OKERA

WHITEPAPER

Unified Governance for Amazon S3 Data Lakes

Core Capabilities and Best Practices for Effective Governance

Introduction

Data governance ensures data quality exists throughout the complete lifecycle of data. Historically, governance has been primarily concerned about risk mitigation, protecting against data breaches, securing data and, in many respects, restricting business users from accessing unauthorized data. With the emergence of data lakes with a focus on business agility, data governance has now evolved to empowering users to find, trust, and consume data sets so organizations can realize their vision of becoming data-driven incurring any business risk.

A key outcome of data governance is comprehensive auditability of the system, including full visibility into data usage. This data includes tracking datasets by how often they are used, as well as ensuring their usage complies with industry privacy standards

What is Data Governance?

Simply put, governance for a data lake encompasses: (1) knowing what data is available, (2) allowing the right kind of access, and (3) understanding how the data is being used. This often includes tracking the movement of data throughout the data lake, from ingestion, to transformation/ETL, then consumption, and to its eventual removal or archiving. Various data governance tools can help show where data comes from, how it is identified and made discoverable, how it is changed to suit an operational or business requirement, and who is able to see and use it.

A key outcome of data governance is comprehensive auditability of the system, including full visibility into data usage. This data includes tracking datasets by how often they are used, as well as ensuring their usage complies with industry privacy standards and regulation as well as the business's internal operational guidelines

Key Concerns for Effective Data Governance

Effective governance involves a complex range of requirements, so it's common to build a foundation then grow in sophistication as other needed attributes are identified. There are core capabilities every enterprise should consider in a governance solution, including:

- **Capturing and managing metadata**
- **Enforcing access control**
- **Anonymizing sensitive data**
- **Capturing data lineage**
- **Capturing audit activity**
- **Tracking data quality**

CAPTURING AND MANAGING METADATA

To build consistent, repeatable governance practices for data, you need a method for describing its content (i.e., metadata). There are three main types of metadata that are useful for data lakes:

CAPTURING AND MANAGING METADATA

METADATA TYPE	EXAMPLES
TECHNICAL	FORMAT (JSON, AVRO) & STRUCTURE (FIELDS, DATA TYPES)
OPERATIONAL	DATASET SIZE, NUMBER OF RECORDS, PERFORMANCE METRICS
BUSINESS	NAMING FOR BUSINESS VALUE, CATEGORICAL TAGS, QUALITY ATNALYSIS, METHODS TO PROTECT SENSITIVE DATA (E.G., TOKENIZING)

Technical metadata, as the table implies, is driven by your storage model. Operational metadata is an intermediate layer that provides the base you need for auditing, estimating workloads, and tracking data lineage. Business metadata supplies the context that application developers and data consumers need. To unify your governance model, business context must relate to technical and operational metadata without requiring business users to understand them. The "Metadata Management" section below offers more detail on this point.

ENFORCING ACCESS CONTROL

Data lake access control is particularly challenging because of the diversity of data, the volume of data, and the multitude of analytic engines that can access the data. Despite these complexities, data lake access control is a simple problem: data consumers must be able to access only the data they're allowed to access. Unfortunately, this is a very complex problem to solve.

The solution is complex because data lake access control needs to be enforced at many different granularities: at the file-level (e.g., X-ray images), table-level (e.g., sales data), field-level (e.g., social security number), record-level (e.g., only California records), and cell-level (e.g., last four digits of social security number). To make things even more complicated, data consumers need to work with a variety of analytic tools—each with their own data models (tables, files, RDDs, etc.)—to access this data.

Furthermore, data lakes are more permissive by design, with the separation of storage and compute and a, schema-on-read approach. Consequently, although access control must be specified in very granular ways, it must be flexible enough to deliver protected data to a variety of analytic tools.

Data lakes are more permissive by design, sometimes using services with a schema-on-read approach. Consequently, although access control must be specified in very granular ways, it must be flexible enough to deliver protected data to a variety of analytic tools—many of which don't know how to work directly with tables.

ANONYMIZING SENSITIVE DATA

In some cases, users want to see certain fields, but privacy regulations require that those specific field values remain hidden. This can include credit card numbers, email addresses, phone numbers, and social security numbers. In many cases, these values must be replaced with a tokenized version of the original, such as a unique hashed value that can be used for a tally. At other times, you may need to anonymize the value so that the underlying data can't be used to infer other facts. These are difficult problems in any multi-tenant environment where data sets are shared, and they are all but impossible to address without a centralized governance model.

CAPTURING DATA LINEAGE

Data lineage captures the data's original data sources, what happens to it, and where it moves over time. Lineage captures the sources of a data set and the movement of that data over time, from ingest to egress. Lineage serves several important purposes:

- **Trust - Upstream lineage, or provenance, shows the data sets and transformations that generated a new data set. By providing insight into the data sources of a particular data set, along with the transformations that generated that data set, lineage helps end users determine whether they should trust a data set as reliable and complete.**
- **Impact - Downstream lineage, or Impact analysis, helps data engineers understand which data sets would be impacted by changes to particular data sets. For example, lineage can help a data engineer determine the transformations that would need to be updated when a column is added or modified in a particular upstream data set.**
- **Compliance - Detailed, or column-level, lineage is a critical requirement for many data regulations, such as PCI, HIPAA, GDPR, and CCPA. By capturing the movement of data at a granular level, lineage arms compliance groups with vital information in the event of an audit.**

Recording the actions performed on each dataset makes it possible to review potential exposures of sensitive data, such as changes to access rights or removing a field's masking.

Tools that support lineage often include metadata tags that map technical and operational names to business meaning. It is up to governance to curate and preserve these meanings for consistency across different lines of business.

Tools that support lineage often include metadata tags that map technical and operational names to business meaning. It is up to governance to curate and preserve these meanings for consistency across different lines of business.

CAPTURING AUDIT ACTIVITY

Who accessed sensitive data at the time of a security breach? Is there a spike in denied access to the data lake? What data assets did a particular employee work with before he or she resigned? The only way to answer questions like this is through comprehensive audit logs. However, audit logs are traditionally compute-specific—there's one for Hive activity, another for Spark activity, another for Presto activity, and another one for every additional compute engine. Even worse, each audit log contains different granularities of information. Without a better approach to auditing, it's either extremely difficult or altogether impossible to figure out who accessed sensitive data at the time of a security breach.

TRACKING DATA QUALITY

Data lakes, unlike data warehouses, impose no structure on data. Changing storage formats, adding compression, and converting field data types are typical refinement and enrichment activities, but they're never required. It is possible data may be validated for quality at this stage, but it is also possible this work is part of a data engineering or ad hoc process after ingest. Consequently, it's important to make sure that data consumers always understand the quality of data sets they plan to analyze.

This refinement process helps build metadata models from the technical layer to the operational layer. At this layer, a consistent, generalized metadata model simplifies the task of measuring the data's quality. A unified data governance model establishes common metrics for all data sets. Over time, it becomes more sophisticated as the business learns how strong data quality contributes to making better business decisions.

Fundamental Capabilities for a Governed System

Three of the elements we've discussed above are foundational to creating a unified governance model: metadata management, access control, and auditing capability. Without careful attention in designing these capabilities, achieving any other governance benefits becomes laborious, expensive, and difficult to maintain.

METADATA MANAGEMENT

To realize good governance in a data lake environment, you need a metadata foundation that starts with technical elements and works up towards business meaning and value. This direction helps you understand the data granularity required for effective access control and helps discern the best configuration for access patterns.

To realize good governance in a data lake environment, you need a metadata foundation that starts with technical elements and works up towards business meaning and value.

Working from the top-down, as you might when designing a B2B application, can be a reasonable alternative because doing so makes it easier to identify data-exchange requirements up front. For example, with a top-down design, you would enumerate the data elements required then orchestrate their retrieval. The business value of this approach can be communicated in a straightforward manner and early in the design process.

Conversely, a top-down (or business-first) definition tends to view governance as an exception-driven bureaucracy at best, and a gauntlet of operational blockers at worst. In such cases, the appearance of an inflexible governance model can be a self-fulfilling prophecy. It becomes more likely that business units, in the name of agility, resort to closely-held copies of enterprise data and metadata. A data catalog with many entries that are almost identical, except for their business names, is a common consequence.

Maintaining balance between robust governance and agile business practices still requires a mindful approach, but getting there smoothly depends on getting the technical metadata under control first.

A data governance model is responsible for describing the visibility of data to each defined role—administrator, data steward, data analyst, business user, and so on—to whom data stewards can then assign appropriate access.

POLICY-BASED ACCESS CONTROL

Fitting new data to a normalized type, such as a basic table format, makes it much easier for catalog users to understand the value of an unfamiliar data set. This is also an ideal contact point for factoring away technical elements—such as location, compression type, and row or columnar storage format—so the end user does not have to factor these into their use of data.

A data governance model is responsible for describing the visibility of data to each defined role—administrator, data steward, data analyst, business user, and so on—to whom data stewards can then assign appropriate access. With access policies initially defined at this relatively coarse grain, it's much simpler to establish policy at a finer granularity later, including sensitive data points and complex, multiple-role assignments for a given consumer.

AUDITING

To be effective, an auditing system must be baked in to any governance architecture, and must be comprehensive-supporting analysis through technical, operational, and business layers. It must also track application consumption and operational changes.

Your auditing capability should also offer a way to export audit data in a form suitable for use by third-party tools. As the system grows in sophistication, administrators will want audit data to support metrics for monitoring SLAs, tuning system utilization, and optimizing workload performance.

Benefits of a Unified Governance Strategy

Armed with these core capabilities—a metadata management practice, policy-based access controls, and an auditing subsystem—plus an achievable roadmap to bring the remaining components into play, a well-rounded data governance strategy will yield more than the sum of its parts. At Okera, the endpoint for all this work is a robust, unified governance model that easily scales to new analytic tools, storage systems, and data formats.

Consider what's possible in a data lake where the concerns of storage and compute services are brought together through a central data platform service layer:

- **A model for catalog services that enables easy discovery and analysis of data sets**
- **Support for multiple tenants and multiple compute frameworks**
- **Integrated services to audit privacy, security, and regulatory compliance**
- **Governance that supports business agility—users can create new data sets and try out new tools without engaging security or IT personnel**
- **Self-service for existing data sets, backed by quality metrics and usage statistics**
- **Comprehensive usage insights to understand business value and plan for growth**
- **Chargebacks - evaluating costs on per-department or per-team basis**

One Okera customer (a Fortune 500 retailer) manages, audits, and reports on over a trillion data records in production every day.

In the face of high-stakes challenges to comply with GDPR and CCPA, enterprises need a unified governance model to mitigate risk - but they can't afford to compromise on business agility. At the time of this writing, one Okera customer (a Fortune 500 retailer) manages, audits, and reports on over a trillion data records in production every day.

The total compliance liability this customer faced annually—up to $1.4B in penalties for improperly protecting customer data—shows the value of a unified governance strategy. Without a comprehensive method for implementing and reporting on privacy compliance, this customer's business units would have been forced to act on their own. Using Okera, they avoided redundancies not only in storage and compute costs, but also in dedicating staff to solving the same problem for each line of business.

Conclusion

Modelling and developing governance for your data lake is no small or simple task. At the outset, the architecture everyone wants should help reduce cost, eliminate duplication of effort, and promote agile practice without sacrificing security or privacy. These goals encourage wider adoption, which opens the door to deeper insights into more efficient operations and deriving greater business value from your data.

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