MONASH University

FIT5145 Assignment 4

Business and Data Case Study Report

DATA ANALYTICS AT NETFLIX

(Evolution from a Video Store to a Movie Recommender System)

Ajay Ganapathy

Student ID: 29822270

Email: agan0012@student.monash.edu

CONTENTS

1. PROJECT DESCRIPTION

1.1. PROJECT PURPOSE

This case study shines a light on Netflix which we all know, the popular online movie store. Netflix has made lives easier by creating and suggesting content based on user preferences. But how does Netflix actually uses analytics and data science to be the best movie recommendation engine?

The purpose of this project is to understand the use of analytics at Netflix, how Netflix deals with huge volumes of data and acts as a recommender system for the users based on their preferences, thereby making multi-million dollar decisions

1.2. TRADITIONAL MOVIE BUSINESS

People who watch a lot of movies once used to spend a lot of their time during weekends at the video stores. As soon as a movie is released, they used to pop in at the store, looking for the DVDs of that specific movie and then they used to realize that the latest releases are all sold out. They still used to leave the store with 2 or 3 other movies and enjoy those movies with chips and popcorns because why not? It's a weekend.

1.3. NETFLIX, THE VIDEO STORE

It all started on April 14th, 1998 when a startup in California developed and launched a website known as Netflix.com. They had a different business idea in mind. Instead of people coming all the way to the store, they developed the website in such a way that people were able to search the virtual releases online, rent a movie for a week for few dollars through their website itself and after 2 to 3 working days, an envelope with a DVD tucked used to arrive at people's mailboxes. After a week, people used to return the same packaging via mail.

Fusing the tech of Silicon Valley with every aspect of Hollywood, Netflix became a global TV service and most famous online video store. Having more than 117 million subscribers in every country of the world and a workforce of more than 5500 people,

Netflix generated more than 11$ billion in revenue in the year 2017. Today, Netflix's net revenue is around 135$ billion more than its rivals.

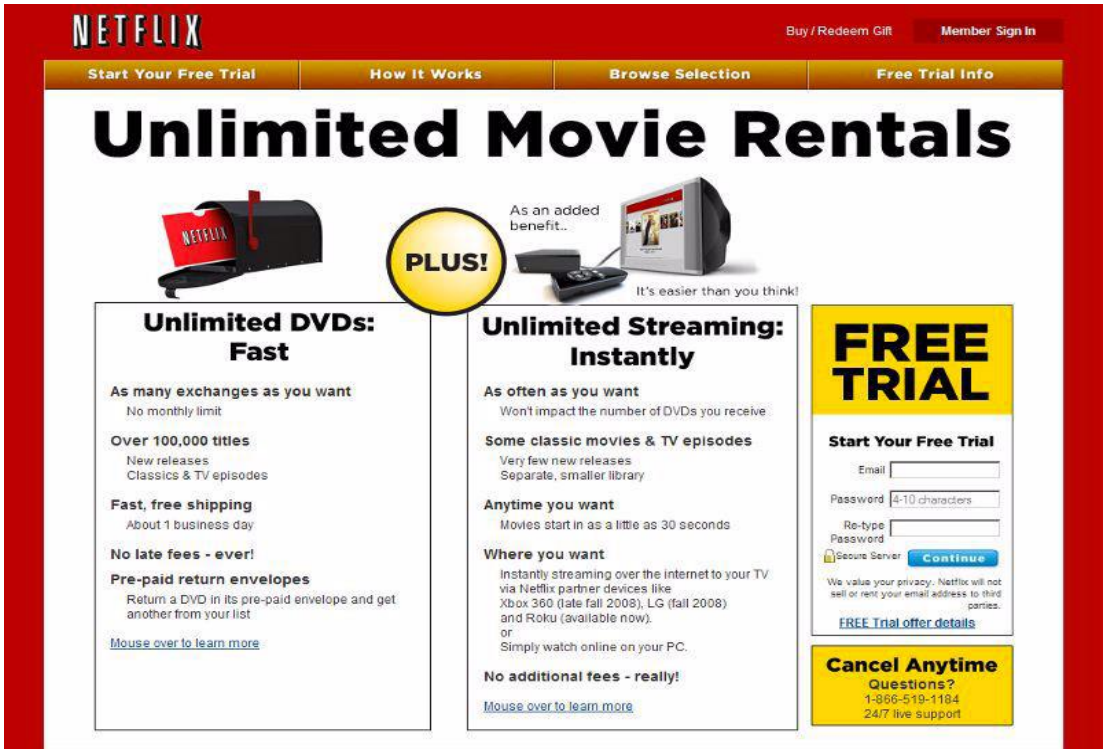

Figure 1 Netflix's Website (1998)

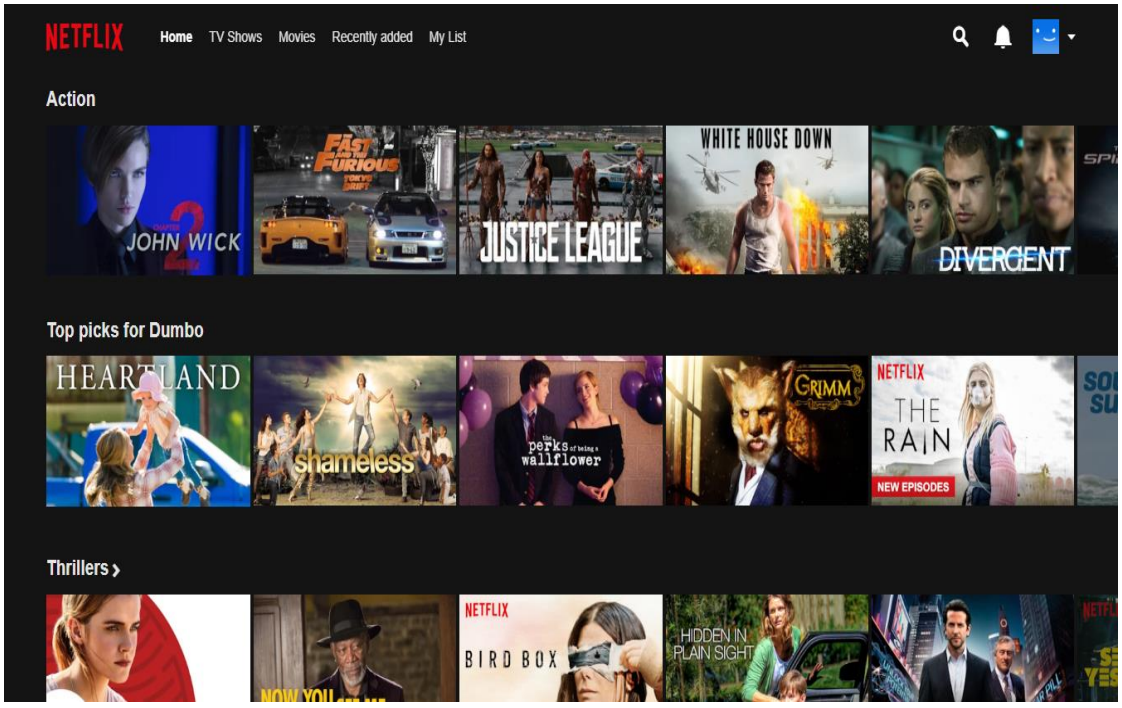

Figure 2 Netflix's Website (2019)

1.4. ANALYTICS AT NETFLIX

The crux of analytics is to provide a base for companies to gain user insights so that they can use these insights and increase the optimization thereby delivering an efficient product.

Having such a large subscriber base of more than 130 million subscribers around the globe, would result in tremendous flow of user data. Netflix uses this data to make decisions which ultimately leads to user satisfaction. Let's take an example, if you are watching a TV series say 'How to get away with Murder', Netflix is able to see how much you have completed that TV series at a particular point in time.

They further ask questions such as 'How many users completed a particular TV series', 'How big is time gap between users finishing a movie/TV series/episode and users beginning next episode or a movie'. They collect such kind of user data and see the trends in form of charts/graphs. For instance, if there was a cancelled show by the name 'Jessica Jones' and through the trends, Netflix comes to know that 70% of the users have binge watched all the seasons available of that show, then Netflix would be interested in restarting that show. They would be confident that users will watch the next season.

But Netflix dives deeper into data. Here's a look of user activities that Netflix has an eye on:

1. When you increase the speed of the video or rewind the video
2. The timestamp and the location you watched the content and the content related information
3. On which device, you watched the content and many more.
4. When you pause and leave the content
5. The ratings you give and the searches you do.

By gathering and analyzing such data, they pull out the user insights and improve their service in a much efficient manner ultimately leading to customer satisfaction.

(The All-In-One SEO Tool)

1.5. ANALYTICAL ROLES

Looking at the current opportunities at https://jobs.netflix.com/teams/data, the main analytical roles of Netflix are as follows:

Data Scientist

A person who is a Data Scientist at Netflix has immense experience in problem solving through statistics and Maths, who has a passion for solving real life problems by developing meaningful stakeholder expertise and deep domain knowledge, who is proficient in SQL and statistical programming in Python or R. A data scientist at Netflix would create algorithms that improve a service, mostly using machine learning and A/B testing. Content and recommendation algorithms are probably the core problem Netflix is tackling, so a data scientist would tackle the actual algorithm that is used to feed recommendations to users. They might run A/B tests on a new recommendation system they've created to figure out if users spend more time on Netflix with these recommendations. They might introduce more of one type of content into the ecosystem and test whether users watch that content more (and if that affects their total time watched).

Data Analyst

A data analyst at Netflix is the one who develops clean datasets for reporting, the one who is an expert in Google Sheets, Excel, Air table and Tableau. The mandatory programming skill includes Python/R. A data analyst at Netflix would perform analysis for the consumption of business partners. They would be tasked with identifying trends and building pipelines and dashboards for mass consumption. They would analyze improvements made to products and how much impact they have had, and possibly speculate on areas where the business should focus next. For example, a project a data analyst might do would be: "What is the impact of Wi-Fi coverage in Japan on Netflix usage? Do Japanese users stream Netflix on their data plan more or less often than US users? How else is their behavior different?"

Machine Learning Engineer

A machine learning engineer at Netflix is the one who uses machine learning algorithms so that the machine learns and makes predictions. Their core skills include Python/R and various machine learning algorithms such as neural networks, tensor flow, etc.

Data Engineer

A data engineer at Netflix is the one who collaborates with Machine Learning Engineers to build and test solutions. They are expertise in batch processing in distributed systems and proficient in Python/R as well.

Through this, we can conclude that one of the core skill for any data driven analytical role at Netflix is Python/R.

2. BUSINESS MODEL

Netflix business model can be explained in terms of how it benefits the business in terms of revenue and the users.

2.1. HOW NETFLIX GENERATES REVENUE?

Netflix revenue model is quite simple. It generates revenue by asking users to subscribe to their platform. They offer three levels of subscriptions i.e. basic, standard and premium. Once you have subscribed, you get to stream all the TV series and movies as per different genres as present in their catalogue.

Netflix has slowly changed the manner in which we consume traditional media. Starting from series like Titans to Black Mirror, Netflix has ruled the media industry with more than 150 members across the globe.

By setting up three plans of subscription i.e. basic, standard and premium ranging from 10 to 18$, Netflix has become a multi-billion dollar unicorn with more than 150$ billion revenue.

2.2. HOW THE BUSINESS BENEFITS USERS?

The current business model benefits the users in several ways as listed below:

Technology: Netflix allows us to watch content with ease on different kinds of d devices such as laptop/PCs/Tablet/smartphones, etc.

Comfort: In today's busy world, people don't have sufficient time to go to video store and purchase a DVD. People want comfort for the money they put in and Netflix

delivers it with ease wherein the content is presented to them in a personalized manner through their website/mobile application.

On demand: You can watch the content anywhere in the world, anytime you wish to.

Subscription: All it requires is you to subscribe for a low cost monthly fee.

Data driven: Netflix is data driven. It is not only a recommendation engine but also creates content that fits user preferences in an efficient manner.

2.3. NETFLIX CHALLENGES

The key challenges that NETFLIX faces globally are as follows:

1. **Content:** The more Netflix expands its market globally, the more it would need to create and produce content that is in line with the regional markets since it runs into local competitors that already have a regional content. Producing more content would lead to higher costs.

2. **Price:** Netflix's International expansion means it has to come up with different pricing structures to suit the market sensitivity. For example, Turkish users can get a subscription for 3.27$ a month while subscribers in Japan, Argentina has to pay not more than 6$ a month for streaming Netflix. You also have to consider the price of the subscription plan by taking the region's average income into account.

3. **Infrastructure:** Lack of fast paced internet across the globe is another challenge for Netflix especially in developing markets. This is beyond the company's control but this issue can harm long term growth of Netflix. Also, the internet services are expensive which poses another problem.

   NETFLIX is however tackling these challenges by setting up dedicated servers in developing markets so that the users can stream video faster and by addressing content and price related challenges.

3. DATA CHARACTERISTICS AND PROCESSING

3.1. DATA MODEL

From the last several years, there has been a tremendous increase in the streams of Netflix starting from thousands of users who watch ocassionaly to billions of people watching every hour. Every time, when a user starts a series or a movie, a view gets generated in their systems and collection of several events summarizing that view is collected. Having a robust data model which scales efficiently and is able to manage and process such huge chunks of viewing data is a critical factor for success.

Netflix's current data model is developed for wide range of cases starting from experiences of the user to data analytics. The following three cases enriches user experience.

1. **Watched Titles**

   As long as the user is subscribed, the systems know the entire viewing history of the user. This history data is fed to their recommendation algorithm so that the user is able to get a preference of title for whatever mood they are in. It also feeds the 'recent titles you have watched' row in the UI.

2. **The last title watched**

   Netflix records how much the user has watched a particular series or a movie and where the user left off. This enables the users to continue watching that particular movie or series on the same or another device the next time they log in.

3. **Account Sharing**

   Account sharing with friends and family members means everyone gets to watch their favorites. It also means that when the account's concurrent screen limit is reached, you might have to ask someone to stop watching. To support this, the data model of Netflix collects signals periodically throughout all the views which describes whether a user is still watching or stopped watching.

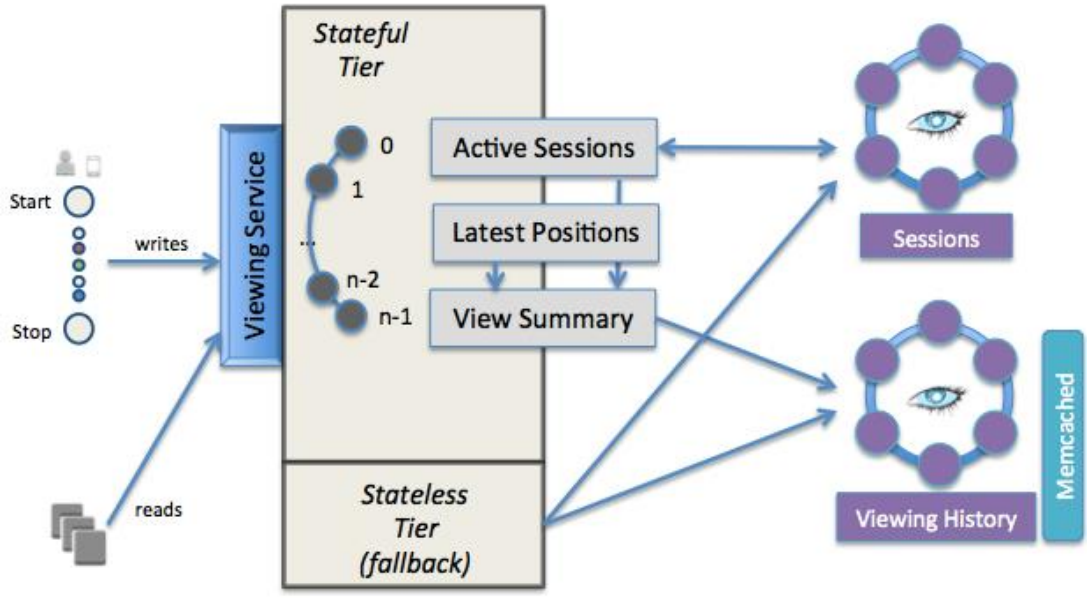

Figure 3 Viewing Data Model of Netflix

The core interface of the current Data Model of Netflix is the viewing service that is partitioned into two separate tiers i.e. the stateful and stateless tier. The latest data for all the currently running views stored in the memory is collected in the stateful tier. Data is segmented into N stateful nodes on the basis of user's account it. When the nodes come online, they go through a slot selection process which determines the data partition they belong to. Cassandra acts as a primary data store for all the persistent data. Memcached layer is added on the top of Cassandra which guarantees a low latency read path for materialized views of data.

(A Medium Corporation, 2015)

3.2. DATA CHARACTERISTICS

This section describes the characteristics of the data which is used to create the above data model. Below is a tabular representation which throws a light on the V's of the data at Netflix.

Characteristics of Data	
Volume	Volume refers to the huge amounts of data that is produced every second. At Netflix, large amounts of data is generated every second. From the point, user logs in at the Netflix website/application, till the user completes watching a movie

	or TV series and logs out from the application, thousands of API requests goes to backend, more than thousand tera bytes of data containing user events are stored. This data drives Netflix to make decisions and suggests contents to users based on their preferences.
Velocity	Velocity refers to the rate at which the data is generated. At Netflix, every second, requests made by the users across the globe to watch a particular movie or TV series are handled in a blink of an eye efficiently and the content is made available to the users in a seamless manner.
Variety	Variety refers to various types of data. Netflix stores all the different formats of data such as video files (.mpg, etc.), text files, spreadsheets, etc. Netflix stores several structured/unstructured data with ease in their databases.
Veracity	Veracity refers to the quality of data that is stored in the databases. Netflix has several batch processing jobs which improves the quality of poor data being stored.

3.3. DATA STORAGE AND PROCESSING

1. Data Storage

Long ago, Netflix used to host their digital content on third-party service providers but a couple of years back, they started to develop their own content delivery network (CDN). Now, all of the movies and TV series that you watch is cached and delivered through their customized network. This cache can then be streamed anytime, anywhere you want. The global pooled network of servers allow the media you stream to be sourced close by cutting down on bandwidth and increasing speed. The storage is basically collection of hard drives in a server. 36 drives are used that can withstand 100TB of data. These servers are able to store and stream between 10,000 and 20,000 movies simultaneously. Netflix has approx. thousands of these servers spread across the globe. Each server collects the content which is then transmitted to various devices.

Netflix has more than Petabyte worth of content which are sorted and organized like a library. During the off hours i.e. between midnight and lunch, Netflix prepopulates its servers with most famous movies/TV series there by reducing the bandwidth during the peak hours. The content is collected overtime and then streamed when called by the user. The media still needs to reach the users from the CDN. So, the ISPs (Internet Service Providers) connect to the Internet at large data centers around the world to keep everything connected. The providers can also choose to have Netflix install a CDN on the site, reducing bandwidth costs.

Further, Netflix is also boosted with AI. When a user logs into his/her account and selects a movie or TV series to watch, Netflix then tracks your location from where the user logged in and then decides which location in relation to you the particular movie should be streamed from. It quickly decides and in a blink of an eye, the content has travelled from the source through the data centers across the Internet, and into your home. The movies are stored and then transported thousands of miles to reach the user. It repeats the same process million times a day until a user binge watches all the latest episodes of Jessica Jones or How to get away with Murder.

(Uptrends, 2016)

2. Data Processing

This section describes the evolution of Netflix data processing pipeline over the years. The aim of data processing pipeline is to collect, aggregate, process and move data at cloud. Hundreds of event streams flow through this pipeline. For instance, Video viewing event, UI event, error logs, performance activities, trouble shooting and diagnostic activities. The sole purpose of original pipeline was to aggregate and upload events to Hadoop/Hive for batch processing.

Version 1.0 Apache Chukwa Pipeline

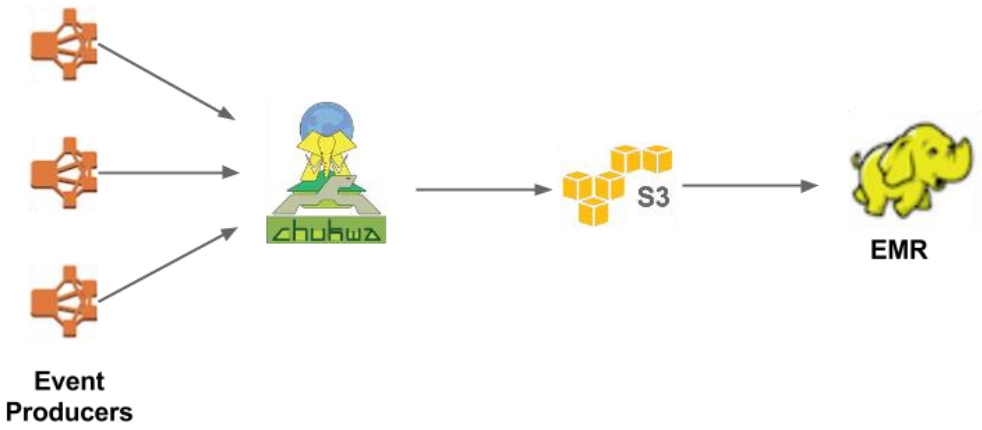

Figure 4 V1.0 Apache Chukwa Pipeline

This pipeline is simple. Apache Chukwa records the user events and writes them to Amazon S3 DB in Hadoop sequential file format. These event producers are basically Netflix UI (Web/Mobile) which is sending data to Apache Chukwa. Apache Chukwa further sends the data into Amazon S3 DB. So they are working with Amazon as well. Then, they use elastic map reduce to do batch jobs which gives a better analysis and results. The problem with this pipeline was the frequency of getting results was slow.

Version 1.5 Chukwa pipeline with real time branch

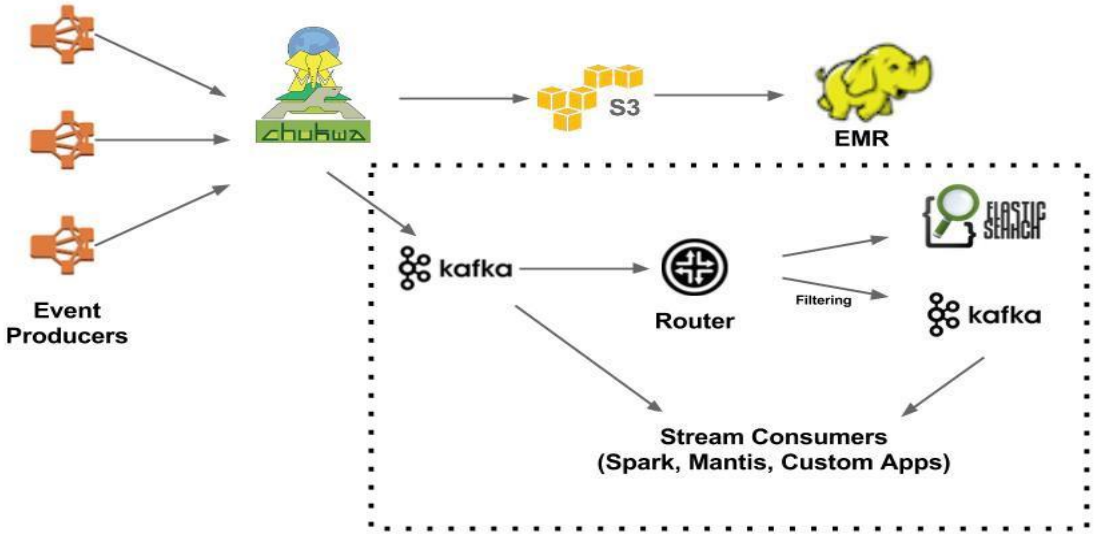

Figure 5 V1.5 Chukwa pipeline with real time branch

In this pipeline, they introduced Apache Kafka – a distributed streaming platform. This led to growing demand for real time analytics at Netflix. They added a real time branch where the data that comes in gets written into Kafka stream and then the router routes the data from Kafka to elastic search or another Kafka stream. Finally, the data is consumed by stream consumers such as Spark or custom applications. Basically the data here is being collected and splitted into streams.

V2.0 Keystone Pipeline (Kafka Frontend)

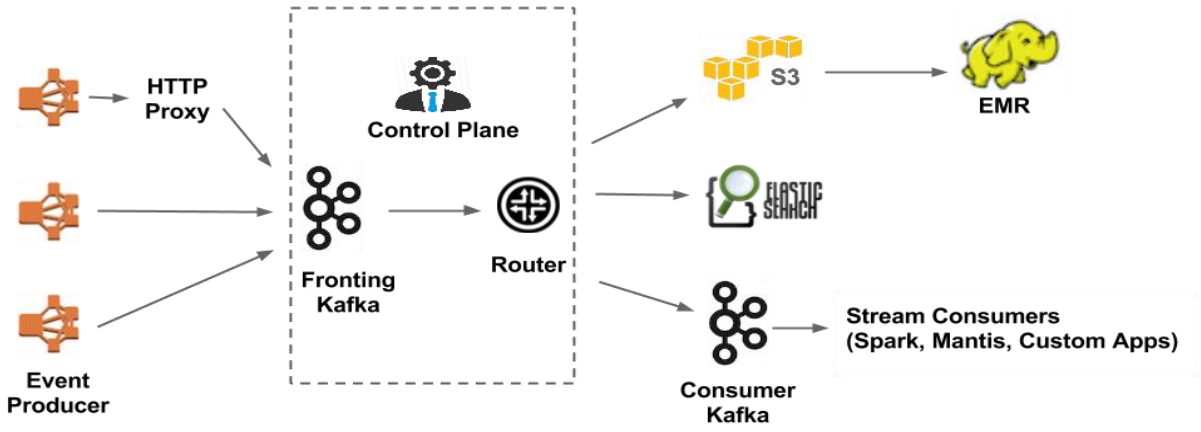

Figure 6 V2.0 Keystone Pipeline (Kafka Frontend)

This is the latest pipeline Netflix follows to collect, transport, aggregate, process and visualize events. Here, they completely removed Apache Chukwa. The event producers makes HTTP requests to APIs and APIs post the data into front facing Kafka stream which are collection interfaces or collectors. The routing service routes the data into different streams such as S3, Elastic Search or other Kafka streams.

There are three major components in this pipeline:

Data Ingestion: There are two ways for applications to ingest the data i.e. using a Java library and writing to Apache Kafka directly or send to an HTTP proxy which then writes to Kafka.

Data Buffering: Kafka serves as a replicated persistent messaging queue.

Data Routing: The routers are responsible for moving the data from fronting Kafka to various sinks i.e. S3, Elastic Search, other Kafka streams.

(A Medium Corporation, 2016)

4. RESOURCES

The potential resources can be categorized into data sources, data services and tools.

The data warehouse at Netflix comprises of huge datasets stored in databases such as Amazon S3, Druid, Elastic Search, Snowflake and My Sql. The platform supports Spark, Presto, Pig and Hive for consuming, processing and producing datasets. Using these wide range of data sources, the data platform interoperate across these datasets as one single data warehouse using Metacat.

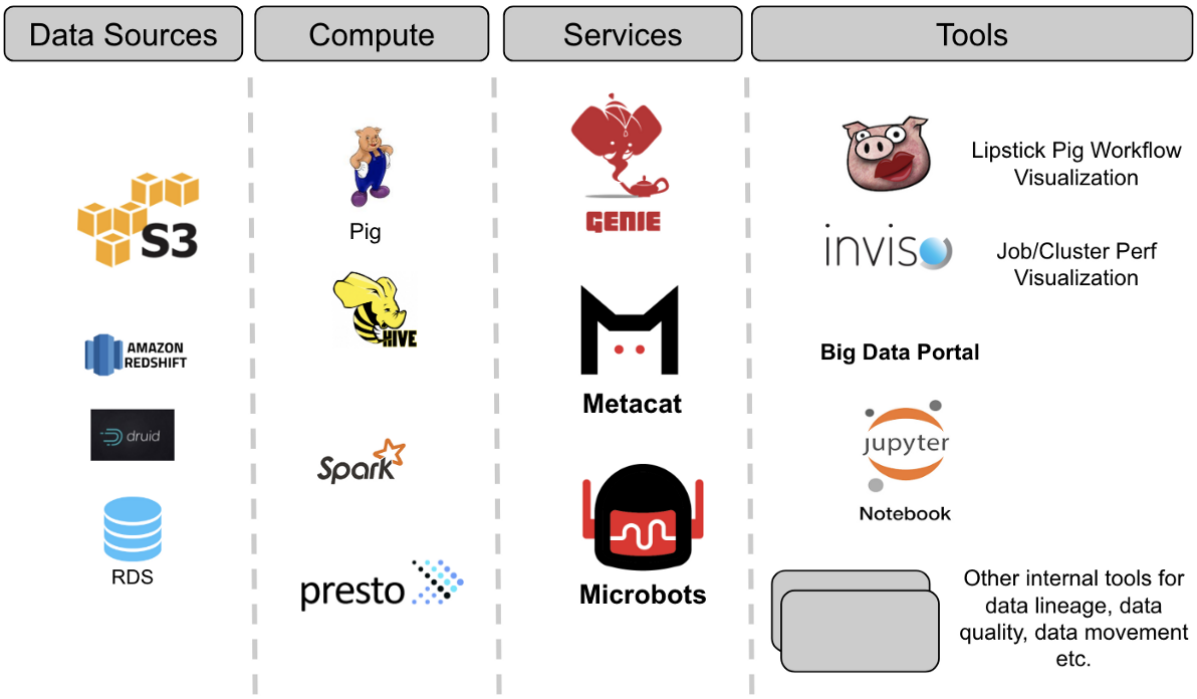

Figure 7 Sources, Services and Tools

The core architecture of Netflix's platform comprises of three major services i.e. Genie, the execution service, Metacat, the metadata service and Microbots, the event service. Netflix used Pig as their ETL language and Hive as their ad-hoc query language. Since Pig did not have a metadata system, a system called Metacat was developed that acts as a federated metadata access layer for all the data stores they support. Metacat is a centralized service that various computing engines of Netflix use to access different data sets.

The major tools used are Jupyter Notebook, Lipstick Pig Visualization Framework, Inviso: Visualizing Hadoop Performance, and other internal tools.

(A Medium Corporation, 2018)

5. DATA ANALYSIS

The Recommendation Algorithm

Everything is a recommendation. Recommendations are a great value added service to the users to personalize their content. Personalization begins from the home page of Netflix which comprises of collection of videos arranged in the horizontal rows. Each row displays a title which describes the collection of videos in that row with an intended

meaning such as ‘Top 10 movies’ or ‘Last Watched’. Majority of the personalization is based on the way users select the rows, how the users determine which videos or movies to include in those rows and in what order to place them.

Let’s take the first example the top 10 row, this is the best guess of Netflix at the 10 titles you will most enjoy. Netflix’s personalization aims to handle everyone with different likes and tastes. Hence, when you see that top 10 row, you most likely see the content for kids, adults and the whole family. Even for a single person, Netflix appeals to display recommendations based on their moods and interests. Netflix is not only focusing on accuracy but also for diversity.

Figure 8 Netflix Diversity

Second major element when it comes to personalization is awareness. Netflix aims to keep their users informed about how they are making an effort to adapt their tastes. This not only builds trust among the system but also encourages the users to give feedback that results in a better recommendation. Netflix decides to recommend a movie or TV show not because it suits their business needs but it matches the user information they have. Information such as ratings, taste preferences, viewing history or your friend’s recommendation.

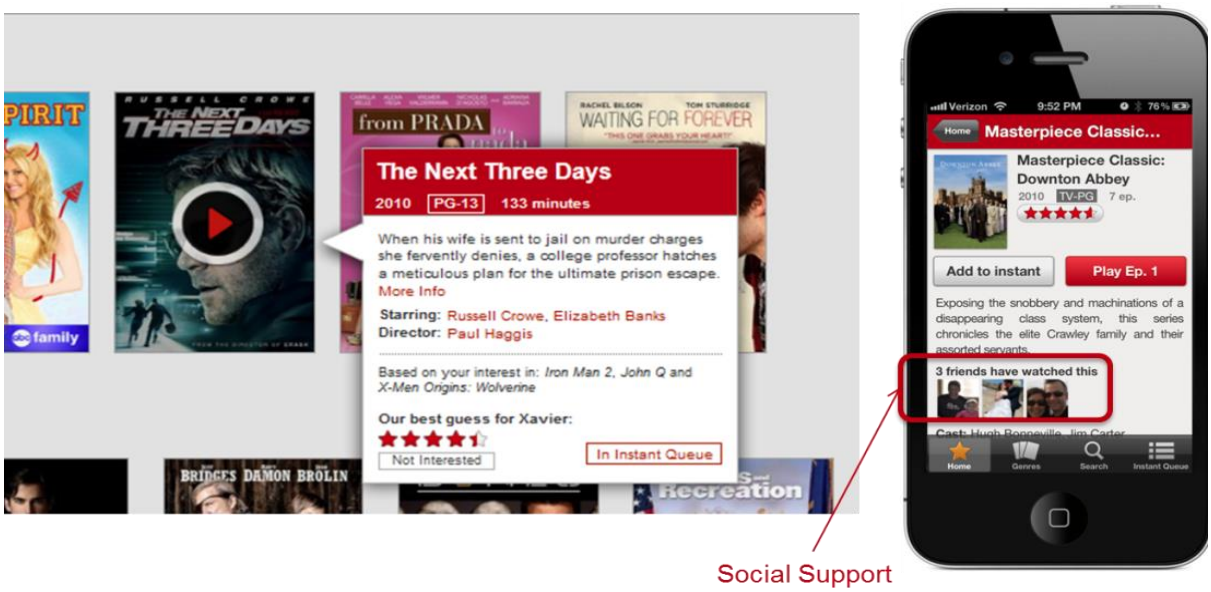

Figure 9 Netflix Awareness

As far as recommendations from your friends are concerned, knowing about your friends not only helps Netflix with their personalization algorithm but also allows for different rows that depend mostly on your social circle to generate recommendations.

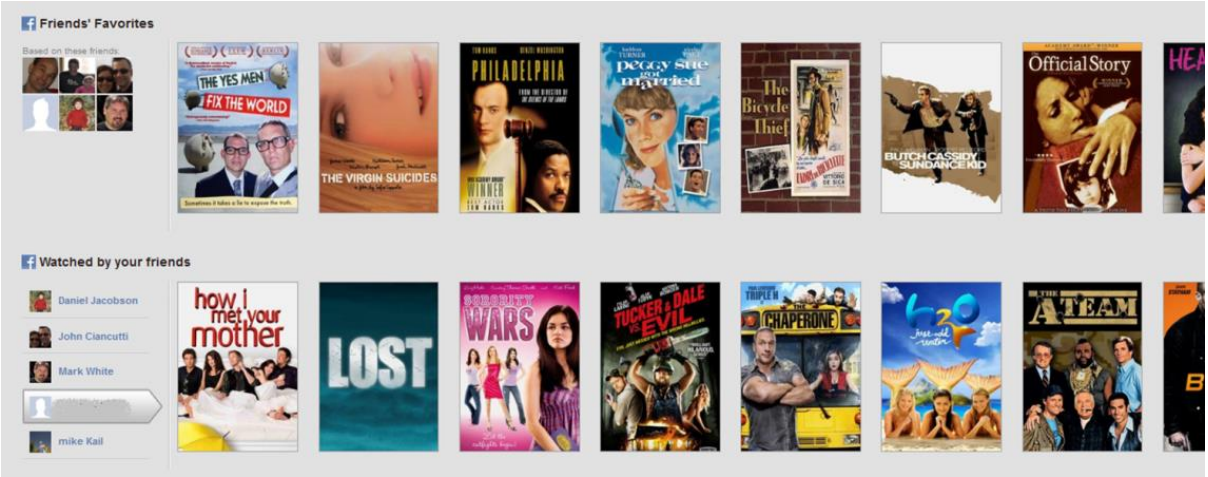

Figure 10 Social Circle

Some of the most recognizable personalization is the collection of genre rows which range from Comedy to Action. Each row has three layers when it comes to personalization. The choice of genre, the subset of titles, and the ratings given. Users connect with these rows so well and based on this, Netflix keeps the most tailored rows higher on the page than the lower.

Figure 11 Freshness of the content

One more important source of personalization is similarity. There can be similar movies or similar users and can be in many dimensions such as metadata, ratings or viewing data. These similarities are blended and used as features in their recommendation models. Similarity is used in multiple contexts for example in response to a user's action such as searching or adding a title to the queue. It can also be used to generate rows of adhoc genres based on similarity of titles that the user has interacted with recently.

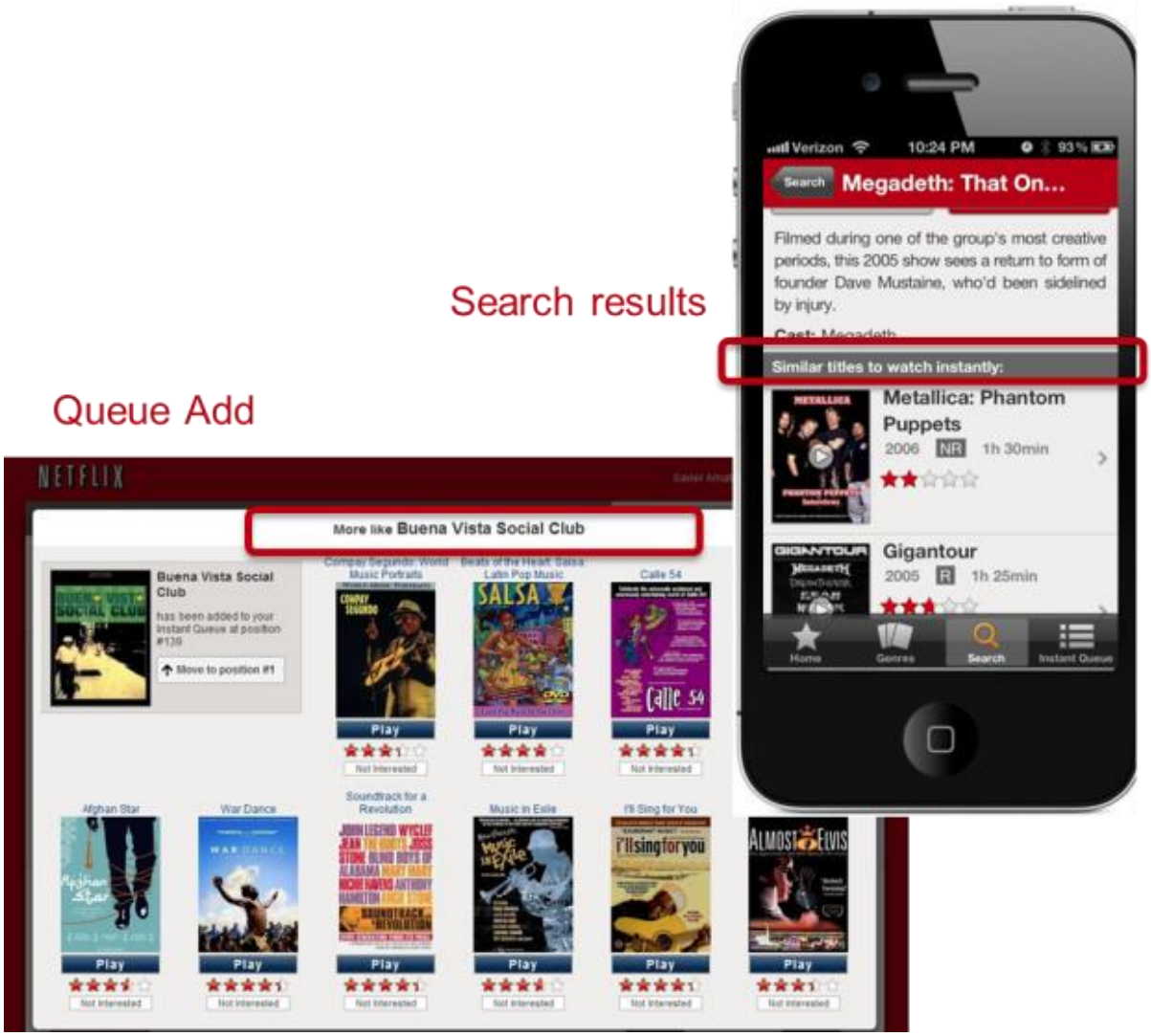

Figure 12 Similar Content

Another important thing is the ranking which decides what order the items needs to be placed in the row. The goal of their ranking system is to find the best possible ordering of a set of items for user real time. Netflix decompose ranking into scoring, sorting and filtering sets of movies for presentation to the user. The business objective is to maximize user satisfaction and monthly user retention. Netflix therefore optimizes its algorithm to provide the best scores to the titles that the user is most likely to watch and enjoy.

(A Medium Corporation, 2012)

6. CONCLUSION

From this case study, we can conclude that Netflix has been extremely successful in adapting the changing needs of customers over time. When online streaming became more and more popular, Netflix embraced the change and used it to their advantage to become the company they are today. The ability to provide more content than any physical DVD store has helped them become a multi-million company today. By changing the business models and the processing pipelines, Netflix has developed over time to become a successful and attention gaining business in the attention economy.

7. BIBLIOGRAPHY

A Medium Corporation. (2012, 4 6). Retrieved from https://medium.com/netflix-techblog/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429

A Medium Corporation. (2015, 1 27). Retrieved from https://medium.com/netflix-techblog/netflixs-viewing-data-how-we-know-where-you-are-in-house-of-cards-608dd61077da

A Medium Corporation. (2016, 2 15). Retrieved from https://medium.com/netflix-techblog/evolution-of-the-netflix-data-pipeline-da246ca36905

A Medium Corporation. (2018, 6 15). Retrieved from https://medium.com/netflix-techblog/metacat-making-big-data-discoverable-and-meaningful-at-netflix-56fb36a53520

The All-In-One SEO Tool. (n.d.). Retrieved from https://neilpatel.com/blog/how-netflix-uses-analytics/

Uptrends. (2016, 2 1). Retrieved from https://blog.uptrends.com/technology/binge-watching-how-netflix-content-is-stored-and-streamed/