

Tarea 03: Ecosistema Hadoop

Juan Manuel García Moyano
IABD
Informática y comunicaciones

Índice

1. Caso Práctico.....	4
2. ¿Qué te pedimos que hagas?.....	4
Práctica 1. Hive.....	4
Pregunta 1.....	5
Pregunta 2.....	6
Pregunta 3.....	6
Pregunta 4.....	7
Pregunta 5.....	8
Pregunta 6.....	9
Pregunta 7.....	13
Pregunta 8.....	14
Pregunta 9.....	15
Práctica 2. Spark.....	17
Pregunta 1.....	17
Pregunta 2.....	18
Pregunta 3.....	18
Pregunta 4.....	19
Pregunta 5.....	20
Pregunta 6.....	21

Índice de figuras

Figura 1: Base de datos, la tabla externa y sus atributos.....	5
Figura 2: Consulta con el total de filas de la tabla.....	6
Figura 3: Consulta con el total de cuerpos distintos.....	7
Figura 4: Consulta con el total de plazas de maestros.....	8
Figura 5: Consulta con todos los maestros que tiene Jaén.....	9
Figura 6: Consulta con el total de plazas de profesores técnicos.....	10
Figura 7: Resultado de la consulta.....	10
Figura 8: Resultado de la consulta.....	11
Figura 9: Resultado de la consulta.....	12
Figura 10: Resultado de la consulta.....	12
Figura 11: Consulta con la plazas de profesores técnicos en el Zaidín-Vergeles.....	13
Figura 12: Consulta para saber el total de plazas por cuerpo.....	13
Figura 13: Resultado de la consulta.....	14
Figura 14: Consulta para saber el total de plazas por cuerpo de Sevilla.....	14
Figura 15: Resultado de la consulta.....	15
Figura 16: Modifico el modo de partición de Hive a nonstrict.....	15
Figura 17: Consulta para la creación de la tabla administrada.....	16
Figura 18: Añado todos los valores de la tabla externa a la tabla administrada.....	16
Figura 19: Muestro el contenido del directorio plantilla2223_administrada.....	16
Figura 20: Creación de una sesión con pyspark.sql y con soporte a Hive.....	17
Figura 21: Importación de los datasets y el número de fila de cada uno.....	18
Figura 22: Total de plazas de los cursos 21/22 y 22/23, junto la diferencia de plazas de ambos.....	19
Figura 23: Muestro la provincia que ha tenido un mayor y menor crecimiento de plantilla.....	20
Figura 24: Gráfica de barras apiladas con las plazas de cada provincia y curso.....	21
Figura 25: Muestro la sección de archivos para ver si se ha creado correctamente la tabla.....	22

1. Caso Práctico

La Federación de Enseñanza del sindicato mayoritario X de Andalucía quiere afrontar el estudio de la plantilla orgánica de los centros públicos andaluces.

El objetivo inmediato es poder analizar la evolución de esta plantilla en los últimos años, comparar esta evolución con la de los centros concertados y privados, y para el futuro, aplicar aprendizaje automático para predecir el número de profesores necesarios en un curso escolar. Este conocimiento les situaría con una gran ventaja en las negociaciones con el gobierno andaluz.

Para cumplir estos objetivos han contratado a Raúl. Raúl ha trabajado muchos años como programador full stack en la empresa privada y, para reciclarse, ha realizado el Curso de Especialización en Inteligencia Artificial y Big Data.

Para empezar, Raúl va a estudiar la plantilla orgánica de los centros docentes públicos. Después afrontará la recolección de datos de los centros privados, escolarización, natalidad, etc., como analizarlos y visualizarlos.

La plantilla orgánica de los Centros públicos se publica anualmente en el BOJA en formato pdf. En el portal de Datos abiertos de la Junta de Andalucía podemos encontrar un archivo con esta plantilla en formato CSV. El archivo se actualiza anualmente. En el sindicato cuentan con los archivos del curso 21/22 y 22/23.

Estos dos archivos no son muy grandes, pero ante la previsión del futuro crecimiento de datos, Raúl va a trabajar con un clúster de AWS con Hive y Spark para analizar estos archivos y para hacer una primera comparativa de la plantilla de los dos últimos cursos escolares.

2. ¿Qué te pedimos que hagas?

Para esta tarea es necesario crear un clúster ERM en el laboratorio AWS que se te ha proporcionado en el módulo y un cuaderno Google Colab.

La entrega final será un fichero PDF con las preguntas que se formulan y las respuestas que se solicitan.

ADVERTENCIA: No llenes el PDF de capturas innecesarias. Solo se evaluarán aquellas capturas que se pidan explícitamente y que tengan un tamaño legible en el PDF.

Práctica 1. Hive.

Resumen. En el clúster, con Hive y Hue, crearemos una tabla externa sobre el archivo `plantilla2223.csv` que se aporta como recurso. Sobre esta tabla realizaremos una serie de consultas interactivas que nos ayudarán a comprender la información del archivo. A continuación crearemos una tabla interna Hive particionada por el campo CUERPO.

Pregunta 1.

A través de la herramienta Hue, crea una base de datos con tu **nombre completo**, por ejemplo, ArmandoBroncaSegura_2425. En esta base de datos crearás una **tabla externa** Hive con los datos del archivo plantilla2223.csv.

Respuesta: Responde con la consulta de creación de la tabla y aporta una captura de pantalla con la información de la tabla creada (icono i, show details) desplegada. Debe observarse claramente el nombre de la base de datos, el nombre de la tabla y el contenido de la primera línea.

USE juanmanuelgarciamoyano_2425;

CREATE EXTERNAL TABLE IF NOT EXISTS plantilla2223 (

CUERPO STRING,

PROVINCIA STRING,

LOCALIDAD STRING,

CENTRO STRING,

PUESTO STRING,

PLAZAS INT

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/user/iabd-bda/plantilla_organica_2022_2023_dir/';

Column (6)	Type	Description	Sample
cuerpo	string	CUERPO	MAESTROS
provincia	string	PROVINCIA	ALMERÍA
localidad	string	LOCALIDAD	ABLA
centro	string	CENTRO	04000018-C.E.I.P.
puesto	string	PUESTO	00597031-EDUCACIÓN INFANTIL
plazas	int	NULL	2

Figura 1: Base de datos, la tabla externa y sus atributos.

Pregunta 2.

Escribe una consulta para saber **cuántas filas** tiene la tabla creada sin la cabecera. Responde con la consulta y el número de fila que has obtenido.

```
1 SELECT COUNT(*) AS 'Total filas'
2 FROM plantilla2223
3 WHERE CUERPO != 'CUERPO';
4
5
```

```
INFO : Map 1: 1/1      Reducer 2: 0(1)/1
INFO : Map 1: 1/1      Reducer 2: 1/1
INFO : Completed executing command(queryId=hive_20250309
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock
```

Query History	Saved Queries	Results (1)
total filas		
1	35656	

Figura 2: Consulta con el total de filas de la tabla.

Pregunta 3.

Escribe la consulta necesaria para mostrar los **distintos cuerpos** de la tabla. Responde con la consulta y el número de cuerpos obtenidos.

```
1 SELECT count(DISTINCT CUERPO) AS 'Total de cuerpos distintos'
2 FROM plantilla2223
3 WHERE CUERPO != 'CUERPO';
4
5
6
```

```
INFO : Map 1: 1/1      Reducer 2: 0/1/2      Reducer 3: 0/1
INFO : Map 1: 1/1      Reducer 2: 2/2      Reducer 3: 1/1
INFO : Completed executing command(queryId=hive_20250309173436_146bbf7f-6
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
```



Query History	Saved Queries	Results (1)
total de cuerpos distintos		
		
		
1	11	

Figura 3: Consulta con el total de cuerpos distintos.

Pregunta 4.

Escribe una consulta para saber cuántas **plazas de maestros** hay en la tabla. Responde con la consulta y el número de plazas de maestros que hayas obtenido.

```
1
2 SELECT SUM(PLAZAS) AS `Total plazas de maestros`
3 FROM plantilla2223
4 WHERE CUERPO LIKE '%MAESTROS%' AND CUERPO != 'CUERPO';
5
```

```
INFO : Map 1: 0/1/1 Reducer 2: 0/1
INFO : Map 1: 1/1 Reducer 2: 0/1
INFO : Map 1: 1/1 Reducer 2: 0(+1)/1
INFO : Map 1: 1/1 Reducer 2: 1/1
INFO : Completed executing command(queryId=hive_20250309173829_f432
8 seconds
```

Query History

Saved Queries

Results (1)

total plazas de maestros



id	total plazas de maestros
1	38423

Figura 4: Consulta con el total de plazas de maestros.

Pregunta 5.

Escribe una consulta para saber cuántas **plazas de maestros** hay en la provincia de **Jaén**. Responde con la consulta y el número de plazas de maestros que hayas obtenido.


```
1 SELECT sum(PLAZAS) AS `Maestros en Jaén`  
2 FROM plantilla2223  
3 WHERE cuerpo LIKE '%MAESTROS%'  
4     AND provincia = 'JAÉN'  
5     AND cuerpo != 'CUERPO';  
6
```

```
INFO : Map 1: 1/1   Reducer 2: 1/1  
INFO : Completed executing command(queryId=hive_202  
9 seconds  
INFO : OK  
INFO : Concurrency mode is disabled, not creating a
```



Query History	Saved Queries	Results
maestros en jaén		
 		
1	3014	

Figura 5: Consulta con todos los maestros que tiene Jaén.

Pregunta 6.

Escribe una consulta que muestre los centros de **Granada** y el **número de plazas de Profesores técnicos** de cada uno de ellos. Responde con la consulta, el **número de resultados** que hayas obtenido y las plazas de Profesores técnicos que hay en el **Zaidín-Vergeles**.

```

1 SELECT CENTRO, SUM(PLAZAS) AS `Total plazas de Profesores técnicos`
2 FROM plantilla2223
3 WHERE CUERPO LIKE '%PROFESORES TEC. DE FORMACION PROFESIONAL%'
4 AND PROVINCIA = 'GRANADA'
5 AND CUERPO != 'CUERPO'
6 GROUP BY CENTRO;

```

INFO : Compiling command(queryId=hive_20250309174954_7fe3af16-831f-42d2-ad58-200e9c6d2660): SELECT SUM(PLAZAS) AS `Total plazas de Profesores técnicos` FROM plantilla2223 WHERE CUERPO LIKE '%PROFESORES TEC. DE FORMACION PROFESIONAL%' AND PROVINCIA = 'GRANADA' AND CUERPO != 'CUERPO' GROUP BY CENTRO;

Query History

Saved Queries

Results (59)

	centro	total plazas de profesores técnicos
1	18000601-I.E.S. ANTIGUA SEXI	2
2	18001123-I.E.S. PEDRO JIMÉNEZ MONTOYA	21

Figura 6: Consulta con el total de plazas de profesores técnicos.

	centro	total plazas de profesores técnicos
3	18002243-I.E.S. FEDERICO GARCÍA LORCA	6
4	18004288-I.E.S. POLITÉCNICO HERMENEGILDO LANZ	40
5	18004291-I.E.S. PADRE MANJÓN	5
6	18004355-C.P.I.F.P. HURTADO DE MENDOZA	26
7	18004801-I.E.S. PADRE POVEDA	4
8	18005153-I.E.S. ALQUIVIRA	8
9	18005980-I.E.S. VIRGEN DE LA CARIDAD	4
10	18007046-I.E.S. FRANCISCO JAVIER DE BURGOS	5
11	18008464-I.E.S. ULYSSEA	1
12	18008841-I.E.S. CARTUJA	6
13	18009249-I.E.S. SEVERO OCHOA	4
14	18009389-I.E.S. ALBAYZÍN	11
15	18009419-I.E.S. VALLE DE LECRÍN	6
16	18009432-I.E.S. CERRO DE LOS INFANTES	10
17	18009444-I.E.S. HISPANIDAD	12

Figura 7: Resultado de la consulta.

	centro	total plazas de profesores técnicos
18	18009778-I.E.S. ALONSO CANO	3
19	18700013-I.E.S. FRAY LUIS DE GRANADA	1
20	18700232-I.E.S. ALBA LONGA	1
21	18700499-I.E.S. PUERTA DEL MAR	2
22	18700611-I.E.S. FERNANDO DE LOS RÍOS	1
23	18700621-I.E.S. ISABEL LA CATÓLICA	2
24	18700724-I.E.S. LA LAGUNA	2
25	18700761-I.E.S. ALPUJARRA	5
26	18000039-I.E.S. ARICEL	3
27	18000787-I.E.S. LUIS BUENO CRESPO	19
28	18000908-I.E.S. ILIBERIS	5
29	18001147-I.E.S. JOSÉ DE MORA	1
30	18001834-I.E.S. EMILIO MUÑOZ	7
31	18004264-I.E.S. PADRE SUÁREZ	4
32	18004276-I.E.S. ÁNGEL GANIVET	12



Figura 8: Resultado de la consulta.

	centro	total plazas de profesores técnicos
33	18004458-I.E.S. VIRGEN DE LAS NIEVES	36
34	18005141-I.E.S. LA SAGRA	3
35	18005992-I.E.S. MORAIMA	7
36	18007022-I.E.S. LA ZAFRA	27
37	18008257-I.E.S. JIMÉNEZ DE QUESADA	2
38	18009213-I.E.S. ACCI	15
39	18009377-C.P.I.F.P. AYNADAMAR	55
40	18009407-I.E.S. VEGA DE ATARFE	4
41	18009961-I.E.S. MEDITERRÁNEO	8
42	18700037-I.E.S. PEDRO SOTO DE ROJAS	4
43	18700098-I.E.S. ZAIDÍN-VERGELES	27
44	18700293-I.E.S. FRANCISCO AYALA	4
45	18700301-I.E.S. AMÉRICO CASTRO	5
46	18700311-I.E.S. LA MADRAZA	3
47	18700347-I.E.S. LA CONTRAVIESA	1

Figura 9: Resultado de la consulta.

48	18700359-I.E.S. ALCREBITE	1
49	18700414-I.E.S. ALFAGUARA	4
50	18700426-I.E.S. FRANCISCO GINER DE LOS RÍOS	2
51	18700441-I.E.S. MIGUEL DE CERVANTES	1
52	18700451-I.E.S. ALHAMA	6
53	18700463-I.E.S. AL-ÁNDALUS	3
54	18700475-I.E.S. MONTES ORIENTALES	8
55	18700487-I.E.S. HIPONOVA	3
56	18700542-I.E.S. VELETA	1
57	18700566-I.E.S. LA PAZ	3
58	18700694-I.E.S. BULYANA	1
59	18700773-I.E.S. DIEGO DE SILOÉ	7

Figura 10: Resultado de la consulta.

```
1 SELECT SUM(PLAZAS) AS `Plazas en Zaidín-Vergeles`  
2 FROM plantilla2223  
3 WHERE CUERPO LIKE '%PROFESORES TEC. DE FORMACION PROFESIONAL%'  
4     AND PROVINCIA = 'GRANADA'  
5     AND CENTRO = '18700098-I.E.S. ZAIDÍN-VERGELES'  
6     AND CUERPO != 'CUERPO';  
7
```

```
AND PROVINCIA = 'GRANADA'  
AND CENTRO = '18700098-I.E.S. ZAIDÍN-VERGELES'  
AND CUERPO != 'CUERPO'  
INFO : Concurrency mode is disabled, not creating a lock manager  
INFO : Semantic Analysis Completed (retrial = false)
```

Query History

Saved Queries

Results (1)

plazas en zaidín-vergeles



1	27
---	----

Figura 11: Consulta con la plazas de profesores técnicos en el Zaidín-Vergeles.

Pregunta 7.

Escribe una consulta para mostrar el **total de plazas de Andalucía** agrupada por **cuero**. Responde con la consulta y el número de maestros que hayas obtenido.

```
1 SELECT CUERPO, SUM(PLAZAS) AS `Total plazas por cuerpo`  
2 FROM plantilla2223  
3 WHERE CUERPO != 'CUERPO'  
4 GROUP BY CUERPO;
```

Figura 12: Consulta para saber el total de plazas por cuerpo.

	cuerpo	total plazas por cuerpo
1	CATEDRATICOS DE MUSICA Y ARTES ESCENICAS	280
2	INSPECTORES DE EDUCACION	280
3	PERSONAL VARIO ASUMIDO	3
4	PROFESORES DE ENSEÑANZA SECUNDARIA	35401
5	PROFESORES DE MÚSICA Y ARTES ESCÉNICAS	2014
6	PROFESORES ESCUELAS OFICIALES DE IDIOMAS	592
7	EDUCADORES PERMANENTES DE ADULTOS	5
8	MAESTROS	<u>38353</u>
9	MAESTROS DE TALLER ARTES PLASTICAS Y DIS	70
10	PROFESORES DE ARTES PLASTICAS Y DISEÑO	278
11	PROFESORES TEC. DE FORMACION PROFESIONAL	4049

Figura 13: Resultado de la consulta.

Pregunta 8.

Escribe una consulta para mostrar el **total de plazas de Sevilla** agrupada por **cuerpo**. Responde con la consulta y el número de maestros de Sevilla.

```
SELECT CUERPO, SUM(PLAZAS) AS `Total plazas por cuerpo`  
FROM plantilla2223  
WHERE CUERPO != 'CUERPO'  
      AND PROVINCIA = 'SEVILLA'  
GROUP BY CUERPO;
```

Figura 14: Consulta para saber el total de plazas por cuerpo de Sevilla.

	cuerpo	total plazas por cuerpo
1	CATEDRATICOS DE MUSICA Y ARTES ESCENICAS	87
2	INSPECTORES DE EDUCACION	58
3	PERSONAL VARIO ASUMIDO	1
4	PROFESORES DE ENSEÑANZA SECUNDARIA	8162
5	PROFESORES DE MÚSICA Y ARTES ESCÉNICAS	356
6	PROFESORES ESCUELAS OFICIALES DE IDIOMAS	86
7	MAESTROS	8693
8	MAESTROS DE TALLER ARTES PLASTICAS Y DIS	17
9	PROFESORES DE ARTES PLASTICAS Y DISEÑO	57
10	PROFESORES TEC. DE FORMACION PROFESIONAL	882

Figura 15: Resultado de la consulta.

Pregunta 9.

Crea una **tabla administrada** Hive con la tabla externa que has creado en la pregunta 1. Esta tabla estará particionada por el campo cuerpo.

Respuesta: Responde con las **sentencias** que has utilizado para crear esta tabla y aporta una captura de pantalla del directorio `/user/hive/warehouse/aquitubasededatos.db/aquitutablaadministrada` en la que se observen claramente las particiones creadas en HDFS.

```
1|SET hive.exec.dynamic.partition.mode=nonstrict;  
2|
```

Figura 16: Modifico el modo de partición de Hive a nonstrict.


```

1 CREATE TABLE plantilla2223_administrada (
2     PROVINCIA STRING,
3     LOCALIDAD STRING,
4     CENTRO STRING,
5     PUESTO STRING,
6     PLAZAS INT
7 )
8 PARTITIONED BY (CUERPO STRING)
9 ROW FORMAT DELIMITED
10 FIELDS TERMINATED BY ','
11 STORED AS TEXTFILE;

```

Figura 17: Consulta para la creación de la tabla administrada.

```

1 INSERT OVERWRITE TABLE plantilla2223_administrada PARTITION (CUERPO)
2 SELECT PROVINCIA, LOCALIDAD, CENTRO, PUESTO, PLAZAS, CUERPO
3 FROM plantilla2223;

```

Figura 18: Añado todos los valores de la tabla externa a la tabla administrada.

```

[hadoop@ip-172-31-55-204 ~]$ hdfs dfs -ls /user/hive/warehouse/juanmanuelgarciamoyano_2425.db/plantilla2223_administrada
Found 12 items
drwxr-xr-x - juanmi4000 hdfsadmingroup 0 2025-03-09 18:16 /user/hive/warehouse/juanmanuelgarciamoyano_2425.db/
plantilla2223_administrada/cuerpo=CATEDRATICOS DE MUSICA Y ARTES ESCENICAS
drwxr-xr-x - juanmi4000 hdfsadmingroup 0 2025-03-09 18:16 /user/hive/warehouse/juanmanuelgarciamoyano_2425.db/
plantilla2223_administrada/cuerpo=CUERPO
drwxr-xr-x - juanmi4000 hdfsadmingroup 0 2025-03-09 18:16 /user/hive/warehouse/juanmanuelgarciamoyano_2425.db/
plantilla2223_administrada/cuerpo=EDUCADORES PERMANENTES DE ADULTOS
drwxr-xr-x - juanmi4000 hdfsadmingroup 0 2025-03-09 18:16 /user/hive/warehouse/juanmanuelgarciamoyano_2425.db/
plantilla2223_administrada/cuerpo=INSPECTORES DE EDUCACION
drwxr-xr-x - juanmi4000 hdfsadmingroup 0 2025-03-09 18:16 /user/hive/warehouse/juanmanuelgarciamoyano_2425.db/
plantilla2223_administrada/cuerpo=MAESTROS
drwxr-xr-x - juanmi4000 hdfsadmingroup 0 2025-03-09 18:16 /user/hive/warehouse/juanmanuelgarciamoyano_2425.db/
plantilla2223_administrada/cuerpo=MAESTROS DE TALLER ARTES PLASTICAS Y DIS
drwxr-xr-x - juanmi4000 hdfsadmingroup 0 2025-03-09 18:16 /user/hive/warehouse/juanmanuelgarciamoyano_2425.db/
plantilla2223_administrada/cuerpo=PERSONAL VARIO ASUMIDO
drwxr-xr-x - juanmi4000 hdfsadmingroup 0 2025-03-09 18:16 /user/hive/warehouse/juanmanuelgarciamoyano_2425.db/
plantilla2223_administrada/cuerpo=PROFESORES DE ARTES PLASTICAS Y DISEÑO
drwxr-xr-x - juanmi4000 hdfsadmingroup 0 2025-03-09 18:16 /user/hive/warehouse/juanmanuelgarciamoyano_2425.db/
plantilla2223_administrada/cuerpo=PROFESORES DE ENSEÑANZA SECUNDARIA
drwxr-xr-x - juanmi4000 hdfsadmingroup 0 2025-03-09 18:16 /user/hive/warehouse/juanmanuelgarciamoyano_2425.db/
plantilla2223_administrada/cuerpo=PROFESORES DE MÚSICA Y ARTES ESCÉNICAS
drwxr-xr-x - juanmi4000 hdfsadmingroup 0 2025-03-09 18:16 /user/hive/warehouse/juanmanuelgarciamoyano_2425.db/
plantilla2223_administrada/cuerpo=PROFESORES ESCUELAS OFICIALES DE IDIOMAS
drwxr-xr-x - juanmi4000 hdfsadmingroup 0 2025-03-09 18:16 /user/hive/warehouse/juanmanuelgarciamoyano_2425.db/
plantilla2223_administrada/cuerpo=PROFESORES TEC. DE FORMACION PROFESIONAL
[hadoop@ip-172-31-55-204 ~]$

```

Figura 19: Muestro el contenido del directorio plantilla2223_administrada

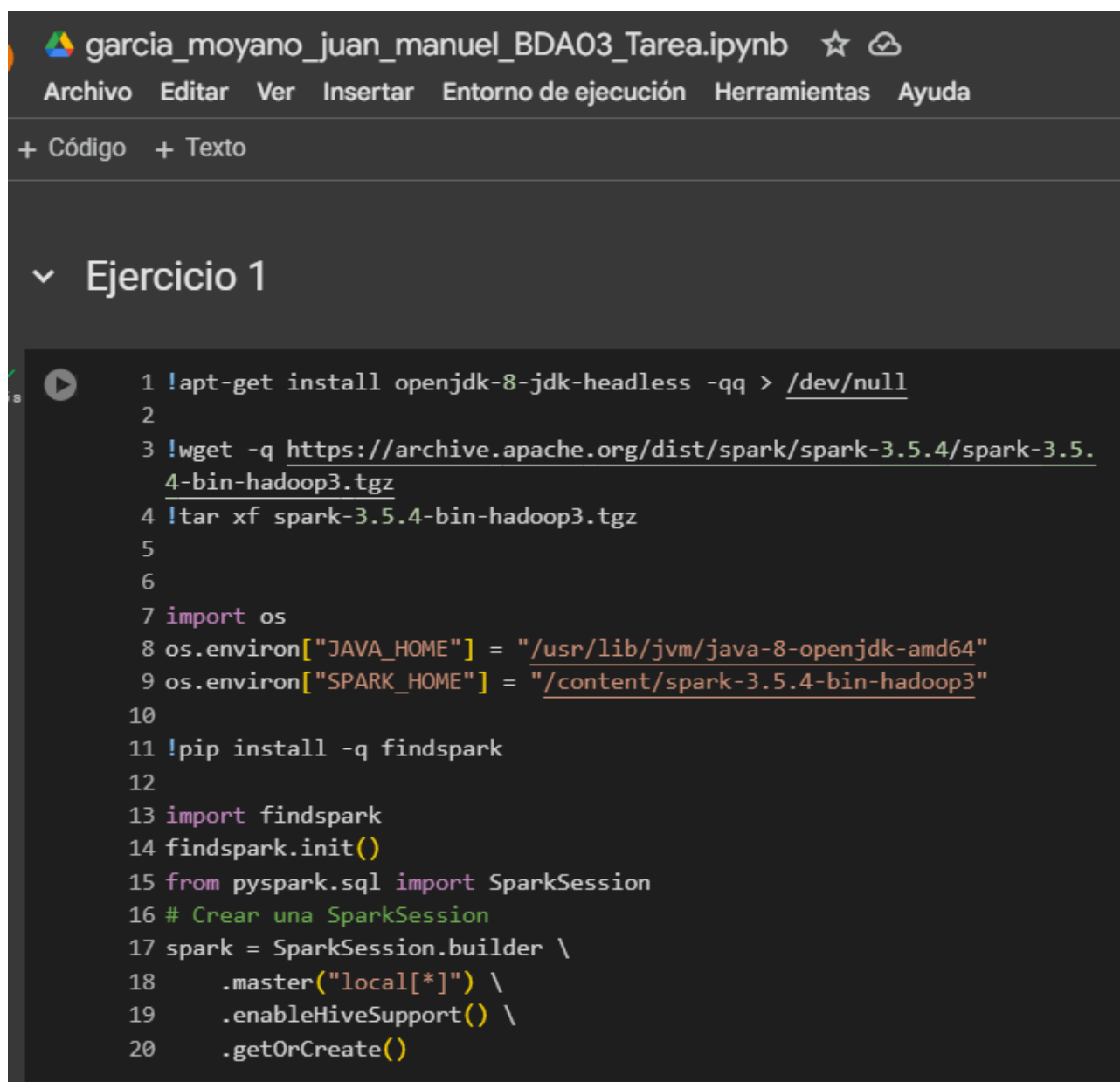
Práctica 2. Spark

Esta práctica la implementarás en un cuaderno de Google Colab con Hadoop y Spark instalados.

Pregunta 1.

Crea una sesión pyspark.sql con soporte Hive.

Respuesta: Las sentencias necesarias para crear esta sesión.



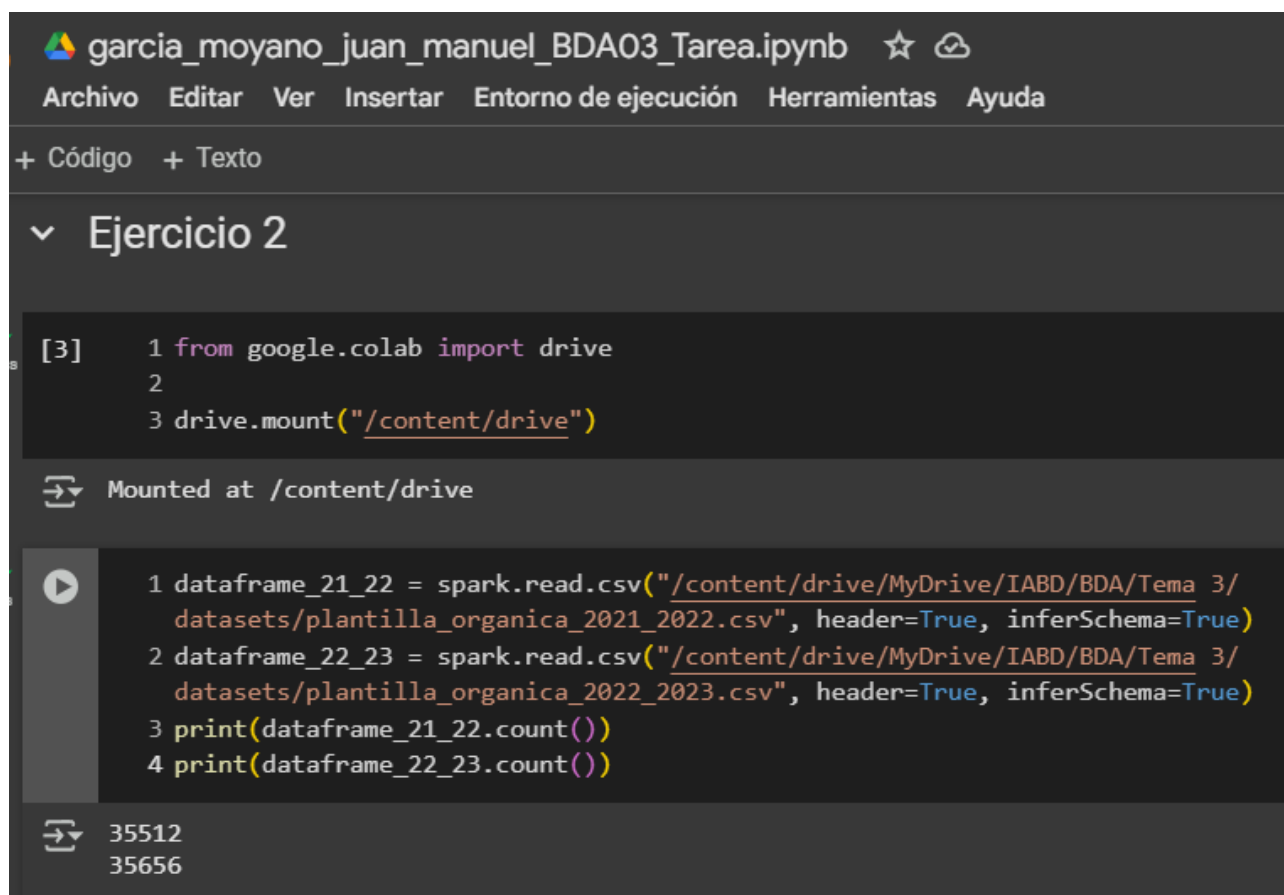
```
1 !apt-get install openjdk-8-jdk-headless -qq > /dev/null
2
3 !wget -q https://archive.apache.org/dist/spark/spark-3.5.4/spark-3.5.4-bin-hadoop3.tgz
4 !tar xf spark-3.5.4-bin-hadoop3.tgz
5
6
7 import os
8 os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
9 os.environ["SPARK_HOME"] = "/content/spark-3.5.4-bin-hadoop3"
10
11 !pip install -q findspark
12
13 import findspark
14 findspark.init()
15 from pyspark.sql import SparkSession
16 # Crear una SparkSession
17 spark = SparkSession.builder \
18     .master("local[*]") \
19     .enableHiveSupport() \
20     .getOrCreate()
```

Figura 20: Creación de una sesión con pyspark.sql y con soporte a Hive.

Pregunta 2.

Crea un DataFrame con el archivo `plantilla2122.csv`. Crea un DataFrame con el archivo `2223.csv`. Cuenta el número de filas de cada uno de los DataFrame.

Respuesta: Instrucción de lectura del primer archivo. Instrucción de lectura del segundo archivo.



The screenshot shows a Jupyter Notebook interface with the title `garcia_moyano_juan_manuel_BDA03_Tarea.ipynb`. The menu bar includes 'Archivo', 'Editar', 'Ver', 'Insertar', 'Entorno de ejecución', 'Herramientas', and 'Ayuda'. Below the menu, there are tabs for '+ Código' and '+ Texto'. The notebook content is titled 'Ejercicio 2'. It contains two code cells. The first cell, labeled '[3]', imports the `drive` module from `google.colab` and mounts the drive at `/content/drive`. The second cell contains four lines of code: it reads two CSV files into DataFrames (`dataframe_21_22` and `dataframe_22_23`), and then prints the row counts for each. The output of the second cell shows the row counts: 35512 for the first DataFrame and 35656 for the second.

```
[3] 1 from google.colab import drive
    2
    3 drive.mount("/content/drive")

Mounted at /content/drive

1 dataframe_21_22 = spark.read.csv("/content/drive/MyDrive/IABD/BDA/Tema 3/
  datasets/plantilla_organica_2021_2022.csv", header=True, inferSchema=True)
2 dataframe_22_23 = spark.read.csv("/content/drive/MyDrive/IABD/BDA/Tema 3/
  datasets/plantilla_organica_2022_2023.csv", header=True, inferSchema=True)
3 print(dataframe_21_22.count())
4 print(dataframe_22_23.count())

35512
35656
```

Figura 21: Importación de los datasets y el número de fila de cada uno.

Pregunta 3.

Mediante instrucciones `SELECT`, obtén un nuevo DataFrame que contenga las provincias, el número de plazas del curso 21-22 por provincia, el número de plazas del curso 22-23 por provincia y la diferencia de plazas entre ambos cursos, ordenado por provincia.

Respuesta: Escribe las instrucciones de creación del DataFrame y el resultado de mostrar el nuevo DataFrame.

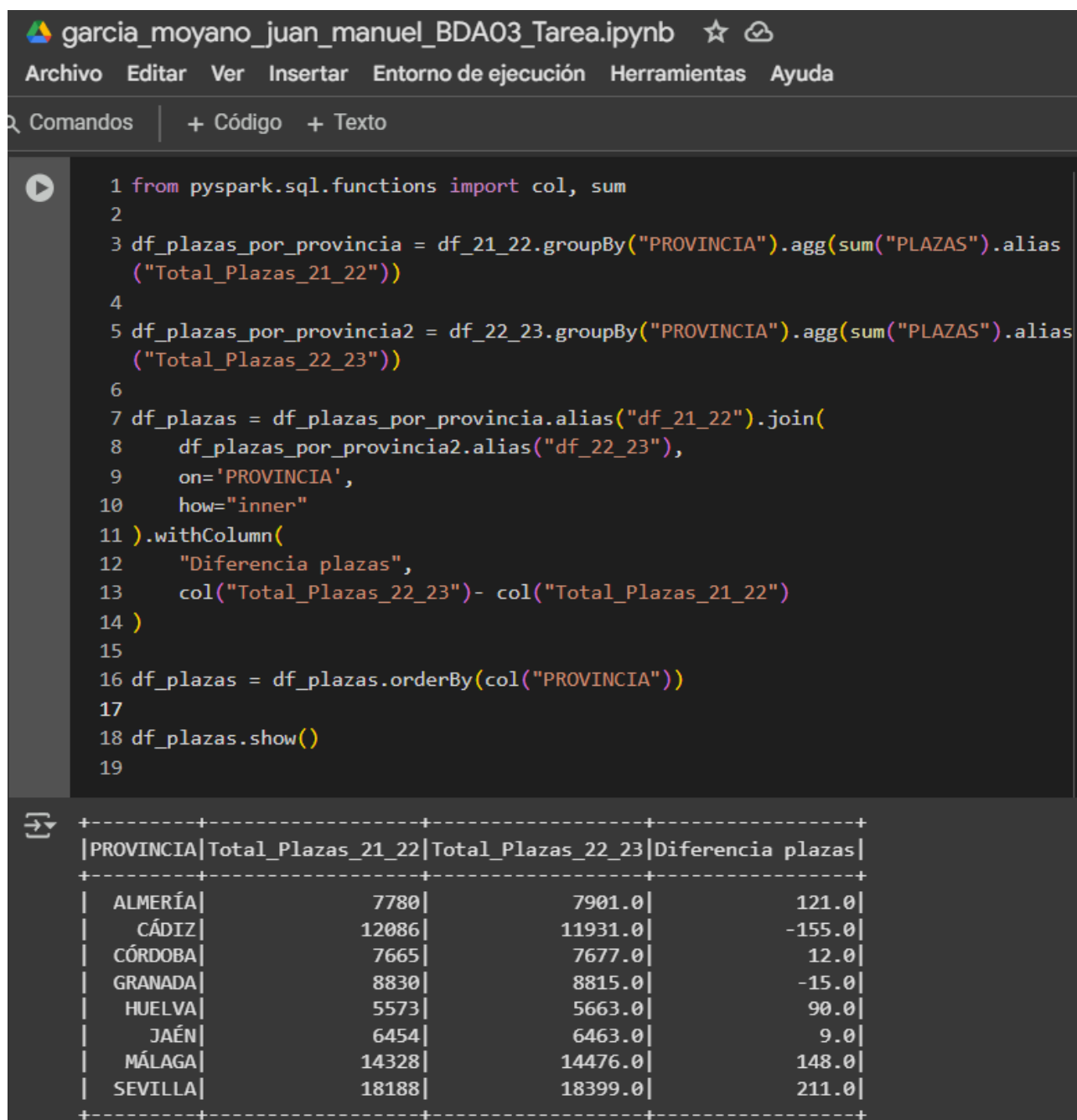
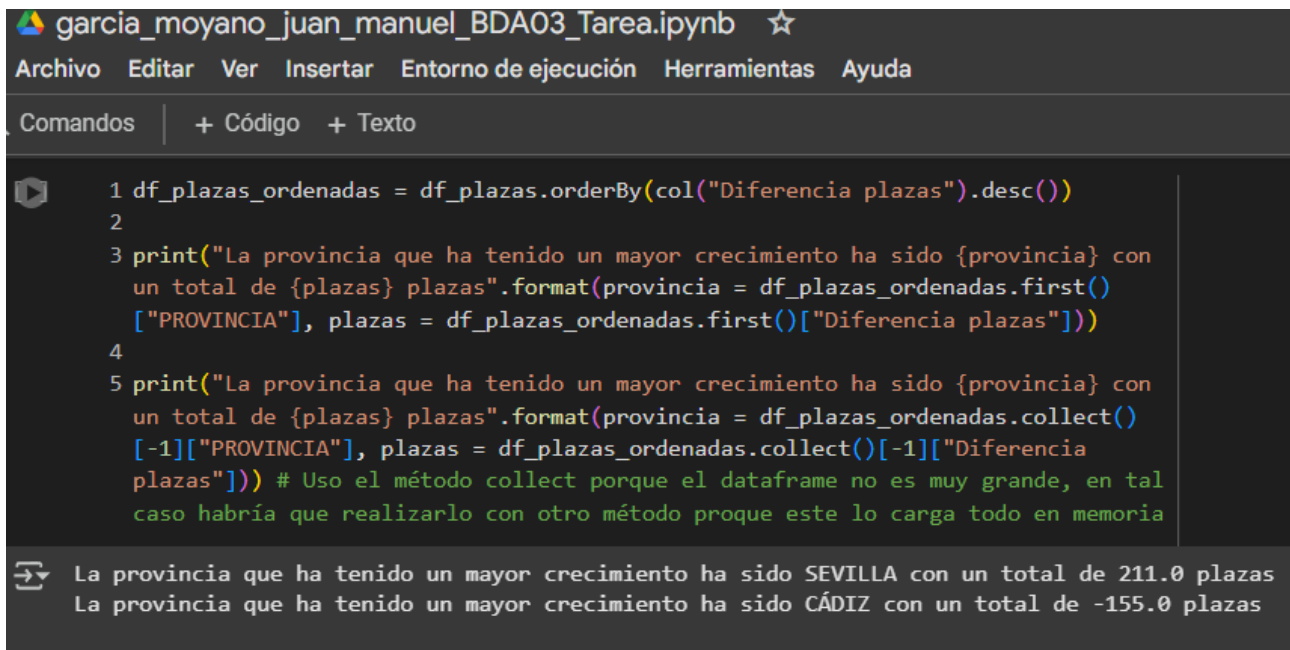


Figura 22: Total de plazas de los cursos 21/22 y 22/23, junto la diferencia de plazas de ambos.

Pregunta 4.

Escribe qué provincia es la que ha tenido un mayor crecimiento de plantilla entre los dos cursos.

Escribe qué provincia es la que ha tenido un menor crecimiento de plantilla entre los dos cursos.



```
garcia_moyano_juan_manuel_BDA03_Tarea.ipynb ☆
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda
Comandos + Código + Texto

1 df_plazas_ordenadas = df_plazas.orderBy(col("Diferencia plazas").desc())
2
3 print("La provincia que ha tenido un mayor crecimiento ha sido {provincia} con
  un total de {plazas} plazas".format(provincia = df_plazas_ordenadas.first()
  ["PROVINCIA"], plazas = df_plazas_ordenadas.first()["Diferencia plazas"]))
4
5 print("La provincia que ha tenido un mayor crecimiento ha sido {provincia} con
  un total de {plazas} plazas".format(provincia = df_plazas_ordenadas.collect()
  [-1]["PROVINCIA"], plazas = df_plazas_ordenadas.collect()[-1]["Diferencia
  plazas"])) # Uso el método collect porque el dataframe no es muy grande, en tal
  caso habría que realizarlo con otro método porque este lo carga todo en memoria

La provincia que ha tenido un mayor crecimiento ha sido SEVILLA con un total de 211.0 plazas
La provincia que ha tenido un mayor crecimiento ha sido CÁDIZ con un total de -155.0 plazas
```

Figura 23: Muestro la provincia que ha tenido un mayor y menor crecimiento de plantilla.

Pregunta 5.

Muestra una gráfica de barras apiladas con las plazas de ambos cursos por provincia.

Respuesta: Pega en el PDF una captura de la gráfica obtenida.

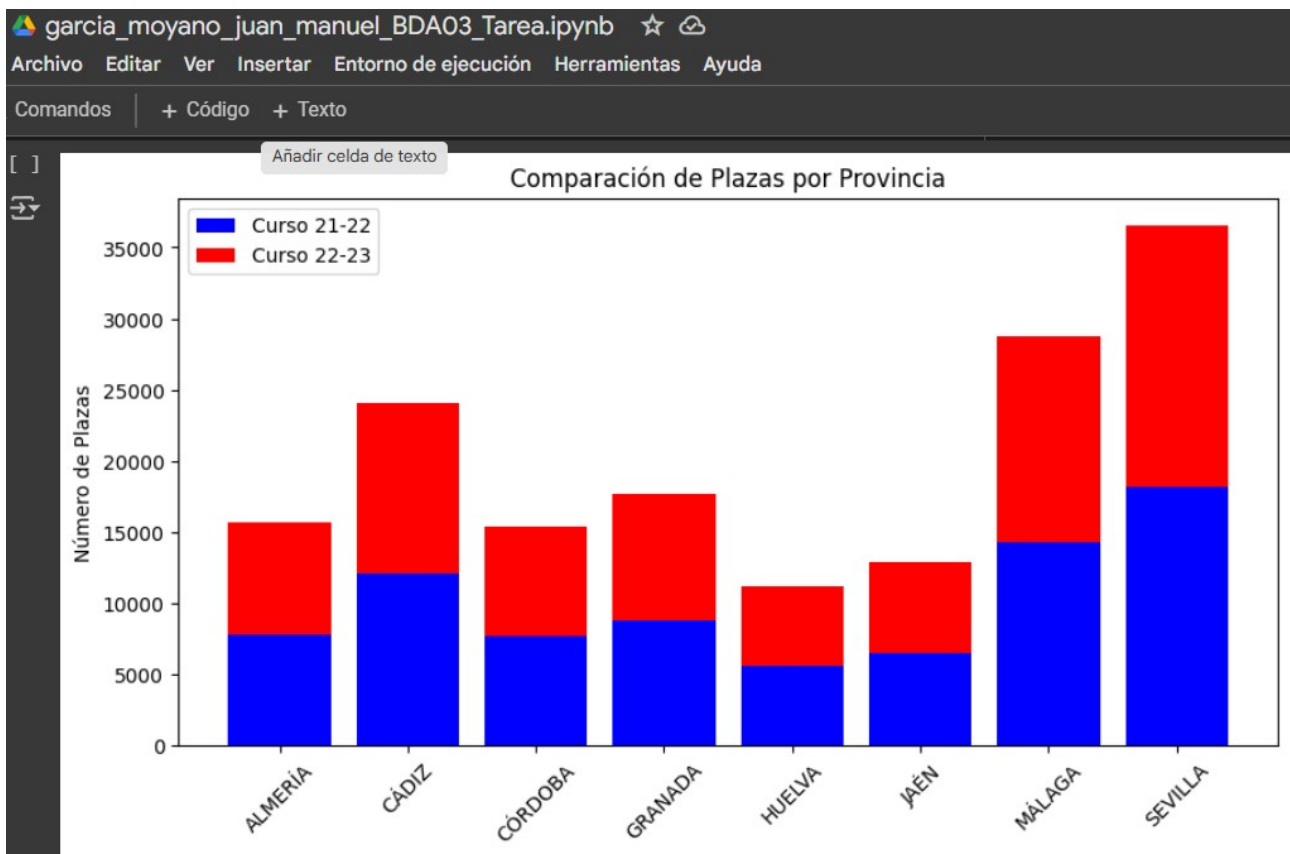


Figura 24: Gráfica de barras apiladas con las plazas de cada provincia y curso.

Pregunta 6.

Guarda en Hive el DataFrame creado en la pregunta 3.

Respuesta: Escribe la instrucción para guardar el DataFrame en Hive y adjunta una captura de pantalla de la sección de archivos con el directorio spark-warehouse desplegado.

```
df_plazas.write.mode("overwrite").format("csv").saveAsTable("plazas_21_22_23")
```

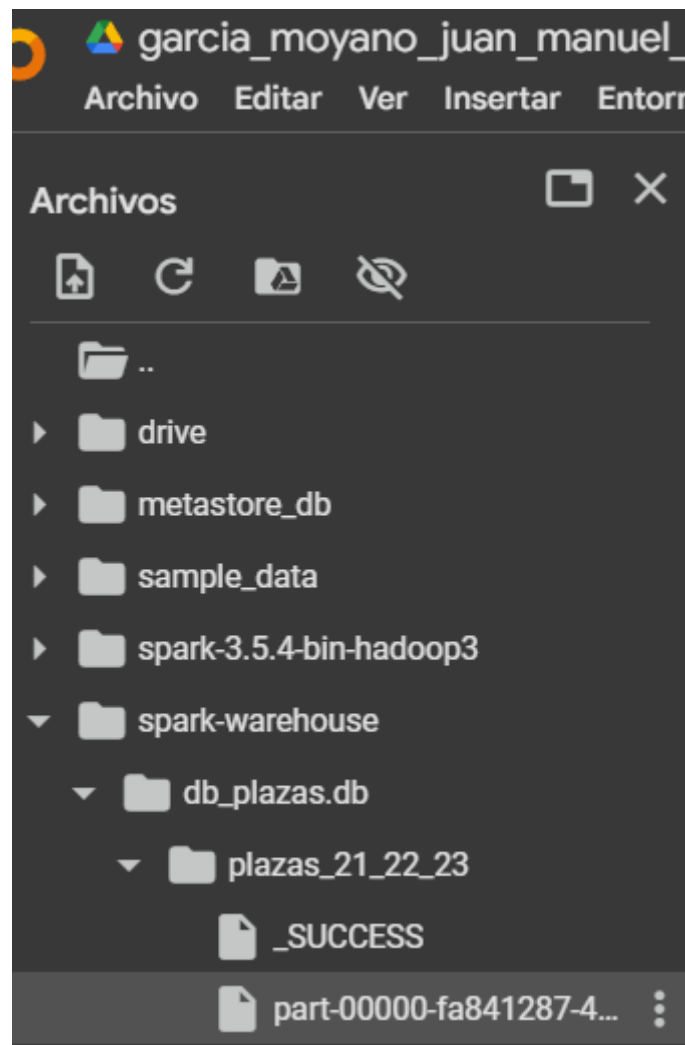


Figura 25: Muestro la sección de archivos para ver si se ha creado correctamente la tabla.