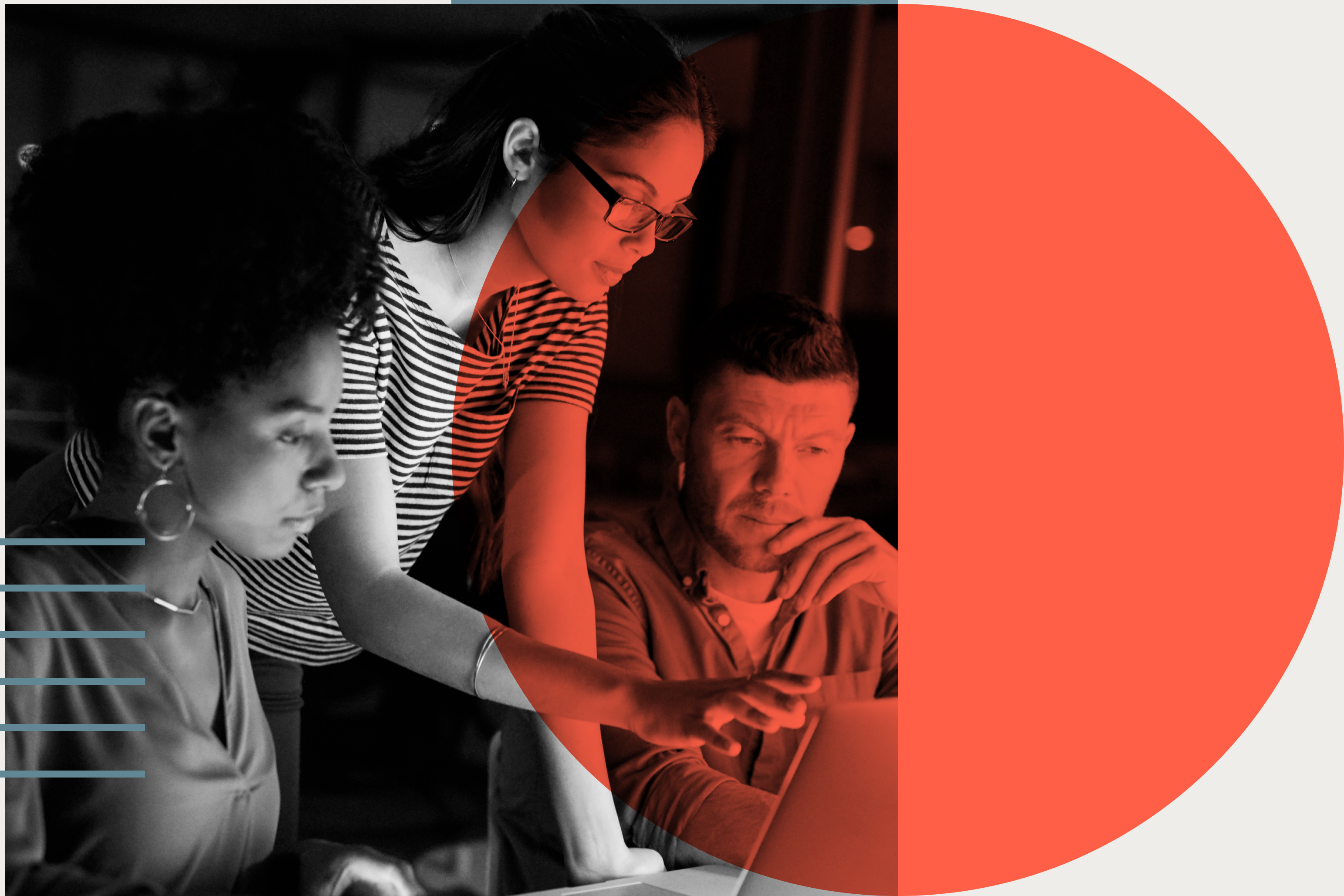
Executive Guide

Transform and Scale Your Organization With Data and AI

A guide for CIOs, CDOs, and data and AI executives



databricks

Contents

AUTHOR:
Chris D'Agostino
Global Field CTO
Databricks

EDITORS:
Manveer Sahota
Jessica Barbieri
Toby Balfre

CHAPTER 1:

Executive Summary

It's not surprising that many industry experts say data is the most valuable resource in the modern economy — some even go so far as to describe it as the "new oil." But at Databricks, we think of data as water. Its core compound never changes, and it can be transformed to whatever use case is desired, with the ability to get it back to its original form. Furthermore, just as water is essential to life, data is now essential to survival, competitive differentiation and innovation for every business. Clearly, the impact and importance of data are growing exponentially in both our professional and personal lives, while artificial intelligence (AI) is being infused in more of our daily digital interactions. The explosion in data availability over the last decade and the forecast for growth at a compounded annual growth rate (CAGR) of 23% over 2020–2025 — combined with low-cost cloud storage, compute, open source software and machine learning (ML) environments — have caused a major shift in how organizations leverage data and AI to improve data governance and the user experience, plus satisfy more AI/ML-based use cases to drive future growth.

Every organization is working to improve business outcomes while effectively managing a variety of risks — including economic, compliance, security and fraud, financial, reputational, operational and competitive risk. Your organization's data and the systems that process it play a critical role in not only enabling your financial goals but also in minimizing these seven key business risks.

Businesses have realized that their legacy information technology (IT) platforms are not able to scale and meet the increasing demands for better data analytics. As a result, they are looking to transform how their organizations use and process data. Successful data transformation initiatives for data, analytics and AI involve not only the design of hardware and software systems but also the alignment of people, processes and platforms. These initiatives always require a major financial investment and, therefore, need to yield a significant return on investment (ROI) — one that starts in months, not years.

To guide these initiatives, many organizations are adding the role of chief data officer (CDO) to their C-suite. Despite this structural change and focused resources, 87% of organizations still face many challenges to deliver on their data strategy — including how to deploy a modern data architecture, leverage data efficiently and securely, stay compliant with an ever-increasing set of regulations, hire the right talent, and identify and execute on AI opportunities.

Data and AI leaders are faced with the challenge of future-proofing their architecture and platform investments. The Lakehouse implementation from Databricks combines the best features of EDWs and data lakes by enabling all their workloads using open source and open standards — avoiding the vendor lock-in, black box design and proprietary data formats of other cloud vendors.

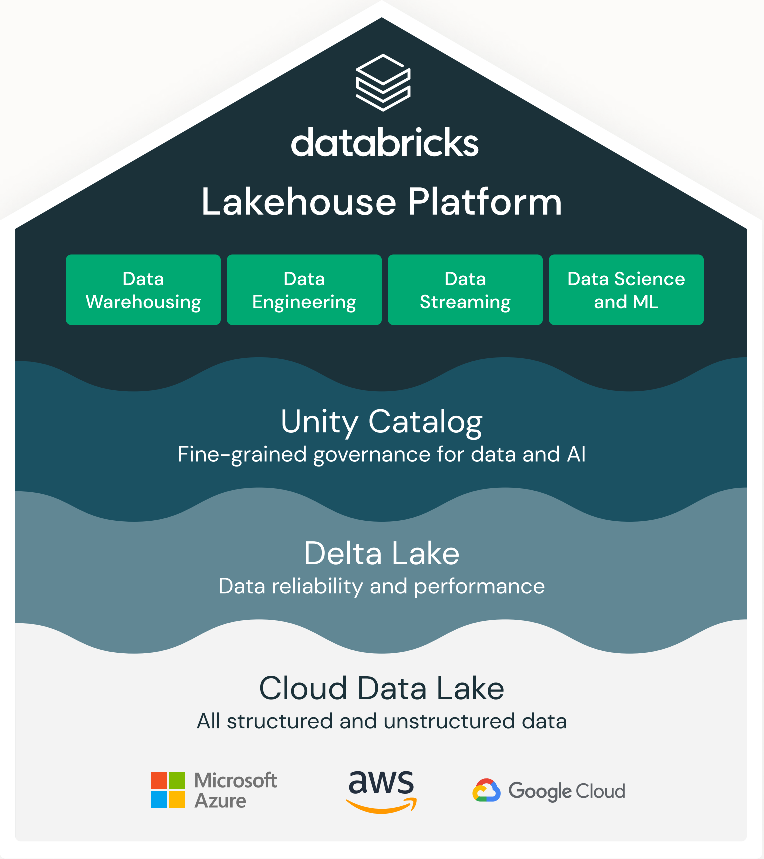


Figure 1:
The Databricks Lakehouse Platform

To successfully lead data and AI transformation initiatives, organizations need to develop and execute a comprehensive strategy that enables them to easily deploy a modern data architecture, unlock the full potential of all their data, and future-proof their investments to provide the greatest ROI. Today, organizations have the option of moving away from closed, proprietary systems offered by a variety of cloud vendors and adopting a strategy that emphasizes open, nonproprietary solutions built using industry standards.

At Databricks, we have helped over 7,000 companies achieve data, analytics and AI breakthroughs, and we've hired industry experts and thought leaders to help organizations better understand the steps involved in successful digital transformation initiatives. We are the first vendor to propose the data lakehouse architecture, which decouples data storage from compute while providing the best price/performance metrics for all your data workloads — including data warehousing. We have captured the lessons learned and summarized them in this series of Executive Guides — which are designed to serve as blueprints for CIOs, CDOs, CTOs and other data and technology executives to implement successful digital transformation initiatives for data, analytics and AI using a *modern data stack*. Databricks is the first company to deliver a unified data platform that realizes the data lakehouse architecture and enables the data personas in your organization to run their data, analytics and AI workloads in a simple, open and collaborative environment, as shown in Figure 1.

The lakehouse architecture benefits organizations in several ways:

1. It leverages low-cost cloud object stores to store ALL enterprise data.
2. It provides the ability to run different data workloads efficiently and in a cost-effective manner.
3. It uses open formats and standards that provide greater data portability — thus avoiding vendor lock-in.

Our intention is to present key considerations and equip you with the knowledge to ask informed questions, make the most critical decisions early in the process, and develop the comprehensive strategy that most organizations lack.

In addition, we have created an easy-to-follow Data and AI Maturity Model and provided a comprehensive professional services offering that organizations can leverage to measure their readiness, reskill their staff and track progress as they embark on their data transformation initiative.

CHAPTER 2:

Define the Strategy

The most critical step to enable data, analytics and AI at scale is to develop a comprehensive and executable strategy for how your organization will leverage people, processes and platforms to drive measurable business results against your corporate priorities. The strategy serves as a set of principles that every member of your organization can refer to when making decisions. The strategy should cover the roles and responsibilities of teams within your organization for how you capture, store, curate and process data to run your business — including the internal and external resources (labor and budget) needed to be successful.

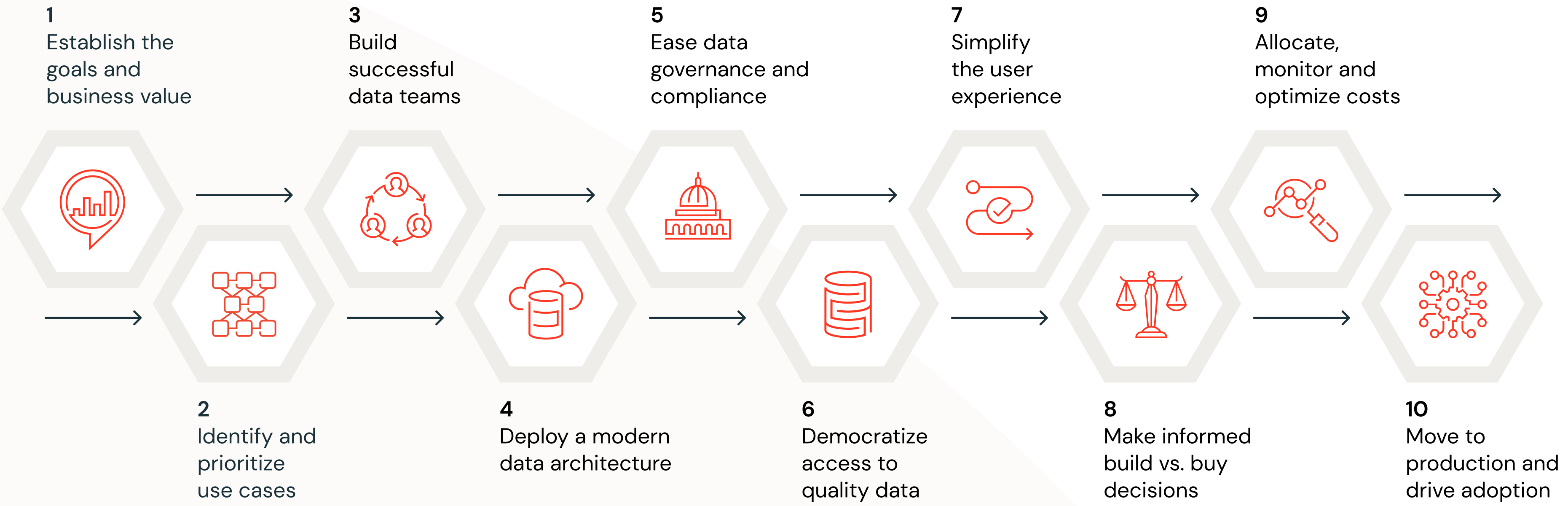

1
Establish the goals and business value
2
Identify and prioritize use cases
3
Build successful data teams
4
Deploy a modern data architecture
5
Ease data governance and compliance
6
Democratize access to quality data
7
Simplify the user experience
8
Make informed build vs. buy decisions
9
Allocate, monitor and optimize costs
10
Move to production and drive adoption

Figure 2:
The 10 steps to a winning data and AI strategy

This guide takes a stepwise approach to addressing each of these 10 topics.

Here are 10 key considerations

1. Secure buy-in and alignment on the overall business goals, timeline and appetite for the initiative.
2. Identify, evaluate and prioritize use cases that actually provide a significant ROI.
3. Create high-performing teams and empower your business analyst, data scientist, machine learning and data engineering talent.
4. Future-proof your technology investment with a modern data architecture.
5. Ensure you satisfy the European Union's General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA) and other emerging data compliance and governance regulations.
6. Implement needed policies, procedures and technology to guarantee data quality and enable secure data access and the sharing of all your data across the organization.
7. Streamline the user experience (UX), improve collaboration and simplify the complexity of your tooling.
8. Make informed build vs. buy decisions and ensure you are focusing your limited resources on the most important problems.
9. Establish the initial budgets and allocate and optimize costs based on SLAs and usage patterns.
10. Codify best practices for moving into production and how to measure progress, rate of adoption and user satisfaction.

The strategy should clearly answer these 10 topics and more, and should be captured in a living document, owned and governed by the CDO and made available for everyone in the organization to review and provide feedback on. The strategy will evolve based on the changing market/political conditions, evolving business, the technology landscape or a combination of any of these — but it should serve as the North Star for how you will navigate the many decisions and trade-offs that you will need to make over the course of the transformation.

Studies have shown that data scientists spend 80% of their time collecting and compiling data sets and only 20% of their time developing insights and algorithms. Organizations that are able to invert these numbers benefit in two ways — happier employees and improved time to market for use cases. These employers create more favorable working environments and lower the risk of burnout and the resulting regrettable attrition.

1. Establish the goals and business value

Most organizations on a data, analytics and AI journey establish a set of goals for the resulting investment. The goals generally fall into one of three categories:

1. **Business outcomes**
2. **People**
3. **Technology**

In terms of business outcomes, organizations need to adapt more quickly to market opportunities and emerging risks, and their legacy-based information systems make that difficult to achieve. As a result, business leaders see the digital transformation as an opportunity to build a new technology foundation from which to run their business and increase business value. One that is more agile, scalable, secure and easier to use — making the organization better positioned to adapt, innovate and thrive in the modern and dynamic economy.

For organizations today, people are one of their most valuable assets — you cannot succeed in data, analytics and AI without them. The battle for top talent is as fierce as ever, and the way that people work impacts your ability to hire and retain the skills you need to succeed. It is important to make sure that employees work in a frictionless data environment, to the extent possible, so they feel productive each day and can do their best work.

Finally, from a technology perspective, organizations have grown tired of the high costs associated with complex system architectures, vendor lock-in, and proprietary solutions that are slow to evolve. The industry trend is to move away from large capital expenditures (capex) to pay for network and server capacity in advance — and toward a "just-in-time" and "pay-for-what-you-use" operating expense (opex) approach. Your data analytics environment should support this trend as well — using open standards, low-cost storage and on-demand compute that efficiently spins up to perform data workloads and spins down once they are complete.

Executive buy-in and support

Large organizations are difficult to change — but it's not impossible. In order to be successful, you need to have unwavering buy-in and support from the highest levels of management — including the CEO and board of directors. With this support, you have the leverage you need to develop the strategy, decide on an architecture and implement a solution that can truly change the way your business is run. Without it, you have a very expensive science project that has little hope of succeeding. Why? Because the majority of people in your organization are busy doing their day jobs. The added work to support the initiative must be offset by a clear articulation of the resulting benefits — not only for the business but for the personnel within it. The transformation should result in a positive change to how people do their jobs on a daily basis.

Transformation for data, analytics and AI needs to be a company-wide initiative that has the support from all the leaders. Even if the approach is to enable data and AI one business unit (BU) at a time, the plan needs to be something that is fully embraced in order to succeed. Ideally, the senior-most executives serve as vocal proponents.

Evolve to an AI-first company — not just a data-first company

Data and AI transformations should truly transform the way organizations use data, not just evolve it. For decades, businesses have operated using traditional business processes and leveraged Structured Query Language (SQL) and business intelligence (BI) tools to query, manipulate and report on a subset of their data. There are five major challenges with this approach:

1. A true self-assessment of where your organization is on the AI maturity curve. Most organizations will use pockets of success with analytics and AI to move higher up the maturity curve, but in reality the ability to replicate and scale the results is nearly impossible.

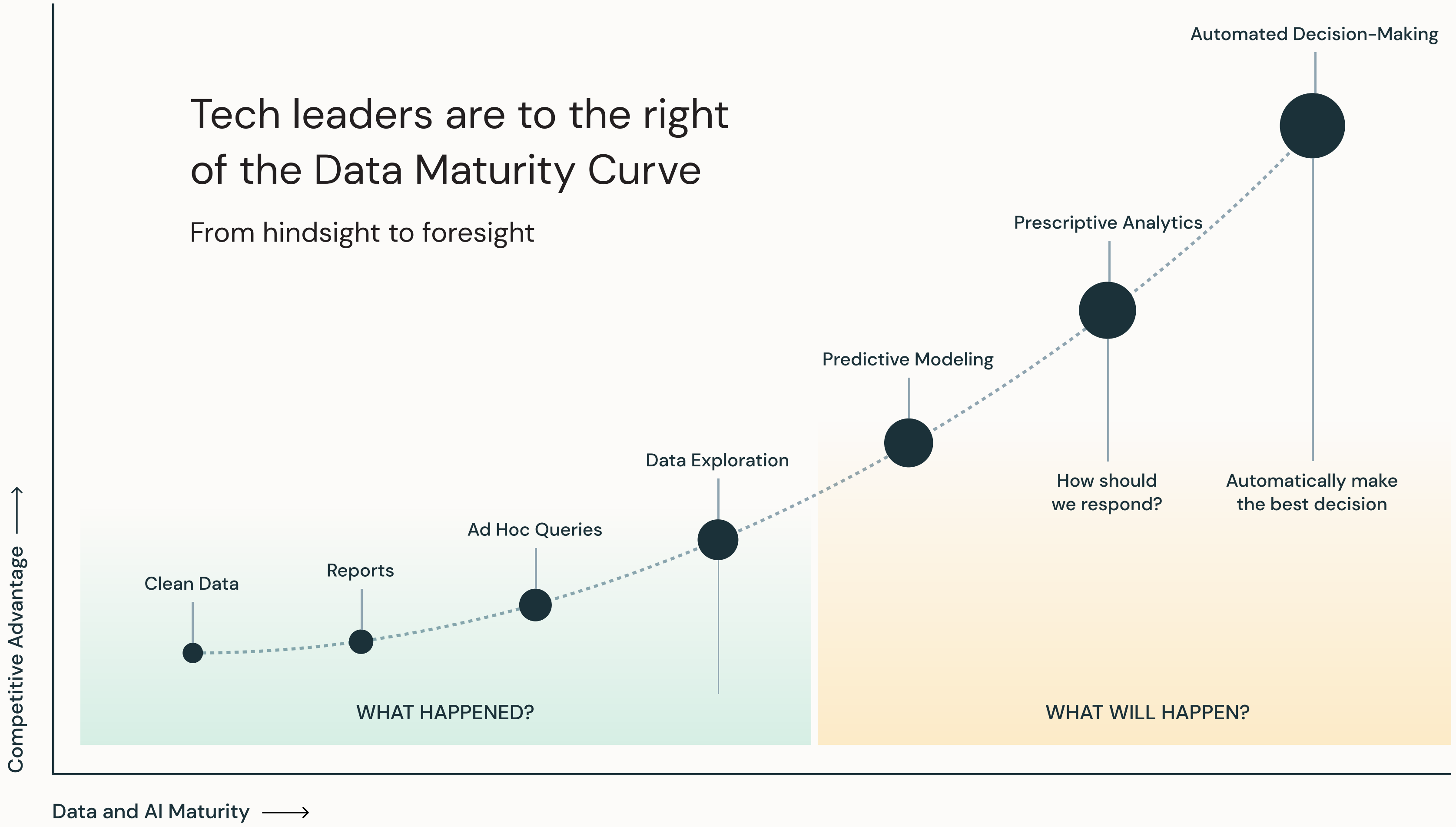


Figure 3:
The Data Maturity Curve

2. Data volumes and types have outgrown even the most modern approaches to SQL-based data processing.
3. These large data volumes also make it nearly impossible for your workforce to continue to programmatically state, in a priority manner, how data insights can be achieved or how the business should react to changing data.
4. Organizations need to reduce the costs of processing all this data. You simply cannot afford to hire the number of people needed to respond to every piece of data flowing into your environment. Machines scale, people do not.
5. Advances in machine learning and AI have simplified the steps and reduced the expertise needed to gain game-changing insights. For these reasons, plus many others, the organizations that thrive in the 21st century will do so based on their ability to leverage all the data at their disposal. Traditional ways of processing and managing data will not work. Using ML and AI will empower your workforce to leverage data to make better decisions for managing risk, helping your organization succeed in the modern economy.

Go "all in" on the cloud

The COVID-19 pandemic has caused rapid adoption of cloud-based solutions for collaboration and videoconferencing — and organizations are now using this time to reevaluate their use of on-premises and cloud-based services. The cloud vendors provide many benefits to organizations, including Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) solutions. These benefits, especially when combined with the use of open source software (OSS), increase the speed at which organizations can use the latest technologies while also reducing their capex in these budget-conscious times.

For AWS, Microsoft, Google and other cloud providers, the game is about data acquisition. The more corporate data that resides in a specific cloud, the more sticky the customer is to the vendor. At the same time, multicloud support is both a selling point and an aspirational goal for many organizations. Companies are well aware of vendor lock-in and want to abstract their applications so they can be moved across clouds if there is a compelling business reason.

"We are on an amazing journey. Being among the fastest-growing enterprise software cloud companies on record was unimaginable when we started Databricks. To get here, we've stayed focused on the three big bets we made when founding the company — cloud, open source and machine learning. Fast-forward seven years, thousands of data teams around the globe are working better together on Databricks."

Ali Ghodsi
Co-founder and CEO
Databricks

Approaching your technology choices with a multicloud point of view gives the organization more sovereignty over the data — flexibility to run workloads anywhere, ease of integration when acquiring businesses that run on different cloud providers and simplified compliance with emerging regulations that may require companies to be multicloud — as part of a mandate to reduce risk to the consumer's personal information.

As a result, data portability and the ability to run workloads on different cloud providers are becoming increasingly important.

Modernize business applications

As organizations begin to accelerate the adoption of the cloud, they should avoid a simple "lift and shift" approach. The majority of on-premises applications are not built with the cloud in mind. They usually differ in the way that they handle security, resiliency, scalability and failover. Their application designs often store data in ways that make it difficult to adhere to regulatory requirements such as the GDPR and CCPA standards. Finally, the features and capabilities of the application may be monolithic in nature and, therefore, tightly coupled. In contrast, modern cloud applications are modular in design and use RESTful web services and APIs to easily provide access to an application's functionality.

Cloud-based architectures, commodity databases and software application development frameworks make it easier for developers to build scalable, secure end-to-end applications to run all your internal business processes. Building n-tiered applications (e.g., mobile and web-based applications with RESTful APIs and a backing database) has become straightforward with the latest tooling available to your application development teams.

As a first step, organizations should inventory their business-critical applications, prioritize them based on business impact and modernize them in a consistent manner for cloud-based deployments. It is these applications that generate and store a significant amount of the data consumed within an organization. Using a consistent approach to cloud-based application design makes it easier to extract data when it is needed.

The next step is to identify which applications are viewed as the system of record (SOR) for a given data set. A good architectural principle is to only allow data sets to be stored inside their declared SOR and not allow other applications within your environment to store copies of the data — unless absolutely necessary for performance reasons. In this case, it is best to "cache" the data for use in the non-SOR application and sync the data from the actual SOR.

Data from these SORs should be made available in three ways:

1. Expose a set of RESTful APIs for applications to invoke at any given time.
2. Ensure that copies of the data land in the data lake.
3. Change data capture (CDC) and other business events should be streamed in real time for immediate consumption by downstream applications.

Move toward real-time decisioning

The value of data should be viewed through two different lenses. The first is to view data in the aggregate, and the second is to view data as an individual event. This so-called "time value of data" is an important concept in the world of data, analytics and AI. To be effective, you need to be able to leverage both — on the same data platform.

On the one hand, data in aggregate becomes more valuable over time — as you collect more of it. The aggregate data provides the ability to look back in time and see the complete history of an aspect of your business and to discover trends. Real-time data is most valuable the moment it is captured. In contrast, a newly created or arriving data event gives you the opportunity to make decisions — in the moment — that can positively affect your ability to reduce risk, better service your customers or lower your operating costs. The goal is to act immediately — with reliability and accuracy — upon the arrival of a new streaming event. This "time value of data" is shown in Figure 4 on the next page.

The Databricks Lakehouse Platform allows you to combine real-time streaming and batch processing using one architecture and a consistent set of programming APIs.

For example, real-time processing of clickstream data from your customer-facing mobile application can indicate when the customer is having trouble and may need to call into your call center. This insight gives you the opportunity to interject with a digital assistant or to pass on "just-in-time" information to your call center agents — improving the customer experience and lowering customer churn.

Data, analytics and AI rely on the "time value of data" — a powerful concept that allows you to train your machine learning models using historical data and provides you with the ability to make real-time decisions as new events take place. For example, credit card fraud models can use deep historical data about a given customer's buying patterns (location, day of week, time of day, retailer, average purchase amount, etc.) to build rich models that are then executed for each new credit card transaction. This real-time execution, combined with historical data, enables the best possible customer experience.

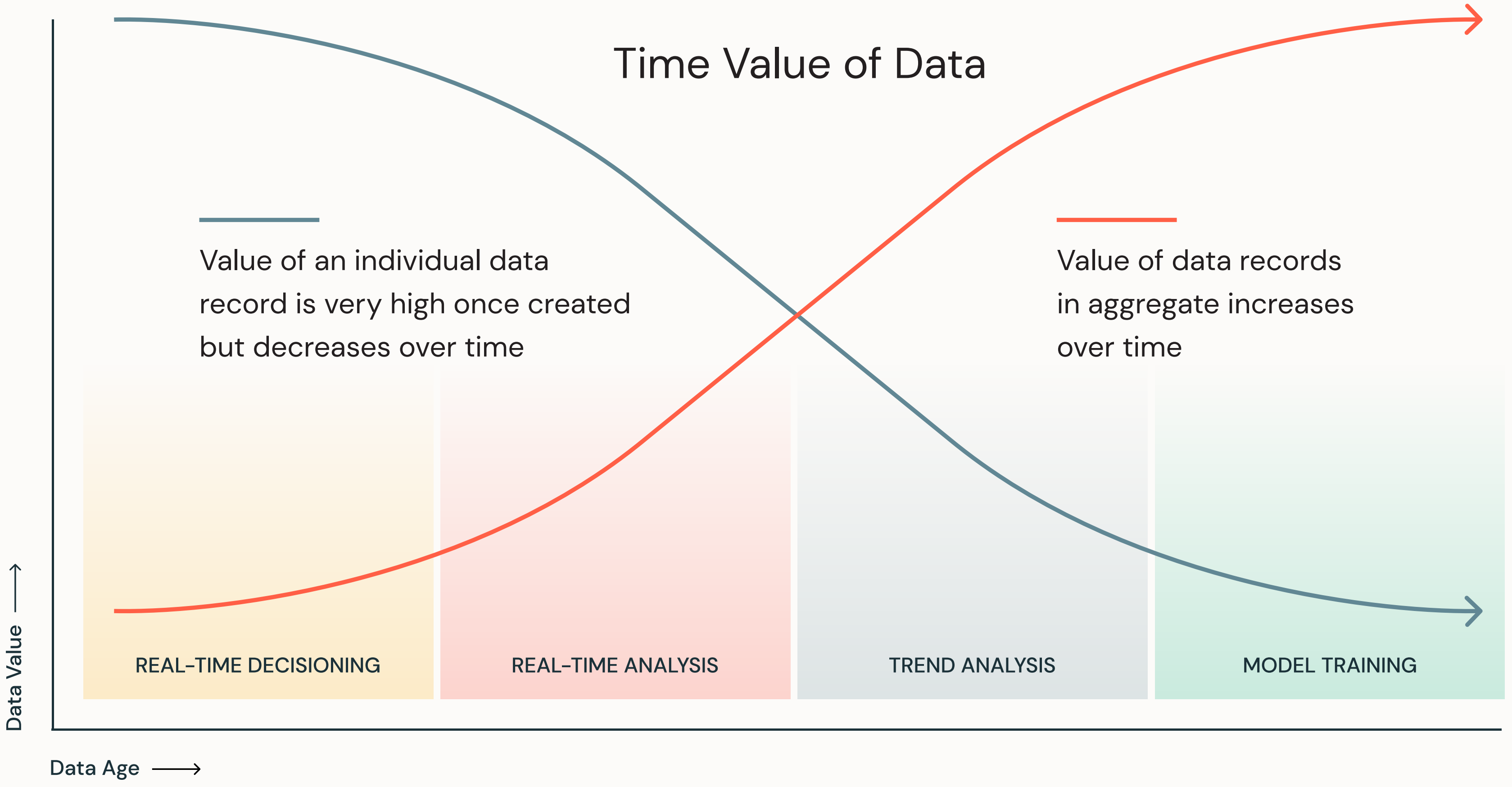


Figure 4:
Time Value of Data

"We believe that the data lakehouse architecture presents an opportunity comparable to the one we saw during early years of the data warehouse market. The unique ability of the lakehouse to manage data in an open environment, blend all varieties of data from all parts of the enterprise and combine the data science focus of the data lake with the end-user analytics of the data warehouse will unlock incredible value for organizations."

Bill Inmon
The father of the data warehouse

Land *all* data in a data lake

In order to effectively drive data, analytics and AI adoption, relevant data needs to be made available to the user as quickly as possible. Data is often siloed in various business applications and is hard and/or slow to access. Likewise, organizations can no longer afford to wait for data to be loaded into data stores like a data warehouse, with predefined schemas that are designed to allow you to ask very specific questions about that data only. What do you do when you want to ask a different question? To further complicate matters, how do you handle new data sets that cannot easily be manipulated to fit into your predefined data stores? How do you find new insights as quickly as possible?

The overall goal is to gain insights from the data as quickly as possible — which can happen at any step along the data pipeline — including raw, refined and curated data states.

This phenomenon has led to the concept known as the four Vs of data — specifically, *volume*, *velocity*, *variety* and *veracity*. Data-, analytics- and AI-driven organizations need to be able to store and process all their data, regardless of size, shape or speed. In addition, data lineage and provenance are critical to knowing whether or not you can trust the data.

Change the way people work

When done correctly, organizations get value from data, analytics and AI in three ways — infrastructure savings, productivity gains and business-impacting use cases. Productivity gains require a true focus on minimizing the number of steps needed to produce results with data. This can be accomplished by:

1. Making data more accessible and ensuring it can be trusted
2. Minimizing the number of tools/systems needed to perform work
3. Creating a flywheel effect by leveraging the work of others

With Databricks, users can collaborate and perform their work more efficiently, regardless of their persona or role. The user experience is designed to support the workloads of data analysts, SQL developers, data engineers, data scientists and machine learning professionals.

In large organizations, it's understandable why application and data silos are prevalent. Each business unit is laser-focused on achieving their goals, and the use of information technology is viewed as an enabler. Systems and applications get built over time to satisfy specific needs within a line of business. As a result, it's not surprising to learn that employees must jump through a large number of hoops to get access to the data they need to do their jobs. It should be as simple as getting your identity and PC.

A primary goal of your data and AI transformation should be to focus on improving the user experience — in other words, improving how your entire organization interacts with data. Data must be easily discoverable with default access to users based on their role(s) — with a simple process to compliantly request access to data sets that are currently restricted. The tooling you make available should satisfy the principal needs of the various personas — data engineers, data scientists, machine learning engineers, business analysts, etc. Finally, the results of the work performed by a user or system upstream should be made available to users and systems downstream as "data assets" that can drive business value.

Organizations that maximize the productivity of their workforce and enable employees to do their best work under optimal conditions are the ones that have the greatest chance to recruit and retain top talent.

Minimize time in the "seam"

As you begin your data transformation, it is important to know that the longer it takes, the more risk and cost you introduce into your organization. The stepwise approach to migrating your existing data ecosystem to a modern data stack will require you to operate in two environments simultaneously, the old and the new, for some period of time. This will have a series of momentary adverse effects on your business:

- It will increase your operational costs substantially, as you will run two sets of infrastructure
- It will increase your data governance risk, since you will have multiple copies of your data sitting in two very different ecosystems

- It increases the cyberattack footprint and vectors, as the platforms will likely have very different security models and cyber defenses
- It will cause strain on your IT workforce due to the challenges of running multiple environments
- It will require precise communications to ensure that your business partners know which environment to use and for what data workloads

To mitigate some of the strain on the IT workforce, some organizations hire staff augmentation firms to "keep the lights on" for the legacy systems while the new systems are being implemented and rolled out. It's important to remember this is a critical but short-lived experience for business continuity.

Shut down legacy platforms

In keeping with the goal of minimizing time in the seam, the project plan and timeline must include the steps and sequencing for shutting down legacy platforms. For example, many companies migrate their on-premises Apache Hadoop data lake to a cloud-based object store. The approach for shutting down the on-premises Hadoop system is generally as follows:

1. Identify the stakeholders (business and IT) who own the jobs that run in the Hadoop environment.
2. Declare that no changes can be made to the Hadoop environment — with the exception of emergency fixes or absolutely critical new business use cases.
3. Inventory the data flow paths that feed data into the Hadoop environment.
4. Identify the source systems that feed the data.
5. Inventory the data that is currently stored in the Hadoop environment and understand the rate of change.
6. Inventory the software processes (aka jobs) that handle the data and understand the output of the jobs.
7. Determine the downstream consumers of the output from the jobs.

8. Prioritize the jobs to move to the modern data architecture.
9. One by one, port the data input, job execution, job output and downstream consumers to the new architecture.
10. Run legacy and new jobs in parallel for a set amount of time — in order to validate that things are working smoothly.
11. Shut down the legacy data feeds, job execution and consumption. Wait. Look for smoke.
12. Rinse and repeat — until all jobs are migrated.
13. Shut down the Hadoop cluster.

A similar model can also be applied to legacy on-premises enterprise data warehouses.

You can follow the same process for other legacy systems in your environment. Some of these systems may be more complex and require the participation of more stakeholders to identify the fastest way to rationalize the data and processes. It is important, however, to make sure that the organization has the fortitude to hold the line when there is pressure to make changes to the legacy environments or extend their lifespan. Setting firm dates for when these legacy systems will be retired will serve as a forcing function for teams when they onboard to the new modern data architecture. Having the executive buy-in from page 9 plays a crucial role in seeing the shutdown of legacy platforms through.

2. Identify and prioritize use cases

An important next step in enabling data, analytics and AI to transform your business is to identify use cases that drive business value — while prioritizing the ones that are achievable under the current conditions (people, processes, data and infrastructure). There are typically hundreds of use cases within an organization that could benefit from better data and AI — but not all use cases are of equal importance or feasibility. Leaders require a systematic approach for identifying, evaluating, prioritizing and implementing use cases.

Establish the list of potential use cases

The first step is to ideate by bringing together various stakeholders from across the organization and understand the overall business drivers — especially those that are monitored by the CEO and board of directors. The second step is to identify use case opportunities in collaboration with business stakeholders, and understand the business processes and the data required to implement the use case. After steps one and two, the next step is to prioritize these cases by calculating the expected ROI. To avoid this becoming a pet project within the data/IT teams, it's important to have a line of business champion at the executive level.

There needs to be a balance between use cases that are complex and ones that are considered low-hanging fruit. For example, determining if a web visitor is an existing or net new customer requires a fairly straightforward algorithm that uses web browser cookie data and the correlation of the devices used by a given individual or household. However, developing a sophisticated credit card fraud model that takes into account geospatial, temporal, merchant and customer-purchasing behavior requires a broader set of data to perform the analytics.

In terms of performance, thought should be given to the speed at which the use case must execute. In general, the greater the performance, the higher the cost. Therefore, it's worth considering grouping use cases into three categories:

1. Sub-second response
2. Multi-second response
3. Multi-minute response

Being pragmatic about the true service level agreement (SLA) will save time and money by avoiding over-engineering the design and infrastructure.

Thinking in terms of "data assets"

Machine learning algorithms require data — data that is readily available, of high quality and relevant — to perform the experiments, train the models, and then execute the model when it is deployed to production. The quality and veracity of the data used to perform these machine learning steps are key to deploying models into production that produce a tangible ROI.

It is critical to understand what steps are needed in order to make the data available for a given use case. One point to consider is to prioritize use cases that make use of similar or adjacent data. If your engineering teams need to perform work to make data available for one use case, then look for opportunities to have the engineers do incremental work in order to surface data for adjacent use cases.

Mature data and AI companies embrace the concept of "data assets" or "data products" to indicate the importance of adopting a design strategy and data asset roadmap for the organization. Taking this approach helps stakeholders avoid fit-for-purpose data sets that drive only a single use case — and raise the level of thinking to focus on data assets that can fuel many more business functions. The "data asset" roadmap helps data source owners understand the priority and complexity of the data assets that need to be created. Using this approach, data becomes part of the fabric of the company, evolves the culture, and influences the design of business applications and other systems within the organization.

Determine the highest impact/priority

As shown in Figure 5, organizations can evaluate a given use case using a scorecard approach that takes into account three factors: strategic importance, feasibility and tangible ROI. Strategic importance measures whether or not the use case helps meet immediate corporate goals and has the potential to drive growth or reduce risk. Feasibility measures whether or not the organization has the data and IT infrastructure, plus the data science talent readily available, to implement the use case. The ROI score indicates whether or not the organization can easily measure the impact to the P/L.

🟥 = Scored by business stakeholders 🟦 = Scored by technology stakeholders		SCORING GUIDELINES (RELATIVE SCORING)		
		1 = LOW SCORE, DO LATER	5 = AVERAGE, NICE TO HAVE	10 = HIGH, MUST HAVE
Strategic Importance Score How important is it to business success?	Business Alignment	Not required for any corporate goals	Not required for immediate corporate goals	Required for immediate corporate goals
	Business Driver	Does not drive growth/profitability (P&L) or competitiveness	Could drive some growth/profitability (P&L)	Significantly drives growth/profitability (P&L) and competitiveness
	IT Foundation	No BI/IT dependencies	BI/IT best practice	BI/IT foundational element
Feasibility Score What is the current data and AI readiness?	Data Access and Trust Adjusting Based on Availability	Low awareness of available data (internal and external) or the problems it can solve	Some ingestion and exploration of large-scale data is possible	Large-scale data is available for exploration in the cloud
	Delivery (Data Engineers, Data Scientists, Data Analysts)	Limited in-house resources	Hiring plan for data science and engineering resources, few available in-house	Scaled data science, engineering, cloud and deployment organization
	Architecture	Current thinking on architecture resembles on-prem traditional data warehousing solution with batch processes rather than a data lakehouse approach	Architecture has been built and tested, some use cases are underway with multiple data sources now available in the cloud	The platform is utilized at scale across the business and is able to evolve to meet the demands of new business lines and services driven by data
ROI Score How tangible and large is the ROI?	ROI Potential	Mostly productivity gains, “soft intangible benefits”	Some P&L impact, not easily tangible	Significant P&L impact, “hard measured benefits”

Figure 5:
Methodology for scoring use cases

Ensure business and technology leadership alignment

Prioritizing use cases requires striking a balance between offensive- and defensive-oriented use cases. It is important for executives to evaluate use cases in terms of opportunity growth (offensive) and risk reduction (defensive). For example, data governance and compliance use cases should take priority over offensive-oriented use cases when the cost of a data breach or noncompliance is higher than the acquisition of a new customer.

The Databricks Professional Services team can help customers identify revenue-generating and cost-saving opportunities for data and AI use cases that provide a significant ROI when adopting the lakehouse architecture.

3. Build successful data teams

In order to succeed with data, analytics and AI, companies must find and organize the right talent into high-performing teams — ones that can execute against a well-defined strategy with the proper tools, processes, training and leadership. Digital transformations require executive-level support and are likely to fail without it — especially in large organizations.

However, it's not enough to simply hire the best data and AI talent — the organization must want to succeed, at an enterprise level. In other words, they must also evolve their company culture into one that embraces data, data literacy, collaboration, experimentation and agile principles. We define these companies as "data native."

Chief information officers and chief data officers — two sides of the data coin

Data native companies generally have a single, accountable executive who is responsible for areas such as data science, business analytics, data strategy, data governance and data management. The data management aspects include registering data sets in a data catalog, tracing data lineage as data sets flow through the environment, performing data quality checks and scanning for sensitive data in the clear.

Many organizations are rapidly adding the chief data officer (CDO) role to their executive ranks in order to oversee and manage these responsibilities. The CDO works closely with CIOs and other business stakeholders to establish the overall project plan, design and implementation — and to align project management, product management, business analysis, data engineering, data scientist and machine learning talent.

The CDO and CIO will need to build a broad coalition of support from stakeholders who are incentivized to make the transformation a success and help drive organization-wide adoption. To do this, the stakeholders must understand the benefits of — and their role and responsibilities in — supporting the initiative.

There are two organizational constructs that are found in most successful data native companies. The first is the creation of an *AI/ML center of excellence* (COE) that is designed to establish in-house expertise around ML and AI, and which is then used to educate the rest of the organization on best practices. The second is the formation of a *data and AI transformation steering committee* that will oversee and guide decisions and priorities for the transformative data, analytics and AI initiatives, plus help remove obstacles.

Furthermore, CDOs need to bring their CIOs along early in the journey.

Creating an AI/ML COE

Data science is a fast-evolving discipline with an ever-growing set of frameworks and algorithms to enable everything from statistical analysis to supervised learning to deep learning using neural networks. While it is difficult to establish specific and exact boundaries between the various disciplines, for the purposes of this document, we use "data science" as an umbrella term to cover machine learning and artificial intelligence. However, the general distinction is that data science is used to produce insights, machine learning is used to produce predictions, and artificial intelligence is used to produce actions. In contrast, while a data scientist is expected to forecast the future based on past patterns, data analysts extract meaningful insights from various data sources. A data scientist creates questions, while a data analyst finds answers to the existing set of questions.

Organizations wanting to build a data science competency should consider hiring talent into a centralized organization, or COE, for the purposes of establishing the tools, techniques and processes for performing data science. The COE works with the rest of the organization to educate and promote the appropriate use of data science for various use cases.

A common approach is to have the COE report into the CDO, but still have data scientists dotted line into the business units or department. Using this approach, you achieve two goals:

- The data scientists are closer to the business stakeholders, have a better understanding of the data within a business unit and can help identify use cases that drive value
- Having the data scientists reporting into the CDO provides a structure that encourages collaboration and consistency in how work is performed among the cohort and brings that to the entire organization

Data and AI transformation steering committee

The purpose of the steering committee is to provide governance and guidance to the data transformation initiative. The CDO and CIO should co-chair the committee along with one business executive who can be a vocal advocate and help drive adoption. The level of executive engagement is critical to success of the initiative.

The steering committee should meet regularly with leaders from across the organization to hear status reports and resolve any conflicts and remove obstacles, if possible. The leaders should represent a broad group of stakeholders, including:

Program/project management: To report the status of progress for deploying the new data ecosystem and driving adoption through use cases

Business partners: To provide insight and feedback on how easy or difficult it is to drive adoption of the platform

Engineering: To report the status of the implementation and what technology trade-offs need to be made

Data science: To report on the progress made by the COE on educating the organization about use cases for ML and to report the status of various implementations

It is important to fully understand which environments and accounts your data is stored in. The goal is to minimize the number of copies of your data and to keep the data within your cloud account — and not the vendor's.

Make sure the architecture and security model for protecting data is well understood.

InfoSec: To review the overall security, including network, storage, application and data encryption and tokenization

Architecture: To oversee that the implementation adheres to architectural standards and guardrails

Risk, compliance and legal: To oversee the approach to data governance and ethics in ML

User experience: To serve as the voice of the end users who will perform their jobs using the new data ecosystem

Communication: To provide up-to-date communications to the organization about next steps and how to drive adoption

Partnering with architecture and InfoSec

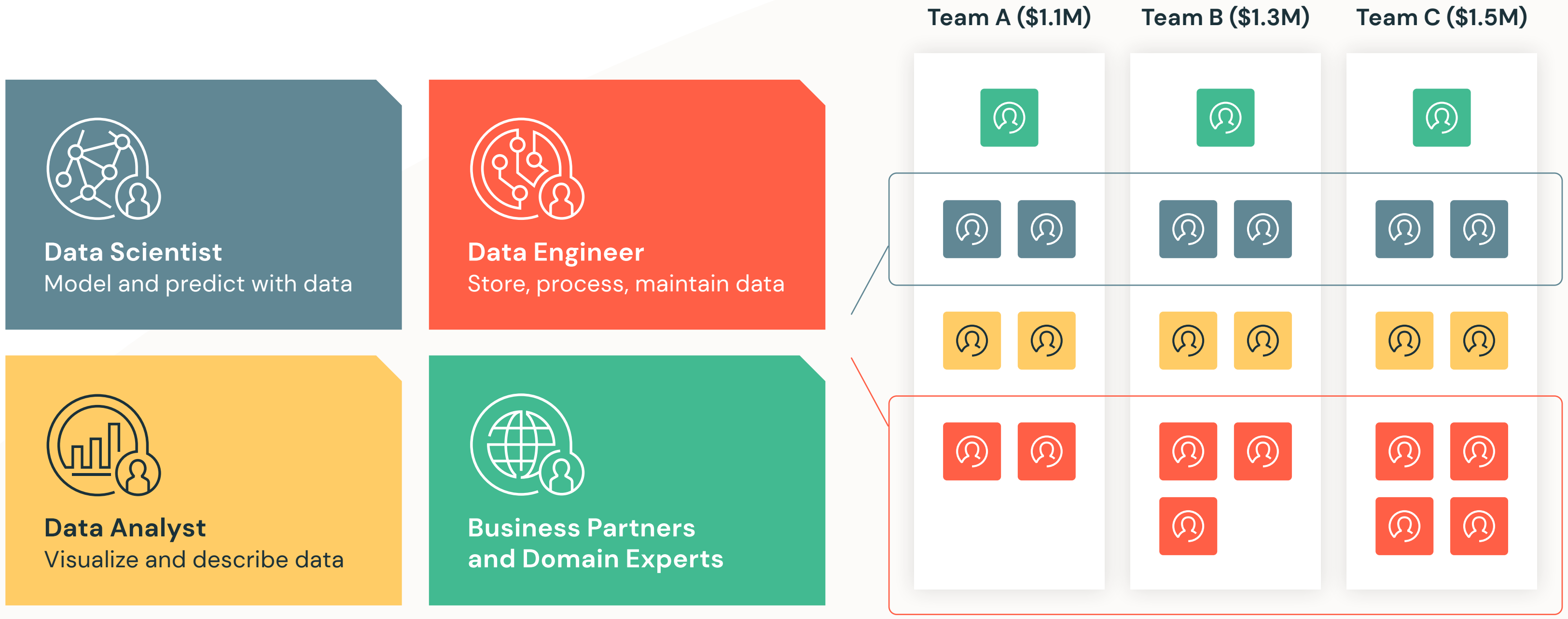
Early on, the CDO and CIO should engage the engineering and architecture community within the organization to ensure that everyone understands the technical implications of the overall strategy. This minimizes the chances that the engineering teams will build separate and competing data platforms. In regulated industries that require a named enterprise architect (EA), this will be a key relationship to foster. The EA is responsible for validating that the overall technology design and data management features support the performance and regulatory compliance requirements — specifically, whether the proposed design can meet the anticipated SLAs of the most demanding use cases and support the volume, velocity, variety and veracity (four Vs) of the data environment.

From an InfoSec perspective, the CDO must work to ensure that the proper controls and security are applied to the new data ecosystem and that the authentication, authorization and access control methods meet all the data governance requirements. An industry best practice is to enable self-service registration of data sets, by the data owner, and support the assignment of security groups or roles to help automate the access control process. This allows data sets to be accessible only to the personnel that belong to a given group. The group membership could be based primarily on job function or role within the organization. This approach provides fast onboarding of new employees, but caution should be taken not to proliferate too many access control groups — in other words, do not get too fine grained with group permissions, as they will become increasingly difficult to manage. A better strategy is to be more coarse-grained and use row- and column-level security sparingly.

Centralized vs. federated labor strategy

In most organizations today, managers work in silos, making decisions with the best intentions but focused on their own functional areas. The primary risk to the status quo is that there will be multiple competing and conflicting approaches to creating enterprise data and AI platforms. This duplication of effort will waste time and money and potentially erode the confidence and motivation of the various teams. While it certainly is beneficial to compare and contrast different approaches to implementing an architecture, the approaches should be strictly managed, with everyone designing for the same goals and requirements — as described in this strategy document and adhering to the architectural principles and best practices.

Even still, the roles of the CDO and CIO together should deliver a data analytics and AI platform with the least amount of complexity as possible, and one that can easily scale across the organization. It is very challenging to merge disparate data platform efforts into a single, cohesive design. It is best to get out in front of this wave of innovation and take input from the various teams to create a single, centralized platform. Having the data engineering teams centralized, reporting into a CIO, makes it easier to design a modern data stack — while ensuring that there is no duplication of effort when implementing the platform components. Figure 6 shows one possible structure.



🟦 Centralize data scientists under CDO — embed in lines of business for day-to-day tasking

🟥 Centralize data engineers under CIO/CTO — initially as an enterprise function

Figure 6:
Centralized teams with matrixed responsibilities

Hiring, training and upskilling your talent

While this guide does not cover recruiting strategies, it is important to note that data engineering and data science talent is very difficult to find in this competitive market. As a result, every organization should consider what training and upskilling opportunities exist for their current staff. A large number of online courses, at relatively low cost, teach the fundamentals of data science and AI. It will still be important to augment your existing staff with experienced data scientists and machine learning experts. You will then need to establish clear training paths, resources and timelines to upskill your talent.

Using the COE construct, it is easier to upskill a mix of data science talent by having the experts mentor the less experienced staff. The majority of Ph.D.-level talent comes from academia and has a vested interest in educating others. It's important to set up the structure and allow time in the schedule for knowledge transfer, experimentation and a safe environment in which to fail. A key aspect in accelerating the experience of your talent is to enable data science using production-like data and creating a collaborative environment for code sharing.

4. Deploy a modern data stack

The modern data architecture can most easily be described as the evolution of the enterprise data warehouse (EDW) from the 1980s and the Hadoop-style data lakes from the mid-2000s. The capabilities, limitations and lessons learned from working with these two legacy data architectures inspired the next generation of data architecture — what the industry now refers to as the lakehouse.

Figure 7 shows how the architectures have evolved as networking, storage, memory and CPU performance have improved over time.

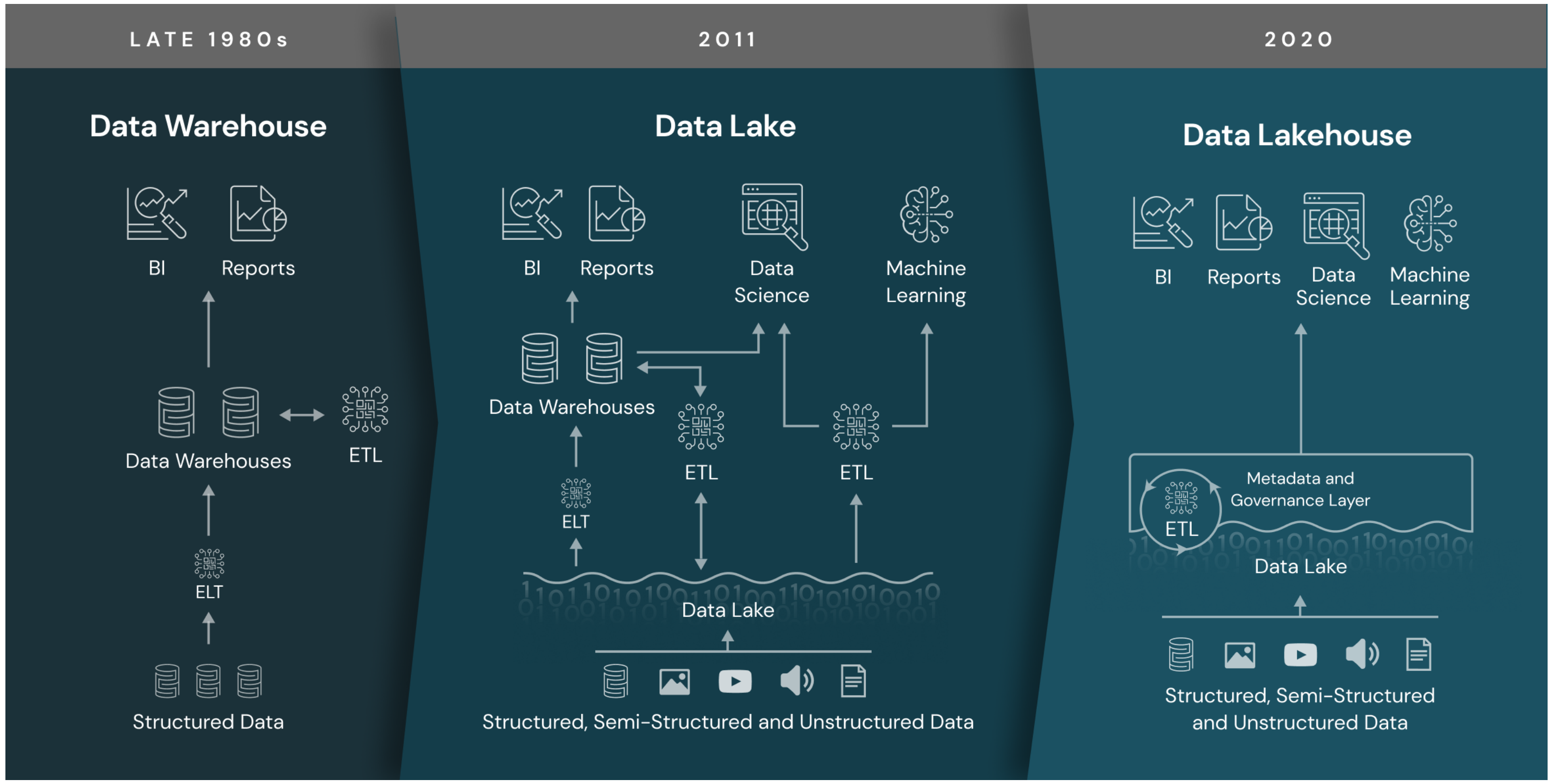


Figure 7:
A brief history of data architectures

The Databricks training, documentation and certification available to customers is industry-leading, and our Solution Accelerators provide exemplar code for organizations to hit the ground running with data and AI.

Evolving beyond the enterprise data warehouse and data lake

The EDW provided organizations with the ability to easily load structured and semi-structured data into well-organized tables — like rows and columns in a spreadsheet — and execute Structured Query Language (SQL) queries and generate business intelligence (BI) reports to measure the health and performance of the business. Though the EDW coupled storage and compute, it provided organizations with the ability to catalog data, apply robust security and audit, monitor costs and support a large number of simultaneous users — while still being performant. The EDW served its purpose for decades. However, most of the recent advances in AI have been in better models to process unstructured data (text, images, video, audio), but these are precisely the types of data that an EDW is not optimized for.

Therefore, in the mid-2000s, organizations wanted to take advantage of new data sets — *ones that contained unstructured data* — and apply new analytics — *ones that leveraged emerging data science algorithms*. In order to accomplish this, massive investments in on-premises data lakes occurred — most often leveraging Apache Hadoop and its distributed file system, known as HDFS, running on low-cost, commodity hardware. The Hadoop-style data lake provided the separation of compute from storage that organizations were seeking — thus eliminating the risk of vendor lock-in and opening the doors to a wide range of new analytics. Despite all these benefits, the architecture proved to be difficult to use, with a complex programming model known as MapReduce, and the performance fell short of the majority of real-time use cases.

Over time, Hadoop workloads were often migrated to Apache Spark™ workloads, which run 100x faster by processing data in-memory across a cluster — with the ability to massively scale. The Spark programming model was also simpler to use and provided a consistent set of application programming interfaces (APIs) for languages such as Python, SQL, R, Java and Scala. Spark was the first major step in separating compute from storage and providing the scale needed for distributed workloads.

A data lakehouse combines the best of data lakes and data warehouses, enabling BI and ML on all data on a simple, open and multicloud modern data stack.

Cloud-based data lakes

More than a decade ago, the cloud opened a new frontier for data storage. Cloud object stores like Amazon S3 and Azure Data Lake Storage (ADLS) have become some of the largest, most cost-effective storage systems in the world — which make them an attractive platform to serve as the next generation of data lakes. Object stores excel at massively parallel reads — an essential requirement for modern data warehouses.

However, data lakes lack some critical features: They do not support transactions, they do not enforce data quality, and their lack of consistency/isolation makes it almost impossible to mix appends and reads, and batch and streaming jobs. Also, performance is hampered by expensive metadata operations — for example, efficiently listing the millions of files (objects) that make up most large data lakes.

Lakehouse — the modern data architecture

What if it were possible to combine the best of both worlds? The performance, concurrency and data management of EDWs with the scalability, low cost and workload flexibility of the data lake. This is exactly the target architecture described by CDOs, CIOs and CTOs when asked how they would envision reducing the complexity of their current data ecosystems while enabling data and AI, at scale. The building blocks of this architecture are shown in Figure 8 and are what inspired the innovations that make the lakehouse architecture possible.

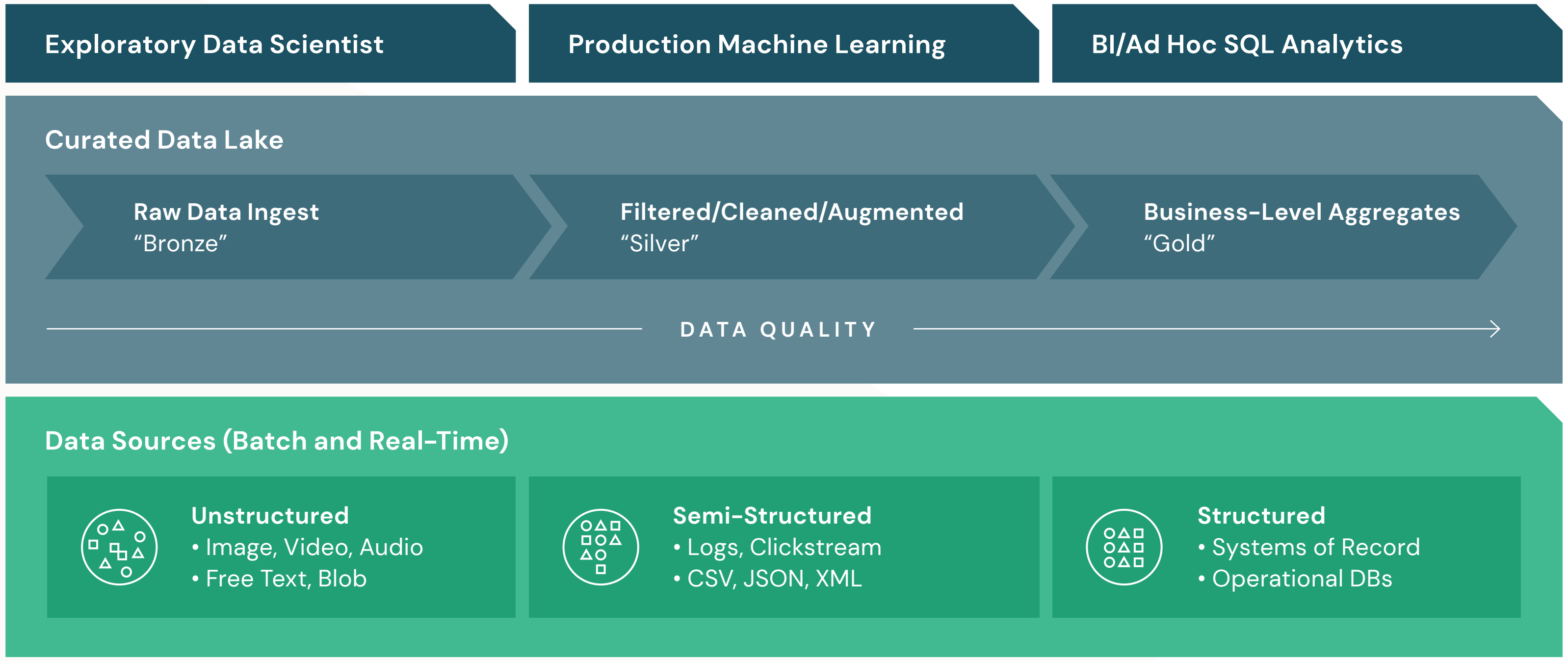


Figure 8:
The building blocks for a modern data architecture

The lakehouse architecture provides a flexible, high-performance design for diverse data applications, including real-time streaming, batch processing, data warehousing, data science and machine learning. This target-state architecture supports loading all the data types that might be interesting to an organization — structured, semi-structured and unstructured — and provides a single processing layer, using consistent APIs across programming languages, to curate data while applying rigorous data management techniques.

The move toward a single, consistent approach to data pipelining and refinement saves organizations time, money and duplication of effort. Data arrives in a landing zone and is then moved through a series of curation and refinement steps resulting in highly consumable and trusted data for downstream use cases. The architecture makes possible the efficient creation of "data assets" for the organization by taking a stepwise approach to improving data.

Databricks released Delta Lake to the open source community in 2019. Delta Lake provides all the data lifecycle management functions that are needed to make cloud-based object stores reliable and performant. This design allows clients to update multiple objects at once, replace a subset of the objects with another, etc., in a serializable manner that still achieves high parallel read/write performance from the objects — while offering advanced capabilities like time travel (e.g., query point-in-time snapshots or rollback of erroneous updates), automatic data layout optimization, upserts, caching and audit logs.

Lakehouse key features

To effectively migrate organizations to the lakehouse architecture, here's a list of key features that must be available for stakeholders to run business-critical production workloads:

- **Reliable data pipelines:** The lakehouse architecture simplifies the ETL development and management with declarative pipeline development, automatic data testing and deep visibility for monitoring and recovery.
- **Transaction support:** In an enterprise lakehouse, many data pipelines will often be reading and writing data concurrently. Support for ACID transactions ensures consistency as multiple parties concurrently read or write data, typically using SQL.
- **Schema enforcement and governance:** The lakehouse should have a way to support schema enforcement and evolution, supporting DW schema paradigms such as star/snowflake schemas. The system should be able to reason about data integrity, and it should have robust governance and auditing mechanisms.
- **Fine-grained governance for data and AI:** The first fine-grained, centralized security model for data lakes across clouds — based on the ANSI SQL open standards. The lakehouse enables organizations to unify data and AI assets by centrally sharing, auditing, securing and managing structured and unstructured data like tables, files, models and dashboards in concert with existing data, storage and catalogs.
- **Storage is decoupled from compute:** In practice this means storage and compute use separate clusters, thus these systems are able to scale to many more concurrent users and larger data sizes. Some modern data warehouses also have this property.
- **Openness:** The storage formats they use are open and standardized, such as Parquet, and they provide an API so a variety of tools and engines, including machine learning and Python/R libraries, can efficiently access the data directly.

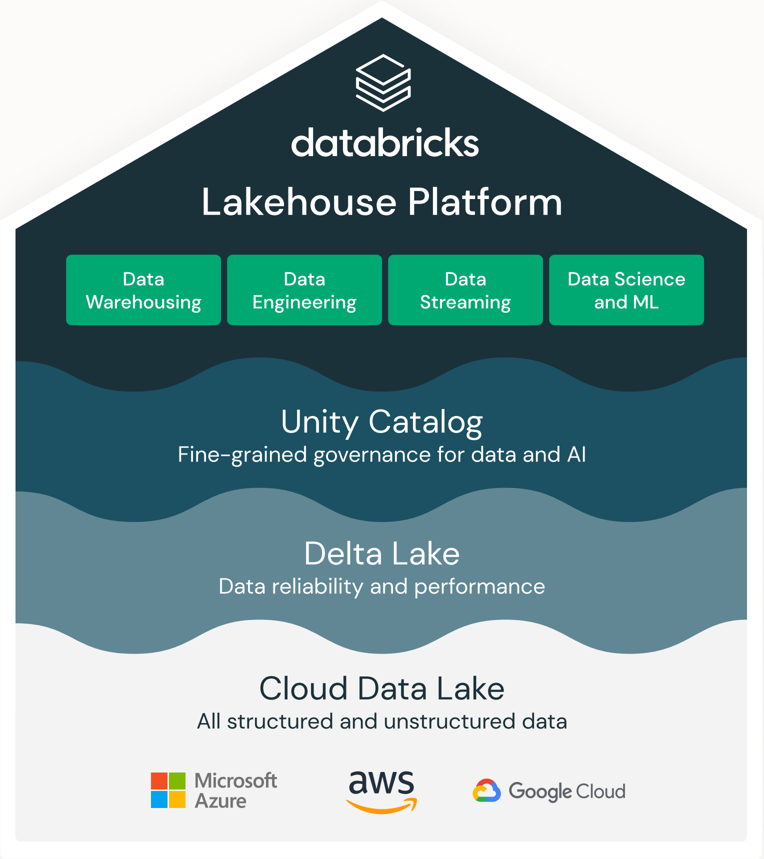


Figure 9:
Delta Lake is the open data storage layer that delivers reliability, security and performance on your data lake — for both streaming and batch operations

- **Support for diverse data types ranging from unstructured to structured data:** The lakehouse can be used to store, refine, analyze and access data types needed for many new data applications, including images, video, audio, semi-structured data and text.
- **Support for diverse workloads:** This includes data science, machine learning, SQL and analytics. Multiple tools might be needed to support all these workloads, but they all rely on the same data repository.
- **End-to-end streaming:** Real-time reports are the norm in many enterprises. Support for streaming eliminates the need for separate systems dedicated to serving real-time data applications.
- **BI support:** Lakehouses enable the use of BI tools directly on the source data. This reduces staleness, improves recency, reduces latency and lowers the cost of having to operationalize two copies of the data in both a data lake and a warehouse.
- **Multicloud:** The Databricks Lakehouse Platform offers you a consistent management, security and governance experience across all clouds. You don't need to invest in reinventing processes for every cloud platform that you're using to support your data and AI efforts. Instead, your data teams can simply focus on putting all your data to work to discover new insights and create business value.

Databricks is the only cloud-native vendor to be recognized as a Leader in both 2021 Magic Quadrant reports: **Cloud Database Management Systems** and **Data Science and Machine Learning Platforms**

These are the key attributes of lakehouses. Enterprise-grade systems require additional features. Tools for security and access control are basic requirements. Data governance capabilities, including auditing, retention and lineage, have become essential, particularly in light of recent privacy regulations. Tools that enable data discovery such as data catalogs and data usage metrics are also needed. With a lakehouse, such enterprise features only need to be implemented, tested and administered for a single system.

Databricks — innovation driving performance

Advanced analytics and machine learning on unstructured and large-scale data are two of the most strategic priorities for enterprises today — and the growth of unstructured data is going to increase exponentially — so it makes sense for CIOs and CDOs to think about positioning their data lake as the center of their data infrastructure. The main challenge is whether or not it can perform reliably and fast enough to meet the SLAs of the various workloads — especially SQL-based analytics.

Databricks has focused its engineering efforts on incorporating a wide range of industry-leading software and hardware improvements in order to implement the first lakehouse solution. Our approach capitalizes on the computing advances of the Apache Spark framework and the latest networking, storage and CPU technologies to provide the performance customers need to simplify their architecture. These innovations combine to provide a single architecture that can store and process all the data sets within an organization — supporting the range of analytics outlined above.

BI and SQL workloads

Perhaps the most significant challenge for the lakehouse architecture is the ability to support SQL queries for star/snowflake schemas in support of BI workloads. Part of the reason EDWs have remained a major part of the data ecosystem is because they provide low-latency, high-concurrency query support. In order to compete with the EDW, optimizations must be found within the lakehouse architecture that provide satisfactory query performance for the majority of BI workloads. Fortunately, advances in query plan, query execution, statistical analysis of files in the object store, and hardware and software improvements make it possible to deliver on this promise.

A word about the data mesh architecture

In 2019, another architectural concept, called the data mesh, was introduced. This architecture addresses what some designers identify as weaknesses of a centralized data lake. Namely, that you fill the data lake using a series of extract, transform, load (ETL) processes — which unnecessarily adds complexity. The data mesh approach avoids centralizing data in one location and encourages the source systems to create "data products" or "data assets" that are served up directly to consumers for data and AI workloads. The designers advocate for a federated approach to data and AI — while using enterprise policies to govern how source systems make data assets available.

There are several challenges with this approach. First, the data mesh assumes that each source system can dynamically scale to meet the demands of the consumers — particularly challenging when data assets become "hot spots" within the ecosystem. Second, centralized policies oftentimes leave the implementation details to the individual teams. This has the potential of inconsistent implementations, which may lead to performance degradations and differing cost profiles. Finally, the data mesh approach assumes that each source system team has the necessary skills, or can acquire them, to build robust data products.

The lakehouse architecture is not at odds with the data mesh philosophy — as ingesting higher-quality data from the source systems reduces the curation steps needed inside the data lake itself.

What are the basic building blocks of a sound data governance approach?

5. Improve data governance and compliance

Data governance is perhaps the most challenging aspect of data transformation initiatives. Every stakeholder recognizes the importance of making data readily available, of high quality and relevant to help drive business value. Likewise, organizations understand the risks of failing to get it right — the potential for undetected data breaches, negative impact on the brand and the potential for significant fines in regulated environments. However, organizations shouldn't perceive data governance or a defensive data strategy as a blocker or deterrent to business value. In fact, many organizations have leveraged their strong stance on data governance as a competitive differentiator to earn and maintain customer trust, ensure sound data and privacy practices, and protect their data assets

Why data governance fails

While most people agree that data governance is a set of principles, practices and tooling that helps manage the complete lifecycle of your data, what is often not discussed is what constitutes a pragmatic approach — one that balances realistic policies with automation and scalability.

Too often the policies developed around data governance define very strict data management principles — for example, the development of an enterprise-wide ontological model that all data must adhere to. Organizations can spend months, if not years, trying to define the perfect set of policies. The engineering effort to automate the enforcement of the new policies is not prioritized, or takes too long, due to the complexity of the requirements. Meanwhile, data continues to flow through the organization without a consistent approach to governance, and too much of the effort is done manually and fraught with human error.

A pragmatic approach to data governance

At a high level, organizations should enable the following data management capabilities:

- **Identify all sources of data**
  - Identify all data-producing and data-storing applications
  - Identify the systems of record (SOR) for each data set
  - Label data sets as internal or external (third party)
  - Identify where sensitive data is stored — GDPR/CCPA scope
  - Limit which operational data stores (ODSs) can re-store SOR data
- **Catalog data sets**
  - Register all data sets in a centralized data catalog
  - Create a lightweight, self-service data registration process
  - Limit manual entry as much as possible
  - Record the schema, if any, for the data set
  - Use an inference engine or tool to extract the data set schema
  - Add business and technical metadata to make it meaningful
  - Use machine learning to classify data sets
  - Use crowdsourcing to validate the machine-based results
- **Track data lineage**
  - Track data set flow and what systems act on data
  - Create an enumerated list of action values for specific operations
  - Emit lineage events via streaming layer and aggregate in data lake lineage event schema: <data setID, systemID, action, timestamp>
  - Optional: Add a source code repository URL for action traceability

By minimizing the number of copies of your data and moving to a single data processing layer where all your data governance controls can run together, you improve your chances of staying in compliance and detecting a data breach.

- **Perform data quality checks**
  - Create a rules library that is centrally managed and versioned
  - Update the rules library periodically with new rules
  - Use a combination of checks — null/not null, regex, valid values
  - Perform schema enforcement checks against data set registration
- **Scan for sensitive data**
  - Establish a tokenization strategy for sensitive data — GDPR/CCPA
  - Tokenize all sensitive data stored in the data lake — avoid cleartext
  - Use fixed-length tokens to preserve analytic value
  - Determine the approach for token lookup/resolution when needed
  - Ensure that any central token stores are secure with rotating keys
  - Identify which data elements from GDPR/CCPA to include in scans
  - Efficiently scan for sensitive data in cleartext using the rules library
- **Establish approved data flow patterns**
  - Determine pathways for data flow (source —> target)
  - Limit the ways to get SOR data (APIs, streaming, data lake, etc.)
  - Determine read/write patterns for the data lake
  - Strictly enforce data flow pathways to/from data lake
  - Detect violations and anomalies using lineage event analysis
  - Identify offending systems and shut down or grant exception
  - Record data flow exceptions and set a remediation deadline
- **Centralize data access controls**
  - Establish a common governance model for all data and AI assets
  - Centrally define access policies for all data and AI assets
  - Enable fine-grained access controls at row and column levels
  - Centrally enforce access policies across all workloads — BI, analytics, ML

- **Make data discovery easy**
  - Establish a data discovery model
  - Use manual or automatic data classification
  - Provide a visual interface for data discovery across your data estate
  - Simplify data discovery with rich keyword- or business glossary-based search
- **Centralize data access auditing**
  - Establish a framework or best practices for access auditing
  - Capture audit logs for all CRUD operations performed on data
  - Make auditing reports easily accessible to data stewards/admins for ensuring compliance

This is not intended to be an exhaustive list of features and requirements but rather a framework to evaluate your data governance approach. There will be violations at runtime, so it will be important to have procedures in place for how to handle these violations. In some cases, you may want to be very strict and shut down the data flow of the offending system. In other cases, you may want to quarantine the data until the offending system is fixed. Finally, some SLAs may require the data to flow regardless of a violation. In these cases, the receiving systems must have their own methodology for dealing with bad data.

Hidden cost of data governance

There are numerous examples of high-profile data breaches and failure to comply with consumer data protection legislation. You don't have to look very far to see reports of substantial fines levied against organizations that were not able to fully protect the data within their data ecosystem. As organizations produce and collect more and more data, it's important to remember that while storage is cheap, failing to enforce proper data governance is very, very expensive.

In order to catalog, lineage trace, quality check, and scan your data effectively, you will need a lot of compute power when you consider the massive amounts of data that exist in your organization. Each time you copy a piece of data to load it into another tool or platform, you need to determine what data governance techniques exist there and how you ensure that you truly know where all your data resides. Imagine the scenario where data flows through your environment and is loaded into multiple platforms using various ETL processes. How do you handle the situation when you discover that sensitive data is in cleartext? Without a consistent set of data governance tools, you may not be able to remediate the problem before it's flagged for violation.

Having a smaller attack surface and fewer ingress/egress routes helps guard your data and protect your organization's brand and balance sheet.

The bottom line is that the more complex your data ecosystem architecture is, the more difficult and costly it is to get data governance right.

6. Democratize access to quality data

Effective data and AI solutions rely more on the amount of quality data available than on the sophistication or complexity of the model or algorithm. Google published a paper titled "The Unreasonable Effectiveness of Data" demonstrating this point. The takeaway is that organizations should focus their efforts on making sure data scientists have access to the widest selection of relevant and high-quality data to perform their jobs — which is to create new opportunities for revenue growth, cost reduction and risk reduction.

The 80/20 data science dilemma

Most existing data environments have their data stored primarily in different operational data stores within a given business unit (BU) — creating several challenges:

- Most business units deploy use cases that are based only on their own data — without taking advantage of cross-BU opportunities
- The schemas are generally not well understood outside of BU or department — with only the database designers and power users being able to make efficient use of the data. This is referred to as the "tribal knowledge" phenomenon.
- The approval process and different system-level security models make it difficult and time-consuming for data scientists to gain the proper access to the data they need

In order to perform analysis, users are forced to log in to multiple systems to collect their data. This is most often done using single-node data science and generates unnecessary copies of data stored on local disk drives, various network shares or user-controlled cloud storage. In some cases, the data is copied to "user spaces" within production platform environments. This has the strong potential of degrading the overall performance for true production workloads.

To make matters worse, these copies of data are generally much smaller than the full-size data sets that would be needed in order to get the best model performance for your ML and AI workloads.

The Databricks Lakehouse Platform brings together all the data and AI personas into one environment and makes it easy to collaborate, share code and insights, and operate against the same view of data.

Small data sets reduce the effectiveness of exploration, experimentation, model development and model training — resulting in inaccurate models when deployed into production and used with full-size data sets.

As a result, data science teams are spending 80% of their time wrangling data sets and only 20% of their time performing analytic work — work that may need to be redone once they have access to the full-size data sets. This is a serious problem for organizations that want to remain competitive and generate game-changing results.

Another factor contributing to reduced productivity is the way in which end users are typically granted access to data. Security policies usually require both coarse-grained and fine-grained data protections. In other words, granting access at a data set level but limiting access to specific rows and columns (fine-grained) within the data set.

Rationalize data access roles

The most common approach to providing coarse-grained and fine-grained access is to use what's known as role-based access control (RBAC). Individual users log on to system-level accounts or via a single sign-on (SSO) authentication and access control solution.

Users can access data by being added to one or more Lightweight Directory Access Protocol (LDAP) groups. There are different strategies for identifying and creating these groups — but typically, they are done on a system-by-system basis, with a 1:1 mapping for each coarse- and fine-grained access control combination. This approach to data access usually produces a proliferation of user groups. It is not unusual to see several thousand discrete security groups for large organizations — despite having a much smaller number of defined job functions.

This approach creates one of the biggest security challenges in large organizations. When personnel leave the company, it is fairly straightforward to remove them from the various security groups. However, when personnel move around within the organization, their old security group assignments often remain intact and new ones are assigned based on their new job function. This leads to personnel continuing to have access to data that they no longer have a "need to know."

Data classification

Having all your data sets stored in a single, well-managed data lake gives you the ability to use partition strategies to segment your data based on "need to know." Some organizations create a partition based on which business unit owns the data and which one owns the data classification. For example, in a financial services company, credit card customers' data could be stored separately from that of debit card customers, and access to GDPR/CCPA-related fields could be handled using classification labels.

The simplest approach to data classification is to use three labels:

- **Public data:** Data that can be freely disclosed to the public. This would include your annual report, press releases, etc.
- **Internal data:** Data that has low security requirements but should not be shared with the public or competitors. This would include strategy briefings and market or customer segmentation research.
- **Restricted data:** Highly sensitive data regarding customers or internal business operations. Disclosure could negatively affect operations and put the organization at financial or legal risk. Restricted data requires the highest level of security protection.

Some organizations introduce additional labels, but care should be taken to make sure that everyone clearly understands how to apply them.

The data classification requirements should be clearly documented and mapped to any legal or regulatory requirements. For example, CCPA is so sweeping that it includes 11 categories of personal information — and defines "personal information" as "information that identifies, relates to, describes, is capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household."

Just examining one CCPA category, *Customer Records Information*, we see that the following information is to be protected: name, signature, social security number, physical characteristics or description, address, telephone number, passport number, driver's license or state identification card number, insurance policy number, education, employment, employment history, bank account number, credit or debit card number, other financial information, medical information, and health insurance information.

There are generally three different approaches in industry to performing data classification:

1. **Content-based:** Scans or inspects and interprets files to find sensitive information. This is generally done using regular expressions and lookup tables to map values to actual entities stored inside the organization (e.g., customer SSN).
2. **Context-based:** Evaluates the source of the data (e.g., application, location or creator) to determine the sensitivity of the data.
3. **User-based:** Relies on a manual, end-user selection of each data set or element and requires expert domain knowledge to ensure accuracy.

Taking all this into account, an organization could implement a streamlined set of roles for RBAC that uses the convention <domain><entity><data set | data asset><classification> where "domain" might be the business unit within an organization, "entity" is the noun that the role is valid for, "data set" or "data asset" is the ID, and "classification" is one of the three values (public, internal, restricted).

There is a "deny all default" policy that does not allow access to any data unless there is a corresponding role assignment. Wild cards can be used to grant access to eliminate the need to enumerate every combination.

Adopting the Databricks Lakehouse Platform allows you to add data sets into a well-managed data lake using low-cost object stores, and makes it easy to partition data based on domain, entity, data set and classification levels to provide fine-grained (row-level and column-level) security.

For example, <credit-card><customers><transactions> <restricted> gives a user or a system access to all the data fields that describe a credit card transaction for a customer, including the 16-digit credit card number. Whereas <credit-card><customers><transactions><role><membership> would allow the user or system access only to nonsensitive data regarding the transaction.

This gives organizations the chance to rationalize their security groups by using a domain naming convention to provide coarse-grained and fine-grained access without the need for creating tons of LDAP groups. It also dramatically eases the administration of granting access to data for a given user.

Everyone working from the same view of data

The modern data stack, when combined with a simplified security group approach and a robust data governance methodology, gives organizations an opportunity to rethink how data is accessed — and greatly improves time to market for their analytic use cases. All analytic workloads can now operate from a single, shared view of your data.

Combining this with a sensitive data tokenization strategy can make it straightforward to empower data scientists to do their job and shift the 80/20 ratio in their favor. It's now easier to work with full-size data sets that both obfuscate NPI/PII information and preserve analytic value.

Now, data discovery is easier because data sets have been registered in the catalog with full descriptions and business metadata — with some organizations going as far as showing realistic sample data for a particular data set. If a user does not have access to the underlying data files, having data in one physical location eases the burden of granting access, and then it's easier to deploy access-control policies and collect/analyze audit logs to monitor data usage and to look for bad actors.

Data security, validation and curation — in one place

The modern data architecture using Databricks Lakehouse makes it easy to take a consistent approach to protecting, validating and improving your organization's data. Data governance policies can be enforced using the built-in features of schema validation, expectations and pipelines — the three main steps to data curation. Databricks enables moving data through well-defined states: Raw —> Refined —> Curated or, as we refer to it at Databricks, Bronze —> Silver —> Gold.

The raw data is known as "Bronze-level" data and serves as the landing zone for all your important analytic data. Bronze data functions as the starting point for a series of curation steps that filter, clean and augment the data for use by downstream systems. The first major refinement results in data being stored in "Silver-level" tables within the data lake. These tables carry all the benefits of the Delta Lake product — for example, ACID transactions and time travel. The final step in the process is to produce business-level aggregates, or "Gold-level" tables, that combine data sets from across the organization. It's a set of data used to improve customer service across the full line of products, perform GDPR/CCPA reporting or look for opportunities to cross-sell to increase customer retention. For the first time, organizations can truly optimize data curation and ETL — eliminating unnecessary copies of data and the duplication of effort that often happens in ETL jobs with legacy data ecosystems. This "solve once, access many times" approach speeds time to market, improves the user experience and helps retain talent.

Extend the impact of your data with secure data sharing

Data sharing is crucial to drive business value in today's digital economy. More and more organizations are now looking to securely share trusted data with their partners/suppliers, internal lines of business or customers to drive collaboration, improve internal efficiency and generate new revenue streams with data monetization. Additionally, organizations are interested in leveraging external data to drive new product innovations and services.

Business executives must establish and promote a data sharing culture in their organizations to build competitive advantage.

7. Dramatically increase productivity of your workforce

Now that you have deployed a modern data stack and have landed all your analytical data in a well-managed data lake with a rationalized approach to access control, the next question is, “What tools should I provide to the user community so they can be most effective at using the new data ecosystem?”

Design thinking: working backward from the user experience

Design thinking is a human-centered approach to innovation — focused on understanding customer needs, rapid prototyping and generating creative ideas — that will transform the way you develop products, services, processes and organizations. Design thinking was introduced as a technique to not only improve but also bring joy to the way people work. The essence of design thinking is to determine what motivates people to do their job, where their current pain points are and what could be improved to make their jobs enjoyable.

Moving beyond best of breed

If you look across a large enterprise, you will find no shortage of database design, ETL, data cleansing, model training and model deployment tools. Many organizations take a “best of breed” approach in providing tooling for their end users. This typically occurs because leaders genuinely want to empower business units, departments and teams to select the tool that best suits their specific needs — so-called federated tool selection. Data science tooling, in particular, tends not to be procured at the “enterprise” level at first — given the high cost of rolling it out to the entire user population.

When tool selection becomes localized, there are a few things to consider:

- Tools are generally thought of as discrete components within an ecosystem and, therefore, interchangeable with criteria that are established within a specific tool category. The tool with the best overall score gets selected.
- The selection criteria for a tool usually contains a subjective list of "must-have" features based on personal preference or adoption within a department, or because a given tool is better suited to support a current business process
- Discrete tools tend to leapfrog one another and add features based on market demand rather quickly
- Evaluations that are performed over many months likely become outdated by the time the tool has moved into production
- The "enterprise" requirements are often limited to ensuring that the tool fits into the overall architecture and security environment but nothing more
- It's rare that the tools are evaluated in terms of simplifying the overall architecture, reducing the number of tools in play or streamlining the user experience
- The vendor backing the tool is evaluated in terms of risk, but not enough focus is spent on the partnership model, the ability to influence the roadmap and professional services support

For these reasons and more, it's worth considering an architecture and procurement strategy that centers on selecting a data platform that enables seamless integration with point solutions rather than a suite of discrete tools that require integration work and may no longer be category leaders over the long haul.

Databricks is a leading data and AI company — partly due to the innovations in the open source software that runs our platform — and as a result of listening to the needs of thousands of customers and having our engineers work side by side with customer teams to deliver real business value using data and AI.

Keep in mind that data platforms work well because the vendor took an opinionated point of view of how data processing, validation and curation should work. It's the integration between the discrete functions of the platform that saves time, conserves effort and improves the user experience. Many companies try to take on the integration of different technology stacks, which increases risk, cost and complexity. The consequences of not doing the integration properly can be serious — in terms of security, compliance, efficiency, cost, etc.

So, find a vendor that you can develop a true partnership with — one that is more likely to take feedback and incorporate your requirements into their platform product roadmap. This will require some give-and-take from both parties — sometimes calling for an organization to adjust their processes to better fit how the platform works. There are many instances where a given business process could be simplified or recast to work with the platform, as is. Sometimes it will require the vendor to add features that support your processes. The vendor will always be market driven and will want to build features in such a way that they apply to the broadest set of customers.

The final point to consider is that it takes a substantial amount of time to become an expert user of a given tool. Users must make a significant investment to learn how the tool works and the most efficient way of performing their job. The more discrete tools in an environment, the more challenging this becomes.

Minimizing the number of tools and their different interfaces, styles of interaction and approach to security and collaboration helps improve the user experience and decreases time to market.

Unified platform, unified personas

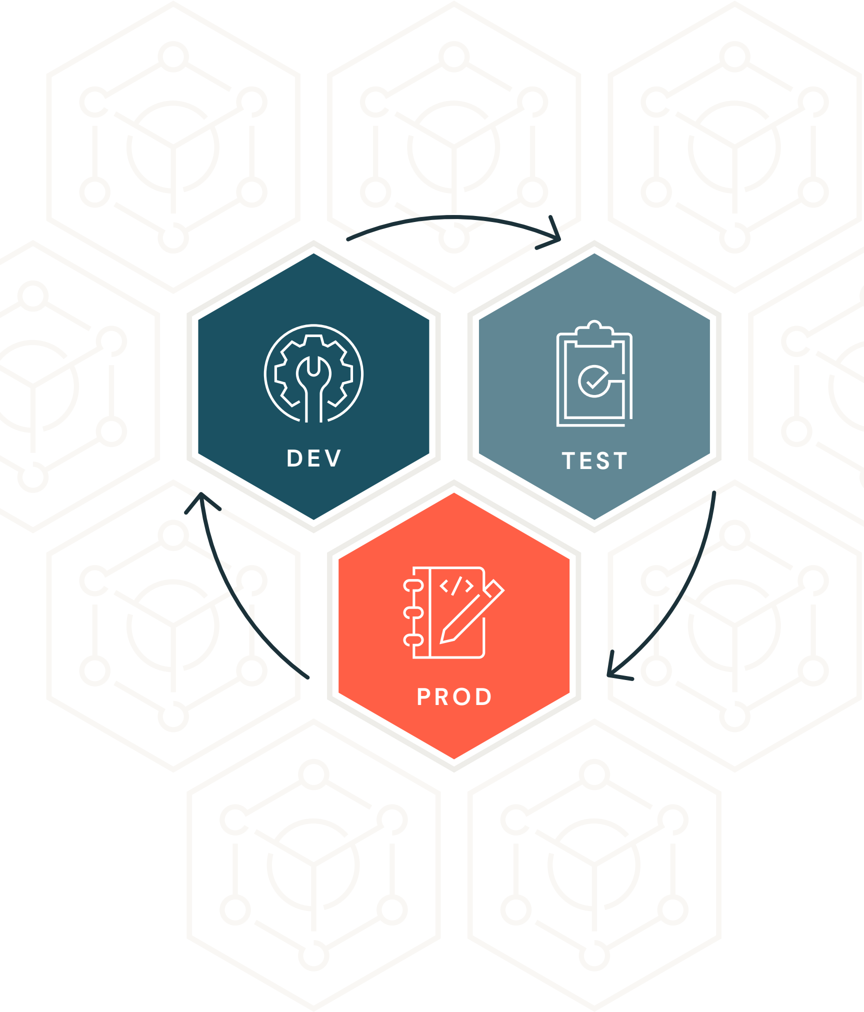
Deploying a unified data platform — like the Databricks Lakehouse Platform, which implements a modern data stack — will provide an integrated suite of tools for the full range of personas in your organization, including business analysts, SQL developers, data engineers and data scientists. You will immediately increase productivity and reduce risk because you'll be better able to share the key aspects of data pipelining — including ingestion, partitioning, curation, SQL analytics, reporting, and model development and deployment. All the work streams function off a single view of the data, and the handoffs between subsystems are well managed.

Data processing happens in one auditable environment, and the number of copies of data is kept to an absolute minimum — with each user benefiting from the data assets created by others. Redundant work is eliminated.

The 80/20 dilemma for data scientists shifts to a healthier ratio, and they now are able to spend more time working with rather than collecting the data. It's difficult to decide what algorithm will work best — shifting the 80/20 ratio allows the data scientist to try out multiple algorithms to solve a problem.

Another challenge is that enterprise data changes rapidly. New fields are added or existing fields are typed differently — for example, changing a string to an integer. This has a cascading effect, and the downstream consumers must be able to adjust by monitoring the execution and detecting the changes. The data scientist, in turn, must update and test new models on the new data. Your data platform should make the detection and remediation easier, not harder.

For the data engineers, their primary focus is extracting data from source systems and moving it into the new data ecosystem. The data pipeline function can be simplified with a unified data platform because the programming model and APIs are consistent across programming languages (e.g., Scala, Python). This results in improved operations and maintenance (O&M). The runtime environment is easier to troubleshoot and debug since the compute layer is consistent, and the logging and auditing associated with the data processing and data management is centralized and of more value.



Maximize the productivity of your workforce

Once you have a data platform that brings together your full range of personas, you should focus on the next step for increasing productivity — namely, self-service environments.

In large organizations, there needs to be a strategy for how solutions are promoted up through the runtime environments for development, testing and production. These environments need to be nearly identical to one another — using the same version of software while limiting the number, size and horsepower of the compute nodes. To the extent possible, development and test should be performed with realistic test/synthetic data. One strategy to support this is to tap into the flow of production data and siphon off a small percentage that is then changed in randomized fashion — obfuscating the real data but keeping the same general shape and range of values.

The **DEV** environment should be accessible to everyone without any organizational red tape. The DEV environments should be small and controlled with policies that spin them up and tear them down efficiently. Every aspect of the DEV infrastructure should be treated as ephemeral. Nothing should exist in the environment that cannot be destroyed and easily rebuilt.

The **TEST** environment should mimic the PROD environment as much as possible, including the monitoring tools — within obvious cost/budget constraints. The use of the TEST environment can be requested by the developers, but the process is governed using a workflow/sign-off approval approach — signed off by management.

Moving to **PROD** is the final step, and there usually is a "separation of duties" that is required so that developers cannot randomly promote software to run in production. Again, this process should be strictly governed using a workflow/sign-off approval approach — signed off by management as well. Many organizations fully automate the steps, except the sign-offs, and support the notion of continuous deployments.

8. Make informed build vs. buy decisions

A key piece of the strategy will involve the decision around which components of the data ecosystem are built by the in-house engineering team and which components are purchased through a vendor relationship. There is increased emphasis within engineering teams on taking a "builder" approach. In other words, the engineering teams prefer to develop their own solutions in-house rather than rely on vendor products.

Competitive advantage

This "roll your own" approach has some advantages — including being able to establish the overall product vision, prioritize features and directly allocate the resources to build the software. However, it is important to keep in mind which aspects of your development effort give you the most competitive advantage.

Spend some time working with the data transformation steering committee and other stakeholders to debate the pros and cons of building out various pieces of the data ecosystem. The primary factor should come down to whether or not a given solution offers true competitive advantage for the organization. Does building this piece of software make it harder for your competitors to compete with you? If the answer is no, then it is better to focus your engineering and data science resources on deriving insights from your data.

Beware: becoming your own software vendor

As many engineering leaders know, building your own software is an exciting challenge. However, it does come with added responsibility — namely, managing the overall project timeline and costs, and being responsible for the design, implementation, testing, documentation, training, and ongoing maintenance and updates. You basically are becoming your own software vendor for every component of the ecosystem that you build yourself. When you consider the cost of a standard-sized team, it is not uncommon to spend several million dollars per year building out individual component parts of the new data system. This doesn't include the cost to operate and maintain the software once it is in production.

Databricks is built on top of popular open source software that it created. Engineering teams can improve the underpinnings of the Databricks platform by submitting code via pull request and becoming committers to the projects. The benefit to organizations is that their engineers contribute to the feature set of the data platform while Databricks remains responsible for all integration and performance testing plus all the runtime support, including failover and disaster recovery.

To offset the anticipated development costs, engineering teams will oftentimes make the argument that they are starting with open source software and extending it to meet the "unique requirements" of your organization. It's worth pressure testing this approach and making sure that a) the requirements truly are unique and b) the development offers the competitive advantage that you need.

Even software built on top of open source still requires significant investment in integration and testing. The integration work is particularly challenging because of the large number of open source libraries that are required in the data science space. The question becomes, "Is this really the area that you want your engineering teams focused on?" Or would it be better to "outsource" this component to a third party?

How long will it take? Can the organization afford to wait?

Even if you decide the software component provides a competitive advantage and is something worth building in-house, the next question that you should ask is, "How long will it take?" There is definitely a time-to-market consideration, and the build vs. buy decision needs to also account for the impact to the business due to the anticipated delivery schedule. Keep in mind that software development projects usually take longer and cost more money than initially planned.

The organization should understand the impact to the overall performance and capabilities of the daily ecosystem for any features tied to the in-house development effort. Your business partners likely do not care how the data ecosystem is implemented as long as it works, meets their needs, is performant, is reliable and is delivered on time. Carefully weigh the trade-offs among competitive advantage, cost, features and schedule.

Don’t forget about the data

Perhaps the single most important feature of a modern data stack is its ability to help make data sets and “data assets” consumable to the end users or systems. Data insights, model training and model execution cannot happen in a reliable manner unless the data they depend on can be trusted and is of good quality. In large organizations, revenue opportunities and the ability to reduce risk often depend on merging data sets from multiple lines of business or departments. Focusing your data engineering and data science efforts on curating data and creating robust and reliable pipelines likely provides the best chance at creating true competitive advantage.

The amount of work required to properly catalog, schema enforce, quality check, partition, secure and serve up data for analysis should not be underestimated. The value of this work is equally important to the business. The ability to curate data to enable game-changing insights should be the focus of the work led by the CDO and CIO. This has much more to do with the data than it does with the ability to have your engineers innovate on components that don’t bring true competitive advantage.

The Databricks platform optimizes costs for your data and AI workloads by intelligently provisioning infrastructure only as you need it. Customers can establish policies that govern the size of clusters based on DEV, TEST, PROD environments or anticipated workloads.

9. Allocate, monitor and optimize costs

Beginning in 1987, Southwest Airlines famously standardized on flying a single airplane type — the Boeing 737 class of aircraft. This decision allowed the airline to save on both operations and maintenance — requiring only one type of simulator to train pilots, streamlining their spare parts supply chain and maintaining a more manageable parts inventory. Their pilots and maintenance crews were effectively interchangeable in case anyone ever called in sick or missed a connection. The key takeaway is that in order to reduce costs and increase efficiency, Southwest created their own version of a unified platform — getting all their flight-related personas to collaborate and operate from the same point of view. Lessons learned on the platform could be easily shared and reused by other members of the team. The more the team used the unified platform, the more they collaborated and their level of expertise increased.

Reduce complexity, reduce costs

The architectures of enterprise data warehouses (EDWs) and data lakes were either more limited or more complex — resulting in increased time to market and increased costs. This was mainly due to the requirement to perform ETL to explore data in the EDW or the need to split data using multiple pipelines for the data lake. The data lakehouse architecture simplifies the cost allocation because all the processing, serving and analytics are performed in a single compute layer.

Organizations can rightsize the data environments and control costs using policies. The centralized and consistent approach to security, auditing and monitoring makes it easier to spot inefficiencies and bottlenecks in the data ecosystem. Performance improvements can be gained quickly as more platform expertise is developed within the workforce.

Databricks monitors and records usage and allows organizations to easily track costs on a data and AI workload basis. This provides the ability to implement an enterprise-wide chargeback mode and put in place appropriate spending limits.

Centralized funding model

As previously mentioned, data transformation initiatives require substantial funding. Centralizing the budget under the CDO provides consistency and visibility into how funds are allocated and spent — increasing the likelihood of a positive ROI. Funding at the beginning of the initiative will be significantly higher than the funding in the out-years. It's not uncommon to see 3- to 5-year project plans for larger organizations. Funding for years 1 and 2 is often reduced in years 3 and 4 and further reduced in year 5 — until it reaches a steady state that is more sustainable.

The budget takes into account the cost of the data engineering function, commercial software licenses and building out the center of excellence to accelerate the data science capabilities of the organization. Again, the CDO must partner closely with the CIO and the enterprise architect to make sure that the resources are focused on the overall implementation plan and to make sound build vs. buy decisions.

It's common to see the full budget controlled by the CDO, with a significant portion allocated to resources in the CIO's organization to perform the data engineering tasks. The data science community reports into the CDO and is matrixed into the lines of business in order to better understand the business drivers and the data sets. Finally, investing in data governance cannot wait until the company has suffered from a major regulatory challenge, a data breach or some other serious defense-related problem. CDOs should spend the necessary time to educate leaders throughout the organization on the value of data governance.

Chargeback models

To establish the centralized budget to fund the data transformation initiative, some organizations impose a "tax" on each part of the organization — based on size as well as profit and loss. This base-level funding should be used to build the data engineering and data science teams needed to deploy the building blocks of the new data ecosystem. However, as different teams, departments and business units begin using the new data ecosystem, the infrastructure costs, both compute and storage, will begin to grow. The costs will not be evenly distributed, due to different levels of usage from the various parts of the organization. The groups with the heavier usage should obviously cover their pro rata share of the costs. This requires the ability to monitor and track usage — not only based on compute but also on the amount of data generated and consumed. This so-called chargeback model is an effective and fair way to cover the cost deltas over and above the base-level funding.

Plus, not all the departments or lines of business will require the same level of compute power or fault tolerance. The architecture should support the ability to separate out the runtime portions of the data ecosystem and isolate the workloads based on the specific SLAs for the use cases in each environment. Some workloads cannot fail and their SLAs will require full redundancy, thus increasing the number of nodes in the cluster or even requiring multiple clusters operating in different cloud regions. In contrast, less critical workloads that can fail and be restarted can run on less costly infrastructure. This makes it easier to better manage the ecosystem by avoiding a one-size-fits-all approach and allocating costs to where the performance is needed most.

In 2018, Databricks released MLflow — an open source platform to manage the ML lifecycle, including experimentation, reproducibility, deployment and a central model registry. MLflow is included in the Databricks Lakehouse Platform and accelerates the adoption of machine learning and AI in organizations.

10. Move to production and scale adoption

Now that you've completed the hard work outlined in the first nine steps, it is time to put the new data ecosystem to use. In order to get truly game-changing results, organizations must be really disciplined at managing and using data to enable use cases that drive business value. They must also establish a clear set of metrics to measure adoption and track the net promoter score (NPS) so that the user experience continues to improve over time.

If you build it, they will come

Keep in mind that your business partners are likely the ones to do the heavy lifting when it comes to data set registration. Without a robust set of relevant, quality data to use, the data ecosystem will be useless. A high level of automation for the registration process is important because it's not uncommon to see thousands of data sets in large organizations. The business and technical metadata plus the data quality rules will help guarantee that the data lake is filled with consumable data. The lineage solution should provide a visualization that shows the data movement and verifies that the approved data flow paths are being followed.

Some key metrics to keep an eye on are:

- Percentage of source systems contributing data to the ecosystem
- Percentage of real-time streaming relative to API and batch transfers
- Percentage of registered data sets with full business and technical metadata
- Volume of data written to the data lake
- Percentage of raw data that enters a data curation pipeline
- Volume of data consumed from the data lake
- Number of tables defined and populated with curated data
- Number of models trained with data from the data lake
- Lineage reports and anomaly detection incidents
- Number of users running Python, SQL, Scala and R workloads

Communication plan

Communication is critical throughout the data transformation initiative — however, it is particularly important once you move into production. Time is precious and you want to avoid rework, if at all possible. Organizations often overlook the emotional and cultural toll that a long transformation process takes on the workforce. The seam between the legacy environment and the new data ecosystem is an expensive and exhausting place to be — because your business partners are busy supporting two data worlds. Most users just want to know when the new environment will be ready. They don't want to work with partially completed features, especially while performing double duty.

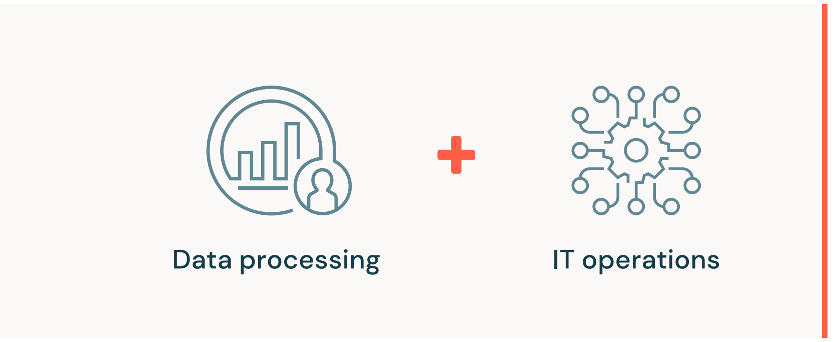
Establish a solid communication plan and set expectations for when features will come online. Make sure there is detailed documentation, training and a support/help desk to field users' questions.

DevOps — software development + IT operations

Mature organizations develop a series of processes and standards for how software and data are developed, managed and delivered. The term "DevOps" comes from the software engineering world and refers to developing and operating large-scale software systems. DevOps defines how an organization, its developers, operations staff and other stakeholders establish the goal of delivering quality software reliably and repeatedly. In short, DevOps is a culture that consists of two practices: continuous integration (CI) and continuous delivery (CD).

The CI portion of the process is the practice of frequently integrating newly written or changed code with the existing code repository. As software is written, it is continuously saved back to the source code repository, merged with other changes, built, integrated and tested — and this should occur frequently enough that the window between commit and build is narrow enough that no errors can occur without developers noticing them and correcting them immediately.

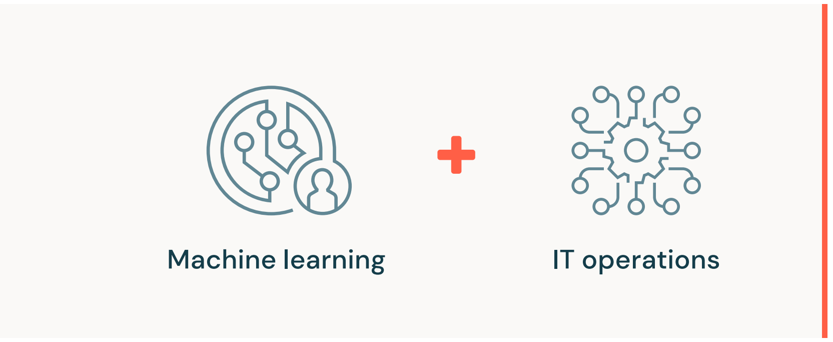
This is particularly important for large, distributed teams to ensure that the software is always in a working state — despite the frequent changes from various developers. Only software that passes the CI steps is deployed — resulting in shortened development cycles, increased deployment velocity and the creation of dependable releases.



DataOps — data processing + IT operations

DataOps is a relatively new focus area for the data engineering and data science communities. Its goal is to use the well-established processes from DevOps to consistently and reliably improve the quality of data used to power data and AI use cases. DataOps automates and streamlines the lifecycle management tasks needed for large volumes of data — basically, ensuring that the volume, velocity, variety and veracity of the data are taken into account as data flows through the environment. DataOps aims to reduce the end-to-end cycle time of data analytics — from idea, to exploration, to visualizations and to the creation of new data sets, data assets and models that create value.

For DataOps to be effective, it must encourage collaboration, innovation and reuse among the stakeholders, and the data tooling should be designed to support the workflow and make all aspects of data curation and ETL more efficient.



MLOps — machine learning + IT operations

Not surprisingly, the term “MLOps” takes the DevOps approach and applies it to the machine learning and deep learning space — automating or streamlining the core workflow for data scientists. MLOps is a bit unique when compared with DevOps and DataOps because the approach to deploying effective machine learning models is far more iterative and requires much more experimentation — data scientists try different features, parameters and models in a tight iteration cycle. In all these iterations, they must manage the code base, understand the data used to perform the training and create reproducible results. The logging aspect of the ML development lifecycle is critical.

MLOps aims to manage deployment of machine learning and deep learning models in large-scale production environments while also focusing on business and regulatory requirements. The ideal MLOps environment would include data science tools where models are constructed and analytical engines where computations are performed.

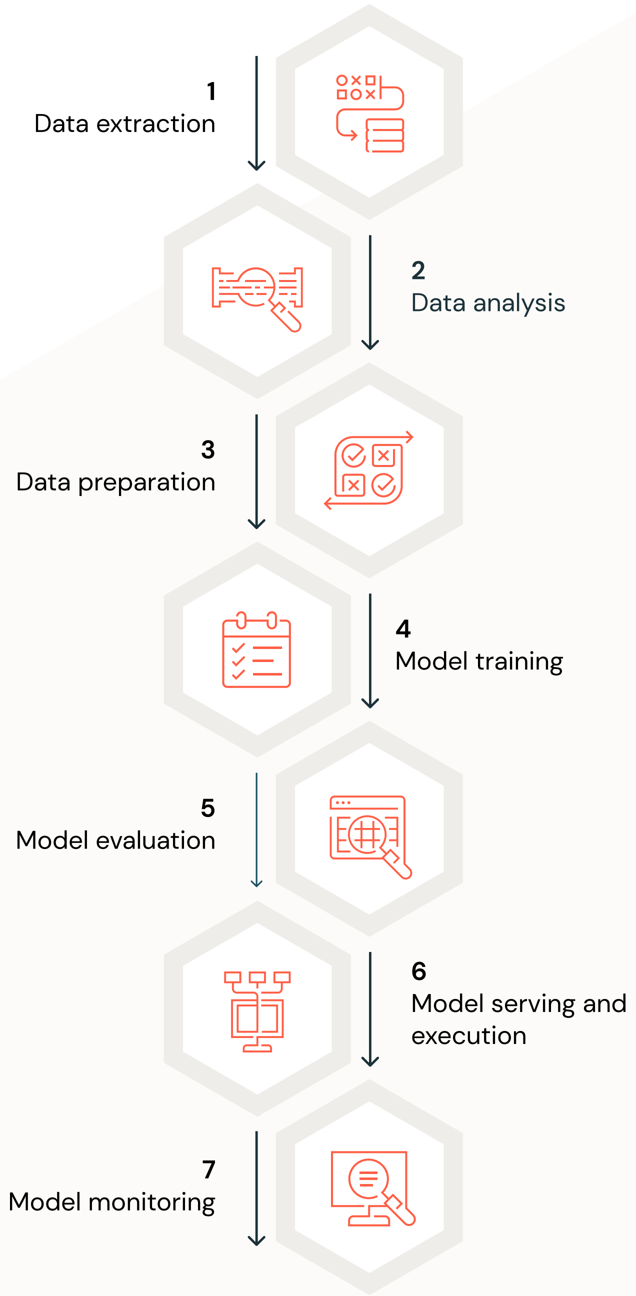


Figure 10:
Workflow for deploying production ML models

The overall workflow for deploying production ML models is shown in Figure 10.

Unlike most software applications that execute a series of discrete operations, ML platforms are not deterministic and are highly dependent on the statistical profile of the data they use. ML platforms can suffer performance degradation of the system due to changing data profiles. Therefore, the model has to be refreshed even if it currently "works" — leading to more iterations of the ML workflow. The ML platform should natively support this style of iterative data science.

Ethics in AI

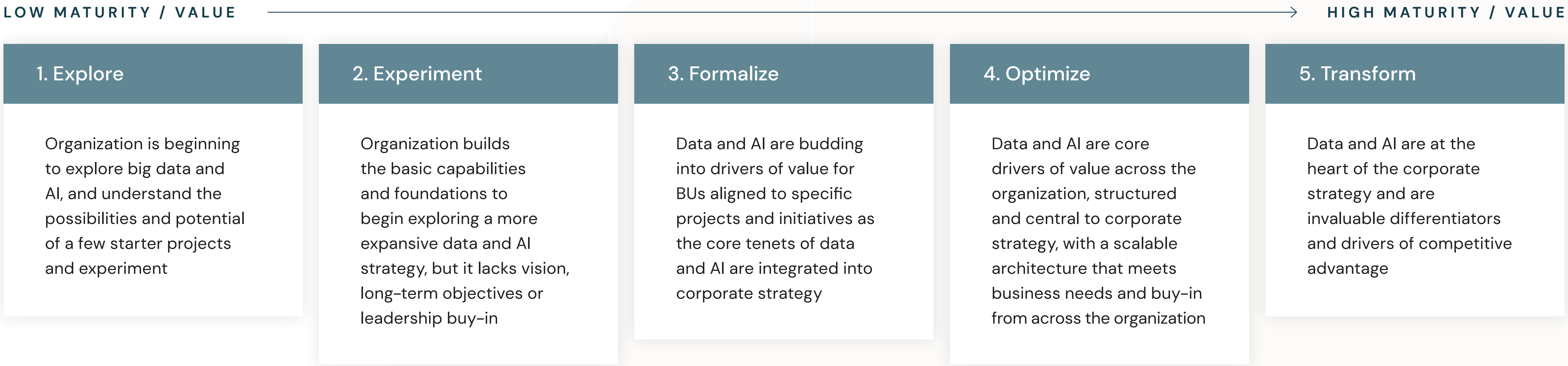
As more organizations deploy data and AI solutions, there is growing concern around a number of issues related to ethics — in particular, how do you ensure the data and algorithms used to make decisions are fair and ethical, and that the outcomes have the appropriate impact on the target audience? Organizations must ensure that the "black box" algorithms that produce results have the transparency, interpretability and explainability to satisfy legal and regulatory safeguards.

The vast majority of AI work still involves software development by human beings and the use of curated data sets. There is the obvious potential for bias and the application of AI in domains that are ethically questionable. CDOs are faced with the added challenge of needing to be able to defend the use of AI, explain how it works and describe the impact of its existence on the target audience — whether internal workers or customers.

Data and AI Maturity Model

When data and AI become part of the fabric of the company and the stakeholders in the organization adopt a data asset and AI mindset, the company moves further along a well-defined maturity curve, as shown in Figure 11.

Top-Line Categories and Ranking Criteria



LOW MATURITY / VALUE → HIGH MATURITY / VALUE

1. Explore

Organization is beginning to explore big data and AI, and understand the possibilities and potential of a few starter projects and experiment

2. Experiment

Organization builds the basic capabilities and foundations to begin exploring a more expansive data and AI strategy, but it lacks vision, long-term objectives or leadership buy-in

3. Formalize

Data and AI are budding into drivers of value for BUs aligned to specific projects and initiatives as the core tenets of data and AI are integrated into corporate strategy

4. Optimize

Data and AI are core drivers of value across the organization, structured and central to corporate strategy, with a scalable architecture that meets business needs and buy-in from across the organization

5. Transform

Data and AI are at the heart of the corporate strategy and are invaluable differentiators and drivers of competitive advantage

Figure 11:
The Data and AI Maturity Model

Databricks partners with its customers to enable them to do an internal self-assessment. The output of the self-assessment allows organizations to:

- Understand the current state of their journey to data and AI maturity
- Identify key gaps in realizing (more) value from data and AI
- Plot a path to increase maturity with specific actions
- Identify Databricks resources who can help support their journey

CHAPTER 3:

Conclusion

After a decade in which most enterprises took a hybrid approach to their data architecture — and struggled with the complexity, cost and compromise that come with supporting both data warehouses and data lakes — the lakehouse paradigm represents a breakthrough. Choosing the right modern data stack will be critical to future-proofing your investment and enabling data and AI at scale. The simple, open and multicloud architecture of the Databricks Lakehouse Platform delivers the simplicity and scalability you need to unleash the power of your data teams to collaborate like never before — in real time, with all their data, for every use case.

For more information, please visit Databricks or contact us.

ABOUT THE AUTHOR

Chris D'Agostino is the Global Field CTO at Databricks, having joined the company in January 2020. His role is to provide thought leadership and serve as a trusted advisor to our top customers, globally.

Prior to Databricks, Chris ran a 1,000-person data engineering function for a top 10 U.S. bank. In that role, he led a team that was responsible for building out a modern data architecture that emphasized the key attributes of the lakehouse architecture.

Chris has also held leadership roles at a number of technology companies.

About Databricks

Databricks is the data and AI company. More than 7,000 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.

Sign up for a free trial

databricks