Privacera and Starburst

How To Securely Accelerate Data Science Initiatives, While Balancing Governance And Compliance

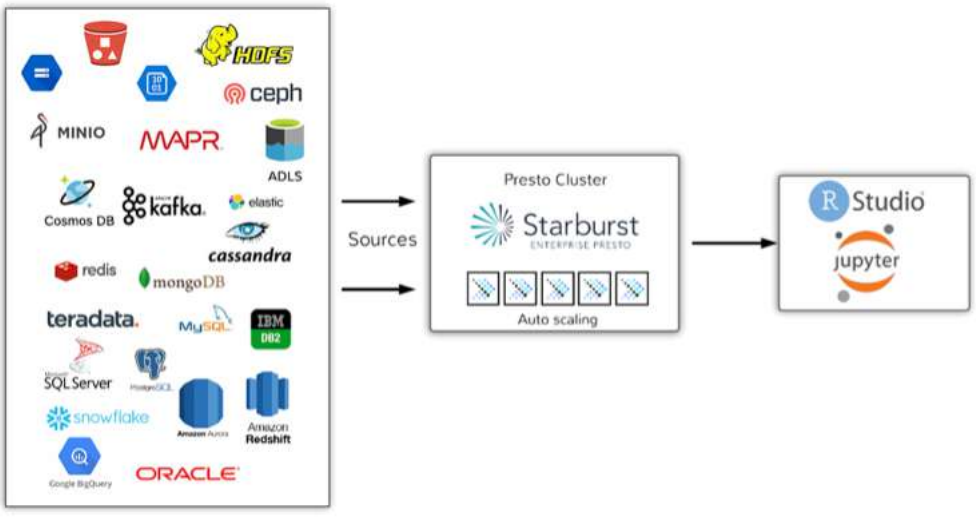
Whitepaper
February 2021

Content

Before the explosion of data and the rise of cloud services, the Internet of Things (IoT), and numerous other new sources of data, data scientists could build their machine learning models by querying data from a single, centralized repository, such as a data lake or data warehouse. But today's modern computing world is undergoing a massive digital transformation; the volume of data is increasing exponentially, resulting in data distributed across various sources that makes it virtually impossible for data to be brought into a single repository to be analyzed by data scientists and analysts.

Due to this inherently distributed nature of modern data, data scientists and analysts find themselves spending 50% of their time wrangling, loading, and cleansing data – which can stifle data science initiatives, frustrate data teams, and negatively impact business operations[1].

With technology like Starburst Enteprise and its rapid federated query engine, data scientists and analysts can query data in distributed datasets without moving all of it to a central location–enabling minimal data movement, fast search results, better quality of data, and less manual burdens for data teams.



[1] Source: 2020 State of Data Science, Anaconda, https://know.anaconda.com/rs/387-XNW-688/images/Anaconda-SODS-Report-2020-Final.pdf

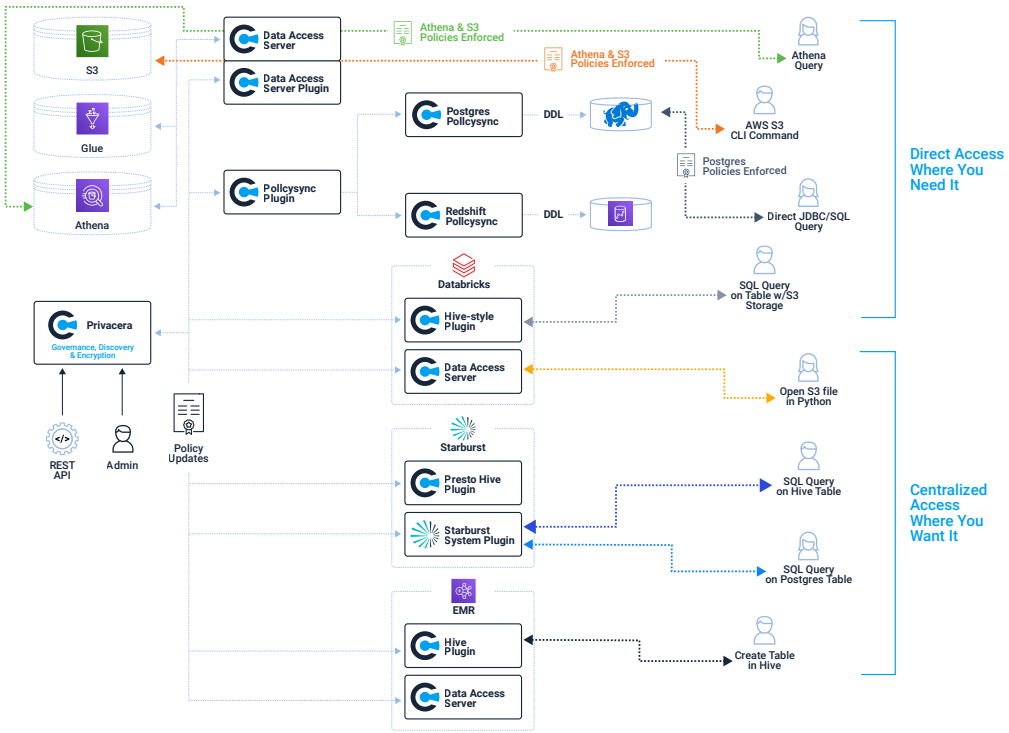
However, with the benefits of rapid federated analytics comes the additional complexity of ensuring that users' access to data fully respects the privacy, governance, risk and compliance constraints required by industry regulations. To truly empower data teams and securely democratize data, while still remaining compliant, enterprises need a solution for consistent access controls and governance policies across cloud data services from a single user interface.

The Privacera-Starburst integration balances this dual mandate by enabling enterprises to implement fine-grained access controls across their federated query environments from a single location with consistent security and privacy policies to protect sensitive data from unauthorized access. Data teams are empowered with fast access to data, with comprehensive auditing and reporting capabilities to meet privacy and industry regulations like CCPA, HIPAA, LGPD, GDPR, and more.

1 Privacera-Starburst Architecture Overview

Privacera and Starburst Enterprise are integrated via an Apache Ranger plug-in (a lightweight component embedded natively) that provides data teams the trusted capabilities of Apache Ranger across their cloud-first data platforms and on-premises data repositories. This plug-in validates users' permissions to access specific data, while enabling data transparency across users.

Data teams can query multiple data sources across both their on-premises and cloud infrastructures – enabling immediate analysis of siloed data without costly data warehouse appliances, significantly reducing the need to move or copy data, and streamlining access to data for rapid analysis and quicker time-to-value. With Privacera Platform's automated governance capabilities, data is protected whether at rest or in motion, and governance policies are immediately enforced.

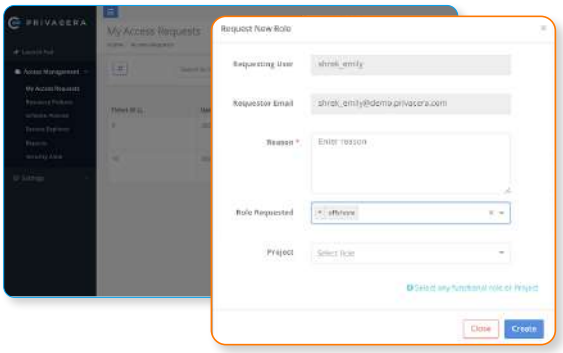


Starburst clusters are configured to use policies managed by Privacera. If Starburst is already integrated with Apache Ranger, existing policies can be easily migrated. Privacera's tag-based policies can automatically protect sensitive elements in data against inadvertent exposure. Policies in Privacera can be extended beyond Starburst to protect data even when it is accessed outside the distributed query environment, such as BI queries on data warehouses, or direct access to cloud storage like AWS S3, Azure ADLS, or Google Cloud Storage.

2 Key Benefits for Data Teams

Fast, Secure Access To Data With Fine-Grained Access Control

With fine-grained access control, each data resource has its own access control policy, enabling data with different access requirements to coexist in the same storage or analytics environment. With native support for Starburst, Privacera's Platform allows data administrators to create role-, attribute-, and tag-based policies to control data access at the file, row, and column level–as well as implement dynamic data masking and filtering – enabling data scientists and analysts to query data rapidly while still protecting against authorized access (e.g., sensitive data or personally identifiable information). With the safeguards fine-grained access controls provide, data teams can:

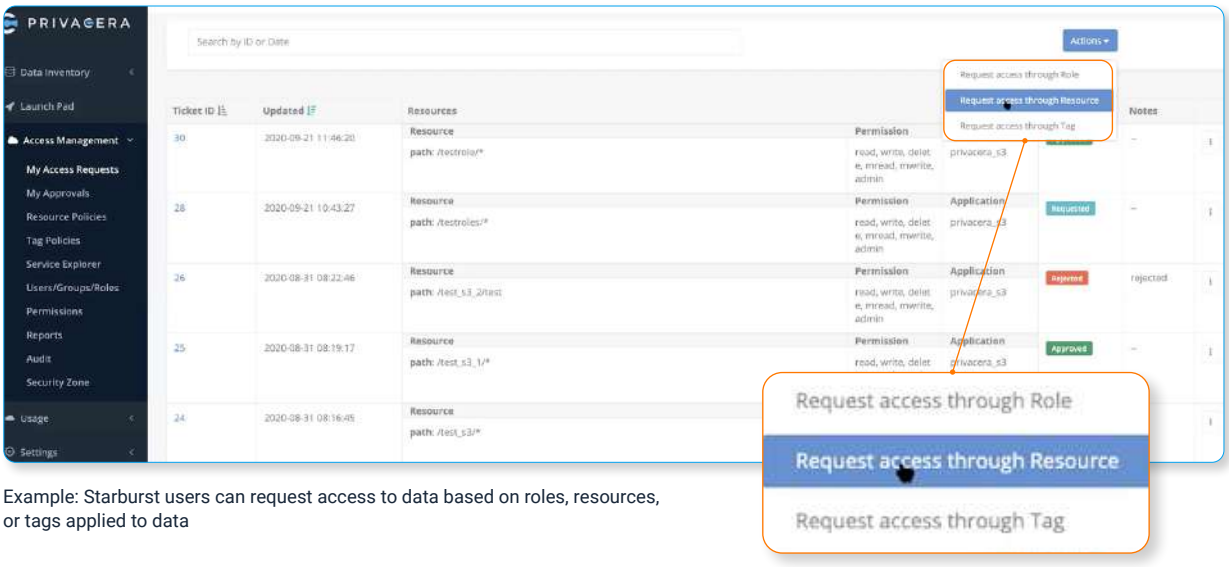


Example: Starburst users can select functional roles or projects

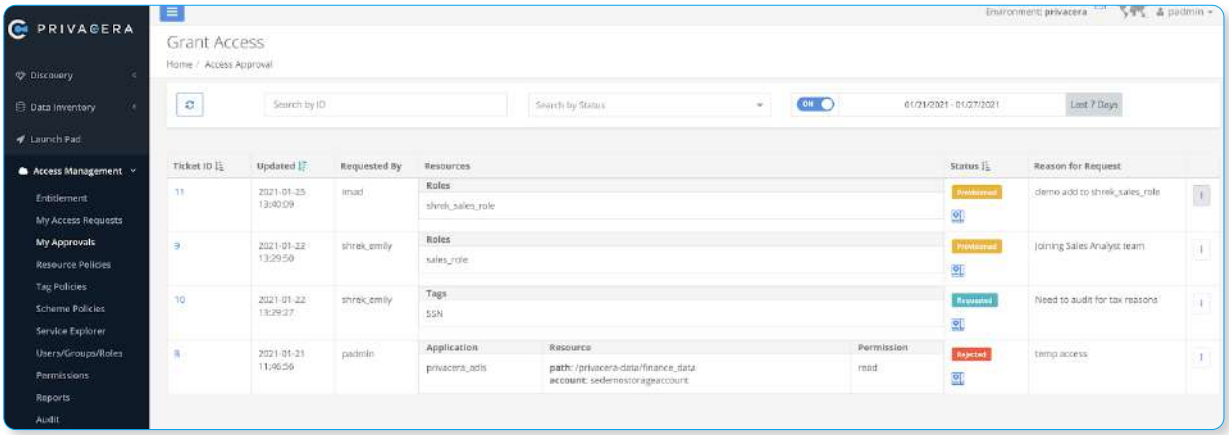
- Quickly access available datasets they have permissions for, alleviating the time-consuming process of manually requesting access from each separate data owner (self-service analytics)
- Deliver high-quality insights and analytics faster with the ability to access any portion of data, no matter where it resides

Additionally, with Privacera's Access Workflows, data teams using Starburst can further accelerate and customize their data access by enabling bulk access requests based on their functional roles, projects, data sharing agreements, or engagements. Data teams can also request access to specific sets of data resources (e.g., "à la carte" data) or classifications-enabling:

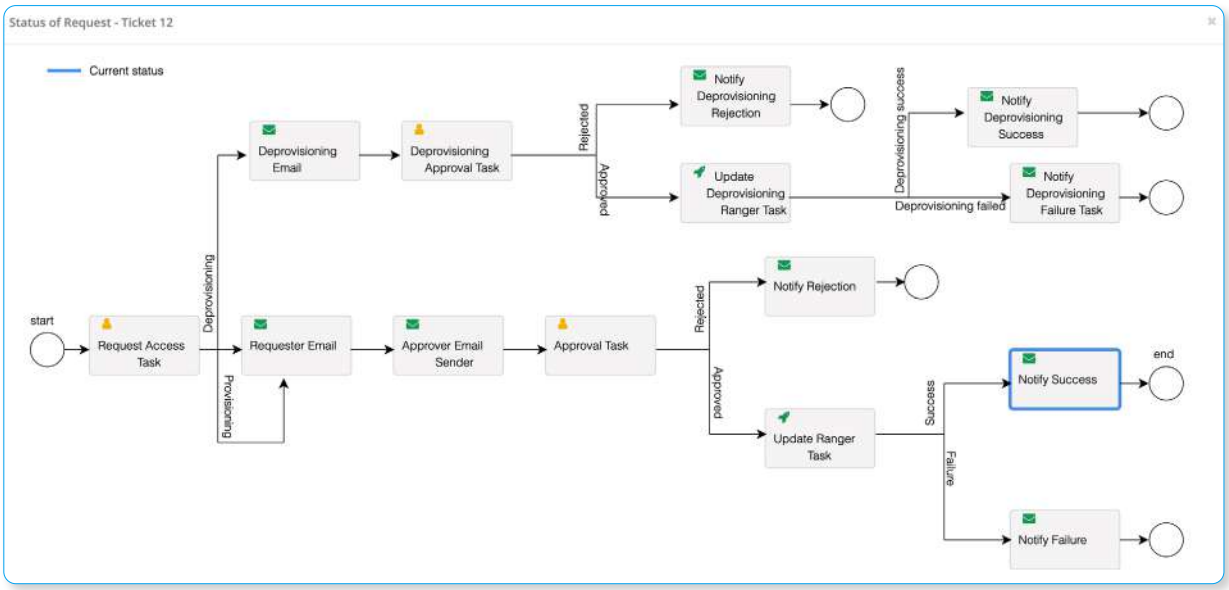
- Improved cross-functional collaboration for project-based or time-bound assignments
- Decreased onboarding times for new data users
- Faster approvals for requested data access
- Less manual access request processes



Example: Starburst users can request access to data based on roles, resources, or tags applied to data



Example: Data administrators can grant access approvals with a single click



Example: Data Access Request Approval Process

Less Time Copying And Pasting Data, More Time Analyzing It With Automated Governance And Compliance

Data compliance is an integral part of a data-driven enterprise and a key component in analytics, especially as industry regulations continue to become more stringent and expand into new domains (e.g., Brazil's newest regulation rolled out in 2020, LGPD). If compliance is not managed well, data teams may, for fear of regulatory actions, be blocked from legitimate research, while simultaneously (and invisibly) creating an increased risk, due to poor visibility and inconsistent policy enforcement. As a result, enterprises risk incurring hefty monthly penalties, security breaches, legal ramifications, revenue loss, or losing the trust of customers.

To ensure data compliance and enable data teams to maximize access to and quality of data, the Privacera Platform can migrate existing on-premises policies to public cloud services and enforce them immediately in Starburst, reducing manual burdens for data and compliance teams by alleviating:

- Rewriting of policies from scratch
- Copying and pasting data
- Inconsistencies from on-premise data sources to cloud data sources
- Risks of non-compliance with regulations like CCPA, GDPR, LGPD, HIPAA, and more
- Manually tracking processes for the right to be forgotten, right to erasure, and right to access

Apache Ranger already brings a complete audit picture to Starburst data access by collecting and collating data access attempts, login sessions, plugin state, and a complete history of policy changes.

Ranger provides a central audit location for all access requests across all services, and its comprehensive audits framework provides rich event data, along with contextual metadata such as data classifications of the resources accessed, IP addresses, locales, specific policies, and versions that granted or denied access requests. Ranger also provides users with the flexibility to leverage raw event data to do additional post-processing and visualization.

Privacera extends Ranger's capabilities with the addition of access and administrative audits, as well as policy states from directly-accessed cloud native services, completing the picture for compliance and legal teams. Privacera also provides:

- Automatic scanning for sensitive data
- Access controls for direct access to cloud resources
- Synchronization with LDAP, Active Directory, Azure AD, and other sources of user and group data
- Enterprise-grade encryption
- Centralized audit, even for direct access to cloud resources

True Democratization Of Secure Data

Enterprises cannot afford to have a bottleneck at the gateway to their data; secure democratization is critical to enable data scientists and analysts to make faster decisions, build more agile teams, and gain greater competitive advantage over slower, data-limiting enterprises. Privacera and Starburst give enterprises the ability to break down data silos–the first step in empowering data teams–and centralize access and compliance, so data teams can spend time on what truly matters: getting fast, secure access to data to deliver as much value as possible to their enterprises and customers.

The push for data democratization is not the future; it is already here, and it is critical to successfully managing data and realizing its value. Enterprises that recognize this and adopt it early into their mission-critical operations will succeed, because they are empowering their teams with the tools and knowledge to make intelligent decisions and provide better experiences for their internal and external customers.

To learn more about how Privacera and Starburst enable secure data democratization without sacrificing governance and compliance, watch our on-demand webinar here, or contact us for a demo.