

# **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 trabajo del apartado C. ....	5
3.) Finalmente hemos de crear un trigger 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

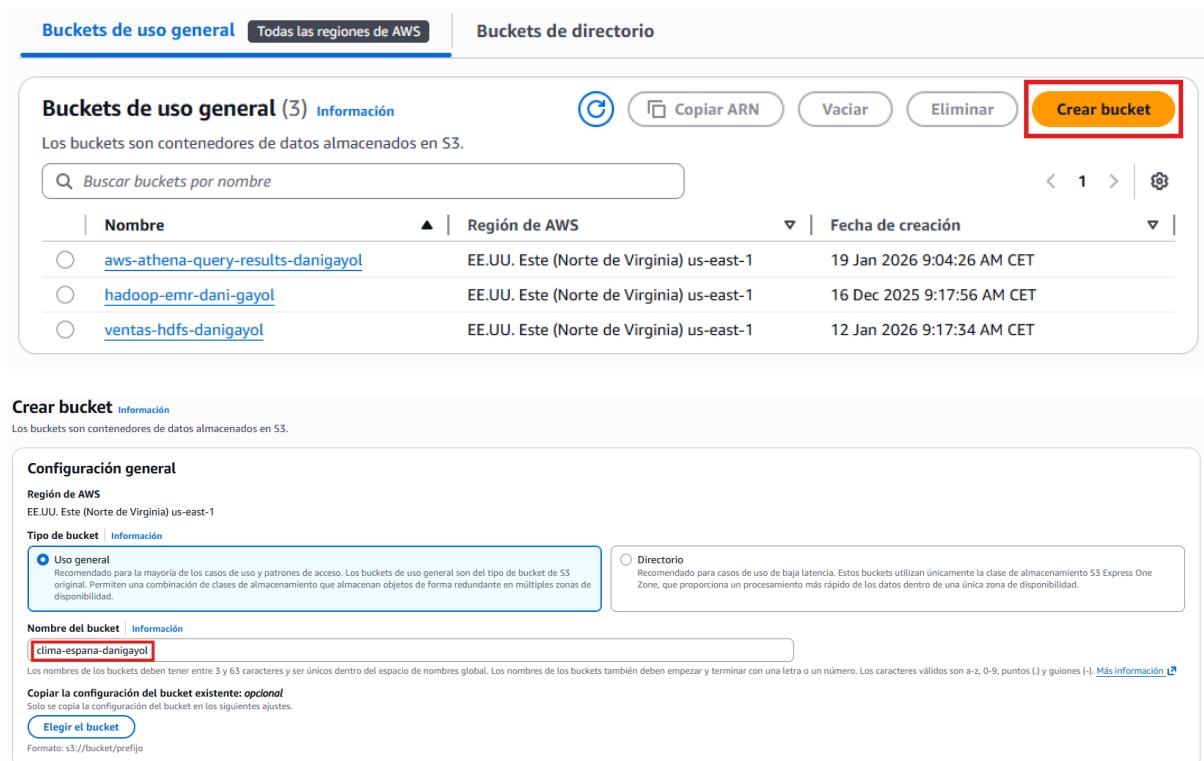
## 1.) Crea un bucket S3 con una carpeta dentro, por ejemplo, clima/espana

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



The screenshot shows the AWS search results for "S3". The "S3" service card is highlighted with a red box. The card displays the service name "S3" with the subtitle "Almacenamiento escalable en la nube". Other services listed include "S3 Glacier" and "AWS Snow Family". On the left, there's a sidebar with categories like "Servicios (9)", "Características (40)", etc.

Ahora le damos al botón de “Crear Bucket”



The screenshot shows the AWS S3 buckets list and the "Create bucket" wizard. In the buckets list, three buckets are shown: "aws-athena-query-results-danigayol", "hadoop-emr-dani-gayol", and "ventas-hdfs-danigayol". The "Crear bucket" button is highlighted with a red box. In the "Create bucket" wizard, the "Configuración general" section is visible, showing the region "EE.UU. Este (Norte de Virginia) us-east-1" and the "Uso general" bucket type selected. The bucket name input field contains "clima-espana-danigayol".

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

**clima-espana-danigayol** Información

**Objetos (0)** Crear carpeta Cargar

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

Buscar objetos por prefijo

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
No hay objetos No tiene objetos en este bucket.				
<a href="#">Cargar</a>				

**Crear carpeta** Información

Utilice carpetas para agrupar los objetos en buckets. Al crear una carpeta, S3 creará un objeto con el nombre que usted especifique seguido de una barra inclinada (/). Este objeto luego aparecerá como una carpeta en la consola. [Más información](#)

**Su política de bucket podría bloquear la creación de carpetas**  
Si su política de bucket impide cargar objetos sin etiquetas, metadatos o beneficiarios específicos de la lista de control de acceso (ACL), no podrá crear una carpeta con esta configuración. En su lugar, puede utilizar la [configuración de carga](#) para cargar una carpeta vacía y especificar la configuración adecuada.

**Carpeta**

Nombre de la carpeta  /

Los nombres de las carpetas no pueden contener "/". Consulte las [reglas de nomenclatura](#).

**Cifrado del lado del servidor** Información

El cifrado del lado del servidor protege los datos en reposo.

La siguiente configuración de cifrado se aplica únicamente al objeto de carpeta y no a los objetos de subcarpeta.

**Cifrado del lado del servidor**

No especificar una clave de cifrado  
La configuración del bucket para el cifrado predeterminado se utiliza para cifrar el objeto de carpeta al almacenarlo en Amazon S3.

Especificar una clave de cifrado  
La clave de cifrado especificada se utiliza para cifrar el objeto de carpeta antes de almacenarlo en Amazon S3.

**⚠ Si la política del bucket requiere que los objetos se cifren con una clave de cifrado específica, deberá especificar la misma clave de cifrado al crear una carpeta. De lo contrario, se producirá un error al crear la carpeta.**

[Cancelar](#) Crear carpeta

## 2.) Mediante un comando AWS CLI, copia los archivos csv con las mediciones de todas las estaciones meteorológicas 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

Amazon S3 > Buckets > clima-espana-danigayol > clima/ > espana/

**espana/**

Objetos Propiedades

**Objetos (207)**

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

Copiar URI de S3 Copiar URL Descargar Eliminar Acciones Crear carpeta Cargar

Buscar objetos por prefijo

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
SP000003195.csv	csv	19 Jan 2026 10:13:41 AM CET	3.8 MB	Estándar
SP000004452.csv	csv	19 Jan 2026 10:13:41 AM CET	3.2 MB	Estándar
SP000006155.csv	csv	19 Jan 2026 10:13:41 AM CET	3.6 MB	Estándar
SP000007038.csv	csv	19 Jan 2026 10:13:41 AM CET	4.0 MB	Estándar
SP000008027.csv	csv	19 Jan 2026 10:13:41 AM CET	4.2 MB	Estándar
SP000008181.csv	csv	19 Jan 2026 10:13:41 AM CET	3.8 MB	Estándar
SP000008202.csv	csv	19 Jan 2026 10:13:41 AM CET	3.5 MB	Estándar
SP000008215.csv	csv	19 Jan 2026 10:13:41 AM CET	3.5 MB	Estándar
SP000008280.csv	csv	19 Jan 2026 10:13:41 AM CET	4.4 MB	Estándar
SP000008410.csv	csv	19 Jan 2026 10:13:41 AM CET	3.0 MB	Estándar
SP000008416.csv	csv	19 Jan 2026 10:13:41 AM CET	3.3 MB	Estándar
SP000009434.csv	csv	19 Jan 2026 10:13:41 AM CET	2.8 MB	Estándar
SP000009981.csv	csv	19 Jan 2026 10:13:41 AM CET	4.9 MB	Estándar
SP000060010.csv	csv	19 Jan 2026 10:13:41 AM CET	4.0 MB	Estándar

### 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”,

aws AWS Glue Preguntar a Amazon Q

Servicios Ver los 159 resultados

- AWS Glue** AWS Glue es un servicio de integración de datos sin servidor.
- AWS Glue DataBrew** Herramienta de preparación de datos visuales para limpiar y normalizar datos para a...
- AWS Private Certificate Authority** Servicio de la entidad de certificación privada administrada

Amazon S3 Servicios (159) Características (429) Publicaciones de blog (811) Documentación (53.493) Artículos de conocimiento (1339) Tutoriales (217) Eventos (3) Marketplace (23.120)

Buckets Buckets de uso general Buckets de desarrollo Buckets de trabajo Buckets vectoriales Seguridad y acceso

En el menú de la izquierda, entramos en “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

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

## Create a database

Create a database in the AWS Glue Data Catalog.

### Database details

#### Name

espana

Database name is required, in lowercase characters, and no longer than 255 characters.

#### Description - *optional*

Enter text

Descriptions can be up to 2048 characters long.

### Database settings

#### Location - *optional*

Set the URI location for use by clients of the Data Catalog.

An S3 location is required for managed tables and Zero-ETL integrations.

**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”

## 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

Al “Crawler” le ponemos la siguiente configuración:

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-espana

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.

#### Tags - *optional*

Use tags to organize and identify your resources.

## 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).

▼ C

[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.

X [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.

Type	Data source	Parameters
<input type="radio"/> 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](#)

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  
 LabRole [View](#)

[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.

**Lake Formation configuration - optional**

Allow the crawler to use Lake Formation credentials for crawling the data source. [Learn more](#).

Use Lake Formation credentials for crawling S3 data source  
 Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, only those data sources associated to the account. Only applicable to S3, Glue Catalog, Iceberg, and Hudi data sources.

## Set output and scheduling

**Output configuration** [Info](#)

Target database  
 espana [View](#)

[Clear selection](#) [Add database](#)

Table name prefix - optional  
 espcsv\_

Maximum table threshold - optional  
 This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.  
 Type a number greater than 0

[Advanced options](#)

**Crawler schedule**

You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. [Learn more](#).

Frequency  
 On demand [View](#)

[Cancel](#) [Previous](#) [Next](#)

## Y así quedaría finalmente:

**Review and create**

**Step 1: Set crawler properties**

**Set crawler properties**

Name	Description	Tags
crawler-espana	-	-

**Step 2: Choose data sources and classifiers**

**Data sources (1)** [Info](#)

The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
S3	s3://clima-espana-danigayol/clima/espana/	Recrawl all

**Step 3: Configure security settings**

**Configure security settings**

IAM role	Security configuration	Lake Formation configuration
LabRole	-	-

**Step 4: Set output and scheduling**

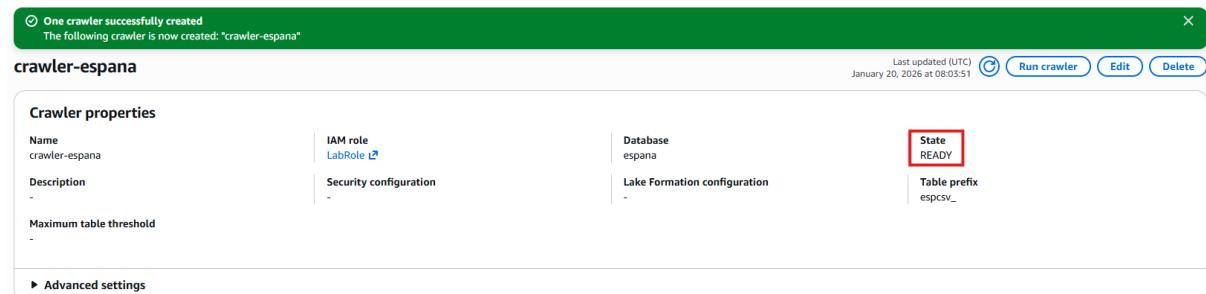
**Set output and scheduling**

Database	Table prefix - optional	Maximum table threshold - optional	Schedule
espana	espcsv_	-	On demand

[Cancel](#) [Previous](#) [Create crawler](#)

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

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



The screenshot shows the 'crawler-espana' configuration page. At the top, a green banner indicates 'One crawler successfully created' with the message 'The following crawler is now created: "crawler-espana"'. Below the banner, the crawler name 'crawler-espana' is displayed. On the right, there are buttons for 'Run crawler', 'Edit', and 'Delete'. The main area shows 'Crawler properties' with sections for Name (crawler-espana), IAM role (LabRole), Database (espana), Security configuration, Lake Formation configuration, and State (READY). A red box highlights the 'State READY' field. At the bottom left, there is a link to 'Advanced settings'.

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

# Apartado B

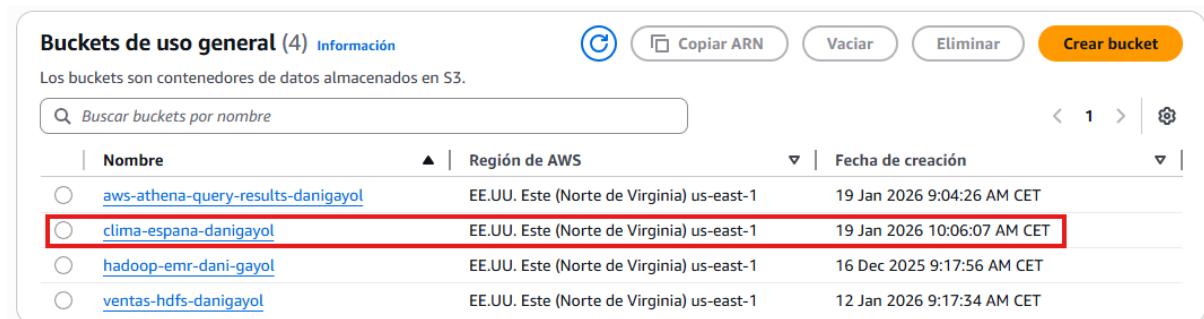
1.) Crea una carpeta dentro del bucket anterior (clima) con el nombre parquet. por ejemplo, clima/parquet

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



The screenshot shows the AWS search results for "S3". The "S3" service card is highlighted with a red box. The service card includes the icon, name, and description: "Almacenamiento escalable en la nube". Other services listed include S3 Glacier and AWS Snow Family.

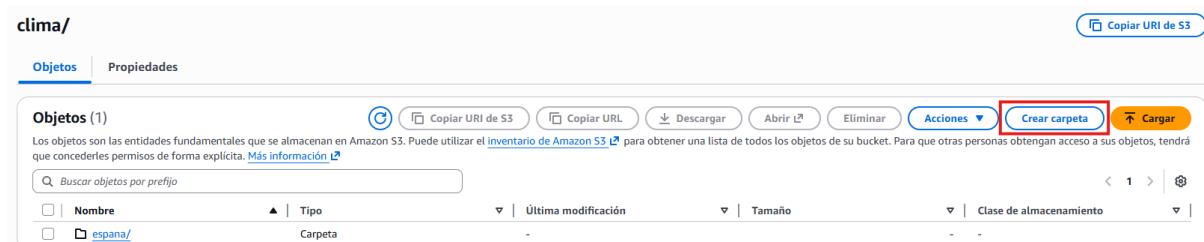
Entramos dentro del bucket que tenemos creado:



The screenshot shows the "Buckets de uso general" list. The "clima-espagna-danigayol" bucket is selected and highlighted with a red box. The table columns are Nombre, Región de AWS, and Fecha de creación. The bucket details show it was created on 19 Jan 2026 at 10:06:07 AM CET.

Nombre	Región de AWS	Fecha de creación
aws-athena-query-results-danigayol	EE.UU. Este (Norte de Virginia) us-east-1	19 Jan 2026 9:04:26 AM CET
<b>clima-espagna-danigayol</b>	EE.UU. Este (Norte de Virginia) us-east-1	19 Jan 2026 10:06:07 AM CET
hadoop-emr-dani-gayol	EE.UU. Este (Norte de Virginia) us-east-1	16 Dec 2025 9:17:56 AM CET
ventas-hdfs-danigayol	EE.UU. Este (Norte de Virginia) us-east-1	12 Jan 2026 9:17:34 AM CET

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



The screenshot shows the "Objetos" list for the "clima/" bucket. The "espana/" folder is selected and highlighted with a red box. The "Crear carpeta" button is also highlighted with a red box. The table columns are Nombre, Tipo, Última modificación, Tamaño, and Clase de almacenamiento.

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
espana/	Carpeta	-	-	-

**Crear carpeta** Información

Utilice carpetas para agrupar los objetos en buckets. Al crear una carpeta, S3 creará un objeto con el nombre que usted especifique seguido de una barra inclinada (/). Este objeto luego aparecerá como una carpeta en la consola. [Más información](#)

**Su política de bucket podría bloquear la creación de carpetas**  
Si su política de bucket impide cargar objetos sin etiquetas, metadatos o beneficiarios específicos de la lista de control de acceso (ACL), no podrá crear una carpeta con esta configuración. En su lugar, puede utilizar la [configuración de carga](#) para cargar una carpeta vacía y especificar la configuración adecuada.

**Carpeta**

Nombre de la carpeta  
 /

Los nombres de las carpetas no pueden contener "/". Consulte las reglas de nomenclatura.

**Cifrado del lado del servidor** Información  
El cifrado del lado del servidor protege los datos en reposo.

La siguiente configuración de cifrado se aplica únicamente al objeto de carpeta y no a los objetos de subcarpeta.

**Cifrado del lado del servidor**

No especificar una clave de cifrado  
La configuración del bucket para el cifrado predeterminado se utiliza para cifrar el objeto de carpeta al almacenarlo en Amazon S3.

Especificar una clave de cifrado  
La clave de cifrado especificada se utiliza para cifrar el objeto de carpeta antes de almacenarlo en Amazon S3.

**Aviso:** Si la política del bucket requiere que los objetos se cifren con una clave de cifrado específica, deberá especificar la misma clave de cifrado al crear una carpeta. De lo contrario, se producirá un error al crear la carpeta.

[Cancelar](#) [Crear carpeta](#)

**Se creó correctamente la carpeta "parquet"**

clima/

[Copiar URI de S3](#)

**Objetos** **Propiedades**

**Objetos (2)**

<input type="checkbox"/>	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	espana/	Carpetas	-	-	-
<input type="checkbox"/>	parquet/	Carpetas	-	-	-

**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”

Servicios

[Ver los 159 resultados](#)

- AWS Glue**  
AWS Glue es un servicio de integración de datos sin servidor.
- AWS Glue DataBrew**  
Herramienta de preparación de datos visuales para limpiar y normalizar datos para a...
- AWS Private Certificate Authority**  
Servicio de la entidad de certificación privada administrada

**Amazon S3**

**Buckets**

**Seguridad y acceso**

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

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

AWS Glue Studio [Info](#)

Create job [Info](#)

Visual ETL

Author in a visual interface focused on data flow.

Notebook

Author using an interactive code notebook.

Script editor

Author code with a script editor.

Example jobs [Info](#) Create example job

Your jobs (0) [Info](#)

Actions Run job

< 1 > Filter

No jobs

You have not created a job yet.

Create job from a blank graph

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

Visual | Script | Job details | Runs | Data q

+ Add nodes X

Search sources, transforms and targets

Sources | Transforms | Targets | Popular

AWS Glue Data Catalog  
AWS Glue Data Catalog table as the data source.

Amazon S3  
JSON, CSV, or Parquet files stored in S3.

Amazon Kinesis  
Read from an Amazon Kinesis Data Stream.

Apache Kafka  
Read from an Apache Kafka or Amazon MSK topic.

Relational DB  
AWS Glue Data Catalog table with a relational database as the data source.

Amazon Redshift  
Read your data from Amazon Redshift.

MySQL  
AWS Glue Data Catalog table with MySQL as the data source.

PostgreSQL  
AWS Glue Data Catalog table with PostgreSQL as the data source.

Oracle SQL  
AWS Glue Data Catalog table with Oracle SQL as the data source.

Manage Connections ↗

Lo configuramos de la siguiente manera:

**Data source properties - Data Catalog**

**Name**  
AWS Glue Data Catalog

**Database**  
Choose a database.  
clima ▾ C

▶ Use runtime parameters

**Table**  
ghcn\_csv ▾ C

▶ Use runtime parameters

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 ↗

**Transform**

**Name**  
Change Schema

**Node parents**  
Choose which nodes will provide inputs for this one.  
Choose one or more parent node

AWS Glue Data Catalog X  
Catalog - DataSource

**Change Schema (Apply mapping)**

Source key	Target key	Data type	Drop
id	id	string ▼	<input type="checkbox"/>
date	fecha	long ▼	<input type="checkbox"/>
element	tipo_medicion	string ▼	<input type="checkbox"/>
data_value	tipo_medicion	long ▼	<input type="checkbox"/>
m_flag	marca_medicion	string ▼	<input type="checkbox"/>
q_flag	marca_calidad	string ▼	<input type="checkbox"/>
s_flag	fuente	string ▼	<input type="checkbox"/>
obs_time	hora_observacion	long ▼	<input type="checkbox"/>
partition_0	particion_0	string ▼	<input type="checkbox"/>

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

Visual | Script | Job details | Runs | Data c

+ Add nodes X

Search sources, transforms and targets

Sources | **Transforms** | Targets | Popular

AWS Glue Data Catalog  
AWS Glue Data Catalog table as the data target.

**Amazon S3**  
S3 bucket by specifying a bucket path as the data target.

Amazon Redshift  
Write your data to Amazon Redshift.

MySQL  
AWS Glue Data Catalog table with MySQL as the data target.

PostgreSQL  
AWS Glue Data Catalog table with PostgreSQL as the data target.

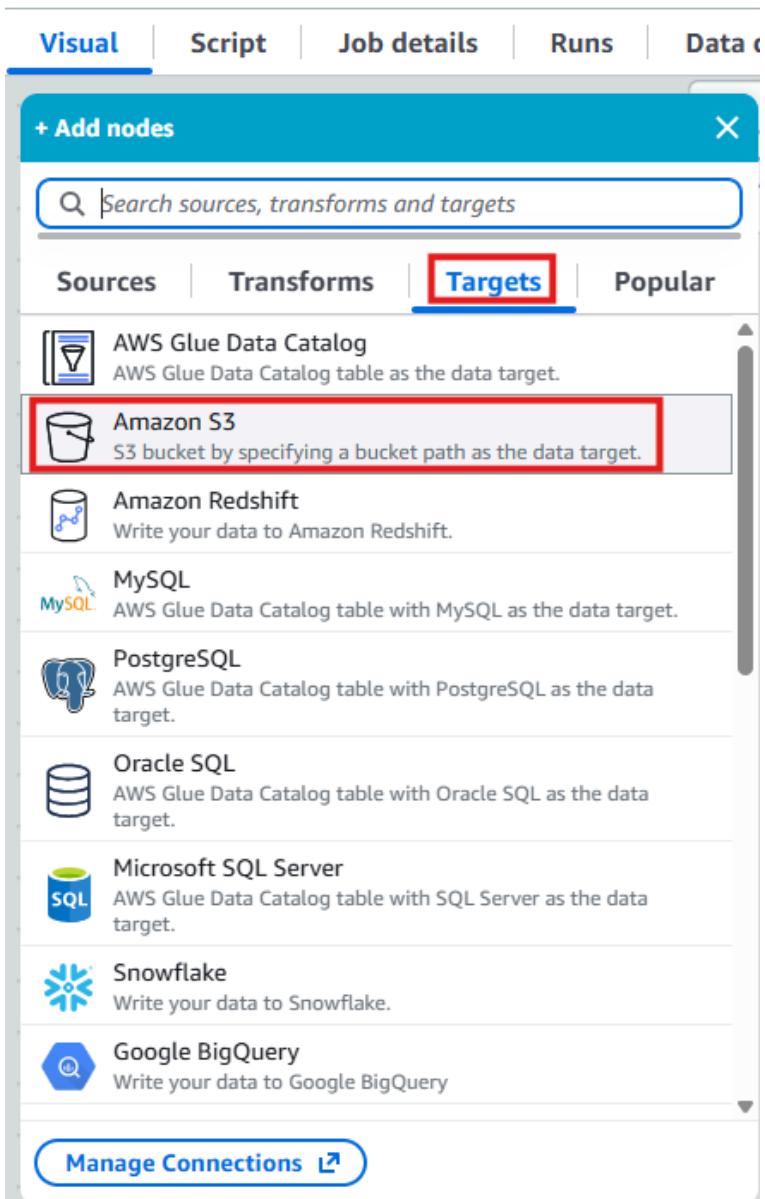
Oracle SQL  
AWS Glue Data Catalog table with Oracle SQL as the data target.

Microsoft SQL Server  
AWS Glue Data Catalog table with SQL Server as the data target.

Snowflake  
Write your data to Snowflake.

Google BigQuery  
Write your data to Google BigQuery

Manage Connections ↗



Y lo configuramos de la siguiente manera:

**Data target properties - S3**

**Name**  
Amazon S3

**Node parents**  
Choose which nodes will provide inputs for this one.  
*Choose one or more parent node*

**Format**  
Parquet

**Compression Type**  
Snappy

**S3 Target Location**  
Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).  
s3://clima-espana-danigayol/clima/parquet/ [View](#) [Browse S3](#)

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

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



Y nos aparecerá aquí;

Your jobs (1) <a href="#">Info</a>						<a href="#">Actions</a>	<a href="#">Run job</a>
<input type="checkbox"/> <a href="#">Filter jobs by property</a>						Last modified	AWS Glue version
Job name	Type	Created by	Last modified	AWS Glue version	Action		
campos_en_espanol	Glue ETL	Visual	20/1/2026, 10:19:52	5.0	-		

# 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”

## 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

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

The screenshot shows the 'Set crawler properties' step of the AWS Glue Create Crawler wizard. On the left, a sidebar lists five steps: Step 1 (Set crawler properties, marked with a circle), Step 2 (Choose data sources and classifiers), Step 3 (Configure security settings), Step 4 (Set output and scheduling), and Step 5 (Review and create). The main area is titled 'Set crawler properties' and contains a 'Crawler details' section. It includes a 'Name' field with the value 'crawler-parquet-espana' and a 'Description - optional' field with the placeholder 'Enter a description'. A note below the 'Name' field states: 'Name can be up to 255 characters long. Some character set including control characters are prohibited.' Another note below the 'Description' field states: '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.

Type	Data source	Parameters
You don't have any data sources. <a href="#">Add a data source</a>		

### 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 [X](#) [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.

Type	Data source	Parameters
<input checked="" type="radio"/> S3	s3://clima-espana-danigayol/clima/parquet/	Recrawl all!

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  
 LabRole  

[Create new IAM role](#) [Update chosen IAM role](#) [View](#)  

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

**Lake Formation configuration - optional**

Allow the crawler to use Lake Formation credentials for crawling the data source. [Learn more](#)  

Use Lake Formation credentials for crawling S3 data source  
 Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, crawl only those data sources associated to the account. Only applicable to S3, Glue Catalog, Iceberg, and Hudi data sources.

## Set output and scheduling

**Output configuration** [Info](#)

Target database  
 clima  

[Clear selection](#) [Add database](#)  

Table name prefix - optional  
 espparq\_  

**Maximum table threshold - optional**  
 This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will a depend on the data schema.

**Advanced options**

**Crawler schedule**  
 You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. [Learn more](#)  

Frequency  
 On demand  

**Finalmente, nos quedaría de la siguiente manera:**

**Review and create**

**Step 1: Set crawler properties**

**Set crawler properties**

Name	Description	Tags
crawler-parquet-espana	-	-

**Step 2: Choose data sources and classifiers**

**Data sources (1)** [Info](#)

The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
S3	s3://clima-espana-danigayol/clima/parquet/	Recrawl all

**Step 3: Configure security settings**

**Configure security settings**

IAM role	Security configuration	Lake Formation configuration
LabRole	-	-

**Step 4: Set output and scheduling**

**Set output and scheduling**

Database	Table prefix - optional	Maximum table threshold - optional	Schedule
clima	espparq_	-	On demand

[Cancel](#) [Previous](#) [Create crawler](#)

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

## Ahora creamos el “crawler” pero NO lo ejecutamos

One crawler successfully created  
The following crawler is now created: "crawler-parquet-espana"

crawler-parquet-espana

Last updated (UTC) January 21, 2026 at 10:10:26    Run crawler    Edit    Delete

Crawler properties			
Name crawler-parquet-espana	IAM role <a href="#">LabRole</a>	Database clima	State READY
Description -	Security configuration -	Lake Formation configuration -	Table prefix espparq_
► Advanced settings			

# 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”

The screenshot shows the AWS Glue navigation menu. The main menu items include Getting started, ETL jobs, Visual ETL, Notebooks, Job run monitoring, Data Catalog tables, Data connections, Workflows (orchestration), Zero-ETL integrations (marked as New), Data Catalog, Data Integration and ETL, Legacy pages, and Triggers. The 'Triggers' option is highlighted with a red box.

- 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
- ▼ Data Integration and ETL
  - Zero-ETL integrations New
  - ETL jobs
    - Visual ETL
    - Notebooks
    - Job run monitoring
  - Interactive Sessions
  - Data classification tools
    - Sensitive data detection
    - Record Matching
  - Triggers**
  - Workflows (orchestration)
  - Blueprints
  - Security configurations
  - Cost management New
- Legacy pages

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

Step 1  
 Set trigger properties  
 Choose jobs or crawlers to activate  
 Review and create

## Set trigger properties

**Trigger details**

Name  
  
Name can be up to 255 characters long. Some character set including control characters are prohibited.

Description - optional  
  
Descriptions can be up to 2048 characters long.

**Trigger type**

On demand  
Fire the trigger immediately when started.

Schedule  
Fire the trigger on a timer.

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

Type	Name	Status
You don't have any watched resources		
<a href="#">Add a watched resource</a>		

## Add condition

**Resource type**

The type of resource to apply this condition to

**Crawler name**  
 

Crawler to apply this condition to

**Status**

The status transition that the trigger listens for

**Add** Cancel

**Watched resources (1)**  
List of conditions that will start the trigger

Type	Name	Status
<input type="radio"/> Crawler	<a href="#">crawler-espana</a> 	 Succeeded

Step 1  
 Set trigger properties  
 Choose jobs or crawlers to activate  
 Review and create

### Choose jobs or crawlers to activate

**Resources to trigger (0)**  
List of resources to start once the trigger activates

Type	Name	Parameters
You don't have any target resources		

[Edit](#) [Remove](#) [Add a target resource](#)

[Cancel](#) [Previous](#) [Next](#)

### Add target

**Resource type**  
 Job

The type of resource for this trigger target

**Job name**  
 [C](#)

Job to start when this trigger fires

► Parameters passed down to job "campos\_en\_espanol" when started - *optional*

[Cancel](#) [Add](#)

### Choose jobs or crawlers to activate

**Resources to trigger (1)**  
List of resources to start once the trigger activates

Type	Name	Parameters
<input checked="" type="radio"/> Job	<a href="#">campos_en_espanol</a>	-

[Edit](#) [Remove](#) [Add a target resource](#)

[Cancel](#) [Previous](#) [Next](#)

**Nos tendría que quedar configurado de la siguiente manera:**

**Review and create**

**Step 1: Set trigger properties**

[Edit](#)

**Trigger details**

Name espa_ab	Description -	Tags -
-----------------	------------------	-----------

**Watched resources (1)**  
List of conditions that will start the trigger

Type Crawler	Name <a href="#">crawler-espana</a>	Status Succeeded
-----------------	--	---------------------

**Step 2: Choose jobs or crawlers to activate**

**Resources to trigger (1)**  
List of resources to start once the trigger activates

Type	Name	Parameters
Job	<a href="#">campos_en_espanol</a>	-

Enable trigger on creation

[Edit](#) [Cancel](#) [Previous](#) [Create](#)

**Trigger successfully created**  
The following trigger was created: "espa\_ab"

Last updated (UTC) January 21, 2026 at 10:55:54 [Edit trigger](#)

**espa\_ab**

**Trigger properties**

Name espa_ab	Description -
Trigger type Conditional	Status <span style="color: green;">Activated</span>
Associated workflow -	

**Target resources** | Watched resources | Tags

**Resources to trigger (1)**  
List of resources to start once the trigger activates

Type	Name	Parameters
Job	campos_en_espanol	-

**2.) Crea un disparador (trigger) -puedes llamarlo espa\_bc - que después de finalizado el trabajo del apartado B lance el trabajo del apartado C.**

Ahora vamos a crear otro “trigger” con la siguiente configuración:

**Step 1**

**Set trigger properties**

Step 2  
Choose jobs or crawlers to activate

Step 3  
Review and create

**Set trigger properties**

**Trigger details**

**Name**  
 Name can be up to 255 characters long. Some character set including control characters are prohibited.

**Description - optional**  
 Descriptions can be up to 2048 characters long.

**Trigger type**

On demand  
Fire the trigger immediately when started.

Schedule  
Fire the trigger on a timer.

**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

Type	Name	Status
You don't have any watched resources		
<a href="#">Add a watched resource</a>		

### Add condition

**Resource type**

**Job**

The type of resource to apply this condition to

**Job name**

**campos\_en\_espanol**

Job to apply this condition to

**Status**

**Succeeded**

The status transition that the trigger listens for

**Add**

Watched resources (1)		
List of conditions that will start the trigger		
Type	Name	Status
<input type="radio"/> Job	campos_en_espanol	Succeeded

Step 1 Set trigger properties  
 Step 2 Choose jobs or crawlers to activate  
 Step 3 Review and create

### Choose jobs or crawlers to activate

**Resources to trigger (0)**

List of resources to start once the trigger activates

**Add a target resource**

**Add a target resource**

**Cancel** **Previous** **Next**

### Add target

**Resource type**

**Crawler**

The type of resource for this trigger target

**Crawler name**

**crawler-parquet-espana**

Crawler to start when this trigger fires

**Add**

Choose jobs or crawlers to activate		
Resources to trigger (1)		
List of resources to start once the trigger activates		
Type	Name	Parameters
<input type="radio"/> Crawler	crawler-parquet-espana	-

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

**Review and create**

**Step 1: Set trigger properties**

Trigger details		
Name	Description	Tags
espa_bc	-	-

**Watched resources (1)**  
List of conditions that will start the trigger

Type	Name	Status
Job	<a href="#">campos_en_espanol</a>	Succeeded

**Step 2: Choose jobs or crawlers to activate**

Resources to trigger (1)		
List of resources to start once the trigger activates		
Type	Name	Parameters
Crawler	<a href="#">crawler-parquet-espana</a>	-

Enable trigger on creation

[Cancel](#) [Previous](#) [Create](#)

Trigger successfully created  
The following trigger was created: "espa\_bc"

**espa\_bc**

Last updated (UTC) January 21, 2026 at 10:41:51 [Edit trigger](#)

**Trigger properties**

Name	Description
espa_bc	-

**Trigger type**  
Conditional

**Associated workflow**  
-

[Target resources](#) [Watched resources](#) [Tags](#)

**Resources to trigger (1)**  
List of resources to start once the trigger activates

Type	Name	Parameters
Crawler	<a href="#">crawler-parquet-espana</a>	-

**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:**

Step 1  
 Set trigger properties  
 Step 2  
 Choose jobs or crawlers to activate  
 Step 3  
 Review and create

## Set trigger properties

### Trigger details

Name  
  
Name can be up to 255 characters long. Some character set including control characters are prohibited.

Description - optional  
  
Descriptions can be up to 2048 characters long.

### Trigger type

On demand  
Fire the trigger immediately when started.

Schedule  
Fire the trigger on a timer.

Job or crawler event  
Fire the trigger when job or crawler events match your watched list.

Step 1  
 Set trigger properties  
 Step 2  
 Choose jobs or crawlers to activate  
 Step 3  
 Review and create

## Choose jobs or crawlers to activate

### Resources to trigger (0)

List of resources to start once the trigger activates

Type	Name	Parameters
You don't have any target resources		

[Edit](#) [Remove](#) [Add a target resource](#)

Cancel [Previous](#) [Next](#)

## Add target

Resource type

The type of resource for this trigger target

Crawler name  
 [Edit](#) [Remove](#) [Add](#)

Crawler to start when this trigger fires

Cancel [Add](#)

## Choose jobs or crawlers to activate

### Resources to trigger (1)

List of resources to start once the trigger activates

Type	Name	Parameters
<input checked="" type="radio"/> Crawler	<a href="#">crawler-espana</a>	-

[Edit](#) [Remove](#) [Add a target resource](#)

**Nos tiene que quedar configurado de la siguiente manera:**

**Review and create**

**Step 1: Set trigger properties**

**Trigger details**

Name espa_inicio	Description	Tags
---------------------	-------------	------

**Step 2: Choose jobs or crawlers to activate**

**Resources to trigger (1)**

Type	Name	Parameters
Crawler	crawler-espana	-

**Success Message:** Trigger successfully created  
The following trigger was created: "espa\_inicio"

**Trigger Properties:**

**Name:** espa\_inicio  
**Status:** Created  
**Type:** On demand

**Target resources:**

**Resources to trigger (1)**

Type	Name	Parameters
Crawler	crawler-espana	-

## 4.) Arranca este manualmente este último disparador.

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

**Triggers**

A trigger starts a job when it fires.

**Triggers (1/3)**

Name	Status	Type	Parameters	Targets
espa_ab	Activated	Conditional	1 condition	1 job: campo
espa_bc	Activated	Conditional	1 condition	1 crawler: crawler-espana
<b>espa_inicio</b>	Created	On demand	-	1 crawler: crawler-espana

**Action Menu:** Action ▲ (highlighted), Add trigger, Edit trigger, Deactivate trigger, Activate trigger, Start trigger (highlighted), Delete trigger

**Success Message:** Trigger successfully started  
The following trigger was started: "espa\_inicio".

**Triggers**

A trigger starts a job when it fires.

**Triggers (1/3)**

Name	Status	Type	Parameters	Targets
espa_ab	Activated	Conditional	1 condition	1 job: campo
espa_bc	Activated	Conditional	1 condition	1 crawler: crawler-parquet-espana
<b>espa_inicio</b>	Created	On demand	-	1 crawler: crawler-espana

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) <a href="#">Info</a>							Last updated (UTC) <a href="#">Action</a> <a href="#">Run</a> <a href="#">Create crawler</a>
View and manage all available crawlers.							Table changes from last r...
<input type="checkbox"/>		Name	State	Schedule	Last run	Last run timestamp	Log
<input type="checkbox"/>	crawler-clima	Ready	Succeeded		January 15, 2026 at 09:10:...	<a href="#">View log </a>	1 created
<input type="checkbox"/>	crawler-espana	Ready	Succeeded		January 21, 2026 at 10:49:...	<a href="#">View log </a>	1 created
<input type="checkbox"/>	crawler-parquet-espana	Ready	Succeeded		January 21, 2026 at 10:56:...	<a href="#">View log </a>	-

# 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:

Crawlers						
View and manage all available crawlers.						
Crawlers (3) Info						
Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last r...
crawler-clima	Ready		Succeeded	January 15, 2026 at 09:10:...	<a href="#">View log</a>	1 created
crawler-espana	Ready		Succeeded	January 21, 2026 at 10:49:...	<a href="#">View log</a>	1 created
crawler-parquet-espana	Ready		Succeeded	January 21, 2026 at 10:56:...	<a href="#">View log</a>	-

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

parquet/						
Objetos   Propiedades						
Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el <a href="#">inventario de Amazon S3</a> para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. <a href="#">Más información</a>						
Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento		
run-1768993456911-part-block-0-r-00000-snappy.parquet	parquet	21 Jan 2026 12:14:38 PM CET	146.3 MB	Estándar		
run-1768993456911-part-block-0-r-00001-snappy.parquet	parquet	21 Jan 2026 12:15:46 PM CET	143.7 MB	Estándar		
run-1768993456911-part-block-0-r-00002-snappy.parquet	parquet	21 Jan 2026 12:15:32 PM CET	154.9 MB	Estándar		
run-1768993456911-part-block-0-r-00003-snappy.parquet	parquet	21 Jan 2026 12:14:59 PM CET	145.7 MB	Estándar		
run-1768993456911-part-block-0-r-00004-snappy.parquet	parquet	21 Jan 2026 12:15:12 PM CET	148.3 MB	Estándar		
run-1768993456911-part-block-0-r-00005-snappy.parquet	parquet	21 Jan 2026 12:16:04 PM CET	161.6 MB	Estándar		
run-1768993456911-part-block-0-r-00006-snappy.parquet	parquet	21 Jan 2026 12:13:14 PM CET	145.4 MB	Estándar		
run-1768993456911-part-block-0-r-00007-snappy.parquet	parquet	21 Jan 2026 12:15:53 PM CET	162.7 MB	Estándar		
run-1768993456911-part-block-0-r-00008-snappy.parquet	parquet	21 Jan 2026 12:15:05 PM CET	186.4 MB	Estándar		
run-1768993456911-part-block-0-r-00009-snappy.parquet	parquet	21 Jan 2026 12:13:15 PM CET	155.4 MB	Estándar		

## 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 ahí las tablas que existen

Tables						
View and manage all available tables.						
Tables (3)						
Name	Database	Location	Classification	Deprecated	View data	Data quality
espcsv_espana	espana	s3://clima-espana-danigayc	CSV	-	<a href="#">Table data</a>	<a href="#">View data quality</a>
espparq_parquet	clima	s3://clima-espana-danigay	Parquet	-	<a href="#">Table data</a>	<a href="#">View data quality</a>
ghcn_csv	clima	s3://noaa-ghcn-pds/csv/	CSV	-	<a href="#">Table data</a>	<a href="#">View data quality</a>

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

### Schema (9)

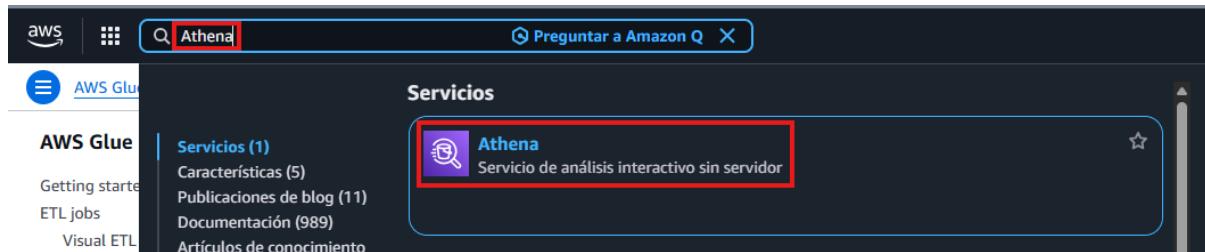
View and manage the table schema.

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

A screenshot of the AWS Athena console. The left sidebar is titled "Datos". Under "Origen de datos", the dropdown menu "AwsDataCatalog" is highlighted with a red box. Under "Catálogo", the dropdown menu "Ningún elemento" is shown. Under "Base de datos", the dropdown menu "espana" is highlighted with a red box. In the center, there's a "Tablas y vistas" section with a "Crear" button and a search bar. Below it, a table list shows "Tablas (1)" with "espcsv\_espana" listed, and "Vistas (0)".

**Datos**

Origen de datos: AwsDataCatalog

Catálogo: Ningún elemento

Base de datos: clima

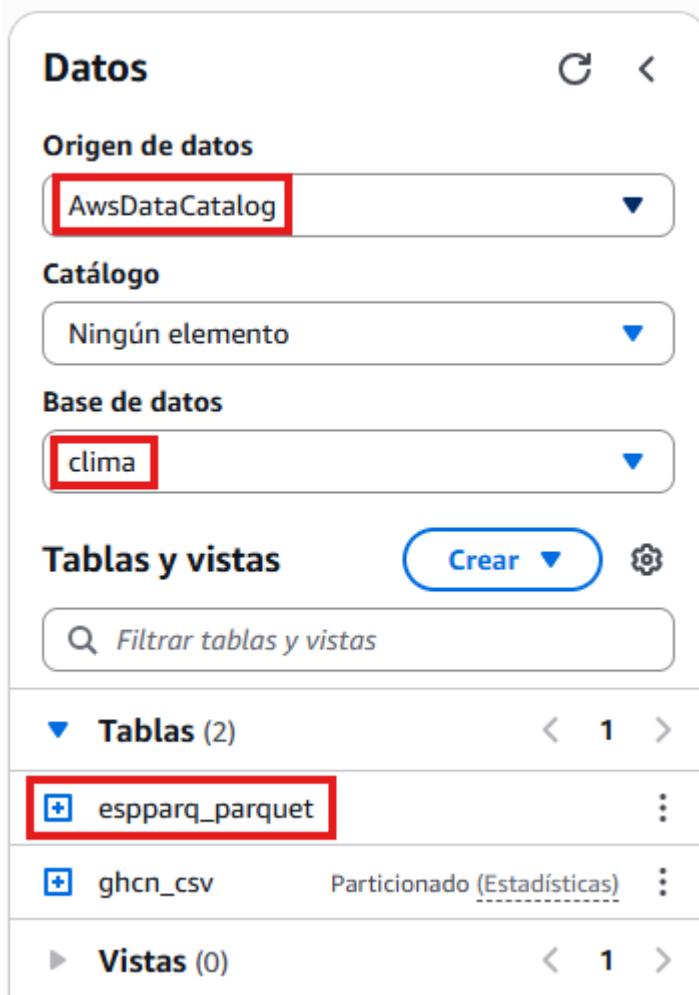
Tablas y vistas: Crear ▾

Filtrar tablas y vistas

▼ Tablas (2) < 1 >

- + espparq\_parquet
- + ghcn\_csv Particionado (Estadísticas)

► Vistas (0) < 1 >



## 1.) ¿Cuántas mediciones tenemos de España?

Tabla

Consulta 1 :

```
1 SELECT COUNT(*) AS total_mediciones
2 FROM espcsv_espana;
```

Resultados de la consulta Estado de la consulta

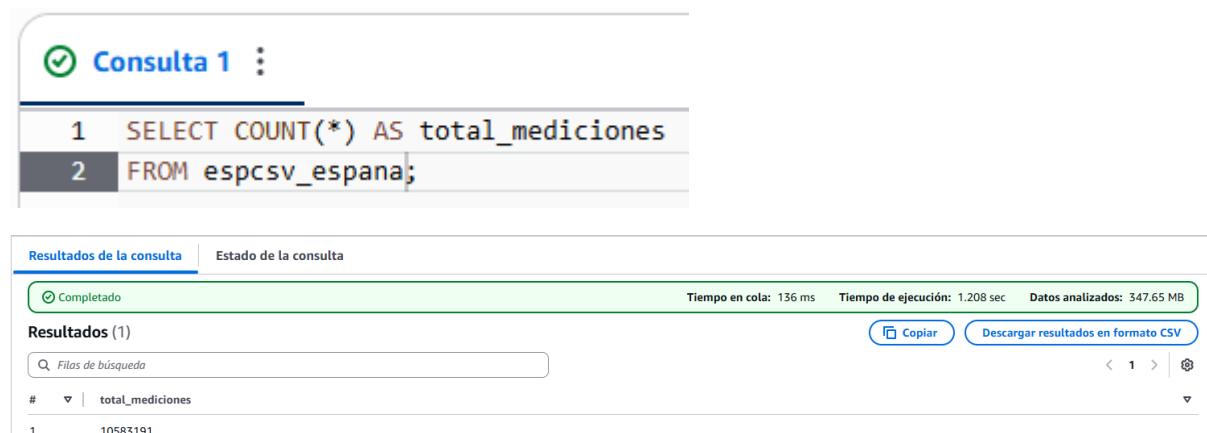
Completado Tiempo en cola: 136 ms Tiempo de ejecución: 1.208 sec Datos analizados: 347.65 MB

Resultados (1)

Filas de búsqueda

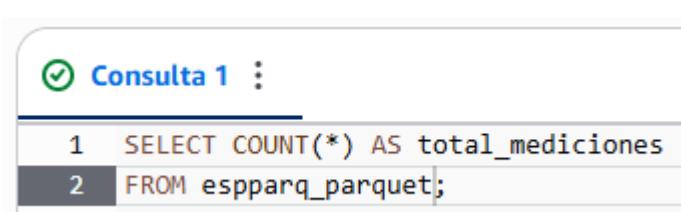
Copiar Descargar resultados en formato CSV

#	total_mediciones
1	10583191



Consulta 1 :

```
1 SELECT COUNT(*) AS total_mediciones
2 FROM espparq_parquet;
```



Resultados de la consulta	Estado de la consulta
<input checked="" type="checkbox"/> Completado	Tiempo en cola: 104 ms Tiempo de ejecución: 938 ms Datos analizados: -
<b>Resultados (1)</b>	
Filas de búsqueda	
# ▾   total_mediciones	<input type="button" value="Copiar"/> <input type="button" value="Descargar resultados en formato CSV"/>
1 7357877863	< 1 > ⚙ ▾

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

### Consulta 1 :

```
1 SELECT COUNT(*) AS mediciones_asturias
2 FROM espcsv_espana
3 WHERE id IN ('SPE00119792', 'SPE00119801', 'SPE00119819', 'SPE00119828');
```

Resultados de la consulta	Estado de la consulta
<input checked="" type="checkbox"/> Completado	Tiempo en cola: 87 ms Tiempo de ejecución: 1.451 sec Datos analizados: 347.65 MB
<b>Resultados (1)</b>	
Filas de búsqueda	
# ▾   mediciones_asturias	<input type="button" value="Copiar"/> <input type="button" value="Descargar resultados en formato CSV"/>
1 272023	< 1 > ⚙ ▾

### Consulta 1 :

```
1 SELECT COUNT(*) AS mediciones_asturias
2 FROM espparq_parquet
3 WHERE id IN ('SPE00119792', 'SPE00119801', 'SPE00119819', 'SPE00119828');
```

Resultados de la consulta	Estado de la consulta
<input checked="" type="checkbox"/> Completado	Tiempo en cola: 105 ms Tiempo de ejecución: 5.149 sec Datos analizados: 19.64 GB
<b>Resultados (1)</b>	
Filas de búsqueda	
# ▾   mediciones_asturias	<input type="button" value="Copiar"/> <input type="button" value="Descargar resultados en formato CSV"/>
1 497644	< 1 > ⚙ ▾

## 3.) ¿Cuántas mediciones tenemos de Oviedo?

### Consulta 1 :

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

Resultados de la consulta | Estado de la consulta

Completo | Tiempo en cola: 105 ms | Tiempo de ejecución: 1.175 sec | Datos analizados: 347.65 MB

**Resultados (1)**

Filas de búsqueda

# | mediciones\_oviedo

1	73047
---	-------

Consulta 1 :

```

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

```

Resultados de la consulta | Estado de la consulta

Completo | Tiempo en cola: 102 ms | Tiempo de ejecución: 4.146 sec | Datos analizados: 16.94 GB

**Resultados (1)**

Filas de búsqueda

# | mediciones\_oviedo

1	134966
---	--------

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

Consulta 1 :

```

1 SELECT MIN(date) AS fecha_mas_antigua
2 FROM espespana;

```

Resultados de la consulta | Estado de la consulta

Completo | Tiempo en cola: 111 ms | Tiempo de ejecución: 1.047 sec | Datos analizados: 347.65 MB

**Resultados (1)**

Filas de búsqueda

# | fecha\_mas\_antigua

1	18961101
---	----------

Consulta 1 :

```

1 SELECT MIN(fecha) AS fecha_mas_antigua
2 FROM espparq_parquet;

```

Resultados de la consulta	Estado de la consulta
<input checked="" type="checkbox"/> Completado	Tiempo en cola: 171 ms Tiempo de ejecución: 12.598 sec Datos analizados: 1.29 GB
<b>Resultados (1)</b>	
Filas de búsqueda	
#	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
<input checked="" type="checkbox"/> Completado	Tiempo en cola: 112 ms Tiempo de ejecución: 2.99 sec Datos analizados: 347.65 MB
<b>Resultados (1)</b>	
Filas de búsqueda	
#	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
<input checked="" type="checkbox"/> Completado	Tiempo en cola: 107 ms Tiempo de ejecución: 3.201 sec Datos analizados: 17.62 GB
<b>Resultados (1)</b>	
Filas de búsqueda	
#	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