

Executive Summary

The Internet Behavior Experiment is a large-scale longitudinal study designed to examine participants' online activity, including browsing and purchasing behavior, in order to better understand the impacts of online tracking. This capstone project focuses on a core technical challenge within that broader effort: extracting structured purchase information from large volumes of raw, unstructured browsing and email data collected under strict privacy and computational constraints. The extracted data must be accurate enough to support meaningful analysis, efficient enough to operate at scale, and robust to the variability present in real-world web and email content.

Over the course of this semester, we conducted an extensive investigation of both the data itself and the constraints under which it is collected and processed. This work included manual inspection of HTML pages and emails, analysis of vendor-specific design patterns, and research into web standards, anti-scraping practices, and large-scale data processing techniques. Through this process, we identified significant challenges, such as vendor variability, template drift, anti-scraping measures, and privacy-preserving data transformations, as well as important simplifying factors, including a highly skewed vendor distribution and the benefit of operating within a fixed temporal window. These observations directly informed our methodological choices.

Rather than relying on a single extraction technique, we developed a staged extraction pipeline through extensive iteration and empirical testing, balancing determinism, flexibility, and computational efficiency. The final approach prioritizes a rule-based extractor for both HTML and email data, leveraging structural, semantic, and format-based signals present in the raw files. For cases where our rule-based extraction fails, we have found incorporating a selective large language model (LLM) fallback operating on compact, text-based representations to be a viable fallback solution. This fallback is used sparingly and only when necessary, mitigating the cost and reliability concerns associated with LLMs. Monitoring and evaluation are treated as first-class components of the system, with extraction success rates tracked by vendor and data field to detect template drift and guide targeted updates over time.

In addition to extraction, we designed the downstream storage and analytics strategy. Extracted purchase data will be stored in Apache Parquet and analyzed using Apache Spark, aligning with existing project infrastructure and enabling efficient large-scale aggregation and analysis.

The outcome of this project is not a fully automated or universally applicable extractor, nor was that the goal. Instead, we deliver a practical and extensible pipeline that is well-suited to the constraints and characteristics of the Internet Behavior Experiment data. The system achieves reliable extraction for high-frequency vendors, scales to large participant datasets, and provides clear mechanisms for handling failures and evolving data formats, forming a solid foundation for continued development and analysis.