

TRABAJO FINAL

RESULTADOS DE LABORATORIOS

Curso: Diploma Data Engineer - Edición 11

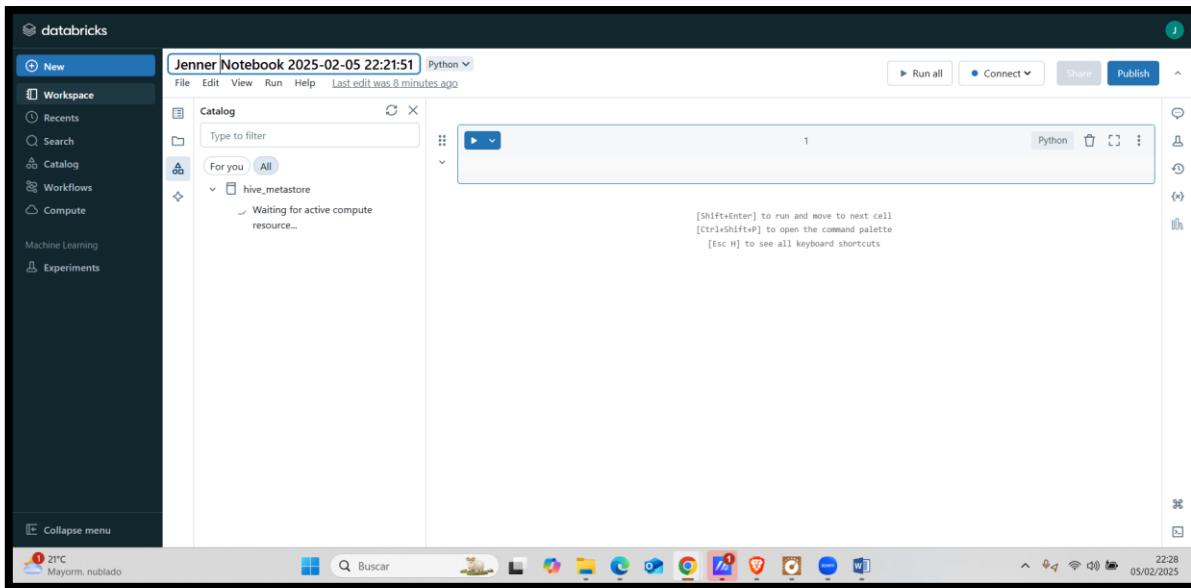
Alumno: Jenner Conco Dextre

ÍNDICE

SESIÓN 01	3
SESIÓN 02	3
SESIÓN 03	7
SESIÓN 04	10
SESIÓN 05	19
SESIÓN 06	22
SESIÓN 07	27
SESIÓN 08	30
SESIÓN 09	42
SESIÓN 10	44
SESIÓN 11	49
SESIÓN 12	52

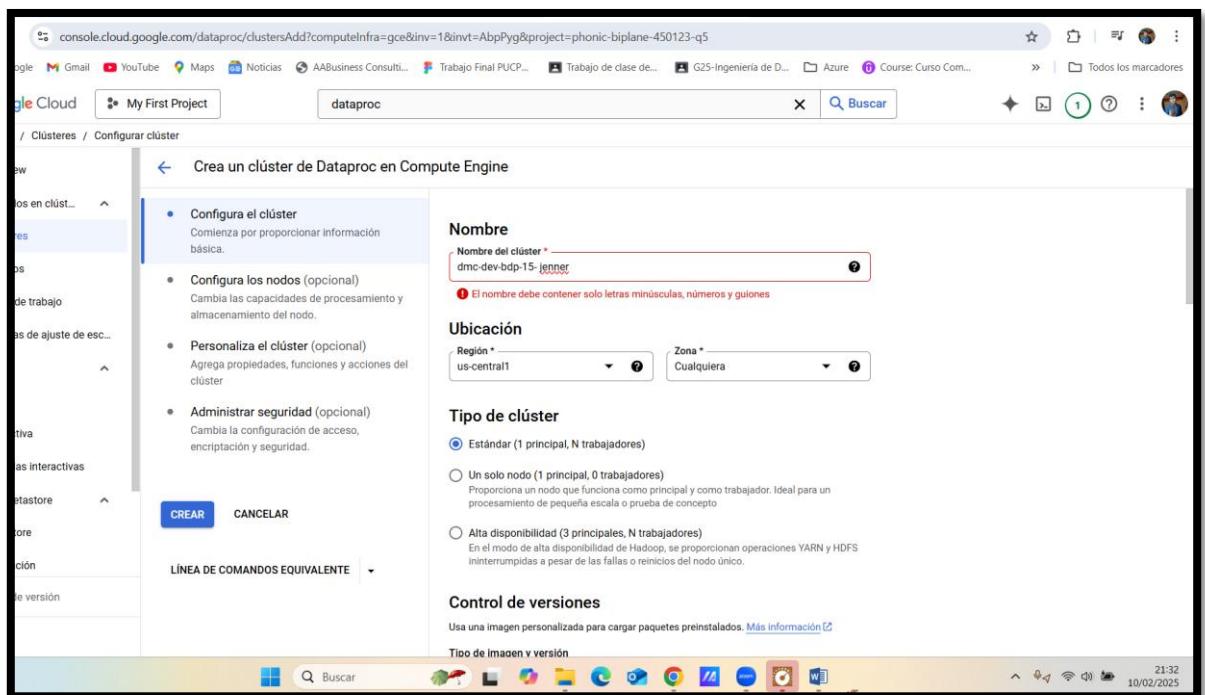
SESIÓN 01

1. Laboratorio – Creación de un Cluster en Databricks.



SESIÓN 02

1. Creando cluster en GCP.



Google Cloud / Clústeres / Configurar clúster

Crea un clúster de Dataproc en Compute Engine

Configura el clúster

Configura los nodos (opcional)

Personaliza el clúster (opcional)

Administrar seguridad (opcional)

Nodo de administrador

De uso general (selected)

Optimizado para procesamiento

Con optimización de memoria

GPU

Tipos de máquinas para cargas de trabajo comunes, optimizados en función del costo y la flexibilidad

Serie N2

n2-standard-2 (2 CPU virtuales, 1 núcleo, 8 GB de memoria)

vCPU 2

Memory 8 GB

PLATAFORMA DE CPU Y GPU

Tamaño del disco principal * 50 GB

Primary disk type * SSD Persistent Disk

Cantidad de SSD locales * x 375GB

Interfaz de SSD local SCSI

CREAR CANCELAR

Google Cloud / Clústeres / Configurar clúster

Crea un clúster de Dataproc en Compute Engine

Configura el clúster

Configura los nodos (opcional)

Personaliza el clúster (opcional)

Administrar seguridad (opcional)

Solo IP internas

Configurar todas las instancias para tener solo direcciones IP internas. [Learn more](#)

Etiquetas

+ AGREGAR ETIQUETAS

Propiedades del clúster

Usa las propiedades del clúster para agregar o modificar archivos de configuración cuando creas un clúster.

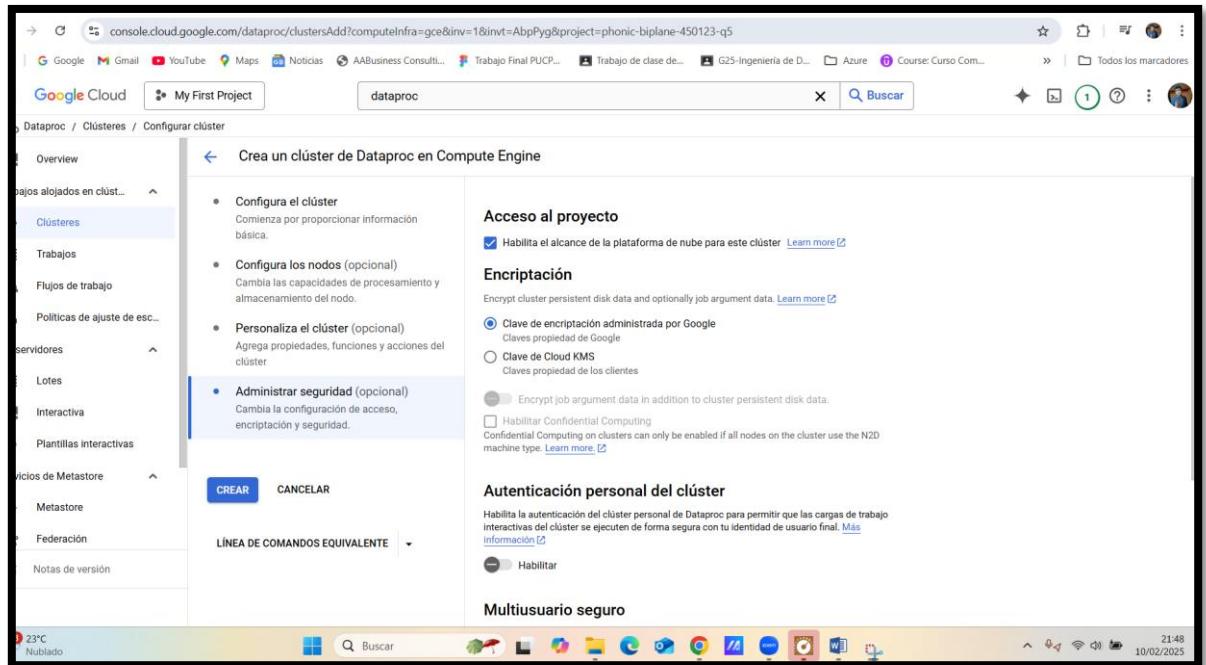
+ AGREGAR PROPIEDADES

Acciones de inicialización

Utiliza acciones de inicialización para personalizar la configuración, instalar aplicaciones o realizar otras modificaciones en tu clúster. Selecciona secuencias de comandos o ejecutables que Cloud Dataproc ejecutará cuando aprovisiona tu clúster.

+ AGREGAR ACCIÓN DE INICIALIZACIÓN

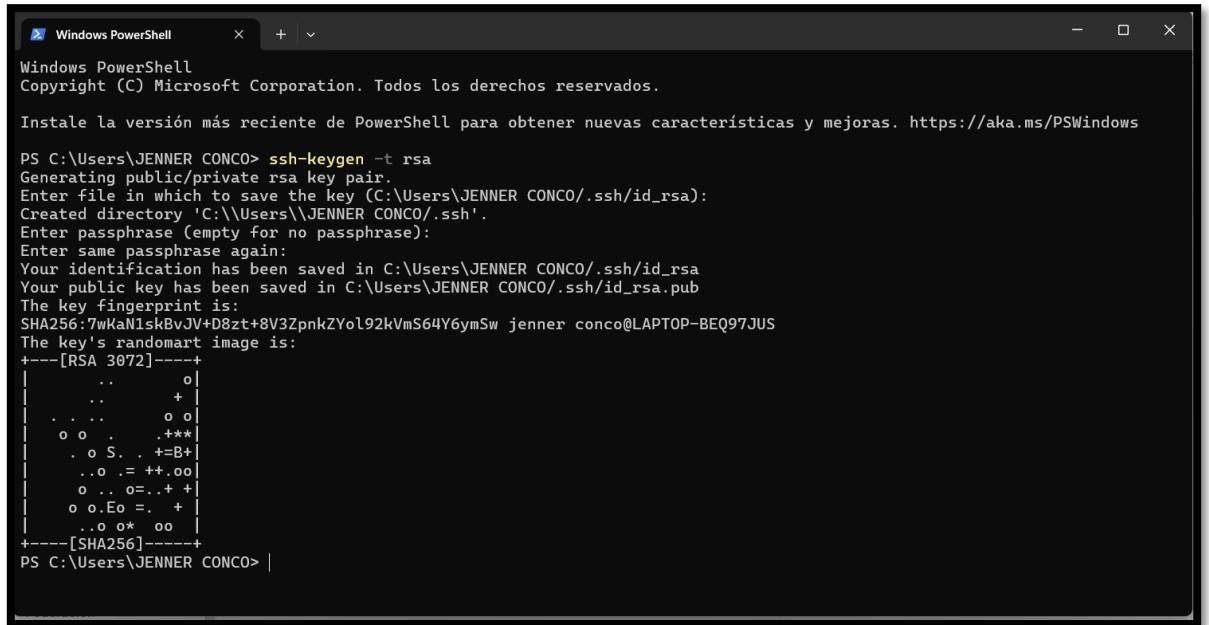
CREAR CANCELAR



- Aprovisionamiento de cluster.

Nombre	Estado	Región	Zona	Total de nodos
dmc-dev-bdp-15	En ejecución	us-central1	us-central1-f	3

- Creación de Llave pública.



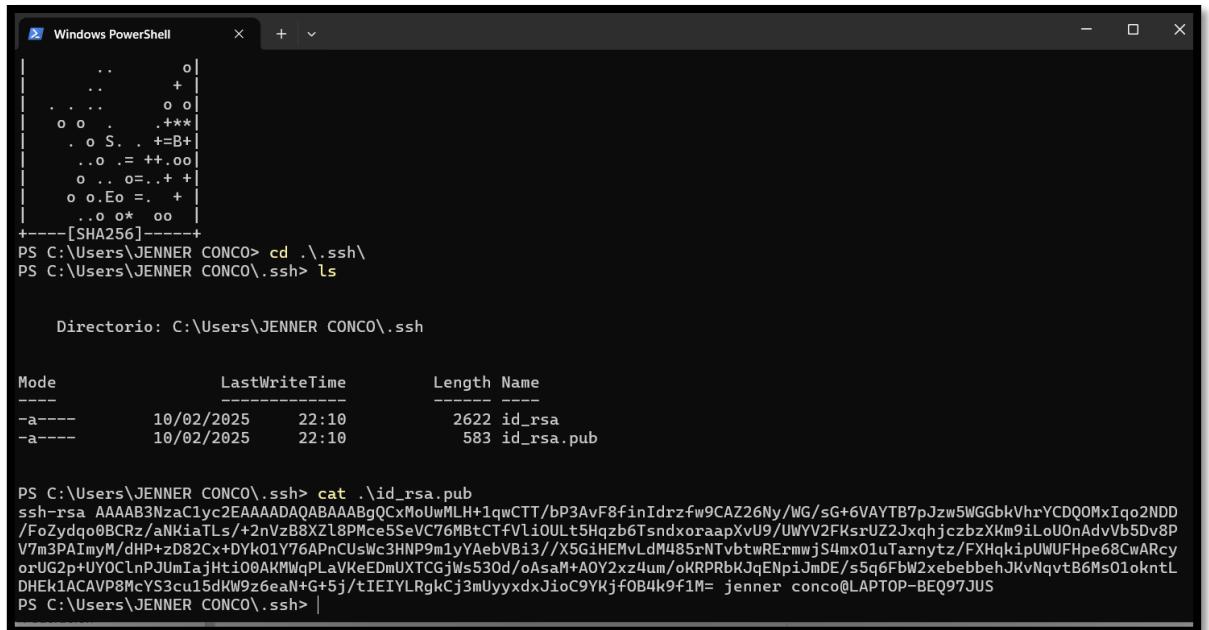
```

Windows PowerShell
Copyright (C) Microsoft Corporation. Todos los derechos reservados.

Instale la versión más reciente de PowerShell para obtener nuevas características y mejoras. https://aka.ms/PSWindows

PS C:\Users\JENNER CONCO> ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (C:/Users/JENNER CONCO/.ssh/id_rsa):
Created directory 'C:/\Users\JENNER CONCO/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in C:/Users/JENNER CONCO/.ssh/id_rsa
Your public key has been saved in C:/Users/JENNER CONCO/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:7wKaN1skBvJV+D8zt+8V3ZpnkZYol92kVmS64Y6ymSw jenner conco@LAPTOP-BEQ97JUS
The key's randomart image is:
+---[RSA 3072]---+
| .. o |
| .. + |
| . . o o |
| o o . .+**|
| . o S. . +=B+|
| ..o .= ++.oo|
| o .. o=..+ |
| o o.Eo =. + |
| ..o o* oo |
+---[SHA256]---+
PS C:\Users\JENNER CONCO> |

```



```

Windows PowerShell
Copyright (C) Microsoft Corporation. Todos los derechos reservados.

PS C:\Users\JENNER CONCO> cd .\.ssh\
PS C:\Users\JENNER CONCO\.ssh> ls

Directorio: C:\Users\JENNER CONCO\.ssh

Mode                LastWriteTime         Length Name
----                -----          ----  --
-a---        10/02/2025      22:10       2622 id_rsa
-a---        10/02/2025      22:10       583 id_rsa.pub

PS C:\Users\JENNER CONCO\.ssh> cat .\id_rsa.pub
ssh-rsa AAAAB3NzaC1yc2EAAAQABAAABgQCxMoUwMLH+1qwCTT/bP3AvF8finIdrzfw9CAZ26Ny/WG/sG+6VAYTB7pJzw5WGgbkVhrYCDQ0MxIqo2NDD
/FoZydqo0BCRz/aNKiaTLs/+2nVzB8XZl8PMce5SeVC76MbTCfVliOULt5Hqzb6TsndxoraapXvU9/UWVV2FKsrUZ2JxqhjczbXkm9iLoUOnAdvVb5Dv8P
V7m3PAImyM/dHP+zD82Cx+DYk01Y76APnCUwC3HNp9m1yYAebVBi3//X5GiHEMvLdM485rNTvbktwERmwjS4mx01uTarnytz/FXHqkipUWFHpe68CwARcy
orUG2pt+UYOClnPJUmIajHti00AKMwqPLaVKeeDmUXTCGjWs530d/oAsaM+A0Y2xz4um/oKRPPrbKJqEnpiJmDE/s5q6FbW2xebebbehJKvNqvtB6Ms01okntL
DHEk1ACAVP8McYS3cu15dKW9z6eaN+G+5j/tIEIYLrgkCj3mUyyxdxJioC9YKjfOB4k9f1M= jenner conco@LAPTOP-BEQ97JUS
PS C:\Users\JENNER CONCO\.ssh> |

```

- Se logró realizar la conexión con moba.

```

34.173.131.140
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
Quick connect...
/home/jenner/
Name
.. .ssh .bash_logout .bashrc .profile .Xauthority .comindos.txt empresas2050201.csv empresas2050202.csv empresas2050203.txt empresas20504.csv persona.data

login as: jenner
Authenticating with public key "jenner"
MobaXterm Personal Edition v24.1
(SSH client, X server and network tools)

> SSH session to jenner@34.173.131.140
* Direct SSH : ✓
* SSH compression : ✓
* SSH-browser : ✓
* X11-forwarding : ✓ (remote display is forwarded through SSH)

> For more info, ctrl+click on help or visit our website.

Linux dmc-dev-bdp-15-m 6.1.0-30-cloud-amd64 #1 SMP PREEMPT_DYNAMIC Debian 6.1.124-1 (2025-01-12) 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.
/usr/bin/xauth: file /home/jenner/.Xauthority does not exist
jenner@dmc-dev-bdp-15-m:~$ cd /
bin boot copyright dev etc hadoop home lib lib64 lost+found media mnt opt proc root run sbin srv sys tmp usr var
jenner@dmc-dev-bdp-15-m:~$ cd
jenner@dmc-dev-bdp-15-m:~$ pwd
/home/jenner
jenner@dmc-dev-bdp-15-m:~$ ls /
jenner@dmc-dev-bdp-15-m:~$ ls /

```

SESIÓN 03

- Se logró la creación de carpeta jennerconco en moba.

```

Sessions /home/jenner/
Name
.. .ssh .bash_logout .bashrc .profile .Xauthority .comindos.txt empresas2050201.csv empresas2050202.csv empresas2050203.txt empresas20504.csv persona.data

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.
/usr/bin/xauth: file /home/jenner/.Xauthority does not exist
jenner@dmc-dev-bdp-15-m:~$ cd /
jenner@dmc-dev-bdp-15-m:~$ ls
bin boot copyright dev etc hadoop home lib lib64 lost+found media mnt opt proc root run sbin srv sys tmp usr var
jenner@dmc-dev-bdp-15-m:~$ cd
jenner@dmc-dev-bdp-15-m:~$ pwd
/home/jenner
jenner@dmc-dev-bdp-15-m:~$ ls /
bin boot copyright dev etc hadoop home lib lib64 lost+found media mnt opt proc root run sbin srv sys tmp usr var
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /
Found 3 items
drwxrwxrwt - hdfs hadoop 0 2025-02-16 21:06 /tmp
drwxrwxrwt - hdfs hadoop 0 2025-02-16 21:05 /user
drwxrwxrwt - hdfs hadoop 0 2025-02-16 21:05 /var
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -mkdir /jennerconco
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /
Found 4 items
drwxr-xr-x - jenner hadoop 0 2025-02-17 22:00 /jennerconco
drwxr-xr-x - jenner hadoop 0 2025-02-16 21:06 /tmp
drwxr-xr-x - jenner hadoop 0 2025-02-16 21:05 /user
drwxr-xr-x - jenner hadoop 0 2025-02-16 21:05 /var
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /
Found 6 items
drwxr-xr-x - jenner hadoop 0 2025-02-17 22:01 /carpeta1
drwxr-xr-x - jenner hadoop 0 2025-02-17 22:02 /carpeta2
drwxr-xr-x - jenner hadoop 0 2025-02-17 22:00 /jennerconco
drwxr-xr-x - hdfs hadoop 0 2025-02-16 21:06 /tmp
drwxr-xr-x - hdfs hadoop 0 2025-02-16 21:05 /user
drwxr-xr-x - hdfs hadoop 0 2025-02-16 21:05 /var
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -mkdir /jennerconco/carpeta1
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -mkdir /jennerconco/carpeta2
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /jennerconco
Found 2 items
drwxr-xr-x - jenner hadoop 0 2025-02-17 22:03 /jennerconco/carpeta1
drwxr-xr-x - jenner hadoop 0 2025-02-17 22:03 /jennerconco/carpeta2

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net

```

- Se realizó la eliminación de archivos.

The screenshot shows the MobaXterm interface with a terminal window open. The terminal window title is '34.173.131.140'. The command entered was 'hdfs dfs -rm -f /jennerconco/persona.data'. The terminal output shows the file was deleted successfully, along with other files in the directory. The file system view on the left shows a folder named 'Name' containing various files like '.bash_logout', '.profile', '.Xauthority', 'comandos.txt', 'empresa20250201.csv', 'empresa20250202.csv', 'empresa20250203.txt', 'empresa202504.csv', and 'persona.data'.

```

34.173.131.140
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
Quick connect...
[2 34.173.131.140]
/home/jenner/
└── Name
    ├── .bash_logout
    ├── .profile
    ├── .Xauthority
    ├── comandos.txt
    ├── empresa20250201.csv
    ├── empresa20250202.csv
    ├── empresa20250203.txt
    ├── empresa202504.csv
    └── persona.data

88|Zeus|1-525-361-5753|non_magna.Nam@nuncid.org|2008-09-15|29|21556|7
89|Kelly|430-9134|turpis.In@egestasfusce.ca|2016-04-01|55|10110|6
90|Stjourney|598-8806|sit.amet@sem.net|2002-05-02|42|9145|6
91|Erica|1-320-339-2705|tristique.senectus@leifendnondapibus.org|2015-02-23|32|8934|6
92|Lesley|973-4982|felis.adipiscing.fringilla@Pronvelnsl.co.uk|2004-02-29|19|23547|4
93|Althea|1-163-702-1244|sit.amet.null@elit.co.uk|2002-09-01|24|8818|1
94|Amir|1-221-717-0093|dapibus.liquida@faucibuslectus.com|2001-02-11|18|20980|4
95|Jayme|1-641-113-8418|ornare.Fusce@inciduntcongueutris.co.uk|2002-05-31|58|23975|4
96|Amos|729-4665|non.lacunatapharetra.net|2017-11-27|42|15855|2
97|Flavia|1-559-270-7164|erat.vel.pede@sedtortor.co.uk|2004-11-14|27|13473|3
98|Uli|772-9996|sagittis@isl.com|2005-04-17|57|13361|5
99|Ray|1-420-314-2886|ac.risus.Morbi@elluseuauug.org|2011-12-30|26|5570|1
100|Cynthia|148-9696|justo.nec.ante@tinciduntvehicula.org|2008-01-23|57|8682|5jenner@dmc-dev-bdp-15-m:~$ ls
comandos.txt  empresa20250201.csv  empresa20250202.csv  empresa20250203.txt  empresa202504.csv  persona.data
jenner@dmc-dev-bdp-15-m:~$ pwd
/home/jenner
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -rm -f /jennerconco/persona.data
Deleted /jennerconco/persona.data
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /jennerconco
Found 9 items
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:12 /jennerconco/_ERROR
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:12 /jennerconco/_SUCCESS
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:11 /jennerconco/archivo_vacio.txt
drwxr-xr-x  - jenner hadoop     0 2025-02-17 22:09 /jennerconco/carpeta1
drwxr-xr-x  - jenner hadoop     0 2025-02-17 22:03 /jennerconco/carpeta2
-rw-r--r--  2 jenner hadoop     0 2025-02-17 23:07 /jennerconco/empresa20250201.csv
-rw-r--r--  2 jenner hadoop     0 2025-02-17 23:07 /jennerconco/empresa20250202.csv
-rw-r--r--  2 jenner hadoop     0 2025-02-17 23:12 /jennerconco/empresa202504.csv
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -rm -f /jennerconco/empresa*
Deleted /jennerconco/empresa20250201.csv
Deleted /jennerconco/empresa20250202.csv
Deleted /jennerconco/empresa202504.csv
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /jennerconco
Found 6 items
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:12 /jennerconco/_ERROR
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:12 /jennerconco/_SUCCESS
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:11 /jennerconco/archivo_vacio.txt
drwxr-xr-x  - jenner hadoop     0 2025-02-17 22:09 /jennerconco/carpeta1
drwxr-xr-x  - jenner hadoop     0 2025-02-17 22:03 /jennerconco/carpeta2
-rw-r--r--  2 jenner hadoop     0 2025-02-17 23:12 /jennerconco/empresa20250203.txt
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -rm -r -f /jennerconco/carpeta1
Deleted /jennerconco/carpeta1
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /jennerconco
Found 5 items
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:12 /jennerconco/_ERROR
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:12 /jennerconco/_SUCCESS
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:11 /jennerconco/archivo_vacio.txt
drwxr-xr-x  - jenner hadoop     0 2025-02-17 22:03 /jennerconco/carpeta2
-rw-r--r--  2 jenner hadoop     0 2025-02-17 23:12 /jennerconco/empresa20250203.txt
jenner@dmc-dev-bdp-15-m:~$
```

- Eliminando carpeta1

The screenshot shows the MobaXterm interface with a terminal window open. The command entered was 'hdfs dfs -rm -r -f /jennerconco/carpeta1'. The terminal output shows the directory was deleted successfully, along with other files in the directory. The file system view on the left shows a folder named 'Name' containing various files like '.bash_logout', '.profile', '.Xauthority', 'comandos.txt', 'empresa20250201.csv', 'empresa20250202.csv', 'empresa20250203.txt', 'empresa202504.csv', and 'persona.data'.

```

Deleted /jennerconco/empresa202504.csv
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /jennerconco
Found 6 items
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:12 /jennerconco/_ERROR
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:12 /jennerconco/_SUCCESS
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:11 /jennerconco/archivo_vacio.txt
drwxr-xr-x  - jenner hadoop     0 2025-02-17 22:09 /jennerconco/carpeta1
drwxr-xr-x  - jenner hadoop     0 2025-02-17 22:03 /jennerconco/carpeta2
-rw-r--r--  2 jenner hadoop     0 2025-02-17 23:12 /jennerconco/empresa20250203.txt
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -rm -r -f /jennerconco/carpeta1
Deleted /jennerconco/carpeta1
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /jennerconco
Found 5 items
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:12 /jennerconco/_ERROR
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:12 /jennerconco/_SUCCESS
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:11 /jennerconco/archivo_vacio.txt
drwxr-xr-x  - jenner hadoop     0 2025-02-17 22:03 /jennerconco/carpeta2
-rw-r--r--  2 jenner hadoop     0 2025-02-17 23:12 /jennerconco/empresa20250203.txt
jenner@dmc-dev-bdp-15-m:~$
```

- Visualizando los permisos.

```

jenner@jenner-OptiPlex-5090:~/hadoop$ hdfs dfs -ls /jennerconco/carpeta2
Found 5 items
-rw-r--r--  2 jenner hadoop      0 2025-02-17 23:12 /jennerconco/empresa20250203.txt
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:12 /jennerconco/_ERROR
-rw-r--r--  2 jenner hadoop      0 2025-02-17 22:12 /jennerconco/_SUCCESS
drwxr----  - dmc    bigdata    0 2025-02-17 22:03 /jennerconco/carpeta2
-rw-r--r--  2 jenner hadoop      0 2025-02-17 23:12 /jennerconco/empresa20250203.txt
jenner@jenner-OptiPlex-5090:~/hadoop$ hdfs dfs -chmod 740 /jennerconco/carpeta2
jenner@jenner-OptiPlex-5090:~/hadoop$ hdfs dfs -ls /jennerconco
base support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net

```

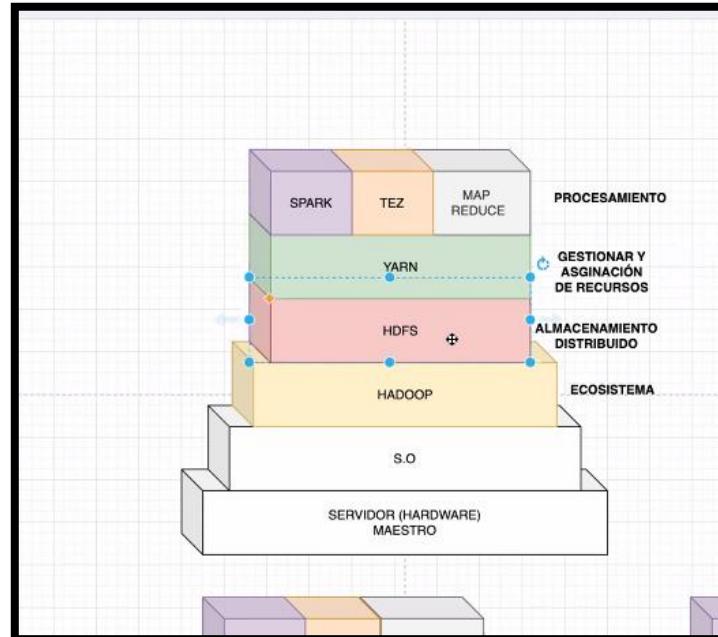
- Comandos utilizados, en carpeta comandos.txt

```

* comandos.txt
1 hdfs dfs -put /home/jenner/persona.data /jennerconco
2
3
4 hdfs dfs -put /home/jenner/empresax.csv /jennerconco
5
6
7 hdfs dfs -put /home/jenner/empresax /jennerconco
8
9 hdfs dfs -cat /home/jennerconco/persona.data
10
11
12 hdfs dfs -rm -f /jennerconco/persona.data
13
14 hdfs dfs -chown dmc:bigdata /jennerconco/carpeta2
15
16 hdfs dfs -chown -R dmc:bigdata /jennerconco/carpeta2

```

- Arquitectura a utilizar.



SESIÓN 04

- Asignacion de usuario y grupo.

```
usage: hadoop [generic options] -setfacl [-R] [-u|-g] [-m|-x] pathspec pathspec|||-set pathspec pathspec
jenner@dmc-dev-bdp-15-m:~$ ^C
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -setfacl -R -m user:leonor:rw- /jennerconco/archivo_vacio.txt
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -setfacl -R -m group:ulima:r-x /jennerconco/archivo_vacio.txt
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /jennerconco
Found 5 items
-rw-r--r--  2 jenner hadoop          0 2025-02-17 22:12 /jennerconco/_ERROR
-rw-r--r--  2 jenner hadoop          0 2025-02-17 22:12 /jennerconco/_SUCCESS
-rw-rwxr--+ 2 jenner hadoop          0 2025-02-17 22:11 /jennerconco/archivo_vacio.txt
drwxr-----  - dmc    bigdata        0 2025-02-17 22:03 /jennerconco/carpeta2
-rw-r--r--  2 jenner hadoop          0 2025-02-17 23:12 /jennerconco/empresa20250203.txt
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -getfacl /jennerconco/archivo_vacio.txt
# file: /jennerconco/archivo_vacio.txt
# owner: jenner
# group: hadoop
user::rw-
user:leonor:rw-
group::r--
group:ulima:r-x
mask::rwx
other::r--

jenner@dmc-dev-bdp-15-m:~$ █
support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net
```

- Subiendo archivo persona.data.

```
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -put /home/jenner/persona.data /jennerconco
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /jennerconco
Found 6 items
-rw-r--r--  2 jenner hadoop          0 2025-02-17 22:12 /jennerconco/_ERROR
-rw-r--r--  2 jenner hadoop          0 2025-02-17 22:12 /jennerconco/_SUCCESS
-rw-rwxr--+ 2 jenner hadoop          0 2025-02-17 22:11 /jennerconco/archivo_vacio.txt
drwxr-----  - dmc    bigdata        0 2025-02-17 22:03 /jennerconco/carpeta2
-rw-r--r--  2 jenner hadoop          0 2025-02-17 23:12 /jennerconco/empresa20250203.txt
-rw-r--r--  2 jenner hadoop          7282 2025-02-18 01:38 /jennerconco/persona.data
jenner@dmc-dev-bdp-15-m:~$ █
support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net
```

```
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -put /home/jenner/persona.data /jennerconco
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /jennerconco
Found 6 items
-rw-r--r--  2 jenner hadoop          0 2025-02-17 22:12 /jennerconco/_ERROR
-rw-r--r--  2 jenner hadoop          0 2025-02-17 22:12 /jennerconco/_SUCCESS
-rw-rwxr--+ 2 jenner hadoop          0 2025-02-17 22:11 /jennerconco/archivo_vacio.txt
drwxr-----  - dmc    bigdata        0 2025-02-17 22:03 /jennerconco/carpeta2
-rw-r--r--  2 jenner hadoop          0 2025-02-17 23:12 /jennerconco/empresa20250203.txt
-rw-r--r--  2 jenner hadoop          7282 2025-02-18 01:38 /jennerconco/persona.data
jenner@dmc-dev-bdp-15-m:~$ cksum /home/jenner/persona.data
229400302 7282 /home/jenner/persona.data
jenner@dmc-dev-bdp-15-m:~$ █
support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net
```

```

jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -put /home/jenner/persona.data /jennerconco
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /jennerconco
Found 6 items
-rw-r--r--  2 jenner hadoop          0 2025-02-17 22:12 /jennerconco/_ERROR
-rw-r--r--  2 jenner hadoop          0 2025-02-17 22:12 /jennerconco/_SUCCESS
-rw-rwxr--+ 2 jenner hadoop          0 2025-02-17 22:11 /jennerconco/archivo_vacio.txt
drwxr-----  - dmc    bigdata        0 2025-02-17 22:03 /jennerconco/carpeta2
-rw-r--r--  2 jenner hadoop          0 2025-02-17 23:12 /jennerconco/empresa20250203.txt
-rw-r--r--  2 jenner hadoop        7282 2025-02-18 01:38 /jennerconco/persona.data
jenner@dmc-dev-bdp-15-m:~$ cksum /home/jenner/persona.data
229400302 7282 /home/jenner/persona.data
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -cat /home/jenner/persona.data | cksum
cat: `/home/jenner/persona.data': No such file or directory
4294967295 0
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -cat /jennerconco/persona.data | cksum
229400302 7282
jenner@dmc-dev-bdp-15-m:~$
```

port MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

```

4294967295 0
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -cat /jennerconco/persona.data | cksum
229400302 7282
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -du -s -h '/*'
0      0      /carpeta1
0      0      /carpeta2
7.1 K  14.2 K  /jennerconco
0      0      /tmp
0      0      /user
0      0      /var
jenner@dmc-dev-bdp-15-m:~$
```

port MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

```

@      @      /var
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -setrep -w 3 -R /jennerconco
setrep: '-R': No such file or directory
Replication 3 set: /jennerconco/_ERROR
Replication 3 set: /jennerconco/_SUCCESS
Replication 3 set: /jennerconco/archivo_vacio.txt
Replication 3 set: /jennerconco/empresa20250203.txt
Replication 3 set: /jennerconco/persona.data
Waiting for /jennerconco/_ERROR ... done
Waiting for /jennerconco/_SUCCESS ... done
Waiting for /jennerconco/archivo_vacio.txt ... done
Waiting for /jennerconco/empresa20250203.txt ... done
Waiting for /jennerconco/persona.data .... done
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /jennerconco
Found 6 items
-rw-r--r--  3 jenner hadoop          0 2025-02-17 22:12 /jennerconco/_ERROR
-rw-r--r--  3 jenner hadoop          0 2025-02-17 22:12 /jennerconco/_SUCCESS
-rw-rwxr--+ 3 jenner hadoop          0 2025-02-17 22:11 /jennerconco/archivo_vacio.txt
drwxr-----  - dmc    bigdata        0 2025-02-17 22:03 /jennerconco/carpeta2
-rw-r--r--  3 jenner hadoop          0 2025-02-17 23:12 /jennerconco/empresa20250203.txt
-rw-r--r--  3 jenner hadoop        7282 2025-02-18 01:38 /jennerconco/persona.data
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -du -s -h '/*'
0      0      /carpeta1
0      0      /carpeta2
7.1 K  21.3 K  /jennerconco
0      0      /tmp
0      0      /user
0      0      /var
jenner@dmc-dev-bdp-15-m:~$
```

port MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

- Comandos para HIVE, realizando la conexión a HIVE.

```

jenner@dmc-dev-bdp-15-m:~$ beeline -u jdbc:hive2://
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/tez/lib/sl4j-reload4j-1.7.36.jar!/_org/sl4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/sl4j-reload4j-1.7.36.jar!/_org/sl4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/tez/lib/sl4j-reload4j-1.7.36.jar!/_org/sl4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/sl4j-reload4j-1.7.36.jar!/_org/sl4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
Connecting to jdbc:hive2://
Hive Session ID = 8ebfia36-ef90-4e87-9776-8129994a6c33
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.hive.common.StringInternUtils (file:/usr/lib/hive/lib/hive-common-3.1.3.jar) to field java.net.URI.str
ing
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.hive.common.StringInternUtils
WARNING: Use --illegal-access=warn to enable warnings or further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
25/02/18 02:15:48 [main]: WARN session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of Hi
veAuthorizerFactory.
Connected to: Apache Hive (version 3.1.3)
Driver: Hive JDBC (version 3.1.3)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.3 by Apache Hive
0: jdbc:hive2://> 
```



```

beeline version 3.1.3 by Apache Hive
0: jdbc:hive2://> SHOW DATABASES;
OK
+-----+
| database_name |
+-----+
| default       |
+-----+
1 row selected (1.131 seconds)
0: jdbc:hive2://> 
```

- Realizando consulta show databases.

```

1 row selected (1.131 seconds)
0: jdbc:hive2://> CREATE DATABASE bigdata;
OK
No rows affected (1.573 seconds)
0: jdbc:hive2://> SHOW DATABASES;
OK
+-----+
| database_name |
+-----+
| bigdata        |
| default        |
+-----+
2 rows selected (0.082 seconds)
0: jdbc:hive2://> 
```

port MobaXterm by subscribing to the professional edition here: <https://m>

- Resultado de la consulta.

```

2 rows selected (0.082 seconds)
0: jdbc:hive2://> CREATE SCHEMA miusuario_test;
OK
No rows affected (0.113 seconds)
0: jdbc:hive2://> SHOW DATABASES;
OK
+-----+
| database_name |
+-----+
| bigdata      |
| default      |
| miusuario_test |
+-----+
3 rows selected (0.07 seconds)
0: jdbc:hive2://> ■

```

Import MobaXterm by subscribing to the professional edition here: <https://mobaxterm.moba>

- Creación de tabla MIUSUARIO_TEST.PERSONA, por medio de HIVE.

```

3 rows selected (0.07 seconds)
0: jdbc:hive2://> CREATE TABLE MIUSUARIO_TEST.PERSONA(
    . . . . . > ID STRING,
    . . . . . > NOMBRE STRING,
    . . . . . > TELEFONO STRING,
    . . . . . > CORREO STRING,
    . . . . . > FECHA_INGRESO STRING,
    . . . . . > EDAD INT,
    . . . . . > SALARIO DOUBLE,
    . . . . . > ID_EMPRESA STRING
    . . . . . > )
    . . . . . > ROW FORMAT DELIMITED
    . . . . . > FIELDS TERMINATED BY '|'
    . . . . . > LINES TERMINATED BY '\n'
    . . . . . > STORED AS TEXTFILE;
OK
No rows affected (0.845 seconds)
0: jdbc:hive2://> SHOW DATABASES;
OK
+-----+
| database_name |
+-----+
| bigdata      |
| default      |
| miusuario_test |
+-----+
3 rows selected (0.078 seconds)
0: jdbc:hive2://> SHOW TABLES IN MIUSUARIO_TEST;
OK
+-----+
| tab_name   |
+-----+
| persona   |
+-----+
1 row selected (0.084 seconds)
0: jdbc:hive2://> ■

```

Import MobaXterm by subscribing to the professional edition here: <https://mobaxterm.moba>

- Validando existencia de la base de datos en Linux.

```
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Tue Feb 18 00:38:31 2025 from 200.48.170.38
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/hive/warehouse
Found 2 items
drwxr-xr-x  - jenner hadoop      0 2025-02-18 02:22 /user/hive/warehouse/bigdata.db
drwxr-xr-x  - jenner hadoop      0 2025-02-18 02:29 /user/hive/warehouse/miusuario_test.db
jenner@dmc-dev-bdp-15-m:~$
```

- Viendo el contenido

```
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/hive/warehouse/miusuario_test.db
Found 1 items
drwxr-xr-x  - jenner hadoop      0 2025-02-18 02:29 /user/hive/warehouse/miusuario_test.db/persona
jenner@dmc-dev-bdp-15-m:~$
```

- Migrando persona.data

```
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Tue Feb 18 00:38:31 2025 from 200.48.170.38
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/hive/warehouse
Found 2 items
drwxr-xr-x  - jenner hadoop      0 2025-02-18 02:22 /user/hive/warehouse/bigdata.db
drwxr-xr-x  - jenner hadoop      0 2025-02-18 02:29 /user/hive/warehouse/miusuario_test.db
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/hive/warehouse/miusuario_test.db
Found 1 items
drwxr-xr-x  - jenner hadoop      0 2025-02-18 02:29 /user/hive/warehouse/miusuario_test.db/persona
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/hive/warehouse/miusuario_test.db/persona
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -put /home/jenner/persona.data /user/hive/warehouse/miusuario_test.db/persona
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/hive/warehouse/miusuario_test.db/persona
Found 1 items
-rw-r--r--  2 jenner hadoop    7282 2025-02-18 02:46 /user/hive/warehouse/miusuario_test.db/persona/persona.data
jenner@dmc-dev-bdp-15-m:~$
```

- Visualizando los tipos de datos de los campos de la tabla persona.

```
0: jdbc:hive2://> DESC MIUSUARIO_TEST.PERSONA;
OK
+-----+-----+-----+
| col_name | data_type | comment |
+-----+-----+-----+
| id       | string    |          |
| nombre   | string    |          |
| telefono | string    |          |
| correo   | string    |          |
| fecha_ingreso | string |          |
| edad     | int      |          |
| salario  | double   |          |
| id_empresa | string |          |
+-----+-----+-----+
8 rows selected (0.178 seconds)
0: jdbc:hive2://>
```

- Aplicando desc formatted para visualizar la composición de la tabla persona.

col_name	<th>comment</th>	comment
# col_name		
id	string	
nombre	string	
telefono	string	
correo	string	
fecha_ingreso	string	
edad	int	
salario	double	
id_empresa	string	
NULL	NULL	NULL
# Detailed Table Information		
Database:	miusuario_test	
OwnerType:	USER	
Owner:	jenner	
CreateTime:	Tue Feb 18 02:29:07 UTC 2025	
LastAccessTime:	UNKNOWN	
Retention:	0	
Location:	hdfs://dmc-dev-bdp-15-m/user/hive/warehouse/miusuario_test.db/persona NULL	
Table Type:	MANAGED_TABLE	
Table Parameters:	NULL	
COLUMN_STATS_ACCURATE	{ "BASIC_STATS": "true", "COLUMN_STATS": { "correo": "true", "edad": "true", "fecha_ingreso": "true", "id": "true", "id_empresa": "true", "nombre": "true", "salario": "true", "telefono": "true" } }	
bucketing_version	2	
numFiles	0	
numRows	0	
rawDataSize	0	
totalSize	0	
transient_lastDdlTime	1739845747	
NULL	NULL	NULL
# Storage Information		
SerDe Library:	org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	
InputFormat:	org.apache.hadoop.mapred.TextInputFormat	
OutputFormat:	org.apache.hadoop.hive ql.io.HiveIgnoreKeyTextOutputFormat NULL	
Compressed:	No	
Num Buckets:	-1	
Bucket Columns:	[]	
Sort Columns:	[]	
Storage Desc Params:	NULL	
field.delim		
line.delim		\n
serialization.format		

col_name	<th>comment</th>	comment
# col_name		
id	string	
nombre	string	
telefono	string	
correo	string	
fecha_ingreso	string	
edad	int	
salario	double	
id_empresa	string	
NULL	NULL	NULL
# Detailed Table Information		
Database:	miusuario_test	
OwnerType:	USER	
Owner:	jenner	
CreateTime:	Tue Feb 18 02:29:07 UTC 2025	
LastAccessTime:	UNKNOWN	
Retention:	0	
Location:	hdfs://dmc-dev-bdp-15-m/user/hive/warehouse/miusuario_test.db/persona NULL	
Table Type:	MANAGED_TABLE	
Table Parameters:	NULL	
COLUMN_STATS_ACCURATE	{ "BASIC_STATS": "true", "COLUMN_STATS": { "correo": "true", "edad": "true", "fecha_ingreso": "true", "id": "true", "id_empresa": "true", "nombre": "true", "salario": "true", "telefono": "true" } }	
bucketing_version	2	
numFiles	0	
numRows	0	
rawDataSize	0	
totalSize	0	
transient_lastDdlTime	1739845747	
NULL	NULL	NULL
# Storage Information		
SerDe Library:	org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	
InputFormat:	org.apache.hadoop.mapred.TextInputFormat	
OutputFormat:	org.apache.hadoop.hive ql.io.HiveIgnoreKeyTextOutputFormat NULL	
Compressed:	No	
Num Buckets:	-1	
Bucket Columns:	[]	
Sort Columns:	[]	
Storage Desc Params:	NULL	
field.delim		
line.delim		\n
serialization.format		

- Realizando consulta select * from, para validar la ingestión correcta de la información.

```
40 rows selected (0.232 seconds)
0: jdbc:hive2://> select * from miusuario_test.persona;
OK
+-----+-----+-----+-----+-----+
| persona.id | persona.nombre | persona.telefono | persona.correo | persona.fecha_ingreso | persona.edad | persona.s
|-----+-----+-----+-----+-----+
| ID | NOMBRE | TELEFONO | CORREO | FECHA_INGRESO | NULL | NULL
| 1 | Carl | 1-745-633-9145 | arcu.Sed.et@ante.co.uk | 2004-04-23 | 32 | 20095.0
| 2 | 5 | Priscilla | 155-2498 | Donec.egestas.Aliquam@volutpatnunc.edu | 2019-02-17 | 34 | 9298.0
| 3 | 2 | Jocelyn | 1-204-956-8594 | amet.diam@lobortis.co.uk | 2002-08-01 | 27 | 10853.0
| 4 | 3 | Aidan | 1-719-862-9385 | euismod.et.commodo@nibhacinaorci.edu | 2018-11-06 | 29 | 3387.0
| 5 | 10 | Leandra | 839-8044 | at@premiumetrurum.com | 2002-10-10 | 41 | 22102.0
| 6 | 1 | Bert | 797-4453 | a.felis.ullamcorper@arcu.org | 2017-04-25 | 70 | 7800.0
| 7 | 7 | Mark | 1-680-102-6792 | Quisque.ac@placerat.ca | 2006-04-21 | 52 | 8112.0
| 8 | 5 | Jonah | 214-2975 | eu.ultrices.sit@vitae.ca | 2017-10-07 | 23 | 17040.0
| 9 | 5 | Hanae | 935-2277 | eu@Nunc.ca | 2003-05-25 | 69 | 6834.0
| 10 | 3 | Cadman | 1-866-561-2701 | orci.adipiscing.non@semperNam.ca | 2001-05-19 | 19 | 7996.0
| 11 | 7 | Melyssa | 596-7736 | vel@vulputateposuerevlpitate.net | 2008-10-14 | 48 | 4913.0
| 12 | 8 | Tanner | 1-739-776-7897 | arcu.Aliquam.ultrices@sociis.com | 2011-05-10 | 24 | 19943.0
| 13 | 8 | Trevor | 512-1955 | Nunc.quis.arcu@egestasa.org | 2010-08-06 | 34 | 9501.0
| 14 | 5 | Allen | 733-2795 | felis.Donec@mecleo.org | 2005-03-07 | 59 | 16289.0
| 15 | 2 | Wanda | 359-6973 | Nam.nulla.magna@In.org | 2005-08-21 | 27 | 1539.0
| 5 |
```

```
101 rows selected (2.071 seconds)
0: jdbc:hive2://> select * from miusuario_test.persona where nombre='Carl';
Query ID = jenner_20250218030001_7efc164c-bd06-4d57-a470-637552736e69
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739823135750_0001)
OK
+-----+-----+-----+-----+-----+-----+-----+-----+
| persona.id | persona.nombre | persona.telefono | persona.correo | persona.fecha_ingreso | persona.edad | persona.salario | persona.id_empresa
|-----+-----+-----+-----+-----+-----+-----+-----+
| 1 | Carl | 1-745-633-9145 | arcu.Sed.et@ante.co.uk | 2004-04-23 | 32 | 20095.0 | 5
|-----+-----+-----+-----+-----+-----+-----+-----+
1 row selected (31.422 seconds)
0: jdbc:hive2://> ■
```

```
---+
1 row selected (31.422 seconds)
0: jdbc:hive2://> select * from miusuario_test.persona where nombre='Carl' or nombre='Wanda';
Query ID = jenner_20250218030114_2afc8a5f-3eb8-405e-aa06-1b713bceccbc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739823135750_0001)
OK
+-----+-----+-----+-----+-----+-----+-----+-----+
| persona.id | persona.nombre | persona.telefono | persona.correo | persona.fecha_ingreso | persona.edad | persona.salario | persona.id_empresa
|-----+-----+-----+-----+-----+-----+-----+-----+
| 1 | Carl | 1-745-633-9145 | arcu.Sed.et@ante.co.uk | 2004-04-23 | 32 | 20095.0 | 5
| 15 | Wanda | 359-6973 | Nam.nulla.magna@In.org | 2005-08-21 | 27 | 1539.0 | 5
| 42 | Wanda | 941-3970 | auctor.velit@sem.com | 2011-12-12 | 42 | 5419.0 | 9
|-----+-----+-----+-----+-----+-----+-----+-----+
3 rows selected (9.043 seconds)
0: jdbc:hive2://> ■
```

- Creación de otra tabla, apuntando a un location.

```

4 rows selected (0.097 seconds)
0: jdbc:hive2://> CREATE TABLE MIUSUARIO_TEST2.PERSONA(
. . . . . > ID STRING,
. . . . . > NOMBRE STRING,
. . . . . > TELEFONO STRING,
. . . . . > CORREO STRING,
. . . . . > FECHA_INGRESO STRING,
. . . . . > EDAD INT,
. . . . . > SALARIO DOUBLE,
. . . . . > ID_EMPRESA STRING
. . . . . > )
. . . . . > ROW FORMAT DELIMITED
. . . . . > FIELDS TERMINATED BY '|'
. . . . . > LINES TERMINATED BY '\n'
. . . . . > STORED AS TEXTFILE
. . . . . > LOCATION '/user/miusuario/bd/miusuario_test2/persona';
OK
No rows affected (0.102 seconds)
0: jdbc:hive2://> show tables in miusuario_test2;
OK
+-----+
| tab_name |
+-----+
| persona |
+-----+
1 row selected (0.062 seconds)
0: jdbc:hive2://> █

```

```

0: jdbc:hive2://> desc formatted miusuario_test2.persona;
OK
+-----+-----+-----+
| col_name | data_type | comment |
+-----+-----+-----+
| # col_name | data_type | comment |
| id | string | NULL |
| nombre | string | NULL |
| telefono | string | NULL |
| correo | string | NULL |
| fecha_ingreso | string | NULL |
| edad | int | NULL |
| salario | double | NULL |
| id_empresa | string | NULL |
| _ | NULL | NULL |
| # Detailed Table Information | | |
| Database: | miusuario_test2 | |
| OwnerType: | USER | |
| Owner: | janner | |
| CreateTime: | Tue Feb 18 03:21:17 UTC 2025 | |
| LastAccessTime: | UNKNOWN | |
| Retention: | 0 | |
| Location: | hdfs://dmc-dev-bdp-15-m/user/miusuario/bd/miusuario_test2/persona | |
| Table Type: | MANAGED_TABLE | |
| Table Parameters: | NULL | |
| COLUMN_STATS_ACCURATE | {"BASIC_STATS": "true", "COLUMN_STATS": [{"correo": "true", "edad": ":"}, {"fecha_ingreso": "true", "id": "true", "id_empresa": "true", "nombre": "true"}, {"salario": "true", "telefono": "true"}]} | |
| bucketing_version | 2 | |
| numFiles | 0 | |
| numRows | 0 | |
| rawDataSize | 0 | |
| totalSize | 0 | |
| transient_lastDdlTime | 1739848877 | |
| # Storage Information | | |
| SerDe Library: | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe | |
| InputFormat: | org.apache.hadoop.mapred.TextInputFormat | |
| OutputFormat: | org.apache.hadoop.hive ql.io.HiveIgnoreKeyTextOutputFormat | |
| Compressed: | No | |
| Num Buckets: | -1 | |

```

# col_name	data_type	comment
id	string	
nombre	string	
telefono	string	
correo	string	
fecha_ingreso	string	
edad	int	
salario	double	
id_empresa	string	
NULL	NULL	NULL
NULL	NULL	NULL
Database:	miusuario_test2	USER
OwnerType:		
Owner:	jenner	
CreateTime:	Tue Feb 18 03:21:17 UTC 2025	
LastAccessTime:	UNKNOWN	
Retention:	0	
Location:	hdfs://dmc-dev-bdp-15-m/user/miusuario/bd/miusuario_test2/persona	NULL
Table Type:	MANAGED_TABLE	
Table Parameters:	NULL	
"": "true", "fecha_ingreso": "true", "id": "true", "id_empresa": "true", "nombre": "true", "correo": "true", "edad": "true", "telefono": "true", "salario": "true"	{"BASIC_STATS": "true", "COLUMN_STATS": [{"correo": "true", "edad": "true", "fecha_ingreso": "true", "id": "true", "id_empresa": "true", "nombre": "true", "telefono": "true"}]}	
COLUMN_STATS_ACCURATE		
bucketing_version	2	
numFiles	0	
numRows	0	
rawDataSize	0	
totalSize	0	
transient_lastDdlTime	1739848877	
NULL	NULL	NULL
# Storage Information		
SerDe Library:	org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	
InputFormat:	org.apache.hadoop.mapred.TextInputFormat	
OutputFormat:	org.apache.hadoop.hive.io.HiveIgnoreKeyTextOutputFormat	NULL
Compressed:	No	
Num Buckets:	-1	
Bucket Columns:	[]	
Sort Columns:	[]	
Storage Desc Params:	NULL	
field.delim		
line.delim		
serialization.format	\n	

port MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

- Validando cantidad de registros.

40 rows selected (0.151 seconds)
0: jdbc:hive2://> SELECT * FROM MIUSUARIO_TEST2.PERSONA LIMIT 10;
OK
+-----+-----+-----+-----+-----+-----+-----+-----+
persona.id persona.nombre persona.telefono persona.correo persona.fecha_ingreso persona.edad persona.salario persona.id_empresa
+-----+-----+-----+-----+-----+-----+-----+-----+
No rows selected (0.206 seconds)
0: jdbc:hive2://> ■

port MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

- Carga con hive y limit 10.

No rows selected (0.206 seconds)
0: jdbc:hive2://> LOAD DATA LOCAL INPATH '/home/jenner/persona.data' INTO TABLE MIUSUARIO_TEST2.PERSONA;
OK
Loading data to table miusuario_test2.persona
OK
No rows affected (0.357 seconds)
0: jdbc:hive2://> SELECT * FROM MIUSUARIO_TEST2.PERSONA LIMIT 10;
OK
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
persona.id persona.nombre persona.telefono persona.correo persona.fecha_ingreso persona.edad persona.salario persona.id_empresa
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
ID_EMPRESA NOMBRE TELEFONO CORREO FECHA_INGRESO NULL NULL ID
+-----+-----+-----+-----+-----+-----+-----+-----+
1 Carl 1-745-633-9145 arcu.Sed.et@ante.co.uk 2004-04-23 32 20095.0 5
+-----+-----+-----+-----+-----+-----+-----+-----+
2 Priscilla 155-2498 Donec.egestas.Aliquam@voluptatnunc.edu 2019-02-17 34 9298.0 2
+-----+-----+-----+-----+-----+-----+-----+-----+
3 Jocelyn 1-204-956-8594 amet.diam@lobortis.co.uk 2002-08-01 27 10853.0 3
+-----+-----+-----+-----+-----+-----+-----+-----+
4 Aidan 1-719-862-9385 euismod.et.commodo@nibhlacliniaorci.edu 2018-11-06 29 3387.0 10
+-----+-----+-----+-----+-----+-----+-----+-----+
5 Leandra 839-8844 at@pretiumetrutrum.com 2002-10-10 41 22102.0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
6 Bert 797-4453 a.felis.ullamcorper@arcu.org 2017-04-25 70 7800.0 7
+-----+-----+-----+-----+-----+-----+-----+-----+
7 Mark 1-680-102-6792 Quisque.ac@placerat.ca 2006-04-21 52 8112.0 5
+-----+-----+-----+-----+-----+-----+-----+-----+
8 Jonah 214-2975 eu.ultrices.sit@vitae.ca 2017-10-07 23 17040.0 5
+-----+-----+-----+-----+-----+-----+-----+-----+
9 Hanae 935-2277 eu@Nunc.ca 2003-05-25 69 6834.0 3
+-----+-----+-----+-----+-----+-----+-----+-----+
10 rows selected (0.238 seconds)
0: jdbc:hive2://> ■

SESIÓN 05

- Validando existencia de mi tabla persona.

```
Beeline version 3.1.3 by Apache Hive
0: jdbc:hive2://> show databases;
OK
+-----+
| database_name |
+-----+
| bigdata      |
| default      |
| miusuario_test |
| miusuario_test2 |
+-----+
4 rows selected (1.262 seconds)
0: jdbc:hive2://> SHOW TABLES IN MIUSUARIO_TEST2;
OK
+-----+
| tab_name |
+-----+
| persona |
+-----+
1 row selected (0.1 seconds)
0: jdbc:hive2://> SELECT * FROM MIUSUARIO_TEST2.PERSONA LIMIT 5;
OK
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| persona.id | persona.nombre | persona.telefono | persona.correo | persona.fecha_ingreso | persona.edad | persona.salario | persona.id_empresa |
+-----+-----+-----+-----+-----+-----+-----+-----+
| 1           | Carl          | 1-745-633-9145 | arcu.Sed.et@ante.co.uk | 2004-04-23        | 32          | 20095.0       | 5               |
| 2           | Priscilla     | 155-2498       | Donec.egestas.Aliquam@voluptatnunc.edu | 2019-02-17        | 34          | 9298.0        | 2               |
| 3           | Jocelyn       | 1-204-956-8594 | amet.diam@lobortis.co.uk    | 2002-08-01        | 27          | 10853.0       | 3               |
| 4           | Aidan         | 1-719-862-9385 | euismod.et.commodo@nibhaciniaorci.edu | 2018-11-06        | 29          | 3387.0        | 10              |
+-----+-----+-----+-----+-----+-----+-----+-----+
5 rows selected (2.596 seconds)
0: jdbc:hive2://>
```

- Eliminación mediante drop table de la tabla persona.

```
5 rows selected (2.506 seconds)
0: jdbc:hive2://> DROP TABLE MIUSUARIO_TEST2.PERSONA;
OK
No rows affected (2.627 seconds)
0: jdbc:hive2://> SELECT * FROM MIUSUARIO_TEST2.PERSONA LIMIT 5;
25/02/24 21:09:28 [dca4a556-7fb3-4777-96c4-e727c59b8b52 main]: ERROR parse.CalcitePlanner: org.apache.hadoop.hive.ql.parse.SemanticException: Line 1:14 Table not found 'PERSONA'
at org.apache.hadoop.hive.ql.parse.SemanticAnalyzer.getMetaData(SemanticAnalyzer.java:2156)
at org.apache.hadoop.hive.ql.parse.SemanticAnalyzer.getMetaData(SemanticAnalyzer.java:2076)
at org.apache.hadoop.hive.ql.parse.SemanticAnalyzer.genResolvedParseTree(SemanticAnalyzer.java:12048)
at org.apache.hadoop.hive.ql.parse.SemanticAnalyzer.analyzeInternal(SemanticAnalyzer.java:12144)
at org.apache.hadoop.hive.ql.parse.CalcitePlanner.analyzeInternal(CalcitePlanner.java:330)
at org.apache.hadoop.hive.ql.parse.BaseSemanticAnalyzer.analyze(BaseSemanticAnalyzer.java:285)
at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:659)
at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1826)
at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1773)
at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1768)
```

- Se ha borrado la carpeta y el archivo dentro de ella.

```
-rw-r--r-- 2 jenner hadoop 1202 2023-02-10 05:21 /user/miusuario/bd/miusuario_test2
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/miusuario/bd/miusuario_test2/persona
ls: `/user/miusuario/bd/miusuario_test2/persona': No such file or directory
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/miusuario/bd/miusuario_test2/
jenner@dmc-dev-bdp-15-m:~$
```

- Creando tabla external .

```

0: jdbc:hive2://> CREATE EXTERNAL TABLE MIUSUARIO_TEST2.PERSONA(
. . . . . > ID STRING,
. . . . . > NOMBRE STRING,
. . . . . > TELEFONO STRING,
. . . . . > CORREO STRING,
. . . . . > FECHA_INGRESO STRING,
. . . . . > EDAD INT,
. . . . . > SALARIO DOUBLE,
. . . . . > ID_EMPRESA STRING
. . . . . > )
. . . . . > ROW FORMAT DELIMITED
. . . . . > FIELDS TERMINATED BY '|'
. . . . . > LINES TERMINATED BY '\n'
. . . . . > STORED AS TEXTFILE
. . . . . > LOCATION '/user/miusuario/bd/miusuario_test2/persona';
OK
No rows affected (0.282 seconds)
0: jdbc:hive2://> LOAD DATA LOCAL INPATH '/home/jenner/persona.data' INTO TABLE
. . . . . > MIUSUARIO_TEST2.PERSONA;
Loading data to table miusuario_test2.persona
OK
No rows affected (1.598 seconds)
0: jdbc:hive2://> █

```

- Comprobando la data.

```

0: jdbc:hive2://> SELECT * FROM MIUSUARIO_TEST2.PERSONA LIMIT 10;
OK
+-----+-----+-----+-----+-----+-----+-----+-----+
| persona.id | persona.nombre | persona.telefono | persona.correo | persona.fecha_ingreso | persona.edad | persona.salario | persona.id_empresa |
+-----+-----+-----+-----+-----+-----+-----+-----+
| ID_EMPRESA | NOMBRE | TELEFONO | CORREO | FECHA_INGRESO | EDAD | SALARIO | ID |
+-----+-----+-----+-----+-----+-----+-----+-----+
| 1 | Carl | 1-745-633-9145 | arcu.Sed.et@ante.co.uk | 2004-04-23 | 32 | 20095.0 | 5 |
| 2 | Priscilla | 155-2498 | Donec.egestas.Aliquam@volutpatnunc.edu | 2019-02-17 | 34 | 9298.0 | 2 |
| 3 | Jocelyn | 1-204-956-8594 | amet.diam@lobortis.co.uk | 2002-08-01 | 27 | 10853.0 | 3 |
| 4 | Aidan | 1-719-862-9385 | euismod.et.commodo@nibhlacliniaorci.edu | 2018-11-06 | 29 | 3387.0 | 10 |
| 5 | Leandra | 839-8044 | at@pretiumetrurrum.com | 2002-10-10 | 41 | 22102.0 | 1 |
| 6 | Bert | 797-4453 | a.felis.ultramcorper@arcu.org | 2017-04-25 | 70 | 7800.0 | 7 |
| 7 | Mark | 1-680-102-6792 | Quisque.ac@placerat.ca | 2006-04-21 | 52 | 8112.0 | 5 |
| 8 | Jonah | 214-2975 | eu.ultrices.sit@vitae.ca | 2017-10-07 | 23 | 17040.0 | 5 |
| 9 | Hanae | 935-2277 | eu@Nunc.ca | 2003-05-25 | 69 | 6834.0 | 3 |
+-----+-----+-----+-----+-----+-----+-----+-----+
10 rows selected (0.444 seconds)
0: jdbc:hive2://> █

```

```

jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/miusuario/bd/miusuario_test2/
Found 1 items
drwxr-xr-x - jenner hadoop 0 2025-02-24 21:20 /user/miusuario/bd/miusuario_test2/persona
jenner@dmc-dev-bdp-15-m:~$ █

```

- Se elimina la tabla en hive, pero en Linux aún existe.

```
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/miusuario/bd/miusuario_test2/
Found 1 items
drwxr-xr-x - jenner hadoop 0 2025-02-24 21:20 /user/miusuario/bd/miusuario_test2/persona
jenner@dmc-dev-bdp-15-m:~$
```

- Sin validador IF NOT EXISTS.

```
0: jdbc:hive2://> CREATE DATABASE DMC;
OK
No rows affected (0.135 seconds)
0: jdbc:hive2://> CREATE DATABASE DMC;
25/02/24 21:44:18 [HiveServer2-Background-Pool: Thread-81]: ERROR exec.DDLTask: Failed
org.apache.hadoop.hive.ql.metadata.HiveException: Database DMC already exists
at org.apache.hadoop.hive.ql.exec.DDLTask.createDatabase(DDLTask.java:4835) ~[hive-exec-3.1.3.jar:3.1.3]
at org.apache.hadoop.hive.ql.exec.DDLTask.execute(DDLTask.java:393) ~[hive-exec-3.1.3.jar:3.1.3]
at org.apache.hadoop.hive.ql.exec.Task.executeTask(Task.java:205) ~[hive-exec-3.1.3.jar:3.1.3]
at org.apache.hadoop.hive.ql.exec.TaskRunner.runSequential(TaskRunner.java:100) ~[hive-exec-3.1.3.jar:3.1.3]
at org.apache.hadoop.hive.ql.Driver.launchTask(Driver.java:2664) ~[hive-exec-3.1.3.jar:3.1.3]
at org.apache.hadoop.hive.ql.Driver.execute(Driver.java:2335) ~[hive-exec-3.1.3.jar:3.1.3]
at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:2011) ~[hive-exec-3.1.3.jar:3.1.3]
at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1709) ~[hive-exec-3.1.3.jar:3.1.3]
at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1703) ~[hive-exec-3.1.3.jar:3.1.3]
at org.apache.hadoop.hive.ql.reexec.ReExecDriver.run(ReExecDriver.java:157) ~[hive-exec-3.1.3.jar:3.1.3]
at org.apache.hive.service.cli.operation.SQLOperation.runQuery(SQLOperation.java:227) ~[hive-service-3.1.3.jar:3.1.3]
```

- Con validador IF NOT EXISTS.

```
e=42000,code=1)
0: jdbc:hive2://> CREATE DATABASE IF NOT EXISTS DMC;
OK
No rows affected (0.044 seconds)
0: jdbc:hive2://>
```

SESIÓN 06

- Cargando el archivo Poblano_Capa_Workload.sql

```
No rows affected (0.116 seconds)
0: jdbc:hive2://>
0: jdbc:hive2://>
0: jdbc:hive2://>
0: jdbc:hive2://> -- Subida de datos
0: jdbc:hive2://>
0: jdbc:hive2://> LOAD DATA LOCAL INPATH '/home/${hiveconf:PARAM_USERNAME}/dataset/transacciones.data'
0: jdbc:hive2://> . . . . . >
0: jdbc:hive2://> . . . . . > INTO TABLE ${hiveconf:ENV}_workload.TRANSACCION;
Loading data to table dev_workload.transaccion
OK
No rows affected (0.399 seconds)
0: jdbc:hive2://>
0: jdbc:hive2://>
0: jdbc:hive2://>
0: jdbc:hive2://> -- Impresión de datos
0: jdbc:hive2://>
0: jdbc:hive2://> SELECT * FROM ${hiveconf:ENV}_workload.TRANSACCION LIMIT 10;
OK
+-----+-----+-----+-----+
| transaccion.id_persona | transaccion.id_empresa | transaccion.monto | transaccion.fecha |
+-----+-----+-----+-----+
| 18 | 3 | 1383 | 2018-01-21 |
| 30 | 6 | 2331 | 2018-01-21 |
| 47 | 2 | 2280 | 2018-01-21 |
| 28 | 1 | 730 | 2018-01-21 |
| 91 | 4 | 3081 | 2018-01-21 |
| 74 | 8 | 2409 | 2018-01-21 |
| 41 | 2 | 3754 | 2018-01-22 |
| 42 | 9 | 4079 | 2018-01-22 |
| 24 | 6 | 4475 | 2018-01-22 |
| 67 | 9 | 561 | 2018-01-22 |
+-----+-----+-----+-----+
10 rows selected (0.396 seconds)
0: jdbc:hive2://>
0: jdbc:hive2://>
0: jdbc:hive2://> Closing: 0: jdbc:hive2://
25/02/26 22:21:38 [shutdown-hook-0]: WARN util.ShutdownHookManager: Shutdown in progress, cannot cancel a deleteOnExit
25/02/26 22:21:38 [shutdown-hook-0]: WARN util.ShutdownHookManager: Shutdown in progress, cannot cancel a deleteOnExit
...[redacted]
```

- Haciendo un show databases;

```
Beeline version 3.1.3 by Apache Hive
0: jdbc:hive2://> show databases;
OK
+-----+
| database_name |
+-----+
| bigdata |
| default |
| dev_workload |
| dmc |
| miusuario_test |
| miusuario_test2 |
+-----+
6 rows selected (1.046 seconds)
0: jdbc:hive2://>
```

```

0: jdbc:hive2://> show tables in dev_workload
OK
+-----+
| tab_name |
+-----+
| empresa   |
| persona   |
| transaccion |
+-----+
3 rows selected (0.102 seconds)
0: jdbc:hive2://>

```

- Subida de los 3 archivos con formato .avsc.

```

jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/jenner/datalake/schema/dev_LANDING/
Found 3 items
-rw-r--r--  2 jenner hadoop      160 2025-02-26 23:19 /user/jenner/datalake/schema/dev_LANDING/empresa.avsc
-rw-r--r--  2 jenner hadoop      498 2025-02-26 23:19 /user/jenner/datalake/schema/dev_LANDING/persona.avsc
-rw-r--r--  2 jenner hadoop     226 2025-02-26 23:19 /user/jenner/datalake/schema/dev_LANDING/transaccion.avsc
jenner@dmc-dev-bdp-15-m:~$ 

```

- Resultado de ejecución de capa Landing.

```

+-----+-----+-----+-----+
| transaccion.id_persona | transaccion.id_empresa | transaccion.monto | transaccion.fecha |
+-----+-----+-----+-----+
| 18                  | 3                   | 1383            | 2018-01-21       |
| 30                  | 6                   | 2331            | 2018-01-21       |
| 47                  | 2                   | 2280            | 2018-01-21       |
| 28                  | 1                   | 730             | 2018-01-21       |
| 91                  | 4                   | 3081            | 2018-01-21       |
| 74                  | 8                   | 2409            | 2018-01-21       |
| 83                  | 5                   | 2079            | 2018-01-21       |
| 48                  | 4                   | 2543            | 2018-01-21       |
| 15                  | 6                   | 1434            | 2018-01-21       |
| 89                  | 4                   | 780             | 2018-01-21       |
+-----+-----+-----+-----+
10 rows selected (0.31 seconds)
0: jdbc:hive2://>
0: jdbc:hive2://>
0: jdbc:hive2://>
0: jdbc:hive2://>
0: jdbc:hive2://> -- Verificamos las particiones
0: jdbc:hive2://>
0: jdbc:hive2://> SHOW PARTITIONS ${hiveconf:ENV}_LANDING.TRANSACCION;
OK
+-----+
| partition |
+-----+
| fecha=2018-01-21 |
| fecha=2018-01-22 |
| fecha=2018-01-23 |
+-----+
3 rows selected (0.169 seconds)
0: jdbc:hive2://>
0: jdbc:hive2://>

```

- Creación de capa CURATED.

```
0: jdbc:hive2://> SELECT * FROM ${hiveconf:ENV}_curated.PERSONA LIMIT 10;
OK
+-----+-----+-----+-----+-----+-----+-----+-----+
| persona.id | persona.nombre | persona.telefono | persona.correo | persona.fecha_ingreso | persona.edad | persona.salario | persona.id_empresa |
+-----+-----+-----+-----+-----+-----+-----+-----+
| 1 | Carl | 1-745-633-9145 | arcu.Sed.et@ante.co.uk | 2004-04-23 | 32 | 20095.0 | 5 |
| 2 | Priscilla | 155-2498 | Donec.egestas.Aliquam@voluptpatnunc.edu | 2019-02-17 | 34 | 9298.0 | 2 |
| 3 | Jocelyn | 1-204-956-8594 | amet.diam@lobortis.co.uk | 2002-08-01 | 27 | 10853.0 | 3 |
| 4 | Aidan | 1-719-862-9385 | euismod.et.commodo@nibhlaclinaorci.edu | 2018-11-06 | 29 | 3387.0 | 16 |
| 5 | Leandra | 839-8044 | at@pretiumpetrum.com | 2002-10-10 | 41 | 22102.0 | 1 |
| 6 | Bert | 797-4453 | a.felis.ullamcorper@arcu.org | 2017-04-25 | 70 | 7800.0 | 7 |
| 7 | Mark | 1-680-102-6792 | Quisque.ac@placerat.ca | 2006-04-21 | 52 | 8112.0 | 5 |
| 8 | Jonah | 214-2975 | eu.ultrices.sit@vitae.ca | 2017-10-07 | 23 | 17040.0 | 5 |
| 9 | Hanae | 935-2277 | eu@Nunc.ca | 2003-05-25 | 69 | 6834.0 | 3 |
| 10 | Cadman | 1-866-561-2701 | orci.adipiscing.non@semperNam.ca | 2001-05-19 | 19 | 7996.0 | 7 |
+-----+-----+-----+-----+-----+-----+-----+-----+
```

```
0: jdbc:hive2://>
0: jdbc:hive2://> SELECT * FROM ${hiveconf:ENV}_curated.EMPRESA LIMIT 10;
OK
+-----+-----+
| empresa.id | empresa.nombre |
+-----+-----+
| 1 | Walmart |
| 2 | Microsoft |
| 3 | Apple |
| 4 | Toyota |
| 5 | Amazon |
| 6 | Google |
| 7 | Samsung |
| 8 | HP |
| 9 | IBM |
| 10 | Sony |
+-----+-----+
10 rows selected (0.36 seconds)
0: jdbc:hive2://>
```

```
0: jdbc:hive2://>
0: jdbc:hive2://> SELECT * FROM ${hiveconf:ENV}_curated.TRANSACCION LIMIT 10;
OK
+-----+-----+-----+-----+
| transaccion.id_persona | transaccion.id_empresa | transaccion.monto | transaccion.fecha |
+-----+-----+-----+-----+
| 18 | 3 | 1383.0 | 2018-01-21 |
| 30 | 6 | 2331.0 | 2018-01-21 |
| 47 | 2 | 2280.0 | 2018-01-21 |
| 28 | 1 | 730.0 | 2018-01-21 |
| 91 | 4 | 3081.0 | 2018-01-21 |
| 74 | 8 | 2409.0 | 2018-01-21 |
| 83 | 5 | 2079.0 | 2018-01-21 |
| 48 | 4 | 2543.0 | 2018-01-21 |
| 15 | 6 | 1434.0 | 2018-01-21 |
| 89 | 4 | 780.0 | 2018-01-21 |
+-----+-----+-----+-----+
10 rows selected (0.327 seconds)
0: jdbc:hive2://>
0: jdbc:hive2://>
0: jdbc:hive2://>
0: jdbc:hive2://> -- Verificamos las particiones
0: jdbc:hive2://>
0: jdbc:hive2://> SHOW PARTITIONS ${hiveconf:ENV}_curated.TRANSACCION;
OK
```

- Ejecucion de capa Funcional.

transaccion_enriquecida.id_persona	transaccion_enriquecida.nombre_persona	transaccion_enriquecida.edad_persona	transaccion_enriquecida.salario_persona	transaccion_enriquecida.trabajo_persona	transaccion_enriquecida.monto_transaccion	transaccion_enriquecida.empresa_transaccion	transaccion_enriquecida.fecha_transaccion
18	Samsung	Owen	1383.0	34	Apple	4759.0	2018-01-21
30	HP	Clayton	2331.0	52	Google	9505.0	2018-01-21
47	Sony	Vernon	2280.0	35	Microsoft	7109.0	2018-01-21
28	HP	Stephen	730.0	53	Walmart	9469.0	2018-01-21
91	Google	Erica	3081.0	32	Toyota	8934.0	2018-01-21
74	Sony	Kaitlin	2409.0	56	HP	6515.0	2018-01-21
83	Microsoft	Giselle	2079.0	45	Amazon	2503.0	2018-01-21
48	HP	Illiana	2543.0	18	Toyota	1454.0	2018-01-21
15	Amazon	Wanda	1434.0	27	Google	1539.0	2018-01-21
89	Google	Kelly	780.0	55	Toyota	10110.0	2018-01-21

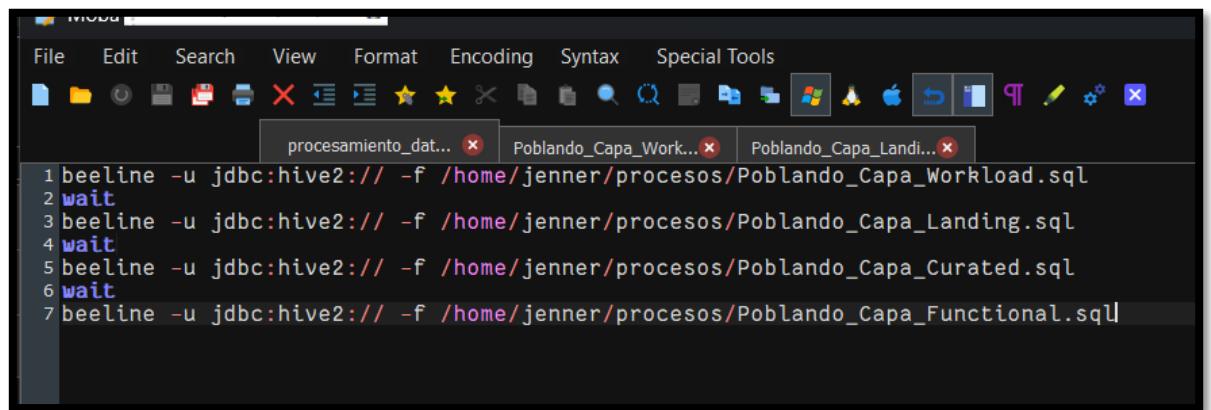
tmp_transaccion_enriquecida_3.id_persona	tmp_transaccion_enriquecida_3.nombre_persona	tmp_transaccion_enriquecida_3.edad_persona	tmp_transaccion_enriquecida_3.salario_persona	tmp_transaccion_enriquecida_3.trabajo_persona	tmp_transaccion_enriquecida_3.monto_transaccion	tmp_transaccion_enriquecida_3.empresa_transaccion	tmp_transaccion_enriquecida_3.fecha_transaccion
41	Toyota	Wynne	3754.0	31	19522.0		2018-01-22
42	Microsoft	Wanda	4079.0	42	5419.0		2018-01-22
24	IBM	Amaya	4475.0	24	1801.0		2018-01-22
67	HP	Buffy	561.0	38	15116.0		2018-01-22
9	Google	Hanae	3765.0	69	6834.0		2018-01-22
97	Apple	Flavia	3669.0	27	13473.0		2018-01-22
91	Toyota	Erica	3497.0	32	8934.0		2018-01-22
22	Apple	Kibo	2058.0	22	7449.0		2018-01-22
10	Amazon	Cadman	2027.0	19	7996.0		2018-01-22
48	Microsoft	Illiana	3833.0	18	1454.0		2018-01-22
	Samsung						
	Apple						
	HP						
	Walmart						

tmp_transaccion_enriquecida_2.id_persona tmp_transaccion_enriquecida_2.nombre_persona tmp_transaccion_enriquecida_2.edad_persona tmp_transaccion_enriquecida_2.salario_persona tmp_transaccion_enriquecida_2.trabajo_persona tmp_transaccion_enriquecida_2.monto_transaccion tmp_transaccion_enriquecida_2.fecha_transaccion tmp_transaccion_enriquecida_2.id_empresa_transaccion						
+-----+-----+-----+-----+-----+-----+-----+						
41 2 Toyota Wynne 31 19522.0 2018-01-22						
42 9 IBM Wanda 42 5419.0 2018-01-22						
24 6 HP Amaya 24 1801.0 2018-01-22						
67 9 Microsoft Buffy 38 15116.0 2018-01-22						
9 4 Apple Hanae 69 6834.0 2018-01-22						
97 3 Apple Flavia 27 13473.0 2018-01-22						
91 5 Google Erica 32 8934.0 2018-01-22						
22 4 Microsoft Kibo 22 7449.0 2018-01-22						
10 3 Samsung Cadman 19 7996.0 2018-01-22						
48 1 HP Illiana 18 1454.0 2018-01-22						
+-----+-----+-----+-----+-----+-----+-----+						

tmp_transaccion_enriquecida_1.id_persona tmp_transaccion_enriquecida_1.nombre_persona tmp_transaccion_enriquecida_1.edad_persona tmp_transaccion_enriquecida_1.salario_persona tmp_transaccion_enriquecida_1.id_empresa_persona tmp_transaccion_enriquecida_1.monto_transaccion tmp_transaccion_enriquecida_1.fecha_transaccion tmp_transaccion_enriquecida_1.id_empresa_transaccion						
+-----+-----+-----+-----+-----+-----+-----+						
41 4 Wynne 31 19522.0 2018-01-22						
42 9 Wanda 42 5419.0 2018-01-22						
24 8 Amaya 24 1801.0 2018-01-22						
67 6 Buffy 38 15116.0 2018-01-22						
9 3 Hanae 69 6834.0 2018-01-22						
97 3 Flavia 27 13473.0 2018-01-22						
91 6 Erica 32 8934.0 2018-01-22						
22 2 Kibo 22 7449.0 2018-01-22						
10 7 Cadman 19 7996.0 2018-01-22						
48 3 Illiana 18 1454.0 2018-01-22						
1 8						
+-----+-----+-----+-----+-----+-----+-----+						

🕒 dmc-dev-bdp-15-m 🔋 4% 🖱 5.49 GB / 7.76 GB ⚡ 0.07 Mb/s 🔩 0.04 Mb/s 🕒 141 min 📺 jenner (x2) 📲 /: 33% 🏁 /boot/efi: 10%

- Archivo procesamiento_datalake.sh



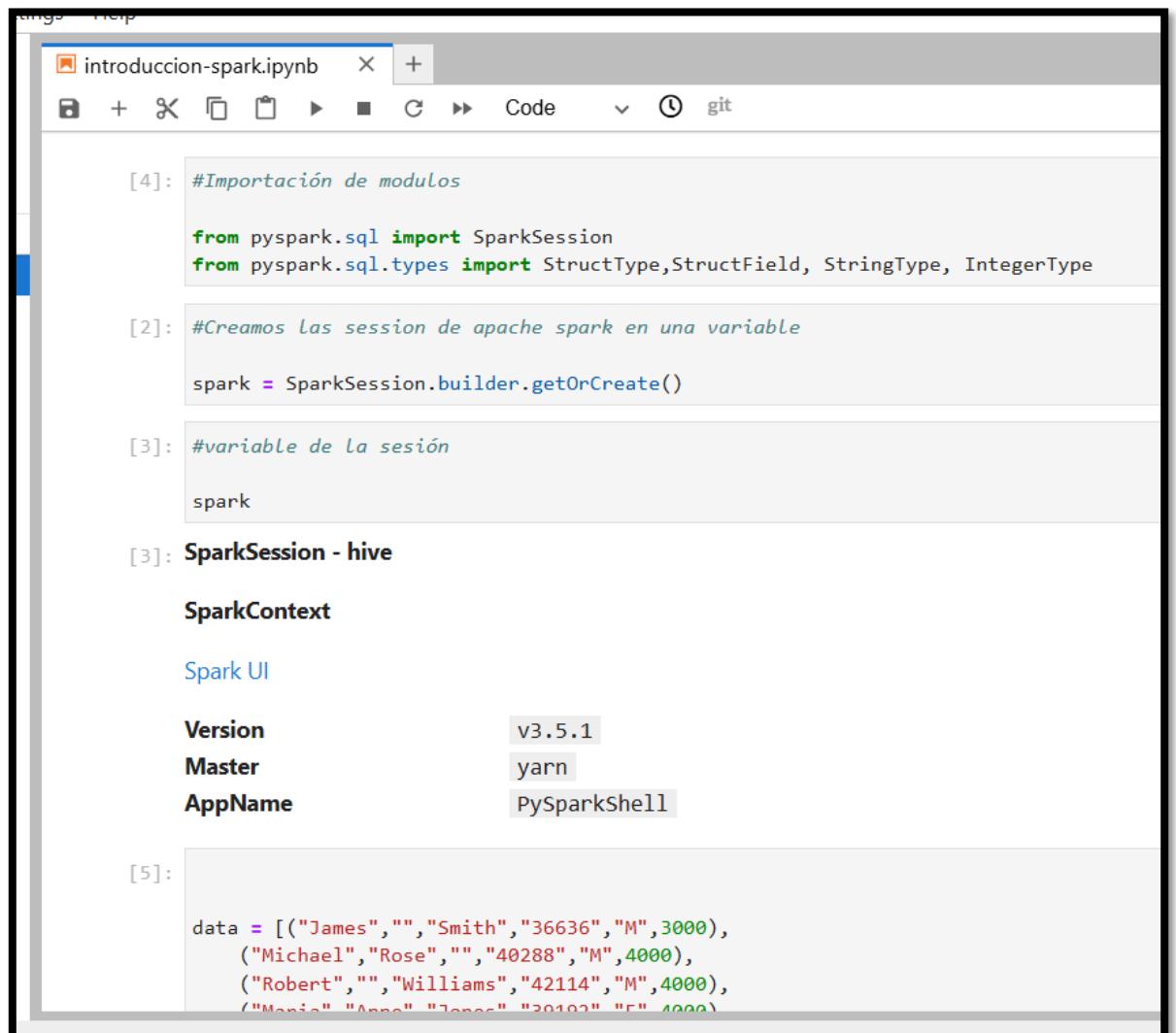
```

1 beeline -u jdbc:hive2:// -f /home/jenner/procesos/Poblano_Capa_Workload.sql
2 wait
3 beeline -u jdbc:hive2:// -f /home/jenner/procesos/Poblano_Capa_Landing.sql
4 wait
5 beeline -u jdbc:hive2:// -f /home/jenner/procesos/Poblano_Capa_Curated.sql
6 wait
7 beeline -u jdbc:hive2:// -f /home/jenner/procesos/Poblano_Capa_Functional.sql

```

SESIÓN 07

- Código pyspark en jupyterlab.



```

[4]: #Importación de modulos
from pyspark.sql import SparkSession
from pyspark.sql.types import StructType,StructField, StringType, IntegerType

[2]: #Creamos la session de apache spark en una variable
spark = SparkSession.builder.getOrCreate()

[3]: #variable de la sesión
spark

[3]: SparkSession - hive
      SparkContext
      Spark UI
      Version v3.5.1
      Master yarn
      AppName PySparkShell

[5]:
data = [("James","","Smith","36636","M",3000),
       ("Michael","Rose","","40288","M",4000),
       ("Robert","","Williams","42114","M",4000),
       ("Maria","Anne","Jones","39192","F",4000)
      ]

```

```

data = [("James","","Smith","36636","M",3000),
        ("Michael","Rose","","40288","M",4000),
        ("Robert","","Williams","42114","M",4000),
        ("Maria","Anne","Jones","39192","F",4000),
        ("Jen","Mary","Brown","","F",-1)
    ]

schema = StructType([ \
    StructField("firstname",StringType(),True), \
    StructField("middlename",StringType(),True), \
    StructField("lastname",StringType(),True), \
    StructField("id", StringType(), True), \
    StructField("gender", StringType(), True), \
    StructField("salary", IntegerType(), True) \
])

df = spark.createDataFrame(data=data,schema=schema)
df.printSchema()
df.show(truncate=False)

root
 |-- firstname: string (nullable = true)
 |-- middlename: string (nullable = true)
 |-- lastname: string (nullable = true)
 |-- id: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- salary: integer (nullable = true)

```

- Subida de archivo persona.data en cloud storage.

The screenshot shows the 'Detalles del bucket' (Bucket Details) page in Google Cloud Storage. The bucket is named 'dmc_datalake_dde_11_jacondex'. The 'OBJETOS' (Objects) tab is selected, displaying a list of files. One file, 'persona.data', is visible in the 'archivos/' folder. The file has a size of 7.1 KB and a type of application/octet-stream. There are buttons for 'CREAR CARPETA' (Create Folder), 'SUBIR' (Upload), and 'TRANSFERIR LOS DATOS' (Transfer Data).

- Creación de estructura de la tabla e insertando data de cloud storage

```

ruta = 'gs://dmc_datalake_dde_11_jacondex/archivos/persona.data'

df_columns = StructType([
    StructField("ID", StringType(), True),
    StructField("NOMBRE", StringType(), True),
    StructField("TELEFONO", StringType(), True),
    StructField("CORREO", StringType(), True),
    StructField("FECHA_INGRESO", StringType(), True),
    StructField("EDAD", IntegerType(), True),
    StructField("SALARIO", DoubleType(), True),
    StructField("ID_EMPRESA", StringType(), True),
])

df_with_schema = spark.read.format("CSV").option("header", "true").option("delimiter", "|").schema(df_columns).load(ruta)

df_with_schema.show(10)

```

ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO	EDAD	SALARIO	ID_EMPRESA
1	Carl	1-745-633-9145	arcu.Sed.et@ante....	2004-04-23	32	20095.0	5
2	Priscilla	155-2498	Donec.egestas.Ali...	2019-02-17	34	9298.0	2
3	Jocelyn	1-204-956-8594	amet.diam@loborti...	2002-08-01	27	10853.0	3
4	Aidan	1-719-862-9385	euismod.et.commod...	2018-11-06	29	3387.0	10
5	Leandra	839-8044	at@premiumetrutru...	2002-10-10	41	22102.0	1
6	Bert	797-4453	a.felis.ullamcorp...	2017-04-25	70	7800.0	7
7	Mark	1-680-102-6792	Quisque.ac@placer...	2006-04-21	52	8112.0	5
8	Jonah	214-2975	eu.ultrices.sit@v...	2017-10-07	23	17040.0	5
9	Hanae	935-2277	eu@Nunc.ca	2003-05-25	69	6834.0	3
10	Cadman	1-866-561-2701	orci.adipiscing.n...	2001-05-19	19	7996.0	7

- Imprimiendo la estructura del schema

```
[14]: df_with_schema.printSchema()

root
|-- ID: string (nullable = true)
|-- NOMBRE: string (nullable = true)
|-- TELEFONO: string (nullable = true)
|-- CORREO: string (nullable = true)
|-- FECHA_INGRESO: string (nullable = true)
|-- EDAD: integer (nullable = true)
|-- SALARIO: double (nullable = true)
|-- ID_EMPRESA: string (nullable = true)
```

SESIÓN 08

- Ejercicios de Select

```
[15]: df.printSchema()

root
|-- firstname: string (nullable = true)
|-- middlename: string (nullable = true)
|-- lastname: string (nullable = true)
|-- id: string (nullable = true)
|-- gender: string (nullable = true)
|-- salary: integer (nullable = true)

[18]: df_with_schema.select("NOMBRE","FECHA_INGRESO").show(4)

+-----+-----+
|   NOMBRE|FECHA_INGRESO|
+-----+-----+
|    Carl|  2004-04-23|
|Priscilla|  2019-02-17|
|  Jocelyn|  2002-08-01|
|    Aidan|  2018-11-06|
+-----+-----+
only showing top 4 rows

[20]: df_with_schema.select(df_with_schema.NOMBRE,df_with_schema.FECHA_INGRESO).show(4)

+-----+-----+
|   NOMBRE|FECHA_INGRESO|
+-----+-----+
|    Carl|  2004-04-23|
|Priscilla|  2019-02-17|
|  Jocelyn|  2002-08-01|
|    Aidan|  2018-11-06|
+-----+-----+
only showing top 4 rows
```

- Función COL.

```
[23]: df_with_schema.select(col("NOMBRE"),col("FECHA_INGRESO")).show(4)

+-----+-----+
| NOMBRE|FECHA_INGRESO|
+-----+-----+
| Carl| 2004-04-23|
| Priscilla| 2019-02-17|
| Jocelyn| 2002-08-01|
| Aidan| 2018-11-06|
+-----+-----+
only showing top 4 rows

[ ]:
```

- Guardando en un nuevo DF.

```
[24]: df_nuevo = df_with_schema.select(col("NOMBRE"),col("FECHA_INGRESO"))

[25]: df_nuevo.show(2)

+-----+-----+
| NOMBRE|FECHA_INGRESO|
+-----+-----+
| Carl| 2004-04-23|
| Priscilla| 2019-02-17|
+-----+-----+
only showing top 2 rows

[ ]:
```

- Función WITH COLUMN.

```
: df_with_schema.withColumn("SALARIO",col("SALARIO").cast("Integer")).show(10)

+-----+-----+-----+-----+-----+-----+-----+-----+
| ID| NOMBRE| TELEFONO| CORREO|FECHA_INGRESO|EDAD|SALARIO|ID_EMPRESA|
+-----+-----+-----+-----+-----+-----+-----+-----+
| 1| Carl|1-745-633-9145|arcu.Sed.et@ante....| 2004-04-23| 32| 20095| 5|
| 2| Priscilla| 155-2498|Donec.egestas.Ali...| 2019-02-17| 34| 9298| 2|
| 3| Jocelyn|1-204-956-8594|amet.diam@loborti...| 2002-08-01| 27| 10853| 3|
| 4| Aidan|1-719-862-9385|euismod.et.commod...| 2018-11-06| 29| 3387| 10|
| 5| Leandra| 839-8044|at@pretiumetrutru...| 2002-10-10| 41| 22102| 1|
| 6| Bert| 797-4453|a.felis.ullamcorp...| 2017-04-25| 70| 7800| 7|
| 7| Mark|1-680-102-6792|Quisque.ac@placer...| 2006-04-21| 52| 8112| 5|
| 8| Jonah| 214-2975|eu.ultrices.sit@v...| 2017-10-07| 23| 17040| 5|
| 9| Hanae| 935-2277| eu@Nunc.ca| 2003-05-25| 69| 6834| 3|
| 10| Cadman|1-866-561-2701|orci.adipiscing.n...| 2001-05-19| 19| 7996| 7|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 10 rows
```

- Operaciones con WITH COLUMN.

```
df_with_schema.withColumn("SALARIO",col("SALARIO")*100).show(10)

+---+-----+-----+-----+-----+-----+
| ID| NOMBRE| TELEFONO| CORREO|FECHA_INGRESO|EDAD| SALARIO|ID_EMPRESA|
+---+-----+-----+-----+-----+-----+
| 1| Carl|1-745-633-9145|arcu.Sed.et@ante....| 2004-04-23| 32| 2009500.0| 5|
| 2| Priscilla| 155-2498|Donec.egestas.Ali...| 2019-02-17| 34| 929800.0| 2|
| 3| Jocelyn|1-204-956-8594|amet.diam@loborti...| 2002-08-01| 27| 1085300.0| 3|
| 4| Aidan|1-719-862-9385|euismod.et.commod...| 2018-11-06| 29| 338700.0| 10|
| 5| Leandra| 839-8044|at@pretiunmetrtru...| 2002-10-10| 41| 2210200.0| 1|
| 6| Bert| 797-4453|a.felis.ullamcorp...| 2017-04-25| 70| 780000.0| 7|
| 7| Mark|1-680-102-6792|Quisque.ac@placer...| 2006-04-21| 52| 811200.0| 5|
| 8| Jonah| 214-2975|eu.ultrices.sit@v...| 2017-10-07| 23| 1704000.0| 5|
| 9| Hanae| 935-2277| eu@Nunc.ca| 2003-05-25| 69| 683400.0| 3|
| 10| Cadman|1-866-561-2701|orci.adipiscing.n...| 2001-05-19| 19| 799600.0| 7|
+---+-----+-----+-----+-----+-----+
only showing top 10 rows
```

- Nuevo campo con WITH COLUMN.

```
: df_with_schema.withColumn("NUEVO SALARIO",col("SALARIO")*1.05).show(3)

+---+-----+-----+-----+-----+-----+-----+
| ID| NOMBRE| TELEFONO| CORREO|FECHA_INGRESO|EDAD| SALARIO|ID_EMPRESA|NUEVO SALARIO|
+---+-----+-----+-----+-----+-----+-----+
| 1| Carl|1-745-633-9145|arcu.Sed.et@ante....| 2004-04-23| 32| 20095.0| 5| 21099.75|
| 2| Priscilla| 155-2498|Donec.egestas.Ali...| 2019-02-17| 34| 9298.0| 2| 9762.9|
| 3| Jocelyn|1-204-956-8594|amet.diam@loborti...| 2002-08-01| 27| 10853.0| 3| 11395.65|
+---+-----+-----+-----+-----+-----+-----+
only showing top 3 rows
```

- Creando campo con ciertas condiciones.

```

:e_salario_bajo = 5000
:o_salario_alto = 30000

df.schema.withColumn("TIPO_SALARIO", when(col("SALARIO")<5000,"Salario Bajo") \
    .when(((col("SALARIO")>=5000) & (col("SALARIO")<inicio_salario_alto)),"Salario Medio") \
    .otherwise("Salario Alto")).show(10)

```

ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO	EDAD	SALARIO	ID_EMPRESA	TIPO_SALARIO
1	Carl	1-745-633-9145	arcu.Sed.et@ante....	2004-04-23	32	20095.0		5 Salario Medio
2	Priscilla	155-2498	Donec.egestas.Ali...	2019-02-17	34	9298.0		2 Salario Medio
3	Jocelyn	1-204-956-8594	amet.diam@loborti...	2002-08-01	27	10853.0		3 Salario Medio
4	Aidan	1-719-862-9385	euismod.et.commod...	2018-11-06	29	3387.0	10	10 Salario Bajo
5	Leandra	839-8044	at@pretiumetrutru...	2002-10-10	41	22102.0		1 Salario Medio
6	Bert	797-4453	a.felis.ullamcorp...	2017-04-25	70	7800.0		7 Salario Medio
7	Mark	1-680-102-6792	Quisque.ac@placer...	2006-04-21	52	8112.0		5 Salario Medio
8	Jonah	214-2975	eu.ultrices.sit@v...	2017-10-07	23	17040.0		5 Salario Medio
9	Hanae	935-2277	eu@Nunc.ca	2003-05-25	69	6834.0		3 Salario Medio
10	Cadman	1-866-561-2701	orci.adipiscing.n...	2001-05-19	19	7996.0		7 Salario Medio

- Función WithColumnRenamed.

```

df_with_schema.withColumnRenamed("SALARIO","SALARY").show()

```

ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO	EDAD	SALARY	ID_EMPRESA
1	Carl	1-745-633-9145	arcu.Sed.et@ante....	2004-04-23	32	20095.0	5
2	Priscilla	155-2498	Donec.egestas.Ali...	2019-02-17	34	9298.0	2
3	Jocelyn	1-204-956-8594	amet.diam@loborti...	2002-08-01	27	10853.0	3
4	Aidan	1-719-862-9385	euismod.et.commod...	2018-11-06	29	3387.0	10
5	Leandra	839-8044	at@pretiumetrutru...	2002-10-10	41	22102.0	1
6	Bert	797-4453	a.felis.ullamcorp...	2017-04-25	70	7800.0	7
7	Mark	1-680-102-6792	Quisque.ac@placer...	2006-04-21	52	8112.0	5
8	Jonah	214-2975	eu.ultrices.sit@v...	2017-10-07	23	17040.0	5
9	Hanae	935-2277	eu@Nunc.ca	2003-05-25	69	6834.0	3
10	Cadman	1-866-561-2701	orci.adipiscing.n...	2001-05-19	19	7996.0	7
11	Melyssa	596-7736	vel@vulputateposu...	2008-10-14	48	4913.0	8
12	Tanner	1-739-776-7897	arcu.Aliquam.ultr...	2011-05-10	24	19943.0	8
13	Trevor	512-1955	Nunc.quis.arcu@eg...	2010-08-06	34	9501.0	5
14	Allen	733-2795	felis.Donec@necle...	2005-03-07	59	16289.0	2
15	Wanda	359-6973	Nam.nulla.magna@I...	2005-08-21	27	1539.0	5
16	Alphonse	241-8700	Ullamcorper.sed...	2006-10-25	26	2277.0	2

- Eliminando campo CORREO

```
: df_with_schema.drop("CORREO").show()
```

ID	NOMBRE	TELEFONO	FECHA_INGRESO	EDAD	SALARIO	ID_EMPRESA
1	Carl	1-745-633-9145	2004-04-23	32	20095.0	5
2	Priscilla	155-2498	2019-02-17	34	9298.0	2
3	Jocelyn	1-204-956-8594	2002-08-01	27	10853.0	3
4	Aidan	1-719-862-9385	2018-11-06	29	3387.0	10
5	Leandra	839-8044	2002-10-10	41	22102.0	1
6	Bert	797-4453	2017-04-25	70	7800.0	7
7	Mark	1-680-102-6792	2006-04-21	52	8112.0	5
8	Jonah	214-2975	2017-10-07	23	17040.0	5
9	Hanae	935-2277	2003-05-25	69	6834.0	3
10	Cadman	1-866-561-2701	2001-05-19	19	7996.0	7
11	Melyssa	596-7736	2008-10-14	48	4913.0	8
12	Tanner	1-739-776-7897	2011-05-10	24	19943.0	8
13	Trevor	512-1955	2010-08-06	34	9501.0	5
14	Allen	733-2795	2005-03-07	59	16289.0	2
15	Wanda	359-6973	2005-08-21	27	1539.0	5

- Función lit.

```
: df_with_schema.withColumn("PERIODO", lit("202503")).show(19)
```

ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO	EDAD	SALARIO	ID_EMPRESA	PERIODO
1	Carl	1-745-633-9145	arcu.Sed.et@ante....	2004-04-23	32	20095.0	5	202503
2	Priscilla	155-2498	Donec.egestas.Ali...	2019-02-17	34	9298.0	2	202503
3	Jocelyn	1-204-956-8594	amet.diam@loborti...	2002-08-01	27	10853.0	3	202503
4	Aidan	1-719-862-9385	eiusmod.et.commod...	2018-11-06	29	3387.0	10	202503
5	Leandra	839-8044	at@pretiumetrutru...	2002-10-10	41	22102.0	1	202503
6	Bert	797-4453	a.felis.ullamcorp...	2017-04-25	70	7800.0	7	202503
7	Mark	1-680-102-6792	Quisque.ac@placer...	2006-04-21	52	8112.0	5	202503
8	Jonah	214-2975	eu.ultrices.sit@v...	2017-10-07	23	17040.0	5	202503
9	Hanae	935-2277	eu@Nunc.ca	2003-05-25	69	6834.0	3	202503
10	Cadman	1-866-561-2701	orci.adipiscing.n...	2001-05-19	19	7996.0	7	202503
11	Melyssa	596-7736	vel@vulputateposu...	2008-10-14	48	4913.0	8	202503
12	Tanner	1-739-776-7897	arcu.Aliquam.ultr...	2011-05-10	24	19943.0	8	202503
13	Trevor	512-1955	Nunc.quis.arcu@eg...	2010-08-06	34	9501.0	5	202503
14	Allen	733-2795	felis.Donec@necle...	2005-03-07	59	16289.0	2	202503

- Funcion de fecha.

```
[45]: fecha_actual=datetime.now()
print("hora del servidor: ",fecha_actual)
fecha_peru= fecha_actual - timedelta(hours=5)
print("hora de Perú: ",fecha_peru)
periodo = fecha_actual.strftime('%Y%m')
print(periodo)

hora del servidor: 2025-03-04 02:31:19.587169
hora de Perú: 2025-03-03 21:31:19.587169
202503
```

```
[ ]:
```

- Función FILTER.

```
: df_with_schema.filter(col("ID_EMPRESA") ==5).show(4)

+---+-----+-----+-----+-----+-----+
| ID|NOMBRE|    TELEFONO|          CORREO|FECHA_INGRESO|EDAD|SALARIO|ID_EMPRESA|
+---+-----+-----+-----+-----+-----+
|  1| Carl|1-745-633-9145|arcu.Sed.et@ante....| 2004-04-23| 32|20095.0|      5|
|  7| Mark|1-680-102-6792|Quisque.ac@placer...| 2006-04-21| 52| 8112.0|      5|
|  8| Jonah| 214-2975|eu.ultrices.sit@v...| 2017-10-07| 23|17040.0|      5|
| 13|Trevor| 512-1955|Nunc.quis.arcu@eg...| 2010-08-06| 34| 9501.0|      5|
+---+-----+-----+-----+-----+-----+
only showing top 4 rows
```

```
df_with_schema.filter((col("ID_EMPRESA") ==5) & (col("SALARIO">>10000)).show()

+---+-----+-----+-----+-----+-----+
| ID|NOMBRE|    TELEFONO|          CORREO|FECHA_INGRESO|EDAD|SALARIO|ID_EMPRESA|
+---+-----+-----+-----+-----+-----+
|  1| Carl|1-745-633-9145|arcu.Sed.et@ante....| 2004-04-23| 32|20095.0|      5|
|  8| Jonah| 214-2975|eu.ultrices.sit@v...| 2017-10-07| 23|17040.0|      5|
| 50| Ross|1-587-285-1837|at.risus@milacini...| 2009-11-03| 31|19092.0|      5|
| 59|Quemby| 930-5882|lorem.ut.aliquam@...| 2017-10-04| 26|12092.0|      5|
| 86| Jack| 860-9554|parturient.montes...| 2017-03-10| 58|14473.0|      5|
+---+-----+-----+-----+-----+-----+
```

- Función ISIN.

ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO	EDAD	SALARIO	ID_EMPRESA
1	Carl	1-745-633-9145	arcu.Sed.et@ante....	2004-04-23	32	20095.0	5
2	Priscilla	155-2498	Donec.egestas.Ali...	2019-02-17	34	9298.0	2
3	Jocelyn	1-204-956-8594	amet.diam@loborti...	2002-08-01	27	10853.0	3
7	Mark	1-680-102-6792	Quisque.ac@placer...	2006-04-21	52	8112.0	5
8	Jonah	214-2975	eu.ultrices.sit@v...	2017-10-07	23	17040.0	5
9	Hanae	935-2277	eu@Nunc.ca	2003-05-25	69	6834.0	3
13	Trevor	512-1955	Nunc.quis.arcu@eg...	2010-08-06	34	9501.0	5
14	Allen	733-2795	felis.Donec@necle...	2005-03-07	59	16289.0	2
15	Wanda	359-6973	Nam.nulla.magna@I...	2005-08-21	27	1539.0	5
16	Alden	341-8522	odio@morbitristiq...	2006-12-05	26	3377.0	2
22	Kibo	970-8006	scelerisque.lorem...	2012-04-25	22	7449.0	2
29	Jana	1-564-106-5562	sed.dolor.Fusce@S...	2008-10-18	39	6483.0	2
31	Rylee	306-5447	Sed.nunc@turbis.edu	2001-09-17	47	21591.0	3
33	Jin	1-620-779-3366	est.Nunc.ullamcor...	2016-11-07	42	22038.0	2
35	Aurora	1-865-751-3479	magna@Cras.net	2017-10-21	54	4588.0	5
37	Inga	767-6448	Mauris.quis.turpi...	2016-12-31	41	8562.0	3
40	Ross	387-0945	amet.faucibus@ips...	2009-07-27	67	14285.0	3
43	Yetta	986-0220	vitae@dapibusrutr...	2008-03-24	61	21452.0	2
50	Ross	1-587-285-1837	at.risus@milacini...	2009-11-03	31	19092.0	5
51	Damon	368-7630	nunc@dapibusquamq...	2016-08-11	49	2669.0	5

- Función ISIN negado

ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO	EDAD	SALARIO	ID_EMPRESA
4	Aidan	1-719-862-9385	euismod.et.commod...	2018-11-06	29	3387.0	10
5	Leandra	839-8044	at@premiumetruru...	2002-10-10	41	22102.0	1
6	Bert	797-4453	a.felis.ullamcor...	2017-04-25	70	7800.0	7
10	Cadman	1-866-561-2701	orci.adipiscing.n...	2001-05-19	19	7996.0	7
11	Melyssa	596-7736	vel@vulputateposu...	2008-10-14	48	4913.0	8
12	Tanner	1-739-776-7897	arcu.Aliquam.ultr...	2011-05-10	24	19943.0	8
17	Omar	720-1543	Phasellus.vitae.m...	2014-06-24	60	6851.0	6
18	Owen	1-167-335-7541	sociis@erat.com	2002-04-09	34	4759.0	7
19	Laura	1-974-623-2057	mollis@ornare.ca	2017-03-09	70	17403.0	4
20	Emery	1-672-840-0264	at.nisi@vel.org	2004-02-27	24	18752.0	9
21	Carissa	1-300-877-0859	dignissim.pharetr...	2011-10-16	31	1952.0	10
23	Samson	1-430-188-6663	urna.justo.faucib...	2011-12-15	51	8099.0	6
24	Amaya	1-448-826-9497	ullamcorper.magna...	2000-09-18	24	1801.0	8
25	Pearl	1-850-202-3373	vel.convallis@rho...	2018-12-21	52	14756.0	6
26	Brenden	1-455-726-9413	elit.pede.malesua...	2000-03-17	33	20549.0	7
27	Alexander	912-0676	semper.auctor.Mau...	2000-08-16	55	13813.0	6
28	Stephen	326-2020	arcu.Aliquam.ultr...	2016-09-04	53	9469.0	8
30	Clayton	1-599-614-5185	est.Nunc@dictumeu...	2008-10-08	52	9505.0	8
32	Gisela	406-8031	Praesent.luctus@d...	2002-08-21	67	6497.0	1
34	Lila	1-703-777-4150	viverra@sit.edu	2015-02-26	48	24305.0	4

- Función LIKE.

df_with_schema.filter(col("NOMBRE").like("%ar%")).show()							
ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO	EDAD	SALARIO	ID_EMPRESA
1	Carl	1-745-633-9145	arcu.Sed.et@ante....	2004-04-23	32	20095.0	5
7	Mark	1-680-102-6792	Quisque.ac@placer...	2006-04-21	52	8112.0	5
17	Omar	720-1543	Phasellus.vitae.m...	2014-06-24	60	6851.0	6
21	Carissa	1-300-877-0859	dignissim.pharetr...	2011-10-16	31	1952.0	10
25	Pearl	1-850-202-3373	vel.convallis@rho...	2018-12-21	52	14756.0	6
39	Carolyn	846-7060	metus.Aenean.sed@...	2013-05-29	64	22838.0	6
54	Lars	1-554-600-0855	commodo@Nam.edu	2005-06-22	25	20573.0	1
60	Bernard	492-8823	vel.faucibus@Done...	2005-04-15	27	10825.0	2
76	Omar	1-325-245-9578	elit.erat@utodiov...	2012-11-19	34	12163.0	6
87	Karly	1-644-725-7241	tempor.erat@feugi...	2011-06-12	25	3715.0	1

- Función Distinct.

```
+-----+-----+
|employee_name|department|salary|
+-----+-----+
|James        |Sales      |3000  |
|Michael     |Sales      |4600  |
|Robert       |Sales      |4100  |
|Maria        |Finance    |3000  |
|James        |Sales      |3000  |
|Scott        |Finance    |3300  |
|Jen          |Finance    |3900  |
|Jeff          |Marketing  |3000  |
|Kumar        |Marketing  |2000  |
|Saif          |Sales      |4100  |
+-----+-----+
```

: 10

```
[Stage 29:=====]
+-----+-----+
|employee_name|department|salary|
+-----+-----+
|      James| Sales | 3000|
|     Robert| Sales | 4100|
|      Maria| Finance| 3000|
| Michael| Sales | 4600|
|      Saif| Sales | 4100|
|     Scott| Finance| 3300|
|     Jeff| Marketing| 3000|
|      Jen| Finance| 3900|
|     Kumar| Marketing| 2000|
+-----+-----+
```

- Función DropDuplicates.

```
df.dropDuplicates().count()

9

df.dropDuplicates().show()

+-----+-----+-----+
|employee_name|department|salary|
+-----+-----+-----+
|      James|    Sales| 3000|
|     Robert|    Sales| 4100|
|      Maria|  Finance| 3000|
|   Michael|    Sales| 4600|
|      Saif|    Sales| 4100|
|     Scott|  Finance| 3300|
|      Jeff| Marketing| 3000|
|      Jen|  Finance| 3900|
|     Kumar| Marketing| 2000|
+-----+-----+-----+
```

```
: df.dropDuplicates(["department","Salary"]).count()

8

: df.dropDuplicates(["department","Salary"]).show()

[Stage 47:=====]
+-----+-----+-----+
|employee_name|department|salary|
+-----+-----+-----+
|      Maria|  Finance| 3000|
|     Scott|  Finance| 3300|
|      Jen|  Finance| 3900|
|     Kumar| Marketing| 2000|
|      Jeff| Marketing| 3000|
|      James|    Sales| 3000|
|     Robert|    Sales| 4100|
|   Michael|    Sales| 4600|
+-----+-----+-----+
```

- Función OrderBy.

ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO	EDAD	SALARIO	ID_EMPRESA
32	Gisela	406-8031	Praesent.luctus@d...	2002-08-21	67	6497.0	1
70	Suki	1-654-418-8537	varius.orci.in@co...	2010-05-15	43	12029.0	1
51	Leandra	839-8044	at@premiumetruru...	2002-10-10	41	22102.0	1
99	Ray	1-420-314-2886	ac.risus.Morbi@te...	2011-12-30	26	5570.0	1
54	Lars	1-554-600-0855	commodo@Nam.edu	2005-06-22	25	20573.0	1
87	Karly	1-644-725-7241	tempor.erat@feugi...	2011-06-12	25	3715.0	1
93	Althea	1-163-702-1244	sit.amet.null@el...	2002-09-01	24	8818.0	1
43	Yetta	986-0220	vitae@dapibusrutr...	2008-03-24	61	21452.0	2
14	Allen	733-2795	felis.Donec@necle...	2005-03-07	59	16289.0	2
78	Lenore	1-817-973-5592	Nunc.pulvinar.arc...	2014-07-07	57	1483.0	2
66	Adrian	220-8905	enim@mollisPhasel...	2018-03-15	46	22953.0	2
83	Giselle	1-515-799-6913	ipsum.Suspendisse...	2002-10-31	45	2503.0	2
33	Jin	1-620-779-3366	est.Nunc.uliamcor...	2016-11-07	42	22038.0	2
96	Amos	729-4665	non.lacinia@phare...	2017-11-27	42	15855.0	2
29	Jana	1-564-106-5562	sed.dolor.Fusce@S...	2008-10-18	39	6483.0	2
67	Buffy	1-538-190-2276	amet@at.ca	2005-10-21	38	15116.0	2
2	Priscilla	155-2498	Donec.egestas.Ali...	2019-02-17	34	9298.0	2
60	Bernard	492-8823	vel.faucibus@Done...	2005-04-15	27	10825.0	2
16	Alden	341-8522	odio@morbitristiq...	2006-12-05	26	3377.0	2
22	Kibo	970-8006	scelerisque.lorem...	2012-04-25	22	7449.0	2
81	Joy	1-379-384-0646	mi.Aliquam@nis1Nu...	2015-11-15	19	1256.0	2
9	Hanae	935-2277	e@Nunc.ca	2003-05-25	69	6834.0	3
40	Ross	387-0945	amet.faucibus@ips...	2009-07-27	67	14285.0	3
31	Rylee	306-5447	Sed.nunc@turpis.edu	2001-09-17	47	21591.0	3
37	Inga	767-6448	Mauris.quis.turpi...	2016-12-31	41	8562.0	3
85	Kennedy	596-0668	nibh.sit@famesac.ca	2011-04-25	41	7272.0	3
58	Igor	1-361-603-8276	mattis.Integer.eu...	2003-07-18	37	6191.0	3
69	Halee	1-208-996-8549	interdum.feugiat@...	2008-12-25	30	16782.0	3
64	Graiden	874-4897	massa.Mauris.vest...	2010-06-10	28	22037.0	3
2	2	1-304-956-8561	1-1-02-1-1-1-1	2002-09-04	27	140952.0	2

-Función GROUP BY.

ID_EMPRESA	sum(SALARIO)
7	106710.0
9	91678.0
3	151700.0
6	135243.0
1	79304.0
8	73319.0
5	136609.0
10	82012.0
4	155503.0
2	156377.0

- Función AGG para agregar más agrupaciones.

```
df.groupBy("ID_EMPRESA").agg(sum("SALARIO").alias("PLANILLA"), \
                           avg("EDAD").alias("PROM_EDAD"), \
                           max("SALARIO").alias("MAX_SALARY") \
).show()
```

ID_EMPRESA	PLANILLA	PROM_EDAD	MAX_SALARY
7	106710.0	34.55555555555556	21556.0
9	91678.0	37.666666666666664	23051.0
3	151700.0	39.63636363636363	23820.0
6	135243.0	50.0	22838.0
1	79304.0	35.857142857142854	22102.0
8	73319.0	39.888888888888886	19943.0
5	136609.0	41.214285714285715	20095.0
10	82012.0	40.888888888888886	24575.0
4	155503.0	38.875	24305.0
2	156377.0	39.785714285714285	22953.0

- Función Where en group by.

```
: df.groupBy("ID_EMPRESA").agg(sum("SALARIO").alias("PLANILLA"), \
                           avg("EDAD").alias("PROM_EDAD"), \
                           max("SALARIO").alias("MAX_SALARY") \
).where(col("PLANILLA")>=10000).show(10)
```

ID_EMPRESA	PLANILLA	PROM_EDAD	MAX_SALARY
7	106710.0	34.55555555555556	21556.0
9	91678.0	37.666666666666664	23051.0
3	151700.0	39.63636363636363	23820.0
6	135243.0	50.0	22838.0
1	79304.0	35.857142857142854	22102.0
8	73319.0	39.888888888888886	19943.0
5	136609.0	41.214285714285715	20095.0
10	82012.0	40.888888888888886	24575.0
4	155503.0	38.875	24305.0
2	156377.0	39.785714285714285	22953.0

- Función PARTITIONBY.

```
]#PARTITIONBY ----- FUNCIONA AL MOMENTO DE ESCRIBIR EN DISCO df.write.partitionby("campo_particionado")
#REPARTITION ----- CAMBIAR EL NRO DE PARTICIONES DE UN DATAFRAME DENTRO DE LA MEMORIA df.repartition(4)
#COALESCE ----- REDUCIR U OPTIMIZAR EL NRO DE PARTICIONES SIN USAR UN SHUFFLE df.coalesce()

5]: ruta_guardado = 'gs://dmc_datalake_dde_11_jacondex/archivos/persona_output/'
df.write.mode('overwrite').partitionBy("ID_EMPRESA").format("parquet").save(ruta_guardado)
```

- Inserción en nube de Google.

Detalles del bucket

us (varias regiones en Estados Unidos) Standard No público Borrar de forma no definitiva

OBJETOS CONFIGURACIÓN PERMISOS PROTECCIÓN CICLO DE VIDA OBSERVABILIDAD INFORMES DE

Navegador de carpetas

Depósitos > dmc_datalake_dde_11_jacondex > archivos > persona_output

CREAR CARPETA SUBIR TRANSFERIR LOS DATOS

Filtrar solo por prefijo de nombre Filtro Mostrar Solo objetos activos

Nombre	Tamaño	Tipo
ID_EMPRESA=1/	ca	Carpeta
ID_EMPRESA=10/	Lp	Carpeta
ID_EMPRESA=2/	e	Carpeta
ID_EMPRESA=3/	at	Carpeta
ID_EMPRESA=4/	s	Carpeta
ID_EMPRESA=5/	—	Carpeta
ID_EMPRESA=6/	—	Carpeta
ID_EMPRESA=7/	—	Carpeta
ID_EMPRESA=8/	—	Carpeta
ID_EMPRESA=9/	—	Carpeta
_SUCCESS	0 B	application/octet-stream

SESIÓN 09

CREACION DE CLUSTER Y NOTEBOOK

- Creación de claves en GCP.

```

{} phonic-biplane-450123-q5-3a1bab1eefbd.json 1 X
C: > Users > JENNER CONCO > Documents > DMC > Diploma DE > Curso 3 - Big Data Processing > Material Sesión 9-20250310 > {} phonic-biplane-450123-q5-3a1bab1eefbd.json > ...
1 {
2   "type": "service_account",
3   "project_id": "phonic-biplane-450123-q5",
4   "private_key_id": "3a1bab1eefbd82acf78b9ddbd7210551264cc24",
5   "private_key": "-----BEGIN PRIVATE KEY-----\nMIIEvgIBADANBgkqhkiG9w0BAQEFAASCBKgwggSkAgEAAoIBAQCeVrsi133sjN51\\nndwx6sstHQxuwKAGv3ATe5yECSpDxtqdhcZxowdaai9RciNPBVZjnaJ-\n-----END PRIVATE KEY-----\n",
6   "client_email": "databricks-server-community@phonic-biplane-450123-q5.iam.gserviceaccount.com",
7   "client_id": "101974024077057966670",
8   "auth_uri": "https://accounts.google.com/o/oauth2/auth",
9   "token_uri": "https://oauth2.googleapis.com/token",
10  "auth_provider_x509_cert_url": "https://www.googleapis.com/oauth2/v1/certs",
11  "client_x509_cert_url": "https://www.googleapis.com/robot/v1/metadata/x509/databricks-server-community%40phonic-biplane-450123-q5.iam.gserviceaccount.com",
12  "universe_domain": "googleapis.com"
13 }
14
15 spark.hadoop.google.cloud.auth.service.account.enable true
16 spark.hadoop.fs.gs.auth.service.account.email databricks-server-community@phonic-biplane-450123-q5.iam.gserviceaccount.com
17 spark.hadoop.fs.gs.project.id phonic-biplane-450123-q5
18 spark.hadoop.fs.gs.auth.service.account.private.key -----BEGIN PRIVATE KEY-----\nMIIEvgIBADANBgkqhkiG9w0BAQEFAASCBKgwggSkAgEAAoIBAQCeVrsi133sjN51\\nndwx6sstHQxuwKAGv3ATe5yECSpDxtqdhcZxowdaai9RciNPBVZjnaJ-\n-----END PRIVATE KEY-----\nspark.hadoop.fs.gs.auth.service.account.private.key.id 3a1bab1eefbd82acf78b9ddbd7210551264cc24

```

- Ejecución y uso de cluster con notebook.

```

Python ▾ Run all cluster-db Share Publish
1
▶ ✓ 1 minute ago (<1s) 1
#Importación de modulos
from pyspark.sql import SparkSession
from pyspark.sql.types import StructType, StructField, StringType, IntegerType, DoubleType

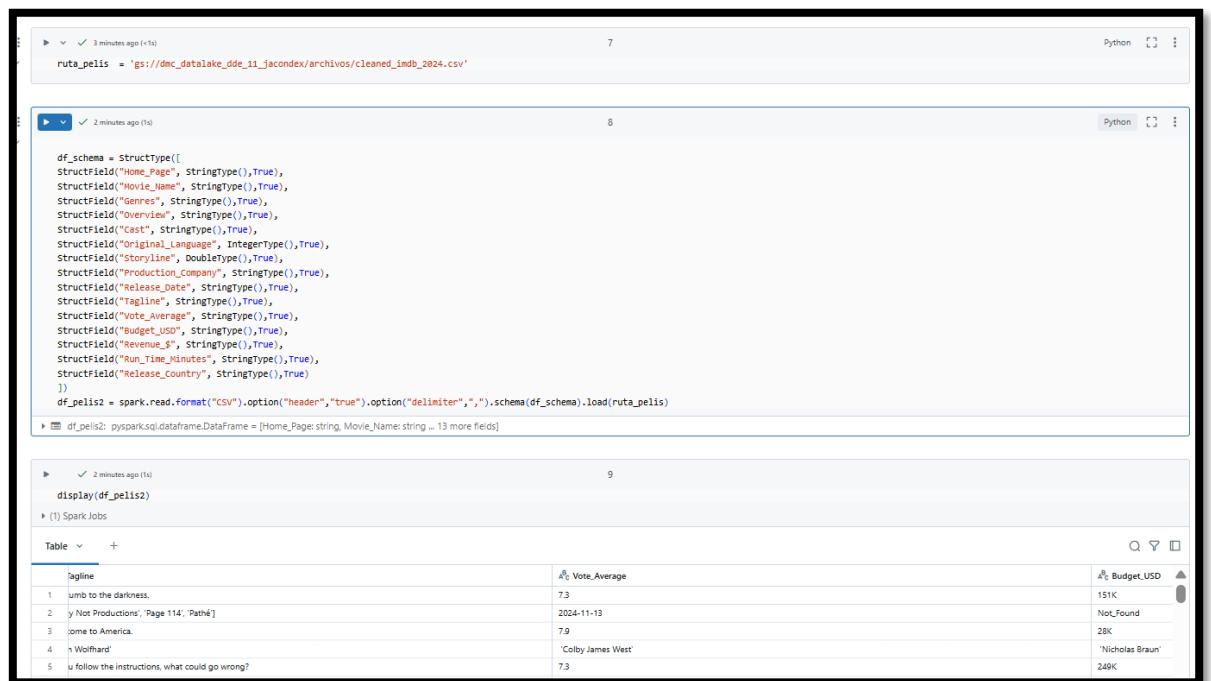
#Creamos la sesión de apache spark en una variable
spark = SparkSession.builder.getOrCreate()

2
▶ ✓ 1 minute ago (<1s) 2
ruta = 'gs://dmc_datalake_dde_11_jacondex/archivos/persona.data'

```

	ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO
1	1	Carl	1-745-633-9145	arcu.Sed.et@ante.co.uk	2004-04-23
2	2	Priscilla	155-2498	Donec.egestas.Aliquam@voluptatnunc.edu	2019-02-17
3	3	Jocelyn	1-204-956-8594	amet.diam@lobortis.co.uk	2002-08-01
4	4	Aidan	1-719-862-9385	euismod.et.commodo@nibhInacionaorci.edu	2018-11-06
5	5	Leandra	839-8044	at@pretiumetrutrum.com	2002-10-10
6	6	Bert	797-4453	a.felis.ullamcorper@arcu.org	2017-04-25
7	7	Mark	1-680-102-6792	Quisque.ac@placerat.ca	2006-04-21
8	8	Jonah	214-2975	eu.ultrices.sit@vitae.ca	2017-10-07

- Laboratorio con dataset descargado de kaggle.



The screenshot shows a Jupyter Notebook interface with three code cells and their corresponding outputs.

```

7
rutas_pelis = 'gs://dmc_datalake_dde_i1_jacondex/archivos/cleaned_imdb_2024.csv'

8
df_schema = StructType([
    StructField("Home_Page", StringType(), True),
    StructField("Movie_Name", StringType(), True),
    StructField("Genres", StringType(), True),
    StructField("Overview", StringType(), True),
    StructField("cast", StringType(), True),
    StructField("Original_Language", IntegerType(), True),
    StructField("Storyline", DoubleType(), True),
    StructField("Production_company", StringType(), True),
    StructField("Release_date", StringType(), True),
    StructField("Tagline", StringType(), True),
    StructField("Vote_Average", StringType(), True),
    StructField("Budget_USD", StringType(), True),
    StructField("Revenue_$", StringType(), True),
    StructField("Run_Time_Minutes", StringType(), True),
    StructField("Release_Country", StringType(), True)
])
df_pelis2 = spark.read.format("csv").option("header","true").option("delimiter",",").schema(df_schema).load(ruta_pelis)

9
display(df_pelis2)

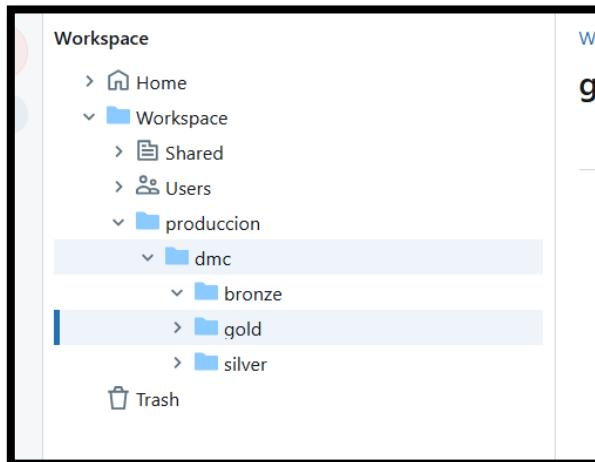
```

The third cell displays the first five rows of the DataFrame:

	Tagline	% Vote_Average	Budget_USD
1	unto to the darkness.	7.3	151K
2	y Not Productions, 'Page 114, 'Pathé'	2024-11-13	Not_Found
3	ome to America.	7.9	28K
4	'Wolfhard'	'Colby James West'	'Nicholas Braun'
5	u follow the instructions, what could go wrong?	7.3	246K

SESIÓN 10

- Creacion de carpetas en Databricks.



Name	Type	Owner
hubspot	Folder	ja.condex@gmail.com
salesforce	Folder	ja.condex@gmail.com
sap	Folder	ja.condex@gmail.com
sbs	Folder	ja.condex@gmail.com
sunat	Folder	ja.condex@gmail.com

- Creación de estructuras de carpetas en Cloud Storage

Nombre	Tamaño	Tipo	Fecha
bronze/	—	Carpeta	—
gold/	—	Carpeta	—
silver/	—	Carpeta	—

- Creando las variables en el notebook de databricks.

```

1
#Importación
from pyspark.sql import SparkSession
from pyspark.sql.types import StructType, StructField, StringType

2
#Variables
spark = SparkSession.builder.getOrCreate()
#Archivo en Cloud Storage - Google Cloud Platform
nombre_bucket = "dmc_datalake_dde_11_jacondex"
path_lakehouse = f"gs://{name_bucket}/produccion/dmc"
path_persona_landing = f"{path_lakehouse}/landing/personas/personas.data"
path_persona_bronze = f"{path_lakehouse}/bronze/personas/"

3
print(ruta_persona_bronze)
gs://dmc_datalake_dde_11_jacondex/produccion/dmc/bronze/personas/

```

- Cargando DF_PERSONAS

```

df_personas = spark.read.format("CSV").option("header", "true").option("delimiter", "|").schema(df_schema).load(path_persona_landing)

df_personas: pyspark.sql.dataframe.DataFrame = [ID: string, NOMBRE: string ... 6 more fields]

```

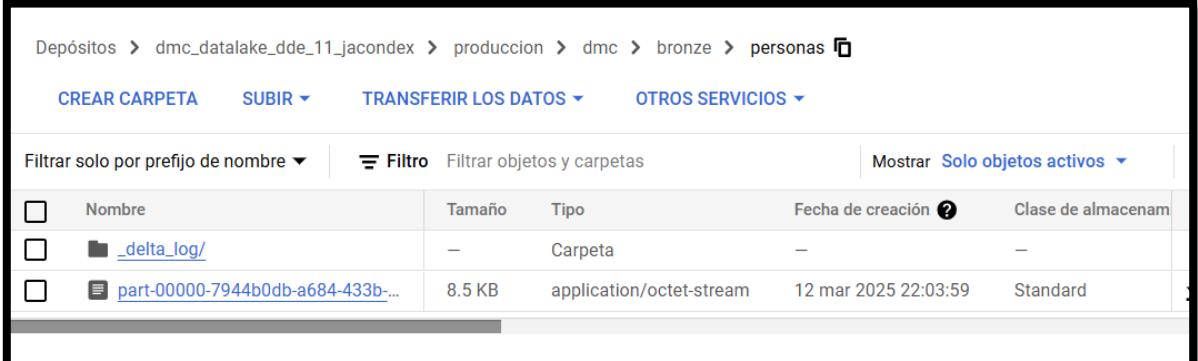
ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO	EDAD	SALAI
1	Carl	1-745-633-9145	arcu.Sed.et@ante.co.uk	2004-04-23	32	20095
2	Priscilla	155-2498	Donec.egestas.Aliquam@voluptatnunc.edu	2019-02-17	34	9298
3	Jocelyn	1-204-956-8594	amet.diam@lobortis.co.uk	2002-08-01	27	10853
4	Aidan	1-719-862-9385	euismod.et.commodo@nibhacinaorci.edu	2018-11-06	29	3387
5	Leandra	839-8044	at@prettiumetrurum.com	2002-10-10	41	22102
6	Bert	797-4453	a.felis.ulamcorper@arcu.org	2017-04-25	70	7800
7	Mark	1-680-102-6792	Quisque.ac@placerat.ca	2006-04-21	52	8112
8	Jonah	214-2975	eu.ultrices.sit@vitae.ca	2017-10-07	23	17040
9	Hanae	935-2277	eu@Nunc.ca	2003-05-25	69	6834
10	Cadman	1-866-561-2701	orci.adipiscing.non@semperNam.ca	2001-05-19	19	7996
11	Melyssa	596-7736	vel@vulputateposuerevulputate.net	2008-10-14	48	4913
12	Tanner	1-739-776-7897	arcu.Aliquam.ultrices@sociis.com	2011-05-10	24	19943
13	Trevor	512-1955	Nunc.quis.arcu@egestasa.org	2010-08-06	34	9501
14	Allen	733-2795	felis.Donec@nedleo.org	2005-03-07	59	16289

- Subiendo archivo delta a capa bronce.



```
+ Code + Text
1 minute ago (25s)
df_personas.write.mode("overwrite").format("delta").save(path_persona_bronze)

▶ (6) Spark Jobs
```



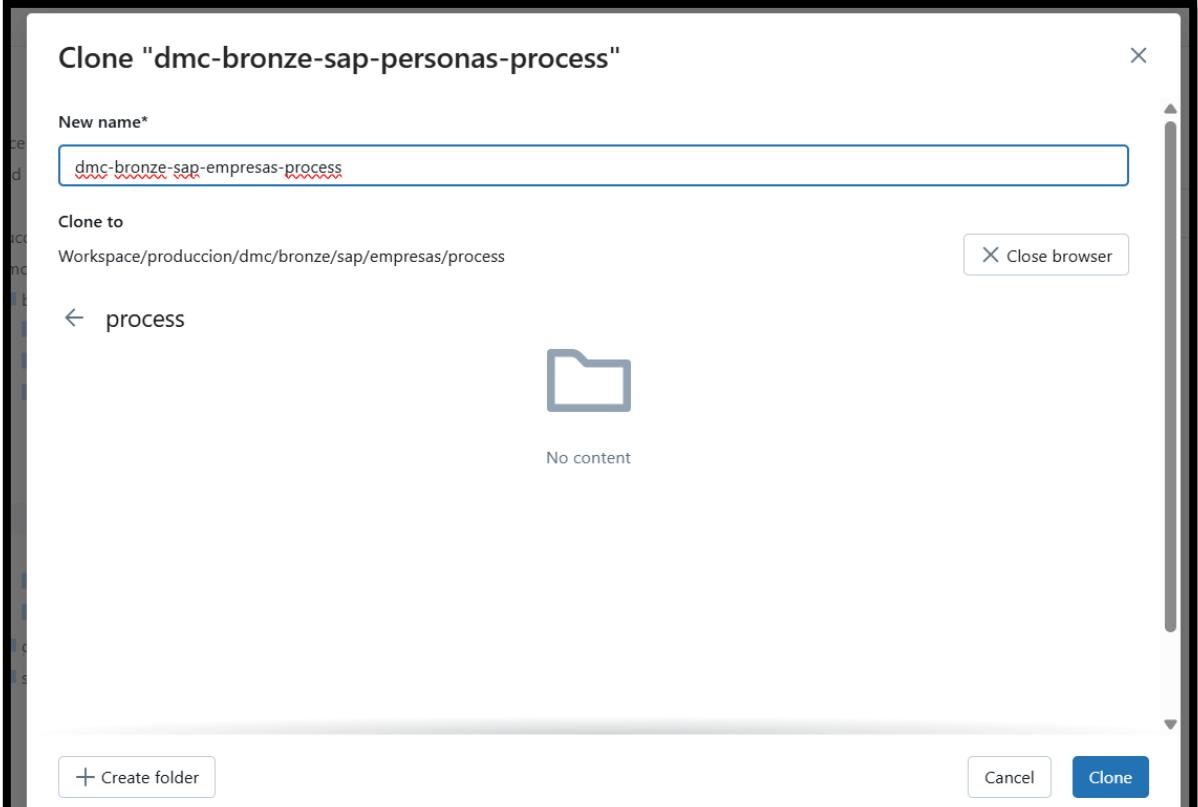
Depósitos > dmc_datalake_dde_11_jacondex > producción > dmc > bronze > personas

CREAR CARPETA SUBIR ▾ TRANSFERIR LOS DATOS ▾ OTROS SERVICIOS ▾

Filtrar solo por prefijo de nombre ▾ Filtro Filtrar objetos y carpetas Mostrar Solo objetos activos ▾

	Nombre	Tamaño	Tipo	Fecha de creación	Clase de almacenamiento
<input type="checkbox"/>	_delta_log/	—	Carpeta	—	—
<input type="checkbox"/>	part-00000-7944b0db-a684-433b-...	8.5 KB	application/octet-stream	12 mar 2025 22:03:59	Standard

- Clonando notebook.



Clone "dmc-bronze-sap-personas-process"

New name*
dmc-bronze-sap-empresas-process

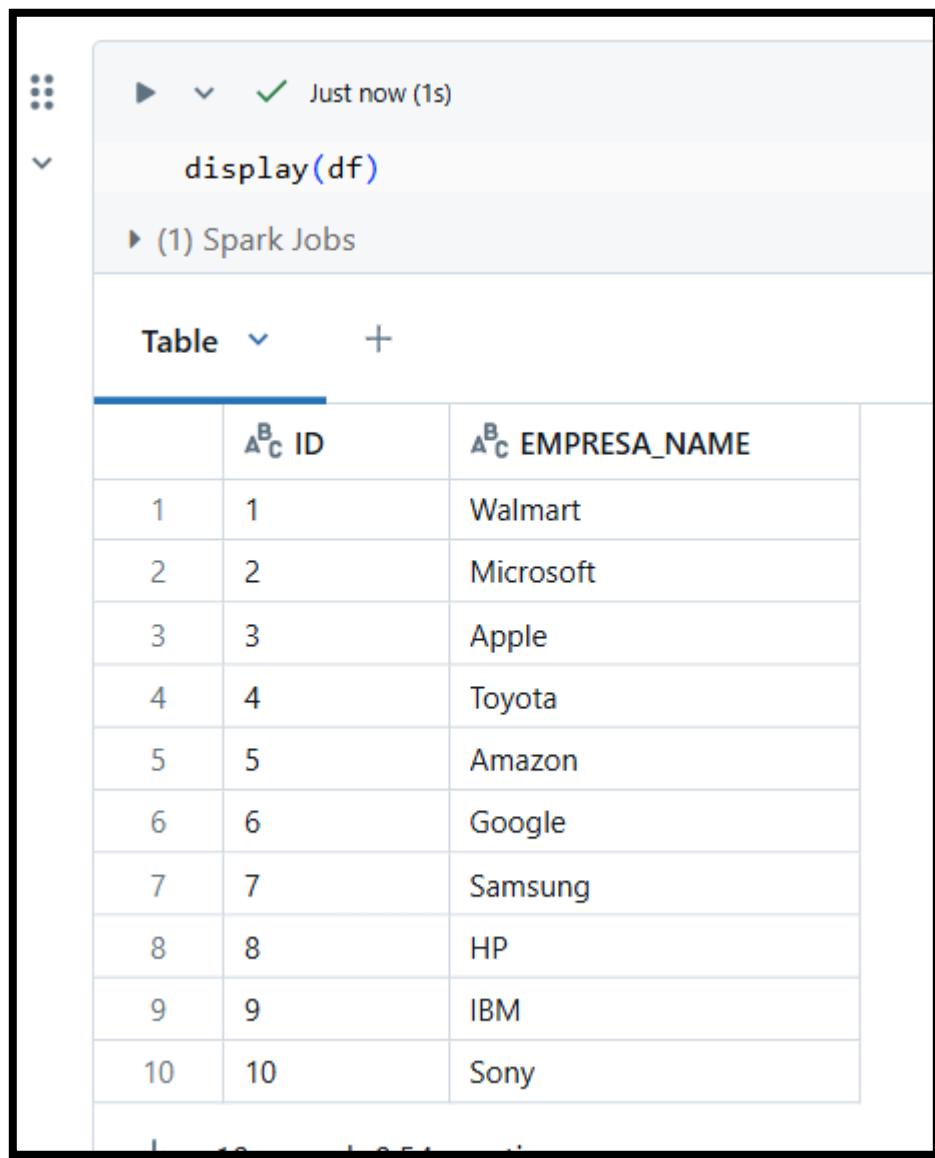
Clone to
Workspace/producción/dmc/bronze/sap/empresas/process

← process

No content

+ Create folder Cancel Clone

- Cargando Empresas.



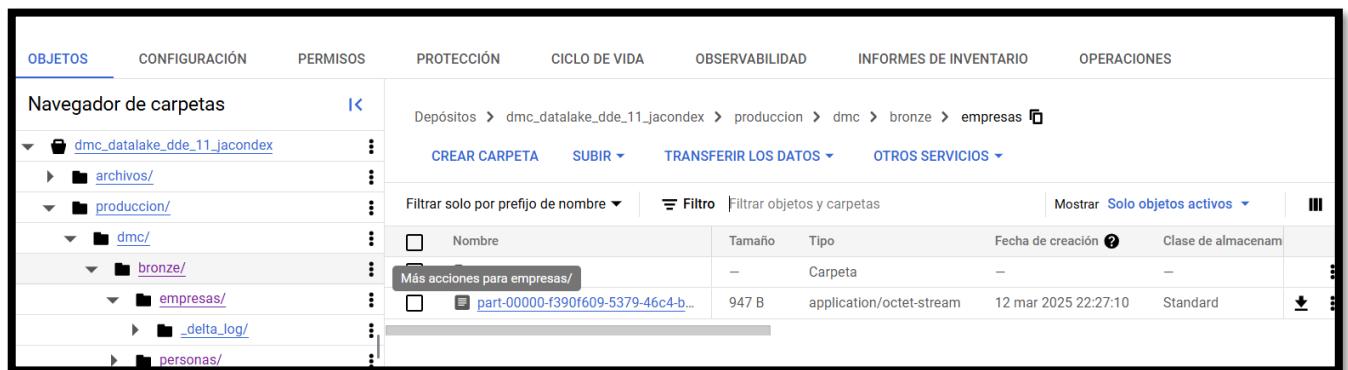
Just now (1s)

```
display(df)
```

(1) Spark Jobs

Table +

	A ^B _C ID	A ^B _C EMPRESA_NAME
1	1	Walmart
2	2	Microsoft
3	3	Apple
4	4	Toyota
5	5	Amazon
6	6	Google
7	7	Samsung
8	8	HP
9	9	IBM
10	10	Sony



OBJETOS CONFIGURACIÓN PERMISOS PROTECCIÓN CICLO DE VIDA OBSERVABILIDAD INFORMES DE INVENTARIO OPERACIONES

Navegador de carpetas I

Depósitos > dmc_datalake_dde_11_jacondex > produccion > dmc > bronze > empresas F

CREAR CARPETA SUBIR TRANSFERIR LOS DATOS OTROS SERVICIOS

Filtrar solo por prefijo de nombre F Filtro Filtrar objetos y carpetas Mostrar Solo objetos activos

Nombre	Tamaño	Tipo	Fecha de creación	Clase de almacenamiento
Más acciones para empresas/	—	Carpeta	—	—
part-00000-f390f609-5379-46c4-b...	947 B	application/octet-stream	12 mar 2025 22:27:10	Standard

- Cargando transacciones.

Just now (1s)

```
display(df)
```

(1) Spark Jobs

Table +

	ID_PERSONA	ID_EMPRESA	MONTO	FECHA
1	18	3	1383	2018-01-21
2	30	6	2331	2018-01-21
3	47	2	2280	2018-01-21
4	28	1	730	2018-01-21
5	91	4	3081	2018-01-21
6	74	8	2409	2018-01-21
7	41	2	3754	2018-01-22
8	42	9	4079	2018-01-22
9	24	6	4475	2018-01-22
10	67	9	561	2018-01-22
11	9	4	3765	2018-01-22
12	97	3	3669	2018-01-22
13	91	5	3497	2018-01-22
14	61	2	725	2018-01-22

Depósitos > dmc_datalake_dde_11_jacondex > producción > dmc > bronze > transacciones

CREAR CARPETA SUBIR ▾ TRANSFERIR LOS DATOS ▾ OTROS SERVICIOS ▾

Filtrar solo por prefijo de nombre ▾ Filtro Mostrar Solo ob

<input type="checkbox"/> Nombre	Tamaño	Tipo	Fecha de creación
<input type="checkbox"/> _delta_log/	—	Carpeta	—
<input type="checkbox"/> part-00000-77587818-cf60-4691...	635.7 KB	application/octet-stream	12 mar 2025 22:32:11
<input type="checkbox"/> part-00001-ccee2a70-22c1-4235...	158.4 KB	application/octet-stream	12 mar 2025 22:32:11

SESIÓN 11

- Cargando de capa bronce.

```
#Leer el archivo de origen
df = spark.read.format("delta").option("header","true").load(path_bronze)
display(df)

▶ (2) Spark Jobs

▶ df: pyspark.sql.dataframe.DataFrame = [ID: string, NOMBRE: string ... 6 more fields]
```

Table +

	ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO	
1	1	Carl	1-745-633-9145	arcu.Sed.et@ante.co.uk	2004-04-23	32
2	2	Priscilla	155-2498	Donec.egestas.Aliquam@voluptatnunc.edu	2019-02-17	34
3	3	Jocelyn	1-204-956-8594	amet.diam@lobortis.co.uk	2002-08-01	27
4	4	Aidan	1-719-862-9385	euismod.et.commodo@nibhacinaorci.edu	2018-11-06	29
5	5	Leandra	839-8044	at@pretiumetrurum.com	2002-10-10	41
6	6	Bert	797-4453	a.felis.ullamcorper@arcu.org	2017-04-25	70
7	7	Mark	1-680-102-6792	Quisque.ac@placerat.ca	2006-04-21	52
8	8	Jonah	214-2975	eu.ultrices.sit@vitae.ca	2017-10-07	23
9	9	Hanae	935-2277	eu@Nunc.ca	2003-05-25	69
10	10	Cadman	1-866-561-2701	orci.adipiscing.non@semperNam.ca	2001-05-19	19
11	11	Melyssa	596-7736	vel@vulputateposuerevulputate.net	2008-10-14	48
12	12	Tanner	1-739-776-7897	arcu.Aliquam.ultrices@sociis.com	2011-05-10	24

- Transformaciones y casteos en capa silver.

```
#La Transformación de datos
df_t = df.withColumn("TELEFONO", regexp_replace('telefono', '-', ''))\
    .withColumn("PERIODO", date_format(add_months(current_date(), -1), "yyyy-MM"))\
    .withColumn("SEGMENTO", when(col("SALARIO")<3500, "Masivo").when((col("SALARIO")>=3500) & (col("SALARIO")<=10000), "Premium").otherwise("Beyond"))

#Casteo de datos
df_c = df_t.withColumn("ID", col("ID").cast(IntegerType())) \
    .withColumn("ID_EMPRESA", col("ID_EMPRESA").cast(IntegerType())) \
    .withColumn("EDAD", col("EDAD").cast(IntegerType())) \
    .withColumn("SALARIO", col("SALARIO").cast(DoubleType())) \
    .withColumn("FECHA_INGRESO", to_date(col("FECHA_INGRESO"), "yyyy-MM-dd")) \
    .withColumn("ANIO", year(col("FECHA_INGRESO"))) \
    .withColumn("MES", month(col("FECHA_INGRESO"))) \
    .withColumn("DIA", dayofmonth(col("FECHA_INGRESO")))

▶ df_c: pyspark.sql.dataframe.DataFrame
ID: integer
NOMBRE: string
TELEFONO: string
CORREO: string
FECHA_INGRESO: date
EDAD: integer
SALARIO: double
ID_EMPRESA: integer
PERIODO: string
SEGMENTO: string
ANIO: integer
MES: integer
DIA: integer
```

- Data partitionada en capa Silver.

Nombre	Tamaño	Tipo
PERIODO=202502/	—	Carpeta
_delta_log/	—	Carpeta

- Cargando en capa silver – empresas, particionado por periodo.

Nombre	Tamaño	Tipo	Fecha de creación
PERIODO=202502/	—	Carpeta	—
_delta_log/	—	Carpeta	—

- Casteando en capa Silver – Transacciones.

```
#La Transformación de datos

#Casteo de datos
df_c = df.withColumn("ID_PERSONA",col("ID_PERSONA").cast(IntegerType()))
      .withColumn("ID_EMPRESA",col("ID_EMPRESA").cast(IntegerType()))
      .withColumn("MONTO",col("MONTO").cast(DoubleType()))
      .withColumn("FECHA",to_date(col("FECHA"),"yyyy-MM-dd"))
      .withColumn("ANIO",year(col("FECHA")))
      .withColumn("MES",month(col("FECHA")))
      .withColumn("DIA",dayofmonth(col("FECHA")))

df_c: pyspark.sql.dataframe.DataFrame = [ID_PERSONA: integer, ID_EMPRESA: integer ... 5 more fields]
```

display(df_c)

(1) Spark Jobs

	ID_PERSONA	ID_EMPRESA	MONTO	FECHA	ANIO	MES	DIA
1	18	3	1383	2018-01-21	2018	1	21
2	30	6	2331	2018-01-21	2018	1	21
3	47	2	2280	2018-01-21	2018	1	21
4	28	1	730	2018-01-21	2018	1	21
5	91	4	3081	2018-01-21	2018	1	21
6	74	8	2409	2018-01-21	2018	1	21
7	41	2	3754	2018-01-22	2018	1	22
8	42	9	4079	2018-01-22	2018	1	22
9	24	6	4475	2018-01-22	2018	1	22

- Guardo en cloud storage – transacciones.

OBJETOS CONFIGURACIÓN PERMISOS PROTECCIÓN CICLO DE VIDA OBSERVABILIDAD NUEVO INFORMES DE INVENTA

Navegador de carpetas

Depósitos > dmc_datalake_dde_11_jacondex > produccion > dmc > silver > transacciones

CREAR CARPETA SUBIR TRANSFERIR LOS DATOS OTROS SERVICIOS

Filtrar solo por prefijo de nombre Filtro Filtrar objetos y carpetas

Nombre	Tamaño	Tipo	Fecha de creación
ANIO=2018/	—	Carpeta	—
_delta_log/	—	Carpeta	—

SESIÓN 12

- Capa GOLD.

```
dmc-gold-reporting-analisis-salario-process Python

File Edit View Run Help Last edit was now

Workspace analysis_x_salario dmc-gold-reporting-analisis-salario-process

1
from pyspark.sql import SparkSession
#from pyspark.sql.types import StructType, StructField, StringType, IntegerType, DoubleType
#from pyspark.sql.functions import regexp_replace, date_format, current_date, add_months, when, col, to_date, year, month, dayofmonth

2
#Variables
spark = SparkSession.builder.getOrCreate()

3
#Archivo en Cloud Storage - Google Cloud Platform
name_bucket = "dmc_datalake_dde_11_jacondex"
path_lakehouse = f"gs://{name_bucket}/produccion/dmc"

#TABLAS INPUT
path_silver_personas = f"{path_lakehouse}/silver/personas/"
path_silver_empresas = f"{path_lakehouse}/silver/empresas/"

#Tabla output
path_gold = f"{path_lakehouse}/gold/machine-learning/analisis_x_salario/"
```

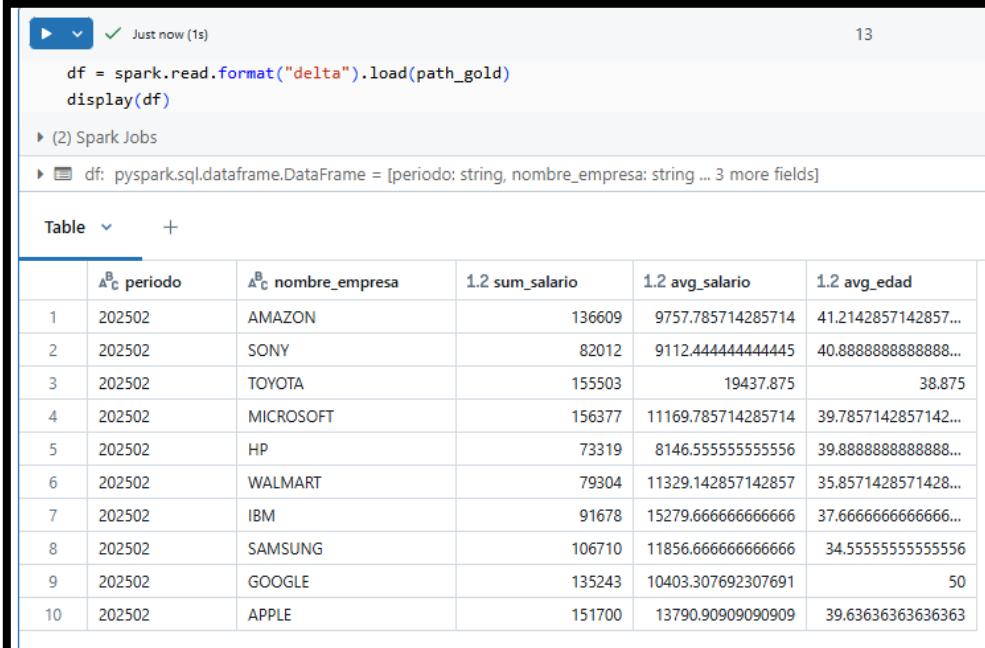
- Uso de select en tablas creadas.

```
7
sql = "SELECT * FROM tb_personas p inner join tb_empresas e on e.ID = p.ID_EMPRESA order by p.id_empresa"
df_result_1 = spark.sql(sql)

8
display(df_result_1)
(2) Spark Jobs
```

ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO	EDAD	SALARIO	ID_EMPRESA
1	Leandra	8398044	at@preiumetrurum.com	2002-10-10	41	22102	1
2	Gisela	4068031	Praesent.luctus@dui.co.uk	2002-08-21	67	6497	1
3	Lars	15546000855	commodo@Nam.edu	2005-06-22	25	20573	1
4	Suki	16544188537	varius.orci@congueit.co.uk	2010-05-15	43	12029	1
5	Karly	16447257241	tempor.erat@feugiatnon.ca	2011-06-12	25	3715	1
6	Althea	11637021244	sit.amet.nulla@elit.co.uk	2002-09-01	24	8818	1
7	Ray	14203142886	ac.risus.Morbi@telluseaugue.org	2011-12-30	26	5570	1
8	Priscilla	1552498	Donec.egestas.Aliquam@voluptatnunc.edu	2019-02-17	34	9298	2
9	Allen	7332795	felis.Donec@necleo.org	2005-03-07	59	16289	2
10	Alden	3418522	odio@morbitristiquesenectus.ca	2006-12-05	26	3377	2
11	Kibo	9708006	scelerisque.lorem@mattis.org	2012-04-25	22	7449	2
12	Jana	15641065562	sed.dolor.Fusce@Sedet.co.uk	2008-10-18	39	6483	2
13	Jin	16207793366	est.Nunc.uliamcorper@nullalInteger.co.uk	2016-11-07	42	22038	2
14	Yetta	9860220	vitae@dapibusrutrumjusto.co.uk	2008-03-24	61	21452	2
15							

- Poblando capa final GOLD.



```

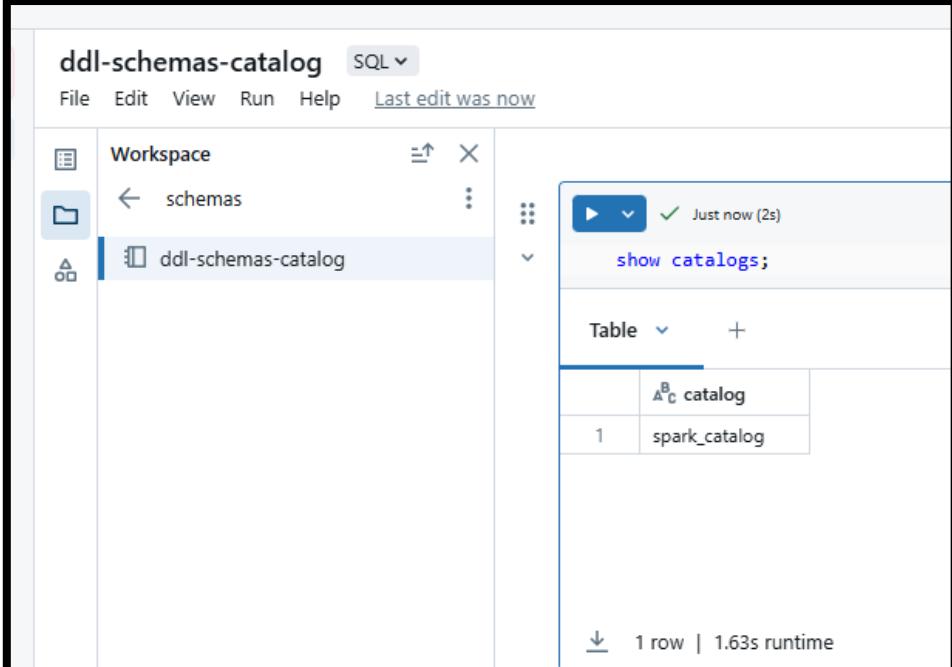
Just now (1s) 13
df = spark.read.format("delta").load(path_gold)
display(df)

▶ (2) Spark Jobs
▶ df: pyspark.sql.dataframe.DataFrame = [periodo: string, nombre_empresa: string ... 3 more fields]

Table +
```

	periodo	nombre_empresa	1.2 sum_salario	1.2 avg_salario	1.2 avg_edad
1	202502	AMAZON	136609	9757.785714285714	41.2142857142857...
2	202502	SONY	82012	9112.444444444445	40.888888888888...
3	202502	TOYOTA	155503	19437.875	38.875
4	202502	MICROSOFT	156377	11169.785714285714	39.7857142857142...
5	202502	HP	73319	8146.555555555556	39.888888888888...
6	202502	WALMART	79304	11329.142857142857	35.8571428571428...
7	202502	IBM	91678	15279.666666666666	37.666666666666...
8	202502	SAMSUNG	106710	11856.666666666666	34.55555555555556
9	202502	GOOGLE	135243	10403.307692307691	50
10	202502	APPLE	151700	13790.90909090909	39.63636363636363

- Creación de notebook DDL.



ddl-schemas-catalog SQL

File Edit View Run Help Last edit was now

Workspace

schemas

ddl-schemas-catalog

show catalogs;

Table +

catalog

1 spark_catalog

1 row | 1.63s runtime

The image shows two separate Databricks SQL query results:

Query 1 (Top):

```
Just now (<1s)
show databases;
```

Table +

databaseName
default

↓ 1 row | 0.29s runtime

Query 2 (Bottom):

```
Just now (<1s)
show tables in default;
```

Table +

database	tableName	isTemporary
----------	-----------	-------------

No rows returned

- Creacion de 3 capas en databricks.

The screenshot shows the Databricks Data browser interface. On the left, under 'Databases', the 'bronze' database is selected. The 'Tables' section shows 'No Tables'. On the right, there are three entries in the notebook:

- XISTS bronze
alake_dde_11_jacondex/produccion/dmc/bronze';
- XISTS silver
alake_dde_11_jacondex/produccion/dmc/silver';
- XISTS gold
alake_dde_11_jacondex/produccion/dmc/gold';

- Creación de tabla bronze.personas.

A screenshot of a Databricks notebook cell. The code is:

```
create external table bronze.personas(
    ID string,
    NOMBRE STRING,
    TELEFONO STRING,
    CORREO string,
    FECHA_INGRESO string,
    EDAD STRING,
    SALARIO STRING,
    ID_EMPRESA string
)
using delta
location "gs://dmc_datalake_dde_11_jacondex/produccion/dmc/bronze/personas/"
```

The cell status is "Just now (19s)" with a green checkmark. Below the code, it says "(3) Spark Jobs". At the bottom, there is an "OK" button.

A screenshot of a Databricks notebook cell. The code is:

```
select id_empresa, count(1) from bronze.personas group by ID_EMPRESA
```

The cell status is "Just now (2s)" with a green checkmark. Below the code, it says "(2) Spark Jobs".

Below the code, there is a table visualization:

	A ^B C id_empresa	i ² 3 count(1)
1	7	9
2	3	11
3	8	9
4	5	14
5	6	13
6	9	6
7	1	7
8	10	9
9	4	8
10	2	14

At the bottom of the table, it says "10 rows | 2.17s runtime".

- Creación de tabla bronze.empresas.

The screenshot shows a Databricks notebook interface with two code cells and a results section.

Top Cell:

```
▶ ▾ ✓ Just now (5s)
create external table bronze.empresas(
| ID string,
| EMPRESA_NAME STRING
)
using delta
location "gs://dmc_datalake_dde_11_jacondex/produccion/dmc/bronze/empresas/"

▶ (3) Spark Jobs
OK
```

Bottom Cell:

```
▶ ▾ ✓ Just now (1s)
SELECT * FROM bronze.empresas
▶ (2) Spark Jobs
```

Results Section:

Table ▼ +

	ID	EMPRESA_NAME
1	1	Walmart
2	2	Microsoft
3	3	Apple
4	4	Toyota
5	5	Amazon
6	6	Google
7	7	Samsung
8	8	HP
9	9	IBM
10	10	Sony

↓ 10 rows | 1.35s runtime

- Creación de tabla bronze.transacciones.

The screenshot shows two code cells in a Databricks notebook.

Code Cell 1:

```
create external table bronze.transacciones(
    ID_PERSONA string,
    ID_EMPRESA STRING,
    MONTO STRING,
    FECHA STRING
)
using delta
location "gs://dmc_datalake_dde_11_jacondex/produccion/dmc/bronze/transacciones/"
```

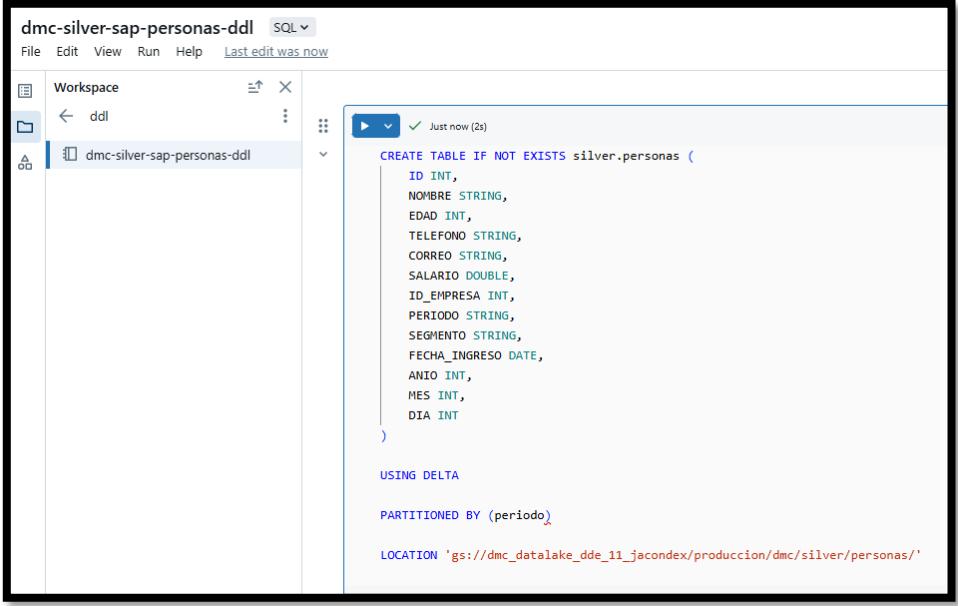
Code Cell 2:

```
select * from bronze.transacciones
```

The second cell displays the results of the query as a table:

	ID_PERSONA	ID_EMPRESA	MONTO	FECHA
1	18	3	1383	2018-01-21
2	30	6	2331	2018-01-21
3	47	2	2280	2018-01-21
4	28	1	730	2018-01-21
5	91	4	3081	2018-01-21
6	74	8	2409	2018-01-21
7	41	2	3754	2018-01-22
8	42	0	4070	2018-01-22

- Creando tabla silver.personas.



```

dmc-silver-sap-personas-ddl SQL
File Edit View Run Help Last edit was now

Workspace
ddl
dmc-silver-sap-personas-ddl

CREATE TABLE IF NOT EXISTS silver.personas (
    ID INT,
    NOMBRE STRING,
    EDAD INT,
    TELEFONO STRING,
    CORREO STRING,
    SALARIO DOUBLE,
    ID_EMPRESA INT,
    PERIODO STRING,
    SEGMENTO STRING,
    FECHA_INGRESO DATE,
    ANIO INT,
    MES INT,
    DIA INT
)

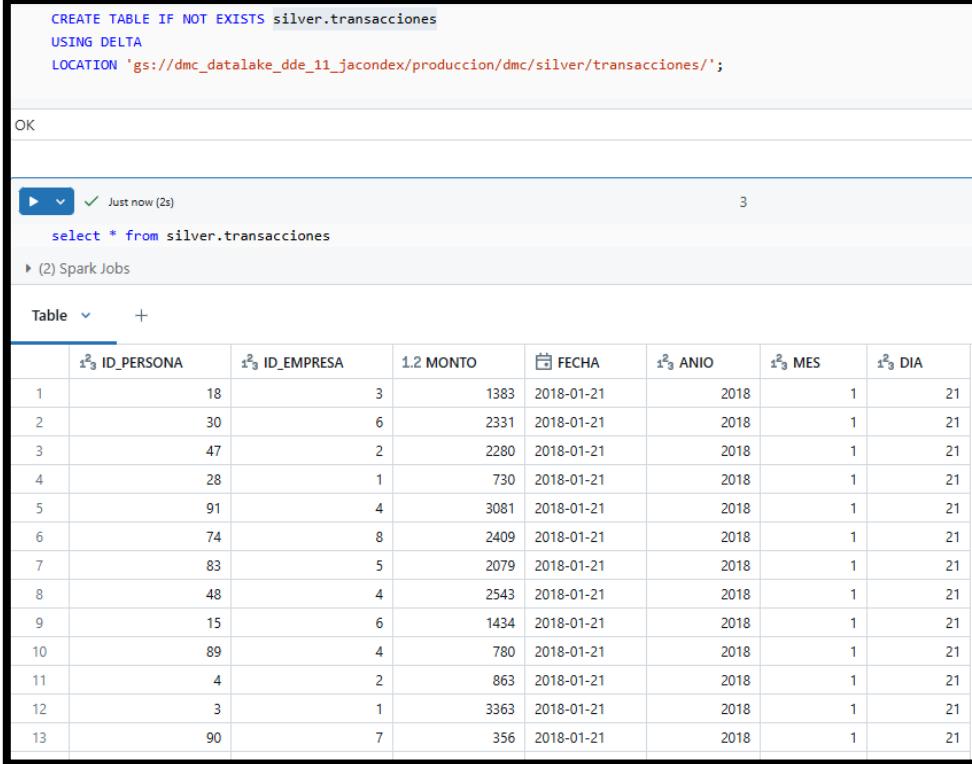
USING DELTA

PARTITIONED BY (periodo)

LOCATION 'gs://dmc_datalake_dde_11_jacondex/produccion/dmc/silver/personas/'

```

- Creando tabla silver.transacciones.



```

CREATE TABLE IF NOT EXISTS silver.transacciones
USING DELTA
LOCATION 'gs://dmc_datalake_dde_11_jacondex/produccion/dmc/silver/transacciones/';

OK

Just now (2s) 3
select * from silver.transacciones
▶ (2) Spark Jobs

Table +
```

	ID_PERSONA	ID_EMPRESA	MONTO	FECHA	ANIO	MES	DIA
1	18	3	1383	2018-01-21	2018	1	21
2	30	6	2331	2018-01-21	2018	1	21
3	47	2	2280	2018-01-21	2018	1	21
4	28	1	730	2018-01-21	2018	1	21
5	91	4	3081	2018-01-21	2018	1	21
6	74	8	2409	2018-01-21	2018	1	21
7	83	5	2079	2018-01-21	2018	1	21
8	48	4	2543	2018-01-21	2018	1	21
9	15	6	1434	2018-01-21	2018	1	21
10	89	4	780	2018-01-21	2018	1	21
11	4	2	863	2018-01-21	2018	1	21
12	3	1	3363	2018-01-21	2018	1	21
13	90	7	356	2018-01-21	2018	1	21