**20200623学习报告**

1813068郭冠琛

**学习内容**

* 虚拟机与宿主机间互传文件，共享文件夹在/mnt/hgfs中

1. 安装vmtools（VMware 安装vmtools 解压文件夹 sudo ./vmware-install.pl）
2. 虚拟机设置-选项-启用共享文件夹
3. 查看是否共享成功vmware-hgfsclient查看当前被共享文件夹
4. 安装open-vm-tools并挂载，才能显示当前文件夹

yum install open-vm-tools

vmhgfs-fuse .host:/ /mnt/hgfs

直接从宿主机拖拽文件进入虚拟机会造成文件损坏

* java安装

解压jdk文件

卸载默认安装的jdk（使用rpm -qa | grep java查看当前安装的jdk并卸载）rpm -e –nodeps 安装文件卸载

配置环境变量 vim /etc/profile在文件中加入如下路径

export JAVA\_HOME=jdk解压的路径

export CLASSPATH=$:CLASSPATH:$JAVA\_HOME/lib/

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

source /etc/profile重新加载

* Hadoop安装（解压目录在/opt）
  1. 配置所有主机虚拟ip、ssh、主机名host，profile文件

export HADOOP\_HOME=Hadoop安装目录

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

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

export HADOOP\_CONF\_DIR=${HADOOP\_HOME}/etc/hadoop

* 1. 创建相关目录

mkdir -p /opt/hadoop-3.2.1/tmp /opt/hadoop-3.2.1/hdfs /opt/hadoop-3.2.1/hdfs/data /opt/hadoop-3.2.1/hdfs/name

* 1. 修改配置文件

vim /opt/hadoop-3.2.1/etc/hadoop/hadoop-env.sh

# 注意：自己安装的java目录

export JAVA\_HOME=自己的java目录

# 文件最后追加

export HDFS\_NAMENODE\_USER=root

export HDFS\_DATANODE\_USER=root

export HDFS\_SECONDARYNAMENODE\_USER=root

export YARN\_RESOURCEMANAGER\_USER=root

export YARN\_NODEMANAGER\_USER=root

vim /opt/hadoop-3.2.1/etc/hadoop/core-site.xml

# 按照以下内容修改

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://主节点主机名:9000</value>

<description>HDFS的URI，文件系统://namenode标识:端口号</description>

</property>

<property>

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

<value>/opt/hadoop-3.2.1/tmp</value>

<description>namenode上本地的hadoop临时文件夹</description>

</property>

</configuration>

vim /opt/hadoop-3.2.1/etc/hadoop/hdfs-site.xml

# 修改如下配置

<configuration>

<property>

<name>dfs.replication</name>

<value>3</value>

<description></description>

</property>

</configuration>

vim /opt/hadoop-3.2.1/etc/hadoop/mapred-site.xml

<configuration>

<property>

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

<value>yarn</value>

</property>

</configuration>

vim /opt/hadoop-3.2.1/etc/hadoop/yarn-site.xml

<configuration>

<property>

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

<value>mapreduce\_shuffle</value>

</property>

</configuration>

vim /opt/hadoop-3.2.1/etc/hadoop/workers

所有作为DataNode节点的主机名

* 1. 将配置好的安装文件传输到其他主机（-r表示传输文件夹）

scp -r 安装文件目录/文件夹 root@主机名：安装文件夹所在位置

* 1. 主机格式化

/opt/hadoop-3.2.1/bin/hdfs namenode -format 出现has been successfully formatted则成功

* 1. 启动

/opt/program-files/hadoop/hadoop-3.2.1/sbin/start-all.sh

启动后使用jps查询应出现进程：

NodeManager （11|12|13机器上，也就是slave1，slave2，slave3节点上）

DataNode （11|12|13机器上，也就是slave1，slave2，slave3节点上）

NameNode （10机器上，也就是hadoop主节点上）

SecondaryNameNode （10机器上，也就是hadoop主节点上）

ResourceManager （10机器上，也就是hadoop主节点上）

* ssh：ssh-keygen -t rsa产生公钥和私钥 hosts文件配置能识别的主机名 ssh-copy-id 主机名将文件传输到对应电脑 则该电脑可远程免密登录对方电脑 三台机器互相copy-id即可实现互相免密登录
* 每次启动需要重新source profile文件才能识别java命令：在~/.bashrc文件末尾加source /etc/profile
* Hadoop启动resourcemanager和nodemanager报错：Caused by: java.lang.NoClassDefFoundError: javax/activation/DataSource

Caused by: java.lang.ClassNotFoundException: javax.activation.DataSource

下载所缺文件：执行命令cd ${HADOOP\_HOME}/share/hadoop/yarn/lib

wget <https://repo1.maven.org/maven2/javax/activation/activation/1.1.1/activation-1.1.1.jar>

* 安装后：

http://主机ip:9870 （nameNode information 存在于主节点上 ）

http://主机ip:8088/cluster （All Applications 集群运行状态，以及任务执行服务）

http://slave\_ip:8042/node （NodeManager 数据节点管理服务，存在于DataNode数据节点上）

http://slave\_ip:9864/datanode.html （DataNode，数据存储， 存在于DataNode数据节点上）

* 安装成功后访问集群运行状态不显示其他节点

临时关闭防火墙systemctl stop firewalld.service

* 试运行Wordcount

1. 创建示例文件

vim /root/test.txt并输入任意测试内容

1. 上传文件到hdfs

hdfs dfs -mkdir /user/root

hdfs dfs -put /root/test.txt /user/root/

1. 找到官方自带jar

cd /opt/hadoop-3.2.1/share/hadoop/mapreduce/

1. 运行命令

hadoop jar hadoop-mapreduce-examples-3.2.1.jar wordcount /user/root/test.txt /root/output

Wordcount为案例名 第一个参数是hdfs上的文件路径 第二个参数是hdfs结果输出路径（必须不存在）

1. 查看运行结果

hdfs dfs -lsr /root/output

输出success文件则执行成功

1. 验证

hdfs dfs -cat /root/output/part-r-00000

r为reduce输出 改为m则为mapper输出 整个文件即结果文件

* 遇到问题：运行卡在job running

查看log发现unsetHostnameOrPost错误

将以下代码添加到所有配置的marped-site.xml

<property>

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

<value>master:8032</value>

</property>

<property>

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

<value>master:8030</value>

</property>

<property>

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

<value>master:8031</value>

</property>

其中master为主机名

* 遇到问题：无法加载主类

找不到或无法加载主类org.apache.hadoop.mapreduce.v2.app.MRAppMaster

在命令行中Hadoop classpath

在yarn-site.xml中添加代码

<property>

<name>yarn.application.classpath</name>

<value>输入刚才返回的Hadoop classpath路径</value>

</property>

* 仍卡在job running

1. 在yarn-site.xml中添加

<property>

<name>yarn.nodemanager.resource.memory-mb</name>

<value>20480</value>

</property>

<property>

<name>yarn.scheduler.minimum-allocation-mb</name>

<value>2048</value>

</property>

<property>

<name>yarn.nodemanager.vmem-pmem-ratio</name>

<value>2.1</value>

</property>

出现此故障的原因可能是，在每个虚拟机分配的内存和CPU资源太少，不能满足Hadoop运行所需的默认资源需求

1. 更改marped-site.xml

将

<property>

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

<value>yarn</value>

</property>

改为

<property>

<name>mapreduce.job.tracker</name>

<value>hdfs://master的ip:8001</value>

<final>true</final>

</property>

**需要解决的问题**

进度过慢，周三需要把两天的内容全部完成