

Cloud Computing – Lab

Description of wordcount2.py

What wordcount2.py does is take text as input and parses it into exact words. In essence, it performs word counting on a file. The imports are necessary modules for building data processing pipelines. The 'WordExtractingDoFn' class is used to process lines of text in the input file and parses it into words. The 'def run' function will run the main class using the Apache pipelines. The argument parser parses arguments to specify input and 2 output file locations. Then, the pipeline is constructed and configured. With pipeline running, the program reads from input file and saves it. Then, it will begin extracting and counting. It will split each line into separate words using the 'WordExtractingDoFn' class and converts all words to lowercase. Then it splits into two outputs/pipelines, one pipeline filters words starting from a-f. It PairsWithOne and GroupAndSum. The second pipeline extracts first letter and also does PairWithOne2 and GroupAndSum. Then, both pipelines will format the result. Then will write the file to text. As a result, there will be two different outputs.

WordCount Examples

←

Bucket details

REFRESH

LEARN

to parent page

analog-premise-414918-bucket

Location

Storage class

Public access

Protection

us (multiple regions in United States)

Standard

Not public

None

OBJECTS

CONFIGURATION

PERMISSION

PROTECTION

LIFECYCLE

OBSERVABILITY

INVENTORY REPORTS

Buckets > analog-premise-414918-bucket > result

UPLOAD FILES

UPLOAD FOLDER

CREATE FOLDER

TRANSFER DATA

MANAGE HOLDS

EDIT RETENTION

DOWNLOAD







DELETE

Filter by name prefix only

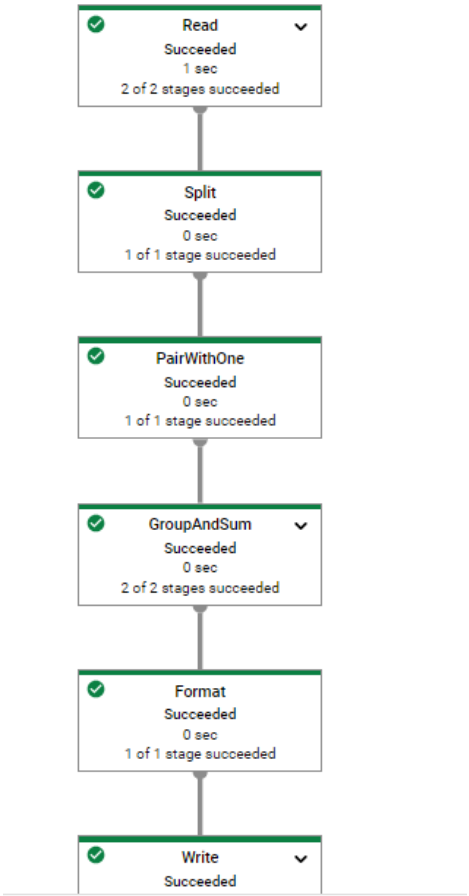
Filter

Filter objects and folders

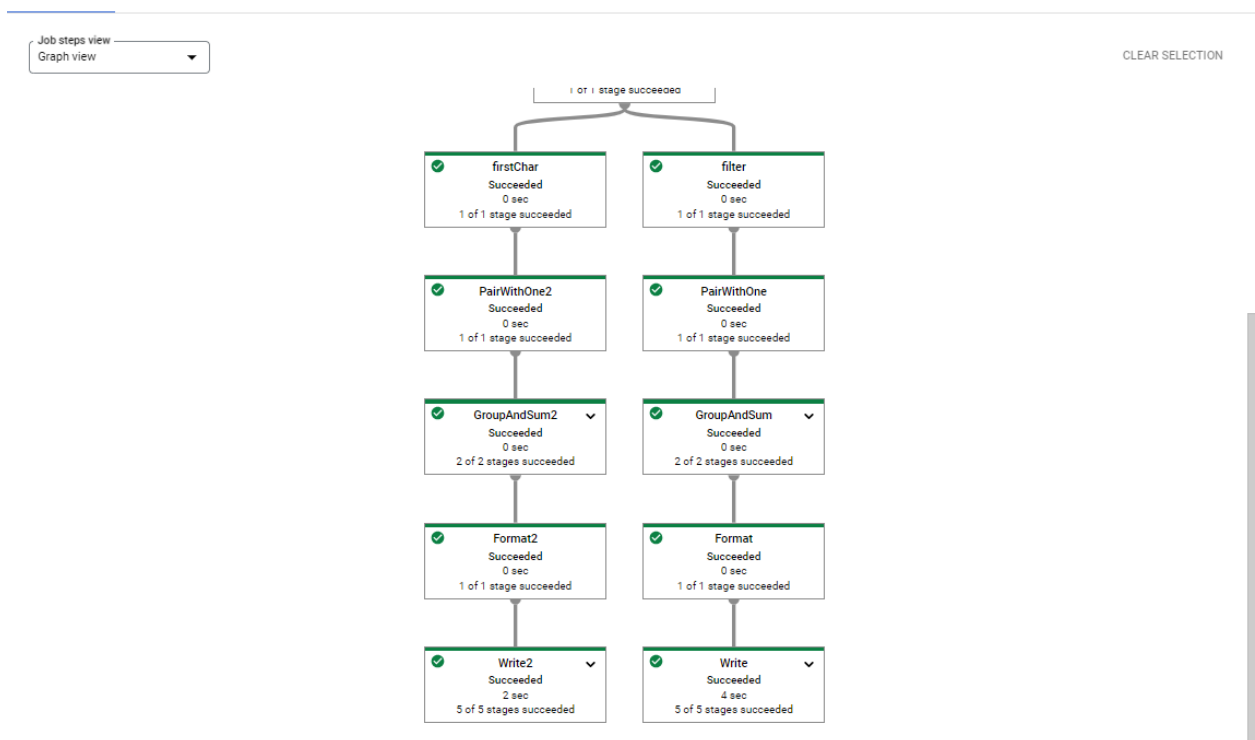
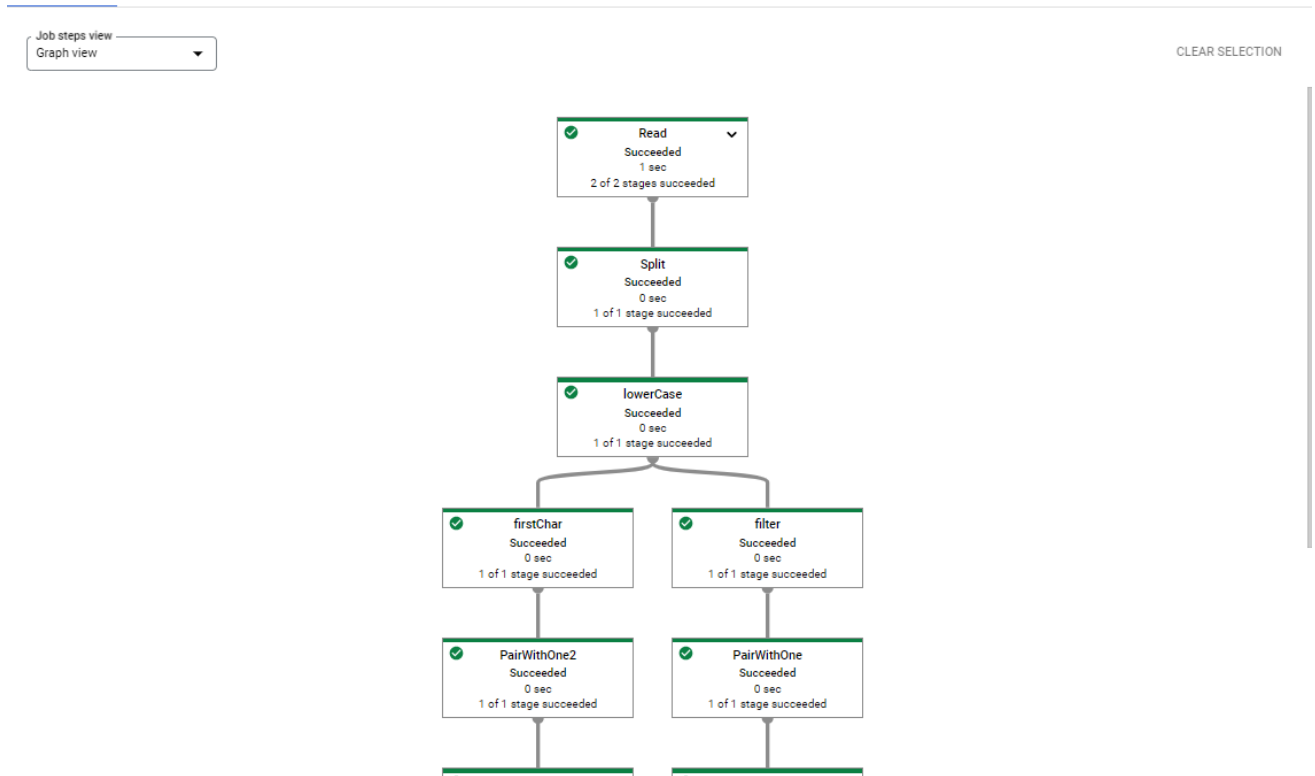
Show deleted data

<input type="checkbox"/>	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encrypti
<input type="checkbox"/>	 outputs-00000-of-00001	13.9 KB	text/plain	20 Feb 2024, 13:52:57	Standard	20 Feb 2024, 13:52:57	Not public	—	Google-  
<input type="checkbox"/>	 outputs2-00000-of-00001	189 B	text/plain	20 Feb 2024, 13:52:54	Standard	20 Feb 2024, 13:52:54	Not public	—	Google-  

Bucket File Outputs



Data Flow Diagram for wordcount.py



Dataflow Diagram for wordcount2.py

MNIST Examples

Your BigQuery projects will have new capabilities after 14 February 2024. Services and roles will be enabled automatically to help with these changes. [Learn more](#) DISMISS

Explorer

+ ADD

I<

Type to search

?

Viewing resources.

SHOW STARRED ONLY

analog-premise-414918

External connections

MNIST

Images

Predict

Untitled 3

RUN

SAVE

DOWNLOAD

SHARE

SCHEDULE

MORE

Query completed.

1 SELECT * FROM `analog-premise-414918.MNIST.Predict` LIMIT 1000

Press Alt+F1 for accessibility options.

Query results

SAVE RESULTS

EXPLORE DATA

JOB INFORMATION

RESULTS

CHART

JSON

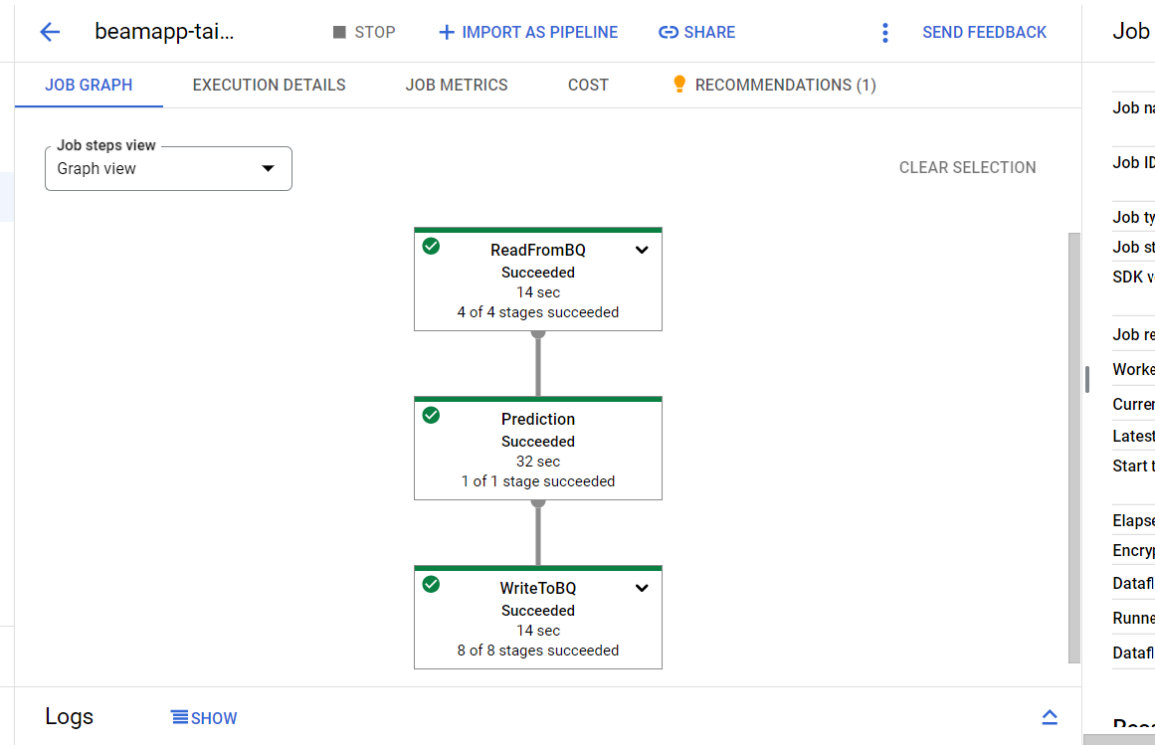
EXECUTION DETAILS

EXECUTION GRAPH

Row	ID	P0	P1	P2	P3	P4	P5	P6
1	2707	1.0	8.506401429393...	3.808996851262...	6.511568428995...	5.193195726435...	1.397785831308...	9.98
2	2824	1.0	5.001731450890...	3.952270688500...	2.805631432473...	1.422910834297...	1.122149942411...	6.33
3	2942	1.0	1.449114342316...	8.970644671535...	8.255261454488...	5.208099813772...	2.719874325762...	9.44
4	2996	1.0	1.523458212338...	1.828361106959...	3.050696500972...	8.937530129800...	5.510777945083...	2.87
5	3044	1.0	3.754634474964...	2.739663607087...	1.490486878884...	4.114576721329...	3.771219794779...	7.42
6	3052	1.0	7.718183697138...	1.281993400326...	1.424809578927...	4.289918798866...	3.833227680584...	4.98
7	3179	1.0	9.842973892139...	1.112963610161...	5.492211181424...	9.541629215659...	1.324008719501...	1.98
8	3226	1.0	9.589175292571...	6.668702812184...	4.503230758803...	3.433702839205...	1.511858520319...	2.34

Results per page: 50 1 - 50 of 1000

SUMMARY



MNIST Example 1 Dataflow Diagram

Topics

CREATE TOPIC

DELETE

HIDE INFO PANEL

Simplify data lake pipelines with new Pub/Sub Cloud Storage subscriptions

Storage subscriptions

NEW

You can now streamline your data ingestion pipelines with Cloud Storage subscriptions, enabling you to write raw streaming data into Cloud Storage without any transformations in between. To get started, create a new Cloud Storage subscription for a Pub/Sub topic.

CREATE CLOUD STORAGE SUBSCRIPTION

LEARN MORE

LIST

METRICS

Filter

Filter topics

<input type="checkbox"/>	Topic ID ↑	Encryption key	Topic name	Retention
<input type="checkbox"/>	mnist_image	Google-managed	projects/analog-premise-414918/topics/mnist_image	—
<input type="checkbox"/>	mnist_predict	Google-managed	projects/analog-premise-414918/topics/mnist_predict	—

Select a topic

PERMISSIONS

LABELS

STORAGE POLICY

Please select at least one resource.

Topics Creation

beamapp-tai...

STOP

CREATE SNAPSHOT

IMPORT AS PIPELINE

SHARE

ARCHIVE

TRIGGER

SEND FEEDBACK

beamapp-taimourahad12-0221045947-937009-20dvm

JOB METRICS

COST

RECOMMENDATIONS (1)

AUTO-SCALING

Job steps view

Graph view

CLEAR SELECTION

Read from Pub

Running

1 stage

toDict

Running

1 stage

Prediction

Running

1 stage

to byte

Running

1 stage

to Pub

Running

1 stage

Job info

>

Job name

beamapp-taimourahad12-0221045947-937009-ybj20dvm

Job ID

2024-02-20_20_59_48-16072889123091640790

Job type

Streaming

Job status

Running

SDK version

Apache Beam Python 3.9 SDK 2.54.0

Job region

northamerica-northeast2

Worker location

northamerica-northeast2

Current workers

1

Latest worker status

Worker pool started.

Straggler status

No active straggler

Start time

20 February 2024 at 23:59:49 GMT-5

Elapsed time

2 min 2 sec

Encryption type

Google-managed

Dataflow Prime

Disabled

Runner v2

Enabled

Streaming Engine

Enabled

Vertical auto-scaling

Disabled

Resource metrics

^

Current vCPUs

2

Total vCPU time

0.033 vCPU hr

Current memory

7.5 GB

Total memory time

0.125 GB hr

Current HDD PD

30 GB

Total HDD PD time

0.5 GB hr

Current SSD PD

0 B

Total SSD PD time

0 GB hr

Pipeline options

^

beam_plugins

['apache_beam.io.filesystem.FileSystem', 'apache_beam.io.hadoopfilesystem.HadoopFileSystem']

MNIST Example 2 Dataflow Diagram

Videos

WordCount:

https://drive.google.com/file/d/1CqXCHB9jXUYFiFxhIFzWEvE4UHx_cIJS/view?usp=sharing

MNIST:

https://drive.google.com/file/d/1N6w7N3MVVMhOA2_6rq_mpDZUUYQHTWJC/view?usp=sharing

Design

GitHub Design: <https://github.com/TaimourArshad1/cloud-lab2>