1. Watch the following video about Google Cloud Dataflow https://youtu.be/KalJ0VuEM7s (Links to an external site.)
2. Watch the following video Describing how to apply MapReduce to count the words within a certain document. https://youtu.be/re6c_ee7uTc (Links to an external site.)
3. Follow the following video to set up the GCP environment for Dataflow and run wordcount examples. https://youtu.be/re6c_ee7uTc (Links to an external site.)
    - Video Demo: https://drive.google.com/file/d/1RUnxRnh4xs3QYBYnyluIzAvNI9cZw_Pz/view?usp=sharing

4. 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. https://youtu.be/9ZDj9KDGtEs (Links to an external site.)
    - Video Demo: https://drive.google.com/file/d/1WSGmnLMNNuYuQs0-34VSjbbJHQ_Ib1J3/view?usp=sharing

5. (optional) The following video describes how to use BigQuery and Google PubSub as sources and destinations for the Dataflow pipeline. https://youtu.be/bqY46hS6Y7U (Links to an external site.)
6. 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.
    - The other processing service is DataPrep.
    - Differences:
        - Automation of a cluster is done by DataPrep and DataFlow, while DataProc relies on manual setup for a cluster.
        - Suitable for BigQuery, while the other processing service is more suitable for Apache and Hadoop.
    - Advantages:
        - Improve data visualization as the user interface is displayed in a tabular format.
        - High security as the dataset is protected due to the access management applied.
        - Utilized machine learning to remove unnecessary data to reduce the size of the datasets.
    - Disadvantage:
        - Not easy to use as DataFlow.
        - Heavily used for systems that are focused on user interface.
        - Not working well with Apache Hadoop for the processing service.
    - Limitations:
        - There is maximum of 1000 datasets can be stored in the database

    - There is a lack of memory size that can be used for storing the datasets.
    - Does not support UTF-32.

7. 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 decided 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.