

Hadoop WordCount Assignment

WORDCOUNT

Step1:

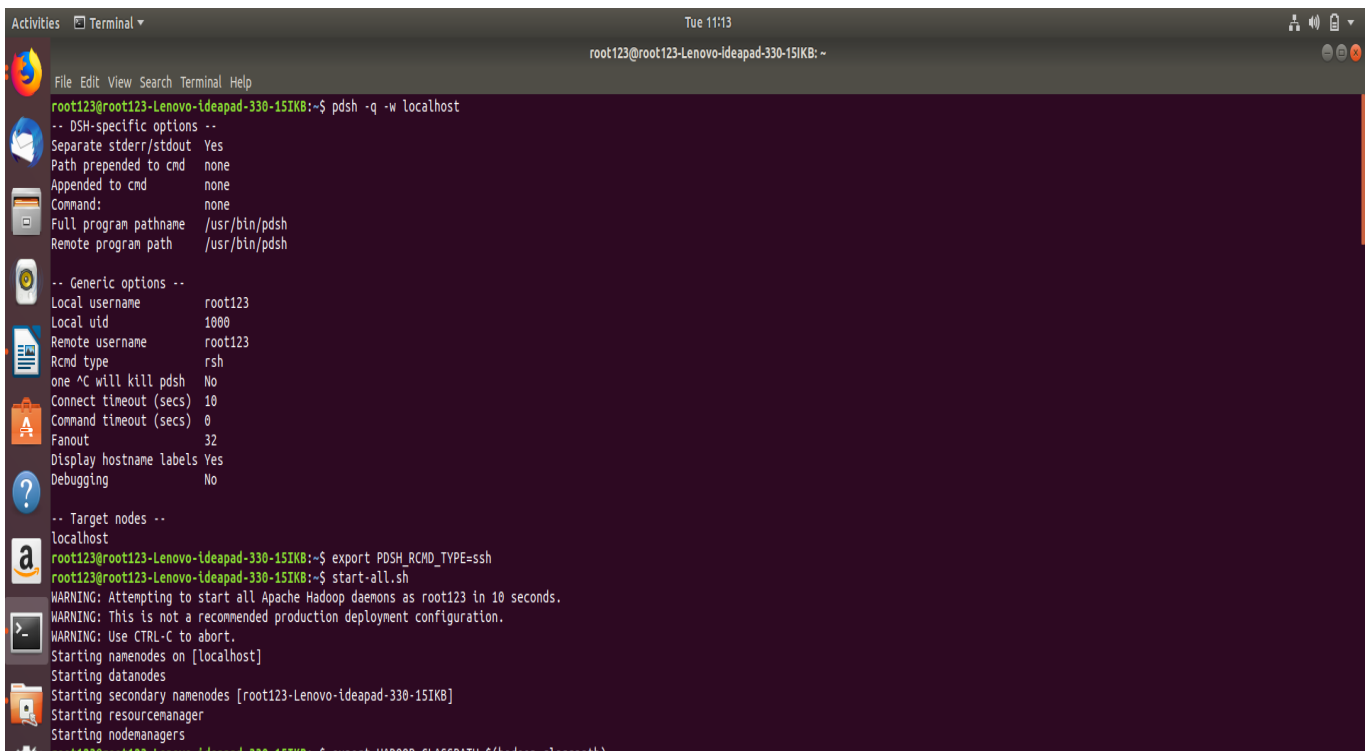
Start hadoop-3.0.3

Commands

-pdsh -q -w localhost

-export PDSH_RCMD_TYPE=ssh

-start-all.sh



```
Activities Terminal Tue 11:13
root123@root123-Lenovo-Ideapad-330-15IKB: ~
root123@root123-Lenovo-Ideapad-330-15IKB:~$ pdsh -q -w localhost
-- DSH-specific options --
Separate stderr/stdout Yes
Path prepended to cmd none
Appended to cmd none
Command: none
Full program pathname /usr/bin/pdsh
Remote program path /usr/bin/pdsh

-- Generic options --
Local username root123
Local uid 1000
Remote username root123
Rcmd type rsh
one ^C will kill pdsh No
Connect timeout (secs) 10
Command timeout (secs) 0
Fanout 32
Display hostname labels Yes
Debugging No

-- Target nodes --
localhost
root123@root123-Lenovo-Ideapad-330-15IKB:~$ export PDSH_RCMD_TYPE=ssh
root123@root123-Lenovo-Ideapad-330-15IKB:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as root123 in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [root123-Lenovo-Ideapad-330-15IKB]
Starting resourcemanager
Starting nodemanagers
root123@root123-Lenovo-Ideapad-330-15IKB:~$ export HADOOP_CLASSPATH=$(hadoop classpath)
```

Text file-



Step2:

Commands:

- export HADOOP_CLASSPATH=\$(hadoop classpath)
- echo \$HADOOP_CLASSPATH
- hadoop fs -mkdir
- hadoop fs -mldir /run

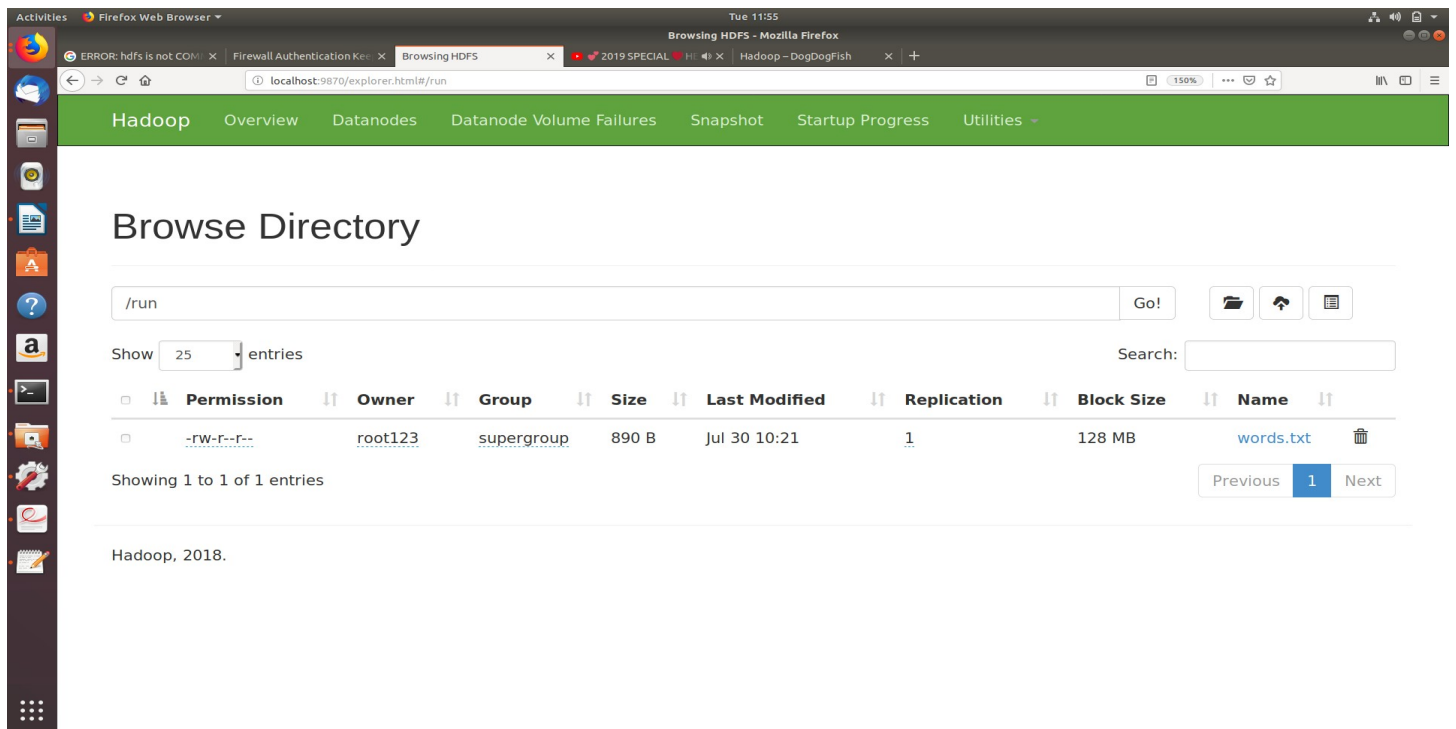
```
Activities Terminal Tue 11:13
root123@root123-Lenovo-Ideapad-330-151KB: ~
File Edit View Search Terminal Help
Command [generic options] [command options]

root123@root123-Lenovo-Ideapad-330-151KB:~$ hadoop fs -mldir /run
-mldir: Unknown command
Usage: hadoop fs [generic options]
[-appendToFile <localsrc> ... <dst>]
[-cat [-ignoreCrc] <src> ...]
[-checksum <src> ...]
[-chgrp [-R] GROUP PATH...]
[-chmod [-R] <MODE>[,<MODE>... | OCTALMODE] PATH...]
[-chown [-R] [OWNER][:[GROUP]] PATH...]
[-copyFromLocal [-f] [-p] [-l] [-d] [-t <thread count>] <localsrc> ... <dst>]
[-copyToLocal [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
[-count [-q] [-h] [-v] [-t <storage type>]] [-u] [-x] [-e] <path> ...]
[-cp [-f] [-p | -p[topax]] [-d] <src> ... <dst>]
[-createSnapshot <snapshotDir> [<snapshotName>]]
[-deleteSnapshot <snapshotDir> <snapshotName>]
[-df [-h] [<path> ...]]
[-du [-s] [-h] [-v] [-X] <path> ...]
[-expunge]
[-find <path> ... <expression> ...]
[-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
[-getfacl [-R] <path>]
[-getfattr [-R] [-n name | -d] [-e en] <path>]
[-getmerge [-nl] [-skip-empty-file] <src> <localdst>]
[-help [cmd ...]]
[-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [-e] [<path> ...]]
[-mkdir [-p] <path> ...]
[-moveFromLocal <localsrc> ... <dst>]
[-moveToLocal <src> <localdst>]
[-mv <src> ... <dst>]
[-put [-f] [-p] [-l] [-d] <localsrc> ... <dst>]
[-renameSnapshot <snapshotDir> <oldName> <newName>]
[-rm [-f] [-r] [-R] [-skipTrash] [-safely] <src> ...]
[-rmdir [--ignore-fail-on-non-empty] <dir> ...]
[-setfacl [-R] [{-b|-k} {-m|-x <acl_spec>} <path>][[-set <acl_spec> <path>]]
[-setfattr [-n name [-v value] | -x name] <path>]
[-setrep [-R] [-w] <rep> <path> ...]
[-stat [format] <path> ...]
[-tail [-f] <file>]
[-test [-d] [-d] <path>]
[-text [-ignoreCrc] <src> ...]
[-touchz <path> ...]
[-truncate [-w] <length> <path> ...]
[-usage [cmd ...]]

Generic options supported are:
-conf <configuration file>      specify an application configuration file
-D <property=value>              define a value for a given property
-fs <file:///|hdfs://namenode:port> specify default filesystem URL to use, overrides 'fs.defaultFS' property from configurations.
-jt <local|resourceManager:port> specify a ResourceManager
-files <file1,...>               specify a comma-separated list of files to be copied to the map reduce cluster
-libjars <jar1,...>             specify a comma-separated list of jar files to be included in the classpath
-archives <archive1,...>       specify a comma-separated list of archives to be unarchived on the compute machines

The general command line syntax is:
```

File allocation in localhost:9870



Step3:
write mapper code in .py format

```
#!/usr/bin/python3
import sys

for line in sys.stdin:
    for word in line.strip().split():
        print(word, 1)
```

Step4:
Write reducer code in .py format

```
#!/usr/bin/python3
import sys

current_word = None
current_count = 1

for line in sys.stdin:
    word, count = line.strip().split(' ')
    if current_word:
        if word == current_word:
            current_count += int(count)
        else:
            print(current_word, current_count)
            current_count = 1
    current_word = word

if current_count > 1:
    print(current_word, current_count)
```

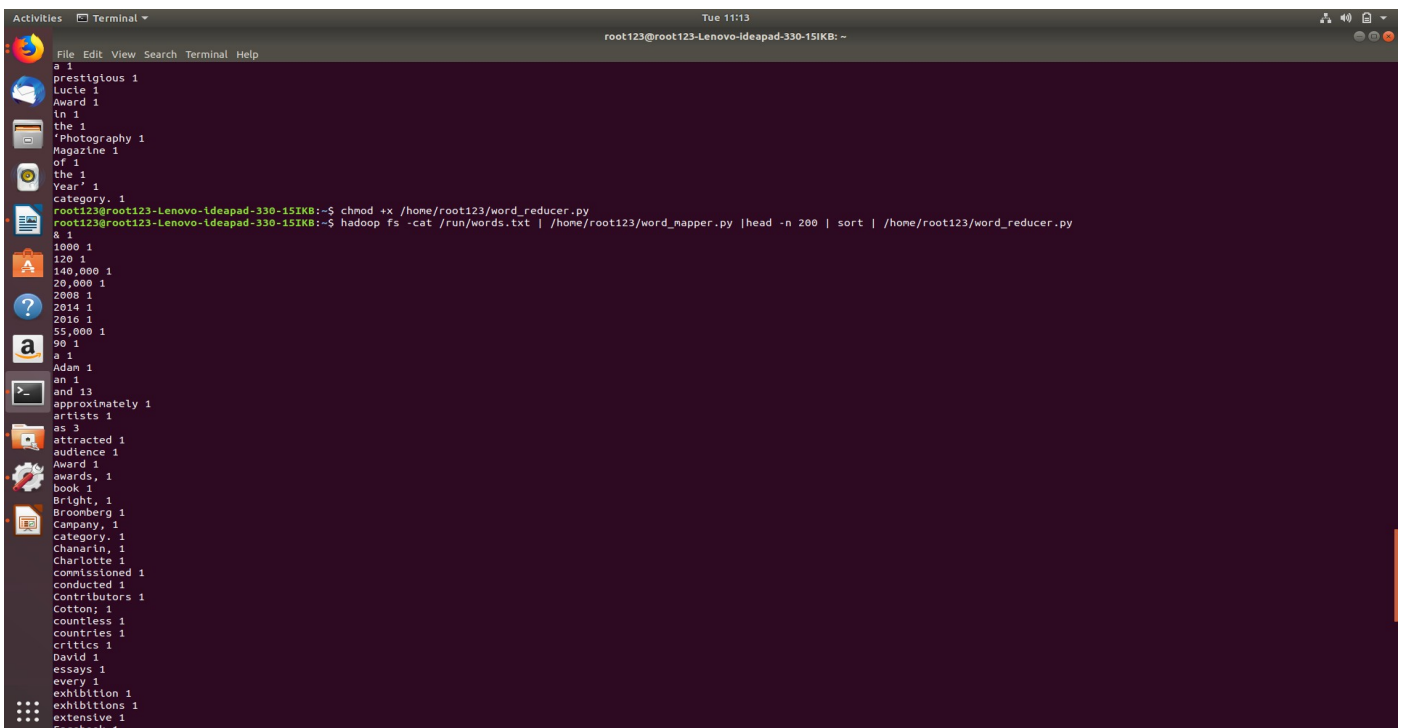
Step5:

commands:

```
-hadoop fs -put '/home/root123/Desktop/run/words.txt' /run  
-/usr/local/hadoop-3.0.3/etc/hadoop fs -mkdir  
-chmod +x '/home/root123/word_mapper.py'  
-hadoop fs -cat /run/words.txt | /home/root123/word_mapper.py  
-chmod +x /home/root123/word_reducer.py
```

Step6:

To get output of reducer after mapping.



The screenshot shows a terminal window with the following content:

```
Activities Terminal  
Tue 11:13  
root123@root123-Lenovo-Ideapad-330-15IKB: ~  
a 1  
prestigious 1  
Lucie 1  
Award 1  
in 1  
the 1  
Photography 1  
Magazine 1  
of 1  
the 1  
Year 1  
category. 1  
root123@root123-Lenovo-Ideapad-330-15IKB:~$ chmod +x /home/root123/word_reducer.py  
root123@root123-Lenovo-Ideapad-330-15IKB:~$ hadoop fs -cat /run/words.txt | /home/root123/word_mapper.py | head -n 200 | sort | /home/root123/word_reducer.py  
a 1  
1000 1  
120 1  
140,000 1  
20,000 1  
2000 1  
2014 1  
2016 1  
55,000 1  
90 1  
a 1  
Adam 1  
an 1  
and 13  
approximately 1  
artists 1  
as 3  
attracted 1  
audience 1  
Award 1  
awards, 1  
book 1  
Bright, 1  
Broomberg 1  
Company, 1  
category, 1  
Chanarin, 1  
Charlotte 1  
commissioned 1  
conducted 1  
Contributors 1  
Cotton; 1  
countless 1  
countries 1  
critics 1  
David 1  
essays 1  
every 1  
exhibition 1  
exhibitions 1  
extensive 1  
Fairbank 1
```

