PR_08.2 Dani Gayol Rodríguez

PR_08.2 Dani Gayol Rodríguez 1

Apartado A 1

1.) Crea un bucket S3 con una carpeta dentro, por ejemplo, clima/espana 2

2.) Mediante un comando AWS CLI, copia los archivos csv con las mediciones de todas las estaciones metereológicas de España en él. 2

3.) Crea una base de datos en del Data Catalog que se llame espana. 2

4.) Crea un Crawler que nos permita agregar a esa base de los ficheros de las estaciones meteorológicas de España (Pon como prefijo a la tabla espcsv_. 2

5.) Guarda el Crawler pero no lo ejecutes. 2

Apartado B 2

1.) Crea una carpeta dentro del bucket anterior (clima) con el nombre parquet. por ejemplo, clima/parquet 3

2.) Crea un trabajo mediante Visual ETL que nos permita cambiar el esquema de los CSV's que acabamos de importar poniendo los nombres de los campos en español y guardando los datos en formato parquet en la carpeta del punto anterior. 3

3.) Guarda el trabajo, pero no lo ejecutes. 3

Apartado C 3

1.) Crea un Crawler AWS GLUE que nos explore el bucket del ejercicio anterior (parquet) generando la tabla correspondiente en la base de datos clima. Ponle de prefijo a la tabla espparq_. 4

2.) Guarda el rastreador, pero no lo ejecutes. 4

Apartado D 4

1.) Crea un disparador (trigger) -puedes llamarlo espa_ab - que después de finalizado el crawler del apartado A lance el trabajo del apartado B. 5

2.) Crea un disparador (trigger) -puedes llamarlo espa_bc - que después de finalizado el trabajo del apartado B lance el trabao del apartado C. 5

3.) Finalmente hemos de crear un triggerr bajo demanda que nos arranque el crawler inicial (en nuestro caso el del apartado A) 5

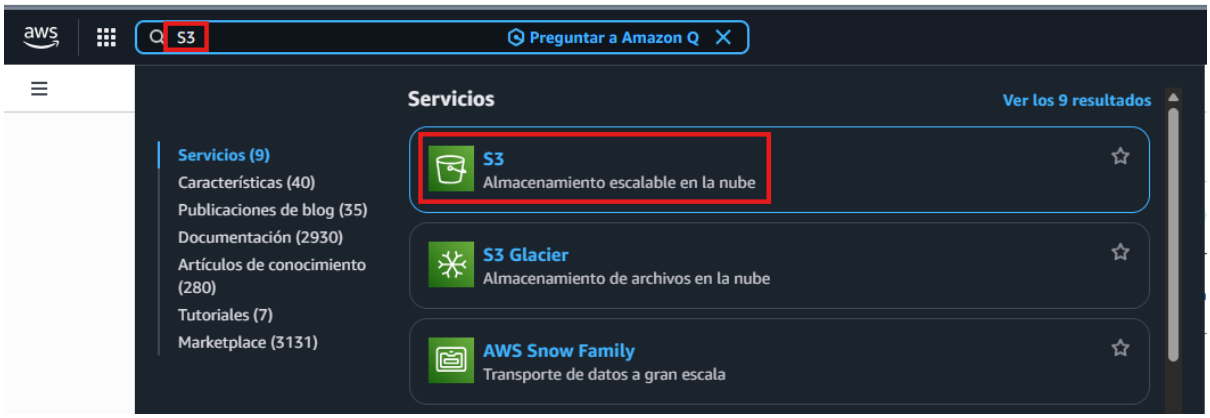
4.) Arranca este manualmente este último disparador. 5

Apartado E 5
1.) Muestra los archivos creados. 6
2.) Muestra las tablas y campos creados. 6
Apartado F 6
1.) ¿Cuántas mediciones tenemos de España? 7
2.) Sabiendo los códigos de las 4 estaciones de Asturias ¿Cuántas mediciones tenemos de Asturias? 7
3.) ¿Cuántas mediciones tenemos de Oviedo? 7
4.) ¿Cuál es la medición más antigua de España, Asturias y Oviedo? 7
5.) Haz una tabla comparativa con los tiempos de ejecución de las consultas sobre las tres diferentes tablas (las de la práctica anterior y las dos de esta práctica) ¿Cuáles han sido las más veloces? 7

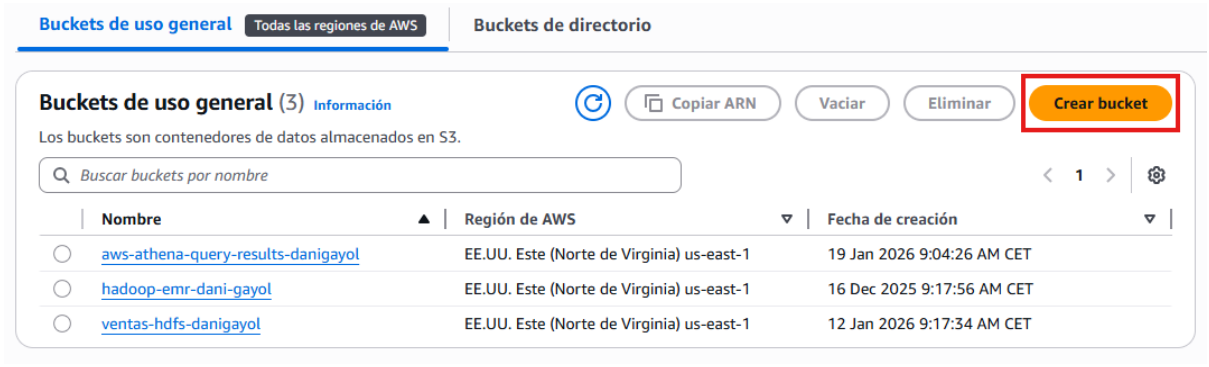
Apartado A

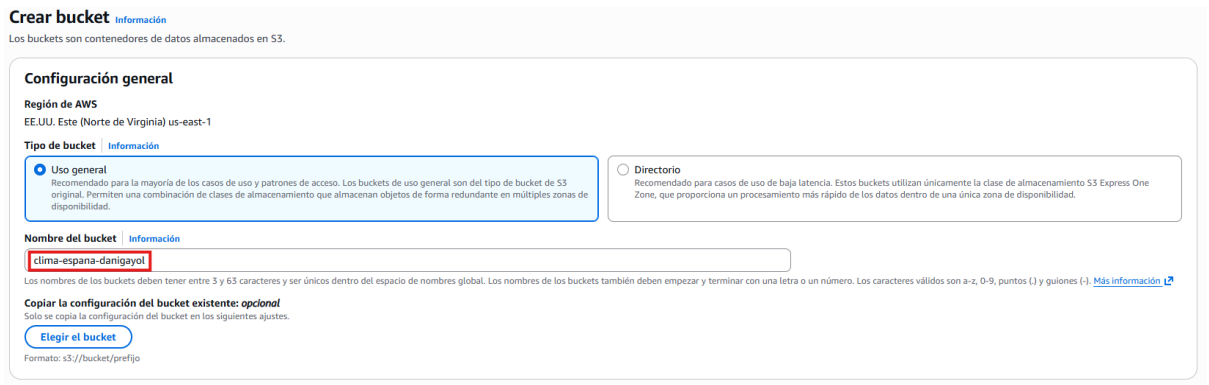
1.) Crea un bucket S3 con una carpeta dentro, por ejemplo, clima/espana

Para crear el bucket, buscamos “S3” en la barra de búsqueda de AWS

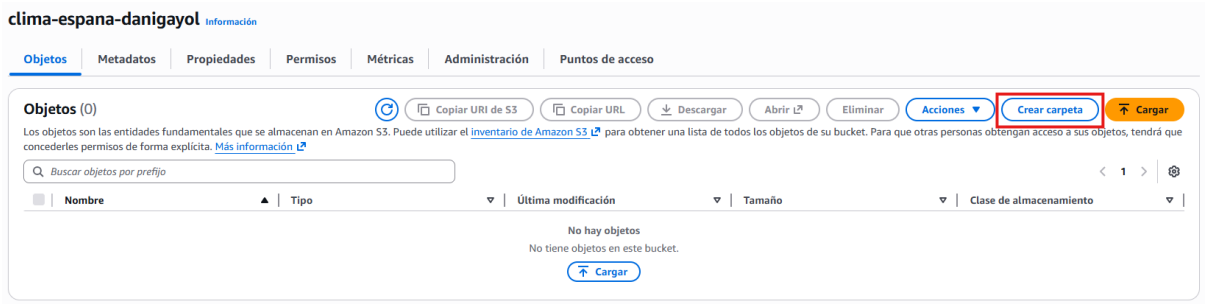


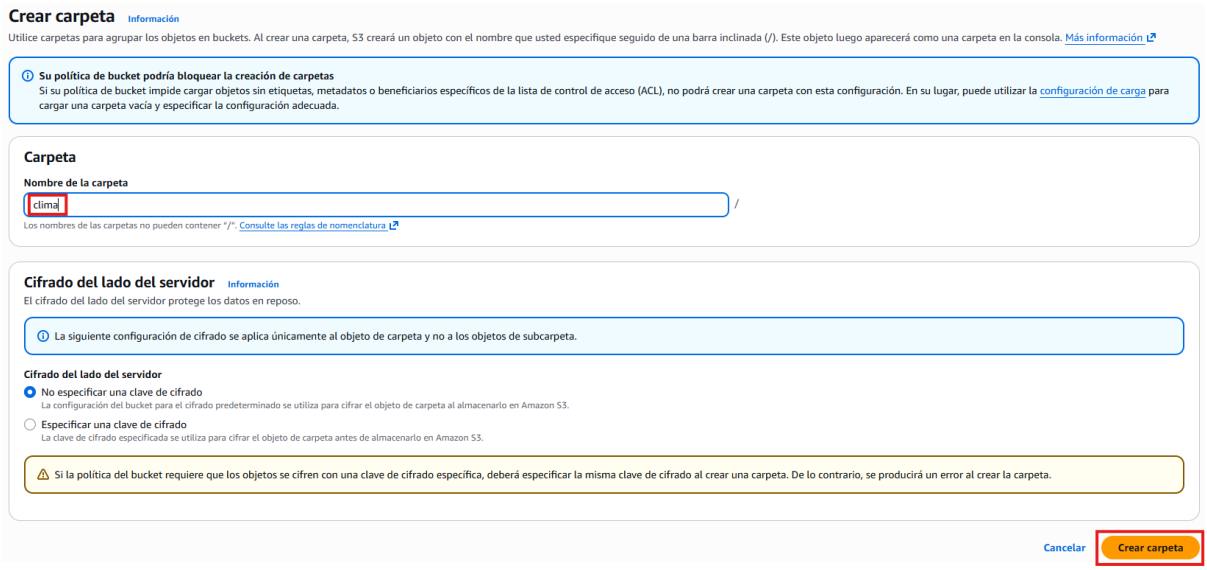
Ahora le damos al botón de “Crear Bucket”





Ahora una vez creado el bucket, le vamos a crear la carpeta dentro



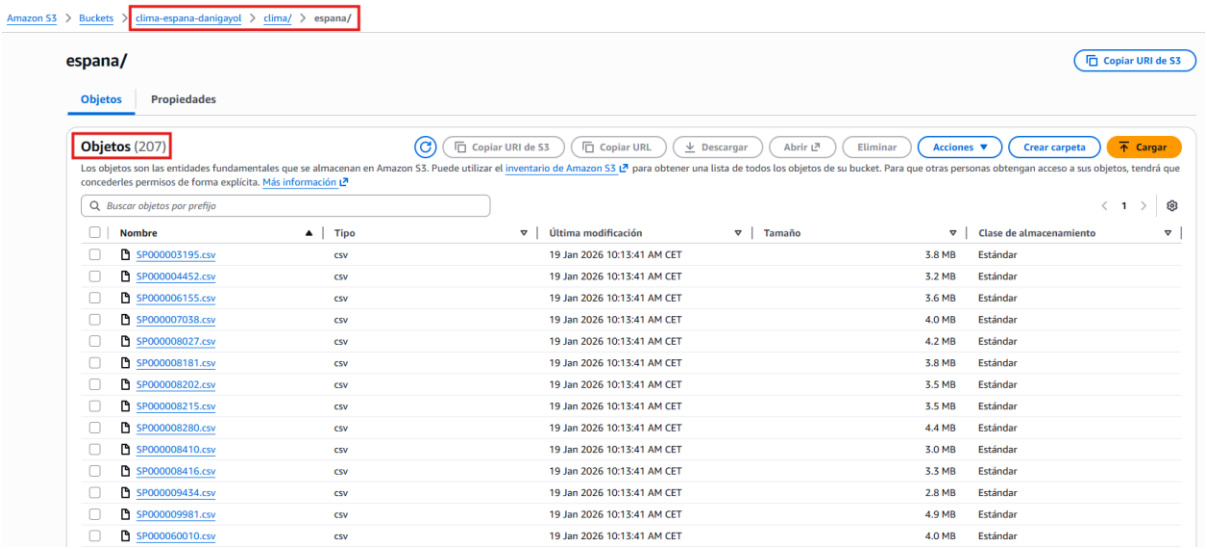


2.) Mediante un comando AWS CLI, copia los archivos csv con las mediciones de todas las estaciones metereológicas de España en él.

Para ejecutar el comando, abrimos “CMD” y ponemos lo siguiente:

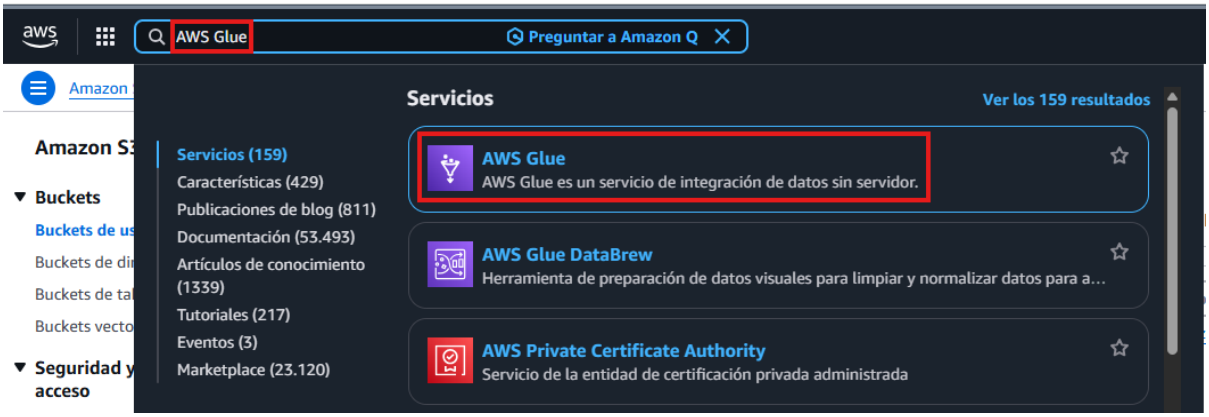
```
C:\Users\Mañana\.aws>aws s3 cp s3://noaa-ghcn-pds/csv/by_station/ s3://clima-espana-danigayol/clima/espana/ --recursive
--exclude "*" --include "SP*.csv"
copy: s3://noaa-ghcn-pds/csv/by_station/SP000008202.csv to s3://clima-espana-danigayol/clima/espana/SP000008202.csv
copy: s3://noaa-ghcn-pds/csv/by_station/SP000004452.csv to s3://clima-espana-danigayol/clima/espana/SP000004452.csv
copy: s3://noaa-ghcn-pds/csv/by_station/SP000008181.csv to s3://clima-espana-danigayol/clima/espana/SP000008181.csv
copy: s3://noaa-ghcn-pds/csv/by_station/SP000008027.csv to s3://clima-espana-danigayol/clima/espana/SP000008027.csv
copy: s3://noaa-ghcn-pds/csv/by_station/SP000003195.csv to s3://clima-espana-danigayol/clima/espana/SP000003195.csv
copy: s3://noaa-ghcn-pds/csv/by_station/SP000007038.csv to s3://clima-espana-danigayol/clima/espana/SP000007038.csv
copy: s3://noaa-ghcn-pds/csv/by_station/SP000006155.csv to s3://clima-espana-danigayol/clima/espana/SP000006155.csv
copy: s3://noaa-ghcn-pds/csv/by_station/SP000008280.csv to s3://clima-espana-danigayol/clima/espana/SP000008280.csv
```

Y ahora para verificar, nos vamos al bucket que creamos en “S3” y entramos en la carpeta donde copiamos los archivos

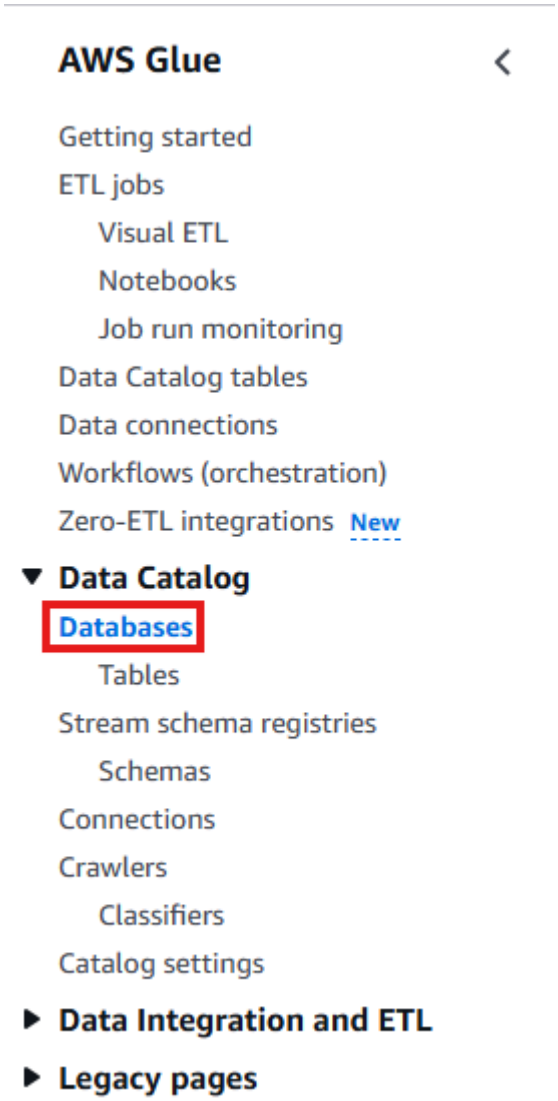


3.) Crea una base de datos en del Data Catalog que se llame espana.

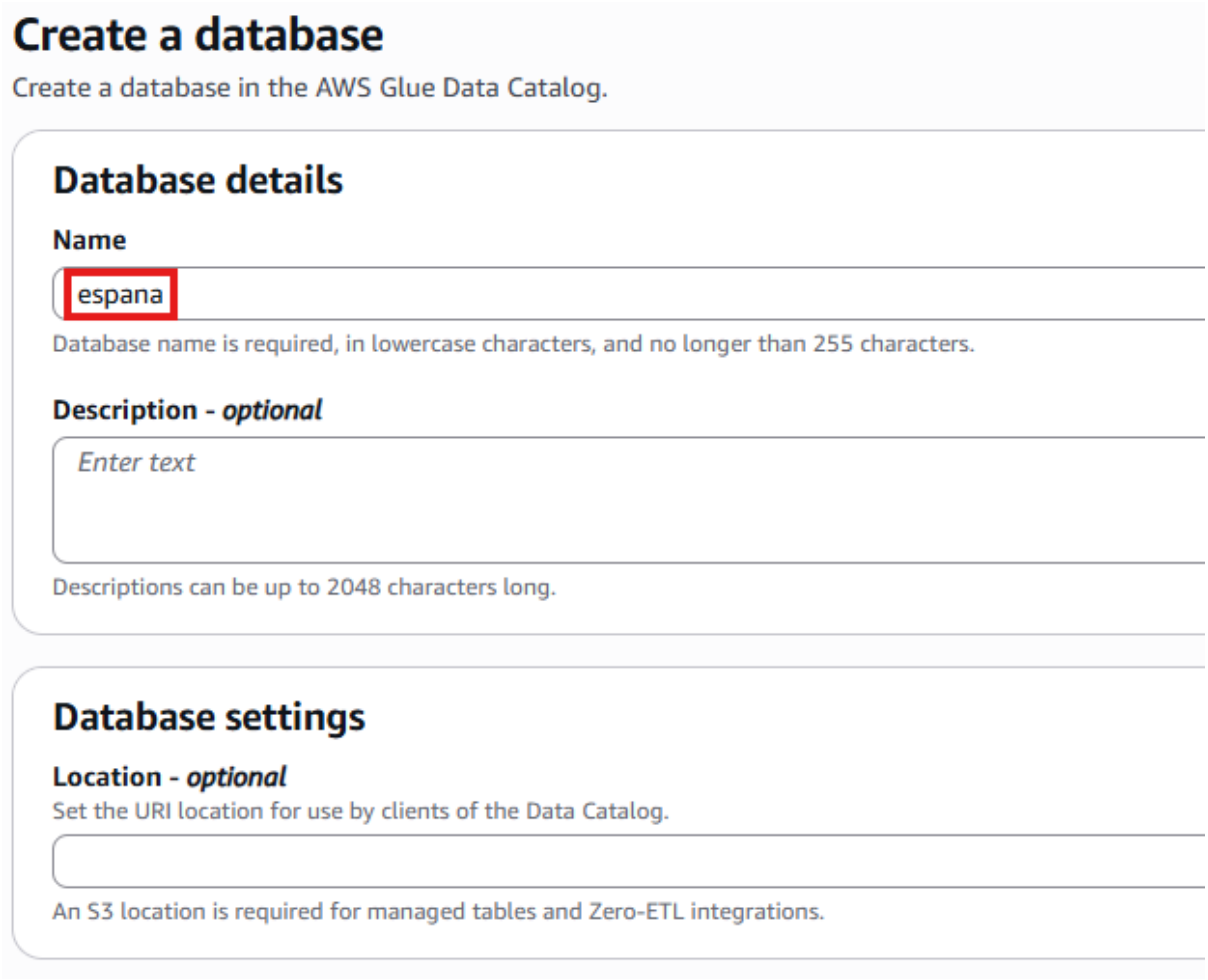
Para crear la base de datos, nos dirigimos a “AWS Glue”,



En el menú de la izquierda, entramos en “Databases”

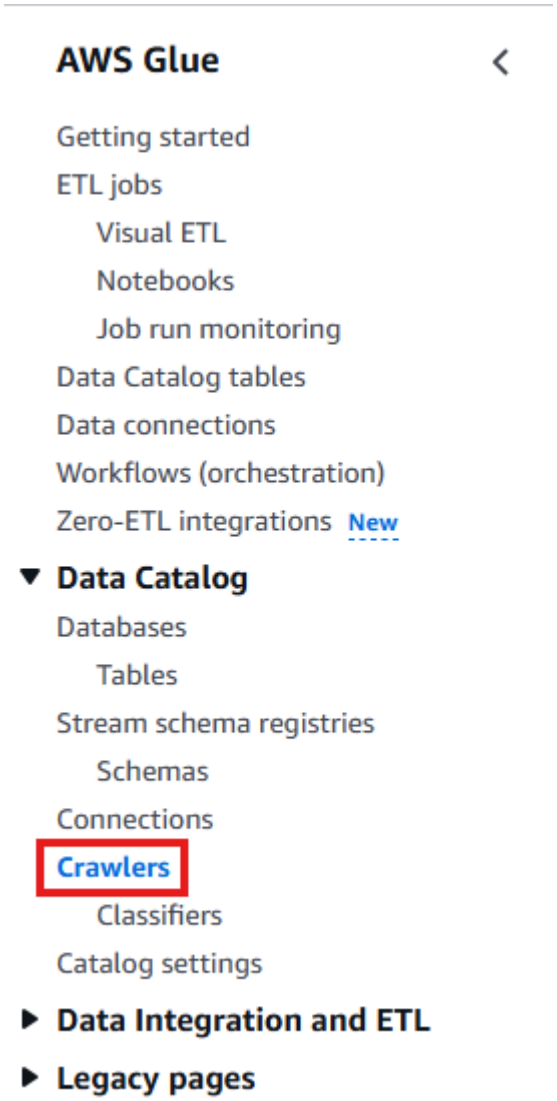


Una vez dentro, le damos al botón de “Add Database”

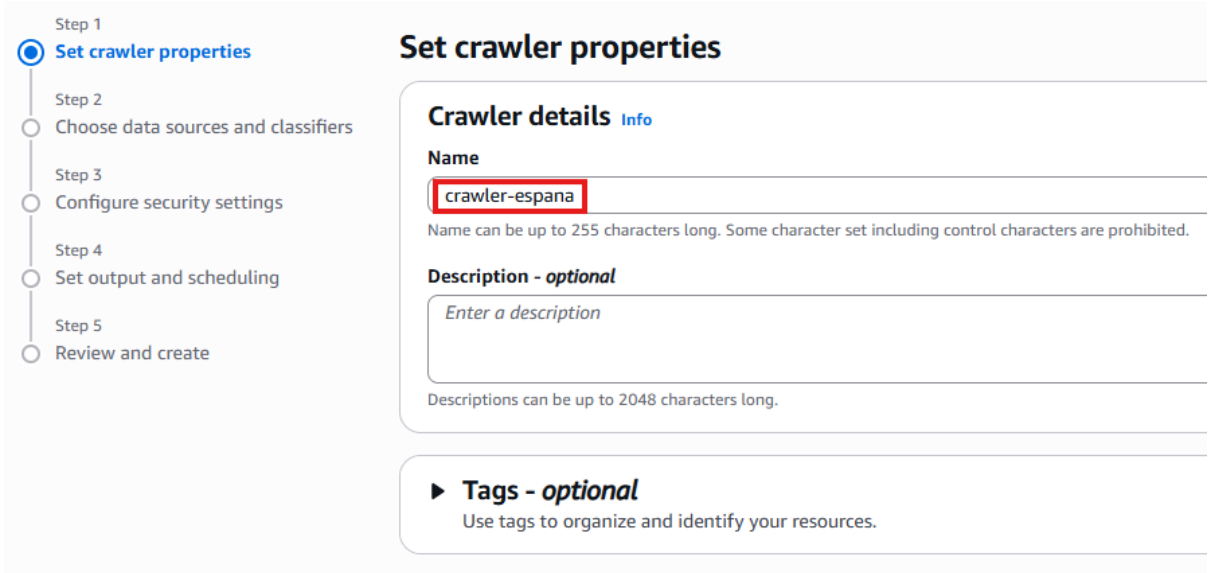


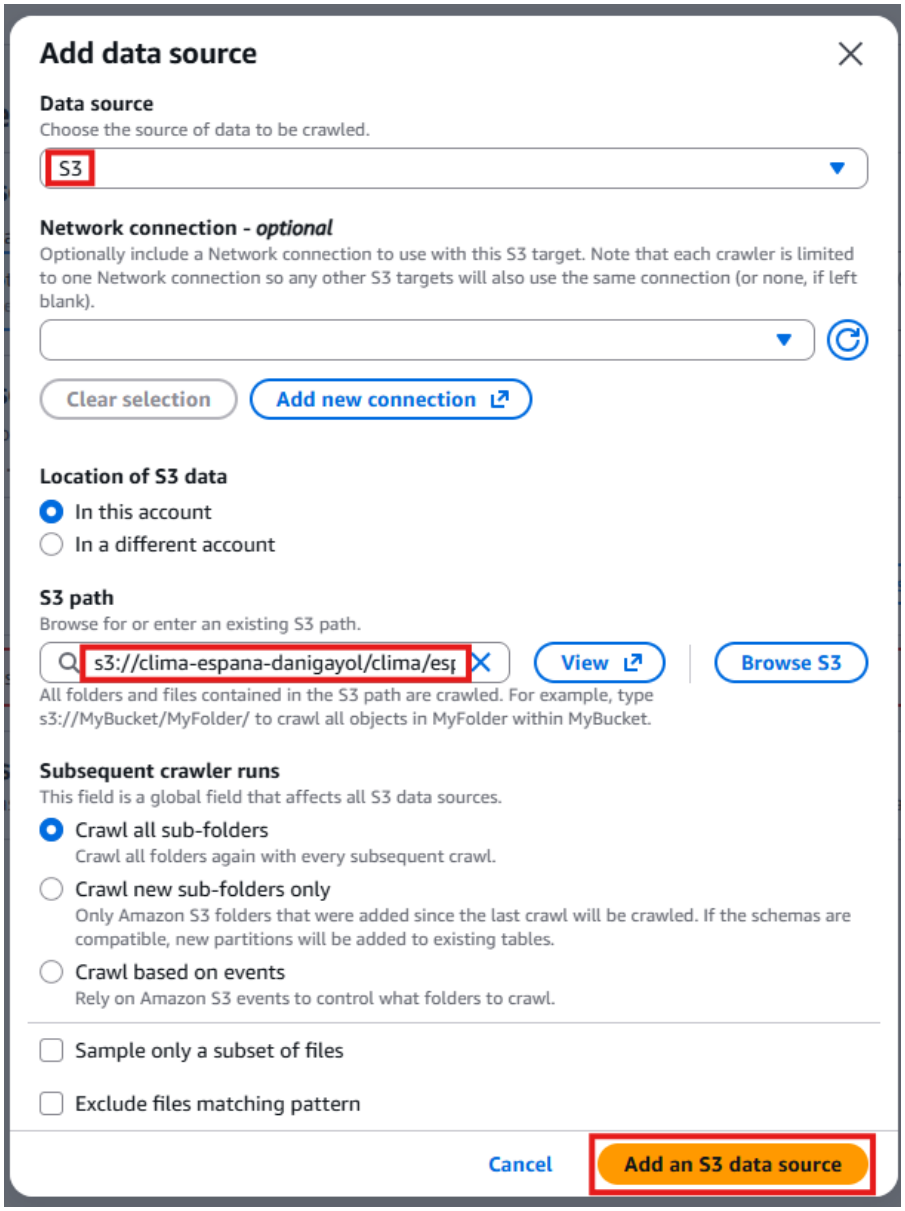
4.) Crea un Crawler que nos permita agregar a esa base de los ficheros de las estaciones meteorológicas de España (Pon como prefijo a la tabla espcsv_.

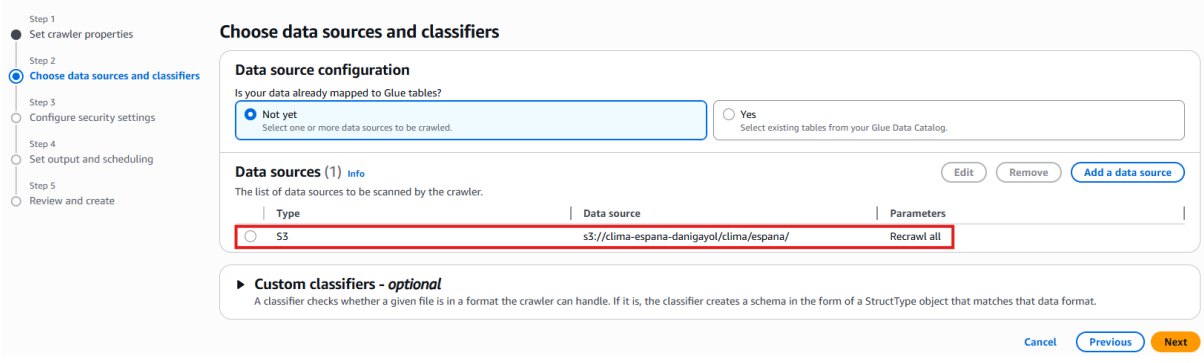
Ahora en el menú de la izquierda, entramos en “Crawlers” y le damos al botón de “Create Crawler”

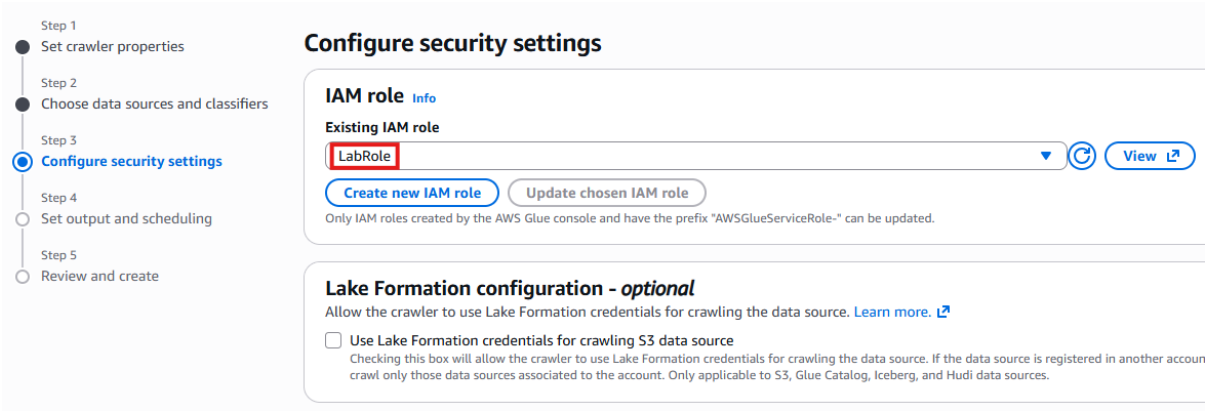


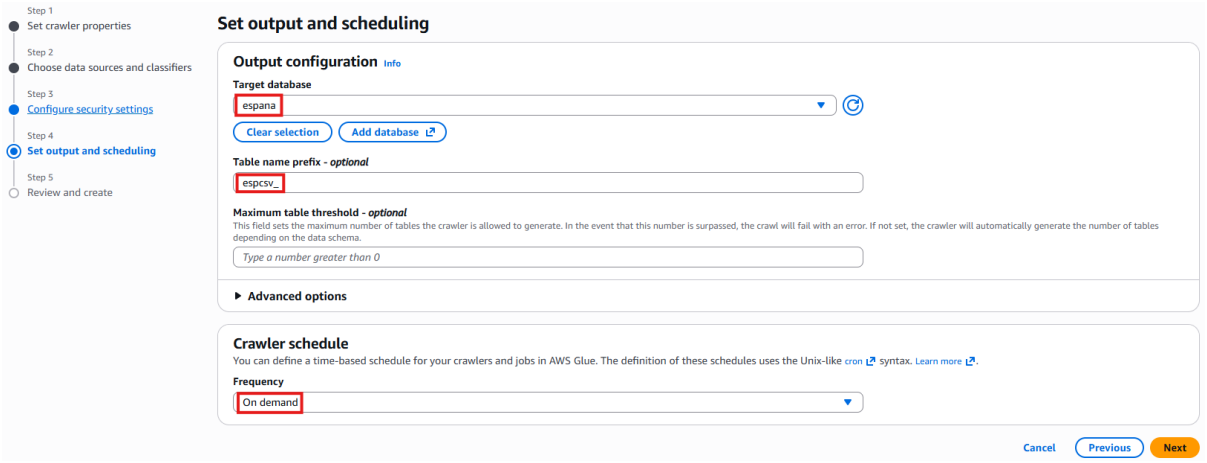
Al "Crawler" le ponemos la siguiente configuración:



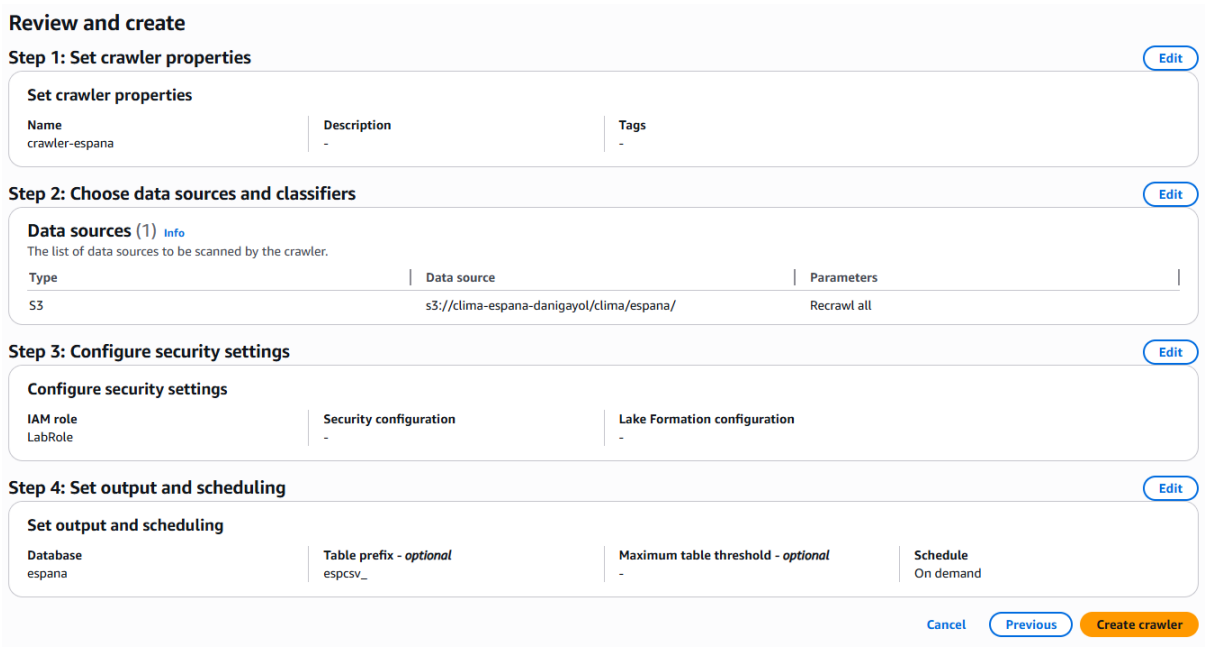

Add data source
Data source
Choose the source of data to be crawled.
S3
Network connection - optional
Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).
Clear selection
Add new connection
Location of S3 data
In this account
In a different account
S3 path
Browse for or enter an existing S3 path.
s3://clima-espana-danigayol/clima/esp
View
Browse S3
All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.
Subsequent crawler runs
This field is a global field that affects all S3 data sources.
Crawl all sub-folders
Crawl all folders again with every subsequent crawl.
Crawl new sub-folders only
Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.
Crawl based on events
Rely on Amazon S3 events to control what folders to crawl.
Sample only a subset of files
Exclude files matching pattern
Cancel
Add an S3 data source


Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Choose data sources and classifiers
Data source configuration
Is your data already mapped to Glue tables?
Not yet
Select one or more data sources to be crawled.
Yes
Select existing tables from your Glue Data Catalog.
Data sources (1) Info
The list of data sources to be scanned by the crawler.
Edit
Remove
Add a data source
Type
Data source
Parameters
S3
s3://clima-espana-danigayol/clima/espana/
Recrawl all
Custom classifiers - optional
A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.
Cancel
Previous
Next



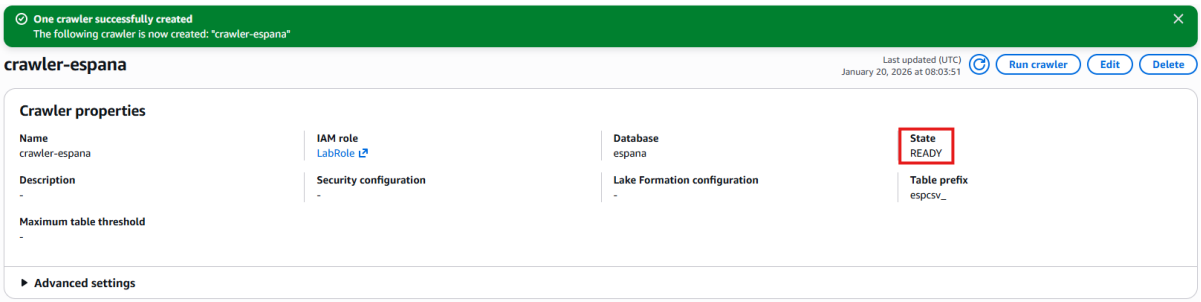


Y así quedaría finalmente:



5.) Guarda el Crawler pero no lo ejecutes.

Una vez terminado de configurar el “Crawler”, le damos al botón de “Create Crawler”

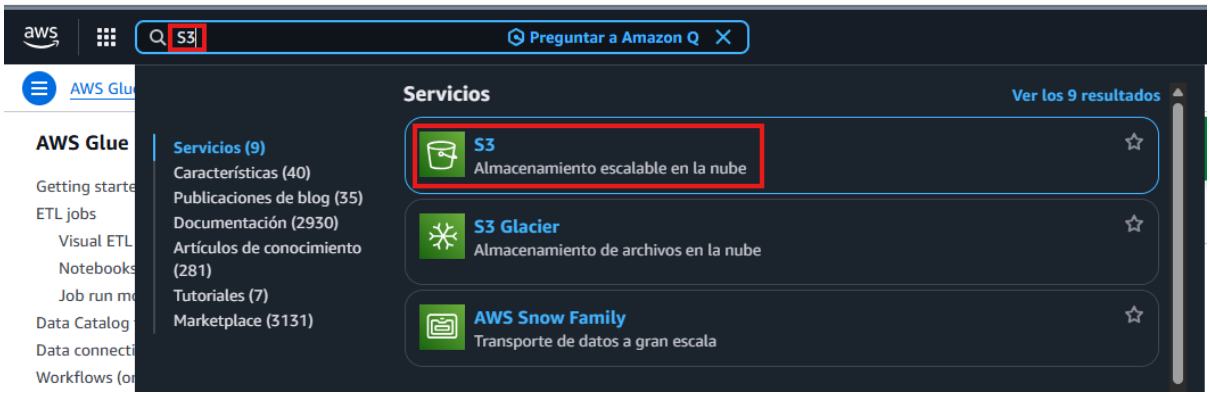


IMPORTANTE, le dejamos en estado “Ready”, NO PULSAMOS EL BOTÖN “RUN CRAWLER”

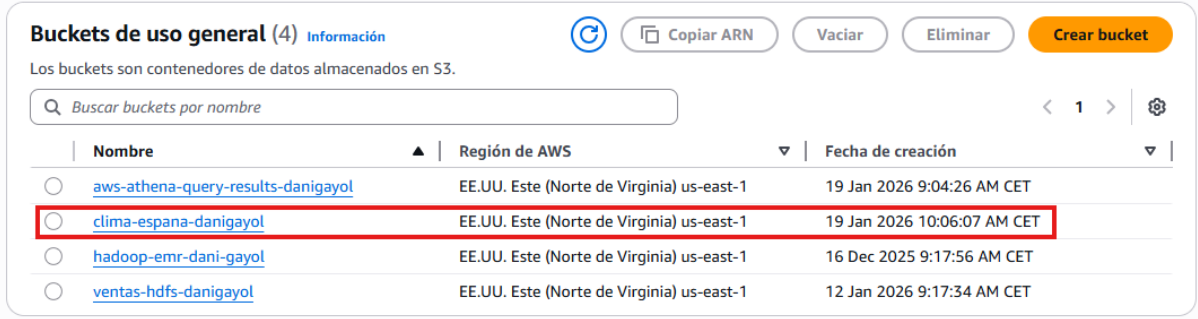
Apartado B

1.) Crea una carpeta dentro del bucket anterior (clima) con el nombre parquet. por ejemplo, clima/parquet

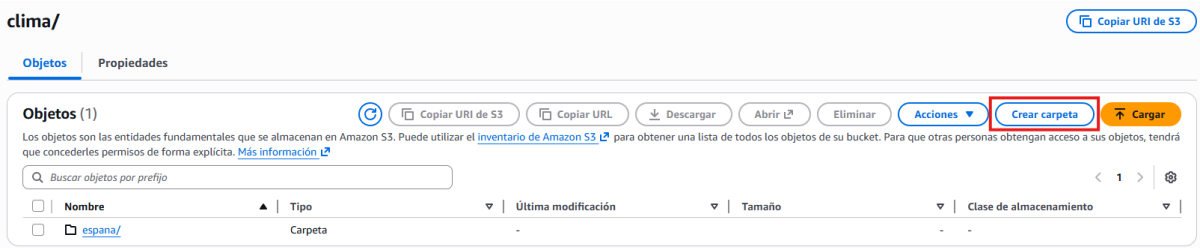
Nos volvemos a dirigir a “S3” en la barra de búsqueda de AWS

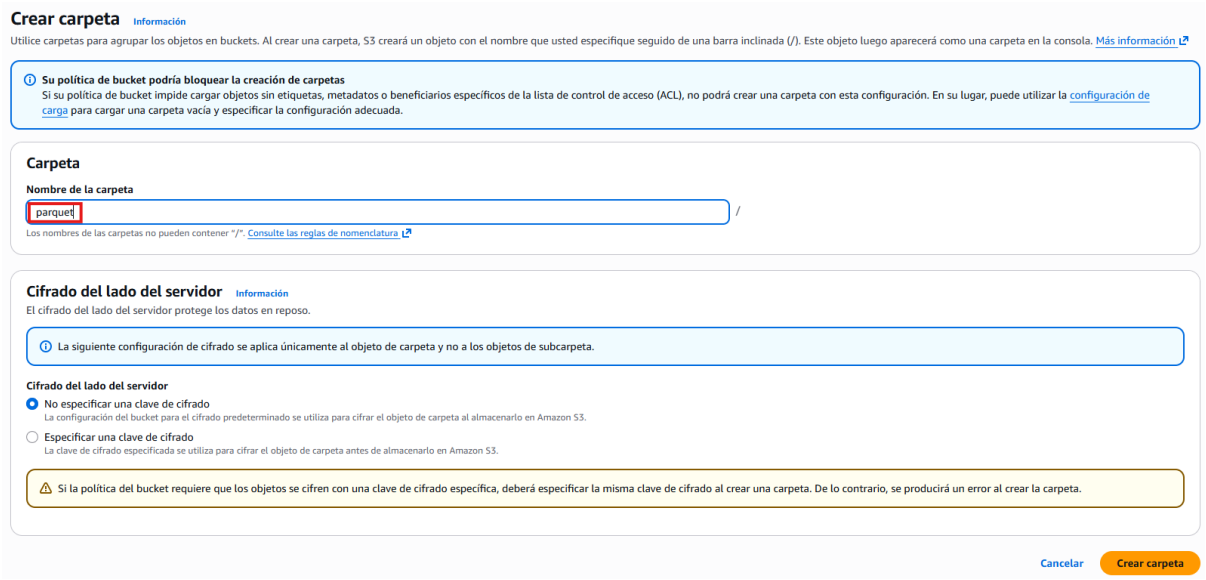


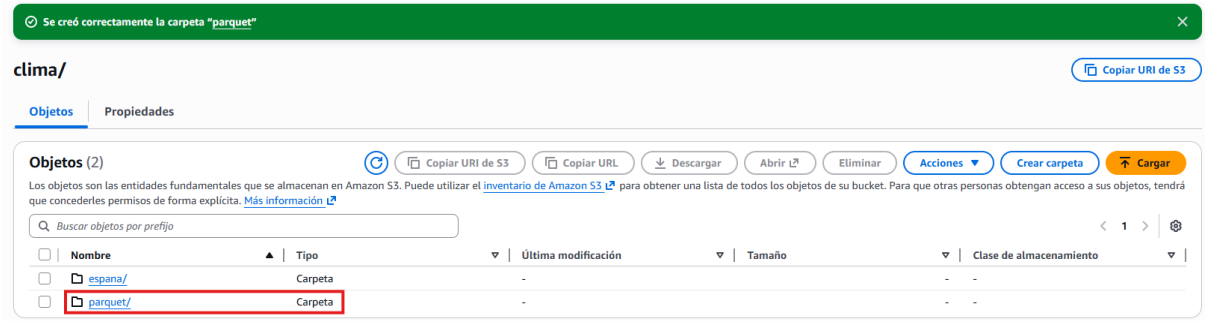
Entramos dentro del bucket que tenemos creado:



Le damos a “Crear Carpeta” (dentro de la carpeta “clima”)

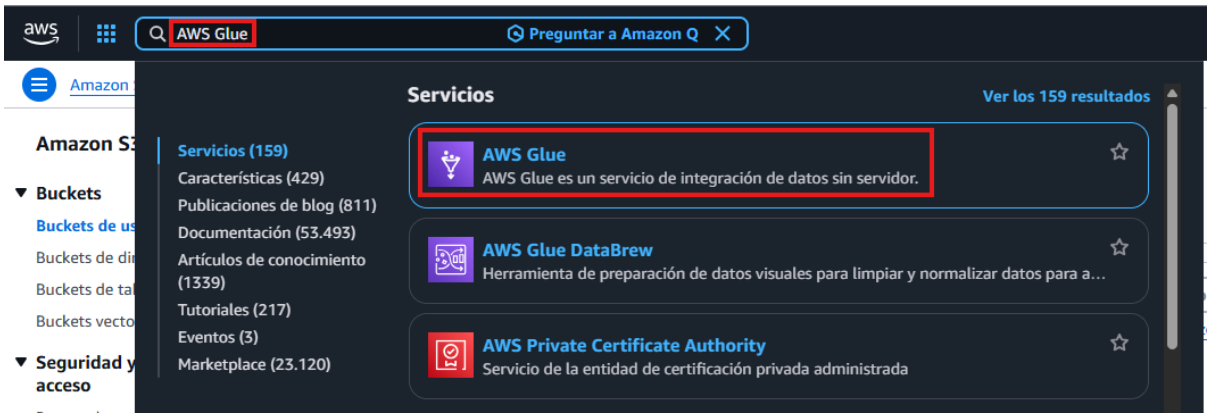




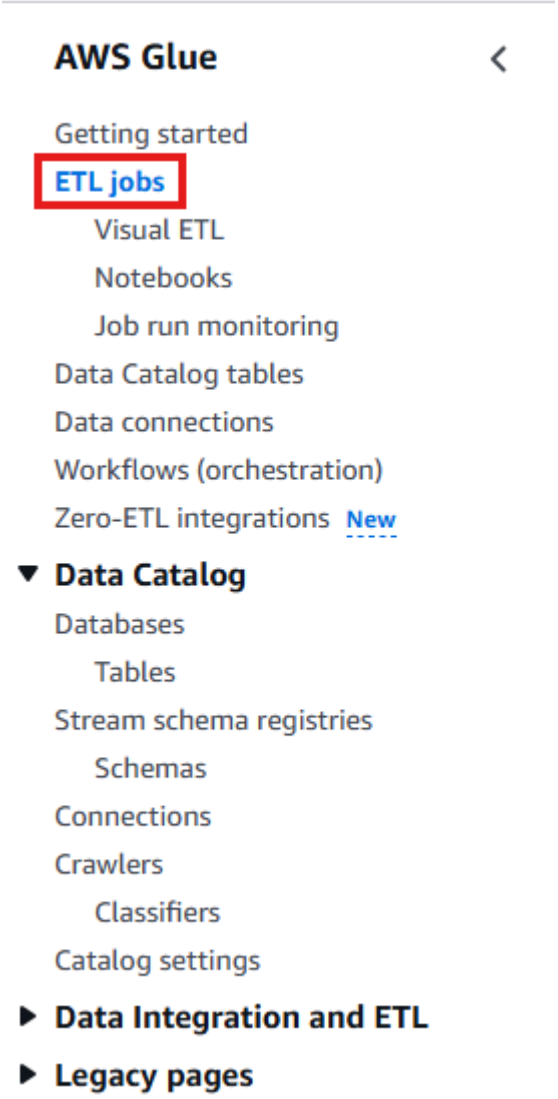


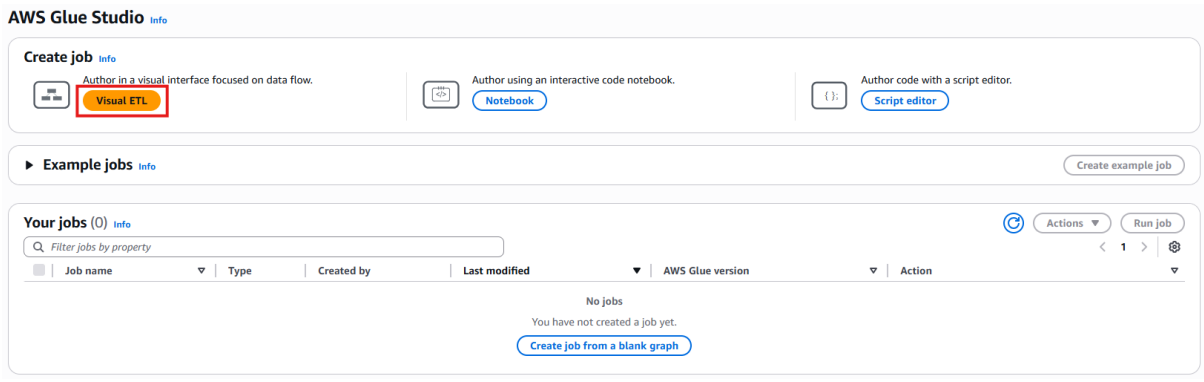
2.) Crea un trabajo mediante Visual ETL que nos permita cambiar el esquema de los CSV's que acabamos de importar poniendo los nombres de los campos en español y guardando los datos en formato parquet en la carpeta del punto anterior.

Nos volvemos a dirigir a "AWS Glue"

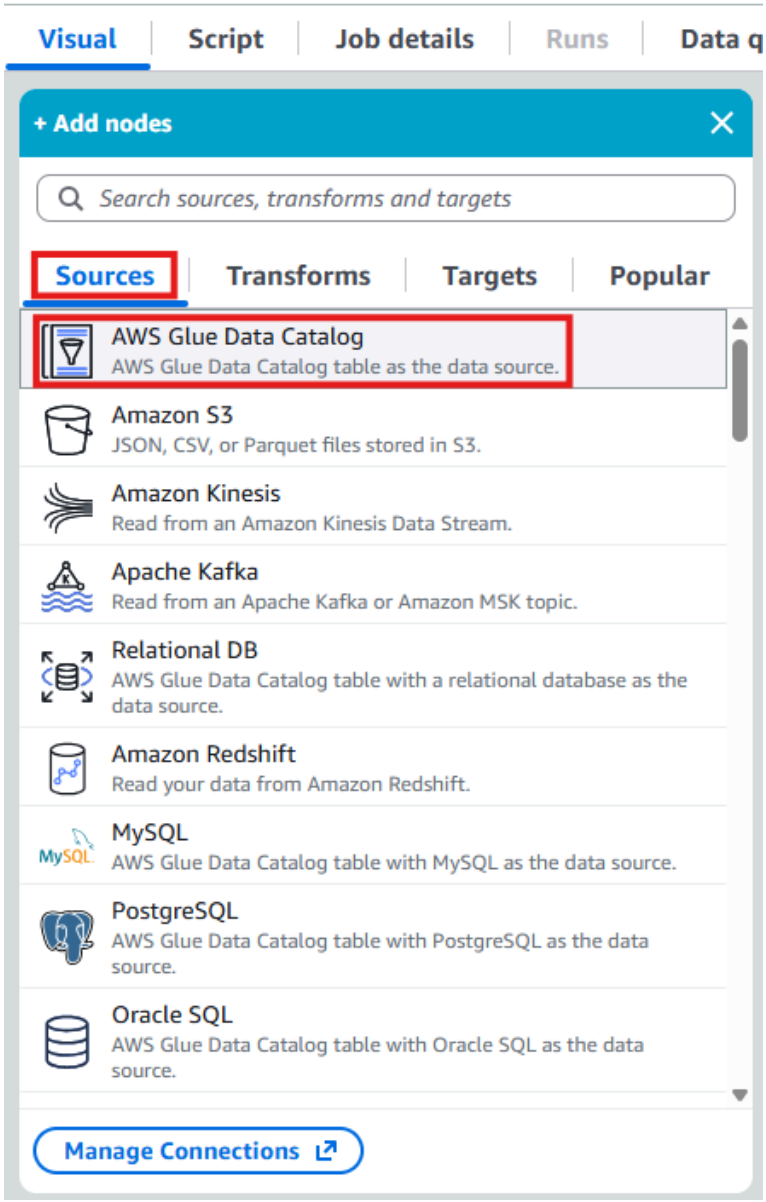


En el manú de la izquierda, entramos en “ETL Jobs” y dentro de “ETL Jobs”, entramos en “Visual ETL”

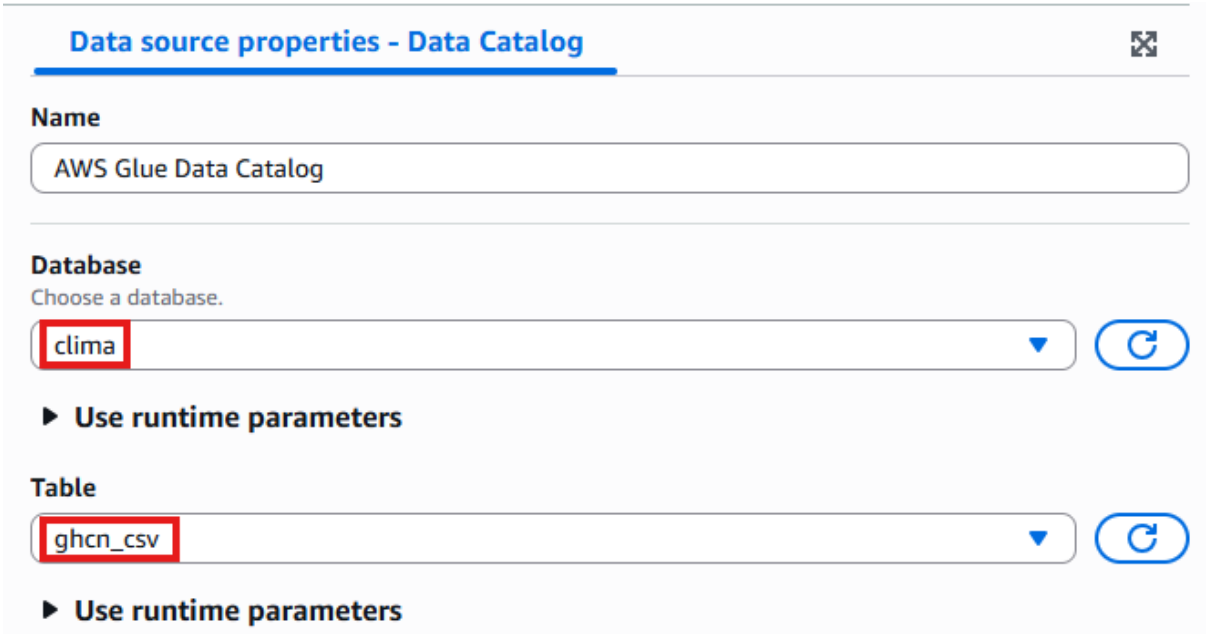




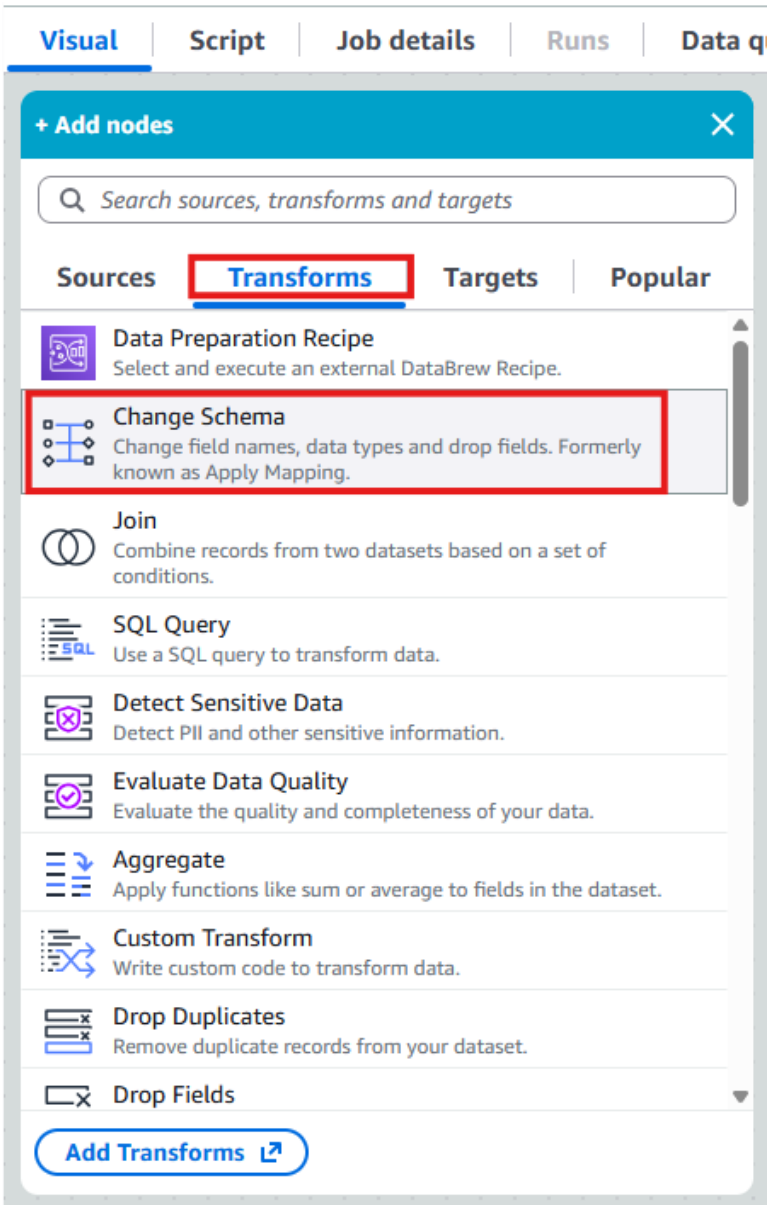
Seleccionamos el menú de “Sources” y dentro de él, la opción "Change Schema”

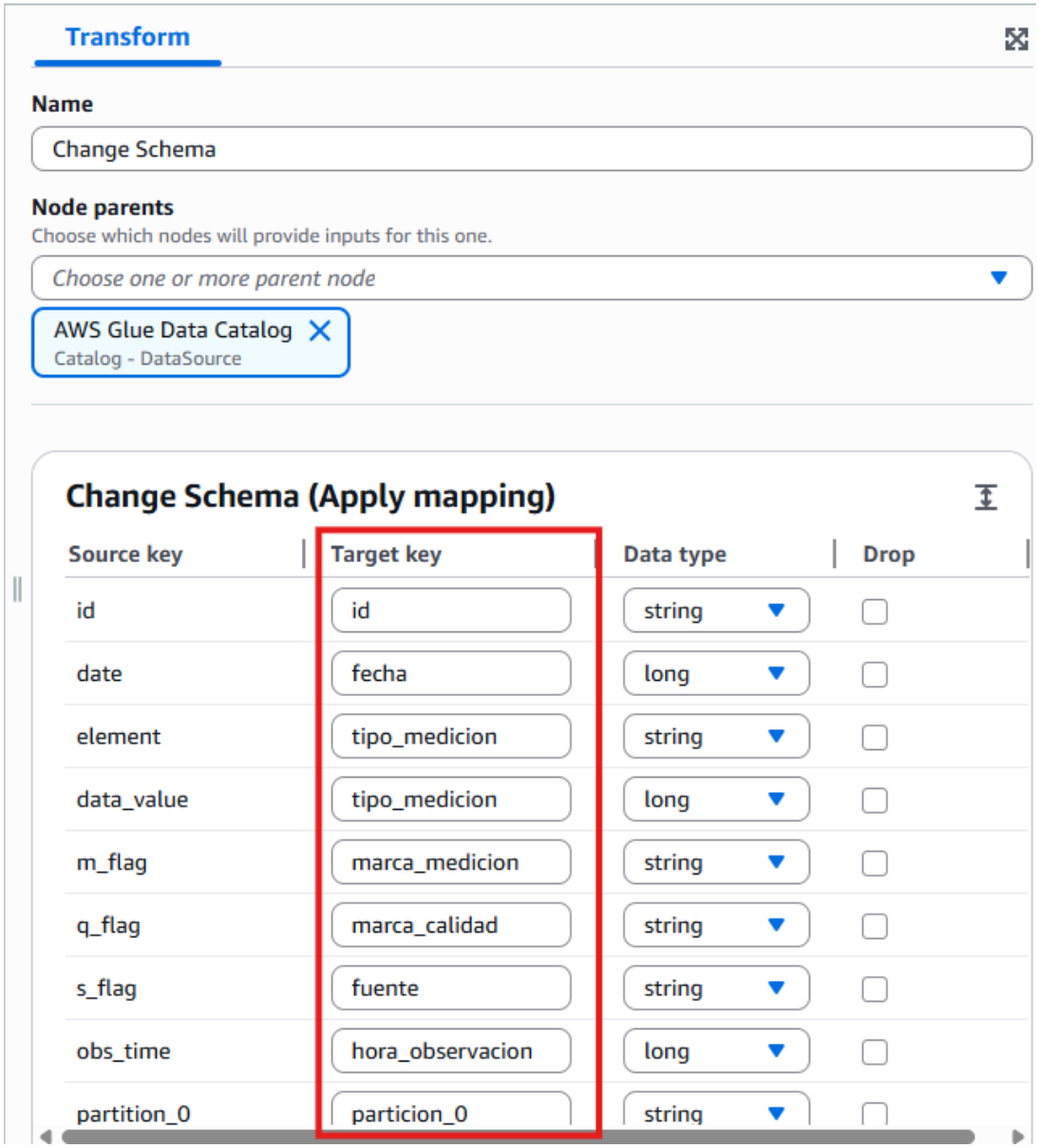


Lo configuramos de la siguiente manera:

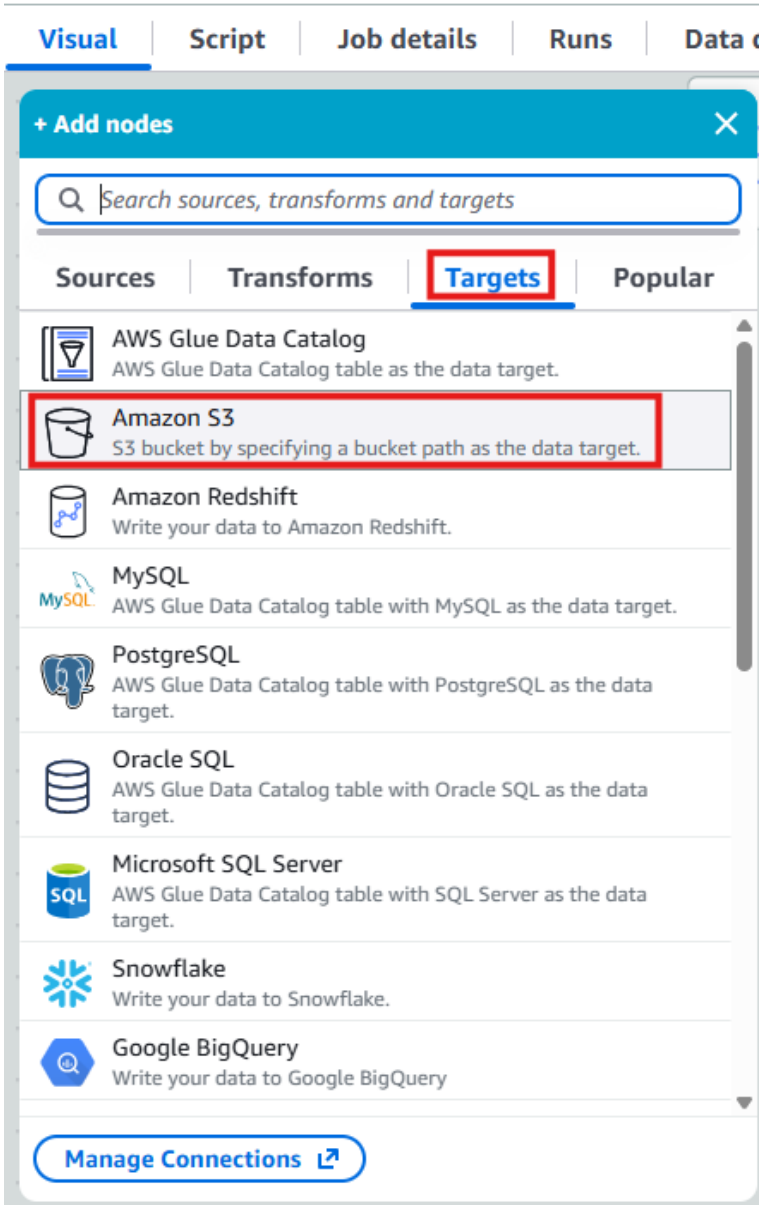


Ahora, seleccionamos el menú de “Transforms” y dentro de él, la opción "Change Schema”

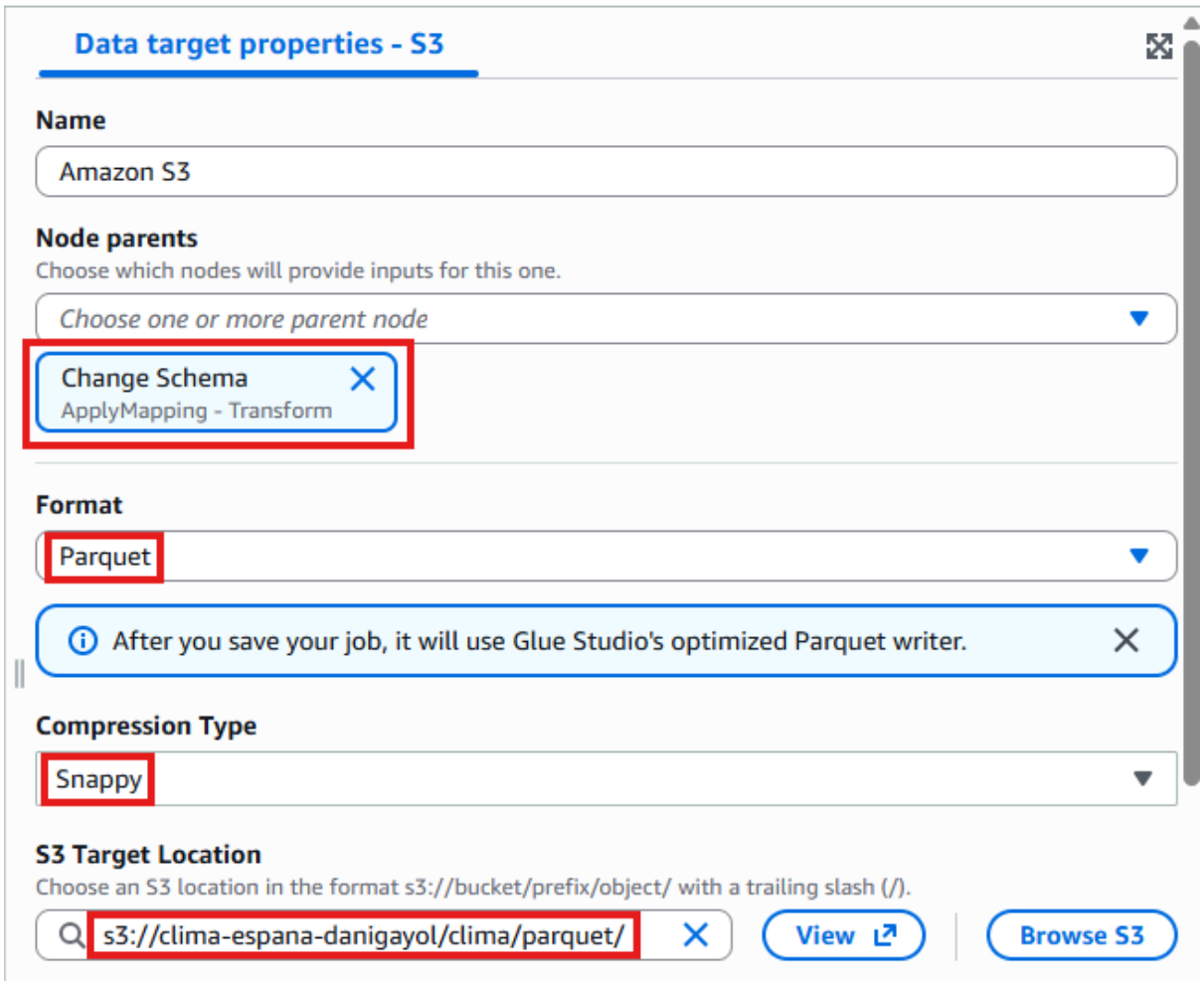

Visual
Script
Job details
Runs
Data q
+ Add nodes
Search sources, transforms and targets
Sources
Transforms
Targets
Popular
Data Preparation Recipe
Select and execute an external DataBrew Recipe.
Change Schema
Change field names, data types and drop fields. Formerly known as Apply Mapping.
Join
Combine records from two datasets based on a set of conditions.
SQL Query
Use a SQL query to transform data.
Detect Sensitive Data
Detect PII and other sensitive information.
Evaluate Data Quality
Evaluate the quality and completeness of your data.
Aggregate
Apply functions like sum or average to fields in the dataset.
Custom Transform
Write custom code to transform data.
Drop Duplicates
Remove duplicate records from your dataset.
Drop Fields
Add Transforms



Ahora, seleccionamos el menú de “Targets” y dentro de él, la opción "Amazon S3”



Y lo configuramos de la siguiente manera:



3.) Guarda el trabajo, pero no lo ejecutes.

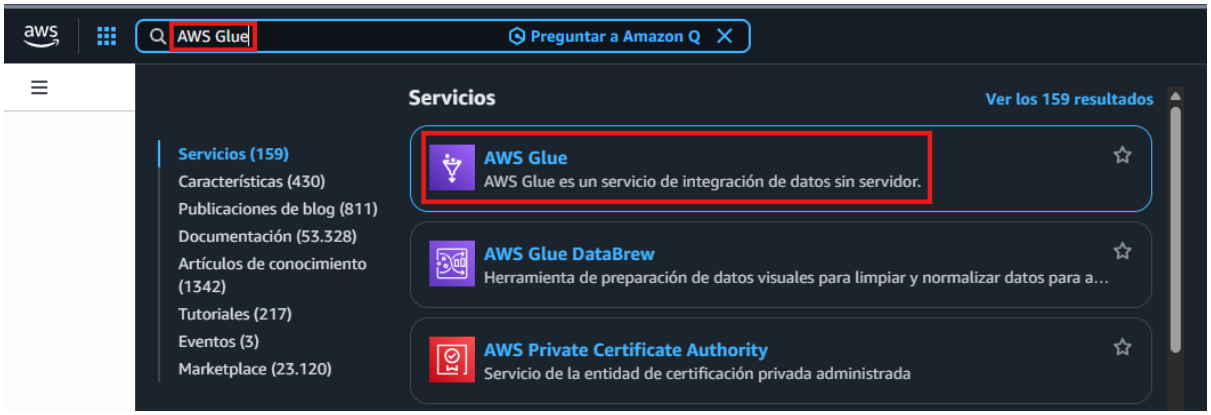
Finalmente, le ponemos un nombre y le damos a “save”

Y nos aparecerá aquí;

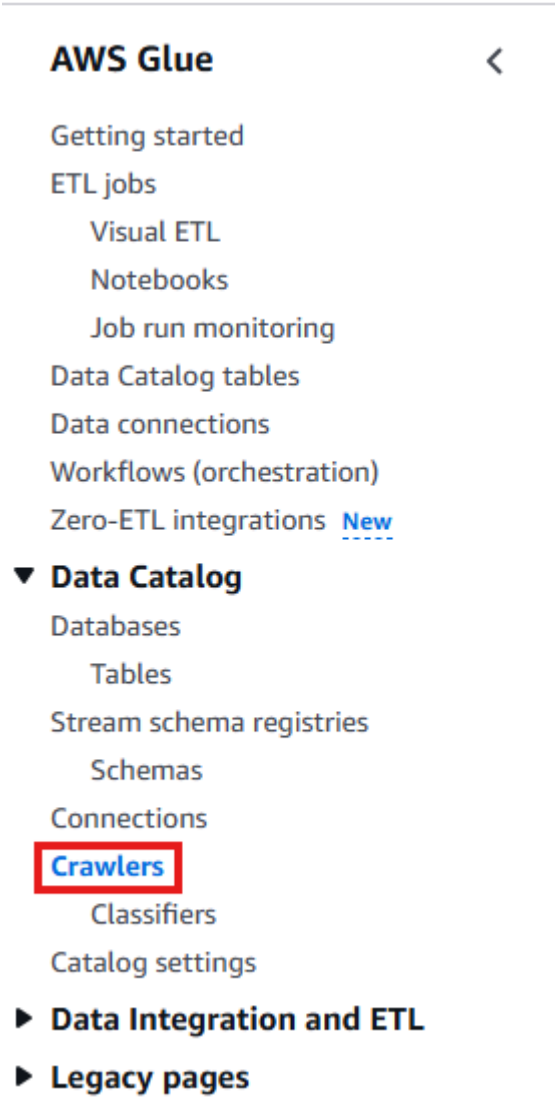
Apartado C

1.) Crea un Crawler AWS GLUE que nos explore el bucket del ejercicio anterior (parquet) generando la tabla correspondiente en la base de datos clima. Ponle de prefijo a la tabla espparq_.

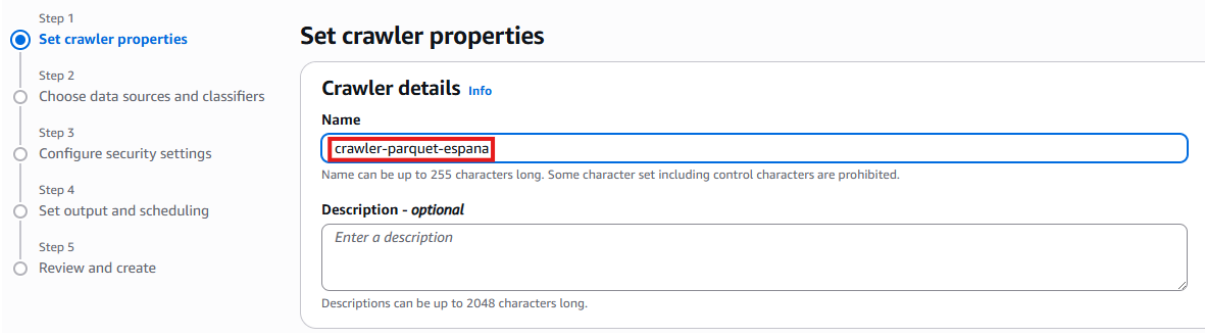
Para crear un “Crawler”, nos vamos a “AWS Glue”, para ello entramos en la barra de búsqueda y ponemos esto:

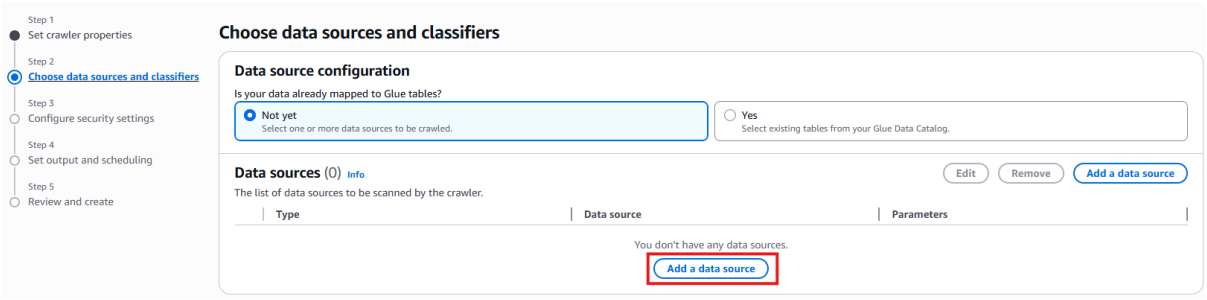


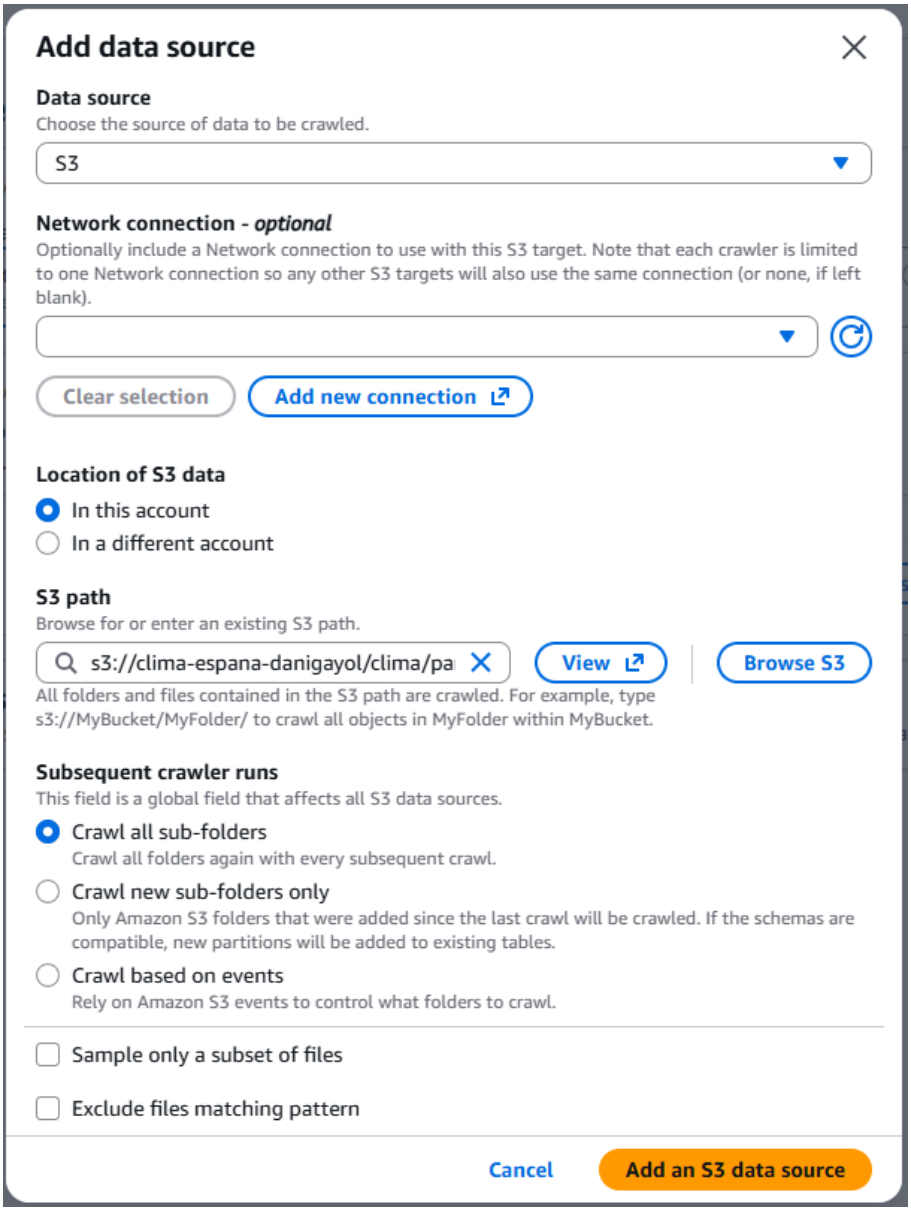
Una vez dentro, en el menú de la izquierda, entramos en “Crawlers”

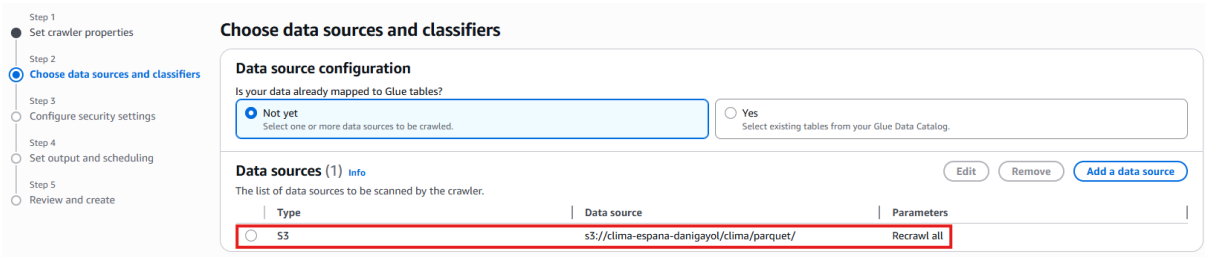


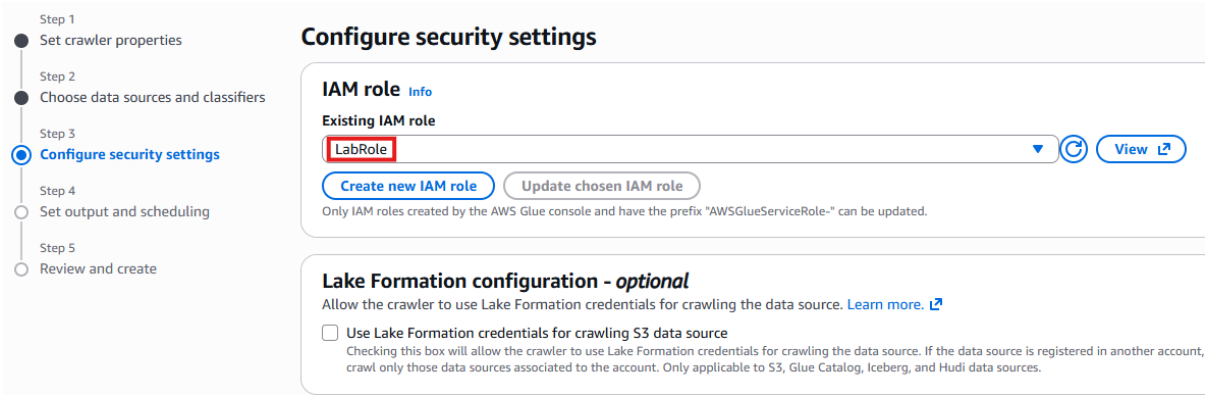
Una vez dentro, le damos al botón de "Create Crawler" y lo configuramos de la siguiente manera:

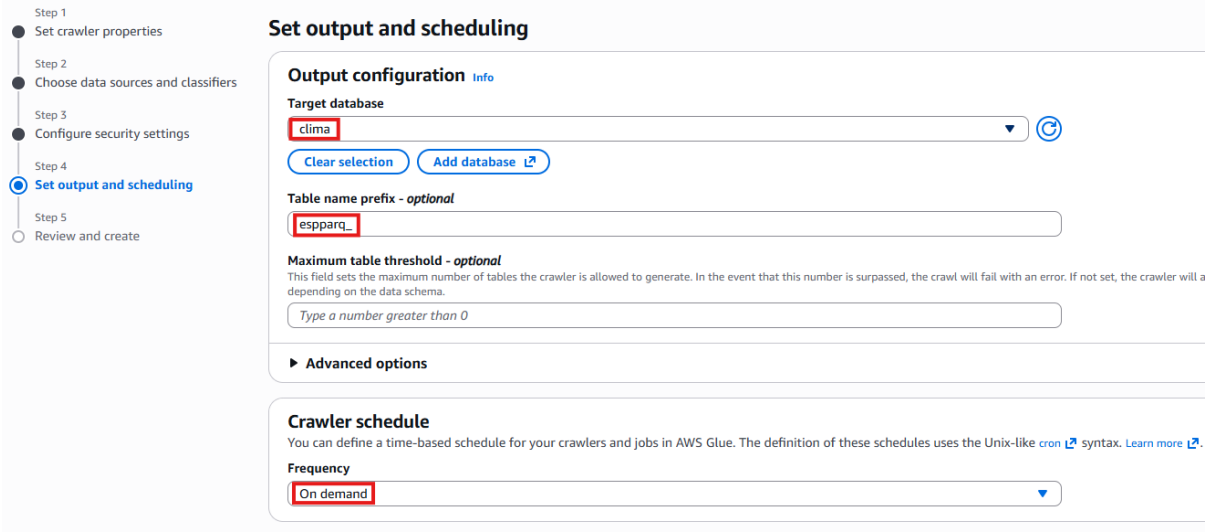



Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Choose data sources and classifiers
Data source configuration
Is your data already mapped to Glue tables?
Not yet
Select one or more data sources to be crawled.
Yes
Select existing tables from your Glue Data Catalog.
Data sources (0) Info
The list of data sources to be scanned by the crawler.
Edit
Remove
Add a data source
Type
Data source
Parameters
You don't have any data sources.
Add a data source

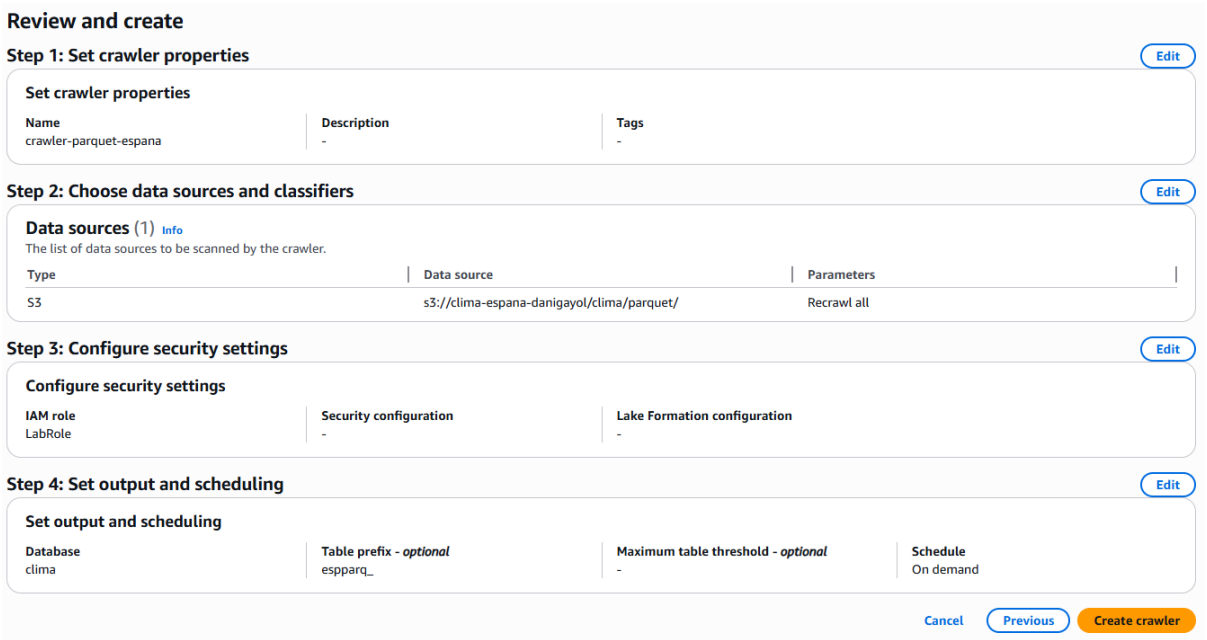

Add data source
Data source
Choose the source of data to be crawled.
S3
Network connection - optional
Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).
Clear selection
Add new connection
Location of S3 data
In this account
In a different account
S3 path
Browse for or enter an existing S3 path.
s3://clima-espana-danigayol/clima/pa
View
Browse S3
All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.
Subsequent crawler runs
This field is a global field that affects all S3 data sources.
Crawl all sub-folders
Crawl all folders again with every subsequent crawl.
Crawl new sub-folders only
Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.
Crawl based on events
Rely on Amazon S3 events to control what folders to crawl.
Sample only a subset of files
Exclude files matching pattern
Cancel
Add an S3 data source


Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Choose data sources and classifiers
Data source configuration
Is your data already mapped to Glue tables?
Not yet
Select one or more data sources to be crawled.
Yes
Select existing tables from your Glue Data Catalog.
Data sources (1) Info
The list of data sources to be scanned by the crawler.
Edit
Remove
Add a data source
Type
Data source
Parameters
S3
s3://clima-espana-danigayol/clima/parquet/
Recrawl all



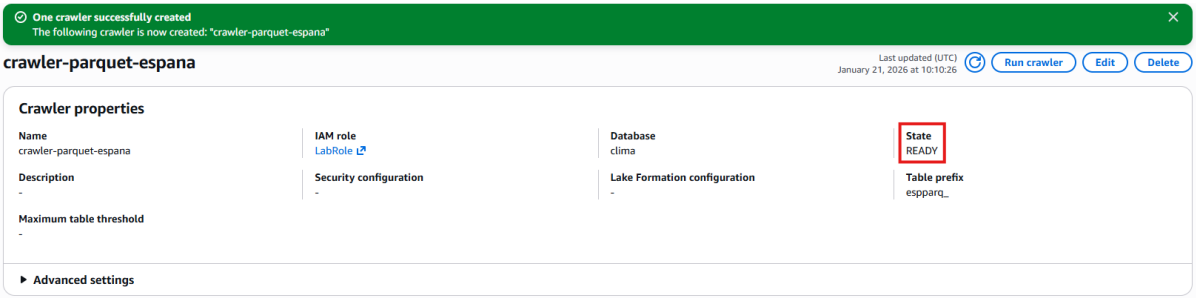


Finalmente, nos quedaría de la siguiente manera:



2.) Guarda el rastreador, pero no lo ejecutes.

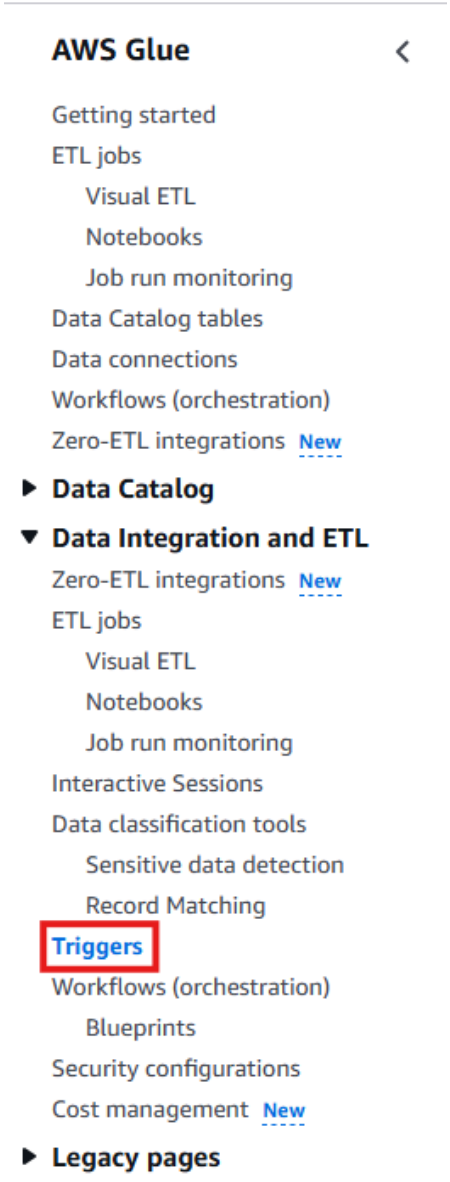
Ahora creamos el “crawler” pero NO lo ejecutamos



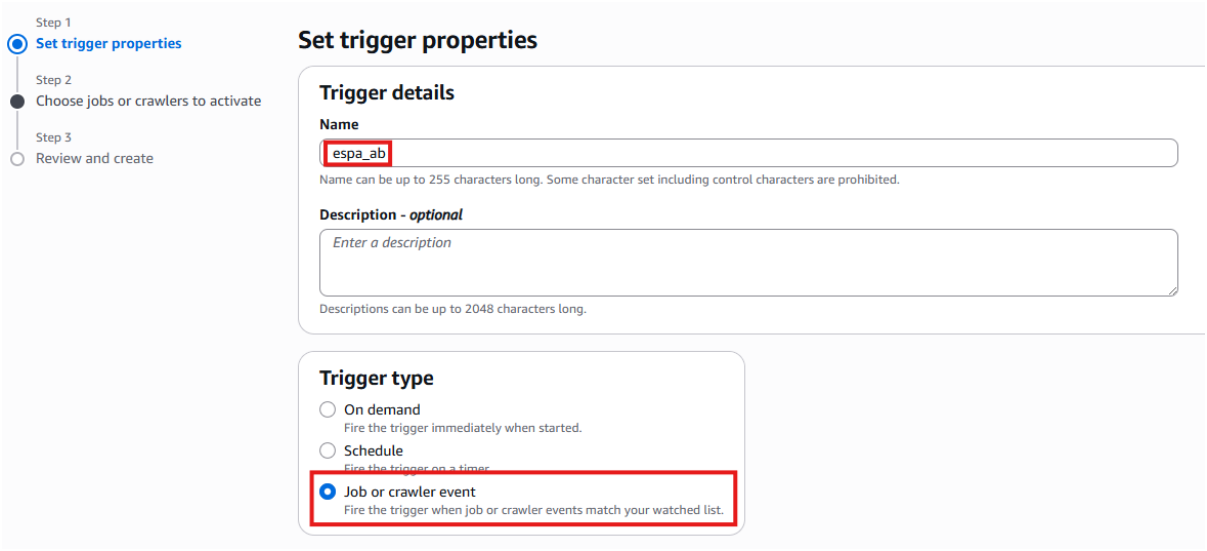
Apartado D

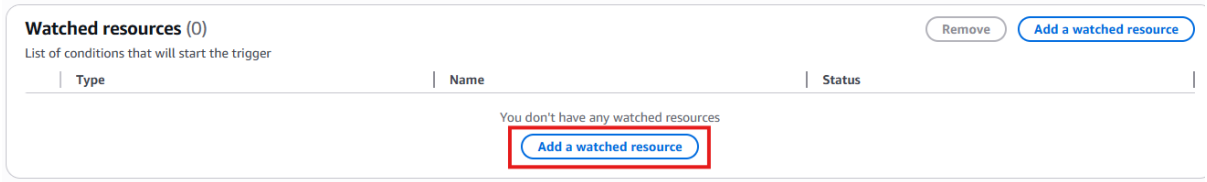
1.) Crea un disparador (trigger) -puedes llamarlo espa_ab - que después de finalizado el crawler del apartado A lance el trabajo del apartado B.

Dentro de "AWS Glue", en el menú de la izquierda, nos vamos al apartado de "Triggers"



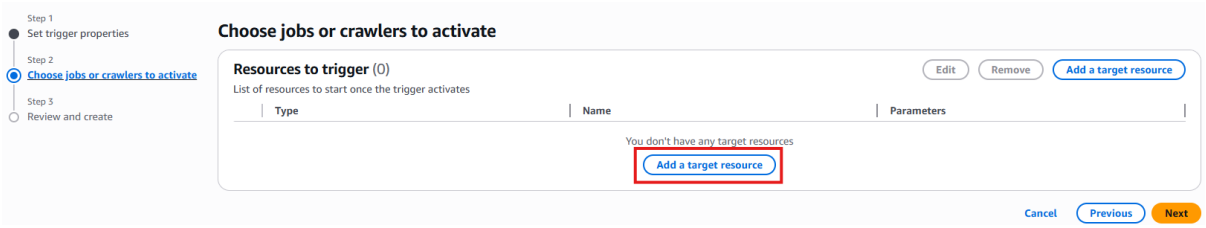
Una vez dentro de él, le damos al botón de "Add trigger" y lo configuramos de la siguiente manera:


Step 1
Set trigger properties
Step 2
Choose jobs or crawlers to activate
Step 3
Review and create
Set trigger properties
Trigger details
Name
espa_ab
Name can be up to 255 characters long. Some character set including control characters are prohibited.
Description - optional
Enter a description
Descriptions can be up to 2048 characters long.
Trigger type
On demand
Fire the trigger immediately when started.
Schedule
Job or crawler event
Fire the trigger when job or crawler events match your watched list.

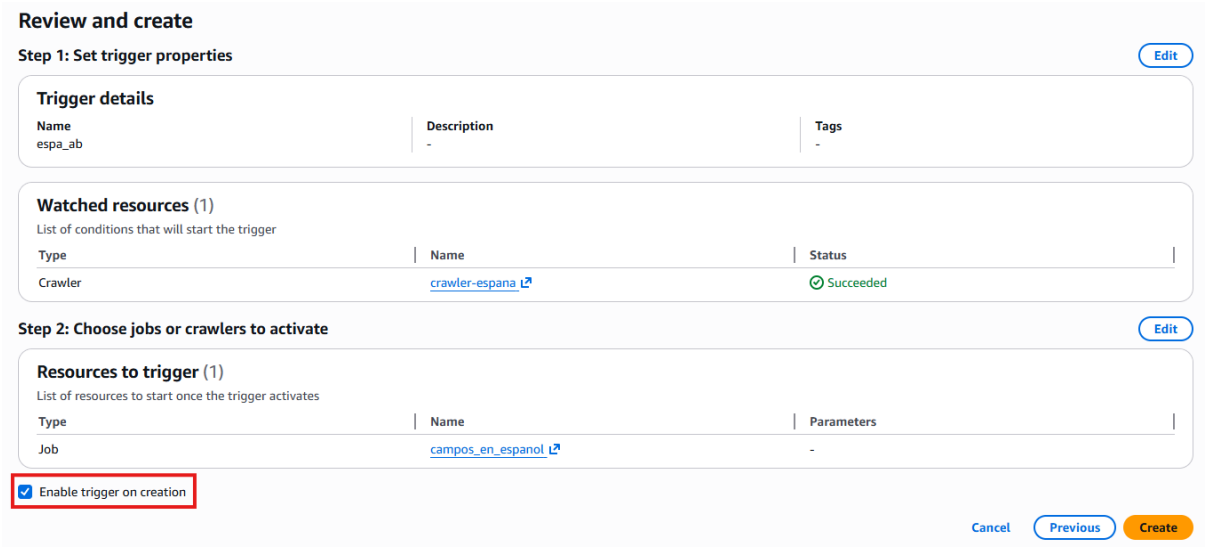

Watched resources (0)
List of conditions that will start the trigger
Remove
Add a watched resource
Type
Name
Status
You don't have any watched resources
Add a watched resource

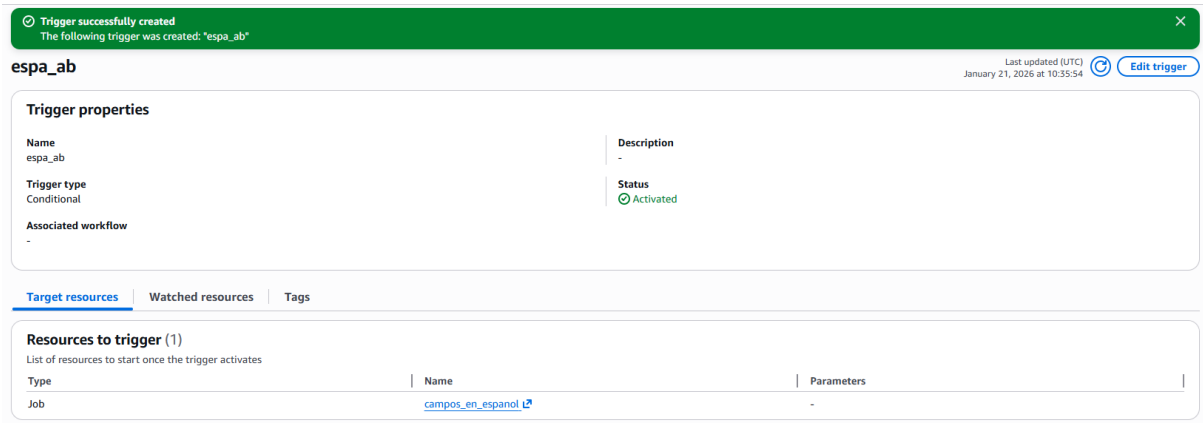

Add condition
Resource type
Crawler
The type of resource to apply this condition to
Crawler name
crawler-espana
Crawler to apply this condition to
Status
Succeeded
The status transition that the trigger listens for
Cancel
Add


Watched resources (1)
List of conditions that will start the trigger
Remove
Add a watched resource
Type
Name
Status
Crawler
crawler-espana
Succeeded



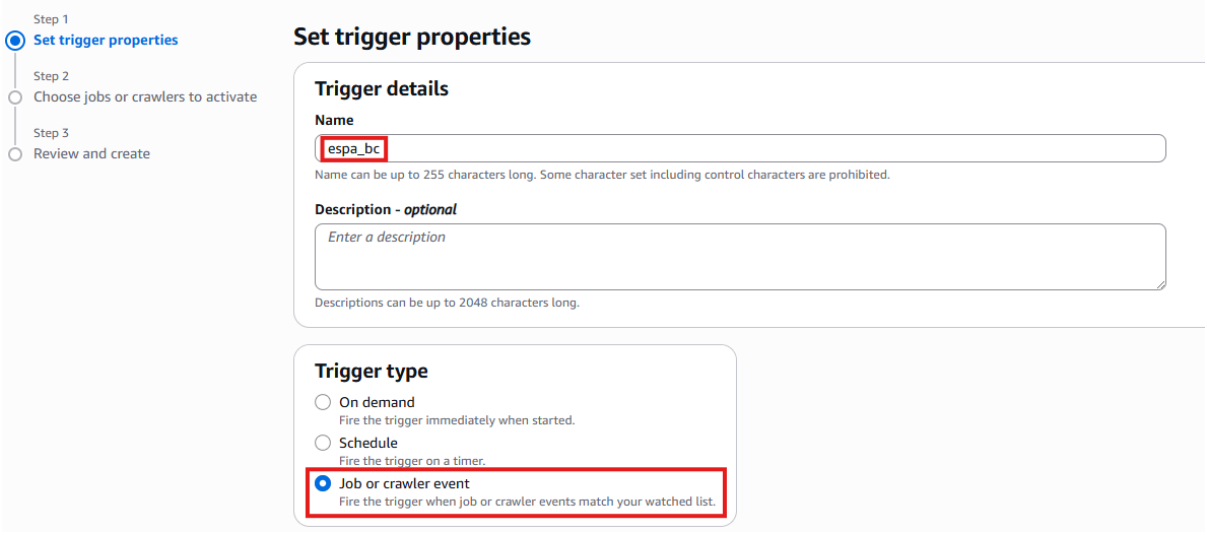
Nos tendría que quedar configurado de la siguiente manera:

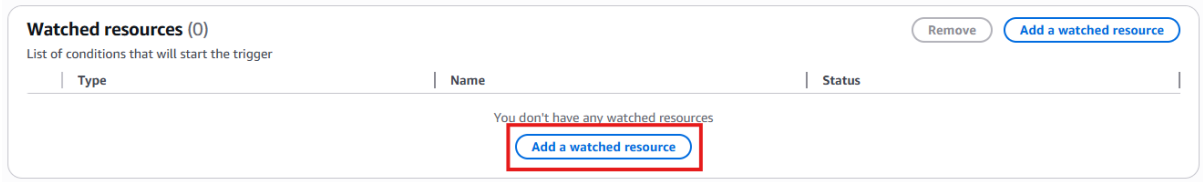


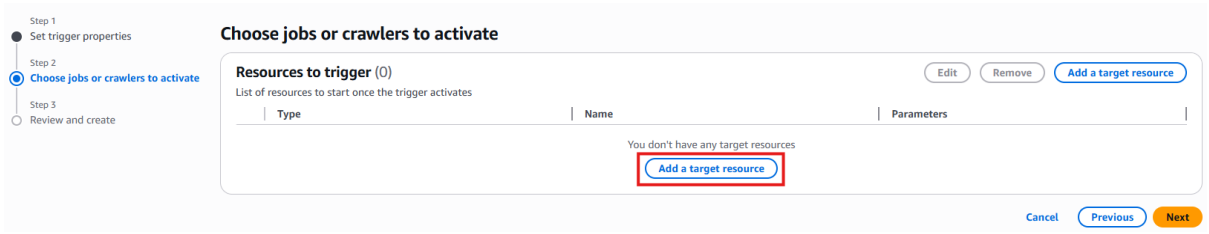


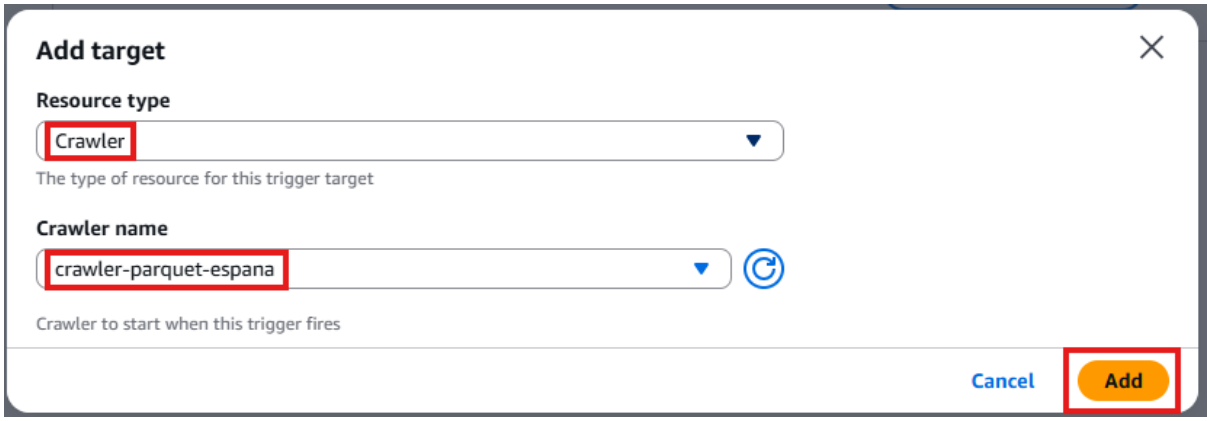
2.) Crea un disparador (trigger) -puedes llamarlo espa_bc - que después de finalizado el trabajo del apartado B lance el trabao del apartado C.

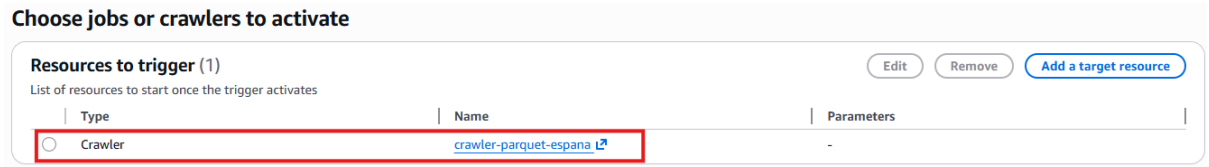
Ahora vamos a crear otro "trigger" con la siguiente configuración:



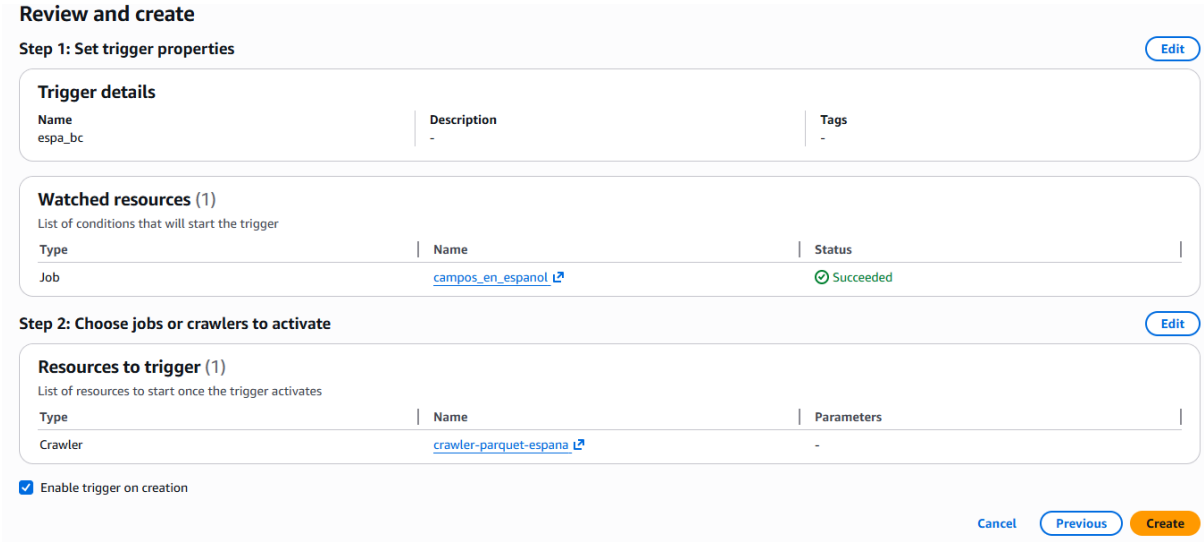


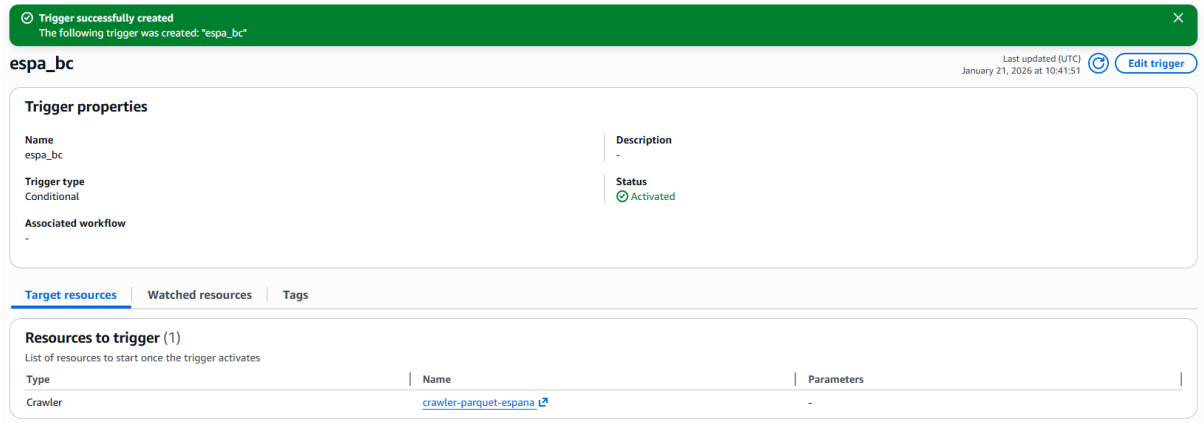






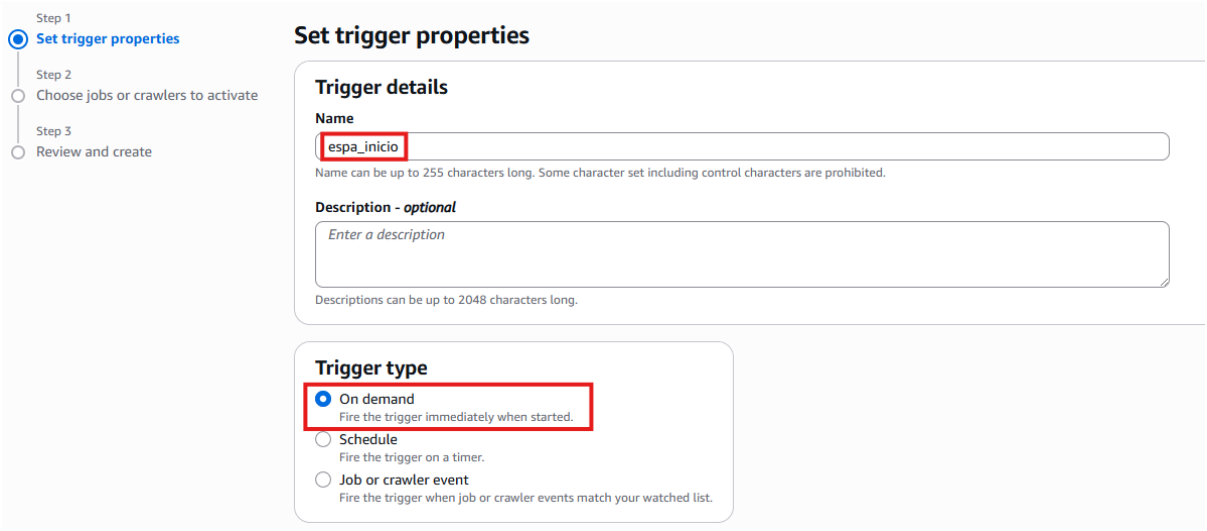
Finalmente, nos tendría que quedar configurado de la siguiente manera:

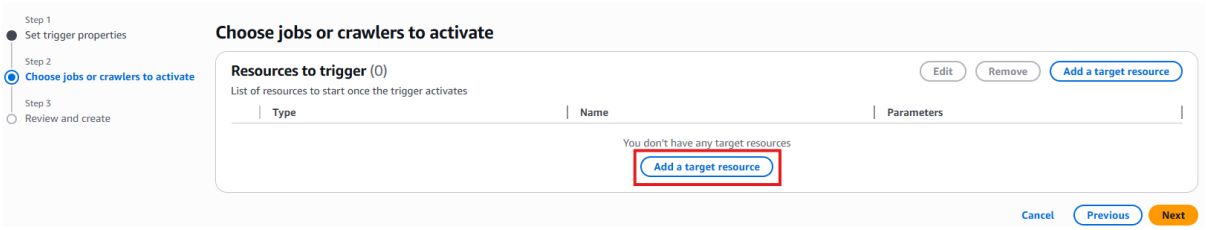


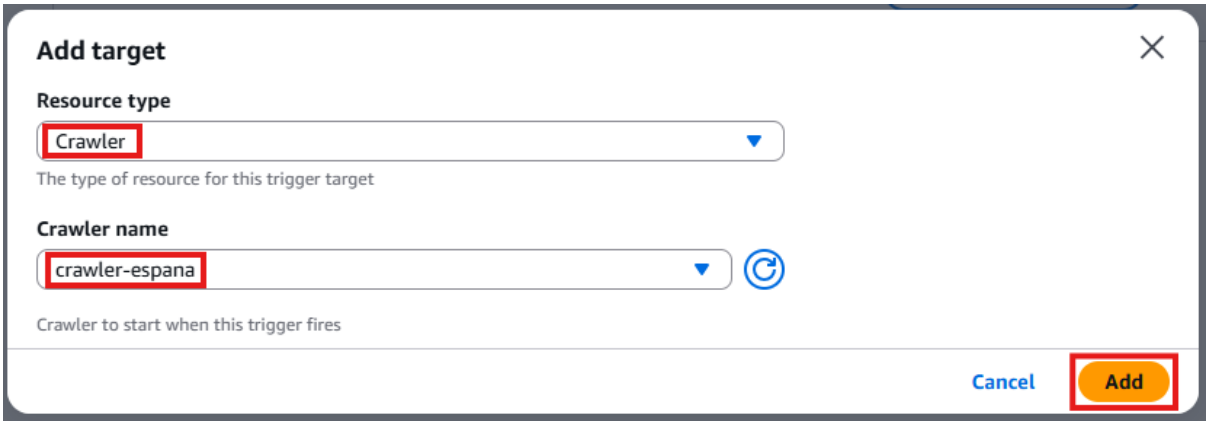


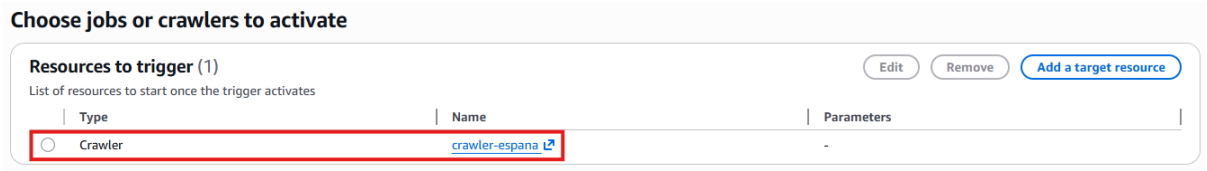
3.) Finalmente hemos de crear un triggerr bajo demanda que nos arranque el crawler inicial (en nuestro caso el del apartado A)

Volvemos a darle al botón de “add trigger” para crear el último trigger, y configurarlo de la siguiente manera:

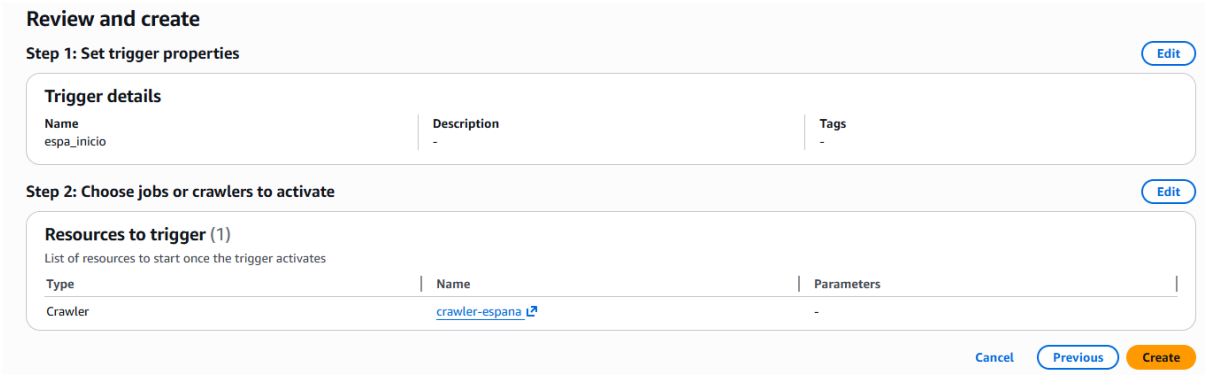


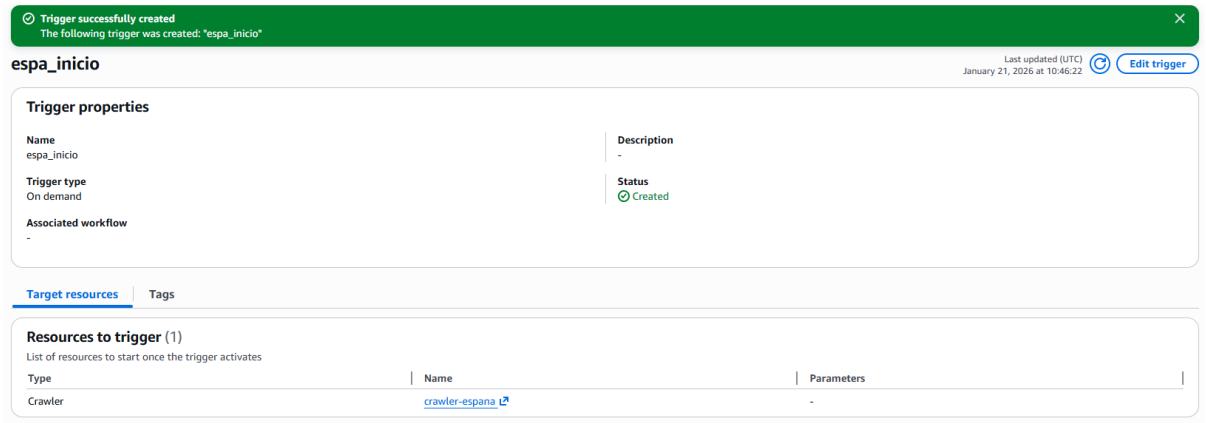






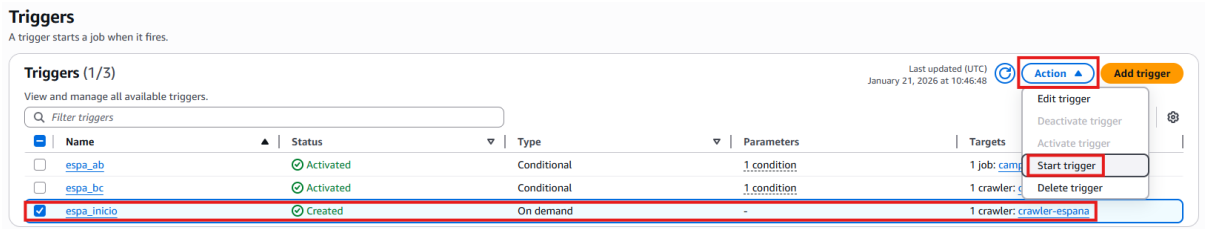
Nos tiene que quedar configurado de la siguiente manera:

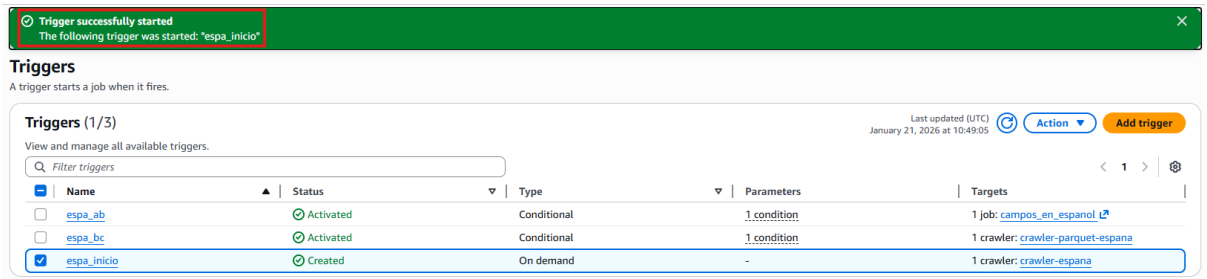




4.) Arranca este manualmente este último disparador.

Para arrancarlo manualmente, seleccionamos el "trigger" y le damos a "start trigger" en el manú de "Action"





Ahora para verificar que se hizo correctamente, nos vamos al menú de la izquierda y entramos en "Crawlers", una vez dentro, nos aparecerá lo siguiente:

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (3) Info

View and manage all available crawlers.

Last updated (UTC)
January 21, 2026 at 10:58:39

Action | Run | Create crawler

Filter crawlers

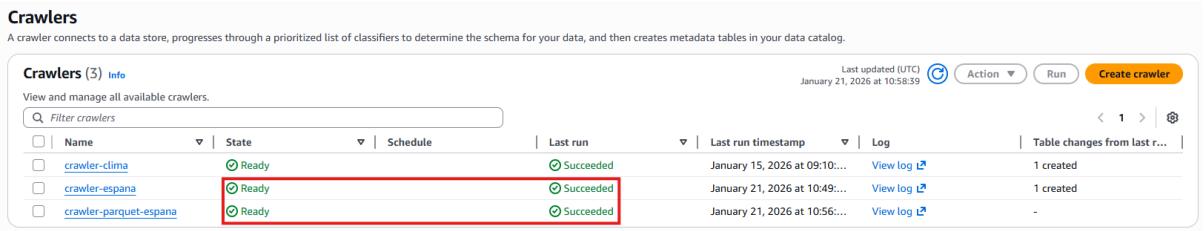
1

Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last r...
crawler-clima	Ready		Succeeded	January 15, 2026 at 09:10:...	View log	1 created
crawler-espana	Ready		Succeeded	January 21, 2026 at 10:49:...	View log	1 created
crawler-parquet-espana	Ready		Succeeded	January 21, 2026 at 10:56:...	View log	-

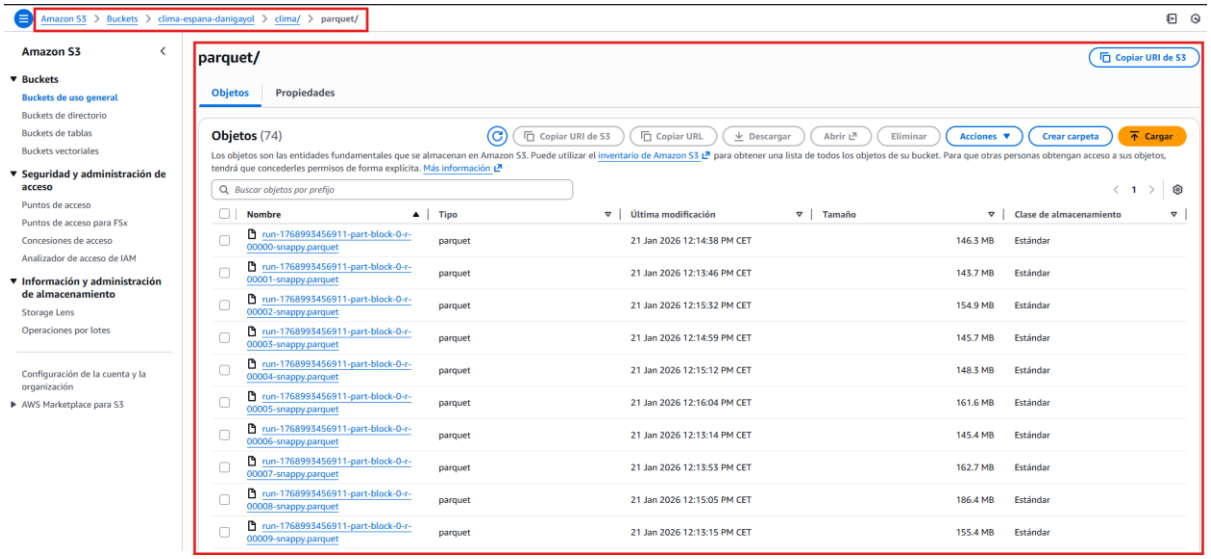
Apartado E

1.) Muestra los archivos creados.

Como hicimos en el ejercicio anterior, entramos en el apartado de “Crawlers” y nos tiene que aparecer lo siguiente:

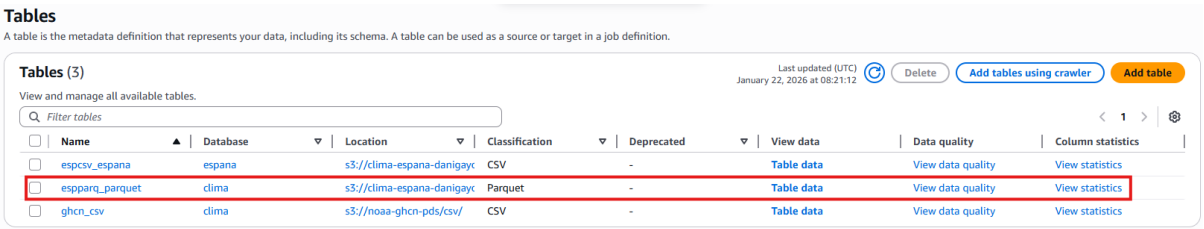


Si queremos comprobar los archivos del bucket en “S3”, no dirigimos a la siguiente ruta: “Amazon S3/Buckets/clima-espana-danigayol/clima/parquet”

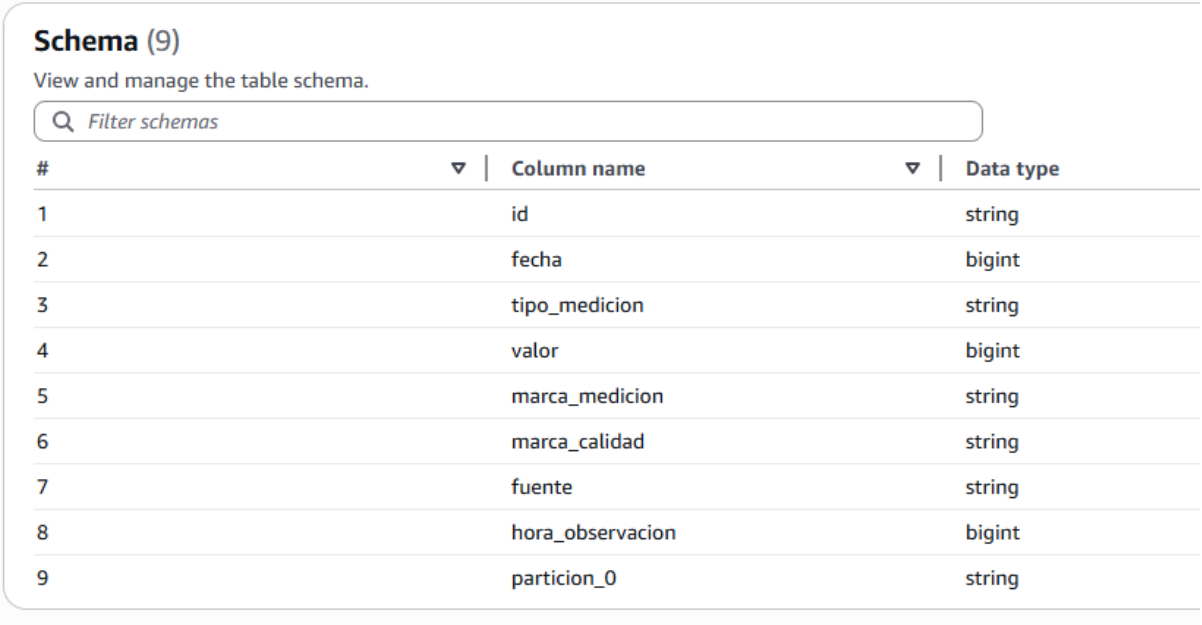


2.) Muestra las tablas y campos creados.

Ahora para mostrar la tabla creada, nos dirigimos de nuevo a “AWS Glue” y en el menú de la izquierda, entramos en donde pone “Tables”, una vez dentro, nos aparecerán ahi las tablas que existen



A continuación, entramos en la tabla “espparq_parquet” y ahí nos aparecerán los campos cambiados al español de la tabla:



Schema (9)

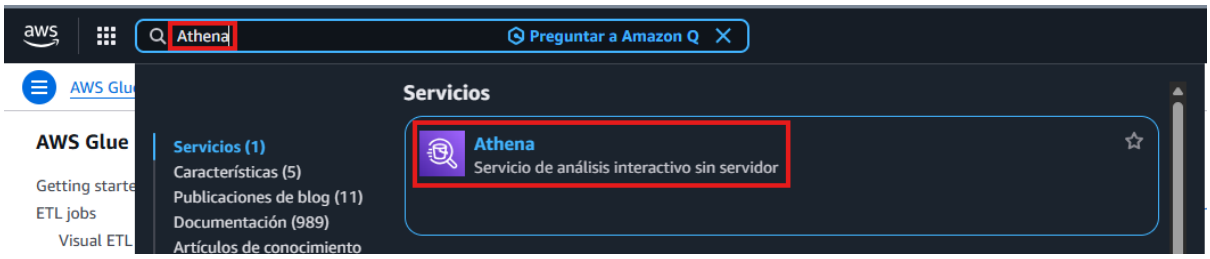
View and manage the table schema.

Filter schemas

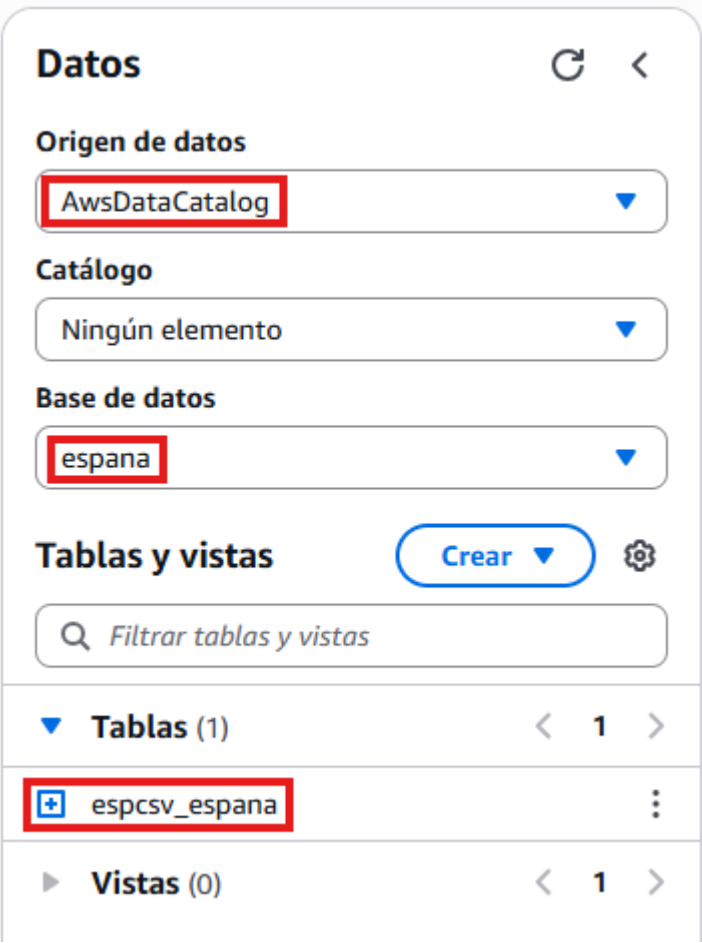
#	Column name	Data type
1	id	string
2	fecha	bigint
3	tipo_medicion	string
4	valor	bigint
5	marca_medicion	string
6	marca_calidad	string
7	fuente	string
8	hora_observacion	bigint
9	particion_0	string

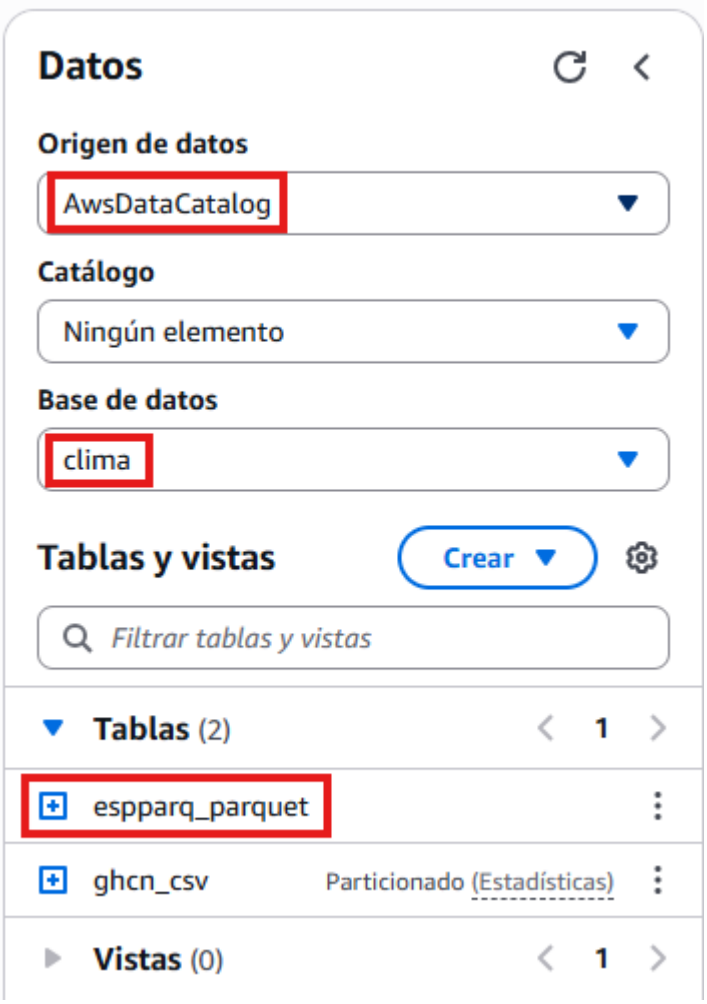
Apartado F

Para realizar las consultas, vamos a entrar en “Athena”, para ello, buscamos en la barra de búsqueda “Athena” y entramos



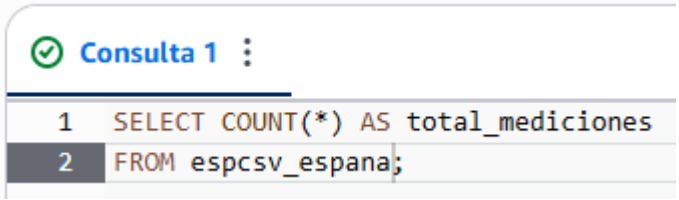
Una vez dentro, completamos los datos necesarios para hacer las consultas que queramos

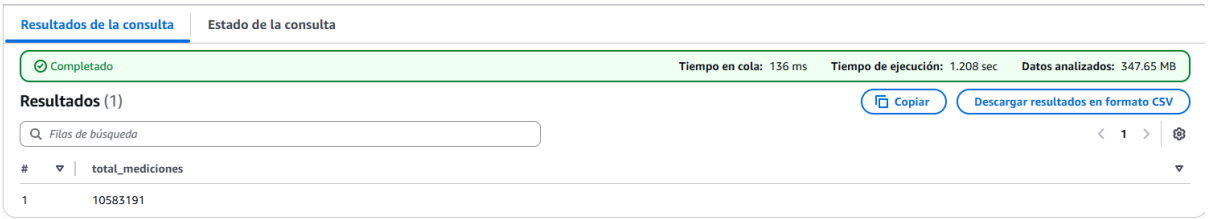


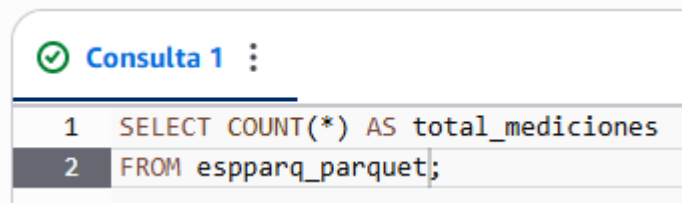


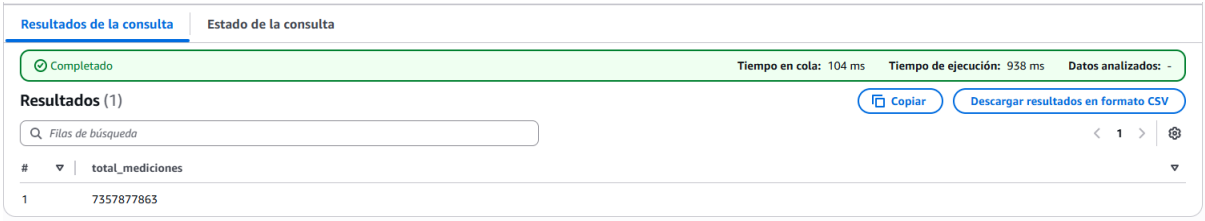
1.) ¿Cuántas mediciones tenemos de España?

Tabla

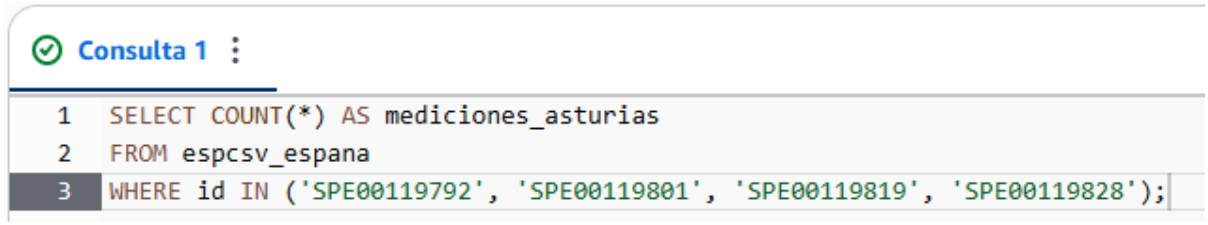


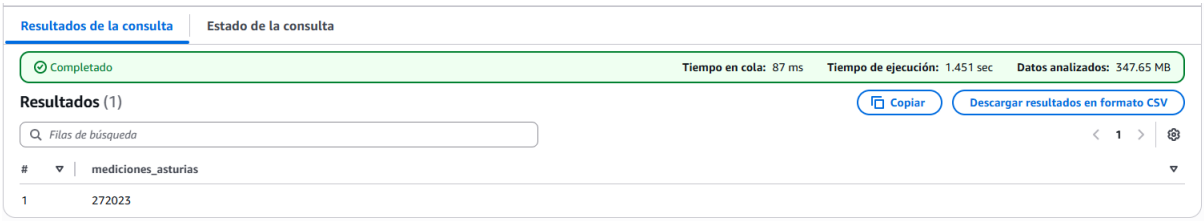


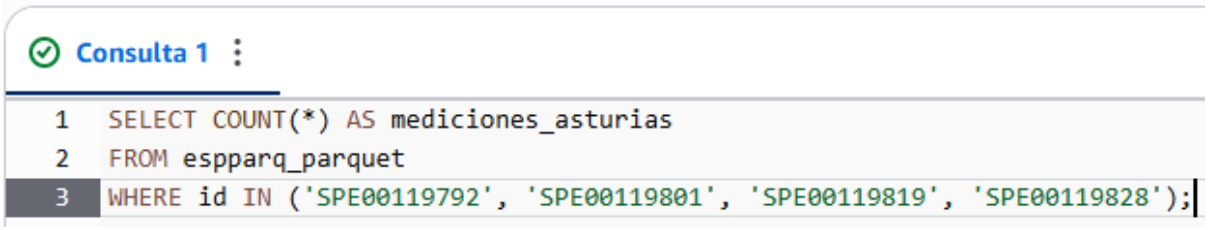


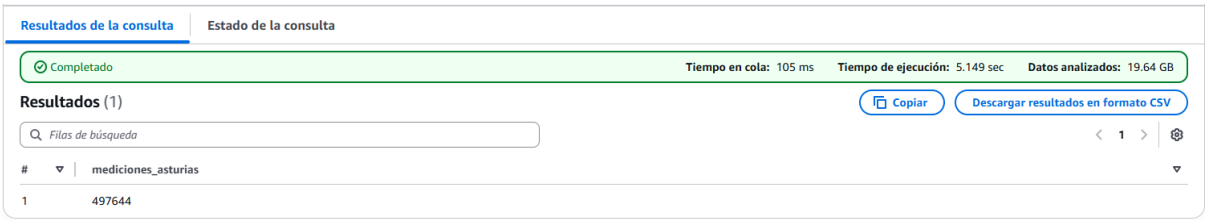


2.) Sabiendo los códigos de las 4 estaciones de Asturias ¿Cuántas mediciones tenemos de Asturias?

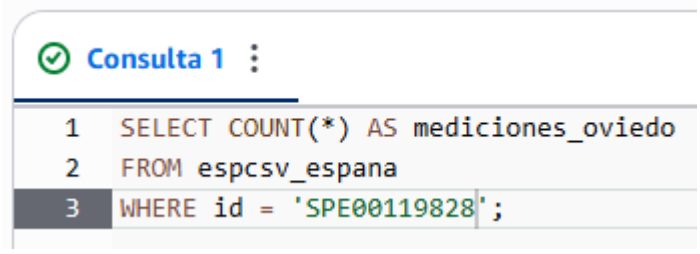


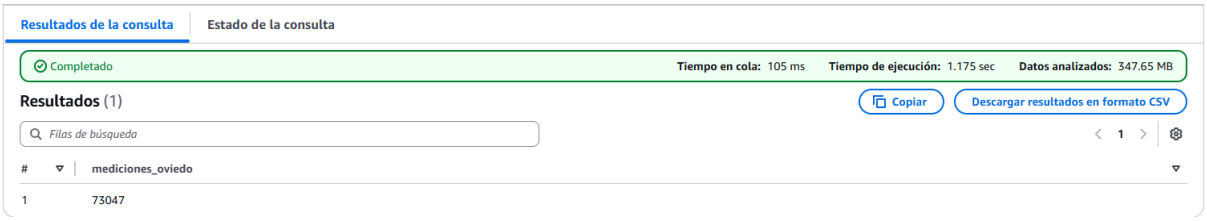


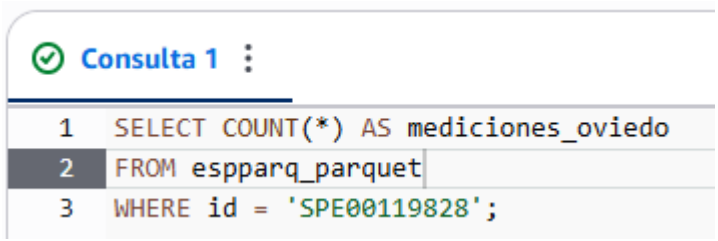


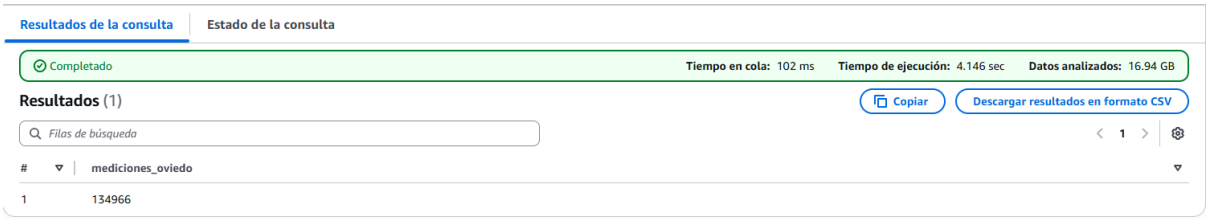


3.) ¿Cuántas mediciones tenemos de Oviedo?

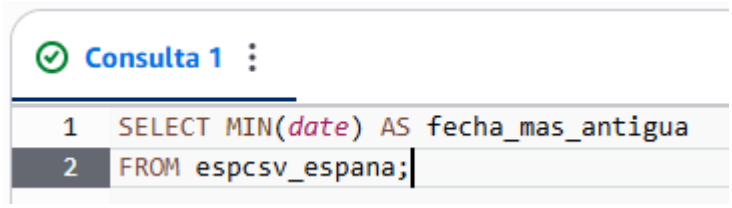


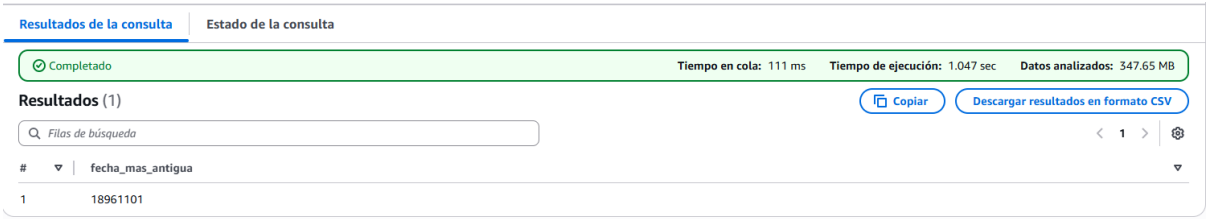


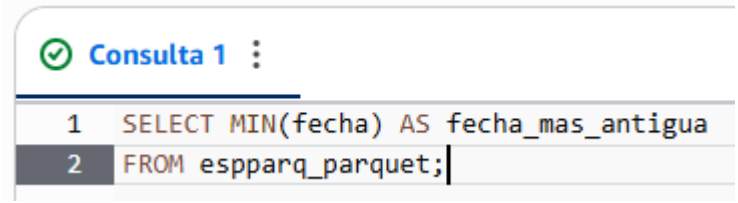


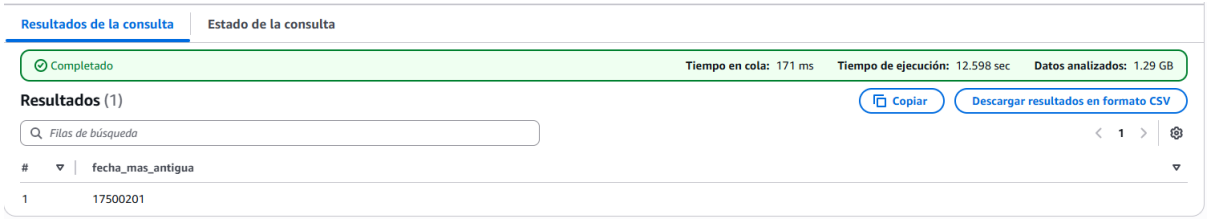


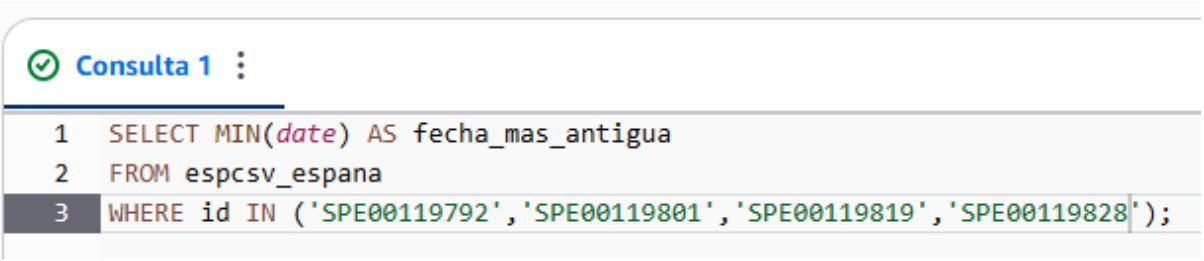
4.) ¿Cuál es la medición más antigua de España, Asturias y Oviedo?

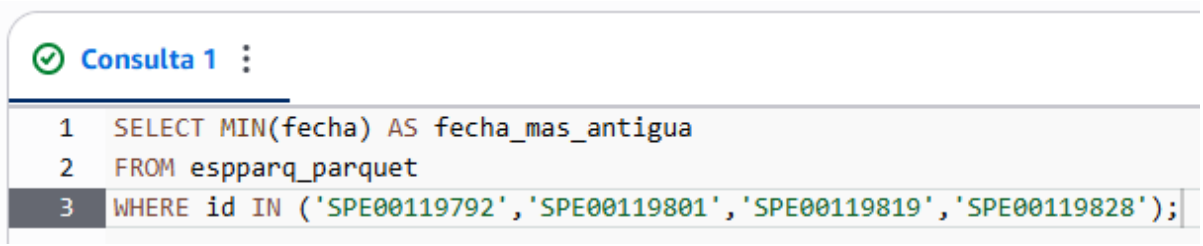


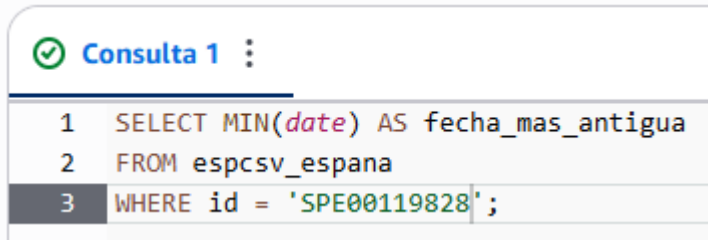


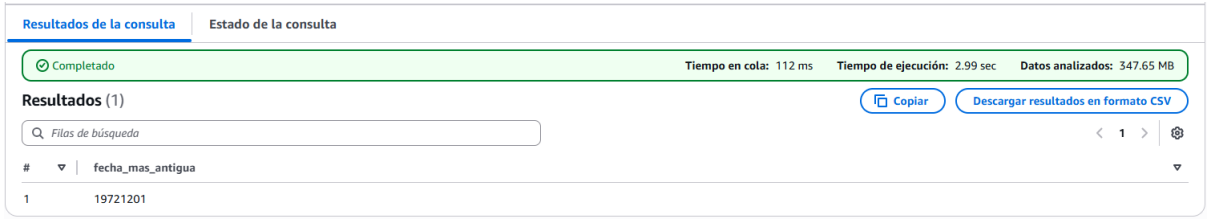


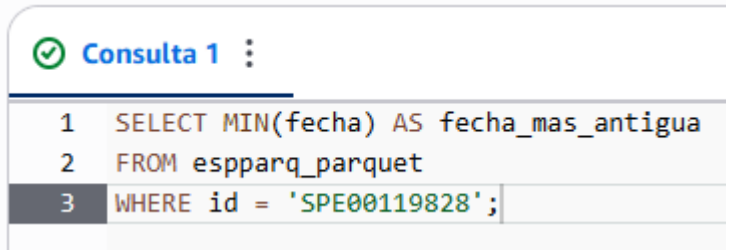

Resultados de la consulta
Estado de la consulta
Completado
Tiempo en cola: 171 ms
Tiempo de ejecución: 12.598 sec
Datos analizados: 1.29 GB
Resultados (1)
Copiar
Descargar resultados en formato CSV
Filas de búsqueda
1
fecha_mas_antigua
1 17500201

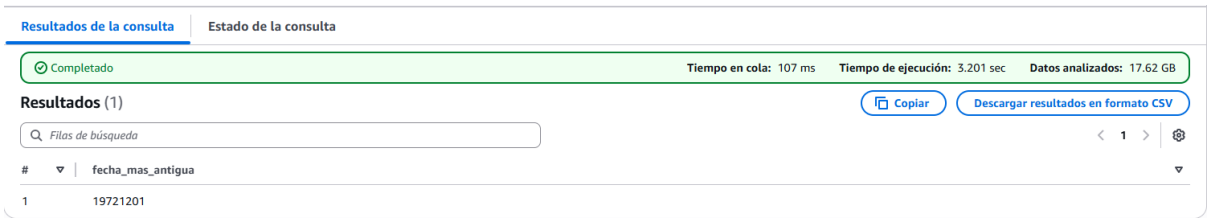

Consulta 1
1 SELECT MIN(date) AS fecha_mas_antigua
2 FROM espcsv_espana
3 WHERE id IN ('SPE00119792','SPE00119801','SPE00119819','SPE00119828');


Consulta 1
1 SELECT MIN(fecha) AS fecha_mas_antigua
2 FROM espparq_parquet
3 WHERE id IN ('SPE00119792','SPE00119801','SPE00119819','SPE00119828');


Consulta 1
1 SELECT MIN(date) AS fecha_mas_antigua
2 FROM espcsv_espana
3 WHERE id = 'SPE00119828';


Resultados de la consulta
Estado de la consulta
Completado
Tiempo en cola: 112 ms
Tiempo de ejecución: 2.99 sec
Datos analizados: 347.65 MB
Resultados (1)
Copiar
Descargar resultados en formato CSV
Filas de búsqueda
1
fecha_mas_antigua
1 19721201


Consulta 1
1 SELECT MIN(fecha) AS fecha_mas_antigua
2 FROM espparq_parquet
3 WHERE id = 'SPE00119828';


Resultados de la consulta
Estado de la consulta
Completado
Tiempo en cola: 107 ms
Tiempo de ejecución: 3.201 sec
Datos analizados: 17.62 GB
Resultados (1)
Copiar
Descargar resultados en formato CSV
Filas de búsqueda
1
fecha_mas_antigua
1 19721201

5.) Haz una tabla comparativa con los tiempos de ejecución de las consultas sobre las tres diferentes tablas (las de la práctica anterior y las dos de esta práctica) ¿Cuáles han sido las más veloces?

Consulta	CSV original	CSV espcsv	Parquet espparq
1	13,3 segundos	1,2 segundos	0,9 segundos
2	13,4 segundos	1,4 segundos	5,1 segundos
3	13,1 segundos	1,1 segundos	4,1 segundos
4	14,8 segundos	1,0 segundos	12,5 segundos