

Case Study: Netflix Big Data Automating
the Video Delivery Business

Megan Bell

Spring 2017

Table of Contents

Introduction	3
Scope of Big Data.....	3
Brief History	4
Importance to Backend Operations	5
Serving Data Scientists.....	6
How Netflix Achieved Big Data Infrastructure.....	6
Automating Big Data Infrastructure.....	7
Impact of Big Data	8
Business Investment	8
Growing Need for Talent.....	9
Big Data as a Service.....	9
Summary.....	10
References	11

Introduction

Originally founded as a mail-order DVD rental company, Netflix has transitioned to a leader in on-demand streaming video with more than 93 million subscribers across 190 countries (Netflix, 2017). Netflix success can be attributed to its business model of broad video choice delivered in a user-friendly and consumable format for a low, fixed monthly price. Subscriber members have access to Netflix content from their device at a time and location of their choice that has available Internet access (Netflix, 2017; Twentyman, 2014). To maintain long-term member engagement and loyalty, Netflix makes significant infrastructure investments such as recommendation engines that customize content or improve backend systems operations (Netflix, 2017). At the scale of data amassed by Netflix, big data aggregation and automation is essential to business viability and enduring success.

This paper presents a case study of Netflix big data automation technology and its importance to Netflix. First, a summary of Netflix big data perspective is provided. Next, the need for big data automation is discussed and an example of real-time data warehousing is provided to demonstrate the impact of big data infrastructure. Big data automation has impacted Netflix business, and the impacts on business investment and culture are summarized. Finally, this case study explores Netflix big data as a service.

Scope of Big Data

According to Nirmal Govind, Director of Streaming Science and Algorithms, core Netflix services are built on data (2015, New). To this end, Netflix is a

comprehensive consumer of its own data in delivering shows that can be “[watched] as much as they want, anytime, anywhere, on nearly any internet-connected screen” (Netflix, 2017). Big data automation provides real-time analytics that are used to deliver customer-facing service, monitor infrastructure performance and service delivery, analyze social relationships, and support internal data science teams with readily available data for analytics (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011). Big data automation has also transformed Netflix data into a self-service engine that powers most Netflix services and empowers teams of data scientists (Harper, 2016; Skelly, 2016; Simon, n.d.).

Big Data’s Role and Evolution

A brief review of Netflix services provides context to understand Netflix decision to employ big data as an infrastructure service. Big data encompasses the aggregation, management, and potentially dissemination of most data at any “volume, velocity, and variety” for services such as streaming video, enabling members to “play, pause and resume watching” for any show, or optimizing software processing in production (Lycett, 2013; Netflix, 2017). Netflix has massive data stores that require big data architecture for almost any conceivable task. For example, the Netflix video catalogue stores more than 3 petabytes of data, which is 3,000,000 GB (Wang, Xu, Chen, & Chen, 2016). Netflix must also capture and maintain data for its millions of active viewers’ preferences and show histories and streams more than 125 million hours of video daily (Netflix, 2015). Also, Netflix must re-package and deliver video content for more than 1,000 device types with different streaming video capabilities (Weeks & Gianos, 2017). Big data automation

is particularly critical for backend operations such as streaming video and enabling the Netflix team of data scientists.

Importance to Backend Operations

While customer recommendations may be the most obvious reason for Netflix big data automation, it is perhaps more important for managing the backend infrastructure. As an example, Figure 1 below presents a high-level diagram of backend systems where a Netflix member is able to log into Netflix from a device and then access video content in the Netflix video library. Netflix systems identify video files that match a member's request and device type and then send the appropriate files from the closest content distributed network (CDN) provider (Adhikari, Guo, Hao, Varvello, Hilt, Steiner, & Zhang, 2012). The seamless delivery process is transparently on-demand for the end-user. Netflix can scale to 10,000 or 20,000 servers on demand depending on the number of shows accessed (Vance, 2013).

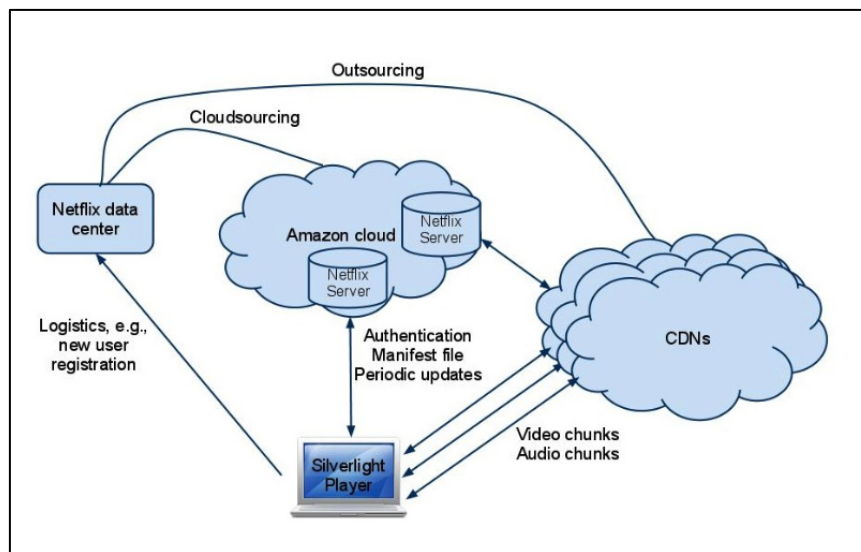


Figure 1. Simplified view of Netflix architecture. Reprinted from “Unreeling Netflix: Understanding and Improving Multi-CDN Movie Delivery”, from Adhikari, Guo, Hao, Varvello, Hilt, Steiner, & Zhang, 2012, INFOCOM, 2012 Proceedings IEEE (pp. 1620-1628). Copyright 2017 IEEE.

Serving Data Scientists

Former Vice President of Data Science, Eric Colson, identified Netflix as having “unwieldy” data (Cox, 2013). Data scientists are limited without supporting big data architecture for their diverse roles across Netflix. Netflix data science teams are broken into specialized groups such as computer science, content, recommendations, and video streaming to focus on diverse scope of data (Skelly, 2016). Each team requires efficient access to big data for completion of tasks such as recommendations development.

How Netflix Achieved Big Data Infrastructure

To better understand the impact of big data automation, Tom Gianos who manages the Netflix big data team and senior software engineer Dan Weeks shared their experiences on the automation of the Netflix big data platform real-time analysis (Gianos & Weeks, 2017). Both work with the Netflix data warehouse (DW), which Weeks describes as the Netflix “source of truth”. The real-time data warehouse (DW) system processed more than 500 billion daily collected system events and holds more than 60 petabytes (PB) of data, has daily reads of 3 PB, and daily writes 500 terabytes (TB). Read means extract information, and write means add information

The big data infrastructure that supports the DW is presented in Figure 2. In order to move from real-time streaming data to the DW, a two-part process was defined. First, applications events data Netflix is collected and channeled through data collection, cleanup, and transformation. Once transformed, events data is placed in the DW. Second, Netflix collects dimension data from what are known as Cassandra databases. Cassandra databases house data such as member information for Netflix services such as video viewing activity. Located in the Amazon cloud with infrastructure-level design, “the architecture delivers real-time analytics for data scientists across the Netflix business”, according to Weeks (2017).

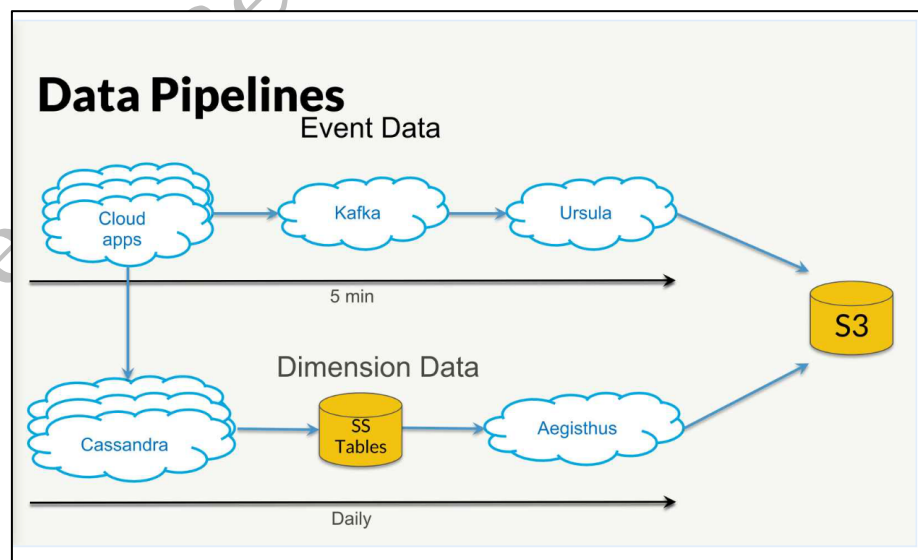


Figure 2. Netflix event data collection for data warehouse. Reprinted from presentation provided by Gianos & Weeks, 2017. Copyright 2017 Netflix, Inc.

Automating Big Data Infrastructure

Delivering DW functionality requires other components with the Netflix big data infrastructure. The infrastructure is important due to the volume and complexity of DW data resources. With more than 300 data scientists needing

access to differing data resources, automation tools and services were necessary to provide flexible access and other analytics resource capabilities such as task scheduling. Netflix automation of big data enables the Netflix analytics platform to handle more than 45 thousand tasks per day.

Impact of Big Data

Business Investment

Big data automation was part of substantial investment required to streaming video. This included significant R&D investment. Between 2012 and 2016, Netflix was spending \$.09 per dollar earned on research and development (R&D), and R&D expense was outpacing net income as observed in Figure 3.

Netflix business performance appears to support the impact of Netflix R&D investment. By 2015, Netflix membership exceeded 80 million members who streamed more than 125 million hours of video daily (Netflix, 2015; Netflix, 2017). Netflix subscribership also grew by 16% between 2015 and 2017. In 2016, Netflix was named marketer of the year by AdAge for its marketing excellence (Poggi, 2016).

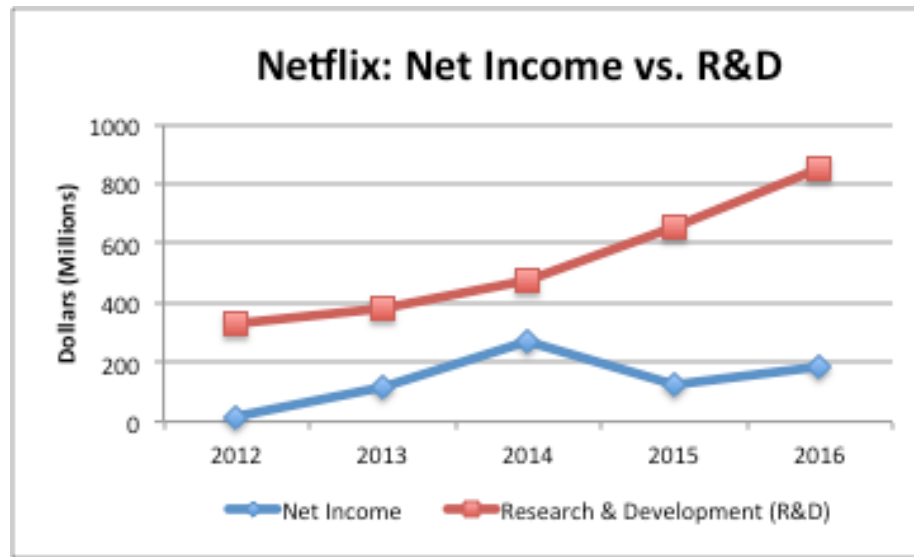


Figure 3. Netflix net income versus research and development expense.

Created from "Annual Financials for Netflix Inc.", from FactSet Research

Systems Inc. Retrieved, on April 1, 2017, from

<http://www.marketwatch.com/investing/stock/nflx/financials>

Growing Need for Talent

As Netflix transitioned from video rentals to stream-video services, Netflix realized the need to change its employee composition to include top-technical talent. To this end, Netflix underwent cultural transformation to becoming a digitally integrated business, and human resources were reshaped to focus on digital talent and release employees without needed skill sets (Kaufman & Horton, 2014; McCord, 2014; Nocera, 2016). Over time, Netflix was able to attract top-talent with 10% higher salaries and perks such as generous parental leave (Alba, 2015).

Big Data as a Service

With its expansive big data infrastructure and large data science teams, big data is used in furtherance of efficient streaming video services to deliver customer-

preferred content. As previously discussed, big data automation enabled development of a real-time data warehouse. Another benefit of big data automation is the ability to better understand member viewing psychology and preferences to evolve the Netflix business model (Erevelles, Fukawa, & Swayne, 2016). A well-documented example of this phenomenon was the Netflix purchasing rights to “House of Cards”—a British TV show that Netflix repurposed into an award-winning series (Carr, 2013). From a backend systems perspective, big data automation enables engineers to identify efficiencies such as algorithms that deliver bandwidth compression. As an example, Netflix identified a method of recoding streaming video in 2014 that resulted in 20 percent bandwidth reduction (Dillet, 2015). By harnessing big data automation, Netflix has developed a unique business model where introspection drives self-improvement and improvement in customer-oriented services.

Summary

Netflix has accomplished the feat of delivering video on-demand and becoming a “leading internet television network “ (Netflix, 2017). Where other companies failed, Netflix architected an automated big data infrastructure to deliver streaming video at consumer-agreeable price-points and overcome traditional video rental and cable-exclusive programming channels (Curtin, Holt, & Sanson, 2014). Big data automation evolved to self-service infrastructure for overall services management and to provide the basis for real-time analytics against petabytes of data for diverse teams of data scientists. Big data has empowered Netflix to excel in the streaming video and entertainment markets.

References

- Adhikari, V. K., Guo, Y., Hao, F., Varvello, M., Hilt, V., Steiner, M., & Zhang, Z. L. (2012, March). Unreeling Netflix: Understanding and improving multi-CDN movie delivery. In INFOCOM, 2012 Proceedings IEEE (pp. 1620-1628). IEEE.
- Alba, D. (2015, Dec. 9). Netflix adds hourly workers to its generous parental leave plan. Wired. Retrieved from <https://www.wired.com/2015/12/netflix-adds-hourly-workers-to-its-generous-parental-leave-plan/>
- Carr, D. (2013, Feb. 24). Giving viewers what they want. The New York Times. Retrieved from http://www.nytimes.com/2013/02/25/business/media/for-house-of-cards-using-big-data-to-guarantee-its-popularity.html?_r=1&%20
- Cox, R. (Mar 8, 2013). Former Netflix data scientist finds smaller data sets “less daunting” for Fashion Algorithms (Interview). Retrieved from <http://siliconangle.com/blog/2013/03/08/former-netflix-data-scientist-finds-smaller-data-sets-less-daunting-for-fashion-algorithms/>
- Curtin, M., Holt, J., & Sanson, K. (Eds.). (2014). Distribution revolution: conversations about the digital future of film and television. Univ of California Press.
- Dillet, R. (2015). Netflix’s ongoing quest to save bandwidth. TechCrunch. Retrieved from <https://techcrunch.com/2015/12/15/netflixs-ongoing-quest-to-save-bandwidth/>
- Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research*, 69(2), 897-904.

FactSet Research Systems Inc. (n.d.) Annual Financials for Netflix Inc. Retrieved, on April 1, 2017, from

<http://www.marketwatch.com/investing/stock/nflx/financials>

Gianos, T. & Weeks, D. (2017). Netflix: Petabyte scale analytics infrastructure in the cloud (Video presentation). InfoIQ. Retrieved from

<https://www.infoq.com/presentations/netflix-big-data-infrastructure>

Gianos, T. & Weeks, D. (2017). Netflix: Petabyte scale analytics infrastructure in the cloud (Presentation). Qconf. Retrieved from

https://qconf.com/sf2016/system/files/presentation-slides/netflix-cloud_analytics.pdf

Harper, J. (2016, Jan. 12). Automation is the new reality for big data initiatives.

DataInformed. Retrieved from <http://data-informed.com/automation-new-reality-for-big-data-initiatives/>

Kaufman, I. & Horton, C. (2014). Digital marketing: integrating strategy and tactics with values, a guidebook for executives, managers, and students. New York London: Routledge, Taylor & Francis Group

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. MIT Sloan management review, 52(2), 21.

Lycett, M. (2013). 'Datafication': Making sense of (big) data in a complex world. *European Journal of Information Systems*, 22(4), 381.

McCord, P. (2014). How Netflix reinvented HR. Harvard Business Review. Retrieved from <https://hbr.org/2014/01/how-netflix-reinvented-hr>

Netflix, Inc. (2015). Form 10-K 2015. Retrieved from <https://ir.netflix.com/annuals.cfm>

- Netflix, Inc. (2017, Jan. 18). Q416 Letter to shareholders . Retrieved from
<http://files.shareholder.com/downloads/NFLX/4154650369x0x924415/A5ACACF9-9C17-44E6-B74A-628CE049C1B0/Q416ShareholderLetter.pdf>
- New, J. (2015, Sep 28). 5 Q's for Nirmal Govind, Director of Streaming Science and Algorithms at Netflix. Retrieved from
<https://www.datainnovation.org/2015/09/5-qs-for-nirmal-govind-director-of-streaming-science-and-algorithms-at-netflix/>
- Nocera, J. (2016, Jun. 15). Can Netflix survive in the world it created. New York Times. https://www.nytimes.com/2016/06/19/magazine/can-netflix-survive-in-the-new-world-it-created.html?_r=0
- Poggi. (2016). Marketer of the year: Netflix. AdAge. Retrieved from
<http://adage.com/article/news/ad-age-s-2016-marketer-a-list/306928/>
- Simon, P. (n.d.). Big data lessons from Netflix. Wired Magazine. Retrieved from
<https://www.wired.com/insights/2014/03/big-data-lessons-netflix/>.
(Reprinted from The visual organization: Data visualization, big data, and the quest for better decisions, by P. Simon ,2014, Wiley Press.)
- Skelly, L. (2016, Feb. 29). Interview with Todd Holloway (Audio). Retrieved from
<https://www.thisismetis.com/blog/speaker-series-todd-holloway>
- Twentyman, J. (2014). Forget orange; for Netflix, Big Data is the new black. Diginomica. Retrieved from <http://diginomica.com/2014/07/11/forget-orange-netflix-big-data-black/>
- Vance, A. (2013). Netflix, Reed Hastings survive missteps to join Silicon Valley's elite. Bloomberg. Retrieved from
<https://www.bloomberg.com/news/articles/2013-05-09/netflix-reed-hastings-survive-missteps-to-join-silicon-valleys-elite#r=rss>

Wang, C., Xu, S., Chen, L., & Chen, X. (2016, June). Exposing library data with big data technology: A review. In *Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on* (pp. 1-6). IEEE.