databricks

Databricks Platform Security

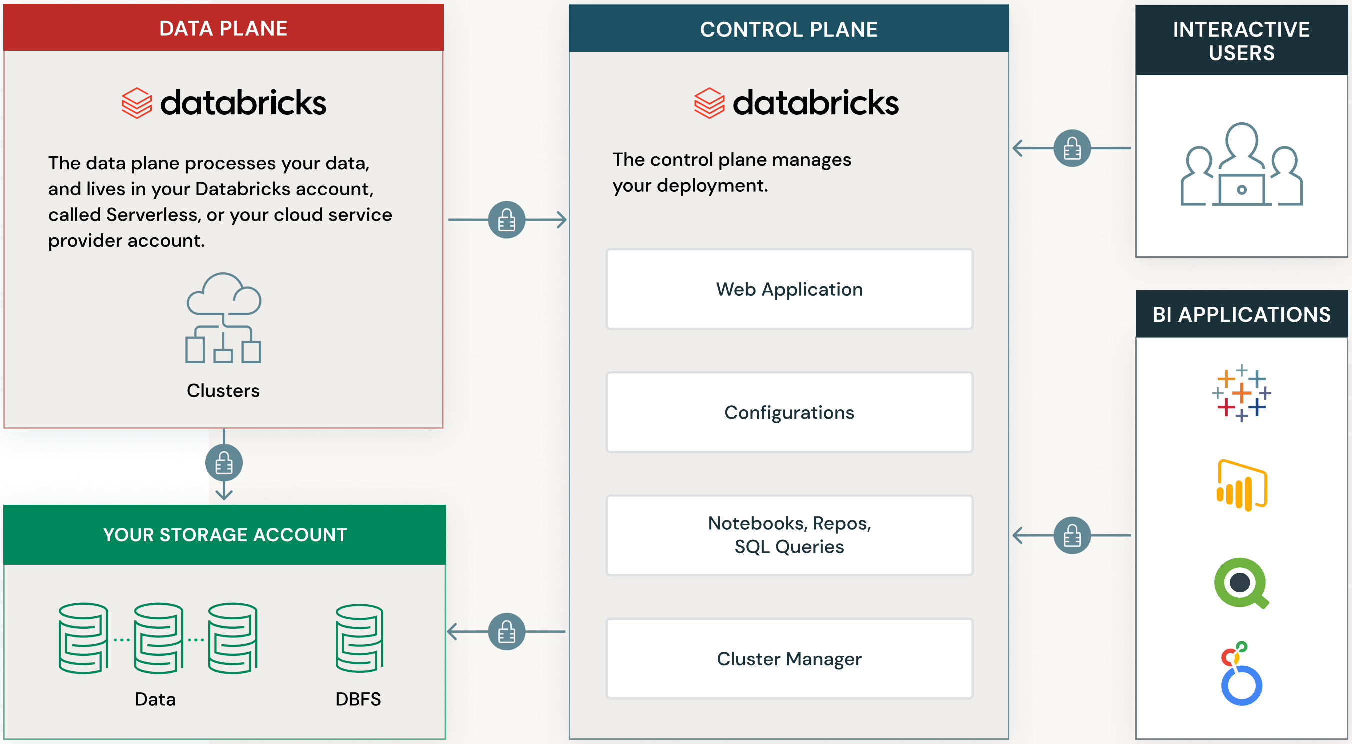
The Databricks platform provides
unparalleled security for your data and users

Thousands of customers trust Databricks with their most sensitive data to analyze and build data products using machine learning (ML). With significant investment in building a highly secure and scalable platform, Databricks delivers end-to-end platform security for data and users. This document provides an overview of the Databricks platform architecture, design choices and platform security features that enable your data teams to securely access relevant data while enforcing your data governance policies. Security is a broad topic, and you can find additional details in our Security & Trust Center.

NOTE: This document assumes at least the Premium tier Databricks subscription. Databricks is available on AWS, GCP and Azure. Learn more in the Databricks Platform Overview.

Architecture

The Databricks architecture is split into two planes to simplify your permissions, avoid data duplication and reduce risk.



DATA PLANE

databricks

The data plane processes your data, and lives in your Databricks account, called Serverless, or your cloud service provider account.

Clusters

YOUR STORAGE ACCOUNT

Data

DBFS

CONTROL PLANE

databricks

The control plane manages your deployment.

Web Application

Configurations

Notebooks, Repos, SQL Queries

Cluster Manager

INTERACTIVE USERS

BI APPLICATIONS

Step-by-Step Example

Suppose you have a data engineer that signs in to Databricks and writes a notebook that transforms raw data in Kafka to a normalized data set sent to storage such as Amazon S3 or Azure Data Lake Storage or Google Cloud Storage. The following steps make that happen:

1. Your single sign-on (such as Okta) seamlessly authenticates the data engineer via SAML to the Databricks web UI in the control plane. Native authentication is also available.
2. As the data engineer writes code, their web browser sends it to the control plane. JDBC/ODBC requests also follow the same path.
3. When ready, the control plane creates a Databricks cluster in the data plane, and sends the data engineer's code.
4. The cluster pulls from Kafka, transforms the data and writes it to your storage.
5. The cluster reports status and any outputs back to the control plane.

The data engineer doesn't need to worry about many of the details — they simply write the code and Databricks runs it.

Network and server security

Below, we'll review networking, servers and how Databricks interacts with your cloud service provider account.

Networking

Whether you choose Classic or Serverless compute, 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 customer-managed VPC or VNet. The Serverless data plane network infrastructure is managed by Databricks with additional network boundaries between workspaces and between clusters.

Local firewalls complement security groups and subnet firewall policies block unexpected inbound connections.

Customers at the Enterprise tier can also configure IP access lists or private networking (such as PrivateLink) to limit which IP addresses can connect to the web UI or REST API — for example, to allow only VPN or office IPs.

Servers

Databricks clusters automatically run the latest hardened system image. Users cannot choose older (less secure) images or code. For AWS and Azure deployments, images are typically updated every 2–4 weeks. GCP is responsible for their system image.

Databricks runs scans for every release, including:

1. System image scanning for vulnerabilities
2. Container OS and library scanning
3. Static and dynamic code scanning

Databricks code is peer reviewed by developers with security training. Significant design documents go through comprehensive security reviews. Scans run fully authenticated, with all checks enabled. Issues are tracked against the timeline shown in this table.

SEVERITY	REMEDIATION TIME
Critical	< 14 days
High	< 30 days
Medium	< 60 days
Low	When appropriate

Importantly, Databricks clusters are typically short-lived (often terminated after a job completes) and do not persist data after they terminate. Serverless workloads and customers using our Compliance Security Profile can enforce max cluster durations to regularly use the latest patches. Clusters typically share the same permission level (excluding high concurrency or Databricks SQL clusters, where more robust security controls are in place). Your code is launched in an unprivileged container to maintain system stability. This security design provides protection against persistent attackers and privilege escalation.

Databricks access

Databricks' access to your environment is limited to cloud service provider APIs for our automation and support access.

For customers who do not use Serverless compute, you will grant automated access for the Databricks control plane to configure resources in your environment using the cloud service provider APIs when you deploy. The specific APIs vary based on the cloud: an AWS cross-account IAM role, Azure-owned automation or GKE automation. These do not grant access to your data sets (see the next section).

By default Databricks personnel do not have access to customer workspaces or production environments. Databricks staff may request temporary access in order to investigate an outage or security event, or to support your deployment. Customers can choose to disable that access. Furthermore, Databricks staff activity is included in our audit logs delivery; access requires a multi-factor authentication prompt and employees must be on our VPN.

Identity and data governance

Databricks supports robust ACLs and SCIM. AWS customers can configure SAML 2.0 and block non-SSO logins. Azure Databricks and Databricks on GCP automatically integrate with Azure Active Directory or GCP identity.

Unity Catalog provides centralized access control, auditing, lineage and data discovery capabilities across Databricks workspaces:

Define once, secure everywhere: Unity Catalog offers a single place to administer data access policies that apply across all workspaces and personas.

Standards-compliant security model: Unity Catalog's security model is based on standard ANSI SQL and allows administrators to grant permissions in their existing data lake using familiar syntax, at the level of catalogs, databases (also called schemas), tables and views.

Built-in auditing and lineage: Unity Catalog automatically captures user-level audit logs that record access to your data. Unity Catalog also captures lineage data that tracks how data assets are created and used across all languages and personas.

Compliance

Databricks supports the following compliance standards on our multi-tenant platform:

SOC 1 Type II, SOC 2 Type II, SOC 3
ISO 27001
ISO 27017
ISO 27018

Certain clouds support Databricks deployment options for FedRAMP High, HITRUST, HIPAA and PCI. Databricks is GDPR-compliant as an organization, and the Databricks platform is GDPR-ready.

Encryption and auditing

Databricks provides encryption, isolation and auditing.

Databricks encryption capabilities are in place both at rest and in motion

For data-at-rest encryption:

- Control plane is encrypted
- Data plane supports local encryption
- Customers can use encrypted storage buckets
- Customers at some tiers can configure customer-managed keys for managed services

For data-in-motion encryption:

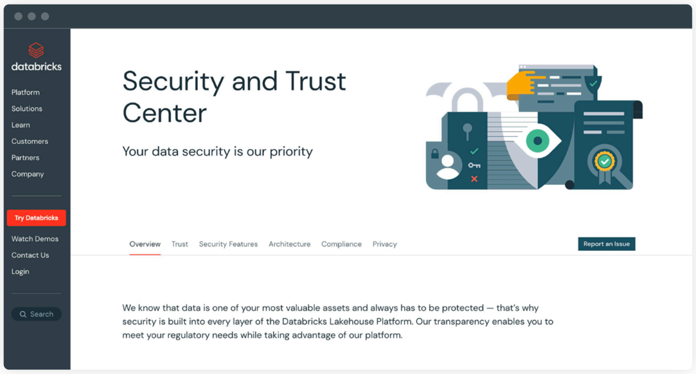
- Control plane <-> data plane is encrypted
- Offers optional intra-cluster encryption
- Customer code can be written to avoid unencrypted services (e.g., FTP)

Activities of Databricks users, including data access and commands, are logged and can be delivered automatically to a cloud storage bucket. Customers can also monitor provisioning activities by monitoring cloud audit logs.

Learn more

For more information, see our Security and Trust Center, our documentation and our detailed Enterprise Security Guide. Databricks provides an enterprise-ready cloud platform that is built on a strong platform security posture for organizations small and large, and across all industries.

We're happy to discuss your specific needs in more detail — please reach out to your Databricks representative or email sales@databricks.com.



Security and Trust Center

Your data security is our priority

We know that data is one of your most valuable assets and always has to be protected — that's why security is built into every layer of the Databricks Lakehouse Platform. Our transparency enables you to meet your regulatory needs while taking advantage of our platform.

About Databricks

Databricks is the lakehouse company. More than 9,000 organizations worldwide — including Comcast, Condé Nast and over 50% 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.