

## 1. Moving files to bucket

← Bucket details REFRESH HELP ASSISTANT LEARN

**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 data

<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 URL</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 URL</a>	

## 2. Codes

```
File Edit Selection View Go Run Terminal Help ga4_code.py - Visual Studio Code

ga4_code.py X
D:\> Other > #1 Bsc 2 > Codes_big_data > Assignment 4 > ga4_code.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("assignment4").getOrCreate()
6
7 customer_data = spark.read.csv("gs://my_bucket_20/Customer Master Dataframe - Information.csv", header=True, inferSchema= True )
8 updates = spark.read.csv("gs://my_bucket_20/Customer Master Dataframe - Updates.csv", header=True, inferSchema=True)
9
10 customer_data.show()
11 updates.show()
12
13 updated = updates.join(customer_data, on ="Name")
14 updated.show()
15
16 updated = updated.drop("DOB")
17 updated = updated.withColumnRenamed('updated_DOB','DOB')
18
19 updated = updated.withColumn("validity_start", lit(current_date()))
20
21 new_record = updates.join(customer_data, on = "Name", how="right_outer")
22
23 new_record.show()
24
25 null_count = new_record.filter(col("updated_DOB").isNull()).count()
26
27 print(null_count)
28
29 new_record = new_record.withColumn('validity_end', when(isnull(col('updated_DOB')), col('validity_end')))
30 new_record = new_record.drop("updated_DOB")
31 new_record = new_record.withColumn("validity_end", when(new_record.validity_end.isNull(), lit(current_date())) otherwise(new_record.validity_end))
32 new_record.show()
33 updated.show()
34
35 type2 = new_record.unionByName(updated)
36 type2.show()
37
38 type2.write.format("csv").option("header", True).mode("overwrite").save("gs://my_bucket_20/output.csv")
```

3. The CSV Files

Paste

Copy

Format Painter

Clipboard

B

I

U

Font

POSSIBLE DATA LOSS

Some features might be lost if you save this workbook in the c

E14

M8

4. Creating Clusters

My First Project

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

Search

Cluster details

SUBMIT JOB

REFRESH

START

STOP

DELETE

VIEW LOGS

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

2cfcfb0-2646-4524-86c0-814f5c8a711b

Type

Dataproc Cluster

Status

Running

MONITORING

JOBS

VM INSTANCES

CONFIGURATION

WEB INTERFACES

Filter

Filter jobs

Job ID	Status	Region	Time	Start time
--------	--------	--------	------	------------

## 5. Creating Job & Output

My First Project

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

Search

Job details

CLONE

DELETE

STOP

REFRESH

Job ID	GA4-job-f2da32c6
Job UUID	a20d7202-6d09-4687-bda7-3f675845259a
Type	Dataproc Job
Status	Succeeded

MONITORING

CONFIGURATION

Output

LINE WRAP: OFF

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 03:55:27 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: Successfully repaired 'gs://my\_buck
23/04/13 03:55:28 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@6ab39f8f[HTTP/1.1, (http/1.1)]{0.0.0.0:0}

## 6. The final Output.csv

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

>

output.csv

UPLOAD FILES

UPLOAD FOLDER

CREATE FOLDER

TRANSFER DATA

MANAGE HOLDS




DOWNLOAD

DELETE

Filter by name prefix only

Filter

Filter objects and folders

<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, 9:25:28 AM	Standard	Apr 13, 2023, 9:25:28 AM	Not public
<input type="checkbox"/>	 <a href="#">part-00000-9bffc554-c6f0-4854-9...</a>	166 B	application/octet-stream	Apr 13, 2023, 9:25:26 AM	Standard	Apr 13, 2023, 9:25:26 AM	Not public
<input type="checkbox"/>	 <a href="#">part-00001-9bffc554-c6f0-4854-9...</a>	83 B	application/octet-stream	Apr 13, 2023, 9:25:26 AM	Standard	Apr 13, 2023, 9:25:26 AM	Not public