Prueba Data Engineering Spin By Oxxo

Prueba

El objetivo principal de la prueba es entender patrones de diseño y mejores prácticas implementadas en el desarrollo de un pipeline de datos usando un caso de uso real.
El problema proporcionado toma como base este dataset publicado en Kaggle. El dataset es usado para implementar un modelo de riesgo crediticio usando información real. Esta prueba consiste en tomar **este** kernel y traducir el código fuente a un pipeline de Airflow (DAG).

Consideraciones

- La prueba consiste en tomar el kernel proporcionado y adaptarlo a un DAG, no hace falta implementar un nuevo proceso o evaluar la eficacia de predicción del mismo o hacer ajustes en el modelo.

- Se considerará principalmente como es que el candidato ataca el problema de dividir eficientemente las tareas dentro del DAG de Airflow creado.
  .
- La versión de Airflow a usar queda a consideración del candidato.

- El primer paso del DAG debe leer la información.

- El último paso del DAG debe escribir un dataframe en formato parquet con la clasificación completa del dataset de la misma forma que describe el kernel. **Bonus: ¿Podrías hacer que el DAG escriba el archivo final en AWS S3 o que ingeste directo en una base de datos?**

- Analizando el contenido del data frame, asume que este proceso será replicado n número de veces. ¿sería mejor considerar guardarlo en una base de datos en lugar parquets? ¿que propondrías para poder consultar millones de resultados asumiendo que este proceso los generaría? ¿Se te ocurre alguna solución mejor a guardarlo en una base de datos?.

- El aspecto principal a evaluar será el patrón de diseño en el DAG.

- La prueba no debería tardar más de una semana en desarrollarse.

- La elección de cómo presentar el proyecto (Repo Git, zip via mail, etc.) queda a consideración del candidato.

- Para garantizar la calidad de los datos y haciendo uso de tu sentido de Data Steward propón una estrategía de Data Governance considerando Diccionarios de Datos y generando el propio del ejercicio a realizar.