SEAS 6402 Capstone Project

Event Stream Analysis and Audit Trail Management: A Framework for Machine Learning-Driven Recommendations for Trustworthy AI systems

Paul Gimeno

The George Washington University

M.S Data Analytics Program, Spring 2023

Table of Contents

Event Stream Analysis and Audit Trail Management: A Framework for Machine Learning-Driven Recommendations for Trustworthy AI systems.

I. Abstract

This capstone project presents a framework for processing simple event streams, generating machine learning-driven recommendations, creating action items using a large language model (LLms), and maintaining customizable audit trails tailored to a user's role as a data steward in a fundraising organization. With the rapid growth of data and increasing complexity in decision-making processes, there is a need for an integrated approach to harness the power of big data, machine learning recommendations and LLM's for processing for real-time insights and efficient action planning.

The proposed project contains four components: (1) event stream processing to identify and analyze donor activity data, (2) a machine learning model for generating recommendations or probabilistic predictions, (3) LLM-based action item creation to facilitate effective communication and task management, and (4) an audit trail component for ensuring transparency and accountability to ensure sustainable data governance. This framework is designed to be adaptable to enterprise business intelligence use cases and enables organizations to leverage data-driven insights with insight into their data's data and model training lineage.

To evaluate the effectiveness of the framework, we implement it in a case study focusing on a specific industry application in fundraising. The results demonstrate the framework's ability to process event streams efficiently, generate accurate recommendations, create coherent and actionable items using LLM, and maintain comprehensive audit trails according to user preferences. We discuss the importance of audit trail management and the role of data governance in building trustworthy decision support systems, and

how best to leverage a standard framework such as the proposed project as a real-life example of practical application in the industry.

II. Introduction

The rapid advancement of technology has led to an exponential growth in data generation, making it increasingly challenging for organizations to process and analyze information in real-time. Event stream analysis and audit trail management have emerged as crucial aspects of modern data-driven decision-making processes, enabling organizations to stay agile, informed, and accountable. This capstone project introduces an applied framework focused on event stream analysis and audit trail management, which incorporates machine learning-driven recommendations and action item generation through large language models (LLMs).

The framework proposed in this project is designed to address four main objectives: (1) efficiently process event streams to identify patterns, trends, and anomalies in real-time data, (2) generate context-aware recommendations using machine learning algorithms, (3) create coherent and actionable items by leveraging LLMs, and (4) maintain audit trails to ensure transparency, traceability, and compliance with user preferences.

We'll apply this framework to a case sample in an advancement organization where potential prospects are evaluated for giving propensity given their background, wealth rating, giving history and interactions with staff. This industry use case leverages a blend of data inputs from a variety of sources: user-generated data from donors and staff, external data from the web about donors and public information about a prospect’s assets. We’ll generate data from these sources, make recommendations and evaluate the data lineage behind the models and the LLM parameters.

For the project architect, an AWS serverless infrastructure is used for fast and scalable deployment. For our base language model, we’ll leverage OpenAI’s chatGPT completion API for summarization tasks and embedding endpoints for vectorizing chunks of text and performing semantic search on related documents. While ChatGPT can sometimes demonstrate poor performance in terms of objective evaluation metrics for explainable recommendation tasks, it outperforms most state-of-the-art methods in human evaluations. [1]

Lastly, we piece data sourcing, recommender systems and LLM application together by organizing data lineage and model lineage through a common pipeline that collects artifacts for the purposes of auditing data usage, data access and validation.

III. The case for data centric governance for responsible AI deployment

As artificial intelligence (AI) and machine learning applications continue to play an increasingly prominent role in an organization’s core operation, it becomes crucial to ensure that its development and use are guided by ethical principles and standards that promote fairness, transparency, and accountability. The field of AI governance is concerned with designing and implementing sound principles and standards to ensure that AI technologies are used responsibly and do not cause unintended harm to individuals or society as a whole. However, the current approaches to AI governance have limitations that hinder their effectiveness in addressing potential harms. It is important to explore alternative approaches to AI

[1] Liu et al., J. (2023, April 20). *[2304.10149] Is ChatGPT a Good Recommender? A Preliminary Study*. arXiv. Retrieved April 24, 2023, from https://arxiv.org/abs/2304.10149

governance that can better operationalize governance requirements and enable a way to reproduce evaluation of its use with minimal cost.

Challenges of AI and data governance

The current approaches to AI governance consist mainly of manual review and documentation processes. While such reviews are necessary for many systems, they are not sufficient to systematically address all potential harms [2](McGregor & Hostetler, 2023) The challenges associated with data governance in the context of large-scale language technology include diverse needs and goals of each stakeholder, navigating varying local laws, promoting agency of all data stewards, contending with multiple legal contexts, and addressing existing power imbalances. These challenges require coordination across stakeholders and alignment with ethical and legal standards to ensure effective data management practices. [3] (Yacine at al., 2022, #)

AI, LLM's and model interpretability

The field of AI and machine learning processing has experienced a surge in popularity recently due to the increasing capabilities of large language models. However, this popularity has led to a false sense of perceived performance, as increased performance in a certain setting does not necessarily indicate that the interpretation or explanation provided by the model is faithful. It is important to distinguish between model performance and how trustworthy the value it provides because it could lead to unintended biases in our products and decision support systems. In most instances, Increased performance is indicative of a

[2] McGregor, S., & Hostetler, J. (2023, February 14). *[2302.07872] Data-Centric Governance*. arXiv. Retrieved April 24, 2023, from https://arxiv.org/abs/2302.07872

[3] Yacine at al., J. (2022). Data Governance in the Age of Large-Scale Data-Driven Language Technology. *2022 ACM Conference on Fairness, Accountability, and Transparency*, (2022). https://arxiv.org/abs/2206.03216v2

correlation between the plausibility of the explanations and the model's performance, rather than faithfulness[4] (Jacovi & Goldberg, 2020). This highlights the need to evaluate faithfulness independently from plausibility to ensure that the explanations provided by the model accurately represent the model's reasoning process, rather than simply providing plausible explanations.

Using AI to manage AI

Blending smart organizational policy with AI tooling to manage AI through a unified framework can be valuable for data governance in a number of ways. Firstly, they can facilitate collaboration between multiple stakeholders by providing a common platform for data management and analysis. LLM tools can be leveraged to bridge the data and business domain silos across an organization by making data understandable to anyone. This fosters an environment where multiple parties can contribute to the development and refinement of a set of processes: 1) from refining data sourcing catalogs to 2) model training and 3) transparent use of LLM in providing products. This collaborative approach can lead to more comprehensive data governance frameworks that are tailored to the specific needs of the stakeholders involved.

Data governance in AI is a complex and multifaceted process that requires both human and technological tools to work together. The current approaches to AI governance that rely solely on manual review and documentation processes are insufficient in systematically addressing all potential harms. The challenges associated with data governance in the context of large-scale language technology highlight the need for coordination across stakeholders and alignment with ethical and legal requirements to ensure effective data management practices. Therefore, it is essential to strike a balance between the human element and autonomous tooling to promote agency for data stewards, navigate local laws and legal contexts, and

[4] Jacovi, A., & Goldberg, Y. (2020, April 7). *[2004.03685] Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?* arXiv. Retrieved April 25, 2023, from https://arxiv.org/abs/2004.03685

address existing power imbalances. By leveraging both human and technological capabilities, we can ensure that AI technologies are developed and used in a responsible and ethical manner that benefits all stakeholders involved.

IV. Implementing a trustworthy AI system with AWS

The project consists of the following components. 1) the event source 2) the data integration layer 3) a unified pipeline for master data/metadata management 4) an API service that processes recommendations and 5) the presentation layer interface. The event source, data integration layer and master data management mirrors common data warehousing architecture where events, entities and objects will derive from dimension and fact tables that unites different source systems into one unified architecture and organizes business processes logically in a data lake[5]. (Ross & Kimball, 2013, #)

Architecture

The project adopts a serverless AWS architecture where application data is ingested through AWS Data Migration Service and first cataloged into S3 buckets (figure 1). The collection constitutes the 'raw' zone where data is re-routed to various data pipelines that are first orchestrated via lambda functions. The initial lambda function processes raw data designed to be stored in the audit trail catalog. The data storage of choice chosen for this configuration is DynamoDB primarily for its ease of accessibility and unstrict requirements for pre-defined data schemas. The audit trail needs to be robust and malleable for when new features needed to be added at any given time which is something not easily feasible with relational database systems. New data is then picked up by a second lambda function into AWS glue which serves as the data integration layer that is responsible for transforming and standardizing the datasets into the

[5] Ross, M., & Kimball, R. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Wiley.

raw zone. For the final data catalog layer, a relational database is chosen as the final layer for consumption (MySQL). Our Amazon instance serves as the analytics data lake for which data scientists of ML practitioners run their experiments. This serverless AWS architecture provides a scalable, efficient, and cost-effective solution for ingesting, cataloging, and transforming application data. By leveraging AWS services like AWS Data Migration Service, S3, Glue, Lambda, and DynamoDB[6]. This architecture can easily handle large volumes of data while keeping the infrastructure costs low. Machine learning tasks that integrate with common model repository libraries such as MLFlow, Sagemaker, Azure can easily write training artifacts into our DynamoDB instance where all pre-catalogic tasks are handled.

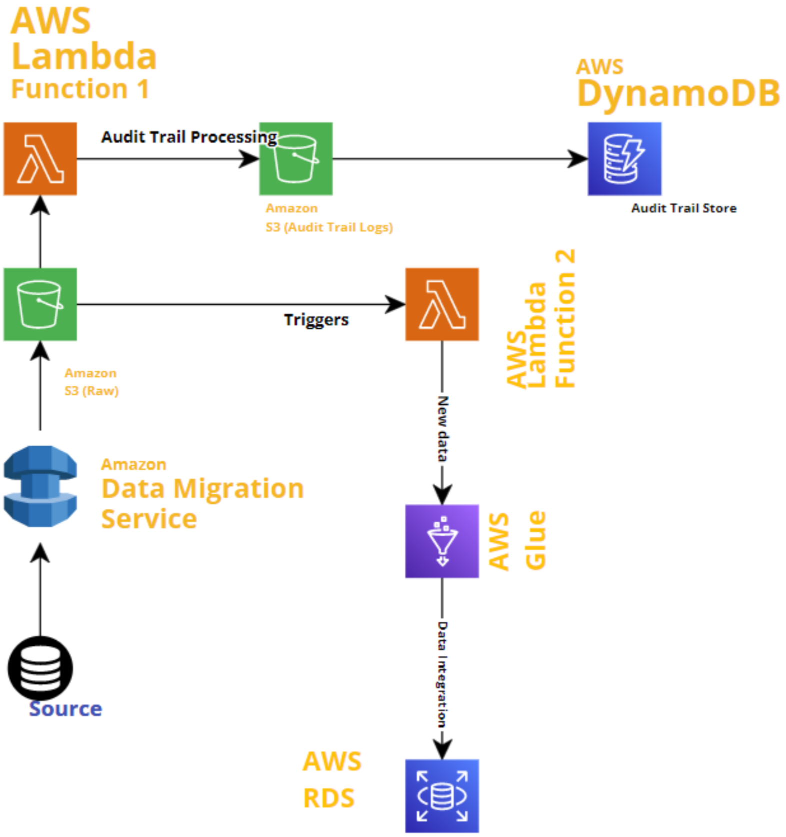

Figure 1: Project Serverless AWS infrastructure

[6] *Serverless | AWS Architecture Center*. (n.d.). Amazon AWS. Retrieved April 29, 2023, from https://aws.amazon.com/architecture/serverless/

Model training, inference and Large Language Models

The standard inference lifecycle for a model trained in our unified prospect activity stream involves three phases: model training, model evaluation, and making an inference. During the model training phase, various models are trained on the prospect activity dataset using various experiment parameters such as learning rate, batch size, and regularization techniques and other hyperparameters specific to the algorithm being applied. The models are then validated to hold out validation sets to ensure that experiments are not overfitted to the training data.

In the model evaluation phase, the model is evaluated on a holdout set to measure its performance on new, unseen data. This evaluation is done using various metrics (accuracy, F1 scores, precision, recall,among many others). During this phase, models are evaluated against a baseline and better performing models are used for inference.

During the inference phase, the trained model is used to predict the probability of a prospect making a gift. The model outputs a probability score, which can be used to make a decision on how to approach the prospect. For instance, if the probability of making a gift is high, an e-mail can be generated using a large language model to persuade the prospect to donate. All training parameters and metadata, including the model weights, validation and evaluation metrics, and inference results, are stored in a DynamoDB instance to ensure traceability and reproducibility.

Large Language Models and Predictive Models in Fundraising Analytics

Machine learning has become a valuable tool for fundraising organizations looking to anticipate donor behavior and make targeted solicitations. By leveraging the power of machine learning algorithms, fundraising organizations can analyze large datasets to identify patterns and insights that can help them optimize their fundraising efforts.

One technique that has proven effective in predicting donor behavior is clustering analysis. Clustering analysis involves grouping donors based on their giving history, demographics, and other relevant factors. By analyzing these groups, organizations can gain insights into donor preferences and behavior, which can be used to make more effective fundraising appeals. Popular donor screening vendors such as DonorSearch utilize publicly available information about donors to cluster similarly profiled donors in a given locale. In addition to clustering, the rise of large language models and state of the art neural network architectures, embeddings can be a game changer in making sense of disparate unstructured data that make up our donor attributes.

Large language model (LLM) embeddings are useful because they provide a way to represent text data in a high-dimensional vector space, which captures the underlying semantics of the text[7] (Tunstall et al., 2022, #). These embeddings are generated by training a neural network on a large collection of text, such as structured texts found in books or unstructured texts obtained from Twitter's API, or in our specific case, a collection of communication between organization staff and donor prospects. The resulting embeddings can be used in a wide range of natural language processing (NLP) tasks, such as text classification, sentiment analysis, and machine translation.

One of the key advantages of LLM embeddings is their ability to capture the meaning of words in context. Traditional text representation techniques, such as bag-of-words or TF-IDF, evaluate each word in isolation, ignoring the context in which it appears. In contrast, LLM embeddings can consider the entire sentence or paragraph in which a word appears, capturing the semantic relationships between words and

[7] Tunstall, L., Werra, L. v., & Wolf, T. (2022). *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O'Reilly Media, Incorporated.

their context. This allows for more accurate and nuanced representation of text data, leading to improved performance and allows us to use more unstructured data to cluster donor prospects that are semantically related.

Moreover, LLM architectures can be used to generate text, such as donor “thank you emails”, event invitations, or solicitation scripts. These architectures can be combined with a decoder neural network architecture to generate text that is semantically meaningful and is aware of prior contexts fed into it. With the availability of embedding models offerings from OpenAI and state of the art embedding models available through platforms like Huggingface, users can take advantage of these LLM embeddings to generate high-quality and contextually relevant text search or generate text for a wide range of applications. Overall, LLM architectures and LLM vector embeddings provide a powerful tool for representing and generating text data, and their use is likely to become increasingly widespread in the field of fundraising.

Beyond clustering and usage of LLM, another mode of technique is predictive modeling. This involves building a statistical model that predicts donor behavior based on historical data. These models can be used to anticipate donor behavior, such as the likelihood of making a gift or the amount of a potential gift. Fundraising organizations can use these predictions to make targeted solicitations that are more likely to result in a successful outcome.

Using both predictive modeling and large language models for clustering can be highly beneficial in predicting donor behavior. Predictive modeling can help fundraising organizations anticipate donor behavior by predicting the likelihood of a donor making a gift or the size of the potential donation. This can help organizations make informed decisions about how to target their fundraising efforts to maximize their success. Additionally, predictive modeling can provide insights into which factors are most important in driving donor behavior, such as demographics, past giving history, or the frequency of engagement with the organization.

On the other hand, large language models for clustering can help fundraising organizations identify patterns and insights in the language used by donors. These patterns can be used to group donors based on their interests and preferences, allowing organizations to tailor their fundraising appeals to specific groups of donors. By using large language models for clustering, fundraising organizations can gain a more nuanced understanding of donor behavior, which can help them make more targeted and effective fundraising appeals.

V. Implementation

The event sourcing and the data integration layer

An end-to-end implementation of a data reporting project involves several steps to ensure that data is collected, transformed, and analyzed in a streamlined manner. The first step is to gather data from the source application for all prospect and donor activities. Once the data is collected, the next step is to use AWS migration service to siphon data from the application and into AWS S3 buckets. This allows for secure and scalable data storage, ensuring that the data is accessible for future analysis.

Data integration and a unified pipeline for master data/metadata management

To further enhance the scalability of the data pipelines, a serverless AWS architecture is employed. This architecture enables the project to scale up jobs as needed, ensuring that the data pipelines can handle increased data volumes without impacting performance. Additionally, two data pipelines are used, with one pipeline designed to store dataset metadata and data transforms in DynamoDB, a NoSQL database. The second pipeline is designed to transform the data into a Kimball data model via facts and dimensions to build a data catalog using AWS Glue and Amazon RDS.

Data catalog for machine learning tasks

Once the final data store is complete, data analysts can utilize the curated data in the warehouse and perform model training using a unified framework such as Databricks or SageMaker. In this implementation, model training artifacts are saved to the same DynamoDB database used in the previous steps. Additionally, text embeddings are employed to perform clustering and similarity search for finding the best prospects using large language models. The embedding parameters and all generative AI prompts (if using ChatGPT) are saved into the audit trail repository stored in DynamoDB. This provides a comprehensive and unified framework for analyzing data and tracking progress, ensuring that data is accurately represented and analyzed for maximum insights.

Fetching Audit Trail logs

An audit trail can be fetched using a DynamoDB source by calling into an API service which is built using Flask. The API service acts as a middleware layer between the front-end interface and the database, allowing users to query the audit trail and retrieve relevant metadata. For example, when a user accesses the front-end interface, data is served to them with data source ID's. These data source ID's can then be used to fetch DynamoDB objects that contain relevant data, such as model training metadata, access levels, and data freshness.

By leveraging DynamoDB as a source of truth for the audit trail, users can access relevant metadata in real-time and with minimal latency. The API service can be designed to handle complex queries, allowing users to filter and search for specific data based on their needs. Additionally, by centralizing the audit trail in DynamoDB, data access and control can be easily managed and audited. This approach enables users to gain valuable insights into the AI system and ensure that it is operating in a transparent and trustworthy manner.

VI. Conclusion

An audit trail is a critical component of any trustworthy AI system, providing transparency and accountability throughout the data pipeline. The audit trail contains important metadata, such as model training parameters, potential biases in the dataset, and any issues with accuracy or performance. This information can be used to identify potential issues and improve the accuracy and fairness of the AI system over time.

Having a transparent data ecosystem requires a unified data pipeline, with all data and metadata stored in a centralized location. This enables easy access to information at all stages of the data pipeline, facilitating analysis and improving overall transparency. Having a unified structure available to stakeholders is key in taming all the complexities in data governance around AI systems because we can show how siloed pieces of data can be related to each other. In addition, we can harness and fine-tune organizational processes via the unified audit trail pipeline by leveraging reinforcement learning and pre-defining costs to certain bad artifacts we see anywhere in the data lineage.

However, there are several strategies that may pose challenges for maintaining a transparent data ecosystem. For example, there may be data sources that are difficult to integrate into the unified data pipeline, such as legacy systems or third-party data providers. Additionally, there may be challenges in ensuring data privacy and security, particularly when dealing with sensitive information

[8](Samarawickrama, 2022). To address these challenges, it is important to establish clear policies and procedures for data governance and security, as well as ongoing monitoring and auditing of the data pipeline to ensure compliance with established standards.

Ultimately, maintaining a transparent and trustworthy AI system requires ongoing effort and vigilance, with a focus on continually improving the data pipeline and addressing any issues that may arise. By leveraging the power of an audit trail and a unified data pipeline, organizations can build more accurate and trustworthy AI systems that can provide valuable insights and improve decision-making across a wide range of applications.

[8] Samarawickrama, M. (2022, September 28). *[2210.08984] AI Governance and Ethics Framework for Sustainable AI and Sustainability*. arXiv. Retrieved April 26, 2023, from https://arxiv.org/abs/2210.08984

References

Jacovi, A., & Goldberg, Y. (2020, April 7). *[2004.03685] Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?* arXiv. Retrieved April 25, 2023, from https://arxiv.org/abs/2004.03685

Liu et al., J. (2023, April 20). *[2304.10149] Is ChatGPT a Good Recommender? A Preliminary Study*. arXiv. Retrieved April 24, 2023, from https://arxiv.org/abs/2304.10149

McGregor, S., & Hostetler, J. (2023, February 14). *[2302.07872] Data-Centric Governance*. arXiv. Retrieved April 24, 2023, from https://arxiv.org/abs/2302.07872

OpenAI. (n.d.). *OpenAI API Reference*. OpenAI API Reference. Retrieved April 29, 2023, from https://platform.openai.com/docs/api-reference

Ross, M., & Kimball, R. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Wiley.

Samarawickrama, M. (2022, September 28). *[2210.08984] AI Governance and Ethics Framework for Sustainable AI and Sustainability*. arXiv. Retrieved April 26, 2023, from https://arxiv.org/abs/2210.08984

Serverless | AWS Architecture Center. (n.d.). Amazon AWS. Retrieved April 29, 2023, from https://aws.amazon.com/architecture/serverless/

Tunstall, L., Werra, L. v., & Wolf, T. (2022). *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O'Reilly Media, Incorporated.

Yacine at al., J. (2022). Data Governance in the Age of Large-Scale Data-Driven Language Technology. *2022 ACM Conference on Fairness, Accountability, and Transparency*, (2022). https://arxiv.org/abs/2206.03216v2