

# **A3: Design & Implementation of E2E Pipeline using Snowflake and Tableau**

**Business Intelligence**

Thursday, 17th of April 2025

By: Habib Mohamed Sultan Saleem

MSc Business Analytics (MBAN)

Table of Contents

[A3: Design & Implementation of E2E Pipeline using Snowflake and Tableau 1](#_Toc195826306)

[Business Problem Statement & Dataset Details 3](#_Toc195826307)

[ETL Pipeline and Machine Learning Process 4](#_Toc195826308)

[SQL + Tableau Insights 5](#_Toc195826309)

[Conclusion 9](#_Toc195826310)

[Appendix 9](#_Toc195826311)

[Reference 16](#_Toc195826312)

## Business Problem Statement & Dataset Details

In the fast-evolving streaming business, players such as Netflix have the perennial task of streamlining their content strategy at both global and local levels. With a catalogue as vast as its film and TV offerings, it is critical to know how different types of content perform differently in different countries, formats, and time frames to drive viewership, aid user retention, and inform investments in original content.

This project is to build an end-to-end data pipeline and business intelligence solution to facilitate data-driven decision-making at Netflix. With viewership trends analysis, drop-off behaviours, and predicted performance, to answer key strategic questions:

• Which are the most consistent performing shows worldwide?

• What type of content encourages long-term engagement?

• How does content perform differently by region and nation?

•How can Netflix strategically forecast content marketing and production through an analysis of viewer trends?

**Dataset Information**

It bases its findings on official Netflix Top 10, weekly-released data that contains:

* global\_alltime: Global "All-Time" hits, based on the original 91-day view (40 rows).
* global\_weekly: Weekly top 10 global rankings for titles (25K+ rows)
* country\_weekly: Country Top 10 weekly titles (120K+ rows)

Every dataset contains features such as:

* Title, Format, Category
* Week, Weekly Rank
* Hours of Viewing (91-day and weekly)
* Country specific appearance data

All three datasets are stored and cleansed within Snowflake's cloud data warehouse environment, enabling SQL analytics and dashboard integration.

## ETL Pipeline and Machine Learning Process

The project built an end-to-end machine learning and ETL methodology using Snowflake, SQL, Python, and Tableau. The aim was to create a scalable and reliable architecture that can Store raw Netflix datasets, transform them to use in analytics, and enable predictive business decisions through visualizations in the form of dashboards.

**Data extraction**

The datasets were imported in tsv format and included both weekly and all-time top 10 titles from Netflix. These files specifically consisted of global\_alltime.tsv, global\_weekly.tsv, and country\_weekly.tsv. All these files were loaded into Snowflake using a custom file format definition (tsv\_format) combining with the COPY INTO command. The data were then loaded into raw staging tables (raw\_global\_alltime, raw\_global\_weekly, raw\_country\_weekly) with error handling enabled using the ON\_ERROR = 'CONTINUE' parameter. The date parsing was controlled using the TRY\_TO\_DATE function to allow for type conversion.

**Data Modification**

Each raw table was then transformed using SQL into clean, analysis-ready tables. The transformation involved trimming and lowercasing strings, converting week columns into date format, and filtering out nulls. Cleaned tables like clean\_global\_alltime, clean\_global\_weekly, and clean\_country\_weekly were structured with schemas to simplify further analysis. Additional transformations included computing weekly views, runtime normalization, and aligning column naming conventions across datasets. These cleaned tables were used for both SQL-driven analytics and dashboard integration.

**Importing and Integrating Data**

The processed tables were connected to Tableau as stand-alone data sources, thus creating the base for several interactive dashboards. Machine learning predictions and simulation output were, in turn, sent back to Snowflake through the Python method, session.write\_pandas(). Integrating the machine learning output as live data sources in Tableau enabled the simultaneous display of predictions and forecasts along with descriptive analytics.

**Machine Learning Models**

The predictive analyses were incorporated into the dashboard by applying machine learning methodologies. The first was a K-Means clustering model, trained on the clean\_global\_alltime dataset. Features such as runtime, views\_first\_91\_days, and content format were used to group shows into distinct behavioural clusters. This enabled strategic segmentation of content. The model produced 4 clusters, and the outputs were visualized through a scatter plot in Tableau. In addition to the clustering approach, a Prophet time series forecast model was used to forecast the expected top shows on a week-to-week schedule, by both title and country. The resulting dataset (PREDICTED\_TOP\_SHOWS\_NEXT\_WEEK) supported trend analysis, the display of confidence intervals, and the performance of the scenario simulation model in Tableau.

## SQL + Tableau Insights

**Page 1: Top Performers and Global Trends**

**A screen shot of a black screen

AI-generated content may be incorrect.**

The dashboard interface offers an overview of the performance of Netflix on its platform. It starts with three KPIs: the total number of unique content titles, the overall global views for all titles, and the title with the most viewership on the platform. These statistics were obtained through SQL queries that aggregated unique values of show\_title and summed weekly\_views from the clean\_global\_weekly dataset. The most viewed title was found by ranking the shows based on their total lifetime view counts.

The “Most Consistent Movies & Shows” bar chart outlines the titles that have been most consistently in the Top 10 for the highest number of cumulative weeks worldwide. This analysis uses the MAX(cumulative\_weeks\_in\_top\_10) measure from the weekly dataset. Below it, a line chart displays the weekly viewing by genre, allowing for an exploration of content engagement trends over time. These kinds of data offer critical information about user preference shifts—particularly the ebbs and flows in popularity of movies, series, and regional content. Furthermore, a pie chart splits the content (e.g., movies versus TV and English versus Non-English), helping Netflix understand the overall content mix and the allocation of audience engagement worldwide.

**Page 2: Content Strategy Insights**

**A screenshot of a black screen

AI-generated content may be incorrect.**

The dashboard is intended to enable the execution of a data-driven strategy by the content and production teams at Netflix. One of the key measures used by Netflix to evaluate the success of recent titles is the all-time Top 10 titles, ordered by 91-day view counts, with the data sourced from the clean\_global\_alltime table.

The scatter plot entitled "Runtime Vs Views" investigates the relationship of television show duration to different measures of viewer engagement. Comparison of shorter shows to their longer equivalents opens relevant questions, thus allowing Netflix to optimize the runtimes of its upcoming productions. Apart from the graphical illustrations shown, the section explores the application of machine learning methods. The K-Means clustering, for instance, groups the shows in accordance with their duration and numbers of views, the resulting clusters displayed in a bubble chart. In each cluster, the characteristics of high-performing shows are outlined. A regression-driven scenario model known as "What If" provides Netflix with the ability to project potential performance based on theoretical variables such as length and type of format. The knowledge helps creative teams to evaluate the potential success of different pieces of content before their production.

**Page 3: Content Performance**

**A screen shot of a computer

AI-generated content may be incorrect.**

This dashboard offers extensive information regarding content performance across regions. It begins with a shaded world map showing the number of unique titles consumed within each nation. This figure, achieved by calculating COUNTD(show\_title) of the clean\_country\_weekly dataset, represents content variety as well as market saturation levels.

Below the map, a trend chart showing a decline in performance provides the weekly rankings of different titles. The chart, sorted by show name based on the weekly\_rank dataset, provides a clear representation of viewer attrition trends and the specific times at which audiences switch away from a show. Beside it is an interactive bubble chart titled “Content Longevity – Top Regional Titles,” that utilizes the length of time a title has spent in the Top 10 of each respective country. The analysis done during that process helps to identify regionally popular titles that are of cultural importance, potentially to be strategically used in upcoming campaigns or adaptations.

**Page 4: Forecasted Global Top Performance**

**A screen shot of a screen

AI-generated content may be incorrect.**

The final dashboard shows predictive analytics created using Prophet, a forecasting model for time-series data. Models for every nation were trained using past weekly rank data pulled out from the clean\_country\_weekly dataset. Later, the output obtained was loaded into Snowflake, where it was set up as a new table: PREDICTED\_TOP\_SHOWS\_NEXT\_WEEK. The top of the dashboard shows a bar chart that indicates the predicted Top 10 shows in the world for the upcoming week based on forecast rank scores. This tool provides useful insight into current content trends. Below that graphical representation, a world map indicates the forecasted performance of individual shows in different countries. Results can be filtered by show and country by users, allowing them to compare the performance of specific content in different markets. Bringing machine learning into the equation adds a forecasting element to the decision-making process, helping Netflix optimize its spending in the next budget cycle through advanced predictions.

## Conclusion

The BI solution demonstrated the performance of Netflix content through viewing data. Trends such as viewing, week over week, and format performance were displayed in dashboards. SQL analysis, machine learning, and Tableau converted raw data into insights.

Key Highlights:

* Squid Game and Bridgerton generate continued interest.
* Series exceeds films in viewing run-time ratios.
* Viewer preferences vary by country, and local productions have higher Top 10 consistency.
* Utilized Machine Learning to forecast top performers and performance of the content.

Recommendations

* Invest in Series: Profitable because of heavy viewership.
* Regional launches: Concentrate productions based on regional expertise.
* Apply ML for Planning: Utilize ML predictions to optimize campaigns.
* Monitor Drop-Off Trends: Early drop-offs lead to re-engagement efforts.

## Appendix

SQL Codes:  
CREATE DATABASE netflix\_project\_db;

CREATE SCHEMA netflix\_project\_db.netflix\_analytics;

CREATE OR REPLACE STAGE netflix\_stage

FILE\_FORMAT = (TYPE = 'CSV' FIELD\_OPTIONALLY\_ENCLOSED\_BY = '"' SKIP\_HEADER = 1);

CREATE OR REPLACE TABLE raw\_global\_weekly (

week DATE,

category STRING,

weekly\_rank NUMBER,

show\_title STRING,

season\_title STRING,

weekly\_hours\_viewed NUMBER,

runtime NUMBER,

weekly\_views NUMBER,

cumulative\_weeks\_in\_top\_10 NUMBER,

is\_staggered\_launch BOOLEAN,

runtime\_override\_flag STRING,

episode\_launch\_dtls STRING

);

CREATE OR REPLACE TABLE raw\_global\_alltime (

category STRING,

rank NUMBER,

show\_title STRING,

season\_title STRING,

hours\_viewed\_first\_91\_days NUMBER,

runtime NUMBER,

views\_first\_91\_days NUMBER

);

CREATE OR REPLACE TABLE raw\_country\_weekly (

country\_name STRING,

country\_iso2 STRING,

week DATE,

category STRING,

weekly\_rank NUMBER,

show\_title STRING,

season\_title STRING,

cumulative\_weeks\_in\_top\_10 NUMBER

);

CREATE OR REPLACE FILE FORMAT tsv\_format

TYPE = 'CSV'

FIELD\_DELIMITER = '\t'

SKIP\_HEADER = 1

FIELD\_OPTIONALLY\_ENCLOSED\_BY = '"';

TRUNCATE TABLE raw\_global\_weekly;

TRUNCATE TABLE raw\_global\_alltime;

TRUNCATE TABLE raw\_country\_weekly;

CREATE OR REPLACE TABLE clean\_global\_weekly AS

SELECT

LOWER(TRIM(show\_title)) AS show\_title,

LOWER(TRIM(category)) AS category,

weekly\_rank,

TRY\_TO\_DATE(week) AS week,

weekly\_hours\_viewed,

weekly\_views,

cumulative\_weeks\_in\_top\_10,

season\_title,

runtime,

is\_staggered\_launch,

runtime\_override\_flag,

episode\_launch\_dtls

FROM raw\_global\_weekly

WHERE show\_title IS NOT NULL;

CREATE OR REPLACE TABLE clean\_global\_alltime AS

SELECT

LOWER(TRIM(show\_title)) AS show\_title,

LOWER(TRIM(category)) AS category,

rank,

hours\_viewed\_first\_91\_days,

views\_first\_91\_days,

runtime,

season\_title

FROM raw\_global\_alltime

WHERE show\_title IS NOT NULL;

CREATE OR REPLACE TABLE clean\_country\_weekly AS

SELECT

country\_name,

country\_iso2,

TRY\_TO\_DATE(week) AS week,

LOWER(TRIM(category)) AS category,

weekly\_rank,

LOWER(TRIM(show\_title)) AS show\_title,

season\_title,

cumulative\_weeks\_in\_top\_10

FROM raw\_country\_weekly

WHERE show\_title IS NOT NULL;

-- FETCHING TOP 10 SHOWS BASED ON WEEKLY VIEWS FOR A PARTICULAR WEEK

SELECT show\_title, weekly\_views, weekly\_hours\_viewed

FROM clean\_global\_weekly

WHERE week = '2023-11-19'

ORDER BY weekly\_views DESC

LIMIT 10;

-- AGGREGATING TOTAL VIEWS GLOBALLY PER WEEK FOR TREND ANALYSIS

SELECT week, SUM(weekly\_views) AS total\_views

FROM clean\_global\_weekly

GROUP BY week

ORDER BY week;

-- SHOWING WEEKLY VIEWS TREND FOR A SELECTED SHOW

SELECT week, show\_title, weekly\_views

FROM clean\_global\_weekly

WHERE show\_title = 'squid game'

ORDER BY week;

-- FETCHING SHOWS WITH HIGHEST VIEWS IN FIRST 91 DAYS AFTER RELEASE

SELECT show\_title, views\_first\_91\_days, hours\_viewed\_first\_91\_days

FROM clean\_global\_alltime

ORDER BY views\_first\_91\_days DESC

LIMIT 10;

-- CALCULATING AVERAGE RUNTIME OF ALL-TIME TOP SHOWS

SELECT AVG(runtime) AS avg\_runtime

FROM clean\_global\_alltime;

-- COUNTING TOTAL WEEKS A PARTICULAR SHOW STAYED IN TOP 10 FOR EACH COUNTRY

SELECT country\_name, COUNT(\*) AS total\_weeks\_in\_top10

FROM clean\_country\_weekly

WHERE show\_title = 'money heist'

GROUP BY country\_name

ORDER BY total\_weeks\_in\_top10 DESC;

-- FINDING SHOWS OR MOVIES WHICH STAYED MAXIMUM WEEKS IN GLOBAL TOP 10

SELECT show\_title, MAX(cumulative\_weeks\_in\_top\_10) AS max\_weeks

FROM clean\_global\_weekly

GROUP BY show\_title

ORDER BY max\_weeks DESC

LIMIT 10;

-- SUM OF TOTAL VIEWS BASED ON CONTENT CATEGORY

SELECT category, SUM(weekly\_views) AS total\_views

FROM clean\_global\_weekly

GROUP BY category

ORDER BY total\_views DESC;

-- PREPARING WEEKLY DATA FOR FORECAST VISUALIZATION IN TABLEAU

SELECT week, SUM(weekly\_views) AS total\_views

FROM clean\_global\_weekly

GROUP BY week

ORDER BY week;

## Reference

OpenAI. (2023). *ChatGPT* (Mar 14 version) [Large language model]. https://chat.openai.com/

Grammarly Inc. (2023). *Grammarly* (Version 1.0) [AI-powered writing assistant software]. https://www.grammarly.com/