

# Diplomado\_DS\_Mod13\_Pyspark\_sesion1

September 9, 2023

```
[1]: spark
```

```
[1]: <pyspark.sql.session.SparkSession at 0x7f9f63db8bb0>
```

```
[2]: !pwd
```

```
/
```

```
[3]: ! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/03/2023-01.csv
      ! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/03/2023-02.csv
      ! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/04/2023-03.csv
      ! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/05/2023-04.csv
      ! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/06/2023-05.csv
      ! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/07/2023_06.csv
      ! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/08/2023_07.csv
      ! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/09/2023_08.csv
```

```
--2023-09-09 19:50:28-- https://ecobici.cdmx.gob.mx/wp-
content/uploads/2023/03/2023-01.csv
Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)... 157.230.72.27
Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443...
connected.
HTTP request sent, awaiting response... 200 OK
Length: 41926344 (40M) [text/csv]
Saving to: '2023-01.csv'
```

```
2023-01.csv          100%[=====>]  39.98M  33.9MB/s   in 1.2s
```

```
2023-09-09 19:50:30 (33.9 MB/s) - '2023-01.csv' saved [41926344/41926344]
```

```
--2023-09-09 19:50:30-- https://ecobici.cdmx.gob.mx/wp-
content/uploads/2023/03/2023-02.csv
```

Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)... 157.230.72.27  
Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443...  
connected.

HTTP request sent, awaiting response... 200 OK

Length: 47532884 (45M) [text/csv]

Saving to: '2023-02.csv'

2023-02.csv 100%[=====>] 45.33M 42.6MB/s in 1.1s

2023-09-09 19:50:31 (42.6 MB/s) - '2023-02.csv' saved [47532884/47532884]

--2023-09-09 19:50:31-- https://ecobici.cdmx.gob.mx/wp-  
content/uploads/2023/04/2023-03.csv

Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)... 157.230.72.27

Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443...  
connected.

HTTP request sent, awaiting response... 200 OK

Length: 58449187 (56M) [text/csv]

Saving to: '2023-03.csv'

2023-03.csv 100%[=====>] 55.74M 34.4MB/s in 1.6s

2023-09-09 19:50:33 (34.4 MB/s) - '2023-03.csv' saved [58449187/58449187]

--2023-09-09 19:50:33-- https://ecobici.cdmx.gob.mx/wp-  
content/uploads/2023/05/2023-04.csv

Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)... 157.230.72.27

Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443...  
connected.

HTTP request sent, awaiting response... 200 OK

Length: 58119473 (55M) [text/csv]

Saving to: '2023-04.csv'

2023-04.csv 100%[=====>] 55.43M 48.4MB/s in 1.1s

2023-09-09 19:50:35 (48.4 MB/s) - '2023-04.csv' saved [58119473/58119473]

--2023-09-09 19:50:35-- https://ecobici.cdmx.gob.mx/wp-  
content/uploads/2023/06/2023-05.csv

Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)... 157.230.72.27

Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443...  
connected.

HTTP request sent, awaiting response... 200 OK

Length: 67532130 (64M) [text/csv]

Saving to: '2023-05.csv'

2023-05.csv 100%[=====>] 64.40M 48.6MB/s in 1.3s

2023-09-09 19:50:36 (48.6 MB/s) - '2023-05.csv' saved [67532130/67532130]

```
--2023-09-09 19:50:36-- https://ecobici.cdmx.gob.mx/wp-
content/uploads/2023/07/2023_06.csv
Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)... 157.230.72.27
Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443...
connected.
HTTP request sent, awaiting response... 200 OK
Length: 72151312 (69M) [text/csv]
Saving to: '2023_06.csv'
```

2023\_06.csv 100%[=====>] 68.81M 29.4MB/s in 2.3s

2023-09-09 19:50:39 (29.4 MB/s) - '2023\_06.csv' saved [72151312/72151312]

```
--2023-09-09 19:50:39-- https://ecobici.cdmx.gob.mx/wp-
content/uploads/2023/08/2023_07.csv
Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)... 157.230.72.27
Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443...
connected.
HTTP request sent, awaiting response... 200 OK
Length: 71416339 (68M) [text/csv]
Saving to: '2023_07.csv'
```

2023\_07.csv 100%[=====>] 68.11M 52.8MB/s in 1.3s

2023-09-09 19:50:41 (52.8 MB/s) - '2023\_07.csv' saved [71416339/71416339]

```
--2023-09-09 19:50:41-- https://ecobici.cdmx.gob.mx/wp-
content/uploads/2023/09/2023_08.csv
Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)... 157.230.72.27
Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443...
connected.
HTTP request sent, awaiting response... 200 OK
Length: 79235355 (76M) [text/csv]
Saving to: '2023_08.csv'
```

2023\_08.csv 100%[=====>] 75.56M 54.8MB/s in 1.4s

2023-09-09 19:50:42 (54.8 MB/s) - '2023\_08.csv' saved [79235355/79235355]

```
[4]: ##Crear directorios HDFS, subir archivos y eliminar archivos locales
! hdfs dfs -mkdir -p /tmp/dcd/ecobici/
! hdfs dfs -put *.csv /tmp/dcd/ecobici/
! rm *.csv
```

[5]: *##Ver el contenido HDFS*

```
! hdfs dfs -ls -R /tmp/dcd/ecobici/
```

```
-rw-r--r--  2 root hadoop  41926344 2023-09-09 19:52
/tmp/dcd/ecobici/2023-01.csv
-rw-r--r--  2 root hadoop  47532884 2023-09-09 19:52
/tmp/dcd/ecobici/2023-02.csv
-rw-r--r--  2 root hadoop  58449187 2023-09-09 19:52
/tmp/dcd/ecobici/2023-03.csv
-rw-r--r--  2 root hadoop  58119473 2023-09-09 19:52
/tmp/dcd/ecobici/2023-04.csv
-rw-r--r--  2 root hadoop  67532130 2023-09-09 19:52
/tmp/dcd/ecobici/2023-05.csv
-rw-r--r--  2 root hadoop  72151312 2023-09-09 19:52
/tmp/dcd/ecobici/2023_06.csv
-rw-r--r--  2 root hadoop  71416339 2023-09-09 19:52
/tmp/dcd/ecobici/2023_07.csv
-rw-r--r--  2 root hadoop  79235355 2023-09-09 19:52
/tmp/dcd/ecobici/2023_08.csv
```

[6]: *##Crear un DF con spark*

```
eb_df = spark.read.csv('hdfs:///tmp/dcd/ecobici/', header=True,
↳inferSchema=True)
```

[7]: *##Ver los 10 primeros renglones*

```
eb_df.show(10)
```

```
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
|Genero_Usuario|Edad_Usuario|  Bici|Ciclo_Estacion_Retiro|Fecha_Retiro|
Hora_Retiro|Ciclo_EstacionArribo|Fecha Arribo|      Hora_Arribo|
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
|          M|          24|2252039|          007|
31/07/2023|2023-09-09 23:52:...|          064| 01/08/2023|2023-09-09
01:00:...|
|          M|          33|8626897|          206|
31/07/2023|2023-09-09 23:48:...|          212| 01/08/2023|2023-09-09
01:00:...|
|          M|          34|4940557|          215|
31/07/2023|2023-09-09 23:55:...|          212| 01/08/2023|2023-09-09
01:00:...|
|          F|          30|2036523|          291|
31/07/2023|2023-09-09 23:32:...|          082| 01/08/2023|2023-09-09
01:01:...|
|          F|          23|8079220|          546|
```

```

31/07/2023|2023-09-09 23:51:...|          498| 01/08/2023|2023-09-09
01:01:...|
|          M|          20|7019979|          306|
31/07/2023|2023-09-09 23:47:...|          375| 01/08/2023|2023-09-09
01:01:...|
|          F|          30|2998797|          065|
31/07/2023|2023-09-09 23:42:...|          016| 01/08/2023|2023-09-09
01:01:...|
|          M|          33|2085892|          221|
31/07/2023|2023-09-09 23:37:...|          029| 01/08/2023|2023-09-09
01:01:...|
|          M|          30|5081708|          394|
31/07/2023|2023-09-09 23:58:...|          372| 01/08/2023|2023-09-09
01:01:...|
|          M|          29|4998803|          375|
31/07/2023|2023-09-09 23:57:...|          398| 01/08/2023|2023-09-09
01:01:...|
+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+
only showing top 10 rows

```

```

[8]: #Ver el numero de renglones, numero de columnas y las columnas del DF
print(f"Renglones {eb_df.count()}")
print(f"Columnas {len(eb_df.columns)}")
print(f"Columnas: {eb_df.columns}")

```

```

[Stage 3:=====> (5 + 1) / 6]

```

```

Renglones 6947149
Columnas 9
Columnas: ['Genero_Usuario', 'Edad_Usuario', 'Bici', 'Ciclo_Estacion_Retiro',
'Fecha_Retiro', 'Hora_Retiro', 'Ciclo_EstacionArribo', 'Fecha_Arribo',
'Hora_Arribo']

```

```

[9]: ##Crea un nuevo DF con solo 4 columnas
sub_eb_df = eb_df.select(['Genero_Usuario', 'Edad_Usuario', 'Bici',
↪ 'Fecha_Retiro'])
sub_eb_df.show(10)
print(f"Renglones {sub_eb_df.count()}")

```

```

+-----+-----+-----+-----+
|Genero_Usuario|Edad_Usuario|    Bici|Fecha_Retiro|
+-----+-----+-----+-----+
|          M|          24|2252039| 31/07/2023|
|          M|          33|8626897| 31/07/2023|
|          M|          34|4940557| 31/07/2023|
|          F|          30|2036523| 31/07/2023|

```

	F	23	8079220	31/07/2023
	M	20	7019979	31/07/2023
	F	30	2998797	31/07/2023
	M	33	2085892	31/07/2023
	M	30	5081708	31/07/2023
	M	29	4998803	31/07/2023

+-----+-----+-----+-----+

only showing top 10 rows

[Stage 7:=====> (5 + 1) / 6]

Renglones 6947149

```
[10]: ##Generar un nuevo DF quitando duplicados
sub_nd_eb_df = sub_eb_df.dropDuplicates()
print(f"Renglones {sub_nd_eb_df.count()}")
```

[Stage 12:=====> (3 + 1) / 4]

Renglones 6010044

```
[11]: ##Ver estadisticos basicos
sub_nd_eb_df.describe().show()
```

[Stage 18:=====> (3 + 1) / 4]

	Genero_Usuario	Edad_Usuario	Bici	Fecha_Retiro
count	6010044	6010044	6010044	6010044
mean	null	34.62895452592612	5471698.783170473	null
stddev	null	10.010392880603918	2026433.8338036423	null
min	?	100	2001188	01/01/2023
max	0	NULL	8998171	31/12/2022

```
[12]: sub_nd_eb_df.write.save("gs://dcd05-sir-bucket/dcd/pyspark/ecobici/",
format='csv', header=True)
```

```
[ ]:
```



---

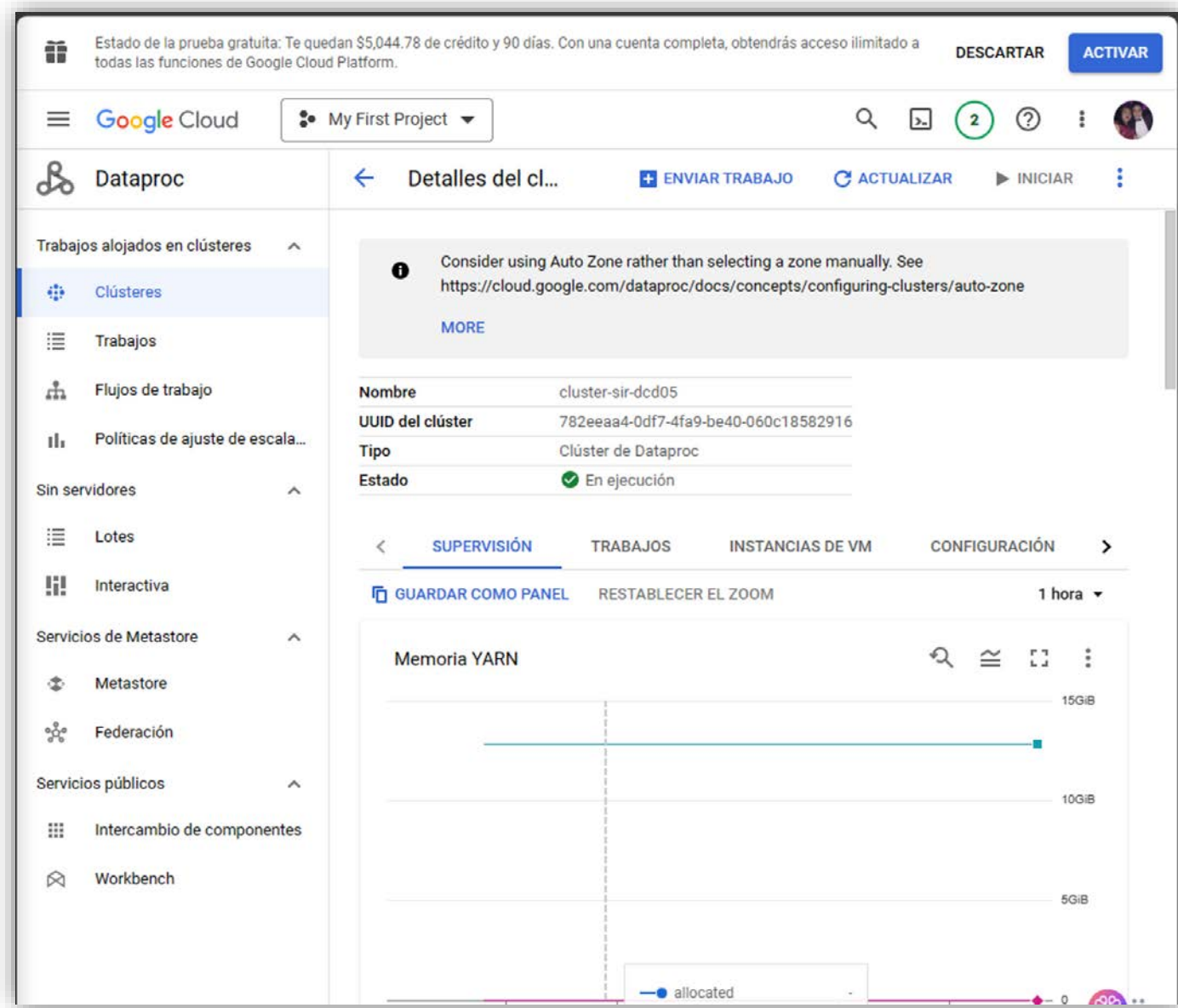
# **Diplomado en Ciencia de Datos UNAM**

## **Módulo 13 Datos Masivos**

### **Septiembre de 2023**

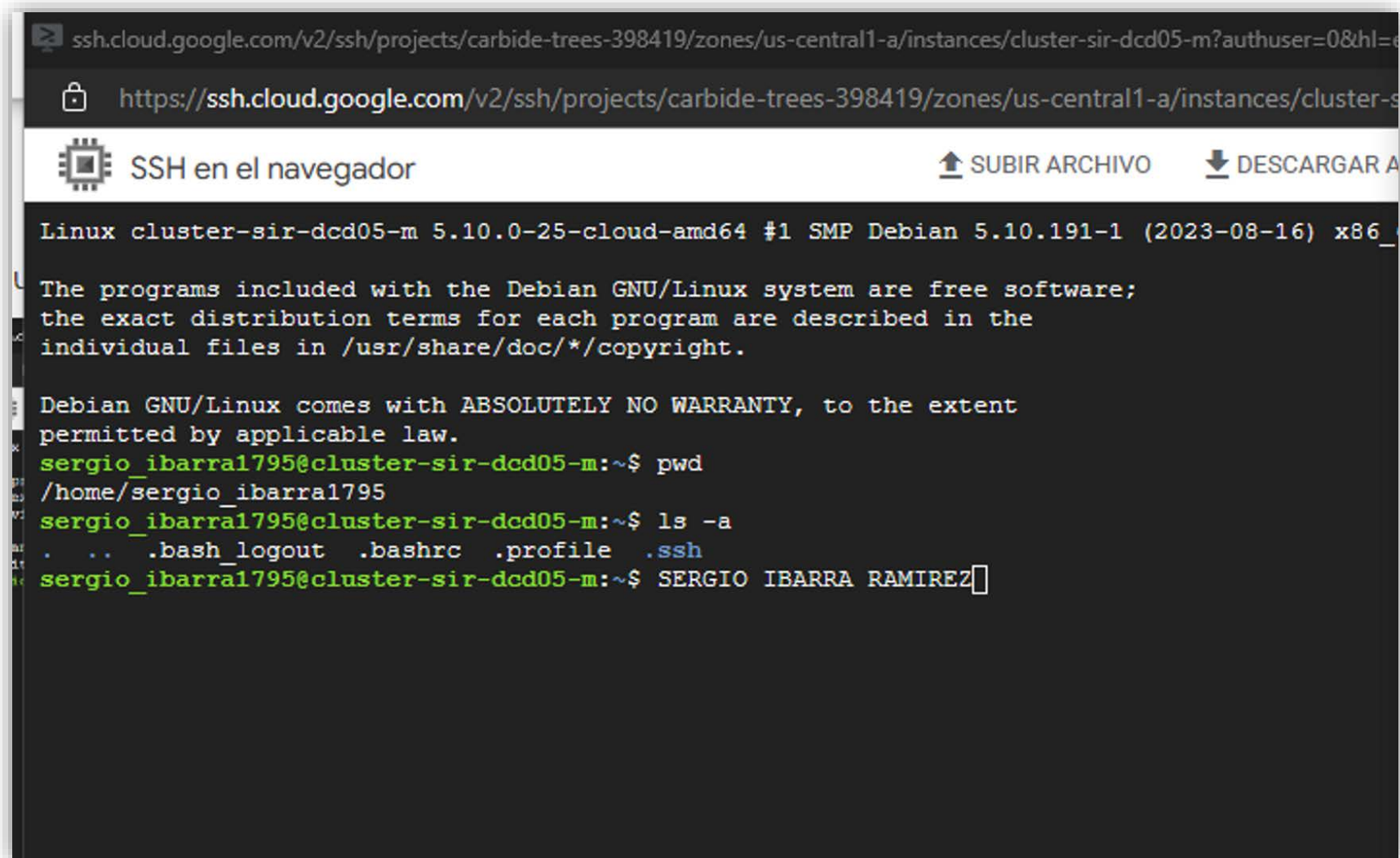
**Sergio Ibarra Ramírez**

# Pantalla de Cluster en Ejecucion





# Pantalla de SSH del Master



The screenshot shows a web browser window with the address bar displaying the URL: `https://ssh.cloud.google.com/v2/ssh/projects/carbide-trees-398419/zones/us-central1-a/instances/cluster-sir-dcd05-m?authuser=0&hl=es`. The browser's title bar reads "SSH en el navegador". In the top right corner, there are two buttons: "SUBIR ARCHIVO" (Upload File) and "DESCARGAR ARCHIVO" (Download File). The main content area is a terminal window with a black background and white text. The terminal output is as follows:

```
Linux cluster-sir-dcd05-m 5.10.0-25-cloud-amd64 #1 SMP Debian 5.10.191-1 (2023-08-16) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
sergio_ibarra1795@cluster-sir-dcd05-m:~$ pwd
/home/sergio_ibarra1795
sergio_ibarra1795@cluster-sir-dcd05-m:~$ ls -la
.  ..  .bash_logout  .bashrc  .profile  .ssh
sergio_ibarra1795@cluster-sir-dcd05-m:~$ SERGIO IBARRA RAMIREZ
```

# MapReduce Job History

carbide-trees-398419 > cluster-sir-1-105 [Sign out](#)

Logged in as: dr.who

## MapReduce Job job\_1694277344266\_0001

**Application**

- Job**
  - [Overview](#)
  - [Counters](#)
  - [Configuration](#)
  - [Map tasks](#)
  - [Reduce tasks](#)
- Tools**

**Job Overview**

**Job Name:** word count  
**User Name:** sergio\_ibarra1795  
**Queue:** default  
**State:** SUCCEEDED  
**Uberized:** false  
**Submitted:** Sat Sep 09 18:08:48 UTC 2023  
**Started:** Sat Sep 09 18:09:03 UTC 2023  
**Finished:** Sat Sep 09 18:09:28 UTC 2023  
**Elapsed:** 25sec  
**Diagnostics:**  
**Average Map Time:** 9sec  
**Average Shuffle Time:** 7sec  
**Average Merge Time:** 0sec  
**Average Reduce Time:** 0sec

**ApplicationMaster**

Attempt Number	Start Time	Node	Logs
1	Sat Sep 09 18:08:56 UTC 2023	cluster-sir-dcd05-w-0.us-central1-a.c.carbide-trees-398419.internal:8042	<a href="#">/gateway/default/jobhistory/logs</a>

Task Type	Total	Complete	
<b>Map</b>	1	1	
<b>Reduce</b>	3	3	
Attempt Type	Failed	Killed	Successful
<b>Maps</b>	0	0	1
<b>Reduces</b>	0	0	3

carbide-trees-398419 > cluster-sir-1-105 [Sign out](#)

Hadoop

[Overview](#) [Datanodes](#) [Datanode Volume Failures](#) [Snapshot](#) [Startup Progress](#) [Utilities](#)

## Browse Directory

/tmp/dcd [Go!](#) [Folder](#) [Upload](#) [Download](#)

Show  entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
<input type="checkbox"/>	drwxr-xr-x	<a href="#">sergio_ibarra1795</a>	<a href="#">hadoop</a>	0 B	Sep 09 12:39	0	0 B	<a href="#">sirilo</a> <a href="#">Trash</a>
<input type="checkbox"/>	drwxr-xr-x	<a href="#">sergio_ibarra1795</a>	<a href="#">hadoop</a>	0 B	Sep 09 12:09	0	0 B	<a href="#">wordcount</a> <a href="#">Trash</a>

Showing 1 to 2 of 2 entries [Previous](#) [1](#) [Next](#)

Hadoop, 2022.

# Jupyter Ecobici

The screenshot shows the Google Cloud Console interface. The left sidebar displays the 'Cloud Storage' menu with options for Buckets, Supervisión, and Configuración. The main panel shows the 'Detalles del bucket' for 'dcd05-sir-bucket'. Below the bucket details, there are tabs for OBJETOS, CONFIGURACIÓN, PERMISOS, PROTECCIÓN, and CICLO DE VIDA. The 'OBJETOS' tab is active, showing a list of objects. The list includes a file named '\_SUCCESS' and several 'part' files. The selected file is 'part-00003-77345103-55b8-4b8f-...' with a size of 35.1 MB.

Nombre	Tamaño	Tipo	Fecha de c
_SUCCESS	0 B	application/octet-stream	9 sept 20;
part-00000-77345103-55b8-4b8f-...	34.4 MB	application/octet-stream	9 sept 20;
part-00001-77345103-55b8-4b8f-...	34.4 MB	application/octet-stream	9 sept 20;
part-00002-77345103-55b8-4b8f-...	33.7 MB	application/octet-stream	9 sept 20;
part-00003-77345103-55b8-4b8f-...	35.1 MB	application/octet-stream	9 sept 20;

```
In [10]: ##Generar un nuevo DF quitando duplicados
sub_nd_eb_df = sub_eb_df.dropDuplicates()
print(f"Renglones {sub_nd_eb_df.count()}")

[Stage 12:=====> (3 + 1) / 4]

Renglones 6010044

In [11]: ##Ver estadisticos basicos
sub_nd_eb_df.describe().show()
```

summary	Genero_Usuario	Edad_Usuario	Bici	Fecha_Retiro
count	6010044	6010044	6010044	6010044
mean	null	34.62895452592612	5471698.783170473	null
stddev	null	10.010392880603918	2026433.8338036423	null
min	?	100	2001188	01/01/2023
max	0	NULL	8998171	31/12/2022

```
In [9]: ##Crea un nuevo DF con solo 4 columnas
sub_eb_df = eb_df.select(['Genero_Usuario', 'Edad_Usuario', 'Bici', 'Fecha_Retiro'])
sub_eb_df.show(10)
print(f"Renglones {sub_eb_df.count()}")
```

Genero_Usuario	Edad_Usuario	Bici	Fecha_Retiro
M	24	2252039	31/07/2023
M	33	8626897	31/07/2023
M	34	4940557	31/07/2023
F	30	2036523	31/07/2023
F	23	8079220	31/07/2023
M	20	7019979	31/07/2023
F	30	2998797	31/07/2023
M	33	2085802	31/07/2023
M	30	5081708	31/07/2023
M	29	4998803	31/07/2023

only showing top 10 rows

```
[Stage 7:=====> (5 + 1) / 6]

Renglones 6947149
```