VCU College of Engineering

CS 25-322 AI-generated planning insights powered by Clickstream data

Project Proposal

Prepared for

Mahesh Nair / Tyler Jordan / Emily Corxall

Capital One

By

Priya Choudhary, Bindi Patel, Carissa Trieu, Ivan Emdee

Under the supervision of

Thomas Gyeera

10/11/2024

Executive Summary

The purpose of this project is to collect and analyze real-time clickstream data to provide AI-driven insights for optimizing roadmap planning. Analysis of user interactions on a web application similar to the Capital One Help Center page will allow stakeholders to make better informed decisions with an emphasis on metrics like Average Handle Time and user engagement.

The primary objectives and deliverables of the project include the creation of an intuitive user interface, analysis of clickstream, and implementation of AI/ML models to offer actionable recommendations. A scalable database will also be utilized to handle clickstream data. Using the recommendations, epics, and stories will be automatically generated in JIRA for efficient decision-making.

The final project deliverable will be a fully functional prototype, ready for demonstration to Capital One. Academic deliverables include a project proposal, a preliminary design report, a fall semester poster and presentation by November 2024, a final design report, and a final EXPO poster/ presentation by Spring 2025. This project is relevant for Capital One's Agent Servicing team, as it will allow stakeholders to gain a better understanding of user behavior, improve platform efficiency, and support long-term strategic planning. Ultimately, this project aims to both enhance user experience and operational efficiency.

Table of Contents

Section A. Problem Statement

In today's fast-paced digital world, platform stakeholders need to make informed, data-driven decisions to enhance user experiences, optimize application performance, and plan for future growth. Clickstream data, which captures user interactions within an application, provides valuable insights into user behavior, but many organizations struggle to fully leverage this information to guide their product roadmaps.

Clickstream is a sequence of pages to describe the history of a user's session. This type of data has proven useful in web usage mining and in generating real-time predictions (Bucklin & Sismeiro, 2009). Currently, existing solutions often fall short when it comes to processing clickstream data at scale because it is done by hand. This takes a significant amount of time to do. Delivering AI-powered recommendations, and integrating seamlessly with tools like JIRA for project management will help shorten the time taken. This project aims to bridge that gap by developing a web application that not only captures clickstream data but also uses AI to analyze it and provide actionable recommendations for improving platform features and reducing inefficiencies.

The system will address key stakeholder questions, such as identifying areas to reduce Average Handle Time (AHT), pinpointing where users spend the most time in the application, and highlighting critical features. Additionally, it will integrate with JIRA to automatically generate mock Epics and stories based on accepted recommendations, helping streamline the implementation process. By creating this solution, the project equips stakeholders with a powerful tool to turn clickstream data into meaningful insights, allowing for smarter, data-driven decisions that support strategic roadmap planning.

Our project client is Capital One, they have tasked us with this project because they want to see if it is more efficient to have an AI look over their clickstream data. Rather than having a human look over it. After doing much research, we believe that it will be and hope to prove it with our prototype. The articles we read said that it was faster and more accurate to have an AI look over the data.

The general field of study that the project falls under is data analytics and artificial intelligence, specifically clickstream analysis and artificial intelligence-driven product roadmap planning. By analyzing the sequence of clicks taken by users on a website, clickstream data can provide valuable customer insights and enhanced business intelligence. Other use cases include web movement investigation, statistical surveying, and programming testing (Hanamanthrao & Thejaswini, 2017). Additionally, AI and machine learning models are transforming how product roadmap planning is done. One key advantage of its implementation is the acceleration of timeline creation. Furthermore, AI and ML algorithms have already been implemented in other industries like traffic management and urban planning to make real-time decisions, based on data collected from GPS and sensors (Khare & Arora, 2024).

Capital One, the sponsor company, is a leading company in the banking industry with a great emphasis on information and technology. Specifically, the Agent Servicing team serves to enhance customer support by providing agents with the necessary tools to effectively assist customers. They provide services that seamlessly handle customer inquiries by enabling agents to access relevant real-time information such as customer account details, transaction histories, and service records. Currently, the platform does not have any integration with artificial intelligence.

Clickstream data has been used to generate AI insights for project and product planning, but it is less common in industries related to customer support platforms for large-scale digital services in the banking industry. For instance, historically, clickstream data has been utilized for optimizing websites and predicting user behavior. In a paper titled “Real-Time Clickstream Data Analytics and Visualization”, Hanamanthrao & Thejaswini (2017) highlighted the benefits of real-time data processing over traditional batch processing techniques for clickstream analysis. Data collected from an online learning platform website included course enrollment, navigation patterns, and time spent per page. The data was stored and processed using Apache Hadoop, an open-source big data storage platform. A real-time data pipeline was implemented using Apache Kafka for data streaming, Apache Spark for data processing, and ElasticSearch for data indexing and searching. Real-time processing led to the creation of immediate insights, such as the identification of popular courses and optimal navigation pathways on the portal. The results also demonstrated that real-time processing improved course recommendation accuracy and resource allocation, compared to traditional batch processing. Previous approaches required more time to process data before any insights could be applied. Manipulation of real-time data processing sets the groundwork for future predictive analytics related to improvements based on live user behavior. Furthermore, this approach allows for future implementation of ML models which create personalized recommendations.

While Hanamanthrao & Thejaswini demonstrated the clear effectiveness of real-time processing, there was no implementation of AI or ML algorithms. Gumber et al.conducted a study, “Predicting Customer Behavior by Analyzing Clickstream Data”, which emphasized the importance of clickstream in understanding user behavior and applying it to develop a machine learning model. In this academic journal, they also reference Hanamanthrao & Thejaswini in the literature survey, noting that they were unable to collect data from multiple sources. Clickstream data was collected from an e-commerce site because shoppers often navigate through multiple pages before making a purchase decision. Unstructured data was obtained from Kaggle, an open-source platform providing datasets, containing information such as event time, event type (view, cart, purchase), and category information. Preprocessing was conducted before implementing the machine learning algorithm to make it usable. The machine learning model chosen, XGBoost, can handle large datasets well while processing them quickly with fewer computing resources. It also utilizes regularization techniques to prevent overfitting, a common machine learning behavior that can potentially give inaccurate predictions, more efficiently than other boosting algorithms. Using the Extreme Gradient Boosting (XGBoost) algorithm, customer behavior predictions were made with 85.9% accuracy and 91.04% recall (Gumber et al., 2021).

Though the industry in which the research was performed is vastly different from agent servicing platforms, similar principles could be applied to enhance customer support platforms.

"Neural Networks for Customer Classification Through Clickstream Analysis" written by Erika Severeyn, Alexandra La Cruz, Roberto Matute, and Juan Estrada in 2023 goes over how most organizations currently interpret large amounts of data within their databases and how they hope to move toward AI algorithms in the future. The article uses the term business intelligence (BI), which is the strategies, technologies, and tools that an organization uses to analyze data. They then use that information to try and gain insights to conduct future business decisions. By analyzing this data organizations can attempt to make predictions about customers behavior, market trends, and other business-impacting factors. The newest addition to companies BI's has been the use of AI to break down these large data sets. Leveraging clickstream data is a valuable resource that can significantly aid businesses in enhancing their productivity through the implementation of artificial intelligence for data analysis. The main use of this so far has been to personalize online experiences for customers. The type of AI being used is an AI that has a neural network. They mimic having a human brain to learn as they get fed more information. By using these neural networks businesses can more accurately predict user behaviors. For the research done in the paper, they used Clickstream data from IMOLKO C.A.'s website. IMOLKO C.A. is a service company that helps businesses increase their profits, retain clients, and reduce the churn rate. The dataset was collected by Google Analytics which is a platform that can collect real-team user interactions. The researchers' goal was to see which users would become customers based on the dataset. They trained the neural network on the pre-processed dataset. They did five different training techniques, it was performed for 90% training and 10% testing, 80% training and 20% testing, 70% training and 30% testing, 60% training and 40% testing, and 50% training and 50% testing. The result was that the different techniques did not affect the AIs accuracy. It stayed consistent throughout all the training. The model demonstrated that it could accurately identify positive cases, cases where the user did become a customer, 85% of the time. It was much better at identifying when a user would not become a customer, achieving an accuracy of 91%.

Another article that we looked into was focused less on trying to make a profit and more on helping users correctly use their website. The article is called "A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks" and it was written by Esther Ulitzsch, Vincent Ulitzsch, Qiwei He, and Oliver Lüdtke in 2022. They wanted to see if they could get an AI to predict whether or not a user would fail an interactive problem based on their actions on the website. The thought process was that there was sufficient information for predicting outcomes based on which actions they performed early on and how long it took them. They wanted to be able to predict the outcome as early and accurately as possible. This was done to see if there was a pattern in the way a user would attempt to solve a problem and their success rate. If there was, the website could be redesigned to try and raise the percentage of successful problem-solving. They tested users with two different problems in their study. They called these problems "Lamp Return" and "Meeting Rooms". "Lamp return" involves going through an online shop to return a lamp. "Meeting Rooms" involved going through an online website to reserve rooms based on different requests.

The data from 6,791 “Lamp Return” problems and 6,629 “Meeting Rooms” problems were used for analysis. The success rate for both problems was around 50%. Those who failed generally did fewer actions and spent less time on the problems. The data was preprocessed to make it easier for the AI to understand when it was trying to predict outcomes. This researcher also used XGBoost to try and predict the outcome. The AI was trained on the data and it made its predictions based on many formulas set up by the researchers. The results were that the AI was able to achieve excellent classification performance for the “Lamp Returns” problem. The results were not as clear for the “Meeting Rooms” problem. This comes from the AI’s most predictive feature being time elapsed until action. This means how long the user would spend reading the problem before doing anything. It was much more of a telling feature for the “Lamp Returns” problem than it was for the “Meeting Rooms” problem. Overall though the AI was very accurate at predicting the outcome of the test very early on in the testing. The difference between the accuracy in the problems only shows how important the criteria the AI follows for its predictions are.

Section B. Engineering Design Requirements

This section outlines the core goals, objectives, design specifications, and constraints that guide the development of this project. It provides a structured framework for defining the problem space, informed by client needs, and outlines the conditions and limitations within which the design will operate. By clearly establishing these requirements, the project ensures alignment with stakeholder expectations and creates a roadmap for delivering a successful solution.

The engineering design requirements address several critical aspects that guide the development of the project. First, the Project Goals section describes the overarching goals from the client’s perspective, focusing on the high-level outcomes the client expects. These goals do not provide specific details about the design but highlight the purpose and intent behind the project.

Next, the Design objectives focus on specific, measurable, and time-bound goals that define what the design will achieve. These objectives are drawn from the client’s needs and translated into actionable tasks for the design, ensuring that the solution meets the expected functionality.

The Design Specifications and Constraints section lists the measurable and testable limitations that must be met for the design to be considered successful. It includes performance requirements, hardware capabilities, cost constraints, and data handling regulations, ensuring the design adheres to both functional and legal expectations.

B.1 Project Goals (i.e. Client Needs)

The project aims to help Capital One stakeholders make informed, data-driven decisions using clickstream data. This project will address the underutilization of clickstream data in

roadmap planning and will empower stakeholders to gain actionable insights into user behavior. the primary goals of the project are:

- To leverage real-time clickstream data to provide AT-driven insights for platform optimization.
- To build a web application that mimics Capital One's Help Center and allows stakeholders to visualize data in real time.
- To automate the generation of recommendations for improving key metrics like Average Handle Time (AHT) and user engagement.
- To seamlessly integrate the insights and recommendations into JIRA, allowing for the automatic creation of epics or stories that align with AI-driven suggestions.

B.2 Design Objectives

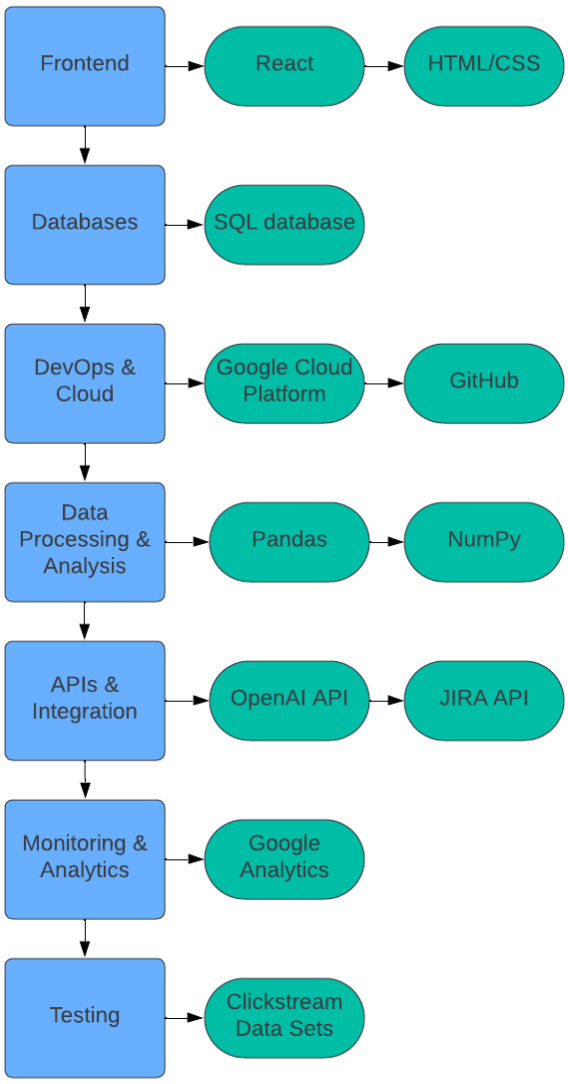
- The design will capture real-time clickstream data from users interacting with a web application that mimics the Capital One Help Center
- The design will provide AI-generated insights based on the clickstream data to answer queries such as, "How can I reduce AHT in the X container?" or "What are my critical features?"
- The design will integrate with JIRA via APIs to allow stakeholders to automatically create epics or stories based on the recommendations provided by the AI model.
- The design will offer a user-friendly, responsive interface that provides easy navigation, filtering, and data visualization using charts and graphs to help users analyze trends and performance metrics.
- The design will ensure that the platform operates effectively across different devices (desktop, mobile, tablet) to accommodate diverse users.

B.3 Design Specifications and Constraints

- The application must handle real-time clickstream data processing with a latency of no more than 5 seconds to ensure accurate insights
- The AI model must analyze clickstream data and generate recommendations with an accuracy rate of at least 85%. the system will utilize OpenAI's GPT for generating insights.
- The system must handle up to 1,000 concurrent users and process data efficiently without performance degradation. This ensures the platform's scalability for larger user bases.
- The application must interface with JIRA via REST APIs to create epics or stories automatically from AI recommendations within 3 seconds of acceptance.
- Clickstream data must be stored in a scalable database capable of handling data growth of at least 1GB/day without affecting system performance.

- The system must comply with Capital One's privacy regulations, ensuring that no personal identifiable information (PII) is exposed or stored. Data encryption will be implemented for both in-transit and at-rest data.
- The UI should allow for filtering of data by various parameters (e.g., date range, container type, user actions) and ensure responsiveness on mobile, tablet, and desktop devices. Performance tests will ensure UI responsiveness within 2 seconds across devices.
- The system should have an uptime of at least 99%, ensuring minimal downtime for continuous access to insights and recommendations.

Tech Stack Diagram:



Section C. Scope of Work

The scope of this project involves building an AI-powered web application to analyze clickstream data for platform performance recommendations, with integration into JIRA for project management. The key objectives include capturing and analyzing real-time data, generating actionable insights, and streamlining roadmap planning. This project is critical for platform stakeholders to make data-driven decisions to improve user engagement and other key metrics.

Boundaries for the project include a focus on web application development with specific deliverables and milestones as outlined below. The team will operate under the Agile methodology, utilizing iterative development to adapt as insights from earlier project phases are incorporated into subsequent stages. The team will work closely with the faculty advisor and project sponsor to ensure timely completion and adherence to the scope, preventing scope creep.

Stakeholder Involvement:

The project sponsor and faculty advisor play a crucial role in verifying and approving each stage of the project. Regular reviews and feedback sessions (weekly Zoom meetings) will be conducted to ensure that the project remains aligned with the expectations and requirements set by the stakeholders. This will help identify any potential issues early on, ensuring any necessary adjustments are made without deviating from the project's scope. Active stakeholder engagement will mitigate the risk of scope creep and help ensure that all deliverables meet the desired quality standards.

C.1 Deliverables

Project Deliverables:

- Web Application: A functioning web app with a user-friendly interface mimicking the Capital One Help Center page. The app will capture and display real-time clickstream data, providing actionable AI-driven insights.
- Clickstream Data Analysis: Real-time data processing and visualization using AI to recommend improvements, including metrics like Average Handle Time (AHT) and user engagement.
- JIRA Integration: Automatic creation of mock Epics and stories based on AI recommendations.
- AI/ML Recommendation Engine: Integration of OpenAI's API for generating insights from clickstream data.
- Database: A scalable SQL database for storing structured clickstream data.
- Final Prototype: A complete, functional prototype that can be demonstrated for review.

Academic Deliverables:

- Team Contract
- Project Proposal
- Preliminary Design Report
- Fall Poster and Presentation (November 2024)
- Final Design Report
- Capstone EXPO Poster and Presentation (Spring 2025)

Risks and Mitigations:

- Access to campus: Some team members may not have consistent access to campus facilities due to living off-campus, but the majority of work can be completed remotely.
- Remote Work: Resources like shared drives, GitHub, and remote collaboration tools (VS Code, JIRA, and Slack) will be used to mitigate risks.
- Third-party vendor delays: The use of existing APIs and software mitigates the need for extended lead times or delays related to third-party ordering.

C.2 Milestones

Milestone	Estimated Completion Date
Team Contract - Establish a foundation for team collaboration and communication. The document lists individual team member strengths, roles, and responsibilities, time commitments, and the communication/meeting structure.	September 6, 2024
Project Proposal - Frame the project early on by outlining key details such as background information, objectives, and the unmet engineering need being addressed. Includes sections on background, literature review, project goals, objectives, constraints, project scope, deliverables, organizational structure, and a proposed timeline.	October 11, 2024

Prototype Web Application Build - Develop a basic web application version to mimic the Capital One Help Center page, integrating key features and functionalities to create a working model for testing and feedback.	November 15, 2024
Fall Design Poster - Present preliminary design results in a public forum at the end of the semester, where team members can showcase their work to sponsors, faculty, administrators, and peers.	November 15, 2024
Preliminary Design Report - Provide a comprehensive summary of the team's design efforts over the semester, with more detail than the project proposal.	December 9, 2024
Data Collection and AI Integration - Implement a system for collecting relevant data, followed by integrating AI/ML models to analyze and interpret the data, providing insights for future development.	February 2025
AI Recommendations in JIRA - Deploy the AI-powered solution, enabling it to analyze clickstream data and provide actionable recommendations. These will be integrated with JIRA to create mock Epics and stories for roadmap planning.	March 2025
Final Design Report Submission - Deliver a detailed report summarizing the team's work, including final design, implementation results, and project outcomes, showcasing the full scope of the project.	April 2025
Capstone Design Expo - Present the final project at a public expo, demonstrating the completed work to industry sponsors, faculty, and	April 24-25, 2025

peers, emphasizing technical achievements and project outcomes.	

C.3 Resources

1. Software:
    a. Integrated Development Environments (IDEs): Primarily using VS Code for development.
    b. Version Control: GitHub will be used to manage source code, track changes, and collaborate.
    c. Data Analysis Platforms: Pandas and NumPy libraries for processing and analyzing clickstream data.
2. Cloud Computing Services:
    a. Amazon Web Services (AWS) or Google Cloud Platform (GCP): To scale real-time data processing and storage as needed, especially during peak usage or testing.
3. APIs and Libraries:
    a. OpenAI API: For generating AI-driven recommendations based on clickstream data.
    b. JIRA API: To automate the creation of stories and epics in JIRA based on AI insights.
4. Databases:
    a. PostgreSQL Database: For storing structured clickstream data for real-time access and analysis.
5. Testing and Analytics Tools:
    a. Clickstream Data Sets: Operational datasets for testing AI recommendations and simulating user interactions on the platform.
    b. Google Analytics: To track and analyze web traffic, providing additional insights into user behavior and augmenting clickstream data for more robust analysis.

At this stage of the project, we do not anticipate purchasing any additional software or resources. However, as the project progresses and requirements evolve, we may reassess our needs and consider acquiring specific tools or services if necessary to ensure successful completion.

References

[1] VCU Writing Center. (2021, September 8). *APA Citation: A guide to formatting in APA style.* Retrieved September 2, 2024. https://writing.vcu.edu/student-resources/apa-citations/

[2] Teach Engineering. *Engineering Design Process*. TeachEngineering.org. Retrieved September 2, 2024. https://www.teachengineering.org/populartopics/designprocess

[3]J. Electrical Systems 20-2 (2024):1600-1608.Real-Time Clickstream Analytics with Apache. Retrieved October 10, 2024. https://www.researchgate.net/profile/Tulshihar-Patil/publication/382193441_Real-Time_Clickstream_Analytics_with_Apache/links/6691464a3e0edb1e0fe0ce1c/Real-Time-Clickstream-Analytics-with-Apache.pdf

[4]Hanamanthrao, R., & Thejaswini, S. (2017). Real-time clickstream data analytics and visualization. 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). https://doi.org/10.1109/rteict.2017.8256978

[5]Ehsan, A., Abuhaliqa, M. A. M. E., Catal, C., & Mishra, D. (2022). RESTful API Testing Methodologies: Rationale, Challenges, and Solution Directions. Applied Sciences, 12(9), 4369. https://doi.org/10.3390/app12094369

[6]Yusof, M. K., Man, M., & Ismail, A. (2022). Design and Implement of REST API for Data Integration. 2022 International Conference on Engineering and Emerging Technologies (ICEET). https://doi.org/10.1109/iceet56468.2022.10007414

[7]M. Gumber, A. Jain and A. L. Amutha, "Predicting Customer Behavior by Analyzing Clickstream Data," *2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, Chennai, India, 2021, pp. 1-6, doi: 10.1109/ICCCSP52374.2021.9465526.

[8]Liu Y, Fan S, Xu S, Sajjanhar A, Yeom S, Wei Y. Predicting Student Performance Using Clickstream Data and Machine Learning. Education Sciences. 2023; 13(1):17. https://doi.org/10.3390/educsci13010017

[9]Ulitzsch, E., Ulitzsch, V., He, Q. et al. A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks. Behav Res 55, 1392–1412 (2023). https://doi.org/10.3758/s13428-022-01844-1

[10]Severeyn, E., Alexandra La Cruz, Matute, R., & Estrada, J. (2023). Neural Networks for Customer Classification Through Clickstream Analysis. https://doi.org/10.1109/etcm58927.2023.10309081