Designing an ETL Pipeline for IMDB Movie Data Analysis

Stephen Nwoye

In this project we are tasked with Extracting, Transforming, Loading, and Analyzing IMDB Movie datasets. Our datasets, sourced from GitHub.com, focus on various movie franchise brands within the IMDB.

In this project, we will follow the following steps: re-model the metadata of our dataset, come up with a business question to answer, discover and gather possible missing datasets to solve the business question, and then extract, transform, load, and analyze the data to address the business question.

Re-modelling the Metadata

This table explains the different tables and their individual column descriptions. It also suggests names for some tables to adopt a unified name across the database for columns with the same values.

TABLE NAME	COLUMNS	DESCRIPTION	CHANGED COLUMN NAME
Domestic box office franchise	Franchise	This is a collection of Movies	
	Number of movies	This is the numbers of movies associated with a particular franchise.	
	Domestic Box Office	This is the revenue earned from showing in theaters located within a particular country; in this case the USA.	
	Infl adj domestic box office	This shows the inflation adjustment domestic box office value, which is the equivalent of revenue, given the present inflation	
	Worldwide_Box_Office	This is the revenue earned from showing in theaters located around the world.	
	First_Year	This is the year, that the first movie in a particular franchise was released.	
	Last_Year	This is the year; that the last movie in the franchise is expected to be released or was released.	
	No_of_Years	This is the numbers of years from the first movie release to the last movie release in a particular franchise.	
Marvel Cinematic Universal Franchise domestic box office	Release_Date	This is the date for the release of a particular movie in the Marvel Comics franchise.	
	Title	This is the name of the movie	
	Production_Budget	This is the budget for the movie production	
	Opening_Weekend	This is the amount of money the movie earned from ticket sales during the first weekend of its release in theaters.	
	Domestic_Box_Office	This is the revenue earned from showing the movie in theaters located within the USA.	
	Worldwide_Box_Office	This is the revenue the movie earned from showing in theaters located around the world.	
Brands (US & Canada)	Brand	This refers to the recognizable production company associated with a particular movie or franchise.	
	Total_Gross	This is the total revenue generated by the brand.	
	Releases	This is the number of movies that the brand has released	**No_of_releases**
	No_1_Release	This is the top movie released by the brand.	
	Lifetime_Gross	This refers to the total amount of money the No_1_release has earned at the box office over its entire theatrical run.	
Brand_ Marvel Comics	Rank	This is the order in which a movie performed revenue-wise in comparison to other movies.	
	Release	This is the **title** of the movie.	**Title**
	Lifetime_Gross	This refers to the total amount of money a movie has earned at the box office over its entire theatrical run.	
	Max_Theaters	This refers to the highest number of theaters in which a movie was screened during its theatrical release period.	
	Opening_Gross	This refers to the total box office revenue a movie earns during its opening weekend in theaters.	
	Open_Theaters	This refers to the number of theaters in which a movie was simultaneously screened during its first week release period.	
	Release_Date	This is the date in which a movie was released	
	Distributor	This is a company responsible for the marketing, promotion, and distribution of the movie to theaters, among other channels.	
Top 20 Movies franchise	franchise	This is the name of a collection that has intellectual properties associated with it.	
	Rank	This is the order in which a movie performed revenue-wise in comparison to other movies.	
	Release	This is the **title** of the movie.	**Title**
	Lifetime_gross	This refers to the total amount of money a movie has earned at the box office over its entire theatrical run.	
	Max_theater	This refers to the highest number of theaters in which a movie was screened during its theatrical release period.	
	Open_theater	This refers to the number of theaters in which a movie was simultaneously screened during its first week release period.	
	Opening_gross	This refers to the total box office revenue a movie earns during its opening weekend in theaters.	
	Release_date	This is the date in which a movie was released	
	Distributor	This is a company responsible for the marketing, promotion, and distribution of the movie to theaters, among other channels.	
Franchises (US & Canada)	Franchise	This is the name of a collection that has intellectual properties associated with it.	
	Total_Revenue	This is the total revenue generated by the franchise.	

	Top_Release	This is the top movie released by the brand.	
	Lifetime_gross	This refers to the total revenue generated by a movie during its theatrical run.	
Movie|_Tags	User_Id	No Information on this.	
	movie_Id	This is the identification number of each movie	
	tags	This is the likened to the metadata associated to a particular movie	
	timestamp	This is the duration of the movie	
World Wide Box Office All Time Top 1000	Rank	This is the order in which a movie performed revenue-wise in comparison to other movies.	
	Title	This is the name of the movie	
	Worldwide_lifetime_gross	This refers to the total revenue generated by a movie during its theatrical run, in the whole world.	
	Domestic_lifetime_gross	This refers to the total revenue generated by a movie during its theatrical run, in the USA	
	Domestic%	The domestic percentage represents the proportion of total revenue attributed to the domestic market, in comparison to the overall revenue.	
	foreign_lifetime_gross	This refers to the total revenue generated by a movie during its theatrical run, in other foreign countries.	
	foreign%	The foreign percentage represents the proportion of total revenue attributed to the foreign market, in comparison to the overall revenue.	
	Year	This is the year that the particular movie was released.	
Movie_Id	movie_Id	This is the identification number of each movie	
	title	This is the name of the movie	
	genres	This is the movie category based on their common themes, styles, or subject matter	
Domestic Box Office Daily - The Avengers	Date	This is the date that the movie was shown in the theater.	
	Rank	No Information	
	Gross	This is the total amount of revenue generated for a particular day	**daily_Gross**
	%YD	This is the percentage change in revenue generated for a particular day compared to the revenue generated in the previous day.	**1D%_change**
	%LW	This is the percentage change in revenue generated over a 7-days period calculated on the eighth day.	**7D%Change**
	Theaters	This is the number of theaters were the movie was shown on a particular day	
	Per_theater	This is the average revenue generated per theater	**Revenue_per_theater**
	Total gross	This is the accumulated revenue generated since from the first day	
	Days	This is the progressive number days the theater has been showing the movie.	