La finalidad de esta práctica es crear un Trabajos ETL (*ETL Jobs*) y disparadores en AWS Glue.

INTRODUCCIÓN

a) Vete haciendo capturas de pantalla de los pasos que des para resolver los ejercicios y añadiéndoles comentarios explicativos.

CONTENIDO
APARTADO A

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

2.- Mediante un comando AWS CLI, copia los archivos *csv* con las mediciones de todas las estaciones metereológicas 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".**

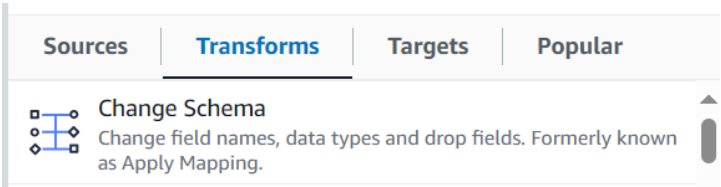
3.- Crea una base de datos en del *Data Catalog* que se llame ***espana.***

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

5.- Guarda el *Crawler* pero **no lo ejecutes.**

CONTENIDO
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.



3. Guarda el trabajo, **pero no lo ejecutes.**

CONTENIDO
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_.***

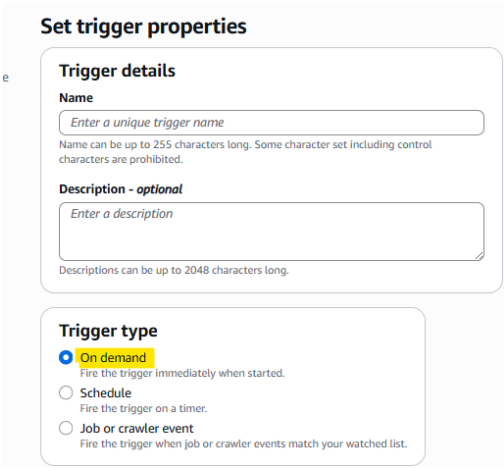
2.- Guarda el rastreador, **pero no lo ejecutes.**

CONTENIDO
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 *trabao* del apartado C.

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



4.- Arranca este manualmente este último disparador.

CONTENIDO

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.

CONTENIDO

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?

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

3.- ¿Cuántas mediciones tenemos de Oviedo?

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

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?