Centro Universitario de Occidente
Ingeniería en Ciencias y Sistemas
Inteligencia Artificial
Ing. Marco Garcia

Manual Técnico
ETL-Inteligente

Fatima Odra Daniela Tezo Sum
201831039

Quetzaltenango, abril 2022

Índice

SmartETL

SmartEtl es una aplicación especialmente diseñada para ser utilizada en un sistema operativo Windows con PostgreSQL como sistema de base de datos.

Requerimientos Mínimos

Hardware

SmartETL no es una aplicación que consuma muchos recursos, sin embargo se recomienda un procesador con frecuencia mínima de 2.4 Ghz y una memoria ram mínima de 4GB

Software

Para la ejecución de SmartETL se necesita como mínimo Python en su versión 3.10.1, contar con las librerías adecuadas, tener instalado Postgresql en su versión 14 y a su vez tener la base de datos.

Herramientas utilizadas para el desarrollo

- **Python**

Python es un lenguaje de programación interpretado, orientado a objetos y de alto nivel con semántica dinámica. Sus estructuras de datos integradas de alto nivel, combinadas con la escritura dinámica y el enlace dinámico, lo hacen muy atractivo para el desarrollo rápido de aplicaciones, así como para su uso como lenguaje de scripting o pegamento para conectar componentes existentes.

- **PostgreSQL**

PostgreSQL es un potente sistema de base de datos relacional de objetos de código abierto que utiliza y amplía el lenguaje SQL combinado con muchas características que almacenan y escalan de forma segura las cargas de trabajo de datos más complicadas.

- **Visual Studio Code**

Es un editor de código fuente desarrollado por Microsoft para Windows, Linux, macOS y Web. Incluye soporte para la depuración, control integrado de Git, resaltado de sintaxis, finalización inteligente de código, fragmentos y refactorización de código. También es personalizable, por lo que los usuarios pueden cambiar el tema del editor, los atajos de teclado y las preferencias. Es gratuito y de código abierto, aunque la descarga oficial está bajo software privativo e incluye características personalizadas por Microsoft.

- **Github**

Github es un portal creado para alojar el código de las aplicaciones de cualquier desarrollador, y que fue comprado por Microsoft en junio del 2018. La plataforma está creada para que los desarrolladores suban el código de sus aplicaciones y herramientas, y que como usuario no sólo se pueda descargar la aplicación, sino también entrar al perfil para leer sobre ella o colaborar con su desarrollo.

Librerías utilizadas

Tkinter

Tkinter es el paquete GUI (interfaz gráfica de usuario) estándar de facto de Python. Es una capa delgada orientada a objetos en la parte superior de Tcl / Tk. Tkinter no es el único kit de herramientas de GuiProgramming para Python. Sin embargo, es el más utilizado.

Psycopg2

Psycopg2 es un adaptador PostgreSQL para el lenguaje Python implementado utilizando libpq, la librería oficial del cliente PostgreSQL. Su código de programa es poco, rápido y estable.

Pandas

Pandas es una muy popular librería de código abierto dentro de los desarrolladores de Python, y sobre todo dentro del ámbito de Data Science y Machine Learning, ya que ofrece unas estructuras muy poderosas y flexibles que facilitan la manipulación y tratamiento de datos.

Pandas surgió como necesidad de aunar en una única librería todo lo necesario para que un analista de datos pudiese tener en una misma herramienta todas las funcionalidades que necesitaba en su día a día, como son: cargar datos, modelar, analizar, manipular y prepararlos.

Sqlalchemy

SQLAlchemy es un Object-Relational Mapper / Mapping-tool, o un ORM, es decir una librería que los desarrolladores utilizan para crear bases de datos y manipular sus datos sin la necesidad de conocer / usar SQL.

Existen otras alternativas como SQL Alchemy o Peewee, y otros lenguajes tienen sus propios ORMs como PHP Eloquent o Java Hibernate.

Numpy

NumPy es una librería de Python especializada en el cálculo numérico y el análisis de datos, especialmente para un gran volumen de datos.

Incorpora una nueva clase de objetos llamados arrays que permite representar colecciones de datos de un mismo tipo en varias dimensiones, y funciones muy eficientes para su manipulación.

Validate_email

Validate_email es un paquete para Python que revisa si un email es válido, propiamente formateado y si puede llegar a existir.

SmartETL

Instalación de Aplicaciones

Requisitos generales pre-instalación

Python

Se debe descargar Python en el siguiente enlace Download Python | Python.org , e instalarlo por medio del ejecutable descargado.

PostgreSQL

Se debe descargar PostgreSQL en el siguiente enlace PostgreSQL: Downloads, e instalarlo por medio del ejecutable descargado.

Descargar SmartETL

Se debe de descargar la carpeta ejecutable-SmartETL desde el siguiente enlace SmarETL

Creación de Base de Datos

Dentro de la carpeta ejecutable-SmartETL/DataBase se encuentra el archivo StructureDB.sql en donde se encuentra el código sql para la creación correspondiente.

Instalación de dependencias

Dentro de la carpeta ejecutable-SmarETL se encuentra el archivo requirements..txt mediante el comando

pip3 install -r requirements.txt

Estructura de la aplicación

Backend

Controller

En este carpeta se encuentra la conexión entre el backend y el frontend

DataBase

En esta carpeta se encuentra el script para la creación de Base de Datos

ETL

En esta carpeta se encuentran todos los archivos requeridos para el proceso de ETL

Connector

En esta carpeta se encuentra la conexión a la base de datos

Extract

- ExtractFile.py: Este archivo contiene la lógica para extraer los datos desde los archivos csv.
- ExtractDB.py: Este archivo contiene la lógica para extraer los datos desde la base de datos.

Load

En este archivo se encuentra la lógica para cargar los datos hacia la base de datos

Transform

- Joiner.py: Este archivo contiene la lógica para agrupar los datos con respecto a las tablas de la base de datos.
- Transformer.py: Este archivo contiene la logica para transformar, formatear y filtrar los datos para la posterior agrupación de los mismos

Wrong Data

Esta carpeta contiene archivos con los datos no aceptados y duplicados.

Frontend

MainScreen.py

este archivo contiene el código de la pantalla de la aplicación

Assets

Esta carpeta contiene imágenes utilizadas en la pantalla de la aplicación

Bibliografía

- A. (s. f.). Todo lo que necesitas saber sobre SQLAlchemy. BreatheCode. https://content.breatheco.de/lesson/todo-lo-necesario-para-empezar-usar-sqlalchemy

- Alberca, A. S. (2020, 4 octubre). La librería Numpy. Aprende con Alf. https://aprendeconalf.es/docencia/python/manual/numpy

- /Chacón, J. L. (2022, 5 abril). Introducción a Pandas, la librería de Python para trabajar con datos. Profile Software Services. https://profile.es/blog/pandas-python

- /[Python] psycopg2 se conecta a PostgreSQL e inserta datos json en la base de datos - programador clic. (s. f.). Programador clic. https://programmerclick.com/article/5541137286/

- S. (2015, 24 julio). GitHub - syrusakbary/validate_email: Validate_email verify if an email address is valid and really exists. GitHub. https://github.com/syrusakbary/validate_email