Arquitectura de Proyecto ETL para BigQuery

1. Fuente de Datos (Archivo CSV):

- **Descripción:** Nuestros datos se originan en un archivo CSV, que estará alojado localmente.
- **Flujo:** Con un script en Python, se lee el contenido del archivo CSV para iniciar el proceso ETL.

2. Proceso ETL en Python:

- **Descripción:** Núcleo del proceso donde los datos se extraen, transforman y cargan.
    - **Extracción:** Con Python, utilizo la biblioteca **`pandas`** para leer el archivo CSV y cargar su contenido en un DataFrame.
    - **Transformación:** Una vez cargados en el DataFrame, se emplean funciones de `pandas` para limpiar y transformar los datos. Esto incluye:
        - Eliminación de registros duplicados.
        - Completar valores faltantes usando técnicas como la imputación.
        - Realizar cualquier otra operación necesaria para asegurar la calidad y coherencia de los datos.
    - **Carga:** Una vez transformados, los datos estarán listos para ser cargados en BigQuery.
- **Flujo:** Los datos transformados son enviados a BigQuery mediante la SDK de GCP para Python.

3. BigQuery (Base de Datos):

- **Descripción:** En la nube de Google los datos transformados se cargan y consultan.
    - Diseñaré una tabla con un esquema adecuado para recibir y alojar los datos transformados.
    - Usaré consultas SQL para segmentar, agrupar y analizar la información según los requisitos.

4. Automatización del Proceso ETL:

- **Descripción:** Implementaré un mecanismo de automatización para garantizar que la ingestión de datos se realice de manera periódica.
    - **Apache Airflow**: Crearé un DAG (Directed Acyclic Graph) que defina la secuencia y programación del proceso ETL.
- **Flujo:** Se programará la ejecución del script ETL para que se ejecute automáticamente cada 24 horas.

5. Control de Versiones y Documentación:

- **Descripción:** Garantizo el seguimiento y control de las versiones del código.
  - Establezco una estructura de directorios clara y se emplean buenas prácticas para la ramificación (branching) y fusión (merging).
  - Crearé un archivo `README.md` en el repositorio principal proporcionando instrucciones detalladas sobre cómo configurar, ejecutar y entender el flujo de trabajo del proyecto.