搭建大数据平台

1. 集群搭建准备（以3 台机器为例）

## 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 |

## 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)

解压安装包

cd /home/hzgc

tar -xf bigdata\_hzgc.tar.gz

## 3.安装包bigdata\_hzgc目录结构和简要说明



software 大数据平台组件安装包

│ 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2.11\_0.11.0.1

│ scala.tgz 版本2.11.8

│ spark.tgz 版本2.2.0

│ zookeeper.tar.gz 版本3.5.1-alpha

二．配置主机名和主机ip映射

## 1，修改主机s153的/etc/hosts文件

vim /etc/hosts

写入一下内容，并保存退出(hzgc可配在任意节点下)

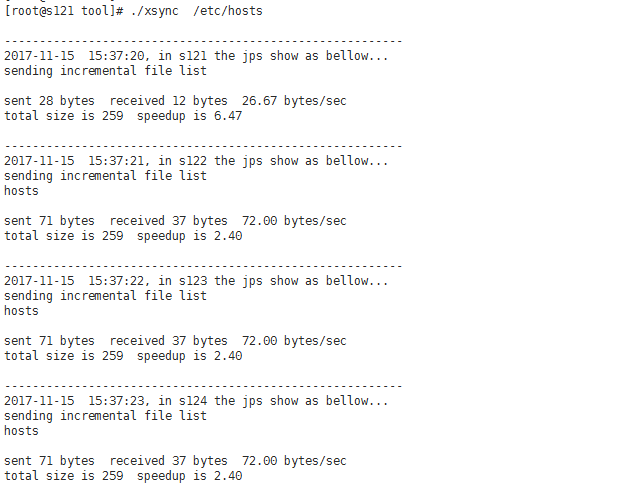
172.18.18.153 s153 hzgc

172.18.18.154 s154

172.18.18.155 s155



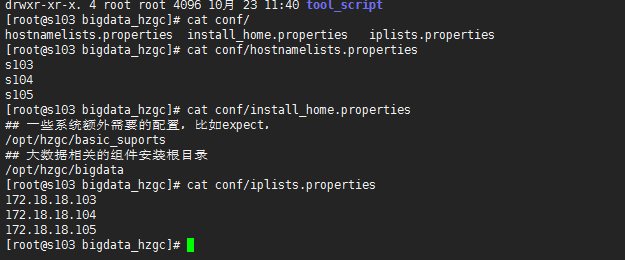
## 2，分发至s154和s155（根据提示输入yes 和密码）



## 3，配置bigdata\_hzgc/conf 下配置文件

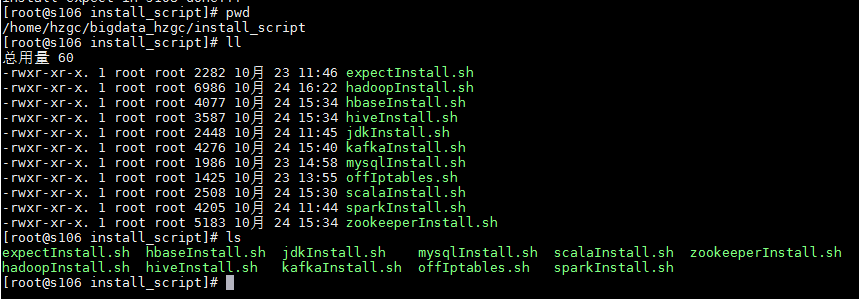
1.只需要修改hostname.properties 文件和 iplists.properties

2.Install\_home.properties文件尽量不要修改

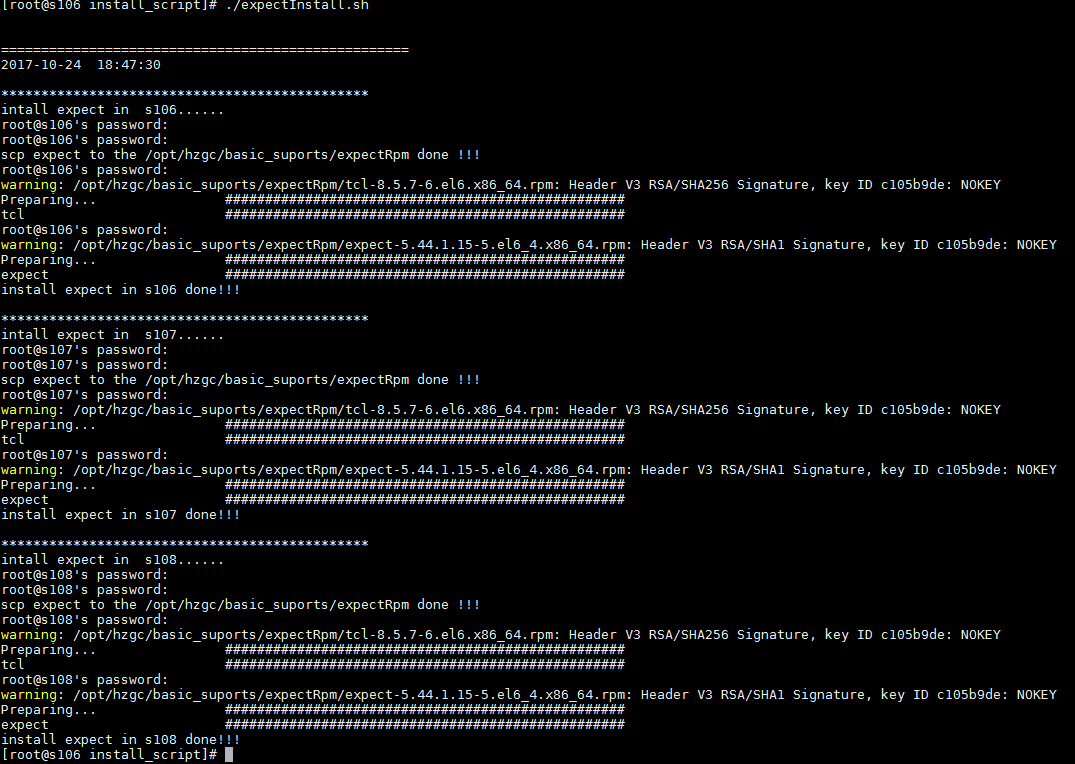


## 4，安装expect包（根据提示输入yes 和机器密码）

（1）脚本所在目录，执行./expectInstall.sh



（2）根据提示输入yes和密码

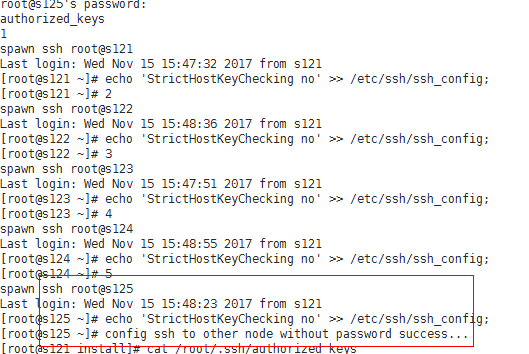


三．配置ssh免密码登录

进入到install 目录，

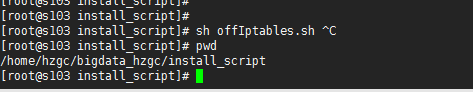
sh config-no-password.sh

期间不要操作机器，等待配置完成。

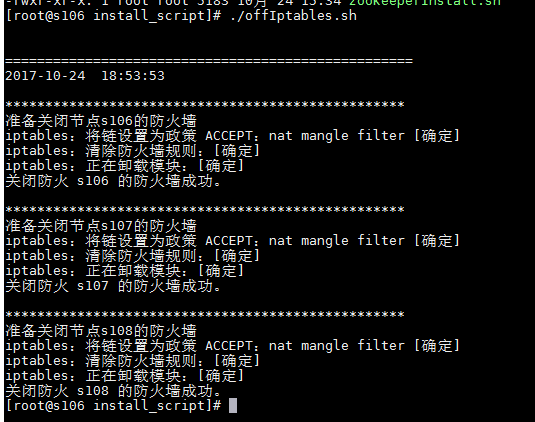


四．关闭防火墙

（1）脚本所在目录

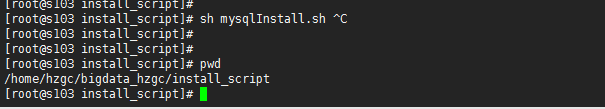


（2）脚本执行的结果

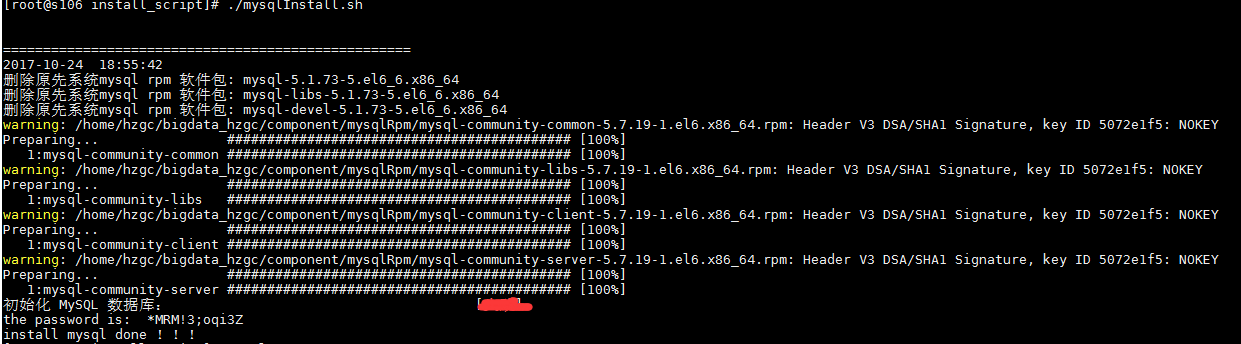


五．安装mysql（在安装包所在节点上安装）

（1）脚本所在目录



1. 脚本执行结果



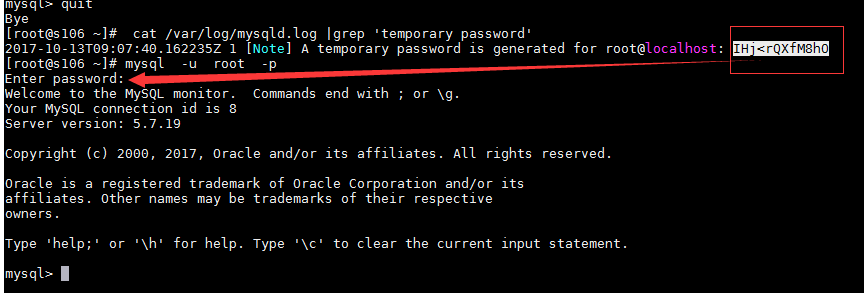
1. 启动mysql服务



（4）修改默认密码为自定义密码(密码获取如图上。)

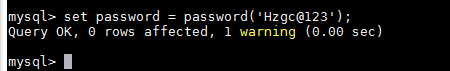
登陆msyql,执行以下命令并将文件中的密码输入

mysql -u root -p



修改密码

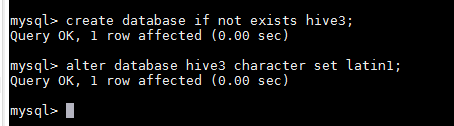
mysql> SET PASSWORD = PASSWORD('Hzgc@123');



5.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;



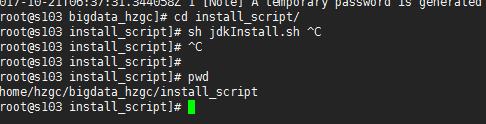
6.授权远程连接

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

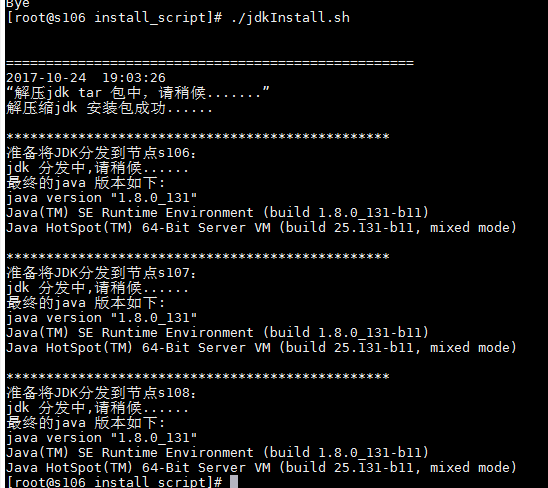
mysql> flush privileges;

六．安装jdk

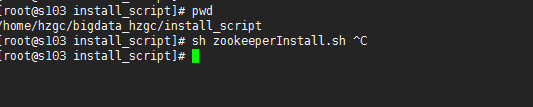
（1）脚本路径



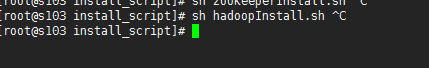
脚本执行结果



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

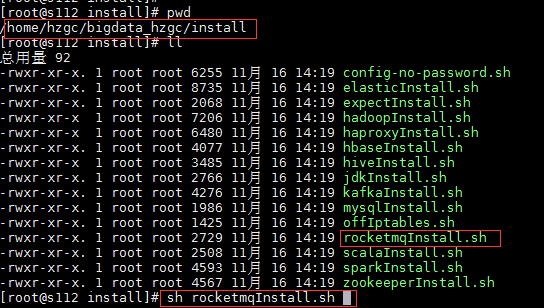
十一．安装scala

十二．安装kafka

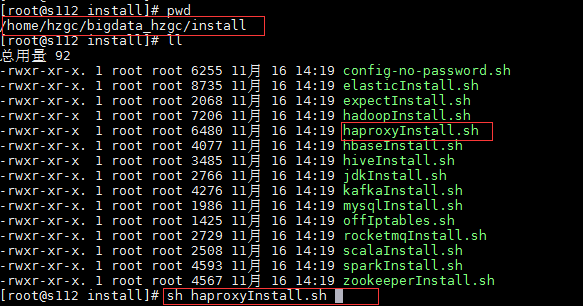
十三．安装spark



1. 安装roketmq



1. 安装Haproxy



# 十六．安装ElasticSearch

在



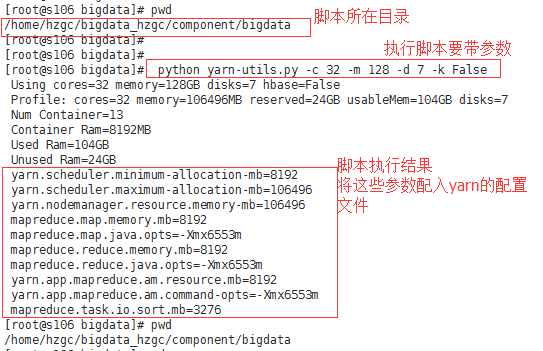
# 十七．拷贝环境变量以及相关环境变量到各个目录

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：**

1. 进入集群安装包目录/home/hzgc/bigdata\_hzgc/component/
2. 执行sh yarn-utils.py：执行时带上参数，e.x：python yarn-utils.py -c 32 -m 128 -d 7 -k False

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）

1. 进入Hadoop安装目录



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

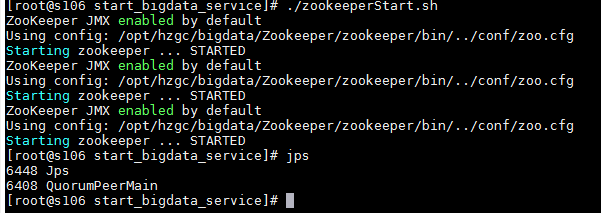


说明：yarn-utils.py这是一个根据输入的参数，自动计算yarn应该如何分配内存和cpu的脚本，算出结果后，将脚本所得结果配入yarn的配置文件。

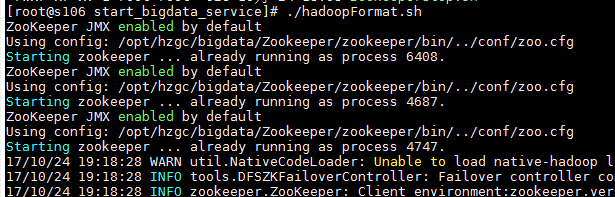
# 二十．启动服务

1. **启动zookeeper**

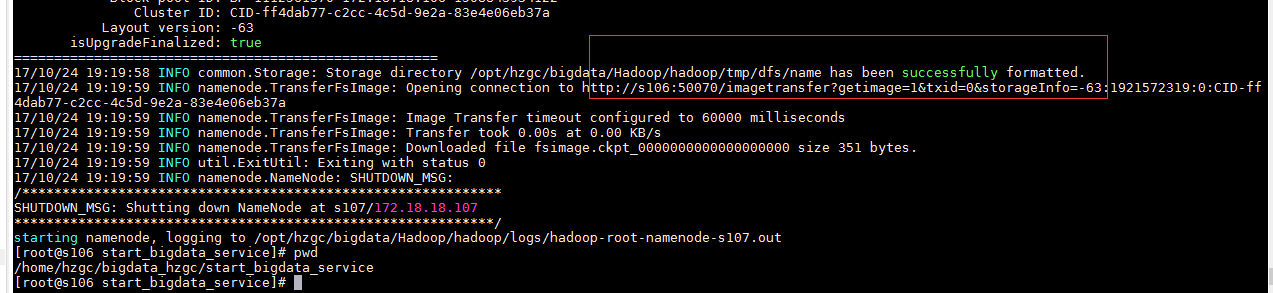
在s153主机上执行

**2. 启动hadoop**

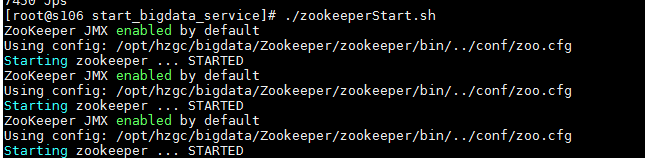
1. 第一次启动hadoop请按以下顺序执行



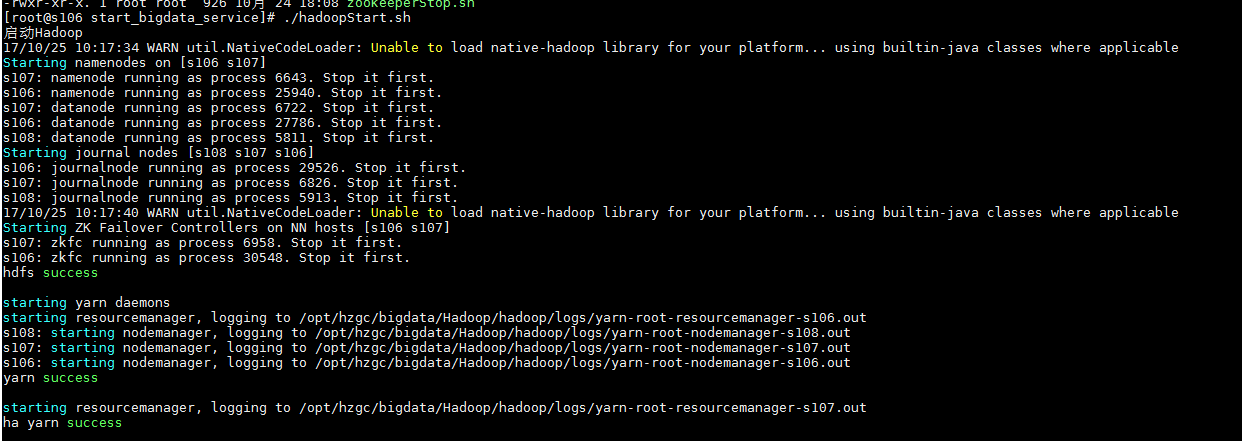
（2）脚本执行结果



（3）启动zookeeper

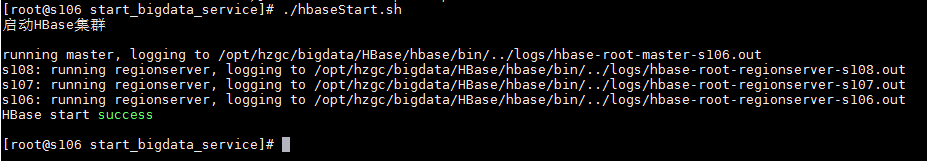


1. 启动hadoop



1. **启动hbase**

在s153主机上执行



1. **Hive bin目录下执行**

sh schematool -initSchema -dbType mysql

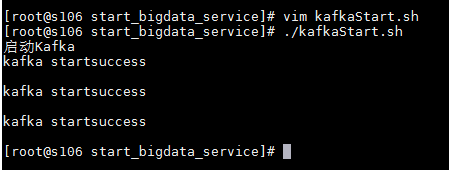
1. **启动hive（启动后每个节点有两个RunJar 进程）**

（1）在s153主机上执行，启动hiveserver2 metastore

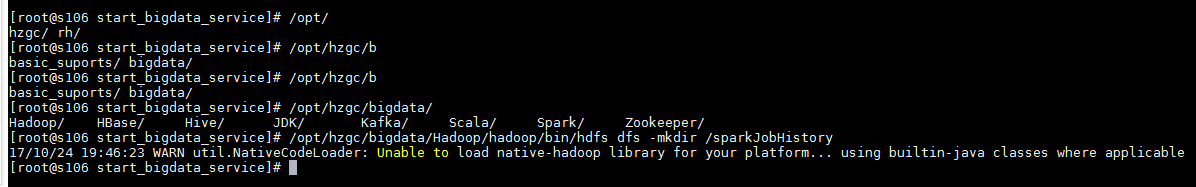


1. **启动kafka**

在s153主机上执行



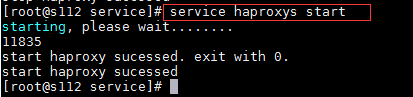
1. **使用spark jdbc**
2. 第一次使用spark请先在hdfs创建sparkJobHistory



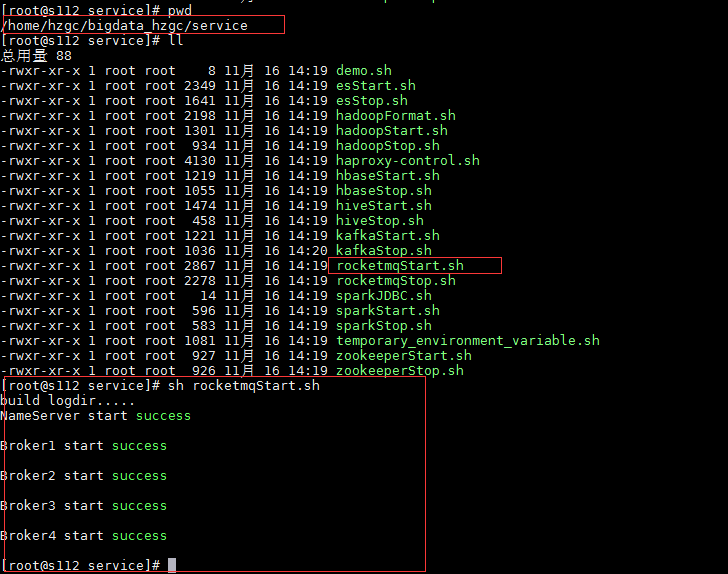
1. 启动thriftserver服务

sh /opt/hzgc/bigdata/Spark/spark/sbin/start-thriftserver.sh 1G 1G 1 1 1

1. **启动Haproxy**



1. **启动rocketmq**

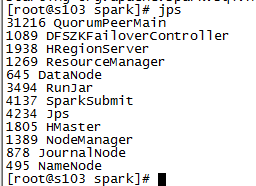


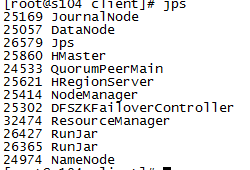
**10. 启动elasticsearch**

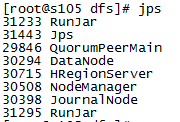
在service目录下，执行sh esStart.sh



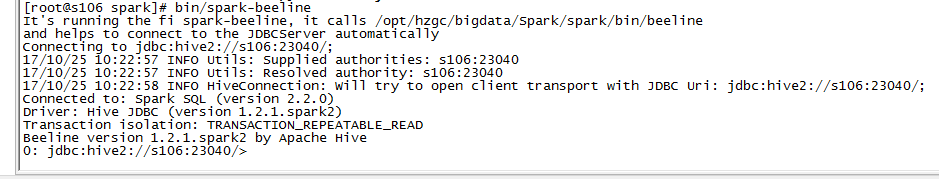
**11. 服务验证（每台主机所起的服务）**







1. 连接jdbc

 **常见安装问题**

## 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

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