Project Milestone 04: Data Processing - DataFlow and Apache Beam

- In Amazon's cloud environment (AWS), Amazon Data Pipeline is used as its data processing software
- Here's the following comparisons between Google Cloud Dataflow, Google Cloud DataProc, and Amazon Data Pipeline

	Google Cloud **DataFlow**	Google Cloud **DataProc**	Amazon Web Services **Data Pipeline**
Batch Processing	Yes	Yes	Yes
Stream Processing	Yes	No	Yes
Provisioning	Automatic	Manual	Automatic
Data Writes	PubSub, GCS, BigQuery, SQL, etc.	Vertex AI, BigQuery, DataPlex	DynamoDB, SQL, RedShift tables
Data Sharing	Yes	Yes	Yes

- Advantages:
  - DataFlow
    - It is fully managed, so data is secured and provides automated provisioning/management of processing resources
    - Horizontally scalable to ensure all resources are utilized efficiently
    - It is very reliable and consistent
  - DataProc
    - Very secure as it provides multiple security integrations such as Apache Ranger, Kerberos, and Personal Authentication
    - It is flexible to use, as it is serverless and uses Kubernetes to manage the clusters
    - It is cost-effective compared to other options, as it uses 57% less than the average total cost of ownership of on-premise data lakes
  - Data Pipeline
    - Very reliable as it uses a distributed infrastructure
      - Distributed infrastructure provides fault tolerance and data security
    - Can be flexible to suit the customer's needs
      - Provides a wide variety of functionality and features that the customer can choose to implement with their system

        - It is easy to use as the structure of Data Pipeline can be designed through a drag and drop console
            - No extra written logic is required
            - As long as preconditions are met, nothing else is required
- Disadvantages/Limitations
    - DataFlow
        - Not preferred if Hadoop dependencies are required
        - Not as fast compared to other data processing software
            - But not as significant
    - DataProc
        - DataProc doesn't offer stream processing (hot path)
        - DataProc needs manual provisioning of clusters
            - Takes time to configure clusters
            - Requires more resources to do so
        - Does not provide portability as
    - Data Pipeline
        - Only Amazon Web Services (AWS) platform can access and utilize Data Pipeline
        - Other frameworks might have to be used to create and set preconditions if user finds it to hard
        - Implementing Data Pipeline and other on-premise resources can be hard to configure with the system
        - Creating and setting preconditions can be lengthy and complex for newcomers
            - Other frameworks like Airflow can be used to simplify the process

- **Application**: Continuous Glucose Monitoring System
    - **Impact**: This application would provide continuous blood glucose levels by having a sensor attached to the skin
    - The data from the sensor is sent to the cloud for data processing
        - Stream processing in particular would be used for sending real time data to the user's mobile application
            - The user can view real time data on their application to see their current blood glucose levels
        - Batch processing can be used to store the data retrieved from the sensor to a data warehouse/lake
            - The data can then be viewed by the user to view their history of blood glucose levels

        - A report can also be generated based off the data from batch processing and can be sent to the user's medical doctor for assessment
- **Dataset**: The dataset can either be a JSON file or a AVRO file depending on the type of data processing software the file is suitable for
    - The data can be organized in the dataset where the data would be located under the sensor ID group, and the data can be updated to the file accordingly during every sensor reading interval
- **List of other tools**:
    - PubSub can be used for the sensor data ingestion and management
    - BigQuery can be used to query user's data located in the data warehouse
    - IoT Core can be used to maintain and diagnose sensor connection and sensor management