Whitepaper | Summer 2022

Data Security Governance for Databricks with Privacera

PRIVACERA + DATABRICKS OVERVIEW

privacera

Contents

Introduction 03

Privacera Security Governance Platform Overview 04

Scan and Tag Sensitive Data 06

Defining and Enforcing Data Access Control Policies 07

- Fine-Grained Access Control 07

Monitoring and Reporting 08

Security for Databricks Clusters 09

Case Studies 10

- Global Consumer Product Company 10
- Telecommunications Company 10

Conclusion 11

Introduction

It's no secret that deriving business value from data with analytics is a top CIO and CDO priority. But another top concern in organizations, and rightly so, is data governance and security.

The reality is that enterprise IT and data teams face a challenging dual mandate to meet this lofty goal. These teams are under increasing pressure to make data widely available tothe business users to support various digital transformation initiatives. Data scientists and business analysts need seamless access to data via the tools and applications of their choice. As enterprises embark on the journey to be data-driven, they require seamless and broad access to data using powerful analytics platforms such as Databricks. The IT infrastructure teams are forced to seek a balance between the mandate to make more data widely available with the competing directive to ensure that the enterprise's use of data is in compliance with all applicable external privacy regulations and industry standards, as well as internal data usage best practices. Creating and enforcing data access control policies to ensure that only authorized users have access to the data, as well as implementing mechanisms to enable monitoring and auditing of access patterns is foundational to meeting this dual mandate.

These two mandates are in tension with one another and must be balanced. The companies that get this balance wrong by overly restricting access to sensitive data for data scientists and analysts risk missing out on valuable insights that could lead to competitive advantage. On the other end of the spectrum, enterprises that do not apply the proper data access governance controls risk unauthorized users accessing sensitive data, which can result in a lack of regulatory compliance and financial and/or reputational harm to the enterprise and in some extreme cases its ability to operate.

Privacera's mission is to help organizations strike this delicate balance for their enterprise data and to that end provides a platform that strives to deliver this capability to enterprises. Privacera's centralized data access governance platform is based on the open-source project Apache Ranger. Privacera has extended Ranger's capabilities beyond traditional Big Data environments to cloud-native services and leading analytics platforms such as AWS, Azure, GCP, and Databricks to allow enterprise access and data security management across the entire data and analytics ecosystem. Together, Privacera and Databricks enable enterprises to safely and securely make data accessible for processing, advanced machine learning, and artificial intelligence by managing the complete data access governance lifecycle.

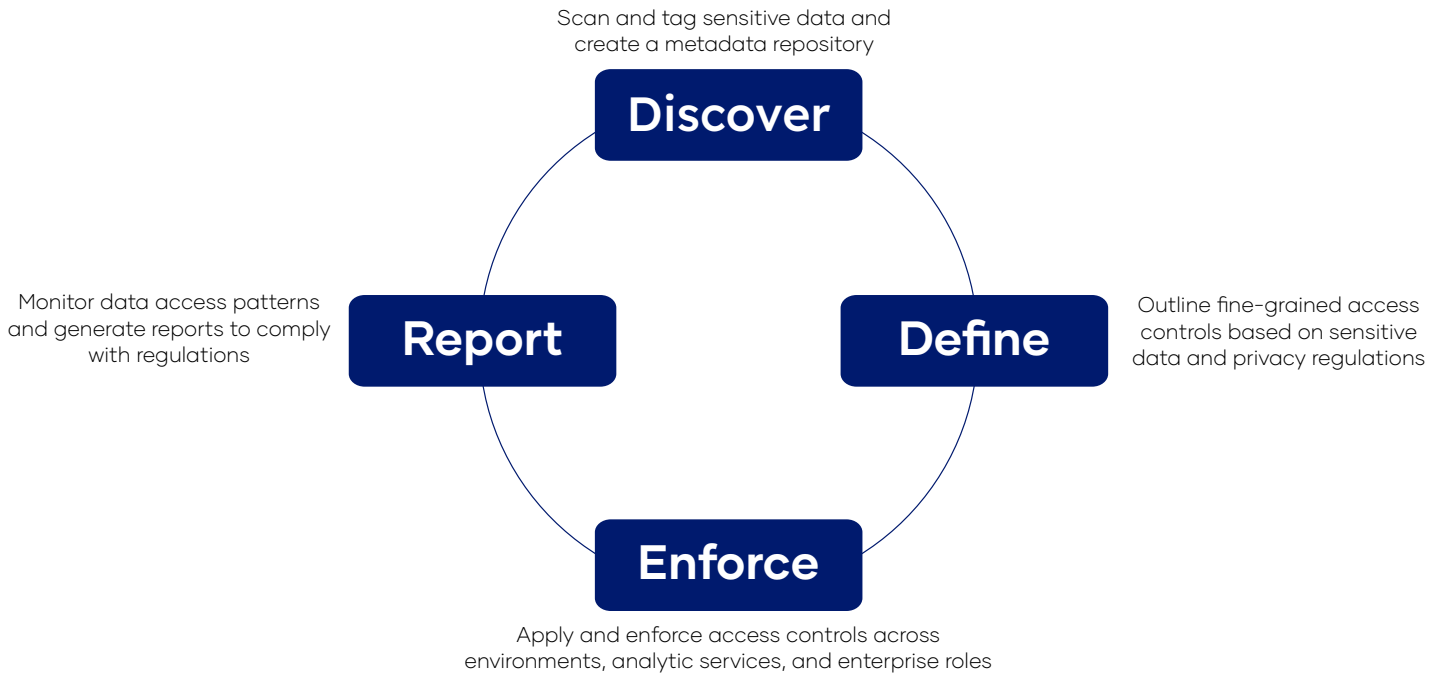

Figure 1: Data Access Governance Lifecycle

By managing the entire data access governance lifecycle with Privacera, enterprise IT and data teams can confidently make more data available via Databricks for data science and machine learning use cases with the assurance that data is only being accessed by authorized users in compliance with applicable privacy regulations and policies.

Privacera Security Governance Platform Overview

Privacera users, such as data architects and data platform administrators, interact with the platform via the Privacera portal. Users can view sensitive data tags, create and define access control policies, automatically enforce those policies, and monitor data access behavior all from a centralized interface.

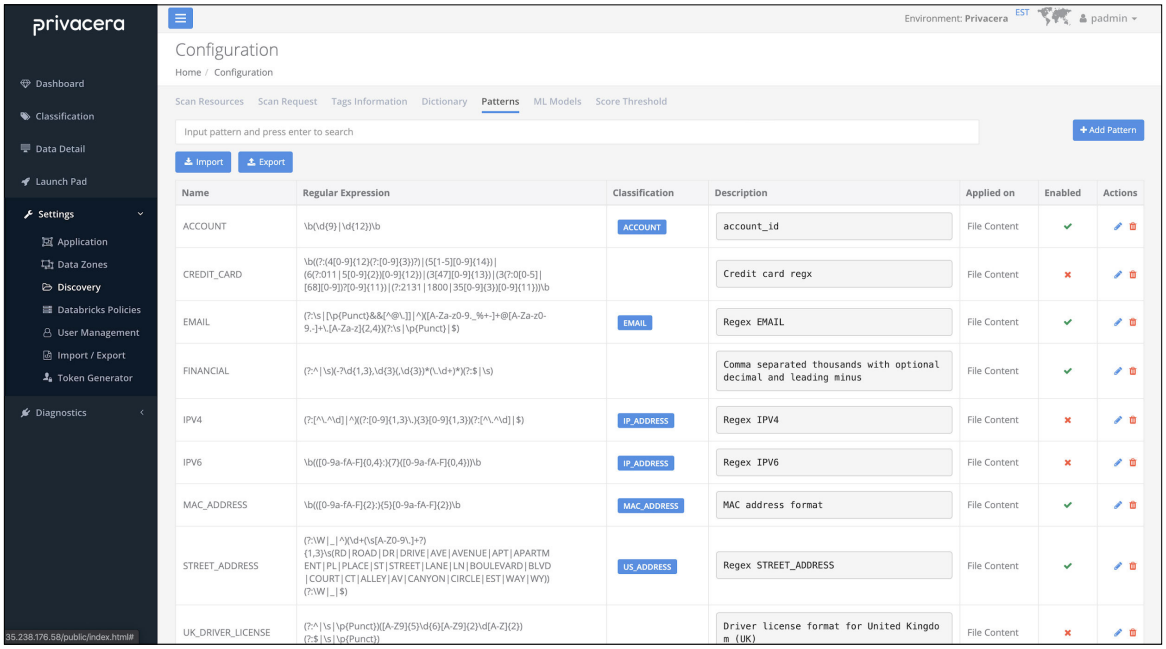

Figure 2: Privacera Portal

Privacera supports fine-grained data access control for Databricks Spark clusters running in high concurrency mode with Python and SQL languages. The Privacera Ranger plugin for Spark runs within the Databricks cluster and provides access control for all user requests. The architecture of the plugin is similar to the Ranger plugins for Apache Hive, Apache HBase, HDFS, and Apache Kafka, and includes a library that is loaded at startup time that runs within the Spark Driver. This ensures that all data is accessed via the Ranger plugin for authorization.

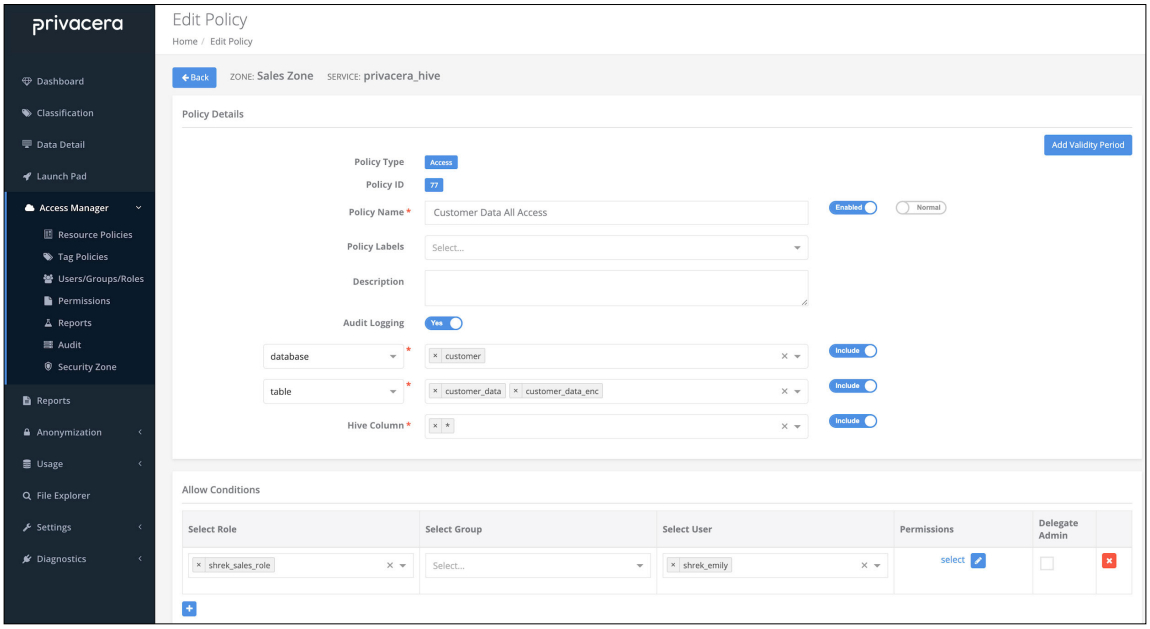

Figure 3: Privacera supports fine-grain data access control for Databricks clusters

Scan and Tag Sensitive Data

Databricks is the leading analytics platform that provides shared usage infrastructure for data science, data engineering and analyst teams plus is compatible with multiple storage environments. Privacera automatically connects to cloud and on-premises storage services and databases that serve as the storage layer for Databricks deployments. This includes Amazon S3, Azure Cloud Storage, and Google Cloud Storage.

Once connected to the storage environment (including Delta tables), Privacera performs an initial scan of data stored at rest and then continuously scans new data in near real-time as it enters the environment. As it scans the data, Privacera uses one of three methods to identify and tag sensitive data - pattern matching, machine learning models, and dictionary or lookup tables - depending on the use case and data type. Various rules and algorithms can be composed to further refine the detection.

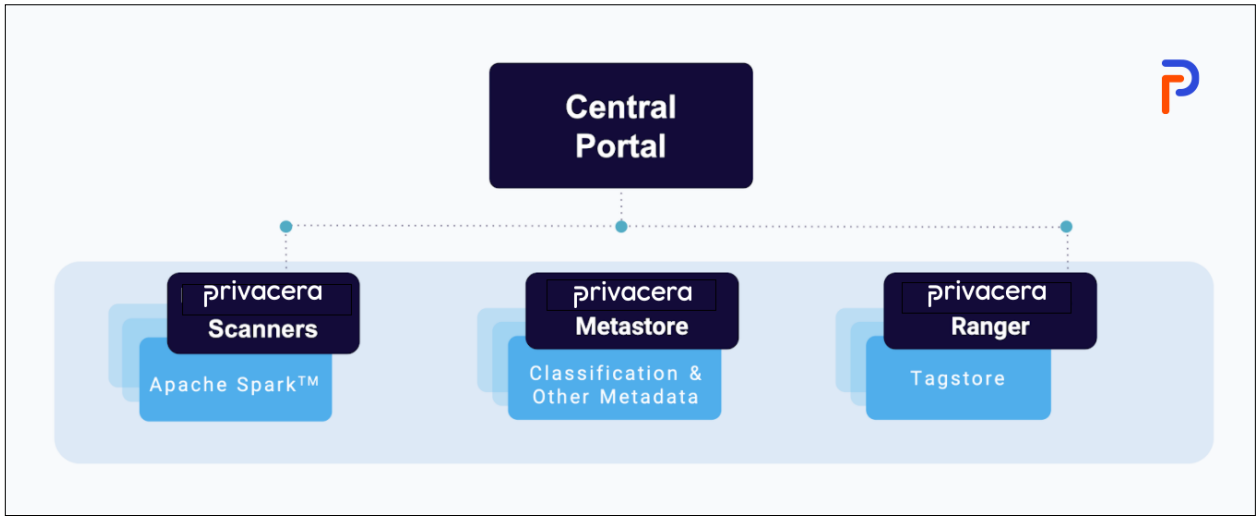

Figure 4: Privacera Data Discovery Architecture

Once identified, the tags are automatically applied to the data and stored in a metadata store, or sensitive data catalog, which can be reviewed and reported against to prove compliance. Privacera also provides the flexibility for administrators to review and approve the workflows as part of the data classification curation.

Defining and Enforcing Data Access Control Policies

Fine-Grained Access Control

Privacera is based on Apache Ranger, an open source project for creating, defining, and enforcing data access control policies and providing a comprehensive non-repudiable trail of audit events. Privacera has extended Ranger to work seamlessly with cloud databases and analytics services as well as relational databases. This includes Databricks running on AWS, Microsoft Azure, and Google Cloud.

Administrators create and define fine-grained data access control policies in the Privacera portal. Refer to the figure below. The following data access control functionality is supported in Databricks clusters enabled with Python and/or SQL languages:

Python/ SQL Cluster
Fine Grained Access Control • Row, Column • Row Level Filtering • Column Masking
DBFS/ S3/ ADLS file level access control
Attribute-based (ABAC), Tag-based, and role-based access control (RBAC)
Uses Ranger Plugin along with High Concurrency and Process Isolation

Monitoring and Reporting

Once data access control policies are defined, created and enforced, it is important for IT and data platform teams to have visibility into access behaviors, including when access is denied to unauthorized users. To facilitate this, the Privacera Ranger plugin keeps an audit of every access request that is made and the result of each request. The audit events have a normalized schema with rich event metadata. The event metadata includes information about who tried to access what data, when, and from which environment along with contextual information such as its classification and which tenant, security zone or cluster the data is accessed. This information is temporarily cached locally and uploaded regularly to an indexing service on the Ranger server. These audits are available in near real-time to query in the Privacera portal and are searchable by a number of attributes for forensic analysis by internal and external compliance auditors.

The audit logs can also be forwarded to or transformed for consumption by downstream systems such as enterprise messaging platforms (Kafka, AWS Kinesis, Azure EventHubs etc.), SIEM and CyberSecurity systems (Splunk) or written in optimized formats such as ORC for direct querying.

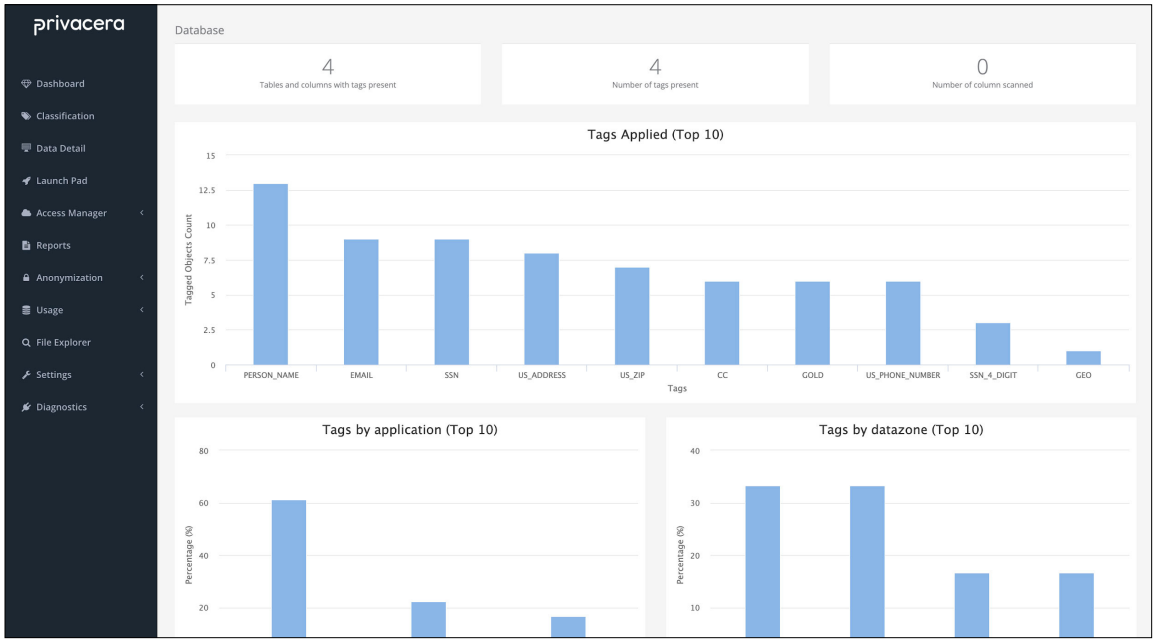

Figure 5: Privacera provides data teams with instant visibility to data assets for regulatory compliance requirements.

Security for Databricks Clusters

Privacera leverages Databricks' built-in advanced security settings to harden the cluster for python and SQL workloads. This includes the following:

- **Network isolation:** This prevents normal users from accessing IAM roles within the cluster
- **Process isolation:** In a hardened environment, notebooks are run as a different Linux user for everyend user. This restricts what folders and files the user can access, to prevent cross- contamination between notebooks
- **Interpreter restriction:** Certain interpreters are restricted and can't be used in the Databricks cluster
- **Python command white listing:** This ensures that only "safe" Python commands can be run in the cluster
- **Shell restriction:** Access to local file interpreters nless accessed via Spark read commands. The Ranger plugin authorizes all Spark read commands

In addition, Privacera has added the following features to help simplify security in a Databricks environment:

- Manage Databricks cluster policies using Privacera
- Map Databricks username in email address format to AD/LDAP username format
- Retrieve groups for users from AD/LDAP via Ranger admin during policy evaluation. This removes the requirement to synchronize Databricks with all the groups from AD/LDAP just for authorization purposes
- S3 and ADLS policies from Ranger are automatically applied in Databricks clusters when files are read/written using Spark load commands
- SAML authentication is enabled to Databrick from the Privacera portal. Privacera can be configured to authenticate using AD/LDAP, OAuth, SAML, Okta, etc.

Case Studies

Global Consumer Product Company

Challenge

A large, multi-national athletic footwear and apparel company has one of the largest deployments of Databricks on AWS. The company's analytics platform team was mandated by its internal privacy and governance organization to ensure access to sensitive data is managed and governed at all times. This required fine-grained row and column-level access control in Databricks and other services.

Solution

After experimenting with Apache Ranger, the company's analytics platform team turned to Privacera to provide centralized data access governance for its Databricks environment. The company deployed Privacera on AWS, taking advantage of AWS's native Spark capabilities and database services to store policies and metadata.

Benefits

Today, administrators on the analytics platform team use the Privacera portal to manage row, column and file-level access control policies across AWS services, including Databricks. The analytics platform team also uses the portal to monitor which users are accessing what data and to generate reports for fuller visibility.

Thanks to Privacera, the company's analytics platform team is able to onboard more users in a much shorter timeframe and at the same time reduce the number of access policies. This provides the data team with the assurance that the right data access control policies are in place for them to enable new use cases.

Telecommunications Company

Challenge

A leading media and telecommunications company was building the next-generation data platform across AWS and its on-premises data center for collecting and analyzing data from various sources, including cable set top boxes. The company also needed to comply with the newly enacted California Consumer Privacy Act and other compliance regulations.

Solution

The company's data platform team turned to Privacera to provide centralized data access governance on its new data platform, of which Databricks is a central component. The company deployed Privacera on AWS, taking advantage of AWS's native Spark capabilities and database services to store policies and metadata, to manage access governance policies across on-premises and cloud data analytics systems.

privacera

Benefits

Today, the data platform team uses Privacera with Databricks to continuously scan data stored in Amazon S3 to detect personally identifiable information (PII) and build a sensitive data catalog. It creates and enforces row, column and file-level access control policies for its Databricks deployment, and addresses CCPA requirements to anonymize data on request and report on results to show compliance. With Privacera and Databricks, the data team now has centralized data access governance across all its data workloads. This helped the company build internal trust in the security and governance of its data platform and scale to meet the growing needs of the business.

Conclusion

Databricks and Privacera provide faster time to value and empower enterprises to maximize the value of data by ensuring consistent governance, security, and compliance across all data science, machine learning, and artificial intelligence workloads. Privacera offers:

- Privacera offers rich data discovery & compliance workflows that natively leverages Spark from Databricks
- Centralized management with distributed enforcement via plugin based model
- Extends Databricks security model to provide fine grained access control with high availability and a unified View of all access patterns across Delta Lake.
- Enables non-intrusive, airtight, yet transparent security without impacting application changes or user behavior
- Centralize audits, policy analytics, and entitlements management
- Strong engineering partnership that builds and certifies a tightly integrated solution
- GDPR, CCPA & other compliance workflows based on using Databricks for ETL operations and leveraging DeltaLake to support update/deletes