Glue Crawler

1) Apartado A

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

```
PS C:\Users\Mañana> 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.

```
2026-01-13 18:42:31    1358655 ASN00075014.csv
2026-01-13 18:42:30     511831 ASN00075015.csv
2026-01-13 18:42:30     260207 ASN00075016.csv
2026-01-13 18:42:31     221954 ASN00075017.csv
2026-01-13 18:42:31    1397669 ASN00075018.csv

PS C:\Users\Mañana> aws s3 cp s3://noaa-ghcn-pds/csv/by_station/ASN00075013.csv "C:\Users\Mañana\Desktop\metereologia"
download: s3://noaa-ghcn-pds/csv/by_station/ASN00075013.csv to Desktop\metereologia\ASN00075013.csv
```

```
2026-01-13 18:27:30 1327477186 2018.csv
2026-01-13 18:17:47 1315282445 2019.csv
2026-01-13 18:57:11 1328605862 2020.csv
2026-01-13 18:37:24 1362730612 2021.csv
2026-01-13 19:07:09 1363495913 2022.csv
2026-01-13 19:18:35 1365842617 2023.csv
2026-01-13 19:20:59 1328562850 2024.csv
2026-01-13 18:17:04 1157539420 2025.csv
2026-01-13 18:15:15   19228431 2026.csv
PS C:\Users\Mañana> aws s3 cp s3://noaa-ghcn-pds/csv/by_year/2025.csv "C:\Users\Mañana\Desktop\metereologia"
download: s3://noaa-ghcn-pds/csv/by_year/2025.csv to Desktop\metereologia\2025.csv
```

Por estación:

```
Mañana@A26P42 MINGW64 ~/Desktop/metereologia
$ head ASN00075013.csv
ID,DATE,ELEMENT,DATA_VALUE,M_FLAG,Q_FLAG,S_FLAG,OBS_TIME
ASN00075013,18860101,PRCP,0,,,a,
ASN00075013,18860102,PRCP,0,,,a,
ASN00075013,18860103,PRCP,0,,,a,
ASN00075013,18860104,PRCP,0,,,a,
ASN00075013,18860105,PRCP,0,,,a,
ASN00075013,18860106,PRCP,0,,,a,
ASN00075013,18860107,PRCP,269,,,a,
ASN00075013,18860108,PRCP,0,,,a,
ASN00075013,18860109,PRCP,0,,,a,
```

Por año:

```
Mañana@A26P42 MINGW64 ~/Desktop/metereologia
$ head 2025.csv
ID,DATE,ELEMENT,DATA_VALUE,M_FLAG,Q_FLAG,S_FLAG,OBS_TIME
IN002050100,20250101,TMIN,140,,,S,
IN002050100,20250101,TAVG,196,H,,S,
IN003010600,20250101,TMAX,264,,,S,
IN003010600,20250101,TMIN,146,,,S,
IN003010600,20250101,TAVG,203,H,,S,
IN003020100,20250101,TMIN,124,,,S,
IN003020100,20250101,TAVG,181,H,,S,
IN003030100,20250101,TMIN,139,,,S,
IN003030100,20250101,TAVG,162,H,,S,
```

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

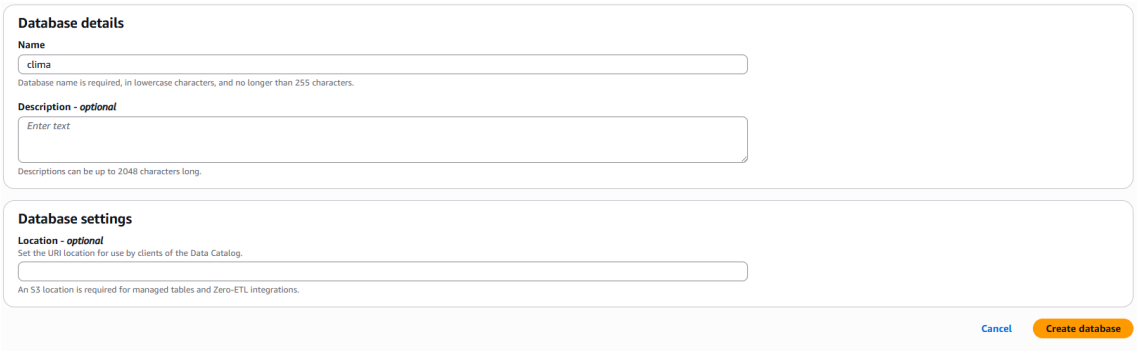
Se muestra la información de datos de estaciones meteorológicas. Uno muestra los datos recogidos por las propias estaciones y el otro según año.

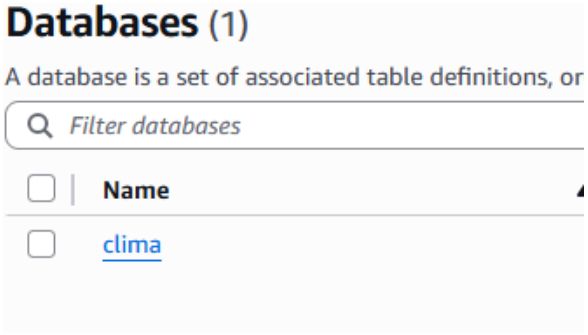
Los atributos guardados en ambos son los mismos:

- ID: Identifica la entrada de datos.
- DATE: Fecha del dato.
- ELEMENT: Indica el tipo de elemento.
- DATA_VALUE: El valor del elemento.
- M-FLAG: Marca de mediciones.
- Q-FLAG: Marca de calidad.
- S-FLAG: Marca de fuente.
- OBS-TIME: Hora de la muestra.

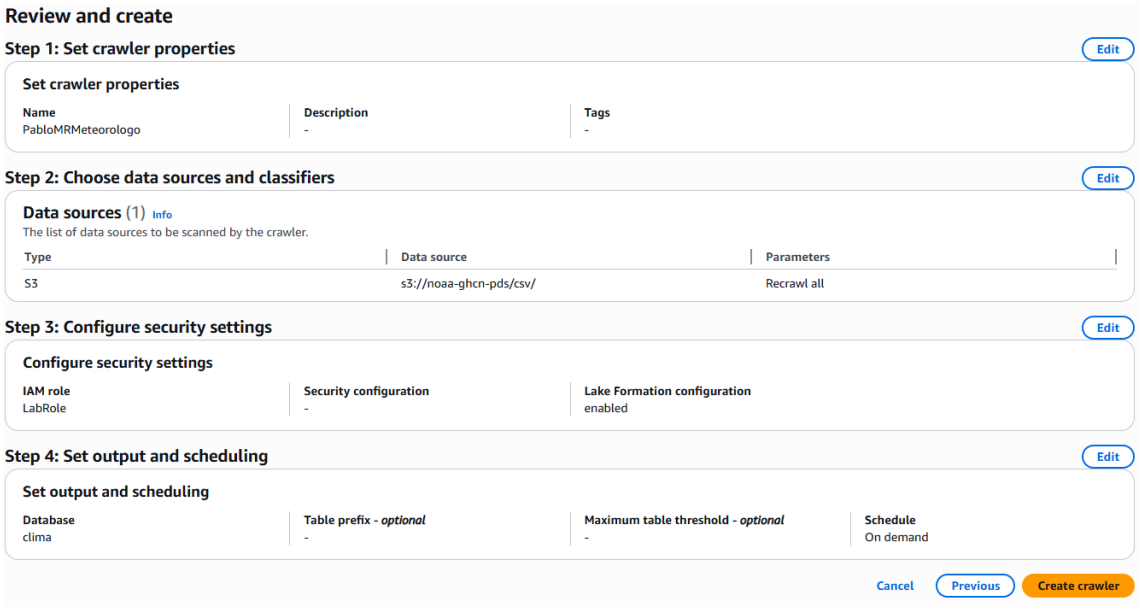
2) Apartado B

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

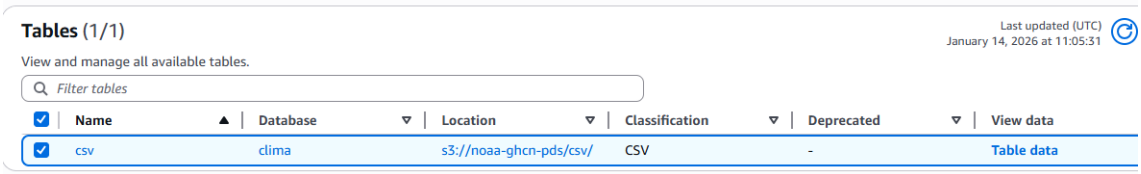




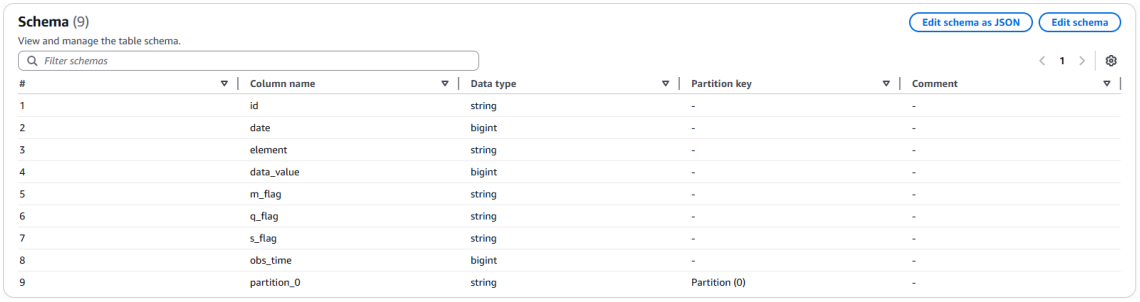
2) Crea un Crawler AWS GLUE que nos explore el bucket del ejercicio anterior generando las tablas en la base de datos que acabas de crear.



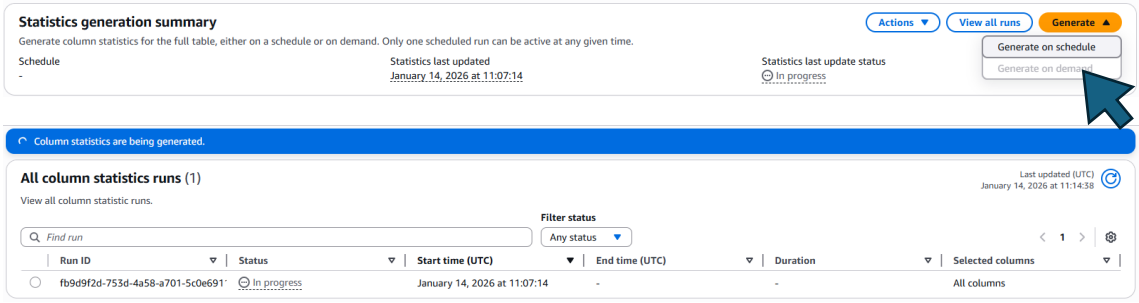
Ha generado esta tabla.



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.

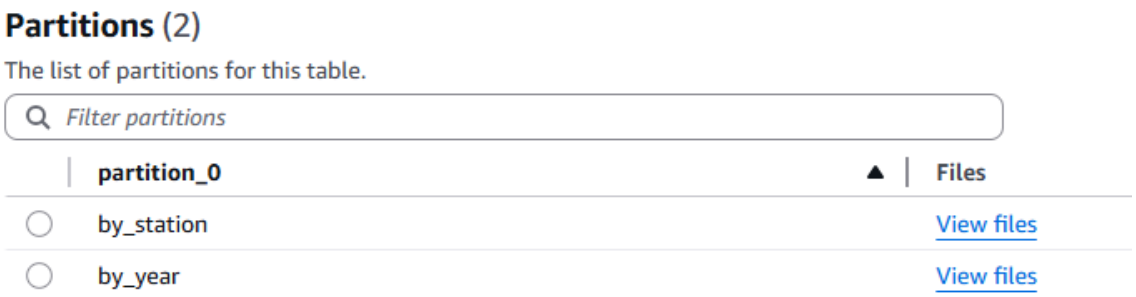


Generamos el resumen estadístico



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

   Sí, tiene dos particiones. Por los campos “by_station” y “by_year” provenientes de la división del dataset.

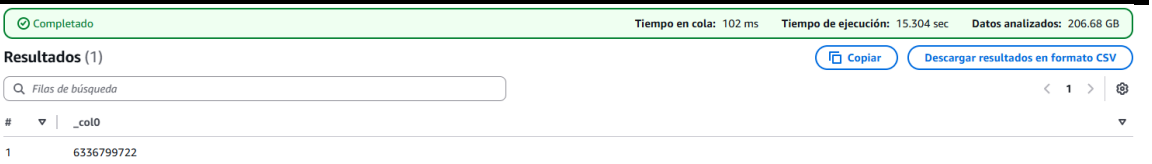


3) Apartado C

Desde ATHENA, intenta realizar las siguientes consultas mostrando sus resultados y tiempos de ejecución:

1) ¿Cuántos registros tiene la tabla?

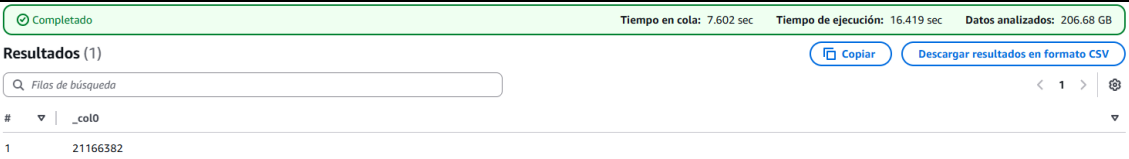
```
SELECT count(*) FROM "AwsDataCatalog"."clima"."csv";
```



2) ¿Cuántas mediciones tenemos de España?

   Para realizar esta maniobra, necesitaremos investigar la documentación. En ella descubrimos que los dos primeros caracteres del id representan el país donde se encuentra la estación, en el caso de España es “SP”.

```
SELECT count(*) FROM "AwsDataCatalog"."clima"."csv"
WHERE REGEXP_LIKE(id, '^SP');
```

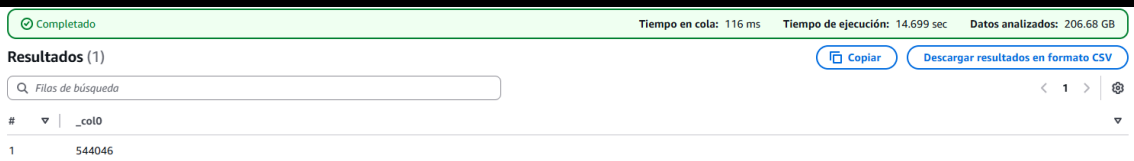


3) Sabiendo los códigos de las 4 estaciones de Asturias ¿Cuántas mediciones tenemos de Asturias?

   Los códigos son “SPE00119792”, “SPE00119801”, “SPE00119819” y “SPE00119828”

```
SELECT count(*) FROM "AwsDataCatalog"."clima"."csv"
WHERE id = 'SPE00119792'
```

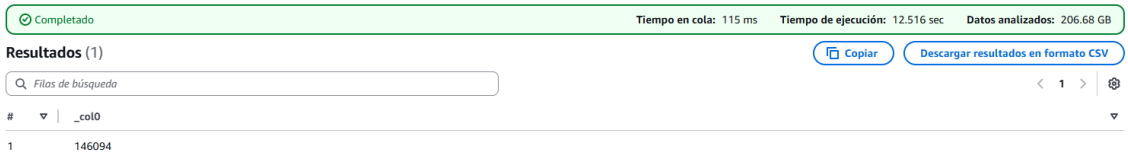
```
OR id = 'SPE00119801'
OR id = 'SPE00119819'
OR id = 'SPE00119828';
```



4) ¿Cuántas mediciones tenemos de Oviedo?

Usaremos el id “SPE00119828”

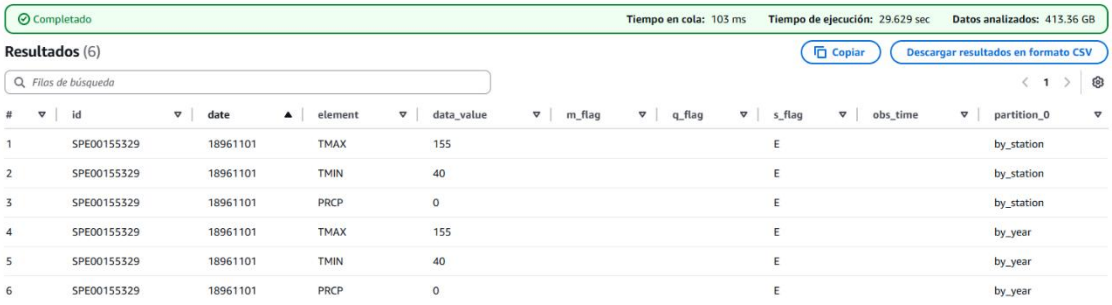
```
SELECT count(*) FROM "AwsDataCatalog"."clima"."csv"
WHERE id = 'SPE00119828';
```



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

España:

```
SELECT * FROM "AwsDataCatalog"."clima"."csv"
WHERE REGEXP_LIKE(id, '^SP')
AND "date" = (
    SELECT MIN("date") FROM "AwsDataCatalog"."clima"."csv"
    WHERE REGEXP_LIKE(id, '^SP')
);
```



Asturias:

```
SELECT * FROM "AwsDataCatalog"."clima"."csv"
WHERE (
    id = 'SPE00119792'
    OR id = 'SPE00119801'
    OR id = 'SPE00119819'
    OR id = 'SPE00119828'
)
AND "date" = (
    SELECT MIN("date") FROM "AwsDataCatalog"."clima"."csv"
    WHERE id = 'SPE00119792'
```

```
    OR id = 'SPE00119801'
    OR id = 'SPE00119819'
    OR id = 'SPE00119828'
);
```

Completado | Tiempo en cola: 119 ms | Tiempo de ejecución: 29.153 sec | Datos analizados: 413.36 GB

Resultados (6)

Copiar | Descargar resultados en formato CSV

Filas de búsqueda

#	id	date	element	data_value	m_flag	q_flag	s_flag	obs_time	partition_0
1	SPE00119801	19381001	TMAX	192			E		by_station
2	SPE00119801	19381001	TMIN	135			E		by_station
3	SPE00119801	19381001	PRCP	1			E		by_station
4	SPE00119801	19381001	TMAX	192			E		by_year
5	SPE00119801	19381001	TMIN	135			E		by_year
6	SPE00119801	19381001	PRCP	1			E		by_year

Oviedo:

```
SELECT * FROM "AwsDataCatalog"."clima"."csv"
WHERE id = 'SPE00119828'
AND "date" = (
    SELECT MIN("date") FROM "AwsDataCatalog"."clima"."csv"
    WHERE id = 'SPE00119828'
);
```

Completado | Tiempo en cola: 115 ms | Tiempo de ejecución: 35.196 sec | Datos analizados: 413.36 GB

Resultados (6)

Copiar | Descargar resultados en formato CSV

Filas de búsqueda

#	id	date	element	data_value	m_flag	q_flag	s_flag	obs_time	partition_0
1	SPE00119828	19721201	TMAX	130			E		by_station
2	SPE00119828	19721201	TMIN	38			E		by_station
3	SPE00119828	19721201	PRCP	0			E		by_station
4	SPE00119828	19721201	TMAX	130			E		by_year
5	SPE00119828	19721201	TMIN	38			E		by_year
6	SPE00119828	19721201	PRCP	0			E		by_year