郑凯琳 205220025

实验 1：单机 Hadoop 系统与程序开发工具

任务 1：单机 Hadoop 系统与 WordCount 程序

系统安装运行情况

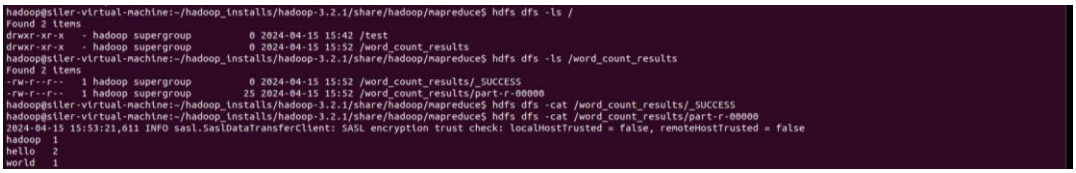
1. 单机操作系统安装
   在 VMWare 中创建 Linux（Ubuntu 22.04.4）虚拟机。
2. 安装 Java
   jdk1.7.0
3. 创建用户
   hadoop
4. 解压安装 Hadoop
   安装的 Hadoop 版本为 3.2.1
5. 配置环境变量

```
PATH=$PATH:$HOME/bin
export JAVA_HOME=/usr/java
export HADOOP_HOME=/home/hadoop/hadoop_installs/hadoop-3.2.1
export PATH=$JAVA_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$PATH
export CLASSPATH=$JAVA_HOME/lib:.
```

6. 免密码 SSH 访问配置
   生成 SSH 认证文件并将密钥复制到 /.ssh/authorized_keys 文件中。
7. 修改 Hadoop 配置文件
   参考材料《深入理解大数据》2.2.5 节配置 Hadoop 环境。
8. 格式化 NameNode
   格式化成功，返回一堆有关 NameNode 的启动信息，其中有一句“…. has been successfully formatted.”。
9. 启动 HDFS 和 MapReduce
   执行 start-all.sh 后用 jps 指令查看进程信息。

```
hadoop@siler-virtual-machine:~$ jps
10288 ResourceManager
9716 NameNode
10725 Jps
10405 NodeManager
10085 SecondaryNameNode
9898 DataNode
```

10. 停止 HDFS 和 MapReduce
    执行 stop-all.sh。
11. 运行 WordCount 测试：
    file1.txt：hello hadoop hello world
    file2.txt：goodbye hadoop



实验数据说明（下载的什么网页数据，多少个 HTML 或 text 文件）

选自维基百科三个页面进行统计，地址如下：

https://en.wikipedia.org/wiki/Chair

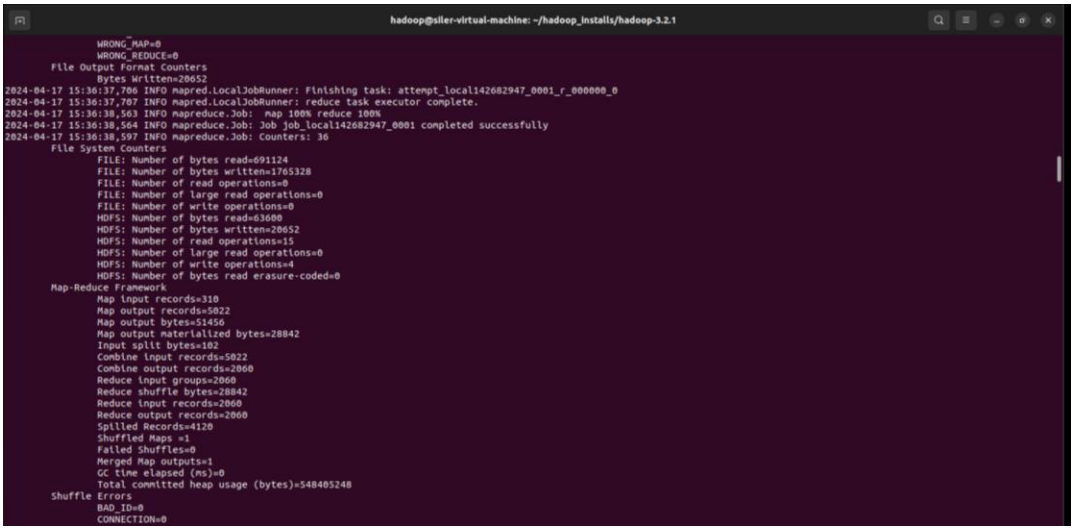
https://en.wikipedia.org/wiki/Table_(furniture)#

https://en.wikipedia.org/wiki/Couch#Image_gallery

其中，我将它们的文字部分 crtl + A 复制到了三个记事本里，分别是 chair.txt，table.txt 和 couch.txt。

实验输出结果开头部分的屏幕拷贝

1. result1（chair）

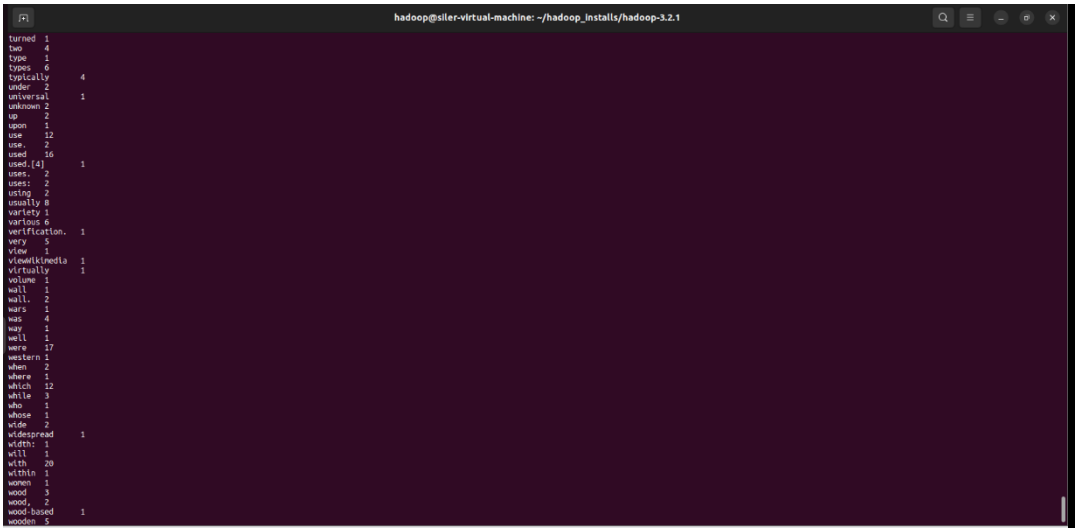


```
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=31800
        File Output Format Counters
                Bytes Written=20652
hadoop@siler-virtual-machine:~/hadoop_installs/hadoop-3.2.1/share/hadoop/mapreduce$ hdfs dfs -ls /result1
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2024-04-17 15:36 /result1/_SUCCESS
-rw-r--r--   1 hadoop supergroup      20652 2024-04-17 15:36 /result1/part-r-00000
hadoop@siler-virtual-machine:~/hadoop_installs/hadoop-3.2.1/share/hadoop/mapreduce$ hdfs dfs -cat /result1/part-r-0000
cat: `/result1/part-r-0000': No such file or directory
hadoop@siler-virtual-machine:~/hadoop_installs/hadoop-3.2.1/share/hadoop/mapreduce$ hdfs dfs -cat /result1/part-r-00000
2024-04-17 15:37:37,461 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
"06-12a.html".  1
"15.5.1.4".     1
"26     1
"A      1
"Anthropocene:  1
"Architecture   1
"Armchair".     1
"BIFMA".        1
"Basic  1
"CGSB   1
"Chair" 1
"Chair".        1
"Christo:       1
"Collection     2
"Definitions    1
"Eisteddfod".   1
"Endowed        1
"Ergonomic      1
"From   1
```

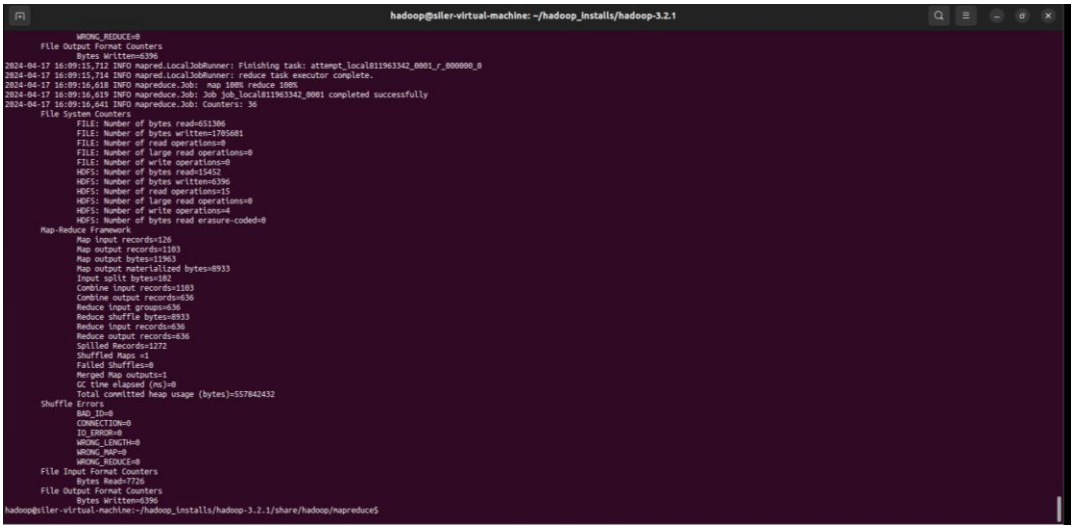
```
hadoop@siler-virtual-machine: ~/hadoop_installs/hadoop-3.2.1
Broken  1
Buddhism        1
Buddhist        2
Business        1
By      6
C.      1
CAN/CGSB        2
CRC     1
Cambridge       1
Canada's        1
Canada,[7]      1
Canadian        3
Caning, 1
Car     1
Caster  1
Categories:     1
Center  1
Central 1
Centre  2
Chair   15
Chair". 2
Chair,  5
Chair.[citation 1
Chair:  1
Chair?"".       1
Chair?".        1
Chairs  10
Chairs".        2
Chairs, 2
Chairs. 4
Chairs: 1
ChairsFurniture 1
Chaise.[35]     1
Chamber 1
China   3
China,  1
China.  1
China:  1
Chinese 2
Christian       1
Christo 3
Cirulis,        1
```

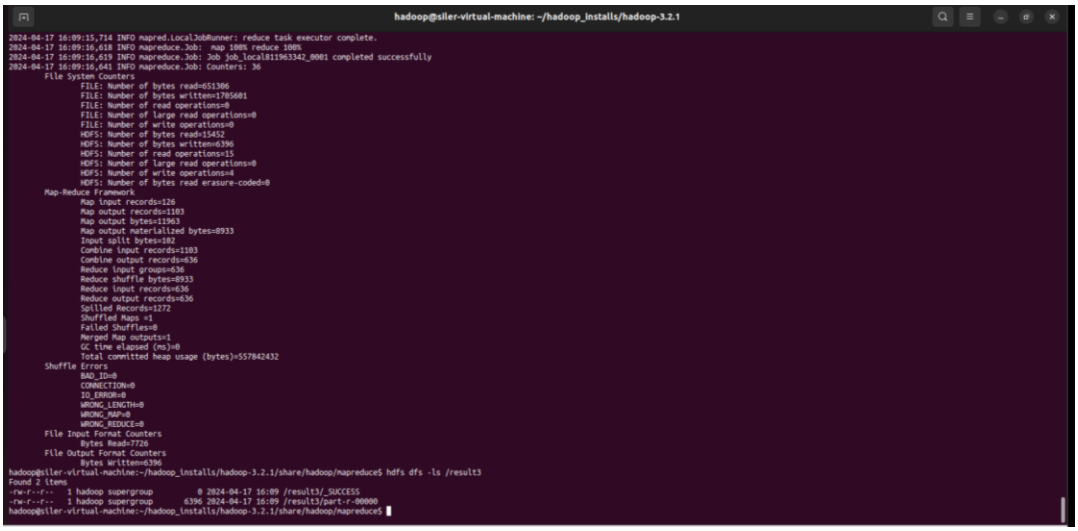
2. result2（table）

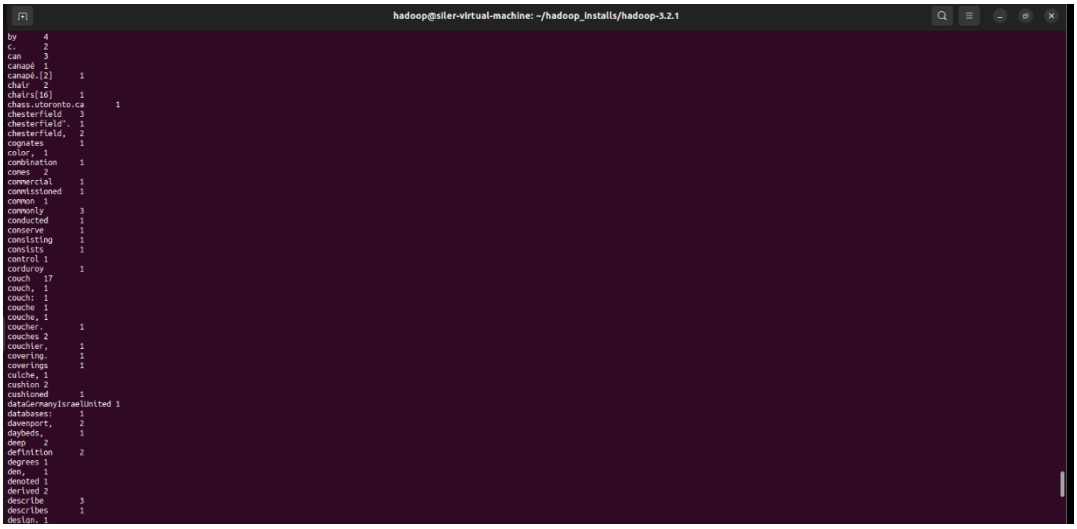
```
hadoop@siler-virtual-machine: ~/hadoop_installs/hadoop-3.2.1
standing        2
standing,       1
statementMobile 1
steel   2
stone   2
stool   1
stool.  1
stools  1
storage,        1
stored  2
storing 1
stretchers.     2
sturdy  1
style,  1
subject.        1
subsection      1
such    2
support.        2
supported       2
supporting      1
supports        1
surface 5
surface,        1
surfaces        2
tabele, 1
table   49
table"  1
table)  1
table),1
table,  6
table.  8
table.[7]       1
table:  1
table;  2
tables  59
tables) 1
tables, 10
tables. 2
tables[8]       1
tabletop        2
tabula  1
talented        1
talk    1
tall    1
tall,   1
tea     2
tea,    2
technical       1
template        1
tennis  3
terms   1
```

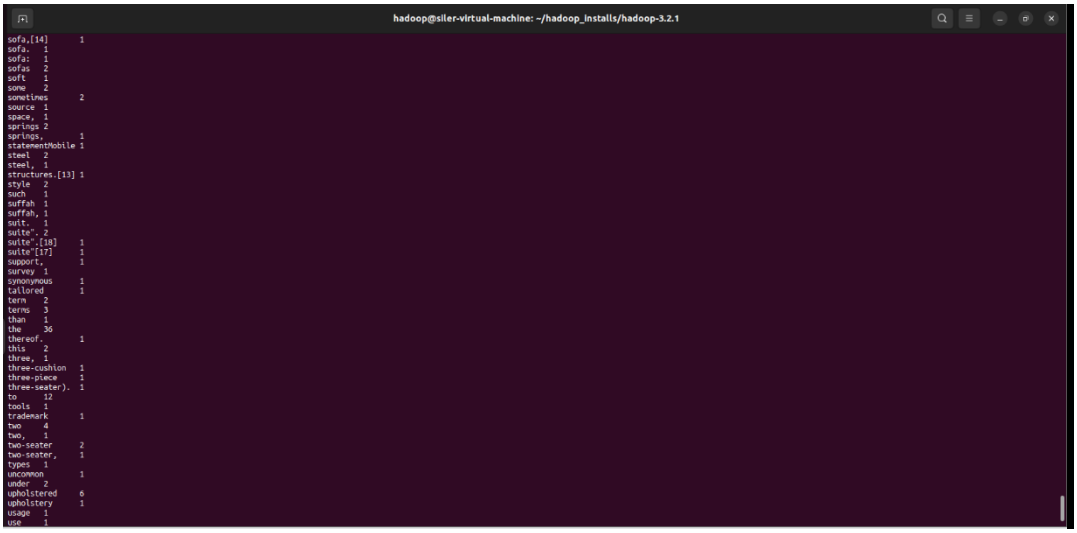


3. result3（couch）

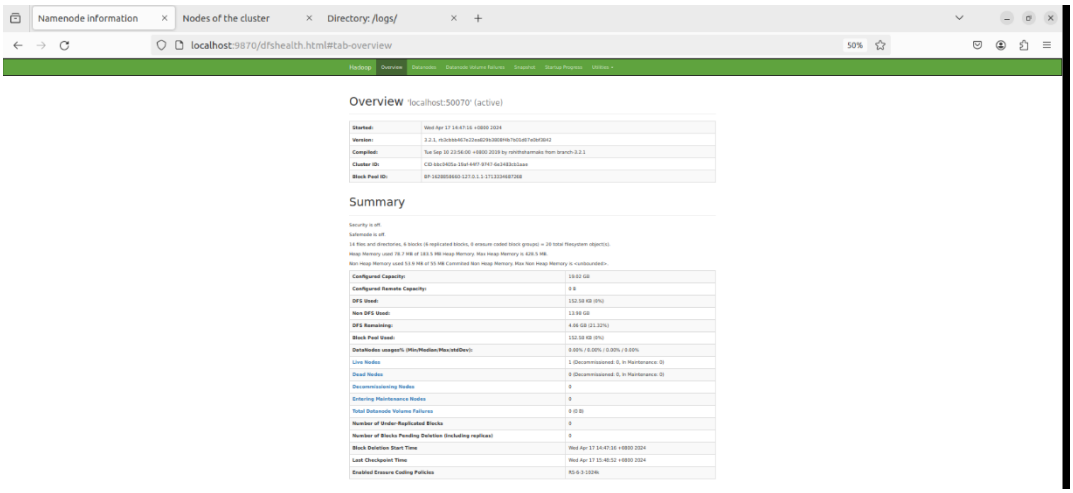


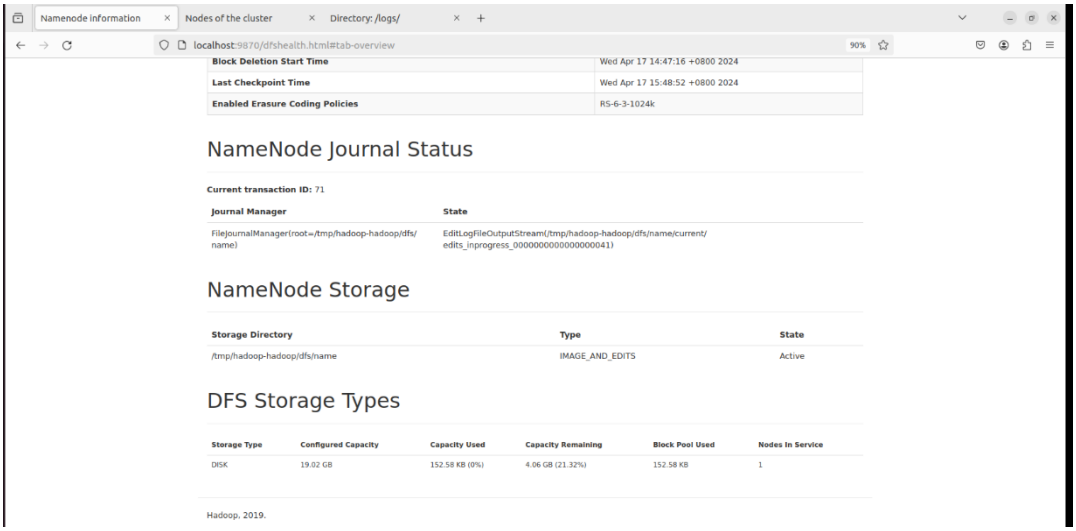


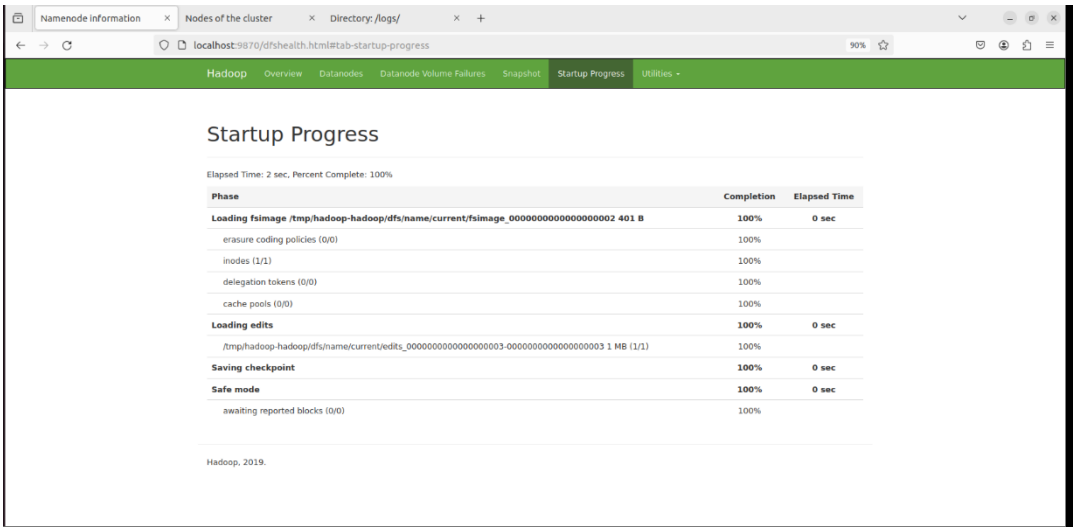


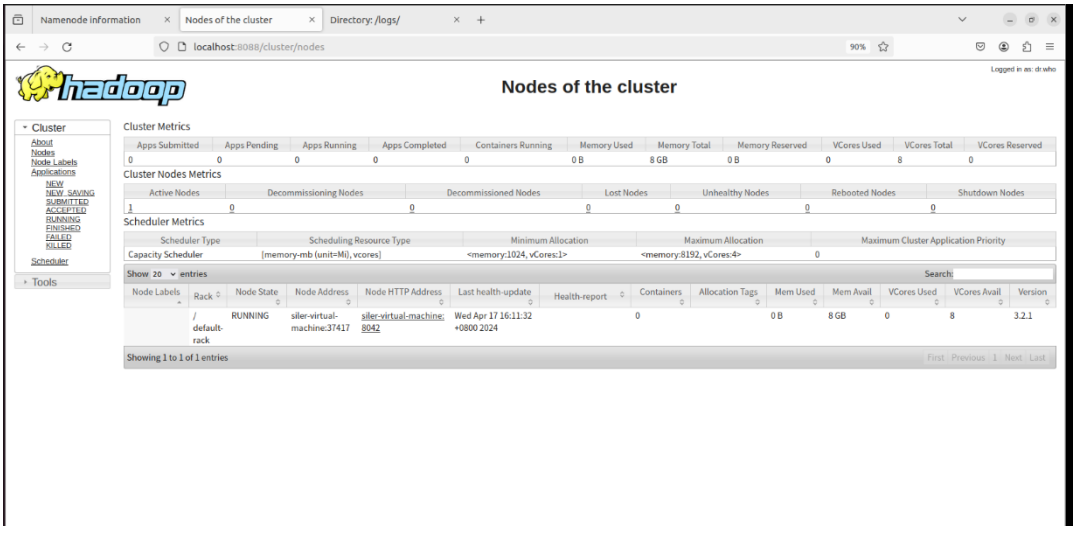


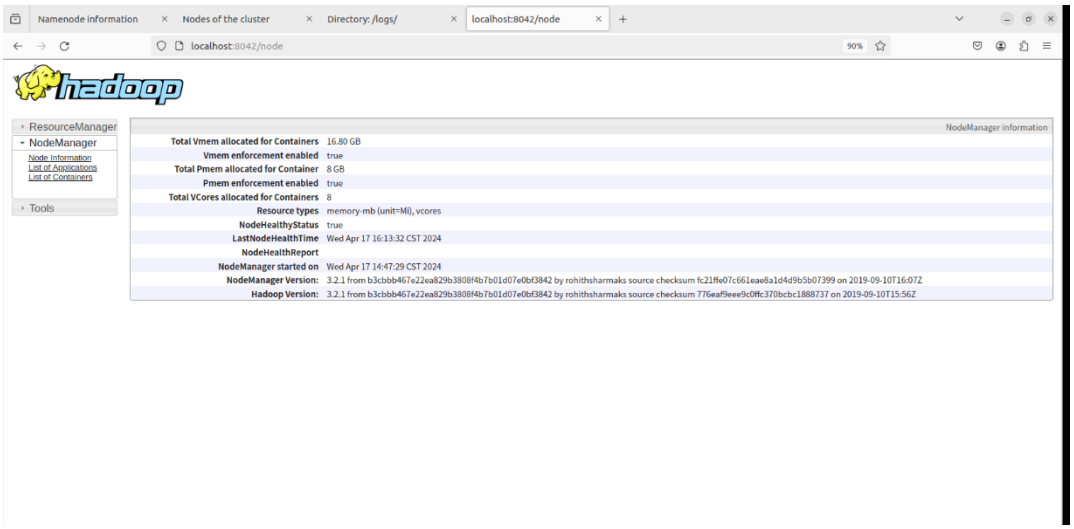
程序运行后在 Hadoop Web 作业状态查看界面上的作业运行状态屏幕拷贝

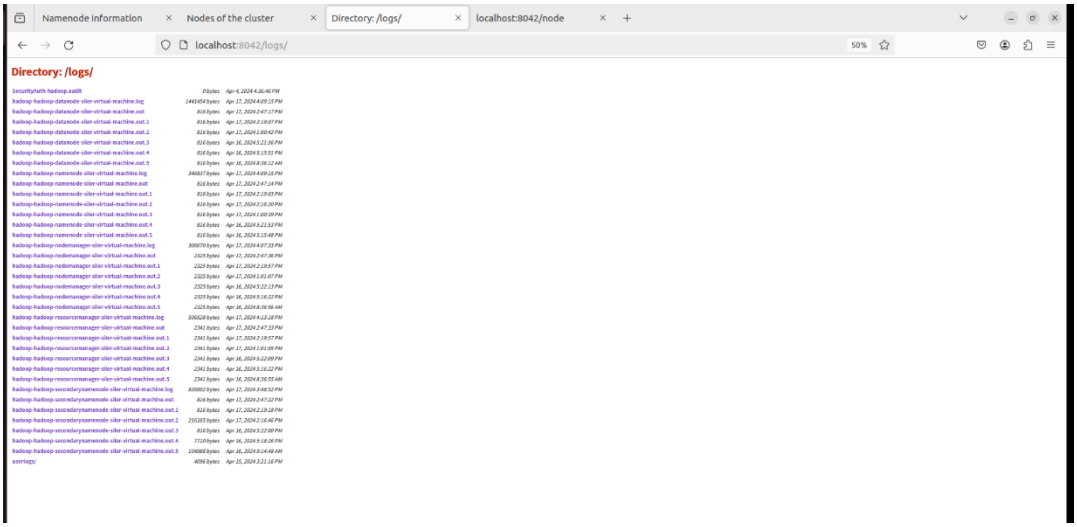



Namenode information
Nodes of the cluster
Directory: /logs/
localhost:9870/dfshealth.html#tab-overview
Block Deletion Start Time
Wed Apr 17 14:47:16 +0800 2024
Last Checkpoint Time
Wed Apr 17 15:48:52 +0800 2024
Enabled Erasure Coding Policies
RS-6-3-1024k
NameNode Journal Status
Current transaction ID: 71
Journal Manager
State
FileJournalManager(root=/tmp/hadoop-hadoop/dfs/name)
EditLogFileOutputStream(/tmp/hadoop-hadoop/dfs/name/current/edits_inprogress_0000000000000000041)
NameNode Storage
Storage Directory
Type
State
/tmp/hadoop-hadoop/dfs/name
IMAGE_AND_EDITS
Active
DFS Storage Types
Storage Type
Configured Capacity
Capacity Used
Capacity Remaining
Block Pool Used
Nodes In Service
DISK
19.02 GB
152.58 KB (0%)
4.06 GB (21.32%)
152.58 KB
1
Hadoop, 2019.


Namenode information
Nodes of the cluster
Directory: /logs/
localhost:9870/dfshealth.html#tab-startup-progress
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Startup Progress
Elapsed Time: 2 sec, Percent Complete: 100%
Phase
Completion
Elapsed Time
Loading fsimage /tmp/hadoop-hadoop/dfs/name/current/fsimage_0000000000000000002 401 B
100%
0 sec
erasure coding policies (0/0)
100%
inodes (1/1)
100%
delegation tokens (0/0)
100%
cache pools (0/0)
100%
Loading edits
100%
0 sec
/tmp/hadoop-hadoop/dfs/name/current/edits_0000000000000000003-0000000000000000003 1 MB (1/1)
100%
Saving checkpoint
100%
0 sec
Safe mode
100%
0 sec
awaiting reported blocks (0/0)
100%
Hadoop, 2019.


Namenode information
Nodes of the cluster
Directory: /logs/
localhost:8088/cluster/nodes
Logged in as: dr.who
Nodes of the cluster
Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools
Cluster Metrics
Apps Submitted
Apps Pending
Apps Running
Apps Completed
Containers Running
Memory Used
Memory Total
Memory Reserved
VCores Used
VCores Total
VCores Reserved
0 0 0 0 0 0 B 8 GB 0 B 0 8 0
Cluster Nodes Metrics
Active Nodes
Decommissioning Nodes
Decommissioned Nodes
Lost Nodes
Unhealthy Nodes
Rebooted Nodes
Shutdown Nodes
1 0 0 0 0 0
Scheduler Metrics
Scheduler Type
Scheduling Resource Type
Minimum Allocation
Maximum Allocation
Maximum Cluster Application Priority
Capacity Scheduler
[memory-mb (unit=Mi), vcores]
<memory:1024, vCores:1>
<memory:8192, vCores:4>
0
Show 20 entries
Search:
Node Labels
Rack
Node State
Node Address
Node HTTP Address
Last health-update
Health-report
Containers
Allocation Tags
Mem Used
Mem Avail
VCores Used
VCores Avail
Version
/default-rack
RUNNING
siler-virtual-machine:37417
siler-virtual-machine:8042
Wed Apr 17 16:11:32 +0800 2024
0
0 B
8 GB
0
8
3.2.1
Showing 1 to 1 of 1 entries
First Previous 1 Next Last





实验体会

发现实验后的单词统计混入了一些英文单词+符号，不过因为是 WordCount 实现问题，所以忽略了。