# Ingeniería de Datos: Integración y Proceso ETL de Fuentes Externas para Aplicación Web

## Descripción Técnica del Trabajo Realizado

Nuestra participación en el proyecto se enfocó en tareas propias del área de ingeniería de datos, especialmente en la identificación, extracción, tratamiento y entrega de información estructurada proveniente de fuentes externas. Se diseñó e implementó un flujo ETL (Extract, Transform, Load) orientado a garantizar la legalidad, consistencia y disponibilidad de los datos requeridos por la aplicación.

### 1. Identificación y selección de fuentes de datos

Se investigaron diversas fuentes web y APIs públicas con información relevante para los objetivos del proyecto. Se evaluaron criterios como estructura de los datos, frecuencia de actualización, formato y confiabilidad de las fuentes.

### 2. Verificación legal de uso de datos

Se analizaron los términos de uso, licencias y políticas de cada fuente. Se determinó la viabilidad legal de utilizar, transformar y republicar los datos en la aplicación, respetando las condiciones de atribución y limitaciones de uso cuando correspondía.

### 3. Extracción de datos (Extract)

Se desarrollaron scripts de web scraping y se configuraron integraciones con APIs para automatizar la recolección de datos. Se adaptaron los métodos de extracción a la estructura y restricciones técnicas de cada fuente.

### 4. Transformación de datos (Transform)

Se utilizaron herramientas como Pandas para convertir los datos a estructuras tabulares (dataframes). Se aplicaron procesos de limpieza, estandarización y enriquecimiento para asegurar la coherencia y utilidad de los datos. Se documentó

el flujo de transformación para facilitar su mantenimiento y trazabilidad.

**5. Carga y almacenamiento (Load)**

Se almacenaron los datos transformados en una base de datos PostgreSQL, hospedada en la plataforma en la nube Render. Se estructuró la base para permitir acceso eficiente por parte de los distintos equipos del proyecto. Actualmente, las actualizaciones de los datos se realizan bajo demanda, según necesidades del equipo.

Próximas mejoras: se contempla la creación de un flujo automatizado mediante Apache Airflow, organizando las tareas del pipeline ETL en un DAG para asegurar actualizaciones periódicas, trazables y escalables.

**6. Colaboración y entrega**

Se compartieron los datasets estructurados con los equipos de desarrollo, análisis y diseño. Se brindó soporte técnico y documentación sobre el uso y estructura de los datos para facilitar su integración en la aplicación.

**Resultado y valor aportado**

El trabajo permitió establecer una arquitectura sólida de ingeniería de datos, basada en un flujo ETL adaptable, legalmente conforme y centrado en la calidad de la información. Esto contribuyó significativamente al avance del proyecto, proporcionando a los distintos equipos una base de datos confiable y lista para su uso.