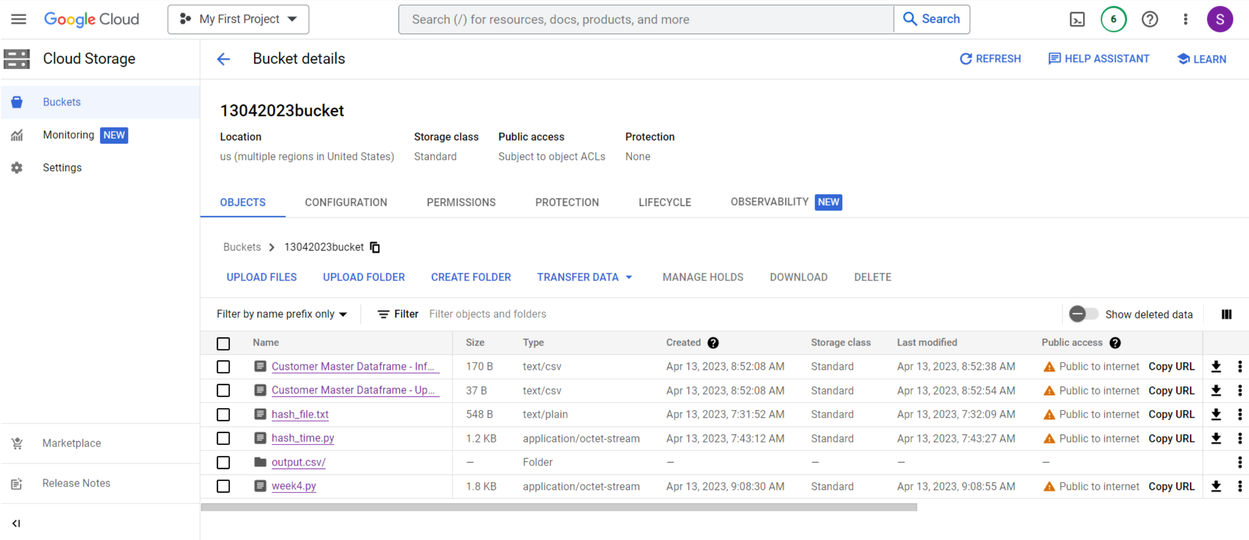
Big Data Graded Assignment 4

❖ Step 1 : Upload all files under bucket with public access

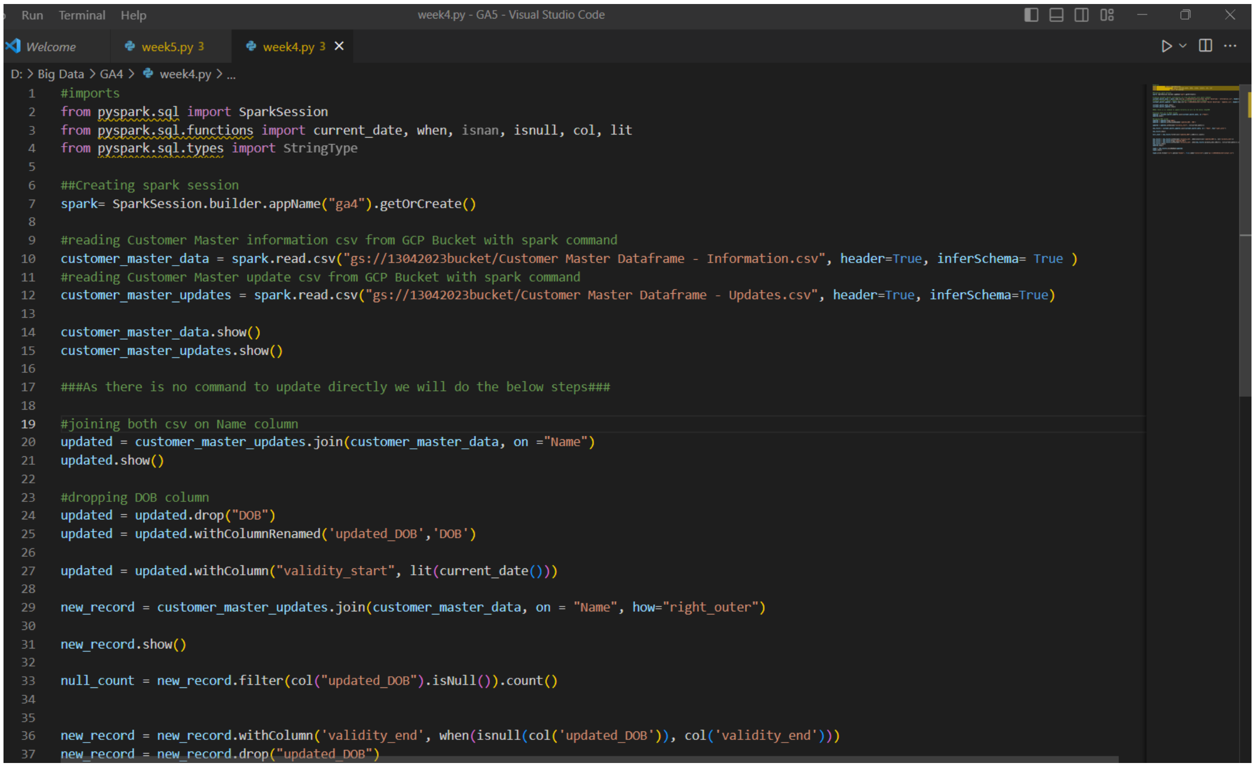


13042023bucket

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Subject to object ACLs	None

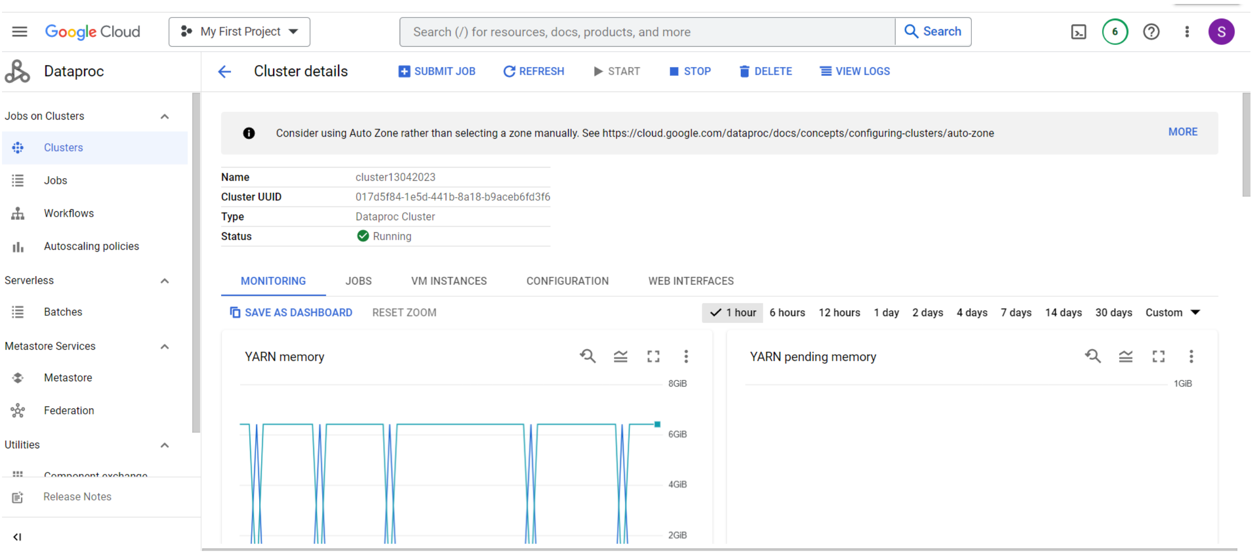
□	Name	Size	Type	Created	Storage class	Last modified	Public access
□	Customer Master Dataframe - Inf...	170 B	text/csv	Apr 13, 2023, 8:52:08 AM	Standard	Apr 13, 2023, 8:52:38 AM	Public to internet Copy URL
□	Customer Master Dataframe - Up...	37 B	text/csv	Apr 13, 2023, 8:52:08 AM	Standard	Apr 13, 2023, 8:52:54 AM	Public to internet Copy URL
□	hash_file.txt	548 B	text/plain	Apr 13, 2023, 7:31:52 AM	Standard	Apr 13, 2023, 7:32:09 AM	Public to internet Copy URL
□	hash_time.py	1.2 KB	application/octet-stream	Apr 13, 2023, 7:43:12 AM	Standard	Apr 13, 2023, 7:43:27 AM	Public to internet Copy URL
□	output.csv/	—	Folder	—	—	—	—
□	week4.py	1.8 KB	application/octet-stream	Apr 13, 2023, 9:08:30 AM	Standard	Apr 13, 2023, 9:08:55 AM	Public to internet Copy URL

❖ Python file


Run
Terminal
Help
week4.py - GA5 - Visual Studio Code
Welcome
week5.py 3
week4.py 3
D: > Big Data > GA4 > week4.py > ...

```python
#imports
from pyspark.sql import SparkSession
from pyspark.sql.functions import current_date, when, isnan, isnull, col, lit
from pyspark.sql.types import StringType

##Creating spark session
spark= SparkSession.builder.appName("ga4").getOrCreate()

#reading Customer Master information csv from GCP Bucket with spark command
customer_master_data = spark.read.csv("gs://13042023bucket/Customer Master Dataframe - Information.csv", header=True, inferSchema= True )
#reading Customer Master update csv from GCP Bucket with spark command
customer_master_updates = spark.read.csv("gs://13042023bucket/Customer Master Dataframe - Updates.csv", header=True, inferSchema=True)

customer_master_data.show()
customer_master_updates.show()

###As there is no command to update directly we will do the below steps###

#joining both csv on Name column
updated = customer_master_updates.join(customer_master_data, on ="Name")
updated.show()

#dropping DOB column
updated = updated.drop("DOB")
updated = updated.withColumnRenamed('updated_DOB','DOB')

updated = updated.withColumn("validity_start", lit(current_date()))

new_record = customer_master_updates.join(customer_master_data, on = "Name", how="right_outer")

new_record.show()

null_count = new_record.filter(col("updated_DOB").isNull()).count()


new_record = new_record.withColumn('validity_end', when(isnull(col('updated_DOB')), col('validity_end')))
new_record = new_record.drop("updated_DOB")
```

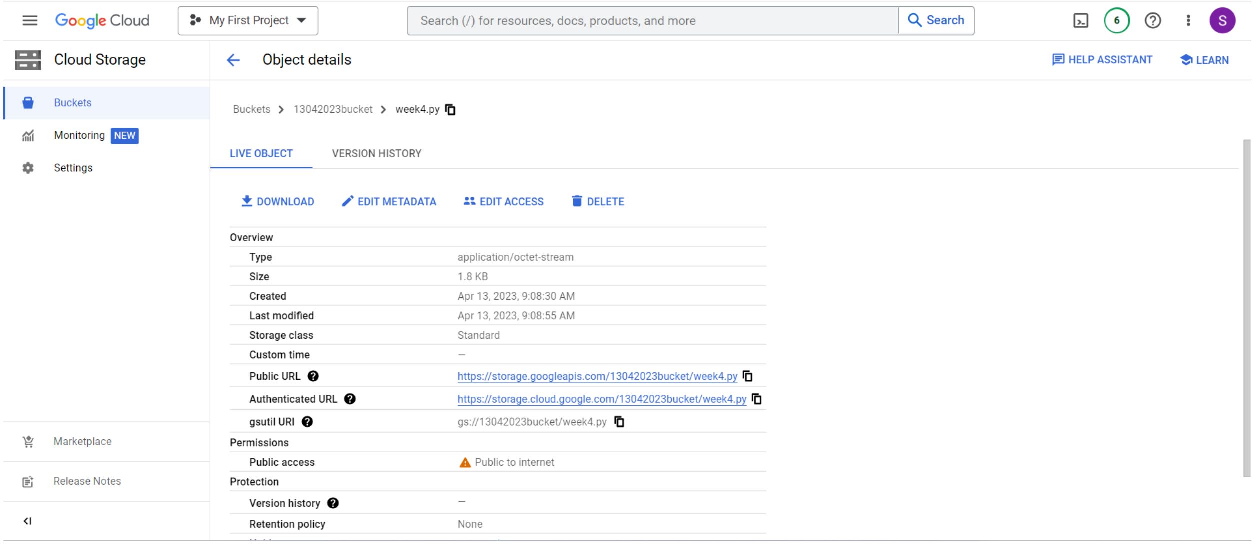
❖ Create a cluster which is running



Consider using Auto Zone rather than selecting a zone manually. See https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone

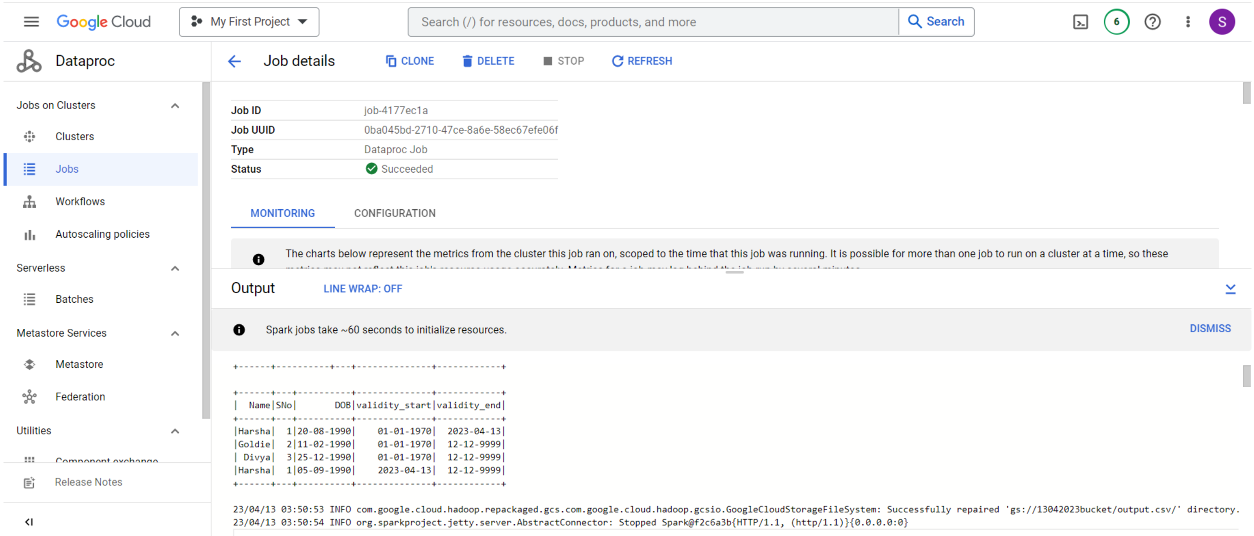
Name	cluster13042023
Cluster UUID	017d5f84-1e5d-441b-8a18-b9aceb6fd3f6
Type	Dataproc Cluster
Status	✓ Running

❖ Get gsutil url from file



Overview	
Type	application/octet-stream
Size	1.8 KB
Created	Apr 13, 2023, 9:08:30 AM
Last modified	Apr 13, 2023, 9:08:55 AM
Storage class	Standard
Custom time	—
Public URL	https://storage.googleapis.com/13042023bucket/week4.py
Authenticated URL	https://storage.cloud.google.com/13042023bucket/week4.py
gsutil URI	gs://13042023bucket/week4.py
Permissions	
Public access	⚠ Public to internet
Protection	
Version history	—
Retention policy	None

❖ Job succeeded with correct output

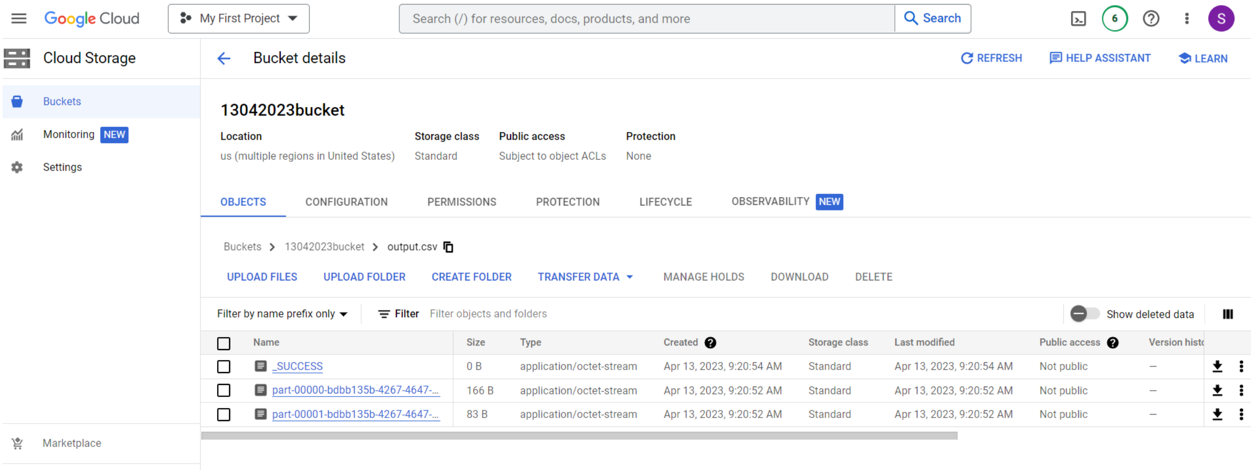


Job ID	job-4177ec1a
Job UUID	0ba045bd-2710-47ce-8a6e-58ec67efe06f
Type	Dataproc Job
Status	✓ Succeeded

The charts below represent the metrics from the cluster this job ran on, scoped to the time that this job was running. It is possible for more than one job to run on a cluster at a time, so these

Spark jobs take ~60 seconds to initialize resources.

```
+------+---------+---+-------------+-----------+


+------+---+----------+-------------+-----------+
|  Name|SNo|       DOB|validity_start|validity_end|
+------+---+----------+-------------+-----------+
|Harsha|  1|20-08-1990|    01-01-1970|  2023-04-13|
|Goldie|  2|11-02-1990|    01-01-1970|  12-12-9999|
| Divya|  3|25-12-1990|    01-01-1970|  12-12-9999|
|Harsha|  1|05-09-1990|    2023-04-13|  12-12-9999|
+------+---+----------+-------------+-----------+

23/04/13 03:50:53 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: Successfully repaired 'gs://13042023bucket/output.csv/' directory.
23/04/13 03:50:54 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@f2c6a3b{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
```



13042023bucket

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Subject to object ACLs	None

Buckets > 13042023bucket > output.csv

□	Name	Size	Type	Created	Storage class	Last modified	Public access	Version histo
□	_SUCCESS	0 B	application/octet-stream	Apr 13, 2023, 9:20:54 AM	Standard	Apr 13, 2023, 9:20:54 AM	Not public	–
□	part-00000-bdbb135b-4267-4647-...	166 B	application/octet-stream	Apr 13, 2023, 9:20:52 AM	Standard	Apr 13, 2023, 9:20:52 AM	Not public	–
□	part-00001-bdbb135b-4267-4647-...	83 B	application/octet-stream	Apr 13, 2023, 9:20:52 AM	Standard	Apr 13, 2023, 9:20:52 AM	Not public	–