OKERA

WHITEPAPER

Securing Amazon S3 Data Lakes

Core Capabilities and Benefits of Comprehensive Data Security

Introduction

Many successful companies have recognized that Data is a true differentiator for their business. To leverage the power of data, companies have deployed many data scientists and analysts to derive value from this data. However, data analysts and other consumers that work with data to support business initiatives now demand the ability to use different kinds of analytics frameworks and tools that are best suited for the workloads they are trying to run.

Cloud-based infrastructure makes it even easier to build and operate such platforms and is making the promise of self-service data lakes possible.

Modern data lake capabilities that offer flexibility at both the storage and analytics framework layer are attractive, and fast becoming a reality. It enables data analysts and scientists to do more with data, and in the process, better serve their business users. Cloud-based infrastructure makes it even easier to build and operate such platforms and is making the promise of self-service data lakes possible. This requires several different technologies (storage, streaming, and analytics frameworks) work in concert to provide different kinds of abstractions (such as files of different formats, streams, SQL, NoSQL, etc) users can interact with.

While powerful, these varied abstractions are generally not very intuitive or usable for the mainstream analyst. This diversity of systems creates challenges for defining metadata and access policies, and enforcing them consistently across the board is an operational nightmare. The only way to enable agility without compromising on enterprise requirements for security, auditability, and usability, is with custom engineering that often forces you to create multiple copies of data for different workloads. This drives up inefficiencies, costs, and risk - in addition to slowing analyst productivity, which ultimately limits business agility.

Core Security Capabilities

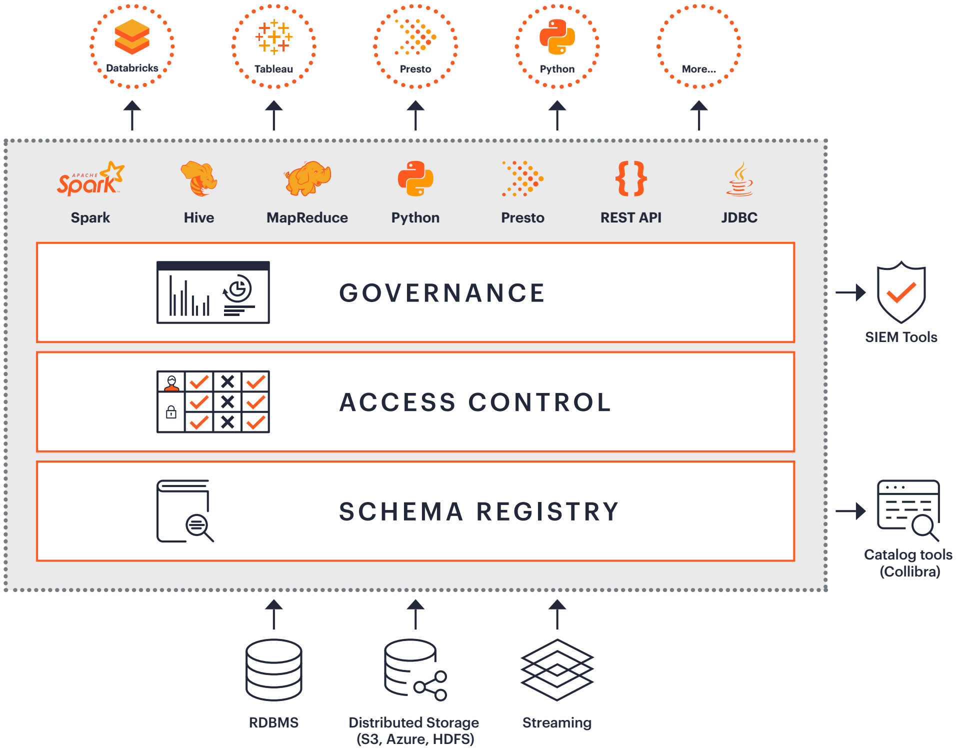
The **Okera Active Data Access Platform (ODAP)** solves this problem by providing data consumers with self-service access to data from a variety of storage and streaming systems. The Okera platform was designed with **high performance, fine-grained security and auditability** in mind. Users can employ any analytics tool they want using industry-standard APIs.

The Okera Active Data Access Platform consists of two main components:

- **Catalog Services - Contain the Schema Registry, Policy Engine and Audit Engine. Together they manage the fine-grained security policies, audit trail, and metadata.**
- **Data Access Service - enforces the security policies by dynamically providing restricted views to end users. The same user, accessing the same dataset, using any tool will always have the exact same policies applied.**

Okera's easy-to-use abstraction of logical datasets is presented as tables which removes the complexity of different APIs, file formats, streams, and coarse-grained access methods behind the scenes.

Okera's easy-to-use abstraction of logical datasets is presented as tables which removes the complexity of different APIs, file formats, streams, and coarse-grained access methods behind the scenes. This allows ODAP to perform much like a relational database does for all its internal components. In addition, it is modular in nature and exposes common APIs to make it easy for platform teams to integrate ODAP with their environment, and gives them the choice to use other services as they require.



Okera Catalog Services

Okera Catalog Services are a unified, common set of services that provide vital details to users and the Okera platform itself. It stores dataset definitions, access policies, and any other metadata that you may choose to include, that can be shared across different storage systems, streaming systems, and analytics tools. Currently the following services are included: **Schema Registry, Audit Engine, and Policy Engine**.

These services provide the following primary functions:

- **Dataset registration and publishing**
- **Defining fine-grained access policies down to the individual cell level, including user-defined anonymization, tokenization, masking as well as other security related functions**
- **Dataset search and access for analytics thereafter**
- **Comprehensive auditing and reporting for every metadata operation and any access that is processed by the platform**

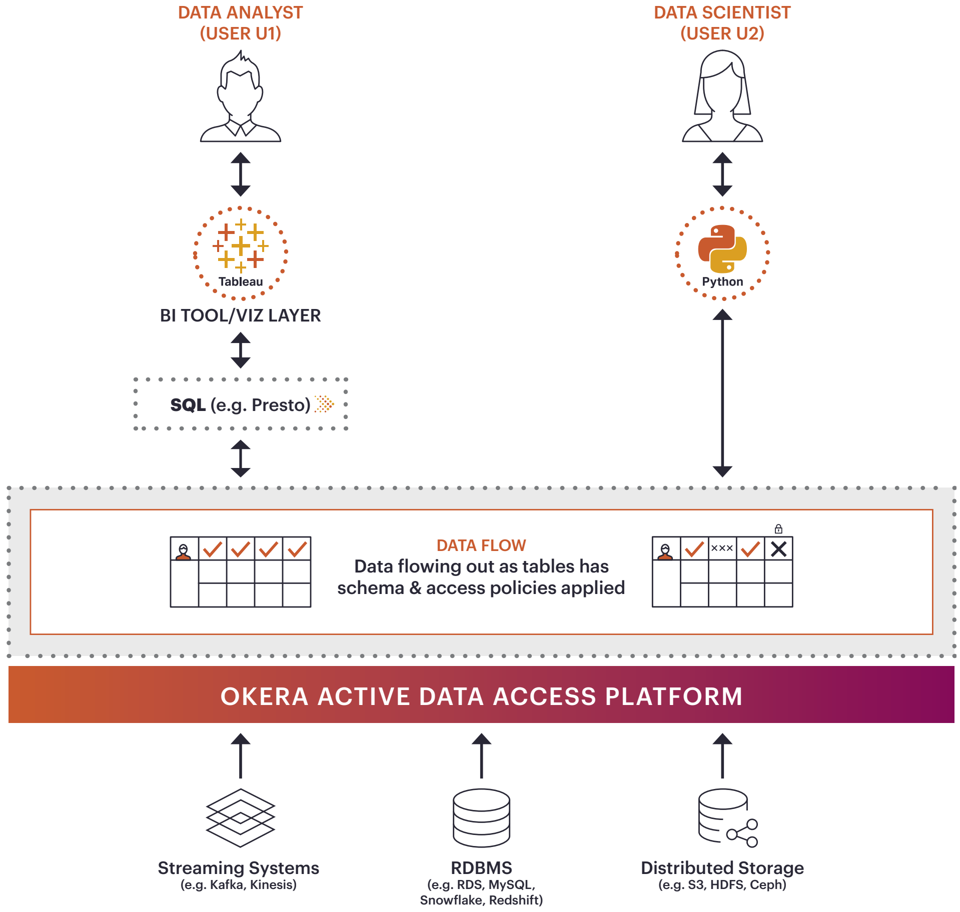
Typically, a single instance of the Okera Catalog Services are deployed and shared across multiple teams. The services expose standard APIs of the Hive Metastore (for schemas) and REST APIs to interact with the metadata. This makes it the long-running, common metastore for different Hadoop components, regardless of the infrastructure they run in (on-premise or cloud) and what analytics tool they use (Amazon EMR, Cloudera, HWX, MapR, Databricks, and so on). REST APIs can be used to integrate with any other systems and workflows that may already be in place.

Okera Data Access Service

The Okera Data Access Service is a scalable, fault-tolerant distributed service that makes it easier for businesses to use multiple analytics tools on their data lake. Okera's Data Access Service handles the heavy I/O while providing data to the analytics tools after applying schema, fine-grained security policies, and other transformations (for instance, UDFs, tokenization, masking) with high performance. Data provisioned in this form is easily consumable and delivered as a familiar abstraction of tables, or in the form of files in a preferred user format. Different analytics tools like Spark, Python, SQL engines, BI tools, and spreadsheets can all interact with this service. With Okera, every tool works with the same view of the data after individual user security policies have been applied.

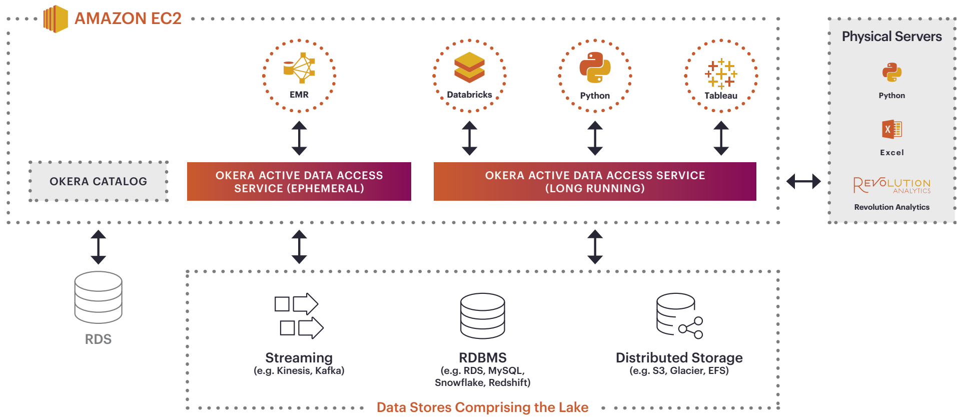
Users with different use cases and different choices of tools are able to interact with the Okera platform. Regardless of what tool or access API they use, they will get the same view of the Data.

Okera's Data Access Service handles the heavy I/O while providing data to the analytics tools after applying schema, fine-grained security policies, and other transformations (for instance, UDFs, tokenization, masking) with high performance.



In some instances, users will want to run multiple instances of Okera's access service running inside their environment. These instances may be ephemeral, while others may be persistent. In other cases, instances may act as independent services, whereas others may be co-located with an analytics framework. You can choose between flexible deployment models based on your performance and isolation requirements.

A typical Okera deployment is shown in the figure below - this type of deployment can be in your data center, on the cloud, or spanning both to provide a hybrid data access platform.



Key Security Benefits of ODAP

Okera's Active Data Access Platform is the missing link that enables true data democracy by providing secure, reliable, self-service access to data. With Okera, companies can enable mainstream analysts to easily find, understand, and securely access data for analytics with minimal dependence on data engineering and IT. This experience is similar to what a well-tuned data warehouse provides, but with the flexibility, power, and self-service capabilities of a true, cloud-based platform. Key benefits include:

FLEXIBILITY

A single place to store and enforce schemas and access policies, Okera delivers complementary audit access for any infrastructure or platform being used. Users also gain the agility to choose from best-of-breed technologies that match their unique data requirements.

LOW COST

Okera lowers the time and cost required to enable different kinds of users or types of workloads. By significantly reducing the amount of manual plumbing required, Okera makes operating and maintaining multiple compute tools significantly easier to manage.

> *With Okera, companies can enable mainstream analysts to easily find, understand, and securely access data for analytics with minimal dependence on data engineering and IT.*

LOW RISK

Okera eliminates the need for creating multiple copies of data for different use cases and users. This makes governing access much easier and reduces overall risk for violating compliance guidelines like GDPR and CCPA.

SPEED

Okera increases the speed with which users can run analytic queries by reducing the complexity and time required to provision data. Easy-to-use abstractions are so powerful that data consumers can immediately become more productive.

PORTABILITY

With Okera, users gain the ability to port workloads between different infrastructures and platforms because it automatically produces a technology agnostic abstraction layer for all underlying systems.

Conclusion

When architected properly for governance and security, data lakes can offer remarkable business agility. From a security perspective, understanding that analytic tool usage, user behavior, and data will all change—being prepared for this ongoing pace of change is necessary if data producers and stewards want to support data consumers with their tools of choice.

Enterprises are now learning to view data lakes as a series of capabilities instead of a linear technology stack. This evolution in how businesses produce, consume, protect, and govern their data is creating new opportunities and new pressure for data professionals. To unlock the full power of data lakes, companies require a unified view into all the attendant technologies that their data lake relies on, including storage, streaming, and analytics frameworks. With a single view into the security of these heterogeneous technologies, businesses can gain the security and governance capabilities they require to maintain business agility as their data lake continues to grow.

To learn more about how Okera can help you achieve your data lake security and governance goals, contact us today at **info@okera.com**.

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