## **The Business/Organization & Opportunity**

Netflix is a subscription-based streaming service that allows its members to watch tv shows and movies without commercials on an internet-connected device (ref. netflix.com). There is an opportunity to optimize the selection of tv shows and movies for the business to increase member subscriptions and/or stock prices by analyzing the current lists of shows/movies, list of Netflix original shows/movies, the IMDB scores of the shows/movies and the stock prices data.

## 

### **Problem**

The system helps us to find out what the most popular genre of movies/shows are in the current Netflix shows/movies selection; whether the popular videos align with high IMDB scores; whether Netflix originals have higher IMDB scores; and if there is a positive relationship between high score movies/shows to the stock prices.

### **Updated Dimensional Model**

Diagram

Description automatically generated

### **Updated Architecture Diagram**

A picture containing text, orange, kitchen appliance

Description automatically generated

1. (Updated) We choose these data sources for our project:
   1. Netflix original list ([https://www.kaggle.com/abhimanyudasarwar/netflix-originals](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.kaggle.com_abhimanyudasarwar_netflix-2Doriginals&d=DwMFAw&c=dTXc8cCP8suVpClwB1HRHQACHN4UFMgL7MtSjCbKyts&r=uCemPd7JMUYoUMs7kirjWxYabMtp-x_PVYDU_TDDh80&m=WzoiuvOVNGCPFvQgJR8O8ELAcS7LMDxVU23XGLa5sZs&s=WUkCkdEc1snKWQH2lmdvUkz7SMVm2aeku3bUOe8ozF4&e=))
   2. Netflix stock data ([https://www.kaggle.com/aayushmishra1512/netflix-stock-data](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.kaggle.com_aayushmishra1512_netflix-2Dstock-2Ddata&d=DwMFAw&c=dTXc8cCP8suVpClwB1HRHQACHN4UFMgL7MtSjCbKyts&r=uCemPd7JMUYoUMs7kirjWxYabMtp-x_PVYDU_TDDh80&m=WzoiuvOVNGCPFvQgJR8O8ELAcS7LMDxVU23XGLa5sZs&s=9j8a84Xq_KYb7szekO0BmZVNcza5z_QpI6_MedHa2Uo&e=))
   3. IMDB scores ([https://www.kaggle.com/sarahjeeeze/imdbfile](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.kaggle.com_sarahjeeeze_imdbfile&d=DwMFAw&c=dTXc8cCP8suVpClwB1HRHQACHN4UFMgL7MtSjCbKyts&r=uCemPd7JMUYoUMs7kirjWxYabMtp-x_PVYDU_TDDh80&m=WzoiuvOVNGCPFvQgJR8O8ELAcS7LMDxVU23XGLa5sZs&s=Exi9rbBgmhUuX5b9gPyRayOsmmWgaVSAvjsBpEtjbIU&e=)

The Netflix movies/shows list has identical information to the IMDB scores file so it is removed from the original data sources list.

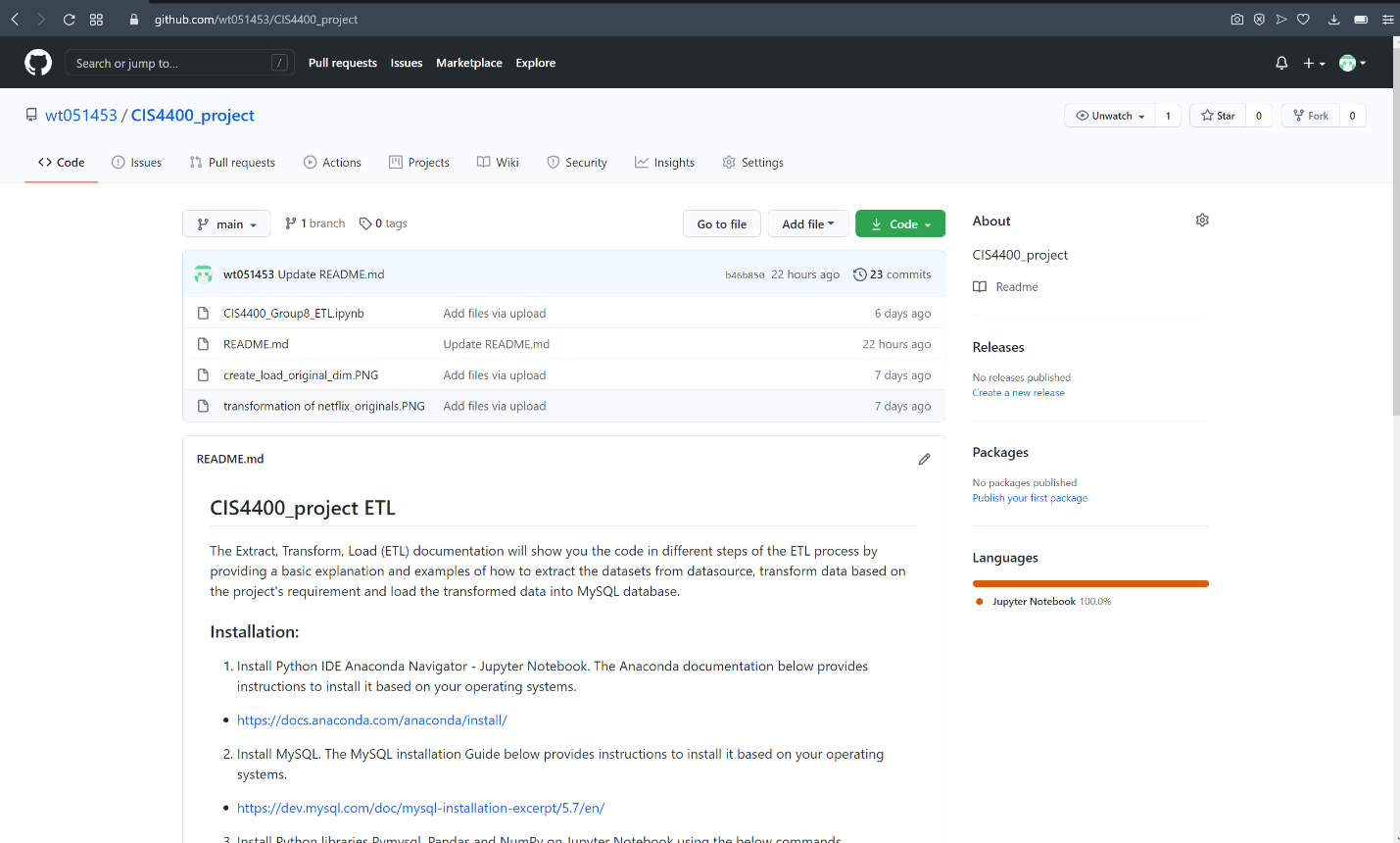
1. (Updated) We import the data by downloading the flat files to our computers, then we use python scripts run on our laptops to clean the data, merge data, and upload the merged (transformed) data into our data warehouse.
2. (Updated) We run our data warehouse on MySQL server for this project.
3. We use Tableau to build a dashboard of graphs, which query our MySQL instance, to visualize our KPIs.

### **Updated Detailed Design**

The datasets we found from Kaggle.com allow us to analyze and study the relationships between Netflix’s movies and shows selection to IMDB scores and their stock prices. For ETL steps, we chose to use python because it is one of the popular general-purpose programming languages that is capable of performing the tasks and the team is familiar with. We chose to run our data warehouse on MySQL server because it is a popular relational database management system that provides secure and fast data storage, fast query speed and the team is also familiar with it. Finally, we chose Tableau to create visualizations because it is a powerful visualization tool that offers ease of use and many visualization possibilities. Any future members to the team may already have experiences using these technologies or can easily pick up how to use them because there are many instructions or documentations available for these technologies which will significantly reduce the training time and cost.

**Technical documentation of the ETL processes**

The link to the technical documentation is:<https://github.com/wt051453/CIS4400_project>



**Final schema**

The final schema for our project is a Star schema which contains two fact tables and four dimensional tables.

Graphical user interface, text, application, email

Description automatically generated

**Graphical user interface, text, application

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Dashboard application**

Use MySQL workbench to monitor performance

Graphical user interface, application

Description automatically generated

**Connect Tableau to MySQL**

A picture containing table

Description automatically generated

The analytics/visualization in the dashboard help answer the questions in the Problem section.

What the most popular genre of movies/shows are in the current Netflix shows/movies selection and whether the popular videos align with high IMDB scores? The answer is no, the most popular shows on the left don’t always correspond to high IMDB scores on the right.

Chart

Description automatically generated

Settings to create the visualization: from facts\_imdb\_rating table, I put the ‘Listed in’ to Rows as ‘Dimension’ and put Counts of ‘Listed in’ and Average of ‘Rating’ to the Columns and finally displayed the data in bar graph.

Do Netflix originals have higher IMDB scores? The answer is yes. The originals have slightly lower max score (9.5 vs 9.7) than not originals but have higher 75th percentile, median, 25th percentile and mix score than not originals.

Graphical user interface, application

Description automatically generated

Settings to create the visualization: from facts\_imdb\_rating table, I put the ‘Rating’ to the Rows and put ‘Original Title’ as the filters (selecting all except null) to create the boxplot.

Graphical user interface, application

Description automatically generated

Settings to create the visualization: from facts\_imdb\_rating table, I put the ‘Rating’ to the Rows and put ‘Title’ (selecting all) and ‘Original Title’ (selecting null only) as the filters so the data only reflects non Netflix originals to create the boxplot.

Is there a positive relationship between high score movies/shows to the stock prices? The answer is no, there is no positive correlation between the average IMDB scores to stock prices by year.

Chart, bar chart

Description automatically generated

Settings to create the visualization: from facts\_imdb\_rating table, I put the Average ‘Rating’ to the Rows, from facts\_stock\_prices table, I put the Sum ‘Adj Close’ to the Rows, and from date\_dim table, I put ‘Year’ to the Columns to create the dual Y-axis graph to show average rating and total of adjusted closing prices by year.

A map to show the average IMDB scores by country.

Map

Description automatically generated

Settings to create visualization: I put the Tableau auto generated Longitude and Latitude to Columns and Rows respectively and put Average ‘Rating’ in 'Color', Counts ‘Rating’ (facts\_imdb\_rating) and Country (show\_dim) as ‘Dimension’ in the Detail to display the world map showing average IMDB rating by country.

**Conclusion**

The analytics built in Tableau using data transformed by the data source helped answer the problems we identified during the planning section and learned from data analysis that there are room to improve the movie/show selections in genres that received high IMDB rating, produce more high quality Netflix original movies/shows and increase movies/shows from countries that have higher than average IMDB ratings to attract more and retain existing subscriptions.

During this assignment, I got the opportunity to work on tasks that are practical and identical to real life project and it was a tremendous and valuable experience. I have learned how to figure out what directions to seek answers when I was stuck, effectively ask and follow-up with questions, perform researches (mostly online) using keywords and quickly acquired the necessary skillsets to solve various problems I ran into. I believe this will be beneficial for me as a Data Analyst or any other positions in the future. If I had to do it all over again, I would still be proactively getting involved in as much of the project as I could, working diligently with and motivating all the stakeholders.