Cheng-Han Yang, Chia-Lin Hsieh, Rahaf Alhazmi, Weixin Tang

Professor Yao Shen

CS 644 852

February 5 2020

Ads Based on What You Buy

According to Mckinsey Company, “90 percent of the digital data ever created in the world has been generated in just the past two years, ...”. We are living at the age of the most convenient way to generate and store data. How to efficiently utilize those data to improve human's life facing huge challenges, but also bringing with many opportunities. In the following passages, we will analyze how to use big data to optimize advertising, and our big data advertising research will surround Amazon. Besides, we will work on our research-based on the following aspects of how big data advertising works on Amazon, four main characters(4 V's) in our big data domain, Big data-related questions, existing solutions, the solution proposed by our team.

1. Application Domain

Ever look at an item on Amazon and then later see an AD for that same item on another website or be surprised with a recommendation that matches our hobby. Amazon accomplishes these feats by mining data from its million consumer accounts to recommend products based on user behavior. The company keeps track not just of what customers buy but also what items save, view, and even what they remove from their carts. Data from other Amazon own sources can reveal what shows movies and books people prefer, what they say to Alexa as well as what external sites they view that contain Amazon Ads combined with demographic data like age, gender, payment type, and location that is tied to every user’s account. Amazon can build a picture of a customer’s preferences; they use all these data to optimize how billion items are distributed to fulfillment centers around the world. If their data indicates a given county contains lots of new parents for example, they might stock a popular brand of the diaper at the closest distribution center or if an individual customer has added and removed a specific model of TV from their cart several times, Amazon might recommend a cheap flat screen to that customer and even pre-ship it to a center nearby, so next time we get the feeling that an online retailer like Amazon is reading our mind. Actually, they are only reading our data.

How does Amazon recommend our next purchase? The answer is big data. How does big data help the recommendation engine of Amazon? It involves three stages. First is events, Amazon tracks and stores data on all customer behavior and activity on the site. For every click the shopper makes, a record of the event is logged in the database, the entry is stored as something like “userA clicked a productX details once”. Events are captured for all kinds of actions like users liking a product, adding product to cart, and purchasing a product. The second is ratings which are important because they reveal what a user feels about a product. Recommendation systems can assign an implicit value to different kinds of user actions. For

example, the maximum rating is five stars, Purchase is four stars, Like is three stars, Click is two stars, and so on. Recommendation systems can also take into account ratings and feedback users provide. The third is filtering products which is based on ratings and other user data. Recommendation systems use three types of filtering. The first is collaborative filtering. All the visitor's choices are compared and they get a recommendation. For example, if a userX likes products A, B, C, and D; and a userY likes products A, B, C, D, and E. Then it's likely userX will also like product E. Second is user-based filtering, the user's browsing history, likes, purchases, and ratings are taken into account before providing recommendations. The third is the hybrid approach. Many companies use a hybrid approach, which consists of both the mentioned approaches.

2. Four V's of Big Data

Volume:

Big data technology giants like Amazon get real-time, structured, and unstructured data, lying between terabytes and zettabytes every second from millions of customers especially smartphone users from across the globe. They do near real-time data processing and after running machine learning algorithms to do data analysis on big data, they make decisions to provide the best customer experience.

Velocity:

Increasingly, businesses have stringent requirements from the time data is generated, to the time actionable insights are delivered to the users. Therefore, data needs to be collected, stored, processed, and analyzed within relatively short windows – ranging from daily to real-time.

Variety:

Includes data from a wide range of sources and formats (e.g. web logs, social media interactions, ecommerce and online transactions, financial transactions, etc).

Veracity:

Veracity refers to the quality of the data that is being analyzed. High veracity data has many records that are valuable to analyze and that contribute in a meaningful way to the overall results. Low veracity data, on the other hand, contains a high percentage of meaningless data. Data that is high volume, high velocity, and high variety must be processed with advanced tools (analytics and algorithms) to reveal meaningful information. Because of these characteristics of the data, the knowledge domain that deals with the storage, processing, and analysis of these data sets has been labeled Big Data.

3. Big Data Related Questions

Data silos

A data silo is a repository of fixed data that remains under the control of one department and is isolated from the rest of the organization, much like grain in a farm silo is closed off from outside elements. Data silos tend to arise naturally in large organizations because each organizational unit has different goals, priorities and responsibilities. (Definition from TechTarget) To be more specific, according to the different goals between each department and the different habits of each employee, the exact same data will get inconsistent labels while processing by different hands. Furthermore, when the same data is stored under different labels, the impact is not only the storage space is prodigally occupied, but also derived many versions when different regions update new information, leading following users to confusion during subsequent access.

Having numerous users all around the world is one of the reasons why Amazon dominates e-commerce. While a large amount of data gives Amazon an advantage, countless problems also plague it, such as how to store those data. If Amazon chooses to store the data in a single location, it will make the data transmission in other regions too slow. The server security is low and space is easier to be full. Hence, just like other renowned companies such as Google, Amazon set up servers spread all over the world to store and back up all precious data, but the resulting problem is that scattered servers may cause data to be out of sync. At Amazon's scale, a miscalculated metric, such as cost per unit, or delayed data can have a huge impact.

Privacy

Although big data has brought customized advertising to help us more easily buy the products we want, but behind the convenience is the sacrifice of personal privacy information, such as location information, browsing history, personal photos, etc., which will permanently stored in Amazon's database. Therefore, many users may use laws and other means to ask Amazon to stop collecting their information.

Data Security

When the user agrees to provide information to Amazon as accurate personalized advertising and other services, Amazon is obliged to ensure that these data are not abused. This data security includes protecting the data stored in the server from being accessed by outsiders, being used for other private purposes, and being exposed to its confidential content.

Shortage of Skilled People

With the development of internet technology more mature, data transfer nowadays is growing exponentially. Compared with the data processing talents are completely insufficient, whether it is to analyze data or solve program vulnerabilities, it needs professional people to handle it, not only powerful The accumulated experience of background knowledge is even more important.

As it is one of the companies required to handle the most information in the world, Amazon especially lacks talents.

4. Existing solutions

Data silos

As data soils have different problems, such as unsynchronized data and inconsistent data, some solutions might solve them. One of these solutions is the data lake, a centralized repository that lets you store all your data, whether structured or unstructured. So, it will collect all data into one location. You can also create links between information with different label names but represent the same thing to solve the inconsistent problem. Also, there are other benefits of the data lake.

In 2019, Amazon decided to build a data lake named the Galaxy data lake to help it overcome its data soil problems and improve different areas. Especially the advertisement area because data lake gives Amazon the ability to analyze the data easily. To explain that, data lake stores all kinds and formats of data, making it easy to run different analytics types to uncover insights and come up with suitable customized ads. In the end, The Galaxy data lake might be the perfect solution for Amazon.

Data Security

Every company needs to secure their customers data, and they need to make sure that data transformation is processed securely, which means that the data cannot be read or taken by any third party.

Amazon provides strong encryption for its customers data, whether the data in rest or transit. Also, all its header information is obscured to ensure that the website servers details are hidden and prevent hackers from attacking its resources and making it difficult to intercept its information. Moreover, Amazon uses encryption protocols and software to ensure that the data are secured during the transmission, which is part of the customized ads process. Finally, Amazon applied many other security techniques to secure its customers data.

Privacy

There are many possible solutions to the privacy problem. For example, companies can stop using customers data or limit the data collected from their customers. Therefore, companies might solve part of the privacy problems. But Amazon took another path to solve its privacy problem, and it might not be considered a solution for some of its customers. First, Amazon describes what data are collected and to whom on its website, and the purposes behind using its customers data. Then, it clarified that you accept the practices described in its Privacy Notice by using its services. Also, Amazon stated that it uses state of the art privacy-enhancing technology. This privacy-enhancing technology is designed to support privacy and data protection. This way of solving Amazon’s privacy problems associated with the customized advertisements might not be the desirable solution. However, Amazon offers an option to its customers that you can choose not to provide your information. Still, as a result of that, you will not have full access to Amazon’s services advantages, such as the personalized advertisements.

Shortage of Skilled People

Nowadays, manufacturers are facing the challenge of building and maintaining a skilled workforce. Amazon is one of the manufacturers that is facing a shortage of qualified people. Therefore, Amazon needs to fill this gap, and most of this gap is related to a lack of technical skills. How will it overcome this shortage of skilled people? There are two actions that Amazon has taken.

First, Amazon uses many benefits to retain and attract talents, including raising the minimum wage to $15, provides up to 20 weeks of vacation for delivery mothers, including a comprehensive follow-up rework plan, and many benefits applicable to all full-time or part-time workers.

The second action is launching training programs. One of these programs is to upskill 100,000 of its employees in the United States. For this program, Amazon announced that it plans to spend $700M to retrain employees for two roles: software developers and data scientists. Finally, Amazon intends to upskill its employees to keep up with its demands to reach their goals.

5. Solution Proposed By Your Team

We will present our team solution from a data scientist perspective. Providing our idea to grapple with the questions mentioned above and explaining how we handle the big data advertising problem, and our solution will be surrounded with 5 p's.

Compared with Amazon, which uses a private data lake to store all data as a centralized concept, our group believes that using decentralized blockchain technology to store data will have better results. To begin with, when the number of users of the blockchain has accumulated to a certain level, the speed of data upload and download will be equal to or even surpassing Amazon's data lake. In addition, if the hacker cannot control more than 50% of the users' devices at the same time, none of the data content could be tampered with. Last but not least, under the SHA encryption algorithm, even a single piece of data is intercepted, a Brute-force attack could not break it and get its information, which ensures the security and privacy of the data.

Propose

To achieve this challenge application, our team would have a clear goal. Optimize users' searching results when the users try to find their ideal product. Present the product most likely bought by the user based on users preference which is calculated by our model. While a user is browsing a product, the system would make a quick decision and recommend a similar product that may be more suitable for this user or other related product may want it. Overall, our team's ambition is to increase product selling and saving customer's searching and finding the time. So, both seller and buyer would gain happiness with the assistant from our application.

People

Since it's a challenging project. Handling large amounts of diverse data would not just require using big data related techniques but also need to know a related scientific and business analysis. Hence, we need to build a team with diverse skills including statistics, programming, and market selling. All of the teammates are required to know the fundamental knowledge in all those areas, besides each of the teammates have to focus on developing to be one of the domain experts.

Process:

Data Collection:

The data we use from the user's history track, personal information, and product's relevant information to predict the user's preferences. Those history data would include users' search history, product browsing history, shopping history. Personal information would include the customer's age, sex, address, payment method, etc. Product information may include product name, brand, category, price, production rates, storage location, available condition, and so on. We may ignore using customers' evaluation of the product and using the product rating to build our model. Since those comments are pretty unorganized. We currently have a limited ability in working on natural language processing. So, we will leave it for future development.

Data Cleaning

Furthermore, in order to make sure the quality of the data we collected. We will largely reduce the weight of the data in our model if it's too old. Besides, we would remove some extreme cases. For example, some users have unusual behavior. They intentionally give terrible feedback to almost all the products they bought. Or, they bought a lot of products and returned those products. In analysis statistics, we would remove those outliers for better analyzing our model.

Data Process/ Analysis

In processing the giant data, we would try to use some classification and clustering techniques. We would classify users into multidimensional groups based on what product they bought, which product they browsed before. We would classify products into different categories, prices suit different groups of people. So, that product information could be used to map customers with more suitable products. Moreover, we will try to figure out the relationship among products. We call the valence character of the data. For example, some customers may buy baby bottle nipples. They may also need diapers. A complex algorithm or advanced model would be hard to implement in a distributed system with function parallelize. So, we could start from a simple model, even if a simple model couldn't reduce some bias with increasing amounts of data.

Platforms:

Instead of managing our server machine, we would use Iaas (infrastructure as a service) provided by GCP (Google Cloud Platform). So, we don't worry about the hardware and focus on design and coding our project. Besides, GCP allows us to easily scale up our service if our business increases rapidly. We could deploy the server close to our customers to reduce latency and increase customer's experience.

Programming ability :

We would prefer using the technique Hadoop with Spark. Using the Hadoop file system, the data would be triple duplicated and saved in multiple data nodes. Such a redundancy structure provides features of fault tolerance. So, we wouldn't worry much about single node failure. User search feedback shouldn't take too long. If it takes a long time, maybe a few seconds, we will lose opportunities. So, we prefer to use spark which saves the data in memory rather than the traditional relational database which is saved on disk. We would make sure our algorithm is suitable for MapReduce, so it can function parallel instead of task-parallel.

In conclusion, we observed some methods or mechanisms Amazon used to improve their advertising service and maximize their profit. We feel how big the difficulty is through knowing four V's and related big data obstacles or problems. In the end, we analyze the existing solutions we found, and summary our solution proposal from 5 p's. We believe huge challenges bring large opportunities. Through effectively using big data, humans' life quality could be unexpectedly improved, and the application built by our team will have great achievement.

Works Cited

"Amazon.com Privacy Notice" , *Amazon*, January 1, 2020 , https://www.amazon.com/gp/help/customer/display.html?nodeId=GX7NJQ4ZB8MHFRNJ

Greg Petro, "The Talent War: Walmart And Amazon Compete For A Better Workplace", *Forbes* , Feb 21 2020 , https://www.forbes.com/sites/gregpetro/2020/02/21/the-talent-war-walmart-and-amazon-compete-for-a-better-workplace/?sh=5d2b7545a536

Heather Landi, "Amazon plans to spend $700M to retrain a third of its workforce for data, analytics roles", *Fierce Healthcare*, Jul 12 2019 , https://www.fiercehealthcare.com/tech/amazon-plans-to-spend-700m-to-retrain-a-third-its-workforce-for-data-and-analytics-roles

Jennifer Wills, "6 Ways Amazon Uses Big Data To Stalk You" , *Investopedia* , Oct 5 2020, https://www.investopedia.com/articles/insights/090716/7-ways-amazon-uses-big-data-stalk-you-amzn.asp

Nick Heath, "Tech skills shortage: Amazon trains veterans to fill the gap", *TechRepublic*, January 12 2017, https://www.techrepublic.com/article/tech-skills-shortage-amazon-trains-veterans-to-fill-the-gap/

Nicolaus, Ari, et al. "Straight talk about big data" , *Mckinsey* , October 28, 2016 , https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/straight-talk-about-big-data

"Protecting data privacy", *Amazon*, July 9 2018 , https://www.aboutamazon.com/news/amazon-ai/protecting-data-privacy

UpGuard Team, "How does Amazon handle cybersecurity?", *UpGuard*, Aug 5 2020, https://www.upguard.com/blog/prime-day-how-amazon-handles-cybersecurity

Werner Vogels, "How Amazon is solving big-data challenges with data lakes" , *SiliconAngle*, January 30 2020, https://siliconangle.com/2020/01/30/amazon-solving-big-data-challenges-data-lakes/