Junto con databricks, AWS es el principal servicio del que se basa este proyecto. En particular utilizamos S3 para el almacenamiento de datos (las tablas csv de GDELT y consultas de Athena), Glue para juntar todos los archivos csv en una sola tabla consultable, Athena para transformar los datos crudos de la tabla generada por Glue en tablas agregadas o condensadas para su posterior uso en databricks, y servidores cloud EC2 para potenciar el poder de procesamiento y descargas de los datos.

**S3**

El bucket correspondiente al proyecto es **esg-gdelt**

1. La principal carpeta es gkg-data (gkg por el nombre que reciben los archivos csv de gdelt), acá se almacenan todas las tablas de la base de datos, separados en 3 carpetas:

“english”: contiene todos los datos principales de gdelt, desde el 2015 al 2021. Se puede notar que el numero en cada archivo corresponde a la fecha de creación del mismo.

“english-2015-2018”: igual que el anterior, datos provenientes de la base de datos “english” de gdelt (la principal) exclusivamente entre 2015 y 2018. Estos datos complementan el set anterior, pues estaban faltantes, no son duplicados.

“translation”: contiene los datos de la base de datos de nombre “translation” de gdelt. Esta base traduce las noticias de todo el mundo para aplicar sus algoritmos de NLP y generar las métricas que poblan las tablas. Es menos precisa pero mucho más amplia. Datos desde 2015 al 2021.

1. Query-results

Es la carpeta donde se almacenan las tablas generadas por las consultas de Athena. Las subcarpetas están segmentadas por fecha, por lo que se recomienda periódicamente eliminar aquellas que se hayan generado en un pasado y no se estén utilizando.

1. Test-folder

Nombre autoexplicativo, el fin de esta carpeta es probar funcionalidades de Python y S3 en conjunto.

1. Time-series

Tablas correspondientes al resultado final de los notebooks 2 y 3

1. StorageLens

Carpeta de sistema de databricks. Irrelevante

**Glue**

El primer uso que le damos a Glue es el de unificar todas las tablas csv de gdelt en una única base de datos consultable por Athena, mediante un Crawler: “gdelt-crawler”. Este está direccionado a la carpeta que contiene todos los archivos csv de interés, y de manera automática los unifica.

Adicionalmente tenemos los crawlers org-esg y org-market para incorporar las tablas de resultados de los notebooks de databricks en AWS y consultarlas en Athena para crear views que sean accesibles desde Tableau.

**Athena**

Usamos Athena para transformar los datos de gdelt a las tablas intermedias que se generan en los notebooks originales de los autores, esto pues es mucho más rápido procesarlas en AWS mientras siguen ahí (en S3), que migrar nuestros CSV de gdelt a databricks y luego procesarlos.

Es por ello que para ejecutar los notebooks antes se deben generar las tablas en el siguiente orden:

gdelt\_silver\_pre1

[gdelt\_silver\_pre2](javascript:void(0))

[gdelt\_silver](javascript:void(0))

[temp\_gold](javascript:void(0))

[gdelt\_gold](javascript:void(0))

[temp\_scores](javascript:void(0))

[scores](javascript:void(0))

**EC2**

De EC2 solo nos concierne un servidor, “Gdelt-prop”, responsable de descargar los csv de gdelt y almacenarlos en S3. Para ello revisar la carpeta server-key del proyecto, en github y seguir las instrucciones de la documentación.