



**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)

Table of Contents

Chapter 1: Introduction	3
Chapter 2: Abstract	3
Chapter 3: Dataset	4
Chapter 4: Task A – Database Implementation	5-20
Chapter 5: Task B – Data Export, Migration and Backup	21-39
References	40

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)

Torrini, Mira
To: Mohammad Norouzifard

Thank you 😊

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

Kia ora Mira,

It is approved. Good luck with your assessment.

Ngi mihi,
Mohammed

NZSE
New Zealand Skills and Education College
www.nzse.ac.nz
info@nzse.ac.nz
Phone: 022 468 6667
Head Office: 2023 Great North Road, Nairn Park, Auckland 0600, NZ
Postal Code: 0600 018 018 018

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: Torrini, Mira <764707793@nse.ac.nz>
Sent: Sunday, March 3, 2024 9:46 PM
To: Mohammad Norouzifard <mohammad@nse.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>

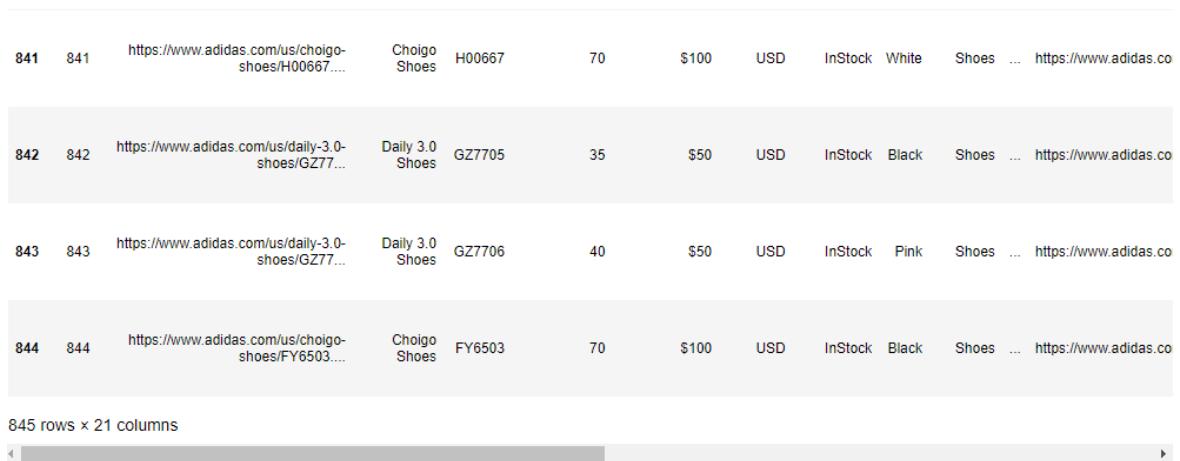
Chapter 4: Task A

- 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 x 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

index		url	name	sku	selling_price	original_price	currency	availability	color	category	...	source_website
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-lace-mountain-bike-shoes/G...	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-competition-climbing-shoes/G...	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-polo-shirt/G...	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/G...	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

b) Transforming the messy data to a tidy form

- Checking for duplicates

duplicates_all = adidas_messydf[adidas_messydf.duplicated()]																
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 ($) in original_price column
adidas_messydf['original_price'] = adidas_messydf['original_price'].str.replace("$", "")
```

index		url	name	sku	selling_price	original_price	currency	availability	color	category	...	source_website
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-lace-mountain-bike-shoes/GC7946....	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/GC7946....	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-competition-climbing-shoes/FV4744....	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-polo-shirt/GM0239....	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/FX2858....	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/GZ7705....	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/GZ7706....	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 x 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	0	https://www.adidas.com/us/beach-shorts/FJ5089...	Beach Shorts	FJ5089	40	49	USD	InStock	Black	Clothing	...	https://www.adidas.co
1	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	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	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	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	840	https://www.adidas.com/us/supernova-plus-shoes...	Supernova+ Shoes	FX2858	72	120	USD	InStock	White	Shoes	...	https://www.adidas.co
841	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	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-lace-mountain-bike-shoes/BC0770....	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/GC7946....	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-competition-climbing-shoes/FV4744....	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-polo-shirt/GM0239....	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 data
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,...
...	...
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 rows of the tidy dataset
adidas_tidydf.head (10)
```

	index	url	name	sku	selling_price	original_price	currency	availability	color	category	source	source_website
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.com
1	1	https://www.adidas.com/us/five-ten-kestrel-lace-mountain-bike-shoes	Five Ten Kestrel Lace Mountain Bike Shoes	BC0770	150	184	USD	InStock	Grey	Shoes	adidas United States	https://www.adidas.com
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.com
3	3	https://www.adidas.com/us/five-ten-hiangle-pro-competition-climbing-shoes	Five Ten Hiangle Pro Competition Climbing Shoes	FV4744	160	196	USD	InStock	Black	Shoes	adidas United States	https://www.adidas.com
4	4	https://www.adidas.com/us/mesh-broken-stripe-polo-shirt	Mesh Broken-Stripe Polo Shirt	GM0239	65	80	USD	InStock	Blue	Clothing	adidas United States	https://www.adidas.com
5	5	https://www.adidas.com/us/eqt-spikeless-golf-shoes	EQT Spikeless Golf Shoes	FX7449	110	135	USD	InStock	Grey	Shoes	adidas United States	https://www.adidas.com
6	6	https://www.adidas.com/us/adicross-hybrid-shorts	Adicross Hybrid Shorts	GM5505	80	98	USD	InStock	Black	Clothing	adidas United States	https://www.adidas.com
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.com

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 x 18 columns											

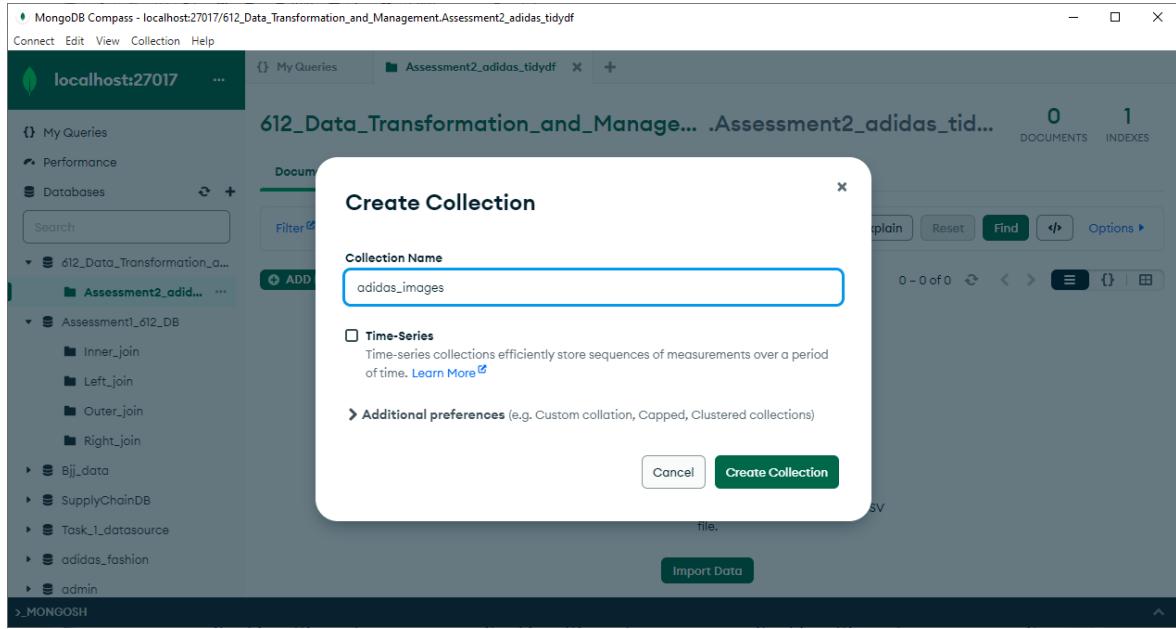
e) Creating the database using MongoDB (NoSQL)

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

The screenshot shows the MongoDB Compass interface. On the left, there's a sidebar with 'My Queries', 'Performance', 'Databases', and a search bar. The 'Databases' section lists several databases, including 'Assessment1_612_DB' which has sub-folders for various joins. In the center, a modal window titled 'Create Database' is open. It has fields for 'Database Name' (containing '612_Data_Transformation_and_Management') and 'Collection Name' (containing 'Assessment2_adidas_tidydf'). Below these fields is a checkbox for 'Time-Series' with a descriptive note about time-series collections. At the bottom of the modal are 'Cancel' and 'Create Database' buttons. The overall interface is dark-themed.

1.2 adidas_images collection



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

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

The screenshot shows the MongoDB Compass interface. The left sidebar lists databases and collections, including 'Assessment2_adidas_tidydf'. The main area displays a single document in the 'Assessment2_adidas_tidydf' collection. The document details a product: Beach Shorts, FJ5089, Black, Clothing, US, USA, etc. A red arrow points to the 'Refresh' icon in the top right corner of the document preview area, with the text 'click the refresh icon' written below it.

```

_id: ObjectId('65e6422ae0c68893a7384193')
index: 0
url: "https://www.adidas.com/us/beach-shorts/FJ5089.html"
name: "Beach Shorts"
sku: "FJ5089"
selling_price: 49
original_price: 49
currency: "USD"
availability: "InStock"
color: "Black"
category: "Clothing"
source: "adidas United States"
source_website: "https://www.adidas.com"
breadcrumbs: "Women/Clothing"
description: "Splashing in the surf. Making memories with your friends. Beach days a..."
brand: "adidas"
country: "USA"
  
```

Created database and columns.

The screenshot shows the MongoDB Compass interface. The left sidebar lists databases and collections, including '612_Data_Transformation_and_Management'. The main area displays storage statistics for two collections: 'adidas_images' and 'Assessment2_adidas_tidydf'. Both collections have a storage size of approximately 380 kB, 6.5 K documents, and 1 index, resulting in a total index size of about 81 kB and 24 kB respectively.

Collection	Storage size	Documents	Avg. document size	Indexes	Total index size
adidas_images	380.93 kB	6.5 K	194.00 B	1	81.92 kB
Assessment2_adidas_tidydf	249.86 kB	845	859.00 B	1	24.58 kB

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

	category String	source String	source_website String	breadcrumbs String	description String	brand String	country String
1	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"Splashing in the surf. ..."	"adidas"	"USA"
2	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"Lace up and get after it. ..."	"adidas"	"USA"
3	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Kids/Clothing"	"Clean and crisp, this au..."	"adidas"	"USA"
4	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Five Ten/Shoes"	"The Htangle Pro takes ou..."	"adidas"	"USA"
5	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Men/Clothing"	"Step up to the tee rela..."	"adidas"	"USA"
6	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Women/Shoes"	"Put comfort first. Then..."	"adidas"	"USA"
7	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Kids/Clothing"	"On the golf course, com..."	"adidas"	"USA"
8	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"You can't always be a fu..."	"adidas"	"USA"
9	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"You can show your conce..."	"adidas"	"USA"
10	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Men/Clothing"	"You can't always be a fu..."	"adidas"	"USA"
11	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"Sometimes confidence cou..."	"adidas"	"USA"
12	"Accessories"	"adidas United States"	"https://www.adidas.com"	"Men/Accessories"	"Stop searching for the u..."	"adidas"	"USA"
13	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Kids/Clothing"	"The future's a blank ca..."	"adidas"	"USA"
14	"Accessories"	"adidas United States"	"https://www.adidas.com"	"Men/Accessories"	"To get serious about tr..."	"adidas"	"USA"
15	"Accessories"	"adidas United States"	"https://www.adidas.com"	"Men/Accessories"	"Whether you prefer to du..."	"adidas"	"USA"
16	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"What's on the agenda? W..."	"adidas"	"USA"
17	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"What's on the agenda? W..."	"adidas"	"USA"
18	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Kids/Shoes"	"Ready to shine. Whether..."	"adidas"	"USA"
19	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"What's on the agenda? W..."	"adidas"	"USA"
20	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"Sometimes confidence cou..."	"adidas"	"USA"

*After

	category String	source String	source_website String	breadcrumbs String	description String	brand String	country String
1	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"Splashing in the surf. ..."	"adidas"	"USA"
2	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Kids/Clothing"	"Clean and crisp, this au..."	"adidas"	"USA"
3	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Men/Clothing"	"Step up to the tee rela..."	"adidas"	"USA"
4	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Men/Clothing"	"On the golf course, com..."	"adidas"	"USA"
5	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Kids/Clothing"	"You can't always be a fu..."	"adidas"	"USA"
6	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"You can show your conce..."	"adidas"	"USA"
7	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Men/Clothing"	"You can't always be a fu..."	"adidas"	"USA"
8	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"Sometimes confidence cou..."	"adidas"	"USA"
9	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Kids/Clothing"	"The future's a blank ca..."	"adidas"	"USA"
10	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"What's on the agenda? W..."	"adidas"	"USA"
11	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"What's on the agenda? W..."	"adidas"	"USA"
12	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"What's on the agenda? W..."	"adidas"	"USA"
13	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"Sometimes confidence cou..."	"adidas"	"USA"
14	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"What's on the agenda? W..."	"adidas"	"USA"
15	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"If ease is what you're ..."	"adidas"	"USA"
16	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"Every day brings a new ..."	"adidas"	"USA"
17	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"There are a lot of thin..."	"adidas"	"USA"
18	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"Every day brings a new ..."	"adidas"	"USA"
19	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Men/Clothing"	"There are times to go h..."	"adidas"	"USA"
20	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Women/Clothing"	"Dash out the door feeli..."	"adidas"	"USA"

- Another sample

The screenshot shows the MongoDB Compass interface with a query results table titled "Assessment2_adidas_tidydf". The table has 845 documents and 1 index. The columns include: string, category, String, source, String, source_website, String, breadcrumbs, String, description, String, brand, String, and country, String. The data shows various shoe models from Adidas, such as "adidas United States", "Five Ten/Shoes", "Hiking Pro", etc., with descriptions like "Lace up and get after it.", "Put comfort first. Then lace up and get after it.", and "Ready to shine, whether you're on the court or the street." The table also includes columns for MaxTimeMS (40000), Skip (0), and Limit (0).

string	category	String	source	String	source_website	String	breadcrumbs	String	description	String	brand	String	country
1	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Women/Shoes"	"https://www.adidas.com"	"Lace up and get after it."	"adidas"	"USA"	"Lace up and get after it."	"adidas"	"USA"	"USA"	
2	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Five Ten/Shoes"	"https://www.adidas.com"	"Put comfort first. Then lace up and get after it."	"adidas"	"USA"	"Put comfort first. Then lace up and get after it."	"adidas"	"USA"	"USA"	
3	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Women/Shoes"	"https://www.adidas.com"	"Ready to shine, whether you're on the court or the street."	"adidas"	"USA"	"Ready to shine, whether you're on the court or the street."	"adidas"	"USA"	"USA"	
4	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Adidas/Shoes"	"https://www.adidas.com"	"Ready to shine, whether you're on the court or the street."	"adidas"	"USA"	"Ready to shine, whether you're on the court or the street."	"adidas"	"USA"	"USA"	
5	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Soccer/Shoes"	"https://www.adidas.com"	"You've mutated to give you more."	"adidas"	"USA"	"You've mutated to give you more."	"adidas"	"USA"	"USA"	
6	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Men/Shoes"	"https://www.adidas.com"	"For quick morning miles."	"adidas"	"USA"	"For quick morning miles."	"adidas"	"USA"	"USA"	
7	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Men/Shoes"	"https://www.adidas.com"	"Simplicity, versatility."	"adidas"	"USA"	"Simplicity, versatility."	"adidas"	"USA"	"USA"	
8	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Women/Shoes"	"https://www.adidas.com"	"When you hit the pool, ..."	"adidas"	"USA"	"When you hit the pool, ..."	"adidas"	"USA"	"USA"	
9	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Men/Shoes"	"https://www.adidas.com"	"In the cage. On the court. In the cage. On the court."	"adidas"	"USA"	"In the cage. On the court. In the cage. On the court."	"adidas"	"USA"	"USA"	
10	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Men/Shoes"	"https://www.adidas.com"	"Trends change, but feet..."	"adidas"	"USA"	"Trends change, but feet..."	"adidas"	"USA"	"USA"	
11	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Kids/Shoes"	"https://www.adidas.com"	"For the toddler in you..."	"adidas"	"USA"	"For the toddler in you..."	"adidas"	"USA"	"USA"	
12	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Essentials/Shoes"	"https://www.adidas.com"	"You never blend in. So..."	"adidas"	"USA"	"You never blend in. So..."	"adidas"	"USA"	"USA"	
13	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Originals/Shoes"	"https://www.adidas.com"	"The ZX stepped onto the..."	"adidas"	"USA"	"The ZX stepped onto the..."	"adidas"	"USA"	"USA"	
14	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Men/Shoes"	"https://www.adidas.com"	"Simplicity, versatility."	"adidas"	"USA"	"Simplicity, versatility."	"adidas"	"USA"	"USA"	
15	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Women/Shoes"	"https://www.adidas.com"	"ZXience expands the ZX..."	"adidas"	"USA"	"ZXience expands the ZX..."	"adidas"	"USA"	"USA"	
16	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Women/Shoes"	"https://www.adidas.com"	"These adidas shoes bring..."	"adidas"	"USA"	"These adidas shoes bring..."	"adidas"	"USA"	"USA"	
17	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Men/Shoes"	"https://www.adidas.com"	"In the '80s, the ZX served..."	"adidas"	"USA"	"In the '80s, the ZX served..."	"adidas"	"USA"	"USA"	
18	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Women/Shoes"	"https://www.adidas.com"	"ZXience expands the ZX..."	"adidas"	"USA"	"ZXience expands the ZX..."	"adidas"	"USA"	"USA"	
19	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Men/Shoes"	"https://www.adidas.com"	"Look fast. Feel fast. Be..."	"adidas"	"USA"	"Look fast. Feel fast. Be..."	"adidas"	"USA"	"USA"	
20	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Men/Shoes"	"https://www.adidas.com"	"Peer into pure satisfaction."	"adidas"	"USA"	"Peer into pure satisfaction."	"adidas"	"USA"	"USA"	

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

The screenshot shows the MongoDB Compass interface with a query results table titled "Assessment2_adidas_tidydf". The table has 845 documents and 1 index. The columns include: e, String, breadcrumbs, String, description, String, brand, String, country, String, language, String, average_rating, Double, and reviews_count, Int32. The data shows various women's clothing items from Adidas, such as "adidas.com", "Men/Clothing", "Hurricane pride looks...", "adidas", "USA", "en", 1, 1. The table also includes columns for MaxTimeMS (40000), Skip (0), and Limit (0).

e	String	breadcrumbs	String	description	String	brand	String	country	String	language	String	average_rating	Double	reviews_count	Int32
1	"adidas.com"	"Men/Clothing"	"Hurricane pride looks..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	1	1	1	1	1
2	"adidas.com"	"Men/Clothing"	"It's all how you look at..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	3	3	2	2	2
3	"adidas.com"	"Women/Clothing"	"From ice skating to col..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	3.7	3	3	3	3
4	"adidas.com"	"Women/Clothing"	"From ice skating to col..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	3.7	3	3	3	3
5	"adidas.com"	"Women/Clothing"	"Lunge, twist and stretc..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	3.7	20	20	20	20
6	"adidas.com"	"Women/Clothing"	"Lunge, twist and stretc..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	3.7	20	20	20	20
7	"adidas.com"	"Women/Clothing"	"These adidas maternity ..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	3.7	134	134	134	134
8	"adidas.com"	"Women/Clothing"	"These adidas maternity ..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	3.9	134	134	134	134
9	"adidas.com"	"Women/Clothing"	"Sweat it out in style w..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	4	11	11	11	11
10	"adidas.com"	"Kids/Clothing"	"Celebrate female empow..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	4	2	2	2	2
11	"adidas.com"	"Women/Clothing"	"You deserve comfort, an..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	4	6	6	6	6
12	"adidas.com"	"Women/Clothing"	"Crush leg day in these ..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	4	25	25	25	25
13	"adidas.com"	"Kids/Clothing"	"You don't have to do or..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	4	1	1	1	1
14	"adidas.com"	"Women/Clothing"	"Sport isn't a time or p..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	4	4	4	4	4
15	"adidas.com"	"Women/Clothing"	"Nearly light as air. Th..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	4	2	2	2	2
16	"adidas.com"	"Women/Clothing"	"Get the best of ease an..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	4	2	2	2	2
17	"adidas.com"	"Women/Clothing"	"It's all in the details..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	4	1	1	1	1
18	"adidas.com"	"Women/Clothing"	"It doesn't matter what w..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	4	2	2	2	2
19	"adidas.com"	"Women/Clothing"	"Get the best of ease an..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	4	2	2	2	2
20	"adidas.com"	"Women/Clothing"	"Don't save these adida..."	"adidas"	"USA"	"adidas"	"USA"	"USA"	"en"	"en"	4.1	14	14	14	14

- Descending

The screenshot shows the MongoDB Compass interface with the following details:

- Project:** { field: 0 }
- Sort:** { average_rating:-1}
- Collection:** {}
- Documents:** 845
- Indexes:** 1
- MaxTimeMS:** 60000
- Limit:** 0
- Table Headers:** e String, breadcrumb String, description String, brand String, country String, language String, average_rating Double, reviews_count Int32
- Table Data:** A grid of 20 rows showing document details. For example, row 1: e="adidas.com", breadcrumb="Men/Clothing", description="There are a lot of thin...", brand="adidas", country="USA", language="en", average_rating=5.0, reviews_count=21.

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

The screenshot shows the MongoDB Compass interface with the following details:

- Collection:** Assessment2_adidas_tidydf
- adidas_images:**
 - Storage size:** 300.93 kB
 - Documents:** 654
 - Avg. document size:** 194.00 B
 - Indexes:** 1
 - Total index size:** 81.92 kB
- Assessment2_adidas_tidydf:**
 - Storage size:** 249.66 kB
 - Documents:** 845
 - Avg. document size:** 859.00 B
 - Indexes:** 1
 - Total index size:** 24.50 kB

j) Performing grouping operation.

1. Click the name of the collection; Assessment2_adidas_tidydf

The screenshot shows the MongoDB Compass interface with the following details:

- Collection:** Assessment2_adidas_tidydf
- Aggregations:** Pipeline (empty)
- Preview:** Stage 1: \$group
- Stages:** Stage disabled. Results not passed in the pipeline.
- WIZARD:**

The preview shows three sample documents from the collection:

```

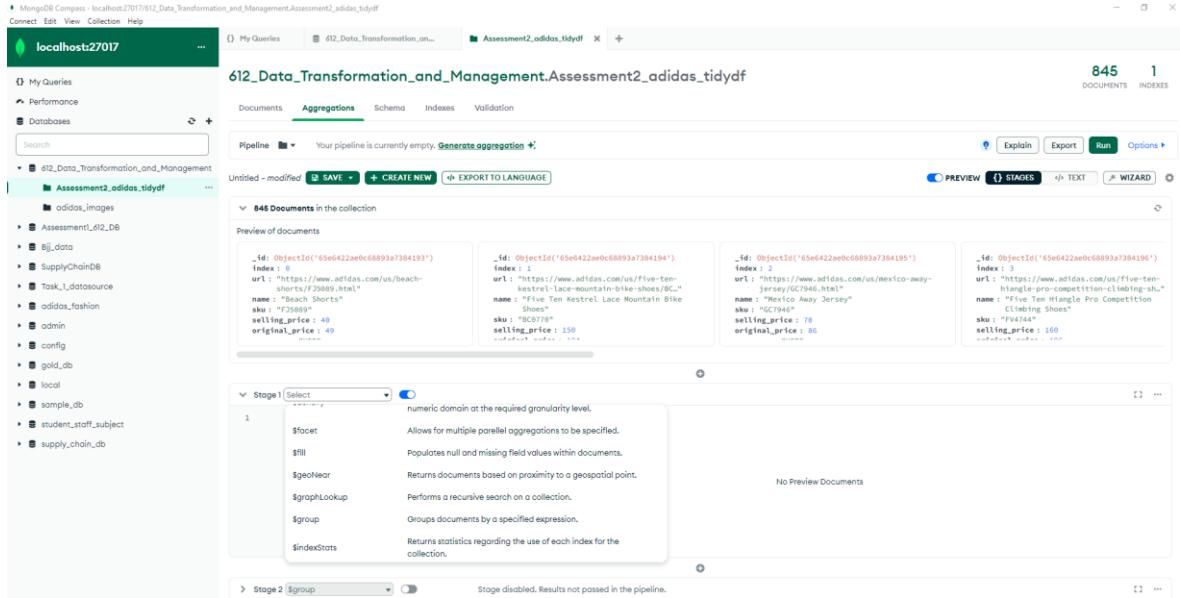
{
  "_id": ObjectId("65e6422ae0c68893a7384193"),
  "Index": 0,
  "url": "https://www.adidas.com/us/beach-shorts/f35689.html",
  "name": "Five Ten Shorts",
  "sku": "F75689",
  "selling_price": 40,
  "original_price": 49
}

{
  "_id": ObjectId("65e6422ae0c68893a7384194"),
  "Index": 1,
  "url": "https://www.adidas.com/us/five-ten-kestrel-lace-mountain-bike-shoes/BC_",
  "name": "Five Ten Kestrel Lace Mountain Bike Shoes",
  "sku": "BC8778",
  "selling_price": 150
}

{
  "_id": ObjectId("65e6422ae0c68893a7384195"),
  "Index": 2,
  "url": "https://www.adidas.com/us/five-ten-climbing-shoes/go-away-jersey",
  "name": "Five Ten Hangle Pro Competition Climbing Shoes",
  "sku": "FV4744",
  "selling_price": 160
}

```

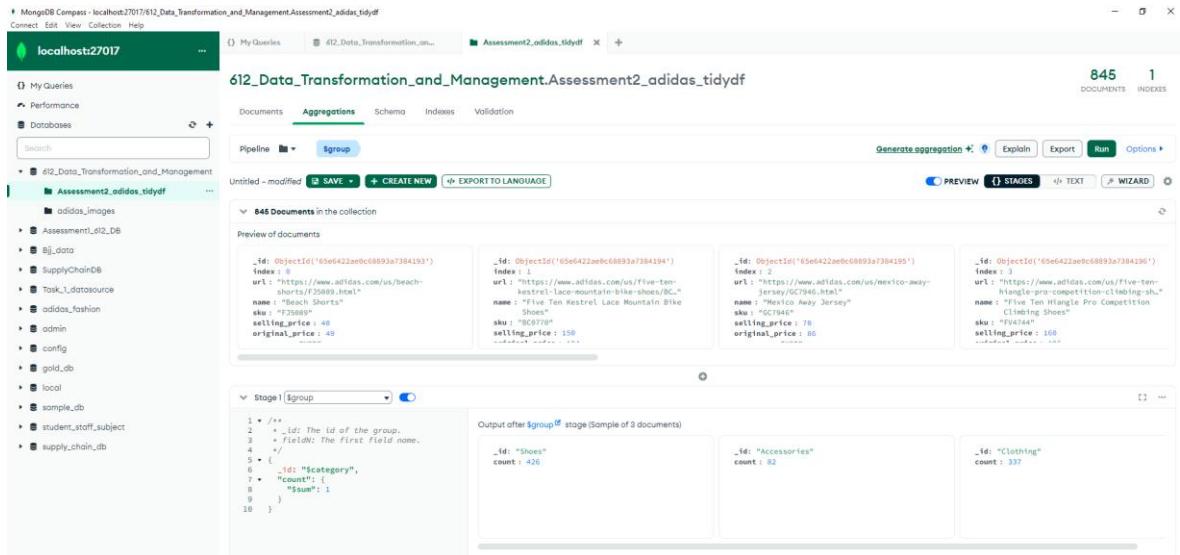
2. In the Aggregations Tab, click + Add Stage and select \$group to aggregate the collection.
Make sure to enable the Stage (using this button):  to pass the result in the pipeline.



The screenshot shows the MongoDB Compass interface with the 'Aggregations' tab selected for the 'Assessment2_adidas_tidydf' collection. The pipeline is empty. A dropdown menu under 'Stage 1' is open, showing options like '\$facet', '\$fill', '\$geoNear', '\$graphLookup', '\$group', and '\$indexStats'. The '\$group' option is highlighted with a blue border, indicating it is the current stage being configured. The preview pane shows four document samples from the collection.

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

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



The screenshot shows the MongoDB Compass interface with the 'Aggregations' tab selected for the 'Assessment2_adidas_tidydf' collection. The '\$group' stage is selected and highlighted with a blue border. The preview pane shows four document samples from the collection. The output after the '\$group' stage is shown as a sample of 3 documents, which includes the grouped data.

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'}.

string	availability String	color String	category String	source String	source_website String	breadcrumbs String	description
1	"OutOfStock"	"Black"	"Clothing"	"adidas United States"	"https://www.adidas.com"	"Kids/Clothing"	"Little one ✓ 4 9 9 9 9"
2	"OutOfStock"	"Pink"	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Women/Shoes"	"Since the ✓ 4 9 9 9 9"
3	"OutOfStock"	"Black"	"Shoes"	"adidas United States"	"https://www.adidas.com"	"Women/Shoes"	"These add ✓ 4 9 9 9 9"

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

```
{ availability: 'OutOfStock' }
```

```
1 ▾ {  
2 ▾   $set: {  
3     availability: 'InStock'  
4   },  
5 }
```

Save **Cancel** **Update 3 documents**

Click Update 3 documents and refresh the page.

MongoDB Compass - localhost:27017/612_Data_Transformation_and_Management.Assessment2_adidas_tidydf

Connect Edit View Collection Help

localhost:27017 ...

My Queries d12_Data_Transformation_on... Assessment2_adidas_tidydf +

612_Data_Transformation_and_Management.Assessment2_adidas_tidydf

Documents Aggregations Schema Indexes Validation

Filter (eval:availability != "OutofStock")

Project { field: 0 }

Sort { field: -1 } or [{ 'field': -1 }]

MaxTimeMS 60000

Collation {}

Skip 0 Limit 0

Generate query Explain Reset Find Options ▾

ADD DATA EXPORT DATA UPDATE DELETE

1-3 of 3

	_id	objectid	index	Int32	url	String	name	String	sku	String	selling_price	Int32	original_price	Int32
1	ObjectID('65e6422ae8c688...	239	"https://www.adidas.com/...	"Originals x Kevin Lyons...	"H22615"	44	55							
2	ObjectID('65e6422ae8c688...	459	"https://www.adidas.com/...	"NMD_R1 Spectoo Shoes"	"FZ3288"	75	150							
3	ObjectID('65e6422ae8c688...	569	"https://www.adidas.com/...	"Racer TR21 Shoes"	"GS5776"	53	75							

REFRESH

3 documents have been updated.

845 DOCUMENTS INDEXES 1

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.

MongoDB Compass - localhost:27017/612_Data_Transformation_and_Management.Assessment2_adidas_tidydf

Connect Edit View Collection Help

localhost:27017 ...

My Queries

Performance

Databases

612_Data_Transformation_and_Management

Assessment2_adidas_tidydf

adidas_images

Assessment1_d12_DB

8jJ_db

SupplyChainDB

Task_L1_datasource

adidas_fashion

admin

config

gold_db

local

sample_db

student_staff_subject

supply_chain_db

My Queries

612_Data_Transformation_and_Management

Assessment2_adidas_tidydf

Documents Aggregations Schema Indexes Validation

Filter (Availability: 'OutOfStock')

Project (field: 0)

Sort (field: -1) or (['field', -1])

Collation ()

Generate query Explain Reset Find Options

MaxTimeMS 40000

Skip 0 Limit 0

ADD DATA EXPORT DATA UPDATE DELETE

No results

Try modifying your query to get results.

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.

MongoDB Compass - localhost:27017/612_Data_Transformation_and_Management.adidas_tidydf

localhost:27017

My Queries Performance Databases

612_Data_Transformation_and_Management.adidas_tidydf

845 1 DOCUMENTS INDEXES

Documents Aggregations Schema Indexes Validation

Filter Type a query: { field: 'value' } or **Generate query**

Project { field: 0 }

Sort { field: -1 } or { field: -1 }

Collection { locale: 'simple' }

EXPORT DATA Export query results Export the full collection

_id	ObjectID	url	String	name	String	sku	String	selling_price	Int32	original_price	Int32
1	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Beach Shorts"		"F36899"		49	49		
2	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Five Ten Kestrel Lace Mu..		"GC8778"		158	184		
3	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Mexico Away Jersey"		"GC7946"		79	86		
4	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Five Ten Hangle Pro Co..		"FY4744"		108	136		
5	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Mesh Broken-Stripe Polo..		"G08239"		65	88		
6	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"EQT Spikless Golf Shoe..		"FX7449"		119	135		
7	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Adicross Hybrid Shorts"		"G05305"		89	98		
8	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Tiro 21 Windbreaker"		"GP4879"		68	74		
9	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Classic 3-Stripes Swims..		"F33923"		49	49		
10	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Tiro 21 Windbreaker"		"GP4879"		65	89		
11	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Formation Sculpt Blkr ..		"G01127"		60	74		
12	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Athletic Cushioned Crew..		"B93219"		29	25		
13	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Inter Miami Cf Home Aut..		"H08029"		79	86		
14	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Cushioned Mid-Crew Sock..		"C05831"		14	18		
15	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Tour Camo-Print Hat"		"G04799"		39	37		
16	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Essentials Loose Logo T..		"H07758"		29	25		
17	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Essentials Loose Logo T..		"H07757"		29	25		
18	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Superstar Shoes"		"F31434"		59	62		
19	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Essentials Loose Logo T..		"H07756"		29	25		
20	ObjectID('65e6422ae0c68893a73844e01')	"https://www.adidas.com/..		"Formation Sculpt Tights"		"G09137"		48	89		

MongoDB Compass - localhost:27017/612_Data_Transformation_and_Management.adidas_images

localhost:27017

My Queries Performance Databases

612_Data_Transformation_and_Management.adidas_images

6.5k 1 DOCUMENTS INDEXES

Documents Aggregations Schema Indexes Validation

Filter Type a query: { field: 'value' } or **Generate query**

Project { field: 0 }

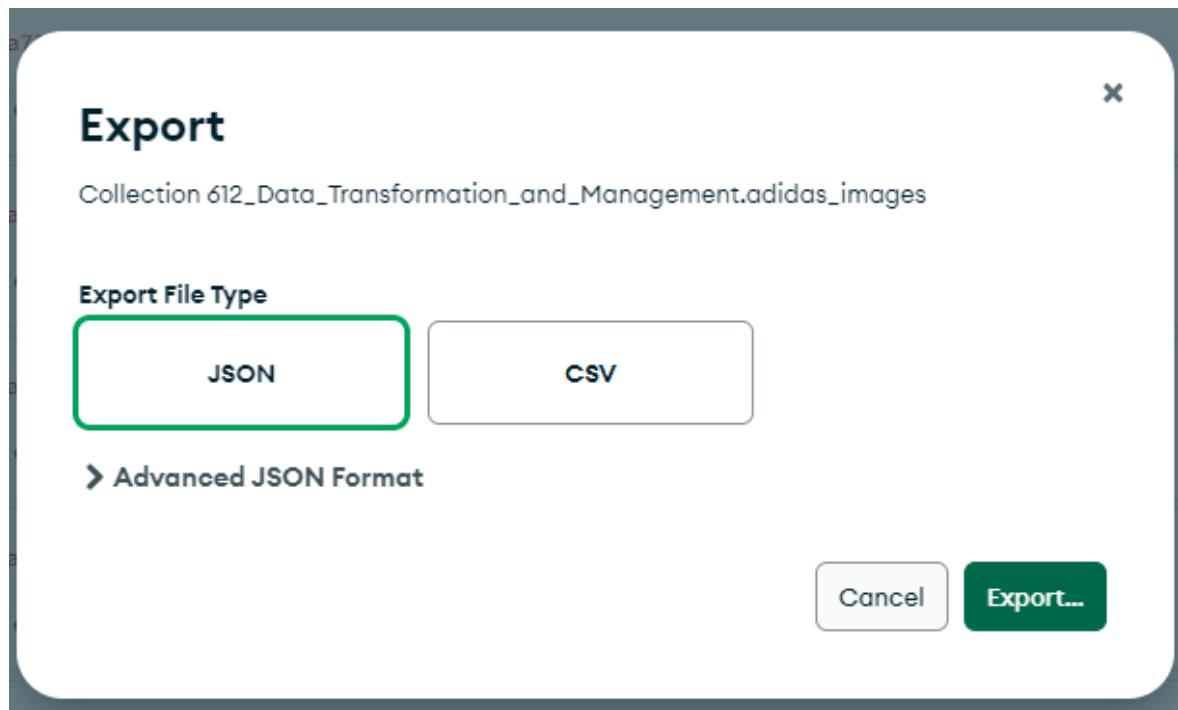
Sort { field: -1 } or { field: -1 }

Collection { locale: 'simple' }

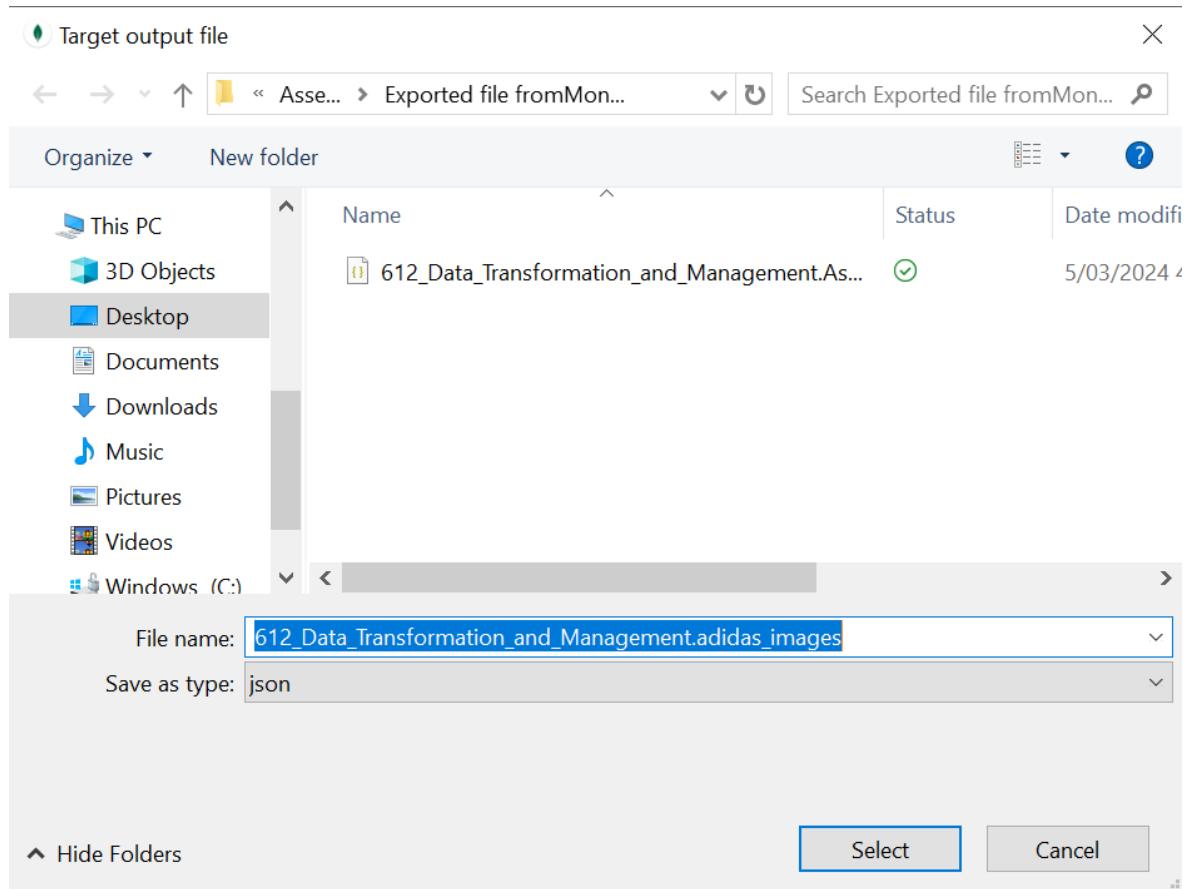
EXPORT DATA Export query results Export the full collection

_id	ObjectID	sku	Images
1	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
2	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
3	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
4	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
5	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
6	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
7	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
8	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
9	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
10	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
11	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
12	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
13	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
14	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
15	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
16	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
17	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
18	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
19	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."
20	ObjectID('65e6422ae0c68893a73844e01')	"F36899"	"images/w_600,f_auto,q_auto,adc4c860bb0d841e8a.."

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.

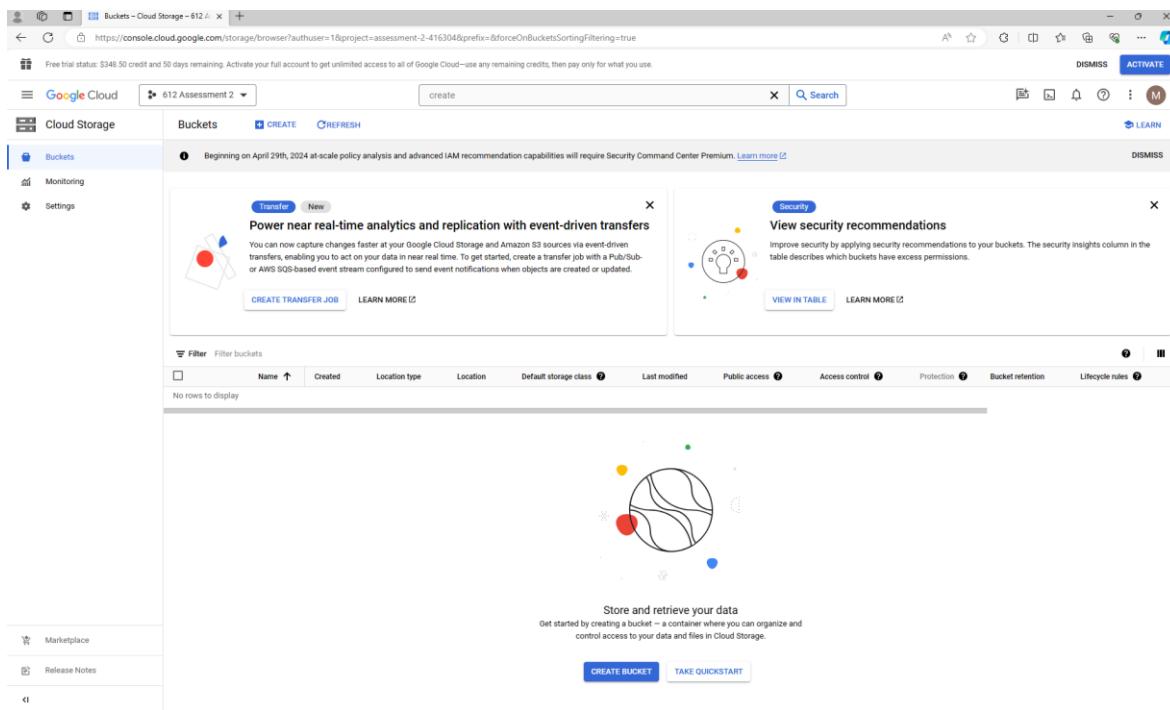
You will be notified once you have successfully created a project on the upper right side of the page.

The screenshot shows the Google Cloud Platform dashboard for the project '707 Activity Alice Dataset'. The left sidebar lists pinned products: APIs & Services, Billing, IAM & Admin, Marketplace, Compute Engine, Kubernetes Engine, Cloud Storage, BigQuery, VPC network, Cloud Run, SQL, Security, and Google Maps Platform. The main dashboard area has sections for Project info, Compute Engine (CPU %), API APIs (Requests (requests/sec)), and Getting Started. On the right, there is a Notifications sidebar showing activity logs such as 'Create Project: 612 Assessment 2' and 'Create VM instance "Instance-20240303-204951" and its boot disk "Instance-20240303-204951"'.

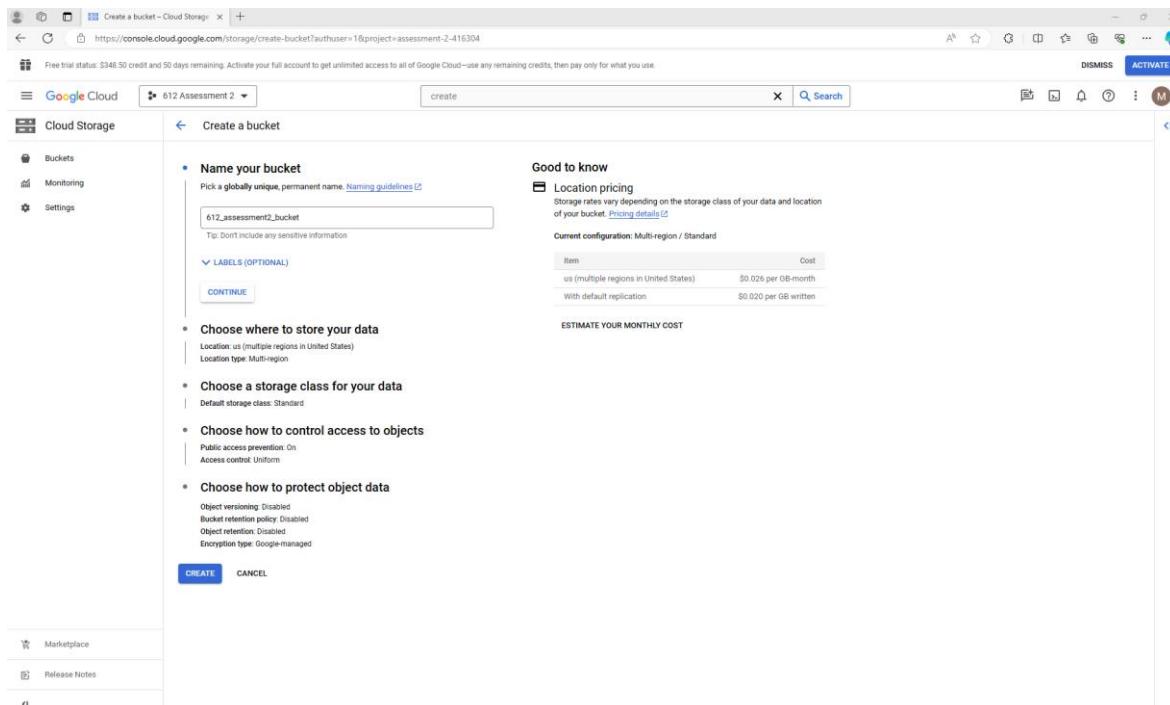
2. After creating the project, you can now create a bucket. From the navigation menu, go to the cloud storage and select Buckets.

The screenshot shows the Google Cloud Platform dashboard for the project '612 Assessment 2'. The left sidebar shows the 'Cloud Storage' section selected, with 'Buckets' highlighted. A modal window at the bottom of the screen displays the message 'Now viewing project "612 Assessment 2" in organization "No organization"'. The main dashboard area includes sections for Project info, API APIs, Google Cloud Platform status, Billing, Monitoring, and Error Reporting.

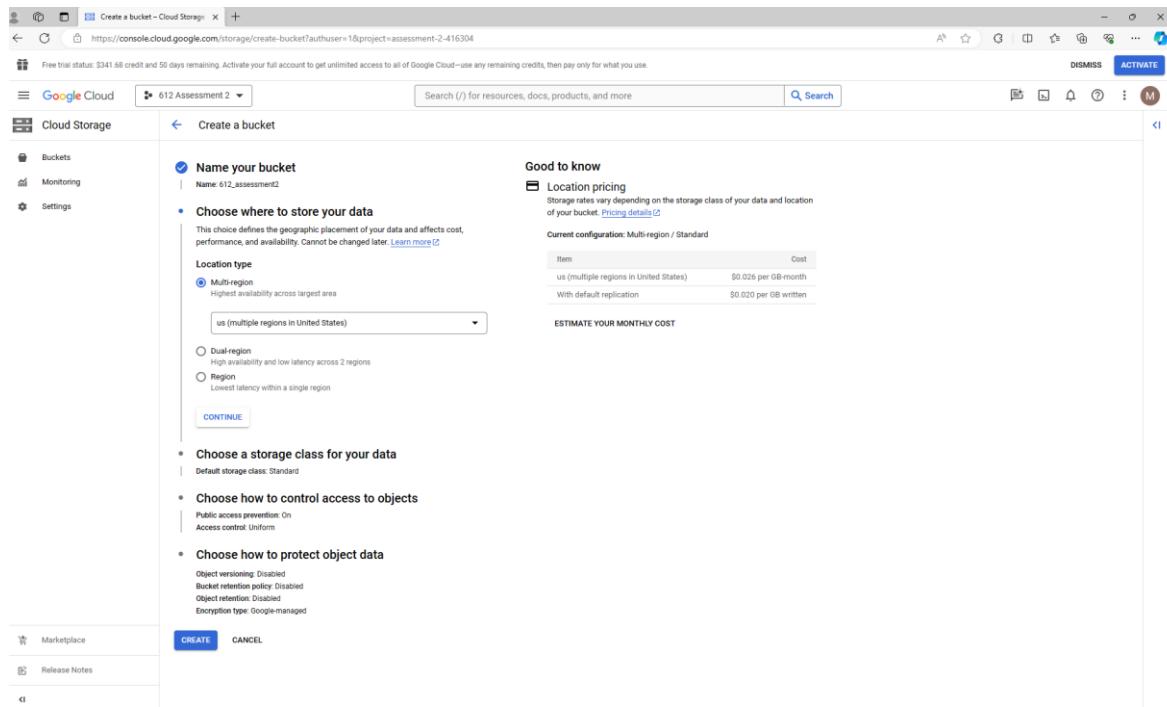
Click CREATE BUCKET



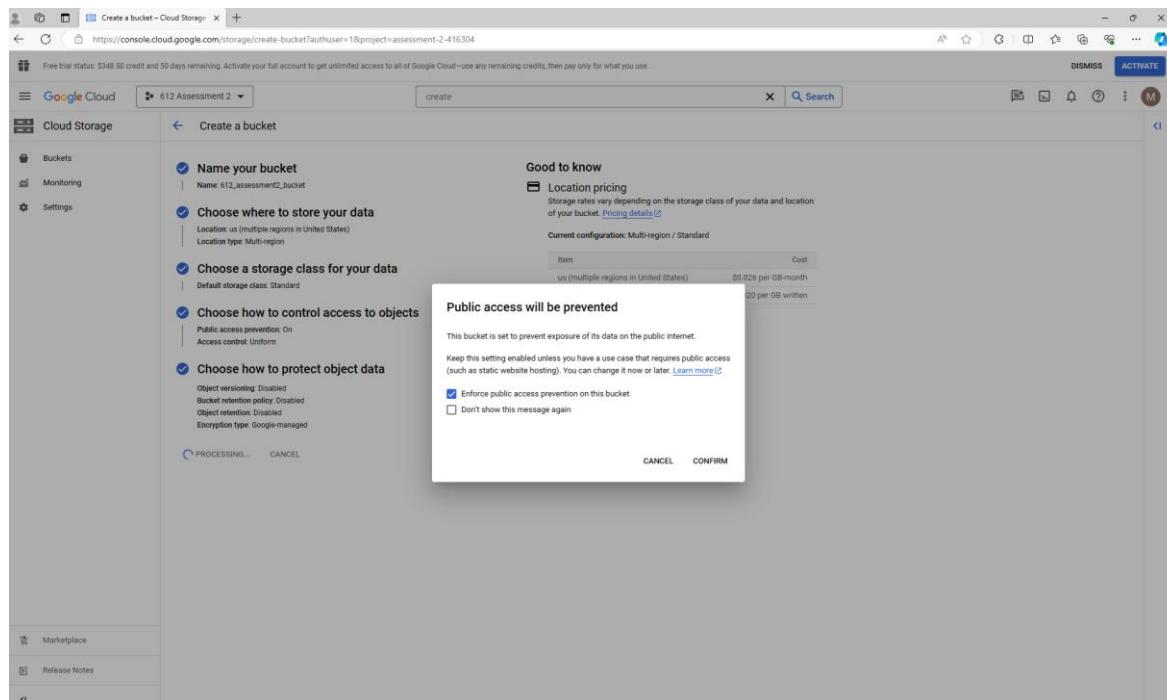
Name your bucket and click CONTINUE.



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:

The screenshot shows the Google Cloud Storage console with a newly created bucket named "612_assessment2_bucket". The bucket details page is displayed, showing basic information such as location (us), storage class (Standard), and public access (Not public). The objects tab is empty, displaying a large plus sign icon. A modal window at the bottom right says "Created bucket 612_assessment2_bucket".

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

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

The screenshot shows the Google Cloud Storage console with the "UPLOAD FILES" interface. A file selection dialog is open, showing a folder path "612_Data_Transformation_and_Management.xls". The dialog includes fields for "File name" (set to "612_Data_Transformation_and_Management.xls"), "Upload from mobile" (checkbox), and "Open" (button).

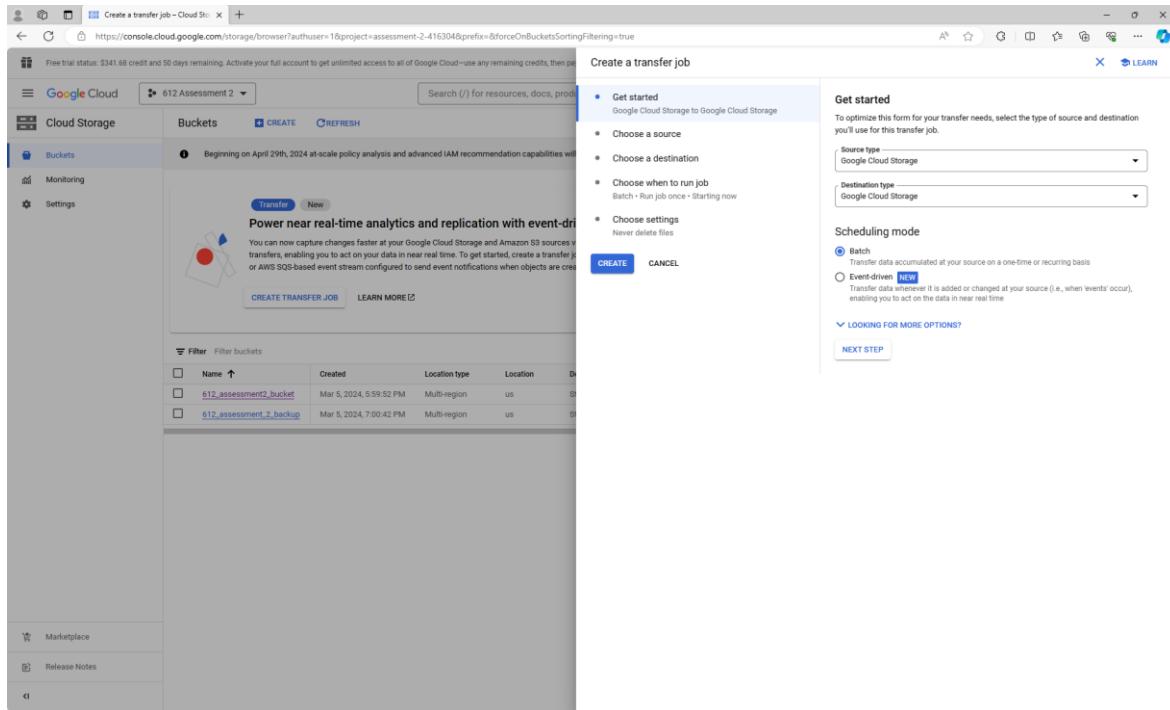
Below is the data uploaded from my local computer to the GCP bucket. I put it in the 612_assessment2_bucket collection.

Name	Type	Created	Storage class	Last modified	Public access	Encryption
612_Data_Transformation_and_Management_Adress_toJSON_1.json	application/json	Mar 5, 2024, 6:05:41 PM	Standard	Mar 5, 2024, 6:05:41 PM	Not public	Google-managed
612_Data_Transformation_and_Management_Adress_toJSON_2.json	application/json	Mar 5, 2024, 6:05:41 PM	Standard	Mar 5, 2024, 6:05:39 PM	Not public	Google-managed

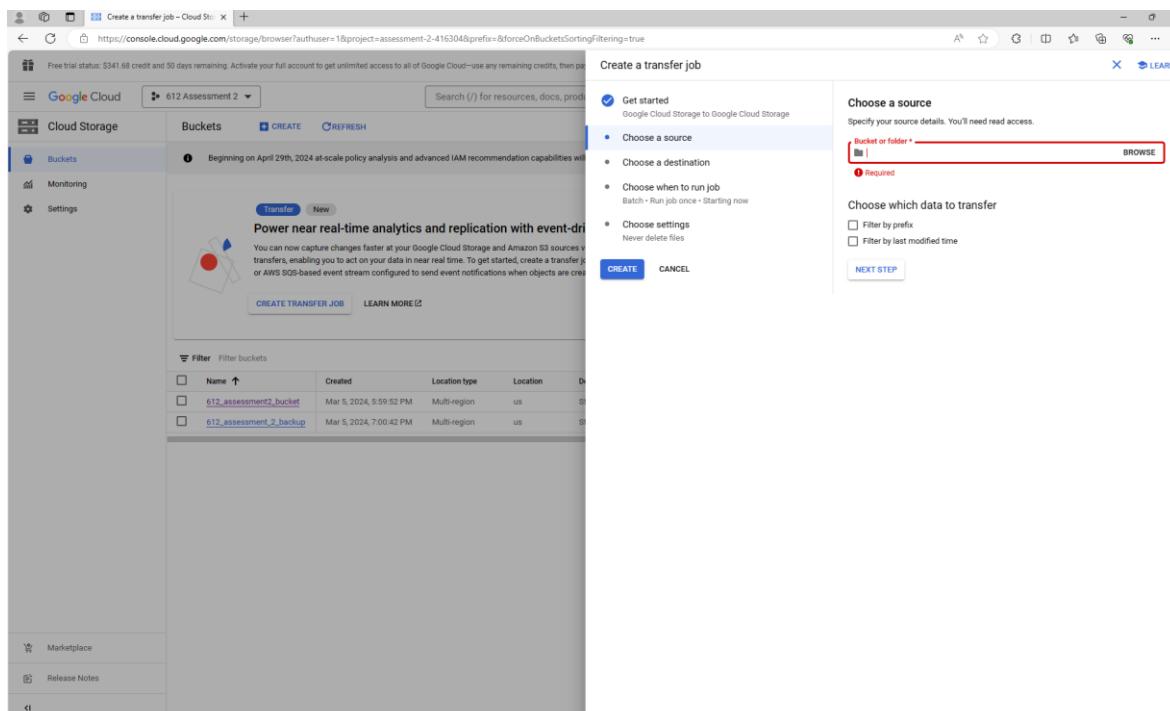
c. Following the same procedure in creating a bucket, I created another bucket to be used as backup and named it 612_assessment_2_backup.

Name	Created	Location type	Location	Default storage class	Last modified	Public access	Access control	Protection	Bucket retention
612_assessment2_bucket	Mar 5, 2024, 5:59:52 PM	Multi-region	us	Standard	Mar 5, 2024, 7:02:00 PM	Not public	Uniform	None	None
612_assessment_2_backup	Mar 5, 2024, 7:00:42 PM	Multi-region	us	Standard	Mar 5, 2024, 7:02:01 PM	Not public	Uniform	None	None

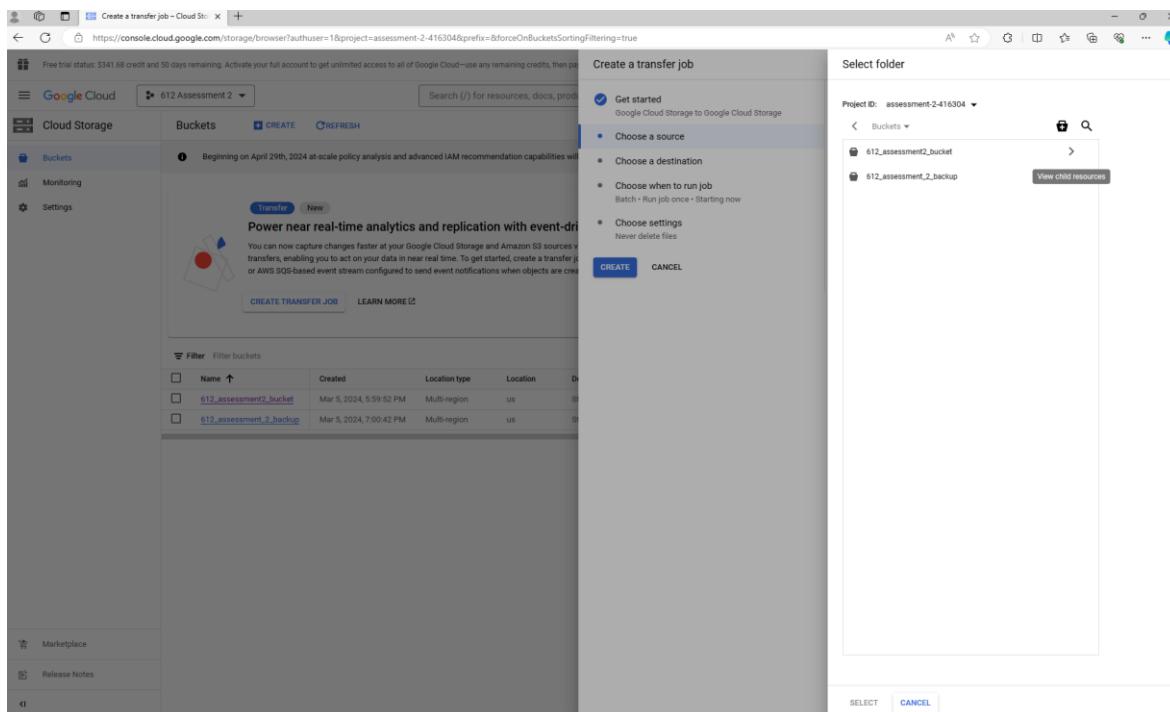
After creating the backup bucket, click the CREATE TRANSFER JOB



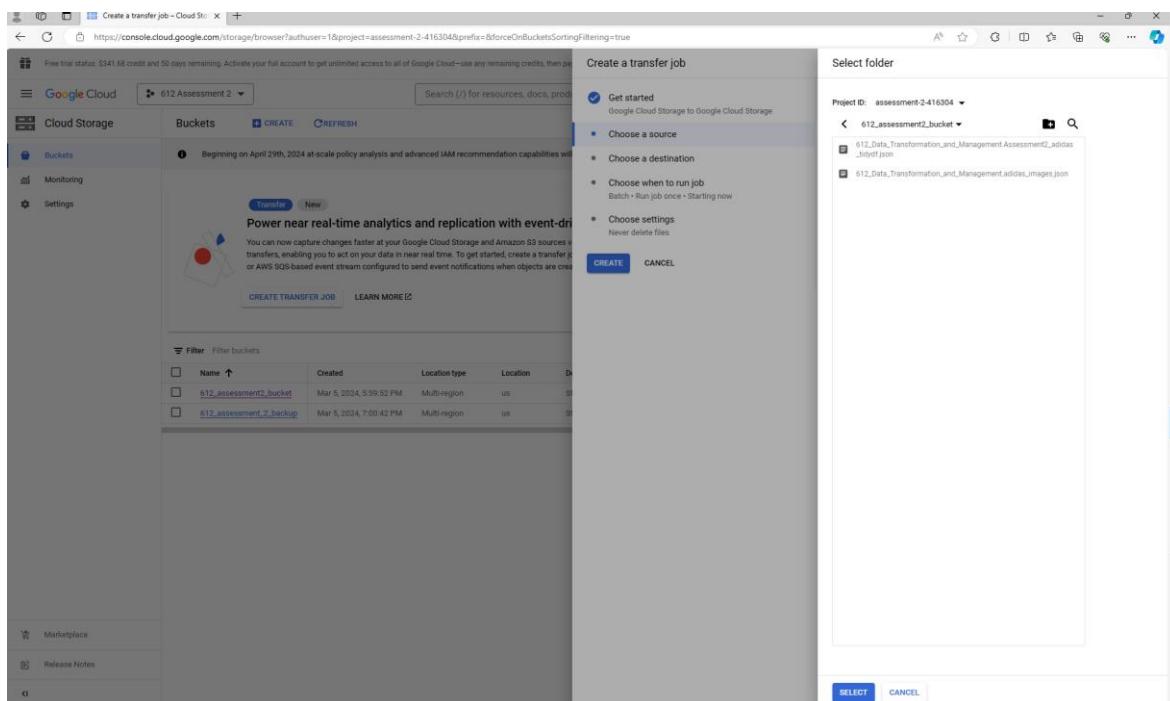
My data source and backup are in the same Google Cloud Storage, so I put Google Cloud Storage in Source type and Destination type and choose Batch in scheduling mode. Click NEXT STEP.



Choose a source to be backed up.

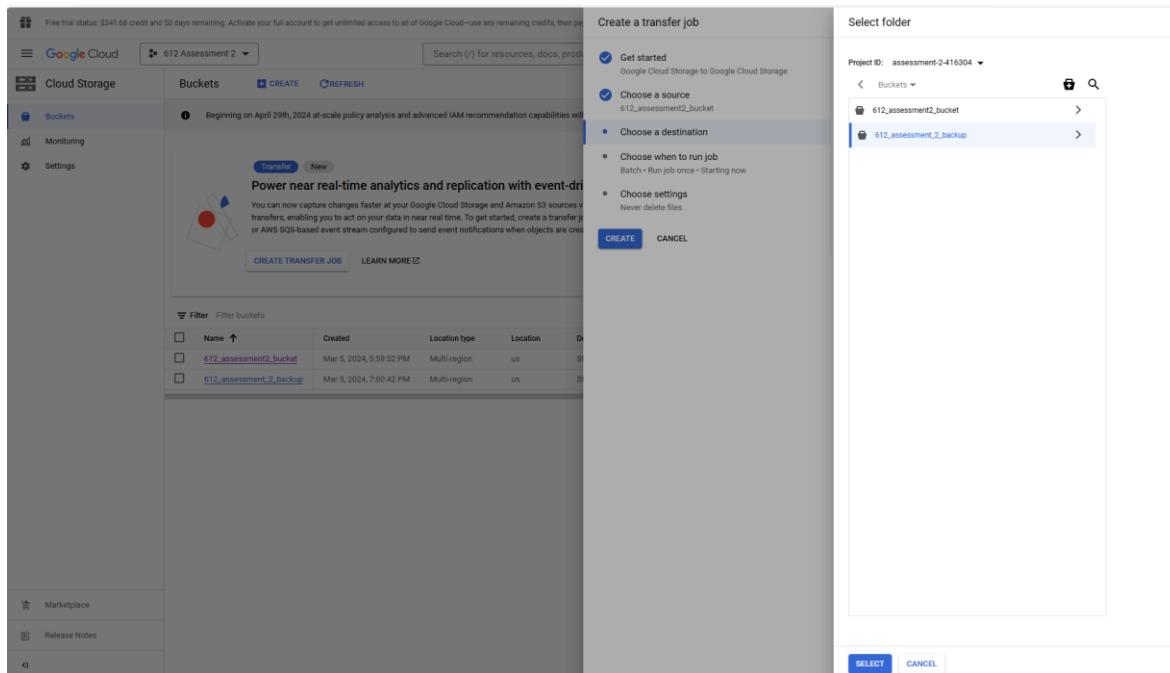


Then click SELECT.

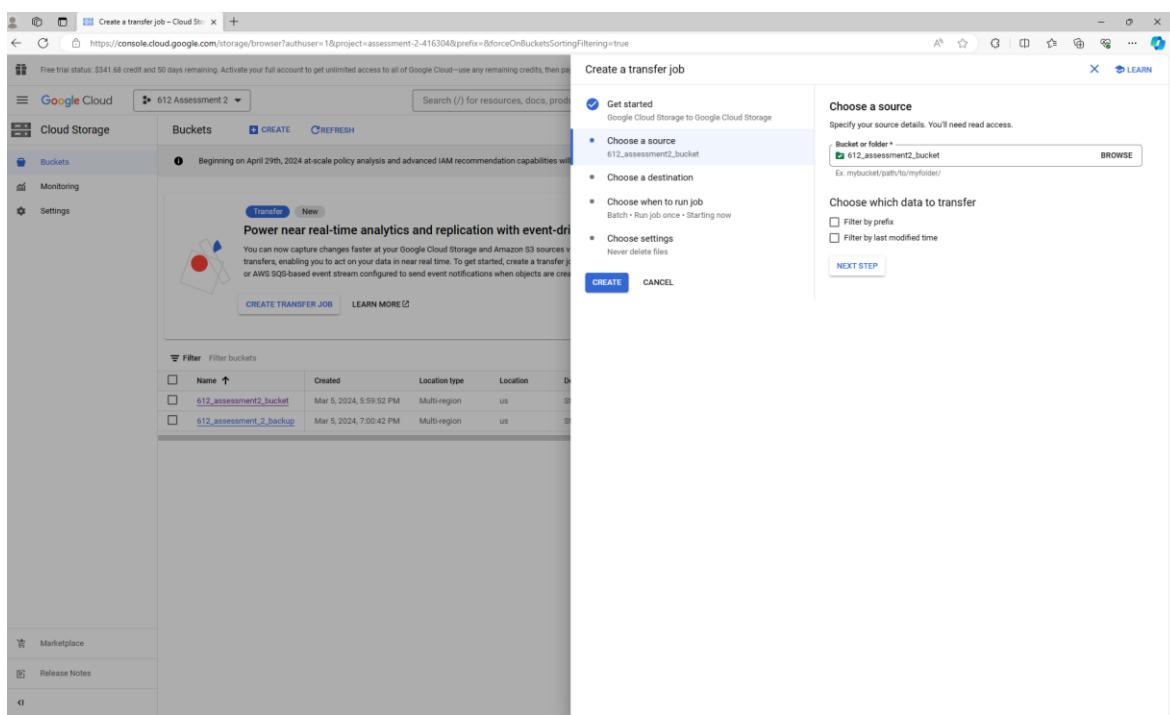


After clicking SELECT, you will be prompted to this screen. You will be given a choice to filter the content from the data source. If you want everything to be transferred, you may disregard it and click NEXT STEP.

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



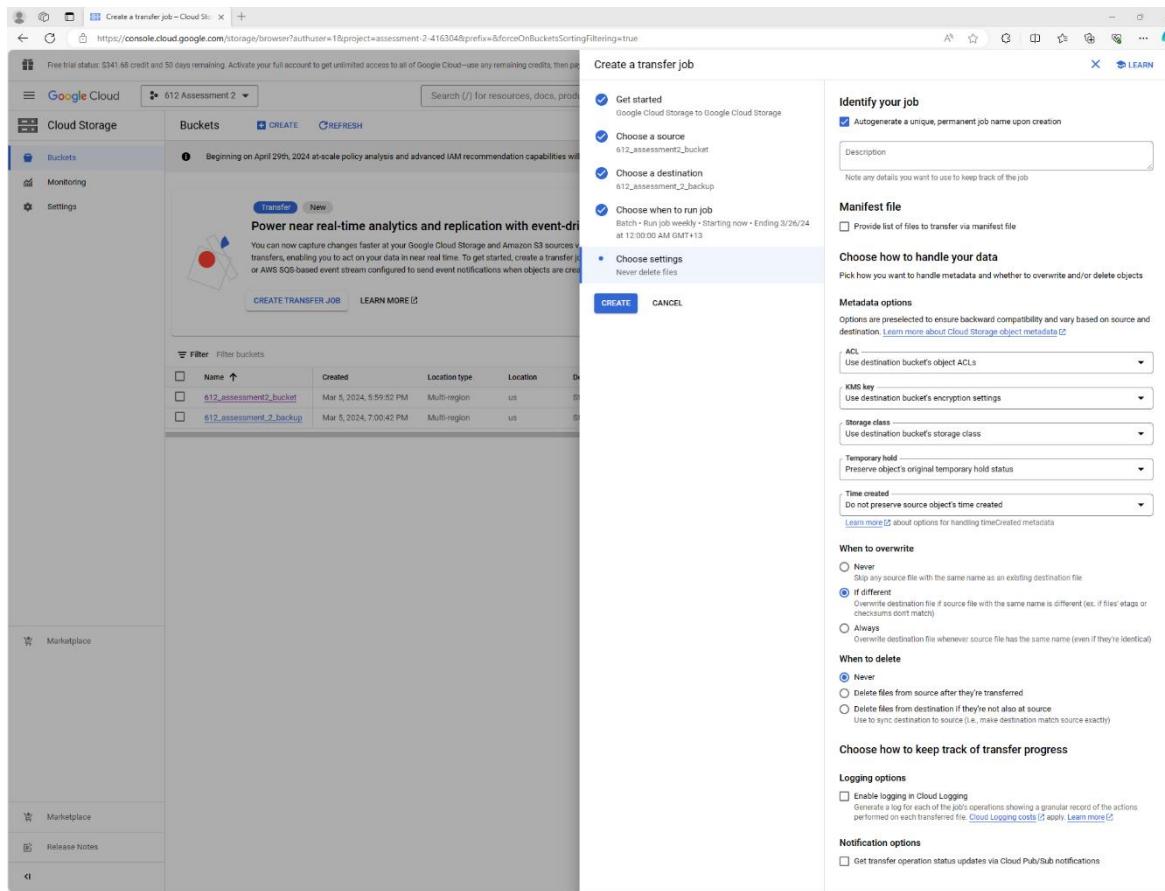
Click NEXT STEP



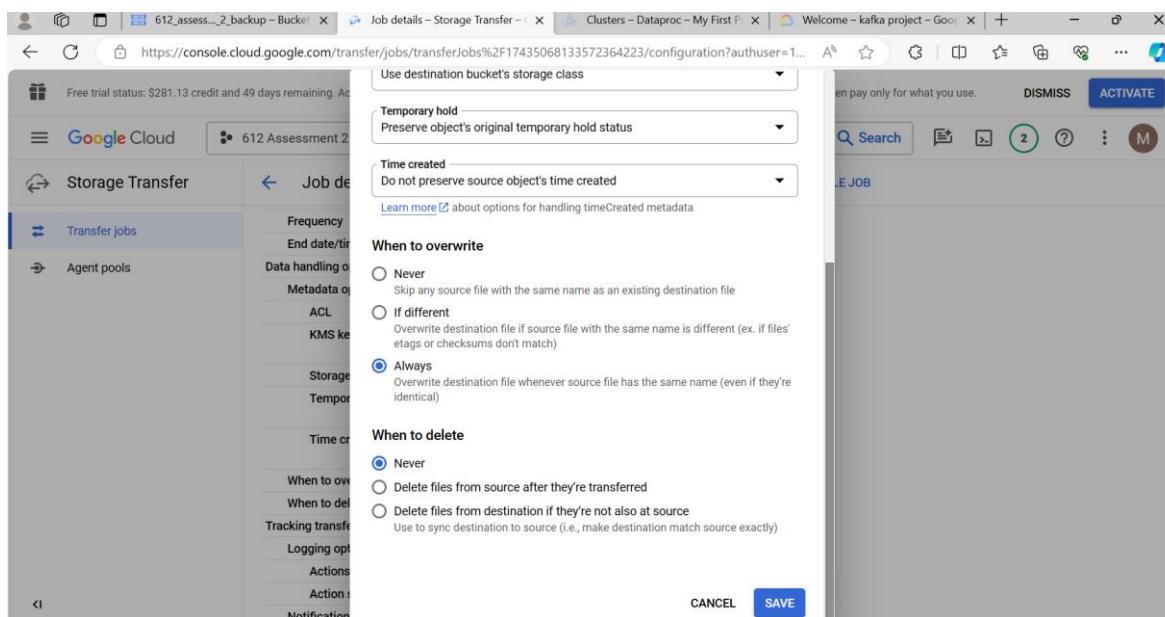
You may now choose when to run the job.

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.

The screenshot shows the Google Cloud Storage Buckets screen. On the left sidebar, there are links for Marketplace and Release Notes. The main area displays two buckets: `612_assessment2_bucket` and `612_assessment2_backup`. A modal window titled "Transfer" is open, featuring a section about "Power near real-time analytics and replication with event-driven transfers". It includes a "CREATE TRANSFER JOB" button and a "LEARN MORE" link. Another modal window titled "Security" is also visible, with a "View security recommendations" section and a "VIEW IN TABLE" button. At the bottom of the screen, a message box states: "Transfer job 'transferJobs/2421659450839609322' created successfully. You can monitor or manage the job in Storage Transfer Service." Below this message is a "GO TO JOB" button.

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

The screenshot shows the Job details - Storage Transfer screen for the transfer job `2421659450839609322`. The left sidebar has a "Transfer jobs" section. The main area displays the "Job information" for the job, including the full resource name (`transferJobs/2421659450839609322`), type (Google Cloud Storage), name (`612_assessment2_bucket`), and folder path (empty). The "Source" and "Destination" sections also show the same details. Below this, there are four monitoring charts: "Summary of bytes copied", "Summary of objects copied", "Bandwidth of bytes copied", and "Rate of objects copied", all showing "No data is available for the selected time frame".

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

This screenshot shows the 'Job details - Storage Transfer' page in the Google Cloud console. The job ID is 2421659450839609322. The 'Job information' section shows the full resource name as transferJobs/2421659450839609322, type as Google Cloud Storage, and folder path as 612_assessment2_bucket. The destination is also a Google Cloud Storage bucket named 612_assessment_2_backup. The 'Monitoring' tab is selected, showing four time-series charts: 'Summary of bytes copied', 'Summary of objects copied', 'Bandwidth of bytes copied', and 'Rate of objects copied'. All charts indicate 'No data is available for the selected time frame'. A message at the bottom states 'Running job now. Upcoming runs will follow the job's schedule.'

Checking if the backup has been successfully done.

This screenshot shows the 'Bucket details' page for the '612_assessment_2_backup' bucket. The bucket location is US, storage class is Standard, and public access is Not public. The 'OBJECTS' tab is selected, displaying a list of objects. The list includes two files: '612_Data_Transformation_and_Management.Assessment2_adidas_tidydf_.json' (size 791.1 KB) and '612_Data_Transformation_and_Management.adidas_images.json' (size 1.5 MB). Both files were created on March 5, 2024, at 7:02:14 PM. The Windows taskbar at the bottom shows various open applications and system status.

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.

Name	Version history	Encryption	Object retention
612_Data_Transformation_and_Management.Assessment2_adidas_tidydf...	1 noncurrent version	Google-managed	—
612_Data_Transformation_and_Management.adidas_images.json	1 noncurrent version	Google-managed	—

Turn on object versioning?

With object versioning on, live and noncurrent versions will be stored in the same bucket and storage class by default.

Save on version costs by adding lifecycle rules

Object lifecycle rules keep versioning costs under control. Without any lifecycle rules, versioning will be unlimited. Rules can be added or modified at any time.

[Learn more](#)

Add recommended lifecycle rules to manage version costs

Retention (for objects)

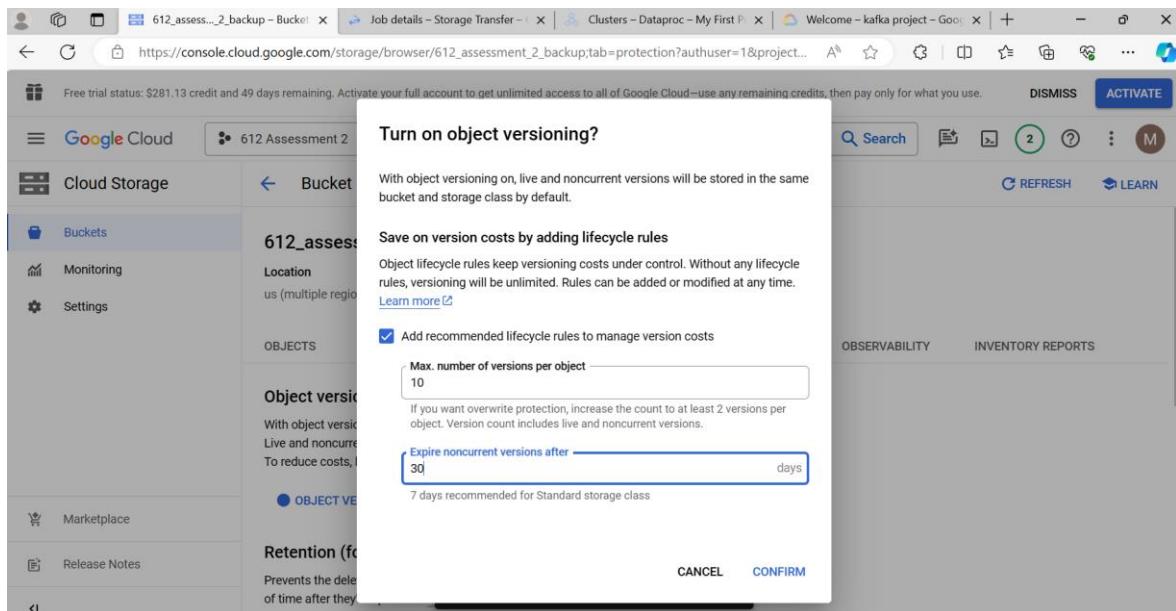
Prevents the deletion or modification of the objects in this bucket for a specified period of time after they're uploaded. [Learn more](#)

Bucket retention policy

Retain objects in this bucket for a uniform retention period.

CANCEL 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.



Bucket details

612_assessment_2_backup

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	Object versioning

PROTECTION

Object versioning (for data recovery)

With object versioning on, you can restore objects that have been overwritten or deleted. Live and noncurrent versions are stored in the same bucket and storage class by default. To reduce costs, limit the number of versions by adding a lifecycle rule. [Learn more](#)

OBJECT VERSIONING ON

You have 2 lifecycle rules applied to noncurrent versions.

[MANAGE RULES](#)

Retention (for compliance)

Prevents the deletion or modification of the objects in this bucket for a specified period of time after they're uploaded. [Learn more](#)

Bucket retention policy

Best to guarantee compliance for all objects in this bucket for a uniform retention period. The optional step of setting the retention mode to "locked" ensures that no one can remove or shorten the retention period.

[+ SET RETENTION POLICY](#)

Object retention

Best to guarantee compliance at the individual object level to meet specific retention needs. The bucket must have this option enabled before objects in it can have their own retention configuration.

OBJECT RETENTION NOT ENABLED

Default event-based hold option

When this option is enabled, event-based holds are placed on objects when they're uploaded to the bucket. This supports an event-based retention strategy: 'start the clock'

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

Name	Storage class	Last modified	Public access	Version history	Encryption
612_Data_Transformation_and_Management.Assessment2_adidas_tidydf.json	Standard	Mar 6, 2024, 2:34:38 PM	Not public	1 noncurrent version	Google-managed
612_Data_Transformation_and_Management.adidas_images.json	Standard	Mar 6, 2024, 2:34:38 PM	Not public	1 noncurrent version	Google-managed

Below is the backup history of the 612_Data_Transformation_and_Management Assessment2_adidas.tidydf.

Object version	CRC32C hash	Storage class	Size
a3ae	3447666961	Standard	791.1 KB
Mar 6, 2024, 2:30:31 PM	a3ae	3447666961	791.1 KB

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.

The screenshot shows the Google Cloud Kubernetes Engine interface. At the top, there is a navigation bar with the Google Cloud logo, a search bar containing 'Search (/) for resources, docs, products, and m...', and various icons including a user profile with '2' notifications, a help icon, and a menu icon. Below the navigation bar, the main title is 'Kubernetes Engine'. A sidebar on the left lists several categories: 'Resource Management' (Overview, Clusters, Workloads, Teams, Applications, Secrets & ConfigMaps, Storage, Marketplace, Release Notes), 'Learn about Enterprise', and a 'Failed to load' message. The 'Storage' category is currently selected, indicated by a blue background. The 'Failed to load' message states: 'There was an error while loading /kubernetes/persistentvolumeclaim?referrer=search&authuser=1&project=activity-alice-dataset. Please try again. It may be a browser or network issue. Go to the [loading issues help page](#) to troubleshoot the issue.'

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>