南京信息工程大学　实验（实习）报告

实验（实习）名称　实验四：Hadoop 伪分布式模式环境搭建　（实习）日期　5-18　得分＿＿＿＿＿＿

指导教师　孙乐　专业　软件工程（中外合作办学）　年级　2020 级　班次　2

姓名　颜晓雨　学号　202083020070

实验目的：学习分布式计算框架 hadoop 的使用

实验内容：

1. HDFS 及 YARN 环境搭建
2. Hadoop 集群启动
3. 测试 MapReduce 案例

实验环境：Linux 系统、VMware

实验步骤：

1.HDFS 及 YARN 环境搭建

core-site.xml 文件是 Hadoop 的核心配置文件，在这里需指定几个配置项，其中必须指定的是访问的端口；每一个配置项都有其默认值；使用 vim 编辑文件

```
[root@localhost hadoop]# vim core-site.xml
```

修改成如下图所示，保存退出：

```
<configuration>
<!-- 指定 NameNode 的地址 -->
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://hadoop:9820</value>
</property>
<!-- 指定 hadoop 数据的存储目录 -->
    <property>
        <name>hadoop.tmp.dir</name>
        <value>/opt/module/hadoop-3.1.3/data</value>
   </property>
<!-- 配置 HDFS 网页登录使用的静态用户为 root -->
   <property>
        <name>hadoop.http.staticuser.user</name>
        <value>root</value>
  </property>
</configuration>
```

将信息修改成如下图所示，保存退出即可。

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->


<configuration>
<!-- 指定NameNode的地址 -->
     <property>
        <name>fs.defaultFS</name>
        <value>hdfs://hadoop:9820</value>
    </property>
<!-- 指定hadoop数据的存储目录 -->
    <property>
        <name>hadoop.tmp.dir</name>
        <value>/opt/module/hadoop-3.1.3/data</value>
    </property>
<!-- 配置HDFS网页登录使用的静态用户为root -->
    <property>
        <name>hadoop.http.staticuser.user</name>
        <value>root</value>
   </property>
</configuration>
~
~
~
~
~
~
~
~
~
"/opt/module/hadoop-3.1.3/etc/hadoop/core-site.xml" 36L, 1246C
```

2. SSH 免密码登录配置

① 生成密钥

```
[root@hadoop ~]# ssh-keygen -t rsa
```

整个过程一直回车即可：

```
[root@hadoop ~]# ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/root/.ssh/id_rsa):
```

② 拷贝公钥：

这里 root 代表用户名， hadoop 代表 ip 地址，因为我们已经将 hadoop 名称与 ip 地址做了映射。

```
[hadoop@centos ~]$ ssh-copy-id root@hadoop
# 输入一次 root 用户的密码即可通过验证
```

3.hadoop 集群启动

在 hadoop 集群启动之前，先做一个格式化的操作，格式化主要是去生成集群自己需要的信息，比如主节点的唯一 id 等，主要是通过 id 来确认哪些主机归集群管理，也就是说是通过 id 来与主机进行绑定和对应的；

1) 格式化 namenode

第一次使用 Hadoop 时需要进行初始化，该操作只需要执行一次，完成后会根据 core-site.xml 中的配置，在对应的目录下自动创建相应的文件夹

```
[root@hadoop ~]# hdfs namenode -format
```

2) 启动 Hadoop：

启动 hadoop 前，我们可以使用 jps（Java Virtual Machine Process Status Tool）命令查看一下当前系统有哪些 java 进程

```
[root@hadoop ~]# jps
```

```
[root@hadoop ~]# jps
8614 Jps
```

a) 启动 HDFS：

```
[root@hadoop ~]# start-dfs.sh
```

```
[root@hadoop ~]# start-dfs.sh
Starting namenodes on [hadoop]
上一次登录：六 10月 22 22:28:46 CST 2022pts/1 上
Starting datanodes
上一次登录：六 10月 22 22:34:32 CST 2022pts/1 上
Starting secondary namenodes [hadoop]
上一次登录：六 10月 22 22:34:34 CST 2022pts/1 上
```

使用 jps 命令查看，发现多了两个进程：

```
[root@hadoop ~]# jps
8898 NameNode
9478 Jps
9309 SecondaryNameNode
```

TIPs：可能会报如下错误：

```
[root@hadoop hadoop-3.1.3]# start-dfs.sh
Starting namenodes on [hadoop]
ERROR: Attempting to operate on hdfs namenode as root
ERROR: but there is no HDFS_NAMENODE_USER defined. Aborting operation.
Starting datanodes
ERROR: Attempting to operate on hdfs datanode as root
ERROR: but there is no HDFS_DATANODE_USER defined. Aborting operation.
Starting secondary namenodes [hadoop]
ERROR: Attempting to operate on hdfs secondarynamenode as root
ERROR: but there is no HDFS_SECONDARYNAMENODE_USER defined. Aborting operation.
```

解决办法：在/etc/profile.d/my_env.sh 文件后面添加如下内容：

```
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
```

最终环境变量文件如下：

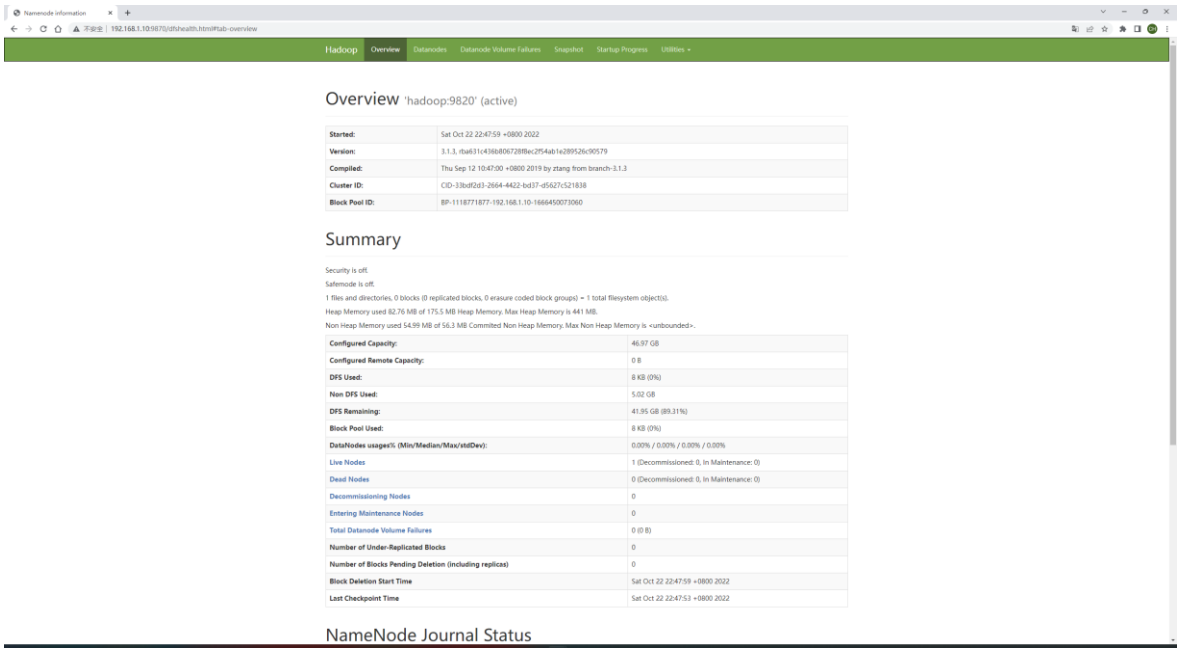
```
#JAVA_HOME
export JAVA_HOME=/opt/module/jdk1.8.0_212
export PATH=$PATH:$JAVA_HOME/bin
#HADOOP_HOME
export HADOOP_HOME=/opt/module/hadoop-3.1.3
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin

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
```

让环境变量生效：

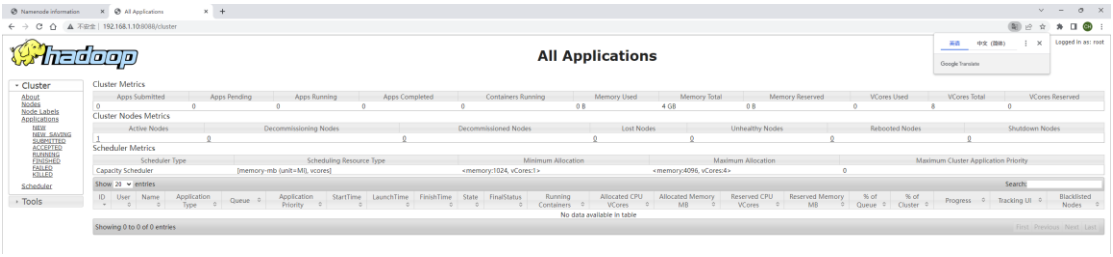
```
[root@hadoop ~]# source /etc/profile
```

修改好后再次启动 hdfs 即可。

4.访问 UI 界面



5.官方 WordCount 测试案例



1) 切换路径到`/opt/module/hadoop-3.1.3/share/hadoop/mapreduce/`目录下，

```
[root@hadoop ~]# cd /opt/module/hadoop-3.1.3/share/hadoop/mapreduce/
```

```
[root@hadoop ~]# cd /opt/module/hadoop-3.1.3/share/hadoop/mapreduce/
[root@hadoop mapreduce]# ll
总用量 5576
-rw-r--r--. 1 haihc haihc  612175 9月  12 2019 hadoop-mapreduce-client-app-3.1.3.jar
-rw-r--r--. 1 haihc haihc  804003 9月  12 2019 hadoop-mapreduce-client-common-3.1.3.jar
-rw-r--r--. 1 haihc haihc 1655414 9月  12 2019 hadoop-mapreduce-client-core-3.1.3.jar
-rw-r--r--. 1 haihc haihc  215372 9月  12 2019 hadoop-mapreduce-client-hs-3.1.3.jar
-rw-r--r--. 1 haihc haihc   45334 9月  12 2019 hadoop-mapreduce-client-hs-plugins-3.1.3.jar
-rw-r--r--. 1 haihc haihc   85396 9月  12 2019 hadoop-mapreduce-client-jobclient-3.1.3.jar
-rw-r--r--. 1 haihc haihc 1659884 9月  12 2019 hadoop-mapreduce-client-jobclient-3.1.3-tests.jar
-rw-r--r--. 1 haihc haihc  126143 9月  12 2019 hadoop-mapreduce-client-nativetask-3.1.3.jar
-rw-r--r--. 1 haihc haihc   97155 9月  12 2019 hadoop-mapreduce-client-shuffle-3.1.3.jar
-rw-r--r--. 1 haihc haihc   57652 9月  12 2019 hadoop-mapreduce-client-uploader-3.1.3.jar
-rw-r--r--. 1 haihc haihc  316382 9月  12 2019 hadoop-mapreduce-examples-3.1.3.jar
drwxr-xr-x. 2 haihc haihc    4096 9月  12 2019 jdiff
drwxr-xr-x. 2 haihc haihc      57 9月  12 2019 lib
drwxr-xr-x. 2 haihc haihc      30 9月  12 2019 lib-examples
drwxr-xr-x. 2 haihc haihc    4096 9月  12 2019 sources
```

2) 在 mapreduce 下创建一个测试文件 data.txt

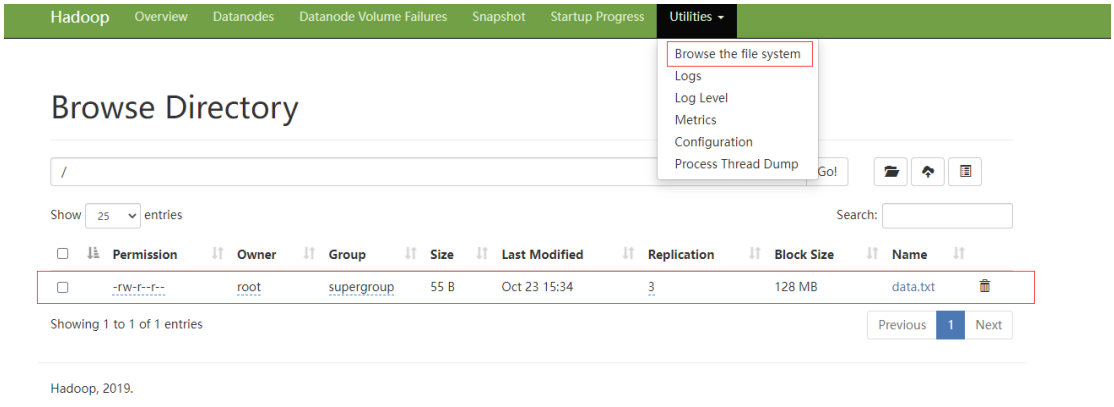
```
hello hadoop
hello java
hello spark
flume flink hadoop
```

```
[root@hadoop mapreduce]# vim data.txt
[root@hadoop mapreduce]# cat data.txt
hello hadoop
hello java
hello spark
flume flink hadoop
```

3) 将此文件上传至 HDFS 文件系统内

```
[root@hadoop mapreduce]# hadoop fs -put data.txt /
```

在 HDFS Web UI 界面上也能看到我们上传了一个文件 data.txt:



4) 执行作业，测试 wordcount：

```
[root@hadoop mapreduce]# hadoop jar
hadoop-mapreduce-examples-3.1.3.jar wordcount /data.txt /output
```

5) 查看结果：

```
[root@hadoop mapreduce]# hadoop fs -cat /output/part-r-00000
```

```
[root@hadoop mapreduce]# hadoop fs -cat /output/part-r-00000
2022-10-23 15:40:44,198 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
flink   1
flume   1
hadoop  2
hello   3
java    1
spark   1
```

在 HDFS Web UI 界面上也能看到多了个文件。

6.hadoop 集群关闭

Hdfs 和 yarn 需要分别关闭（无先后顺序）

1) Hdfs 的关闭

```
[root@hadoop ~]# stop-dfs.sh
```

2) Yarn 的关闭

```
[root@hadoop ~]# stop-yarn.sh
```

关虚拟机前记得把 hadoop 集群关了，不然下次启动可能会报错。

检查下有没有关闭：

```
[root@hadoop ~]# jps
```

```
[root@hadoop ~]# jps
11672 Jps
```

实验时间：2 机时