Automated GDP Extractor – ETL Pipeline Project

Tools: Python, BeautifulSoup, pandas, SQLite, CSV, Logging
Role: Simulated Junior Data Engineer (Project-Based)
Project Type: Data Engineering / Web Scraping / ETL Automation
Date: April 2025

Project Overview

In a simulated role as a Junior Data Engineer for a global business expansion scenario, I designed and implemented an automated ETL (Extract, Transform, Load) pipeline to collect and process country-level GDP data from the International Monetary Fund (via an archived Wikipedia page). This project mimics real-world requirements where companies need up-to-date economic data to support global decision-making.

Project Breakdown

1. Data Extraction

- Scraped structured GDP data from an archived Wikipedia page using `requests` and `BeautifulSoup`.
- Parsed relevant country and GDP fields into a pandas DataFrame.

2. Data Transformation

- Cleaned and converted GDP values from formatted strings to floats.
- Transformed values from millions to billions and rounded to 2 decimal places.
- Renamed the GDP column for clarity.

3. Data Loading

- Saved the transformed dataset to a `.csv` file (`Countries_by_GDP.csv`).
- Loaded the data into a SQLite database (`World_Economies.db`) using pandas' `to_sql()` method.

4. Querying and Logging

- Queried all countries with GDP ≥ $100B using SQL.
- Implemented logging to track the status of each ETL phase in `etl_project_log.txt`.

Key Outputs

- CSV file of all countries with GDP values
- SQLite database with GDP table for querying
- Logged ETL execution timeline for transparency and reproducibility