杭州电子科技大学

HANGZHOU DIANZI UNIVERSITY

2021/2022学年第二学期《大数据平台》课程四次作业

专　　业：信息与计算科学

班　　级：20073112

学　　号：20071230

姓　　名：武　琦

任课老师：邵新平

完成时间：2022年5月

作业一：在虚拟 VmBox 上安装 Linux

第一步：安装 VirtualBox 虚拟机软件

1. 下载 VirtualBox-6.1.18-142142-Win.rar
2. 运行 VirtualBox 安装程序
3. 单击“Next”按钮开始安装
4. 选择储存位置（按个人喜好，本人选择的是 E:\VirtualBox），单击“Next”
5. 完成设置后，单击“Install”按钮开始安装
6. 完成后单击“Finish”，启动 VirtualBox

第二步：在 VirtualBox 创建虚拟机

1. 单击“新建”，名称为“master”，文件夹位置为“E:\VirtualBox”，类型为“Linux”，版本为“Ubuntu（64-bit）”，单击“下一步”
2. 内存大小为 1024MB，单击“下一步”
3. 点击“现在创建虚拟硬盘”，单击“创建”
4. 虚拟硬盘文件类型选择为“VDI”，单击“下一步”
5. 存储在物理硬盘上选择“固定大小”，单击“下一步”
6. 选择虚拟硬盘大小为“20.00GB”，单击“创建”

第三步：安装 Ubuntu Linux 操作系统

1. 下载 ubuntu-14.04.4-desktop-amd64.iso
2. 在 VirtualBox 页面上点击已建好的“master”，单击“设置”
3. 单击“系统”，将启动顺序改为“光驱，软驱，硬盘”

4. 单击“网络”，点击“网卡 2”，选择“ √(启用网络连接)”，连接方式为“仅主机(Host-Only)网络”，点击“高级”，混杂模式选择为“全部允许”
5. 单击“存储”，选择下方的“加号”中的 USB 设备，在刚刚创建的“控制器：USB”中点击第一个“加号”，单击“注册”，选择下载好的“ubuntu-14.04.4-desktop-amd64.iso”，单击“选择”；点击“控制器：IDE”，重复上面操作
6. 点击“OK”

第四步：开始安装 Ubuntu

1. 启动虚拟机
2. 安装语言版本，选择“English”，单击“Install Ubuntu”
3. 选择全部“ √ ”，单击“continue”
4. 单击第一个选项，点击安装，一直选“continue”
5. 键盘布局为“English”，单击“continue”
6. 创建姓名和密码，计算机名为“Hadoop”，选择“登录时需要密码”，单击“continue”
7. 安装完成，强制退出，在设置中选择“存储”，删除“控制器：IDE”中的“ubuntu-14.04.4-desktop-amd64.iso”，点击“OK”
8. 单击“启动”，使用“Ctrl+Alt+F1”进入命令行，输入用户名、密码，按下“Enter”，Ubuntu Linux 操作系统安装完成

简单介绍 Linux 系统

Linux，全称 GNU/Linux，是一套免费使用和自由传播的类 Unix 操作系统，是一个基于 POSIX 的多用户、多任务、支持多线程和多 CPU 的操作系统。伴随着互联网的发展，Linux 得到了来自全世界软件爱好者、组织、公司的支持。它除了在服务器方面保持着强劲的发展势头以外，在个人电脑、嵌入式系统上都有着长足的进步。使用者不仅可以直观地获取该操作系统的实现机制，而且可以根据自身的需要来修改完善 Linux，使其最大化地适应用户的需要。

Linux 不仅系统性能稳定，而且是开源软件。其核心防火墙组件性能高效、配置简单，保证了系统的安全。在很多企业网络中，为了追求速度和安全，Linux 不仅仅是被网络运维人员当作服务器使用，甚至当作网络防火墙，这是 Linux 的一大亮点。

Linux 具有开放源码、没有版权、技术社区用户多等特点，开放源码使得用户可以自由裁剪，灵活性高，功能强大，成本低。尤其系统中内嵌网络协议栈，经过适当的配置就可实现路由器的功能。这些特点使得 Linux 成为开发路由交换设备的理想开发平台。

作业二：搭建 Hadoop 集群和 Spark 平台

搭建 Hadoop 集群

第一步：Xshell 远程访问工具和 Xftp 文件传输

1. 下载 Xshell 和 Xftp，并分别安装
2. 打开 Xshell——文件：新建——输入主机(IP 地址)、用户名和密码———一次性连接
3. 打开 Xftp——相同步骤

第二步：SSH 免密登录

1.sudo apt-get install update （更新）

2.ssh-keygen（产生 SSH Key 密钥）

3.cd ~/.ssh

4.cp id_dsa.pub authorized_keys （将产生的 Key 放置到许可证文件中）

第三步：安装 JDK

1. mkdir software （新建文件夹）
2. 使用 Xftp 将下载好的 jdk-8u131-linux-x64.tar.gz 移动到虚拟机下
3. jar xvf jdk-8u131-linux-x64.tar.gz ~/software/ （解压）
4. touch mybash.sh(创建环境变量包)

   nano mybash.sh
5. export JAVA_HOME=/home/xuan/software/jdk1.8.0_131

   export PATH=$JAVA_HOME/bin:$PATH

export

CLASSPATH=.:$JAVA_HOME/lib/dt.jar:$JAVA_HOME/lib/tools.jar （复制到 mybash.sh 中）

6. Ctrl O——Enter——Ctrl X （保存并退出）
7. source mybash.sh （以后每次打开 Linux 都要进行的操作）
8. java -version （测试 JDK 是否安装成功）

第四步：安装 Hadoop

1. 打开 Xftp 将下载好的 hadoop-2.6.4.tar.gz 移动到虚拟机下
2. jar xvf hadoop-2.6.4.tar.gz ~/software/ （解压）
3. nano mybash.sh
4. export HADOOP_HOME=/home/xuan/software/hadoop-2.6.4

   export PATH=$HADOOP_HOME/sbin:$HADOOP_HOME/bin:$PATH

   （添加到 mybash.sh 中）
5. 保存并退出
6. source mybash.sh
7. cd $HADOOP_HOME/etc/Hadoop
8. nano core-site.xml （将下图代码复制到文件后保存并退出，注意文件位置）

```
<configuration>
        <property>
             <name>hadoop.tmp.dir</name>
             <value>file:/home/xuan/software/hadoop-2.6.4/tmp</value>
             <description>Abase for other temporary directories.</description>
        </property>
        <property>
             <name>fs.defaultFS</name>
             <value>hdfs://localhost:9000</value>
        </property>
</configuration>
```

9. nano hdfs-site.xml （相同操作）

```
<configuration>
        <property>
             <name>dfs.replication</name>
             <value>1</value>
        </property>
        <property>
             <name>dfs.namenode.name.dir</name>
             <value>file:/home/xuan/software/hadoop-2.6.4/namedir</value>
        </property>
        <property>
             <name>dfs.datanode.data.dir</name>
             <value>file:/home/xuan/software/hadoop-2.6.4/datadir</value>
        </property>
</configuration>
```

第五步：启动 Hadoop

1. hadoop namenode -format （格式化数据，只使用一次）

2. $HADOOP_HOME/sbin/start-all.sh （同时启动 HDFS、Yarn）

3. jps （查看已启动的进程）

```
1715 NameNode
2313 NodeManager
1993 SecondaryNameNode
1834 DataNode
2347 Jps
2189 ResourceManager
```

4. 如图，安装成功

搭建 Spark 平台

第一步：安装 Scala

1. 用 Xftp 将下载好的 scala-2.11.6.tar.gz 移动到虚拟机下

2. tar xvf scala-2.11.6.tar.gz ~/software/ （解压）

3. nano mybash.sh

4. export SCALA_HOME=/home/xuan/software/scala-2.11.6

   export PATH=$SCALA_HOME/bin:$PATH （添加到 mybash.sh 中）

5. source mybash.sh

6. Scala （启动 Scala）

7. Ctrl C （退出 Scala）

第二步：安装 Spark

1.用 Xftp 将下载好的 spark-2.4.5-bin-hadoop2.6.tar.gz 移动到虚拟机下

2.tar spark-2.4.5-bin-hadoop2.6.tar.gz ~/software （解压）

3.nano mybash.sh

4.export SPARK_HOME=/home/xuan/software/spark-2.4.5-bin-hadoop2.6

export PATH=$SPARK_HOME/bin:$SPARK_HOME/sbin:$PATH（添加到 mybash.sh 中）

5.source mybash.sh

6.pyspark （启动 pyspark）

作业三：Pyspark 统计文档字数

1. 打开文件

```
data=sc.textFile("/home/xuan/bigdata/The Old Man and the Sea.txt")
```

2. 句子分割为列表

```
data1=data.flatMap(lambda line:line.split(" "))
```

3. 去除标点符号

```
import re
```

```
data2=data1.map(lambda x:re.sub('[\W_]+','',x))
```

4. 统计单词个数

```
data3=data2.map(lambda word:(word,1))
```

```
data4=data3.reduceByKey(lambda x,y:x+y)
```

5. 结果保存

```
f=open("/home/xuan/bigdata/result.txt","w")
```

```
for i,j in data4.collect():
```

```
        f.write(i+'\t'+str(j)+'\n')
```

```
f.close()
```

6. 部分结果展示(共有 2571 个单词)

pardon 1
comparatively 1
knelt 1
yellow 11
four 5
sleep 17
appetite 1
skeleton 1
looking 8
feeding 2

作业四：RDD 求平均分

求每个学生的平均分

1. 导入文件和库

```
import re
```

```
data=sc.textFile("/home/xuan/bigdata/score.txt")
```

2. 将数字和字母外的字符转化为空格

```
data1=data.map(lambda x:re.sub('[\W_]+',' ',x))
```

3. 字符串分割为列表

```
data2=data1.map(lambda x:x.split())
```

4. 选择学生的序号和分数组成元组

```
data3=data2.map(lambda x:(x[0],x[2]))
```

5. 以学生的序号计数

```
data4=data3.map(lambda x:(x[0],(int(x[1]),1)))
```

6. 求每个人的总分和总课程数

```
data5=data4.reduceByKey(lambda x,y:(x[0]+y[0],x[1]+y[1]))
```

7. 求每个人的平均分

```
data6=data5.map(lambda x:(x[0],float(x[1][0])/x[1][1]))
```

结果：

[(u'1', 92.0), (u'3', 84.66666666666667), (u'2', 88.0)]

求每门课的平均分

将元组改为课程名和分数，其余代码不变

```
data3=data2.map(lambda x:(x[1],x[2]))
```

结果：

[(u'mathematic', 91.33333333333333), (u'chinese', 81.66666666666667),

(u'english', 91.66666666666667)]