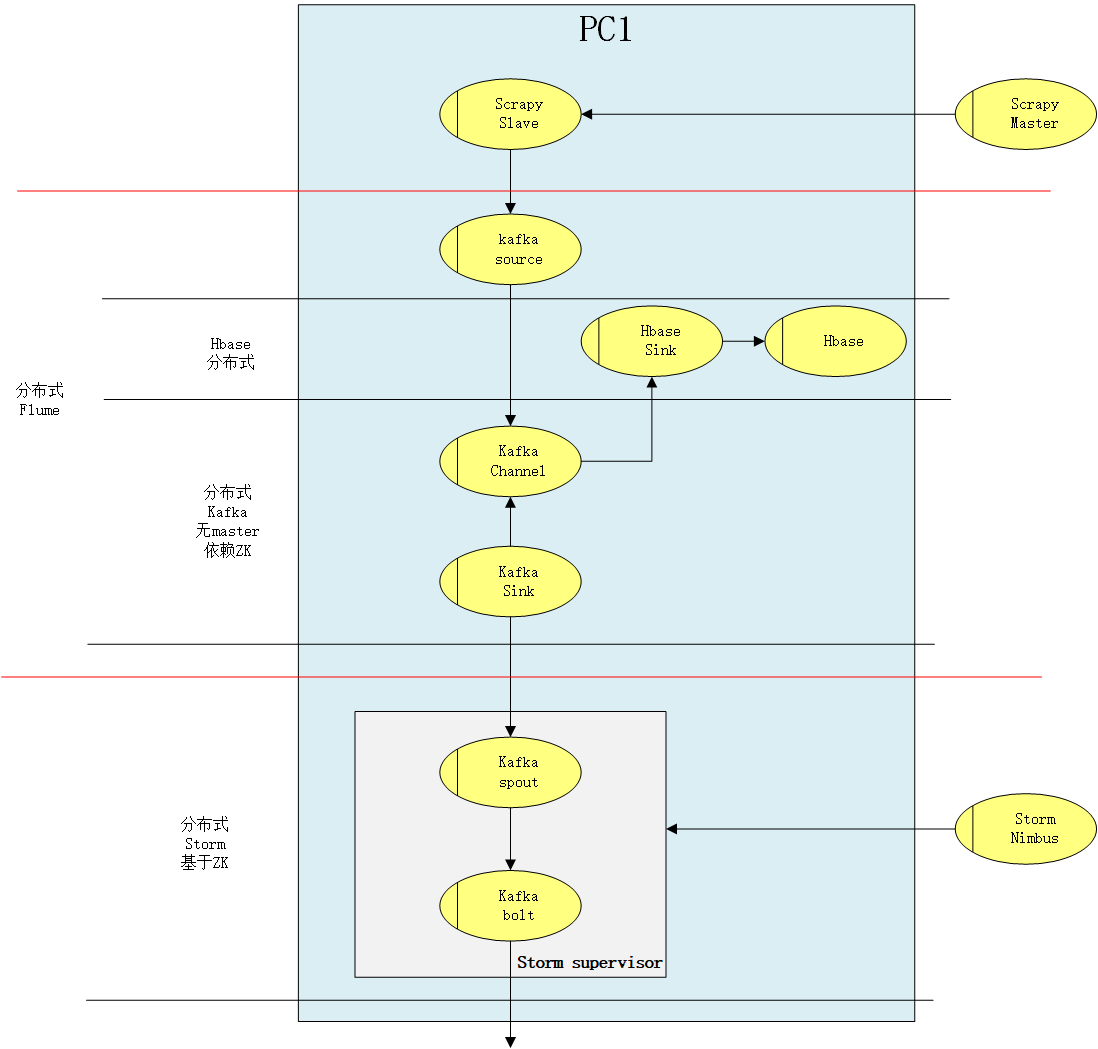
最终框架：



我们使用5台服务器来实现这个分布式爬虫框架，这5台服务器系统为centos6.4 x64。首先我们进行了集群的基础配置，即进行时间的同步，关闭防火墙，修改host与network，之后配置了ssh免密登录。

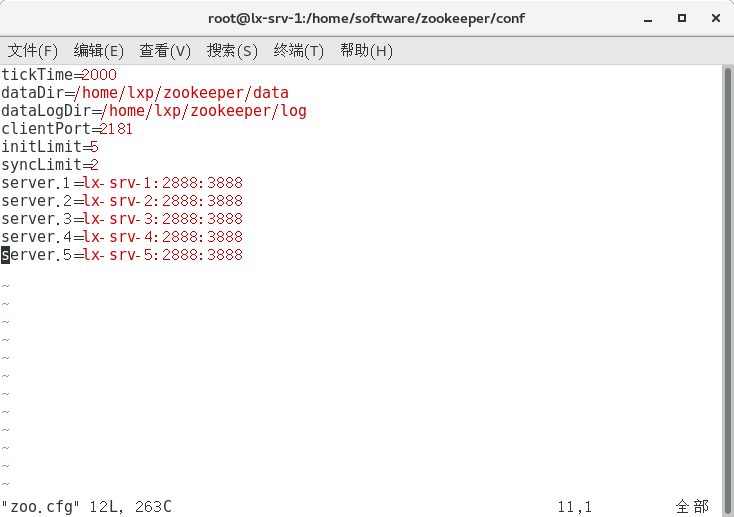
我们没有使用支持最新版jdk1.9，而是使用了jdk1.8，因为jdk1.8已经成熟，jdk1.9可能会出现一些兼容问题。我们卸载了系统自带的openjdk1.7，然后安装jdk1.8，并修改了符号变量。对每台机器进行了环境变量的配置。如下是一个的示例



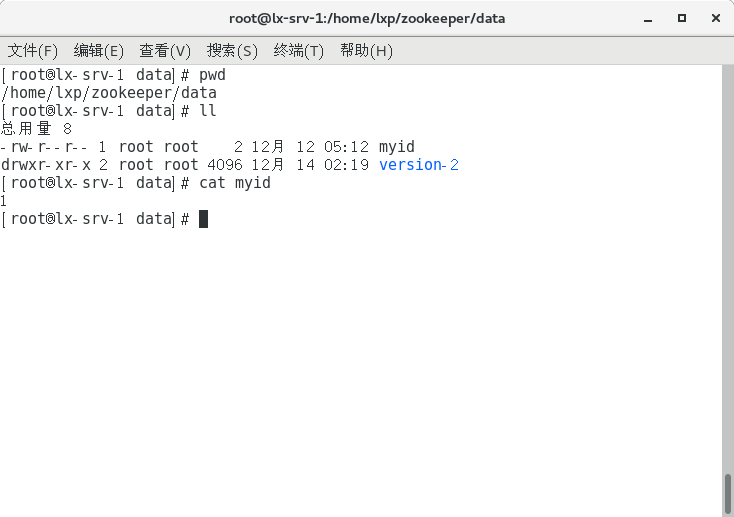
# Zookeeper:

首先安装zookeeper，并进行配置。

$ZK\_HOME/conf/zoo.cfg，对集群内每个节点进行相同的配置。

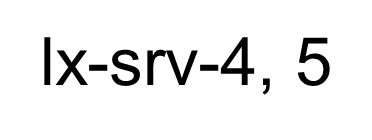


然后在每台机器的dataDir中创建myid文件，内容为每个zookeeper的uuid。



配置完毕，我们试着运行zookeeper：

在每台机器上执行 $zkServer.sh start，全部启动后，执行$zkServer.sh status



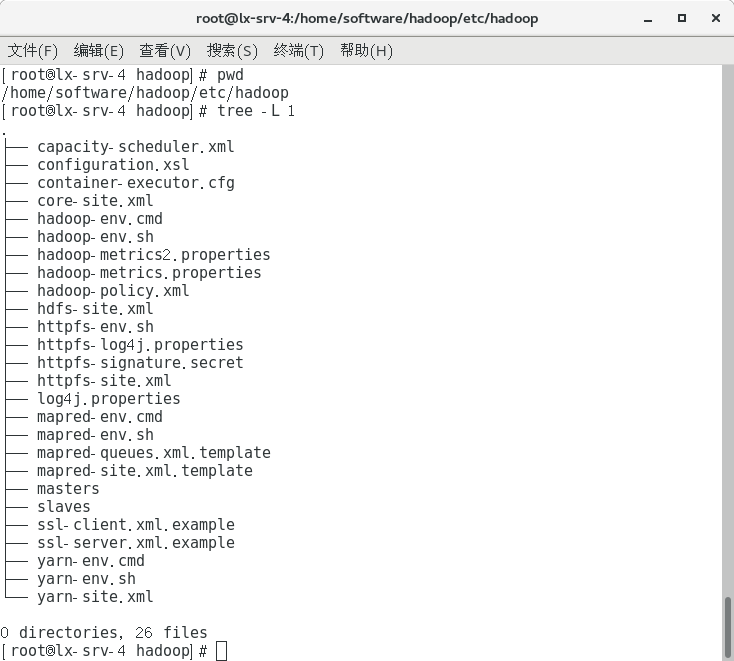
说明zookeeper运行成功。

若没有按照配置全部启动zookeeper则会报错。根据输出，我们能看到其中一个节点被选举为了leader，其余节点为follower。

# Hadoop Hbase

然后我们开始配置hadoop，我们以lx-svr-1为hadoop master节点，lx-srv-4, 5为hadoop slave节点。经过惨痛的教训（无数次失败），我们把storm、hbase等等组件的jar包复制到hadoop目录下面，并对冲突包进行删除（主要是log4j.jar）。如果没有复制，就会发生莫名其妙的错误。

我们开始修改配置文件。



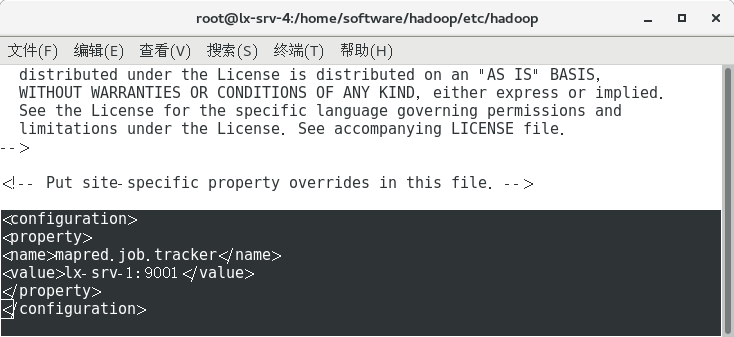
Lx-srv-4(1,5)

$HADOOP\_HOME/etc/hadoop/core-site.xml



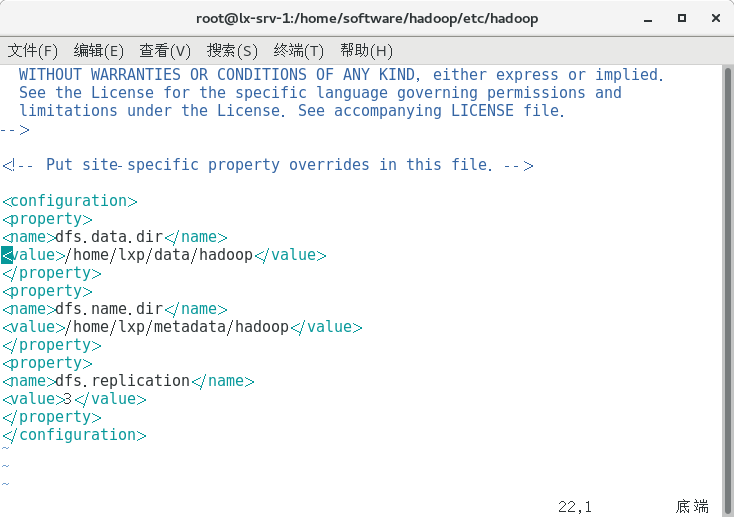
Lx-srv-4(1,5)

$HADOOP\_HOME/etc/hadoop/mapred-site.xml.template



Lx-srv-4(1,5)

$HADOOP\_HOME/etc/hadoop/hdfs-site.xml



Lx-srv-4(1,5)

cat $HADOOP\_HOME/etc/hadoop/masters

lx-srv-1

cat $HADOOP\_HOME/etc/hadoop/slaves

lx-srv-4

lx-srv-5

按照设想，lx-svr-1为hadoop的master节点，所以这个节点的hbase并不存储爬虫数据，只存储元数据。并且我们取消hbase使用自带的zk进行托管。

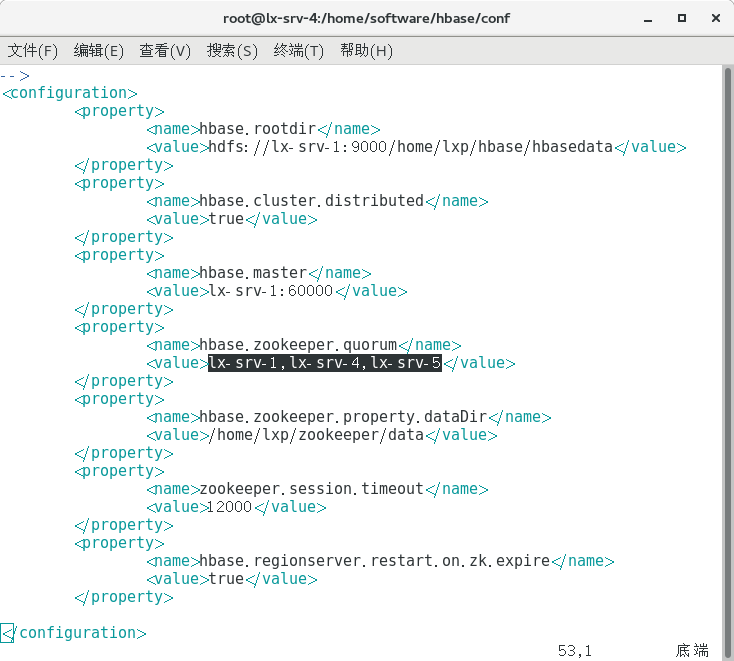
Lx-srv-4(1,5)

cat $HBASE\_HOME/conf/hbase-env.sh

export HBASE\_MANAGES\_ZK=false

Lx-srv-4(1,5)

cat $HBASE\_HOME/conf/hbase-site.xml



Lx-srv-4(1,5)

cat $HBASE\_HOME/conf/regionservers

lx-srv-4

lx-srv-5

实际运行中我们遇到了很多问题，首先是需要按照配置去mkdir相应的文件夹。如果文件夹不存在，就会报错。

另外当各台服务器的时间相差比较大的时候，就会导致一些进程启动失败。所以需要不断同步时间。还有一个错误为ERROR: Can't get master address from ZooKeeper; znode data == null，这是因为没有给hbase和hadoop的文件赋予权限，导致每次重新启动后，都会查询失败，必须要使用zkCli.sh进入zk后台删除原有数据。

如果对配置进行修改，则需要删除原有的文件夹，重新进行format。

Lx-srv-1(4, 5)

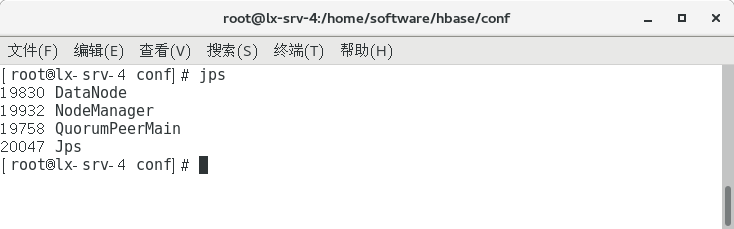
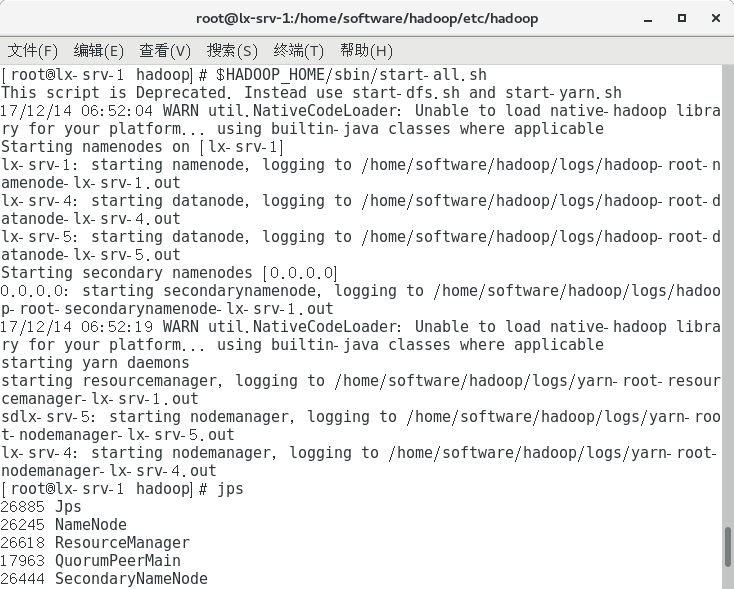
hdfs namenode -format

Hadoop master(lx-srv-1)

$HADOOP\_HOME/sbin/start-all.sh

成功后lx-srv-1下有resourceManager和SecondaryNameNode和NameNode

lx-srv-4, 5下有 DataNode和NodeManager

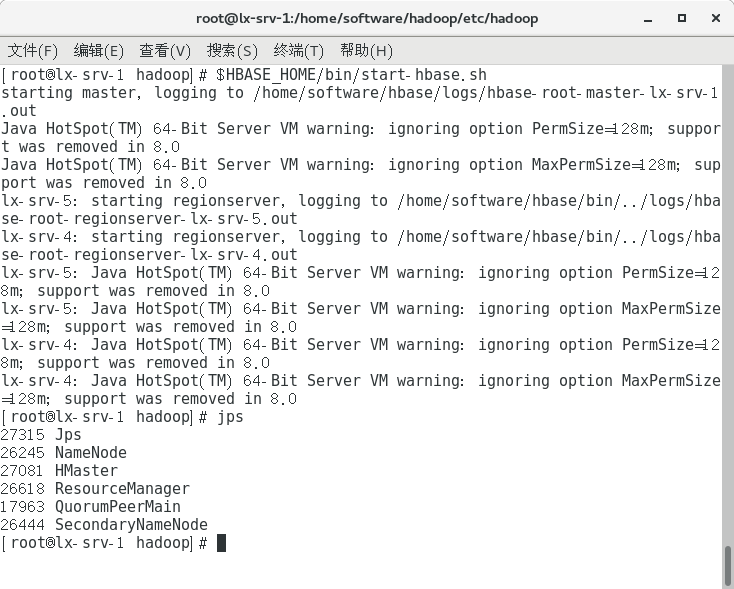


Hbase master(lx-srv-1)

$HBASE\_HOME/bin/start-hbase.sh

成功后lx-srv-1下有HMaster

lx-srv-4, 5下有 HRegionServer

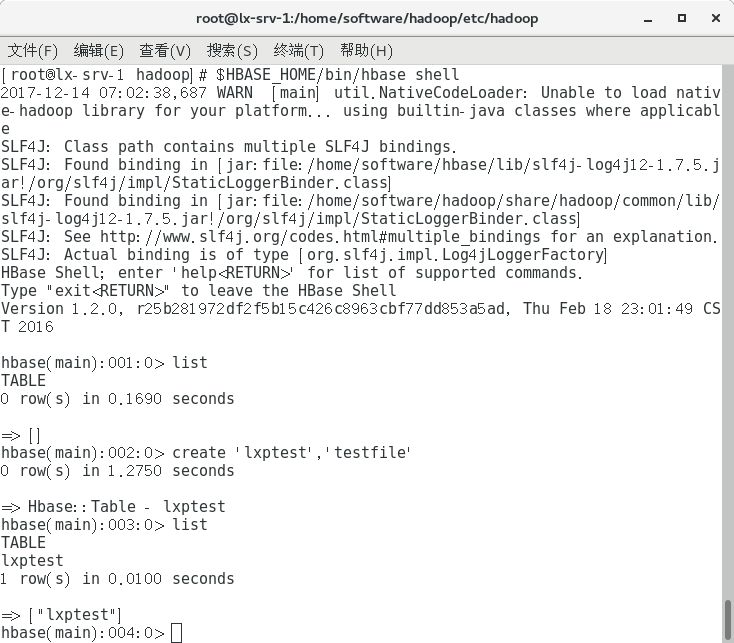




然后启动hbase命令行

$HBASE\_HOME/bin/hbase shell

第一次启动时需要 在hbase master节点里create 'lxptest','testfile'（这里与flume中的配置相同）



早期草稿

本周我们没有先做爬虫，我们做了比较简单的hbase，hadoop，zookeeper和flume， 我们测试了flume的很多种配置

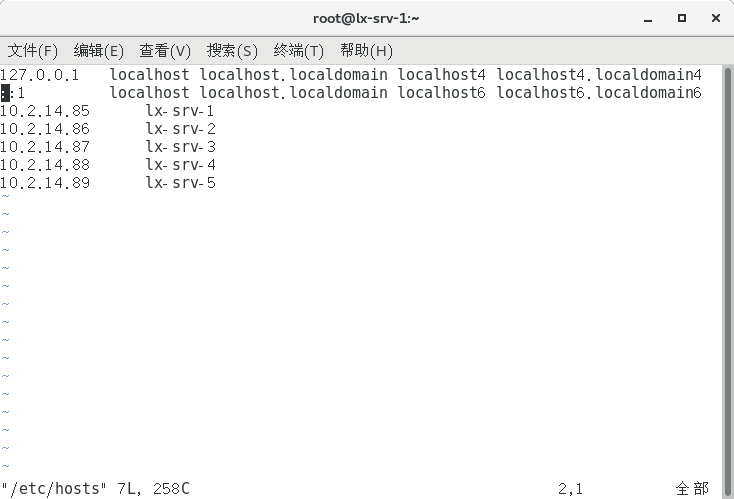
初始调研：另附文档；可能晚于周报

预先准备：

1. 同步时间

时间不一致可能会报region is not online错误

1. 设置host



1. 设置network
2. 关闭防火墙

避免服务器之间的通讯bug问题

1. Ssh免密登录
2. 修改环境变量



还有修改java符号链接，因为linux系统自带open-jdk-1.7版本，然后我们配置需要的是一个1.8的版本，但我们安装后没有改链接，导致第一顺位还是系统自带的jdk

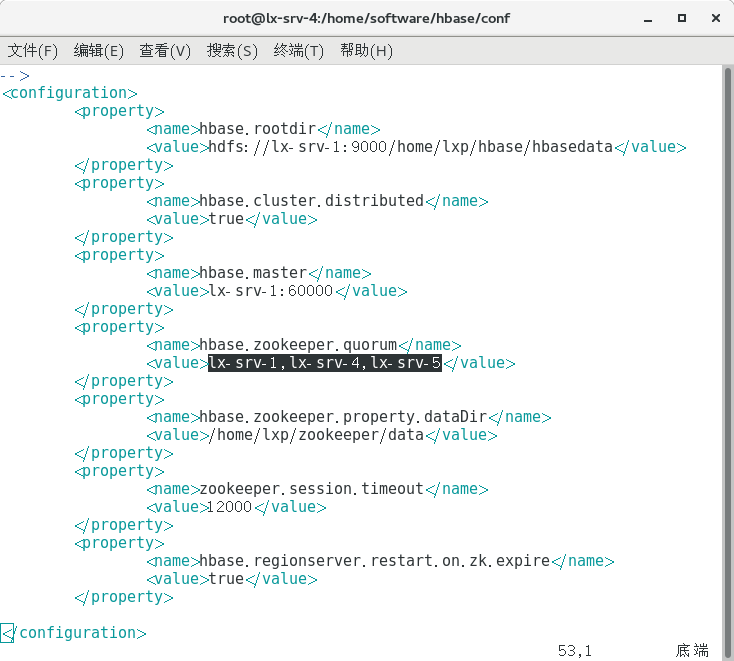
Hbase配置：

Lx-srv-4(1,5)

cat $HBASE\_HOME/conf/hbase-env.sh

export HBASE\_MANAGES\_ZK=false

注 是否托管zookeeper



cat $HBASE\_HOME/conf/hbase-site.xml

注 黑色部分可能有错(lx-srv-1只保存元数据)

Lx-srv-4(1,5)

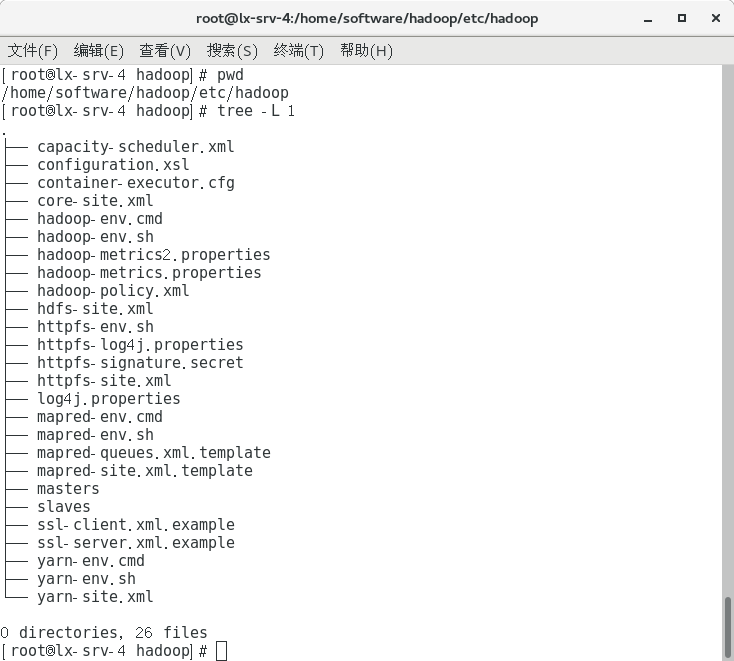
cat $HBASE\_HOME/conf/regionservers

Lx-srv-4

Lx-srv-5

注 这里是实际保存数据的节点

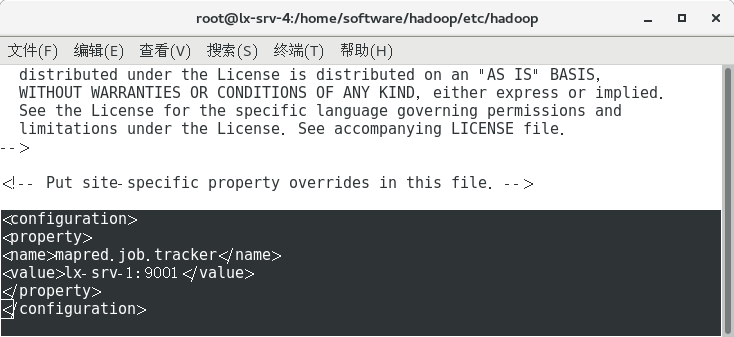
Hadoop配置：

 Lx-srv-4(1,5)

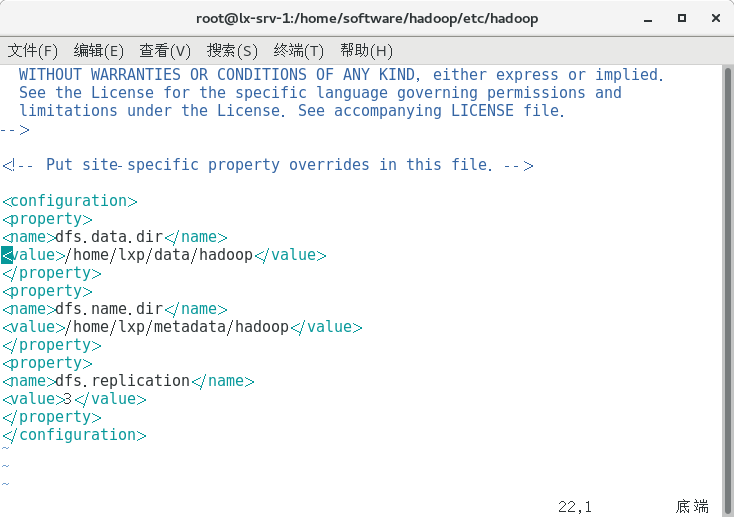
$HADOOP\_HOME/etc/hadoop/

 Lx-srv-4(1,5)

$HADOOP\_HOME/etc/hadoop/core-site.xml

 Lx-srv-4(1,5)

$HADOOP\_HOME/etc/hadoop/mapred-site.xml.template

 Lx-srv-4(1,5)

$HADOOP\_HOME/etc/hadoop/hdfs-site.xml

Lx-srv-4(1,5)

cat $HADOOP\_HOME/etc/hadoop/masters

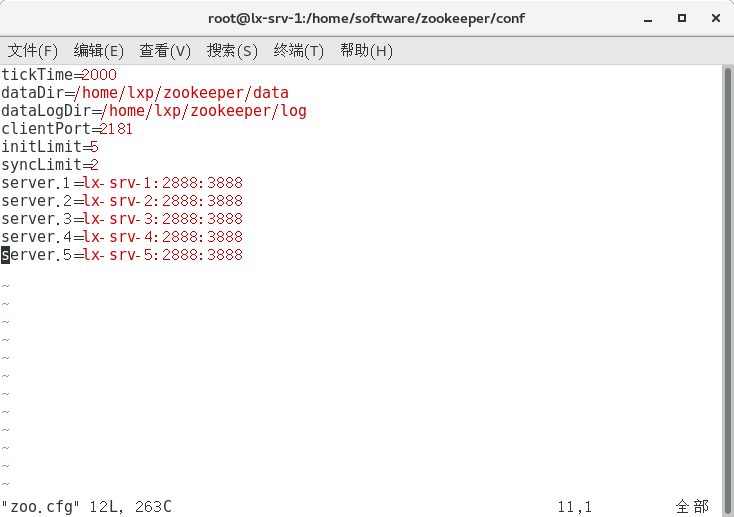
lx-srv-1

cat $HADOOP\_HOME/etc/hadoop/slaves

Lx-srv-4

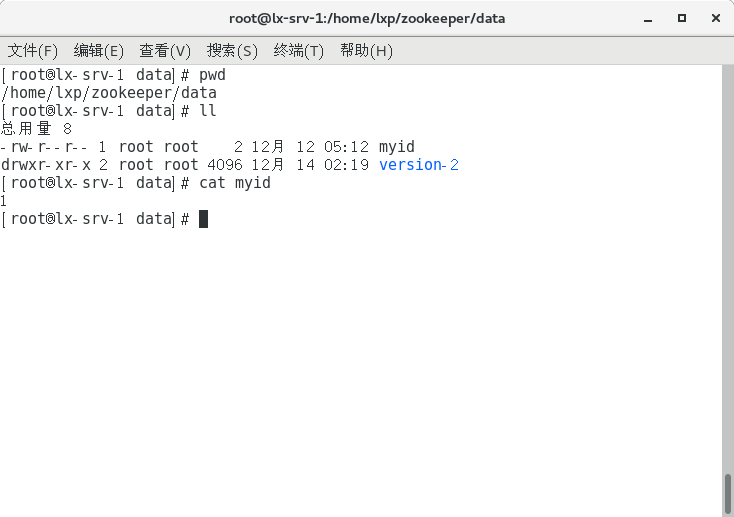
lx-srv-5

Zookeeper配置：

 lx-srv-1

$ZK\_HOME/conf/zoo.cfg

所有的5个节点配置相同

 lx-srv-1 /home/lxp/zookeeper/data

需要在上一张ppt的配置文件夹中新建myid, 内容为zookeeper的唯一的id, lx-srv-2,3,4,5分别为2,3,4,5