“Pi calculation” in exercise 3

```
yutao@surface-laptop:~$ gcloud dataproc clusters create example-cluster --region=us-east1
Waiting on operation [projects/silken-water-362100/regions/us-east1/operations/7fdee4ff-2da1-3dc4-aa88-d75af94d32b6].
Waiting for cluster creation operation...working.
WARNING: No image specified. Using the default image version. It is recommended to select a specific image version in produ
ction, as the default image version may change at any time.
Waiting for cluster creation operation...done.
Created [https://dataproc.googleapis.com/v1/projects/silken-water-362100/regions/us-east1/clusters/example-cluster] Cluster
 placed in zone [us-east1-d].
```

```
yutao@surface-laptop:~$ gcloud dataproc jobs submit spark --cluster example-cluster     --region=us-east1     --class org.a
pache.spark.examples.SparkPi     --jars file:///usr/lib/spark/examples/jars/spark-examples.jar -- 1000
Job [e05aeb15389b4ecca25ea81e1100abc5] submitted.
Waiting for job output...
22/09/13 11:28:09 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/09/13 11:28:09 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/09/13 11:28:09 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
22/09/13 11:28:09 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
22/09/13 11:28:09 INFO org.sparkproject.jetty.util.log: Logging initialized @3675ms to org.sparkproject.jetty.util.log.Slf4
jLog
22/09/13 11:28:09 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git:
b881a572662e1943a14ae12e7e1207989f218b74; jvm 1.8.0_332-b09
22/09/13 11:28:09 INFO org.sparkproject.jetty.server.Server: Started @3808ms
22/09/13 11:28:09 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@333cb916{HTTP/1.1, (http/1.
1)}{0.0.0.0:38865}
22/09/13 11:28:10 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at example-cluster-m/10.142.0.1
1:8032
22/09/13 11:28:10 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at example-cluster-
m/10.142.0.11:10200
22/09/13 11:28:11 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
22/09/13 11:28:11 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/09/13 11:28:12 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_16630683900
49_0001
22/09/13 11:28:13 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at example-cluster-m/10.142.0.1
1:8030
22/09/13 11:28:16 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignorin
g exception of type GoogleJsonResponseException; verified object already exists with desired state.
Pi is roughly 3.141798231417982
22/09/13 11:28:33 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@333cb916{HTTP/1.1, (http/1.1)}{0.0.0.
0:0}
Job [e05aeb15389b4ecca25ea81e1100abc5] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-east1-882795771611-ccxr9twc/google-cloud-dataproc-metainfo/cb6de1c2-87ad-45
c6-8abe-97d441e7098d/jobs/e05aeb15389b4ecca25ea81e1100abc5/
driverOutputResourceUri: gs://dataproc-staging-us-east1-882795771611-ccxr9twc/google-cloud-dataproc-metainfo/cb6de1c2-87ad-
45c6-8abe-97d441e7098d/jobs/e05aeb15389b4ecca25ea81e1100abc5/driveroutput
jobUuid: ef7eea40-a337-37e3-b144-df6f5866e377
```

```
yutao@surface-laptop:~$ gcloud dataproc clusters update example-cluster     --region=us-east1     --num-workers 3
Waiting on operation [projects/silken-water-362100/regions/us-east1/operations/a45b4dd6-252a-3fc9-84f4-a57c66589be7].
Waiting for cluster update operation...done.
Updated [https://dataproc.googleapis.com/v1/projects/silken-water-362100/regions/us-east1/clusters/example-cluster].
```

```
yutao@surface-laptop:~$ gcloud dataproc clusters update example-cluster     --region=us-east1     --num-workers 2
Waiting on operation [projects/silken-water-362100/regions/us-east1/operations/b1042f57-6ad5-3c20-bd4e-62da406c8788].
Waiting for cluster update operation...done.
Updated [https://dataproc.googleapis.com/v1/projects/silken-water-362100/regions/us-east1/clusters/example-cluster].
```

```
yutao@surface-laptop:~$ gcloud dataproc clusters delete example-cluster     --region=us-east1
The cluster 'example-cluster' and all attached disks will be deleted.

Do you want to continue (Y/n)?  y

Waiting on operation [projects/silken-water-362100/regions/us-east1/operations/8fe43d7b-b566-3af0-ba13-451aa65edd33].
Waiting for cluster deletion operation...done.
Deleted [https://dataproc.googleapis.com/v1/projects/silken-water-362100/regions/us-east1/clusters/example-cluster].
```

```
def inside(p):
    x, y = random.random(), random.random()
    return x*x + y*y < 1

count = sc.parallelize(range(0, NUM_SAMPLES)) \
             .filter(inside).count()
print("Pi is roughly %f" % (4.0 * count / NUM_SAMPLES))
```

Transformations

- filter

Actions

- count

RDD operation that triggers the program to execute

- count

“word count” in exercise 4

```
yutao@surface-laptop:~$ gsutil cp gs://pub/shakespeare/rose.txt     gs://eecs6893_data/input/rose.txt
Copying gs://pub/shakespeare/rose.txt [Content-Type=text/plain]...
/ [1 files][   84.0 B/   84.0 B]
Operation completed over 1 objects/84.0 B.
```

```
yutao@surface-laptop:~$ PROJECT=silken-water-362100
yutao@surface-laptop:~$ BUCKET_NAME=eecs6893_data
yutao@surface-laptop:~$ CLUSTER=wordcount
yutao@surface-laptop:~$ REGION=us-east1
yutao@surface-laptop:~$ gcloud dataproc clusters create ${CLUSTER} \
    --project=${PROJECT} \
    --region=${REGION} \
    --single-node
Waiting on operation [projects/silken-water-362100/regions/us-east1/operations/d593d7fb-dc0e-3c62-86c5-890c15b08da7].
Waiting for cluster creation operation...
WARNING: No image specified. Using the default image version. It is recommended to select a specific image version in produ
ction, as the default image version may change at any time.
Waiting for cluster creation operation...done.
Created [https://dataproc.googleapis.com/v1/projects/silken-water-362100/regions/us-east1/clusters/wordcount] Cluster place
d in zone [us-east1-d].
```

```
yutao@surface-laptop:~/Downloads$ gcloud dataproc jobs submit pyspark word-count.py     --cluster=${CLUSTER}     --region=$
{REGION}     -- gs://${BUCKET_NAME}/input/ gs://${BUCKET_NAME}/output/
Job [816adfd70a404623bb4c7e6a3c5cc09f] submitted.
Waiting for job output...
22/09/13 13:29:56 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/09/13 13:29:56 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/09/13 13:29:56 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
22/09/13 13:29:56 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
22/09/13 13:29:56 INFO org.sparkproject.jetty.util.log: Logging initialized @3967ms to org.sparkproject.jetty.util.log.Slf4
jLog
22/09/13 13:29:56 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git:
b881a572662e1943a14ae12e7e1207989f218b74; jvm 1.8.0_332-b09
22/09/13 13:29:56 INFO org.sparkproject.jetty.server.Server: Started @4085ms
22/09/13 13:29:57 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@213e519b{HTTP/1.1, (http/1.
1)}{0.0.0.0:37135}
22/09/13 13:29:57 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at wordcount-m/10.142.0.13:8032
22/09/13 13:29:58 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at wordcount-m/10.1
42.0.13:10200
22/09/13 13:29:59 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
22/09/13 13:29:59 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/09/13 13:30:01 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_16630756084
23_0001
22/09/13 13:30:02 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at wordcount-m/10.142.0.13:8030
22/09/13 13:30:04 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignorin
g exception of type GoogleJsonResponseException; verified object already exists with desired state.
22/09/13 13:30:07 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1
22/09/13 13:30:24 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: S
uccessfully repaired 'gs://eecs6893_data/output/' directory.
22/09/13 13:30:24 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@213e519b{HTTP/1.1, (http/1.1)}{0.0.0.
0:0}
Job [816adfd70a404623bb4c7e6a3c5cc09f] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-east1-882795771611-ccxr9twc/google-cloud-dataproc-metainfo/8366895e-68ce-42
2c-964b-50605fe8a274/jobs/816adfd70a404623bb4c7e6a3c5cc09f/
driverOutputResourceUri: gs://dataproc-staging-us-east1-882795771611-ccxr9twc/google-cloud-dataproc-metainfo/8366895e-68ce-
422c-964b-50605fe8a274/jobs/816adfd70a404623bb4c7e6a3c5cc09f/driveroutput
jobUuid: e8065237-915d-3002-abef-172e4b00a586
placement:
```

```
placement:
  clusterName: wordcount
  clusterUuid: 8366895e-68ce-422c-964b-50605fe8a274
pysparkJob:
  args:
  - gs://eecs6893_data/input/
  - gs://eecs6893_data/output/
  mainPythonFileUri: gs://dataproc-staging-us-east1-882795771611-ccxr9twc/google-cloud-dataproc-metainfo/8366895e-68ce-422c
-964b-50605fe8a274/jobs/816adfd70a404623bb4c7e6a3c5cc09f/staging/word-count.py
reference:
  jobId: 816adfd70a404623bb4c7e6a3c5cc09f
  projectId: silken-water-362100
status:
  state: DONE
  stateStartTime: '2022-09-13T13:30:26.848345Z'
statusHistory:
- state: PENDING
  stateStartTime: '2022-09-13T13:29:50.942371Z'
- state: SETUP_DONE
  stateStartTime: '2022-09-13T13:29:50.999744Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2022-09-13T13:29:51.411510Z'
yarnApplications:
- name: word-count.py
  progress: 1.0
  state: FINISHED
  trackingUrl: http://wordcount-m:8088/proxy/application_1663075608423_0001/
```

```
yutao@surface-laptop:~/Downloads$ gsutil cat gs://${BUCKET_NAME}/output/*
("What's", 1)
('in', 1)
('name?', 1)
('That', 1)
('we', 1)
('call', 1)
('rose', 1)
('other', 1)
('name', 1)
('would', 1)
('smell', 1)
('as', 1)
('sweet.', 1)
('a', 2)
('which', 1)
('By', 1)
('any', 1)
```

```
yutao@surface-laptop:~/Downloads$ gcloud dataproc clusters delete wordcount    --region=us-east1
The cluster 'wordcount' and all attached disks will be deleted.

Do you want to continue (Y/n)?  y

Waiting on operation [projects/silken-water-362100/regions/us-east1/operations/9f16f2ba-8815-313b-a9d5-70cba3ee4cd5].
Waiting for cluster deletion operation...done.
Deleted [https://dataproc.googleapis.com/v1/projects/silken-water-362100/regions/us-east1/clusters/wordcount].
```

```
#!/usr/bin/env python

import pyspark
import sys

if len(sys.argv) != 3:
```

```
  raise Exception("Exactly 2 arguments are required: <inputUri> <outputUri>")

inputUri=sys.argv[1]
outputUri=sys.argv[2]

sc = pyspark.SparkContext()
lines = sc.textFile(sys.argv[1])
words = lines.flatMap(lambda line: line.split())
wordCounts = words.map(lambda word: (word, 1)).reduceByKey(lambda count1,
count2: count1 + count2)
wordCounts.saveAsTextFile(sys.argv[2])
```

Transformations

- flatMap
- map
- reduceByKey

Actions

- saveAsTextFile

RDD operation that triggers the program to execute

- reduceByKey

NYC Bike expert

```sql
SELECT COUNT(*)
FROM `silken-water-362100.citibike.bike_data`
WHERE -74.04 <= longitude AND longitude <= -73.94
```

Stations with longitude between -73.94 and -74.04: **698**

```sql
SELECT SUM(num_bikes_available)
FROM `silken-water-362100.citibike.bike_data`
WHERE region_id = 71
```

Total number of bikes available in region_id 71: **11885**

```sql
SELECT MAX(capacity)
FROM `silken-water-362100.citibike.bike_data`
```

Largest capacity for a station: **79**

```sql
SELECT station_id
FROM `silken-water-362100.citibike.bike_data`
WHERE capacity = (
SELECT MAX(capacity)
FROM `silken-water-362100.citibike.bike_data`
)
```

Station_id of the stations that have the largest capacity:

Row	station_id
1	445
2	422
3	501

Understanding William Shakespeare

TOP5

filteredshakes.py | shakes.py

```python
#!/usr/bin/env python

import pyspark
import sys

if len(sys.argv) != 3:
  raise Exception("Exactly 2 arguments are required: <inputUri> <outputUri>")

inputUri=sys.argv[1]
outputUri=sys.argv[2]

sc = pyspark.SparkContext()
lines = sc.textFile(sys.argv[1])
words = lines.flatMap(lambda line: line.split())
wordCounts = words.map(lambda word: (word, 1)).reduceByKey(lambda count1, count2: count1 + count2)
top5 = wordCounts.sortBy(lambda pair: -pair[1]).take(5)
print(top5)
saveRdd = sc.parallelize(top5)
saveRdd.saveAsTextFile(outputUri)
```

```
yutao@surface-laptop:~/Downloads$ gcloud dataproc jobs submit pyspark shakes.py --cluster=filter --region=us-east1 -- gs://
eecs6893_data/shakes.txt gs://eecs6893_data/top5
Job [c8d966c8bfca491cbade92543003fae5] submitted.
Waiting for job output...
22/09/22 13:48:13 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/09/22 13:48:13 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/09/22 13:48:13 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
22/09/22 13:48:13 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
22/09/22 13:48:13 INFO org.sparkproject.jetty.util.log: Logging initialized @3760ms to org.sparkproject.jetty.util.log.Slf4
jLog
22/09/22 13:48:13 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git:
b881a572662e1943a14ae12e7e1207989f218b74; jvm 1.8.0_332-b09
22/09/22 13:48:13 INFO org.sparkproject.jetty.server.Server: Started @3889ms
22/09/22 13:48:13 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@405b9505{HTTP/1.1, (http/1.
1)}{0.0.0.0:42957}
22/09/22 13:48:14 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at filter-m/10.142.0.22:8032
22/09/22 13:48:14 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at filter-m/10.142.
0.22:10200
22/09/22 13:48:16 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
22/09/22 13:48:16 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/09/22 13:48:16 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_16638500207
52_0034
22/09/22 13:48:17 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at filter-m/10.142.0.22:8030
22/09/22 13:48:20 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignorin
g exception of type GoogleJsonResponseException; verified object already exists with desired state.
22/09/22 13:48:23 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1
[('the', 620), ('and', 427), ('of', 396), ('to', 367), ('I', 326)]
22/09/22 13:48:39 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: S
uccessfully repaired 'gs://eecs6893_data/top5/' directory.
22/09/22 13:48:39 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@405b9505{HTTP/1.1, (http/1.1)}{0.0.0.
0:0}
Job [c8d966c8bfca491cbade92543003fae5] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-east1-882795771611-ccxr9twc/google-cloud-dataproc-metainfo/8ae2c167-ff6b-4f
14-9929-3be59de04af4/jobs/c8d966c8bfca491cbade92543003fae5/
driverOutputResourceUri: gs://dataproc-staging-us-east1-882795771611-ccxr9twc/google-cloud-dataproc-metainfo/8ae2c167-ff6b-
4f14-9929-3be59de04af4/jobs/c8d966c8bfca491cbade92543003fae5/driveroutput
```

```
jobUuid: 5408b09a-9508-3131-88d2-996158ddacbc
placement:
  clusterName: filter
  clusterUuid: 8ae2c167-ff6b-4f14-9929-3be59de04af4
pysparkJob:
  args:
  - gs://eecs6893_data/shakes.txt
  - gs://eecs6893_data/top5
  mainPythonFileUri: gs://dataproc-staging-us-east1-882795771611-ccxr9twc/google-cloud-dataproc-metainfo/8ae2c167-ff6b-4f14
-9929-3be59de04af4/jobs/c8d966c8bfca491cbade92543003fae5/staging/shakes.py
reference:
  jobId: c8d966c8bfca491cbade92543003fae5
  projectId: silken-water-362100
status:
  state: DONE
  stateStartTime: '2022-09-22T13:48:40.948729Z'
statusHistory:
- state: PENDING
  stateStartTime: '2022-09-22T13:48:08.540812Z'
- state: SETUP_DONE
  stateStartTime: '2022-09-22T13:48:08.584046Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2022-09-22T13:48:08.826212Z'
yarnApplications:
- name: shakes.py
  progress: 1.0
  state: FINISHED
  trackingUrl: http://filter-m:8088/proxy/application_1663850020752_0034/
```

TOP10 words

filteredshakes.py shakes.py

```python
#!/usr/bin/env python

import pyspark
import sys

if len(sys.argv) != 3:
  raise Exception("Exactly 2 arguments are required: <inputUri> <outputUri>")

inputUri=sys.argv[1]
outputUri=sys.argv[2]

sc = pyspark.SparkContext()
lines = sc.textFile(sys.argv[1])
words = lines.flatMap(lambda line: line.split())
wordCounts = words.map(lambda word: (word, 1)).reduceByKey(lambda count1, count2: count1 + count2)
top10 = wordCounts.sortBy(lambda pair: -pair[1]).take(10)
print(top10)
saveRdd = sc.parallelize(top10)
saveRdd.saveAsTextFile(outputUri)
```

```
yutao@surface-laptop:~/Downloads$ gcloud dataproc jobs submit pyspark shakes.py --cluster=filter --region=us-east1 -- gs://
eecs6893_data/shakes.txt gs://eecs6893_data/top10
Job [455c80c0523646e1afa91a435bbdece0] submitted.
Waiting for job output...
22/09/22 13:43:09 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/09/22 13:43:09 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/09/22 13:43:09 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
22/09/22 13:43:09 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
22/09/22 13:43:10 INFO org.sparkproject.jetty.util.log: Logging initialized @4203ms to org.sparkproject.jetty.util.log.Slf4
jLog
22/09/22 13:43:10 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git:
b881a572662e1943a14ae12e7e1207989f218b74; jvm 1.8.0_332-b09
22/09/22 13:43:10 INFO org.sparkproject.jetty.server.Server: Started @4330ms
22/09/22 13:43:10 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@17efc31{HTTP/1.1, (http/1.1
)}{0.0.0.0:35103}
22/09/22 13:43:10 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at filter-m/10.142.0.22:8032
22/09/22 13:43:11 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at filter-m/10.142.
0.22:10200
22/09/22 13:43:12 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
22/09/22 13:43:12 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/09/22 13:43:13 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_16638500207
52_0033
22/09/22 13:43:14 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at filter-m/10.142.0.22:8030
22/09/22 13:43:16 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignorin
g exception of type GoogleJsonResponseException; verified object already exists with desired state.
22/09/22 13:43:18 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1
[('the', 620), ('and', 427), ('of', 396), ('to', 367), ('I', 326), ('a', 256), ('you', 193), ('in', 190), ('is', 185), ('my
', 170)]
22/09/22 13:43:35 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: S
uccessfully repaired 'gs://eecs6893_data/top10/' directory.
22/09/22 13:43:35 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@17efc31{HTTP/1.1, (http/1.1)}{0.0.0.0
:0}
Job [455c80c0523646e1afa91a435bbdece0] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-east1-882795771611-ccxr9twc/google-cloud-dataproc-metainfo/8ae2c167-ff6b-4f
14-9929-3be59de04af4/jobs/455c80c0523646e1afa91a435bbdece0/
driverOutputResourceUri: gs://dataproc-staging-us-east1-882795771611-ccxr9twc/google-cloud-dataproc-metainfo/8ae2c167-ff6b-
```

```
4f14-9929-3be59de04af4/jobs/455c80c0523646e1afa91a435bbdece0/driveroutput
jobUuid: 6bc9d190-00e2-3627-bb38-74e167069e7f
placement:
  clusterName: filter
  clusterUuid: 8ae2c167-ff6b-4f14-9929-3be59de04af4
pysparkJob:
  args:
  - gs://eecs6893_data/shakes.txt
  - gs://eecs6893_data/top10
  mainPythonFileUri: gs://dataproc-staging-us-east1-882795771611-ccxr9twc/google-cloud-dataproc-metainfo/8ae2c167-ff6b-4f14
-9929-3be59de04af4/jobs/455c80c0523646e1afa91a435bbdece0/staging/shakes.py
reference:
  jobId: 455c80c0523646e1afa91a435bbdece0
  projectId: silken-water-362100
status:
  state: DONE
  stateStartTime: '2022-09-22T13:43:40.108673Z'
statusHistory:
- state: PENDING
  stateStartTime: '2022-09-22T13:43:04.926461Z'
- state: SETUP_DONE
  stateStartTime: '2022-09-22T13:43:05.020594Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2022-09-22T13:43:05.246346Z'
yarnApplications:
- name: shakes.py
  progress: 1.0
  state: FINISHED
  trackingUrl: http://filter-m:8088/proxy/application_1663850020752_0033/
```

Top10 filtered words

filteredshakes.py shakes.py

```python
#!/usr/bin/env python

import pyspark
import sys
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
import os

if len(sys.argv) != 3:
  raise Exception("Exactly 2 arguments are required: <inputUri> <outputUri>")

inputUri=sys.argv[1]
outputUri=sys.argv[2]

sc = pyspark.SparkContext()
lines = sc.textFile(sys.argv[1])
words = lines.flatMap(lambda line: line.split(" ")).filter(lambda x: x != "")
user_paths = os.environ['PYTHONPATH'].split(os.pathsep)
stop = [w.title() for w in stopwords.words('english')]
stop += stopwords.words('english')
stopWords = sc.parallelize(stop)
filteredWords = words.subtract(stopWords)
wordCounts = filteredWords.map(lambda word: (word, 1)).reduceByKey(lambda count1, count2: count1 + count2)
frequencyMap = wordCounts.sortBy(lambda pair: -pair[1]).take(10)
print(frequencyMap)
saveRdd = sc.parallelize(frequencyMap)
saveRdd.saveAsTextFile(outputUri)
```

```
yutao@surface-laptop:~/Downloads$ gcloud dataproc jobs submit pyspark filteredshakes.py --cluster=filter --region=us-east1
-- gs://eecs6893_data/shakes.txt gs://eecs6893_data/filteredTop10
Job [178450cc956d47a5b4f37a07ee076d83] submitted.
Waiting for job output...
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
22/09/22 13:34:54 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/09/22 13:34:54 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/09/22 13:34:54 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
22/09/22 13:34:54 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
22/09/22 13:34:54 INFO org.sparkproject.jetty.util.log: Logging initialized @4795ms to org.sparkproject.jetty.util.log.Slf4
jLog
22/09/22 13:34:54 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git:
b881a572662e1943a14ae12e7e1207989f218b74; jvm 1.8.0_332-b09
22/09/22 13:34:54 INFO org.sparkproject.jetty.server.Server: Started @4931ms
22/09/22 13:34:54 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@7d1a6566{HTTP/1.1, (http/1.
1)}{0.0.0.0:38525}
22/09/22 13:34:55 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at filter-m/10.142.0.22:8032
22/09/22 13:34:55 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at filter-m/10.142.
0.22:10200
22/09/22 13:34:56 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
22/09/22 13:34:56 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/09/22 13:34:58 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_16638500207
52_0031
22/09/22 13:34:59 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at filter-m/10.142.0.22:8030
22/09/22 13:35:01 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignorin
g exception of type GoogleJsonResponseException; verified object already exists with desired state.
22/09/22 13:35:03 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1
[('Macb.', 137), ('haue', 114), ('Enter', 73), ('thou', 61), ('Macd.', 58), ('shall', 47), ('vpon', 47), ('thy', 46), ('yet
', 45), ('thee', 43)]
22/09/22 13:35:20 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: S
uccessfully repaired 'gs://eecs6893_data/filteredTop10/' directory.
22/09/22 13:35:20 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@7d1a6566{HTTP/1.1, (http/1.1)}{0.0.0.
0:0}
Job [178450cc956d47a5b4f37a07ee076d83] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-east1-882795771611-ccxr9twc/google-cloud-dataproc-metainfo/8ae2c167-ff6b-4f
```

```
14-9929-3be59de04af4/jobs/178450cc956d47a5b4f37a07ee076d83/
driverOutputResourceUri: gs://dataproc-staging-us-east1-882795771611-ccxr9twc/google-cloud-dataproc-metainfo/8ae2c167-ff6b-
4f14-9929-3be59de04af4/jobs/178450cc956d47a5b4f37a07ee076d83/driveroutput
jobUuid: 2a47d96b-9881-318f-a91f-324b4fe98cbd
placement:
  clusterName: filter
  clusterUuid: 8ae2c167-ff6b-4f14-9929-3be59de04af4
pysparkJob:
  args:
  - gs://eecs6893_data/shakes.txt
  - gs://eecs6893_data/filteredTop10
  mainPythonFileUri: gs://dataproc-staging-us-east1-882795771611-ccxr9twc/google-cloud-dataproc-metainfo/8ae2c167-ff6b-4f14
-9929-3be59de04af4/jobs/178450cc956d47a5b4f37a07ee076d83/staging/filteredshakes.py
reference:
  jobId: 178450cc956d47a5b4f37a07ee076d83
  projectId: silken-water-362100
status:
  state: DONE
  stateStartTime: '2022-09-22T13:35:24.815641Z'
statusHistory:
- state: PENDING
  stateStartTime: '2022-09-22T13:34:48.432592Z'
- state: SETUP_DONE
  stateStartTime: '2022-09-22T13:34:48.477903Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2022-09-22T13:34:48.744426Z'
yarnApplications:
- name: filteredshakes.py
  progress: 1.0
  state: FINISHED
  trackingUrl: http://filter-m:8088/proxy/application_1663850020752_0031/
```