Table of Contents

Overview

Thinking about watching a movie or an episode from your favourite series tonight? If so, the chances are you are thinking about doing it on Netflix. Netflix has become ubiquitous in the video streaming industry. While Youtube is still the biggest provider of streamed content in the world, Netflix is the largest provider of streamed movies and series, streaming over one billion hours of content per month (https://www.youtube.com/watch?v=lQGHsBOZJBw). Since its expansion into streaming services outside of the USA in 2011, Netflix has grown from just over 20 million subscribers, to over 200 million, with each subscriber watching, on average, 3.5 hours of video content per day (Dean, 2021). This paper investigates Netflix's approach to dealing with big data, its transfer and storage, following their transition from housing their own services in massive data centres to their implementation of a cloud computing solution, based mainly in the Amazon Web Services (AWS) cloud.

What is Big Data?

Big data is the term used to describe the ever-increasing amount of structured and unstructured data which proves challenging to process in real-time using current data science paradigms and processes. One of the issues surrounding what constitutes big data, and what does not, is the fact that there is no singular accepted definition, with many of the stakeholders providing their own variation. There are, however, some commonly accepted characteristics of big data. A Meta (now Gartner) report from 2001 (Douglas, 2001) detailing and predicting the exponential growth of data as a result of increasing internet usage listed three main problematic

characteristics of data: volume, velocity and variety. These have since become known as the three V's of big data, and that for data to be considered as big data it should display these characteristics. In other words, it is the size of data, the speed at which it is produced and transferred, and the wide variety of formats in which it is represented that define it as big data. Ward and Barker (2013) conducted a review of various large companies' definitions of big data and found that supplementary to the size and complexity of the data, the technologies used to process it were also what marked it out as big data, resulting in their proposal of the following definition:

> "**Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning.**"

What is the "cloud"?

Rather than building and maintaining physical servers and large data centres, technological advances in the last fifteen years have made it possible to access computing power to process, store, and manage data, on a flexible basis from a provider such as Microsoft Azure, Google Cloud, or Amazon Web Services (AWS).

Though cloud computing is a term with numerous definitions, there is broad agreement of a set of characteristics that a cloud should have (Buyya, Broberg, Goscinski, 2010). These include but are not limited to a scalable capacity, where resources can be added to or subtracted from the network on demand, a pay-as-you-go model, where the customer only pays for the resources it is currently using (as other resources can be redirected towards other customers when not in use), a self-service interface, where the customer can control and interact with the

resources its using, and finally that the resources are virtualized, i.e. there are no physical resources which need to be purchased or stored and instead resources are accessed remotely.

Cloud computing services can be split into three separate categories, Infrastructure as a Service (IaaS), the dynamic provision of virtual machines and flexible on-demand data storage (Bokhari et al., 2016), Platform as a Service (PaaS), where a user can lease a development environment on which to develop, run and test their own software (Rimal et al., 2009), and Software as a Service (SaaS) such as email, web-conferencing, and word processing software that removes the need for an application to be installed on the host computer (Rimal et al., 2009).

Netflix's data storage solution

Netflix began as a dvd rental company. They built their own data centres and managed their own servers. In August 2008, they suffered a major database crash which resulted in their inability to ship DVDs to their customers for a three day period. As the company was expanding with a rapidly increasing customer base, the decision was made at a corporate level that Netflix would prefer to concentrate on providing videos to customers than on building an ever larger infrastructure to manage their customers' data. They began the switch to AWS in 2008, and had fully migrated all of their code and services seven years later in 2016, when they shut down the last of their physical data centres (Izrailevsky, Vlaovic & Mehsenberg, 2016).

There are three main components to Netflix's infrastructure. They employ AWS EC2 instances to deploy the back-end to their product, that is, everything you see and do

on the site up until a user clicks the “Play video” button. Amazon S3 databases are used to store the content users stream, thousands of hours of movies, documentaries and series. And once the video begins streaming, Netflix employs its own CDN network to transfer the video data from the databases to the user’s machine.

Benefits of Cloud Computing for Netflix

Scalability

Cloud computing is seen as an unlimited resource as it can be scaled horizontally by adding more virtual machines when they are needed. This allowed Netflix to concentrate on a rapid expansion of their customer base without having to consider whether their software infrastructure could cope with the extra workload or whether they would run out of resources like server space.

Affordability

Despite the initial cost of migrating all of their software and data, using the cloud has proven to be a more economical solution than building and maintaining their own networks. As one of the foundations of cloud computing is only paying for what you use, Netflix can cope with the increased streaming demand of just after dinner in the USA without having to pay for all of the resources necessary to cope with this at 4am in the morning when there is less traffic.

Assigning Development Resources

Netflix realised early in their streaming days that they were good at providing video streaming services to customers, but not as competent at building and maintaining networks of servers and storage devices. Since moving the infrastructure to the cloud, Netflix software engineers can dedicate their time to the user experience, improving the recommendation algorithm and streamlining the customer journey, rather than monitoring and maintaining their network, servers, and storage. This part of the business is now maintained remotely (at a cost) by AWS. (Ciancutti, 2014).

Fault Tolerance

Netflix state that their content library is stored across three different regions on Amazon S3 databases. This means that in the event of one of the regions experiencing difficulties, service can be restored for all customers by transferring the affected streams to originate from one of the unaffected sites. They test this process every month and say it can be completed in 6 minutes.

Potential Difficulties with Netflix approach

Vendor lock-in

One of the major drawbacks when migrating computing operations to the cloud is the problem of vendor lock-in (Opara-Martins et. al, 2016). Vendor lock-in is the phenomenon of feeling tied to one product or service because of the potential difficulties and costs that would be encountered should you try to access another

provider. As many cloud providers develop their own proprietary softwares for operating successfully in the cloud, the interoperability and portability between clouds can be compromised. The cost and complexity of switching from one cloud provider to another can be prohibitive and lead companies to be 'stuck' with their initial provider. In the case of Netflix, much of their architecture and systems have been purpose-built to work on AWS resources. This puts AWS in a strong negotiating position when contacts are being updated and leaves Netflix potentially vulnerable to changes in market conditions.

Fault Tolerance

Though fault tolerance has been described as a benefit of cloud computing systems, this is only the case when part of the cloud goes down, e.g. if the Dublin region of AWS S3 database fails, resources using that can be transferred to use the resources from another region within the cloud. However, there is the possibility of the entire cloud encountering difficulties or, as happened in AWS's EAST-1 region in 2020, the region could fall, and the transfer might not happen.
(https://www.theregister.com/2020/11/30/why_less_is_more_in_cloud/
)

Proposed Solution

Vergilio (2020) proposed a multi-cloud containerised architecture that would mitigate both of the drawbacks listed above.

Netflix have already stated that they use containers to deploy their code and they have invested heavily in their own container management platform, Titus (Joshi et. al, 2018). Rather than Netflix being almost entirely reliant on AWS to provide them with computing resources to deploy their code, it would be possible to spread their resources across the clouds of a number of providers. This would result in a massive improvement in fault tolerance as Netflix could move resources between different providers should any of them encounter difficulties.

References

Bokhari, M.U., Shallal, Q.M., Tamandani, Y.K., (2016) Cloud computing service models: a comparative study. In: IEEE Int. Conf. Comput. Sustain. Glob. Dev. INDIACom, pp. 16–18.

Buyya, R., Broberg, J., Goscinski, A. (2010) *Cloud computing: Principles and paradigms*, Chichester, UK: Wiley Publishing

Ciancutti, J. (2010) Four Reasons We Choose Amazon as our Cloud Computing Platform, The Netflix Tech Blog. Available from: https://netflixtechblog.com/four-reasons-we-choose-amazons-cloud-as-our-computing-platform-4aceb692afec [Accessed 1 July 2021].

Dean, B. (2021) Netflix Subscriber and Growth Statistics: How Many People Watch Netflix in 2021? Available from: https://backlinko.com/netflix-users#netflix-statistics [Accessed 25 June 2021].

Goodwins, R. (2020) AWS going AWOL last week is exactly why less is more in cloud server land. Available from: https://www.theregister.com/2020/11/30/why_less_is_more_in_cloud/ [Accessed 30 June 2021].

Izrailevsky, Y., Vlaovic, S. & Mehsenberg, R. (2016) Completing the Cloud Migration, The Netflix Tech Blog. Available from: https://about .netflix.com/en/news/completing-the-netflix-cloud-migration [Accessed 21 July 2021].

Joshi, A., Leung, A., Dwyer, C., Kung, F., Dhillon, S., Bak, T., Spyker, A., Bozarth, T., (2018) Titus, the Netflix container management platform, is now open source, The Netflix Tech Blog. Available from: https://netflixtechblog.com/titus-the-netflix-container-management-platform-is-now-open-source-f868c9fb5436 [Accessed July 23, 2021].

Kaur, K., Sharma, DR. S. & Kahlon, DR. K. S. (2017) Interoperability and Portability Approaches in Inter-Connected Clouds: A Review. *ACM Computing Surveys,* 50 (4) October, p. 49:1-49:40.

Krishnan, S., Tse, E. (2013) Hadoop Platform as a Service in the Cloud, *The Netflix Tech Blog*. Available from: https://netflixtechblog.com /hadoop-platform-as-a-service-in-the-cloud-c23f35f965e7 [Accessed 21 July 2021].

Laney, D. (2001), 3D Data Management: Controlling Data Volume, Velocity, and Variety. Technical report, META Group.

Opara-Martins, J., Sahandi, R. & Tian, F. (2016) Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective. *J Cloud Comp* 5, 4 (2016). https://doi.org/10.1186/s13677-016-0054-z

Rimal, B.P., Choi, E., Lumb, I. (2009) A taxonomy and survey of cloud computing systems. In: NCM 2009 - 5th Int. Jt. Conf. INC, IMS, IDC, pp. 44–51.

Vergilio, T. (2020) ‘A Unified Vendor-Agnostic Solution for Big Data Stream Processing in a Multi-Cloud Environment’, PhD Thesis, Leeds Beckett University. Leeds.

Ward, J., and Barker, A. (2013) Undefined By Data: A Survey of Big Data Definitions. *arXiv e-prints*, p.arXiv:1309.5821. https://backlinko.com/netflix-users#netflix-statistics [Accessed 17 July 2021]