New Math
DATA

CASE STUDY

Optimizing Data Operations: Customer's Transition to AI-Driven Extraction and Data Lakehouse Efficiency

Executive Summary

Customer, leveraging cutting-edge technology in data analytics, partnered with us to overcome challenges in their expansive data collection and processing operations. Tasked with extracting and refining vast datasets from myriad sources, the complexity and maintenance of their system escalated, alongside rising costs associated with data warehousing. Our solution harnessed AI and Language Models to develop a universal extractor, significantly reducing the need for custom coding and streamlining the integration of new data sources. Furthermore, we facilitated a strategic migration from a costly data warehouse system to a cost-effective Data Lakehouse system. This transformation not only optimized operational efficiency and cost but also enhanced agility in onboarding new data points, demonstrating a marked decrease in development and maintenance expenses while accelerating data sharing capabilities.

Customer Description

Founded by experts in the data analytics industry, the customer uses cutting-edge technology and efficient workflow design to collect, curate, and distribute foundational reference datasets.

Description of Service

Our client has been gathering information from numerous websites to collect public datasets. They also pull in data from various third-party sources. To ensure the data we provide is top-notch — accurate and comprehensive — it undergoes several layers of processing, filtering, and application of business logic.

Our system operates on a multitude of container instances to fetch data from thousands of different websites. Each category of website demands custom scripting to extract the specific information needed for our purposes. Creating and maintaining this custom code is resource-intensive. As our client adds new sources, the code becomes more intricate and challenging to manage.

Recently, we've identified numerous websites for data extraction and encountered issues with existing sources that need fixing. This involves substantial efforts in discovering the right patterns for data extraction, as well as developing, testing, and integrating new data sources. The task backlog is substantial, requiring a team of developers to accomplish. Currently, our focus is on maintaining their existing system and incorporating new websites and data sources as needed.

Another challenge we're addressing is the rising cost of utilizing a data warehouse for all transformation tasks. To manage this, we are engaged in migrating our client's data from the current data warehouse system to a more cost-effective Data Lake House system. This migration effort is essential to optimize costs and maintain an efficient data-sharing platform.

Description of Solution

Our team, comprised of a few architects and multiple engineers, collaborated with the company to develop and maintain their application and data platform. Our focus was on finding a better and more cost-effective solution to reduce the total cost of ownership. We aimed to streamline the onboarding process for new websites and data sources, all while maintaining the data platform efficiently with fewer resources and developers.

We utilized AI to enhance agility and quickly integrate new data sources. The application was seamlessly connected with Language Models (LLMs) to create a universal extractor. This extractor can pull data from both new and existing websites without the need for custom code. Our objective was to leverage the capabilities of

LLMs to identify all relevant information and any additional data from sources, minimizing the development and maintenance efforts.

The second major task involved migrating from a data warehouse system to a Data Lakehouse system. This move was intended to consolidate all data assets in one place at a lower cost. The migration process included moving data from Snowflake to Databricks Lakehouse, and the transformations were shifted to Dbt and PySpark. Our NMD team successfully designed a migration plan, executed the migration, and thoroughly validated all aspects of the system. We smoothly transitioned to the new platform with minimal downtime.

Description of Outcome

The integration of AI into our application exceeded expectations, allowing us to effortlessly onboard new websites and extract additional data points with minimal effort. This initiative resulted in a substantial reduction in ongoing development and maintenance costs, enabling our client to operate at an accelerated pace.

Furthermore, our client achieved a further reduction in their total cost of ownership by transitioning from a data warehouse to a data lakehouse system. This move not only streamlined internal processes but also facilitated a more efficient and rapid sharing of data assets with their customers.

Lessons Learned

As the Language Models (LLMs) offerings are still in the early stages, we encountered several challenges with API limitations. Expanding the application beyond a certain point proved to be difficult due to the restrictions imposed by the LLMs API endpoint.

Current Relationship

NMD is consistently collaborating with customer to enhance automation within their system, aiming to create a more efficient and robust application and data platform.