

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 xzf 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

```

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>
<property>
<name>dfs.replication</name> <value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name> <value>/home/totoro/hadoop-3.2.2/tmp/dfs/nam
e</value>
</property>
<property>
<name>dfs.datanode.data.dir</name> <value>/home/totoro/hadoop-3.2.2/tmp/dfs/dat
a</value>
</property>
</configuration>
"hadoop-3.2.2/etc/hadoop/hdfs-site.xml" 29L, 1067C          29,1          Bot

```

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

```

```

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>
<property>
<name>hadoop.tmp.dir</name> <value>/home/totoro/hadoop-3.2.2/tmp</value> <descrip
tion>Abase for other temporary directories.</description>
</property>
<property>
<name>fs.defaultFS</name> <value>hdfs://localhost:9000</value>
</property>
</configuration>
"hadoop-3.2.2/etc/hadoop/core-site.xml" 26L, 1022C          21,61      Bot

```

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.debug"

# 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

[illegible]

图 5: hdfs format

```

input/https-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 .
```

\$ 1s

```
gost  hadoop-3.2.2  part-r-00000  WordCount.jar
```

```
$ head part-r-00000
```

"*" 21

```

STARTUP_MSG: build = Unknown -r 7a3bc90b05f257c8ace2f76d74264906f0f7a932; compiled by 'hexiaobao' on 2021-01-03T09:26Z
STARTUP_MSG: java = 1.8.0_292
*****
2021-05-31 18:27:40,281 INFO namenode.NameNode: registered UNIX signal handlers for [TERM, HUP, INT]
2021-05-31 18:27:40,372 INFO namenode.NameNode: createNameNode [-format]
2021-05-31 18:27:40,788 INFO common.Util: Assuming 'file' scheme for path /home/totoro/hadoop-3.2.2/tmp/dfs/name in configuration.
2021-05-31 18:27:40,789 INFO common.Util: Assuming 'file' scheme for path /home/totoro/hadoop-3.2.2/tmp/dfs/name in configuration.
Formatting using clusterid: CID-f3aeb66-7611-4a8d-bbb6-0e03154912fc
2021-05-31 18:27:40,825 INFO namenode.FSEditLog: Edit logging is async:true
2021-05-31 18:27:40,850 INFO namenode.FSNamesystem: KeyProvider: null
2021-05-31 18:27:40,851 INFO namenode.FSNamesystem: fsLock is fair: true
2021-05-31 18:27:40,852 INFO namenode.FSNamesystem: Detailed lock hold time metrics enabled: false
2021-05-31 18:27:40,858 INFO namenode.FSNamesystem: fsOwner = totoro (auth:SIMPLE)
2021-05-31 18:27:40,858 INFO namenode.FSNamesystem: supergroup = supergroup
2021-05-31 18:27:40,858 INFO namenode.FSNamesystem: isPermissionEnabled = true
2021-05-31 18:27:40,858 INFO namenode.FSNamesystem: HA Enabled: false
2021-05-31 18:27:40,903 INFO common.Util: dfs.datanode.fileio.profiling.sampling.percentage set to 0. Disabling file IO profiling
2021-05-31 18:27:40,916 INFO blockmanagement.DatanodeManager: dfs.block.invalidate.limit: configured=1000, counted=60, effected=1000
2021-05-31 18:27:40,916 INFO blockmanagement.DatanodeManager: dfs.namenode.datanode.registration.ip-hostname-check=true
2021-05-31 18:27:40,922 INFO blockmanagement.BlockManager: dfs.namenode.startup.delay.block.deletion.sec is set to 000:00:00:00.000
2021-05-31 18:27:40,922 INFO blockmanagement.BlockManager: The block deletion will start around 2021 May 31 18:27:40
2021-05-31 18:27:40,924 INFO util.GSet: Computing capacity for map BlocksMap
2021-05-31 18:27:40,924 INFO util.GSet: VM type = 64-bit
2021-05-31 18:27:40,925 INFO util.GSet: 2.0% max memory 1.7 GB = 35.4 MB
2021-05-31 18:27:40,926 INFO util.GSet: capacity = 2^22 = 4194304 entries
2021-05-31 18:27:40,947 INFO blockmanagement.BlockManager: Storage policy satisfier is disabled
2021-05-31 18:27:40,947 INFO blockmanagement.BlockManager: dfs.block.access.token.enable = false
2021-05-31 18:27:40,956 INFO Configuration.deprecation: No unit for dfs.namenode.safemode.extension(30000) assuming MILLISECONDS
2021-05-31 18:27:40,956 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.threshold-pct = 0.9990000128746033
2021-05-31 18:27:40,956 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.min.datanodes = 0
2021-05-31 18:27:40,956 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.extension = 30000
2021-05-31 18:27:40,957 INFO blockmanagement.BlockManager: defaultReplication = 1
2021-05-31 18:27:40,957 INFO blockmanagement.BlockManager: maxReplication = 512
2021-05-31 18:27:40,957 INFO blockmanagement.BlockManager: minReplication = 1
2021-05-31 18:27:40,957 INFO blockmanagement.BlockManager: maxReplicationStreams = 2
2021-05-31 18:27:40,957 INFO blockmanagement.BlockManager: redundancyRecheckInterval = 3000ms
2021-05-31 18:27:40,957 INFO blockmanagement.BlockManager: encryptDataTransfer = false
2021-05-31 18:27:40,958 INFO blockmanagement.BlockManager: maxNumLocksToLog = 1000
2021-05-31 18:27:41,030 INFO namenode.FSDirectory: GLOBAL serial map: bits=29 maxEntries=536870911
2021-05-31 18:27:41,039 INFO namenode.FSDirectory: USER serial map: bits=24 maxEntries=16777215
2021-05-31 18:27:41,039 INFO namenode.FSDirectory: GROUP serial map: bits=24 maxEntries=16777215
2021-05-31 18:27:41,039 INFO namenode.FSDirectory: XATTR serial map: bits=24 maxEntries=16777215
2021-05-31 18:27:41,052 INFO util.GSet: Computing capacity for map INodeMap
2021-05-31 18:27:41,052 INFO util.GSet: VM type = 64-bit
2021-05-31 18:27:41,052 INFO util.GSet: 1.0% max memory 1.7 GB = 17.7 MB
2021-05-31 18:27:41,052 INFO util.GSet: capacity = 2^21 = 2097152 entries
2021-05-31 18:27:41,053 INFO namenode.FSDirectory: ACLs enabled? false
2021-05-31 18:27:41,054 INFO namenode.FSDirectory: POSIX ACL inheritance enabled? true
2021-05-31 18:27:41,054 INFO namenode.FSDirectory: Xattrs enabled? true
2021-05-31 18:27:41,054 INFO namenode.NameNode: Caching file names occurring more than 10 times
2021-05-31 18:27:41,058 INFO snapshot.SnapshotManager: Loaded config captureOpenFiles: false, skipCaptureAccessTimeOnlyChange: false,
2021-05-31 18:27:41,060 INFO snapshot.SnapshotManager: Skiplist is disabled
2021-05-31 18:27:41,064 INFO util.GSet: Computing capacity for map cachedBlocks
2021-05-31 18:27:41,064 INFO util.GSet: VM type = 64-bit
2021-05-31 18:27:41,064 INFO util.GSet: 0.25% max memory 1.7 GB = 4.4 MB
2021-05-31 18:27:41,064 INFO util.GSet: capacity = 2^19 = 524288 entries
2021-05-31 18:27:41,071 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2021-05-31 18:27:41,072 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2021-05-31 18:27:41,072 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2021-05-31 18:27:41,075 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2021-05-31 18:27:41,075 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 60000ms
2021-05-31 18:27:41,077 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2021-05-31 18:27:41,077 INFO util.GSet: VM type = 64-bit
2021-05-31 18:27:41,077 INFO util.GSet: 0.029999999329447746% max memory 1.7 GB = 543.4 KB
2021-05-31 18:27:41,077 INFO util.GSet: capacity = 2^16 = 65536 entries
2021-05-31 18:27:41,102 INFO namenode.FSImage: Allocated new BlockPoolId: BP-690611542-127.0.1.1-1622456961092
2021-05-31 18:27:41,114 INFO common.Storage: Storage directory /home/totoro/hadoop-3.2.2/tmp/dfs/name has been successfully formatted.
2021-05-31 18:27:41,140 INFO namenode.FSImageFormatProtobuf: Saving image file /home/totoro/hadoop-3.2.2/tmp/dfs/name/current/fsimage.ckpt_000
2021-05-31 18:27:41,234 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2021-05-31 18:27:41,239 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2021-05-31 18:27:41,239 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at ubuntu2004/127.0.1.1
*****/
# totoro @ ubuntu2004 in ~/hadoop-3.2.2 [18:27:41]
$

```

图 6: hdfs format 输出结果

```

$ sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ubuntu2004]

# totoro @ ubuntu2004 in ~/hadoop-3.2.2 [18:39:18]

```

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/https-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

```

```

$ hadoop-3.2.2/bin/hdfs dfs -rm -r output
Deleted output

# totoro @ ubuntu2004 in ~ [19:27:48]
$ hadoop-3.2.2/bin/hdfs dfs -ls hdfs/input
Found 9 items
-rw-r--r-- 1 totoro supergroup 9213 2021-05-31 18:59 hdfs/input/capacity-scheduler.xml
-rw-r--r-- 1 totoro supergroup 1022 2021-05-31 18:59 hdfs/input/core-site.xml
-rw-r--r-- 1 totoro supergroup 11392 2021-05-31 18:59 hdfs/input/hadoop-policy.xml
-rw-r--r-- 1 totoro supergroup 1067 2021-05-31 18:59 hdfs/input/hdfs-site.xml
-rw-r--r-- 1 totoro supergroup 620 2021-05-31 18:59 hdfs/input/https-site.xml
-rw-r--r-- 1 totoro supergroup 3518 2021-05-31 18:59 hdfs/input/kms-acls.xml
-rw-r--r-- 1 totoro supergroup 682 2021-05-31 18:59 hdfs/input/kms-site.xml
-rw-r--r-- 1 totoro supergroup 758 2021-05-31 18:59 hdfs/input/mapred-site.xml
-rw-r--r-- 1 totoro supergroup 690 2021-05-31 18:59 hdfs/input/yarn-site.xml

# totoro @ ubuntu2004 in ~ [19:28:00]
$ hadoop-3.2.2/bin/hadoop jar LenCount.jar LenCount hdfs/input/hdfs-site.xml output
2021-05-31 19:28:04,240 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2021-05-31 19:28:04,327 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2021-05-31 19:28:04,327 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2021-05-31 19:28:04,558 INFO input.FileInputFormat: Total input files to process : 1
2021-05-31 19:28:04,601 INFO mapreduce.JobSubmitter: number of splits:1
2021-05-31 19:28:04,701 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2129489676_0001
2021-05-31 19:28:04,701 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-05-31 19:28:04,798 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2021-05-31 19:28:04,799 INFO mapreduce.Job: Running job: job_local2129489676_0001
2021-05-31 19:28:04,800 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2021-05-31 19:28:04,806 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2021-05-31 19:28:04,806 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary for
2021-05-31 19:28:04,807 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.o
2021-05-31 19:28:04,838 INFO mapred.LocalJobRunner: Waiting for map tasks
2021-05-31 19:28:04,839 INFO mapred.LocalJobRunner: Starting task: attempt_local2129489676_0001_m_000000_0
2021-05-31 19:28:04,862 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2021-05-31 19:28:04,862 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary fo
2021-05-31 19:28:04,880 INFO mapred.Task: Using ResourceCalculatorProcessTree: [ ]
2021-05-31 19:28:04,884 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/totoro/hdfs/in
2021-05-31 19:28:04,974 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2021-05-31 19:28:04,974 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2021-05-31 19:28:04,974 INFO mapred.MapTask: soft limit at 83886080
2021-05-31 19:28:04,974 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2021-05-31 19:28:04,974 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2021-05-31 19:28:04,979 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapT
2021-05-31 19:28:05,076 INFO mapred.LocalJobRunner:
2021-05-31 19:28:05,078 INFO mapred.MapTask: Starting flush of map output
2021-05-31 19:28:05,078 INFO mapred.MapTask: Spilling map output
2021-05-31 19:28:05,078 INFO mapred.MapTask: bufstart = 0; bufend = 709; bufvoid = 104857600
2021-05-31 19:28:05,078 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26213948(104855792);
2021-05-31 19:28:05,092 INFO mapred.MapTask: Finished spill 0
2021-05-31 19:28:05,102 INFO mapred.Task: Task:attempt_local2129489676_0001_m_000000_0 is done. And is
2021-05-31 19:28:05,106 INFO mapred.LocalJobRunner: map
2021-05-31 19:28:05,106 INFO mapred.Task: Task 'attempt_local2129489676_0001_m_000000_0' done.
2021-05-31 19:28:05,112 INFO mapred.Task: Final Counters for attempt_local2129489676_0001_m_000000_0: Co
File System Counters
FILE: Number of bytes read=3516

```

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]
```