REPORT REPRINT

Privacera tackles elephant-sized privacy challenges by extending key Apache projects

FEBRUARY 18 2020

By Paige Bartley, Rachel Dunning

Open source can be a valuable foundation for security and privacy of data yet open source projects can be limited in scope. Cloud-centric Privacera looks to solve this by extending the Apache Ranger and Apache Atlas projects with IP geared toward solving privacy and compliance use cases across a variety of data environments.

451 Research®

Introduction

With regulations such as GDPR and CCPA emerging and evolving at a rapid pace, organizations are struggling to take control of their large volumes of distributed data so that they can fulfill basic requirements such as retrieval of personal information and management of data access permissions. Many data platform providers offer robust data governance layers that can help address these needs; however, these capabilities are typically constrained to the vendor's specific environment. Yet organizations today have IT environments that are more diversified than ever. According to 451 Research's Voice of the Enterprise: Cloud, Hosting and Managed Services, Workloads and Key Projects – Quarterly Advisory Report from June 2019, 72% of organizations using the public cloud have more than one vendor in place; 8% have more than three. With data everywhere, there's a strong enterprise desire for a 'single pane of glass' that can provide strong, compliant privacy controls regardless of the system or application in which data resides natively.

Privacera is looking to extend the existing strengths of open source security and governance projects – Apache Ranger and Apache Atlas -- to provide organizations with a single view that provides data discovery and classification, centralized access management and anonymization/masking regardless of an organization's existing IT investments.

451 TAKE

Conceptually, open source works. Crowdsourcing has a multiplicative effect on innovation and can quickly weed out flaws in code. But human psychology is also at play. Even the humblest contributor desires some recognition, so development of less glamorous functionality – such as security and governance – often gets less attention. For open source projects that are highly active, lopsided contribution from large vendors can potentially skew the direction. There is a time and a place for unique IP to extend open source functionality; the realm of data privacy may be a good example. Today's IT ecosystems are more complex than ever, and as major data management vendors continually seek to broaden their platforms, organizations are becoming frustrated with data governance and privacy layers within these products that can exert control on a portion (but not all) of their data. Privacera's value proposition – taking the best parts of open source projects, extending them even further for privacy use cases and consolidating the capabilities into a single view – will likely resonate well with organizations that are struggling to maintain control of highly heterogeneous hybrid and multicloud data environments.

Context

Privacera was founded in 2016 by CEO Balaji Ganesan and CTO Don Bosco Durai to essentially extend the functionality of open source data governance and security frameworks Apache Ranger and Apache Atlas into the modern cloud domain. Ganesan and Durai had previously founded the data security company XA Secure in 2012, which offered centralized data governance, security and access management tools for the Hadoop stack. Once XA Secure had been acquired by Hortonworks in 2014, its technology assets and tooling were donated to the Apache Software Foundation and incubated as the open source project Apache Ranger, perhaps now best known for its granular data access and auditing through a console to manage policies and set up user entitlements. At Hortonworks, Ganesan was also closely involved with developing another project, Apache Atlas, to provide the scalable

metadata management framework that would help organizations meet requirements around data governance, lineage and regulatory compliance. These building blocks became foundational to the Privacera platform, which became generally available in 2017.

In addition to extending organizations' abilities to maintain GDPR, CCPA and related data privacy compliance requirements, Privacera assists organizations in migrating from Hadoop to cloud-native services including Google Dataproc, Big Query, Big Table, Amazon EMR, AWS S3 Buckets, Amazon Redshift, AWS Dynamo DB, AWS Glue, AWS Athena, Snowflake, ADLS, Azure Synapse, Azure Databricks, Azure HD Insights and Cosmos DB. Facilitated data democratization is another major aspect and benefit of Privacera's platform, particularly for customers operating in highly regulated industries: better insight into sensitive data assets – and more scalable control for access rights – allows general business users to go about their daily work without cumbersome permissioning workflows.

Privacera's headquarters are in Fremont, California, and the company is still fairly early in its growth trajectory; by the end of 2019, it had roughly 50 employees, and had raised $4m in early series A funding. For 2020, the company is exploring the possibility of additional funding, given the fervor and activity of the privacy software market, a trend we discussed in our 2020 Tech M&A Outlook: Application software report.

Technology

Architecturally, Privacera sits on top of Apache Ranger and provides an Apache Atlas compatible layer, essentially extending them and adding additional granularity via IP for environments such as the cloud that those open source projects were not designed to handle natively. Elastic infrastructure means compute can be spun up and down as needed, with Kubernetes-based and Dockerized deployments. Centralized access management capabilities for data are the company's core competency; however, data discovery and classification at cloud scale, as well as flexible anonymization and masking, are also key strengths. The capabilities are currently packaged in three module-like offerings that comprise the Privacera platform, which can be deployed equally well on-premises or in the cloud; for example, on bare metal, VMs or containers in private cloud/datacenters or using cloud-native services from AWS, Google Cloud Platform (GCP) or Microsoft Azure.

Discovery and classification

Discovery and classification capabilities span 100+ structured and unstructured data formats from a variety of cloud and on-premises data sources out of the box, with the ability to scan cloud RDBMS, NoSQL, data warehouses, object stores and file system storage. Rather than just scanning metadata for sensitive data-detection purposes, the data content is examined as well. Discovery of sensitive data is both pattern- and dictionary-based – catching all common sensitive data types such as SSN – as well as through machine learning and NLP-based techniques. Sensitivity can be adjusted for detection and classification algorithms, proximity-based matching is supported and additional libraries may be used to expand recognized data types. Because workflow support is built into the module, data stewards and other stakeholders can appropriately coordinate when manual review or override of classification is needed. Binding the module is a sensitive data catalog based on Apache Atlas, which catalogs data classifications using a metastore; additionally, APIs allow the sensitive data catalog to be integrated with existing enterprise data catalog investments. The Privacera catalog provides visibility into data use, allowing for various workflow actions such as blocking/alerting if data classified to configurable 'zones' is moved elsewhere or encrypting, tokenizing, quarantining, stripping, redacting or masking sensitive data as it moves across such zones.

Centralized access management

Privacera's centralized data access management capabilities extend the Apache Ranger project's native ability to administer data access permissions in Hadoop to a diverse number of other environments, including the cloud. Full integration with various IAM systems from cloud vendors and via standardized authentication protocols (e.g., SAML, Oauth, OpenID, SCIM) are fully supported as well as integration with user repositories and directory services via LDAP and Active Directory allows existing user roles to be mapped into Privacera, minimizing duplicative work. Data access policies can be set based on a variety of parameters: resource/data type, role of user, classification of metadata or other properties of the data. With a central audit console provided for all data that has had access policies applied to it, the organization can get a 'single pane of glass' view that exposes every access attempt and other relevant action along with rich event metadata across the data ecosystem. Architecturally, the Centralized Access Management module uses one of three primary mechanisms for connecting to data locations: a plug-in is used for Apache Hive and Spark as well as some commercial providers such as Databricks, Privacera Data Access Server acts as a lightweight and distributed proxy for many other environments and Privacera Policy Sync can push down directly to databases and interact natively with infrastructure. The sum of these capabilities allows organizations to centrally control data access permissions, regardless of where data resides, with a single portal.

Anonymization and masking

Privacera again extends the capabilities of Apache Ranger with its Anonymization and Masking module, not only by expanding its control far beyond Hadoop, but also by making its controls more granular; unlike Apache Ranger's native capabilities, Privacera enables encryption at the column level. Encryption capabilities can be offered via API, allowing them to be called within existing pipelines and key management for encryption is integrated with popular cloud-based key vaults, allowing organizations additional flexibility and centralization of security control. The concept of 'schemes' for Privacera's anonymization and masking allow organizations to essentially set detailed rules and triggers for how and when certain types of data are masked or protected. A variety of masking options such as regex-based, partial reveal/blur, redaction, hashing and tokenization are provided out of the box. Furthermore, masking schemes can be extended by adding any custom logic via functions and a variety of standardized encryption schemes such as NIST standard AES and FPE (Format Preserving Encryption) are available to enable organizations to protect the privacy of sensitive data while still conducting meaningful analysis for other purposes.

Competition

Being fundamentally based on Apache open source projects, Privacera will naturally face some competition from DIY enthusiast organizations that like to tackle the challenge of assembling open source into home-grown systems. Given Privacera's Hortonworks heritage, the company is not unlikely to face Cloudera, particularly Cloudera's SDX governance layer, which leverages acquired Hortonworks assets in addition to the Apache Ranger and Apache Atlas projects. It should be noted, however, that Privacera's approach is designed to alleviate some of Cloudera's platform-specific constraints and provide complementary extensions that are designed to work natively for access control of cloud data sources.

Given Privacera's core competency in managing data access permissions and rules, the company faces competition from the data access governance segment. Some of the key providers in this space include Immuta, Netwrix, Okera, PlainID and STEALTHbits. A major incumbent in this space would be Varonis, which has notable strength and expertise in the access governance and security of unstructured data sources. Some new emergent players in unified data access governance – seeking to provide a single pane of glass for access control across systems – would include Cyral and Satori Cyber. IDaaS vendors, such as Okta, can also nominally overlap into this space.

Data discovery and data classification category vendors frequently overlap with each other, as detailed most recently in the Data Security Market Map 2018. Mainstay vendors in the data discovery category include Dataguise, Digital Guardian, PKWare, Spirion (formerly Identity Finder), TITUS and Varonis in addition to many household names such as AWS, CA Technologies, Google, IBM, Microsoft and Symantec. Vendors that specialize in data classification, in addition to most of the aforementioned providers, include Boldon James. Some of the notable innovators in these segments include those that can leverage machine learning or automated technology to detect and classify or associate sensitive data with identities or entities for privacy use cases. Examples include 1Touch.io, BigID, Integris Software, Io-Tahoe, NetApp (via the Cognigo acquisition) and SECURITI.ai.

Many products have embedded pseudonymization capabilities that are catered toward privacy use cases. Major providers such as IBM, Informatica and Oracle use integration across their product portfolios to enable capabilities such as data masking and obfuscation. Many of the providers mentioned in the data access governance space, such as Immuta, offer ways to modulate sensitive data so that it may still be used without compromising privacy. Privitar also offers de-identification of data. Homomorphic encryption providers such as Duality Technologies allow data to be analyzed while encrypted, not revealing its details to those interpreting analytics results. Synthetic dataset generation using AI is another area of innovation that allows organizations to take private datasets and generate comparable non-private datasets with nearly identical statistical properties; upstarts like Mostly AI play in this space. Others that specialize in the anonymized analytics of personal data include Cryptonumerics and Very Good Security.

Others in the PrivacyOps market category as outlined by the Data Management Market Map 2019, which were not already mentioned above, compete to some extent and include DataGrail, Infosys, OneTrust and TrustArc (which recently acquired competitor Nymity).

SWOT Analysis

STRENGTHS

Unlike its pure proprietary competitors, Privacera is extending the strengths of existing open source projects, giving an element of transparency to the company's technology and direction. Competency with big data should not be taken for granted, given the heritage of the Apache Ranger and Atlas technology. Because the product is agnostic and cloud-enabled, it aims to give a single pane of glass experience for data management across diverse IT ecosystems.

WEAKNESSES

Privacera's core competency is in data access governance for privacy use cases, although the breadth of its capabilities today means that it is targeting multiple stakeholders within a given organization, trying to be a little bit of everything to everyone. To date, messaging and positioning need to be tightened somewhat so that Privacera can more accurately convey a business value proposition more specific than 'governance and privacy capabilities for hybrid/cloud environments.'

OPPORTUNITIES

Regulations continue to intensify. Additionally, data ecosystems within organizations are only getting more heterogeneous. Yet there is already consolidation occurring in the PrivacyOps space as large data management vendors make acquisitions to flesh out governance and privacy capabilities. If Privacera can maintain its neutrality and agnosticism, it will be an attractive choice for organizations that struggle to control their data outside of major platforms.

THREATS

The flip side of the opportunity listed here is that Privacera itself may eventually become a target for acquisition by a large data management provider; a development that could potentially undermine its role as a neutral, binding layer for privacy governance and management of data across diverse data locations. And because the company overlaps into so many existing segments, incumbent vendors in those markets have every reason to set their sights on Privacera as a competitive threat.