DIPLOMA DATA ENGINNER

Big Data Processing

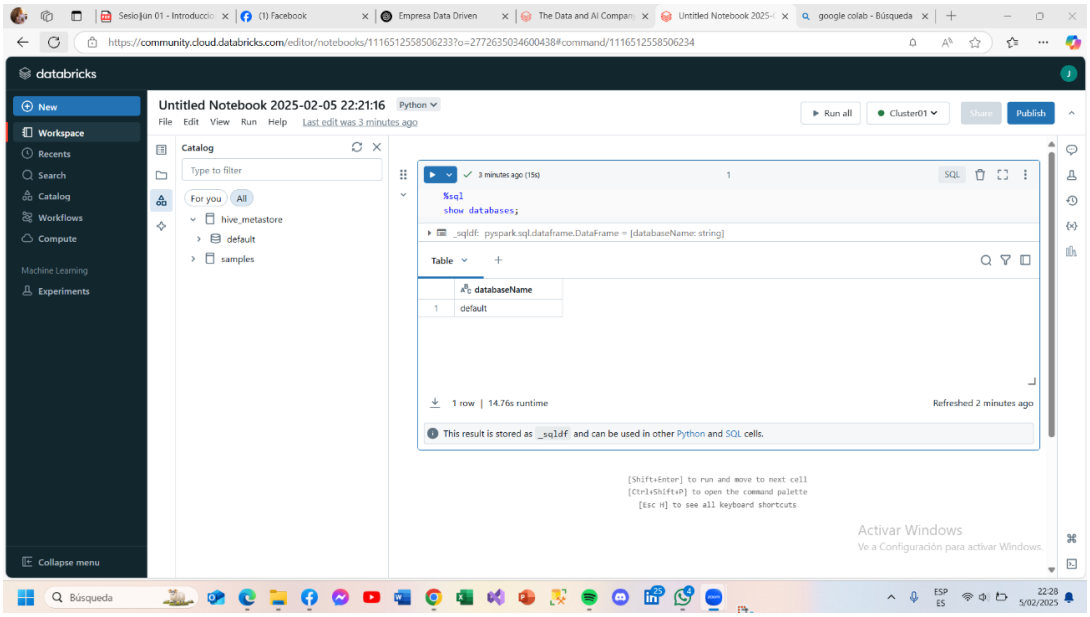
Edicion 11 – Noviembre 2024

Proyecto Final

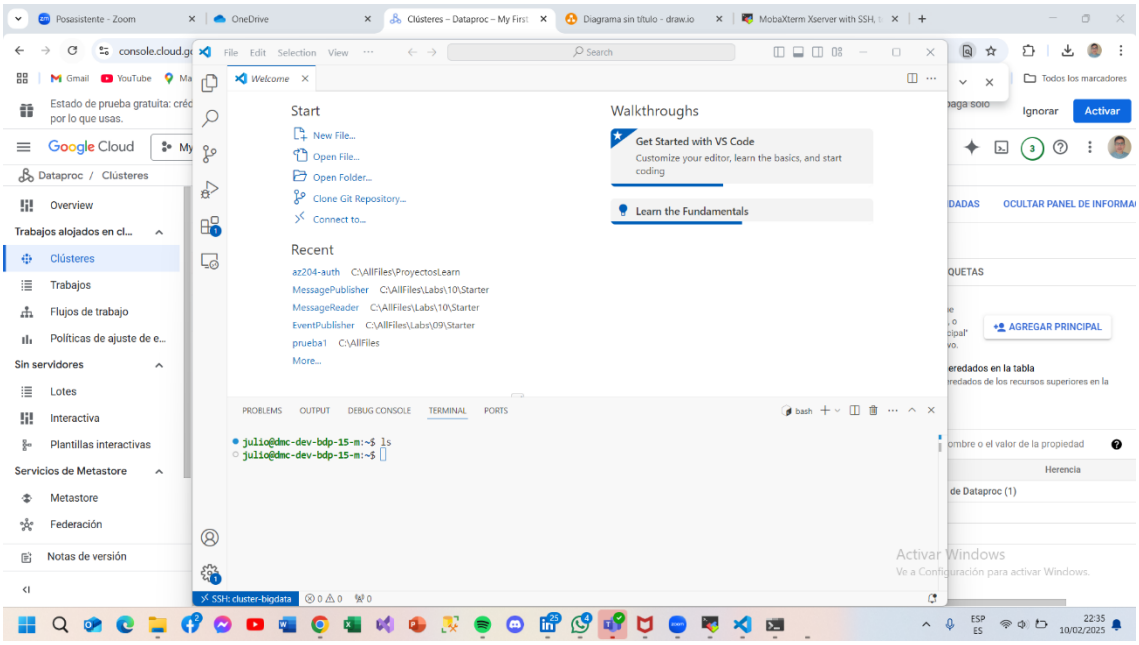
Alumno: Julio Alexander Vasquez Pacheco
Curso: Big Data procesing
Trabajo: Captura de pantalla de laboratorios en cada sesión y adjuntar los archivos.
Github: https://github.com/julio2019vp/curso-dmc-big-data-proccesing

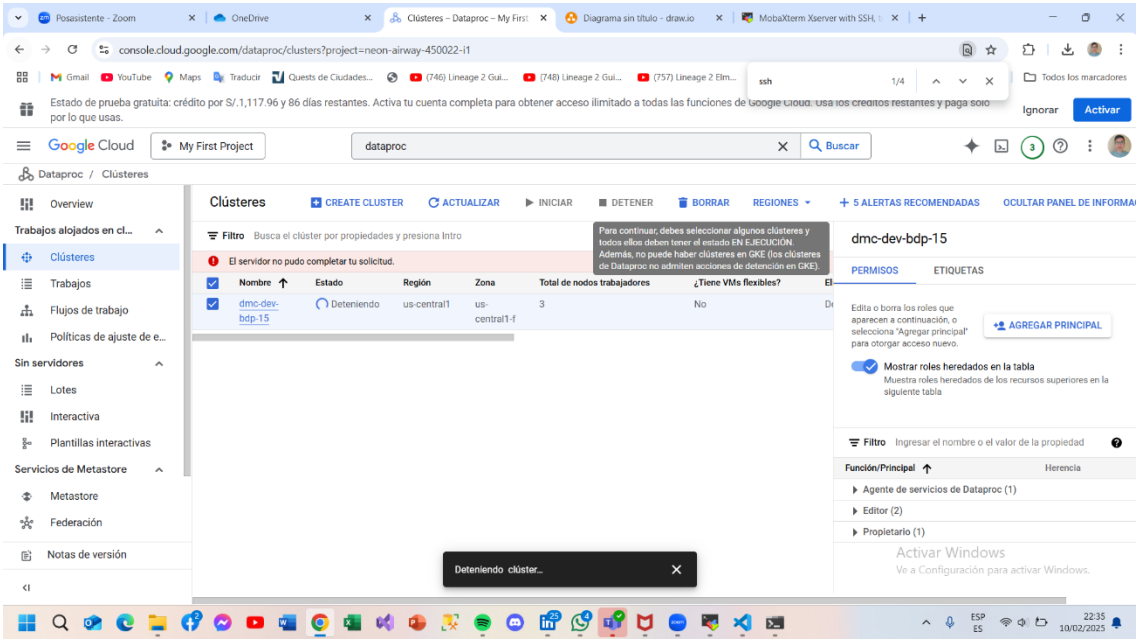
Captura de Pantalla de Laboratorios en cada sesión

1. Sesión 1

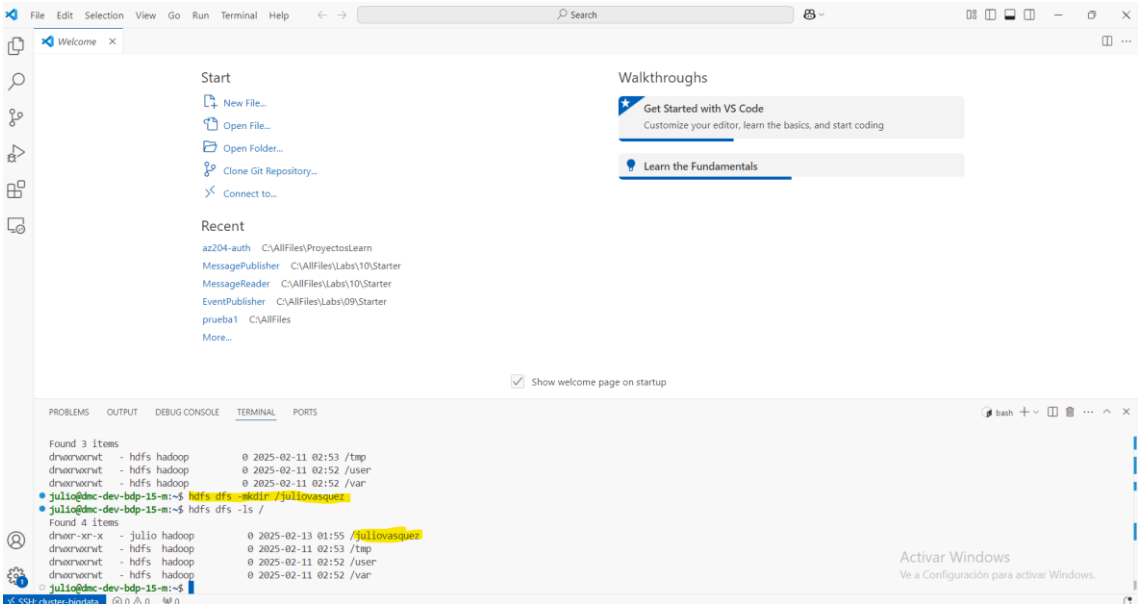


2. Sesión 2

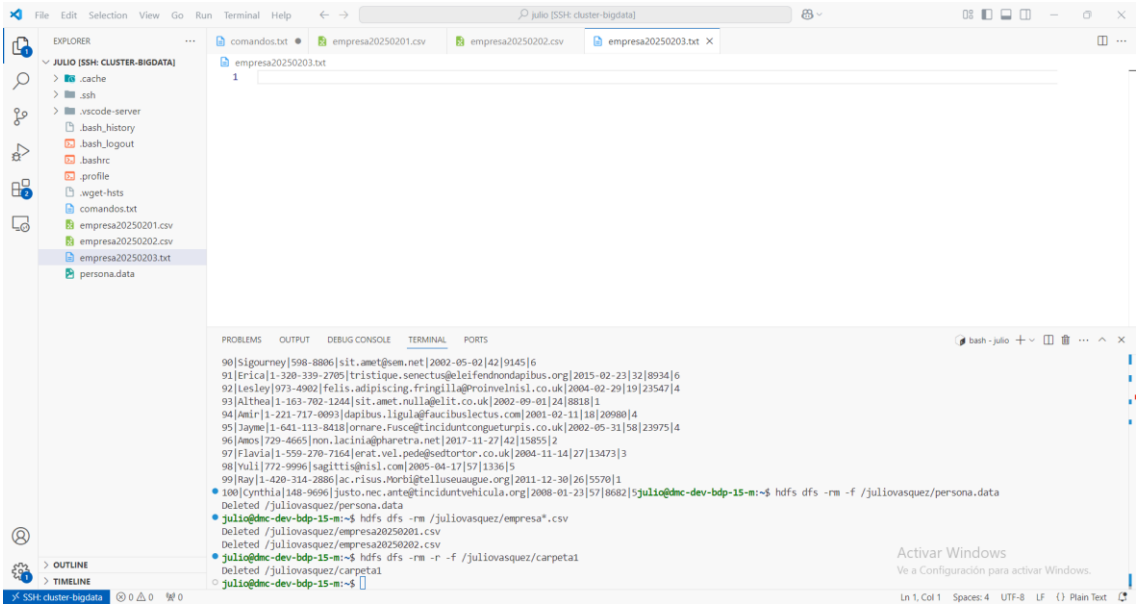


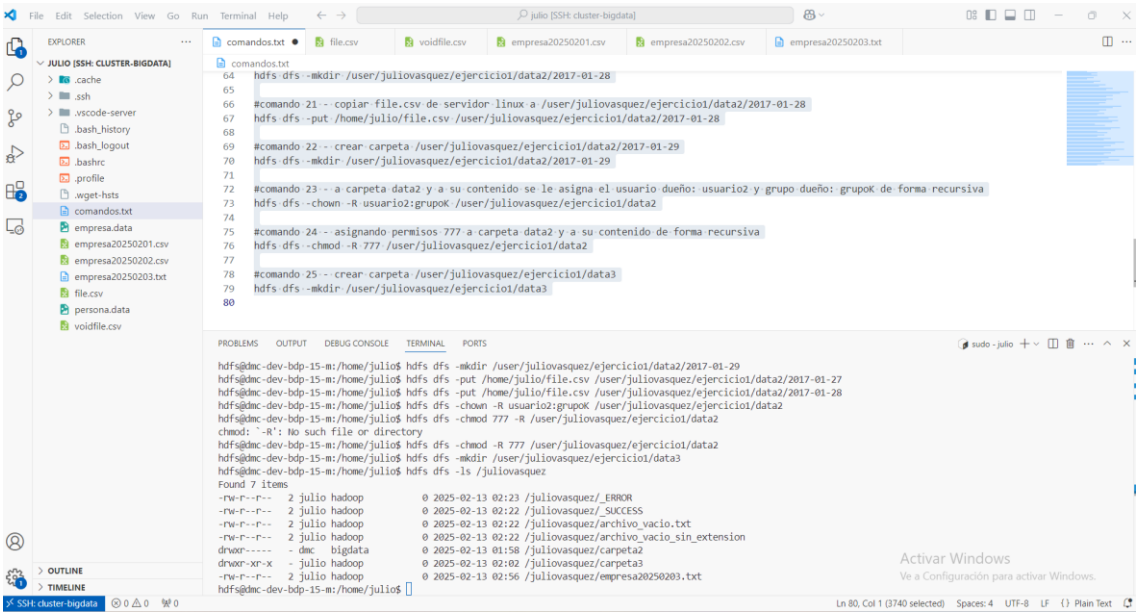

Clústeres – Dataproc – My First
console.cloud.google.com/dataproc/clusters?project=neon-airway-450022-i1
Estado de prueba gratuita: crédito por S/.1,117.96 y 86 días restantes. Activa tu cuenta completa para obtener acceso ilimitado a todas las funciones de Google Cloud. Usa los créditos restantes y paga solo por lo que usas.
Ignorar
Activar
Google Cloud
My First Project
dataproc
Buscar
Dataproc / Clústeres
Overview
Trabajos alojados en cl...
Clústeres
Trabajos
Flujos de trabajo
Políticas de ajuste de e...
Sin servidores
Lotes
Interactiva
Plantillas interactivas
Servicios de Metastore
Metastore
Federación
Notas de versión
Clústeres
CREATE CLUSTER
ACTUALIZAR
INICIAR
DETENER
BORRAR
REGIONES
5 ALERTAS RECOMENDADAS
OCULTAR PANEL DE INFORMA
Filtro Busca el clúster por propiedades y presiona Intro
El servidor no pudo completar tu solicitud.
Para continuar, debes seleccionar algunos clústeres y todos ellos deben tener el estado EN EJECUCIÓN. Además, no puede haber clústeres en GKE (los clústeres de Dataproc no admiten acciones de detención en GKE).
Nombre
Estado
Región
Zona
Total de nodos trabajadores
¿Tiene VMs flexibles?
dmc-dev-bdp-15
Deteniendo
us-central1
us-central1-f
3
No
dmc-dev-bdp-15
PERMISOS
ETIQUETAS
Edita o borra los roles que aparecen a continuación, o selecciona "Agregar principal" para otorgar acceso nuevo.
AGREGAR PRINCIPAL
Mostrar roles heredados en la tabla
Muestra roles heredados de los recursos superiores en la siguiente tabla
Filtro Ingresar el nombre o el valor de la propiedad
Función/Principal
Herencia
Agente de servicios de Dataproc (1)
Editor (2)
Propietario (1)
Deteniendo clúster...
Activar Windows
Ve a Configuración para activar Windows.
ESP ES
22:35
10/02/2025

3. Sesión 3

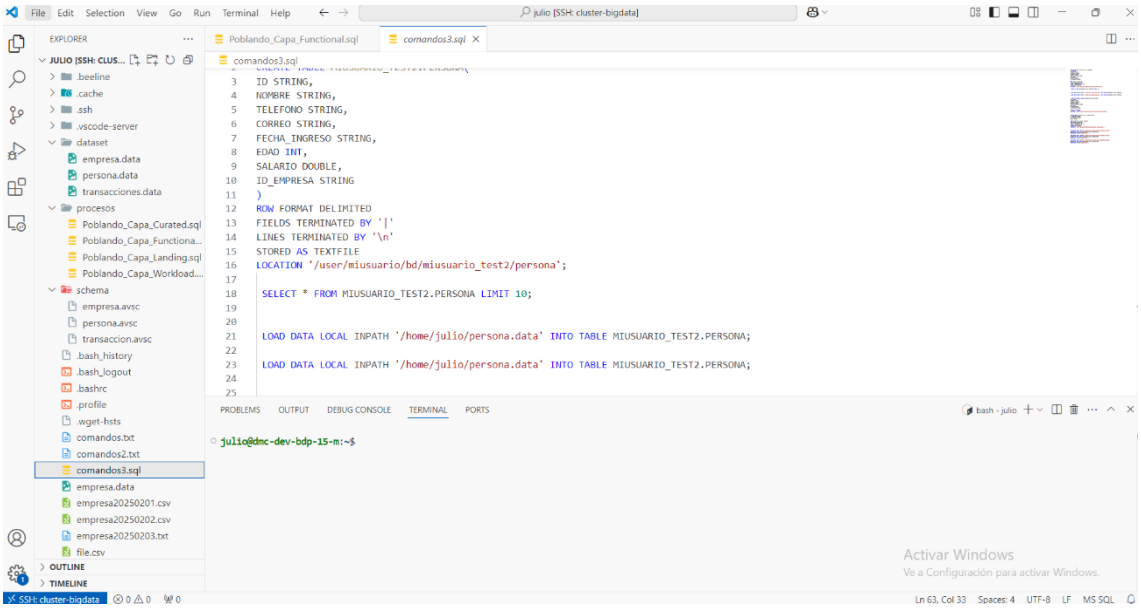

File Edit Selection View Go Run Terminal Help
Search
Welcome
Start
New File...
Open File...
Open Folder...
Clone Git Repository...
Connect to...
Recent
az204-auth C:\AllFiles\ProyectosLearn
MessagePublisher C:\AllFiles\Labs\10\Starter
MessageReader C:\AllFiles\Labs\10\Starter
EventPublisher C:\AllFiles\Labs\09\Starter
prueba1 C:\AllFiles
More...
Walkthroughs
Get Started with VS Code
Customize your editor, learn the basics, and start coding
Learn the Fundamentals
Show welcome page on startup
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
Found 3 items
drwxrwxrwt - hdfs hadoop 0 2025-02-11 02:53 /tmp
drwxrwxrwt - hdfs hadoop 0 2025-02-11 02:52 /user
drwxrwxrwt - hdfs hadoop 0 2025-02-11 02:52 /var
julio@dmc-dev-bdp-15-m:~$ hdfs dfs -mkdir /juliovasquez
julio@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /
Found 4 items
drwxr-xr-x - julio hadoop 0 2025-02-13 01:55 /juliovasquez
drwxrwxrwt - hdfs hadoop 0 2025-02-11 02:53 /tmp
drwxrwxrwt - hdfs hadoop 0 2025-02-11 02:52 /user
drwxrwxrwt - hdfs hadoop 0 2025-02-11 02:52 /var
julio@dmc-dev-bdp-15-m:~$
Activar Windows
Ve a Configuración para activar Windows.

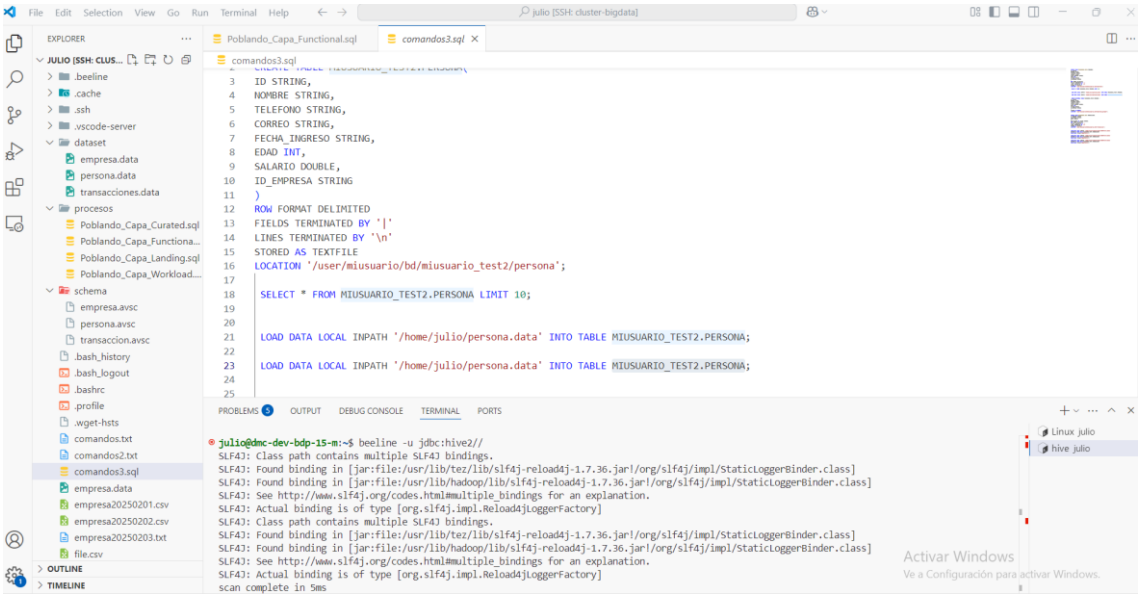
```
julio@dmc-dev-bdp-15-m:~$ ls
persona.data
julio@dmc-dev-bdp-15-m:~$ pwd
/home/julio
julio@dmc-dev-bdp-15-m:~$ hdfs dfs -put /home/julio/persona.data /juliovasquez
julio@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /juliovasquez
Found 8 items
-rw-r--r--   2 julio hadoop          0 2025-02-13 02:23 /juliovasquez/_ERROR
-rw-r--r--   2 julio hadoop          0 2025-02-13 02:22 /juliovasquez/_SUCCESS
-rw-r--r--   2 julio hadoop          0 2025-02-13 02:22 /juliovasquez/archivo_vacio.txt
-rw-r--r--   2 julio hadoop          0 2025-02-13 02:22 /juliovasquez/archivo_vacio_sin_extension
drwxr-xr-x   - julio hadoop          0 2025-02-13 02:00 /juliovasquez/carpeta1
drwxr-xr-x   - julio hadoop          0 2025-02-13 01:58 /juliovasquez/carpeta2
drwxr-xr-x   - julio hadoop          0 2025-02-13 02:02 /juliovasquez/carpeta3
-rw-r--r--   2 julio hadoop       7282 2025-02-13 02:46 /juliovasquez/persona.data
julio@dmc-dev-bdp-15-m:~$
```



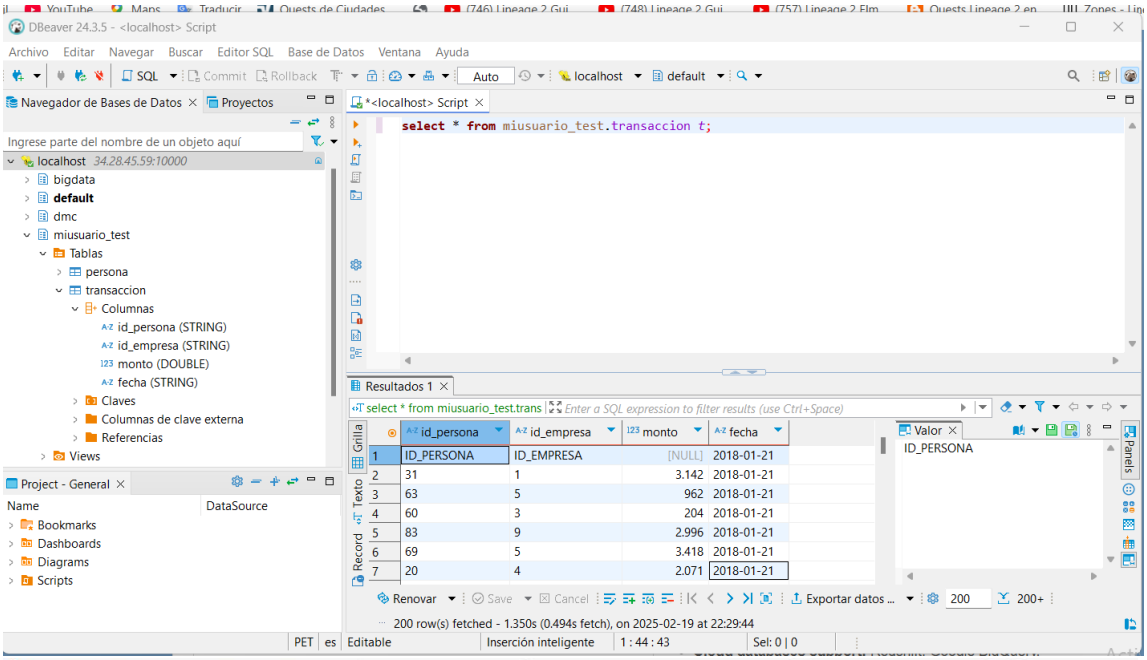


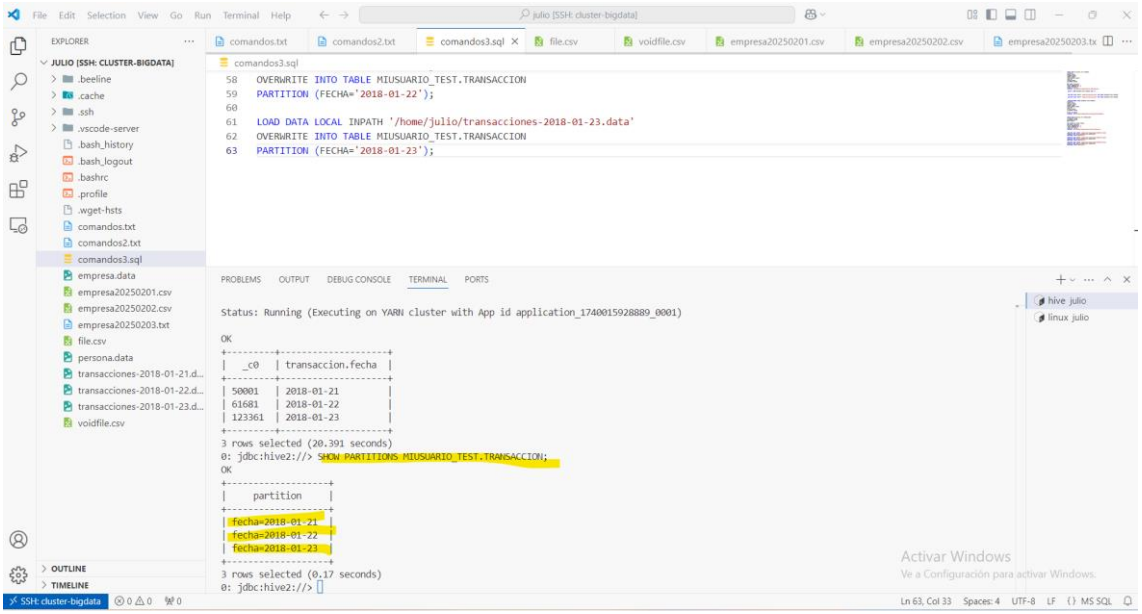
4. Sesión 4

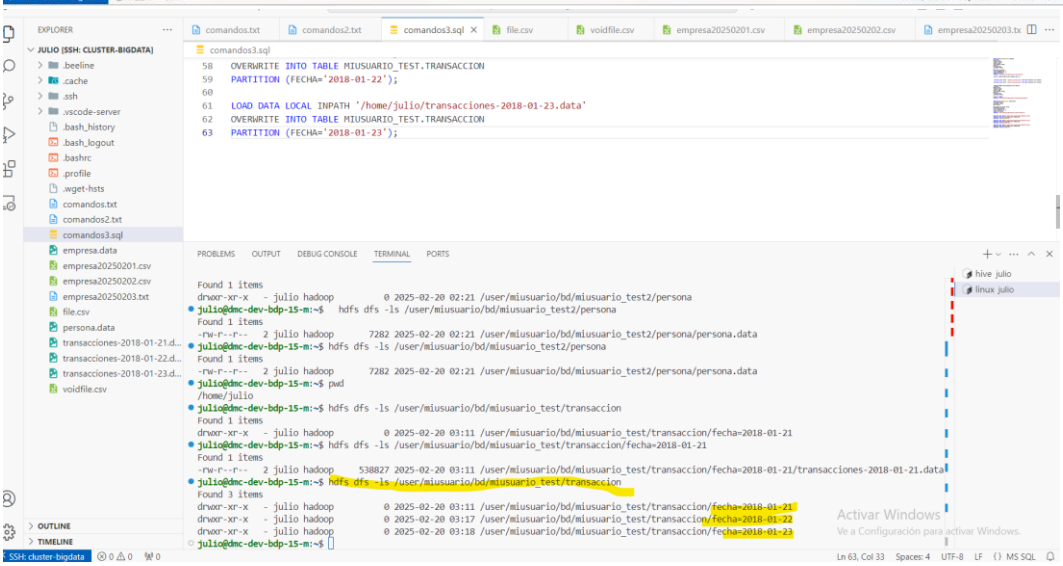

File Edit Selection View Go Run Terminal Help
julio [SSH: cluster-bigdata]
EXPLORER
Poblando_Capa_Functional.sql
comandos3.sql
JULIO [SSH: CLUS...
.beeline
.cache
.ssh
.vscode-server
dataset
empresa.data
persona.data
transacciones.data
procesos
Poblando_Capa_Curated.sql
Poblando_Capa_Functiona...
Poblando_Capa_Landing.sql
Poblando_Capa_Workload....
schema
empresa.avsc
persona.avsc
transaccion.avsc
.bash_history
.bash_logout
.bashrc
.profile
.wget-hsts
comandos.txt
comandos2.txt
comandos3.sql
empresa.data
empresa20250201.csv
empresa20250202.csv
empresa20250203.txt
file.csv
OUTLINE
TIMELINE
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';
SELECT * FROM MIUSUARIO_TEST2.PERSONA LIMIT 10;
LOAD DATA LOCAL INPATH '/home/julio/persona.data' INTO TABLE MIUSUARIO_TEST2.PERSONA;
LOAD DATA LOCAL INPATH '/home/julio/persona.data' INTO TABLE MIUSUARIO_TEST2.PERSONA;
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
bash - julio
julio@dmc-dev-bdp-15-m:~$
Activar Windows
Ve a Configuración para activar Windows.
SSH: cluster-bigdata
Ln 63, Col 33 Spaces: 4 UTF-8 LF MS SQL


File Edit Selection View Go Run Terminal Help
julio [SSH: cluster-bigdata]
EXPLORER
Poblando_Capa_Functional.sql
comandos3.sql
JULIO [SSH: CLUS...
comandos3.sql
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';
SELECT * FROM MIUSUARIO_TEST2.PERSONA LIMIT 10;
LOAD DATA LOCAL INPATH '/home/julio/persona.data' INTO TABLE MIUSUARIO_TEST2.PERSONA;
LOAD DATA LOCAL INPATH '/home/julio/persona.data' INTO TABLE MIUSUARIO_TEST2.PERSONA;
PROBLEMS 5 OUTPUT DEBUG CONSOLE TERMINAL PORTS
Linux julio
hive julio
julio@dmc-dev-bdp-15-m:~$ beeline -u jdbc:hive2//
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/tez/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/tez/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
scan complete in 5ms
Activar Windows
Ve a Configuración para activar Windows.

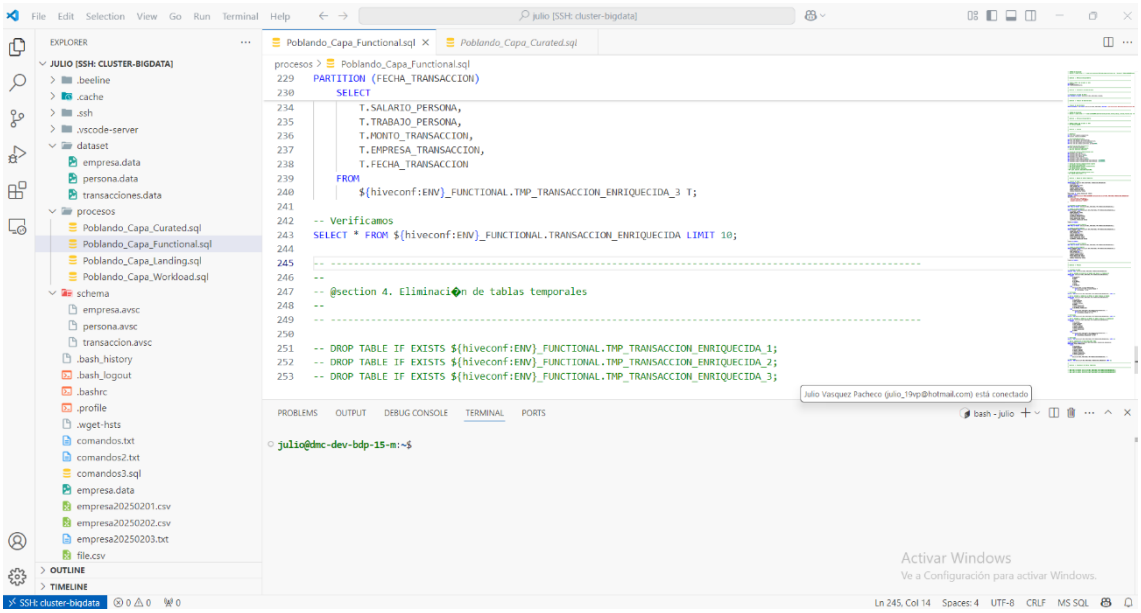
5. Sesión 5

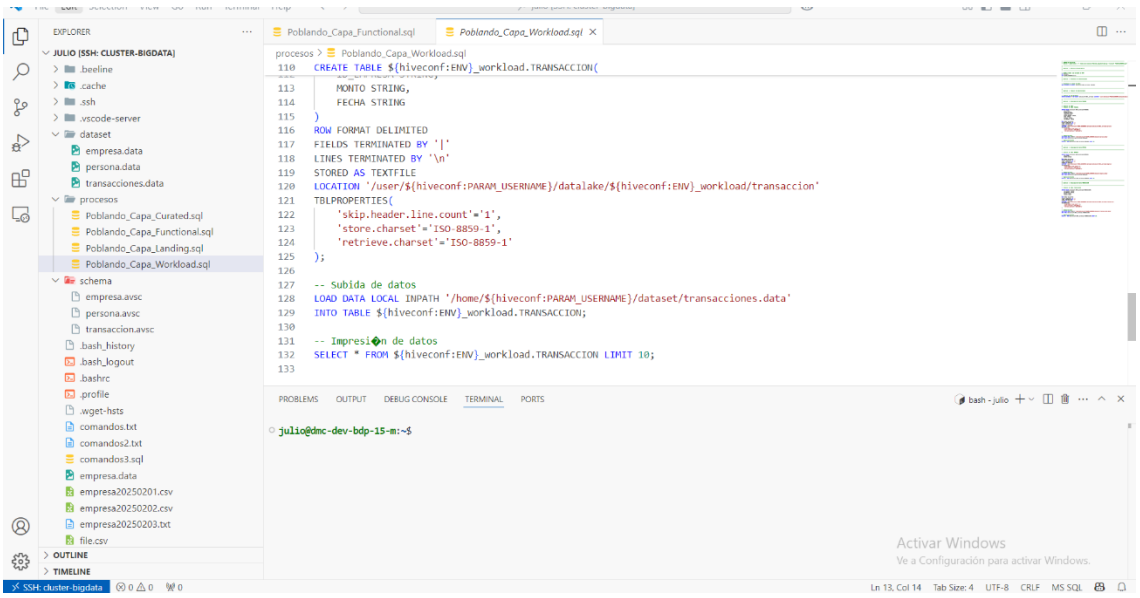

DBeaver 24.3.5 - <localhost> Script
Archivo Editar Navegar Buscar Editor SQL Base de Datos Ventana Ayuda
Navegador de Bases de Datos
Proyectos
localhost 34.28.45.59:10000
bigdata
default
dmc
miusuario_test
Tablas
persona
transaccion
Columnas
id_persona (STRING)
id_empresa (STRING)
monto (DOUBLE)
fecha (STRING)
Claves
Columnas de clave externa
Referencias
Views
Project - General
Name
DataSource
Bookmarks
Dashboards
Diagrams
Scripts
*<localhost> Script
select * from miusuario_test.transaccion t;
Resultados 1
select * from miusuario_test.trans
Enter a SQL expression to filter results (use Ctrl+Space)
id_persona
id_empresa
monto
fecha
1 ID_PERSONA ID_EMPRESA [NULL] 2018-01-21
2 31 1 3.142 2018-01-21
3 63 5 962 2018-01-21
4 60 3 204 2018-01-21
5 83 9 2.996 2018-01-21
6 69 5 3.418 2018-01-21
7 20 4 2.071 2018-01-21
Valor
ID_PERSONA
Renovar
Save
Cancel
Exportar datos ...
200 row(s) fetched - 1.350s (0.494s fetch), on 2025-02-19 at 22:29:44
PET es Editable Inserción inteligente 1 : 44 : 43 Sel: 0 | 0

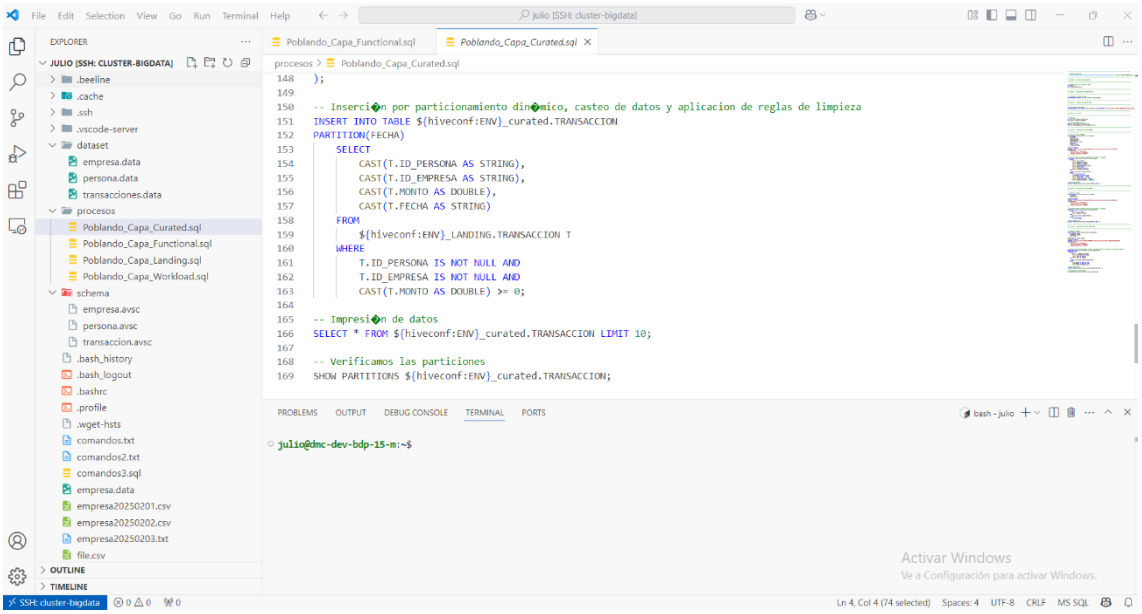

comandos3.sql
58 OVERWRITE INTO TABLE MIUSUARIO_TEST.TRANSACCION
59 PARTITION (FECHA='2018-01-22');
60
61 LOAD DATA LOCAL INPATH '/home/julio/transacciones-2018-01-23.data'
62 OVERWRITE INTO TABLE MIUSUARIO_TEST.TRANSACCION
63 PARTITION (FECHA='2018-01-23');
Status: Running (Executing on YARN cluster with App id application_1740015928889_0001)
OK
_c0	transaccion.fecha
50001	2018-01-21
61681	2018-01-22
123361	2018-01-23
3 rows selected (20.391 seconds)
0: jdbc:hive2://> SHOW PARTITIONS MIUSUARIO_TEST.TRANSACCION;
OK
partition
fecha=2018-01-21
fecha=2018-01-22
fecha=2018-01-23
3 rows selected (0.17 seconds)
0: jdbc:hive2://>
Activar Windows

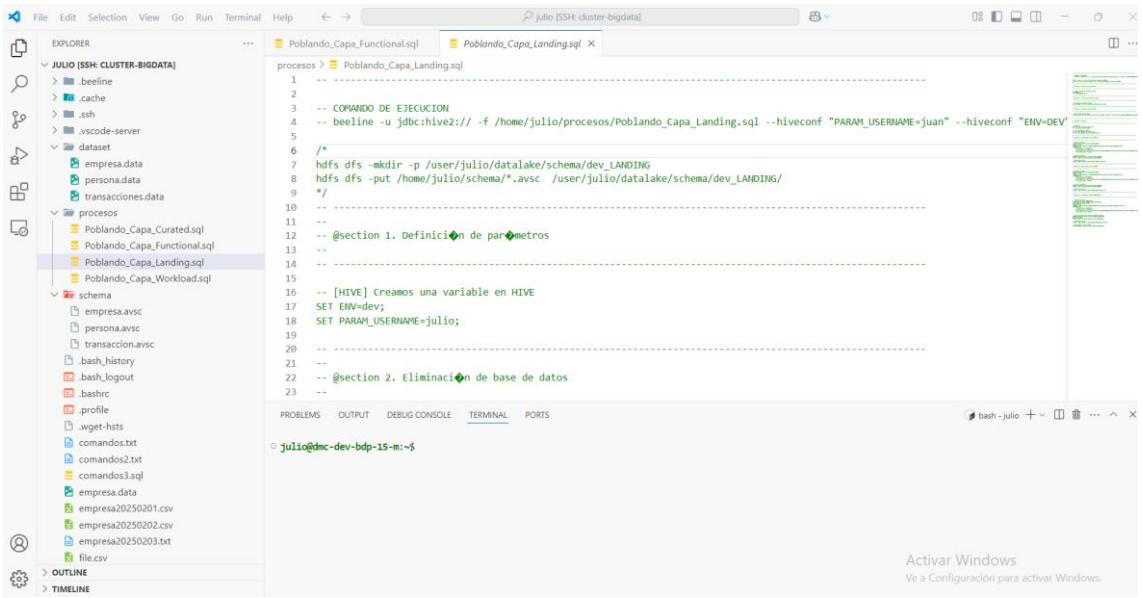

comandos3.sql
58 OVERWRITE INTO TABLE MIUSUARIO_TEST.TRANSACCION
59 PARTITION (FECHA='2018-01-22');
61 LOAD DATA LOCAL INPATH '/home/julio/transacciones-2018-01-23.data'
62 OVERWRITE INTO TABLE MIUSUARIO_TEST.TRANSACCION
63 PARTITION (FECHA='2018-01-23');
julio@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/miusuario/bd/miusuario_test2/persona
julio@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/miusuario/bd/miusuario_test2/persona/persona.data
julio@dmc-dev-bdp-15-m:~$ pwd
/home/julio
julio@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/miusuario/bd/miusuario_test/transaccion
julio@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/miusuario/bd/miusuario_test/transaccion/fecha=2018-01-21
julio@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/miusuario/bd/miusuario_test/transaccion/fecha=2018-01-21/transacciones-2018-01-21.data
julio@dmc-dev-bdp-15-m:~$ hdfs dfs -ls /user/miusuario/bd/miusuario_test/transaccion
drwxr-xr-x - julio hadoop 0 2025-02-20 03:11 /user/miusuario/bd/miusuario_test/transaccion/fecha=2018-01-21
drwxr-xr-x - julio hadoop 0 2025-02-20 03:17 /user/miusuario/bd/miusuario_test/transaccion/fecha=2018-01-22
drwxr-xr-x - julio hadoop 0 2025-02-20 03:18 /user/miusuario/bd/miusuario_test/transaccion/fecha=2018-01-23
Activar Windows

6. Sesión 6

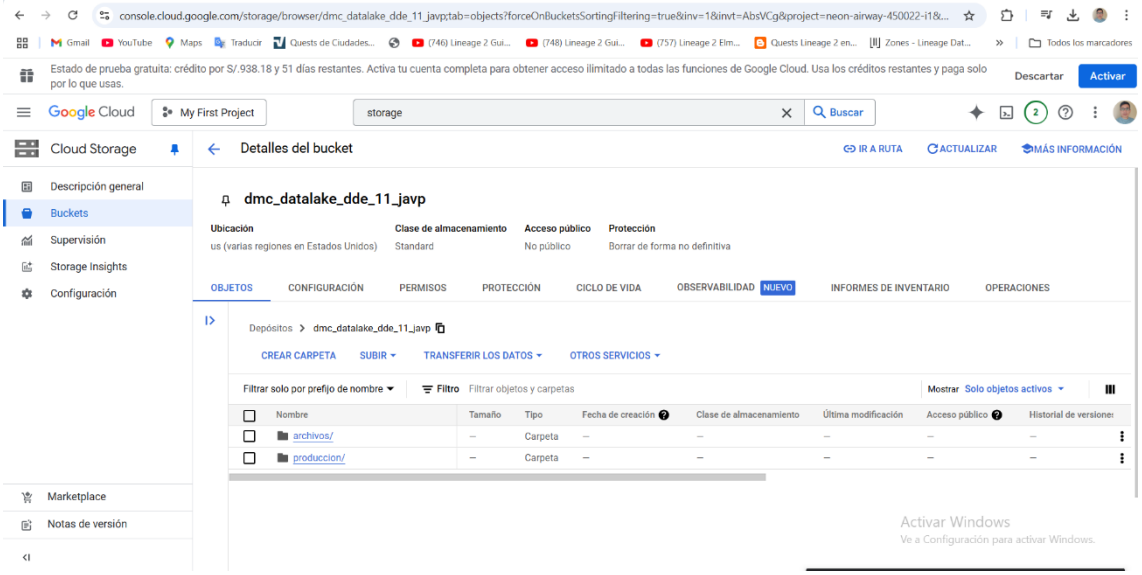

Poblando_Capa_Functional.sql
Poblando_Capa_Curated.sql
procesos > Poblando_Capa_Functional.sql
PARTITION (FECHA_TRANSACCION)
SELECT
T.SALARIO_PERSONA,
T.TRABAJO_PERSONA,
T.MONTO_TRANSACCION,
T.EMPRESA_TRANSACCION,
T.FECHA_TRANSACCION
FROM
${hiveconf:ENV}_FUNCTIONAL.TMP_TRANSACCION_ENRIQUECIDA_3 T;
-- Verificamos
SELECT * FROM ${hiveconf:ENV}_FUNCTIONAL.TRANSACCION_ENRIQUECIDA LIMIT 10;
-- @section 4. Eliminaci&#65533;n de tablas temporales
-- DROP TABLE IF EXISTS ${hiveconf:ENV}_FUNCTIONAL.TMP_TRANSACCION_ENRIQUECIDA_1;
-- DROP TABLE IF EXISTS ${hiveconf:ENV}_FUNCTIONAL.TMP_TRANSACCION_ENRIQUECIDA_2;
-- DROP TABLE IF EXISTS ${hiveconf:ENV}_FUNCTIONAL.TMP_TRANSACCION_ENRIQUECIDA_3;
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
julio@dmc-dev-bdp-15-m:~$
Activar Windows

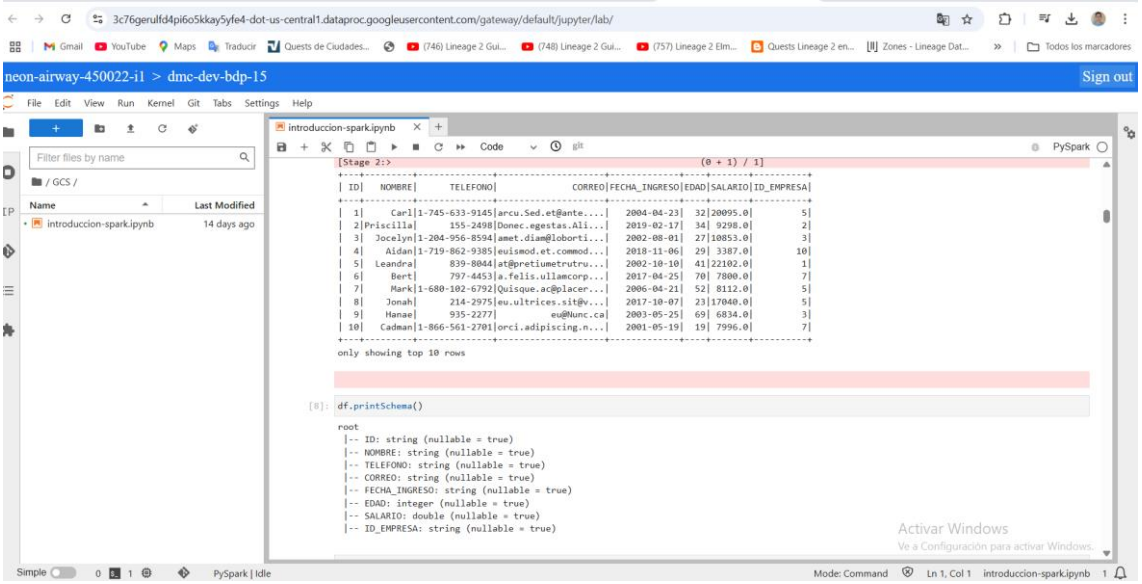

Poblando_Capa_Functional.sql
Poblando_Capa_Workload.sql
procesos > Poblando_Capa_Workload.sql
CREATE TABLE ${hiveconf:ENV}_workload.TRANSACCION(
MONTO STRING,
FECHA STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE
LOCATION '/user/${hiveconf:PARAM_USERNAME}/datalake/${hiveconf:ENV}_workload/transaccion'
TBLPROPERTIES(
'skip.header.line.count'='1',
'store.charset'='ISO-8859-1',
'retrieve.charset'='ISO-8859-1'
);
-- Subida de datos
LOAD DATA LOCAL INPATH '/home/${hiveconf:PARAM_USERNAME}/dataset/transacciones.data'
INTO TABLE ${hiveconf:ENV}_workload.TRANSACCION;
-- Impresi&#65533;n de datos
SELECT * FROM ${hiveconf:ENV}_workload.TRANSACCION LIMIT 10;
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
julio@dmc-dev-bdp-15-m:~$
Activar Windows

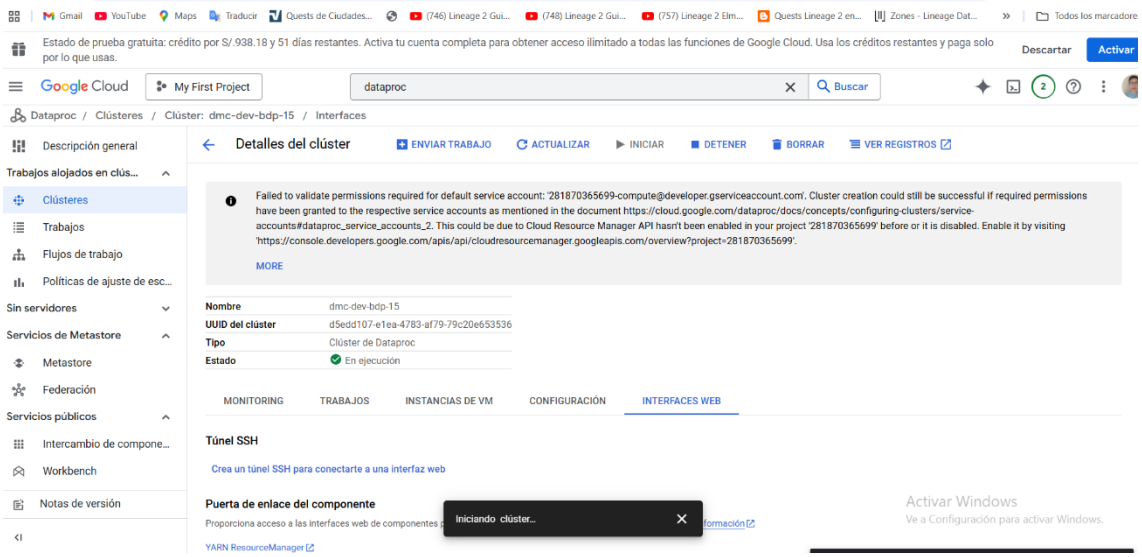

File Edit Selection View Go Run Terminal Help
julio [SSH: cluster-bigdata]
EXPLORER
JULIO [SSH: CLUSTER-BIGDATA]
Poblando_Capa_Functional.sql
Poblando_Capa_Curated.sql
procesos > Poblando_Capa_Curated.sql
.beeline
.cache
.ssh
.vscode-server
dataset
empresa.data
persona.data
transacciones.data
procesos
Poblando_Capa_Curated.sql
Poblando_Capa_Functional.sql
Poblando_Capa_Landing.sql
Poblando_Capa_Workload.sql
schema
empresa.avsc
persona.avsc
transaccion.avsc
.bash_history
.bash_logout
.bashrc
.profile
.wget-hsts
comandos.txt
comandos2.txt
comandos3.sql
empresa.data
empresa20250201.csv
empresa20250202.csv
empresa20250203.txt
file.csv
OUTLINE
TIMELINE
);
-- Inserci&#65533;n por particionamiento din&#65533;mico, casteo de datos y aplicacion de reglas de limpieza
INSERT INTO TABLE ${hiveconf:ENV}_curated.TRANSACCION
PARTITION(FECHA)
SELECT
CAST(T.ID_PERSONA AS STRING),
CAST(T.ID_EMPRESA AS STRING),
CAST(T.MONTO AS DOUBLE),
CAST(T.FECHA AS STRING)
FROM
${hiveconf:ENV}_LANDING.TRANSACCION T
WHERE
T.ID_PERSONA IS NOT NULL AND
T.ID_EMPRESA IS NOT NULL AND
CAST(T.MONTO AS DOUBLE) >= 0;
-- Impresi&#65533;n de datos
SELECT * FROM ${hiveconf:ENV}_curated.TRANSACCION LIMIT 10;
-- Verificamos las particiones
SHOW PARTITIONS ${hiveconf:ENV}_curated.TRANSACCION;
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
bash - julio
julio@dmc-dev-bdp-15-m:~$
Activar Windows
Ve a Configuración para activar Windows.
SSH: cluster-bigdata
Ln 4, Col 4 (74 selected) Spaces: 4 UTF-8 CRLF MS SQL

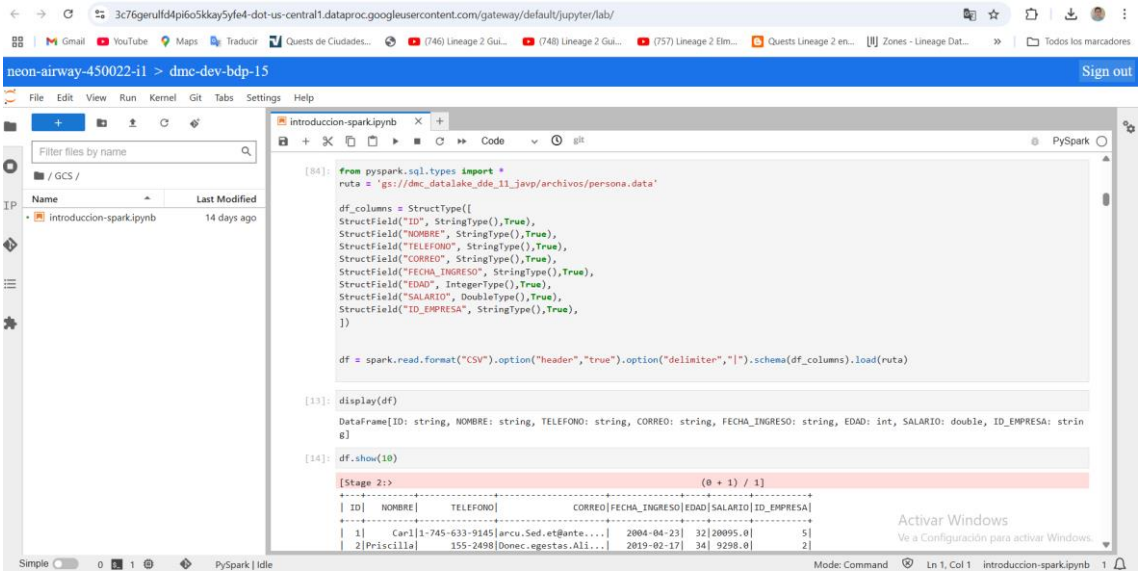

File Edit Selection View Go Run Terminal Help
julio [SSH: cluster-bigdata]
EXPLORER
JULIO [SSH: CLUSTER-BIGDATA]
Poblando_Capa_Functional.sql
Poblando_Capa_Landing.sql
procesos > Poblando_Capa_Landing.sql
.beeline
.cache
.ssh
.vscode-server
dataset
empresa.data
persona.data
transacciones.data
procesos
Poblando_Capa_Curated.sql
Poblando_Capa_Functional.sql
Poblando_Capa_Landing.sql
Poblando_Capa_Workload.sql
schema
empresa.avsc
persona.avsc
transaccion.avsc
.bash_history
.bash_logout
.bashrc
.profile
.wget-hsts
comandos.txt
comandos2.txt
comandos3.sql
empresa.data
empresa20250201.csv
empresa20250202.csv
empresa20250203.txt
file.csv
OUTLINE
TIMELINE
-- COMANDO DE EJECUCION
-- beeline -u jdbc:hive2:// -f /home/julio/procesos/Poblando_Capa_Landing.sql --hiveconf "PARAM_USERNAME=juan" --hiveconf "ENV=DEV"
/*
hdfs dfs -mkdir -p /user/julio/datalake/schema/dev_LANDING
hdfs dfs -put /home/julio/schema/*.avsc /user/julio/datalake/schema/dev_LANDING/
*/
-- @section 1. Definici&#65533;n de par&#65533;metros
-- [HIVE] Creamos una variable en HIVE
SET ENV=dev;
SET PARAM_USERNAME=julio;
-- @section 2. Eliminaci&#65533;n de base de datos
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
bash - julio
julio@dmc-dev-bdp-15-m:~$
Activar Windows
Ve a Configuración para activar Windows.

7. Sesión 7

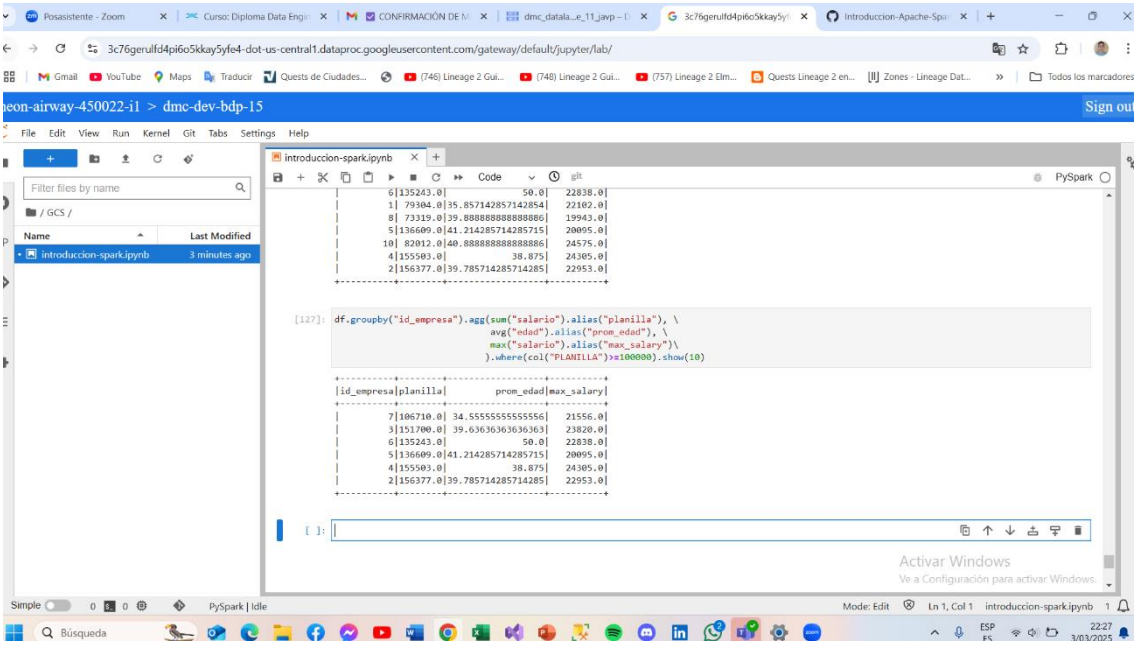

Cloud Storage
Detalles del bucket
dmc_datalake_dde_11_javp
Ubicación
us (varias regiones en Estados Unidos)
Clase de almacenamiento
Standard
Acceso público
No público
Protección
Borrar de forma no definitiva
OBJETOS
CONFIGURACIÓN
PERMISOS
PROTECCIÓN
CICLO DE VIDA
OBSERVABILIDAD
INFORMES DE INVENTARIO
OPERACIONES
CREAR CARPETA
SUBIR
TRANSFERIR LOS DATOS
OTROS SERVICIOS
archivos/
produccion/

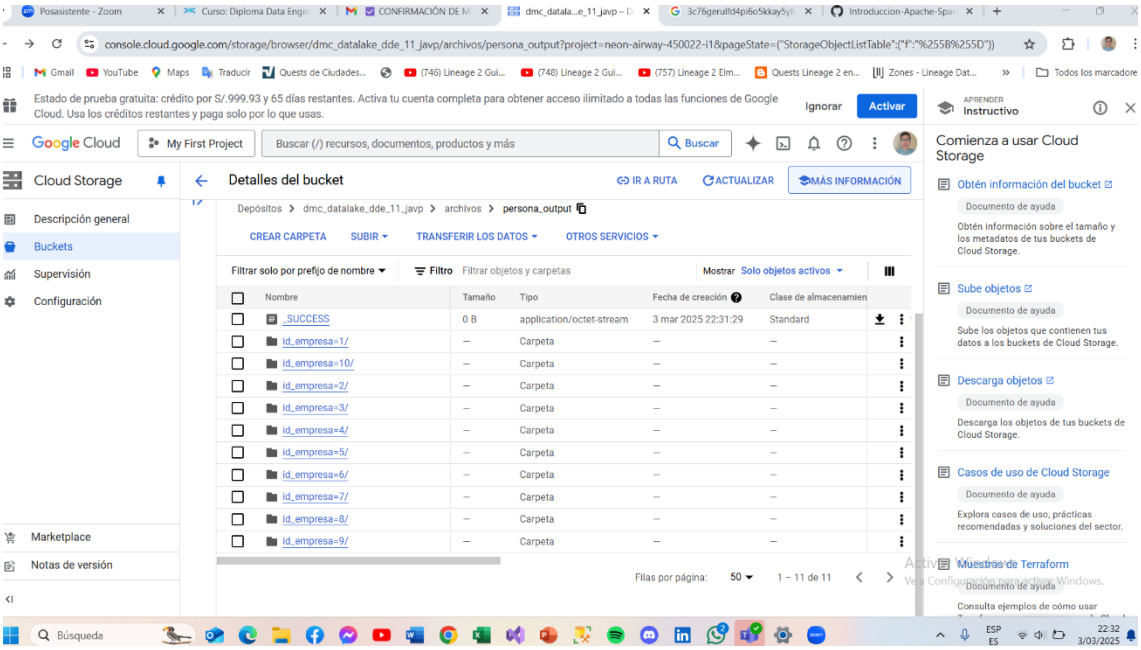

neon-airway-450022-i1 > dmc-dev-bdp-15
Sign out
introduccion-spark.ipynb
PySpark
only showing top 10 rows
df.printSchema()
root
|-- ID: string (nullable = true)
|-- NOMBRE: string (nullable = true)
|-- TELEFONO: string (nullable = true)
|-- CORREO: string (nullable = true)
|-- FECHA_INGRESO: string (nullable = true)
|-- EDAD: integer (nullable = true)
|-- SALARIO: double (nullable = true)
|-- ID_EMPRESA: string (nullable = true)


Dataproc
Clústeres
Interfaces
Detalles del clúster
ENVIAR TRABAJO
ACTUALIZAR
INICIAR
DETENER
BORRAR
VER REGISTROS
Nombre
dmc-dev-bdp-15
UUID del clúster
d5edd107-e1ea-4783-af79-79c20e653536
Tipo
Clúster de Dataproc
Estado
En ejecución
MONITORING
TRABAJOS
INSTANCIAS DE VM
CONFIGURACIÓN
INTERFACES WEB
Túnel SSH
Puerta de enlace del componente
Iniciando clúster...

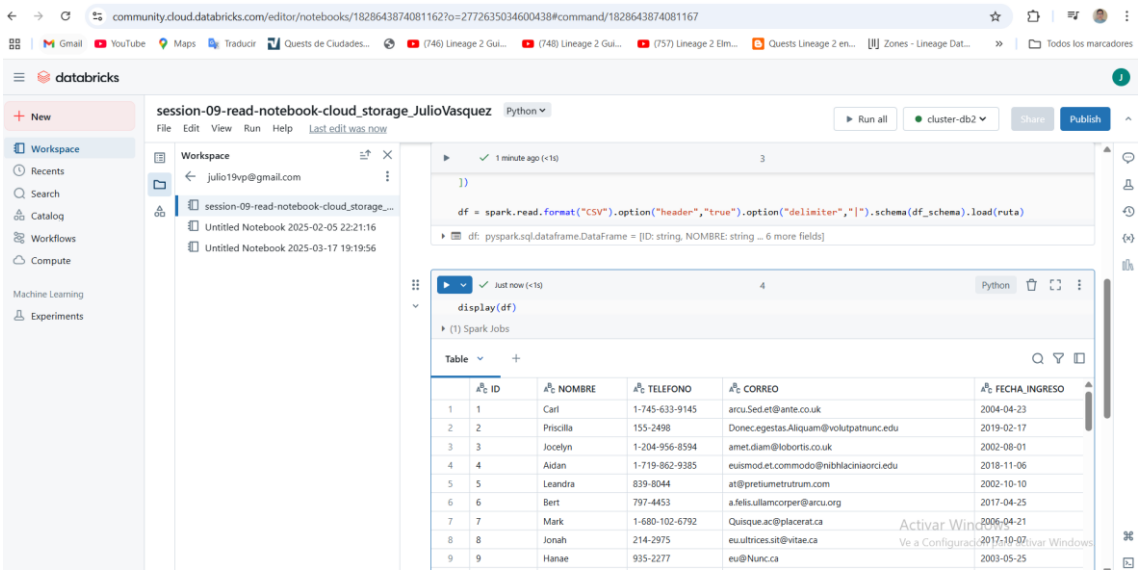

3c76gerulfd4pi6o5kkay5yfe4-dot-us-central1.dataproc.googleusercontent.com/gateway/default/jupyter/lab/
neon-airway-450022-i1 > dmc-dev-bdp-15
Sign out
File Edit View Run Kernel Git Tabs Settings Help
introduccion-spark.ipynb
Code
PySpark
Filter files by name
/ GCS /
Name
Last Modified
introduccion-spark.ipynb
14 days ago
[84]:
from pyspark.sql.types import *
ruta = 'gs://dmc_datalake_dde_11_javp/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 = spark.read.format("CSV").option("header","true").option("delimiter","|").schema(df_columns).load(ruta)
[13]:
display(df)
DataFrame[ID: string, NOMBRE: string, TELEFONO: string, CORREO: string, FECHA_INGRESO: string, EDAD: int, SALARIO: double, ID_EMPRESA: strin
g]
[14]:
df.show(10)
[Stage 2:> (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
Activar Windows
Ve a Configuración para activar Windows.
Simple
PySpark | Idle
Mode: Command
Ln 1, Col 1
introduccion-spark.ipynb

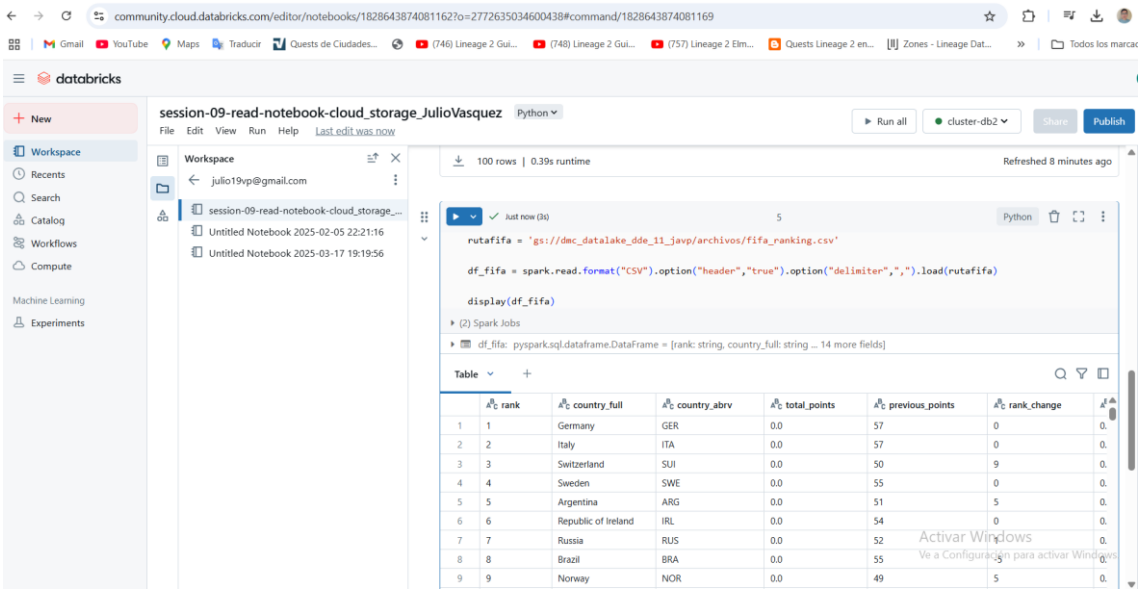
8. Sesión 8


3c76gerulfd4pi6o5kkay5yfe4-dot-us-central1.dataproc.googleusercontent.com/gateway/default/jupyter/lab/
neon-airway-450022-i1 > dmc-dev-bdp-15
Sign out
File Edit View Run Kernel Git Tabs Settings Help
introduccion-spark.ipynb
Code
PySpark
Filter files by name
/ GCS /
Name
Last Modified
introduccion-spark.ipynb
3 minutes ago
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
[127]:
df.groupby("id_empresa").agg(sum("salario").alias("planilla"), \
avg("edad").alias("prom_edad"), \
max("salario").alias("max_salary")\
).where(col("PLANTILLA")>=100000).show(10)
id_empresa	planilla	prom_edad	max_salary
7	106710.0	34.55555555555556	21556.0
3	151700.0	39.63636363636363	23820.0
6	135243.0	50.0	22838.0
5	136609.0	41.214285714285715	20095.0
4	155503.0	38.875	24305.0
2	156377.0	39.785714285714285	22953.0
[ ]:
Activar Windows
Ve a Configuración para activar Windows.
Simple
PySpark | Idle
Mode: Edit
Ln 1, Col 1
introduccion-spark.ipynb
Búsqueda
ESP
22:27
3/03/2025

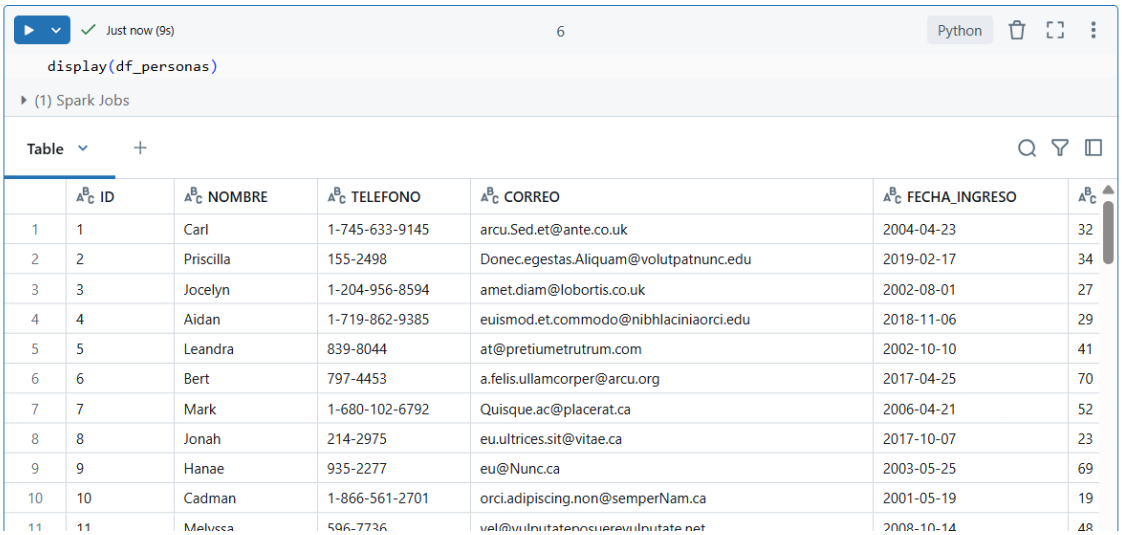

console.cloud.google.com/storage/browser/dmc_datalake_dde_11_javp/archivos/persona_output?project=neon-airway-450022-i1&pageState=("StorageObjectListTable":("f":"%255B%255D"))
Estado de prueba gratuita: crédito por S/.999.93 y 65 días restantes. Activa tu cuenta completa para obtener acceso ilimitado a todas las funciones de Google Cloud. Usa los créditos restantes y paga solo por lo que usas.
Ignorar
Activar
Google Cloud
My First Project
Buscar (/) recursos, documentos, productos y más
Buscar
Cloud Storage
Descripción general
Buckets
Supervisión
Configuración
Marketplace
Notas de versión
Detalles del bucket
IR A RUTA
ACTUALIZAR
MÁS INFORMACIÓN
Depósitos > dmc_datalake_dde_11_javp > archivos > persona_output
CREAR CARPETA
SUBIR
TRANSFERIR LOS DATOS
OTROS SERVICIOS
Filtrar solo por prefijo de nombre
Filtro
Filtrar objetos y carpetas
Mostrar
Solo objetos activos
Nombre
Tamaño
Tipo
Fecha de creación
Clase de almacenamien
_SUCCESS
0 B
application/octet-stream
3 mar 2025 22:31:29
Standard
id_empresa=1/
id_empresa=10/
id_empresa=2/
id_empresa=3/
id_empresa=4/
id_empresa=5/
id_empresa=6/
id_empresa=7/
id_empresa=8/
id_empresa=9/
Carpeta
Filas por página:
50
1 – 11 de 11
APRENDER
Instructivo
Comienza a usar Cloud Storage
Obtén información del bucket
Documento de ayuda
Obtén información sobre el tamaño y los metadatos de tus buckets de Cloud Storage.
Sube objetos
Sube los objetos que contienen tus datos a los buckets de Cloud Storage.
Descarga objetos
Descarga los objetos de tus buckets de Cloud Storage.
Casos de uso de Cloud Storage
Explora casos de uso, prácticas recomendadas y soluciones del sector.
Muestras de Terraform
Consulta ejemplos de cómo usar
Búsqueda

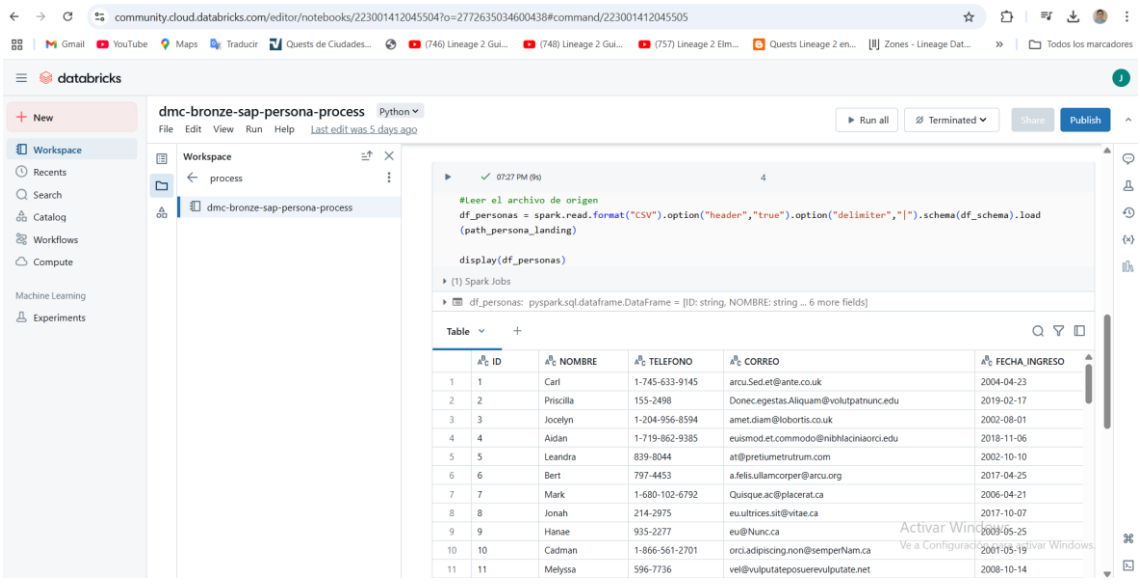
9. Sesión 9

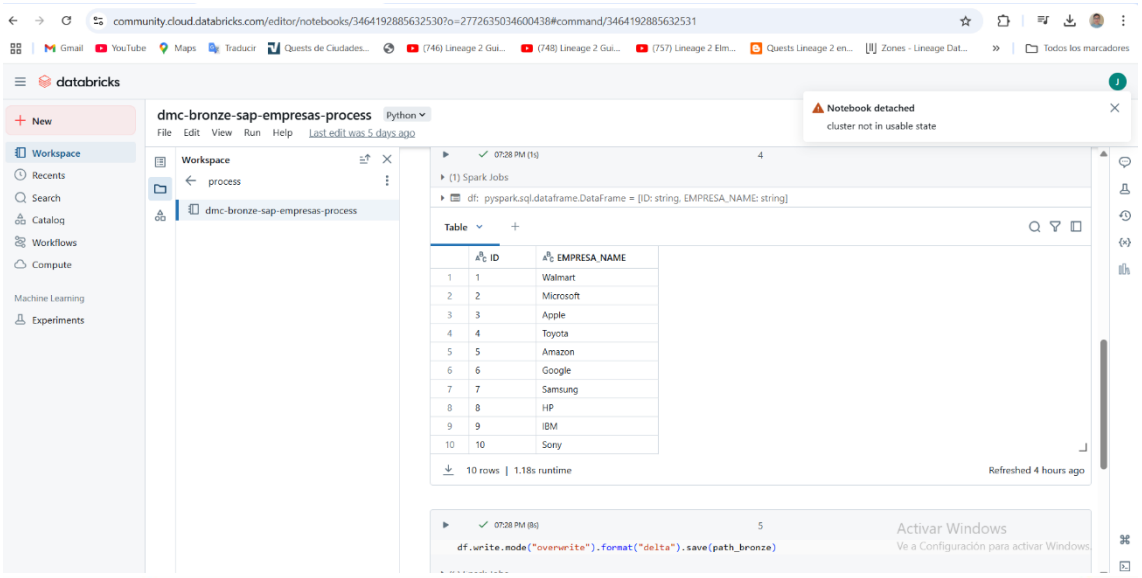

community.cloud.databricks.com/editor/notebooks/1828643874081162?o=2772635034600438#command/1828643874081167
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
session-09-read-notebook-cloud_storage_JulioVasquez
Python
File Edit View Run Help Last edit was now
Run all
cluster-db2
Share
Publish
Workspace
julio19vp@gmail.com
session-09-read-notebook-cloud_storage_...
Untitled Notebook 2025-02-05 22:21:16
Untitled Notebook 2025-03-17 19:19:56
df = spark.read.format("CSV").option("header","true").option("delimiter","|").schema(df_schema).load(ruta)
df: pyspark.sql.dataframe.DataFrame = [ID: string, NOMBRE: string ... 6 more fields]
display(df)
(1) Spark Jobs
Table
ID
NOMBRE
TELEFONO
CORREO
FECHA_INGRESO
1 Carl 1-745-633-9145 arcu.Sed.et@ante.co.uk 2004-04-23
2 Priscilla 155-2498 Donec.egestas.Aliquam@volutpatnunc.edu 2019-02-17
3 Jocelyn 1-204-956-8594 amet.diam@lobortis.co.uk 2002-08-01
4 Aidan 1-719-862-9385 euismod.et.commodo@nibhlaciniaorci.edu 2018-11-06
5 Leandra 839-8044 at@pretiumetrutrum.com 2002-10-10
6 Bert 797-4453 a.felis.ullamcorper@arcu.org 2017-04-25
7 Mark 1-680-102-6792 Quisque.ac@placerat.ca 2006-04-21
8 Jonah 214-2975 eu.ultrices.sit@vitae.ca 2017-10-07
9 Hanae 935-2277 eu@Nunc.ca 2003-05-25

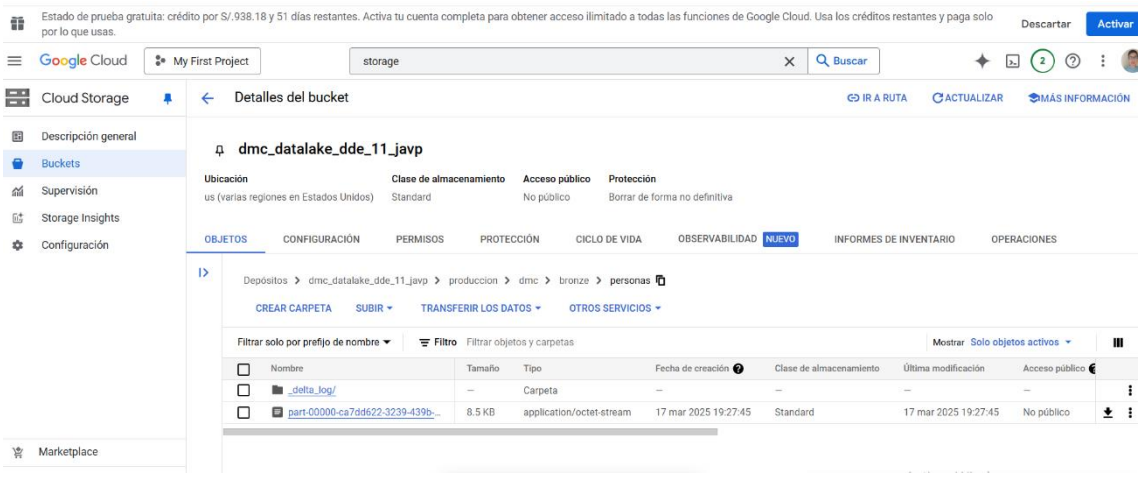

community.cloud.databricks.com/editor/notebooks/1828643874081162?o=2772635034600438#command/1828643874081169
databricks
session-09-read-notebook-cloud_storage_JulioVasquez
Python
File Edit View Run Help Last edit was now
Run all
cluster-db2
Share
Publish
Workspace
julio19vp@gmail.com
session-09-read-notebook-cloud_storage_...
Untitled Notebook 2025-02-05 22:21:16
Untitled Notebook 2025-03-17 19:19:56
100 rows | 0.39s runtime
Refreshed 8 minutes ago
Just now (3s)
rutafifa = 'gs://dmc_datalake_dde_11_javp/archivos/fifa_ranking.csv'
df_fifa = spark.read.format("CSV").option("header","true").option("delimiter",",").load(rutafifa)
display(df_fifa)
(2) Spark Jobs
df_fifa: pyspark.sql.dataframe.DataFrame = [rank: string, country_full: string ... 14 more fields]
Table
rank
country_full
country_abrv
total_points
previous_points
rank_change
1 Germany GER 0.0 57 0
2 Italy ITA 0.0 57 0
3 Switzerland SUI 0.0 50 9
4 Sweden SWE 0.0 55 0
5 Argentina ARG 0.0 51 5
6 Republic of Ireland IRL 0.0 54 0
7 Russia RUS 0.0 52 5
8 Brazil BRA 0.0 55 5
9 Norway NOR 0.0 49 5
Activar Windows

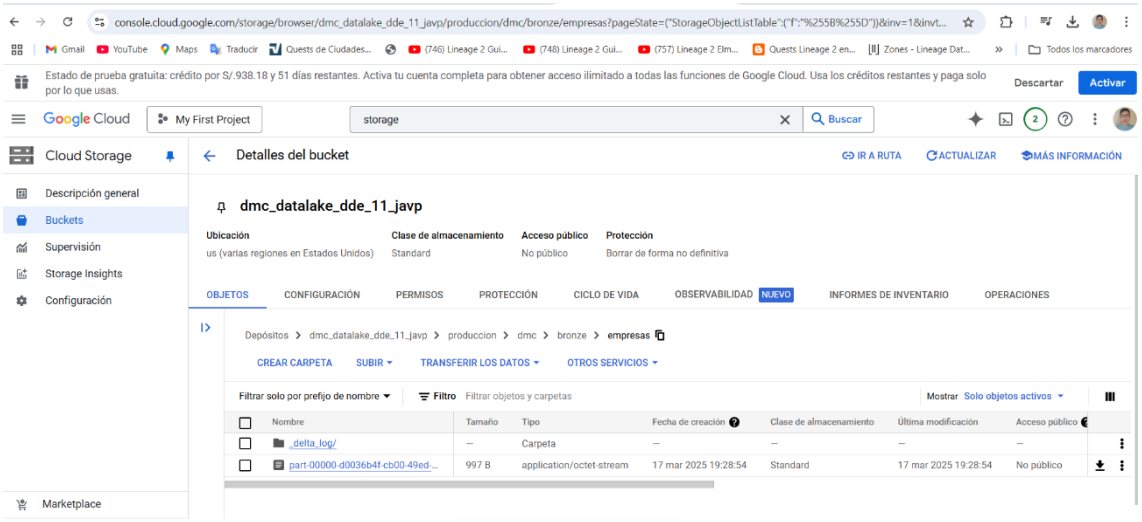
10.Sesión 10


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

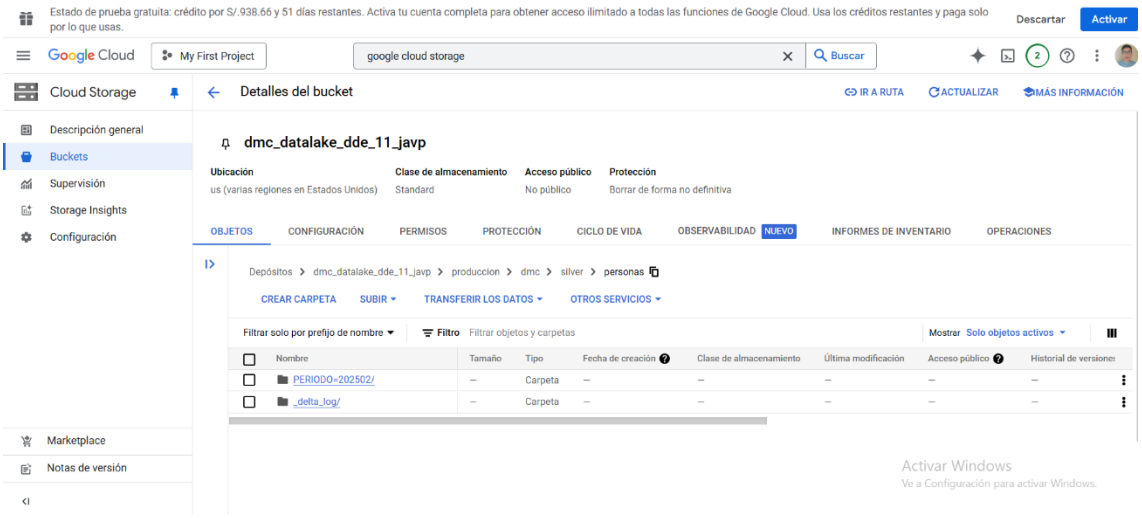

community.cloud.databricks.com/editor/notebooks/2230014120455042o=2772635034600438#command/2230014120455505
databricks
dmc-bronze-sap-persona-process
Python
File Edit View Run Help Last edit was 5 days ago
Run all
Terminated
Share
Publish
Workspace
process
dmc-bronze-sap-persona-process
07:27 PM (9s)
#Leer el archivo de origen
df_personas = spark.read.format("CSV").option("header","true").option("delimiter","|").schema(df_schema).load(path_persona_landing)
display(df_personas)
(1) Spark Jobs
df_personas: pyspark.sql.dataframe.DataFrame = [ID: string, NOMBRE: string ... 6 more fields]
Table
ID
NOMBRE
TELEFONO
CORREO
FECHA_INGRESO
1 Carl 1-745-633-9145 arcu.Sed.et@ante.co.uk 2004-04-23
2 Priscilla 155-2498 Donec.egestas.Aliquam@volutpatnunc.edu 2019-02-17
3 Jocelyn 1-204-956-8594 amet.diam@lobortis.co.uk 2002-08-01
4 Aidan 1-719-862-9385 euismod.et.commodo@nibhlaciniaorci.edu 2018-11-06
5 Leandra 839-8044 at@pretiumetrutrum.com 2002-10-10
6 Bert 797-4453 a.felis.ullamcorper@arcu.org 2017-04-25
7 Mark 1-680-102-6792 Quisque.ac@placerat.ca 2006-04-21
8 Jonah 214-2975 eu.ultrices.sit@vitae.ca 2017-10-07
9 Hanae 935-2277 eu@Nunc.ca 2003-05-25
10 Cadman 1-866-561-2701 orci.adipiscing.non@semperNam.ca 2001-05-19
11 Melyssa 596-7736 vel@vulputateposuerevulputate.net 2008-10-14
Activar Windows

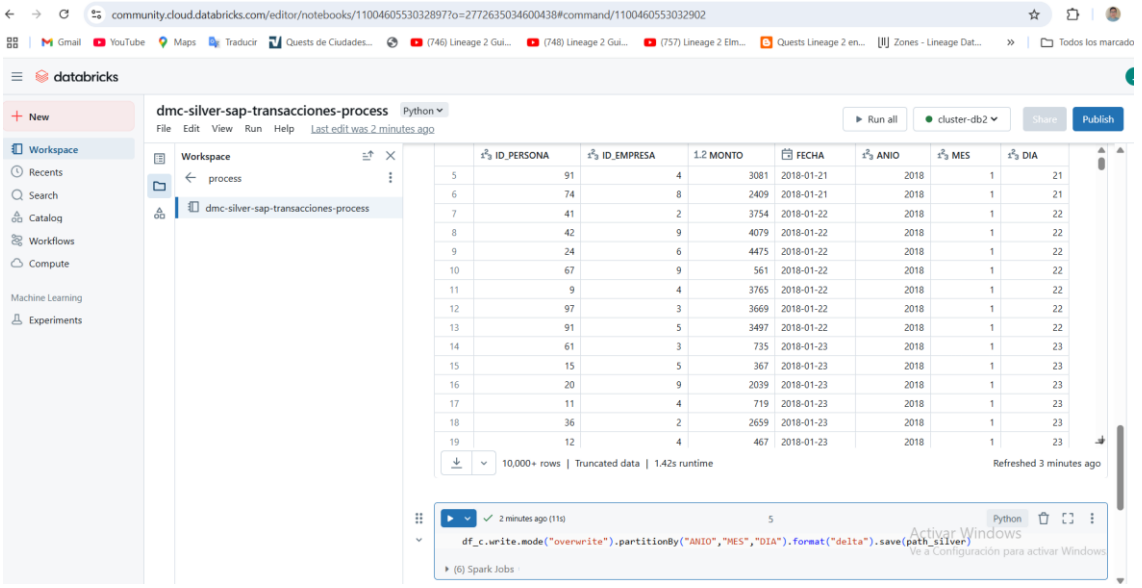

community.cloud.databricks.com/editor/notebooks/3464192885632530?o=2772635034600438#command/3464192885632531
databricks
dmc-bronze-sap-empresas-process
Notebook detached
cluster not in usable state
Workspace
process
dmc-bronze-sap-empresas-process
(1) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [ID: string, EMPRESA_NAME: string]
ID
EMPRESA_NAME
1 Walmart
2 Microsoft
3 Apple
4 Toyota
5 Amazon
6 Google
7 Samsung
8 HP
9 IBM
10 Sony
10 rows | 1.18s runtime
Refreshed 4 hours ago
df.write.mode("overwrite").format("delta").save(path_bronze)
Activar Windows

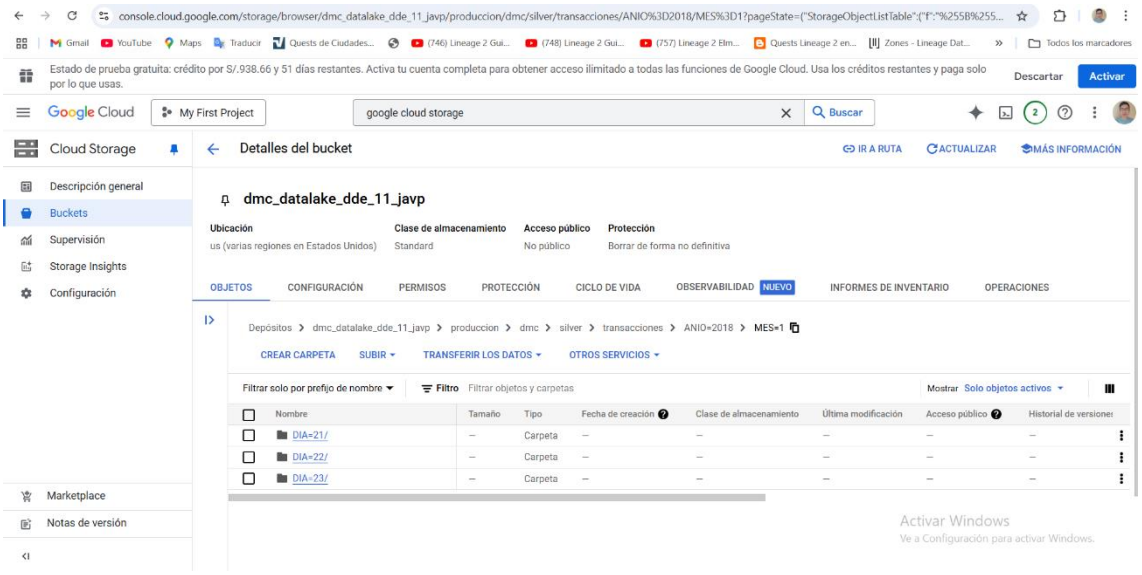

Estado de prueba gratuita: crédito por S/.938.18 y 51 días restantes. Activa tu cuenta completa para obtener acceso ilimitado a todas las funciones de Google Cloud. Usa los créditos restantes y paga solo por lo que usas.
Descartar
Activar
Google Cloud
My First Project
storage
Buscar
Cloud Storage
Detalles del bucket
IR A RUTA
ACTUALIZAR
MÁS INFORMACIÓN
dmc_datalake_dde_11_javp
Ubicación
us (varias regiones en Estados Unidos)
Clase de almacenamiento
Standard
Acceso público
No público
Protección
Borrar de forma no definitiva
OBJETOS
CONFIGURACIÓN
PERMISOS
PROTECCIÓN
CICLO DE VIDA
OBSERVABILIDAD
NUEVO
INFORMES DE INVENTARIO
OPERACIONES
Depósitos > dmc_datalake_dde_11_javp > produccion > dmc > bronze > personas
CREAR CARPETA
SUBIR
TRANSFERIR LOS DATOS
OTROS SERVICIOS
Filtrar solo por prefijo de nombre
Filtro
Filtrar objetos y carpetas
Mostrar
Solo objetos activos
Nombre
Tamaño
Tipo
Fecha de creación
Clase de almacenamiento
Última modificación
Acceso público
_delta_log/
Carpeta
part-00000-ca7dd622-3239-439b-...
8.5 KB
application/octet-stream
17 mar 2025 19:27:45
Standard
17 mar 2025 19:27:45
No público
Descripción general
Buckets
Supervisión
Storage Insights
Configuración
Marketplace


console.cloud.google.com/storage/browser/dmc_datalake_dde_11_javp/produccion/dmc/bronze/empresas?pageState=("StorageObjectListTable":("f":"%255B%255D"))&inv=1&invt...
Estado de prueba gratuita: crédito por S/.938.18 y 51 días restantes. Activa tu cuenta completa para obtener acceso ilimitado a todas las funciones de Google Cloud. Usa los créditos restantes y paga solo por lo que usas.
Descartar
Activar
Google Cloud
My First Project
storage
Buscar
Cloud Storage
Detalles del bucket
IR A RUTA
ACTUALIZAR
MÁS INFORMACIÓN
dmc_datalake_dde_11_javp
Ubicación
us (varias regiones en Estados Unidos)
Clase de almacenamiento
Standard
Acceso público
No público
Protección
Borrar de forma no definitiva
OBJETOS
CONFIGURACIÓN
PERMISOS
PROTECCIÓN
CICLO DE VIDA
OBSERVABILIDAD
NUEVO
INFORMES DE INVENTARIO
OPERACIONES
Depósitos > dmc_datalake_dde_11_javp > produccion > dmc > bronze > empresas
CREAR CARPETA
SUBIR
TRANSFERIR LOS DATOS
OTROS SERVICIOS
Filtrar solo por prefijo de nombre
Filtro
Filtrar objetos y carpetas
Mostrar
Solo objetos activos
Nombre
Tamaño
Tipo
Fecha de creación
Clase de almacenamiento
Última modificación
Acceso público
_delta_log/
Carpeta
part-00000-d0036b4f-cb00-49ed-...
997 B
application/octet-stream
17 mar 2025 19:28:54
Standard
17 mar 2025 19:28:54
No público
Descripción general
Buckets
Supervisión
Storage Insights
Configuración
Marketplace

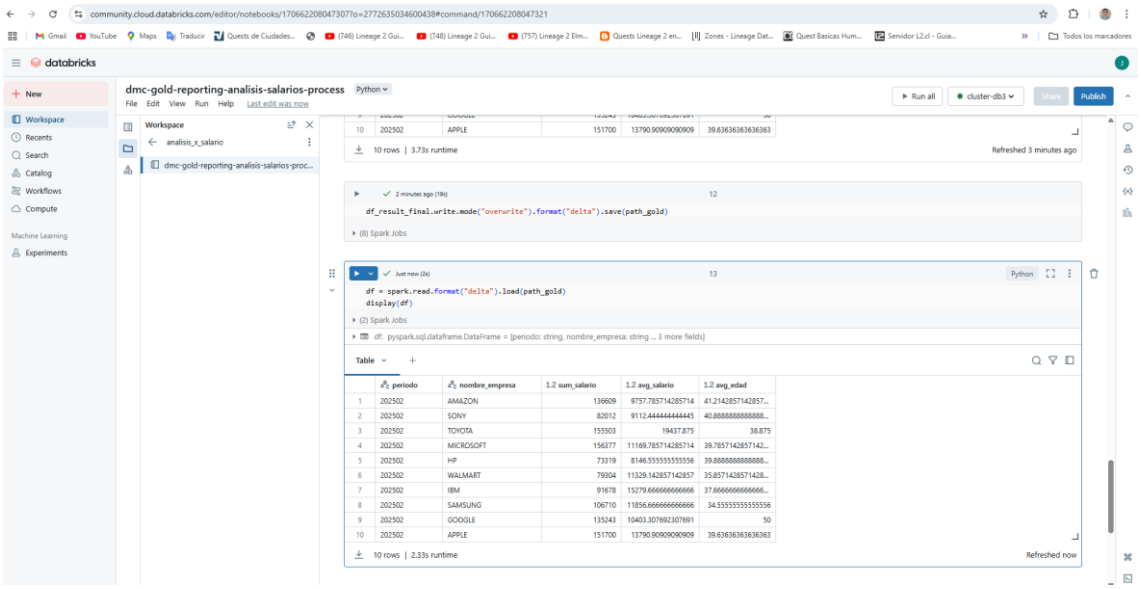
11.Sesión 11


Estado de prueba gratuita: crédito por S/.938.66 y 51 días restantes. Activa tu cuenta completa para obtener acceso ilimitado a todas las funciones de Google Cloud. Usa los créditos restantes y paga solo por lo que usas.
Descartar
Activar
Google Cloud
My First Project
google cloud storage
Buscar
Cloud Storage
Detalles del bucket
IR A RUTA
ACTUALIZAR
MÁS INFORMACIÓN
Descripción general
Buckets
Supervisión
Storage Insights
Configuración
dmc_datalake_dde_11_javp
Ubicación
us (varias regiones en Estados Unidos)
Clase de almacenamiento
Standard
Acceso público
No público
Protección
Borrar de forma no definitiva
OBJETOS
CONFIGURACIÓN
PERMISOS
PROTECCIÓN
CICLO DE VIDA
OBSERVABILIDAD
NUEVO
INFORMES DE INVENTARIO
OPERACIONES
Depósitos > dmc_datalake_dde_11_javp > produccion > dmc > silver > personas
CREAR CARPETA
SUBIR
TRANSFERIR LOS DATOS
OTROS SERVICIOS
Filtrar solo por prefijo de nombre
Filtro
Filtrar objetos y carpetas
Mostrar
Solo objetos activos
Nombre
Tamaño
Tipo
Fecha de creación
Clase de almacenamiento
Última modificación
Acceso público
Historial de versiones
PERIODO=202502/
Carpeta
_delta_log/
Carpeta
Marketplace
Notas de versión
Activar Windows
Ve a Configuración para activar Windows.


community.cloud.databricks.com/editor/notebooks/1100460553032897?o=2772635034600438#command/1100460553032902
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
dmc-silver-sap-transacciones-process
Python
File Edit View Run Help Last edit was 2 minutes ago
Run all
cluster-db2
Share
Publish
Workspace
process
dmc-silver-sap-transacciones-process
ID_PERSONA
ID_EMPRESA
1.2 MONTO
FECHA
ANIO
MES
DIA
10,000+ rows | Truncated data | 1.42s runtime
Refreshed 3 minutes ago
2 minutes ago (11s)
df_c.write.mode("overwrite").partitionBy("ANIO","MES","DIA").format("delta").save(path_silver)
(6) Spark Jobs
Activar Windows
Ve a Configuración para activar Windows.


console.cloud.google.com/storage/browser/dmc_datalake_dde_11_javp/produccion/dmc/silver/transacciones/ANIO%3D2018/MES%3D1?pageState=("StorageObjectListTable":("f":"%255B%255...
Estado de prueba gratuita: crédito por S/.938.66 y 51 días restantes. Activa tu cuenta completa para obtener acceso ilimitado a todas las funciones de Google Cloud. Usa los créditos restantes y paga solo por lo que usas.
Descartar
Activar
Google Cloud
My First Project
google cloud storage
Buscar
Cloud Storage
Detalles del bucket
IR A RUTA
ACTUALIZAR
MÁS INFORMACIÓN
Descripción general
Buckets
Supervisión
Storage Insights
Configuración
dmc_datalake_dde_11_javp
Ubicación
us (varias regiones en Estados Unidos)
Clase de almacenamiento
Standard
Acceso público
No público
Protección
Borrar de forma no definitiva
OBJETOS
CONFIGURACIÓN
PERMISOS
PROTECCIÓN
CICLO DE VIDA
OBSERVABILIDAD
NUEVO
INFORMES DE INVENTARIO
OPERACIONES
Depósitos > dmc_datalake_dde_11_javp > produccion > dmc > silver > transacciones > ANIO=2018 > MES=1
CREAR CARPETA
SUBIR
TRANSFERIR LOS DATOS
OTROS SERVICIOS
Filtrar solo por prefijo de nombre
Filtro
Filtrar objetos y carpetas
Mostrar
Solo objetos activos
Nombre
Tamaño
Tipo
Fecha de creación
Clase de almacenamiento
Última modificación
Acceso público
Historial de versiones
DIA=21/
Carpeta
DIA=22/
Carpeta
DIA=23/
Carpeta
Marketplace
Notas de versión
Activar Windows
Ve a Configuración para activar Windows.

12.Sesión 12



Archivos de los laboratorios zipeados:

En github: https://github.com/julio2019vp/curso-dmc-big-data-proccesing

1. **HDFS**

   HDFS.zip

2. **Apache Hive**

   Apache Hive.zip

3. **Apache Spark**

   Apache Spark.zip

4. **LakeHouse**

   LakeHouse.zip