Tercer Final de Big Data

Objetivo del Taller:

El objetivo de este taller es integrar las tecnologías usadas en las fases anteriores (Kafka para mensajería, Dask para transformación de datos) con Apache Airflow para la orquestación, y trabajar con SQL (PostgreSQL o MySQL) para generar un data warehouse final.

Entrega Final (30% de la nota del curso)

Fecha de entrega: Viernes, 4 de mayo, en clase y plataforma teams.

Entregables via teams: Códigos, Documentación, Video cortico de la prueba.

Instrucciones Generales:

1. **Orquestación del Pipeline (Airflow):**
    - Debes implementar un flujo de trabajo en Apache Airflow que orqueste todas las tareas desde la ingesta de datos con Kafka hasta la carga en la base de datos final. Las tareas de transformación, carga y generación de informes deberán estar orquestadas.
    - Crea DAGs (Directed Acyclic Graphs) que representen las fases de ingesta, procesamiento, almacenamiento en la base de datos, generación de gráficas y entrenamiento-predicción de un modelo de ML.
2. **Ingesta y Procesamiento:**
    - **Kafka** (de la entrega anterior): Mantener la ingesta de datos utilizando **múltiples** productores y consumidores.
    - **Pandas o Dask**: Utiliza Dask para realizar la transformación final de los datos, asegurando que los datos estén listos para ser almacenados en la base de datos SQL.
3. **Almacenamiento (SQL Data Warehouse):**
    - **SQL (PostgreSQL o MySQL)**: Almacenar los datos transformados en una base de datos SQL. La base debe estar bien estructurada para facilitar consultas y análisis.
    - Incluir un esquema claro de las tablas utilizadas, con detalles de los tipos de datos y relaciones entre las tablas.
    - **Data Warehouse**: Genera un pequeño data warehouse para análisis de los datos en tiempo real o batch, permitiendo la ejecución de consultas complejas.

4. **Visualizaciones y Análisis (otra tarea DEL dag):**
    - **Visualización de Datos**: Genera al menos dos visualizaciones a partir de los datos almacenados en la base de datos SQL. Puedes usar herramientas como Matplotlib, Seaborn, o cualquier biblioteca de visualización en Python.
    - **Modelos de Machine Learning**: Entrena un modelo de Machine Learning sobre los datos almacenados en el data warehouse. Puedes usar cualquier algoritmo (regresión, clasificación, clustering, etc.). NO IMPORTA SI EL SCORE ES BAJITO.
5. **Documentación (2 PÁG MAX) y demostración (video corto):**
    - **Flujo del Proceso**: Explica detalladamente el flujo de trabajo desde la ingesta hasta la generación de informes. Usa diagramas de flujo que ilustren cómo interactúan Kafka, Dask, Airflow, SQL y las visualizaciones.
    - **Explicación de Airflow**: Describe cómo has configurado los DAGs en Apache Airflow y cómo se aseguran de que las tareas se ejecuten correctamente.

Resumen Criterios de Evaluación:

1. **Orquestación con Apache Airflow**: (2 unidades)
    - Calidad y funcionalidad de los DAGs creados, coordinación adecuada entre tareas.
    - Programe los tiempos entre dags para que sea algo que se pueda ver en clase.
2. **Integración en el DAG Kafka, Dask**: (1 unidad)
    - Flujo de datos continuo y eficiente desde la ingesta hasta la transformación.
3. **Almacenamiento en SQL (PostgreSQL/MySQL) en el DAG**: (1 unidad)
    - Diseño de la base de datos (esquema, relaciones) y la correcta inserción de los datos transformados.
4. **Visualizaciones y Modelo de Machine Learning en el DAG**: (1 unidad)
    - No importa el tipo de gráfica o modelo, o si da un buen resultado, lo único es que sea un proceso más del DAG.
5. **Documentación del Proceso/video corto de que funciona**: (No entregar resta 1 unidad)
    - Calidad de la explicación escrita, diagramas de flujo, esquemas y justificación de decisiones técnicas.