Proyecto ETL: Análisis de ventas.

Extracción, transformación y visualización de datos para el periodo 2023 - 2024.

Andrea Reyes Mejía
Cargo: Candidato a líder analista de datos.
Fecha de entrega: 22/11/2024

Documento técnico del proyecto ETL

Introducción Técnica

Este documento describe el desarrollo técnico del proyecto ETL para el análisis de oportunidades de negocio, implementado con Python y herramientas de visualización. El flujo ETL (Extracción, Transformación y Carga) fue diseñado para procesar datos de ventas del periodo 2023-2024, generando insights clave para la toma de decisiones.

Estructura del Proyecto

El proyecto está organizado de la siguiente manera:

Directorios y Archivos

```
/project
  |
  |-- data/
  |    |-- raw/              # Datos originales sin procesar
  |    |    |-- BD_OPORTUNIDADES_23_24.csv
  |    |
  |    |-- processed/        # Datos procesados y enriquecidos
  |         |-- _NUEVA_BD_OPORTUNIDADES_23_24.xlsx
  |
  |-- src/                   # Código fuente del proyecto
  |    |-- Reto_code.py      # Script principal con el flujo ETL
  |
  |-- dashboard/             # Dashboards y capturas de pantalla
  |    |-- proyecto_power_bi.pbix
  |    |--P1_Zona_ventas2024.png
  |    |--P2_Crecimiento_empresas.png
  |    |--P3_Crecimiento_asesores.png
  |    |--P4_Zona_decremento.png
  |    |--P5_Cliente_zona.png
  |    |--P6_Relacion.png
  |-- docs/             #Documentación del proyecto
  |    |-- Reto.pdf
  |    |-- Manual de usuario.pdf
  |    |-- Diagrama de flujo del proceso ETL.pdf
  |    |-- Reporte de Insights.pdf
  |    |-- Documentación técnica del proyecto ETL.pdf
  |
  |-- README.md              # Documentación general
```

Descripción de los Archivos Clave

- **`Reto_code.py`**: Contiene todo el código para realizar el flujo ETL, incluyendo la extracción, transformación, y exportación de datos.
- **`_NUEVA_BD_OPORTUNIDADES_23_24.xlsx`:** Archivo generado que incluye los datos procesados y enriquecidos.
- **`Reto_dashboard.pbix`:** Archivo de Power BI con dashboards interactivos.

Flujo ETL Detallado

1. Extracción

- **Función Principal**: `cargar_datos`
- **Descripción**: Los datos se cargan desde un archivo CSV utilizando la biblioteca de pandas. Se maneja el formato y codificación para garantizar la integración exitosa.

2. Transformación

Incluye varias subetapas:

a. Limpieza de Datos

- **Función**: `limpiar_datos`
- **Acciones Clave**:
  - Manejo de valores nulos (relleno o eliminación según contexto).
  - Eliminación de registros duplicados.
  - Normalización de formatos (fechas, texto y valores numéricos).

b. Normalización

- **Función**: `normalizar_datos`
- **Descripción**: Conversión de importes a una divisa estándar (MXN), normalizar columna de FechaCierre a tipo date y cálculo de importes en USD y EUR.

c. Generación de Columnas Derivadas

- **Función**: `generar_columnas`
- **Columnas Generadas**:
  - Rangos de importe (Bajo, Medio, Alto).
  - Clasificaciones por zona (Importante u Otras).
  - Segmentación temporal (año, mes, trimestre).

3. Cálculo de Métricas

- **Función**: `calcular_agrupaciones`
    - **Descripción**: Genera datos agregados como densidad de ingresos por zona, empresa y propietario.
- **Función**: `calcular_crecimientos`
    - **Descripción**: Genera datos agregados como crecimientos por zona, empresa y propietario.

4. Carga

- **Función**: `exportar_datos_con_formato`
    - **Descripción**: Exporta los datos procesados a un archivo Excel.
- **Función**: `exportar_agrega_hojas_crecimiento`
    - **Descripción**: Exporta las hojas adicionales con análisis de crecimiento.

Requisitos y Configuración

Herramientas Utilizadas

- **Python 3.x**
- Librerías:
    - pandas
    - openpyxl
    - word2number

Configuración del Entorno

Instala las dependencias:

```
pip install pandas openpyxl word2number
```

1. Verifica las rutas de los archivos en el script principal:

```
ruta_archivo = 'data/raw/BD_OPORTUNIDADES_23_24.csv'

ruta_salida='data/processed/_NUEVA_BD_OPORTUNIDADES_23_24.xlsx'
```

2. Ejecuta el script:

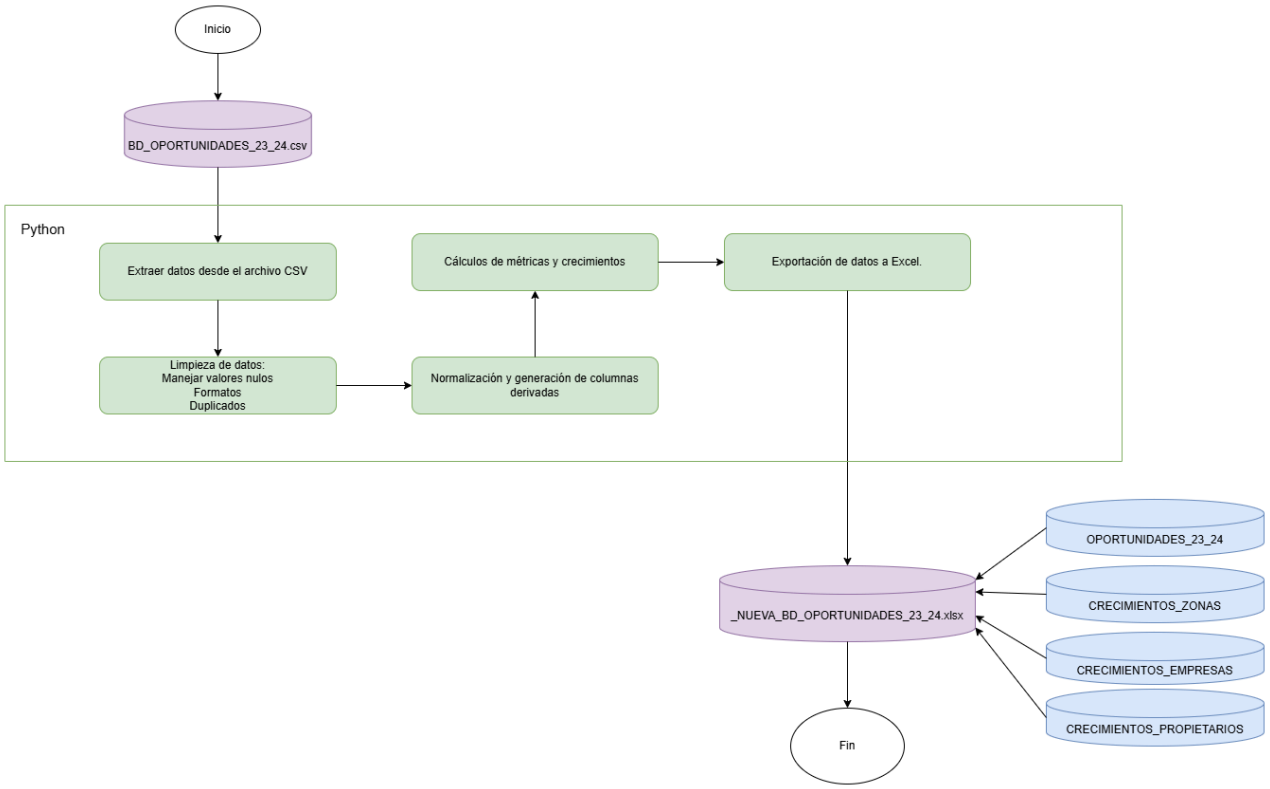
```
python src/Reto_code.py
```

Al momento de ejecutar el script puede tardar unos segundos en mostrar mensajes en la consola, esto no significa que no está ejecutándose, solo es cuestión de esperar unos momentos.

Diagramas

Diagrama de Flujo del Proceso ETL

Representa el flujo de los datos a través de las etapas de extracción, transformación y carga. Este diagrama ilustra cómo se gestionan los datos desde su origen hasta su visualización final.



Pruebas Realizadas

Validaciones de Limpieza

- **Duplicados**: Se eliminaron X registros duplicados.
- **Valores Nulos**:
    - Importe: Reemplazados por 0.
    - Participantes: Asignados como 0 en "Sin Datos".

Resultados de Pruebas

- La transformación generó columnas consistentes y correctamente normalizadas.
- Las exportaciones a Excel fueron validadas manualmente para confirmar la integridad de los datos.

Conclusiones Técnicas

- **Desafíos**:
    - Manejo de cálculos de crecimiento infinito.
    - Optimización de procesamiento para grandes volúmenes de datos.
- **Soluciones Implementadas**:
    - Reglas claras para valores nulos y normalización.
    - Exportación en formato Excel con estilos personalizados para facilitar su uso.
- **Futuras Mejoras**:
    - Escalabilidad para grandes bases de datos.
    - Automatización de la generación de dashboards a partir de datos procesados.