Scalable Ad-Targeting Technology using AWS

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Nitin Maddi
Spring 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Rosanne Vrugtman, Department of Computer Science

ABSTRACT

Currently, companies collect massive amounts of data on their users but have no effective way to digest and utilize this information without investing millions of dollars in infrastructure. By using AWS cloud technology, I was able to develop an efficient ad-targeting solution that costs a fraction of the price typically required. I utilized AWS Simple Queue Service (SQS) to simplify and orchestrate the data analysis workflow by breaking down the problem into smaller parts. With the help of Java, these smaller processes are examined to match people to an advertisement based on the information collected. Finally, these advertisements are pushed to an advertisement platform, such as Google or Facebook, and a person's account on these platforms will receive the advertisement. The finished product allowed companies to deliver targeted advertisements at a fraction of the cost, and not have a max cap on the amount of data being processed. Next steps would include adding more advertising platform options and further increase the efficiency of the data analysis process.

1. INTRODUCTION

Currently, companies collect data on their users in any way they can. For example, E-Commerce companies track and keep records of items a customer adds to their cart or items the person just hovered over (Subramanian, 2022). They can even track data that represent a customer's current behavior by analyzing downgrades in subscriptions, order cancellations and much more. With this data, they detect which customers are most likely to buy certain products, and the company can send targeted advertisements to them through platforms such as Google and Facebook. Although this analysis is extremely powerful, the amount of data the company generally collects can be on the scale of petabytes which is an impossible amount of information to work with. To evaluate that amount of data, the company would need to invest millions of dollars into a data analysis platform to do in-depth data analysis on such a large scale. Additionally, separate pipelines would need to be developed for each advertisement platform.

2. RELATED WORKS

Previous research regarding ad-targeting systems provides insight into the generic techniques utilized in the field and data analytics necessary to develop effective systems. Iyer, et. al (2005) write about the general theory behind the process and how companies can collect the data on their customers in various ways. Furthermore, they discuss "the ability to target specific segments within a market," which is the key fundamental behind all ad-targeting software. Overall, this research provided a high-level overview of how an ad-targeting system would work in practice. Salesforce's system takes many of the principles explained by Iyer and executes them at an expert level. Every technique aimed to increase efficiency is perfected, and the scalability is much greater. The entire system is hosted on the AWS cloud, which gives it almost an unlimited capacity. The benefits of being a cloud service include a wide range of tools that allow for simple implementation, which was not discussed in this article.

A newer addition to the field of ad-targeting is machine learning. Choi and Lim (2020) discuss machine learning techniques which could increase the advertisement match rate

while simultaneously decreasing fraud. They suggest a deep belief network combined with logistical regression in order to maximize the relation of user information and click logs, but Salesforce decided to go with a much lighter weight algorithm, k-nearest-neighbors, due to the throughput the system needed to handle.

3. PROJECT DESIGN

My project utilized Amazon Web Services (AWS) to act as the frame of our solution. AWS is a cloud service provider, meaning it can provide as much power and storage as necessary for our system to run, while we pay on an hourly basis based on the amount of resources used (Barney & Gillis, 2022). No matter how much data is sent through this system, my team was able to automatically scale up resources on an as-needed basis.

In addition, my team utilized AWS Simple Queue Service (SQS) to simplify and orchestrate the data analysis workflow. SQS allowed decoupling of the various systems that interconnected to build this pipeline to prevent bottlenecking from occurring. This process started by splitting the original ad-targeting job into hundreds of smaller-sized processes. With the help of Java, these smaller processes are examined and linked to the advertisement most suitable based on previously conducted data analysis. An issue arises here, as the only identifying information we have are simply a list of email addresses and names that are sent from the company using our product. Therefore, the system matches these accounts to real accounts, such as a Google or Facebook, on an advertisement platform. This is done by creating an Audience, which is essentially a list of email addresses to which a specific advertisement will be sent.

From here, our system will communicate with a special class, known as the PartnerAPIRepository, which will link our request to a specific advertisement platform the customer wishes to push advertisements to. This abstraction is necessary to allow for greater scalability on the advent of new advertisement platforms to be added. The PartnerAPIRepository will make a request is made to Google, Facebook, or any other platform, and the email addresses in the Audiences will be linked to accounts in the corresponding advertisement platform.

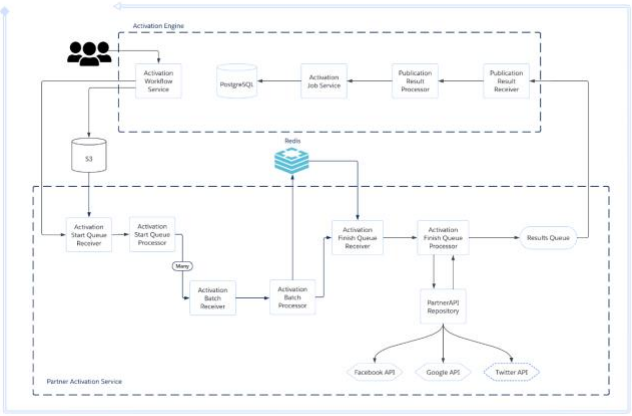

Figure 1. System Architecture with SQS Queues

Now, the specific advertisement that was chosen for these accounts will be pushed, and it will show up when the account is logged into. After the results of an advertisement batch are published back to the Audience, we retrieve that information and store it in our own internal database. This data is integral to provide to our customers to show them how well their advertisements are doing, as well as for our own analysis on how well our data analysis is working. This process follows a similar structure, and uses SQS to decouple the process, and finally input the data to our PostgreSQL database.

4. RESULTS

The results of this system vary greatly, as a variety of different companies utilize this system. Each of these companies pays a different amount, and each previously spent different amounts of money on ad-targeting.

However, the ad-targeting efficiency, which was calculated by the number of clicks divided by overall advertisements sent, increased by 40% with the new data analysis workflow. Additionally, the newly-designed system withheld the stress of nearly double the amount of load than if we were to use on-premises servers rather than AWS.

5. CONCLUSION

This system allows companies to have access to high quality targeted advertisements without needing to spend millions of dollars in capital and maintenance fees. Additionally, the process of sending targeted advertisements has been simplified and streamlined to essentially let a company send their data and automatically push advertisements to users. Implementing an efficient system required elaborate planning and architecting, while being supplemented by industry leading technology, such as AWS. The meticulous engineering also provided easy ways to expand the system in the future, allowing for other advertisement platforms to be added if necessary. Overall, this project introduced me to real-life code bases with millions of lines of code and displayed the importance of planning a design carefully, as making changes down the line is much more difficult than designing for a situation from the beginning.

6. FUTURE WORK

There are many different tasks to complete in order to further improve. Some of the more important one would be to add the Twitter advertisement platform, because currently the only available platforms are Google and Facebook. This would expand the reach of the customers who use this service and make the product more accessible to a wider range of users. Another improvement would be to make the system more efficient. Some possible ways of achieving this would be to use Caches, such as Redis, to provide reused results much faster than recomputing the entire process. Also, analyzing the bottlenecks of the system, such as the data processing section, and working to unclog these sections would prove useful for the future. Last, the data analysis that is currently being employed works by conducting rather simple analyses such as K-means or K-Nearest Neighbors. By implementing more sophisticated algorithms, such as Neural Networks, the click count and overall advertisement efficiency may increase further.

REFERENCES

[1] Choi, J.-A., & Lim, K. (2020). Identifying machine learning techniques for classification of Target Advertising. *ICT Express*, *6*(3), 175–180. https://doi.org/10.1016/j.icte.2020.04.012

[2] Iyer, G., Soberman, D., & Villas-Boas, J. M. (2005). The targeting of advertising. *Marketing Science, 24*(3), 461+. https://link.gale.com/apps/doc/A137861142/AONE?u=anon~cf7e64bb&sid=googleScholar&xid=a2705892

[3] Barney, N., & Gillis, A. S. (2022, October 6). *What is AWS (Amazon Web Services) and how does it work?* SearchAWS. Retrieved October 30, 2022, from https://www.techtarget.com/searchaws/definition/Amazon-Web-Services