



## Big Data

**La finalidad de esta práctica es crear un rastreador (*Crawler*) AWS Glue para cargar nuestro Data Catalog.**

### INTRODUCCIÓN

- a) Utilizaremos un *bucket* público con información de estaciones meteorológicas de todo el mundo en formato csv que recoge datos desde el año 1763: <s3://noaa-ghcn-pds/csv/>
- b) En el siguiente enlace tenemos la descripción de los campos de dichos archivos: [open-data-docs/docs/noaa/noaa-ghcn/README.md at main · awslabs/open-data-docs · GitHub](https://open-data-docs/docs/noaa/noaa-ghcn/README.md)
- c) Vete haciendo capturas de pantalla de los pasos que des para resolver los ejercicios y añadiéndoles comentarios explicativos.

### CONTENIDO

#### APARTADO A

- 1.- Desde Aws CLI explora el contenido del *bucket* <s3://noaa-ghcn-pds/csv/>.
- 2.- Descarga uno cualquiera de los archivos que contiene en cada una de sus carpetas y muestra las primeras líneas de ellos.
- 3.- ¿Qué contiene cada uno de los dos tipos de archivos?

### CONTENIDO

#### APARTADO B

- 1.- Crea una base de datos en AWE 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.
- 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.
- 4.- ¿Está particionada la tabla? ¿Por qué campos?

### INTRODUCCIÓN

En el archivo <http://noaa-ghcn-pds.s3.amazonaws.com/ghcnd-stations.txt> tenemos una descripción de las estaciones meteorológicas.

Las españolas, como dice en la documentación, son las que comienzan por “**SP**” en el campo ID (el código de las 4 de Asturias son los que se ven en las imágenes de abajo):

50285	SPE00119792	43.5667	-6.0442	127.0	ASTURIAS/AVILES
50286	SPE00119801	43.5381	-5.6417	22.0	GIJON LA MERCED
50287	SPE00119819	43.5606	-5.6983	5.0	GIJON MUSEL
50288	SPE00119828	43.3542	-5.8728	336.0	OVIEDO



### CONTENIDO

#### APARTADO C

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

- 1.- ¿Cuántos registros tiene la tabla?
- 2.- ¿Cuántas mediciones tenemos de España?
- 3.- Sabiendo los códigos de las 4 estaciones de Asturias ¿Cuántas mediciones tenemos de Asturias?
- 4.- ¿Cuántas mediciones tenemos de Oviedo?
- 5.- ¿Cuál es la medición más antigua de España, Asturias y Oviedo?