

Fannie Mae Automates Data Lake Design Process to Accelerate Insight

Waterline Data's Extensive APIs Automate Data Cataloging from High-Velocity Applications

Fannie Mae Backstory

Fannie Mae is a leader in providing housing finance for homebuyers and renters across the United States. They help make fixed-rate mortgage and affordable rental housing possible for millions of Americans.

The Challenge

Founded in 1938, Fannie Mae deals with the same challenges as many mature corporations, including legacy environments and data silos. As a \$110 billion company that is the leading source of financing for mortgage lenders, Fannie Mae's increasingly data-centric business wanted to transition to an agile, more responsive data lake. They sought to create a modern data environment that ensured the right data got to the right person at the right time.

Fannie Mae set a governance standard whereby every dataset and field in the data lake was completely documented. In fact, each dataset goes through a design process where it gets curated and assigned a unique identifier, which stays with it no matter where it gets copied. Each dataset also has an elaborate set of properties that have to be filled out before the identifier can be issued.

While this process made the data more accessible, it was taking too much time. It could take days, weeks, or even months from the time the design was approved to when data was actually generated and transferred from the IMR design time system into the data lake.

Meanwhile, several of Fannie Mae's high-velocity apps continued to generate more than 10 million new files every day, clogging up the slow design process even further. These new files also needed to be integrated into the data lake, which required an API-based automated solution.

Challenges

- Efficiently pre-populating data lake with all necessary data set properties
- Automation required an API-based solution
- High daily volume ingestion of datasets

Results

- Automated solution capable of cataloging greater than 10 million files per day with all associated properties
- Searchable UI with custom search properties to deliver desired data quickly and efficiently

Benefits

- ✓ Self-service "marketplace" data catalog for business users
- ✓ Cataloged custom properties attached to each dataset
- ✓ APIs to automatically process datasets
- ✓ Millions of files cataloged per day

Applications Integrated

AWS

The Solution

As part of its wholesale data transformation to a modern data infrastructure, Fannie Mae integrated Waterline Data's extensive APIs to enable high-volume applications that generate millions of files daily for pre-registration of ingested datasets.

The interface allows validation and management of metadata for different roles, including metadata analysts, data stewards, data governors, and business data officers. Business data officers at Fannie Mae "ensure that data is fully owned and cared for by business leaders, and that new initiatives consider the creation, ongoing quality, and effective usage of data from the outset."¹

Metadata versioning allows capture and display of technical metadata provided by the ingesting application, including file location, file size, file format, time of ingestion, partition, etc.

The solution can catch unresolved schema evolution in order to produce discrepancy reports between reported and inferred schema. That data is then made available to business end-users in a robust self-service "marketplace" UI that features complex custom properties presented simply and cohesively to enable end users to quickly find and utilize the data.



Fannie Mae

“With Waterline we were able to fully automate and accelerate the cataloging and searchability of data to deliver game changing value to the business

Prakash Jaganathan
Data Management Leader, Fannie Mae

Ultimately, this results in faster analytics and insights, which translates to better business outcomes. As Fannie Mae Chief Data Officer Scott Richardson says, "We are engaged in thinking about business strategy through the lens of furthering our mission and improving the customer experience with data."¹

¹"How Fannie Mae is Creating a Modern Data Environment," *Forbes*

Waterline Data automates data discovery, compliance and the ability to take action on data by using a combination of artificial intelligence, machine learning, ratings and reviews, and tribal knowledge to deliver a AI-driven Data Catalog. Our customers spend less time searching for data and more time using it to derive value while complying with data governance mandates such as GDPR. Our data catalog is best of breed for large enterprises with big data implementations from multiple sources or migrations to the cloud.

