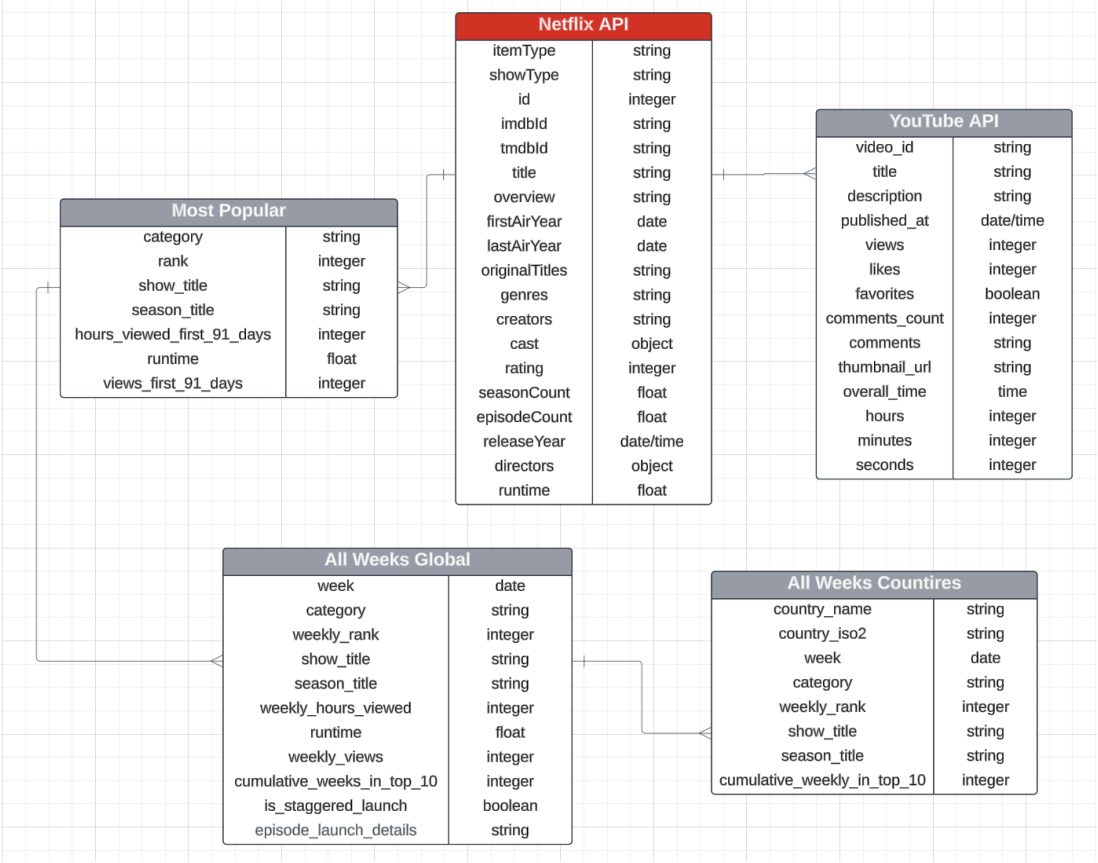
BA882 Project Deliverable 1

Data Model *(ERD Diagram on Lucid Charts)*

Our project aims to provide insights that could inform Netflix's content acquisition, production, and marketing strategies, ultimately helping to optimize viewer engagement and subscription growth. To achieve this, we extracted multidimensional data related to movies and shows on Netflix from three different sources: the Netflix Top 10 weekly data website, the YouTube Data API v3, and the Streaming Availability API. We then ingested this data into five tables that we designed.



Data Pipeline *(Cloud Run Functions on GCP)*

To facilitate debugging during the initial phase of our pipeline construction, we opted to create separate pipelines for our three different data sources. At this point, we have confirmed that all pipelines are functioning correctly. In the next phase, we will explore the possibility of merging them into a single pipeline to enhance efficiency.

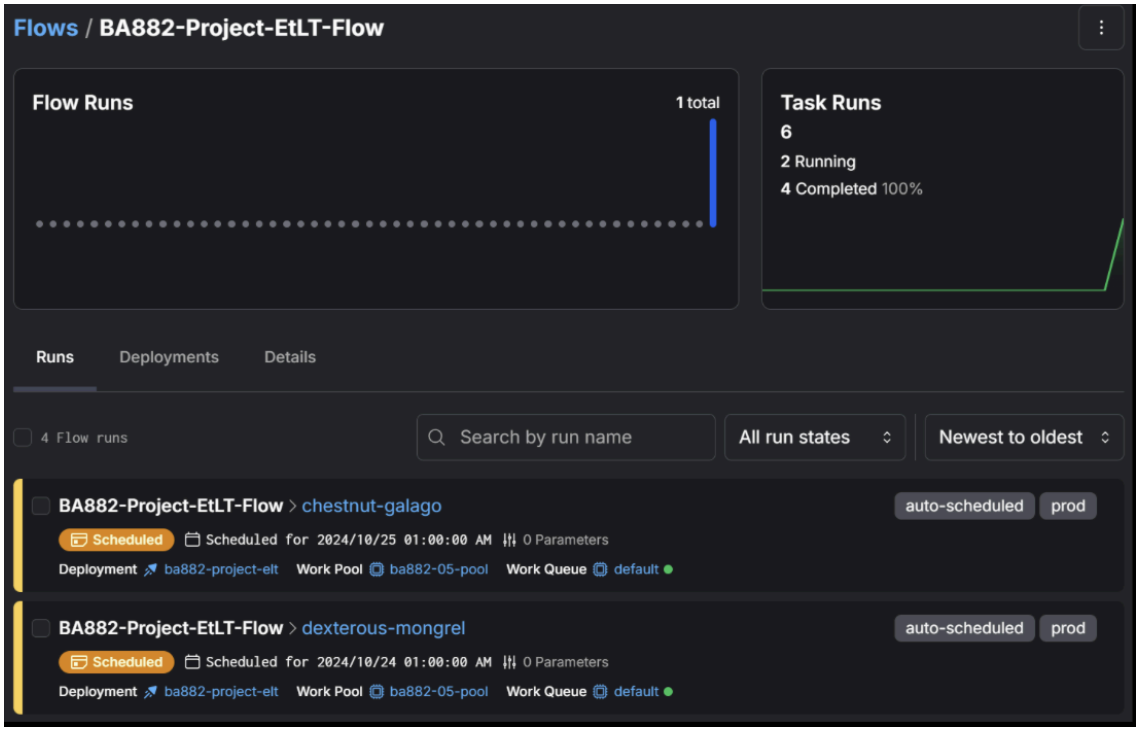
For the different pipelines, our designs vary. For the video data from the Netflix channel on YouTube, we chose the ETLT (Extract, Transform, Load, Transform) approach. This

data includes descriptions with unrelated Netflix promotional content. After extracting the data, we first transformed it to remove this unnecessary information. We also converted the video length from the “PT2M5S” format to a standard time format, separating hours, minutes, and seconds. Once this data was loaded into the tables in MotherDuck, we further transformed the title column to keep only the actual title of each show for better connections between the tables.

For the Netflix Weekly Top 10 data, we opted for the ELT (extract, load, transform) approach. We used a single script to extract data from three different links on the website: the Weekly Top 10 lists of the most-watched TV shows and films, the Weekly Top 10 lists of the most-watched TV shows and films in various countries, and the Top 10 most popular titles in each category based on views within their first 91 days on Netflix. We loaded the data from GCS into the MotherDuck tables and performed various transformations based on the requirements for later analysis.

For the Netflix show data extracted from the Streaming Availability API, we employed an ELT (Extract, Load, Transform) design. After extracting the data, we loaded it directly into MotherDuck for further transformation and analysis.

Data Orchestration *(Scheduling Flows on Prefect)*



Cloud-based Reporting *(Dashboarding on Superset)*

In our project, we used MotherDuck as the data warehouse and Apache Superset to create visualizations. By linking MotherDuck with Apache Superset, we ensured seamless data flow and real-time updates in visualizations. This setup allowed us to derive meaningful insights from our data stored in the cloud data warehouse. The dashboard below provides some great insights from the Netflix and YouTube databases.

The dashboard offers a clear snapshot of how Netflix content is performing and how viewers are engaging with it. It features a line graph that tracks the weekly popularity of shows like “Bridgerton” and “Avatar: The Last Airbender,” revealing trends in viewer interest. A treemap highlights the most liked genres on YouTube, with Drama and Comedy at the forefront. The bar chart showcases top shows by engagement, with “Bridgerton” leading in views and likes. A histogram shows that shorter videos are more common, aligning with viewer preferences for concise content. Lastly, a line graph compares viewership across categories, showing that English TV content captures the most viewing hours, underscoring its strong audience appeal.

Looking Ahead

Looking ahead, we can enhance this pipeline setup by incorporating machine learning operations (MLOps) to boost predictive analytics and refine content recommendations. Future developments may also focus on gaining more detailed insights into regional viewing habits and integrating additional data sources. This will help improve strategic decision-making for content acquisition and marketing strategies.