Thesis Project Portfolio

REAL-TIME STREAMING FEATURE GENERATION

(Technical Report)

RESPONSIBLE RESEARCH AND INNOVATION OF RECOMMENDATION SYSTEMS

(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 in Computer Science

Param Damle

Spring 2024

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

In the digital age, companies are increasingly leveraging online advertising to target a wider variety of users and tailor content to maximize the likelihood of the user engaging with their products. As data-driven marketing demonstrates better personalization for online ads, companies pursue transformation pipelines to efficiently convert data collected from a user's web traffic and browsing behavior into marketing predictions. A distributed streaming pipeline not only provides robust data processing, but its real-time operation allows the company to market ads to a user before they leave the website. However, due to the technical complexity of connecting data between endpoints, as well as the expertise required to deploy such an application, many companies search for a simpler way to deliver marketing predictions in real time. To this end, we developed a streaming pipeline using Apache Flink that processed user data across nodes in an Elastic Map Reduce (EMR) cluster running on Amazon Web Services (AWS), transformed raw information into feature sets, and outputted these features to a model for marketing prediction. This pipeline provided reliable feature extraction from data streams over 1Gb/s in throughput and improved the system latency from 24 hours to sub-millisecond orders of magnitude. Expansions upon our setup can provide additional efficiency by taking advantage of pipeline-model synergy and by deploying hardware accelerators for feature extraction.

The volume of data collection required to run an effective recommendation algorithm coupled with the manner in which it is exposed to the developers of the system poses numerous risks to the end user. In addition to deceptive collection of one's browsing patterns, such information can also be breached by hackers, sold to third parties, or even exposed indirectly through recommendations to other users in one's circles. Beyond privacy concerns, these recommendation algorithms can influence user behavior towards misinformation and extremism

on social media or towards unsafe or unethically produced merchandise on ecommerce websites. The Responsible Research and Innovation framework indicates that a consensus of regulatory bodies, system developers, and public advocates must agree on future directions of emerging technology like recommendation systems so that their positive benefits can be maximally exploited while their costs to societal welfare are mitigated. While Europe, where this policy was introduced, has taken the lead on data privacy and online user protections with legislation like GDPR, technological actors in the United States have acted with limited regulatory oversight. In this case, it is upon the engineers with direct influence over the systems to integrate guardrails against the aforementioned risks and proactively protect their end users. This concept, termed "middle modulation", is key to addressing these concerns between the extremes of capable but faraway legislation and a concerned but disenfranchised population of users. I analyze whether the research and development community has adequately discussed and pursued middle modulation by analyzing the records of the Recommendation Systems conference, which has brought together the most prominent voices in the topic for over a decade. By using natural language processing techniques such as word frequency and sentiment analysis, I discovered a noticeable increase in attention to topics relating to privacy and societal stability, indicating that the research community is making an effort to effectively modulate the technology and protect the users living in a society deeply intertwined with recommendation systems and their ramifications.