ETL Pipeline For Movie Analysis

James Poh Hao[1], Saw Wei Han[2], Shan Yi[3], Low Mei Lin[4], Ke Jiayi[5]

1. National University Of Singapore (e0726810@u.nus.edu)
2. National University Of Singapore (e0684511@u.nus.edu)
3. National University Of Singapore (shanyi@u.nus.edu)
4. National University Of Singapore (e0775506@u.nus.edu)
5. National University Of Singapore (e0775753@u.nus.edu)

This document outlines our analysis and ETL pipeline creation for Movie Analysis for the company Netflix. The project aims to develop an automated pipeline that collects, processes, and stores movie data and metadata from multiple sources for movie trend analysis, and popularity prediction.
Project Url : https://github.com/jpoh1999/IS3107

Keywords: Movies, Netflix, ETL, Data Engineering, Machine Learning

Edited by:
James Poh Hao

Reviewed by:
Saw Wei Han, Low Mei Lin

Copy-edited by:
Shan Yi, Ke Jiayi

Received: 2024-04-24
Published: 2024-04-26

1 Problem Statement

Movie popularity is an important factor in measuring the success of a movie. As film production is costly, it is crucial to create a movie that appeals to the majority of the audience so that the film production company can maximize their profits. Therefore, the film industry should not only innovate in terms of content and technology but also create movies that best cater to the increasingly picky audience's taste. Consequently, we would like to analyze and understand patterns in the movie industry over time. This analysis will empower industry stakeholders to make informed decisions on future productions, marketing strategies, and content curation. In addition to the film industry, it can also be applied to other TV series or online video platforms for greater impact. Furthermore, audiences will also gain satisfaction and experience provoked sentiments through engaging with dynamic film entertainment.

2 Data Overview

Netflix, one of the most beneficial stakeholders in the movie industry for our project, is both a filmmaker and a film streaming service provider. To better facilitate our end users, we will be using Netflix data as our primary source of data. Given the limited data available, we will also be gathering data from the free TMDB source dataset to enhance and enrich the Netflix data.

2.1 Netflix Dataset

This dataset was sourced from Kaggle Netflix Movies and TV Shows. It comprises a comprehensive list of movies and TV shows available on Netflix and is updated regularly. It contains 8,807 entries as of the date this report was written. This will facilitate long-term movie data insight generation. The fields provided can assist in performing analyses on directors, cast actors, countries, and duration trends over time. Refer to **Table 6.1** for a sample of our dataset.

2.2 TMDB Dataset

This Full TMDB Movies Dataset 2024 (1M Movies) from Kaggle consists of data from The Movie Database (TMDB) covering over 930k movies up to the year 2024. It is valuable for its extensive coverage of movies, including budgets, revenues, votes, popularity, and more. This dataset adds more quantitative information, enabling us to do deeper analysis on the movie industry's trends, financial insights, and production patterns.Refer to **Table 6.2** for a sample of our dataset.

3 Methodology

3.1 Data Preparation and Cleaning

3.1.1 Basic Cleaning and Standardization during Data Transfer to the Data Warehouse

This initial phase involves foundational data preparation tasks to ensure the integrity and consistency of the data as it moves from the data lake to the data warehouse. Key activities in this phase include:

- **Handling Missing Values**
  Identification and treatment of missing or null values within the dataset to mitigate potential data quality issues and ensure completeness.
- **Standardizing Formats**
  Conversion of various data formats (such as date-time formats) into a consistent standard format across the dataset. This helps to streamline data processing and analysis.
- **Ensuring Basic Data Quality Checks**
  Implementation of basic data quality checks to identify and rectify any anomalies or inconsistencies in the data. This may involve validating data against predefined criteria or business rules.
- **Data Profiling**
  Conducting data profiling to gain insights into the structure, content, and quality of the dataset. This aids in identifying potential data issues and guiding subsequent cleaning and transformation efforts.

3.1.2 Specific Transformations for Data Marts

Once the data is loaded into the data warehouse, it undergoes more specialized transformations tailored to the requirements of individual data marts. This phase involves applying advanced cleaning and transformation techniques to prepare the data for specific analytical or business needs. Key aspects of this phase include:

- **Detailed Cleaning**
  Performing more granular data cleaning processes targeted at addressing specific data anomalies or inconsistencies relevant to the analytical context of each data mart. This may include outlier detection and removal, data imputation techniques, and error correction mechanisms.
- **Specialized Transformations**
  Implementing transformations that are customized to the business logic or analytical requirements of each data mart. This could involve aggregations, calculations, derivations, and enrichments to derive insights and support decision-making processes.

- **Dimensional Modeling**
  Designing and implementing dimensional models, such as star or snowflake schemas, to optimize data retrieval and analysis within the context of each data mart's dimensional structure.
- **Data Partitioning and Indexing**
  Partitioning large datasets and creating appropriate indexes to enhance query performance and optimize data retrieval for analytical queries.

3.2 Database Design and Schema

After analyzing the business requirements, we came up with the following database design and schema below which we think will benefits the respective departments with their operational needs.

3.2.1 Data Warehouse

This database design and schema at **Fig 6.1** for the Datawarehouse provide a structured framework for data engineers to efficiently manage and maintain the movie data. By organizing data into tables such as movies_ratings, movies_casts, and movies_finance, the data engineers can ensure data integrity and consistency. This structured approach streamlines data ingestion, storage, and retrieval processes, enabling data engineers to perform tasks such as data cleaning, transformation, and loading with ease. The separation of into the following schema allows data engineers to easily integrate new data from future sources into the datawarehouse.

Furthermore, the database schema facilitates scalability and flexibility, allowing data engineers to accommodate evolving data requirements and handle increasing volumes of movie data over time. Features such as normalization and indexing optimize database performance, ensuring fast query execution and data access for downstream applications.

3.2.2 Dashboard Mart

The database schema at **Fig 6.2** for our dashboard mart were chosen specifically to enable the operations department to conduct comprehensive dashboard analysis on the movie data. By normalizing the tables for movie genres, directors, actors, and countries, the schema provides a rich dataset for generating insights and visualizations. Through dashboard analytics, the operations department can also track key performance indicators (KPIs) such as revenue, audience demographics, and movie popularity trends.

This schema's relational structure facilitates complex queries and joins, allowing the operations department to explore correlations between different movie attributes and identify patterns and anomalies. Additionally, features such as data mart creation and aggregation functions enable the aggregation of data at various levels of granularity, empowering the operations department to derive actionable insights for strategic decision-making.

3.2.3 Machine Learning Mart

For the machine learning department, the database design and schema at **Fig 6.16** serve as the foundation for predicting the profitability of movies. By denormalizing the three movie data stored in the datawarehouse into a single table, the machine learning department can easily perform data

cleaning and preprocessing into a data type more suited for machine learning training.

The schema's well-defined relationships and data types facilitate seamless feature engineering, data scaling and normalization, enabling the extraction of meaningful features from the movie dataset and facilitates easy pre-processing to train machine learning algorithms. Features such as data partitioning and sampling also support model training and evaluation.

Training data generated from each machine learning pipeline processes are also stored in the local device as parquet files. The parquet format was carefully chosen after noticing the increased in the dimensionality of the data from doing one-hot encoding, resulting in **10x** more column features than rows.

3.3 Pipeline Design and Architecture

Airflow emerges as a formidable automation orchestrator renowned for its multifaceted advantages in optimizing business operations. It distinguishes itself through its robust capabilities, offering a comprehensive solution for orchestrating intricate workflows.

Airflow empowers organizations to effortlessly design, schedule, and oversee workflows, presenting a centralized platform for streamlined management. Through its user-friendly interface, users can delineate workflows as Directed Acyclic Graphs (DAGs), facilitating clear visualization of dependencies and task sequences within the pipeline. This graphical representation simplifies comprehension of data flow and process intricacies, enhancing overall operational efficiency and facilitating seamless troubleshooting. For our project, we will use a structure which mirrors the business world as closely as possible to integrate what we learn with practical examples.

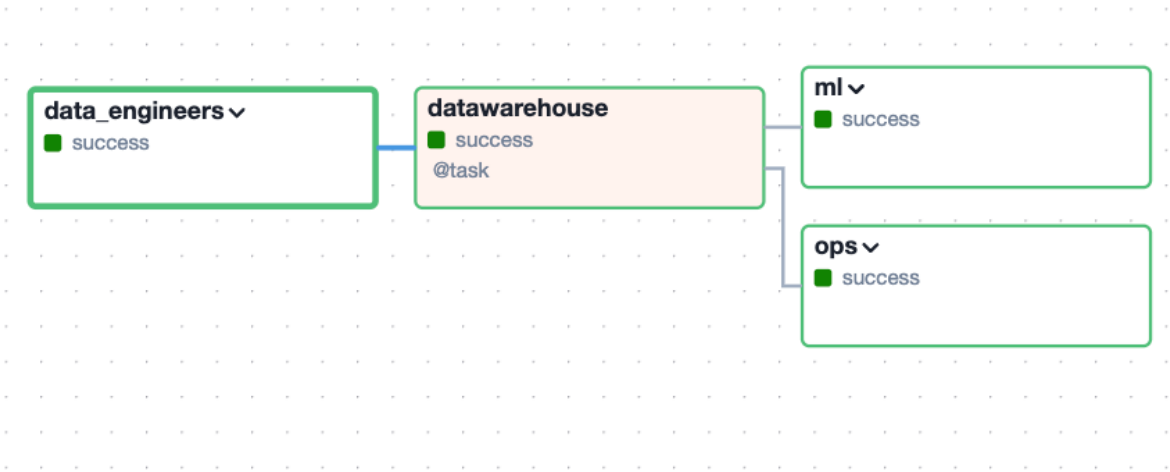

Figure 3.1: Airflow Business Structure

3.3.1 Data Engineering

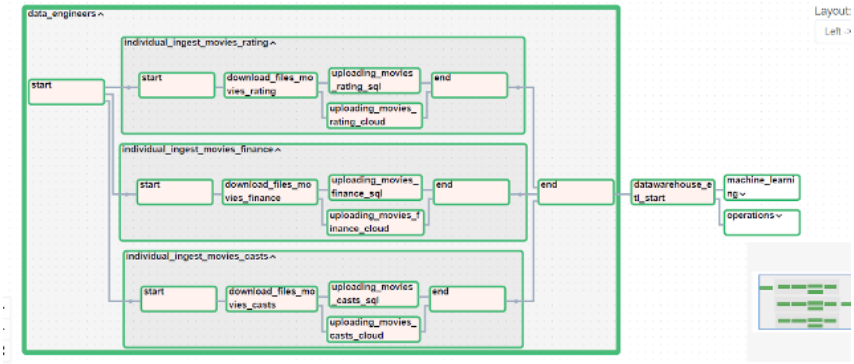

Figure 3.2: Data Engineer Workflow

In our data engineering process, we have strategically opted for MySQL as the foundation for both our data warehouse and data marts. Leveraging its reliability and scalability, MySQL serves as an optimal platform for storing and managing our structured data. Concurrently, we utilize Google Cloud Storage as our designated data lake, providing a secure and flexible solution for backup and archival purposes.

The initial phase, 'ingestion', encompasses the downloading of relevant datasets and sending them to the respective datalake and datawarehouses. Simple data transformations was also done using SQL queries during the initial ingestion to ensure that the datawarehouse have the necessary schemas to facilitate the other data marts. The ingestion procedure to the datawarehouse was also done parallelly to optimize performance.

In essence, our approach integrates the robust capabilities of MySQL and Google Cloud Storage to orchestrate a cohesive and efficient ETL process, empowering data engineers to architect and maintain a consistent and synchronized data infrastructure conducive for insightful analytics and informed decision-making by the other departments.

3.3.2 Operations

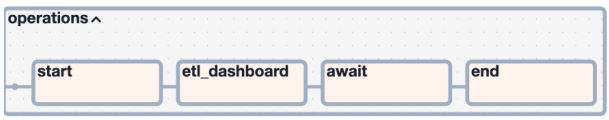

Figure 3.3: Operations Workflow

In our operational workflow, the department retrieves data from the data warehouse, performs necessary transformations to conform to the required schema, and subsequently loads the processed data into their designated databases within MySQL. The data stored in the operations database serves as the foundation for analytical works conducted by the team using Tableau. This structured approach enables the team to generate insightful analyses that cater to client needs, fostering informed decision-making and enhancing overall service delivery. The application of our operation process can be found in **Section 3.4.1**

3.3.3 Machine Learning

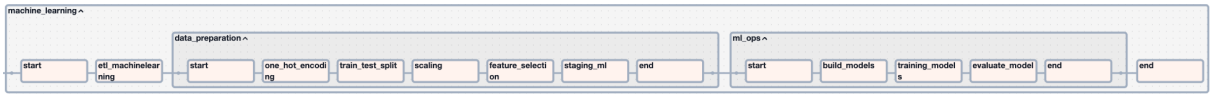

Figure 3.4: Machine Learning Workflow

In our machine learning workflow, the department is first integrates the three datasets sourced from the data warehouse into a denormalized format optimized for streamlined machine learning processes. This integration process aims to consolidate disparate data sources, facilitating easier access and manipulation for subsequent analytical tasks.

Following data retrieval, the department has also devised a meticulous pipeline for data preparation, encompassing comprehensive cleaning and preprocessing steps. This preparatory phase is essential for ensuring the quality and integrity of the data prior to its utilization in machine learning training.

Subsequently, the preprocessed data is directed into the machine learning pipeline, where it undergoes rigorous training procedures tailored to the specific objectives of the project. This phase involves the application of diverse machine learning algorithms and techniques, with a focus on optimizing model performance and accuracy.

To maintain transparency, reproducibility, and scalability throughout the workflow, all the predicted results and artifacts generated during the training process are meticulously logged into MLFlow—a dedicated platform for managing the machine learning lifecycle. This logging mechanism serves to establish a comprehensive record of all experiments, enabling seamless tracking of model iterations, parameter configurations, and performance metrics.

By adhering to this meticulously orchestrated workflow—from initial data preparation to the deployment of models in production—our department ensures the consistency, reliability, and quality of our machine learning operations, ultimately fostering the delivery of actionable insights and value-added solutions. More information on the application of the ML process can be found in **Section 3.4.2**.

3.4 Downstream Applications

3.4.1 Trends Analysis Visualizations with Tableau

We have explored a positive correlation between budget and revenue. Higher-budget films tend to generate higher revenue, and it is a common expectation in the industry due to the production quality, marketing effort and often star-power associated with bigger budgets. However, there could be exceptions to the rule, such as “Joker”, which appears to have generated high revenue on a smaller budget compared to some other films with similar earnings. This suggests that a strong story, higher rating has a strong correlation with revenue independent of the budget. In contrast, underperformers like “John Carter” and “The Flash” show large budgets with lower revenue, indicating that a high investment does not always guarantee success. Overall, the association between revenue and budget will still hold for the

majority of the time.

Another insight we derived is that high revenue is correlated with show rating, but show rating is not correlated with high budgets. Show rating in this case means the popularity from TMDB. Movies like "Avengers: Infinity War" and "Black Panther" are notable for their high revenue and relatively larger bubble, which suggests a high show rating. However, show ratings are not correlated with high budgets. Low budget movies could outperform high budget movies in show rating.

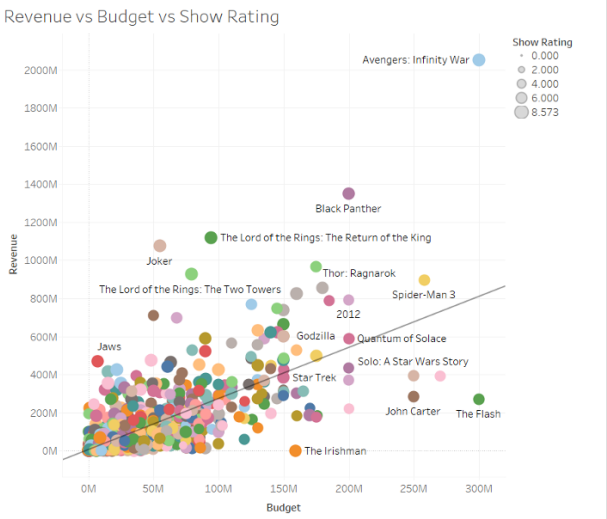

Figure 3.5: Eda Revenue_Budget

Films released in high-attendance seasons like summer breaks and holidays may perform better as it maximizes audience availability and willingness to spend on entertainment, further boosting the profitability. For example, even the number of movies published for August to December are similar, but December significantly has gained much more revenue than the rest due to Christmas.

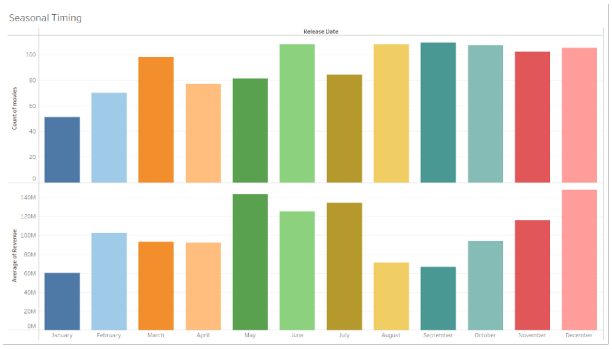

Figure 3.6: Seasonal EDA

Genres "Drama", followed by "Comedy", "Thriller" and "Action" have the most movies produced over the years, even during periods where overall movie production is decreasing. This seems to suggest that these genres are likely stable in popularity or demand among audiences. Movies under these genres generally appeal to a broad audience, and their consistent production also indicates that filmmakers

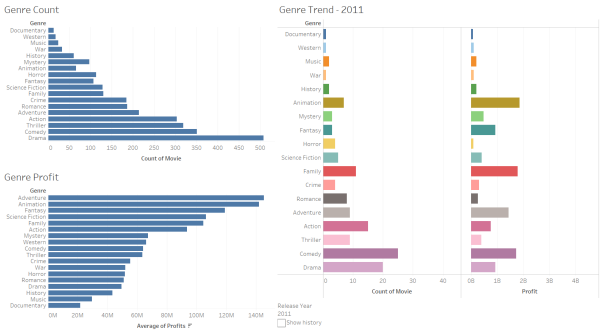

Figure 3.7: Trends in Genre

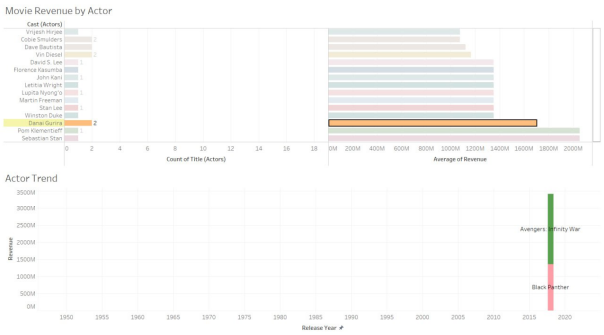

Figure 3.8: Actor Analysis

believe there is a reliable market for these types of movies. These could be considered as a "safe" choice for the industry leaders.

Genres "Adventure", "Animation" and "Fantasy" yield the highest profits but are not produced as frequently as other genres such as "Drama", "Comedy" and "Thriller". This could be due to their higher production costs from the need for extensive visual effects, elaborate sets. Especially in the case of animation, it requires intensive labor and technology requirements. Because of the significant investment involved, companies might be less willing to frequently invest in these types of films compared to more consistently profitable genres with lower overheads, despite the high profit potential. However, when these films do succeed, they often generate substantial profits, possibly due to their broad, global appeal. Therefore, these could be considered as an "opportunity" choice for the industry leaders to earn money.

We examined the frequency of actors' film appearances, the associated box office revenues. This chat has contextualizes the economic significance of these actors, by presenting the average revenue yielded by films in which they starred. Additionally, the chart also traced the volume of work actors have secured across different periods, which could be indicative of their career growth, industry demand, and market shifts. This analysis aims to identify the potential profitability of actors in the industry. Insights gathered here are crucial for industry stakeholders to enhance strategic decision-making in production and casting processes.

3.4.2 Machine Learning Application

Our machine learning department aims to forecast the return on investment (ROI) of movies, a pivotal metric gauging the profitability of a movie. ROI is computed as the ratio of a movie's profits to its budget, with a threshold of 10 delineating positive (True) and negative (False) outcomes. For clarity, movies with an ROI equal to or exceeding 10 are classified as profitable (positive outcome), while those falling below are deemed not profitable (negative outcome). Therefore, the ROI target serves as a binary label for our supervised machine-learning endeavours.

In the preprocessing phase, we curated the feature set used for training the machine learning models. Employing one-hot encoding, categorical data underwent transformation for compatibility with numerical data from the original dataset. Concurrently, redundant columns, such as the description column, were identified and subsequently dropped to enhance model training efficiency.

Feature selection played a pivotal role in optimizing model performance. Leveraging the Random Forest Classifier, features were ranked based on importance scores, derived from the Random Forest model's tree-based approach. The top-ranking features were then selected to serve as inputs for subsequent modelling tasks, ensuring focus on the most influential predictors.

Our ensemble learning approach integrated various classification algorithms, including Support Vector Machine, K-nearest neighbour classifier, AdaBoost classifier, and Multi-Layer Perceptron classifier. Each model contributed unique perspectives and predictive capabilities to the ensemble. These diverse outputs were then aggregated and fed into a Logistic Regression model, which served as the final estimator. This comprehensive ensemble strategy enabled robust predictions of movie profitability, offering a binary classification outcome based on the predetermined ROI threshold.

To facilitate consistency and reliability of our ML operations, we also connected our ML workflow into MLFlow to track our model versions and prediction history. Refer to **Fig 6.17** and **Fig6.18** for the history and performance of our ML models.

4 Performance Analysis

4.1 Pipeline Performance

The Gantt charts provide valuable insights into the batch ETL performance of the data pipeline. The task duration seem to be very efficient, with tasks completed in relatively short intervals. For instance, the final task concluded within a minute indicates a relatively fast speed during the pipeline execution. Furthermore, the structure of the tasks effectively utilized the resources through parallelism. The strategic arrangement enables the concurrent execution of tasks whenever feasible, optimizes resource utilization and enhances the overall throughput. The green bars denote task completion and highlight the reliability of the pipeline. The absence of errors, as shown by the success of tasks, tells us that the pipeline's execution is robust and dependable. Hence, we can tell from the Gantt charts the effectiveness and efficiency of the pipeline operation. Refer to **Fig 6.19** and **Fig 6.20**

4.1.1 Speed

Our ETL pipeline is highly efficient, typically completing the entire data processing cycle within minutes. This rapid processing capability ensures that data analysis and subsequent actions are based on the most current information, facilitating timely decision-making.

4.1.2 Data Update and Completeness

The pipeline is designed to regularly update data, capturing the latest market and consumer trends to keep our analyses relevant. Furthermore, the system is robust, capable of incorporating previously missing data retroactively if such data is provided at a later date upstream. This feature significantly enhances the reliability and completeness of our data sets, ensuring that our models and analyses remain accurate over time.

4.1.3 Model Accuracy and Retraining

The accuracy of our machine learning models is a critical component of the pipeline. We maintain high standards of accuracy by implementing continuous retraining protocols. The models are periodically updated with the latest data, which allows them to adapt to new trends and changes in the environment. This ongoing retraining process not only sustains but often improves the predictive accuracy of our models over time.

4.1.4 Reliability and Future Enhancements

To further enhance the reliability of the model, we are exploring advanced data integration techniques and considering the incorporation of additional data sources that can provide more comprehensive insights. These improvements will aim to refine the predictions further and provide even faster data processing capabilities in future iterations of the pipeline.

4.2 Business Integration

The predictive analytics model could be integrated into the investment decision-making process for potential movie projects. It may assist investors in determining whether to invest in a movie and what returns might be expected. A user-friendly interface can be developed to input variables such as budget and casting. Producers can simulate different scenarios and see how these changes could affect the potential return on investment. Hence, this model empowers film producers with data-driven insights, reducing the risks associated with film financing. The dashboard could be useful for decisions related to movie copyright purchases for streaming services. Stream service providers can identify and select films that are likely to attract a large audience, thereby maximising viewership and subscription growth for streaming platforms.

4.3 Insights from Dashboard

Building on the discussion in Section 3.4.1, film producers like Netflix could base their decisions on statistical analysis. Moreover, they should monitor trends to capture shifts in audience preferences and adjust their films accordingly. For example, producers working with low budgets should prioritize crafting stories that attract audiences. Choosing the optimal release timing and focusing on genres that typically yield high returns can also increase their chances of success. Conversely, films with less

engaging plots could benefit from hiring well-known actors or investing in sophisticated visual effects. More strategic moves can be obtained by experimenting with dashboard data and qualitative analysis on case studies. Then, the decision can be verified with a prediction model.

5 Conclusion

In conclusion, our project revolutionizes the analysis of movie profitability, empowering filmmakers, including industry giants like Netflix, to confidently navigate the ever-evolving landscape of content creation. By leveraging cutting-edge machine learning techniques, we also offer actionable insights that drive strategic decision-making, ultimately shaping the future of entertainment with blockbuster success. Lights, camera, data-driven action!

6 Appendices

Field Name	Data Type	Description	Example Entry
show_id	String	Unique ID for every Movie/TV Show	s1, s2
type	String	Identifier - A Movie or TV Show	Movie, TV Show
title	String	Title of the Movie / TV Show	Dick Johnson Is Dead, Blood & Water
director	String	Director of the Movie	Kirsten Johnson, -
cast	String	Actors involved in the movie/show	Kirsten Johnson, Ama Qamata, Khosi Ngema, Gail Mabalane, etc.
country	String	Country where the movie/show was produced	United States, South Africa
date_added	Date	Date it was added on Netflix	September 25, 2021, September 24, 2021
release_year	Integer	Actual Release year of the movie/show	2020, 2021
rating	String	TV Rating of the movie/show	PG-13, TV-MA
duration	String	Total Duration - in minutes or number of seasons	90 min, 2 Seasons

Table 6.1: Netflix Dataset

Field Name	Data Type	Description	Example Entry
id	Integer	Unique identifier for each movie	1226636, 1247951

title	String	Title of the movie	Party Animals, Best Bros
vote_average	Float	Average vote or rating given by viewers	0.0, 6.0
vote_count	Integer	Total count of votes received for the movie	0
status	String	The status of the movie	In Production, Released
release_date	String	Date when the movie was released	2008-07-16, 2014-11-05
revenue	Integer	Total revenue generated by the movie	825532764, 1004558444
runtime	Integer	Duration of the movie in minutes	169, 152
adult	Boolean	Indicates if the movie is suitable only for adult audiences	False, True
backdrop_path	String	URL of the backdrop image for the movie	nMKdUUepRSBhyas.jpg
budget	Integer	Budget allocated for the movie	185000000, 237000000
homepage	String	Official homepage URL of the movie	https://www.warnerbros.com
imdb_id	String	IMDb ID of the movie	tt0468569
original_language	String	Original language in which the movie was produced	en, fr
original_title	String	Original title of the movie.	Interstellar, Inception
overview	String	Brief description or summary of the movie.	Harry, Ron and Hermione walk away from their last year at Hogwarts to find and …
popularity	Float	Popularity score of the movie.	83.952, 140.241
poster_path	String	URL of the movie poster image.	/oYuLEt3zVCKq57qu2F8dT7NIa6f.jpg
tagline	String	Catchphrase or memorable line associated with the movie.	Your mind is the scene of the crime.

genres	String	List of genres the movie belongs to.	Action, Science Fiction, Adventure
production_companies	String	List of production companies involved in the movie.	Legendary Pictures, Syncopy, Warner Bros. Pictures
production_countries	String	List of countries involved in the movie production.	United Kingdom, United States of America
spoken_languages	String	List of languages spoken in the movie.	English, French, Japanese, Swahili
keywords	String	Keywords associated with the movie.	rescue, mission, paris, virtual reality, philosophy, …

Table 6.2: TMDB Dataset

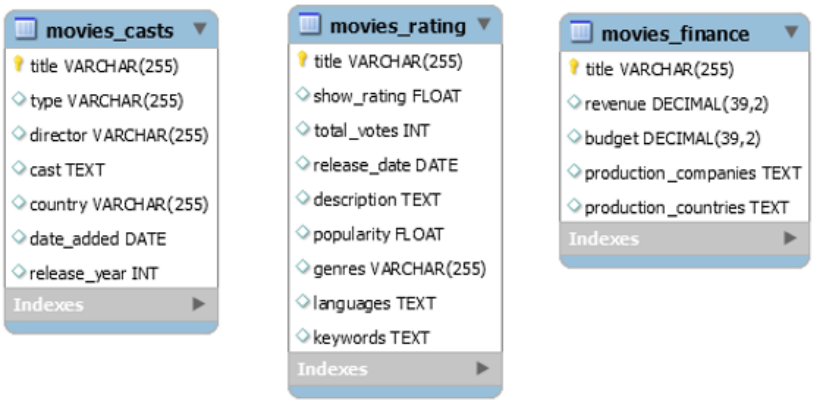

Figure 6.1: Datawarehouse Schema

title	type	director	cast	country	date_added	release_year	show_rating	release_date	genres
13 Sins	Movie	Daniel Stamm	Mark Webber, Rutina Wesley, Devon Graye, To…	United States	2019-01-13	2014	6.267	2014-04-11	Horror, Thriller
16 Blocks	Movie	Richard Donner	Bruce Willis, Mos Def, David Morse, Jenna Stern…	United States, Germany	2019-11-01	2006	6.424	2006-03-01	Action, Adventure, Crime, Thriller
17 Again	Movie	Burr Steers	Zac Efron, Leslie Mann, Matthew Perry, Thomas…	United States	2021-01-01	2009	6.287	2009-03-11	Comedy
20 Feet From Stardom	Movie	Morgan Neville	Darlene Love, Merry Clayton, Lisa Fischer, Táta…	United States	2018-09-22	2013	7.233	2013-06-14	Documentary, Music
2012	Movie	Roland Emmerich	John Cusack, Amanda Peet, Chiwetel Ejiofor, T…	United States	2021-04-01	2009	5.819	2009-10-10	Action, Adventure, Science Fiction
20th Century Women	Movie	Mike Mills	Annette Bening, Elle Fanning, Greta Gerwig, Lu…	United States	2019-06-28	2016	7.348	2016-12-28	Drama
21	Movie	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aa…	United States	2020-01-01	2008	6.7	2008-03-27	Drama, Crime
21 & Over	Movie	Jon Lucas, Scott Moore	Miles Teller, Skylar Astin, Justin Chon, Sarah Wr…	United States	2019-04-16	2013	5.815	2013-03-01	Comedy
23:59	Movie	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence …	Singapore	2018-12-20	2011	4.768	2011-11-03	Horror
28 Days	Movie	Betty Thomas	Sandra Bullock, Viggo Mortensen, Dominic West…	United States	2020-09-30	2000	6.12	2000-04-06	Comedy, Drama
3 Days to Kill	Movie	McG	Kevin Costner, Amber Heard, Hailee Steinfeld, …	United States, France, …	2020-12-01	2014	6.068	2014-02-14	Action, Drama, Thriller
3 Generations	Movie	Gaby Dellal	Elle Fanning, Naomi Watts, Susan Sarandon, Ta…	United States	2017-08-28	2015	6.391	2016-10-27	Comedy, Drama

Figure 6.3: Movie Table Under Data Mart

languages	keywords	revenue	budget	production_companies	production_countries	profit	roi
English	hotel, hostage, blackmail, detective, new orlea…	13809.00	5000000.00	Dimension Films, IM Global, Automatik Entertain…	United States of America	-4986191.00	-99.723820
English	new york city, mission of murder, witness, betr…	65664721.00	55000000.00	Alcon Entertainment, Cheyenne Enterprises, Eq…	Germany, United States of America	10664721.00	19.390402
English	high school, fake identity, bullying, high school …	136316880.00	20000000.00	New Line Cinema	United States of America	116316880.00	581.584400
English, Spanish		5892466.00	1000000.00	Tremolo Productions, Gil Friesen Productions	United States of America	4892466.00	489.246600
Tibetan, German, Russian, French, Mandarin, L…	race against time, maya civilization, civilization, …	791217826.00	200000000.00	Columbia Pictures, Centropolis Entertainment, F…	United States of America	591217826.00	295.608913
English	parent child relationship, 1970s, balcony, femini…	5664764.00	7000000.00	Annapurna Pictures, Archer Gray, Modern People	United States of America	-1335236.00	-19.074800
English	friendship, card game, casino, gambling, profes…	159808370.00	35000000.00	Columbia Pictures, Relativity Media	United States of America	124808370.00	356.595343
English	alcohol, birthday, friends, debauchery	48065672.00	13000000.00	Mandeville Films, Relativity Media, SkyLand Ent…	United States of America	35065672.00	269.735938
Mandarin, English, Malay	singapore, jungle, soldier, ghost, chinese ouija	1208479.00	600000.00	Goryiah Pictures, Clover Films, Grand Brilliance, …	Malaysia, Singapore	608479.00	101.413167
English	drug addiction, car crash, alcoholism, drug reha…	6219894.00	43000000.00	Columbia Pictures	United States of America	-36780106.00	-85.535130
English	daughter, cia, wife, killing, retirement, family, s…	53260230.00	28000000.00	Wonderland Sound and Vision, EuropaCorp, Rel…	France, United States of America	25260230.00	90.215107
English	single mother, gender roles, lgbt, woman direct…	680351.00	10000000.00	IM Global, Big Beach	United States of America	-9319649.00	-93.196490

Figure 6.4: Movie Table Under Data Mart

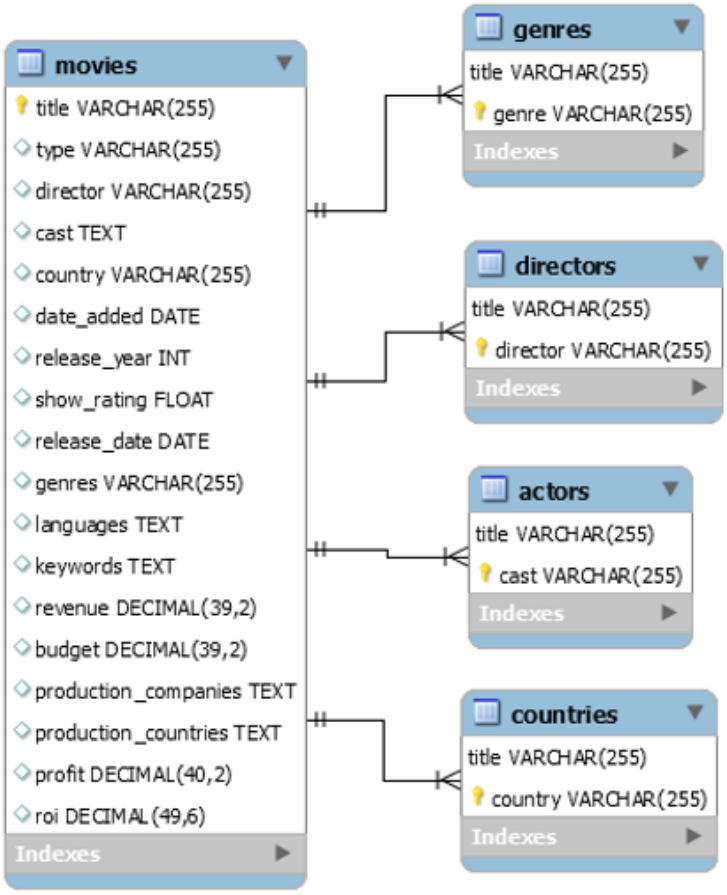

Figure 6.2: DashboardMart Schema

title	type	director	cast	country	date_added	release_year
¡Ay, mi madre!	Movie	Frank Ariza	Estefanía de los Santos, Secun de la Rosa, Tere...	Spain	2019-07-19	2019
'76	Movie	Izu Ojukwu	Ramsey Nouah, Rita Dominic, Chidi Mokeme, Ibi...	Nigeria	2021-08-04	2016
'89	Movie		Lee Dixon, Ian Wright, Paul Merson	United Kingdom	2018-05-16	2017
(T)ERROR	Movie	Lyric R. Cabral, David Felix Sutcliffe		United States	2016-06-30	2015
(Un)Well	TV Show			United States	2020-08-12	2020
#Alive	Movie	Cho Il	Yoo Ah-in, Park Shin-hye	South Korea	2020-09-08	2020
#AnneFrank - Parallel Stories	Movie	Sabina Fedeli, Anna Migotto	Helen Mirren, Gengher Gatti	Italy	2020-07-01	2019
#blackAF	TV Show		Kenya Barris, Rashida Jones, Iman Benson, Ge...	United States	2020-04-17	2020
#cats_the_mewvie	Movie	Michael Margolis		Canada	2020-02-05	2020
#FriendButMarried	Movie	Rako Prijanto	Adipati Dolken, Vanesha Prescilla, Rendi Jhon, B...	Indonesia	2020-05-21	2018
#FriendButMarried 2	Movie	Rako Prijanto	Adipati Dolken, Mawar de Jongh, Sari Nila, Vonn...	Indonesia	2020-06-28	2020
#realityhigh	Movie	Fernando Lebrija	Nesta Cooper, Kate Walsh, John Michael Higgin...	United States	2017-09-08	2017
#Roxy	Movie	Michael Kennedy	Jake Short, Sarah Fisher, Booboo Stewart, Dan...	Canada	2019-04-10	2018

Figure 6.5: Movie Cast Table Under Data Warehouse

title	revenue	budget	production_companies	production_countries
	0.00	0.00		
Dazzling Shadows	0.00	0.00		
dining room egg	0.00	0.00		
Forever Bro	0.00	0.00		
Give Chains A Chance	0.00	0.00		
Illegal Tourist	0.00	0.00		
Memoria Data	0.00	0.00		
O kyrios pou xairetouse mihanika to parelthon tou	0.00	0.00		
The Day People Left	0.00	0.00		
Without A Soul	0.00	0.00		
The Sky is my Ceiling	0.00	0.00		
Од ба Гэгээ	0.00	0.00	Really Nice Content, ...	Mongolia, United Kin...
Хэвтрийн хүн	0.00	0.00	Nomadia Pictures, Gu...	Mongolia

Figure 6.6: Movie Finance Table Under Data Warehouse

title	show_rating	total_votes	release_date	description	popularity	genres	languages	keywords
	0	0	0000-00-00		0.6	Documentary		
Dazzling Shadows	0	0	0000-00-00	Seeing through the camera's lens when form di...	0			
dining room egg	0	0	0000-00-00		0			
Forever Bro	0	0	0000-00-00		0			
Give Chains A Chance	0	0	0000-00-00	'Give Chains A Chance' is a performance center...	0			
Illegal Tourist	0	0	0000-00-00	Soran is a Kurdish refugee from Iran. Fulvia is a...	0			
Memoria Data	0	0	0000-00-00	Memoria Data collects the moments of connecti...	0			
O kyrios pou xairetouse mihanika to parelthon tou	0	0	0000-00-00		0			
The Day People Left	0	0	0000-00-00	When life in cities began to change due to Covi...	0			
Without A Soul	0	0	0000-00-00	Without A Soul is an experimental visual poem c...	0			
The Sky is my Ceiling	0	0	0000-00-00	The story by JG Ballard « The Concentration Cit...	0			
Од ба Гэгээ	6.5	2	2019-03-29	Any number of superlatives describe Mongolia's ...	0.708	Comedy, Dr...	Mongolian	
Хэвтрийн хүн	0	0	2020-10-20	A multi-generational story of love and loss cent...	0.607	Drama	Mongolian	

Figure 6.7: Movie Rating Table Under Data Warehouse

title	cast
13 Sins	Devon Graye
13 Sins	Mark Webber
13 Sins	Pruitt Taylor Vince
13 Sins	Ron Perlman
13 Sins	Rutina Wesley
13 Sins	Tom Bower
16 Blocks	Bruce Willis
16 Blocks	Casey Sander
16 Blocks	Cylk Cozart
16 Blocks	David Morse
16 Blocks	David Zayas
16 Blocks	Jenna Stern
16 Blocks	Mos Def

Figure 6.8: Movie Actors Table Under Data Mart

title	country
13 Sins	United States
16 Blocks	Germany
16 Blocks	United States
17 Again	United States
20 Feet From Stardom	United States
2012	United States
20th Century Women	United States
21	United States
21 & Over	United States
23:59	Singapore
28 Days	United States
3 Days to Kill	France
3 Days to Kill	Serbia

Figure 6.9: Movie Country Table Under Data Mart

title	director
13 Sins	Daniel Stamm
16 Blocks	Richard Donner
17 Again	Burr Steers
20 Feet From Stardom	Morgan Neville
2012	Roland Emmerich
20th Century Women	Mike Mills
21	Robert Luketic
21 & Over	Jon Lucas
21 & Over	Scott Moore
23:59	Gilbert Chan
28 Days	Betty Thomas
3 Days to Kill	McG
3 Generations	Gaby Dellal

Figure 6.10: Directors Table Under Data Mart

title	genre
13 Sins	Horror
13 Sins	Thriller
16 Blocks	Action
16 Blocks	Adventure
16 Blocks	Crime
16 Blocks	Thriller
17 Again	Comedy
20 Feet From Stardom	Documentary
20 Feet From Stardom	Music
2012	Action
2012	Adventure
2012	Science Fiction
20th Century Women	Drama

Figure 6.11: Genres Table Under Data Mart

director	title	cast	release_year	keywords	show_rating	total_votes	popularity	genres
Daniel Stamm	13 Sins	Mark Webber, Rutina Wesley, Devon Graye, To...	2014	hotel, hostage, blackmail, detective, new orlea...	6.267	995	16.647	Horror, Thriller
Richard Donner	16 Blocks	Bruce Willis, Mos Def, David Morse, Jenna Stern...	2006	new york city, mission of murder, witness, betr...	6.424	1792	18.368	Action, Adventure, Crime, Thriller
Burr Steers	17 Again	Zac Efron, Leslie Mann, Matthew Perry, Thomas...	2009	high school, fake identity, bullying, high school ...	6.287	4788	32.875	Comedy
Morgan Neville	20 Feet From Stardom	Darlene Love, Merry Clayton, Lisa Fischer, Táta...	2013		7.233	223	13.659	Documentary, Music
Roland Emmerich	2012	John Cusack, Amanda Peet, Chiwetel Ejiofor, T...	2009	race against time, maya civilization, civilization, ...	5.819	11300	51.914	Action, Adventure, Science Fiction
Mike Mills	20th Century Women	Annette Bening, Elle Fanning, Greta Gerwig, Lu...	2016	parent child relationship, 1970s, balcony, femini...	7.348	1052	10.702	Drama
Robert Luketic	21	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aa...	2008	friendship, card game, casino, gambling, profes...	6.7	4317	25.256	Drama, Crime
Jon Lucas, Scott Moore	21 & Over	Miles Teller, Skylar Astin, Justin Chon, Sarah Wr...	2013	alcohol, birthday, friends, debauchery	5.815	1432	12.952	Comedy
Gilbert Chan	23:59	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	2011	singapore, jungle, soldier, ghost, chinese ouija	4.768	41	3.772	Horror
Betty Thomas	28 Days	Sandra Bullock, Viggo Mortensen, Dominic West...	2000	drug addiction, car crash, alcoholism, drug reha...	6.12	898	12.279	Comedy, Drama
McG	3 Days to Kill	Kevin Costner, Amber Heard, Hailee Steinfeld, ...	2014	daughter, cia, wife, killing, retirement, family, s...	6.068	1998	25.75	Action, Drama, Thriller
Gaby Dellal	3 Generations	Elle Fanning, Naomi Watts, Susan Sarandon, Ta...	2015	single mother, gender roles, lgbt, woman direct...	6.391	417	8.001	Comedy, Drama

Figure 6.12: Machine Learning Data Table

description	production_companies	production_countries	revenue	budget	profit	roi
Drowning in debt as he's about to get married, ...	Dimension Films, IM Global, Automatik Entertain...	United States of America	13809.00	5000000.00	-4986191.00	-99.723820
An aging cop is assigned the ordinary task of es...	Alcon Entertainment, Cheyenne Enterprises, Eq...	Germany, United States of America	65664721.00	55000000.00	10664721.00	19.390402
On the brink of a midlife crisis, 30-something Mik...	New Line Cinema	United States of America	136316880.00	20000000.00	116316880.00	581.584400
Backup singers live in a world that lies just beyo...	Tremolo Productions, Gil Friesen Productions	United States of America	5892466.00	1000000.00	4892466.00	489.246600
Dr. Adrian Helmsley, part of a worldwide geoph...	Columbia Pictures, Centropolis Entertainment, F...	United States of America	791217826.00	200000000.00	591217826.00	295.608913
In 1979 Santa Barbara, California, Dorothea Fie...	Annapurna Pictures, Archer Gray, Modern People	United States of America	5664764.00	7000000.00	-1335236.00	-19.074800
Ben Campbell is a young, highly intelligent, stud...	Columbia Pictures, Relativity Media	United States of America	159808370.00	35000000.00	124808370.00	356.595343
Brilliant student Jeff Chang has the most import...	Mandeville Films, Relativity Media, SkyLand Ent...	United States of America	48065672.00	13000000.00	35065672.00	269.735938
An army recruit was found dead during a 24km ...	Gorylah Pictures, Clover Films, Grand Brilliance, ...	Malaysia, Singapore	1208479.00	600000.00	608479.00	101.413167
After getting into a car accident while drunk on ...	Columbia Pictures	United States of America	6219894.00	43000000.00	-36780106.00	-85.535130
A dangerous international spy is determined to ...	Wonderland Sound and Vision, EuropaCorp, Rel...	France, United States of America	53260230.00	28000000.00	25260230.00	90.215107
A teenager transitions from female to male, and...	IM Global, Big Beach	United States of America	680351.00	10000000.00	-9319649.00	-93.196490

Figure 6.13: Machine Learning Data Table

roi
1
0
1
0
1
1
0
1
1
1
1
1
1

Figure 6.14: Machine Learning Prediction Result Table

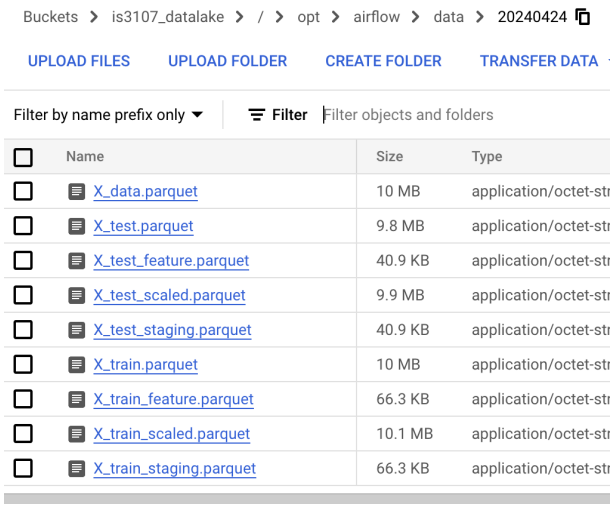

Figure 6.15: Caption

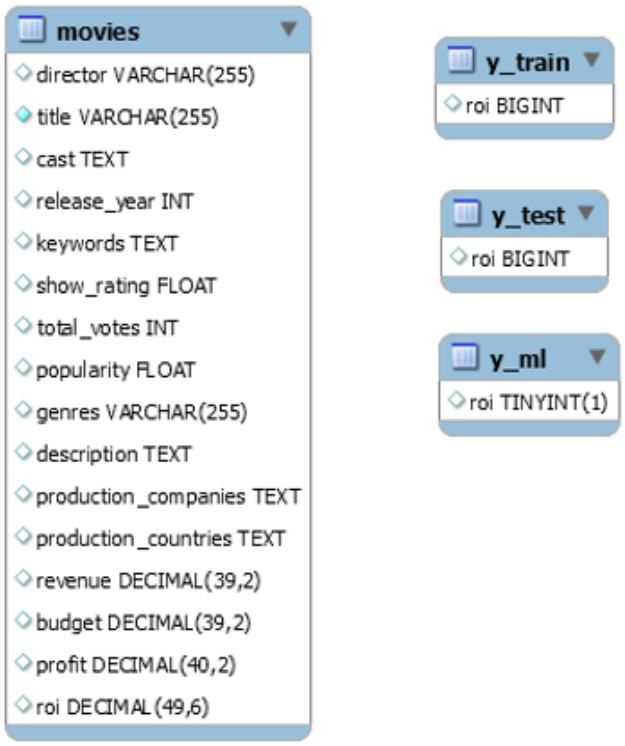

Figure 6.16: Machine Learning Schema

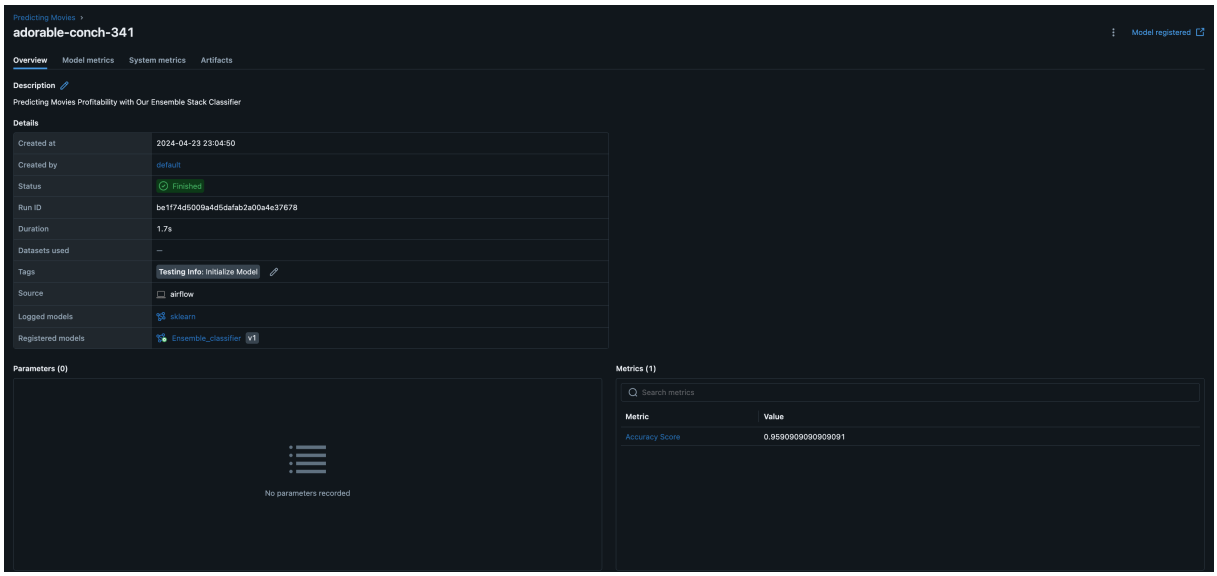

Figure 6.17: MLFlow

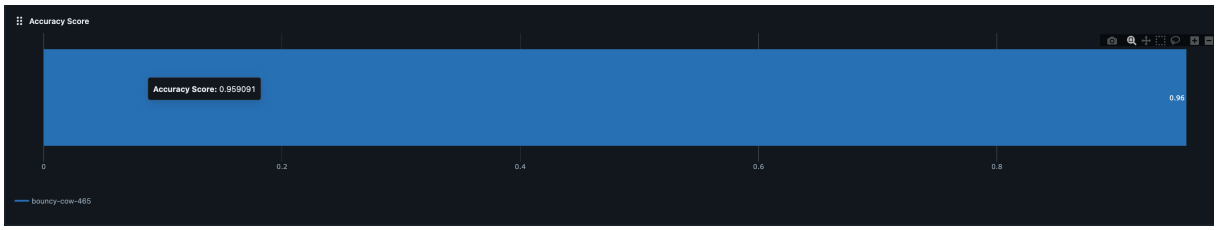

Figure 6.18: MLFlow Metrics

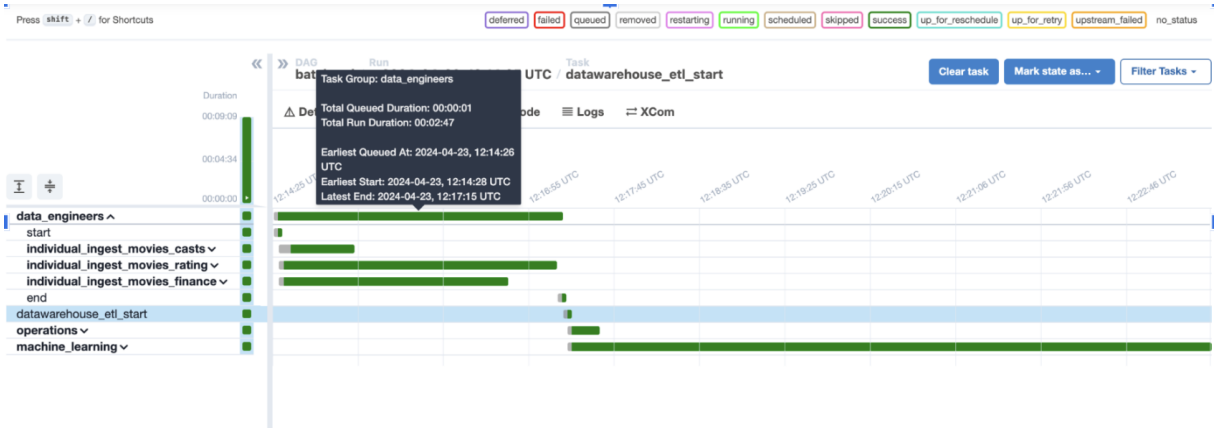

Figure 6.19: Gantt Chart: Batch ETL Start

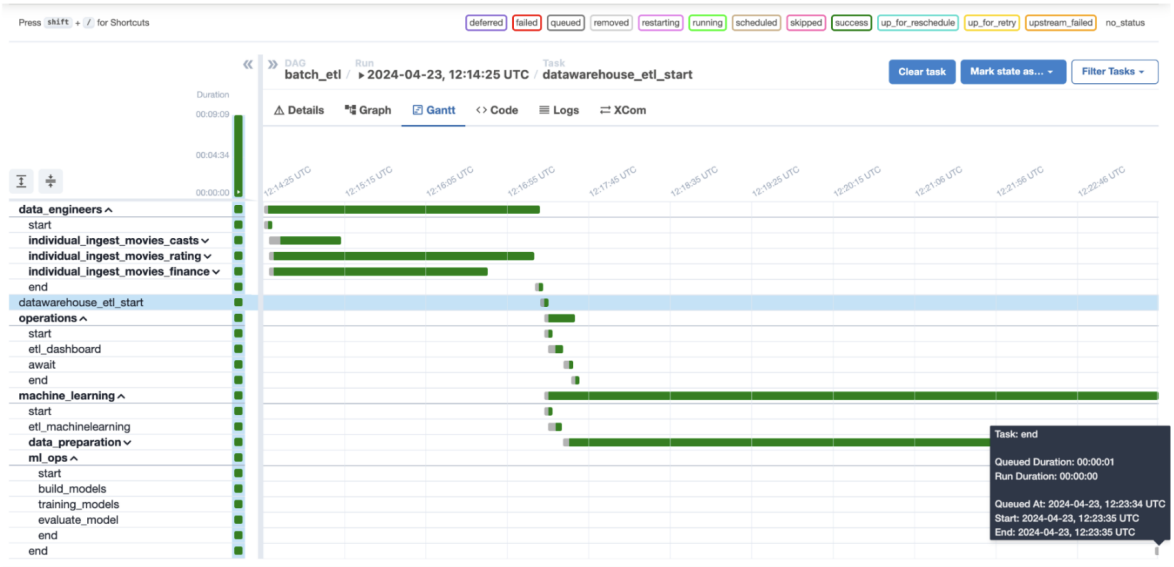

Figure 6.20: Gantt Chart: Batch ETL End