

电 子 科 技 大 学

实 验 报 告

学生姓名：Lolipop 学号：2018091202000 指导教师：xx

实验地点：信软学院楼西 400 实验时间：2020.11.18

一、实验名称：Hadoop 伪分布式环境模式

二、实验学时：4 学时

三、实验目的：

1. 掌握 Hadoop 的伪分布式环境搭建；
2. 熟悉 Hadoop、Linux 的基本命令。

四、实验原理：

分布式指将项目拆分（按业务或者服务），将项目部署在不同的机器上运行，对机器性能要求下降。伪分布式不是真正的分布式：伪分布式是将多态机器的任务放到一台机器运行。

例如，将淘宝分多模块后，一个模块一个模块放到一台机器中运行，多台机器的時候是同时运行，速度快。一台机器中运行，速度慢、且多个模块不能并行处理，必须得一个任务一个任务地完成，其他任务只能等待。

五、实验内容：

实现 Hadoop 伪分布式环境的搭建，在此基础上进一步学习和使用 Hadoop 与 Linux 的基本命令。

六、实验器材（设备、元器件）：

1. Ubuntu 20.04

2. jdk 1.8
3. hadoop 2.10.1

七、实验步骤：

1. 配置 core-site.xml
2. 配置 yarn-site.xml
3. 创建和配置 mapred-site.xml
4. 配置 hdfs-site.xml
5. 格式化 hdfs
6. 启动 Hadoop
7. 运行 Hadoop 伪分布式实例

八、实验结果与分析（含重要数据结果分析或核心代码流程分析）

1. 配置 core-site.xml

- a) /usr/local/hadoop/etc/hadoop/core-site.xml 包含了 hadoop 启动时的配置信息。编辑器中打开此文件。

```
$ sudo gedit /usr/local/hadoop/etc/hadoop/core-site.xml
```

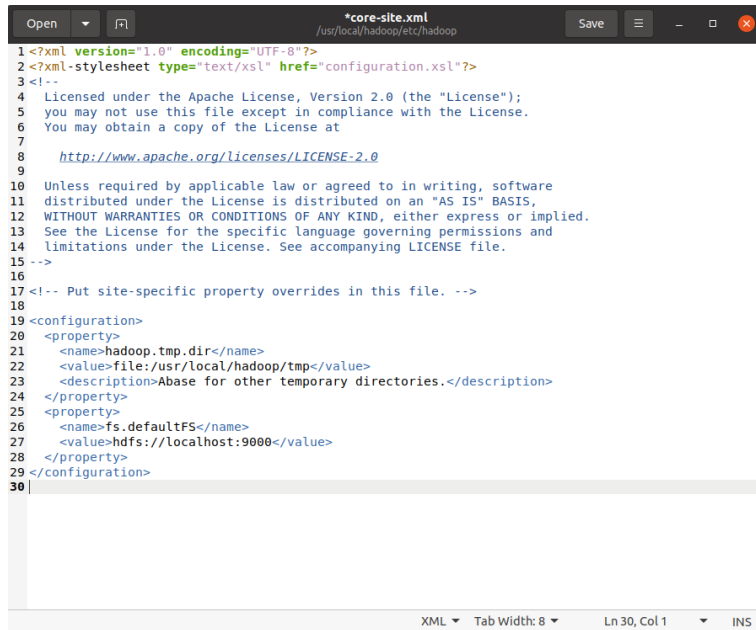
- b) 添加内容并保存。如图 1-1 所示。

2. 配置 yarn-site.xml

- a) /usr/local/hadoop/etc/hadoop/yarn-site.xml 包含了 MapReduce 启动时的配置信息。编辑器中打开此文件。

```
$ sudo gedit /usr/local/hadoop/etc/hadoop/yarn-site.xml
```

- b) 添加内容并保存。如图 2 所示。



```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4 Licensed under the Apache License, Version 2.0 (the "License");
5 you may not use this file except in compliance with the License.
6 You may obtain a copy of the License at
7
8 http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
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>
30
```

图 1 配置 core-site.xml



```
1 <?xml version="1.0"?>
2 <!--
3 Licensed under the Apache License, Version 2.0 (the "License");
4 you may not use this file except in compliance with the License.
5 You may obtain a copy of the License at
6
7 http://www.apache.org/licenses/LICENSE-2.0
8
9 Unless required by applicable law or agreed to in writing, software
10 distributed under the License is distributed on an "AS IS" BASIS,
11 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
12 See the License for the specific language governing permissions and
13 limitations under the License. See accompanying LICENSE file.
14 -->
15 <configuration>
16 <property>
17 <name>yarn.nodemanager.aux-services</name>
18 <value>mapreduce_shuffle</value>
19 </property>
20 <property>
21 <name>yarn.nodemanager.aux-services.mapreduce_shuffle.class</name>
22 <value>org.apache.hadoop.mapred.ShuffleHandler</value>
23 </property>
24 </configuration>
25
```

图 2 配置 yarn-site.xml

3. 创建和配置 mapred-site.xml

- a) 默认情况下, /usr/local/hadoop/etc/hadoop/ 文件夹下有 mapred.xml.template 文件, 复制该文件, 并命名为 mapred.xml, 该文件用于指定 MapReduce

使用的框架。复制并重命名。

```
$ sudo cp mapred-site.xml.template mapred-site.xml
```

b) 编辑器中打开此文件。

```
$ sudo gedit mapred-site.xml
```

c) 添加内容并保存。如图 3 所示。

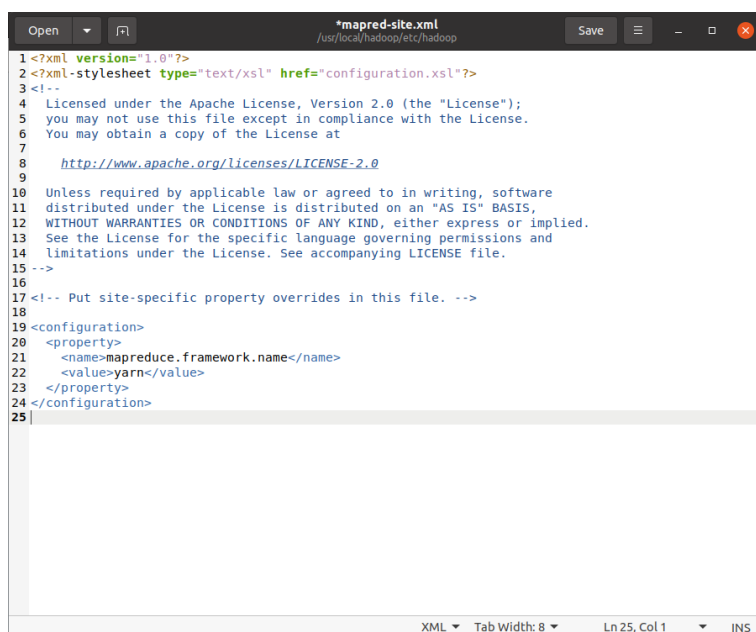


图 3 配置 mapred-site.xml

4. 配置 hdfs-site.xml

a) /usr/local/hadoop/etc/hadoop/hdfs-site.xml 用来配置集群中每台主机都可用, 指定主机上作为 namenode 和 datanode 的目录,在/usr/local/hadoop 下创建文件夹。

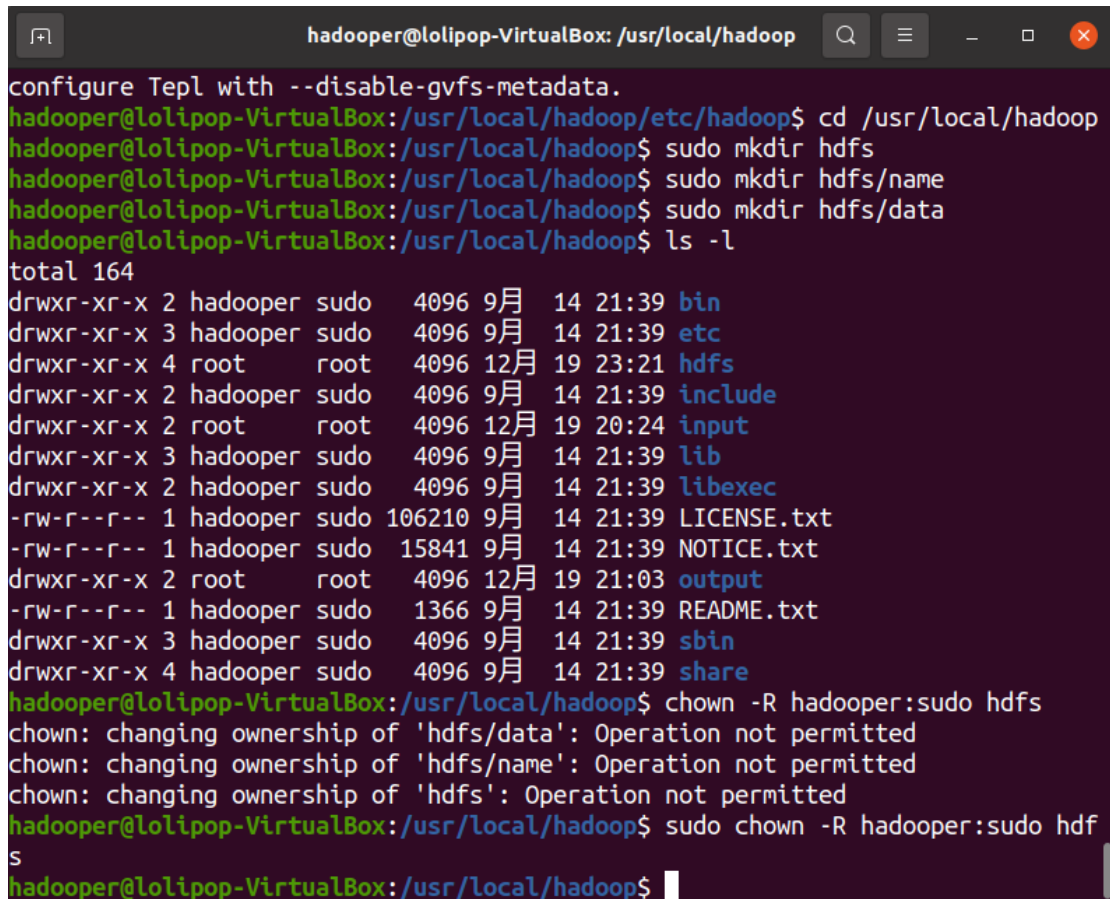
```
$ sudo mkdir hdfs
$ sudo mkdir hdfs/name
$ sudo mkdir hdfs/data
```

b) 检查文件夹授权是否是 hadoop。

```
ls -l
```

- c) 否则执行下述命令。结果如图 4-1 所示。

```
chown -R hadoop:hadoop /usr/local/hadoop/hdfs
```



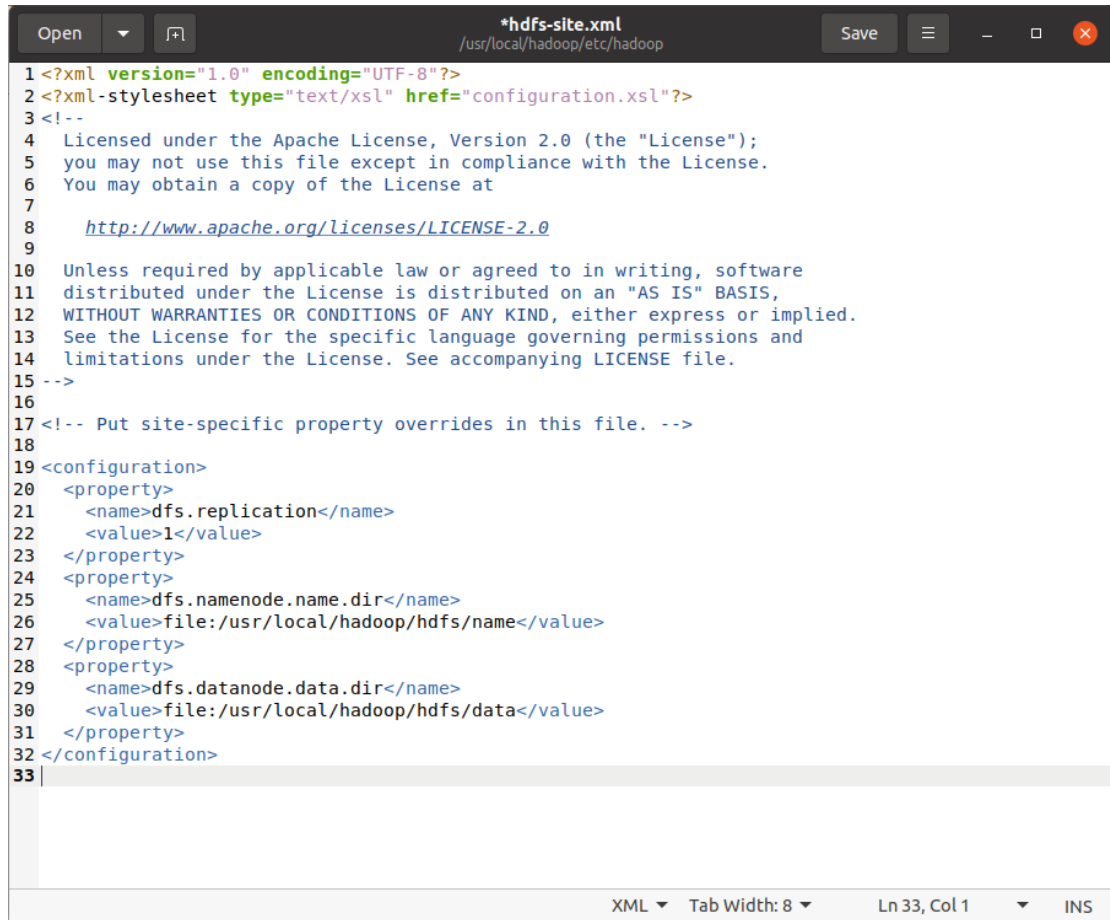
```
hadoop@hadoop-VirtualBox: /usr/local/hadoop
configure Tepl with --disable-gvfs-metadata.
hadoop@hadoop-VirtualBox:/usr/local/hadoop/etc/hadoop$ cd /usr/local/hadoop
hadoop@hadoop-VirtualBox:/usr/local/hadoop$ sudo mkdir hdfs
hadoop@hadoop-VirtualBox:/usr/local/hadoop$ sudo mkdir hdfs/name
hadoop@hadoop-VirtualBox:/usr/local/hadoop$ sudo mkdir hdfs/data
hadoop@hadoop-VirtualBox:/usr/local/hadoop$ ls -l
total 164
drwxr-xr-x 2 hadoop sudo 4096 9月 14 21:39 bin
drwxr-xr-x 3 hadoop sudo 4096 9月 14 21:39 etc
drwxr-xr-x 4 root root 4096 12月 19 23:21 hdfs
drwxr-xr-x 2 hadoop sudo 4096 9月 14 21:39 include
drwxr-xr-x 2 root root 4096 12月 19 20:24 input
drwxr-xr-x 3 hadoop sudo 4096 9月 14 21:39 lib
drwxr-xr-x 2 hadoop sudo 4096 9月 14 21:39 libexec
-rw-r--r-- 1 hadoop sudo 106210 9月 14 21:39 LICENSE.txt
-rw-r--r-- 1 hadoop sudo 15841 9月 14 21:39 NOTICE.txt
drwxr-xr-x 2 root root 4096 12月 19 21:03 output
-rw-r--r-- 1 hadoop sudo 1366 9月 14 21:39 README.txt
drwxr-xr-x 3 hadoop sudo 4096 9月 14 21:39 sbin
drwxr-xr-x 4 hadoop sudo 4096 9月 14 21:39 share
hadoop@hadoop-VirtualBox:/usr/local/hadoop$ chown -R hadoop:hadoop hdfs
chown: changing ownership of 'hdfs/data': Operation not permitted
chown: changing ownership of 'hdfs/name': Operation not permitted
chown: changing ownership of 'hdfs': Operation not permitted
hadoop@hadoop-VirtualBox:/usr/local/hadoop$ sudo chown -R hadoop:hadoop hdfs
hadoop@hadoop-VirtualBox:/usr/local/hadoop$
```

图 4-1 文件夹授权

- d) 编辑器打开 hdfs-site.xml。

```
$ sudo gedit /usr/local/hadoop/etc/hadoop/hdfs-site.xml
```

- e) 添加内容并保存。如图 4-2 所示。



```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4 Licensed under the Apache License, Version 2.0 (the "License");
5 you may not use this file except in compliance with the License.
6 You may obtain a copy of the License at
7
8 http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20   <property>
21     <name>dfs.replication</name>
22     <value>1</value>
23   </property>
24   <property>
25     <name>dfs.namenode.name.dir</name>
26     <value>file:/usr/local/hadoop/hdfs/name</value>
27   </property>
28   <property>
29     <name>dfs.datanode.data.dir</name>
30     <value>file:/usr/local/hadoop/hdfs/data</value>
31   </property>
32 </configuration>
33 |
```

XML Tab Width: 8 Ln 33, Col 1 INS

图 4-2 配置 hdfs-site.xml

5. 格式化 hdfs

- a) 执行一次格式化 hdfs 的命令。结果如图 5-1 和图 5-2 所示。

```
hadoop@lolipop-VirtualBox: /usr/local/hadoop
hadoop@lolipop-VirtualBox:~$ cd /usr/local/hadoop
hadoop@lolipop-VirtualBox:/usr/local/hadoop$ bin/hdfs namenode -format
20/12/19 23:33:25 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = lolipop-VirtualBox/127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 2.10.1
STARTUP_MSG: classpath = /usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/hadoop/common/lib/commons-net-3.1.jar:/usr/local/hadoop/share/hadoop/common/lib/jaxb-api-2.2.2.jar:/usr/local/hadoop/share/hadoop/common/lib/guava-11.0.2.jar:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar:/usr/local/hadoop/share/hadoop/common/lib/jsch-0.1.55.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-configuration-1.6.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-io-2.4.jar:/usr/local/hadoop/share/hadoop/common/lib/gson-2.2.4.jar:/usr/local/hadoop/share/hadoop/common/lib/audience-annotations-0.5.0.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-lang-2.6.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-cli-1.2.jar:/usr/local/hadoop/share/hadoop/common/lib/jettison-1.1.jar:/usr/local/hadoop/share/hadoop/common/lib/curator-recipes-2.13.0.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-codec-1.4.jar:/usr/local/hadoop/share/hadoop/common/lib/jackson-xc-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/jaxb-impl-2.2.3-1.jar:/usr/local/hadoop/share/hadoop/common/lib/jersey-json-1.9.jar:/usr/local/hadoop/share/hadoop/common/lib/java-xmlbuilder-0.4.jar:/usr/local/hadoop/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/httpclient-4.5.2.jar:/usr/local/hadoop/share/hadoop/common/lib/woodstox-core-5.0.3.jar:/usr/local/hadoop/share/hadoop/common/lib/protobuf-java-2.5.0.jar:/usr/local/hadoop/share/hadoop/common/lib/protobuf-java-2.5.0.jar:/usr/local/hadoop/share/hadoop/common/lib/protobuf-java-2.5.0.jar/*****/
```

图 5-1 格式化 hdfs（前半）

```
20/12/19 23:33:26 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
20/12/19 23:33:26 INFO util.GSet: Computing capacity for map NameNodeRetryCache
20/12/19 23:33:26 INFO util.GSet: VM type = 64-bit
20/12/19 23:33:26 INFO util.GSet: 0.029999999329447746% max memory 889 MB = 273.1 KB
20/12/19 23:33:26 INFO util.GSet: capacity = 2^15 = 32768 entries
20/12/19 23:33:26 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1100400170-127.0.1.1-1608392006499
20/12/19 23:33:26 INFO common.Storage: Storage directory /usr/local/hadoop/hdfs/name has been successfully formatted.
20/12/19 23:33:26 INFO namenode.FSImageFormatProtobuf: Saving image file /usr/local/hadoop/hdfs/name/current/fsimage.ckpt_000000000000000000 using no compression
20/12/19 23:33:26 INFO namenode.FSImageFormatProtobuf: Image file /usr/local/hadoop/hdfs/name/current/fsimage.ckpt_000000000000000000 of size 327 bytes saved in 0 seconds .
20/12/19 23:33:26 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
20/12/19 23:33:26 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid = 0 when meet shutdown.
20/12/19 23:33:26 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at lolipop-VirtualBox/127.0.1.1
*****/
hadoop@lolipop-VirtualBox:/usr/local/hadoop$
```

图 5-2 格式化 hdfs（后半）

6. 启动 Hadoop

- a) 检查所有抽到/usr/local/hadoop/下文件的权限，特别是/hdfs/data 和 /hdfs/name，保证为之前所建立的 hadoop 的权限。结果如图 6-1 所示。

- b) 启动单节点的集群。

```
$ ssh hadoop@localhost  
$ sbin/start-dfs.sh
```

- c) 执行 jps 命令，会看到 Hadoop 相关的进程 DataNode, Namenode, SecondaryNamenode。如图 6-2 所示。

```
hadoop@lolipop-VirtualBox:/usr/local/hadoop$ sbin/start-dfs.sh  
Starting namenodes on [localhost]  
hadoop@localhost's password:  
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hadoop  
er-namenode-lolipop-VirtualBox.out  
hadoop@localhost's password:  
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hadoop  
er-datanode-lolipop-VirtualBox.out  
Starting secondary namenodes [0.0.0.0]  
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.  
ECDSA key fingerprint is SHA256:x8z3aaUgwAPV63B47tN2G4CXQeYLPJJJoqGdtrNAZbF0.  
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes  
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hos  
ts.  
hadoop@0.0.0.0's password:  
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-  
hadoop-secondarynamenode-lolipop-VirtualBox.out  
hadoop@lolipop-VirtualBox:/usr/local/hadoop$ jps  
35728 Jps  
35415 DataNode  
35240 NameNode  
35625 SecondaryNameNode  
hadoop@lolipop-VirtualBox:/usr/local/hadoop$
```

图 6-2 查看 Hadoop 相关进程（前半）

- d) 执行下述命令，再执行 jps 命令，会看到 Hadoop 相关的进程:ResourceManager, Nodemanager。如图 6-3 所示。

```
$ sbin/start-yarn.sh
```



```

hadoop@lolipop-VirtualBox:/usr/local/hadoop$ sbin/start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hadoop-reso
urcemanager-lolipop-VirtualBox.out
hadoop@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hadoop
er-nodemanager-lolipop-VirtualBox.out
hadoop@lolipop-VirtualBox:/usr/local/hadoop$ jps
35415 DataNode
35240 NameNode
35625 SecondaryNameNode
36219 Jps
35772 ResourceManager
36108 NodeManager
hadoop@lolipop-VirtualBox:/usr/local/hadoop$

```

图 6-3 查看 Hadoop 相关进程（后半）

e) 浏览器打开 <http://localhost:50070/>, 会看到 hdfs 管理页面, 如图 6-4 所示。

Activities Firefox Web Browser 12月 19 23:38

Namenode Information x +

localhost:50070/dfshealth.html#tab-overview

Started:	Sat Dec 19 23:36:01 +0800 2020
Version:	2.10.1, r1827467c9a56f133025f28557bfc2c562d78e816
Compiled:	Mon Sep 14 21:17:00 +0800 2020 by centos from branch-2.10.1
Cluster ID:	CID-46d9c55d-1c95-445a-89d8-7b022bb89e55
Block Pool ID:	BP-1100400170-127.0.1.1-1608392006499

Summary

Security is off.
Safemode is off.
1 files and directories, 0 blocks = 1 total filesystem object(s).
Heap Memory used 114.37 MB of 159 MB Heap Memory. Max Heap Memory is 889 MB.
Non Heap Memory used 42.34 MB of 43.97 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	97.44 GB
DFS Used:	24 KB (0%)
Non DFS Used:	10.48 GB
DFS Remaining:	81.97 GB (84.12%)
Block Pool Used:	24 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)

图 6-4 hdfs 管理页面

f) 浏览器打开 <http://localhost:8088/>, 会看到 hadoop 进程管理页面, 如图 6-5

所示。

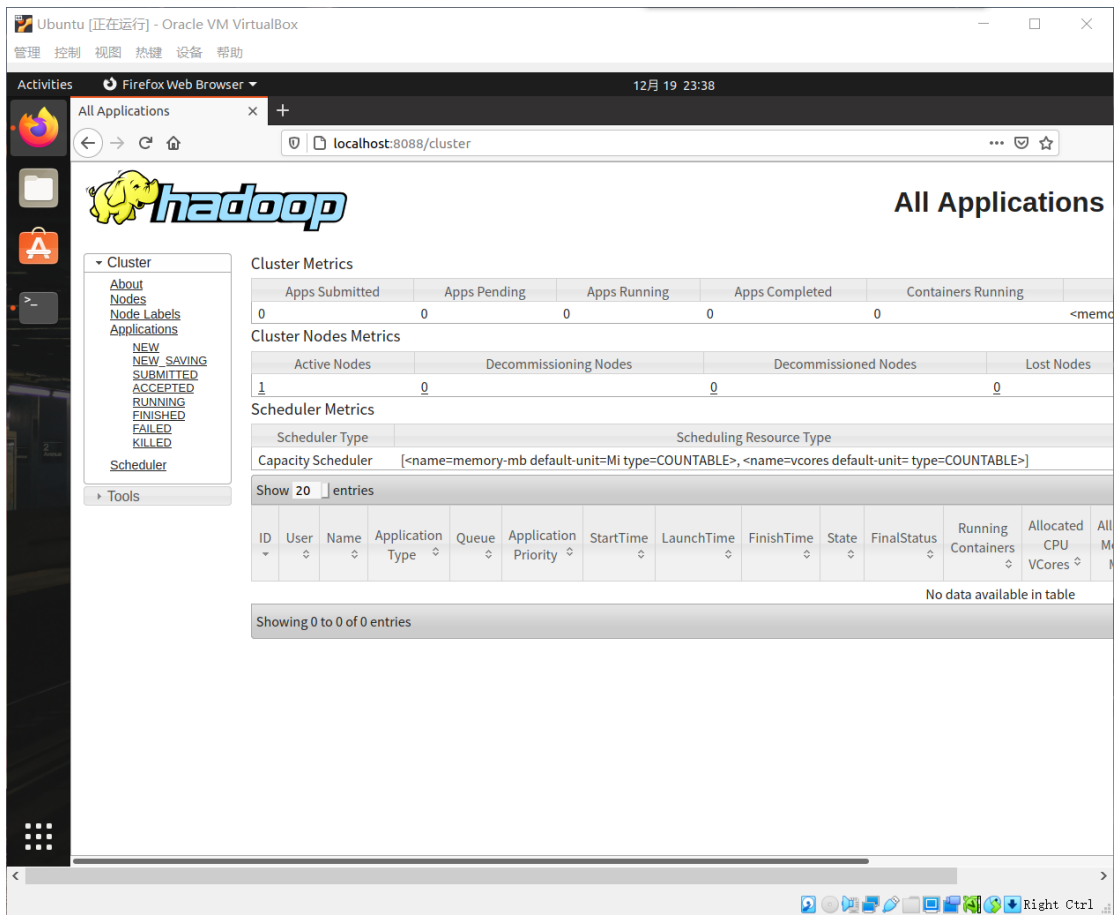


图 6-5 hadoop 进程管理页面

7. 运行 Hadoop 伪分布式实例

- a) 伪分布式读取的是 HDFS 上的数据。要使用 HDFS，首先需要创建用户目录。


```
$ bin/hdfs dfs -mkdir -p /user/hadooper
```

- b) 接着将 input 中的文件作为输入文件复制到分布式文件系统中，即将 /usr/local/ hadoop /input 复制到分布式文件系统上的 /user/ hadoop /input 中。上一步已创建了用户目录 /user/ hadoop ，因此命令中就可以使用相对目录如 input，其对应的绝对路径就是 /user/ hadoop /input。

```
$ bin/hdfs dfs -mkdir input  
$ bin/hdfs dfs -put input/*.txt input
```

- c) 复制完成后，可以通过如下命令查看文件列表。结果如图 7-1 所示。

```
$ bin/hdfs dfs -ls input
```



```
hadoop@lolipop-VirtualBox:/usr/local/hadoop$ bin/hdfs dfs -put input/*.txt i  
nput  
hadoop@lolipop-VirtualBox:/usr/local/hadoop$ bin/hdfs dfs -ls input  
Found 1 items  
-rw-r--r--  1 hadoop supergroup      1366 2020-12-19 23:43 input/README.tx  
t  
hadoop@lolipop-VirtualBox:/usr/local/hadoop$
```

图 7-1 查看复制结果

- d) 伪分布式运行 MapReduce 作业的方式跟单机模式相同，区别在于伪分布式读取的是 HDFS 中的文件。执行命令下述命令并查看运行结果。如图 7-2 和图 7-3 所示。

```
$ bin/hadoop jar  
share/hadoop/mapreduce/sources/hadoop-mapreduce-examples-2.7.3-sources  
.jar org.apache.hadoop.examples.WordCount input output  
$ bin/hdfs dfs -cat output/*
```

- e) Hadoop 运行程序时，默认输出目录不能存在，因此运行需要执行如下命令删除 output 文件夹。如图 7-4 所示。

```
$ bin/hdfs dfs -rm -r /user/zhangsan/output
```

```
hadoop@lolipop-VirtualBox: /usr/local/hadoop
-rw-r--r-- 1 hadoop supergroup 1366 2020-12-19 23:43 input/README.txt
hadoop@lolipop-VirtualBox: /usr/local/hadoop$ bin/hadoop jar share/hadoop/mapreduce/s
sources/hadoop-mapreduce-examples-2.10.1-sources.jar org.apache.hadoop.examples.WordCou
nt input output
20/12/19 23:50:36 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/12/19 23:50:37 INFO input.FileInputFormat: Total input files to process : 1
20/12/19 23:50:38 INFO mapreduce.JobSubmitter: number of splits:1
20/12/19 23:50:38 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16083922
38964_0001
20/12/19 23:50:38 INFO conf.Configuration: resource-types.xml not found
20/12/19 23:50:38 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
20/12/19 23:50:38 INFO resource.ResourceUtils: Adding resource type - name = memory-mb
, units = Mi, type = COUNTABLE
20/12/19 23:50:38 INFO resource.ResourceUtils: Adding resource type - name = vcores, u
nits = , type = COUNTABLE
20/12/19 23:50:39 INFO impl.YarnClientImpl: Submitted application application_16083922
38964_0001
20/12/19 23:50:39 INFO mapreduce.Job: The url to track the job: http://lolipop-Virtual
Box:8088/proxy/application_1608392238964_0001/
20/12/19 23:50:39 INFO mapreduce.Job: Running job: job_1608392238964_0001
20/12/19 23:50:47 INFO mapreduce.Job: Job job_1608392238964_0001 running in uber mode
: false
20/12/19 23:50:47 INFO mapreduce.Job: map 0% reduce 0%
20/12/19 23:50:55 INFO mapreduce.Job: map 100% reduce 0%
20/12/19 23:51:00 INFO mapreduce.Job: map 100% reduce 100%
20/12/19 23:51:02 INFO mapreduce.Job: Job job_1608392238964_0001 completed successfull
y
20/12/19 23:51:03 INFO mapreduce.Job: Counters: 49
    File System Counters
        FILE: Number of bytes read=1836
        FILE: Number of bytes written=421167
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
```

图 7-2 执行测试命令

```
hadoop@lolipop-VirtualBox: /usr/local/hadoop$ bin/hdfs dfs -cat output/*
(BIS), 1
(ECCN) 1
(TSU) 1
(see 1
5D002.C.1, 1
740.13) 1
<http://www.wassenaar.org/> 1
Administration 1
Apache 1
BEFORE 1
BIS 1
Bureau 1
Commerce, 1
Commodity 1
Control 1
Core 1
Department 1
ENC 1
Exception 1
Export 2
For 1
Foundation 1
```

图 7-3 输出测试结果

```
hadoop@lolipop-VirtualBox: /usr/local/hadoop$ bin/hdfs dfs -rm -r /user/hadooper/outp
ut
Deleted /user/hadooper/output
hadoop@lolipop-VirtualBox: /usr/local/hadoop$
```

图 7-4 删除伪分布式文件系统中的 output 文件夹

九、总结及心得体会：

// removed

十、对本实验过程及方法、手段的改进建议：

// removed

报告评分：

指导教师签字：