



# **DIPLOMA DATA ENGINEER**

## **Big Data Processing**

Edicion 11 – Noviembre 2024

# Proyecto Final

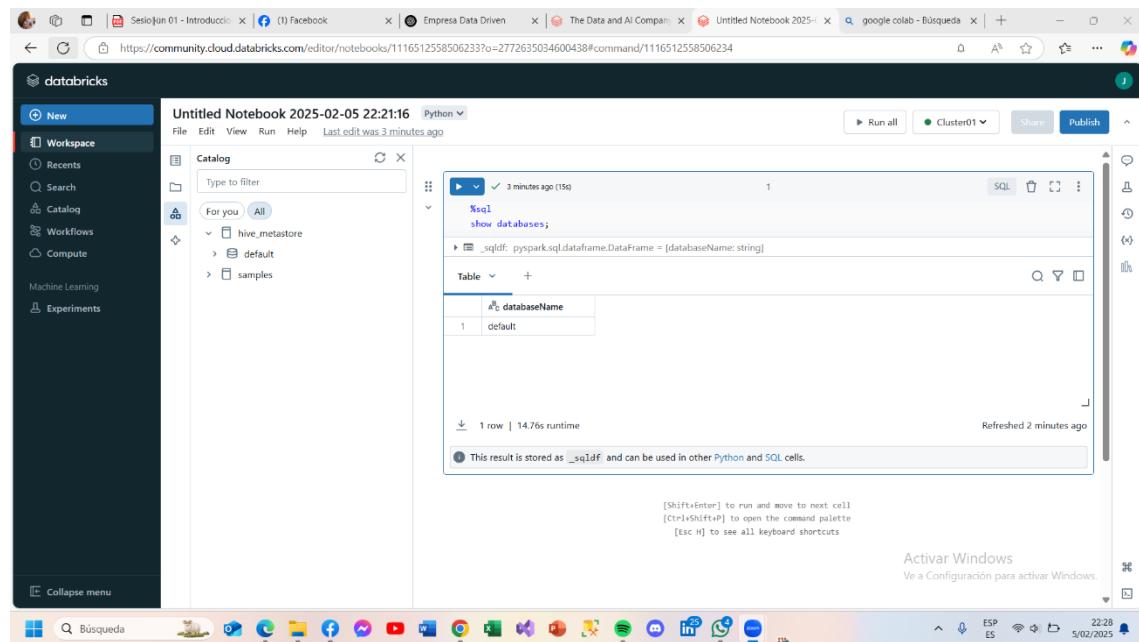
**Alumno: Julio Alexander Vasquez Pacheco**

**Curso: Big Data procesing**

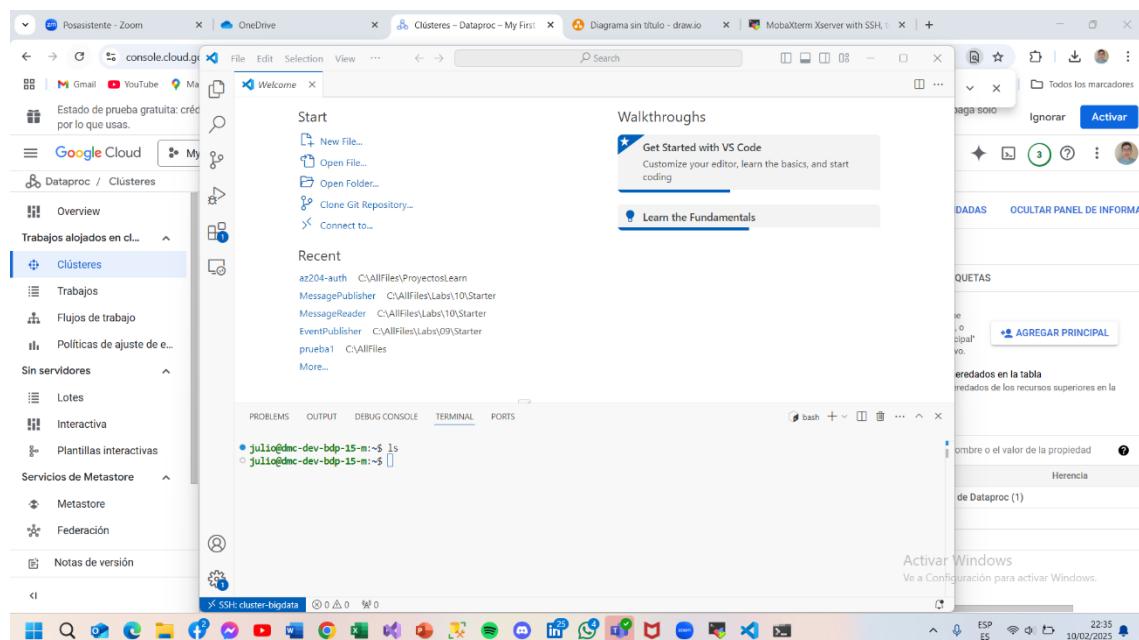
**Trabajo: Captura de pantalla de laboratorios en cada sesión y adjuntar los archivos.**

## Captura de Pantalla de Laboratorios en cada sesión

### 1. Sesión 1



### 2. Sesión 2



The screenshot shows the Google Cloud DataProc Clusters interface. On the left, there's a sidebar with options like Overview, Clusters, Trabajos, Flujos de trabajo, and Políticas de ajuste de e... In the main area, there's a table titled 'Clústeres' with one row:

Nombre	Estado	Región	Zona	Total de nodos trabajadores	Tiene VMs flexibles?
dmc-dev-bdp-15	Deteniendo	us-central1	us-central1-f	3	No

A message box says: "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)." Below the table, there's a section for 'PERMISOS' with a button 'AGREGAR PRINCIPAL'.

### 3. Sesión 3

The screenshot shows the Visual Studio Code (VS Code) interface. On the left, there's a sidebar with icons for file operations like New File, Open File, Open Folder, Clone Git Repository, and Connect to... Below that is a 'Recent' section listing files and folders. At the bottom of the sidebar, there's a checkbox 'Show welcome page on startup' which is checked.

The main area has tabs for PROBLEMS, OUTPUT, DEBUG CONSOLE, TERMINAL, and PORTS. The TERMINAL tab is active, showing a terminal session:

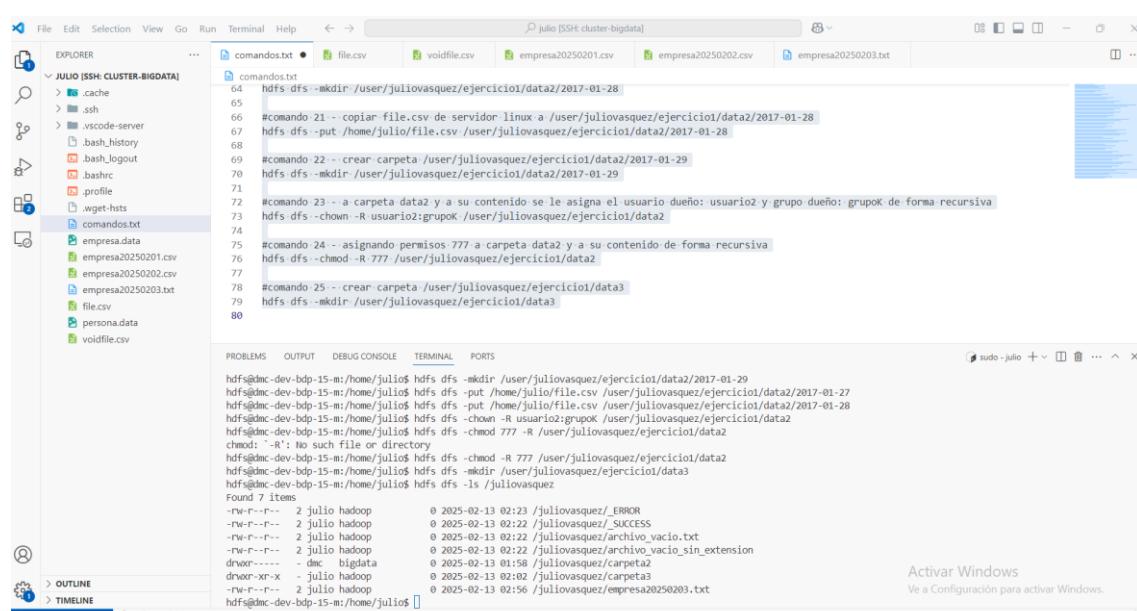
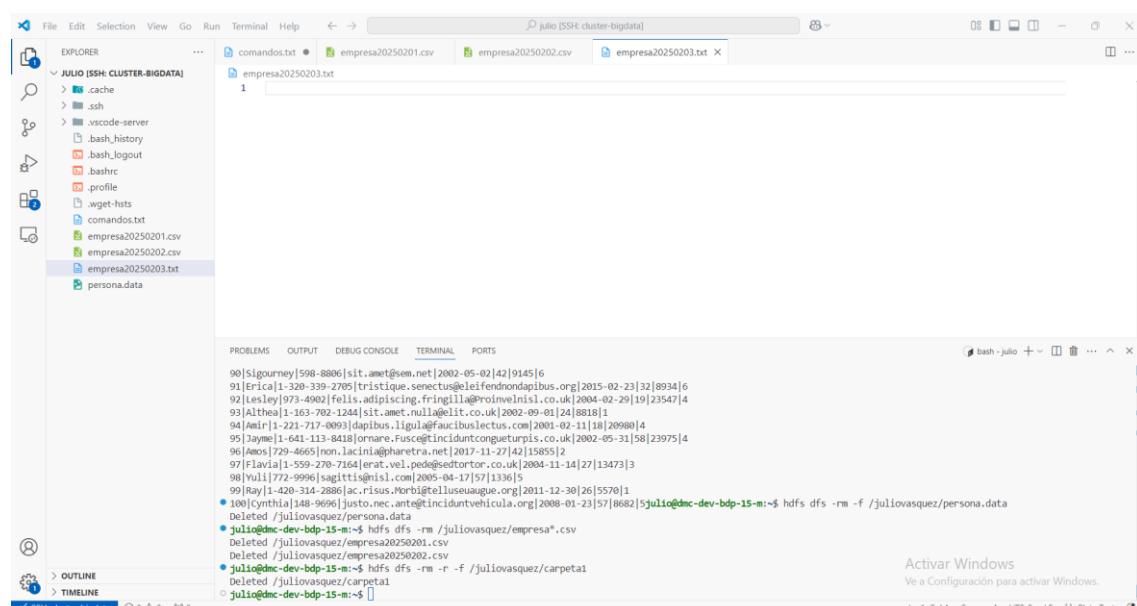
```

Found 3 items
dneovernet - hdfs hadoop    0 2025-02-11 02:53 /tmp
dneovernet - hdfs hadoop    0 2025-02-11 02:52 /user
dneovernet - hdfs hadoop    0 2025-02-11 02:52 /var
● julio@dmc-dev-bdp-15:~$ hdfs dfs -mkdir /juliovasesque
● julio@dmc-dev-bdp-15:~$ hdfs dfs -ls /
Found 4 items
drwxr-x  - julio hadoop    0 2025-02-13 01:55 /juliovasesque
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:~$ 
  
```

To the right of the terminal, there's a 'Walkthroughs' section with links to 'Get Started with VS Code' and 'Learn the Fundamentals'. At the bottom right, there's a note 'Activar Windows' with a link '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



```
julio@euc-dev-bdp-15-m:~$ ./comandos3.sql
```

The screenshot shows a terminal window with the following content:

```
Julio [SSH: cluster-bigdata]
```

```
File Edit Selection View Go Run Terminal Help ← →
```

```
explorer
```

```
JULIO (SSH: CLUS...)
```

```
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...
```

```
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
```

```
File Explorer
```

```
Terminal
```

```
PROBLEMS OUTPUT DEBUG CONSOLE PORTS
```

```
bash - julio
```

```
Activar Windows
```

```
Vea la configuración para activar Windows.
```

```
Ln 63 Col 33 Spaces: 4 UTF-8 LF MS SQL Q
```

```
julio@edc-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-reload4-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4-1.7.36.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-reload4-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4-1.7.36.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]  
socorro@edc-dev-bdp-15-m:~$
```

## 5. Sesión 5

DBBeaver 24.3.5 - <localhost> Script

Archivo Editor Navegar Buscar Editor SQL Base de Datos Ventana Ayuda

Navegador de Bases de Datos X Proyectos \*localhost> Script X

Ingrese parte del nombre de un objeto aquí

localhost 34.28.45.59:10000

- > bigdata
- > default
- > dmc
- miusuario\_test
  - Tablas
    - persona
    - transaccion
  - Columnas
    - AZ id\_persona (STRING)
    - AZ id\_empresa (STRING)
    - I23 monto (DOUBLE)
    - AZ fecha (STRING)
  - Claves
  - Columnas de clave externa
  - Referencias
  - Views

Project - General X

Name DataSource

Bookmarks Dashboards Diagrams Scripts

PET es Editable Inserción inteligente 1:44:43 Sel: 0 | 0

Resultados 1 X

```
select * from miusuario_test.transaccion t;
```

ID_PERSONA	ID_EMPRESA	VALOR	FECHA
1	[NULL]	3.142	2018-01-21
2	31	5	962 2018-01-21
3	63	3	204 2018-01-21
4	60	9	2.996 2018-01-21
5	83	5	3.418 2018-01-21
6	69	4	2.071 [2018-01-21]
7	20		

Renovar Save Cancel Exportar datos ... 200 200+ ... 200 row(s) fetched - 1.350s (0.949s fetch), on 2025-02-19 at 22:29:44

File Edit Selection View Go Run Terminal Help <- > julio [SSH cluster-bigdata]

EXPLORER comando3.sql comando2.txt comando3.sql file.csv voidfile.csv empresa20250201.csv empresa20250202.csv empresa20250203.txt file.csv persona.data transacciones-2018-01-21.d... transacciones-2018-01-22.d... transacciones-2018-01-23.d... voidfile.csv

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

Status: Running (Executing on YARN cluster with App id application\_1740015928889\_0001)

```
+---+ | _C0 | transaccion.fecha | +---+
| 50001 | 2018-01-21 |
| 61681 | 2018-01-22 |
| 123361 | 2018-01-23 |
+---+
3 rows selected (26.391 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 Ve a Configuración para activar Windows.

Ln 63, Col 33 Spaces: 4 UTF-8 LF {} MS SQL

File Edit Selection View Go Run Terminal Help <- > julio [SSH cluster-bigdata]

EXPLORER comando3.sql comando2.txt comando3.sql file.csv voidfile.csv empresa20250201.csv empresa20250202.csv empresa20250203.txt file.csv persona.data transacciones-2018-01-21.d... transacciones-2018-01-22.d... transacciones-2018-01-23.d... voidfile.csv

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

Found 1 items
driver-xr-x - julio hadoop 0 2025-02-20 02:21 /user/miusuario/bd/miusuario\_test2/persona
● juliodmc-dev-bdp-15-m-8 hdfs dfs -ls /user/miusuario/bd/miusuario\_test2/persona

Found 1 items
driver-xr-x - julio hadoop 7282 2025-02-20 02:21 /user/miusuario/bd/miusuario\_test2/persona/persona.data
● juliodmc-dev-bdp-15-m-8 hdfs dfs -ls /user/miusuario/bd/miusuario\_test2/persona/persona.data

Found 1 items
driver-xr-x - julio hadoop 7282 2025-02-20 02:21 /user/miusuario/bd/miusuario\_test2/persona/persona.data
● juliodmc-dev-bdp-15-m-8 pad /home/julio
● juliodmc-dev-bdp-15-m-8 hdfs dfs -ls /user/miusuario/bd/miusuario\_test/transaccion

Found 1 items
driver-xr-x - julio hadoop 0 2025-02-20 03:11 /user/miusuario/bd/miusuario\_test/transaccion/fecha=2018-01-21
● juliodmc-dev-bdp-15-m-8 hdfs dfs -ls /user/miusuario/bd/miusuario\_test/transaccion/fecha=2018-01-21

Found 1 items
driver-xr-x - julio hadoop 538827 2025-02-20 03:11 /user/miusuario/bd/miusuario\_test/transaccion/fecha=2018-01-21/transacciones-2018-01-21.data
● juliodmc-dev-bdp-15-m-8 hdfs dfs -ls /user/miusuario/bd/miusuario\_test/transaccion

Found 3 items
driver-xr-x - julio hadoop 0 2025-02-20 03:11 /user/miusuario/bd/miusuario\_test/transaccion/fecha=2018-01-21
driver-xr-x - julio hadoop 0 2025-02-20 03:17 /user/miusuario/bd/miusuario\_test/transaccion/fecha=2018-01-22
driver-xr-x - julio hadoop 0 2025-02-20 03:18 /user/miusuario/bd/miusuario\_test/transaccion/fecha=2018-01-23

● juliodmc-dev-bdp-15-m-8 [2]

Activar Windows Ve a Configuración para activar Windows.

Ln 63, Col 33 Spaces: 4 UTF-8 LF {} MS SQL

## 6. Sesión 6

The screenshot shows the Data Studio interface with the following components:

- EXPLORER**: Shows the project structure under "JULIO [SSH: CLUSTER-BIGDATA]".
- EDITOR**: Displays two tabs: "Poblando\_Capa\_Functional.sql" and "Poblando\_Capa\_Curated.sql". The "Poblando\_Capa\_Functional.sql" tab contains the following code:

```
procesos > Poblando_Capa_Functional.sql
229 PARTITION (FECHA_TRANSACCION)
230   SELECT
231     T.SALARIO_PERSONA,
232     T.TRABAJO_PERSONA,
233     T.MONTO_TRANSACCION,
234     T.EMPRESA_TRANSACCION,
235     T.FECHA_TRANSACCION
236   FROM
237     $[hiveconf:ENV]_FUNCTIONAL.TMP_TRANSACCION_ENRIQUECIDA_3 T;
238
239 -- Verificamos
240 SELECT * FROM ${hiveconf:ENV}_FUNCTIONAL.TRANSACCION_ENRIQUECIDA LIMIT 10;
241
242 -- @sección 4. Eliminación de tablas temporales
243
244
245
246
247
248
249
250
251 -- DROP TABLE IF EXISTS ${hiveconf:ENV}_FUNCTIONAL.TMP_TRANSACCION_ENRIQUECIDA_1;
252 -- DROP TABLE IF EXISTS ${hiveconf:ENV}_FUNCTIONAL.TMP_TRANSACCION_ENRIQUECIDA_2;
253 -- DROP TABLE IF EXISTS ${hiveconf:ENV}_FUNCTIONAL.TMP_TRANSACCION_ENRIQUECIDA_3;
```

- TERMINAL**: Shows the command "julio@dmc-dev-bdp-15-m:~\$".
- STATUS BAR**: Shows "Julio Vásquez Pacheco (julio\_tvip@hotmail.com) está conectado".
- NOTIFICATIONS**: Shows "Activar Windows" and "Ve la Configuración para activar Windows".
- STATISTICS**: Shows "Ln 245, Col 14 | Spaces: 4 | UTF-8 | CRLF | MS SQL | 🌐".

The screenshot shows the Data Studio interface with the following components:

- EXPLORER**: Shows the project structure under "JULIO [SSH: CLUSTER-BIGDATA]".
- EDITOR**: Displays two tabs: "Poblando\_Capa\_Functional.sql" and "Poblando\_Capa\_Workload.sql". The "Poblando\_Capa\_Workload.sql" tab contains the following code:

```
procesos > Poblando_Capa_Workload.sql
110 CREATE TABLE ${hiveconf:ENV}_workload.TRANSACCION(
111   MONTO STRING,
112   FECHA STRING
113 )
114 ROW FORMAT DELIMITED
115   FIELDS TERMINATED BY '|'
116   LINES TERMINATED BY '\n'
117   STORED AS TEXTFILE
118   LOCATION '/user/${hiveconf:PARAM_USERNAME}/datalake/${hiveconf:ENV}_workload/transaccion'
119   TBLPROPERTIES(
120     'skip.header.line.count'=1',
121     'store.charset'='ISO-8859-1',
122     'retrieve.charset'='ISO-8859-1'
123   );
124
125
126
127 -- Subida de datos
128 LOAD DATA LOCAL INPATH '/home/${hiveconf:PARAM_USERNAME}/dataset/transacciones.data'
129 INTO TABLE ${hiveconf:ENV}_workload.TRANSACCION;
130
131 -- Impresión de datos
132 SELECT * FROM ${hiveconf:ENV}_workload.TRANSACCION LIMIT 10;
133
```

- TERMINAL**: Shows the command "julio@dmc-dev-bdp-15-m:~\$".
- STATUS BAR**: Shows "Activar Windows" and "Ve la Configuración para activar Windows".
- STATISTICS**: Shows "Ln 13, Col 14 | Tab Size: 4 | UTF-8 | CRLF | MS SQL | 🌐".

File Edit Selection View Go Run Terminal Help ⏪ ⏴ ⏵ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿

REPORETERO JULIO [SSH: CLUSTER-BIGDATA] ... Poblano\_Capa\_Functional.sql Poblano\_Capa\_Curated.sql

procesos > Poblano\_Capa\_Curated.sql

```

148  );
149
150 -- Inserción por particionamiento dinámico, casteo de datos y aplicación de reglas de limpieza
151 INSERT INTO TABLE ${hiveconf:ENV}_curated.TRANSACCION
152 PARTITION(FECHA)
153   SELECT
154     CAST(T.ID_PERSONA AS STRING),
155     CAST(T.ID_EMPRESA AS STRING),
156     CAST(T.MONTO AS DOUBLE),
157     CAST(T.FECHA AS STRING)
158   FROM
159     ${hiveconf:ENV}_LANDING.TRANSACCION T
160   WHERE
161     T.ID_PERSONA IS NOT NULL AND
162     T.ID_EMPRESA IS NOT NULL AND
163     CAST(T.MONTO AS DOUBLE) >= 0;
164
165 -- Impresión de datos
166 SELECT * FROM ${hiveconf:ENV}_curated.TRANSACCION LIMIT 10;
167
168 -- Verificamos las particiones
169 SHOW PARTITIONS ${hiveconf:ENV}_curated.TRANSACCION;

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

julio@dmn-dev-bdp-15-m:~\$

Activar Windows  
Ve a Configuración para activar Windows.

Ln 4, Col 4 (74 selected) Spaces: 4 UTF-8 CRLF MS SQL

File Edit Selection View Go Run Terminal Help ⏪ ⏴ ⏵ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿

REPORETERO JULIO [SSH: CLUSTER-BIGDATA] ... Poblano\_Capa\_Functional.sql Poblano\_Capa\_Landing.sql

procesos > Poblano\_Capa\_Landing.sql

```

1
2
3 -- COMANDO DE EJECUCIÓN
4 -- beeline -u jdbc:hive2:// -f /home/julio/procesos/Poblano_Capa_Landing.sql --hiveconf "PARAM_USERNAME=juan" --hiveconf "ENV=DEV"
5
6 /*
7 hdfs dfs -mkdir -p /user/julio/datalake/schema/dev_LANDING
8 hdfs dfs -put /home/julio/schema/*.avsc /user/julio/datalake/schema/dev_LANDING/
9 */
10 -----
11 --
12 -- @section 1. Definición de parámetros
13 --
14 -----
15 -- [HIVE] Creamos una variable en HIVE
16 SET ENV=dev;
17 SET PARAM_USERNAME=julio;
18
19 --
20 --
21 --
22 -- @section 2. Eliminación de base de datos
23 --

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

julio@dmn-dev-bdp-15-m:~\$

Activar Windows  
Ve a Configuración para activar Windows.

Ln 4, Col 4 (74 selected) Spaces: 4 UTF-8 CRLF MS SQL

## 7. Sesión 7

Google Cloud My First Project storage Buscar Detalles del bucket dmc\_datalake\_dde\_11\_javp

Cloud Storage Buckets

Ubicación Clase de almacenamiento Acceso público Protección

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

Depósitos > dmc\_datalake\_dde\_11\_javp

CREAR CARPETA SUBIR TRANSFERIR LOS DATOS OTROS SERVICIOS

Filtrar solo por prefijo de nombre Filtro Filtrar objetos y carpetas

Nombre	Tamaño	Tipo	Fecha de creación	Clase de almacenamiento	Última modificación	Acceso público	Historial de versión
archivos/	-	Carpetas	-	-	-	-	-
producción/	-	Carpetas	-	-	-	-	-

Activar Windows Ve a Configuración para activar Windows.

File Edit View Run Kernel Git Tabs Settings Help

introducción-spark.ipynb

Filter files by name

Name Last Modified

introducción-spark.ipynb 14 days ago

[Stage 2/2] (0 + 1) / 1

ID	NOMBRE	TELÉFONO	CORREO	FECHA_INGRESO	EDAD	SALARIO	ID_EMPRESA
1	Carl	1-745-633-9145	arcu.Sed.et@ante...	2004-04-23	32	20995.0	5
2	Priscilla	155-2498	bоне.егеста.Али...	2019-02-17	34	9298.0	2
3	Jocelyn	1-284-956-8351	and.diam@liborti...	2012-06-10	24	10850.0	31
4	Aidan	1-719-862-9385	eu.uliam@et.commod...	2018-11-06	29	3387.0	10
5	Leopold	1-884-3844	ut.sed@ipsum.org...	2009-10-10	41	10202.0	11
6	Bertl	797-6453	a.felis.ulamcorp...	2017-04-25	70	7800.0	7
7	Mark	1-680-102-6792	loisque.a@placer...	2006-04-21	52	8112.0	5
8	Jonah	214-2975	eu.ultrices.sit@...	2017-10-07	23	17040.0	5
9	Hanae	935-2277	eu@uncn.ca	2003-05-25	69	6834.0	3
10	Cadman	1-866-561-2701	orci.adipiscing.n...	2001-05-19	19	7996.0	7

only showing top 10 rows

[#]: df.printSchema()

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

Activar Windows Ve a Configuración para activar Windows.

Google Cloud My First Project dataproc

Dataproc / Clústeres / Clúster: dmc-dev-bdp-15 / Interfaces

Detalles del clúster ENVIAR TRABAJO ACTUALIZAR INICIAR DETENER BORRAR VER REGISTROS

Failed to validate permissions required for default service account: 281870365699-compute@developer.gserviceaccount.com. Cluster creation could still be successful if required permissions have been granted to the respective service accounts as mentioned in the document https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/service-accounts#dataproc\_service\_accounts\_2. This could be due to Cloud Resource Manager API hasn't been enabled in your project "281870365699" before or it is disabled. Enable it by visiting https://console.developers.google.com/apis/api/cloudresourcemanager.googleapis.com/overview?project=281870365699.

Nombre dmc-dev-bdp-15  
UUID del clúster dSedd107-e1e9-4783-af79-79c20e653536  
Tipo Clúster de Dataproc  
Estado En ejecución

MONITORING TRABAJOS INSTANCIAS DE VM CONFIGURACIÓN INTERFAZES WEB

Túnel SSH Crea un túnel SSH para conectarte a una interfaz web

Puerta de enlace del componente Proporciona acceso a las interfaces web de componentes

Iniciando clúster... Activar Windows Ve a Configuración para activar Windows.

YARN ResourceManager

The screenshot shows a Jupyter Notebook interface running on Google Cloud DataProc. The notebook file is named "introduccion-spark.ipynb". The current cell (cell 14) contains the following Python code:

```
[84]: from pyspark.sql.types import *
ruta = "gs://dmc_datelake_ode_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: string)

[14]: df.show(10)
Stage 2: (0 + 1) / 1
+----+-----+-----+-----+-----+-----+
| ID | NOMBRE | TELEFONO | CORREO | FECHA_INGRESO | EDAD | SALARIO | ID_EMPRESA |
+----+-----+-----+-----+-----+-----+
| 1 | Carl1-745-633-9145|arcu.Sed et@ante....| 2004-04-23 | 32 | 20095.0 | 5 | 
| 2 | Priscilla | 155-2498|Donec.egestas.Ai....| 2019-02-17 | 34 | 9298.0 | 2 |
+----+-----+-----+-----+-----+-----+
```

The notebook interface includes a sidebar for file navigation, a code editor, and a results pane where the output of the `df.show(10)` command is displayed as a table. A tooltip "Activar Windows" is visible on the right.

## 8. Sesión 8

The screenshot shows a Jupyter Notebook interface running on Google Cloud DataProc. The notebook file is named "introduccion-spark.ipynb". The current cell (cell 127) contains the following Python code:

```
[127]: df.groupby("id_empresa").agg(sum("salario").alias("planilla"), \
    avg("edad").alias("prom_edad"), \
    max("salario").alias("max_salary")) \
    .where((col("PLANILLA")>=100000).show(10)
```

The results of the aggregation are displayed in a table:

id_empresa	planilla	prom_edad	max_salary
1	79304.0	35.857142857142854	22102.0
8	73319.0	39.888888888888866	19943.0
5	136609.0	41.214285714285715	20095.0
10	82012.0	40.888888888888866	24575.0
4	155503.0	38.875	24305.0
2	156377.0	39.785714285714285	22953.0

The notebook interface includes a sidebar for file navigation, a code editor, and a results pane where the output of the `df.show(10)` command is displayed as a table. A tooltip "Activar Windows" is visible on the right.

The screenshot shows the Google Cloud Storage console. On the left, there's a sidebar with options like 'Descripción general', 'Buckets', 'Supervisión', 'Configuración', 'Marketplace', and 'Notas de versión'. The main area is titled 'Detalles del bucket' and shows a list of objects. The list includes several folders named 'id\_empresa-1/' through 'id\_empresa-10/'. The table columns are 'Nombre', 'Tamaño', 'Tipo', 'Fecha de creación', and 'Clase de almacenamiento'. At the bottom, there are filters for 'Nombre', 'Tamaño', 'Tipo', and 'Filtrar objetos y carpetas', and a dropdown for 'Mostrar'.

## 9. Sesión 9

The screenshot shows a Databricks notebook titled 'session-09-read-notebook-cloud\_storage\_JulioVasquez'. The left sidebar has sections for 'Workspace', 'Machine Learning', and 'Experiments'. The notebook content shows Python code for reading a CSV file and displaying its contents. The code is:

```

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

```

The right side shows the output of the 'display(df)' command, which is a table with 9 rows. The columns are labeled 'ID', 'NOMBRE', 'TELEFONO', 'CORREO', and 'FECHA\_INGRESO'. The data is as follows:

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@voluptatumn.edu	2019-02-17
3	Jocelyn	1-204-956-8594	amet.diam@loboris.co.uk	2002-08-01
4	Aidan	1-719-862-9385	eiusmod.et.commodo@nihilacciniorci.edu	2018-11-06
5	Leandra	839-8044	at@pretiumetrurum.com	2002-10-10
6	Bert	797-4453	a.feisullamcorper@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

**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

100 rows | 0.39s runtime Refreshed 8 minutes ago

```

rutasfifa = 'gs://dmc_datalake_dde_11_javp/archivos/fifa_ranking.csv'

df_fifa = spark.read.format("CSV").option("header","true").option("delimiter",",").load(rutasfifa)

display(df_fifa)

```

(2) Spark Jobs

df\_fifa: pyspark.sql.dataframe.DataFrame = [rank: string, country\_full: string ... 14 more fields]

#	rank	country_full	country_abrv	total_points	previous_points	rank_change
1	1	Germany	GER	0.0	57	0
2	2	Italy	ITA	0.0	57	0
3	3	Switzerland	SUI	0.0	50	9
4	4	Sweden	SWE	0.0	55	0
5	5	Argentina	ARG	0.0	51	5
6	6	Republic of Ireland	IRL	0.0	54	0
7	7	Russia	RUS	0.0	52	0
8	8	Brazil	BRA	0.0	55	0
9	9	Norway	NOR	0.0	49	5

## 10. Sesión 10

Just now (9s)

display(df\_personas)

(1) Spark Jobs

Table

#	ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO
1	1	Carl	1-745-633-9145	arcu.Sed.et@ante.co.uk	2004-04-23
2	2	Priscilla	155-2498	Donec.egestas.Aliquam@voluptatnunc.edu	2019-02-17
3	3	Jocelyn	1-204-956-8594	amet.diam@lobortis.co.uk	2002-08-01
4	4	Aidan	1-719-862-9385	euismod.et.commodo@nibhacinaorci.edu	2018-11-06
5	5	Leandra	839-8044	at@pretiumetrurum.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
9	9	Hanae	935-2277	eu@Nunc.ca	2003-05-25
10	10	Cadman	1-866-561-2701	orci.adipiscing.non@semperNam.ca	2001-05-19
11	11	Melyssa	506-7736	vel@vulputateposuerelutuata.net	2008-10-14

community.cloud.databricks.com/editor/notebooks/223001412045504?o=2772635034600438#command/223001412045505

**dmc-bronze-sap-persona-process** Python

File Edit View Run Help Last edit was 5 days ago

Run all Terminated Share Publish

Workspace

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]

#	ID	NOMBRE	TELEFONO	CORREO	FECHA_INGRESO
1	1	Carl	1-745-633-9145	arcu.Sed.et@ante.co.uk	2004-04-23
2	2	Priscilla	155-2498	Donec.egestas.Aliquam@voluptatnunc.edu	2019-02-17
3	3	Jocelyn	1-204-956-8594	amet.diam@lobortis.co.uk	2002-08-01
4	4	Aidan	1-719-862-9385	euismod.et.commodo@nibhacinaorci.edu	2018-11-06
5	5	Leandra	839-8044	at@pretiumetrurum.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
9	9	Hanae	935-2277	eu@Nunc.ca	2003-05-25
10	10	Cadman	1-866-561-2701	orci.adipiscing.non@semperNam.ca	2001-05-19
11	11	Melyssa	506-7736	vel@vulputateposuerelutuata.net	2008-10-14

Screenshot of a Databricks notebook titled "dmc-bronze-sap-empresas-process". The notebook is running in Python. It shows a table with 10 rows of data:

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

The notebook has a status message: "Notebook detached cluster not in usable state". Below the notebook, there is a snippet of code:

```
df.write.mode("overwrite").format("delta").save(path_bronze)
```

Screenshot of Google Cloud Storage showing the details of a bucket named "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:**

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

Screenshot of Google Cloud Storage showing the details of the same 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:**

Nombre	Tamaño	Tipo	Fecha de creación	Clase de almacenamiento	Última modificación	Acceso público
_delta_log/	—	Carpeta	—	—	—	—
part-00000-d0036bf4-cb00-49ed-...	997 B	application/octet-stream	17 mar 2025 19:28:54	Standard	17 mar 2025 19:28:54	No público

## 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.

Google Cloud My First Project google cloud storage Buscar

Cloud Storage Detalles del bucket dmc\_datalake\_dde\_11\_javp

Ubicación Clase de almacenamiento Acceso público Protección

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

Depósitos > dmc\_datalake\_dde\_11\_javp > producción > 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 versión
PERIODO-202502/	-	Carpetas	-	-	-	-	-
_delta_log/	-	Carpetas	-	-	-	-	-

Activar Windows Ve a Configuración para activar Windows.

community.cloud.databricks.com/editor/notebooks/1100460553032897?o=2772635034600438#command/1100460553032902

Gmail YouTube Maps Traducir Quests Ciudades... (746) Lineage 2 Gui... (748) Lineage 2 Gui... (757) Lineage 2 Elm... Quests Lineage 2 en... Zones - Lineage Dat... Todos los marcados

databricks

+ New

File Edit View Run Help Last edit was 2 minutes ago

Run all cluster-db2 Share Publish

Workspace Recents Search Catalog Workflows Compute Machine Learning Experiments

dmc-silver-sap-transacciones-process Python

ID_PERSONA	ID_EMPRESA	MONTO	FECHA	ANIO	MES	DIA
91	4	3081	2018-01-21	2018	1	21
74	8	2409	2018-01-21	2018	1	21
41	2	3754	2018-01-22	2018	1	22
42	9	4079	2018-01-22	2018	1	22
24	6	4475	2018-01-22	2018	1	22
67	9	561	2018-01-22	2018	1	22
9	4	3765	2018-01-22	2018	1	22
97	3	3669	2018-01-22	2018	1	22
91	5	3497	2018-01-22	2018	1	22
61	3	735	2018-01-23	2018	1	23
15	5	367	2018-01-23	2018	1	23
20	9	2039	2018-01-23	2018	1	23
11	4	719	2018-01-23	2018	1	23
36	2	2659	2018-01-23	2018	1	23
12	4	467	2018-01-23	2018	1	23

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) Activar Windows

(6) Spark Jobs

console.cloud.google.com/storage/browser/dmc\_datalake\_dde\_11\_javp/producción/dmc/silver/transacciones/ANIO=2018/MES=1?pageState={"StorageObjectListTable":{}}

Gmail YouTube Maps Traducir Quests Ciudades... (746) Lineage 2 Gui... (748) Lineage 2 Gui... (757) Lineage 2 Elm... Quests Lineage 2 en... Zones - Lineage Dat... Todos los marcados

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.

Google Cloud My First Project google cloud storage Buscar

Cloud Storage Detalles del bucket dmc\_datalake\_dde\_11\_javp

Ubicación Clase de almacenamiento Acceso público Protección

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

Depósitos > dmc\_datalake\_dde\_11\_javp > producción > dmc > 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 versión
DIA=21/	-	Carpetas	-	-	-	-	-
DIA=22/	-	Carpetas	-	-	-	-	-
DIA=23/	-	Carpetas	-	-	-	-	-

Activar Windows Ve a Configuración para activar Windows.

## 12. Sesión 12

The screenshot shows a Databricks workspace with a notebook titled "dmc-gold-reporting-analysis-salarios-process". The notebook contains the following Python code:

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

df_result_final.write.mode("overwrite").format("delta").save(path_gold)
```

Below the code, a table is displayed with the following data:

#	periodo	nombre_empresa	1.2 sum_salario	1.2 avg_salario	1.2 avg_edad
1	202502	AMAZON	136609	9757.785714285714	41.214285714285714...
2	202502	SONY	82012	9112.444444444445	40.88888888888888...
3	202502	TOHOTA	155903	19437.875	38.875
4	202502	MICROSOFT	156877	11169.785714285714	39.78571428571428...
5	202502	HP	73319	8145.555555555556	39.88888888888888...
6	202502	WALMART	79304	13234.142857142857	35.8571428571428...
7	202502	IBM	91676	13279.666666666666	37.66666666666666...
8	202502	SAMSUNG	106710	11856.666666666666	34.55555555555555...
9	202502	GOOGLE	135241	10403.307693307691	50
10	202502	APPLE	151700	13790.90909090909	39.63636363636363...

### Archivos de los laboratorios zipeados:

#### 1. HDFS



HDFS.zip

#### 2. Apache Hive



Apache Hive.zip

#### 3. Apache Spark



Apache Spark.zip

#### 4. LakeHouse



LakeHouse.zip