Enable cloud Data Fusion API from API & Services

Table

Description automatically generatedCreate data fusion instance- After that note the service account in the new instance.

Next, in IAM click Grant access and add the data fusion service account in the principal. Add Data Fusion API service agent and admin.

Graphical user interface, text, application

Description automatically generated

Create a bucket to store the data.  
Then Go to Data Fusion and click on view instance, Select wrangler (many options there). Then select GCS default storage to ingest our data.

Finally, we can use the desired plugins to create a pipeline, deploy it, and run it.

We can Import json file of previously built pipeline to the Data Fusion studio.

Note: If face failure when running, try to click on the configure button in the UI and change dataproc machines 4 to Auto scale, depending on the dataset size and processes.

Dataproc is a Data fusion runtime environment.

Table

Description automatically generated with low confidence

Graphical user interface, email

Description automatically generated

Graphical user interface, application

Description automatically generated

A picture containing diagram

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

<https://www.kaggle.com/datasets/arjunascagnetto/ptbxl-atrial-fibrillation-detection>

INPUT – uniqueID.csv, index\_dup.csv

OUTPUT – final\_pipeline.csv (downloaded few records from BigQuery table final)

Note: Regular monitoring can reduce the costs.

Table

Description automatically generated