Informe – Tarea 3

Base Datos a gran escala

Diego Moyano 202004509-7
Luis Zegarra 202073628-6
Nicolás Cancino 202004680-8

Desafíos que surgieron durante el desarrollo de la tarea:

1. Dificultad de que Apache Airflow detectara el DAG debido a errores de sintaxis, pero se resolvió este problema al estudiar el modelo subido en Aula, lo que permitió entender la estructura correcta del archivo y ajustar la definición de tareas y dependencias de forma adecuada.
2. Configuración del entorno mediante devcontainers, al principio el ambiente de Docker no se configuraba correctamente, pero después de varios intentos y guiarnos por el contenido del curso, se logró establecer un entorno Docker funcional en los equipos, con todos los servicios necesarios levantados (Airflow, Kafka y Apache Beam) y correctamente interconectados para ejecutar y probar la solución de forma local.
3. Simular adecuadamente el flujo completo (lectura de JSONL → conversión → escritura en Avro → ejecución → notificación a Kafka) en el entorno local requirió mucho esfuerzo de configuración, especialmente para que sea reproducible por terceros a través del `README`.

Conclusiones generales

A pesar de las dificultades iniciales, como los errores de sintaxis que impedían que Apache Airflow detectara correctamente el DAG, la configuración compleja del entorno con devcontainers y los problemas de integración con Kafka, se logró cumplir satisfactoriamente con los objetivos de la tarea. La revisión de los recursos proporcionados en Aula fue clave para corregir errores estructurales en el DAG, y la persistencia en la configuración del entorno Docker permitió establecer una plataforma funcional con todos los servicios necesarios.

Además, el desafío de enviar notificaciones a Kafka nos obligó a entender en mayor profundidad cómo funcionan sus componentes y cómo asegurar la conectividad adecuada entre servicios dentro del entorno. Simular todo el flujo de procesamiento desde la lectura de archivos JSONL hasta la escritura en formato Avro y la notificación

final supuso un reto técnico importante, especialmente pensando en que fuese fácilmente reproducible por terceros mediante el archivo README.

En resumen, este trabajo no solo cumplió con los requerimientos técnicos, sino que también fue una valiosa oportunidad para familiarizarnos con herramientas ampliamente utilizadas en la industria como Apache Beam, Airflow, Kafka y Docker. La experiencia nos permitió afianzar conceptos de integración, orquestación de procesos y despliegue de entornos reproducibles, lo cual es fundamental para enfrentar proyectos reales a gran escala.

Se hizo uso de la Inteligencia Artificial, más específicamente ChatGPT-4o para realizar una mejor redacción del informe, se usó el siguiente prompt

“Escríbeme y redacta bien los siguientes párrafos entendiendo que el contexto del problema es aplicar nuestros conocimientos en Apache Beam y Apache Airflow para construir una ETL que permita convertir ciertos datos de Json a Avro, para su posterior uso en el entrenamiento de modelos de Machine Learning: (aquí van los párrafos con nuestras ideas).”