satori

Data Protection and Governance for Data-Driven Organizations

Data is at the center of everything. From business intelligence to advanced analytics, across every sector of business, the creation, storage, and use of data is the engine that drives today's competitive advantage and innovation. Much of this has been possible thanks to leaps in cloud data storage and access technologies allowing companies to make data a strategic asset.

As a result, data has been democratized, and the number of employees, applications, and third parties actively engaging with data is rapidly increasing. At the same time concerns over data privacy have moved regulation forward, forcing organizations to handle our data in a secure, respectful, and compliant way. However, existing methods and tools do not yet provide the requisite level of visibility and control. They also do not scale to meet the demands of today's use-cases.

We propose a new approach - one that relies on complete data flow visibility, policy enforcement, and rich data access context, and that can scale to meet both today's and tomorrow's needs.

Changes in Data Access, Technology, and Regulation Bring New Challenges to Security and Privacy Teams

Evolution of Data Access

In 1970, Dr. Edgar Codd published the research paper that would introduce SEQUEL, the Structured English Query Language [1]. Later abbreviated to SQL (Structured Query Language), the language has become the de-facto standard for communicating with a relational database management system. While SQL has endured the test of time, as evident by the many tools and platforms that still support it 50 years later, the way we consume business information has changed significantly.

In the past, businesses relied on analysts who were proficient with SQL to query information directly from a database. They would achieve this using their client-server interfaces and authenticating their access using the database's user management system. However, as the demand for data grew within organizations, more complex clients were invented to make the analyst's job easier. This largely entailed abstracting away SQL and database connectivity using graphical user interfaces.

As these clients became more widely used in organizations, it became easier to provision at scale by deploying them as web applications. As enterprises migrated to the cloud, these deployments turned into delivery as a service. The shift from a few single-user desktop clients accessing data in the organization to a large number of users using web-based or even mobile applications has made it more difficult for organizations to provision users using the database's user management system. Instead of provisioning every user in both the database and the application layers, organizations have shifted authentication towards applications and started using service accounts to connect applications to databases.

Service Accounts are Over-Privileged by Design

A service account is a special type of identity that applications use to access resources on behalf of their users. A service account does not represent any particular user, it represents any user that requires access to the resource. In the case of data stores , that 1 means a service account typically has access to all of the data in the data store.

This leads to two significant consequences. First, security controls for data access have moved from the data store layer to the application layer, meaning that the responsibility for building, maintaining, and monitoring security controls has moved from security teams to engineering teams.

Second, data has become much more exposed - either by security vulnerabilities in applications, such as bugs in their security controls or SQL injections, or by using the service account's credentials and connecting to the data store directly, circumventing the application altogether. Regardless of the attack vector, we have left the gate to our most sensitive information largely unmanned.

From Data Warehouses to Big Data and Data Lakes

By 2010, business intelligence was widely adopted in most business sectors and serviced by a mature portfolio of BI and Analytics solutions. The emergence of the web, however, has generated demand for a more advanced framework for handling large data sets. In a 2004 research paper, Google described its MapReduce framework [2] and, not long thereafter, Hadoop, its open-source equivalent became very popular with companies operating at web-scale.

1 Databases, data warehouses, data lakes and other systems like caches, search engines and message queues, to name a few

The cloud has enabled organizations to build large computing and storage infrastructure more easily than before, and new pricing models for operating large data repositories in the cloud are now contingent on the queries they run, not the amount of data stored. Consequently, a path of least resistance was formed for almost any company with an online business to build a petabyte-scale datastore. Combined with greater demand from sales, marketing, customer success, and data-driven strategies, data is now largely accessible to many internal stakeholders; ensuring that access is granted to those who can be trusted with it in a secure, compliant,and effective way poses great challenges to security teams and cloud architects.

New Legislation Challenges Organizations to Rethink their Data Strategies

With digital transformation, businesses are moving online, and staggering amounts of personal information are being generated, stored, processed, and exchanged on a daily basis. This has not occurred without a number of significant high-profile breaches and cases of mismanagement, and a 2017 survey shows that many Americans lack trust in the institutions that are supposed to secure that data [3].

Fueled by concerns over the security and privacy of personal and generally sensitive information, new legislation has been introduced in many jurisdictions, such as GDPR in the European Union and CCPA in California. Despite only coming into force in 2018 and 2019 respectively, both have already become established standards in the security and privacy industry.

With substantial fines and clear definitions of the rights customers have with respect to their data, as well as the requirements organizations must adhere to when collecting, storing, and using that data, these new legislations have significantly changed the way organizations think about data security, privacy, and governance.

Existing Methods for Protecting Sensitive Information

The challenge of protecting sensitive information is not new and solutions that address various aspects of the problem do exist. However, each solution addresses only parts of the problem and a holistic approach to securing sensitive information is required.

Data Cataloging and Classification

Most data governance initiatives start with trying to understand where data resides in the organization and what type of data is being generated, processed, stored, and read. In most cases, this process requires getting all stakeholders to cooperate and share the information they know to build a map of all data flows. This includes who is accessing data, what kind of data they are accessing, and where that data is stored. For large organizations, this in and of itself poses a huge challenge, as team members are distributed geographically and across different time zones. More often than not, these initiatives are slow to start and often fail midway.

Another impediment to this approach's success is that data is a moving target - especially in cloud environments where producing new data stores is fast and easy and traditional IT governance is less effective. By the time these initiatives generate results, the context and circumstances are likely to have changed, and the organization is likely to unknowingly possess even more sensitive information.

Finally, even after gaining knowledge of where sensitive information is, organizations still struggle to make that information actionable without correlating it with the actual access patterns that show whether or not sensitive data has been exposed. Relying on Data Loss Prevention (DLP) solutions to provide that context is often too little and too late, as there are many methods to exfiltrate data once exposed.

Access Control and Permissions Management

Once organizations have an idea of what sensitive information they possess and where it is, it makes sense to establish guardrails and perimeters to limit access to only those who need it. While there is an abundance of tools that help organizations manage access to resources, such as data stores, they are not data-aware. Attempts to define access control on specific parts of a data store's schema are extremely challenging: semi-structured and unstructured data stores do not have schemas and, where schema-based data stores are concerned, the process of managing fine-grained permissions at the table and column level for each user for each use-case is an insurmountable challenge at scale.

Masking, Encryption, and Tokenization

Some organizations follow a strategy of duplicating sensitive data and applying various techniques to de-risk it at rest, such as masking, encryption, and tokenization - f or example, making a copy of the production database available for a development team for debugging while masking any sensitive information in it.

While effective for specific use-cases, that approach fails at scale due to a lack of agility and the overhead it incurs. Creating a copy of the data for every use-case, applying the required transformations to protect it, and granting access to it is a slow process. Whenever a new field is required in the cloned data store, the duplication process needs to be adjusted and re-run. Moreover, this approach by design creates more data, which leads to more risk, larger operational overhead, and increased infrastructure costs.

Built piece by piece over the last few decades, existing strategies for securing sensitive information are ineffective and inefficient. They result in significant operational overhead and only solve parts of the problem. Given that organizations are adopting new data store technologies in the cloud and handling more data than ever before under increasingly strict privacy regulations, a new approach is in order.

A New Approach for Data Protection and Governance

Comprehensive data protection and governance solutions need to answer the following fundamental questions: Where is the data? Which data is sensitive? Where is it? Who is accessing the data? What are they doing with it?

When thinking about a better approach for securing sensitive information, a successful solution needs to be able to cope with a number of key challenges.

Support for New Data Technologies

As development teams adopt new data stores at an ever-growing pace, a data protection and governance solution must be able to support new technologies easily. Having solutions that are only relevant to some elements of the ecosystem, or that are not a good fit in a multi-cloud environment, guarantees that organizations will have to invest more in the future to implement other solutions to address these gaps - or risk remaining exposed.

Deployment in Existing Environments

Most organizations are not looking to re-architect their data infrastructure for the mere sake of adopting a security solution. Solutions that deliver security value by replacing parts of data infrastructure, like storage or query engines, or rely on the extensive deployment of application or endpoint agents, would be hard to implement and difficult to appreciate.

Protecting Data at its Source

Instead of using static data transformation, for example, replacing all email addresses in a data set with masked ones , solutions can utilize dynamic data transformations to protect sensitive data at its source, for example, by masking email addresses when they are being queried by a user that should not be exposed to Personally Identifiable Information (PII).

This ensures that organizations only store the data they need, do not increase infrastructure costs and operational complexity, and allow a more agile approach to providing access to data by configuring who can access it instead of running offline duplication processes.

Activity Aware

It is not enough to know where your sensitive data is and how its distribution changes over time. Without being able to tell who is accessing it, and enforce policies that basis, organizations do not have enough actionable information to respond and remediate incidents.

Take, for example, the case of a data warehouse that was loaded with sensitive information by mistake due to an ETL process change or a new version of the application. Without leveraging the actual activity of data access and answering the question of whether anyone may have been exposed to sensitive data, organizations do not have a complete view of what happened. By the time that data has been exfiltrated or more broadly shared, there is not much to do in the way of minimizing the damage.

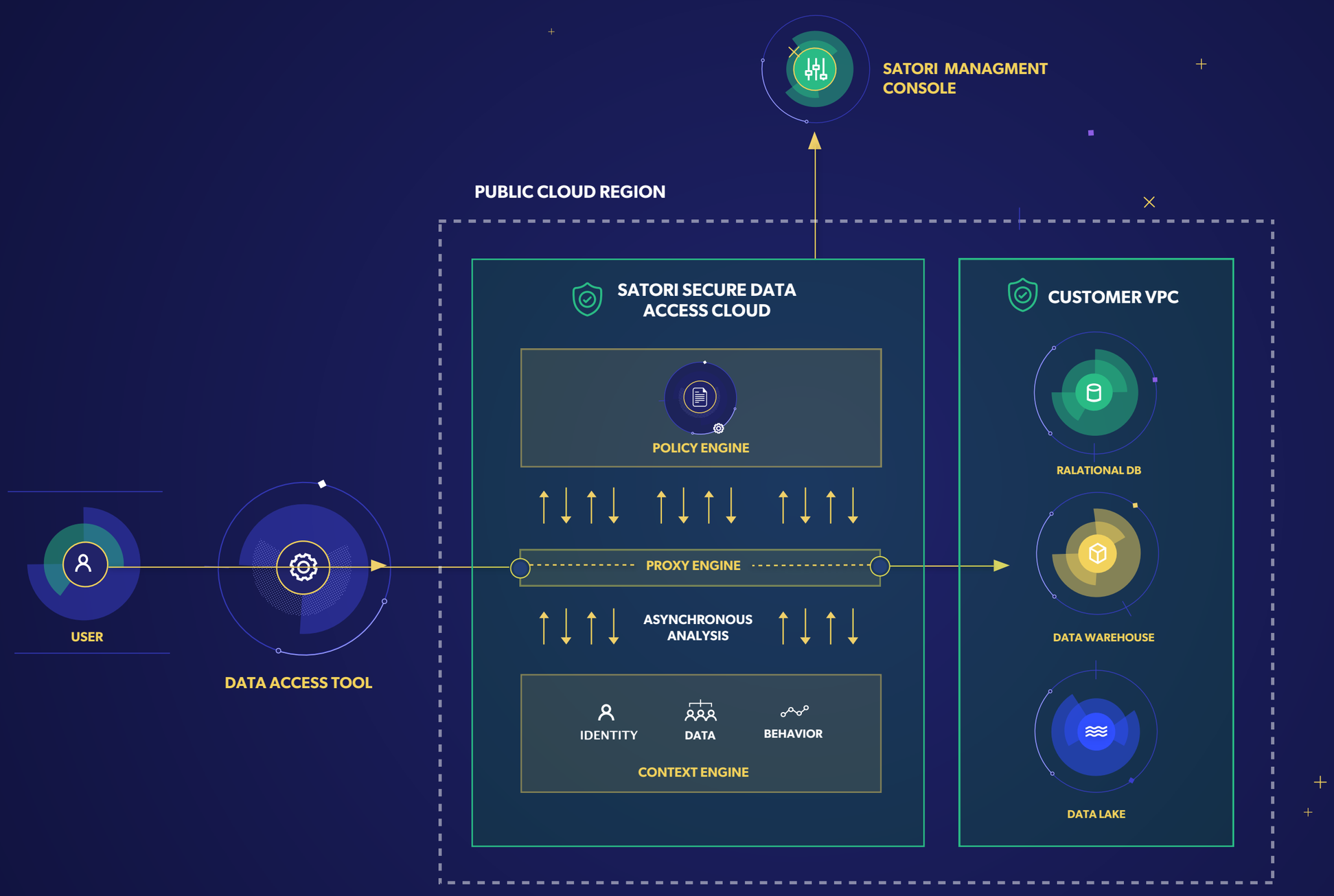
Delivering More than just Security Value

Solutions that are able to deliver value to other stakeholders in the organization, as well as the security team, stand a better chance of adoption, because implementing a solution requires coordination and approval from more than just the CISO.

Key capabilities helping development teams manage database loads today include: terminating queries that are causing the database to grind to a halt; distributing queries to different data stores instances based on rules that can be updated dynamically or triggering a hook when certain fields in a data set change; or calling a serverless compute function so that engineering teams can debug systems more easily or trigger a workflow.

Secure Data Access Cloud

Satori presents a new approach that provides continuous visibility into data and data flows, implement requisite security controls, and enforce compliance and privacy policies while making legitimate access easy, fast, and efficient. It is a single platform that can fully understand, manage, and secure data access and data usage consistently across all current and future cloud and legacy data stores. Satori achieves this by combining user and application identity with real-time data discovery, classification, and behavior analysis, as well as providing full visibility into data stores and usage while enforcing data access security, privacy, and compliance policies.



About Satori Cyber

Satori Cyber is revolutionizing data protection and governance. It's Secure Data Access Cloud seamlessly integrates into any environment to deliver complete data-flow visibility utilizing activity-based discovery and classification. The platform provides context-aware and granular data access and privacy policies across all enterprise cloud or hybrid data stores. With Satori Cyber, organizations and their security teams can confidently ensure that data security, privacy and compliance are in place, enabling data-driven innovation and competitive advantage.

References

[1] History of SQL https://docs.oracle.com/cd/B12037_01/server.101/b10759/intro001.htm

[2] Google MapReduce research paper: https://ai.google/research/pubs/pub62

[3] https://www.pewinternet.org/2017/01/26/1-americans-experiences-with-data-security/