stay unique
ESTANDO CONTIGO

Prueba Data Engineer Junior

Problemática:

La gerencia de Stay Unique ha solicitado un análisis exhaustivo para evaluar el desempeño de nuestras propiedades vacacionales en comparación con las del mercado en Barcelona. Para tomar decisiones estratégicas y mejorar nuestro posicionamiento competitivo, necesitamos contar con datos de propiedades externas, así como la información interna de nuestras propiedades y reservas, en un formato listo para el análisis.

El equipo de datos tiene que realizar lo siguiente:

Ejercicios:

1 - Realice un scraping de una página de propiedades vacacionales (ej. Booking o Airbnb) para extraer la información que considere relevante. Puede incluir datos como el nombre de la propiedad, ubicación, precio por noche, número de habitaciones, puntuación de reseñas, entre otros.

A continuación, se detallan las instrucciones a seguir:

- Extraiga al menos 100 propiedades o reseñas de Barcelona.
- Incluya detalles como nombre de la propiedad, dirección, precio por noche, puntuación de reseñas, y número de habitaciones.
- Almacene los datos en un archivo CSV o JSON.
- Documente el proceso de scraping, incluyendo cómo manejó la paginación (si aplica), posibles bloqueos de acceso (ej. captchas o limitaciones de solicitud) y cualquier otro desafío que haya encontrado.

2 - Recibirá dos datasets; uno de propiedades (Properties) y otro de reservas (Bookings), con información de propiedades vacacionales recopilada en el último año, se debe realizar procesos de ETL y EDA sobre los mismos con el fin de dejarlos tener un solo archivo que sea consumible para análisis.

3 - Los datos obtenidos mediante scraping y el dataset limpio deben consolidarse en uno o varios archivos CSV o cargarse en una base de datos, según lo considere adecuado. Puede optar por subir un único dataset consolidado o varios datasets separados, dependiendo de la estructura y el análisis que considere más conveniente.

A continuación, se detallan las instrucciones a seguir:

- Consolide los datos en uno o más archivos CSV o cárguelos en una base de datos (puede elegir entre MySQL, PostgreSQL o BigQuery).
- Si lo considera relevante, puede explorar el uso de APIs para complementar o enriquecer los datos obtenidos durante el scraping. Documente el proceso si decide utilizar alguna API adicional.

Entregables en GitHub:

- **Código de scraping**: Todo el código para realizar el scraping debe subirse a GitHub, documentado con comentarios y explicando cómo ejecutarlo.
- **Código de limpieza**: El código utilizado para limpiar y transformar los datos.
- **Dataset final**: El archivo CSV limpio y el archivo con los datos obtenidos del scraping.
- **README**: Un archivo detallado que explique:
    - Cómo configuró el entorno.
    - Cómo ejecutó los scripts.
    - Detalles de las decisiones de limpieza de datos.
    - Descripción del pipeline de ETL implementado.
    - Cualquier reto o problema encontrado y cómo lo resolvió durante el proceso.

Envíe su trabajo hasta lo que haya avanzado, incluso si no puede contestar todo.