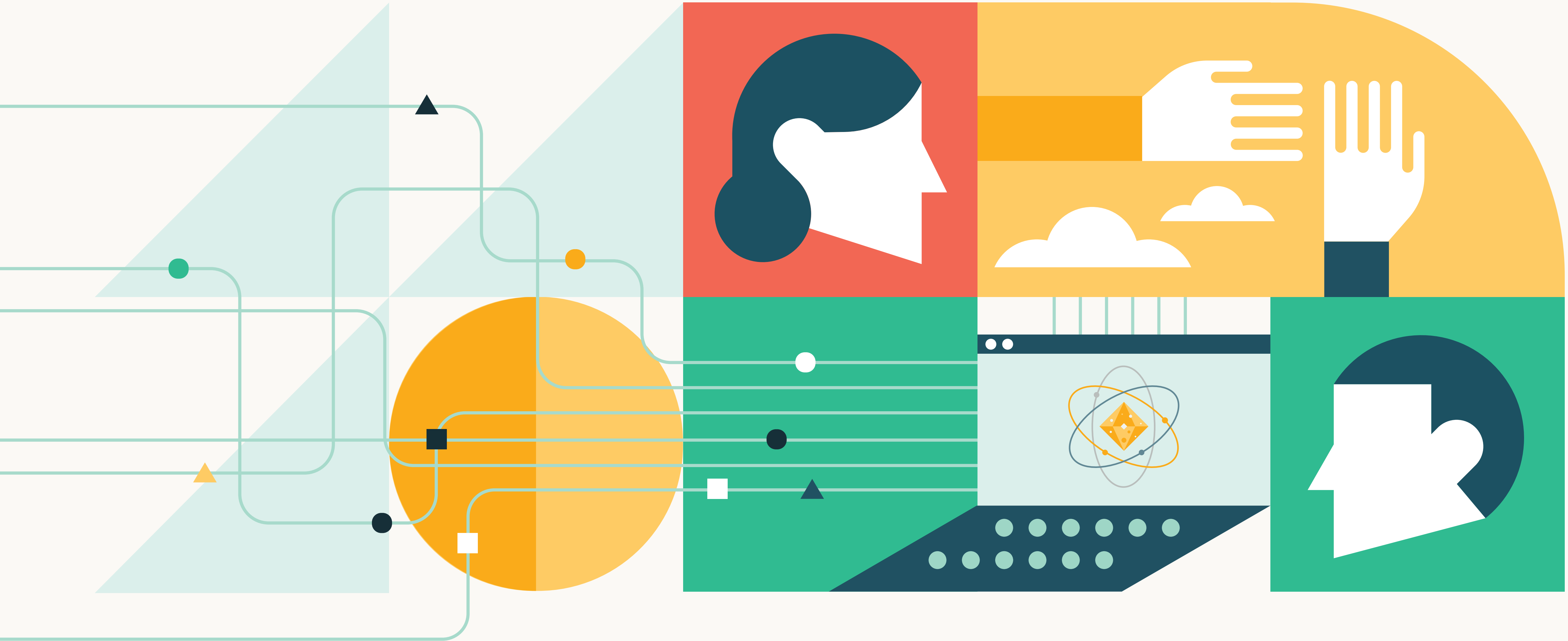
eBook

From ETL to Gen AI

Uniting Data Engineers and Data Scientists for AI Innovation



databricks

Contents

Key Terms

Generative AI
Artificial intelligence that learns the patterns and structure of your input data or existing online data and generates new content in response to prompts.

Data intelligence
The deep understanding of enterprise data — including contents and metadata, as well as how it is used (queries, reports, lineage, etc.,) — made possible through the embedding of generative AI into data platforms.

Data science
An interdisciplinary field that combines statistics, specialized programming, advanced analytics, AI and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization's data.

Data engineering
The process of designing and building systems for collecting, storing, transforming and orchestrating data at scale to convert raw data into usable and accessible information.

ML engineering
The process of designing and building software models that can automate AI and machine learning models to perform tasks without instructions.

Executive Summary

Generative AI (GenAI) has the potential to democratize AI and transform every industry and every function of the enterprise to support every employee and to engage every customer.

Investment is exploding as enterprises look for ways to convert their valuable data into intellectual property. The urgency for GenAI is apparent. Accenture[1] found that generative AI has the potential to impact 44% of all working hours across industries in the U.S., enable productivity enhancements across 900 different types of jobs and create at least $8 trillion in global economic value. McKinsey[2] predicts that generative AI and related technologies could automate activities that currently take up 60%–70% of worker time.

The Databricks 2023 State of Data + AI report shows organizations are putting substantially more models into production (411% YoY growth) while also increasing their ML experimentation (54% YoY growth). But for all their experimenting and testing, organizations still lack confidence in their AI models. The same report shows that for every three experimental models, roughly one is put into production.

Data engineers and data scientists struggle with difficult handoffs between teams due to disparate platforms, tools and processes from data ingestion and prep to experimentation and production. A lack of unified processes and access controls between systems for governance, auditability/traceability and regulatory compliance introduces risks of undesirable model outputs, or worse. The same lack of unification makes AI models more difficult to build and expensive to run at scale.

Generative AI is inextricably linked to data; good data is what makes an AI model intelligent. When the needed data spans many siloed platforms, the challenges of data accuracy, security and control are multiplied.

Databricks' mission is to democratize data and AI, allowing organizations to use their own unique data to build or fine-tune their own machine learning and generative AI models so they can produce new insights that lead to business innovation.

This eBook shows how the data engineering and data science disciplines are becoming more intertwined as organizations prioritize high-quality AI built on the solid foundation of reliable, secure data, and provides an overview of best practices, systems and tooling to unify both disciplines under a single platform.

[1] Accenture Technology Vision 2024: "Human by Design" Technologies Will Reinvent Industries and Redefine Leaders by Supercharging Productivity and Creativity

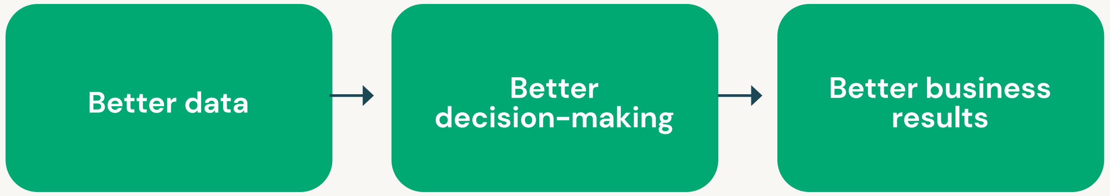
[2] McKinsey & Company, 2023 Report: "The economic potential of generative AI: The next productivity frontier"

A Data-Centric Approach for the Age of AI

The winners in every industry will be data and AI companies.

Today, understanding data and AI has become a prerequisite for competitive differentiation and long-term success. A fresh, accurate, comprehensive grasp of your unique data — as it is understood, analyzed and democratized by your own AI — is what will set you apart from your peers.

We're all pretty comfortable with the following idea:



Better data → Better decision-making → Better business results

This isn't exactly controversial. Better data is the key to unlocking a deeper understanding of customers, prospects, market forces, internal dynamics and so on.

This is why data engineering exists. The discipline of data engineering exists to deliver better data to the business.

But sometimes organizations take that arrow between "better data" and "better decision-making" for granted. It's not a given that we're able to engineer excellent data pipelines and *automatically* emerge with better decisions.

In between "better data" and "better decision-making" lies the critical function of *extracting more value* from that data. Here, the discipline of data science flourishes. If "data is the new oil" (something we've collectively been saying for almost two decades now), then data science is the refinery necessary for converting it into something useful, and generative AI is supercharging the potential of every refinery worldwide.

Despite the interdependence of data engineering and data science in the value chain, they've only recently converged into a cohesive operating model for data and AI.

Initially, most data science software tools emerged from a model-centric approach to model management, which left data engineering teams to manage the critical data pipeline services and production environments.
This spread the access control, testing and documentation for the entire flow of data across multiple platforms and unnecessarily increased the complexity and risk for any ML application. It's no surprise we commonly hear aphorisms like *"garbage in, garbage out"* and *"80% of a data scientist's time is spent cleaning data"*.

In MIT's Technology Review Insights survey, 2022, 72% of executives agreed that **data problems** are the most likely factor to jeopardize their AI/ML goals. Like any data science project, generative AI is reactionary, meaning the quality of any model is constrained by the quality of data coming in.

Bringing AI models into production means understanding the data, the model and the development environment. Increasingly popular is the *data-centric approach* that brings the data and AI all into one unified workflow and simplifies the lifecycle of a given project with standardized tools, frameworks and governance.

A data-centric AI environment unifies all types of data, and in doing so, makes it fundamentally easier to build, deploy and manage large language model (LLM) applications.

There are four requirements for a data-centric approach to AI:

Trustworthy data from reliable data pipelines
Data engineering and platform teams easily deliver the right data to data scientists exactly when it is needed.

Democratized access to data
Pipeline development is quick and easy; ingestion, transformation and orchestration are managed on a single integrated platform; through data intelligence and built-in AI assistance, anyone in the organization can access the data they need and extract value from it.

Optimized cost/performance
Infrastructure management and pipeline dependencies are automatically optimized in the background, delivering ETL pipelines at the best possible cost/performance and making it easy to scale.

Unified governance for data and AI at scale
Singular, centralized oversight and security across all of an organization's unique, proprietary data and AI.

With the right people, processes and data stack in place, this approach is more than achievable today. In the following chapters, we'll discuss areas where data engineers and data scientists must work closer together within this framework to ensure optimal AI project outcomes, and other areas where data engineers can make it easier for data scientists to self-serve the data they need.

> **We are confident that a data-centric approach will enable us to decrease bias, increase efficiency, reduce costs in our AI efforts, and create accurate industry-leading models and AI agents.**
>
> — Greg Rokita, AVP
> Technology, Edmunds

A data intelligence platform for all teams

A data-centric approach to building AI models requires (1) a unified data platform, (2) built on data intelligence.

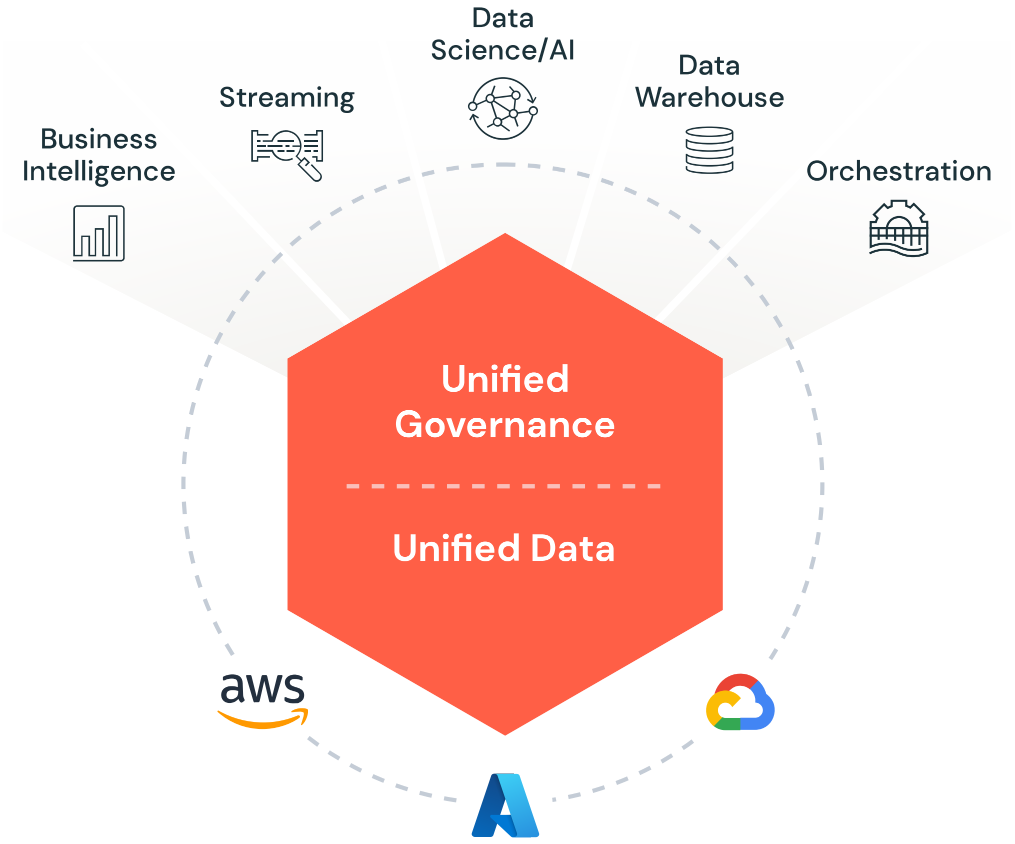
A Unified Data Platform

Most organizations struggle to realize the vision of bringing data and AI together, simply because different parts of their data estate are siloed and decoupled from one another:

- Data warehouses where you put all your structured data
- BI platforms to visualize and understand what's going in the business
- Data science and ML platforms for advanced use cases using data in your data lake
- ETL and orchestration to prepare and move data
- Real-time systems for streaming use cases
- Generative AI (!)

With siloed systems comes a siloed view of the world, and AI initiatives that are not fully integrated with the data they need. Instead, today's data and AI needs require a platform built on unification:

- Unified storage (one copy of all structured, unstructured and semi-structured data)
- Unified governance (securely discover, access and collaborate on trusted data and AI assets)
- Unified data and AI (simplify your data estate by eliminating the silos that historically complicate data and AI)
- It's for this reason that lakehouse architecture has become the de facto standard across data teams worldwide. In fact, today 74% of global enterprises have adopted a lakehouse, and almost all of the remainder intend to have one within the next three years. Lakehouse architecture is the open, unified foundation for all an organization's data and AI.



Data intelligence

The next evolution of lakehouse architecture is the infusion of data intelligence. If the core value of a lakehouse is unification, the additive value of a data intelligence platform is democratization.

A data intelligence platform is necessarily rooted in a lakehouse. The unification of data and AI is foundational to making a data intelligence platform work.

The next layer is the intelligence engine — a generative AI model that is built to (1) fully understand the unique semantics and structure of your data, and (2) use that understanding across everything in the platform. Here are some examples of how it works:

- Instead of manually indexing and partitioning data themselves, data engineers can ask the intelligence engine to do it for them
- Instead of needing to know code to build a query, anyone in the organization can interrogate a dataset in natural language like English, Spanish or Japanese
- Instead of using a monolithic public LLM for enterprise use cases, data scientists and ML engineers can quickly and cheaply build their own LLMs to their own standards and their own compliance rules

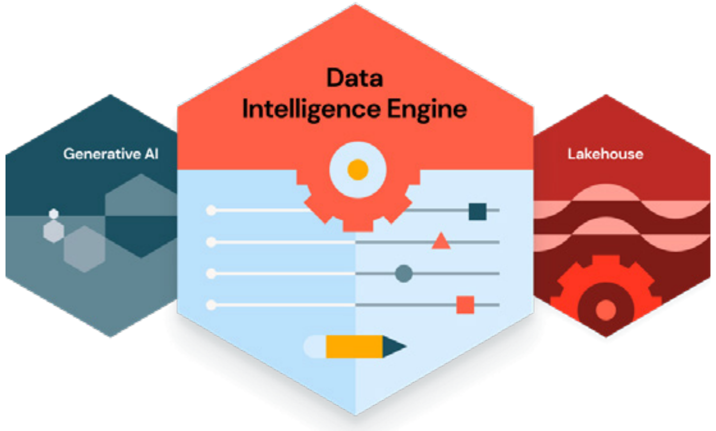
It's much easier for data engineers and data scientists to collaborate on the data needed for AI initiatives when they are operating off a single platform that encourages, empowers and almost enforces seamless collaboration.

The process flow below shows how the Databricks Data Intelligence Platform brings the data engineering and generative AI lifecycles together, allowing teams to work on the same data in the same environment. Data engineers identify data sources and ingest the data into the Bronze layer — part of a data design pattern called medallion architecture — and then join and clean the data in the Silver layer.

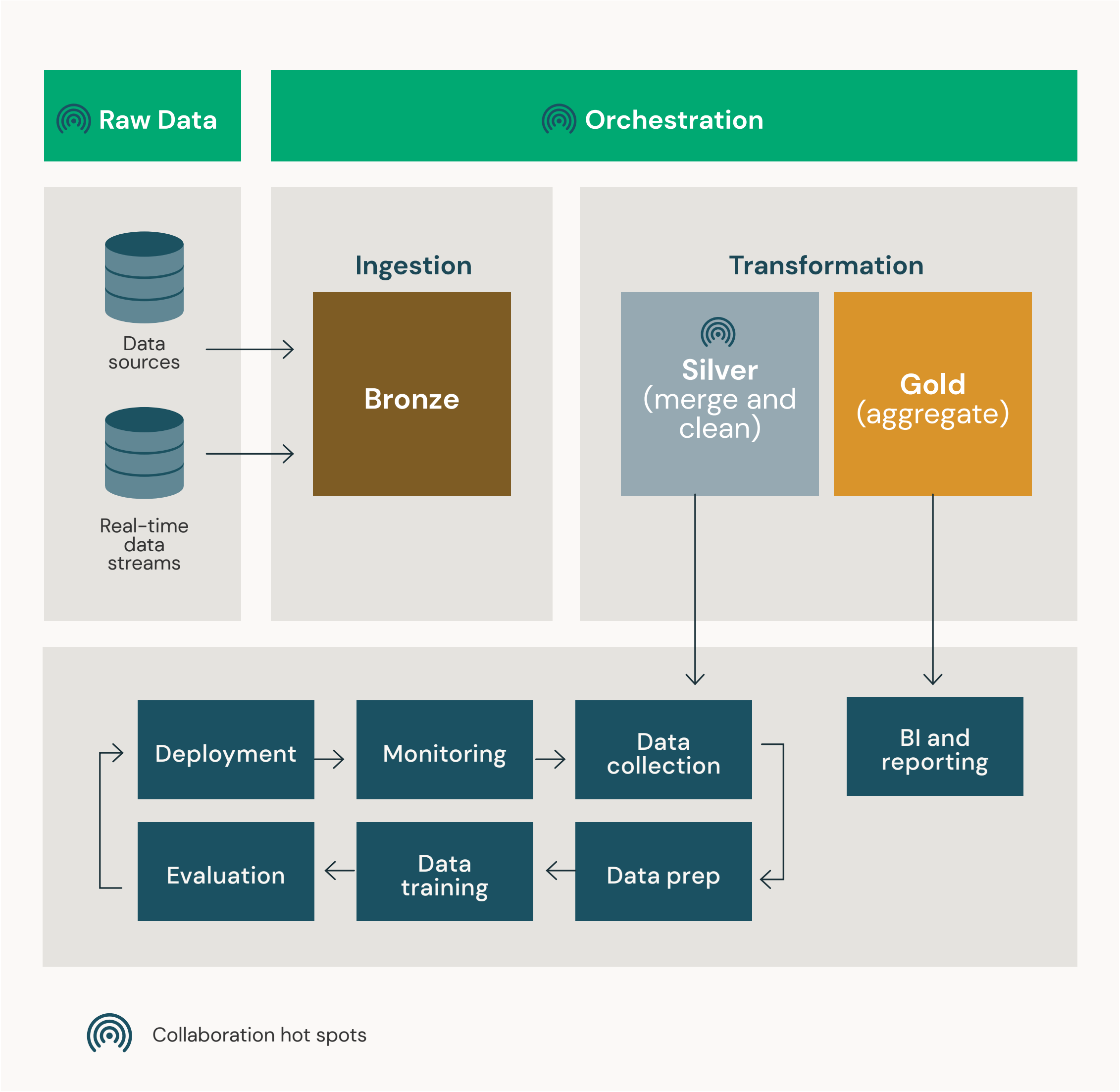
Data engineers also orchestrate this data pipeline to deliver data according to service-level agreements at an agreed-upon cost to the business. They might aggregate data into the Gold layer for BI and reporting, or consolidate data for data scientists to prepare in the generative AI lifecycle for model training, deployment and management.

The Databricks Data Intelligence Platform enforces collaboration at three key points. Data scientists can help identify the needed data sources for their models to accelerate the data collection and preparation process. The data science team can also help define what data needs to be merged and cleaned in the silver zone to improve the quality of data collected for their models for better performance once they are in production.

And to ensure that they have the right data where and when it is needed, the data science team can also collaborate on pipeline orchestration to reduce operational friction.



The following flowchart illustrates a typical workflow from source data to production models, with callouts for "collaboration hot spots" between data engineers and data scientists:



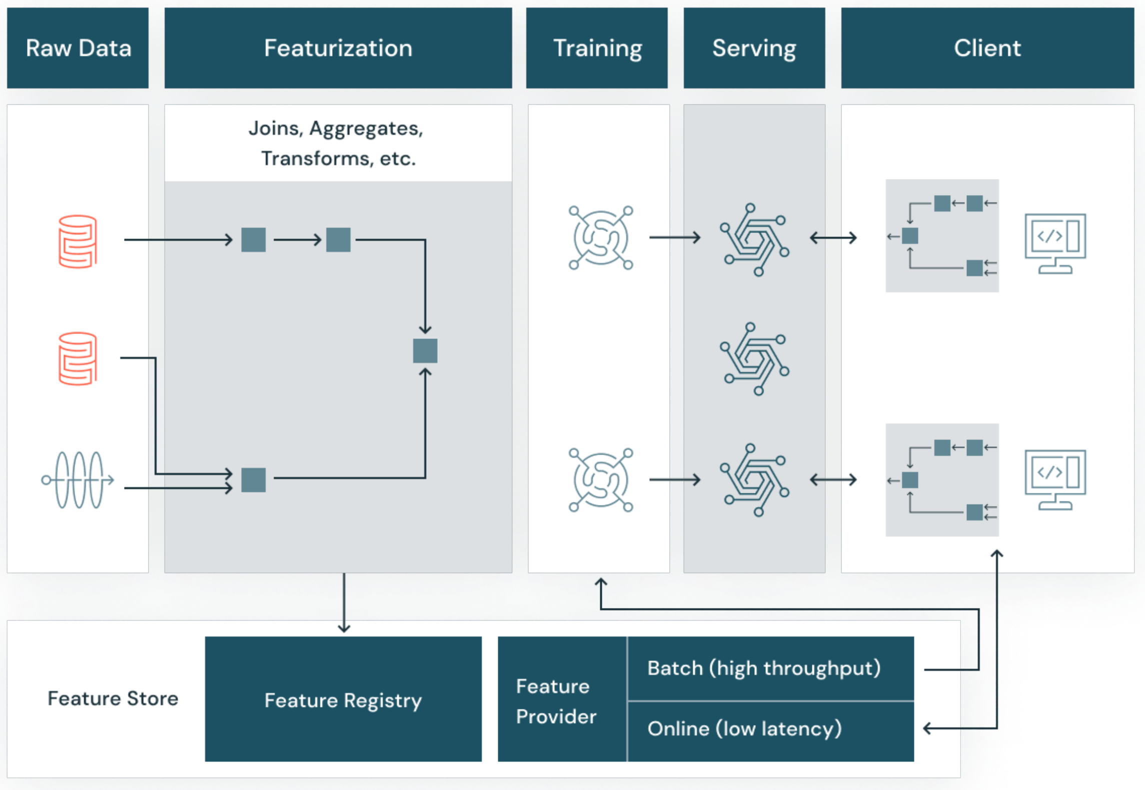
> **[With DLT pipelines] the team collaborates beautifully now, working together every day to divvy up the pipeline into their own stories and workloads.**
>
> — Dr. Chris Inkpen, Global Solutions Architect, Honeywell Energy and Environmental Solutions

Databricks tools like Delta Live Tables codify best practices for data engineer + data scientist collaboration into a simple data pipelining framework. With just a few lines of simple code, users in any part of the organization can easily spin off their own pipelines from a core pipeline like the one conceptualized in the diagram above.

Turning data into features

End-to-end machine learning pipelines need a smooth transition between data engineers — who ingest and transform raw data from your data stores into features — and data scientists, who use the features to build their training sets.

The Databricks Feature Store provides a searchable record of all features published by the data engineering team. Data scientists can search for features in batch mode for higher throughput or online mode for serving models at low latency. Features used in model training are automatically tracked with the model and, during model inference, the model itself retrieves them directly from the feature store. Why is that important?



> “[With DLT pipelines] the time required to define and develop a streaming pipeline has gone from days to hours.
>
> — Yue Zhang, Staff Software Engineer, Block

First, it enables data scientists to easily find and share features, even features that another team developed. If implementations are done in different code environments, it can lead to errors and delays, without a centralized repository to ensure that the same code used to develop the features is used for model training and inference.

When you create a feature table in Databricks, the data sources used to create the feature table are saved and accessible. For each feature in a feature table, you can also access the models, notebooks, jobs and endpoints that use the feature. So, when you use features from Feature Store, the model is packaged with feature metadata to automatically retrieve or join features to score new data, making model deployment and updates much easier.

Prototype to production on a single platform

Databricks Mosaic AI provides unified tooling to build, deploy and monitor AI and ML solutions — from building predictive models to the latest GenAI and large language models. It is completely integrated with the rest of the Data Intelligence Platform. This enables better collaboration between data scientists and data engineers to move models from prototype to production quickly and ensures data lineage and compliance with strong security/access controls, centralized governance and auditability of ML-related assets.

It is crucial in the age of generative AI to have observability on model quality, data quality and system reliability in a unified platform to alert, diagnose and debug issues easily and automatically redeploy models to maintain business and production SLAs. To do that, you need to build your own LLMs and keep your data private.

By bringing together your own data, AI models, LLM operations (LLMOps), monitoring and governance on the Databricks Data Intelligence Platform, your data science team can accelerate the production of their generative AI models with:

- Flexibility to use existing models or to train their model using your data
- Monitoring, evaluating and logging their model and prompt performance, and securely serving models, features and functions in real time
- Built-in capabilities for the entire AI lifecycle and underlying monitoring and governance

New features that will help the data science team more easily implement generative AI use cases include:

- Vector Search, a database that is optimized to store and retrieve embeddings, which are mathematical representations of the semantic content of data, typically text or image data
- LLM-optimized Model Serving, including support for GPU serving endpoints, provisioned throughput and per-token models, as well as external models
- MLflow enhancements such as a prompt engineering UI, evaluation tools, native flavors for LLMs and system metrics
- Lakehouse Monitoring
- Model sharing with MLflow, Unity Catalog and Model Serving integrations

Complete control

Data is what makes a Generative AI model intelligent, and the domain-specific enterprise data that you already have is the most valuable for training differentiated models specific to your use cases. Using an off-the-shelf model that everyone else is using does not lead to differentiation and without ownership, organizations cannot address any issue related to the model or how it was trained.

When considering a strategy for deploying generative AI, you'll want to decide how important ownership, control, compliance and security are to your business. You can use your unique enterprise data to augment, fine-tune or pretrain models, or you can augment an open source model or SaaS model through retrieval augmented generation (RAG).

Either way, you own both the model and the data with the Databricks approach.

Production quality

Your data science team relies on your data engineers for clean, curated data to deploy production-quality generative AI applications that are accurate, safe and governed. Your solution should enable rich tools for understanding the quality of their data and model outputs, along with an underlying solution to optimize the entire GenAI process (such as data preparation, retrieval models, language models, ranking and post-processing pipelines, prompt engineering and training models on custom enterprise data). Your solution should also allow for faster deployment across multiple use cases to let the data team orchestrate reliable production workflows to move multiple models and use cases easily from POCs into production in a standardized and governed manner.

Databricks handles the management of all aspects of the ML lifecycle on a single platform, from data ingestion and featurization to model building, tuning and production. LLMOps practices can be automated to reduce the time and complexity needed to build and deploy a model. Built-in data management and governance capabilities provide monitoring of models around model quality, hallucination, toxicity, etc., and end-to-end lineage and auditability from data to production model.

Lower cost

The unified Databricks Data Intelligence Platform empowers data engineers to simplify and efficiently manage and govern all data types and help the data science team productionalize their own GenAI models cost-effectively using their own data. Databricks' automated capabilities streamline platform administration and reduce infrastructure and storage costs. A Nucleus Research report found Databricks customers achieved 482% ROI over three years (on average), with a payback period as short as four months. This accelerated their time to value for big data projects by 52%, improved data team productivity by almost 50%, and achieved up to 60% process improvements across data ingestion, ETL, BI and MLOps.

For data scientists, Mosaic AI has a proven track record of lowering costs by up to 10x. They can train multibillion parameter models in days, not weeks, and you don't have to spend millions and have a fleet of engineers to deliver your model. Databricks can also help lower cost by making purpose-built, smaller sized models available that can be augmented or fine-tuned with your data, delivering similar performance as larger foundation models at a fraction of the cost.

Data Sharing and Collaboration

Developing, deploying and managing AI models is a cross-functional effort, requiring project management and collaboration among data engineering, data science and business teams.

These teams have historically faced several collaboration challenges including:

- **Difficult handoffs** between teams due to disparate tools and processes (from data preparation to experimentation to operationalization)
- **Lack of access controls** for governance, auditability/traceability, and regulatory compliance, and managing risks of undesirable model outputs
- **Direct cloud storage sharing,** which can lead to compliance issues since it requires data replication and provides no centrally managed audit of data movement
- **Commercial data sharing platforms** which are expensive and inflexible and promote vendor lock-in

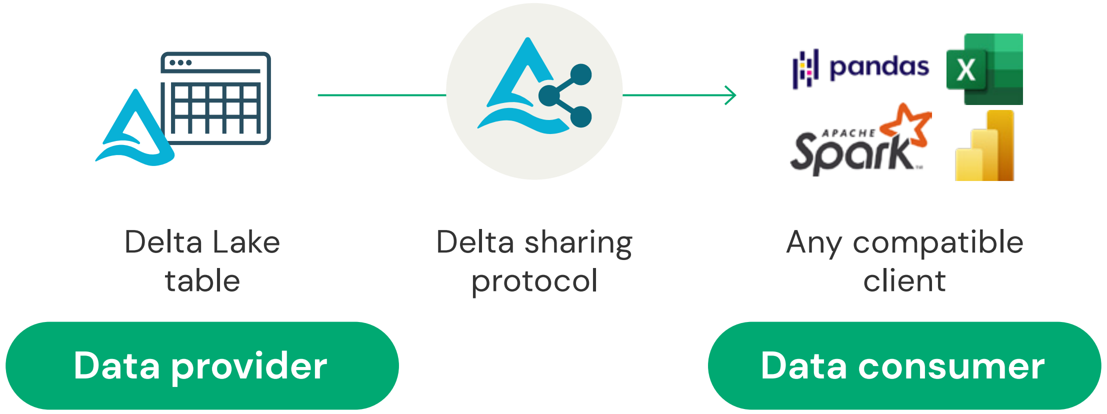
Some homegrown methods of data sharing such as SFTP and REST APIs are difficult to manage, maintain and scale — particularly for larger datasets since they require data replication.

Data sharing on the Data Intelligence Platform enables trusted collaboration for data providers and data consumers that drives business impact. Databricks is uniquely positioned to simplify data and AI model sharing and collaboration across the enterprise by providing one open and secure sharing platform for all your data, analytics and AI. Collaboration can encompass several use cases — sharing within your organization across different clouds or regions, or externally with partners and customers on any platform. With Databricks, you can accelerate the discovery, evaluation and usage of both internal and external data and AI products, simplify collaboration complexity and reduce costs. These products include datasets, AI models, volumes, dashboards and solutions.

Databricks also simplifies access management for these previously siloed data assets and supports access control of models, code, compute and credentials. These robust capabilities enable users to co-edit and co-view notebooks in their workspace in compliance with their organization's security policies.

Delta Sharing

For collaboration to work across different clouds, platforms and regions, there needs to be an open protocol so data recipients don't need to be on the same platform, the same cloud, or even on the cloud at all. Databricks and the Linux Foundation developed Delta Sharing to provide the first open source approach to data sharing across data, analytics and AI. With Delta Sharing, you can share live datasets, AI models, applications and notebooks across platforms, clouds, and regions without the need for replication or vendor lock-in.



> Delta Sharing makes it easy to securely share data with business units and subsidiaries without copying or replicating it, enabling us to share data without the recipient having an identity in our workspace.
>
> — Robert Hamlet, Lead Data Engineer, Enterprise Data Services, Cox Automotive

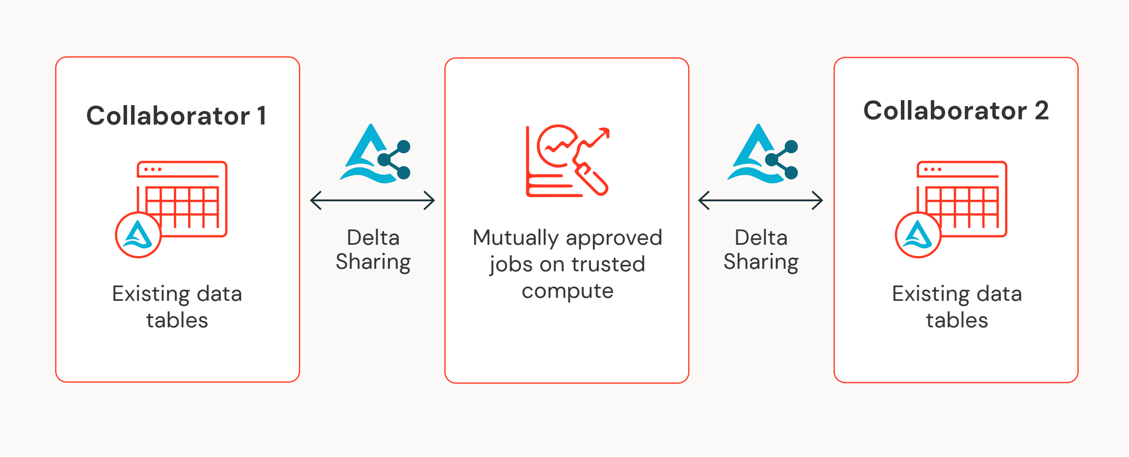
An open protocol for secure data sharing allows you to share and consume data from any platform or vendor that supports the Delta Sharing protocol. This helps put your data to work more quickly to discover insights faster. Delta Sharing is also integrated with Databricks Unity Catalog to securely govern, track and audit access to your shared datasets. You benefit from always consuming the latest version of data without the need for architectural parity and from reduced integration cost.

Delta Sharing also powers the Databricks Marketplace, an open marketplace for data, analytics and AI, such as AI models, notebooks, applications and dashboards — without proprietary platform dependencies, complicated ETL or expensive replication. It can simplify and accelerate your evaluation of external data and help put your data to work faster.

Secure collaboration while safeguarding data privacy

In cases where organizations need to securely share and consume data externally with customers, suppliers and partners within an isolated environment, Databricks Clean Rooms can help. With Databricks Clean Rooms, you can easily collaborate in a secure environment with your customers or partners on any cloud in a privacy-safe way. Also built on Delta Sharing with the principles of flexibility, scalability and interoperability, Clean Rooms allows for multicloud and multiregion support and participants can collaborate in a secure hosted environment without having to copy their data into a new location.

Clean Rooms provides a secure environment for data sharing



Clean Rooms allows for simple use cases such as joins, as well as complex computations such as machine learning of respective datasets, all while ensuring that no party has direct access or visibility to each other's data or proprietary packages or AI models. Your team can run computations in any language — SQL, R, Scala, Java, Python — and easily scale to multiple participants and reduce time to insights with predefined templates for common clean room use cases. Databricks Clean Rooms is currently in private preview.

Community bank tackles complex data environment with Delta Sharing

To scale its banking-as-a-service offering, Coastal Community Bank needed a modern, future-proof data platform to support its stringent customer data sharing and compliance requirements. "How do we harness incoming data from about 20 partners with technology environments that we don't control?" asks Barb MacLean, Coastal's Senior Vice President and Head of Technology Operations and Implementation. "The data's never going to be clean. We decided to make dealing with that complexity our strength and take on that burden. That's where we saw the true power of Databricks' capabilities. We couldn't have done this without the tools their platform gives us."

Reduce complexity for better collaboration

The Databricks Platform simplifies collaboration complexity and reduces costs with DatabricksIQ, a generative AI-powered data intelligence engine. It understands the semantics of your data and your business to empower nontechnical users to discover valuable insights on their own, enable AI assistants to streamline development, and to make queries faster by incorporating predictions in the Photon engine, optimizing autoscaling and minimizing cost through workload predictions, and simplifying maintenance by automating common configuration and tuning tasks.

For search use cases, DatabricksIQ employs a large language model (LLM) specifically tuned for enterprise data, drawing from example schemas across a variety of industries. Any employee in your organization can search, understand and query data in natural language. DatabricksIQ uses information about your data, usage patterns and org chart to understand your business's jargon and unique data environment.

Model Management

To trust your models and move them into production quickly, you want to see consistent, accurate outputs based on your data. One of the primary challenges for both data engineers and data scientists is the absence of a central repository to collaborate, share code and manage the model lifecycle.

With MLflow and the model registry capabilities now built into Unity Catalog, data teams can manage model deployment across execution environments, designate which versions are active for a given purpose via aliases, see the history of prior versions and trace back to the experiments, configuration, code and data that was used to create each model version.

Central hub for model management

When you have a central place to discover and share models, teams can more easily collaborate on moving them from experimentation to online testing and production, integrate them with approval and governance workflows and CI/CD pipelines and monitor ML deployments and their performance.

To achieve quality standards, your models must be reproducible. Managed MLflow is a Databricks service for managing the entire machine learning lifecycle. Based on the popular open source MLflow library, it simplifies experiment tracking and reproducibility, model management, and deployment. Tightly integrated with Unity Catalog, Managed MLflow helps securely share, manage and compare experiment results along with corresponding artifacts and code versions, as well as track lineage to upstream and downstream feature tables and model outputs.

An MLflow model is a standard format for packaging machine learning models that can be used in a variety of downstream tools. The format defines a convention that lets you save a model in different “flavors” that can be understood by different downstream tools.

Managed MLflow can track metrics, parameters and artifacts as part of experiments; package models and reproducible ML projects; and deploy models to power batch, streaming or real-time serving use cases.

Models in Unity Catalog

Databricks recommends using Models in Unity Catalog for improved governance, easy sharing across workspaces and environments, and more flexible MLOps workflows. You can create a registered model for each environment in your MLOps workflow, utilizing three-level namespaces and permissions of Unity Catalog to express governance. With Models in Unity Catalog, you reference model versions by named aliases, and you can create up to 10 custom references to model versions.

A centralized registry for models across an organization allows data science teams to:

- Discover registered models, experiment runs and associated code with a registered model
- Transition models to deployment stages
- Deploy different versions of a registered model in different aliases, allowing MLOps engineers to deploy and conduct testing of different model versions
- Control granular access and permission for models, model versions and model metadata, including aliases, using the same governance tools used for the rest of the platform: Unity Catalog

Unified deployment and governance for all AI models

Databricks Model Serving is a unified service for deploying, governing, querying and monitoring models fine-tuned or pre-deployed by Databricks like Llama 2, MosaicML MPT or BGE, or from any other model provider like Azure OpenAI, AWS Bedrock, AWS SageMaker and Anthropic.

A unified approach makes it easier to experiment with and productionize models from any cloud or provider to find the best candidate for your real-time application. You can do A/B testing of different models and monitor model quality on live production data once they are deployed. Model Serving also has pre-deployed models such as Llama2 70B, allowing you to jump-start developing AI applications like retrieval augmented generation (RAG) and provide pay-per-token access or pay-for-provisioned compute for throughput guarantees.

With an approach deeply grounded in open source, you can experiment with and switch models and techniques, taking full advantage of the latest innovations without additional effort or risk of lock-in. You can also easily retrain your models to take advantage of new data as it is generated.

JetBlue able to rapidly launch ML models

JetBlue depends on real-time data to make proactive decisions encompassing weather, flight crews, aircraft sensors and air traffic control. And the airline strategically depends on AI and ML to handle that volume of data. With the Databricks Data Intelligence Platform serving as the central hub for all streaming use cases, JetBlue efficiently delivers several ML and analytics products/insights by processing thousands of attributes in real time.

Using this architecture, JetBlue has sped AI and ML deployments across a wide range of use cases spanning four lines of business, each with its own AI and ML team. JetBlue's data engineers and data scientists integrate streaming pipelines, ML training using MLflow, ML API serving using ML registry and more in one cohesive platform. Using real time streams of data from weather, aircraft sensors, FAA data feeds, JetBlue operations and more, JetBlue built the world's first AI and ML operating system orchestrating a digital twin, known as BlueSky for efficient and safe operations.

Investing in Your Success With Education and Training

Build a strong data and AI foundation with Databricks training and certification to demonstrate your competence and accelerate your career.

Databricks Academy offers data and AI learning content accessible to all, from technical and business leaders to data practitioners, such as generative AI engineers to data scientists and machine learning engineers.

Databricks offers free and paid training content across skill levels and roles taught by expert instructors, including:

- Role-based learning pathways
- On-demand courses
- Instructor-led courses
- Certification
- Databricks Academy Labs
- Cohort-based learning (Blended Learning, Skills at Scale)

Assess your knowledge of the Databricks Data Intelligence Platform and the underlying methods required to successfully implement quality projects. Databricks Certification and Badging help you gain industry recognition, competitive differentiation, greater productivity and results, and a tangible measure of your educational investment.

Next Steps

Take the free on-demand Generative AI Fundamentals course — Build foundational knowledge of generative AI, including large language models (LLMs). There is also a full Generative AI certification pathway.

Explore the free on-demand Get Started With Data Engineering course — This course provides four short tutorial videos to help you learn data engineering on the Databricks Data Intelligence Platform.

Conclusion

The best GenAI models in the world will not succeed without good data.

This new age of generative AI requires a data-centric and collaborative approach where data engineering and data science teams work together on their own data in the same place at the same time to ensure the accuracy, quality and governance of their end-to-end LLM solutions.

That's why the Databricks Data Intelligence Platform is built around the core philosophy that good generative AI is inextricably linked to good data, rich tools that allow these teams to understand the quality of their data and model outputs, and an underlying platform that lets them optimize all aspects of the application.

A unified data platform is the best way for teams to efficiently and securely share and collaborate on all types of structured and unstructured data. It puts all aspects of your data engineering and generative AI lifecycles on the same platform, from data ingestion to production with data integration tools and automation to reduce the complexity and cost of building and deploying a model.

With a central place to discover and share models, teams can more easily collaborate on moving them from experimentation to online testing and production. And when they use their own unique data to build differentiated generative AI solutions, they're not only improving business outcomes with better performance, they're also reducing the associated risks and costs.

With the Databricks Data Intelligence Platform, teams own their generative AI models, securely trained or augmented with their own enterprise data. The unified platform provides automatic monitoring and lineage tracking of all models, features, and data to ensure system reliability, model quality and upstream data quality. The unified platform provides automatic monitoring and lineage tracking of all models, features and data to ensure system reliability, model quality and upstream data quality.

And best of all, the platform enables secure data collaboration and analysis on the same data to ensure data consistency and smooth transitions from data ingestion to development and deployment.

Generative AI on Databricks is democratizing access to artificial intelligence, sparking the beginnings of truly enterprise-wide AI.

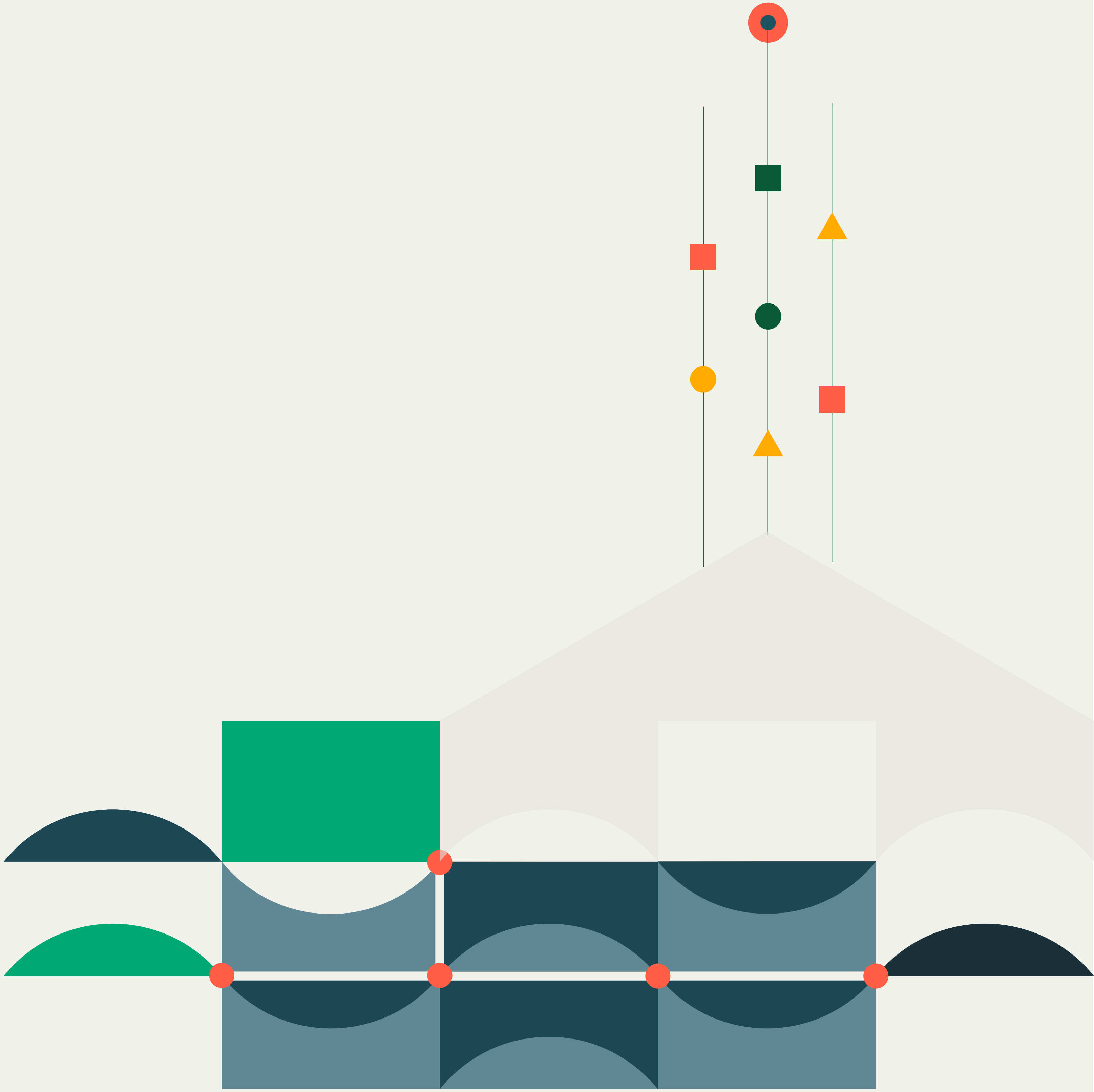
See how the Databricks Data Intelligence Platform can become the engine to accelerate productivity, innovation, and successful business outcomes with generative AI solutions.

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.

Start your free trial



databricks