Note: - Virtualenv Address = /home/username/opt/python-2.7.7/bin/python

Run Pipeline Directly

python -m apache\_beam.examples.wordcount --input /path/to/inputfile --output /path/to/write/counts

Dataflow Pipeline

# As part of the initial setup, install Google Cloud Platform specific extra components. Make sure you

# complete the setup steps at https://beam.apache.org/documentation/runners/dataflow/#setup

pip install apache-beam[gcp]

python -m apache\_beam.examples.wordcount --input gs://dataflow-samples/shakespeare/kinglear.txt \

--output gs://<your-gcs-bucket>/counts \

--runner DataflowRunner \

--project your-gcp-project \

--temp\_location gs://<your-gcs-bucket>/tmp/

**DOCUMENTATION**

Instead of processing all elements simultaneously, the elements in a PCollection are processed in bundles.

Dependent-parallelism between transforms

ParDo transforms that are in sequence may be *dependently parallel* if the runner chooses to execute the consuming transform on the producing transform’s output elements without altering the bundling.

## **Branching PCollections**

It’s important to understand that transforms do not consume PCollections; instead, they consider each individual element of a PCollection and create a new PCollection as output. This way, you can do different things to different elements in the same PCollection.

To execute this pipeline locally, first edit the code to specify the output location. Output location could be a local file path or an output prefix on GCS. (Only update the output location marked with the first CHANGE comment.)

To execute this pipeline remotely, first edit the code to set your project ID, runner type, the staging location, the temp location, and the output location. The specified GCS bucket(s) must already exist. (Update all the places marked with a CHANGE comment.)

Then, run the pipeline as described in the README. It will be deployed and run using the Google Cloud Dataflow Service. No args are required to run the pipeline. You can see the results in your output bucket in the GCS browser.

|  |
| --- |
|  |

python -m dummy --input test.csv --output output/

python -m dummy --input gs://hardik\_rategain/test.csv \

--output gs://hardik\_rategain/output \

--runner DataflowRunner \

--project airbnb-150508 \

--temp\_location gs://hardik\_rategain/tmp \

--job\_name=inamedit