```
root@hadoop-master:~# ./run-wordcount.sh
20/04/11 02:27:05 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.2:8032
20/04/11 02:27:06 INFO input.FileInputFormat: Total input paths to process : 2
20/04/11 02:27:06 INFO mapreduce.JobSubmitter: number of splits:2
20/04/11 02:27:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1586571991386_0001
20/04/11 02:27:07 INFO impl.YarnClientImpl: Submitted application application_1586571991386_0001
20/04/11 02:27:07 INFO mapreduce.Job: The url to track the job: http://hadoop-master:8088/proxy/application_1586571991386_
0001/
20/04/11 02:27:07 INFO mapreduce.Job: Running job: job_1586571991386_0001
20/04/11 02:27:16 INFO mapreduce.Job: Job job_1586571991386_0001 running in uber mode : false
20/04/11 02:27:16 INFO mapreduce.Job:  map 0% reduce 0%
20/04/11 02:27:27 INFO mapreduce.Job:  map 50% reduce 0%
20/04/11 02:27:28 INFO mapreduce.Job:  map 100% reduce 0%
20/04/11 02:27:34 INFO mapreduce.Job:  map 100% reduce 100%
20/04/11 02:27:34 INFO mapreduce.Job: Job job_1586571991386_0001 completed successfully
20/04/11 02:27:34 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=56
                FILE: Number of bytes written=352398
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=258
                HDFS: Number of bytes written=26
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=17577
                Total time spent by all reduces in occupied slots (ms)=4279
                Total time spent by all map tasks (ms)=17577
                Total time spent by all reduce tasks (ms)=4279
                Total vcore-milliseconds taken by all map tasks=17577
                Total vcore-milliseconds taken by all reduce tasks=4279
                Total megabyte-milliseconds taken by all map tasks=17998848
                Total megabyte-milliseconds taken by all reduce tasks=4381696
        Map-Reduce Framework
                Map input records=2
                Map output records=4
                Map output bytes=42
                Map output materialized bytes=62
                Input split bytes=232
                Combine input records=4
                Combine output records=4
                Reduce input groups=3
                Reduce shuffle bytes=62
                Reduce input records=4
                Reduce output records=3
                Spilled Records=8
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=132
                CPU time spent (ms)=1920
                Physical memory (bytes) snapshot=894713856
                Virtual memory (bytes) snapshot=2629046272
                Total committed heap usage (bytes)=524812288
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
```

```
        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=26
        File Output Format Counters
                Bytes Written=26

input file1.txt:
Hello Hadoop

input file2.txt:
Hello Docker

wordcount output:
Docker  1
Hadoop  1
Hello   2
root@hadoop-master:~#
```

1. First, put the local file to the Hadoop clusters. Then in Hadoop, each MapReduce task is initialized as a Job, and each Job can be divided into two phases: the map phase and the reduce phase. These two phases are represented by two functions, the map function and the reduce function. The map function receives an input in the form of <key, value>, and then also produces an intermediate output in the form of <key, value>. The Hadoop function receives an input in the form of <key, (list of="" values) =">, and then processes

the value set. Each reduce produces 0 or 1 output, and the output of reduce is also in the form of <key, value>. </key, value></key,(list></key, value></key, value>

Start container -> build master and slaves->put the file to Hadoop->call Job ->call Task->call map and reduce -> handle the data -> output

2. Upload the local file to the Hadoop cluster input file
3. 2 mapper and 1 reducer
4. Map 17577 and reduce 4279
5. Output file in HDFS and the content is that :

File name: part-r-00000

```
wordcount output:
Docker  1
Hadoop  1
Hello   2
root@hadoop-master:~#
```

Task2

1. 1 master and 4 slaves.

```
yyang199506@assignment1-yy:~/hadoop-cluster-docker$ sudo ./start-container.sh
start hadoop-master container...
start hadoop-slave1 container...
start hadoop-slave2 container...
start hadoop-slave3 container...
start hadoop-slave4 container...
root@hadoop-master:~#
```

2. The master used for controlling the whole cluster resource and the slaves only manage it own node resource.
3. It use 3 mapper and 1 reducer.
   In the start-container.sh, change the node number 3 to 5. And in the run-wordcount.sh add “mv text.txt input” after the mkdir input.

```
# create input files
mkdir input
echo "Hello Docker" >input/file2.txt
echo "Hello Hadoop" >input/file1.txt
mv text.txt input
```

Activities Terminal Sun 16:36 英

root@hadoop-master: ~

File Edit View Search Terminal Help

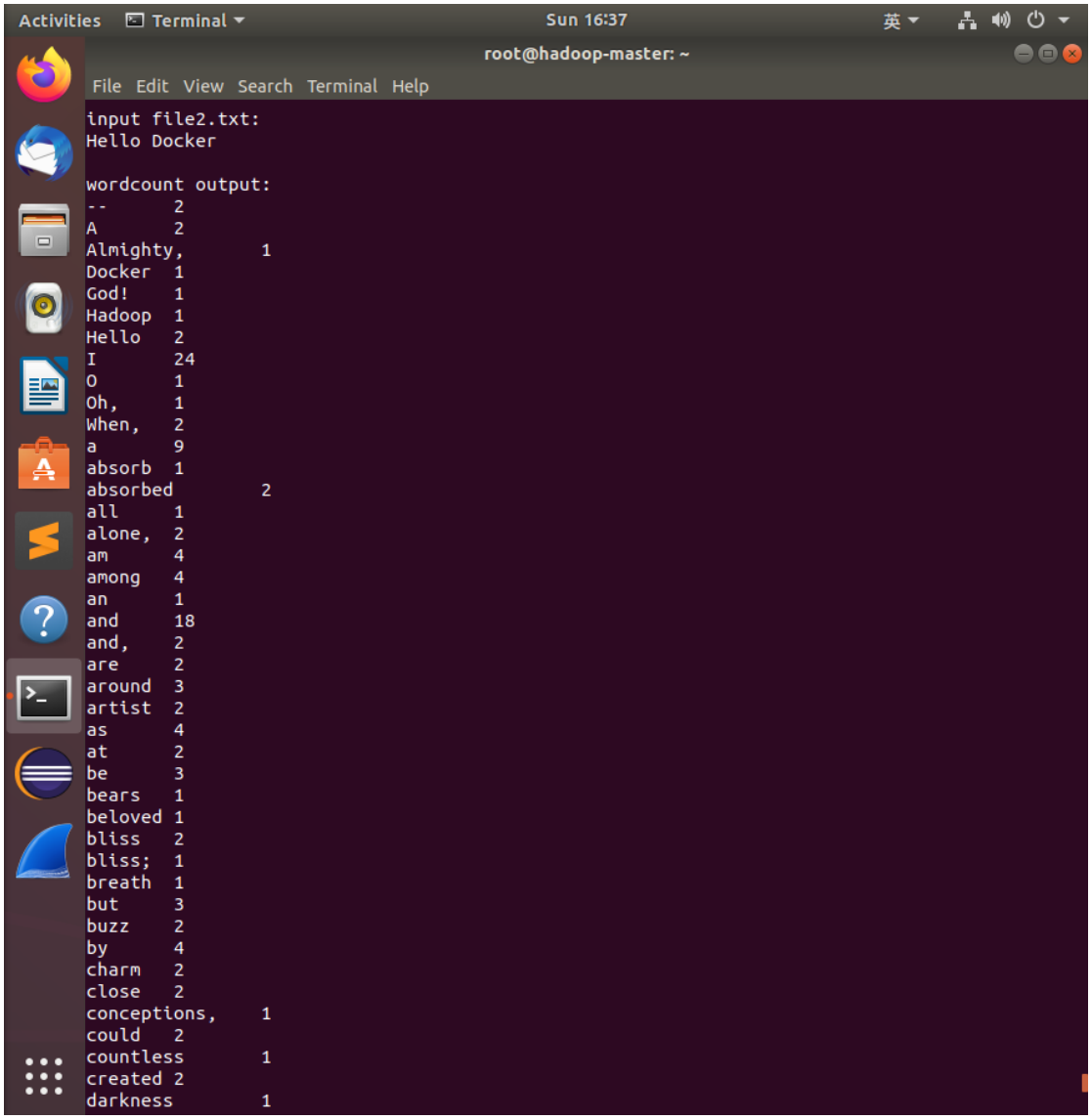
```
update-alternatives: using /usr/bin/vim.basic to provide /usr/bin/ex (ex) in auto mode
update-alternatives: using /usr/bin/vim.basic to provide /usr/bin/editor (editor) in auto
mode
Processing triggers for libc-bin (2.19-0ubuntu6.15) ...
root@hadoop-master:~# vi run-wordcount.sh
root@hadoop-master:~# ./run-wordcount.sh
20/04/12 20:32:16 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.
18.0.2:8032
20/04/12 20:32:17 INFO input.FileInputFormat: Total input paths to process : 3
20/04/12 20:32:17 INFO mapreduce.JobSubmitter: number of splits:3
20/04/12 20:32:17 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_158672335148
1_0001
20/04/12 20:32:17 INFO impl.YarnClientImpl: Submitted application application_158672335148
1_0001
20/04/12 20:32:17 INFO mapreduce.Job: The url to track the job: http://hadoop-master:8088/
proxy/application_1586723351481_0001/
20/04/12 20:32:17 INFO mapreduce.Job: Running job: job_1586723351481_0001
20/04/12 20:32:24 INFO mapreduce.Job: Job job_1586723351481_0001 running in uber mode : fa
lse
20/04/12 20:32:24 INFO mapreduce.Job:  map 0% reduce 0%
20/04/12 20:32:29 INFO mapreduce.Job:  map 33% reduce 0%
20/04/12 20:32:30 INFO mapreduce.Job:  map 100% reduce 0%
20/04/12 20:32:34 INFO mapreduce.Job:  map 100% reduce 100%
20/04/12 20:32:34 INFO mapreduce.Job: Job job_1586723351481_0001 completed successfully
20/04/12 20:32:34 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=2418
                FILE: Number of bytes written=474573
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2990
                HDFS: Number of bytes written=1629
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=3
                Launched reduce tasks=1
                Data-local map tasks=2
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=11982
                Total time spent by all reduces in occupied slots (ms)=2149
                Total time spent by all map tasks (ms)=11982
                Total time spent by all reduce tasks (ms)=2149
                Total vcore-milliseconds taken by all map tasks=11982
```

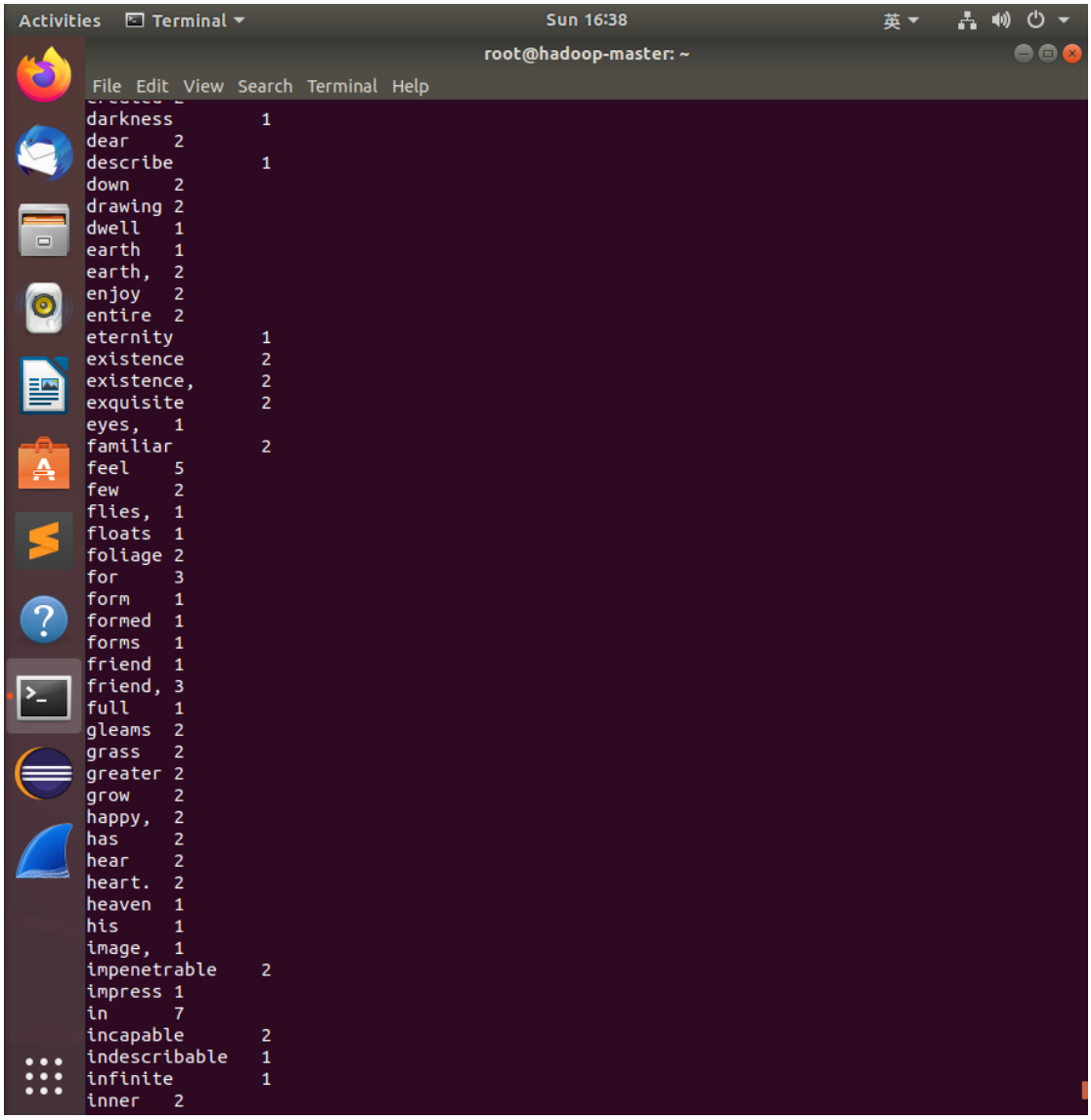
Activities Terminal Sun 16:37 英

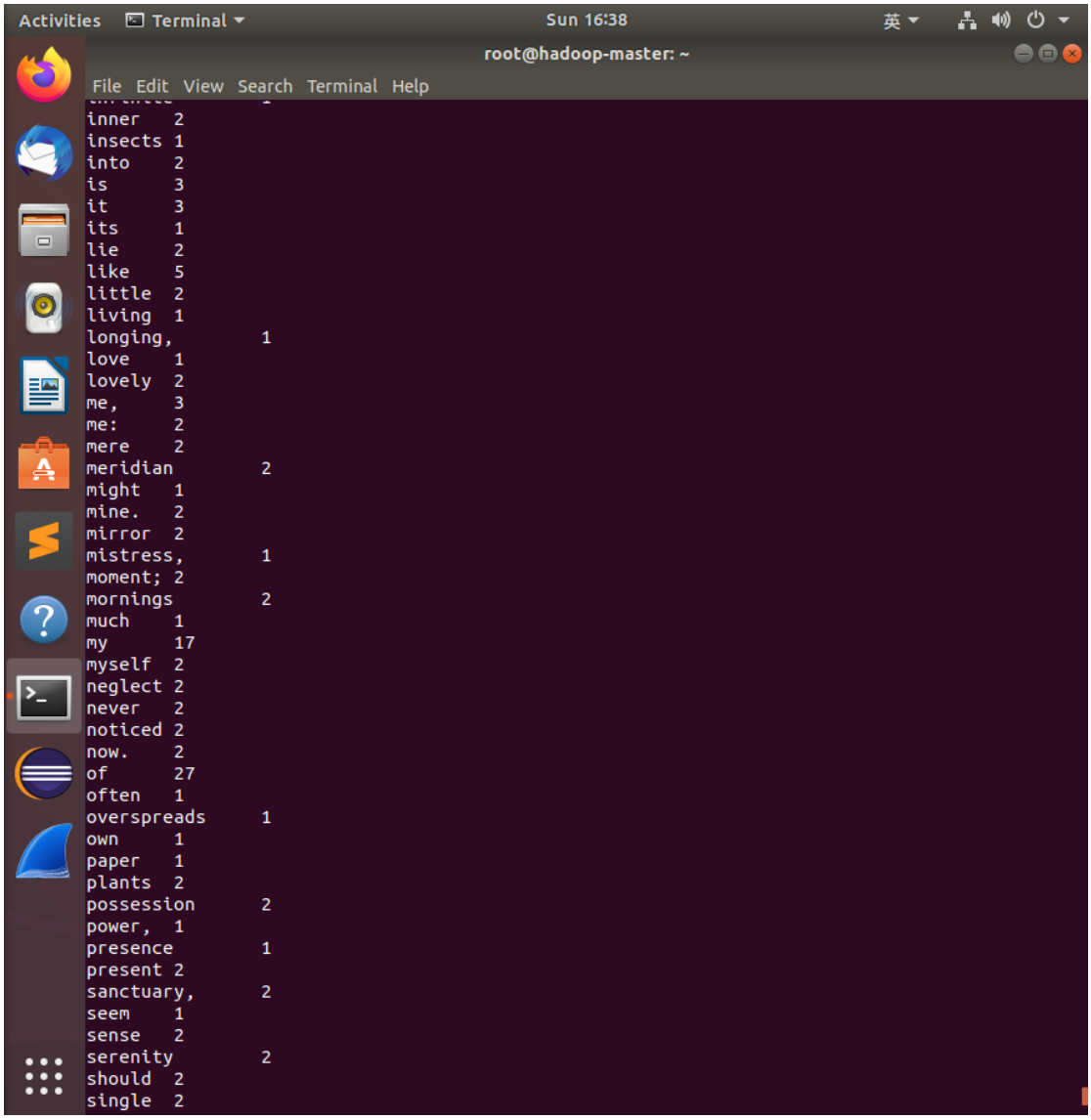
root@hadoop-master: ~

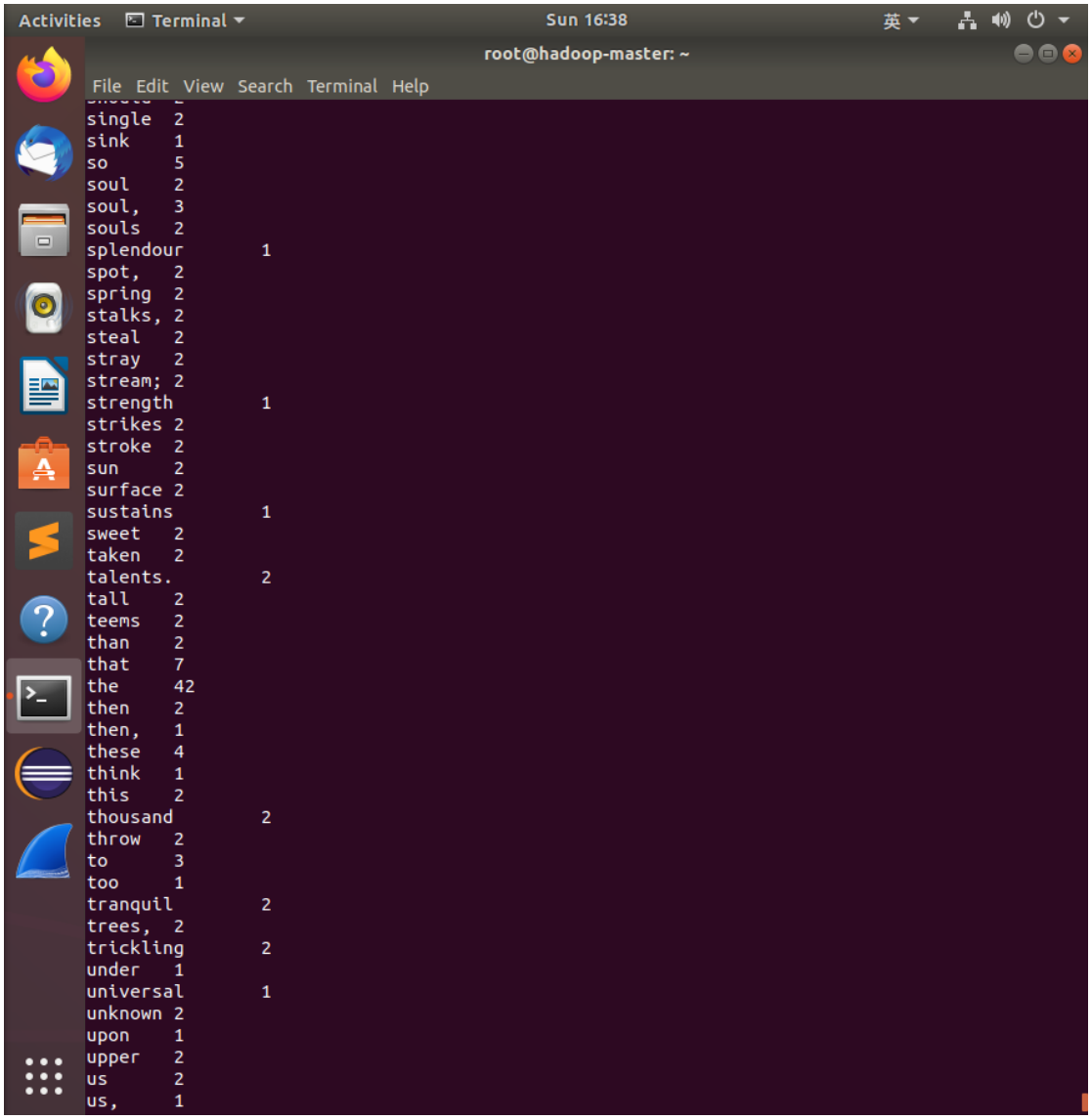
File Edit View Search Terminal Help

```
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=3
                Launched reduce tasks=1
                Data-local map tasks=2
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=11982
                Total time spent by all reduces in occupied slots (ms)=2149
                Total time spent by all map tasks (ms)=11982
                Total time spent by all reduce tasks (ms)=2149
                Total vcore-milliseconds taken by all map tasks=11982
                Total vcore-milliseconds taken by all reduce tasks=2149
                Total megabyte-milliseconds taken by all map tasks=12269568
                Total megabyte-milliseconds taken by all reduce tasks=2200576
        Map-Reduce Framework
                Map input records=3
                Map output records=497
                Map output bytes=4631
                Map output materialized bytes=2430
                Input split bytes=347
                Combine input records=497
                Combine output records=195
                Reduce input groups=194
                Reduce shuffle bytes=2430
                Reduce input records=195
                Reduce output records=194
                Spilled Records=390
                Shuffled Maps =3
                Failed Shuffles=0
                Merged Map outputs=3
                GC time elapsed (ms)=118
                CPU time spent (ms)=1810
                Physical memory (bytes) snapshot=1208229888
                Virtual memory (bytes) snapshot=3511894016
                Total committed heap usage (bytes)=764936192
        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=2643
        File Output Format Counters
                Bytes Written=1629
```


Activities
Terminal
Sun 16:37
root@hadoop-master: ~
File Edit View Search Terminal Help
input file2.txt:
Hello Docker
wordcount output:
-- 2
A 2
Almighty, 1
Docker 1
God! 1
Hadoop 1
Hello 2
I 24
O 1
Oh, 1
When, 2
a 9
absorb 1
absorbed 2
all 1
alone, 2
am 4
among 4
an 1
and 18
and, 2
are 2
around 3
artist 2
as 4
at 2
be 3
bears 1
beloved 1
bliss 2
bliss; 1
breath 1
but 3
buzz 2
by 4
charm 2
close 2
conceptions, 1
could 2
countless 1
created 2
darkness 1


Activities
Terminal
Sun 16:38
root@hadoop-master: ~
File Edit View Search Terminal Help
darkness 1
dear 2
describe 1
down 2
drawing 2
dwell 1
earth 1
earth, 2
enjoy 2
entire 2
eternity 1
existence 2
existence, 2
exquisite 2
eyes, 1
familiar 2
feel 5
few 2
flies, 1
floats 1
foliage 2
for 3
form 1
formed 1
forms 1
friend 1
friend, 3
full 1
gleams 2
grass 2
greater 2
grow 2
happy, 2
has 2
hear 2
heart. 2
heaven 1
his 1
image, 1
impenetrable 2
impress 1
in 7
incapable 2
indescribable 1
infinite 1
inner 2


Activities
Terminal
Sun 16:38
root@hadoop-master: ~
File Edit View Search Terminal Help
inner 2
insects 1
into 2
is 3
it 3
its 1
lie 2
like 5
little 2
living 1
longing, 1
love 1
lovely 2
me, 3
me: 2
mere 2
meridian 2
might 1
mine. 2
mirror 2
mistress, 1
moment; 2
mornings 2
much 1
my 17
myself 2
neglect 2
never 2
noticed 2
now. 2
of 27
often 1
overspreads 1
own 1
paper 1
plants 2
possession 2
power, 1
presence 1
present 2
sanctuary, 2
seem 1
sense 2
serenity 2
should 2
single 2



```
single  2
sink    1
so      5
soul    2
soul,   3
souls   2
splendour       1
spot,   2
spring  2
stalks, 2
steal   2
stray   2
stream; 2
strength        1
strikes 2
stroke  2
sun     2
surface 2
sustains        1
sweet   2
taken   2
talents.        2
tall    2
teems   2
than    2
that    7
the     42
then    2
then,   1
these   4
think   1
this    2
thousand        2
throw   2
to      3
too     1
tranquil        2
trees,  2
trickling       2
under   1
universal       1
unknown 2
upon    1
upper   2
us      2
us,     1
```

```
us      2
us,     1
valley  2
vapour  2
visions!        1
warm    1
was     4
weight  1
when    3
which   5
while   2
who     1
whole   2
with    7
within  1
wonderful       2
world   2
would   1
yet     2
root@hadoop-master:~#
```

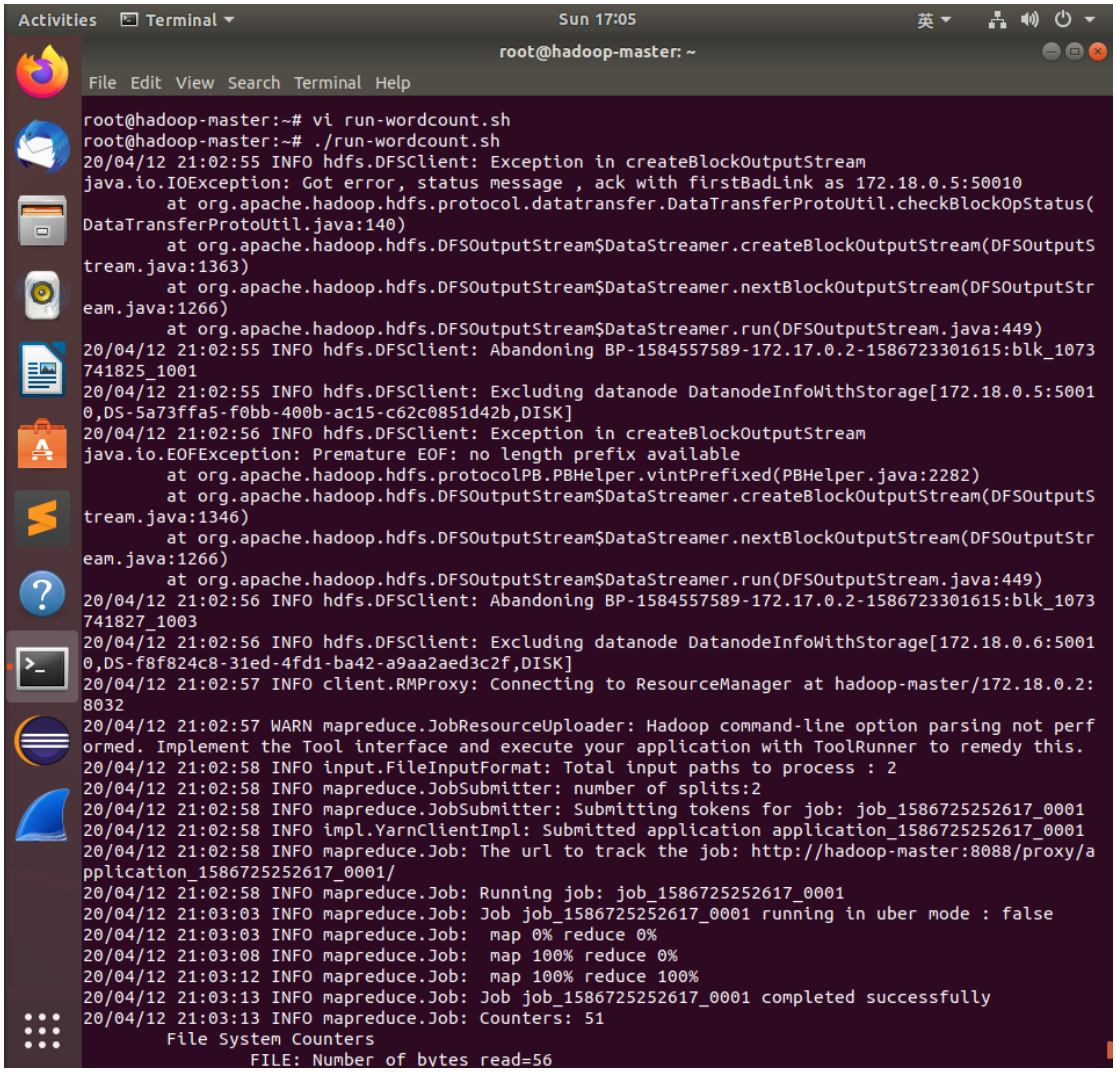
4. The total time of map is 11982 and total time of reduce is 2149
5. “the” occurred 42 times and “of” 27 times.

Task 3

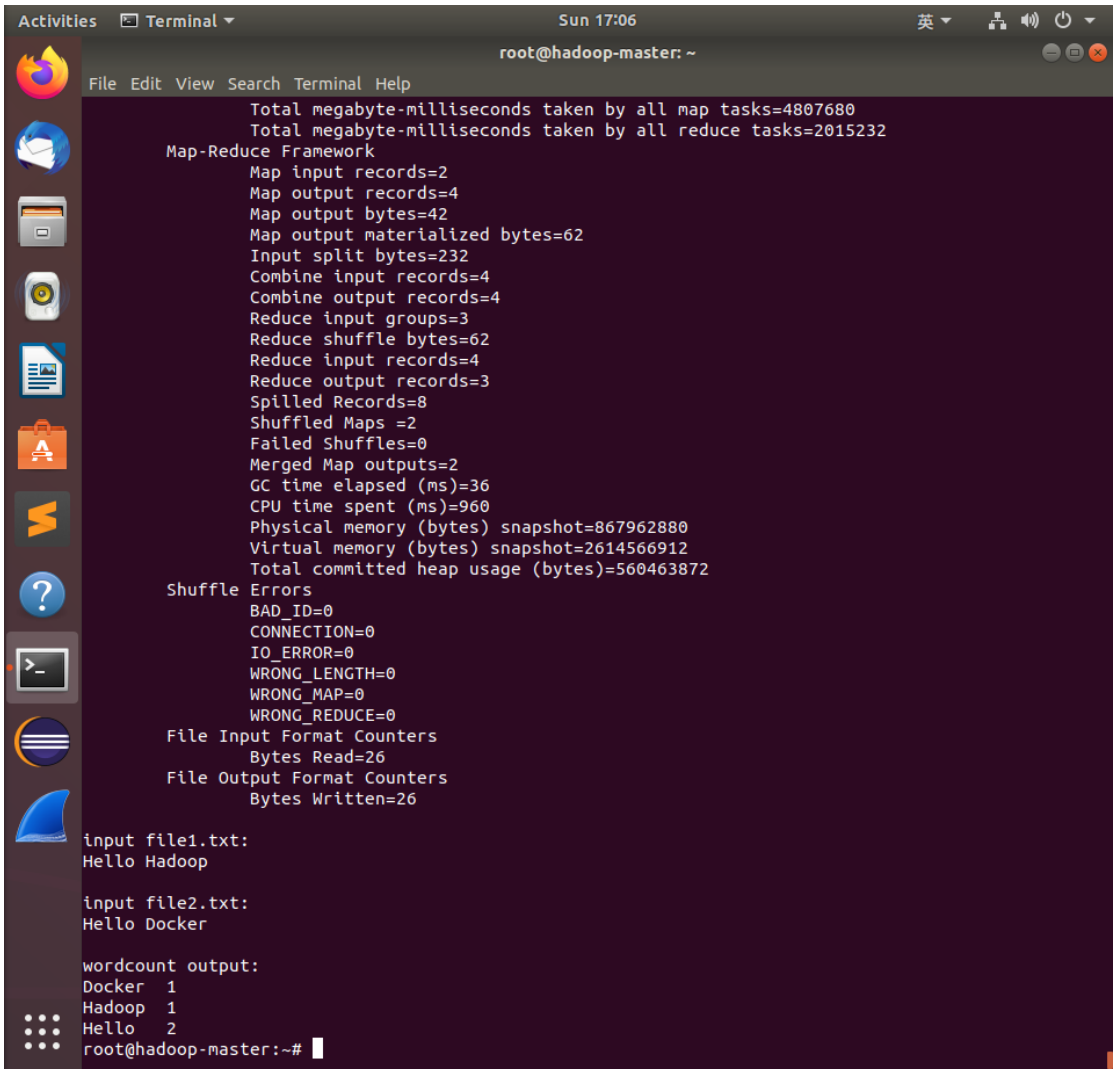
1. The Map process we need to inherit org. apache. Hadoop. Graphs Mapper classes in the package, and rewrite the Map method. By adding two sentences to the map method to output the key and value values to the console, you can see that the value in the map method stores a line from the text file (marked with a carriage return), and the key value is the offset of the line's initial letter relative to the first address of the text file. The StringTokenizer class then splits each row into individual words, outputs <word,1> as the result of the map method, and hands over the rest to the MapReduce framework.</word,1>

2. Reduce processes need the org.apache hadoop. The graphs Reducer class in the package, and rewrite the Reduce method. In the Map process output <key,values>, the key is a single word, and values is the list composed of the counting values of the corresponding words. The output of the Map is the input of Reduce. Therefore, as long as the Reduce method traverses values and sums, the total number of times of a certain word can be obtained.</key,values>

3. 4 mapper and 3 reducer



```
20/04/12 21:03:13 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=56
                FILE: Number of bytes written=351699
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=258
                HDFS: Number of bytes written=26
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=4695
                Total time spent by all reduces in occupied slots (ms)=1968
                Total time spent by all map tasks (ms)=4695
                Total time spent by all reduce tasks (ms)=1968
                Total vcore-milliseconds taken by all map tasks=4695
                Total vcore-milliseconds taken by all reduce tasks=1968
                Total megabyte-milliseconds taken by all map tasks=4807680
                Total megabyte-milliseconds taken by all reduce tasks=2015232
        Map-Reduce Framework
                Map input records=2
                Map output records=4
                Map output bytes=42
                Map output materialized bytes=62
                Input split bytes=232
                Combine input records=4
                Combine output records=4
                Reduce input groups=3
                Reduce shuffle bytes=62
                Reduce input records=4
                Reduce output records=3
                Spilled Records=8
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=36
                CPU time spent (ms)=960
                Physical memory (bytes) snapshot=867962880
                Virtual memory (bytes) snapshot=2614566912
                Total committed heap usage (bytes)=560463872
```



4. The time of map is 4695 and the time of reduce is 1968.

Task 4