Project Overview

This project is designed to extract, transform, and load (ETL) Instacart data from MySQL to Snowflake. The project also includes exploratory data analysis (EDA) and insights generation from the Instacart dataset.

Scripts and Notebooks

1. ETL Pipeline

- Mysql_to_snowflake.py
    - Extracts data from a MySQL database.
    - Transforms data as needed.
    - Loads the data into a Snowflake database.
- Load_data.py
    - Loads raw data into the processing environment.
- Transform_data.py
    - Cleans and processes data for analysis.
    - Performs feature engineering where required.
- Create_snowflake_db.py
    - Creates the necessary tables and schemas in Snowflake.

2. Exploratory Data Analysis (EDA)

- eda.ipynb
    - Performs initial exploration of the dataset.
    - Generates summary statistics and visualizations.

3. Insights Generation

- insights.ipynb
    - Analyzes user purchase behavior.
    - Provides insights on order frequency, reorders, and product popularity.
    - Visualizes trends in customer purchasing patterns.

Data Flow (how it works):

1. Extract data from MySQL (mysql_to_snowflake.py).
2. Load raw data into the processing environment (load_data.py).
3. Transform and clean data (transform_data.py).
4. Create necessary database structures in Snowflake (create_snowflake_db.py).

5. Perform EDA (eda.ipynb).
6. Generate insights from cleaned data (insights.ipynb).

Technologies Used:

- Databases: MySQL, Snowflake
- Programming Language: Python
- Libraries: pandas, matplotlib, seaborn, snowflake-connector-python
- Environment: WSL, Jupyter Notebook

Documentación del Pipeline en MageAI

1. Descripción del Pipeline

Este pipeline extrae datos de una base de datos MySQL y los carga en Snowflake usando MageAI. Se utiliza pymysql para la conexión con MySQL y mage_ai.io.snowflake para manejar la carga en Snowflake.

2. Configuración de Credenciales

Se cargan manualmente las credenciales desde archivos YAML:

- MySQL: /root/instacart_project/instacart_pipeline/my_connections/mysql.yaml
- Snowflake: /root/instacart_project/instacart_pipeline/my_connections/snowflake.yaml

Ambas configuraciones se validan para asegurar que fueron correctamente cargadas como diccionarios.

3. Conexión a las Bases de Datos

- MySQL: Se establece la conexión utilizando pymysql.connect, y se maneja cualquier error en la conexión.
- Snowflake: Se inicializa la conexión usando la clase Snowflake de MageAI, permitiendo exportar datos directamente.

4. Extracción de Datos desde MySQL

Se define una lista de tablas a extraer:

- instacart_orders
- products

- order_products
- aisles
- departments

Para cada tabla:

1. Se ejecuta una consulta SELECT * FROM {table}.
2. Los datos se transforman en un DataFrame de Pandas.
3. Se imprime un mensaje de confirmación.

5. Carga de Datos en Snowflake

La carga se realiza en lotes (BATCH_SIZE = 10,000):

1. Se divide el DataFrame en segmentos de 10,000 filas.
2. Cada lote se exporta a Snowflake usando export.
3. Se imprimen mensajes para indicar el progreso de la carga.
4. Se maneja cualquier error durante la inserción.

6. Cierre de Conexiones

Una vez transferidos los datos:

- Se cierra la conexión con MySQL.
- Se imprime un mensaje de finalización exitoso.

7. Posibles Mejoras

- Paralelización: Usar procesamiento en paralelo para acelerar la extracción y carga.
- Validación de Datos: Agregar verificaciones para evitar la carga de datos corruptos.
- Logging Avanzado: Guardar logs detallados en un archivo en lugar de imprimirlos en consola.
- Automatización con MageAI: Configurar flujos de trabajo para ejecución periódica.

Este pipeline permite transferir eficientemente datos desde MySQL a Snowflake, asegurando robustez y escalabilidad.