

A globe with a circuit board pattern and the text 'BIG DATA' overlaid. The globe is dark with a grid of lines and various electronic symbols like a microchip, a battery, and a signal wave. The text 'BIG DATA' is in a large, light blue, sans-serif font.

BIG DATA

PROYECTO FINAL PRIMER AÑO BIG DATA

ESPECIALIZACIÓN BIG DATA & IA

GERARDO BLÁZQUEZ MORENO

Proyecto Final. Especialización Big Data & IA

ÍNDICE

1. Introducción.....	01
1.1. Ingesta con S3	02
1.2. AWS glue	03
1.3. AWS glue visual ETL	05
2. Almacenamiento.....	09
2.1. Almacenamiento con Athena	09
3. Almacenamiento.....	12
3.1. Almacenamiento con Spark.....	12
3.1. Almacenamiento con Jupyter Google drive.....	15
4. Análisis.....	16
4.1. Análisis con power BI.....	16
5. Conclusión.....	20

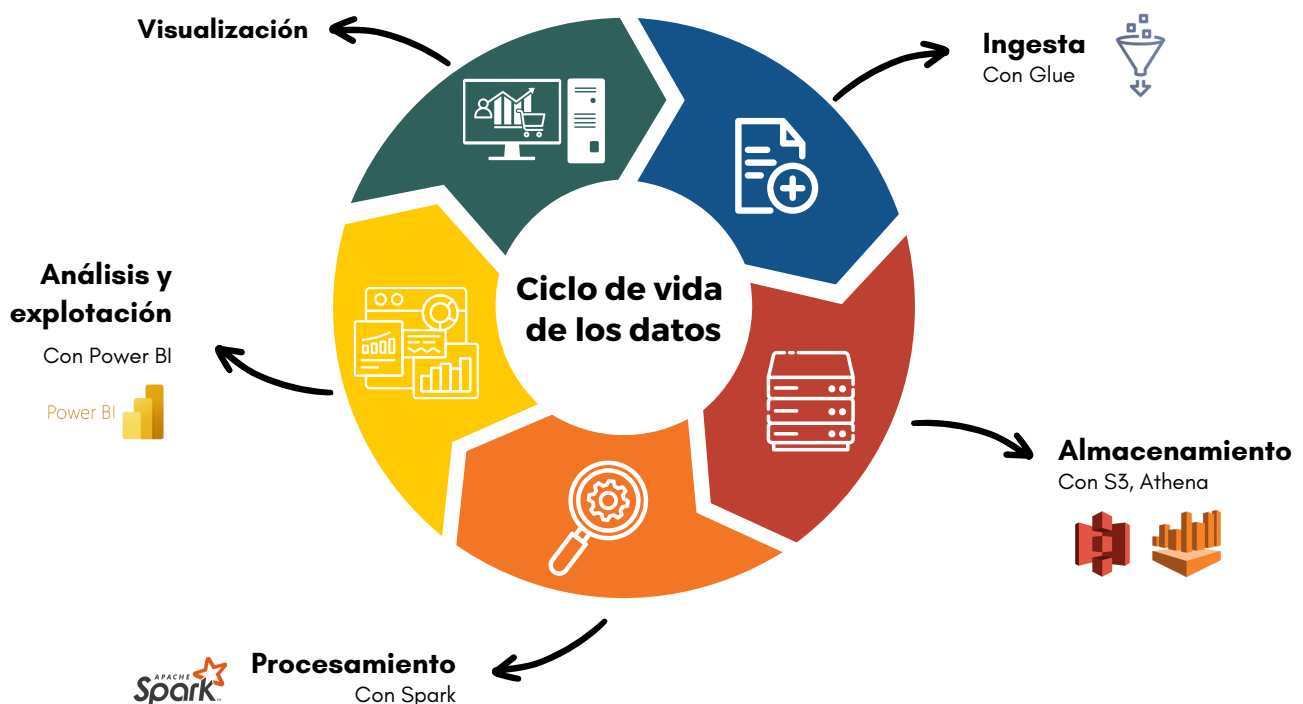
Proyecto Final. Especialización Big Data & IA

Caso de uso

Este proyecto tiene como objetivo **analizar** y **optimizar** la gestión de eventos (*conciertos, partidos de fútbol, festivales y musicales*) de un cliente que tiene un **estadio** inteligente.

Se recopilarán y procesarán datos relacionados con: *entradas vendidas, entradas finalmente utilizadas, consumo en puntos de venta, costes de sueldo de empleados, mantenimiento, procedencia principal de los asistentes, opiniones de asistentes...*

A través de las distintas etapas del ciclo de Big Data, se pretende mejorar la planificación de futuros eventos, detectar patrones de comportamiento de los asistentes, analizando que eventos han sido más rentables.



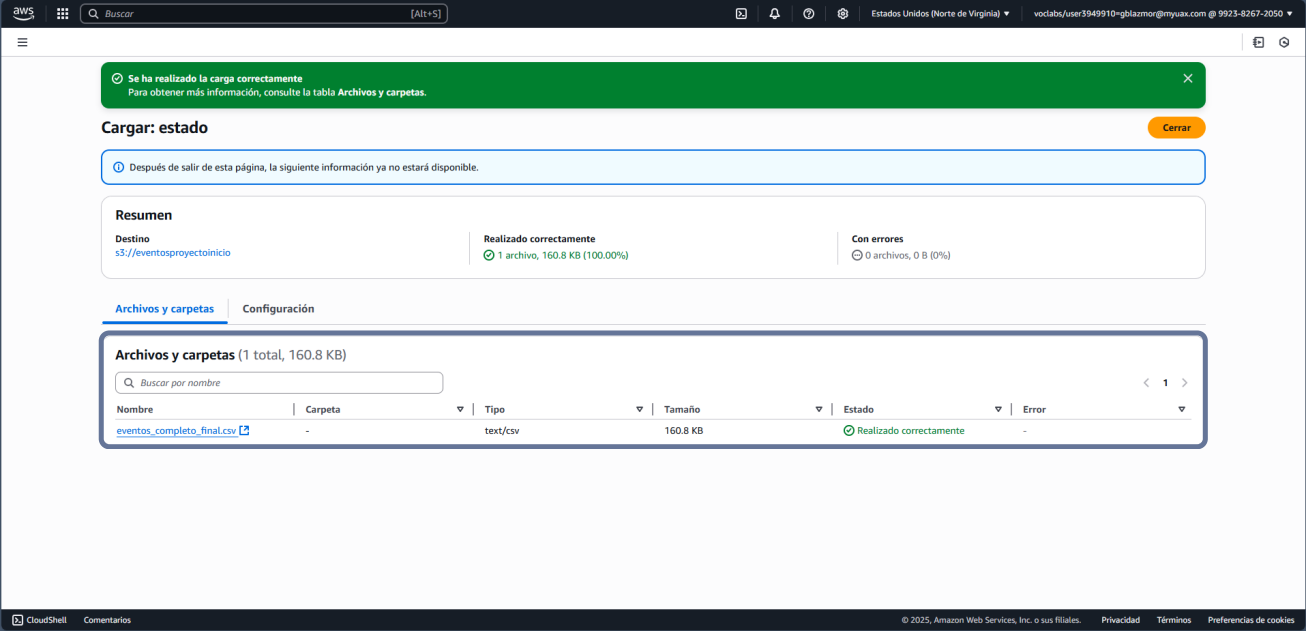
Desarrollaré el proyecto con las siguientes herramientas:

- **Ingesta:** ingesta de dato en **batch** (AWS **GLUE**).
 - **Glue:** Data Catalog y Glue Studio (con **Glue Data Quality**).
- **Almacenamiento:** Utilizaré para almacenamiento **AWS S3** pero lo complementaré con **Athena**. Se requiere hacer una operación sobre el almacenamiento, ejemplo implementar un **ciclo de vida en S3, crear tablas de Athena** con datos de S3 y hacer una **query**, utilizar Redshift desde S3, crear un modelo de ML con Redshift.
- **Procesamiento:** En este caso con **spark** en AWS EMR.
- **Análisis y Explotación:** Se requiere hacer un cuadro de mandos de visualización: en este caso con **Power BI**.

Proyecto Final. Especialización Big Data & IA

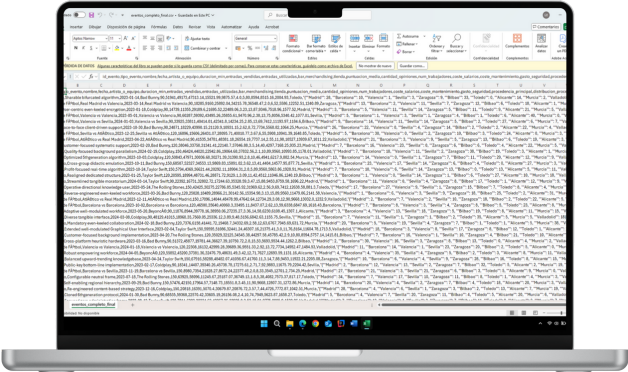
1. Ingesta con AWS s3

Ingesta de datos con **S3**, con el archivo `eventos_completo_final.csv`



Columnas iniciales

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	id_evento	tipo_evento	nombre	fecha	artista_o_equipo	duracion_min	entradas_vendidas	entradas_utilizadas	bar	merchandising	tienda	puntuacion_media	cantidad_opiniones	num_trabajadores	coste_salarios	coste_mantenimiento	gasto_seguridad	procedencia_principal	distribucion_procedencias			



Proyecto Final. Especialización Big Data & IA

1. Ingesta con AWS GLUE

El ciclo de vida de los datos en AWS S3 se refiere a la configuración de reglas que definen acciones que Amazon S3 aplica a un grupo de objetos. Las reglas del ciclo de vida se ejecutan una vez al día.

Para administrar los objetos de forma que se almacenen de manera rentable durante todo el ciclo de vida, configure el ciclo de vida. Una configuración de ciclo de vida es un conjunto de reglas que definen acciones que Amazon S3 aplica a un grupo de objetos. Las reglas del ciclo de vida se ejecutan una vez al día.

Reglas de ciclo de vida

Utilice reglas de ciclo de vida para definir las acciones que desea que Amazon S3 realice durante la vida útil de un objeto, como la transición de objetos a otra clase de almacenamiento, su archivado o su eliminación después de un periodo de tiempo especificado. [Obtenga más información](#)

Nombre de la regla de ciclo de vida | Estado | Ámbito | Acciones de la versión actual... | Acciones de versiones no actuales... | Marcadores de eliminación... | Cargas de multipartes...

No hay reglas de ciclo de vida.

No hay reglas del ciclo de vida para este bucket.

[Crear la regla del ciclo de vida](#)

Reglas de replicación (0)

Utilice las reglas de replicación para definir las opciones que desea que Amazon S3 aplique durante la replicación, como el cifrado del lado del servidor, la titularidad de la réplica, la transición de réplicas a otra clase de almacenamiento y mucho más. [Más información](#)

Nombre de la regla de replicación | Estado | Bucket de destino | Región de destino | Prioridad | Ámbito | Clase de almacenamiento | Propietario de la réplica | Control del tiempo de replicación | Objetos cifrados por KMS (SSE-KMS o DSSE-KMS) | Sincronización de modificación de réplicas

No hay reglas de replicación.

No tiene ninguna regla en la configuración de replicación.

[Crear regla de replicación](#)

Establezco que a los 30 días los datos pasen a otra temperatura del dato y a los 10 años de no ser utilizados sean eliminados automáticamente

Configuración del ciclo de vida

actuales se transferirán.

[Agregar transición](#)

Hacer que venzan las versiones actuales de los objetos

Para los buckets con control de versiones habilitado, Amazon S3 agrega un marcador de eliminación y la versión actual de un objeto se retiene como una versión no actual. En el caso de los buckets sin control de versiones, Amazon S3 elimina permanentemente el objeto. [Obtenga más información](#)

Días después de la creación del objeto

3600

Revisar acciones de transición y vencimiento

Acciones de la versión actual

Día 0

- Objetos cargados

↓

Día 30

- Los objetos se transfieren a Estándar - Acceso poco frecuente

↓

Día 3600

- Los objetos vencen

Acciones de versiones no actuales

Día 0

- Los objetos se desactualizan

↓

Día 30

- Se conservan las 1 versiones no actuales más recientes
- Todas transfieren a Estándar - Acceso poco frecuente

[Cancelar](#) [Crear regla](#)

Proyecto Final. Especialización Big Data & IA

1. Ingesta con AWS GLUE

Ejecución del **crawler**, para **esquematizar**



esquemaproyectoeventos

Crawler properties

Name esquemaproyectoeventos	IAM role LabRole	Database basededatosproyectoeventos	State READY
Description -	Security configuration -	Lake Formation configuration -	Table prefix -
Maximum table threshold -			

Advanced settings

Crawler runs (1)

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
June 4, 2025 at 09:47:32	June 4, 2025 at 09:48:46	01 min 14 s	Completed	-	1 table change, 0 partition changes

- Visualización de la base de datos generada con el crawler, a través de la base de datos creada

eventosproyectoinicio

Table details

Name eventosproyectoinicio	Classification CSV	Deprecated -
Database basededatosproyectoeventos	Location s3://eventosproyectoinicio/	Column statistics No statistics
Description -	Connection -	
Last updated June 4, 2025 at 09:48:46		

Schema (19)

#	Column name	Data type	Partition key	Comment
1	id_evento	bigint	-	-
2	tipo_evento	string	-	-
3	nombre	string	-	-

Proyecto Final. Especialización Big Data & IA

1. Ingesta con AWS GLUE visual ETL

Planteamos los siguientes cambios:



- **Renombrar** columnas.
- Crear nuevas **columnas calculadas**.
- **Conversión de tipos** → cambio fecha a date y puntuación media a double.
- Formato de fecha, para filtrar más adelante por año más fácilmente.

Utilizamos ahora el AWS GLUE visual **ETL**



- Volcamos el bucket almacenado en **S3**

The screenshot shows the AWS Glue Visual ETL console. The main workspace displays a 'Data source - S3 bucket' node connected to 'Amazon S3'. The 'Data preview' section shows a table with columns: id_evento, tipo_evento, nombre, fecha, artista_o_equipo, and duracion_min. The 'Data source properties - S3' panel on the right is open, showing the configuration for the S3 source. The 'S3 source type' is set to 'S3 location'. The 'S3 URL' is 's3://eventosproyectoinicio'. The 'Data format' is 'CSV' with a 'Comma (,)' delimiter. The 'Escape character' is 'Double quote (")'. The 'Quote character' is 'Double quote (")'. The 'First line of source file contains column headers' checkbox is checked.

- **Renombrar** columnas



The screenshot shows the AWS Glue Visual ETL console with a 'Transform - RenameField' node added to the workflow. The 'Data preview' section shows the same table as before. The 'Transform' panel on the right is open, showing the configuration for the 'RenameField' node. The 'Name' is 'cambiar nombre a coste total de salarios'. The 'Node parents' are 'Amazon S3 - DataSource'. The 'Rename field' section shows the 'Current field name' as 'coste_salarios' and the 'New field name' as 'coste_total_salarios'.

Proyecto Final. Especialización Big Data & IA

1. Ingesta con AWS GLUE visual ETL

Planteamos los siguientes cambios:



- **Renombrar** columnas.
- Crear nuevas **columnas calculadas**.
- **Conversión de tipos** → cambio fecha a date y puntuación media a double.
- Formato de fecha, para filtrar más adelante por año más fácilmente.

Utilizamos ahora el AWS GLUE visual **ETL**

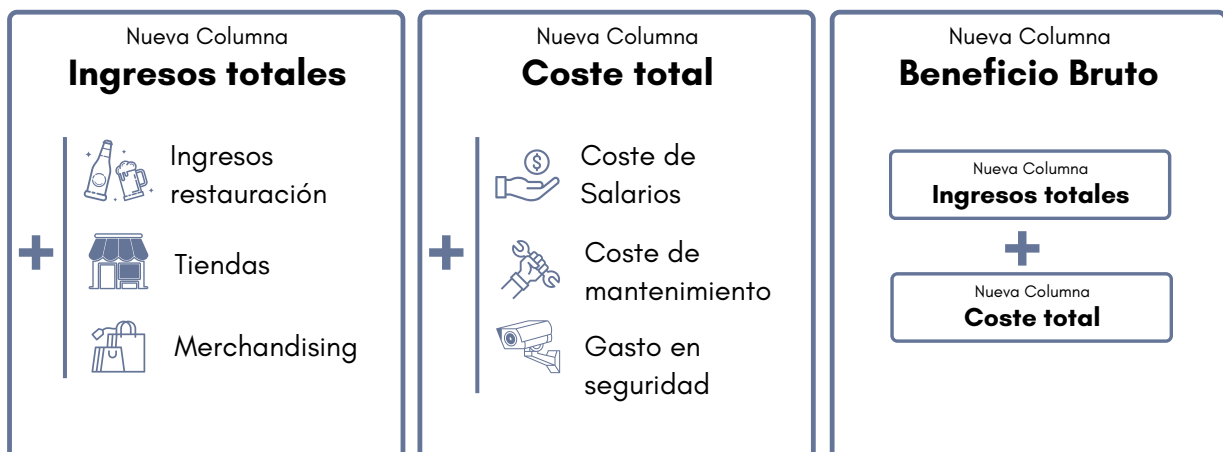


- Crear nuevas columnas calculadas, útiles para saber nuevos parámetros

Adjunto enlaces que he utilizado para poder realizar el código correctamente

- https://docs.aws.amazon.com/es_es/glue/latest/dg/edit-nodes-chapter.html
- <https://www.knowmadmood.com/es/blog/transformaciones-personalizadas-con-aws-glue>
- <https://www.youtube.com/watch?v=QsmNGmh-qQI>

Objetivo:



Proyecto Final. Especialización Big Data & IA

1. Ingesta con AWS GLUE visual ETL

Planteamos los siguientes cambios:



- **Renombrar** columnas.
- Crear nuevas **columnas calculadas**.
- **Conversión de tipos** → cambio fecha a date y puntuación media a double.
- Formato de fecha, para filtrar más adelante por año más fácilmente.

Utilizamos ahora el AWS GLUE visual **ETL**



- Conversión de tipos → cambio fecha a date y puntuación media a double

Source key	Target key	Data type	Drop
tipo_evento	tipo_evento	string	☐
nombre	nombre	string	☐
fecha	fecha	date	☐
artista_o_equipo	artista_o_equi	string	☐
duracion_min	duracion_min	string	☐
entradas_vendidas	entradas_venc	string	☐
entradas_utilizadas	entradas_utili	string	☐
bar	bar	string	☐
merchandising	merchandising	string	☐
tienda	tienda	string	☐
puntuacion_media	puntuacion_m	double	☐
cantidad_opiniones	cantidad_opin	string	☐
num_trabajadores	num_trabajad	string	☐
coste_mantenimiento	coste_manten	string	☐
gasto_seguridad	gasto_segurid	string	☐
procedencia_principal	procedencia_p	string	☐

- Implemento un último nodo de "select from collection" debido a que mi flujo es una colección y me da **error** a la hora de cargar las transformaciones en el bucket de destino.

id_evento	tipo_evento	nombre	fecha	artista_o_equipo	duracion_min

Proyecto Final. Especialización Big Data & IA

1. Ingesta con AWS GLUE visual ETL

Planteamos los siguientes cambios:



- **Renombrar** columnas.
- Crear nuevas **columnas calculadas**.
- **Conversión de tipos** → cambio fecha a date y puntuación media a double.
- Formato de fecha, para filtrar más adelante por año más fácilmente.

Utilizamos ahora el AWS GLUE visual **ETL**



- Vuelco las transformaciones utilizadas en el bucket de destino

The screenshot shows the AWS Glue console interface. The main workspace displays a visual ETL job named 'proyectoeventos'. The job consists of two nodes: 'Transform - SelectFromCollection' and 'Data target - S3 bucket Amazon S3'. The 'Data target properties - S3' panel is open on the right, showing the configuration for the 'Data target - S3 bucket Amazon S3' node. The configuration includes the following fields:

- Name:** Amazon S3
- Node parents:** Select From Collection
- Format:** CSV
- Compression Type:** None
- S3 Target Location:** s3://eventosproyectoeventos
- Data Catalog update options:** Do not update the Data Catalog
- File partitioning:** Choose if you want the output to be saved into autogenerated or in a predesignated number of partitions.

A warning message at the bottom of the console states: 'Target node not supported. You have selected a data target node which is not supported for data preview. Please select another type of node instead.'

- Ejecuto el primer "RUN"

The screenshot shows the AWS Glue console interface. The main workspace displays the 'Job runs (1/1)' for the 'proyectoeventos' job. The job status is 'Succeeded'. The 'Run details' panel is open at the bottom, showing the following information:

- Job name:** proyectoeventos
- Id:** j_844987a0e905ef978e2169d23cc4ecca4a29c761a98ec730e4f8294695
- Run status:** Succeeded
- Start time (Local):** 06/04/2025 12:54:15
- End time (Local):** 06/04/2025 12:55:33
- Duration:** 1 m 5 s
- Capacity (DPU):** 10 DPU
- Worker type:** G.1X
- Glue version:** 5.0
- Start-up time:** 13 seconds
- Max capacity:** 10 DPU
- Last modified on (Local):** 06/04/2025 12:55:33
- Log group name:** /aws-glue/jobs
- Number of workers:** 10

Proyecto Final. Especialización Big Data & IA

2. Almacenamiento con Athena

Utilizamos **Athena** para enriquecer el almacenamiento de S3

The screenshot shows the Amazon Athena console interface. On the left, the 'Datos' sidebar is visible, showing the 'Origen de datos' as 'AwsDataCatalog', 'Catálogo' as 'Ningún elemento', and 'Base de datos' as 'basedatosproyectoeventos'. The 'Tablas y vistas' section shows a list of tables: 'eventos_proyecto', 'eventosproyectoinicio', and 'nombre_tabla'. The main area displays the SQL editor for 'Consulta 1' with the following code:

```
1 CREATE EXTERNAL TABLE eventos_proyecto (
2   id_evento string,
3   tipo_evento string,
4   nombre string,
5   fecha date,
6   artista_o_equipo string,
7   duracion_min int,
8   entradas_vendidas int,
9   entradas_utilizadas int,
10  bar double,
11  merchandising double,
12  tienda double,
13  puntuacion_media double,
14  cantidad_opiniones int,
15  num_trabajadores int,
16 )
17 LOCATION 's3://basedatosproyectoeventos/eventos_proyecto/'
18 PARTITIONED BY (fecha date)
```

Below the editor, the 'Ejecutar de nuevo' button is highlighted. The 'Resultados de la consulta' tab shows the query status as 'Completado' with a message: 'La consulta se ha realizado correctamente.' The execution details show 'Tiempo en cola: 89 ms', 'Tiempo de ejecución: 399 ms', and 'Datos analizados: -'.

- Visualizo las tablas

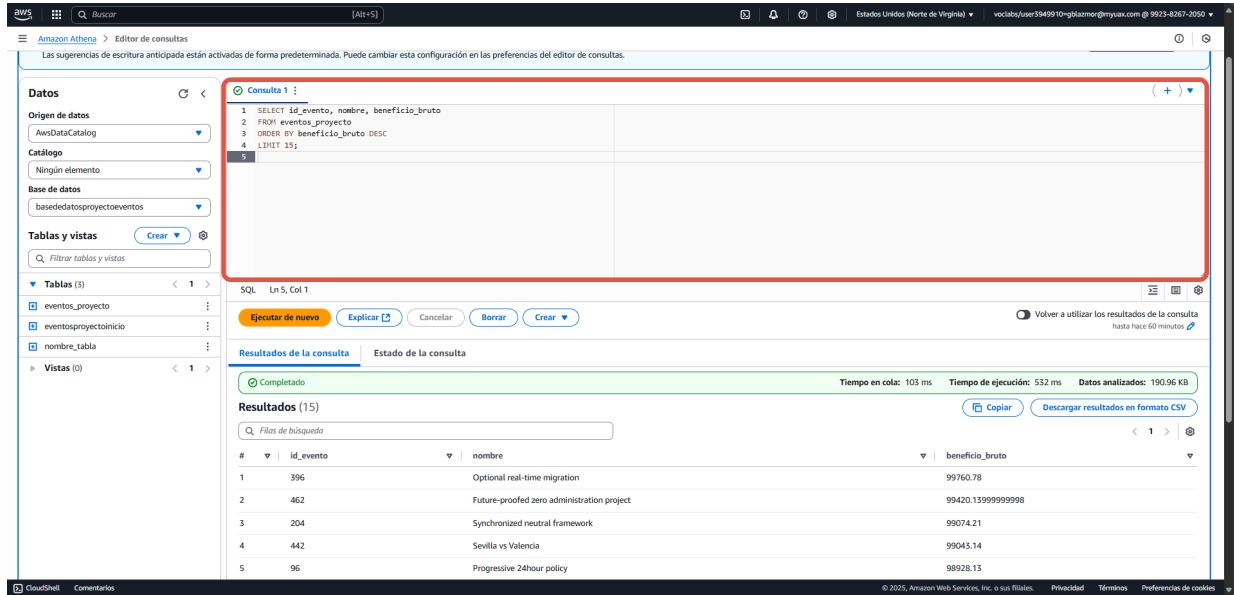
The screenshot shows the Amazon Athena console interface with the query results for the 'eventos_proyecto' table. The SQL editor shows the query: 'SELECT * FROM eventos_proyecto LIMIT 30;'. The 'Resultados de la consulta' tab shows the query status as 'Completado' with a message: 'La consulta se ha realizado correctamente.' The execution details show 'Tiempo en cola: 89 ms', 'Tiempo de ejecución: 399 ms', and 'Datos analizados: -'. The results are displayed in a table with 15 columns: 'id_evento', 'tipo_evento', 'nombre', 'fecha', 'artista_o_equipo', 'duracion_min', 'entradas_vendidas', 'entradas_utilizadas', 'bar', 'merchandising', 'tienda', 'puntuacion_media', 'cantidad_opiniones', 'num_trabajadores', and 'fecha_particion'. The table contains 30 rows of data, with the first row highlighted in red.

Proyecto Final. Especialización Big Data & IA

2. Almacenamiento con Athena

Utilizamos **Athena** para visualizar futuras vistas que haremos en powerBI

- Eventos que más beneficios han dado



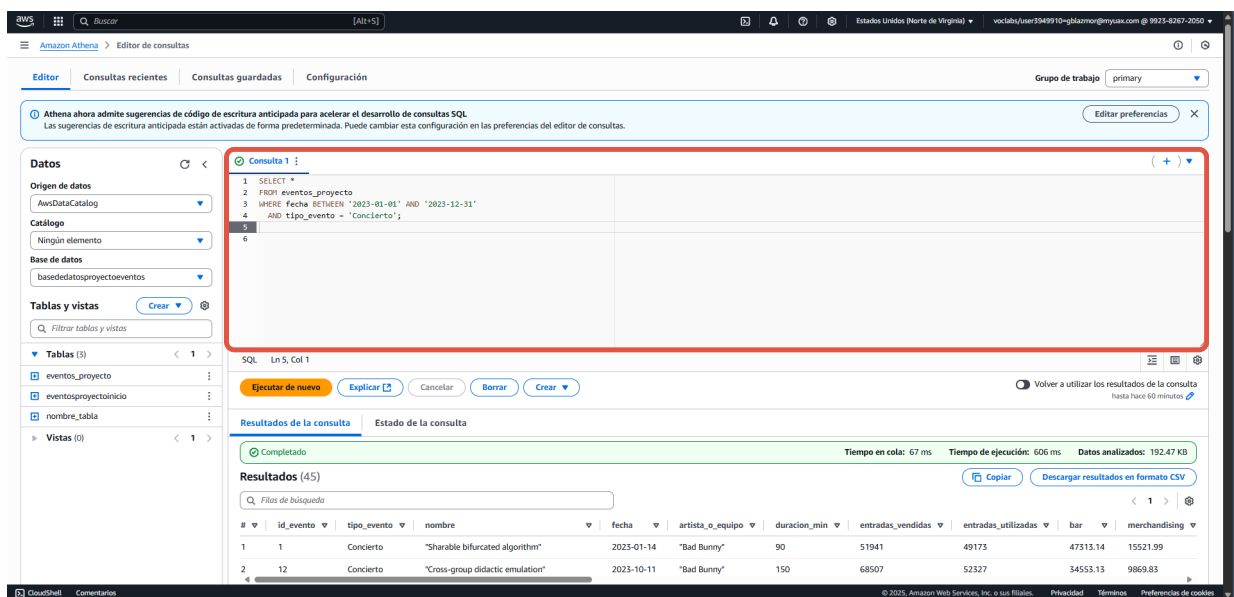
The screenshot shows the Amazon Athena console interface. On the left, the 'Datos' sidebar is visible with 'Origen de datos' set to 'AwsDataCatalog', 'Catálogo' set to 'Ningún elemento', and 'Base de datos' set to 'basedatosproyectoeventos'. The 'Tablas y vistas' section shows a list of tables including 'eventos_proyecto', 'eventosprojectoinicio', and 'nombre_tabla'. The main area displays a SQL query for 'Consulta 1':

```
1 SELECT id_evento, nombre, beneficio_bruto
2 FROM eventos_proyecto
3 ORDER BY beneficio_bruto DESC
4 LIMIT 15;
```

The query has been executed successfully. The 'Resultados de la consulta' section shows the following data:

#	id_evento	nombre	beneficio_bruto
1	396	Optional real-time migration	99760.78
2	462	Future-proofed zero administration project	99420.13999999998
3	204	Synchronized neutral framework	99074.21
4	442	Sevilla vs Valencia	99043.14
5	96	Progressive 24-hour policy	98928.13

- Eventos de tipo "concierto", por fechas, tipo de evento, artista, etc...



The screenshot shows the Amazon Athena console interface. On the left, the 'Datos' sidebar is visible with 'Origen de datos' set to 'AwsDataCatalog', 'Catálogo' set to 'Ningún elemento', and 'Base de datos' set to 'basedatosproyectoeventos'. The 'Tablas y vistas' section shows a list of tables including 'eventos_proyecto', 'eventosprojectoinicio', and 'nombre_tabla'. The main area displays a SQL query for 'Consulta 1':

```
1 SELECT *
2 FROM eventos_proyecto
3 WHERE fecha BETWEEN '2023-01-01' AND '2023-12-31'
4 AND tipo_evento = 'Concierto';
```

The query has been executed successfully. The 'Resultados de la consulta' section shows the following data:

#	id_evento	tipo_evento	nombre	fecha	artista_o_equipo	duracion_min	entradas_vendidas	entradas_utilizadas	bar	merchandising
1	1	Concierto	"Sharable bifurcated algorithm"	2023-01-14	"Bad Bunny"	90	51941	49173	47313.14	15521.99
2	12	Concierto	"Cross-group didactic emulation"	2023-10-11	"Bad Bunny"	150	68507	52327	34553.13	9869.83

Proyecto Final. Especialización Big Data & IA

2. Almacenamiento con Athena

Utilizamos **Athena** para visualizar futuras vistas que haremos en powerBI

- Suma, promedio, máximo, mínimo, etc.:

The screenshot shows the Amazon Athena console interface. On the left, there's a sidebar with 'Datos' (Data) and 'Tablas y vistas' (Tables and views). The main area displays a SQL query for 'Consulta 1' (Query 1) with the following code:

```
1 SELECT
2   tipo_evento,
3   COUNT(*) AS total_eventos,
4   AVG(CAST(beneficio_bruto AS double)) AS promedio_beneficio
5 FROM eventos_proyecto
6 WHERE beneficio_bruto <> '' AND beneficio_bruto IS NOT NULL
7 GROUP BY tipo_evento;
```

Below the query, the results are shown in a table format. The status is 'Completado' (Completed) with a time to complete of 103 ms and a time to execute of 697 ms. The results are as follows:

#	tipo_evento	total_eventos	promedio_beneficio
1	Partido de Fútbol	128	64648.45648437498
2	Festival	123	62576.776422764226
3	Concierto	158	60538.18620253161
4	Musical	133	61750.32984962405

- Ranking de los eventos más rentables de 2024:

The screenshot shows the Amazon Athena console interface. On the left, there's a sidebar with 'Datos' (Data) and 'Tablas y vistas' (Tables and views). The main area displays a SQL query for 'Consulta 1' (Query 1) with the following code:

```
1 SELECT *
2 FROM eventos_proyecto
3 WHERE fecha BETWEEN '2024-01-01' AND '2024-12-31'
```

Below the query, the results are shown in a table format. The status is 'Completado' (Completed) with a time to complete of 116 ms and a time to execute of 640 ms. The results are as follows:

#	id_evento	tipo_evento	nombre	fecha	artista_o_equipo	duracion_min	entradas_vendidas	entradas_utilizadas	bar
1	5	"Partido de Fútbol"	"Valencia vs Sevilla"	2024-01-03	"Valencia vs Sevilla"	90	33925	33811	40414.0
2	10	Musical	"Quality-focused background parallelism"	2024-02-19	Coldplay	150	46424	44020	22942.9

Proyecto Final. Especialización Big Data & IA

3. Procesamiento con Spark

Utilizamos **spark** para el procesamiento del dataset obtenido de aws glue visual etl

- o Inicio de spark en aws ERM, utilizando la tecnología Jupyter



```

[1]: from pyspark.sql import SparkSession
spark = (
    SparkSession.builder
    .appName("Python Spark SQL basic example")
    .config("spark.some.config.option", "some-value")
    .getOrCreate()
)

Last executed at 2025-06-05 00:17:05 in 12ms

[2]: df = spark.read.option("header", True).option("inferSchema", True).csv("s3://eventosproyectofinal/run-1749034515535-part-r-00000")

Last executed at 2025-06-05 00:15:48 in 21.58s

[4]: df.show()

Last executed at 2025-06-05 00:17:30 in 15.42s
    
```

[id_evento]	tipo_evento	nombre	fecha	artista_o_equipo	duracion_min	entradas_vendidas	entradas_utilizadas	bar	merchandising	tienda	puntuacion_medio	beneficio_bruto
1	Concierto	Shirley D'Arcy...	2023-01-14	Bad Bunny	90	51941	49173	47313.14	15521.99	9875.57	4.0	
80		8515.16	2004.93	Toledo								
2	Partido de Fútbol	Real Madrid vs Va...	2023-03-14	Real Madrid vs Va...	90	10285	9160	25092.04	34215.78	30548.47	2.0	
52		12252.51	1240.89	Zaragoza								
3	Festival	User-centric even...	2023-01-18	Coldplay	90	14739	11355	20189.6	21695.52	22489.06	3.23	
87		7510.90	1577.52	Madrid								
4	Partido de Fútbol	Valencia vs Valencia	2023-05-01	Valencia vs Valencia	90	60287	59392	43495.26	35855.61	9470.96	2.38	
75		5340.42	1077.01	Sevilla								
5	Partido de Fútbol	Valencia vs Sevilla	2024-01-03	Valencia vs Sevilla	90	33925	33831	140454.01	42341.8	14234.25	2.85	
69		11193.97	1184.6	Sevilla								
6	Festival	Face-to-face cle...	2023-10-30	Bad Bunny	90	24871	18229	43990.15	21120.9	10931.15	2.62	
72		5568.82	1064.25	Madrid								

- o Cálculo del beneficio por tipo de evento y muestro el esquema del dataset, para revisar si hay tipos de datos equivocados

```

[115]: from pyspark.sql.functions import year, to_date
df = df.withColumn("año", year(to_date(col("fecha"), "yyyy-MM-dd")))

Last executed at 2025-06-05 10:26:51 in 133ms

[117]: # Cálculo el beneficio medio por tipo de evento y año, para luego hacer un análisis con power BI de su evolución
df_beneficio = df.groupBy("tipo_evento", "año").agg(
    round(avg("beneficio_bruto"), 2).alias("beneficio_medio"),
    round(sum("entradas_utilizadas"), 0).alias("total_asistentes"),
    count("id_evento").alias("n_eventos")
)

df.printSchema()
df.show()

Last executed at 2025-06-05 10:27:22 in 13.34s
    
```

```

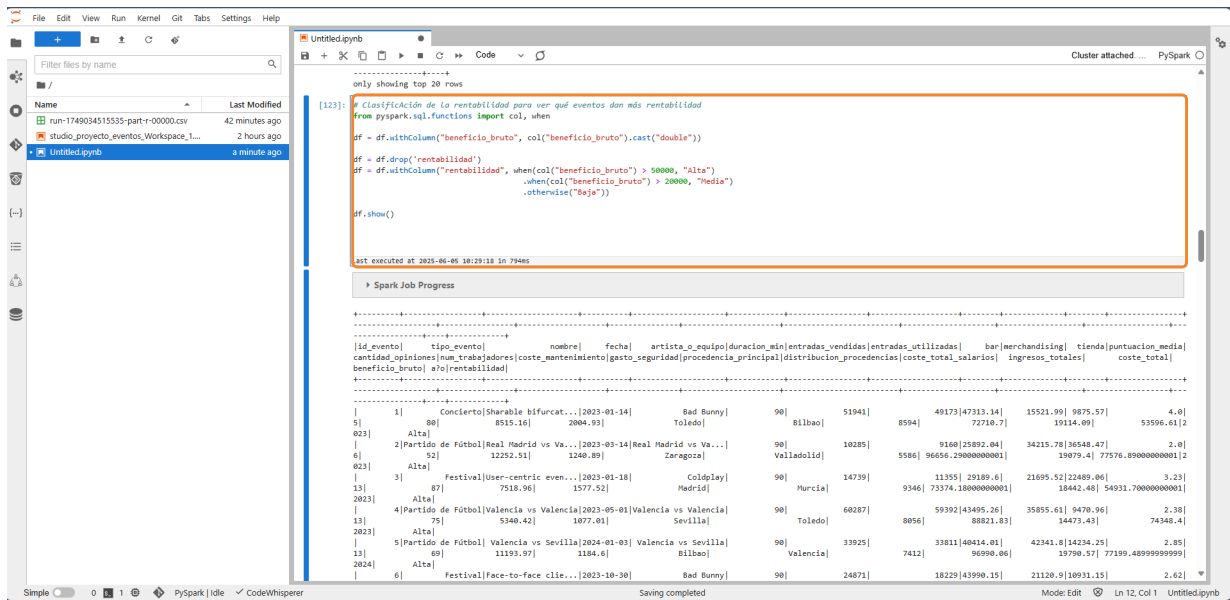
root
 |-- id_evento: integer (nullable = true)
 |-- tipo_evento: string (nullable = true)
 |-- nombre: string (nullable = true)
 |-- fecha: date (nullable = true)
 |-- artista_o_equipo: string (nullable = true)
 |-- duracion_min: integer (nullable = true)
 |-- entradas_vendidas: integer (nullable = true)
 |-- entradas_utilizadas: integer (nullable = true)
 |-- bar: double (nullable = true)
 |-- merchandising: double (nullable = true)
 |-- tienda: double (nullable = true)
 |-- puntuacion_medio: double (nullable = true)
 |-- cantidad_opiniones: integer (nullable = true)
 |-- num_trabajadores: integer (nullable = true)
 |-- coste_mantenimiento: double (nullable = true)
 |-- gasto_seguridad: double (nullable = true)
 |-- procedencia_principal: string (nullable = true)
 |-- distribucion_procedencias: string (nullable = true)
 |-- coste_total_salarios: integer (nullable = true)
    
```

Proyecto Final. Especialización Big Data & IA

3. Procesamiento con Spark

Utilizamos **spark** para el procesamiento del dataset obtenido de aws glue visual etl

- Clasifico los eventos según su rentabilidad



```
[123]: # Clasificación de la rentabilidad para ver qué eventos dan más rentabilidad
from pyspark.sql.functions import col, when

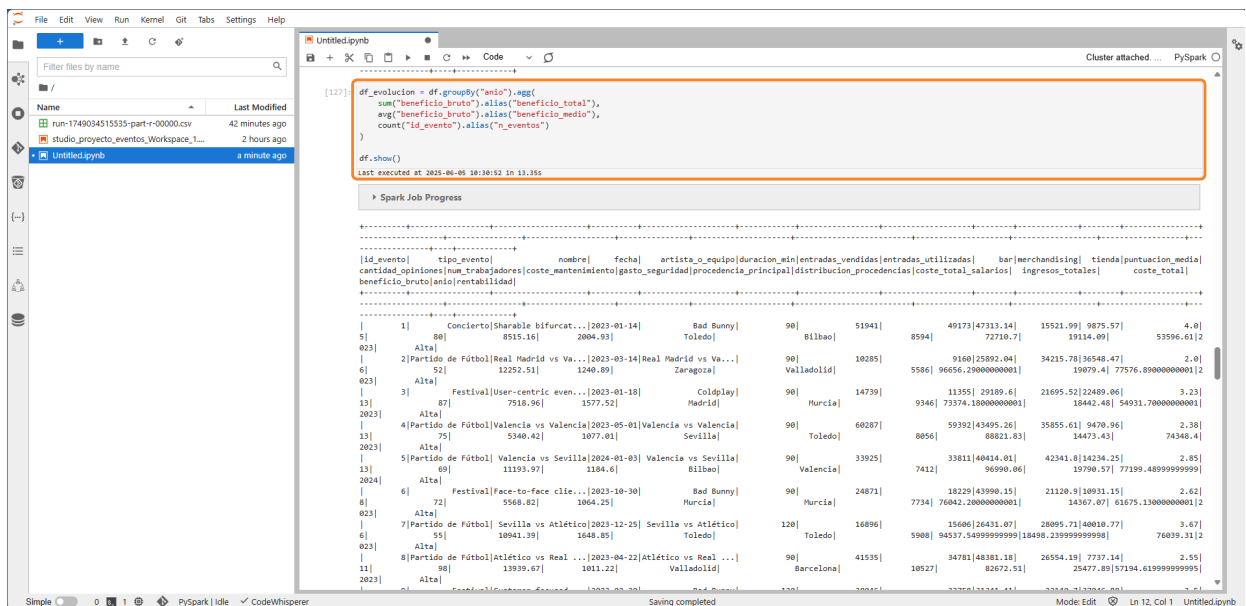
df = df.withColumn("beneficio_bruto", col("beneficio_bruto").cast("double"))
df = df.drop("rentabilidad")
df = df.withColumn("rentabilidad", when(col("beneficio_bruto") > 50000, "Alta")
                                   .when(col("beneficio_bruto") > 20000, "Media")
                                   .otherwise("Baja"))

df.show()
```

last executed at 2025-06-05 18:29:18 in 79ms

[id_evento]	[tipo_evento]	[nombre]	[fecha]	[artista_o_equipo]	[duracion_min]	[entradas_vendidas]	[entradas_utilizadas]	[bar]	[merchandising]	[tienda]	[puntuacion_media]	[cantidad_opiniones]	[num_trabajadores]	[coste_mantenimiento]	[gasto_seguridad]	[procedencia_principal]	[distribucion_procedencias]	[coste_total_salarios]	[ingresos_totales]	[coste_total]	[beneficio_bruto]	[rentabilidad]
1	Concierto	Sharable bifurcat...	2023-01-14	Bad Bunny	90	51941	49173	47313.14	15521.99	9875.57	4.0	53596.61	2	8515.16	2004.93	Toledo	Bilbao	8594	72710.7	19114.09	53596.61	Alta
2	Partido de Fútbol	Real Madrid vs Va...	2023-03-14	Real Madrid vs Va...	90	10285	9160	25892.04	34215.78	36548.47	2.0	19079.4	77576.89000000001	12252.51	1240.89	Zaragoza	Valladolid	5586	96656.29000000001	19079.4	77576.89000000001	Media
3	Festival	User-centric even...	2023-01-18	Coldplay	90	14739	11355	29189.6	21695.52	22489.06	3.23	18442.48	54931.70000000001	7518.96	1577.52	Madrid	Murcia	9346	73374.10000000001	18442.48	54931.70000000001	Alta
4	Partido de Fútbol	Valencia vs Valencia	2023-05-01	Valencia vs Valencia	90	60287	59392	43495.26	35855.61	9470.96	2.38	14473.43	74348.4	5340.42	1077.01	Sevilla	Toledo	8056	88821.83	14473.43	74348.4	Alta
5	Partido de Fútbol	Valencia vs Sevilla	2024-01-03	Valencia vs Sevilla	90	33925	33811	40414.01	42341.8	14234.25	2.85	19790.57	77199.48999999999	11193.97	1184.6	Bilbao	Valencia	7412	96990.06	19790.57	77199.48999999999	Alta
6	Festival	Face-to-face cle...	2023-10-30	Bad Bunny	90	24871	18229	43990.15	21120.9	10931.15	2.62	14367.07	61675.13000000001	5560.82	1064.25	Murcia	Murcia	7734	76042.20000000001	14367.07	61675.13000000001	Alta

- Evolución del beneficio bruto por año



```
[127]: df_evolucion = df.groupBy("anio").agg(
    sum("beneficio_bruto").alias("beneficio_total"),
    avg("beneficio_bruto").alias("beneficio_medio"),
    count("id_evento").alias("n_eventos")
)

df_evolucion.show()
```

last executed at 2025-06-05 18:30:52 in 13.35s

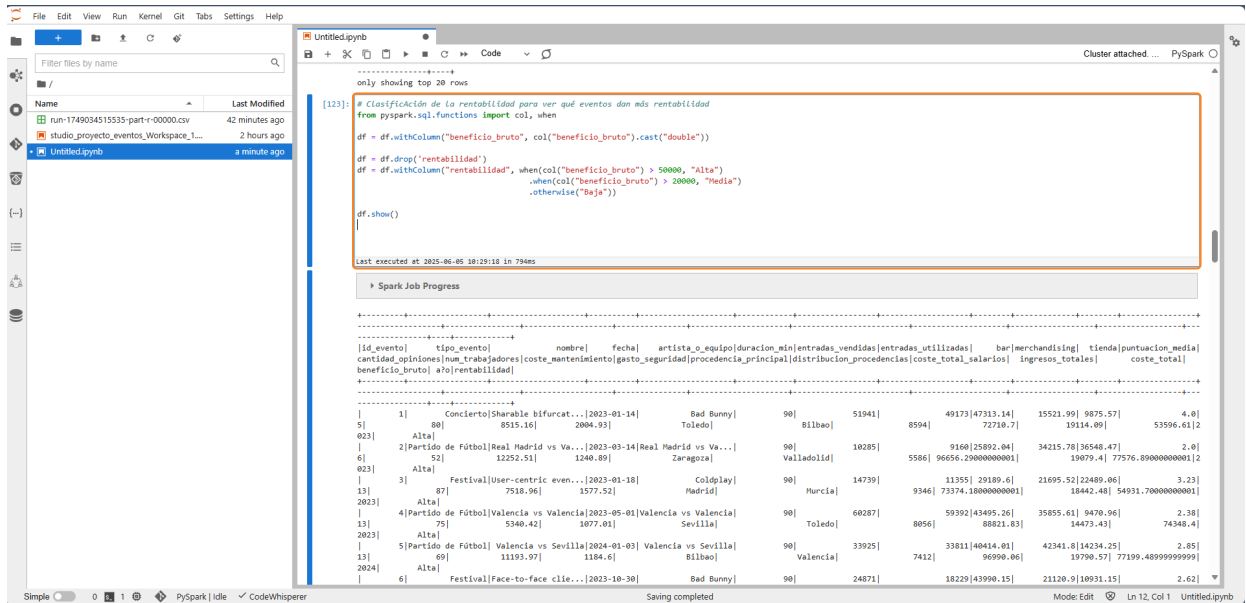
[id_evento]	[tipo_evento]	[nombre]	[fecha]	[artista_o_equipo]	[duracion_min]	[entradas_vendidas]	[entradas_utilizadas]	[bar]	[merchandising]	[tienda]	[puntuacion_media]	[cantidad_opiniones]	[num_trabajadores]	[coste_mantenimiento]	[gasto_seguridad]	[procedencia_principal]	[distribucion_procedencias]	[coste_total_salarios]	[ingresos_totales]	[coste_total]	[beneficio_bruto[anio]]	
1	Concierto	Sharable bifurcat...	2023-01-14	Bad Bunny	90	51941	49173	47313.14	15521.99	9875.57	4.0	53596.61	2	8515.16	2004.93	Toledo	Bilbao	8594	72710.7	19114.09	53596.61	2023
2	Partido de Fútbol	Real Madrid vs Va...	2023-03-14	Real Madrid vs Va...	90	10285	9160	25892.04	34215.78	36548.47	2.0	19079.4	77576.89000000001	12252.51	1240.89	Zaragoza	Valladolid	5586	96656.29000000001	19079.4	77576.89000000001	2023
3	Festival	User-centric even...	2023-01-18	Coldplay	90	14739	11355	29189.6	21695.52	22489.06	3.23	18442.48	54931.70000000001	7518.96	1577.52	Madrid	Murcia	9346	73374.10000000001	18442.48	54931.70000000001	2023
4	Partido de Fútbol	Valencia vs Valencia	2023-05-01	Valencia vs Valencia	90	60287	59392	43495.26	35855.61	9470.96	2.38	14473.43	74348.4	5340.42	1077.01	Sevilla	Toledo	8056	88821.83	14473.43	74348.4	2023
5	Partido de Fútbol	Valencia vs Sevilla	2024-01-03	Valencia vs Sevilla	90	33925	33811	40414.01	42341.8	14234.25	2.85	19790.57	77199.48999999999	11193.97	1184.6	Bilbao	Valencia	7412	96990.06	19790.57	77199.48999999999	2024
6	Festival	Face-to-face cle...	2023-10-30	Bad Bunny	90	24871	18229	43990.15	21120.9	10931.15	2.62	14367.07	61675.13000000001	5560.82	1064.25	Murcia	Murcia	7734	76042.20000000001	14367.07	61675.13000000001	2023
7	Partido de Fútbol	Sevilla vs Atlético	2023-12-25	Sevilla vs Atlético	120	16896	15606	26431.07	28095.71	40010.77	3.67	76039.31	2	18941.39	1648.05	Toledo	Toledo	5908	94537.54999999999	18498.239999999998	76039.31	2023
8	Partido de Fútbol	Atlético vs Real ...	2023-04-22	Atlético vs Real ...	90	41535	34781	48381.18	26554.19	7737.14	2.55	25477.89	57194.619999999995	13939.67	1011.22	Valladolid	Barcelona	10527	82672.51	25477.89	57194.619999999995	2023

Proyecto Final. Especialización Big Data & IA

3. Procesamiento con

Utilizamos **spark** para el procesamiento del dataset obtenido de aws glue visual etl

- Clasifico los eventos según su rentabilidad



```
[123]: # Clasificación de la rentabilidad para ver qué eventos dan más rentabilidad
from pyspark.sql.functions import col, when

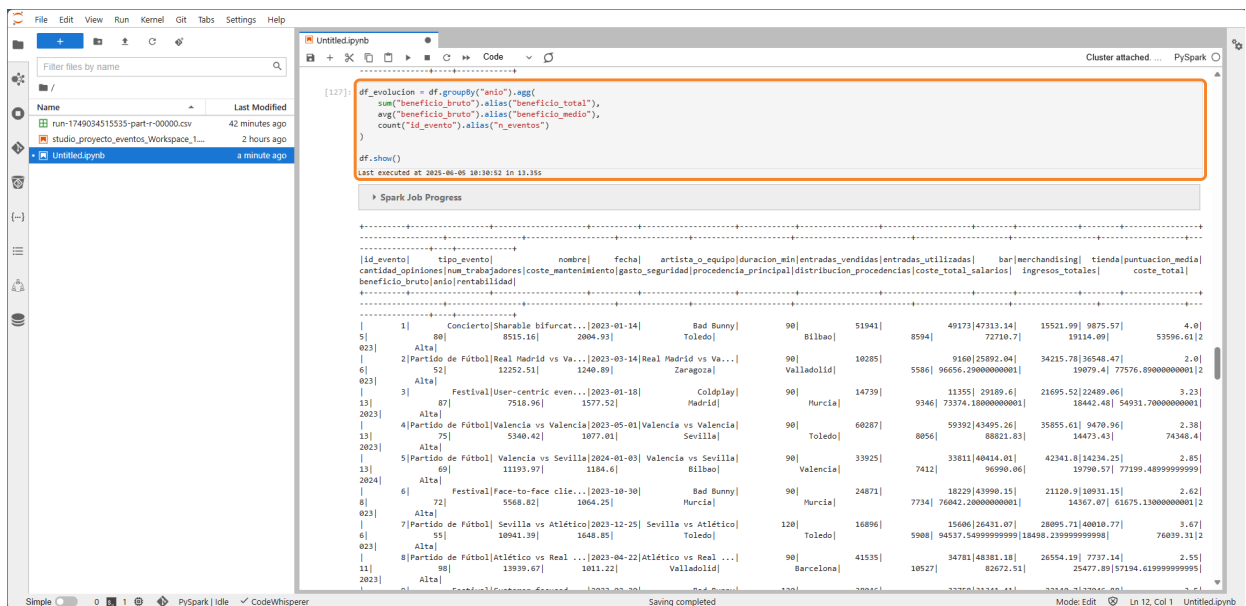
df = df.withColumn("beneficio_bruto", col("beneficio_bruto").cast("double"))
df = df.drop("rentabilidad")
df = df.withColumn("rentabilidad", when(col("beneficio_bruto") > 50000, "Alta")
    .when(col("beneficio_bruto") > 20000, "Media")
    .otherwise("Baja"))

df.show()
```

Last executed at 2025-06-05 18:29:18 in 794ms

[id_evento]	tipo_evento]	nombre]	fecha]	artista_o_equipo]	duracion_min]	entradas_vendidas]	entradas_utilizadas]	bar]	merchandising]	tienda]	puntuacion_media]	coste_total]	
1	Alta]	Concierto Sharable bifurcat...	2023-01-14]	Bad Bunny]	90]	51941]	49173 47313.14]	15521.99]	9875.57]	4.0]	53596.61]		
5	Alta]	8515.16]	2004.93]	Toledo]	Bilbao]	8594]	72710.7]	19114.09]	53596.61]	2.0]	53596.61]		
023]	Alta]	2 Partido de Fútbol Real Madrid vs Va...	2023-03-14]	Real Madrid vs Va...	90]	10285]	9160 25892.04]	34215.78]	36548.47]	2.0]	53596.61]		
6]	Alta]	52]	12252.51]	1240.89]	Zaragoza]	Valladolid]	5586]	96656.29000000001]	19079.4]	77576.89000000001]	2.0]	53596.61]	
023]	Alta]	3]	7518.96]	1577.52]	Madrid]	Murcia]	14739]	9346]	73374.18000000001]	18442.48]	54931.70000000001]	2.0]	53596.61]
13]	Alta]	87]	7518.96]	1577.52]	Madrid]	Murcia]	14739]	9346]	73374.18000000001]	18442.48]	54931.70000000001]	2.0]	53596.61]
2023]	Alta]	4 Partido de Fútbol Valencia vs Valencia]	2023-05-01]	Valencia vs Valencia]	90]	60287]	59392 43495.26]	35855.61]	9470.96]	2.38]	74348.4]		
13]	Alta]	75]	5340.42]	1077.01]	Sevilla]	Toledo]	8056]	88821.83]	14473.43]	74348.4]	2.38]	74348.4]	
2023]	Alta]	5 Partido de Fútbol Valencia vs Sevilla]	2024-01-03]	Valencia vs Sevilla]	90]	33925]	33811 40414.01]	42341.8]	14234.25]	2.85]	74348.4]		
13]	Alta]	69]	11193.97]	1184.6]	Bilbao]	Valencia]	7412]	96990.06]	19790.57]	77199.48999999999]	2.85]	74348.4]	
2024]	Alta]	6]	5568.82]	1064.25]	Bad Bunny]	90]	24871]	18229 43990.15]	21120.9]	10931.15]	2.62]	74348.4]	
023]	Alta]	72]	5568.82]	1064.25]	Murcia]	Murcia]	7734]	76842.20000000001]	14367.07]	61675.13000000001]	2.62]	74348.4]	
023]	Alta]	7 Partido de Fútbol Sevilla vs Atlético]	2023-12-25]	Sevilla vs Atlético]	120]	16896]	15606 26431.07]	28095.71]	40010.77]	3.67]	74348.4]		
6]	Alta]	55]	10941.39]	1648.05]	Toledo]	Toledo]	5908]	94537.54999999999]	18498.23999999999]	76039.31]	2.62]	74348.4]	
023]	Alta]	8 Partido de Fútbol Atlético vs Real ...]	2023-04-22]	Atlético vs Real ...]	90]	41535]	34781 48381.18]	26554.19]	7737.14]	2.55]	74348.4]		
11]	Alta]	98]	13939.67]	1011.22]	Valladolid]	Barcelona]	10527]	82672.51]	25477.89]	57194.61999999999]	2.55]	74348.4]	
2023]	Alta]	8]	13939.67]	1011.22]	Valladolid]	Barcelona]	10527]	82672.51]	25477.89]	57194.61999999999]	2.55]	74348.4]	

- Evolución del beneficio bruto por año



```
[127]: df_evolucion = df.groupBy("anio").agg(
    sum("beneficio_bruto").alias("beneficio_total"),
    avg("beneficio_bruto").alias("beneficio_medio"),
    count("id_evento").alias("n_eventos")
)

df_evolucion.show()
```

Last executed at 2025-06-05 18:30:52 in 13.35s

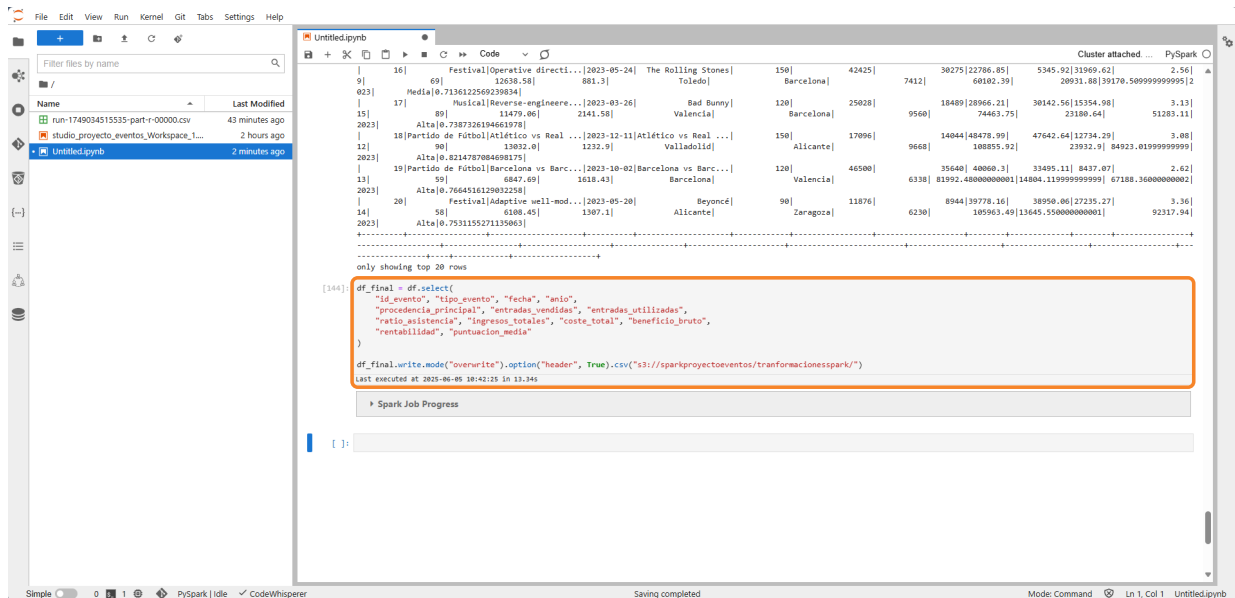
[id_evento]	tipo_evento]	nombre]	fecha]	artista_o_equipo]	duracion_min]	entradas_vendidas]	entradas_utilizadas]	bar]	merchandising]	tienda]	puntuacion_media]	coste_total]	
1	Alta]	Concierto Sharable bifurcat...	2023-01-14]	Bad Bunny]	90]	51941]	49173 47313.14]	15521.99]	9875.57]	4.0]	53596.61]		
5	Alta]	8515.16]	2004.93]	Toledo]	Bilbao]	8594]	72710.7]	19114.09]	53596.61]	2.0]	53596.61]		
023]	Alta]	2 Partido de Fútbol Real Madrid vs Va...	2023-03-14]	Real Madrid vs Va...	90]	10285]	9160 25892.04]	34215.78]	36548.47]	2.0]	53596.61]		
6]	Alta]	52]	12252.51]	1240.89]	Zaragoza]	Valladolid]	5586]	96656.29000000001]	19079.4]	77576.89000000001]	2.0]	53596.61]	
023]	Alta]	3]	7518.96]	1577.52]	Madrid]	Murcia]	14739]	9346]	73374.18000000001]	18442.48]	54931.70000000001]	2.0]	53596.61]
13]	Alta]	87]	7518.96]	1577.52]	Madrid]	Murcia]	14739]	9346]	73374.18000000001]	18442.48]	54931.70000000001]	2.0]	53596.61]
2023]	Alta]	4 Partido de Fútbol Valencia vs Valencia]	2023-05-01]	Valencia vs Valencia]	90]	60287]	59392 43495.26]	35855.61]	9470.96]	2.38]	74348.4]		
13]	Alta]	75]	5340.42]	1077.01]	Sevilla]	Toledo]	8056]	88821.83]	14473.43]	74348.4]	2.38]	74348.4]	
2023]	Alta]	5 Partido de Fútbol Valencia vs Sevilla]	2024-01-03]	Valencia vs Sevilla]	90]	33925]	33811 40414.01]	42341.8]	14234.25]	2.85]	74348.4]		
13]	Alta]	69]	11193.97]	1184.6]	Bilbao]	Valencia]	7412]	96990.06]	19790.57]	77199.48999999999]	2.85]	74348.4]	
2024]	Alta]	6]	5568.82]	1064.25]	Bad Bunny]	90]	24871]	18229 43990.15]	21120.9]	10931.15]	2.62]	74348.4]	
023]	Alta]	72]	5568.82]	1064.25]	Murcia]	Murcia]	7734]	76842.20000000001]	14367.07]	61675.13000000001]	2.62]	74348.4]	
023]	Alta]	7 Partido de Fútbol Sevilla vs Atlético]	2023-12-25]	Sevilla vs Atlético]	120]	16896]	15606 26431.07]	28095.71]	40010.77]	3.67]	74348.4]		
6]	Alta]	55]	10941.39]	1648.05]	Toledo]	Toledo]	5908]	94537.54999999999]	18498.23999999999]	76039.31]	2.62]	74348.4]	
023]	Alta]	8 Partido de Fútbol Atlético vs Real ...]	2023-04-22]	Atlético vs Real ...]	90]	41535]	34781 48381.18]	26554.19]	7737.14]	2.55]	74348.4]		
11]	Alta]	98]	13939.67]	1011.22]	Valladolid]	Barcelona]	10527]	82672.51]	25477.89]	57194.61999999999]	2.55]	74348.4]	
2023]	Alta]	8]	13939.67]	1011.22]	Valladolid]	Barcelona]	10527]	82672.51]	25477.89]	57194.61999999999]	2.55]	74348.4]	

Proyecto Final. Especialización Big Data & IA

3. Procesamiento con Spark

Utilizamos **spark** para el procesamiento del dataset obtenido de aws glue visual etl

- Guardo todos los cambios y selecciono los campos que me interesan para usarlos en powerBI

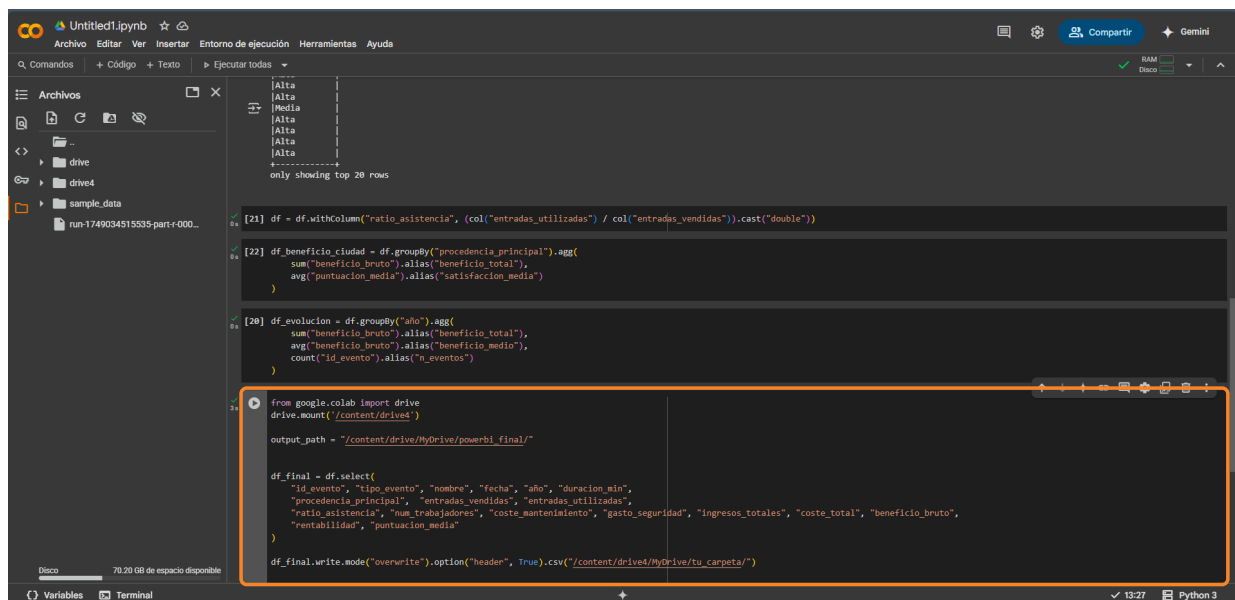


```

[14]: df_final = df.select(
    "id_evento", "tipo_evento", "fecha", "año",
    "procedencia_principal", "entradas_vendidas", "entradas_utilizadas",
    "ratio_asistencia", "ingresos_totales", "coste_total", "beneficio_bruto",
    "rentabilidad", "puntuacion_media"
)

df_final.write.mode("overwrite").option("header", True).csv("s3://sparkproyectoeventos/transformacionesspark/")
  
```

- Tuve que usar la versión de Jupyter de drive para rescatar más columnas y tener más datos con los que trabajar en Power BI



```

from google.colab import drive
drive.mount('/content/drive')

output_path = "/content/drive/MyDrive/powerbi_final/"

df_final = df.select(
    "id_evento", "tipo_evento", "nombre", "fecha", "año", "duracion_min",
    "procedencia_principal", "entradas_vendidas", "entradas_utilizadas",
    "ratio_asistencia", "num_trabajadores", "coste_mantenimiento", "gasto_seguridad", "ingresos_totales", "coste_total", "beneficio_bruto",
    "rentabilidad", "puntuacion_media"
)

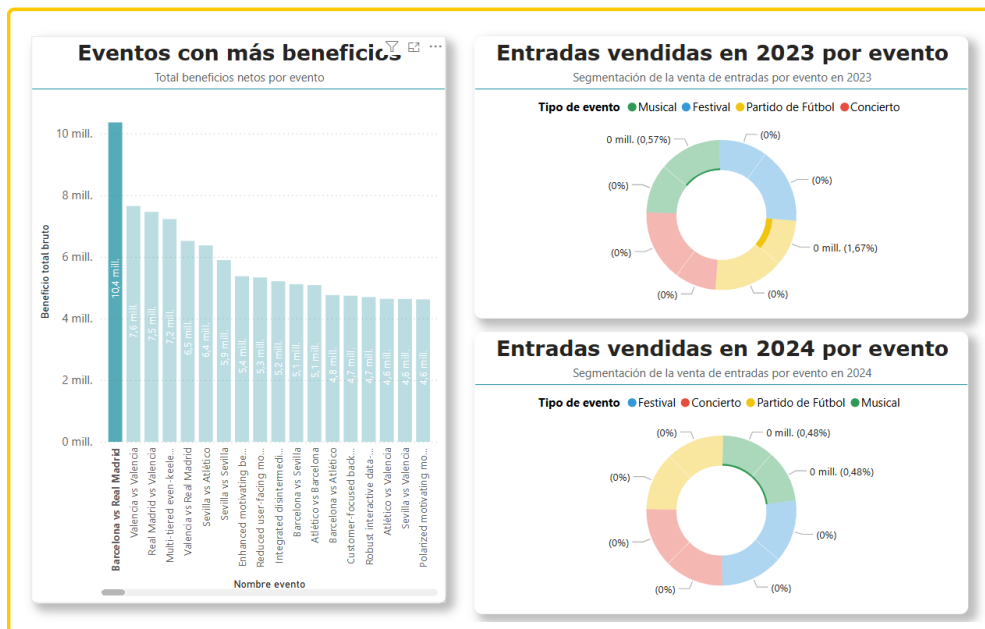
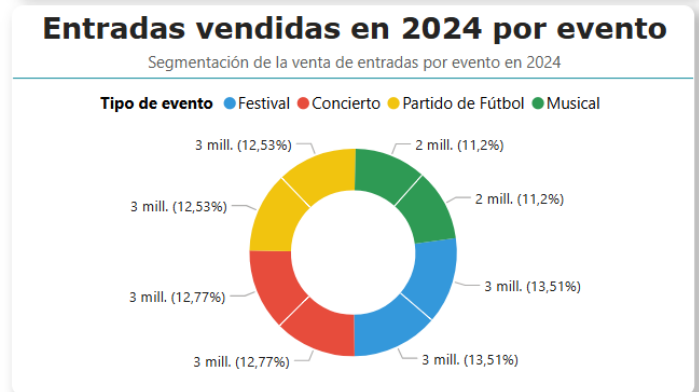
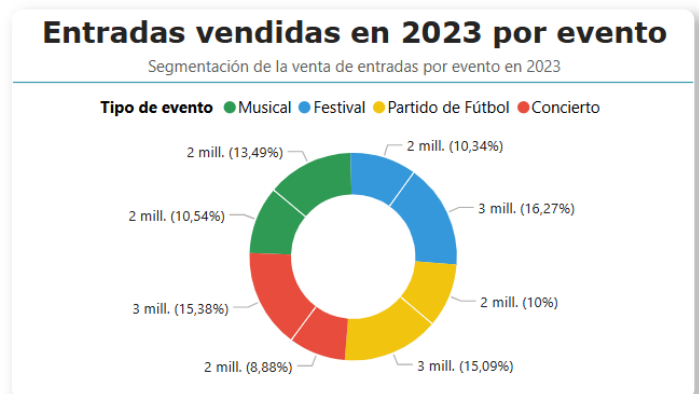
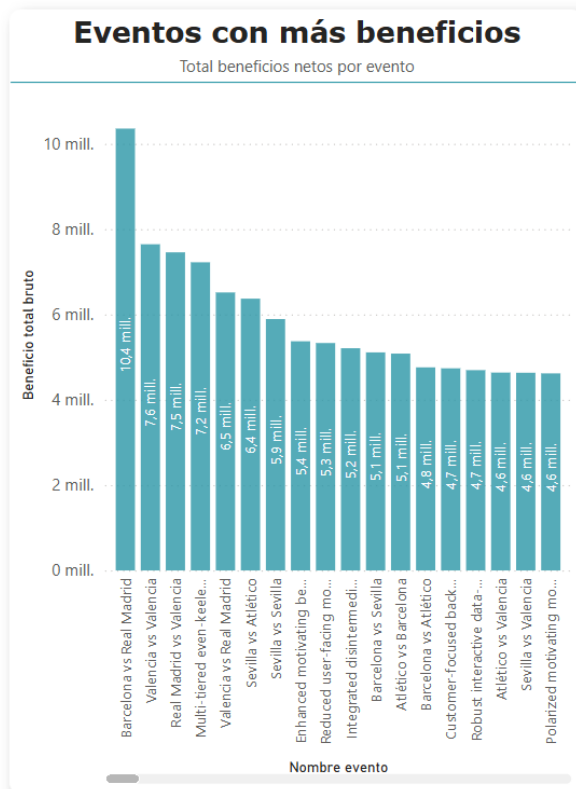
df_final.write.mode("overwrite").option("header", True).csv("/content/drive/MyDrive/tu_carpetita/")
  
```

Proyecto Final. Especialización Big Data & IA

4. Análisis con powerBI

Utilizamos **powerBI** para hacer el **análisis** y poder observar, qué eventos han dado mejor rendimiento, qué ciudades han comprado más entradas, desglose de gastos...

- En este panel, podemos observar los eventos que han dado más beneficios y un desglose de entradas vendidas según el tipo de evento.

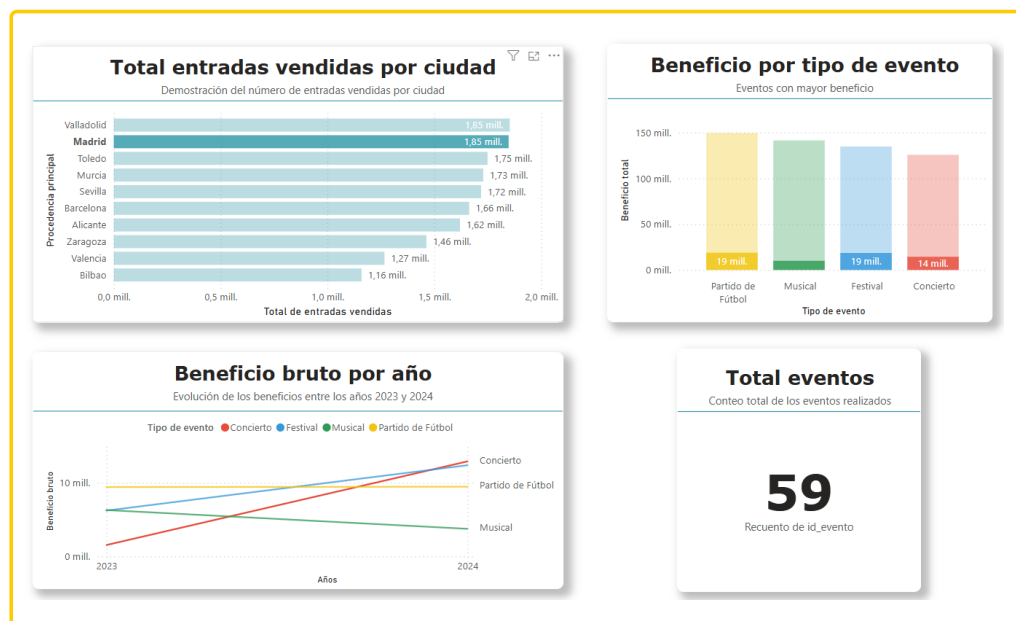
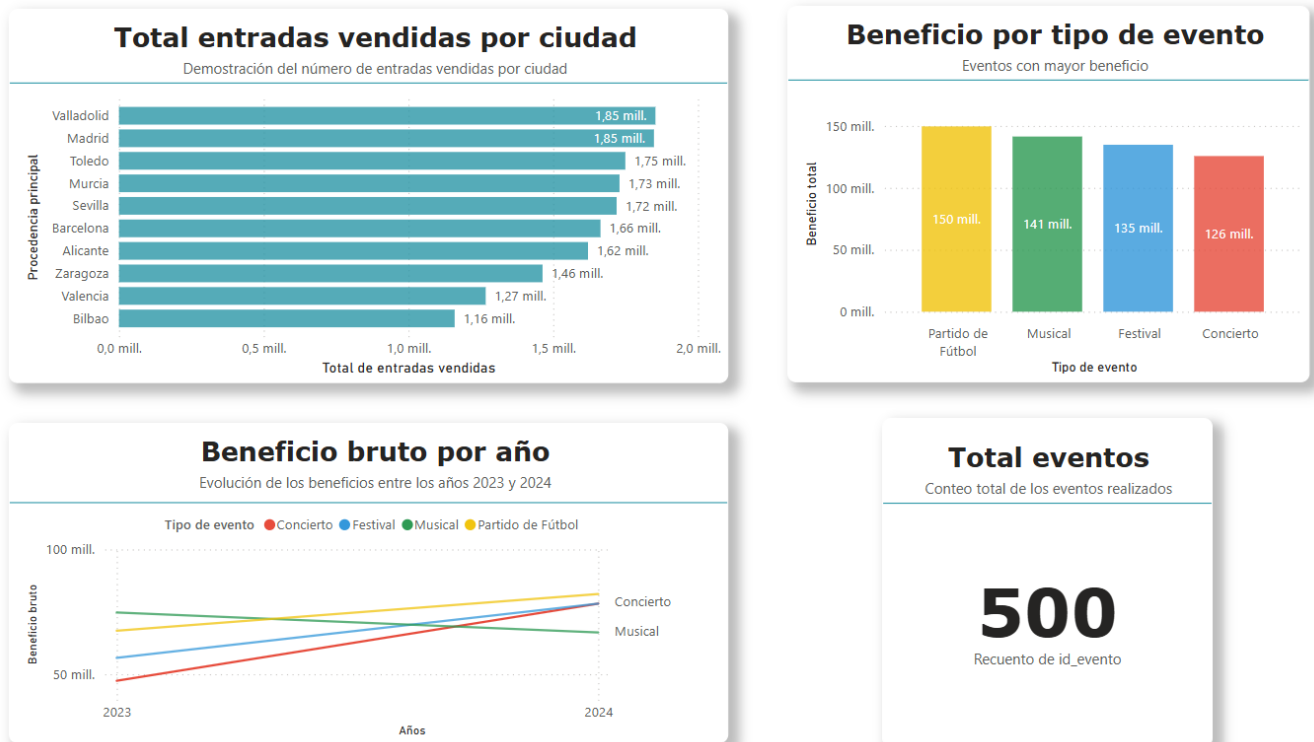


Proyecto Final. Especialización Big Data & IA

4. Análisis con powerBI

Utilizamos **powerBI** para hacer el **análisis** y poder observar, qué eventos han dado mejor rendimiento, qué ciudades han comprado más entradas, desglose de gastos...

- En este panel, podemos observar en qué ciudades se han vendido más entradas, el beneficio según el tipo de evento, la evolución del beneficio bruto por año y el total de eventos.

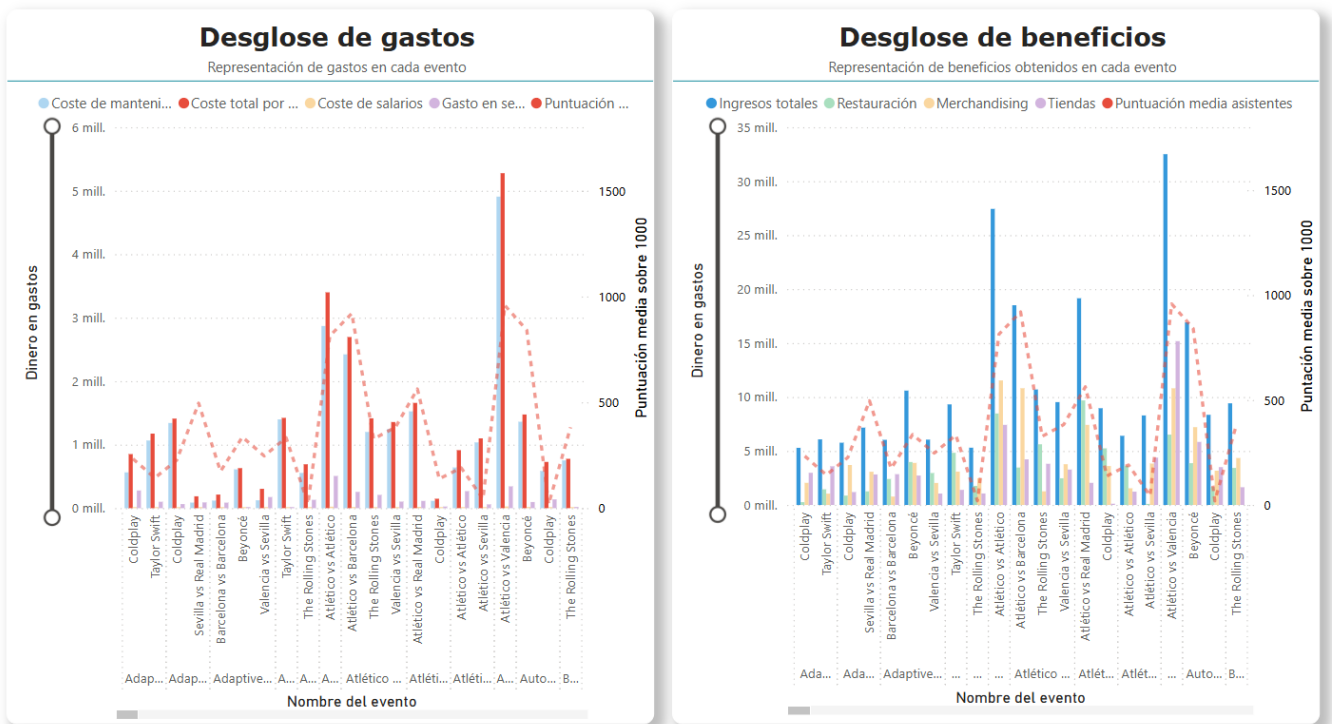


Proyecto Final. Especialización Big Data & IA

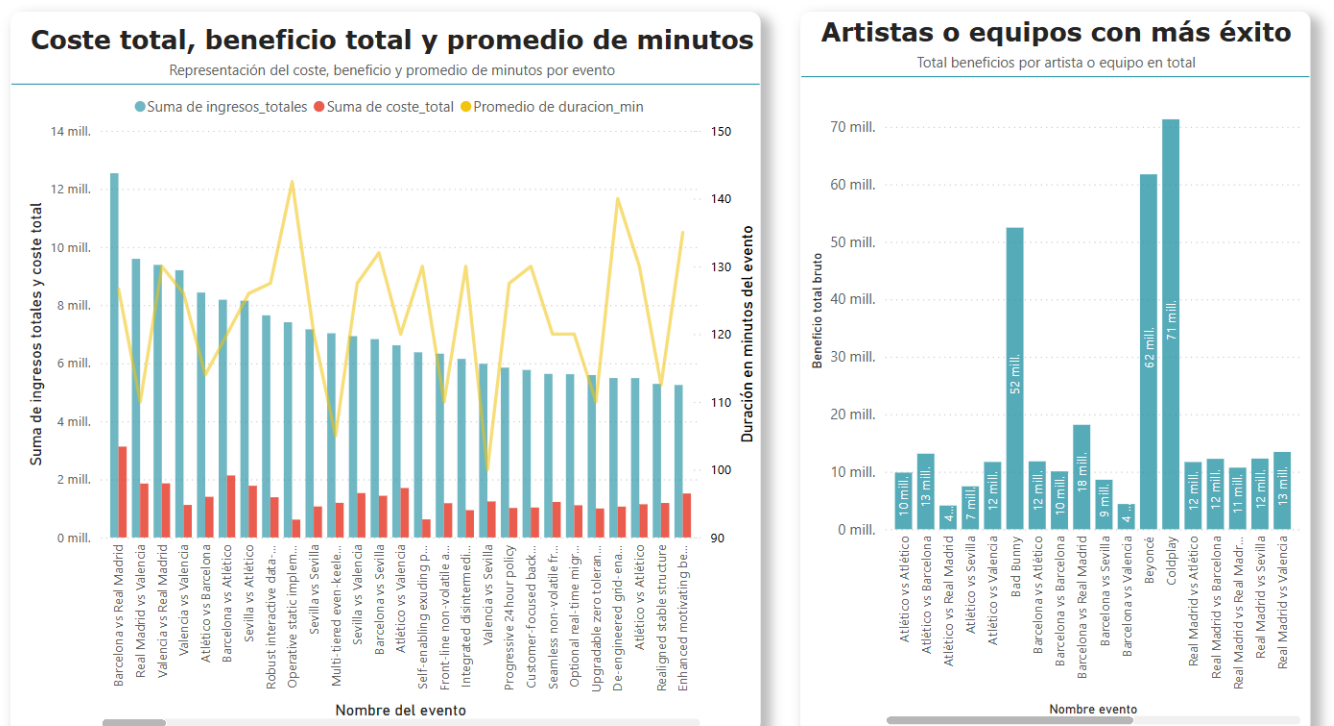
4. Análisis con powerBI

Utilizamos **powerBI** para hacer el **análisis** y poder observar, qué eventos han dado mejor rendimiento, qué ciudades han comprado más entradas, desglose de gastos...

- En este panel, podemos hacer un análisis del desglose gastos y beneficios, según su fuente de procedencia. Además de la puntuación media de los asistentes para encontrar una relación entre el beneficio y gasto y al satisfacción de los asistentes



- En este segundo panel, podemos ver el coste total, beneficio bruto y los minutos que ha durado el evento. Además de que artistas o equipos son más rentables debido al beneficio neto que dejan.

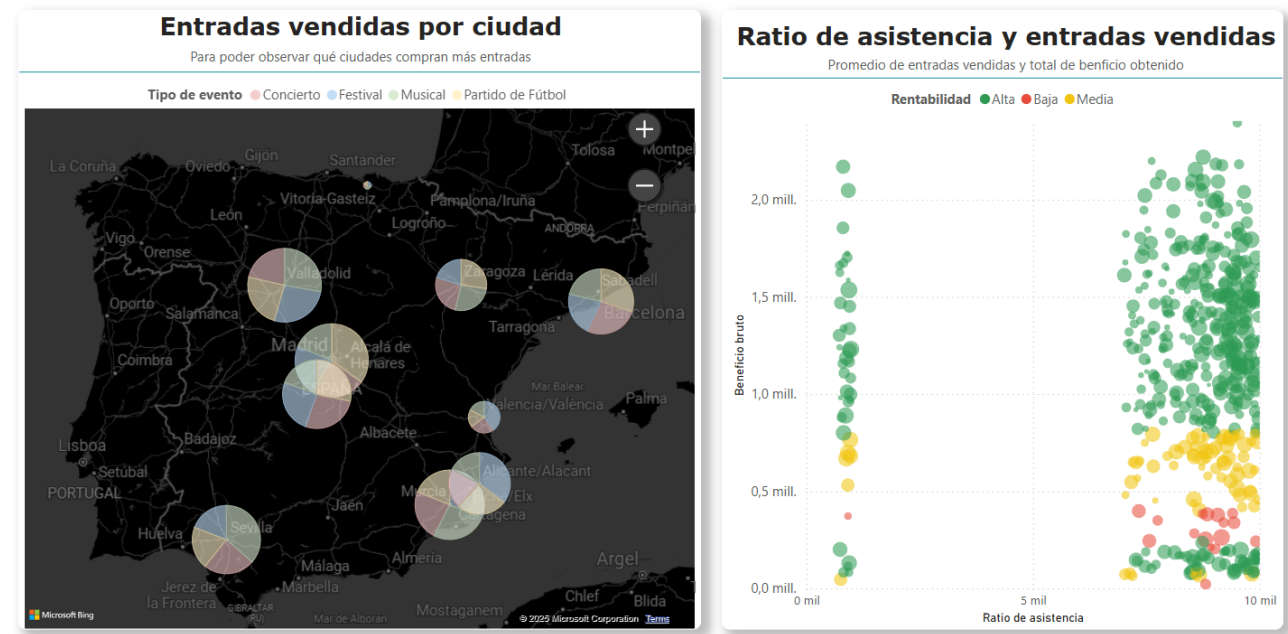


Proyecto Final. Especialización Big Data & IA

4. Análisis con powerBI

Utilizamos **powerBI** para hacer el **análisis** y poder observar, qué eventos han dado mejor rendimiento, qué ciudades han comprado más entradas, desglose de gastos...

- En este panel, muestro las ciudades dónde más entradas se han vendido y las divido según el tipo de evento. En el gráfico de la derecha, a través de un gráfico de dispersión muestro el promedio de entradas vendidas, el ratio de asistencia y el beneficio bruto obtenido.



- Por último muestro un panel con a información más importante, mostrando si el evento ha sido rentable o no lo ha sido.

Tabla de eventos y su información importante												
Tabla con la información más representativa por evento realizado												
Año	2023						2024					
nombre	Tipo de evento	Artista o equipo	Ratio de asistencia	Procedencia principal	Beneficio bruto	Rentabilidad	Tipo de evento	Artista o equipo	Ratio de asistencia	Procedencia principal	Beneficio bruto	Rentabilidad
Barcelona vs Sevilla	Partido de Fútbol	Barcelona vs Sevilla	85,34%	Alicante	3805272	Alta	Concierto	Real Madrid vs Sevilla	115,21%	Murcia	1305744	Alta
Barcelona vs Valencia	Partido de Fútbol	Barcelona vs Valencia	117,15%	Alicante	1935370	Alta						
Business-focused systematic projection	Festival	Taylor Swift	95,58%	Barcelona	349831	Baja	Concierto	Bad Bunny	58,77%	Toledo	978383	Alta
Centralized actuating paradigm	Festival	Taylor Swift	113,24%	Valencia	1241749	Alta	Musical	Atlético vs Sevilla	120,06%	Sevilla	1379602	Media
Centralized composite implementation	Musical	Coldplay	11,68%	Sevilla	1223769	Alta	Festival	Real Madrid vs Valencia	120,99%	Sevilla	1794366	Alta
Centralized discrete alliance	Festival	The Rolling Stones	108,40%	Sevilla	1660824	Alta	Concierto	Beyoncé	11,06%	Bilbao	1200259	Media
Centralized value-added functionalities	Concierto	Bad Bunny	97,37%	Valencia	1474544	Alta						
Cloned 6thgeneration protocol							Musical	Bad Bunny	106,84%	Toledo	1342011	Alta
Cloned didactic open system							Festival	Bad Bunny	97,95%	Alicante	1253492	Alta
Cloned human-resource implementation	Musical	Bad Bunny	10,10%	Toledo	83270	Alta	Festival	Sevilla vs Atlético	58,85%	Toledo	1873676	Alta
Cloned well-modulated monitoring	Musical	Beyoncé	98,32%	Murcia	1158359	Alta	Musical	Real Madrid vs Valencia	114,07%	Toledo	192763	Alta
Cloned zero administration focus	Concierto	Bad Bunny	114,86%	Murcia	1331481	Alta						
Total	Concierto	Atlético vs Atlético	93,95%	Alicante	246053540	Alta	Concierto	Atlético vs Atlético	104,68%	Alicante	305474846	Alta

Proyecto Final. Especialización Big Data & IA

5. Conclusión

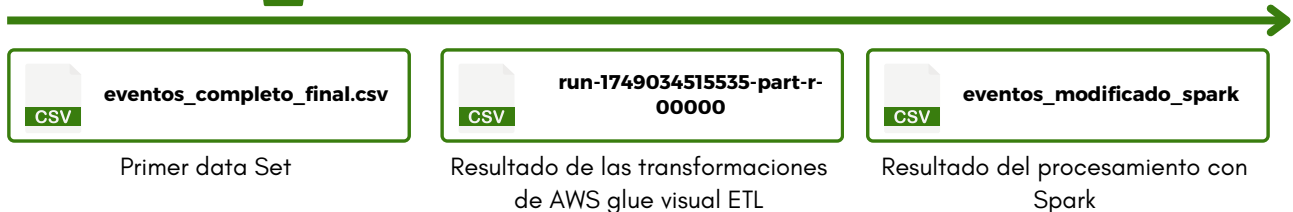
Gracias al ciclo completo de Big Data, hemos podido **ingerir, almacenar, procesar, analizar y visualizar** los datos de los eventos celebrados en un estadio de fútbol, obteniendo así información importante a cerca de la **rentabilidad** y el **comportamiento** del público. Este proyecto ha permitido generar indicadores de negocio, como el **beneficio neto** por evento, el **tipo de espectador más rentable** o la **evolución anual** del rendimiento económico...

A nivel personal, este proyecto me ha servido para comprender de forma práctica cómo se conectan las distintas fases del **ciclo Big Data** y cómo cada herramienta (**S3, Glue, EMR, Athena, Power BI**) aporta valor en cada etapa. He aprendido la importancia de diseñar **transformaciones con lógica de negocio**, como sacar nuevas columnas con ingresos totales, beneficio bruto... y de visualizar los datos de forma clara para facilitar la toma de decisiones.

Como posibles mejoras, me gustaría **incorporar modelos predictivos** en futuras versiones del proyecto, que permitan anticipar la rentabilidad de eventos antes de su celebración. También podría **ampliarse el análisis** incluyendo variables externas como el **clima**, la **competencia** u otros factores económicos que afecten la asistencia como opiniones en reseñas o redes sociales.

En definitiva, ha sido una experiencia muy enriquecedora que demuestra el potencial del Big Data para resolver problemas reales y tomar decisiones basadas en datos.

◦ Datos utilizados



◦ Fuentes consultadas

- <https://docs.aws.amazon.com/es-es/glue/latest/dg/edit-nodes-chapter.html>
- <https://www.knowmadmood.com/es/blog/transformaciones-personalizadas-con-aws-glue>
- <https://www.youtube.com/watch?v=OsmNGmh-qQI>
- <https://catalog.us-east-1.prod.workshops.aws/workshops/ee59d21b-4cb8-4b3d-a629-24537cf37bb5/en-US/lab2>
- <https://www.youtube.com/embed/pJF2kYGtO4A?feature=oembed>
- <https://spark.apache.org/docs/latest/quick-start.html>
- <https://www.youtube.com/embed/6OEwdJbnDY8?feature=oembed>
- <https://www.youtube.com/watch?v=6Vr9ZUntCyE>
- <https://www.youtube.com/embed/AzbPeJmbw2U?feature=oembed>
- <https://learn.microsoft.com/es-es/power-bi/fundamentals/service-get-started>
- <https://www.youtube.com/watch?v=pwJuFbyhZFE>