Desafío Técnico - Data *Engineer* ⚡

Consideraciones

Deadline

Entendemos cada situación en particular y es por eso que preferimos que lo hagas a tu tiempo, aunque en no más de una semana desde que recibiste el desafío. Lo más importante es que la solución tenga explicaciones que demuestren tus decisiones y los gráficos y diagramas tengan tu sello.

Puntos a evaluar

- Interpretación de la consigna
- Orden y código documentado
- Código simple y efectivo
- Calidad del entregable
- Si pensás otro punto que pueda sumar, sentite libre de agregarlo y comentarnos cuál es y por qué lo sumaste.

Consigna

La evaluación se divide conceptualmente en 4 secciones:

1. Programático

Se solicita programar una función en Python para bajar datos desde un S3. La función debe permitir que:

- Se pueda cambiar el repositorio de datos de forma flexible.
- Se pueda cambiar la ruta/nombre de los archivos de forma flexible.

- Se permita la descarga de archivos como csv.

Atención: Revisar anexos al final del documento.

2. QA

- Hacer análisis descriptivo y control de datos con Python.
- Explicar hallazgos y posibles problemas en un Jupyter Notebook o archivo con el código. Proponer posibles soluciones de forma escrita y/o profundizando en el video.

3. Modelado de datos

- Partiendo de la salida generada en el punto anterior (QA) armar el modelo de datos.
- Generar el script DDL para la creación de cada una de las tablas representadas en el modelo datos.
- Subir los archivos a una base de datos a elección.
- Crear la cantidad de tablas e índices que consideres necesario.

4. SQL

Responder en base al modelo de datos realizado en el punto anterior:

- Considerando únicamente la plataforma de Netflix, ¿qué actor aparece más veces?
- Top 10 de actores participantes considerando ambas plataformas en el año actual. Se aprecia flexibilidad.
- Crear un Stored Proceadure que tome como parámetro un año y devuelva una tabla con las 5 películas con mayor duración en minutos.

Entregables

- Archivos de código (SQL y Python) que permitan recrear la solución de los diferentes puntos desarrollados.
- Diagrama del modelo de datos.
- Resultados de las queries.

- Video de máximo 10 minutos explicando la resolución.

Anexos

Para conectarse a S3 y acceder a los archivos:

Key AKIA2NU5TZR6RVMXSOKK

Secret 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

Archivos

https://desafio-rkd.s3.amazonaws.com/disney_plus_titles.csv

https://desafio-rkd.s3.amazonaws.com/netflix_titles.csv