HW7 - David Euijoon Kim - Apache Beam

https://github.com/dk-davidekim/Google-Cloud-Computing.git

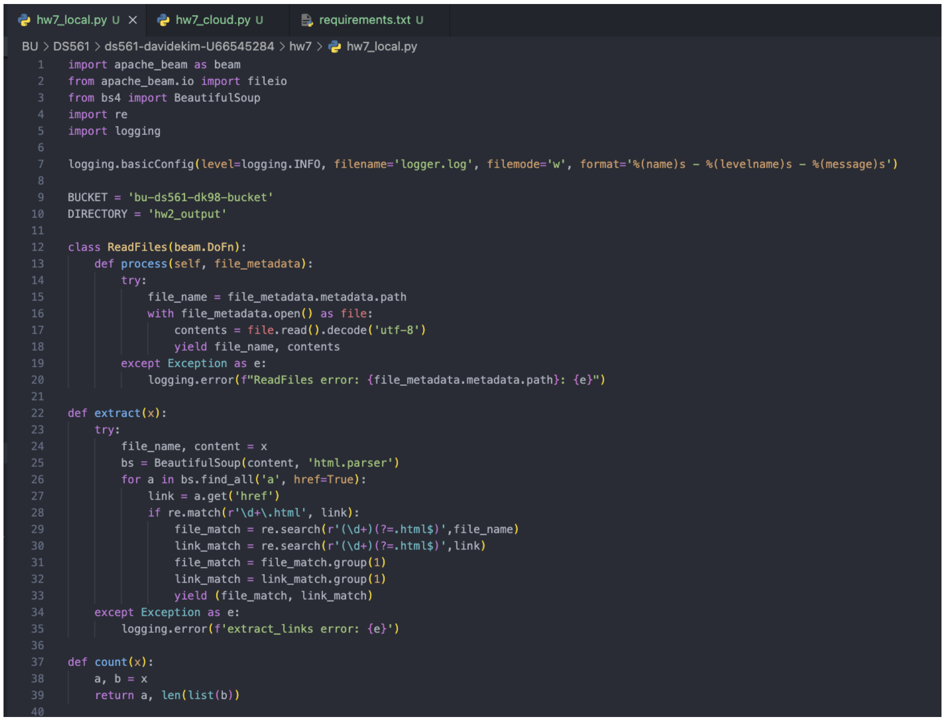
1. Install Dependencies

```
pip install ‘apache-beam[gcp]’
pip install beautifulsoup4
```

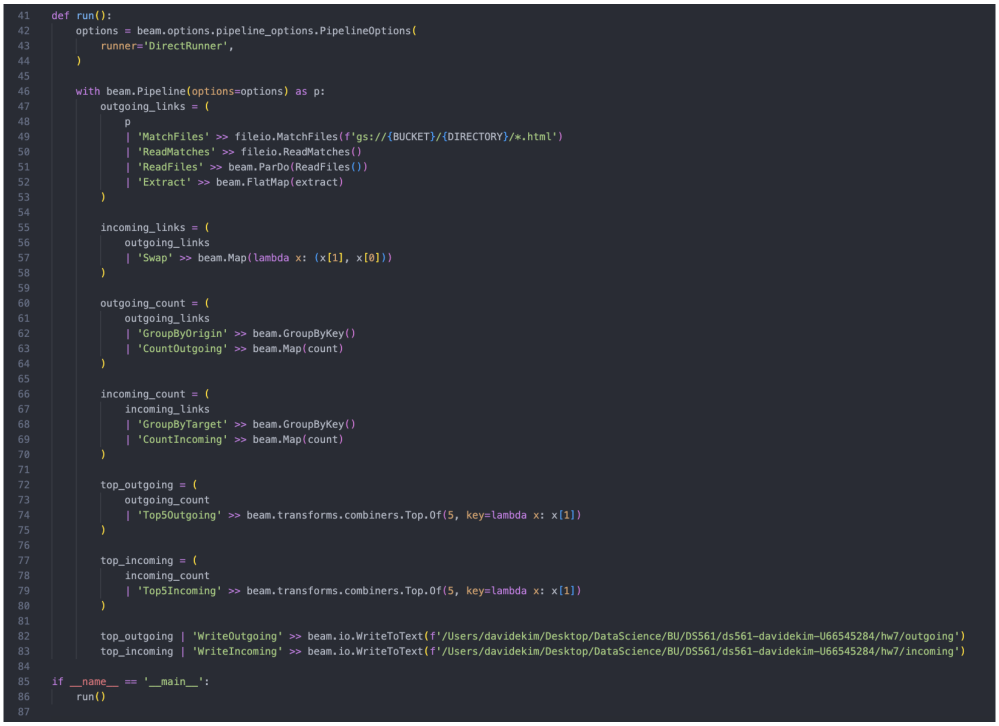
2. Enable API and Link Service Account

```
gcloud config set project ds-561

for i in dataflow compute_component logging storage_component storage_api bigquery pubsub datastore.googleapis.com cloudresourcemanager.googleapis.com; do gcloud services enable $i; done

for i in roles/dataflow.admin roles/dataflow.worker roles/storage.objectAdmin; do
gcloud projects add-iam-policy-binding ds-561
--member="serviceAccount:634775913953-compute@developer.gserviceaccount.com"
--role=$i;
done
```

3. Write Code - hw7_local.py

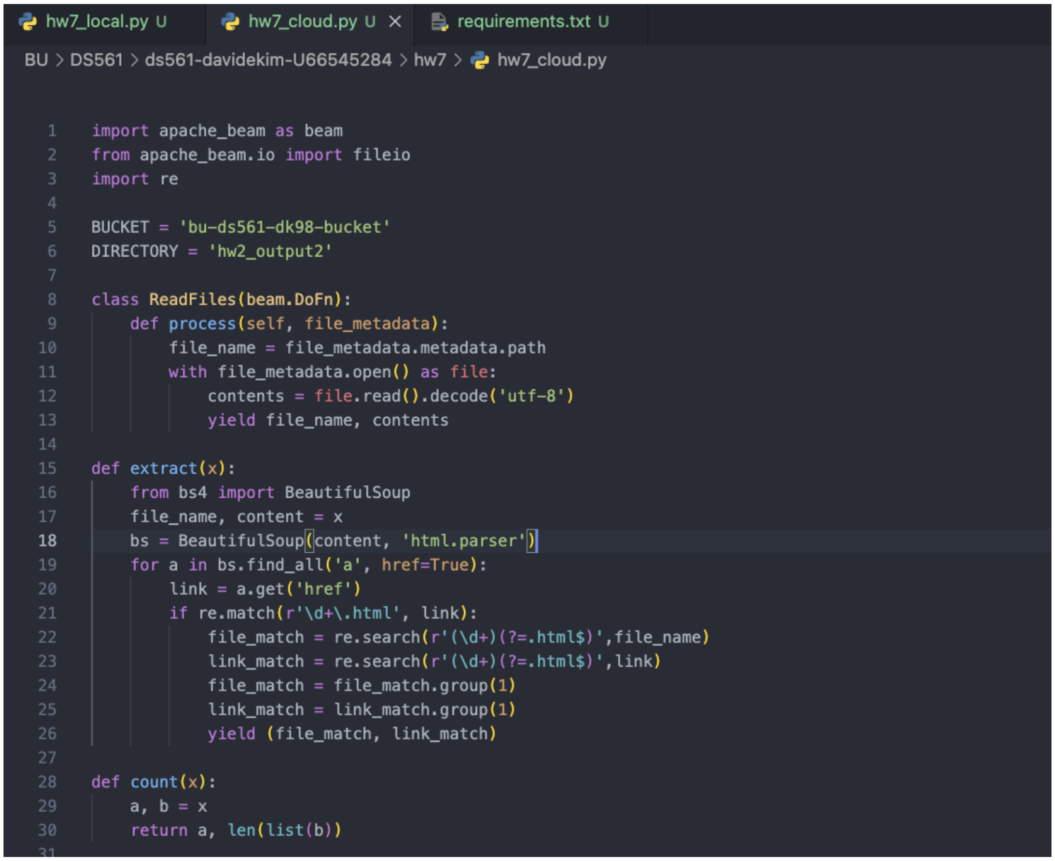


```python
import apache_beam as beam
from apache_beam.io import fileio
from bs4 import BeautifulSoup
import re
import logging

logging.basicConfig(level=logging.INFO, filename='logger.log', filemode='w', format='%(name)s - %(levelname)s - %(message)s')

BUCKET = 'bu-ds561-dk98-bucket'
DIRECTORY = 'hw2_output'

class ReadFiles(beam.DoFn):
    def process(self, file_metadata):
        try:
            file_name = file_metadata.metadata.path
            with file_metadata.open() as file:
                contents = file.read().decode('utf-8')
                yield file_name, contents
        except Exception as e:
            logging.error(f"ReadFiles error: {file_metadata.metadata.path}: {e}")

def extract(x):
    try:
        file_name, content = x
        bs = BeautifulSoup(content, 'html.parser')
        for a in bs.find_all('a', href=True):
            link = a.get('href')
            if re.match(r'\d+\.html', link):
                file_match = re.search(r'(\d+)(?=.html$)',file_name)
                link_match = re.search(r'(\d+)(?=.html$)',link)
                file_match = file_match.group(1)
                link_match = link_match.group(1)
                yield (file_match, link_match)
    except Exception as e:
        logging.error(f'extract_links error: {e}')

def count(x):
    a, b = x
    return a, len(list(b))

```

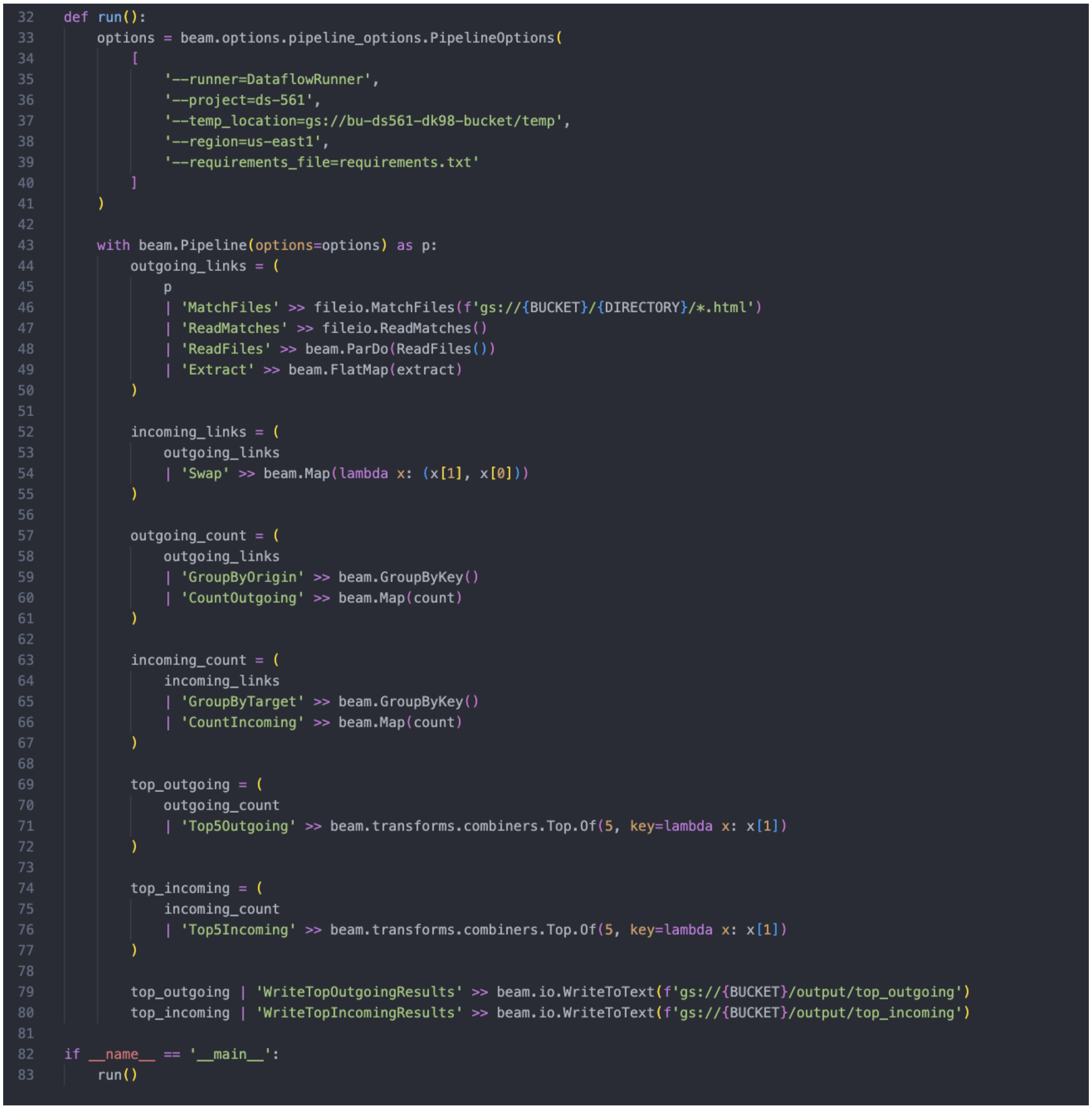


```python
def run():
    options = beam.options.pipeline_options.PipelineOptions(
        runner='DirectRunner',
    )

    with beam.Pipeline(options=options) as p:
        outgoing_links = (
            p
            | 'MatchFiles' >> fileio.MatchFiles(f'gs://{BUCKET}/{DIRECTORY}/*.html')
            | 'ReadMatches' >> fileio.ReadMatches()
            | 'ReadFiles' >> beam.ParDo(ReadFiles())
            | 'Extract' >> beam.FlatMap(extract)
        )

        incoming_links = (
            outgoing_links
            | 'Swap' >> beam.Map(lambda x: (x[1], x[0]))
        )

        outgoing_count = (
            outgoing_links
            | 'GroupByOrigin' >> beam.GroupByKey()
            | 'CountOutgoing' >> beam.Map(count)
        )

        incoming_count = (
            incoming_links
            | 'GroupByTarget' >> beam.GroupByKey()
            | 'CountIncoming' >> beam.Map(count)
        )

        top_outgoing = (
            outgoing_count
            | 'Top5Outgoing' >> beam.transforms.combiners.Top.Of(5, key=lambda x: x[1])
        )

        top_incoming = (
            incoming_count
            | 'Top5Incoming' >> beam.transforms.combiners.Top.Of(5, key=lambda x: x[1])
        )

        top_outgoing | 'WriteOutgoing' >> beam.io.WriteToText(f'/Users/davidekim/Desktop/DataScience/BU/DS561/ds561-davidekim-U66545284/hw7/outgoing')
        top_incoming | 'WriteIncoming' >> beam.io.WriteToText(f'/Users/davidekim/Desktop/DataScience/BU/DS561/ds561-davidekim-U66545284/hw7/incoming')

if __name__ == '__main__':
    run()
```

4. Write Code - hw7_cloud.py

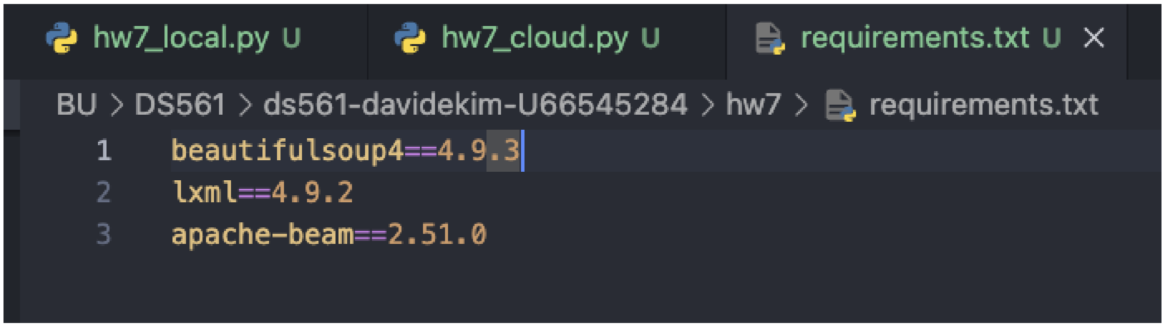


```python
import apache_beam as beam
from apache_beam.io import fileio
import re

BUCKET = 'bu-ds561-dk98-bucket'
DIRECTORY = 'hw2_output2'

class ReadFiles(beam.DoFn):
    def process(self, file_metadata):
        file_name = file_metadata.metadata.path
        with file_metadata.open() as file:
            contents = file.read().decode('utf-8')
            yield file_name, contents

def extract(x):
    from bs4 import BeautifulSoup
    file_name, content = x
    bs = BeautifulSoup(content, 'html.parser')
    for a in bs.find_all('a', href=True):
        link = a.get('href')
        if re.match(r'\d+\.html', link):
            file_match = re.search(r'(\d+)(?=.html$)',file_name)
            link_match = re.search(r'(\d+)(?=.html$)',link)
            file_match = file_match.group(1)
            link_match = link_match.group(1)
            yield (file_match, link_match)

def count(x):
    a, b = x
    return a, len(list(b))
```



```python
def run():
    options = beam.options.pipeline_options.PipelineOptions(
        [
            '--runner=DataflowRunner',
            '--project=ds-561',
            '--temp_location=gs://bu-ds561-dk98-bucket/temp',
            '--region=us-east1',
            '--requirements_file=requirements.txt'
        ]
    )

    with beam.Pipeline(options=options) as p:
        outgoing_links = (
            p
            | 'MatchFiles' >> fileio.MatchFiles(f'gs://{BUCKET}/{DIRECTORY}/*.html')
            | 'ReadMatches' >> fileio.ReadMatches()
            | 'ReadFiles' >> beam.ParDo(ReadFiles())
            | 'Extract' >> beam.FlatMap(extract)
        )

        incoming_links = (
            outgoing_links
            | 'Swap' >> beam.Map(lambda x: (x[1], x[0]))
        )

        outgoing_count = (
            outgoing_links
            | 'GroupByOrigin' >> beam.GroupByKey()
            | 'CountOutgoing' >> beam.Map(count)
        )

        incoming_count = (
            incoming_links
            | 'GroupByTarget' >> beam.GroupByKey()
            | 'CountIncoming' >> beam.Map(count)
        )

        top_outgoing = (
            outgoing_count
            | 'Top5Outgoing' >> beam.transforms.combiners.Top.Of(5, key=lambda x: x[1])
        )

        top_incoming = (
            incoming_count
            | 'Top5Incoming' >> beam.transforms.combiners.Top.Of(5, key=lambda x: x[1])
        )

        top_outgoing | 'WriteTopOutgoingResults' >> beam.io.WriteToText(f'gs://{BUCKET}/output/top_outgoing')
        top_incoming | 'WriteTopIncomingResults' >> beam.io.WriteToText(f'gs://{BUCKET}/output/top_incoming')

if __name__ == '__main__':
    run()
```

5. Write Code - requirements.txt



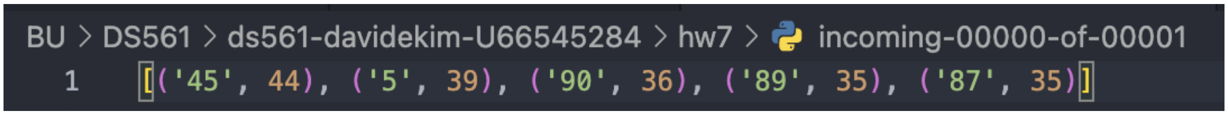
hw7_local.py U
hw7_cloud.py U
requirements.txt U ×
BU > DS561 > ds561-davidekim-U66545284 > hw7 > requirements.txt

```
beautifulsoup4==4.9.3
lxml==4.9.2
apache-beam==2.51.0
```

6. Debugging with Smaller Dataset

Created a new directory in my bucket with only 100 files ranging from 1.html to 100.html

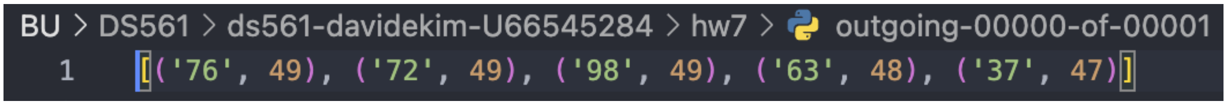
(Local - DirectRunner)
Incoming



BU > DS561 > ds561-davidekim-U66545284 > hw7 > incoming-00000-of-00001

```
[('45', 44), ('5', 39), ('90', 36), ('89', 35), ('87', 35)]
```

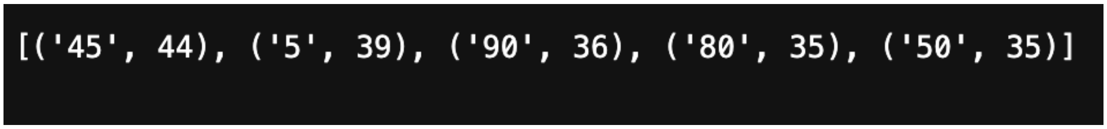
Outgoing



BU > DS561 > ds561-davidekim-U66545284 > hw7 > outgoing-00000-of-00001

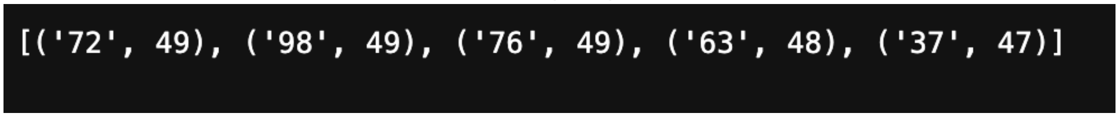
```
[('76', 49), ('72', 49), ('98', 49), ('63', 48), ('37', 47)]
```

(Cloud - DataflowRunner)
Incoming



```
[('45', 44), ('5', 39), ('90', 36), ('80', 35), ('50', 35)]
```

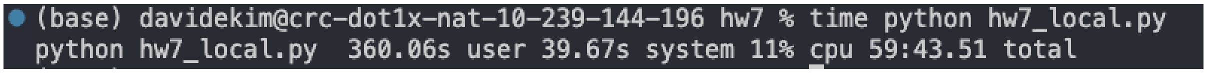
Outgoing



```
[('72', 49), ('98', 49), ('76', 49), ('63', 48), ('37', 47)]
```

7. Local - DirectRunner

time python hw7_local.py

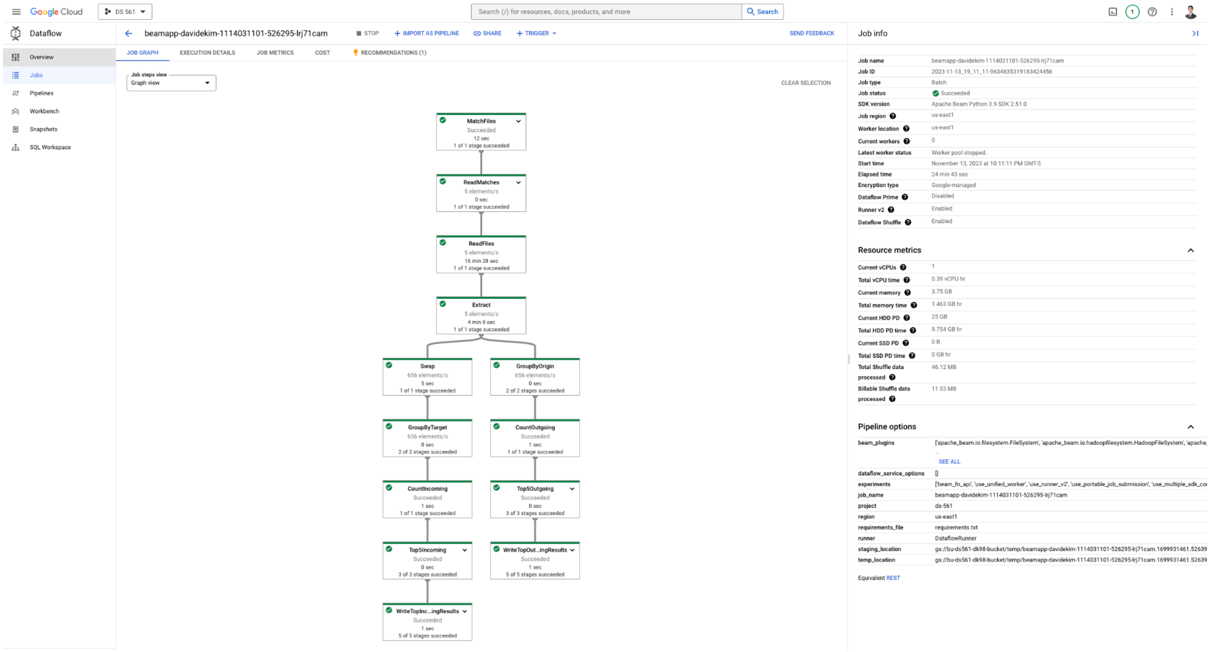


```
● (base) davidekim@crc-dot1x-nat-10-239-144-196 hw7 % time python hw7_local.py
python hw7_local.py  360.06s user 39.67s system 11% cpu 59:43.51 total
```

Runtime = 59:43.51

8. Cloud - DataflowRunner

python hw7_cloud.py



beamapp-davidekim-1114031101-526295-lrj71cam

Job info

Job name	beamapp-davidekim-1114031101-526295-lrj71cam
Job ID	2023-11-13_19_11_11-9634835319183424456
Job type	Batch
Job status	Succeeded
SDK version	Apache Beam Python 3.9 SDK 2.51.0
Job region	us-east1
Worker location	us-east1
Current workers	0
Latest worker status	Worker pool stopped.
Start time	November 13, 2023 at 10:11:11 PM GMT-5
Elapsed time	24 min 43 sec
Encryption type	Google-managed
Dataflow Prime	Disabled
Runner v2	Enabled
Dataflow Shuffle	Enabled

Resource metrics

Current vCPUs	1
Total vCPU time	0.39 vCPU hr
Current memory	3.75 GB
Total memory time	1.463 GB hr
Current HDD PD	25 GB
Total HDD PD time	9.754 GB hr
Current SSD PD	0 B
Total SSD PD time	0 GB hr
Total Shuffle data processed	46.12 MB
Billable Shuffle data processed	11.53 MB

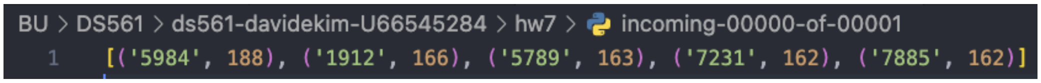
Pipeline options

beam_plugins	['apache_beam.io.filesystem.FileSystem', 'apache_beam.io.hadoopfilesystem.HadoopFileSystem', 'apache_
dataflow_service_options	[]
experiments	['beam_fn_api', 'use_unified_worker', 'use_runner_v2', 'use_portable_job_submission', 'use_multiple_sdk_co
job_name	beamapp-davidekim-1114031101-526295-lrj71cam
project	ds-561
region	us-east1
requirements_file	requirements.txt
runner	DataflowRunner
staging_location	gs://bu-ds561-dk98-bucket/temp/beamapp-davidekim-1114031101-526295-lrj71cam.1699931461.52639
temp_location	gs://bu-ds561-dk98-bucket/temp/beamapp-davidekim-1114031101-526295-lrj71cam.1699931461.52639

Runtime = 24 min 43 sec

9. Output Result - Local - DirectRunner

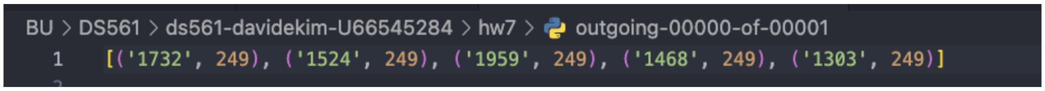
Incoming



BU > DS561 > ds561-davidekim-U66545284 > hw7 > incoming-00000-of-00001

```
[('5984', 188), ('1912', 166), ('5789', 163), ('7231', 162), ('7885', 162)]
```

Outgoing

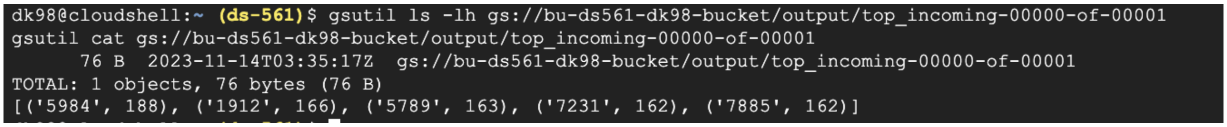


BU > DS561 > ds561-davidekim-U66545284 > hw7 > outgoing-00000-of-00001

```
[('1732', 249), ('1524', 249), ('1959', 249), ('1468', 249), ('1303', 249)]
```

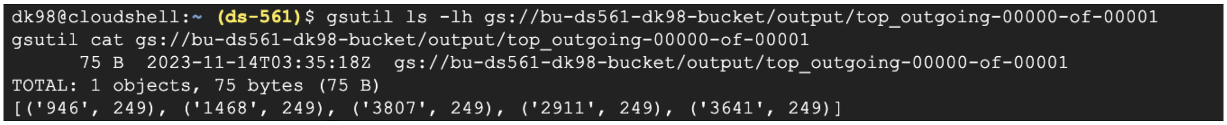
10. Output Result - Cloud - DataflowRunner

Incoming
gsutil ls -lh gs://bu-ds561-dk98-bucket/output/top_incoming-00000-of-00001
gsutil cat gs://bu-ds561-dk98-bucket/output/top_incoming-00000-of-00001



```
dk98@cloudshell:~ (ds-561)$ gsutil ls -lh gs://bu-ds561-dk98-bucket/output/top_incoming-00000-of-00001
gsutil cat gs://bu-ds561-dk98-bucket/output/top_incoming-00000-of-00001
         76 B  2023-11-14T03:35:17Z  gs://bu-ds561-dk98-bucket/output/top_incoming-00000-of-00001
TOTAL: 1 objects, 76 bytes (76 B)
[('5984', 188), ('1912', 166), ('5789', 163), ('7231', 162), ('7885', 162)]
```

Outgoing
gsutil ls -lh gs://bu-ds561-dk98-bucket/output/top_outgoing-00000-of-00001
gsutil cat gs://bu-ds561-dk98-bucket/output/top_outgoing-00000-of-00001



```
dk98@cloudshell:~ (ds-561)$ gsutil ls -lh gs://bu-ds561-dk98-bucket/output/top_outgoing-00000-of-00001
gsutil cat gs://bu-ds561-dk98-bucket/output/top_outgoing-00000-of-00001
      75 B  2023-11-14T03:35:18Z  gs://bu-ds561-dk98-bucket/output/top_outgoing-00000-of-00001
TOTAL: 1 objects, 75 bytes (75 B)
[('946', 249), ('1468', 249), ('3807', 249), ('2911', 249), ('3641', 249)]
```

END