INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

Big Data Analytics

CSC 3303
[Sem II 2018/2019]

Assignment 1

Group members :

Abdirahman Abdullahi	1432401
Asem Hamood Al-abdali	1513599
MHD Khaled Maen	1523591
Alrefaei Mohammad	1617111

NETFLIX

How Netflix Used Big Data To Show Us The Movies We Wan

Background

According to research firm Sandvine, the flowing of movie and TV service Netflix now accounts for more than a third of all downstream internet traffic during peak evening hours in the US. Netflix's service now has over 65 million members in over 50 countries entertaining more than 100 million hours of TV shows and movies a day. The data was collected from millions of subscribers and observed in a trial to understand people behaviors. What makes Netflix’s data a Big Data Company is the incorporation of this data with cutting-edge analytical technique.

What Problem Is Big Data Helping To Solve?

Netflix has over 100 million subscribers and with that comes a wealth of data they can analyze to improve the user experience. Big data has helped Netflix massively in its mission to become the king of the stream.

Big data helps Netflix decide which programs will be of interest to you and the recommendation system actually influences 80% of the content we watch on Netflix. The company even gave away a $1 million prize in 2009 to the group who came up with the best algorithm for predicting how customers would like a movie based on previous ratings. The algorithms help Netflix save $1 billion a year in value from customer retention.

Once upon a time, legendary screenwriter William Goldman said, “Nobody, nobody – not now, not ever – knows the least goddamn thing about what is or isn’t going to work at the box office.” That was of course before the time the internet, and big data and big data analytics decided to make an appearance and change everything.

The writer of Goldman was speaking before the arrival of the Internet and Big Data and, since then, Netflix has been determined to prove him wrong by building a business around predicting exactly what we’ll enjoy watching.

How Is Big Data Used In Practice?

“How Netflix Uses Big Data to Drive Success,” that highlights Netflix’s use of big data, specifically interesting statistics, how Netflix gathers big data, and how Netflix uses big data.

A quick look at Netflix’s job opportunities in the data field is to give you an idea of how seriously data and analytics are taken. experts are enlisted to join teams who skilled in implementing analytical skills to particular business areas: personalization analytics, messaging analytics, content delivery analytics, device analytics and so on. However, although Big Data is used across every aspect of the Netflix business, their holy grail has always been to predict what customers will enjoy watching. Big Data analytics is the fuel that fires the “recommendation engines” designed to serve this purpose.

Predicting what subscribers will want to watch next is another big business for networks recently, distributors and producers. The recommendation algorithms and content decisions are fed by data on what titles customers watch, what time of day movies are watched, time spent selecting movies, how often playback is stopped, either by the user or owing to network limitations, and ratings are given. In order to analyze the quality of experience, Netflix collects data on delays caused by buffering, the rebuffer rate, and bitrate, which affects the picture quality, as well as customer location. Going forward, Netflix is exploring Spark for streaming, machine learning, and analytic use cases, and they're continuing to develop new additions for their own open−source suite.

Netflix organized a big competition called the Netflix Prize and offering $1 million USD to the winner that could come up with the best algorithm for predicting how their customers would rate a movie based on their previous ratings. The winning entry was finally announced in 2009 and, although the algorithms are constantly revised and added to, the principles are still a key element of the recommendation engine.

In the beginning, there were limited in analysts by the lack of information on their customers – only four data classifiers which are customer ID, movie ID, rating and the date the movie was watched. However, there were data available for analysis as soon as streaming became the fast delivery way to customers, many new data points on their customers became accessible. This new data made Netflix build models and different algorithms to predict the main goal was to make customers consistently be served with movies they enjoy most. Satisfied customers are more likely to continue their subscriptions.

recently, Netflix has begun to call themselves as a content creator, not only distribution and streaming for movie studios and other networks. Their strategy here has also been firmly driven by their data – which showed that their subscribers had a voracious appetite for content directed by David Fincher and starring Kevin Spacey. After outbidding networks including HBO and ABC for the rights to House of Cards, they were so confident it fitted their predictive model for the "perfect TV show" that they bucked the convention of producing a pilot and immediately commissioned two seasons comprising 26 episodes. Every aspect of the production under the control of Netflix was informed by data – even the range of colors used on the cover image for the series was selected to draw viewers in.

One way that data has changed the streaming video experience is autoplay. Analytics showed that customers often leave after a show is over, so many streaming sites began to automatically start a new episode of a show as soon as the last one ended. For binge watchers, autoplay eased the transition from episode to episode. And, once you've finished watching a show, it tees up another program that's been selected to appeal to your tastes. If all goes well, the binge process repeats itself, you stay glued to the screen, and the number of hours that you watch rises.

What Were The Results?

In April 2015, Netflix's send a letter to all shareholders shows their Big Data strategy was paying off. 4.9 millions of new subscribers were added in first quarter 2015, matched to 4 million in the same duration in 2014. Netflix linked all the success to their "ever-improving content", including House of Cards and Orange is the New Black. This original content is driving new member gaining and customer reservation. In fact, most of Netflix members approximately 90% have engaged with this original content. Obviously, a large part of this success is their ability to predict what viewers will enjoy. Their final measure is: how many hours do customers spend using the service? In the first quarter of 2015 alone, Netflix members watched 10 billion hours of content. If Netflix's Big Data strategy continues to improve, that number will be increased.

What Data Was Used?

Based on what titles customers watch the recommendation system algorithms and content decisions are supplied by data, the time of day movies are being watched, the time spent of selecting movies, if the playback is stopped (either by the user or due to network limitations) and the ratings which are given by the customers. Netflix collects data on delays caused by buffering, and bitrate (which affects the picture quality), as well as customer location, in order to analyze the quality of experience.

What Are The Technical Details?

Even Though their large catalog of movies and TV shows is hosted in the cloud on Amazon Web Services (AWS), it is also mirrored around the world by ISPs and other hosts. Additionally, improving user experience by decrease lateness when streaming content globally, this reduces costs for the ISPs – saving the customers from the cost of downloading the data from the Netflix server before transmitting it on to the viewers at home.

The size of their catalog exceeds three petabytes in 2013. This tremendous amount of data is considered by the need of holding many of their titles up to 120 different video formats, because of the number of different devices that Netflix playback offers.

Their systems used Oracle databases primarily, but they convert to NoSQL and Cassandra to allow more complex, Big Data-driven analysis of unstructured data.

The Netflix data infrastructure contains Big Data technologies like Hadoop, Hive, and Pig plus traditional business intelligence tools like Teradata and MicroStrategy. Netflix's have their open-source applications and services Lipstick and Genie. And, like all of Netflix's core infrastructure are also included, all of them run in the AWS cloud. Furthermore, Netflix is exploring Spark for streaming, machine learning, and analytic use cases, they keep going to develop new additions for their own open-source Composition.

Any Challenges That Had To Be Overcome?

A lot of the metadata collected by Netflix for example, actors who viewer likes to watch and the time of day they watch films or TV, it is easily quantified structured data, Netflix recognizes a lot of valuable data stored in the chaos, unstructured of video and audio content.

Making this data obtainable for computer analysis to unlock its value, somehow it had to be quantitative. Netflix achieves that by paying teams of thousands of viewers sitting through hours of content, accurately tagging elements they found in them.

These paid viewers marked up themes after reading a 32-page handbook, issues, and trims that took a place on the screen and made tough moral choices are a hero experiencing a religious absorption or a strong female character. Based on this data, Netflix has specified approximately 80,000 (micro-genres) like comedy films featuring talking animals or historical themes. Netflix can easily now identify what films customer like to watch much more accurately without depending on the customer's likes, and can also use this to predict what the customers will watch. This method gives the unstructured, messy data the outline of a structure that can Evaluate the quantity of the fundamental principles of Big Data.

By creating routines that can take a snapshot of the content (Jpeg format) and analyze everything happening on the screen using advanced technologies such as facial recognition and color analysis, Netflix is capable of automating this process. These snapshots can be taken either at a scheduled period or when a user takes a particular action like pausing or stopping playback.

What Are The Key Learning Points And Takeaways?

Prediction is what viewers will want to watch next and that is becoming a successful business for networks recently, distributors and producers in all aspects that Netflix now consider the top in the media industry. Apple and other big companies can also be counted on to be improving and refining their own analytics. Predictive content programming is a field in which we can expect to see continued innovation, driven by fierce competition.

Netflix now has its own ideal of changing the normal TV to personalized TV, where any customers can have their own schedule of movies to consume, based on their preferences and favorites.

REFERENCES

1. http://techblog.netflix.com/
2. http://www.netflixprize.com/http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html
3. http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/
4. http://www.wired.com/insights/2014/03/big-data-lessons-netflix/
5. http://files.shareholder.com/downloads/NFLX/47469957x0x821407/DB785B50-90FE-44DA-9F5B-37DBF0DCD0E1/Q1_15_Earnings_Letter_final_tables.pdf
6. https://insidebigdata.com/2018/01/20/netflix-uses-big-data-drive-success/
7. https://blog.datahut.co/how-netflix-leverages-big-data-analytics-to-drive-success/
8. **https://onlinelibrary.wiley.com/doi/10.1002/9781119278825.ch3**
9. **https://www.ceo.com/data-and-business/the-new-age-of-media-metrics**

LINKEDIN

How to use big data to fuel the success of social media.

Background

LinkedIn is the world's largest on-line expert network, with more than 410 million individuals in over 200 countries.LinkedIn connects specialists by using enabling them to construct a network of their connections and the connections of their connections. The site was launched by Reid Hoffman in 2003, making it one of the oldest social media networks in the world.

What Problem Is Big Data Helping To Solve?

Competition among social networks is fiercer than ever and what's warm one month may also no longer be the next. LinkedIn need to make sure their website remains an imperative tool for busy professionals, assisting them emerge as greater productive and successful, whether they're the use of the premium (paid-for) service or the free carrier .
As such, Big Data is at the very heart of LinkedIn's operations and decision making, helping them furnish the excellent feasible service for the site's hundreds of thousands of members.

What Were The Results?

LinkedIn's success metrics include income and variety of members, both of which proceed to upward push 12 months on year. LinkedIn gained forty million new individuals in the first half of of 2015 and, at the time of writing, the company's most recent quarterly revenue stood at over $700 million (up from around $640 in the previous quarter). There's no doubt that Big Data performs a giant function in the company's endured success.

What Data Was Used?

LinkedIn tune every move their customers make on the site, from the whole thing appreciated and shared to every job clicked on and each contact messaged. The organisation serve tens of hundreds of Web pages each and every second of every day. All these requests contain fetching data from LinkedIn's backend systems, which in flip handle millions of queries per second. With permission, LinkedIn additionally acquire facts on users' Email contacts.

What Are The Technical Details?

Hadoop form the core of LinkedIn's Big Data infrastructure, and are used for each advert hoc and batch queries. The corporation have a large investment in Hadoop, with thousands of machines strolling map/reduce jobs. Other key components of the LinkedIn Big Data jigsaw encompass Oracle, Pig, Hive, Kafka, Java and MySQL. Multiple statistics centres are particularly important to LinkedIn, in order to make certain high availability and keep away from a single factor of failure. Today, LinkedIn run out of three essential facts centres.

LinkedIn have also developed their personal open-source equipment for Big Data get right of entry to and analytics. Kafka commenced lifestyles this way, and other trends include Voldemort and Espresso (for records storage) and Pinot (for analytics). Open-source technology like this is important to LinkedIn due to the fact they sense it creates higher code (and a better product) in the long run.

In addition, the company have an astounding group of in-house records scientists – around 150 at contemporary estimates. Not only do the crew work to enhance LinkedIn merchandise and resolve troubles for members, they also put up at most important conferences and contribute to the open-source community. In fact, the group are prompted to actively pursue research in a wide variety of areas, together with computational advertising, computing device learning and infrastructure, textual content mining and sentiment analysis, safety and SPAM.

Any Challenges That Had To Be Overcome?

When you assume that LinkedIn started out with just 2700 members in their first week, big information growth is one obvious venture LinkedIn continuously have to overcome – the

organization now have to be in a position to take care of and apprehend large amounts of records every day. The solution to this is in investing in incredibly scalable systems, and making sure that the records is nevertheless granular ample to supply useful insights.
Hadoop grant the back-end energy and scalability needed to cope with the volumes of data, and LinkedIn's user interface approves their personnel to slice and cube the data in lots of exclusive ways.

From a business enterprise that employed fewer than a thousand employees five years ago, LinkedIn have grown to appoint nearly 9000 people. This locations considerable demand on the analytics team. Perhaps in response to this, LinkedIn currently reorganized their information science group so that the choice sciences section (which analyses data usage and key product metrics) now comes underneath the company's chief economic officer, while the product facts science section (which develops the LinkedIn facets that generate hundreds of records for analysis) is now phase of engineering. As such, facts science is now extra built-in than ever at LinkedIn, with analysts becoming extra carefully aligned with company functions.

It can also come as a shock to study that hiring group of workers is also a challenge, even for a giant like LinkedIn. Speaking to CNBC.com, LinkedIn's head of statistics recruiting, Sherry Shah, confirmed they were searching to employ extra than 100 statistics scientists in 2015 (a 50% increase from 2014). But opposition for the first-rate facts scientists is tough, in particular in California, and Shah admitted that "there is continually a bidding war". Although extra human beings are entering the field, it's probably this abilities gap – the place demand for records scientists outstrips furnish – will proceed for a few years yet.

In addition, LinkedIn haven't escaped the privateness backlash. In June 2015, the organisation agreed to pay $13 million to settle a type action lawsuit ensuing from sending more than one Email invitations to users' contact lists. As a result of the settlement, LinkedIn will now explicitly nation that their "Add Connections" device imports tackle books, and the website will enable these who use the device to select which contacts will obtain automated invites and follow-up emails.

What Are The Key Learning Points And Takeaways?

As one of the oldest social media networks and still going strong, LinkedIn provide a lesson to all businesses in how Big Data can lead to huge growth. Their capability to make hints and hints to users is especially enviable (and is also used efficiently by means of different corporations featured in this book, such as Etsy and Airbnb). But LinkedIn also furnish an example of the want for transparency when the use of individuals' records – and the backlash that can show up when humans sense a agency isn't being completely transparent. I assume we can expect to see greater complaints like this in opposition to corporations in future so it's necessary to be crystal clear with your clients what statistics you are gathering and how you intend to use it.

REFERENCES

1. https://engineering.linkedin.com/big-data
2. https://engineering.linkedin.com/architecture/brief-history-scaling-linkedin
3. http://www.mediapost.com/publications/article/251911/linkedin-to-pay-13-million-to-settle-battle-over.html