Documentación del Segundo Sprint: Workflow y Proceso ETL en AWS

Resumen

En el segundo sprint de nuestro proyecto de análisis de flujos migratorios y sus impactos, nos enfocamos en la implementación de un workflow eficiente para el procesamiento de datos y la realización del proceso de Extracción, Transformación y Carga (ETL) de varios datasets relevantes. Este sprint fue esencial para preparar los datos que necesitamos para nuestro análisis posterior.

Descripción General del Workflow

El workflow implementado se basó en servicios y tecnologías de AWS (Amazon Web Services) para garantizar la escalabilidad y la eficiencia del proceso de ETL. A continuación, se describen los componentes clave y el paso a paso del workflow:

1. AWS Glue

- Utilizamos AWS Glue como servicio central para la realización del ETL. AWS Glue es un servicio de procesamiento de datos completamente administrado que automatiza gran parte del trabajo pesado asociado con la preparación de datos.

2. Extracción de Datos

- Para la extracción de datos, configuramos AWS Glue para conectarse a las fuentes de datos, que incluyen los siguientes datasets:

  - 'undesa': Datos de flujos migratorios de diferentes regiones del mundo.

  - 'consolidado_latinoamerica': Datos socioeconómicos de América Latina.

  - 'canada': Datos socioeconómicos de Canadá.

  - 'consolidado_regiones': Datos socioeconómicos de América Latina y el Caribe, Canadá y Estados Unidos.

  - 'eeuu': Datos socioeconómicos de Estados Unidos.

3. Transformación de Datos

- En esta etapa, utilizamos AWS Glue para aplicar transformaciones a los datos extraídos según nuestras necesidades específicas. Esto incluyó la limpieza, la agregación y la reestructuración de datos para facilitar su análisis posterior.

4. Carga de Datos

- Una vez que los datos se transformaron con éxito, los cargamos en un almacenamiento de datos adecuado para su posterior análisis. Utilizamos Amazon S3 para almacenar los datos procesados debido a su escalabilidad y disponibilidad.

5. Preparación para el Análisis

- Los datos procesados se dejaron listos y disponibles en Amazon S3 para su uso posterior en el análisis de flujos migratorios y sus impactos.

Tecnologías y Herramientas Asociadas

- **AWS Glue**: Como mencionamos anteriormente, AWS Glue fue la piedra angular de nuestro workflow de ETL.
- **Amazon S3**: Utilizamos Amazon S3 para almacenar los datos procesados debido a su capacidad de escalabilidad y su fácil accesibilidad.
- **Python**: En AWS Glue, escribimos scripts en Python para aplicar las transformaciones necesarias a los datos.
- **AWS IAM (Identity and Access Management)**: Configuramos permisos y roles de IAM para garantizar la seguridad de los datos y el acceso a los servicios de AWS.

Este segundo sprint fue fundamental para preparar los datos que necesitamos para el análisis en profundidad de los flujos migratorios y sus impactos. Los datos procesados y almacenados en Amazon S3 ahora están listos para ser utilizados en las etapas posteriores del proyecto.

En el próximo sprint, nos centraremos en la creación de modelos predictivos y la visualización de datos para extraer información valiosa de estos datasets.