**ETL Workflow for the World's Largest Banks Data Using Python**

**Project Description**

This project demonstrates a Python-based ETL (Extract, Transform, Load) workflow for processing data about the world's largest banks. The project utilizes various libraries including:

* **Pandas** for data manipulation
* **Beautiful Soup** for web scraping
* **NumPy** for numerical computations
* **MySQL** for database management
* **Matplotlib** for data visualization

**Workflow Stages**

1. **Data Extraction:** Scrape data from a Wikipedia page listing the largest banks using Beautiful Soup.
2. **Staging Area:** Store the scraped data in a Pandas DataFrame for further processing.
3. **Data Transformation:** Clean and manipulate the data using Pandas and NumPy. Additional data sources like CSV files (e.g., exchange rates) can be incorporated here.
4. **Data Loading:** Load the processed data into a MySQL database for efficient storage and querying.
5. **Data Visualization:** Query the MySQL database and generate visualizations using Matplotlib.

**Tech Stack**

* Python
* Beautiful Soup
* Pandas
* MySQL
* Matplotlib

**Dataset**

* **URL for Data Scraping:** List of Largest Banks <https://web.archive.org/web/20230908091635/https:/en.wikipedia.org/wiki/List_of_largest_banks>

**Steps**

1. **Project Setup:** Install required libraries and configure access to the MySQL database.
2. **Data Extraction (Task 1):** Scrape data from the Wikipedia URL using Beautiful Soup and store it in a Pandas DataFrame.
3. **Data Transformation (Task 2):** Preprocess the scraped data using Pandas and NumPy. This may involve cleaning, handling missing values, and incorporating additional data from sources like the exchange\_rate.csv file.
4. **Loading to CSV (Task 3):** Save the transformed data as a CSV file for backup or further analysis.
5. **Loading to Database (Task 4):** Load the transformed data into a MySQL database table.
6. **Database Queries (Task 5):** Implement functions to query the database for specific insights or analysis.

**Benefits for Beginners**

This project offers valuable hands-on experience with key data engineering concepts and tools, making it ideal for beginners to learn:

* Web scraping with Beautiful Soup
* Data transformation and cleaning with Pandas and NumPy
* Data management with MySQL
* Data visualization with Matplotlib

The step-by-step approach allows beginners to follow along, build their skills progressively, and gain a comprehensive understanding of the ETL process.

**Attached files:**

* exhange\_rate.csv
* Task 1 and 2 Data\_Scraping\_and\_Transformation.ipynb
* Task 3 transformed\_data.csv
* Task 4 visualize\_data.xlsx
* Task 5 SQL\_Queries\_On\_Transformed\_data.sql