**ETL (Extract, Transform, Load)**

**is a fundamental process used in data management systems, especially for building data warehouses. It consists of three main stages:**

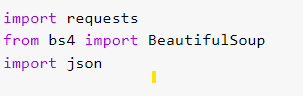
**Extract:  
Data is collected from various sources, such as databases, CSV files, or other systems. The goal is to pull raw data from its original sources while maintaining its quality and validity**.

**Task 1: Data Extracting (Scraping):**

**Code Explanation: Extracting Links for IMDb's Top 250 Movies This script uses the requests library to fetch the IMDb webpage containing the Top 250 movies, BeautifulSoup to parse and extract specific elements from the webpage, and json to process structured data in JSON format. Below is a detailed explanation:**

**A)**

1. **Importing Libraries:**

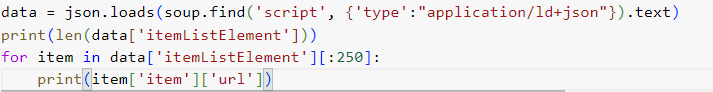
****

* **requests: Used to send HTTP requests to fetch web pages.**
* **BeautifulSoup: Used to parse HTML content and extract desired elements.**
* **json: Used to handle JSON-formatted data and convert it into Python structures like dictionaries.**

**2.Setting Up HTTP Headers:**

**3.Fetching the IMDb Top 250 Page:**

**4. Extracting JSON Data:**

**5. Printing the Number of Movies and Extracting URLs:**

**6.Output:**

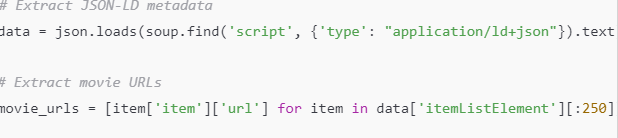
1. **The total number of movies in the list is displayed.**
2. **The script prints the IMDb URLs for the Top 250 movies.**

****

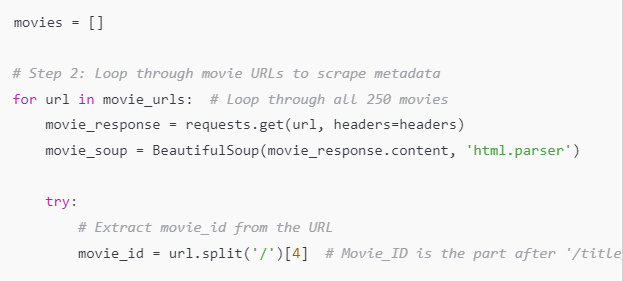
**B)1**

1. **Fetch IMDb Top 250 Metadata:**

* **Purpose: Prevents IMDb from identifying the script as a bot. These headers mimic a real browser.**

**2: Extract Movie URLs from JSON-LD Metadata:**

* **json.loads: Parses the JSON-LD data embedded in the page, which contains metadata about each movie.**
* **movie\_urls: Extracts the URLs of the top 250 movies from the JSON data to be scraped later.**

**3: Loop through Movie URLs to Scrape Metadata:**

* **movie\_urls loop: Iterates through each movie URL to fetch detailed information about the movie.**
* **movie\_id: Extracts the unique movie ID from the URL by splitting the URL and accessing the correct segment.**

**4: Extract Additional Metadata (Title, Rating, Director, etc.):**

****

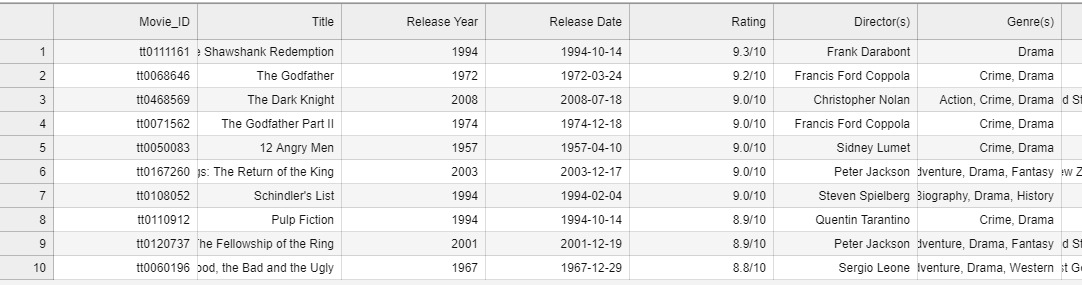
* title: Scrapes the movie title from the page.
* release\_year: Extracts the release year from the JSON-LD data if available.
* release\_date: Extracts the full release date from the JSON-LD data.

**5: Save the Data into a DataFrame and CSV File:**

****

* **movies\_df: Converts the list of movie data into a Pandas DataFrame for easier handling and analysis.**
* **to\_csv: Saves the DataFrame to a CSV file for later use or analysis.**

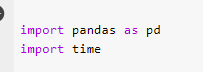
**A screenshot of a computer

Description automatically generated**

**B)2**

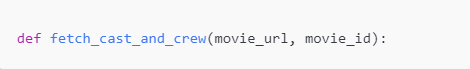
**Explanation of the Code: Extracting Cast and Crew Data for IMDb Movies This script is designed to scrape cast and crew information for a list of IMDb movies, store the data in a structured format using Pandas, and save it to a CSV file. Below is a detailed explanation of the script:**

1. **Importing Required Libraries:**

****

* **time**: Used to introduce delays between web requests to avoid overloading the server.

1. **Function to Fetch Cast and Crew Details :**

****

**3 . List of IMDb Movie URLs:**

****

1. **Loop to Process Each Movie:**

****

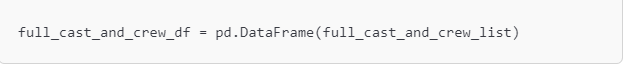
 Loops through each movie URL.

 Extracts the movie ID from the URL (e.g., tt0111161).

 Calls the fetch\_cast\_and\_crew function to get cast and crew details.

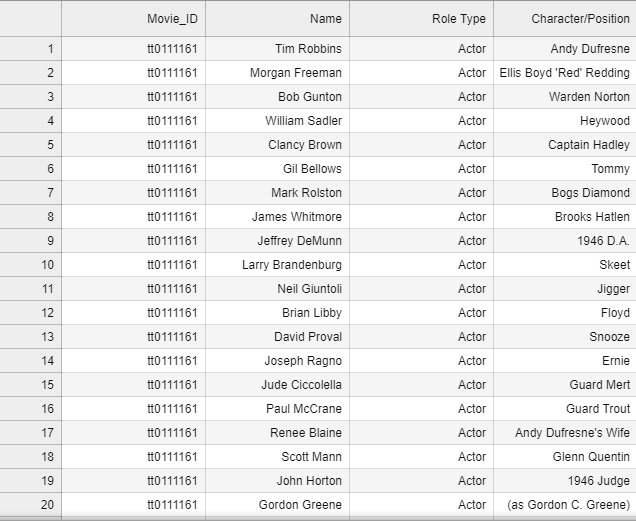
 Appends the results to the full\_cast\_and\_crew\_list.

**4.Storing and Saving Data:**

****

* Converts the list of cast and crew data into a Pandas DataFrame for easy manipulation and analysis.
* Saves the DataFrame to a CSV file for future use.

**5.Example Output:**



**Task2: Data Preprocessing and Transformation**

**Transform:  
In this stage, the data is cleaned, standardized, and formatted to meet the requirements of the data warehouse. This includes improving data quality, handling missing values, converting formats, and integrating data from multiple sources.**

**cleaned imbd movies**

**1: Load the Scraped Data**

****

**2: Clean the Data:**

* **Handle Missing Values:**
* **Remove Duplicates:**

****

* **Correcting Inconsistencies (Budget and Gross Revenue):**

****These lines remove any non-numeric characters (such as dollar signs or commas) from the Budget and Gross Worldwide Revenue columns, leaving only the numeric values.

The Budget and Gross Worldwide Revenue columns are converted to numeric values. Any errors during conversion are handled, and invalid entries are replaced with "N/A".

* **Correct the Release Year:**
* **Remove Extra Spaces in String Fields:**
* **Save the Cleaned Data:**

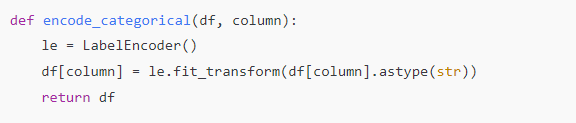
****

* **A screenshot of a computer

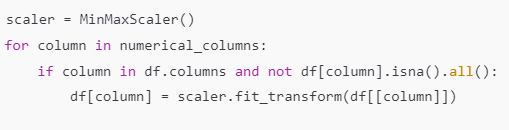
  Description automatically generatedPreview the Cleaned Data:**

**Processing the imbd table:**

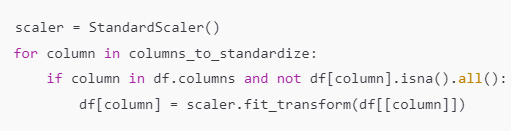
* **Encode Categorical Variables Using LabelEncoder**

****

* **Normalize Numerical Variables (Min-Max Scaling)**



* **Standardize Numerical Variables (Z-score)**

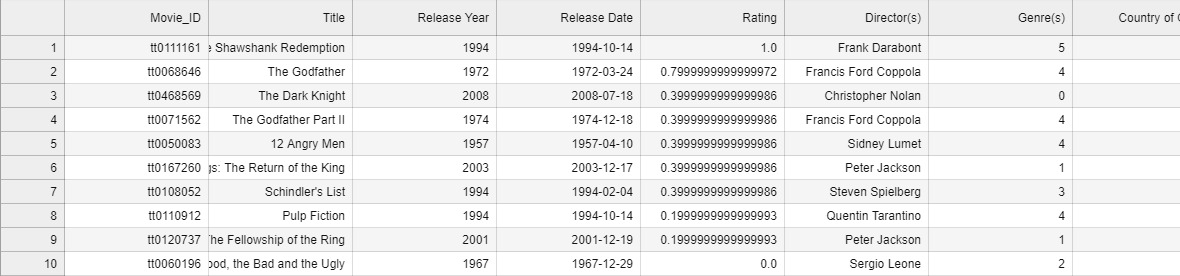


* **Handle Missing Values**
* **Save Processed Data**

****

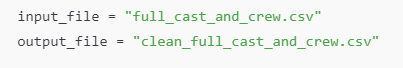
**Output:**

**A screenshot of a computer

Description automatically generated**

**clean crew and cast :**

**1: Importing Libraries and Loading the Data**

****

**2: Dropping Duplicate Rows**

**A close-up of a computer screen

Description automatically generated**

**3: Handling Missing Values**

**A screenshot of a computer

Description automatically generated**

**4: Standardizing Column Names**

**A screen shot of a computer program

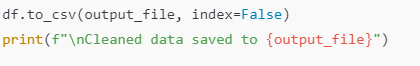
Description automatically generated**

**5: Handling Inconsistent Data**

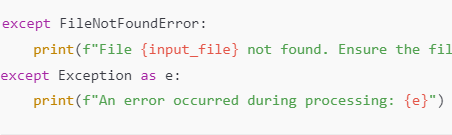
**A screen shot of a computer code

Description automatically generated**

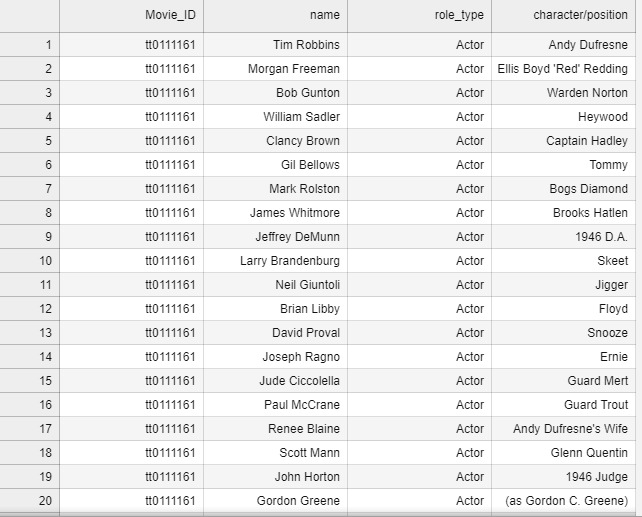
**6: Saving the Cleaned Data**

****

**7: Error Handling**

****

**Output:**

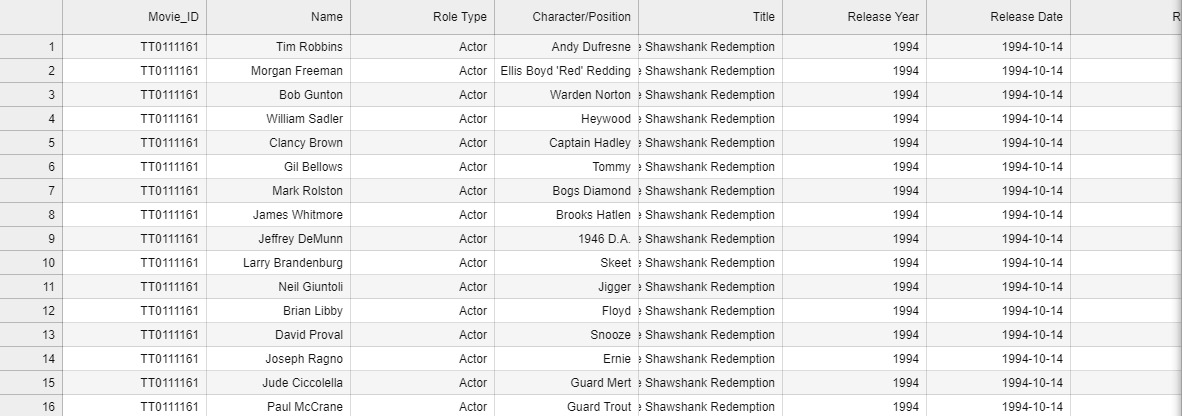
****

**Integration cod :**

****

**Output:**

A screenshot of a computer

Description automatically generated****

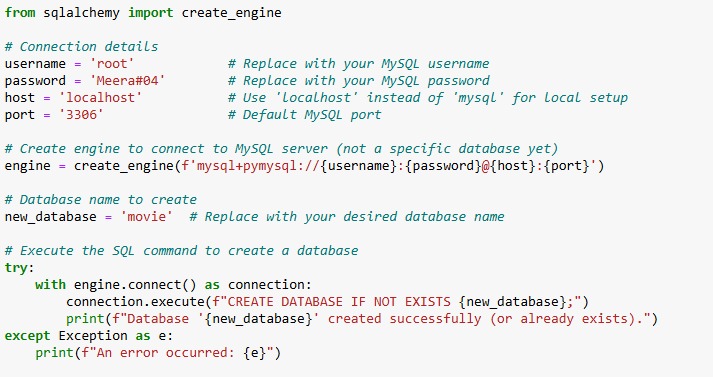
Task3: Loading Data and Executing Queries in MySQL Database Server

**Load:**  
After extraction and transformation, the processed data is loaded into the target data warehouse or database, making it ready for analysis or querying.



Output:



****

Output:

****

A screen shot of a computer program

Description automatically generated

