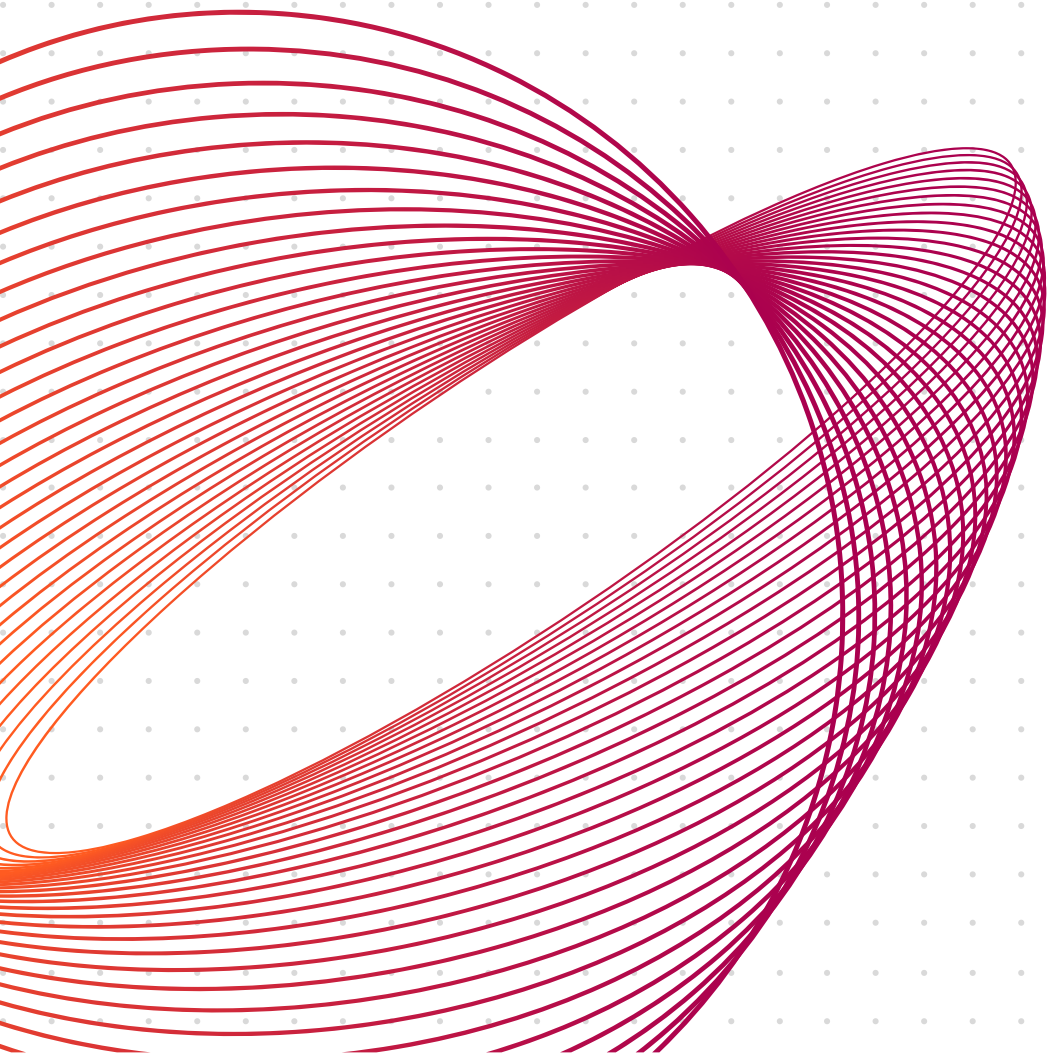
OKERA

WHITEPAPER

Six Critical Requirements for Scaling Secure Data Access



Executive Summary

Modern data platforms continue to grow in complexity to meet the changing needs of data consumers. Data analysts and data scientists demand faster access to data, but IT and governance are stuck at the last mile figuring out how to give access to the data in a secure, standardized way across a wide variety of analytic tools.

In this paper, we'll examine the six key requirements you need to consider in order to design such a framework and scale secure data access:

The ideal access control framework is dynamic, unifies metadata and audit logs, integrates seamlessly with your data platform, and requires minimal change management to implement.

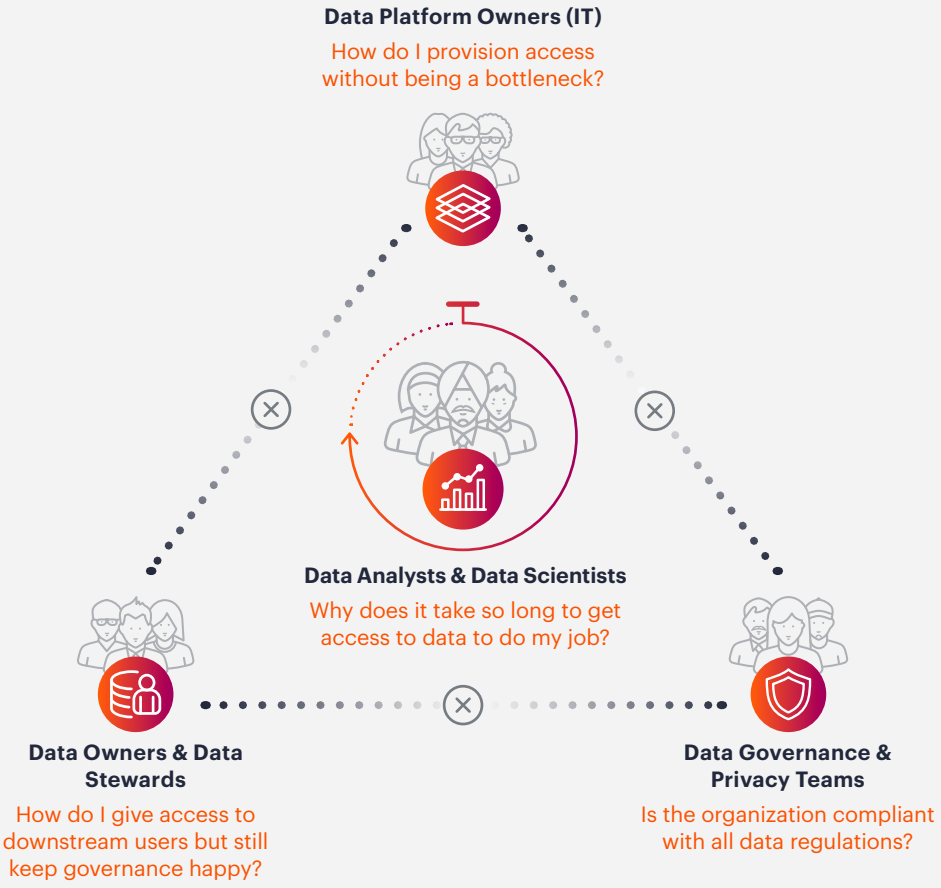
- Leverage attribute-based access control (ABAC) to scale policy management. ABAC, unlike role-based access control (RBAC), allows you to define flexible access policies by leveraging attributes from multiple systems in order to make a context-aware decision regarding any individual request for access.
- Enforce access policies dynamically. Dynamic enforcement is key to increasing the granularity of the policies without increasing complexity in the overall system, as well as ensuring your organization remains agile in responding to changing governance requirements.
- Create a unified metadata layer. Unifying metadata on a continuous basis establishes a "single source of truth" with respect to data, avoiding "metadata drift" and enables effective data governance processes such as data classification. Seek to automate metadata management early.
- Enable distributed stewardship. Your access control framework should seek to enable an organizational model of having more centralized IT and governance teams, and distributed business units.
- Ensure easy centralized auditing. Having visibility into (1) where sensitive data lives (2) who is accessing it, and (3) who has access to it, is critical to making intelligent access decisions. Invest in a basic visibility mechanism early on in your data platform, even if you can't control access to everything just yet.
- Future-proof your integrations. A flexible, API-driven architecture will ensure that your access control framework is future-proof and can adapt with the needs of your data platform. Ensuring this flexible foundation also means you don't need to have your entire architecture figured out from day one, but can easily add things as you go.

The ideal access control framework is dynamic, unifies metadata and audit logs, integrates seamlessly with your data platform, and requires minimal change management to implement.

Modern Data Platforms are Complex

Organizations of all sizes leverage data in order to better understand and delight their customers, achieve competitive advantage, and improve operational efficiency. To meet these needs, the enterprise data platform needs to account for a growing amount of complexity. One of the biggest challenges facing data platform teams today is how to make data accessible from the range of disparate storage systems (data lakes, data warehouses, relational databases, etc) whilst meeting increasingly complex data governance and compliance requirements due to emerging privacy legislation (GDPR, CCPA, etc).

This complexity is further exacerbated by the disconnect between key stakeholder groups: the centralized data platform and data governance teams, the data stewards sitting in the lines of business, and the data analysts trying to get access to data. If this complexity is not successfully abstracted away, it significantly affects everyone's productivity and limits the amount of data being made available.



Key users of the data platform are disconnected

Pushing Through the Last Mile of Delivering Value

Without this critical alignment across people and technology, organizations are still stuck at the "last mile" of delivering value. In order to generate business value, data consumers need to be able to **find** the right dataset, **understand** its context, **trust** its quality, **access** it in the tool of their choice with the right data access and governance **policies** applied.

Data analysts are stuck at the last mile trying to get access to data

Organizations seeking to accelerate time-to-insight on their data platform need to close this critical gap. Scaling access control successfully throughout the organization requires a solution that engages all four stakeholder groups, and ensures:

Organizations seeking to accelerate time-to-insight on their data platform need to close this critical gap.

- **Data consumers can easily get access to data regardless of where it's stored, with all access policies dynamically and consistently applied.**
- **All stakeholders have the correct access to perform their tasks.**
- **The user experience is intuitive and makes performing those tasks seamless.**
- **There is full visibility into everything happening in the system.**

... all without compromising on security.

The following sections detail the six key areas needed to architect such a system.

Six Critical Requirements for Scaling Secure Data Access

When designing or architecting your data platform, consider these key requirements for how to effectively scale access control throughout the organization:

1. **Leverage attribute-based access control**
2. **Enforce access policies dynamically**
3. **Create a unified metadata layer**
4. **Enable distributed stewardship**
5. **Ensure easy centralized auditing**
6. **Future-proof your integrations**

1. Leverage Attribute-Based Access Control (ABAC)

Most organizations start creating access control policies using role-based access control (RBAC). For example, a data analyst sitting in the credit card line-of-business unit (LOB) in a financial organization could have the role card_analyst, which is granted access to data within the transactions database.

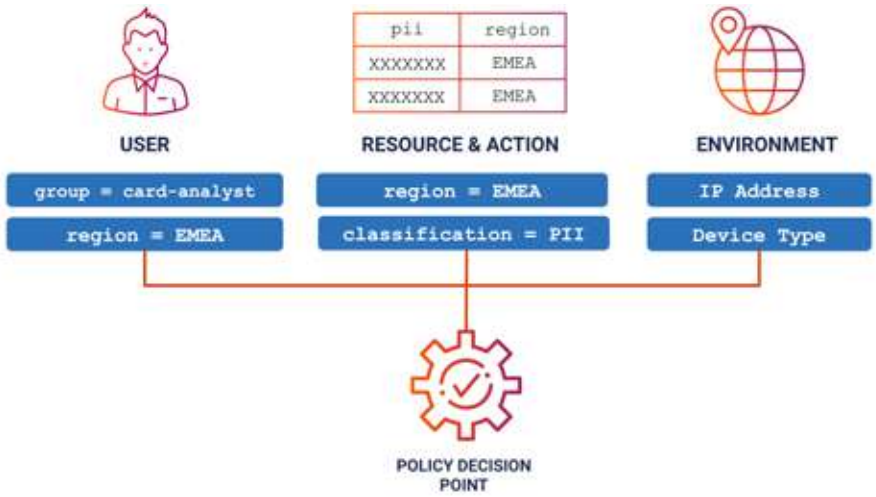
RBAC is useful for simple use cases, but since roles are manual and inherently static, every new use case requires the creation of a new role with new permissions granted to that user. As the data platform grows in scale and complexity, this eventually results in a very painful policy environment called "role explosion". Furthermore, each system has its own standards of defining and managing permissions on roles, and often RBAC is limited to coarse-grained access (e.g to an entire table or file).

RBAC is useful for simple use cases, but since roles are manual and inherently static, they cannot scale with your data platform.

Attribute-based access control (ABAC), on the other hand, allows you to define flexible access policies by leveraging attributes from multiple systems in order to make a context-aware decision regarding any individual request for access. ABAC, being a superset of RBAC, is able to support the complexity of very granular policy requirements, and expand data access to more people and use cases.

Within an ABAC system, there are three main categories of attributes that can be used to define policies:

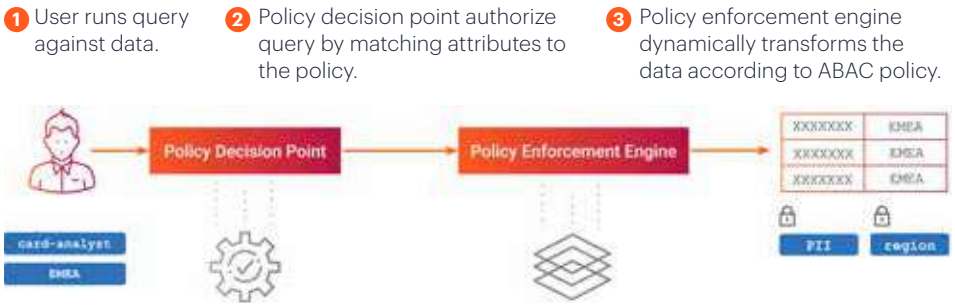
- **User attributes: about the user trying to access the data, e.g. department, region, level etc; e.g the region of an analyst user could be "EMEA"**
- **Resource attributes - this could be thought of as both data and metadata e.g. the name of a column "region" or the value of that column "EMEA". The also includes business metadata such as data classification; e.g. "pii".**
- **Environmental attributes - information about the request environment such as time, location, application, IP address; e.g. users can only access data between 9am and 5pm.**



An example ABAC policy: Members of the Card analyst group can access customer transaction data read-only for their region only, but only if they are in the office; and any data classified as restricted is always masked.

2. Enforce Access Policies Dynamically

Most existing solutions for policy enforcement still require maintaining multiple copies of each dataset, and the cost of creating and maintaining these can quickly add up. Simply leveraging ABAC to define policies doesn't completely alleviate the pain if when the attributes are evaluated against the access policy at the decision point, you're still pointing toward a static copy.



Example of dynamic ABAC policy enforcement. The analyst in the card LOB and EMEA region will only see rows from the EMEA region and personally identifiable information is masked.

Dynamic enforcement is key to increasing the granularity of access policies without increasing complexity in the overall system.

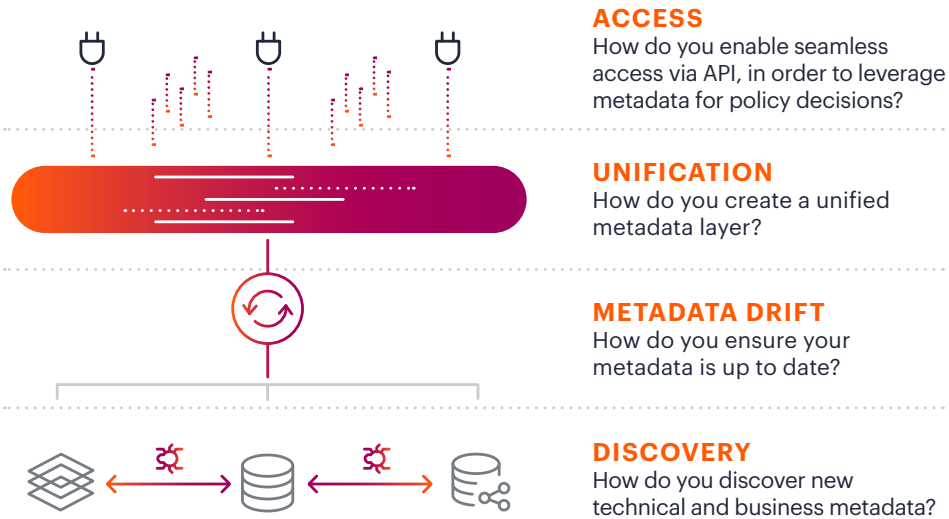
Once you've done the hard work of defining attributes and policies, you want this to be pushed down to the enforcement engine to dynamically filter and transform the data. This can mean redacting a column, or applying data transformations like anonymization, tokenization, masking, or even advanced techniques such as differential privacy. Dynamic enforcement is key to increasing the granularity of access policies without increasing complexity in the overall system. It's also key to ensuring the organization remains heavily responsive to changing governance requirements. For example, if the organization decides to change how customer date of birth is shown, this change should be seamlessly propagated through all data access.

Access at the enterprise level means many hundreds or thousands of users accessing data. Any dynamic enforcement engine needs to be architected in a distributed and horizontally scalable way, working in a symbiotic relationship with the tools querying data. Ensuring that your policy enforcement always leverages **the most performant method of enforcement** without compromising on security, consistency, or end-user experience is key. Otherwise, it further reinforces data consumers' perception that access control slows them down.

3. Create a Unified Metadata Layer

If ABAC is the engine needed to drive scalable, secure data access, metadata is the engine fuel. It provides visibility into the what and where for the organization's datasets, and is required to construct attribute-based access control policies. A richer layer of metadata will enable you to create more granular and relevant access policies with it.

When it comes to your architecting your metadata lifecycle, there are four key areas you need to be thinking about:



If ABAC is the engine needed to drive scalable, secure data access, metadata is the engine fuel.

The challenge is that metadata, just like data, typically exists in multiple places in the enterprise, and these are **often owned by different teams**. Each analytical engine often requires its own technical metastore, whereas governance teams maintain the business context and classifications within a business catalog like Collibra or Alation. Therefore, organizations need to federate and unify their metadata so that the complete set is available in real time for governance and access control policies. Inherently, this unification needs to be an abstract layer since it would be unreasonable, and almost impossible, to expect to have metadata always defined in a single place.

Unifying metadata on a continuous basis establishes a single source of truth with respect to data. This helps to avoid "metadata drift" or "schema drift" (inconsistency in data management) over time, and enables effective data governance and business processes such as data classification, or tagging, across the organization. It also establishes a unified data taxonomy, making data discovery and access easier for data consumers.

Many forward-thinking organizations have invested in metadata management tools that use artificial intelligence to automate parts of the metadata lifecycle; for example, identifying sensitive data types and applying the appropriate data classification, automating data discovery and schema inference, and automatically detecting metadata drift.

4. Enable Distributed Stewardship

Scaling secure data access is not just a matter of scaling the types of policies and enforcement methods. The process of policy decision-making must also be able to scale, because the types of data available and the business requirements to leverage it are so diverse and complex. In the same way that your enforcement engine could be a bottleneck (if not properly architected), the lack of an access model and user experience that enables non-technical users to manage these policies gets in the way of an organization's ability to scale access control.

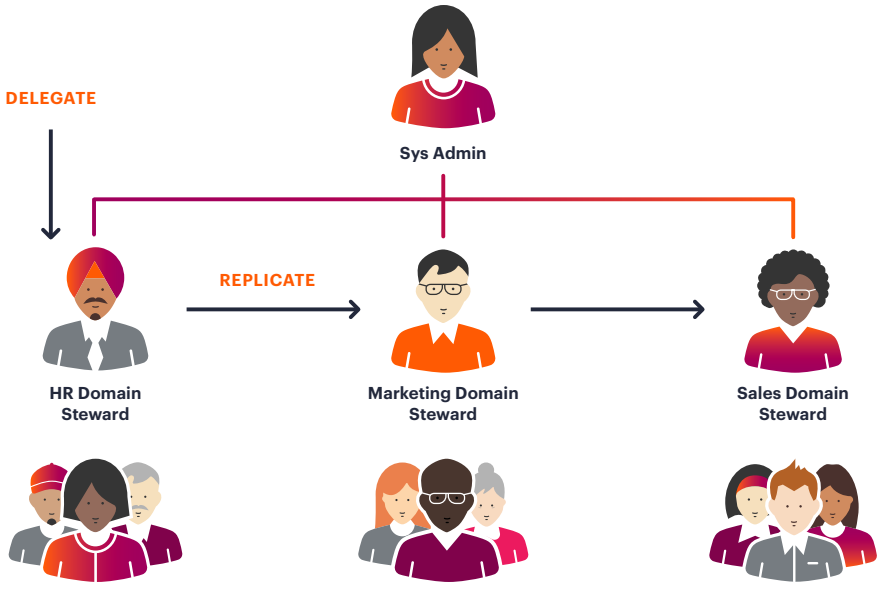
Typically within an organization you see a centralized IT or platform team, a separate centralized governance team, and distributed business units. The challenge occurs when IT manages the centralized access infrastructure, but the data itself sits in the business units, along with the stewards and owners who best understand that data.

Data access management should seek to **embrace this model, not obstruct it**. Many access management tools require complex change management and the development of bespoke processes and workflows in order to be effective. In order to avoid this, you should be asking "how can this access model adapt to my organization?" very early on.

Data access management should embrace the distributed stewardship model, not obstruct it.

To enable "distributed stewardship", you should be looking at how easily your access system supports two key areas:

- **Delegating the management of data and access policies to people in the lines of business (data stewards and administrators) who understand the data or governance requirements.**
- **Replicating centralized governance standards across groups in the organization, and ensuring that change can be propagated consistently throughout the organization.**



Distributed data stewardship

A centralized governance system is still required to enable rapid response to regulatory change; for example, if a new type of information (say, email address) is now deemed to be PII, the policy update can be made once and applied globally. But it must have a component of flexibility and accessibility to those outside of the data platform team, so that non-technical users can handle the day-to-day onboarding of new datasets and granting permissions to their business users. This will provide the organization with a tremendous amount of business agility.

5. Ensure Easy Centralized Auditing

Having visibility into (1) where sensitive data lives, (2) who is accessing it, and (3) who can access it is critical for making intelligent access decisions.

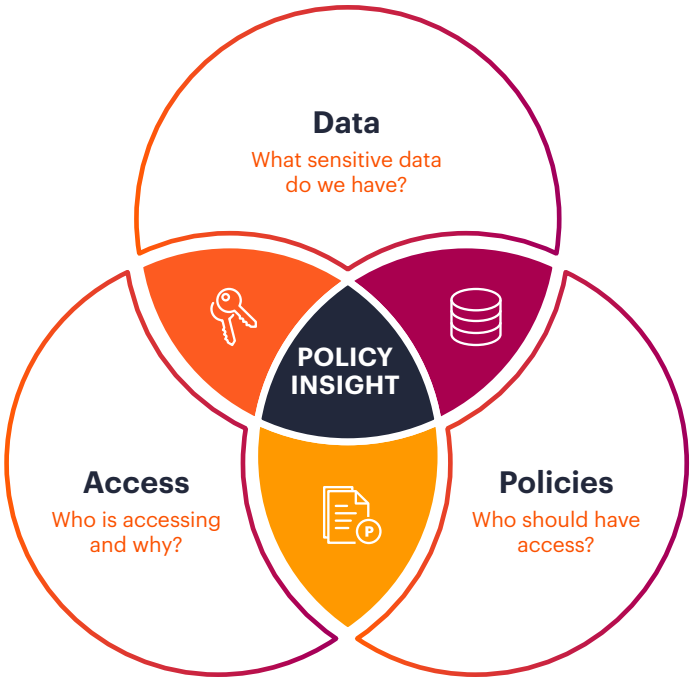
Auditing is a consistent challenge for governance teams, since there is no single standard across the variety of tools in the modern enterprise environment. Collating audit logs across various systems so that governance teams can answer basic questions like the ones above is painful, and definitely doesn't scale.

Once you know what data you have and start to understand how people are using it, you can design more effective access policies around it.

The governance team, despite setting the policies at the top level, has no way to easily understand whether or not their policies are being enforced at the time of data access and the organization's data is actually being protected. The ideal solution would replace the static process of generating one-time reports with continuous insights into how the data is being used.

Therefore, centralized auditing with a consistent schema is critical for generating reports, and can also enable automated data breach alerts through a single integration with the enterprise SIEM. Unfortunately, without a unified access management system, some amount of custom plumbing and ETL will be necessary, though there are tools with integrations that can make it easier. One key consideration is to make sure that your audit log schema can enable governance folks to answer audit questions, since many log management solutions are more focused on application logs.

You should invest in a basic visibility mechanism early in your data platform journey, to help data stewards and governance teams understand data usage and help demonstrate the value of your platform. Even if you can't control access to everything just yet, once you know what data you have and start to understand how people are using it, you can design more effective access policies around it.



Policy insight is a continuum and interesting insights sit at the overlap of key questions

Ultimately, a flexible, API-driven architecture will ensure that your access control framework is future-proof and can adapt with the needs of your data platform.

6. Future-Proof Your Integrations

Integrating with the broader environment has been a consistent theme throughout this whitepaper. Your data platform will likely change over time as data sources and tools evolve, so your access control framework must be adaptable and support flexible integrations across the data fabric.

One huge advantage of using ABAC for access control is that attributes can come from existing systems within the organization. Of course, you need to ensure that attributes can be fetched in a performant way in order to make dynamic policy decisions.

Ultimately, a flexible, API-driven architecture will ensure that your access control framework is future-proof and can adapt with the needs of your data platform. Ensuring this flexible foundation also means you don't need to have your entire architecture figured out from day one. Instead, you can start with a few key tools and use cases, and add more as you learn more about how your organization uses data.

Some organizations choose to focus on open-source for this reason, since they have the option to customize integrations to their needs. However, a key consideration here is that building and maintaining these integrations can quickly become a full time job. In the ideal scenario, the data platform team should remain lean and have low operational overhead. Investing time into engineering and maintaining integrations is unlikely to provide differentiation to your organization, especially with several high quality integration tools in the ecosystem.

Below are some of the key areas you can start to focus on to ensure performant integrations to scale access control.

1

USER AND DATA ATTRIBUTES

Leverage attributes from existing systems in your organization such as identity management systems and business catalogs.

Microsoft Active Directory collibra

2

SEAMLESS AUTHENTICATION EXPERIENCE

End users of data shouldn't need to manage multiple credentials to access data across different tools.

okta LDAP

3

UNIFIED AUDITING

Integrate data access logs from your tools to gain visibility into how sensitive data is being accessed.

splunk> sumo logic

4

CONSISTENT ENFORCEMENT ACROSS TOOLS

A single access policy needs to be consistently enforced across the data repositories and analytics tools used in your organization.

APACHE Spark presto

Like with any big initiative, it's important to take a step back and leverage a design-to-value approach.

Getting Started

Like with any big initiative, it's important to take a step back and leverage a design-to-value approach when trying to scale secure data access. This means finding the highest value data domains that need access to sensitive data and enabling or unblocking them first, as well as trying to establish visibility on how data is being used today in order to prioritize action.

Companies can begin their journey by answering these questions:

- **Is my RBAC approach working seamlessly, and are there any pain points users have brought up?**
- **Do I understand who is using what data assets for what use cases?**
- **What would be a phased approach to implementing an access control environment that covers my high-risk data assets across the organization?**
- **Does my organization currently use or is planning on using a data catalog? How can I integrate with this to implement ABAC policies?**
- **How does my current access control framework adapt to my organization? Is it enabling the distributed stewardship model, or obstructing it?**
- **How agile is my governance strategy? What happens if the organization needs to respond to a changing regulation?**
- **How future-proof is my access control framework for my data fabric? Does it have a flexible underlying API-driven foundation?**

Conclusion

Organizations are making significant investments in their data platforms in order to unlock new innovation; however data efforts will continue to be blocked at the last mile without an underlying framework for truly scalable and secure data access. Okera's belief is that scaling secure data access can be a tremendous enabler of agility within the organization. You can leverage the six principles discussed in this paper to ensure that you're staying ahead of the curve and designing the right underlying access framework that will make all your stakeholders successful.

Faster time to insight for your data consumers, without having to compromise on security.

Governance teams will feel confident that their policies are being enforced, while data stewards are empowered to manage day-to-day access. IT is free to focus on enabling their end users with the best tools, rather than constantly worrying about access control and complex integrations. Ultimately, this means much faster time-to-insight for your data platform without having to compromise on security.

If you're ready to get started building a modern data platform with scalable, secure data access, **let's talk**.

ABOUT OKERA

Okera enables the management of data access and governance at scale for today's modern data lakes. Built on the belief that companies can do more with their data, Okera's Active Data Access Platform (ODAP) allows agility and governance to co-exist and gives data consumers, owners and stewards the confidence to unlock the power of their data for innovation and growth. Okera can be deployed in as little as one day to facilitate the provisioning, accessing, governing and auditing of data in today's multi-data format, and multi-tool world.

Learn more at www.okera.com or contact us at info@okera.com