School of IT & Business Technologies
Graduate Diploma in Data Analytics
Cover Sheet and Student Declaration

This sheet must be signed by the student and attached to the submitted assessment.

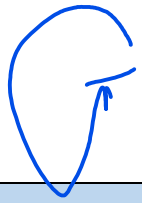
Course Title:	**Data Transformation and Management**	**Course code:**	**GDDA-612**
Student Name:	Mira Torririt	**Student ID:**	764707793
Assessment No & Type:	Assessment 2- Project-2	**Cohort:**	GDDA7123C
Due Date:	06/03/24	**Date Submitted:**	06/03/24
Tutor's Name:	Mohammad Norouzifard		
Assessment Weighting	40%		
Total Marks	100		

Student Declaration:

I declare that:

- I have read the New Zealand School of Education Ltd policies and regulations on assessments and understand what plagiarism is.
- I am aware of the penalties for cheating and plagiarism as laid down by the New Zealand School of Education Ltd.
- This is an original assessment and is entirely my own work.
- Where I have quoted or made use of the ideas of other writers, I have acknowledged the source.
- This assessment has been prepared exclusively for this course and has not been or will not be submitted as assessed work in any other course.
- It has been explained to me that this assessment may be used by NZSE Ltd, for internal and/or external moderation.
- If I am late in handing in this assessment without prior approval (see student regulations in handbook), marks will be deducted, to a maximum of 50%.

Student signature:



Date: 06/03/24

Tutor only to complete		
Assessment result:	**Mark /100**	**Grade**

Assessment 2: GDDA612 – Data Transformation and Management

Project 2: Data Migration, Integration and Backup

Mira Torririt

GDDA 612

Lecturer: Dr. Mohammad Norouzifard

School of Technology

Graduate Diploma in Data Analytics (Level 7)

March 06, 2024

Table of Contents

Chapter 1: Introduction

Data migration is the transfer or movement of data from one location to another. It is part of data management to improve data governance. Some of the uses of migration are for system upgrades, storing files in cloud environments, data consolidation, and archiving. The process includes data wrangling and conversion, which are critical in data preparation to ensure the reliability and accuracy of information. Data conversion is the process of data transformation from one format to another. This is important when moving data to an upgraded new structure. On the other hand, data integration is the merging of data from different sources, such as databases, applications, and systems.

According to The Investopedia Team (2022), the data migration is risky to the business continuity when it is not appropriately done such as the loss of data, which is considered the worst case. Other risks include downtime, compatibility issues, system performance issues, extensive data, format diversity, and different data approaches within the organization. To manage the risks, companies also emphasize the value of backups, sequential order of data transfer, and, if possible, creating an environment where the old and new systems run simultaneously during the migration period. This is to avoid downtime in the system, which is very cost-efficient in business operations.

A migration plan should be done to ensure business continuity and avoid downtime. The factors to be considered include the timeline of the migration, required downtime, and the risks that may arise to the business due to technical issues, data errors, and application performance (Gillis et al., n.d).

Just like data migration, there are also challenges in data integration, such as delays in delivering the data, security risks, resourcing constraints, data quality issues, and lack of actionability. (Gitlin, n.d). The delays in the workflow occurred when the process was done manually. The security risks involve integrating confidential data. Data integration causes resourcing constraints considering the time and cost, so it is suggested to invest in a platform offering a low-maintenance user experience rather than in-house implementation. The poor-quality data caused by various forms and formats and duplicates can be overwhelming and increase the chance of errors. Even after addressing all the first issues, there's still no assurance that the team can effectively adapt to the new system (Gitlin,n.d).

Chapter 2: Abstract

In this assessment, I will demonstrate the steps of data migration, from data preparation to backup.

The data was in CSV file when extracted from the data source. It was loaded in Python, converted to parquet, and cleaned as part of preparation in migration. A first normal form (1NF) was done in the images' column, leading to another data frame. The rationality for normalization is to separate the images inside each row to facilitate the data manipulation or query when we import it into the database.

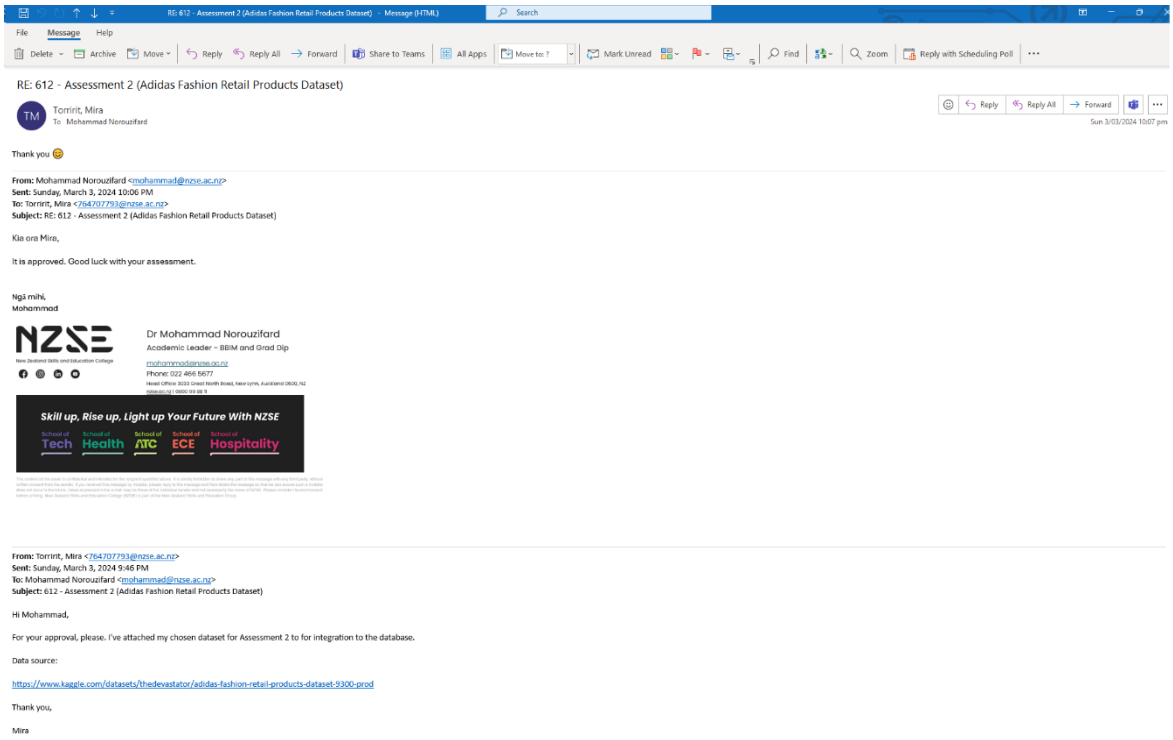
MongoDB was used due to its schemaless, flexibility, system simplicity, and scalability. Various data manipulations were done to test the success of data importation, such as data retrieval and display, sorting, record counting, grouping, and updating.

The next step was migrating the data in Google Cloud Platform and creating backup in two separate buckets inside one project. The backup was scheduled weekly to ensure the updates' continuity.

Chapter 3: Dataset

I found a dataset that matches the given scenario on the Kaggle website and had it approved by my lecturer before starting the assessment.

Below is the email approval.



RE: 612 - Assessment 2 (Adidas Fashion Retail Products Dataset) - Message (HTML)

File Message Help

Delete Archive Move Reply Reply All Forward Share to Teams All Apps Move to: ? Mark Unread Find Zoom Reply with Scheduling Poll

RE: 612 - Assessment 2 (Adidas Fashion Retail Products Dataset)

Torrirt, Mira
To Mohammad Norouzifard

Reply Reply All Forward

Sun 3/03/2024 10:07 pm

Thank you

From: Mohammad Norouzifard <mohammad@nzse.ac.nz>
Sent: Sunday, March 3, 2024 10:06 PM
To: Torrirt, Mira <764707793@nzse.ac.nz>
Subject: RE: 612 - Assessment 2 (Adidas Fashion Retail Products Dataset)

Kia ora Mira,

It is approved. Good luck with your assessment.

Ngā mihi,
Mohammad

NZSE
New Zealand Skills and Education College

Dr Mohammad Norouzifard
Academic Leader - BBIM and Grad Dip
mohammad@nzse.ac.nz
Phone: 022 466 5677

Skill up, Rise up, Light up Your Future With NZSE
School of Tech School of Health School of ATC School of ECE School of Hospitality

From: Torrirt, Mira <764707793@nzse.ac.nz>
Sent: Sunday, March 3, 2024 9:46 PM
To: Mohammad Norouzifard <mohammad@nzse.ac.nz>
Subject: 612 - Assessment 2 (Adidas Fashion Retail Products Dataset)

Hi Mohammad,

For your approval, please. I've attached my chosen dataset for Assessment 2 to for integration to the database.

Data source:

https://www.kaggle.com/datasets/thedevastator/adidas-fashion-retail-products-dataset-9300-prod

Thank you,

Mira

The dataset is about Adidas online retail. It consists of the following information: url, name (product name), sku, selling price, original price, currency, availability, color, category, source (Adidas United States), source website, breadcrumbs (navigation trail), description, brand, images (in string format), country, language, average rating, reviews count and crawled at.

Data source: https://www.kaggle.com/datasets/thedevastator/adidas-fashion-retail-products-dataset-9300-prod

Chapter 4: Task A

a) Reading the CSV file and converting it to a parquet file.

- Loading the dataset to the data frame using Python. I imported the library pandas to read the data.

841	841	https://www.adidas.com/us/choigo-shoes/H00667....	Choigo Shoes	H00667	70	$100	USD	InStock	White	Shoes	...	https://www.adidas.co
842	842	https://www.adidas.com/us/daily-3.0-shoes/GZ77...	Daily 3.0 Shoes	GZ7705	35	$50	USD	InStock	Black	Shoes	...	https://www.adidas.co
843	843	https://www.adidas.com/us/daily-3.0-shoes/GZ77...	Daily 3.0 Shoes	GZ7706	40	$50	USD	InStock	Pink	Shoes	...	https://www.adidas.co
844	844	https://www.adidas.com/us/choigo-shoes/FY6503....	Choigo Shoes	FY6503	70	$100	USD	InStock	Black	Shoes	...	https://www.adidas.co

845 rows × 21 columns

- Converting the CSV file to parquet with execution time

```
# Converting the csv file to parquet
import time

start_time = time.time()

adidas_messydf.to_parquet('adidas_messydf.parquet')
adidas_messydf

end_time = time.time()
execution_time = end_time - start_time

print(f"Execution time: {execution_time} seconds")
```

```
Execution time: 0.017109394073486328 seconds
```

- Reading the parquet file

```
# Reading the parquet file

adidas_messyparquetdf= pd.read_parquet('adidas_messydf.parquet')
adidas_messyparquetdf
```

	index	url	name	sku	selling_price	original_price	currency	availability	color	category	...	source_websit
0	0	https://www.adidas.com/us/beach-shorts/FJ5089....	Beach Shorts	FJ5089	40	None	USD	InStock	Black	Clothing	...	https://www.adidas.co
1	1	https://www.adidas.com/us/five-ten-kestrel-lac...	Five Ten Kestrel Lace Mountain Bike Shoes	BC0770	150	None	USD	InStock	Grey	Shoes	...	https://www.adidas.co
2	2	https://www.adidas.com/us/mexico-away-jersey/G...	Mexico Away Jersey	GC7946	70	None	USD	InStock	White	Clothing	...	https://www.adidas.co
3	3	https://www.adidas.com/us/five-ten-hiangle-pro...	Five Ten Hiangle Pro Competition Climbing Shoes	FV4744	160	None	USD	InStock	Black	Shoes	...	https://www.adidas.co
4	4	https://www.adidas.com/us/mesh-broken-stripe-p...	Mesh Broken-Stripe Polo Shirt	GM0239	65	None	USD	InStock	Blue	Clothing	...	https://www.adidas.co
...	...	...	...	...	...	...	...	...	...	...	...	
840	840	https://www.adidas.com/us/supernova-plus-shoes...	Supernova+ Shoes	FX2858	72	$120	USD	InStock	White	Shoes	...	https://www.adidas.co
841	841	https://www.adidas.com/us/choigo-shoes/H00667....	Choigo Shoes	H00667	70	$100	USD	InStock	White	Shoes	...	https://www.adidas.co
842	842	https://www.adidas.com/us/daily-3.0-shoes/GZ77...	Daily 3.0 Shoes	GZ7705	35	$50	USD	InStock	Black	Shoes	...	https://www.adidas.cc
843	843	https://www.adidas.com/us/daily-3.0-shoes/GZ77...	Daily 3.0 Shoes	GZ7706	40	$50	USD	InStock	Pink	Shoes	...	https://www.adidas.cc
844	844	https://www.adidas.com/us/choigo-shoes/FY6503....	Choigo Shoes	FY6503	70	$100	USD	InStock	Black	Shoes	...	https://www.adidas.cc

845 rows × 21 columns

b) Transforming the messy data to a tidy form

- Checking for duplicates

```
duplicates_all = adidas_messydf[adidas_messydf.duplicated()]
duplicates_all
```

index	url	name	sku	selling_price	original_price	currency	availability	color	category	...	source_website	breadcrumbs	description	brand	images	cou

0 rows × 21 columns

- Removing the dollar sign in the original price column to avoid redundancy due to the presence of the currency column.

```
# Removing the dollar sign ($) in original_price column
adidas_messydf['original_price'] = adidas_messydf['original_price'].str.replace("$", "")
adidas_messydf
```

	index	url	name	sku	selling_price	original_price	currency	availability	color	category	...	source_websit
0	0	https://www.adidas.com/us/beach-shorts/FJ5089....	Beach Shorts	FJ5089	40	NaN	USD	InStock	Black	Clothing	...	https://www.adidas.co
1	1	https://www.adidas.com/us/five-ten-kestrel-lac...	Five Ten Kestrel Lace Mountain Bike Shoes	BC0770	150	NaN	USD	InStock	Grey	Shoes	...	https://www.adidas.co
2	2	https://www.adidas.com/us/mexico-away-jersey/G...	Mexico Away Jersey	GC7946	70	NaN	USD	InStock	White	Clothing	...	https://www.adidas.co
3	3	https://www.adidas.com/us/five-ten-hiangle-pro...	Five Ten Hiangle Pro Competition Climbing Shoes	FV4744	160	NaN	USD	InStock	Black	Shoes	...	https://www.adidas.co
4	4	https://www.adidas.com/us/mesh-broken-stripe-p...	Mesh Broken-Stripe Polo Shirt	GM0239	65	NaN	USD	InStock	Blue	Clothing	...	https://www.adidas.co
...	...	...	...	...	...	...	...	...	...	...	...	
840	840	https://www.adidas.com/us/supernova-plus-shoes...	Supernova+ Shoes	FX2858	72	120	USD	InStock	White	Shoes	...	https://www.adidas.co
841	841	https://www.adidas.com/us/choigo-shoes/H00667....	Choigo Shoes	H00667	70	100	USD	InStock	White	Shoes	...	https://www.adidas.co
842	842	https://www.adidas.com/us/daily-3.0-shoes/GZ77...	Daily 3.0 Shoes	GZ7705	35	50	USD	InStock	Black	Shoes	...	https://www.adidas.co
843	843	https://www.adidas.com/us/daily-3.0-shoes/GZ77...	Daily 3.0 Shoes	GZ7706	40	50	USD	InStock	Pink	Shoes	...	https://www.adidas.co
844	844	https://www.adidas.com/us/choigo-shoes/FY6503....	Choigo Shoes	FY6503	70	100	USD	InStock	Black	Shoes	...	https://www.adidas.co

845 rows × 21 columns

- Identify the null values.

```
# Identifying the null values

adidas_messydf.isnull().sum()
```

```
index              0
url                0
name               0
sku                0
selling_price      0
original_price    16
currency           0
availability       0
color              0
category           0
source             0
source_website     0
breadcrumbs        0
description        0
brand              0
images             0
country            0
language           0
average_rating     0
reviews_count      0
crawled_at         0
dtype: int64
```

- Replacing the null values with an additional 22.30% markup price based on the products' average markup value. Using the library math, it was rounded up to its nearest value for the consistency of the format.

```
import math
adidas_messydf['original_price'] = adidas_messydf['original_price'].astype(float).apply(math.ceil)
adidas_messydf
```

	index	url	name	sku	selling_price	original_price	currency	availability	color	category	...	source_websi
0	0	https://www.adidas.com/us/beach-shorts/FJ5089....	Beach Shorts	FJ5089	40	49	USD	InStock	Black	Clothing	...	https://www.adidas.co
1	1	https://www.adidas.com/us/five-ten-kestrel-lac...	Five Ten Kestrel Lace Mountain Bike Shoes	BC0770	150	184	USD	InStock	Grey	Shoes	...	https://www.adidas.co
2	2	https://www.adidas.com/us/mexico-away-jersey/G...	Mexico Away Jersey	GC7946	70	86	USD	InStock	White	Clothing	...	https://www.adidas.co
3	3	https://www.adidas.com/us/five-ten-hiangle-pro...	Five Ten Hiangle Pro Competition Climbing Shoes	FV4744	160	196	USD	InStock	Black	Shoes	...	https://www.adidas.co
4	4	https://www.adidas.com/us/mesh-broken-stripe-p...	Mesh Broken-Stripe Polo Shirt	GM0239	65	80	USD	InStock	Blue	Clothing	...	https://www.adidas.co
...	...	...	...	...	...	...	...	...	...	...	...	
840	840	https://www.adidas.com/us/supernova-plus-shoes...	Supernova+ Shoes	FX2858	72	120	USD	InStock	White	Shoes	...	https://www.adidas.co
841	841	https://www.adidas.com/us/choigo-shoes/H00667....	Choigo Shoes	H00667	70	100	USD	InStock	White	Shoes	...	https://www.adidas.co
842	842	https://www.adidas.com/us/daily-3.0-shoes/GZ77...	Daily 3.0 Shoes	GZ7705	35	50	USD	InStock	Black	Shoes	...	https://www.adidas.co

- Checking to see if the null values have been successfully managed.

```
# Checking to see if the null values have been successfully managed.

adidas_messydf.isnull().sum()
```

```
index             0
url               0
name              0
sku               0
selling_price     0
original_price    0
currency          0
availability      0
color             0
category          0
source            0
source_website    0
breadcrumbs       0
description       0
brand             0
images            0
country           0
language          0
average_rating    0
reviews_count     0
dtype: int64
```

- Dropping the 'crawled at' column as it has no value in the analysis.

```
#dropping crawled at column

adidas_messydf.drop('crawled_at', axis=1, inplace=True)
adidas_messydf
```

	index	url	name	sku	selling_price	original_price	currency	availability	color	category	source	source_v
0	0	https://www.adidas.com/us/beach-shorts/FJ5089....	Beach Shorts	FJ5089	40	49	USD	InStock	Black	Clothing	adidas United States	https://www.adid
1	1	https://www.adidas.com/us/five-ten-kestrel-lac...	Five Ten Kestrel Lace Mountain Bike Shoes	BC0770	150	184	USD	InStock	Grey	Shoes	adidas United States	https://www.adid
2	2	https://www.adidas.com/us/mexico-away-jersey/G...	Mexico Away Jersey	GC7946	70	86	USD	InStock	White	Clothing	adidas United States	https://www.adid
3	3	https://www.adidas.com/us/five-ten-hiangle-pro...	Five Ten Hiangle Pro Competition Climbing Shoes	FV4744	160	196	USD	InStock	Black	Shoes	adidas United States	https://www.adid
4	4	https://www.adidas.com/us/mesh-broken-stripe-p...	Mesh Broken-Stripe Polo Shirt	GM0239	65	80	USD	InStock	Blue	Clothing	adidas United States	https://www.adid

- Checking to see if the dropping of the column is successfully done.

```
# checking
adidas_messydf.columns
```

```
Index(['index', 'url', 'name', 'sku', 'selling_price', 'original_price',
       'currency', 'availability', 'color', 'category', 'source',
       'source_website', 'breadcrumbs', 'description', 'brand', 'images',
       'country', 'language', 'average_rating', 'reviews_count'],
      dtype='object')
```

- Applying normalization (1NF) on the images' column, as it contains multiple values. The values were split into a new data frame called 'adidas_image_df' to avoid duplication. The adidas_image_df data frame was linked to the adidas_messydf data frame via the sku column to establish a relationship. Explode function will create multiple rows from the array of images containing a single value while maintaining the details from the rest of the columns. The image column in adidas_messydf was dropped as the adidas_image_df data frame will hold all the images.

```
# Normalization of the images column

adidas_messydf['images'] = adidas_messydf['images'].str.split('~')
# Create new dataframe for images after calling explode function.
# Explode function will create multiple rows from the array of images, each containing a single value while maintaining the deta
adidas_image_df = adidas_messydf.explode('images')
adidas_image_df = adidas_image_df[['sku','images']]
adidas_tidydf = adidas_messydf.drop(columns=['images'], axis=1) # Will drop images column as the adidas_image_df dataframe will h
adidas_image_df
```

	sku	images
0	FJ5089	https://assets.adidas.com/images/w_600,f_auto,...
0	FJ5089	https://assets.adidas.com/images/w_600,f_auto,...
0	FJ5089	https://assets.adidas.com/images/w_600,f_auto,...
0	FJ5089	https://assets.adidas.com/images/w_600,f_auto,...
0	FJ5089	https://assets.adidas.com/images/w_600,f_auto,...
...	...	...
844	FY6503	https://assets.adidas.com/images/w_600,f_auto,...
844	FY6503	https://assets.adidas.com/images/w_600,f_auto,...
844	FY6503	https://assets.adidas.com/images/w_600,f_auto,...
844	FY6503	https://assets.adidas.com/images/w_600,f_auto,...
844	FY6503	https://assets.adidas.com/images/w_600,f_auto,...

6466 rows × 2 columns

- Merging the two data frames via left join to show that they are linked by sku column.

```
# Shows dataframe containing the images is linked to the original dataframe by its sku
merge_df = pd.merge(adidas_image_df, adidas_tidydf, on="sku", how="left")
merge_df[['sku', 'images', 'name']]
```

	sku	images	name
0	FJ5089	https://assets.adidas.com/images/w_600,f_auto,...	Beach Shorts
1	FJ5089	https://assets.adidas.com/images/w_600,f_auto,...	Beach Shorts
2	FJ5089	https://assets.adidas.com/images/w_600,f_auto,...	Beach Shorts
3	FJ5089	https://assets.adidas.com/images/w_600,f_auto,...	Beach Shorts
4	FJ5089	https://assets.adidas.com/images/w_600,f_auto,...	Beach Shorts
...	...	...	...
6461	FY6503	https://assets.adidas.com/images/w_600,f_auto,...	Choigo Shoes
6462	FY6503	https://assets.adidas.com/images/w_600,f_auto,...	Choigo Shoes
6463	FY6503	https://assets.adidas.com/images/w_600,f_auto,...	Choigo Shoes
6464	FY6503	https://assets.adidas.com/images/w_600,f_auto,...	Choigo Shoes
6465	FY6503	https://assets.adidas.com/images/w_600,f_auto,...	Choigo Shoes

6466 rows × 3 columns

- Showing that the images column has been removed.

```
# Showing the images column is already removed
adidas_tidydf.columns
```

```
Index(['index', 'url', 'name', 'sku', 'selling_price', 'original_price',
       'currency', 'availability', 'color', 'category', 'source',
       'source_website', 'breadcrumbs', 'description', 'brand', 'country',
       'language', 'average_rating', 'reviews_count'],
      dtype='object')
```

c. Displaying the initial rows of the tidy dataset using function df.head()

```
# c. Displaying the initial ros of the tidy dataset

adidas_tidydf.head (10)
```

	index	url	name	sku	selling_price	original_price	currency	availability	color	category	source	source_webs
0	0	https://www.adidas.com/us/beach-shorts/FJ5089....	Beach Shorts	FJ5089	40	49	USD	InStock	Black	Clothing	adidas United States	https://www.adidas.c
1	1	https://www.adidas.com/us/five-ten-kestrel-lac...	Five Ten Kestrel Lace Mountain Bike Shoes	BC0770	150	184	USD	InStock	Grey	Shoes	adidas United States	https://www.adidas.c
2	2	https://www.adidas.com/us/mexico-away-jersey/G...	Mexico Away Jersey	GC7946	70	86	USD	InStock	White	Clothing	adidas United States	https://www.adidas.c
3	3	https://www.adidas.com/us/five-ten-hiangle-pro...	Five Ten Hiangle Pro Competition Climbing Shoes	FV4744	160	196	USD	InStock	Black	Shoes	adidas United States	https://www.adidas.c
4	4	https://www.adidas.com/us/mesh-broken-stripe-p...	Mesh Broken-Stripe Polo Shirt	GM0239	65	80	USD	InStock	Blue	Clothing	adidas United States	https://www.adidas.c
5	5	https://www.adidas.com/us/eqt-spikeless-golf-s...	EQT Spikeless Golf Shoes	FX7449	110	135	USD	InStock	Grey	Shoes	adidas United States	https://www.adidas.c
6	6	https://www.adidas.com/us/adicross-hybrid-shor...	Adicross Hybrid Shorts	GM5505	80	98	USD	InStock	Black	Clothing	adidas United States	https://www.adidas.c
7	7	https://www.adidas.com/us/tiro-21-windbreaker/...	Tiro 21 Windbreaker	GP4975	60	74	USD	InStock	Black	Clothing	adidas United States	https://www.adidas.c

d) Filtering the clothing column

```
# d) Filtering the clothing column
clothing_df = adidas_tidydf[adidas_tidydf['category'] == 'Clothing']
clothing_df
```

	url	name	sku	selling_price	original_price	currency	availability	color	category	source	source_website
0	https://www.adidas.com/us/beach-shorts/FJ5089....	Beach Shorts	FJ5089	40	49	USD	InStock	Black	Clothing	adidas United States	https://www.adidas.com
2	https://www.adidas.com/us/mexico-away-jersey/G...	Mexico Away Jersey	GC7946	70	86	USD	InStock	White	Clothing	adidas United States	https://www.adidas.com
4	https://www.adidas.com/us/mesh-broken-stripe-p...	Mesh Broken-Stripe Polo Shirt	GM0239	65	80	USD	InStock	Blue	Clothing	adidas United States	https://www.adidas.com
6	https://www.adidas.com/us/adicross-hybrid-shor...	Adicross Hybrid Shorts	GM5505	80	98	USD	InStock	Black	Clothing	adidas United States	https://www.adidas.com
7	https://www.adidas.com/us/tiro-21-windbreaker/...	Tiro 21 Windbreaker	GP4975	60	74	USD	InStock	Black	Clothing	adidas United States	https://www.adidas.com
...	...	...	...	...	...	...	...	...	...	...	...
827	https://www.adidas.com/us/fast-primeblue-tee/H...	Fast Primeblue Tee	H32236	32	40	USD	InStock	White	Clothing	adidas United States	https://www.adidas.com

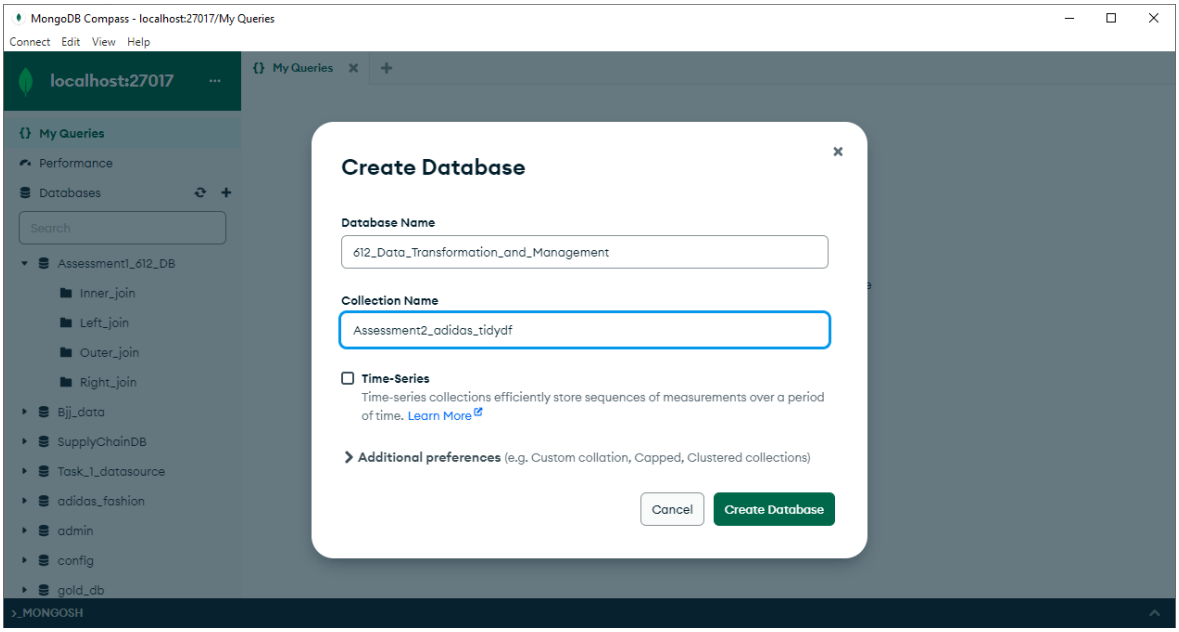
828	https://www.adidas.com/us/fast-primeblue-tee/H...	Fast Primeblue Tee	H11276	32	40	USD	InStock	Green	Clothing	adidas United States	https://www.adidas.com
829	https://www.adidas.com/us/short-piping-high-wa...	Short Piping High-Waist Tights	H17923	28	40	USD	InStock	Black	Clothing	adidas United States	https://www.adidas.com
830	https://www.adidas.com/us/adicolor-essentials-...	Adicolor Essentials Trefoil Swim Shorts	H35501	28	35	USD	InStock	Yellow	Clothing	adidas United States	https://www.adidas.com
831	https://www.adidas.com/us/marble-logo-graphic-...	Marble Logo Graphic Print Hoodie	H22628	32	45	USD	InStock	White	Clothing	adidas United States	https://www.adidas.com

337 rows × 18 columns

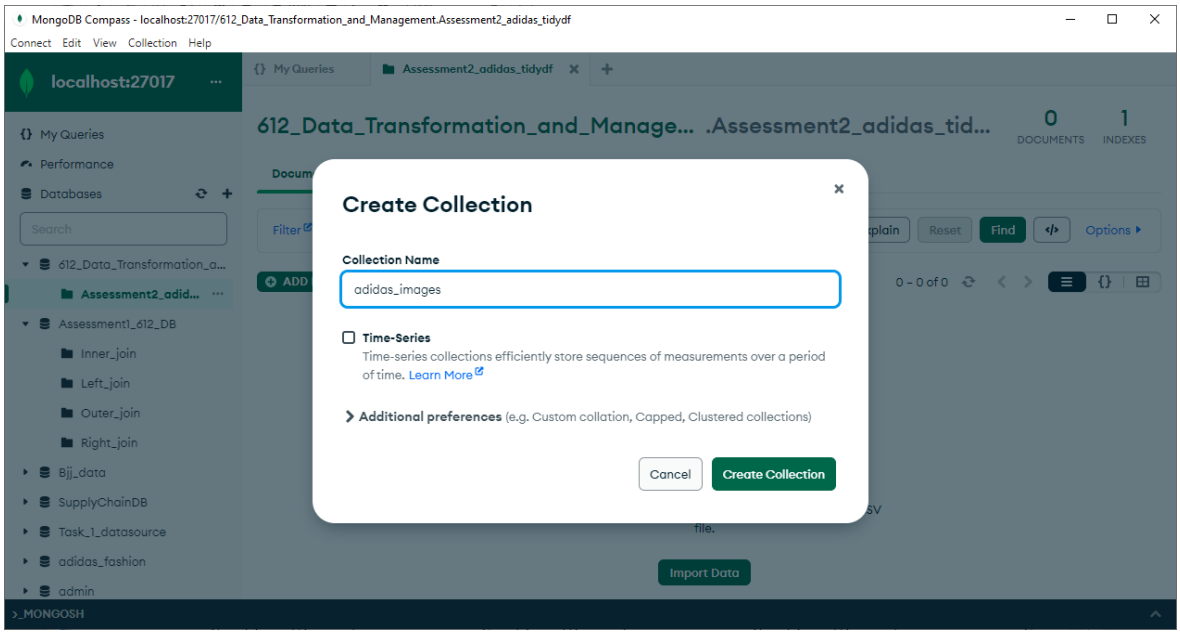
e) Creating the database using MongoDB (NoSQL)

1. Connect to MongoDB and create a database name and collections. Click Databases and create a database. For this assessment, I created a database named 612 Data Transformation and Management and two collections separately. The first collection named Assessment2_adidas_tidydf was done simultaneously with the database. The second collection was created by clicking the + sign (create collection button) beside the database name (612_ Data_ Transformation_ and_ Management), I named it adidas_images.

1.1 Assessment2_adidas_tidydf collection



1.2 adidas_images collection

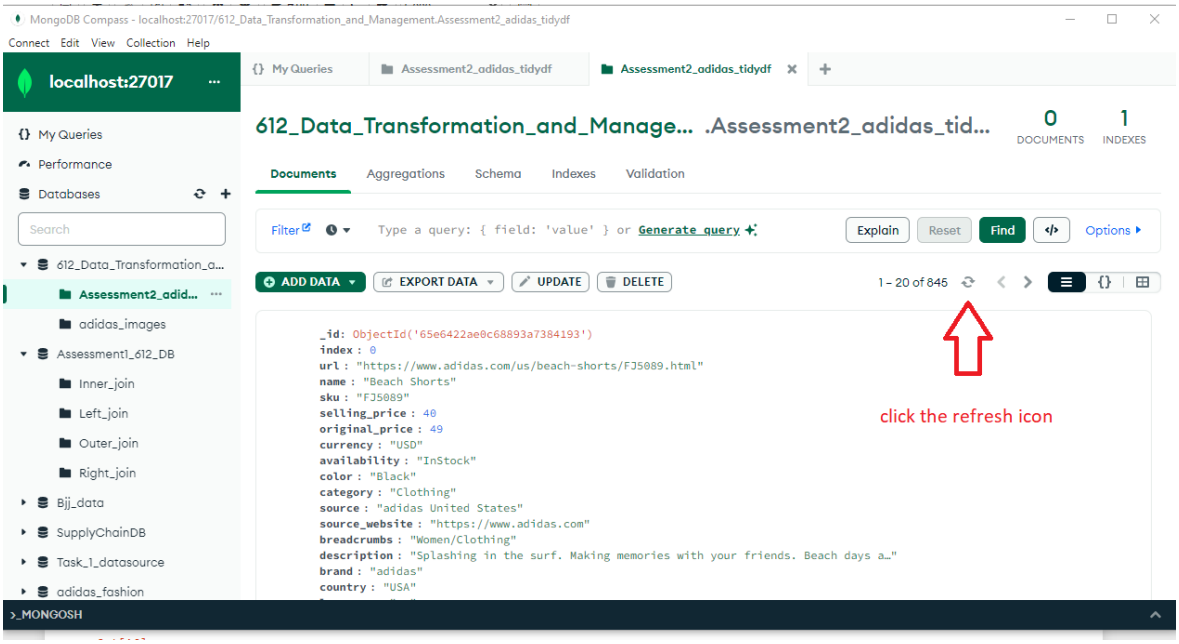


2. Create a python code to import the data to MongoDB.

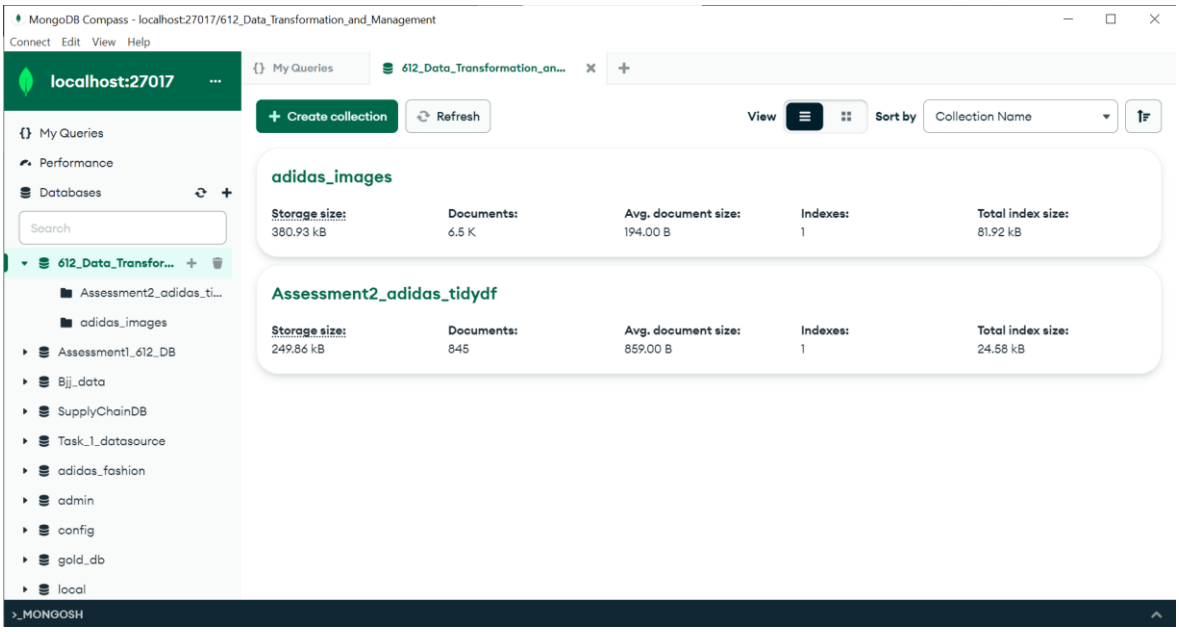
```
# f) Importing the dataset in the database
adidas_records = adidas_tidydf.to_dict(orient='records')
db = client['612_Data_Transformation_and_Management']
collection = db.Assessment2_adidas_tidydf
collection.insert_many(adidas_records)

adidas_images = adidas_image_df.to_dict(orient='records')
db = client['612_Data_Transformation_and_Management']
imgcollection = db.adidas_images
imgcollection.insert_many(adidas_images)
```

```
InsertManyResult([ObjectId('65e6422ae0c68893a73844e0'), ObjectId('65e6422ae0c68893a73844e1'), ObjectId('65e6422ae0c68893a7384
4e2'), ObjectId('65e6422ae0c68893a73844e3'), ObjectId('65e6422ae0c68893a73844e4'), ObjectId('65e6422ae0c68893a73844e5'), Obje
ctId('65e6422ae0c68893a73844e6'), ObjectId('65e6422ae0c68893a73844e7'), ObjectId('65e6422ae0c68893a73844e8'), ObjectId('65e64
22ae0c68893a73844e9'), ObjectId('65e6422ae0c68893a73844ea'), ObjectId('65e6422ae0c68893a73844eb'), ObjectId('65e6422ae0c68893
a73844ec'), ObjectId('65e6422ae0c68893a73844ed'), ObjectId('65e6422ae0c68893a73844ee'), ObjectId('65e6422ae0c68893a73844ef'),
ObjectId('65e6422ae0c68893a73844f0'), ObjectId('65e6422ae0c68893a73844f1'), ObjectId('65e6422ae0c68893a73844f2'), ObjectId('6
5e6422ae0c68893a73844f3'), ObjectId('65e6422ae0c68893a73844f4'), ObjectId('65e6422ae0c68893a73844f5'), ObjectId('65e6422ae0c6
8893a73844f6'), ObjectId('65e6422ae0c68893a73844f7'), ObjectId('65e6422ae0c68893a73844f8'), ObjectId('65e6422ae0c68893a73844f
9'), ObjectId('65e6422ae0c68893a73844fa'), ObjectId('65e6422ae0c68893a73844fb'), ObjectId('65e6422ae0c68893a73844fc'), Object
Id('65e6422ae0c68893a73844fd'), ObjectId('65e6422ae0c68893a73844fe'), ObjectId('65e6422ae0c68893a73844ff'), ObjectId('65e6422
ae0c68893a7384500'), ObjectId('65e6422ae0c68893a7384501'), ObjectId('65e6422ae0c68893a7384502'), ObjectId('65e6422ae0c68893a7
384503'), ObjectId('65e6422ae0c68893a7384504'), ObjectId('65e6422ae0c68893a7384505'), ObjectId('65e6422ae0c68893a7384506'), O
bjectId('65e6422ae0c68893a7384507'), ObjectId('65e6422ae0c68893a7384508'), ObjectId('65e6422ae0c68893a7384509'), ObjectId('65
e6422ae0c68893a738450a'), ObjectId('65e6422ae0c68893a738450b'), ObjectId('65e6422ae0c68893a738450c'), ObjectId('65e6422ae0c68
893a738450d'), ObjectId('65e6422ae0c68893a738450e'), ObjectId('65e6422ae0c68893a738450f'), ObjectId('65e6422ae0c68893a738451
0'), ObjectId('65e6422ae0c68893a7384511'), ObjectId('65e6422ae0c68893a7384512'), ObjectId('65e6422ae0c68893a7384513'), Object
Id('65e6422ae0c68893a7384514'), ObjectId('65e6422ae0c68893a7384515'), ObjectId('65e6422ae0c68893a7384516'), ObjectId('65e6422
ae0c68893a7384517'), ObjectId('65e6422ae0c68893a7384518'), ObjectId('65e6422ae0c68893a7384519'), ObjectId('65e6422ae0c68893a7
38451a'), ObjectId('65e6422ae0c68893a738451b'), ObjectId('65e6422ae0c68893a738451c'), ObjectId('65e6422ae0c68893a738451d'), O
```

Note: Click refresh to show the latest data, because MongoDB does not show data in real time.



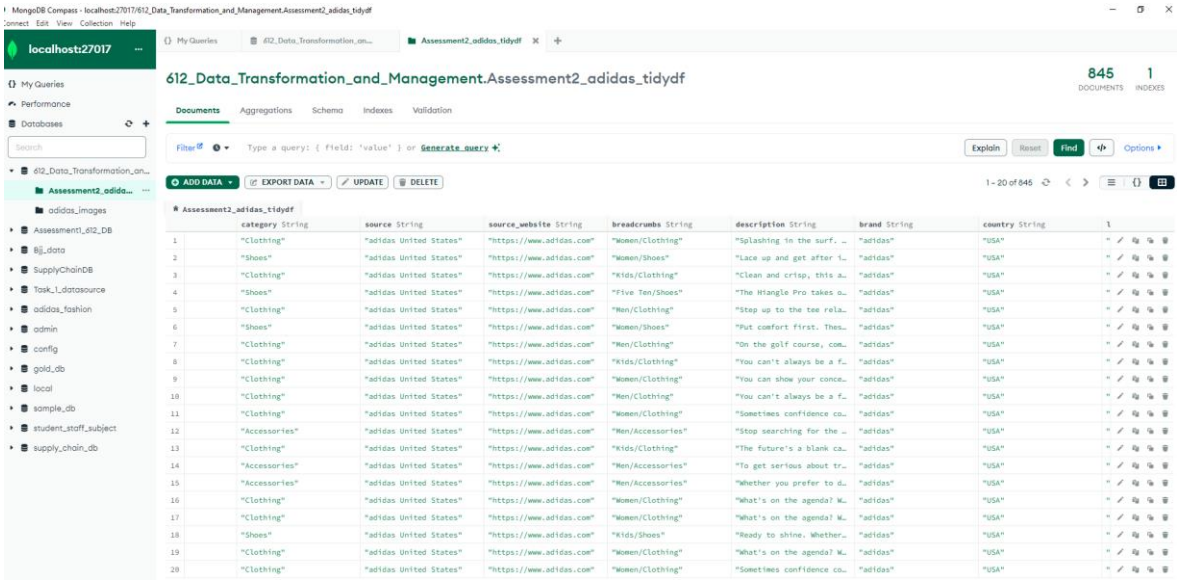
Created database and columns.



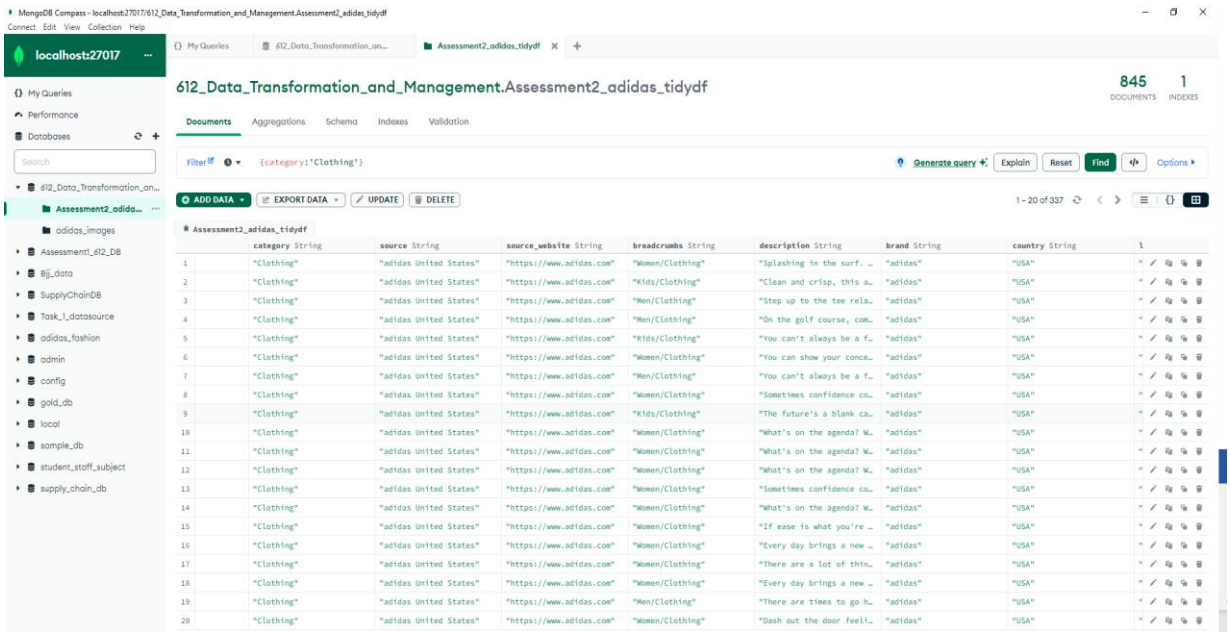
g) Retrieving and displaying documents from the collection in MongoDB.

1. From Assessment2_adidas_tidydf collection, in the **filter** box, write the query in this format: {field: 'value'}. The code is {category: 'Clothing'}, then click **find.**

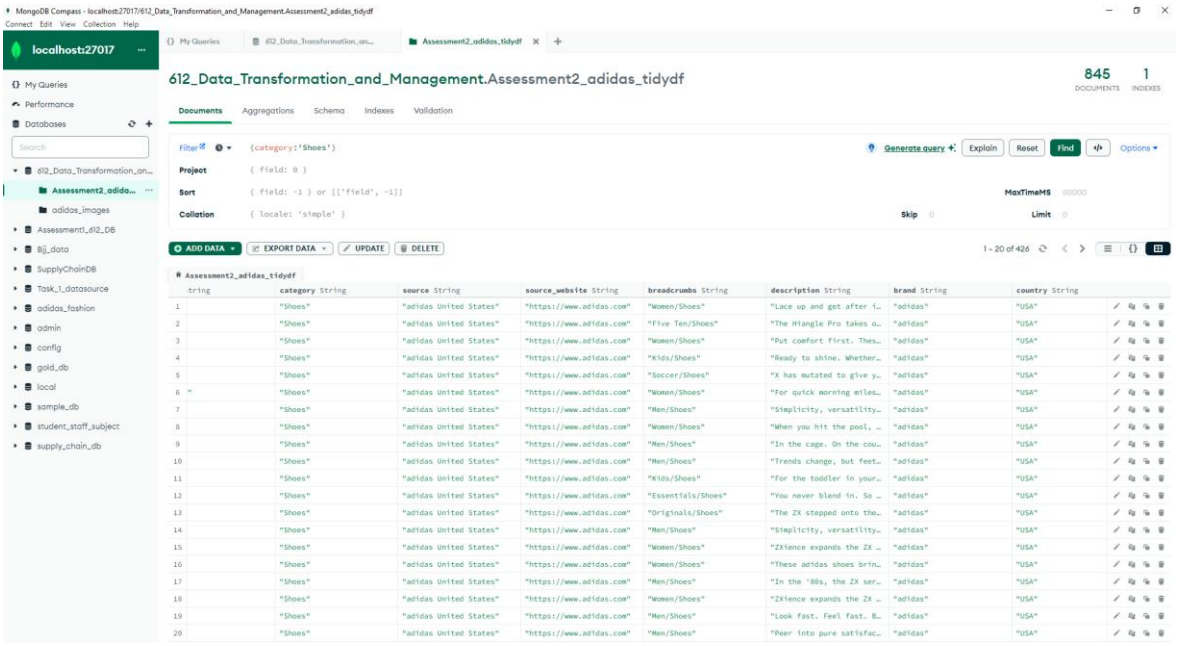
* Before



*After



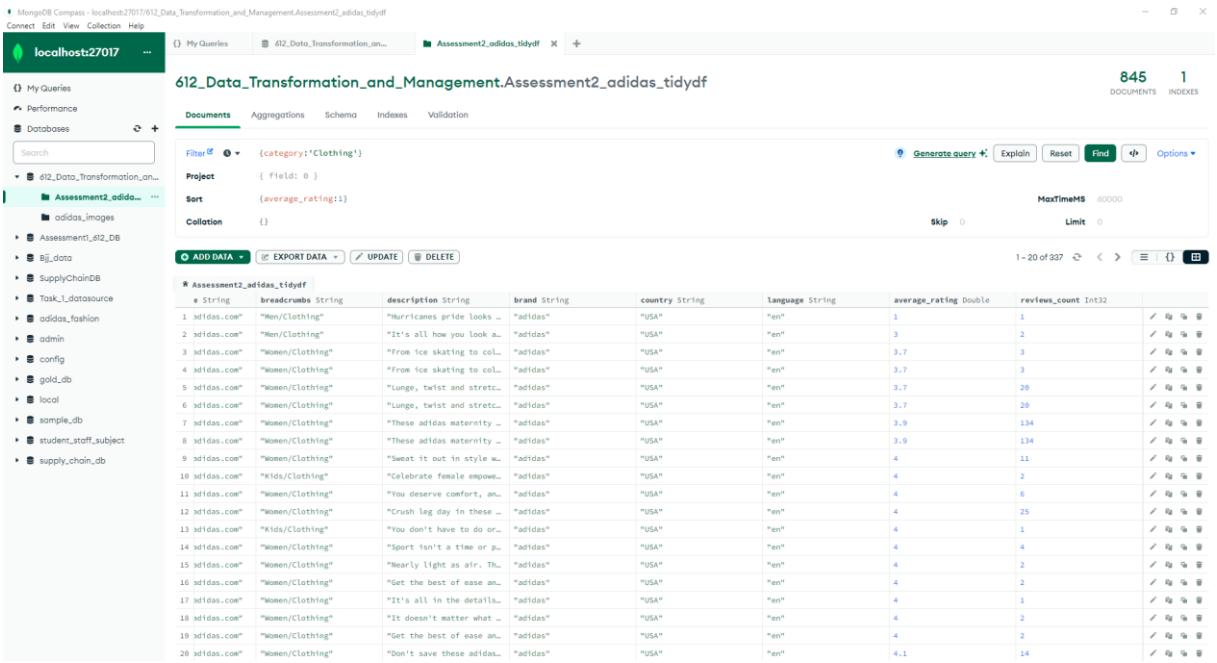
- Another sample



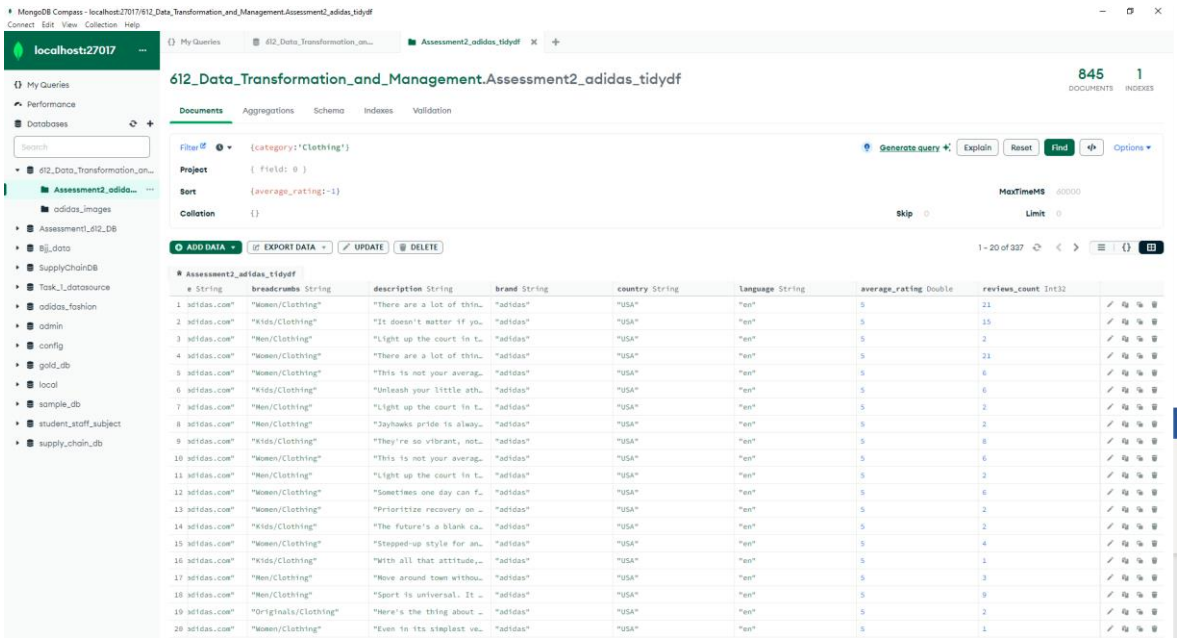
h) Sorting the documents based on a given condition.

* Sorting the clothing category by average rating in ascending (1) and descending (-1) to check which one needs improvement, and which ones satisfy the customers. From the clothing category, click **the option function, select sort,** and type your query. The query is { average_rating:1}, then click Find.

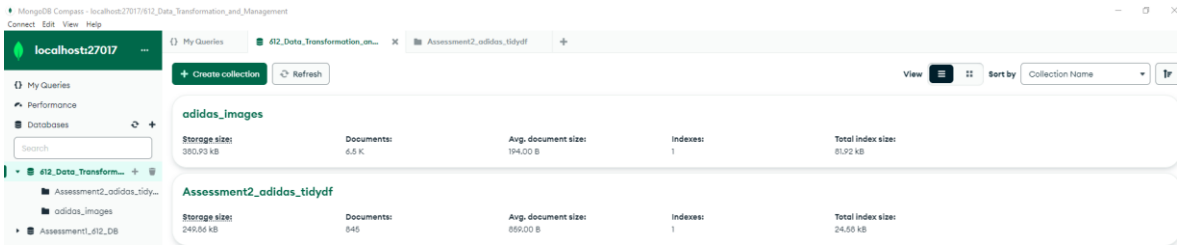
- Ascending



- Descending

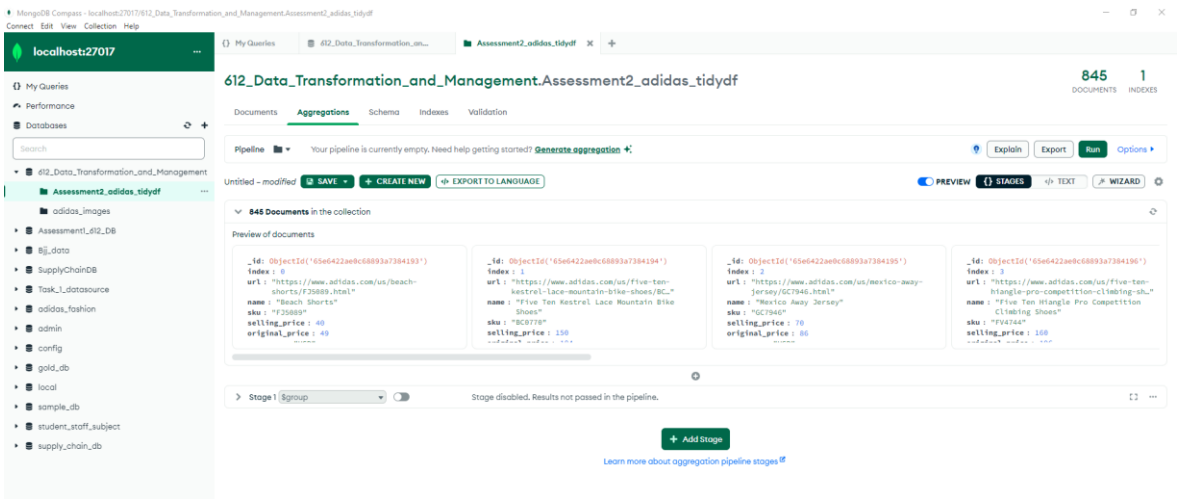


i) You can see the count of documents in the collection by checking it in the database.

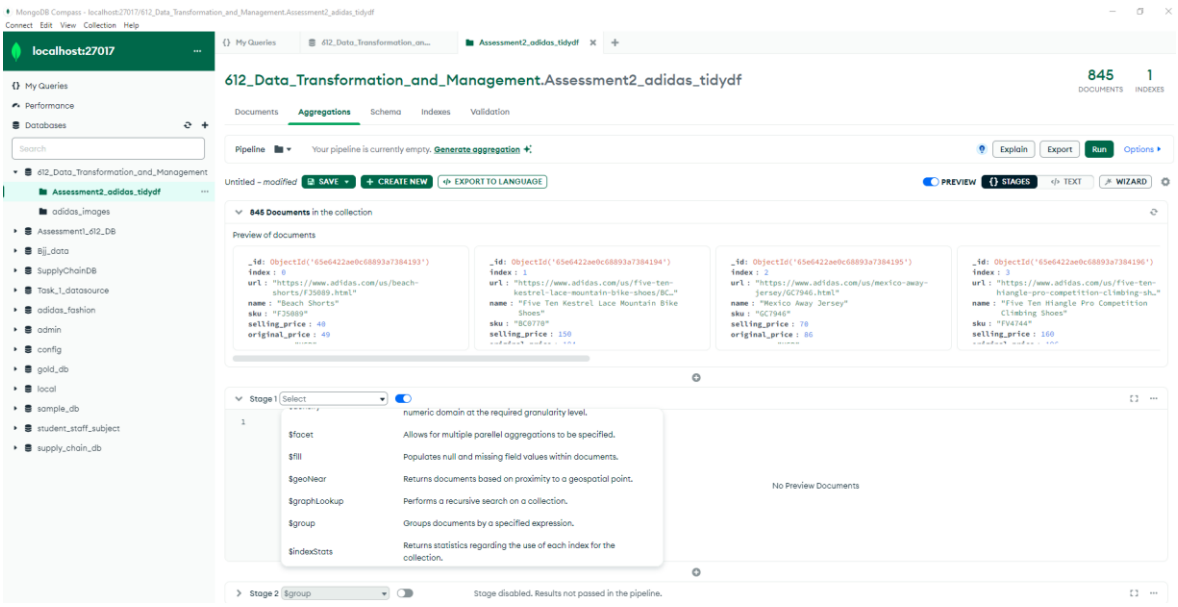


j) Performing grouping operation.

1. Click the name of the collection; Assessment2_adidas_tidydf

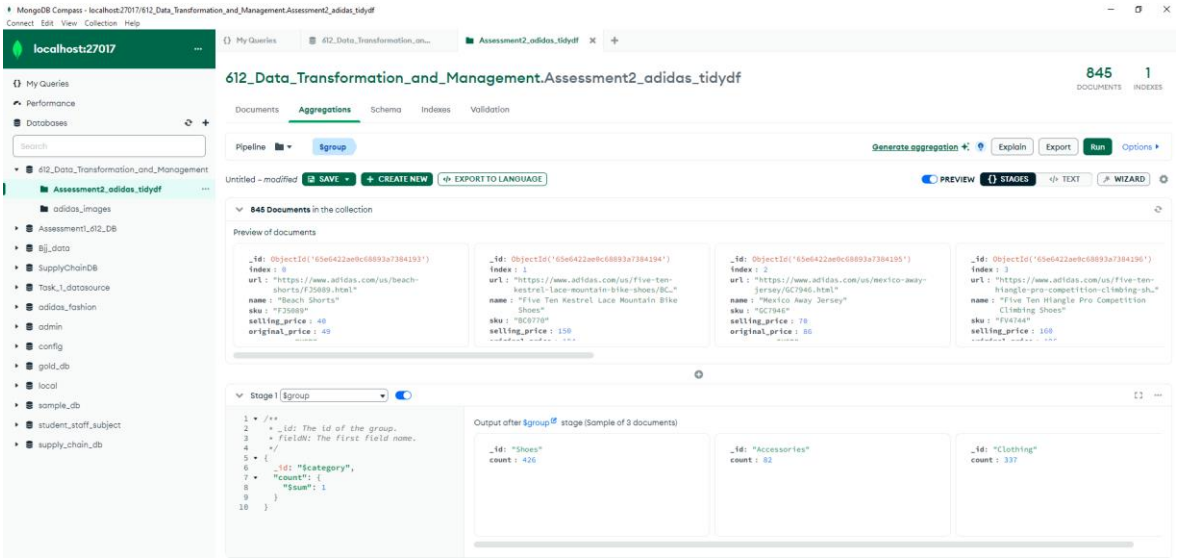


2. In the Aggregations Tab, click + Add Stage and select $group to aggregate the collection. Make sure to enable the Stage (using this button:  to pass the result in the pipeline.



Type the query in Stage 1 box using the syntax below.

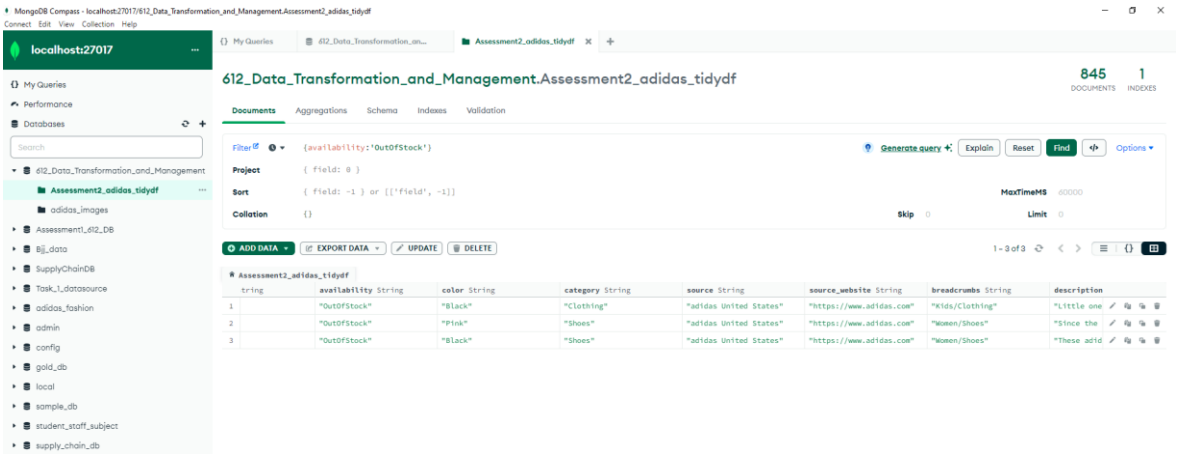
```
_id: "$category",
  "count": {
    "$sum": 1
}
```



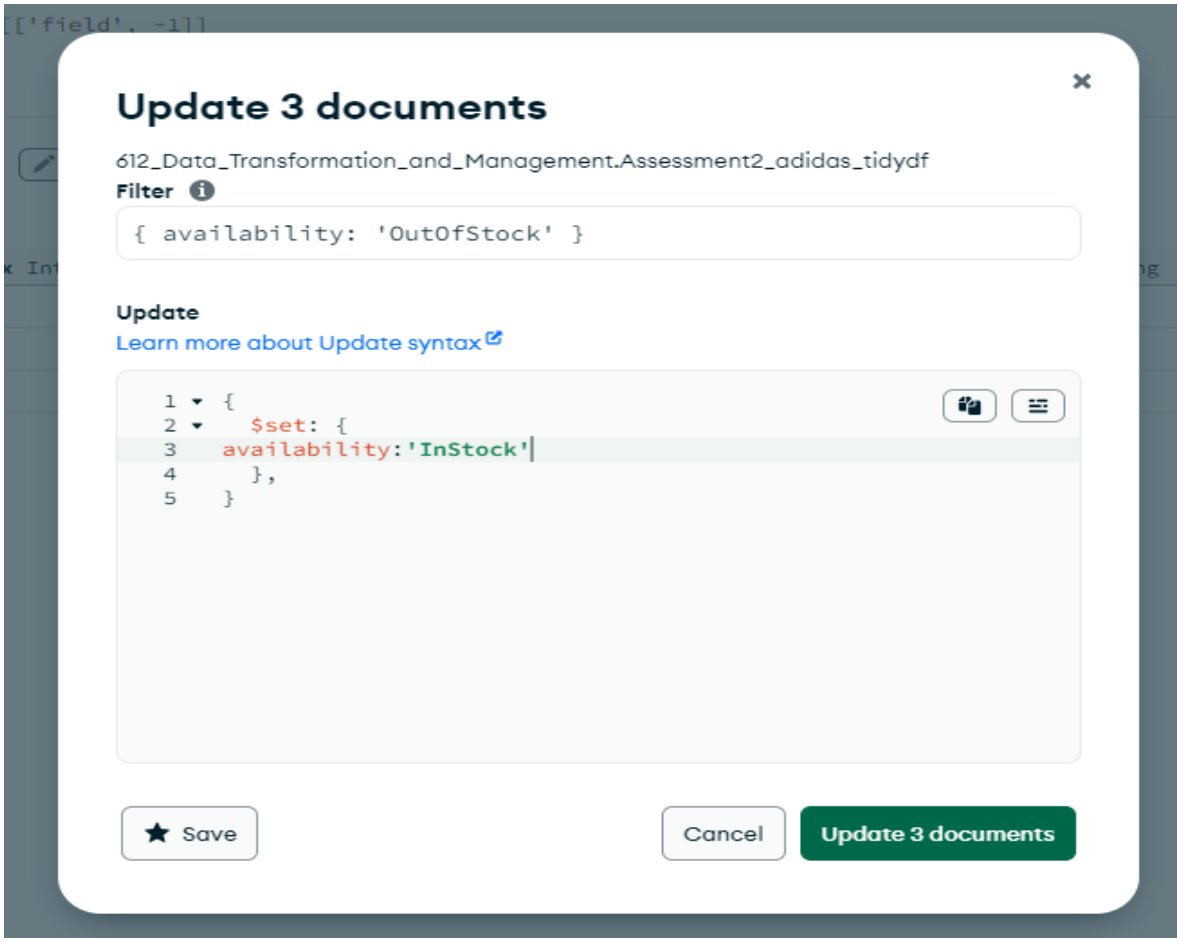
The data was grouped according to category and it counts the number of documents.

k) To update the document in MongoDB, go to the collection (Assessment2_adidas_tidydf), click on Documents, then click on UPDATE.

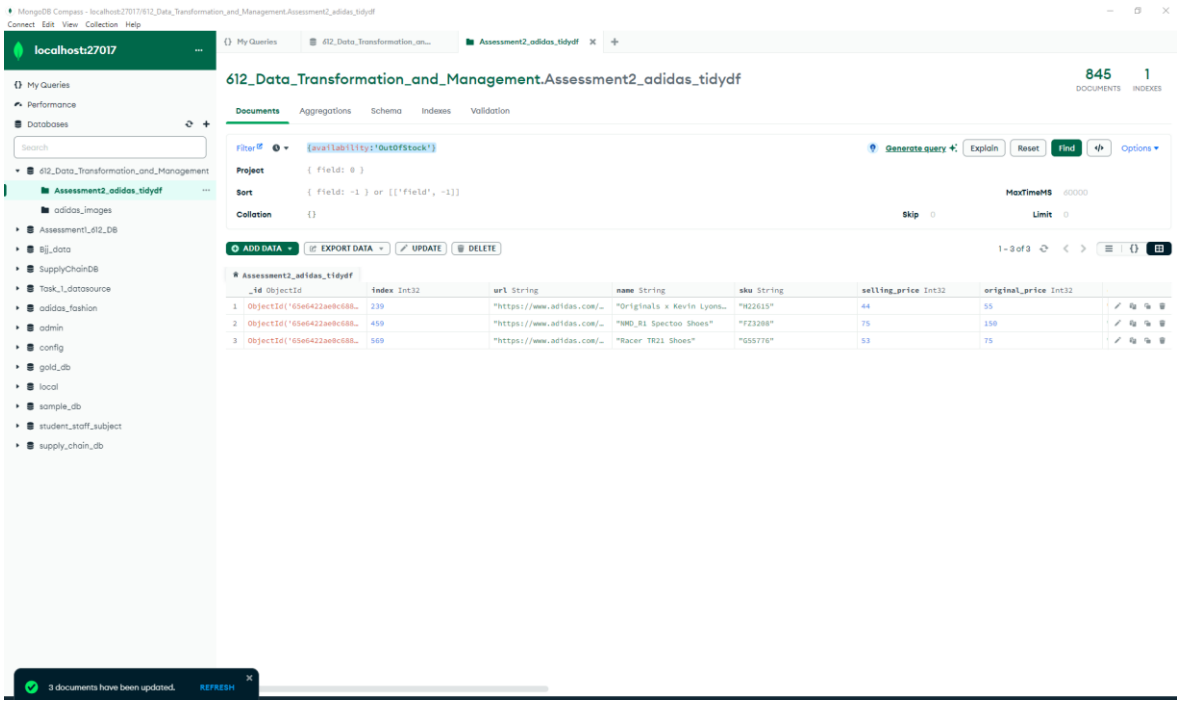
Assuming we need to update the following products from out-of-stock to in stock. Create a query to filter the data to find the item to be updated. In this case, I used the syntax {availability:'OutOfStock'}.



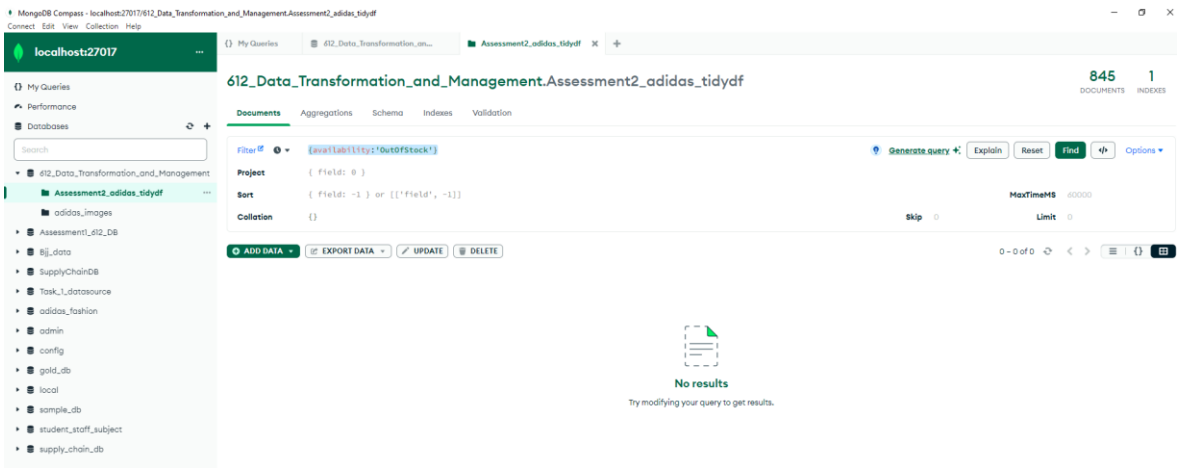
Click the UPDATE button to show you the pop-up form where you can write your query to update the documents.



Click Update 3 documents and refresh the page.



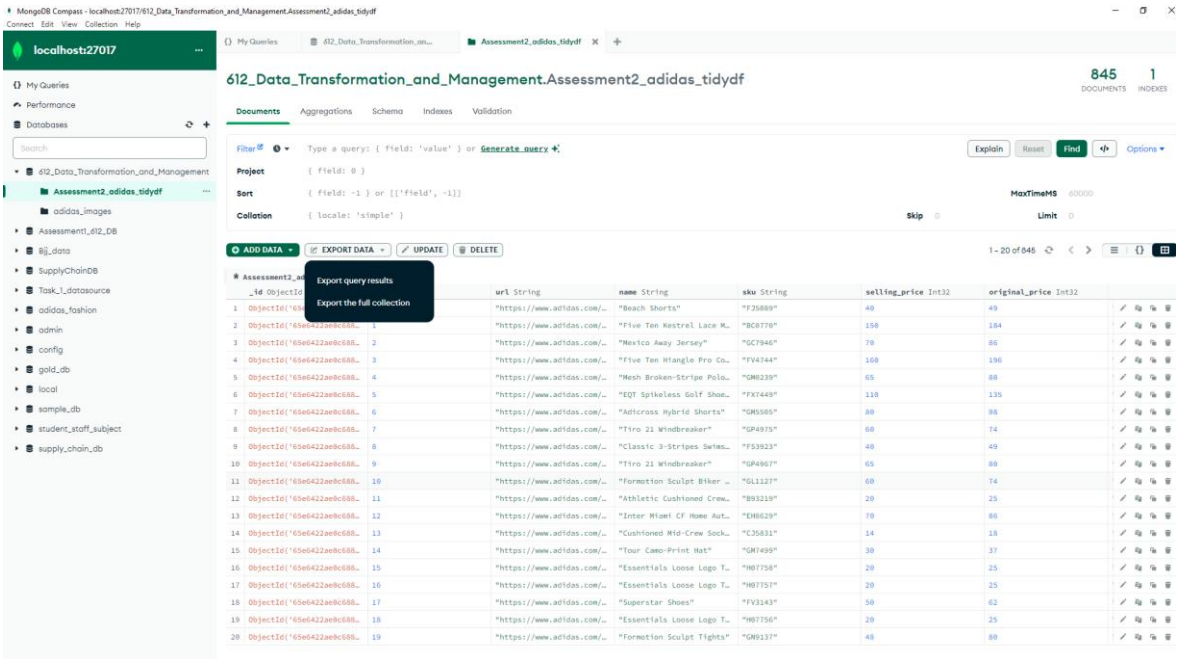
Once you click the refresh button, the previous result will be gone on the current page as it was already updated. The filter stayed the same, so no result was found. This means that all the three items in Out of Stock is now in In Stock status.

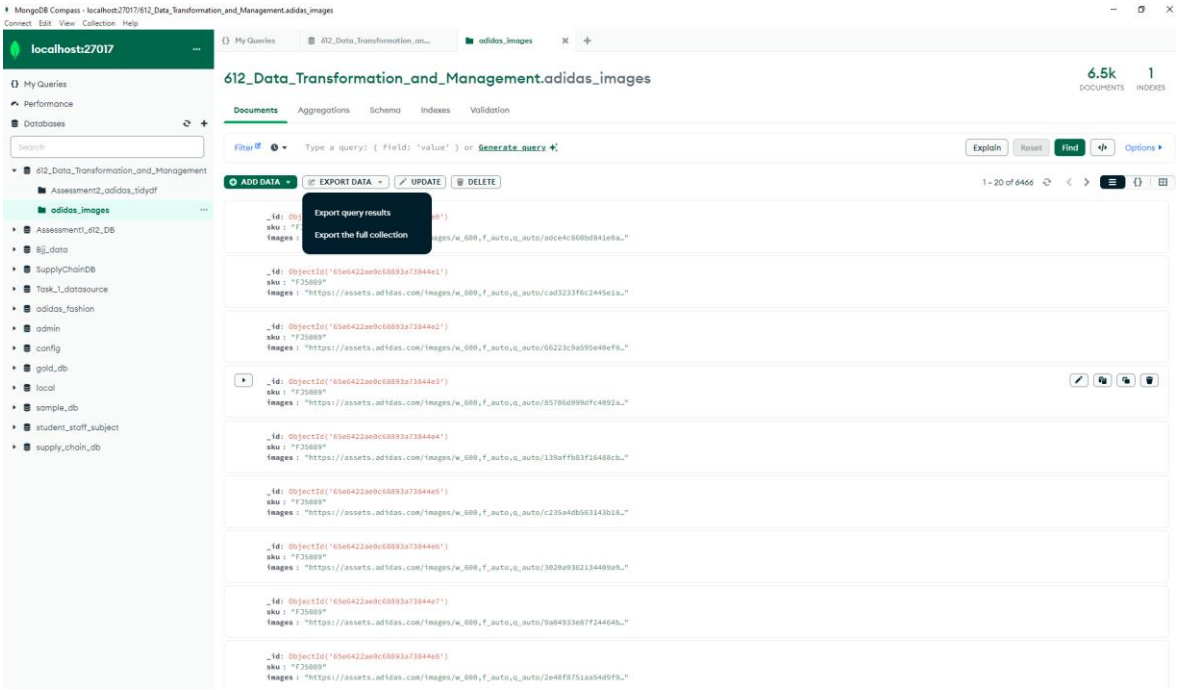


Chapter 5: Task B

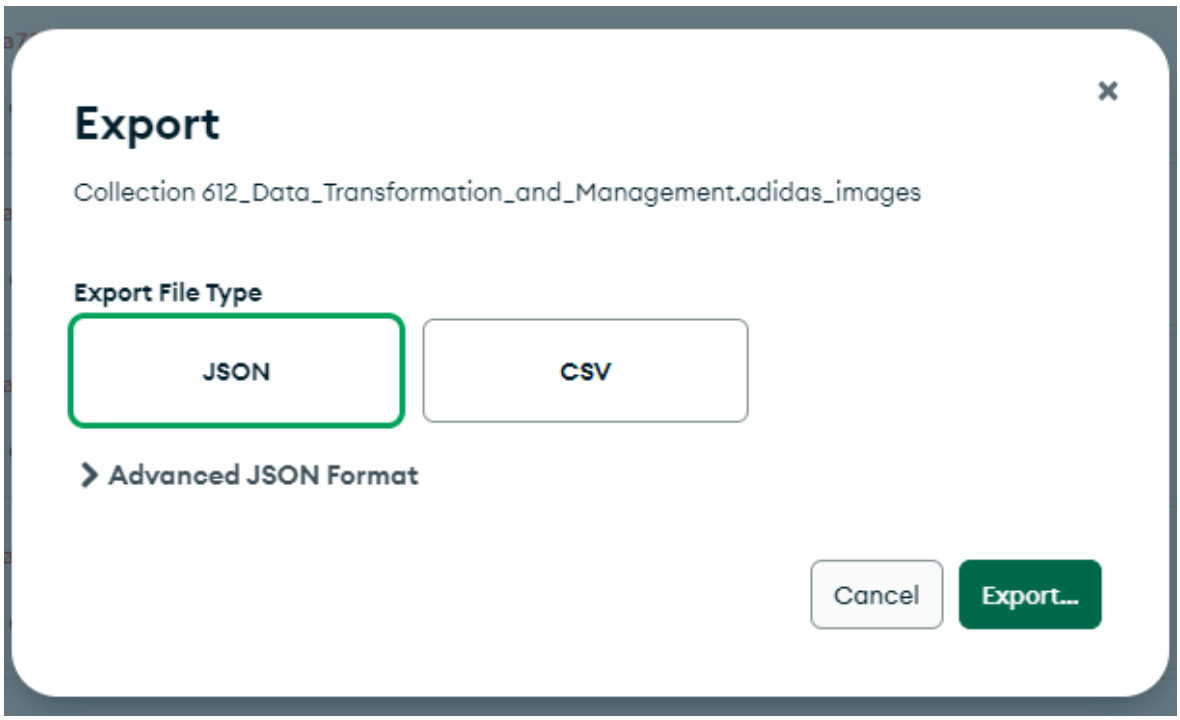
a) Exporting the entire collection from MongoDB.

1. Click the EXPORT data (note: You will be given a choice if want to export the query result or the full collection) and choose Export the entire collection.

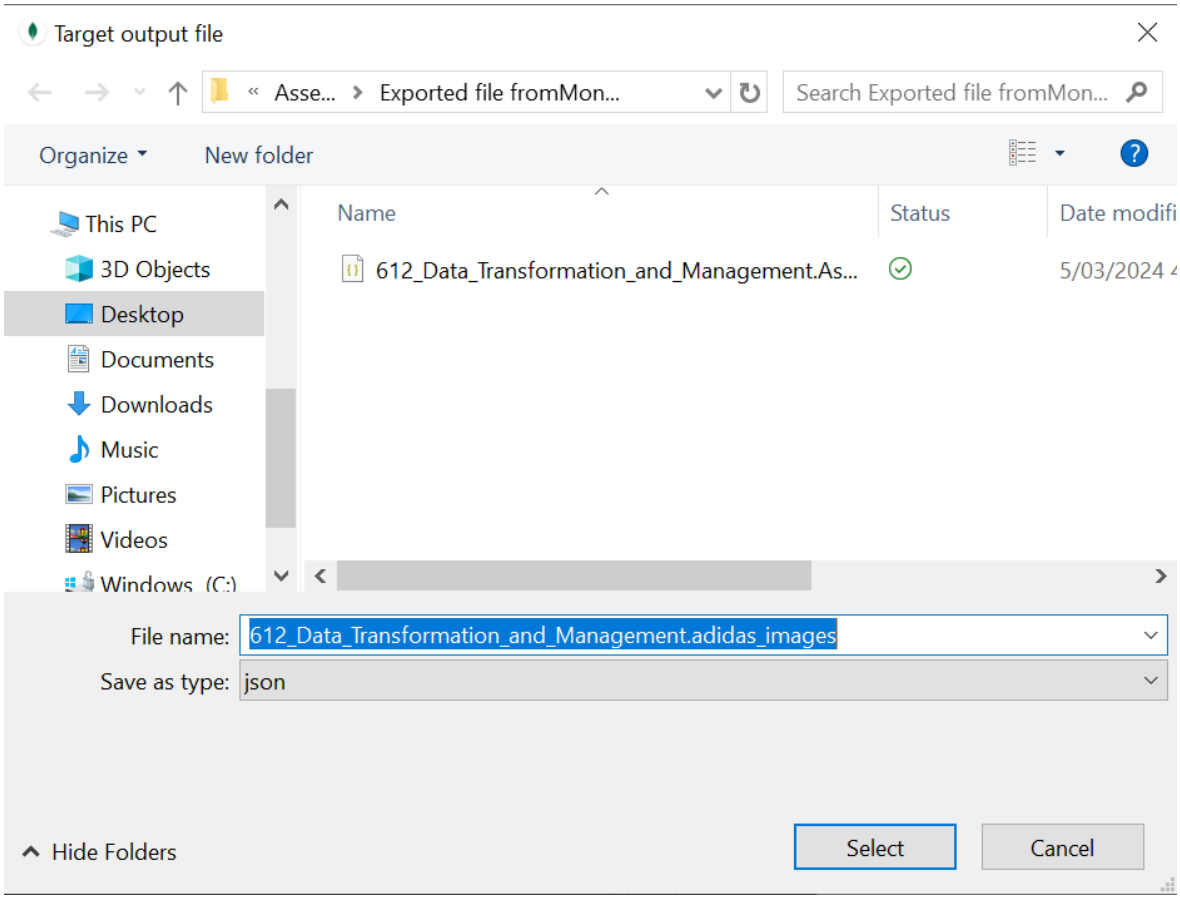




In exporting data, you will be given a choice to export it in JSON or CSV format. In my case, I chose the JSON as it has the same format as the parquet.

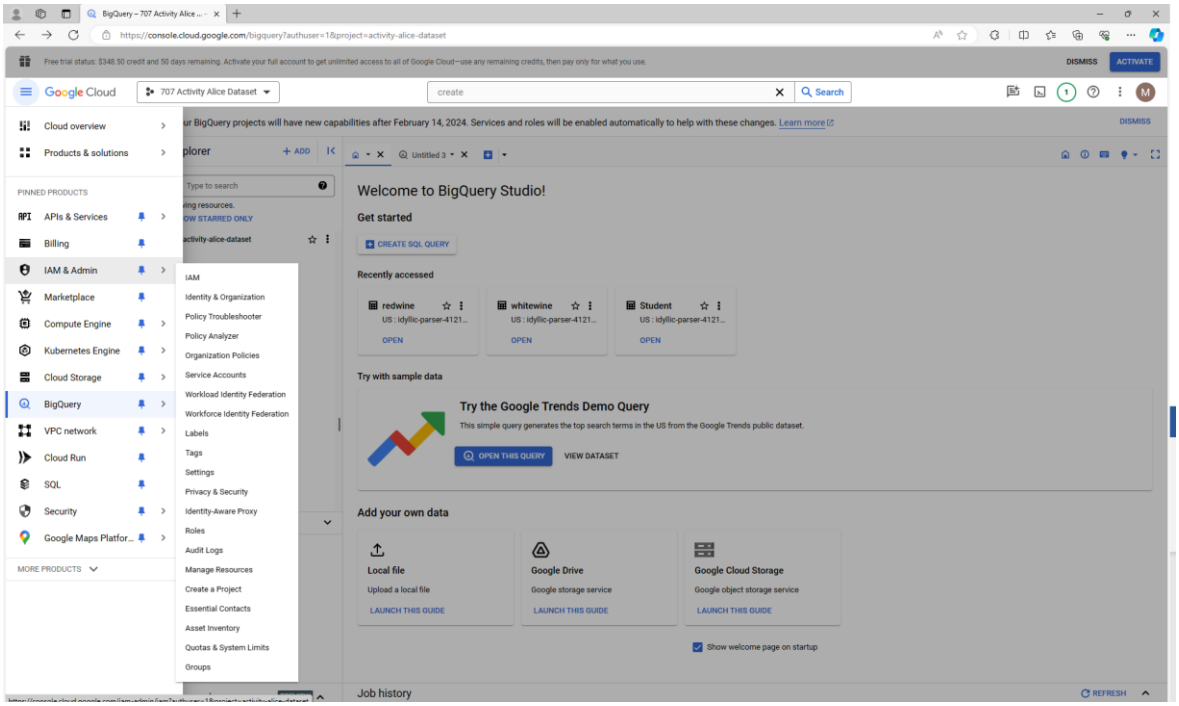


Exporting to my local computer.

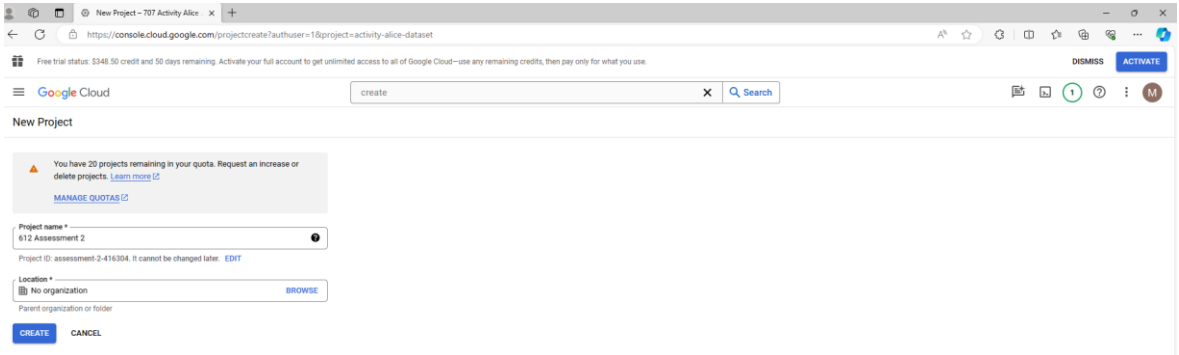


b) Establishing the connection to the cloud service and uploading the local file to the cloud storage bucket using the Google Cloud Platform (GCP)

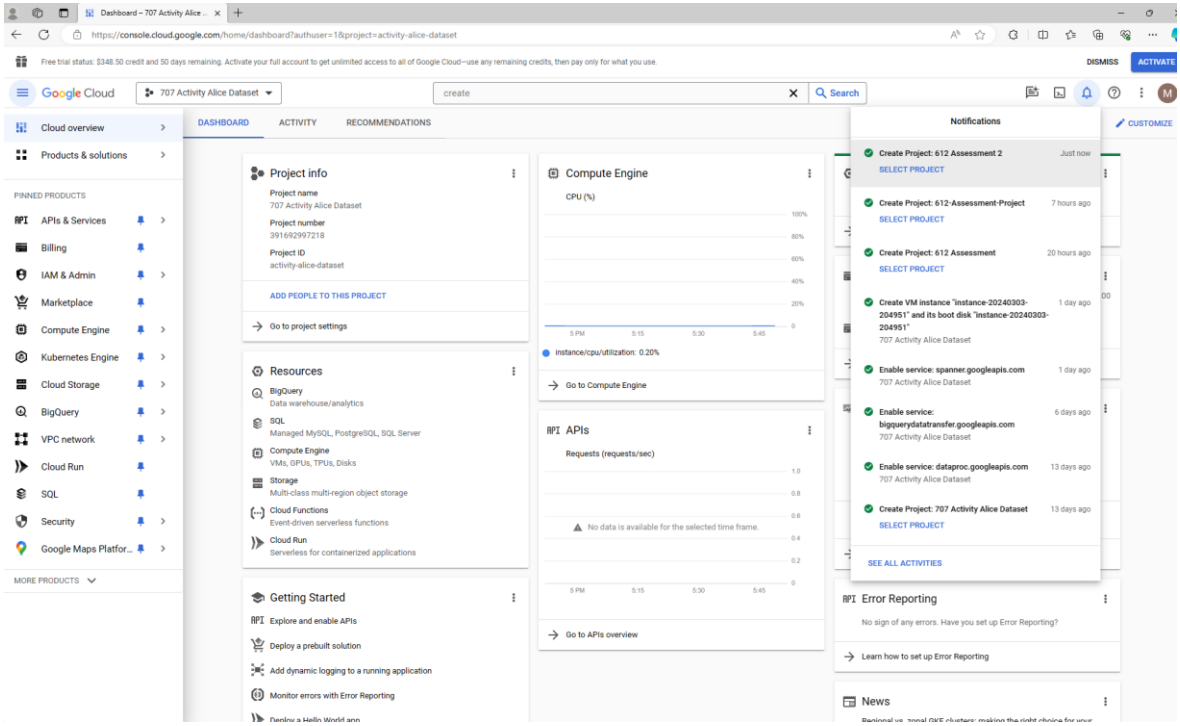
1. Click in Google Cloud, go to IAM & Admin, and create a project.



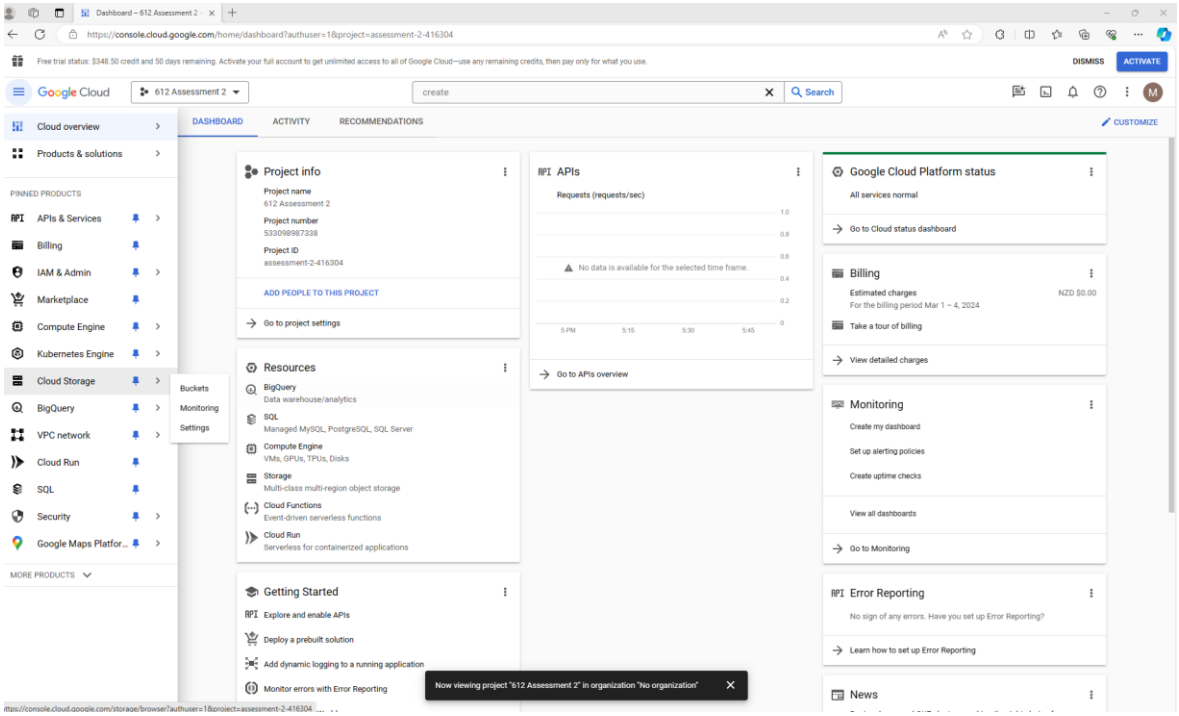
Naming the project.



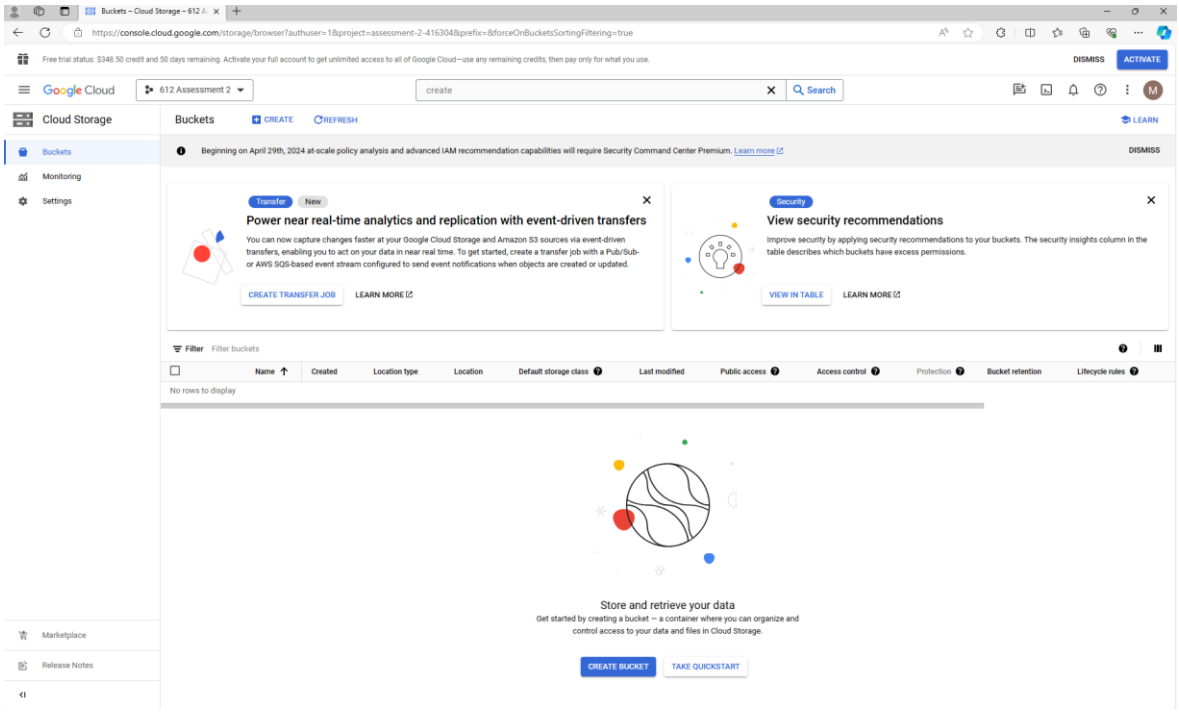
You will be notified once you have successfully created a project on the upper right side of the page.



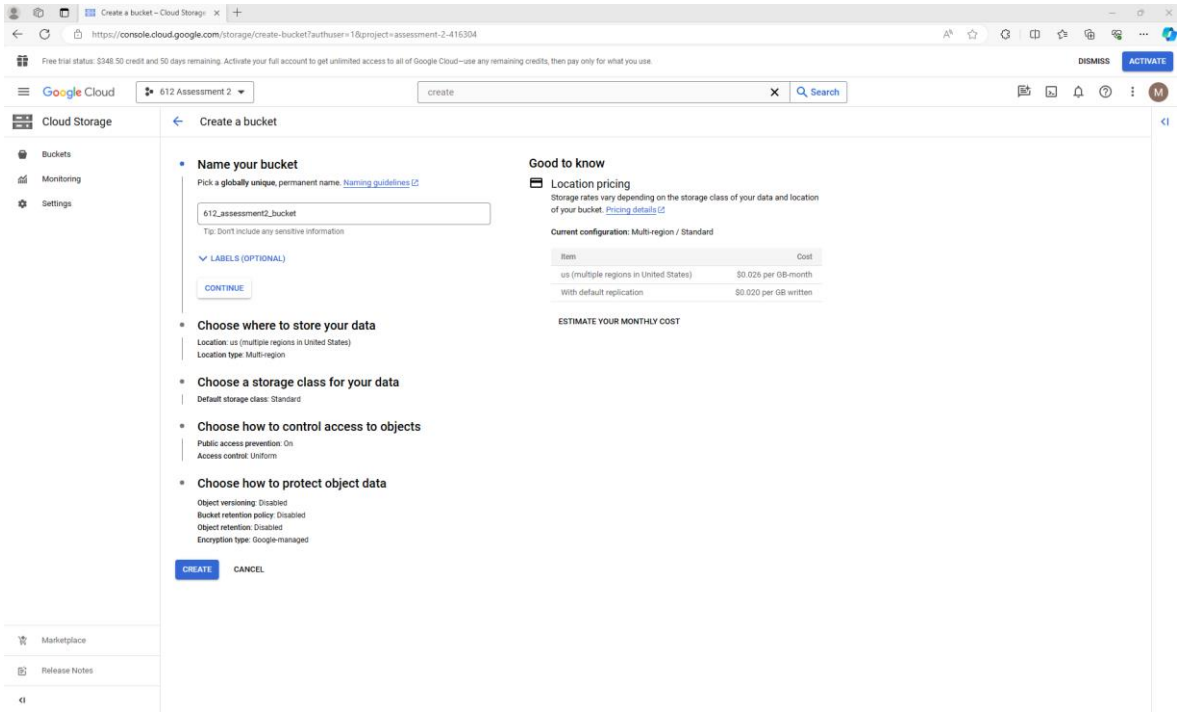
2. After creating the project, you can now create a bucket. From the navigation menu, go to the cloud storage and select Buckets.



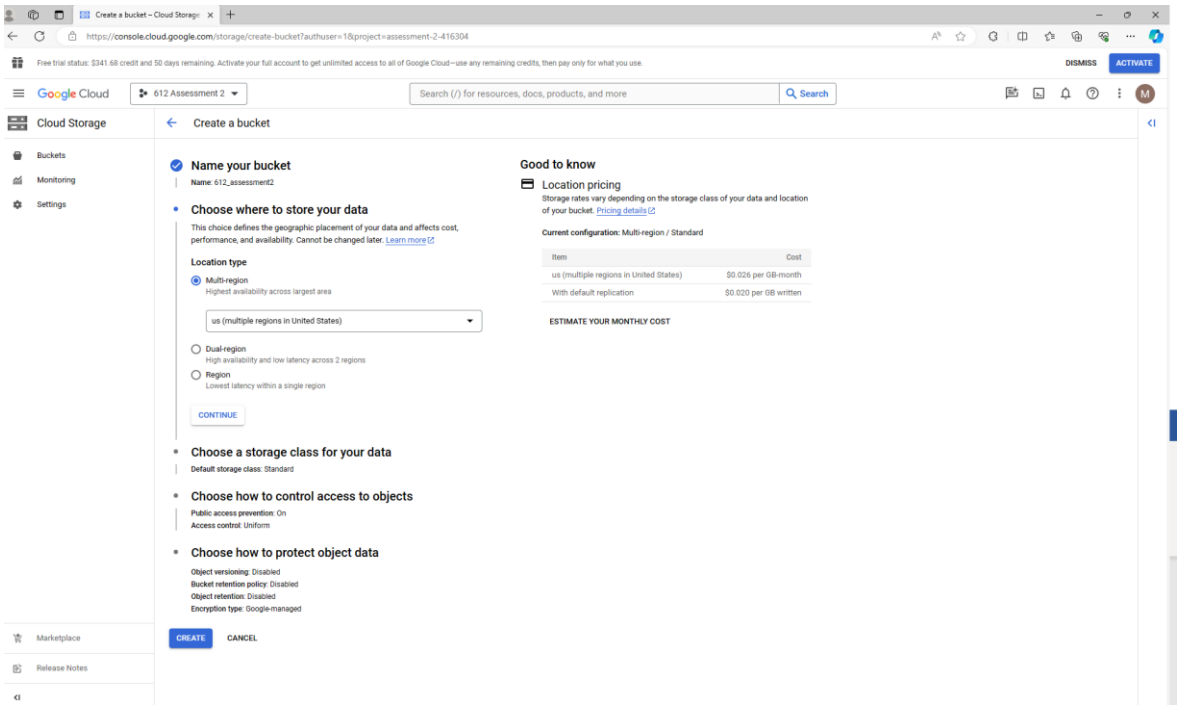
Click CREATE BUCKET



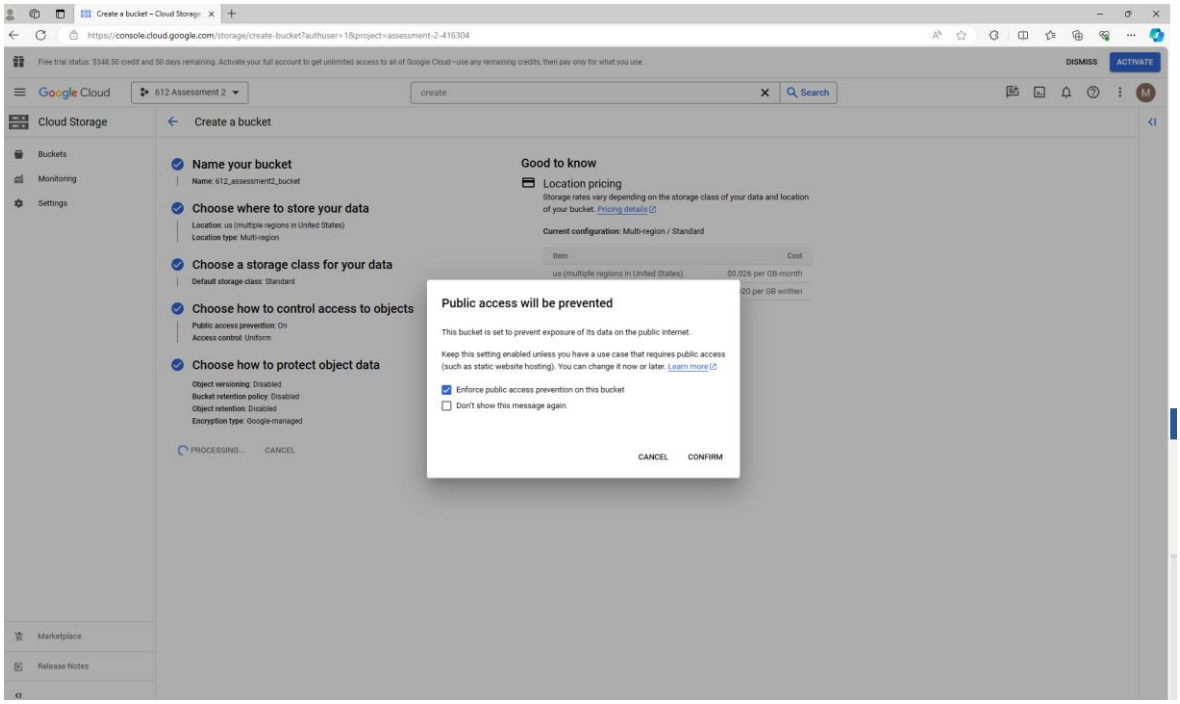
Name your bucket and click CONTINUE.



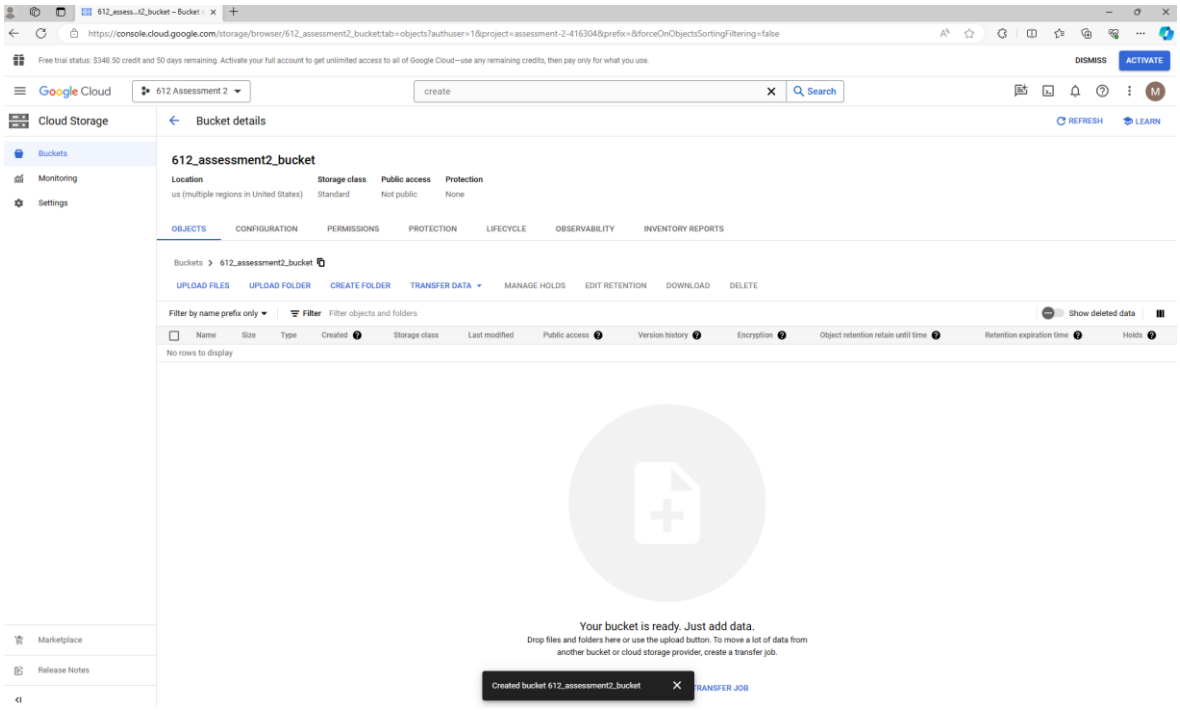
Choose where you want to store your data and click CREATE.



You will be prompted about public access, check to enforce public access, then click CONFIRM.

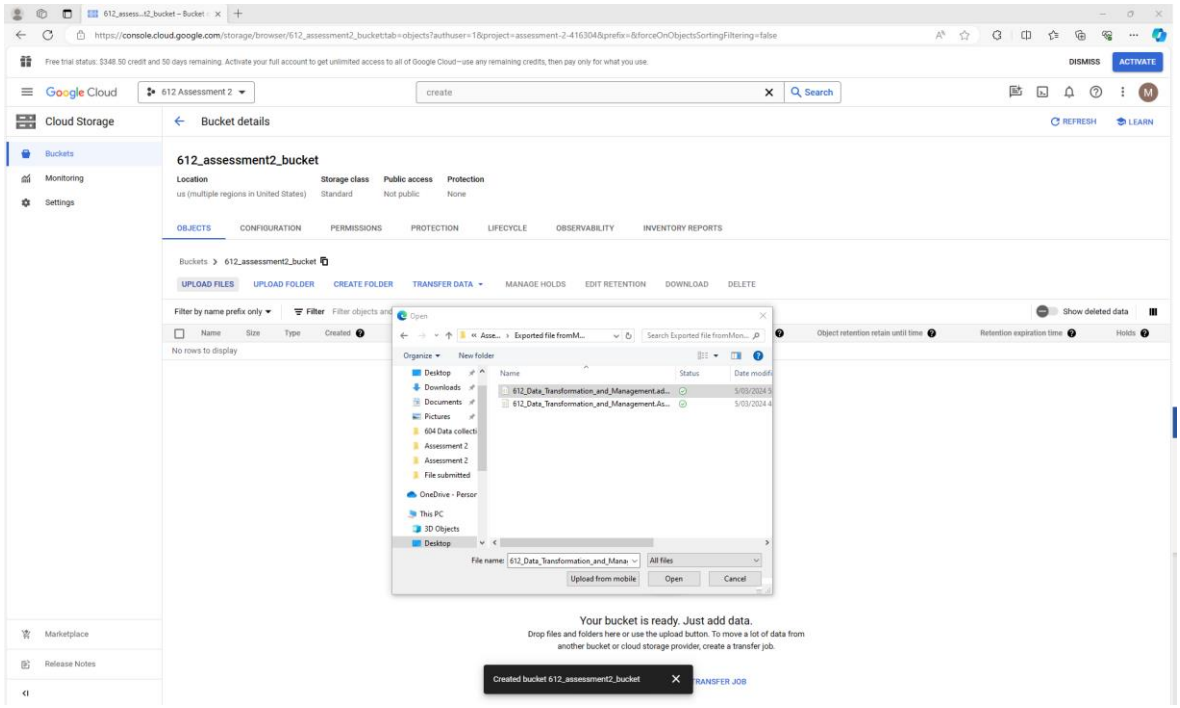


Below is the sample created bucket:

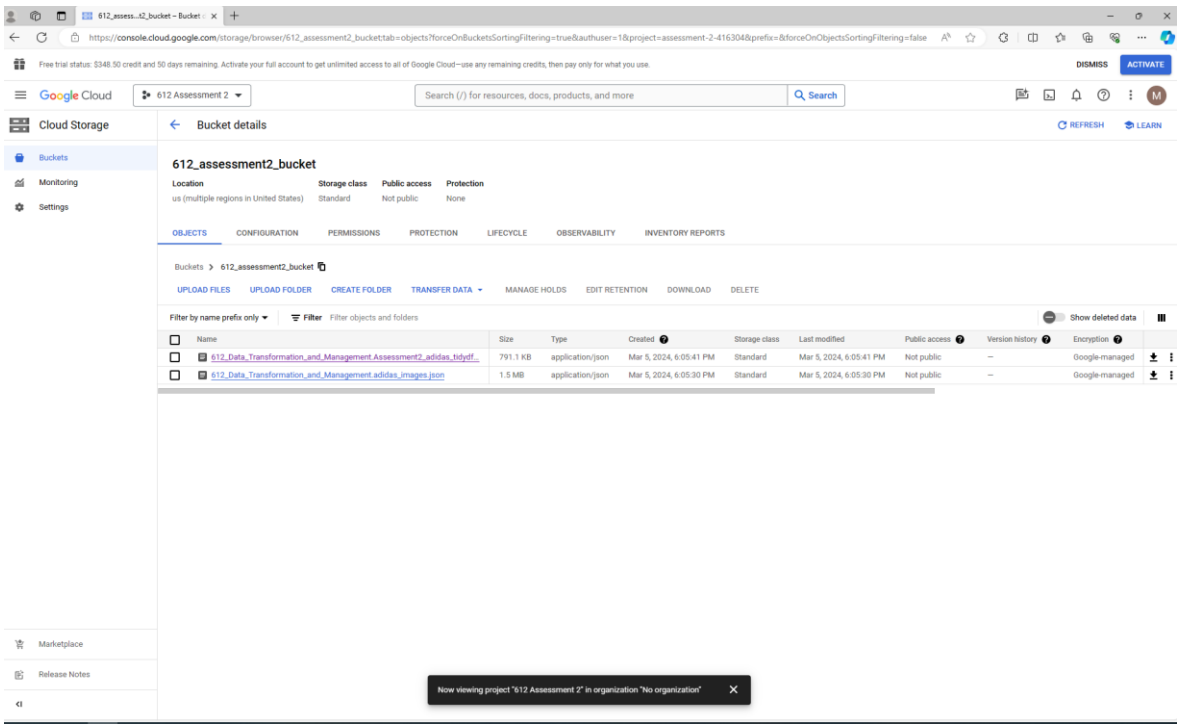


Once your bucket is created, you can now upload files (click upload files).

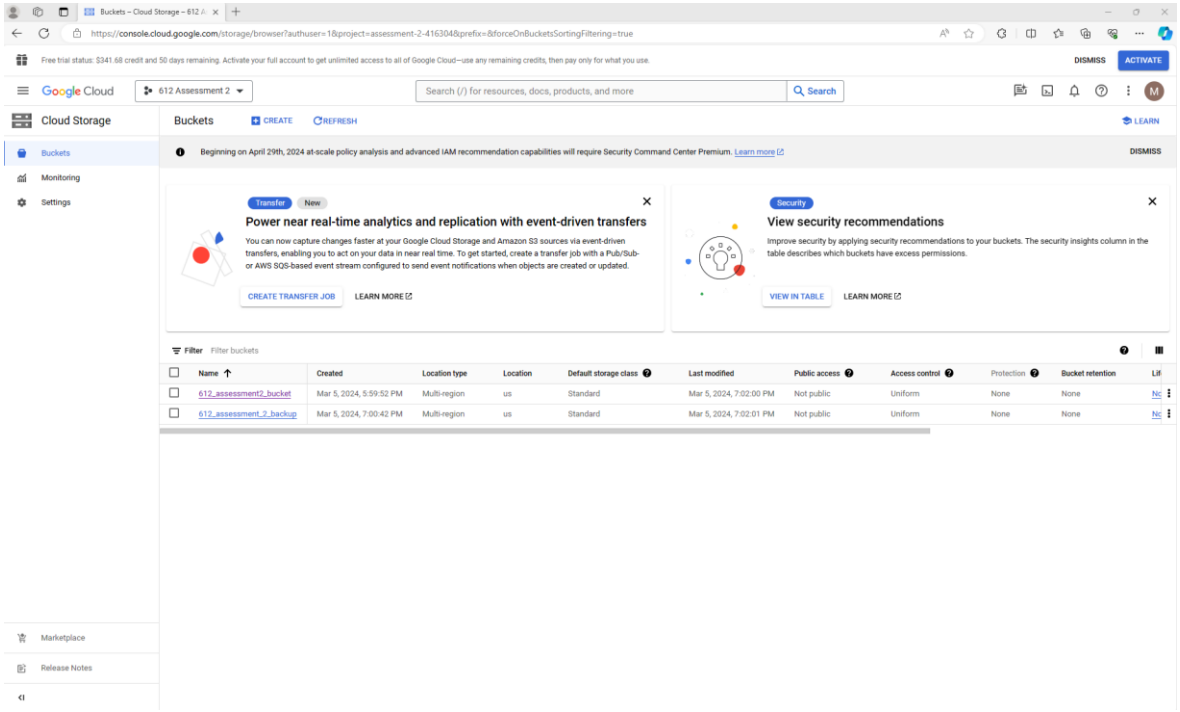
Uploading the exported file (from MongoDB) from my local computer.



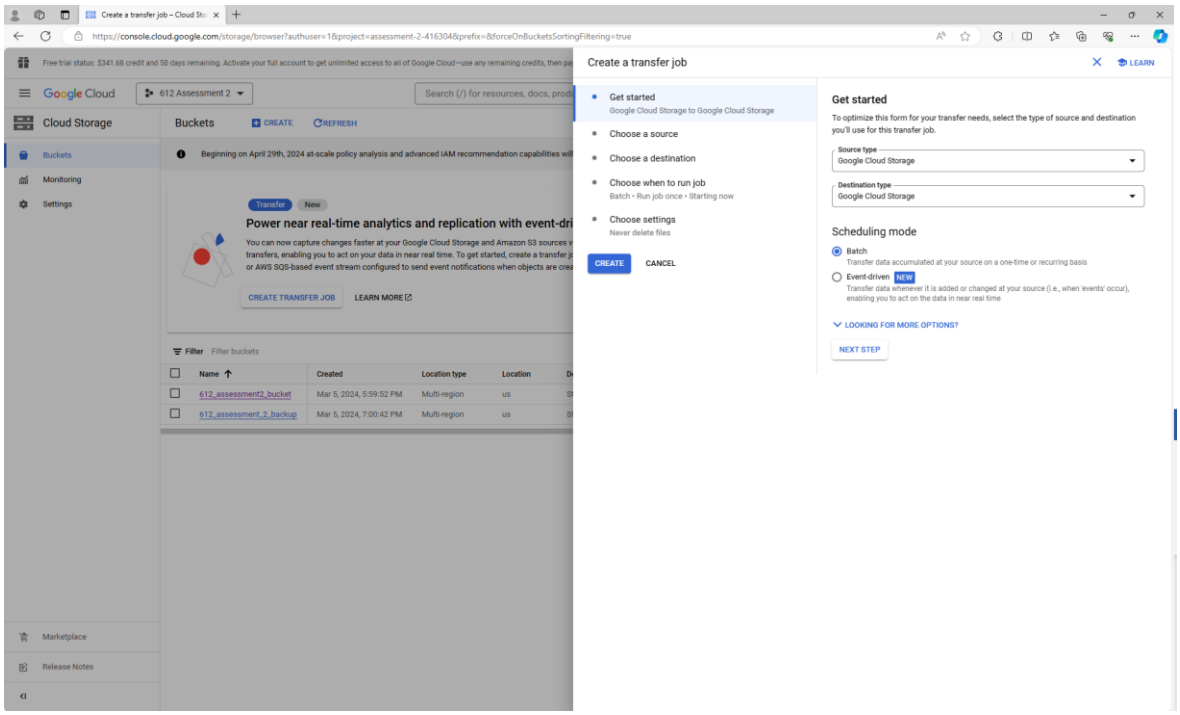
Below is the data uploaded from my local computer to the GCP bucket. I put it in the 612_assessment2_bucket collection.



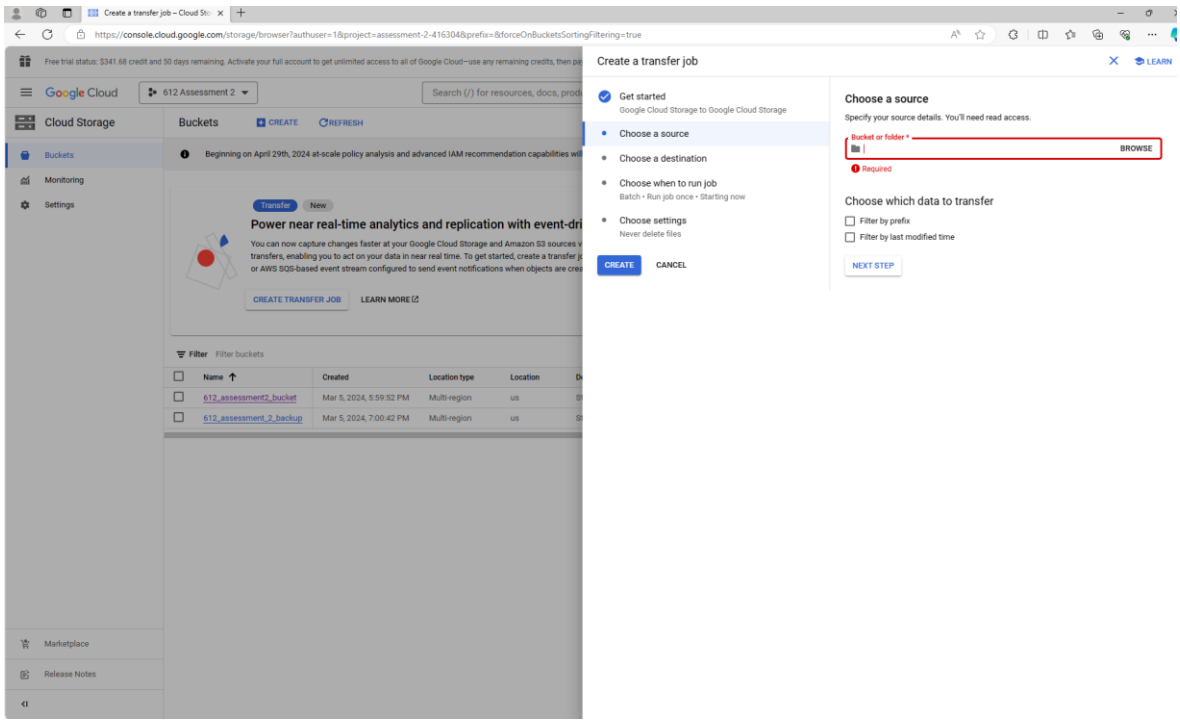
c. Following the same procedure in creating a bucket, I created another bucket to be used as backup and named it 612_assessment_2_backup.



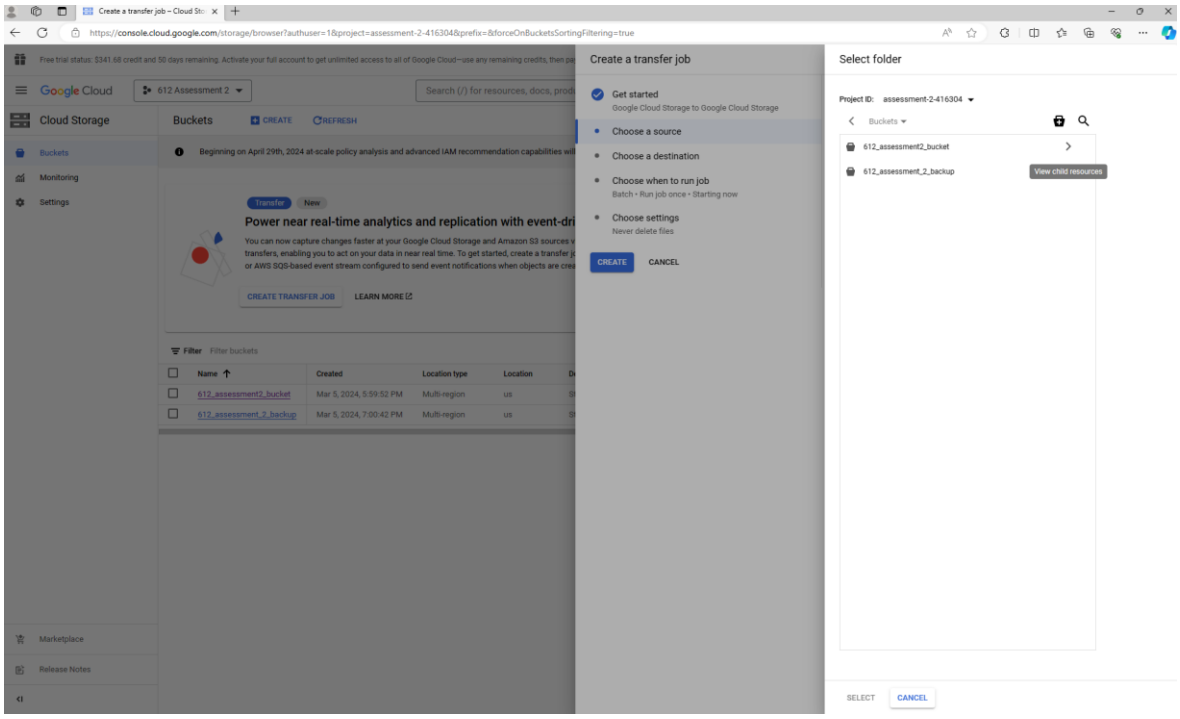
After creating the backup bucket, click the CREATE TRANSFER JOB



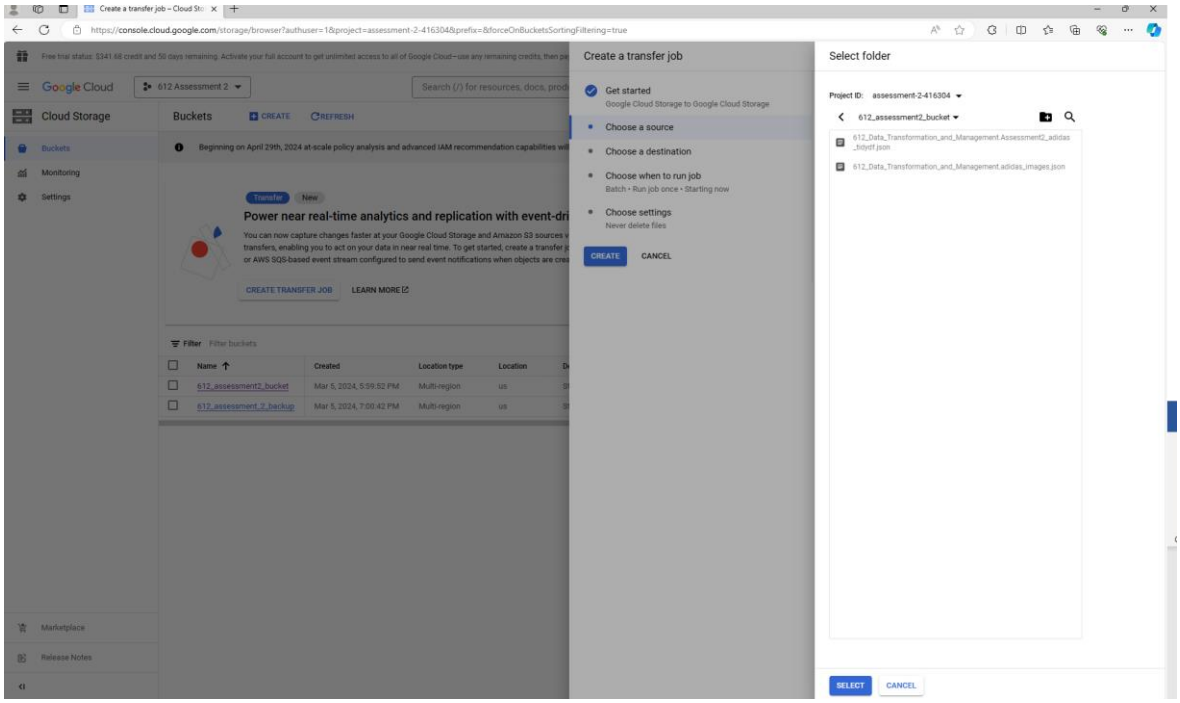
My data source and backup are in the same Google Cloud Storage, so I put Google Cloud Storage in Source type and Destination type and choose Batch in scheduling mode. Click NEXT STEP.



Choose a source to be backed up.

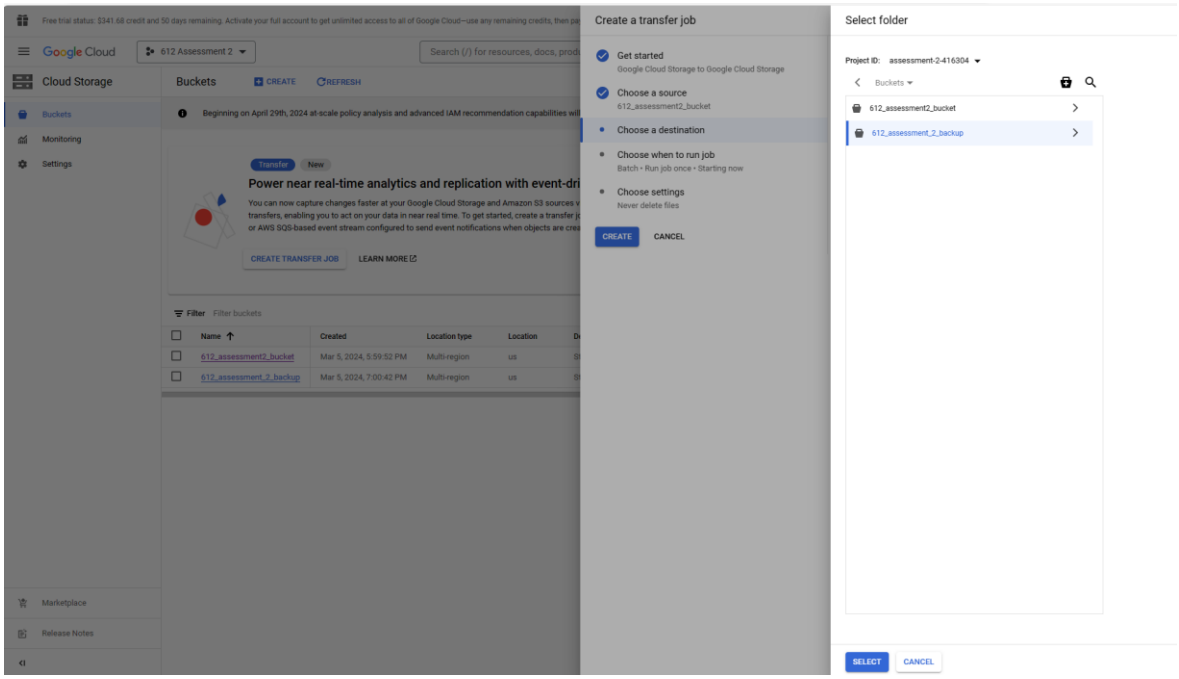


Then click SELECT.

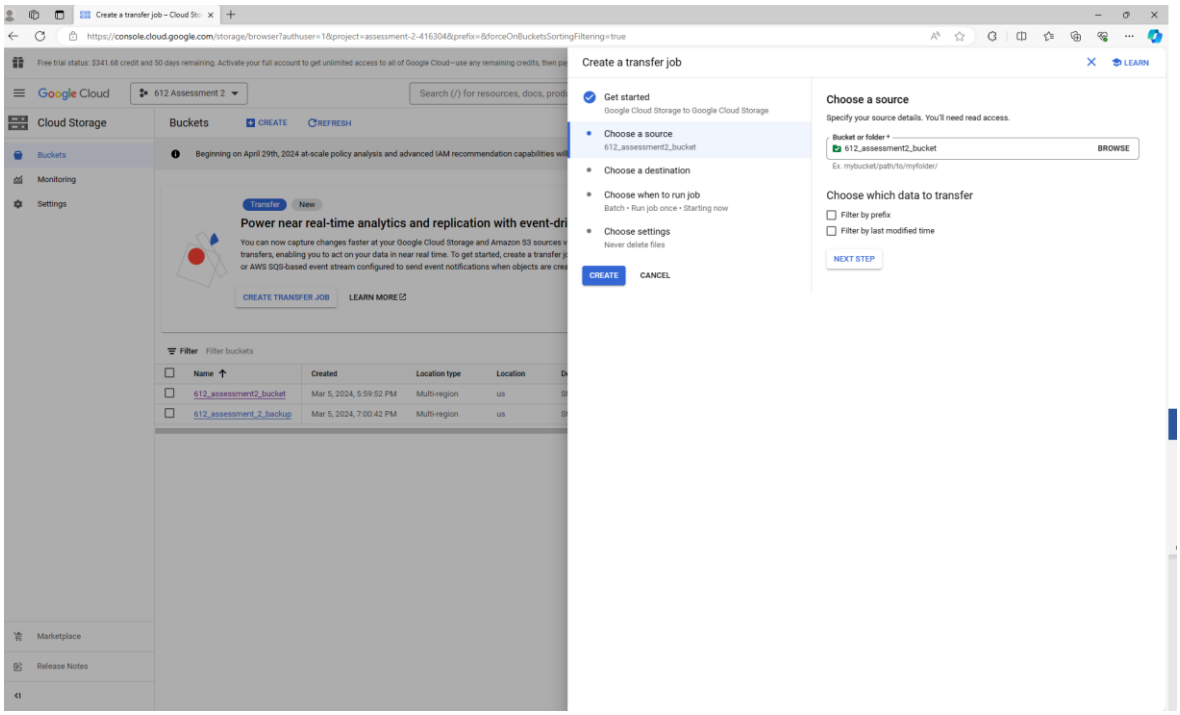


After clicking SELECT, you will be prompted to this screen. You will be given a choice to filter the content from the data source. If you want everything to be transferred, you may disregard it and click NEXT STEP.

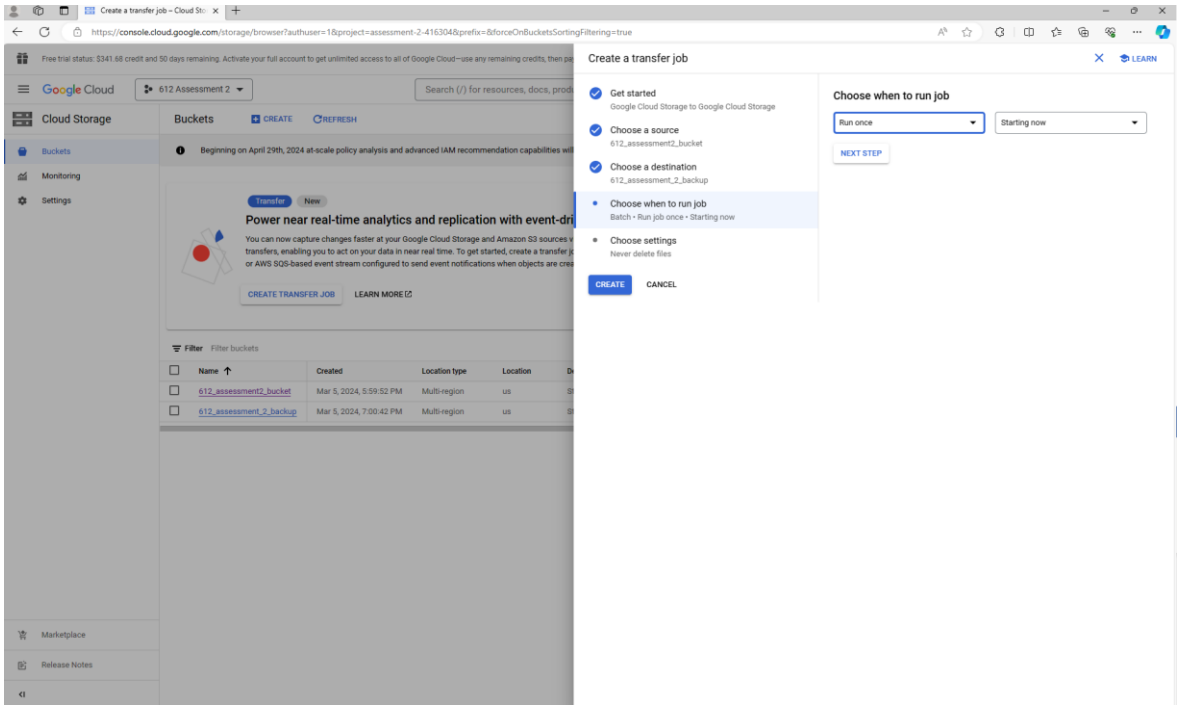
You will be asked to choose a destination, this time, I chose the 612_assessment_2_backup, and then click SELECT.



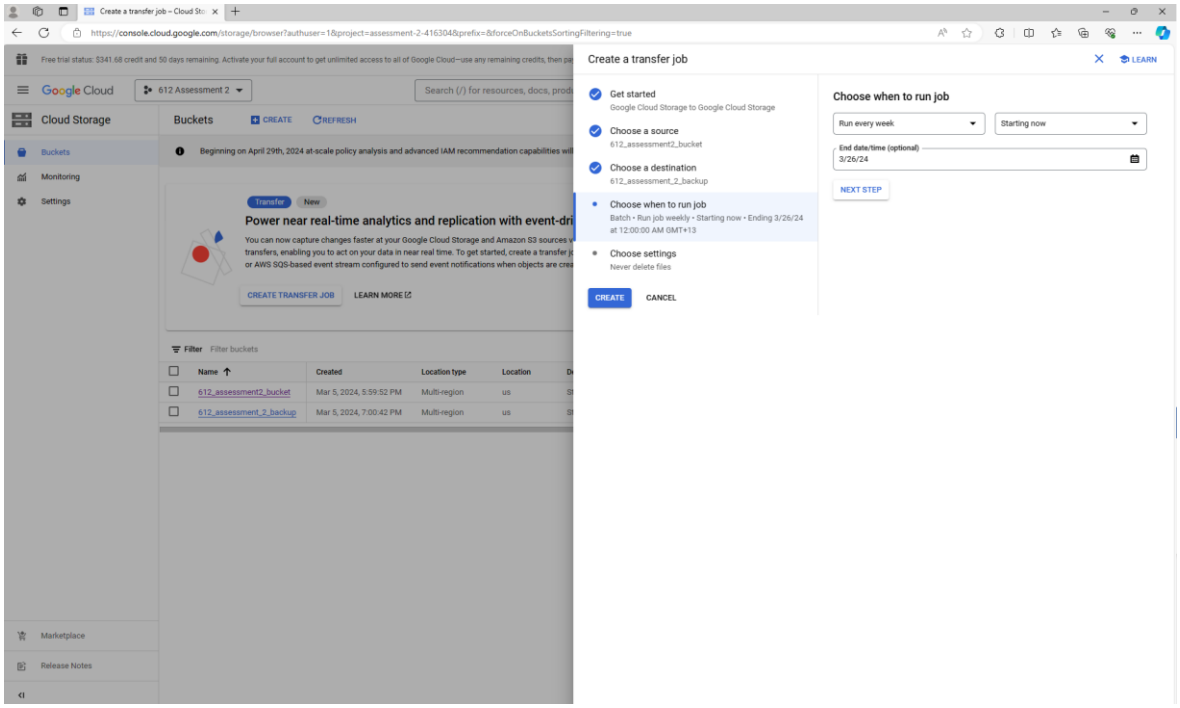
Click NEXT STEP



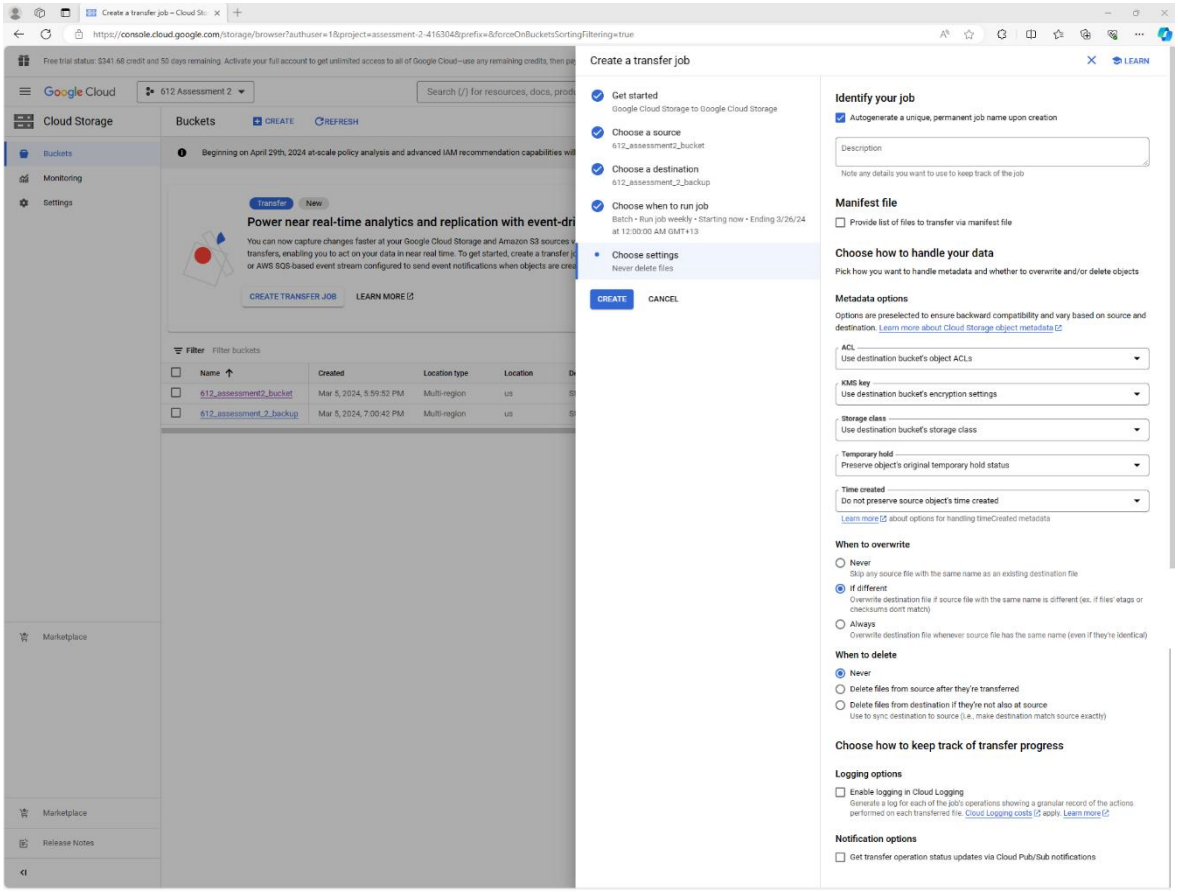
You may now choose when to run the job.



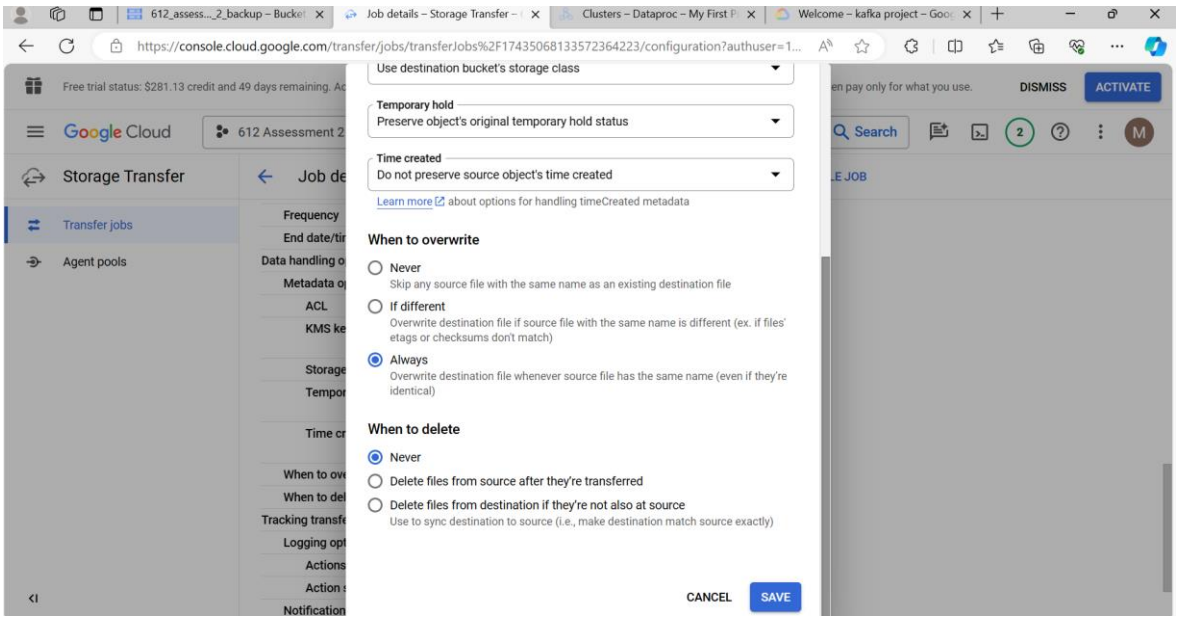
Once you created a schedule, click NEXT STEP.



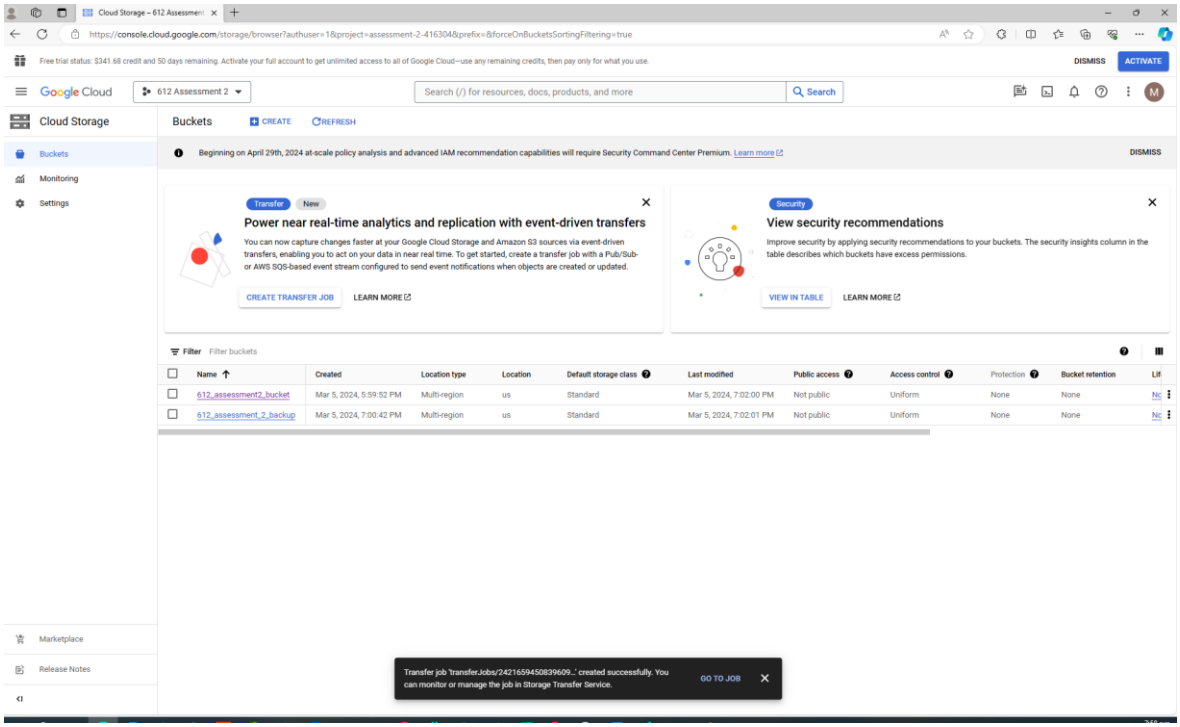
Check the default settings.



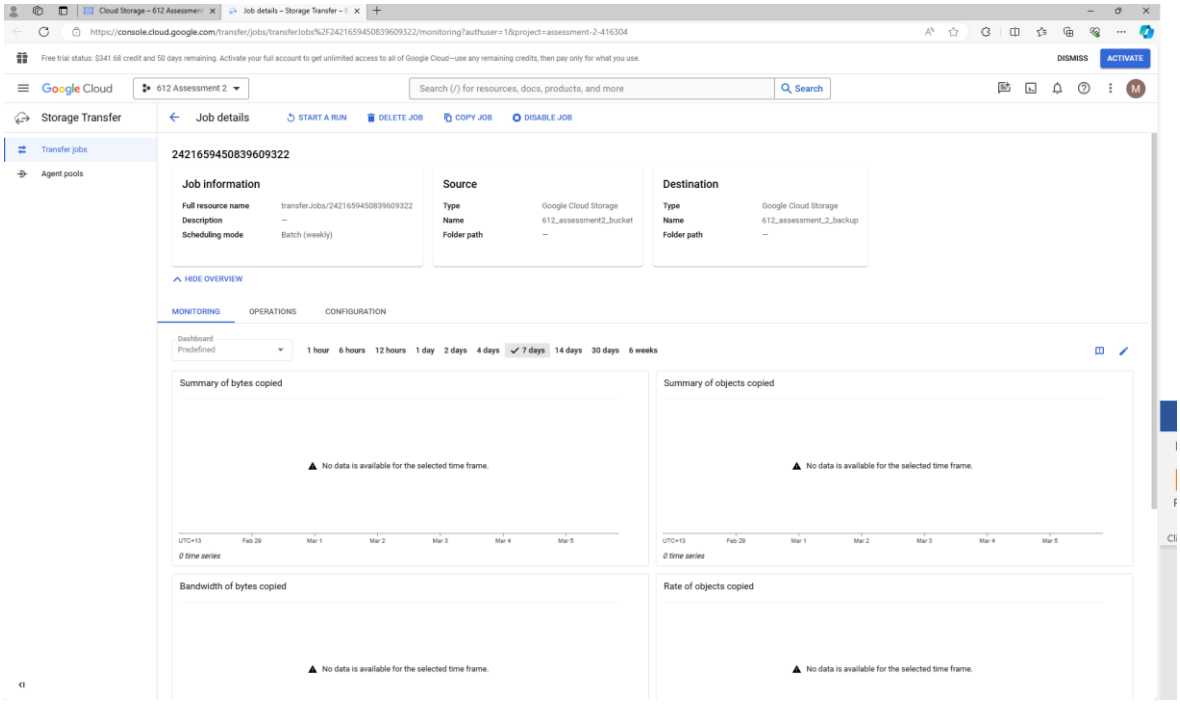
In the default settings, change the overwrite option to Always. So we can see the backup history.



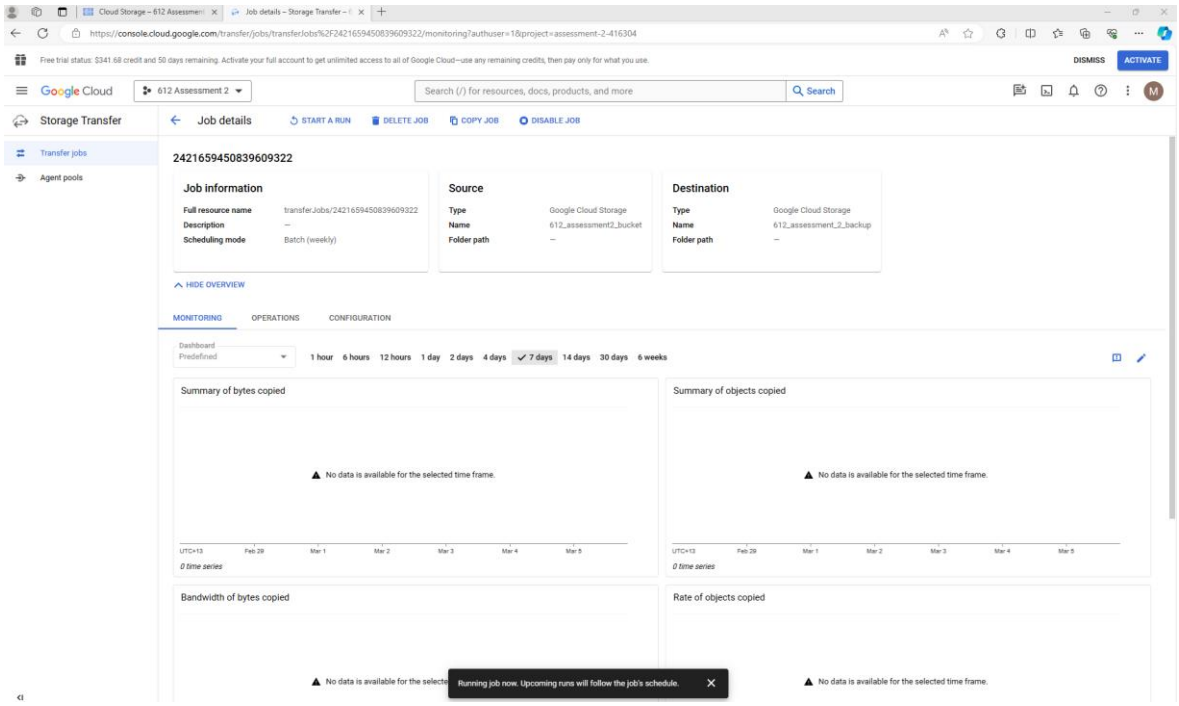
Lastly, click CREATE. You will be prompted to the Buckets screen. Click GO TO JOB at the bottom page.



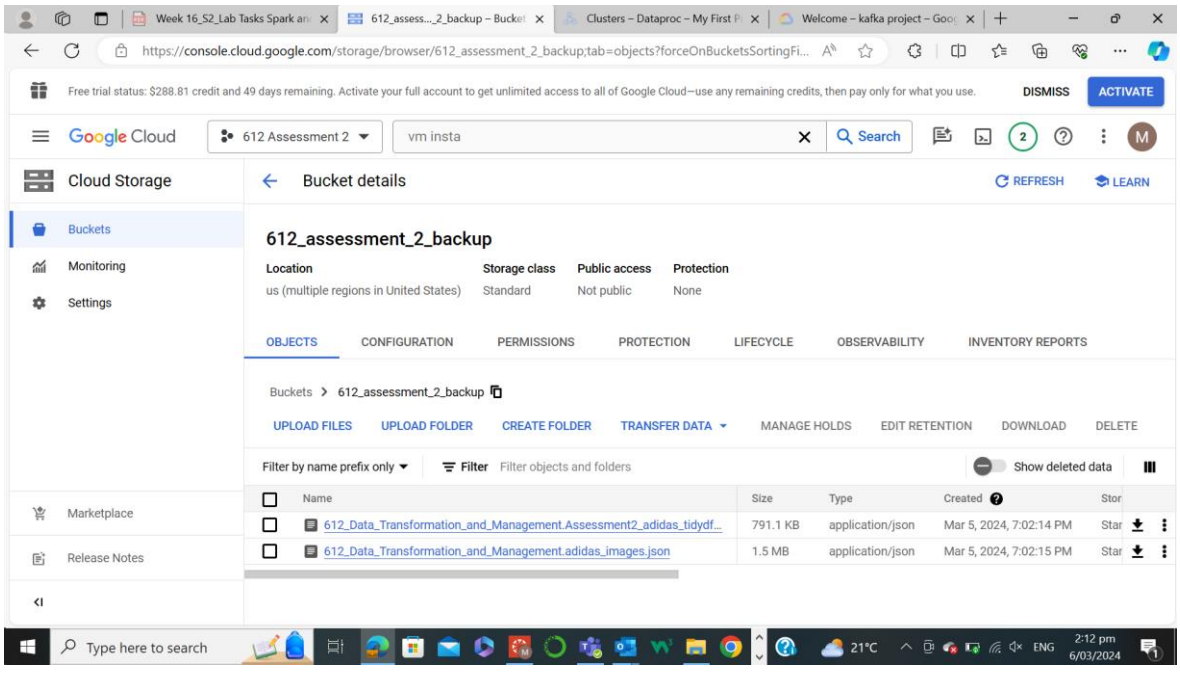
You will be prompted to the screen below. Click **START A RUN.**



The upcoming runs will follow the job's schedule. Click START A RUN.

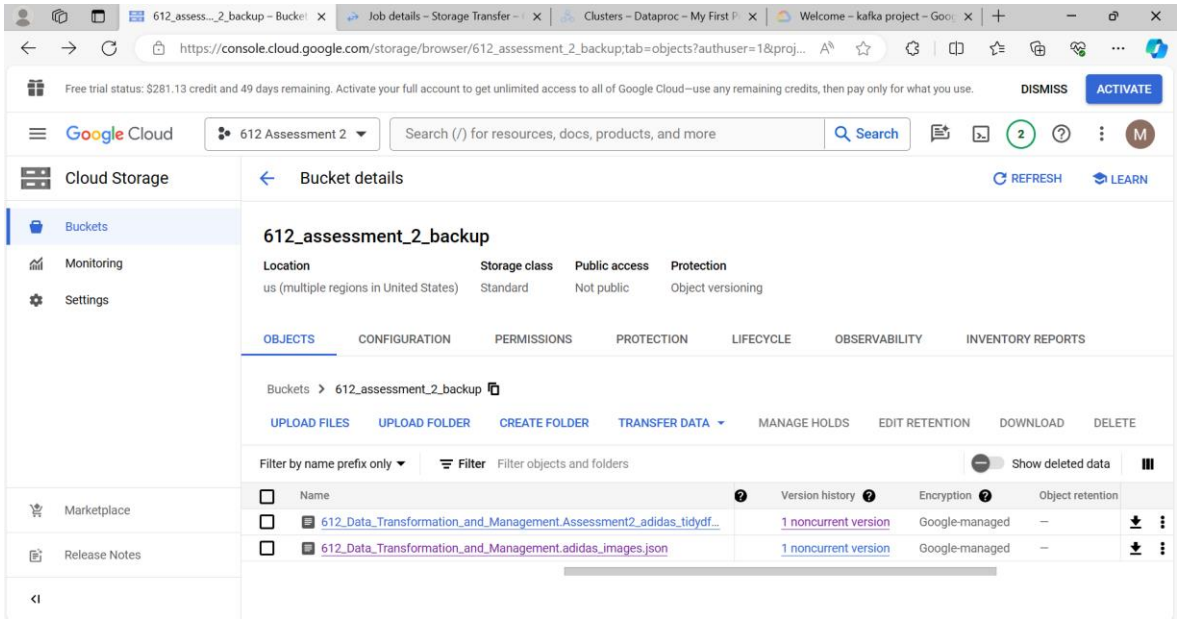


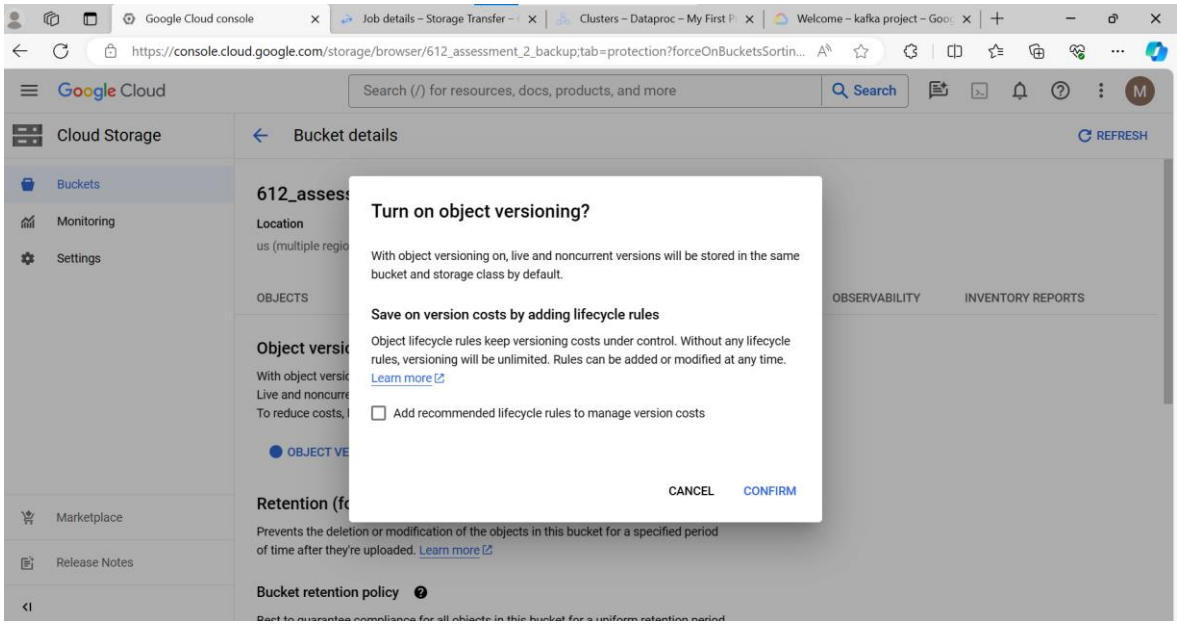
Checking if the backup has been successfully done.



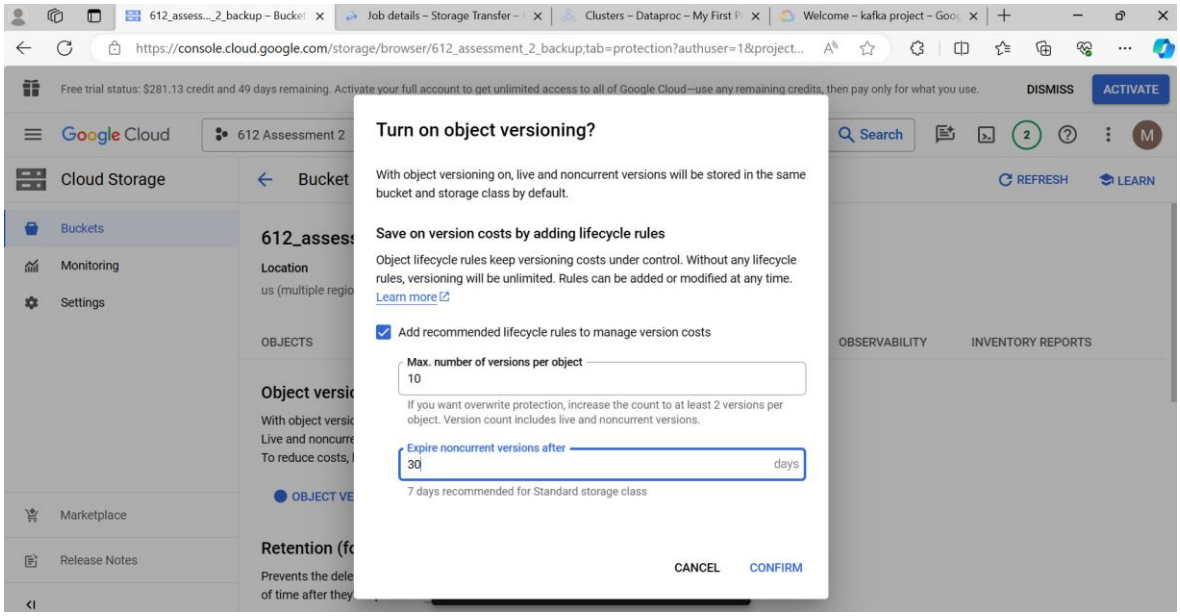
Object Versioning – is used to maintain an archive of the files, giving you the option to retrieve accidentally deleted data.

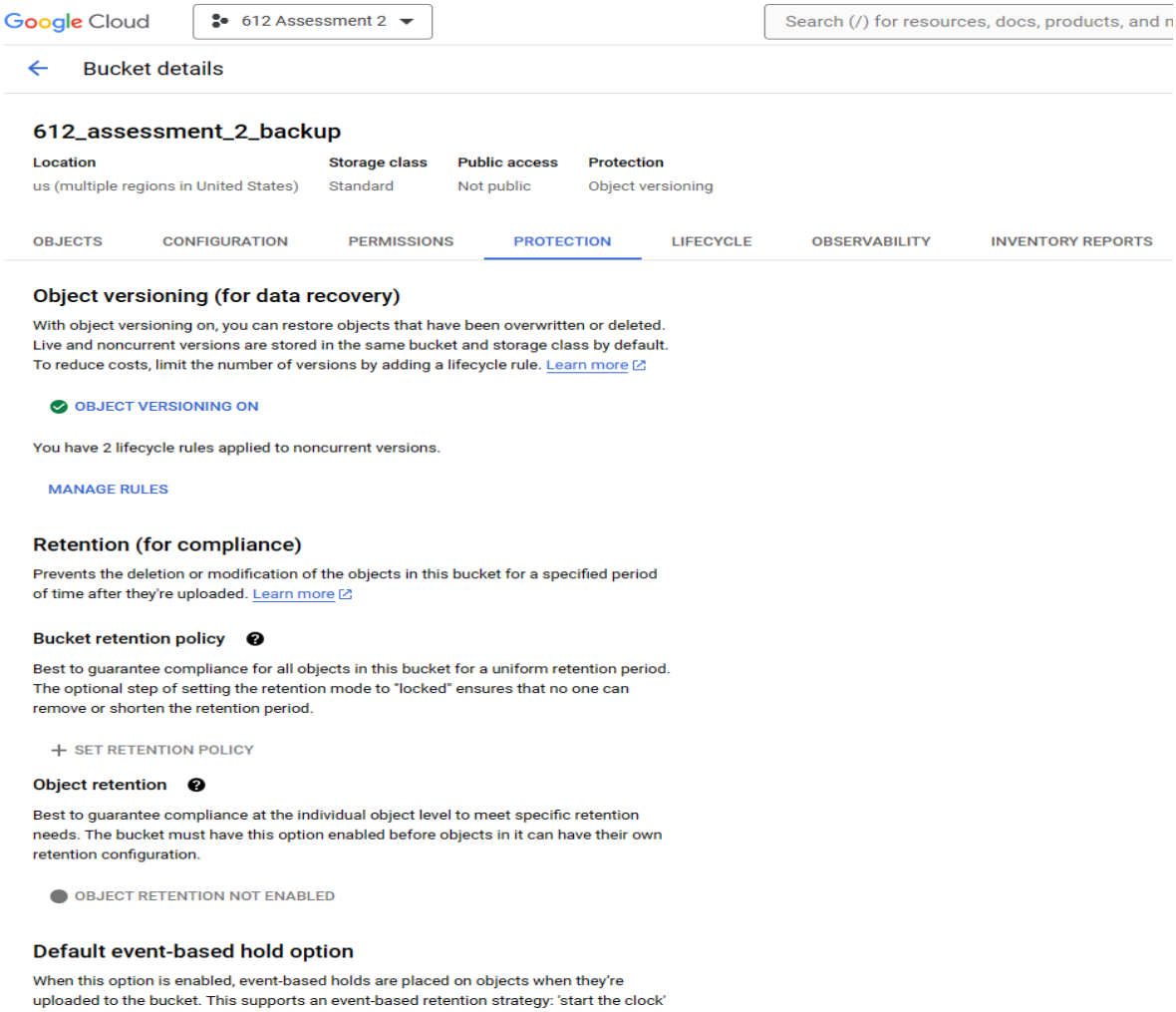
To enable the Object Versioning, go to the 612_assessment_2_backup bucket, click the PROTECTION tab, choose the file, and click CONFIRM.



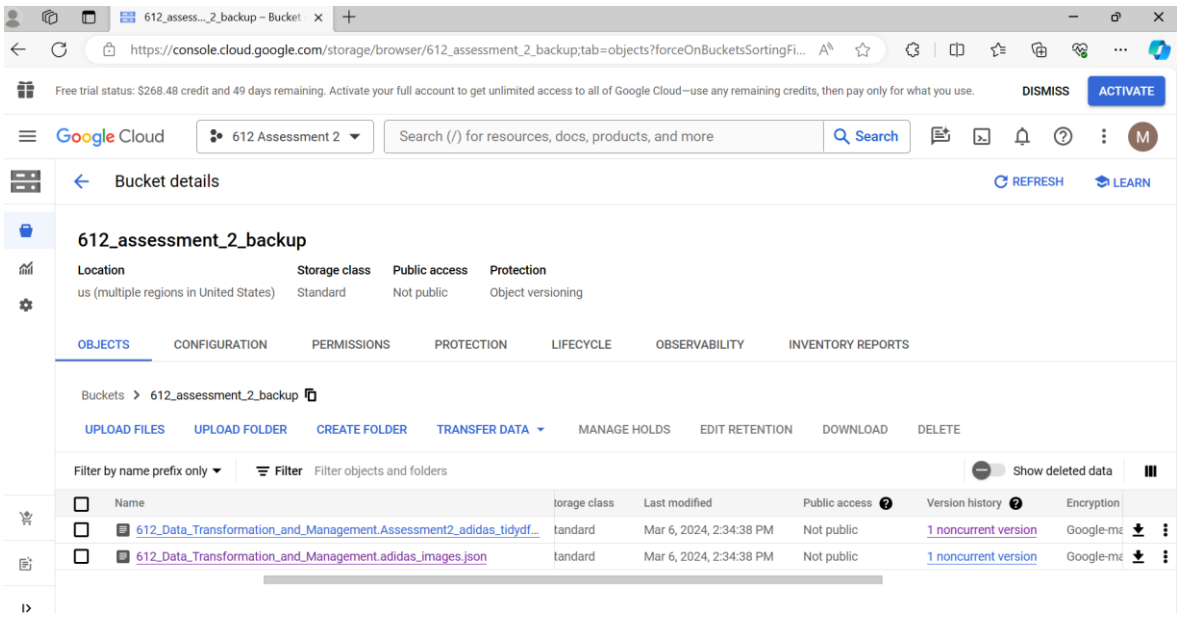


Choose the recommended life cycle rules for maximum versions per object and the noncurrent version expiry. The example below means you can only hold up to 10 backup versions of the file. The oldest version will be automatically removed if you exceed more than 10. After 30 days, the old files will be removed, and it will only save the latest file.



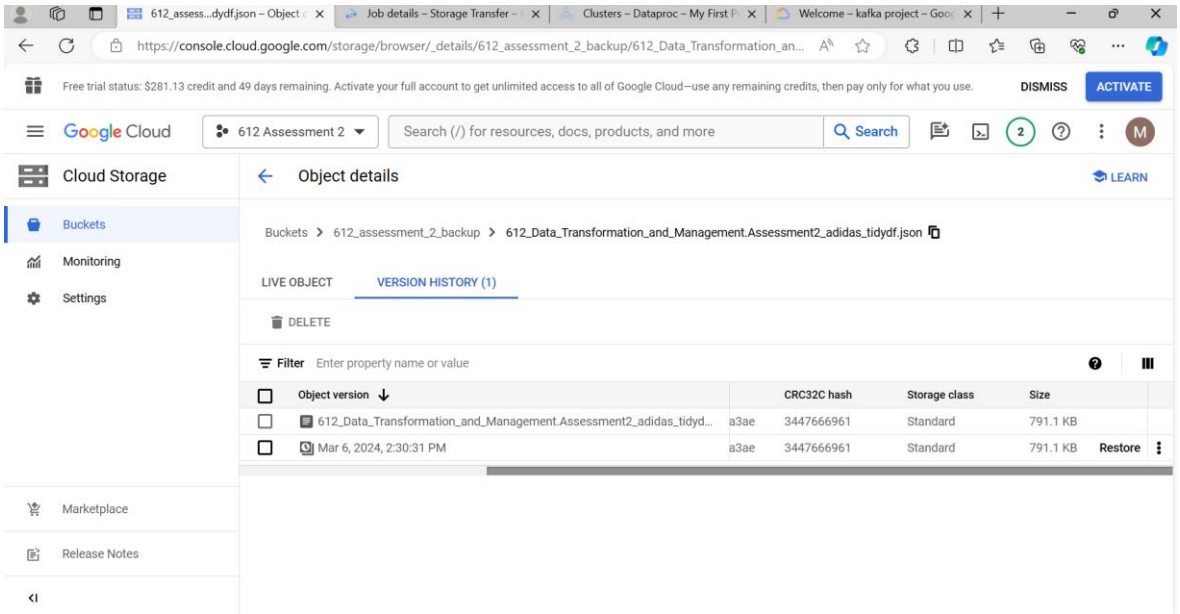


Check the Version History in the backup bucket to check if the Object Versioning has been done.

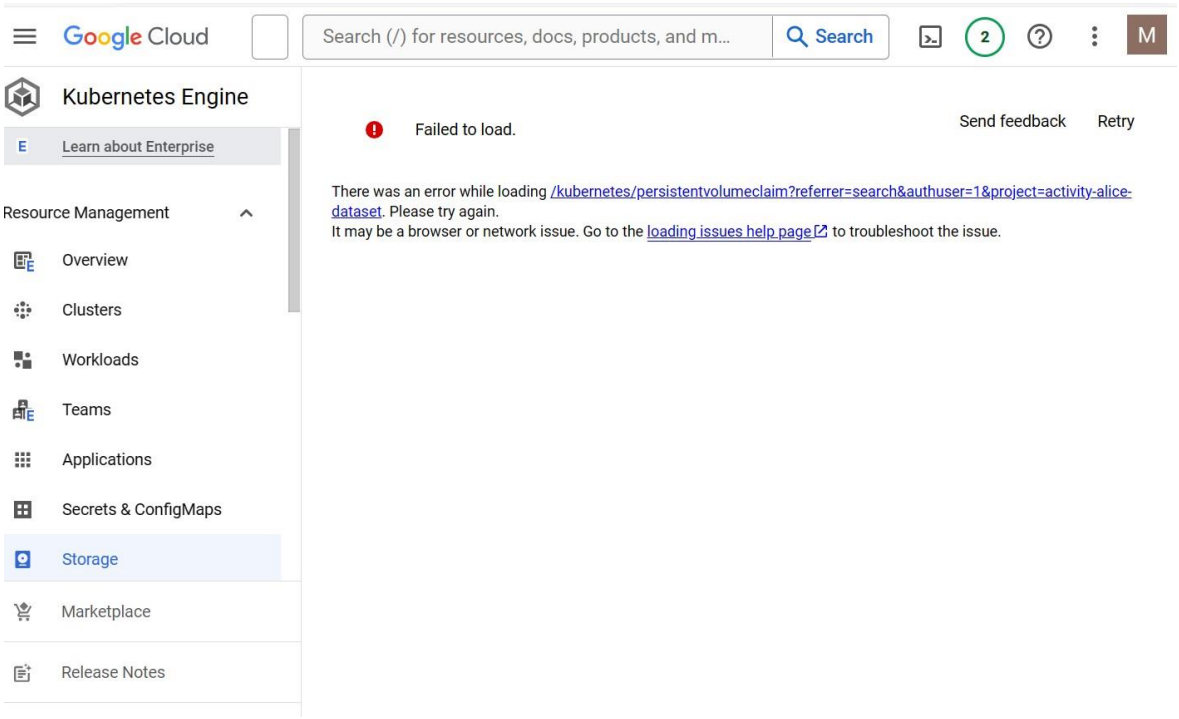


Below is the backup history of the 612_Data_Transformation_and _Management Assessment2_adidas.tidydf.

61



d) Error handling – the first attempt to do the backup using an old file was unsuccessful due to the absence of the bucket. The wrong procedure has been done, going straight to Kubernetes Engine without creating a bucket. After researching Google, I learned how to create a storage bucket and a bucket backup and followed the correct procedure for bucket creation.



References

Gillis, Alexander S., Lelii, Sonia., & Hefner, Kim.(n.d). Data Migration. Tech Target Storage. https://www.techtarget.com/searchstorage/definition/data-migration

Gitlin, Jon. (n.d). Integration and Automation: 5 Data Integration Challenges to Look Out For (and the solutions for overcoming them). Workato. https://www.workato.com/the-connector/data-integration-challenges/

The Devastator. (n.d.). Adidas Fashion Retail Products Dataset. Kaggle https://www.kaggle.com/datasets/thedevastator/adidas-fashion-retail-products-dataset-9300-prod

The Investopedia Team. (2022, November 13). Data Migration. Investopedia. https://www.investopedia.com/terms/d/data-migration.asp