学习总结

1812975 曹沛鑫 2020.6.23

1. MapReduce处理数据流程

·map函数从原始数据中提取有用数据并将其作为输出被发送

·输出经过MapReduce框架处理--排序分组

·reduce函数根据需求得到最后结果

1. 配置网络Ip地址（参考：<https://blog.csdn.net/yuioplv/article/details/79631745?utm_medium=distribute.pc_relevant_t0.none-task-blog-BlogCommendFromMachineLearnPai2-1.nonecase&depth_1-utm_source=distribute.pc_relevant_t0.none-task-blog-BlogCommendFromMachineLearnPai2-1.nonecase>）
2. 安装Hadoop
3. 环境准备

·关闭selinux：cat /etc/selinux/config|grep 'SELINUX=disabled' #永久关闭

·关闭防火墙：systemctl stop firewalld  #关闭防火墙

systemctl disable firewalld  #禁止开机启动

·修改hostsname：hostnamectl  set-hostname master #其他节点修改成hadoop1,hadoop2

·安装ntp：[root@master ~]# yum install ntp -y  #安装ntp

[root@master sbin]# crontab -e       #设置同步时间

0 \* \* \* \* /usr/sbin/ntpdate cn.pool.ntp.org

·设置ssh互信：ssh-keygen #一路回车，生成秘钥

·将公钥拷贝到其他worker节点：ssh-copy-id root@worker1

ssh-copy-id root@worker2

ssh-copy-id root@worker3

## 2.**解压Hadoop安装包（只在master做）**

确保network网络已经配置好，使用Xftp等类似工具进行上传，把hadoop-2.7.5.tar.gz上传到/opt/hadoop目录内。

上传完成后，在master主机上执行以下代码：cd /opt/hadoop

进入/opt/hadoop目录后，执行解压缩命令：tar -zxvf hadoop-2.7.5.tar.gz

回车后系统开始解压，屏幕会不断滚动解压过程，执行成功后，系统在hadoop目录自动创建hadoop-2.7.5子目录。

然后修改文件夹名称为“hadoop”，即hadoop安装目录，执行修改文件夹名称命令：mv hadoop-2.7.5 hadoop

## **3.配置env文件（只在master做）**

配置jdk文件

执行命令：

vi /opt/hadoop/hadoop/etc/hadoop/hadoop-env.sh

1

找到 “ export JAVA\_HOME ” 这行，用来配置jdk路径

修改为：export JAVA\_HOME=/usr/local/java/jdk1.8.0\_162/

## 4.**配置核心组件文件（只在master做）**

执行编辑core-site.xml文件的命令：vi /opt/hadoop/hadoop/etc/hadoop/core-site.xml

需要在<configuration>和</configuration>之间加入的代码：

<property>

<name>fs.defaultFS</name>

<value>hdfs://master:9000</value>

</property>

<property>

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

<value>/opt/hadoop/hadoopdata</value>

</property>

## 5.**配置文件系统（只在master做）**

执行编辑hdfs-site.xml文件的命令：vi /opt/hadoop/hadoop/etc/hadoop/hdfs-site.xml

需要在<configuration>和</configuration>之间加入的代码：

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

## 6.**配置 yarn-site.xml 文件（只在master做）**

执行编辑yarn-site.xml文件的命令：vi /opt/hadoop/hadoop/etc/hadoop/yarn-site.xml

需要在<configuration>和</configuration>之间加入的代码：

<property>

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

<value>mapreduce\_shuffle</value>

</property>

<property>

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

<value>master:18040</value>

</property>

<property>

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

<value>master:18030</value>

</property>

<property>

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

<value>master:18025</value>

</property>

<property>

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

<value>master:18141</value>

</property>

<property>

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

<value>master:18088</value>

</property>

## 7.**配置MapReduce计算框架文件（只在master做）**

执行命令：vi /opt/hadoop/hadoop/etc/hadoop/mapred-site.xml

需要在<configuration>和</configuration>之间加入的代码：

<property>

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

<value>yarn</value>

</property>

## 8.**配置master的hadoops文件（只在master做）**

执行编辑slaves文件命令：vi /opt/hadoop/hadoop/etc/hadoop/workers

应当加入以下代码：hadoop1

hadoop2

## **复制master上的Hadoop到hadoop2\hadoop3节点（只在master做）**

scp -r /opt/hadoop root@hadoop2:/opt

scp -r /opt/hadoop root@hadoop3:/opt

## **Hadoop集群的启动-配置操作系统环境变量（三个节点都做）**

同时开启三个虚拟机卡死，不得不重启先写完总结后再尝试

1. 总结

配置ip网络后不能使用，耽误大量时间，明天会继续尝试

重新复习了vi的使用