

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

```
hadoop@siler-virtual-machine:~/hadoop_installs/hadoop-3.2.1/share/hadoop/mapreduce$ hdfs dfs -cat /word_count_results_2/part-r-00000
2024-04-15 15:54:24,628 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remotemostTrusted = false
goodbye 1
hadoop 1
```

```
hadoop@stiller-virtual-machine:~/hadoop_installs/hadoop-3.2.1/share/hadoop/mapreduce$ hdfs dfs -ls /
Found 2 items
drwxr-xr-x  - hadoop supergroup      0 2024-04-15 15:42 /test
drwxr-xr-x  - hadoop supergroup      0 2024-04-15 15:52 /word_count_results
hadoop@stiller-virtual-machine:~/hadoop_installs/hadoop-3.2.1/share/hadoop/mapreduce$ hdfs dfs -ls /word_count_results
Found 2 items
-rw-r--r--  1 hadoop supergroup      0 2024-04-15 15:52 /word_count_results/_SUCCESS
-rw-r--r--  1 hadoop supergroup    25 2024-04-15 15:52 /word_count_results/part-r-000000
hadoop@stiller-virtual-machine:~/hadoop_installs/hadoop-3.2.1/share/hadoop/mapreduce$ hdfs dfs -cat /word_count_results/_SUCCESS
hadoop@stiller-virtual-machine:~/hadoop_installs/hadoop-3.2.1/share/hadoop/mapreduce$ hdfs dfs -cat /word_count_results/part-r-000000
2024-04-15 15:53:21,611 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
hadoop 1
hello 2
world 1
```

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/Table_(furniture)#)

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

其中，我将它们的文字部分 ctrl + 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

```
hadoop@stiller-virtual-machine:~$ hadoop jar hadoop-mapreduce-examples-3.2.1.jar wordcount /test2/chair.txt
JAR does not exist or is not a normal file: /home/hadoop/hadoop-mapreduce-examples-3.2.1.jar
hadoop@stiller-virtual-machine:~$ hadoop jar hadoop-mapreduce-examples-3.2.1.jar wordcount /test2/chair.txt /result1
JAR does not exist or is not a normal file: /home/hadoop/hadoop-mapreduce-examples-3.2.1.jar
hadoop@stiller-virtual-machine:~$ cd hadoop_installs
hadoop@stiller-virtual-machine:~/hadoop_installs$ cd hadoop-3.2.1
hadoop@stiller-virtual-machine:~/hadoop_installs/hadoop-3.2.1$ hadoop jar hadoop-mapreduce-examples-3.2.1.jar wordcount /test2/chair.txt /result1
JAR does not exist or is not a normal file: /home/hadoop/hadoop_installs/hadoop-3.2.1/hadoop-mapreduce-examples-3.2.1.jar
hadoop@stiller-virtual-machine:~/hadoop_installs/hadoop-3.2.1$ hdfs dfs -ls /test
Found 3 items
-rw-r--r--  1 hadoop supergroup    31800 2024-04-17 14:51 /test/chair.txt
-rw-r--r--  1 hadoop supergroup     7726 2024-04-17 14:51 /test/couch.txt
-rw-r--r--  1 hadoop supergroup    15758 2024-04-17 14:52 /test/table.txt
hadoop@stiller-virtual-machine:~/hadoop_installs/hadoop-3.2.1$
```

1. result1 (chair)

```
hadoop@stiller-virtual-machine: ~/hadoop_installs/hadoop-3.2.1
WRONG_MAP=0
WRONG_REDUCE=0
File Output Format Counters
  Bytes Written=20652
2024-04-17 15:36:37,706 INFO mapred.LocalJobRunner: Finishing task: attempt_local142682947_0001_r_000000_0
2024-04-17 15:36:37,707 INFO mapred.LocalJobRunner: reduce task executor complete.
2024-04-17 15:36:38,563 INFO mapreduce.Job: map 100% reduce 100%
2024-04-17 15:36:38,564 INFO mapreduce.Job: Job job_local142682947_0001 completed successfully
2024-04-17 15:36:38,597 INFO mapreduce.Job: Counters: 36
File System Counters
  FILE: Number of bytes read=691124
  FILE: Number of bytes written=1765328
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=63600
  HDFS: Number of bytes written=20652
  HDFS: Number of read operations=15
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=4
  HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=310
  Map output records=5022
  Map output bytes=51560
  Map output materialized bytes=28842
  Input split bytes=102
  Combine input records=5022
  Combine output records=2060
  Reduce input groups=2060
  Reduce shuffle bytes=28842
  Reduce input records=2060
  Reduce output records=2060
  Spilled Records=4120
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=0
  Total committed heap usage (bytes)=548405248
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
```

```

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-00000
cat: /result1/part-r-00000: 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, remotemostTrusted = false
"66-12a.html". 1
"15.5.1.4". 1
"26 1
"A 1
"Anthropocene: 1
"Architecture 1
"Armchair". 1
"BITMA". 1
"Basic 1
"CGSB 1
"Chair" 1
"Chair". 1
"Christo: 1
"Collection 2
"Definitions 1
"Elsteddfod". 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
Chairsrunnture 1
Chaise,[35] 1
Chamber 1
China 3
China, 1
China, 1
China: 1
Chinese 2
Christian 1
Christo 3
Cirruls, 1

```

2. result2 (table)

```

hadoop@siler-virtual-machine: ~/hadoop_installs/hadoop-3.2.1
standing 2
standing, 2
statementtable 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
table, 1
table, 49
table, 1
table, 1
table, 1
table, 6
table, 8
table,[7] 1
table: 1
table: 2
table, 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
temple 3
terms 1

```

```
hadoop@siler-virtual-machine: ~/hadoop_installs/hadoop-3.2.1
turned 1
two 4
type 1
types 6
typically 4
under 2
universal 1
unknown 2
up 2
upon 1
use 12
uses 2
used 16
used, 1
uses, 2
uses: 2
using 2
usually 8
variety 1
various 6
verification, 1
very 5
view 1
viewed/media 1
virtually 1
volume 1
wall 1
wall, 2
was 1
was 4
way 1
well 1
were 17
western 1
when 2
where 1
which 12
while 3
who 1
whose 1
wide 1
widespread 1
width 1
will 1
with 20
within 1
woolen 1
wood 3
wood, 2
wood-based 1
wooden 5
```

3. result3 (couch)

```
hadoop@siler-virtual-machine: ~/hadoop_installs/hadoop-3.2.1
MRJobReducer:
File Output Format Counters
  Bytes Written:396
2024-04-17 16:09:15,714 INFO Mapred.LocalJobRunner: Finishing task: attempt_local11963342_0001_r_000000_0
2024-04-17 16:09:16,618 INFO Mapred.LocalJobRunner: reduce task executor complete.
2024-04-17 16:09:16,619 INFO Mapreduce.Job: map 100% reduce 100%
2024-04-17 16:09:16,619 INFO Mapreduce.Job: Job job_local11963342_0001 completed successfully
2024-04-17 16:09:16,641 INFO Mapreduce.Job: Counters: 36
File System Counters
  FILE: Number of bytes read=651306
  FILE: Number of bytes written=1705681
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=15452
  HDFS: Number of bytes written=6396
  HDFS: Number of read operations=15
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=4
  HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=126
  Map output records=1103
  Map output bytes=11963
  Map output materialized bytes=8933
  Input split bytes=102
  Combine input records=1103
  Combine output records=636
  Reduce input groups=636
  Reduce shuffle bytes=8933
  Reduce input records=636
  Reduce output records=636
  Spilled Records=1272
  Shuffled Map <1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=0
  Total committed heap usage (bytes)=557842432
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read:728
File Output Format Counters
  Bytes Written:396
hadoop@siler-virtual-machine:~/hadoop_installs/hadoop-3.2.1/share/hadoop/mapreduce$
```

```
hadoop@siler-virtual-machine: ~/hadoop_installs/hadoop-3.2.1
2024-04-17 16:09:15,714 INFO Mapred.LocalJobRunner: reduce task executor complete.
2024-04-17 16:09:16,618 INFO Mapreduce.Job: map 100% reduce 100%
2024-04-17 16:09:16,619 INFO Mapreduce.Job: Job job_local11963342_0001 completed successfully
2024-04-17 16:09:16,641 INFO Mapreduce.Job: Counters: 36
File System Counters
  FILE: Number of bytes read=651306
  FILE: Number of bytes written=1705681
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=15452
  HDFS: Number of bytes written=6396
  HDFS: Number of read operations=15
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=4
  HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=126
  Map output records=1103
  Map output bytes=11963
  Map output materialized bytes=8933
  Input split bytes=102
  Combine input records=1103
  Combine output records=636
  Reduce input groups=636
  Reduce shuffle bytes=8933
  Reduce input records=636
  Reduce output records=636
  Spilled Records=1272
  Shuffled Map <1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=0
  Total committed heap usage (bytes)=557842432
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read:728
File Output Format Counters
  Bytes Written:396
hadoop@siler-virtual-machine:~/hadoop_installs/hadoop-3.2.1/share/hadoop/mapreduce$ hdfs -ls /result3
Found 2 items
-rw-r--r-- 1 hadoop supergroup 0 2024-04-17 16:09 /result3/_SUCCESS
-rw-r--r-- 1 hadoop supergroup 6396 2024-04-17 16:09 /result3/part-r-000000
hadoop@siler-virtual-machine:~/hadoop_installs/hadoop-3.2.1/share/hadoop/mapreduce$
```


Namenode information

Nodes of the cluster

Directory: /logs/

localhost:9870/dfshealth.html#tab-overview

90%

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

90%

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_000000000000000002 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_0000000000000000003-000000000000000003 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

90%

hadoop

Nodes of the cluster

Logged in as: d:who

Cluster

About

Nodes

Node Labels

Applications

NEW

NEWLY SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

PULLED

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 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	0

Scheduler Metrics

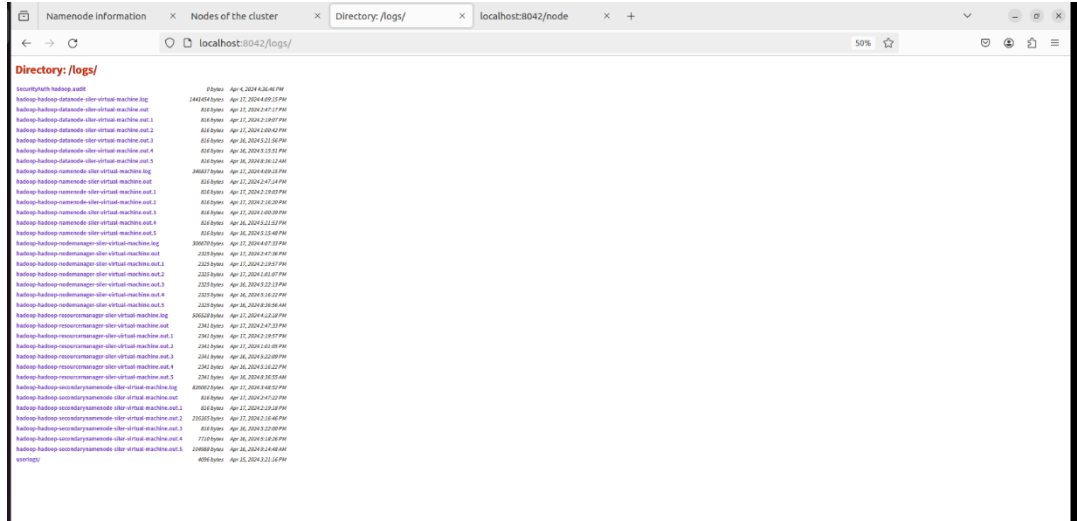
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit-Mb), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries

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	sller-virtual-machine-37417	sller-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。