TRABAJO FINAL

RESULTADOS DE LABORATORIOS

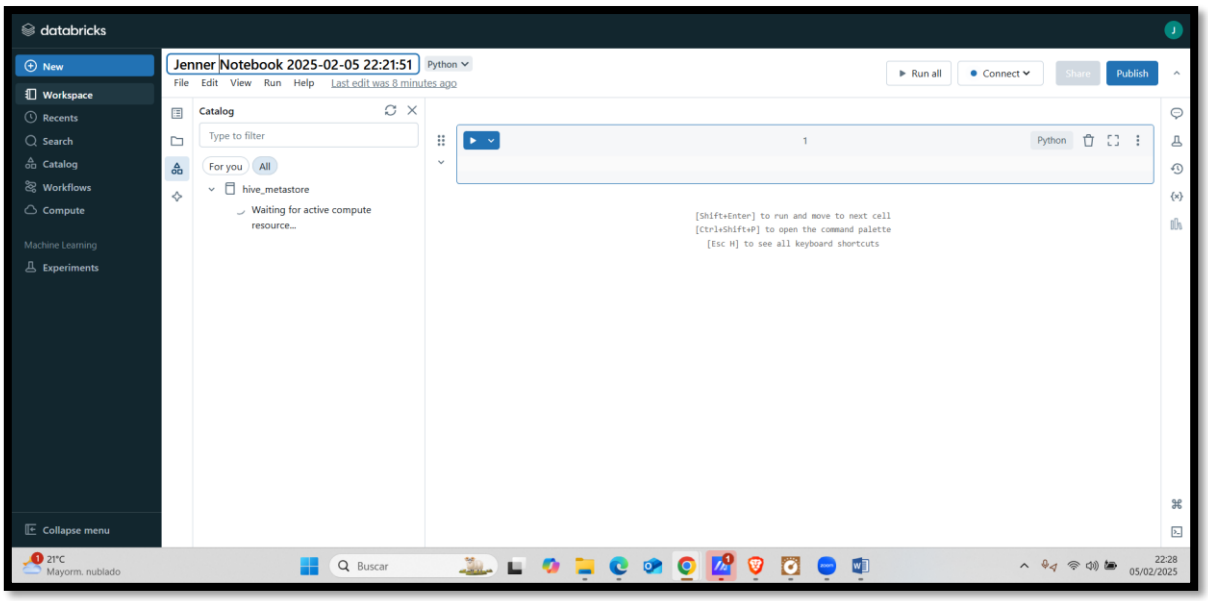
Curso: Diploma Data Engineer - Edición 11

Alumno: Jenner Conco Dextre

ÍNDICE

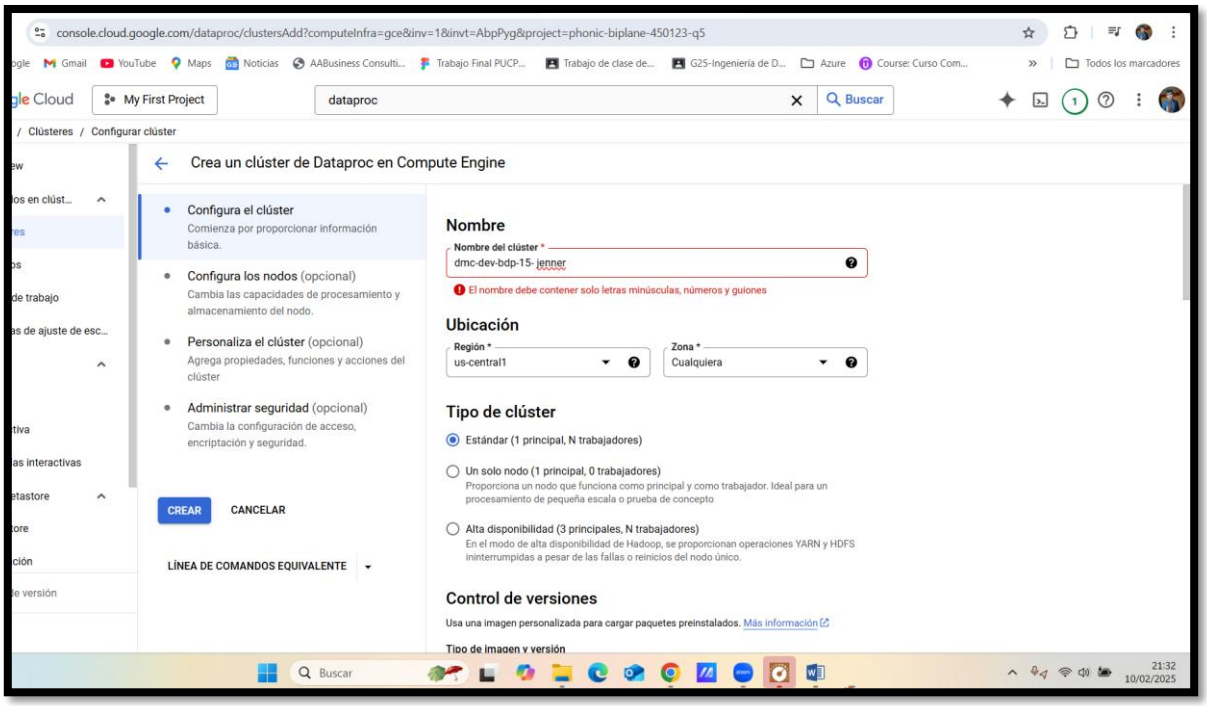
SESIÓN 01

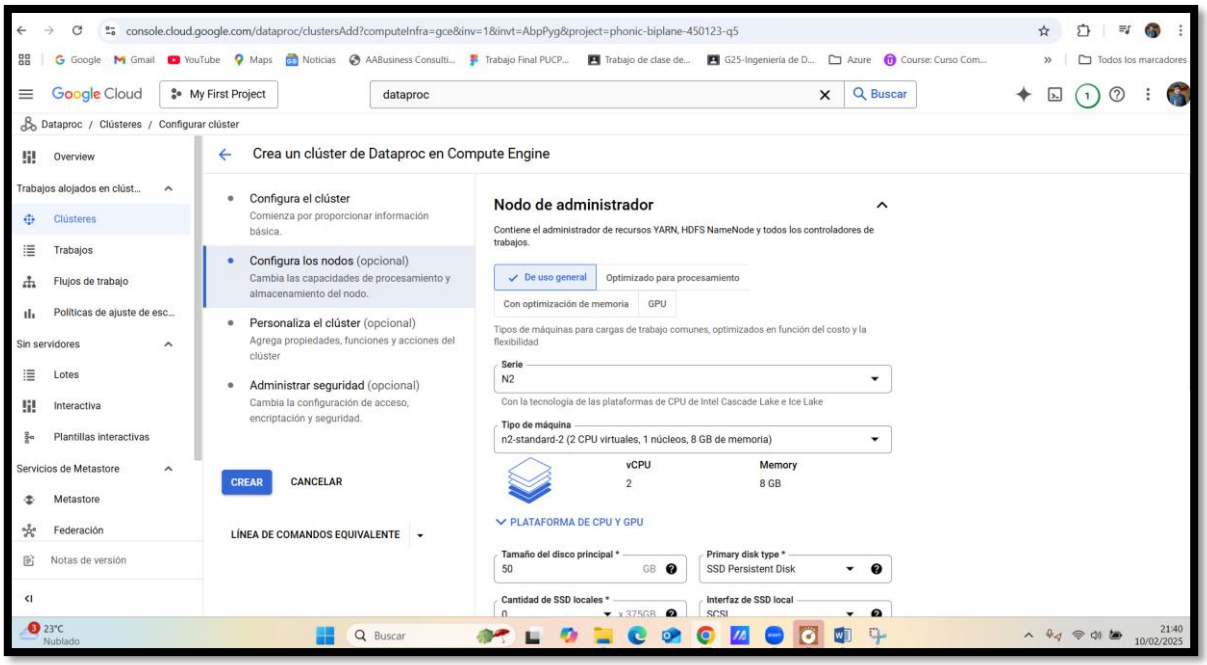
1. Laboratorio – Creación de un Cluster en Databricks.

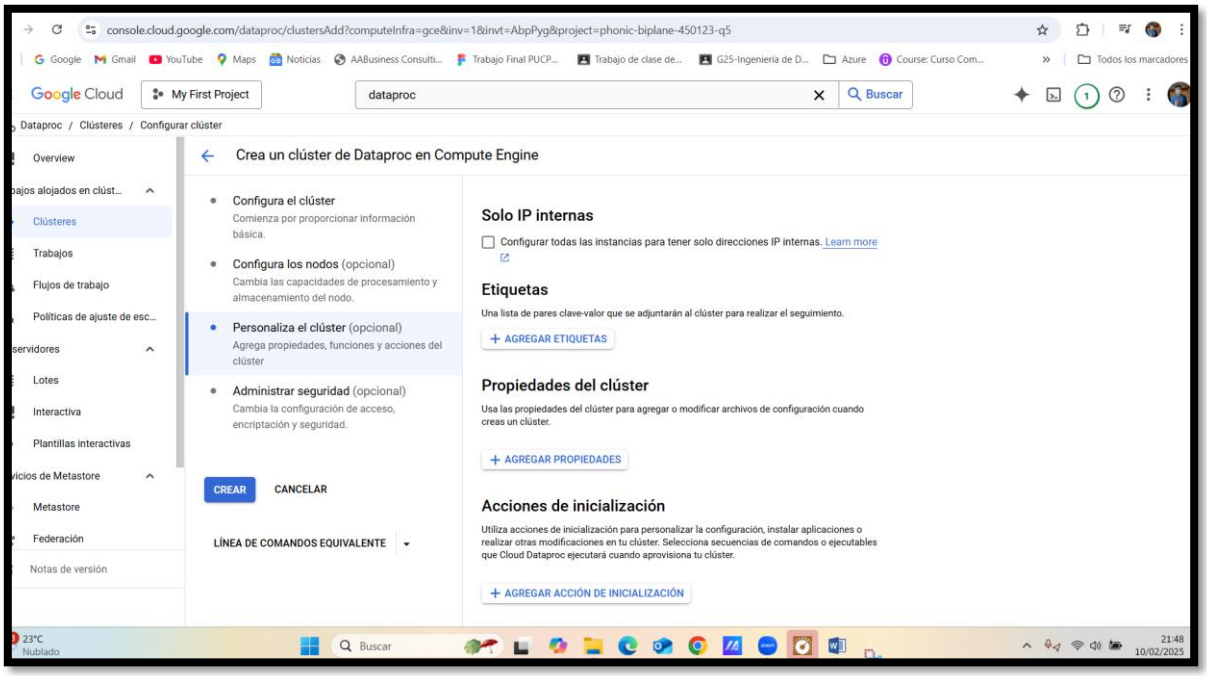


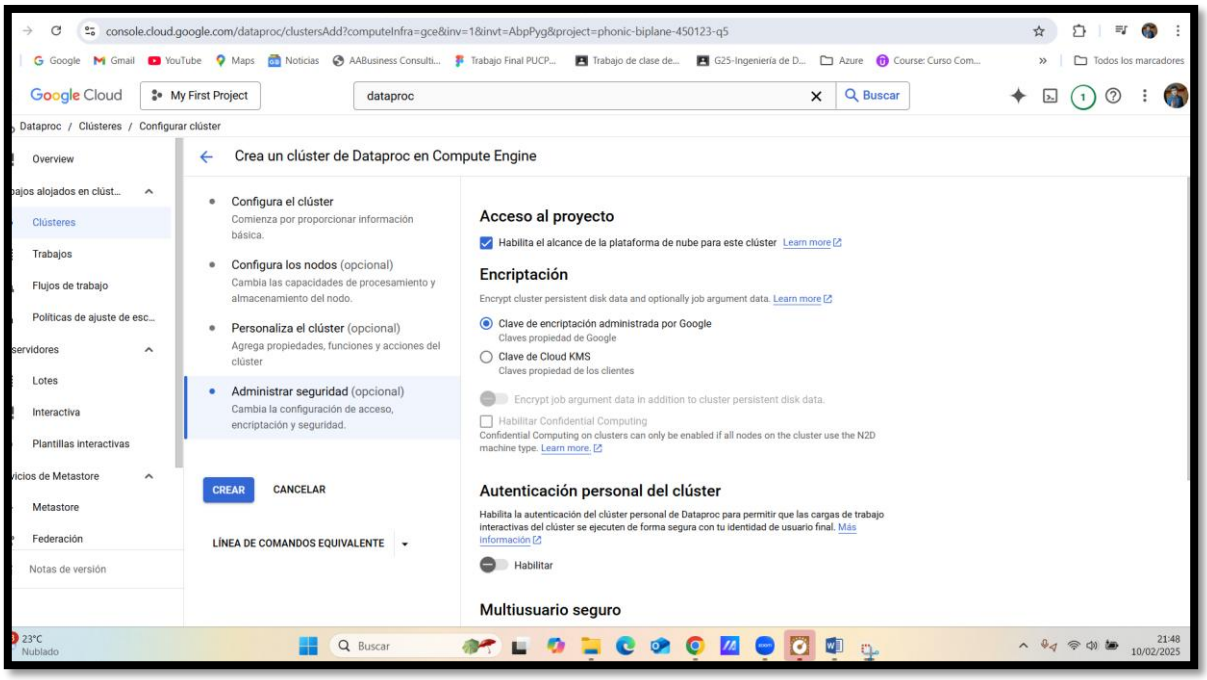
SESIÓN 02

1. Creando cluster en GCP.

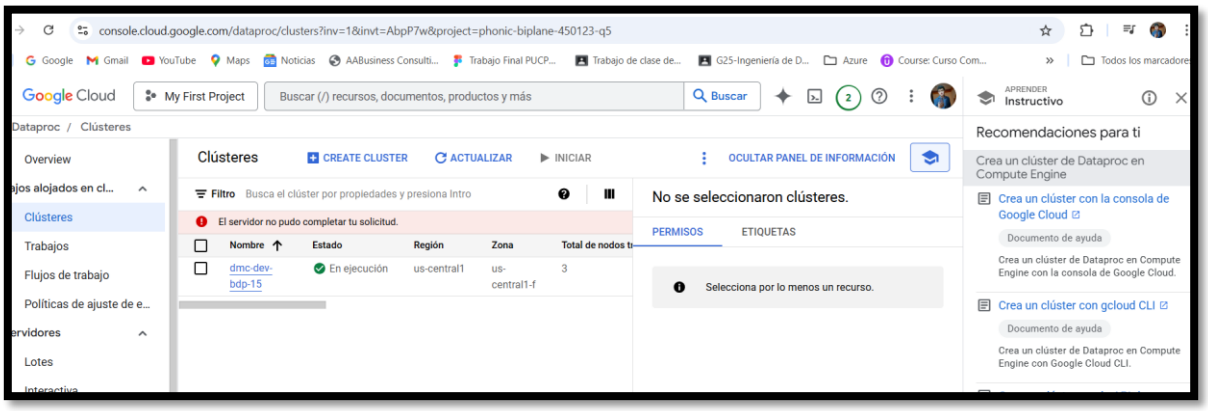



console.cloud.google.com/dataproc/clustersAdd?computeInfra=gce&inv=1&invt=AbpPyg&project=phonic-biplane-450123-q5
Google Cloud
My First Project
dataproc
Buscar
Dataproc / Clústeres / Configurar clúster
Overview
Trabajos alojados en clúst...
Clústeres
Trabajos
Flujos de trabajo
Políticas de ajuste de esc...
Sin servidores
Lotes
Interactiva
Plantillas interactivas
Servicios de Metastore
Metastore
Federación
Notas de versión
Crea un clúster de Dataproc en Compute Engine
Configura el clúster
Comienza por proporcionar información básica.
Configura los nodos (opcional)
Cambia las capacidades de procesamiento y almacenamiento del nodo.
Personaliza el clúster (opcional)
Agrega propiedades, funciones y acciones del clúster
Administrar seguridad (opcional)
Cambia la configuración de acceso, encriptación y seguridad.
CREAR
CANCELAR
LÍNEA DE COMANDOS EQUIVALENTE
Nodo de administrador
Contiene el administrador de recursos YARN, HDFS NameNode y todos los controladores de trabajos.
De uso general
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
Con la tecnología de las plataformas de CPU de Intel Cascade Lake e Ice Lake
Tipo de máquina
n2-standard-2 (2 CPU virtuales, 1 núcleos, 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 *
Interfaz de SSD local
SCSI
23°C
Nublado
Buscar
21:40
10/02/2025

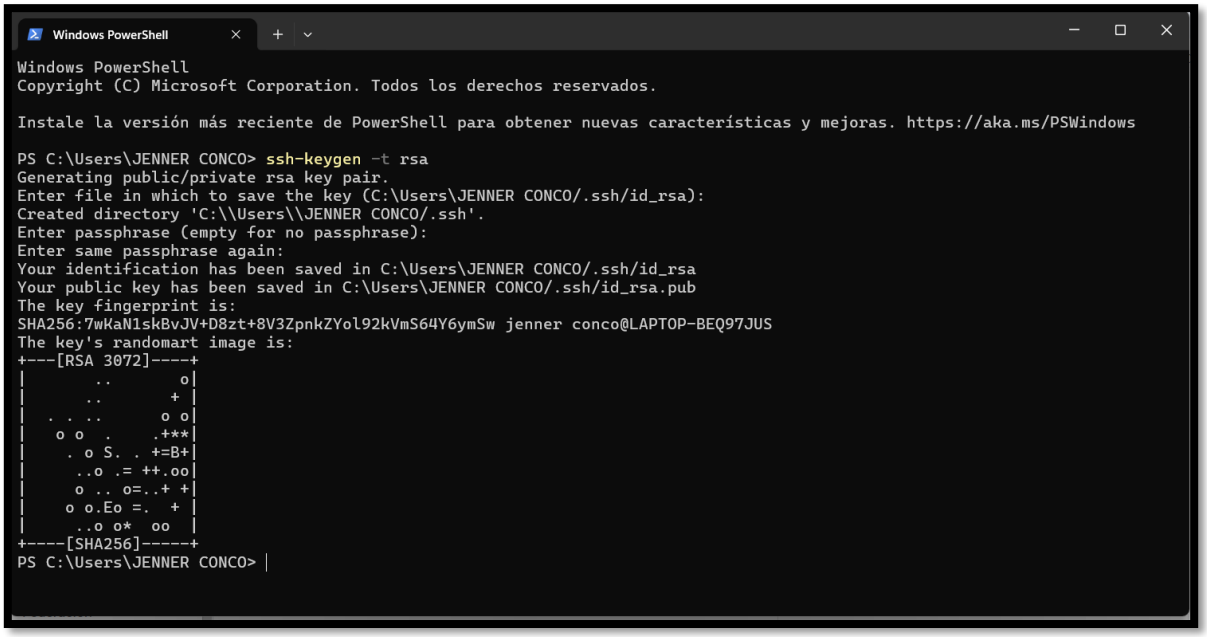

console.cloud.google.com/dataproc/clustersAdd?computeInfra=gce&inv=1&invt=AbpPyg&project=phonic-biplane-450123-q5
Google Cloud
My First Project
dataproc
Buscar
Dataproc / Clústeres / Configurar clúster
Overview
Clústeres
Trabajos
Flujos de trabajo
Políticas de ajuste de esc...
Lotes
Interactiva
Plantillas interactivas
Metastore
Federación
Notas de versión
Crea un clúster de Dataproc en Compute Engine
Configura el clúster
Comienza por proporcionar información básica.
Configura los nodos (opcional)
Cambia las capacidades de procesamiento y almacenamiento del nodo.
Personaliza el clúster (opcional)
Agrega propiedades, funciones y acciones del clúster
Administrar seguridad (opcional)
Cambia la configuración de acceso, encriptación y seguridad.
CREAR
CANCELAR
LÍNEA DE COMANDOS EQUIVALENTE
Solo IP internas
Configurar todas las instancias para tener solo direcciones IP internas. Learn more
Etiquetas
Una lista de pares clave-valor que se adjuntarán al clúster para realizar el seguimiento.
+ 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
23°C
Nublado
Buscar
21:48
10/02/2025

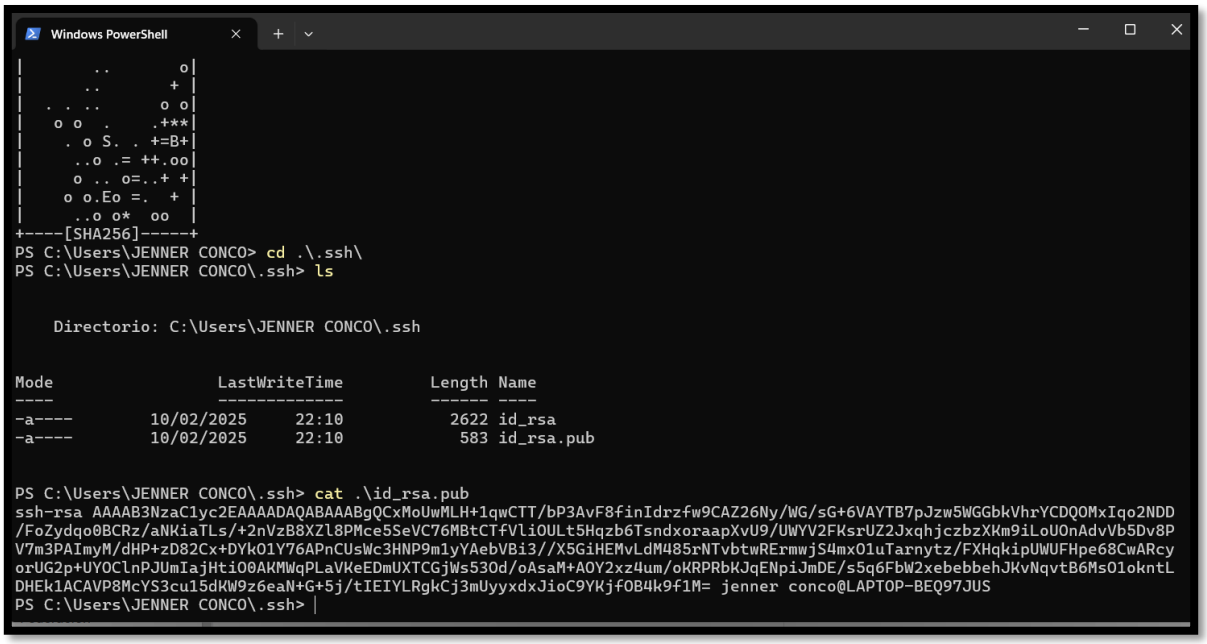


- Aprovisionamiento de cluster.

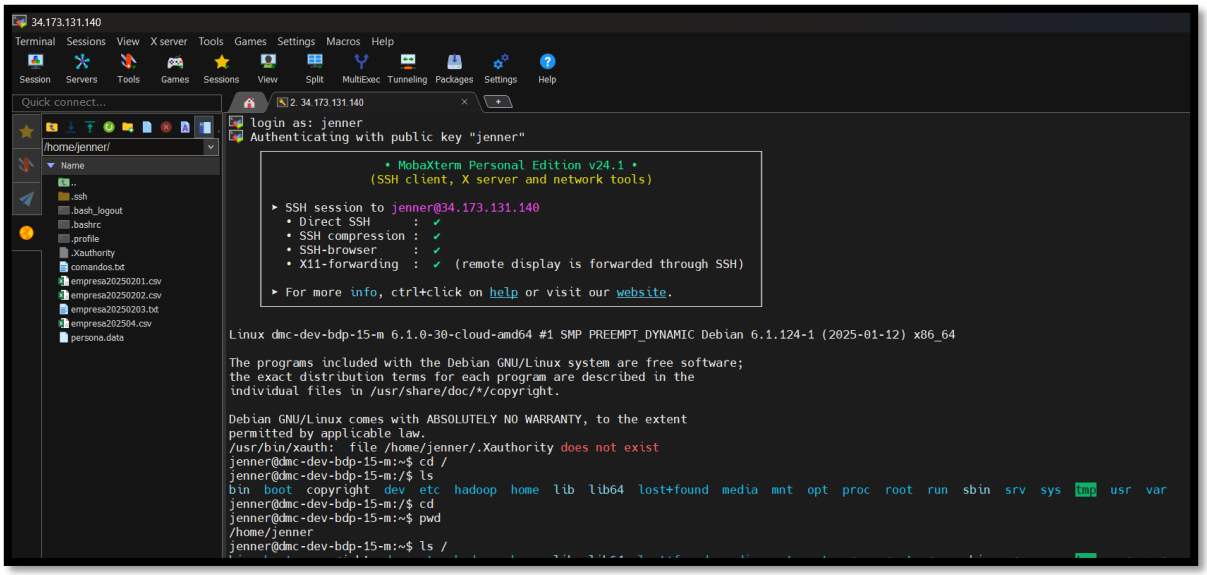


- Creación de Llave pública.



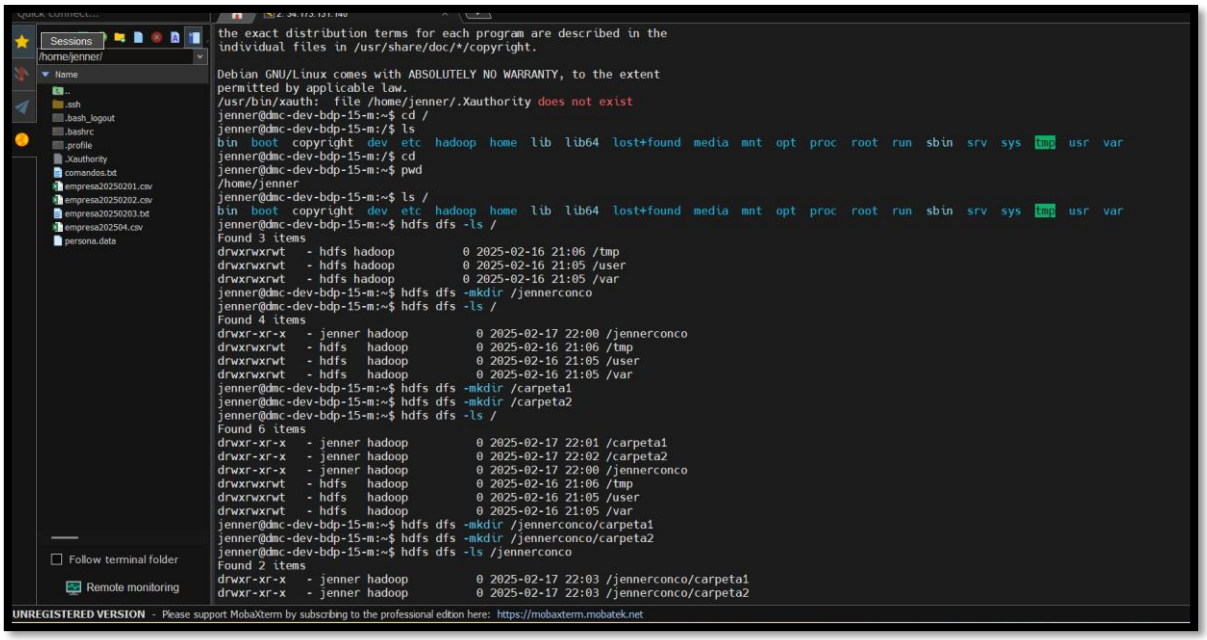


- Se logró realizar la conexión con moba.

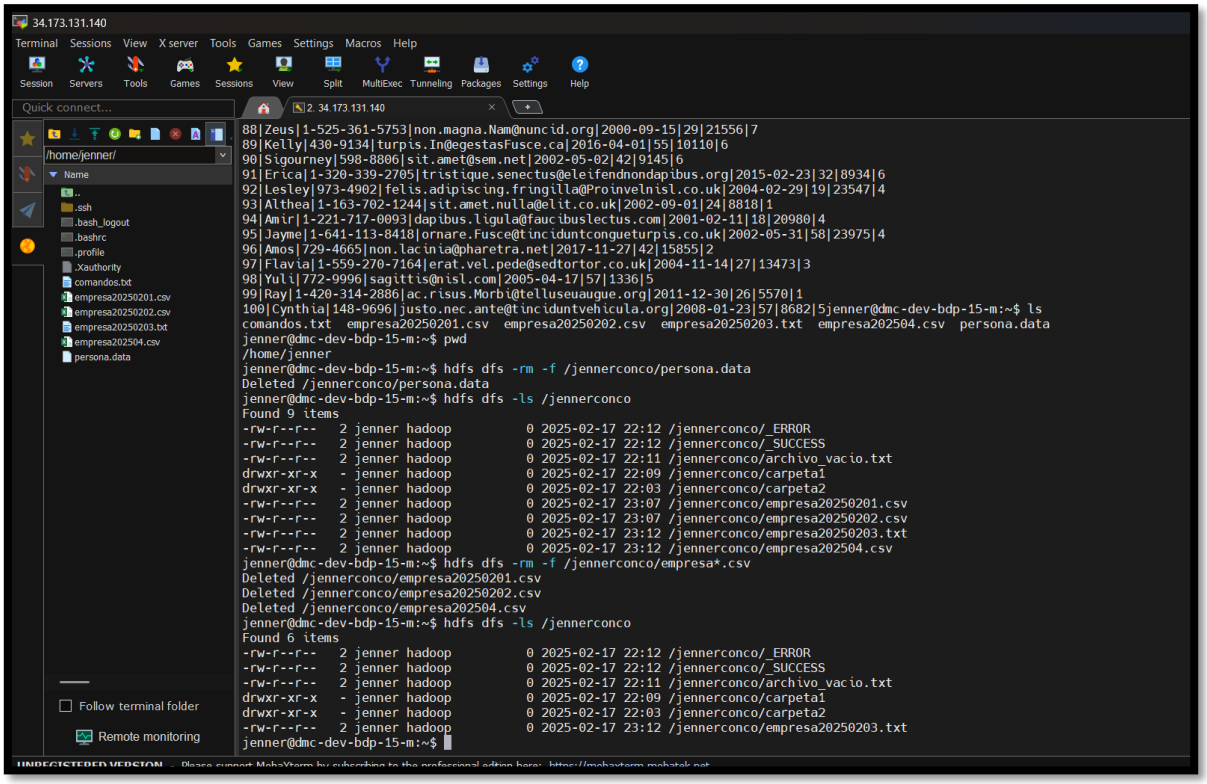


SESIÓN 03

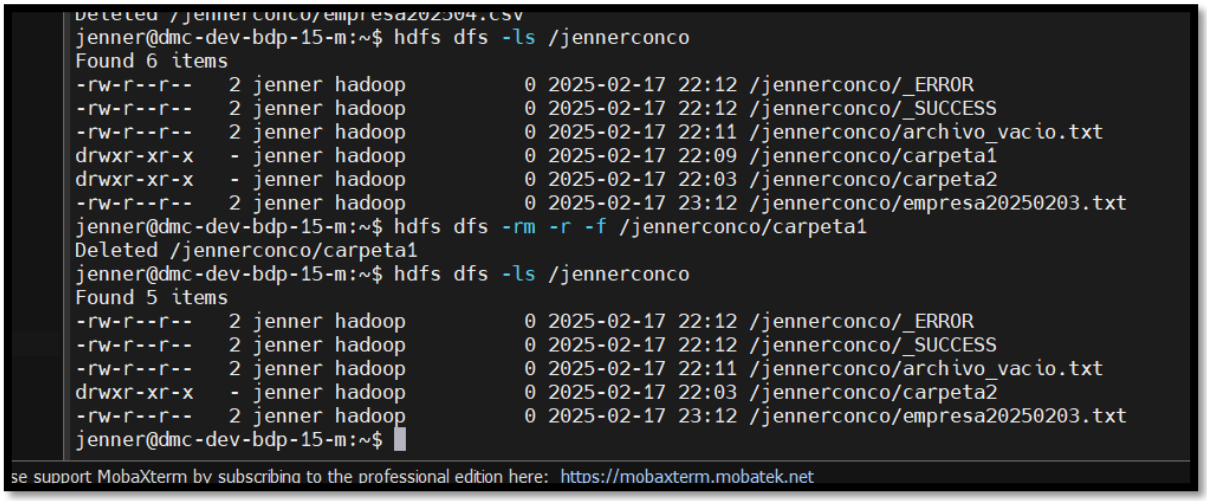
- Se logró la creación de carpeta jennerconco en moba.



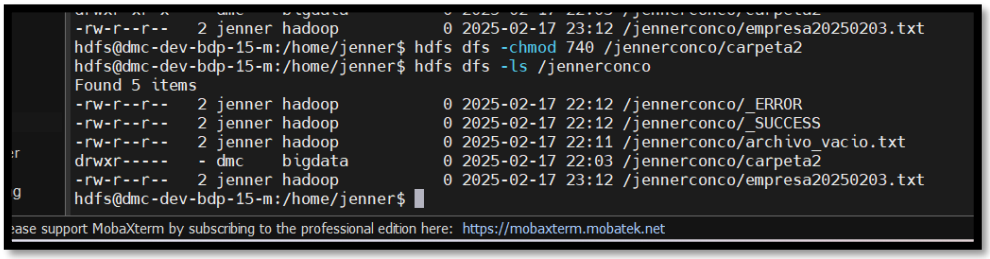
- Se realizó la eliminación de archivos.



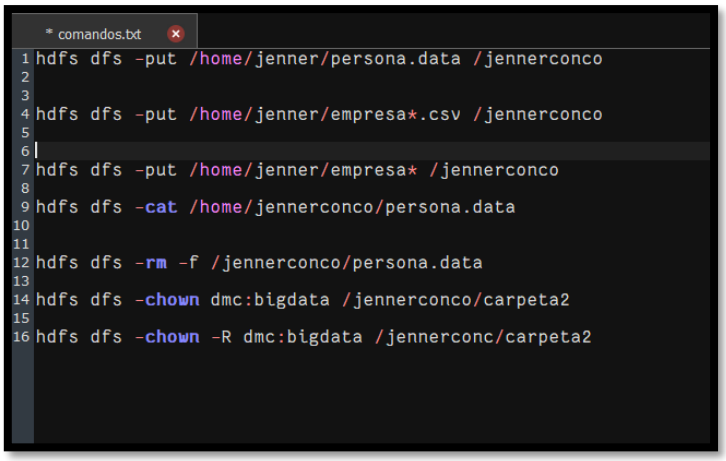
- Eliminando carpeta1



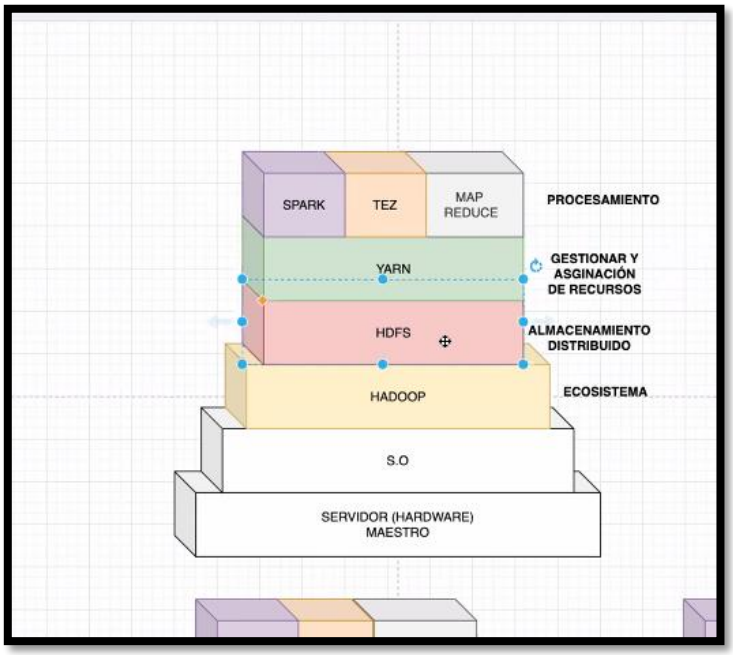
- Visualizando los permisos.



- Comandos utilizados, en carpeta comandos.txt

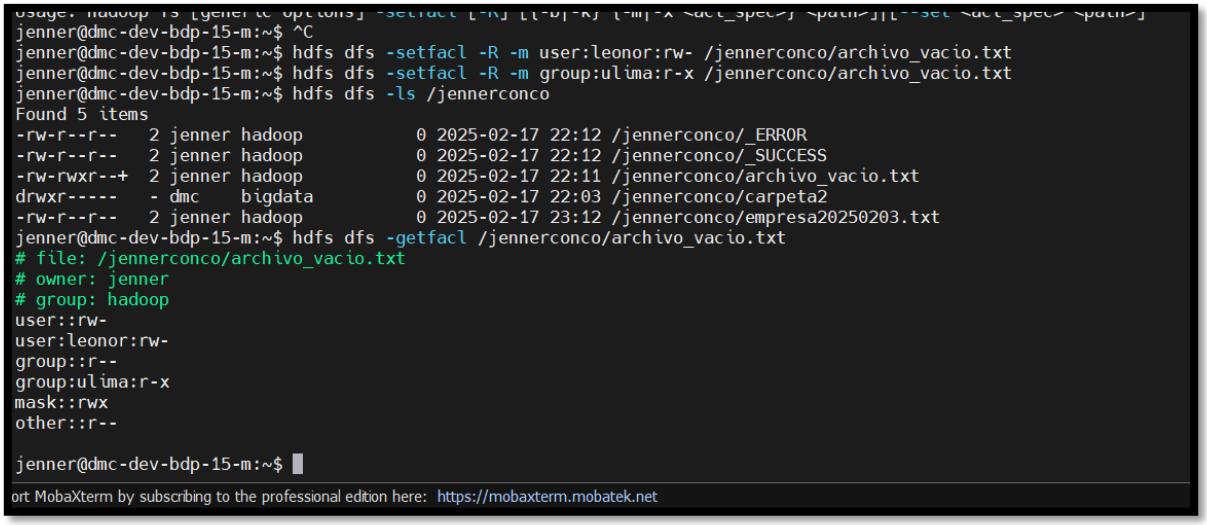


- Arquitectura a utilizar.

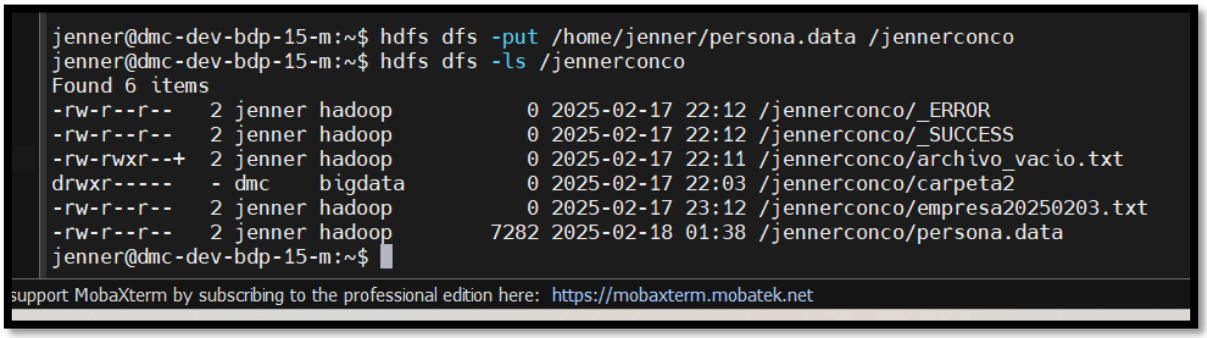


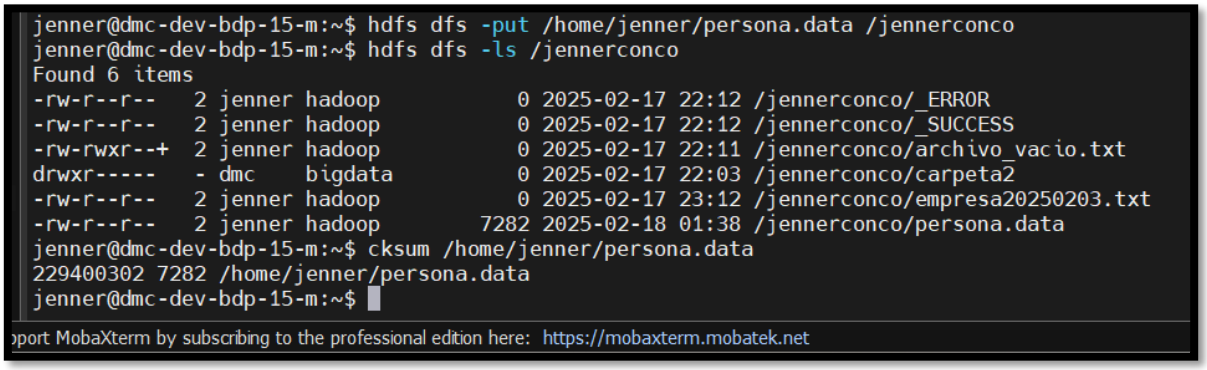
SESIÓN 04

- Asignacion de usuario y grupo.



- Subiendo archivo persona.data.



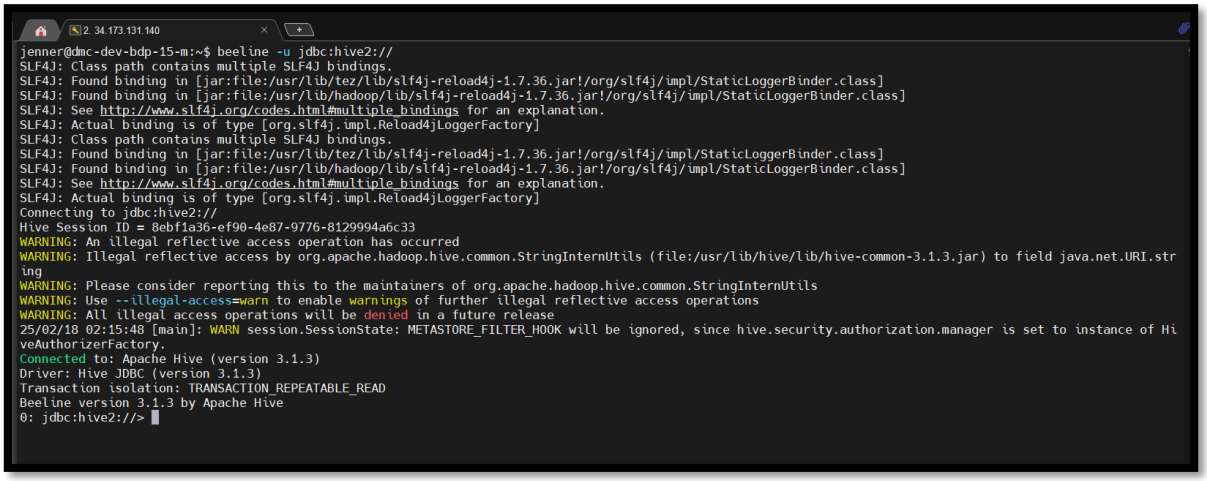


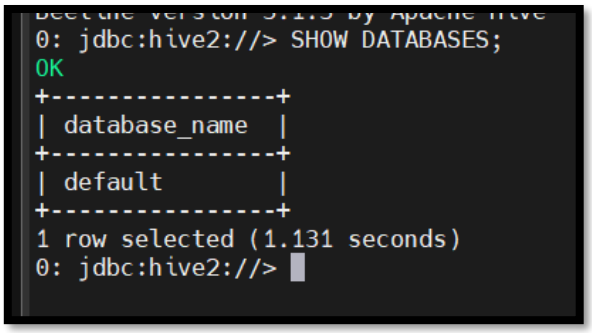
```
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -put /home/jenner/persona.data /jennerconco
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /jennerconco
Found 6 items
-rw-r--r--   2 jenner hadoop          0 2025-02-17 22:12 /jennerconco/_ERROR
-rw-r--r--   2 jenner hadoop          0 2025-02-17 22:12 /jennerconco/_SUCCESS
-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:~$ 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:~$
pport MobaXterm by subscribing to the professional edition here:  https://mobaxterm.mobatek.net
```

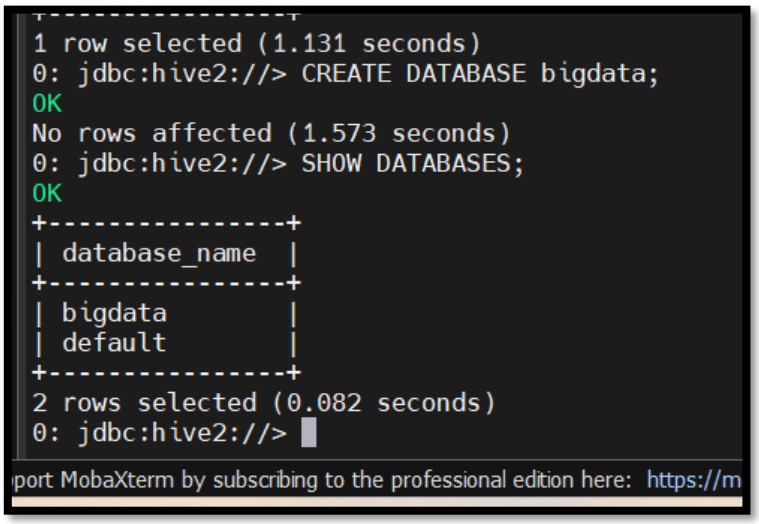
```
0      0       /var
jenner@dmc-dev-bdp-15-m:~$ hdfs dfs -setrep -w 3 -R /jennerconco
setrep: `-R': No such file or directory
Replication 3 set: /jennerconco/_ERROR
Replication 3 set: /jennerconco/_SUCCESS
Replication 3 set: /jennerconco/archivo_vacio.txt
Replication 3 set: /jennerconco/empresa20250203.txt
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:~$
upport MobaXterm by subscribing to the professional edition here:  https://mobaxterm.mobatek.net
```

- Comandos para HIVE, realizando la conexión a HIVE.

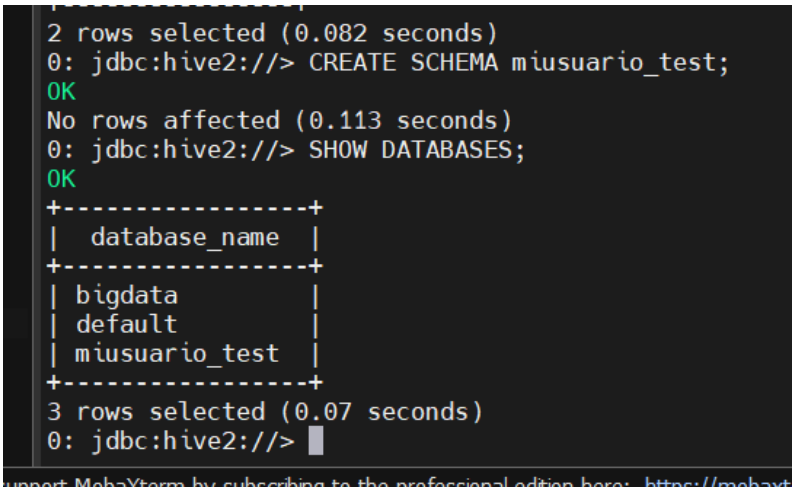




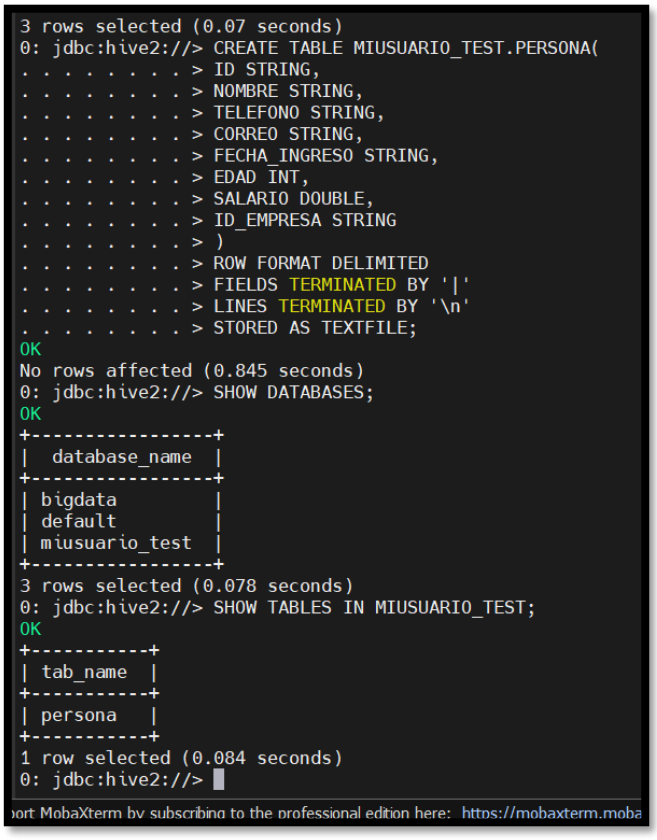
- Realizando consulta show databases.



- Resultado de la consulta.



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



- 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:~$
```

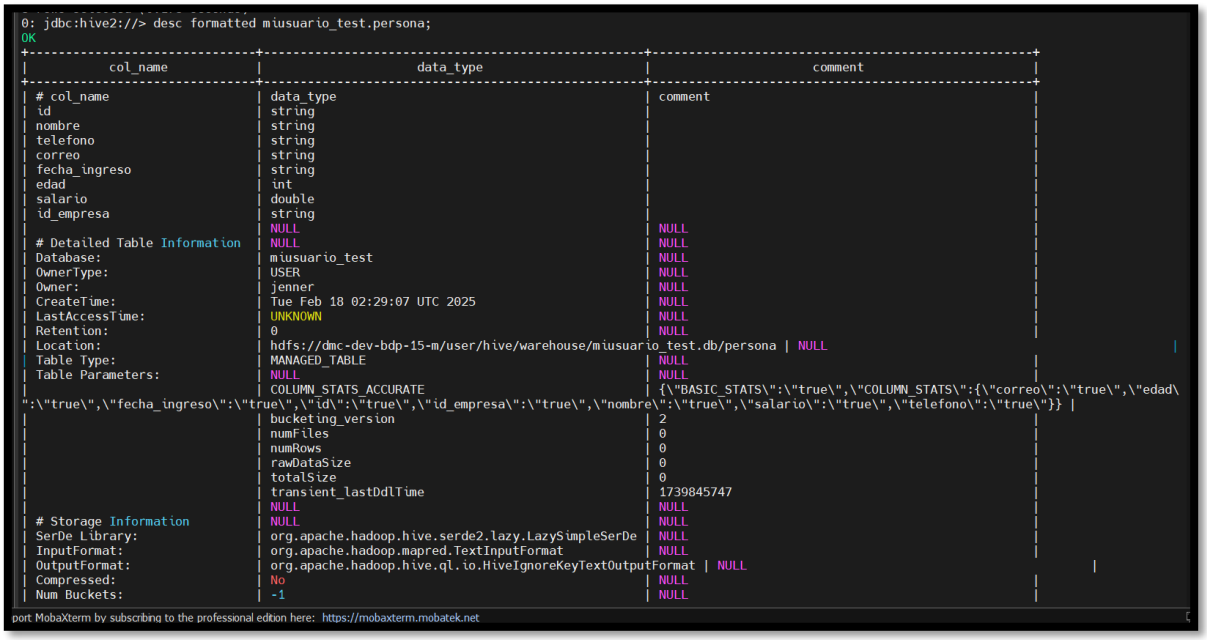
- Migrando persona.data

```
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:~$ ^C
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:~$
```

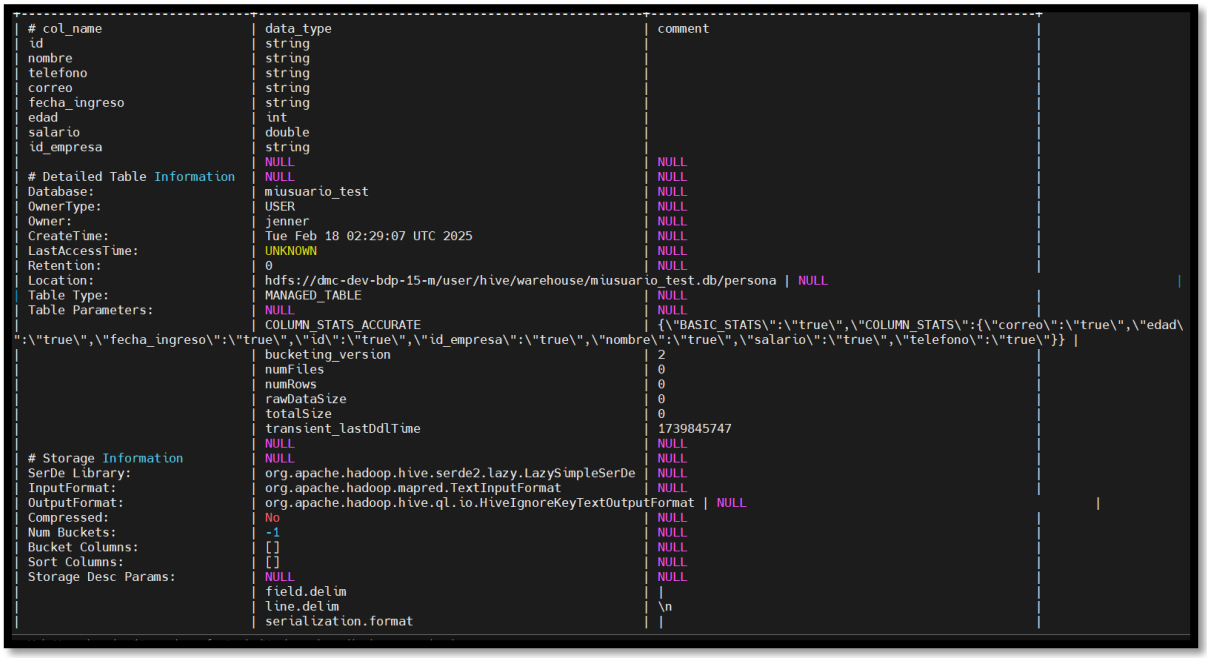
- Visualizando los tipos de datos de los campos de la tabla persona.

```
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.

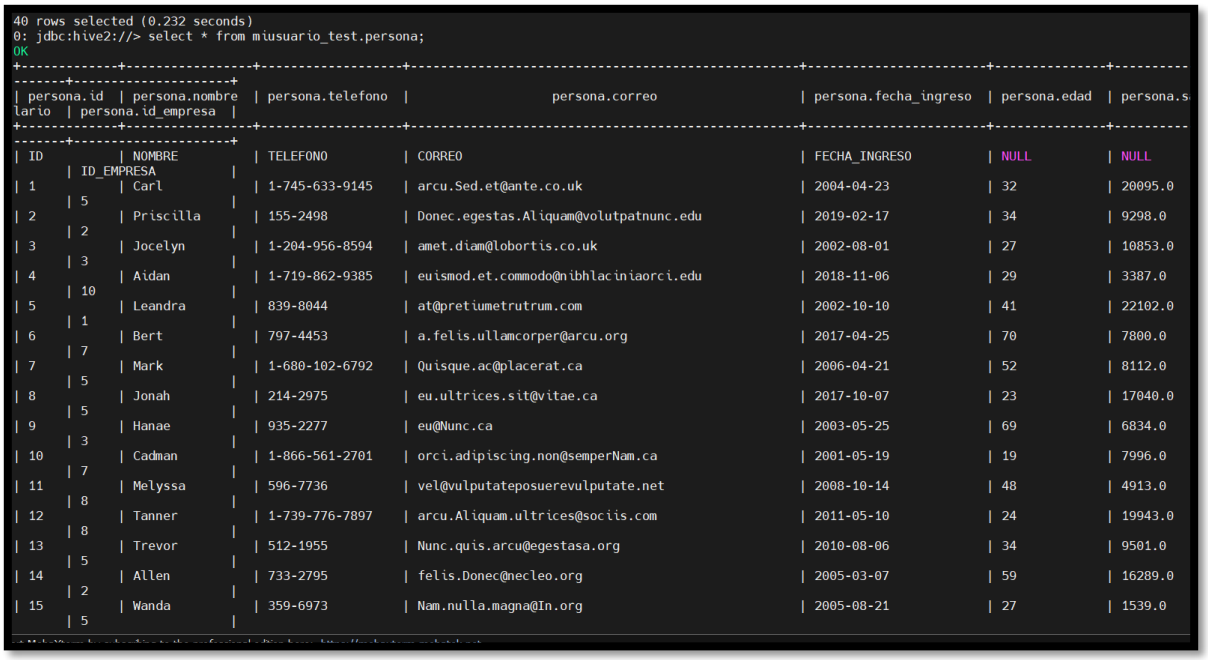


```
0: jdbc:hive2://> desc formatted miusuario_test.persona;
OK
+-------------------------------+----------------------------------------------------+----------------------------------------------------+
|           col_name            |                     data_type                      |                      comment                       |
+-------------------------------+----------------------------------------------------+----------------------------------------------------+
| # 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                                               |
| # Detailed Table Information  | NULL                                               | NULL                                               |
| Database:                     | miusuario_test                                     | NULL                                               |
| OwnerType:                    | USER                                               | NULL                                               |
| Owner:                        | jenner                                             | NULL                                               |
| CreateTime:                   | Tue Feb 18 02:29:07 UTC 2025                       | NULL                                               |
| LastAccessTime:               | UNKNOWN                                            | NULL                                               |
| Retention:                    | 0                                                  | NULL                                               |
| Location:                     | hdfs://dmc-dev-bdp-15-m/user/hive/warehouse/miusuario_test.db/persona | NULL                                               |
| Table Type:                   | MANAGED_TABLE                                      | NULL                                               |
| Table Parameters:             | NULL                                               | 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                                               |
| # Storage Information         | NULL                                               | NULL                                               |
| SerDe Library:                | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe | NULL                                               |
| InputFormat:                  | org.apache.hadoop.mapred.TextInputFormat           | NULL                                               |
| OutputFormat:                 | org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat | NULL                                               |
| Compressed:                   | No                                                 | NULL                                               |
| Num Buckets:                  | -1                                                 | NULL                                               |
port MobaXterm by subscribing to the professional edition here:  https://mobaxterm.mobatek.net
```



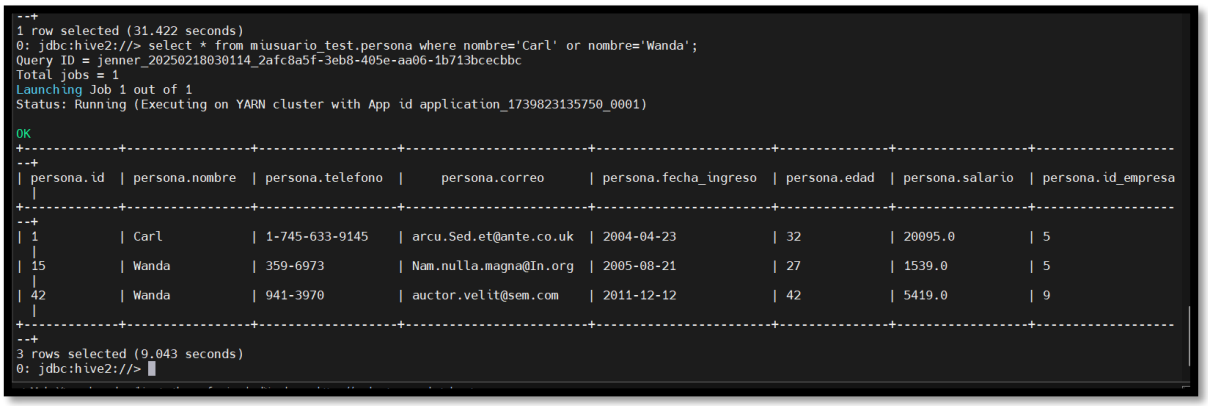
```
+-------------------------------+----------------------------------------------------+----------------------------------------------------+
| # col_name                    | data_type                                          | comment                                            |
| id                            | string                                             |                                                    |
| nombre                        | string                                             |                                                    |
| telefono                      | string                                             |                                                    |
| correo                        | string                                             |                                                    |
| fecha_ingreso                 | string                                             |                                                    |
| edad                          | int                                                |                                                    |
| salario                       | double                                             |                                                    |
| id_empresa                    | string                                             |                                                    |
|                               | NULL                                               | NULL                                               |
| # Detailed Table Information  | NULL                                               | NULL                                               |
| Database:                     | miusuario_test                                     | NULL                                               |
| OwnerType:                    | USER                                               | NULL                                               |
| Owner:                        | jenner                                             | NULL                                               |
| CreateTime:                   | Tue Feb 18 02:29:07 UTC 2025                       | NULL                                               |
| LastAccessTime:               | UNKNOWN                                            | NULL                                               |
| Retention:                    | 0                                                  | NULL                                               |
| Location:                     | hdfs://dmc-dev-bdp-15-m/user/hive/warehouse/miusuario_test.db/persona | NULL                                               |
| Table Type:                   | MANAGED_TABLE                                      | NULL                                               |
| Table Parameters:             | NULL                                               | 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                                               |
| # Storage Information         | NULL                                               | NULL                                               |
| SerDe Library:                | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe | NULL                                               |
| InputFormat:                  | org.apache.hadoop.mapred.TextInputFormat           | NULL                                               |
| OutputFormat:                 | org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat | NULL                                               |
| Compressed:                   | No                                                 | NULL                                               |
| Num Buckets:                  | -1                                                 | NULL                                               |
| Bucket Columns:               | []                                                 | NULL                                               |
| Sort Columns:                 | []                                                 | NULL                                               |
| Storage Desc Params:          | NULL                                               | NULL                                               |
|                               | field.delim                                        |                                                    |
|                               | line.delim                                         | \n                                                 |
|                               | serialization.format                               |                                                    |
```

- Realizando consulta select * from, para validar la ingesta 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
lario  | persona.id_empresa  |
+-------------+-----------------+-------------------+-----------------------------------------------+----------------------+---------------+-----------
-------+---------------------+
| ID          | NOMBRE          | TELEFONO          | CORREO                                        | FECHA_INGRESO          | NULL          | NULL
       | 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@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@nibhlaciniaorci.edu        | 2018-11-06             | 29            | 3387.0
       | 10                  |
| 5           | Leandra         | 839-8044          | 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          | Cadman          | 1-866-561-2701    | orci.adipiscing.non@semperNam.ca              | 2001-05-19             | 19            | 7996.0
       | 7                   |
| 11          | Melyssa         | 596-7736          | vel@vulputateposuerevulputate.net             | 2008-10-14             | 48            | 4913.0
       | 8                   |
| 12          | Tanner          | 1-739-776-7897    | arcu.Aliquam.ultrices@sociis.com              | 2011-05-10             | 24            | 19943.0
       | 8                   |
| 13          | Trevor          | 512-1955          | Nunc.quis.arcu@egestasa.org                   | 2010-08-06             | 34            | 9501.0
       | 5                   |
| 14          | Allen           | 733-2795          | felis.Donec@necleo.org                        | 2005-03-07             | 59            | 16289.0
       | 2                   |
| 15          | 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-aa06-1b713bcecbbc
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                                             |                                                    |
| nombre                        | string                                             |                                                    |
| telefono                      | string                                             |                                                    |
| correo                        | string                                             |                                                    |
| fecha_ingreso                 | string                                             |                                                    |
| edad                          | int                                                |                                                    |
| salario                       | double                                             |                                                    |
| id_empresa                    | string                                             |                                                    |
|                               | NULL                                               | NULL                                               |
| # Detailed Table Information  | NULL                                               | NULL                                               |
| Database:                     | miusuario_test2                                    | NULL                                               |
| OwnerType:                    | USER                                               | NULL                                               |
| Owner:                        | jenner                                             | NULL                                               |
| CreateTime:                   | Tue Feb 18 03:21:17 UTC 2025                       | NULL                                               |
| LastAccessTime:               | UNKNOWN                                            | NULL                                               |
| Retention:                    | 0                                                  | NULL                                               |
| Location:                     | hdfs://dmc-dev-bdp-15-m/user/miusuario/bd/miusuario_test2/persona | NULL                                               |
| Table Type:                   | MANAGED_TABLE                                      | NULL                                               |
| Table Parameters:             | NULL                                               | 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                              | 1739848877                                         |
|                               | NULL                                               | NULL                                               |
| # Storage Information         | NULL                                               | NULL                                               |
| SerDe Library:                | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe | NULL                                               |
| InputFormat:                  | org.apache.hadoop.mapred.TextInputFormat           | NULL                                               |
| OutputFormat:                 | org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat | NULL                                               |
| Compressed:                   | No                                                 | NULL                                               |
| Num Buckets:                  | -1                                                 | NULL                                               |
```

```
+-------------------------------+----------------------------------------------------+----------------------------------------------------+
| # 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                                               |
| # Detailed Table Information  | NULL                                               | NULL                                               |
| Database:                     | miusuario_test2                                    | NULL                                               |
| OwnerType:                    | USER                                               | NULL                                               |
| Owner:                        | jenner                                             | NULL                                               |
| CreateTime:                   | Tue Feb 18 03:21:17 UTC 2025                       | NULL                                               |
| LastAccessTime:               | UNKNOWN                                            | NULL                                               |
| Retention:                    | 0                                                  | NULL                                               |
| Location:                     | hdfs://dmc-dev-bdp-15-m/user/miusuario/bd/miusuario_test2/persona | NULL                                               |
| Table Type:                   | MANAGED_TABLE                                      | NULL                                               |
| Table Parameters:             | NULL                                               | 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                              | 1739848877                                         |
|                               | NULL                                               | NULL                                               |
| # Storage Information         | NULL                                               | NULL                                               |
| SerDe Library:                | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe | NULL                                               |
| InputFormat:                  | org.apache.hadoop.mapred.TextInputFormat           | NULL                                               |
| OutputFormat:                 | org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat | NULL                                               |
| Compressed:                   | No                                                 | NULL                                               |
| Num Buckets:                  | -1                                                 | NULL                                               |
| Bucket Columns:               | []                                                 | NULL                                               |
| Sort Columns:                 | []                                                 | NULL                                               |
| Storage Desc Params:          | NULL                                               | NULL                                               |
|                               | field.delim                                        |                                                    |
|                               | line.delim                                         | \n                                                 |
|                               | serialization.format                               |                                                    |
ort MobaXterm by subscribing to the professional edition here:  https://mobaxterm.mobatek.net
```

- Validando cantidad de registros.

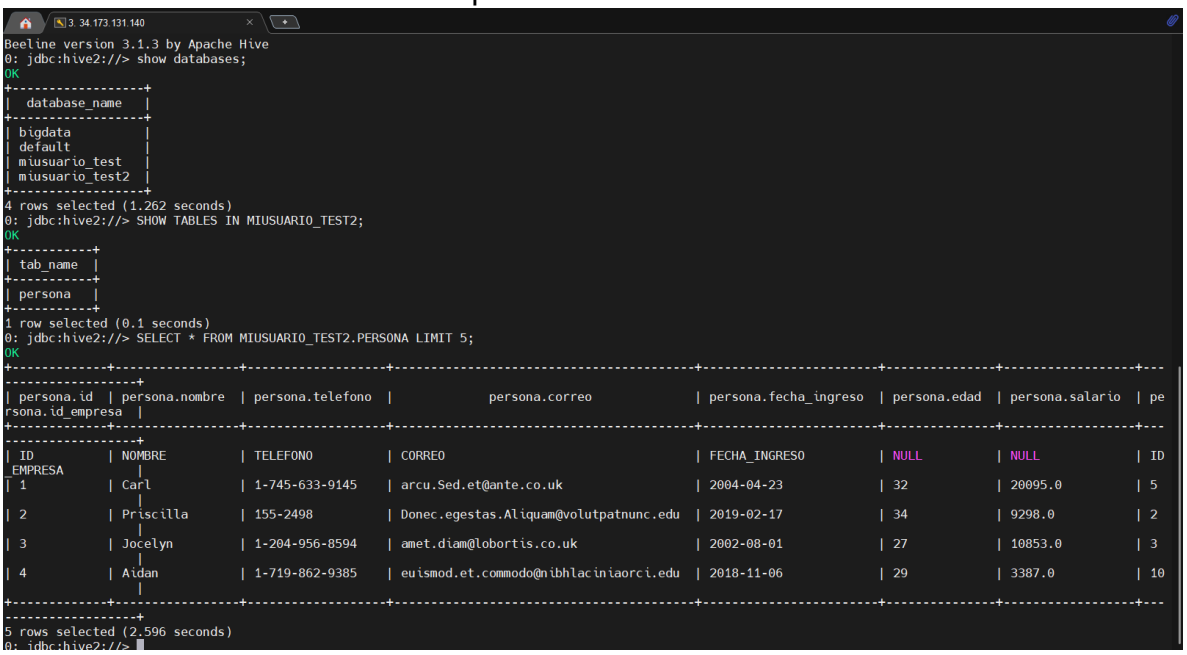
```
40 rows selected (0.151 seconds)
0: jdbc:hive2://>  SELECT * FROM MIUSUARIO_TEST2.PERSONA LIMIT 10;
OK
+-------------+-----------------+-------------------+-----------------+-----------------------+---------------+-----------------+----------------------+
| persona.id  | persona.nombre  | persona.telefono  | 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;
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  | pe
rsona.id_empresa  |
+-------------+-----------------+-------------------+-----------------------------------------+-----------------------+---------------+-----------------+---
-----------------+
| ID          | NOMBRE          | TELEFONO          | CORREO                                  | FECHA_INGRESO         | NULL          | NULL             | 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@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@nibhlaciniaorci.edu  | 2018-11-06            | 29            | 3387.0           | 10
                  |
| 5           | Leandra         | 839-8044          | 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://>
```

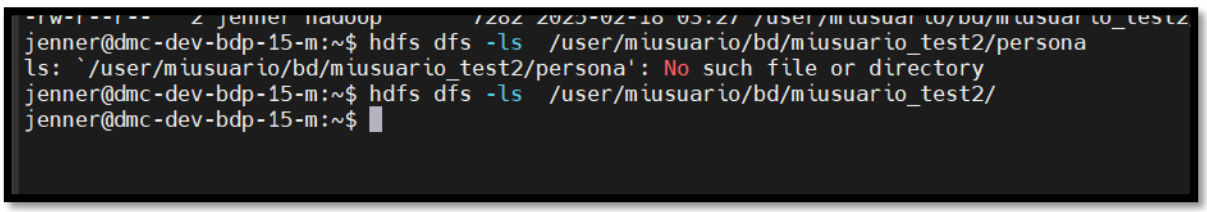
SESIÓN 05

- Validando existencia de mi tabla persona.



- Eliminación mediante drop table de la tabla persona.

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



- Creando tabla external .

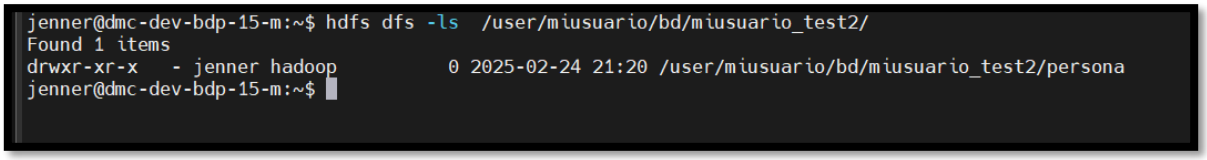
```
0: jdbc:hive2://> CREATE EXTERNAL TABLE MIUSUARIO_TEST2.PERSONA(
. . . . . > ID STRING,
. . . . . > NOMBRE STRING,
. . . . . > TELEFONO STRING,
. . . . . > CORREO STRING,
. . . . . > FECHA_INGRESO STRING,
. . . . . > EDAD INT,
. . . . . > SALARIO DOUBLE,
. . . . . > ID_EMPRESA STRING
. . . . . > )
. . . . . > 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.

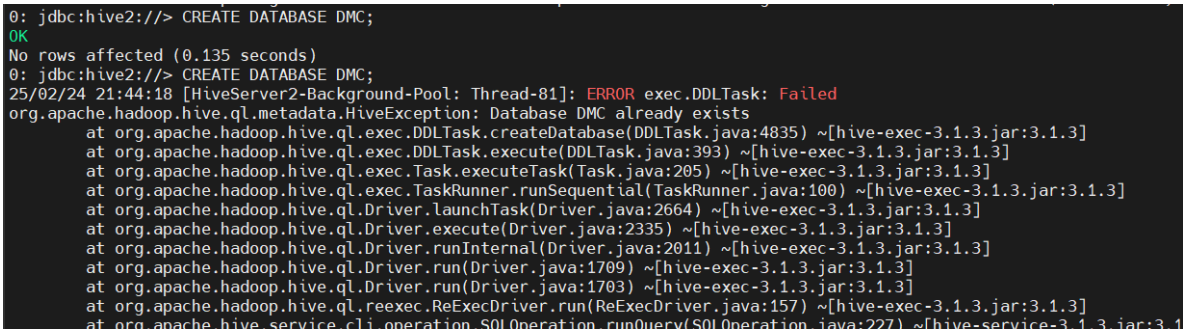
```
0: jdbc:hive2://>   SELECT * FROM MIUSUARIO_TEST2.PERSONA LIMIT 10;
OK
+-------------+-----------------+-------------------+----------------------------------------+-----------------------+---------------+-------------------+---
-----------------+
| persona.id  | persona.nombre  | persona.telefono  |             persona.correo             | persona.fecha_ingreso | persona.edad  | persona.salario   | pe
rsona.id_empresa  |
+-------------+-----------------+-------------------+----------------------------------------+-----------------------+---------------+-------------------+---
-----------------+
| ID          | NOMBRE          | TELEFONO          | CORREO                                 | FECHA_INGRESO         | NULL          | NULL              | 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@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@nibhlaciniaorci.edu | 2018-11-06            | 29            | 3387.0            | 10
|
| 5           | Leandra         | 839-8044          | 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 (0.444 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:~$
```

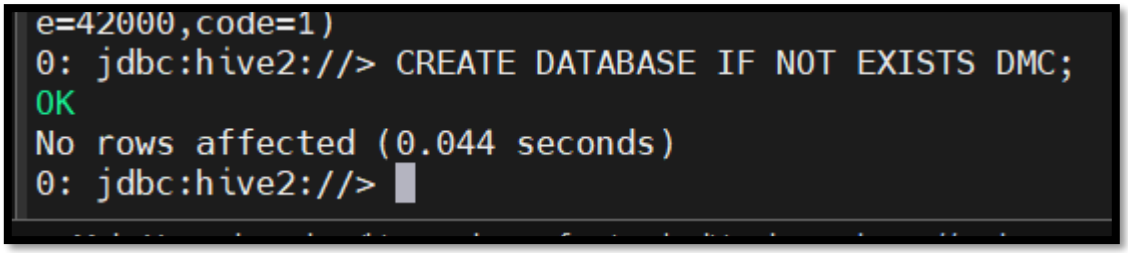
- Se elimina la tabla en hive, pero en Linux aún existe.



- Sin validador IF NOT EXISTS.

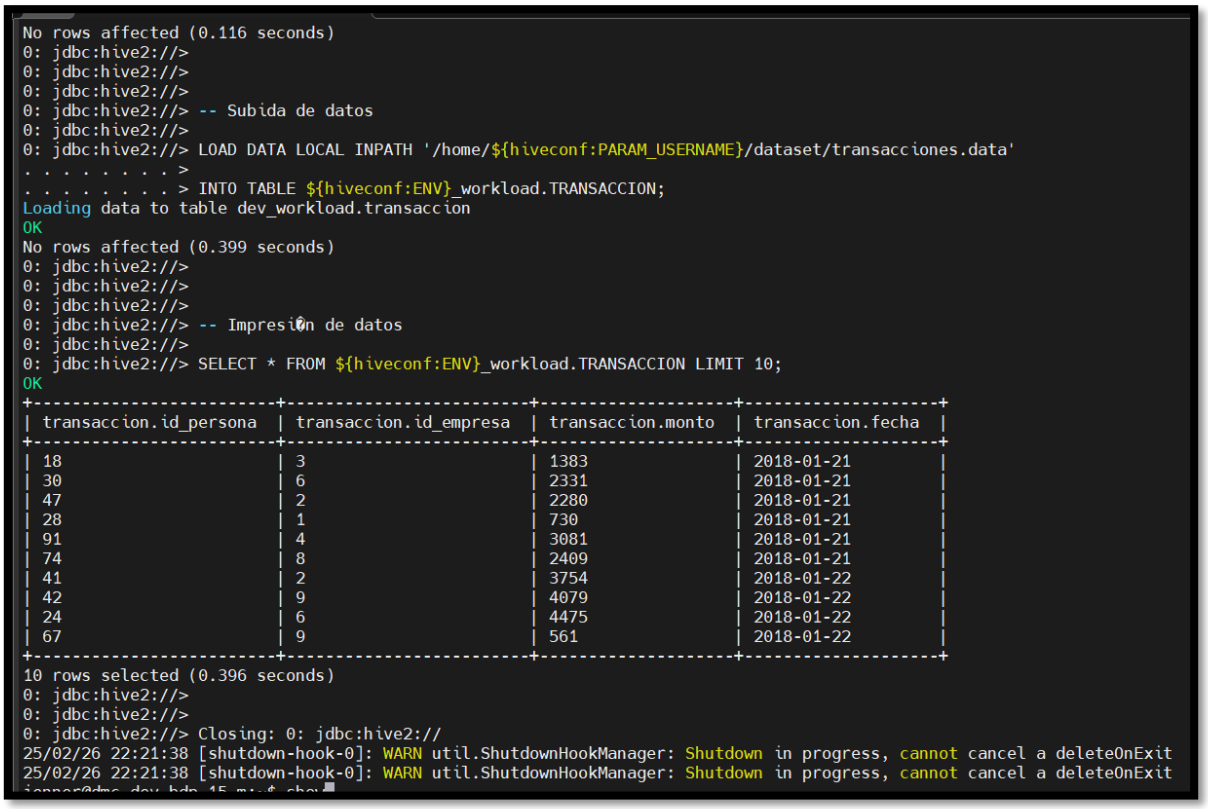


- Con validador IF NOT EXISTS.



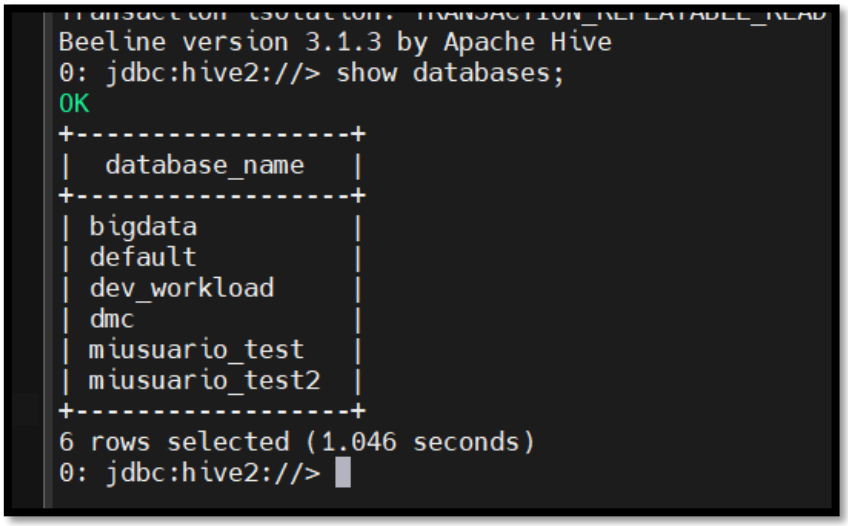
SESIÓN 06

- Cargando el archivo Poblando_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'
. . . . . >
. . . . . > 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://> -- Impresi?n 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
```

- 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://> -- 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://>
```

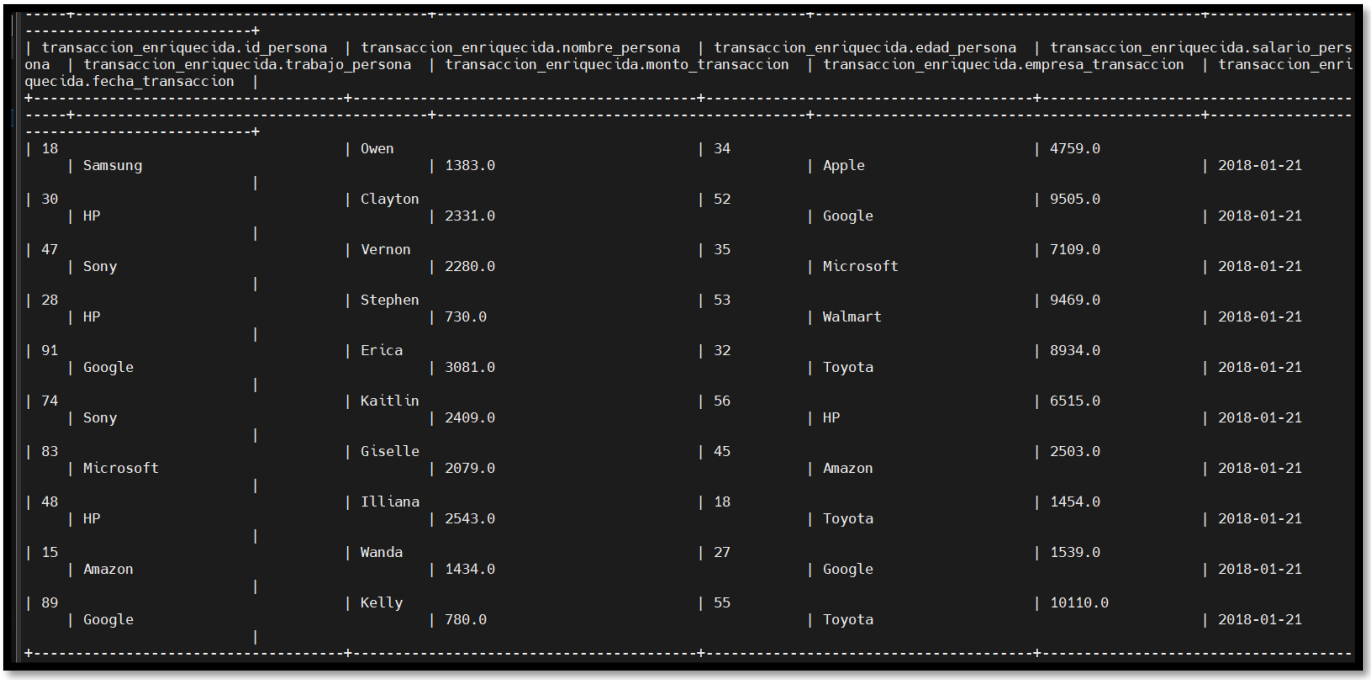
- Creación de capa CURATED.

```
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 | pe
rsona.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@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@nibhlaciniaorci.edu | 2018-11-06          | 29            | 3387.0          | 10
                  |
| 5           | Leandra         | 839-8044          | 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          | Cadman          | 1-866-561-2701    | orci.adipiscing.non@semperNam.ca     | 2001-05-19            | 19            | 7996.0          | 7
                  |
+-------------+-----------------+-------------------+--------------------------------------+-----------------------+---------------+-----------------+---
```

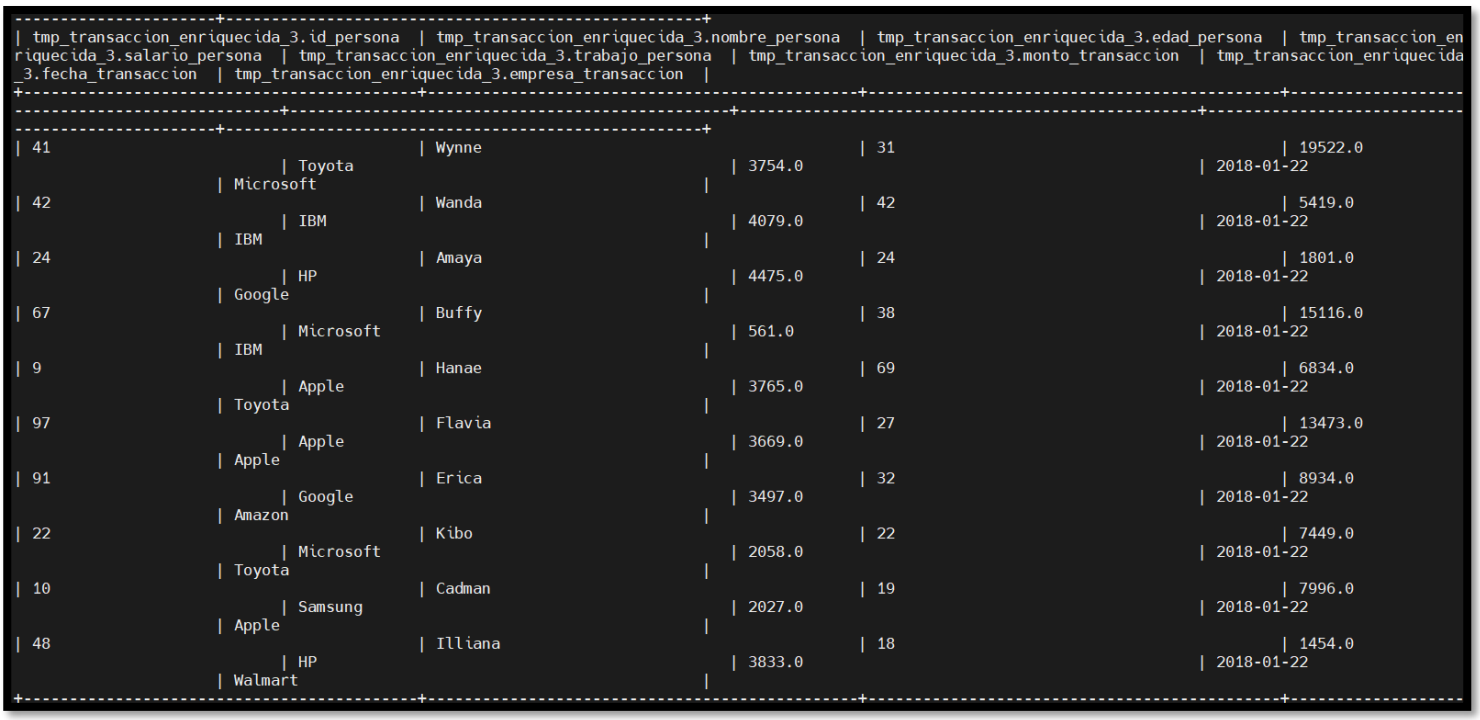
```
0: jdbc:hive2://>
0: jdbc:hive2://> SELECT * FROM ${hiveconf:ENV}_curated.EMPRESA LIMIT 10;
OK
+-------------+-----------------+
| empresa.id  | empresa.nombre  |
+-------------+-----------------+
| 1           | Walmart         |
| 2           | Microsoft       |
| 3           | Apple           |
| 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
```

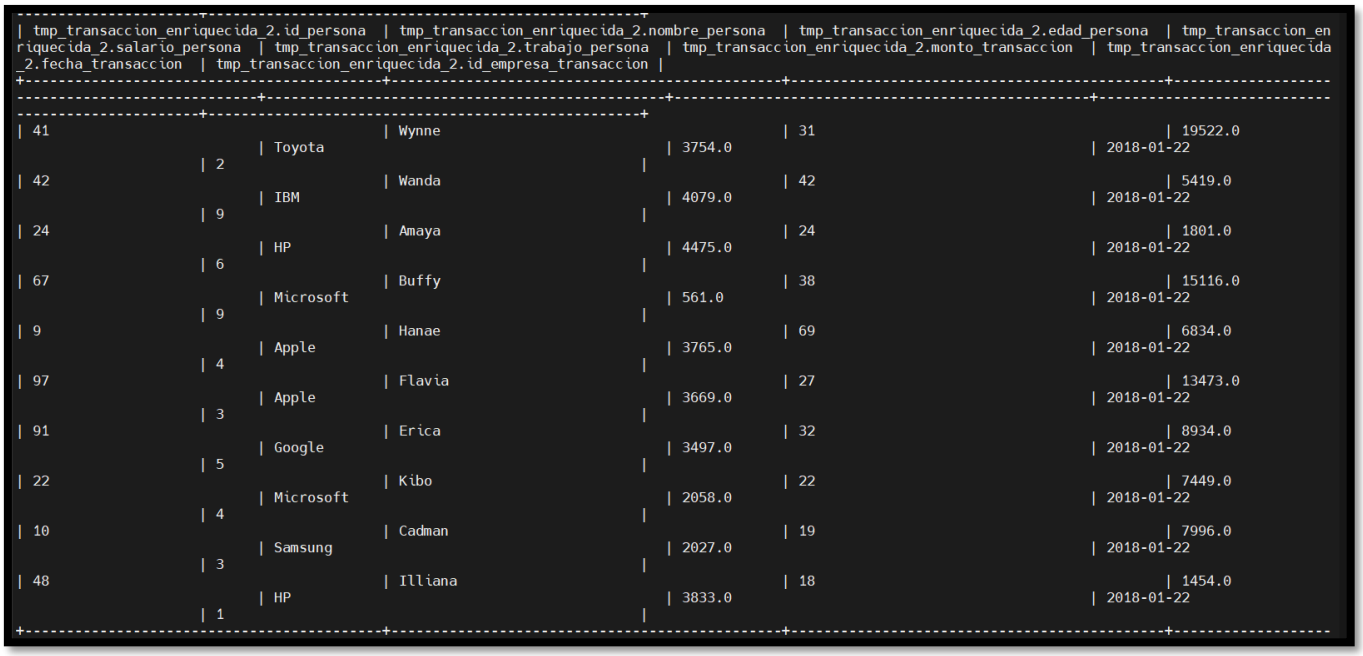
- Ejecucion de capa Funcional.



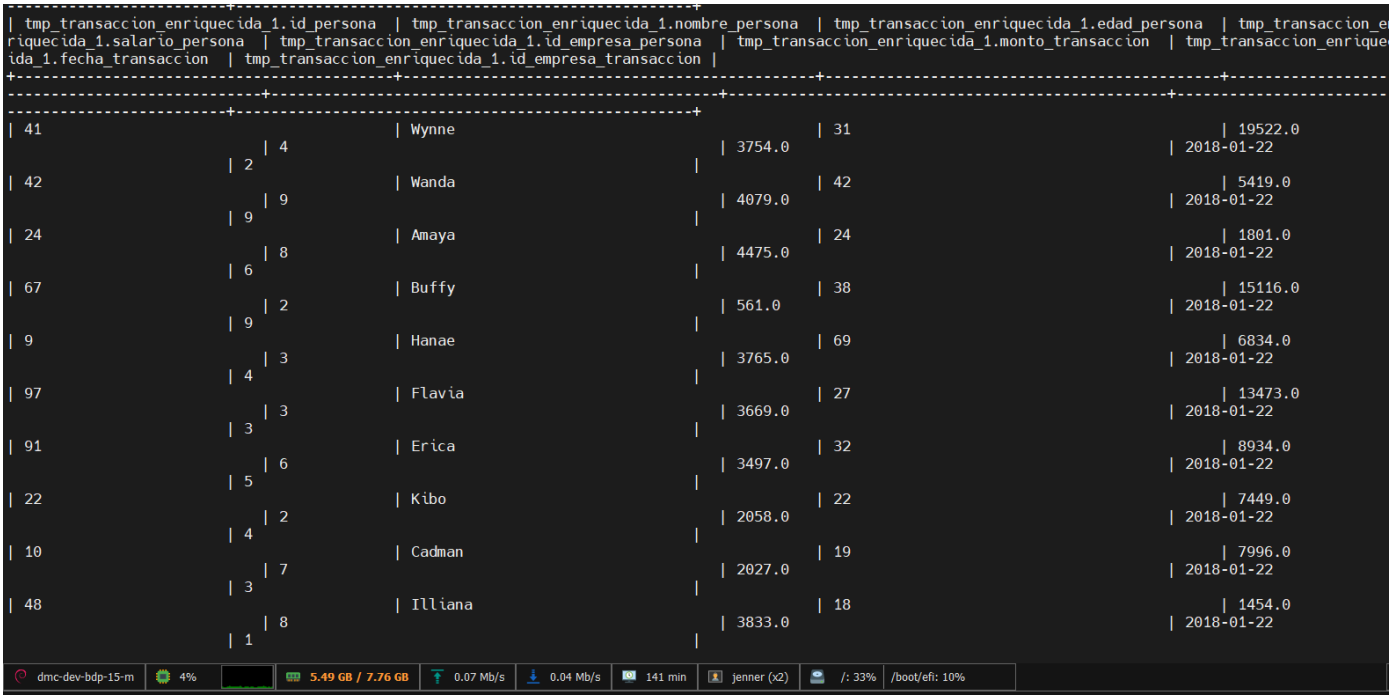
transaccion_enriquecida.id_persona	transaccion_enriquecida.nombre_persona	transaccion_enriquecida.edad_persona	transaccion_enriquecida.salario_persona	transaccion_enriquecida.trabajo_persona	transaccion_enriquecida.monto_transaccion	transaccion_enriquecida.empresa_transaccion	transaccion_enriquecida.fecha_transaccion
18	Owen	34	Apple	4759.0	Samsung	1383.0	2018-01-21
30	Clayton	52	Google	9505.0	HP	2331.0	2018-01-21
47	Vernon	35	Microsoft	7109.0	Sony	2280.0	2018-01-21
28	Stephen	53	Walmart	9469.0	HP	730.0	2018-01-21
91	Erica	32	Toyota	8934.0	Google	3081.0	2018-01-21
74	Kaitlin	56	HP	6515.0	Sony	2409.0	2018-01-21
83	Giselle	45	Amazon	2503.0	Microsoft	2079.0	2018-01-21
48	Illiana	18	Toyota	1454.0	HP	2543.0	2018-01-21
15	Wanda	27	Google	1539.0	Amazon	1434.0	2018-01-21
89	Kelly	55	Toyota	10110.0	Google	780.0	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.fecha_transaccion	tmp_transaccion_enriquecida_3.empresa_transaccion
41	Wynne	31	19522.0	Toyota	3754.0	2018-01-22	Microsoft
42	Wanda	42	5419.0	IBM	4079.0	2018-01-22	IBM
24	Amaya	24	1801.0	HP	4475.0	2018-01-22	Google
67	Buffy	38	15116.0	Microsoft	561.0	2018-01-22	IBM
9	Hanae	69	6834.0	Apple	3765.0	2018-01-22	Toyota
97	Flavia	27	13473.0	Apple	3669.0	2018-01-22	Apple
91	Erica	32	8934.0	Google	3497.0	2018-01-22	Amazon
22	Kibo	22	7449.0	Microsoft	2058.0	2018-01-22	Toyota
10	Cadman	19	7996.0	Samsung	2027.0	2018-01-22	Apple
48	Illiana	18	1454.0	HP	3833.0	2018-01-22	Walmart



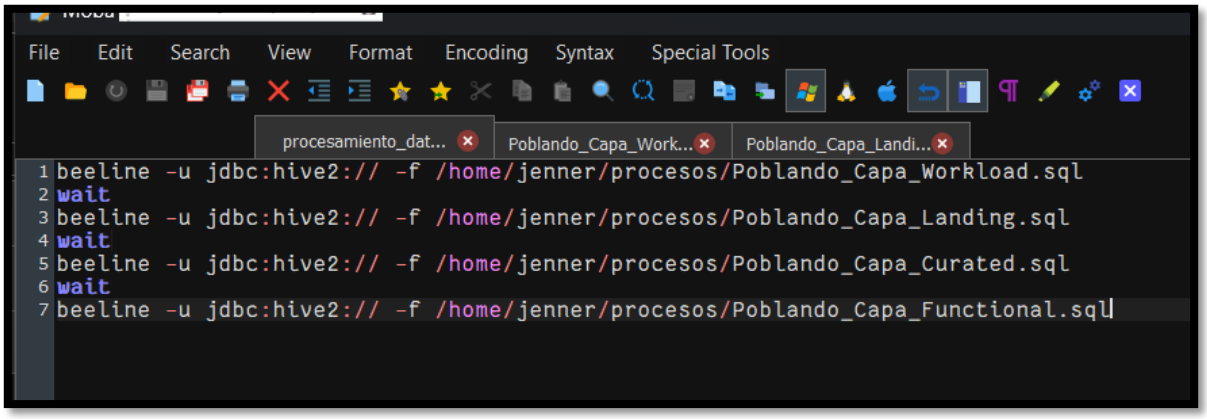
tmp_transaccion_enriquecida_2.id_persona	tmp_transaccion_enriquecida_2.nombre_persona	tmp_transaccion_enriquecida_2.edad_persona	tmp_transaccion_enriquecida_2.salario_persona	tmp_transaccion_enriquecida_2.trabajo_persona	tmp_transaccion_enriquecida_2.monto_transaccion	tmp_transaccion_enriquecida_2.fecha_transaccion	tmp_transaccion_enriquecida_2.id_empresa_transaccion
41	Wynne	31	19522.0	Toyota	3754.0	2018-01-22	2
42	Wanda	42	5419.0	IBM	4079.0	2018-01-22	9
24	Amaya	24	1801.0	HP	4475.0	2018-01-22	6
67	Buffy	38	15116.0	Microsoft	561.0	2018-01-22	9
9	Hanae	69	6834.0	Apple	3765.0	2018-01-22	4
97	Flavia	27	13473.0	Apple	3669.0	2018-01-22	3
91	Erica	32	8934.0	Google	3497.0	2018-01-22	5
22	Kibo	22	7449.0	Microsoft	2058.0	2018-01-22	4
10	Cadman	19	7996.0	Samsung	2027.0	2018-01-22	3
48	Illiana	18	1454.0	HP	3833.0	2018-01-22	1



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	Wynne	31	19522.0	4	3754.0	2018-01-22	2
42	Wanda	42	5419.0	9	4079.0	2018-01-22	9
24	Amaya	24	1801.0	8	4475.0	2018-01-22	6
67	Buffy	38	15116.0	2	561.0	2018-01-22	9
9	Hanae	69	6834.0	3	3765.0	2018-01-22	4
97	Flavia	27	13473.0	3	3669.0	2018-01-22	3
91	Erica	32	8934.0	6	3497.0	2018-01-22	5
22	Kibo	22	7449.0	2	2058.0	2018-01-22	4
10	Cadman	19	7996.0	7	2027.0	2018-01-22	3
48	Illiana	18	1454.0	8	3833.0	2018-01-22	1

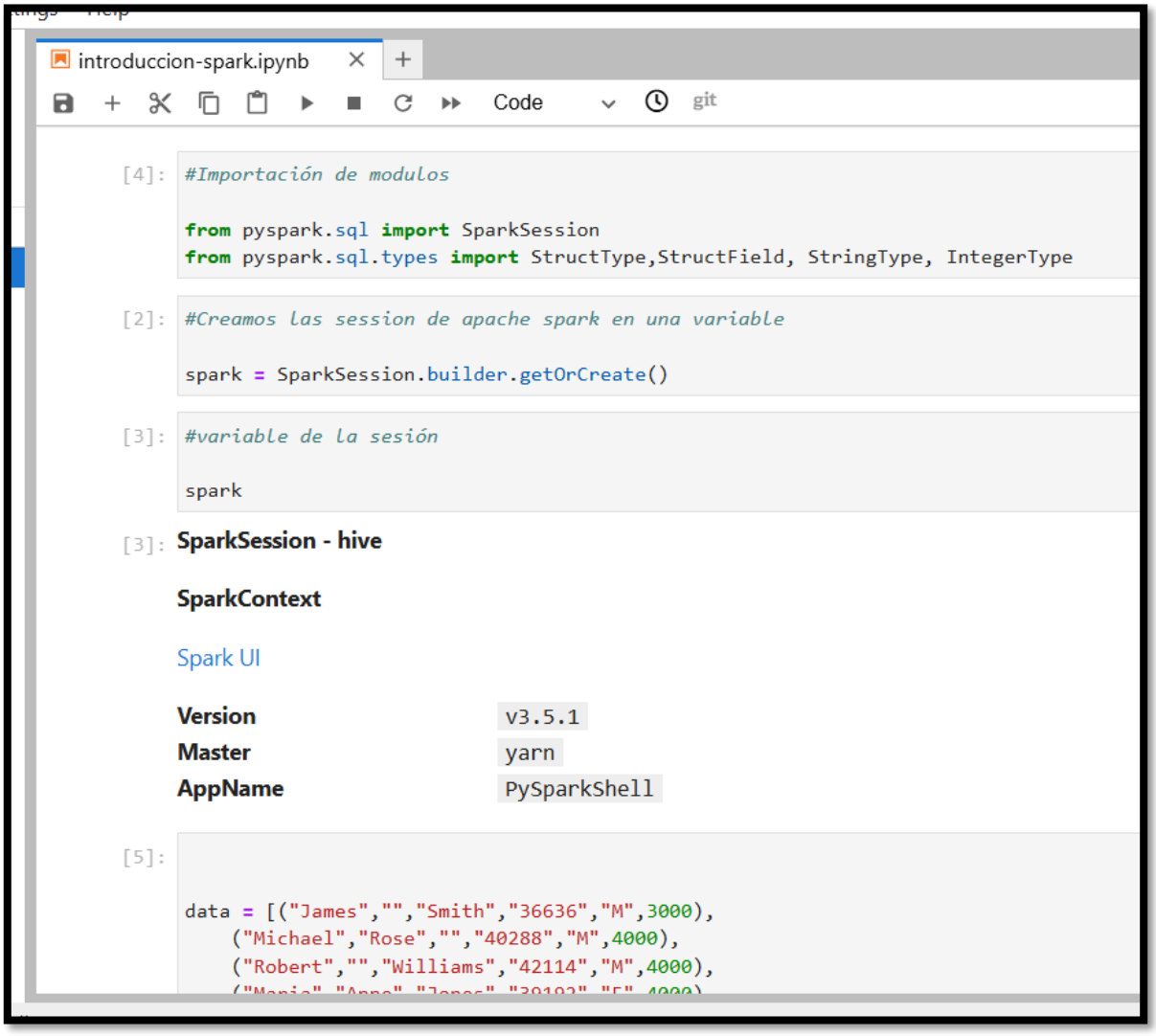
dmc-dev-bdp-15-m | 4% | 5.49 GB / 7.76 GB | 0.07 Mb/s | 0.04 Mb/s | 141 min | jenner (x2) | /: 33% | /boot/efi: 10%

- Archivo procesamiento_datalake.sh



SESIÓN 07

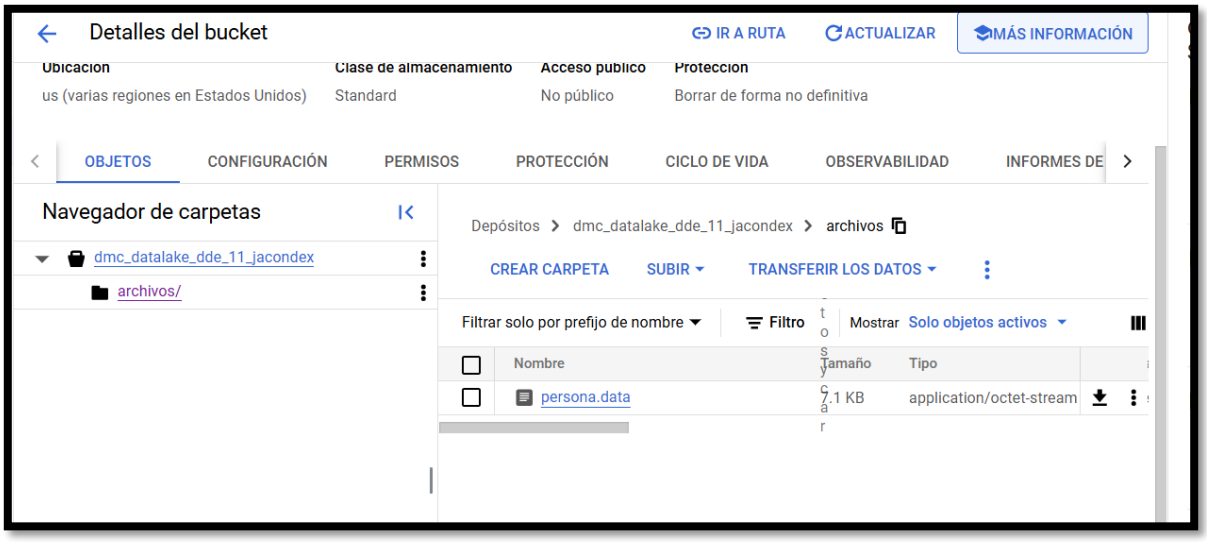
- Codigo pyspark en jupyterlab.



```
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.



- 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)
```

```
[Stage 6:>                                                          (0 + 1) / 1]
+---+---------+--------------+--------------------+-------------+----+-------+----------+
| 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@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|
```

- 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@pretiumetrutru...|   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

.th_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| 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|
```

- 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|euismod.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.

```
LISTA_EMPRESA = [2,3,5]
```

```
df_with_schema.filter(col("ID_EMPRESA").isin(LISTA_EMPRESA)).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|
|  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@turpis.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

```
df_with_schema.filter(~col("ID_EMPRESA").isin(LISTA_EMPRESA)).show()
```

```
+---+---------+--------------+--------------------+-------------+----+-------+----------+
| 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@pretiumetrutru...|   2002-10-10|  41|22102.0|         1|
|  6|     Bert|      797-4453|a.felis.ullamcorp...|   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|
+---+---------+--------------+--------------------+-------------+----+-------+----------+
only showing top 20 rows
```

- 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
```

```
]: df.distinct().show()
```

```
[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.

```
#df.sort(col("ID_EMPRESA").cast("integer")).show(30)
df2= df.orderBy(col("ID_EMPRESA").cast("integer").asc(),col("edad").cast("integer").desc())
```

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
5	Leandra	839-8044	at@pretiumetrutru...	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.nulla@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.ullamcor...	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@nislNu...	2015-11-15	19	1256.0	2
9	Hanae	935-2277	eu@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

-Función GROUP BY.

```
df.groupBy("ID_EMPRESA").sum("SALARIO").show(
```

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.

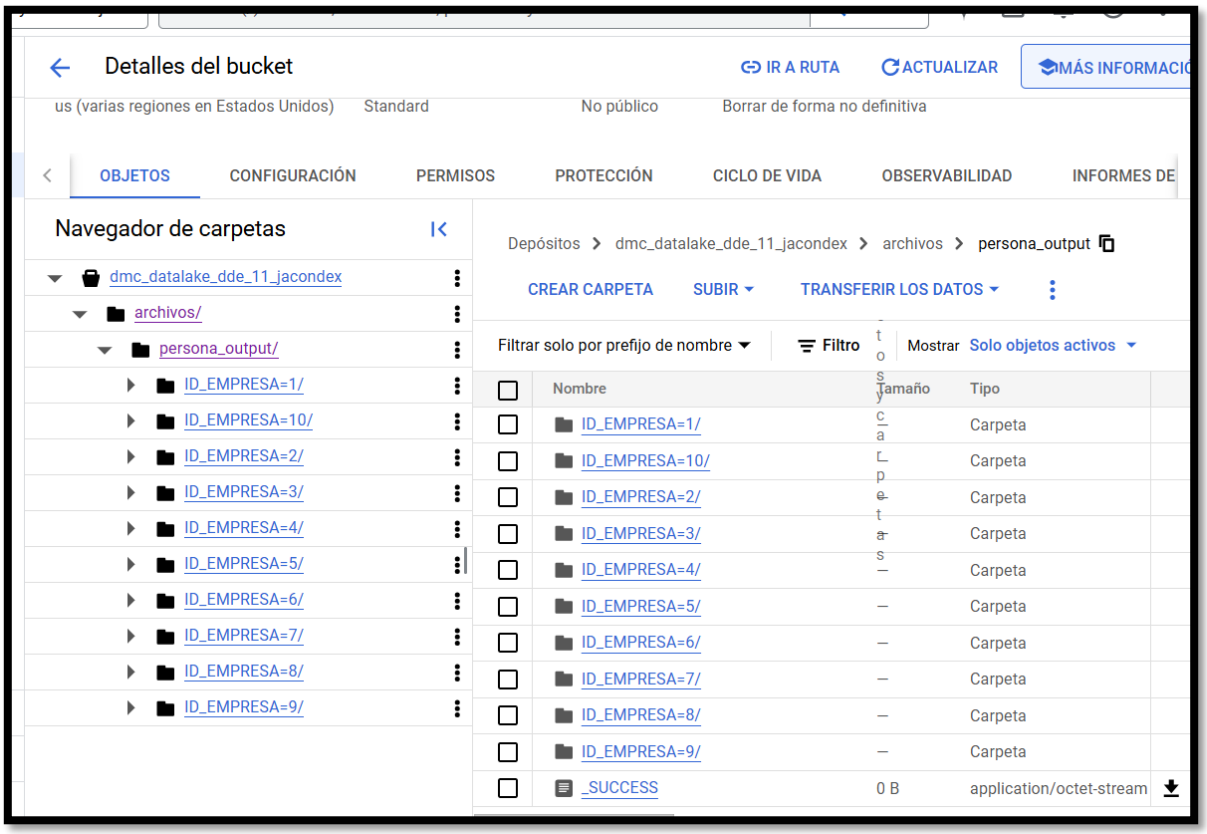
```
df.groupBy("ID_EMPRESA").agg(sum("SALARIO").alias("PLANILLA"), \
                             avg("EDAD").alias("PROM_EDAD"), \
                             max("SALARIO").alias("MAX_SALARY") \
                            ).where(col("PLANILLA")>=10000).show(10)
```

```
+----------+--------+------------------+----------+
|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.

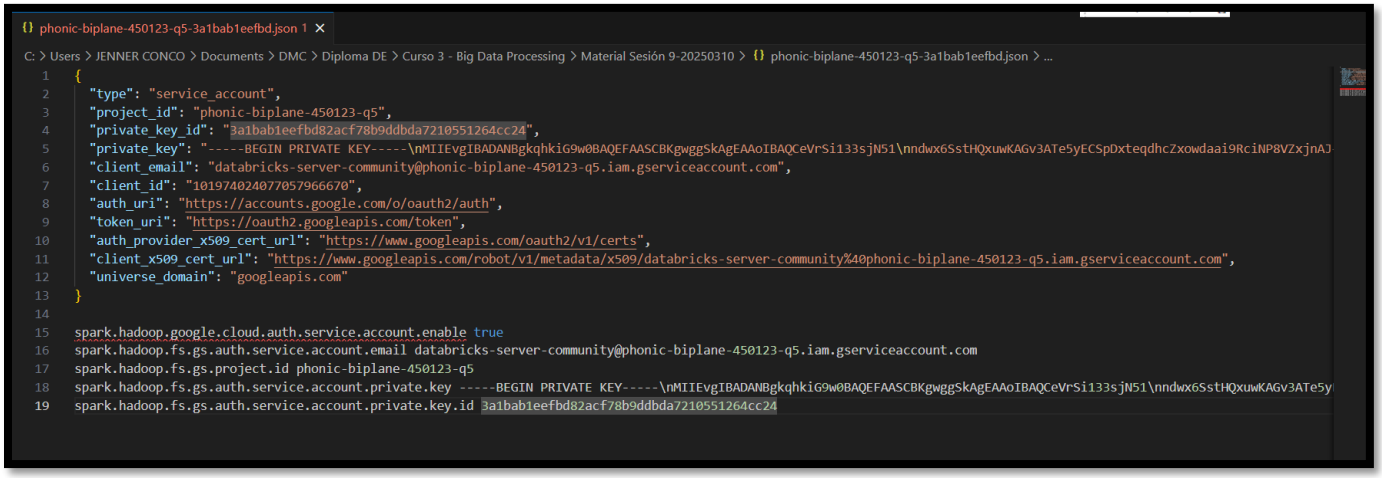
```
]: #PARTITIONBY ------ FUNCIONA AL MOMENTO DE ESCRIBIR EN DISCO df.write.partitionby("campo_particionado")
   #REPARTITION ------ CAMBIAR EL NRO DE PARTICIONES DE UN DATAFRAME DENTRO DE LA MEMORIA df.repartition(4)
   #COALESCE ------ REDUCIR U OPTIMIZAR EL NRO DE PARTICIONES SIN USAR UN SHUFFLE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

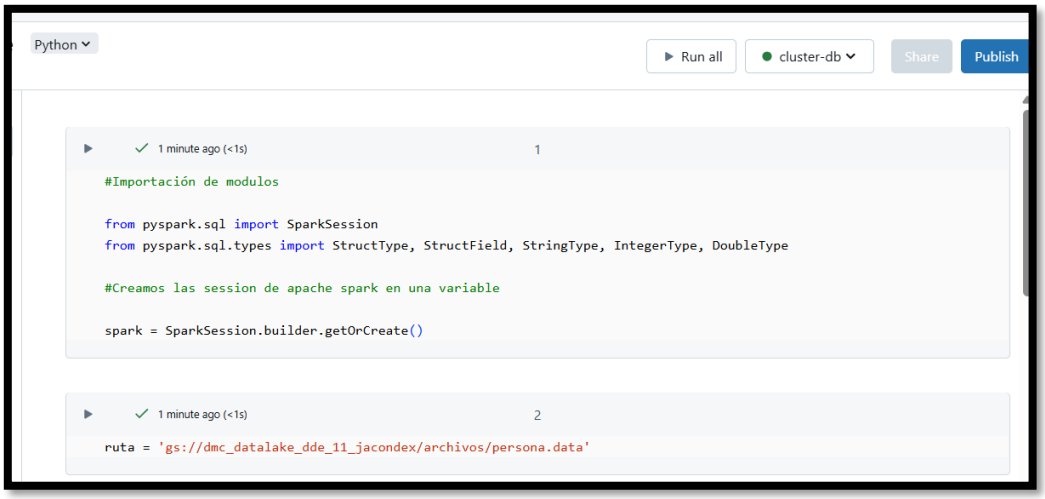


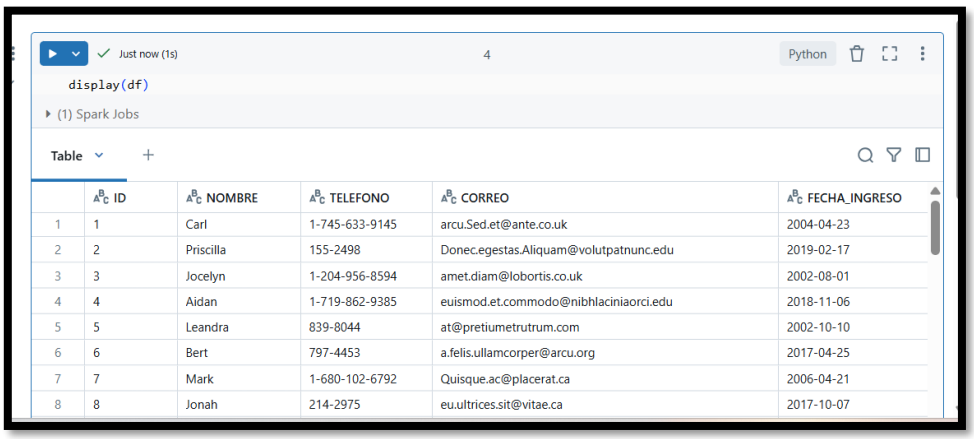
SESIÓN 09
CREACION DE CLUSTER Y NOTEBOOK

- Creación de claves en GCP.



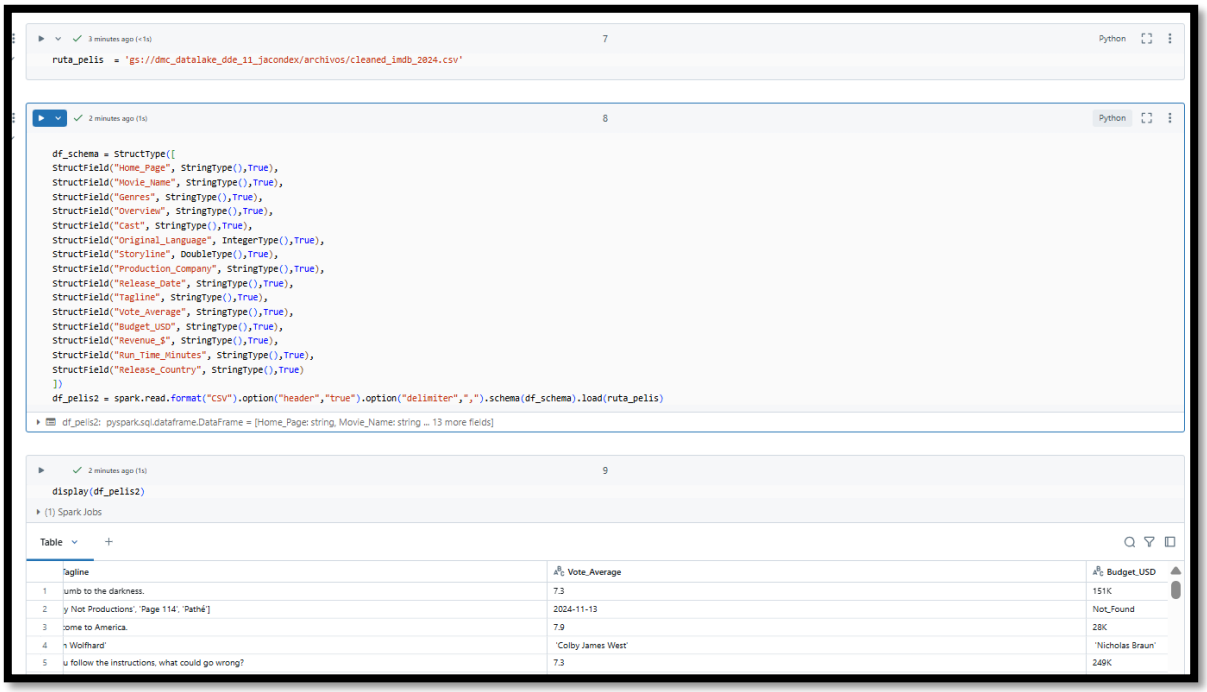
- Ejecución y uso de cluster con notebook.





	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@volutpatnunc.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@nibhlaciniaorci.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.



```
ruta_pelis = 'gs://dmc_datalake_dde_11_jacondex/archivos/cleaned_imdb_2024.csv'
```

```
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)
```

df_pelis2: pyspark.sql.dataframe.DataFrame = [Home_Page: string, Movie_Name: string ... 13 more fields]

```
display(df_pelis2)
```

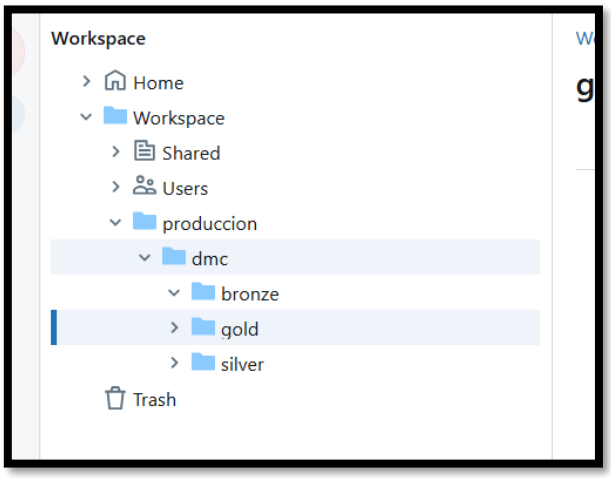
(1) Spark Jobs

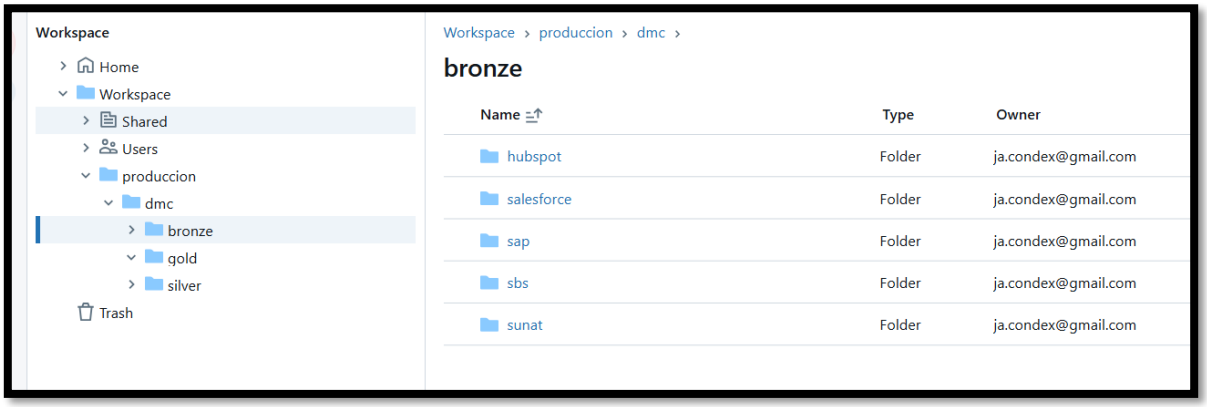
Table

	Tagline	Vote_Average	Budget_USD
1	umb to the darkness.	7.3	151K
2	y Not Productions', 'Page 114', 'Pathé']	2024-11-13	Not_Found
3	come to America.	7.9	28K
4	n Wolfhard'	'Colby James West'	'Nicholas Braun'
5	u follow the instructions, what could go wrong?	7.3	249K

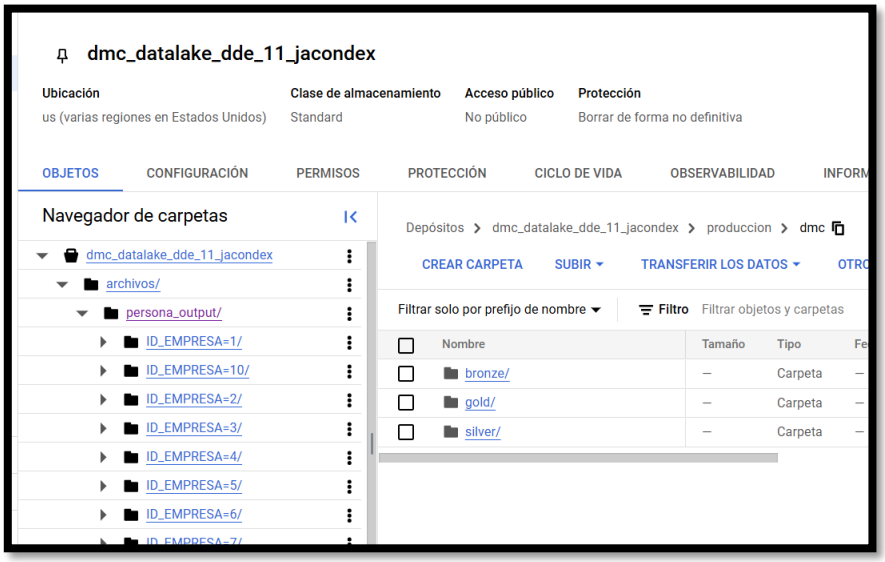
SESIÓN 10

- Creacion de carpetas en Databricks.

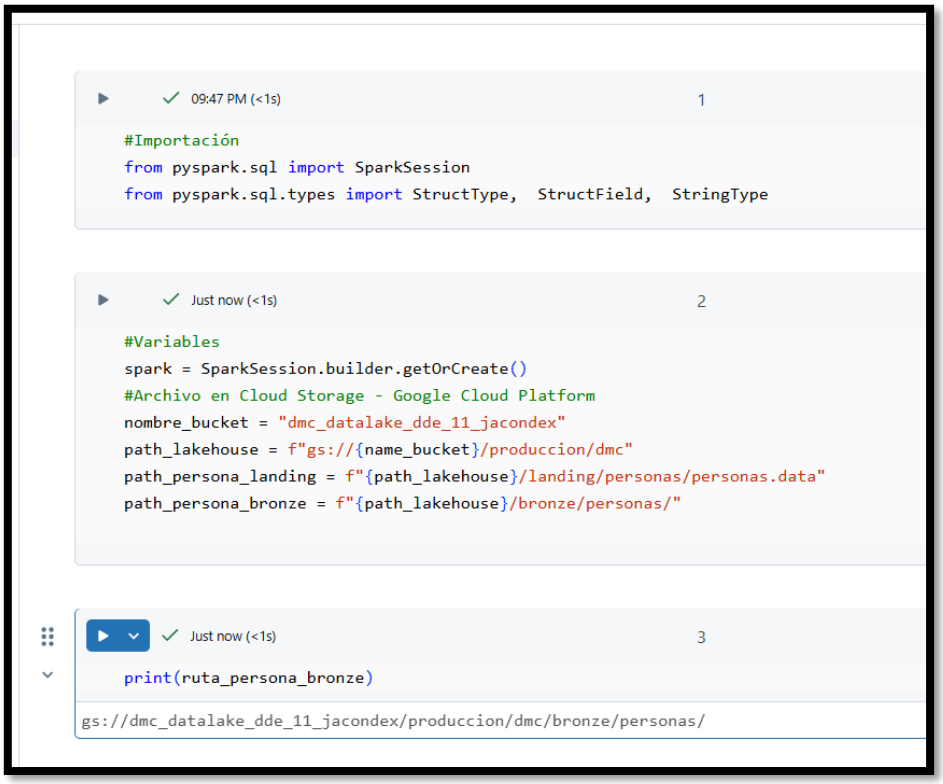




- Creación de estructuras de carpetas en Cloud Storage

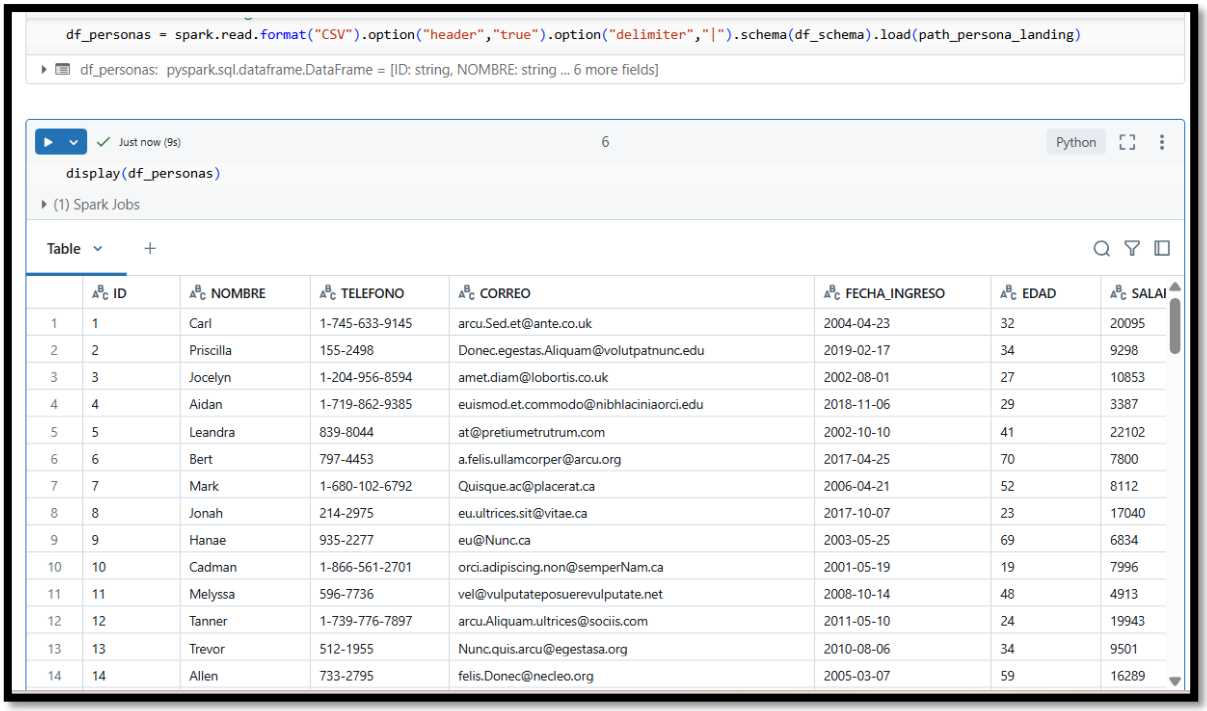


- Creando las variables en el notebook de databricks.



```
09:47 PM (<1s)  1
#Importación
from pyspark.sql import SparkSession
from pyspark.sql.types import StructType,  StructField,  StringType

Just now (<1s)  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/"

Just now (<1s)  3
print(ruta_persona_bronze)
gs://dmc_datalake_dde_11_jacondex/produccion/dmc/bronze/personas/
```

- Cargando DF_PERSONAS

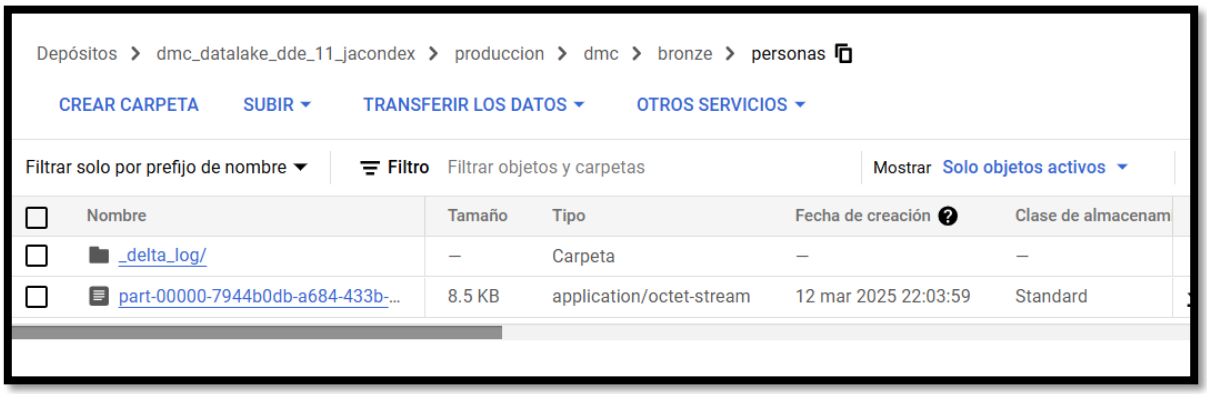


```
df_personas = spark.read.format("CSV").option("header","true").option("delimiter","|").schema(df_schema).load(path_persona_landing)
df_personas:  pyspark.sql.dataframe.DataFrame = [ID: string, NOMBRE: string ... 6 more fields]

Just now (9s)  6  Python
display(df_personas)
(1) Spark Jobs
```

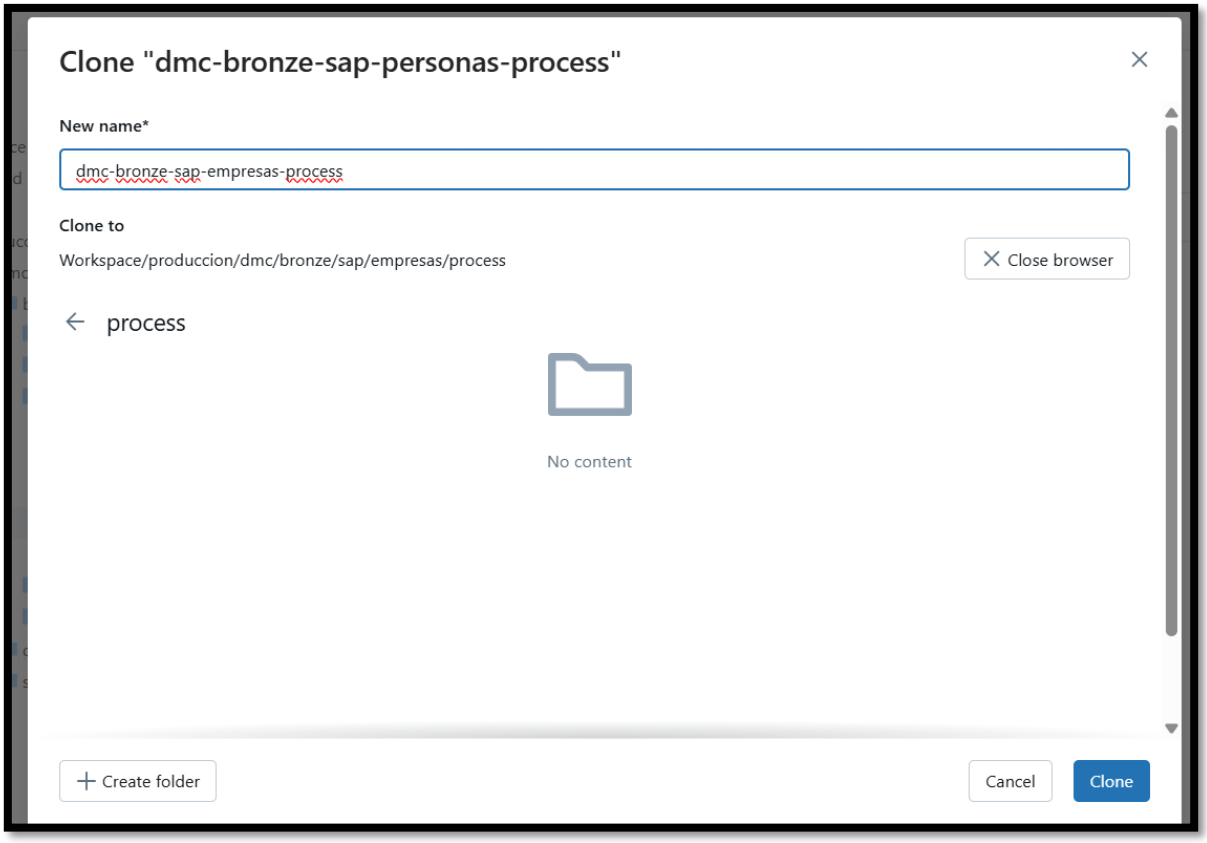
Table

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

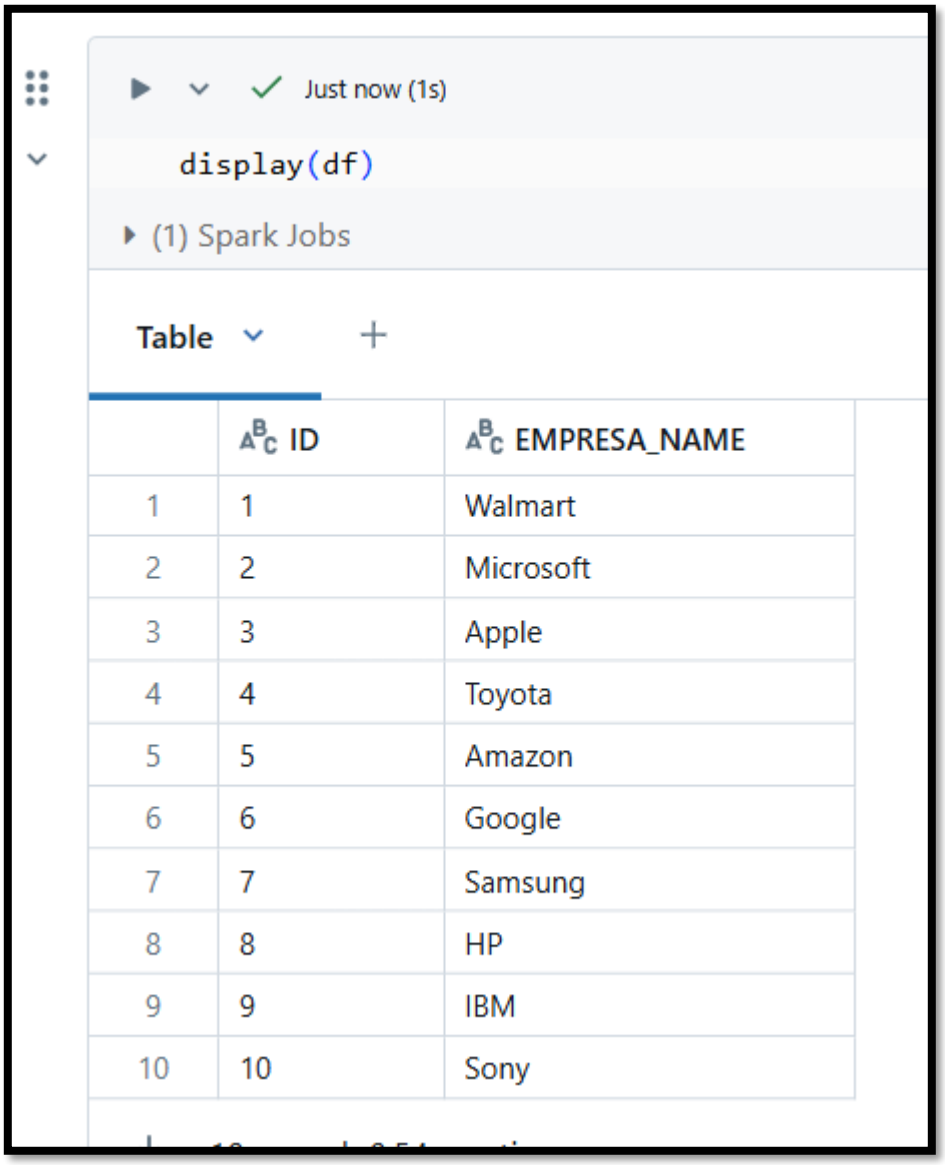
- Subiendo archivo delta a capa bronce.



- Clonando notebook.



- Cargando Empresas.



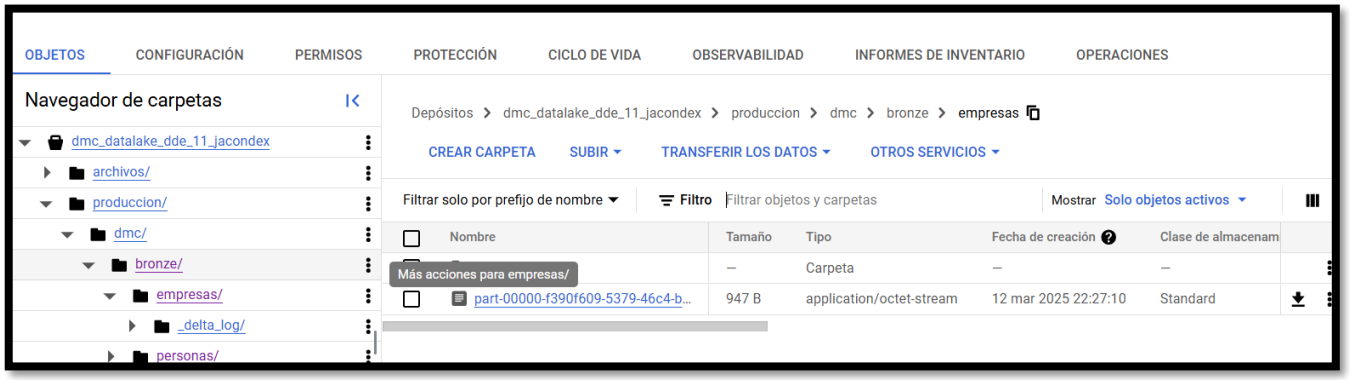
Just now (1s)

```
display(df)
```

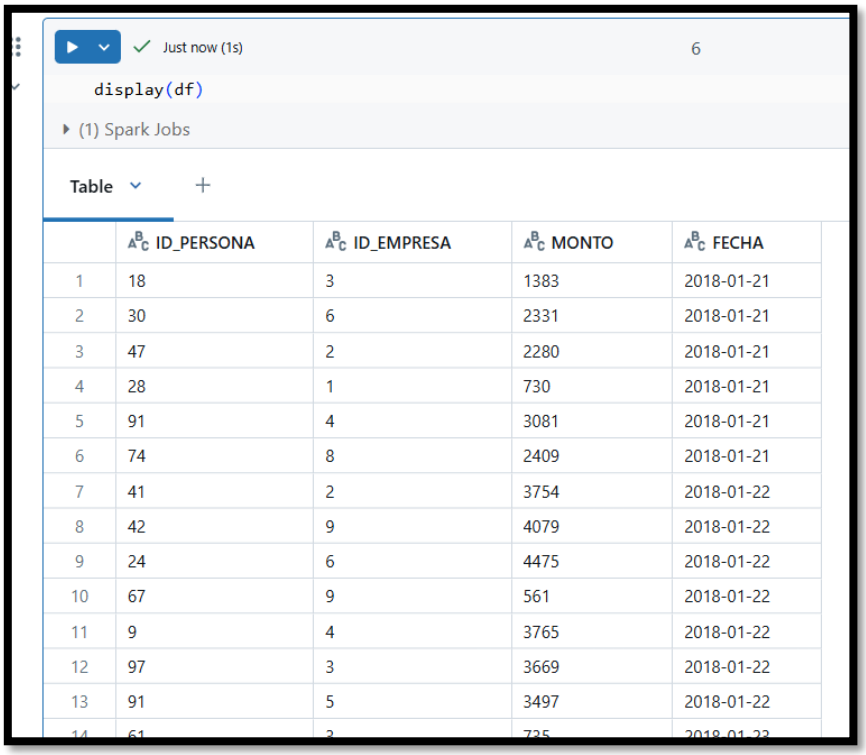
(1) Spark Jobs

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



- Cargando transacciones.



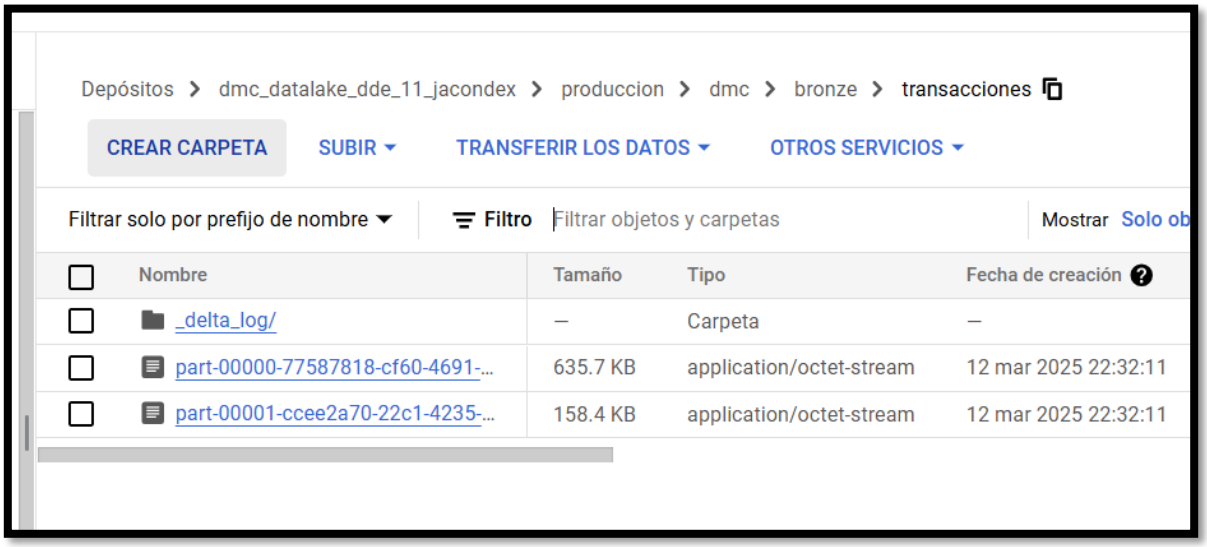
Just now (1s) 6

```
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



Depósitos > dmc_datalake_dde_11_jacondex > produccion > dmc > bronze > transacciones

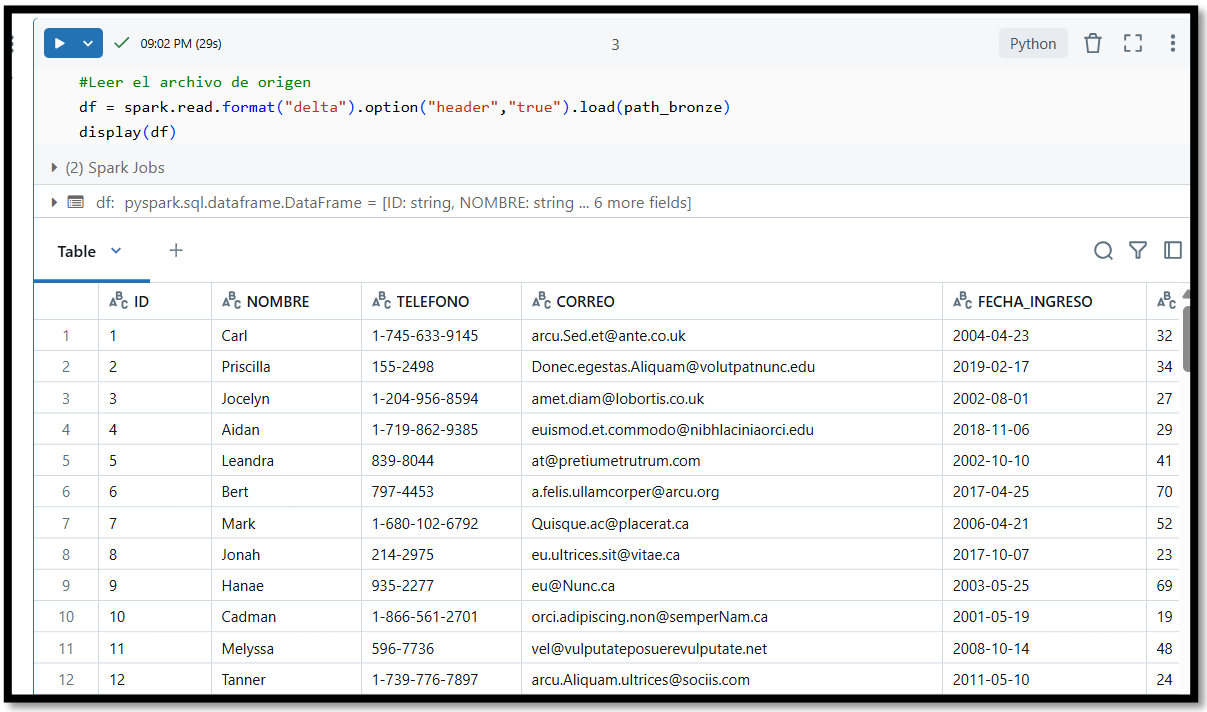
CREAR CARPETA SUBIR TRANSFERIR LOS DATOS OTROS SERVICIOS

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

Nombre	Tamaño	Tipo	Fecha de creación
_delta_log/	–	Carpeta	–
part-00000-77587818-cf60-4691-...	635.7 KB	application/octet-stream	12 mar 2025 22:32:11
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@volutpatnunc.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@nibhlaciniaorci.edu	2018-11-06	29
5	5	Leandra	839-8044	at@pretiumetrutrum.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.

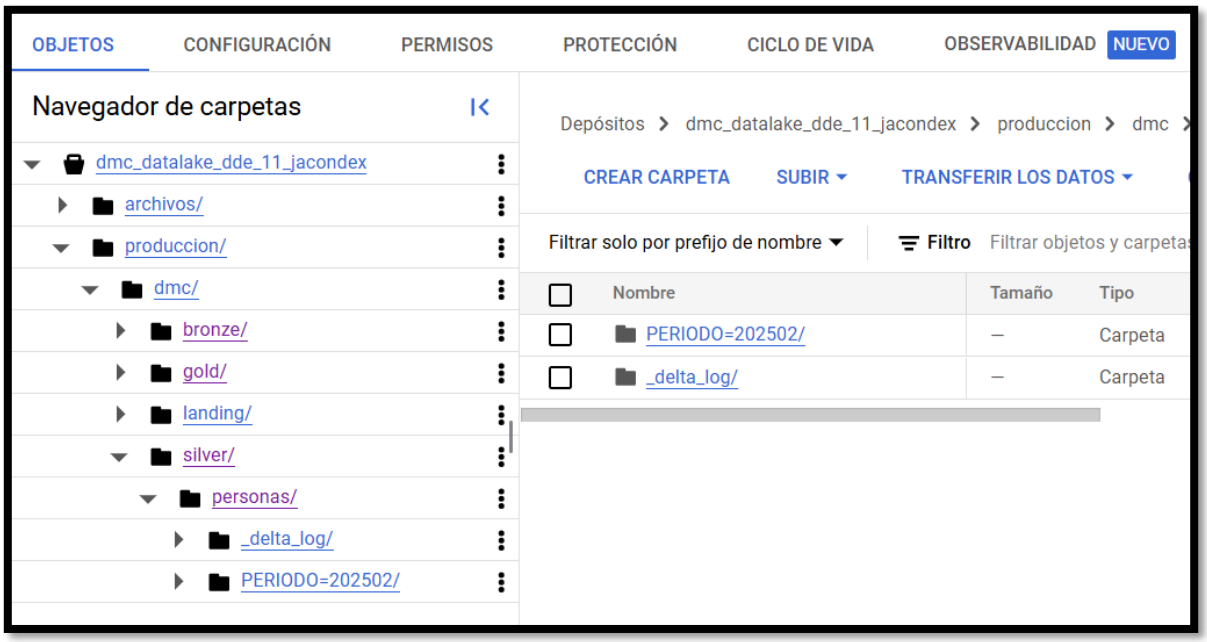
```
#La Transformación de datos
df_t = df.withColumn("TELEFONO", regexp_replace('telefono', '-', ''))\
  .withColumn("PERIODO",date_format(add_months(current_date(),-1),"yyyyMM"))\
  .withColumn("SEGMENTO", when(col("SALARIO")<3500,"Masivo").when((col("SALARIO")>=3500) & (col("SALARIO")<=10000),"Premiun").otherwise("Beyond"))


#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

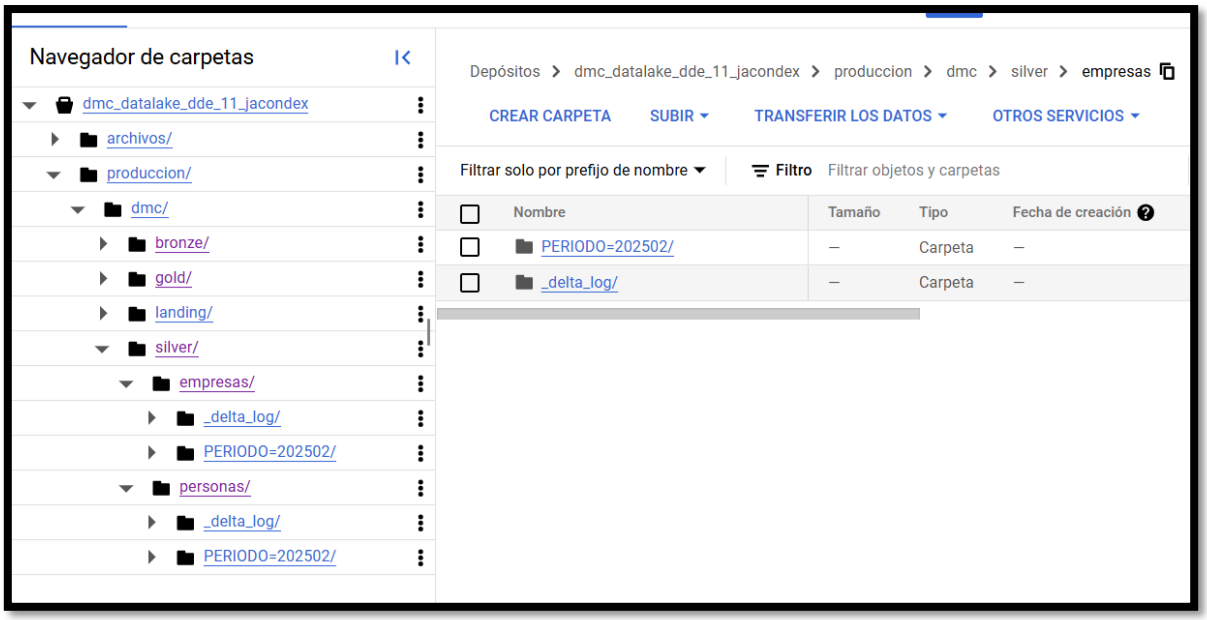
```
ID: integer
NOMBRE: string
TELEFONO: string
CORREO: string
FECHA_INGRESO: date
EDAD: integer
SALARIO: double
ID_EMPRESA: integer
PERIODO: string
SEGMENTO: string
ANIO: integer
MES: integer
DIA: integer
```

-

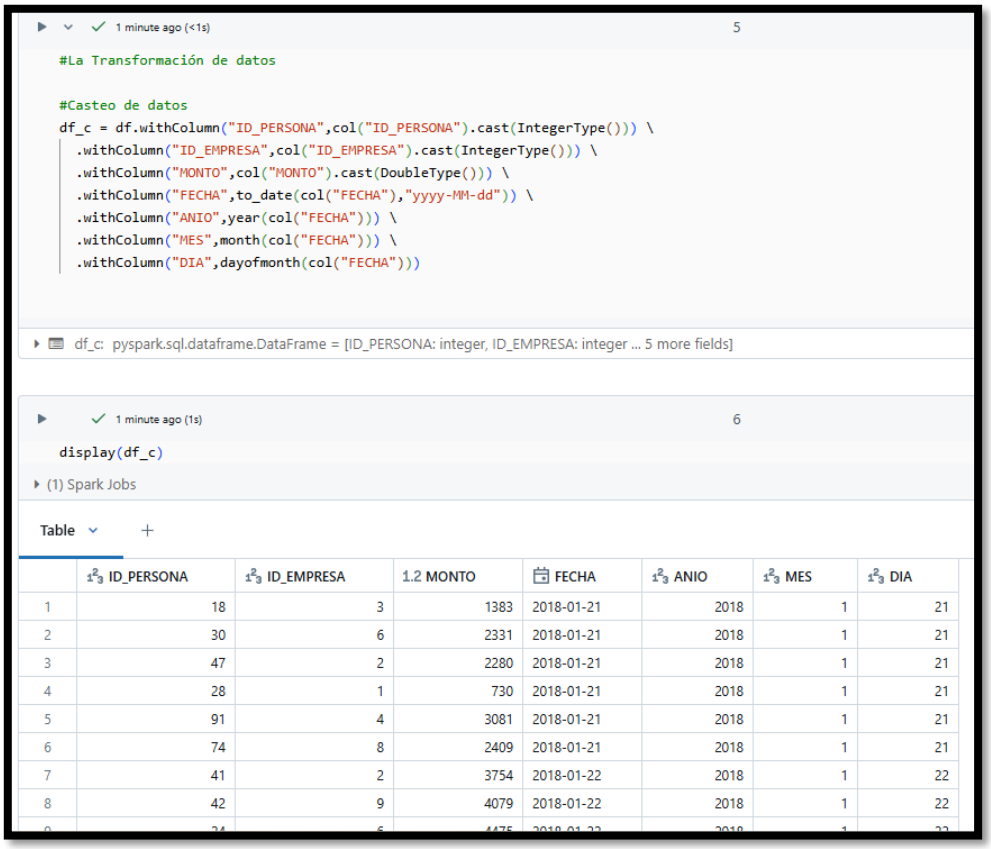
- Data particionada en capa Silver.



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



- 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]

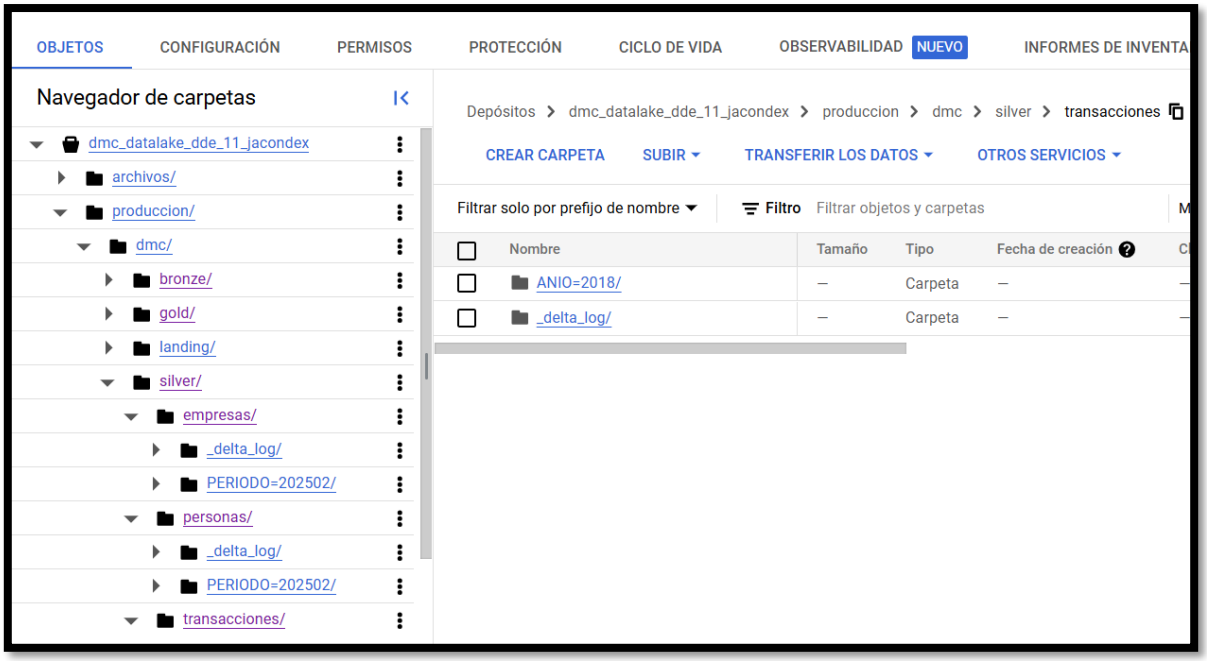
```
display(df_c)
```

(1) Spark Jobs

Table

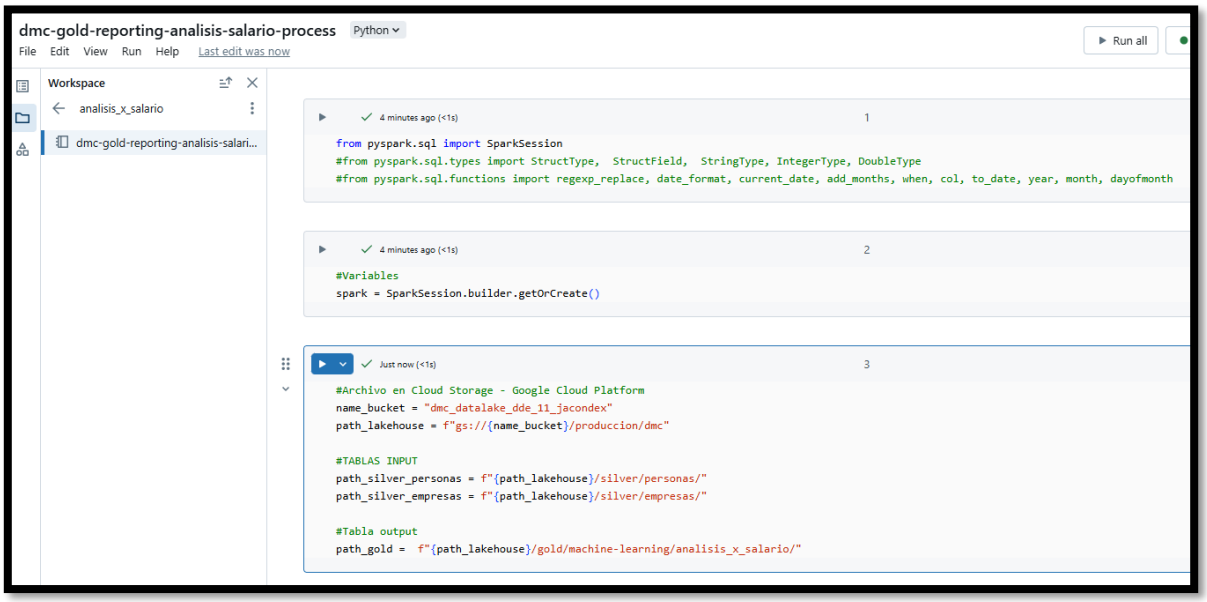
	ID_PERSONA	ID_EMPRESA	MONTO	FECHA	ANIO	MES	DIA
1	18	3	1383	2018-01-21	2018	1	21
2	30	6	2331	2018-01-21	2018	1	21
3	47	2	2280	2018-01-21	2018	1	21
4	28	1	730	2018-01-21	2018	1	21
5	91	4	3081	2018-01-21	2018	1	21
6	74	8	2409	2018-01-21	2018	1	21
7	41	2	3754	2018-01-22	2018	1	22
8	42	9	4079	2018-01-22	2018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

- Guando en cloud storage – transacciones.

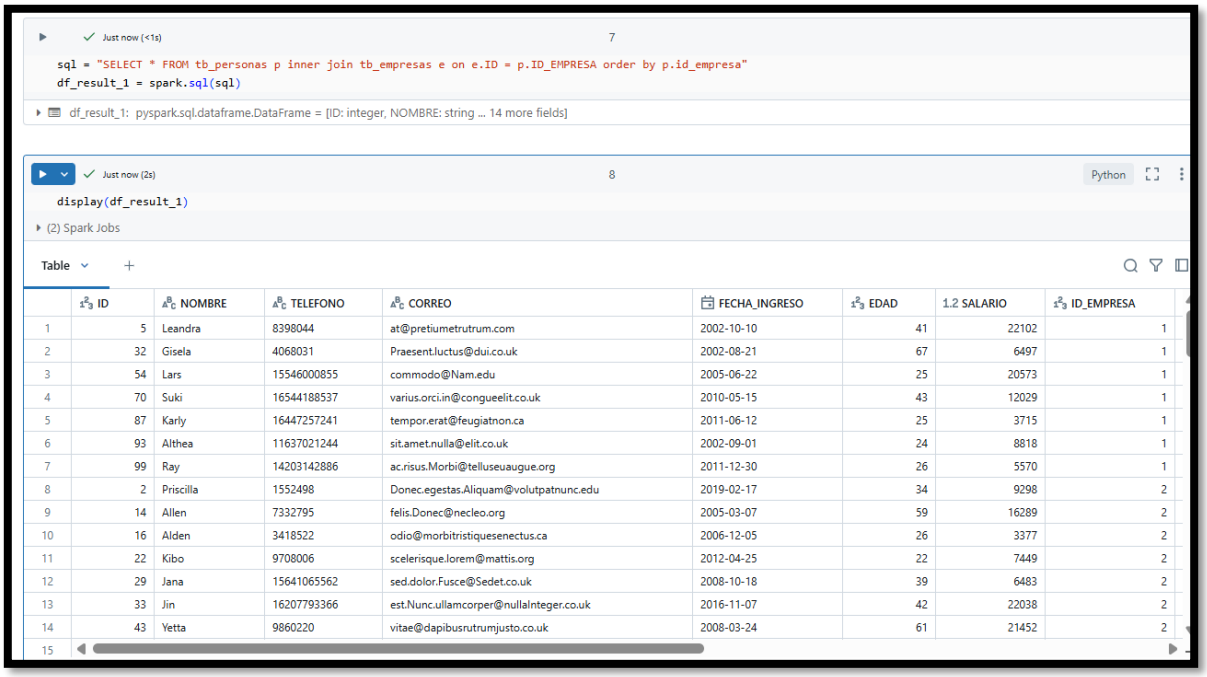


SESIÓN 12

- Capa GOLD.

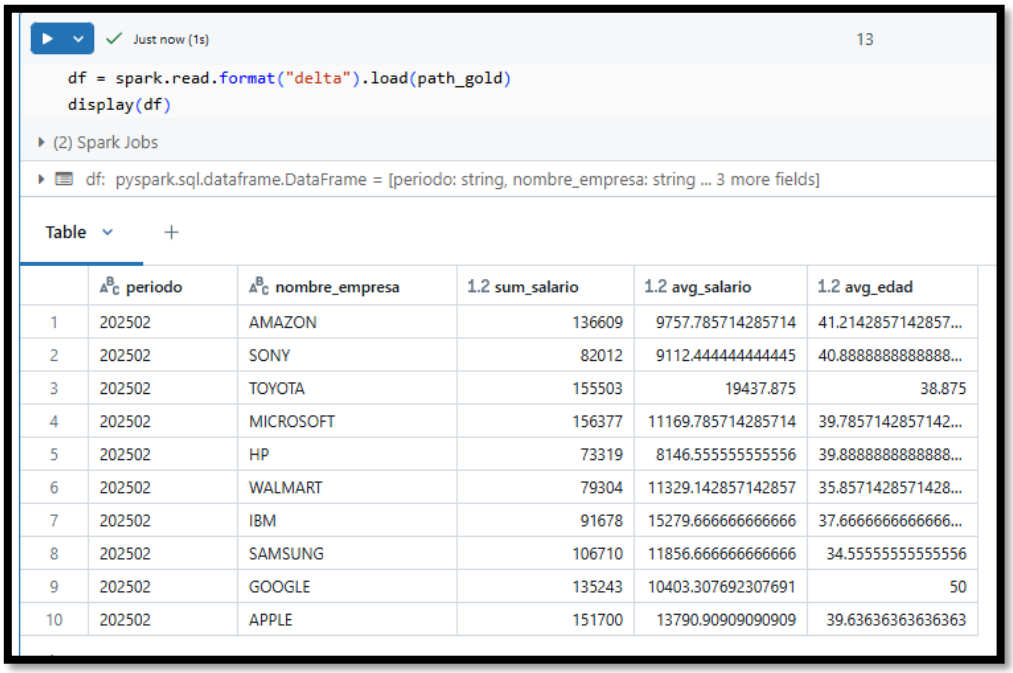


- Uso de select en tablas creadas.



	ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO	EDAD	SALARIO	ID_EMPRESA
1	5	Leandra	8398044	at@pretiumetrutrum.com	2002-10-10	41	22102	1
2	32	Gisela	4068031	Praesent.luctus@dui.co.uk	2002-08-21	67	6497	1
3	54	Lars	15546000855	commodo@Nam.edu	2005-06-22	25	20573	1
4	70	Suki	16544188537	varius.orci.in@congueelit.co.uk	2010-05-15	43	12029	1
5	87	Karly	16447257241	tempor.erat@feugiatnon.ca	2011-06-12	25	3715	1
6	93	Althea	11637021244	sit.amet.nulla@elit.co.uk	2002-09-01	24	8818	1
7	99	Ray	14203142886	ac.risus.Morbi@telluseuaugue.org	2011-12-30	26	5570	1
8	2	Priscilla	1552498	Donec.egestas.Aliquam@volutpatnunc.edu	2019-02-17	34	9298	2
9	14	Allen	7332795	felis.Donec@necleo.org	2005-03-07	59	16289	2
10	16	Alden	3418522	odio@morbitristiquesenectus.ca	2006-12-05	26	3377	2
11	22	Kibo	9708006	scelerisque.lorem@mattis.org	2012-04-25	22	7449	2
12	29	Jana	15641065562	sed.dolor.Fusce@Sedet.co.uk	2008-10-18	39	6483	2
13	33	Jin	16207793366	est.Nunc.ullamcorper@nullaInteger.co.uk	2016-11-07	42	22038	2
14	43	Yetta	9860220	vitae@dapibusrutrumjusto.co.uk	2008-03-24	61	21452	2
15								

- Poblando capa final GOLD.



```
df = spark.read.format("delta").load(path_gold)
display(df)
```

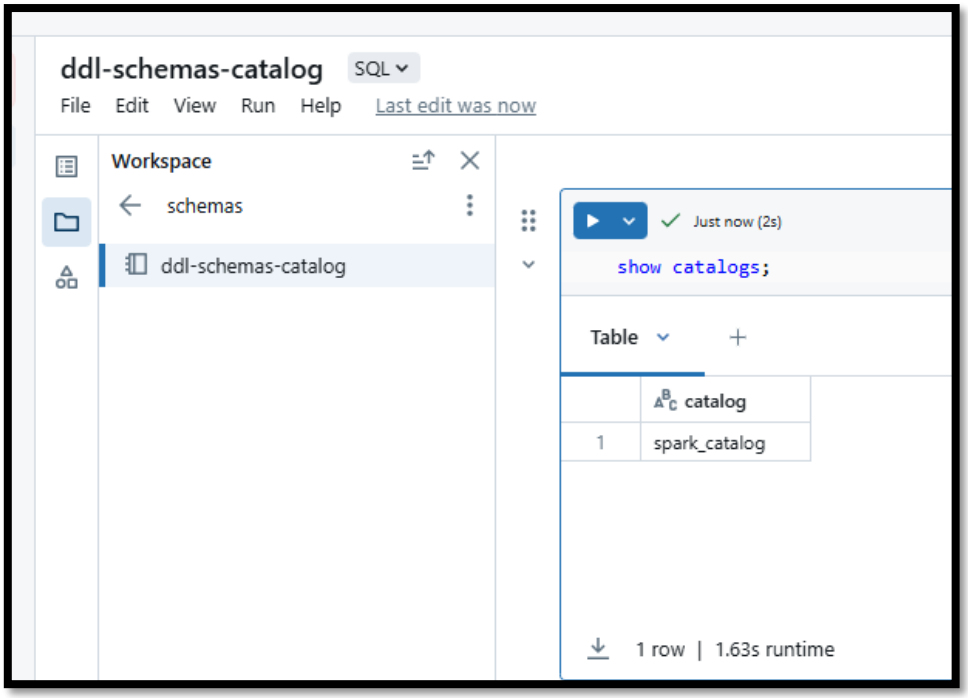
(2) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [periodo: string, nombre_empresa: string ... 3 more fields]

Table

	periodo	nombre_empresa	sum_salario	avg_salario	avg_edad
1	202502	AMAZON	136609	9757.785714285714	41.2142857142857...
2	202502	SONY	82012	9112.444444444445	40.8888888888888...
3	202502	TOYOTA	155503	19437.875	38.875
4	202502	MICROSOFT	156377	11169.785714285714	39.7857142857142...
5	202502	HP	73319	8146.555555555556	39.8888888888888...
6	202502	WALMART	79304	11329.142857142857	35.8571428571428...
7	202502	IBM	91678	15279.666666666666	37.6666666666666...
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

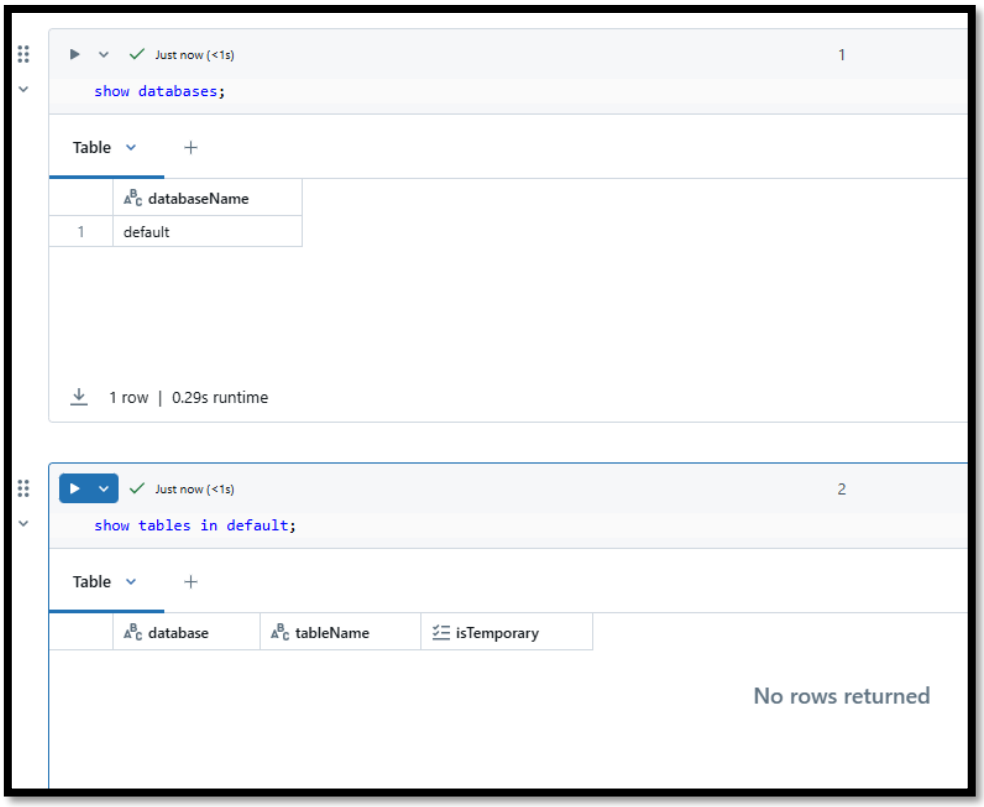
Just now (2s)

```
show catalogs;
```

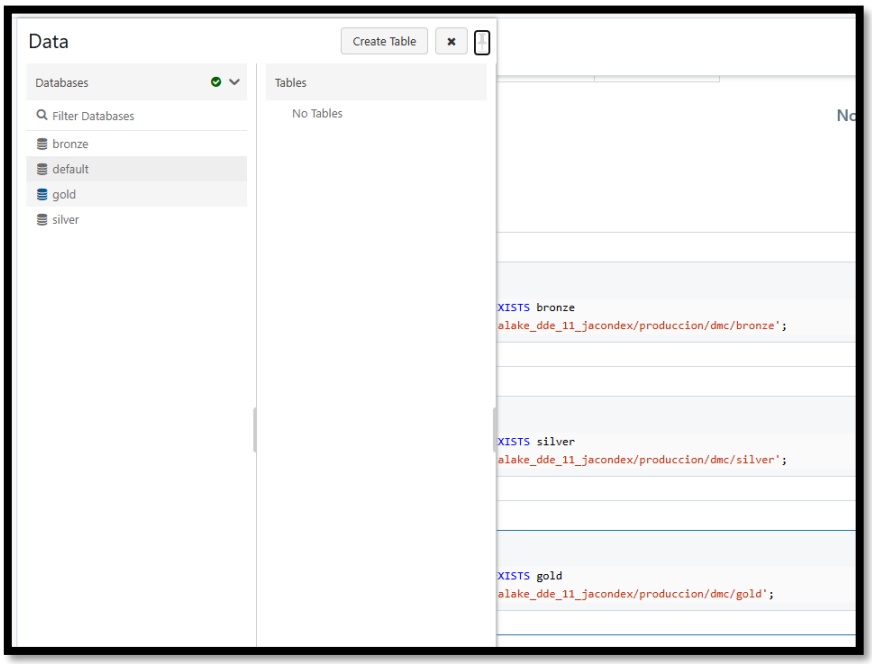
Table

	catalog
1	spark_catalog

1 row | 1.63s runtime

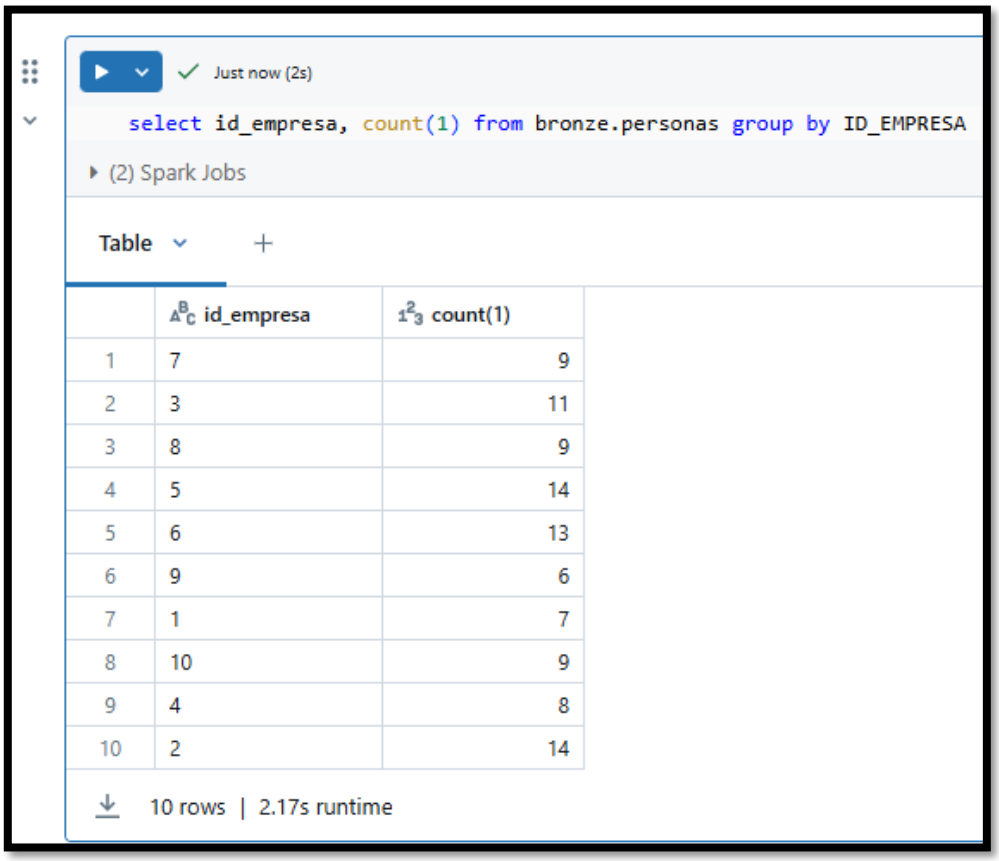


- Creacion de 3 capas en databricks.



- Creación de tabla bronze.personas.

```
Just now (19s)

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/"

(3) Spark Jobs

OK
```

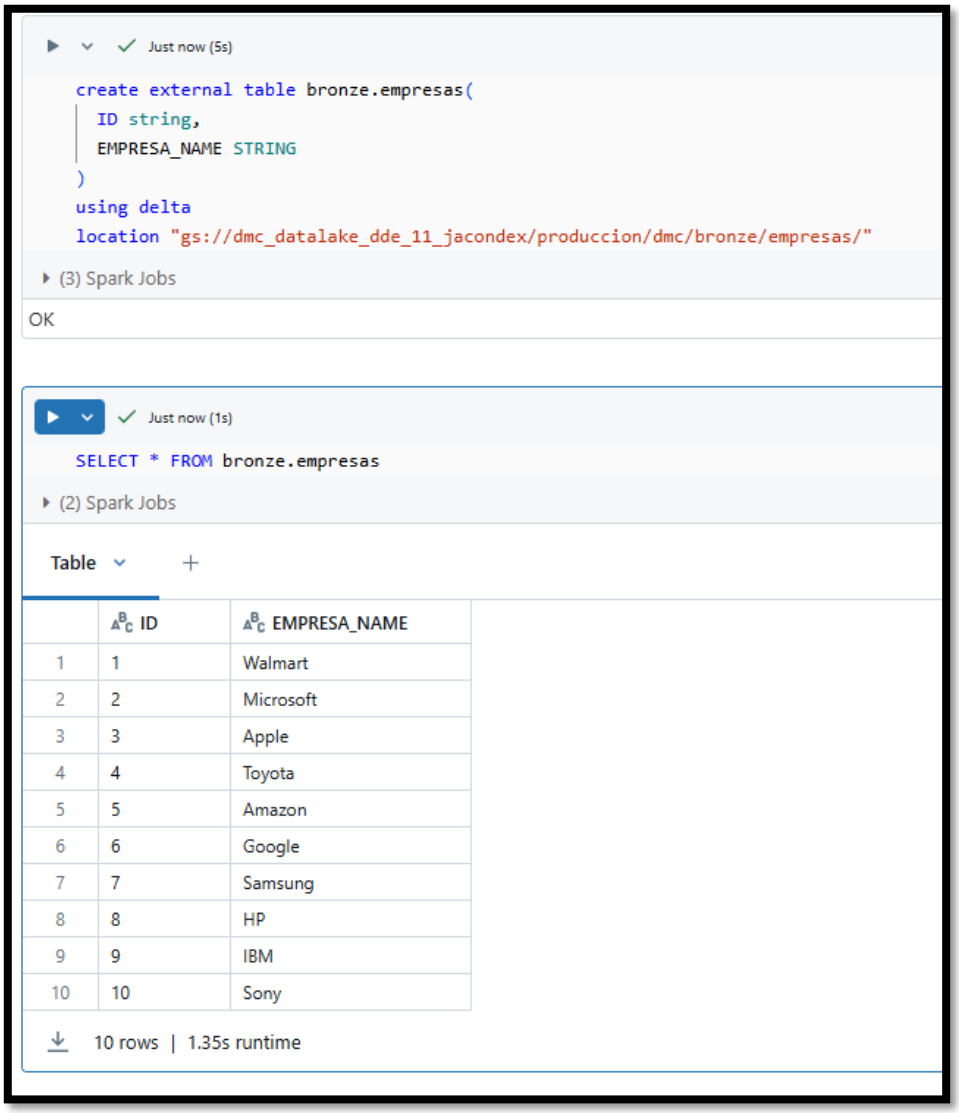


```
Just now (2s)

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

(2) Spark Jobs

Table +
```

	id_empresa	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

10 rows | 2.17s runtime

- Creación de tabla bronze.empresas.



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

Just now (1s)

```
SELECT * FROM bronze.empresas
```

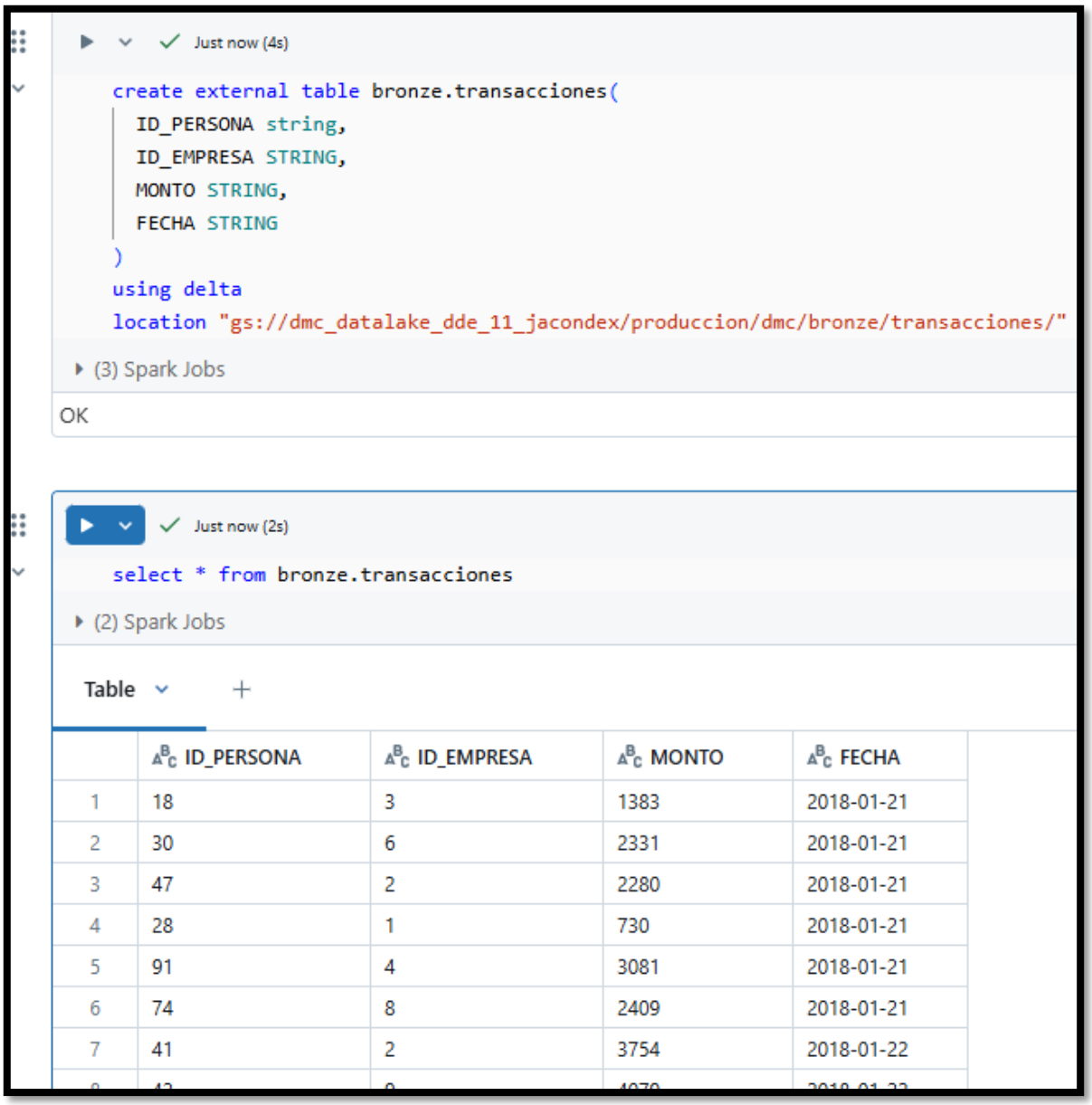
(2) Spark Jobs

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.



Just now (4s)

```
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/"
```

(3) Spark Jobs

OK

Just now (2s)

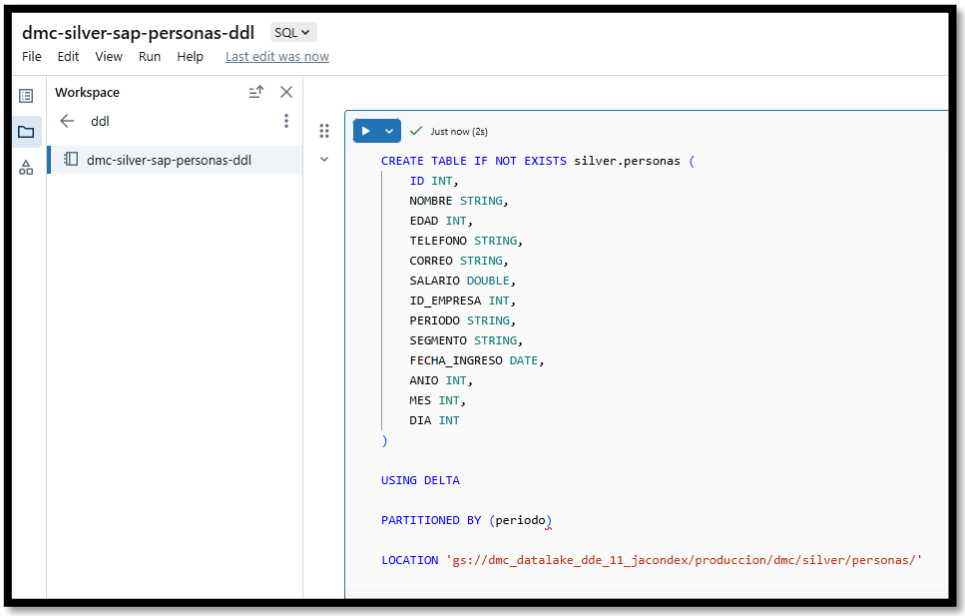
```
select * from bronze.transacciones
```

(2) Spark Jobs

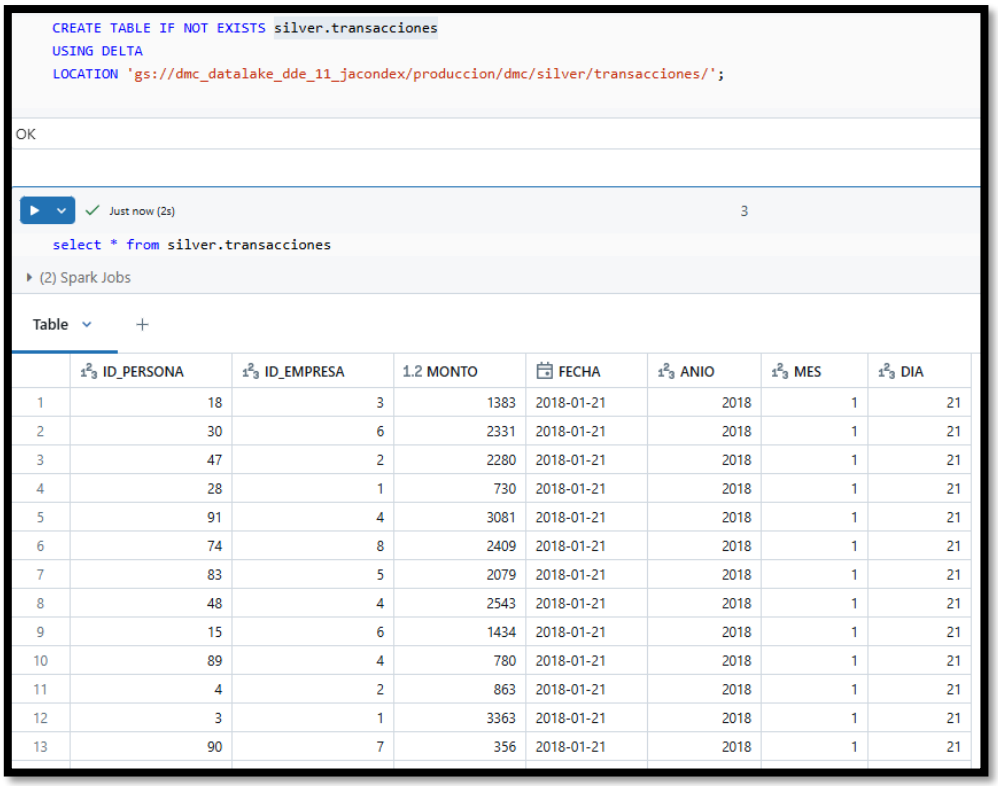
Table

	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

- Creando tabla silver.personas.



- Creando tabla silver.transacciones.



	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