Whitepaper | Summer 2020

Data Governance for Databricks with Privacera, Powered by Apache Ranger

PRIVACERA + DATABRICKS OVERVIEW

CONTENTS

Introduction

It’s no secret that deriving business value from data with analytic techniques like machine learning and artificial intelligence is a top CIO priority. Artificial intelligence, for example, was identified by CIOs as the top disruptive technology to existing business models in Gartner’s 2019 global CIO survey. But another top concern of CIOs, 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 to the 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. 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.

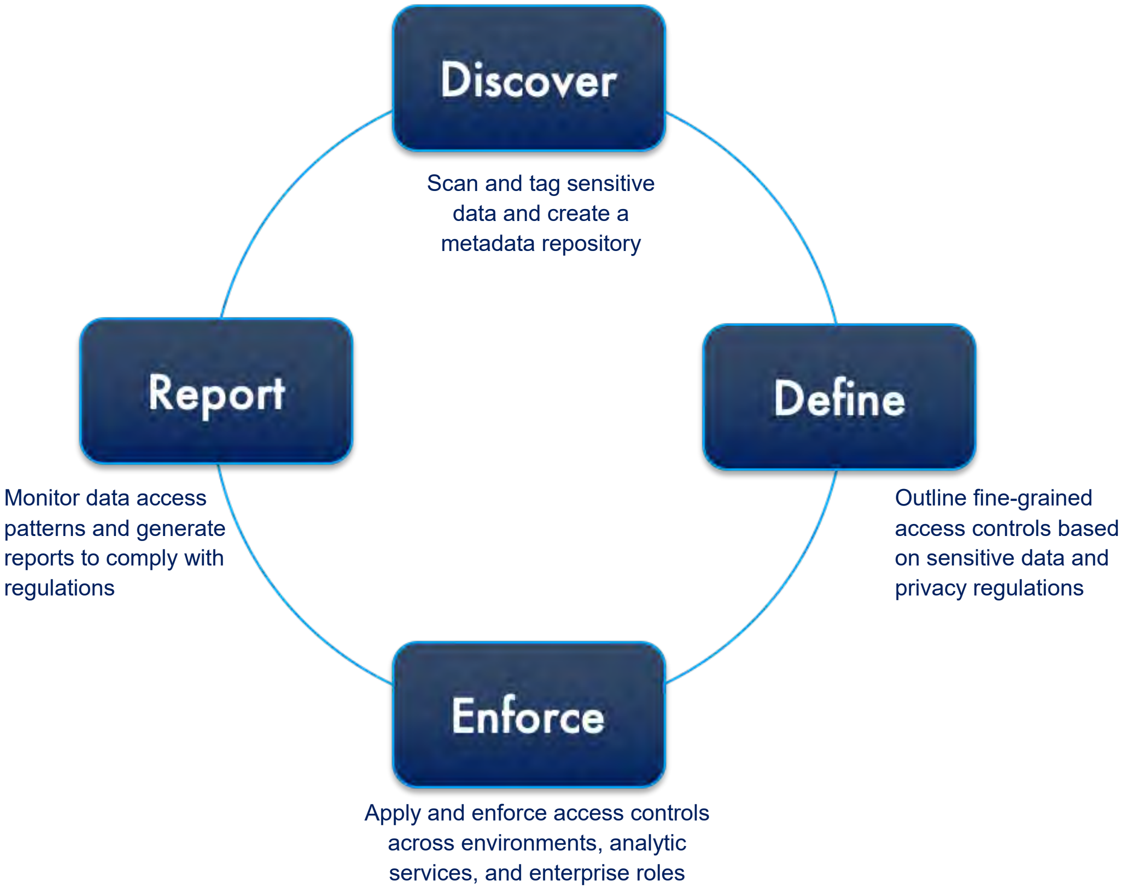

Discover
Scan and tag sensitive data and create a metadata repository
Report
Monitor data access patterns and generate reports to comply with regulations
Define
Outline fine-grained access controls based on sensitive data and privacy regulations
Enforce
Apply and enforce access controls across environments, analytic services, and enterprise roles

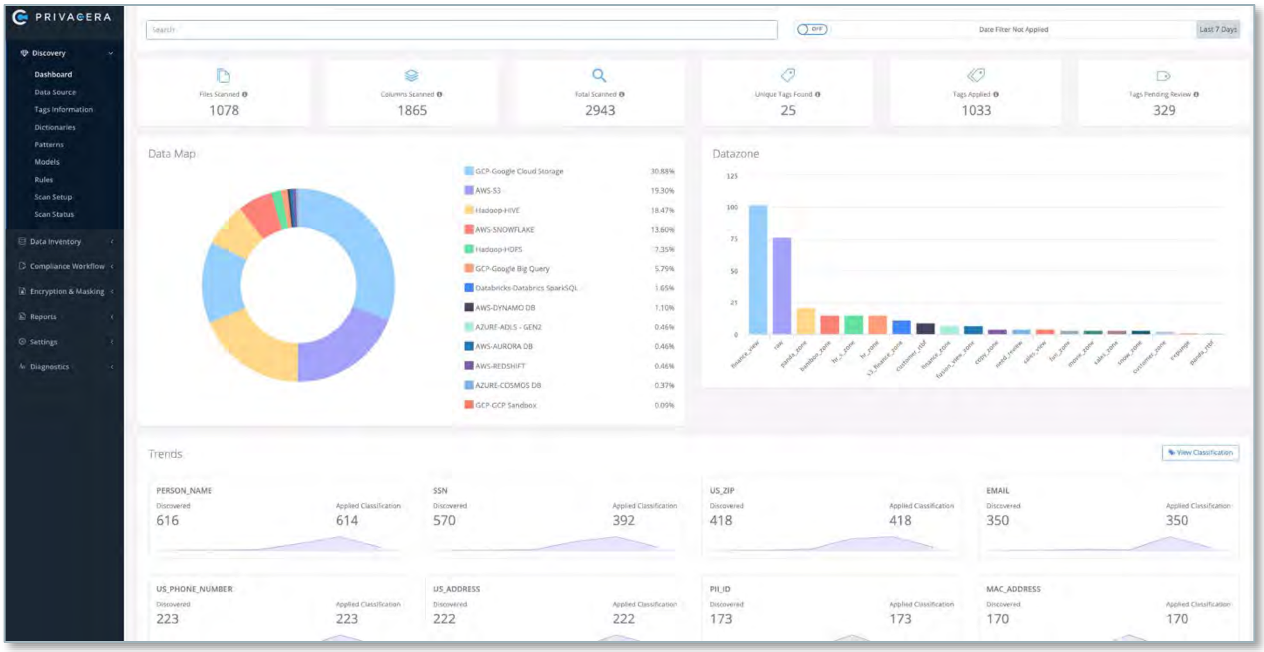
Figure 1: Data Access Governance Lifecycle

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.

Read on to learn how Privacera integrates with Databricks, how to execute the four phases of the data access governance lifecycle, and how enterprises in various industries are leveraging the two solutions together to uncover insights while maintaining privacy compliance.

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.



Files Scanned 1078

Columns Scanned 1865

Total Scanned 2943

Unique Tags Found 25

Tags Applied 1033

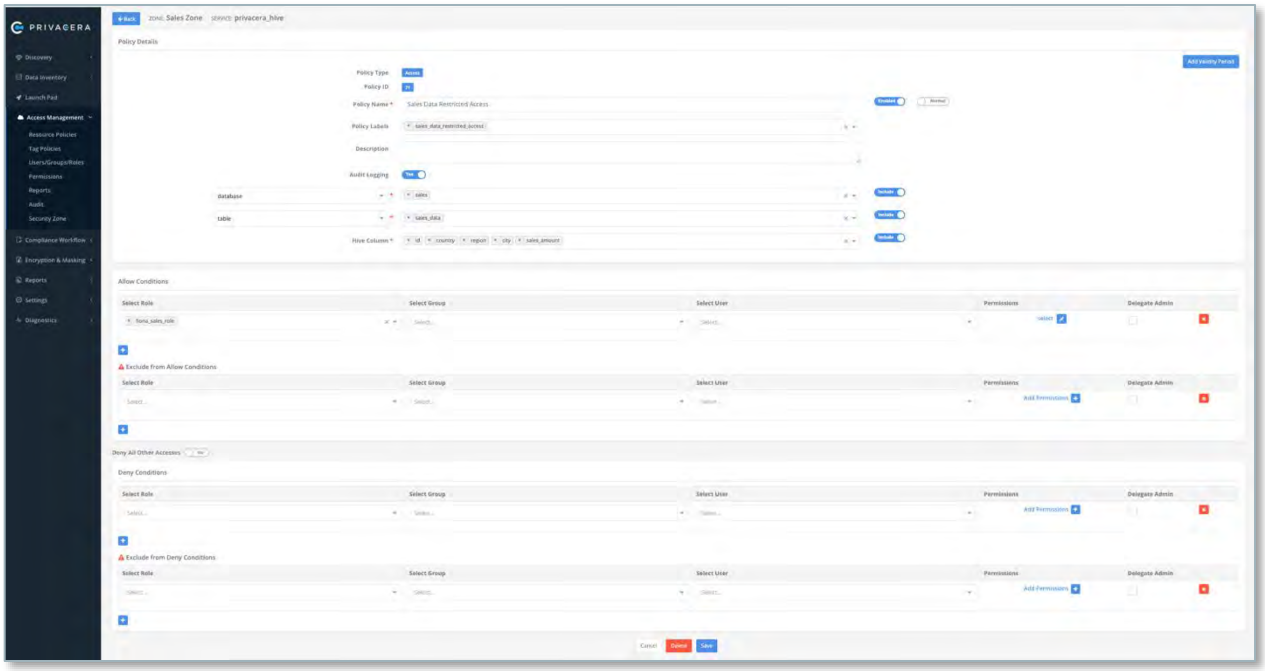
Tags Pending Review 329

GCP-Google Cloud Storage	30.88%
AWS-S3	19.30%
Hadoop-HIVE	18.47%
AWS-SNOWFLAKE	13.60%
Hadoop-HDFS	7.35%
GCP-Google Big Query	5.79%
Databricks-Databrics SparkSQL	1.65%
AWS-DYNAMO DB	1.10%
AZURE-ADLS - GEN2	0.46%
AWS-AURORA DB	0.46%
AWS-REDSHIFT	0.46%
AZURE-COSMOS DB	0.37%
GCP-GCP Sandbox	0.09%

	Discovered	Applied Classification
PERSON_NAME	616	614
SSN	570	392
US_ZIP	418	418
EMAIL	350	350
US_PHONE_NUMBER	223	223
US_ADDRESS	222	222
PII_ID	173	173
MAC_ADDRESS	170	170

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. Refer to the architecture below. This ensures that all data is accessed via the Ranger plugin for authorization.



ZONE: Sales Zone SERVICE: privacera_hive

Policy Details

Policy Type: Access
Policy ID: 71
Policy Name *: Sales Data Restricted Access (Enabled) (Normal)
Policy Labels: sales_data_restricted_access
Description:
Audit Logging: Yes
database *: sales (Include)
table *: sales_data (Include)
Hive Column *: id, country, region, city, sales_amount (Include)

Allow Conditions

Select Role	Select Group	Select User	Permissions	Delegate Admin
fiona_sales_role	Select...	Select...	select	□

Exclude from Allow Conditions

Select Role	Select Group	Select User	Permissions	Delegate Admin
Select...	Select...	Select...	Add Permissions	□

Deny All Other Accesses: No

Deny Conditions

Select Role	Select Group	Select User	Permissions	Delegate Admin
Select...	Select...	Select...	Add Permissions	□

Exclude from Deny Conditions

Select Role	Select Group	Select User	Permissions	Delegate Admin
Select...	Select...	Select...	Add Permissions	□

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

The Privacera platform is a containerized application that can be deployed on any of the three major public cloud providers - Amazon Web Services, Microsoft Azure, and Google Cloud via Docker Compose or natively within a Kubernetes infrastructure. Privacera has also added additional features to ease security administration within a Databricks environment.

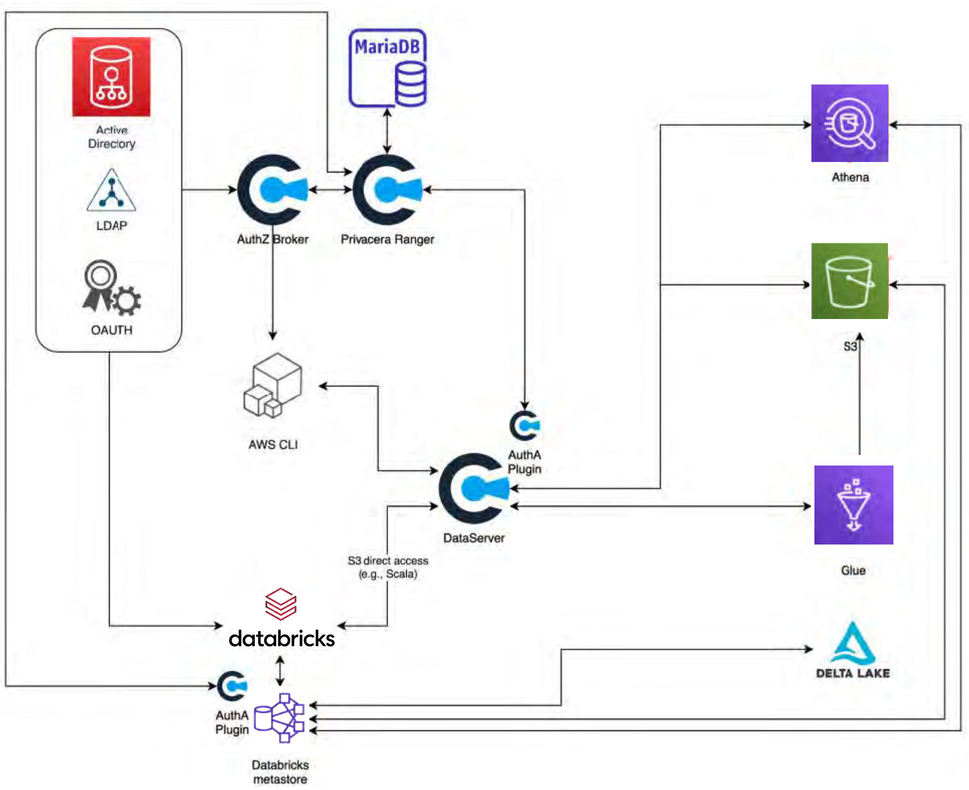


Figure 4: Privacera ensures all data is accessed via the Ranger plugin for authorization

Deployment

The Privacera platform is installed via Docker Compose on VM based environments or within a Kubernetes infrastructure. It can be deployed on any of the three major cloud providers in addition to virtual private cloud (VPC) environments. If Privacera is installed outside of the VPC as part of Databricks deployment, it is recommended to use VPC peering to ensure the security of the communication between the Ranger plugin and the Privacera services.

Privacera supports high availability (HA) and scalability in deployments (e.g. multiple instances behind load balancers). Cloud formation templates and other scripts are also available to simplify and automate the deployment of the Privacera platform.

Ranger Plugin

In order to extend the native access control for Databricks clusters, Privacera platform provides a plug-in model based on Apache Ranger. These plug-ins are lightweight distributed agents that act as the gatekeepers to access various cloud resources. Ranger plugin is embedded within the Spark Driver. When a user executes a SQL query or reads a file from the cloud storage such as S3 or

ADLS, the request is received by Spark Driver. Spark Driver parses the request and generates a logical plan to process the query. Ranger plugin embedded within the Spark driver performs a quick authorization check against the resources that the user is requesting to access. If the user has the required permissions, the plugin then essentially lets the Databricks cluster take over the processing of the query.

The Ranger plugin is installed in Databricks using init scripts provided by Privacera. These init scripts can be deployed globally for all the clusters or locally for the individual clusters. Refer to the figure below. The init script downloads the appropriate Ranger and Privacera libraries and enables the Ranger plugin for the cluster. Databricks calls the init scripts automatically when the cluster is started for the first time to ensure that security is enabled for the lifetime of the cluster.

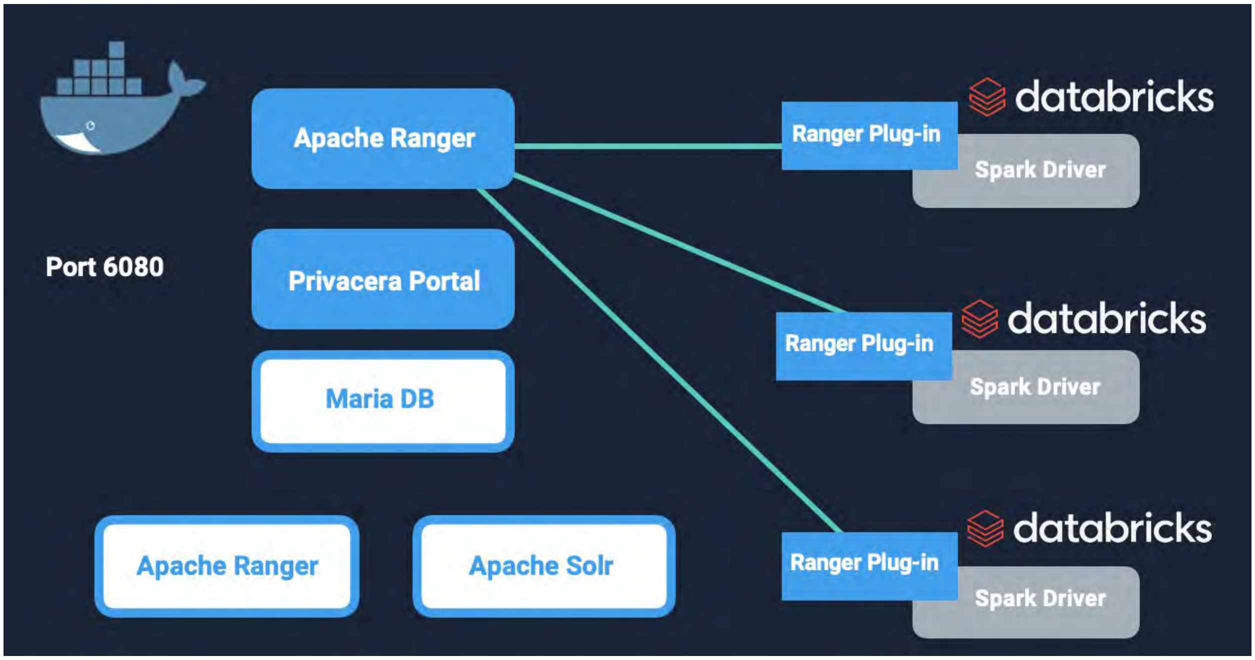


Figure 5: Privacera provides a plug-in model based on Apache Ranger for Databricks clusters

Ranger plugin used in the Privacera platform offers several advantages. Firstly, it doesn’t negatively impact the performance of the cluster because the plugin only performs a quick authorization check and let's Spark access the data directly. In other words, Privacera is not in the data access path. Also, Privacera doesn’t require any additional metadata registration or creation as users can continue to access the tables using the original database and table names and the configured metastore (AWS Glue or Hive metastore for example.) The plugin therefore does not require any modifications to the metastore or any modifications to how the end user queries the data.

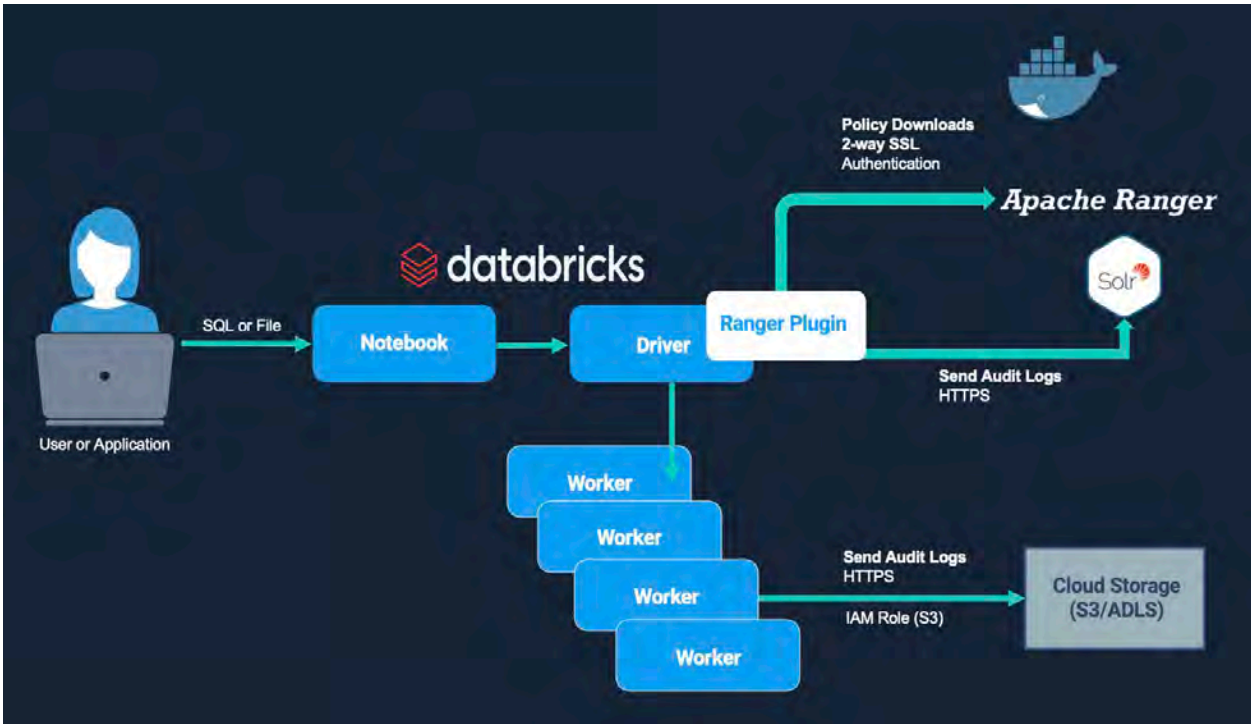


Figure 6: Privacera extends the native access control for Databricks clusters

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 and Azure Data Lake 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.

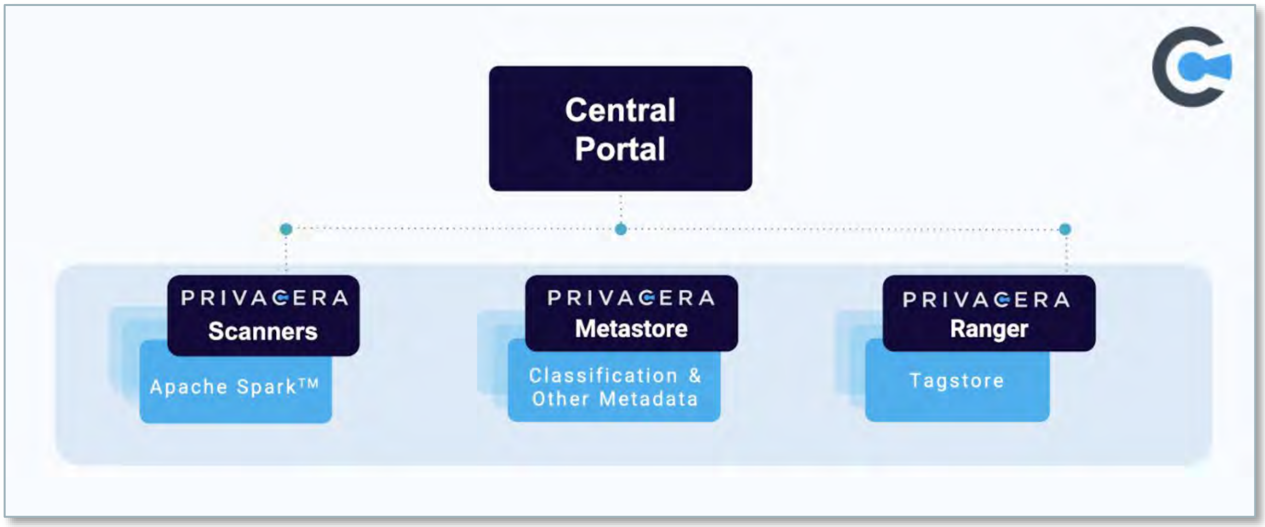


Figure 7: Privacera Data Discovery Architecture

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

Authentication

Before creating, defining and enforcing data access control policies, data consumers must be authenticated. Privacera integrates with popular identity and access management tools, including Microsoft Active Directory and Okta. The Platform supports a wide variety of protocols for users to be authenticated including LDAP, SAML, OAuth, and OpenID. Users can connect to Databricks clusters using username/ password or via SAML authentication. Databricks also supports batch jobs and JDBC connectivity via access tokens generated through the Databricks console.

After the user's identity is authenticated, Databricks passes it to the Ranger plugin. The plugin then uses the identity to perform the following tasks:

- Map the Databricks username (in email format) to a standardized username (generally AD/ LDAP username).
- Retrieve the membership of the user to various groups from Ranger to implement group-based policies.

Fine-Grained Access Control Functionality

As mentioned above, Privacera is based on Apache Ranger, an open source project for creating, defining, and enforcing data access control policies and providing a comprehensive non-reputable 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 or Microsoft Azure.

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
Tag-based and role-based access control (RBAC)
Uses Ranger Plugin along with High Concurrency and Process Isolation

Data Access Control Policy Creation and Enforcement

When a user requests access to a resource either via Spark SQL or file-level access, the Spark plugin uses the synchronized Ranger policies to determine if the user has permissions to access the data resource. The Databricks cluster gains access to the underlying data storage by way of an IAM EC2 instance role (AWS) or an ADLS shared key (Azure). The Spark Driver uses the Privacera plugin to determine if the end user is allowed access to these resources before accessing them on behalf of the user. Refer to the figure below.

If the user is authorized, then the Spark Driver continues to process the request. Otherwise, the request is denied and an error is reported back to the user. The Spark plugin from Privacera does not alter the data access path for the Databricks cluster. The plugin runs in the Spark Driver similar to how other Ranger plugins work, and regularly synchronizes the latest set of access policies from the Ranger server (image 5).

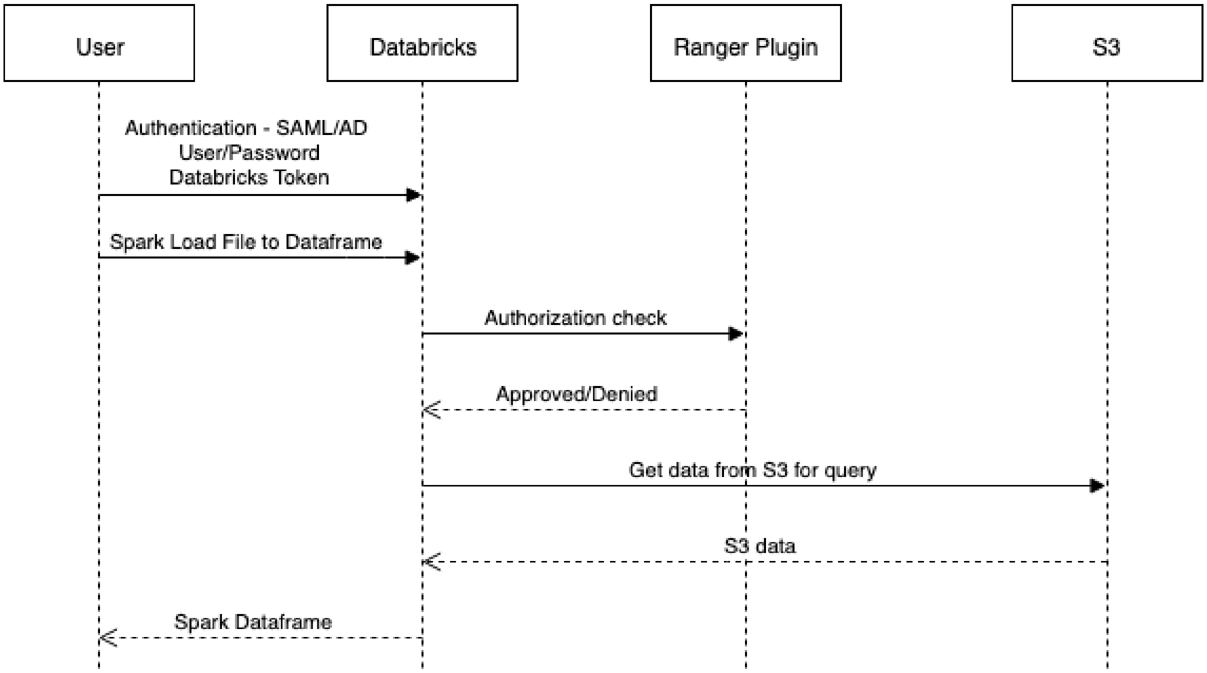


Figure 8: The Spark driver uses the Privacera plugin to authorize the users.

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.

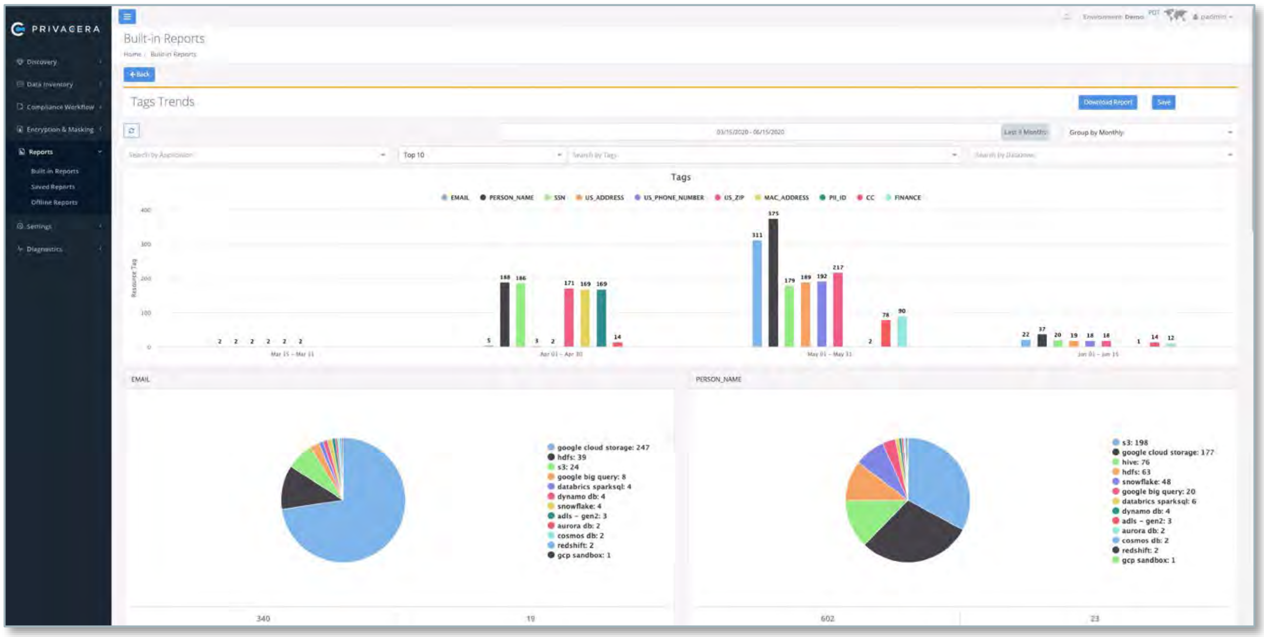


Figure 9: Privacera provides data teams with instant visibility to data assets for regulatory compliance requirements.

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 every end 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 systems and DBFS are blocked from notebook interpreters unless 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 (Beta) 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.

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. Let’s recap how Databricks and Privacera jointly offers enterprise-grade security for Spark clusters:

- 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 Delta Lake to support update/deletes

About Privacera

Founded in 2016 by the creators of Apache Ranger™ and Apache Atlas™, Privacera’s mission is to empower enterprises building data platforms in the cloud to balance data governance and security with data access, discovery, and analytics. Often described as “Apache Ranger in the Cloud”, Privacera provides a centralized data access governance platform that extends Ranger’s capabilities beyond traditional Big Data environments to cloud-native services and leading analytics platforms such as AWS, Azure, GCP and Databricks. Privacera enables IT and data platform teams to make as much data as possible available to the business for analytics while ensuring it is used ethically and in compliance with privacy regulations.

Email us at questions@privacera.com, visit www.privacera.com, or follow at @privacera to learn more.