Mejoras

Olmo Baldoni, Nasser Chaouchi
922443@unizar.es, 922613@unizar.es

Mayo 2024

Durante el desarrollo de nuestro proyecto, introdujimos una serie de cambios y mejoras en el modelo de datos y el proceso ETL para optimizar la calidad de los datos y la eficacia del análisis.

1 Modificaciones en el Modelo de Datos

En primer lugar, en el modelo de datos introdujimos `multiple_concatenate_X_dim_table` y `multiple_X_dim_table` en las dimensiones de género, ocupación y profesión. Esta metodología nos permite gestionar eficazmente múltiples combinaciones de géneros o profesiones asociadas a una persona o una película. Los nombres concatenados, como `multiple_concatenate_profession_name`, identifican unívocamente estas combinaciones, facilitando las consultas y los análisis complejos.

Ventajas de las Modificaciones

- **Flexibilidad y Precisión:** Aumenta la flexibilidad y precisión del análisis.
- **Selección Detallada:** Facilita la selección detallada de películas por género o personas por profesión.
- **Gestión de Filtros:** Mejora la gestión de los filtros para su posterior uso en PowerBI, optimizando la presentación y el análisis de los datos.

2 Mejoras en el Proceso ETL

Hemos creado flujos de trabajo específicos para cada tabla de origen para garantizar que los archivos se leen correctamente y se insertan en la base de datos PostgreSQL sin errores. El uso de flujos de trabajo específicos para cada fuente nos permite manejar correctamente las particularidades de cada archivo de datos, como los diferentes delimitadores y formatos.

Beneficios de los Flujos de Trabajo

- **Automatización y Gestión Eficaz:** Los flujos de trabajo de Apache HOP garantizan una automatización y gestión eficaces del proceso ETL.
- **Minimización de Errores:** Minimiza el riesgo de errores humanos e incoherencias en los datos.
- **Gestión y Supervisión:** Facilita la gestión de registros y la supervisión del proceso ETL, permitiendo identificar y resolver problemas con mayor rapidez y eficacia.

Estos cambios se han aplicado para mejorar la solidez, flexibilidad y rendimiento de nuestro modelo de datos y proceso ETL, garantizando una mejor calidad de los datos y un análisis más preciso y eficaz.