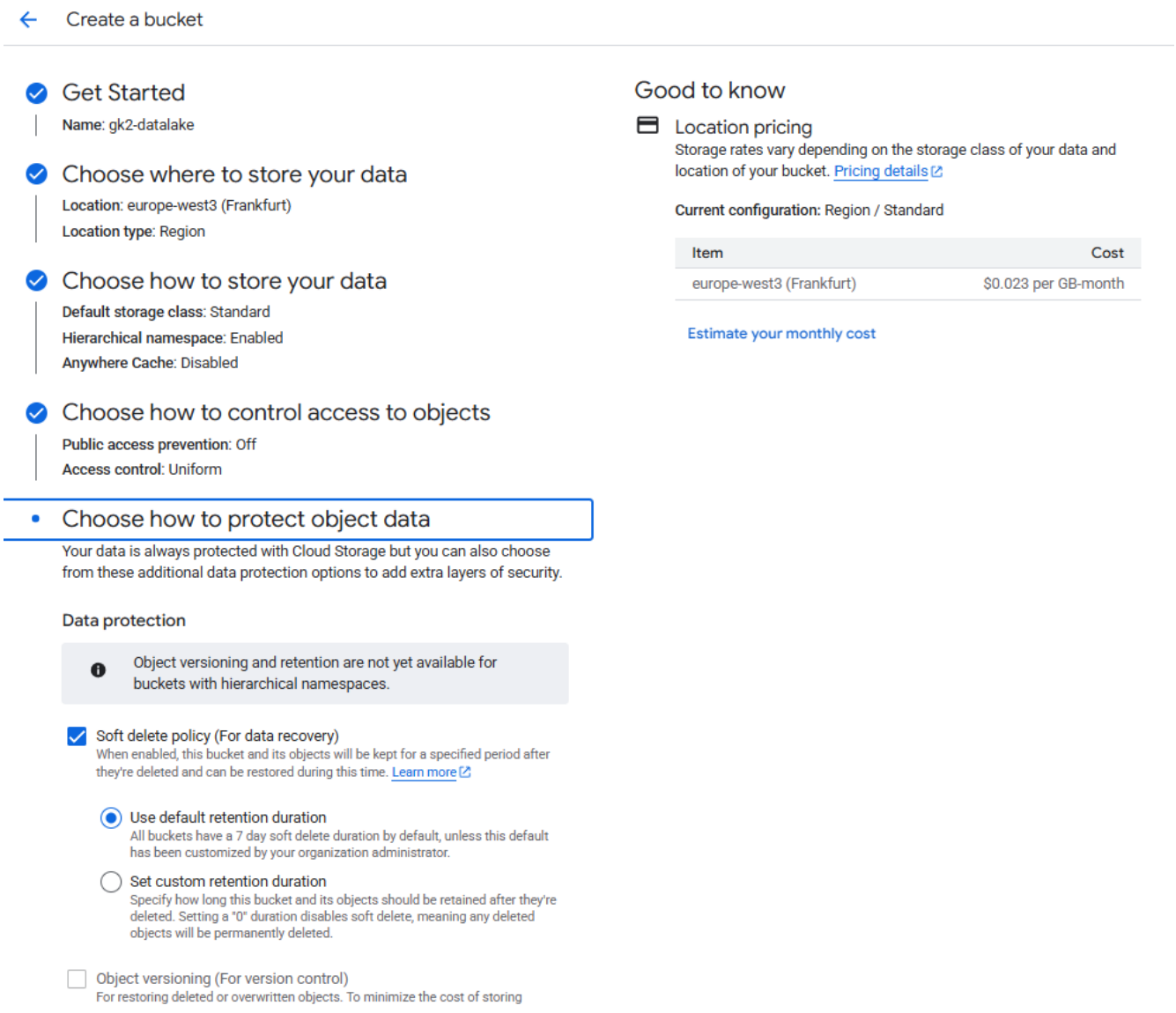
Bucket Creation

Ensure Hierarchical Namespace enabled

Ensure Public Access Prevention: off [For training only]



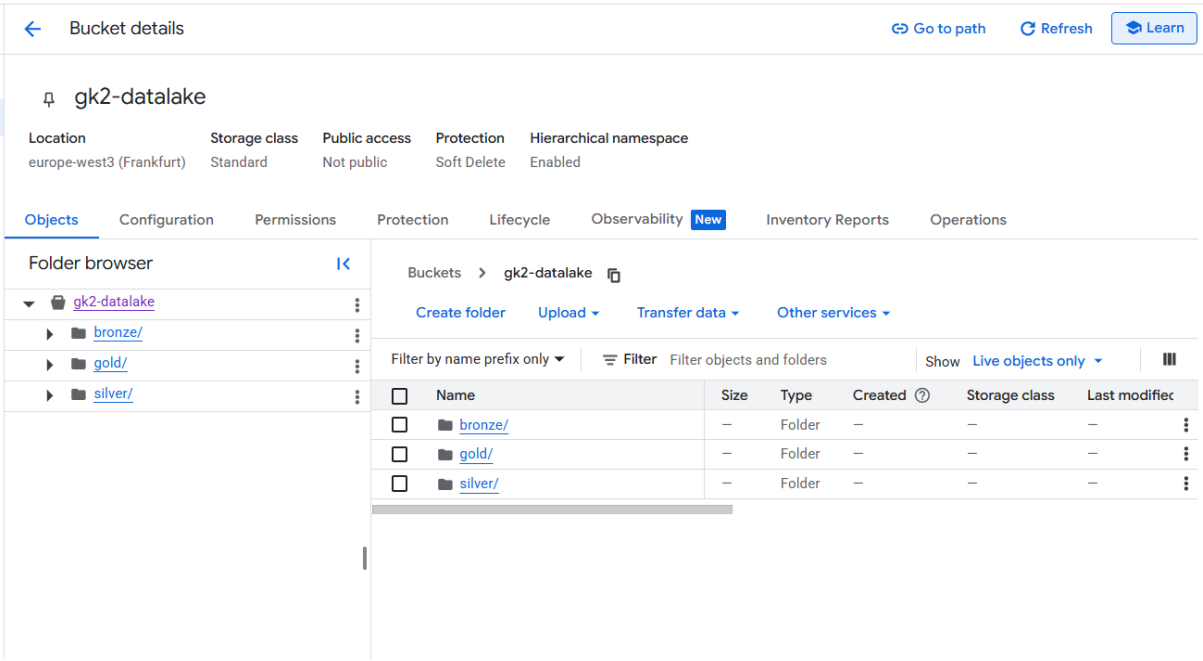
Create Data Lake Folders

Three folders named bronze, silver, gold under data lake

bronze shall have csv, json, logs, unclean, inefficient storage data

silver shall have efficient, cleaned data like parquet, delta table or iceberg format

gold will have analytical data



Download MovieLens dataset

Movielens an open source dataset for movies, ratings, tags
https://grouplens.org/datasets/movielens/

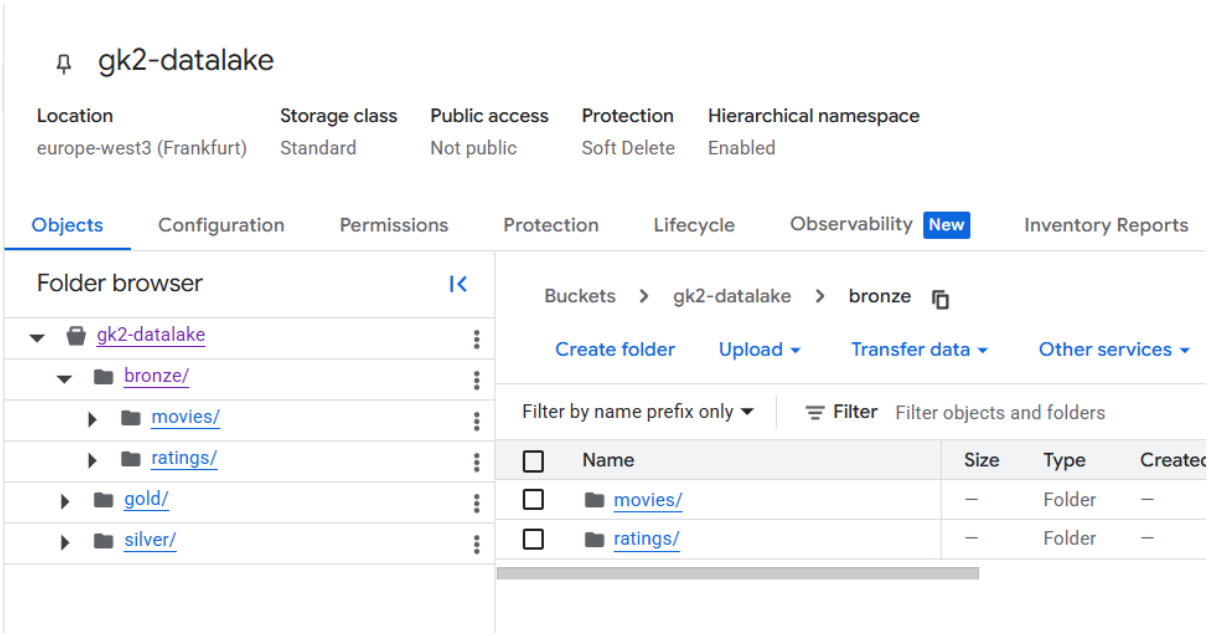
The data set to download ml-latest-small.zip

https://files.grouplens.org/datasets/movielens/ml-latest-small.zip

Extract the zip files, explore the dataset by opening them in an editor without any modification

Data Organizations

Create two directories named **movies**, **ratings**, all in lower case under **bronze** directory

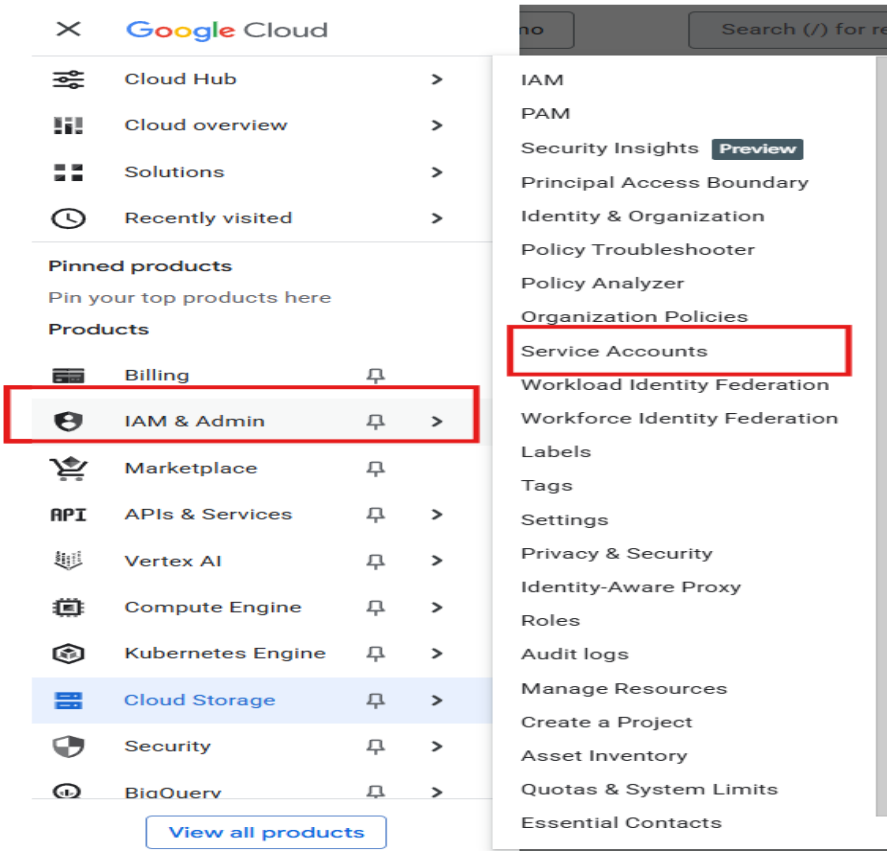


Upload movies.csv into movies directory

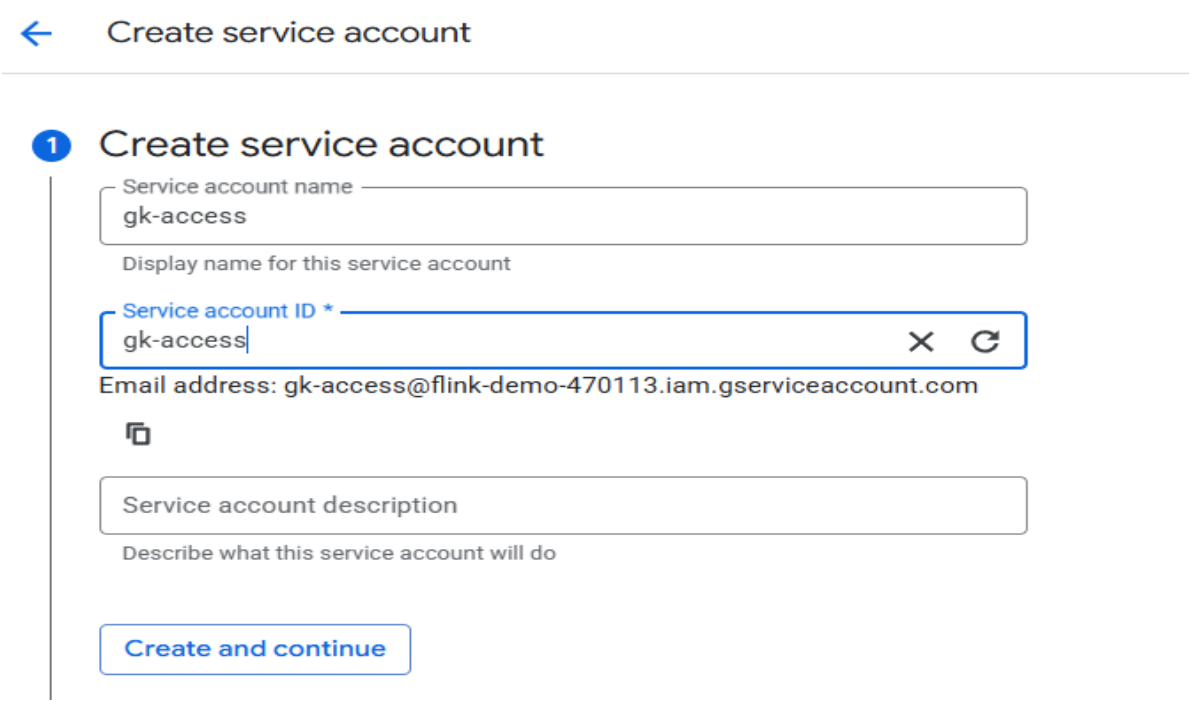
Upload ratings.csv into ratings directory

Permission to access Datalake from Flink

Go to IAM & Admin then Service account

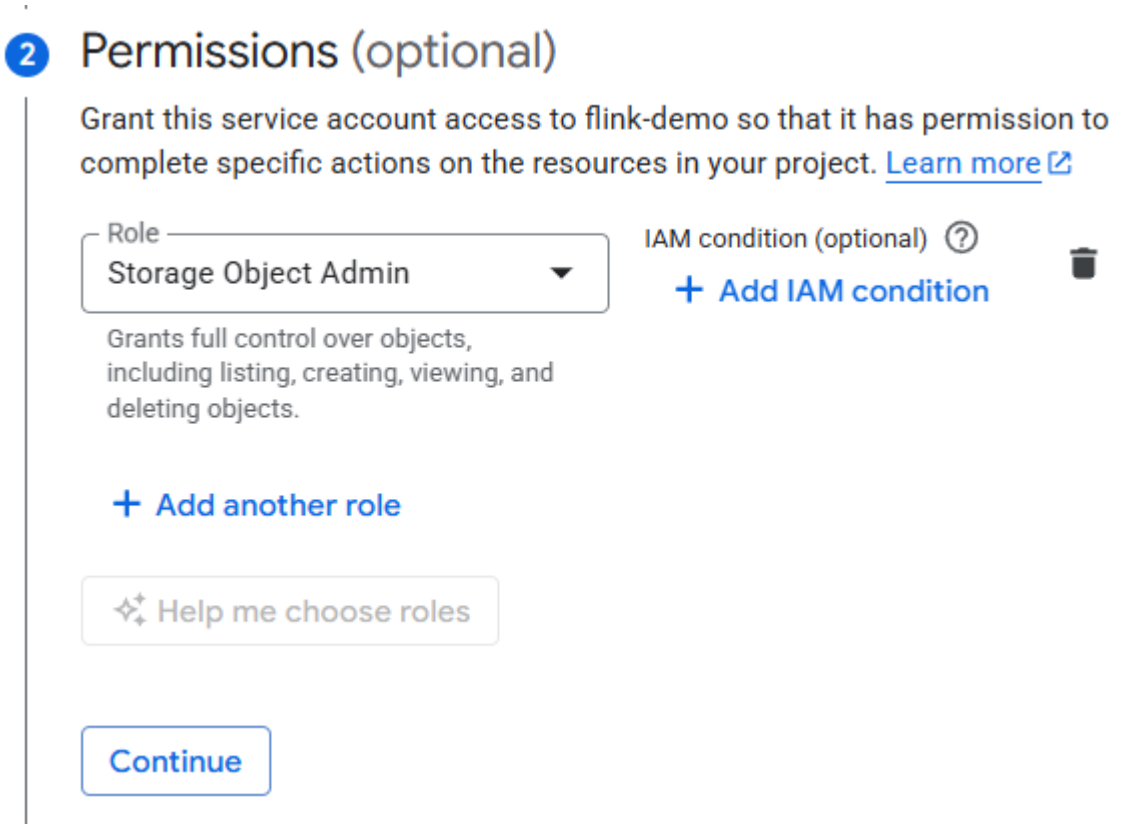


Click on Create Service Account

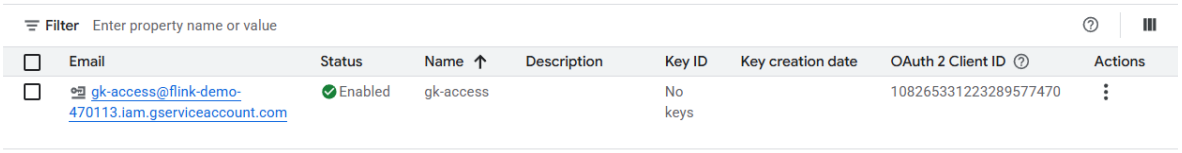


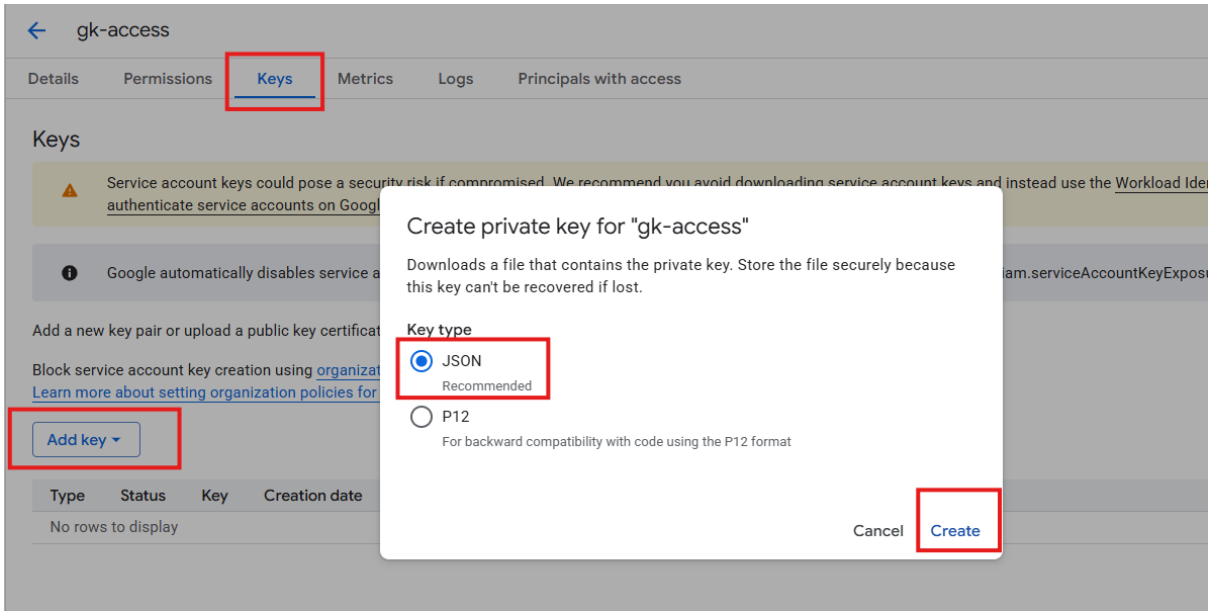
Select editor permission

On Permission, ensure to use roles/storage.objectAdmin



Click on the service account for generating key.json





This will download a json file.

Copy the downloaded json file, paste into **secret** directory of your project

Remove existing key.json if any

Ensure to rename the downloaded json file to key.json