How Rocket Companies modernized their data science solution on AWS

February 21st, 2025

Venkata Santosh Sajjan Alla
Sr. Solutions Architect,
AWS Financial Services

Dian Xu
Sr. Director of Engineering in Data,
Rocket Companies

Joel Hawkins
Principal Data Scientist,
Rocket Companies

Abstract

In this article we delve into the architectural blueprint for the successful modernization of a national-scale data science platform for Rocket Companies, a leader in the U.S. financial and homeownership industry. We showcase how a strategic migration from legacy systems to a high-velocity, cloud-native solution on AWS sets a new standard for the financial services sector. The resulting framework has become a cornerstone of the company's mission to simplify the American homeownership journey, now powering over 10 million automated AI decisions daily and increasing the number of production models by five times. This architectural transformation drastically reduced the speed to delivery for new data-driven projects from eight weeks to under one hour, providing a proven, replicable model for U.S. financial institutions to build the robust, scalable AI infrastructure needed to drive national economic innovation and efficiency.

Introduction

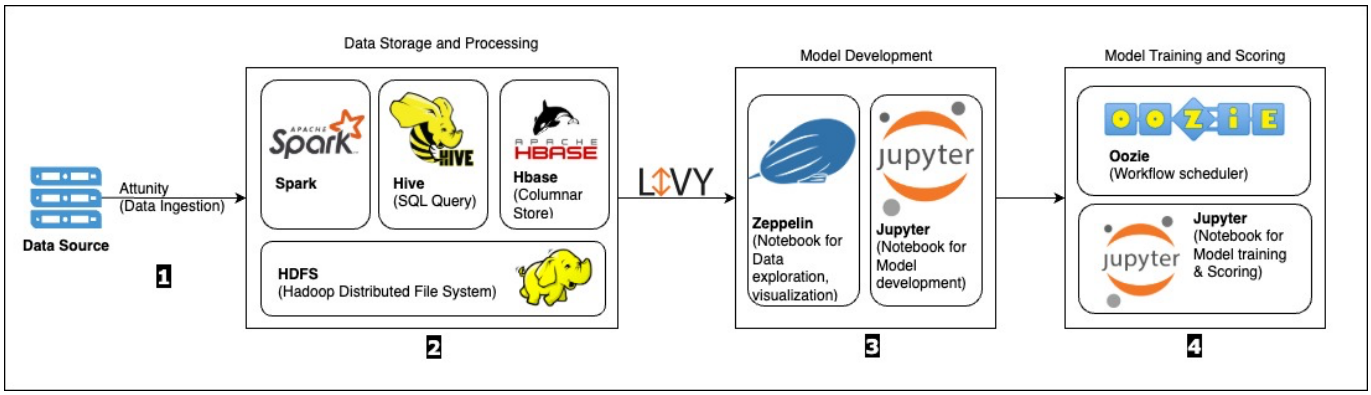
In the modern U.S. financial services landscape, the demand for rapid, data-driven decision-making has exposed the limitations of legacy data science platforms. Many institutions face significant challenges with scalability, operational stability, and the speed of innovation, which hinders their ability to serve customers effectively and compete in a dynamic market. The complexity of the homeownership process, a cornerstone of the American economy, further amplifies these challenges. Overcoming these hurdles requires a fundamental architectural shift from monolithic, on-premises systems to scalable, cloud-native frameworks.

We discuss a detailed case study and architectural blueprint for such a transformation, demonstrating how migrating to a managed services environment on AWS can unlock massive efficiency gains and power AI at a national scale.

Central to this transformation is Amazon SageMaker AI, a fully managed machine learning (ML) service that enables data scientists and developers to build, train, and deploy ML models quickly and at scale. This report will illustrate how SageMaker, in conjunction with foundational services like Amazon Simple Storage Service (Amazon S3) for data storage and Amazon EMR for big data processing, provides a cohesive and powerful ecosystem for modern data science in the demanding financial services industry.

Overview

The following diagram illustrates the high-level reference architecture of the legacy data science environment prior to the modernization effort.



This legacy architecture was comprised of several key components that faced operational challenges.

Data exploration and model development were conducted using well-known machine learning (ML) tools such as Jupyter or Apache Zeppelin [1] notebooks. Apache Hive [2] was used to provide a tabular interface to data stored in HDFS, and to integrate with Apache Spark [3] SQL. Apache HBase [4] was employed to offer real-time key-based access to data. Model training and scoring was performed either from Jupyter notebooks or through jobs scheduled by Apache's Oozie [5] orchestration tool, which was part of the Hadoop [6] implementation.

Despite its functionality, this architecture presented several constraints that limited its effectiveness and hindered innovation. These are detailed in the following section.

Problem Statement: Legacy Architecture Constraints

- **Accessibility limitations:** The data lake was stored in HDFS and only accessible from the Hadoop environment, hindering integration with other data sources. This also led to a backlog of data that needed to be ingested.

- **Steep learning curve for data scientists:** Many of data scientists did not have experience with Spark, which had a more nuanced programming model compared to other popular ML solutions.

- **Responsibility for maintenance and troubleshooting:** Rocket's DevOps/Technology team was responsible for all upgrades, scaling, and troubleshooting of the Hadoop cluster. This resulted in a backlog of issues with both vendors that remained unresolved.

- **Balancing development vs. production demands**: Organizations had to manage work queues between development and production, which were always competing for the same resources.

- **Deployment challenges:** Organizations sought to support more real-time and streaming inferencing use cases, but this was limited by the capabilities of the existing tools.

- **Inadequate data security and DevOps support:** The previous solution lacked robust security measures, and there was limited support for development and operations of the data science work.

Rocket's migration journey

To address the legacy data science environment challenges, Rocket decided to migrate its ML workloads to the Amazon SageMaker AI[7] suite. This would allow us to deliver more personalized experiences and understand our customers better. To promote the success of this migration, we collaborated with the AWS team to create automated and intelligent digital experiences that demonstrated Rocket's understanding of its clients and kept them connected.

We implemented an AWS multi-account strategy, standing up Amazon SageMaker Studio[8] in a build account using a network-isolated Amazon VPC. This allows us to separate development and production environments, while also improving our security stance.

We moved our new work to SageMaker Studio and our legacy Hadoop workloads to Amazon EMR[9], connecting to the old Hadoop cluster using Livy and SageMaker notebooks[10] to ease the transition. This gives us access to a wider range of tools and technologies, enabling us to choose the most appropriate ones for each problem we're trying to solve.

In addition, we moved our data from HDFS to Amazon Simple Storage Service[11] (Amazon S3), and now use Amazon Athena[12] and AWS Lake Formation[13] to provide proper access controls to production data. This makes it easier to access and analyze the data, and to integrate it with other systems. The team also provides secure interactive integration through Amazon Elastic Kubernetes Service[14] (Amazon EKS), further improving the company's security stance.

SageMaker AI has been instrumental in empowering our data science community with the flexibility to choose the most appropriate tools and technologies for each problem, resulting in faster development cycles and higher model accuracy. With SageMaker Studio, our data scientists can seamlessly develop, train, and deploy models without the need for additional infrastructure management.

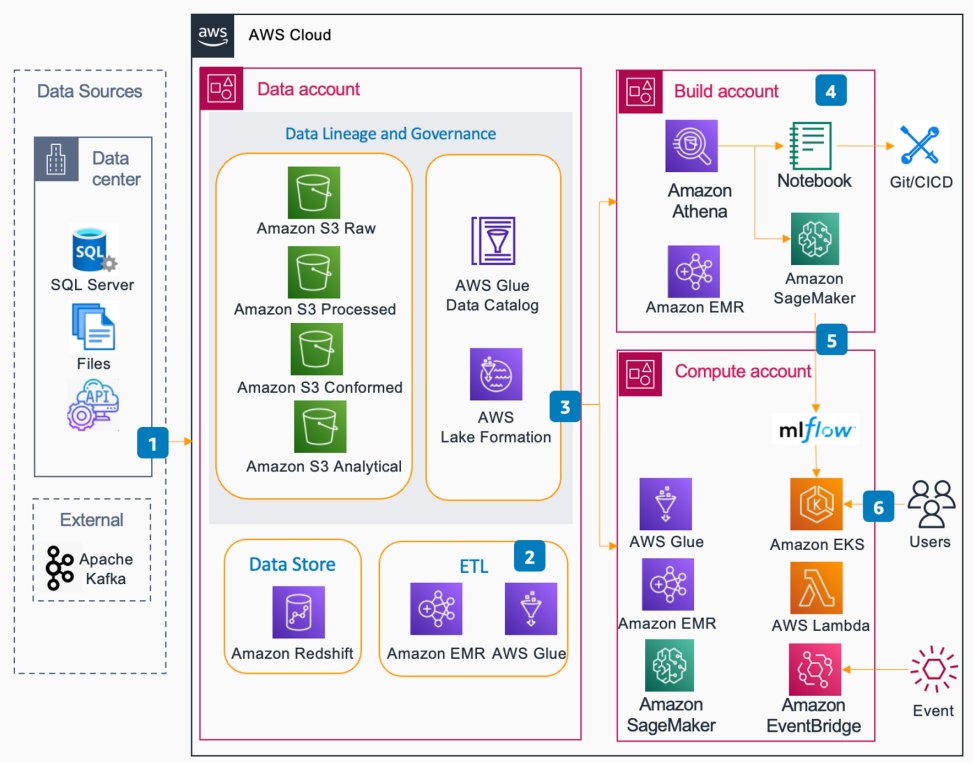
As a result of this modernization effort, SageMaker AI enabled Rocket to scale our data science solution across Rocket Companies and integrate using a hub-and-spoke model. The ability of SageMaker AI to automatically provision and manage instances has allowed us to focus on our data science work rather than infrastructure management, increasing the number of models in production by five times and data scientists' productivity by 80%.

Our data scientists are empowered to use the most appropriate technology for the problem at hand, and our security stance has improved. Rocket can now compartmentalize data and compute, as well as compartmentalize development and production. Additionally, we are able to provide model tracking and lineage using Amazon SageMaker Experiments[15] and artifacts discoverable using the SageMaker model registry and Amazon SageMaker Feature Store[16]. All the data science work has now been migrated onto SageMaker, and all the old Hadoop work has been migrated to Amazon EMR.

Overall, SageMaker AI has played a critical role in enabling Rocket's modernization journey by building a more scalable and flexible ML framework, reducing operational burden, improving model accuracy, and accelerating deployment times.

The successful modernization allowed Rocket to overcome our previous limitations and better support our data science efforts. We were able to improve our security stance, make work more traceable and discoverable, and give our data scientists the flexibility to choose the most appropriate tools and technologies for each problem. This has helped us better serve our customers and drive business growth.

Rocket's new data science solution architecture on AWS is shown in the following diagram.



The solution consists of the following components:

1. Data ingestion: Data is ingested into the data account from on-premises and external sources.
2. Data refinement: Raw data is refined into consumable layers (raw, processed, conformed, and analytical) using a combination of AWS Glue[17] extract, transform, and load (ETL) jobs and EMR jobs.
3. Data access: Refined data is registered in the data account's AWS Glue Data Catalog and exposed to other accounts via Lake Formation. Analytic data is stored in Amazon Redshift[18]. Lake Formation makes this data available to both the build and compute accounts. For the build account, access to production data is restricted to read-only.
4. Development: Data science development is done using SageMaker Studio. Data engineering development is done using AWS Glue Studio. Both disciplines have access to Amazon EMR for Spark development. Data scientists have access to the entire SageMaker ecosystem in the build account.
5. Deployment: SageMaker trained models developed in the build account are registered with an MLFlow instance. Code artifacts for both data science activities and data engineering activities are stored in Git. Deployment initiation is controlled as part of CI/CD.
6. Workflows: We have a number of workflow triggers. For online scoring, we typically provide an external-facing endpoint using Amazon EKS with Istio. We have numerous jobs that are launched by AWS Lambda[19] functions that in turn are triggered by timers or events. Processes that run may include AWS Glue ETL jobs, EMR jobs for additional data transformations or model training and scoring activities, or SageMaker pipelines and jobs performing training or scoring activities.

Migration impact

We've evolved a long way in modernizing our infrastructure and workloads. We started our journey supporting six business channels and 26 models in production, with dozens in development. Deployment times stretched for months and required a team of three system engineers and four ML engineers to keep everything running smoothly. Despite the support of our internal DevOps team, our issue backlog with the vendor was an unenviable 200+.

Today, we are supporting nine organizations and over 20 business channels, with a whopping 210+ models in production and many more in development. Our average deployment time has gone from months to just weeks sometimes even down to mere days! With just one part-time ML engineer for support, our average issue backlog with the vendor is practically non-existent. We now support over 120 data scientists, ML engineers, and analytical roles. Our framework mix has expanded to include 50% SparkML models and a diverse range of other ML frameworks, such as PyTorch and scikit-learn. These advancements have given our data science community the power and flexibility to tackle even more complex and challenging projects with ease.

The following table compares some of our metrics before and after migration.

	Before Migration	After Migration
Speed to Delivery	New data ingestion project took 4–8 weeks	Data-driven ingestion takes under one hour
Operation Stability and Supportability	Over a hundred incidents and tickets in 18 months	Fewer incidents: one per 18 months
Data Science	Data scientists spent 80% of their time waiting on their jobs to run	Seamless data science development experience
Scalability	Unable to scale	Powers 10 million automated data science and AI decisions made daily

Lessons learned

Throughout the journey of modernizing our data science solution, we've learned valuable lessons that we believe could be of great help to other organizations who are planning to undertake similar endeavors.

First, we've come to realize that managed services can be a game changer in optimizing your data science operations.

The isolation of development into its own account while providing read-only access to production data is a highly effective way of enabling data scientists to experiment and iterate on their models without putting your production environment at risk. This is something that we've achieved through the combination of SageMaker AI and Lake Formation.

Another lesson we learned is the importance of training and onboarding for teams. This is particularly true for teams that are moving to a new environment like SageMaker AI. It's crucial to understand the best practices of utilizing the resources and features of SageMaker AI, and to have a solid understanding of how to move from notebooks to jobs.

Lastly, we found that although Amazon EMR still requires some tuning and optimization, the administrative burden is much lighter compared to hosting directly on Amazon EC2. This makes Amazon EMR a more scalable and cost-effective solution for organizations who need to manage large data processing workloads.

Conclusion

This post provided overview of the successful partnership between AWS and Rocket Companies. Through this collaboration, Rocket Companies was able to migrate many ML workloads and implement a scalable ML framework. Ongoing with AWS, Rocket Companies remains committed to innovation and staying at the forefront of customer satisfaction.

Don't let legacy systems hold back your organization's potential. Discover how AWS can assist you in modernizing your data science solution and achieving remarkable results, similar to those achieved by Rocket Companies.

References

[1] Apache Zeppelin - A web-based notebook that enables data-driven, interactive data analytics and collaborative documents with SQL, Scala, Python, R and more. https://zeppelin.apache.org/

[2] Apache Hive - Data warehouse software that facilitates reading, writing, and managing large datasets residing in distributed storage using SQL. https://hive.apache.org/

[3] Apache Spark - A multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters. https://spark.apache.org/

[4] Apache HBase - A non-relational, distributed database modeled after Google's Bigtable. https://hbase.apache.org/

[5] Apache Oozie - A workflow scheduler system to manage Apache Hadoop jobs. https://oozie.apache.org/

[6] Apache Hadoop - A framework that allows for the distributed processing of large data sets across clusters of computers. https://hadoop.apache.org/

[7] Amazon SageMaker AI - A fully managed service to build, train, and deploy machine learning (ML) models. https://aws.amazon.com/sagemaker-ai/

[8] Amazon SageMaker Studio - A single web-based interface for end-to-end ML development https://aws.amazon.com/sagemaker-ai/studio/

[9] Amazon EMR - A cloud big data platform for running large-scale distributed data processing jobs. https://aws.amazon.com/emr/

[10] Amazon SageMaker notebooks - Fully managed notebooks in JupyterLab for exploring data and building ML models https://aws.amazon.com/sagemaker-ai/notebooks/

[11] Amazon Simple Storage Service (Amazon S3) - An object storage service offering industry-leading scalability, data availability, security, and performance. https://aws.amazon.com/s3/

[12] Amazon Athena - A serverless, interactive analytics service built on open-source frameworks. https://aws.amazon.com/athena/

[13] AWS Lake Formation - A service that makes it easy to set up a secure data lake in days. https://aws.amazon.com/lake-formation/

[14] Amazon Elastic Kubernetes Service (Amazon EKS) - A managed container service to run and scale Kubernetes. https://aws.amazon.com/eks/

[15] Machine Learning Experiments using Amazon SageMaker with MLflow - Efficiently manage machine learning models and generative AI application experiments at scale using MLflow https://aws.amazon.com/sagemaker-ai/experiments/

[16] Amazon SageMaker notebooks - Fully managed notebooks in JupyterLab for exploring data and building ML models https://aws.amazon.com/sagemaker-ai/notebooks/

[17] AWS Glue - A serverless data integration service that makes it easy to discover, prepare, and combine data. https://aws.amazon.com/glue/

[18] Amazon Redshift - A fully managed, petabyte-scale data warehouse service in the cloud. https://aws.amazon.com/redshift/

[19] AWS Lambda - A serverless, event-driven compute service. https://aws.amazon.com/lambda/