**How Netflix Implemented the Six Phases of Big Data**

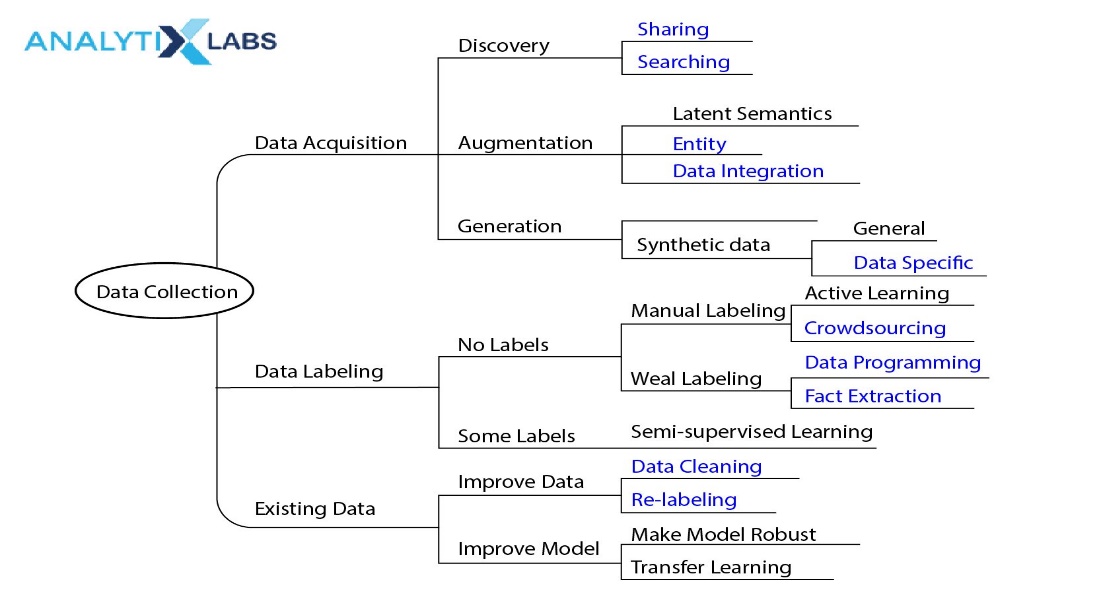
Netflix's incredible story, which started out as a DVD rental service in 1997 and has led it to become the largest video-streaming platform in the world, is evidence of its creativity and adaptability. Netflix has become a household name, revolutionising the way we access and consume entertainment content, with over 148 million paying subscribers worldwide (Quico, C. (2019)). The term "big data," which is popular in today's tech world, refers to the process of gathering and analysing enormous amounts of data in order to provide insights that can be used to guide decisions. Big data has been essential to its success despite the lofty claims and passionate debates about its effects on society and the industry. In order to give millions of users worldwide a smooth and customised streaming experience, Netflix uses the six phases of big data in an efficient manner, which are data generation, data acquisition, data storage, data analysis, data visualisation, and decision making. This study sets out to explore this process.

**Data Generation in Netflix**

A company that distributes audiovisual content on demand and is cognizant of the value of data starts its operations with a strategic business plan that goals are established, which will define their boundaries and KPIs, or key performance indicators, are another name for performance indicators. As stated in Ferní, E. P., et al. (2016), while applying this logic to Netflix, data generation come from understanding the users. Furthermore, van Es, K. (2023). also stated that businesses that are able to extract intelligence from data will have a deeper understanding of us than we have, enabling them to develop strategies that steer us in their direction rather than the direction we would take on our own if left to our own devices. Based on 1T data from 2016, the 81 million subscribers globally equate to 125 million viewing hours per day. They grow to such an extent within the organisation that they transcend the purely numerical. In actuality, every one of these hours spent using an internet-connected device (streameding) is associated with distinct information batches that, when managed and examined, assist Netflix in determining its objectives and understanding its target market. This allows them to come out with a design about the flow of data management process with their aspirations.

As mentioned before Netflix generates a vast amount of data through certain user interactions such as rating and reviews, content metadata, playback intensity and characteristics (Ferní, E. P., et al. (2016)), search queries, and recommendation system’s feedback (Maddodi, S. (2019)). In study of Ferní, E. P., et al. (2016), they mentioned the information that users voluntarily provide such as by rating content, completing enrollment forms or choosing to register using their social profiles is testimonial when compared to the information that is generated, but not voluntarily. When a user uses this monitoring mode, which is conducted in their own surroundings and is not detected by any connected devices, useful unconditioned responses are obtained that allow information about the users' reasons for watching different types of content to be gathered and generated. Besides that, as mentioned in Maddodi, S. (2019), most recommendation systems consider the user's profile to be a crucial factor. A subscriber's interest, past search query history, system interactions, and other information are all included in their subscriber profile. Upon creating a new account or adding a profile to an already-existing one, Netflix will request that the user select a few genres or titles that will serve as the starting points for the recommendation system. If the user bypasses this step, Netflix will automatically load the most popular content onto the user's homepage and this has automatically help Netflix to generate another data.

**Data Acquisition in Netflix**

According to Lyko, K. (2016), collecting, filtering, and cleaning data prior to storing it in a data warehouse or another type of storage solution has been referred to as data acquisition. The four Vs that most frequently govern the acquisition of big data are volume, velocity, variety, and value. As a pioneer in the streaming entertainment sector, Netflix prioritises data acquisition as a key component of its business model. In Netflix's case, data acquisition usually entails obtaining and analysing massive amounts of data from multiple sources in order to gain insights and enhance its offerings.

Source: analytixlabs

In contrast to traditional television, which used small sample sizes to evaluate viewer behaviour, Netflix gathers real-time data about all of its subscribers' viewing habits. The potential of big data to capture the entire domain and make precise and impartial calculations regarding taste has been heightened by this later fact of scale (van Es, K. (2023)). As stated by (Sarandos, 2012), "With streaming, we have insight into every second of the viewing experience, I am aware of the things you have tried and disabled. I am aware of the moment you disabled it. It's extremely advanced”. Thus, this has help Netflix to claim themselves as the company as “know their audiences” in their ways of traditional television did not.

Besides, van Es, K. (2023) has mentioned in his study that A/B testing is a key component of Netflix's data collection and decision-making process. They experiment with various platform variations and collect data through A/B testing to help inform their service improvements. This strategy is in line with their dedication to the scientific method and enables them to validate or invalidate concepts and theories, keeping their platform flexible and adaptable to the demands and preferences of its users. This testing involved filtering and cleaning the data as A/B testing divides user interactions and behaviours into groups according to particular platform iterations, which helps filter and refine data. This procedure guarantees that the information gathered for A/B testing is extremely pertinent to the specific modifications or features under evaluation. In line with Netflix's dedication to data-driven decision-making and continuous improvement, the filtered data is then analysed to make wise decisions and improve the user experience overall.

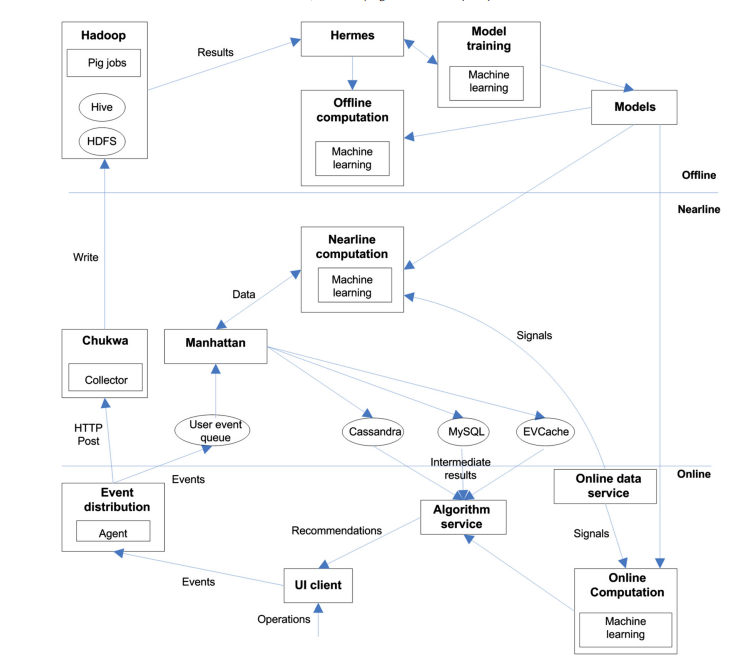
Lastly, the technical layer of data handling for Netflix is called extract, transform, and load, or ETL. It involves the mass data (big data) management process where quality (veracity characteristic) and other technical characteristic-related criteria are applied. This makes it possible to use data that is helpful since it aligns with business objectives and does not contain errors (duplicate, erroneous, or invalid data) before proceed to store the data (Ferní, E. P.,2016). The loading process is framed within the data architecture and defines such aspects as data integration models.

**Data Storage in Netflix**

Large volumes of information can now be stored and generated by businesses thanks to new technologies, which also place a high demand on highly specific technological actions. As a matter of fact, when people and companies interact with new technology, data is generated and we need to have a storage to store all the insightful information. However, this might be the result of a volume problem, in which case working with actual mass data storage (data centers) or the cloud will be required (Ferní, E. P.,2016).

The foundation for data analytics, decision-making, and service enhancements is laid by this stage of data storage. According to Lyko, K. (2016), Data Storage is the persistence and management of data in a scalable way that satisfies the needs of applications that require fast access to the data. As mentioned earlier, after the data went through the ETL process during Data Acquisition process, it’s time to store the data into a storage system. As mentioned in (Ferní, E. P.,2016), There are two options to store the data, either it stored in a huge data repository, can be also known as data warehouse or it can be stored in the cloud. AWS is the cloud infrastructure where Netflix keeps all their processing and cleaning data. As stated in (Tse, 2015), monthly streaming volumes of nearly one billion hours are made possible by this large-scale management. Thus, the importance of effective data storage in assisting Netflix's operations and service delivery is highlighted by this large-scale management. The availability and accessibility of data, regardless of whether it is stored in a traditional repository or the cloud, is essential for decision-making, ongoing user experience optimisation, and meeting the unquenchable demand for streaming content.

**Data Analysis in Netflix**

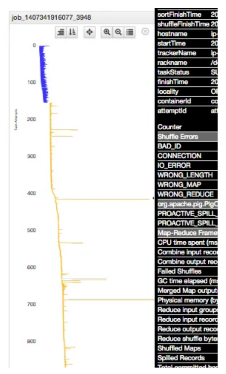
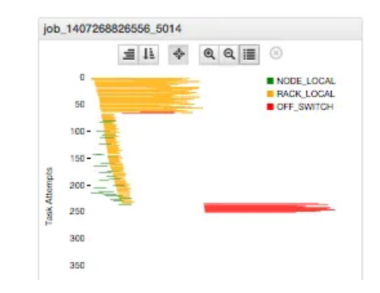
After a wealth of data has been gathered and store, the next critical stage in Netflix's data journey is data analysis where raw information is transformed into actionable insights. Data analysis is essential to create the fitted, data-driven content recommendations that have come to define Netflix as it continues to serve its global audience. As mentioned by Lyko, K. (2016), Data Analysis is concerned with making the raw data acquired amenable to use in decision-making as well as domain-specific usage.

Source: Pääkkönen, P., & Pakkala, D. (2015)

As shown in figure above, Netflix's data analytics infrastructure, which is separated into online, nearline, and offline components to meet different real-time requirements. It runs entirely within Amazon's cloud environment. With a focus on minimising response latency to client applications, online computation manages user operations. Similar to online computation, nearline computation enables the storing of computed results for later use. Timing requirements are most relaxed for offline computation. End users interact with Netflix services in a variety of ways, including playing and rating content. Data, models, and signals are used to generate recommendations based on user behaviour (Maddodi, S. (2019)). Data is pre-processed information stored in a database, whereas signals are live data from online services. Models are typically fine-tuned using incremental machine learning after they have been trained offline. For offline processing, user events are routed through the Chukwa framework, while nearline processing is routed through a user event queue (Pääkkönen, P., & Pakkala, D. (2015)).

Furthermore, as mentioned in Pääkkönen, P., & Pakkala, D. (2015), Netflix makes use of two streaming data sources: an online data service and Netflix user events, both of which are treated as streaming data. Chukwa agents extract the streaming data and Chukwa collectors temporarily store it, with Hadoop HDFS serving as a raw data store. Deep analytics are represented by offline Pig jobs. Hermes is a publish/subscribe system for storing and retrieving results of offline analysis. The user data queue is used as a temporary storage location for streaming data. For nearline computation, data transformation through Offline Pig jobs is considered stream processing, and intermediate analysis results are stored in stream data stores. A stream analysis process is used to model online and nearline computation. Netflix UI clients use the algorithm service as a data store. Besides, models are trained using machine learning algorithms.

**Data Visualisation of Netflix**

The ability of Netflix to transform massive amounts of data into actionable insights is built on data visualisation. Netflix relies on the power of visual representations to make sense of user behaviour, content performance, Job performance and more in an era where information is abundant but attention spans are limited. Finding a job and the corresponding Hadoop resources does not make understanding the performance any easier. The execution of stages of a Hive or Pig script may be serial or parallel, affecting the total runtime. Thus, visualisation is very important in monitoring job performance for every single resources.

**Diagram showing Task Details & Execution Locality**

Source: NetflixTechBlog

For example, a detailed task diagram for each job is shown above the workflow diagram, showing the individual execution of each task attempt. Laying these out in chronological order demonstrates how tasks were assigned and completed. This visual presentation can quickly communicate obvious job issues such as data skew, slow attempts, inconsistent resource allocation, speculative execution, and locality. Users can quickly scan the behaviour of many jobs for problems by visualising job performance in this compact format. Tasks are ordered according to scheduler allocation, which provides information about how many resources were available at the time and how long it took for the attempt to begin. The colour indicates the type or status of the task. Failed or killed tasks even display the failure reason and stack trace, eliminating the need to dig through logs. If you want to look at a specific task log, simply select the task and click the link provided to go directly to that task's log.

**Decision Making of Netflix**

Netflix is a data-driven company that uses business analytics, machine learning, artificial intelligence, and deep learning (Costa, 2020). Its success is dependent on its ability to run analytics across all business functions in order to provide insights to various stakeholder groups such as members, executives, and partners. Smith et al., 2019 stated that Netflix's goal is to give content creators the opportunity to realise their most creative ideas and entertain through innovative storytelling. However, it need carefully selects content for its audiences. Thus, data in the big data is used for decision-making and predictive analytics in areas such as determining which titles to stream, which content to licence or produce, customer journey analytics from prospect to member and what movie we should recommend to every audience.

Netflix will fine-tune its recommendation engine algorithms as well as adapt to changing consumer needs and desires. Netflix will continue to offer low-cost subscription plans while aggressively gaining market share and brand loyalty. It will become a major content creator, creating millions of jobs and attracting millions of viewers. However, it's important to note that Netflix's big data-driven approach isn't without ethical implications. Metrics such as viewing hours and content promotion raise concerns about striking the right balance between maximising user engagement and delivering responsible content, a challenge that is inextricably linked to big data.

Reference:

Quico, C. (2019). Television reshaped by big data: Impacts and implications for Netflix-like platforms in the age of dataism.

Ferní, E. P., Neira, E., & Clares-Gavilí, J. (2016). Data management in audiovisual business: Netflix as a case study. Profesional de la Información, 25(4), 568-577.

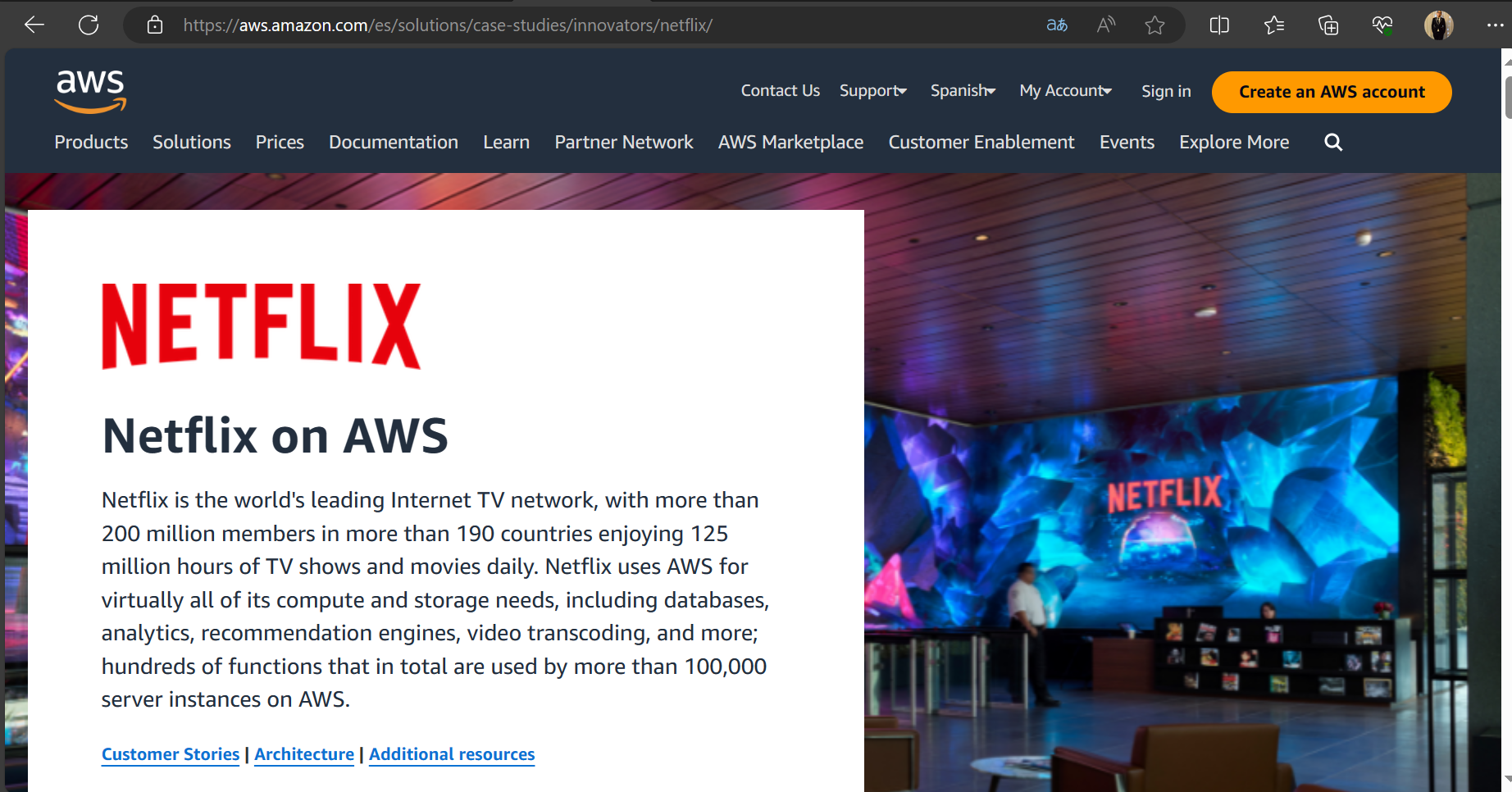
Maddodi, S. (2019). NETFLIX bigdata analytics-the emergence of data driven recommendation. Srivatsa Maddodi, & Krishna Prasad, K.(2019). Netflix Bigdata Analytics-The Emergence of Data Driven Recommendation. International Journal of Case Studies in Business, IT, and Education (IJCSBE), 3(2), 41-51.

van Es, K. (2023). Netflix & Big Data: The Strategic Ambivalence of an Entertainment Company. Television & New Media, 24(6), 656-672.

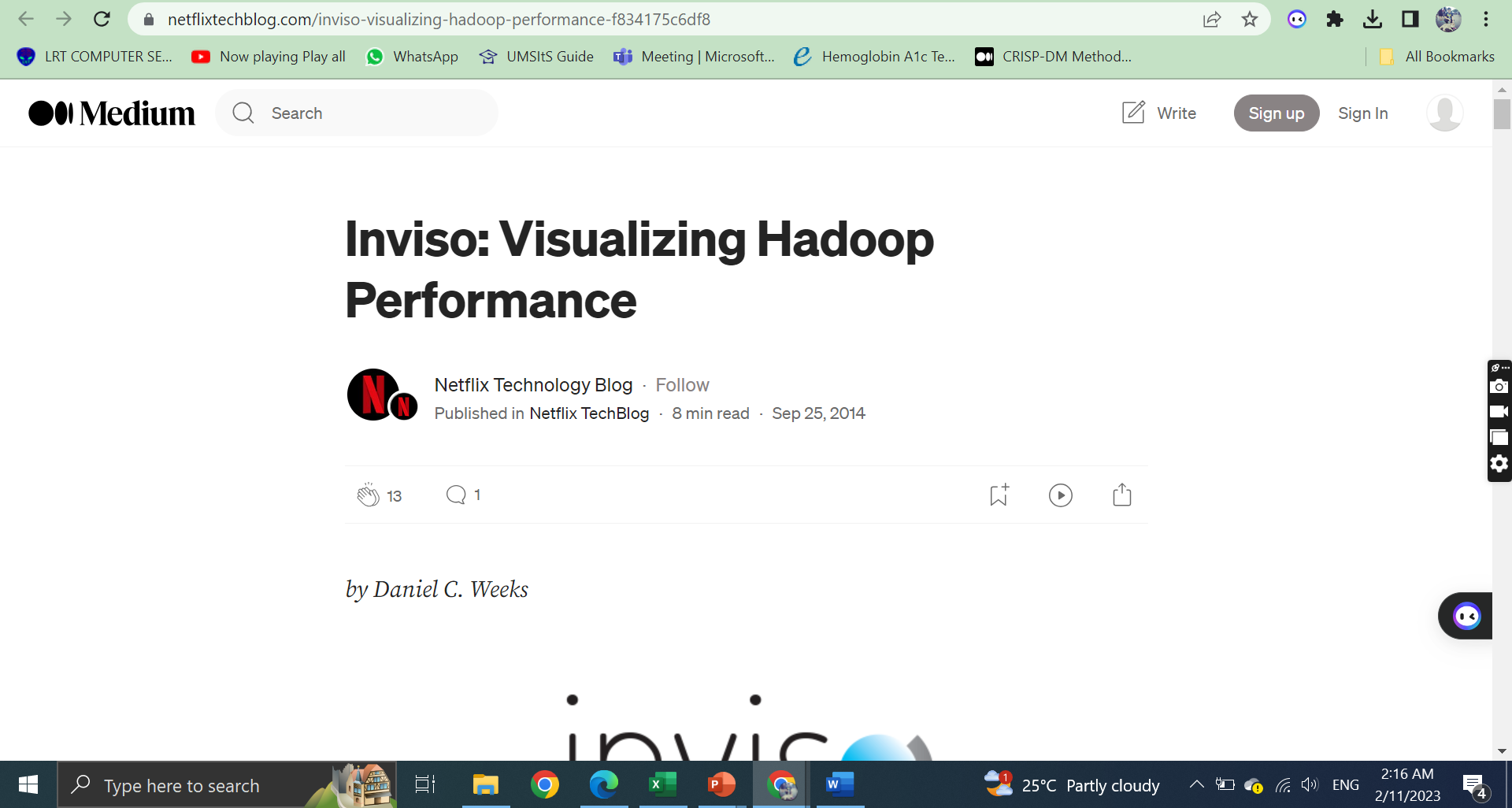
Lyko, K., Nitzschke, M., & Ngonga Ngomo, A. C. (2016). Big data acquisition. New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe, 39-61.

Pääkkönen, P., & Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems. Big data research, 2(4), 166-186.

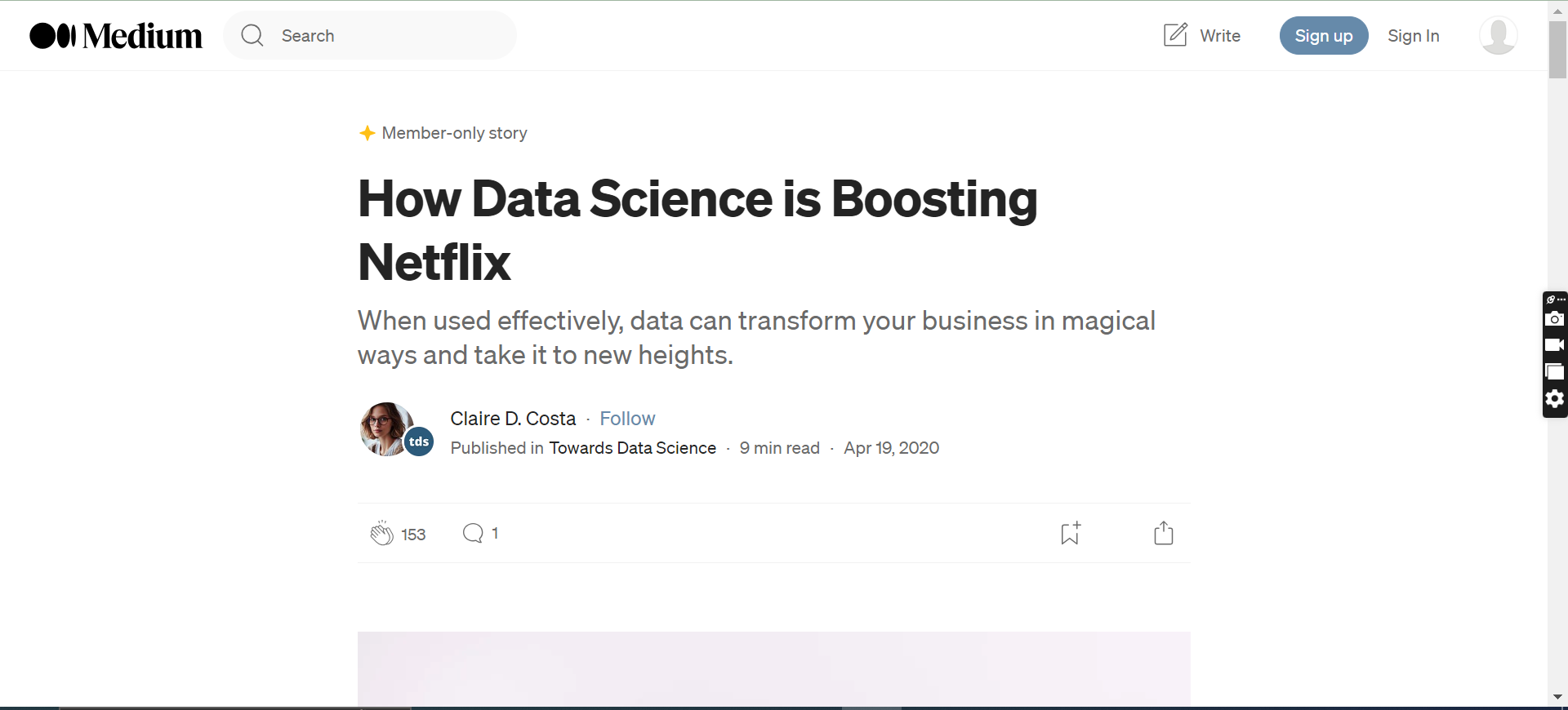
Tse, Eva (2015). “Netflix case study”. Amazon web service. <https://aws.amazon.com/es/solutions/case-studies/netflix>



<https://netflixtechblog.com/inviso-visualizing-hadoop-performance-f834175c6df8>



Costa, C. (2020). How Data Science is boosting Netflix. Towards Data Science. <https://towardsdatascience.com/how-data-science-is-boosting-netflix-785a1cba7e45>



Smith, M. Telang, R. (2019). Netflix and the economics of bundling. Harvard Business Review. <https://hbr.org/2019/02/netflix-and-the-economics-of-bundling>

