并行计算实验四-Hadoop

PB18000221 袁一玮

5 月 31 日

实验题目

1. 按照 Hadoop 安装运行说明文档中的指导，自己搭建伪分布式 Hadoop 环境，熟悉 HDFS 的常用操作，运行 WordCount 程序，得到统计结果
2. 实现一个统计输入文件中各个长度的单词出现频次的程序

实验环境

- CPU:Xeon(R) Silver 4116
- 内存:255G
- 操作系统:Ubuntu 20.04
- 软件平台:Hadoop 分布式文件系统, openjdk 1.8.0_292

实验内容

安装运行 Hadoop

从 Ubuntu 20.04 官方源中装了 openjdk 1.8.0_292

参照 Hadoop 的官方文档：https://hadoop.apache.org/docs/r1.0.4/cn/quickstart.html，先从科大镜像站下载 Hadoop 的安装包：http://mirrors.

ustc.edu.cn/apache/hadoop/common/hadoop-3.2.2/hadoop-3.2.2.tar.gz，
之后再释放路径

```
$ ssh-keygen
$ cp ~/.ssh/id_rsa.pub ~/.ssh/authorized_keys
# 配置自我连接
$ export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
$ tar zxf hadoop-3.2.2.tar.gz
$ hadoop-3.2.2/bin/hadoop version
Hadoop 3.2.2
Source code repository Unknown -r 7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled by hexiaoqiao on 2021-01-03T09:26Z
Compiled with protoc 2.5.0
From source with checksum 5a8f564f46624254b27f6a33126ff4
This command was run using /home/totoro/hadoop-3.2.2/share/hadoop/
    common/hadoop-common-3.2.2.jar
```

```
$ hadoop-3.2.2/bin/hadoop version
Hadoop 3.2.2
Source code repository Unknown -r 7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled by hexiaoqiao on 2021-01-03T09:26Z
Compiled with protoc 2.5.0
From source with checksum 5a8f564f46624254b27f6a33126ff4
This command was run using /home/totoro/hadoop-3.2.2/share/hadoop/common/hadoop-common-3.2.2.jar

# totoro @ ubuntu2004 in ~ [17:03:14]
$ cat hadoop-3.2.2/etc/hadoop/core-site.xml
<?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>
</configuration>

# totoro @ ubuntu2004 in ~ [17:04:52]
$
```

图 1: 安装 Hadoop

配置 Hadoop 伪分布式，修改 core-site.xml 和 hdfs-site.xml 两个配置，如

图

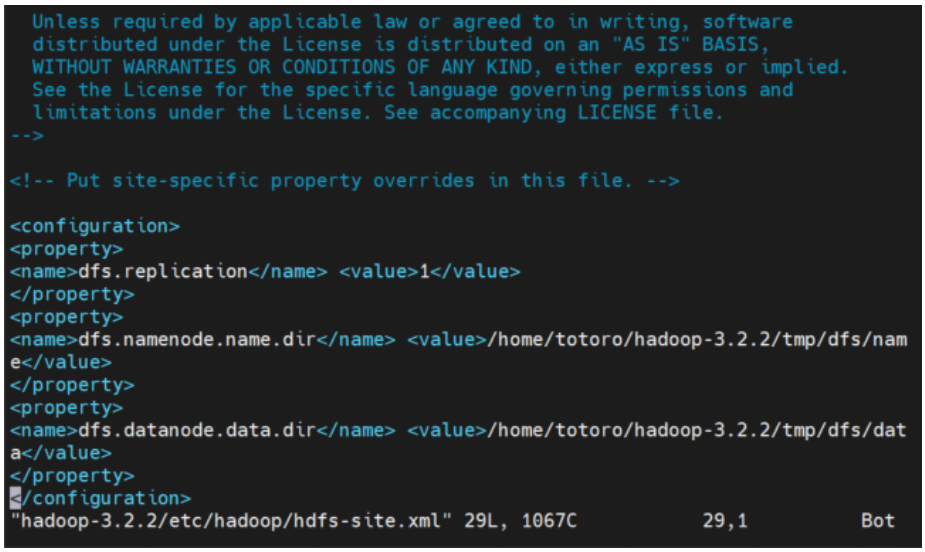


图 2: 配置 hdfs-site

再修改 hadoop-env.sh，加上 JAVA_HOME 环境变量的设置

运行 `bin/hdfs namenode -format` 命令后，输出 namenode 格式化的结果，如图

运行 `sbin/start-dfs.sh` 命令，运行 hdfs，如图

创建 hdfs 内的文件夹

```
$ hadoop-3.2.2/bin/hdfs dfs -mkdir -p hdfs
$ hadoop-3.2.2/bin/hdfs dfs -mkdir hdfs/input
$ hadoop-3.2.2/bin/hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - totoro supergroup          0 2021-05-31 18:47 /user
$ hadoop-3.2.2/bin/hdfs dfs -ls /user
Found 1 items
drwxr-xr-x   - totoro supergroup          0 2021-05-31 18:47 /user/totoro
$ hadoop-3.2.2/bin/hdfs dfs -ls /user/totoro
Found 1 items
```

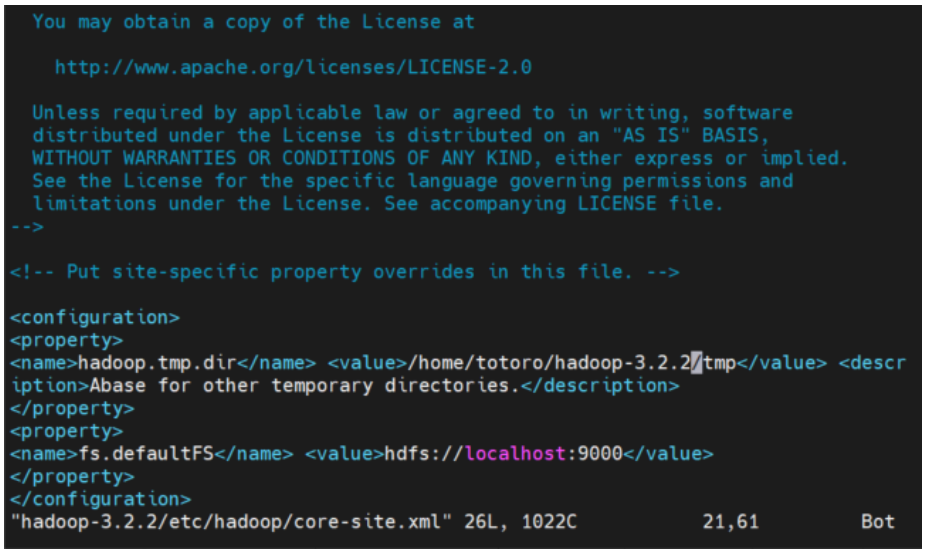
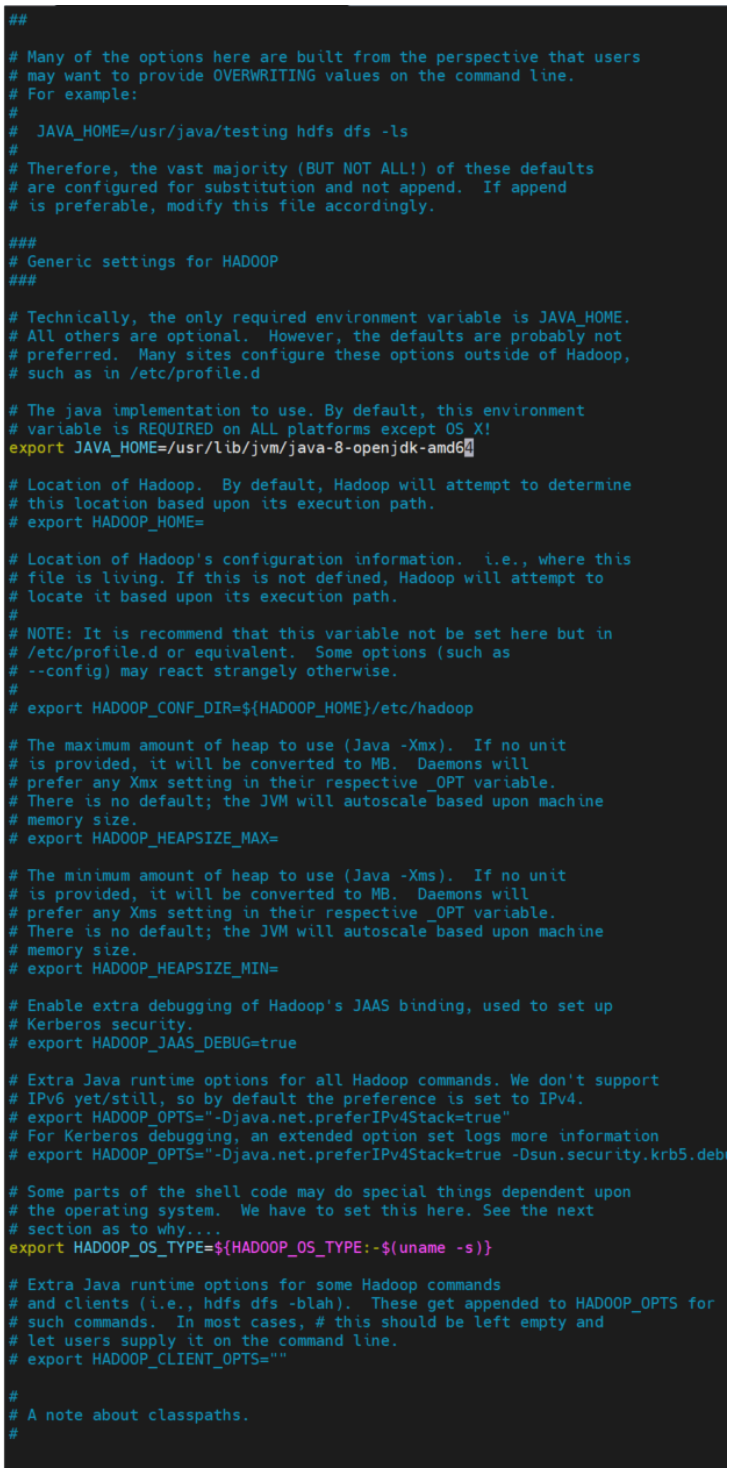


图 3: 配置 core-site

```
drwxr-xr-x   - totoro supergroup          0 2021-05-31 18:48 /user/totoro/hdfs
$ hadoop-3.2.2/bin/hdfs dfs -ls /user/totoro/hdfs
Found 1 items
drwxr-xr-x   - totoro supergroup          0 2021-05-31 18:48 /user/totoro/hdfs/input
```

将主机上文件转移至 hdfs 内

```
$ hadoop-3.2.2/bin/hdfs dfs -put hadoop-3.2.2/etc/hadoop/*.xml hdfs/input
$ hadoop-3.2.2/bin/hdfs dfs -ls /user/totoro/hdfs/input
Found 9 items
-rw-r--r--   1 totoro supergroup       9213 2021-05-31 18:59 /user/totoro/hdfs/
    input/capacity-scheduler.xml
-rw-r--r--   1 totoro supergroup       1022 2021-05-31 18:59 /user/totoro/hdfs/
    input/core-site.xml
-rw-r--r--   1 totoro supergroup      11392 2021-05-31 18:59 /user/totoro/hdfs/
    input/hadoop-policy.xml
-rw-r--r--   1 totoro supergroup       1067 2021-05-31 18:59 /user/totoro/hdfs/
    input/hdfs-site.xml
-rw-r--r--   1 totoro supergroup        620 2021-05-31 18:59 /user/totoro/hdfs/
```



```
##

# Many of the options here are built from the perspective that users
# may want to provide OVERWRITING values on the command line.
# For example:
#
#  JAVA_HOME=/usr/java/testing hdfs dfs -ls
#
# Therefore, the vast majority (BUT NOT ALL!) of these defaults
# are configured for substitution and not append.  If append
# is preferable, modify this file accordingly.

###
# Generic settings for HADOOP
###

# Technically, the only required environment variable is JAVA_HOME.
# All others are optional.  However, the defaults are probably not
# preferred.  Many sites configure these options outside of Hadoop,
# such as in /etc/profile.d

# The java implementation to use. By default, this environment
# variable is REQUIRED on ALL platforms except OS X!
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

# Location of Hadoop.  By default, Hadoop will attempt to determine
# this location based upon its execution path.
# export HADOOP_HOME=

# Location of Hadoop's configuration information.  i.e., where this
# file is living. If this is not defined, Hadoop will attempt to
# locate it based upon its execution path.
#
# NOTE: It is recommend that this variable not be set here but in
# /etc/profile.d or equivalent.  Some options (such as
# --config) may react strangely otherwise.
#
# export HADOOP_CONF_DIR=${HADOOP_HOME}/etc/hadoop

# The maximum amount of heap to use (Java -Xmx).  If no unit
# is provided, it will be converted to MB.  Daemons will
# prefer any Xmx setting in their respective _OPT variable.
# There is no default; the JVM will autoscale based upon machine
# memory size.
# export HADOOP_HEAPSIZE_MAX=

# The minimum amount of heap to use (Java -Xms).  If no unit
# is provided, it will be converted to MB.  Daemons will
# prefer any Xms setting in their respective _OPT variable.
# There is no default; the JVM will autoscale based upon machine
# memory size.
# export HADOOP_HEAPSIZE_MIN=

# Enable extra debugging of Hadoop's JAAS binding, used to set up
# Kerberos security.
# export HADOOP_JAAS_DEBUG=true

# Extra Java runtime options for all Hadoop commands. We don't support
# IPv6 yet/still, so by default the preference is set to IPv4.
# export HADOOP_OPTS="-Djava.net.preferIPv4Stack=true"
# For Kerberos debugging, an extended option set logs more information
# export HADOOP_OPTS="-Djava.net.preferIPv4Stack=true -Dsun.security.krb5.debu

# Some parts of the shell code may do special things dependent upon
# the operating system.  We have to set this here. See the next
# section as to why....
export HADOOP_OS_TYPE=${HADOOP_OS_TYPE:-$(uname -s)}

# Extra Java runtime options for some Hadoop commands
# and clients (i.e., hdfs dfs -blah).  These get appended to HADOOP_OPTS for
# such commands.  In most cases, # this should be left empty and
# let users supply it on the command line.
# export HADOOP_CLIENT_OPTS=""

#
# A note about classpaths.
#
```

图 4: 配置 JAVA_HOME

图 5: hdfs format

```
    input/httpfs-site.xml
-rw-r--r--   1 totoro supergroup       3518 2021-05-31 18:59 /user/totoro/hdfs/
    input/kms-acls.xml
-rw-r--r--   1 totoro supergroup        682 2021-05-31 18:59 /user/totoro/hdfs/
    input/kms-site.xml
-rw-r--r--   1 totoro supergroup        758 2021-05-31 18:59 /user/totoro/hdfs/
    input/mapred-site.xml
-rw-r--r--   1 totoro supergroup        690 2021-05-31 18:59 /user/totoro/hdfs/
    input/yarn-site.xml
```

运行了 `hadoop-3.2.2/bin/hadoop jar WordCount.jar grep hdfs/input output` 程序，成功运行字数记录程序，output 里结果如 word-count 所示

从 HDFS 中取出运行的结果

```
$ hadoop-3.2.2/bin/hadoop fs -get output/part-r-00000 .
$ ls
gost  hadoop-3.2.2  part-r-00000  WordCount.jar
$ head part-r-00000
"*"     21
```

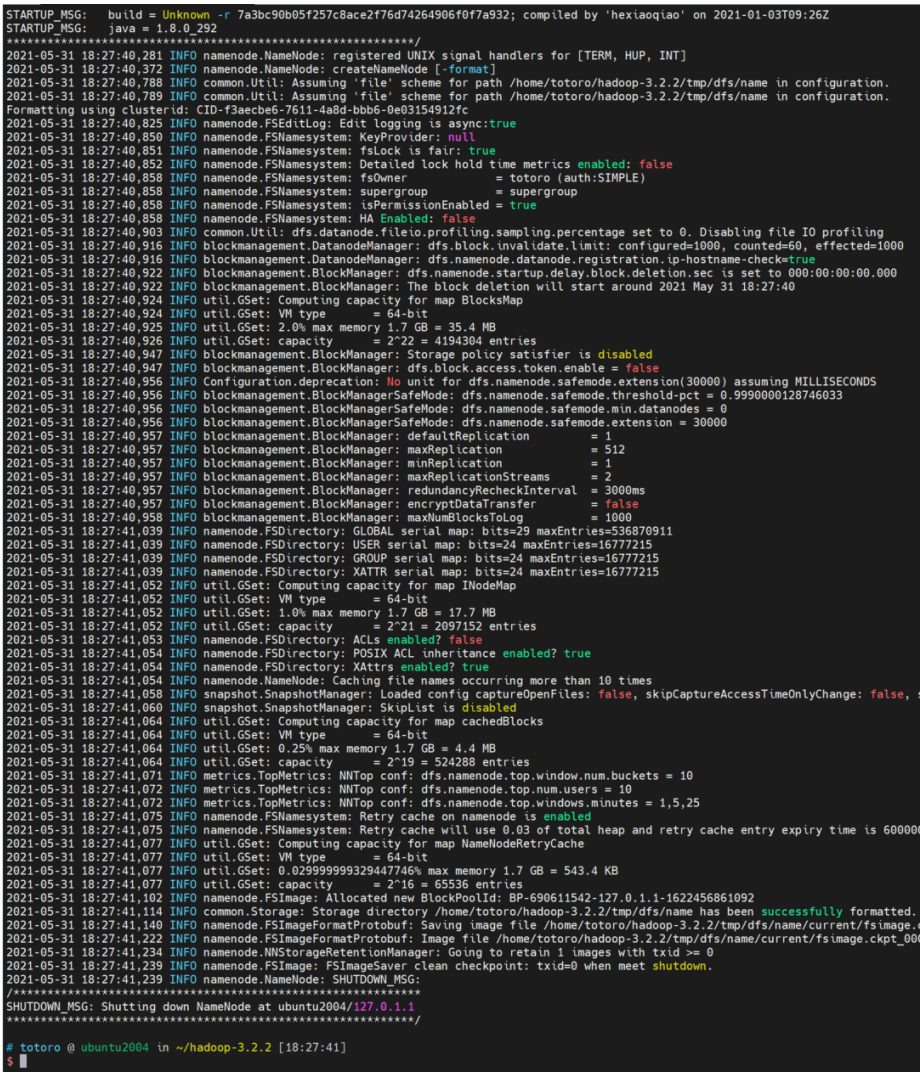


图 6: hdfs format 输出结果

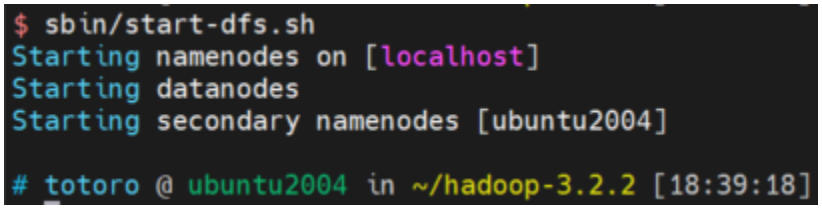


图 7: 运行 hdfs

```
# totoro @ ubuntu2004 in ~ [18:50:56]
$ hadoop-3.2.2/bin/hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - totoro supergroup          0 2021-05-31 18:47 /user

# totoro @ ubuntu2004 in ~ [18:51:10]
$ hadoop-3.2.2/bin/hdfs dfs -ls /user
Found 1 items
drwxr-xr-x   - totoro supergroup          0 2021-05-31 18:47 /user/totoro

# totoro @ ubuntu2004 in ~ [18:51:21]
$ hadoop-3.2.2/bin/hdfs dfs -ls /user/totoro
Found 1 items
drwxr-xr-x   - totoro supergroup          0 2021-05-31 18:48 /user/totoro/hdfs

# totoro @ ubuntu2004 in ~ [18:51:27]
$ hadoop-3.2.2/bin/hdfs dfs -ls /user/totoro/hdfs
Found 1 items
drwxr-xr-x   - totoro supergroup          0 2021-05-31 18:48 /user/totoro/hdfs/input

# totoro @ ubuntu2004 in ~ [18:51:32]
$
```

图 8: hdfs-ls

```
# totoro @ ubuntu2004 in ~ [18:58:53]
$ hadoop-3.2.2/bin/hdfs dfs -put hadoop-3.2.2/etc/hadoop/*.xml hdfs/input

# totoro @ ubuntu2004 in ~ [18:59:19]
$ hadoop-3.2.2/bin/hdfs dfs -ls /user/totoro/hdfs/input
Found 9 items
-rw-r--r--   1 totoro supergroup       9213 2021-05-31 18:59 /user/totoro/hdfs/input/capacity-scheduler.xml
-rw-r--r--   1 totoro supergroup       1022 2021-05-31 18:59 /user/totoro/hdfs/input/core-site.xml
-rw-r--r--   1 totoro supergroup      11392 2021-05-31 18:59 /user/totoro/hdfs/input/hadoop-policy.xml
-rw-r--r--   1 totoro supergroup       1067 2021-05-31 18:59 /user/totoro/hdfs/input/hdfs-site.xml
-rw-r--r--   1 totoro supergroup        620 2021-05-31 18:59 /user/totoro/hdfs/input/httpfs-site.xml
-rw-r--r--   1 totoro supergroup       3518 2021-05-31 18:59 /user/totoro/hdfs/input/kms-acls.xml
-rw-r--r--   1 totoro supergroup        682 2021-05-31 18:59 /user/totoro/hdfs/input/kms-site.xml
-rw-r--r--   1 totoro supergroup        758 2021-05-31 18:59 /user/totoro/hdfs/input/mapred-site.xml
-rw-r--r--   1 totoro supergroup        690 2021-05-31 18:59 /user/totoro/hdfs/input/yarn-site.xml
```

图 9: hdfs-put

```
$ hadoop-3.2.2/bin/hdfs dfs -ls /user/totoro/output
Found 2 items
-rw-r--r--   1 totoro supergroup          0 2021-05-31 19:10 /user/totoro/output/_SUCCESS
-rw-r--r--   1 totoro supergroup      10108 2021-05-31 19:10 /user/totoro/output/part-r-00000

# totoro @ ubuntu2004 in ~ [19:11:40]
$ hadoop-3.2.2/bin/hdfs dfs -cat /user/totoro/output
cat: `/user/totoro/output': Is a directory

# totoro @ ubuntu2004 in ~ [19:11:55] C:1
$ hadoop-3.2.2/bin/hdfs dfs -cat /user/totoro/output/part-r-00000
"*"     21
"AS     9
"License");     9
"alice,bob      21
"clumping"      1
"full_queue_name"       1
"priority".     1
"workflowId"    1
(ASF)   1
(as     1
(or     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
<!--    18
</configuration>        9
</description>  35
</name> 2
</property>     64
<?xml   8
<?xml-stylesheet        4
<configuration> 9
<description>   33
<description>ACL        25
<description>Abase      1
<description>Default    1
<name>default.key.acl.DECRYPT_EEK</name>        1
<name>default.key.acl.GENERATE_EEK</name>       1
<name>default.key.acl.MANAGEMENT</name> 1
<name>default.key.acl.READ</name>       1
<name>dfs.datanode.data.dir</name>      1
<name>dfs.namenode.name.dir</name>      1
<name>dfs.replication</name>    1
<name>fs.defaultFS</name>       1
<name>hadoop.kms.acl.CREATE</name>      1
<name>hadoop.kms.acl.DECRYPT_EEK</name> 1
<name>hadoop.kms.acl.DELETE</name>      1
<name>hadoop.kms.acl.GENERATE_EEK</name>        1
<name>hadoop.kms.acl.GET</name> 1
<name>hadoop.kms.acl.GET_KEYS</name>    1
<name>hadoop.kms.acl.GET_METADATA</name>        1
<name>hadoop.kms.acl.ROLLOVER</name>    1
<name>hadoop.kms.acl.SET_KEY_MATERIAL</name>    1
<name>hadoop.tmp.dir</name>     1
<name>security.admin.operations.protocol.acl</name>     1
<name>security.applicationclient.protocol.acl</name>    1
<name>security.applicationhistory.protocol.acl</name>   1
<name>security.applicationmaster-nodemanager.applicationmaster.protocol.acl</name>      1
<name>security.applicationmaster.protocol.acl</name>    1
<name>security.client.datanode.protocol.acl</name>      1
<name>security.client.protocol.acl</name>       1
<name>security.collector-nodemanager.protocol.acl</name>        1
<name>security.containermanagement.protocol.acl</name>  1
<name>security.datanode.protocol.acl</name>     1
<name>security.distributedscheduling.protocol.acl</name>        1
<name>security.ha.service.protocol.acl</name>   1
<name>security.inter.datanode.protocol.acl</name>       1
```

图 10: word-count

```
"AS     9
"License");     9
"alice,bob      21
"clumping"      1
"full_queue_name"       1
"priority".     1
"workflowId"    1
(ASF)   1
(as     1
```

在 Hadoop 中统计输入文件中各个长度的单词出现频次

根据 WordCount.jar，实现 LenCount.jar 程序运行 Map-Reduce 运算，output 里输出的结果如下

```
$ hadoop-3.2.2/bin/hdfs dfs -cat output/part-r-00000
1       1
10      7
11      8
12      1
13      2
15      2
16      3
18      1
2       14
26      1
28      1
3       25
34      2
4       8
42      1
5       7
53      2
6       7
```

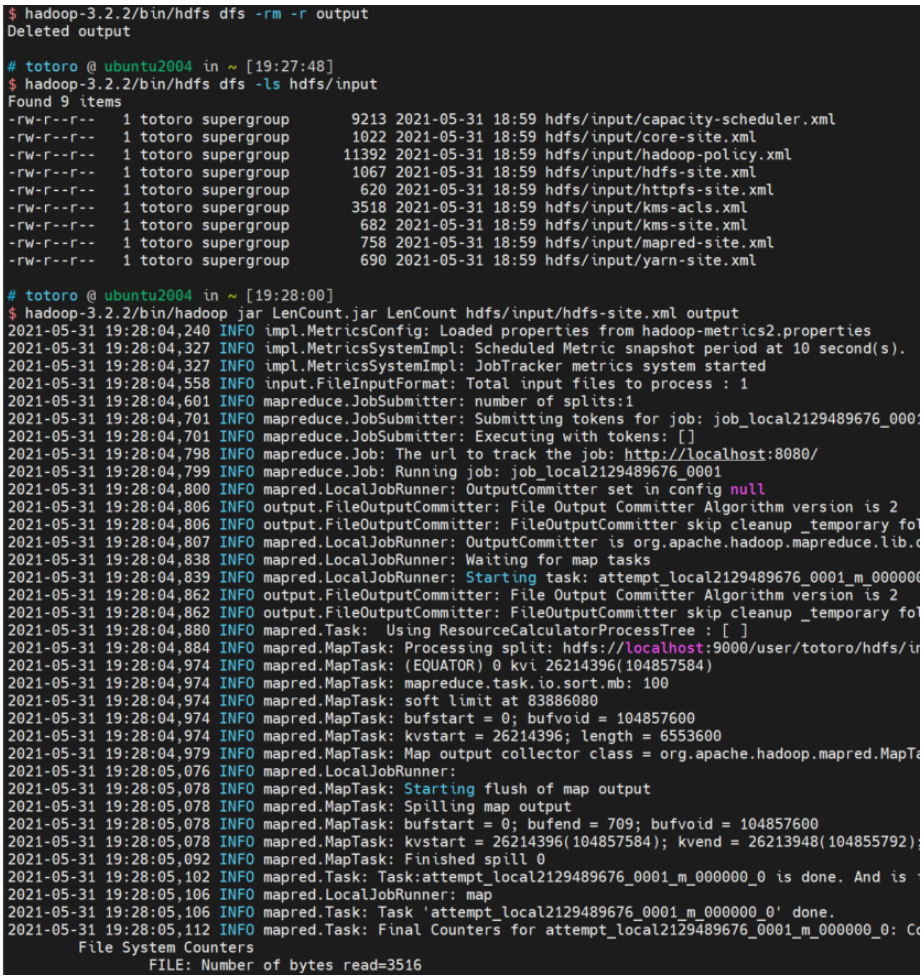


图 11: LenCount

```
7       7
8       11
9       2
```

实验结束后关闭 Hadoop 程序

```
$ hadoop-3.2.2/sbin/stop-dfs.sh
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [ubuntu2004]
```