hadoop环境一文通

土豪选择cdh、bitnamihadoop、hdp等集成环境，直接导入虚拟机。

手动分布式配置需要准备jdk、hadoop。

## hadoop3完整安装

宿主机win1064

虚拟工具vmware14

虚拟机debian9.3 x32

winscp访问linux文件，mobaxerm访问ssh

编辑器：everdit，编码utf无bom

用户名root 密码passwd

1.虚拟环境配置

一主三从，主机nat联网，1g内存；从机仅主机模式，0.5g内存。

master 192.168.111.111 //网卡1：nat联网

slave1 192.168.112.112 //网卡2：仅与主机连接

slave2 192.168.113.113 //网卡3：仅与主机连接

本地建立共享文件夹D:\sharing，映射为各个虚拟机/mnt/hgfs/sharing/

//mount -t vboxsf sharing /mnt/shared

/etc/hosts追加

192.168.111.111 master

192.168.112.112 slave1

192.168.113.113 slave2

2.jdk配置

源代码位置，就地解压安装，加全部权限

/usr/src/hadoop-3.0.0/

/usr/src/jdk1.8.0\_161/

安装sun jdk

/etc/profile追加环境变量

#set java environment

export JAVA\_HOME=/usr/src/jdk1.8.0\_161

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

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

export JSVC\_HOME=.:$JAVA\_HOME/lib/:$JAVA\_HOME/lib/commons-daemon-1.1.0.jar

3.hadoop配置

hadoop安装

/usr/src/hadoop-3.0.0/

bin/hadoop version

/etc/profile追加

export HADOOP\_HOME=/usr/src/hadoop-3.0.0

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

cd /usr/src/hadoop-3.0.0/etc/hadoop/

/usr/src/hadoop-3.0.0/etc/hadoop/hadoop-env.sh追加

export JAVA\_HOME=/usr/src/jdk1.8.0\_161

export HDFS\_DATANODE\_SECURE\_USER=root

export HDFS\_DATANODE\_SECURE\_USER=root

export HDFS\_SECONDARYNAMENODE\_USER=root

export HDFS\_NAMENODE\_USER=root

core-site.xml

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://master:9000</value> //master即为控制节点的主机名，

</property>

<property>

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

<value>/usr/src/hadoop-3.0.0/tmp</value> //新建tmp文件夹

<description>Abase for other temporary directories.</description>

</property>

</configuration>

hdfs-site.xml

<configuration>

<!-- 设置namenode的http通讯地址 -->

<property>

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

<value>master:50070</value>

</property>

<!-- 设置secondarynamenode的http通讯地址 -->

<property>

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

<value>slave1:50090</value>

</property>

<!-- 设置namenode存放的路径 -->

<property>

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

<value>/usr/src/hadoop-3.0.0/name</value>

</property>

<!-- 设置hdfs副本数量 -->

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<!-- 设置datanode存放的路径 -->

<property>

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

<value>/usr/src/hadoop-3.0.0/data</value>

</property>

</configuration>

mapred-site.xml

<configuration>

<property>

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

<value>yarn</value>

</property>

<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>

</configuration>

workers

slave1

slave2

yarn-site.xml

<configuration>

<!-- 设置 resourcemanager 在哪个节点-->

<property>

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

<value>master</value>

</property>

<!-- reducer取数据的方式是mapreduce\_shuffle -->

<property>

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

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>

<value>org.apache.hadoop.mapred.ShuffleHandler</value>

</property>

</configuration>

新建slave文件

slave1

slave2

/root/.bashrc

export HADOOP\_HOME=/usr/src/hadoop-3.0.0

export HADOOP\_INSTALL=$HADOOP\_HOME

export HADOOP\_MAPRED\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_HOME=$HADOOP\_HOME

export HADOOP\_HDFS\_HOME=$HADOOP\_HOME

export YARN\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native

export PATH=$PATH:$HADOOP\_HOME/sbin:$HADOOP\_HOME/bin:/usr/src/hadoop-3.0.0/bin:/usr/src/hadoop-3.0.0/sbin

vim sbin/start-dfs.sh

vim sbin/stop-dfs.sh

两处顶部增加以下内容

HDFS\_DATANODE\_USER=root

HADOOP\_SECURE\_DN\_USER=hdfs

HDFS\_NAMENODE\_USER=root

HDFS\_SECONDARYNAMENODE\_USER=root

sbin/start-yarn.sh

sbin/stop-yarn.sh

两处增加以下内容

YARN\_RESOURCEMANAGER\_USER=root

YARN\_NODEMANAGER\_USER=root

HADOOP\_SECURE\_DN\_USER=yarn

HDFS\_NAMENODE\_USER=root

4.多机器配置

主机复制到从机，配置网卡，/etc/hostname改名，各个ssh公钥共享

ssh免密登录

ssh-keygen -t rsa生成公钥

Your identification has been saved in /root/.ssh/id\_rsa.

Your public key has been saved in /root/.ssh/id\_rsa.pub.

cp id\_rsa.pub authorized\_keys

此文件内容追加到同位置新建文件authorized\_keys中

cat /root/.ssh/id\_rsa.pub >> /mnt/hgfs/sharing/authorized\_keys

其余两台相同操作，最后将追加了所有公钥的/mnt/hgfs/sharing/authorized\_keys文件分发至各台虚拟机的 /root/.ssh/，实现互相ssh免去密码登录

5.hadoop初始化

cd /usr/src/hadoop-3.0.0/bin 运行hdfs namenode -format

cd /usr/src/hadoop-3.0.0/sbin 下运行hadoop start-all.sh开启hadoop

虚拟机master浏览器中访问master:8088可以看到hadoop控制面板

## hadoop3简易安装

完整安装之后配置的文件夹可以打包，留给其他机器上简易部署。

宿主机win1064

虚拟工具vmware14

虚拟机debian9.3 x32,网易镜像

用户名root 密码passwd 默认root目录

winscp访问linux文件，mobaxerm访问ssh

编辑器：everdit/vscode，编码utf无bom

1.虚拟环境配置

一主三从，0.5g内存。

vmware->编辑->各个虚拟机分配的网卡为

名称 ip 网卡

master 192.168.111.111 //网卡1：nat联网

slave1 192.168.112.112 //网卡2：仅与主机连接

slave2 192.168.113.113 //网卡3：仅与主机连接

2.开启ssh

图形化界面进入

/etc/ssh/sshd.conf

PermitRootLogin without-password 修改为

PermitRootLogin yes

之后用mobaxterm、winscp通过连接虚拟机

vmware安装增强工具，运行/media/cdrom0/run\_upgrader.sh

本地建立共享文件夹D:\sharing，映射为各个虚拟机/mnt/hgfs/sharing/

替换/etc/hosts

3.安装jdk

解压jdk至/usr/src/jdk1.8.0\_161/

替换/etc/profile

运行 apt-get install jsvc

4.安装hadoop

解压hadoop 至/usr/src/hadoop-3.0.0/

运行：

cd /usr/src/hadoop-3.0.0/

bin/hadoop version

替换 /usr/src/hadoop-3.0.0/

5.复制从机

将master复制（vmware中叫克隆）为slave1、slave2，分配网卡

/etc/hostname分别改为master、slave1、slave2

6.分发sshkey

运行ssh-keygen 生成公钥

Your identification has been saved in /root/.ssh/id\_rsa.

Your public key has been saved in /root/.ssh/id\_rsa.pub.

运行cp id\_rsa.pub authorized\_keys

此文件内容追加到同位置新建文件authorized\_keys中

运行cat /root/.ssh/id\_rsa.pub >> /mnt/hgfs/sharing/authorized\_keys

将/mnt/hgfs/sharing/authorized\_keys分发至各台虚拟机的 /root/.ssh/

7.hadoop启动

master初始化，只进行一次

cd /usr/src/hadoop-3.0.0/bin 运行hdfs namenode -format

hadoop启动 master中

cd /usr/src/hadoop-3.0.0/sbin 下运行hadoop start-all.sh开启hadoop

虚拟机master浏览器中访问master:8088可以看到hadoop控制面板

## 常用命令

常见流程，eclipse导出项目的jar，jar传至master的linux路径，再执行hadoop命令运行jar，或者eclipse上带参运行。

#注意，hadoop jar 后一个参数是本地路径，次一个参数是hadoop路径

hadoop fs -mkdir /input

hadoop fs -ls /input

hadoop fs -put "/home/cloudera/workspace/jar/weibo1000.txt" /input

hadoop fs -put /home/cloudera/workspace/jar/weibo.jar /input

hadoop fs -put /home/cloudera/workspace/jar/sentiment.jar /input

hadoop jar /home/cloudera/workspace/jar/weibo.jar Main /input/weibo1000.txt /output

hadoop jar /home/cloudera/workspace/jar/sentiment.jar Main /input/weibo1000.txt /output

hadoop jar sentiment.jar Main /input/weibo1000.txt /output

hadoop fs -rm /input/weibo.jar

hadoop fs -rm /input/sentiment.jar