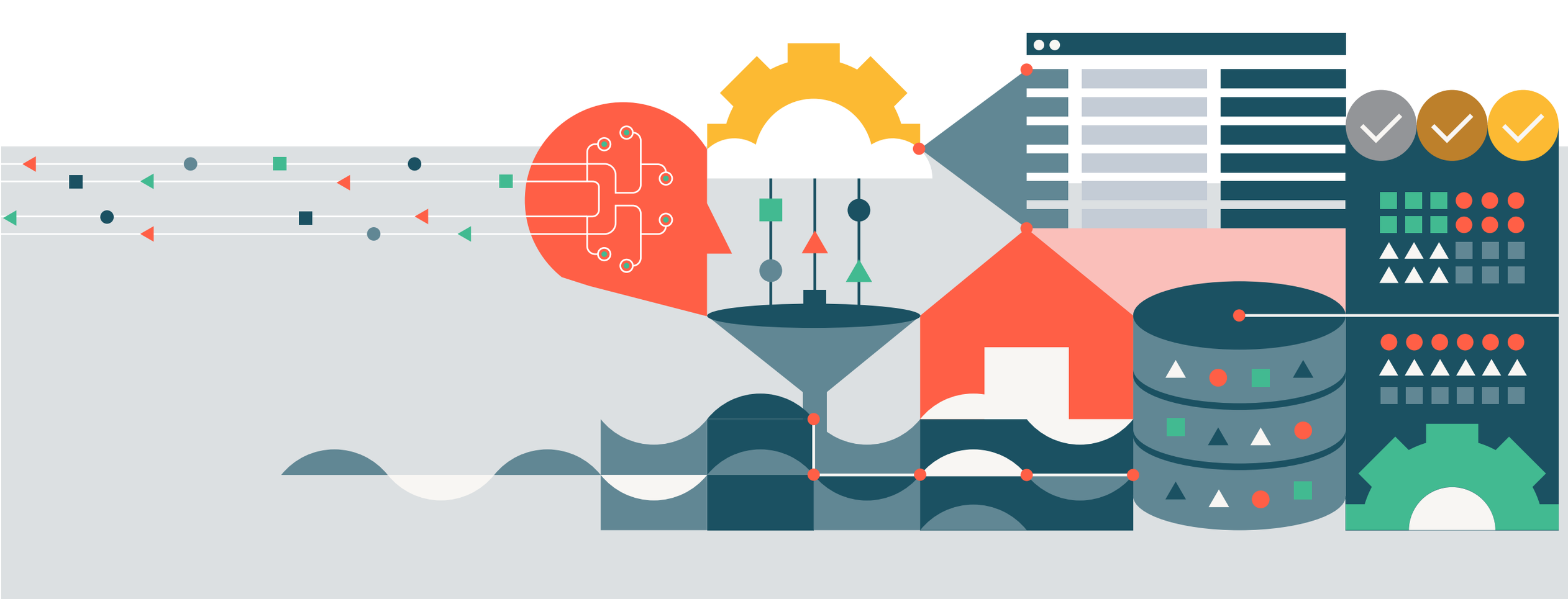
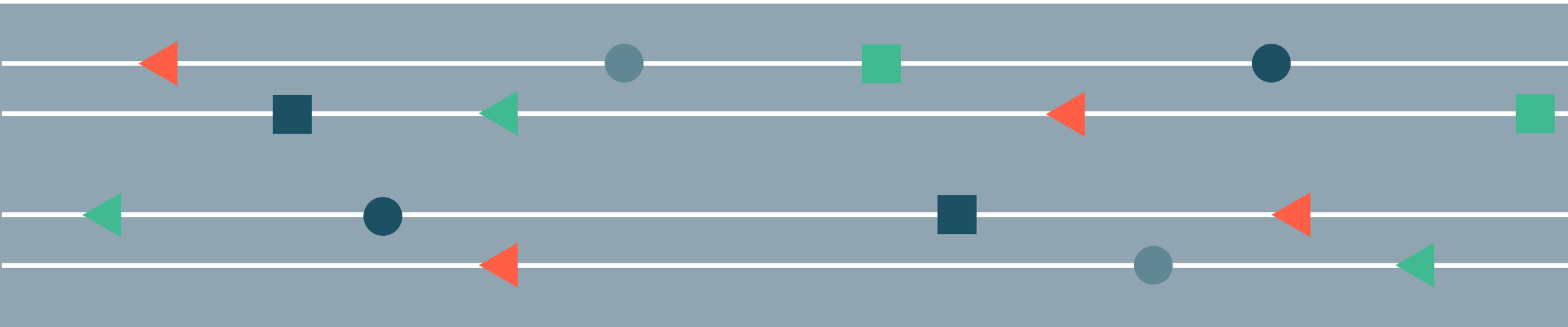
databricks

eBook

Governance:
The Unseen Foundation of AI Success



Contents



As data continues to explode and rapid technological advancements accelerate, the relationship between data governance and artificial intelligence (AI) has become a critical linchpin for success. How data organizations manage, secure and leverage data directly impacts AI implementation outcomes and considerations. **Simply put, you can't have AI without high-quality data, and you can't have high-quality data without data governance.**

While the concept is straightforward, the interplay between robust data governance frameworks and the successful deployment of AI strategies is difficult for many organizations to achieve. AI capabilities and machine learning (ML) models function differently than traditional, fixed record systems, requiring a modern data infrastructure and advanced tooling for efficient processing and reliable results. For these technologies to perform as expected, organizations must train AI and ML systems on trustworthy data to ensure accuracy.

Unfortunately, most enterprises face a dual-platform dilemma — with the absence of a comprehensive and actionable data and AI governance strategy at the core of the problem. With disparate systems encompassing the entire range of data applications, spanning from business intelligence to ML, enterprises are struggling to provide accurate and complete data and ML assets in a secure and governed way.

When data and AI are treated as separate entities and not governed together, organizations often encounter significant challenges. The segregation of data from AI processes creates silos that impede the free flow of information, which is critical for AI to learn, adapt and make accurate predictions. This disjointed approach can lead to inconsistencies in data handling, increased vulnerability to breaches and difficulty maintaining compliance with data protection regulations. As AI models rely heavily on large datasets to train algorithms, the lack of integration can result in AI that is unable to effectively access or utilize the latest data. Organizations are realizing that they must have access to all relevant data — of any type, from any source, in real time — governed comprehensively and shared across an industry ecosystem. When organizations can achieve this with the right data and AI foundation, they can turn AI strategy into reality and have the beginnings of data intelligence. But true data intelligence also empowers organizations to democratize the power of AI across their entire organization.

CIOs agree that a strong data strategy, with a robust governance foundation, is critical for AI initiatives to succeed.

This has brought into focus the importance of how governance is implemented and the need for built-in intelligence native to governance itself.

98%

Of CIOs believe a single built-in governance model for data and AI is critical

96%

Of CIOs believe a single system for structured and unstructured data used for BI and AI is important

#1

Democratizing data and AI is the #1 challenge to achieving generative AI goals across enterprises and government agencies

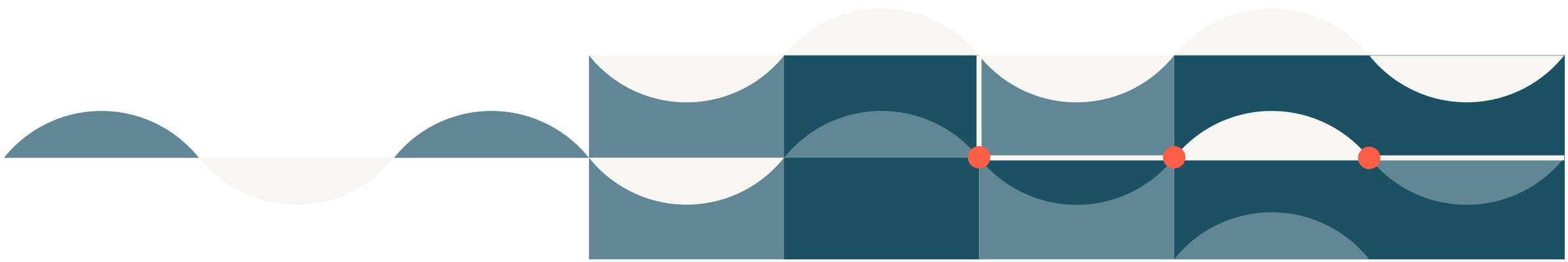
Source: MIT Tech Review Insights: Laying the foundation for data and AI-led growth

Unity Catalog: Unified Governance for Data and AI

To develop a trustworthy, efficient and scalable AI strategy, we are seeing businesses adopting a data lakehouse architecture that provides built-in, customizable data governance protections. In fact, based on the recent MIT Tech Review Insights Report, 74% have already made the move to lakehouse architecture. Data practitioners need a comprehensive data platform that can accommodate the size and scale of massive datasets and effectively gain value from that data through AI and ML modeling. Organizations can achieve this by:

- **Unified visibility into data and AI.** Ensure easy discoverability and seamless collaboration through the unification of data and AI assets, and the ability to catalog data from various sources.
- **Enforcing governance and security.** Secure data assets with a centralized approach to enforcing fine-grained access controls, auditing and governance policies — simplifying the overall management of data and AI resources.
- **Automated monitoring of data and AI.** Ensure high-quality data and fair, unbiased ML models with AI-powered monitoring that proactively identifies errors, conducts root cause analysis, and upholds the quality standards of both data and AI pipelines.
- **Open data sharing across the organization.** Easily share data and AI assets seamlessly across teams. This collaborative sharing fosters improved teamwork and flexibility in utilizing shared resources for enhanced productivity.

As the industry's first unified governance solution for data and AI on the lakehouse, Unity Catalog seamlessly manages structured and unstructured data, ML models, notebooks, dashboards, and files on any cloud or platform. Data scientists, analysts and engineers can securely discover, access, and collaborate on trusted data and AI assets, leveraging AI to boost productivity and unlock the full potential of the lakehouse environment. Organizations across industries can accelerate data and AI initiatives while ensuring regulatory compliance with security and simplicity.



Unifying governance *and* recognizing industry differences

Each industry has specific circumstances that influence the approaches organizations take to governing data. Take, for example, the strict rules for healthcare providers on how sensitive patient data must be secured, or financial institutions' complex reporting requirements across a global regulatory landscape. The nuances of industry ecosystems comprising different partners, systems of record, and compliance considerations put governance squarely at the center of how organizations share data to operate and innovate, without sacrificing security.

Keep reading to see how business leaders within different industries achieve three universal goals where data governance is critical — innovation acceleration, customer trust and compliance, and operational efficiency. See how data governance with Unity Catalog fuels AI within real-world use cases and imagine how your organization could benefit from a similar strategy.

Business Goal #1: Innovation Acceleration

Companies must constantly move toward better, faster, and more available products and services to remain competitive in any market. Data governance is the cornerstone of today's greatest innovation accelerator — data and AI — because it provides a structured framework to ensure data quality, security and compliant use. Using Unity Catalog on the Databricks Data Intelligence Platform, organizations can securely democratize access to high-quality data at scale and facilitate cross-team collaboration. This combination of access and capabilities fosters an environment conducive to AI experimentation and development.

Clear data governance policies, standards and automation enable teams to navigate data complexities efficiently, reducing time spent on data cleaning and enhancing the availability of reliable datasets for AI models. Business leaders gain accurate, timely and relevant insights for strategic decision-making to move innovation in the right direction from the top down. Moreover, with the single-permission model in Unity Catalog, organizations mitigate the risks associated with data misuse or bias, fostering trust in AI systems. The Data Intelligence Platform provides a single point of access to securely consolidate and query data from multiple sources. By promoting transparency, compliance and accountability in managing data, the Databricks Platform and Unity Catalog encourage and facilitate exploring new AI applications to help organizations stand out in their market.

Use case: Secure data sharing and collaboration in manufacturing

The manufacturing industry generates massive volumes of complex unstructured data via sensors, images, video and telemetry across vehicles, factories, buildings and workers. Organizations need the ability to stream data in real time and fuse it with contextual data sources to respond to events meaningfully. Data often comes from various platforms, including legacy systems, making it complicated to gain complete insights or share data across teams.

Rivian, the Electric Adventure Vehicle (EAV) manufacturer, moved off of its legacy cloud tooling to the Databricks Data Intelligence Platform to remove the data silos and barriers to collaboration that stall innovation. Unity Catalog enables end-to-end version-controlled governance of users and sensitive datasets with auditability across workspaces. With a simple but secure framework, Rivian has unified its data into a common view for downstream analytics and ML to support AI use cases.

On the Databricks Platform, Rivian increased runtime performance by 30%–50%, leading to faster insights and analytics and improved model performance for AI projects.

Although 40% of Rivian's data users are nontechnical business users, the company expanded Databricks Platform users from 5 to 250 — a 50x increase — in only one year. Rivian uses AI for predictive and preventative maintenance, remote diagnostics, and service fulfillment by scaling its capacity to deliver valuable data insights with speed, efficiency and reliability. These new capabilities have also unlocked new use cases for utilizing ML and AI systems to optimize battery efficiency, increase the accuracy of autonomous driving systems, and serve commercial depots with vehicle health dashboards for early and ongoing services.

RIVIAN

> Read the case study to learn more.

"We now have a centralized data catalog and access management across various teams and workspaces, which has opened up unprecedented opportunities for collaboration and innovation."

— Venkat Sivasubramanian, Senior Director of Big Data, Rivian

databricks

Use case:
Smarter product development in retail and consumer goods

Consumer-focused industries like retail and consumer goods (RCG) have both an AI advantage and a curse due to the wealth of business, customer and market data available. By establishing data quality standards and ensuring a unified view of information across various touchpoints, governance enables retailers to gather comprehensive insights into consumer behaviors, preferences and market trends. This high-quality data serves as the bedrock for AI-driven analytics and predictive models, aiding in identifying emerging trends, demand forecasting and personalized product recommendations.

Anker opted to unify over 200 terabytes of data from its business, product, marketing and sales sources on the Databricks Data Intelligence Platform to foster collaboration with Unity Catalog governance. Swapping complicated user account access to a data warehouse for granular access controls and auditability in Delta Lake — the default storage layer for all operations on Databricks — Anker could confidently make data available to data analytics and business teams for intelligent decision-making.

Since consolidating on the Databricks Platform, Anker can now quickly ingest and visualize large amounts of data to uncover information, trends and patterns. The speed and unobstructed views help the device charging company rely on local insights to inform expansion strategies to other markets. Over 500 Anker employees use the Databricks Platform and have generated over 1,000 reports and dashboards to inform organization-wide outcomes. Furthering innovation, Anker is integrating large language models (LLMs) for use cases like auto-generated SQL commands and code review that can accelerate productivity and time to insights.

Anker Innovations

> Read the case study to learn more.

"The Databricks Data Intelligence Platform made it possible for us to take insights and innovation from local markets and then surface these to our global product, marketing and manufacturing teams. Teams now have an easy and secure way to share data and build AI solutions that can be scaled worldwide."

— Sean Zuo, Chief Information Officer, Anker Innovations

Use case: Data-driven decision-making in financial services

Data governance in the financial services sector establishes data collection, storage and usage protocols to ensure compliance with stringent regulatory frameworks. By implementing strong governance, financial institutions cultivate a foundation of trust in their data, enabling AI systems to analyze large and complex datasets swiftly and accurately. These systems aid in generating precise risk assessments, optimizing investment strategies, identifying market trends and detecting fraud.

The International Finance Corporation (IFC) created MALENA, an AI-powered ESG analytics platform where users can extract meaningful insights from unstructured ESG data at scale for rapid analysis, increased productivity and building confidence. IFC successfully scaled the MALENA platform using the Databricks Data Intelligence Platform, so emerging market investors can quickly extract insights from unstructured data to identify better-quality investments.

Collaboration within the Databricks Platform fueled MALENA's development and supported AI model training for natural language processing (NLP) and LLMs. Now that MALENA can analyze 19,000 sentences per minute — a 950x increase compared to the average human reader — what used to take months takes moments. The IFC has developed 10,000 company profiles and expanded insights for over 180 markets.

IFC International Finance Corporation
WORLD BANK GROUP
Creating Markets, Creating Opportunities

Read the case study to learn more.

"The operational efficiencies that we've gained using the Databricks Data Intelligence Platform have paid off in spades. We can now lean on the power of GPU computing and AI to support our global ESG initiatives — leading to better-quality investments and a more sustainable future."

— Laura McCanlies, Chief Information Officer, IFC

Business Goal #2: Customer Trust and Compliance

Fostering public trust and regulatory compliance is a non-negotiable that demands data governance be a top priority for organizations across industries. Through data governance and security policies, organizations demonstrate their commitment to safeguarding sensitive customer information to instill customer confidence regarding data handling and utilization.

Beyond building rapport with customers, organizations must meet industry regulations and global standards to avoid legal and regulatory penalties, reputational damage, revenue loss, and business disruption.

The Databricks Data Intelligence Platform and Unity Catalog simplify compliance with advanced tooling to track data lineage, fine-tune access controls, generate audit trails, and automatically detect and tag sensitive data. Data practitioners can use Unity Catalog to centrally track data lineage and data asset usage across the platform for organization-wide data transparency. These built-in protections reduce the likelihood of errors or misuse, reinforcing customer trust and long-term compliance with data protection and privacy requirements.

Governance for AI
Unifying data and AI catalogs under one roof
Learn more

Use case:
Efficient data governance at enterprise scale in healthcare and life sciences

The infinite amount of personally identifiable information (PII) managed by the healthcare and life sciences (HLS) industry must be treated with the utmost security to protect privacy while enabling advancements through data science. Operating per contractual, legal and regulatory standards like HIPAA and FERPA can be complex, but when you embed data governance into an organization's data culture, it becomes routine. Building on the result of that strategy — accurate, consolidated and visible data — the HLS industry can train AI models to contribute to lifesaving interventions with diagnostic precision, personalized treatments and drug discovery.

Pioneering new drug-making processes and developing lifesaving medicines make data and AI pivotal to Amgen's business strategy as the world's largest independent biotech company. To unlock the true potential of its data and AI across various departments, Amgen needed a governance solution capable of both regulatory-grade oversight and open-access collaboration. Amgen integrated Unity Catalog on the Databricks Data Intelligence Platform to streamline operational efficiency for compliant and trustworthy data management that would accelerate drug discovery.

Unity Catalog gave Amgen the flexibility and specificity necessary to manage data governance properly while increasing collaboration for its 10,000 active users. Data teams went from managing more than 120 roles down to just one or two principal roles using the fine-grained access controls in Unity Catalog. Easy auditability and query histories resulted in a 50% efficiency improvement in audit management and contributed to a 20% reduction in Databricks' cost. Now able to scale its data ecosystem efficiently and compliantly, delivering data reliability and availability through data governance, Amgen can expand into AI use cases that directly impact drug discovery.

AMGEN

> Read the blog to learn more.

"Integration of the Unity Catalog has enhanced our ability to implement precise and intricate governance policies for Amgen's restricted datasets. This remarkable achievement has sparked immense enthusiasm within our data engineering department, leading to increased investment in our data platform, with Unity Catalog serving as the central Metastore and access management service."

— Jaison Dominic, Senior Manager, Information Systems, Amgen

databricks

Use case:
Faster threat detection in communications, media and entertainment

The communications, media and entertainment (CME) industry is expanding alongside technological innovations that highlight new data governance needs in accordance with growing regulations. For successful AI implementation to occur, CME organizations must first meet these standards for high-quality data and customer privacy to guarantee the accuracy and relevance of content recommendations, personalization strategies and targeted advertising. Transparency through data governance fosters trust by empowering consumers with control over their preferences and privacy settings while providing insights into how their data shapes content curation and advertising.

As a massive customer data manager, BlackBerry is tasked with investigating and understanding the attack vectors used by bad actors to complete successful infiltration. Armed with these insights, the organization optimizes its network strength to protect customer safety and respond to attacks swiftly. BlackBerry consolidated its data on the Databricks Data Intelligence Platform to provide data analysts, data scientists and threat researchers with an end-to-end view of their data for simple, secure collaboration that meets data privacy laws and telecommunications regulations.

Read the blog to learn more.

After breaking down data silos, BlackBerry iteratively improved its data processing pipelines to speed data ingestion and reduce query latency by more than 20%. That boost is passed on to BlackBerry customers through rapid threat detection that enhances the security and recovery of data. The Databricks Platform delivered the governance controls BlackBerry needed to ramp up the strength and speed of its threat detection tasks with confidence and a foundation that can expand and scale for AI and ML use cases.

BlackBerry

"It allowed us to create a unified view of our data estate, simplifying collaboration across teams. We now have a standard approach to manage access permissions and audit files or tables in our lake, with the ability to define fine-grained access controls on rows and columns. Automated data lineage helped us see where the data is coming from to pinpoint the source of a potential threat and to understand which research projects or teams are leaving the data for threat detection."

— Justin Lai, Data Architect, BlackBerry

Use case: Fast and compliant product development in transportation

Transportation is a huge data industry with many moving parts — from passengers, ticketing and machinery to weather, communications and regulations — and everything generates data that informs the next best action. To manage these moving targets efficiently, transportation organizations must be transparent with customers who are often impacted in real time by service interruptions. Gaining control of such a large and interconnected system means implementing governance software to leverage AI for customer notifications, predictive and preventative maintenance, weather forecasting, route optimization, and regular compliance oversight.

Aviation giant JetBlue is no stranger to AI and ML and uses the Databricks Data Intelligence Platform to take advantage of recent advancements in generative AI for faster product development. Deploying new ML products often demands exhaustive change management processes, particularly due to oversight by Federal Aviation Regulations and other laws protecting highly sensitive data. These processes include a series of workshops, training and product feedback that can slow iteration. JetBlue sought to empower users and speed time to market using LLM technology to gain access to role-specific KPIs and information, including help using natural language they're familiar with.

Read the blog to learn more.

Data governance was the biggest obstacle due to the criticality of role-based access monitoring in highly regulated businesses like aviation. Unity Catalog gives JetBlue role-based access to documents in the vector database document store so that any JetBlue user can quickly access the same chatbot for accurate information. This organizational flow of strong data governance catalyzes internal teams to produce more, better and faster without risking inaccurate results, noncompliance or a breach of public trust. JetBlue has over 10 million products in production within the Databricks Platform, including dynamic pricing, customer recommendation engines, supply chain optimization and customer sentiment NLP.

jetBlue

"With the Databricks Data Intelligence Platform serving as the central hub for all streaming use cases, JetBlue efficiently delivers several ML and analytics products and insights by processing thousands of attributes in real time."

— Sai Ravuru, Senior Manager of Data Science and Analytics, JetBlue

databricks

Business Goal #3: Operational Efficiency

Streamlining processes and automating tasks are among the most anticipated AI use cases across every industry due to the phenomenal productivity gains realized with operational efficiency. Paramount to achieving desired outcomes is the seamless flow of credible data between teams, devices, factories, production lines and distribution channels. The Databricks Data Intelligence Platform and Unity Catalog connect these dependent systems at scale by defining roles, responsibilities and workflows to enable democratization. Through cross-team collaboration, data teams can build and implement successful AI models faster and with more precision.

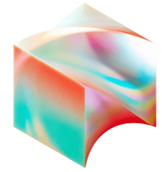
Leveraging unified visibility into data and AI with Unity Catalog immediately encourages exploration with a single point of access. Here, teams leverage generative AI to automatically document their data tables and columns. This documentation then provides additional context to the data and fuels intelligent search to find, understand and extract meaning from data using natural language. Essential to implementing efficiency is identifying errors where bottlenecks are disrupting profitability. Unity Catalog uses AI-powered, automatic monitoring and observability to diagnose errors quickly and uphold data and ML model quality. Proactive alerts automatically detect PII data, track model drift and resolve issues within pipelines to maintain accuracy and integrity. Operational efficiency is built on a solid foundation for strategic initiatives, and Unity Catalog aligns data assets with organizational goals so that outcomes can be achieved.

Use case:
Reducing data infrastructure costs in financial services

Data governance expedites AI implementation in financial services by supplying models with reliable and compliant data from various complex sources, consequently enhancing operational efficiency. Standardized data governance policies and procedures minimize errors and discrepancies in financial data, reducing the time and resources spent on rectifying inaccuracies. Data governance also prioritizes compliance so financial institutions can meet regulatory requirements and avoid noncompliance penalties without hindering data maturity. With these proactive protections, AI-driven algorithms can fulfill use cases for optimizing resource allocation, enabling proactive decision-making, and automating routine tasks like loan approvals, credit scoring and portfolio management.

As a global technology company with various financial services subsidiaries, Block requires proper implementation and uniformity of data governance policies to ensure compliance with privacy laws. Needing to enable secure and compliant access to PII data, Block adopted the Databricks Data Intelligence Platform and Unity Catalog for centralized data, AI and governance at scale. The comprehensive view of data across different business units simplified access permission management within Unity Catalog's single-permission model. With this flexibility Block is distributing cost attribution among teams based on storage locations for catalogs and schemas. Now, different business units can maintain their distributed data governance policies while maintaining a streamlined process.

Substantial improvements in compute performance, cost optimization and data team efficiency reduced Block's compute costs by 12x and enabled use cases that were previously unattainable due to scaling limitations. Using Unity Catalog to create a dynamic marketplace for data exchange between business units played a crucial role in cutting the data egress costs associated with cross-cloud provider data transfer by 20%. Block manages 2PB of data on the Databricks Platform with approximately 70 different teams across business units, and 300 active power users — all contributing to and benefiting from secure access to reliable data.



BLOCK

Read the case study to learn more.

"Unity Catalog distills permissions management into its important pieces by removing the complexity introduced by IAM policies and other data control platforms, allowing us to focus on what is important to our business use case."

— Joseph Kesting, Software Engineer, Block

databricks

Use case: Fueling a hyperefficient supply chain in retail and consumer goods

From product development and manufacturing to distribution, marketing and sales, RCG organizations rely on a steady stream of data to optimize supply chains, predict market trends, manage inventory, influence buying behaviors and drive revenue. Data governance is the link between these processes, providing a secure, compliant and comprehensive view of data sources and usage for smart and efficient operational decision-making. Assuring authorized personnel access through strong governance optimizes workflows, data pipelines and operational processes to augment AI applications for faster time to insights and responses to market shifts and customer demands.

Barilla is the largest pasta producer in the world, with over 300 different food products made in more than 90 factories and distributed across 100 countries and counting. Growing distribution, product demand, and over 1TB of data ingested each day — from manufacturing logs and sensor data, ESG reports, inventory, financial metrics, and marketing media spend — motivated Barilla to modernize its data ecosystem on the Databricks Data Intelligence Platform. Ninety percent of Barilla's business runs on the Databricks Platform, with more than 40 data products in production.

Unity Catalog's simple access management tools give Barilla a single-permission model to easily define access policies on data assets that support the decision-making of over 2,000 internal data users, half of whom are building dashboards to improve supply chain operations. As a result, Barilla can solve business challenges swiftly, saving millions of dollars in production costs. Manufacturing cost deployment models extract and allocate industrial costs and quantify plant losses, as well as support predictive and proactive maintenance. To streamline supply chain management, Barilla implemented a traceability system using the Databricks Platform that analyzes supplier performance to improve inventory management. Additionally, marketing teams are understanding customer behavior and preferences to identify new markets and segment customers for more revenue opportunities.

Read the case study to learn more.

"Moving to Databricks was a game changer for us. We use all the key Databricks technologies, including Unity Catalog to easily discover, understand and use data more securely through catalogs of data assets that are discoverable through shared tables and views."

— Graziano Belmonte, Global Big Data and Advanced Analytics Director, Barilla

Use case: Optimizing data management productivity in advertising and marketing

Competition is fierce for consumer attention, and advertising and marketing professionals depend on many datasets and sources to identify and segment buyers, track engagement, optimize campaigns, and, ultimately, drive sales. Joining internal customer lifecycle data with external market trends allows organizations to strategize long-term, using predictive forecasting, personalization and dynamic testing for continuous improvement. Data governance makes these AI use cases possible by supplying the right people with accurate data in an open arena where collaboration and innovation accelerate business outcomes.

MezzoMedia provides various specialized digital marketing solutions to Korean organizations, having executed over 1,600 campaigns for advertisers and billing more than $620 billion last year. The priority for MezzoMedia is creating new value and achieving innovation, so it modernized its data infrastructure on the Databricks Data Intelligence Platform to explore, analyze and monitor large-scale data with high efficiency. On the Databricks Platform, data teams can improve collaboration to significantly increase overall data management productivity with advanced tooling for data linkage, individual data management, and authority control using Unity Catalog.

Since implementing the centralized platform, MezzoMedia has reduced costs by 80%, simplifying complex processes and supporting more flexible and diverse workflows. Data engineers can actively respond to various use cases by integrating internal and external data and setting individual management and account privilege controls. Scheduled and automated data processing and analysis tasks lower repetition and redundancy, contributing to a 60% reduction in processing and collection time for 100TB of data per month.

Leveraging these operational improvements, MezzoMedia has also reduced data analysis and modeling implementation time by 60%, accelerating its ability to drive AI outcomes throughout the organization. By simplifying and democratizing data access issues with Unity Catalog, MezzoMedia is expanding AI to drive growth and maintain a competitive edge in the advertising market.

MezzoMedia

Read the case study to learn more.

"MezzoMedia has constructed an integrated data platform based on Databricks after long contemplation on how to efficiently process vast amounts of data for advertising execution. Based on such contemplation, we provided ad planning through real-time reports and data analytics to clients and media, enabling customers to reduce costs and maximize performance."

— Gi-hwan Kim, Director of Data Solutions, MezzoMedia

databricks

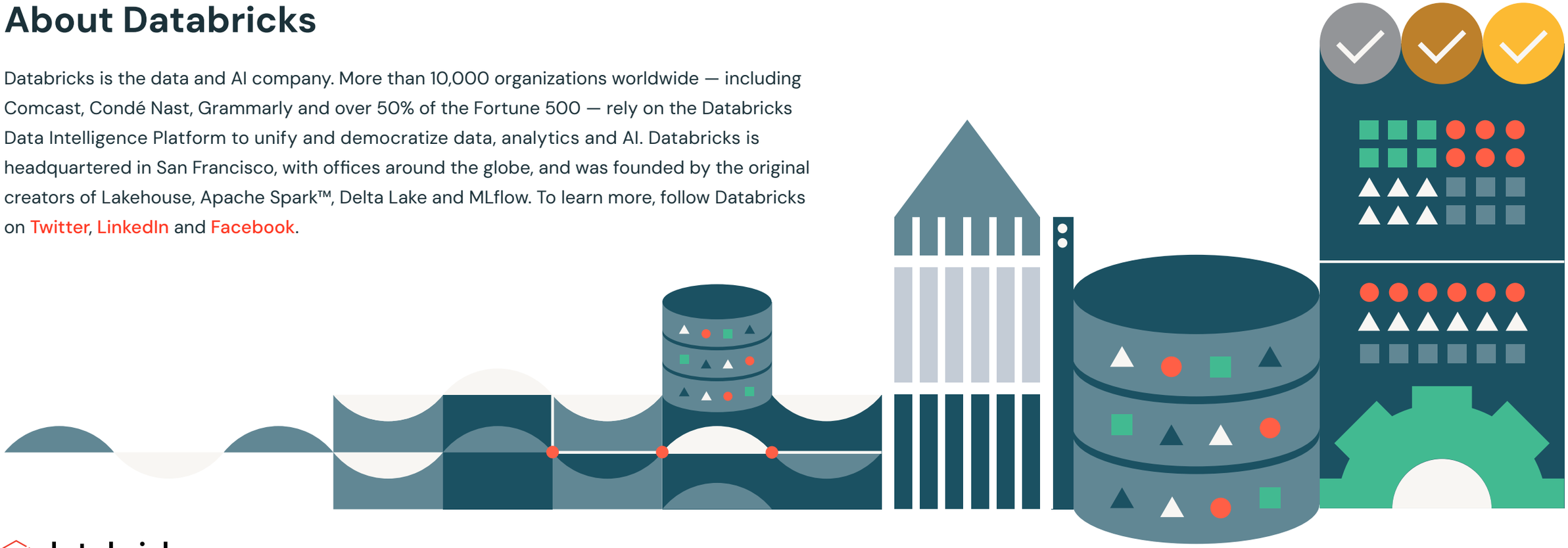
Data Quality Basics for AI-Powered Innovation

Data governance impacts every organization's ability and capacity to reach business outcomes because of the implications for data maturity. Organizations are wasting time and money spinning their AI wheels without clear and defined standards, proactive security, and unobstructed views of internal and external data. Only through embedded policies, automation and consolidation can organizations ensure the quality of their data before training AI and ML systems to depend on it.

The Databricks Data Intelligence Platform and Unity Catalog give data teams confidence through data reliability to expand and accelerate AI use cases. Armed with the simplicity, ease and security of the Databricks Platform, organizations across industries are experiencing significant gains in data team productivity, fostering a data-driven culture prioritizing AI innovation. Realize the potential of your data with built-in, comprehensive data governance on the Databricks Data Intelligence Platform to keep your organization growing and ahead of the competition.

About Databricks

Databricks is the data and AI company. More than 10,000 organizations worldwide — including Comcast, Condé Nast, Grammarly and over 50% of the Fortune 500 — rely on the Databricks Data Intelligence Platform to unify and democratize data, analytics and AI. Databricks is headquartered in San Francisco, with offices around the globe, and was founded by the original creators of Lakehouse, Apache Spark™, Delta Lake and MLflow. To learn more, follow Databricks on Twitter, LinkedIn and Facebook.



databricks