Hands-on Lab: Acquiring and Processing Information on the World's Largest Banks

Estimated Time: 60 mins

In this project, you will put all the skills acquired throughout the course and your knowledge of basic Python to test. You will work on real-world data and perform the operations of Extraction, Transformation, and Loading (ETL) as required.

Disclaimer:

> Cloud IDE is not a persistent platform, and you will lose your progress every time you restart this lab. We recommend saving a copy of your file on your local machine as a protective measure against data loss.

Project Scenario:

You have been hired as a data engineer by research organization. Your boss has asked you to create a code that can be used to compile the list of the top 10 largest banks in the world ranked by market capitalization in billion USD. Further, the data needs to be transformed and stored in GBP, EUR and INR as well, in accordance with the exchange rate information that has been made available to you as a CSV file. The processed information table is to be saved locally in a CSV format and as a database table.

Your job is to create an automated system to generate this information so that the same can be executed in every financial quarter to prepare the report.

Particulars of the code to be made have been shared below.

Parameter	Value
Code name	`banks_project.py`
Data URL	`https://web.archive.org/web/20230908091635 /https://en.wikipedia.org/wiki/List_of_largest_banks`
Exchange rate CSV path	https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMSkillsNetwork-PY0221EN-Coursera/labs/v2/exchange_rate.csv
Table Attributes (upon Extraction only)	`Name, MC_USD_Billion`
Table Attributes (final)	`Name, MC_USD_Billion, MC_GBP_Billion, MC_EUR_Billion, MC_INR_Billion`
Output CSV Path	`./Largest_banks_data.csv`
Database name	`Banks.db`
Table name	`Largest_banks`
Log file	`code_log.txt`

Project tasks

Task 1:
Write a function `log_progress()` to log the progress of the code at different stages in a file `code_log.txt`. Use the list of log points provided to create log entries as every stage of the code.

Task 2:
Extract the tabular information from the given URL under the heading 'By market capitalization' and save it to a dataframe.
a. Inspect the webpage and identify the position and pattern of the tabular information in the HTML code
b. Write the code for a function `extract()` to perform the required data extraction.
c. Execute a function call to `extract()` to verify the output.

Task 3:
Transform the dataframe by adding columns for Market Capitalization in GBP, EUR and INR, rounded to 2 decimal places, based on the exchange rate information shared as a CSV file.
a. Write the code for a function `transform()` to perform the said task.
b. Execute a function call to `transform()` and verify the output.

Task 4:
Load the transformed dataframe to an output CSV file. Write a function `load_to_csv()`, execute a function call and verify the output.

Task 5:
Load the transformed dataframe to an SQL database server as a table. Write a function `load_to_db()`, execute a function call and verify the output.

Task 6:
Run queries on the database table. Write a function `load_to_db()`, execute a given set of queries and verify the output.

Task 7:
Verify that the log entries have been completed at all stages by checking the contents of the file `code_log.txt`.