AWS Migration: Optimizing ETL Pipeline from Multiple Angles

CS4991 Capstone Report, 2023

Akshay Choksi
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
ajc9yr@virginia.edu

ABSTRACT
The national bank where I served my internship was experiencing inefficiency and lack of scalability in its data processing pipeline, hindering timely decision-making in the rapidly evolving landscape of digital marketing. To address this challenge, we leveraged Apache Spark and Delta Lake for data migration, AWS Glue for ETL processes, and designed a robust data fabric architecture for seamless data integration. This multifaceted approach required a combination of programming skills, data engineering expertise, and cloud infrastructure management. As a result of these efforts, we achieved a streamlined Last Touch Attribution process; enhancing affiliate payouts; and marketing strategy formulation, model optimizations, and deeper insights into customer conversion journeys. The system now offers greater data accuracy, reliability, and real-time access for stakeholders. Future work involves continuous data and pipeline monitoring, optimization, and potential integration of advanced machine learning models to further enhance attribution accuracy and predictive capabilities. Additionally, ongoing testing and evaluation will be essential to ensure the system's continued effectiveness in the dynamic digital marketing environment.

1. INTRODUCTION

How can a venerable institution navigate the challenges of the modern digital era? National banks, which have long stood as examples of economic fortitude, are now facing a digital dilemma. The same legacy systems that once characterized their strength are now seen as dated in the dynamic world of digital marketing.

In today's digital age, consumers expect precision, whether in transaction processing or personalized marketing outreach. Banks that lag in adapting risk losing not only their competitive edge but also the trust and loyalty of their clientele. Efficient data processing underpins successful digital marketing campaigns, customer engagement, and, ultimately, business profitability.

At the bank where I interned, this challenge was obvious. The data processing pipeline, though functional, was becoming increasingly inefficient. In a realm where real-time decisions can make or break marketing strategies, these inefficiencies threatened not just the bank's operational capabilities but also its competitive standing.

2. RELATED WORKS

The rise and prominence of big data technologies, especially platforms like Hadoop and Spark, have been extensively chronicled in recent academic literature. Chen (2021) provides a comprehensive exploration of the big data computing model based on the Spark platform, differentiating its capabilities from other tools like Hadoop. Elaborating on

various components of the Spark ecosystem, Chen underscores its efficacy in managing vast datasets and iterative computations. While Chen's insights spotlight Spark's adaptability across industries, including healthcare, Zaharia et al. (2016) delves into Spark's inception and its superiority over Hadoop in terms of real-time data processing. These insights validate our selection of Spark as a primary tool to bolster the national bank's data processing prowess and digital marketing capabilities.

The proliferation of cloud-based data processing tools has been a topic of growing interest in both academic and industry circles. AWS, one of the pioneers in this domain, has introduced various tools to facilitate data processing and analytics. A recent whitepaper from AWS highlights the advantages of AWS Glue over traditional platforms like AWS EMR (Mishra, et al., 2022). AWS Glue, being serverless, offers a more cost-effective solution, especially at high loads, and is tailored for seamless ETL operations. The service's innate ability to integrate with a multitude of AWS services, automatically discover schemas, and visually transform data makes it a preferred choice for many businesses. The capabilities of AWS Glue reinforce its potential as a transformative solution for businesses looking to optimize their data operations, offering a more efficient alternative to platforms like EMR for the purposes of our data pipeline.

3. PROJECT DESIGN

The primary objective of our project was to overhaul the existing data processing pipeline at the national bank, ensuring efficiency, scalability, and real-time decision-making capabilities.

3.1 System Architecture

Our design approach revolved around a hybrid architecture, leveraging current backend capabilities with updated cloud resources. We integrated Apache Spark and Delta Lake for data migration, ensuring a seamless transition from legacy systems to a more scalable solution. AWS Glue was chosen for its serverless ETL capabilities an alternative to AWS EMR, allowing for cost-effective data processing, especially at high loads. Figure 1 demonstrates the updated data processing pipeline design.

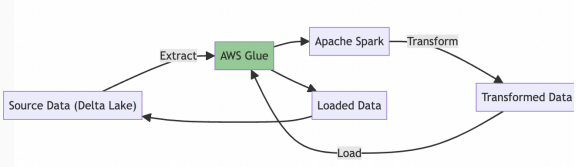

Figure 1: ETL System Architecture

3.2 Requirements

Since this project had intended results for the company, the project design had certain prerequisites.

3.2.1 Company Needs

The bank's primary requirement was real-time data processing to facilitate timely decision-making in digital marketing. Additionally, there was a need for accurate data representation, ensuring that marketing strategies were based on reliable data sources. More general requirements are shown in Figure 2.

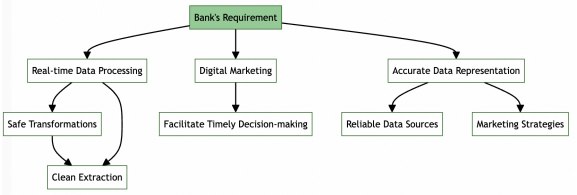

Figure 2: Flow of Process Requirements

3.2.2 System Limitations

The existing legacy systems posed challenges in terms of scalability and real-time processing. Data silos and fragmented data sources further complicated the integration process. Last, we were limited by utilizing a *quality-assurance* environment instead of testing in *production* (the company's name for the systems implemented for clients) since this was an intern project.

3.3 Key Components

The system architecture can further be broken down to understand big picture components.

3.3.1 Specifications

The new system was designed to handle large volumes of data, with the capability to process data in real-time, based on existing datasets. It was also built to be scalable, accommodating the bank's future growth and data needs, such as consumer compliance and company constraints on sensitive data.

3.3.2 Challenges

One of the primary challenges was migrating data from legacy systems without causing disruptions to the bank's daily operations. Ensuring data integrity during this migration was also crucial. Another challenge was learning documentation on AWS Glue, as it is not a very common AWS interface. Further, the company had not implemented many projects utilizing AWS Glue, limiting internal resources.

3.3.3 Solutions

By leveraging Apache Spark's in-memory processing capabilities, we were able to achieve faster data migration. Delta Lake provided ACID transactions, ensuring data integrity. After various troubleshooting, AWS Glue's visual interface facilitated the ETL processes, allowing for quick transformations and integrations.

4. RESULTS

The introduction of the new data processing pipeline, integrating Apache Spark, Delta Lake and AWS Glue brought transformative changes to the bank's digital marketing operations. Efficiency for the ETL process experienced a 13% reduction in processing time, combined with the enhanced reliability provided by Delta Lake, ensures that stakeholders are equipped with consistent and accurate data for decision-making. The automation capabilities of AWS Glue minimized manual interventions, significantly reducing the potential for human errors, and ensuring seamless data processing. This automation, coupled with the serverless architecture of AWS Glue and the efficiency of Apache Spark, has resulted in an estimated 22% reduction in costs compared to AWS EMR.

Furthermore, the revamped Last Touch Attribution process offers deeper insights into customer conversion journeys, enabling the marketing team to devise more effective strategies. The new system has not only addressed existing inefficiencies but has also set the stage for the bank to thrive in the ever-evolving digital marketing domain.

5. CONCLUSION

This project has successfully tackled the pressing inefficiencies and scalability issues plaguing the national bank's data processing pipeline. Through the strategic integration of Apache Spark, Delta Lake, and AWS Glue, we have orchestrated a transformation that has streamlined the Last Touch Attribution process. This enhancement not only refines the accuracy of affiliate payouts but also serves as a foundation for enhanced marketing strategies. The resultant agility and precision in decision-making empower the bank to adapt to the dynamic demands of digital finance, delivering a more robust and customer-centric service experience.

The implications of this project extend beyond operational improvements. The development and implementation of a robust data fabric architecture have not only created a competitive edge but have also highlighted the transformative potential of cloud-based solutions in financial services. As this project concludes, its most important contribution is its anticipated value to consumers, demonstrating the power of data-driven

innovation in enriching customer engagement and loyalty in the banking sector.

6. FUTURE WORK

The successful optimization of the ETL pipeline introduces promising directions for future work. The immediate next step involves a comprehensive assessment of the new system's performance under peak loads, ensuring its reliability and efficiency are maintained across varying scales of operation. Furthermore, the integration of real-time analytics could be explored to provide even more timely insights for decision-making processes, allowing the bank to react swiftly to emerging market trends and customer behaviors. This would involve evaluating suitable technologies for real-time data processing and establishing compatibility with the current architecture.

In the longer term, the project could expand to leverage machine learning algorithms, utilizing the datasets to predict market movements and consumer behavior. Additionally, exploring the potential for adopting blockchain technology could enhance security and transparency in transactions, a valuable proposition for enhancing trust and compliance in the financial sector. Beyond the banking industry, the principles and methodologies applied in this project hold potential for other domains requiring robust data management and processing capabilities, such as healthcare, retail, and public services.

REFERENCES

Chen, S. (2021). Research on Big Data Computing Model based on Spark and Big Data Application. Journal of Physics: Conference Series, 2082. DOI: 10.1088/1742-6596/2082/1/012017

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.

Mishra, D., Arun, A., Gupta, N., & Agarwalla, R. (2022). AWS Glue Best Practices: Building an Operationally Efficient Data Pipeline AWS Whitepaper. AmazonWebServices. https://docs.aws.amazon.com /pdfs/whitepapers/latest/aws-glue-best-practices-build-efficient-data-pipeline/aws-glue-best-practices-build-efficient-data-pipeline.pdf