Movie Box Office Collection Data ETL

Overview

One of the key projects is the exercise is to see the effectiveness in building code to extract, transform and load data of movie data sets.

Goals

1. Create Code to Crawl data from https://boxofcecollection.in/for various movies
2. Clean data to extract relevant information
3. Loop through multiple movies based on the config file
4. Store output in SQL table

Specifcations

Box Office Collection is a site which has information of the latest Indian movies and their India Gross Collection. Their day wise prediction is available on pages with URL structure in the following format:
https://boxofcecollection.in/luka-chuppi-box-ofce-collection-day-wisewhich is a combination of

Base url = https://boxofcecollection.in/
Movie name = luka-chuppi
Tail url = -box-office-collection-day-wise

This data on the website is available in following format, where Day stands for day from release of the film

Total ₹92.05 Cr

Day 1 ₹8.01 Cr

Day 2 ₹10.08 Cr

Day 3 ₹14.04 Cr

Day 4 ₹7.90 Cr

This data has to be read for five movies (defined in a config file), and stored in a different format in a local SQL table.

Data Crawl

To crawl and parse the page HTML, Beautiful Soup library can be used. The header row of the data will have to be removed.

File Structures

Config file

The input config file will have details on movies and their release date to be used in URL generation for boxofcecollection.in and to store data in SQL table

The SQL table should have the following fields:

Movie Name

Days from Release

Date

Box Office Collection

Note: The SQL table creation should also happen via script.

Data Transformation

Following changes need to be made to arrive at the above columns

Movie Name- This data should store the name of movies after removing “-” from input data

Days from Release - This value should be the frst feld acquired from boxofcecollection.in (Day1, Day2) without any change

Date - Based on the release date, this date indicates the exact date of the collection in YYYY-MM-DD. If the movie release date is 2019-03-01 and Days from release is 2, then the date should be stored as 2019-03-02 (MM is 03, DD is 02 and YYYY is 2019 as it is the second day)

Box Office Collection - This should be stored as a float variable after taking out the currency and other signs. So, 17.90 Cr will be stored as 179000000

Collating Data of Multiple Movies

The code should be written in a manner that the data of all movies stored in the config file is stored in one table with all columns as mentioned above.