EVIDENCIA DE APRENDIZAJE 1 - CREACIÓN DE UNA BASE DE DATOS ANALÍTICA

Isabela Arango Verona

Institución Universitaria Digital de Antioquia

BigData (PREICA2501B020109)

Prof. Andres Felipe Callejas Jaramillo

9 de mayo del 2025

Índice

Índice de figuras

Introducción

El presente informe describe el proceso de diseño e implementación de un sistema de gestión de base de datos (SGBD) para el almacenamiento, procesamiento y análisis de datos demográficos a nivel mundial. La necesidad de comprender las dinámicas poblacionales de los diferentes países y dependencias es crucial para la toma de decisiones en diversos campos, como la planificación urbana, la gestión de recursos, la formulación de políticas públicas y la investigación social. Este proyecto se enfoca en la creación de una base de datos robusta y un esquema bien definido, utilizando información obtenida de una fuente web confiable, para facilitar el análisis y la generación de conocimiento a partir de estos datos.

Descripción del problema

La información demográfica mundial es vasta y se encuentra dispersa en diversas fuentes. Para realizar análisis comparativos y longitudinales eficientes, es necesario consolidar estos datos en un sistema estructurado. El problema abordado en este proyecto es la falta de una base de datos centralizada y optimizada para el almacenamiento y la consulta de información demográfica de los países y dependencias del mundo. Esto dificulta la extracción de información relevante, la identificación de tendencias y la generación de informes para la toma de decisiones.

La necesidad de un sistema que permita la captura, el almacenamiento eficiente, el procesamiento para la obtención de indicadores clave y la visualización de estos datos es fundamental para comprender mejor las dinámicas poblacionales globales.

Objetivos

Objetivo general

Diseñar e implementar un sistema de gestión de base de datos relacional para el almacenamiento, procesamiento y análisis de datos demográficos mundiales, utilizando información obtenida de fuentes web públicas.

Objetivos específicos

- Capturar datos demográficos actualizados de países y dependencias desde la fuente web especificada (Worldometer).
- Diseñar un esquema de base de datos relacional eficiente para almacenar los datos demográficos, garantizando la integridad y la consistencia de la información.
- Implementar la base de datos utilizando un Sistema de Gestión de Bases de Datos (SGBD) adecuado.

Descripción de los datos disponibles

Los datos utilizados en este proyecto se obtuvieron de la página web https://www.worldometers.info/world-population/population-by-country/. Esta página proporciona una tabla con información detallada sobre la población de los países y dependencias del mundo, incluyendo los siguientes campos:

- **#:** Número de orden del país en la tabla.
- **País (o dependencia):** Nombre del país o dependencia.
- **Población (2024):** Población estimada para el año 2024.
- **Cambio anual:** Tasa de cambio anual de la población (%).
- **Cambio neto:** Cambio neto en la población.
- **Densidad** (P/Km^2)**:** Densidad de población (personas por kilómetro cuadrado).
- **Superficie** (Km^2)**:** Superficie terrestre en kilómetros cuadrados.
- **Migrantes (neto):** Número neto de migrantes.
- **Tasa fertilidad:** Tasa de fertilidad (nacimientos por mujer).
- **Edad media:** Edad mediana de la población.
- **Poblacion urbana %:** Porcentaje de la población que reside en áreas urbanas.
- **Participacion mundial:** Porcentaje de la población mundial que representa este país.

Estos datos ofrecen una visión general de la situación demográfica actual y las tendencias de cambio a nivel global (Worldometer, 2025).

Solución propuesta

Elección del SGBD

Para este proyecto, se ha seleccionado SQLite como el Sistema de Gestión de Bases de Datos (SGBD). SQLite es una biblioteca de C que proporciona una base de datos SQL pequeña, rápida, autocontenida, de alta fiabilidad y completamente integrada. Se eligió SQLite por las siguientes razones:

- **Simplicidad:** Es fácil de configurar y utilizar, lo que facilita la implementación para un proyecto de esta escala.
- **Portabilidad:** La base de datos se almacena en un único archivo, lo que facilita su manipulación y transporte.
- **Integración:** Puede integrarse fácilmente con lenguajes de programación como Python, que se utiliza en el código proporcionado para la captura y manipulación de datos.
- **Rendimiento:** Para el volumen de datos esperado en este proyecto, SQLite ofrece un rendimiento adecuado para las operaciones de lectura y escritura.

Esquema diseñado

Se propone un esquema de base de datos relacional que consta de las siguientes tablas (ver también la Figura 1 y Figura 2):

1. Country (Tabla Principal): Almacena los datos demográficos brutos obtenidos de la fuente web.
    - num (INTEGER, PRIMARY KEY)
    - pais (TEXT)
    - poblacion_2024 (INTEGER)
    - cambio_anual (REAL)
    - cambio_neto (REAL)

- densidad_p_km2 (REAL)
- superficie_km2 (REAL)
- migrantes_neto (REAL)
- tasa_fertilidad (REAL)
- edad_mediana (REAL)
- porcentaje_poblacion_urbana (REAL)
- participacion_mundial (REAL)
- fecha_creacion (DATETIME)
- fecha_update (DATETIME)

2. AuditFields (Tabla Abstracta): Define los campos de auditoría comunes.
    - fecha_creacion (DATETIME)
    - fecha_update (DATETIME)
3. ProcessedData (Tabla de Datos Procesados - KPIs): Almacena los indicadores clave de rendimiento calculados a partir de los datos brutos.
    - id (INTEGER, PRIMARY KEY AUTOINCREMENT)
    - country_num (INTEGER, FOREIGN KEY REFERENCES Country(num))
    - año (INTEGER)
    - tasa_crecimiento_anual (REAL) - Directamente del cambio_anual.
    - variacion_densidad (REAL) - Directamente de densidad_p_km2.
    - potencial_crecimiento_urbano (REAL) - Directamente de porcentaje_poblacion_urbana.
    - relacion_migrantes_poblacion (REAL) - Calculado como migrantes_neto / poblacion_2024.
    - fecha_creacion (DATETIME)
    - fecha_update (DATETIME)

4. ReportView (Vista de Reporte): Presenta una selección de datos e indicadores para un análisis temporal simulado.
    - id (INTEGER, PRIMARY KEY AUTOINCREMENT)
    - año (INTEGER)
    - pais (TEXT)
    - tasa_crecimiento_anual (REAL)
    - variacion_densidad (REAL)
    - potencial_crecimiento_urbano (REAL)
    - relacion_migrantes_poblacion (REAL)
    - fecha_creacion (DATETIME)
    - fecha_update (DATETIME)

Las relaciones entre las tablas son las siguientes:

La tabla ProcessedData tiene una relación de uno a muchos con la tabla Country a través de la clave foránea country_num, permitiendo asociar los KPIs calculados con cada país.

La tabla ReportView también se relaciona con la tabla Country para mostrar los datos e indicadores por país y año.

Todas las tablas incluyen los campos de auditoría fecha_creacion y fecha_update para rastrear la creación y modificación de los registros.

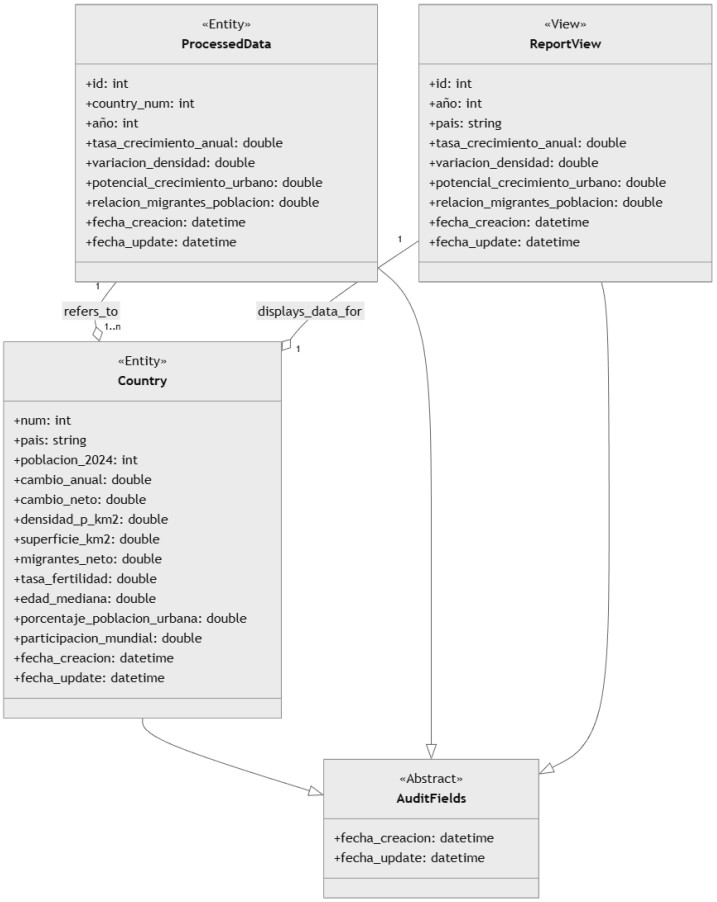


Figura 1. Diagrama de clases.

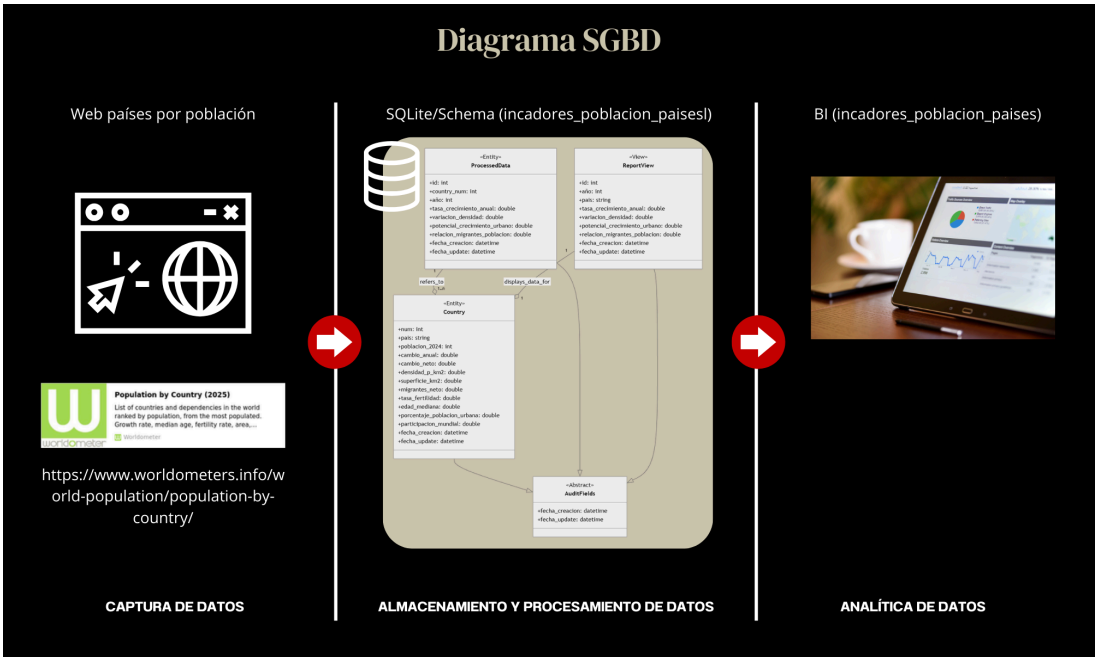
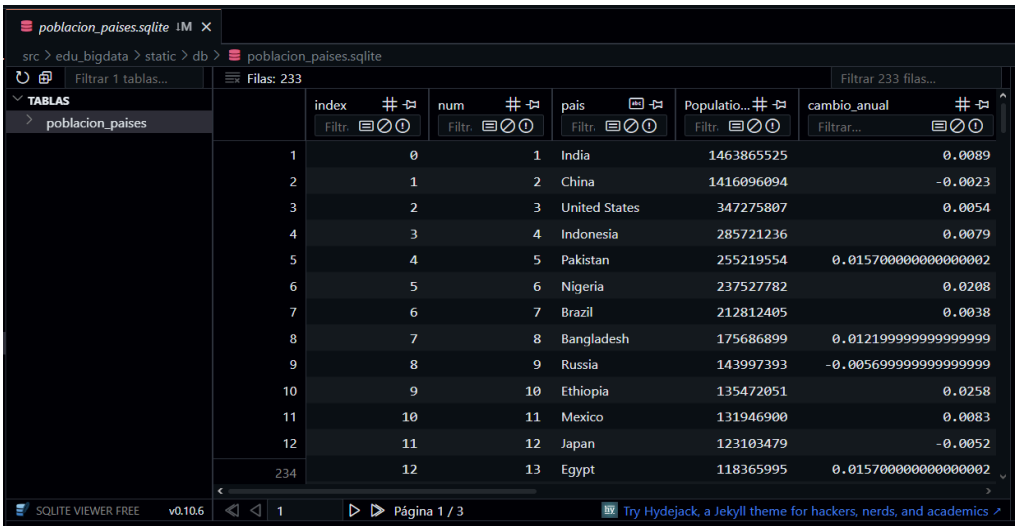


Figura 2. Diagrama SGBD.

Metodología empleada

1. **Captura de Datos:** Se utilizó el script de Python *dataweb.py* para acceder a la página web de Worldometer y extraer la tabla de datos demográficos utilizando la biblioteca pandas y requests. Se implementaron encabezados de agente de usuario para simular una navegación web real y evitar problemas de bloqueo por parte del servidor.
2. **Limpieza y Transformación de Datos:** Los datos extraídos se limpiaron utilizando la función *limpiar_datos* en la clase *DataWeb*. Esta función realizó las siguientes operaciones:
    - Eliminación del símbolo '%' y comas de las columnas numéricas.
    - Reemplazo del símbolo '−' con '-' para representar valores negativos.
    - Conversión de las columnas relevantes a tipos de datos numéricos (float o int).
    - Normalización de los porcentajes dividiéndolos por 100.
3. **Almacenamiento de Datos:** Se utilizó la clase *DataBase* y su método *insert_data* para almacenar los datos limpios en una base de datos SQLite llamada *poblacion_paises.sqlite*. La tabla *poblacion_paises* (que corresponde a la tabla *Country* en el diseño) se creó o sobrescribió con los nuevos datos (Figura 3).



	index	num	pais	Populatio...	cambio_anual
1	0	1	India	1463865525	0.0089
2	1	2	China	1416096094	-0.0023
3	2	3	United States	347275807	0.0054
4	3	4	Indonesia	285721236	0.0079
5	4	5	Pakistan	255219554	0.015700000000000002
6	5	6	Nigeria	237527782	0.0208
7	6	7	Brazil	212812405	0.0038
8	7	8	Bangladesh	175686899	0.012199999999999999
9	8	9	Russia	143997393	-0.005699999999999999
10	9	10	Ethiopia	135472051	0.0258
11	10	11	Mexico	131946900	0.0083
12	11	12	Japan	123103479	-0.0052
234	12	13	Egypt	118365995	0.015700000000000002

Figura 3. Base de datos creada.

4. **Diseño e Implementación del Esquema de la Base de Datos:** Se diseñó el esquema relacional descrito en la sección anterior, incluyendo la tabla principal Country, la tabla de KPIs *ProcessedData* y la vista de reporte *ReportView*. Aunque el código proporcionado principalmente se enfoca en la creación y manipulación de la tabla *Country*, el diseño propuesto considera la creación de tablas adicionales para el procesamiento y la presentación de datos.
5. **Documentación:** Se elaboró el presente informe para describir el proceso completo, desde la identificación del problema hasta la propuesta de la solución y los resultados esperados.

Resultados y conclusiones

Resultados

- Se logró capturar y limpiar los datos demográficos de la página web de Worldometer utilizando el script *dataweb.py*.
- Los datos limpios se almacenaron exitosamente en una base de datos SQLite llamada *poblacion_paises.sqlite* a través de la clase *DataBase*.
- Se diseñó un esquema de base de datos relacional que incluye tablas para los datos brutos (*Country*), los indicadores clave de rendimiento (*ProcessedData*) y una vista para la generación de reportes (*ReportView*), incorporando campos de auditoría para el seguimiento de los datos.
- Se definieron cuatro KPIs relevantes para el análisis demográfico que podrían ser calculados y almacenados en la tabla *ProcessedData*.
- Se conceptualizó la creación de una vista (*ReportView*) para presentar los datos e indicadores de manera organizada para el análisis temporal.

Conclusiones

El proyecto ha demostrado la viabilidad de crear un sistema de gestión de base de datos para el almacenamiento y el potencial análisis de datos demográficos mundiales obtenidos de fuentes web públicas. La utilización de SQLite como SGBD ofrece una solución simple y eficiente para la escala de datos manejada. El diseño del esquema de la base de datos, que incluye tablas para los datos brutos, los KPIs procesados y una vista de reporte, proporciona una estructura sólida para futuras etapas de análisis y visualización.

La implementación de KPIs permite transformar los datos brutos en información más significativa para la toma de decisiones. La inclusión de campos de auditoría en todas las tablas garantiza la trazabilidad y la integridad de los datos.

Si bien el código proporcionado se centra en la captura y el almacenamiento inicial de los datos, el diseño propuesto sienta las bases para la expansión del sistema con funcionalidades de cálculo de KPIs y la creación de vistas de reporte dinámicas.

Bibliografía

Worldometer. (2025). *Population by Country (2025)*. Worldometer. Retrieved May 8, 2025, from https://www.worldometers.info/world-population/population-by-country/

Anexos

Repositorio de la implementación

https://github.com/Isa-av/bigdata_2025_1_2