DATA+AI SUMMIT 2022

# Lakehouse Security

## Security Best Practices and Threat Model for the Lakehouse

**David Veuve**

Head of Security Field Engineering, Databricks

**Arun Pamulapati**

Sr. Staff Security Field Engineer, Databricks

ORGANIZED BY databricks

# whoarewe

- David Veuve
  - Head of Security Field Engineering
- At work I am a
  - Security nerd
  - Content creator
  - Enabler

- I help with
  - Databricks platform security
  - Security detection and response
- I survived quarantine thanks to

- Arun Pamulapati
  - Sr. Staff Security Field Engineer
- At work I am a
  - Friends of the field
  - Builder
  - Collaborator

- I help with
  - Databricks platform security
  - Security use cases (Okta, DNS …log analytics.)
- I survived quarantine thanks to

# Thesis

Databricks has built the security features
Databricks has helped 1000s of customers
Databricks has seen what works

Let’s just tell you what works

# 35 minutes to cover:

- 60+ Slides
- 17 Page Whitepaper
- Self-service tool

# What you will learn today

- Databricks Lakehouse Architecture
- Top threats impacting Databricks
- Example controls
- Where to get the full whitepaper
- How to analyze an existing deployment

# tl;dr

## Security Architecture and Controls

The security & trust center outlines our architecture and enumerates the key security features.

Overview: databricks.com/trust/whitepaper
Detail: databricks.com/trust

## Prescriptive Best Practices Guides

Databricks has captured our best practices into a doc with “common” and “high-security” models and checkboxes. *Check those boxes.*

databricks.com/trust/Security-Best-Practices_Databricks-on-AWS.pdf

## Meet Compliance Needs

Have your security team download our due diligence package (ISO certs, Pen Test) and reach out to your account team for SOC 2, Enterprise Security Guide

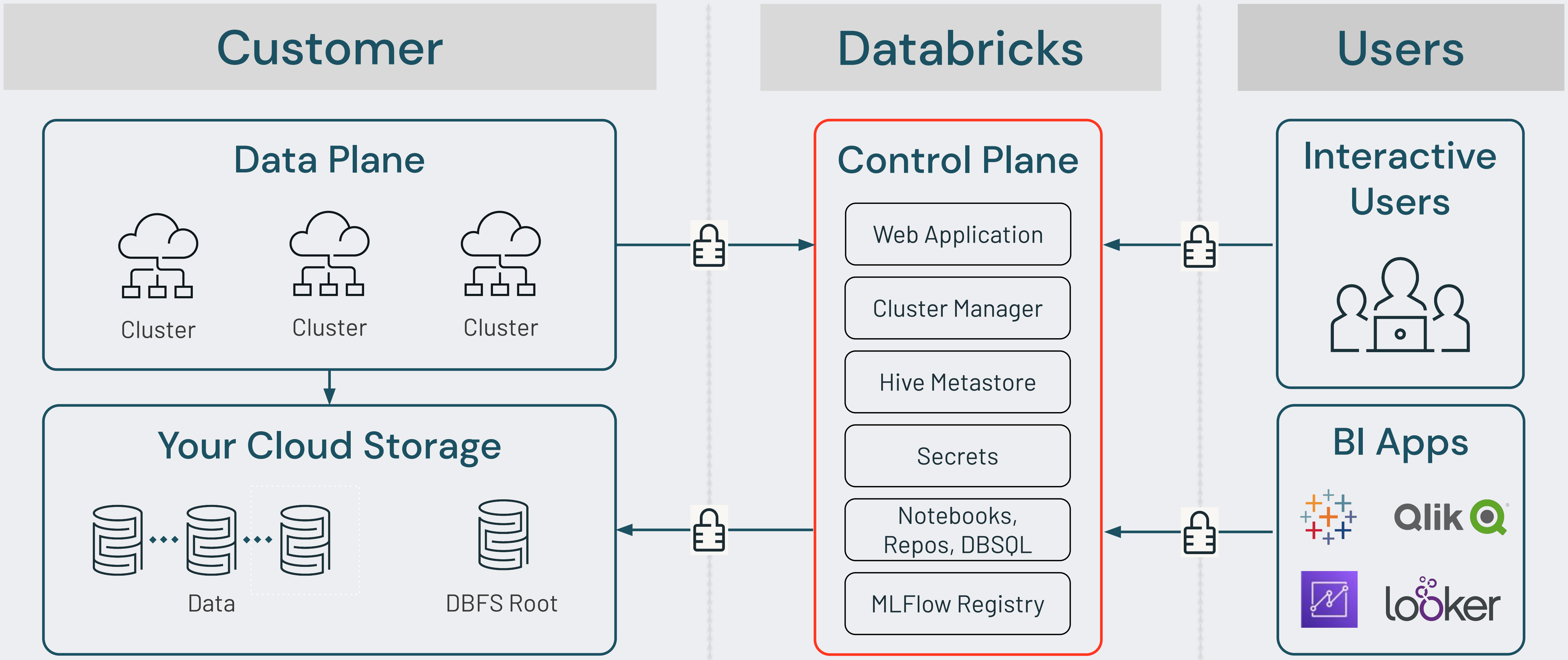
Compliance: databricks.com/trust#compliance

## Analyze your deployment

The Workspace Analysis tool now contains a security section! Check your workspace against our most common best practices.
Demo at the end of this talk!

Coming soon to the Databricks Blog!

# Databricks Lakehouse Architecture



Customer

Databricks

Users

Data Plane

Your Cloud Storage

Control Plane

- Web Application
- Cluster Manager
- Hive Metastore
- Secrets
- Notebooks, Repos, DBSQL
- MLFlow Registry

Interactive Users

BI Apps

# What Are The Threats?

# Account Takeover

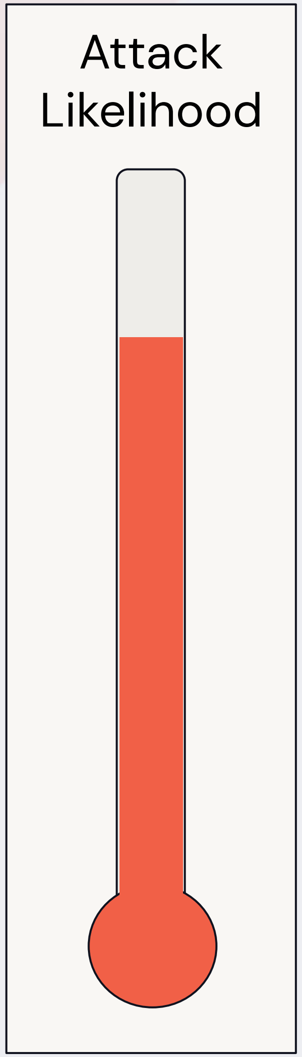
Attackers gain the credentials/access of your users

## Risk Overview

- Customers often analyze sensitive datasets
- Compromised end-user credentials grant access
  - Phishing, brute force, etc.

## Best Practices to Mitigate

- Required two-factor auth on your identity provider
  - Consider FIDO Key
- Manage local passwords
- Use SCIM to deprovision
- IP Access Lists or PrivateLink
- Monitor Audit Log
- Limit token lifetime



# Accidental Insider Exposure

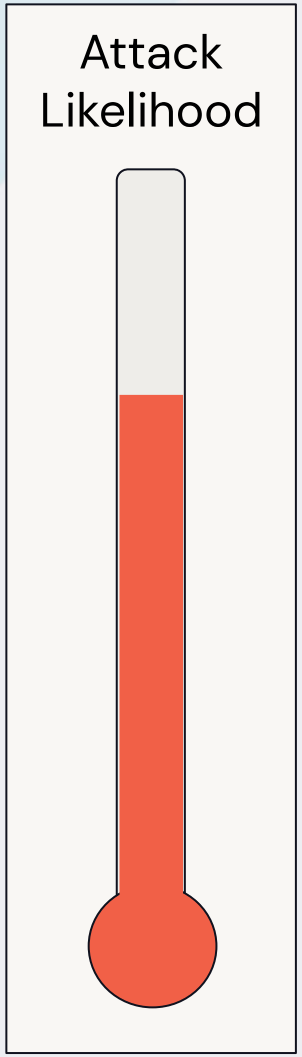
## Resource abuse or accidental copy by insiders

### Risk Overview

- Someone believes their job is easier without pesky security controls
- Data is copied where it shouldn't be, or where ACLs aren't applied

### Best Practices to Mitigate

- Backup your data and code
- Run most sensitive ETL through a CI/CD process (code review)
- Utilize secured models like Table ACLs for limited exposure
- Limit data in DBFS and monitor for large datasets
- Deploy data exfiltration protections



# Data Exfiltration

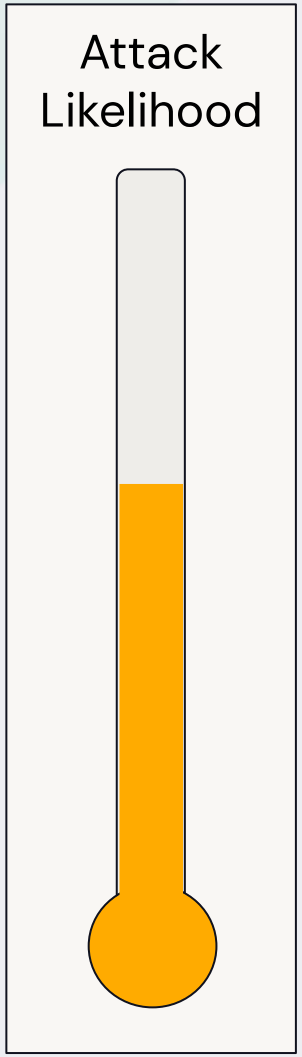
Data stolen by an attacker or malicious insider

## Risk Overview

- Risk is of a user sending sensitive data out to some external location
- Data plane needs connectivity, by default has full outbound
- Also consider access to cloud storage (S3 buckets, etc.)

## Best Practices to Mitigate

- Use Customer-managed VPC / VNet Injection (baseline)
- Route traffic through a firewall or proxy to limit destination
- [AWS] VPC Endpoint Policies
- Limit access to sensitive data
- Configure data exfiltration settings in the console



# Also covered in the docs

## Resource Abuse

- Customer cloud infrastructure hijacking for crypto mining
- Accidental/abusive waste of customer resources

## Compromise of Databricks, Inc.

- Compromise of Databricks Inc user or system could result in compromise of customer environment

# What is a security team to do?

# We want to securely authenticate users

Account Takeover

DE

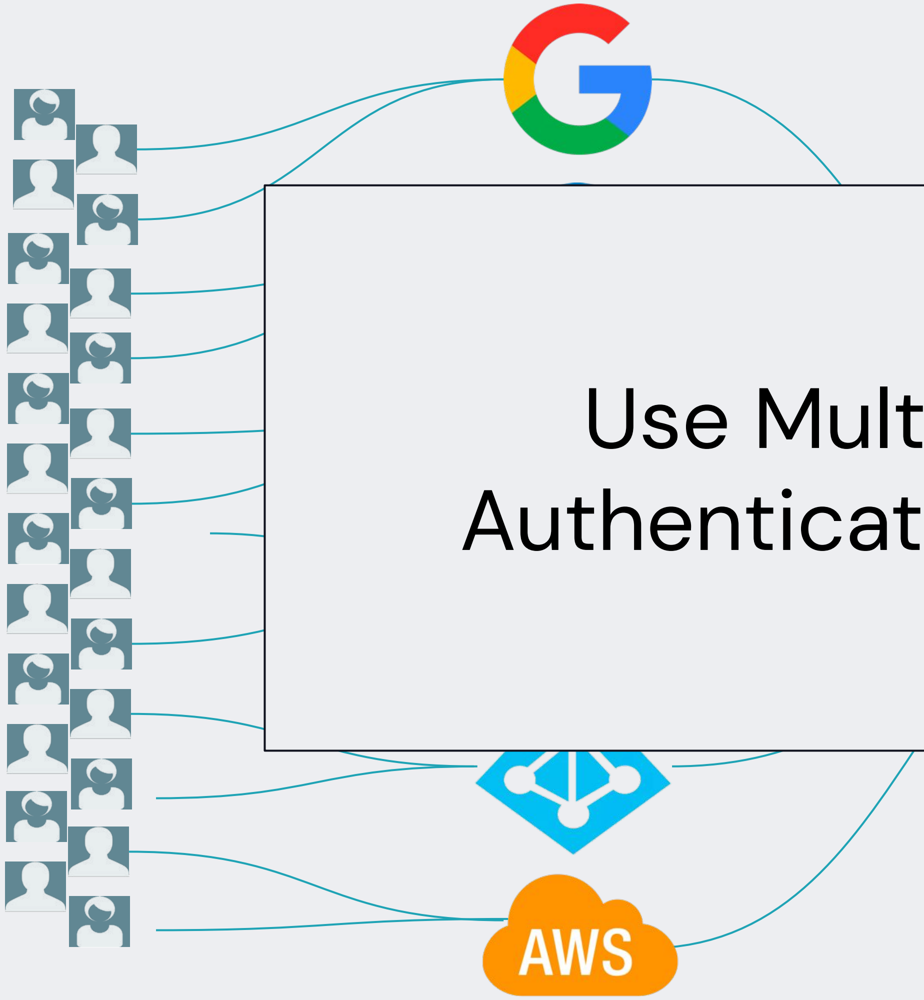
AIE

RA

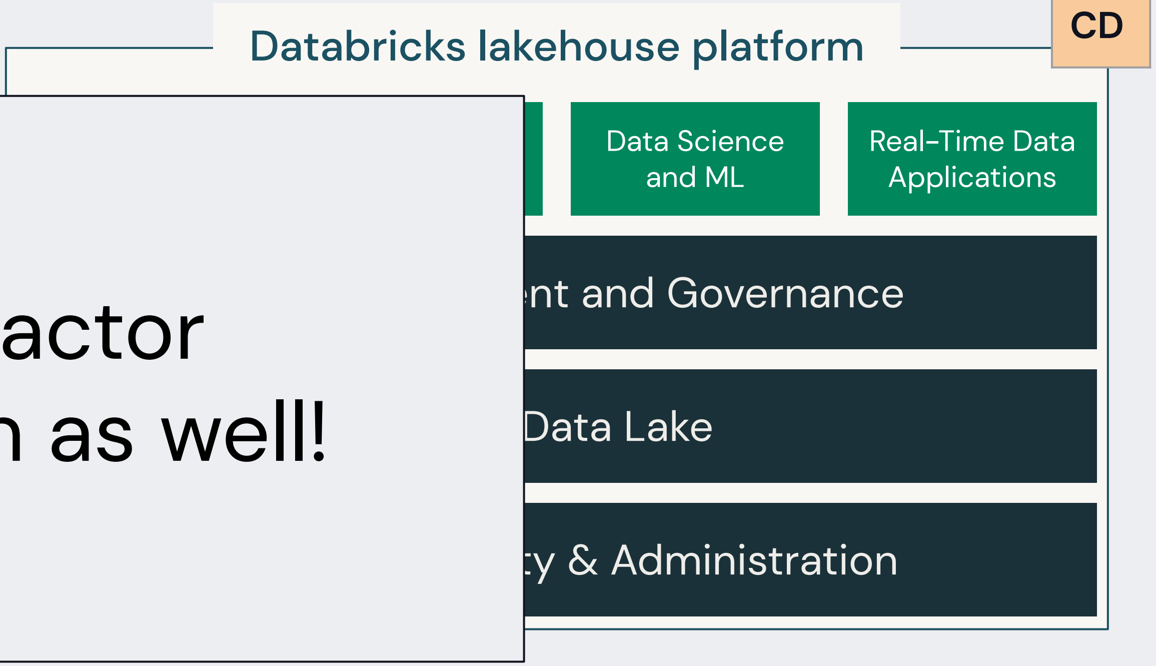
CD

# Authenticate via single sign-on

Users



Use Multi-factor Authentication as well!



- SAMLv2, SCIM support
- Documented integrations with six partners

Account Takeover

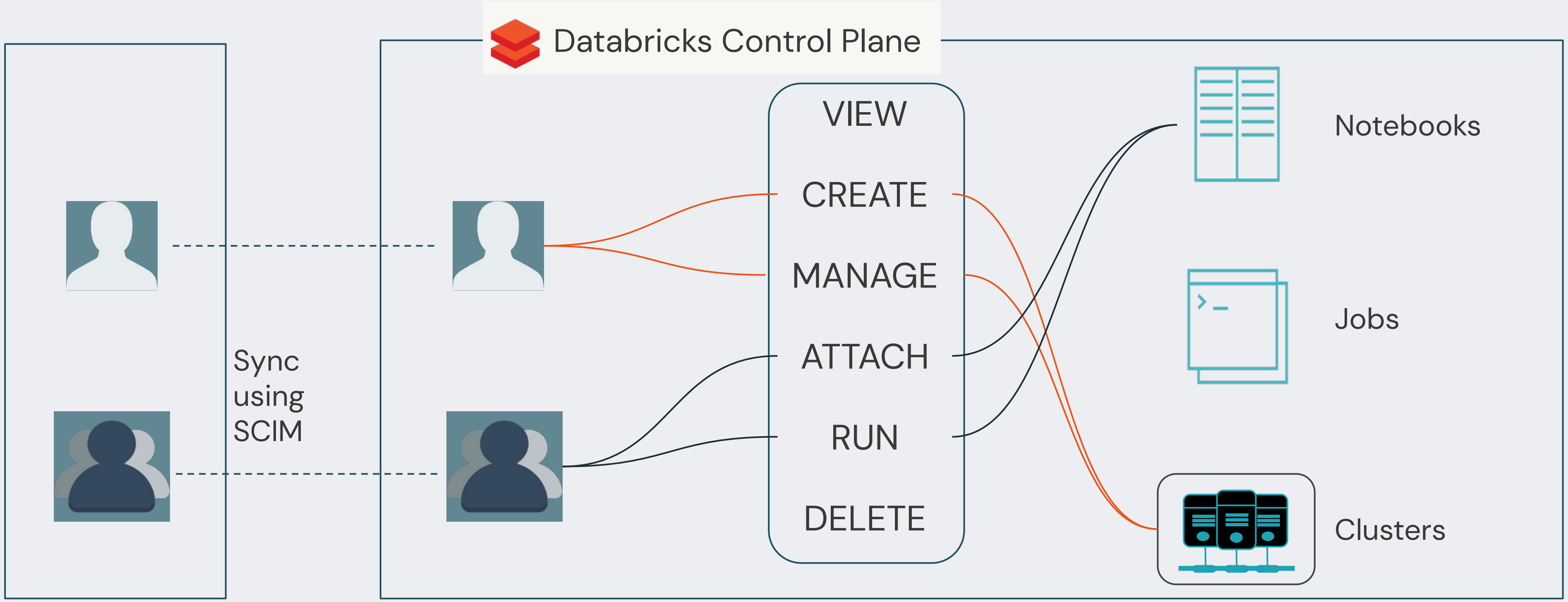
DE

Accidental Insider Exposure

RA

CD

# SCIM & Role-based Access Control



Admins set ACLs via the UI or Permissions API

ACLs for:

- Jobs
- Notebooks
- Clusters
- Pools
- Tables
- Workspaces
- Secrets
- ML Experiments

# We want to encrypt data

# Encrypt buckets and restrict access

AT

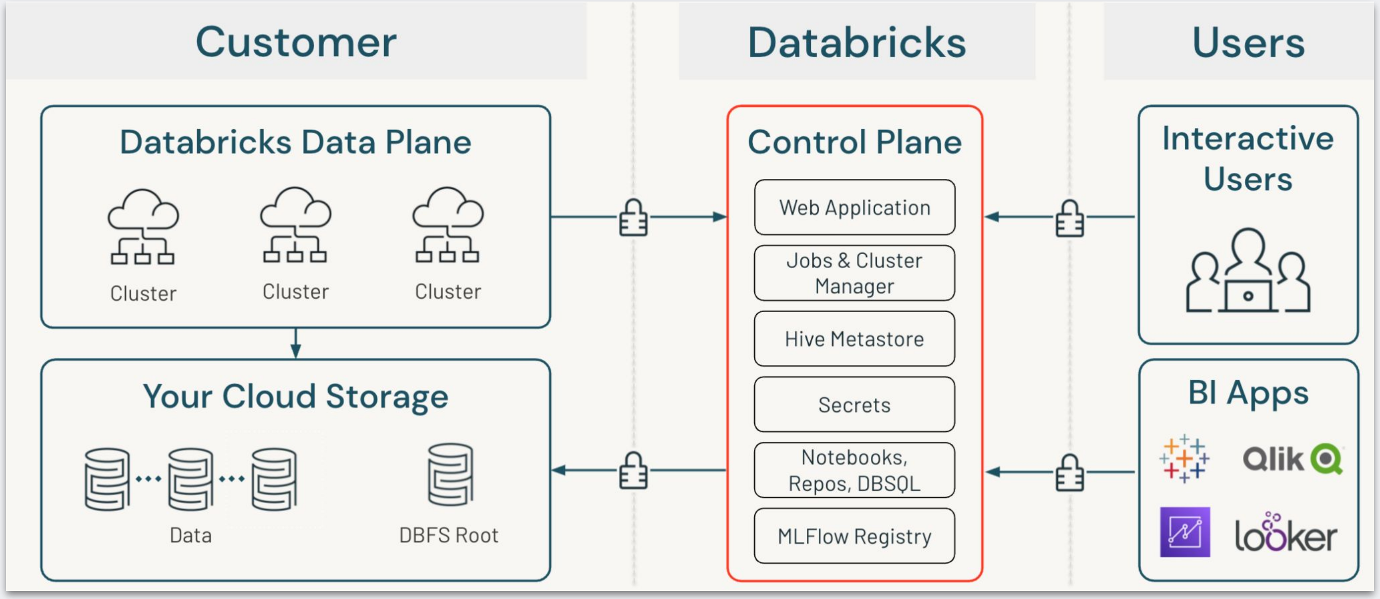
Data Exfiltration

AIE

RA

CD

- Encrypt your buckets
- Don't allow the public to access your buckets



# Customer-managed keys

(AWS, Azure)

| AT |
|---|
| DE |
| AIE |
| RA |

Compromise of Databricks Inc.

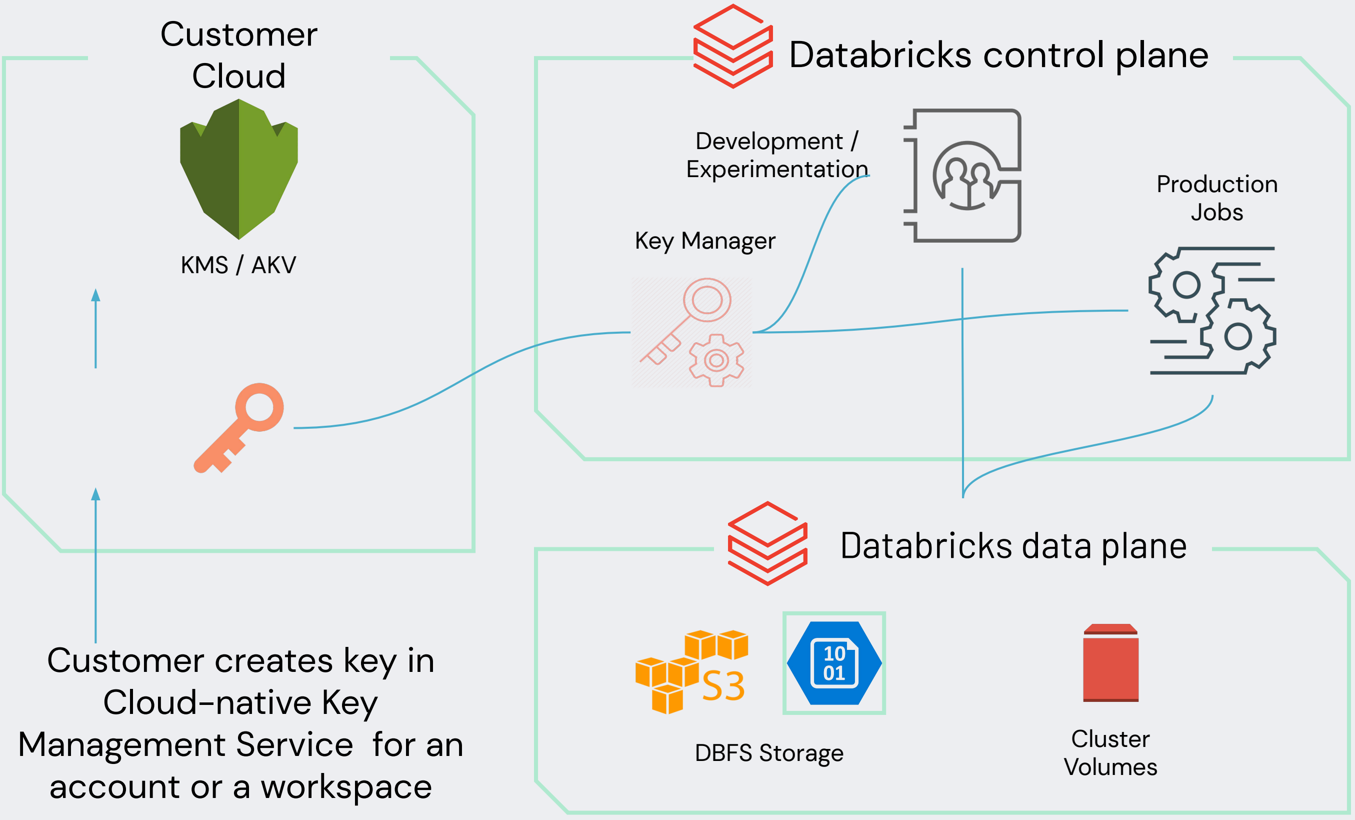
Customers can enable customer-managed keys

- CMK for Managed Services

  Priority control plane data

- CMK for Workspace Storage

  DBFS root storage and (AWS-only) EBS volumes



Customer creates key in Cloud-native Key Management Service for an account or a workspace

We want to isolate different workloads

AT

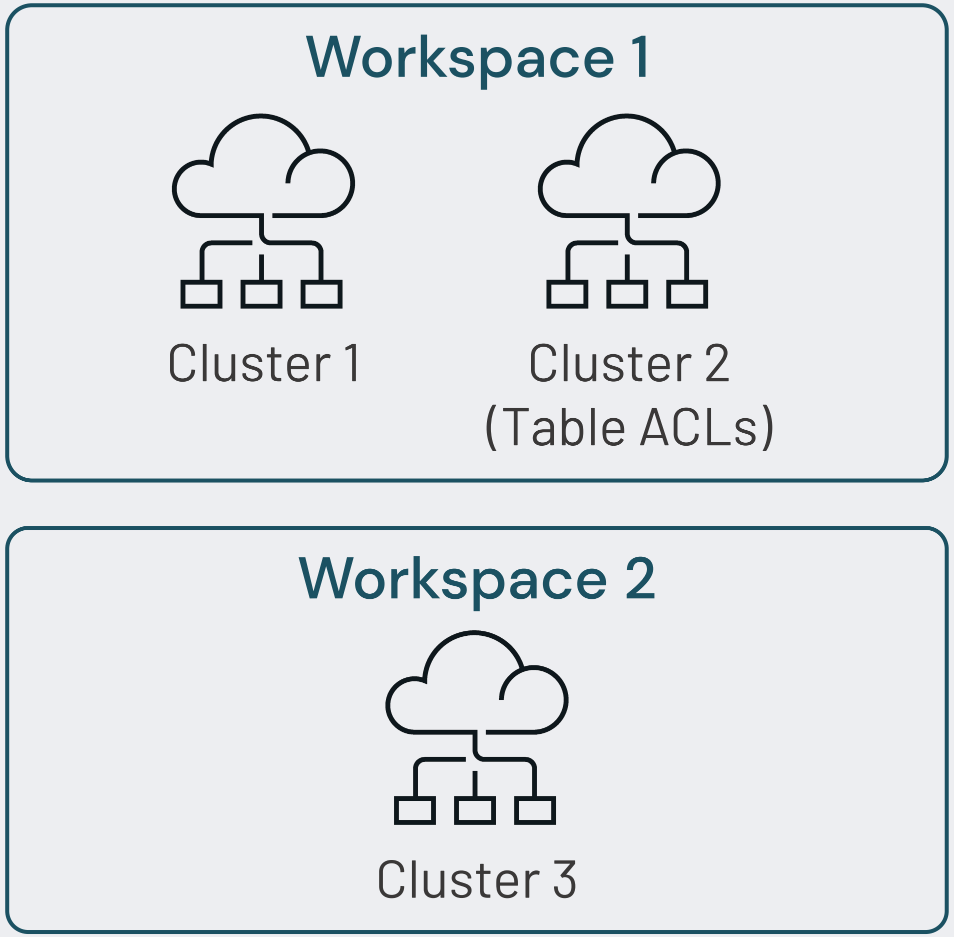
Data Exfiltration

AIE

RA

CD

# Design workload isolation



Easily isolate based on your needs

1. Table ACLs supports users with different privileges
2. Multiple workspaces isolate groups who won’t collaborate
3. Limit usage of standard clusters

AT

Data Exfiltration

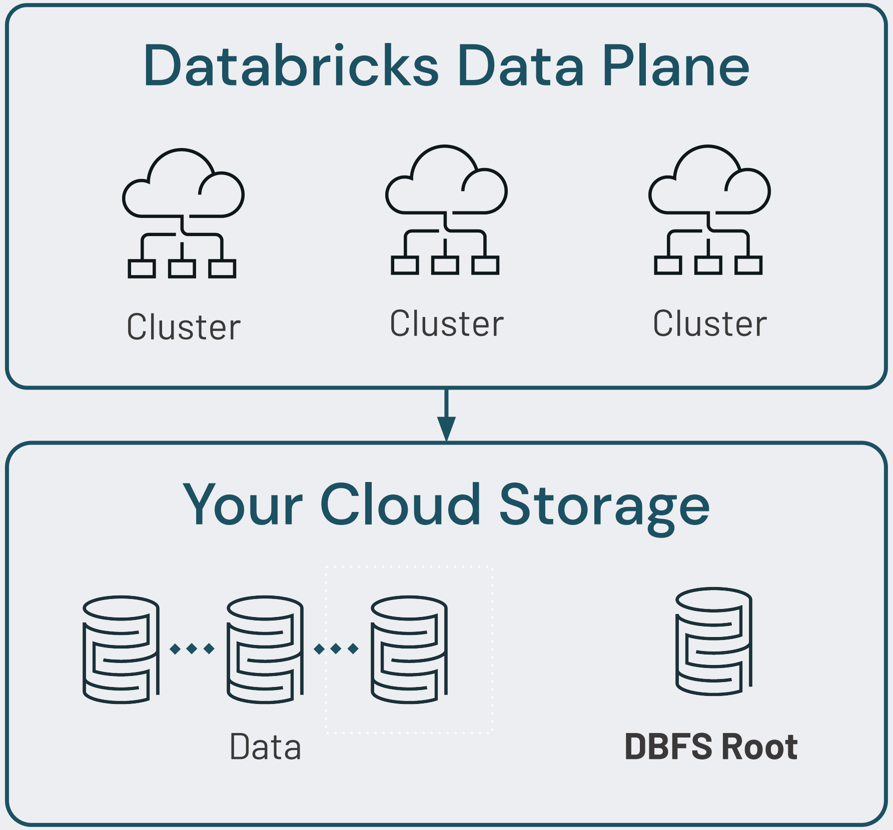
AIE

RA

CD

# Avoid storing production data in DBFS

- DBFS (Databricks Filesystem) is accessible to all workspace users
- Instead, store data in buckets / ADLS
- AWS customers can use bucket policy to limit access



We want to prevent data loss

AT

Data Exfiltration

AIE

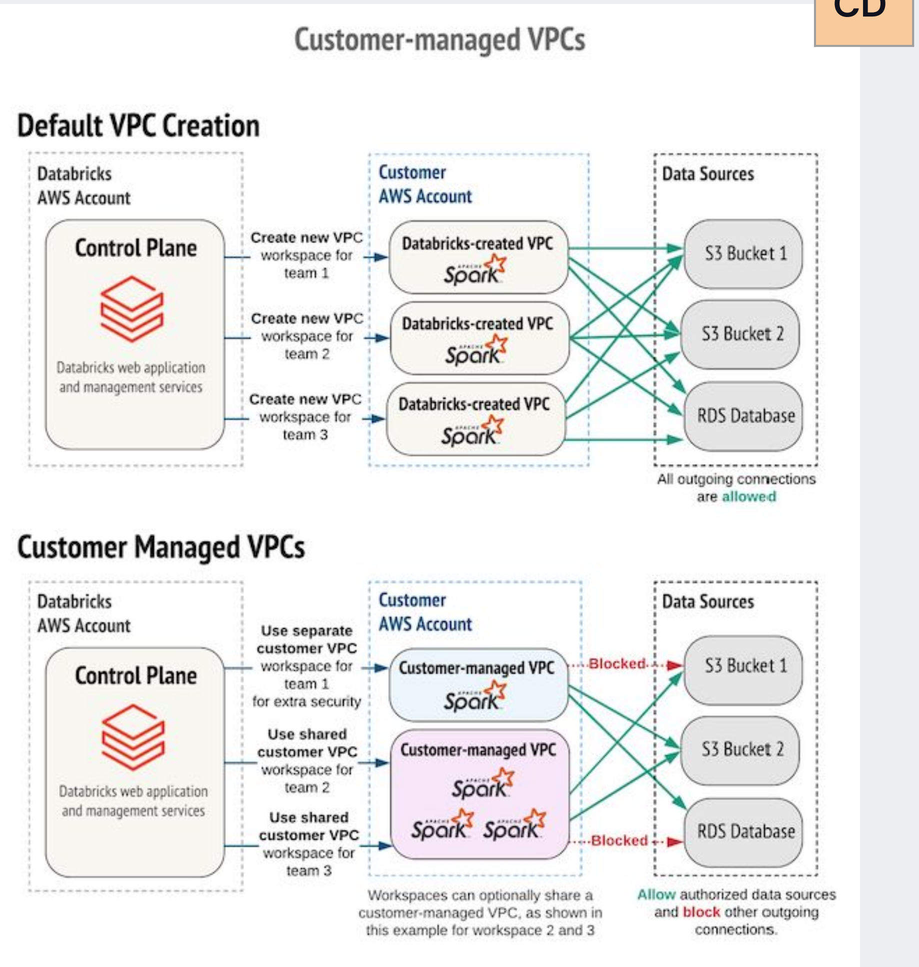
RA

CD

# Deploy with a customer-managed VPC

## Azure name: VNet Injection

- Limit outgoing connections: Use a firewall or proxy to limit outbound traffic.
- (AWS-only) Lower privilege level: maintain more control of your own AWS account.
- Simplified network operations: Better network space utilization.
- Consolidation of VPCs: Multiple data planes can share a single VPC



Customer-managed VPCs
Default VPC Creation
Databricks AWS Account
Control Plane
Databricks web application and management services
Create new VPC workspace for team 1
Create new VPC workspace for team 2
Create new VPC workspace for team 3
Customer AWS Account
Databricks-created VPC
Spark
Databricks-created VPC
Spark
Databricks-created VPC
Spark
Data Sources
S3 Bucket 1
S3 Bucket 2
RDS Database
All outgoing connections are allowed
Customer Managed VPCs
Databricks AWS Account
Control Plane
Databricks web application and management services
Use separate customer VPC workspace for team 1 for extra security
Use shared customer VPC workspace for team 2
Use shared customer VPC workspace for team 3
Customer AWS Account
Customer-managed VPC
Spark
Customer-managed VPC
Spark
Spark
Spark
Blocked
Blocked
Data Sources
S3 Bucket 1
S3 Bucket 2
RDS Database
Workspaces can optionally share a customer-managed VPC, as shown in this example for workspace 2 and 3
Allow authorized data sources and block other outgoing connections.

AT

Data Exfiltration

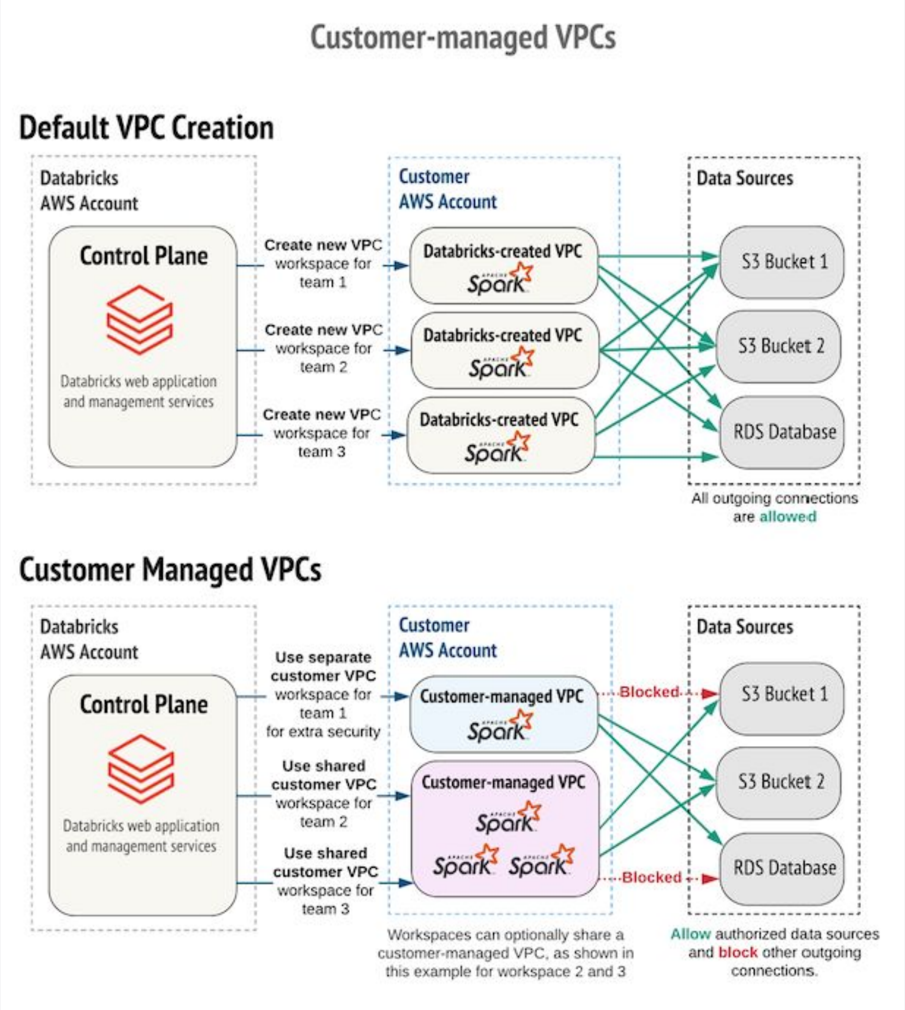
AIE

RA

CD

# Implement data loss protections

- Lock down outbound access with customer managed VPC
- Restrict outbound access
- Restrict access to storage



Customer-managed VPCs

Default VPC Creation
Databricks AWS Account
Control Plane
Databricks web application and management services
Create new VPC workspace for team 1
Create new VPC workspace for team 2
Create new VPC workspace for team 3
Customer AWS Account
Databricks-created VPC
Spark
Data Sources
S3 Bucket 1
S3 Bucket 2
RDS Database

All outgoing connections are allowed

Customer Managed VPCs
Databricks AWS Account
Control Plane
Databricks web application and management services
Use separate customer VPC workspace for team 1 for extra security
Use shared customer VPC workspace for team 2
Use shared customer VPC workspace for team 3
Customer AWS Account
Customer-managed VPC
Spark
Blocked
Data Sources
S3 Bucket 1
S3 Bucket 2
RDS Database

Workspaces can optionally share a customer-managed VPC, as shown in this example for workspace 2 and 3

Allow authorized data sources and block other outgoing connections.

We want to have a record of what happens with our data and detect user compromise

# Databricks audit log

Account Takeover

Data Exfiltration

Accidental Insider Exposure

Resource Abuse

Compromise of Databricks Inc.

## Audit Logging

- Customers can configure near-real-time logging
  - (AWS/GCP) to a bucket owned by the customer
  - (Azure) to diagnostic logging
- (AWS) Cloudtrail logs also includes provisioning activities

## System Logs

- Understand system activities via system logs, including stdout, stderr, etc.
- Use metrics to understand utilization and health

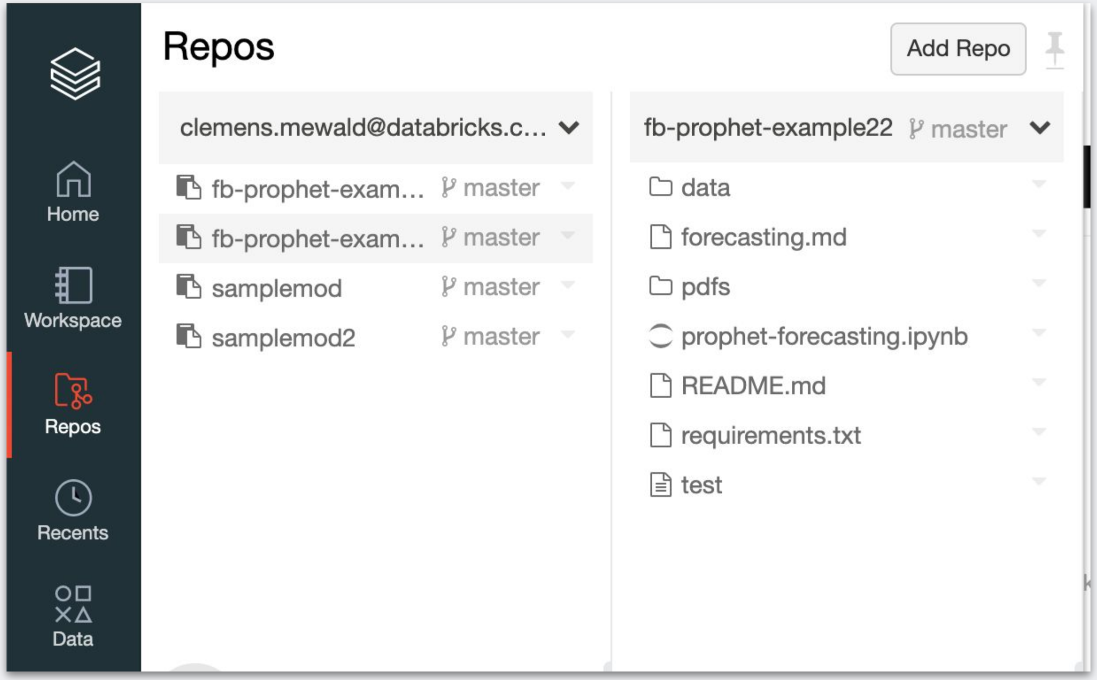
Ready-to-use analysis notebooks
on our blog!
(Linked from whitepaper)

# We want to avoid loss of data

AT
DE
AIE
RA
CD

# Backup your control plane data

- Backup via the Databricks migration tool
- Databricks Repos moves your code storage directly to git



Repos

Add Repo

| clemens.mewald@databricks.c... | |
|---|---|
| fb-prophet-exam... | master |
| fb-prophet-exam... | master |
| samplemod | master |
| samplemod2 | master |

fb-prophet-example22 master

- data
- forecasting.md
- pdfs
- prophet-forecasting.ipynb
- README.md
- requirements.txt
- test

# Manage code run via CI/CD

Mature organizations will often build their production workloads via CI/CD:

- CI/CD is very compatible in a Databricks environment.
- For security:
  - Integrate code scanning, better provide for permissions, perform linting, and more
  - Scan for passwords, token in the code

# We want to avoid leaking passwords

Account Takeover

DE

AIE

RA

CD

# Store and use secrets securely

- Securely store API Keys or other credentials and reference by location
- Secrets help avoid hard-coded credentials in code, but doesn’t hide them from authorized users

```
api_key = dbutils.secrets.get(
                    scope="myScope",
                    key="mySecret"
          )
```

# Use AWS Nitro instances

AWS Nitro instances provide two major security benefits:

- NVMe disks automatically encrypt at rest (local SSDs for shuffle data)
  - Azure can enable software encryption
  - GCP automatically has encryption
- Many Nitro instances automatically encrypt data in transit between hosts
  - Azure can enable intra-cluster encryption in software
  - GCP automatically has inter-host encryption

# Control user network access

Account Takeover

Data Exfiltration

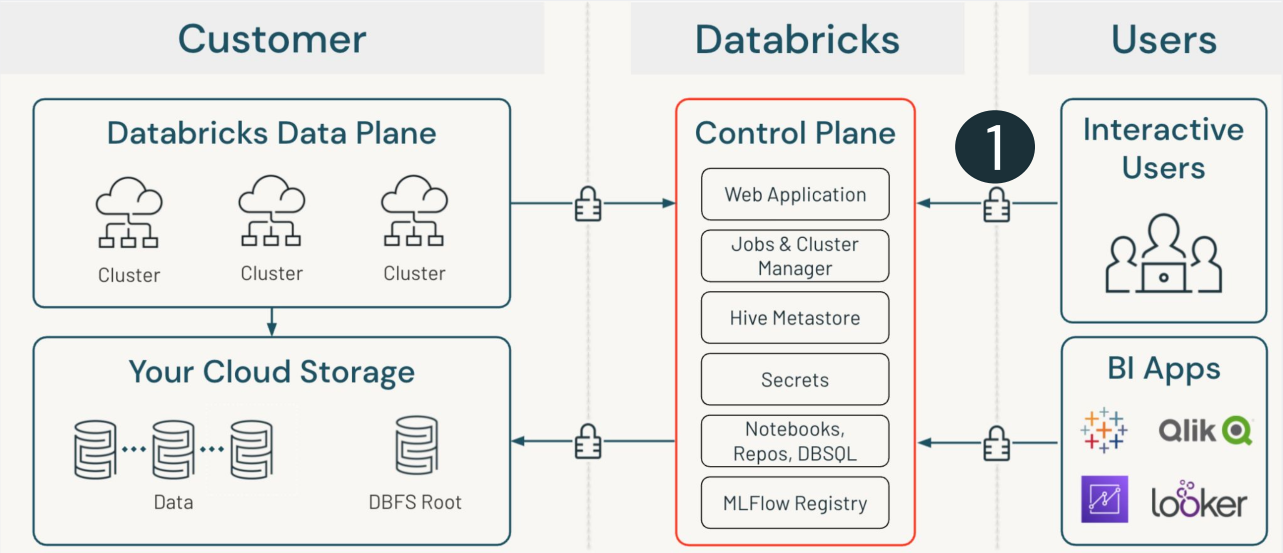
AIE

RA

CD

Two options:

- IP Access Lists Enterprise
- User to Control Plane Private Link (Preview) Enterprise
  - Not available on GCP



Account Takeover

Data Exfiltration

AIE

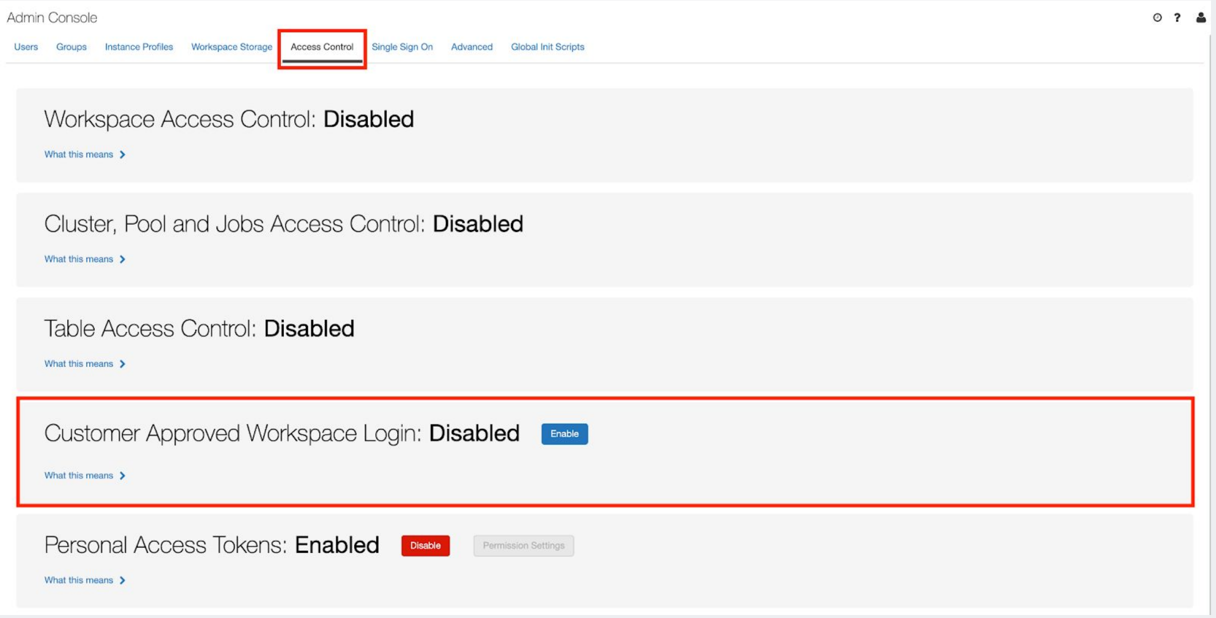
RA

CD

# Configure the admin console settings

Manage key configuration settings

- Can Databricks staff access your workspace for support?
- Can notebooks be exported?
- Can results be downloaded?
- Where are query results stored?
- Are ACLs enabled?



Workspace Access Control: **Disabled**
What this means ›

Cluster, Pool and Jobs Access Control: **Disabled**
What this means ›

Table Access Control: **Disabled**
What this means ›

Customer Approved Workspace Login: **Disabled**
What this means ›

Personal Access Tokens: **Enabled**
What this means ›

Account Takeover

DE

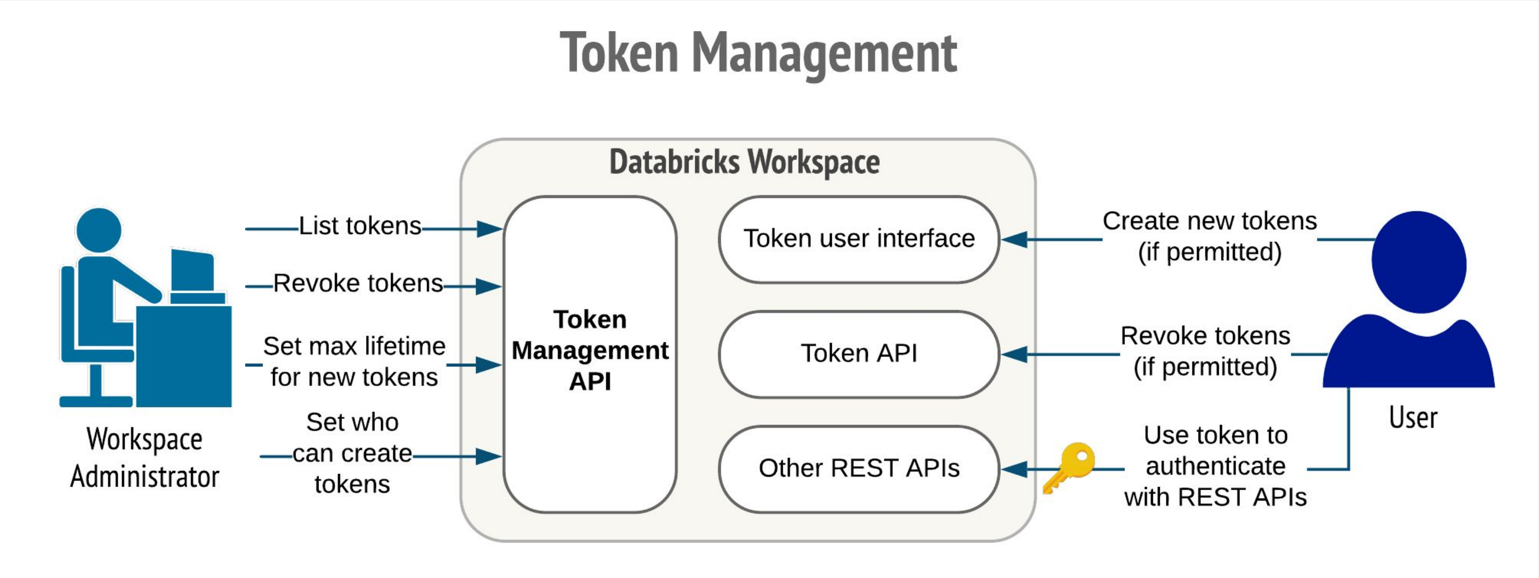
AIE

RA

CD

# Token management

- Enable or disable personal access tokens (PATs) for some or all users
- Configure a max token lifetime for new tokens


Token Management
Databricks Workspace
List tokens
Revoke tokens
Set max lifetime for new tokens
Set who can create tokens
Workspace Administrator
Token Management API
Token user interface
Token API
Other REST APIs
Create new tokens (if permitted)
Revoke tokens (if permitted)
Use token to authenticate with REST APIs
User

We want to utilize private networking

# Configure Back-End Private Link

AT

Data Exfiltration

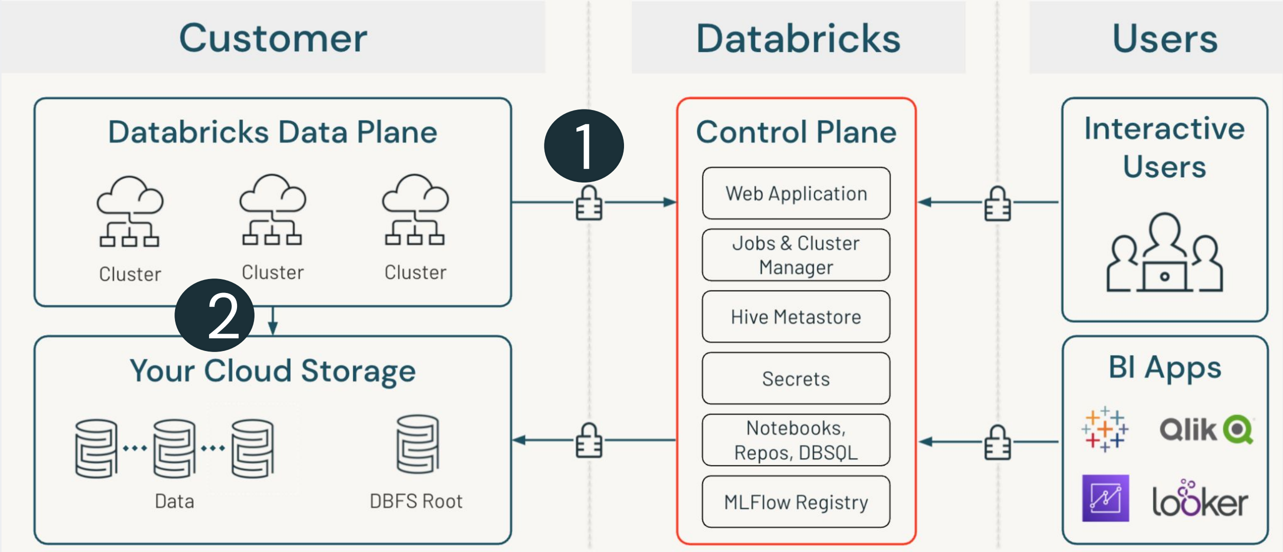
AIE

RA

CD

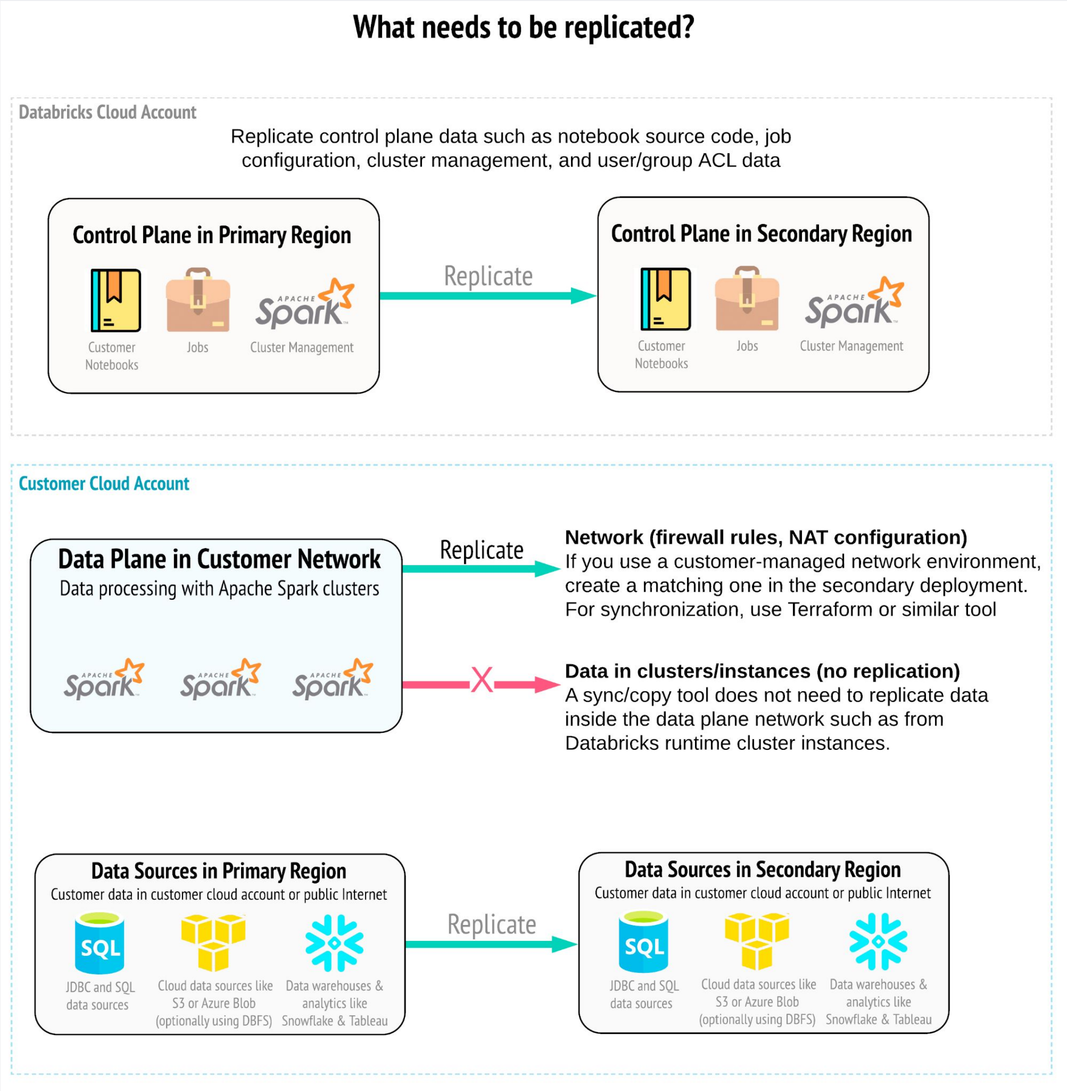
Multiple options:

1. Data Plane to Control Plane
   - By default, the cloud service provider backbone is very secure
   - (AWS / Azure) Use Data Plane to Control Plane PrivateLink (Preview) Enterprise
2. Data Plane to Storage Services
   - Public Endpoints
   - Private Endpoints
   - Private Link to your data sources



# Configure a DR site

- Understand your business needs
- Choose a process that meets your business needs
- Prep workspaces and do a one-time copy
- Prepare your data sources
- Implement and test your solution



**What needs to be replicated?**

Replicate control plane data such as notebook source code, job configuration, cluster management, and user/group ACL data

**Network (firewall rules, NAT configuration)**
If you use a customer-managed network environment, create a matching one in the secondary deployment. For synchronization, use Terraform or similar tool

**Data in clusters/instances (no replication)**
A sync/copy tool does not need to replicate data inside the data plane network such as from Databricks runtime cluster instances.

Databricks Cloud Account
Control Plane in Primary Region
Customer Notebooks
Jobs
Cluster Management
Replicate
Control Plane in Secondary Region
Customer Notebooks
Jobs
Cluster Management
Customer Cloud Account
Data Plane in Customer Network
Data processing with Apache Spark clusters
Replicate
X
Data Sources in Primary Region
Customer data in customer cloud account or public Internet
JDBC and SQL data sources
Cloud data sources like S3 or Azure Blob (optionally using DBFS)
Data warehouses & analytics like Snowflake & Tableau
Replicate
Data Sources in Secondary Region
Customer data in customer cloud account or public Internet
JDBC and SQL data sources
Cloud data sources like S3 or Azure Blob (optionally using DBFS)
Data warehouses & analytics like Snowflake & Tableau

We want to control configurations and costs

# Configure cluster policies

AT

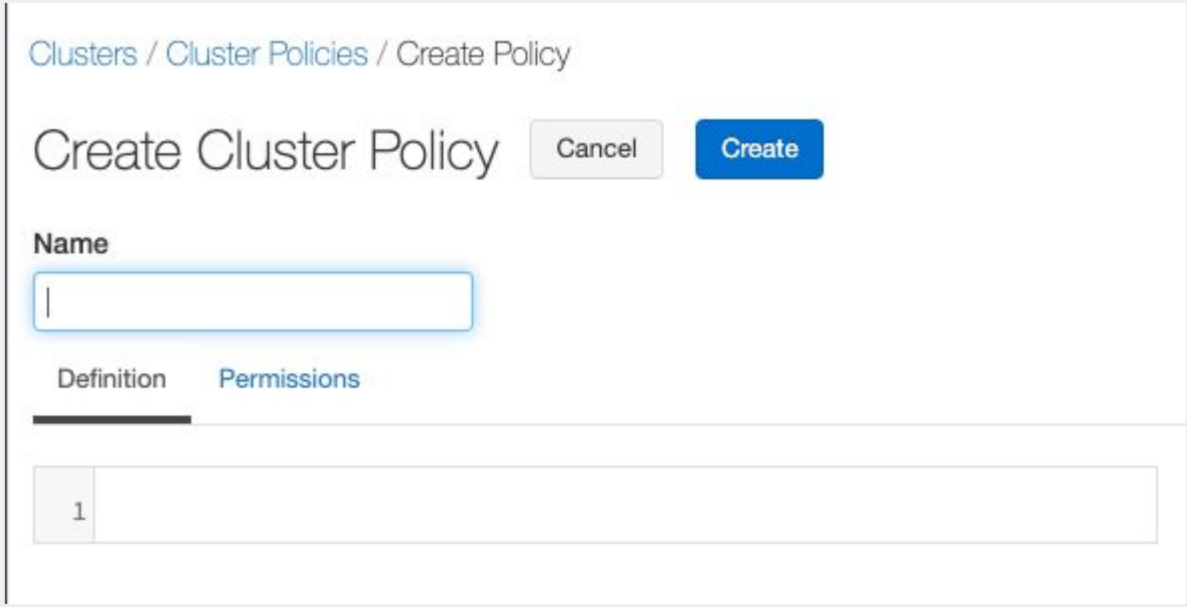
DE

AIE

Resource Abuse

CD

- Limit users to create clusters with prescribed settings.
- Simplify the user interface for your users
- Control cost by limiting per cluster maximum cost



# Configure tagging to monitor

AT

DE

AIE

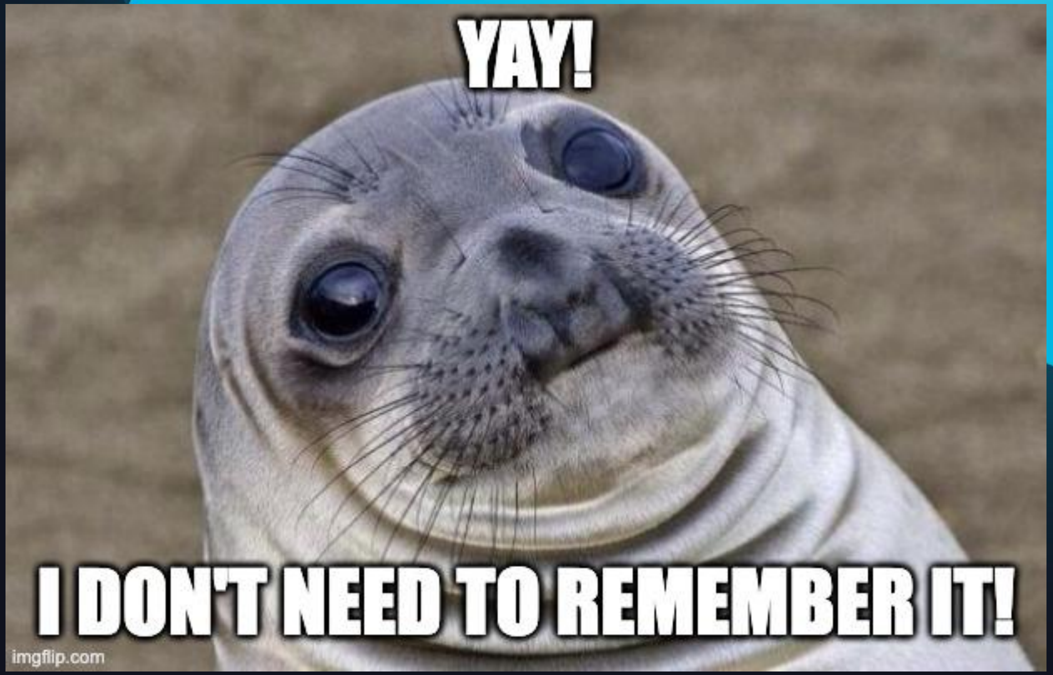
Resource Abuse

CD

- Monitor cost and accurately attribute Databricks usage to your organization's business units and teams (for chargebacks, for example), you can tag clusters and pools.
- These tags propagate both to detailed DBU usage reports and to your cloud service provider (e.g., AWS EC2 and AWS EBS)

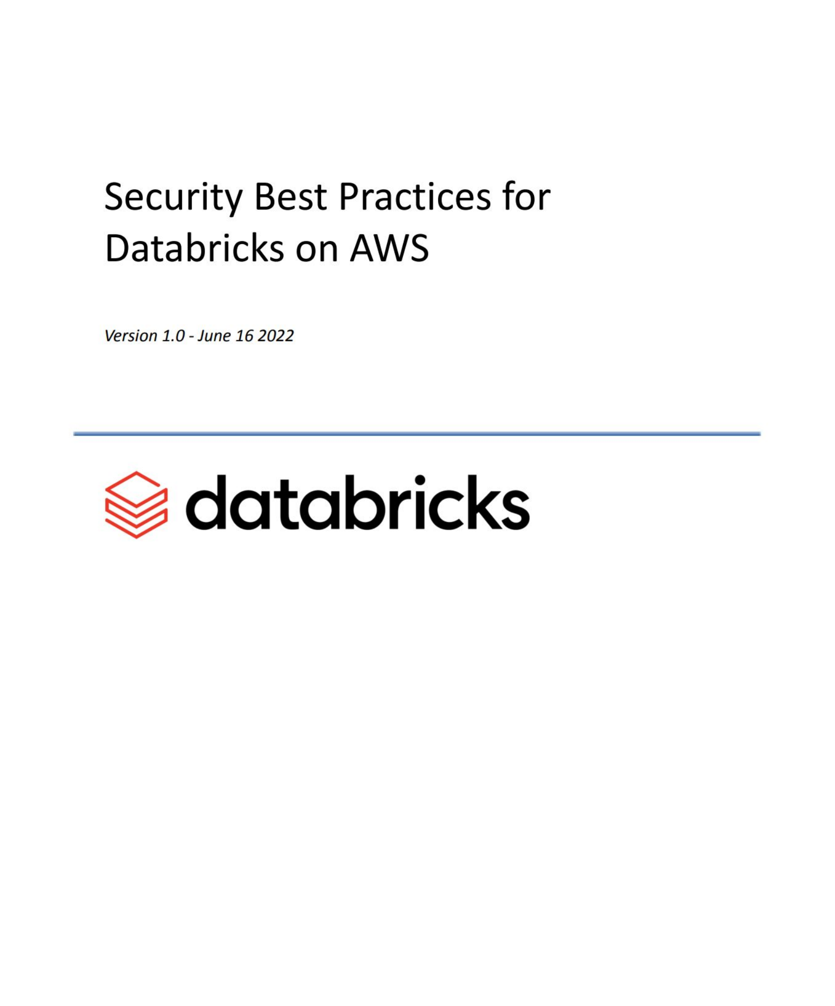
HELP!!
I REMEMBER NOTHING!
imgflip.com

Don’t worry seal.
We’ve got you.



# Security Best Practices Documentation

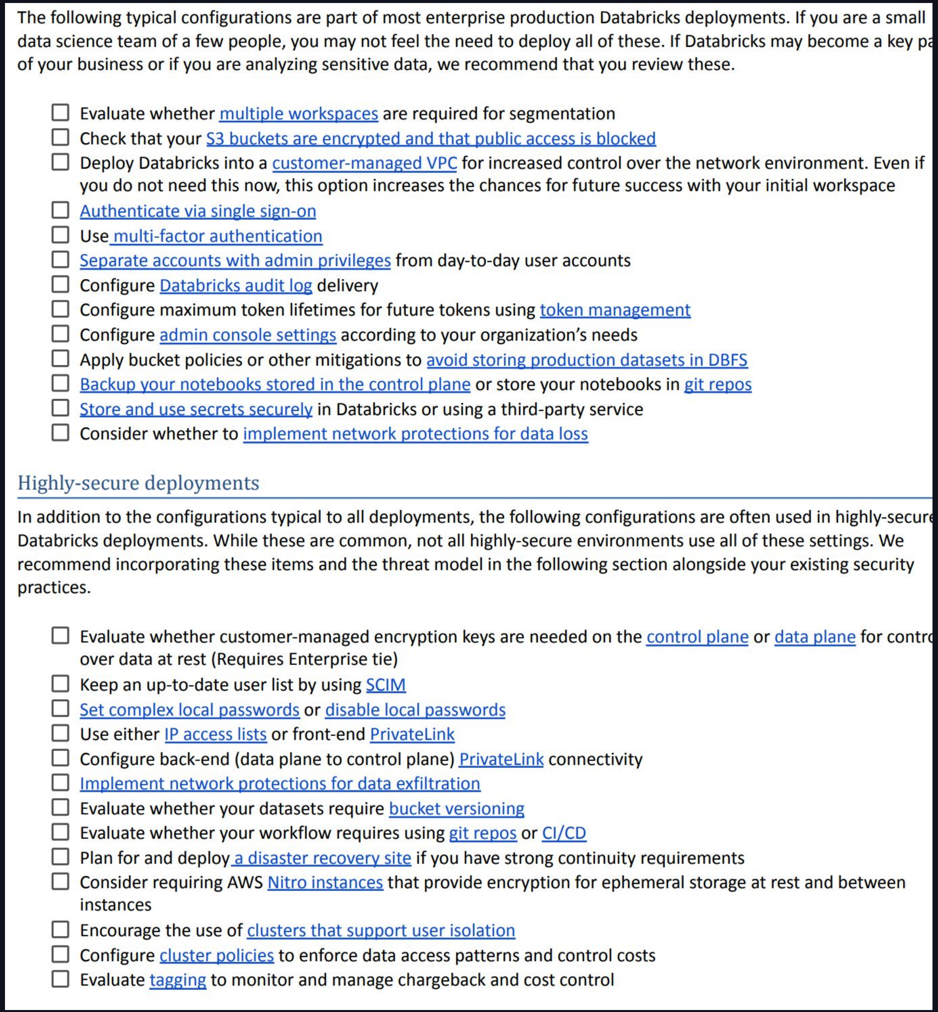
## Including a checklist! databricks.com/trust



Security Best Practices for Databricks on AWS

*Version 1.0 - June 16 2022*

databricks



The following typical configurations are part of most enterprise production Databricks deployments. If you are a small data science team of a few people, you may not feel the need to deploy all of these. If Databricks may become a key pa of your business or if you are analyzing sensitive data, we recommend that you review these.

- ☐ Evaluate whether multiple workspaces are required for segmentation
- ☐ Check that your S3 buckets are encrypted and that public access is blocked
- ☐ Deploy Databricks into a customer-managed VPC for increased control over the network environment. Even if you do not need this now, this option increases the chances for future success with your initial workspace
- ☐ Authenticate via single sign-on
- ☐ Use multi-factor authentication
- ☐ Separate accounts with admin privileges from day-to-day user accounts
- ☐ Configure Databricks audit log delivery
- ☐ Configure maximum token lifetimes for future tokens using token management
- ☐ Configure admin console settings according to your organization's needs
- ☐ Apply bucket policies or other mitigations to avoid storing production datasets in DBFS
- ☐ Backup your notebooks stored in the control plane or store your notebooks in git repos
- ☐ Store and use secrets securely in Databricks or using a third-party service
- ☐ Consider whether to implement network protections for data loss

### Highly-secure deployments

In addition to the configurations typical to all deployments, the following configurations are often used in highly-secure Databricks deployments. While these are common, not all highly-secure environments use all of these settings. We recommend incorporating these items and the threat model in the following section alongside your existing security practices.

- ☐ Evaluate whether customer-managed encryption keys are needed on the control plane or data plane for contro over data at rest (Requires Enterprise tie)
- ☐ Keep an up-to-date user list by using SCIM
- ☐ Set complex local passwords or disable local passwords
- ☐ Use either IP access lists or front-end PrivateLink
- ☐ Configure back-end (data plane to control plane) PrivateLink connectivity
- ☐ Implement network protections for data exfiltration
- ☐ Evaluate whether your datasets require bucket versioning
- ☐ Evaluate whether your workflow requires using git repos or CI/CD
- ☐ Plan for and deploy a disaster recovery site if you have strong continuity requirements
- ☐ Consider requiring AWS Nitro instances that provide encryption for ephemeral storage at rest and between instances
- ☐ Encourage the use of clusters that support user isolation
- ☐ Configure cluster policies to enforce data access patterns and control costs
- ☐ Evaluate tagging to monitor and manage chargeback and cost control

AWS only for now...

But your deployment is probably fine, right?

Want to check anything real quick?

# ANNOUNCING

## SWAT

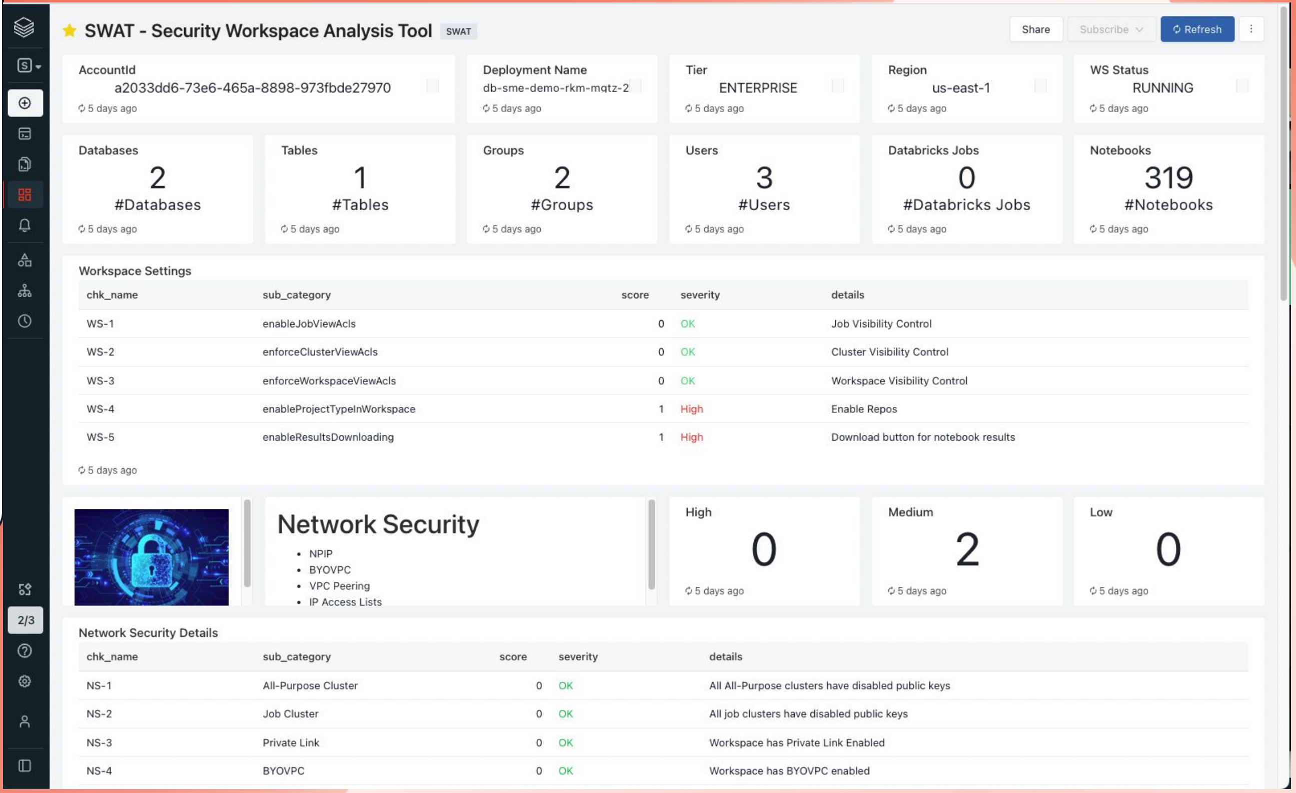
Security Workspace Analysis Tool!

# Security Workspace Analysis Tool

- Compare workspace configurations against specific best practices
- Flag deviations and score your workspace over a period of time
- Get mitigation references
- Want to try this in your workspace?

Contact us at: [cybersecurity@databricks.com](mailto:cybersecurity@databricks.com)

(AWS) Beta starting shortly...



SWAT - Security Workspace Analysis Tool SWAT

AccountId: a2033dd6-73e6-465a-8898-973fbde27970 (5 days ago)

Deployment Name: db-sme-demo-rkm-mqtz-2 (5 days ago)

Tier: ENTERPRISE (5 days ago)

Region: us-east-1 (5 days ago)

WS Status: RUNNING (5 days ago)

Databases: 2 #Databases (5 days ago)

Tables: 1 #Tables (5 days ago)

Groups: 2 #Groups (5 days ago)

Users: 3 #Users (5 days ago)

Databricks Jobs: 0 #Databricks Jobs (5 days ago)

Notebooks: 319 #Notebooks (5 days ago)

Workspace Settings

| chk_name | sub_category | score | severity | details |
|---|---|---|---|---|
| WS-1 | enableJobViewAcls | 0 | OK | Job Visibility Control |
| WS-2 | enforceClusterViewAcls | 0 | OK | Cluster Visibility Control |
| WS-3 | enforceWorkspaceViewAcls | 0 | OK | Workspace Visibility Control |
| WS-4 | enableProjectTypeInWorkspace | 1 | High | Enable Repos |
| WS-5 | enableResultsDownloading | 1 | High | Download button for notebook results |

5 days ago

Network Security

- NPIP
- BYOVPC
- VPC Peering
- IP Access Lists

High: 0 (5 days ago)

Medium: 2 (5 days ago)

Low: 0 (5 days ago)

Network Security Details

| chk_name | sub_category | score | severity | details |
|---|---|---|---|---|
| NS-1 | All-Purpose Cluster | 0 | OK | All All-Purpose clusters have disabled public keys |
| NS-2 | Job Cluster | 0 | OK | All job clusters have disabled public keys |
| NS-3 | Private Link | 0 | OK | Workspace has Private Link Enabled |
| NS-4 | BYOVPC | 0 | OK | Workspace has BYOVPC enabled |

Anindita Mahapatra

Ramdas Murali

Arun Pamulapati

# Recap

# tl;dr

## Security Architecture and Controls

The security & trust center outlines our architecture and enumerates the key security detections.

Overview: databricks.com/trust/whitepaper
Detail: databricks.com/trust

## Prescriptive Best Practices Guides

Databricks has captured our best practices into a doc with “common” and “high-security” models and checkboxes. *Check those boxes.*

databricks.com/trust/Security-Best-Practices_Databricks-on-AWS.pdf

## Meet Compliance Needs

Have your security team download our due diligence package (ISO certs, Pen Test) and reach out to your account team for SOC 2, Enterprise Security Guide

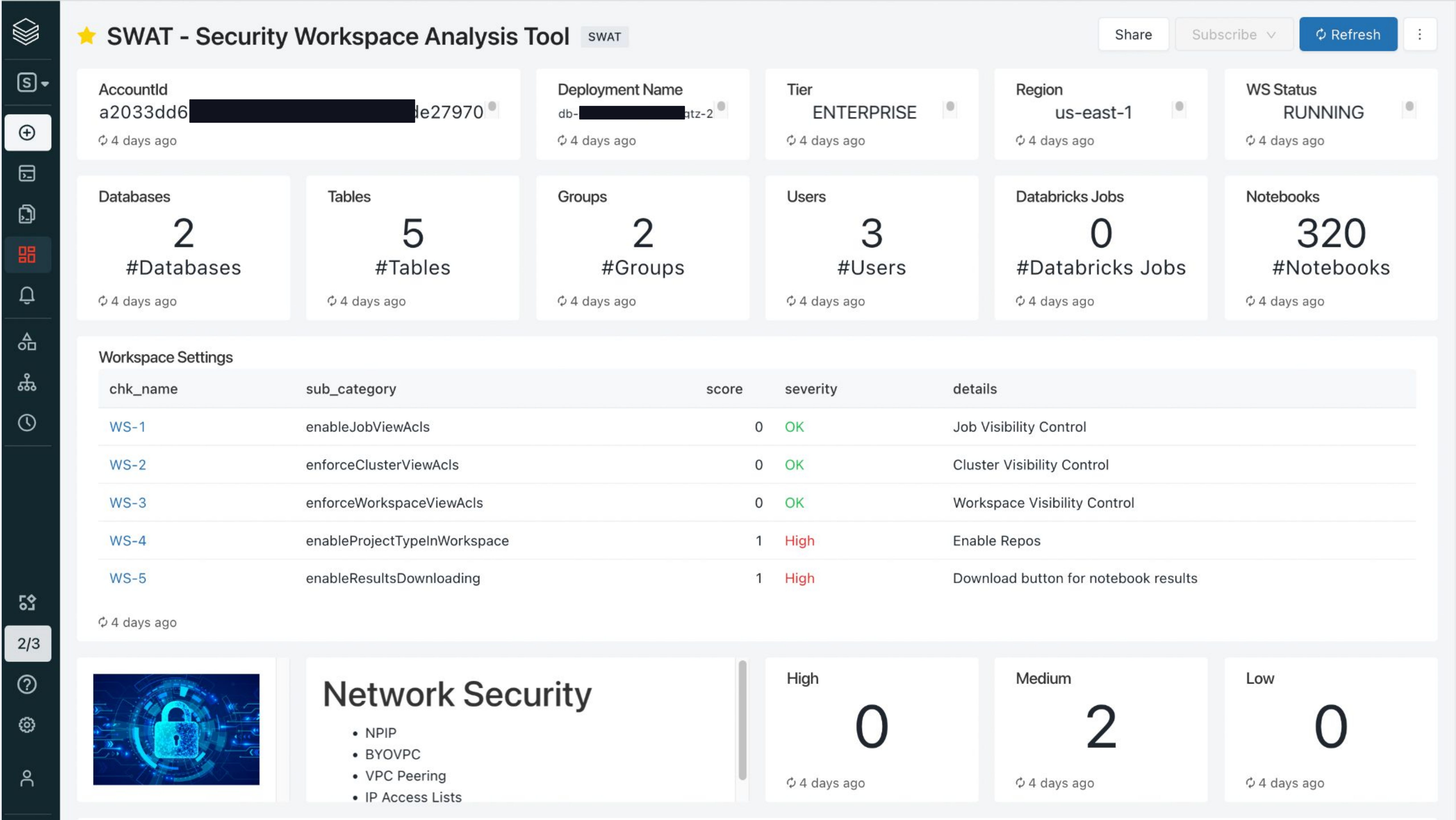
Compliance: databricks.com/trust#compliance

## Analyze your deployment

The Workspace Analysis tool now contains a security section! Check your workspace against our most common best practices.

Coming soon to the Databricks Blog!

# Thank you!



# SWAT - Security Workspace Analysis Tool

SWAT

AccountId
a2033dd6[redacted]de27970
4 days ago

Deployment Name
db-[redacted]qtz-2
4 days ago

Tier
ENTERPRISE
4 days ago

Region
us-east-1
4 days ago

WS Status
RUNNING
4 days ago

Databases
2
#Databases
4 days ago

Tables
5
#Tables
4 days ago

Groups
2
#Groups
4 days ago

Users
3
#Users
4 days ago

Databricks Jobs
0
#Databricks Jobs
4 days ago

Notebooks
320
#Notebooks
4 days ago

Workspace Settings

| chk_name | sub_category | score | severity | details |
|---|---|---|---|---|
| WS-1 | enableJobViewAcls | 0 | OK | Job Visibility Control |
| WS-2 | enforceClusterViewAcls | 0 | OK | Cluster Visibility Control |
| WS-3 | enforceWorkspaceViewAcls | 0 | OK | Workspace Visibility Control |
| WS-4 | enableProjectTypeInWorkspace | 1 | High | Enable Repos |
| WS-5 | enableResultsDownloading | 1 | High | Download button for notebook results |

4 days ago

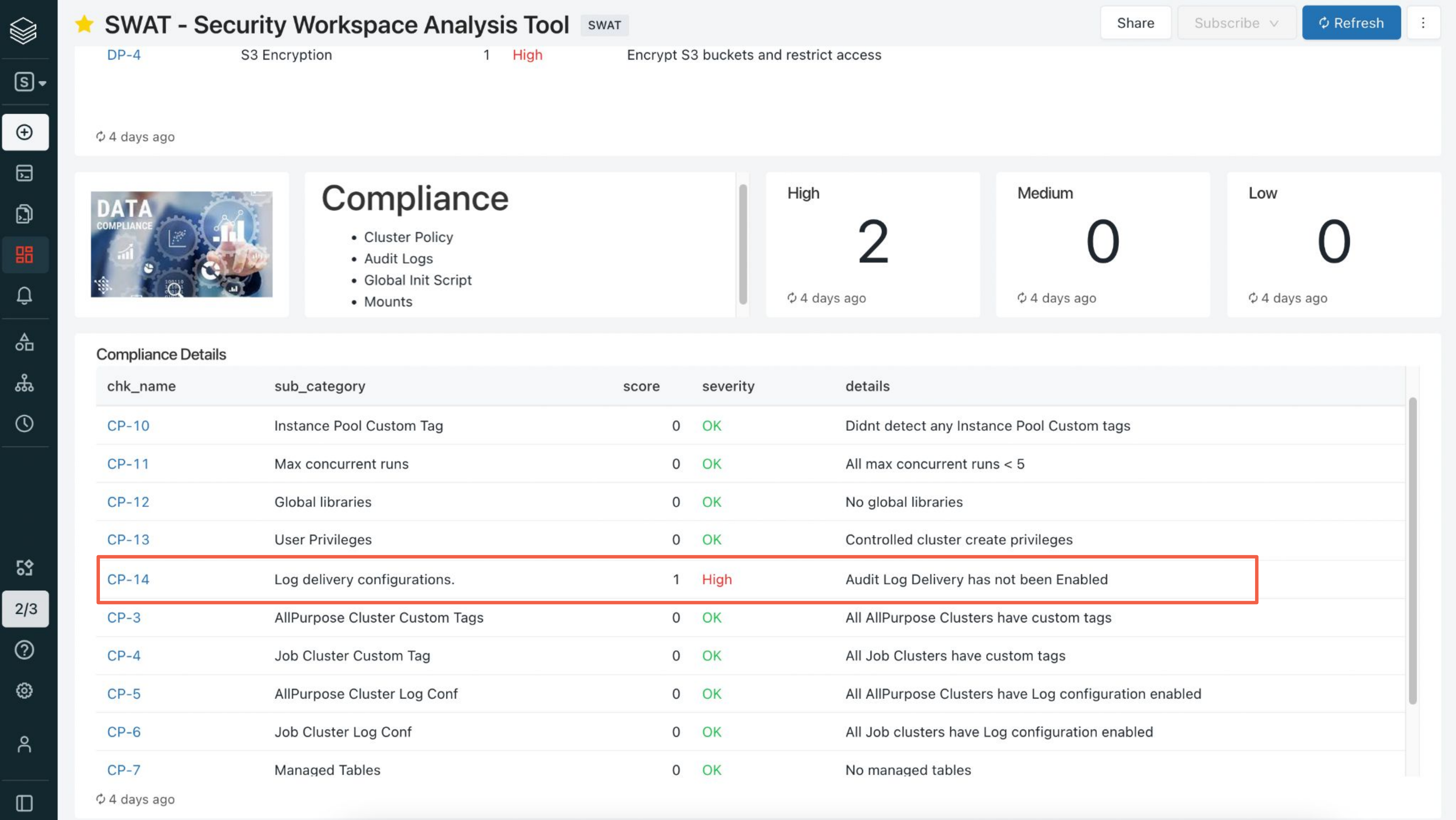
## Network Security

- NPIP
- BYOVPC
- VPC Peering
- IP Access Lists

High
0
4 days ago

Medium
2
4 days ago

Low
0
4 days ago



# SWAT - Security Workspace Analysis Tool

SWAT

| | | | | |
|---|---|---|---|---|
| DP-4 | S3 Encryption | 1 | High | Encrypt S3 buckets and restrict access |

4 days ago

## Compliance

- Cluster Policy
- Audit Logs
- Global Init Script
- Mounts

High

2

4 days ago

Medium

0

4 days ago

Low

0

4 days ago

### Compliance Details

| chk_name | sub_category | score | severity | details |
|---|---|---|---|---|
| CP-10 | Instance Pool Custom Tag | 0 | OK | Didnt detect any Instance Pool Custom tags |
| CP-11 | Max concurrent runs | 0 | OK | All max concurrent runs < 5 |
| CP-12 | Global libraries | 0 | OK | No global libraries |
| CP-13 | User Privileges | 0 | OK | Controlled cluster create privileges |
| CP-14 | Log delivery configurations. | 1 | High | Audit Log Delivery has not been Enabled |
| CP-3 | AllPurpose Cluster Custom Tags | 0 | OK | All AllPurpose Clusters have custom tags |
| CP-4 | Job Cluster Custom Tag | 0 | OK | All Job Clusters have custom tags |
| CP-5 | AllPurpose Cluster Log Conf | 0 | OK | All AllPurpose Clusters have Log configuration enabled |
| CP-6 | Job Cluster Log Conf | 0 | OK | All Job clusters have Log configuration enabled |
| CP-7 | Managed Tables | 0 | OK | No managed tables |

4 days ago

# **Announcing:** Security Workspace Analysis Tool!

# Use AWS Nitro instances

AWS Nitro instances provide two major security benefits:

- NVMe disks automatically encrypt at rest (local SSDs for shuffle data)
  - Azure can enable software encryption
  - GCP automatically has encryption
- Many Nitro instances automatically encrypt data in transit between hosts
  - Azure can enable intra-cluster encryption in software
  - GCP automatically has inter-host encryption

Account Takeover

Data Exfiltration

AIE

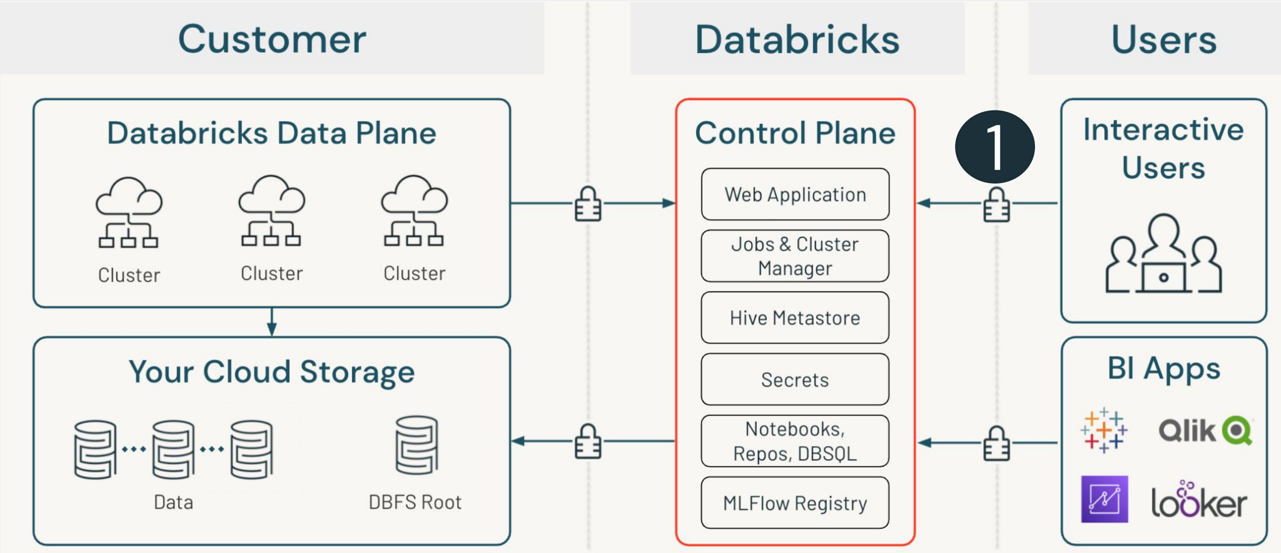
RA

CD

# Control user network access

Two options:

- IP Access Lists Enterprise
- User to Control Plane Private Link (Preview) Enterprise
  - Not available on GCP



Account Takeover

Data Exfiltration

AIE

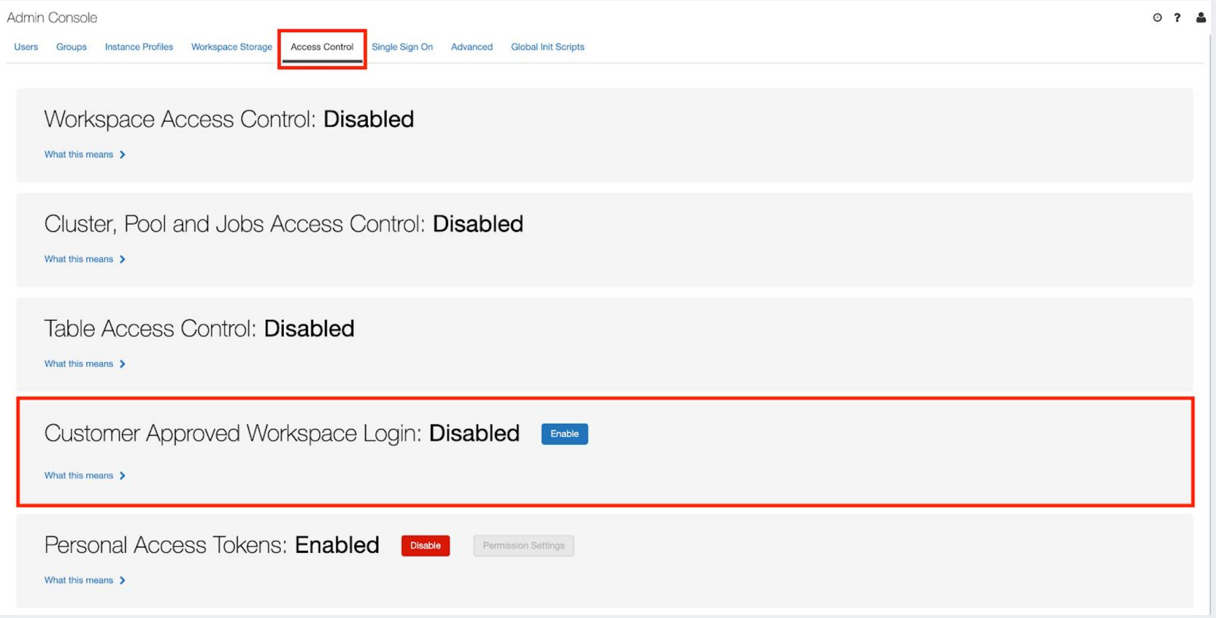
RA

CD

# Configure the admin console settings

Manage key configuration settings

- Can Databricks staff access your workspace for support?
- Can notebooks be exported?
- Can results be downloaded?
- Where are query results stored?
- Are ACLs enabled?



Workspace Access Control: **Disabled**

What this means >

Cluster, Pool and Jobs Access Control: **Disabled**

What this means >

Table Access Control: **Disabled**

What this means >

Customer Approved Workspace Login: **Disabled**

What this means >

Personal Access Tokens: **Enabled**

What this means >

DATA+AI SUMMIT 2022

Account Takeover

DE

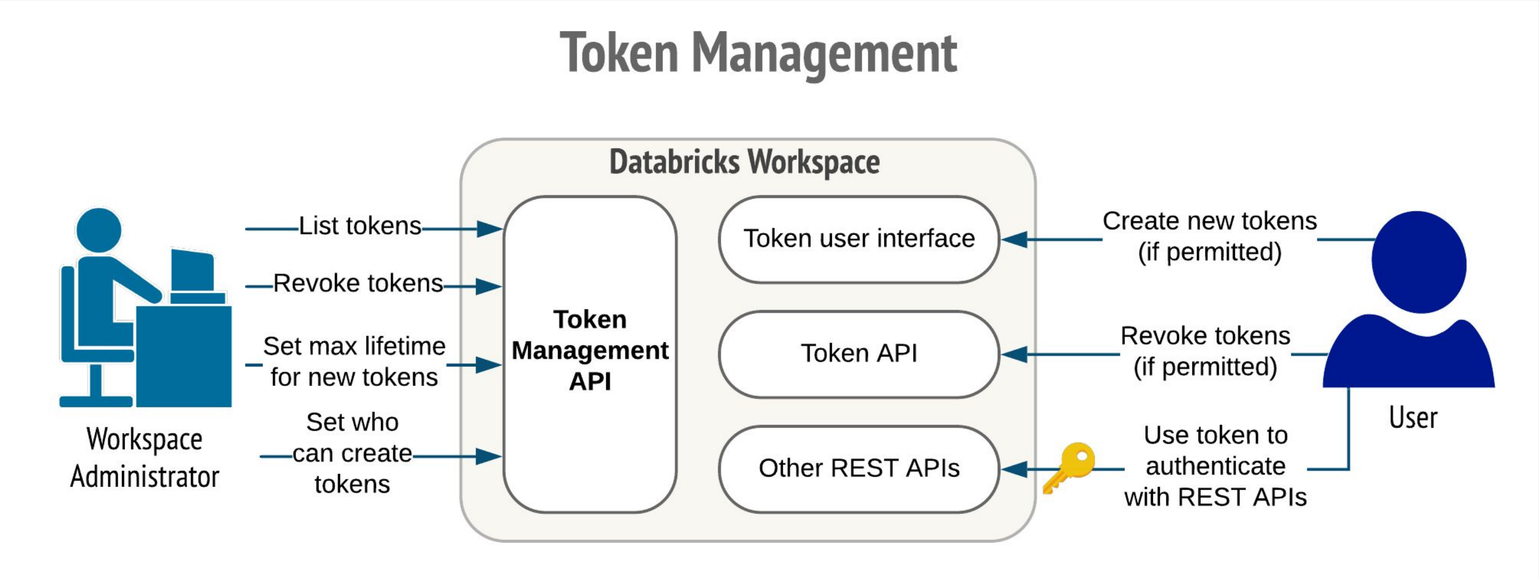
AIE

RA

CD

# Token management

- Enable or disable personal access tokens (PATs) for some or all users
- Configure a max token lifetime for new tokens


Token Management
Databricks Workspace
List tokens
Revoke tokens
Set max lifetime for new tokens
Set who can create tokens
Workspace Administrator
Token Management API
Token user interface
Token API
Other REST APIs
Create new tokens (if permitted)
Revoke tokens (if permitted)
Use token to authenticate with REST APIs
User

## We want to utilize private networking

AT

Data Exfiltration

AIE

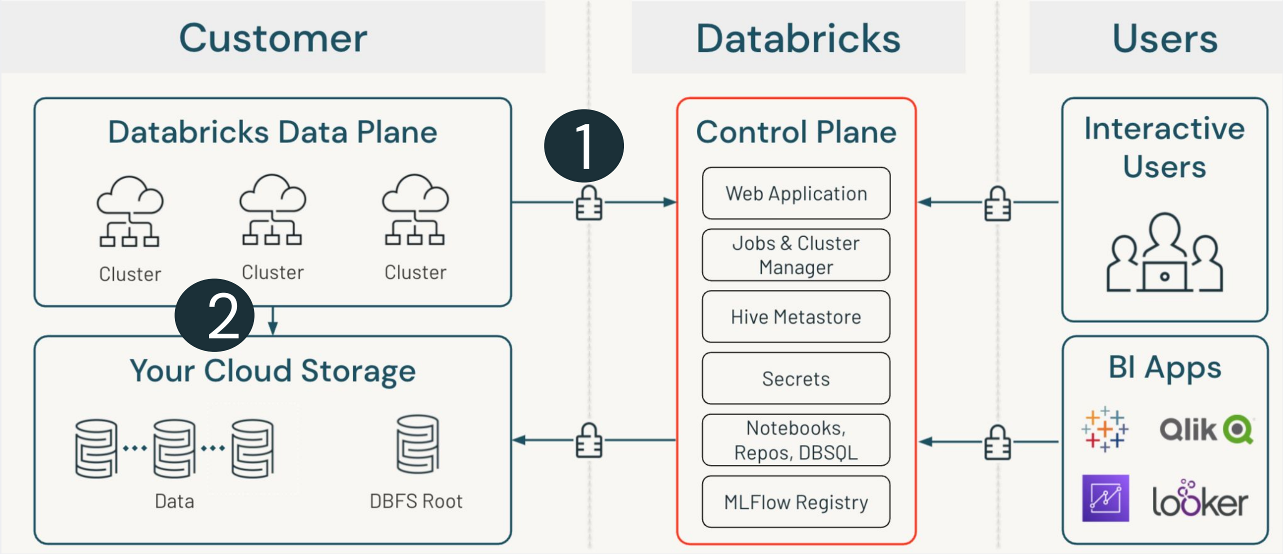
RA

CD

# Configure Back-End Private Link

Multiple options:

1. Data Plane to Control Plane
   - By default, the cloud service provider backbone is very secure
   - (AWS / Azure) Use Data Plane to Control Plane PrivateLink (Preview) Enterprise
2. Data Plane to Storage Services
   - Public Endpoints
   - Private Endpoints
   - Private Link to your data sources



Customer

Databricks

Users

Databricks Data Plane

Cluster

Cluster

Cluster

1

2

Your Cloud Storage

Data

DBFS Root

Control Plane

Web Application

Jobs & Cluster Manager

Hive Metastore

Secrets

Notebooks, Repos, DBSQL

MLFlow Registry

Interactive Users

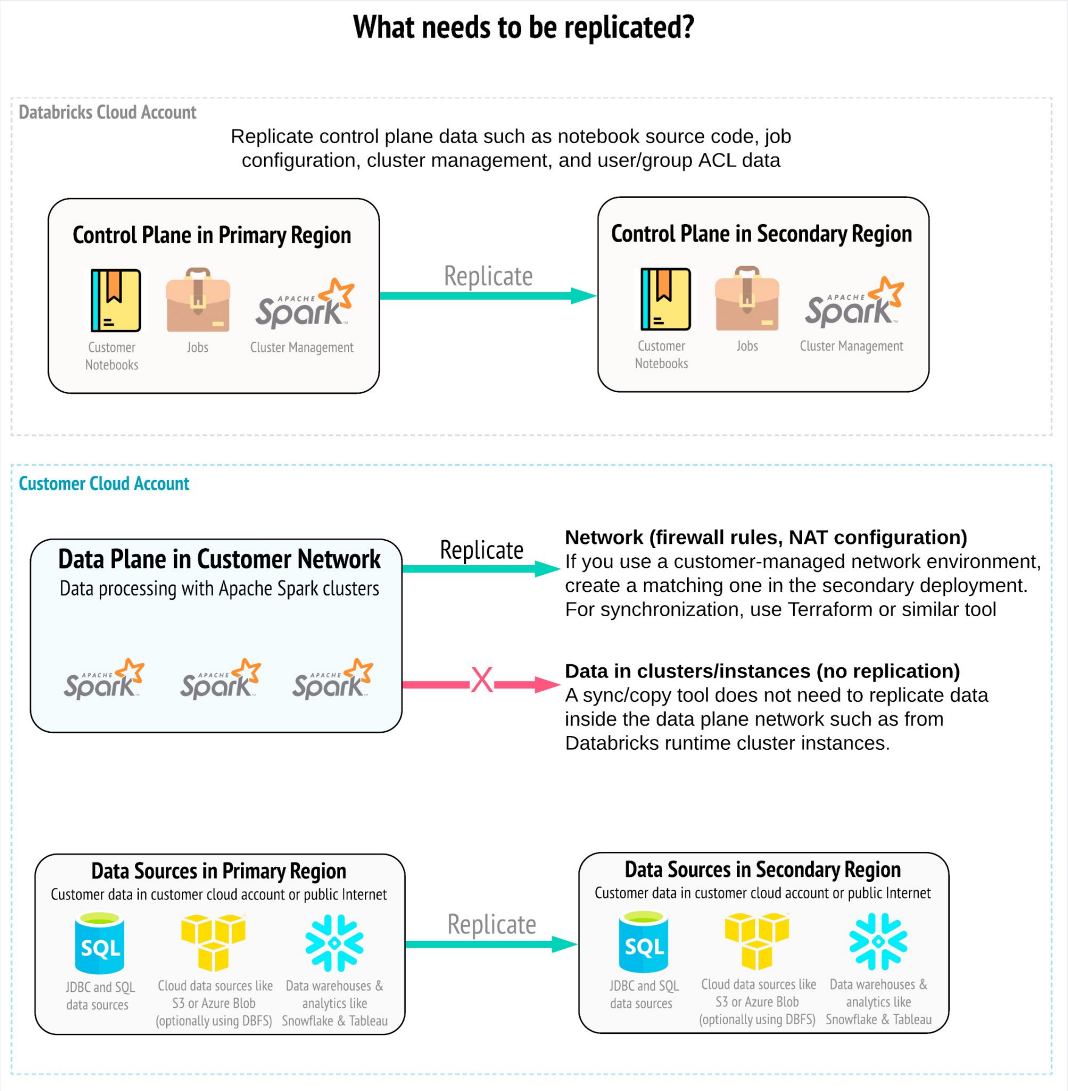
BI Apps

Qlik

looker

# Configure a DR site

- Understand your business needs
- Choose a process that meets your business needs
- Prep workspaces and do a one-time copy
- Prepare your data sources
- Implement and test your solution



**What needs to be replicated?**

Replicate control plane data such as notebook source code, job configuration, cluster management, and user/group ACL data

**Network (firewall rules, NAT configuration)**
If you use a customer-managed network environment, create a matching one in the secondary deployment. For synchronization, use Terraform or similar tool

**Data in clusters/instances (no replication)**
A sync/copy tool does not need to replicate data inside the data plane network such as from Databricks runtime cluster instances.

Databricks Cloud Account
Control Plane in Primary Region
Customer Notebooks
Jobs
Cluster Management
Replicate
Control Plane in Secondary Region
Customer Notebooks
Jobs
Cluster Management
Customer Cloud Account
Data Plane in Customer Network
Data processing with Apache Spark clusters
Replicate
X
Data Sources in Primary Region
Customer data in customer cloud account or public Internet
SQL
JDBC and SQL data sources
Cloud data sources like S3 or Azure Blob (optionally using DBFS)
Data warehouses & analytics like Snowflake & Tableau
Replicate
Data Sources in Secondary Region
Customer data in customer cloud account or public Internet
SQL
JDBC and SQL data sources
Cloud data sources like S3 or Azure Blob (optionally using DBFS)
Data warehouses & analytics like Snowflake & Tableau

# We want to control configurations and costs

AT

DE

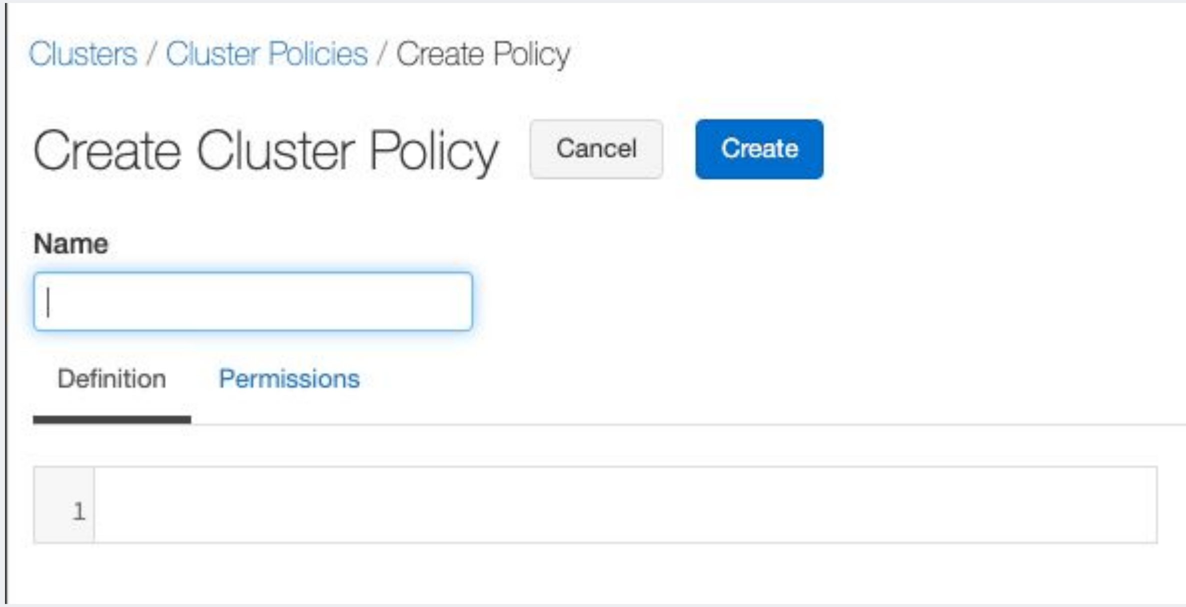
AIE

Resource Abuse

CD

# Configure cluster policies

- Limit users to create clusters with prescribed settings.
- Simplify the user interface for your users
- Control cost by limiting per cluster maximum cost



Clusters / Cluster Policies / Create Policy

Create Cluster Policy

AT

DE

AIE

Resource Abuse

CD

# Configure tagging to monitor

- Monitor cost and accurately attribute Databricks usage to your organization’s business units and teams (for chargebacks, for example), you can tag clusters and pools.
- These tags propagate both to detailed DBU usage reports and to your cloud service provider (e.g., AWS EC2 and AWS EBS)

DATA+AI SUMMIT 2022

David Veuve

Head of Security Field Engineering, Databricks

Arun Pamulapati

Sr. Staff Security Field Engineer, Databricks

# Lakehouse Security

Security Best Practices and Threat Model for the Lakehouse

June 29th 10:35 AM

Moscone South

# Resource Abuse

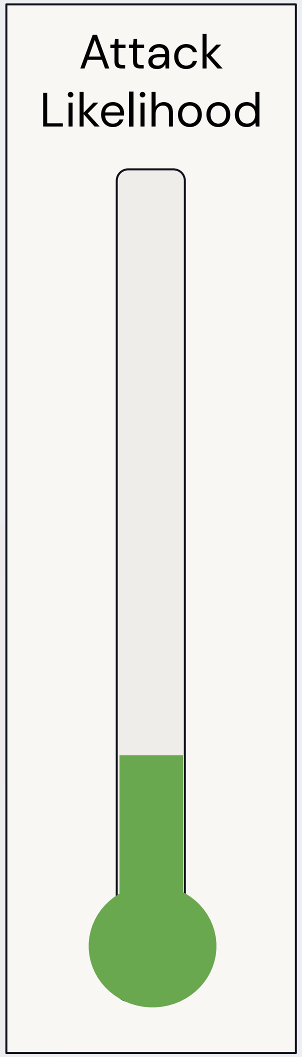
Attackers gain access to customer compute resources

## Risk Overview

- Customer cloud infrastructure hijacking for crypto mining
- Accidental/abusive waste of customer resources

## Best Practices to Mitigate

- Cloud native protections:
  - Restricted x IAM role
  - Service quotas
  - Cloud monitoring
  - CloudTrail
- Databricks protections:
  - Cluster policies
  - Cluster ACLs
  - Library control
  - Databricks Audit logs



# Compromise of Databricks Inc

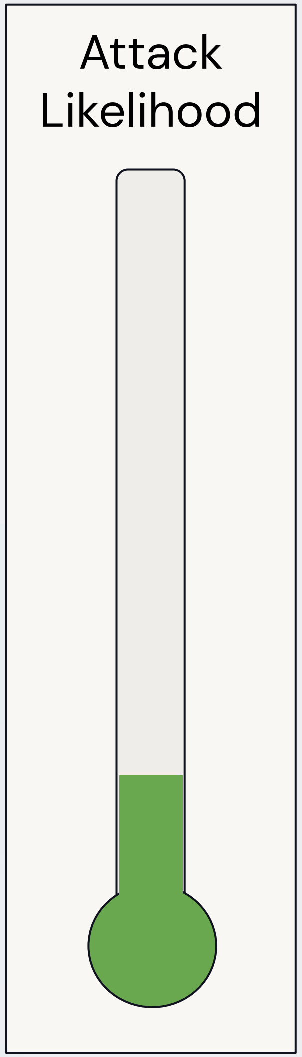
## Attacker gains customer environment via Databricks

### Risk Overview

- Compromise of Databricks Inc user or system could result in compromise of customer environment

### Best Practices to Mitigate

- Monitor Databricks audit logs
- Consider CAWL
- Restrict cross account IAM role
- Monitor CloudTrail
- Break glass controls:
  - Customer managed key for managed services
  - Ability to disable Databricks cross account IAM role



Account Takeover

DE

AIE

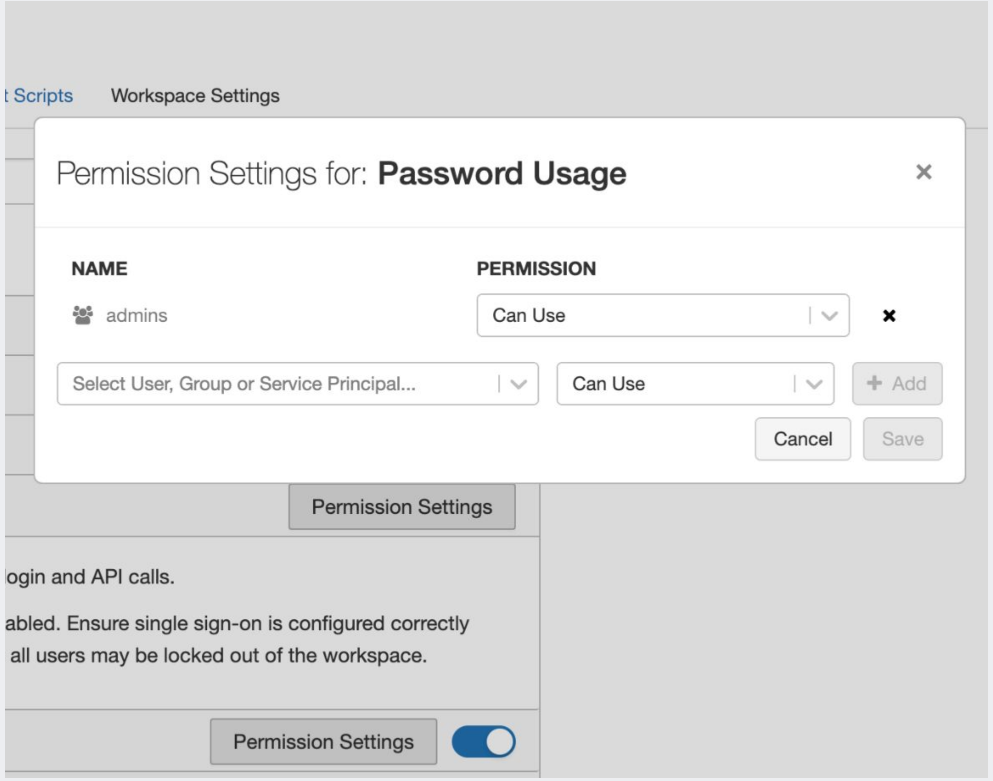
RA

CD

# Secure local passwords

(AWS)

- When MFA isn't available, use very long and complex passwords, and securely store them
- Disable local password via password access control



Permission Settings for: **Password Usage**

| NAME | PERMISSION |
|---|---|
| admins | Can Use |
| Select User, Group or Service Principal... | Can Use |

ogin and API calls.

abled. Ensure single sign-on is configured correctly all users may be locked out of the workspace.