

Take Home Exercise

One of the research teams has requested to scrape data from Baker Hughes (BH) in order to understand changes in the weekly rig report. Your task is to create a schema that will house the data and an ETL to extract the data from the Excel file provided and also to design some SQL tables that will house the data.

Task 1 - PostgreSQL Schema/Table Creation:

Write a SQL query to create a schema that will house the data from both tabs in the provided Excel file.

- Design PostgreSQL tables to appropriately store the data retrieved from the provided Excel file. Consider the data types, relationships, and any necessary constraints.

Output Required:

- A raw SQL script that creates a schema called "**web_scrapes**" and tables to store the data.
- Use Python/SQLAlchemy to create a schema called "**web_scrapes**" and tables to store the data.

Task 2 - Python ETL Development:

Create an ETL pipeline to read the "Baker Hughes" Excel file, transform/clean data as required and then load the data into the tables you have created. You are expected to extract data from both tabs of the Excel file.

- Using Python, develop an ETL pipeline to extract the data from the provided Excel file.
- Transform the extracted data as needed, such as cleaning, filtering, or aggregating, to ensure its compatibility with the SQL tables.
- Load the transformed data into the SQL tables designed in the previous task.

Output Required:

- Python file(s) that contain your code to Extract, Transform and Load the "Baker Hughes" Excel file into the PostgreSQL tables you created.

Task 3 - Packaging ETL Scripts (Bonus Question):

- Package the developed ETL scripts to ensure portability and ease of execution on any machine.
- Include clear instructions on how to set up and run the ETL process, along with any necessary dependencies or configurations.

Feel free to use any libraries or tools that you find suitable for completing the tasks. Remember to include comments and notes within your code to enhance readability and provide explanations where necessary.

Please ensure that the files you send over are in a **.zip** file format.