



Nuestro flujo de trabajo y tecnologías implementadas



El equipo se encuentra en una etapa inicial del proyecto, con una orientación clara hacia los objetivos, pero con limitaciones de tiempo. Se busca optimizar el flujo de trabajo para aprovechar al máximo el tiempo disponible mientras se desarrolla la infraestructura en la nube.

Realizar como primera instancia el ETL completo desde Colab → ¿Con qué finalidad?

Se propone utilizar Colab como entorno de desarrollo para el ETL completo, con el objetivo de brindar al resto del equipo datos limpios y listos para su análisis y procesamiento. Esto permitirá a cada miembro del equipo trabajar en sus tareas de forma independiente, mientras que los ingenieros de datos se enfocan en la implementación de la infraestructura en la nube.

Las herramientas que tenemos pensado usar para montar esta infraestructura son las siguientes:

- almacén de datos → **Google Cloud Storage** (5GB Free)

- procesamiento de datos → **Google Cloud Functions** (2M ejecuciones Free)
- análisis de datos → **Google BigQuery** (1TB Free)

Donde se utilizará Google Cloud Storage para la captura automática de datos en un bucket específico. Luego, se implementará un procesamiento automatizado de datos con Google Cloud Functions, utilizando triggers de Google Cloud Storage. La herramienta Google BigQuery se reservará para la ingestión automática de los datos procesados por Google Cloud Functions, facilitando consultas eficientes.

Una vez que los datos han sido procesados y cargados en BigQuery, se conectarán a dashboards o modelos de aprendizaje automático para análisis y visualización más profundos.

La elección de estas herramientas se basa en su capacidad para ofrecer un procesamiento instantáneo, escalabilidad y eficiencia. La combinación de Google Cloud Functions y los triggers de Google Cloud Storage, junto con la ingestión automática en BigQuery, proporciona una solución completa para el análisis de datos en tiempo real, permitiendo tomar decisiones informadas de manera ágil.

Una vez establecida la infraestructura, se procederá a establecer las conexiones necesarias para automatizar la mayoría de los procesos, asegurando así una gestión eficiente y sin contratiempos.