PRIVACERA

Your Data Has Moved to the Cloud... What About Your Governance Policies?

Buyers Guide to Data Governance in the Cloud

Whitepaper
January 2021

Content

questions@privacera.com | privacera.com | 510.413.7300 privacera @privacera

1 Importance of Cloud Computing

Cloud computing represents one of the most profound IT trends of this decade. Market forecasts, statistics, and projections all point to the importance of the cloud-based services as a juggernaut. According to Gartner, the world-wide public cloud services market, which is estimated to be about $266 billion in 2020, is projected to grow to $335 billion by 2022. McKinsey & Company expects about 35 percent of all enterprise workloads to be on the public cloud by 2021, and anticipates 40 percent of companies will use two or more infrastructure-as-a-service (IaaS) and software-as-a-service (SaaS) providers[1].

The movement towards the cloud started as a strategy to reduce IT infrastructure costs and increase efficiency. Public cloud providers offer data storage and compute services to customers over the internet. By moving to the cloud, businesses are able to eliminate the significant capital expenditure of building and maintaining their own data centers. In doing so, companies can bring their applications online faster. Moving their IT infrastructure to the cloud also enables companies to utilize their resources more efficiently; they can dynamically adjust computing resources devoted to their applications based on the demand these applications are experiencing at any given time. This means businesses can add or remove storage and compute resources on-demand and independently – to operate elastically, if you will.

It comes as no surprise companies are moving to the cloud in droves, and it's safe to say that cloud computing is here to stay, with its reach only increasing with time. As a response, technology vendors and service providers must adjust their business models and product strategies and adopt the cloud-first mantra.

McKinsey[2] cites the following performance improvements to be realized by companies that develop world-class operations in the cloud:

75%
reduction in IT complexity

Up to 20%
efficiency gains within 12 to 18 months of implementation

Up to 50%
improvement in reliability and availability

~4X
faster capacity deployment

5 -10%
reduction in overall cost year-over-year

[1] Source: Gartner, November 2019

[2] Transforming infrastructure operations for a hybrid-cloud world, October 2019

2 Evolution of Cloud Computing

Migrating to the cloud, which started out primarily as a cost reduction play for IT, is now a full-blown business strategy with implications for the survival and competitive advantage of companies. For the past ten years, leading cloud vendors have expanded their offerings from infrastructure to higher-level services and business applications. For example, AWS – which started out with three foundational services, Amazon Elastic Compute Cloud (Amazon EC2), Amazon Simple Storage Service (Amazon S3), and Amazon Simple Queue Services (SQS) – now offers more than 212 services for storage, computing, databases, analytics, application services, and Internet of Things (IoT).

More recently, cloud vendors have moved up the stack to provide packaged applications for developers, data analysts, and data scientists. These analytics and data science services alleviate the task of managing the underlying infrastructure by automatically provisioning compute and storage resources required to perform a particular type of workload, then decommissioning those resources when the task is completed.

It is fairly common for a business running analytics and machine learning workloads in the cloud to subscribe to multiple services from the public cloud and 3rd party service providers. For example, a company can subscribe to Snowflake data warehouse, Databricks analytical platform, and EC2 on AWS at the same time. This requires data infrastructure administrators to manage users' access to each of them individually which in turn prolongs the time to onboard users, eads to policy proliferation, lowers administrator productivity and increases the probability of making an error.

3 Parallels with Data Lake

Developers, solution architects, and DevOps personnel who have implemented on-premises data lakes should see some familiar patterns as they build cloud-based infrastructure for their companies. Hadoop-based data lakes began to appear on the IT scene in the early 2000s. The implementations of Hadoop-based data lake consisted of a number of data processing engines, such as Apache Hive, Apache Spark, Apache Kafka, and Apache HBase operating on top of a distributed storage layer such as Hadoop Distributed File System (HDFS). These engines represent independent open source projects not designed to work together. Hadoop distributions from Hortonworks and Cloudera emerged to provide integrated platforms to ensure that batch, streaming, real-time, and machine learning workloads can be processed using a single platform.

Hadoop-powered data lakes can provide a robust foundation for an enterprise to generate insights based on new and diverse sources of data. However, the data lakes had some shortcomings. First, the Hadoop ecosystem consists of numerous open-source compute engines that were developed by teams of developers working independently, each engine offers its own unique mechanism to administer security or access control. Users across multiple business units can access the data lake freely using one of the engines they had access to, thereby increasing risks of exposure to unauthorized users. However, any internal or external breach of enterprise-wide data in the data lake can be catastrophic. Implications can range from privacy violations, regulatory infractions, to damaged corporate image and shareholder value. With the company's business, customers, finances, and reputation at stake, IT leaders needed to ensure that their data lake had high standards of data security and governance in place.

Second, as Big Data in the form of terabytes of structured and unstructured data began to flow into the company's data lakes, the necessary data governance framework and controls were not in place. With the data governance framework still being defined, the schema-on-write capability of Hadoop did not help matters. As the volume of data in the data lake rose, it became increasingly problematic for IT departments to determine the contents and lineage of the data in their lakes. As a result, businesses started to lose confidence in the integrity of the data in their data lakes. That's when data lakes begin to transform into the proverbial data swamps.

4 Ranger to the Rescue

To address the first problem and deliver consistent security administration and management, Hadoop administrators needed a centralized user interface that could define, administer, and manage security policies consistently across all the compute engines of the Hadoop stack. Apache Ranger provided a single-pane of glass for Hadoop administrators to deliver fine-grained access control through a centralized interface that ensures consistent policy administration. Security administrators now have the flexibility to define security policies for a database, table, and column or a file, and administer permissions for specific LDAP-based groups or individual users. Rules-based on dynamic conditions such as time or geography can also be added to an existing policy rule. The Ranger authorization model is highly pluggable and can be easily extended to any data source using a service-based definition.

Fast forward to the present. **If we replace the compute engines of the Hadoop eco-system with the various cloud services that are now routinely part of any company's public cloud infrastructure, you will clearly see the governance and access control issues of the Hadoop data lake are alive and well in the public cloud**. IT teams are now being pulled in opposite directions. They need to balance their corporate mandate of governing and securing enterprise data, while providing trusted data seamlessly to the business for analytics, data science, and collaboration. However, this issue is magnified, because things move at a faster pace in the cloud, and controls can lag. According to a McKinsey survey, 69 percent of organizations cite that *"implementing stringent security guidelines and code review processes can slow developers significantly."*

This is especially troubling, because one of the primary drivers for businesses to migrate the IT infrastructures to the cloud is the speed and agility at which data can be provided to data analysts and scientists to uncover the insights hidden.

5 Governance Challenges of the Cloud

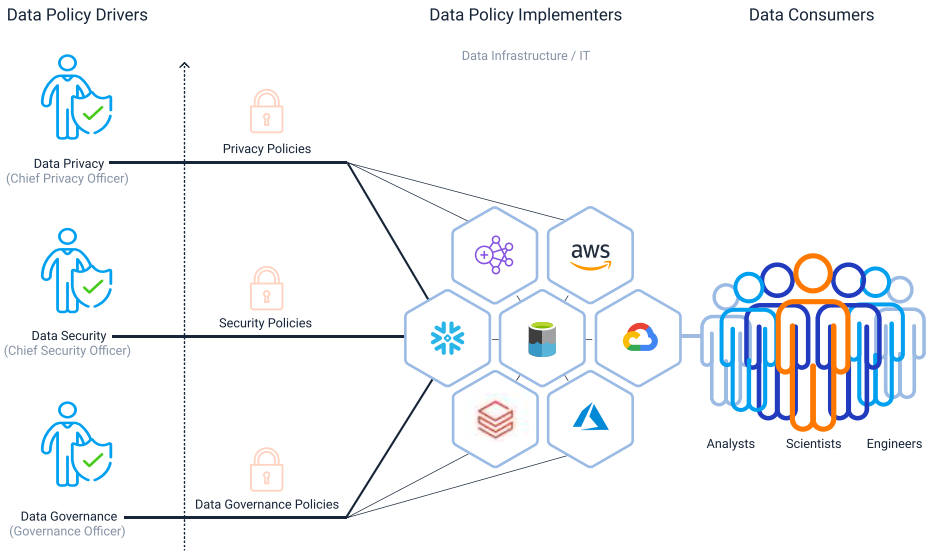
Let's explore the various challenges companies face when migrating to the cloud.

- **Disparate Data Across the Cloud Ecosystem:** Due to the proliferation of data, enterprises lack visibility into their data landscape. The lack of confidence in having knowledge of all available data, its location, ownership, and trust level limits an enterprise's ability to fully leverage its own and third-party data to drive business value.

  When migrating to the cloud, sets of data can reside in a number of repositories based on use cases. These repositories can range from transactional systems, such as AWS S3, which may contain curated data to Snowflake or AWS Redshift for data warehousing and analytics. In each of these repositories, data is stored in a unique format from raw or untransformed data in S3 buckets to aggregated form in Snowflake and Redshift tables.

- **Dual Mandate of Data Infrastructure Teams:** Scaling data governance policies to cover the enterprise is a multidimensional issue. In an environment where consumer privacy regulation is becoming stringent, companies are increasingly investing in data governance initiatives. A number of organizations within the enterprise – including the offices of data privacy, security, and governance – are involved in drafting policies to comply with privacy and industry regulations. These policies dictate who can access what data and under what conditions. The burden of implementing these data access control policies falls on the shoulders of the IT organization, specifically the team responsible for building and maintaining data infrastructure across on-premises repositories and cloud services. At the same time, the data infrastructure team is tasked with sharing data broadly within the enterprise, so that data scientists and analysts can extract new insights from it.

  These opposing mandates are difficult for data infrastructure teams to balance, especially when these implementation teams are not well-versed in regulatory compliance and data governance domains.



Data infrastructure teams must balance the dual mandate of complying with regulation and making data widely available to business.

Complex Governance Tools and Processes: Each public cloud provides its own version of access control and management. For AWS it is Identity and Access Management (IAM), for Azure its Azure Active Directory and Azure Role-Based Access Control or (RBAC), and for Google Cloud it is Cloud Identity and Access Management (IAM). Layer on top the access controls offered by cloud-native solutions like Snowflake, Databricks, and many others, and it becomes evident how complicated managing these services can become. For example, if a company's data resides in S3, Snowflake, and Databricks, the data administrator must navigate to three separate interfaces to grant users access to that data, so it can be appropriately shared with the right external or internal audience.

Cloud services' disparate access control mechanisms make data governance untenable.

As data spans across the edge, on-premises, and multiple cloud environments, the surface area of exposed data increases significantly, along with the probability of an administrator making a costly mistake.

- **Multiple Versions and Copies of Data:** It is a fairly common enterprise practice for the same data to be copied multiple times by different groups. These copies are then manipulated and transformed according to the needs of each group. Over time, copies of data actually become the "source of truth" for each of the groups. The consequences of this practice can be severe for an organization if the data contains sensitive or personally identifiable information (PII). If multiple copies of data containing sensitive information disappear into multiple systems, it is virtually impossible for a company to respond to right-to-be-forgotten requests from consumers or secure data by restricting access to authorized personnel.

  Without fine-grained access control in place, administrators do not have visibility into groups, users, or access privileges, resulting in the organization losing control of the integrity of its data, due to excessive copies and transformations.

- **High Touch Required to Onboard New Data Consumers:** Data infrastructure administrators continuously provide new and existing users access to enterprise data when onboarding them to the data platform. Users now expect to be onboarded to multiple on-premises and cloud services without delay and at high concurrency. The onboarding process must be easy for the new user to request access, yet scalable for the administrator. In order for the users to find value in the platform, they must be able to access their data quickly with the required privileges. If the onboarding process is manual or overbearing, users are unlikely to engage with the system long enough to find value.

6 Privacera's Approach

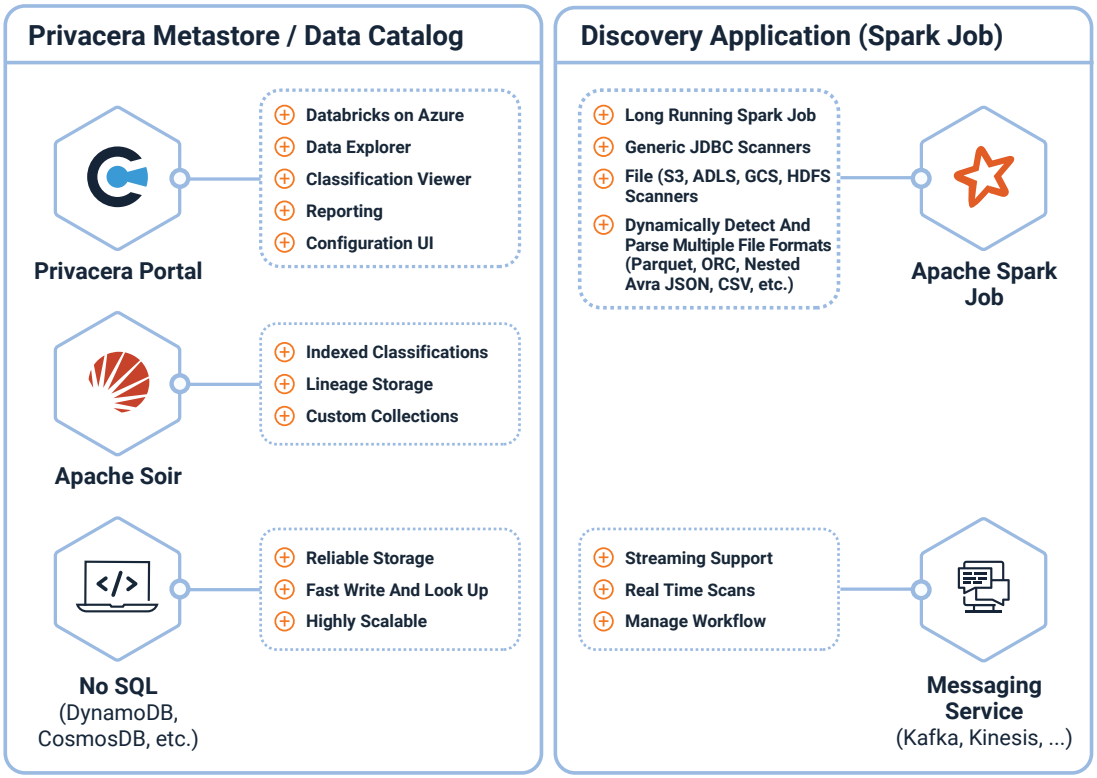
Built by the same team that developed and brought to market Apache Ranger, **Privacera's approach uses Ranger's foundational access control capability proven with big data and Hadoop data lakes, and applies it to cloud services. Essentially, Privacera is a data governance platform built for the cloud.** Privacera's centralized access management empowers administrators to discover, define, and enforce data access policies across heterogeneous cloud and on-premises data services from a single interface.

Privacera's platform is a realization that data discovery, access control, anonymization, and encryption must be centralized in order for enterprises to have effective access governance.

6.1 Data Discovery

Enterprises cannot govern or secure their data if they don't have actionable knowledge of what data exists in which sources, whether it contains any sensitive information, and who is authorized to access it. It is critical to identify and classify sensitive data as it is ingested from the source systems, prior to it being accessed by users. By classifying sensitive data and implementing tags that follow it when it is moved or copied, administrators are enabled to write policies that determine who can and cannot access it.

Privacera's Data Discovery module leverages a number of techniques, including sophisticated rules, pattern matching, dictionaries, algorithms, and machine learning models, to understand the context of sensitive data and accurately classify it. Sensitive data is scanned, identified, and tagged in real-time as it is uploaded to the cloud or object storage. With Privacera's out-of-box reporting, quick generation of custom reports, and alerts when sensitive data is accessed or moved, infrastructure teams have instant visibility of their data assets.

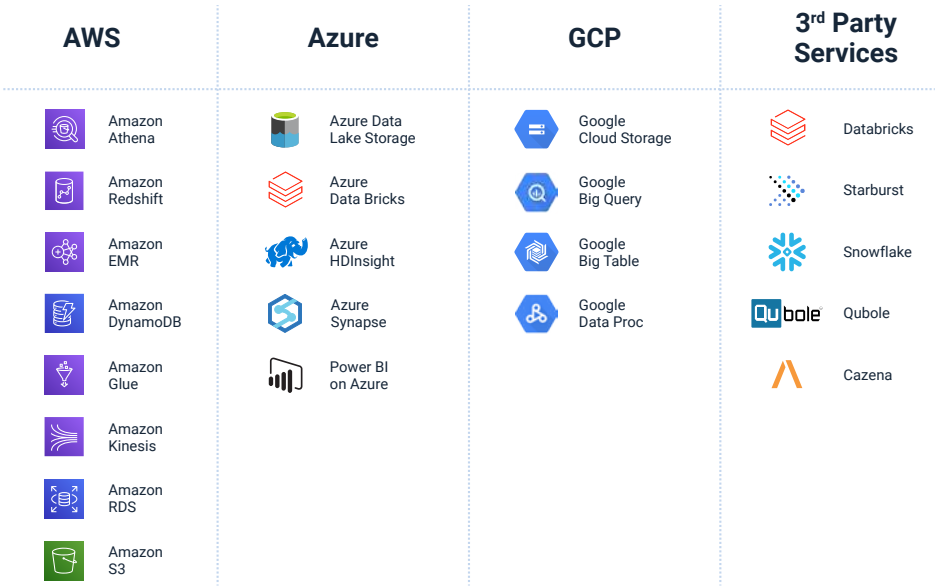


Privacera Data Discovery Components

6.2 Access Control

Once data is classified, the right type of controls must be implemented. These controls are more effective if they are administered from a central location. It is important administrators consider if a single platform can be used to administer access control policies across services. It is extremely important for administrators to consider if a single platform can be used to administer access control policies across services.

The Privacera Platform provides administrators with the familiar Ranger interface to define and administer data access policies for on-premises data lakes, public cloud services, as well as third-party cloud-native services such as Databricks, Snowflake, and others from a single console. Privacera scripts and utilities provide easy configuration of enforcement points across all the cloud and on-premises data services.



Privacera's coverage of various cloud and 3rd party data sources.

The richer the access control platform's ability to administer policies to finer grains of data, the easier it is for infrastructure administrators to grant access to the precise data users need to do their jobs. Privacera's Access Control module provides administrators the flexibility to define access policies at a database, table, column, or file level.

With Privacera, administrators can build access policies based on roles, attributes, and assigned tags. Users from LDAP or AD directories can be associated with specific organizational roles, which can then be assigned access privileges or permissions. Rules based on dynamic conditions, such as time or geography, can also be added to an existing policy rule.

6.3 Data Masking and Row Filtering

It is not enough for a data governance platform to discover sensitive data and apply access control policies to secure it. Data infrastructure teams must also provide mechanisms for data scientists and analysts to extract insights from regulated data. This requires masking sensitive data prior to making it available for analysis and restricting users' access to specific rows based on organizational role or attribute.

The Privacera Platform provides dynamic data masking capabilities via Ranger's intuitive masking policies to protect sensitive content in a variety of flexible formats. This capability enables only authorized users to see data they are permitted to see, while the same data is masked or anonymized for other users or groups. Masking policies can be used to define which specific data fields are masked and how to anonymize or pseudonymize specific data.

Privacera also provides security at the row-level with filtering policies executed as behind-the-scenes query filter conditions that limit the set of displayed rows. These policy-filtering conditions are always in effect and are evaluated against queries to remove the need for security administrators to add filtering predicates manually or create multiple views.

6.4 Encryption

Enterprise data must be protected while it is in motion or at rest. Privacera encryption gateway (PEG) is a robust, scalable application programming interface (API) gateway that protects customers' sensitive data and personally identifiable information, without the need for manual processes and operational burden. PEG provides flexible mapping schemes and policy-based encryption and decryption using NIST standards-based encryption algorithms, such as AES-128, AES-256, hashing, and format preserving encryption (FPE).

7 Privacera's Differentiators

Privacera Differentiators	Customer Benefit
Seamlessly migrate Apache Ranger-based access controls and data governance policies from on-premises repositories to cloud services	⊕ Accelerate data migration and enable analytics ⊕ Onboard users faster and share data broadly ⊕ Comply with privacy and data governance regulations on day 1 ⊕ Leverage new cloud services with the same governance and security as on-premises
Automated data discovery without compromising data security before or after migration	⊕ Ensure sensitive data is classified and inventoried before and after it moves to the cloud ⊕ Build fine-grained policies based on where sensitive data is stored ⊕ Ensure compliance with privacy and industry regulations
Anonymize and encrypt data when moving from on-premises systems to cloud storage	⊕ Move all data, including sensitive data, easily to the cloud ⊕ Enable privacy and data protection when data lands in the cloud ⊕ Leverage sensitive data for analysis by masking or filtering sensitive data fields
Automate workflows to move, quarantine, and anonymize data	⊕ Build automated policies to enable compliance with privacy and other mandates ⊕ Reduce manual overhead ⊕ Accelerate time-to-value for analytics ⊕ Ensure compliance with privacy and industry regulations