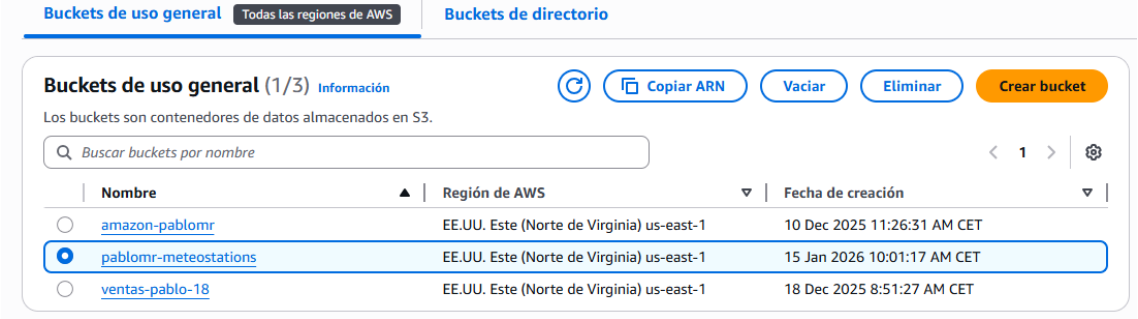
Glue Jobs Triggers

1) APARTADO A

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

Creamos el bucket “pablomr-meteostations”

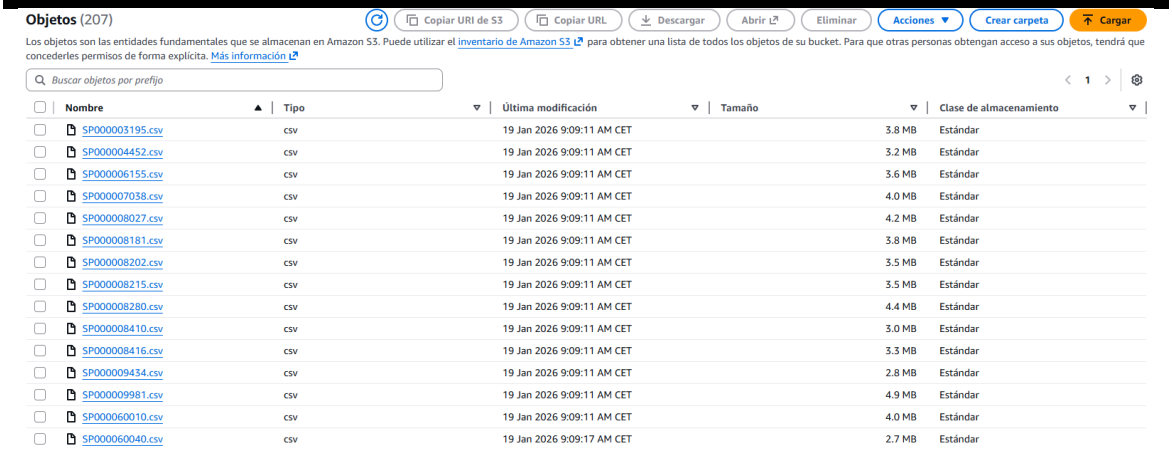



```
C:\Users\Mañana>aws s3 ls s3://pablomr-meteostations/clima/espana/
2026-01-15 10:04:56          0
```

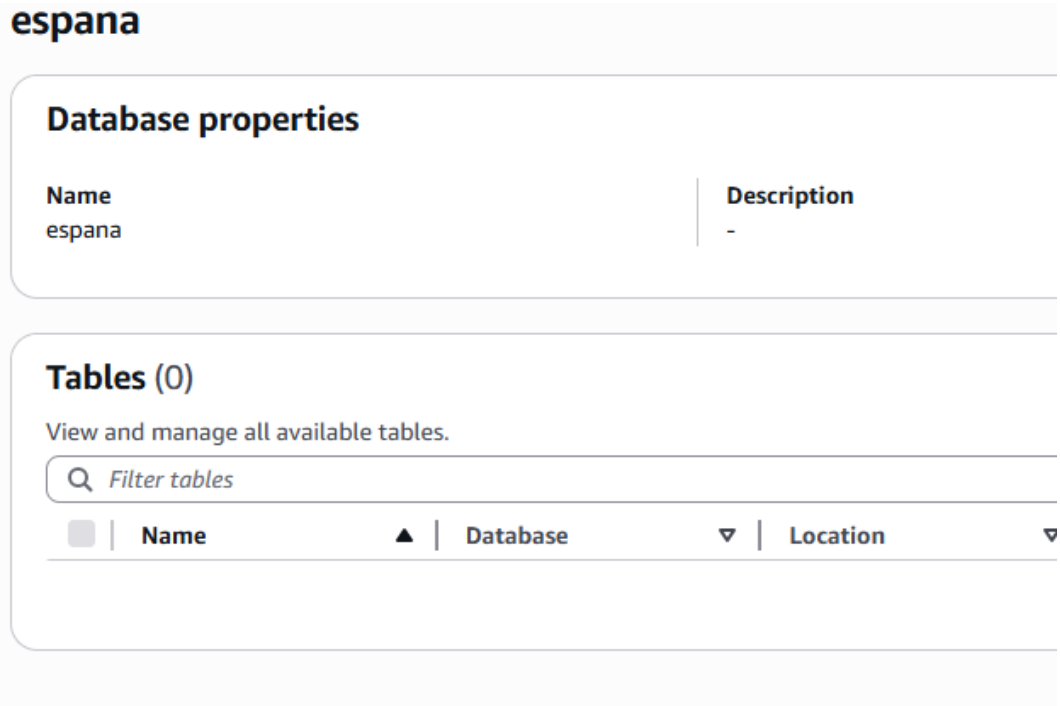
2) Mediante un comando AWS CLI, copia los archivos csv con las mediciones de todas las estaciones meteorológicas de España en él. Recuerda que la ruta era: s3://noaa-ghcn-pds/csv/by_station/). Y que los ID’s (nombre de fichero, por tanto) de las estaciones de España comenzaba por “SP”.

Usamos este comando:

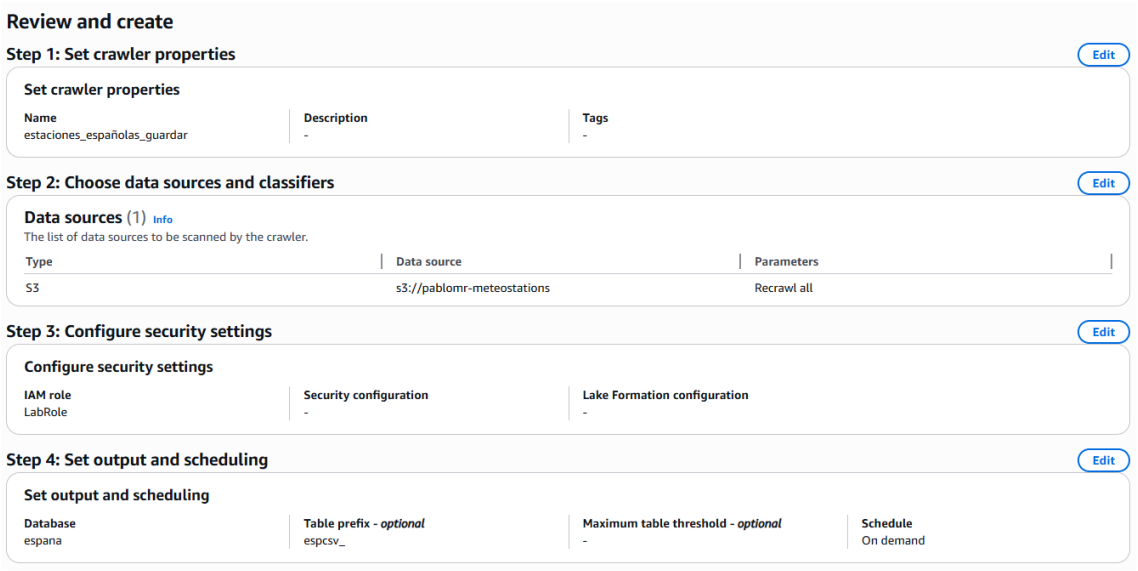
```
aws s3 cp "s3://noaa-ghcn-pds/csv/by_station/" "s3://pablomr-
meteostations/clima/espana/" --recursive --exclude "*" --include "SP*"
```



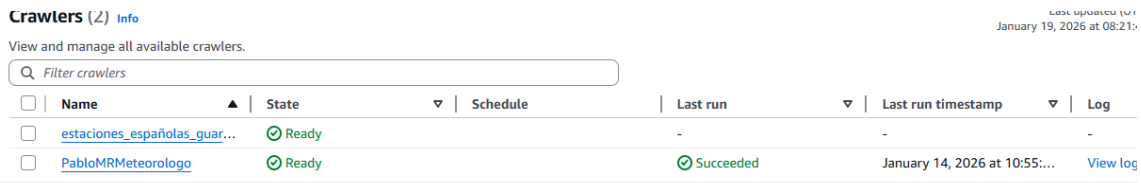
3) Crea una base de datos en del Data Catalog que se llame espana.



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_).

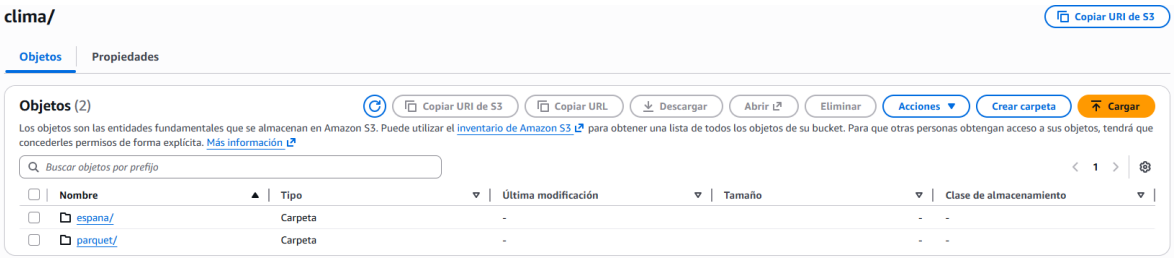


5) Guarda el Crawler, pero no lo ejecutes.



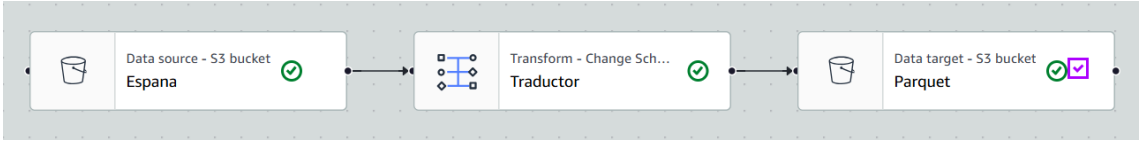
2) APARTADO B

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

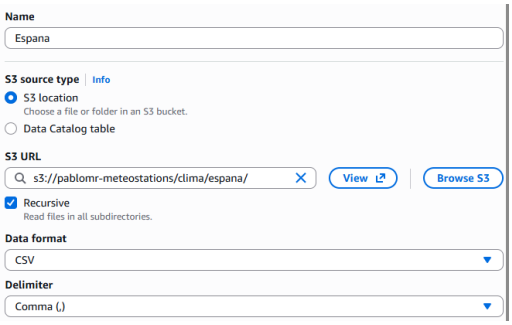


2) Crea un trabajo mediante Visual ETL que nos permita cambiar el esquema de los CSV's que acabamos de importar poniendo los nombres de los campos en español y guardando los datos en formato parquet en la carpeta del punto anterior.

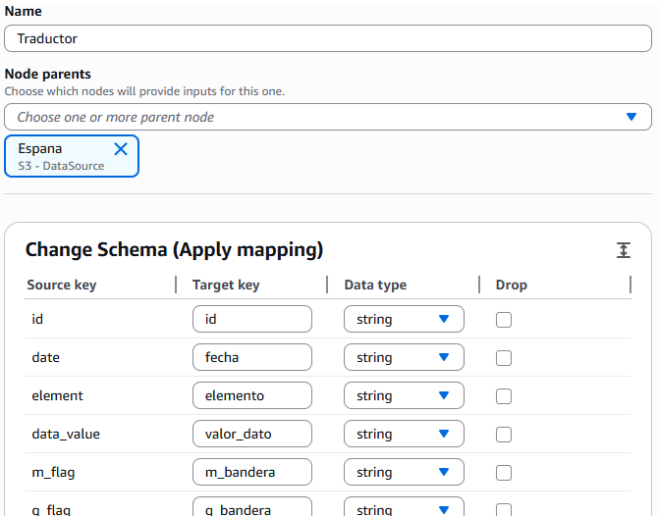
Creamos este esquema:



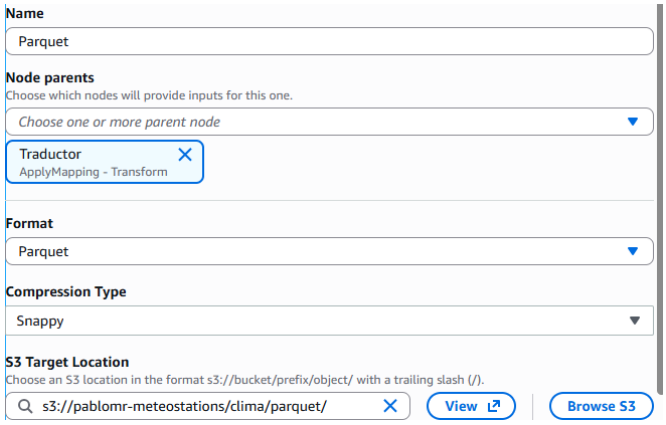
Nodo "Espana":



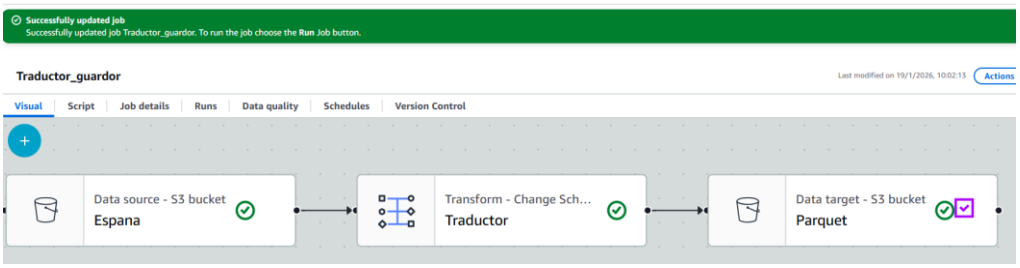
Nodo "Traductor":



Nodo “Parquet”:

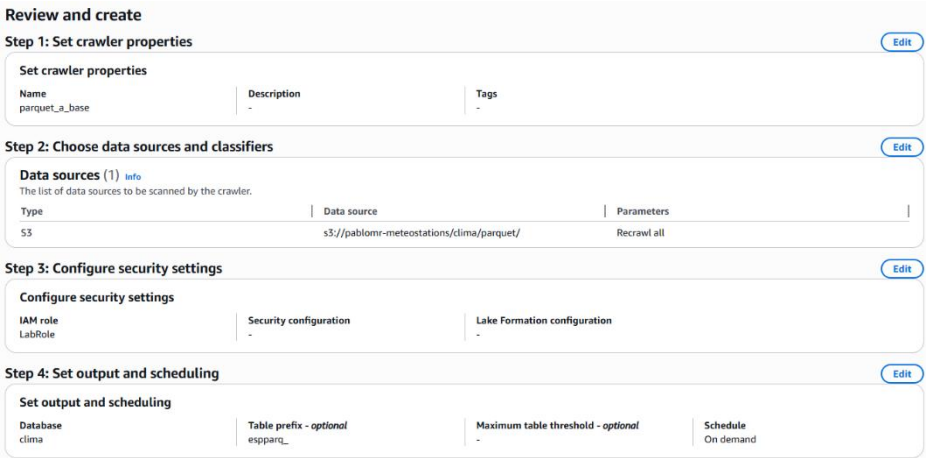


3) Guarda el trabajo, pero no lo ejecutes.

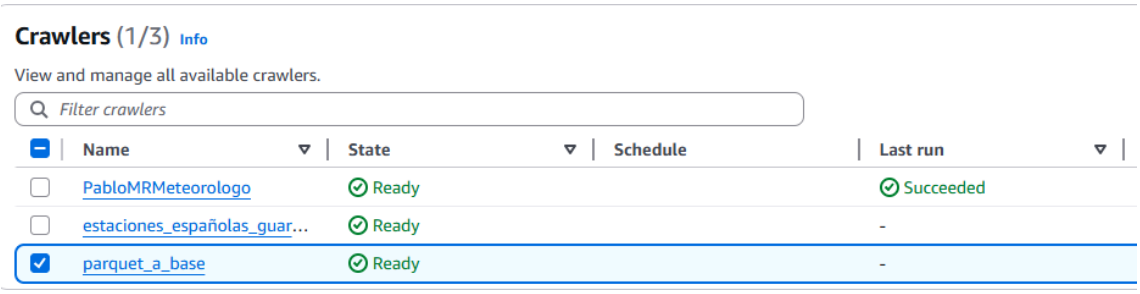


3) APARTADO C

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_.

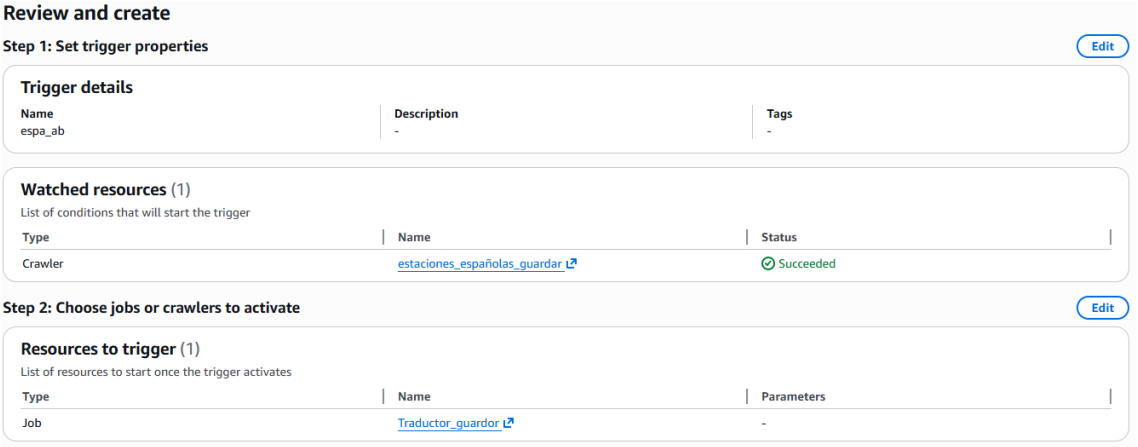


2) Guarda el rastreador, pero no lo ejecutes.

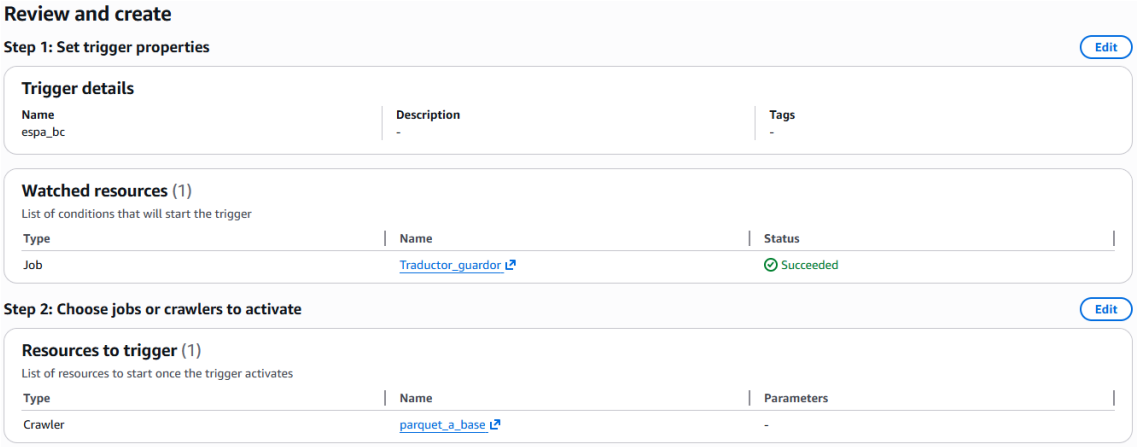


4) APARTADO D

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

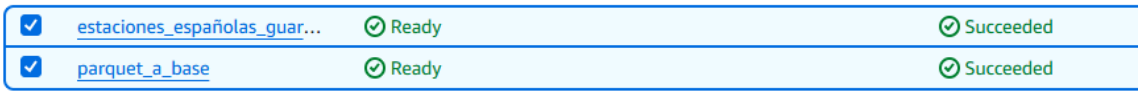


2) Crea un disparador (trigger) -puedes llamarlo espa_bc - que después de finalizado el trabajo del apartado B lance el trabajo del apartado C.



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

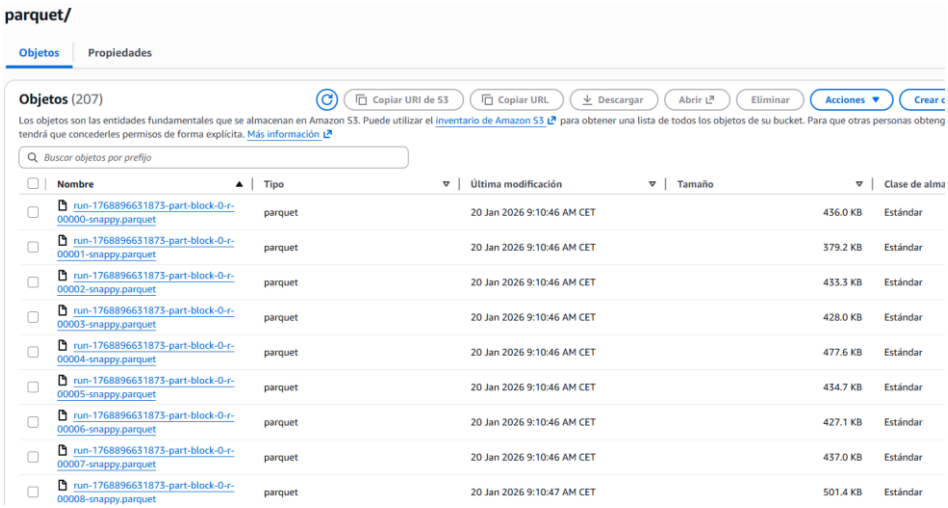
4) Arranca este manualmente este último disparador.



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.



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”:

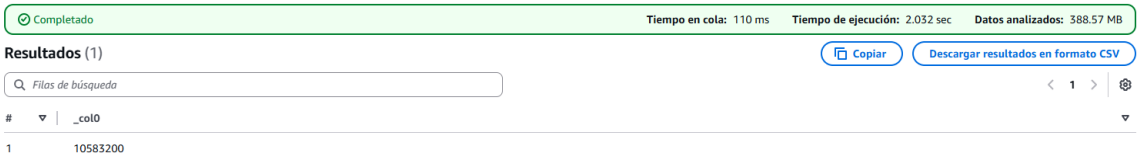
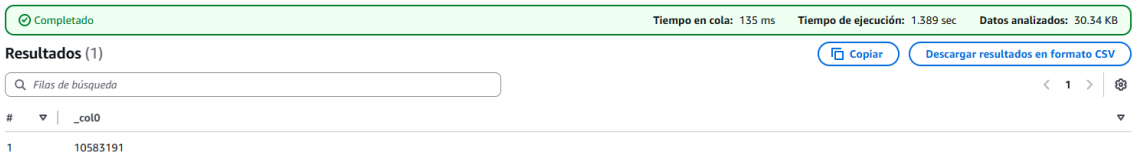


Tabla “espparq ”:



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

Tabla “espcsv”:

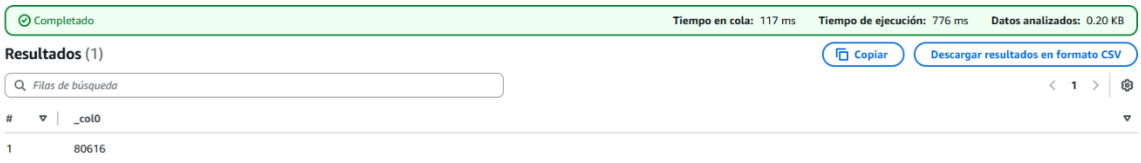
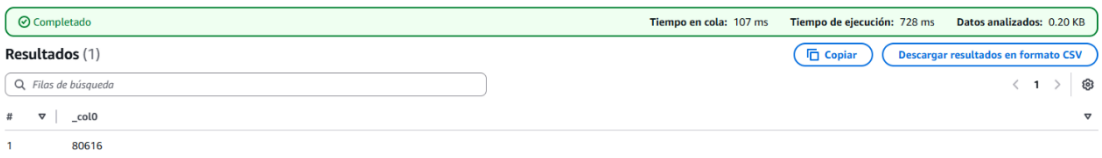


Tabla “espparq ”:



3) ¿Cuántas mediciones tenemos de Oviedo?

Tabla “espcsv”:

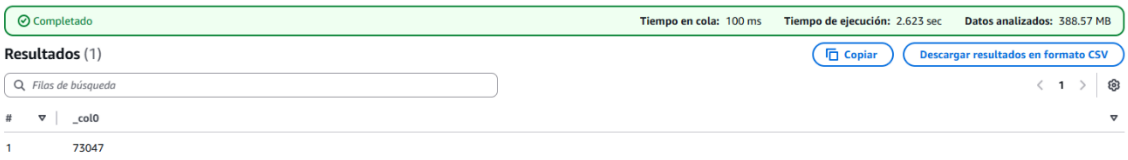
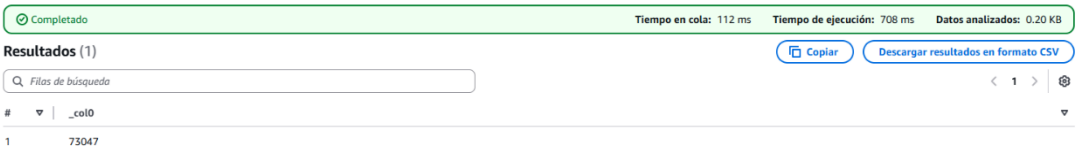


Tabla “espparq ”:



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

- España:

Tabla “espcsv”:

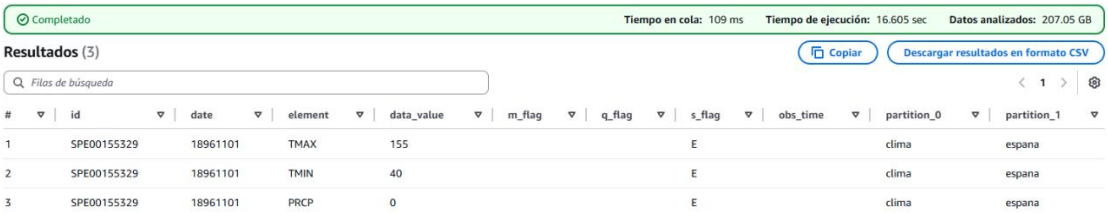
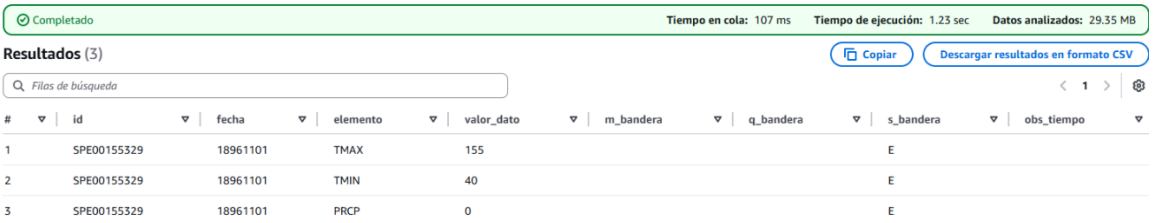


Tabla “espparq ”:



- Asturias:

Tabla “espcsv”:

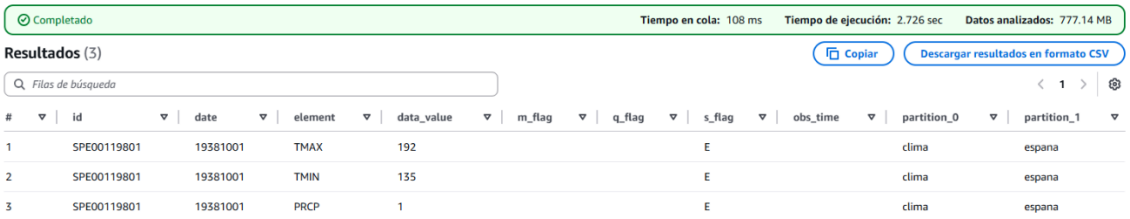
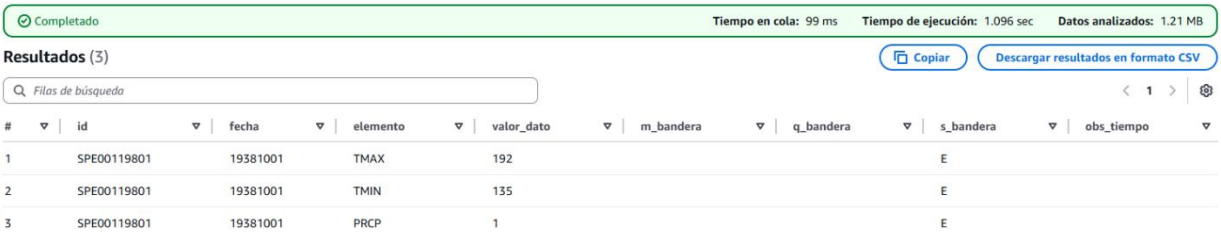


Tabla “espparq ”:



- Oviedo:

Tabla “espcsv”:

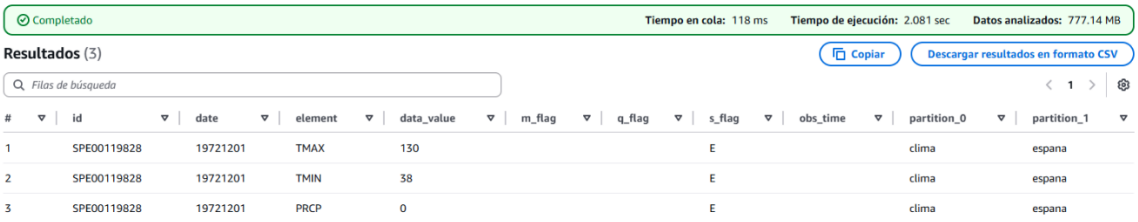
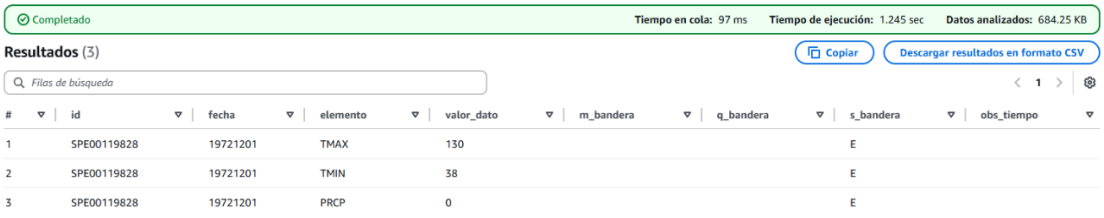


Tabla “espparq ”:



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