**Laugh-Out-Loud – An ETL project: Written Write-Up**

**Extract:**

The data for this ETL project was scraped from the video tabs of the Youtube pages of five popular Late Night comedians. The scrape was specified to include just the videos posted between January 2020 to January 2021. The desired data included the titles of the videos, the runtime of each video, the number of views each video had received over its lifetime, and the comedian hosting in each video. The total lifetime of the video was also available; but this was not considered relevant. As the data was contained as a CSS label inside an HTML anchor tag, extraction was performed in two stages: the first stage specified the parent tag for the relevant HTML tag; and the second stage extracted the contents of the CSS label.

The resulting output was in string format and preliminarily placed inside a list of strings.

**Transform:**

Data in string format presents a unique challenge in that most Python operations in pandas are specifically defined to handle integer or numeric data types. To transform the data into a useable format, Regular Expression was performed to loop through each list of strings to extract the title, runtime, viewership, of each video, along with the comedian hosting the video. Each piece of data was defined as a variable through regular expression then put he variables were placed inside a dictionary of variables for each video, each of these dictionaries was then appended to a list for each comedian; and to a larger list for all the comedians. As such data from about 5,000 videos was scraped; then transformed

Once this process was completed, the larger list for all comedians was converted into a dataframe, although the lists for each individual comedian remined saved as variables As a final transformation, the video length column in the dataframe was converted to a float variable converting the minute and second time string to a number indicating the runtime in seconds.

**Load:**

The final step of the ETL pipeline involved placing the Comedians dataframe into an SQL database through Postgres in a table rather imaginatively named “Comedians.” A SQL database was preferred due to greater familiarity with SQL operations; and the simplicity of the dataframe which negated the primary advantages of NoSQL databases