“大数据工程”课程实验报告

题目：Hadoop 的安装与操作	学号姓名：21377061 范春	日期：2024.03.05

实验环境：

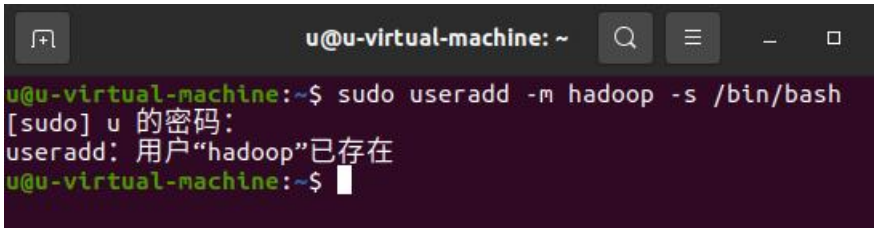
虚拟机软件：VMware

Linux 系统：Ubuntu 20.04

实验内容与完成情况：

一、前置步骤：

1、创建 Hadoop 用户



2、安装和配制 SSH

```
hadoop@u-virtual-machine:~$ sudo apt-get install openssh-server
正在读取软件包列表... 完成
正在分析软件包的依赖关系树
正在读取状态信息... 完成
openssh-server 已经是最新版 (1:8.2p1-4ubuntu0.11)。
升级了 0 个软件包，新安装了 0 个软件包，要卸载 0 个软件包，有 0 个软件包未被升级。
hadoop@u-virtual-machine:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:yfeSgPePGGzorzu6kLn6DHlzVZfO3SmUNiSMZyu1JuM.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
hadoop@localhost's password:
Permission denied, please try again.
hadoop@localhost's password:
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-97-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

 * Introducing Expanded Security Maintenance for Applications.
   Receive updates to over 25,000 software packages with your
   Ubuntu Pro subscription. Free for personal use.

     https://ubuntu.com/pro

扩展安全维护（ESM）Applications 未启用。

0 更新可以立即应用。

启用 ESM Apps 来获取未来的额外安全更新
See https://ubuntu.com/esm or run: sudo pro status

Your Hardware Enablement Stack (HWE) is supported until April 2025.

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.
```

```
hadoop@u-virtual-machine:~$ exit
注销
Connection to localhost closed.
hadoop@u-virtual-machine:~$ cd ~/.ssh/
hadoop@u-virtual-machine:~/.ssh$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:/4+7mOhqeY/GHVOHd71TmGHSAI/qr1ErH5NvvmksDjY hadoop@u-virtual-machine
The key's randomart image is:
+---[RSA 3072]----+
|          ...o   |
|           o. +  |
|          . .+ +.|
|         .  o = +|
|        S .. o .o|
|       . ooo   o |
|       oE.*+    .|
|      o.+XoB+o   |
|     ..==+B=X=.  |
+----[SHA256]-----+
hadoop@u-virtual-machine:~/.ssh$ cat ./id_rsa.pub >> ./authorized_keys
hadoop@u-virtual-machine:~/.ssh$ ssh localhost
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-97-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage
```

3、安装 java 环境

```
hadoop@u-virtual-machine:~$ cd /usr/lib
hadoop@u-virtual-machine:/usr/lib$ sudo mkdir jvm
```

创建/usr/lib/jvm 目录用来存放 jdk 文件。

```
hadoop@u-virtual-machine:/usr/lib/jvm$ ls
jdk1.8.0_162
hadoop@u-virtual-machine:/usr/lib/jvm$ cd ~
hadoop@u-virtual-machine:~$ gedit ~/.bashrc
hadoop@u-virtual-machine:~$ source ~/.bashrc
hadoop@u-virtual-machine:~$ java -version
bash: /usr/lib/jvm/jdk1.8.0_162/bin/java: 权限不够
```

配制环境变量并检查 java 的版本以确定是否安装成功，此处遇到的问题如图所示，显示权限不够。

```
hadoop@u-virtual-machine:/usr/lib/jvm$ ls
jdk1.8.0_162
hadoop@u-virtual-machine:/usr/lib/jvm$ chmod 777 jdk1.8.0_162/bin/java
hadoop@u-virtual-machine:/usr/lib/jvm$ java -version
java version "1.8.0_162"
Java(TM) SE Runtime Environment (build 1.8.0_162-b12)
Java HotSpot(TM) 64-Bit Server VM (build 25.162-b12, mixed mode)
```

查询相关资料，运用上述命令语句赋予权限，从而实现 java 的成功安装，可知其版本为 1.8.0_162。

二、进行 Hadoop 伪分布式安装

```
hadoop@u-virtual-machine:/usr/lib/jvm$ sudo tar -zxf /home/hadoop/桌面/hadoop-3.1.
3.tar.gz -C /usr/local
[sudo] hadoop 的密码：
hadoop@u-virtual-machine:/usr/lib/jvm$ cd /usr/local
hadoop@u-virtual-machine:/usr/local$ ls
bin  etc  games  hadoop-3.1.3  include  lib  man  sbin  share  src
hadoop@u-virtual-machine:/usr/local$ sudo mv ./hadoop-3.1.3/ ./hadoop
hadoop@u-virtual-machine:/usr/local$ ls
bin  etc  games  hadoop  include  lib  man  sbin  share  src
hadoop@u-virtual-machine:/usr/local$ sudo chown -R hadoop ./hadoop
hadoop@u-virtual-machine:/usr/local$ cd /hadoop
bash: cd: /hadoop: 没有那个文件或目录
hadoop@u-virtual-machine:/usr/local$ cd /usr/local/hadoop
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hadoop version
Hadoop 3.1.3
Source code repository https://gitbox.apache.org/repos/asf/hadoop.git -r ba631c436
b806728f8ec2f54ab1e289526c90579
Compiled by ztang on 2019-09-12T02:47Z
Compiled with protoc 2.5.0
From source with checksum ec785077c385118ac91aadde5ec9799
This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-3.1
.3.jar
hadoop@u-virtual-machine:/usr/local/hadoop$
```

将 Hadoop 安装在/usr/local 目录下，因此首先将安装包解压到/usr/local 中，然后将文件名 hadoop-3.1.3 修改为 hadoop，并修改文件的权限，将./hadoop 目录下的所有文件和子目录的所有者修改为 hadoop 用户。由于 Hadoop 解压后即可使用，所以只需通过检查其版本来确定是否安装成功。

```
hadoop@u-virtual-machine:/usr/local/hadoop$ mkdir ./input
hadoop@u-virtual-machine:/usr/local/hadoop$ cp ./etc/hadoop/*.xml ./input
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapred
uce/hadoop-mapreduce-examples-3.1.3.jar grep ./input ./output 'dfs[a-z.]+'
2024-03-05 21:23:06,838 INFO impl.MetricsConfig: loaded properties from hadoop-met
rics2.properties
```

```
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=123
        File Output Format Counters
                Bytes Written=23
hadoop@u-virtual-machine:/usr/local/hadoop$ cat ./output/*
1       dfsadmin
hadoop@u-virtual-machine:/usr/local/hadoop$ rm -r ./output
hadoop@u-virtual-machine:/usr/local/hadoop$ rm -r ./input
```

上面安装的 Hadoop 为单机配制，我运行了 grep 以初步体验其功能。在此过程中，首先创建一个 input 文件夹，并将配制文件复制到里面作为输入文件，将结果输出到 output 中，最终查看运行结果。

下面进行 Hadoop 伪分布式的配置：

修改配置文件 core-site.xml ：

```
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20     <property>
21         <name>hadoop.tmp.dir</name>
22         <value>file:/usr/local/hadoop/tmp</value>
23         <description>Abase for other temporary directories.</description>
24     </property>
25     <property>
26         <name>fs.defaultFS</name>
27         <value>hdfs://localhost:9000</value>
28     </property>
29 </configuration>
```

帮助

修改配置文件 hdfs-site.xml：

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>file:/usr/local/hadoop/tmp/dfs/name</value>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>file:/usr/local/hadoop/tmp/dfs/data</value>
    </property>
</configuration>
```

上述配置完成后，执行 NameNode 的格式化：

```
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs namenode -format
WARNING: /usr/local/hadoop/logs does not exist. Creating.
2024-03-05 21:31:18,342 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = u-virtual-machine/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.1.3
STARTUP_MSG:   classpath = /usr/local/hadoop/etc/hadoop:/usr/local/hadoop/shar
doop/common/lib/curator-client-2.13.0.jar:/usr/local/hadoop/share/hadoop/comm
```

开启 NameNode 和 DataNode 守护进程：

```
hadoop@u-virtual-machine:/usr/local/hadoop$ ./sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [u-virtual-machine]
u-virtual-machine: Warning: Permanently added 'u-virtual-machine' (ECDSA) to the l
ist of known hosts.
hadoop@u-virtual-machine:/usr/local/hadoop$ jps
bash: /usr/lib/jvm/jdk1.8.0_162/bin/jps: 权限不够
hadoop@u-virtual-machine:/usr/local/hadoop$ chmod +x /usr/local/jdk8/bin/jps
chmod: 无法访问 '/usr/local/jdk8/bin/jps': 没有那个文件或目录
hadoop@u-virtual-machine:/usr/local/hadoop$ cd /usr/lib/jvm
hadoop@u-virtual-machine:/usr/lib/jvm$ chmod +x jdk8/bin/jps
chmod: 无法访问 'jdk8/bin/jps': 没有那个文件或目录
hadoop@u-virtual-machine:/usr/lib/jvm$ chmod +x jdk1.8.0_162/bin/jps
hadoop@u-virtual-machine:/usr/lib/jvm$ jps
16912 DataNode
16741 NameNode
17288 Jps
17116 SecondaryNameNode
hadoop@u-virtual-machine:/usr/lib/jvm$
```

此时仍然出现了权限不够的问题，解决原理方法同上。

下面进行 Hadoop 伪分布式实例演示：

```
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs dfs -mkdir -p /user/hadoop
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs dfs -mkdir input
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs dfs -put ./etc/hadoop/*.xml
 input
```

由于伪分布式读取的是 HDFS 上的数据，所以要使用 HDFS，首先需要在 HDFS 中创建用户目录首先创建。另外，将 ./etc/hadoop 中的 xml 文件作为输入文件复制到分布式文件系统中，即将 /usr/local/hadoop/etc/hadoop 复制到分布式文件系统中的 /user/hadoop/input 中。

```
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs dfs -ls input
Found 9 items
-rw-r--r--   1 hadoop supergroup       8260 2024-03-05 22:00 input/capacity-schedu
ler.xml
-rw-r--r--   1 hadoop supergroup       1075 2024-03-05 22:00 input/core-site.xml
-rw-r--r--   1 hadoop supergroup      11392 2024-03-05 22:00 input/hadoop-policy.x
ml
-rw-r--r--   1 hadoop supergroup       1133 2024-03-05 22:00 input/hdfs-site.xml
-rw-r--r--   1 hadoop supergroup        620 2024-03-05 22:00 input/httpfs-site.xml
-rw-r--r--   1 hadoop supergroup       3518 2024-03-05 22:00 input/kms-acls.xml
-rw-r--r--   1 hadoop supergroup        682 2024-03-05 22:00 input/kms-site.xml
-rw-r--r--   1 hadoop supergroup        758 2024-03-05 22:00 input/mapred-site.xml
-rw-r--r--   1 hadoop supergroup        690 2024-03-05 22:00 input/yarn-site.xml
```

通过命令行查看文件列表，发现一共有九个文件。

```
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapred
uce/hadoop-mapreduce-examples-3.1.3.jar wordcount input output
```

用 wordcount 对文件中的单词进行计数。

```
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs dfs -cat output/*
2024-03-05 22:35:16,586 INFO sasl.SaslDataTransferClient: SASL encryption trust ch
eck: localHostTrusted = false, remoteHostTrusted = false
"*"     21
"AS     9
"License");     9
"alice,bob      21
"clumping"      1
(ASF)   1
(root   1
(the    9
-->     18
-1      1
-1,     1
0.0     1
1-MAX_INT.      1
1.      1
1.0.    1
2.0     9
40      2
40+20=60        1
:       2
```

对运行结果进行了查看，由于结果太长，所以只截图了一部分以供参考。

三、使用 Linux 系统中的 Shell 命令进行常用的 Hadoop 操作

由于 Hadoop 用户的登录，启动 Hadoop，为 Hadoop 用户在 HDFS 中创建用户目录“/user/hadoop”等操作在前面的实验中已经涉及，并且进行了相关的操作以及截图，所以在此便不再赘述。

下面在 HDFS 的目录“/user/hadoop”下，创建 test 文件夹，并查看文件列表。由于此时 test 文件夹下没有文件，所以查看文件列表时并没有返回任何文件。

```
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs dfs -mkdir test
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs dfs -ls test
```

将 Linux 操作系统本地的“~/.bashrc”文件上传到 HDFS 的 test 文件夹中，并查看 test 目录下有哪些文件，发现只有我们上传的一个文件。

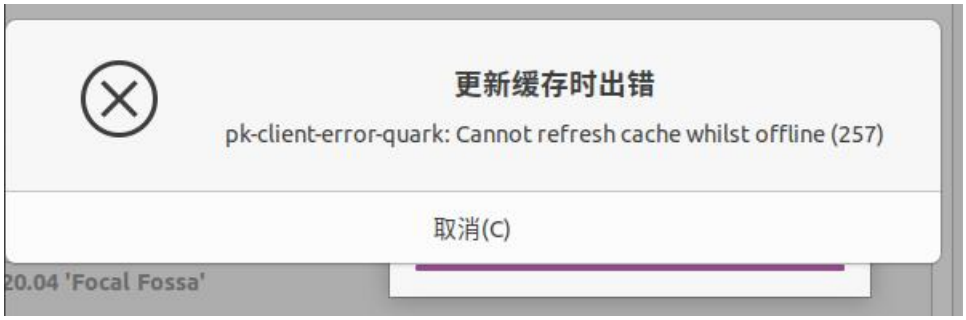
```
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs dfs -put ~/.bashrc test/
2024-03-05 22:44:29,533 INFO sasl.SaslDataTransferClient: SASL encryption trust ch
eck: localHostTrusted = false, remoteHostTrusted = false
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs dfs -ls test
Found 1 items
-rw-r--r--   1 hadoop supergroup       3934 2024-03-05 22:44 test/.bashrc
```

将 HDFS 上的 test 文件夹复制到 Linux 操作系统本地文件系统的 “/usr/local/hadoop”目录下并进行了查看。

```
hadoop@u-virtual-machine:/usr/local/hadoop$ ./bin/hdfs dfs -get test
2024-03-05 23:06:21,374 INFO sasl.SaslDataTransferClient: SASL encryption trust ch
eck: localHostTrusted = false, remoteHostTrusted = false
hadoop@u-virtual-machine:/usr/local/hadoop$ ls /usr/local/hadoop/
bin  etc  include  lib  libexec  LICENSE.txt  logs  NOTICE.txt  README.txt  sbin
share  test  tmp
```

出现的问题：

1、apt 更新未成功，换源出现如下报错：



2、vim 安装未成功。（最终采用 gedit）

```
hadoop@u-virtual-machine:~$ sudo apt-get install vim
[sudo] hadoop 的密码：
正在读取软件包列表... 完成
正在分析软件包的依赖关系树
正在读取状态信息... 完成
没有可用的软件包 vim，但是它被其它的软件包引用了。
这可能意味着这个缺失的软件包可能已被废弃，
或者只能在其他发布源中找到

E: 软件包 vim 没有可安装候选
hadoop@u-virtual-machine:~$
```

3、权限不够

解决方案（列出遇到的问题和解决办法，列出没有解决的问题）：

问题 1 和问题 2 在网上查询了很多方法，尝试了很多遍目前依然没有解决，但不影响后面的实验进程及结果。问题 3 已经解决，重新授权即可，在前文已经详细说明。