Tarea evaluable CE_5075 8.1 Apartado 1

Big data aplicado

David Ramírez Ruiz
20/05/25

Índice

APARTADO 1 **2**
1 - CONFIGURACIÓN DEL CLÚSTER **2**
2 - QUERIES **7**

APARTADO 1

Vas a trabajar con un dataset que contiene todos los partidos de fútbol entre selecciones nacionales desde 1872 hasta la actualidad.

De los cuatro archivos, solo te interesan dos:

1. **results.csv**: contiene la información general de todos los partidos (equipos, marcador, campeonato, sede, etc.).

2. **goalscorers.csv**: contiene detalles de los goles anotados (partido, equipo, jugador, minuto, si fue penal o en propia portería).

Debes:

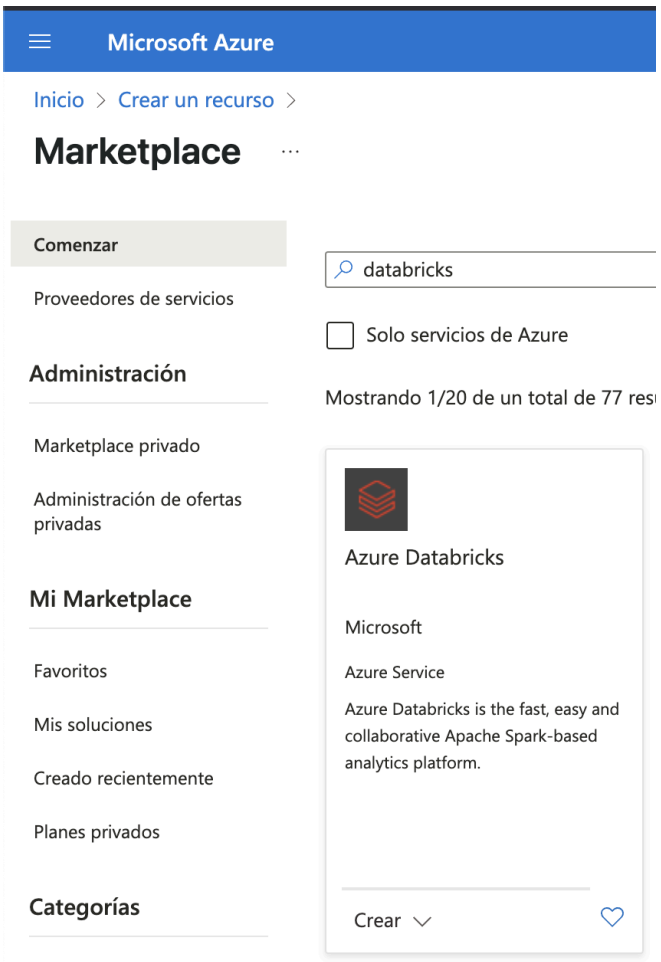
- Usar **Azure Databricks**.
- Crear un **cuaderno (notebook)** que responda a **las mismas preguntas** que en la tarea anterior (con Hive).
- Puedes usar **Python con Spark SQL** o directamente **SQL**.
- Asegúrate de **cargar, preparar y analizar los datos correctamente** para responder a cada pregunta.

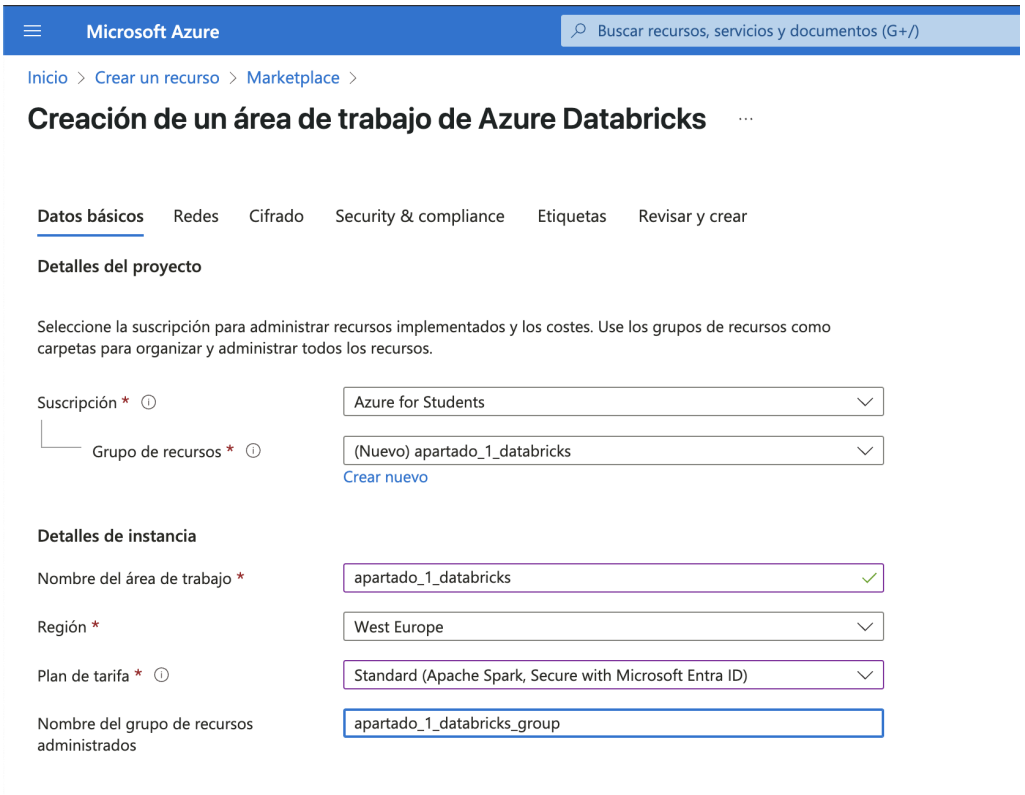
---

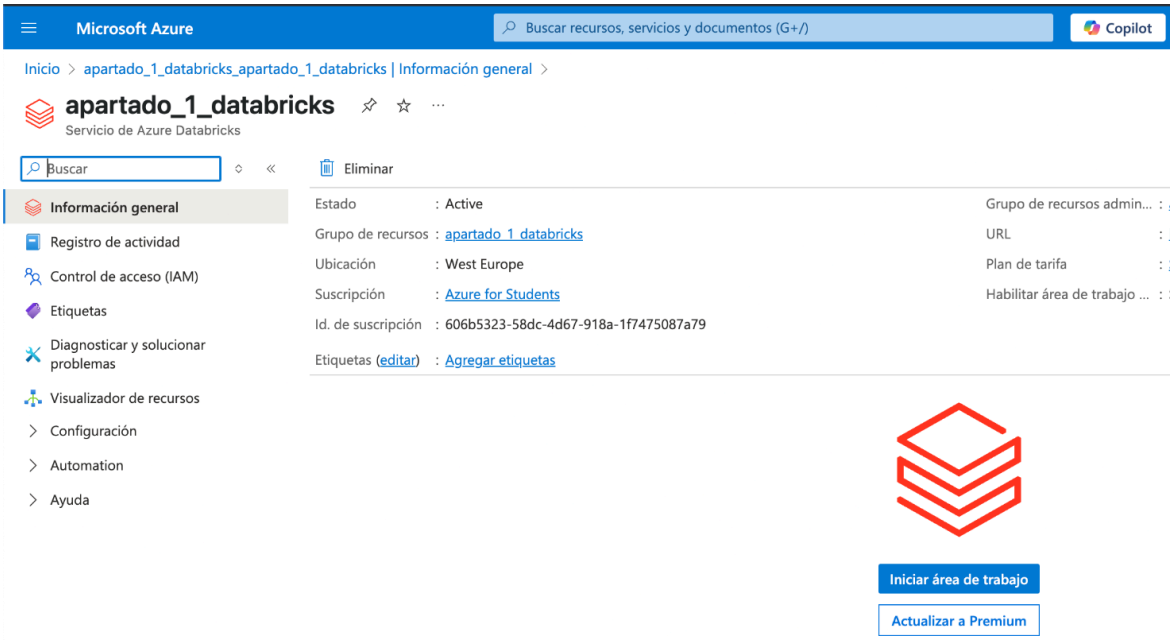
1 - CONFIGURACIÓN DEL CLÚSTER

Empezamos configurando el clúster en Databricks con las instrucciones aprendidas en los apuntes..

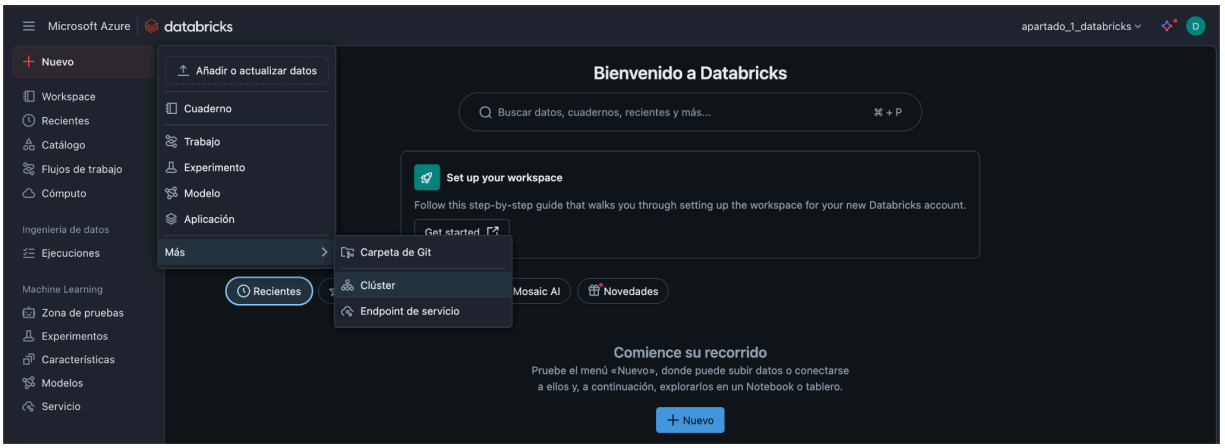
Primero accedemos al Azure Portal y vamos a buscar “databricks” en el marketplace. Ajustamos y rellenamos los datos necesarios y creamos el recurso. Luego pulsamos “acceder al recurso” y podemos iniciar el área de trabajo de databricks.


Microsoft Azure
Inicio > Crear un recurso >
Marketplace
Comenzar
Proveedores de servicios
Administración
Marketplace privado
Administración de ofertas privadas
Mi Marketplace
Favoritos
Mis soluciones
Creado recientemente
Planes privados
Categorías
databricks
Solo servicios de Azure
Mostrando 1/20 de un total de 77 res
Azure Databricks
Microsoft
Azure Service
Azure Databricks is the fast, easy and collaborative Apache Spark-based analytics platform.
Crear

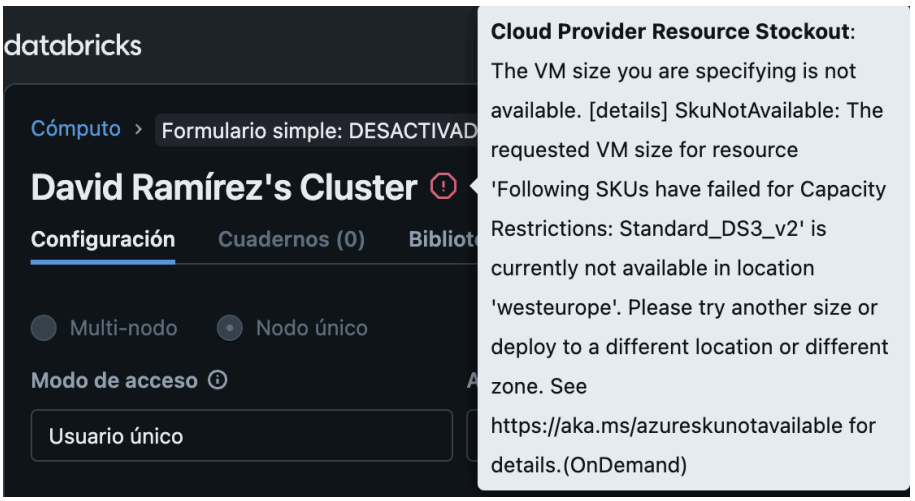

Microsoft Azure
Buscar recursos, servicios y documentos (G+/)
Inicio > Crear un recurso > Marketplace >
Creación de un área de trabajo de Azure Databricks
Datos básicos
Redes
Cifrado
Security & compliance
Etiquetas
Revisar y crear
Detalles del proyecto
Seleccione la suscripción para administrar recursos implementados y los costes. Use los grupos de recursos como carpetas para organizar y administrar todos los recursos.
Suscripción *
Azure for Students
Grupo de recursos *
(Nuevo) apartado_1_databricks
Crear nuevo
Detalles de instancia
Nombre del área de trabajo *
apartado_1_databricks
Región *
West Europe
Plan de tarifa *
Standard (Apache Spark, Secure with Microsoft Entra ID)
Nombre del grupo de recursos administrados
apartado_1_databricks_group



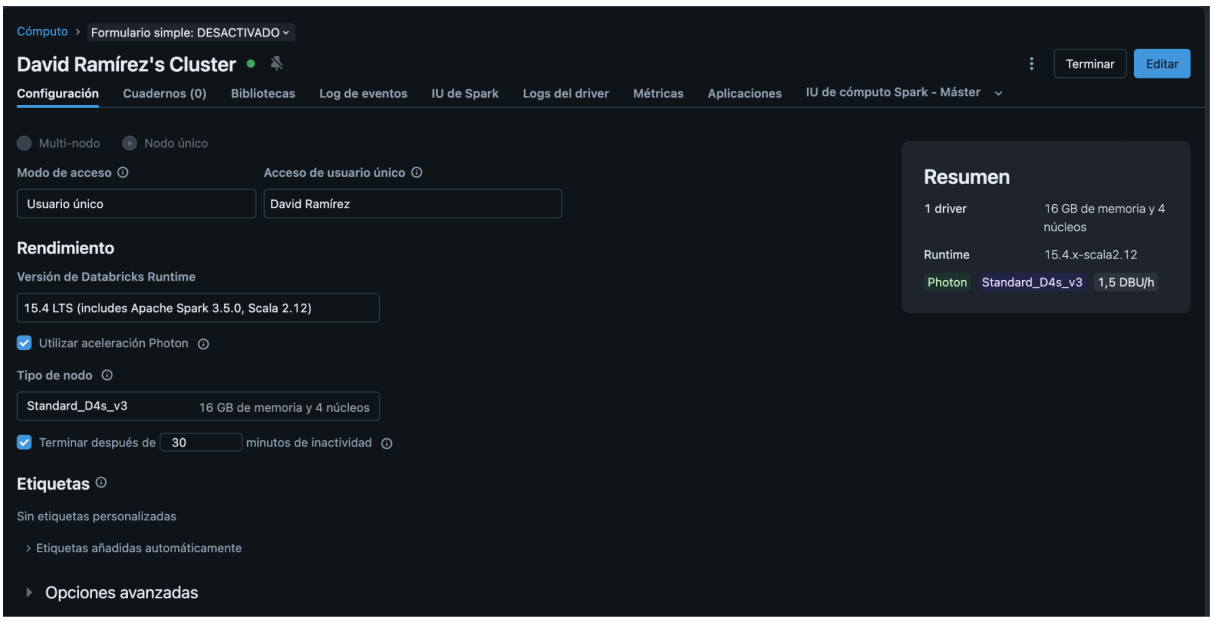
Lo siguiente va a ser crear un clúster de Spark. El clúster solamente dispondrá de un nodo ya que es lo que nos permite nuestra suscripción de Azure for Students. Pulsando en “Nuevo” se abren una serie de opciones, y seleccionamos “Clúster”.

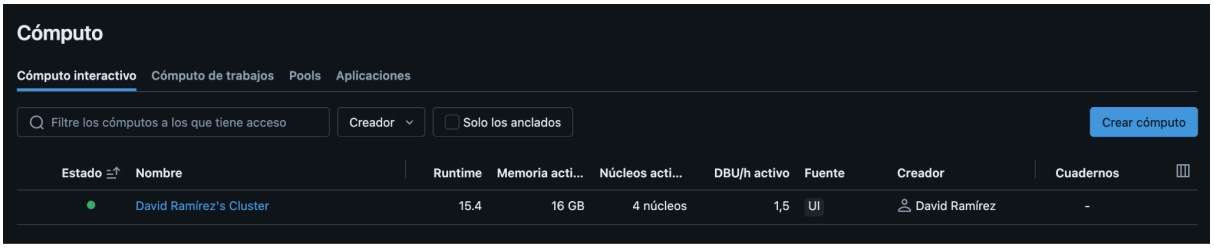


Al seguir la configuración de los apuntes me he encontrado con el siguiente problema:

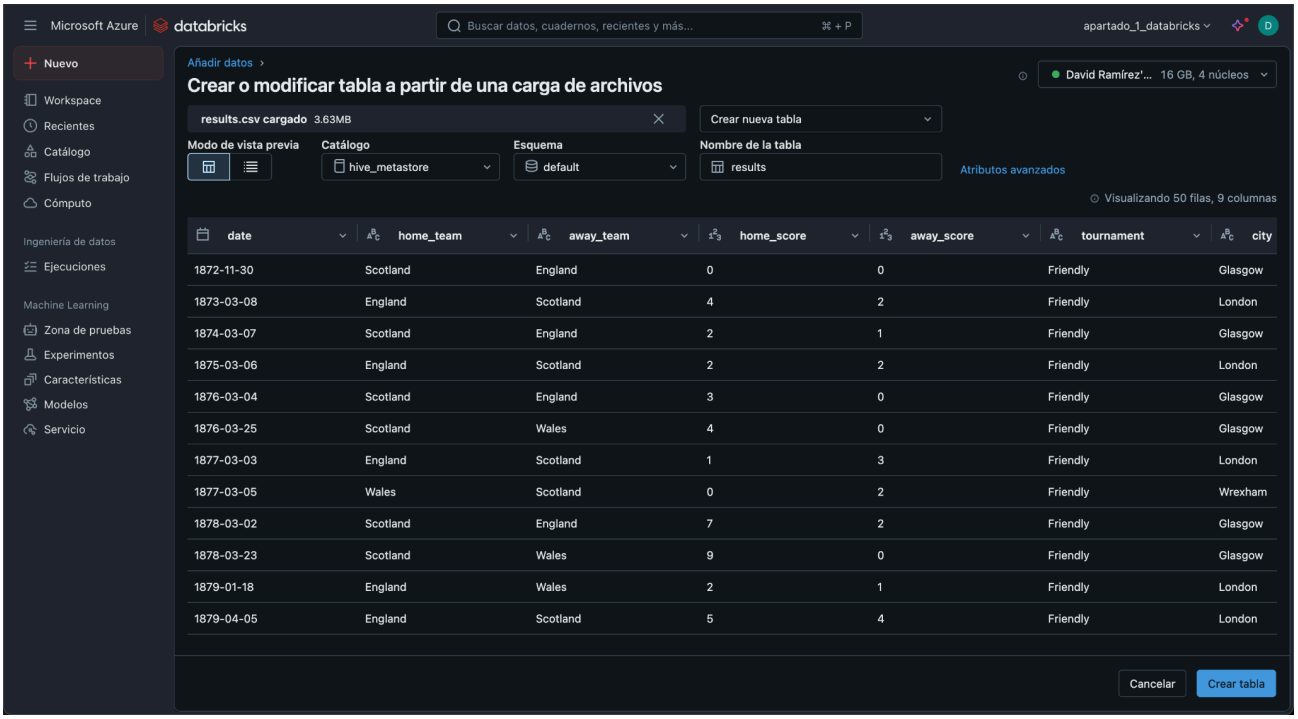


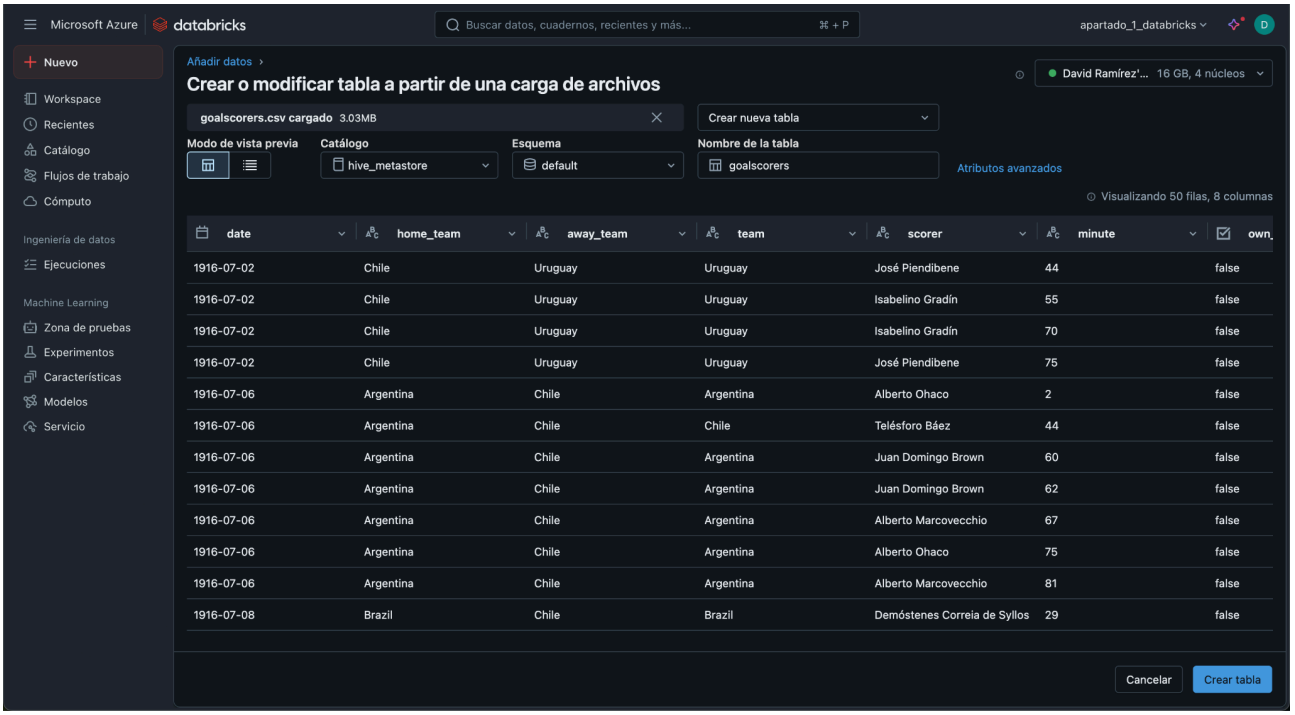
Parece que no es posible usar “Standard_DS3_v2” con localización en West Europe. Por lo tanto he decidido usar “Standard_D4s_v3” con 16 Gb de memoria y 4 núcleos.

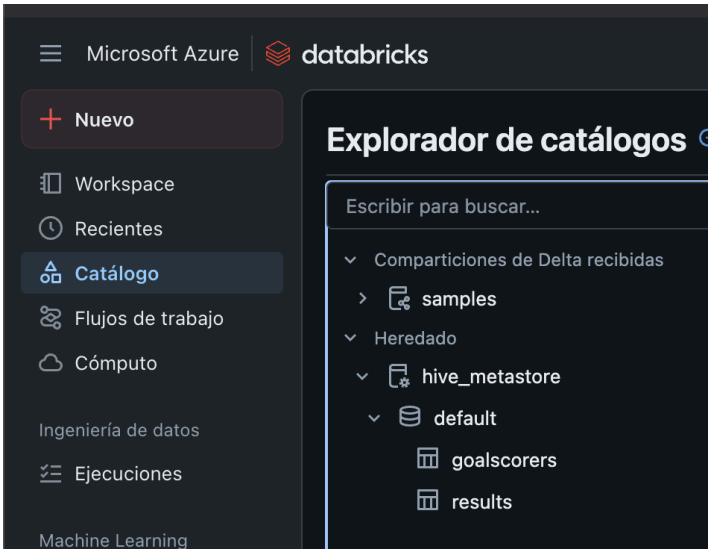




Ahora sí, el clúster está preparado, vamos a añadir los datos.

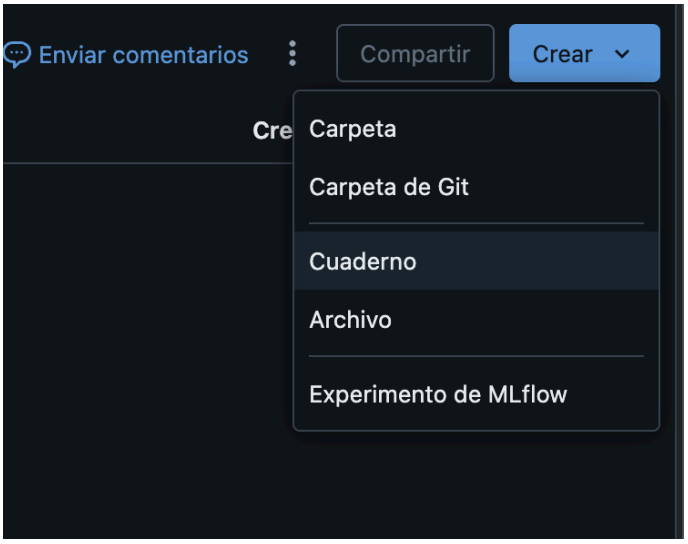

Microsoft Azure
databricks
Buscar datos, cuadernos, recientes y más...
apartado_1_databricks
Nuevo
Workspace
Recientes
Catálogo
Flujos de trabajo
Cómputo
Ingeniería de datos
Ejecuciones
Machine Learning
Zona de pruebas
Experimentos
Características
Modelos
Servicio
Añadir datos
Crear o modificar tabla a partir de una carga de archivos
David Ramírez'... 16 GB, 4 núcleos
results.csv cargado 3.63MB
Crear nueva tabla
Modo de vista previa
Catálogo
hive_metastore
Esquema
default
Nombre de la tabla
results
Atributos avanzados
Visualizando 50 filas, 9 columnas
date home_team away_team home_score away_score tournament city
1872-11-30 Scotland England 0 0 Friendly Glasgow
1873-03-08 England Scotland 4 2 Friendly London
1874-03-07 Scotland England 2 1 Friendly Glasgow
1875-03-06 England Scotland 2 2 Friendly London
1876-03-04 Scotland England 3 0 Friendly Glasgow
1876-03-25 Scotland Wales 4 0 Friendly Glasgow
1877-03-03 England Scotland 1 3 Friendly London
1877-03-05 Wales Scotland 0 2 Friendly Wrexham
1878-03-02 Scotland England 7 2 Friendly Glasgow
1878-03-23 Scotland Wales 9 0 Friendly Glasgow
1879-01-18 England Wales 2 1 Friendly London
1879-04-05 England Scotland 5 4 Friendly London
Cancelar
Crear tabla


Microsoft Azure
databricks
Buscar datos, cuadernos, recientes y más...
apartado_1_databricks
Nuevo
Workspace
Recientes
Catálogo
Flujos de trabajo
Cómputo
Ingeniería de datos
Ejecuciones
Machine Learning
Zona de pruebas
Experimentos
Características
Modelos
Servicio
Añadir datos
Crear o modificar tabla a partir de una carga de archivos
David Ramírez'... 16 GB, 4 núcleos
goalscorers.csv cargado 3.03MB
Crear nueva tabla
Modo de vista previa
Catálogo
hive_metastore
Esquema
default
Nombre de la tabla
goalscorers
Atributos avanzados
Visualizando 50 filas, 8 columnas
date home_team away_team team scorer minute own
1916-07-02 Chile Uruguay Uruguay José Piendibene 44 false
1916-07-02 Chile Uruguay Uruguay Isabelino Gradín 55 false
1916-07-02 Chile Uruguay Uruguay Isabelino Gradín 70 false
1916-07-02 Chile Uruguay Uruguay José Piendibene 75 false
1916-07-06 Argentina Chile Argentina Alberto Ohaco 2 false
1916-07-06 Argentina Chile Chile Telésforo Báez 44 false
1916-07-06 Argentina Chile Argentina Juan Domingo Brown 60 false
1916-07-06 Argentina Chile Argentina Juan Domingo Brown 62 false
1916-07-06 Argentina Chile Argentina Alberto Marcovecchio 67 false
1916-07-06 Argentina Chile Argentina Alberto Ohaco 75 false
1916-07-06 Argentina Chile Argentina Alberto Marcovecchio 81 false
1916-07-08 Brazil Chile Brazil Demóstenes Correia de Syllos 29 false
Cancelar
Crear tabla


Microsoft Azure
databricks
Nuevo
Workspace
Recientes
Catálogo
Flujos de trabajo
Cómputo
Ingeniería de datos
Ejecuciones
Machine Learning
Explorador de catálogos
Escribir para buscar...
Comparticiones de Delta recibidas
samples
Heredado
hive_metastore
default
goalscorers
results

2 - QUERIES

Una vez tenemos todo configurado creamos un cuaderno donde ejecutaremos las consultas de los datos. Pulsamos “Nuevo” y después “Notebook”



Ahora vamos a cargar los datos en los data frames usando Spark

1. Número de goles que ha marcado Lionel Messi (sin contar goles en propia puerta).

```
▶  ✓ Ahora mismo (2 s)                                    5

df_goalscorers.createOrReplaceTempView('goalscorers')
spark.sql('SELECT COUNT(*) AS goals_from_messi FROM goalscorers WHERE scorer = "Lionel Messi" AND own_goal = "FALSE"').show()

▸ (2) trabajos de Spark

+---------------+
|goals_from_messi|
+---------------+
|             55|
+---------------+
```

2. Listado de los 5 partidos más recientes que ha jugado la selección española.

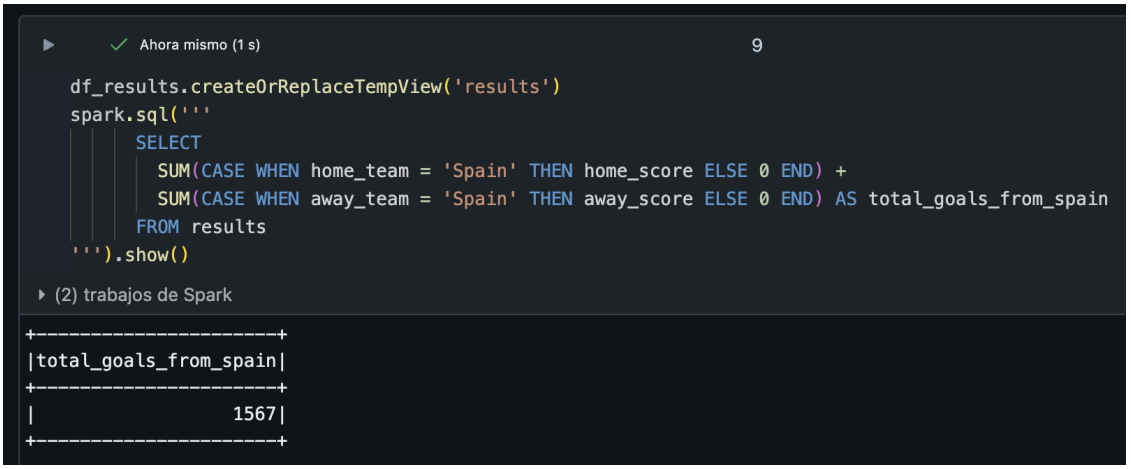
```
▶  ✓ Hace 3 minutos (1 s)                                 7

df_results.createOrReplaceTempView("goalscorers")
spark.sql("SELECT * FROM goalscorers WHERE home_team = 'Spain' OR away_team = 'Spain' ORDER BY date DESC LIMIT 5").show()

▸ (1) trabajos de Spark

+----------+-----------+-----------+----------+----------+----------------+--------------------+-----------+-------+
|      date|  home_team|  away_team|home_score|away_score|      tournament|                city|    country|neutral|
+----------+-----------+-----------+----------+----------+----------------+--------------------+-----------+-------+
|2025-03-23|      Spain|Netherlands|         3|         3|UEFA Nations League|            Valencia|      Spain|  false|
|2025-03-20|Netherlands|      Spain|         2|         2|UEFA Nations League|           Rotterdam|Netherlands|  false|
|2024-11-18|      Spain|Switzerland|         3|         2|UEFA Nations League|Santa Cruz de Ten...|      Spain|  false|
|2024-11-15|    Denmark|      Spain|         1|         2|UEFA Nations League|          Copenhagen|    Denmark|  false|
|2024-10-15|      Spain|     Serbia|         3|         0|UEFA Nations League|             Cordoba|      Spain|  false|
+----------+-----------+-----------+----------+----------+----------------+--------------------+-----------+-------+
```

3. Número de goles que ha marcado España en toda su historia. Esta información debe extraerse del archivo results.csv, ya que goalscorers.csv no contiene todos los goles.



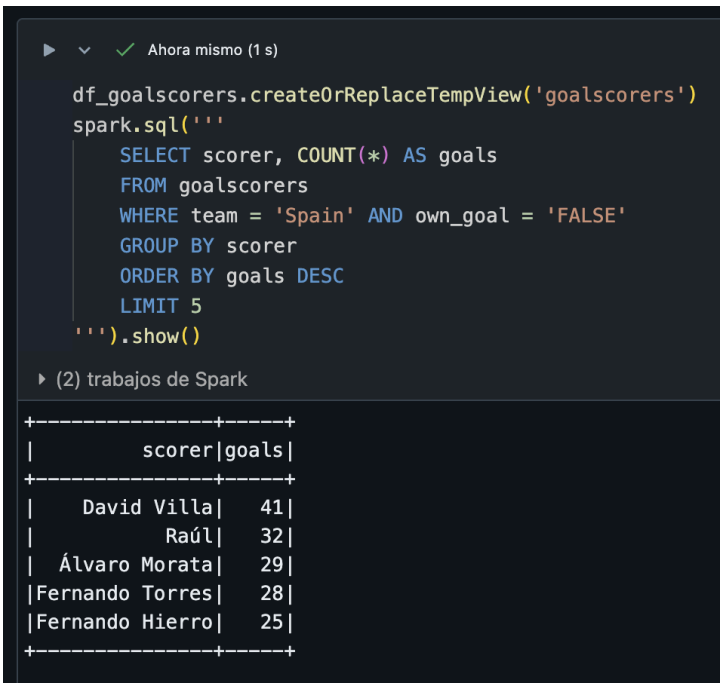
```
▶  ✓ Ahora mismo (1 s)                                    9

df_results.createOrReplaceTempView('results')
spark.sql('''
        SELECT
          SUM(CASE WHEN home_team = 'Spain' THEN home_score ELSE 0 END) +
          SUM(CASE WHEN away_team = 'Spain' THEN away_score ELSE 0 END) AS total_goals_from_spain
        FROM results
''').show()

▸ (2) trabajos de Spark

+---------------------+
|total_goals_from_spain|
+---------------------+
|                 1567|
+---------------------+
```

4. Listado de los 5 máximos goleadores con la selección española (sin contar goles en propia puerta).



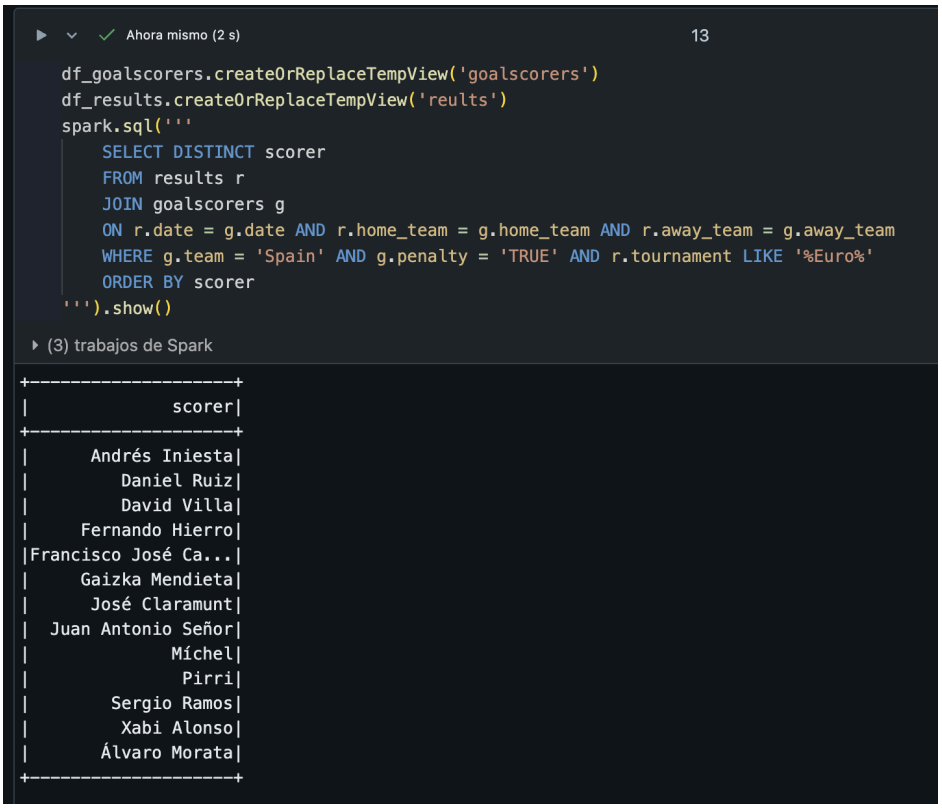
```
Ahora mismo (1 s)

df_goalscorers.createOrReplaceTempView('goalscorers')
spark.sql('''
    SELECT scorer, COUNT(*) AS goals
    FROM goalscorers
    WHERE team = 'Spain' AND own_goal = 'FALSE'
    GROUP BY scorer
    ORDER BY goals DESC
    LIMIT 5
''').show()

(2) trabajos de Spark

+---------------+-----+
|         scorer|goals|
+---------------+-----+
|    David Villa|   41|
|           Raúl|   32|
|  Álvaro Morata|   29|
|Fernando Torres|   28|
|Fernando Hierro|   25|
+---------------+-----+
```

5. Listado de los jugadores españoles que han marcado algún gol de penalti en una Eurocopa (UEFA Euro), ordenados alfabéticamente.



```
Ahora mismo (2 s)                                                    13

df_goalscorers.createOrReplaceTempView('goalscorers')
df_results.createOrReplaceTempView('reults')
spark.sql('''
    SELECT DISTINCT scorer
    FROM results r
    JOIN goalscorers g
    ON r.date = g.date AND r.home_team = g.home_team AND r.away_team = g.away_team
    WHERE g.team = 'Spain' AND g.penalty = 'TRUE' AND r.tournament LIKE '%Euro%'
    ORDER BY scorer
''').show()

(3) trabajos de Spark

+--------------------+
|              scorer|
+--------------------+
|      Andrés Iniesta|
|         Daniel Ruiz|
|         David Villa|
|     Fernando Hierro|
|Francisco José Ca...|
|     Gaizka Mendieta|
|      José Claramunt|
|  Juan Antonio Señor|
|              Míchel|
|               Pirri|
|        Sergio Ramos|
|         Xabi Alonso|
|       Álvaro Morata|
+--------------------+
```

6. Listado de los 5 máximos goleadores de las fases finales de los mundiales (FIFA World Cup) (sin contar goles en propia puerta).