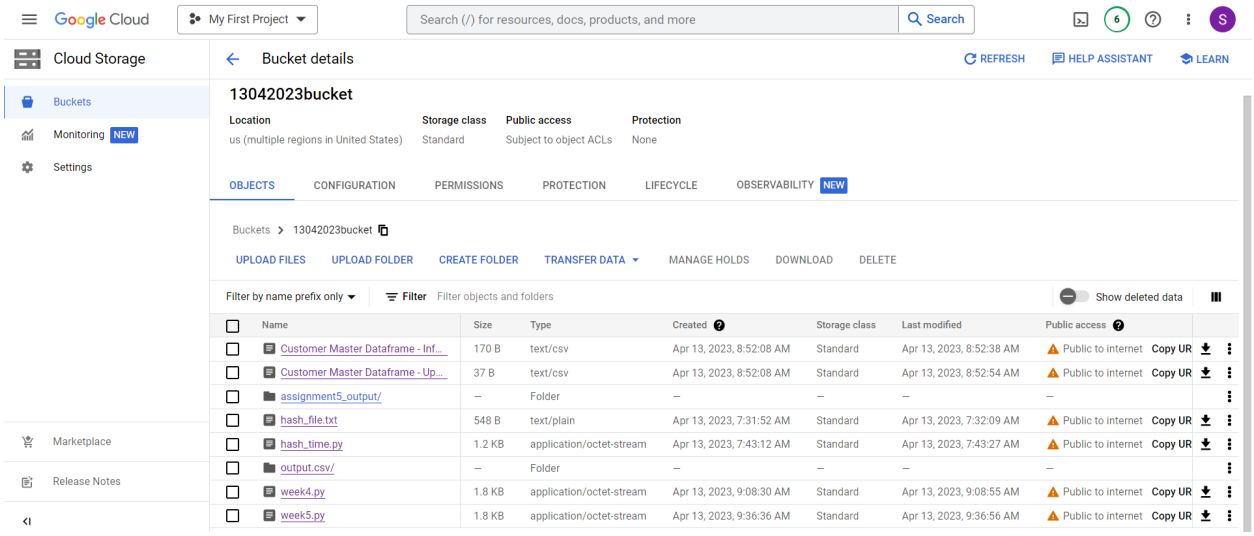
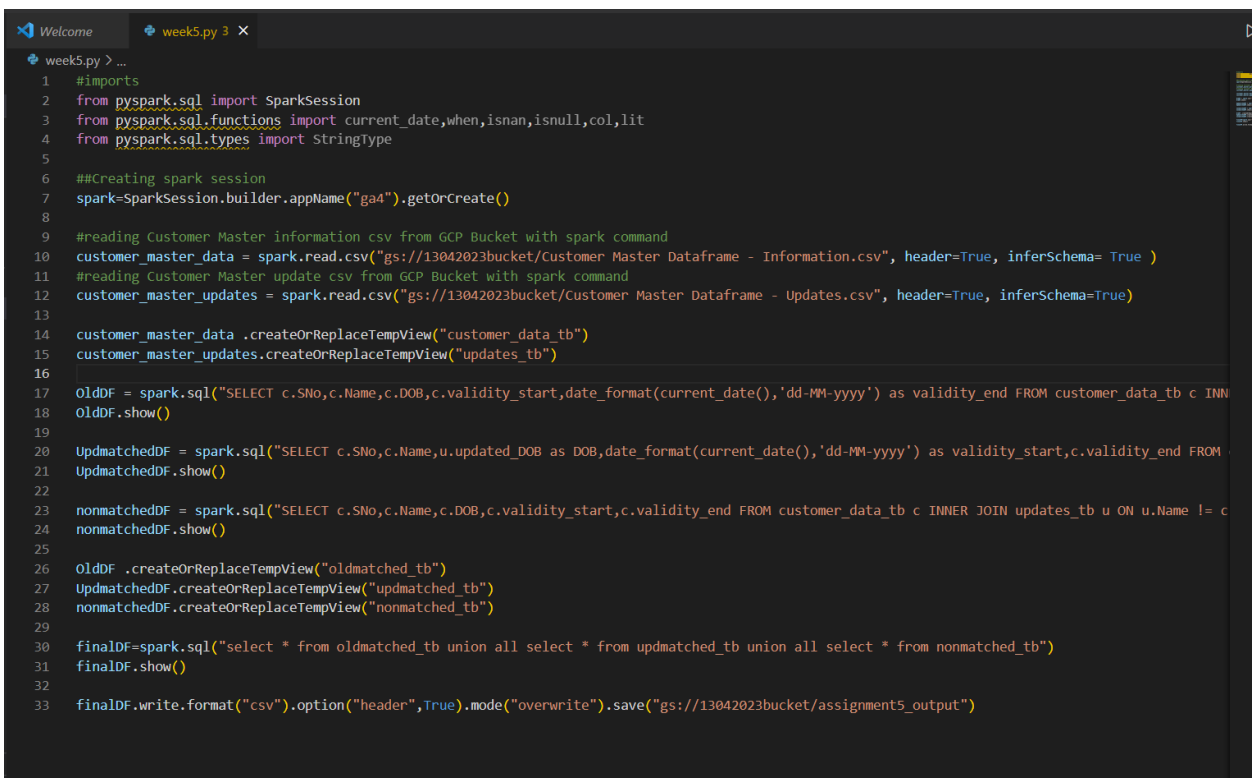
Big Data Graded Assignment 5

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

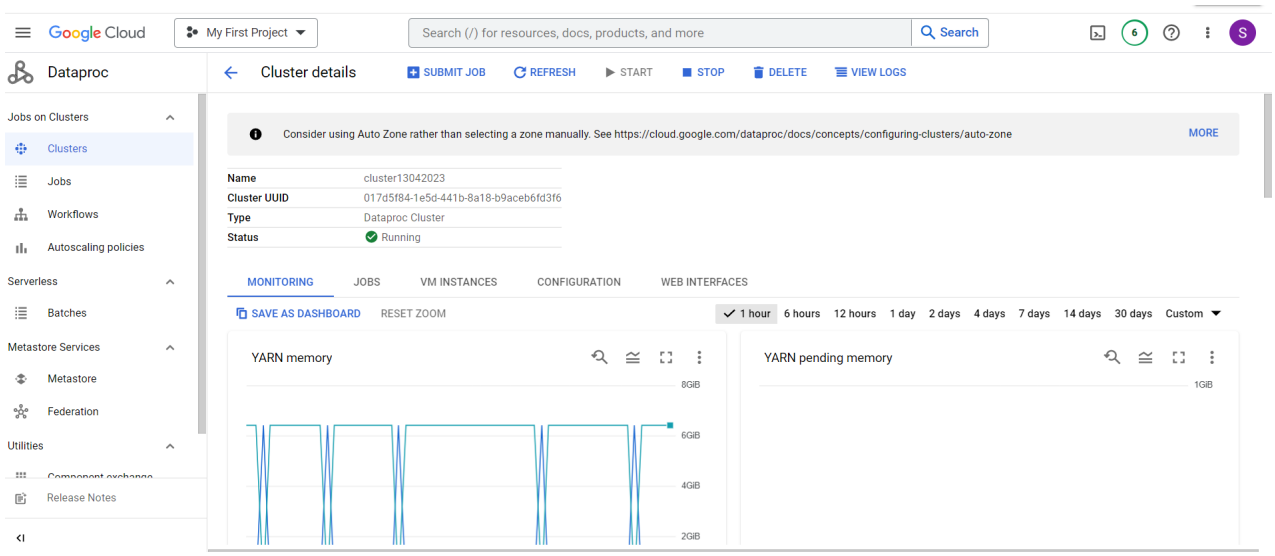


- Python Code file

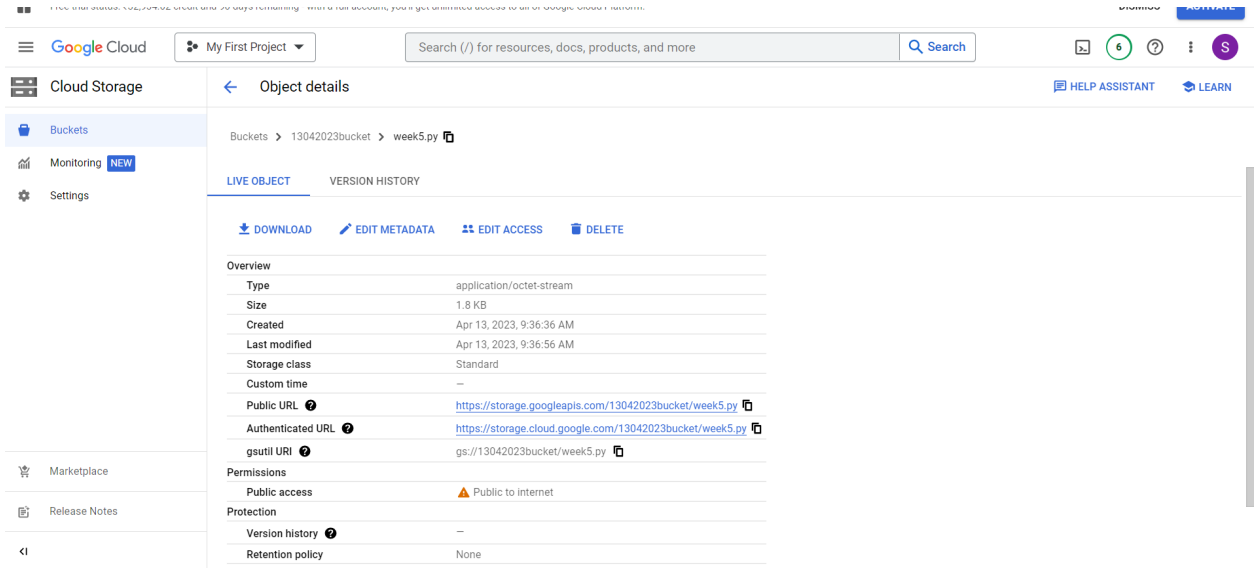


```python
#imports
from pyspark.sql import SparkSession
from pyspark.sql.functions import current_date,when,isnan,isnull,col,lit
from pyspark.sql.types import StringType

##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 .createOrReplaceTempView("customer_data_tb")
customer_master_updates.createOrReplaceTempView("updates_tb")

OldDF = spark.sql("SELECT c.SNo,c.Name,c.DOB,c.validity_start,date_format(current_date(),'dd-MM-yyyy') as validity_end FROM customer_data_tb c INN
OldDF.show()

UpdmatchedDF = spark.sql("SELECT c.SNo,c.Name,u.updated_DOB as DOB,date_format(current_date(),'dd-MM-yyyy') as validity_start,c.validity_end FROM
UpdmatchedDF.show()

nonmatchedDF = spark.sql("SELECT c.SNo,c.Name,c.DOB,c.validity_start,c.validity_end FROM customer_data_tb c INNER JOIN updates_tb u ON u.Name != c
nonmatchedDF.show()

OldDF .createOrReplaceTempView("oldmatched_tb")
UpdmatchedDF.createOrReplaceTempView("updmatched_tb")
nonmatchedDF.createOrReplaceTempView("nonmatched_tb")

finalDF=spark.sql("select * from oldmatched_tb union all select * from updmatched_tb union all select * from nonmatched_tb")
finalDF.show()

finalDF.write.format("csv").option("header",True).mode("overwrite").save("gs://13042023bucket/assignment5_output")
```

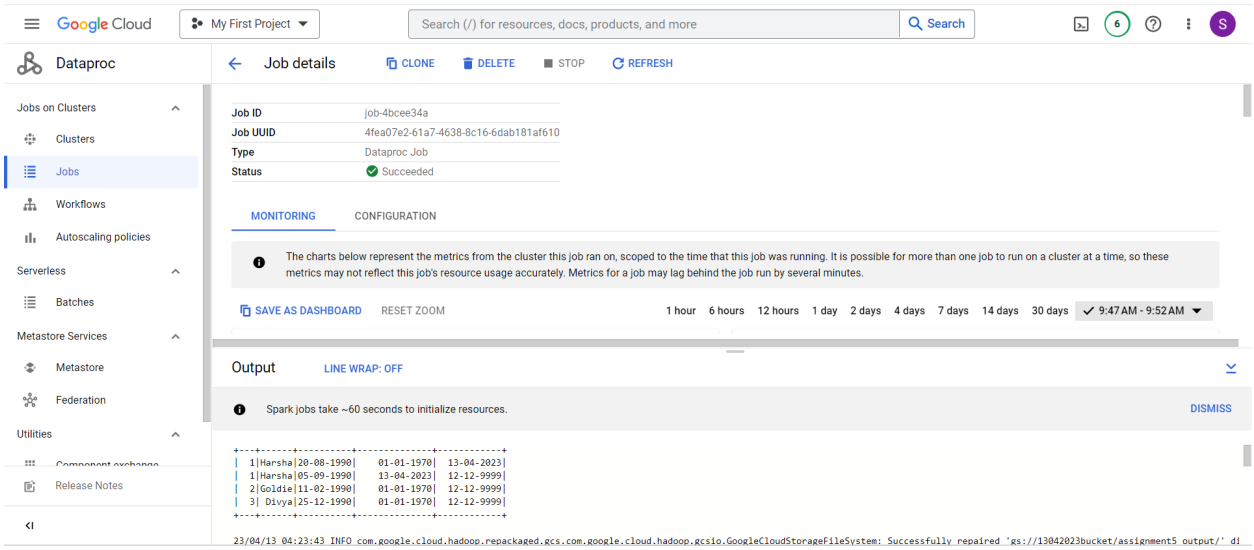
❖ Create a cluster which is running



❖ Get gsutil url from file



❖ Job running successfully



❖ Job with correct output