Bogazici Univesity Master of Science Program in Software Engineering

SWE586 Data Science and Big Data Management

Project Title: Building a Data Pipeline with Lakehouse Architecture on Microsoft Azure Platform

Student Name: Bilge Akpulat

Student Number: 2023719012

TABLE OF CONTENTS

INTRODUCTION

In today’s data-driven world, the ability to efficiently process, analyze, and derive insights from large datasets is critical for organizations across various industries. This project focuses on building an end-to-end data pipeline to analyze Netflix's content dataset, leveraging Azure cloud services to implement a scalable and reliable solution. The project follows a structured data pipeline model, transitioning data through ingestion, processing, storage, and serving layers, to create actionable insights for business and academic purposes. The goal is to identify trends in Netflix content production across regions and categories, such as the proportion of modern vs. classic content and the distribution of content duration and count by country. The pipeline design ensures flexibility, automation, and adaptability to future data needs, adhering to best practices in data engineering.

For the researcher, this study was eye opening as it was the first time for using many tools at hand.

Key Components

Azure Data Factory (ADF): Used for orchestrating and automating data ingestion and transformation processes.

Azure Data Lake Storage Gen2: Serves as the storage layer, structured into Bronze, Silver, and Gold layers for raw, processed, and analytics-ready data.

Azure Synapse Analytics: Enables querying, visualizing, and analyzing data using SQL external tables.

External Tables: Facilitates the serving layer by providing seamless access to processed data stored in the "Gold" layer.

Visualization Tools: SQL-based visualizations in Azure Synapse replace Power BI due to subscription constraints.

Keywords

Data Pipeline, Azure Cloud Services, Data Factory, Data Lake Storage, Synapse Analytics, Content Analysis, Modern vs. Classic Content, Regional Content Trends, Machine Learning Integration

1. ARCHITECTURAL DETAILS

1.a. Work Schemas of the Architeture

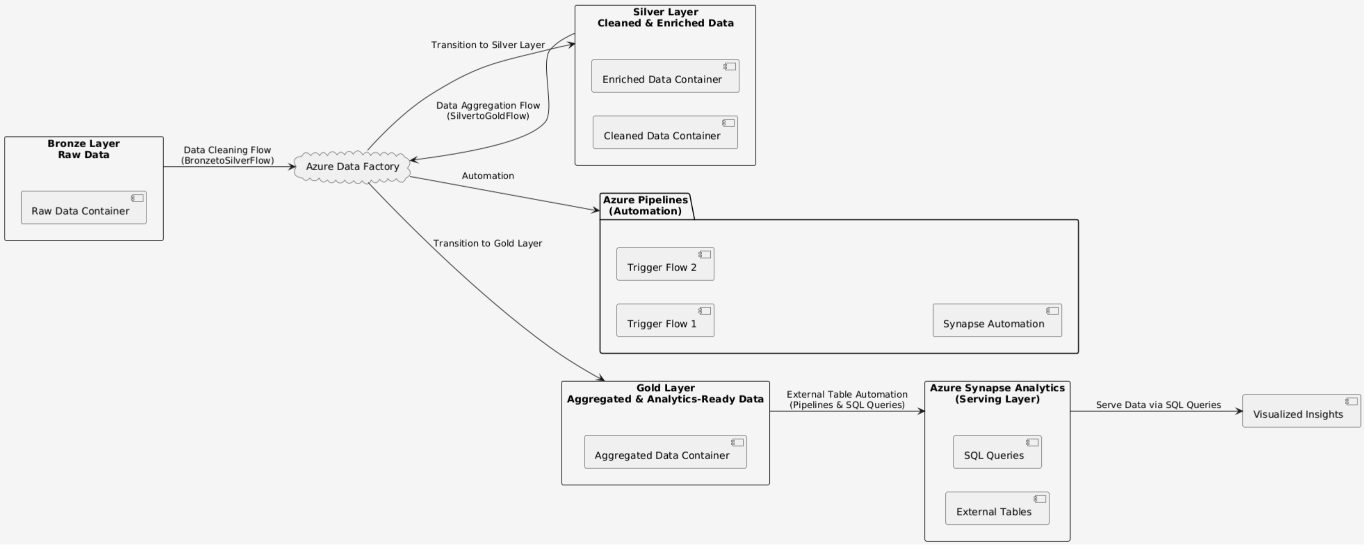


Figure 1. Architectural Details of the Project

Data Containers: Raw Data Container, Cleaned Data Container, Aggregated Data Container.

Azure Data Factory: Handles the transformation and transition between Bronze, Silver, and Gold layers via two distinct flows.

Automation with Pipelines: Pipelines triggering the Data Factory flows and Synapse queries for end-to-end automation.

Azure Synapse Analytics: External tables and queries handle the serving and visualization layer.

Visualization: Synapse is directly used for querying and generating insights, skipping Power BI and Databricks

1.b. Work Schemas of the Architeture for Data Processing Flows

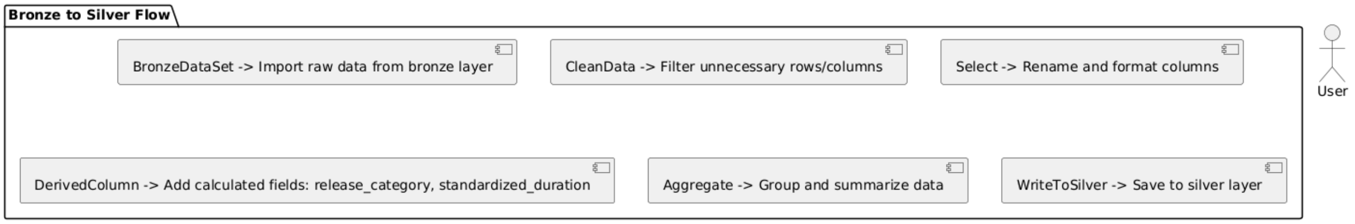

Bronze to Silver Flow
BronzeDataSet -> Import raw data from bronze layer
CleanData -> Filter unnecessary rows/columns
Select -> Rename and format columns
User
DerivedColumn -> Add calculated fields: release_category, standardized_duration
Aggregate -> Group and summarize data
WriteToSilver -> Save to silver layer

Figure 2. Bronze to Silver Flow Use Case Diagram

BronzeDataSet: Imports raw data from the "bronze" layer in Azure Data Lake Storage.

CleanData: Filters unnecessary rows and columns based on specified expressions.

Select: Renames the columns and formats them for the next transformations.

DerivedColumn: Adds calculated columns (release_category, standardized_duration, etc.).

Aggregate: Groups the data based on attributes such as type and release_category, summarizing relevant metrics.

WriteToSilver: Writes the cleaned and processed data to the "silver" layer in Azure Data Lake Storage.

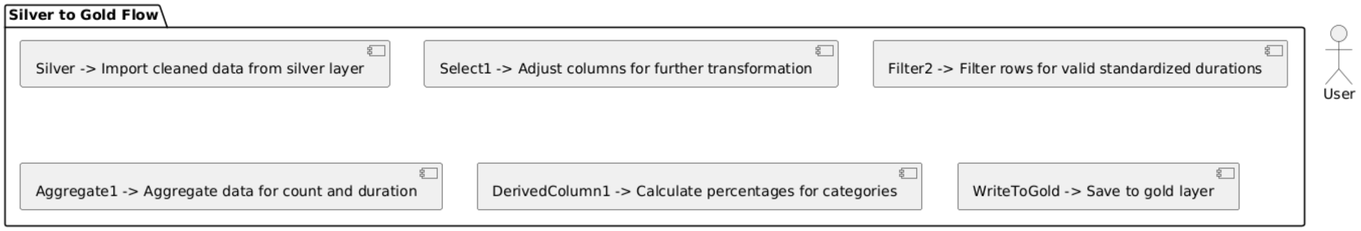

Silver to Gold Flow
Silver -> Import cleaned data from silver layer
Select1 -> Adjust columns for further transformation
Filter2 -> Filter rows for valid standardized durations
User
Aggregate1 -> Aggregate data for count and duration
DerivedColumn1 -> Calculate percentages for categories
WriteToGold -> Save to gold layer

Figure 3. Bronze to Silver Flow Use Case Diagram

Silver: Imports cleaned data from the "silver" layer.

Select1: Applies additional column adjustments and formats the input data.

Filter2: Filters rows based on specific attributes, e.g., filtering by standardized duration.

Aggregate1: Aggregates the data for calculating total counts and durations grouped by primary_country.

DerivedColumn1: Calculates percentages for modern/classic counts and durations.

WriteToGold: Writes aggregated and analytics-ready data to the "gold" layer in Azure Data Lake Storage.

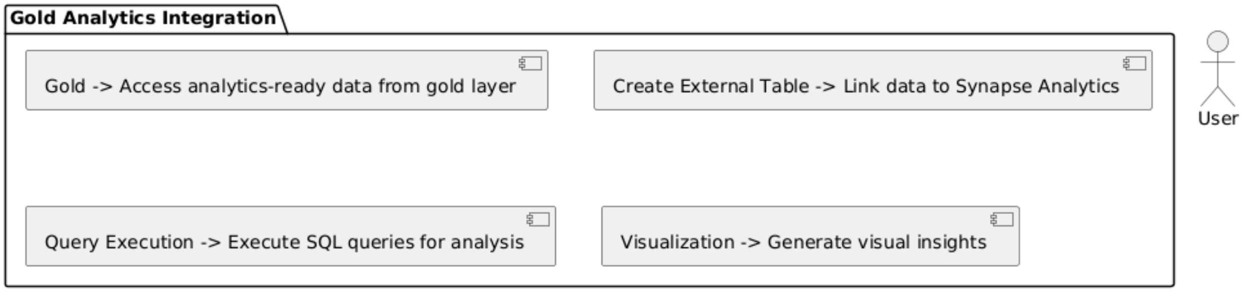

Gold Analytics Integration
Gold -> Access analytics-ready data from gold layer
Create External Table -> Link data to Synapse Analytics
User
Query Execution -> Execute SQL queries for analysis
Visualization -> Generate visual insights

Figure 4. Gold Layer to Analysis Use Case Diagram

Gold: Reads data from the "gold" layer.

Create External Table: Creates external tables in Synapse Analytics, linking the gold data for querying.

Query Execution: Executes SQL queries to derive insights, e.g., modern/classic content analysis by country.

Visualization: Generates bar charts, pie charts, and tables directly in Synapse Analytics based on the queried data.

2. DATA SELECTION AND MENTAL ROAD MAP

In the context of analyzing content trends and their distribution across global platforms, this project focused on Netflix's publicly available dataset. The dataset, encompassing attributes such as primary_country, release_year, type, duration, and release_category, was selected for its richness in metadata and relevance to understanding global media production patterns. The decision to work with this dataset stems from its ability to address key analytical questions about content production, consumption trends, and temporal dynamics.

The goal of this project is not merely to analyze raw counts but also to explore deeper patterns that reveal insights into modern versus classic content creation and how these trends vary across different nations. By defining these analytical objectives from the beginning, the project ensures alignment with broader goals in media analytics, which aim to uncover actionable insights for both audiences and content creators.

With this aim in mind and at heart, the data was selected from the keagle platform, see here.

2.a. Justification for Dataset Selection

The chosen dataset holds a unique position in media analytics due to its comprehensive metadata. Its fields enable the exploration of content production not only by count but also by aggregated metrics such as total duration and categorical analysis. Furthermore, Netflix, as a dominant player in global streaming, provides a representative lens into evolving trends in media production across countries and genres.

The dataset was particularly well-suited to address several research questions, as it includes both temporal data (release year) and categorical data (release_category, type, etc.), allowing for longitudinal and categorical analysis. Additionally, the inclusion of primary_country enables cross-national comparisons, which are critical for understanding how content production varies across geopolitical regions.

2.b. Research Questions and Analytical Focus

Before starting out, the necessity of clear requirements analysis and specification were taken into consideretaion. In other words, what would the researcher like to ask the data, and then create the questions before getting lost in the Azure platform.

The analytical roadmap for this project was designed to focus on ten specific questions that explore the dataset's potential:

1. **Global Content Production by Count**: A bar chart visualization which country produces the most content by count helps reveal the dominant players in the media landscape.
2. **Global Content Production by Total Duration**: A bar chart visualization which country produces the most content by total duration provides additional depth, highlighting nations that prioritize lengthier formats.
3. **Modern Content Creation by Count Percentage**: Identifying countries that focus predominantly on modern content (release_category = Modern) as a percentage of their total count highlights innovative hubs of production.
4. **Modern Content Creation by Duration Percentage**: Measuring the same trend by total duration underscores where innovation aligns with longer-form content production.
5. **Classic Content Creation by Count Percentage**: Highlighting countries that produce the most classic content as a percentage of total count allows for an exploration of heritage or traditional storytelling.
6. **Classic Content Creation by Duration Percentage**: A similar analysis by duration provides a complementary view, often revealing which countries invest more heavily in sustaining long-form classic media.

7. **Content Production Trends Over the Last Decade**: By focusing on content produced within the past 10 years, this bar chart narrows the analysis to more recent trends while omitting questions with redundant insights.
8. **Comparison of Modern vs. Classic Content by Count**: A pie chart answering whether Netflix's catalog today leans more toward modern or classic content by count provides an immediate snapshot of contemporary versus nostalgic focus.
9. **Comparison of Modern vs. Classic Content by Duration**: Analyzing this distribution by total duration rather than count provides an additional dimension, considering the time investment in content.

3. DATA INGESTION LAYER

3.a. Orchestration of Batch Data Ingestion with Azure Data Factory (ADF)

Azure Data Factory was utilized to automate the data ingestion process. Data Factory pipelines were designed to extract raw Netflix data from the source (e.g., an external file or database) and load it into the Azure environment.

This was later deleted as an operation after executing once, as this was seen as the best fit for the ongoing operations, not to confuse the next layers with this one.

3.b. Saving Raw Data into Azure Data Lake Storage Gen2 (Bronze Layer):

A dedicated container named Bronze was created in Azure Data Lake Storage Gen2 to store the raw, unprocessed data. The raw data ingestion ensured all original fields and formats were preserved for traceability and further processing. This storage acted as the foundational layer of the Lakehouse architecture, maintaining data in its rawest form for compliance and auditing purposes.

3.c. Key Steps

1. Data Source Definition: Defined the source of the Netflix dataset, ensuring it was accessible for ADF ingestion.
2. Batch Ingestion Pipeline: Created an ADF pipeline with a Copy Data activity. Scheduled the pipeline to ingest data in batches, ensuring reliability and scalability. The operation was removed after use.
3. Destination Configuration: Configured the destination as Azure Data Lake Storage Gen2 (Bronze container). Verified the successful ingestion of raw data in the Bronze container.

3.d. Challenges and Solutions

Challenge: Ensuring the integrity and completeness of data during ingestion.

Solution: Implemented retry policies in ADF pipelines to handle transient failures during ingestion.

Challenge: Handling varied data formats (e.g., CSV, JSON, Parquet).

Solution: Used schema flexibility within ADF to accommodate different file types.

3.e. Outcome

Raw Netflix dataset was successfully ingested into the Azure ecosystem.

The Bronze container in Azure Data Lake Gen2 served as the source of truth for downstream transformations and analytics.

4. DATA PROCESSING LAYER

4.1. Azure DataBricks Solution

Azure Databricks was considered as a potential solution for processing and transforming the data. However, during the stage of mounting the data, the process encountered a

Quota Exhaustion issue in the Azure subscription. This error, detailed here, rendered the Azure Databricks Python notebooks unusable.

Specifically:

1. The subscription type associated with the Azure account imposed **limitations** on the computational and storage resources.
2. Even after optimizing the pipeline and scaling down resource usage, the **Quota Exhaustion issue** persisted, blocking the use of Databricks for this project.

4.2. Why Azure Data Factory (ADF) Data Flow Was Chosen

Given the challenges with Databricks, **Azure Data Factory Data Flow** was selected for the following reasons:

1. **Ease of Use:** ADF provides an intuitive, graphical interface that simplifies complex data transformation tasks. It allows non-technical users to design and manage ETL pipelines without writing extensive code.
2. **Resemblance to Other Real-World Applications:** Many modern ETL tools, such as Informatica and Talend, utilize similar visual paradigms for defining transformations. This makes ADF Data Flow a familiar tool for professionals.
3. **Error Handling and Iteration:** In Data Flow, debugging and fixing errors are straightforward. Issues in transformations can be visualized and corrected in real-time, enhancing efficiency and reliability.
4. **Seamless Integration with Azure Ecosystem:** ADF Data Flow integrates natively with Azure services such as Data Lake Gen2, Synapse, and Azure SQL, ensuring smooth data movement and processing.

4.3. Implementation of Azure Data Factory

To meet the project’s requirements, the following data processing tasks were planned and implemented:

Cleaning and Preprocessing the Data: Removal of unnecessary rows and columns, filling or removing null values, and ensuring consistency in data formats.

Applying Transformations: Adding derived columns, normalizing inconsistent data fields, and enriching the dataset with new metrics (e.g., standardized duration, release category).

The data was stored in three distinct layers:

1. **Bronze Layer:** Raw data ingested directly into Azure Data Lake Storage.
2. **Silver Layer:** Cleaned and transformed data, ready for advanced analysis.
3. **Gold Layer:** Aggregated and analytics-ready data, structured for querying and reporting.

4.3.a. From Bronze to Silver

The BronzeToSilverFlow pipeline was developed to handle the transition from raw to cleaned data.
The steps involved were:

1. **Importing Raw Data (Bronze Layer):**

   The data from the Bronze container was ingested into the Data Flow. The dataset was loaded with all columns, preserving the raw structure.

2. **Cleaning the Data:**

   Filtering: Rows with invalid or null values in critical columns (e.g., release_year, type) were filtered out using the CleanData activity.

   Renaming Columns: Columns with ambiguous or technical names were renamed to more user-friendly alternatives.

1. show_id was retained for uniqueness.
2. type and duration were left intact but were augmented with additional columns.

3. **Derived Columns:**

   Release Category: A column was added to categorize the content as "Modern" (released in or after 2000) or "Classic" (before 2000).

   Standardized Duration (mins):

   Movies: Extracted numeric values (e.g., 65 mins → 65).

   TV Shows: Converted seasons into an estimated duration (1 season = 12 episodes × 40 minutes).

   Note: Be careful in Azure as standart_duration(min) or standart_duration(minutes) is likely to be misunderstood in format.

   Primary Country: Extracted the first country listed in the country column (e.g., US, UK → US).

4. **Aggregation:**

   Grouped the data by type and release_category to calculate: total count of records and total standardized duration in minutes.

5. **Writing to Silver Layer:**

   The processed data was exported to the **Silver** container in **Parquet** format for optimized querying and storage efficiency.

   As a result:

   - Data inconsistencies were resolved.

- The dataset was enriched with actionable insights (e.g., categorized content, standardized durations).
- Data was ready for further aggregation and transformation in the **SilverToGoldFlow** pipeline.

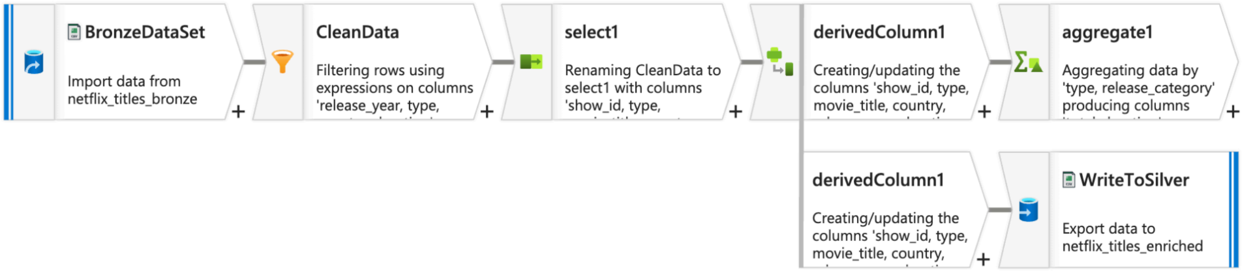

BronzeDataSet
Import data from netflix_titles_bronze
CleanData
Filtering rows using expressions on columns 'release_year, type,
select1
Renaming CleanData to select1 with columns 'show_id, type,
derivedColumn1
Creating/updating the columns 'show_id, type, movie_title, country,
aggregate1
Aggregating data by 'type, release_category' producing columns
derivedColumn1
Creating/updating the columns 'show_id, type, movie_title, country,
WriteToSilver
Export data to netflix_titles_enriched

Figure 5. Bronze to Silver Data Flow in Azure Data Factory

4.3.b. From Silver to Gold Layer

The SilverToGoldFlow pipeline is designed to transition cleaned and enriched data from the Silver Layer to the Gold Layer for aggregated, analytics-ready outputs. Below is a detailed step-by-step explanation of the process

1. Importing Data from the Silver Layer

The first step involves reading the data from the **Silver** container. This data is the cleaned and enriched output of the Bronze to Silver process. The **Silver** dataset is imported and serves as the input for subsequent transformations.

2. Select Transformation

Objective: Rename columns and filter out unnecessary ones for better clarity and relevance in downstream analytics.

Implementation: Columns from the Silver Layer are selected and renamed where necessary for improved readability and usability. This ensures only the required columns move forward in the pipeline.

3. Filter Transformations

There are two distinct filtering operations in this flow:

Filter1: Identify Recent Content (Last 10 Years)

Objective: Retain rows where content was released in the last 10 years.

Expression Used: A derived column (is_last_10_years) was created earlier to flag content released within the last decade. This filter utilizes that column to extract relevant rows.

Filter2: Retain Relevant Duration

Objective: Exclude rows where the standardized duration is null or has invalid values.

Expression Used: Rows with valid values in the standardized_duration(min) column are kept, ensuring accurate duration-based calculations.

4. Aggregation Transformations

Two separate aggregation steps are applied to generate the required analytics:

Aggregate1: Group Data by Primary Country

Objective: Summarize data by the primary_country column to calculate metrics such as: Total count of entries, Total standardized duration of content & Count of modern and classic content for each country.

Output Columns: primary_country, total_count, total_duration, modern_count, classic_count

Aggregate2: Analyze Recent Content by Country

Objective: Focus specifically on content released in the last 10 years, grouped by primary_country.

Metrics: Count of recent entries for each country & Total standardized duration of recent content.

5. Derived Columns

A new transformation step creates additional columns to enrich the dataset further:

Modern Percentage (Count): Percentage of modern content by count relative to the total count.

Classic Percentage (Count): Percentage of classic content by count relative to the total count.

Modern Percentage (Duration): Percentage of modern content by duration relative to the total duration.

Classic Percentage (Duration): Percentage of classic content by duration relative to the total duration.

6. Write to Gold Layer

The final step writes the aggregated and enriched data to the **Gold Layer**.

The data is saved in **Parquet format** for optimized storage and querying.

The Gold Layer now contains analytics-ready data that can be directly used for visualization and reporting in Synapse or other BI tools.

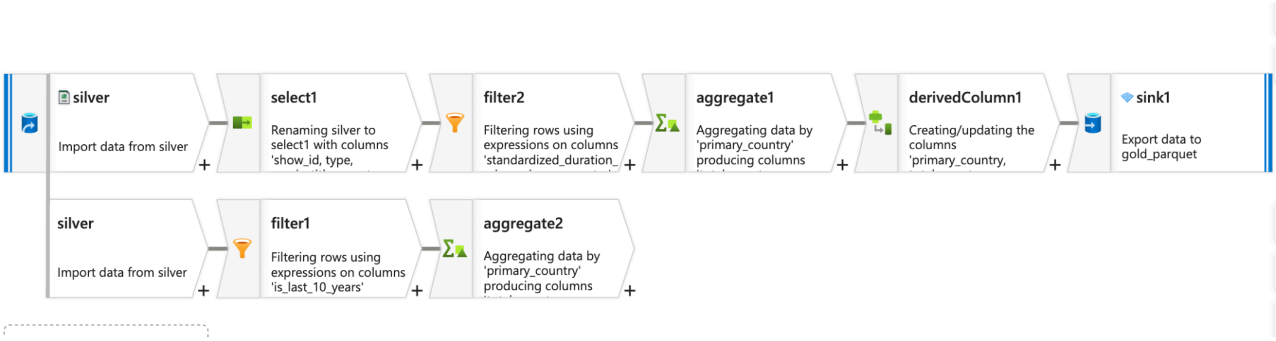

silver
Import data from silver
select1
Renaming silver to select1 with columns 'show_id, type,
filter2
Filtering rows using expressions on columns 'standardized_duration_
aggregate1
Aggregating data by 'primary_country' producing columns
derivedColumn1
Creating/updating the columns 'primary_country,
sink1
Export data to gold_parquet
silver
Import data from silver
filter1
Filtering rows using expressions on columns 'is_last_10_years'
aggregate2
Aggregating data by 'primary_country' producing columns

Figure 6. Silver to Gold Data Flow in Azure Data Factory

Key Highlights: The SilverToGoldFlow ensured that the data was **aggregated** and **enriched**, ready for high-level analytics and visualizations. By using **filtering, aggregation, and derived columns**, this process extracted insights tailored to the project's specific goals, such as identifying content trends by country, release year, and duration.

4.3.c. Bronze to Gold Pipeline

The pipeline demonstared below represents the integration and automation of the BronzeToSilverFlow and SilverToGoldFlow processes within Azure Data Factory. This pipeline enables seamless execution of the two flows in sequence, ensuring that data transitions smoothly from the Bronze Layer (raw data) to the Silver Layer (cleaned and enriched data) and finally to the Gold Layer (aggregated and analytics-ready data). The success dependency between the two activities ensures that the second flow, responsible for preparing the Gold Layer, only executes after the successful completion of the first flow. This design not only simplifies the workflow but also automates the end-to-end data lifecycle, making it robust, scalable, and easy to manage for real-world data processing scenarios.

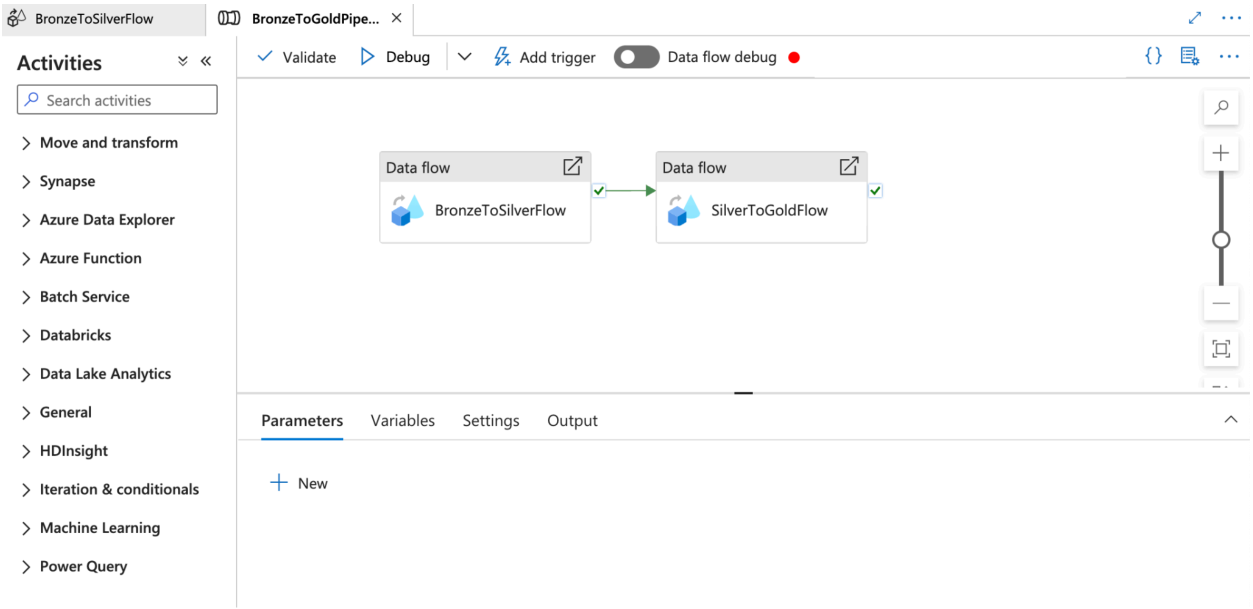


Figure 7. Bronze to Gold Data Flow Pipliene in Azure Data Factory

5. DATA STORAGE LAYER

The Storage Layer design ensures a structured and efficient pipeline for processing and analyzing data using the lakehouse architecture. The three layers—Bronze, Silver, and Gold—each serve distinct purposes in the data lifecycle.

5.a. Bronze Layer

This layer serves as the foundation, storing raw, unprocessed data as ingested from the source systems. The data here is in its original format, maintaining fidelity to ensure traceability. In this project, the Netflix dataset was stored in the "bronze" container in Azure Data Lake Gen2, preserving its original state to serve as a reliable backup for downstream corrections or reprocessing if needed.

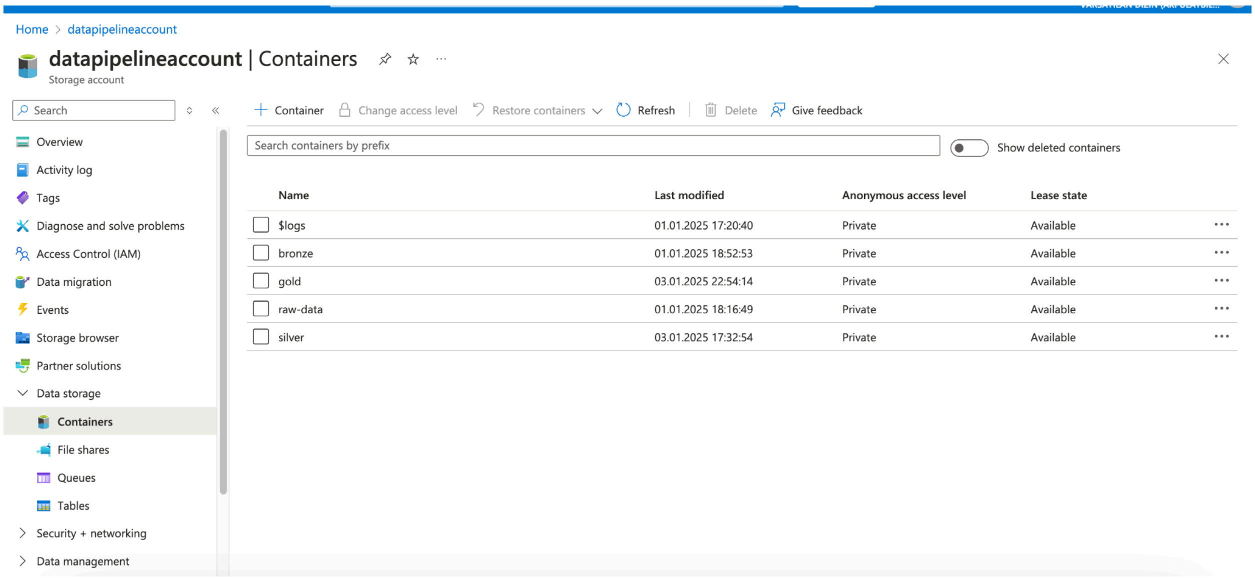
5.b. Silver Layer

The Silver Layer processes the raw data from the Bronze Layer, transforming it into cleaned and enriched datasets. In this step, irrelevant rows and columns were filtered, columns were renamed for clarity, and additional derived columns like release_category and standardized_duration were added. This intermediate layer ensures the data is clean and standardized, enabling smoother processing in subsequent stages. The cleaned data was stored in the "silver" container in Azure Data Lake Gen2.

5.c. Gold Layer

Gold Layer: The Gold Layer contains aggregated and analytics-ready data derived from the Silver Layer. Here, transformations focused on summarizing data for specific analytical questions, such as total counts and durations, and creating percentage distributions for "modern" and "classic" content. This layer is optimized for querying and serves as the backbone for downstream reporting and analytics workflows. The final dataset is stored in the "gold" container in Azure Data Lake Gen2 in Parquet format, ensuring high performance for querying via Synapse Analytics.

These layers collectively form a robust and scalable data storage strategy, supporting both operational and analytical use cases seamlessly.



Name	Last modified	Anonymous access level	Lease state
$logs	01.01.2025 17:20:40	Private	Available
bronze	01.01.2025 18:52:53	Private	Available
gold	03.01.2025 22:54:14	Private	Available
raw-data	01.01.2025 18:16:49	Private	Available
silver	03.01.2025 17:32:54	Private	Available

Figure 8. Data Containers in Azure

6. DATA SERVING LAYER

The serving layer was designed to facilitate downstream analytics and insights through Azure Synapse Analytics. The main objective of this layer was to enable seamless querying of the "gold" layer data, which contains aggregated and analytics-ready data stored as a parquet file in Azure Data Lake Storage Gen2.

6.a. Key Steps in the Serving Layer

1. **Connecting Synapse to Gold Data Layer**

   The gold parquet file was made accessible by creating an external data source (gold_datapipelineaccount_dfs_core_windows_net) in Synapse. This established the linkage between Synapse and the Azure Data Lake.

   An external file format (SynapseParquetFormat) was defined for handling parquet files.

2. **Creating External Tables**

   External tables such as netflix_data5 were created in Synapse to expose the gold layer data for SQL queries. The external table structure precisely matched the columns of the gold data, such as primary_country, total_count, modern_count, and others.

   These tables provided a direct interface for running SQL queries and producing insights.

3. **Writing and Running SQL Queries**

   A set of SQL scripts was written and executed to answer predefined business questions about the dataset. These included:

   1. Finding the country producing the most content by count (1produces_the_most_content_by_count_Bar.sql).

2. Identifying the country producing the most content by total duration (2produces_the_most_content_by_total_duration_Bar.sql).
3. Calculating modern content percentages by count and duration (3modern_content_percentage_by_count.sql and 4modern_content_percentage_by_duration.sql).
4. Aggregating data to compare modern and classic content by count and duration (9totalModernvsClassic.sql and 10totalModernvsClassic_byDuration.sql).
5. Listing the top 5 countries producing classic content with percentages (6top5_by_count_modern.sql).

4. Validation and Visualization

Query results were validated by inspecting data using SELECT TOP 100 statements to ensure data integrity and accuracy.

The results were utilized for generating visualizations like bar charts and pie charts directly within Synapse or exported to downstream tools like Power BI for enhanced visual representation.

5. Benefits of This Approach

Efficiency: External tables in Synapse enable fast querying of large parquet datasets without moving or duplicating data.

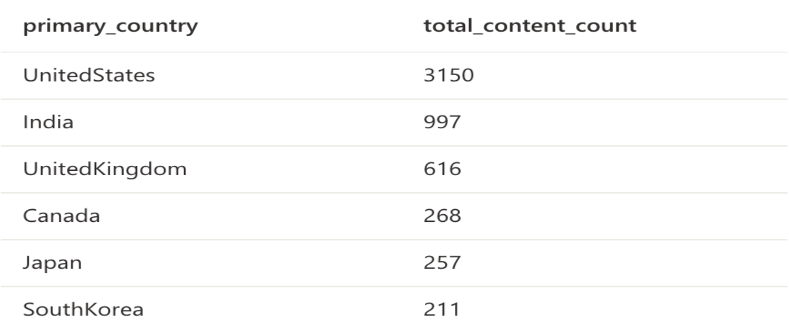
Scalability: The setup leverages Azure's storage and compute scalability to handle larger datasets and complex queries.

Flexibility: SQL-based querying provides flexibility to iterate on business questions, enabling real-time updates to analytics workflows.

7. VISUALIZATION

In this project, visualization played a critical role in deriving insights from the processed data. Although Power BI is a leading tool for creating rich and interactive dashboards, its lack of compatibility with macOS and its licensing cost created significant barriers. As a result, an alternative approach was taken by utilizing Azure Synapse Analytics for direct SQL-based visualizations, which offered both simplicity and efficiency for this project.

7.a. “Which country produces the most content by count?"



primary_country	total_content_count
UnitedStates	3150
India	997
UnitedKingdom	616
Canada	268
Japan	257
SouthKorea	211

Figure 9. The External Table-1

This table represents our values comperatively and numericaly.

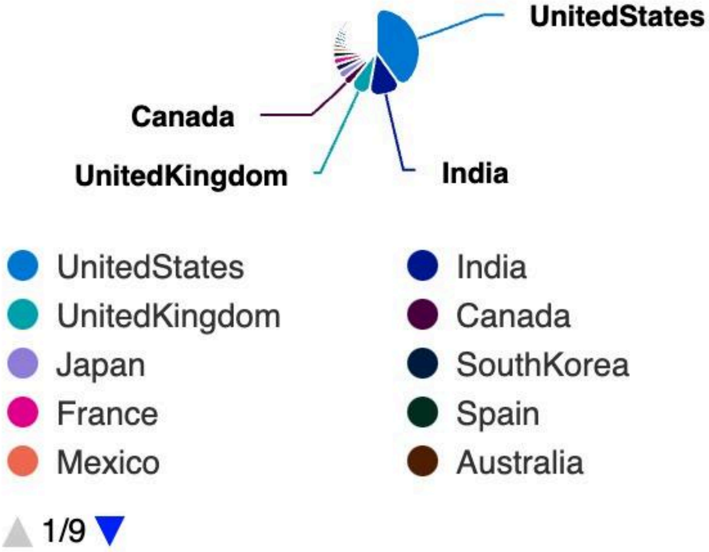


Figure 10. Pie Chart-1

The pie chart provides a high-level overview of content distribution among countries. It highlights that the United States dominates content production, followed by countries such as India, United Kingdom, and Canada. Smaller slices indicate contributions from less prominent countries in the dataset.

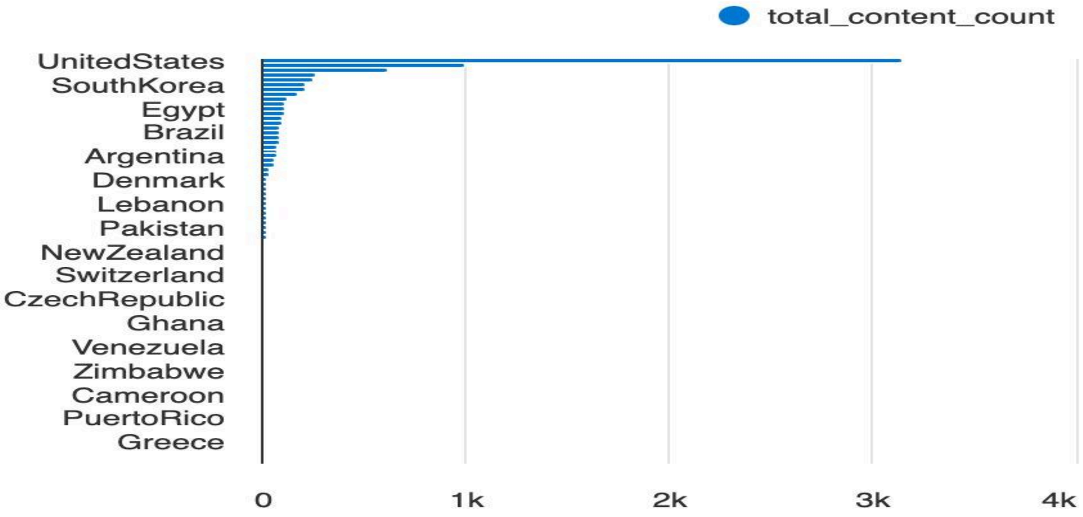


Figure 11. Bar Chart-1

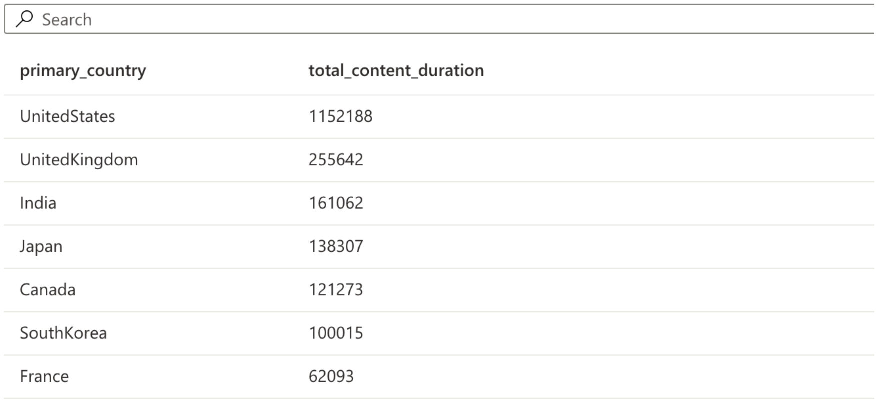
The bar chart offers a more granular breakdown of the total content count by country. The United States stands significantly ahead with an overwhelming count of produced content, reflecting its large entertainment industry and streaming production capabilities.

Other countries, while present, show a steep drop in contribution compared to the United States.

Interpretation

These visualizations collectively demonstrate that the **United States** is the largest contributor to the dataset, underscoring its dominance in the global streaming industry. Countries such as India and the United Kingdom show moderate contributions, indicating their growing role in content production. These insights provide a foundation for further analyses, such as examining content type, duration, and trends over time.

7.b. “Which country produces the most content by volume(duration in mins)?"



primary_country	total_content_duration
UnitedStates	1152188
UnitedKingdom	255642
India	161062
Japan	138307
Canada	121273
SouthKorea	100015
France	62093

Figure 12. The External Table-2

This table provides a numerical breakdown of the total content duration for each country in the dataset. The United States dominates with over 1.1 million total minutes of content, showcasing its strong presence in Netflix's library. Following it are the United Kingdom and India, both with significantly lower figures but still substantial compared to other countries. The insights from this table highlight the disproportionate contribution of the United States to Netflix's catalog, making it a crucial player in content production.

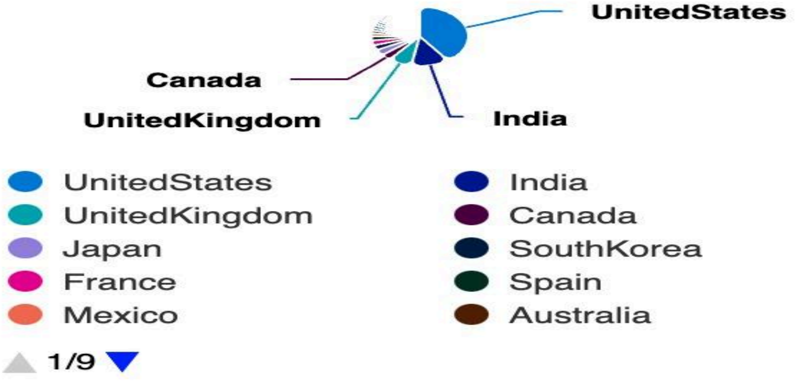


Figure 13. Pie Chart-2

This pie chart visualizes the percentage contribution of various countries to the total content duration on Netflix. The United States takes up the largest slice, clearly showing its dominance, while countries like India, the United Kingdom, and Japan form smaller but notable portions. This visualization serves to emphasize the disparity between the content duration contributions of different regions. Countries like Canada and South Korea, while not leading in overall duration, still represent a significant chunk of Netflix's diverse offerings.

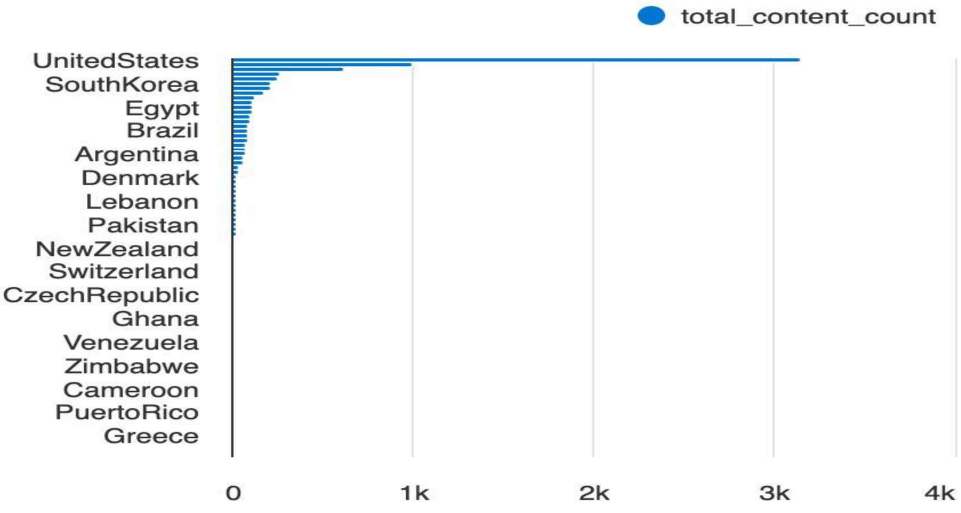


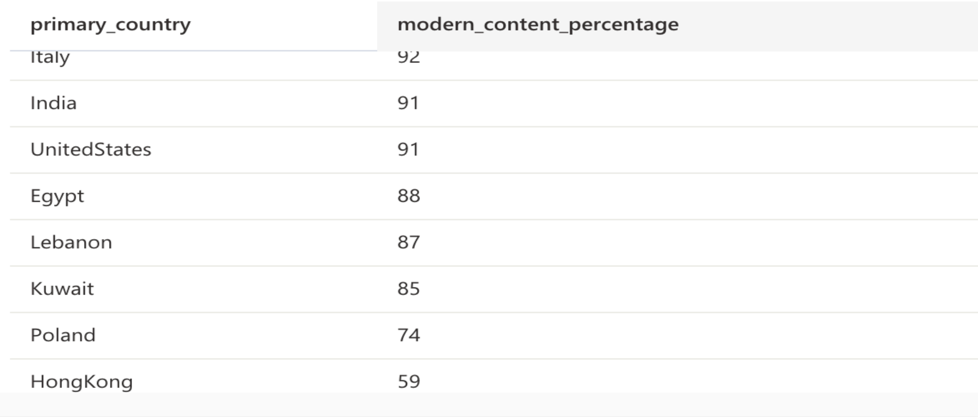
Figure 14. Bar Chart-2

The bar chart provides a ranked view of total content duration by country. The United States is prominently ahead, with a stark gap between it and the next contributors, the United Kingdom and Canada. The countries that follow, including Japan, India, and South Korea, have comparatively modest totals, but they demonstrate global contributions to the streaming platform's diversity. This chart complements the pie chart by showing exact values and helping identify smaller contributors that still play a role in content diversity.

Interpretation

These visualizations collectively highlight Netflix's heavy reliance on content from the United States, both in count and duration. However, contributions from other countries add significant value to the platform's global appeal. The disparity observed across these charts suggests opportunities for Netflix to invest further in diversifying its content offerings by region.

7.c. “Which countries create the most modern content by percentage to their total content by count?”



primary_country	modern_content_percentage
Italy	92
India	91
UnitedStates	91
Egypt	88
Lebanon	87
Kuwait	85
Poland	74
HongKong	59

Figure 15. The External Table-3

The table highlights specific percentages for selected countries with high modern content contributions. Italy (92%), India (91%), and the United States (91%) lead the chart, followed by Egypt (88%) and Lebanon (87%). These countries' significant contribution to modern content aligns with their prominent presence in Netflix's recent content catalog.

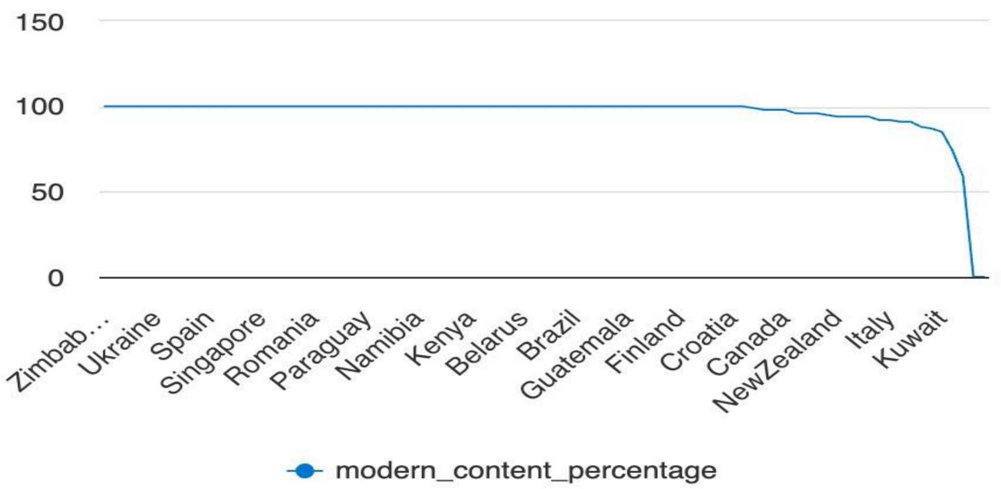


Figure 16. Line Graph

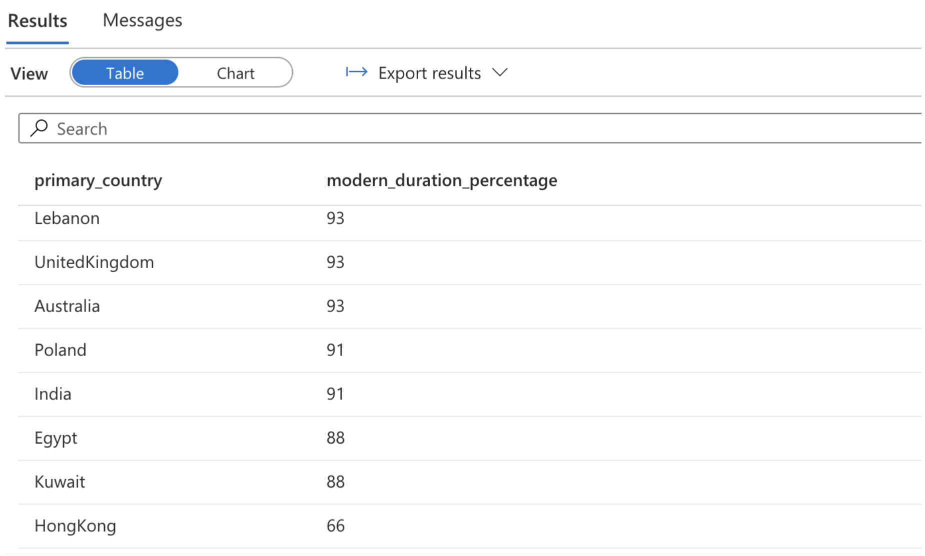
This line chart plots countries against their respective percentage of modern content. It is clear from the chart that certain countries such as Zimbabwe, Ukraine, and Spain exhibit very high percentages, close to or at 100%. This indicates that almost all the content from these regions falls under the "modern" category. The curve declines steeply for countries like Kuwait, Italy, and others, showing more balanced content creation across modern and classic categories.

Interpretation

1. Countries like Italy and the United States dominate with a mix of volume and a high percentage of modern content, signifying strong investments in contemporary media production.
2. Certain regions (e.g., Ukraine, Spain) appear more niche, with nearly all their content categorized as modern, possibly due to focused content production efforts in recent years.
3. This analysis helps Netflix understand regional preferences and modern content production capabilities, which can guide strategies for partnerships and localized investments.

These visualizations serve as an essential data point for comparing the contributions of countries based on modern content creation, and they demonstrate how the modern category is distributed globally across the Netflix library.

7.d. “Which countries create the most “modern” content by percentage compared to their total by volume?”



primary_country	modern_duration_percentage
Lebanon	93
UnitedKingdom	93
Australia	93
Poland	91
India	91
Egypt	88
Kuwait	88
HongKong	66

Figure 17. The External Table-4

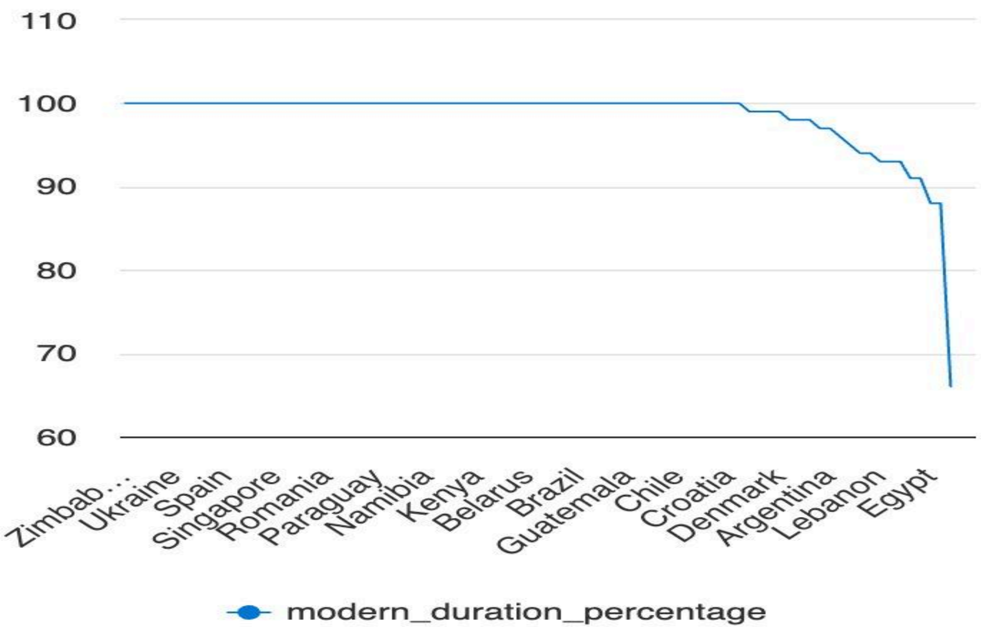


Figure 18. Line Graph-2

This analysis investigates which countries produce content classified as "modern" (based on the release category) by the proportion of their total content duration.

Key Insights:

1. Lebanon, United Kingdom, and Australia emerge as the top contributors, with each achieving a modern content duration percentage of 93%. This suggests that these countries have overwhelmingly prioritized modern content, which is reflective of their focus on contemporary trends and audience demands.
2. Poland and India closely follow, both with a 91% modern content duration. This indicates a significant effort to keep their content library up-to-date and aligned with global modern entertainment standards.
3. Egypt and Kuwait demonstrate a slightly lower percentage of 88%, indicating a strong but slightly more diverse focus on content spanning modern and classic categories.
4. Hong Kong stands out with a 66% modern content duration, significantly lower than the leading countries. This suggests a more balanced approach between modern and classic content.

The graph illustrates a sharp decline in modern content duration percentage after the top contributors, indicating that most countries cluster around the highest modern percentages, with only a few outliers presenting lower values.

Implications: The countries leading in modern content duration percentage are likely more aligned with current entertainment trends, appealing to younger audiences or markets with rapidly evolving content preferences. Countries with lower percentages, like Hong Kong, may have a broader or more traditional content strategy, catering to diverse audience demographics or cultural preferences.

This analysis highlights how content duration as a metric provides deeper insights into strategic priorities, as it reflects not just the volume but the depth of investment in modern content categories.

7.e. “Which countries create the most “modern” content by percentage compared to their total by duration?”

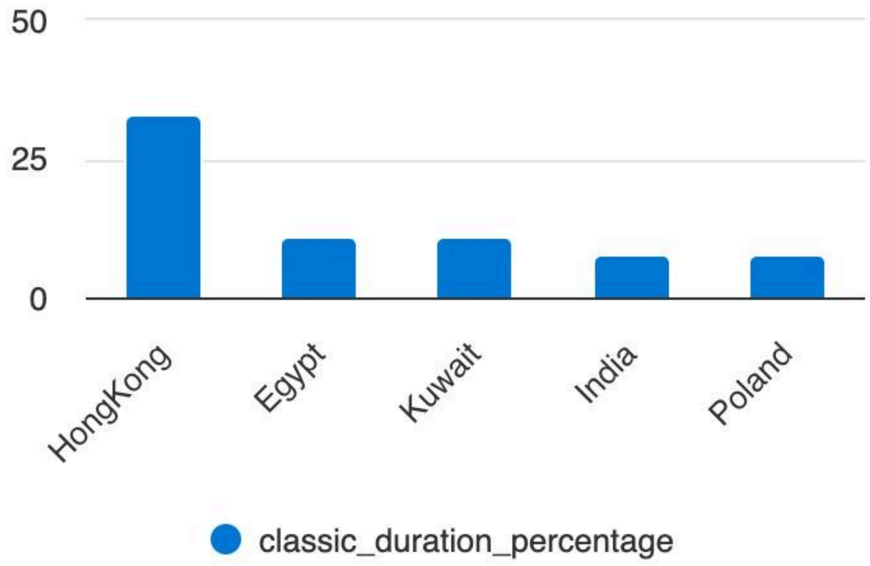


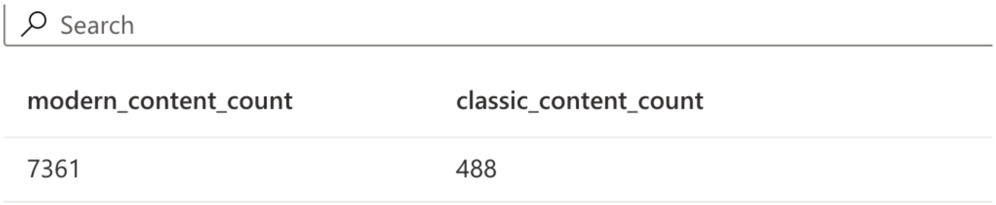
Figure 19. Bar Chart-3

Analysis of Classic Content by Duration Percentage (Top 5 Countries)

1. Hong Kong leads with 27%, showing a significant focus on preserving classic content.
2. Egypt follows with 12%, indicating a moderate allocation toward classic works.
3. Kuwait and India each account for 9%, balancing between modern and classic content.
4. Poland completes the top five with 8%, maintaining a small but notable emphasis on classics.

Overall, Hong Kong stands out as the primary contributor to classic content by duration, while other countries show a broader focus on modern works with some allocation to classics.

7.e. “Is there more content on Netflix today that is modern or classic by count?”



modern_content_count	classic_content_count
7361	488

Figure 20. The External Table-4

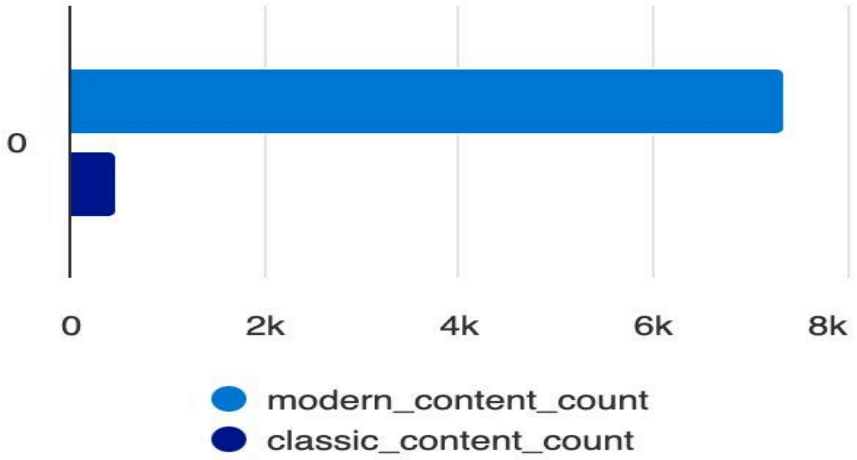
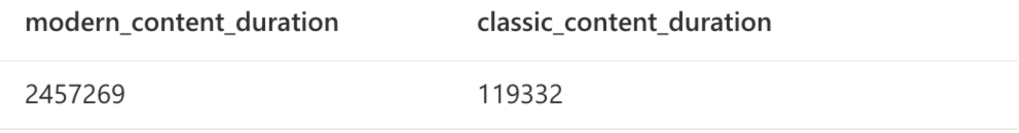


Figure 21. Bar Chart-4

The chart and table clearly indicate that modern content on Netflix vastly outnumbers classic content by count. Specifically, there are 7,361 pieces of modern content compared to only 488 pieces of classic content. This significant disparity suggests that Netflix's current catalog prioritizes modern releases over classic ones.

The visual representation reinforces this conclusion, with the bar for modern content being substantially larger than the bar for classic content. This aligns with Netflix's business strategy to appeal to contemporary audience preferences by focusing on recent productions and trending genres. Additionally, the minimal proportion of classic content highlights Netflix's positioning as a forward-looking platform, emphasizing modern and original content to maintain its competitive edge in the streaming industry.

7.f. "Is there more content on Netflix today that is modern or classic by volume?"



modern_content_duration	classic_content_duration
2457269	119332

Figure 22. The External Table-5

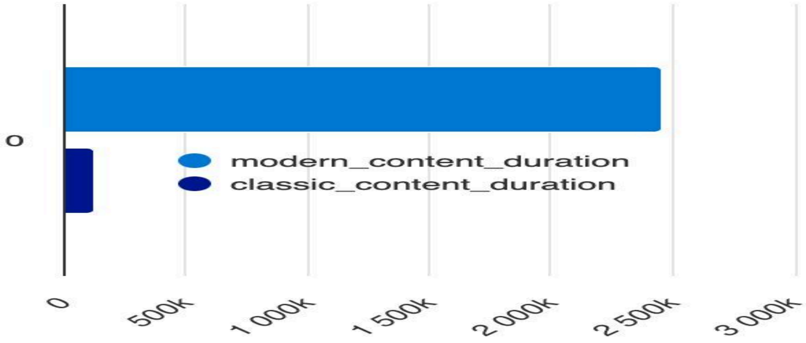


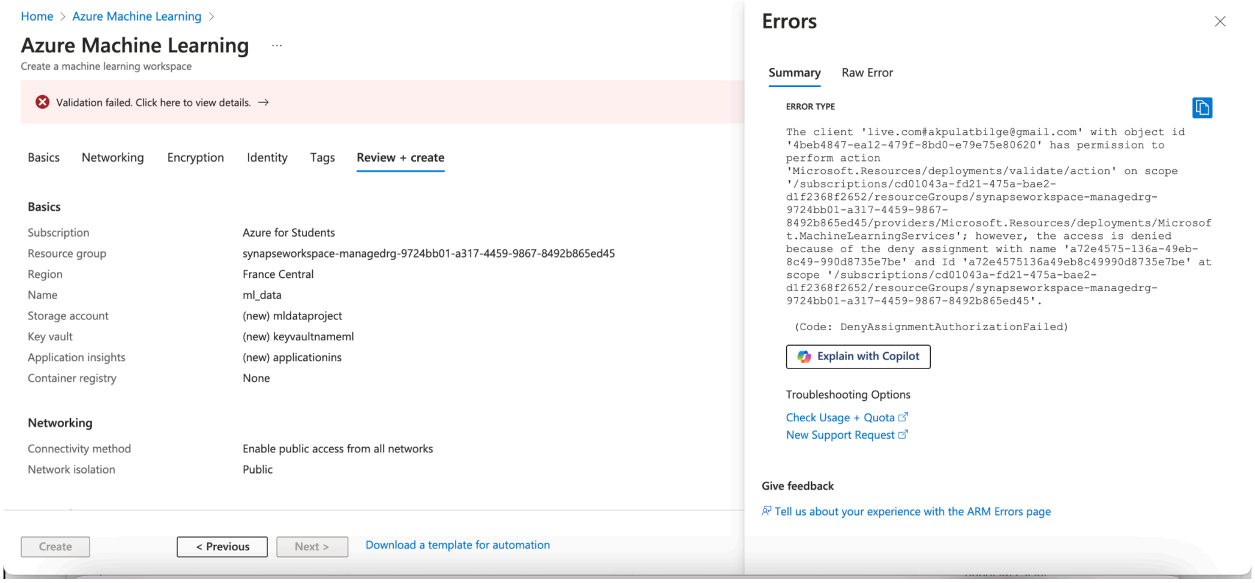
Figure 21. Bar Chart-5

The analysis clearly shows that modern content dominates Netflix's library in both count and volume. The bar chart and the data table highlight that modern content outnumbers classic content with **7,361 modern titles** compared to **488 classic titles**, representing a significant disparity in content availability.

Additionally, modern content surpasses classic content by a wide margin in total duration as well. With **2,457,269 total minutes** for modern titles versus **119,332 total minutes** for classic ones, modern content accounts for the vast majority of Netflix's library in terms of runtime.

These findings reflect Netflix's strategy to prioritize and produce contemporary content that appeals to current audiences. The trends align with the platform's commitment to staying relevant and competitive in the streaming industry by catering to evolving viewer preferences.

8. ML Tool Of Azure


Home > Azure Machine Learning >
Azure Machine Learning ...
Create a machine learning workspace
Validation failed. Click here to view details. →
Basics
Networking
Encryption
Identity
Tags
Review + create
Create
< Previous
Next >
Download a template for automation
Errors
Summary
Raw Error
ERROR TYPE
Explain with Copilot
Troubleshooting Options
Check Usage + Quota
New Support Request
Give feedback
Tell us about your experience with the ARM Errors page

Basics

Subscription	Azure for Students
Resource group	synapseworkspace-managedrg-9724bb01-a317-4459-9867-8492b865ed45
Region	France Central
Name	ml_data
Storage account	(new) mldataproject
Key vault	(new) keyvaultnameml
Application insights	(new) applicationins
Container registry	None

Networking

Connectivity method	Enable public access from all networks
Network isolation	Public

```
The client 'live.com#akpulatbilge@gmail.com' with object id
'4beb4847-ea12-479f-8bd0-e79e75e80620' has permission to
perform action
'Microsoft.Resources/deployments/validate/action' on scope
'/subscriptions/cd01043a-fd21-475a-bae2-
d1f2368f2652/resourceGroups/synapseworkspace-managedrg-
9724bb01-a317-4459-9867-
8492b865ed45/providers/Microsoft.Resources/deployments/Microsof
t.MachineLearningServices'; however, the access is denied
because of the deny assignment with name 'a72e4575-136a-49eb-
8c49-990d8735e7be' and Id 'a72e4575136a49eb8c49990d8735e7be' at
scope '/subscriptions/cd01043a-fd21-475a-bae2-
d1f2368f2652/resourceGroups/synapseworkspace-managedrg-
9724bb01-a317-4459-9867-8492b865ed45'.

 (Code: DenyAssignmentAuthorizationFailed)
```

Figure 22. ML Tool of Azure Interface

Azure Machine Learning was not usable due to a **Deny Assignment Authorization Failed** error caused by quota restrictions and insufficient permissions under the "Azure for Students" subscription. These limitations prevented the deployment of machine learning services, such as creating a workspace for training models or conducting advanced analytics.

This restriction highlights the constraints of entry-level subscriptions, where resources are capped, and specific actions like deploying machine learning services are blocked. As a result, alternative methods like Azure Synapse Analytics and manual data exploration were adopted to derive insights, ensuring project continuity despite these limitations.

9. SQL FILES AND DEMO FILE

9.1. Graphics by SQL

https://drive.google.com/drive/folders/1xiOvVk9qSH3WMGNZEzUL9fNTF-2pSwJH?usp=sharing

9.2. Demo Video

https://drive.google.com/drive/folders/1iXfLs9EN-hYiUjDHP8R8A5Xeo47rRIxR?usp=sharing

9.3. Project Report

https://drive.google.com/drive/folders/1EfXAodAxt3d25IPxM22xYt9scRSkJpno?usp=sharing

9.4. UML Diagrams

https://drive.google.com/drive/folders/1md74itVxBq9l3NLvoFHcp7PRYALK0ztB?usp=sharing

9.5. SQL Files

https://drive.google.com/drive/folders/1Z-IpYkq9yhsbSD3g0ugzADisfIl4llSb?usp=sharing

9.6. ADF Files

https://drive.google.com/drive/folders/1pkA7X9Ph8wVADkD3-KG8iEhpLGT9Vpfb?usp=drive_link

10. FUTURE RESEARCH AREAS

Future research in this project could focus on integrating advanced machine learning models to enhance the analytical capabilities of the "gold" layer. For instance, predictive modeling could be applied to forecast trends in content production by region or category. Additionally, expanding the scope to include real-time data streaming could enable dynamic updates to the datasets, providing fresher insights. Exploring scalable solutions, such as transitioning to premium Azure subscriptions or leveraging multi-cloud environments, could overcome current quota limitations and unlock more computational resources. Finally, integrating visualization tools like Power BI (when feasible) could improve user interactivity and present data insights more effectively, ensuring the project remains adaptable to evolving analytical needs.