Hadoop 环境变量

更改防火墙设置

查看centos7防火墙状systemctl status firewalld.service  
关闭centos7的防火墙systemctl stop firewalld.service

开启防火墙service firewalld start

永久关闭防火墙 Systemctl disable firewalld.service

查看centos版本cat /etc/redhat-release

修改ip地址

找到文件vi /etc/sysconfig/network-scripts/ifcfg-ens33进行修改：

vi /etc/sysconfig/network-scripts/ifcfg-ens33

NET模式下的配置:

DEVICE=ens33 （网络名字）

TYPE=Ethernet （网络类型）

ONBOOT=yes （开机自启）

BOOTPROTO=static (开启手动分配方式 dhcp表示动态 none不启用自己分配ip地址)

IPADDR=192.168.135.11 （自己定义）(ip地址 自己添加 注意网段 ip地址不要冲突)

PREFIX=255.255.255.0 （子网掩码）

GATEWAY=192.168.135.2 （自己定义）（网关 路由器 编辑-虚拟网络编辑器-NAT设置-查看网关）

DNS1=192.168.135.2 （自己定义）（域名解析器）

DNS2=8.8.8.8 (公共的域名解析器)

DEVICE=ens33

TYPE=Ethernet

ONBOOT=yes

BOOTPROTO=static

IPADDR=192.168.239.8

PREFIX=255.255.255.0

GATEWAY=192.168.239.2

DNS1=192.168.239.2

DNS2=8.8.8.8

重启网络服务systemctl restart network

查看ip地址 ip address 或者ip addr

ping自己的IP地址看看是否可以使用

修改主机名 vi /etc/hostname

修改完主机名 重启虚拟机 reboot 重启虚拟机

给ssh映射配置文件 /etc/hosts

192.168.135.21 had1709

192.168.135.22 had1709A

192.168.135.23 had1709B

连接虚拟机，实现虚拟机之间可以免密码登录

要求：三台虚拟机

三台机器每一台都要执行一遍以下内容

设置免密码ssh

ssh-keygen -t rsa 生成公钥和私钥

ssh-copy-id 主机名或主机IP地址 用私钥连接自己的公钥

ssh-copy-id 从机1名或IP地址 用私钥连接从机的公钥

ssh-copy-id 从机2名或IP地址 用私钥连接从机的公钥

上传hadoop和jdk的压缩包并解压hadoop和jdk的压缩包

tar -zxvf 文件名

修改名字 mv 原名 修改名

配置jdk环境变量 vi /etc/profile

export JAVA\_HOME=/home/bigdate/jdk

export HADOOP\_HOME=/home/bigdate/hadoop

export CLASSPATH=.:$JAVA\_HOME/lib/dt.jar:$JAVA\_HOME/lib/tools.jar:

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

使环境变量生效 source /etc/profile

查看是否安装成功 Java -version

配置hadoop环境变量 cd /home/bigdate/hadoop/etc/hadoop

要求配置六个文件

hadoop-env.sh

core-site.xml

hdfs-site.xml

mapred-site.xml(提示 mv mapred-site.xml.template mapred-site.xml )

yarn-site.xml

slaves

文件一：hadoop-env.sh(配置java的地址)

export JAVA\_HOME=/home/bigdate/jdk

文件二：core-site.xml(配置hadoop comm文件)

<property>

<name>fs.defaultFS</name>

<value>hdfs://主机名:8020</value>

</property>

<property>

<name>io.file.buffer.size</name>

<value>4096</value>

</property>

<property>

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

<value>/home/bigdate/tmp</value>

</property>

文件三：hdfs-site.xml（配置虚拟hdfs库）

<property>

<name>dfs.replication</name>

<value>3</value>

</property>

<property>

<name>dfs.block.size</name>

<value>134217728</value>

</property>

<property>

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

<value>file:///home/hadoopdata/dfs/name</value>

</property>

<property>

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

<value>file:///home/hadoopdata/dfs/data</value>

</property>

<property>

<name>fs.checkpoint.dir</name>

<value>file:///home/hadoopdata/checkpoint/dfs/cname</value>

</property>

<property>

<name>fs.checkpoint.edits.dir</name>

<value>file:///home/hadoopdata/checkpoint/dfs/cname</value>

</property>

<property>

<name>dfs.http.address</name>

<value>主机名:50070</value>

</property>

<property>

<name>dfs.secondary.http.address</name>

<value>从机名1:50090</value>

</property>

<property>

<name>dfs.webhdfs.enabled</name>

<value>true</value>

</property>

<property>

<name>dfs.permissions</name>

<value>false</value>

</property>

文件四：（配置mapreduce）mapred-site.xml(提示 mv mapred-site.xml.template mapred-site.xml )

<property>

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

<value>yarn</value>

<final>true</final>

</property>

<property>

<name>mapreduce.jobhistory.address</name>

<value>主机名:10020</value>

</property>

<property>

<name>mapreduce.jobhistory.webapp.address</name>

<value>主机名:19888</value>

</property>

文件五：yarn-site.xml(配置yarn服务)

<property>

<name>yarn.resourcemanager.hostname</name>

<value>主机名</value>

</property>

<property>

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

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.resourcemanager.address</name>

<value>主机名:8032</value>

</property>

<property>

<name>yarn.resourcemanager.scheduler.address</name>

<value>主机名:8030</value>

</property>

<property>

<name>yarn.resourcemanager.resource-tracker.address</name>

<value>主机名:8031</value>

</property>

<property>

<name>yarn.resourcemanager.admin.address</name>

<value>主机名:8033</value>

</property>

<property>

<name>yarn.resourcemanager.webapp.address</name>

<value>主机名:8088</value>

</property>

文件六：slaves（添加3个主机名字）

vi /home/bigdate/hadoop/etc/hadoop/slaves

主机

从机1

从机2

配置完成

发送给其他从机 scp -r 文件地址 从机名:从机文件地址

scp -r /home/bigdate hadoop102:/home/

环境变量也要发送

scp -r /etc/profile hadoop102:/etc/profile

是环境变量生效

source /etc/profile

启动集群

启动集群之前格式化 hadoop namenode -format

如果没有出现错误就开启集群

启动集群 start-all.sh

查看集群节点 Jps（每一台都要查看）

出现以下内容方可完成集群搭建

[root@hadoop101 ~]# jps

10640 Jps

9911 DataNode

9818 NameNode

10268 NodeManager

10173 ResourceManager

[root@hadoop102 ~]# jps

9539 DataNode

9620 SecondaryNameNode

9685 NodeManager

9781 Jps

[root@hadoop103 ~]# jps

17393 NodeManager

17489 Jps

17308 DataNode

备注：要开启集群后才能用浏览器查看和连接，但是要把虚拟机内部的防火墙关了

http://ip号:50070

1.描述HDFS的服务及功能

(1)、HDFS的服务包括：NameNode、DataNode、Secondary NameNode。

(2)、功能：

a、NameNode是主节点，存储文件的元数据如文件名，文件目录结构，

文件属性（生成时间,副本数,文件权限），以及每个文件的块列表和块所

在DataNode等。

b、DataNode 在本地文件系统存储文件块数据，以及块数据的校验和。

c、Secondary NameNode 用来监控HDFS状态的辅助后台程序，每隔一段时间获取HDFS 元数据的快照。

2.解释YARN 的服务功能

(1).ResourceManager

处理客户端请求

启动/监控ApplicationMaster

监控NodeManager

资源分配与调度

(2).NodeManager

单个节点上的资源管理

处理来自ResourceManager的命令

处理来自ApplicationMaster的命令

(3).ApplicationMaster

数据切分

为应用程序申请资源，并分配给内部任务

任务监控与容错

(4).Container

对任务运行环境的抽象，封装了CPU、内存等多维资源以及环境变量、

启动命令等任务运行相关的信息

Client端读取HDFS数据过程分析

1. 客户端或者用户通过调用 FileSystem 对象的 open()方法打开需要读取的文件。

2. FileSystem 通过远程协议调用 NameNode 确定文件的前几个 Block 的位置。

3. NameNode 向客户端返回一个包含数据信息的地址，客户端根据地址创建一个FSDataInputStream 开始对数据进行读取。

4. FSDataInputStream 根据开始时存放的前几个 Blocks 的 DataNode 的地址，连接到最近的 DataNode 上对数据开始从头读取。

5.当读到 Block 的结尾的时候， FSDataInputStream 会关闭到当前 DataNode 的链接，然后查找能够读取下一个 Block 的最好的 DataNode。

6. 读取完成调用 close()方法，关闭 FSDataInputStream。