privacera

Apache Ranger and Privacera:

Key Similarities & Major Privacera Enhancements

Whitepaper

Introduction

Apache Ranger (Ranger) is a highly successful open-source project used by thousands of enterprises worldwide for its robust and extensible data access control framework. Ranger provides comprehensive authorization, audit, and encryption capabilities needed to govern data in Hadoop-based big data infrastructures effectively.

When building Apache Ranger, the goal was to develop an enterprise-ready, centralized platform to define and administer data access controls for on-premises Hadoop data lakes. However, the problem of centralized access control also extended to cloud services. This challenge was the motivation behind the formation of Privacera. Ranger's proven framework lies at the heart of Privacera's unified data governance platform. Privacera supports multiple cloud services — AWS, Microsoft Azure, Google Cloud, Databricks, Snowflake, and more — in addition to enhancements that are critical to securing analytical workloads by enterprises.

The objective of this white paper is to outline the similarities between Ranger and Privacera and discuss the essential product enhancements Privacera has built to address some of the most vital use cases in enterprise digital transformation. After an overview of Apache Ranger and Privacera, the paper will explain why enterprises should consider Privacera's unified access governance platform for their cloud environments, especially if they use Ranger to secure access in their on-premises Hadoop data lakes.

The Origins of Apache Ranger and Privacera

Apache Ranger emerged on the big data scene to help enterprises secure their Hadoop platforms by using a centralized and open-source approach to authorize users' access to various data repositories and compute engines. These repositories included open-source projects such as Apache Hive, Apache Kafka, Apache Spark, Apache NiFi, Apache Atlas, and more. A testament to the success of these open-source projects is that today, all major public cloud providers continue to offer them as part of their offerings, including EMR on AWS, HDInsight in Azure, Dataprocs in GCP in addition to Dremio and PrestoDB.

Apache Ranger at a Glance

Number of Releases (Major & Minor)	17
Number of Committers & Contributors	71
Formation of Apache Ranger as an incubator Project	2014
Recognition of Apache Ranger as a Top-Level Project (TLP)	2017
Lines of Code in the Latest Apache Ranger Release	389,000
Estimated Number of Companies Using Apache Ranger	3000+

The integration of independent open-source engines into a platform provided companies with the flexibility to address their unique use cases as these engines had a different mechanism for allowing access to data for analysis purposes. Administrators had to be well versed in the security mechanism for each engine; they also had to build the same policy repeatedly for each machine in use. This lowered an administrator's productivity and increased the probability of making mistakes or building inconsistent policies across the various engines. Apache Ranger was created to address this inconsistency by defining, administering, and reporting on security policies consistently across Hadoop components.

Similarly, these challenges exist in the cloud but on a far larger scale. Suppose the compute engines of the Hadoop ecosystem are replaced with the various cloud services that are routinely used as part of any company's public cloud infrastructure today. In that scenario, it becomes clear that the governance and access control issues of the Hadoop data lake are alive and well in the public cloud. Not to mention controlling access to other components such as storage repositories (Amazon S3, Azure Data Lake Storage, Google Cloud Storage), query federation (Starburst, Dremio, Presto), and business intelligence tools (Power BI, Tableau) in a complex cloud ecosystem. Enterprises demand a single platform to authorize access to disparate cloud services, and Privacera addresses this need by extending Ranger's enterprise-grade framework to the cloud. Privacera's unified access governance technology provides single-pane visibility

and consistent access management across hybrid- and multi-cloud environments via fine-grained access control, automated sensitive data discovery, dynamic masking and encryption, and continuous audit and monitoring in one scalable platform. Privacera natively integrates with all major cloud vendors and modern data platforms (Databricks, Snowflake, Starburst, Dremio, among others), governing petabytes of enterprise data that helps their customers drive responsible data-powered performance.

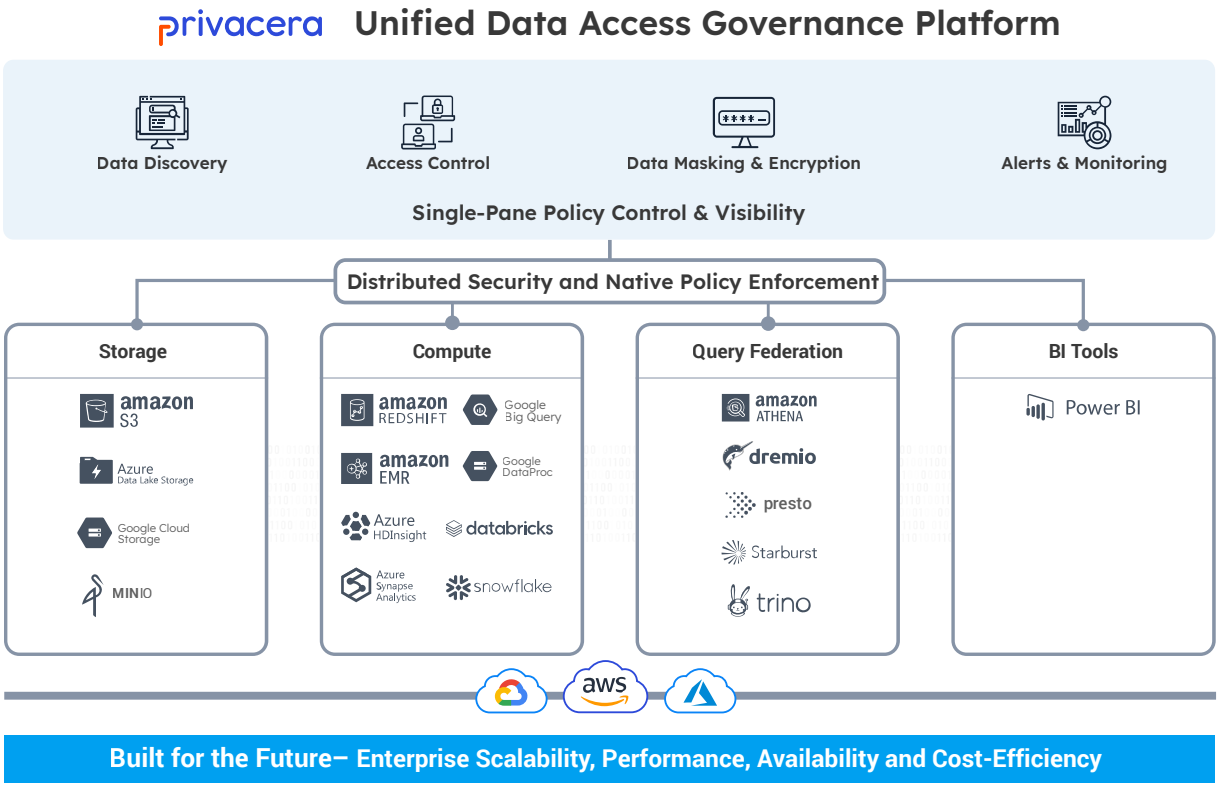

Figure 1: Privacera Access Governance Platform architecture

Built on Top of Apache Ranger's Foundational Access Control Capability

Apache Ranger was the first to introduce the use of plugins purposely built for access control. A plugin architecture featuring native integrations with the data sources provides a lightweight footprint that is easy to layer into complex storage and compute systems. Because the plugins are natively built for the source systems, they don't introduce added complexity, dependency, and overheads. Instead, they can swiftly authorize users to support the performance of thousands of users simultaneously accessing and querying data in production environments at a petabyte scale.

With lightweight plugins, Ranger can enforce the access policies that run natively within the Hadoop components to determine whether a particular user is authorized to access a specific

set of data based on required permissions. Ranger plugins inherit the security policies as outlined in the policy database in the Ranger Policy Server and pull the latest version of the access policies at specified intervals using a REST API from the policy server to enforce the updated guidelines. Plugins remain operational even if the policy server is unavailable and enforce access policies based on the last update from the policy server.

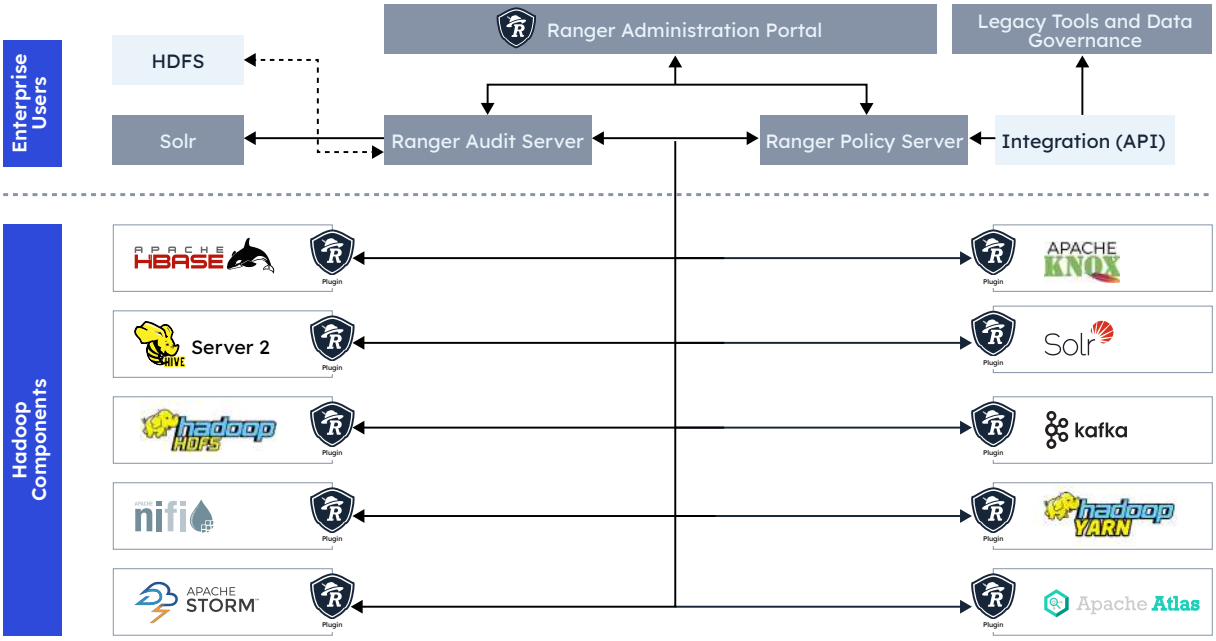

Figure 2: Apache Ranger architecture

Privacera uses Apache Ranger as the authorization engine and enables seamless policy migration from on-premises Hadoop to the cloud. Furthermore, Privacera has extended Apache Ranger with three different approaches to policy enforcement based on the data store and the type of access for the hybrid-cloud environment. All provide consolidated audit logging. And in all cases, data does not have to stream through Privacera's code, so the overhead added by any policy enforcement is kept to a minimum.

- **Access Control via Ranger Policy Enforcement Points (Plugins)**
  Where available, Privacera leverages Apache Ranger-style distributed policy enforcement points for access control. Many data processing engines support these plugins, such as Hive, Spark, and Presto. Policies created and managed in the Privacera portal are distributed to and synchronized with these policy enforcement points. Access control decisions happen at the engine, in line with query execution.
- **Access Control via PolicySync**
  For data sources like RDBMS, where Ranger-style access control plugins are not available, policy synchronization (PolicySync) enforces access policies. PolicySync enables rapid development of connectors which can translate Ranger policies into the native access control framework of the data sources, including Databricks SQL, Snowflake, RDS, Redshift, and more. For example, by translating access policies into GRANT/REVOKE

statements for a relational database and generating views where needed for additional layers of access control like data masking and row filtering, and so on. In addition, when there are policy changes in Privacera, new or updated objects in the data source, or changes to users, groups, and roles, updates are pushed to the data source in real-time to keep it aligned with the latest policies.

- **Access Control via Data Access Server**
  For data in object stores like Amazon S3 or Azure Data Lake Storage (ADLS), access requests flow through Privacera's Data Access Server for policy enforcement. The Data Access Server integration method redirects data access requests to a Privacera 'authentication broker' inserted into the control and data flow. For requests that are allowed based on authentication and other policy checks, the authentication broker generates a signed URL that the requestor can use to fetch the requested data directly from the object-store and audits all-access attempts.

In addition, Privacera provides enterprises with a seamless and automated method to manage user access provisioning and de-provisioning for their data. Using Flowable, a proven open-source business process workflow engine embedded in the Privacera platform, enterprises can automatically control policy creation and updates based on user requests in accordance with their **roles** or **attributes** such as location and other metadata for the data to which they need to gain access. This process replaces the prior manual authoring of access control policies and managing entitlements through manual user intervention of the policy administrators.

With Privacera, data administrators have the flexibility to define access policies at a database, table, column, object, or file level. As a result, data administrators can administer fine-grained access controls for on-premises data lakes, public cloud services, as well as third-party cloud-native services such as Databricks, Snowflake, and others — all from a single console.

	privacera	Projects and Products
Cloud Data Warehouse	✓	Amazon Redshift, Azure Synapse, Snowflake
Cloud Analytics & ML	✓	Amazon SageMaker, Azure ML, jupyterhub, Databricks
Cloud Big Data	✓	Amazon EMR, Azure HDInsight, Delta Lake, Google Cloud Dataproc
Cloud Storage	✓	Amazon S3, Azure Data Lake Storage, Google Cloud Storage
On-Premises Big Data	✓	hadoop, HIVE, APACHE HBASE

Figure 3: Privacera extends Apache Ranger to the cloud

Introduced Data Discovery to Classify Sensitive Data

Sensitive data, such as social security numbers, street addresses, credit card information, and more, require some level of protection or governance to comply with privacy regulations or the enterprise's own internal data usage requirements. With the volume of data created every day (by transactional systems, the Internet of Things, social media, and more), determining what data qualifies as sensitive is a daunting task. Before the era of Big Data, it was possible to manually identify sensitive data flowing into enterprise systems and view sensitive data stored within data infrastructures. But with the sheer volume of data created and stored across numerous sources today, manual efforts won't suffice. Enterprises must embrace automation to understand and secure sensitive data.

Privacera Data Discovery enhances the Ranger functionality, empowering enterprises to leverage automation and sophisticated techniques to understand the context of sensitive data and accurately classify it. When coupled with Privacera Access Manager and Privacera Encryption Gateway, Privacera Data Discovery maximizes the visibility and security of data with a 3-step approach:

1. Firstly, the Data Discovery module scans all data entering the enterprise data ecosystem consisting of analytics services and other storage environments, both on-premises and in the cloud.
2. Then, it uses a variety of techniques to identify and classify sensitive information:
   - Pattern matching with regular expressions
   - Dictionaries to look up data from a whitelist or blacklist
   - Sophisticated heuristics that look at both the data content and the context in which the data is located, such as the table or column name
3. And finally, it automatically creates a sensitive data catalog to provide a unified view of all sensitive data and related classifications stored across all enterprise systems.

By reviewing and updating the classifications generated by the scanners, enterprises can further implement policies to protect sensitive data in conformance with their requirements.
For example, most organizations might want a policy that allows non-privileged users to see only the masked or transformed versions of certain sensitive fields such as SSNs or credit card numbers.

Furthermore, by grouping data sources into administrative data zones, maintenance and control of the assets in these zones can be delegated to the owners of the data in the enterprise's organizational groups. The Data Discovery module also provides a variety of reports to aggregate, summarize, and drill down into the classification results across the entire collection of data assets.

Without the visibility into the types of sensitive data and where they reside in the system, it's nearly impossible to create and enforce smart access control and comply with internal and external regulations. Therefore, organizations first need to identify and tag sensitive data to

reflect its makeup and resulting sensitivity as part of an enterprise effort to manage access to sensitive information. With Privacera Data Discovery, organizations can swiftly identify sensitive PI/PII information in structured and unstructured data. It can also mask sensitive data based on enterprise use cases. All sensitive data classifications sync with Privacera Access Manager, enabling fine-grained access control at the file, table, row, and column levels for various d ata sources.

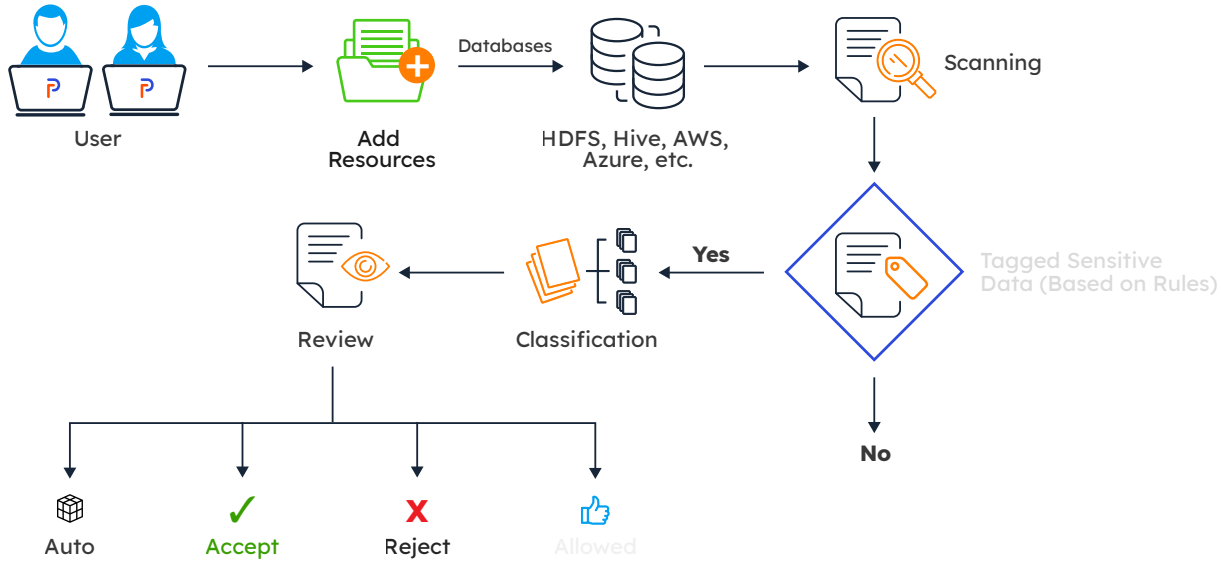

Figure 4: Privacera Data Discovery workflow

Enhanced Encryption to Enable Enterprise Grade Data Protection

Ranger provides a key management service (KMS), an open-source, Hadoop cryptographic essential management service to manage encryption keys for HDFS Transparent Data Encryption. Traditionally, the Hadoop KMS stores keys in a file-based Java keystore. Ranger has extended the native Hadoop KMS functionality to store keys in a secure database and provide a centralized administration of key management through the Ranger admin portal. It also provides an audit trace of all operations performed by Ranger KMS.

Due to the rich functionality and proven scalability of Ranger KMS, Privacera has integrated it as a critical component in the Privacera Encryption module. With Privacera Encryption, organizations can encrypt data at the table, column, row, field, or attribute level instead of the entire data. This granular level of encryption enables the data science and analytics teams to utilize more data to build models and extract insights to drive new business opportunities, leading to increased customer satisfaction and optimized business efficiency. Even if the data is encrypted, it is automatically decrypted for authorized users or applications when they access the data. As a result, the user experience accessing encrypted data on a disk or in the cloud is identical to accessing non-encrypted data.

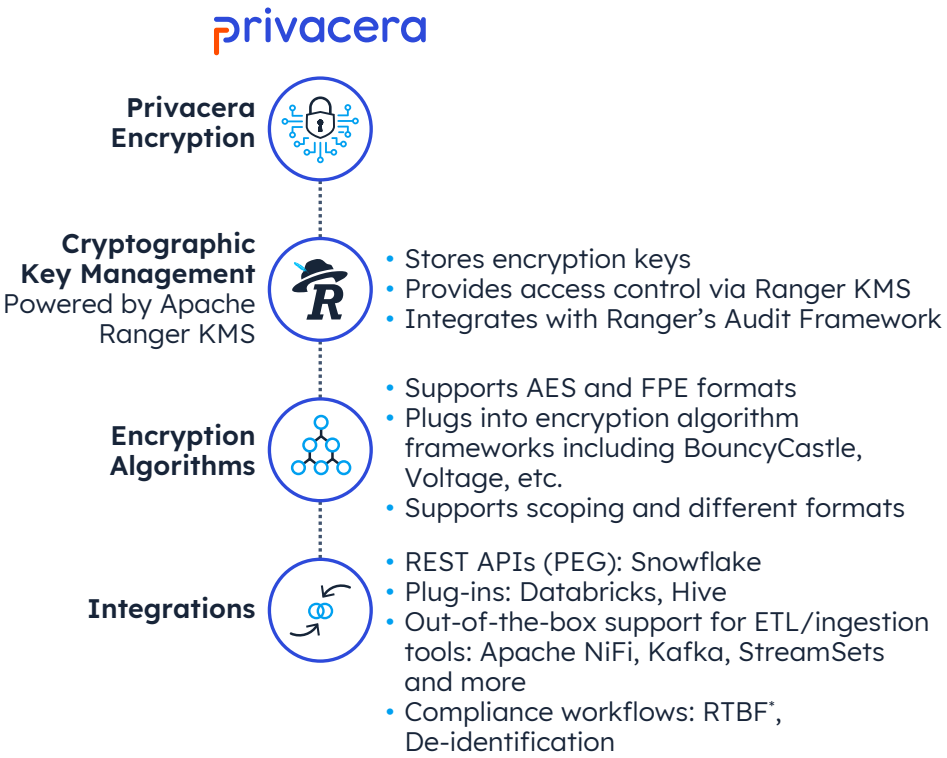

Figure 5: Privacera Encryption at a glance

Privacera Encryption supports Advanced Encryption Standard (AES) and Format-Preserving Encryption (FPE) formats. And to support the dynamic cloud services, data sources, and emerging data and regulatory privacy compliance, Privacera Encryption has four types of offers:

- Compliance workflow for use cases such as the Right To Be Forgotten (RTBF). Users can create policies that map to an encryption scheme. These policies can scan for and encrypt sensitive information residing in cloud object stores and data platforms in AWS, Azure, and Google Cloud Platform.
- Plug-ins for Databricks and Hive.
- Support for ETL/ingestion tools such as StreamSets and more.
- Application programming interfaces (APIs) for encryption and decryption.

The API is a standalone service called Privacera Encryption Gateway (PEG). It significantly lowers the operational burden on infrastructure and security teams as they are not required to install, manage, and update separate encryption and decryption tools. It provides data encryption to protect digital data confidentiality as it is stored and transmitted. Encryption algorithms support security initiatives, including authentication, integrity, and non-repudiation.

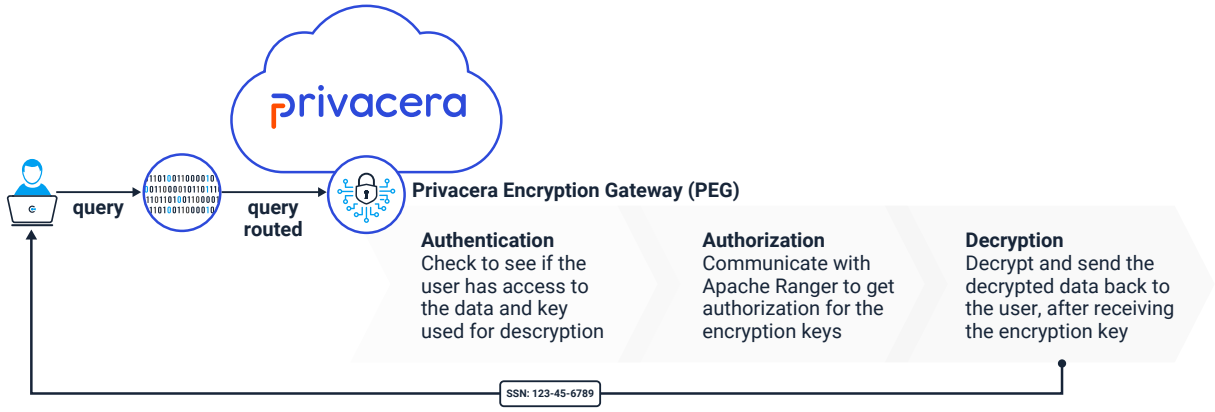

Figure 6: Privacera Encryption Gateway workflow

Secure Data Sharing for Faster, Broader, Deeper Data Analytics

Enterprises use Privacera's centralized access governance solution to address three use cases for driving responsible data-powered performance:

- **Accelerated Migration of Analytical Workloads to the Cloud**
  At Privacera, the migration of on-premises Hadoop data lakes to the cloud is the primary use case among its customers. When companies decide to migrate their analytical workloads to the cloud, they need to migrate data and the usage context to their cloud infrastructure, including access control and data governance policies. Without proper access governance in place, enterprises cannot rapidly democratize data and risk missing critical insights that can add value to their business.

  These companies have invested tremendous effort and resources into building their access control policies to govern their on-premises data lakes. If companies can migrate their existing access control policies to the cloud, they can significantly accelerate the process to onboard their users to the subscribed cloud services and help ensure the success of the company's cloud initiative. It is, therefore, critical that authorized access to data in the cloud is transparent to the users, and data consumers require no behavior-al changes or re-registration of data. Privacera's unique architecture ensures that data analysts and scientists continue to use their previously-built queries without modifications. These queries must reference the same table names and object locations for the data to be accessed seamlessly from the file or the object.

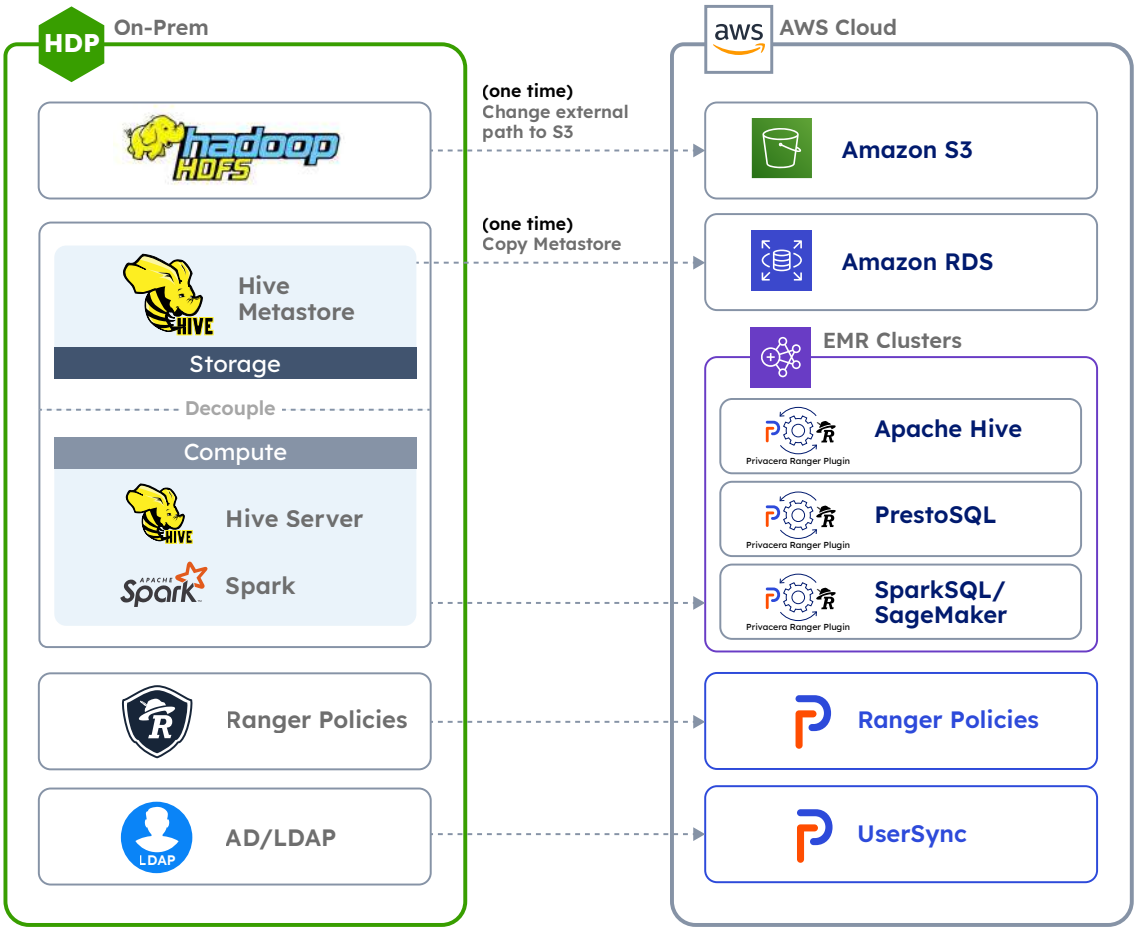

Figure 7: On-prem Hadoop to AWS migration

- **Automated Regulatory Compliance**
  Companies that collect, process, and store customer or personal data must comply with an ever-increasing set of privacy and industry regulation. Failing to do so has the potential to not only impact an organization's business reputation and incur hefty fines, but also, in extreme cases, threaten its ability for continued operations. This concern is magnified across modern hybrid- and multi-cloud architecture in which consistently applying policies across settings and tools is a challenge. As a result, achieving and maintaining regulatory compliance can be an overwhelming endeavor for organizations that lack a structured approach.

  Privacera's centralized data security and governance platform enables analytics teams to access and utilize structured and unstructured data in cloud services and on-premises repositories while meeting data security, privacy, and compliance requirements. The platform automatically detects, catalogs, and processes sensitive data as the platform ingests data. The result is the ability to comply with crucial provisions of privacy and industry regulations such as GDPR, CCPA, Brazil's LGPD, PCI DSS, and HIPAA, respectively.

- **Governed Data Sharing**
  Fractured analytics processes make sharing data internally and with business partners an inefficient process that cannot keep up with the pace of modern business. The solution to granting secure access to the universe of available data to fulfill business objectives while upholding guardrails for regulatory compliance and sensitive data is to employ a unified data access governance/data privacy platform tailored for hybrid and multi-cloud deployments. This modern offering inherently supports the distributed data access governance model in which data stewards familiar with use cases in respective business domains grant end user access by enforcing centralized policies at the local level in individual sources.

  To further drive latency from the self-service analytics process, Privacera introduced a new data sharing approach called Governed Data Sharing that delivers a new level of flexibility and power to accelerate analytical initiatives by grouping functional data–such as sales, marketing, finance, and more–into Data Domains. Access policies are automatically applied to data sets inside Data Domains where data consumers, such as data scientists or business analysts, can browse through an inventory of data sets and request access to them. The use of Data Domains alleviates the operational burden of IT by putting data domain owners in direct contact with data consumers to manage access requests, greatly improving collaboration, flexibility, and responsiveness. This way, uniform policy enforcement occurs regardless of where users or data are, as delegated by stewards or data domain owners who know the data best.

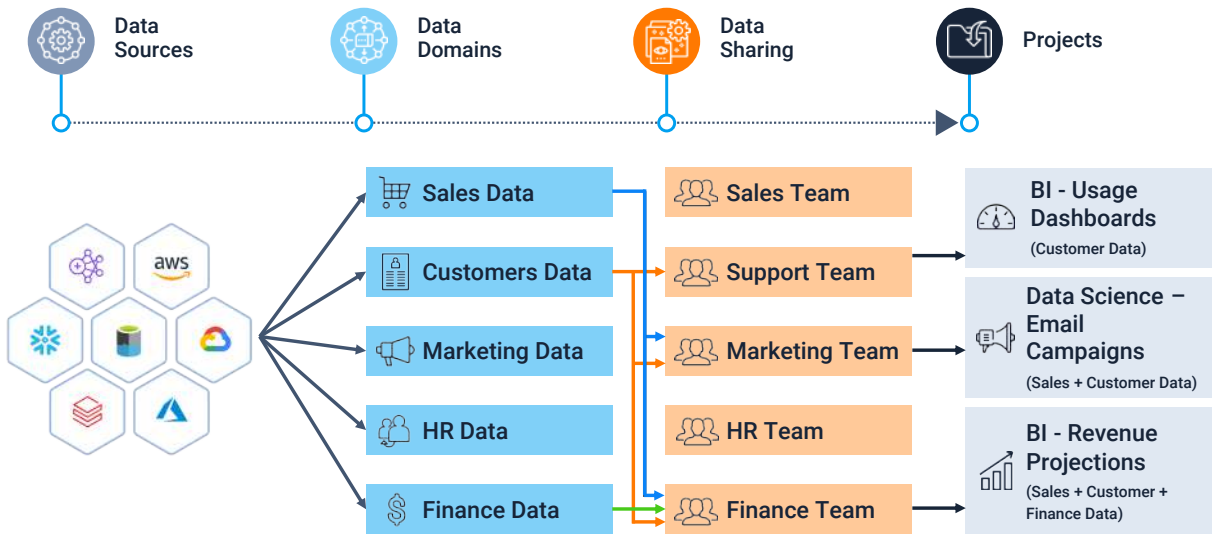

Figure 8: Overview of Governed Data Sharing

Additional Functionalities

In addition to the enhancements mentioned above, Privacera has built several functional capabilities, including:

- **Authentication**
  The Privacera portal natively supports single sign-on (SSO) to web UI via SAML, OAuth, and OpenID Connect.
- **Admin User Interface**
  The Privacera portal is based on React, a modern UI framework that is optimizes for speed and performance. The portal also integrates with the Ranger UI.
- **API Authentication**
  Privacera's version of Ranger API supports the following authentication mechanisms:
  - Kerberos
  - Simple Auth
  - JWT-based
  - Access key/ Secret key
  - Token-based
- **User Sync**
  Privacera's platform inherits users and groups from LDAP, AD, Linux, Azure AD, in addition to any user or group repositories that support SCIM standards. Additional user properties such as email, phone, etc., can be synced from user repositories. Lastly, Privacera exclusively maps user identity between various formats. For example, in Databricks, email user IDs are mapped to a shortname.
- **Tag Sync**
  Privacera provides improved performance by optimizing and locally caching tags to resource mappings.
- **User-Group Mapping for Policy Engine**
  Services that invoke Ranger authorizers in Privacera do not need to provide group info or pass user IDs as part of the request context. Privacera's Ranger plugin automatically looks up the groups from Ranger Admin as part of User Sync. This authorization dramatically simplifies the process of adding new authorizers via plugins/ DAS/ PolicySync for cloud deployments where group info is typically not available.
- **Audits**
  Audits can be streamed to SIEM systems via messaging services–Kafka, AWS Kinesis, Google Pub/Sub, Azure Eventhub – via configuration and support out of the box. Moreover, audits can be sent to Syslog, object store destinations (S3, ADLS, etc.), and written in compressed, optimized formats (ORC, for example) into audit stores via configuration.
- **Integrated Data Browser**
  Privacera File Explorer for AWS, Azure, and GCS enables users to browse the contents of S3, ADLS, and GCS buckets. It can also browse objects or folders and files, opening them if appropriate permissions are available for the user via Ranger policies.

- **Deployment Versatility**
  Customers can deploy Privacera through enterprise-grade containers with high availability. Privacera can also be deployed via Docker Compose or Kubernetes in stand-alone or cloud-managed service modes. The deployment includes all the dependent services like Solr and DB. Privacera can also be installed and deployed as a cloud-native containerized application in addition to VM and software-only packages.

Apache Ranger and Privacera: Conquering Modern Data Access Governance Challenges

Ranger is at the core of most modern data management and analysis tools available today: Amazon EMR (Hadoop), Databricks (Apache Spark), Starburst (Trino), Confluent (Apache Kafka), Glue (Apache Hive Metastore), Google Cloud Dataproc (Hadoop), Azure HDInsight (Hadoop) and many more. Thousands of enterprises accept Ranger as a robust access control platform, and all major cloud vendors like AWS, GCP, and Azure natively integrate with it. It is also the security platform of choice for pure-play data virtualization vendors like Dremio and Starburst.

Due to its robustness, performance, and proven scalability, Privacera has chosen Apache Ranger as its underlying engine for access control and has advanced its capability with many out-of-the-box features to optimize for the cloud data environment. It continues to leverage the power Apache Ranger brings to help future-proof data access governance for its customers who operate in multi-cloud and hybrid data estates to become data-driven organizations.

Learn more about Privacera here or contact us to schedule a call to discuss how we can help your organization meet its dual mandate of balancing data democratization with security to maximize business insights while ensuring privacy and compliance.