Optimizing an ETL Pipeline From Multiple Angles

The Role of Ethical Design in Building Trust in Machine Learning Algorithms

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Akshay Choksi

October 27, 2023

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Rider Foley, Department of Engineering and Society

Rosanne Vrugtman, Department of Computer Science

Introduction

In the current tech-driven world, data is praised as the new oil, a transformative force with the power to reshape industries and drive innovations (Mayer-Schönberger & Cukier, 2013). As Ikegwu et al. (2022) highlight, big data analytics is becoming increasingly central to data-driven industries, offering insights into data sources, challenges, and potential solutions. Geronazzo and Ziegler (2021) emphasize the transformative power of a data-driven approach in marketing analytics, showcasing how machine learning techniques and external data sources can replace traditional expert-driven processes.

Subsequently, the introduction of big data in enterprises has allowed for more accurate reflections of core business services. However, a significant challenge arises when transitioning from intuition-based decision-making to data-based decision-making: the availability of relevant data in digital form, which is a result of data processing. Misut and Juríik (2021) explain that datafication, the process by which objects and procedures are transformed into digital data, is a crucial step in the processing of big data for companies.

I interned with a national bank in a division dedicated to digital marketing, a domain where data plays a pivotal role. Specifically, data is fundamental to their pricing strategies for Google Ads, enabling them to make informed decisions, optimize spending, and maximize return on investment. However, as Ramirez-Asis et al. (2022) discuss, the digital transformation in a competitive environment is critical, and the optimal usage of data remains a challenge.

According to Grandhi et al. (2020), the success of data-driven marketing depends upon how well an organization embraces the practice, emphasizing the importance of both judgment and intuition in the process. In a non-technical approach to data analytics and machine learning, ethical concerns begin to surface. O'Neil (2016) highlights the non-transparent nature of many

machine learning algorithms and the potential biases they can perpetuate. Similarly, Bostrom and Yudkowsky (2014) emphasize the importance of aligning machine learning algorithms with human values.

Within the bank's digital marketing division, the Extract, Transform, Load (ETL) operations in their datafication pipeline were pivotal for ensuring timely and accurate data processing. The ETL process involves extracting data from various sources, loading it into a data warehouse, and then transforming it to fit the bank's specific analytical needs, see Figure 1. However, I observed that the existing ETL operations were not streamlined, leading to bottlenecks and inconsistencies in the data flow, therefore, affecting datafication. This inefficiency often resulted in outdated or mismatched data being used for critical decision-making in their Google Ads campaigns.

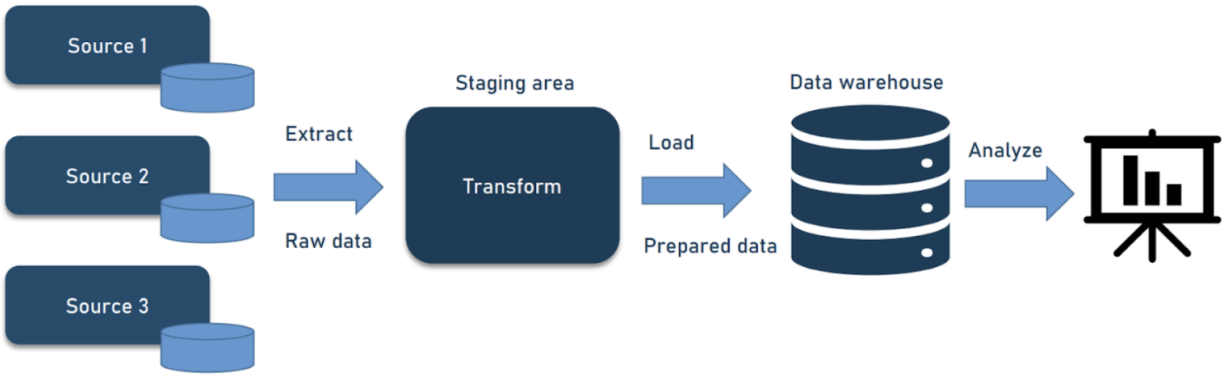

Figure 1. ETL Pipeline Architecture (Source: Altexsoft, 2022)

Optimizing the ETL operations was essential not only to improve the speed and accuracy of data processing but also to ensure that the bank's marketing strategies were based on the most current and relevant data available. This technical project aims to enhance the bank’s data processing pipeline with Delta Lake and AWS Glue technologies. Moreover, it is equally

important to ensure these machine learning technologies are ethically designed, balancing progress with stakeholder trust.

Original System

The bank's initial data processing system was built on the AWS Elastic MapReduce (EMR) platform. AWS EMR, a cloud big data platform, allowed the bank to process vast amounts of data using popular frameworks such as Apache Hadoop and Apache Spark (Chen et al., 2017). While AWS EMR provided the scalability and flexibility to handle large datasets, it came with its own set of challenges.

One of the primary limitations was the bank's reliance on internal data endpoints, which were incompatible with Databricks, an analytics platform known for its collaborative workspace and optimized Apache Spark environment (Chen et al., 2017). This incompatibility posed challenges in terms of data integration, transformation, and real-time processing. The bank relied heavily on custom solutions to bridge the gap between AWS EMR and its internal data sources, leading to increased complexity (Oussous et al., 2022).

Furthermore, unlike AWS Glue, AWS EMR is not serverless. This means that the bank had to manually manage and provision storage, leading to potential inefficiencies in cost and resource utilization (Hassan et al., 2021). The inflexible nature of the original system made it less adaptable to evolving data processing needs, as advanced customizations were time-consuming and costly. The in-house nature of the system's design also meant that it was less familiar to external developers, limiting newer technology access and best coding practices in the evolving big data landscape (Zhou et al., 2017).

Updated System

The redesign principles that guided the enhancement of the bank's data processing pipeline were efficiency and modularity. The architecture was reimagined to integrate cutting-edge technologies like Apache Spark, Delta Lake, and AWS Glue. To see these technologies as part of the ETL Pipeline, see Figure 2. The original system was revamped into distinct modules for each responsibility: Data Extraction, Data Transformation, Data Loading, and Real-time Processing. These modules, while designed to work together, can function independently, allowing for easier updates and maintenance.

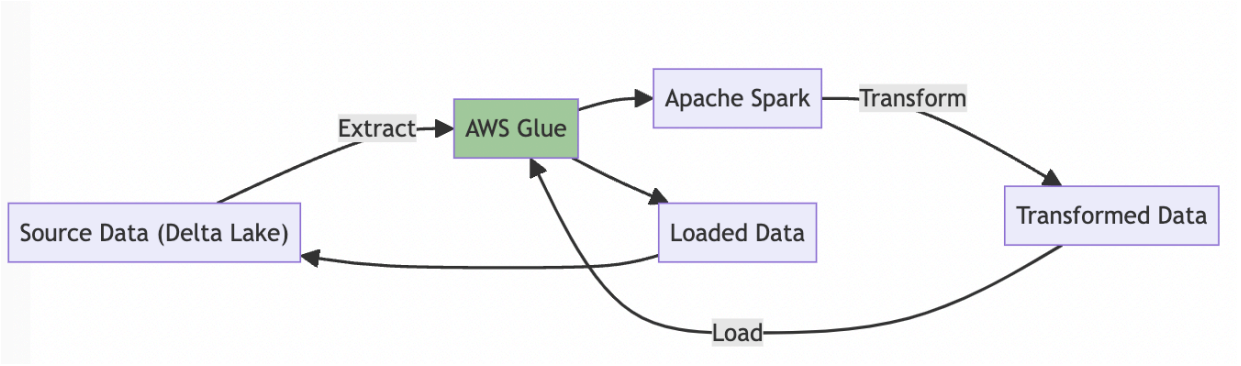

Figure 2. Updated Data Processing Architecture (Source: Choksi, 2023)

A significant upgrade was the integration of Delta Lake for data storage. Delta Lake, an open-source storage layer, brings Atomic, Consistent, Isolation, and Durability (ACID) transactions to Apache Spark and data workloads (Zaharia et al., 2016). This ensures data integrity, especially during concurrent reads and writes, and allows for faster data retrieval. The versioning capabilities of Delta Lake also enable historical data analysis and rollbacks, providing an added layer of data reliability and security.

By leveraging relevant technologies like Apache Spark and AWS Glue, the bank can benefit from the vast global community's knowledge and expertise (Hassan et al., 2021). Apache Spark, known for its in-memory processing capabilities, ensures faster data processing. AWS

Glue, with its serverless ETL capabilities, offers a cost-effective solution for data processing, especially at high loads. As a result, efficiency for the ETL process improved with a 13% reduction in processing time. Combined with the enhanced reliability provided by Delta Lake, this ensures that stakeholders receive consistent and accurate data for decision-making. Integrating these technologies aims to ensure that the bank's data processing operations are not only efficient but also scalable, accommodating future growth and data needs.

Ethical Redesign of Data Algorithms

Machine learning (ML) algorithms are often celebrated as the pinnacle of technological advancement, promising unparalleled efficiency and automation. They are perceived by many as neutral tools, without bias, and purely driven by data (Zhang et al., 2022). However, as more sectors integrate these algorithms into their operations, concerns about their ethical implications have emerged. Large corporations, while utilizing the power of ML, are also at the forefront of these ethical debates, navigating a complex balance between innovation and responsibility (Calvo & Egea-Moreno, 2021).

The Social Construction of Technology (SCOT) framework, as articulated by Pinch and Bijker in 1987, offers a lens through which we can understand this intricate relationship between society and technological advancements. Central to SCOT is the idea that technological development is a process deeply embedded in societal structures and values. Rather than being inevitable or linear, the trajectory of technology is influenced by various stakeholders and their interpretations.

One subtheme of SCOT is *Relevant Social Groups* - different groups of people who interact with technology in distinct ways. Their perceptions and interactions shape the development of technologies. Furthermore, there is *Interpretative Flexibility* - the notion that a

single technology can be interpreted differently by different social groups. The final subtopic of SCOT is *Stabilization and Closur*e - as technological debates unfold and interpretations converge, dominant designs or practices emerge within the technological domain.

There are many relevant social groups in the ML industry. Developers at the forefront of ML innovation view these algorithms from a technical standpoint, focusing on efficiency and optimization (Elish, 2019). End-users, interacting with the tangible outputs of ML algorithms daily, see these algorithms as powerful, often non-transparent forces shaping their experiences (Rader & Gray, 2015). Policymakers, balancing between fostering innovation and protecting public interest, address the societal, ethical, and legal implications (Hoofnagle et al., 2019). This multiplicity of perspectives underscores the SCOT principle of interpretative flexibility, where a single machine learning algorithm might be praised by developers for its efficiency but critiqued by end-users' potentially discriminatory outcomes (Tiwari, 2020).

The introduction of regulations, like the General Data Protection Regulation (GDPR) which created a standard for data privacy laws in all European Union member states, signifies a pivotal moment in the ML landscape. The GDPR has set a precedent for stringent data protection and privacy standards. It emphasizes key principles like consent, transparency, data minimization, and accountability (Lagioia & Sartor, 2020). In the context of ML, GDPR impacts how data is collected, processed, and utilized, directly influencing the design and deployment of ML algorithms, especially in sectors handling sensitive personal data.

Such policies, driven by concerns over transparency and data misuse, represent efforts towards stabilization and closure (Mourby et al., 2021). The GDPR acts as a stabilizing force, setting standards for transparency and data handling. This move towards stabilization reflects a collective response to ML challenges, ensuring technological advancements align with ethical

values. Moreover, the concept of "Ethics as a Service" has been proposed as an approach to bridge the gap between artificial intelligence (AI) ethics principles and practical design, further emphasizing the need for stabilization in the face of ethical challenges (Morley et al., 2021, p. 239).

As ML algorithms become more prevalent, they are deeply embedded within social structures, affecting especially those groups at the receiving end of biases (Tiwari, 2020). Ethical dilemmas in sectors like healthcare, finance, and criminal justice demonstrate the need for alignment between technological advancements and stakeholder values (Calvo & Egea-Moreno, 2021). Therefore, a dialogue involving all stakeholders is important to ensure that machine learning systems are efficient, transparent, and aligned with the end-user's demands. (Zhang et al., 2022).

Research Question and Method

There is an obvious disconnect between the implications of big data usage and machine learning algorithms within industries. A relevant question follows: How can the ethical considerations surrounding machine learning algorithms be integrated into its technological design to increase stakeholder trust?

To investigate this question, a systematic literature review will be undertaken to understand prevailing ethical concerns and the solutions already offered within the ML field, aiming to pinpoint existing patterns and knowledge gaps. This review would include two valuable studies of post-GDPR machine learning development:

1) Transparency in Healthcare (Mourby et al., 2021)
2) Data Minimization in Machine Learning (Goldsteen et al., 2021)

Moreover, structured surveys will be given to three social groups: ML developers, its end-users, and policymakers. Developers offer a lens into the technical challenges of ethical integration, end-users provide insights into expectations and concerns, and regulatory bodies can explain legal decisions and future directions. Key questions will revolve around their primary ethical concerns, awareness of ML tools or frameworks, challenges in integrating ethics, and their prioritization of transparency, fairness, privacy, and accountability in ML. These surveys will be dispersed on ML Linkedin Groups, code forums such as StackOverflow, and through direct outreach to legislators and aides. This approach aims to unearth information not readily available online, such as personal experiences, nuanced user viewpoints, and preliminary regulatory directions.

QDA Miner Lite, a qualitative data analysis tool, will be utilized to analyze survey responses and information from the literature review, focusing on themes like transparency, bias, and privacy. Matrix coding will cross-examine themes against respondent groups, revealing unique concerns and viewpoints. Subsequently, a cluster analysis will group respondents based on thematic similarities, highlighting overarching patterns and stakeholder perspectives that should be embedded into the design of ML algorithms.

Conclusion

The digital evolution, as demonstrated by the bank's transformation, highlights the need for both efficient and ethically sound data systems. On the technical side, an enhanced data processing pipeline integrating Delta Lake, Apache Spark, and AWS Glue, can address bottlenecks and data inconsistencies. This ensures that decision-making has access to accurate and timely data, acting as a cornerstone for the bank's ongoing digital initiatives. Simultaneously, the ethical landscape of machine learning, as analyzed through the SCOT framework,

emphasizes an important relationship between technology and social groups' values. Through an in-depth exploration of stakeholder perceptions, the research anticipates formulating ethically sound design guidelines for machine learning systems. These guidelines aim to enhance system efficiency while ensuring they resonate with stakeholder values, fostering innovation that is both powerful and ethically responsible.

References

Altexsoft (2022) ETL Pipeline. Available at: https://content.altexsoft.com/media/2020/03

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316-334). Cambridge: Cambridge University Press.

Calvo, P. and Egea-Moreno, R. (2021). Ethics Lines and Machine Learning: A Design and Simulation of an Association Rules Algorithm for Exploiting the Data. *Journal of Computer and Communications*, 9, 17-37. https://doi.org/10.4236/jcc.2021.912002

Chen, L., Li, R., Liu, Y., Zhang, R., & Woodbridge, D. M. (2017). Machine learning-based product recommendation using Apache Spark. *Cloud & Big Data Computing*, 1–6. https://doi.org/10.1109/uic-atc.2017.8397470

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, 5, 40–60. https://doi.org/10.17351/ests2019.260

Geronazzo, A., & Ziegler, M. (2021). QMLEX: Data Driven Digital Transformation in marketing analytics. *2021 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata52589.2021.9671890

Goldsteen, A., Ezov, G., Shmelkin, R. et al. Data minimization for GDPR compliance in machine learning models. AI Ethics 2, 477–491 (2022). https://doi.org/10.1007/s43681-021-00095-8

Grandhi, B., Patwa, N., and Saleem, K. (2021). Data-driven marketing for growth and profitability. *EuroMed Journal of Business*, 16(4), 381-398. https://doi.org/10.1108/EMJB-09-2018-0054

Hassan H., Barakat, S.,, & Sarhan Q. (2021). Survey on serverless computing. *Journal of Cloud Computing: Advances, Systems and Applications*, 10(1), 1 - 29. https://doi.org/10.1186/s13677-021-00253-7

Hoofnagle, C. J., van der Sloot, B., & Zuiderveen Borgesius, F. (2019). The European Union General Data Protection Regulation: What It Is And What It Means. *Information & Communications Technology Law*, 28(1), 65-98. https://doi.org/10.2139/ssrn.3254511

Ikegwu, A. C., Nweke, H. F., Anikwe, C. V., Alo, U. R., & Okonkwo, O. R. (2022). Big data analytics for data-driven industry: a review of data sources, tools, challenges, solutions, and research directions. *Cluster Computing - The Journal of Networks, Software Tools and Applications,* 25(5), 3343 - 3387. https://doi.org/10.1007/s10586-022-03568-5

Lagioia, F., Sartor, G. (2020). *The impact of the general data protection regulation on artificial intelligence*, (G.Sartor,edito) Publications Office. https://data.europa.eu/doi/10.2861/293

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt

Misut, M., & Jurik, P. (2021). Datafication as a Necessary Step in the Processing of Big Data in Decision-making Tasks of Business. *Proceedings of CBU in Natural Sciences and ICT*, 2, 75-80. https://doi.org/10.12955/pns.v2.156

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021, June 1). Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds & Machines*, 31(2), 239 - 256. https://doi.org/10.1007/s11023-021-09563-w

Mourby, M., Ó Cathaoir, K., & Collin, C. B. (2021, November 1). Transparency of machine-learning in healthcare: The GDPR & European health law. *Computer Law & Security Review: The International Journal of Technology Law and Practice*, 43, 1 - 14. https://doi.org/10.1016/j.clsr.2021.105611

O'Neil, C. (2017). CHAPTER 5 CIVILIAN CASUALTIES: Justice in the Age of Big Data. In *Weapons of math destruction: How big data increases inequality and threatens democracy* (pp. 76–92). Penguin Random House.

Oussous, A., Benjelloun, F., Ait Lahcen, A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences,* 30(4), 431 - 448. https://doi.org/10.1016/j.jksuci.2017.06.001

Pinch, T. J., & Bijker, W. E. (1984, August 1). The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. *Social Studies of Science*, 14(3), 399 - 441.

Rader, E., & Gray, R. (2015). Understanding user beliefs about algorithmic curation in the facebook news feed. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/2702123.2702174

Ramirez-Asis, H., Silva-Zapata, M., Ramirez-Asis, E., Sharma, T., Durga, S., & Pant, B. (2022). A conceptual analysis on the impact of Big Data Analytics toward on digital marketing transformation. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 1651 - 1655. https://doi.org/10.1109/icacite53722.2 022.9823874

Tiwari, R. (2023). Ethical and societal implications of AI and machine learning. *International Journal of Scientific Research in Engineering and Management*, 07(01), 20 - 27. https://doi.org/10.55041/ijsrem17519

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... & Ghodsi, A. (2016). Apache Spark: A Unified Engine for Big Data Processing. *Communications of the ACM,* 59(11), 56 - 65. https://doi.org/10.1145/2934664

Zhang H., Lee I., Ali S., DiPaola D., Cheng Y., & Breazeal C. (2022). Integrating Ethics and Career Futures with Technical Learning to Promote AI Literacy for Middle School Students: An Exploratory Study. *International Journal of Artificial Intelligence In Education*, 33, 290-324. https://doi.org/10.1007/s40593-022-00293-3

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. (2017, May 10). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, 237, 350 - 361.