Lab - 3: Dataflow

Code for Demo shown in class:

```
import apache_beam as beam
from apache_beam.io import ReadFromText
from apache_beam.io import WriteToText
from apache_beam.options.pipeline_options import PipelineOptions
from apache_beam.options.pipeline_options import GoogleCloudOptions
from apache_beam.options.pipeline_options import StandardOptions
options = PipelineOptions()
google_cloud_options = options.view_as(GoogleCloudOptions)
google_cloud_options.project = 'bdl2022' # Enter your project ID
google_cloud_options.job_name = 'lab3'
google_cloud_options.temp_location = "gs://bdl2022/tmp"
google_cloud_options.region = "us-central1"
options.view_as(StandardOptions).runner = 'DataflowRunner'
p = beam.Pipeline(options=options)
lines = p | 'Read' >> beam.io.ReadFromText('gs://bdl2022/lines_big.txt') |'Write' >>
beam.io.WriteToText('gs://bdl2022/outputs/')
result = p.run()
```

We encourage you to go through the Cloud Dataflow Model documentation before starting the assignment. It will introduce you to some transforms and reducers required in the assignment.

Assignment(Due on 19/02/22 23:59:59):

1. Write a Python code to count lines of the file placed in the BDL2022 bucket (gs://bdl2022/lines_big.txt) using

Dataflow and provide the screenshot of the file that is generated in your bucket. [2]

2. Write a Python code to get the average number of words in a line of the file placed in the BDL2022 bucket (gs://bdl2022/lines_big.txt) using Dataflow provide the screenshot of the file that is generated in your bucket.

   [4]

3. Provide the screenshot for the execution graph created by Dataflow in the background for the pipeline object created for questions 1 and 2. [2]

4. Explain the pipeline used in the first two questions. What issues did you face while trying to make the code work for the first two questions, and how did you resolve them? [2]

5. [Bonus] Trigger a dataflow using GCF for any one of the first two questions. [2]

Create a PDF file containing answers to the above questions. Zip it along with the output files (for the dataflow task), screenshots, and your Python files. Then, submit this zip file on moodle.