# 搭建大数据平台

# 一.集群搭建准备

集群搭建默认搭建N台机器，现以3台机器为例。

## 1.1准备机器

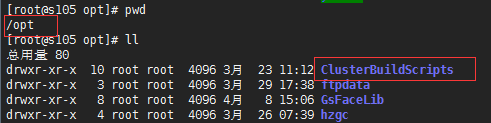
|  |  |
| --- | --- |
| 主机名 | IP |
| s105 | 172.18.18.105 |
| s106 | 172.18.18.106 |
| s107 | 172.18.18.107 |

## 1.2下载集群脚本及安装包

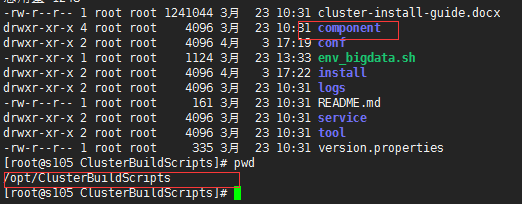
1、进入到其中一台机器例如s105的/opt目录。

cd /opt

2、把集群脚本文件夹ClusterBuildScripts同步到/opt下。



3、再将组件安装包component文件夹同步到/opt/ClusterBuildScripts目录下。



4、修改ClusterBuildScripts的执行权限。

chmod –R 755 /opt/ClusterBuildScripts

# 二.配置集群辅助项

## 2.1配置主机名和主机ip映射

修改主机s105的/etc/hosts文件（ip和主机名根据实际情况填写）

注意：系统自带的hosts配置不要删除！！

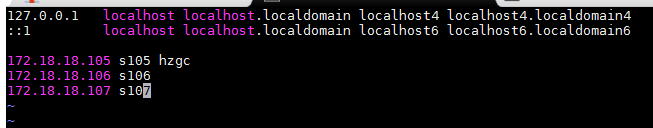
vim /etc/hosts

在文件结尾添加以下内容，并:wq保存退出(hzgc可配在任意节点下)

172.18.18.115 s105 hzgc

172.18.18.116 s106

172.18.18.117 s107

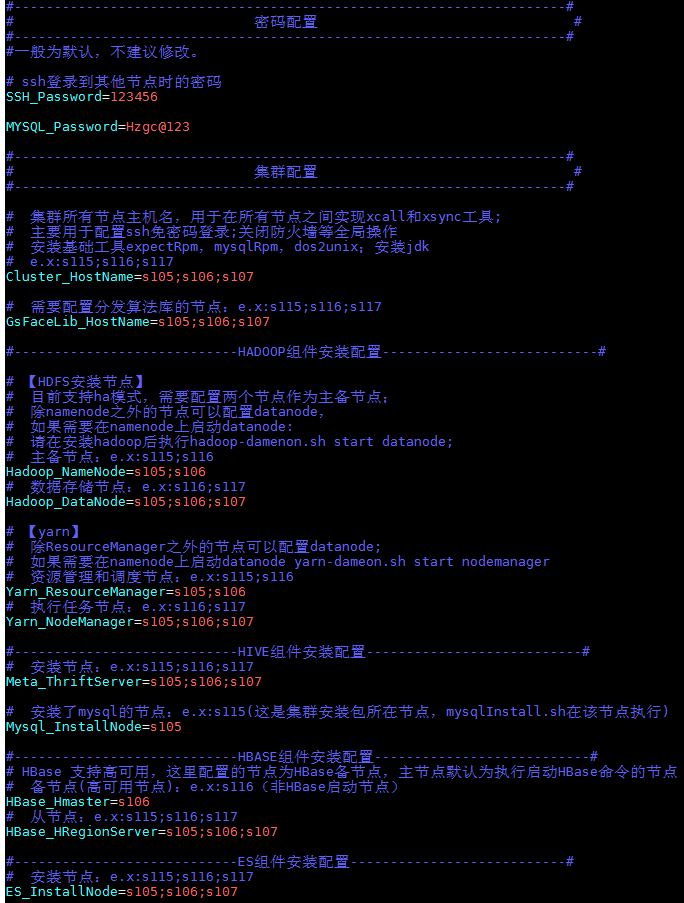


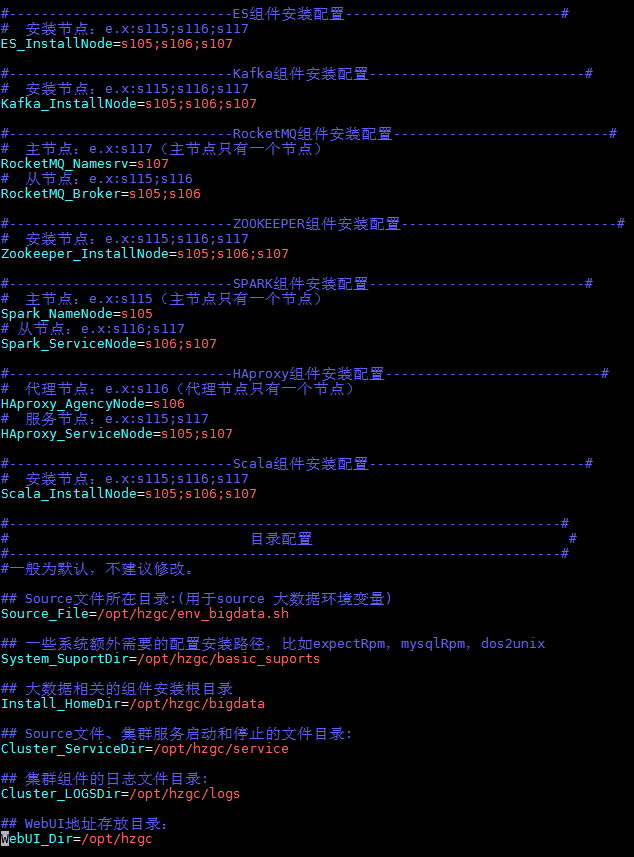
## 2.2修改配置文件

根据配置文件中注释的提示信息修改配置文件。

vim /opt/ClusterBuildScripts/conf/cluster\_conf.properties

部分配置如下：





修改后，将配置文件从dos格式转化为Unix格式：

:set ff=unix

并保存退出

## 2.3安装dos2unix并转换脚本执行格式

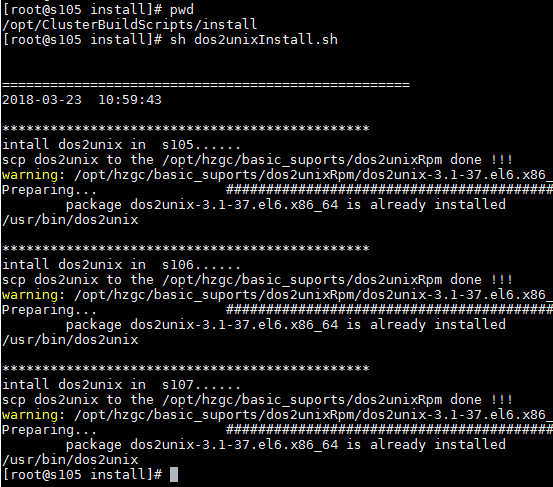
cd /opt/ClusterBuildScripts/install

vim dos2unixInstall.sh

:set ff=unix ##把脚本从dos格式转化为unix格式

:wq ##保存退出

sh dos2unixInstall.sh



转换该文件夹下所有脚本格式：

cd /opt/ClusterBuildScripts/ install

dos2unix \*

cd /opt/ClusterBuildScripts/ service

dos2unix \*

cd /opt/ClusterBuildScripts/ tool

dos2unix \*

（只对需要转化格式的脚本执行转化格式操作）

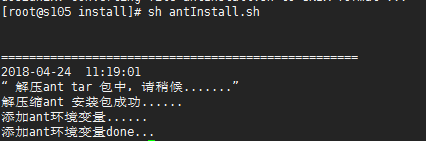
## 2.4安装expect包（在原集群基础上重装集群，该步骤可省略）

### 2.4.1 在当前操作机器（例如s105）上安装Ant工具

cd /opt/ClusterBuildScripts/install

sh antInstall.sh

执行结果如下：

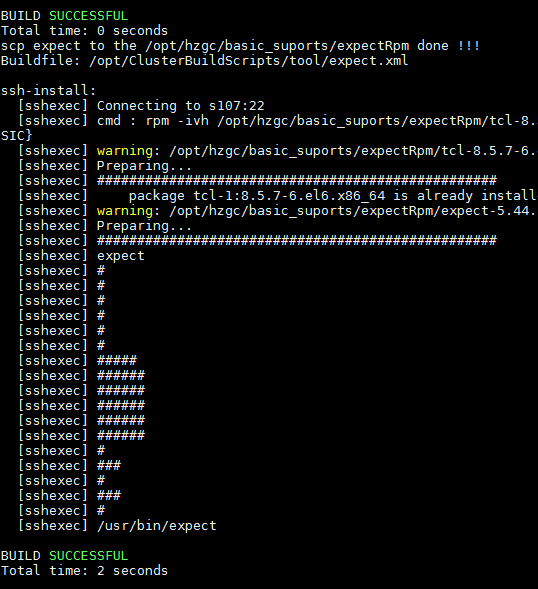


### 2.4.2 安装expect

cd /opt/ClusterBuildScripts/install

sh expectInstall.sh

部分执行结果如下：



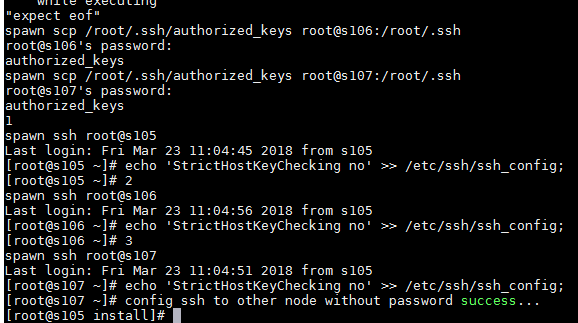
## 2.5配置ssh免密码登录（在原集群基础上重装集群，该步骤可省略）

cd /opt/ClusterBuildScripts/install

sh sshSilentLogin.sh

期间不要操作机器，等待配置完成。

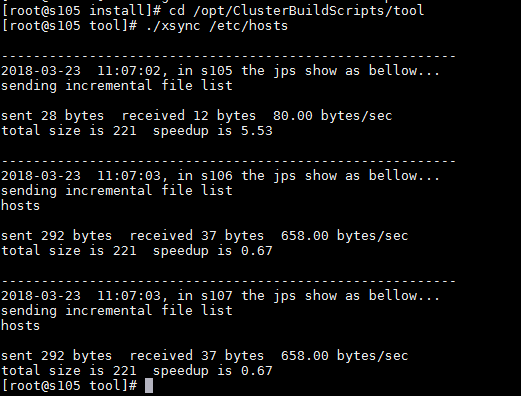
脚本执行部分结果如下：



## 2.6分发hosts文件（在原集群基础上重装集群，该步骤可省略）

cd /opt/ClusterBuildScripts/tool

./xsync /etc/hosts



## 2.7删除系统环境变量(如果是新建集群可不执行此步骤)

cd /opt/ClusterBuildScripts/install

sh delete\_env\_variable.sh



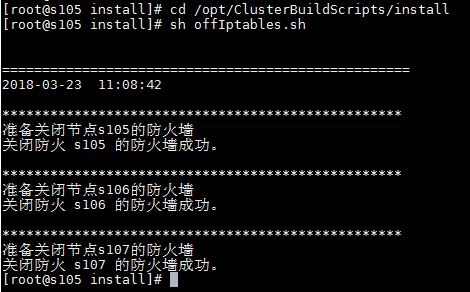
# 三.关闭防火墙及mysql安装（需重启机器）

执行完以下两步，需要重启集群。

## 3.1关闭防火墙（在原集群基础上重装集群，该步骤可省略）

cd /opt/ClusterBuildScripts/install

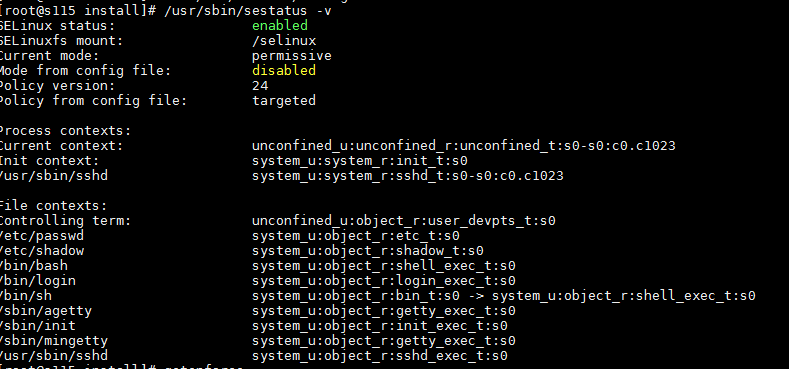
sh offIptables.sh



## 3.2安装Mysql（在安装包所在节点上安装）

### 3.2.1安装mysql前检查SELinux status状态

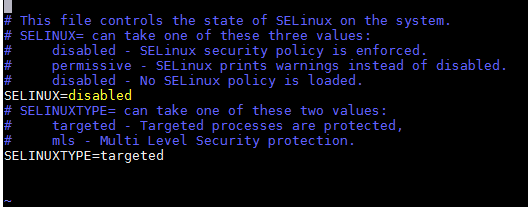
/usr/sbin/sestatus –v



若SELinux status为enabled，需要修改将SELinux status为disabled。执行如下命令：

vim /etc/selinux/config

D:\我的文档\Tencent Files\407932000\Image\C2C\]3UW8T$0[JXE{DLM%OV}_B6.png



### 3.2.2重启集群

修改完后需要重启该节点：reboot（未修改也要重启该节点）

重启后查看：/usr/sbin/sestatus –v



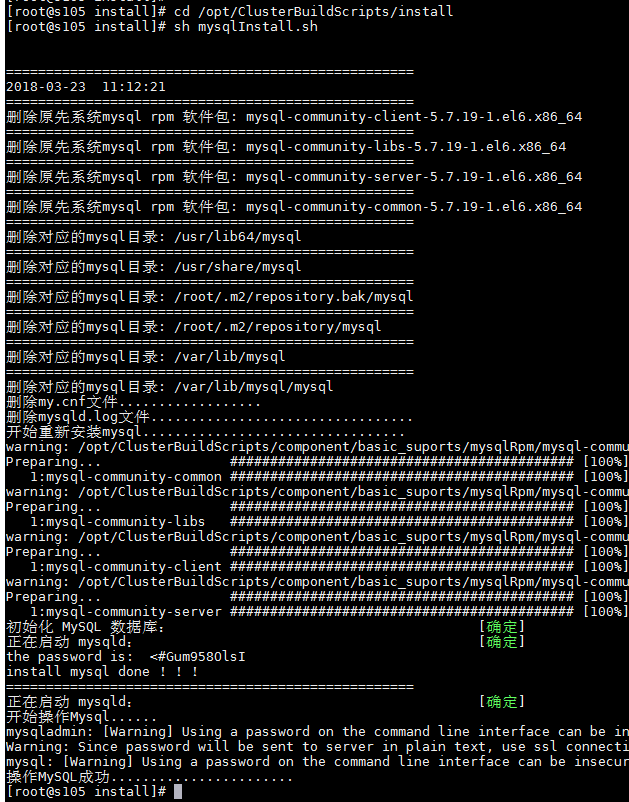
SELinux status为disabled了再安装mysql。

### 3.2.3安装mysql

cd /opt/ClusterBuildScripts/install

sh mysqlInstall.sh

脚本执行结果（涂红处显示失败没有关系）



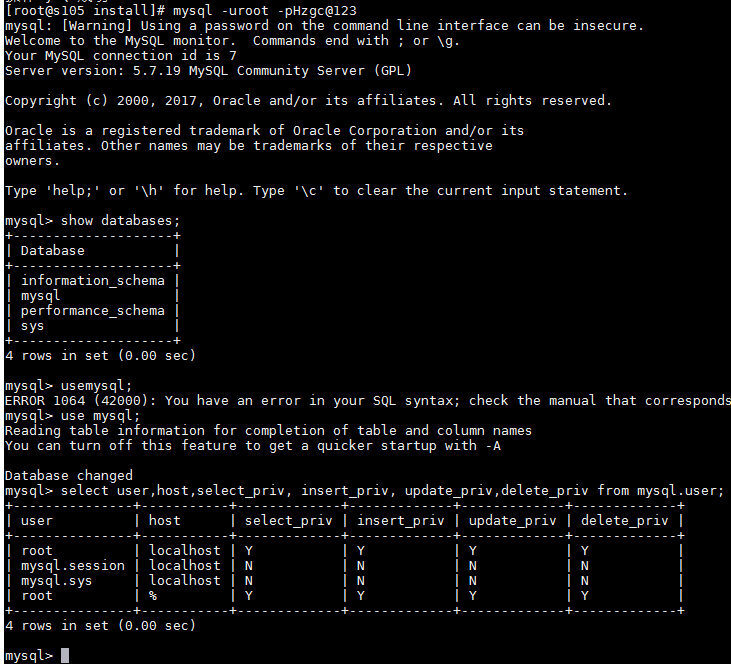
验证是否安装成功：

mysql -uroot -pHzgc@123

show databases;

use mysql;

select user,host,select\_priv, insert\_priv, update\_priv,delete\_priv from mysql.user;



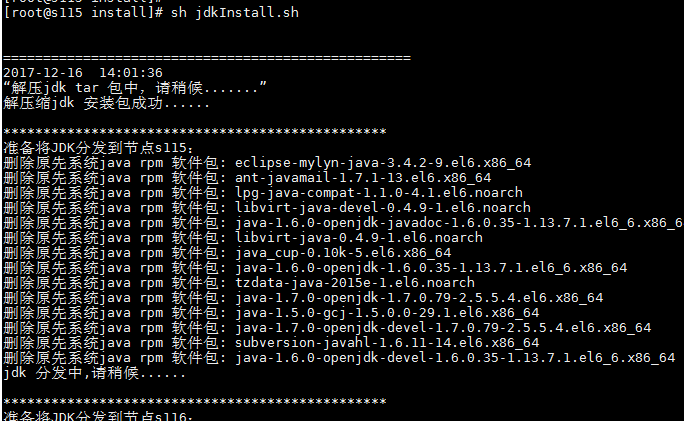
# 四.集群组件安装

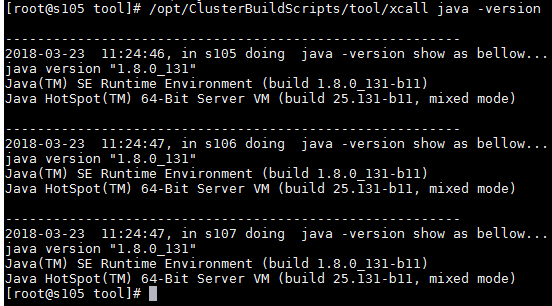
## 4.1 安装jdk

cd /opt/ClusterBuildScripts/install

sh jdkInstall.sh

部分执行结果如下：

/opt/ClusterBuildScripts/tool/xcall java -version

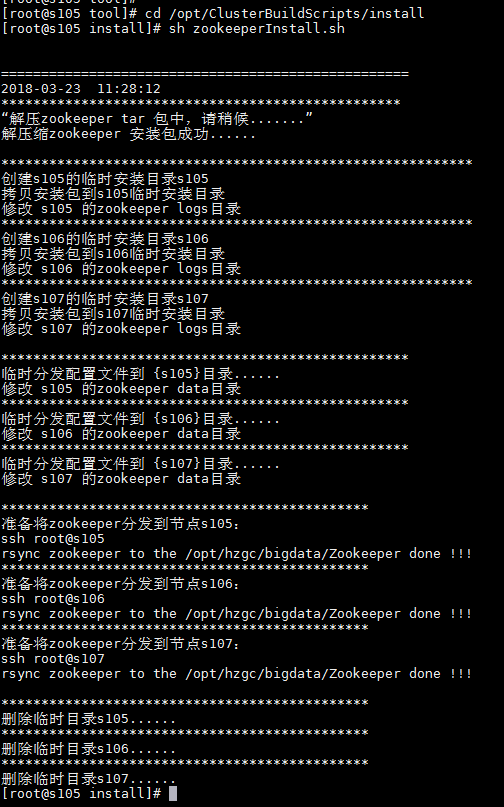


## 4.2.安装zookeeper

cd /opt/ClusterBuildScripts/install

sh zookeeperInstall.sh

部分执行结果如下：

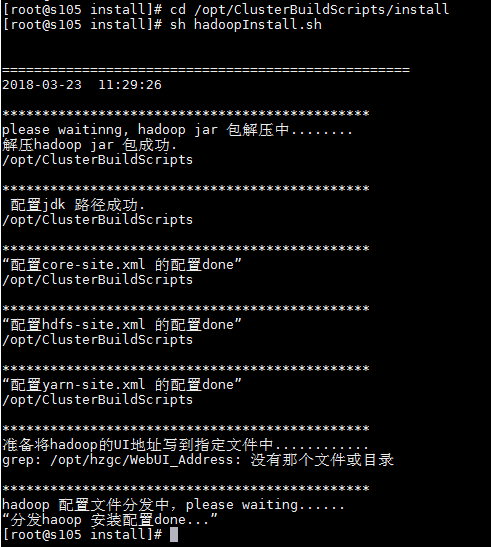


## 4.3.安装hadoop

cd /opt/ClusterBuildScripts/install

sh hadoopInstall.sh

执行结果如下：

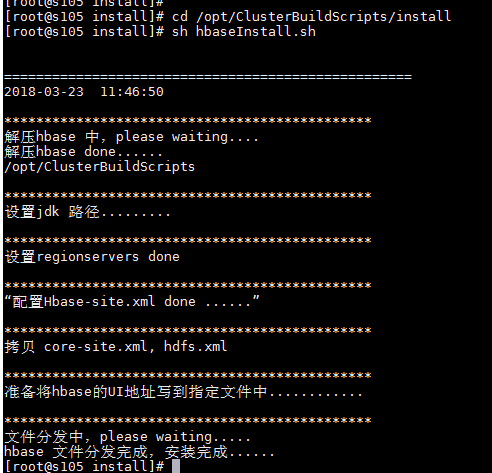


## 4.4.安装hbase

cd /opt/ClusterBuildScripts/install

sh hbaseInstall.sh

部分执行结果如下：

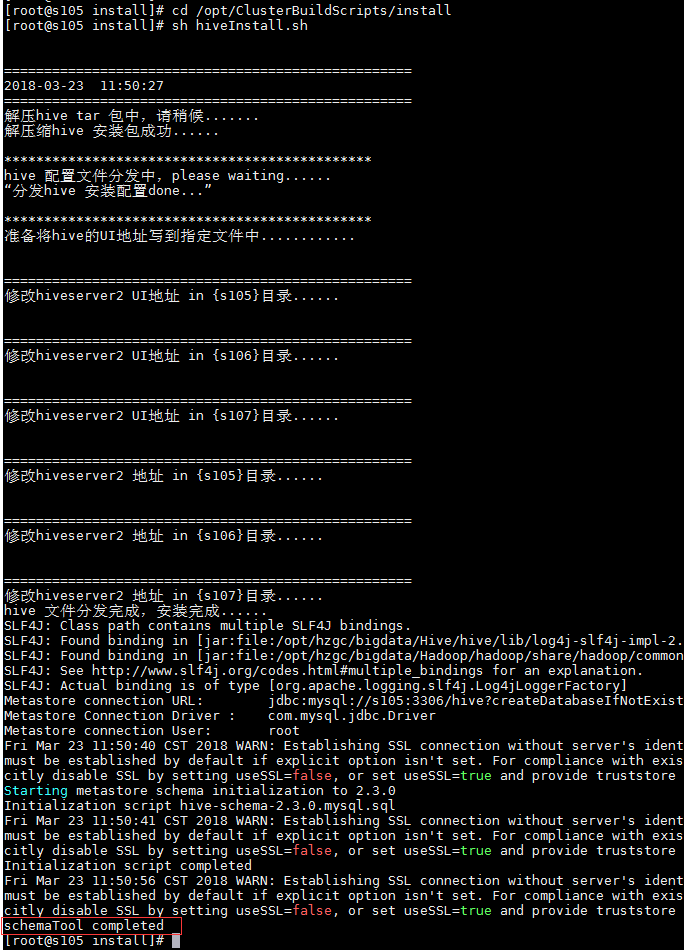


## 4.5.安装hive

cd /opt/ClusterBuildScripts/install

sh hiveInstall.sh

部分执行结果如下：

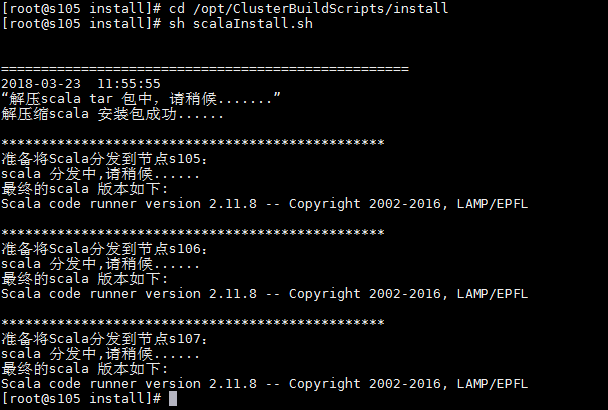


## 4.6.安装scala

cd /opt/ClusterBuildScripts/install

sh scalaInstall.sh

执行结果如下：

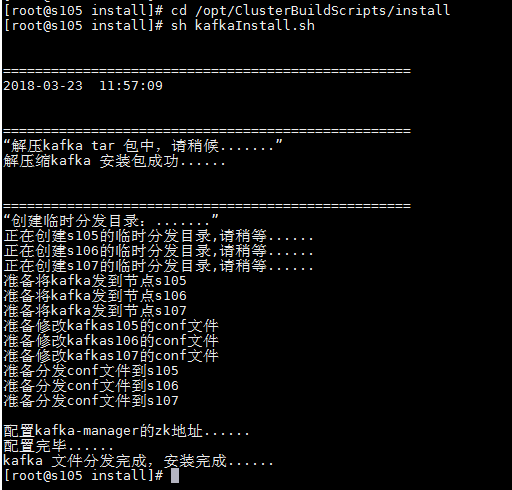


## 4.7.安装kafka

cd /opt/ClusterBuildScripts/install

sh kafkaInstall.sh

执行结果如下：

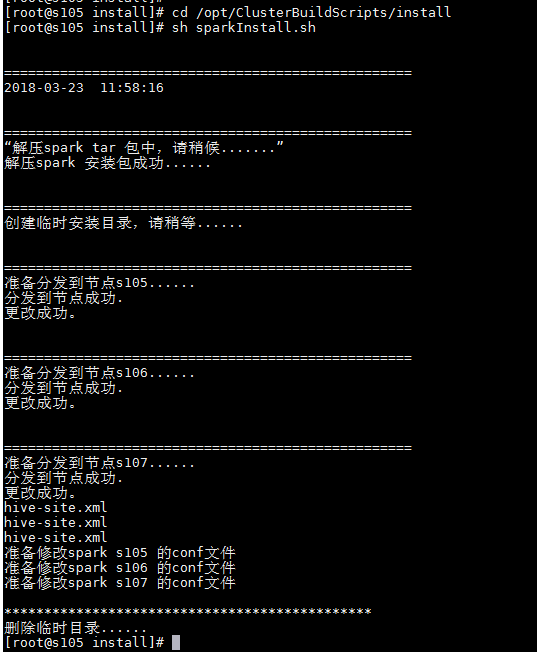


## 4.8.安装spark

cd /opt/ClusterBuildScripts/install

sh sparkInstall.sh

部分执行结果如下：

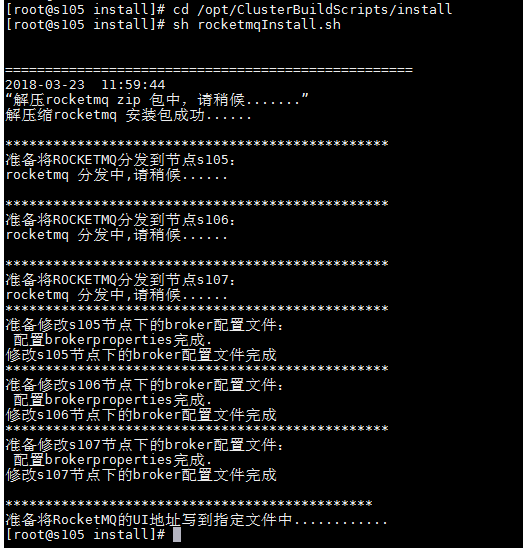


## 4.9.安装roketmq

cd /opt/ClusterBuildScripts/install

sh rocketmqInstall.sh

执行结果如下：

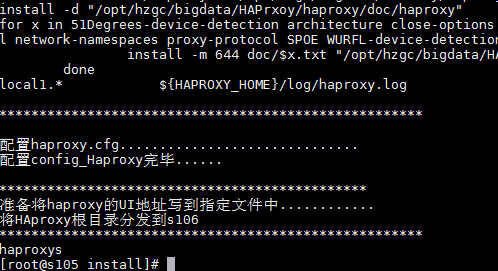


## 4.10.安装Haproxy

cd /opt/ClusterBuildScripts/install

sh haproxyInstall.sh

部分执行结果如下：

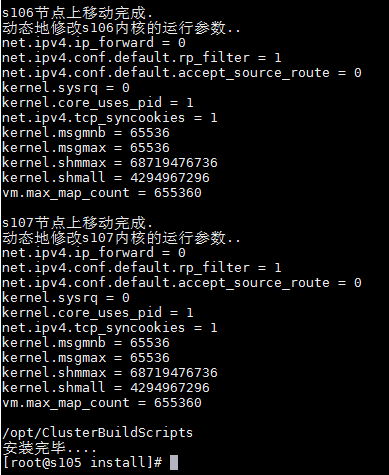


## 4.11.安装ElasticSearch

cd /opt/ClusterBuildScripts/install

sh elasticInstall.sh

部分执行结果如下：

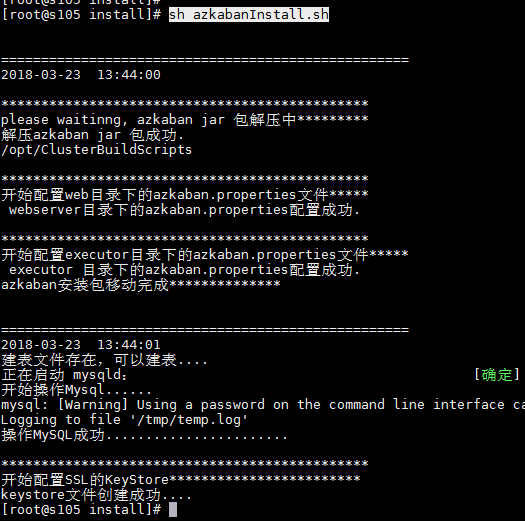


## 4.12.安装Azkaban

cd /opt/ClusterBuildScripts/install

sh azkabanInstall.sh

部分执行结果如下：

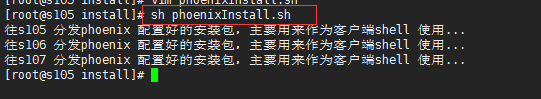


## 4.13.安装Phoenix

cd /opt/ClusterBuildScripts/install

sh phoenixInstall.sh

部分执行结果如下：



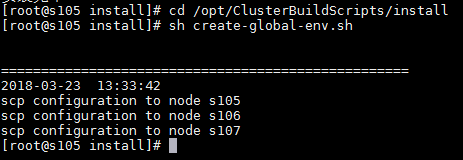
# 五.配置集群环境

## 5.1.拷贝环境变量到各个节点

cd /opt/ClusterBuildScripts/install

sh create-global-env.sh

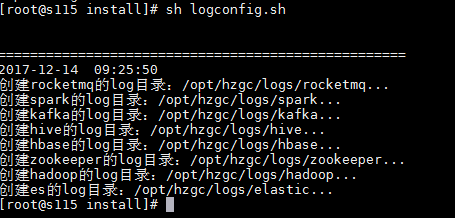
执行结果如下：



## 5.2配置集群组件日志

进入到/opt/ClusterBuildScripts/install目录，执行sh logconfig.sh(需要在执行es启动之前执行)

执行结果如下：

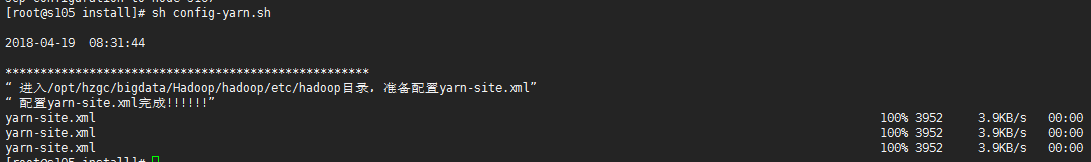


## 5.3.配置yarn的内存和cpu

cd /opt/ClusterBuildScripts/install

sh config-yarn.sh

执行结果如下：



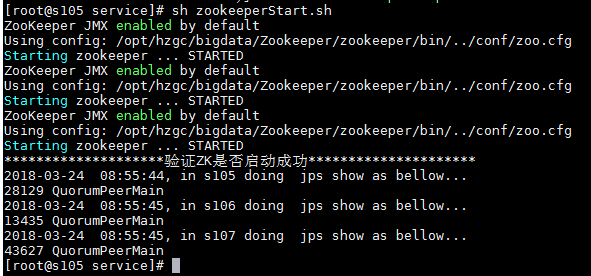
说明：yarn-utils.py这是一个根据输入的参数，自动计算yarn应该如何分配内存和cpu的脚本，算出结果后，将yarn-utils.py脚本所得结果配入yarn的配置文件。执行脚本config-yarn-CPU-RAM.sh就已经涵盖了整个过程，无需手动配置。

# 六.启动集群各组件

## 6.1启动zookeeper

cd /opt/ClusterBuildScripts/service

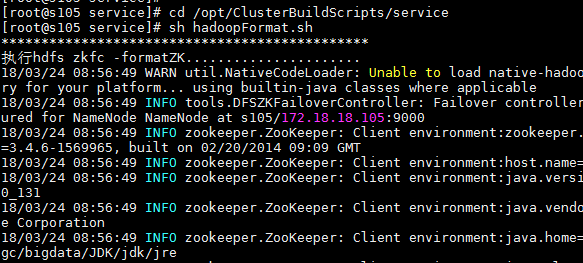
sh zookeeperStart.sh



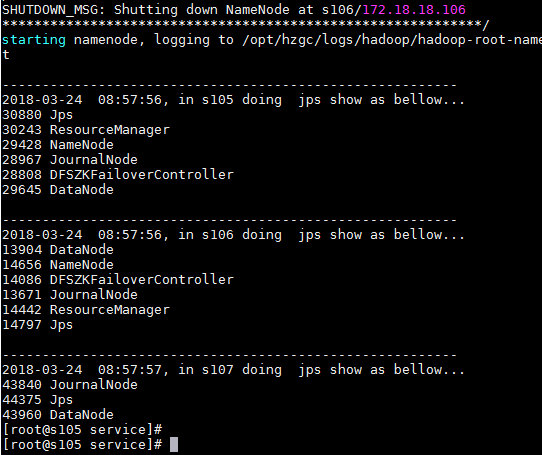
## 6.2启动hadoop

cd /opt/ClusterBuildScripts/service

sh hadoopFormat.sh



脚本部分执行结果

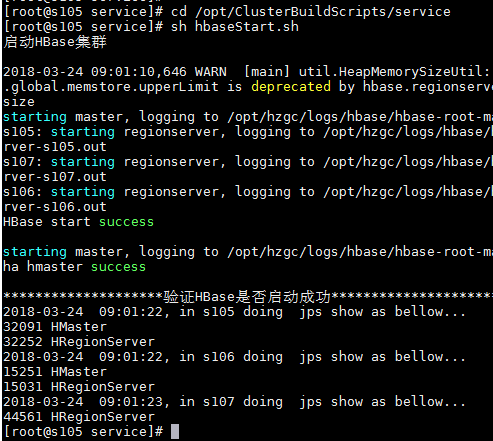


注意：此处不用执行hadoopStart.sh脚本，hadoopStart.sh用于停止hadoop后再启。

## 6.3启动hbase

cd /opt/ClusterBuildScripts/service

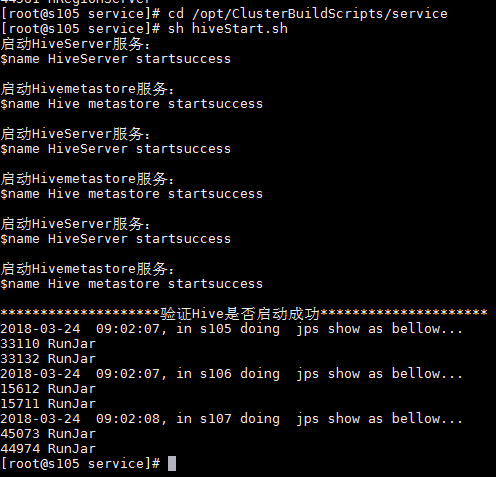
sh hbaseStart.sh



## 6.4启动hive（启动后每个节点有两个RunJar进程）

cd /opt/ClusterBuildScripts/service

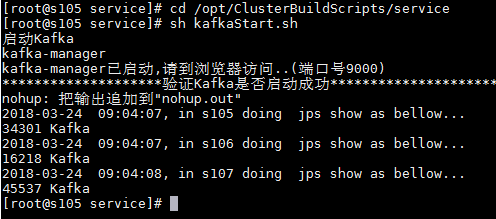
sh hiveStart.sh



## 6.5启动kafka

cd /opt/ClusterBuildScripts/service

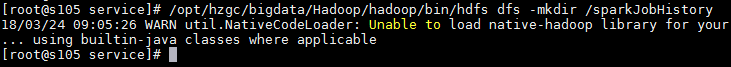
sh kafkaStart.sh



## 6.6使用spark jdbc

第一次使用spark请先在hdfs创建sparkJobHistory

/opt/hzgc/bigdata/Hadoop/hadoop/bin/hdfs dfs -mkdir /sparkJobHistory



## 6.7启动thriftserver服务

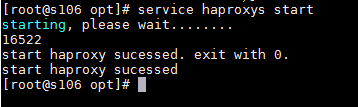
进入到/opt/ClusterBuildScripts/service目录下执行sh thriftServerStart.sh

使用方法：sh thriftServerStart.sh1G 1G 1 1 1（请根据实际情况配置），这五个参数分别表示：driver\_memory；executor\_memory；driver\_cores；executor\_cores；num\_executors

## 6.8启动Haproxy

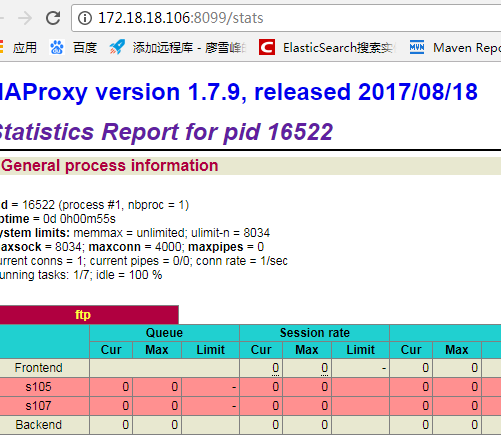
切换至Haproxy代理节点（如s106），（在任意目录下均可）

service haproxys start



在浏览器上验证是否启动：

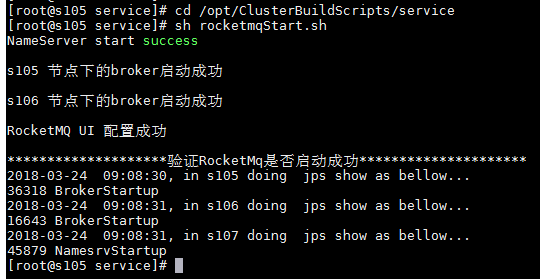
172.18.18.106:8099/stats



## 6.9启动rocketmq

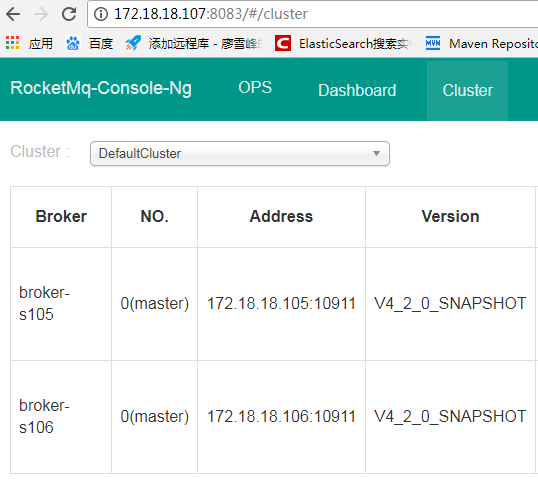
cd /opt/ClusterBuildScripts/service

sh rocketmqStart.sh



在浏览器上验证是否启动：

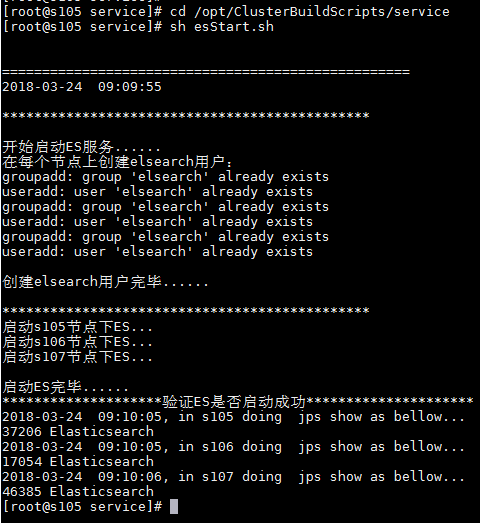
172.18.18.107:8083



## 6.10启动elasticsearch

cd /opt/ClusterBuildScripts/service

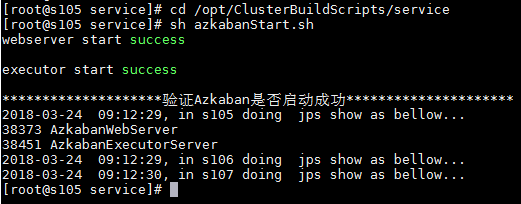
sh esStart.sh



## 6.11启动Azkaban(在安装mysql所在节点启动)

cd /opt/ClusterBuildScripts/service

sh azkabanStart.sh



# 七.服务验证

## 7.1 查看每台主机所起的服务

在任意目录下，执行：

source /opt/hzgc/env\_bigdata.sh

xcall jps

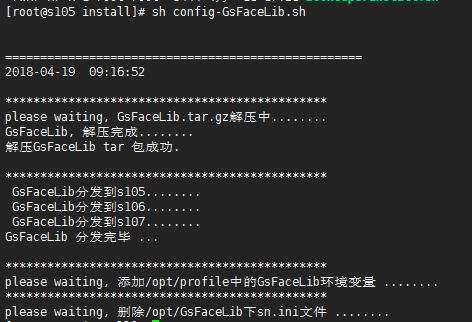


# 八. 配置算法库

## 8.1.分发GsFaceLib

cd /opt/ClusterBuildScripts/install

sh config-GsFaceLib.sh

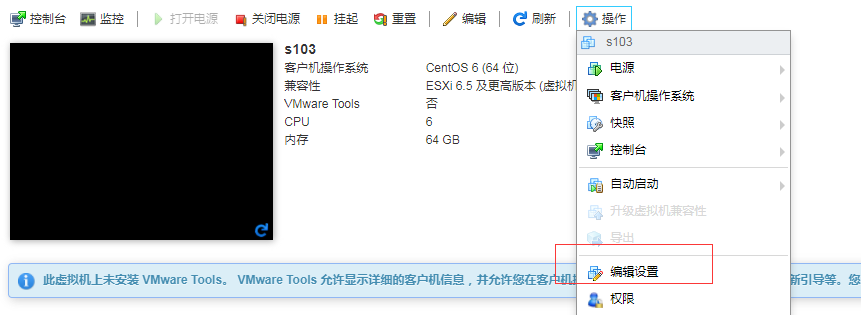


## 8.2.添加加密狗

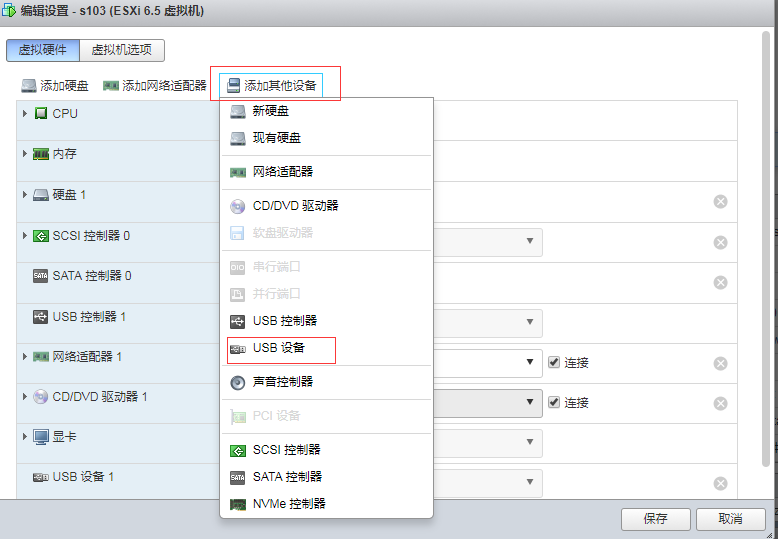
### 8.2.1 添加加密狗驱动（vmware情况下执行，不一定需要）

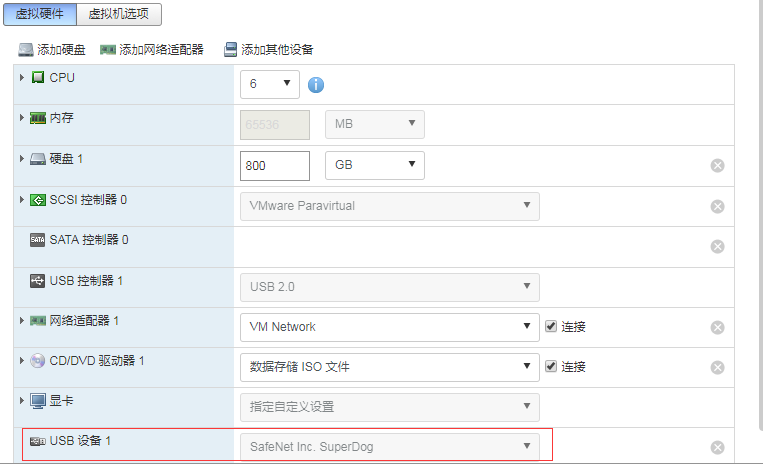
在对应的服务器中添加加密狗

登录对应服务器的vmware，编辑对应虚拟机设置：



添加USB控制器，后选择添加加密狗：

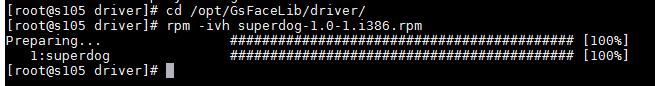




### 8.2.2 添加加密狗驱动（每台需要配置算法的服务器下执行）

cd /opt/GsFaceLib/driver/

rpm -ivh superdog-1.0-1.i386.rpm

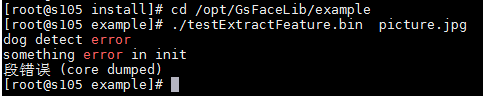


## 8.3.配置算法库

### 8.3.1.进入/opt/GsFaceLib/example目录下

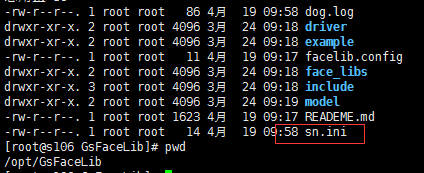
cd /opt/GsFaceLib/example

./testExtractFeature.bin picture.jpg



### 8.3.2.生成新的sn.ini文件

对新产生的sn.ini文件内容发给周琛晖转成相应的密钥后保存到sn.ini;

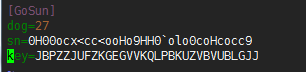


（假如sn.ini文件中只有dog=，没有sn=这一行，需要重新插拔一下加密狗。）

原本的sn.ini文件内容：



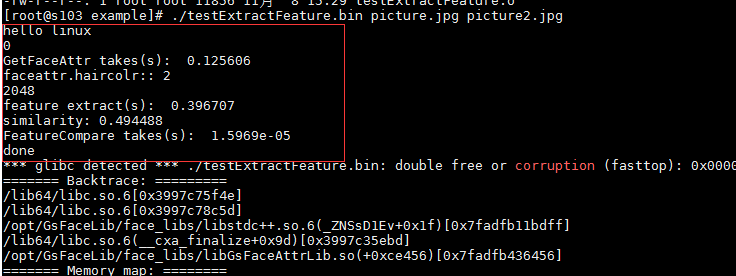
加了密钥后的sn.ini文件内容：



### 8.3.3. 继续进入/opt/GsFaceLib/example执行

cd /opt/GsFaceLib/example

./testExtractFeature.bin picture.jpg



# 九.重装集群

**假如有重装集群的需要，执行以下步骤。**

删除：/opt/hzgc下的所有文件；

建议删除：根目录下的/tmp。

再执行步骤一-八。

在原集群基础上重装集群，其中某些步骤可省略（已在对应的步骤中标注）。

# 十.常见安装问题

## 10.1 hbase错误

File /hbase/.tmp/hbase.version could only be replicated to 0 nodes instead of minReplica

问题原因：多次执行./hdfsnamenode-format 造成，解决办法手动在各个节点中删除tmp文件夹下data目录内容然后从新启动hdfs服务就可以启动了

## 10.2 启动thriftserver

Exception in thread "main" java.io.FileNotFoundException: File does not exist:

hdfs://hzgc/sparkJobHistory

问题原因：启动前请现在hdfs上创建sparkJobHistory目录，

/opt/client/Hadoop/hadoop/bin/hdfsdfs -mkdir /sparkJobHistory

## 10.3 mysql没有被删除

删除mysql时有好几个步骤：

①卸载rpm包，②删除mysql所有目录，③停止服务，测试时需要mysql+hive一起，看看初始化hive是否成功

(1)mysql没有卸载干净会导致hive初始化schema出错

(2)删除mysql的时候要删除/var/log/mysql.log文件，否则会查出多个密码

**问题：**

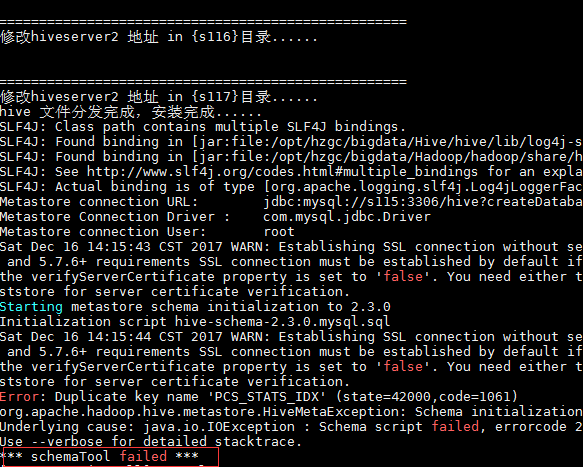
安装Hive时出现：

Error: Duplicate key name 'PCS\_STATS\_IDX' (state=42000,code=1061)

\*\*\* schemaTool failed \*\*\*

**解决方法：**在hive的安装脚本中，初始化元数据步骤${HIVE\_HOME}/bin/schematool -initSchema -dbTypemysql只需在安装了mysql的节点上执行，而不是ssh到每个节点上执行。

## 10.4 连续两次执行hiveInstall脚本报错



**解决方法：**删除mysql中hive数据库，再执行安装脚本