Data Storage 🡪 Bucket 🡪 spark\_ml\_1

DataProc 🡪 Clusters

gsutil cp C:/Users/justi/OneDrive/STUDY/Python/Datproc\_using\_SparkML/natality\_sparkml.py gs://spark\_ml\_1

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

This job failed . Fixed it by executing the below command in the Google CLI

[initialization-actions/connectors at master · GoogleCloudDataproc/initialization-actions (github.com)](https://github.com/GoogleCloudDataproc/initialization-actions/tree/master/connectors)

REGION=us-central1

CLUSTER\_NAME=spark-ml-cluster1

gcloud dataproc clusters create ${CLUSTER\_NAME} \

--region ${REGION} \

--metadata GCS\_CONNECTOR\_VERSION=2.2.2

Job 2:

A screenshot of a computer

Description automatically generated

Had to terminate this job as well as there were 75 million rows. I deleted the table and reran the query and limited the dataset to just 10,000 rows. And reran the job.

A screenshot of a computer

Description automatically generated

Ran successfully!!

Deleted the cluster

Deleted the bucket which I created

A screenshot of a computer

Description automatically generated

Then deleted the table and the dataset. Project still remains.

Creating service account

A screenshot of a computer

Description automatically generated