eBook

A Primer on Using Data + AI for Fraud Prevention

databricks

Abstract

A Primer on Using Data + AI for Fraud Prevention

Fraud is a costly and growing problem. For every $1 of fraud, companies pay 3.36x in chargebacks and replacement and operational costs. Identity fraud losses soared to $56 billion in 2020. Fraudsters continue to increase the scale, speed and sophistication of attacks — from 2019 to 2020, fraud grew at an average rate of 33% — threatening the revenue and growth of companies. This eBook explains how data and machine learning are ideal applications for fraud prevention — especially in the face of modern threats — and how to future-proof your organization using the Databricks Lakehouse Platform. It also explains how financial services leaders are using data and AI to combat fraud and includes two Databricks Solution Accelerators with easy-to-use and best-practice notebooks so you can get a jump start on fraud detection.

Contents

The growing cost of fraud

1

For every $1 of fraud, companies pay 3.36x in chargebacks and replacement and operational costs

Fraud continues to grow at 33% year over year, threatening the revenue and growth of companies

Fraud is a costly and growing problem. For every $1 of fraud, companies pay 3.36x in chargebacks and replacement and operational costs.[1] Identity fraud losses soared to $56 billion in 2020, and with more businesses and end users keeping their sensitive information online, bad actors are devising new tactics to defraud them both. In the past few years, we've seen huge data breaches that have flooded the dark web with sensitive personal and financial information that can be used for account takeover and identity theft. The unfortunate truth is: Fraud will affect your industry.

- $300 billion, or up to 10% of total healthcare expenditure, could be attributed to healthcare fraud in the U.S.[2]
- $1.45 trillion lost turnover as a result of financial crimes, representing 3.5% of global turnover[3]
- $150 billion lost by government agencies to a wide range of fraud, representing 10% of all fraud[4]

Fraud continues to grow at 33% year over year, threatening the revenue and growth of companies. It impacts customer satisfaction, loyalty and the bottom line.[5]

Moreover, fraudsters' methods are becoming more technologically advanced. The Association of Certified Fraud Examiners (ACFE) and PwC have shared how organized crime and large-scale fraudsters are increasing the scale, speed and sophistication of fraud attacks.[6] One of the newest methods involves using machine learning (ML) and other automation techniques to commit fraud that legacy approaches can't detect, including phishing emails, identity theft and forgery, phone and location spoofing, and the emulation of user behavior. Experts and industry leaders are now looking to machine learning and AI to understand and get ahead of fraud. Fraudsters are using AI. Why aren't you?

[1] CNP Fraud Costs U.S. Merchants $3.36 for Every $1 of Direct Fraud Loss
[2] U.S. sentences 14 to combined 74 years in prison for healthcare fraud
[3] Refinitiv Survey Report: Revealing the true cost of financial crime
[4] McKinsey on Government Perspectives: Adopting AI, automation, and advanced analytics in governments
[5] Fraud rate rises 33% during COVID-19 lockdown
[6] ACFE and ABFA Fraud Resources

Why machine learning and AI are ideal applications for fraud prevention

2

Fraudsters constantly adapt their tactics, making them difficult to detect by humans

Fraud detection is the task of predicting unauthorized access to or use of accounts, credit cards and other sensitive information. One salient method of recognizing fraudulent activity is to leverage machine learning models to dynamically detect fraudulent transactions. ML models can be trained on a set of data — for example, account history, credit card transactions, personal information and more.

With the availability of data and with advances in machine learning, fraud prevention is a key area in which machine learning is changing both workflows and outcomes, allowing organizations to stay ahead of criminals who are only growing more technologically advanced. Today's businesses are facing an increasingly sophisticated enemy that attacks, responds and changes tactics extremely quickly. With data analytics and machine learning, companies can get ahead of threats. Below, we discuss the main reasons why machine learning is especially well suited for taking on fraud.

Fraud hides under massive amounts of data

The most effective way to detect fraud is to examine the overall behaviors of end users. Looking at transactions or orders is not enough — we need to follow the events leading up to and after the transaction. This culminates in a lot of structured and unstructured data, and the best way to detect fraud in such huge volumes of data is with machine learning and AI.

Fraud happens quickly

When a machine learning system gets updated in real time, this knowledge can be used within milliseconds to update machine learning models and prevent an attack.

Fraud is always changing

Fraudsters constantly adapt their tactics, making them difficult to detect by humans — and impossible to detect by static rules-based systems, which don't learn. Machine learning, however, can adapt to changing behavior.

Coupled with human talent and experience, data and AI work together to constantly learn and adjust to new user behaviors and trends

Fraud is unnoticeable on the surface

To the naked eye, fraudsters and good users don't appear any different from each other. Machine learning has a deeper and more nuanced way of viewing data, which helps avoid false positives.

Machine learning uses statistical models to look at past outcomes and anomalies to predict future outcomes. A machine learning system can learn, predict and make decisions as data arrives in real time.

The benefits of data and AI for fraud prevention

- **Highly accurate results:** Advanced and data-driven behavioral analysis means fewer false positives
- **Less need for manual review:** Machine learning automates processes in which behaviors can be learned at the individual level and anomalies can be detected
- **Ability to prevent fraud without impeding the user experience:** AI brings automation to the process seamlessly and prevents fraud in advance without burdening users
- **Lower operational costs than other approaches:** With less manual work and more automation, data and AI require fewer resources and preempt losses associated with fraud
- **Frees up time for teams to focus on more strategic tasks:** Most companies are not in the business of fraud detection, and a machine learning fraud prevention process can help them focus on core activities
- **Adapts quickly:** Coupled with human talent and experience, data and AI work together to constantly learn and adjust to new user behaviors and trends

When it comes to operationalizing data and AI to build customer relationships and drive higher returns on equity, fraud should be considered a top priority. Curbing fraudulent or malicious behavior — from fraudulent securities trading to money laundering — is key to mitigating negative revenue impact.

Data and ML challenges in fraud prevention

3

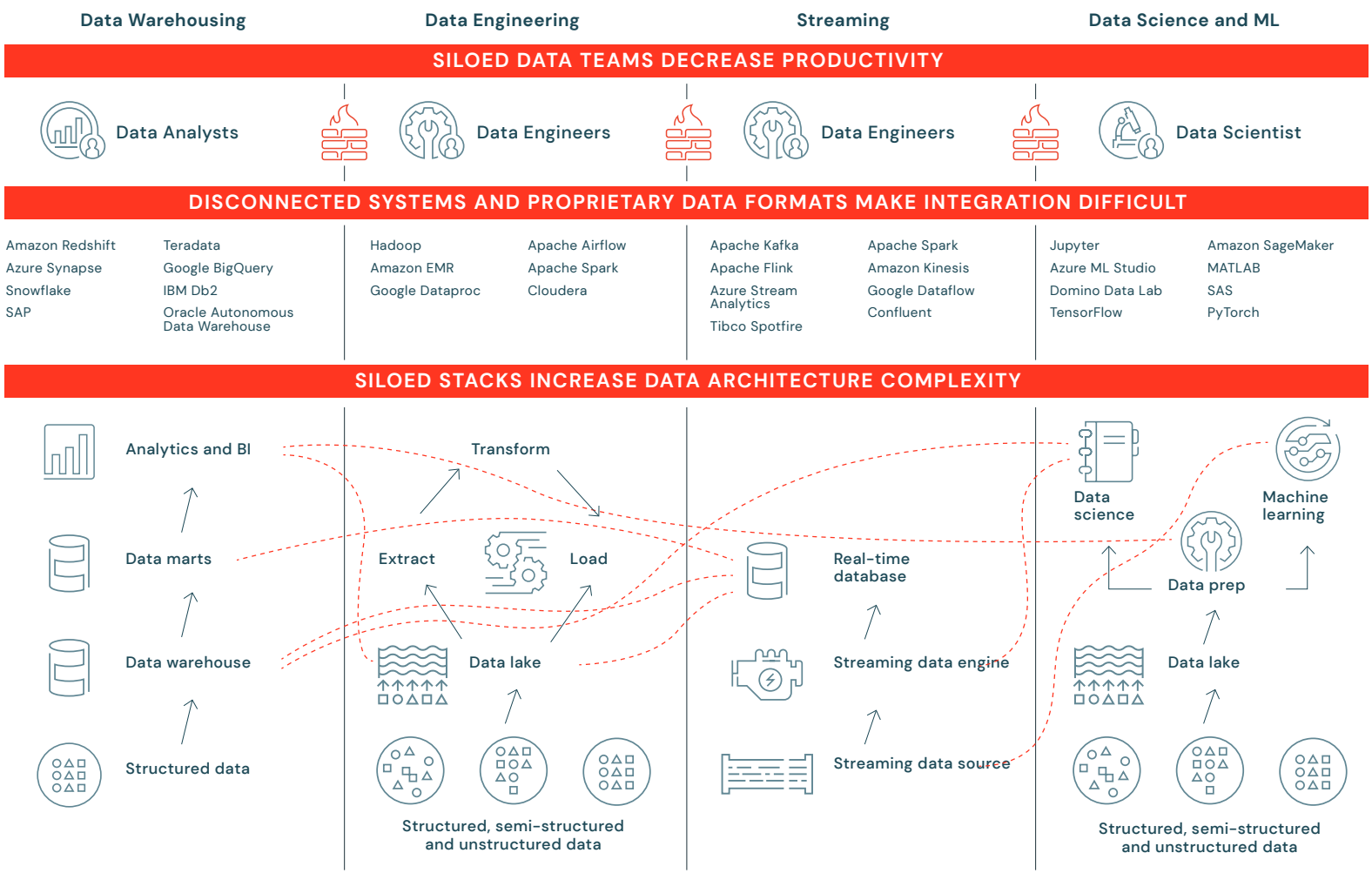
Companies that want to build their own ML infrastructure need to think about supporting massive data growth

We've established that data analytics is the best way to prevent fraud. But there are challenges when it comes to execution, especially regarding the data. As the amount of data increases, so does the complexity of understanding it all. Where will all this data be stored, and in what format? How will it be retrieved and used again? How will the data be prepared for BI and machine learning? These issues are easier to solve when data is limited vs. when it scales by 100x or more. Companies that want to build their own machine learning infrastructure need to think from the beginning about supporting massive data growth — and, along with it — an increasingly complex surface area of problems to tackle. Here are a few objectives to keep in mind when trying to leverage data and AI for fraud prevention:

- Getting high-quality, clean data and maintaining a rich feature store for ever-evolving fraud patterns
- Using multiple vendors that are siloed in capabilities — for example, ETL, data science and scaling on demand
- Enabling multiple data teams to scale their data pipelines and collaborate in the cloud
- Meeting ML challenges by training complex models with hundreds of features on gigabytes of data

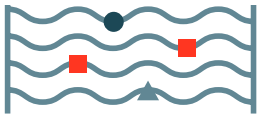
With its high operating costs and lack of agility, a disjointed operating model cannot successfully combat fraud. In order to succeed in the journey of digital transformation, financial services institutions (FSIs) must adapt to today's agile and data-driven times while modernizing their approach to data and analytics.

Today, the fraud ecosystem is disjointed



How a lakehouse approach to data and AI simplifies fraud prevention at scale

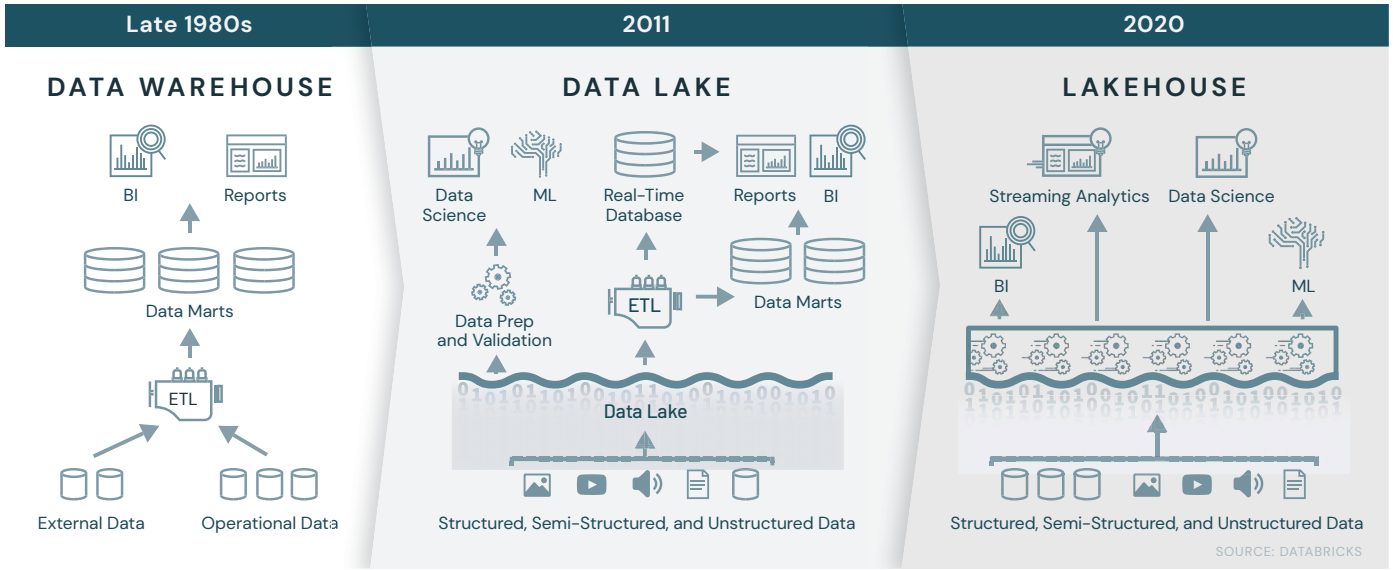
4



A lakehouse architecture is well suited for fraud prevention because it provides a fully managed cloud platform that unifies data engineering, data analysis and data science with the rest of the business

The Databricks Platform leverages the openness and simplicity of the lakehouse — a new, open architecture that combines the best elements of data lakes and data warehouses. The lakehouse system design has similar data structures and data management features to those in a data warehouse, but directly leverages the kind of low-cost storage used for data lakes.

A lakehouse architecture is well suited for fraud prevention because it provides financial services companies with a fully managed cloud platform that accelerates innovation by unifying data engineering, data analysis and data science with the rest of the business. And a lakehouse environment can easily extract and manage the massive, real-time, changing and differing data types that are needed to detect fraud.

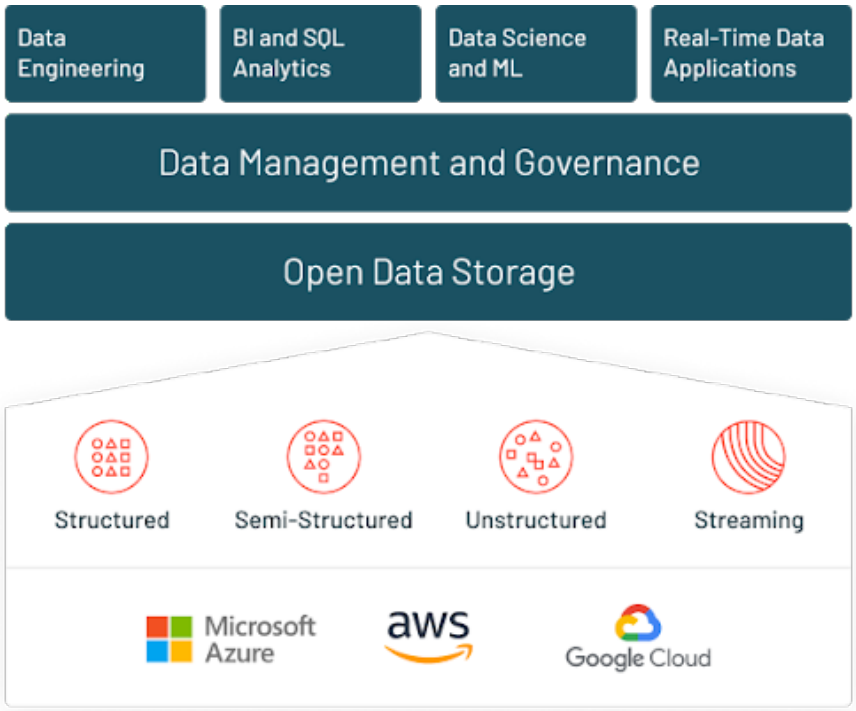


The Databricks Lakehouse Platform combines the best elements of data lakes and data warehouses — delivering the data management and performance typically found in data warehouses and the low-cost, flexible object stores offered by data lakes

Unifying batch and streaming pipelines is key to ML success

One of the most common problems FSIs face is dealing with massive volumes of data — both batch and streaming — across disparate sources. Many FSIs turn to data lakes to aggregate their big data cost-effectively, but this creates a new set of challenges around data management and governance.

Databricks enables organizations to overcome these challenges with a lakehouse architecture powered by Delta Lake. An open source technology, Delta Lake is natively integrated within Databricks to provide reliability and performance. It acts as a storage layer on top of your data lake that enforces data quality with ACID transactions. FSIs can ingest structured, semi-structured and unstructured data, both in batch and streaming, into a single Delta Lake to ensure that the supply of data is clean and usable. The scalability of Databricks enables organizations to then process and query this data for near real-time insights.



How cross-team collaboration boosts AI innovation

Fraud is not solely addressed by AI — it requires domain expertise (e.g., rules) combined with AI-generated insights — so business analysts have to visually inspect patterns and rules efficacy through MI/BI dashboards. To combat fraud, business analysts and data scientists must have the same set of data, which is possible in a lakehouse. Because Databricks' approach to fraud detection often involves a combination of rules and ML, it's well suited to meet the needs of a diverse set of personas required to create rules-based ML models.

The ability for users to collaborate across multiple workspaces while providing isolation at the user level is critical in financial services. Databricks' fraud solutions address the key areas of scalability in the cloud, fraud prevention workflow management and production-grade open source ML frameworks. They allow organizations to build and maintain a modern fraud and financial crimes management infrastructure by increasing alignment between different internal teams.

The Databricks Lakehouse Platform unifies data teams so they can collaborate across the entire data and AI workflow

The power of business insights for fraud analytics

Databricks centralizes your data for easy access. You and your team of data analysts can easily and directly connect and query your most complete and recent data in the data lake — using Delta Lake and Databricks SQL to discover anomalous behaviours and perform root cause analysis. Connectors with popular BI tools like Tableau and Power BI allow your analysts to use their preferred BI visualization and reporting tools for real-time insights that can help stop a malicious attack.

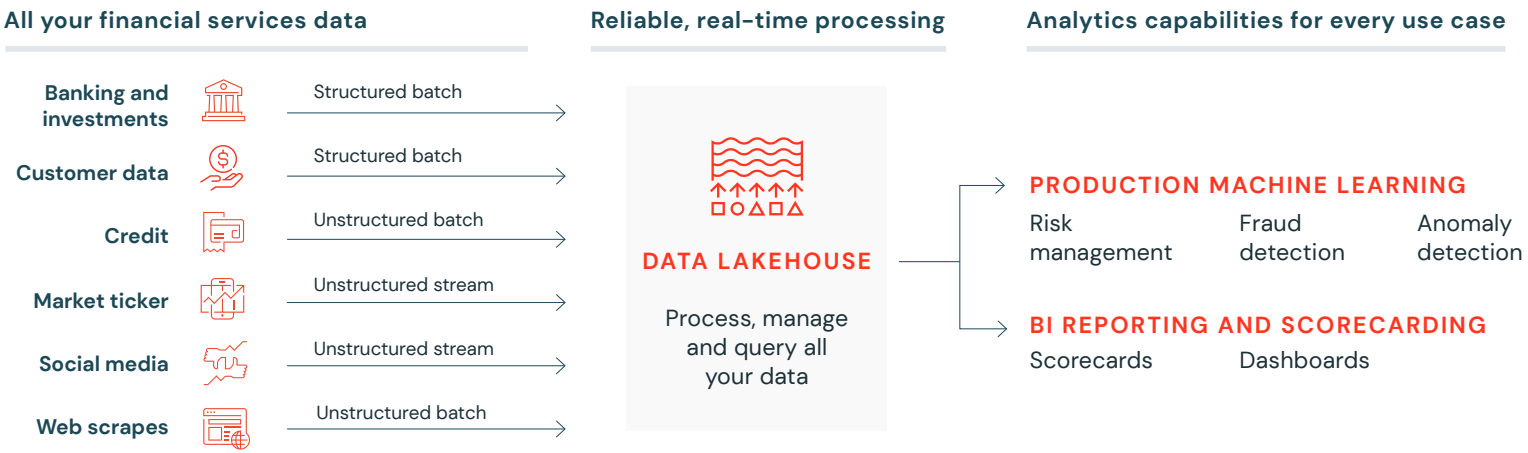
Unlock the potential of your data on a single platform

Financial services institutions can minimize losses and improve customers' trust with fewer false declines. Fraud has diversified rapidly and requires an agile approach. There's no one type of fraud, and fraudsters are constantly changing techniques. We need a platform that allows real-time access to new data and faster delivery to insights and new rules. The Databricks Lakehouse is a high-performance query engine designed for agility and performance on big data workloads and machine learning.

Additionally, any AI-based model must comply with strict regulatory requirements, necessitating as-of code, data and models. By combining MLflow and Delta Lake capabilities, models and rules can be created with a high degree of governance and trust in your data. An independent team of experts can ensure the right parameters are used against the right set of data, resulting in the right outcome.

Unlock the value of data lakes for BI and ML

Databricks provides a Lakehouse Platform that helps financial services institutions democratize data for downstream fraud analytics and AI — minimizing risk while accelerating transformative innovation.



DATA CHALLENGE	THE DATABRICKS LAKEHOUSE FOR FINANCIAL FRAUD PREVENTION
DATA INGEST: Processing batch and streaming data can be slow and error-prone, impacting downstream analytics	Connect traditional data with alternative data insights
DATA LAKE MANAGEMENT: Data silos can limit the ability to gain a complete view of the customer	Easily handle large volumes of data from multiple sources (transactions, geospatial, demographics, etc.) built on a strong privacy foundation
DATA QUERY: Fragmented, siloed and inconsistent data sources for BI and data science	Ability to rapidly and inexpensively experiment, manage and push out at scale from a single platform

Databricks Fraud Prevention Solution Accelerators

5

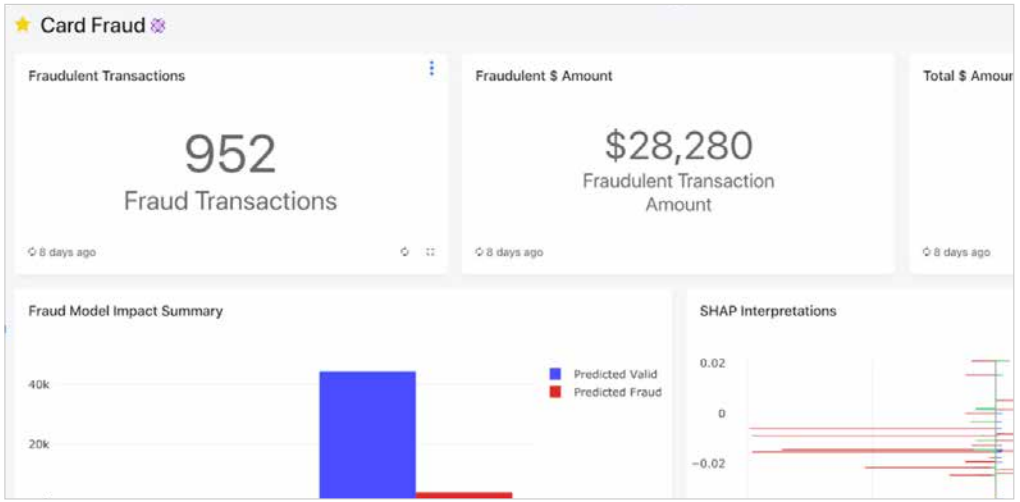
Based on best practices from our work with leading brands, we've developed Solution Accelerators for fraud data analytics and machine learning use cases to save weeks or months of development time for your data engineers and data scientists.

Both accelerators demonstrate the innovative use of anomaly detection with geospatial analysis in fraud prevention, and show how easy it is to get started.

How to build a rules-based AI model to combat financial fraud

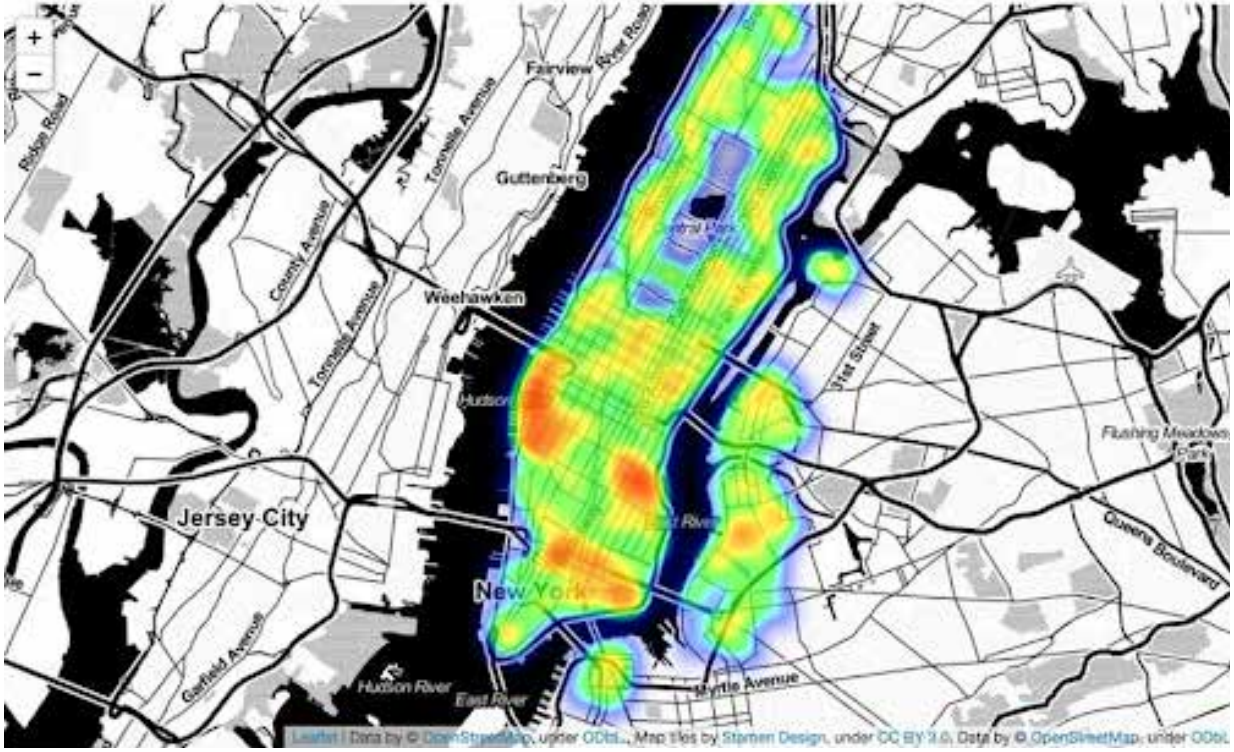
Preempt fraud with rule-based patterns and select ML algorithms for reliable fraud detection. Use anomaly detection and fraud prediction to respond to bad actors rapidly.

- Build trust and loyalty among your customers
- Respond to fraudulent activities fast and reduce operational costs
- Eliminate data complexity and implement fraud prevention at scale



How to identify credit card fraud through AI and geospatial analysis

Learn how geospatial analysis and a lakehouse framework enable organizations to better understand customer spending behaviors and detect abnormal card transaction patterns using machine learning in real time. Geospatial data can enhance fraud prevention to avert losses and build customer trust.



Geospatial clustering helps identify customer spending behaviors

- Leverage geospatial data and rules-based AI to combat financial fraud
- Identify abnormal behavior and detect threats to better protect users
- Personalize the banking experience with granular, customer-centric insights

Solution Accelerators are an easy and simple way to approach fraud.
Visit our Solutions Accelerator page to get started.

Customer Story:

FINRA®

6

"With Databricks, we have one cohesive end-to-end process with one single unified team working on protecting the securities markets."

— Saman Michael Far
Senior Vice President
of Technology at FINRA

The Financial Industry Regulatory Authority (FINRA) is an independent, nongovernmental organization responsible for protecting investors by ensuring that the U.S. securities markets operate in a fair and honest manner. FINRA oversees 12 markets and exchanges, 3,700 firms and more than 600,000 brokers. FINRA deters misconduct by enforcing rules, detecting and preventing wrongdoing in the U.S. markets, and disciplining members who break the rules.

Challenge

- Disjointed systems, workflows and teams caused FINRA to struggle with leveraging big data analytics to detect fraud and protect investors

Use case

- Data ingest and ETL
- Leverage machine learning to detect fraudulent securities trading

Why Databricks?

- Fosters collaboration among data scientists, engineers and analysts to boost innovation
- Provides a fully managed platform that removes infrastructure complexity so they can focus on the data rather than DevOps

With Databricks, teams can quickly iterate on ML models and scale detection efforts to hundreds of billions of market events per day. As a result, FINRA has significantly improved fraud prevention, leading to a more secure financial future for investors in the U.S.

Impact

- Analyze 100 billion stock market events per day to identify fraud and wrongdoing

Customer Story:
coins.ph

7

"With the Databricks unified data analytics platform, our operational costs have been reduced by about 70% due to more efficient cluster management."

— Dmitry Ustimov, Data Architect at Coins.ph

As a key digital payments platform in the Philippines, Coins.ph needs to be able to perform accurate, insightful financial audits and prevent fraud — in real time. With more than 10 million customers accessing digital payment services for local and international remittances, bill payments and online shopping, Coins.ph needed to find a way to move beyond development operational processes and address more advanced business challenges. With Databricks, Coins.ph was able to harness richer data insights to deliver ML-powered fraud detection and anti-money-laundering solutions at greater speed while optimizing financial reconciliation.

Challenge

- Legacy analytics system built on EMR struggled to turn massive data into meaningful insights, requiring a significant effort from their data team to maintain and use the platform for development operational tasks

Use case

- Data ingest, ETL and machine learning
- Develop ML experiments and prototypes to address and deploy fraud detection
- Use the Databricks Platform to enable standardized analytics and create new rule sets for anti-money-laundering compliance

Why Databricks?

- Provides a unified platform for data teams to collaborate on data preparation and analytics and to prototype new models
- Delta Lake ensures consistent data pipelines that feed data downstream for ML
- MLflow provides easy development and tracking of new ML models

Using Delta Lake to ingest large volumes of data in real time, data engineers at Coins.ph are able to bring greater reliability to the data lakes and get up-to-date insights to develop more robust and scalable data pipelines for

70%

operational cost reduction in compute costs

fraud detection. MLflow simplifies and streamlines the ML lifecycle, allowing data teams to easily track ML experiments and quickly develop new prototypes to address fraud detection.

Impact

- 14x reduction in complaints received
- 70% operational cost reduction in compute costs
- 50% infrastructure cost reduction

Conclusion

8

When financial services institutions are able to predict fraudulent activity, they can greatly mitigate revenue loss and customer churn. While machine learning has come very far and is impacting our lives in countless ways, it's still just getting started. Soon, the question of whether to leverage machine learning for fraud won't be a question but a necessity. In the meantime, the lakehouse is an imperative first step to take in fraud prevention. Learn more about Databricks' capabilities and the ways in which the lakehouse can take on fraud.

Stay up to date on the latest events, case studies and solutions for financial services at dbricks.co/fiserv.

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.