搭建大数据平台

# 一．集群搭建准备（以3 台机器为例）

## 1.1准备机器

|  |  |
| --- | --- |
| 主机名 | 主机ip |
| S153 | 172.18.18.153 |
| S154 | 172.18.18.154 |
| S155 | 172.18.18.155 |

## 1.2上传平台安装包bigdata\_hzgc.tar.gz到172.18.18.153(选择ip最小的节点上传)的/home/hzgc目录(目录可换)

**1.2.1解压安装包**

cd /home/hzgc

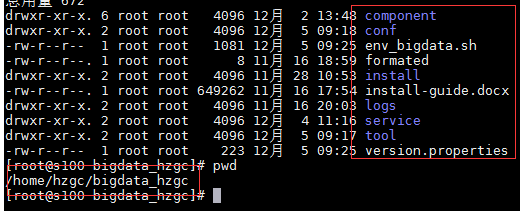
tar -xf bigdata\_hzgc.tar.gz

chmod –R 755 bigdata\_hzgc

## 1.3安装包bigdata\_hzgc目录结构和简要说明

**1.3.1目录结构**

（1）根目录

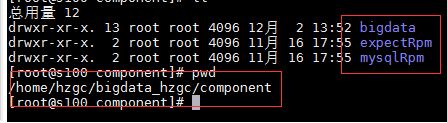


（2）component

包括大数据平台安装包

expect工具RPM包

mysql数据库RPM包



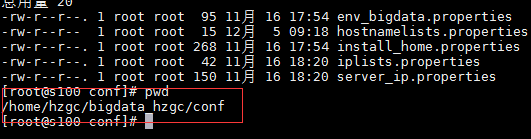
（3）conf

大数据平台使用的环境变量env\_bigdata.properties

主机名和主机ip配置文件hostnamelists.properties和iplists.properties

大数据平台安装目录install\_home.properties

RocketMQ使用的server\_ip.properties

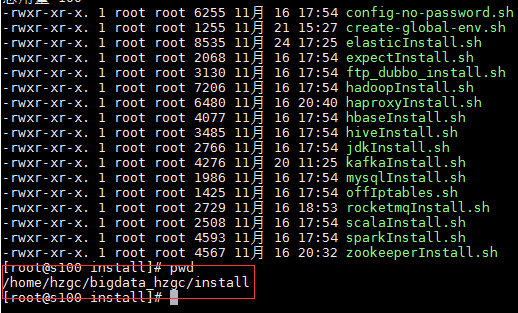


（4）install

大数据平台组件，ftp，dubbo，haproxy，mysql的安装脚本

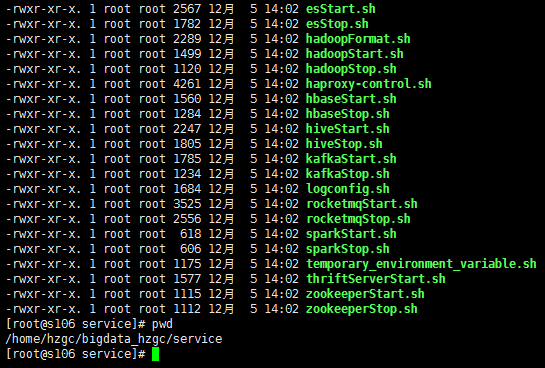
ssh无密钥登录工具脚本config-no-password.sh

expect工具安装脚本expectInstall.sh



（5）service

大数据平台组件启动和停止脚本

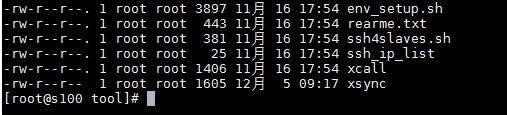


（6）tool

ssh无密钥工具脚本ssh4slaves.sh和ssh无密钥配置文件ssh\_ip\_list

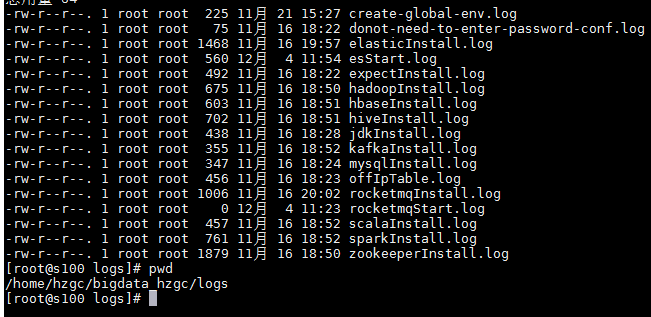
查看集群进程脚本xcall

集群拷贝脚本xsync



（7）logs

集群安装日志目录



（8）version.properties 大数据平台组件安装包信息

│ hadoop.tar.gz 版本2.7.2

│ hbase.tar.gz 版本1.2.6

│ hive.tar.gz 版本2.3.0

│ jdk.tar.gz 版本1.8

│ kafka.tgz 版本1.0.0

│ scala.tgz 版本2.11.8

│ spark.tgz 版本2.2.0

│ zookeeper.tar.gz 版本3.4.10

│ elastic.tar.gz 版本5.5.0

│ rocketmq.zip 版本4.1.0

│ haproxy-1.7.9.tar.gz 版本3.4.10

# 二．配置主机名和主机ip映射

## 2.1修改主机s153的/etc/hosts文件 （ip和主机名根据实际情况填写）

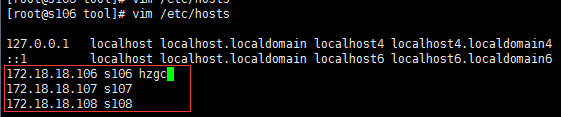
vim /etc/hosts

写入一下内容，并保存退出(hzgc可配在任意节点下)

172.18.18.153 s153 hzgc

172.18.18.154 s154

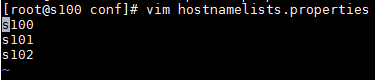
172.18.18.155 s155



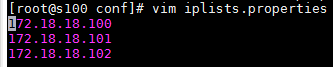
## 2.2分发至s154和s155（根据提示输入yes 和密码）

2.2.1 修改/home/hzgc/bigdata\_hzgc/conf下相关配置文件

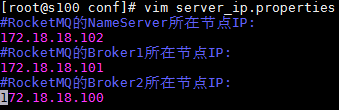
修改hostnamelists.properties



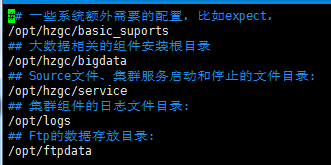
iplists.properties



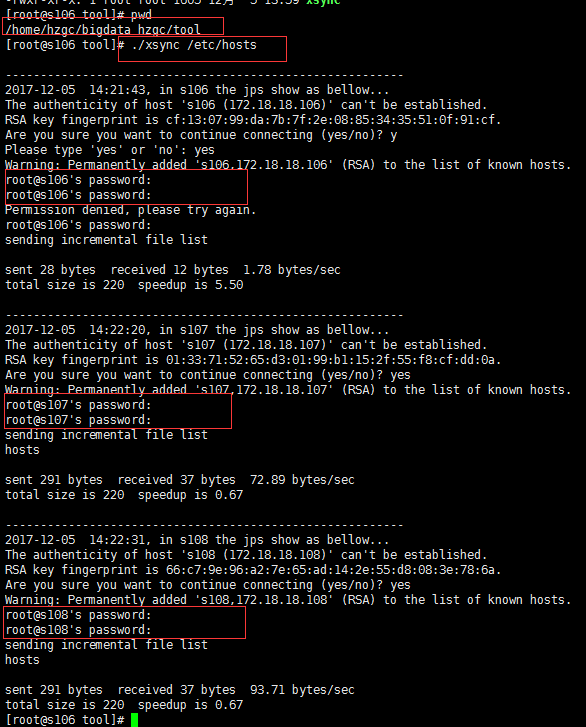
server\_ip.properties



install\_home.properties（请尽量保持默认配置）

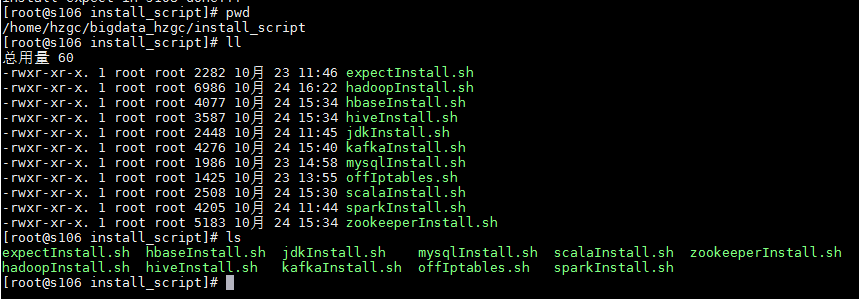


2.2.2 执行/home/hzgc/bigdata\_hzgc/tool下xsync脚本（根据提示输入yes 和机器密码）

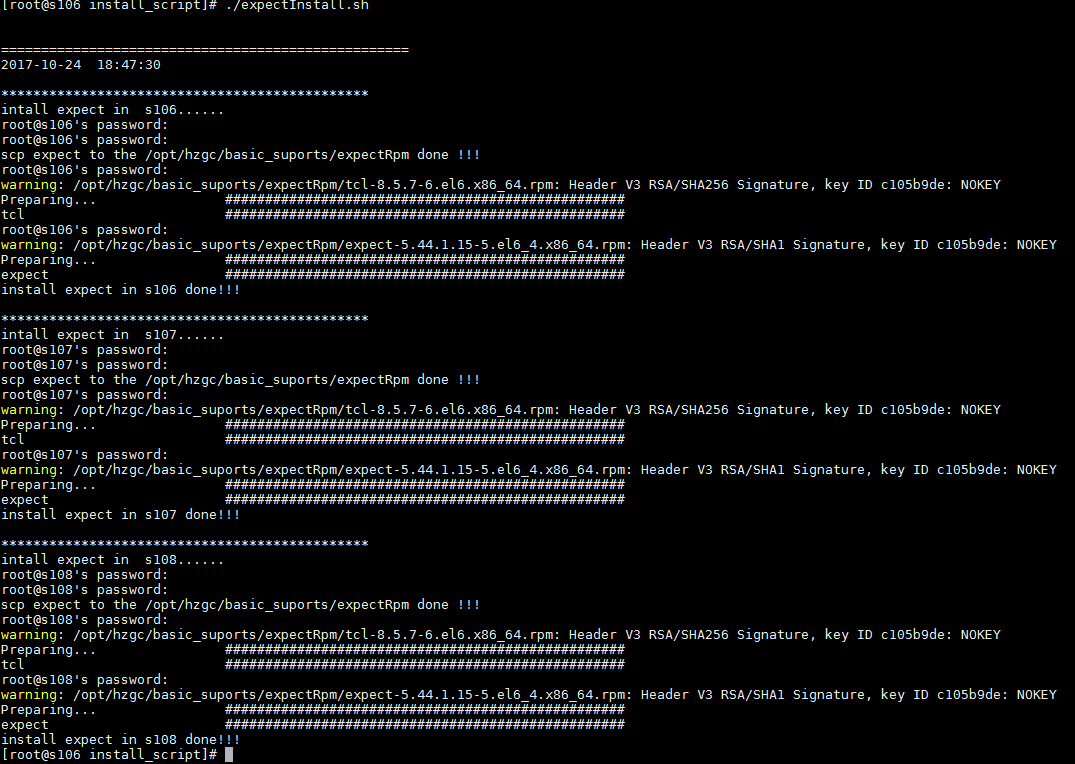


## 2.3安装expect包（根据提示输入yes 和机器密码）

2.3.1脚本所在目录/home/hzgc/bigdata\_hzgc/install，执行./expectInstall.sh



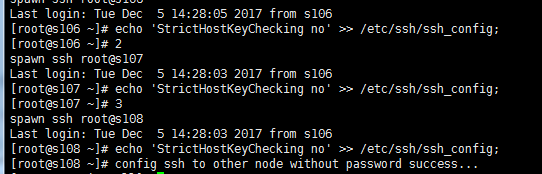
2.3.2根据提示输入yes和密码



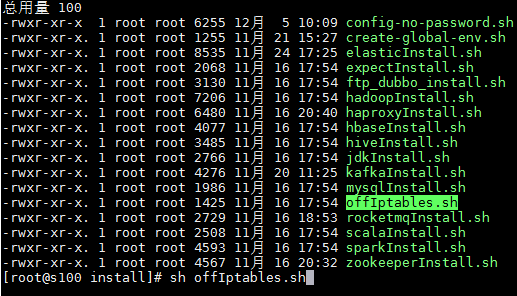
# 三．配置ssh免密码登录

3.1进入到/home/hzgc/bigdata\_hzgc/install目录，sh config-no-password.sh

期间不要操作机器，等待配置完成。

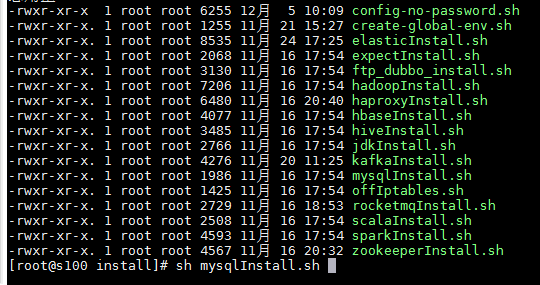


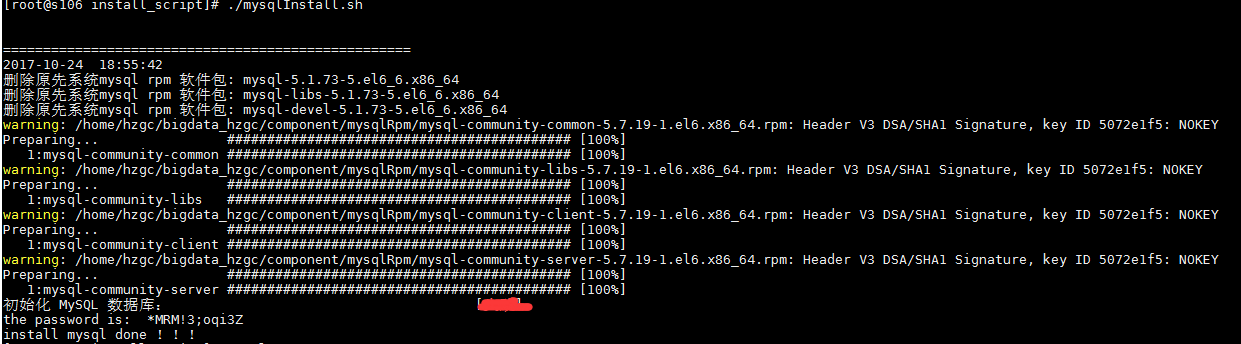
# 四．关闭防火墙



# 五．安装mysql（在安装包所在节点上安装）

5.1脚本所在目录/home/hzgc/bigdata\_hzgc/install

5.5.2脚本执行结果（涂红处显示失败没有关系）



5.3启动mysql服务

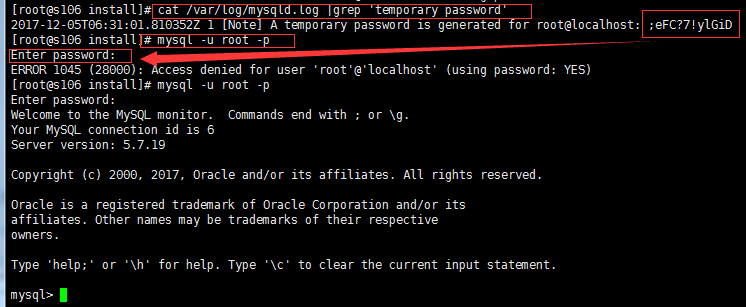


5.4修改默认密码为自定义密码(密码获取如图上。)

先获取密码：执行命令cat /var/log/mysqld.log |grep 'temporary password'

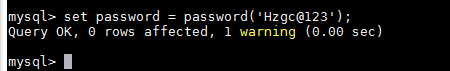
登陆msyql,执行以下命令并将文件中的密码输入

mysql -u root -p



修改密码

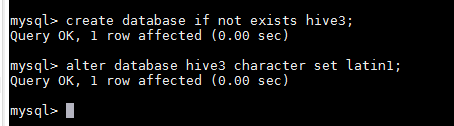
mysql> SET PASSWORD = PASSWORD('Hzgc@123');



5.5创建数据库存放hive元数据

mysql>create database if not exists hive;

mysql>alter database hive character set latin1;



5.6授权远程连接

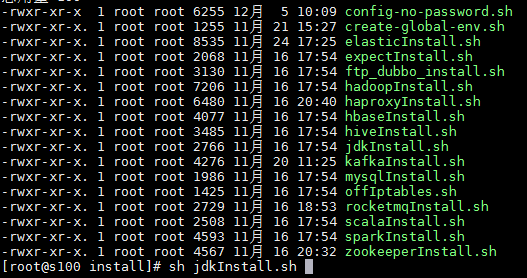
mysql> GRANT ALL PRIVILEGES ON \*.\* TO 'root'@'%' IDENTIFIED BY 'Hzgc@123' WITH GRANT OPTION;

mysql> flush privileges;

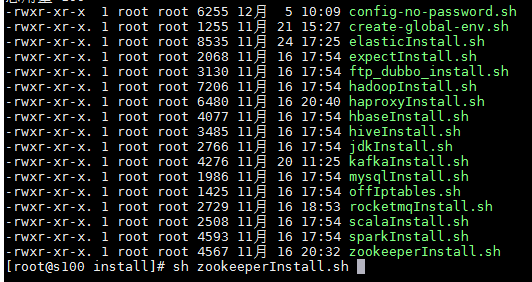
5.7 退出

mysql> \q

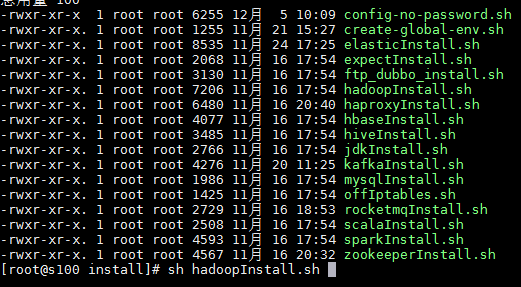
# 六．安装jdk



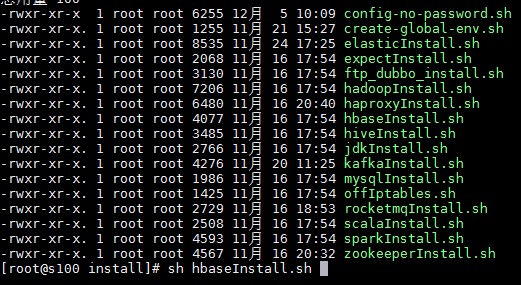
# 七．安装zookeeper



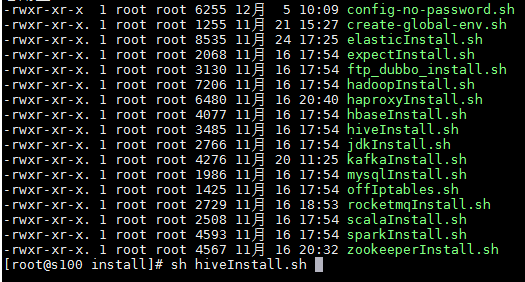
# 八．安装hadoop（默认最小的IP为主节点）



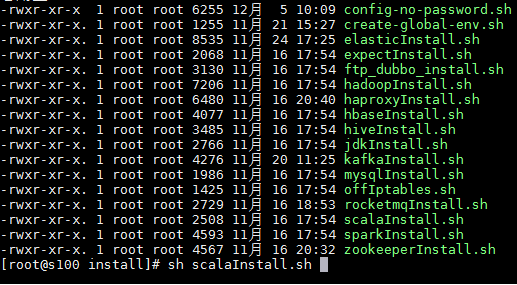
# 九．安装hbase



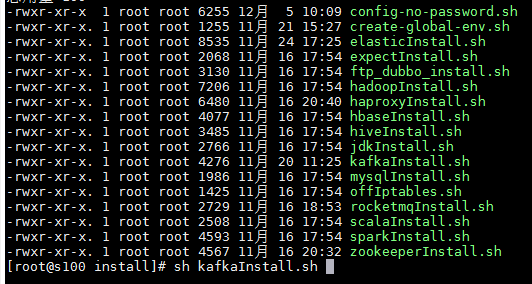
# 十．安装hive



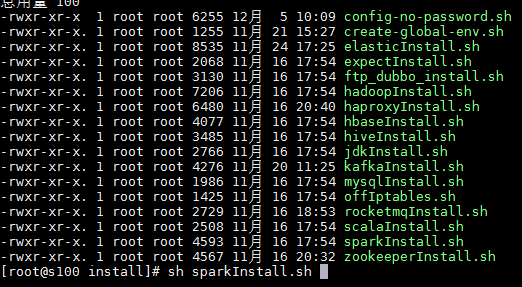
**十一．安装scala**



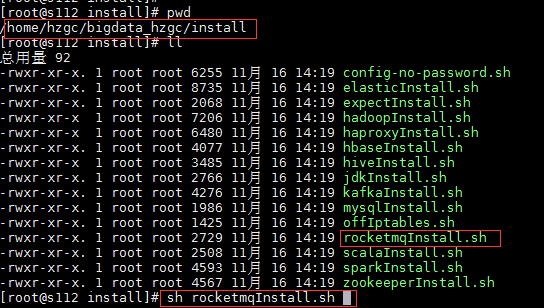
**十二．安装kafka**



**十三．安装spark**

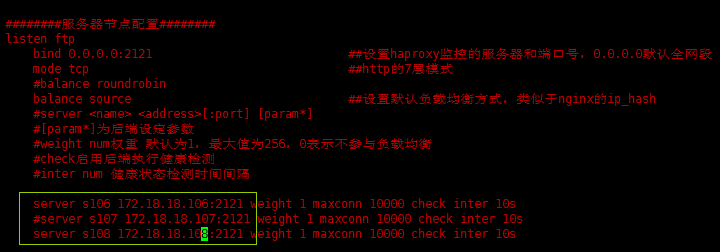


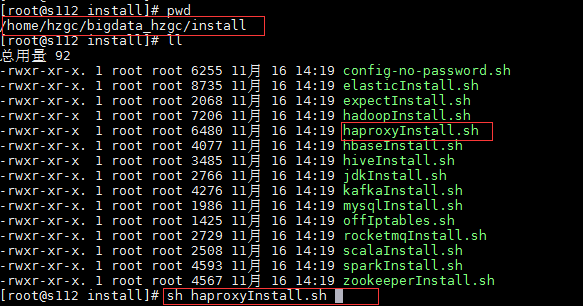
# 十四．安装roketmq



# 十五．安装Haproxy

根据实际的节点需要修改脚本haproxyInstall.sh



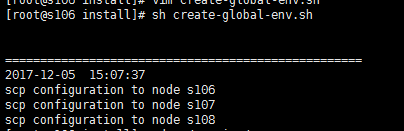


# 十六．安装ElasticSearch



# 十七．拷贝环境变量以及相关环境变量到各个目录

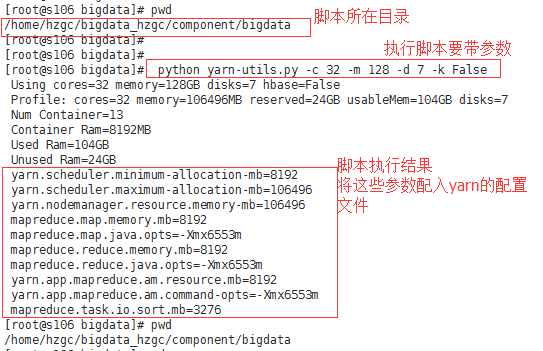
在install 目录下执行sh create-global-env.sh

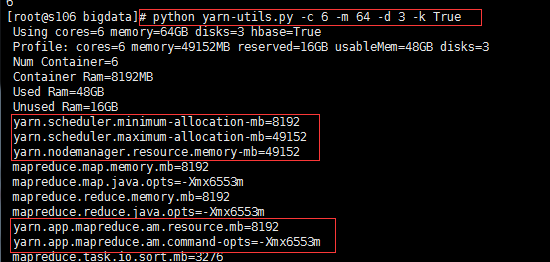


# 十八．配置集群组件日志路径

在service目录下执行sh logconfig.sh（需要在执行es启动之前执行）

# 十九．配置yarn的内存和cpu





**执行步骤：**

**19.1**进入集群安装包目录/home/hzgc/bigdata\_hzgc/component/bigdata

**19.2**执行sh yarn-utils.py：执行时带上参数，e.x：python yarn-utils.py -c 6 -m 64 -d 3 -k True

其中，-c为机器CPU核数，-m为机器总内存，-d为机器上挂载的磁盘个数，-k表示是否使用hbase（值为true或false）

备注：查看机器的CPU核数（逻辑CPU个数）：cat /proc/cpuinfo| grep "processor"| wc –l

查看内存：free –m

查看磁盘个数：df –k

**19.3**进入Hadoop安装目录



**19.4**将脚本执行结果配入hadoop/etc/hadoop/yarn-site.xml的相应部分：



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，将脚本所得结果配入yarn的配置文件。

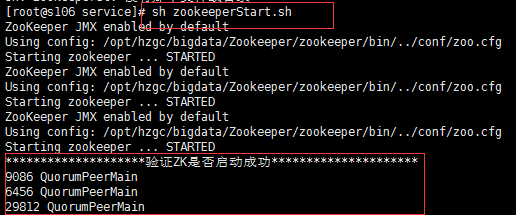
# 二十．启动服务

## 20.1启动zookeeper

进入到service目录下

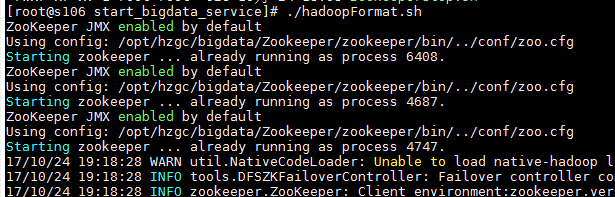


在s153主机上执行

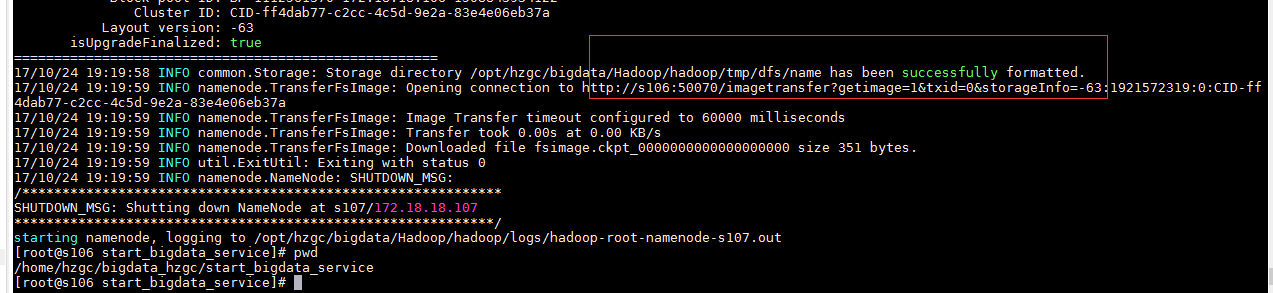


**20.2启动hadoop**

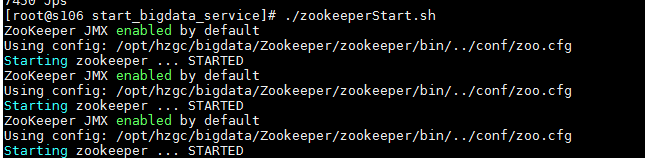
20.2.1第一次启动hadoop请按以下顺序执行



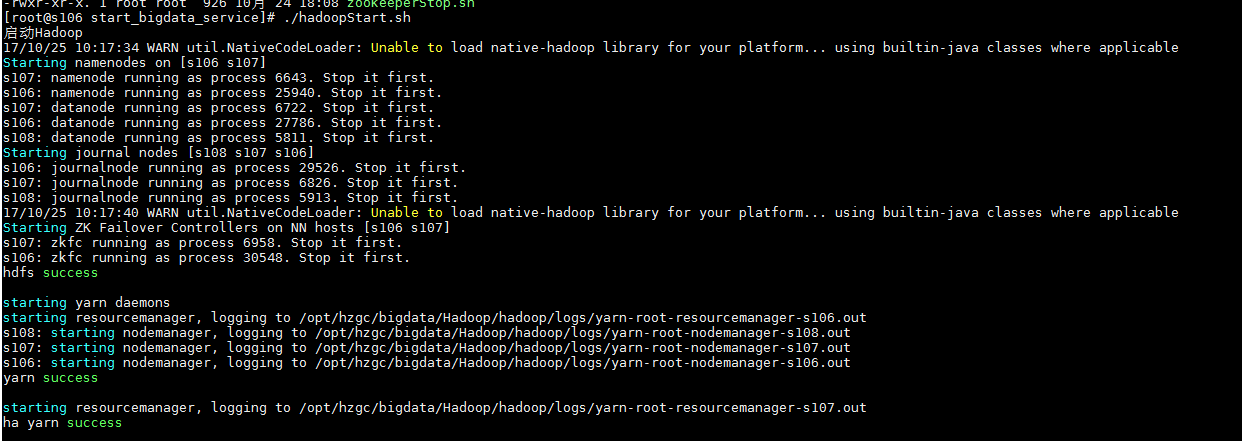
20.2.2脚本执行结果



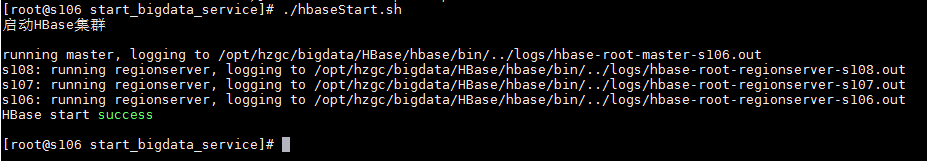
20.2.3启动zookeeper



20.2.4启动hadoop

**20.3启动hbase**

在s153主机上执行

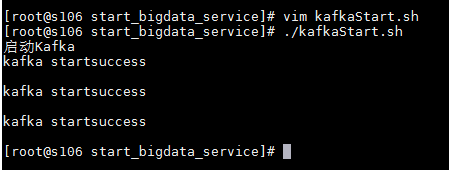


**20.4启动hive（启动后每个节点有两个RunJar 进程）**



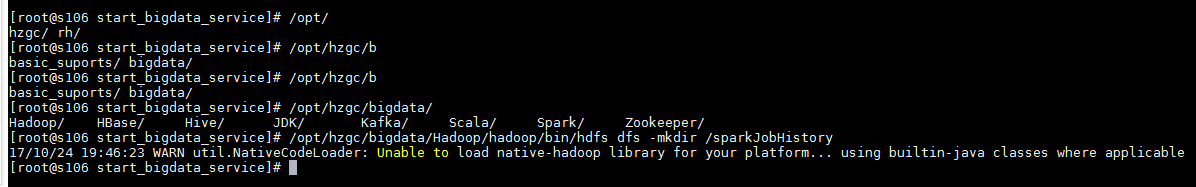
## 20.5启动kafka

在s153主机上执行



## 20.6使用spark jdbc

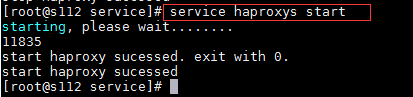
20.6.1第一次使用spark请先在hdfs创建sparkJobHistory



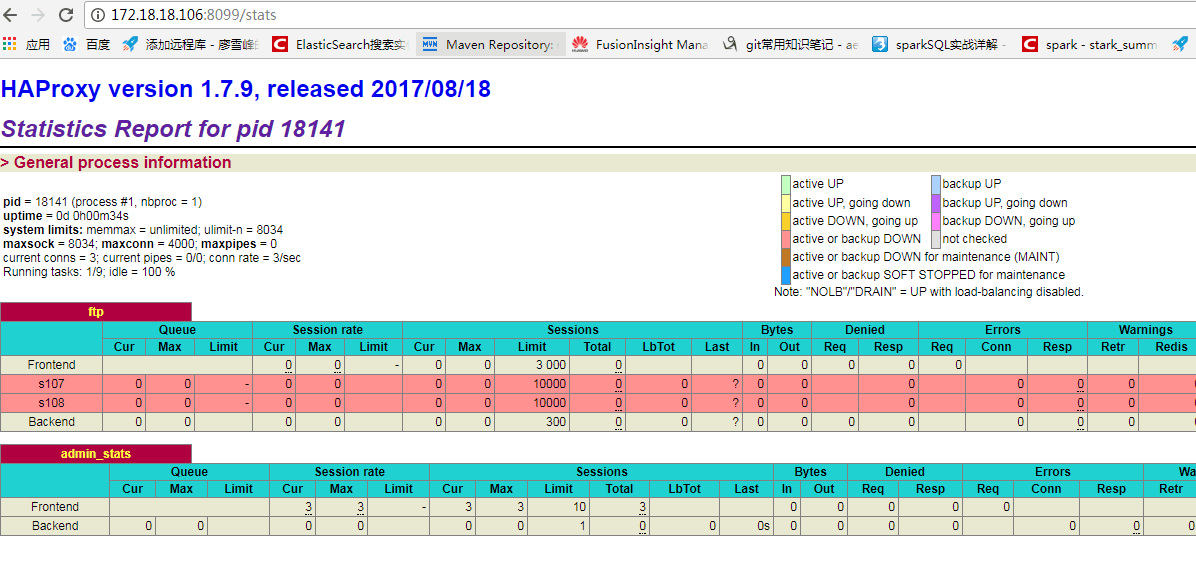
20.6.2启动thriftserver服务

/home/hzgc/bigdata/service/start-thriftserver\_conf.sh 1G 1G 1 1 1（请根据实际情况配置）

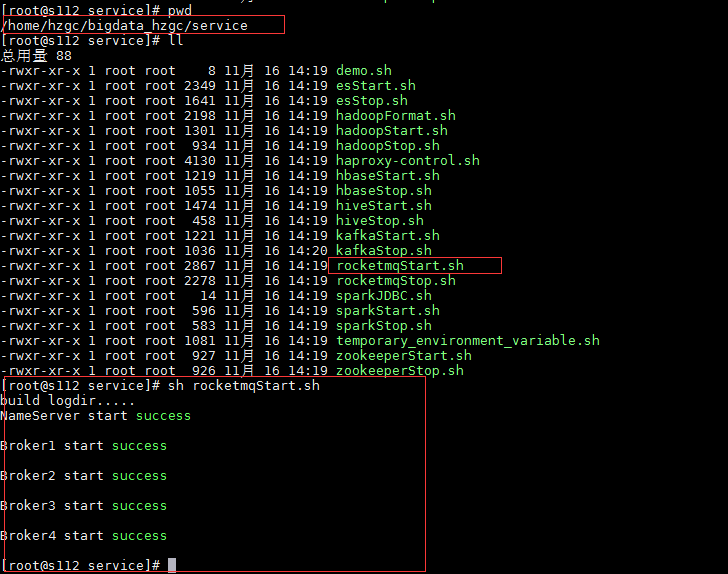
## 20.7启动Haproxy



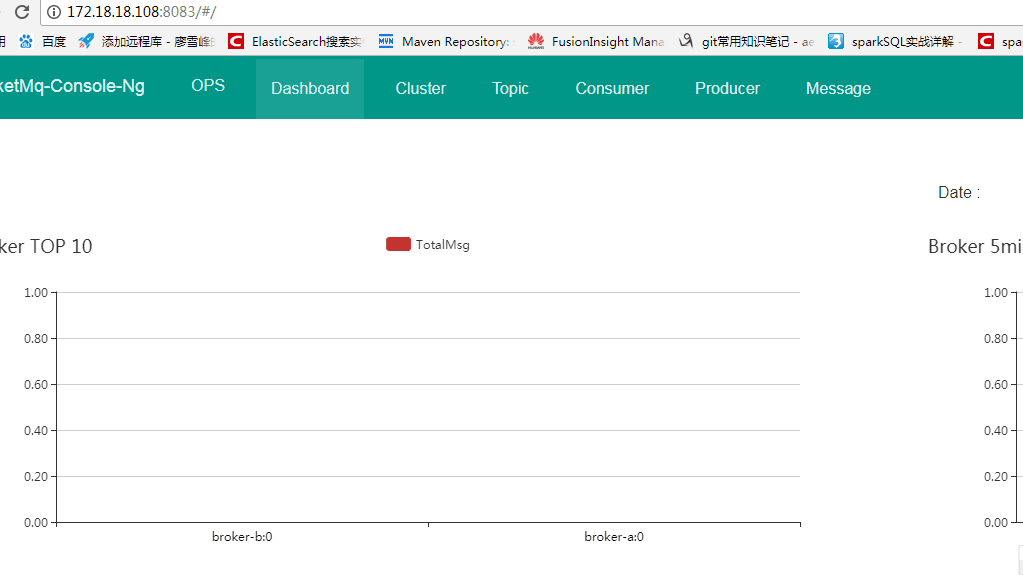
在浏览器上验证是否启动：



## 20.8启动rocketmq



在浏览器上验证是否启动：



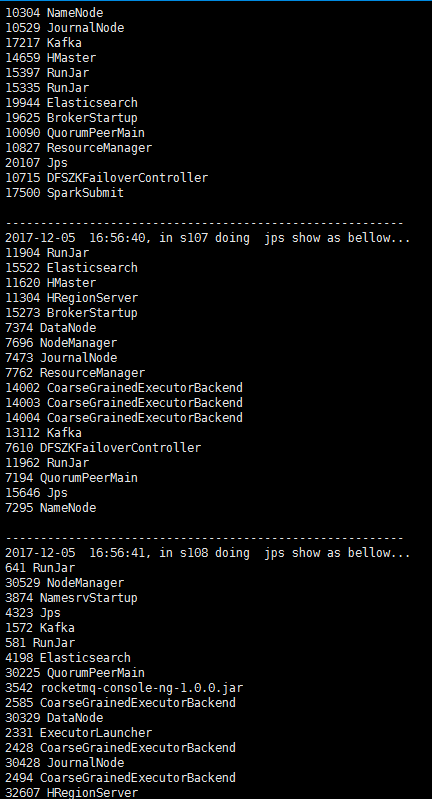
## 20.9启动elasticsearch

在service目录下，执行sh esStart.sh

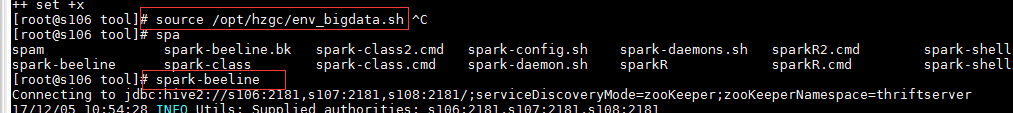


# 二十一.服务验证

## 21.1 每台主机所起的服务



21.2连接jdbc



**二十二.常见安装问题**

## 22.1 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./hdfs namenode -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

## 22.2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/hdfs dfs -mkdir /sparkJobHistory