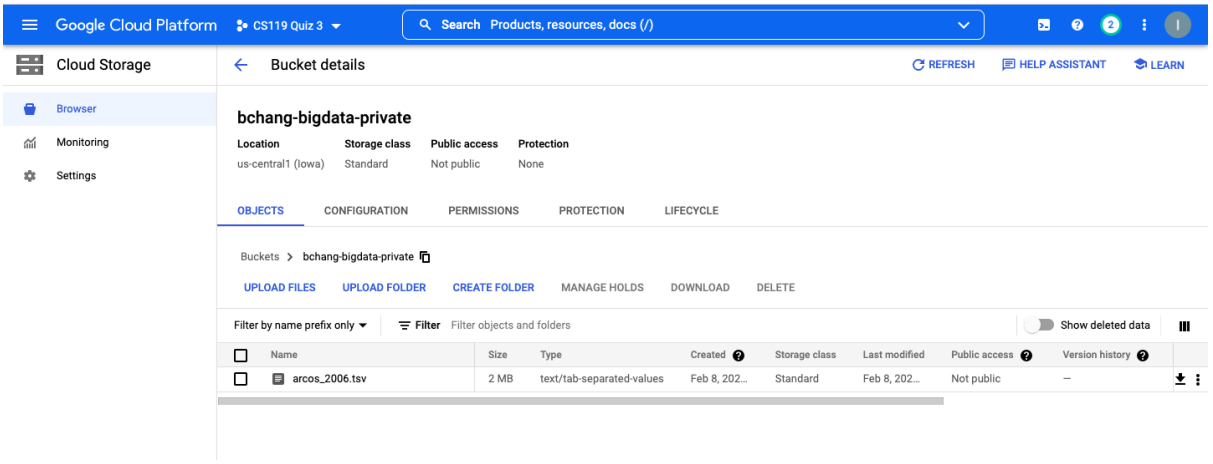
1. Created a google cloud storage bucket named `bchang-bigdata-private`, this is done by directly using the webUI of google cloud platform.



2. Following a similar procedure as in the previous quiz. Firstly, I shuffled and extracted 50000 rows from the original dataset on VM with the command:
`zcat /comp/119/arcos_all_washpost.tsv.gz | shuf -n 50000 > quiz3.csv`

Then I wrote a Python script to obtain the desired dataset of 5000 rows from the year 2006

```
import pandas as pd

# read the column names in, specifying tabs as delimiter
header = pd.read_csv('header.csv', sep='\t')

# read the data in, specifying the column names to be 'header'
df = pd.read_csv('quiz3.csv', delimiter='\t', names = list(header))

# change the format of date in order to extract the year later
df.TRANSACTION_DATE = pd.to_datetime(df.TRANSACTION_DATE, format = '%m%d%Y')

# extract rows of the year 2006 -- got 5926 rows
df_2006 = df[df.TRANSACTION_DATE.dt.year == 2006]

# randomly get 5000 rows
df_2006_5k = df.sample(n=5000)

# write to tsv
df_2006_5k.to_csv('arcos_2006.tsv', index=False, sep='\t')
```

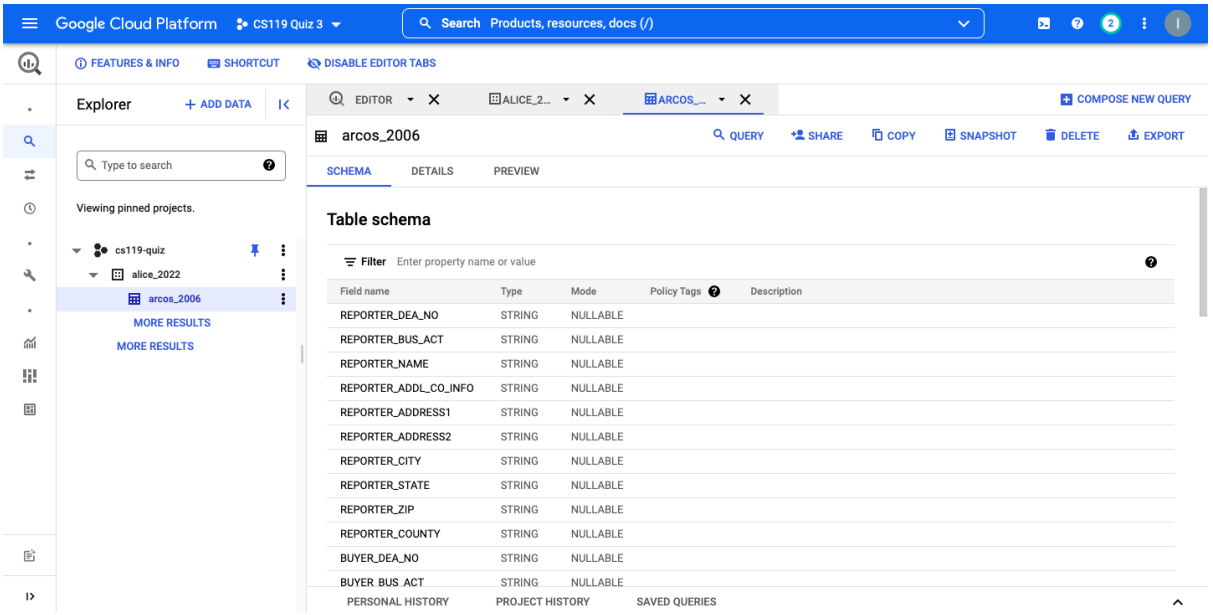
3. Upload the resulting file to GC storage: first I used the command:
`scp bchang01@linux.eecs.tufts.edu:/h/bchang01/arcos_2006.tsv /Users/irenechang/Desktop`
to copy the file from VM to my local directory. Then I clicked “Upload files" on GC storage to upload my file.

   When typing `gsutil ls -la gs://bchang-bigdata-private` in the GC shell to verify if the file is uploaded (authorization is prompted), the output is:
   `2095194 2022-02-08T18:07:25Z gs://bchang-bigdata-private/arcos_2006.tsv#1644343645302351 metageneration=1`

```
TOTAL: 1 objects, 2095194 bytes (2 MiB)
```

So the file is successfully uploaded.

4. Created the dataset alice-2022 and the table arcos-2006 in this dataset. Everything is done by directly using GC webUI. Below is a screenshot of the schema:



5. Ran the following the command in GC Console: bq load --field_delimiter=tab --source_format=CSV cs119-quiz:alice_2022.arcos_2006 gs://bchang-bigdata-private/arcos_2006.tsv

6. The SQL query used:

```
SELECT DISTINCT(DRUG_NAME) FROM `cs119-quiz.alice_2022.arcos_2006`
```

The result obtained:


```
[
  {
    "DRUG_NAME": "HYDROCODONE"
  },
  {
    "DRUG_NAME": "OXYCODONE"
  }
]
```