

PR_08.1 Dani Gayol Rodríguez

PR_08.1 Dani Gayol Rodríguez.....	1
Apartado A.....	1
1.) Desde Aws CLI explora el contenido del bucket s3://noaa-ghcn-pds/csv/.....	2
2.) Descarga uno cualquiera de los archivos que contiene en cada una de sus carpetas y muestra las primeras líneas de ellos.....	2
3.) ¿Qué contiene cada uno de los dos tipos de archivos?	3
Apartado B.....	4
1.) Crea una base de datos en AWE GLUE llamada clima.	5
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.	7
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.....	11
4.) ¿Está particionada la tabla? ¿Por qué campos?	13
Apartado C	13
1.) ¿Cuántos registros tiene la tabla?.....	16
2.) ¿Cuántas mediciones tenemos de España?	16
3.) Sabiendo los códigos de las 4 estaciones de Asturias ¿Cuántas mediciones tenemos de Asturias?	17
4.) ¿Cuántas mediciones tenemos de Oviedo?.....	17
5.) ¿Cuál es la medición más antigua de España, Asturias y Oviedo?	18

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 líneas del archivo “1900.csv” utilizando este comando (al estar en “Windows CMD” el comando “head” no existe así que 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 ghcn-stations section below for an explanation
- 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



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](#)

Database details

Name

clima

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.

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



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.

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).

[View](#) [Edit](#)

[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

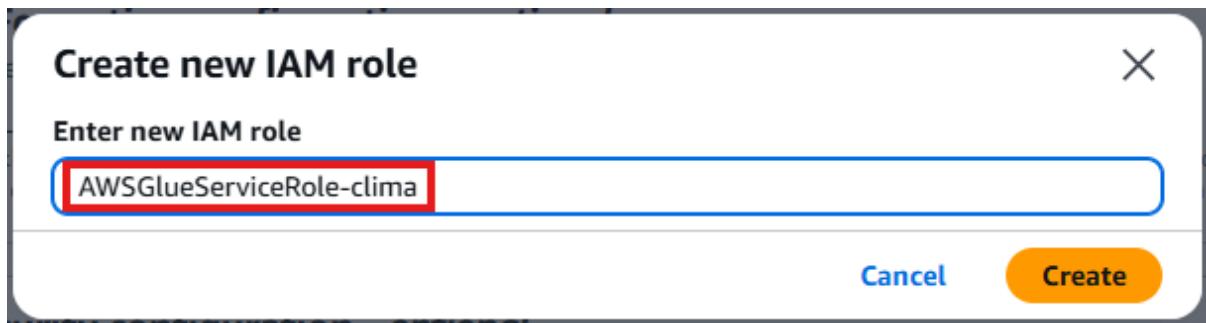
Existing IAM role

Choose an IAM role [View](#) [Edit](#)

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

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.

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 output and scheduling

Output configuration [Info](#)

Target database [clima](#) [C](#)

Clear selection Add database [L](#)

Table name prefix - *optional* [ghcn_](#)

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 auto 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](#)

Finalmente, la configuración nos quedaría de la siguiente manera:

Review and create

Step 1: Set crawler properties

Set crawler properties

Name	Description	Tags
crawler-clima	-	-

[Edit](#)

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://noaa-ghcn-pds/csv/	Recrawl all

[Edit](#)

Step 3: Configure security settings

Configure security settings

IAM role	Security configuration	Lake Formation configuration
LabRole	-	-

[Edit](#)

Step 4: Set output and scheduling

Set output and scheduling

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

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

One crawler successfully created

The following crawler is now created: "crawler-clima"

Last updated (UTC)

January 15, 2026 at 09:08:55

[Run crawler](#)[Edit](#)[Delete](#)

crawler-clima

Crawler properties

Name	IAM role	Database	State
crawler-clima	LabRole Edit	clima	READY
Description	Security configuration	Lake Formation configuration	Table prefix
-	-	-	ghcn_

Advanced settings

Crawler runs (1)

The list of crawler runs for this crawler.

[Filter data](#)[Filter by a date and time range](#)[Stop run](#)[View CloudWatch logs](#)[View run details](#)

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
January 15, 2026 at 09:10:18	January 15, 2026 at 09:17:23	07 min 05 s	Completed	0.374	1 table change, 2 partition changes

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

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

Databases (1/1)

A database is a set of associated table definitions, organized into a logical group.

<input checked="" type="checkbox"/> Name	Description
<input checked="" type="checkbox"/> clima	-

Una vez dentro nos aparecerá lo siguiente:

ghcn_csv

Last updated (UTC)
January 15, 2026 at 09:21:0

[Table overview](#) | [Data quality - new](#)

Table details

Name ghcn_csv	Classification CSV	Deprecated -
Database clima	Location s3://noaa-ghcn-pds/csv/	Column statistics No statistics
Description -	Connection -	
Last updated January 15, 2026 at 09:17:23		

En el apartado de “esquema” no aparece el esquema de las tablas detectadas

[Schema](#) | [Partitions](#) | [Indexes](#) | [Column statistics - new](#)

Schema (9)

View and manage the table schema.

#	Column name	Data type	Partition key	Comment
1	id	string	-	-
2	date	bigint	-	-
3	element	string	-	-
4	data_value	bigint	-	-
5	m_flag	string	-	-
6	q_flag	string	-	-
7	s_flag	string	-	-
8	obs_time	bigint	-	-
9	partition_0	string	Partition (0)	-

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

All column statistics runs (1)

Last updated (UTC) [\(C\)](#)
January 15, 2026 at 10:23:50

[Find run](#)

Run ID	Status	Start time (UTC)	End time (UTC)	Duration	Selected columns
d61fa6fb-7a5e-4e9b-aa7d-41a6265c	Succeeded	January 15, 2026 at 09:28:30	January 15, 2026 at 10:08:00	39 min 30 s	All columns

Column statistics (8) [Info](#)

Last updated (UTC) [\(C\)](#)
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

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	-	4705384759	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

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

The screenshot shows the AWS search interface. In the top navigation bar, the search bar contains the text "Athena". Below the search bar, the results are displayed under the heading "Servicios". A single result, "Athena", is shown in a highlighted box. The "Athena" entry includes a purple icon, the text "Athena Servicio de análisis interactivo sin servidor", and a star icon.

Análisis

Amazon Athena

Comience a consultar datos al instante.

Amazon Athena es un servicio de consulta interactiva que facilita el análisis de datos en Amazon S3 y otros orígenes de datos federados mediante SQL estándar.

Empezar

- Consulte sus datos con Amazon SageMaker Unified Studio
Ejecute y programe consultas en un único entorno de desarrollo de datos e IA.
- Consulte sus datos con la consola de Athena
Use el editor de consultas para analizar datos en S3, en las instalaciones o en otras nubes.

Iniciar el editor de consultas

Cómo funciona

Precios
Región: US East (N. Virginia)
[Calculadora de precios](#)

Ahora vamos a configurarlo rápidamente

The screenshot shows the "Configuración de consultas" (Query Configuration) page. At the top, there are tabs: "Editor", "Consultas recientes", "Consultas guardadas", and "Configuración de consultas", with "Configuración de consultas" being the active tab. To the right, there is a dropdown for "Grupo de trabajo" set to "primary".

Configuración de consultas
La configuración de consultas anula temporalmente los ajustes del grupo de trabajo. Sin embargo, para aplicar dichos ajustes, debe borrar toda la configuración.

Cifrado de resultados de consultas	Cifrado de los resultados de la consulta	Propietario previsto del bucket	Administrador
Ubicación del resultado de la consulta	Cifrado de los resultados de la consulta	Propietario previsto del bucket	Asigna al propietario del bucket el control total sobre los resultados de la consulta Desactivado

Administre la ubicación y codificación de los resultados de la consulta

X

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 X

View ↗

Browse S3

- ✖ Unable to verify if the selected bucket belongs to your current region.

X

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

Datos C <

Origen de datos
AwsDataCatalog

Catálogo
Ningún elemento

Base de datos
clima

Tablas y vistas Crear ▾ ⚙

🔍 Filtrar tablas y vistas

▼ **Tablas (1)** < 1 >

✚ **ghcn_csv** Particionado (Estadísticas) ⋮

▶ **Vistas (0)** < 1 >

1.) ¿Cuántos registros tiene la tabla?

✓ **Consulta 1 :**

```
1 SELECT COUNT(*) AS total_registros
2 FROM ghcn_csv;
```

Resultados (1)

🔍 Filas de búsqueda

#	total_registros
1	6337278525

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 :

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

Resultados (1)

Filas de búsqueda

#	▼	fecha_mas_antigua_asturias
1		19381001

Oviedo:

✓ Consulta 1 :

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

Resultados (1)

Filas de búsqueda

#	▼	fecha_mas_antigua_oviedo
1		19721201