1.配置虚拟机

Net8

宿主机自动获取

配置init3启动 vim /etc/inittab

修改主机名 vim /etc/sysconfig/network

修改IP vim /etc/sysconfig/network-scripts/ifcfg-eth0

修改主机名和IP的映射关系 vi /etc/hosts

关闭防火墙

查看防护墙状态 service iptables status

关闭 service iptables stop

查看防火墙开机启动状态 chkconfig iptables --list

关闭开机启动 chkconfig iptables off

ssh登录

Jdk安装

配置hadoop

解压

配置伪分布式

1）Hadoop-evn.sh

修改JAVA\_HOME 默认的有时候会获取不到

2）core-site.xml

<!—指定hdfs的namenode的通信地址-->

<name> fs.defaultFS</name>

<value>hdfs://主机名:9000</value>

<!-- 指定hadoop运行时产生文件的存放目录 -->

<name>hadoop.tmp.dir</name>

<value>/…. /hadoop-2.4.1/tmp</value>

3）hdfs-site.xml

<!-- 配置HDFS副本的数量 -->

<name>dfs.replication</name>

<value>1</value> //由于一台机器，如果崩了就全崩，故设1。正常搭建时默认3。

<!—配置原信息和数据块的存储地址-->

<name>dfs.namenode.name.dir</name>

<value>/home/hadoop/app/hadoop-2.4.1/data/name</value>

<name>dfs.datanode.data.dir</name>

<value>/home/hadoop/app/hadoop-2.4.1/data/data</value>

4）mapred-site.xml //需修改mapred-site.xml .template

<-- 配置mapreduce运行程序放到哪个资源调度集群去跑-->

<name>mapreduce.framework.name</name>

<value>yarn</value>

{

<property>

<name>mapreduce.admin.user.env</name>

<value>HADOOP\_MAPRED\_HOME=$HADOOP\_COMMON\_HOME</value>

</property>

<property>

<name>yarn.app.mapreduce.am.env</name>

<value>HADOOP\_MAPRED\_HOME=$HADOOP\_COMMON\_HOME</value>

</property>

}

5）yarn-site.xml

<--将yarn. resourcemanager配置到主节点-->

<name>yarn.resourcemanager.hostname</name> //yarn集群里的老大

<value>hadoop-0</value>

<--配置yarn 集群里的从节点-->

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value> //mapreduce里面map产生的中间结果通过shuffle机制传递给reduce

Hadoop加入环境变量中

export HADOOP\_HOME=/…../hadoop-2.4.1

export PATH=…:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin

格式化

Hadoop namenode –format //不要重复执行，最好执行一次

/\*

注：

之前发生过WARN hdfs.DFSUtil: Namenode for null remains unresolved for ID null. Check your hdfs-site.xml file to ensure namenodes are configured properly.

和WARN net.DNS: Unable to determine address of the host-falling back to "localhost" address

原因：之前虚拟机与ip映射没处理好。

\*/

如果成功倒数第五行 status 0

开启hdfs start-dfs.sh

//会读取hadoop/slaves

开启yarn start-yarn.sh

Jps检查是否启动成功

Hdfs提供web服务的访问

http://192.168.2.100:50070 //默认此端口

上传文件进hdfs

hadoop fs –put文件地址 hdfs://主机名:9000/ //传到根目录

从hdfs下载文件

hadoop fs –get 文件地址

删除hadoop下面的文件

hadoop dfs -rmr abc

hadoop查看文件目录命令

bin/hadoop dfs -ls

hadoop查看文件命令

bin/hadoop dfs –lsr

查看hdfs基本统计信息 如同在浏览器下面看基本信息一样

bin/hadoop dfsadmin -report

hadoop安全模式

进入安全模式

hadoop dfsadmin -safemode enter

退出安全模式

hadoop dfsadmin -safemode leave

测试MapReduce

/../hadoop-2.4.1/share/mapreduce //mapreduce的jar包都在这，包括自带的例子程序

运行例子jar包

hadoop jar jar包名 类名 参数

hdfs的实现思想：

1）hdfs是通过分布式集群来存储文件的，为客户端提供了一个便捷的访问方式，就是一个虚拟的目录结构。

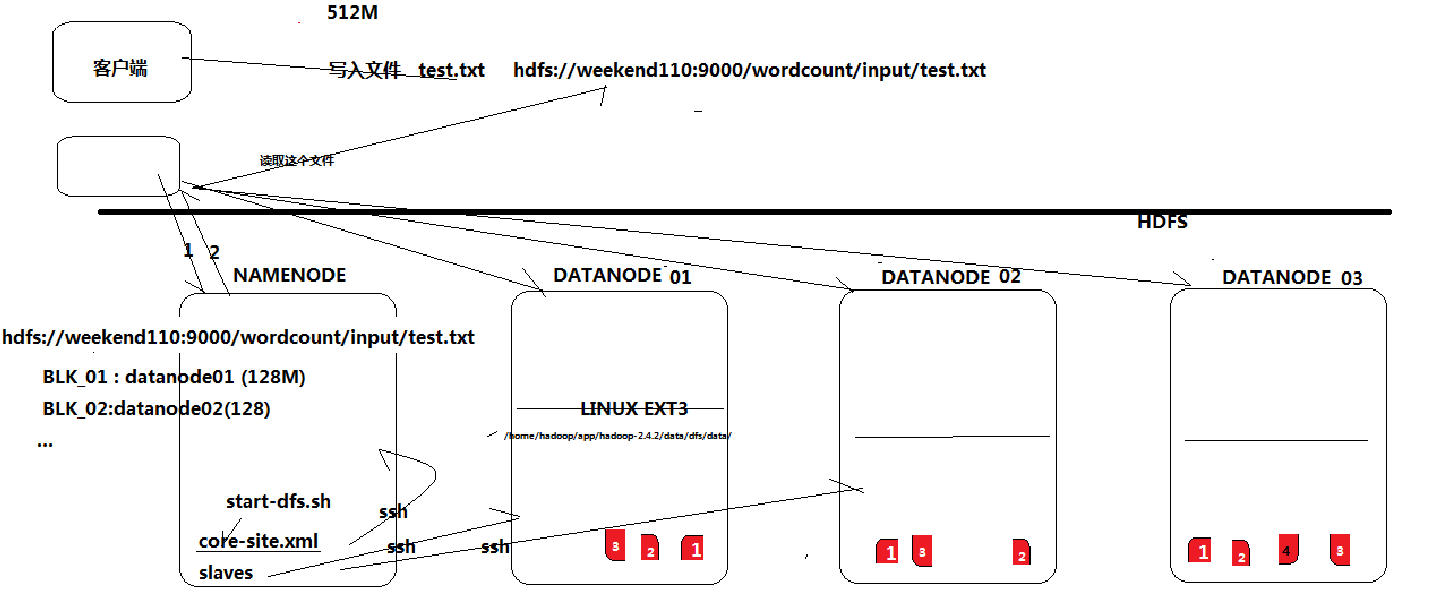
2）文件存储到hdfs集群中去的时候是被切分成block

3）文件的block存放在若干台datanode节点上

4）hdfs文件系统中的文件与真实的block之间有映射关系，由namenode管理

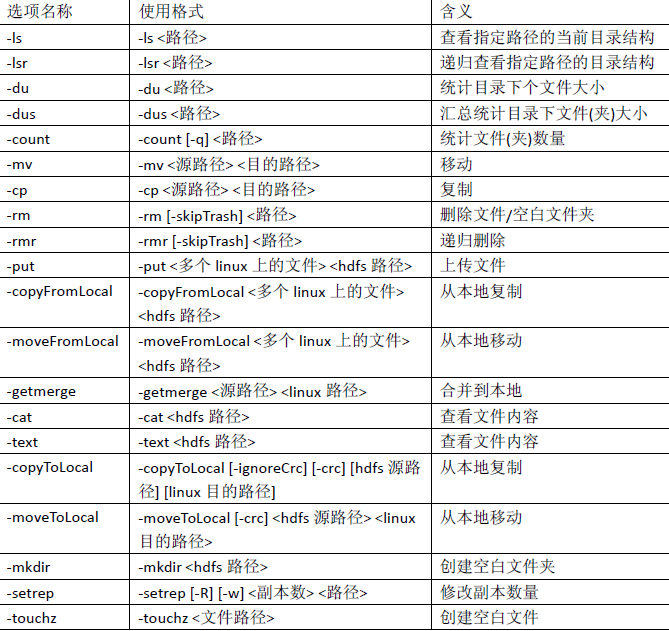
5）每一个block在集群中会存储多个副本，好处是可以提高数据的可靠性，还可以提高访问的吞吐量/并发能力

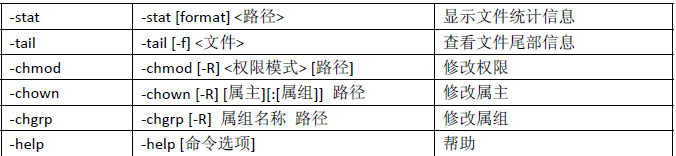
hsfs实现机制



常用HDFS Shell操作

hadoop fs 操作文件（filesystem）





\*\*\*hadoop fs -du -s -h hdfs://hadoop-0:9000/\*

hadoop jar 操作 jar包

hadoop权限检查很弱，chown 设置所属用户时，不检测用户存不存在。

2.创建集群

克隆虚机

管理/克隆/当前状态/完整克隆

修改克隆完成机器的ip和主机名

vim /etc/udev/rules.d/70-persistent-net.rules

物理地址信息对比后，选择对应项目修改eth1—>eth0

编辑网卡配置

vim /etc/sysconfig/network-scripts/ifcfg-eth0

修改物理地址和ip地址

修改主机名

vim /etc/sysconfig/network

hosts里添加映射

vim /etc/hosts

授权 搭建无秘登录集群

cd ~/.shh

生成公钥 ssh-keygen -t rsa (之后全回车)

将公钥传输到hadoop-1上 scp id\_rsa.pub hadoop-1:/home/hadoop

在hadoop-1上将公钥加载到授权列表文件 cd /.ssh

如果没有执行 touch authorized\_keys

修改权限为600 chmod 600 authorized\_keys

将hadoop-0的公钥添加到hadoop-1的授权列表中

cat ../id\_rsa.pub >> ./authorized\_keys

3.