PR_08.1 Dani Gayol Rodríguez

Apartado A

1.) Desde Aws CLI explora el contenido del bucket s3://noaa-ghcn-pds/csv/.

Para explorar el contenido de ese bucket, vamos a usar el siguiente comando:

```
C:\Users\Mañana\.aws>aws s3 ls s3://noaa-ghcn-pds/csv/
                           PRE by_station/
                           PRE by_year/
```

2.) Descarga uno cualquiera de los archivos que contiene en cada una de sus carpetas y muestra las primeras líneas de ellos.

Para la primera carpeta, la de “by_year” voy a mostrar las primeras lineas del archivo “1900.csv” utilizando este comando (al estar en “Windows CMD” el comando “head” no existe asique voy a usar “more”):

```
C:\Users\Mañana\.aws>aws s3 ls s3://noaa-ghcn-pds/csv/by_year/
2025-01-14 13:42:27   11608918 1750.csv
2026-01-14 18:27:25      25333 1763.csv
2026-01-14 18:27:21      25340 1764.csv
2026-01-14 18:27:25      25328 1765.csv
2026-01-14 18:27:29      25378 1766.csv
2026-01-14 18:27:25      25374 1767.csv
2026-01-14 18:27:22      25419 1768.csv
2026-01-14 18:27:29      25310 1769.csv
2026-01-14 18:27:29      25316 1770.csv
```

```
C:\Users\Mañana\.aws>aws s3 cp s3://noaa-ghcn-pds/csv/by_year/1900.csv .
download: s3://noaa-ghcn-pds/csv/by_year/1900.csv to .\1900.csv
```

```
C:\Users\Mañana\.aws>more 1900.csv
ID,DATE,ELEMENT,DATA_VALUE,M_FLAG,Q_FLAG,S_FLAG,OBS_TIME
UPM00033976,19000101,TMAX,0,,,r,
UPM00033976,19000101,TMIN,-55,,,r,
UPM00033976,19000101,PRCP,0,,,r,
UPM00033976,19000101,TAVG,-34,,,r,
UPM00033990,19000101,PRCP,0,,,I,
UPM00033991,19000101,PRCP,0,,,I,
UPM00033999,19000101,TMAX,19,,,E,
UPM00033999,19000101,TMIN,-20,,,E,
UPM00033999,19000101,PRCP,0,,,E,
```

Y ahora voy a hacer lo mismo para la carpeta de "by_station":

```
C:\Users\Mañana\.aws>aws s3 ls s3://noaa-ghcn-pds/csv/by_station/
2026-01-14 18:15:16      41673 ACW00011604.csv
2026-01-14 18:15:16     421621 ACW00011647.csv
2026-01-14 18:16:04    1998959 AE000041196.csv
2026-01-14 18:16:04    1597557 AEM00041194.csv
2026-01-14 18:16:04    1382749 AEM00041217.csv
2026-01-14 18:16:04     909878 AEM00041218.csv
2026-01-14 18:16:04     221370 AF000040930.csv
2026-01-14 18:16:04     450046 AFM00040938.csv
2026-01-14 18:16:04     671460 AFM00040948.csv
2026-01-14 18:16:04     507571 AFM00040990.csv
```

```
C:\Users\Mañana\.aws>aws s3 cp s3://noaa-ghcn-pds/csv/by_station/USW00094728.csv .
download: s3://noaa-ghcn-pds/csv/by_station/USW00094728.csv to .\USW00094728.csv
```

```
C:\Users\Mañana\.aws>more USW00094728.csv
ID,DATE,ELEMENT,DATA_VALUE,M_FLAG,Q_FLAG,S_FLAG,OBS_TIME
USW00094728,18690101,TMAX,-17,,,Z,
USW00094728,18690102,TMAX,-28,,,Z,
USW00094728,18690103,TMAX,17,,,Z,
USW00094728,18690104,TMAX,28,,,Z,
USW00094728,18690105,TMAX,61,,,Z,
USW00094728,18690106,TMAX,33,,,Z,
USW00094728,18690107,TMAX,89,,,Z,
USW00094728,18690108,TMAX,122,,,Z,
USW00094728,18690109,TMAX,89,,,Z,
USW00094728,18690110,TMAX,67,,,Z,
USW00094728,18690111,TMAX,6,,,Z,
USW00094728,18690112,TMAX,28,,,Z,
```

3.) ¿Qué contiene cada uno de los dos tipos de archivos?

En la carpeta de “by_year”, contiene todas las observaciones meteorológicas de todas las estaciones para un año concreto, cada archivo representa un año e incluye múltiples estaciones

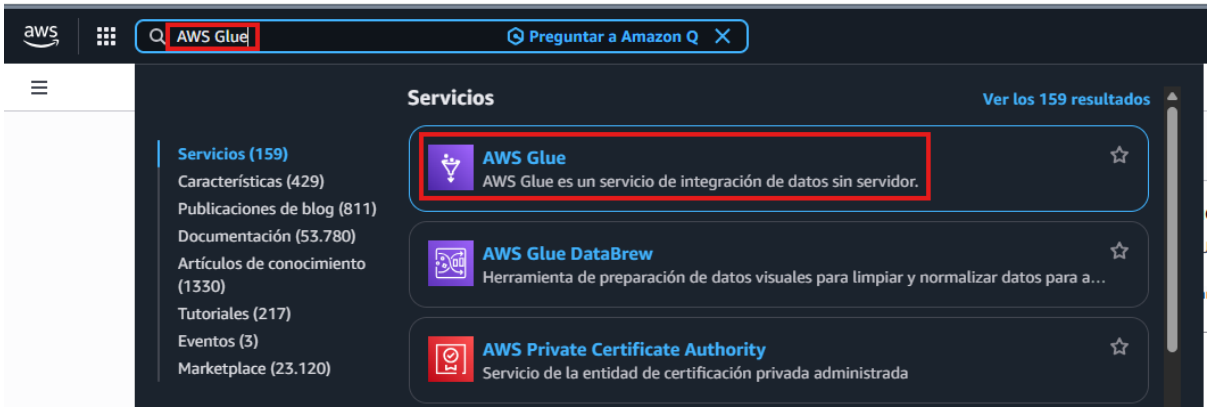
- ID = 11 character station identification code. Please see ghcnd-stations section below for an explantation
- YEAR/MONTH/DAY = 8 character date in YYYYMMDD format (e.g. 19860529 = May 29, 1986)
- ELEMENT = 4 character indicator of element type
- DATA VALUE = 5 character data value for ELEMENT
- M-FLAG = 1 character Measurement Flag
- Q-FLAG = 1 character Quality Flag
- S-FLAG = 1 character Source Flag
- OBS-TIME = 4-character time of observation in hour-minute format (i.e. 0700 =7:00 am)

En la carpeta de “by_station”, contiene todas las observaciones históricas de una sola estación, cada archivo representa una estación meteorológica e Incluye datos de muchos años

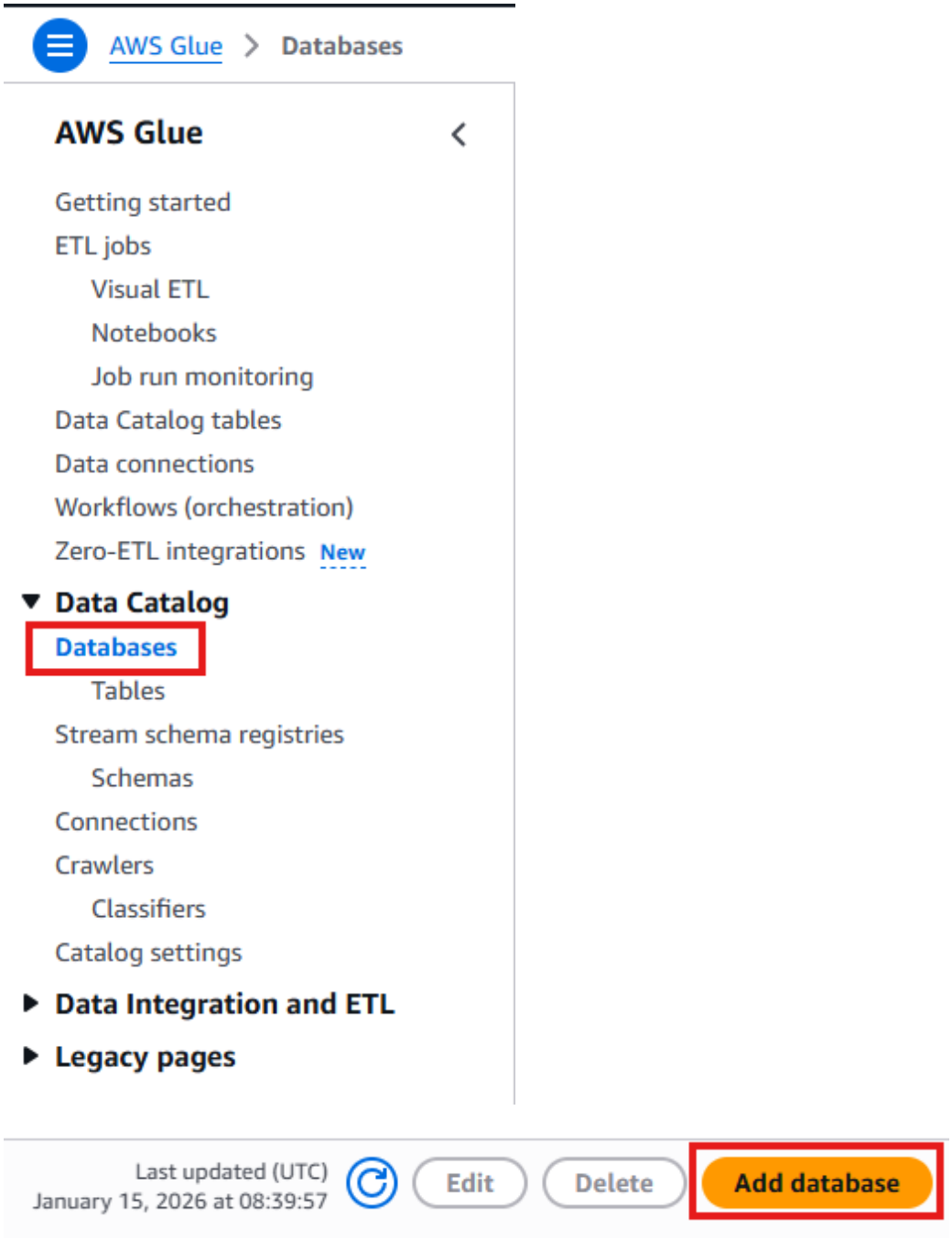
Apartado B

1.) Crea una base de datos en AWE GLUE llamada clima.

Primero tenemos que buscar en la barra de búsqueda de AWS y poner “AWS Glue”

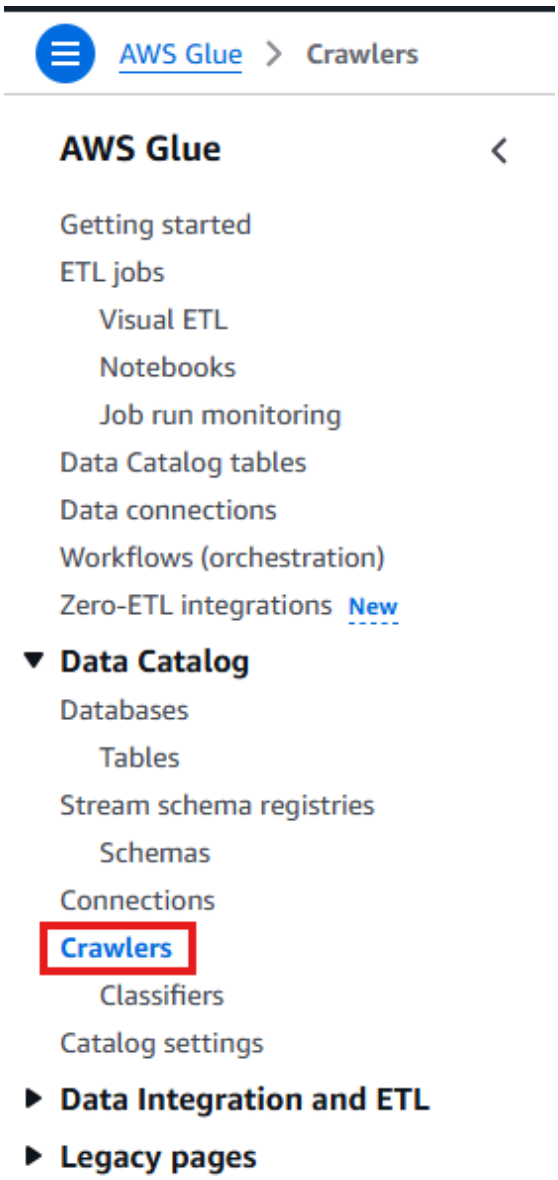


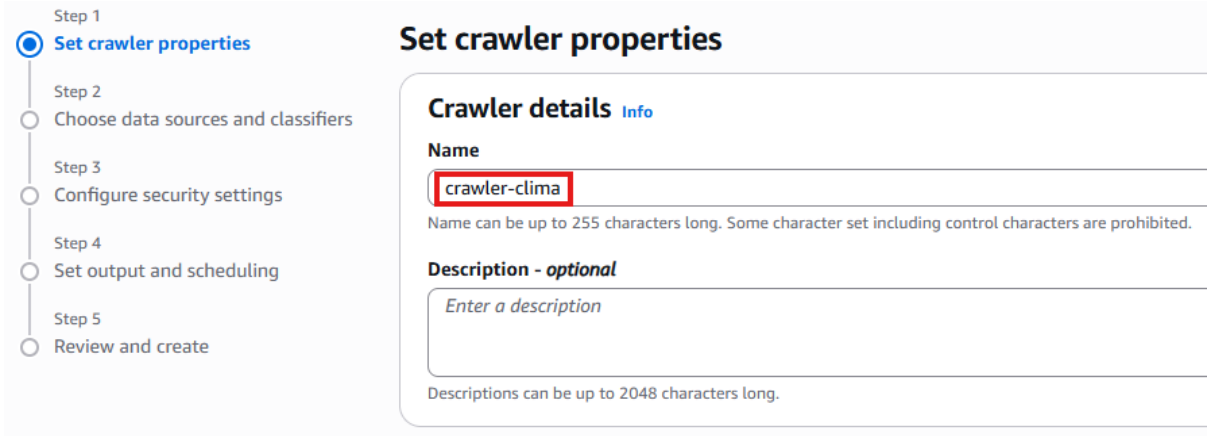
Una vez dentro nos vamos a donde pone “Databases” y una vez dentro le damos al botón de “Add database”

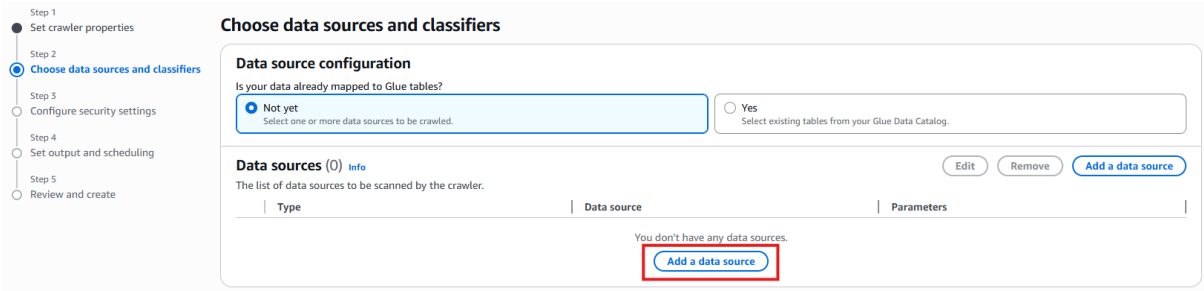

AWS Glue
Databases
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations
New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
Last updated (UTC)
January 15, 2026 at 08:39:57
Edit
Delete
Add database

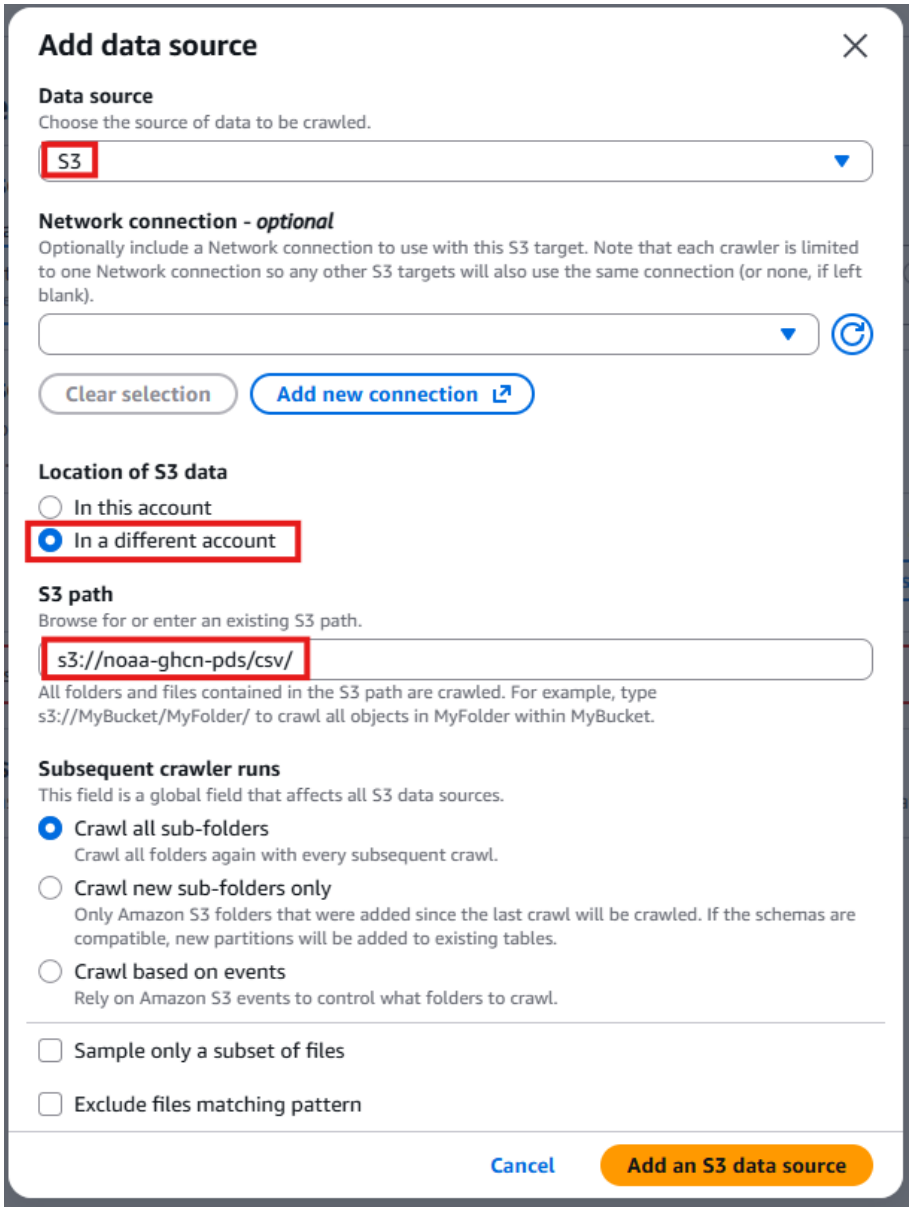
2.) Crea un Crawler AWS GLUE que nos explore el bucket del ejercicio anterior generando las tablas en la base de datos que acabas de crear.

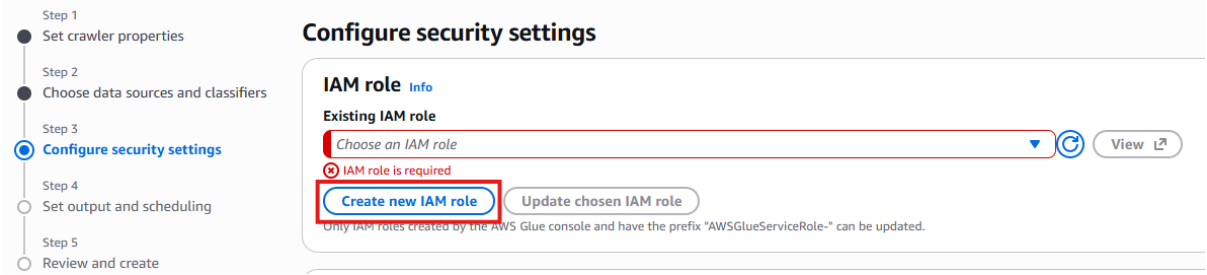
Ahora tenemos que ir a donde pone “Crawler” y darle al botón de “Crear Crawler”

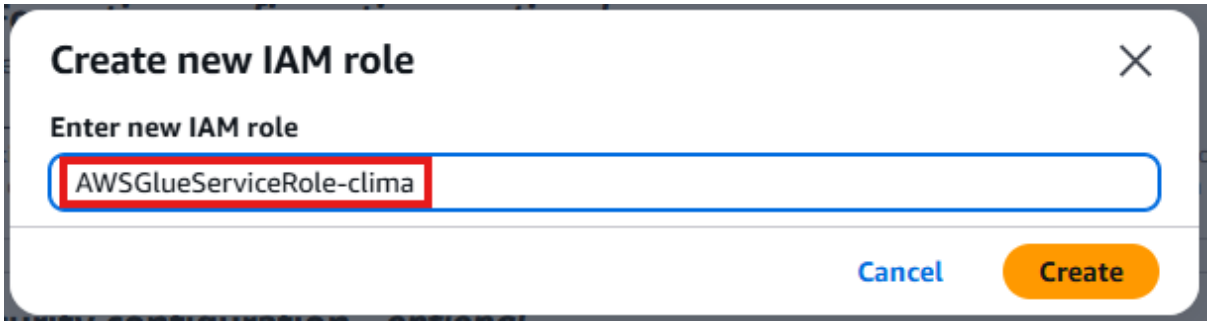

AWS Glue
Crawlers
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations
New
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages


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
Set crawler properties
Crawler details
Info
Name
crawler-clima
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.

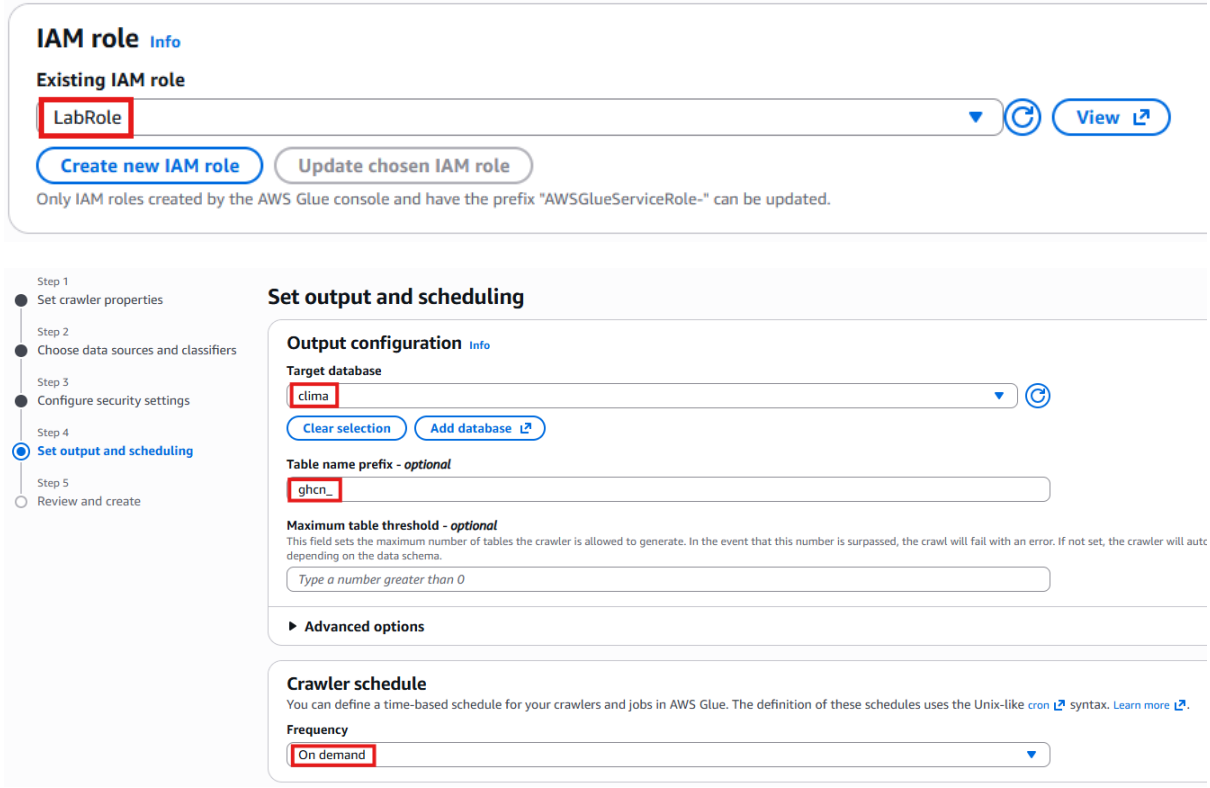

Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Choose data sources and classifiers
Data source configuration
Is your data already mapped to Glue tables?
Not yet
Select one or more data 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://noaa-ghcn-pds/csv/
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

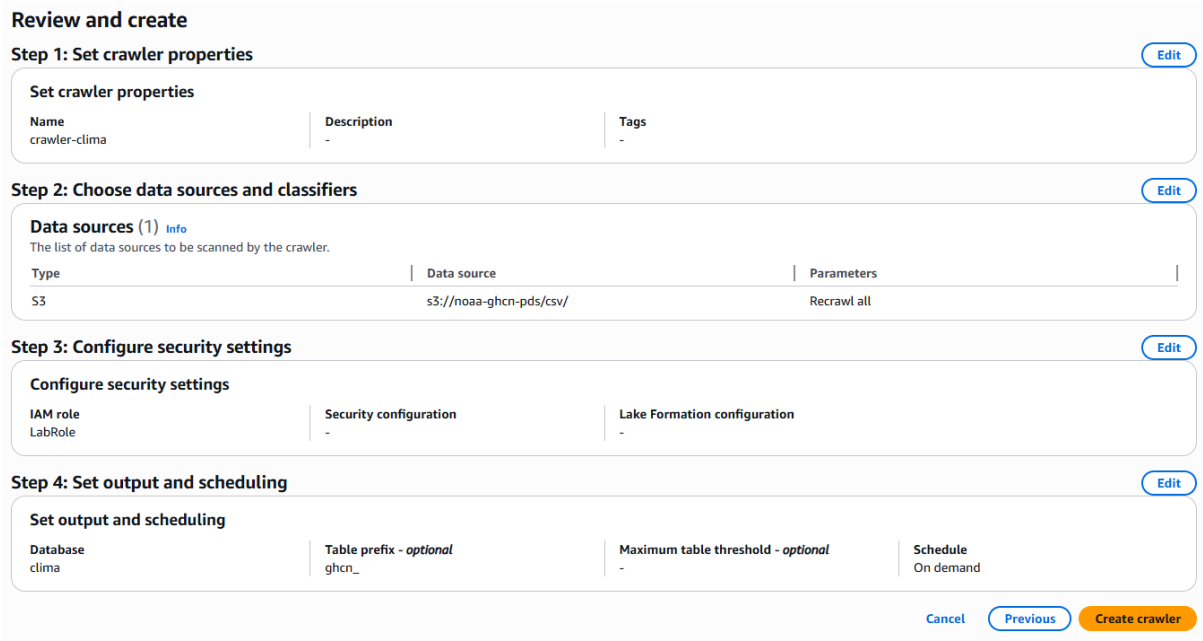

Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Configure security settings
IAM role Info
Existing IAM role
Choose an IAM role
View
IAM role is required
Create new IAM role
Update chosen IAM role
Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

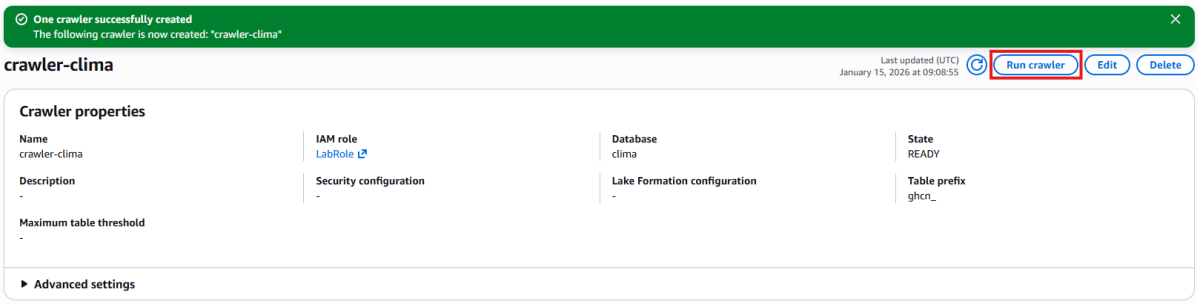


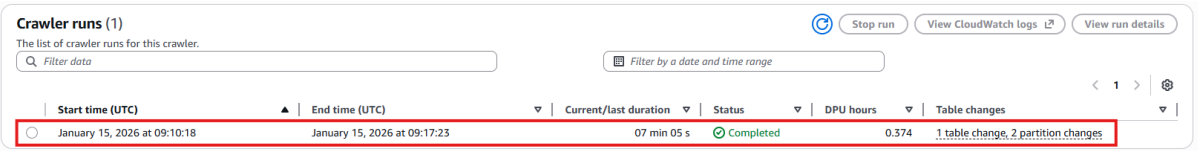
Nos va a dar un error al intentar crear el “rol” ya que no tenemos permisos, por lo tanto, vamos a seleccionar un “rol” ya existente



Finalmente, la configuración nos quedaria de la siguiente manera:

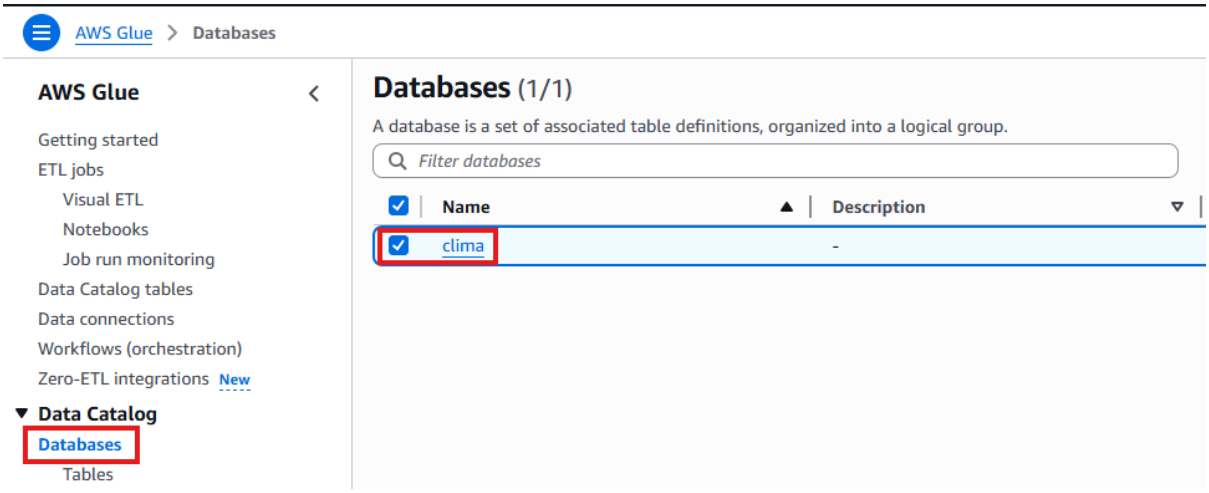




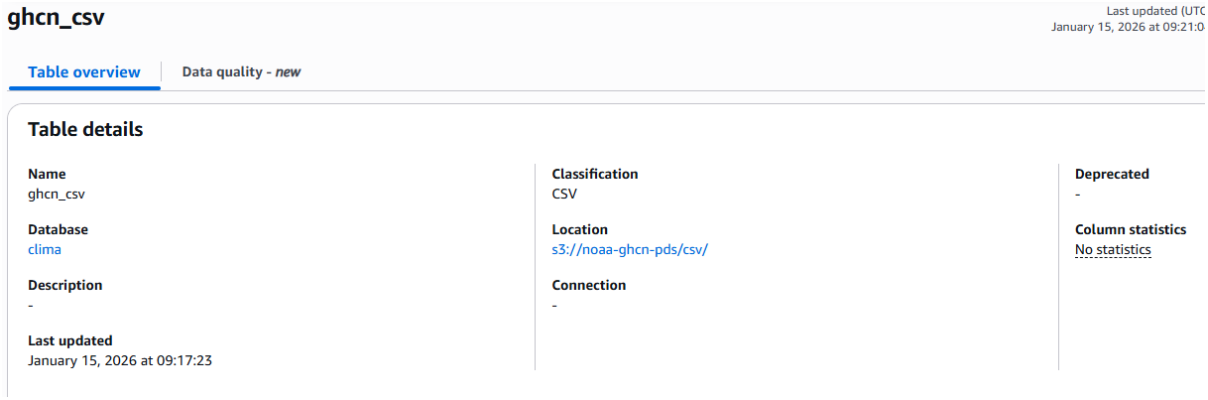


3.) Desde el apartado de Tablas de AWS GLUE, muestra la descripción del esquema de las tablas detectadas y el resumen de estadístico de sus columnas.

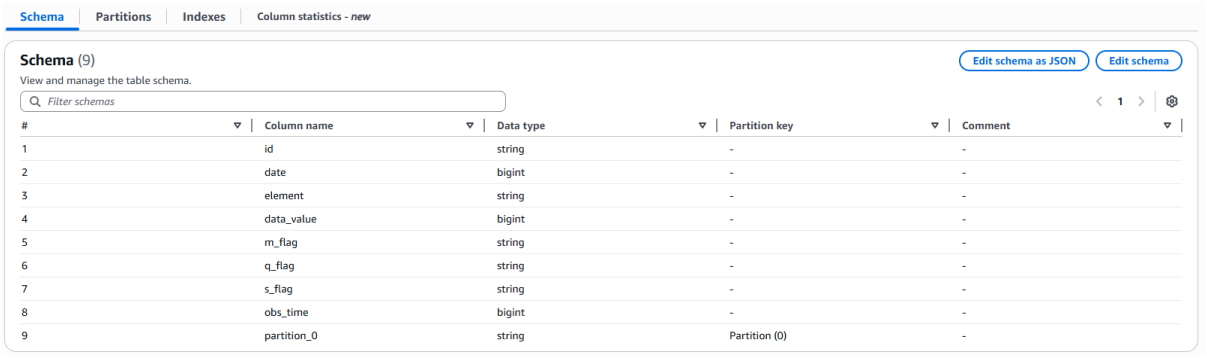
Vamos a volver al menú de “databases” y seleccionar la que se creó



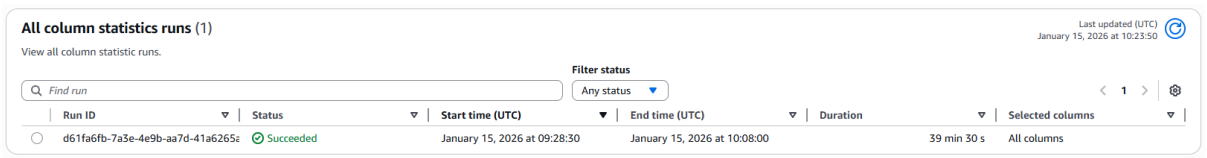
Una vez dentro nos aparecerá lo siguiente:

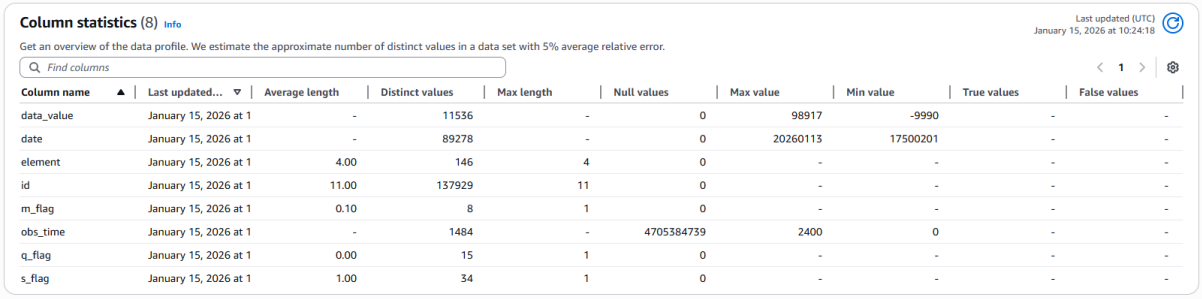


En el apartado de "esquema" no aparece el esquema de las tablas detectadas



Y en el apartado de "Column statistics", le damos a generar estadística y después de un tiempo, no muestra las estadísticas:





Column statistics (8) Info

Last updated (UTC)
January 15, 2026 at 10:24:18

Get an overview of the data profile. We estimate the approximate number of distinct values in a data set with 5% average relative error.

Find columns

1

Column name	Last updated...	Average length	Distinct values	Max length	Null values	Max value	Min value	True values	False values
data_value	January 15, 2026 at 1	-	11536	-	0	98917	-9990	-	-
date	January 15, 2026 at 1	-	89278	-	0	20260113	17500201	-	-
element	January 15, 2026 at 1	4.00	146	4	0	-	-	-	-
id	January 15, 2026 at 1	11.00	137929	11	0	-	-	-	-
m_flag	January 15, 2026 at 1	0.10	8	1	0	-	-	-	-
obs_time	January 15, 2026 at 1	-	1484	-	4705384739	2400	0	-	-
q_flag	January 15, 2026 at 1	0.00	15	1	0	-	-	-	-
s_flag	January 15, 2026 at 1	1.00	34	1	0	-	-	-	-

4.) ¿Está particionada la tabla? ¿Por qué campos?

Para comprobar esto, nos podemos dirigir al apartado de “partitions” y verlo desde ahí:

Schema | Partitions | Indexes | Column statistics - *new*

Partitions (2)

The list of partitions for this table.

Filter partitions

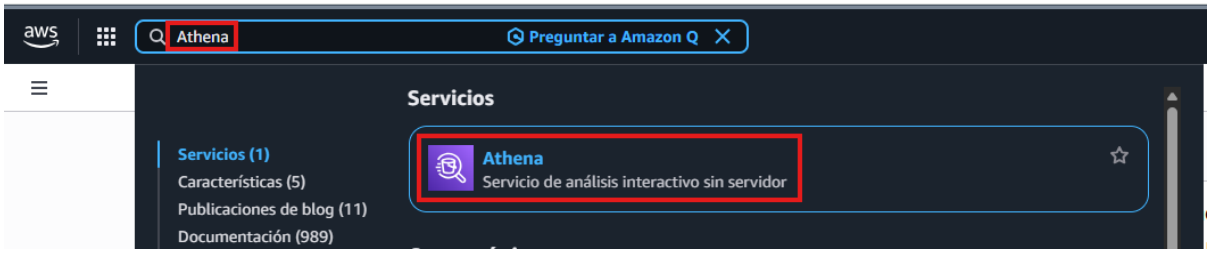
1

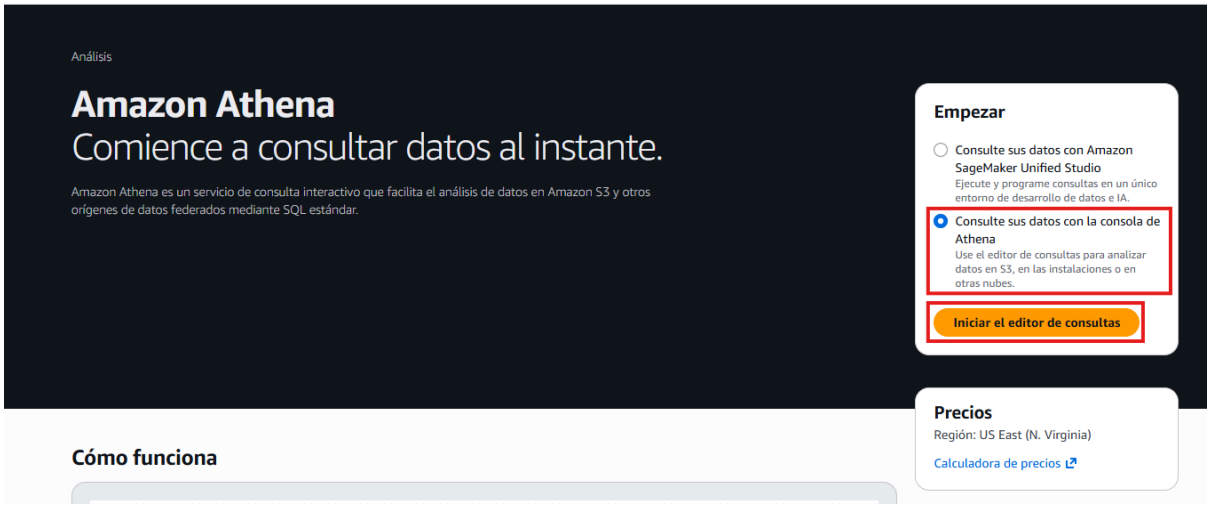
	partition_0	Files	Properties
○	by_station	View files	View Properties
○	by_year	View files	View Properties

Y está estructurada en función de la estructura de carpetas del bucket S3, por lo tanto, tiene dos carpetas, “by_year” y “by_station”

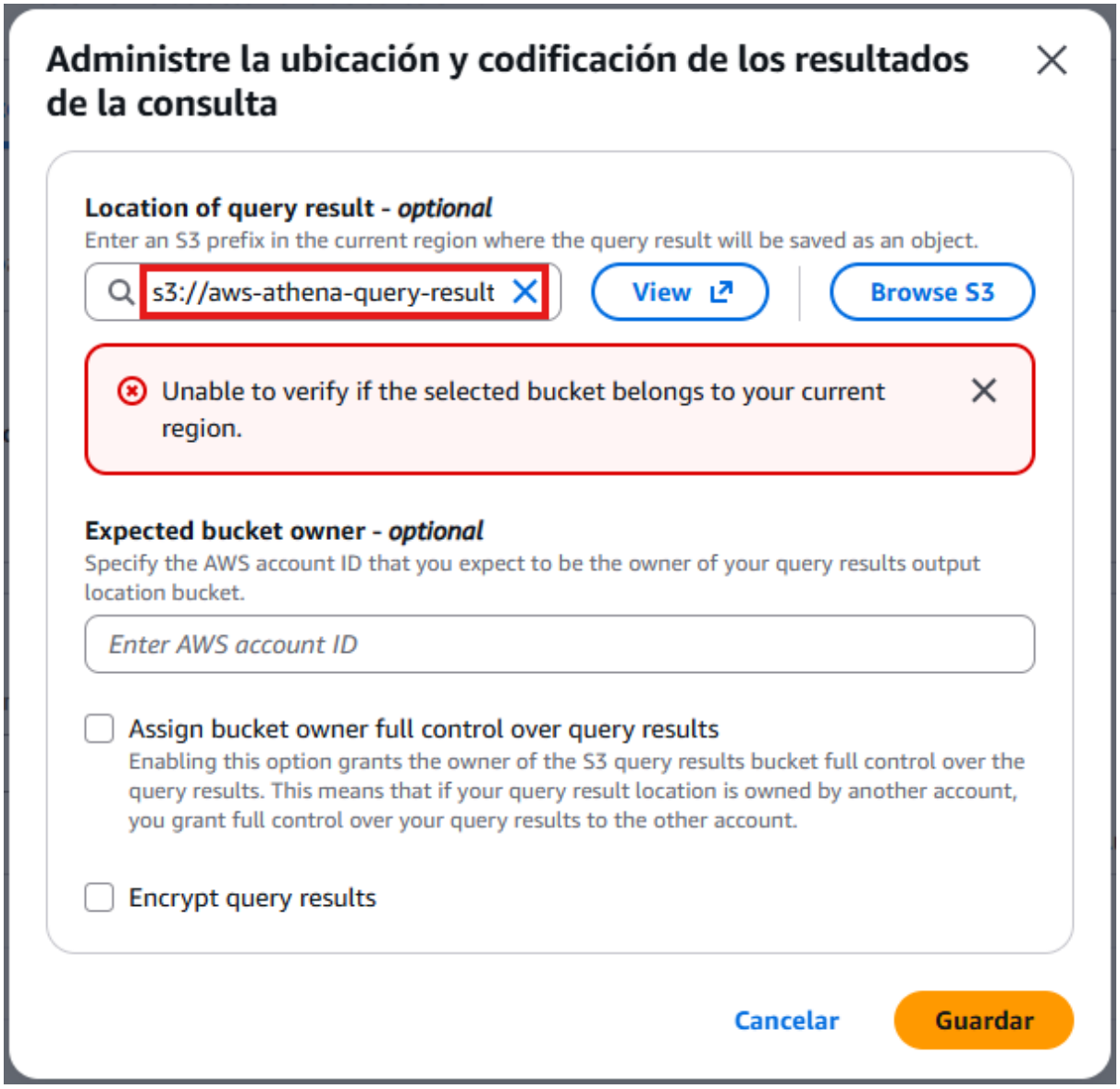
Apartado C

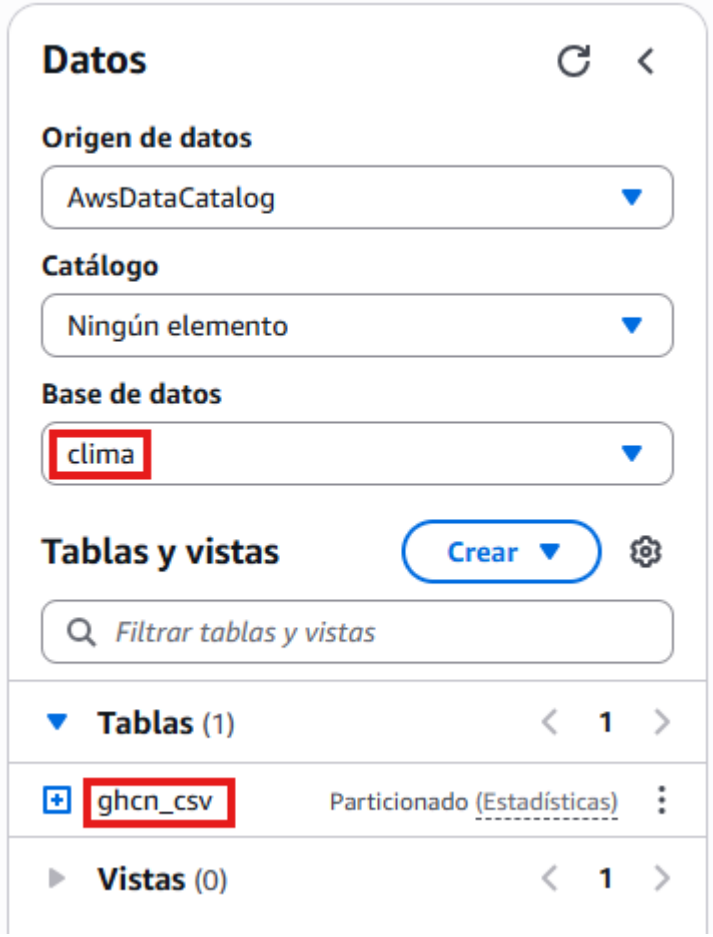
Primero de todo, tenemos que entrar en “Athena”, para ello, lo buscamos en el buscador de AWS





Ahora vamos a configurarlo rápidamente


Administre la ubicación y codificación de los resultados de la consulta
Location of query result - optional
Enter an S3 prefix in the current region where the query result will be saved as an object.
s3://aws-athena-query-result
View
Browse S3
Unable to verify if the selected bucket belongs to your current region.
Expected bucket owner - optional
Specify the AWS account ID that you expect to be the owner of your query results output location bucket.
Enter AWS account ID
Assign bucket owner full control over query results
Enabling this option grants the owner of the S3 query results bucket full control over the query results. This means that if your query result location is owned by another account, you grant full control over your query results to the other account.
Encrypt query results
Cancelar
Guardar



1.) ¿Cuántos registros tiene la tabla?

2.) ¿Cuántas mediciones tenemos de España?

Consulta 1

```
1  SELECT COUNT(*) AS mediciones_espana
2  FROM ghcn_csv
3  WHERE id LIKE 'SP%';
```

Resultados (1)

Filas de búsqueda

#	mediciones_espana
1	21166382

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

Consulta 1

```
1  SELECT COUNT(*) AS mediciones_asturias
2  FROM ghcn_csv
3  WHERE id IN (
4    'SPE00119792',
5    'SPE00119801',
6    'SPE00119819',
7    'SPE00119828'
8  );
```

Resultados (1)

Filas de búsqueda

#	mediciones_asturias
1	544046

4.) ¿Cuántas mediciones tenemos de Oviedo?

Consulta 1

```
1  SELECT COUNT(*) AS mediciones_oviedo
2  FROM ghcn_csv
3  WHERE id = 'SPE00119828';
```

Resultados (1)

Filas de búsqueda

#	mediciones_oviedo
1	146094

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

España:

Consulta 1

```
1  SELECT MIN(date) AS fecha_mas_antigua_espana
2  FROM ghcn_csv
3  WHERE id LIKE 'SP%';
```

Resultados (1)

Filas de búsqueda

#	fecha_mas_antigua_espana
1	18961101

Asturias:

Consulta 1

```sql
SELECT MIN(date) AS fecha_mas_antigua_asturias
FROM ghcn_csv
WHERE id IN (
  'SPE00119792',
  'SPE00119801',
  'SPE00119819',
  'SPE00119828'
);
```

Resultados (1)

Filas de búsqueda

#	fecha_mas_antigua_asturias
1	19381001

Oviedo:

Consulta 1

```sql
SELECT MIN(date) AS fecha_mas_antigua_oviedo
FROM ghcn_csv
WHERE id = 'SPE00119828';
```

Resultados (1)

Filas de búsqueda

#	fecha_mas_antigua_oviedo
1	19721201