

Big Data Graded Assignment 4

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

The screenshot shows the Google Cloud Storage interface for a project named 'My First Project'. The 'Bucket details' page for '13042023bucket' is displayed. The bucket's location is 'us (multiple regions in United States)', storage class is 'Standard', public access is 'Subject to object ACLs', and protection is 'None'. The 'OBJECTS' tab is active, showing a list of files and folders. All objects have 'Public to internet' access, indicated by a yellow triangle icon. The list includes files like 'Customer Master Dataframe - Inf...', 'Customer Master Dataframe - Up...', 'hash_file.txt', 'hash_time.py', 'output.csv', and 'week4.py', along with a folder 'output.csv'.

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
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
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
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
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

❖ Python file

```
Run Terminal Help week4.py - GA4 - Visual Studio Code
Welcome week5.py 3 week4.py 3 X
D:\> Big Data > GA4 > week4.py > ...
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.show()
15 customer_master_updates.show()
16
17 ##As there is no command to update directly we will do the below steps##
18
19 #joining both csv on Name column
20 updated = customer_master_updates.join(customer_master_data, on = "Name")
21 updated.show()
22
23 #dropping DOB column
24 updated = updated.drop("DOB")
25 updated = updated.withColumnRenamed('updated_DOB','DOB')
26
27 updated = updated.withColumn("validity_start", lit(current_date()))
28
29 new_record = customer_master_updates.join(customer_master_data, on = "Name", how="right_outer")
30
31 new_record.show()
32
33 null_count = new_record.filter(col("updated_DOB").isNull()).count()
34
35
36 new_record = new_record.withColumn('validity_end', when(isnull(col('updated_DOB')), col('validity_end')))
37 new_record = new_record.drop("updated_DOB")
```

❖ Create a cluster which is running

The screenshot shows the Google Cloud Dataproc 'Cluster details' page for a cluster named 'cluster13042023'. The cluster is in a 'Running' status. The left sidebar shows navigation options like Clusters, Jobs, Workflows, and Autoscaling policies. The main content area includes a 'MONITORING' tab with two graphs: 'YARN memory' and 'YARN pending memory'. The 'YARN memory' graph shows a sawtooth pattern, indicating memory usage over time. The 'YARN pending memory' graph is currently flat at 0GB. A top banner suggests using Auto Zone for better availability.

Cluster details:

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

Monitoring:

- YARN memory: Graph showing memory usage over time (0 to 8GB).
- YARN pending memory: Graph showing pending memory usage over time (0 to 1GB).

❖ Get gsutil url from file

The screenshot shows the Google Cloud Cloud Storage 'Object details' page for a file named 'week4.py' in a bucket named '13042023bucket'. The page displays various metadata and permissions for the object. The 'gsutil URL' is highlighted as 'gs://13042023bucket/week4.py'. The 'Public URL' is also provided as 'https://storage.googleapis.com/13042023bucket/week4.py'. The 'Authenticated URL' is 'https://storage.cloud.google.com/13042023bucket/week4.py'. The 'Permissions' section shows 'Public access' is set to 'Public to internet'.

Object details:

- Bucket: 13042023bucket
- Object: week4.py
- 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 URL: <gs://13042023bucket/week4.py>
- Permissions: Public access (Public to internet)
- Version history: --
- Retention policy: None

❖ Job succeeded with correct output

Google Cloud

My First Project

Search (/) for resources, docs, products, and more

Search

6

S

Dataproc

Jobs on Clusters

Clusters

Jobs

Workflows

Autoscaling policies

Serverless

Batches

Metastore Services

Metastore

Federation

Utilities

Component exchange

Release Notes

Job details

CLONE

DELETE

STOP

REFRESH

Job ID

job-4177ec1a

Job UUID

0ba045bd-2710-47ce-8a6e-58ec67efe06f

Type

Dataproc Job

Status

Succeeded

MONITORING

CONFIGURATION

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

Output

LINE WRAP: OFF

Spark jobs take ~60 seconds to initialize resources.

DISMISS

+-----+-----+-----+-----+
+ Name | Size | D08 | 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}

Google Cloud

My First Project

Search (/) for resources, docs, products, and more

Search

6

S

Cloud Storage

Buckets

Monitoring

Settings

Bucket details

REFRESH

HELP ASSISTANT

LEARN

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

Buckets > 13042023bucket > output.csv

UPLOAD FILES

UPLOAD FOLDER

CREATE FOLDER

TRANSFER DATA

MANAGE HOLDS

DOWNLOAD

DELETE

Filter by name prefix only

Filter

Filter objects and folders

Show deleted data

	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history
<input type="checkbox"/>	_SUCCESS	0 B	application/octet-stream	Apr 13, 2023, 9:20:54 AM	Standard	Apr 13, 2023, 9:20:54 AM	Not public	--
<input type="checkbox"/>	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	--
<input type="checkbox"/>	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	--

Marketplace