Final Project: Database Applications Development

ETL and Data Visualization for GreenLab

Course: Database Applications Development
Semester: Winter 2023

Introduction

This project allows students to demonstrate their skills in building an application that extracts, transforms, and loads (ETL) data into a Data warehouse, then utilizes the data for a Data plotter application. The final goal is to generate personalized plots for visualization and storage.

Conventions

- **Client:** Refers to GreenLab
- **Consultants:** Refers to student members of each group

Description

GreenLab currently downloads data in CSV format from public sources (e.g., Our World in Data) and processes it manually using MS Excel to generate plots. With global expansion in mind, they want to develop a dynamic website allowing users to generate personalized plots.

Data Files

- **annual-co2-emissions-per-country.csv:** Annual CO2 emissions per country
- **nitrous-oxide-emissions.csv:** Annual nitrous oxide emissions per country
- **methane-emissions.csv:** Annual methane emissions per country
- **population-and-demography.csv:** Population and demographic information
- **gdp-per-capita.csv:** GDP per capita per country

Technical Requirements

- **Primary Language:** Python
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn
- **Environment:** JupyterLab
- **ETL Process:** Automate weekly data pipeline, store production-ready data in a Data Warehouse (e.g., AWS Redshift)
- **Graphs:** Generate static plots using Matplotlib or Seaborn, maintain a historical archive of graphs

Deliverables

Proof of Concept (POC)

- Analysis document exploring data processing technologies
- ETL process that extracts, transforms, and loads data locally
- Automated ETL pipeline running weekly or as needed
- Local plotting of required graphs using production-ready data
- Functions usable as Python scripts outside JupyterLab

Final Submission and Presentation

- Team description, roles, and skills
- Detailed description of technologies used
- Complete working solution including:
  - Data extraction from source files
  - Data transformation for analysis
  - Loading data to storage system
  - Data visualization capabilities
  - User interface for graph generation