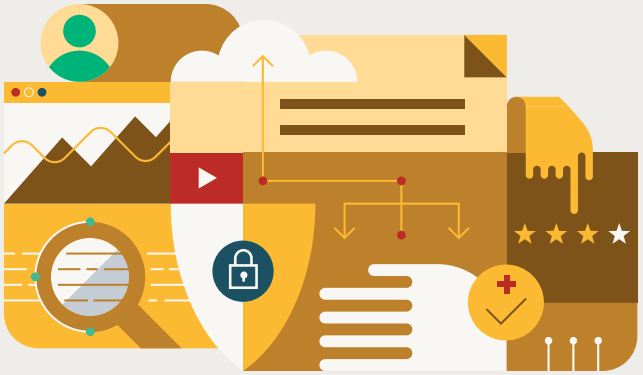
databricks

Databricks Platform Security

The Databricks platform provides unparalleled security for your data and users



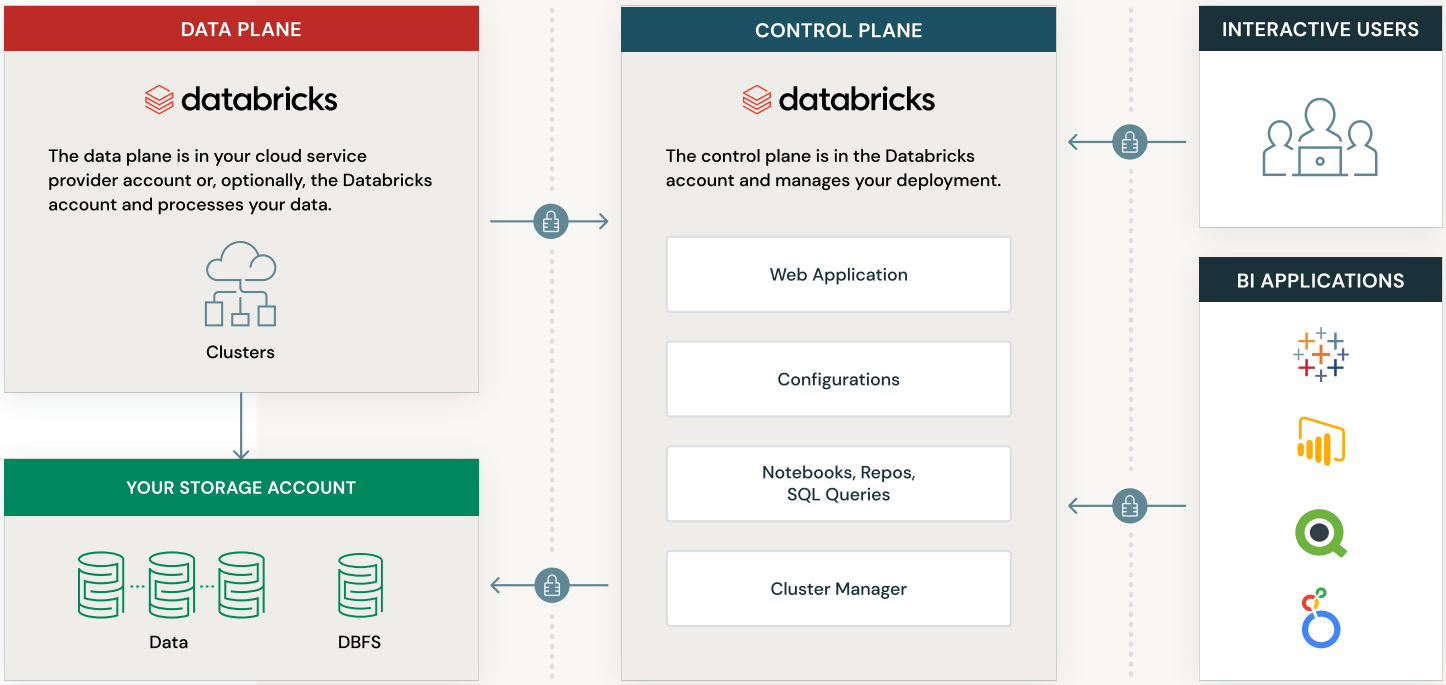
Thousands of customers trust Databricks with their most sensitive data to analyze and build data products using machine learning (ML). With significant investment into building a highly secure and scalable platform, Databricks delivers end-to-end platform security for data and users. This document provides 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.

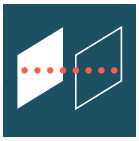
NOTE: This document assumes at least the Premium tier Databricks subscription. Databricks is available on AWS, GCP and Azure. Learn more in the Databricks Platform Overview.

Architecture

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

DATA PLANE	CONTROL PLANE
This is in your cloud service provider account or, optionally, the Databricks account and processes your data. Includes Databricks runtime clusters and networking.	This is in the Databricks cloud service provider account and manages your deployment. Includes web app, jobs, cluster management, notebooks.





End-to-end 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 stored in object storage (such as S3, GCS or Blob store).

Six steps occur in the example

1. Your single sign-on (such as Okta) seamlessly authenticates the data engineer via SAML to the Databricks web UI in the control plane, hosted in the Databricks account. 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, authenticating with a personal access token.
3. When ready, the control plane uses cloud service provider APIs to create a Databricks cluster made of new instances in the data plane, typically in your account. Administrators can apply cluster policies to control costs.
4. Once the instances launch, the cluster manager sends the data engineer's code to the cluster.
5. The cluster pulls from Kafka in your account, transforms the data in your account and writes it to a bucket/blob in your account.
6. The cluster reports status and any outputs back to the cluster manager.

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

Most Databricks customers host the data plane in their own cloud service provider account, but we also have a serverless mode in which the data plane is hosted in a Databricks account. The location of the data plane changes where your code runs, but doesn't impact other parts of the architecture.

Whichever location you choose for hosting the data plane, a key benefit of our model is that the vast majority of your data remains in systems under your control, typically in your account. While certain data, such as notebooks, configurations, Apache Spark™ logs and user information, is present within the control plane, that information is encrypted at rest within the control plane, and communication to and from the control plane is encrypted in transit. You also have choices for where certain data lives: You can host your own store of metadata about your data tables (Hive metastore), store partial interactive query results in your account, and you decide whether to use the Databricks Secrets API.

Network and server security

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

Networking

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 customer-managed VPC or VNet. The serverless data plane network infrastructure is managed by Databricks in a Databricks cloud service provider account and is 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.

Customers at the Enterprise tier can also use the IP access list feature on the control plane to limit which IP addresses can connect to the web UI or REST API — for example, to allow only VPN or office IPs.

Servers

In the data plane, 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. 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.

Automated access allows the Databricks control plane to configure resources in your environment using the cloud service provider APIs. 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).

Databricks has a custom-built system that allows staff to fix issues or support you — for example, when you open a support request and check the box authorizing access to your workspace. Access requires either a support ticket or engineering ticket tied expressly to your workspace, and is limited to a subset of employees and for limited time periods. Additionally, if you've configured audit log delivery, the audit logs show the initial access event and the staff's actions.

Identity and access

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.

Databricks supports a variety of ways to enable users to access their data. Examples include:

The Table ACLs feature uses traditional SQL-based statements to manage access to data and enable fine-grained view-based access.

IAM instance profiles enable AWS clusters to assume an IAM role, so users of that cluster automatically access allowed resources without explicit credentials.

External storage can be mounted or accessed using a securely stored access key.

The Secrets API separates credentials from code when accessing external resources.

Compliance

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

Certain clouds support Databricks deployment options for FedRAMP High, HITRUST, HIPAA and PCI. Databricks Inc. and the Databricks platform are also GDPR and CCPA ready.

SOC 2 Type II
ISO 27001
ISO 27017
ISO 27018

Data security

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

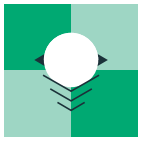
For data-in-motion encryption:

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

Customers can isolate users at multiple levels:

Workspace level

Each team or department can use a separate workspace



High concurrency clusters

Process isolation, JVM whitelisting and limited languages (SQL, Python) allow for the safe coexistence of users of different privilege levels, and is used with Table ACLs

Cluster level

Cluster ACLs can restrict the users who can attach notebooks to a given cluster

Single-user cluster

Users can create a private dedicated cluster

Activities of Databricks users 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 the full 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.

About Databricks

Databricks is the data and AI company. More than 5,000 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.