Data Quality GenAI

Índice

Introducción

Data Quality en su versión original, exigía un proceso manual para la creación de nuevas reglas, lo que implicaba la necesidad de cierto nivel de conocimiento técnico y la posibilidad de cometer errores en la definición de dichas reglas.

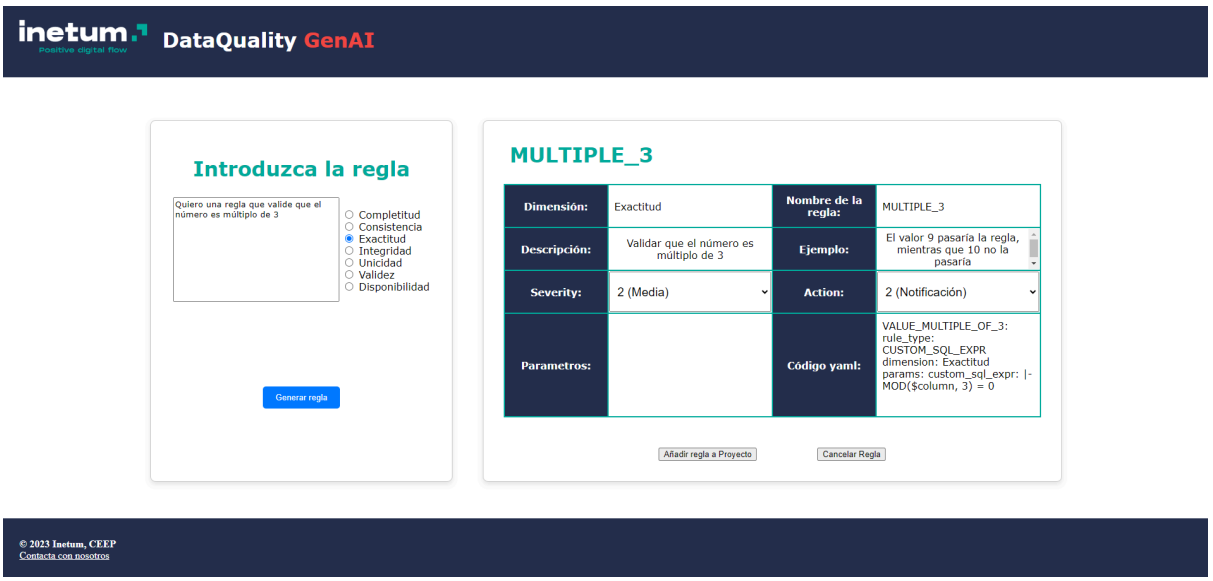
Dado que esta plataforma está dirigida a perfiles tanto técnicos como de negocios, y se buscaba una experiencia de usuario intuitiva y sencilla, se ha desarrollado esta evolución para superar esa dependencia técnica que caracterizaba a la versión inicial de la plataforma de Calidad de Datos. Añadiendo IA generativa conseguimos que el proceso de creación de reglas sea más fácil y accesible para usuarios no técnicos.

Funcionamiento

Esta evolución hace uso de la IA Generativa para generar todos los campos necesarios para el correcto funcionamiento de la nueva regla a partir de una instrucción en lenguaje natural.

Se dispone de una interfaz de entrada, que consiste en una web donde el usuario escribe una instrucción y elige una dimensión de Calidad. Luego presiona sobre el botón de “Generar” y una nueva regla aparecerá por pantalla.

La generación de la regla puede tardar unos segundos. El proceso también se realiza sobre Google Cloud al igual que el resto de la plataforma de Calidad de los Datos.

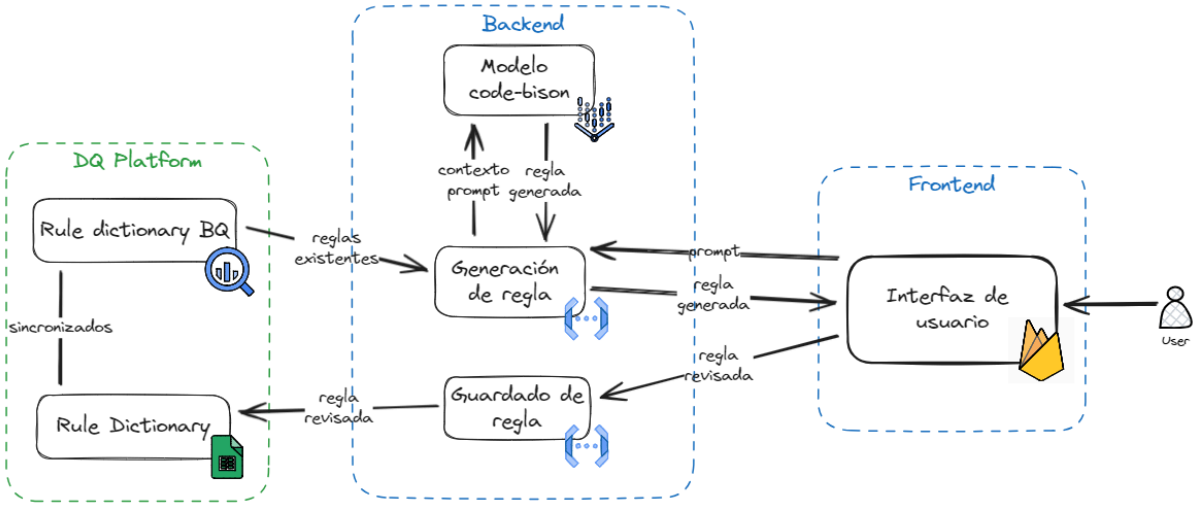


Una vez terminada la creación de la regla aparecerá por pantalla la información referente a la regla creada en base a la instrucción. Mediante IA Generativa se crea un nombre para la regla, descripción, ejemplo y el código yaml. Todos estos campos son editables por el usuario. Se podrá cambiar el nombre de la regla, el ejemplo o incluso ajustar el código para que adecúe más a la instrucción del usuario. También se podrá proporcionar un valor a la severidad y a la acción mediante una lista desplegable.

Una vez satisfecho el usuario con la generación de la regla, se pulsará sobre el botón de “Añadir regla al proyecto” y la regla se insertará automáticamente en el google sheets en el diccionario de Reglas según la dimensión que tenga asociada.

Arquitectura

La arquitectura de Data Quality GenAI consta de dos partes principales: el Frontend y el Backend.



Frontend

El Frontend está desarrollado utilizando tecnologías web estándar como HTML, CSS y JavaScript. Actualmente, está alojado en Firebase, lo que proporciona una plataforma confiable y escalable para la entrega de la interfaz de usuario.

Backend

Cloud Functions

El Backend se compone de dos Cloud Functions que trabajan en conjunto para la generación y gestión de reglas de calidad de datos.

- **Generación de Regla**: Se encarga de gestionar la generación de las reglas de calidad de datos. Esta función accede a BigQuery para recuperar todos los registros de Rule Dictionary, siendo estos todas las reglas disponibles Después formatea estos registros en formato JSON para hacerlos más fáciles de entender por la IA y con esto constituye la capa de contexto que posteriormente enviará al modelo de IA junto con el prompt del usuario. Una vez generada la regla, la función recibe sus parámetros y se los devuelve al Frontend en formato JSON para su presentación y posible edición por parte del usuario.

- **Guardado de Regla**: La segunda Cloud Function se encarga de insertar las reglas creadas por el usuario en el "Rule Dictionary", que está almacenado en Google Sheets. Recibe un JSON que contiene la información de la regla (modificada por el usuario en caso necesario) desde el Frontend y lo inserta en el "Rule Dictionary" en su posición adecuada acorde a su dimensión de calidad y ajusta cada valor de la regla a la columna correspondiente.

Modelo de IA Generativa

Para la generación de la regla utilizamos un modelo de chat code-bison proporcionado por VertexAI al que le agregamos una capa de contexto para poder generar la regla. Esta capa de contexto son todas las reglas en formato JSON recuperadas de BigQuery. Es importante mencionar que el modelo tiene un límite de tokens a pasar en cada interacción con el chat, por lo que se van pasando las reglas de poco en poco hasta llegar a mandar todas.

Integraciones y Consideraciones Técnicas

La herramienta aprovecha la integración entre Google Sheets y BigQuery para mantener actualizado en tiempo real el conjunto de datos utilizado para generar reglas. Cualquier cambio realizado en el "Rule Dictionary" se refleja instantáneamente en BigQuery, lo que garantiza que el modelo de generación de reglas siempre esté actualizado.

El manejo de CORS (Cross-Origin Resource Sharing) se ha abordado en el endpoint de la Cloud Function para garantizar la seguridad y accesibilidad de las solicitudes entre el Frontend y el Backend. Además, se ha evitado el uso de APIs para la conexión entre las Cloud Functions y Firebase, lo que simplifica la arquitectura y reduce la complejidad técnica.

Conclusiones

Data Quality GenAI proporciona una solución innovadora para la generación y gestión de reglas de calidad de datos mediante el uso de tecnologías avanzadas como la inteligencia artificial y la integración con servicios en la nube de Google. Esta herramienta ofrece una interfaz intuitiva para los usuarios y un backend eficiente y escalable para el procesamiento de datos y la generación de reglas de calidad de datos de manera automatizada y en tiempo real.