EEL 6761

Assign1
Word Count

Github link: https://github.com/jeness/cloudComputingAssignment1

Haoran Yu
UFID:18106994

Task1

1. Single word count

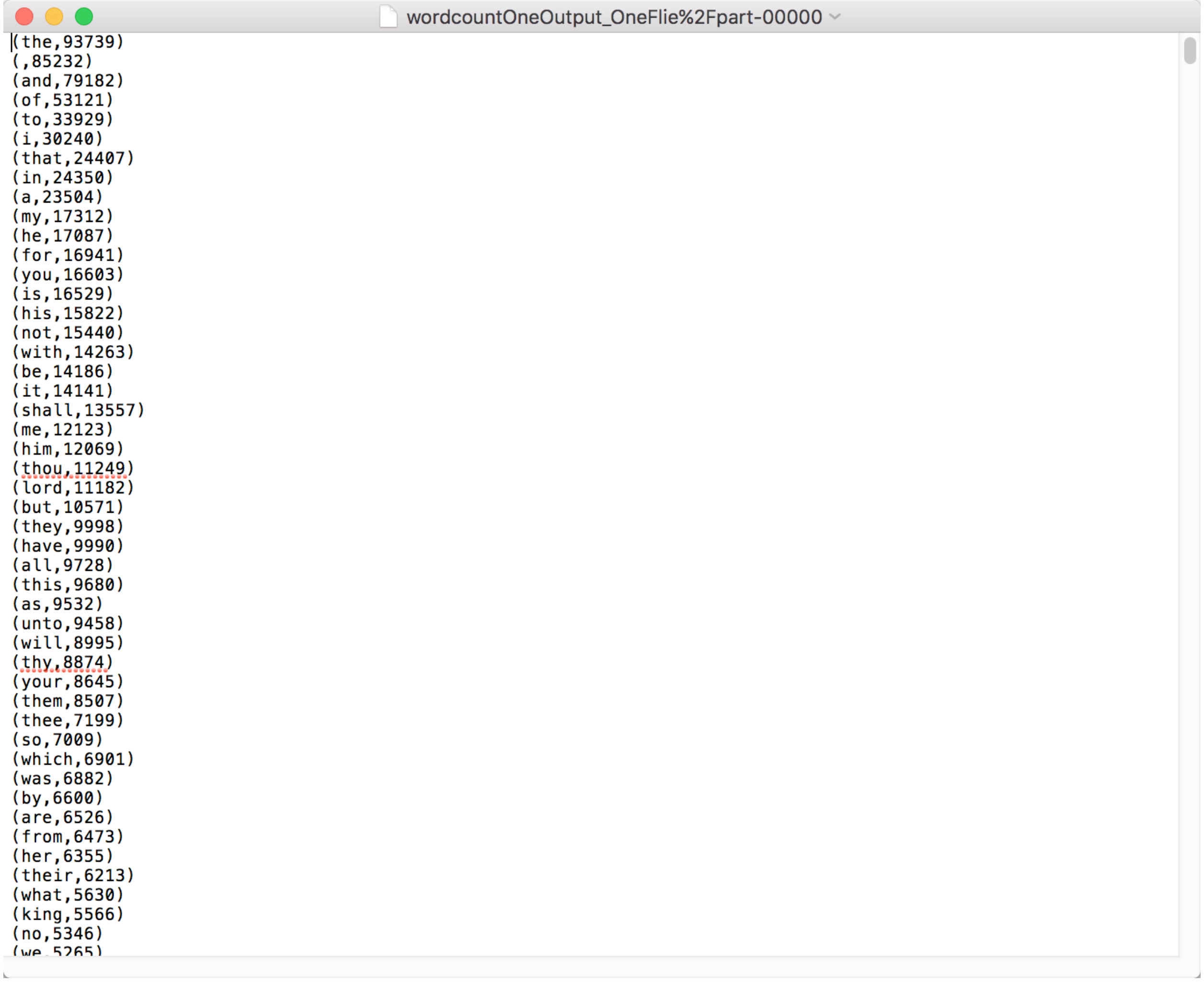
First, load bible data from google cloud storage bucket to RDD, split words with spaces, and then use filter to get non-space elements. Second, map elements to key-value pair like, <word, 1>. Third, reduce pair by the same key and get the result of counting words in descending order.

2. Result

Screenshots shows the result file of single word count which is download from google cloud storage.

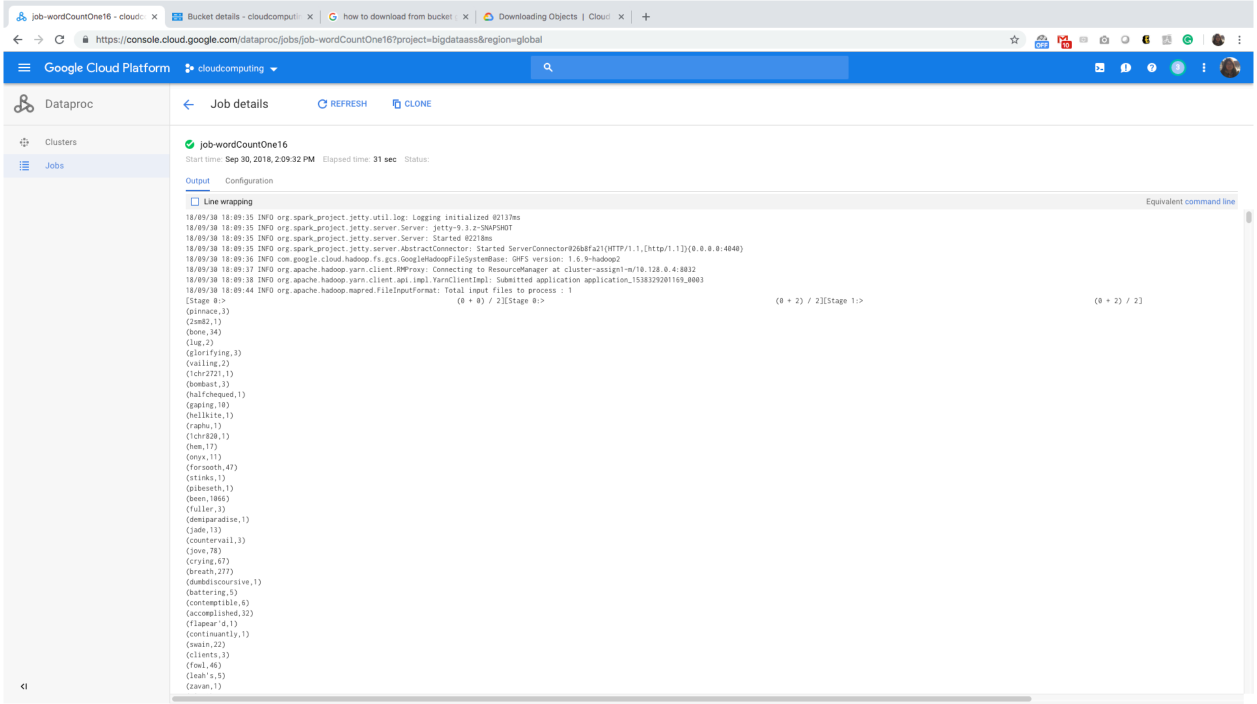
File link:

https://storage.googleapis.com/ufcloudcomputing/wordcountOneOutput_OneFlie/part-00000



```
(the,93739)
(,85232)
(and,79182)
(of,53121)
(to,33929)
(i,30240)
(that,24407)
(in,24350)
(a,23504)
(my,17312)
(he,17087)
(for,16941)
(you,16603)
(is,16529)
(his,15822)
(not,15440)
(with,14263)
(be,14186)
(it,14141)
(shall,13557)
(me,12123)
(him,12069)
(thou,11249)
(lord,11182)
(but,10571)
(they,9998)
(have,9990)
(all,9728)
(this,9680)
(as,9532)
(unto,9458)
(will,8995)
(thy,8874)
(your,8645)
(them,8507)
(thee,7199)
(so,7009)
(which,6901)
(was,6882)
(by,6600)
(are,6526)
(from,6473)
(her,6355)
(their,6213)
(what,5630)
(king,5566)
(no,5346)
(we,5265)
```

3. Output console screenshot


job-wordCountOne16 - cloudc
Bucket details - cloudcomputi
how to download from bucket
Downloading Objects | Cloud
https://console.cloud.google.com/dataproc/jobs/job-wordCountOne16?project=bigdataass&region=global
Google Cloud Platform
cloudcomputing
Dataproc
Clusters
Jobs
Job details
REFRESH
CLONE
Output
Configuration
Equivalent command line

job-wordCountOne16
Start time: Sep 30, 2018, 2:09:32 PM Elapsed time: 31 sec Status:

□ Line wrapping

```
18/09/30 18:09:35 INFO org.spark_project.jetty.util.log: Logging initialized @2137ms
18/09/30 18:09:35 INFO org.spark_project.jetty.server.Server: jetty-9.3.z-SNAPSHOT
18/09/30 18:09:35 INFO org.spark_project.jetty.server.Server: Started @2218ms
18/09/30 18:09:35 INFO org.spark_project.jetty.server.AbstractConnector: Started ServerConnector@26b8fa21{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
18/09/30 18:09:36 INFO com.google.cloud.hadoop.fs.gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.9-hadoop2
18/09/30 18:09:37 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at cluster-assign1-m/10.128.0.4:8032
18/09/30 18:09:38 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1538329201169_0003
18/09/30 18:09:44 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1
[Stage 0:>                                                          (0 + 0) / 2][Stage 0:>                                                          (0 + 2) / 2][Stage 1:>                                                          (0 + 2) / 2]
(pinnace,3)
(2sm82,1)
(bone,34)
(lug,2)
(glorifying,3)
(vailing,2)
(1chr2721,1)
(bombast,3)
(halfchequed,1)
(gaping,10)
(hellkite,1)
(raphu,1)
(1chr820,1)
(hem,17)
(onyx,11)
(forsooth,47)
(stinks,1)
(pibeseth,1)
(been,1066)
(fuller,3)
(demiparadise,1)
(jade,13)
(countervail,3)
(jove,78)
(crying,67)
(breath,277)
(dumbdiscoursive,1)
(battering,5)
(contemptible,6)
(accomplished,32)
(flapear'd,1)
(continuantly,1)
(swain,22)
(clients,3)
(fowl,46)
(leah's,5)
(zavan,1)
```

Task 2

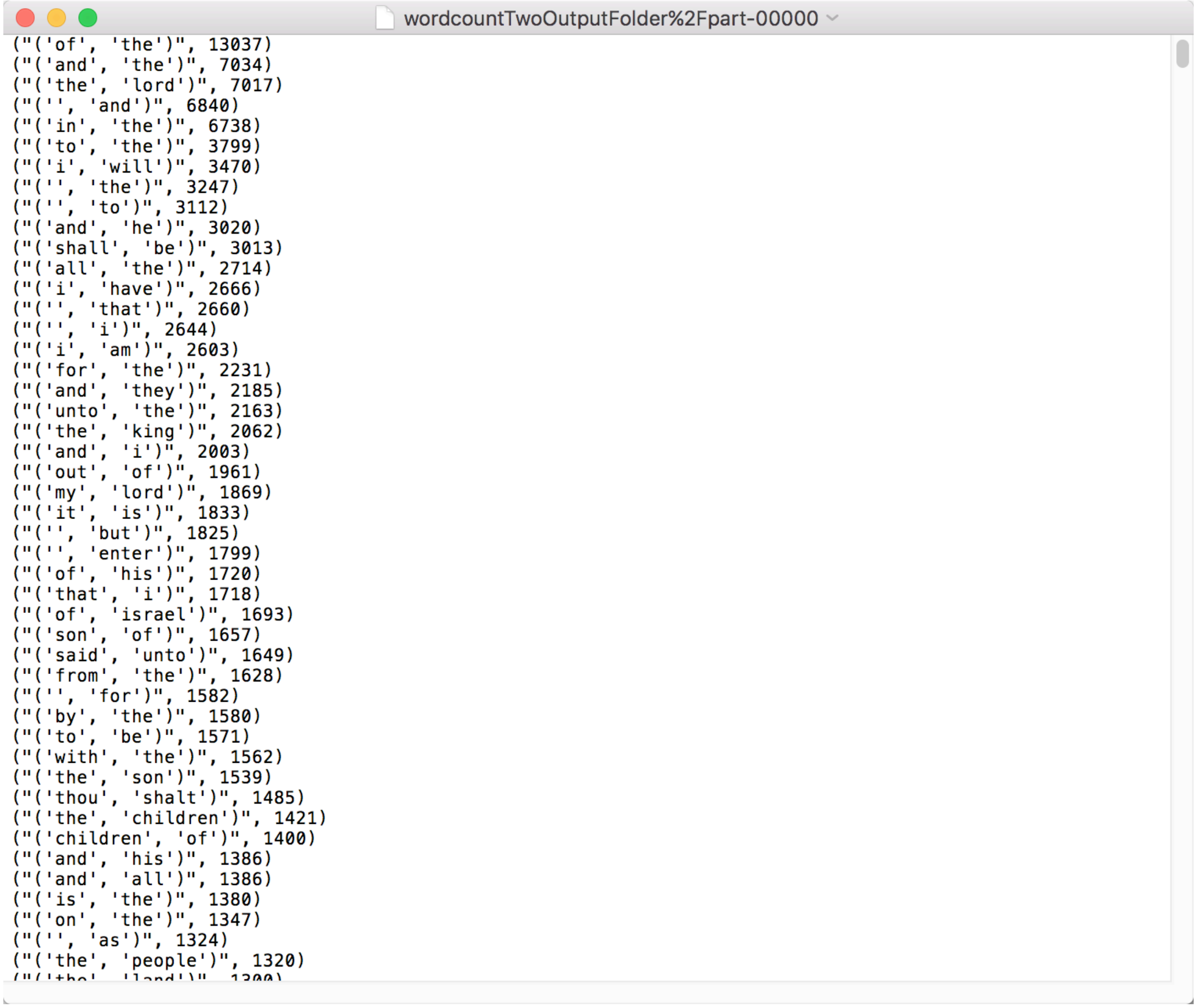
1. Double word count

First, load bible data from google cloud storage bucket to RDD, split words with spaces, then use function to get current word and the word after current word, and store every new double words combination to a new list. Second, map elements to key-value pair like, <<word 1, word 2>, 1>. Third, reduce pair by the same key and get the result of counting words with descending order.

2. Result

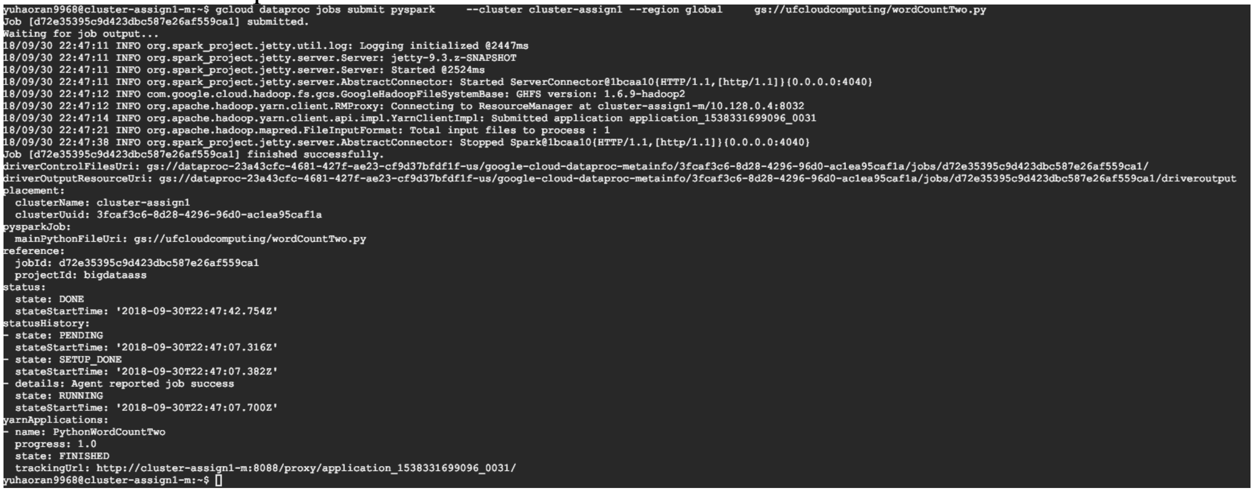
Screenshots shows the result file of double word count which is download from google cloud storage.

File Link:
https://storage.googleapis.com/ufcloudcomputing/wordcountTwoOutputFolder/part-00000



```
("('of', 'the')", 13037)
("('and', 'the')", 7034)
("('the', 'lord')", 7017)
("('', 'and')", 6840)
("('in', 'the')", 6738)
("('to', 'the')", 3799)
("('i', 'will')", 3470)
("('', 'the')", 3247)
("('', 'to')", 3112)
("('and', 'he')", 3020)
("('shall', 'be')", 3013)
("('all', 'the')", 2714)
("('i', 'have')", 2666)
("('', 'that')", 2660)
("('', 'i')", 2644)
("('i', 'am')", 2603)
("('for', 'the')", 2231)
("('and', 'they')", 2185)
("('unto', 'the')", 2163)
("('the', 'king')", 2062)
("('and', 'i')", 2003)
("('out', 'of')", 1961)
("('my', 'lord')", 1869)
("('it', 'is')", 1833)
("('', 'but')", 1825)
("('', 'enter')", 1799)
("('of', 'his')", 1720)
("('that', 'i')", 1718)
("('of', 'israel')", 1693)
("('son', 'of')", 1657)
("('said', 'unto')", 1649)
("('from', 'the')", 1628)
("('', 'for')", 1582)
("('by', 'the')", 1580)
("('to', 'be')", 1571)
("('with', 'the')", 1562)
("('the', 'son')", 1539)
("('thou', 'shalt')", 1485)
("('the', 'children')", 1421)
("('children', 'of')", 1400)
("('and', 'his')", 1386)
("('and', 'all')", 1386)
("('is', 'the')", 1380)
("('on', 'the')", 1347)
("('', 'as')", 1324)
("('the', 'people')", 1320)
("('the', 'land')", 1300)
```

3. Console output screenshot



```
yuhaoran9968@cluster-assign1-m:~$ gcloud dataproc jobs submit pyspark     --cluster cluster-assign1 --region global     gs://ufcloudcomputing/wordCountTwo.py
Job [d72e35395c9d423dbc587e26af559ca1] submitted.
Waiting for job output...
18/09/30 22:47:11 INFO org.spark_project.jetty.util.log: Logging initialized @2447ms
18/09/30 22:47:11 INFO org.spark_project.jetty.server.Server: jetty-9.3.z-SNAPSHOT
18/09/30 22:47:11 INFO org.spark_project.jetty.server.Server: Started @2524ms
18/09/30 22:47:11 INFO org.spark_project.jetty.server.AbstractConnector: Started ServerConnector@1bcaa10{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
18/09/30 22:47:12 INFO com.google.cloud.hadoop.fs.gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.9-hadoop2
18/09/30 22:47:12 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at cluster-assign1-m/10.128.0.4:8032
18/09/30 22:47:14 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1538331699096_0031
18/09/30 22:47:21 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1
18/09/30 22:47:38 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@1bcaa10{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
Job [d72e35395c9d423dbc587e26af559ca1] finished successfully.
driverControlFilesUri: gs://dataproc-23a43cfc-4681-427f-ae23-cf9d37bfdf1f-us/google-cloud-dataproc-metainfo/3fcaf3c6-8d28-4296-96d0-ac1ea95caf1a/jobs/d72e35395c9d423dbc587e26af559ca1/
driverOutputResourceUri: gs://dataproc-23a43cfc-4681-427f-ae23-cf9d37bfdf1f-us/google-cloud-dataproc-metainfo/3fcaf3c6-8d28-4296-96d0-ac1ea95caf1a/jobs/d72e35395c9d423dbc587e26af559ca1/driveroutput
placement:
  clusterName: cluster-assign1
  clusterUuid: 3fcaf3c6-8d28-4296-96d0-ac1ea95caf1a
pysparkJob:
  mainPythonFileUri: gs://ufcloudcomputing/wordCountTwo.py
reference:
  jobId: d72e35395c9d423dbc587e26af559ca1
  projectId: bigdataass
status:
  state: DONE
  stateStartTime: '2018-09-30T22:47:42.754Z'
statusHistory:
- state: PENDING
  stateStartTime: '2018-09-30T22:47:07.316Z'
- state: SETUP_DONE
  stateStartTime: '2018-09-30T22:47:07.382Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2018-09-30T22:47:07.700Z'
yarnApplications:
- name: PythonWordCountTwo
  progress: 1.0
  state: FINISHED
  trackingUrl: http://cluster-assign1-m:8088/proxy/application_1538331699096_0031/
yuhaoran9968@cluster-assign1-m:~$ 
```

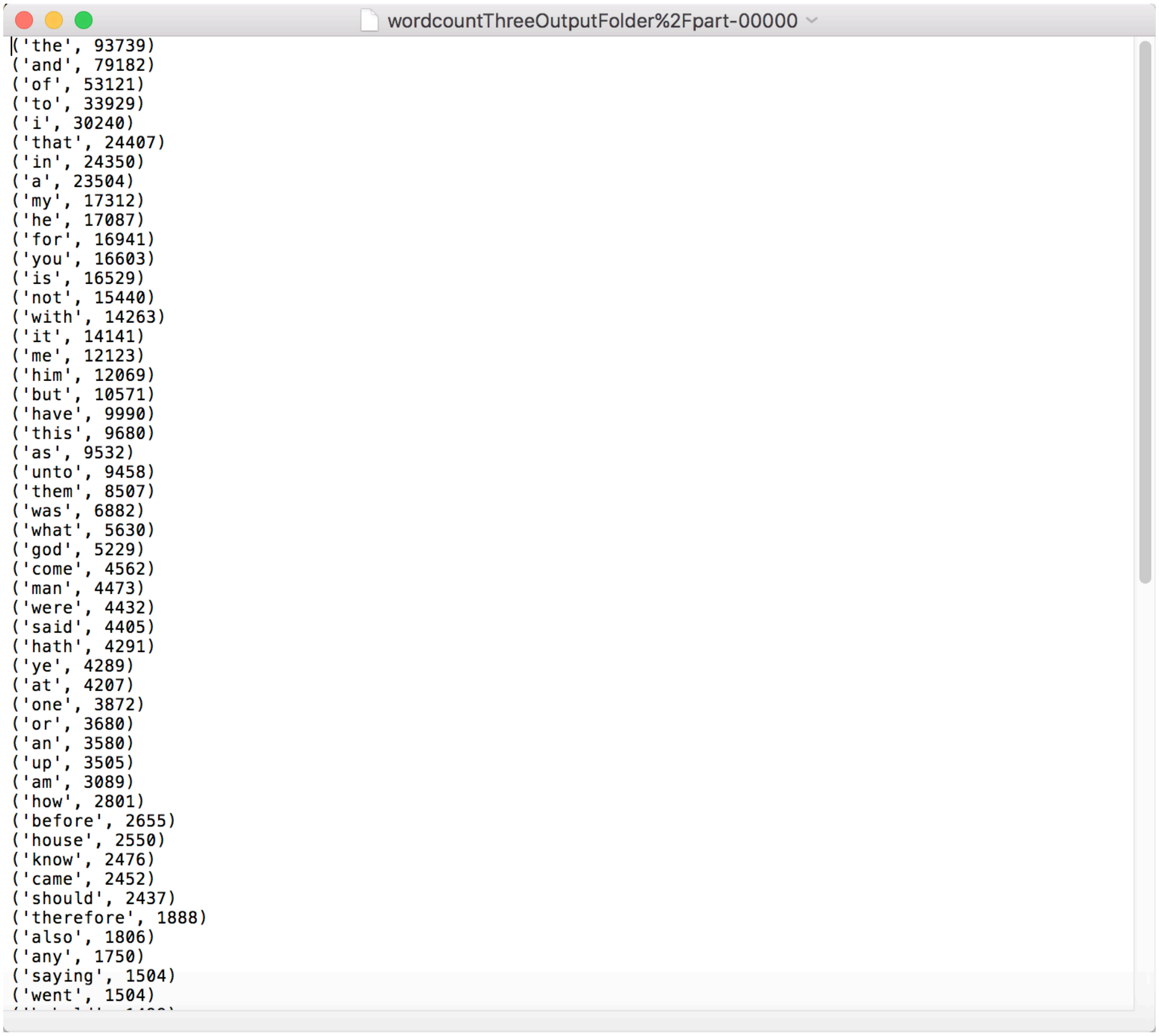
Task 3

1. Word count with distributed cache

First, load bible data and small list data from google cloud storage bucket to 2 different RDD, split words with spaces, and then use filter to get non-space elements. Second, map bible data to key-value pair like, <word, 1> and reduce pair by the same key and get the result of counting words in bible. Third, map small list data to key-value pair like, <word, 0> and join two RDD and get the result of common words and the respect value from bible RDD with descending order.

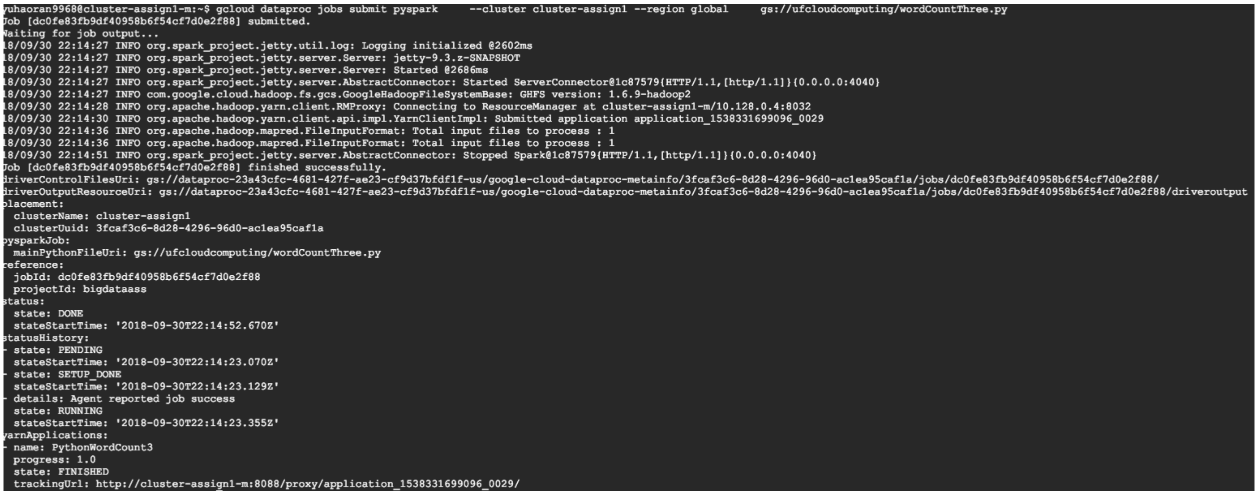
2. Result

Screenshots shows the result file of distributed cache word count which is download from google cloud storage.
Link file:
https://storage.googleapis.com/ufcloudcomputing/wordcountThreeOutputFolder/part-00000



```
('the', 93739)
('and', 79182)
('of', 53121)
('to', 33929)
('i', 30240)
('that', 24407)
('in', 24350)
('a', 23504)
('my', 17312)
('he', 17087)
('for', 16941)
('you', 16603)
('is', 16529)
('not', 15440)
('with', 14263)
('it', 14141)
('me', 12123)
('him', 12069)
('but', 10571)
('have', 9990)
('this', 9680)
('as', 9532)
('unto', 9458)
('them', 8507)
('was', 6882)
('what', 5630)
('god', 5229)
('come', 4562)
('man', 4473)
('were', 4432)
('said', 4405)
('hath', 4291)
('ye', 4289)
('at', 4207)
('one', 3872)
('or', 3680)
('an', 3580)
('up', 3505)
('am', 3089)
('how', 2801)
('before', 2655)
('house', 2550)
('know', 2476)
('came', 2452)
('should', 2437)
('therefore', 1888)
('also', 1806)
('any', 1750)
('saying', 1504)
('went', 1504)
```

3. Console output screenshot

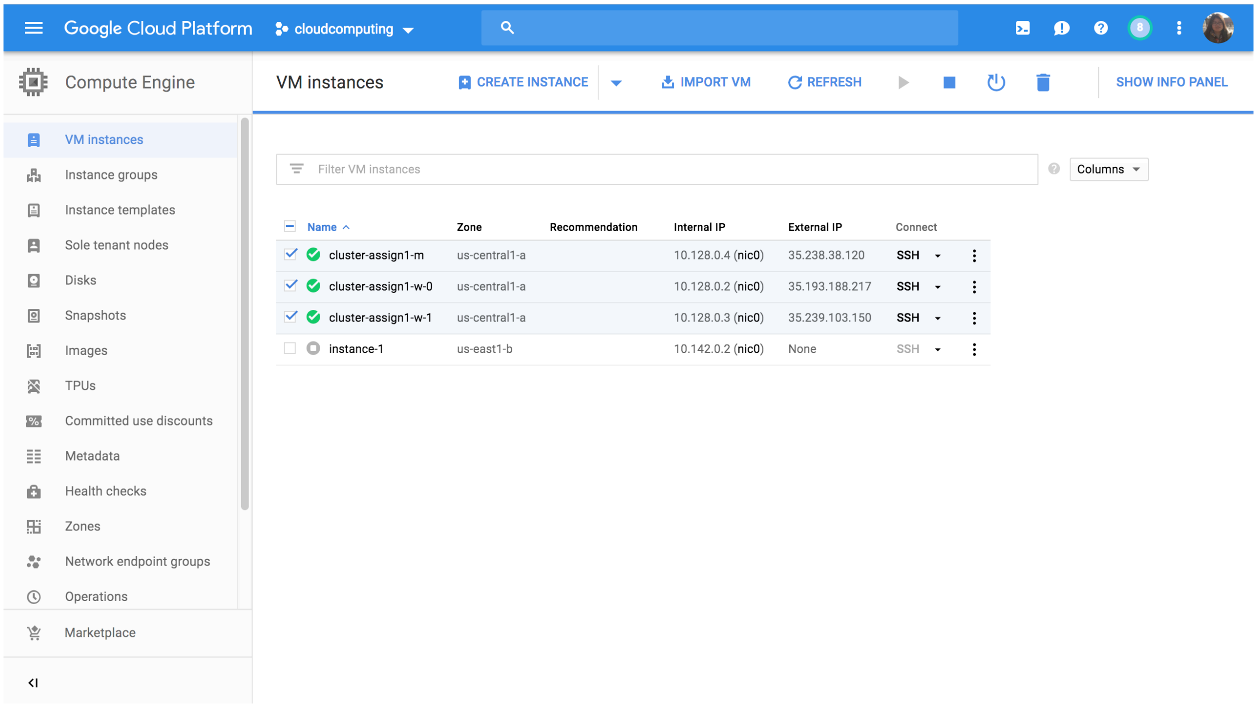


```
yuhaoran9968@cluster-assign1-m:~$ gcloud dataproc jobs submit pyspark     --cluster cluster-assign1 --region global     gs://ufcloudcomputing/wordCountThree.py
Job [dc0fe83fb9df40958b6f54cf7d0e2f88] submitted.
Waiting for job output...
18/09/30 22:14:27 INFO org.spark_project.jetty.util.log: Logging initialized @2602ms
18/09/30 22:14:27 INFO org.spark_project.jetty.server.Server: jetty-9.3.z-SNAPSHOT
18/09/30 22:14:27 INFO org.spark_project.jetty.server.Server: Started @2686ms
18/09/30 22:14:27 INFO org.spark_project.jetty.server.AbstractConnector: Started ServerConnector@1c87579{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
18/09/30 22:14:27 INFO com.google.cloud.hadoop.fs.gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.9-hadoop2
18/09/30 22:14:28 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at cluster-assign1-m/10.128.0.4:8032
18/09/30 22:14:30 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1538331699096_0029
18/09/30 22:14:36 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1
18/09/30 22:14:36 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1
18/09/30 22:14:51 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@1c87579{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
Job [dc0fe83fb9df40958b6f54cf7d0e2f88] finished successfully.
driverControlFilesUri: gs://dataproc-23a43cfc-4681-427f-ae23-cf9d37bfdf1f-us/google-cloud-dataproc-metainfo/3fcaf3c6-8d28-4296-96d0-ac1ea95caf1a/jobs/dc0fe83fb9df40958b6f54cf7d0e2f88/
driverOutputResourceUri: gs://dataproc-23a43cfc-4681-427f-ae23-cf9d37bfdf1f-us/google-cloud-dataproc-metainfo/3fcaf3c6-8d28-4296-96d0-ac1ea95caf1a/jobs/dc0fe83fb9df40958b6f54cf7d0e2f88/driveroutput
placement:
  clusterName: cluster-assign1
  clusterUuid: 3fcaf3c6-8d28-4296-96d0-ac1ea95caf1a
pysparkJob:
  mainPythonFileUri: gs://ufcloudcomputing/wordCountThree.py
reference:
  jobId: dc0fe83fb9df40958b6f54cf7d0e2f88
  projectId: bigdataass
status:
  state: DONE
  stateStartTime: '2018-09-30T22:14:52.670Z'
statusHistory:
- state: PENDING
  stateStartTime: '2018-09-30T22:14:23.070Z'
- state: SETUP_DONE
  stateStartTime: '2018-09-30T22:14:23.129Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2018-09-30T22:14:23.355Z'
yarnApplications:
- name: PythonWordCount3
  progress: 1.0
  state: FINISHED
  trackingUrl: http://cluster-assign1-m:8088/proxy/application_1538331699096_0029/
```

Configurations

1. VM instances

Three VM instances in total. One master node, and two worker nodes.



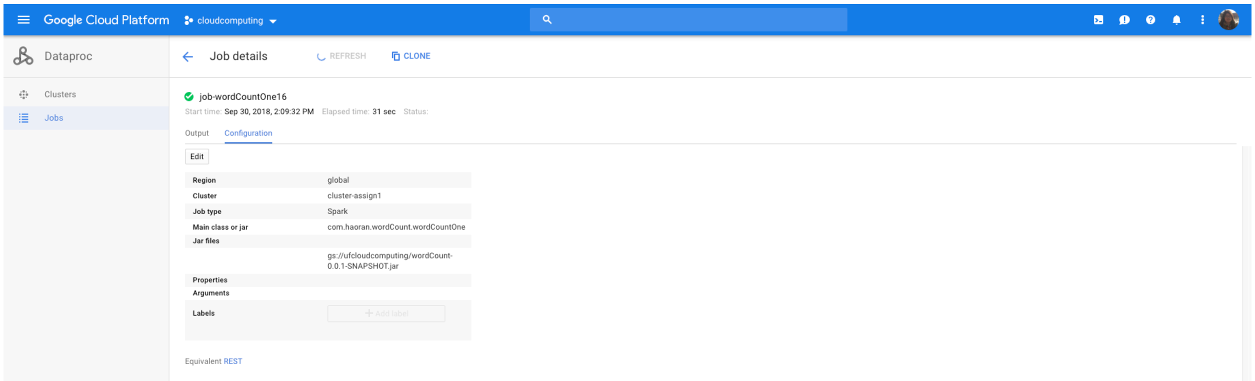
VM instances

	Name	Zone	Recommendation	Internal IP	External IP	Connect
☑	cluster-assign1-m	us-central1-a		10.128.0.4 (nic0)	35.238.38.120	SSH
☑	cluster-assign1-w-0	us-central1-a		10.128.0.2 (nic0)	35.193.188.217	SSH
☑	cluster-assign1-w-1	us-central1-a		10.128.0.3 (nic0)	35.239.103.150	SSH
□	instance-1	us-east1-b		10.142.0.2 (nic0)	None	SSH

2. Submit jobs to run

Task 1

Use web UI, see screenshot for configuration



job-wordCountOne16

Start time: Sep 30, 2018, 2:09:32 PM Elapsed time: 31 sec Status:

Region	global
Cluster	cluster-assign1
Job type	Spark
Main class or jar	com.haoran.wordCount.wordCountOne
Jar files	gs://ufcloudcomputing/wordCount-0.0.1-SNAPSHOT.jar
Properties	
Arguments	
Labels	

Task 2

Use command line tool to submit

```
gcloud dataproc jobs submit pyspark \
    --cluster cluster-assign1 --region global \
    gs://ufcloudcomputing/wordCountTwo.py
```

Task 3

Use command line tool to submit

```
gcloud dataproc jobs submit pyspark \
    --cluster cluster-assign1 --region global \
    gs://ufcloudcomputing/wordCountThree.py
```