Documentación

Introducción

Este documento describe la implementación de un flujo automatizado de Extracción, Transformación y Carga (ETL) utilizando Python para cargar los archivos del conjunto de datos Olist en una base de datos Oracle. Se utilizó Vertabelo para desarrollar el Modelo de Datos Relacional y el diccionario de datos. El ETL extrae información relevante de los archivos CSV, transforma los datos y carga la información en tablas en Oracle.

Herramientas Utilizadas

Python: Lenguaje de programación para el desarrollo del proceso ETL.

Pandas: Biblioteca para manipulación y análisis de datos.

SQLAlchemy: Biblioteca para interactuar con bases de datos SQL desde Python.

Oracle Database: Motor de base de datos relacional.

Vertabelo: Herramienta de modelado de datos en línea.

Proceso ETL

1. Modelado de Datos en Vertabelo: Se diseñó el modelo de datos relacional en Vertabelo, definiendo entidades, atributos y relaciones.
2. Generación del Diccionario de Datos: Vertabelo facilitó la creación del diccionario de datos, documentando las propiedades de las entidades y atributos.
3. Extracción de Datos: Los archivos CSV (olist_customers_dataset.csv, olist_orders_dataset.csv, olist_order_items_dataset.csv) son leídos utilizando la biblioteca Pandas.
4. Transformación de Datos: Se realizan transformaciones específicas para seleccionar solo los campos necesarios en cada DataFrame.

5. Carga de Datos en Oracle: Se establece una conexión a la base de datos Oracle mediante SQLAlchemy.
6. Los DataFrames transformados se cargan en las tablas correspondientes (Customers, Orders, Order_Items) en Oracle.

Resultados

El script ETL se ejecuta con éxito, extrayendo datos de los archivos CSV, transformándolos y cargándolos en las tablas Oracle.

Las tablas en Oracle (Customers, Orders, Order_Items) reflejan el modelo de datos relacional diseñado en Vertabelo.

Diagrama de base de datos