
EEL 6761

Assign1

Word Count

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

Haoran Yu

UFID:18106994

Task1

1. Single word count

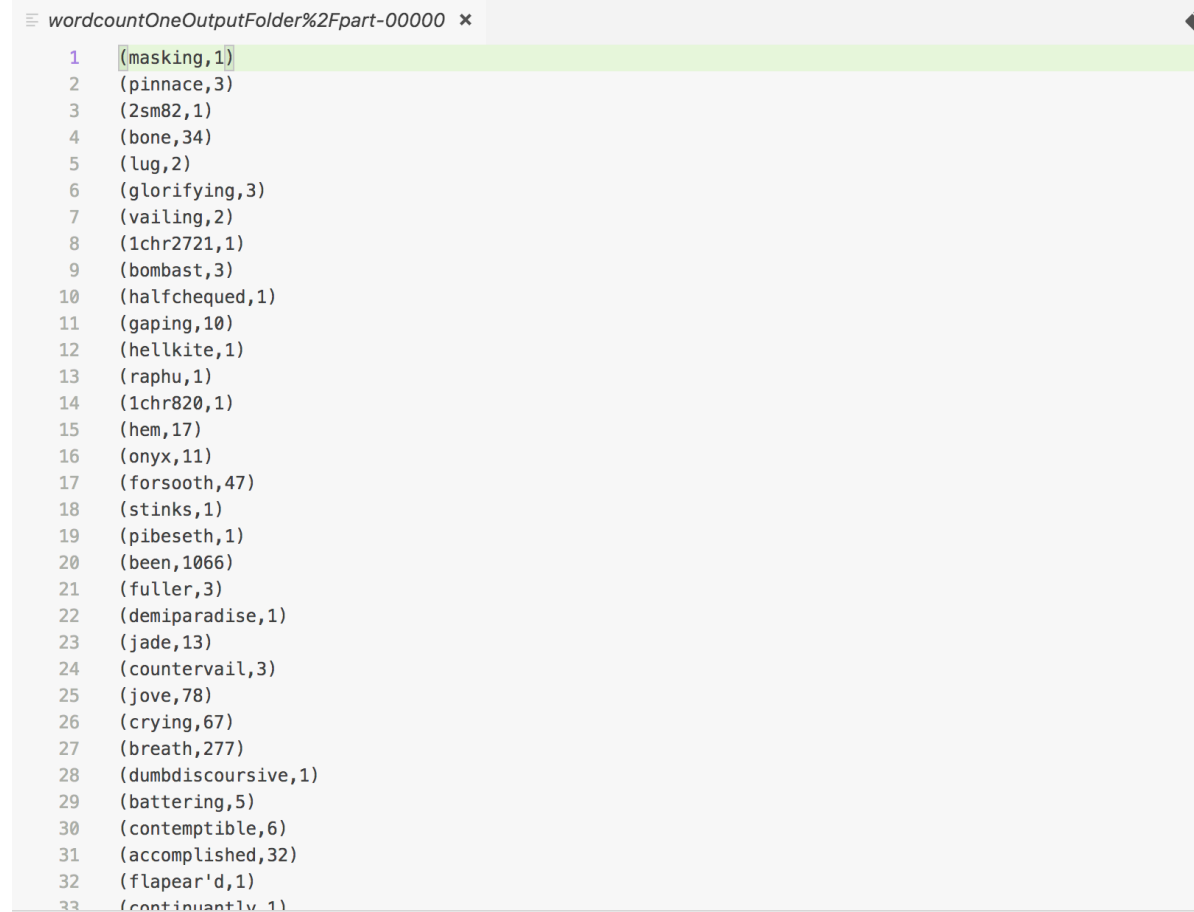
First, load bible data from google cloud storage bucket to RDD, split words with spaces, and then use filter to get non-space elements. Second, map elements to key-value pair like, <word, 1>. Third, reduce pair by the same key and get the result of counting words.

2. Result

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

<https://storage.googleapis.com/ufcloudcomputing/wordcountOneOutputFolder/part-00000>

<https://storage.googleapis.com/ufcloudcomputing/wordcountOneOutputFolder/part-00001>

A screenshot of a text editor window showing a list of words and their counts. The window title is 'wordcountOneOutputFolder%2Fpart-00000'. The text is as follows:

```
1 (masking,1)
2 (pinnacle,3)
3 (2sm82,1)
4 (bone,34)
5 (lug,2)
6 (glorifying,3)
7 (vailing,2)
8 (1chr2721,1)
9 (bombast,3)
10 (halfchequed,1)
11 (gaping,10)
12 (hellkite,1)
13 (raphu,1)
14 (1chr820,1)
15 (hem,17)
16 (onyx,11)
17 (forsooth,47)
18 (stinks,1)
19 (pibeseth,1)
20 (been,1066)
21 (fuller,3)
22 (demiparadise,1)
23 (jade,13)
24 (countervail,3)
25 (jove,78)
26 (crying,67)
27 (breath,277)
28 (dumbdiscursive,1)
29 (battering,5)
30 (contemptible,6)
31 (accomplished,32)
32 (flapear'd,1)
33 (continuantly,1)
```

3. Output console screenshot



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.

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

<https://storage.googleapis.com/ufcloudcomputing/wordcountTwoOutputFolder/part-00000>

```
wordcountTwoOutputFolder%2Fpart-00000 x
1  [('"('give', 'as')", 2)]
2  [('"('my', 'married')", 1)]
3  [('"('candle', 'burns')", 1)]
4  [('"('thy', 'commission')", 1)]
5  [('"('2sm1725', 'and')", 1)]
6  [('"('the', 'backs')", 1)]
7  [('"('bay', 'where')", 1)]
8  [('"('blood', 'than')", 6)]
9  [('"('was', 'fidele')", 1)]
10 [('"('and', 'becometh')", 3)]
11 [('"('need', 'the')", 2)]
12 [('"('grows', 'old')", 4)]
13 [('"('to', 'amplify')", 1)]
14 [('"('which', 'proceed')", 1)]
15 [('"('southwell', '&c')", 1)]
16 [('"('an', 'upper')", 3)]
17 [('"('to', 'servant')", 3)]
18 [('"('confer', 'of')", 2)]
19 [('"('with', 'lady')", 2)]
20 [('"('to', 'trouble')", 12)]
21 [('"('thou', 'eaten')", 1)]
22 [('"('gaunt', 'his')", 1)]
23 [('"('these', 'ill')", 3)]
24 [('"('earth', 'rejoiceth')", 1)]
25 [('"('ethiopia', 'and')", 5)]
26 [('"('chief', 'baker')", 4)]
27 [('"('they', 'know')", 44)]
28 [('"('with', 'arrogancy')", 1)]
29 [('"('hubert', 'do')", 2)]
30 [('"('do', 'divorce')", 1)]
31 [('"('broach', 'this')", 1)]
```

3. Console output screenshot

```
yuhaoran9968@cluster-assign1-m-8 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@1bca10[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: GFS 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.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 22:47:35 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@1bca10[HTTP/1.1,[http/1.1]](0.0.0.0:4040)
Job [d72e35395c9d423dbc587e26af559ca1] finished successfully.
driverControlFilesUri: gs://dataproc-23a43cfc-4681-427f-ae23-cf9d37bdfdf-us/google-cloud-dataproc-metainfo/3fcdf3c6-8d28-4296-96d0-ac1ea95cafla/jobs/d72e35395c9d423dbc587e26af559ca1/
driverOutputResourceUri: gs://dataproc-23a43cfc-4681-427f-ae23-cf9d37bdfdf-us/google-cloud-dataproc-metainfo/3fcdf3c6-8d28-4296-96d0-ac1ea95cafla/jobs/d72e35395c9d423dbc587e26af559ca1/driveroutput
placement:
  clusterName: cluster-assign1
  clusterUuid: 3fcdf3c6-8d28-4296-96d0-ac1ea95cafla
pysparkJob:
  mainPythonFileUri: gs://ufcloudcomputing/wordCountTwo.py
reference:
  jobId: d72e35395c9d423dbc587e26af559ca1
  projectId: bigdataass
status:
  state: DONE
  stateStartTime: '2018-09-30T22:47:42.754Z'
stateHistory:
  - 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
  trackingUri: http://cluster-assign1-m:8088/proxy/application_1538331699096_0031/
yuhaoran9968@cluster-assign1-m-8 []
```

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.

2. Result

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

<https://storage.googleapis.com/ufcloudcomputing/wordcountThreeOutputFolder/part-00000>

<https://storage.googleapis.com/ufcloudcomputing/wordcountThreeOutputFolder/part-00001>

<https://storage.googleapis.com/ufcloudcomputing/wordcountThreeOutputFolder/part-00002>

wordcountThreeOutputFolder%2Fpart-00000 x

```
1 ('therefore', 1888)
2 ('how', 2801)
3 ('came', 2452)
4 ('for', 16941)
5 ('intent', 64)
6 ('nation', 175)
7 ('one', 3872)
8 ('with', 14263)
9 ('am', 3089)
10 ('an', 3580)
11 ('at', 4207)
12 ('company', 293)
13 ('up', 3505)
14 ('ago', 43)
15 ('them', 8507)
16 ('not', 15440)
17 ('that', 24407)
18 ('ye', 4289)
19 ('keep', 867)
20 ('peter', 250)
21 ('he', 17087)
22 ('before', 2655)
23 ('were', 4432)
24 ('common', 177)
25 ('as', 9532)
26 ('my', 17312)
27 ('god', 5229)
```

3. Console output screenshot

```
ruhaosen9988@cluster-assign1-m-8: gocloud 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.-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/0.128.0.4:8032
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-cf9d37bfd1f-us/google-cloud-dataproc-metainfo/3fcdf3c6-8d28-4296-96d0-aclea95cfa1a/jobs/dc0fe83fb9df40958b6f54cf7d0e2f88/driveroutput
DriverOutputResourceUri: gs://dataproc-23a43cfc-4681-427f-ae23-cf9d37bfd1f-us/google-cloud-dataproc-metainfo/3fcdf3c6-8d28-4296-96d0-aclea95cfa1a/jobs/dc0fe83fb9df40958b6f54cf7d0e2f88/driveroutput
Placement:
  clusterName: cluster-assign1
  clusterUuid: 3fcdf3c6-8d28-4296-96d0-aclea95cfa1a
pysparkJob:
  mainPythonFileUri: gs://ufcloudcomputing/wordCountThree.py
reference:
  jobId: dc0fe83fb9df40958b6f54cf7d0e2f88
  projectId: bigdataass
status:
  state: DOWN
  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.

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: ✔

Configuration

- 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

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
```