Prueba Técnica - Analista de Datos

Nombre: Fernando Ernesto Castillo Marroquin
Fecha: 4-6-2025
Proyecto: ETL con dataset Olist hacia SQL Server

Introducción

El presente documento describe la solución desarrollada para la prueba técnica de Analista de Datos, cuyo objetivo es demostrar habilidades de modelado de datos, construcción de un proceso ETL automatizado y documentación técnica.

Se utilizó el dataset **Brazilian E-Commerce Public Dataset by Olist**, el cual contiene información transaccional de una plataforma de comercio electrónico en Brasil.

Selección de archivos

Del conjunto completo de archivos, se seleccionaron los siguientes tres por su relevancia y relación directa:

- **olist_orders_dataset.csv**: contiene los datos generales de cada orden realizada.
- **olist_order_items_dataset.csv**: detalla los artículos incluidos en cada orden.
- **olist_order_customers_dataset.csv**: información del cliente que realizó cada orden.

Esta combinación permite construir un modelo relacional básico que enlaza órdenes con clientes y artículos.

Modelado de datos

Se diseñó un modelo relacional normalizado compuesto por tres tablas principales:

Tabla customers

- Contiene la información del cliente.
- Clave primaria: customer_id.

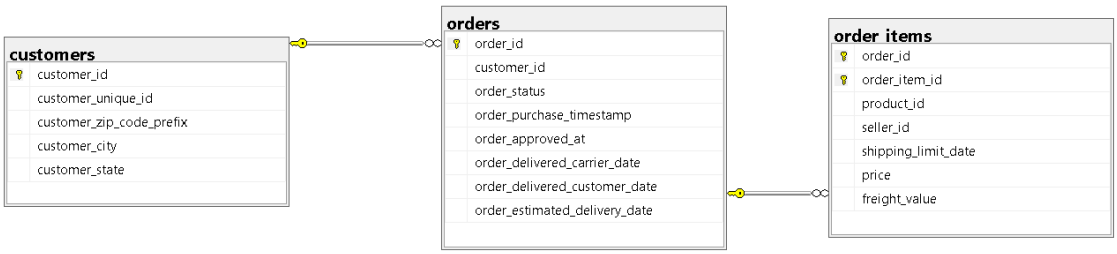
Tabla orders

- Registra cada orden realizada.

- Relación con customers mediante customer_id.
- Clave primaria: order_id.

Tabla order_items

- Detalla cada producto incluido en una orden.
- Relación con orders mediante order_id.
- Clave compuesta primaria: (order_id, order_item_id).



El modelo está implementado en SQL Server con claves primarias y foráneas, como se muestra en el archivo ddl_script.sql.

Proceso ETL

Se desarrolló un script Python (etl_olist_sqlserver.py) que realiza:

- **Extracción:** Lectura de archivos CSV usando pandas.
- **Transformación:** Conversión de columnas de fecha, eliminación de registros nulos.
- **Carga:** Inserción en tablas de SQL Server mediante SQLAlchemy y pyodbc.

El proceso es completamente automático y puede reutilizarse con diferentes versiones del dataset.

Tambien se desarrolló el script Python (etl_olist_sqlserver_conexion.py) para realizar y establecer la conexión con la base de datos (SQL Server).

Tecnologías utilizadas

- Python: Lenguaje de programación versátil para desarrollo web, análisis de datos, automatización y más.
- Pandas: Biblioteca de Python para manipulación y análisis de datos (tablas, series temporales, limpieza de datos).
- SQLAlchemy: Herramienta de Python para interactuar con bases de datos usando ORM (mapeo objeto-relacional) o SQL directo.
- Pyodbc: Biblioteca de Python que permite conectarse a bases de datos (como SQL Server) mediante ODBC.
- SQL Server Express 2019: Versión gratuita de Microsoft SQL Server, ideal para bases de datos pequeñas o desarrollo local.
- VSCode: Editor de código ligero y potente de Microsoft, con extensiones para Python, SQL, etc.
- Power BI: Herramienta de visualización y análisis de datos de Microsoft (dashboards, informes interactivos).

Repositorio en GitHub

Repositorio con todo el código y scripts:
Fernando0131/Prueba-T-cnica: Repositorio creado para realizar y almacenar los archivos de la prueba técnica para analista de datos

El repositorio incluye:

- Scripts ETL (etl_olist_sqlserver.py, etl_olist_sqlserver_conexion.py)
- Script DDL (ddl_script.sql)
- Archivo PBIX (OLIST - Dashboard de Ventas y Pedidos.pbix)
- Este documento (documentacion_tecnica_Fernando_Marroquin.pdf)
- README con instrucciones de ejecución