

Dataflow: Qwik Start – Python

Table of Contents

- 1. Introduction**
- 2. Task 1: Creating a Cloud Storage Bucket**
- 3. Task 2: Installing Apache Beam SDK**
- 4. Task 3: Running a Dataflow Pipeline**
- 5. Task 4: Verifying Pipeline Success**
- 6. Key Learnings**
- 7. Conclusion**

1. Introduction

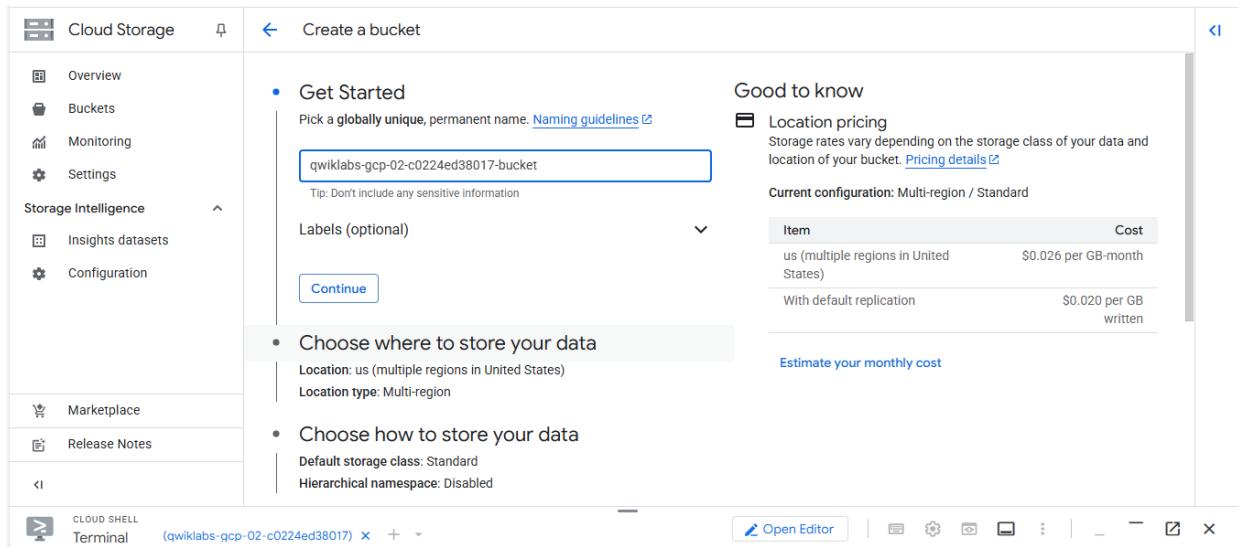
This project demonstrates how to:

- Set up a **Google Cloud Storage (GCS) bucket** for Dataflow outputs.
- Install and test the **Apache Beam SDK (Python)** in a Docker container.
- Run a **WordCount pipeline** locally and remotely via Dataflow.
- Monitor pipeline execution and verify results.

2. Task 1: Creating a Cloud Storage Bucket

Steps Executed:

1. Navigated to **Cloud Storage > Buckets**.
2. Created a bucket with:
 - a. **Name:** [UNIQUE_NAME]-bucket (e.g., myproject-bucket)
 - b. **Location Type:** Multi-region (us)
3. Disabled public access (default).



3. Task 2: Installing Apache Beam SDK

Steps Executed:

- Launched a **Python 3.9 Docker container**:

```
docker run -it -e DEVSHELL_PROJECT_ID=$DEVSHELL_PROJECT_ID python:3.9 /bin/bash
```

- Installed Apache Beam with GCP dependencies:

```
pip install 'apache-beam[gcp]==2.42.0'
```

- Tested the WordCount example **locally**:

```
python -m apache_beam.examples.wordcount --output counts.txt
cat counts.txt
```

```
student_01 9d65d778187b@cloudshell:~ (qwiklabs-gcp-02-c0224ed38017)$ docker run -it -e DEVSHELL_PROJECT_ID=$DEVSHELL_PROJECT_ID python:3.9 /bin/bash
Unable to find image 'python:3.9' locally
3.9: Pulling from library/python
3e6b9d1a9511: Downloading [=====]
37927ed901b1: Download complete
79b2f47ad444: Downloading [=====]
] 23.64MB/48.49MB
e23f099911d6: Waiting
3936689bd0c0: Waiting
d189aca32b40: Waiting
31a5ef1ce157: Waiting
[
```

Created container [id]: 9d65d778187b@cloudshell:~ (qwiklabs-gcp-02-c0224ed38017)

```
student_01 9d65d778187b@cloudshell:~ (qwiklabs-gcp-02-c0224ed38017)$ docker run -it -e DEVSHELL_PROJECT_ID=$DEVSHELL_PROJECT_ID python:3.9 /bin/bash
Unable to find image 'python:3.9' locally
3.9: Pulling from library/python
3e6b9d1a9511: Pull complete
37927ed901b1: Pull complete
79b2f47ad444: Pull complete
e23f099911d6: Pull complete
3936689bd0c0: Pull complete
d189aca32b40: Pull complete
31a5ef1ce157: Pull complete
Digest: sha256:089307feed9d056945980075a2a857c28141354a0158c5a251ebf63291859aa6
Status: Downloaded newer image for python:3.9
root@1e0cbd8fec0f:~ [ ]
```

```
root@1e0cbd8fec0f:~# pip install 'apache-beam[gcp]==2.42.0'
Collecting apache-beam[gcp]==2.42.0
  Downloading apache_beam-2.42.0-cp39-cp39-manylinux2010_x86_64.whl (12.1 MB)
     ━━━━━━━━━━━━━━━━ 12.1/12.1 MB 68.3 MB/s eta 0:00:00
Collecting orjson<4.0
  Downloading orjson-3.10.18-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (132 kB)
     ━━━━━━━━━━━━━━ 132.6/132.6 kB 22.2 MB/s eta 0:00:00
Collecting dill<0.3.2,>=0.3.1.1
  Downloading dill-0.3.1.1.tar.gz (151 kB)
     ━━━━━━━━━━━━ 152.0/152.0 kB 24.3 MB/s eta 0:00:00
Preparing metadata (setup.py) ... done
```

```
WARNING:google_auth_httplib2:httplib2 transport does not support per-request timeout. Set the timeout when constructing the httplib2.Http instance.
WARNING:google_auth_httplib2:httplib2 transport does not support per-request timeout. Set the timeout when constructing the httplib2.Http instance.
INFO:root:Default Python SDK image for environment is apache/beam/python3.9_sdk:2.42.0
INFO:apache beam runners.portability.fn_api.runner.translations:<function annotate_downstream_side_inputs at 0x77fd9aff2790>
INFO:apache beam runners.portability.fn_api.runner.translations:<function fix_side_input_pcoll_coders at 0x77fd9aff28b0>
INFO:apache beam runners.portability.fn_api.runner.translations:<function pack_combiners at 0x77fd9aff2dc0>
INFO:apache beam runners.portability.fn_api.runner.translations:<function lift_combiners at 0x77fd9aff2e50>
INFO:apache beam runners.portability.fn_api.runner.translations:<function expand_sdf at 0x77fd9aff3040>
INFO:apache beam runners.portability.fn_api.runner.translations:<function expand_gbk at 0x77fd9aff30d0>
INFO:apache beam runners.portability.fn_api.runner.translations:<function sink_flattens at 0x77fd9aff31f0>
INFO:apache beam runners.portability.fn_api.runner.translations:<function greedily_fuse at 0x77fd9aff3280>
INFO:apache beam runners.portability.fn_api.runner.translations:<function read_to_impulse at 0x77fd9aff3310>
INFO:apache beam runners.portability.fn_api.runner.translations:<function impulse_to_input at 0x77fd9aff33a0>
INFO:apache beam runners.portability.fn_api.runner.translations:<function sort_stages at 0x77fd9aff35e0>
INFO:apache beam runners.portability.fn_api.runner.translations:<function add_impulse to dangling transforms at 0x77fd9aff3700>
INFO:apache beam runners.portability.fn_api.runner.translations:<function setup_timer_mapping at 0x77fd9aff3850>
INFO:apache beam runners.portability.fn_api.runner.translations:<function populate_data_channel_coders at 0x77fd9aff3870>
INFO:apache beam.runners.worker.statecache:Creating state cache with size 100
INFO:apache beam.runners.portability.fn_api.runner.worker.handlers:Created Worker handler <apache beam.runners.portability.fn_api_runner.worker_handlers.EmbeddedWorkerHandler object at 0x77fd9af77eb0> for environment ref Environment_default_environment_1 (beam:env:embedded_workers:v1, b'')
```

```
root@1e0cbd8fec0f:~# ls
OUTPUT_FILE-00000-of-00001 bin boot dev etc home lib lib64 media mnt opt proc root run sbin srv sys tmp usr var
root@1e0cbd8fec0f:~ [ ]
```

```
wages: 1
cup: 1
deservings: 1
hang'd: 1
Thou'lt: 1
button: 1
faints: 1
Break: 1
Vex: 1
ghost: 1
rack: 1
Stretch: 1
usurp'd: 1
Friends: 1
Rule: 1
gored: 1
journey: 1
weight: 1
ought: 1
oldest: 1
root@1e0cbd8fec0f:/#
```

4. Task 3: Running a Dataflow Pipeline Remotely

Steps Executed:

1. Set the bucket environment variable:

```
BUCKET=gs://[BUCKET_NAME]
```

2. Ran the WordCount pipeline on **Dataflow**:

```
python -m apache_beam.examples.wordcount \
--project $DEVSHELL_PROJECT_ID \
--runner DataflowRunner \
--staging_location $BUCKET/staging \
--temp_location $BUCKET/temp \
--output $BUCKET/results/output \
--region us-central1
```

```
student_01_9d65d778187b@cloudshell:~ (qwiklabs-gcp-02-c0224ed38017)$ BUCKET=gs://qwiklabs-gcp-02-c0224ed38017-bucket
student_01_9d65d778187b@cloudshell:~ (qwiklabs-gcp-02-c0224ed38017)$ python -m apache_beam.examples.WordCount --project $DEVSHELL_PROJECT_ID \
--runner DataflowRunner \
--staging_location $BUCKET/staging \
--temp_location $BUCKET/temp \
--output $BUCKET/results/output \
--region us-west1
/usr/bin/python: Error while finding module specification for 'apache_beam.examples.wordcount' (ModuleNotFoundError: No module named 'apache_beam')
student_01_9d65d778187b@cloudshell:~ (qwiklabs-gcp-02-c0224ed38017)$
```

CLOUD SHELL Terminal (qwiklabs-gcp-02-c0224ed38017) X + Open Editor

```

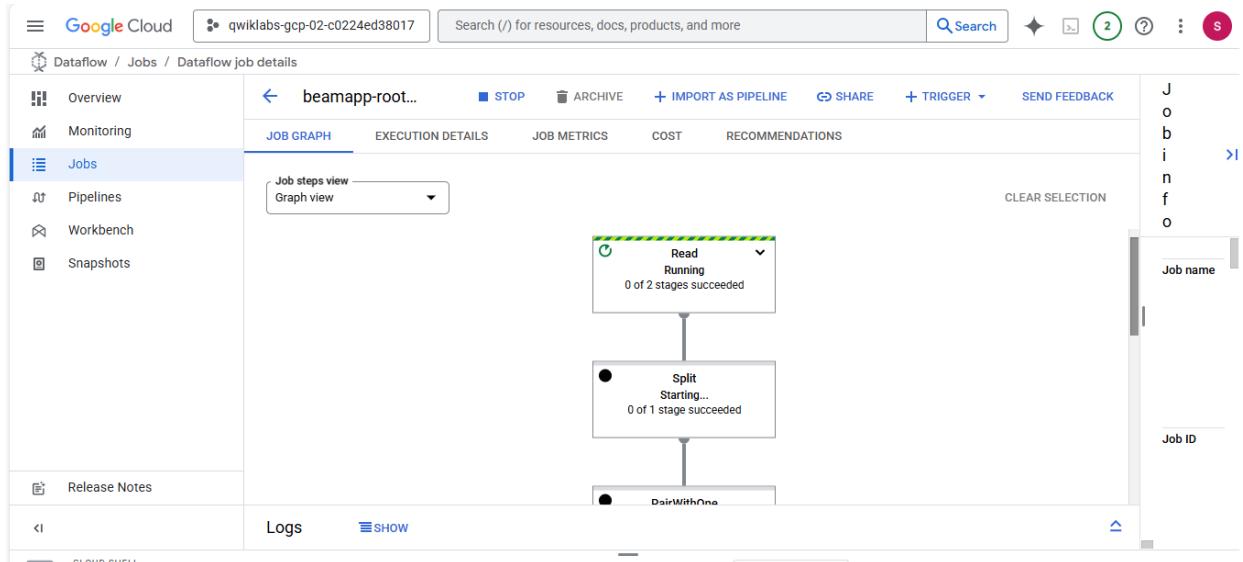
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-06-10T15:54:41.297Z: JOB_MESSAGE_DETAILED: Fusing consumer Write/Write/WriteImpl/WindowInto(WindowIntoFn) into For mat
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-06-10T15:54:41.322Z: JOB_MESSAGE_DETAILED: Fusing consumer Write/Write/WriteImpl/WriteBundles into Write/Write/Wri teImpl/WindowInto(WindowIntoFn)
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-06-10T15:54:41.339Z: JOB_MESSAGE_DETAILED: Fusing consumer Write/Write/WriteImpl/Pair into Write/Write/WriteImpl/W riteBundles
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-06-10T15:54:41.365Z: JOB_MESSAGE_DETAILED: Fusing consumer Write/Write/WriteImpl/GroupByKey/Write into Write/Write/Write /WriteImpl/Pair
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-06-10T15:54:41.386Z: JOB_MESSAGE_DETAILED: Fusing consumer Write/Write/WriteImpl/Extract into Write/Write/WriteImpl/GroupByKey/Read
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-06-10T15:54:41.430Z: JOB_MESSAGE_DEBUG: Workflow config is missing a default resource spec.
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-06-10T15:54:41.451Z: JOB_MESSAGE_DEBUG: Adding StepResource setup and teardown to workflow graph.
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-06-10T15:54:41.465Z: JOB_MESSAGE_DEBUG: Adding workflow start and stop steps.
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-06-10T15:54:41.486Z: JOB_MESSAGE_DEBUG: Assigning stage ids.
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-06-10T15:54:41.615Z: JOB_MESSAGE_DEBUG: Executing wait step start34
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-06-10T15:54:41.710Z: JOB_MESSAGE_BASIC: Executing operation Read/Read/Impulse+Read/Read/Map(<lambda at iobase.py:9 08>)+Read/Read/SDFBoundedSourceReader/ParDo(SDFBoundedSourceDoFn)/PairWithRestriction+Read/Read/SDFBoundedSourceReader/ParDo(SDFBoundedSourceDoFn)/SplitWithSizing
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-06-10T15:54:41.739Z: JOB_MESSAGE_BASIC: Executing operation Write/Write/WriteImpl/DoOnce/Impulse+Write/Write/Write Impl/DoOnce/FlatMap(<lambda at core.py:3401>)+Write/Write/WriteImpl/DoOnce/Map(decode)+Write/Write/WriteImpl/InitializeWrite
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-06-10T15:54:41.743Z: JOB_MESSAGE_DEBUG: Starting worker pool setup.
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-06-10T15:54:41.762Z: JOB_MESSAGE_BASIC: Starting 1 workers in us-west1...
INFO:apache_beam.runners.dataflow.dataflow_runner:Job 2025-06-10_08_54_34-8818699204499638442 is in state JOB_STATE_RUNNING

```

5. Task 4: Verifying Pipeline Success

Steps Executed:

1. Navigated to **Dataflow** in the Cloud Console.
2. Confirmed job status changed to "**Succeeded**".
3. Checked the output in GCS:
 - a. Path: gs://[BUCKET_NAME]/results/output-*
 - b. Contents: Word counts (similar to local test).



6. Key Learnings

- Cloud Storage:** Buckets store pipeline outputs with multi-region redundancy.
- Apache Beam:** Unified model for batch/streaming pipelines.

 **Dataflow:** Serverless execution with auto-scaling workers.

 **Monitoring:** Track jobs via Cloud Console logs.

7. Conclusion

This project successfully:

- Deployed a **WordCount pipeline** on Dataflow.
- Validated end-to-end data processing from local testing to cloud execution.
- Demonstrated GCP's serverless capabilities for large-scale data jobs.

Future Work:

- Process custom datasets (e.g., logs, user-generated content).
- Integrate with **BigQuery** for analytics.