BME AUT Automatizálási és
Alkalmazott
Informatikai Tanszék

Homework specification

Business Intelligence

2022 Fall Semester

NetView

Rita Matos Maranhão Peixoto – IKD60O

rita.peixoto2000@gmail.com

MŰEGYETEM 1782

Introduction

In the current days, Netflix is the most known subscription streaming service. The way the company used its data had a significant role in the remarkable growth that it experienced. It focuses not only on gaining more subscribers but also on improving the subscribers' experience. The focus of this system is to analyse a portion of Netflix's Business verticals: the current competition, the subscription, and the revenue. This aims to evaluate the areas where it is losing profit potential and how to work on them.

How to use and run the system

To run the system developed, all the instructions and files used are in a GitHub repository available at this link. It includes all datasets, files from Pentaho and Tableau as well as the code developed, the configuration of the PostgreSQL data warehouse, all images used in this report and the ReadMe file with the information on how to run and use the system.

Data

The system developed access to six datasets:

- Competition data: list of movies available on various streaming platforms;
- Amazon Prime: listings of all the movies and tv shows available on Amazon Prime;
- Disney +: listings of all the movies and tv shows available on Disney+;
- Netflix: listings of all the movies and tv shows available on Netflix;
- Subscription price: Netflix monthly subscription fees in different countries;
- Netflix subscribers and revenue by country: Netflix's subscription figures and Netflix's revenue ($) in four different regions.

The data sources are all **CSV** files, during the transformation stage are kept in a **staging area** and only after successfully running the ETL tests are these transformed into a relational database and loaded in the data destination – a **data warehouse**, PostgreSQL. This data warehouse is running locally using a Docker container. The data was stored according to the following schema:

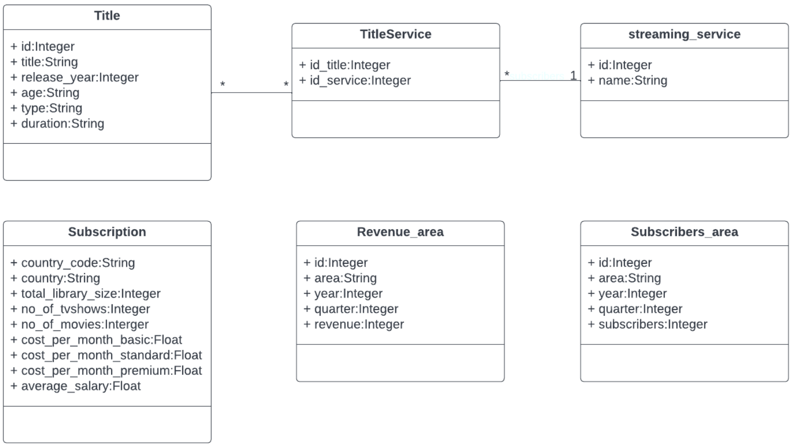

Figure 1: Database schema

ETL

The ETL process contains six jobs and four significant transformations. The jobs are:

- Set up the database: clear the database and create the database tables according to the schema provided;
- Title: populates the database with the titles in the existing datasets;
- Subscription: populates the database with information about the subscriptions;
- Subscribers: populates the database with information about the subscribers per area;
- Revenue: populates the database with information about the revenue per area.
- Complete Pipeline: populates the database with information about the titles, subscriptions, subscribers, and revenue.

The jobs regarding data always have a complete pipeline that sets a variable for the project's path, checks if all datasets will be used exist, check the connection to the data warehouse, cleans the database and creates its schema and then applies all the transformations. These independent jobs were created because if any change occurs in a specific dataset, only the jobs that use it must be rerun.

The initial datasets are stored locally in CSV files. ETL tasks were mainly performed using Pentaho. However, before that, the dataset with the competition data had to be arranged to fix some of its issues that did not allow it to be loaded in Pentaho because it contained both the separator and the enclosure on one of its fields. To solve this, an IPython Notebook was created to perform this task; after this task, the dataset was ready to be loaded into Pentaho.

In addition, another IPython Notebook was created to perform web scraping of the average salaries of the countries. This was accomplished by accessing this website.

In Pentaho, four major transformations were created, where many different tasks were implemented to get the input datasets in the state needed for the presentation layer:

- On the transformation of the titles, all the datasets about the competition are extracted and used. Amazon Prime, Netflix, and Disney+ all have the same structure; therefore, similar tasks were performed by all of them, which included: removing unnecessary characters on ids, renaming columns to better suiting and common names, and creating a field that indicated the service of the title. After that, these tables were merged, having to be chosen the correct fields to keep, eliminating unused ones and renaming the new fields. These datasets were then combined with the competition data, which had a transformation that included removing unnecessary characters, renaming columns to better suiting names, and standardising the format of type and age fields. After the merge, many tasks were performed to achieve the final goal:
    - Decide on what values to keep.
    - Remove unused fields.
    - Rename new ones.
    - Transform all services in one-hot encoding and once again remove unused fields and rename new ones.
    - Standardise the age field so age ratings with the same meaning have the same value.
    - Add the id field.
    - Normalise row so we have a row for each title/service combination and remove the row without value.

Finally, the fields were chosen in a way that only fields with enough information and meaningful for presentation were kept. When merging this table, many fields from the Amazon Prime type of dataset were null because of the merge with the competition data, however, since these were not crucial for the presentation, they were not kept. Three tables were created, one for the titles, one for the streaming services and an association table of both these tables. The data was loaded in the data warehouse, PostgreSQL. All these tasks can be seen in Figure2.

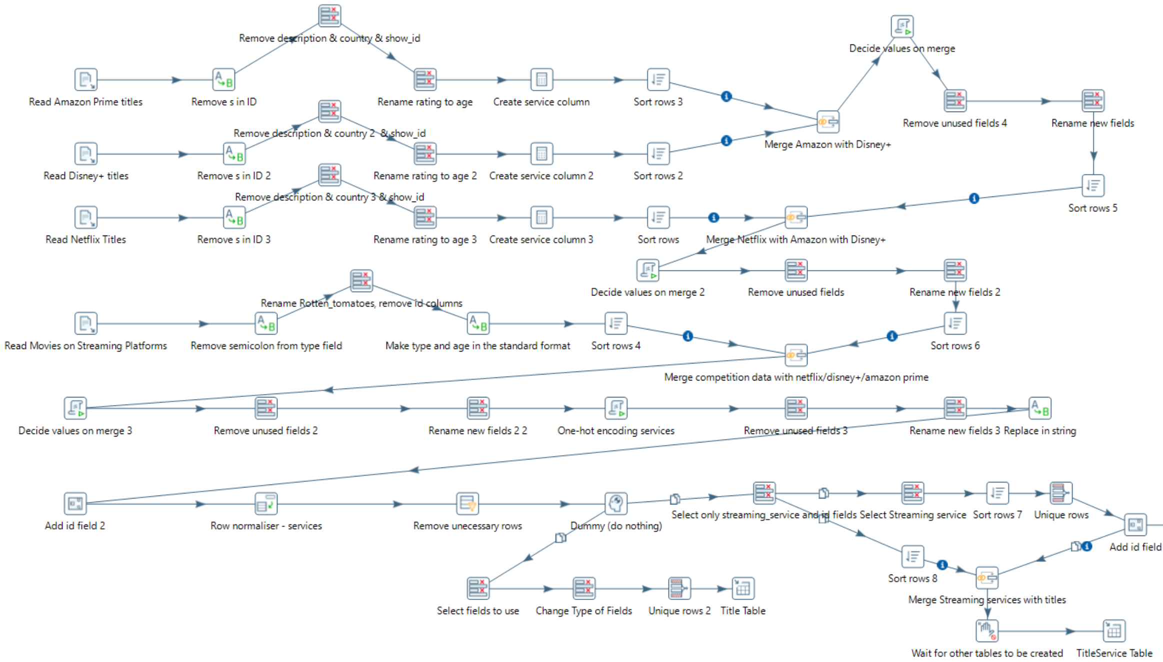

Figure 2:Title's Transformation

- The datasets of countries' average salary and subscription fees are extracted on the subscription transformation. On the countries' average wage, an update was made on the name of the Czech Republic that became Czechia. After that, the two datasets are merged, and the rows are filtered only to keep the countries with existing information. Some countries' information was not scraped because they are small countries, and the information is not readily available. Still, a decision was made to keep them and manually add the country's average salary field value. Unnecessary fields were removed, and the new ones were renamed. Finally, this dataset was loaded in the data warehouse in the *Subscription* Table. This sequence of transformation tasks can be seen in Figure 3.

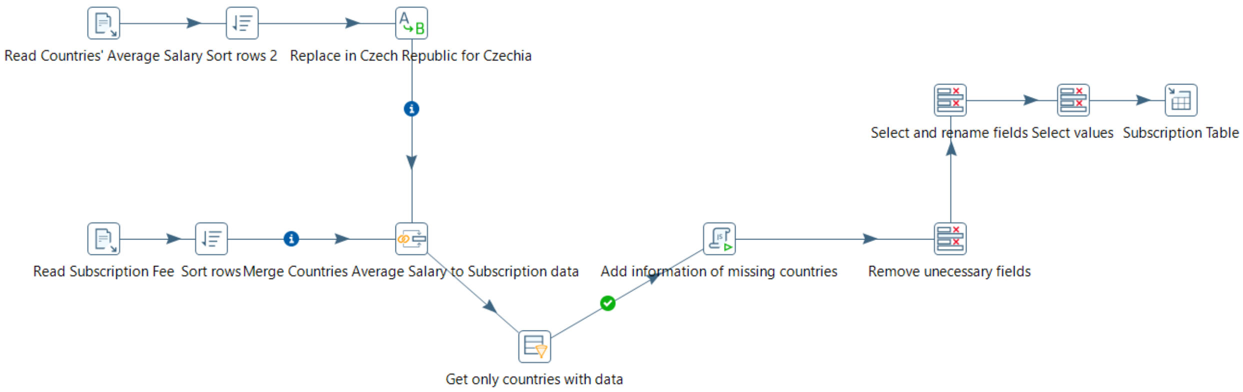
Read Countries' Average Salary
Sort rows 2
Replace in Czech Republic for Czechia
Select and rename fields
Select values
Subscription Table
Read Subscription Fee
Sort rows
Merge Countries Average Salary to Subscription data
Add information of missing countries
Remove unecessary fields
Get only countries with data

Figure 3: Subscription's Transformation

- On the revenue transformation, the rows are normalized in a way that we have a value for each quarter and year combination instead of these being columns. Then, the quarter and year are extracted from the variable name, the unnecessary fields are removed, and the new ones are renamed. An id was added to each row, and finally, the data was loaded in the *Revenue_area* table, as seen in Figure 4.

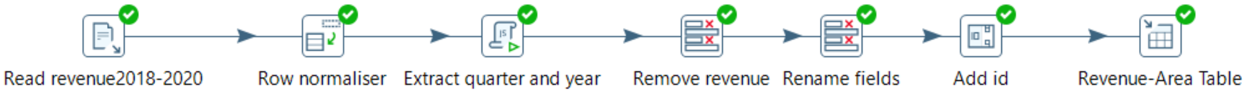

Figure 4: Revenue's Transformation

- On the subscribers' transformation, as can be seen in the following image, similar modifications to the revenue transformation were performed for this dataset. In the end, the data is stored in the *Subscribers_area* table, as seen in Figure 5.

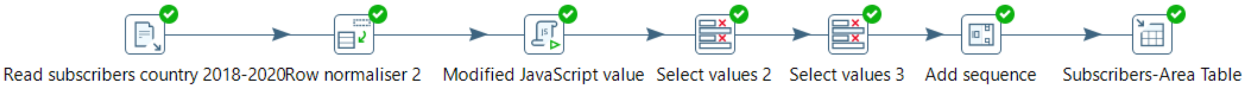

Figure 5: Subscribers Transformation

Presentation

The presentation layer consists of 4 different dashboards, each addressing a single area of analysis. These dashboards contain different visual objects (bar charts, pie charts, maps, etc.) according to what best suits the existing data. This layer was developed using Tableau.

Streaming services offer

This report analyses the competition, comparing the offer from four streaming services: Netflix, Prime Video, Hulu and Disney+. This dashboard aims to access the area where the competition may pose a threat to Netflix and extract the areas where more content is lacking and can help maintain subscribers.

In this, one can see the number of shows available per streaming service, the number of shows per age per streaming service, the years of the shows available per streaming service, the distribution of types of shows, and the dimension comparison per streaming service.

This dashboard allows filtering per streaming service, per type and release year.

Subscription

This report analyses the subscription Netflix in each country. The goal here is to assess if the number of subscribers stays the same because the subscription costs do not fit the country's economic situation, keeping in mind what it offers.

This dashboard shows a coloured map of the average salary of each country, a coloured map with the size of the library available on Netflix in that country, a coloured map with the salary share that Netflix's basic subscription represents and a tree map coloured by the cost per title, which means how much does one film costs on Netflix to the subscriber having in mind the number of shows offered and the basic subscription price. This dashboard allows filter by country and by cost per title.

Revenue

This report analysis the revenue (in USD $) of Netflix from 2018 to 2020 divided in four areas: "Asia-Pacific", "Europe, Middle East and Africa", "Latin America" and "United States and Canada". The goal here is to try to justify the evolution of the revenue and see if it there any way we can tackle it to be greater.

In the dashboard the viewer can see the revenue analysis per year; the revenue analysis per quarter of the year; revenue analysis per year per area and the revenue analysis per area per quarter of the year. This dashboard allows to filter per area, per year and per quarter of the year.

Subscribers

This report analyses the revenue (in USD $) of Netflix from 2018 to 2020, divided into four areas: "Asia-Pacific", "Europe, Middle East and Africa", "Latin America", and "The United States and Canada". The goal here is to try to justify the evolution of the revenue and see if there is any way we can tackle it to be more significant.

In the dashboard, the viewer can see the revenue analysis per year, the revenue analysis per quarter of the year, the revenue analysis per year per area and the revenue analysis per area per quarter of the year. This dashboard allows filtering per area, year, and a quarter of the year.

All visualization are published in Tableau Public, available at this link.

Analysis

From the analysis of the visual elements, many conclusions were derived. In this section they will be explored.

Competition

- Prime Video offers more shows than Netflix. Even though these values are not actual, they remain accurate and even more significant.
- Netflix has a small offer on old movies.
- Netflix has a significant share of movies rated at 17+, being almost one-third of the total offer, which diminishes the offer to teenagers and kids that can result in fewer family subscriptions. On the other hand, Prime Video offers around the same amount for 18+ as 13+, which tackles this gap that Netflix has.
- Netflix's movies share is slightly bigger than Prime Video, but it seems insignificant for further analysis.
- At the moment, Disney+ and Hulu don't pose a significant threat has they have a much smaller offer and diversity. However, they can do better on specific audiences.
- Prime Video is the biggest competitor now and has a better offer in some spectrums. A further analysis of how these are affecting Netflix's revenue is desired.

Subscriptions

- Countries in Latin America have a higher salary share as they have a low average salary, and the same goes for Indonesia. Understandably, making such low subscription prices is hard, but this should also be considered.
- Countries like Liechtenstein, Croatia, San Marino and Denmark have a high cost per title compared to countries in similar economic situations. This discrepancy should be tackled to avoid other schemas that can result in less profit for the company.
- An average salary share of 0.8% is still a high rate. This value should be kept closer to 0.5%.
- If Netflix's goal is to increase revenue worldwide, these conclusions should be given relevance as they are a fundamental reason why people don't use subscription-based services.

Subscribers

- Most of Netflix‘s subscribers are from the United States and Canada area.
- The number of subscribers as being growing steadily in the represented years.
- No decline experienced between the quarters means the revenue kept increasing throughout these years.
- The growth of subscribers in Asia-Pacific has been much slower than the rest, while Europe, the Middle East and Africa have been growing more rapidly.
- This may indicate that the Asia-Pacific area needs more attention to increase the number of subscribers. This can be achieved by producing more Netflix originals in these countries' languages and places.

Revenue

- In only two quarters of 2020, more than half of 2019 revenue has been reached.
- The revenue has been increasing over time but very slowly.
- The United States and Canada pose a big part in Netflix's revenue. Still, their revenue growth stagnated at the end of 2019 and the beginning of 2020, which is strange as it was when Covid 19 started, and the other areas experienced a slight growth. This can be explained as Netflix already being well established in this area and gaining more notoriety in the different regions during quarantine.
- The annual revenue growth is expected to keep growing with the bet being made on Netflix Originals.

Final thoughts

It was not possible to make predictions as the temporal data was minimal. Therefore, it was not helpful to make predictions with these.

ETL tests were performed manually before the data load to the data warehouse, so the data in the data warehouse could be trusted to be correct and be used to create meaningful reports.

There were some more datasets available on another spectrum of Netflix’s Business verticals that would produce an even more complete analysis of Netflix’s Business verticals, which is left as future work.