Prueba técnica Lulo Bank - Data engineers

Versión: 2021-01-27

Actividades

1. Realizar esta prueba utilizando Python (Utilizar buenas prácticas de codificación).
    a. Punto extra pruebas unitarias.

2. Obtener información del siguiente API Rest http://api.tvmaze.com trayendo todas las series que se emitieron en diciembre del 2020.
   Ayuda: para obtener las series emitidas el 29 de mayo del 2020 se utilizó el siguiente llamado http://api.tvmaze.com/schedule/web?date=2020-05-29

3. Almacenar los datos crudos.

4. Con base a los Json obtenidos del API generar diferentes dataframes de pandas que conserven la integridad referencial de los datos del Json.

5. Realizar profiling a los DataFrames y realizar un análisis.
    a. Se espera el resultado del profiling (documento en PDF o HTML) y el análisis de éste.

6. Realizar operaciones de limpieza si es necesario de los datos que están en los dataframes.

7. Realizar operaciones de agregación para obtener:
    a. Runtime promedio.
    b. Conteo de shows de tv por género.

8. Listar los dominios (web) del sitio oficial de los shows.

9. Almacenar los diferentes DataFrames en archivos parquet (con compresión snappy). Tener en cuenta que es importante que se mantenga en tipo de dato en cada columna del parquet.

10. Leer los archivos parquet del punto anterior y almacenar esta información en una base de datos (sqlite) creada por ustedes que respete la integridad referencial.

Entregables:

Un archivo comprimido con lo siguiente:

- Json obtenidos de las consultas al API.
- El archivo del profiling y un archivo adicional del análisis de éste.
- Los parquet generados.
- El archivo de la base de datos SQLite.
- El proyecto de Python que desarrolló el ejercicio