**Problem Statement**

* Design a web scraper to read articles off theverge.com using Python. The web scraper should be able to perform the following tasks:
* Read the headline, get the link of the article, the author, and the date of each of the articles found on theverge.com.
* Store these data in a CSV file titled ddmmyyy\_verge.csv, with the following header id, URL, headline, author, date using the csv library.
* Create an SQLite database to store the same data, and make sure that the id is the primary key using the sqlite3 library.
* Save the articles (and de-duplicate them) daily on the server in a SQL database.

**My Solution**

* First, I start the script by importing all the required libraries such as Beautiful Soup, Pandas, NumPy etc.
* Then I defined a variable website that contains the URL of The Verge website and sends a request to it using the requests get() method.
* The response is then passed to Beautiful Soup constructor along with an HTML parser to create a Beautiful Soup object.
* After that, I read a CSV file named “ddmmyyy\_verge.csv” which is empty but has header names id, url, link, author and date using pandas.
* Next, I extract all links of articles from the website using Beautiful Soup find\_all() method.
* Then I extract all headlines of articles from the website using BeautifulSoup find\_all() method.
* Then I extract all dates of every article from the website using BeautifulSoup find() method.
* Then I extract authors of all articles from the website using BeautifulSoup find\_all() method.
* Then I create a unique ID author of all articles from the website.
* Then I create a dataframe called ‘df1’ with columns ‘id’, ‘url’, ‘headline’, ‘author’, and ‘date’. The values for each column are passed as a list to the pd.DataFrame() function.
* Then I converted the ‘date’ column in dataframe df1 from string format to datetime format using pd.to\_datetime() function.
* Then I extracted only the date part using. dt.strftime('%Y/%m/%d').
* Then I concatenated two dataframes df and df1 along axis 0 (i.e., row-wise) and stored the result in a new dataframe.
* Then I drop all columns that have all missing values (NaN) in the dataframe.
* Then I removed duplicate rows from the dataframe.
* Then I wrote all the contents of the dataframe df1 to a CSV file named ‘ddmmyyy\_verge.csv’.
* Finally, I created a SQLite database and appended the data from a pandas dataframe into a table in the database.