So I just completed some ETL projects recently.

In this project, I web scrapped data of Countries GDP from Wikipedia website. <https://web.archive.org/web/20230902185326/https://en.wikipedia.org/wiki/List_of_countries_by_GDP_%28nominal%29> . The webpage contains information about the GDP of countries for the year 2023. I extracted the data compiled by the IMF.

The libraries I used included:

* Requests: to make http requests to the webpage and to retrieve the html content
* BeautfulSoup: to parse the html of the webpage
* Pandas: to organize the parsed webpage into table and also to transform the data
* Sqlite3: to run queries on the database
* Datetime: to log the processes

Functions were created for the following functions:

* I created functions to extract the required data from the webpage,
* transform the GDP from millions to billions,
* load the data to a csv file,
* load the data to a database
* run queries on the database table
* log the message of each process to a text file, so one can follow the process.

I codes were written and tested in a Jupyter notebook for testing. Jupyter allows for easy localizing of portions of the code which makes it easy to debug codes. After I was sure the codes were free of bugs, I converted the notebook to a python script which was then executed through the terminal.

The source code and related files are available in <https://github.com/Oluwajuwon-O/IBM-data-engineering.git>

A short walkthrough of the project is attached.

Please leave a comment. Thank you.

I’m Oluwajuwon Oyalude. I’m currently looking for entry level, junior roles, internships and opportunities in data engineering and data science. Thank you in advance for any connections, advice, or opportunities you can offer.