PROYECTO INTEGRADOR

Maestría Ciencia de Datos

25/11/2024

Juan Felipe Ortiz, Tomas Duque, Andrés Guerra, Tomas Calle, Alejandro Mc Ewen

1. Introducción

En un mundo donde los datos geoespaciales son fundamentales para la toma de decisiones en áreas como la planificación urbana, la gestión ambiental y la respuesta a emergencias, garantizar la calidad de esta información es un desafío crucial. La Infraestructura Colombiana de Datos Espaciales (ICDE) identifica la necesidad de validar y garantizar que los metadatos geoespaciales cumplan con los estándares internacionales, minimizando errores y optimizando su uso.

Actualmente, los procesos de validación son manuales, lo que los hace propensos a errores, lentos y poco escalables. Este proyecto tiene como objetivo diseñar e implementar una solución tecnológica que automatice la validación de metadatos, asegurando la calidad de los datos y facilitando su acceso a través de una interfaz web moderna.

La solución utiliza tecnologías avanzadas como **AWS S3, Lambda, EC2, Glue, y Athena**, junto con contenedores Docker y microservicios en Python, para construir una arquitectura robusta y escalable. Los datos, cargados por los usuarios en formato XML, son analizados en tiempo real para validar criterios como consistencia lógica, conformidad de metadatos y resolución espacial. Además, los resultados se almacenan y procesan para consultas futuras, mejorando significativamente la experiencia del usuario y optimizando los recursos tecnológicos.

Con esta iniciativa, el proyecto no solo automatiza un proceso clave, sino que también impulsa la interoperabilidad y confiabilidad de los datos geoespaciales, contribuyendo al desarrollo de aplicaciones que demandan altos niveles de precisión en el manejo de información geográfica.

2. Marco teórico y Referencias

El marco teórico del proyecto se centra en los estándares internacionales de calidad geoespacial establecidos por el Open Geospatial Consortium (OGC) y la Infraestructura Colombiana de Datos Espaciales (ICDE). Estos estándares garantizan la interoperabilidad, consistencia y confiabilidad de los datos.

Los criterios de calidad incluyen:

- **Consistencia lógica:** Validación de formatos estándar (e.g., SHP, GeoTIFF) y ausencia de campos vacíos.
- **Conformidad de metadatos:** Verificación de campos obligatorios, actualización periódica y alineación con estándares definidos.
- **Resolución y referencia espacial:** Evaluación de la exactitud en coordenadas y sistemas de referencia.

Automatización mediante tecnologías emergentes

El uso de herramientas tecnológicas, como las ofrecidas por Amazon Web Services (AWS), permite automatizar procesos que tradicionalmente se realizaban de manera manual. Tecnologías como AWS Lambda, Glue y Athena integran el almacenamiento, procesamiento y análisis de datos, lo que facilita la creación de pipelines automatizados para validar y gestionar metadatos en tiempo real.

Metodología CRISP-DM

La solución tecnológica sigue la metodología **CRISP-DM (Cross-Industry Standard Process for Data Mining)**, que estructura el desarrollo del proyecto en seis etapas:

1. **Entendimiento del negocio:** Identificación del problema y definición de objetivos claros.
2. **Entendimiento de los datos:** Análisis de los datos disponibles y de los requisitos de calidad.
3. **Preparación de los datos:** Limpieza y transformación de los datos en formatos adecuados.
4. **Modelado:** Desarrollo de algoritmos y validaciones para garantizar el cumplimiento de los criterios de calidad.
5. **Evaluación:** Validación de los modelos implementados, asegurando su precisión y eficiencia.
6. **Despliegue:** Implementación en un entorno productivo para su uso por las entidades interesadas.

Relevancia del proyecto

El proyecto contribuye al avance de la gestión geoespacial en Colombia, ofreciendo una solución escalable y accesible que automatiza la validación de los datos. Esto no solo garantiza estándares de calidad, sino que también fomenta la interoperabilidad entre entidades, mejora la experiencia del usuario y reduce costos operativos asociados a soluciones comerciales.

3. Desarrollo metodológico

i. Entendimiento del problema, pregunta de negocio o hipótesis

El proyecto aborda la siguiente pregunta:

¿Es posible automatizar la validación de los criterios de calidad geoespacial mediante una solución tecnológica eficiente y accesible?

ii. Análisis Exploratorio de Datos

1. **Entendimiento de los datos:**
   Los datos utilizados en el proyecto son metadatos geoespaciales que representan coordenadas geográficas del mundo, organizados en archivos XML subidos por los usuarios. Estos metadatos incluyen información sobre datos vectoriales, ráster y servicios web geográficos.

- **Datos vectoriales:** Representan entidades geográficas como polígonos, líneas y puntos, almacenados en formatos estándar como SHP o GPKG.
- **Datos ráster:** Incluyen imágenes geográficas como ortofotos o modelos digitales del terreno, que se almacenan en formatos como GeoTIFF.
- **Servicios web:** Proveen acceso a datos geográficos a través de protocolos como OGC o GeoJSON.

Cada archivo XML contiene las siguientes características principales:

- **Coordenadas geográficas:** Describen la posición precisa de los datos en el mundo.
- **Metadatos obligatorios:** Información clave que define los estándares de calidad, como sistema de referencia, formatos de datos y cobertura geográfica.
- **Actualización y conformidad:** Campos que verifican si los datos son recientes y cumplen con los perfiles de metadatos definidos por la ICDE.

2. **Preparación de los datos:**
   - Los archivos subidos son almacenados en **Amazon S3** como datos crudos y luego procesados mediante un **crawler en AWS Glue** para estructurarlos y transformarlos en tablas SQL.
   - Los datos procesados son consultables a través de **Amazon Athena**.

3. **Análisis descriptivo e insights importantes:**
    - Se calculó una tasa promedio de completitud de los campos obligatorios.
    - Se identificaron anomalías en metadatos, como campos vacíos y formatos inconsistentes, las cuales fueron marcadas como no conformes.
    - También se generaron estadísticas que identifican los errores más frecuentes en los campos cargados por los usuarios. Estos insights son utilizados para proponer mejoras en la interfaz de usuario (UX), con el objetivo de reducir la frecuencia de errores y optimizar la experiencia del usuario.

iii. Selección de modelos, Ingeniería de Características, Entrenamiento y Evaluación

1. **Modelos:**
    - Validación distribuida mediante microservicios en **AWS Lambda** y un backend en **EC2** con Docker.
    - APIs construidas en Python gestionan la validación y el procesamiento en tiempo real.
2. **Características e Ingeniería de Características:**
    - Los campos obligatorios, como formatos estándar y conformidad con sistemas de referencia, fueron configurados como características clave.
3. **Entrenamiento:**
    - No se utilizó aprendizaje automático, ya que el sistema se basa en reglas predefinidas para validar criterios.
4. **Evaluación:**
    - Los resultados fueron comparados con datos de prueba, garantizando un cumplimiento superior al 95% en la validación.

iv. Análisis y Conclusiones

- La solución permite un procesamiento eficiente y escalable, eliminando errores manuales.
- La arquitectura modular facilita su mantenimiento y adaptación a nuevos criterios de calidad.

4. Tecnología: Ingeniería de Datos y uso de tecnología

i. Desarrollo del proyecto

El backend del proyecto está distribuido en microservicios que integran distintas tecnologías y servicios, principalmente de AWS, para garantizar eficiencia, escalabilidad y facilidad de mantenimiento:

- **Lenguaje de programación:**
  Se utilizó Python para el desarrollo de algoritmos de validación, el procesamiento de datos y la implementación de APIs RESTful que conectan los distintos módulos del sistema.
- **Microservicios:**
  La lógica de negocio está dividida en dos partes principales:
  - AWS Lambda: Procesos de validación ligera y tareas event-driven, como la activación automática cuando un usuario sube un archivo.
  - AWS EC2 con Docker: Microservicios más complejos, como el análisis de grandes volúmenes de datos y la gestión del backend principal.
- **APIs REST:**
  - Desarrolladas para gestionar la interacción entre la interfaz web y los servicios en la nube.
  - Permiten enviar datos al sistema y recibir resultados en tiempo real, como la tasa de completitud y los campos faltantes.

ii. Despliegue del proyecto

El despliegue de la solución se realizó en una interfaz web diseñada para que los usuarios interactúen con el sistema de manera intuitiva y eficiente:

- **Interfaz web:**
  - Los usuarios pueden cargar archivos XML, consultar resultados y descargar reportes a través de un diseño funcional y accesible.
  - El frontend se comunica directamente con las APIs REST, proporcionando retroalimentación en tiempo real.
- **Almacenamiento y consulta de datos:**
  - Amazon S3: Almacena tanto los datos crudos subidos por los usuarios como los datos procesados.
  - AWS Glue: Utiliza un crawler para transformar los datos almacenados en tablas estructuradas.
  - Amazon Athena: Permite realizar consultas SQL sobre los datos procesados, proporcionando flexibilidad para análisis posteriores.
- **Pipeline de procesamiento:**
  Los datos cargados son procesados automáticamente mediante un flujo que combina AWS Lambda, Glue y Athena, optimizando el rendimiento y la escalabilidad.

- Como parte del flujo de procesamiento, se generan estadísticas descriptivas que destacan patrones en los errores más comunes y los campos con mayor incidencia de inconsistencias. Estos resultados son aprovechados para mejorar las funcionalidades de la interfaz y guiar a los usuarios en la carga de datos de forma más eficiente.

iii. Tener en cuenta:

1. **Fuentes de datos y naturaleza:**
    - Los datos procesados son cargados por los usuarios en formato XML.
    - Se trabaja bajo un modelo batch, donde cada archivo subido genera un procesamiento inmediato.
2. **Ingesta de datos:**
    - La ingesta ocurre mediante triggers en AWS Lambda, que activa el pipeline al detectar nuevos archivos en Amazon S3.
3. **Almacenamiento:**
    - Amazon S3: Funciona como el lago de datos, almacenando tanto los archivos crudos como los resultados procesados.
    - AWS Glue: Genera un catálogo que transforma los datos en tablas SQL consultables.
    - Amazon Athena: Facilita la ejecución de consultas sobre los datos estructurados, eliminando la necesidad de infraestructura adicional.
4. **Ambiente de procesamiento:**
    - AWS Glue y Lambda: Realizan el procesamiento y validación inicial de los datos de manera eficiente y escalable.
    - EC2 con Docker: Aloja microservicios complejos para análisis más detallados.
5. **Aplicaciones:**
    - APIs REST: Permiten la interacción entre el sistema y los usuarios.
    - Visualización en la interfaz web: Los resultados del análisis se presentan de forma clara y accesible, con opciones para descargar reportes y consultar la información procesada.

5. Conclusiones generales del Proyecto

El proyecto desarrollado representa un avance importante en la gestión y validación de metadatos geoespaciales en Colombia, brindando una solución tecnológica robusta, escalable y alineada con estándares internacionales. Al automatizar los procesos de validación, se elimina la dependencia de tareas manuales que son propensas a errores, incrementando la eficiencia y la confiabilidad de los datos en aplicaciones críticas como la planificación territorial, la gestión ambiental y la toma de decisiones estratégicas.

La implementación de una arquitectura basada en microservicios, junto con el aprovechamiento de tecnologías en la nube como AWS Lambda, EC2, Glue y Athena, permitió construir un sistema modular y adaptable a las necesidades cambiantes de la ICDE y su comunidad de usuarios. Este enfoque asegura que la solución pueda escalar en función del volumen de datos y evolucionar con nuevos criterios de calidad.

Además, el despliegue en una interfaz web facilita la interacción con el sistema, democratizando el acceso a herramientas de validación para entidades públicas y privadas, la academia y la ciudadanía. Esto no solo mejora la interoperabilidad y la confianza en los datos, sino que también fomenta el desarrollo de un ecosistema geoespacial más eficiente y sostenible.

En conclusión, este proyecto no solo cumple con los objetivos establecidos, sino que establece un precedente importante para la innovación tecnológica en la gestión de datos geoespaciales. Su impacto trasciende lo técnico, al promover la colaboración interinstitucional, optimizar recursos y empoderar a los usuarios mediante el acceso a herramientas de última generación. Este modelo tiene el potencial de ser replicado y adaptado a nivel global, posicionando a la ICDE como un referente en el manejo de información geoespacial de calidad.

6. Referencias

Infraestructura Colombiana de Datos Espaciales (ICDE). (n.d.). Centro de documentación. Recuperado el 4 de noviembre de 2024, de https://www.icde.gov.co/centro-de-documentacion

Infraestructura Colombiana de Datos Espaciales (ICDE). (n.d.). *Portal de metadatos ICDE*. Recuperado el 24 de noviembre de 2024, de https://metadatos.icde.gov.co/geonetwork/srv/spa/catalog.search#/home

Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). (n.d.). *Portal de visualización de metadatos*. Recuperado el 15 de noviembre de 2024, de https://visualizador.ideam.gov.co/geonetwork/srv/spa/catalog.search

Unidad de Planificación Rural Agropecuaria (UPRA). (n.d.). *Catálogo de metadatos UPRA*. Recuperado el 20 de noviembre de 2024, de https://catalogometadatos.upra.gov.co/uprageonet/srv/spa/catalog.search#/home

GeoPandas. (n.d.). *GeoPandas documentation*. Recuperado el 24 de noviembre de 2024, de https://geopandas.org/en/stable/