

Web Scrape

Extraction

This script uses 'requests' and 'BeautifulSoup' from the 'bs4' library to scrape SpaceX's launch schedule web page. It sends a GET request to the page, and upon receiving the HTML content, it parses the page to find all links leading to individual launch detail pages.

It collects the URLs for each individual launch's detail page by extracting the 'href' attribute from every 'a' tag within the specified class ('article-header')

For each URL, the script performs another GET request to retrieve the individual launch detail page, and then it scrapes the text details from the 'p' tags of the HTML content

Transformation

The scraped details for each launch are structured into a list of dictionaries, where each dictionary contains the URL and corresponding text details of a launch.

The list of dictionaries is then converted into a 'pandas' DataFrame, creating a tabular format of the data that includes columns for the URL and details.

Load

The script creates a connection to the MySQL database using the 'sqlalchemy' library with the provided RDS credentials.

The DataFrame containing the launch details is then inserted into the MySQL database. The 'to_sql' method is used to write records stored in the DataFrame to a SQL table in the database, appending the new data if the table already exists.