Pentaho

Faster Insights Drive Better Results

Challenge
Make millions of files of mission-critical business data rapidly available to business analysts every day.

Solution
Deploy Pentaho Data Catalog to automate profiling and tagging of data sets and provide context for analyses.

Outcome
Automate processes to eliminate data anomalies with AI, accelerate data delivery to analysts, and facilitate compliance.

Fannie Mae

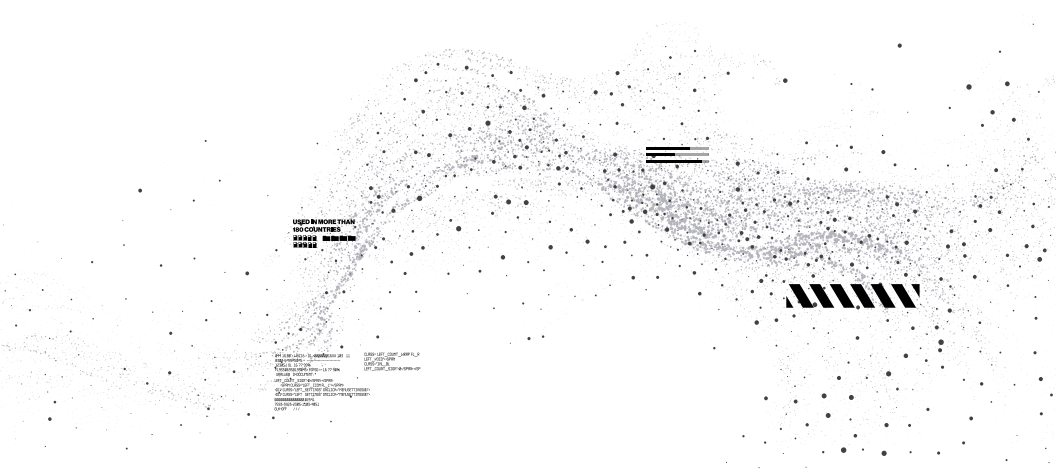### Fannie Mae®

INDUSTRY
Financial Services

LOCATION
United States

PENTAHO PRODUCT(S)
Pentaho Data Catalog

"We were looking for a solution that could handle more than 10 million new files every day to keep our enterprise data lake up to date."

Rohny Kolli, Data Engineering – Advanced Analytics Enablement, Fannie Mae

Challenge

Fannie Mae enabled the acquisition of more than 2 million home purchases and refinancings, and financing of approximately 598,000 rental units across the United States in 2022. Today, Fannie Mae is an increasingly digital and data-centric business. To leverage all its business data across new and legacy applications, as well as break down existing data silos, the company wanted to create an agile and dynamic enterprise data lake.

Rohny Kolli, Data Engineering – Advanced Analytics Enablement at Fannie Mae, says: "Our goal was to build a modern, state-of-the-art data platform for business analysts and decision-makers across the company. We wanted to enable fast, data-driven decisions — which meant we had to make it easier to get the right data to the right people at the right time."

Fannie Mae started by designing a comprehensive process to manage its enterprise data lake. Every single one of its 15,000 datasets went through an initial registration process to assign a unique identifier, and every field had to be documented manually. This approach increased compliance and transparency by helping to identify datasets at every stage of the analytics and reporting process — but the need to add an elaborate set of metadata to every dataset made the process slow.

"With our existing solution, it could take weeks or even months before new datasets would be registered in our data lake and made available to our business analysts and data scientists," adds Rohny Kolli. "To respond faster to new data that is being continuously generated by our high-velocity apps, we had to automate this process. We were looking for a solution that could handle more than 10 million new files every day to keep our enterprise data lake up to date."

> "
> Pentaho Data Catalog gives us real-time insights into how our data is changing over time and helps us ensure that all our data files are stored in the right places to support smooth, standardized operations and compliance with internal guidelines."
>
> **Rohny Kolli**
> Data Engineering Manager – Advanced Analytics Enablement, Fannie Mae

Solution

To help establish a faster and more dynamic data infrastructure, Fannie Mae selected Pentaho Data Catalog as a centralized, data-agnostic tool to accelerate data availability. The software runs fully in the cloud on Amazon Web Services (AWS) across multiple availability zones with auto-scaling to ensure fast performance and business continuity. It processes tens of millions of files and related attributes and aggregates them into thousands of high-level datasets that are easy for the business team to consume and reference for actionable insights.

To transform its data pipeline, Fannie Mae now relies on process automation based on the Pentaho Data Catalog API. This enables the company to connect its wide range of business applications to the enterprise data lake and update datasets on a daily basis.

Pentaho Data Catalog performs an automated pre-registration step, using machine learning and AI to validate and tag metadata and detect sensitive data. It then makes everything immediately available to the company's metadata analysts, data stewards, data governors and business data officers for further processing and analytics.

Built-in metadata versioning helps Fannie Mae keep track of changes in its data sources and better understand the context of its business data. The data-agnostic solution highlights changes in storage location, file size, file format and many other technical details that can help the team to tune and optimize the data processing.

"Pentaho Data Catalog gives us real-time insights into how our data is changing over time and helps us ensure that all our data files are stored in the right places to support smooth, standardized operations and compliance with internal guidelines," says Rohny Kolli. "The solution can catch unresolved schema issues and produce discrepancy reports, helping our various teams ensure high data quality and compliance."

Vestibulum fringilla facilisis nunc id tempus. Sed aliquam eu magna sed interdum. Etiam tincidunt malesuada ex eu viverra. Vivamus vitae estibulum condimentum faucibus nibh, vitae mat. Insert at least 2 stats in the grey bar. This panel should include three numerical values, a title for each and a description.

Outcome

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec ultrices mauris ante, eget Accessing critical business information is now easier than ever. "Using Pentaho Data Catalog, we have created a data-agnostic self-service offering for our business users," adds Rohny Kolli. "Staff can flexibly search our enterprise data lake with a user-friendly and intuitive interface to gain a 360-degree view of our business data. The search results provide a simple overview, so data stewards, business analysts and data scientists can find the right datasets with the custom data properties they need quickly and efficiently."

To unlock further insights and provide meaningful context to business users, Fannie Mae is now using the solution to tag its data — for example, to highlight sensitive and personal information and classify more than 400 key data elements (KDEs).

Ultimately, these solution elements enable faster analytics and insights, which translate into better business outcomes. Rohny Kolli concludes: "With Pentaho Data Catalog, we are integrating millions of files each day into our enterprise data lake. The solution enables data profiling and tagging to gain valuable insights, identify anomalies immediately, and support our data governance management to facilitate compliance."

With Pentaho Data Catalog, we are integrating millions of files each day into our enterprise data lake. The solution enables data profiling and tagging to gain valuable insights, detect anomalies immediately, and support our data governance management to facilitate compliance.

Automated processing of 10 million files per day supported decision-making to provide $684 billion in liquidity to the mortgage market in 2022.

10M files per day

$684B in liquidity per year

About FannieMae

Fannie Mae is a leading financial services company providing lenders with a reliable source of mortgage financing across the United States. By purchasing mortgage loans, the company helps lenders to offer new mortgages to more people. In doing so, Fannie Mae expands access to affordable housing opportunities, supporting renters, homebuyers and homeowners. With its approximately 8,000 employees, Fannie Mae enabled the acquisition and financing of approximately 2.6 million home purchases, refinancings, and rental units in 2022.

About Pentaho

Trusted by over 70% of the Fortune 100, Pentaho's end-to-end data intelligence platform enables the foundational and AI-grade data fitness that drives faster and more confident decision-making.

Pentaho

P-BTD-CS-Fannie-Mae-29Jan25-A