

1) (The code for question 1 & 2 is written in a single file and attached with the ZIP file submission)

The screenshot shows the Google Cloud Platform interface for a bucket named 'naveenrd'. The 'OBJECTS' tab is selected, displaying a list of files. The table below represents the data shown in the screenshot.

Name	Size	Type	Created time	Storage class	Last modified	Public access	Encryption	Retention expiration date	Holds
n_avg_words.txt-00000-of-0001	18 B	text/plain	Feb 27, 2021, 3:21:27 ...	Standard	Feb 27, 2021, 3:21:27 ...	Not public	Google-managed key	—	None
n_lines.txt-00000-of-00001	9 B	text/plain	Feb 27, 2021, 3:21:21 ...	Standard	Feb 27, 2021, 3:21:21 ...	Not public	Google-managed key	—	None

In the above screenshot **n_lines.txt** contains the number of lines in the given data file.
The number of words as obtained after dataflow operation = 38990771.

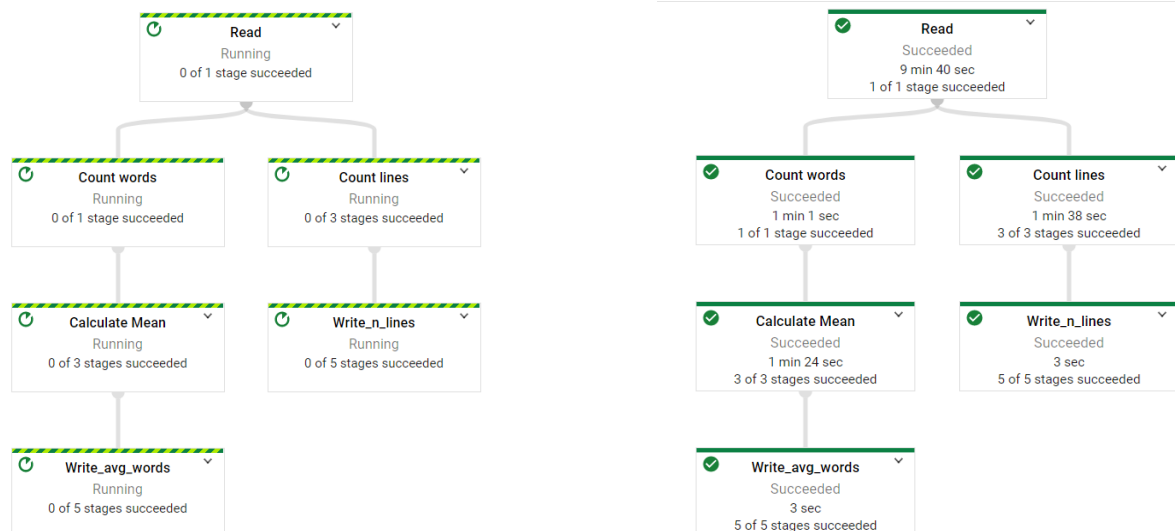
2) (The code for question 1 & 2 is written in a single file and attached with the ZIP file submission)

The screenshot shows the Google Cloud Platform interface for a bucket named 'naveenrd'. The 'OBJECTS' tab is selected, displaying a list of files. The table below represents the data shown in the screenshot.

Name	Size	Type	Created time	Storage class	Last modified	Public access	Encryption	Retention expiration date	Holds
n_avg_words.txt-00000-of-0001	18 B	text/plain	Feb 27, 2021, 3:21:27 ...	Standard	Feb 27, 2021, 3:21:27 ...	Not public	Google-managed key	—	None
n_lines.txt-00000-of-00001	9 B	text/plain	Feb 27, 2021, 3:21:21 ...	Standard	Feb 27, 2021, 3:21:21 ...	Not public	Google-managed key	—	None

In the above screenshot **n_avg_words.txt** contains the average number of words per line in the file.
The average number of words per line obtained by dataflow operation = 2.000363470627447.

3)



- 4) The pipeline is built in such a way that calculating the number of lines & average number of words per line is done in a single run of the dataflow.

The first step is to read the text file placed at ***"gs://iitmbd/out.txt"*** and convert the lines into PCollection.

```
lines = p | 'Read' >> beam.io.ReadFromText('gs://iitmbd/out.txt')
```

Now we have PCollection & PCollections are immutable. That means the same PCollections can be used for various transformation operations. Here we have two tasks at hand: counting number of lines in the text file & counting average number of words per line in the text file. So, the lines PCollection is used for two branches. It can be seen in the execution graph presented in the previous question. The left branch is for calculating the average number of lines in a line & the right branch is for calculating the number is lines in the text file.

```
# The below line of code is to count the number of lines in the text file.
n_lines = lines | 'Count_lines' >> beam.combiners.Count.Globally()
               | 'Write_n_lines' >> beam.io.WriteToText('gs://naveenrd/outputs/n_lines.txt')

# The below line of code is to evaluate the average number of words per line.
n_avg_words = lines | 'Count words' >> beam.FlatMap(Lambda Line: [len(line.split(' '))])
                  | 'Calculate Mean' >> beam.combiners.Mean.Globally()
                  | 'Write_avg_words' >> beam.io.WriteToText('gs://naveenrd/outputs/n_avg_words.txt')
```

Counting Lines: We used ***"beam.combiners.Count.Globally()"*** to count the Number of elements in all the PCollections, which is basically the number of lines. Then wrote it to the cloud bucket as ***"n_lines.txt"***.

Counting Average number of words/line: First we used ***Flatmap inline lambda*** function to split every line into words and output number of words as a PCollection. Then this PCollection is used with ***"beams.combiners.Mean.Globally()"*** to evaluate the average number of lines per line. This number is then written to ***'n_avg_words.txt'*** in the cloud bucket.

Bucket Name = naveenrd

Runner Name = lab-3-lines

- 5) To run google cloud function to calculate the number of lines and average number of words per line, a separate bucket should be used for the text file because if same bucket is used for temporary folder and outputs then it will go on an infinite loop because the outputs are text files.

The **main.py** file & **requirements.txt** are placed in the cloud function folder in the zip file.

Code to deploy the cloud function.

```
gcloud functions deploy lab3_cloudfncn --runtime python37 --trigger-resource lab3_cloudfncn --trigger-event google.storage.object.finalize
```

Trigger Resource Bucket: lab3_cloudfncn

Cloud Function Name: lab3_cloudfncn

The screenshot shows the Google Cloud Platform console interface. At the top, there's a navigation bar with 'Google Cloud Platform' and 'Lab 3'. Below it, the 'Bucket details' page for 'naveenrd' is displayed. The 'OBJECTS' tab is selected, showing a list of objects. The table has columns: Name, Size, Type, Created time, Storage class, Last modified, Public access, Encryption, Retention expiration date, and Holds. One object is listed: 'n_lines.txt-00000-of-00001' with a size of 9 B, type 'text/plain', created on Mar 1, 2021, 8:47:36 PM, storage class 'Standard', last modified on Mar 1, 2021, 8:47:36 PM, public access 'Not public', encryption 'Google-managed key', retention expiration date '-', and holds 'None'.

Name	Size	Type	Created time	Storage class	Last modified	Public access	Encryption	Retention expiration date	Holds
n_lines.txt-00000-of-00001	9 B	text/plain	Mar 1, 2021, 8:47:36 PM	Standard	Mar 1, 2021, 8:47:36 PM	Not public	Google-managed key	-	None

Google Cloud Platform

Lab 3

Search products and resources

Cloud Functions

Function details

EDIT

DELETE

COPY

lab3_cloudfn

Version 3, deployed at Mar 1, 2021, 8:35:46 PM...

METRICS

DETAILS

SOURCE

VARIABLES

TRIGGER

PERMISSIONS

LOGS

TESTING

Logs

Showing 138 messages

Default

Filter

Filter logs

2021-03-01T15:07:13.536Z

lab3_cloudfn

1wedsqj2vud

2021-03-01T15:07:09.316Z: JOB_MESSAGE_DEBUG: Value "Write_n_lines/Write/WriteImpl/GroupByKey/Session" materialized.

2021-03-01T15:07:13.536Z

lab3_cloudfn

1wedsqj2vud

2021-03-01T15:07:09.349Z: JOB_MESSAGE_DEBUG: Value "Count lines/CombineGlobally(CountCombineFn)/CombinePerKey/GroupByKey/Session" materialized.

2021-03-01T15:07:13.536Z

lab3_cloudfn

1wedsqj2vud

2021-03-01T15:07:09.419Z: JOB_MESSAGE_BASIC: Executing operation Read/Read+Count lines/CombineGlobally(CountCombineFn)/KeyWithVoid+Count lines/CombineGlo...

2021-03-01T15:07:13.628Z

lab3_cloudfn

1wedsqj2vud

Job 2021-03-01_07_01-2833102909497602169 is in state JOB_STATE_RUNNING

2021-03-01T15:07:56.875070300Z

lab3_cloudfn

jtiu8ak4i59z

Function execution took 500003 ms, finished with status: 'timeout'

2021-03-01T15:08:38.334Z

lab3_cloudfn

1wedsqj2vud

2021-03-01T15:08:26.158Z: JOB_MESSAGE_DETAILED: Autoscaling: Raised the number of workers to 1 based on the rate of progress in the currently running sta...

2021-03-01T15:08:45.699Z

lab3_cloudfn

1wedsqj2vud

2021-03-01T15:08:45.549Z: JOB_MESSAGE_DETAILED: Workers have started successfully.

2021-03-01T15:08:45.699Z

lab3_cloudfn

1wedsqj2vud

2021-03-01T15:08:45.576Z: JOB_MESSAGE_DETAILED: Workers have started successfully.

2021-03-01T15:08:45.576Z

lab3_cloudfn

1wedsqj2vud

2021-03-01T15:08:45.576Z: JOB_MESSAGE_DETAILED: Workers have started successfully.

2021-03-01T15:12:28.579Z

lab3_cloudfn

1wedsqj2vud

2021-03-01T15:12:19.349Z: JOB_MESSAGE_BASIC: Autoscaling: Resizing worker pool from 1 to 3.

2021-03-01T15:12:25.669Z

lab3_cloudfn

1wedsqj2vud

2021-03-01T15:12:24.575Z: JOB_MESSAGE_DETAILED: Autoscaling: Raised the number of workers to 2 based on the rate of progress in the currently running sta...

2021-03-01T15:12:25.669Z

lab3_cloudfn

1wedsqj2vud

2021-03-01T15:12:24.601Z: JOB_MESSAGE_DETAILED: Resized worker pool to 2, though goal was 3. This could be a quota issue.

2021-03-01T15:12:35.869Z

lab3_cloudfn

1wedsqj2vud

2021-03-01T15:12:34.893Z: JOB_MESSAGE_DETAILED: Autoscaling: Raised the number of workers to 3 based on the rate of progress in the currently running stage(s).

2021-03-01T15:12:34.893Z

lab3_cloudfn

1wedsqj2vud

2021-03-01T15:12:34.893Z: JOB_MESSAGE_DETAILED: Autoscaling: Raised the number of workers to 3 based on the rate of progress in the currently running stage(s).

2021-03-01T15:14:48.943291291Z

lab3_cloudfn

1wedsqj2vud

Function execution took 500003 ms, finished with status: 'timeout'

2021-03-01T15:14:48.943291291Z

lab3_cloudfn

1wedsqj2vud

Function execution took 500003 ms, finished with status: 'timeout'

No newer entries found matching current filter.

Google Cloud Platform

Lab 3

Search products and resources

Jobs

CREATE JOB FROM TEMPLATE

CREATE JOB FROM SQL

Enable sorting

LEARN

Running

Filter jobs

Name	Type	End time	Elapsed time	Start time	Status	SDK version	ID	Region
lab-3-cloud-function	Batch	Mar 1, 2021, 8:48:48 PM	11 min 46 sec	Mar 1, 2021, 8:37:02 PM	Succeeded	2.28.0	2021-03-01_07_07_01-2833102909497602169	us-central1
lab-3-lines	Batch	Feb 27, 2021, 3:22:15 PM	11 min 4 sec	Feb 27, 2021, 3:11:11 PM	Succeeded	2.27.0	2021-02-27_01_41_10-9636390815893976762	us-central1
lab-3-lines	Batch	Feb 27, 2021, 3:06:55 PM	10 min 52 sec	Feb 27, 2021, 2:56:03 PM	Succeeded	2.27.0	2021-02-27_01_26_02-16340977527528683017	us-central1
lab-3-cloud-function	Batch	Feb 27, 2021, 1:25:27 PM	11 min 26 sec	Feb 27, 2021, 1:14:01 PM	Succeeded	2.27.0	2021-02-26_23_44_00-1063080752923718231	us-central1
lab-3-lines	Batch	Feb 26, 2021, 3:59:04 PM	10 min 35 sec	Feb 26, 2021, 3:48:29 PM	Succeeded	2.27.0	2021-02-26_02_18_28-13842842605107911744	us-central1
lab-3-lines	Batch	Feb 26, 2021, 2:42:24 PM	9 min 34 sec	Feb 26, 2021, 2:32:50 PM	Succeeded	2.27.0	2021-02-26_01_02_49-365449294311877729	us-central1
lab-3	Batch	Feb 26, 2021, 2:34:02 PM	8 min 25 sec	Feb 26, 2021, 2:25:37 PM	Canceled	2.27.0	2021-02-26_00_55_36-11239569691126963275	us-central1
lab-3	Batch	Feb 26, 2021, 12:46:35 PM	11 min 4 sec	Feb 26, 2021, 12:35:31 PM	Succeeded	2.27.0	2021-02-25_23_05_30-1162355829438372023	us-central1
lab-3	Batch	Feb 26, 2021, 12:30:04 PM	11 min 40 sec	Feb 26, 2021, 12:18:24 PM	Succeeded	2.27.0	2021-02-25_22_48_23-2747450405106108790	us-central1
lab-3	Batch	Feb 26, 2021, 11:08:50 AM	10 min 59 sec	Feb 26, 2021, 10:57:51 AM	Succeeded	2.27.0	2021-02-25_21_27_50-3741078626652959661	us-central1
lab-3	Batch	Feb 26, 2021, 10:45:55 AM	5 sec	Feb 26, 2021, 10:45:50 AM	Failed	2.27.0	2021-02-25_21_15_47-10765269453099371479	us-central1

Rows per page: 50 1 - 11 of 11

Google Cloud Platform

Lab 3

Search products and resources

Bucket details

REFRESH

LEARN

naveenrd

OBJECTS

CONFIGURATION

PERMISSIONS

RETENTION

LIFECYCLE

Buckets

naveenrd

UPLOAD FILES

UPLOAD FOLDER

CREATE FOLDER

MANAGE HOLDS

DOWNLOAD

DELETE

Filter by name prefix only

Filter

Filter objects and folders

Name	Size	Type	Created time	Storage class	Last modified	Public access	Encryption	Retention expiration date	Holds
outputs/	—	Folder	—	—	—	—	—	—	—
outputs_cloud_function/	—	Folder	—	—	—	—	—	—	—
temp_folder1/	—	Folder	—	—	—	—	—	—	—
temp_folder2/	—	Folder	—	—	—	—	—	—	—
test.txt	15 B	text/plain	Feb 21, 2021, 5:58:25 PM	Standard	Feb 21, 2021, 5:58:25 PM	Not public	Google-managed key	—	None

