

Glue Jobs Triggers

1) APARTADO A

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

Creemos el bucket “pablomr-meteostations”

Buckets de uso general Todas las regiones de AWS **Buckets de directorio**

Buckets de uso general (1/3) Información Copiar ARN Vaciar Eliminar Crear bucket

Los buckets son contenedores de datos almacenados en S3.

Q. Buscar buckets por nombre

Nombre	Región de AWS	Fecha de creación
amazon-pablomr	EE.UU. Este (Norte de Virginia) us-east-1	10 Dec 2025 11:26:31 AM CET
pablomr-meteostations	EE.UU. Este (Norte de Virginia) us-east-1	15 Jan 2026 10:01:17 AM CET
ventas-pablo-18	EE.UU. Este (Norte de Virginia) us-east-1	18 Dec 2025 8:51:27 AM CET

```
C:\Users\Mañana>aws s3 ls s3://pablomr-meteostations/clima/espana/
2026-01-15 10:04:56 0
```

- 2) Mediante un comando AWS CLI, copia los archivos csv con las mediciones de todas las estaciones meteorológicas de España en él. Recuerda que la ruta era: s3://noaa-ghcn-pds/csv/by_station/. Y que los ID's (nombre de fichero, por tanto) de las estaciones de España comenzaba por “SP”.

Usamos este comando:

```
aws s3 cp "s3://noaa-ghcn-pds/csv/by_station/" "s3://pablomr-meteostations/clima/espana/" --recursive --exclude "*" --include "SP*"
```

Objetos (207) Copiar URI de S3 Copiar URL Descargar Abrir L2 Eliminar Acciones 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](#)

Q. Buscar objetos por prefijo

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
SP000003195.csv	csv	19 Jan 2026 9:09:11 AM CET	3.8 MB	Estándar
SP000004452.csv	csv	19 Jan 2026 9:09:11 AM CET	3.2 MB	Estándar
SP000006155.csv	csv	19 Jan 2026 9:09:11 AM CET	3.6 MB	Estándar
SP000007038.csv	csv	19 Jan 2026 9:09:11 AM CET	4.0 MB	Estándar
SP000008027.csv	csv	19 Jan 2026 9:09:11 AM CET	4.2 MB	Estándar
SP000008181.csv	csv	19 Jan 2026 9:09:11 AM CET	3.8 MB	Estándar
SP000008202.csv	csv	19 Jan 2026 9:09:11 AM CET	3.5 MB	Estándar
SP000008215.csv	csv	19 Jan 2026 9:09:11 AM CET	3.5 MB	Estándar
SP000008280.csv	csv	19 Jan 2026 9:09:11 AM CET	4.4 MB	Estándar
SP000008410.csv	csv	19 Jan 2026 9:09:11 AM CET	3.0 MB	Estándar
SP000008416.csv	csv	19 Jan 2026 9:09:11 AM CET	3.3 MB	Estándar
SP000009434.csv	csv	19 Jan 2026 9:09:11 AM CET	2.8 MB	Estándar
SP000009981.csv	csv	19 Jan 2026 9:09:11 AM CET	4.9 MB	Estándar
SP000060010.csv	csv	19 Jan 2026 9:09:11 AM CET	4.0 MB	Estándar
SP000060040.csv	csv	19 Jan 2026 9:09:17 AM CET	2.7 MB	Estándar

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

espana

Database properties

Name	Description
espana	-

Tables (0)

View and manage all available tables.

 Filter tables

<input type="checkbox"/>	Name	Database	Location
--------------------------	------	----------	----------

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

Review and create

Step 1: Set crawler properties

Edit

Set crawler properties

Name	Description	Tags
estaciones_españolas_guardar	-	-

Step 2: Choose data sources and classifiers

Edit

Data sources (1) [Info](#)

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

Type	Data source	Parameters
S3	s3://pablomr-meteostations	Recrawl all

Step 3: Configure security settings

Edit

Configure security settings

IAM role	Security configuration	Lake Formation configuration
LabRole	-	-

Step 4: Set output and scheduling

Edit

Set output and scheduling

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

- 5) Guarda el Crawler, pero no lo ejecutes.

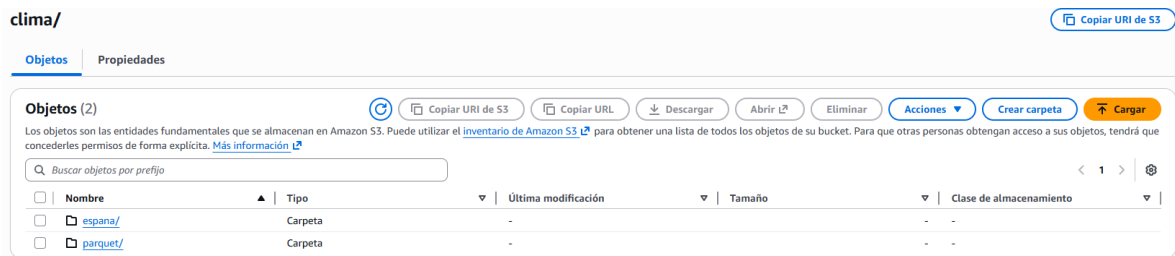
Crawlers (2) [Info](#)
 Last updated (UTC)
January 19, 2026 at 08:21:

View and manage all available crawlers.

<input type="text"/>	Name	State	Schedule	Last run	Last run timestamp	Log
<input type="checkbox"/>	estaciones_españolas_guar...	Ready		-	-	-
<input type="checkbox"/>	PabloMRMeteorologo	Ready		Succeeded	January 14, 2026 at 10:55:...	View log

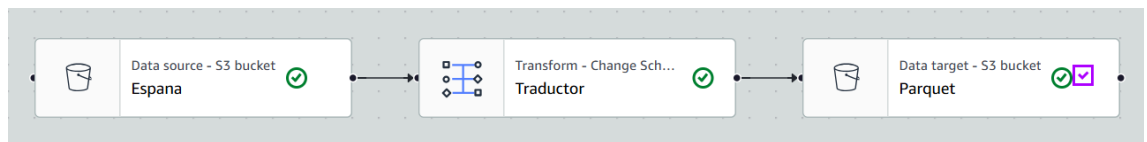
2) APARTADO B

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



- 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.

Creamos este esquema:



Nodo “Espana”:

Name: Espana

S3 source type: S3 location

S3 URL: s3://pablomr-meteostations/clima/espana/

Recursive: ☒ Read files in all subdirectories.

Data format: CSV

Delimiter: Comma (,)

Nodo “Traductor”:

Name: Traductor

Node parents: Espana (S3 - DataSource)

Change Schema (Apply mapping)

Source key	Target key	Data type	Drop
id	id	string	<input type="checkbox"/>
date	fecha	string	<input type="checkbox"/>
element	elemento	string	<input type="checkbox"/>
data_value	valor_dato	string	<input type="checkbox"/>
m_flag	m_bandera	string	<input type="checkbox"/>
q flag	q bandera	string	<input type="checkbox"/>

Nodo “Parquet”:

Name
Parquet

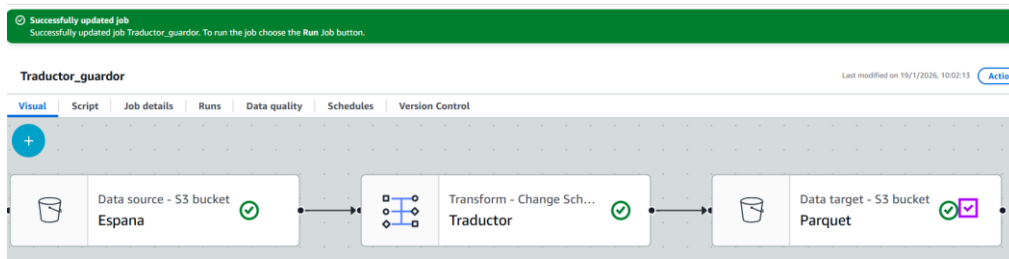
Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node
Traductor
ApplyMapping - Transform

Format
Parquet

Compression Type
Snappy

S3 Target Location
Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).
s3://pablomr-meteostations/clima/parquet/ View Browse S3

3) Guarda el trabajo, pero no lo ejecutes.



3) 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_.

Review and create

Step 1: Set crawler properties Edit

Set crawler properties

Name	Description	Tags
parquet_a_base	-	-

Step 2: Choose data sources and classifiers Edit

Data sources (1) Info
The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
S3	s3://pablomr-meteostations/clima/parquet/	Recrawl all

Step 3: Configure security settings Edit

Configure security settings

IAM role	Security configuration	Lake Formation configuration
LabRole	-	-

Step 4: Set output and scheduling Edit

Set output and scheduling

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

2) Guarda el rastreador, pero no lo ejecutes.

Crawlers (1/3) [Info](#)

View and manage all available crawlers.

<input type="checkbox"/>	Name	State	Schedule	Last run
<input type="checkbox"/>	PabloMRMeteorologo	✓ Ready		✓ Succeeded
<input type="checkbox"/>	estaciones_españolas_guar...	✓ Ready		-
<input checked="" type="checkbox"/>	parquet_a_base	✓ Ready		-

4) 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.

Review and create**Step 1: Set trigger properties** [Edit](#)

Trigger details		
Name	Description	Tags
espa_ab	-	-

Watched resources (1)		
List of conditions that will start the trigger		
Type	Name	Status
Crawler	estaciones_españolas_guar...	✓ Succeeded

Step 2: Choose jobs or crawlers to activate [Edit](#)

Resources to trigger (1)		
List of resources to start once the trigger activates		
Type	Name	Parameters
Job	Traductor_guardar	-

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

Review and create**Step 1: Set trigger properties** [Edit](#)

Trigger details		
Name	Description	Tags
espa_bc	-	-

Watched resources (1)		
List of conditions that will start the trigger		
Type	Name	Status
Job	Traductor_guardar	✓ Succeeded

Step 2: Choose jobs or crawlers to activate [Edit](#)

Resources to trigger (1)		
List of resources to start once the trigger activates		
Type	Name	Parameters
Crawler	parquet_a_base	-

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

GLUE JOBS TRIGGERS

Review and create

Step 1: Set trigger properties

[Edit](#)

Trigger details

Name	Description	Tags
arrancar_todo	-	-

Step 2: Choose jobs or crawlers to activate

[Edit](#)

Resources to trigger (1)

List of resources to start once the trigger activates

Type	Name	Parameters
Crawler	estaciones_españolas_guardar	-

4) Arranca este manualmente este último disparador.

✓	estaciones_españolas_guar...	✓ Ready	✓ Succeeded
✓	parquet_a_base	✓ Ready	✓ Succeeded

Schema (8)

View and manage the table schema.

#	Column name	Data type
1	id	string
2	fecha	string
3	elemento	string
4	valor_dato	string
5	m_bandera	string
6	q_bandera	string
7	s_bandera	string
8	obs_tiempo	string

5) APARTADO E

Deberían de ir ejecutándose todos los trabajos y crawlers. Cuando finalicen todas las tareas tendrían que haberse creado los archivos CSV y Parquet, así como la tablas con los nombres de sus campos en español. Verifica que todo ha ido correctamente.

1) Muestra los archivos creados.

parquet/

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 obteng tendrán que concederles permisos de forma explícita. [Más información](#)

Buscar objetos por prefijo

	Nombre	Tipo	Última modificación	Tamaño	Clase de alma
<input type="checkbox"/>	run-1768896631873-part-block-0-r-00000-snappy.parquet	parquet	20 Jan 2026 9:10:46 AM CET	436.0 KB	Estándar
<input type="checkbox"/>	run-1768896631873-part-block-0-r-00001-snappy.parquet	parquet	20 Jan 2026 9:10:46 AM CET	379.2 KB	Estándar
<input type="checkbox"/>	run-1768896631873-part-block-0-r-00002-snappy.parquet	parquet	20 Jan 2026 9:10:46 AM CET	433.3 KB	Estándar
<input type="checkbox"/>	run-1768896631873-part-block-0-r-00003-snappy.parquet	parquet	20 Jan 2026 9:10:46 AM CET	428.0 KB	Estándar
<input type="checkbox"/>	run-1768896631873-part-block-0-r-00004-snappy.parquet	parquet	20 Jan 2026 9:10:46 AM CET	477.6 KB	Estándar
<input type="checkbox"/>	run-1768896631873-part-block-0-r-00005-snappy.parquet	parquet	20 Jan 2026 9:10:46 AM CET	434.7 KB	Estándar
<input type="checkbox"/>	run-1768896631873-part-block-0-r-00006-snappy.parquet	parquet	20 Jan 2026 9:10:46 AM CET	427.1 KB	Estándar
<input type="checkbox"/>	run-1768896631873-part-block-0-r-00007-snappy.parquet	parquet	20 Jan 2026 9:10:46 AM CET	437.0 KB	Estándar
<input type="checkbox"/>	run-1768896631873-part-block-0-r-00008-snappy.parquet	parquet	20 Jan 2026 9:10:47 AM CET	501.4 KB	Estándar

2) Muestra las tablas y campos creados.

Tabla “espcsv_pablomr_meteostations”:

Schema (10)

View and manage the table schema.

Filter schemas

#	Column name	Data type
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	string
9	partition_0	string
10	partition_1	string

Tabla “espparq_parquet”:

Schema (8)

View and manage the table schema.

Filter schemas

#	Column name	Data type
1	id	string
2	fecha	string
3	elemento	string
4	valor_dato	string
5	m_bandera	string
6	q_bandera	string
7	s_bandera	string
8	obs_tiempo	string

6) APARTADO F

Vete a Athena y ejecuta por duplicado (una vez sobre la tabla `espcsv_` y otra sobre la tabla `espparq_`) las mismas consultas que en la práctica anterior mostrando sus resultados y tiempos de ejecución. Obtén los tiempos obtenidos entonces y ahora sobre las dos tablas.

1) ¿Cuántas mediciones tenemos de España?

Tabla “`espcsv`”:

Completado		Tiempo en cola: 110 ms	Tiempo de ejecución: 2.032 sec	Datos analizados: 388.57 MB
Resultados (1)		Copiar Descargar resultados en formato CSV		
Filas de búsqueda		<input type="text"/>		
#	▼	▼	▼	▼
1		10583200		

Tabla “`espparq`”:

Completado		Tiempo en cola: 135 ms	Tiempo de ejecución: 1.389 sec	Datos analizados: 30.34 KB
Resultados (1)		Copiar Descargar resultados en formato CSV		
Filas de búsqueda		<input type="text"/>		
#	▼	▼	▼	▼
1		10583191		

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

Tabla “`espcsv`”:

Completado		Tiempo en cola: 117 ms	Tiempo de ejecución: 776 ms	Datos analizados: 0.20 KB
Resultados (1)		Copiar Descargar resultados en formato CSV		
Filas de búsqueda		<input type="text"/>		
#	▼	▼	▼	▼
1		80616		

Tabla “`espparq`”:

Completado		Tiempo en cola: 107 ms	Tiempo de ejecución: 728 ms	Datos analizados: 0.20 KB
Resultados (1)		Copiar Descargar resultados en formato CSV		
Filas de búsqueda		<input type="text"/>		
#	▼	▼	▼	▼
1		80616		

3) ¿Cuántas mediciones tenemos de Oviedo?

Tabla “`espcsv`”:

Completado		Tiempo en cola: 100 ms	Tiempo de ejecución: 2.623 sec	Datos analizados: 388.57 MB
Resultados (1)		Copiar Descargar resultados en formato CSV		
Filas de búsqueda		<input type="text"/>		
#	▼	▼	▼	▼
1		73047		

Tabla “`espparq`”:

GLUE JOBS TRIGGERS

Completado

Tiempo en cola: 112 ms

Tiempo de ejecución: 708 ms

Datos analizados: 0.20 KB

Resultados (1)

Copiar

Descargar resultados en formato CSV

Filas de búsqueda

#	col0
1	73047

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

- España:

Tabla “espcsv”:

Completado

Tiempo en cola: 109 ms

Tiempo de ejecución: 16.605 sec

Datos analizados: 207.05 GB

Resultados (3)

Copiar

Descargar resultados en formato CSV

Filas de búsqueda

#	id	date	element	data_value	m_flag	q_flag	s_flag	obs_time	partition_0	partition_1
1	SPE00155329	18961101	TMAX	155			E		clima	espana
2	SPE00155329	18961101	TMIN	40			E		clima	espana
3	SPE00155329	18961101	PRCP	0			E		clima	espana

Tabla “espparq”:

Completado

Tiempo en cola: 107 ms

Tiempo de ejecución: 1.23 sec

Datos analizados: 29.35 MB

Resultados (3)

Copiar

Descargar resultados en formato CSV

Filas de búsqueda

#	id	fecha	elemento	valor_dato	m_bandera	q_bandera	s_bandera	obs_tiempo
1	SPE00155329	18961101	TMAX	155			E	
2	SPE00155329	18961101	TMIN	40			E	
3	SPE00155329	18961101	PRCP	0			E	

- Asturias:

Tabla “espcsv”:

Completado

Tiempo en cola: 108 ms

Tiempo de ejecución: 2.726 sec

Datos analizados: 777.14 MB

Resultados (3)

Copiar

Descargar resultados en formato CSV

Filas de búsqueda

#	id	date	element	data_value	m_flag	q_flag	s_flag	obs_time	partition_0	partition_1
1	SPE00119801	19381001	TMAX	192			E		clima	espana
2	SPE00119801	19381001	TMIN	135			E		clima	espana
3	SPE00119801	19381001	PRCP	1			E		clima	espana

Tabla “espparq”:

Completado

Tiempo en cola: 99 ms

Tiempo de ejecución: 1.096 sec

Datos analizados: 1.21 MB

Resultados (3)

Q

Filas de búsqueda

Copiar

Descargar resultados en formato CSV

#

▼

id

▼

fecha

▼

elemento

▼

valor_dato

▼

m_bandera

▼

q_bandera

▼

s_bandera

▼

obs_tiempo

▼

1

SPE00119801

19381001

TMAX

192

E

2

SPE00119801

19381001

TMIN

135

E

3

SPE00119801

19381001

PRCP

1

E

- Oviedo:

Tabla “espcsv”:

GLUE JOBS TRIGGERS

Completado

Tiempo en cola: 118 ms

Tiempo de ejecución: 2.081 sec

Datos analizados: 777.14 MB

Resultados (3)

Copiar

Descargar resultados en formato CSV

Filas de búsqueda

#	id	date	element	data_value	m_flag	q_flag	s_flag	obs_time	partition_0	partition_1
1	SPE00119828	19721201	TMAX	130			E		clima	espana
2	SPE00119828	19721201	TMIN	38			E		clima	espana
3	SPE00119828	19721201	PRCP	0			E		clima	espana

Tabla “espparq”:

Completado

Tiempo en cola: 97 ms

Tiempo de ejecución: 1.245 sec

Datos analizados: 684.25 KB

Resultados (3)

Copiar

Descargar resultados en formato CSV

Filas de búsqueda

#	id	fecha	elemento	valor_dato	m_bandera	q_bandera	s_bandera	obs_tiempo
1	SPE00119828	19721201	TMAX	130			E	
2	SPE00119828	19721201	TMIN	38			E	
3	SPE00119828	19721201	PRCP	0			E	

- 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?

Claramente es más rápida “espparq” y por detrás “espcsv”. “csv” es increíblemente lenta.

COMPARATIVA	csv	espcsv	espparq
Consulta 1	16,419	2,032	1,389
Consulta 2	14,699	0,776	0,728
Consulta 3	12,516	2,623	0,708
Consulta 4.1 (ES)	29,629	16,605	1,23
Consulta 4.2 (AS)	29,153	2,726	1,096
Consulta 4.3 (OV)	35,196	2,081	1,245
Media	22,94	4,47	1,07