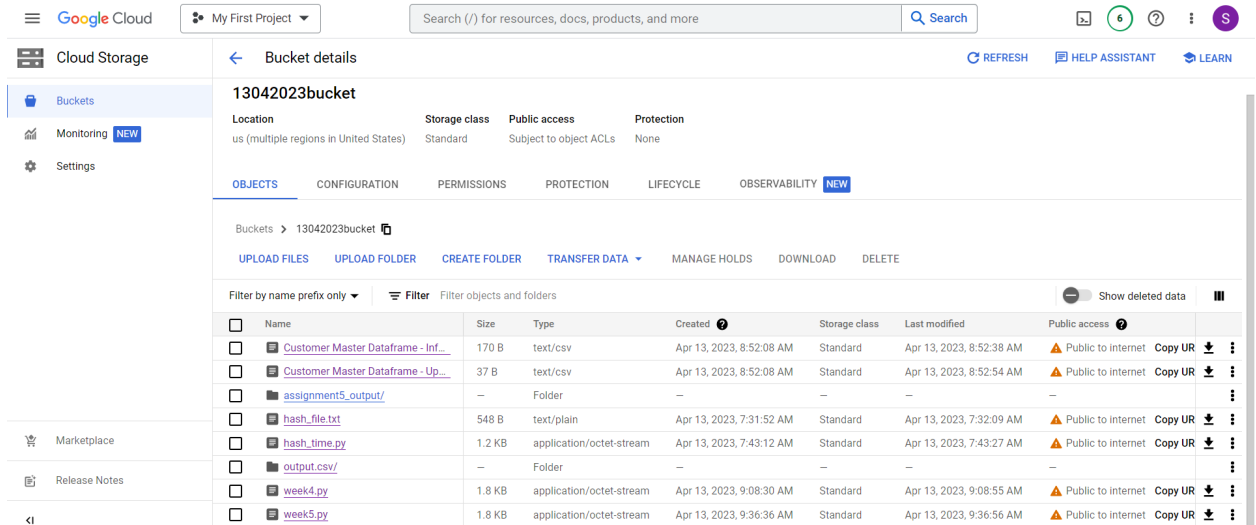


Big Data Graded Assignment 5

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



Cloud Storage

Bucket details

13042023bucket

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

OBJECTS | CONFIGURATION | PERMISSIONS | PROTECTION | LIFECYCLE | OBSERVABILITY NEW

Buckets > 13042023bucket

UPLOAD FILES | UPLOAD FOLDER | CREATE FOLDER | TRANSFER DATA | MANAGE HOLDS | DOWNLOAD | DELETE

Filter by name prefix only | Filter objects and folders | Show deleted data

<input type="checkbox"/>	Name	Size	Type	Created	Storage class	Last modified	Public access	Copy UR	Download	More
<input type="checkbox"/>	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 UR	Download	More
<input type="checkbox"/>	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 UR	Download	More
<input type="checkbox"/>	assignment5_output/	—	Folder	—	—	—	—	—	—	More
<input type="checkbox"/>	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 UR	Download	More
<input type="checkbox"/>	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 UR	Download	More
<input type="checkbox"/>	output.csv/	—	Folder	—	—	—	—	—	—	More
<input type="checkbox"/>	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 UR	Download	More
<input type="checkbox"/>	week5.py	1.8 KB	application/octet-stream	Apr 13, 2023, 9:36:36 AM	Standard	Apr 13, 2023, 9:36:56 AM	Public to internet	Copy UR	Download	More

❖ Python Code file

```
1 #imports
2 from pyspark.sql import SparkSession
3 from pyspark.sql.functions import current_date,when,isnan,isnull,col,lit
4 from pyspark.sql.types import StringType
5
6 ##Creating spark session
7 spark=SparkSession.builder.appName("ga4").getOrCreate()
8
9 #reading Customer Master information csv from GCP Bucket with spark command
10 customer_master_data = spark.read.csv("gs://13042023bucket/Customer Master Dataframe - Information.csv", header=True, inferSchema= True )
11 #reading Customer Master update csv from GCP Bucket with spark command
12 customer_master_updates = spark.read.csv("gs://13042023bucket/Customer Master Dataframe - Updates.csv", header=True, inferSchema=True)
13
14 customer_master_data.createOrReplaceTempView("customer_data_tb")
15 customer_master_updates.createOrReplaceTempView("updates_tb")
16
17 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 INNER JOIN updates_tb u ON c.SNo=u.SNo")
18 OldDF.show()
19
20 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 customer_data_tb c INNER JOIN updates_tb u ON c.SNo=u.SNo")
21 UpdmatchedDF.show()
22
23 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.Name")
24 nonmatchedDF.show()
25
26 OldDF.createOrReplaceTempView("oldmatched_tb")
27 UpdmatchedDF.createOrReplaceTempView("updmatched_tb")
28 nonmatchedDF.createOrReplaceTempView("nonmatched_tb")
29
30 finalDF=spark.sql("select * from oldmatched_tb union all select * from updmatched_tb union all select * from nonmatched_tb")
31 finalDF.show()
32
33 finalDF.write.format("csv").option("header",True).mode("overwrite").save("gs://13042023bucket/assignment5_output")
```

❖ Create a cluster which is running

Google Cloud | My First Project | Search (/) for resources, docs, products, and more

Dataproc | Cluster details | SUBMIT JOB | REFRESH | START | STOP | DELETE | VIEW LOGS

Jobs on Clusters | Clusters | Jobs | Workflows | Autoscaling policies

Serverless | Batches

Metastore Services | Metastore | Federation

Utilities | Component exchange | Release Notes

Consider using Auto Zone rather than selecting a zone manually. See <https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone> [MORE](#)

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

MONITORING | JOBS | VM INSTANCES | CONFIGURATION | WEB INTERFACES

SAVE AS DASHBOARD | RESET ZOOM | 1 hour | 6 hours | 12 hours | 1 day | 2 days | 4 days | 7 days | 14 days | 30 days | Custom

YARN memory

YARN pending memory

❖ Get gsutil url from file

Google Cloud | My First Project | Search (/) for resources, docs, products, and more

Cloud Storage | Object details | HELP ASSISTANT | LEARN

Buckets | Monitoring | Settings

Buckets > 13042023bucket > week5.py

LIVE OBJECT | VERSION HISTORY

DOWNLOAD | EDIT METADATA | EDIT ACCESS | DELETE

Overview	
Type	application/octet-stream
Size	1.8 KB
Created	Apr 13, 2023, 9:36:36 AM
Last modified	Apr 13, 2023, 9:36:56 AM
Storage class	Standard
Custom time	—
Public URL	https://storage.googleapis.com/13042023bucket/week5.py
Authenticated URL	https://storage.cloud.google.com/13042023bucket/week5.py
gsutil URL	gs://13042023bucket/week5.py
Permissions	
Public access	Public to internet
Protection	
Version history	—
Retention policy	None

❖ Job running successfully

The screenshot shows the Google Cloud Dataproc console. The left sidebar lists navigation options: Jobs on Clusters, Clusters, Jobs (selected), Workflows, Autoscaling policies, Serverless, Batches, Metastore Services, Metastore, Federation, Utilities, Component exchange, and Release Notes. The main panel displays 'Job details' for job ID 'job-4bcee34a'. The job status is 'Succeeded'. Below the job details, there are tabs for 'MONITORING' and 'CONFIGURATION'. A message states: '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 metrics may not reflect this job's resource usage accurately. Metrics for a job may lag behind the job run by several minutes.' There are buttons for 'SAVE AS DASHBOARD', 'RESET ZOOM', and a time range selector set to '1 hour'. The 'Output' tab is active, showing a log entry: 'Spark jobs take ~60 seconds to initialize resources.' Below this, a table shows the output of the job, which is a list of names and dates.

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

❖ Job with correct output

The screenshot shows the Google Cloud Dataproc console. The left sidebar lists navigation options: Jobs on Clusters, Clusters, Jobs (selected), Workflows, Autoscaling policies, Serverless, Batches, Metastore Services, Metastore, Federation, Utilities, Component exchange, and Release Notes. The main panel displays 'Job details' for job ID 'job-4bcee34a'. The job status is 'Succeeded'. Below the job details, there are tabs for 'MONITORING' and 'CONFIGURATION'. A message states: 'Spark jobs take ~60 seconds to initialize resources.' Below this, a table shows the output of the job, which is a list of names and dates.

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

The screenshot shows the Google Cloud Cloud Storage console. The left sidebar lists navigation options: Buckets (selected), Monitoring, and Settings. The main panel displays 'Bucket details' for the bucket '13042023bucket'. The bucket location is 'us (multiple regions in United States)', storage class is 'Standard', public access is 'Subject to object ACLs', and protection is 'None'. Below the bucket details, there are tabs for 'OBJECTS', 'CONFIGURATION', 'PERMISSIONS', 'PROTECTION', 'LIFECYCLE', and 'OBSERVABILITY'. The 'OBJECTS' tab is active, showing a list of objects. The objects are listed in a table with columns: Name, Size, Type, Created, Storage class, Last modified, Public access, and Version history.

Name	Size	Type	Created	Storage class	Last modified	Public access	Version history
part-00000-15429ce9-58e8-4da5...	0 B	application/octet-stream	Apr 13, 2023, 9:53:44 AM	Standard	Apr 13, 2023, 9:53:44 AM	Not public	—
part-00001-15429ce9-58e8-4da5...	83 B	application/octet-stream	Apr 13, 2023, 9:53:41 AM	Standard	Apr 13, 2023, 9:53:41 AM	Not public	—
part-00002-15429ce9-58e8-4da5...	83 B	application/octet-stream	Apr 13, 2023, 9:53:41 AM	Standard	Apr 13, 2023, 9:53:41 AM	Not public	—
part-00002-15429ce9-58e8-4da5...	124 B	application/octet-stream	Apr 13, 2023, 9:53:42 AM	Standard	Apr 13, 2023, 9:53:42 AM	Not public	—