub huai@spark01:/home/huai/.ssh/id_rsa.pub.spark03
```



4.在物理机上，按以下操作：

```
1 huai@spark01:~$ cd /home/huai/.ssh
2 huai@spark01:~/.ssh$ ls
3 id_rsa.pub          id_rsa.pub.spark03
4 id_rsa              id_rsa.pub.spark02
```

将公钥导入到综合authorized\_keys文件中， 在目录huai@spark01:~/.ssh\$下使用导入命令：

```
1 cat id_rsa.pub >> authorized_keys
2 cat id_rsa.pub.spark02 >> authorized_keys
3 cat id_rsa.pub.spark03 >> authorized_keys
```

5.将spark01上的authorized\_keys复制到spark02和spark03的/home/huai/.ssh/目录下

```
1 scp authorized_keys huai@spark02:/home/huai/.ssh/authorized_keys
2 scp authorized_keys huai@spark03:/home/huai/.ssh/authorized_keys
```

6.ssh常见错误修改

因为笔者的已经配置好,具体错误就没法再重现了。不过只要按以下操作之后,保证不会再有错误。

1) :

```
1 sudo gedit /etc/ssh/sshd_config
```

打开文件后， 注释”PermitRootLogin without-password”,使用”#”注释  
并在其下面添加: PermitRootLogin yes

2) :

```
1 打开：
2 sudo gedit /etc/ssh/ssh_config
3
4 添加：
5
6 StrictHostKeyChecking no
7 UserKnownHostsFile /dev/null
```

3) :

如果有Agent admitted failure to sign using the key错误抛出， 执行以下命令即可：

```
1 ssh-add
```

4) :

如果ssh登录时,在known\_hosts报出警告,后面显示一个数字,使用以下命令将其删除,比如显示数字”8”.

```
1 sed -i 8d /home/huai/.ssh/known_hosts
```

7.分别在每台上进行以下ssh登录,使用”exit”退出

```
1 ssh spark01
2 ssh spark02
3 ssh spark03
```

二.安装JDK

1.下载jdk， 6和8不做考虑， 8不支持， 6版本太低， 下载地址如下：

<http://download.oracle.com/otn-pub/java/jdk/7u80-b15/jdk-7u80-linux-x64.tar.gz>

2.安装



创建java文件夹:

```
1  sudo mkdir /usr/java
```

解压

```
1  sudo tar -zxvf /home/huai/桌面/jdk-7u80-linux-x64.gz -C /usr/java/
```

修改文件权限

```
1  sudo chmod -R 777 /usr/java/
```

3.修改环境变量

因为笔者集群已经搭建好了,为了让大家不要经常修改环境变量,笔者在此处把所有安装包的环境变量都列出来,供大家参考。

```
1  操作:
2  sudo gedit /etc/profile
3
4  在文件最后添加:
5  export JAVA_HOME=/usr/java/jdk1.7.0_80
6  export JRE_HOME=/usr/java/jdk1.7.0_80/jre
7  export SCALA_HOME=/spark/scala-2.11.6
8  export SPARK_HOME=/spark/spark-1.5.0
9  export HADOOP_HOME=/spark/hadoop-2.6.0
10 export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib:$HADOOP_HOME/lib/r
11 export HIVE_HOME=/spark/hive-1.2.0
12 export IDEA_HOME=/usr/idea/idea-IC-141.178.9
13 export HBASE_HOME=/spark/hbase-1.1.2
14 export ZOOKEEPER_HOME=/spark/zookeeper-3.4.6
15 export PATH=$PATH:$JAVA_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$SCALA_
```

刷新

```
1  source /etc/profile
```

4.验证

使用命令java -version，如果出现java version “1.7.0\_80”,则安装成功

5.设置jdk默认

如果Ubuntu上已经安装过openjdk纳闷需要设置默认jdk。若以上验证成功,可以跳过设置默认jdk。依次操作以下命令:

```
1  sudo update-alternatives --install /usr/bin/java java /usr/java/jdk1.7.0_80
2  sudo update-alternatives --install /usr/bin/javac javac /usr/java/jdk1.7.0_
3  sudo update-alternatives --install /usr/bin/jar jar /usr/java/jdk1.7.0_80/b
4  sudo update-alternatives --install /usr/bin/javah javah /usr/java/jdk1.7.0_
5  sudo update-alternatives --install /usr/bin/javap javap /usr/java/jdk1.7.0_
6  sudo update-alternatives --config java
```

第三部分 Hadoop-2.6.0+ookeeper-3.4.6集群搭建

一.集群规划

主机名	IP	安装软件	运行进程
spark01	192.168.8.101	jdk、hadoop、zookeeper	<b>namenode</b> 、ResourceManager (yarn) QuorumPeerMain、JournalNode、(zkfc)
spark02	192.168.8.102	jdk、hadoop、zookeeper	<b>namenode</b> 、ResourceManager DataNode、JournalNode、QuorumPeerMain
spark02	192.168.8.103	jdk、hadoop、zookeeper	DataNode、NodeManager、JournalNode QuorumPeerMain

文字描述：本集群仅有三台机器：spark01、spark02和spark03。三台都安装hadoop-2.6.0,zookeeper-3.4.6。spark01和spark02共运行了两个namenode，两个datanode运行spark02和spark03上。

二.zooekeeper配置

1.建立spark文件夹

```
1  sudo mkdir /spark
2  修改权限：
3  sudo chmod -R 777 /spark
```

2.解压

```
1  sudo tar -zxvf /home/huai/桌面/zookeeper-3.4.6.tar.gz -C /spark/
```

3.修改zoo.cfg文件

```
1  cd /spark/zookeeper-3.4.6/conf/
2  mv zoo_sample.cfg zoo.cfg
```

4.配置zoo.cfg

```
1  gedit zoo.cfg
2
3  修改：
4  dataDir=/spark/zookeeper-3.4.6/tmp
5  添加：
6  dataLogDir=/spark/zookeeper-3.4.6/logs
7  在最后添加：
8  server.1=spark01:2888:3888
9  server.2=spark02:2888:3888
10 server.3=spark03:2888:3888
11
12 保存退出
```

5.建立文件夹

```
1  mkdir /spark/zookeeper-3.4.6/tmp
2  mkdir /spark/zookeeper-3.4.6/logs
```

6.设置myid

在5) 中建立的tmp文件夹下创建一个空文件，并在其中写入”1“。

```
1  创建文件：
2  touch myid
3
4  写入"1":
5  gedit myid
```

7.修改环境变量

虽然前面已经把所有环境变量都设置好了,但这个步骤是环境配置必须的操作，所以在此处重复一遍

```
1  打开：
```



```
2 sudo gedit /etc/profile
3 添加：
4 export ZOOKEEPER_HOME=/spark/zookeeper-3.4.6
5 export PATH=:$ZOOKEEPER_HOME/bin:$ZOOKEEPER_HOME/conf
6 刷新：
7 source /etc/profile
```

8.将配置好的zookeeper拷贝到其他节点

```
1 scp -r /spark/zookeeper-3.4.6/ huai@spark02:/spark/
2 scp -r /spark/zookeeper-3.4.6/ huai@spark03:/spark/
```

9.修改虚拟机的myid

```
1 执行：
2 cd /spark/zookeeper-3.4.6/tmp
3 打开：
4 gedit myid
5 其中：
6 在spark02 中写入"2"
7 在spark03 中写入"3"
```

10.修改虚拟机的环境变量

按7中的操作分别修改虚拟机spark02和spark03的环境变量。

### 三.搭建hadoop集群

1.下载hadoop

网址：<http://mirror.bit.edu.cn/apache/hadoop/core/>

2.解压

```
1 sudo tar -zxvf /home/huai/桌面/hadoop-2.6.0.tar.gz -C /spark
2 修改权限：
3 sudo chmod -R 777 /spark
```

3.配置hadoop环境文件

注:在操作hadoop-env.sh,注释以下内容

```
1 #export HADOOP_OPTS="$HADOOP_OPTS -
2 Djava.net.preferIPv4Stack=true"
```

```
1 cd到配置文件目录下：
2
3 cd /spark/hadoop-2.6.0/etc/hadoop/
4
5 依次操作：
6
7 gedit hadoop-env.sh
8 gedit yarn-env.sh
9 gedit mapred-env.sh
10
11 依次添加：
12
13 export JAVA_HOME=/usr/java/jdk1.7.0_80
14 export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib:$HADOOP_HOME/lib/
15
16 依次保存退出
```

按以下操作依次修改hadoop其它配置文件，可以直接复制到配置文件。相关文件属性以及细节，可以查阅其他资料。

4.gedit core-site.xml

```
1 <configuration>
```



```
2    <property>
3      <name>fs.defaultFS</name>
4      <value>hdfs://ns1</value>
5    </property>
6    <property>
7      <name>hadoop.tmp.dir</name>
8      <value>/spark/hadoop-2.6.0/tmp</value>
9    </property>
10   <property>
11     <name>ha.zookeeper.quorum</name>    <value>spark01:2181,spark02:2181,s
12   </property>
13 </configuration>
```

##### 5.gedit hdfs-site.xml

```
1 <configuration>
2   <property>
3     <name>dfs.nameservices</name>
4     <value>ns1</value>
5   </property>
6   <property>
7     <name>dfs.ha.namenodes.ns1</name>
8     <value>nn1,nn2</value>
9   </property>
10  <property>
11    <name>dfs.namenode.rpc-address.ns1.nn1</name>
12    <value>spark01:9000</value>
13  </property>
14  <property>
15    <name>dfs.namenode.http-address.ns1.nn1</name>
16    <value>spark01:50070</value>
17  </property>
18  <property>
19    <name>dfs.namenode.rpc-address.ns1.nn2</name>
20    <value>spark02:9000</value>
21  </property>
22  <property>
23    <name>dfs.namenode.http-address.ns1.nn2</name>
24    <value>spark02:50070</value>
25  </property>
26  <property>
27    <name>dfs.namenode.shared.edits.dir</name>  <value>qjournal://spark01
28  </property>
29  <property>
30    <name>dfs.journalnode.edits.dir</name>
31    <value>/spark/hadoop-2.6.0/journal</value>
32  </property>
33  <property>
34    <name>dfs.ha.automatic-failover.enabled</name>
35    <value>true</value>
36  </property>
37  <property>
38    <name>dfs.client.failover.proxy.provider.ns1</name>
39  <value>org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyPro
40  </property>
41  <property>
42    <name>dfs.ha.fencing.methods</name>
43    <value>
44      sshfence
45      shell(/bin/true)
46    </value>
47  </property>
48  <property>
49    <name>dfs.ha.fencing.ssh.private-key-files</name>
50    <value>/home/huai/.ssh/id_rsa</value>
51  </property>
52  <property>
53    <name>dfs.ha.fencing.ssh.connect-timeout</name>
```



```
54     <value>30000</value>
55   </property>
56 </configuration>
```

6.mv mapred-site.xml.template mapred-site.xml

gedit mapred-site.xml

```
1 <configuration>
2   <property>
3     <name>mapreduce.framework.name</name>
4     <value>yarn</value>
5   </property>
6 </configuration>
```

7.gedit yarn-site.xml

```
1 <configuration>
2   <property>
3     <name>yarn.resourcemanager.ha.enabled</name>
4     <value>true</value>
5   </property>
6   <property>
7     <name>yarn.resourcemanager.cluster-id</name>
8     <value>yrc</value>
9   </property>
10  <property>
11    <name>yarn.resourcemanager.ha.rm-ids</name>
12    <value>rm1,rm2</value>
13  </property>
14  <property>
15    <name>yarn.resourcemanager.hostname.rm1</name>
16    <value>spark01</value>
17  </property>
18  <property>
19    <name>yarn.resourcemanager.hostname.rm2</name>
20    <value>spark02</value>
21  </property>
22  <property>
23    <name>yarn.resourcemanager.zk-address</name>
24    <value>spark01:2181,spark02:2181,spark03:2181</value>
25  </property>
26  <property>
27    <name>yarn.nodemanager.aux-services</name>
28    <value>mapreduce_shuffle</value>
29  </property>
30 </configuration>
```

8.gedit slaves

```
1 spark02
2 spark03
```

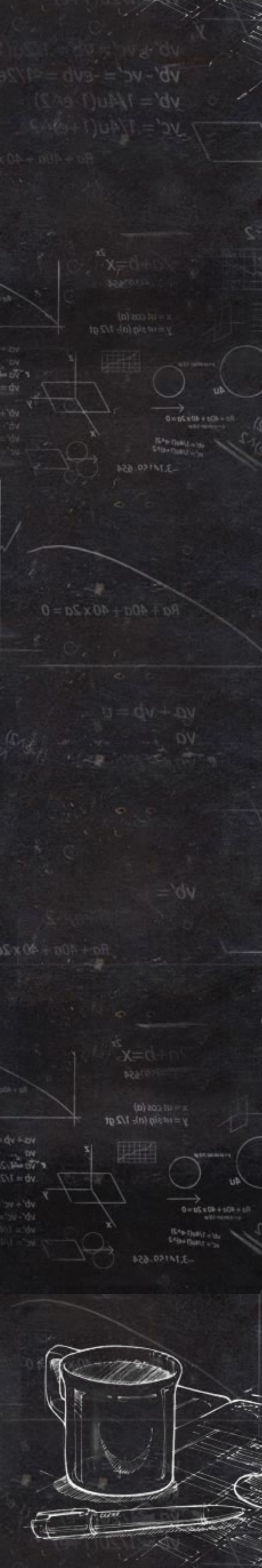
10.复制到其他节点

```
1 scp -r /spark/hadoop-2.6.0 huai@spark02:/spark
2 scp -r /spark/hadoop-2.6.0 huai@spark03:/spark
```

11.修改环境变量,可略过

```
1 打开:
2 sudo gedit /etc/profile
3 添加:
4 export HADOOP_HOME=/spark/hadoop-2.6.0
5 export PATH=$PATH:$JAVA_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
6 刷新:
7 source /etc/profile
```





## 12.验证

输入以下命令，如果出现hadoop对应的版本，则hadoop配置成功。

```
1  hadoop version
```

## 三.集群启动（严格按照下面的步骤）

### 1.启动zookeeper集群（分别在spark01、spark02、spark03上启动zk）

```
1  cd /spark/zookeeper-3.4.6/bin/
2  ./zkServer.sh start
3
4  查看状态：一个leader，两个follower
5  zkServer.sh status
```

### 2.启动journalnode（分别在spark01、spark02、spark03上启动journalnode）

```
1  在三台机子上执行：
2  cd /spark/hadoop-2.6.0/sbin
3  hadoop-daemon.sh start journalnode
```

运行jps命令检验，spark01、spark02、spark03上多了JournalNode进程

### 3.格式化HDFS

在主节点spark01上执行命令

```
1  hdfs namenode -format
```

格式化成功会生成tmp文件,其路径为core-site.xml中的hadoop.tmp.dir配置路径

### 4.将tmp拷到其他节点

注意tmp的路径：

```
1  scp -r tmp      huai@spark02:/spark/hadoop-2.6.0/
2  scp -r tmp      huai@spark03:/spark/hadoop-2.6.0/
```

### 5.格式化ZK

```
1  在spark01上执行：
2  hdfs zkfc -formatZK
```

### 6.启动HDFS

```
1  在spark01上执行：
2  start-dfs.sh
```

### 7.启动YARN.resourcemanager

```
1  start-yarn.sh
```

## 8.验证

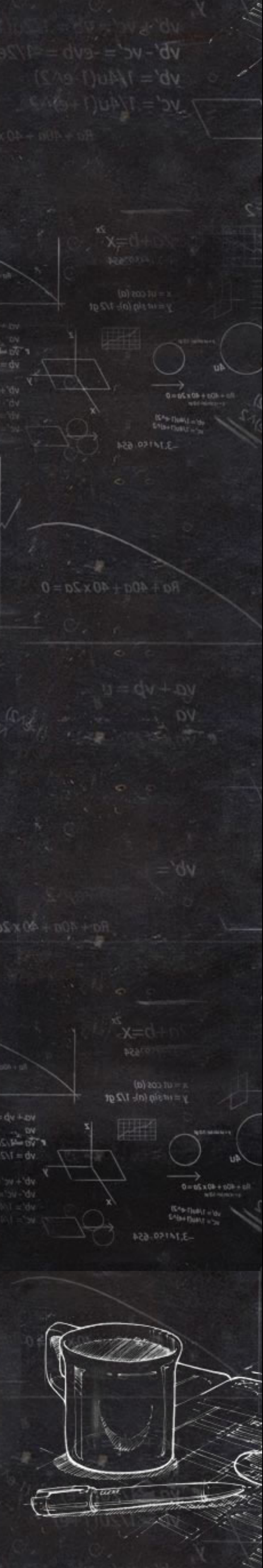
通过以下IP用浏览器访问，一个处于active,一个处于standby，说明集群启动成功。

```
1  http://192.168.8.101:50070
2  NameNode 'spark01:9000' (active)
3  http://192.168.8.102:50070
4  NameNode 'spark02:9000' (standby)
```

## 9.验证HDFS HA

```
1  向hdfs上传一个文件：
```





```
2  hadoop fs -put /etc/profile /profile
3  hadoop fs -ls /
```

```
1  kill掉active的NameNode
2  kill -9 <pid of NN>
3  访问: http://192.168.8.101:50070 无法打开
4  访问:http://192.168.8.102:50070
5  NameNode 'spark02:9000' (active)
```

```
1  执行:
2  hadoop fs -ls /
3  -rw-r--r--    3 root supergroup      1926 2015-05-024 15:36 /profile
```

```
1  手动启动挂掉的那个NameNode, 在spark01上执行
2  hadoop-daemon.sh start namenode
3  访问: http://192.168.8.101:50070
4  显示: NameNode 'spark01:9000' (standby)
5
6  删除上传文件:
7  hadoop fs -rm -r /profile
```

到这，hadoop集群相当于已经成功搭建，剩下的就是要搭建基于hadoop的其他组件，如spark,hive,hbase。相信通过自学搭建hadoop集群的人儿，一定受了不少折腾。作为不是技术大牛的人儿，唯有屡败屡战，才能祭奠内心那份对技术的热血澎湃。

## 第四部分 Spark集群搭建

因为spark 源代码是使用Scala语言，所以需要安装Scala环境。

### 一.安装 Scala

#### 1.下载

网址: <http://www.scala-lang.org/download/>

#### 2.解压

```
1  sudo tar -zxf  sudo tar -zxvf /home/huai/桌面/scala-2.11.6.tgz  -C /spark
```

#### 3.修改环境变量

```
1  打开:
2  sudo gedit /etc/profile
3  添加:
4  export SCALA_HOME=/spark/scala-2.11.6
5  export PATH=PATH:$SCALA_HOME/bin
6  刷新:
7  source /etc/profile
```

#### 4.显示版本

```
1  huai@spark01:~$ scala -version
2  Scala code runner version 2.11.6 -- Copyright 2002-2013, LAMP/EPFL
```

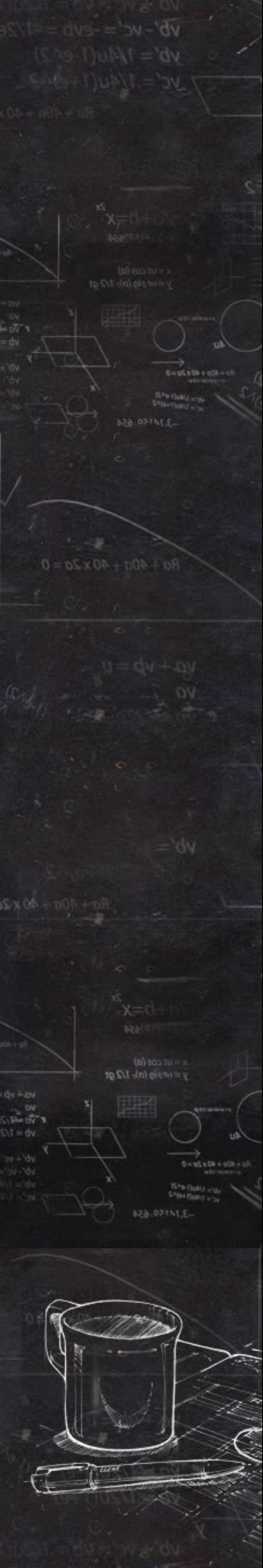
#### 1. 测试

输入”scala”命令，进入scala交互界面

```
1  huai@spark01:~$ scala
2  scala> 8*8
3  res0: Int = 64
```

#### 6.将scala复制到其他节点





```
1 scp -r /spark/scala-2.11.6/ huai@spark02:/spark/
2 scp -r /spark/scala-2.11.6/ huai@spark03:/spark/
```

7.重复 3修改其他节点的环境变量

## 二.spark集群搭建

### 1.下载

网址：<http://mirror.bit.edu.cn/apache/spark/>

### 2.解压

```
1 tar -zxvf /home/huai/桌面/spark-1.5.0-bin-hadoop2.6.tgz -C /spark
```

### 3.文件名修改为：spark-1.5.0

```
1 mv /spark/spark-1.5.0-bin-hadoop2.6 spark-1.5.0
```

### 4.修改环境变量

```
1 打开：
2 sudo gedit /etc/profile
3 添加：
4 export SPARK_HOME=/spark/spark-1.5.0
5 export PATH=PATH:$SPARK_HOME/bin
6 刷新：
7 source /etc/profile
```

### 5.配置spark-env.sh

```
1 进入spark的conf目录：
2 cd /spark/spark-1.5.0/conf/
3 mv spark-env.sh.template spark-env.sh
4
5 打开：
6 sudo gedit spark-env.sh
7
8 添加：
9
10 export JAVA_HOME=/usr/java/jdk1.7.0_80
11 export SCALA_HOME=/spark/scala-2.11.6
12 export SPARK_MASTER_IP=192.168.8.101
13 export SPARK_WORKER_MEMORY=1g
14 export HADOOP_CONF_DIR=/spark/hadoop-2.6.0/etc/hadoop
15 export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib:$HADOOP_HOME/lib/ri
```

### 6.配置slaves

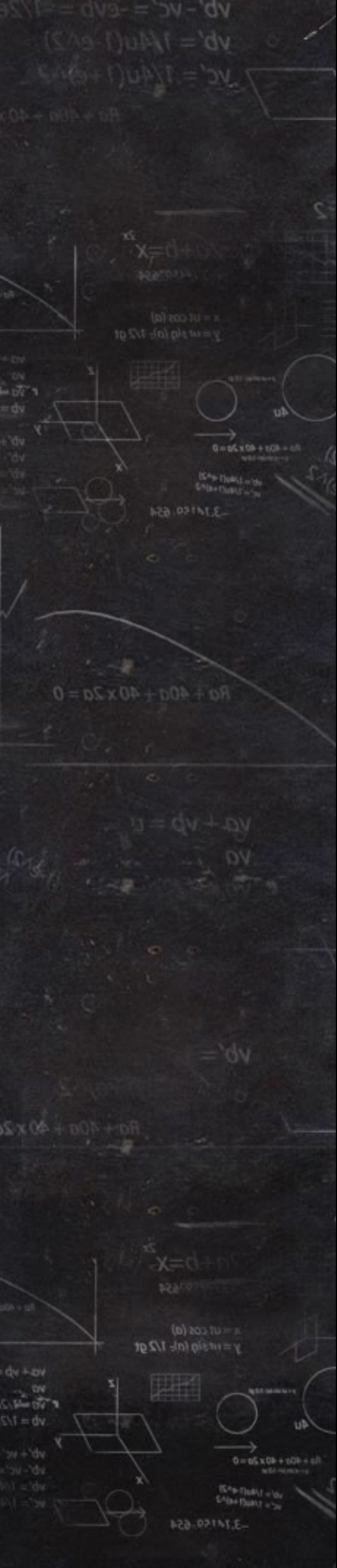
```
1 在spark的conf目录下：
2
3 mv slaves.template slaves
4
5 打开：
6 sudo gedit slaves
7 添加：
8 spark01
9 spark02
10 spark03
```

### 7.把spark复制到其他节点

```
1 scp -r spark-1.5.0 huai@spark02:/spark/
2 scp -r spark-1.5.0 huai@spark03:/spark/
```

### 8.修改虚拟阶环境变量





```
1 sudo gedit /etc/profile
2 添加：
3 export SPARK_HOME=/spark/spark-1.5.0
4 export PATH=PATH:$SPARK_HOME/bin
5 export SCALA_HOME=/spark/scala-2.11.6
6 export PATH=PATH:$SCALA_HOME/bin
7 刷新：
8 source /etc/profile
```

三.启动集群

1.启动zookeeper

```
1 在每台运行：
2 zkServer.sh start
```

2.启动hadoop集群，在hadoop的sbin目录下执行：

```
1 start-all.sh
```

3.启动spark集群

```
1 /spark/spark-1.5.0/bin$./start-all.sh
```

5.Web管理界面

```
1 http://192.168.8.101:8080/
```

Spark有好几个模块，如SparkSQL,Graphx,MLlib。因为Mapreduce处理速率不如Spark快， Spark将成为Mapreduce的替代品。对于Spark，我主要学习了机器学习MLlib和SparkSQL。如果你对大数据感兴趣，那么建议好好研究一下Spark。

第五部分 Hive集群搭建

一.概述

hive安装可分为四部分:

```
1 1)安装mysql
2 2)配置 Metastore
3 3)配置HiveClient
4 4)配置hiveserver
```

搭建分布:mysql,Metastore安装配置在主节点spark01上,HiveClient客户端安装在spark02上。注： HiveClient客户端也可以不配置。

三.安装mysql

mysql只需在物理机spark01上安装， Ubuntu软件中心下载安装mysql服务器

1.配置mysql

```
1 运行：
2 /usr/bin/mysql_secure_installation
3
4 输入密码：
5 删除匿名用户：
6 Remove anonymous users? [Y/n] Y
7 允许用户远程连接：
8 Disallow root login remotely? [Y/n]n
9 移除test数据库：
```





```
10 Remove test database and access to it? [Y/n] Y
```

## 2.登陆mysql

```
1 mysql -u root -p
```

## 3.建立hive数据库,并对其授权,注'huai'为你设立的登录密码

```
1 mysql>create database hive;
2 mysql>GRANT ALL PRIVILEGES ON *.* TO 'root'@'%' IDENTIFIED BY 'huai' WITH (
3 mysql>FLUSH PRIVILEGES;
```

## 4.修改hive数据库字符

一定要修改，不然hive建表时会报错。

```
1 mysql>alter database hive character set latin1
```

## 5.修改允许远程登录mysql

```
1 sudo gedit /etc/mysql/my.cnf
2 注释：
3 #bind-address=127.0.0.1
```

## 三.安装Hive

### 1.下载

网址：<http://mirrors.cnnic.cn/apache/hive/>

### 2.解压

```
1 sudo tar -zxvf /home/huai/桌面apache-hive-1.2.0.tar.gz -C /spark/
2
3 文件名修改为hive-1.2.0
4 mv apache-hive-1.2.0 hive-1.2.0
```

### 3.修改配置文件名

在hive-1.2.0/conf下

```
1 mv hive-default.xml.template hive-site.xml
2 mv hive-log4j.properties.template hive-log4j.properties
3 mv hive-exec-log4j.properties.template hive-exec-log4j.properties
4 mv hive-env.sh.template hive-env.sh
```

### 4.gedit hive-env.sh

```
1 export HADOOP_HOME=/spark/hadoop-2.6.0
2 export HIVE_CONF_DIR=/spark/hive-1.2.0/conf
```

### 5.gedit hive-log4j.properties

添加

```
1 hive.log.dir=/spark/hive-1.2.0/logs
```

### 6.新建logs文件

```
1 mkdir /spark/hive-1.2.0/logs
```

### 7.gedit hive-site.xml

删除所有内容,添加如下内容.注意配置文件中的'huai'为mysql登录密码,记得修改

```
1 <configuration>
2   <property>
3     <name>hive.metastore.warehouse.dir</name>
4     <value>hdfs://ns1/hive/warehouse</value>
```



```
5      </property>
6      <property>
7          <name>hive.exec.scratchdir</name>
8          <value>hdfs://ns1/hive/scratchdir</value>
9      </property>
10     <property>
11         <name>hive.querylog.location</name>
12         <value>/spark/hive-1.2.0/logs</value>
13     </property>
14     <property>
15         <name>javax.jdo.option.ConnectionURL</name>
16         <value>jdbc:mysql://localhost:3306/hive?createDatabaseIfNotExist=true</value>
17     </property>
18     <property>
19         <name>javax.jdo.option.ConnectionDriverName</name>
20         <value>com.mysql.jdbc.Driver</value>
21     </property>
22     <property>
23         <name>javax.jdo.option.ConnectionUserName</name>
24         <value>root</value>
25     </property>
26     <property>
27         <name>javax.jdo.option.ConnectionPassword</name>
28         <value>huai</value>
29     </property>
30     <property>
31         <name>hive.zookeeper.quorum</name>
32         <value>spark01,spark02,spark03</value>
33     </property>
34 </configuration>
```

8.配置环境变量

```
1  打开：
2  sudo gedit /etc/profile
3  添加：
4  export HIVE_HOME=/spark/hive-1.2.0
5  export PATH=$HIVE_HOME/bin
6  刷新：
7  source /etc/profile
```

9.在hive/lib下有个jline的jar，将hadoop内的这个jar包换成一致的，否则会启动hive会报错。

10.hive和mysql安装完成后，将mysql-connector-java-5.1.25-bin.jar连接jar拷贝到hive-1.2.0/lib目录下，在hive-1.2.0/bin下，运行：

```
1  ./hive
```

11.访问 <http://192.168.8.101:50070>

查看hadoop文件系统,多了个hive文件,证明安装成功。

12.修改hive文件权限

```
1  hadoop fs -chmod -R 777 /hive
```

四.配置hiveserver

gedit hive-site.xml

添加:

```
1  <property>
2      <name>hive.server2.thrift.min.worker.threads</name>
3      <value>1</value>
4      <description>Minimum number of Thrift worker threads</description>
5  </property>
6  <property>
```



```
7      <name>hive.server2.thrift.max.worker.threads</name>
8      <value>100</value>
9    </property>
10   <property>
11     <name>hive.server2.thrift.port</name>
12     <value>10000</value>
13   </property>
14   <property>
15     <name>hive.server2.thrift.bind.host</name>
16     <value>spark01</value>
17   </property>
```

五.Hive客户端配置

1.将hive从Master节点(spark01)拷贝到Slave节点(spark02)上

```
1  scp -r /spark/hive-1.2.0  huai@spark02:/spark/
```

2.配置Metastore的连接信息

在hive-1.2.0/conf下

```
1  gedit hive-site.xml
2  添加:
3  <property>
4    <name>hive.metastore.uris</name>
5    <value>thrift://spark01:9083</value>
6  </property>
```

3.修改环境变量

```
1  sudo gedit /etc/profile
2  export HIVE_HOME=/spark/hive-1.2.0
3  export PATH=:$HIVE_HOME/bin
4  source /etc/profile
```

第五部分 Hbase集群搭建

一.配置hbase

1.下载

网址：<http://mirrors.hust.edu.cn/apache/hbase/>

2.解压

```
1  tar -zxf /home/huai/桌面/hbase-1.1.2.tgz -C /spark
```

使用外部zk,配置hbase集群

3.gedit hbase-env.sh

```
1  export JAVA_HOME=/usr/java/jdk1.7.0_80
2  export HBASE_MANAGES_ZK=false
3  export HBASE_PID_DIR=/spark/hbase-1.1.2/pids
4  export HBASE_CLASSPATH=/spark/hadoop-2.6.0/etc/hadoop
5  export HBASE_HOME=/spark/hbase-1.1.2
```

3.gedit hbase-site.xml

```
1  <configuration>
2    <property>
3      <name>hbase.tmp.dir</name>
4      <value>/spark/hbase-1.1.2/tmp</value>
```

```
5      </property>
6      <property>
7          <name>hbase.rootdir</name>
8          <value>hdfs://ns1/hbase</value>
9      </property>
10     <property>
11         <name>hbase.cluster.distributed</name>
12         <value>true</value>
13     </property>
14     <property>
15         <name>zookeeper.session.timeout</name>
16         <value>120000</value>
17     </property>
18     <property>
19         <name>hbase.zookeeper.property.tickTime</name>
20         <value>6000</value>
21     </property>
22     <property>
23         <name>hbase.zookeeper.property.clientPort</name>
24         <value>2181</value>
25     </property>
26     <property>
27         <name>hbase.zookeeper.quorum</name>
28         <value>spark01,spark02,spark03</value>
29     </property>
30     <property>
31         <name>hbase.zookeeper.property.dataDir</name>
32         <value>/spark/zookeeper-3.4.6/tmp</value>
33     </property>
34     <property>
35         <name>dfs.replication</name>
36         <value>2</value>
37     </property>
38     <property>
39         <name>hbase.master.maxclockskew</name>
40         <value>180000</value>
41     </property>
42 </configuration>
```

4.gedit regionservers

```
1  spark02
2  spark03
```

5.把hadoop的hdfs-site.xml和core-site.xml 放到hbase/conf下

6.拷贝hbase到其他节点

```
1  scp -r /spark/hbase-1.1.0  huai@spark02:/spark/
2  scp -r /spark/hbase-1.1.0  huai@spark03:/spark/
```

7.修改环境变量

```
1  sudo gedit /etc/profile
2  export HBASE_HOME=/spark/hbase-1.1.0
3  export PATH=$HBASE_HOME/bin:$PATH
4  source /etc/profile
```

8.将配置好的hbase拷贝到每一个节点并同步时间和修改环境变量

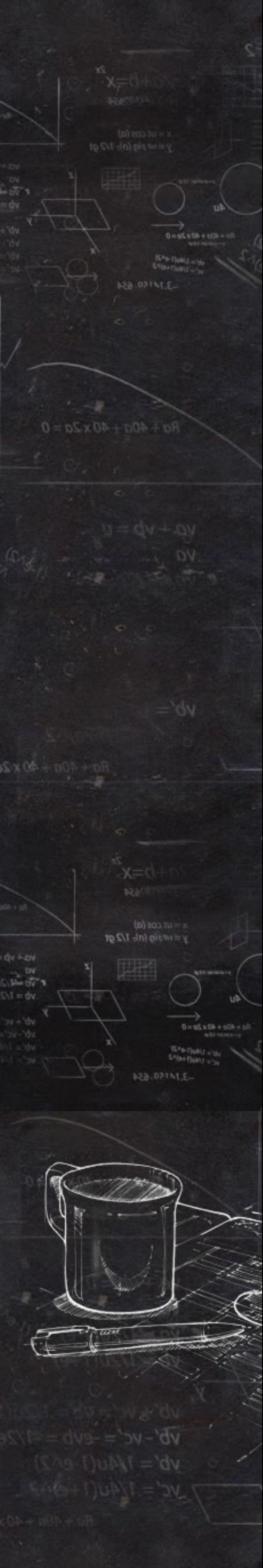
二.启动hbase

1.启动zk

```
1  zkServer.sh start
```

2.启动hadoop集群





```
1 start-dfs.sh
```

3.启动hbase，在主节点上hbase-1.1.0/bin运行

```
1 start-hbase.sh
```

3.通过浏览器访问hbase Web页面

HMaster Web页面

```
1 http://192.168.8.101:16010
```

HRegionServer Web页面

```
1 http://192.168.8.101:16030
```

4.shell 验证

进入hbase bin目录下:

```
1 huai@spark01:/spark/hbase-1.1.0/bin$ ./hbase shell
```

1) 使用list验证:

```
1 hbase(main):001:0> list
2      TABLE
3      user
4      1 row(s) in 1.4940 seconds
5      => ["user"]
6      hbase(main):002:0>
```

2) 建表验证

```
1 create 'user','name','sex'
```

如果list和建表都没出现错误的话,就意味着Hbase 安装成功

### 三.搭建错误排除

1.当Hbase shell时出现这样的错误

```
1 SLF4J: Found binding in [jar:file:/usr/local/hbase-1.1.0/lib/slf4j-log4j12.jar:1]
```

这是因为hadoop和hbase的slf4j-log4j12 jar包冲突,将HBase中的slf4j-log4j12-1.7.5移除

2.Hbase运行要在hadoop非安全模式下使用,使用以下命令查看hadoop安全模式和退出安全模式.

```
1 hdfs dfsadmin -safemode get //查看
2 hdfs dfsadmin -safemode leave //退出
```

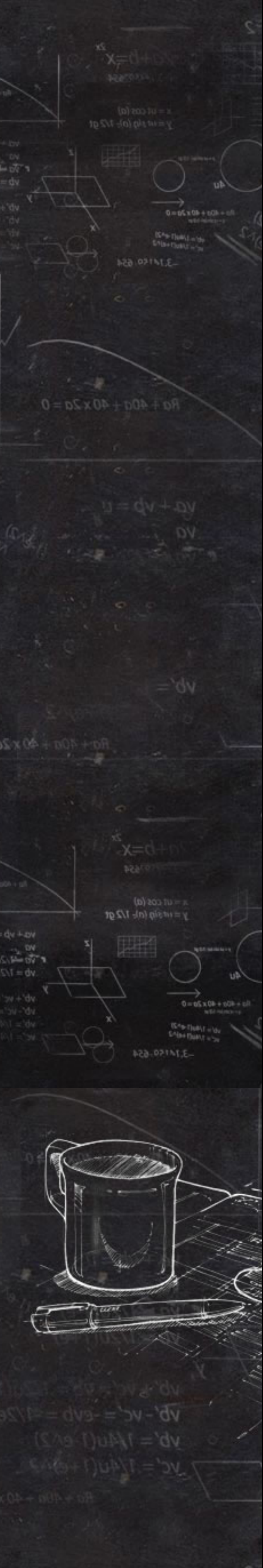
3.安装失败后,要删除hadoop文件系统下的Hbase文件,使用命令

```
1 hadoop fs -rm -r -f /hbase
```

4.这是最关键的。hosts文件中localhost的IP为127.0.0.1,如果有127.0.1.1的话,要将其改为127.0.0.1,以下是笔者hosts文件中的内容:

sudo gedit /etc/hosts

```
1 127.0.0.1 localhost
2
3 192.168.8.101 spark01
4 192.168.8.102 spark02
5 192.168.8.103 spark03
6
```



```
7  #The following lines are desirable for IPv6 capable hosts
8  ::1          ip6-localhost ip6-loopback
9  fe00::0      ip6-localnet
10 ff00::0       ip6-mcastprefix
11 ff02::1       ip6-allnodes
12 ff02::2       ip6-allrouters
```

注:spark01,spark02,spark03为集群主机名

5.修改hbase文件目录权限

```
1  hadoop fs -chmod -R 777 /hbase
```

6.如果你多次安装过不同版本的hbase且使用外部zookeeper,记得清理zookeeper中Hbase中文件,笔者就是因为这个问题,苦苦折磨了一周.

关闭hbase,在zookeeper的bin目录运行zkCli.sh

终端最后显示:

```
1  WATCHER::
2  WatchedEvent state:SyncConnected type:None path:null
3  [zk: localhost:2181(CONNECTED) 0]
```

按以下操作删除Hbase

```
1  [zk: localhost:2181(CONNECTED) 1] ls /
2  [hbase, hadoop-ha, yarn-leader-election, zookeeper]
```

如果有hbase则进行下一步

```
1  [zk: localhost:2181(CONNECTED) 2] rmr /Hbase
```

重启hbase

```
1  start-hbase.sh
```

7.单独启动HMaster，Hregionserve命令

```
1  hbase-daemon.sh  start regionserver
2  hbase-daemon.sh  stop master
3  hbase-daemon.sh  start regionserver
4  hbase-daemon.sh  stop regionserver
```

四.hbase安装总结

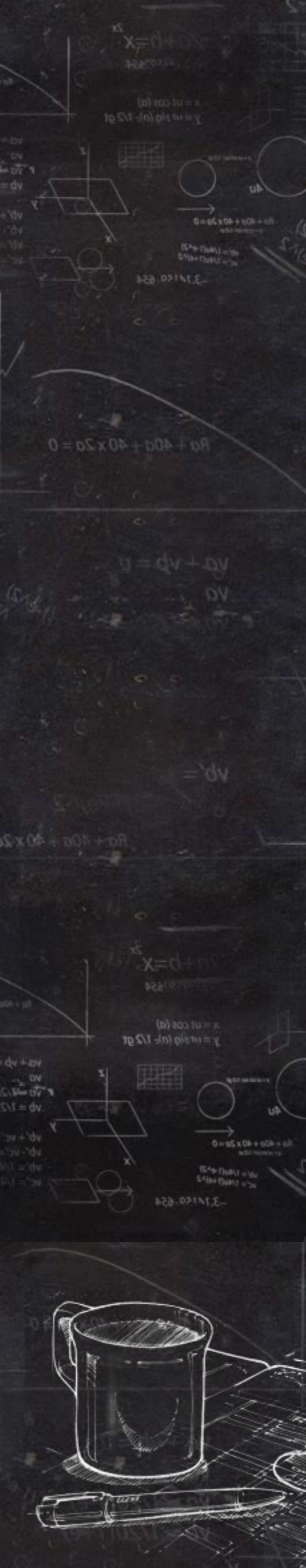
- 1.Hbase与hadoop有版本兼容要求,一般的解决方式都是把Hbase中与hadoop相关的jar包,替换成hadoop版本的jar包。
- 2.集群时间记得要同步,同步方式界面操作调整时区和格式。
- 3.安装会因为环境不同,而出现不同的错误,但整个安装过程应该有这样的心态:屡败屡战,经折腾。

作者后言:

满纸荒唐言，一把辛酸泪。都云作者痴，谁解其中味。第一篇博客到此结束，如果你看到，集群也配到这，证明你配置成功了，恭喜你了。由于是第一次写博客，加上学习时间也比较紧，文章或多或少有些瑕疵，还望见谅。







## 主题推荐

- ubuntu
- vmware
- spark
- hbase
- jdk
- pre
- linux
- hive
- ssh
- pc

## 猜你在找

- 深入浅出Spark

hbase入门精讲

大数据之编程语言：Scala

Hadoop生态系统零基础入门

Hadoop 2.X大数据平台基础

深入浅出Spark

hbase入门精讲

大数据之编程语言：Scala

Hadoop生态系统零基础入门

Hadoop 2.X大数据平台基础
- 运行mapreduce程序时出现异常

docker on spark

linux下面怎样让给一个用户添加对指定文件夹写的

hadoop学习五Hadoop220完全分布式安装详解1

记录一下centos的spark131版本开发环境

## 查看评论

1楼 [caojunxu](#) 2015-09-13 21:14发表



[plain]

```
01. export SPARK_HOME=/spark/spark-1.3.1
02. export PATH=PATH:$SPARK_HOME/bin
```

这里应该是export PATH=\$PATH:\$SPARK\_HOME/bin  
否则会有出事哦，  
谢谢分享，非常实用

您还没有登录,请[\[登录\]](#)或[\[注册\]](#)

\* 以上用户言论只代表其个人观点，不代表CSDN网站的观点或立场

## 核心技术类目

全部主题	Hadoop	AWS	移动游戏	Java	Android	iOS	Swift	智能硬件	Docker	OpenStack		
VPN	Spark	ERP	IE10	Eclipse	CRM	JavaScript	数据库	Ubuntu	NFC	WAP	jQuery	
BI	HTML5	Spring	Apache	.NET	API	HTML	SDK	IIS	Fedora	XML	LBS	Unity
Splashtop	UML	components	Windows Mobile	Rails	QEMU	KDE	Cassandra	CloudStack				
FTC	coremail	OPhone	CouchBase	云计算	iOS6	Rackspace	Web App	SpringSide	Maemo			
Compuware	大数据	apttech	Perl	Tornado	Ruby	Hibernate	ThinkPHP	HBase	Pure	Solr		
Angular	Cloud Foundry	Redis	Scala	Django	Bootstrap							

[公司简介](#) | [招贤纳士](#) | [广告服务](#) | [银行汇款帐号](#) | [联系方式](#) | [版权声明](#) | [法律顾问](#) | [问题报告](#) | [合作伙伴](#) | [论坛反馈](#)

网站客服 杂志客服 微博客服 webmaster@csdn.net 400-600-2320 | 北京创新乐知信息技术有限公司 版权所有 | 江苏乐知网络技术有限公司 提供商务支持

京 ICP 证 070598 号 | Copyright © 1999-2014, CSDN.NET, All Rights Reserved

[公司简介](#) | [招贤纳士](#) | [广告服务](#) | [银行汇款帐号](#) | [联系方式](#) | [版权声明](#) | [法律顾问](#) | [问题报告](#) | [合作伙伴](#) | [论坛反馈](#)

