Databricks Enterprise Security

Safest place to run your AI and Apache Spark workloads

databricks

Table of Contents

1. Databricks approach to Enterprise Security

How to ensure your sensitive data isn't at risk

In today's business climate, the ability to anticipate and meet customer needs is central to success. Forward looking business leaders are looking to unleash the power of Artificial Intelligence (AI) to drive innovation, but this requires bringing together diverse teams and large volumes of data. With attackers getting more sophisticated, securing these complex data workflows is a top priority.

Security challenges in AI and big data

Many organizations face the common challenge of how best to manipulate large volumes of sensitive data to gain meaningful insights in a secure way. Data engineers and security teams struggle to give their data scientists and analysts the speed and access to the data they need to drive AI initiatives while ensuring consistent policy management, data governance, and security compliance.

Many opt to build their own advanced analytics solutions by cobbling together a plethora of data processing (Spark, Hive, Pig etc) and AI/ML tools (SparkML, Tensorflow, PyTorch etc), many of which are open source. This can introduce behaviors that increase security risk. According to Gartner, 80% of organizations will fail to develop a consolidated data security policy across silos. In an effort to address this, some companies over-rotate, by tightly locking down data. This can be costly, hindering their ability to innovate and getting in the way of meeting customer needs.

As the Chief Information Security Officer (CISO) of Databricks, I help customers establish and secure their AI data pipelines. I see the following 3 security challenges over and over again:

- **Teams acting in silos:** For many organizations, technology, people and AI workflows exist in silos. Data engineers and data scientist work with their own toolsets. Often times, these tools are rapidly evolving open-source applications that are poorly integrated across data workflows. This not only kills innovation but also creates large security holes.
- **Inability to deploy securely at scale:** Building a secure scalable architecture is difficult. You have to manage configuration, monitoring, patching, authentication, and security scanning. Enterprise who have compliance requirements (e.g. SOC2, ISO, HIPAA, GDPR) have a more difficult challenge.
- **Security as an afterthought:** AI projects are started with a focus on speed and innovation. Security is often a bolt-on and thought of only later when a major compliance audit is due. At this point, it may be too late to solve some of the fundamental problems with the deployment.

Unified Security Approach for AI & Data

Databricks' Unified Analytics Platform brings together data and Machine Learning(ML) with best-in-class security on the most trusted clouds to accelerate innovation while minimizing risk. We have built the Databricks Unified Analytics Platform with a Security first mindset to solve the following problems:

Data Platform - Knockdown silos while keeping data secure
Databricks offers a unified analytics platform that seamlessly and securely connects disparate teams and data to accelerate innovation.

Deploy and operate securely at scale
As the amounts and types of data, users, tools, workloads and ML models increase, the complexity of securing them increases exponentially.

Our Culture - Security at our Core
With many DIY analytics and AI solutions, speed and innovation come first and security is an afterthought. We built Databricks with data security at its core from Day 1.

In the following sections we are going to take an in-depth look at each area.

Data protection at every level

Databricks has been architected at every layer of our infrastructure to provide advanced security, risk prevention, and management controls for your data, AI and Apache SparkTM workflows. By combining security and convenience, we bring together teams to realize the promise of AI and drive innovations that enable business transformation. Look for more blogs on each of these three Security pillars to be published in the upcoming weeks.

2. Databricks Platform

A security-first approach to AI with Databricks Unified Analytics Platform

DIY Platforms:

A Lack of Cohesion in Security

Many companies today operate on homegrown DIY big data and AI platforms comprised of various open-source tools and technologies. These patchwork platforms pit data scientists against data engineers and put the entire organization at risk of a security breach. On one hand, data scientists demand the latest AI tools and prioritize speed and productivity. They see IT and security as slow, and inflexible. On the other hand, data engineers are incented to maintain pristine data on a secure, supportable, production-ready platform. Moreover, each of these teams requires their own toolsets to deliver on the needs of their job. These tools are often times disjointed and data workflows cannot be tracked end-to-end. This results in data engineering and data science teams working in silos with the following challenges:

From a productivity perspective:

- Sometimes, data is locked down on-prem with a perimeter security approach making scale and elasticity hard, eventually impeding innovation.
- When teams fail to collaborate, use siloed tools and build disjointed data workflows, it becomes virtually impossible to innovate and progress AI projects.

From a security and regulatory perspective:

- Rapidly evolving, open source toolsets are cobbled together. They are not patched, tested, or well maintained – the potential for gaps, and errors by well-intended insiders is high.
- Different systems have different security paradigms, making it a challenge to replicate policies and ensure security fidelity from one toolset to the next.
- Lack of a single workspace makes it harder to get the right data to the right person while controlling who has access to what.

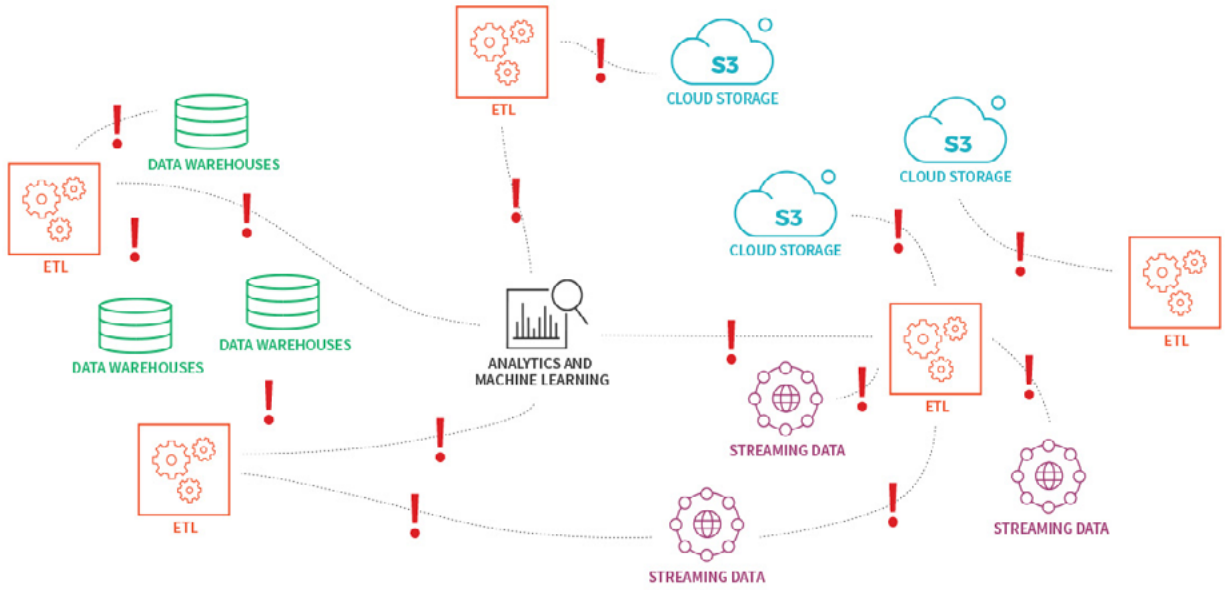


Figure 1: DIY AI platforms – Lack of Cohesion in Security

To overcome these obstacles, data scientists end up hiring data engineers into their teams, and data engineers do the same, busting budgets and creating duplicative functions. In the end, you've got teams that can't work together, are not accountable, and blame each other for creating an untenable situation. This can create significant exposure of a data breach or regulatory infringement putting companies at risk in ways they may not have even considered. I believe, there's a better way to meet the needs of data science teams than DIY AI.

Unified Analytics Platform with security at its Core

Databricks brings data engineering and data science teams together in a unified analytics platform giving the data scientist the agility they want while providing data engineers a consistent, secure and reliable toolset. That unification is key. Unlike DIY solutions, we provide a coherent security model across the entire data workflow. We enable you to set up the security once for your data framework and it can then be used for your data processing and as well as your ML and AI needs.

Knockdown silos while keeping data secure

Our unified analytics platform provides the security you require while enabling teams to work together to drive innovation. At the core of our approach are the following:

1. Security as a Core Design Principle

Databricks is a cloud-native platform that was designed with security as a first-class citizen from day one and is

- **Cloud native using security best practices** – People often say that they keep important data on-prem to keep it safe. This is a dangerous misnomer. Security on-prem is just as difficult as security in the cloud. In addition to the elasticity and the pay as you go advantages of cloud, Databricks offers a fully managed and monitored security offering that utilizes the best cloud practices with world-leading security experts. We have built-in controls to minimize human error that includes isolating data workloads and controlling access through specific VPCs/VNETs, IAM Security groups while encrypting data in-transit and at-rest with fine-grained access controls.

- **Designed to not touch your data** – Data is the business differentiator for your company. We want to ensure you have maximum power over your data to generate business value. With the data staying in your cloud account, we provide you the infrastructure and manage it to ensure you are secure and get maximum value out of your data. With separate data and control planes, the data and the workloads reside in your cloud account and Databricks has no access to it.

- **Integrated with your existing processes** – Being a single platform having secure integrations with familiar products for SSO (Single Sign-On), Data Warehouse and BI tools, Databricks can easily be integrated into your existing environment and processes.

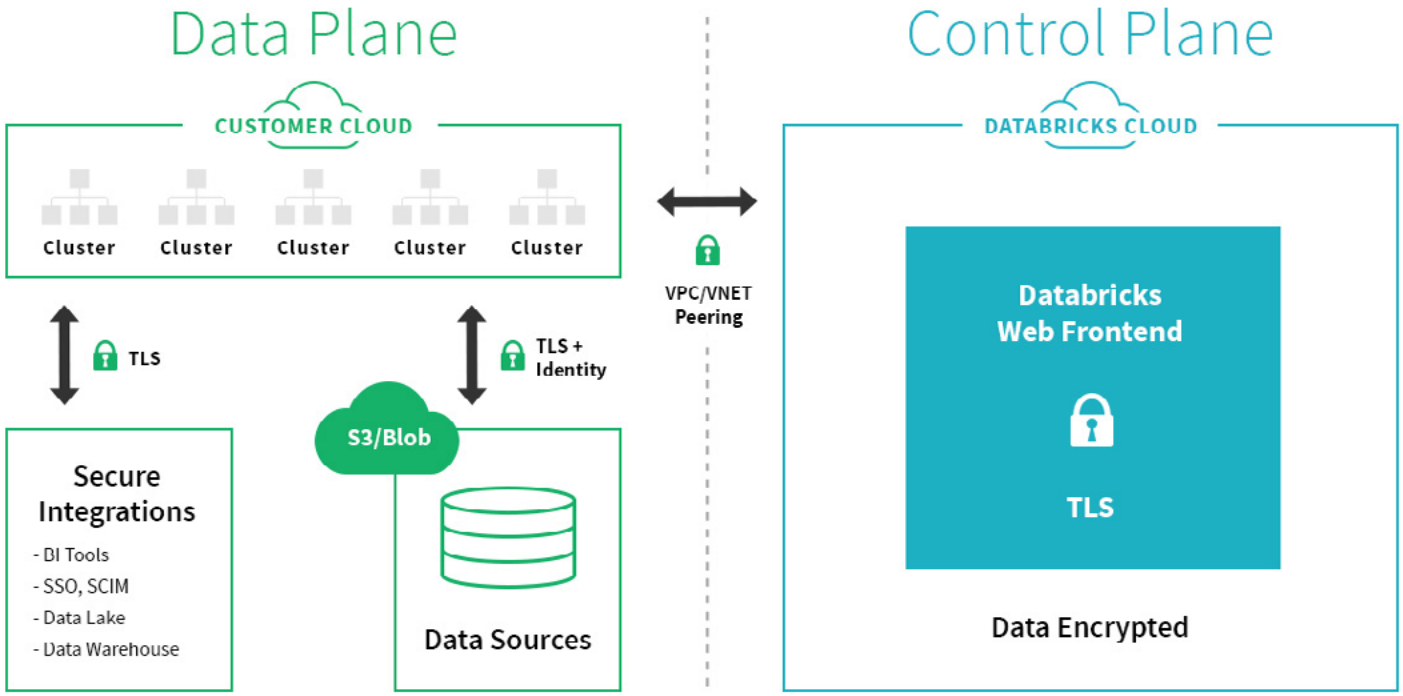


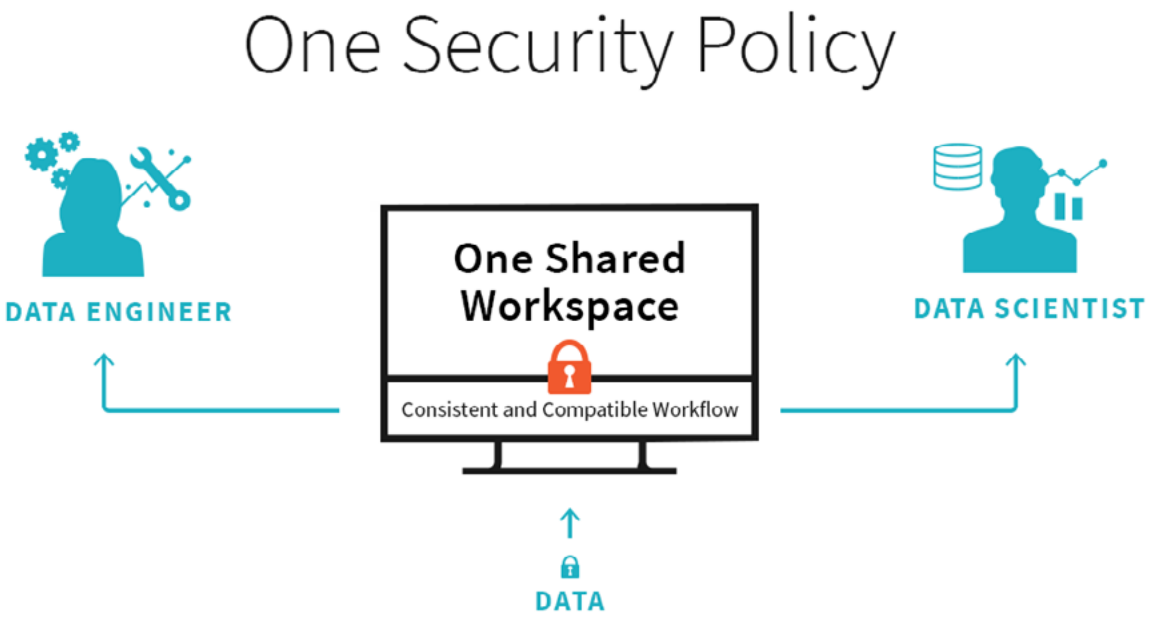
Figure 2: Minimize human error with segregated data and control planes

2. Consistent, Reliable and Compatible Workflows

Data Scientists are constantly looking for the latest software with the latest models. Even a half percent improvement in ML models could impact millions of dollars in revenue. DIY requires teams to integrate their own Data Engineering Tools (Spark, Hive etc) and Data Science tools (SparkML, Tensorflow, Keras etc). Databricks does that work for you by providing a single unified platform with streamlined workflows. No longer do you need to worry about interoperability issues. Patching, configuration is all taken care off and the whole system is pen tested and monitored by world-leading experts. This makes it easy for IT and Security teams to maintain security across the entire workflow with ease.

3. Secure and Transparent Collaboration

Databricks lets both Data Engineering and Data scientist teams work together in a single shared workspace. Databricks notebook can be shared by multiple teams with commenting and versioning much like Google docs. Not only does this enhance collaboration and but is a single interface to control, track and audit user access of data. Fine-grained access controls let you govern data not only at a bucket, file level but also at a row and column level. With this level of control, you could give database access to a wide swath of people but block out specific sensitive columns such as credit card numbers and social security numbers.



When leveraged to its full potential, data can be a true differentiator for your company. We want to empower your teams to generate business value using your preferred frameworks and libraries for AI while ensuring data security and regulatory compliance.

If you are currently working through the challenges posed by DIY AI, we can help identify security gaps in your current data analytics set up and show you how you can address those more effectively with a single platform solution.

3. Deploy and Operate at Scale

Democratize Data! What about Security?

As organizations push to leverage their data to make more intelligent decisions, a core requirement is to "Democratize data". In other words, open up access to previously siloed and restricted data to broader parts of the organization. This often enables a new set of insights that unlock additional sources of revenue and can transform enterprises.

However, this new model of expanded access causes a fair bit of trepidation across organizations, especially in the C-suite. As they hear about data breach reports from enterprises across the spectrum, they look for ways to balance security and governance with access. In particular, access to data from a myriad of services and users is controlled through increasingly complex policies and configurations. The complexity increases the possibility of human error and surface area for attacks, if not done correctly. Earlier this year, IBM published a study in which they showed a dramatic increase in data breaches as a result of cloud service misconfiguration – due in great part to the role of human error.

The Databricks offering is designed with the specific goal of enabling this balance: maximize end-user productivity and access without sacrificing (and in many cases strengthening) security and governance compliance. As part of this, we've focused on automating as many operations as possible. Additionally, we've established "separation of concern" principles to minimize opportunities for human error and the impact of any such misconfigurations.

To provide more detail, here are some considerations we talk through with our customers

Deploy Securely

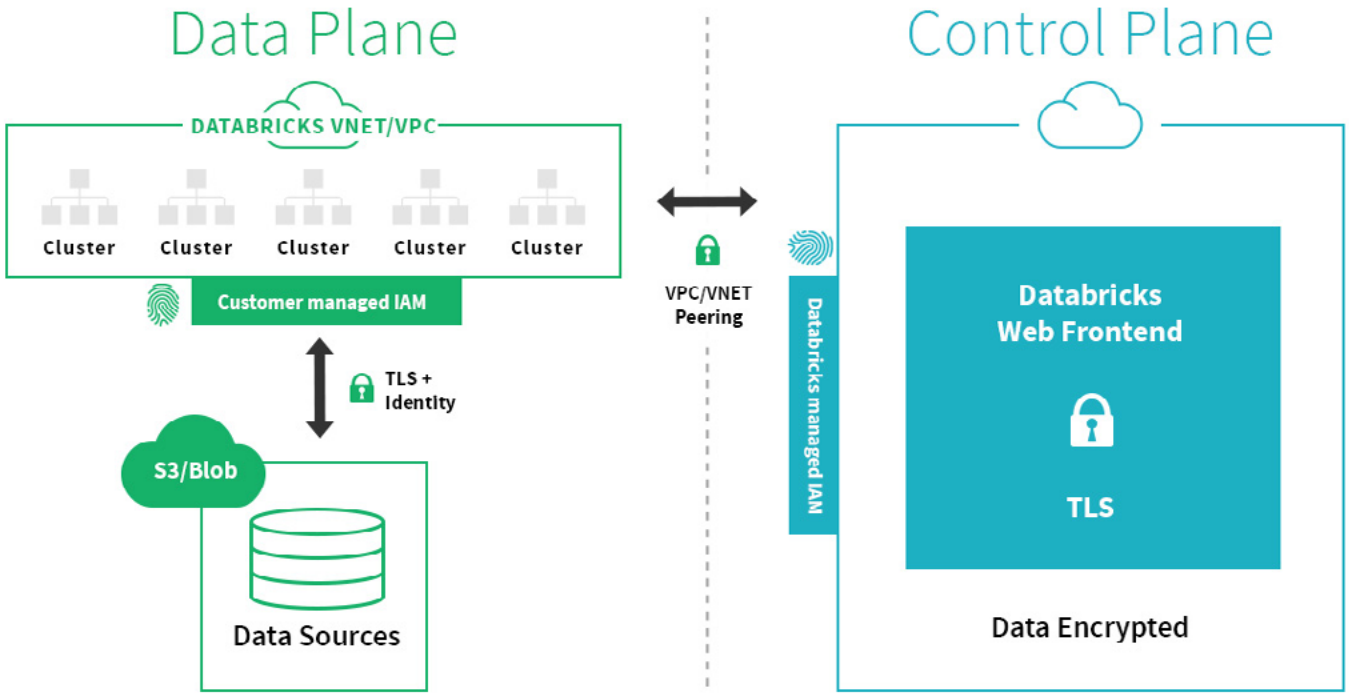
The first step is ensuring no unnecessary security vulnerabilities are introduced during initial deployment:

1. Our Platform – Knockdown silos while keeping data secure

Data scientists and engineers often work in silos using a disparate collection of tools and fragmented data sets. Further AI requires the latest tool (Tensorflow, Pytorch, Keras etc) that enables the best or most efficient results for a given model. These tools are rapidly evolving and staying on top of the latest vulnerabilities and integration errors is cumbersome.

- **Keeping your infrastructure in your account –** Although Databricks provides the benefits of a fully managed SaaS service, all clusters are created and torn down inside of a customer's Azure/AWS account, and data stays where it is. Instead of creating a whole new set of configurations in a different environment, which increases the risk of making a dangerous error, all existing security configurations and monitoring tools can continue to be leveraged.

- **Isolated networking to reduce blast radius** – Databricks creates a dedicated VPC/VNET in the customer's account that only contains Databricks infrastructure. This ensures that all Databricks policies and controls have no access to other production infrastructure in a customers account. Additionally, all communication with Databricks' control infrastructure goes through direct links (while additionally eliminating any Public IPs). This prevents any traffic from traversing the public internet and is also encrypted with mutual TLS v1.2. And finally, multiple security groups are used so that if ever Databricks services need to connect to other customer VPC/VNETs (e.g., containing another production data source), there is no risk of exposing access to those services to the outside world.



Isolated networking and security groups

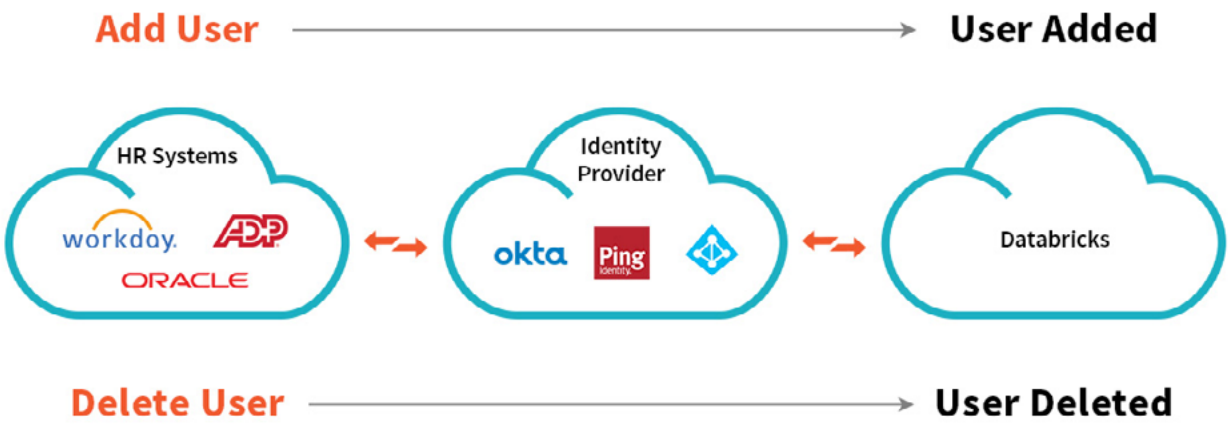
- **Leverage scoped down permissions** – In order to manage customer infrastructure, Databricks solely uses tokens and roles (vs. keys) to eliminate the risk of leaks. Additionally, these roles have extremely limited permissions (e.g., explicitly excluding access to any customer data) to only enable them to set up the Databricks infrastructure.

Operate Securely as you Scale Users

Equally as important as the initial deployment is ongoing operations and maintenance:

- **Automate and minimize human intervention** – Databricks automates general SaaS monitoring and updates, to eliminate human intervention. In the rare cases where human access is needed, it is only provided through a time-bound token-based model. Access is only available to a small subset of core Databricks operators, and all activity is logged. Read-only audit and access logs are delivered directly to customers where they can be fed into existing security monitoring infrastructure.

- **Integrations with popular security tools** – Integration with SSO identity providers (such as Azure Active Directory, Okta, Onelogin etc.) and support for SCIM allow you to easily manage and secure users. New user set up and removal of access during sensitive times when employees may be leaving the business is automated.



SCIM integration with SSO identity provider

- **Strict access controls and network-based isolation** – After deployment, access right provided to the Databricks web application and identity roles can be further restricted to the bare minimum needed for steady-state operations, eliminating any broader permissions needed for deployment. Any exceptions are fully auditable and subject to a formal approval process by the customer.

Govern Data with Confidence

Finally, it is critical to ensure data is always protected

- **Physical and logical data access control** – Databricks provides the ability to leverage ACLs to restrict which users can physically access underlying data files, while also providing fine-grained access control (e.g., row and column based) to logical tables that have been created. For example, a Social Security number column can be hidden while allowing access to the rest of the table.
- **Effortlessly secure end-to-end data and workflow** – Databricks' rich ecosystem enables seamless integrations and secure, encrypted, authenticated communication with popular enterprise big data technologies such as Data Lakes, Data Warehouses, JDBC/ODBC, and BI Tools.
- **Audit Logs** – Databricks provides comprehensive end-to-end audit logs of activities done by the users on the platform, allowing enterprises to monitor the detailed usage patterns of Databricks as the business requires.

- **Effortlessly secure end-to-end data and workflow** – Databricks' rich ecosystem enables seamless integrations and secure, encrypted, authenticated communication with popular enterprise big data technologies such as Data Lakes, Data Warehouses, JDBC/ODBC, and BI Tools.
- **Audit Logs** – Databricks provides comprehensive end-to-end audit logs of activities done by the users on the platform, allowing enterprises to monitor the detailed usage patterns of Databricks as the business requires.

At Databricks, we know how complex and time consuming it is to scale data systems and access while minimizing human error and managing risk. That's why we've worked to make it virtually effortless to deploy and manage our platform while maintaining data governance and access at scale.

4. Partnering with Databricks on Security

Having the best security requires a partnership that's built on technology, trust, and transparency

Artificial intelligence software that can learn and improve human decision-making is transforming business. All sorts of companies are looking to AI to gain an edge over competitors. Unfortunately, everyone is racing to piece together an AI framework, sometimes forging alliances with software vendors that don't prioritize security—clearly a risky proposition.

Is security a priority for your vendors?

The sprint to build the most competitive AI platform invariably involves partnering with organizations that integrate or make software. So, even if you're managing your own security effectively, minimizing risk demands similar levels of vigilance from your software partners. It's a delicate balance—like a dance where each partner has to match or anticipate the other's next move.

In the pre-cloud days, vendors would build software and hand it over to their customers, leaving them responsible to manage it, regardless of whether they had the necessary expertise in that particular area. In today's SaaS era, the vendor manages the software, which requires a strong partnership and deep trust between customer and SaaS provider. For the partnership to succeed in the long term, both parties must be linked by trust and transparency, and having a third party audit your vendor's processes and confirm that their security practices meet or exceed industry standards is key to building both.

Databricks – Security is at our core

Databricks is committed to security first. From day one, we built Databricks as an enterprise software company with a security-first mindset. We've created a culture of security best practices, and we work closely with third-party organizations to provide outside-in testing and validation.

We back up our commitment with our **technology**, earning our customers' trust by ensuring that our development teams apply industry-leading security practices. We have multiple certifications and rigorous third-party testing and validation, along with regular audits ensuring constant **transparency**.

Technology built with a security-first mindset

We've built many facets of security natively into our data platform, including encryption, identity management, role-based access control, data governance, and compliance standards.

We also take your existing security tools investments into account, with the goal of meeting you where you are. This gives you a range of integrations and controls for your existing security tools, including support for your current identity access management solution using SAML 2.0 and SCIM to simplify setting up and managing accounts on the Databricks platform.

Our security-first mindset extends beyond data access to the architecture itself, with hard segregation between Databricks data and controls planes, leaving the data where it is so we can't access it. Customers can choose between single or multi-tenant control planes depending on the level of sophistication needed.

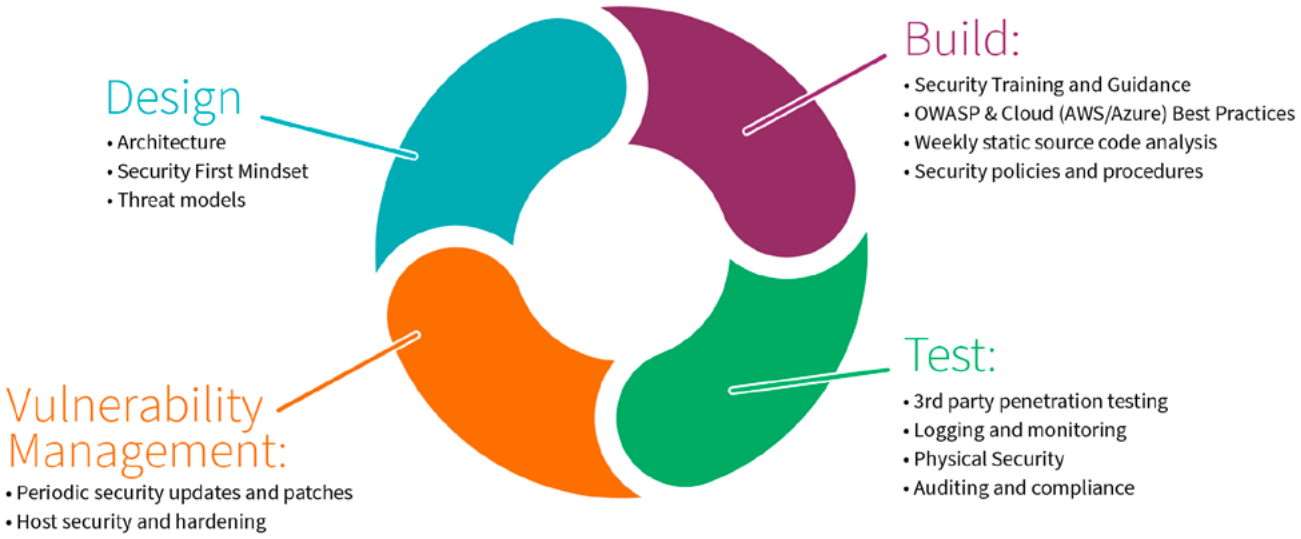
Security-minded teams, you can trust

Earning our customers' trust is our top priority, and we're committed to fostering security-minded development teams. Through rigorous developer training and ongoing education—and application of security best practices—our development team keeps Databricks solutions bulletproof when it comes to security.

As part of our commitment to maintaining our customers' trust and helping them mitigate risk, we also:

- Ensure that security processes and checks are an integral part of development, by following the Secure System Development Life Cycle (SDLC)
- Make sure all of our developers are well-versed in the security principals essential to their roles with quarterly security training
- Focus on threat modeling, which means running risk assessments throughout the development process and implementing preventative security controls as needed
- Take an "always-be-testing" approach that allows us to identify vulnerabilities early on
- Ensure that the platform is free from security defects, by performing comprehensive quality assurance and penetration testing

Finally, our low engineering attrition rates contribute to continuity and adherence to security best practices.



Secure Software Development Lifecycle

Transparency through third-party validation

In addition to maintaining the highest level of data security through industry-leading best practices, we also work with independent, PCAOB-registered firms to audit our program regularly and attest to our certifications. We take transparency seriously and make sure our customers understand clearly the intricacies of our platform by making available detailed architectural documentation.

We meet the unique compliance needs of highly regulated industries. We are in compliance with standards such as ISO 27001, SOC 2 Type 2, and HIPAA, along with validation by third-party penetration testing. Our certifications allow us to serve customers in regulated industries, including the Financial Industry Regulatory Authority (FINRA), Sanford Health, and Shell—along with highly sensitive government agencies.

We're unique in our ability to support customers with GDPR compliance, particularly those using data lakes to store sensitive data that might be subject to a data subject request (DSR). Databricks is also architected so you're always in control of where your data resides.

We are in the process of completing our Privacy Shield certification and have certified services under SOC 2 Type II , PCI DSS and ISO 27001. We also recently attested to ISO 27018, the internationally recognized industry-standard approach for protecting personal data in the cloud.

Partner with us

Security is an increasingly critical part of AI initiatives, which means you need to partner with an organization that doesn't take security shortcuts. Databricks is committed to being your partner over the long haul, with security deeply embedded in our culture.