Automated ETL Pipeline for Real-Time Box Office & Movie Insights

📋 Problem Statement:

In today's dynamic entertainment industry, **real-time movie performance tracking** is crucial for production houses, distributors, and cinema owners to make data-driven decisions. However, **manually collecting and analyzing data** from multiple platforms like **Box Office Mojo** (for box office performance) and **Rotten Tomatoes** (for audience & critic reviews) is time-consuming, inconsistent, and prone to errors.

The **lack of an integrated system** to fetch, clean, and analyze movie data hinders timely insights into box office trends, audience ratings, and movie popularity.

🚀 Objective:

Develop an **automated ETL (Extract, Transform, Load) pipeline** that:

1. **Extracts** the latest movie titles from:
   - **Box Office Mojo** (for top box office performers): https://www.boxofficemojo.com/weekend/
   - **Rotten Tomatoes** (for critically acclaimed movies): https://www.rottentomatoes.com/browse/movies_in_theaters
2. **Fetches detailed movie information** (ratings, release date, box office, etc.) using the **OMDb API**:
   - **OMDb API:** http://www.omdbapi.com/
3. **Transforms** the data by:
   - Cleaning movie titles.
   - Standardizing date formats and genres.
   - Normalizing ratings for comparison.
4. **Loads** the final cleaned data into:
   - A **CSV file** for easy data manipulation.
   - A **PDF report** for business presentation purposes.

⚡ Key Features:

- 🌐 **Web Scraping:** Extract movie titles dynamically from Box Office Mojo & Rotten Tomatoes.
- 📊 **Real-Time Data Integration:** Fetch comprehensive movie details via OMDb API.
- 🔄 **Data Cleaning:** Remove inconsistencies, handle missing values, and normalize formats.
- 💾 **Data Export:** Store results in both CSV and PDF formats for flexible reporting.
- 🔔 **Automated ETL Pipeline:** A single script to automate the entire process.

📊 Sample Workflow Diagram:

🔺 Box Office Mojo

(https://www.boxofficemojo.com/weekend/)

⬇

🔺 Rotten Tomatoes

(https://www.rottentomatoes.com/browse/movies_in_theaters)

⬇

[Extract Movie Titles]

⬇

🔺 OMDb API

(http://www.omdbapi.com/)

⬇

[Transform & Clean Data]

⬇

📋 CSV File | 📄 PDF Report

📈 Real-World Applications:

1. **Cinema Chains:** Optimize movie screening schedules based on real-time trends.
2. **Production Houses:** Monitor box office performance for competitive analysis.
3. **Streaming Services:** Identify trending movies for acquisition decisions.
4. **Data Analysts:** Build dashboards using the cleaned data for predictive analysis.

✅ Deliverables:

1. **Python Script** with:
    - Web scraping modules for Box Office Mojo & Rotten Tomatoes.
    - ETL pipeline with data transformation logic.
    - Integration with OMDb API for movie metadata.
2. **CSV Dataset** with cleaned movie insights.
3. **PDF Report** summarizing key movie details.

🚀 Stretch Goals (Optional Enhancements):

- 📊 **Data Visualization:** Generate charts using Matplotlib for visual trends.
- ⏱️ **Scheduler Integration:** Automate daily/weekly data extraction.

TRANSFORMATION OF ETL PIPELINE

🎞️ Movie Data Transformation Instructions

This document outlines the step-by-step transformation tasks for each movie attribute in the ETL pipeline. The goal is to clean, standardize, and enrich the data to ensure it is ready for analysis.

🎥 1. Title Transformation

- **Remove Special Characters:** Use regular expressions to eliminate non-alphanumeric characters.
- **Standardized Case:** Convert all titles to Title Case for consistency.
- **Trim Whitespaces:** Remove leading/trailing spaces to maintain uniformity.

Example:

- Original: "The Lord of the Rings: The Return of the King!"
- Transformed: "The Lord Of The Rings The Return Of The King"

📅 2. Release Date Transformation

- **Date Formatting:** Replace spaces with hyphens to standardize the date format (e.g., DD MMM YYYY to DD-MMM-YYYY).
- **Convert to Date Object:** Use date parsing to convert text dates into proper date formats.
- **Handle Missing Dates:** Replace missing dates with "Unknown."

Example:

- Original: "25 Dec 2021"
- Transformed: "2021-12-25"

🎭 3. Genre Transformation

- **Convert to Lowercase:** Ensure all genres are in lowercase for consistency.
- **Split Genres:** If multiple genres are present, separate them into a list.
- **Remove Duplicates:** Ensure unique genre entries.

Example:

- Original: "Action, Adventure, Fantasy"

- Transformed: ['action', 'adventure', 'fantasy']

⭐ 4. IMDb Rating Transformation

- **Convert to Numeric:** Change rating from text to a floating-point number.
- **Round Off:** Round ratings to one decimal place.
- **Normalize:** Optionally, normalize ratings on a scale of 0 to 1.

Example:

- Original: "8.789"
- Transformed: "8.8"

🎬 5. Actors Transformation

- **Limit to Top 3:** Display only the top three actors.
- **Trim Spaces:** Remove extra spaces around names.
- **Sort Alphabetically (Optional):** For consistency in display.

Example:

- Original: "Tom Hanks, Robin Wright, Gary Sinise, Mykelti Williamson"
- Transformed: "Tom Hanks, Robin Wright, Gary Sinise"

💰 6. Box Office Transformation

- **Remove Currency Symbols:** Eliminate $, ,, and other non-numeric characters.
- **Convert to Numeric:** Store as an integer for analysis.
- **Handle Missing Data:** Replace missing values with 0 or N/A.

Example:

- Original: "$1,200,000"
- Transformed: 1200000

🏆 7. Awards Transformation

- **Extract Numbers:** Identify and sum all numeric values related to awards won.
- **Standardize Format:** Display total awards won.
- **Handle Missing Awards:** Set to 0 if no data is available.

Example:

- Original: "Won 3 Oscars. Another 5 wins & 10 nominations."
- Transformed: 18 (3 + 5 + 10)

🌐 8. Metascore Transformation

- **Convert to Integer:** Change metascore to an integer for calculations.
- **Normalize:** Convert to a 0-1 scale by dividing by 100.
- **Handle Missing Values:** Replace "N/A" with None.

Example:

- Original: "85"
- Transformed: 0.85

🌎 9. Language Transformation

- **Convert to Lowercase:** Ensure all language names are in lowercase.
- **Standardize Codes:** Optionally convert to ISO language codes.
- **Handle Missing Data:** Replace missing languages with "Unknown."

Example:

- Original: "English, Spanish"
- Transformed: "english, spanish"

🎥 10. Production Transformation

- **Remove Special Characters:** Clean the production company names.
- **Standardize Names:** Correct common misspellings and standardize abbreviations.
- **Handle Missing Data:** Replace missing production companies with "Independent."

Example:

- Original: "Warner Bros. Pictures!"
- Transformed: "Warner Bros Pictures"

These transformation steps will ensure clean, consistent, and analysis-ready data throughout your ETL pipeline. 🌟

Grading Criteria:

Extract - 10 marks

Transform - 30 marks

Load - 10 marks

Submission Instructions:

To submit your assignment, please follow these guidelines:

- Ensure that your assignment is fully completed.
- Push your code/assignment to a GitHub repository.
- Share the repository link by including it in a text, Word, or PDF file format.

Submit the file/text containing the repository link via Vlearn.