This take-home assignment is targeted towards Data Engineer/Cloud Data Scientist position for CGI Inc. The task is a basic data pipeline design that consists of the ETL process. Data dictionary as follows:

* Title: Randomly generated sentences. Title may include special characters.
* ID: Unique Identifier.
* Publication\_DATE: date the posts are published. Format: MM/DD/YYYY
* Score: Calculated score to understand weight of each record
* Type: Roman value for the record

**Task description:** You will need to design and develop the ETL pipeline. There will be different stages in the process. The stages are as follows:

* Save the given file anywhere in your local directory
* Dump 10 records every 5 seconds in this CSV file
* Extract the csv to a middleware using Apache Kafka, Apache NiFi, Apache Airflow or any similar **on-prem or cloud platform** of your choice. Every 5 seconds.
* Once the data is loaded into the middleware, apply the following transformations:
  + Change the date format to: YYYY-MM-DD
  + Remove any special character from the **“title”** column
  + Remove missing values
  + Remove duplicate records
* Once the data is transformed, then load it to a table in a cloud platform. You can select the cloud platform of your choice. Examples: AWS, GCP, Azure etc.

**Note: all the processes need to run sequentially every 5 seconds**

**Note: If you do not have access to a cloud platform, simply create two different folders and consider one as source and the other as destination. You can then transform the data in memory on the fly.**

Once it is loaded into the cloud, please demonstrate the following:

* Dummy data is generated every 5 seconds and appended into the existing csv file at your local directory
* As the file is updated every 5 second, it should automatically be loaded into the middleware
* The transformations will take place
* Transformed data will be loaded into destination (i.e the table in cloud platform)

Please, send the git link once you are done. You will need to prepare a brief presentation to demonstrate the automation of this pipeline. Include a video if you’d like in the presentation. From the time of this document received, you have 7 days to complete the task. Please reach out to [Mustakim.helal@cgi.com](mailto:Mustakim.helal@cgi.com) for any clarification needed.