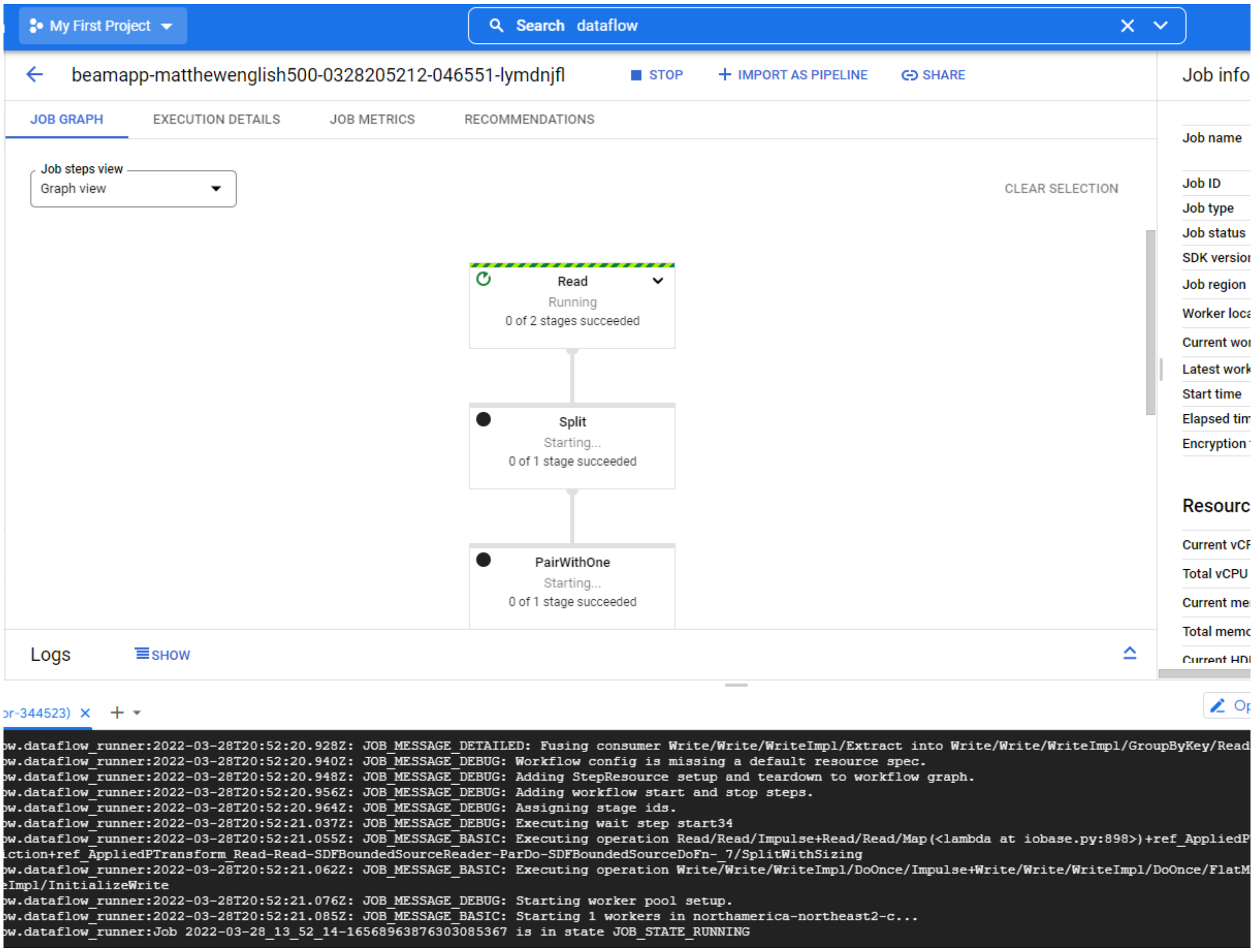
Data Processing: Data flow - Apache Beam

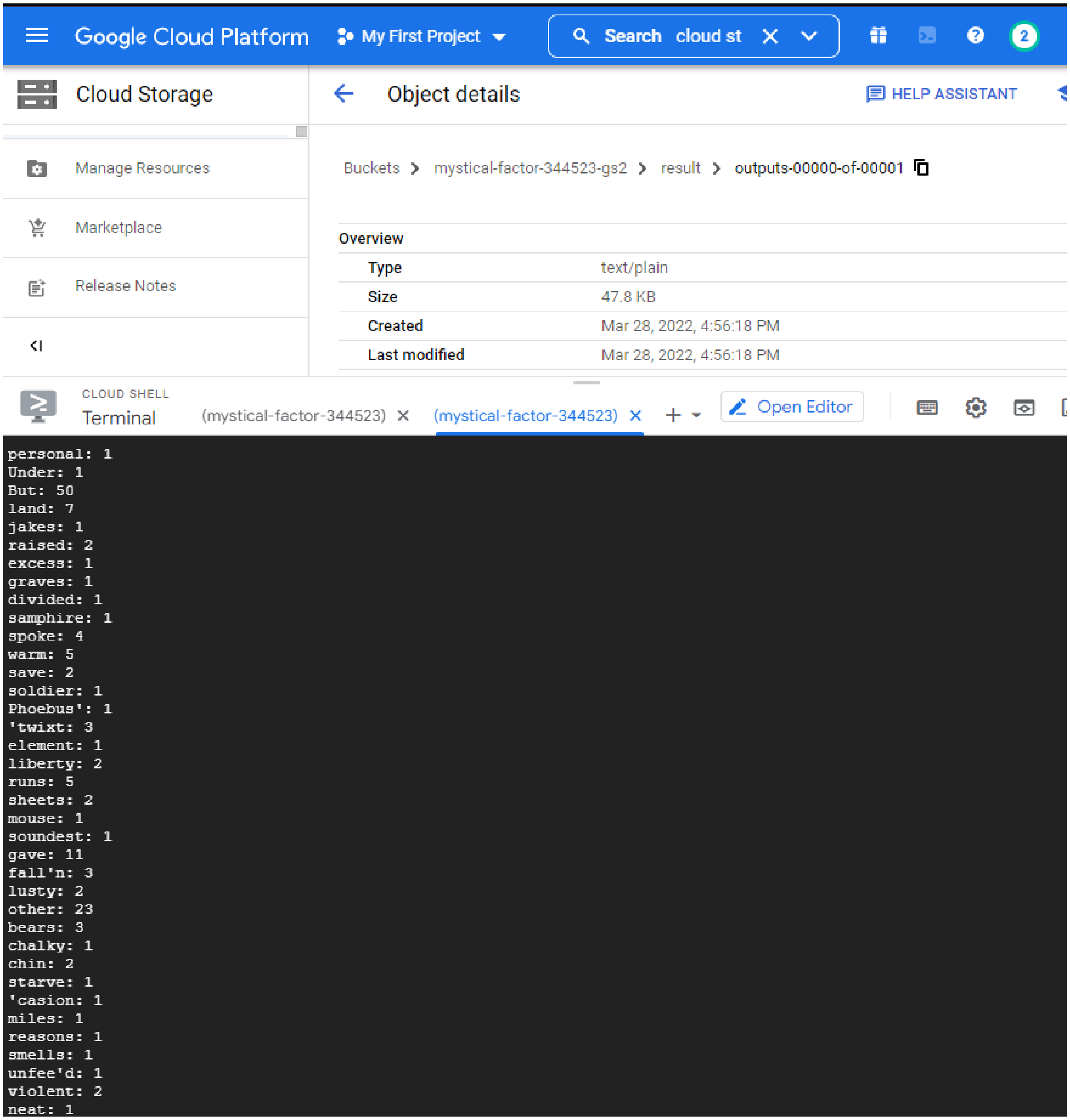
Matthew English (100704553)

Lab Tasks:

Follow the following video to set up the GCP environment for Dataflow and run wordcount examples.



```
ow.dataflow_runner:2022-03-28T20:52:20.928Z: JOB_MESSAGE_DETAILED: Fusing consumer Write/Write/WriteImpl/Extract into Write/Write/WriteImpl/GroupByKey/Read
ow.dataflow_runner:2022-03-28T20:52:20.940Z: JOB_MESSAGE_DEBUG: Workflow config is missing a default resource spec.
ow.dataflow_runner:2022-03-28T20:52:20.948Z: JOB_MESSAGE_DEBUG: Adding StepResource setup and teardown to workflow graph.
ow.dataflow_runner:2022-03-28T20:52:20.956Z: JOB_MESSAGE_DEBUG: Adding workflow start and stop steps.
ow.dataflow_runner:2022-03-28T20:52:20.964Z: JOB_MESSAGE_DEBUG: Assigning stage ids.
ow.dataflow_runner:2022-03-28T20:52:21.037Z: JOB_MESSAGE_DEBUG: Executing wait step start34
ow.dataflow_runner:2022-03-28T20:52:21.055Z: JOB_MESSAGE_BASIC: Executing operation Read/Read/Impulse+Read/Read/Map(<lambda at iobase.py:898>)+ref_AppliedP
iction+ref_AppliedPTransform_Read-Read-SDFBoundedSourceReader-ParDo-SDFBoundedSourceDoFn-_7/SplitWithSizing
ow.dataflow_runner:2022-03-28T20:52:21.062Z: JOB_MESSAGE_BASIC: Executing operation Write/Write/WriteImpl/DoOnce/Impulse+Write/Write/WriteImpl/DoOnce/FlatM
eImpl/InitializeWrite
ow.dataflow_runner:2022-03-28T20:52:21.076Z: JOB_MESSAGE_DEBUG: Starting worker pool setup.
ow.dataflow_runner:2022-03-28T20:52:21.085Z: JOB_MESSAGE_BASIC: Starting 1 workers in northamerica-northeast2-c...
ow.dataflow_runner:Job 2022-03-28_13_52_14-16568963876303085367 is in state JOB_STATE_RUNNING
```



Object details

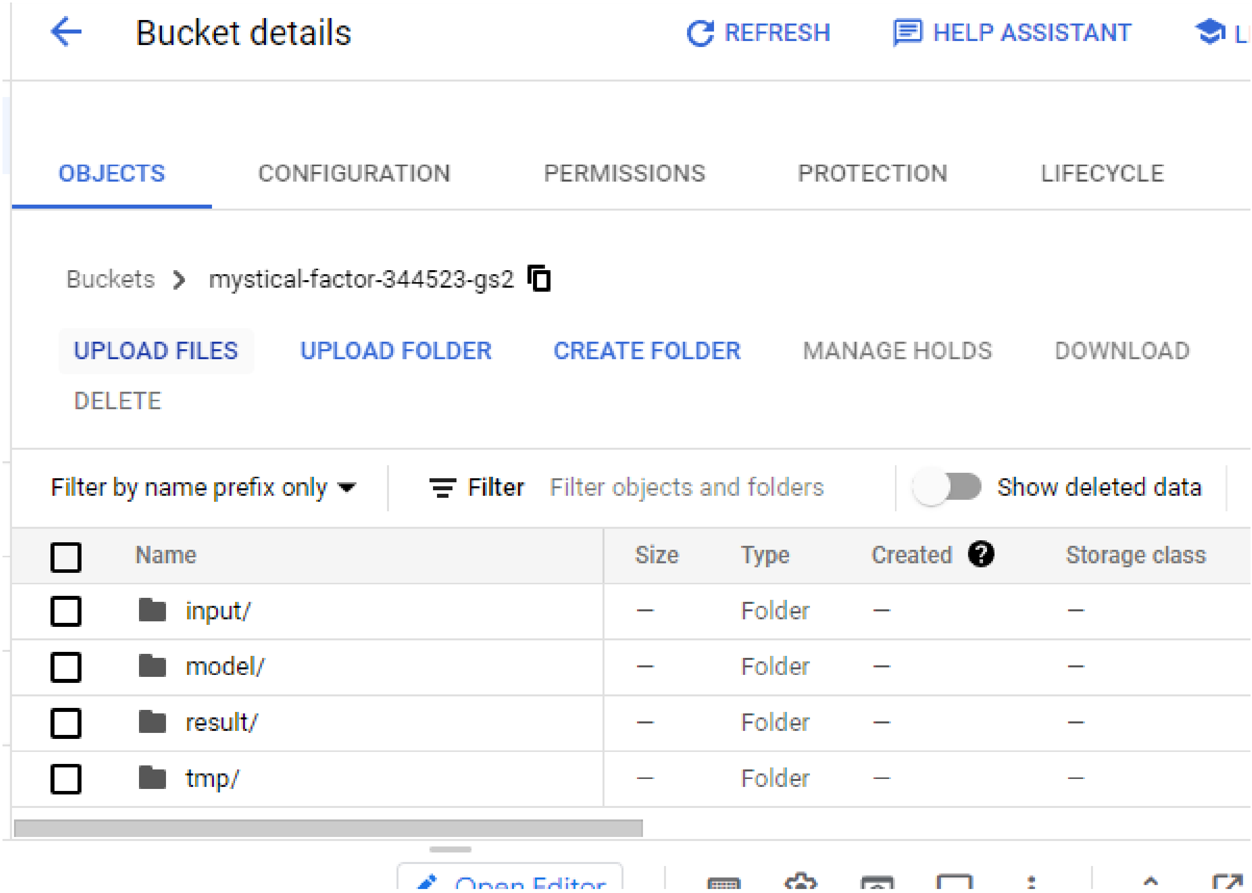
Buckets > mystical-factor-344523-gs2 > result > outputs-00000-of-00001

Overview

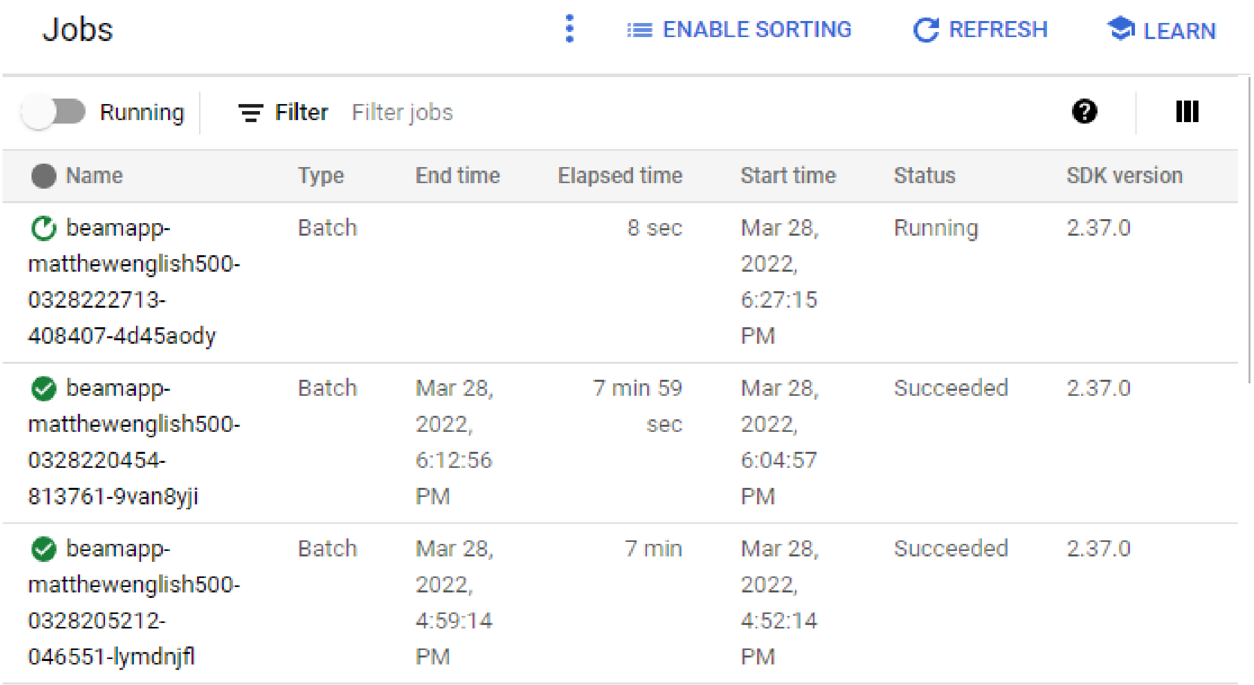
Type	text/plain
Size	47.8 KB
Created	Mar 28, 2022, 4:56:18 PM
Last modified	Mar 28, 2022, 4:56:18 PM

```
personal: 1
Under: 1
But: 50
land: 7
jakes: 1
raised: 2
excess: 1
graves: 1
divided: 1
samphire: 1
spoke: 4
warm: 5
save: 2
soldier: 1
Phoebus': 1
'twixt: 3
element: 1
liberty: 2
runs: 5
sheets: 2
mouse: 1
soundest: 1
gave: 11
fall'n: 3
lusty: 2
other: 23
bears: 3
chalky: 1
chin: 2
starve: 1
'casion: 1
miles: 1
reasons: 1
smells: 1
unfee'd: 1
violent: 2
neat: 1
```

Follow the following videos for various Dataflow examples for Batch and stream processing for the mnist dataset for various source and destination types; text file, MySQL database, and Kafka topics.



□	Name	Size	Type	Created	Storage class
□	input/	–	Folder	–	–
□	model/	–	Folder	–	–
□	result/	–	Folder	–	–
□	tmp/	–	Folder	–	–



Name	Type	End time	Elapsed time	Start time	Status	SDK version
beamapp-matthewenglish500-0328222713-408407-4d45aody	Batch		8 sec	Mar 28, 2022, 6:27:15 PM	Running	2.37.0
beamapp-matthewenglish500-0328220454-813761-9van8yji	Batch	Mar 28, 2022, 6:12:56 PM	7 min 59 sec	Mar 28, 2022, 6:04:57 PM	Succeeded	2.37.0
beamapp-matthewenglish500-0328205212-046551-lymdnjfl	Batch	Mar 28, 2022, 4:59:14 PM	7 min	Mar 28, 2022, 4:52:14 PM	Succeeded	2.37.0

I was not able to complete the final step of this video as the module beam_nuggets was not found when running predictV2. I tried many times to reinstall beam-nuggets but I could not find a solution. I finished watching the video and understand the next steps of using kafka producers and consumers alongside dataflow.

Google Cloud has another processing service called DataProc. Name another processing service that is usually used in the cloud environment (not necessarily GCP). Compare between it and both Dataflow and DataProc. Your comparison may include but is not limited to the major differences, advantages, disadvantages, and limitations.

Processing Service	Dataflow	DataProc	Dataprep
Main use	-Manages and operates batch and stream processing - Automatic provisioning to clusters	-Uses Spark and Hadoop for processing -Supports manual provision to clusters	-Explore data visually by converting the file into JSON,CSV or a table format. - Helps prepare data for further use
Advantages	- Fast - Simple	- Create clusters quickly - Easy to manage - Cost efficient	- Easy to understand the data - Useful for putting use to the data
Limitations/Disadvantages	- Job computation power limationas	- Size of clusters must be known in advance as it cannot be changed	- Only used as a medium for processing data

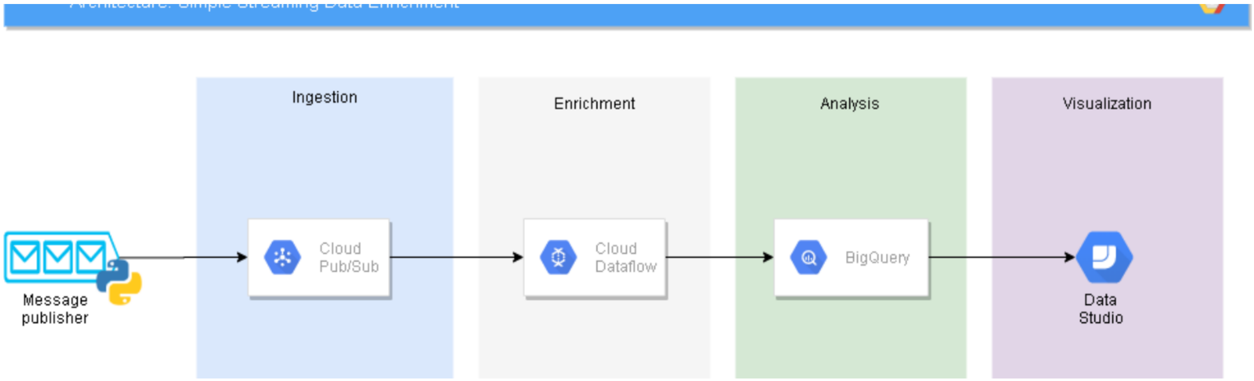
.

Suggest a practical application using both stream and batch processing that can be applied to a given dataset. It's expected to use the dataset uploaded in the third milestone but you can use any other dataset. If you decide to use another dataset, It should maintain both variety and huge volume. Your report should include but not limited to:

- The application.
- Its impact.
- The used dataset (size, schema/structure).
- A graph showing the proposed pipeline(s).
- List of other tools (AI, clustering,…) needed to implement that application.

An application that uses both stream and batch processing could be a fraud detection application. With a large dataset of transactions, you can use stream processing to quickly process data for noticeable fraud red flags while using batch processing to store large amounts of data then search for complex patterns between the transactions. This could have a large impact to quickly detect fraud on the spot or discover fraudulent transactions where we would not have noticed before.

The dataset would likely be a very large dataset of every piece of information of a person, the transaction details and the parties involved in the transaction. The application pipeline could follow a similar pattern to the figure below. The transaction data would get sent to the cloud where it could use BigQuery to help support the large datasets where the transaction data can be visualized to better understand the connections between transaction.


Architecture: Simple Streaming Data Enrichment
Ingestion
Enrichment
Analysis
Visualization
Message publisher
Cloud Pub/Sub
Cloud Dataflow
BigQuery
Data Studio