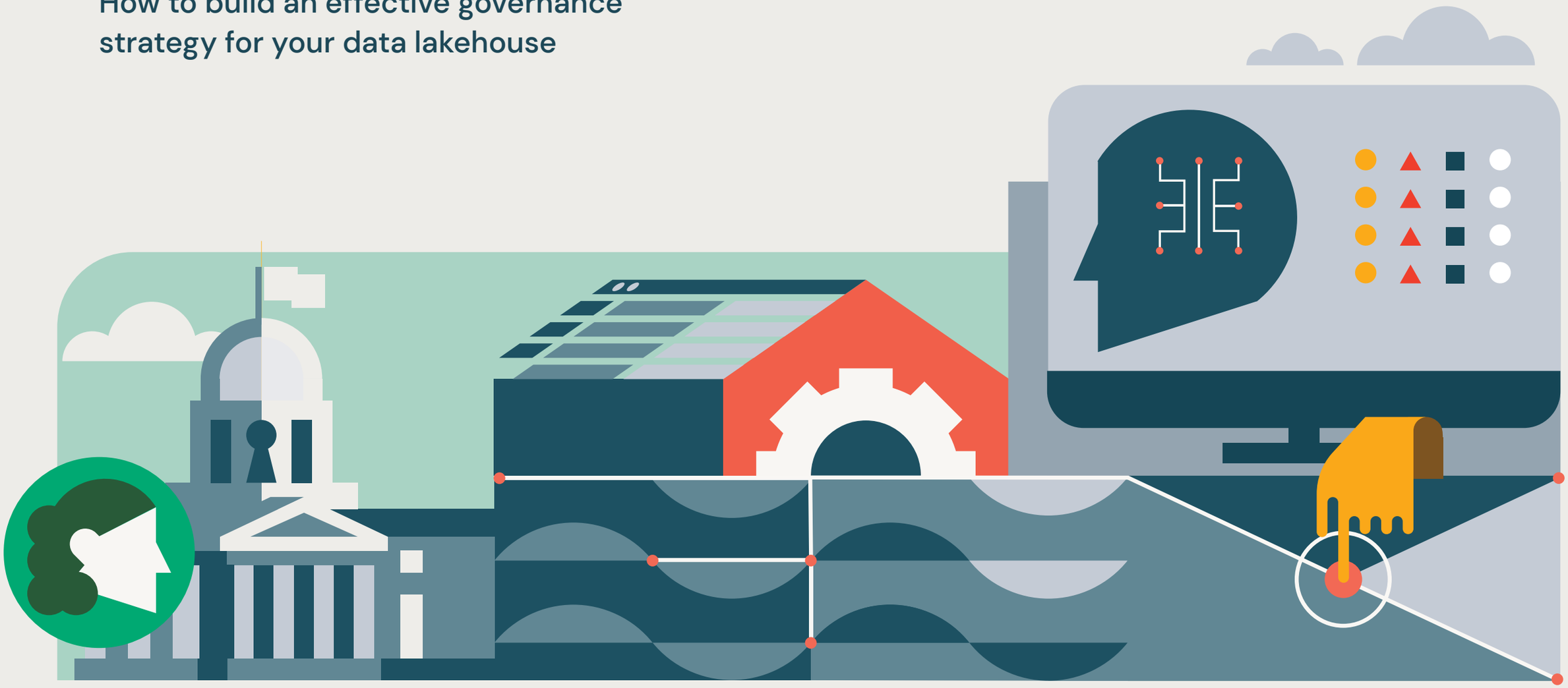
eBook

A Comprehensive Guide to Data and AI Governance

How to build an effective governance strategy for your data lakehouse



databricks

Contents

Executive summary

The dynamic interplay of data, analytics and artificial intelligence (AI) is driving a wave of transformative innovations across diverse sectors, reshaping revenue streams and redefining corporate management paradigms. McKinsey & Company's projections underline the enormity of this potential, estimating that by 2030, analytics and AI could introduce more than $15 trillion in fresh business value. Furthermore, McKinsey's recent State of AI report reveals that, due to advancements in Generative AI, 40 percent of organizations are planning to increase their overall investment in AI. This trend is driving a global upsurge in organizational investments aimed at cultivating data-driven cultures and establishing competitive advantages.

Despite these investments, the challenge of deriving significant value from data and AI initiatives persists. Often, the heart of this challenge lies in the absence of a comprehensive and actionable data and AI governance strategy that encompasses the entire range of data applications, spanning from business intelligence to machine learning (ML). Gartner's insights provide additional depth, predicting that until 2025, a staggering 80% of organizations striving for digital business expansion will encounter obstacles due to outdated approaches to data and analytics governance.

In today's landscape, characterized by a renaissance in Generative AI and Large Language Models (LLMs), the importance of robust data and AI governance is accentuated even further. These advanced technologies empower organizations to create content, simulate scenarios and enhance decision-making. However, they also amplify concerns regarding data privacy, bias mitigation and ethical considerations. As AI becomes increasingly ingrained in core operations, the need for meticulous governance to ensure fairness, accountability and security becomes paramount. McKinsey's Global Survey on AI emphasizes that organizations achieving the highest AI returns have comprehensive AI governance frameworks that cover every stage of the model development process. This sentiment is echoed by Forrester in its 2023 AI Predictions, stating that one in four tech executives will be reporting to their board on AI governance. AI has now become an essential aspect of enterprises and consequently, AI governance is joining cybersecurity and compliance as a topic of discussion at the board level.

Hence, data and AI governance serve as the intricate yet pivotal foundations of digital transformation endeavors. Data teams grapple with an array of challenges, including managing diverse data sources, dismantling silos, upholding data and ML model quality, facilitating trusted data discovery, ensuring secure data access and navigating regulatory landscapes. A forward-looking data governance strategy emerges as the guiding compass, addressing these complexities and empowering organizations to fully unlock the potential of their data investments.

This eBook explores the methodologies that drive successful data and AI governance strategies. Additionally, it uncovers the role of the Databricks Lakehouse Platform as a catalyst for streamlining these transformative efforts. As organizations navigate an era characterized by AI-centric data strategies, the convergence of innovation and governance becomes the foundation for sustainable growth and responsible technological advancement.

Chapter 1
Data Governance

Data is one of the most valuable assets for many organizations, but data governance is the key to unlocking that value.

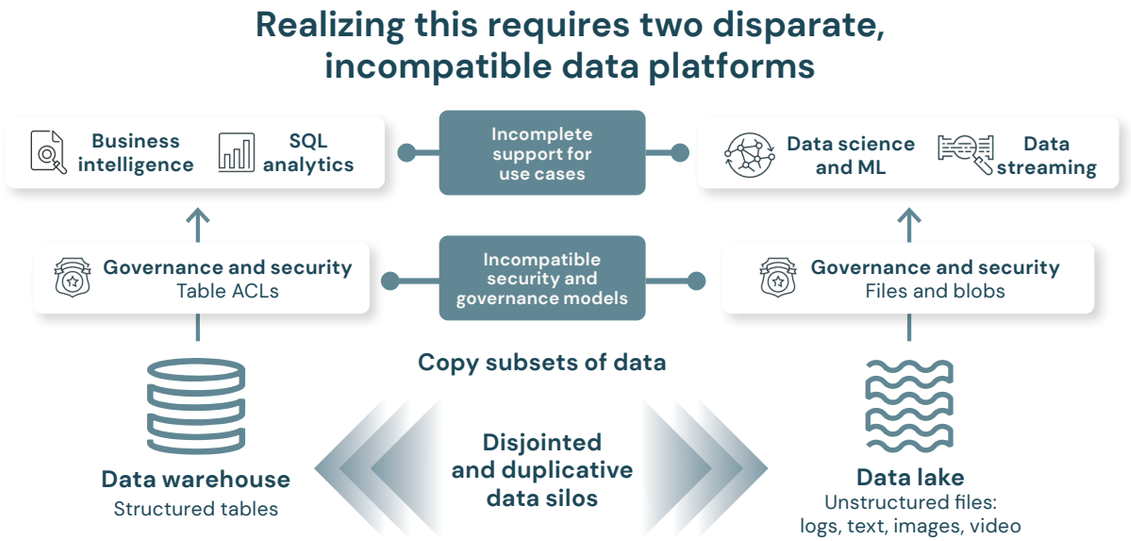
Data governance is a framework of principles, practices and tooling that helps manage the complete lifecycle of your data and aligns data-related requirements to the business strategy. A pragmatic data governance strategy gives data teams superior data management, visibility and auditing of data access patterns across their organization. Implementing an effective data governance solution helps companies protect their data from unauthorized access and ensures that rules are in place to comply with regulatory requirements. 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.

Key data governance challenges

Crafting an effective data governance strategy has become a complex undertaking due to how organizations collect and analyze data. Many organizations grapple with the following issues:

Fragmented data landscape: The presence of data silos across various sources, such as data lakes, warehouses and databases, poses significant challenges for governance teams. These silos hinder the creation of a cohesive view of data, resulting in inefficient data discovery. Most organizations store massive amounts of unstructured data in cloud storage platforms like AWS S3, Azure ADLS and Google Cloud Storage (GCS). In fact, IDC predicts that by 2025, approximately 80% of data in any organization will be unstructured. A subset of this unstructured data is transformed into structured tables within data warehouses for business intelligence purposes. Unfortunately, this data movement creates silos, with data scattered between two systems. Additionally, modern data assets like dashboards, machine learning models and notebooks contribute to this fragmentation. The absence of a unified view hampers data and AI asset discovery, access and effective analysis. This leads to delays in decision-making and inhibits innovation. Managing these silos requires extra resources and investments in data integration technologies, data engineering and governance processes, resulting in higher operational costs.

Complex access management: Enterprises utilize various tools to secure access to diverse data and AI assets, leading to complex, inconsistent and error-prone access management. Different permission structures are applied to data lakes and data warehouses, resulting in inconsistent controls. Furthermore, the tools supporting these platforms have fundamentally different structures, hindering collaboration between the teams responsible for them. This inconsistency affects permissions management, audits, data discovery and sharing. Handling modern data assets like dashboards, machine learning models and notebooks, each with their own permission models, further complicates the issue. When these assets exist across multiple clouds, the complexity grows due to varying cloud-specific access management frameworks.



Inadequate monitoring and visibility: The lack of comprehensive monitoring and visibility into the lifecycle of data and AI assets between systems hampers effective audits, impact analyses and error diagnosis within data and AI pipelines. This inability to track the origin, evolution, transformations, movements and usage of assets undermines data quality assurance.

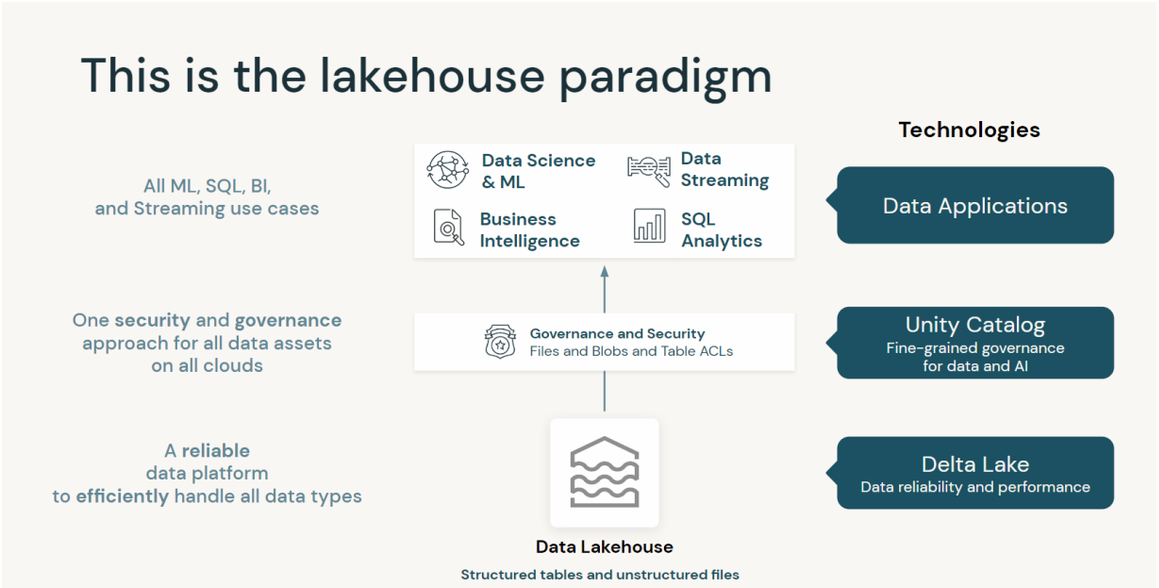
Limited cross-platform sharing and collaboration: The absence of a standardized sharing solution inhibits secure cross-cloud and cross-platform sharing and collaboration of data and AI assets, including machine learning models, notebooks and dashboards. This situation forces enterprises to replicate data across multiple platforms, clouds and regions to facilitate collaboration, resulting in redundancy.

Organizations need a unified approach to simplify data, analytics and AI governance. Gartner predicts that by 2026, 20% of large enterprises will adopt a single data and analytics governance platform to streamline and automate governance efforts. Let's explore how the Databricks Lakehouse Platform addresses these challenges.

Data governance on the lakehouse

What is a data lakehouse?

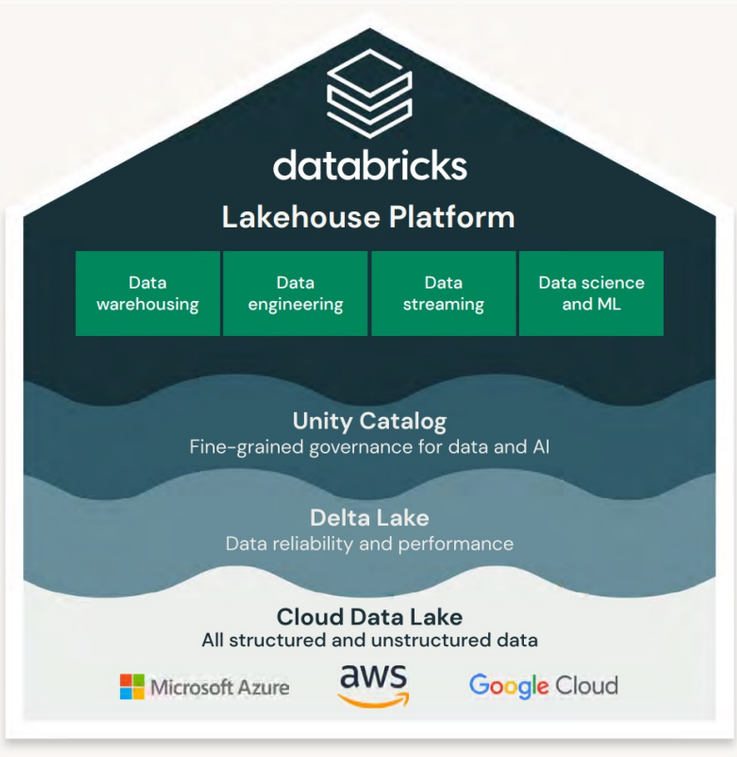
In a 2020 whitepaper, *Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics*, Databricks founders set the vision for a data lakehouse architecture that combines the capabilities of both data warehouses and data lakes to achieve the performance of a data warehouse with the economies and scale of a data lake. This unified approach simplifies your modern data stack by eliminating the data silos that traditionally separate and complicate data engineering, analytics, BI, data science and machine learning.



The Databricks Lakehouse Platform unifies your data warehousing and AI use cases on a single platform. It combines the capabilities of data lakes and data warehouses to deliver reliability and performance without sacrificing open standards and support for both structured and unstructured data. And to simplify data governance, the Databricks Lakehouse Platform offers

Unity Catalog, a unified governance solution for data, analytics and AI on your lakehouse. By minimizing the 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.

In this chapter, we will discuss how you can leverage the Databricks Lakehouse Platform to manage and govern the end-to-end data lifecycle.



Data management

Data management lays the foundation for how you execute your data governance strategy. The practice of data management involves the collection, integration, organization and persistence of trusted data sets to help organizations maximize value. More so now than ever, an organization's value is based on how well it can derive value from the data under its management. Data management also involves helping organizations understand their data usage frequency and provides tools for data lifecycle management.

Data management can be performed on any system that persists data. In the analytics realm, this has historically been done on a data warehouse. Data warehouses consist of tabular data and therefore have managed data sets via constructs such as tables and views, which consist of rows and columns. Data lakes are capable of persisting any type of structured or unstructured data for data science or machine learning use cases. Data lakes are capable of storing file formats like Apache Parquet, images or video, or raw text files for unstructured data. The variety of such data sets requires that all data be managed at the file level.

With a data lakehouse, organizations can persist their data once, and use it for all analytics use cases. This concept alone helps reduce data duplication and minimize the scope of all data that an organization needs to manage.

Data ingestion

Before data can be persisted for future use, it must first be collected. Data can be collected from anywhere, but is primarily collected from the following sources:

- Cloud storage
- Message queues
- Relational databases
- SaaS APIs

Increasingly, data is collected from files written to object storage services that are native to public cloud providers. This data could include anything from a handful to millions of files written per day. The data can include any type of file, including:

- Unstructured (PDF, text, audio, video)
- Semi-structured (JSON)
- Structured (Parquet, Avro)

Databricks provides Auto Loader functionality to automatically detect new files on cloud storage, merge any schema changes, and validate data against the expected schema. Auto Loader provides this functionality across all of the major public cloud providers.

Another common data source, especially for data streaming, is a distributed message queue, such as Apache Kafka. This connection allows for low latency processing of messages in sequential order. In addition to open source queues, each of the major cloud providers has its own native message queue service. All of these can be leveraged in Databricks as just another Apache Spark™ Structured Streaming data source.

One of the most popular data sources to collect data from is a relational database management system (RDBMS). Even though an RDBMS can perform its own data management, data is typically exported from these databases so that analytics can be performed without impacting the online uses of the databases. This data can be collected either through static reads via the Apache Spark JDBC data source, or by applying a listener to the transaction log of the RDBMS. Databricks partners with a number of vendors within Partner Connect to integrate directly with the transaction log of these RDBMS systems so that every change to the data can be captured.

The rise of software as a service (SaaS) means that more and more data is being created and mastered within cloud-based services, rather than on-prem databases. These disparate services need to have their data exported and combined to produce value. Typically, this is done by making a series of API calls to the SaaS provider endpoints. There are a number of vendors who have created hundreds of connectors to easily collect such data using managed services created specifically for the task. They, too, can be found in Partner Connect.

Data persistence and organization

After data has been collected, it needs to be persisted somewhere so that it can be managed and governed. With the rise of the public cloud, that place is increasingly cloud storage services such as AWS S3, Azure ADLS and GCP GCS. Each of these services has its unique characteristics which may or may not provide atomicity and consistency at the file level. Because of this variance, data processing against these storage layers has historically been fraught with data quality issues. To overcome this challenge, Delta Lake was developed as an open-format storage layer to deliver reliability, security and performance on a data lake for both streaming and batch operations.

Delta Lake provides ACID transactions for structured tables, allowing distributed write operations while maintaining a consistent view of the data for reads. This is the foundation for what makes a data lakehouse. Delta Lake is primitive for data integrity, but also serves as an enabler for performance gains when querying the data. The statistics of the underlying files are centralized so that they may be skipped from reads when performing queries. Every transaction against Delta Lake is persisted alongside the data so that a change data feed can be queried for easy change data capture against the tables themselves. Finally, because the transactions are retained, the table can be queried because it was represented at a point in the past using time travel.

Data integration

Data typically needs to be integrated from disparate sources in order to maximize value. This can be done either in a batch or streaming mode to make insights available as quickly as possible. Integrating data typically requires some kind of orchestration so that an end-to-end process can be architected to collect, ingest, combine, transform and deliver the product of all these steps. This is why Databricks provides Workflows as a fully managed solution to orchestrate all of the tasks necessary to integrate data.

An orchestration engine generally understands the dependency graph of operations and whether these tasks successfully complete or not. However, a general-purpose orchestration engine does not have knowledge of the data being processed and cannot make adjustments accordingly. Delta Live Tables is the first extract, transform and load (ETL) framework that uses a simple declarative approach to building reliable data pipelines and automatically managing your infrastructure at scale. Because Delta Live Tables is aware of the amount of data it is about to process, it can automatically trigger infrastructure to scale up or down so that service level objectives are met for data pipelines, regardless of whether it is done in batch or streaming mode.

One type of data integration that can uniquely be performed in a data lakehouse is entity resolution (ER). ER is all about understanding how data represents real-world “things” — or “entities,” for lack of a better word. Common examples are a person, a product, a household, a vendor or a location. ER is typically used in the practice of master data management (MDM) when conforming dimensions in a data mart. A simple example is recognizing that a record pertaining to “Bill Wallace” is in fact the same person being referenced as “William Wallace” from a different system. Linking these records requires applying sophisticated algorithms at scale against a graph of data.

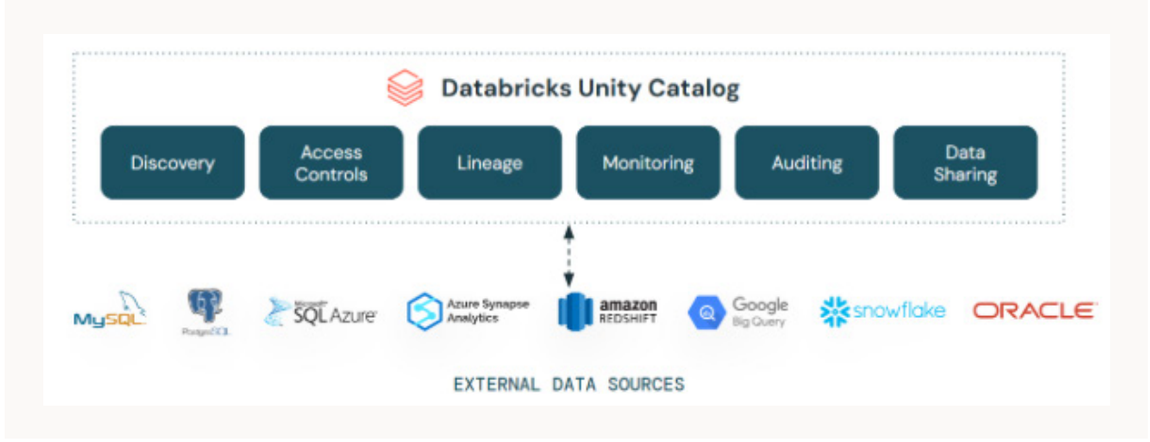
Even at medium scale, say 1 million records, care must be taken with processing techniques. ER is all about comparing records to each other, and it is fundamentally an O(n2) problem. For example, 1 million records could require on the order of 1 trillion comparisons. The Databricks Lakehouse Platform is an excellent choice here because it inherently uses distributed processing and still allows for the use of super-powerful packaged OSS — like Zingg and Splink. At this medium scale, you also need to look into prefiltering techniques like binning, blocking and pre-clustering. At any scale, the Databricks Lakehouse Platform can provide the tools needed to be successful with the integration of ER.

Data federation

Data is often spread out and compartmentalized among numerous heterogeneous data stores. To fully harness the potential of their data, organizations require a centralized access point to query data from diverse data sources. In the conventional approach, data teams typically transfer their data from external sources or systems to a chosen platform (such as a Lakehouse), a process involving ETL procedures. As an alternative, they can depend on query federation/data virtualization vendors to create a unified virtual layer of all their data. Essentially, data virtualization establishes an analytical database of enterprise data without necessitating ETL procedures. This eliminates the necessity for engineers to duplicate, move, or modify the data, empowering data scientists to engage with the data at its source. However, this also introduces additional intricacies and expenses. The integration of data brings about dependencies on data engineering and creates numerous bottlenecks.

Additionally, enterprises manage various governance tools to oversee and secure their scattered data across platforms and data sources. This results in governance that is inconsistent and fragmented across their data infrastructure. Such fragmented governance and lax adherence to compliance standards lead to duplicated efforts and heighten the risk of being unable to monitor and safeguard against unauthorized access or data leaks.

Databricks Lakehouse Platform offers Lakehouse Federation empowering you to build an open, efficient and secure data mesh architecture. With Lakehouse Federation, you can access a uniform data management, discovery and governance experience for all your data across various platforms, including MySQL, PostgreSQL, Amazon Redshift, Snowflake, Azure SQL Database, Azure Synapse, Google BigQuery, Oracle, and more, all within the Databricks environment without the need for traditional ETL. This enables you to establish a unified data access permission model to set and enforce data access rules, ensuring the protection of data across all sources. Implement rules like security at the row and column levels, policies based on tags, centralized auditing consistently across platforms, track data usage and meet compliance requirements with integrated data lineage and auditability features.



Lakehouse Federation

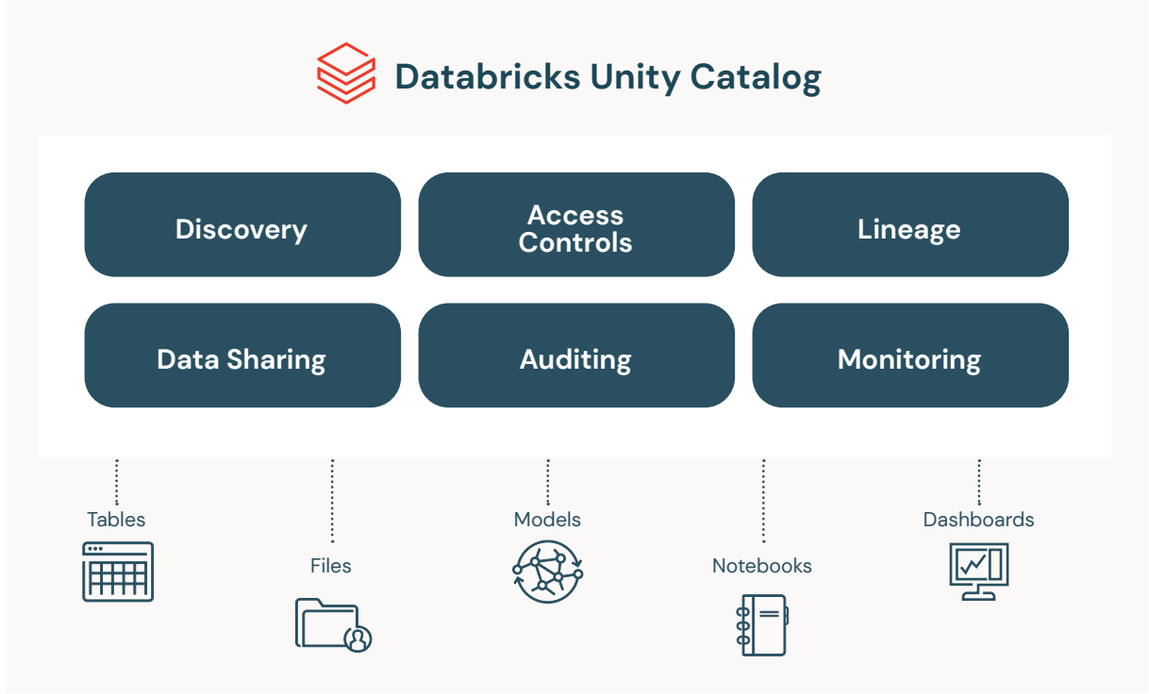
Metadata management

You cannot govern data if you do not know that it exists. Metadata is just data about data. It is data that describes your data and provides semantic value to an otherwise unidentifiable sea of information.

An enterprise's centralized metadata layer is both a representation of that organization's information architecture (how the data is organized for consumption and storage) as well as a semantic layer that describes what that data is and what it is not.

A good data governance program must be tied with a centralized metadata layer that provides a macro view across your enterprise of the data assets available and in use. This is usually organized across segregation points such as data domains, business units, applications or software development lifecycles (SDLC). This centralized metadata layer also needs visibility to a common federated view of identity in your organization so you can ensure that the right users, the right groups, have access to the right data.

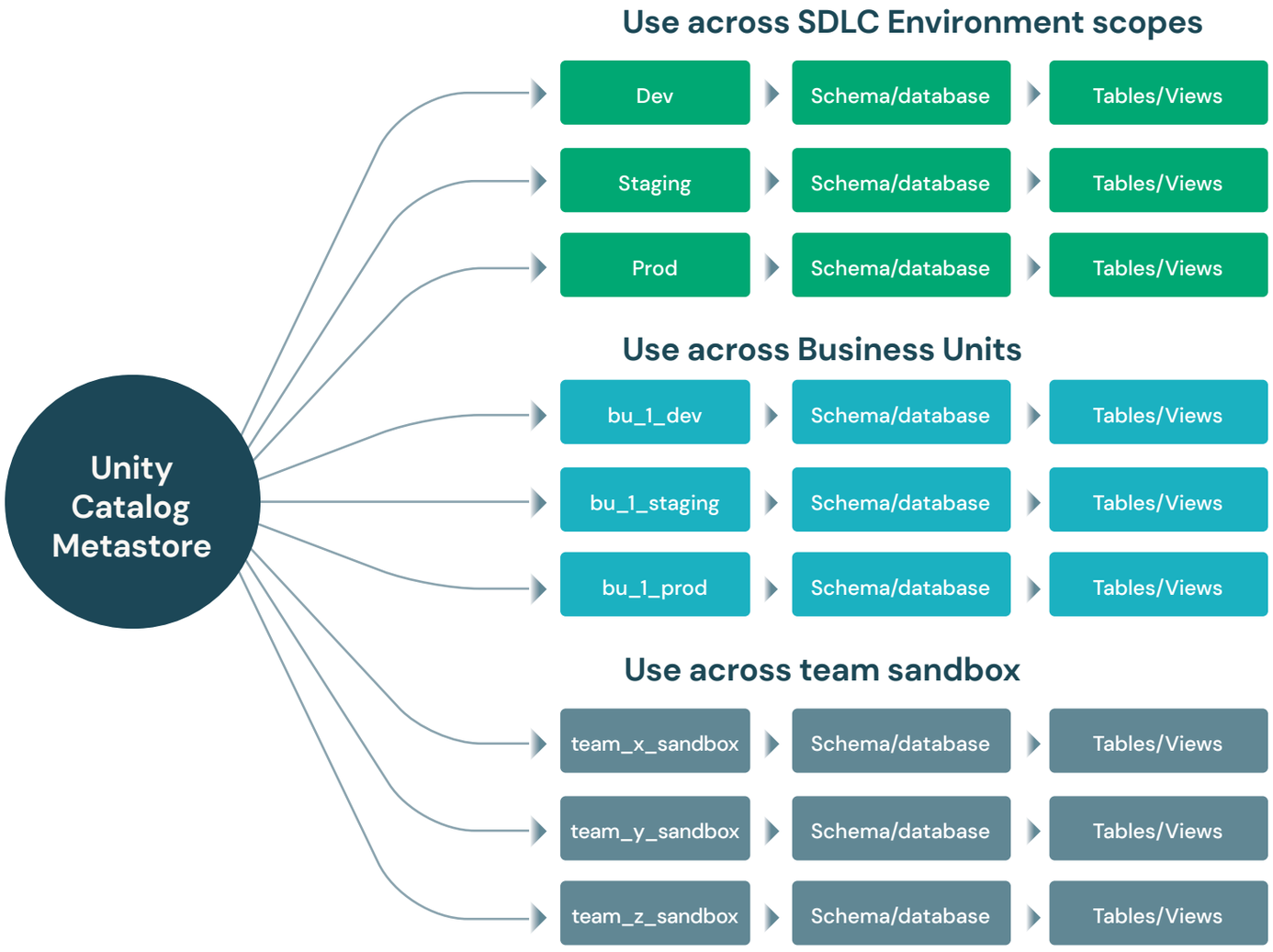
Databricks Unity Catalog offers a centralized metadata layer, called a metastore, that provides the ability to catalog and share data assets across your lakehouse, and across your enterprise's regions and clouds. These data assets could be your files, tables, dashboards, ML models, etc.



A centralized metadata layer ensures that your centralized governance teams have the ability to dictate access controls and retrieve audit information from a single place, greatly reducing organizational risk due to improperly applied access controls. Because Unity Catalog metastores are aligned to cloud provider regions, they implicitly enforce data access along regional boundaries and ensure that your users are not accessing data from other regions unless you explicitly allow it to happen. This ensures and helps you meet your complex data residency requirements.

Unity Catalog allows you to define metadata across logical hierarchies called catalogs, which can be thought of as data domains and can be tied to segregation points in your information architecture. Using access controls on catalogs and within their hierarchies, you can assign business groups to see the relevant parts of your information architecture.

The advantage of having all of your metadata in a single place is that searching for data becomes a lot simpler. Using the Unity Catalog Search features, you can search across tables and columns to find the exact data artifacts that you're looking for. This helps achieve compliance with "right to be forgotten" regulations, which require that you understand exactly what data lives in what data sets in order to search those data sets so that you can remove a specific customer's information.



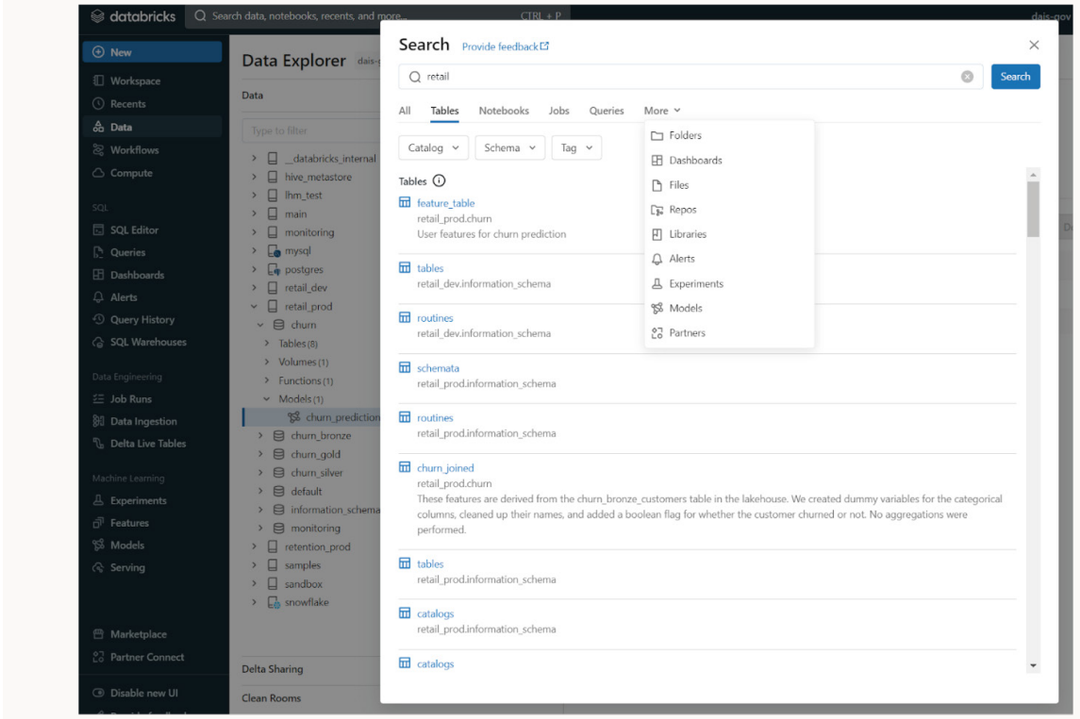
Data discovery and classification

As organizations collect massive amounts of data from various sources and in multiple formats, making the data easily discoverable for analytics, AI, or ML use cases becomes critical to accelerate data democratization and unlock the true value of the data. Per IDC, a typical data team spends around 80% of their time on data discovery, preparation and protection, while only 20% of their time is spent on actual analytics. Additionally, the influx of data isn't limited to files or tables. There are also modern data assets such as dashboards, machine learning models, queries, libraries and notebooks. Therefore, data discovery becomes a key pillar of a robust data governance strategy, enabling data teams to easily find data assets across their organization, collaborate on different projects and innovate quickly and efficiently.

Whenever a new data project is undertaken, the first instinct is to "see what data is available." Metadata must be captured, indexed and made searchable to accomplish this, and that is exactly what Unity Catalog does. Each asset cataloged within Unity Catalog undergoes automatic indexing and meticulous classification through machine learning algorithms, enhancing its discoverability. Users are further empowered to affix tags to any registered asset, amplifying the efficiency of their searches across an array of metadata fields, encompassing table names, column names, comments and distinctive tags. This comprehensive approach ensures swift and precise retrieval of the requisite data, facilitating seamless and targeted analysis.

This unified search experience in Databricks Lakehouse Platform enables users to search not just files and tables, but other assets such as notebooks, libraries, queries, dashboards, etc., resulting in better collaboration among data teams and faster innovation.

This built-in search capability automatically leverages the governance model put in place by Unity Catalog. Users will only see search results for data such as the files, tables and other assets they can access, which serves as a productivity boost for the user and a critical control for data administrators who want to ensure that sensitive data is protected. This can help prevent duplication of data across an organization. Duplicative data sets are problematic because it costs money to persist them, and because of the potential for them to be governed at different security levels.



Built-in search and discovery

Data security

A good data security governance program helps organizations:

- Achieve regulatory and/or organizational compliance over the use of data
- Provide methods to control access to data
- Provide facilities to determine who has access to data
- Provide facilities to understand who has accessed data

In this section, we'll cover some of the drivers for good data security governance, what organizational models for governance look like, and how you can align well-defined processes and technology offerings to create an effective data security governance model for your enterprise's data, specifically looking at lakehouse architectures.

Access management

A good data security governance program provides methods to control access to data. Access controls are the mechanisms that describe which groups or individuals can access what data. These are statements of policy, and they can be extremely granular and specific, right down to definitions of every record of data or file that every individual has access to. Or they can be very expressive and broad, e.g., all finance users can see all financial data.

Any statement of policy on data generally looks like this: <Principal> can/can't perform **<Action>** on **<Resource>**

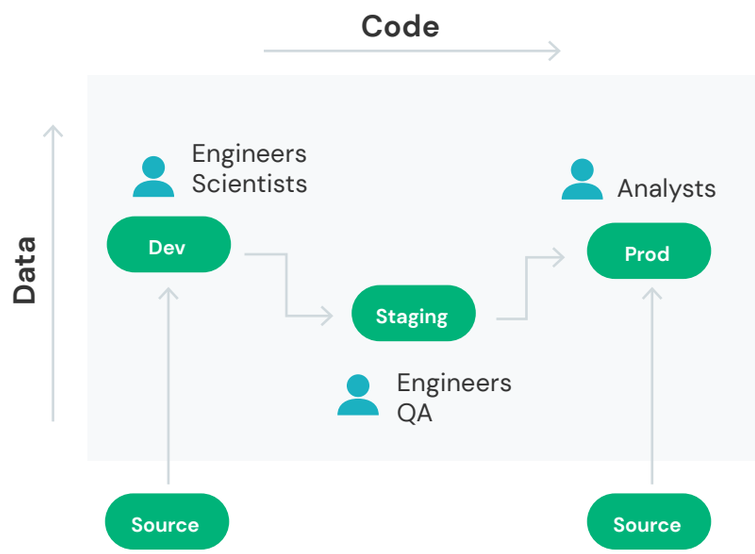
Role/resource-based access control

When you apply policy across roles or groups, you're implementing what is known as RBAC. We can also think of this as resource-based access control, as you're applying differential access controls based on resources as well. The difference is more of an implementation detail, i.e., where the access control record lives, whether on the resource that you are protecting or on the principal. Functionally, for this discussion, they are the same.

When you think about implementing RBAC on a data platform, you're in effect mixing your operational model with business groups to make sure that individuals have the right level of access to data resources at the right point in your process of sourcing, building and analyzing those data insights.

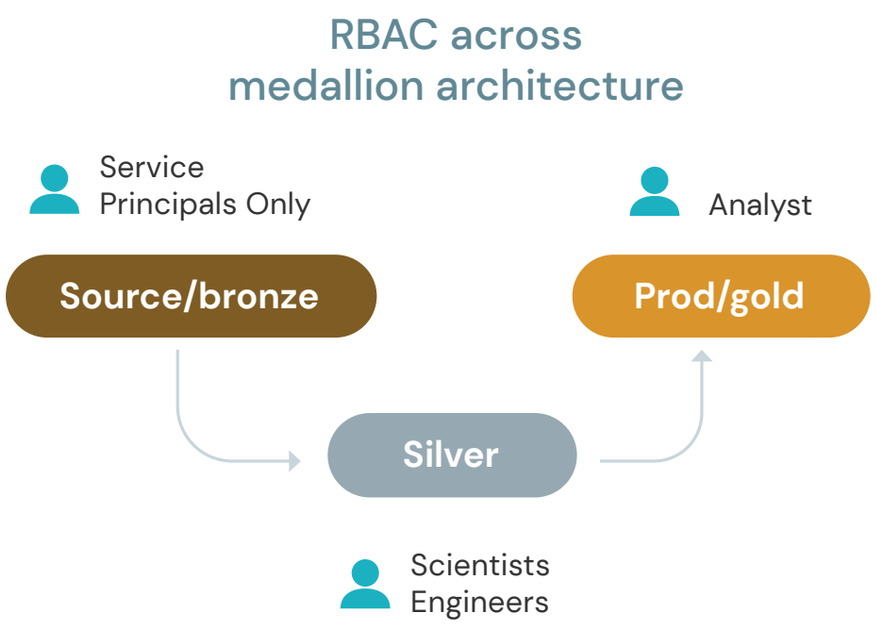
Take, for example, the different types of groups across a typical software development lifecycle.

RBAC across software development lifecycle



In the simplistic case above, you have groups of engineers, scientists and quality assurance engineers that need access to enough source data to develop jobs or insights, and analysts who need to access production data to be able to make business decisions. Assume, for a simple example, that all of your data is contained within a single catalog across environment scopes. Using groups, you can achieve differential access control ensuring that the right teams have access to the right data at the right time in the process.

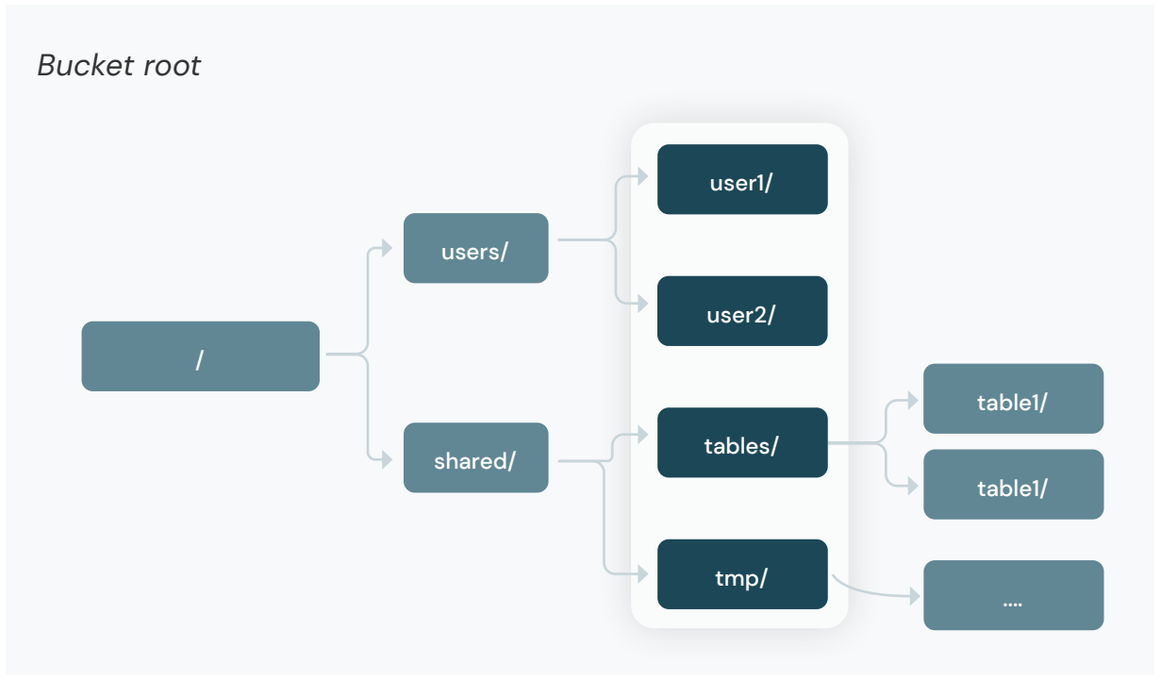
Alternatively, consider RBAC across a medallion architecture. Consider this to be a parallel for any multi-hop architecture, whether you are working with production data throughout the pipeline like a medallion architecture, or you are working with subsets of anonymized or de-identified data throughout the pipeline in the case of an SDLC multi-hop architecture.



Some data platforms and data lakes offer some form of RBAC across coarse-grained objects like tables or files, but not across both. To offer RBAC across file systems, you need the common metadata layer we discussed before.

Remember from earlier that data is more than just registered tables in your catalog; it lives in files, in directories on your data lake and as objects that you need to manage in your data platform.

Take the following example of a layout of a cloud storage container:



Here you have a mix of objects living in the same cloud storage bucket, user or team directories, and places where versioned data may live. Providing access control over this via cloud controls is limited and difficult to manage at scale. To compound that problem, the interfaces to apply access control lists (ACLs) are not homogenous as you move across different cloud providers in the pursuit of multicloud architectures. So you end up creating and managing controls in different places, which relies on costly resources or automation to keep things in sync. To make it worse, because each system works differently, you have to map the intent of each policy you write to every downstream system, which can be an integration challenge even if working with a data governance toolkit.

When you factor in the centralized metadata and centralized user management components offered by Unity Catalog, a data governance user in your enterprise can run simple ANSI SQL GRANT statements on objects to provide access to a group and manage the membership of that group externally. This lets members of the group dictate who has access to SELECT or MODIFY data in your catalog and across your data lakes.

Unity Catalog also offers the same capabilities via REST APIs and Terraform modules to allow integration with existing entitlement request platforms or policies as code platforms. These objects can also be file storage paths in your cloud storage containers.

In the example above, consider the encircled objects to be external locations within Unity Catalog. This allows you to provide differential access control on various parts of your container hierarchy, ensuring that users have access to their sub-directories and that groups have access to read from the shared locations.

Attribute-based access control

Applying access controls on resources and roles can be exceedingly complex when you have thousands of resources and thousands of groups. Attribute-based access control, or ABAC, offers a simplified way to dictate access control based on the semantics of your data. You apply semantic definition by associating attributes, known as tags, to resources, and also users and groups.

This means you can write simpler and more expressive access controls that align well with the business language of enterprise-wide data access policies.

For example, you have a hundred tables with a customer name field, and to protect customer information, only certain business groups are allowed to see this data. Using traditional RBAC, for every business group you must have a corresponding policy for each table. If you have a hundred groups and a hundred tables, this could easily mean 10,000 policies, depending on how you have nested group hierarchies to deal with this level of complexity.

Implementing effective security controls at scale requires that the process be efficient and maintainable so that you do not have an overly large team working to keep your controls in place.

Unity Catalog offers the ability to provide access control on data via attributes using ABAC. This will give data governance professionals the ability to articulate enterprise-wide data policies on certain types of data, without sacrificing individual data owners' ability to dictate governance with role/resource-based access control.

Auditing entitlements and access

Earlier, we discussed two tenets of an effective data security governance program: understanding who has access to what data, and who has recently accessed what data. This is critical for almost all compliance requirements for regulated industries, and a fundamental facet of any security governance program.

Without effective audit mechanisms in place, you do not understand your risk surface area. That is to say, you do not know who can potentially misuse data within your organization. Being able to proactively identify overentitled users and groups and change their accesses accordingly in a centralized manner is a key function of a well-designed audit team within a security or data governance organization.

An effective audit mechanism functionally provides three key pieces of information:

- For any resource, indicate who had access at any point in time
- For any access to a resource, indicate the resource that performed the operation
- For any access to a resource, indicate the operation performed by the resource

Achieving this with existing cloud security controls is a task that requires ETL from access logs, centralization across thousands of different sources and federation across multiple formats. Depending on the capabilities of the source systems, this typically only gets you the first two key pieces of information.

Unity Catalog offers a centralized audit feature that homogenizes access across your metastore, including access to logical objects as well as access to arbitrary cloud storage paths in your data lakes. In addition, Unity Catalog helps you to understand the specific operations, i.e., queries that your users are performing. Unity Catalog also offers a comprehensive UI, APIs as well as system tables that allow you to query the ACLs on resources, so you can effectively audit access by resources or by users to understand audit implications in real time, without waiting for systems to be in sync, or for audit data to be shipped and consolidated.

Combining this information across cloud architectures becomes simpler because you are operating in a common format, and audit data can be readily combined for aggregate views across multiple metastores in different clouds and cloud regions.

Data lineage

Today, organizations deal with an influx of data from multiple sources, and building a better understanding of where data is coming from and how it is consumed becomes paramount to ensure the quality and trustworthiness of the data. Data lineage is a powerful tool that helps data leaders drive better transparency and understanding of data in their organizations. The term "data lineage" describes the transformations and refinements of data from source to insight. Lineage includes capturing all the relevant metadata and events associated with the data in its lifecycle, including the source of the data set, what other data sets were used to create it, who created it and when, what transformations were performed, what other data sets leverage it, and many other events and attributes. With a data lineage solution, data teams get an end-to-end view of how data is transformed and how it flows across their data estate.

As more and more organizations embrace a data-driven culture and set up processes and tools to democratize and scale data and AI, data lineage is becoming an essential pillar of a pragmatic data governance strategy.

To understand the importance of data lineage, we have highlighted some common use cases:

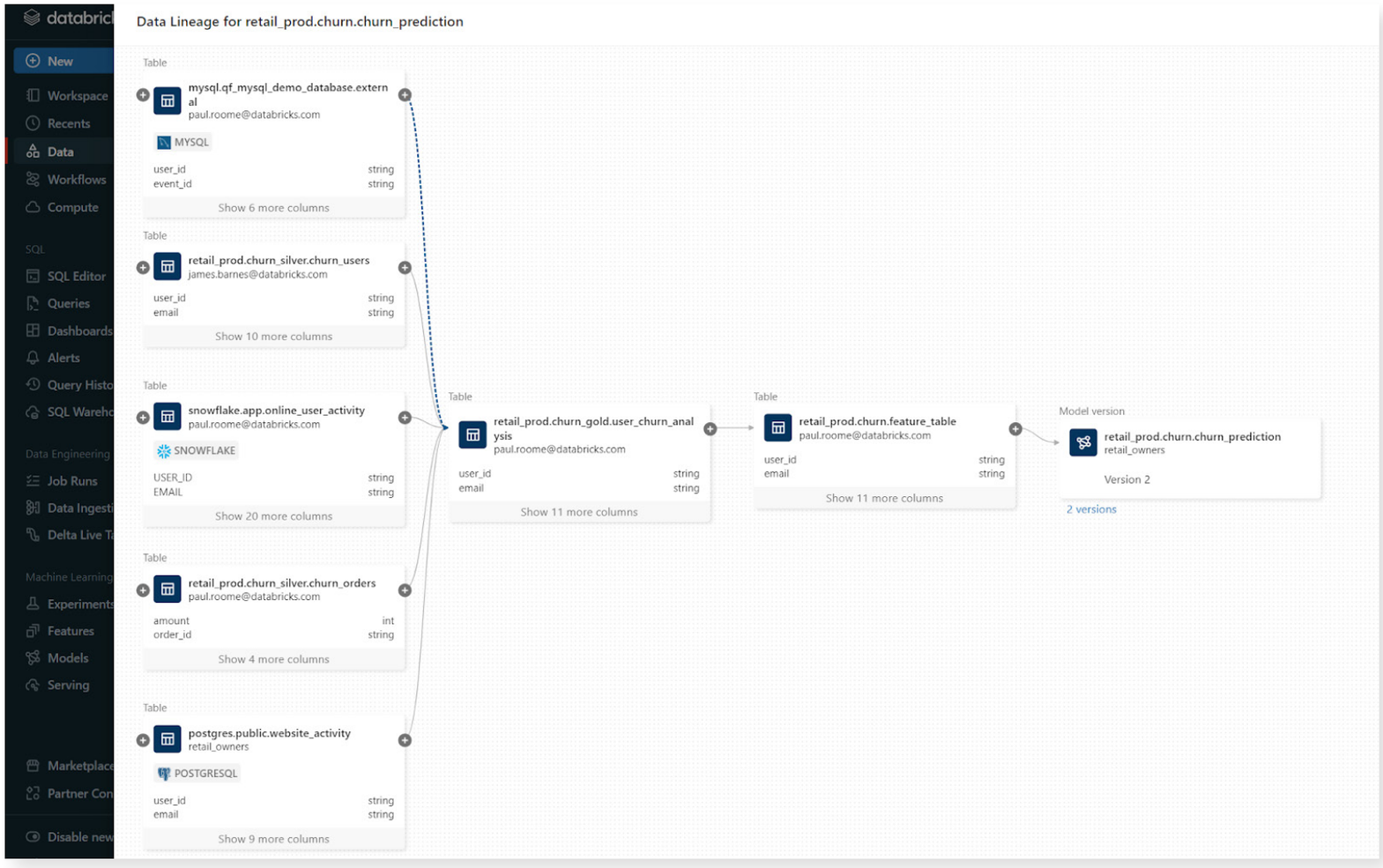
Compliance and audit readiness
Many compliance regulations, such as the General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), Health Insurance Portability and Accountability Act (HIPPA), Basel Committee on Banking Supervision (BCBS) 239, and Sarbanes-Oxley Act (SOX), require organizations to have a clear understanding and visibility of their data flow, and prove the data/reports they are submitting came from a trusted and verified source. This typically means identifying the tables and data sets used in a report or dashboard and tracing the source of these tables and fields, making data traceability a key requirement for their data architecture. Data lineage helps organizations be compliant and audit-ready, thereby alleviating the operational overhead of manually creating the trails of data flows for audit reporting.

Impact analysis/change management
Data goes through multiple updates or revisions over its lifecycle, and understanding the potential impact of any data changes on downstream consumers becomes important from a risk management standpoint. With data lineage, data teams can see all the downstream consumers — applications, dashboards, machine learning models, data sets, etc. — impacted by data changes. They can see the severity of the impact, and notify the relevant stakeholders. Lineage also helps IT teams proactively communicate data migrations to the appropriate teams, ensuring business continuity.

Data quality assurance
Data lineage also empowers data consumers such as data scientists, data engineers and data analysts to be context-aware as they perform analyses, resulting in better quality outcomes. Data stewards can see which data sets are no longer accessed or have become obsolete to retire unnecessary data and ensure data quality for end business users.

Debugging and diagnostics
You can have all the checks and balances in place, but something will eventually break. Data lineage helps data teams perform a root cause analysis of any errors in their data pipelines, applications, dashboards, machine learning models, etc., by tracing the error to its source. This significantly reduces the debugging time, saving days, or, in many cases, months of manual effort.

Unity Catalog provides automated and real-time lineage by automatically capturing lineage generated by operations executed in Databricks. Lineage is not limited to just SQL. It works across all workloads in any language supported by Databricks — Python, SQL, R and Scala. This empowers all personas — data analysts, data scientists and ML experts — to augment their tools with data intelligence and the context surrounding the data, resulting in better insights. Unity Catalog captures data lineage for tables, views, columns, and files. This information is displayed in real time, providing data teams a granular view of how data flows both upstream and downstream from a particular table or column in the Databricks Lakehouse Platform with just a few clicks. Unity Catalog can also capture lineage associated with non-data entities, such as notebooks, workflows and dashboards. This helps with end-to-end visibility into how data is used in your organization.



Data lineage in Unity Catalog

Data quality management and monitoring

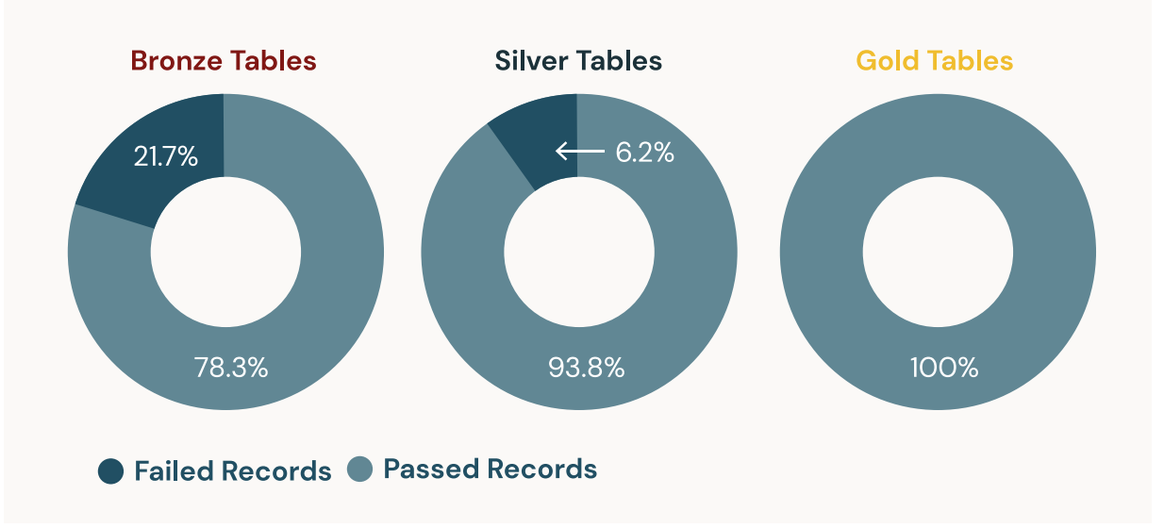
Data needs to be trusted to harness its true value. Bad-quality data leads to inaccurate analytics, poor decision-making and cost overhead. Gartner estimates that poor data quality costs organizations an average $12.9 million every year. An effective data governance strategy must include a focus on data quality so that the provenance of data can be known, rules can be enforced on the data, and changes can be tracked.

When evaluating a data set for quality, a few key attributes should be examined:

- Is the data accurate in every detail? Is it complete?
- Where is it coming from?
- How fresh is the data?
- Does the data set violate any data quality rules?

With automated and real-time data lineage, Unity Catalog makes it easy to quickly understand who owns a data set and where its data originated. Data lineage is automatically captured for Unity Catalog data sets. A user can easily traverse the lineage to see both the upstream data sources that were used to generate the data set and the downstream consumers in the form of other tables, notebooks, dashboards, etc.

In addition to being able to view the provenance of a data set, it is important to understand if the data within the table has violated any rules provided by domain experts. Delta Live Tables (DLT) performs checks against data during runtime when processing the data. These checks are called *expectations*, and are aggregated and viewable for each DLT table. Delta Live Tables helps to ensure accurate and useful BI, data science and machine learning with high-quality data for downstream users. Prevent bad data from flowing into tables through validation and integrity checks and avoid data quality errors with predefined error policies (fail, drop, alert or quarantine data). In addition, you can monitor data quality trends over time to get insight into how your data is evolving and where changes may be necessary.



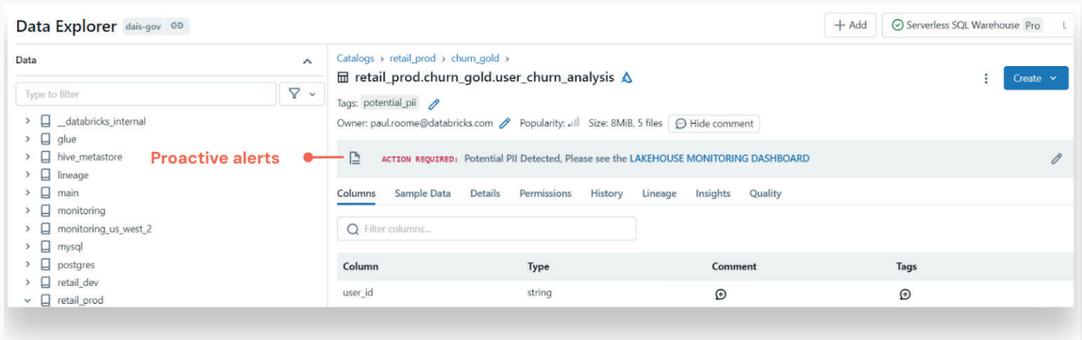
Automatic data quality testing with Delta Live Tables

In the realm of data and ML quality, despite the implementation of meticulous checks and balances, the inevitability of occasional disruptions remains. Therefore, data teams must embrace the task of monitoring data quality over time. Consider a scenario in which a data engineer overlooks utilizing the "update" command on a table, leading to the accumulation of numerous rows with each data load. Consequently, end data users encounter obstacles when querying the data for their specific use cases. This issue often only surfaces when a data user reports it. Similarly, envision a situation where a reference table misses values in a rows following an update, causing downstream queries to fail when merging it with another table. Additionally, contemplate the instance where the performance and accuracy of a trained machine learning model erodes as the production data distribution diverges from the distribution on which the model was trained. This situation is compounded when application development teams deactivate these production models. As a result, data teams grapple with the following challenges:

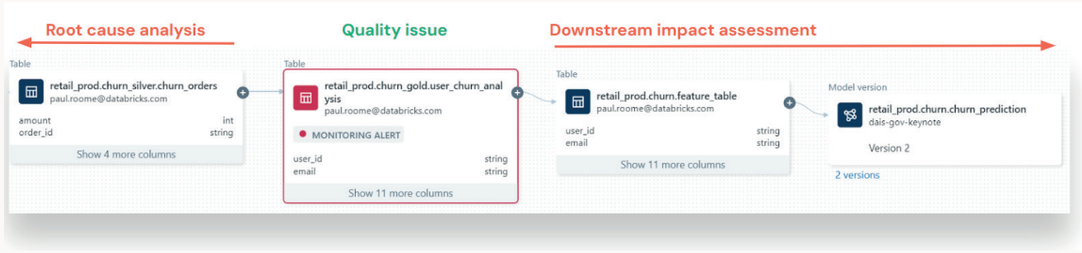
- **Reactive issue management:** Data consumers encounter problems before data owners, leading to inefficiencies in resolution
- **Bottlenecked operations:** Data consumers are deprived of self-service experiences, necessitating their reliance on data engineering teams to rectify quality issues
- **Inefficient processes:** Debugging quality issues and reverting changes can consume weeks, impacting overall operational efficiency

Existing monitoring solutions often pose hurdles. Setting up monitors, crafting dashboards and configuring alerts demand substantial time and resources. Furthermore, data teams employ diverse quality metrics and frameworks, which deter them from adopting multiple solutions and manually creating isolated dashboards for custom metrics. The predicament escalates as data teams manage segregated monitoring tools for data scientists and data engineers, impeding collaboration and debugging efforts.

To confront these challenges, the Databricks Lakehouse Platform introduced Lakehouse Monitoring—an integrated solution for overseeing data and ML model quality. This holistic approach empowers users to configure monitoring parameters for both data and ML models. Proactive alerts are triggered when data quality metrics, such as the percentage of null values in a table, fall below predefined thresholds, or when sensitive data is detected in unmasked form, or even in the presence of significant prediction drift in the ML model. The solution also encompasses auto-generated dashboards. Leveraging the lineage information within Unity Catalog, data teams can conduct root cause analysis and assess the impact of quality issues, thereby enhancing their ability to address and mitigate disruptions effectively.



Proactive alerts on quality issues



Root cause analysis and impact assessment with lineage

databricks

Data intelligence

As organizations accumulate vast amounts of data and diverse assets like ML models, notebooks and dashboards, distilling intelligence from the data landscape becomes crucial. This opportunity not only enables the democratization of data and AI for both technical and non-technical audiences in a scalable manner but also has the potential to accelerate business efforts and create a distinct competitive advantage.

LLMs have heralded the promise of instating language interfaces for data, and the integration of AI assistants by data companies is becoming commonplace. However, in reality, a significant number of these solutions fall short of their intended objectives. Each organization possesses unique datasets, industry-specific jargon, and internal knowledge that are essential for addressing its distinct business inquiries. The mere deployment of an LLM trained on web data to answer questions unique to the business often yields erroneous outcomes. The simplest terms, like “customer” or the fiscal year, can exhibit discrepancies across different companies.

To address this critical need, Databricks introduced LakehouseIQ, an innovative knowledge engine that directly confronts these challenges by autonomously assimilating knowledge about business and data concepts within your enterprise. LakehouseIQ leverages insights from the complete Databricks Lakehouse platform, including Unity Catalog, dashboards, notebooks, data pipelines and documentation. With the aid of this knowledge graph, it discerns the practical application of data, enabling the construction of highly accurate specialized models tailored to your organization.

LakehouseIQ

LakehouseIQ leverages data, usage patterns, and the organizational structure to grasp your business’s specific terminology and unique data environment. This results in markedly improved responses compared to the naive application of LLMs. Any user in your organization, whether technical or non-technical can utilize LakehouseIQ to search, understand and extract insights from the data in natural language. LakehouseIQ's impact is not limited to enhancing queries and troubleshooting; its functionalities extend to a diverse range of natural language interfaces across Databricks. Furthermore, its capabilities are accessible through APIs, granting your customers the power to create their own AI applications that harness this autonomously cultivated knowledge.

Data sharing and collaboration

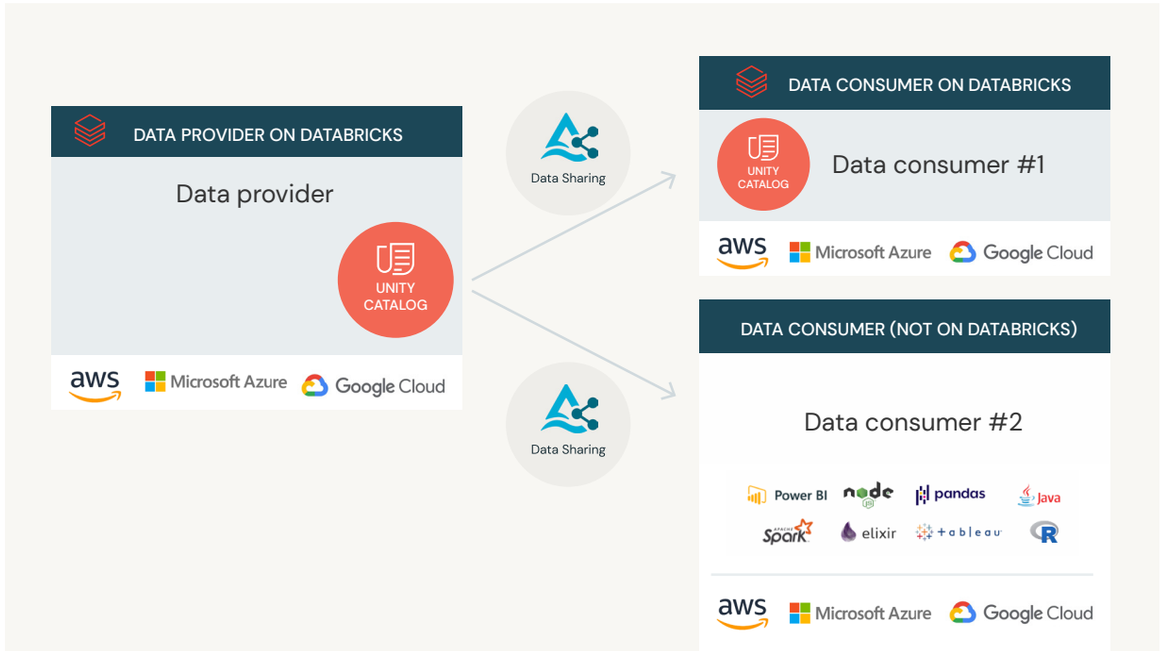
Today's economy revolves around data. Data sharing has a key role to play in business processes across the enterprise, from product development and internal operations to customer experience and compliance. More and more, organizations must exchange data with their business units, customers, suppliers and partners. Security is critical. And yet, efficiency and immediate accessibility are equally important. To be truly data-driven, organizations need a better way to share data. Gartner predicts that by 2023, organizations that promote data sharing will outperform their peers on most business value metrics. Where data sharing was once considered optional, it is now fundamental. More organizations are investing in streamlining internal and external data sharing across the value chain. But they still face major roadblocks — from human inhibition to legacy solutions to vendor lock-in.

We believe the future of data sharing should be characterized by open technology. Data sharing shouldn't be tied to a proprietary technology that introduces unnecessary limitations and financial burdens to the process. This philosophy inspired us to develop and release a new protocol for sharing data: Delta Sharing.

Delta Sharing provides an open solution to securely share live data from your lakehouse to any computing platform. Recipients do not have to be on the Databricks platform or on the same cloud or a cloud at all. Data providers can share live data from where it lives in their cloud storage without replicating it or moving it to another system. This approach reduces the operational cost of data sharing, as data providers don't have to replicate the data multiple times across clouds, regions or data platforms to reach their data consumers.

With Delta Sharing, organizations can easily share existing large-scale data sets based on the open source formats Apache Parquet and Delta Lake. Delta Sharing also provides a robust list of native connectors such as pandas, Apache Spark, Java, Power BI, Python, Tableau, R and many more, providing data consumers the flexibility to query, visualize, transform, ingest or enrich shared data with their tools of choice, if they are not on Databricks.

Delta Sharing is natively integrated with Unity Catalog, enabling organizations to centrally manage and audit shared data across organizations and confidently share data assets while meeting security and compliance needs.



Data sharing with Databricks

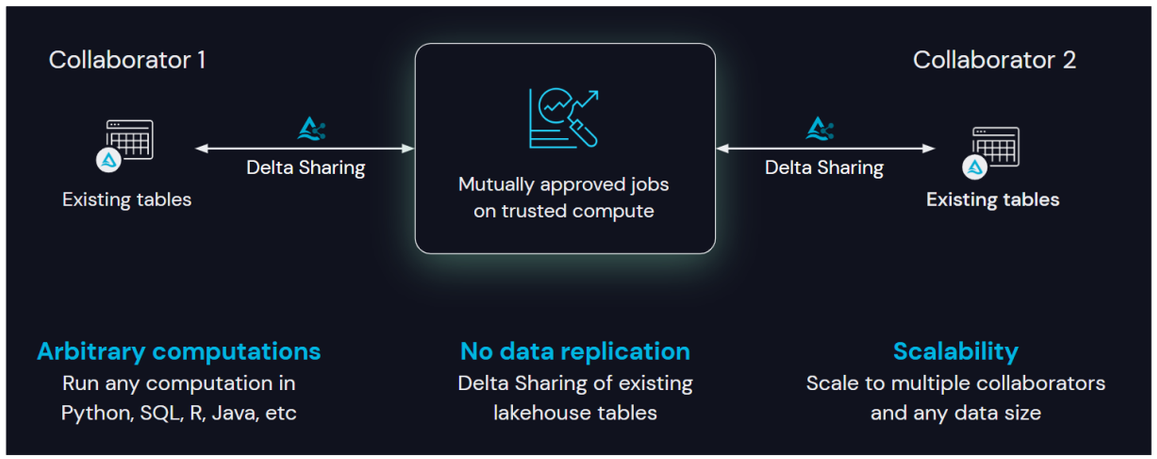
Privacy-safe data clean rooms

With the demand for external data greater than ever, organizations are looking for ways to securely exchange their data and consume external data to foster data-driven innovations. Historically, organizations have leveraged data-sharing solutions to share data with their partners and relied on mutual trust to preserve data privacy. But the organizations relinquish control over the data once it is shared and have little to no visibility into how data is consumed by their partners across various platforms. This exposes potential data misuse and data privacy breaches. With stringent data privacy regulations, it is imperative for organizations to have control and visibility into how their sensitive data is consumed. As a result, organizations need a secure, controlled and private way to collaborate on data, and this is where data cleanrooms come into the picture. In fact, IDC predicts that by 2024, 65% of G2000 enterprises will form data-sharing partnerships with external stakeholders via data clean rooms to increase interdependence while safeguarding data privacy.

A data clean room provides a secure, governed and privacy-safe environment in which multiple participants can join their first-party data and perform analysis on the data without the risk of exposing their data to other participants. Participants have full control of their data and can decide which participants can perform what analysis on their data without exposing any sensitive data such as personally identifiable information (PII).

Data clean room

Powered by open source Delta Sharing, the Databricks Lakehouse Platform provides a flexible data clean room solution allowing businesses to easily collaborate with their customers and partners on any cloud in a privacy-safe way. Participants in the data clean rooms can share and join their existing data and run complex workloads in any language – Python, R, SQL, Java and Scala – on the data while maintaining data privacy. Additionally, data clean room participants don't have to do cost-intensive data replication across clouds or regions with other participants, which simplifies data operations and reduces cost.



Data clean room with the Databricks Lakehouse Platform

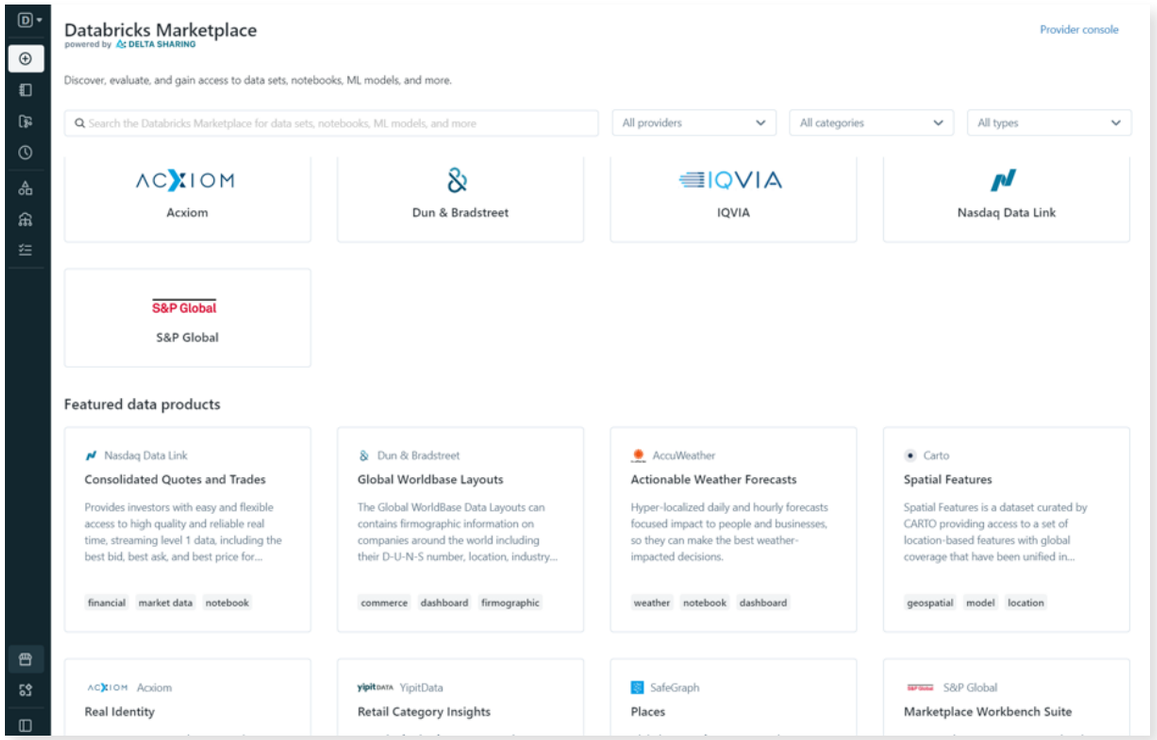
Data Marketplaces

The demand for third-party data as a catalyst for data-driven innovations has reached unprecedented levels. In this landscape, data marketplaces have emerged as pivotal connectors between data providers and consumers, facilitating the seamless exploration and dissemination of diverse data sets.

In parallel, the surge in Generative AI's prominence has triggered substantial industry disruption. Organizations are now under heightened pressure to swiftly engineer their generative AI models and LLM frameworks, leveraging proprietary data reserves to gain a competitive edge and carve a distinctive niche in their offerings.

At the epicenter of this transformative paradigm, the Databricks Marketplace shines as an exemplar. It transcends conventional data exchange platforms by encompassing a broader gamut, encompassing not only data but also notebooks, dashboards and machine learning models. This orchestration empowers data consumers with expedited insights, granting them access to, evaluation of, and engagement with an expansive repertoire of data products originating from a diverse array of third-party vendors.

The ingenuity of this platform extends further, empowering data providers to infuse novel value into their offerings, thereby shortening sales cycles and introducing novel value-added services anchored in their data products. At the core of this marketplace's potency lies its infrastructure's backbone: Delta Sharing. This architectural innovation seamlessly facilitates data product access for consumers, alleviating the necessity of platform lock-in. This open framework fosters an expanded ecosystem, enabling data providers to broaden their market horizon while liberating consumers from vendor-specific constraints.



Databricks Marketplace

Learn more

To learn more about data sharing on Databricks, please read this free eBook on data sharing.

databricks

Chapter 2
AI Governance

The principles of governance — accountability, standardization, compliance, quality and transparency — apply at least as much to AI as to data. AI governance sets out the policies and procedures for the development and application of models in an organization. With proper governance, ML can bring improved value to business processes, efficiency in automating or augmenting decision-making and decreased regulatory, legal and reputation risks.

Unlocking consistent value from ML investments requires that models are:

- Compliant with necessary regulatory and ethical standards
- Reproducible in terms of performance and results
- Explainable and transparent in their development and impact
- Routinely monitored and updated to maintain quality
- Properly cataloged and documented following standardized policies

AI governance on the lakehouse

Compliance and ethics

The minimum bar that ML governance must set is that any models used need to conform to regulations pertinent to the industry or application that they serve. For example, in the United States, the financial services and education sectors are subject to legislation dictating what types of inputs may be included in the development and training of models. These standards are meant to ensure that protected classes of the population receive equitable access to lending, housing, education and other crucial services.

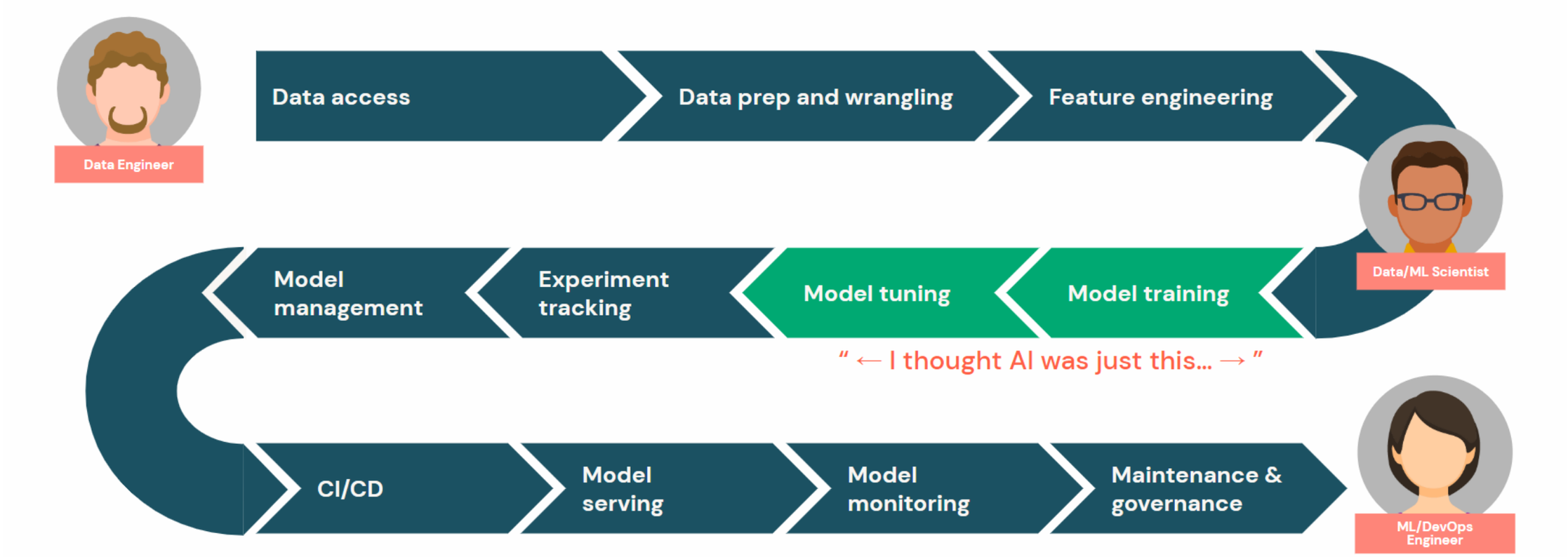
When establishing your ML governance program and policies, consult with legal counsel to ensure that you have a thorough understanding of the regulations you must operate under in your jurisdiction(s). They may vary based on the application. For instance, a retailer operating a private label credit card is subject to the Equal Credit Opportunity Act just as a financial services company would be, despite mostly being in a different industry.

Many of these regulations focus on what types of data inputs (independent variables) may be used when training models. Using Unity Catalog for fine-grained control of data is a secure way to programmatically restrict what can be used as inputs by which teams to ensure compliance.

However, compliance only covers what must be done. Beyond the legal requirements are a host of potential ethical considerations that constitute what should be done. It is in this area where many examples of well-intentioned but poorly governed models have caused reputational headaches for companies of all sizes. A process should be defined as part of your governance program to assess model impacts and identify potential misuses before moving to production.

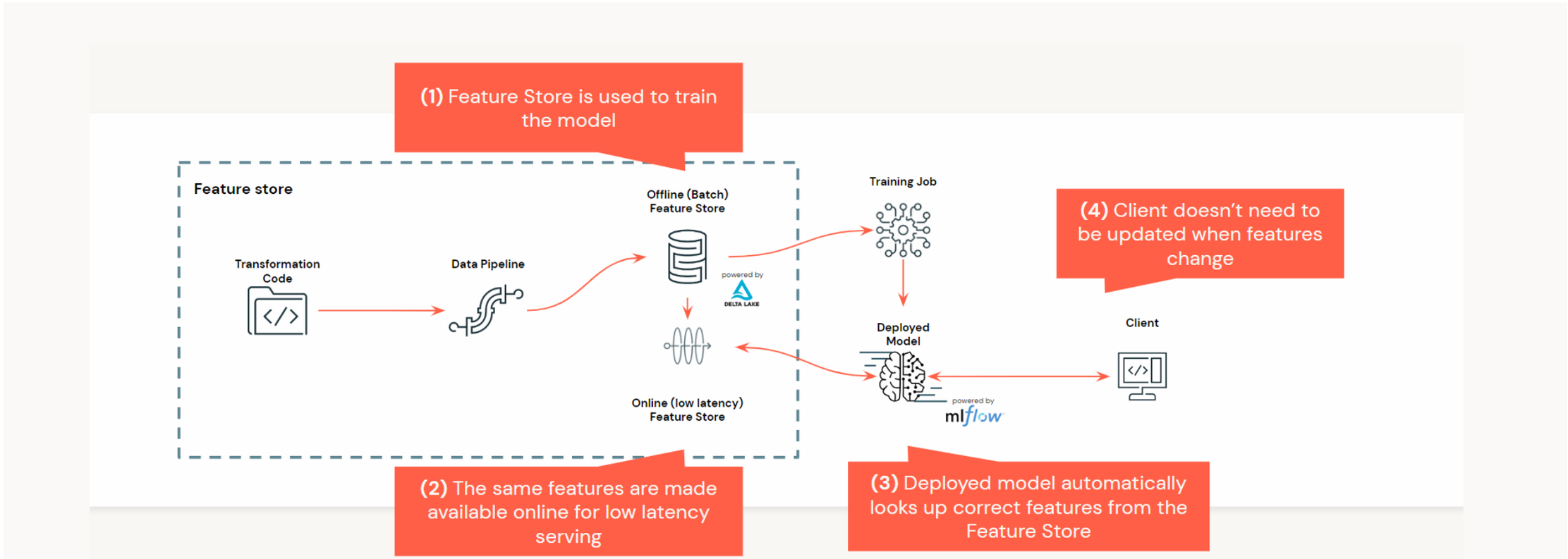
Machine learning reproducibility

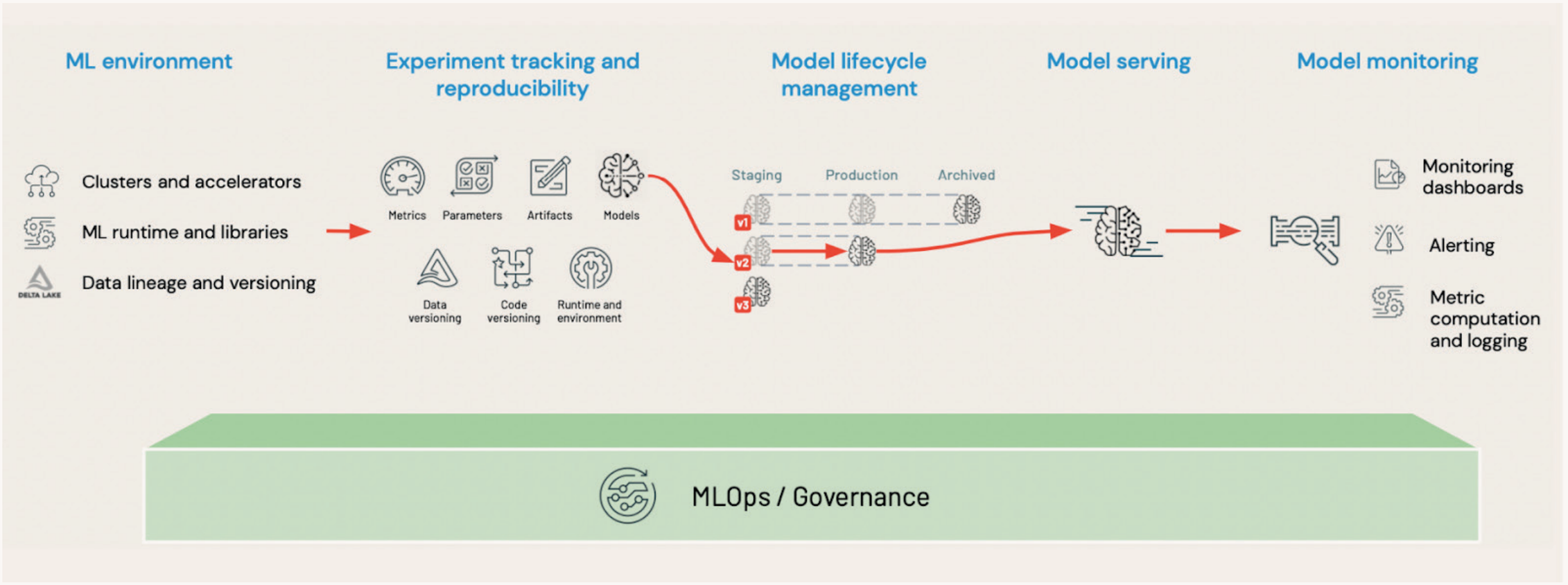
Being able to reproduce the results of machine learning allows for auditing and peer review of the data and methodology used for creating the final model. This improves trust in the performance it will have in production and the ability to meet compliance standards. Due to the multifaceted nature of model development, there are multiple stages of the process that must be tracked in order for the entire study to be reproduced:



1. Feature engineering — The variables, labels and encodings used as inputs to the model should be stored in an accessible, reusable manner. Databricks includes a Feature Store for this purpose that can be referenced in model development. The Feature Store provides lineage tracing tied to your broader data governance so that upstream or downstream impacts from any changes are visible. It also ensures reproducibility of the logic used for feature calculation between the training of the model in a batch mode and the serving of an inference pipeline in a streaming mode without having to create and maintain a separate workflow, which could otherwise introduce discrepancies.

2. Data access, prep and wrangling — Sampling and splitting data during model development is a common way to create and then test models for extensibility. Once a model has been selected, the data that was used to initially train it is often lost, causing problems in trying to recreate how the parameters of the model were determined. When using Managed MLflow on Databricks, data sets are logged alongside experiments so that the end-to-end development cycle of machine learning models can be reproduced easily.





3. Model training — Most machine learning models selected for production are not the first iteration that the engineers developed. Instead, extensive testing and evaluation has gone on to select appropriate methodologies and frameworks, tune parameters and hyperparameters, and then balance speed of execution with optimization of model evaluation metrics. MLflow captures all of these iterations along with the pertinent metrics so that the entire process of training and selecting a model can be reproduced if needed. And with Unity Catalog, ML engineering teams can drive better collaboration as data can be securely accessed across workspaces for model training.

Explainability and transparency

Data governance is involved in explaining and providing transparency into what data is used where and to what end. Similarly, a part of ML governance is concerned with giving insight into how models come up with the conclusions that they do.

However, interpreting machine learning models can be challenging, especially when more advanced types of algorithms are used in their development. Some of the most commonly used frameworks, such as XGBoost or neural networks, are considered too complex for humans to conceptualize. In order to shine a light into the "black box," you need tools for evaluating what has happened inside the box.

As part of MLflow, Databricks integrates bias detection and explainability using Shapley Additive Explanations (SHAP). SHAP provides views of which features played the largest roles and which directions led to the outputs scored. SHAP is a broadly applicable tool for this purpose and can be used on many different types of algorithms. It may be applied either globally (e.g., it looks at the entire population of scores and evaluates for biases and feature impact) or locally (e.g., evaluating the score of one record and determining which features played significant roles in that one outcome).

Many additional frameworks are available for assessing bias and providing transparency into the inner workings of models. These may be added to clusters operating on a Databricks ML Runtime as needed to enhance the explainability of models developed.

Model monitoring

Governance does not end once machine learning models get to production. If anything, increased scrutiny should be placed on models as they are used in "real-world" scenarios and are exposed to new situations. Traditionally, models that were developed using data science and machine learning techniques were expected to degrade in performance and applicability over time. Even with online learning capabilities, models sometimes span shifts in the market or business practices that necessitate revisiting their training or use. It requires continual monitoring to be able to trust and get long-term value from models being used in production.

Some of the concerns addressed around monitoring models include:

Concept drift — This represents a change in the possible results that the model could generate. Concept drift can come from changes in the market, technology, business processes or other external factors.

An example of this could be a sudden allowal of manual review for certain groups of loan applications. If, when the model was developed, the possible outcomes were "approve" or "decline," then the model would only ever classify into these two categories. If subsequently, the business decides to create a team that will manually review loans that do not quite meet the approval threshold but are not terribly far off, the model could be trained for a third option to classify some range of loans into this group.

Upstream data changes — This is broadly thought of as breaks in the data pipelines that generate the features that are used in the model. Using proper lineage tracking with Unity Catalog and Feature Store on Databricks should prevent most outages or breaking impacts. However, it is still possible that a business process change could cause significant problems in a model.

An example of this might be seen in how income is bucketed and encoded as variables. Let's assume that, at the time of model development, there were increments of $25,000 up to $150,000 and an "Other" bucket, which could be selected from a dropdown menu in a customer service screen. Then a change was made to use increments of $30,000 up to $150,000 and keep the Other bucket. Because the old values have disappeared, the new values may not be properly encoded. The result would be a sudden loss of meaningful information and a massive shift in income levels showing as "Other." No actual break in the pipeline would have happened, but the upstream process change could cause the model to behave erratically if income was a significant feature.

Data drift — Data changes over time, even without intentional changes made in upstream processes. Economic conditions, market trends, seasonality, cultural shifts — anything that changes behavior will alter the data collected on that behavior. This is probably the most common cause of model degradation in production and the area given most scrutiny. Data drift can cause a slow and gradual decline or very sudden shocks.

As an example, take a manufacturer that has developed a demand forecasting model that can help ensure that materials are in factories at the right time to make goods and get them to shelves. Then their largest competitor declares bankruptcy and there is a sudden surge in demand. No data on this type of rare event was included in the model, so the performance of the forecast is horribly off. An alert should be put in place for large swings like this so that a new model can be developed taking into consideration the new competitive landscape.

Bias — In the context of machine learning, bias is typically seen in the statistical sense, meaning that there is a mathematically demonstrable imbalance in the data seen versus some ground truth. This could be seen as an imbalance of input to the model (e.g., you know that 50% of your customers are located in Europe, but only 15% of the data coming into the model are from European customers) or outputs (e.g., you expect that 2% of your customers will purchase socks on a given visit, but the recommendation engine you built shows socks to 80% of customers as a top selection for them).

In common parlance, bias is often seen as the inequitable treatment of one group of people versus another. This is one of the most essential forms of bias to monitor for in model performance, as it could cause substantial damages to customer experience and reputation and even have legal ramifications. If you have demographic data available, it is advisable to create a governance policy around checking for disparate impacts across various cross-sections of demographics even if you are not operating in a highly regulated industry where this is required.

As part of your ML governance program, establish guidelines for what thresholds are acceptable for model performance, a cadence for model monitoring, and procedures for alerting when upstream processes are changed or problems are found.

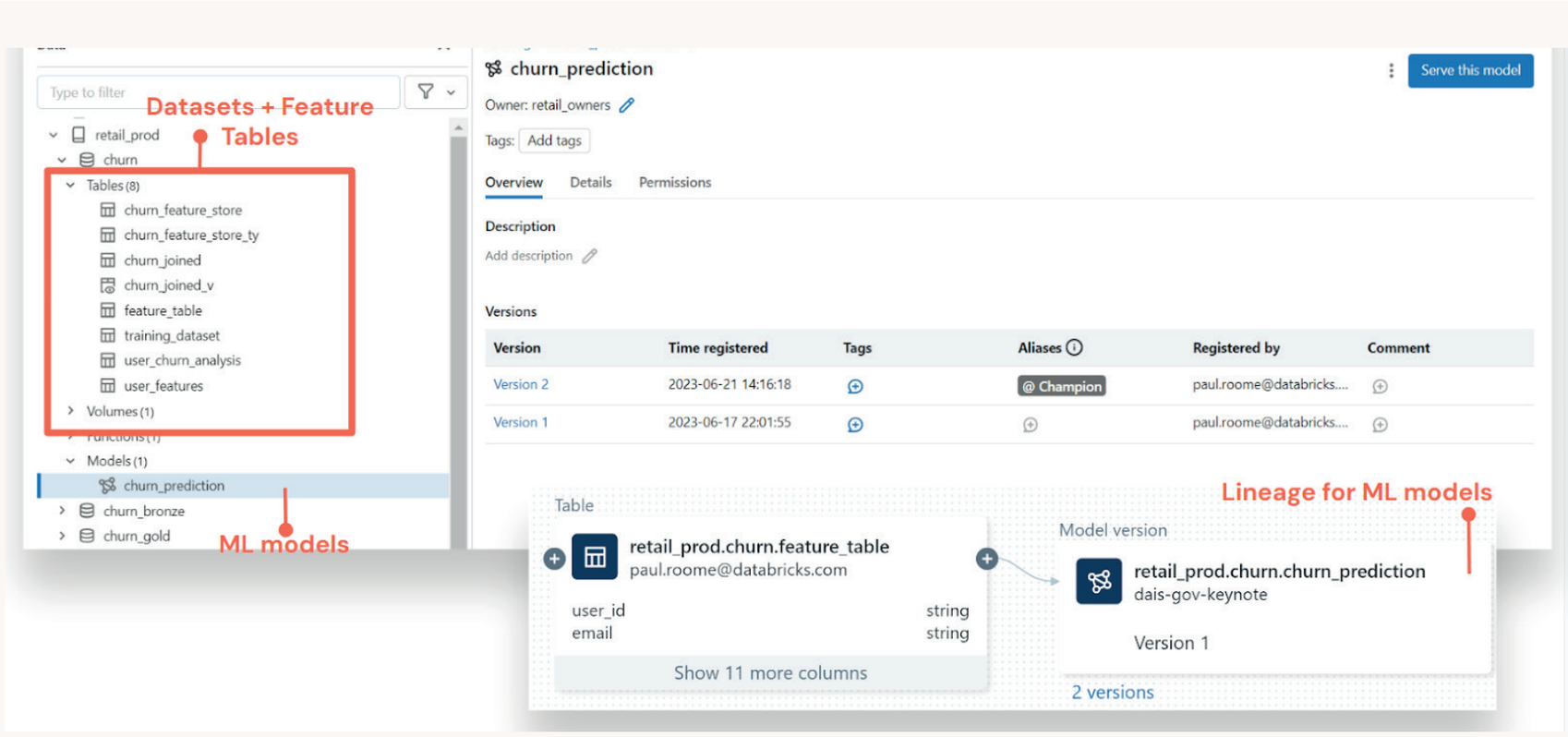
Databricks Lakehouse Monitoring introduces an advanced capability to scrutinize fairness and bias in classification models. It empowers you to evaluate the predictions of a classification model, gauging its consistency across diverse data groups. For instance, you can delve into whether a loan-default classifier exhibits a uniform false-positive rate across applicants from various demographics. Utilizing the lineage information encapsulated within Unity Catalog, you can seamlessly delve into root cause analysis, offering insights into the factors behind any anomalies observed within the ML model's performance.

Cataloging and documentation

Gathering and making available metadata around data and ML is core to governance. Catalogs for ML, in particular, provide insight and guidance on what is available and how it may be reasonably used. This helps data scientists build upon one another's work — taking techniques and expanding upon them as a best practice. Reuse in this way — sometimes a form of transfer learning — accelerates the time to production and value for the company.

Leveraging tools like Databricks Feature Store, MLflow and Unity Catalog is a great start to identifying existing ML assets within the organization. Metadata captured automatically during the model training process provides context for those searching through what has already been developed to use and expand upon. With the support for cataloging and managing access to ML feature tables and ML models in Unity Catalog, governance policy can then be enforced to ensure that the security of sensitive data is maintained and the use is appropriate.

With Feature Engineering supported in Unity Catalog, you no longer need to manage separate data catalogs for AI/ML and data. You can have a single, unified catalog for all data and AI.



ML models, Feature Tables in Unity Catalog

Chapter 3
Architecture Governance

To ensure that business processes are optimally supported by a well-defined underlying IT architecture, an architecture management process provides a useful framework. The goals of this process include:

- Achieving a clear overview of the current architecture
- Defining architecture principles, standards and guidelines
- Analyzing the current business and IT vision and needs
- Defining an appropriate target architecture
- Explaining the value of the target architecture in terms of business performance measures
- Analyzing gaps between the current state and target state architectures, and proposing new projects needed in order to address them
- Creating an architecture road map

In parallel, an architecture governance and compliance process needs to be implemented to ensure the following:

- The defined principles, standards and guidelines are used correctly in every project impacting the underlying systems
- The planned projects to fix the identified gaps are executed in time
- The overall system landscape evolves toward the defined target architecture

Important dimensions to take into account are consistency, security, scalability, standardization and reuse. By making sure initiatives are optimized for these dimensions, architecture governance will increase the cost-effectiveness and efficiency of the underlying system landscape.

Both processes are enterprise-wide efforts and not restricted to IT/architecture departments. For architecture governance to be efficient and successful, it should not follow a decoupled gate approach at the end of an implementation. This leads to extra efforts to fix already implemented violations or to the need to accept and then manage architecture violations. Instead, architecture governance needs to be a collaborative approach from the very beginning of every project. Early understanding of the impacted business processes and requirements allows the technical teams to map these requirements properly to the target architecture and then provide guidance throughout the project to achieve compliance. However, a formal final gate at the end can be considered for documentation reasons.

Architecture governance for the lakehouse

The Databricks Lakehouse Platform equips organizations with a lot of capabilities that make managing data and AI assets easier and fully under control. With such a powerful technology, it is best that organizations apply architecture governance on their data platform that ensures this powerful technology is being well architected and exploited in the best way possible, and is aligned to achieve the organization's data and AI goals. Without architecture governance, an organization might inevitably generate a long-term operational impact, create data silos, or have collaboration and standardization issues that are increasingly difficult to correct at a later stage.

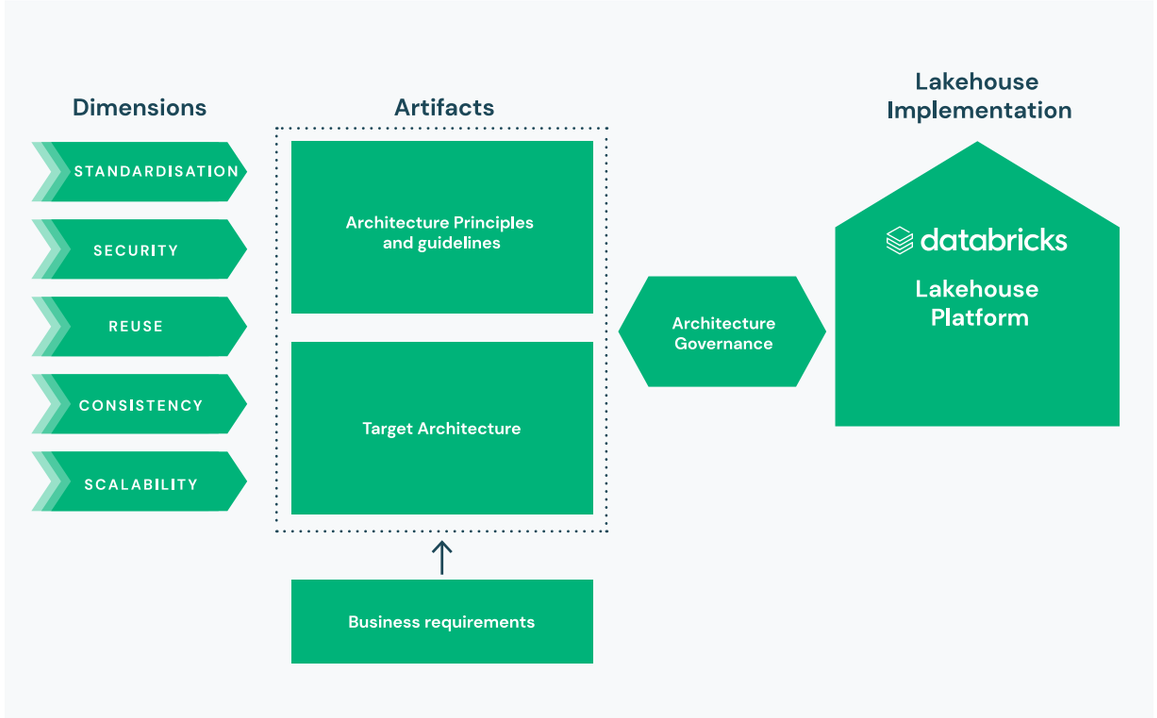
To allow for effective architecture governance for the Databricks Lakehouse Platform, it is recommended that the following artifacts be available.

Achieving a clear overview of the current architecture.

- A set of architecture principles and guidelines tailored to the data domain
- A target architecture for the data domain that is based on the business requirements and follows the agreed-upon architecture principles and guidelines

Every platform architecture should be as simple as possible to make maintenance, integrations and migrations as easy as possible. Simplicity is a prime goal of the Databricks Lakehouse Platform, and by following the above approach, architecture governance will ensure an architecture that is simple and allows for high-quality data and insights.

To frame architecture governance for the lakehouse, the following sections will dive into architecture principles and detail the characteristics of the above-mentioned dimensions.



Architecture principles

Adopting the Databricks Lakehouse technology enables organizations to deliver on their business objectives using data and AI. To ensure this objective is met, data teams are required to adopt a set of architecture principles that deliver on these objectives. We list some examples of these principles below; however, each and every organization should adopt and extend these principles to best fit their business needs.

Principle 1: Offer trusted, high-quality data-as-products

This should be the ultimate goal of the lakehouse architecture. Data teams must apply product thinking to the curated data. That means they view their data assets as products and the rest of the organization's data teams that consume data as their customers. With the Databricks Lakehouse Platform, data engineering teams can curate, build and offer these trusted data-as-products with ease.

Principle 2: Access through a self-service experience

Organizations should, as much as possible, make data accessible to both business and technical teams through a self-service experience without having to rely on tediously long processes, or centralized proxy teams to access the data. The Databricks Lakehouse Platform allows organizations to have this self-service access in place, not only through the built-in Databricks SQL, which offers self-service BI and data warehousing, but also through the integration with a wide set of BI tools that are well integrated into the Databricks partner ecosystem.

Principle 3: Democratize value creation from data

Creating value from data should not be limited to one central team in the organization. Many organizations now are moving toward a data-driven approach, by which all teams inside the organization have access to data sets that help them to make better business decisions and create business value. With the Databricks Lakehouse Platform in place, value creation can be securely distributed to all teams inside the organization. Databricks Unity Catalog, for example, makes sure that teams accessing data are granted proper access levels to avoid unauthorized access.

Principle 4: Prevent data silos

Some organizations suffer from having multiple data silos in place. Perhaps inherited from legacy data management solutions that did not offer proper and easier access for the data teams. This led to copying needed data repeatedly and to innovating and building on top of the siloed data copies. With the Databricks Lakehouse Platform, these data silos will be either integrated into the Lakehouse Platform or removed altogether. The lakehouse offers a set of tools, as well as technologies that ensure no further silos need to be created once the lakehouse is adopted.

Architecture dimensions

Each data platform architecture should be bound by certain dimensions (or properties) that define how this architecture should be maintained, extended or integrated. The Databricks Lakehouse Platform was designed with a set of architectural properties that make the architecture easy to work with. Adopting the Databricks Lakehouse architecture as the organization's data platform will let this architecture automatically inherit these properties. The following are examples of these architectural dimensions:

Dimension 1: Consistency

For an architecture to be consistent, data that flows through this architecture should not hob through multiple, disintegrated products that all require different formats and use different security models. Databricks Lakehouse architecture is inherently consistent: you can perform different kinds of workloads, implement different sets of business use cases and use all the platform components seamlessly. The platform was designed to be fully integrated and collaborative.

Dimension 2: Security and Privacy

Data security is extremely important for businesses. Certain classes of data such as PII, payment card data (PCI) and patient health information require an even greater extent of data security to ensure the data is safe from unauthorized access. With the Databricks Lakehouse Platform, security is a first-class citizen in the platform to ensure data is accessed properly, by authorized individuals. Unity Catalog can track access controls of users and ensure access is fine-grained to the row and column levels.

Dimension 3: Scalability

In many organizations, data is one of the fastest-growing assets inside the business. Some organizations' data almost doubles every year. It is very important that the adopted data platform is scalable to meet the increasing demands on data access. It also should be cost-effective, so in times where demand is low, organizations do not need to pay for the unused resources. The Databricks Lakehouse Platform fits perfectly in this dimension. The platform itself is built as a cloud-native platform-as-a-service, where each compute resource is designed to be elastic and can grow and shrink to meet varying compute demands.

Dimension 4: Standardization

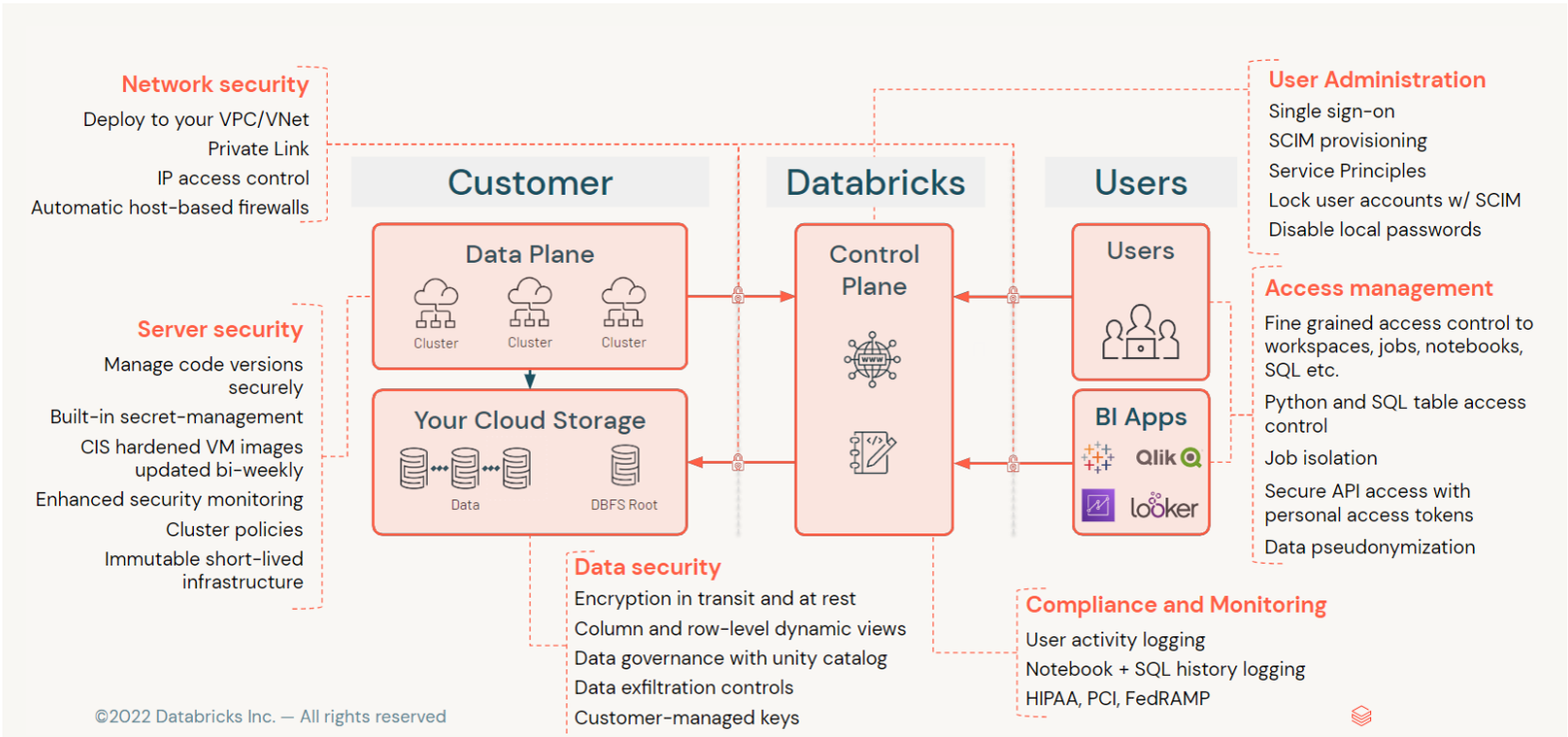
The data platform architecture should be built in a way that follows accepted standards and should be based on open data formats to avoid vendor lock-in. Proprietary data formats might in the short term offer some performance optimization, but in the long run, these formats become difficult to integrate and work with, which increases the total cost of ownership. The Databricks Lakehouse Platform is based on open source formats such as Parquet and Delta. It uses compute engines providing the open source API of Apache Spark and Delta Lake to process and store data. This makes the Lakehouse Platform performant and easy to integrate and work with.

Dimension 5: Reuse

The data platform architecture should be designed in a way that lets data engineers, data scientists and BI analysts create reusable components that fit the platform to avoid duplication of work. The platform should have sufficient tools to integrate with established CI/CD pipelines, and be fully automated to meet the organization's business agility requirements. The Databricks Lakehouse Platform has strong integration with various source control systems, and it also offers a set of APIs and command line tools (CLI) to be fully integrated with CI/CD tools.

Chapter 4
Platform Security and Compliance

We learned how data governance ensures the security, quality, integrity, lineage and availability of data across an organization. However, to meet stringent regulatory and compliance requirements and alleviate the risk of any data breaches, the security of the data platform also becomes a key consideration for data and AI initiatives. In this chapter, we will explore how the Databricks Lakehouse Platform provides end-to-end security and helps you meet compliance requirements.



The Databricks Lakehouse Platform Architecture

Platform security

Today, more and more organizations recognize the importance of making high-quality data readily available for data teams to drive actionable insights and business value. At the same time, organizations also understand the risks of data breaches, which negatively impact brand value and inevitably lead to the erosion of customer trust. Organizations that operate in multicloud environments need a unified, reliable and consistent approach to secure data. We've learned from our customers that a simple and unified approach to data security for the lakehouse is one of the most critical requirements for modern data solutions. With significant investment into building a highly secure and scalable platform, Databricks delivers end-to-end platform security for data and users.

The Databricks Lakehouse Platform architecture is split into two planes — control plane and data plane — to simplify your permissions, avoid data duplication and reduce risk. The control plane is the management plane where Databricks runs the workspace application and manages notebooks, configuration and clusters. Unless you choose to use serverless compute, the data plane runs inside your cloud service provider account, processing your data without taking it out of your account.

You can embed Databricks in your data exfiltration protection architecture using features like customer-managed VPCs/VNets and admin console options that disable export. While certain data, such as your notebooks, configurations, logs and user information, is present within the control plane, that information is encrypted at rest, and communication to and from the control plane is encrypted in transit.

Regardless of where you choose to host the data plane, Databricks networking is straightforward. If you host it yourself, Databricks by default will still configure networking for you, but you can also control data plane networking with your own managed VPC or VNet.

The serverless data plane network infrastructure is managed by Databricks in a Databricks cloud service provider account and shared among customers, with additional network boundaries between workspaces and between clusters.

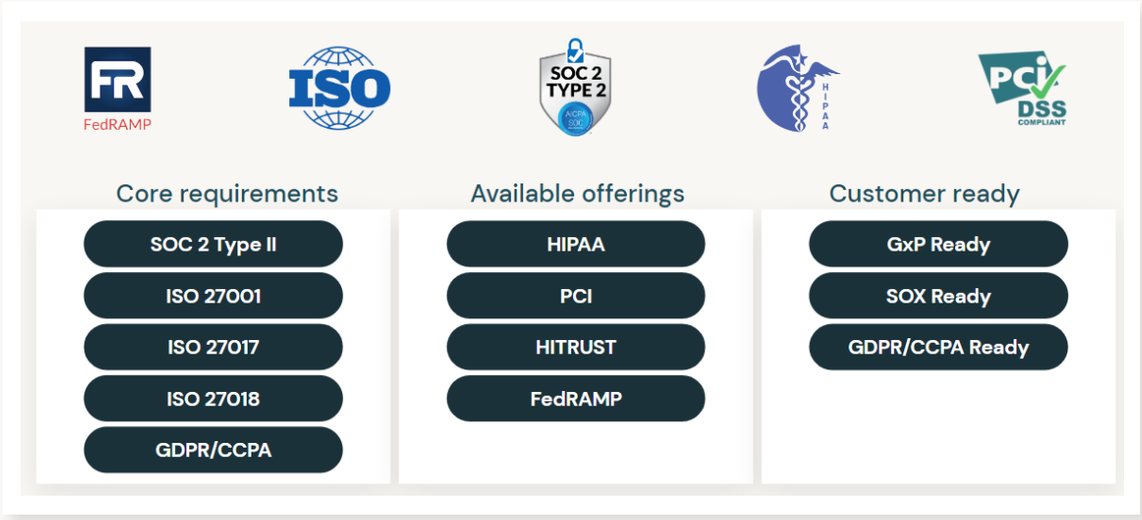
Databricks does not rewrite or change your data structure in your storage, nor does it change or modify any of your security and governance policies. Local firewalls complement security groups and subnet firewall policies to block unexpected inbound connections.

From a data perspective, all data is stored in the customer's cloud object storage, and Databricks offers end-to-end encryption for both data at rest and data in motion.

This is the platform and its security approach at a high level. For more details visit the Databricks Security and Trust Center.

Data compliance

Meeting the regulatory compliance requirements is perhaps one of the key drivers for data governance initiatives. Databricks is trusted by the world's largest organizations to provide a powerful lakehouse platform with high security and scalability. Databricks has put in place controls to meet the unique compliance needs of highly regulated industries. You see all the core compliance certifications such as ISO 27001 and SOC2. But Databricks also has industry-specific compliances such as HIPAA for healthcare, PCI for financial industries, FedRAMP for federal agencies, and so on. This is a growing list — as we see more use cases and customers needing more validations, we are always willing to jump in and work alongside you to help with your compliance needs.



About the authors

Amr Ali is a Senior Solutions Architect at Databricks and has over 18 years of experience in building IT solutions. Amr has helped many organizations build their data and AI platform around lakehouse technology to achieve their business goals, with a strong emphasis on data architecture governance and security.

Bernhard Walter is a Lead Product Specialist for architecture at Databricks and has about 30 years of experience in different roles in the IT industry. His primary focus at Databricks is to help customers to create efficient and future-proof lakehouse architectures that optimally support their business.

Jason Pohl is the Director of Data Management at Databricks. He helps shepherd new products to market in the fields of data warehousing, data engineering and data governance. Jason's career spans architecting legacy data warehouse solutions to modern data lakehouses in the cloud.

Lexy Kassan is a Senior Data and AI Strategist at Databricks, overseeing and guiding customer organizations through holistic digital transformation programs. She is also the founder and host of a data science ethics podcast in which she delves into the responsibilities and societal ramifications of applying advanced analytics across both private and public sectors.

Sachin Thakur is a Principal Product Marketing Manager at Databricks. He leads the go-to-market strategy for Delta Sharing and Unity Catalog, and he is passionate about helping organizations democratize data and AI with the Databricks Lakehouse Platform.

Zeashan Pappa is a Senior Product Specialist for Data Governance and Unity Catalog and works across product/engineering/field teams on go-to-market strategy, adoption and thought leadership. He has 20 years of experience in engineering/architecture/tech leadership at a variety of firms.

Mayur Palta is a Senior Specialist Solution Architect at Databricks and leads a competitive intelligence focus for data management and a win-loss program at Databricks. Mayur's career spans building technology startups, helping Fortune 500 enterprises get more value from the data they have, and building data architectures that scale.

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.

Start your free trial

Contact us for a personalized demo at databricks.com/contact

To learn more about migration, visit databricks.com/migration

databricks