Lakehouse for Financial Services

A big book of use cases

databricks

Contents

Lakehouse for Financial Services

When it comes to innovating with data and AI, the financial services industry lags behind other industries. The lack of innovation is often attributed to stringent regulatory requirements. However, another significant impediment is vendor lock-in to outdated legacy technologies that are difficult to maintain and scale.

Competition from fintechs as well as paradigm shifts, such as open banking, mean data and AI must play an increasingly central role for financial services institutions (FSIs) that wish to survive and thrive. In order to unlock the value of their data and to offer differentiated products to their customers, FSIs need to adopt modern cloud and data analytics technologies — where open formats, scalability and lack of vendor lock-in are core components.

Databricks Lakehouse for Financial Services enables the banking, capital markets and insurance sectors to generate business value from all their data assets. It is purpose-made to address the unique technical, business and regulatory requirements for FSIs by providing additional capabilities on top of the lakehouse platform, such as use-case accelerators, data sharing capabilities, as well as certified implementation partners.

To get ahead of the curve, FSIs will need to turn to solutions that are purpose-built for innovation. Based on open standards and open source technologies, Databricks Lakehouse for Financial Services provides the tools necessary to deliver a range of business solutions, from hyper-personalized customer experiences to AI-driven risk management and data governance practices.

Over the following eight example use cases, covering everything from personalization to fraud detection, we demonstrate how Databricks Lakehouse for Financial Services is critical to maintaining a distinct advantage in the financial services industry. This guide also includes resources in the form of Solution Accelerators and reference architectures to help as you embark on your own journey to drive more sustainable finance and profitable investments through the use of data and AI.

Customers That Innovate With Databricks Lakehouse for Financial Services

Some of the top financial institutions in the world turn to Databricks Lakehouse for Financial Services to bring AI-driven innovation and excellence to their customers.

ABN·AMRO · S&P Global · TD Bank · Northwestern Mutual · AIA · Schroders · Nasdaq · Bread · McKinsey & Company · nubank

databricks

Modern Cloud Data Infrastructure

Overview

Financial services institutions are now turning to cloud and data analytics technologies to help them make sense of vast amounts of data and better adapt to a rapidly shifting market. AI is now playing a key role in driving innovation by giving financial institutions the ability to personalize their services and usher in a more open data economy.

Common challenges

Vendor lock-in to legacy technologies has led to a slowdown in innovation in the financial services industry. These legacy technologies also result in data silos, making it difficult to leverage data effectively. The constraints of managing these unwieldy systems, along with a general lack of investment in R&D, has led to inefficient processes, outdated skill sets and slower advancements.

Lack of customer insights

When data is siloed, teams get an incomplete view of customer segments, preferences and transactional behavior, resulting in missed cross-selling opportunities.

Lack of reliability

Suboptimal processes moving data in and out of silos coupled with proprietary systems lead to poor data quality, low AI efficacy and high operation risk. Because legacy technologies don't tap into unstructured and alternative data sets for insights, the data available to financial institutions can't be relied upon to help make informed investment decisions based on trends, behaviors and risks.

Lack of governance

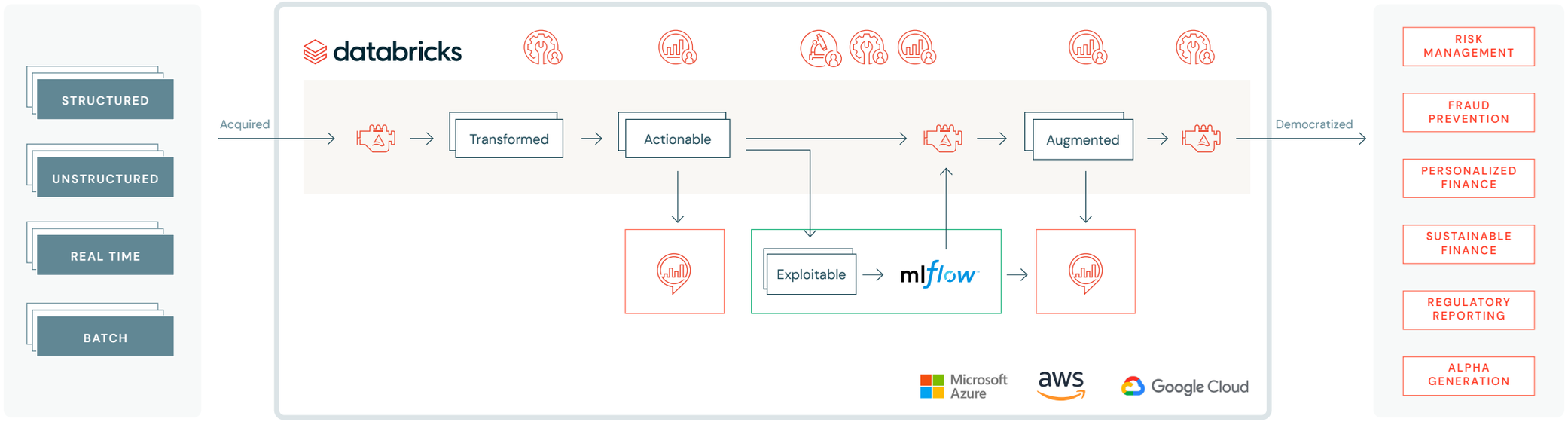
A lack of data agility and model reproducibility makes it challenging for data teams to manage risk. Legacy technologies compound this by leaving institutions with no tools to navigate the complexities of regulatory reporting and compliance.

Lack of agility

Disjointed tools hinder the ability to do real-time analytics and make smarter financial decisions. Traditionally rigid processes and technologies also prevent teams from efficiently sharing data and collaborating on cross-organizational initiatives.

Solution overview

But what if one could marry the concept of a data lake with the best of a data warehouse's functionality — to deliver data lake scale and data lake economics on a platform that enables AI and ensures the reliability of more traditional regulatory data pipelines? Enter Databricks Lakehouse for Financial Services, which gives organizations the ability to drastically simplify the ingestion and cleansing process, extracting insights from raw and complex signals and enabling both machine learning and non-machine learning use cases by combining advanced engineering pipelines and performant SQL capabilities, all within the same platform. With your data stored in an open format, on a low-cost object store in your own cloud, you are also provided with a cost-effective solution that guarantees that you will maintain full ownership and control over your data and foster collaboration between different data personas, quants and business analysts, end to end. Instead of accumulating data quality issues as data traverses far too many legacy ETL processes, this pattern helps organizations increase data quality as data flows through different data pipelines, from raw to curated to AI-augmented layers, in both batch and real time. Data becomes an asset for a business, versatile enough to fuel multiple use cases.



How it works

The reason that Databricks Lakehouse for Financial Services is able to bring the simplicity, flexibility and reusability required for you to move at the speed and agility that the market demands is that it is fundamentally a reimagining of the modern data architecture.

- Any type of data can be stored because, like a data lake, Databricks Lakehouse for Financial Services is built using the low-cost object storage supported by cloud providers. Leveraging this capability helps break down the data silos that hinder efforts to aggregate data for advanced analytics (e.g., personalized finance, fraud prevention), regulatory reporting or compute-intensive risk workloads. Probably even more important is the ability of a lakehouse architecture to travel back in time, ensuring full audit compliance and high governance standards.

- Streaming use cases such as monitoring transaction data are simpler to support because the lakehouse uses Apache Spark™ as the data processing engine and Delta Lake as a storage layer. With Spark, you can toggle between batch and streaming workloads with just a line of code. With Delta Lake, native support for ACID transactions means that you can deploy streaming workloads without the overhead of common reliability and performance issues.

- Machine learning and non-machine learning financial models can be deployed to and supported in production more easily because Databricks Lakehouse for Financial Services uses Delta Lake and MLflow. With Delta Lake, you can ensure your machine learning models remain stable by declaring your expectations for data quality upfront. With MLflow, you can train your model in any language and deploy it anywhere, without the extra overhead from integration work. This minimizes the need for a cumbersome "handoff" between data science practices, independent validation units and operation teams.

Value with Databricks for financial services

By using Databricks to build and support your lakehouse, you can equip your business with even more speed, agility and cost savings.

Collaborative workspace

Databricks provides financial analysts, scientists and engineers a common workspace, with purpose-built tools that promote collaboration over handoffs. For example, a data scientist can get started on an algorithm for climate risk while a data engineer extracts additional signals from unstructured files and risk manager teams provide the requisite expertise.

Integrated platform

Databricks is fully integrated with all the components of the lakehouse architecture. This integration helps free up resources so you can focus on solving business problems instead of hosting software — for example, using MLflow to provide full traceability and lineage from data, code and insights.

Reduced infrastructure costs

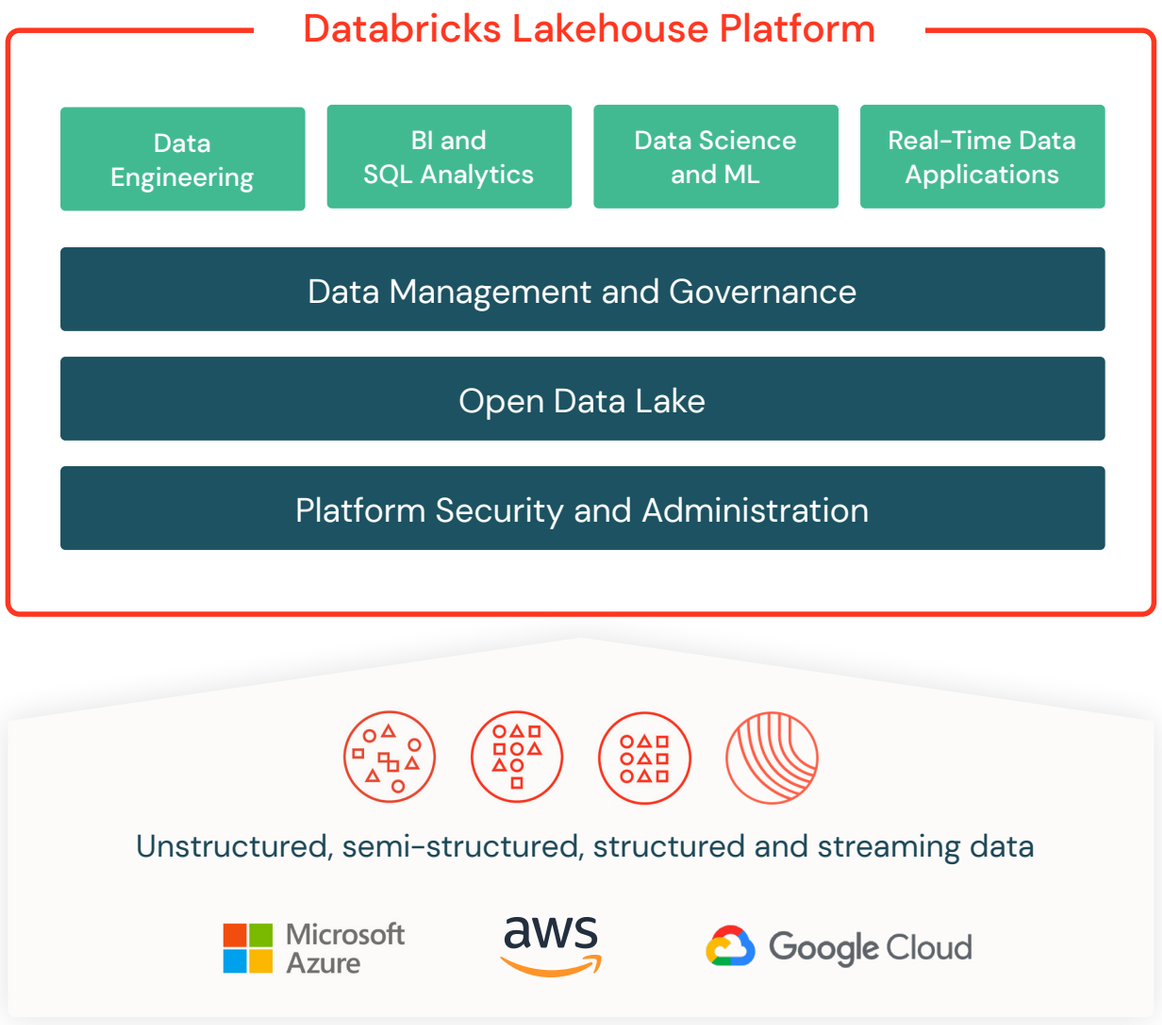
Databricks is deployed across your multiple cloud infrastructures and provides you with the tooling needed to manage your compute resources effectively to satisfy burst compute, such as regulatory reporting, and then scale down once appropriate.

Getting started

The flexibility of Databricks Lakehouse for Financial Services means that you can start with the use case that will be most impactful for your business. As you implement the pattern, you will find that you're able to tackle use cases quicker and easier than before.

Reference architecture

To the right is a common reference architecture for where Databricks fits within a financial services data estate. Databricks commonly handles both batch and streaming workloads as part of the ingestion process, and the data is then refined by a Bronze, Silver and Gold schema that makes data ready for downstream BI, AI or regulatory reporting use cases. Integrated with a vast ecosystem of partners (ETL providers, BI capabilities) and security providers to offer additional governance and row-level security, the lakehouse becomes the true delivery vehicle of business value, from data ingestion to business insights.



USE CASE: Risk Management (Market Risk/Credit Risk/Cybersecurity)

The core function of any FSI is to manage risk. Managing risk within financial services, especially within the banking sector, has increased in complexity over the past several years. The introduction of new frameworks, technologies and business models means the need for sound risk governance is at an all-time high. FSIs today need to build a real-time, scalable and agile modern risk management platform across business lines.

Challenges

Real-time data

Detecting and responding to risk requires the ability to build streaming pipelines for real-time and just-in-time analytics.

Simulations at scale

Risk management and associated calculations, such as FRTB, are highly compute-intensive.

Data quality

Poor data quality can materially impact risk assessment and makes reporting difficult.

Value with Databricks

- Delta Lake–based architecture enables scalable intraday risk calculations/aggregations while adding time travel capabilities
- You can quickly scale Monte Carlo simulations to the millions
- "Aggregations at read" enable "what-if" scenarios and ad hoc analysis

Solution overview

Databricks helps asset managers drastically simplify the E2E data ecosystem, ensuring both timeliness and reliability of market data, scaling Monte Carlo simulations to the millions and enabling clients with an intraday view of risks facing their businesses. With MLflow and the Delta Lake time travel capability, new risk models can be developed in full compliance with regulatory requirements and in complete transparency with internal bank policies.

How to get started

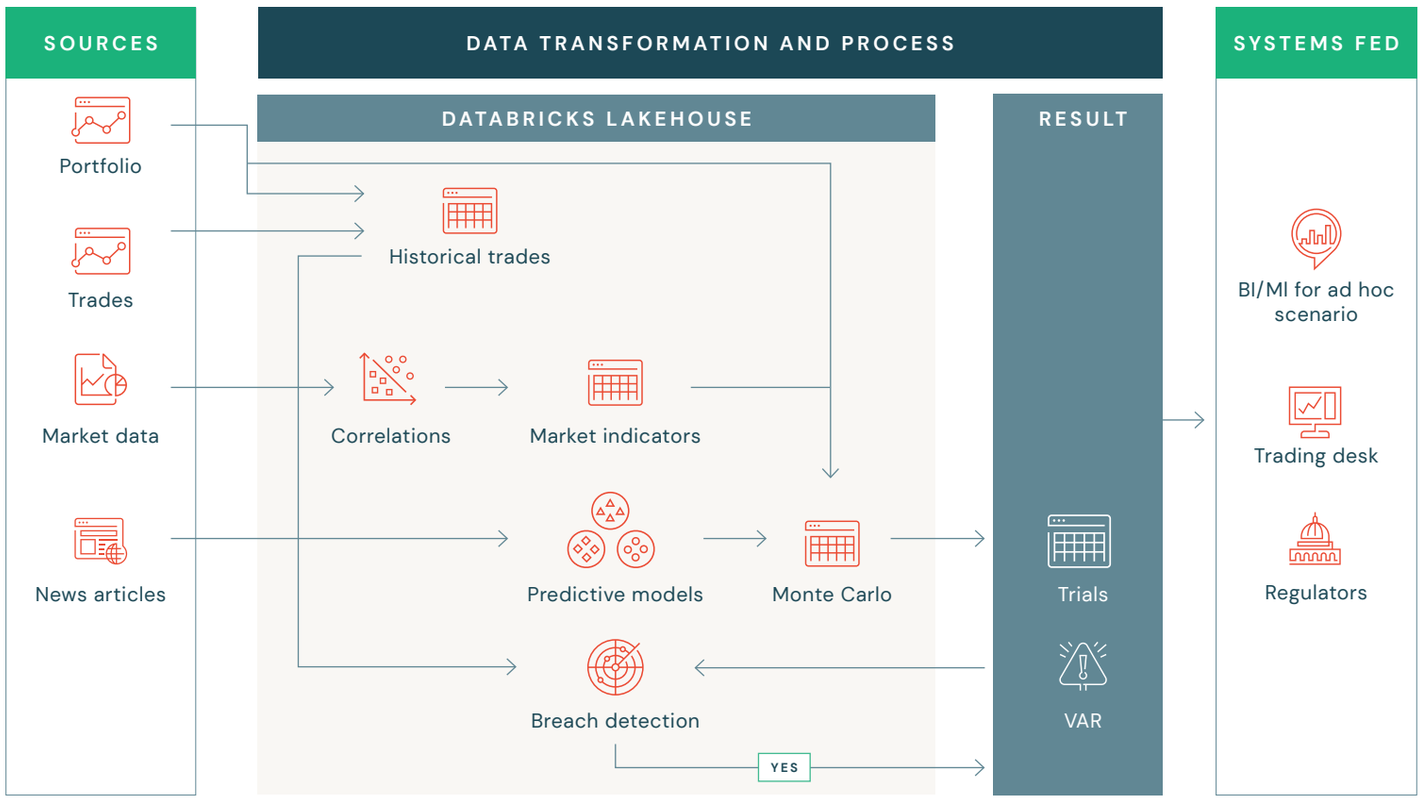
Solution Accelerator: **Modernizing Risk Management**

This modern risk management framework enables intraday views, aggregations on demand and the ability to future-proof and scale risk management. In this two-part blog series, we demonstrate how to modernize traditional value-at-risk calculation through the use of Delta Lake, Apache Spark and MLflow in order to enable a more agile and forward-looking approach to risk management.

Solution Accelerator: **Cybersecurity**

Detecting criminals and nation states through DNS analytics. In order to address common cybersecurity challenges such as deployment complexity, tech limitation and cost, security teams need a real-time data analytics platform that can handle cloud scale, analyze data wherever it is, natively support streaming and batch analytics, and have collaborative content development capabilities.

Reference architecture



Case studies

ABN·AMRO
CUSTOMER STORY >

reonomy
CUSTOMER STORY >

HSBC
CUSTOMER STORY >

databricks

USE CASE: Complaint Management

For financial institutions, reputation is arguably their most important asset. For example, Goldman Sachs' renowned business principles state, "Our assets are our people, capital and reputation. If any of these are ever diminished, the last is the most difficult to restore." In commercial banking, for example, brands that act on consumer complaints and feedback are able to manage the legal, commercial and reputation risks better than their competitors. American Banker magazine published this article which reiterates that non-financial risks, such as complaint management, are critical factors for FSIs to address in a rapidly changing landscape.

Challenges

Diverse data

Unstructured data (call detail records, complaints, NPS surveys) require advanced analytics capabilities (e.g., NLP) that are often disconnected from legacy technologies, tools or reports used by advocacy teams.

Anonymizing at scale

The need to anonymize PII data within customer complaints and other unstructured data is not trivial. It requires data scientists to operate outside their predefined boundaries — moving from R&D to being embedded within core operation teams — which is difficult to support with rigid legacy technologies.

Mining actionable insights

Data silos can contribute to not understanding the context of customer complaints beyond keywords, not linking to other customer 360 touch points (e.g., NPS) and not bringing ML capabilities to operation teams, such as rerouting complaints to the right agent.

Value with Databricks

- Superior CX, rerouting customer complaints to a dedicated agent
- Advocacy team to recommend changes to branch managers and maintain high NPS
- Cost reduction in customer service

Solution overview

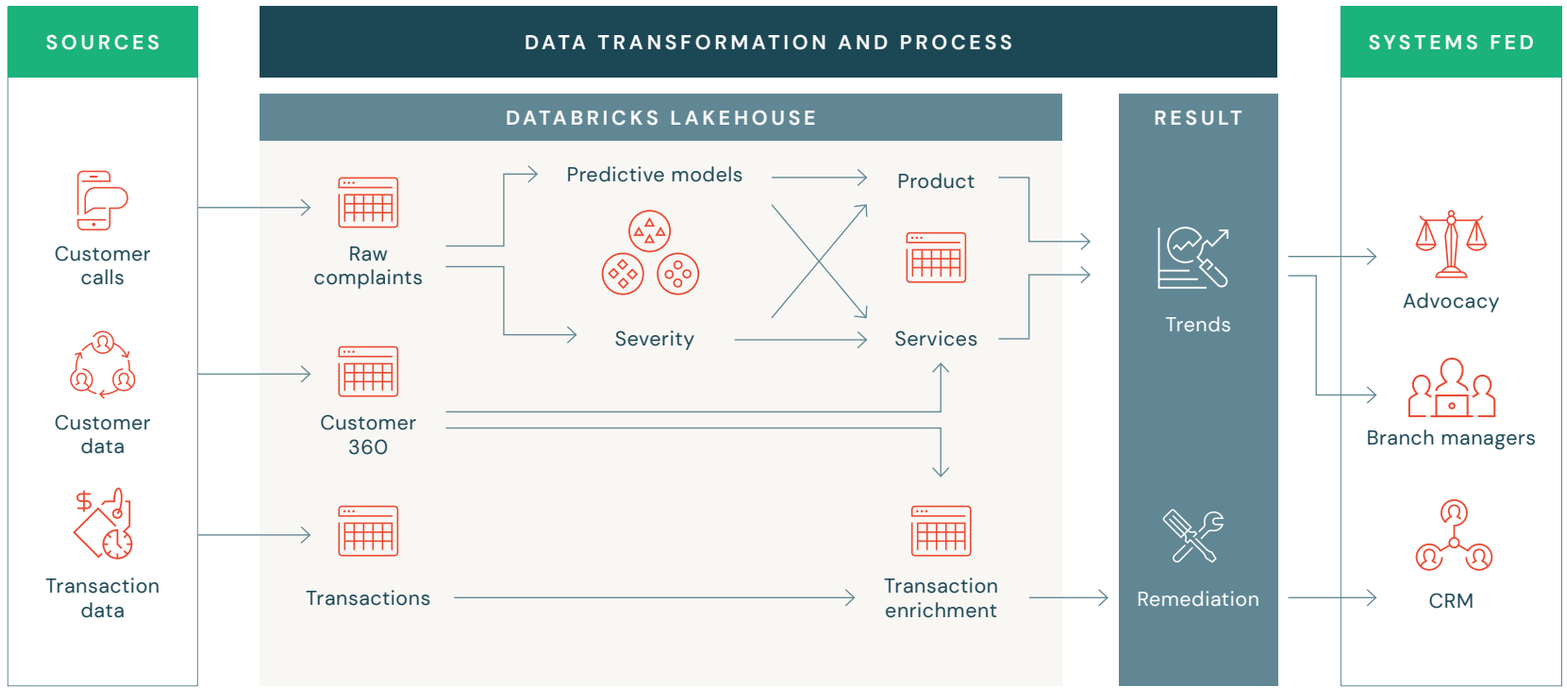
This approach to complaint management enables FSIs to measure brand perception and brings together multiple stakeholders to work collaboratively to drive higher levels of customer satisfaction and trust. This solution shows how to unify data science, operations and lines of business to reduce the friction between complex ETL engineering and anonymization, how to explore unstructured data using best-of-breed libraries and a familiar Python environment (Koalas) and how to act on these insights in a collaborative, iterative (Databricks SQL) and governed manner (MLflow).

How to get started

Solution Accelerator:
Reputation Risk

We demonstrate how to leverage the power of Databricks Lakehouse for Financial Services to solve those challenges, unlock insights and initiate remediation actions. We will look at Delta Lake, which is an open source storage layer that brings reliability and performance to data lakes and easily allows compliance around GDPR and CCPA regulations, whether data is structured or unstructured.

Reference architecture



databricks

USE CASE: ESG Scoring (Sustainable Finance)

The future of finance goes hand in hand with social responsibility, environmental stewardship and corporate ethics. In order to stay competitive, financial services institutions (FSIs) are increasingly disclosing more information about their environmental, social and governance (ESG) performance. By better understanding and quantifying the sustainability and societal impact of any investment in a company or business, FSIs can mitigate reputation risk and maintain trust with both their clients and their shareholders.

Challenges

Understanding ESG is hard

Today, the broad benefits of incorporating ESG goals are well understood by companies, investors and regulators. That's where the consensus ends. To put it another way, the "why" of ESG is clear, but the "how" and "what" are not.

Inconsistent and unreliable data

There is little agreement on what data companies should collect and how they should disclose and analyze it. Furthermore, the majority of data, policies and statements is highly subjective, making it difficult to quantify and measure.

Keeping companies accountable

ESG is meaningless without verification. This is true across all use cases from investing to supply-chain analytics. At the moment, ESG verification is essentially an honor code by which organizations pledge to follow the rules.

Solution overview

For ESG to really take off and make a global impact, data and AI must play a central role in collecting, verifying and analyzing ESG performance. This can be done effectively today only by leveraging technology. This solution shows organizations how they no longer have to rely solely on subjective ESG rating agencies. Instead, through data and AI, they can build an objective and 360 view of their ESG exposure and that of their competitors, investments and suppliers, and they can challenge the ESG status quo and act upon data-driven metrics. Most importantly, this solution shows how to bring complex data science and massive amounts of data for everyone in an organization to trust and use (with MLflow serving and Databricks SQL).

Typical use case data sources include company CSR reports, news analytics, market data, carbon emission records and supply chain management.

How to get started

Solution Accelerator:
ESG

Using Databricks Lakehouse for Financial Services, we demonstrate how Apache Spark, Delta Lake and MLflow can enable asset managers to assess the sustainability of their investments and empower their business with a holistic and data-driven view of their environmental, social and corporate governance strategies. Specifically, we will extract the key ESG initiatives as communicated in yearly PDF reports and compare these with the actual media coverage from news analytics data. We will then learn the connections between companies and understand the positive or negative ESG consequences these connections may have on your business.

Case studies

CREDIT SUISSE

CUSTOMER STORY >

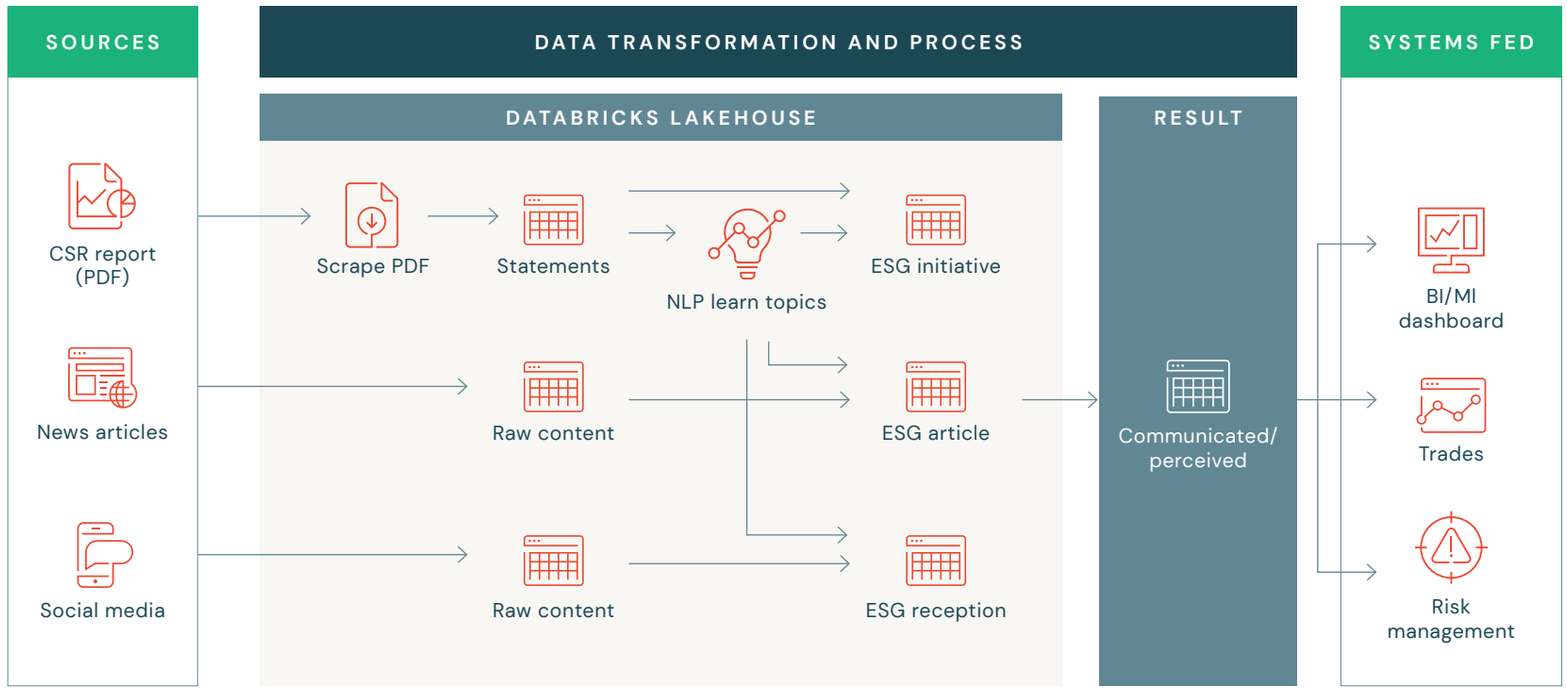
S&P Global

CUSTOMER STORY >

Scaling Data Marketplaces

databricks

Reference architecture



USE CASE: Hyper-Personalization

Accurate, granular and AI-enabled personalization impacts everything from underwriting performance and marketing to cross-selling, churn and customer service. In addition, personalization allows banks to get closer to their customers (by better understanding their lifecycle) and guide production direction. Personalization is ripe for advanced analytics and AI, such as graph analytics and clustering.

Challenges

No machine learning capabilities

Personalization today involves basic analytics on a narrow set of variables, such as age, gender or income.

Limited objective data

Narrow inputs often lead FSIs to view each cohort as a homogenized group, limiting everything from personalization to risk management.

Poor insights

The inability to augment basic customer segmentations with richer (and often unstructured) data for better insights into churn, upsell, etc.

Solution overview

Delta Lake enables organizations to combine various sources of data — for example, internal banking data such as transactions, customer demographics across products and services with digital journeys, and marketing analytics — in a timely and efficient manner, bringing a 360 view of customer interaction. With millions of customers, large organizations are able to leverage AI to extract similar customer patterns and create personalized experiences no longer based on who those customers are but how they bank, bringing a more inclusive approach to banking.

How to get started

Solution Accelerator:
Transaction Enrichment

In this solution, we show how the lakehouse architecture enables banks and open banking aggregators to address the challenge of merchant classification to enrich card transaction data with contextual information and gain further insights on transaction behaviors.

Solution Accelerator:
Customer Segmentation

Learning from the best-in-class industries (retailers and media companies) that drive the future of segmentation and personalization. Those best practices can be applied in the world of retail banking.

Case studies

Starbucks

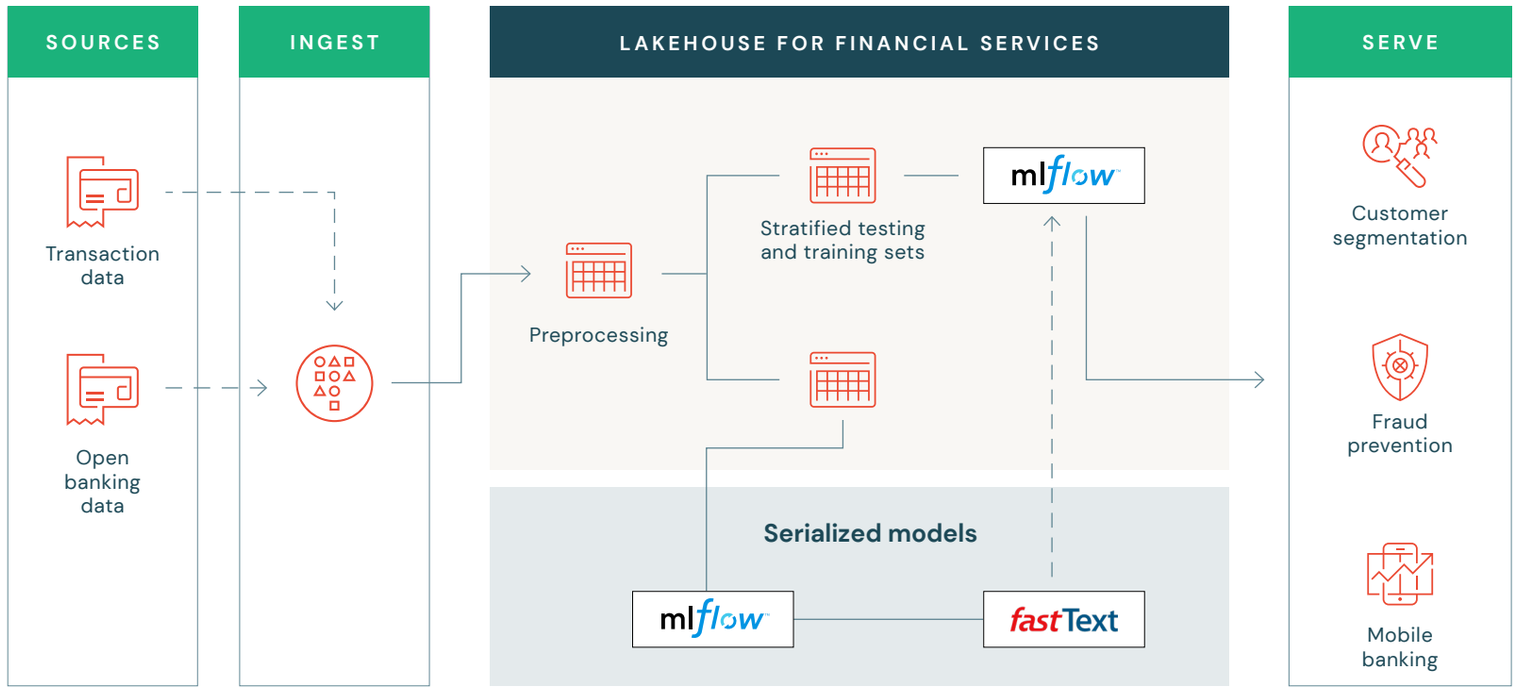
CUSTOMER STORY >

HSBC

CUSTOMER STORY >

databricks

Reference architecture



USE CASE: Compliance (AML/KYC)

Today, broadly speaking, compliance use cases at an FSI are often manual, rule-based and slow to adapt. Meanwhile, fraud gets more and more sophisticated and regulators demand more and more from the banks. AI can help automate the process of client profile verification to enhance due diligence in an organization. Client risk profiles can be verified in mere seconds against several databases and through the process of digital know your customer (KYC) powered by AI. This will help ensure that the companies meet standards for regulatory compliance.

Challenges

Stiff penalties

New regulations, such as PSD2, 5AMLD and GDPR, are raising the compliance bar in an effort to tackle financial crime.

Transparency

Although AI could be used to address many AML use cases, the lack of transparency in the development of ML models offers little explainability to regulators for wider adoption.

False positives

Any reported SAR must be investigated manually, offering no bandwidth to the development of AI solutions that could reduce false positives or contextualize isolated events into fraudulent activities.

Solution overview

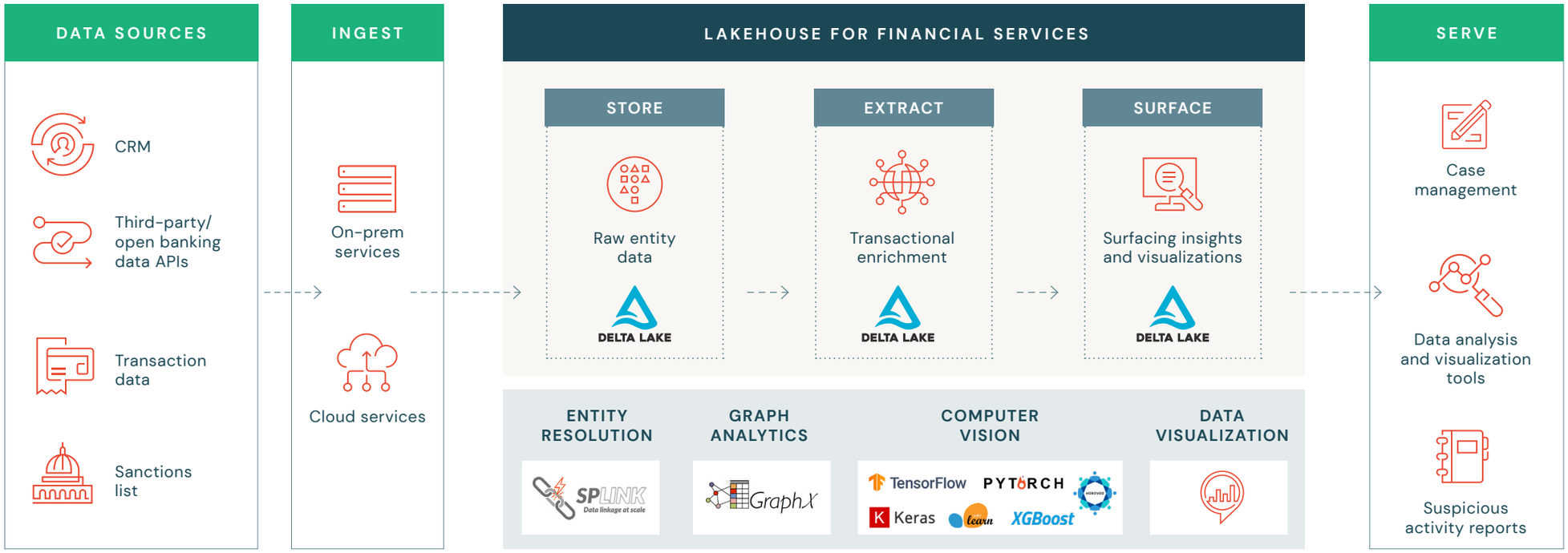
By being able to access all historical data in suspicious activity reports, natural language processing techniques can be used to extract key entities and relationships, helping fraud analysts generalize this approach to better identify common patterns behind malicious activities. Based on that knowledge and existing rule-based detection, data scientists can further identify anomalies from transaction data and move from fully unsupervised anomaly detection to a semisupervised system where each new case contributes to reducing false positives.

How to get started

Blog:
AML Solutions at Scale Using the Databricks Lakehouse Platform

Current AML operations bear little resemblance to those of the last decade. The shift to digital banking, with financial institutions processing billions of transactions daily, has resulted in the ever-increasing scope of money laundering, even with stricter transaction monitoring systems and robust KYC solutions. In this blog, we share a few design patterns from NLP, computer vision, entity resolution and graph analytics that deliver innovative, scalable solutions to adapt to the reality of modern online money laundering threats.

Reference architecture



Case studies

coins.ph
CUSTOMER STORY >

MONETA
CUSTOMER STORY >

zip
CUSTOMER STORY >

databricks

USE CASE: Regulatory Reporting

Regulators around the world expect more from the institutions they regulate. Specifically, timely and accurate reporting. Regulatory reports are only as reliable as the data they're based on. Quickly sourcing, aggregating and reconciling data is a core capability banks must possess in order to satisfy regulators and to respond to new and increasingly complex requirements.

Challenges

Costly and complex to scale

Compute-intensive activities running on-premises and requiring a large amount of available resources even if reports only run on a periodic basis.

Unreliable data

Complexities of data pipelines and system dependencies block visibility into overall data quality. Data consistency issues are often detected late at the time reports are generated.

Disjointed data

Data silos can hinder the ability to aggregate data with complex ETL to serve specific regulatory requirements.

Value with Databricks

- Redefine SLAs: reduce the compute time required
- Create on-demand clusters for compute-intensive queries (no need for on-premises 24x7)
- Simplification of ETL processes to maintain quality control (reduce ops costs)

Solution overview

Regulatory reporting is always tied to a hard deadline (and fines for noncompliance, depending on the regularity of the reporting, can involve a three-strikes-and-you're-out strategy). Processing has to proceed even if data is incomplete. A track record of poor data quality will also be fined, usually indirectly as an adjustment to the capital calculation because of "insufficient controls." Most times it is a battle between input quality and SLA, and the cost to the firm of making those adjustments can be huge. We demonstrate how Delta pipelines, as operating on streaming, would guarantee the transmission of regulatory data in real time to accommodate tight regulatory SLAs, and coupled with Delta back end and a FIRE data model, this framework equally ensures ACID transactions, query optimization (dramatic gain in regulatory computation) and time travel capabilities (auditability). This framework provides analysts with real-time confidence in regulatory data.

How to get started

Blog: **Timeliness and Reliability in the Transmission of Regulatory Reports**

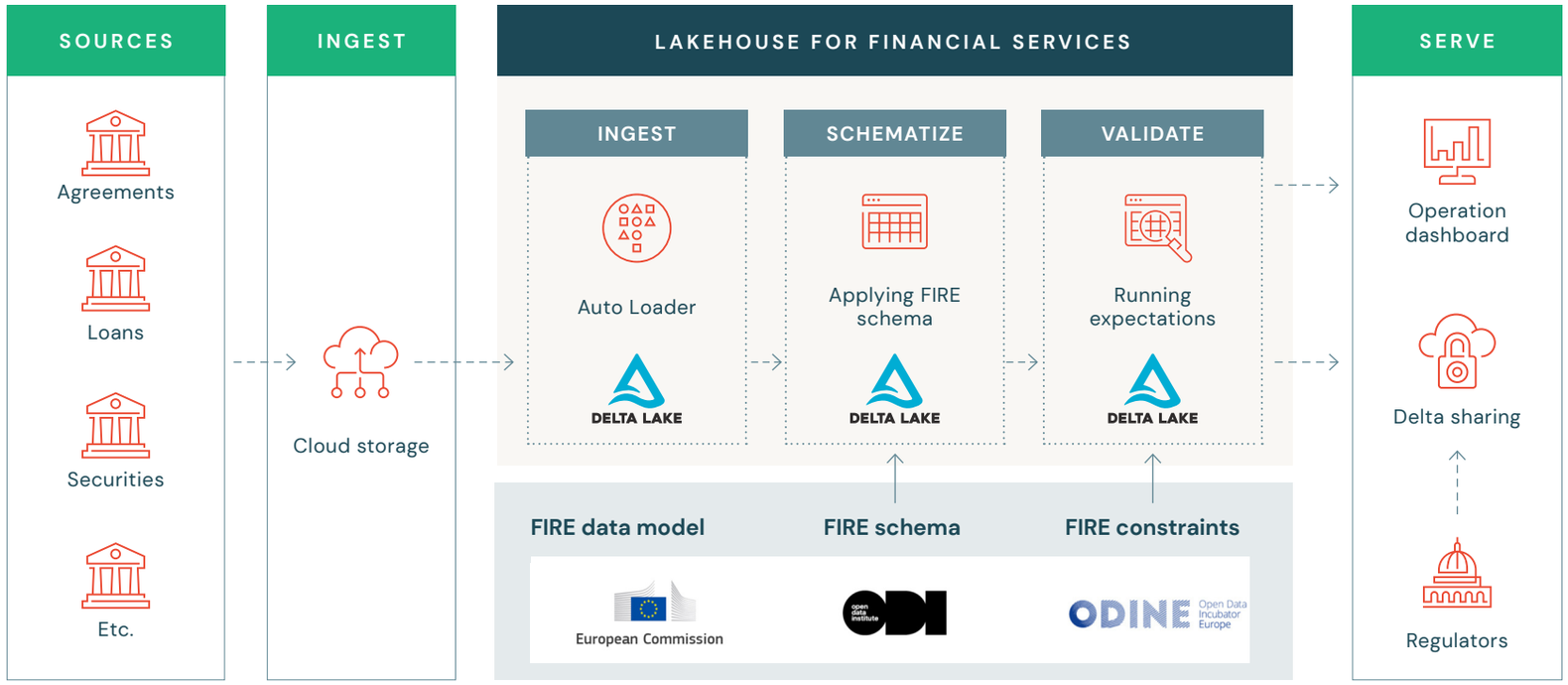
In this regulatory reporting blog, we demonstrate how Delta Live Tables can guarantee the acquisition and processing of regulatory data in real time to accommodate regulatory SLAs. With Delta Sharing and Delta Live Tables combined, analysts gain real-time confidence in the quality of the regulatory data being transmitted.

Case studies

FINRA

CUSTOMER STORY >

Reference architecture



databricks

USE CASE: Fraud Detection

According to McKinsey & Company, financial fraud (both detected and undetected) has become more prevalent and costly than ever. In a widely cited estimate, for every dollar of fraud, financial services institutions lose nearly three dollars once associated costs are added to the fraud loss itself.

Challenges

Lack of machine learning

A rules-based approach is not enough. Bad actors are getting more and more sophisticated with money laundering methods, necessitating an AI-driven approach.

Unreliable data

Getting high-quality, clean data and maintaining a rich feature store is critical for identifying ever-evolving fraud patterns while maintaining a strict record of previous data points.

Analytics at scale

Training complex ML models with hundreds of features on gigabytes of structured, semi-structured and unstructured data can be impossible without a highly scalable and distributed infrastructure.

Value with Databricks

- Rules-based to ML-based, reducing operation costs and maintenance
- Enables quick response to new fraudsters
- Reduces false positives, saving money and maintaining SLAs

Solution overview

The ability to develop complex models with high governance standards and the ability to bridge the gap between science and technology to address a challenge of scale where 40 billion transactions a year are made in the United States. Finally, the ability of Databricks to combine innovation (AI) and legacy (rule-based) within a common and efficient Spark-based orchestration engine.

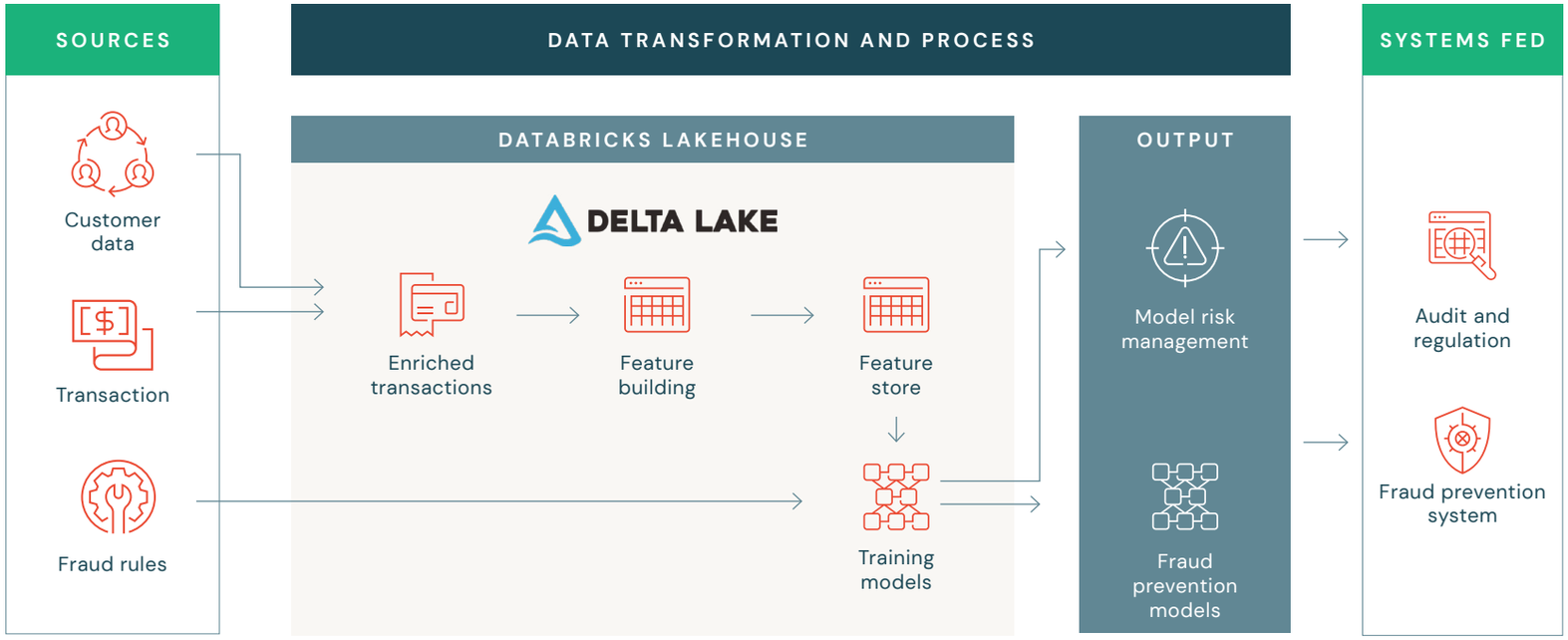
Typical use case data sources include transaction data, application data, customer demographics and geo-analytics data.

How to get started

Solution Accelerator:
Fraud Detection

Due to an ever-changing landscape and customer expectations, building a fraud detection framework often goes beyond just creating a highly accurate machine learning model. Oftentimes it involves a complex-decision science setup that combines a rules engine with a need for a robust and scalable machine learning platform.

Reference architecture



Case studies

MONETA

CUSTOMER STORY >

storebrand

CUSTOMER STORY >

- Financial Technology Company Catches More Fraud With Microsoft Azure AI
- Using Your Data to Stop Credit Card Fraud: Capital One and Other Best Practices
- Credit Card Fraud Detection Using ML in Databricks

databricks

USE CASE: Investment Analytics

Investment analytics enables asset managers to generate alpha, enhance market surveillance and perform complex transaction cost analysis (TCA). Investment analytics involves massive financial time series data (e.g., stock tick data) as well as vast amounts of unstructured/alternative data sets that require an analytical platform that is real-time, scalable and fast.

Challenges

Massive data volumes

The need to ETL billions of financial time series data, such as stock tick data, can be compute-intensive and beyond the scalability of legacy infrastructure.

Gaining a holistic view of data

The inability to unify streaming data as well as years of historical data needed for aggregations and compliance can impact an FSI's ability to provide accurate and informed investment recommendations.

Costly and complex to scale

Analyzing and aggregating tens of thousands of time series in parallel can prove difficult and costly to run on-premises on dedicated servers. Compute-intensive activity, such as a model backtesting, often happens outside of trading hours.

Value with Databricks

- 3x–10x faster than open source Spark on aggregations
- Ingest market data faster — from weeks to hours
- Advanced TCA calculations in parallel

Solution overview

One of the biggest technical challenges underlying problems in financial services is manipulating time series at scale. Another major challenge is centralizing the wide variety of time series data sources, effectively unlocking potential value. Tick data, alternative data sets such as geospatial or transactional data, and fundamental economic data are examples of the rich data sources available to financial institutions, all of which are naturally indexed by timestamp. Solving business problems in finance such as risk, fraud and compliance ultimately rests on being able to aggregate and analyze thousands of time series in parallel.

How to get started

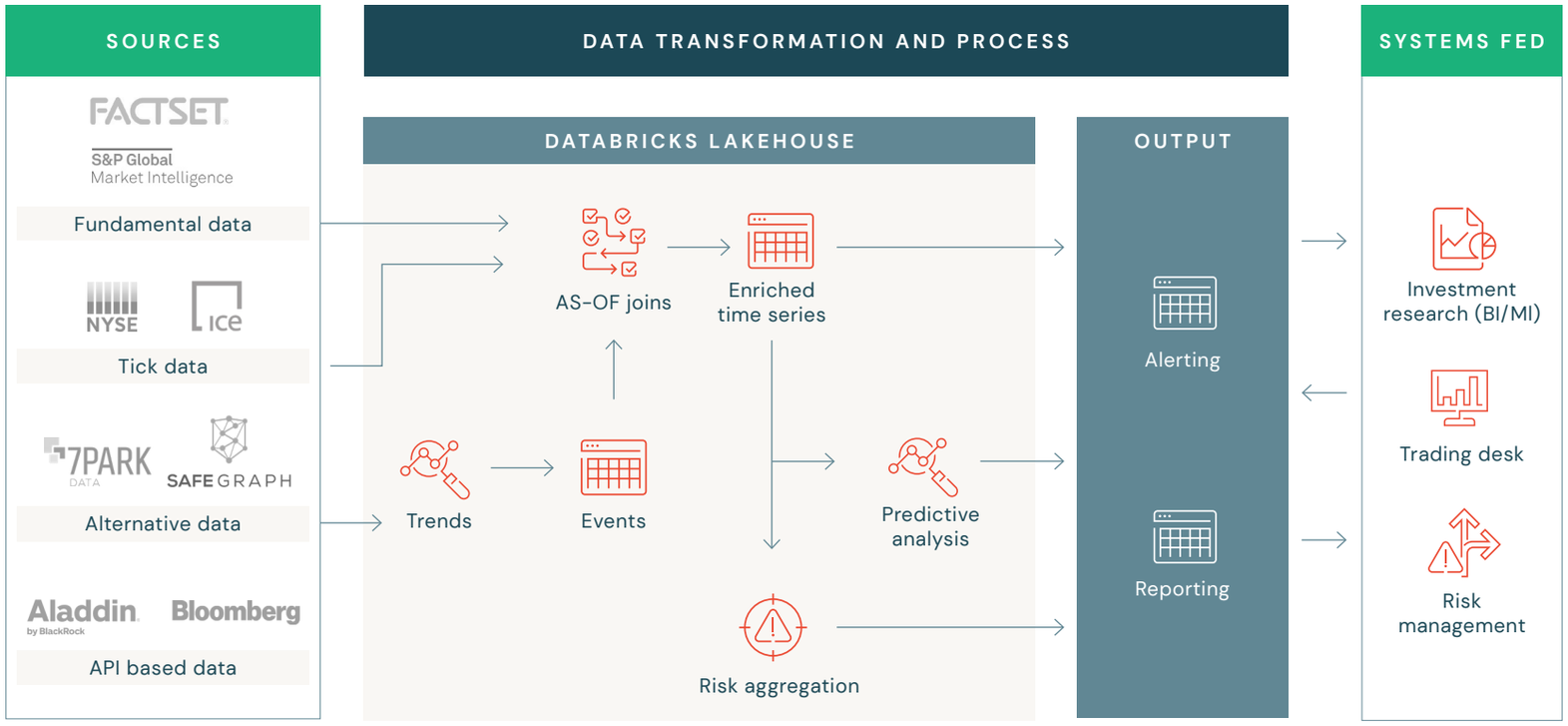
Solution Accelerator:
How to Build: A Modern Investment Data Platform

Leverage common data sets in Delta Lake and productionize use cases, such as financial product volatility forecasting, market surveillance or transaction cost analysis (TCA). Delta Lake enables financial services institutions to focus on product delivery for customers using a modern data architecture.

API for Time Series Manipulation With Spark

The purpose of this project is to make time series manipulation with Spark simpler. Operations covered under this package include AS OF joins, rolling statistics with user-specified window lengths, featurization of time series using lagged values and Delta Lake optimization on time and partition fields.

Reference architecture



Case studies

- Supercharging Your Data Lake With Alternative Data
- Alternative Data Analytics Using SQL
- Using Data + AI to Invest and Mitigate Risk
- Machine Readable Filings for Investment Insights

databricks

Conclusion

Today, data is at the core of every innovation in the financial services industry. Databricks Lakehouse for Financial Services enables financial institutions across banking, insurance, capital markets and wealth management to harness the power of data and analytics to solve strategic challenges and make smarter decisions that minimize risk, deliver superior customer experiences and accelerate innovation.

Get started with a free trial of Databricks Lakehouse for Financial Services and start building better data applications today.

START YOUR FREE TRIAL

Contact us for a personalized demo
databricks.com/contact

databricks

Databricks is the data and AI company. More than 5,000 organizations worldwide — including Comcast, Condé Nast, H&M and over 40% of the Fortune 500 — rely on the Databricks Lakehouse Platform to unify their data, analytics and AI. Databricks is headquartered in San Francisco, with offices around the globe. Founded by the original creators of Apache Spark,™ Delta Lake and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems. To learn more, follow Databricks on Twitter, LinkedIn and Facebook.