**实验报告**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 学 号 | 18020032020 | | 姓 名 | 王子轩 | 专业班级 | | 计算机科学与技术2班 | |
| 课程名称 | 大数据导论 | | | | 学期 | | **2021年秋季学期** | |
| 任课教师 | 刘艳艳,刘洁 | | 完成日期 | 2021/10/2 | | 上机课时间 | | 周五78节 |
| 实验名称 | | 实验一Hadoop实验环境配置 | | | | | | |
| **一、实验要求（10%）**   1. VMware软件的安装 2. 安装Linux系统 3. 配置模板虚拟机 4. 克隆虚拟机模板 5. 在DataNode1上安装JDK 6. 在DataNode1安装Hadoop 7. 完全分布式运行模式，SSH无密登录配置，集群配置   **二、实验内容及步骤（80%）**  1. VMware软件的安装  因本人电脑已经安装过VMware软件，故此步骤跳过。  2. 安装Linux系统  说明文档中有安装虚拟机的博客链接，里边内容很具体，因此在安装此步骤没有太大问题，就是在安装图形界面时遇到了问题，具体见下文中的心得总结。界面如下：    3. 配置模板虚拟机  修改NameNode的ip和主机名以及Windows中虚拟网卡中的ip，教程给的步骤十分具体，没有遇到问题，按照教程即可完成。结果如下图：    修改ifcfg-ens33文件，效果如下图：    安装相关工具包，如epel-release、net-tool等，在安装图形界面前就安装了，不在放效果图了。  修改/etc/sudoers文件，效果图如下：    在/opt目录下创建文件夹，并修改所属主和所属组，效果图如下：    卸载虚拟机自带的JDK，如下图：            关闭NameNode的防火墙，并且关闭开机自启，效果图如下：    4. 在NameNode上安装JDK  在NameNode上安装JDK之前，先下载winscp来将压缩包传到虚拟机中，但是在使用winscp连接NameNode时，一直显示连接超时，一直连接不上，具体解决过程见下面心得总结，在这里放几张图片：  连接超时：    虚拟机可以ping通windows：  NameNode的ssh服务已经打开：    在NameNode上添加20端口：      最后发现是需要重启一波windows里的对应虚拟网卡，可以成功连接。然后使用winscp将jdk和hadoop的压缩包传输到/pot/software中去。  将jdk压缩包解压到module文件夹中，效果图片如下：      配置jdk环境变量并查看是否安装成功，效果图如下：    5. 在NameNode安装Hadoop  将压缩包解压到module文件夹中：    查看是否解压成功：    将hadoop添加到环境变量并查看是否安装成功：    6. 克隆虚拟机模板  克隆虚拟机比较简单，直接点击克隆，选择完整克隆，即可克隆出两台虚拟机，以防万一，在克隆之前保存个快照。  因为在克隆之前就在模板机上安装了新的JDK和Hadoop，因此只需改变克隆出的两台机子的主机名和ip，DataNode2如下图：    修改windows的主机映射文件（host文件），比较简单，不在放图。  7. 完全分布式运行模式，SSH无密登录配置，集群配置  scp安全拷贝，案例实操，效果图如下：      配置SSH：    无秘钥配置：    拷贝公钥不再放图了，在下面的集群配置中会给出免密登录的效果图。  配置四个.xml文件也比较简单，将教程导入三台虚拟机中，使用vim命令打开编辑配置文件，复制教程中的内容到相应的配置文件当中即可。  在分发配置文件的过程中出现问题，分发文件的位置好像出现问题，具体情况在下面的心得总结中有体现。  最后附上启动集群后的几张效果图：        这张图片可以显示出ssh免密登录的效果：    **三、心得总结（写出自己在完成实验过程中遇到的问题、解决方法，以及体会、收获等）（10%）**  在安装好第一台虚拟机后，尝试安装图形界面，但是第一次在下载时，中途显示没有去其他可尝试的镜像，安装失败；然后按照网上教程更新了内核，删除缓存，再重新下载安装图形界面，最终安装成功，后面克隆的两台虚拟机也因此都有图形界面。  在下载使用winscp和Xshell时，死活连接不上虚拟机中的centos，一直显示连接超时。在网上找解决方法，首先是关闭Linux的防火墙，这个在安装Linux的时候就做了，为了以防万一还是查看了防火墙的状态，确定是关闭的，没有解决问题；然后有博客说是在Linux上启动ssh服务，查看sshd状态后，发现是在running，仍未解决问题；再然后是说看看Linux是否开启20端口，我发现我的NameNode并没有这个端口，于是添加这个端口，并且开启端口（期间需要打开firewalld），最后发现还是显示超时；之后又查看了VMnet8中ip和默认网关与虚拟机中虚拟网络编辑器中的网关和子网IP是否对应（其实本来就是对应的，在第一步中就有设置这两个地方），仍未解决问题；最后是将VMnet8这个网卡禁用再重新启用后解决问题。  解压jdk压缩包出现问题，发现是给的教程文档里边的命令有问题，应该加上jdk所在位置，只有文件名是找不到文件的，命令应为“tar -xvf /opt/software/exe/jdk-8u212-linux-x64.tar.gz -C /opt/module”，解压成功。后来发现是我没有先进入software文件夹，命令没有给错。  配置好某台机子的集群文件后要分发到剩下两台机子上，但是分发命令好像有问题，将配置好的hadoop文件夹给分发到目的机的hadoop文件下了，应该是取代这个文件夹的才对。  最后在NameNode启动hdfs和在DataNode1启动yarn时，电脑基本卡得动都动不了，经过五六分钟左右才稍微好点，可以再虚拟机的浏览器中查看HDFS上的数据信息和YARN运行中的Job信息，也不会知道往后的实验会不会更卡。  总之，实验一就遇到了不少问题，虽然花费大量时间在网上查找解决方法（网上内容太繁杂，筛选出正确解答方法很费时间），但最终还是都能解决，最后实验完成时的成就感还是很强的。后面的实验也一定会遇到各种各样的问题，但我还是会尽力去解决问题的。 | | | | | | | | |