

## 1. Moving files to bucket

← Bucket details REFRESH HELP ASSISTANT

**my\_bucket\_20**

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 > my\_bucket\_20

[UPLOAD FILES](#) | [UPLOAD FOLDER](#) | [CREATE FOLDER](#) | [TRANSFER DATA](#) | [MANAGE HOLDS](#) | [DOWNLOAD](#) | [DELETE](#)

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

<input type="checkbox"/>	Name	Size	Type	Created	Storage class	Last modified	Public access
<input type="checkbox"/>	Customer Master Dataframe - Inf...	170 B	text/csv	Apr 13, 2023, 8:36:55 AM	Standard	Apr 13, 2023, 8:36:55 AM	Not public
<input type="checkbox"/>	Customer Master Dataframe - Up...	35 B	text/csv	Apr 13, 2023, 8:00:30 AM	Standard	Apr 13, 2023, 8:00:40 AM	Public to internet <a href="#">Copy I</a>
<input type="checkbox"/>	ga4_code.py	1.4 KB	application/octet-stream	Apr 13, 2023, 8:18:18 AM	Standard	Apr 13, 2023, 8:18:27 AM	Public to internet <a href="#">Copy I</a>
<input type="checkbox"/>	ga5.py	1.6 KB	application/octet-stream	Apr 13, 2023, 8:32:54 AM	Standard	Apr 13, 2023, 8:33:06 AM	Public to internet <a href="#">Copy I</a>
<input type="checkbox"/>	ga5_output/	—	Folder	—	—	—	—

## 2. Codes

```
ga5.py
D:\> Other > #1 Boc 2 > Codes_big_data > Assignment_5 > ga5.py > ...
1 from pyspark.sql import SparkSession
2 from pyspark.sql.functions import current_date, when, isnan, isnull, col, lit
3 from pyspark.sql.types import StringType
4
5 spark=SparkSession.builder.appName("ga4").getOrCreate()
6
7
8 customer_data = spark.read.csv("gs://my_bucket_20/Customer Master Dataframe - Information.csv", header=True, inferSchema=True)
9 updates = spark.read.csv("gs://my_bucket_20/Customer Master Dataframe - Updates.csv", header=True, inferSchema=True)
10
11 customer_data.createOrReplaceTempView("customer_data_tb")
12 updates.createOrReplaceTempView("updates_tb")
13
14 OldmatchedDF = 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 u.Name == c.Name")
15 OldmatchedDF.show()
16
17 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 u.Name == c.Name")
18 UpdmatchedDF.show()
19
20 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")
21 nonmatchedDF.show()
22
23 OldmatchedDF.createOrReplaceTempView("oldmatched_tb")
24 UpdmatchedDF.createOrReplaceTempView("updmatched_tb")
25 nonmatchedDF.createOrReplaceTempView("nonmatched_tb")
26
27 finalDF=spark.sql("select * from oldmatched_tb union all select * from updmatched_tb union all select * from nonmatched_tb")
28 finalDF.show()
29
30 finalDF.write.format("csv").option("header",True).mode("overwrite").save("gs://my_bucket_20/ga5_output")
```

## 3. The CSV Files

	A	B	C	D	E	F	G
1	SNo	Name	DOB	validity_start	validity_end		
2	1	Harsha	20-08-1990	01-01-1970	12-12-9999		
3	2	Goldie	11-02-1990	01-01-1970	12-12-9999		
4	3	Divya	25-12-1990	01-01-1970	12-12-9999		
5							

	M8			
	A	B	C	D
1	Name	updated_DOB		
2	Harsha	05-09-1990		

#### 4. Creating Clusters

Cluster details
+ SUBMIT JOB
↺ REFRESH
▶ START
■ STOP
🗑 DELETE
≡ VIEW LOGS

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

Name	my-cluster-6ebf
Cluster UUID	2cfcfa0-2646-4524-86c0-814f5c8a711b
Type	Dataproc Cluster
Status	<span>✓</span> Running

MONITORING
JOBS
VM INSTANCES
CONFIGURATION
WEB INTERFACES

Filter
Filter jobs

#### 5. Creating Job & Output

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

Job details
📄 CLONE
🗑 DELETE
■ STOP
↺ REFRESH

Job ID	GA5-job-bcca909a
Job UUID	91450c7a-cfc2-4f42-8499-f87ad1fd719a
Type	Dataproc Job
Status	<span>✓</span> Succeeded

MONITORING
CONFIGURATION

**i** 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 metrics may not reflect this job's resource usage accurately. Metrics for a job may lag behind the job run by several minutes.

Output LINE WRAP: OFF

**i** Spark jobs take ~60 seconds to initialize resources.

```

+-----+
|SNo| 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|
+-----+

```

6. The final Output.csv

←

bucket details

REFRESH

HELP ASSISTANCE

my\_bucket\_20

Location

Storage class

Public access

Protection

us (multiple regions in United States)

Standard

Subject to object ACLs

None

OBJECTS

CONFIGURATION

PERMISSIONS

PROTECTION

LIFECYCLE

OBSERVABILITY

NEW

Buckets > my\_bucket\_20 > ga5\_output

UPLOAD FILES

UPLOAD FOLDER

CREATE FOLDER

TRANSFER DATA

MANAGE HOLDS

DOWNLOAD





DELETE

Filter by name prefix only

Filter

Filter objects and folders

Show

<input type="checkbox"/>	Name	Size	Type	Created	Storage class	Last modified	Public access
<input type="checkbox"/>	 <a href="#">_SUCCESS</a>	0 B	application/octet-stream	Apr 13, 2023, 9:31:44 AM	Standard	Apr 13, 2023, 9:31:44 AM	Not public
<input type="checkbox"/>	 <a href="#">part-00000-0fd48d05-8d47-49e5-...</a>	83 B	application/octet-stream	Apr 13, 2023, 9:31:41 AM	Standard	Apr 13, 2023, 9:31:41 AM	Not public
<input type="checkbox"/>	 <a href="#">part-00001-0fd48d05-8d47-49e5-...</a>	83 B	application/octet-stream	Apr 13, 2023, 9:31:41 AM	Standard	Apr 13, 2023, 9:31:41 AM	Not public
<input type="checkbox"/>	 <a href="#">part-00002-0fd48d05-8d47-49e5-...</a>	124 B	application/octet-stream	Apr 13, 2023, 9:31:43 AM	Standard	Apr 13, 2023, 9:31:43 AM	Not public