Thesis Project Portfolio

Amazon Advertising Data: An Automated AWS Pipeline

(Technical Report)

How Memory in History and Technology Shapes Perspectives on Safety and Security

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Johnathan Middleton

Spring, 2024

Department of Computer Science

Table of Contents

Executive Summary

The technical aspect of my thesis focuses on a advertising pipeline I created during an internship at Amazon. This involved the usage of various Amazon Web Services (AWS) tools to create a system that would take in advertising metrics as inputs, aggregate these based on predetermined metrics, and publish a business report that a finance team is able to analyze. In the technical report, I detail the high-level design of the pipeline, which involved two major sections, firstly the components of the pipeline, which is the infrastructure of AWS that the pipeline would be orchestrated upon, as well as the scheduling of the pipeline, which would determine the amount and frequency of resources needed to be dedicated to the pipeline to ensure its successful operation. I then analyze the low-level aspects of the pipeline, as well as the challenges presented in each part of this project-planning phase. The low-level explanation describes the data specifications the pipeline operates upon, which provided a motivation for the STS portion of the thesis. I then summarize the outcome of the successful implementation of the pipeline within the advertising team, and describe the work that will take place in the future to support the pipeline within the team.

Working on a data pipeline provided a novel perspective on the way big data works in the industry, and more specifically how data retention policies exist in the real world. Data retention is the act of holding onto data, whether that is user data or derived metrics based on user data. The insights I gained implemented and researching the data pipeline provided the motivation to focus my STS paper on the connections between data retention and social forgetfulness. Social forgetfulness is the act of willingly or accidentally omitting or forgetting information or data. It takes place all the time in humans and in society, from forgetting a name to historical evidence being erased. In my STS paper, I dive into the challenges that our society faces in the age of

information technology by going against these societal norms of social forgetfulness, and how can we avoid some of the pitfalls of them. To begin, I dive into how we can define information in regards to the age of information, and narrow the scope to electronic data. Then, I discuss why large-scale data retention occurs and when it is necessary and unnecessary to retain information. Then, I discuss the effects that data retention has on a society, both positive and negative. During this discussion, I provide historical anecdotes that show that the issues of data retention do not only occur in a data-centric world like the current one, but throughout history. Following this, I discuss the tangible effects of the loss of data, the inefficiency of data retention, and the collective act of social forgetfulness.

The primary focus of this discussion is to detail why social forgetfulness is necessary in our society, and why it has changed so much in recent times due to the rapid development of data storage, coupled with the slow development of data privacy legislation. I then describe the current state of relevant legislation pertaining to the world of data privacy and more specifically data retention, and detail some of the shortcomings with the current approach that the modern age is taking on regulating the booming data collection industry. To provide a meaningful output, I then propose a specific framework in order to better prepare our modern society for the future of data retention with a focus on prioritizing social forgetfulness. Within this framework are three major proposals that take major inspiration from a similarly focused piece by Jean-François Blanchette and Deborah G. Johnson. These three proposals outline a more specific comprehensive approach that aims to provide a safer set of guidelines to protect the data subjects that are so common in our data-centric world. I think it's incredibly important that we are proactive in the context of data privacy and data retention, as there can be so many unintended consequences when it comes to major decisions and practices that are employed in the data

collection industry. It's important that we protect the user first, and with the lackluster state of legislation relating to data privacy, we need to make major changes rapidly.