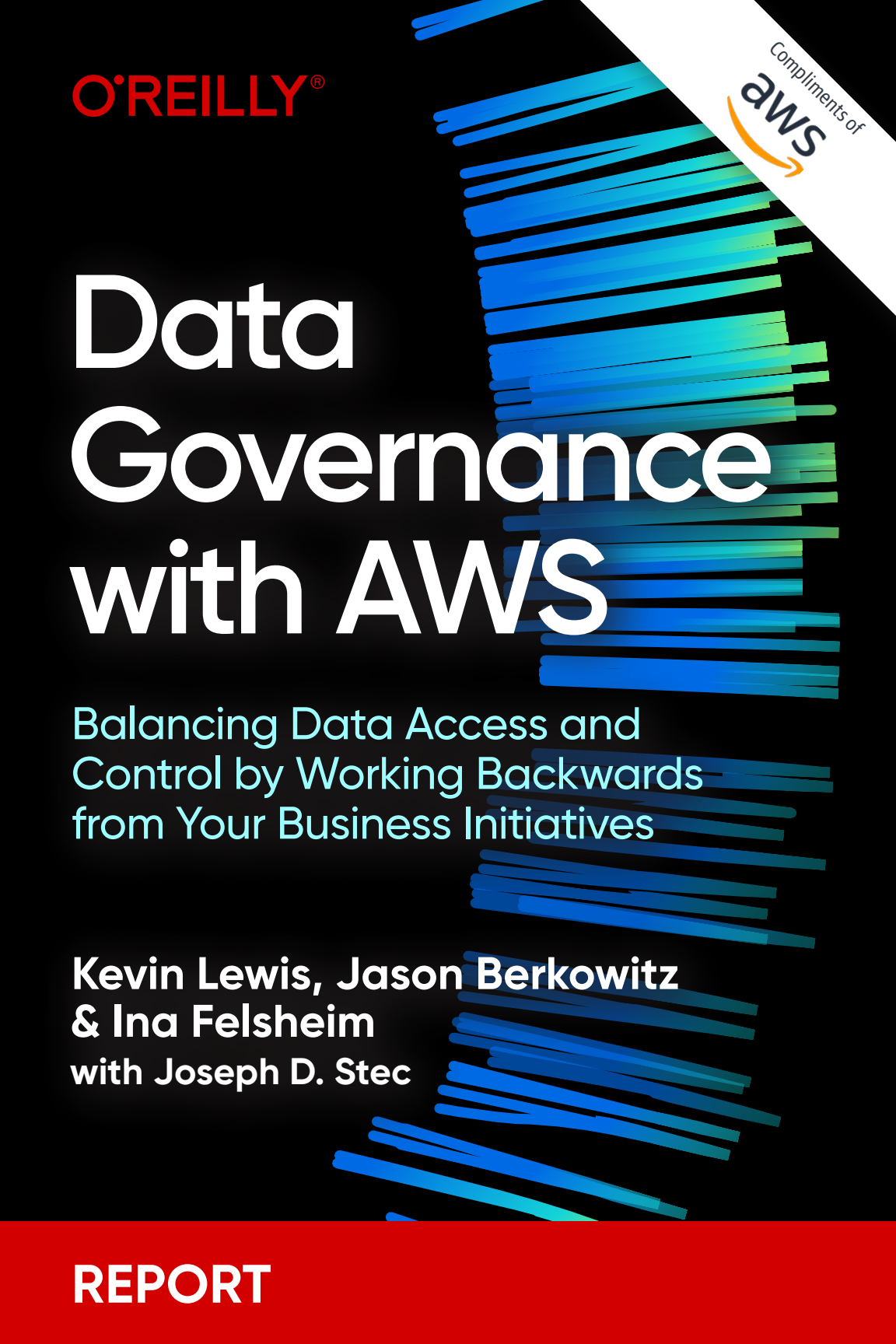
O'REILLY®
Compliments of
aws
Data Governance with AWS
Balancing Data Access and Control by Working Backwards from Your Business Initiatives
Kevin Lewis, Jason Berkowitz & Ina Felsheim
with Joseph D. Stec
REPORT

Data Governance with AWS

Balancing Data Access and Control by Working Backwards from Your Business Initiatives

Kevin Lewis, Jason Berkowitz & Ina Felsheim
with Joseph D. Stec

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Data Governance with AWS
by Kevin Lewis, Jason Berkowitz, and Ina Felsheim with Joseph D. Stec

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Acquisitions Editor: Michelle Smith
Development Editor: Jill Leonard
Production Editor: Clare Laylock
Copyeditor: nSight, Inc.

Proofreader: Rebecca Gordon
Interior Designer: David Futato
Cover Designer: Randy Comer

April 2024: First Edition

Revision History for the First Edition
2024-04-22: First Release

978-1-098-15753-1

[LSI]

Table of Contents

Preface

Data governance is the collection of policies, processes, and systems that organizations use to ensure the quality and appropriate handling of their data throughout its lifecycle for the purpose of generating business value.

Working at Amazon Web Services (AWS) has taught us a lot about helping organizations deliver data-driven solutions. To succeed with these solutions, we've learned that it's crucial to support them with effective data governance in order to ensure businesses execute at speed while not introducing new business risks. It is crucial to make it easy for the right people and applications to securely access and share the right fit-for-purpose data to meet business needs. But to implement data governance successfully, we've also learned that there are a few seemingly subtle actions that can make all the difference in results. By diving into these pages, you will gain insights into how to effectively and pragmatically enhance your data curation, data understanding, and data protection capabilities. These capabilities will empower you to share your data with control and clarity while supporting the most important business initiatives within your enterprise.

Who This Report Is For

This report has been crafted primarily for C-suite executives—specifically chief data officers (CDOs), chief analytics officers (CAOs), and chief information security officers (CISOs)—who are deploying data-driven solutions that need modern data governance. Executives will gain the necessary insights and knowledge to navigate the complexities of managing data governance at scale.

The content within this report will also empower executives to improve the case for comprehensive data governance throughout an organization. Most important, executives will learn that by positioning data governance within funded business initiatives, it's possible to build the right data governance capabilities from the start.

Builders—such as data engineers, data architects, data scientists, and data analysts—will also find value in this report. In particular, they will discover relevant insights and practical guidance on data management practices that directly enable their work.

What You Will Learn

After reading this report, you will have a better understanding of the significance of data governance and how it fosters rapid innovation by optimizing your data resources. Data governance starts by identifying one or more business initiatives to justify the data governance program and drive your data governance priorities and decisions.

Readers will learn:

- Why data governance is challenging
- Why working backwards from business initiatives is critical to success, including:
  - The three pillars of good data governance and the capabilities they require
  - The technical enablers needed to deploy a modern data governance strategy
- Which AWS data and machine learning (ML) governance capabilities are available and how they might fit into your strategy

Equipped with this knowledge, you will be able to develop a strategy that provides direct value to your most important business initiatives while simultaneously maturing the data governance program at every step.

CHAPTER 1

How to Think About Data Governance

Data governance involves establishing robust data and process controls, implementing data standards, and employing effective data-handling practices that optimize data utilization to improve business outcomes while minimizing risk. This fosters trust and enables informed decision making across the organization. But knowing *how* to implement data governance is even more important than knowing *what* data governance is.

Promoters of data governance often justify the program by espousing the value of data governance, focusing on areas such as data quality and consistency, data integration and interoperability, and data access and security. This approach is misguided. Instead, it is better to work backwards from important (that is, funded) business initiatives.

It's vital to think about the true purpose of data governance. The purpose of *data governance* is to ensure that data supports business initiatives. It's that simple. But it's also powerful—and often missed. Every successful data governance program starts by attaching itself to one or more funded business initiatives and delivering the required governance for targeted initiatives—not simply by chasing the value of data governance in and of itself. If you stick to this principle, your data governance program will support the most important strategic goals of the company (via funded business initiatives) while also building coherent, organized, and trustworthy data resources in the process.

Let Business Initiatives Drive Your Data Governance Program

Smart companies understand that data governance initiatives should support—not compete with—their business initiatives.

To illustrate the importance of this seemingly subtle difference, let's consider an example. We once worked with an agricultural firm where an internal audit revealed some crucial data governance shortcomings, such as ineffective data quality management, lack of identified data owners and stewards, limited use of a data catalog, and other issues.

While all of these observations were accurate and important, the problems occurred when the company began attempting to close the gaps that were uncovered in the audit. Instead of aligning data governance capabilities to specific business initiatives, the company addressed the gaps directly, planning the implementation of data quality capability, a data catalog, role assignments, and so on. The company had even clearly articulated a business value proposition and respectable return on investment estimates based on closing these gaps.

Although this approach seemed appropriate, the progress of the data governance program was slow, with little sense of urgency and waning executive support. The results were particularly frustrating because data leaders in the organization were in fact following popular advice they had acquired from a variety of experts.

We advised a simple but profound shift. Instead of proposing the value of data governance directly, we advised the company to identify a funded business initiative and align data governance to support it. It turned out that there was an ongoing initiative to transform the company to *precision farming*.

The data governance program approached the leadership of the precision farming initiative and offered to prioritize its data governance work to support the precision farming business initiative.

The team shifted from focusing on the goal of "good data" or a "more mature data governance capability" to supporting the business initiative of precision farming—with dramatic results.

By aligning systematically, there was:

- A new sense of urgency
- Rededicated focus
- Contagious momentum
- Clear understanding of how data governance can support precision farming

These shifts occurred because the data governance program was now considered vital to the success of the company transformation already under way. Each use case delivered for precision farming relied on the data governance program. For example, the program enabled the integrity and quality of key metrics, such as data about yield, soil, weather, and crops obtained through sensors, drones, satellites, and a variety of internal and external sources. The program also identified new data handling policies to ensure that personally identifiable information (PII) about the farms was not inadvertently used at the risk of violating U.S. Department of Agriculture policies. The initiative was a success.

With the data and associated data management capabilities in place, the company was then in a position to support other business initiatives by reusing and extending data resources.

At Amazon, we call this philosophy "working backwards." The idea is to start with a vision of the final business outcome. In the case of the agricultural firm, the final outcome was associated with the precision farming business initiative, *not* data governance. After the data governance program was positioned as a supporting character in the larger play rather than the star of the show, progress was much easier, more focused, and dramatically more valuable.

What Are the Key Challenges with Data Governance?

Gartner emphasizes that successful digital businesses need solid data governance, a framework that Gartner defines as the specification of decision rights and an accountability that ensures the appropriate behavior in the valuation, creation, consumption, and control of

data and analytics.[1] Gartner also names artificial intelligence (AI) risk, trust, and security management as the top strategic technology trend for 2024. In addition, continuous threat exposure and democratizing AI are in the top 10. Each of these are associated with critical aspects of data governance.

Yet a recent Gartner study also predicts that through 2025, 80% of organizations seeking to scale digital business will fail because they do not take a modern approach to data and analytics governance.[2]

With data governance identified as such a critical priority, why does it so often go so wrong? Even when appropriately aligning data governance to support business initiatives, organizations find implementing data governance challenging for several reasons:

- Data grows exponentially in both size and variability, requiring systems to constantly monitor for data quality or model bias changes in real time.
- Data spreads across multiple purpose-built data stores, and getting access to analyze an organization's data is slow and typically not understood by all team members.
- As data is used by more users for more use cases, it becomes challenging to tie a data or model inference to an acceptable use policy.
- Machine learning (ML) models, model features, and data transformations are not always transparent.
- Data workers are not sure they can trust that generative AI models are returning results that align with acceptable use policies.
- Skill set gaps and staff turnover means fewer employees have historical knowledge of proprietary data or acceptable use policies.
- Ethics policies in data governance become more important as ML and AI are more widely adopted.

1 "Data Governance", Gartner Glossary, accessed March 21, 2024.

2 Laurence Goasduff, "Choose Adaptive Data Governance over One-Size-Fits-All for Greater Flexibility", Gartner, April 11, 2022.

- Proving compliance to a regulation includes collecting the appropriate level of information that ties to that regulation, which is time-consuming.

Let's take a look at how these problems manifest themselves in the real world.

Consider Big Finance, a financial services firm in the United States. Big Finance recognized the importance of data governance and therefore has implemented strict access controls to protect client data. However, its approach was to limit data access even in cases where clients expected their financial advisor to have access to the information. In addition, data resources were disconnected across lines of business (banking, mortgage, credit card, etc.), making a holistic view of customer data difficult. As a result, the relationship with the customer was hindered in several ways:

Limited data access
: With restricted access to client data, financial advisors had difficulty gaining access to data associated with the client's transactions across lines of business, even when explicitly authorized by the customer. The client was then required to personally convey detailed information to the advisor or wait for access to be granted to the advisor, wasting valuable time. Customers became frustrated, creating a higher risk of churn.

Inefficient issue resolution and service experience
: Even when data access was granted, financial services advisors had to navigate through multiple systems and rely on manual processes to gather the necessary information to better understand the customer's history, profile, preferences, and so on. This not only hindered the advisor's ability to counsel the client, but it also made it difficult for the advisor to help resolve issues and act on the client's behalf in other lines of business, such as to negotiate better credit card rates in consideration of the total business with the client. Despite their earnest efforts, clients still experienced disparities and had to repeat information when contacting representatives from other lines of business, leading to even more customer frustration.

Missed cross-sell opportunities
: Lacking a holistic view of product data across lines of business, financial advisors had inadequate information to identify potential cross-selling or upselling opportunities when relevant, such as umbrella insurance policies or mortgage refinances, resulting in lost revenue and suboptimal customer experience.

Inaccurate recommendations
: Robo advisors via ML models that had access to inconsistent and siloed data made inaccurate recommendations to the customer. These suboptimal recommendations created a loss of trust for the customer, who felt that the firm did not understand their financial positions or goals, leading to risk in bad financial advice.

Although Big Finance focused on data governance and protecting customer data, the ineffective application of policy and lack of data integrity derailed the ability for financial advisors to adequately advise the client, leaving the company vulnerable to competitors who could.

The Three Pillars of Good Data Governance

To ensure the data is ready for targeted business initiatives, you need to think holistically about what it means to manage the data effectively and what it means to make sure the data is in the right condition to support the business initiatives.

With the data governance program aligned to support business initiatives, we advocate adopting three fundamental pillars of effective data governance:

Curating your data
: Identify and manage your most valuable data sources so you can limit the proliferation and transformation of critical data assets. Also, ensure that the right data is accurate, is fresh, and has sensitive information identified.

Understanding your data
: Enhance your ability to capture and share the context and meaning of your data and ML models through data profiling, data lineage, automated business summaries, data cataloging, and model governance. This context is managed through metadata.

Protecting your data
: Strike the right balance between data privacy, security, and access. Ensure data is protected through data security, data classification, and data lifecycle management.

A modern data governance framework not only facilitates data accessibility and control but also exhibits the following essential characteristics:

Eliminates friction between using data for decisions and avoiding business risk
: Providing your business with automated approaches to discover, understand, request access, and start using data reduces time to value while also adhering to corporate policies.

Improves decision making and reduces time to value
: Providing your business with high-quality data reduces the risk of making incorrect decisions and reduces time to value. This is accomplished by implementing data quality frameworks, utilizing data validation techniques, and establishing data stewardship roles to help provide more accurate, reliable, and consistent data for applications and decision makers.

Increases trust in your data
: Providing your business with a clear understanding of the quality of the data will help them feel confident in making decisions when using that data. This confidence is accomplished through data quality reporting, data definitions, lineage information, and observability into data usage through data transparency. Here again, you manage this information in your metadata.

Reduces data management costs
: Automating data management such as data cataloging and data profiling reduces data duplication, optimizes data storage and infrastructure utilization, improves query performance, and saves valuable time for data engineers. This results in cost savings by reducing fobtaining conse redundant and overlapping data pipelines, minimizing data management activities, and optimizing data storage and processing costs.

Complies with data privacy and data residency regulations
: Obtaining consent for data usage, implementing data anonymization techniques, deleting data, and adhering to data localization requirements supports compliance with legal and regulatory frameworks, such as the Global Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA).

Overall, good data governance combines the right mix of people, processes, and technology to ensure that data is ready to meet the needs of targeted business initiatives.

Data Governance Considerations for Popular Architectures

Let's take a look at how governing data applies to common elements of enterprise data architectures:

Data lakes
: A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. Historically, data lakes bridged the business intelligence capabilities for data warehouses with storage capabilities more suited for ML. It also brought new data governance challenges with schema-on-read, evolving schemas, semistructured datasets, and unstructured datasets. The emergence of data lakes drove the wide adoption of cross-storage technical catalogs and business catalogs to bring together a better understanding of the data that needs to be governed with permissions.

Transactional data lakes
: A transactional data lake is a type of data lake that not only stores data at scale but also supports transactional operations, ensures that data is accurate and consistent, and allows you to track how data and data structure change over time. Transactional data lakes are supported by open-table formats such as Apache Iceberg, Linux Foundation Delta Lake, and Apache Hudi. With this additional support of the transactional capabilities of traditional databases in a data lake, customers can more easily meet regulatory needs such as the right to be forgotten when they need to delete data.

Data warehouses

A data warehouse is a central repository of information that can be analyzed to make more informed decisions. Data flows into a data warehouse from transactional systems, data lakes, and other sources, typically on a regular cadence. Business analysts, data engineers, data scientists, and decision makers access the data through business intelligence (BI) tools, SQL clients, and other analytics applications. Data warehouses enable organizations to maintain a unified and consistent view of highly structured and refined data. By defining data ownership, access rights, and quality standards, data governance ensures that data in warehouses is well managed, making it easier for users to trust and utilize data for applications and business decision making.

Feature stores

Feature stores are purpose-built repositories to store, share, and manage features for ML models. *Features* are inputs to ML models used during training and inference. For example, in an application that recommends a music playlist, features could include song ratings, listening duration, and listener demographics. Features are used repeatedly by multiple teams, and feature quality is critical to ensure a highly accurate model.

Databases

Databases are used to store data of all kinds, using purpose-built and specialized databases, such as time-series, graph, relational, and vector databases. By defining data policies and auditing procedures and adhering to data quality rules, data governance facilitates data sharing among different applications while ensuring compliance with regulations and organizational standards.

Files

Files store data in widely accessible forms or for special purpose applications such as Microsoft Excel, PDFs, and images. Data governance policies set guidelines on data usage, version control, and data sharing to maintain data accuracy and security across these files.

Data mesh

A data mesh is an architectural framework that enables distributed, decentralized ownership. Organizations have multiple data sources from different lines of business that must be integrated for analytics. A data mesh architecture effectively unites the disparate data sources and links them together through centrally managed data sharing and governance guidelines. Business functions can maintain control over how shared data is accessed, who accesses it, and in what formats it's accessed. A data mesh adds complexities to architecture but also brings efficiency by improving data access, security, and scalability.

Let's dive deeper into the three pillars of data governance, starting with the first one: curating your data.

CHAPTER 2
Curating Your Data

Academics define *data curation* as "the act of discovering a data source(s) of interest, cleaning and transforming the new data, semantically integrating it with other local data sources, and deduplicating the resulting composite."[1]

CDOs think of data curation more broadly as the strategic and systematic process of organizing, managing, and maintaining data to ensure the quality, integrity, and usability of data across the enterprise to meet the needs of a variety of business use cases and applications, from basic reporting to advanced ML and AI.

Both parties agree that data curation involves data collection, validation, transformation, storage, preservation, and dissemination. From the practical perspective of the C-suite, however, data curation needs to go beyond preparing data for individual applications. As the vast amounts of data continue to increase, the ability to automate the data curation process effectively at scale has become an increasingly critical factor for supporting modern, complex, cross-functional business initiatives.

This chapter explores the methods for automating and managing data curation. Let's start by looking at what good data curation at scale looks like.

1 Michael Stonebraker et al., "Data Curation at Scale: The Data Tamer System", 6th Biennial Conference on Innovative Data Systems Research (CIDR '13), January 6–9, 2013, Asilomar, California.

The Value of Curating Data

Effective large-scale data curation forms the cornerstone upon which robust data governance practices are built, enabling organizations to establish trust in their data to meet business needs. Achieving this entails implementing data integration pipelines and data quality validation. Effectively managing both of these steps makes it easy to create data products that support and inform business decisions.

For example, we worked with a homestay accommodation company, Big Homestay, to help develop a digital solution that facilitates the booking of both temporary and extended stays. The platform grew rapidly and evolved to become an excellent example of how to curate data at scale.

Big Homestay managed vast amounts of data related to property listings, guest reviews, bookings, and host information. Because the operations of the online, peer-to-peer platform involved hundreds of services, the company wanted an organized and seamless data consumer experience to navigate the platform, supported by high-quality, integrated data across services. Data leaders achieved this by enabling data producers within various business units to curate their own data through automated, cloud-based services while also proactively facilitating data interoperability across the business units.

Curating data at scale provides many benefits, including:

Business agility
: By decentralizing data ownership and building a platform that automates data curation, you can make your data products readily available for business unit use and shareable across units.

Holistic view of the business
: Interoperable and accessible data across business units allows businesses to have a holistic, 360-degree view of customers, products, suppliers, and other business entities. This is essential to enable business initiatives such as cross-channel customer experience or end-to-end supply chain optimization.

Improved data quality
: Automating data curation ensures that organizations have accurate and reliable data. To achieve high levels of data quality and ensure accurate and informed decision making, implement standardized processes, automated data quality rules, monitoring, and alerting.

Reduced manual effort
: Curating data at scale in the modern landscape necessitates automation across multiple dimensions. This includes assessing the quality of data in the context of targeted applications, automatically monitoring the quality of production data against simple and sophisticated rules, and automating data linkage through cross-referencing and standardization.

Capabilities for Curating Your Data at Scale

> *Quality is never an accident. It's always the result of high intention, sincere effort, intelligent direction, and skillful execution. Quality represents the wise choice of many alternatives, the cumulative experience of many masters of craftsmanship.*
>
> —William A. Foster

These wise words from William A. Foster apply to many spheres of life, including the requisite capabilities for curating your data at scale.

Companies need the following capabilities to curate their data at scale:

- Data integration
- Data quality management
- Master data management

Data Integration

Modern data integration capabilities automate enterprise standards regarding data quality, data classification, data profiling, data cataloging, and data observability.

Data integration is the process of achieving consistent access and delivery for all types of data in the enterprise. All departments in an organization collect large data volumes with varying structures,

formats, and functions. Data integration includes architectural techniques, tools, and practices that unify this disparate data for analytics. As a result, organizations can fully view their data for high-value business intelligence and insights.

Zero-ETL is a set of integrations that eliminates or minimizes the need to build extract, transform, and load (ETL) data pipelines. ETL is the process of combining, cleaning, and normalizing data from different sources to get it ready for analytics, AI, and ML workloads. Zero-ETL integrations facilitate point-to-point data movement without the need to create ETL data pipelines. Zero-ETL can also enable querying across data silos without the need for data movement or people and processes to manage data movement.

The benefits of data integration include the following:

- Understanding transactions from across lines of business associated with customers, products, and other entities
- Enabling consistency in data where data overlaps across business areas
- Supporting faster insights from data through improved navigability

Modern Data Community

Scaling data governance capabilities is about not just the technology on its own but also the automation and processes around the capability. For example, to curate data at scale, companies need to automate their data integration capabilities so that multiple personas can produce, consume, and integrate data.

We refer to this distribution of work as a modern data community. In the same way that organizations have uncovered benefits by decoupling and moving from monolith IT to microservices, the modern data community is an organizational and cultural shift from monolithic data organizations to decoupled responsibility. Rather than a single organization (typically IT) being responsible for all data integration, data quality, management of platforms, and the creation of insights, this model pushes responsibility deeper into organizations, thereby increasing autonomy, ownership, and speed.

Data Quality Management

Data quality management is the practice of ensuring that the quality of data is "fit for purpose." This definition is important because describing data quality as "good" or "bad" is not sufficient to identify and prioritize specific data quality issues. While seeking perfection in data quality is admirable, the effort required for these endeavors is often prohibitive. This effort is yet another reason why it's best to prioritize the issues that will hinder the success of funded initiatives if not addressed. It also highlights the importance of relying on modern techniques to automate data quality rules, data profiles, and monitoring and alerting on data value changes.

For data quality to be fit for purpose, you need to consider several dimensions. Of course, the data should be accurate, but it must also be timely, complete, and consistent to the degree required by targeted business needs. Data quality should also be available and understood in order for data consumers to build trust in data. Consider, for example, the demands of anti-money laundering (AML) compliance, where financial institutions detect and report suspicious activities, including security fraud. Having a wide view of related transactions helps to uncover complex money laundering schemes and reduce false positives that occur with more limited information. And being able to identify this activity quickly, with timely data, helps to report suspected laundering to authorities to prevent further damage.

A mature data quality management capability includes automating the assessment of data quality and recommending a set of rules based on data values. In addition, teams can implement shareable data quality rules that apply across their organization that are enforceable at an organizational level. With the issues noted, the team can take proactive steps to correct any data issues at the source where possible. In addition, as the business solution is developed, capabilities for monitoring the in-scope data quality issues should be out of the box and automated. This way, when the business solution is delivered to production, data quality alerts can be sent from the data analytic ecosystem back to data producers to take corrective action.

Master Data Management

Master data management (MDM) is a process for managing data associated with real-world entities including people, places, and things. Master data is often a data product of its own and is shared, referenced, and accessible across many systems. Compared with transactional data, it requires special data management consideration for several reasons. First, while transactional data tracks unchanging events that occur at a specific time and place, such as a sale or an inventory snapshot, master data entities in the real world tend to change over time. Second, master data about the same entity tends to be stored in multiple systems, requiring decisions about which data is most correct and how to combine and reconcile attributes about each entity from across systems. Master data also tends to be organized into hierarchies, such as product classification schemes or customer segments, requiring proactive management of the hierarchies. Lastly, master data is often enriched through activity in transactional data. For example, a customer segment is often inferred based on a customer's purchase transactions from a point-of-sale system.

Reconciling master data may be done in source systems (source of record), in a specific data product system dedicated to the master data (source of reference), or in a combination of both. The decision on where to manage the mastering of data is dependent on the purpose of said data. For example, one may want to use an Internet of Things (IoT) system to manage master data about devices but have a separate system that manages the associated product master. An example of this is when a manufacturer acquires another similar manufacturer with different systems and wants to consolidate raw material purchasing across both companies. The raw material systems are often purpose built for managing the information about that raw material data, and choosing to manage master data in a source can be best for that corporation. Master data management as a data product system on its own, on the other hand, is chosen when, for example, a corporation needs to perform consolidated reporting across a decentralized organization or when data needs to be shared across multiple applications. In this approach, master data is retained in their separate sources, and entities are matched to create a golden record in the analytic environment.[2]

2 Adam Getz, "Master Data Management: The 'Golden Record'", BI-Insider.com, March 12, 2020.

Technology You Need

When building data governance capabilities to curate your data, technology can accelerate the following capabilities:

Data integration
: Data integration is enabled by streaming applications, data pipelines through ETL processes via programming, and low-code or no-code capabilities. With the introduction of zero-ETL and generative AI for data integration, technology is being created to dramatically reduce or even eliminate the need to build complex ETL processes through automation of pipeline development.

Data quality management
: Data quality tools support automation by enabling data profiling using proven statistical and AI techniques to uncover data pattern issues quickly. These tools also automatically suggest data quality rules for ongoing monitoring through catalogs, alerting mechanisms, and custom applications.

Master data management
: Master data management technology supports organizing data into hierarchies, reconciling and cross-referencing data entities, establishing business workflows to manage master data rationally, and enriching entities via transactional data patterns. AI-enabled entity matching capabilities help to determine records across systems associated with the same entity, such as data about the same customer in multiple business units. AI is also used to monitor related transactional data about entities in order to recommend attributes to be added to a given entity.

Special Consideration: Curating Data for Machine Learning

After an organization masters the process of collecting and managing data for analytics, it must also take additional steps to transform the data into a format that is suitable for ML algorithms to process and analyze. Data governance supports ML because it provides quality data used to inform ML models and foundational models (FMs). Additional steps needed to prepare data for use in ML include the following:

Feature selection
: Identifies the most relevant features (variables) in the dataset that contribute to the prediction or analysis task. This step reduces dimensionality and eliminates noise, leading to improved model performance and efficiency.

Data validation
: Validates the prepared dataset to ensure its quality and integrity. This involves checking for anomalies, assessing statistical properties, and verifying the compatibility of the dataset to specifically address ML model requirements.

Data transformation
: Transforms the data to make it suitable for ML algorithms. This may involve encoding categorical variables, scaling numerical features, or applying mathematical functions to derive new features.

Data normalization
: Normalizes the data to ensure that features are on similar scales. This prevents certain features from dominating the learning process.

Feature reuse
: Creates a shared feature store so that new ML models can take advantage of the features built by previous modeling efforts, thus saving time and continuously improving the quality and breadth of shared features for a variety of use cases.

When building a data governance program, it's important to consider the unique requirements of ML. This will reduce the data management and preparation burden on data scientists, allowing them to focus more of their time on the business value of the models they create and less time on the underlying data needed to train and use the models.

Having explored the significance of data governance and its first pillar of curating data, we will now delve into the second pillar: understanding data.

CHAPTER 3
Understanding Your Data

Understanding your data means that you possess knowledge about the data, such as the content, meaning, quality, and how the data is used. It involves not only grasping the raw facts about the data but also recognizing patterns, trends, and relationships among data. When you truly understand your data, you go beyond surface-level knowledge and gain the capacity to draw informed conclusions, make informed decisions, and communicate the significance of the data to others.

Developing and sharing descriptive information about data, or metadata, across the organization can be difficult.[1] Hence, organizations must implement a process to manage metadata as part of data deployment projects rather than as an afterthought. *Metadata* is information that describes and explains data to provide sufficient context, such as the definition, source, type, owner, and relationships to other datasets. For example, when reporting revenue information, it's important to know the sources, definition, calculations, and transformations that led to the revenue report, as well as who certifies its accuracy.

1 *"Metadata"*, Gartner Glossary, accessed March 24, 2024.

Today, understanding your organization's data is crucial for supporting business initiatives:

- Determining the most appropriate data to use for targeted applications and business decisions
- Building trust in using data for business decisions
- Understanding any challenges in the data, including data quality issues and overlaps of data across sources
- Sharing data demographics for use in the design and optimization of data structures

The Value of Understanding Data

As the number of data users across an organization grows, it can become difficult to find and share the best data assets that will drive decisions. Even when the totality of an organization's data is curated and accounted for, it may still be hard to understand what that data means, especially when there are millions of data points in multiple formats and from various sources. Data also changes rapidly, involving new types of data, new schemas, and new values that need to be continually monitored. Hence, organizations must embark on a well-planned journey to truly discover all of the inherent value locked inside the understanding of their data.

With this in mind, we can now delve into the essential capabilities that organizations require to enhance their data understanding and discovery efforts.

Capabilities for Understanding Your Data

Gaining an understanding of data resources across a large and complex enterprise requires a systematic and holistic approach. The core capabilities needed to fully discover and understand data include the following:

- Data catalog
- Data lineage
- Data profiling

In this section, we'll explore these capabilities and their significance for understanding data at scale.

Data Catalog

As organizations grow in size and complexity, effectively managing and utilizing data becomes increasingly difficult. Data is often scattered across siloed systems and stored in inconsistent formats, lacking context and documentation. This makes discovering, understanding, and using the right data challenging.

A data catalog can help address these challenges by providing a central metadata repository that documents datasets. Data catalog platforms apply ML to automatically find, profile, describe, and tag datasets. This enables users to easily search for relevant data using business or technical terms. More sophisticated data catalogs include workflow experiences for searching for data, finding data, requesting access, and working with data from a single interface. By integrating these customer experiences, technology bridges disjointed data consumption steps into streamlined and governed experiences.

Using a data catalog, stakeholders from across the business can find, understand, and responsibly use data. A modern data catalog connects data producers and consumers, breaking down silos. Rather than going on fruitless hunts for data, employees can focus their time on critical analysis to drive real business value.

A modern data catalog helps enterprises do the following:

Discover and explore data
: Search, browse, and explore available datasets based on various criteria, such as keywords, tags, or metadata attributes. This empowers data consumers to quickly find relevant data for their analysis and decision-making processes.

Understand data context and lineage
: Provide detailed information about each dataset, including its source, structure, ownership, usage history, quality, and relationships to other datasets. Modern data catalogs use AI to infer additional contextual information about the data itself, including business names, business definitions, entity recognition, and data relationships. This contextual information helps data consumers have the necessary insights to trust and interpret the data accurately.

Provide visibility for data access
: Provide visibility into data ownership, usage permissions, and data classification, enabling data assets to be managed and utilized in accordance with regulatory requirements and internal policies.

Enable data observability
: Provide observability into the timeliness, freshness, and accuracy of the data. Data integration capabilities and streaming systems all output information about the performance, runtimes, freshness, successes, and failures. Modern data catalogs present this information to data consumers so they are aware of any real-time system status, earning additional trust in the data they are using.

Promote data democratization and self-service analytics
: Enable business users and analysts to find data, request access, and start to perform self-service analytics. The data catalog allows these users to independently locate and access relevant datasets without needing extensive support from IT or data engineering teams. Additionally, data catalogs enable data owners to approve or deny access requests from a single place.

Emerging Topic: Data Observability

Data observability provides an understanding of how data is used and when it was created. This includes monitoring, execution times, runtimes, usage patterns, and access patterns. As with lineage, this data is valuable for earning trust in how data is used and how fresh it is, as well as for bringing transparency to data delays or processing errors so that data consumers understand the risks when using data in their business decisions.

A well-implemented data catalog helps organizations gain control of their data not just by listing the available data but also by adequately describing the data in detail and enabling easy authorization and access of data for use in applications and analytics.

Data Profiling

Data profiling is the process of examining, analyzing, reviewing, and summarizing datasets to gain insight into the data for data quality analysis and to provide input to ML, data structure design, data quality management, and database optimization. Data profiling involves systematically scanning and analyzing source data to understand its structure, content, and interrelationships. The results of data profiling are also metadata.

Data profiling uses advanced statistical and AI techniques to find patterns and anomalies in data, which can be reviewed by data stewards to determine whether they are legitimate data values or represent potential data quality issues that could hinder success of targeted business initiatives. For example, analysis of customer survey data may indicate inconsistencies across question responses indicating lack of attention to the questions, thus requiring caution when using the data for real business decisions. Data scientists use data profiles to understand data skew, valid values, histograms, and data structures to make more informed decisions on their feature engineering. Data modelers examine data profiling output from identified data sources to uncover data values that were not communicated by end users during data modeling sessions. This information is used to design optimal data structures, guiding choices around indexing, partitioning, and aggregation.

Data Lineage

In large enterprises, data flows through extensive pipelines, undergoing many transformations between raw sources, target applications, and analytics. This can occur in real time and in batch processes. While this processing creates useful data, it also buries its origins under layers of models, scripts, execution history, and hops between systems. This data and metadata falls under the category of data lineage and observability. As previously discussed, metadata has value in helping improve understanding of data. The metadata about data pipelines and transformations have this same intrinsic business value. For example, a transformation used to calculate a metric used on a financial statement needs the same level of trust as the data it produced. This is accomplished through data lineage.

Documenting and exposing data lineage empowers users with essential context to understand the processing the data goes through before being used in an application or to make a decision.

With data lineage, users can examine inputs and outputs at each processing step, examining how data elements were sourced, filtered, modified, or aggregated. Gaps and inconsistencies causing broken lineage expose potential reliability issues. Shared lineage also allows project teams to determine the ripple effects of planned changes in data sources on target systems.

Multiple end personas benefit from well-managed data lineage:

Analysts
: Identifying where data came from and how it was transformed allows analysts to interpret data trustworthiness for specific uses.

Data scientists
: Viewing and tracking the flow of data as it moves from source to destination helps data scientists better understand the quality and origin of data for use in ML models.

Data engineers
: Proactively analyzing changes in pipelines allows data engineers to identify a job's upstream dependencies and downstream usage and thus better evaluate impacts and instigate needed changes to connected systems.

When enabling data lineage, organizations include the results in the descriptive metadata shared through the data catalog, showing the full context of their data by including the true relationships between data sources, targets, and the steps along the way. Data lineage is also metadata.

Technology You Need

Modern technology supports understanding your data from multiple angles, accelerating your ability to capture and share this understanding across the enterprise. Technologies for understanding data include the following:

Data catalog
: This capability helps data producers to automate the definitions of data in order to easily share data. This allows for data consumers to easily find, evaluate, request access, approve access, and access the data they need for business use.

Data profiling
: This capability must support sophisticated statistics, text and image analysis, and visualization to identify data issues and gain an understanding of the data content for a variety of purposes. Typically, data profiling capabilities are built into data integration tools, data quality tools, data lakes, and database systems.

Data lineage
: Data lineage tools automatically document data flows from source through pipelines to end users. Lineage tools capture information about data pipelines themselves, including source structures, transformation rules, hops across data stores, and information about the execution of the pipelines such as run times (start, stop, duration), any errors, and statistics on data volumes processed.

Now that we've looked at the value of discovering and understanding your data, let's take a deeper look at the last pillar of data governance: protecting your data.

CHAPTER 4

Protecting Your Data

Protecting your data requires a delicate equilibrium between enabling data access for informed decision making and implementing robust data controls to protect sensitive information. Balancing data access and control is essential to establishing trust while ensuring compliance so that data consumers have quick and easy access to the data they are authorized to use. It also helps prevent the introduction of additional business risk.

Let's take a look at the key considerations and strategies that empower organizations to achieve this balance of easy access with effective control.

The Value of Protecting Your Data

Organizations store and protect sensitive data, including PII about customers and employees and other internal information such as preliminary financial data or detailed competitive plans and supporting analysis. If this data is improperly disclosed, it can create mistrust and risk the imposition of fines or, at minimum, damage the company's reputation or market position. Yet data must be used every day for the company to function. This includes supporting sales staff with comprehensive information about the customers and publishing timely and accurate financial statements.

Because sensitive data must be used widely for business purposes without excessive delay *and* the data must be safeguarded from inappropriate use, data governance must have efficient process,

procedures, policies, and supporting technologies so that people and groups can be easily authorized to use data for their job functions.

While information security, compliance, and privacy departments possess deep expertise in security practices, regulations, and tools, they need partnership with data governance to develop and apply policies to specific domains of data and to make decisions about the applicability of regulations to these domains. For example, the privacy department brings expertise about what data to classify for PII, data handling policies, and acceptable use policies within the organization. An example policy is to always mask PII data unless it is the customer accessing the data from within a secure application. It is the job of the technology teams to automate definition of the policy for masking, classifying sensitive data, and ensuring the policy is auditable.

Capabilities for Protecting Data While Balancing Access and Control

A comprehensive data governance program includes a full set of security capabilities in partnership with InfoSec and other groups throughout the organization.

The core capabilities that form the foundation for sharing data responsibly include the following:

- Data security
- Data compliance
- Data lifecycle management

Data Security

Data security is the practice of granting data access to the right users and protecting data against corruption and theft.

Security is a broad concept that covers identity and access management, infrastructure protection, data protection, logging and monitoring, and incidence response. For the purpose of this book, we are focused on data security, data compliance, and data lifecycle policies.

This concept encompasses a comprehensive range of information security responsibilities, including:

- Maintaining the physical security of hardware and storage devices (and ensuring cloud providers offer this capability) as well as the logical security of software applications
- Formulating and implementing organizational policies and procedures to ensure data protection
- Enabling identity and access management, including identity system management and integration of identity systems with system authentication and data authorizations
- Managing role-based access controls and purpose-based access controls
- Providing protection not only against cybercriminal activities but also against insider threats and human errors

To achieve this level of security, data security involves deploying tools and technologies that enhance an organization's visibility into the location and usage of critical data. These tools should be equipped to apply various protective measures, such as authentication, authorization, encryption, and data masking. Additionally, automation of reporting streamlines audits and ensures adherence to regulatory requirements.

Data security also includes development of data classification policies that define levels of sensitivity (e.g., public, internal, confidential, highly confidential) and the appropriate protections and tools to secure data at rest and in transit at each of the sensitivity levels. Data owners then use their data curation processes and data catalogs to apply these classifications to their domain of data by determining the sensitivity level their domain (customer, product, sales, etc.) and associated data elements belong to. Next, working with partners in business and IT, they apply the appropriate security practices to the data.

To classify data within the data curation process, we recommend that customers fully automate their data classification policies across all their structured and unstructured data. This typically requires leveraging technology that uses statistics and ML to identify the data and classify it accordingly. Some examples include PII, Payment Card Industry (PCI) data, Health Insurance Portability and

Accountability Act (HIPAA) information, or inappropriate data that should be moderated.

Data Compliance

Data compliance is the practice of following government regulations to ensure that sensitive data is managed in accordance with the public interest and any other interests reflected in regulations.

Regulations can be broadly applicable, such as the GDPR from the European Union (EU), which mandates protection of personal data by, for example, allowing EU citizens to correct personal data and to be made aware of how their data is used. The California Consumer Privacy Act (CCPA) provides similar protections for California residents.

Other regulations may be narrowly focused. For example, the Genetic Information Nondiscrimination Act (GINA) is a US law that prevents insurers from using genetic information to make decisions about a person's eligibility, coverage, underwriting, or premium costs. It also bars employers from making hiring, firing, promotion, or any other employment decisions based on a person's genetic information.

Any modern data governance program must stay up to date on these regulations as they emerge and evolve and determine the applicability to their data assets.

Data Lifecycle Management

A fundamental driver for data lifecycle management is appropriately safeguarding data as it ages.

With vast data growth, persisting everything indefinitely in active primary storage incurs unnecessary costs, yet often data must be retained for long periods for infrequent business use or for regulatory compliance. For example, according to the U.S. Department of Labor, under the Fair Labor Standards Act (FLSA), employers must maintain records for a period of at least three years. Thus, defined lifecycle policies enable automatically transitioning less business-critical data into cost-efficient secondary storage tiers that still provide data retention and protection. Disposing of data that is no longer needed for any purpose reduces vulnerability for stale datasets well past usefulness.

Advanced tools to automate policy-based progression of data across cost-optimized tiers with integrated protection controls provide a governed path that balances accessibility, security, and budget throughout every phase of data's journey to disposition. The sophistication of today's lifecycle capabilities shifts the narrative. Rather than risky data bloat, aging data can reliably and economically be managed to responsibly mitigate threats while avoiding the consequences of needless retention beyond established requirements.

Having explored the essential capabilities for protecting your data, let's delve into the technology needed for data security and compliance.

Technology You Need

To build the data governance capabilities to protect your data, you need technology that supports proficiency in each capability area:

Data security
: Modern data security solutions provide unified visibility and fine-grained access controls that are deeply integrated with data catalogs and data consumption/production technologies such as data warehouses, ETL tools, SQL engines, and data processing engines like Apache Spark. Fine-grained access controls are often integrated with policies for role-based access controls, purpose-based access controls, encryption, masking, and monitoring based on tagging policies as data traverses multiple services and repositories.

  In addition to authorizing data access, organizations use advanced behavioral analytics and ML algorithms to detect credential misuse, abnormal queries, unauthorized access attempts, ransomware activity, and insider threats in real time and to trigger automated responses.

Data compliance
: Technology to support data compliance continuously audits your system usage to help assess risk and compliance with regulations and industry standards. Automated tools help you store and access compliance reports in a self-service portal.

Data lifecycle
: Data lifecycle management technology provides cost-effective storage and retrieval along with automated movement of data based on data access and archival requirements and changes in requirements over time. This type of automated data movement is used for common data storage, such as object storage, and also within specialized storage technologies, such as purpose-built databases.

Special Consideration: Managing Data Permissions in Data Lakes

As organizations build expansive data lakes, the challenge of efficiently managing permissions becomes increasingly critical. Since numerous users and tables are involved, data stewards and administrators require an effective approach to handle permissions at scale.

To ensure data is shared securely and responsibly, organizations must address several challenges:

Fine-grained permissions
: When sharing data, organizations often need to grant access at a granular level, allowing each user to view only specific rows, columns, or even individual cells within a table. *Fine-grained permissions* enable organizations to control data access more precisely, ensuring that sensitive information remains restricted while still facilitating data sharing for legitimate purposes.

Scaling permissions across users and data stores
: As the number of users accessing shared data increases, managing permissions at scale becomes complex. Organizations need to implement scalable solutions to efficiently assign and revoke permissions across a growing user base by authorizing access to individuals, roles, and groups, and access rights should propagate beyond the data lake into data warehouses, other data stores, business intelligence, and ML tools.

Sharing data internally
: Sharing data within the organization requires a well-defined framework for granting access to relevant teams and personnel while ensuring that sensitive data is appropriately protected. Internal data-sharing facilitates collaboration and data-driven decision making across departments.

Sharing data externally
: Sharing data with external parties, such as partners, vendors, or customers, demands additional considerations for data privacy and compliance. For example, capabilities such as clean rooms enable analytics across company boundaries without compromising privacy of customer data.

Auditing
: Regular auditing of data-sharing activities is vital to monitor access, track data usage, and identify any potential security or compliance breaches. Auditing helps maintain data governance standards and provides insights into data-usage patterns, supporting improved risk management.

Make sure the technology that protects your data has robust access controls that can implement fine-grained permissions, scale user access, and enable secure data-sharing practices to foster trust and collaboration while maintaining data security and compliance standards.

Now that we've covered the importance of protecting your data, let's look at how AWS enables data governance.

CHAPTER 5
How AWS Enables Governance

As organizations work to drive business initiatives with data and insights, they learn to balance data access and control by curating, understanding, and protecting their data. They also learn how critical data governance is to their business agility. In this final chapter, we will explore some of the technologies that support and facilitate the three pillars of data governance: curating, understanding, and protecting your data.

In this chapter, we will focus on a subset of AWS services, centering around Amazon DataZone. Amazon DataZone is a data management service that helps you catalog, discover, share, analyze, and govern your data. It is a unified solution, which means you won't have to assemble individual services to implement data governance.

Amazon DataZone allows customers to define data ownership based on what works for them. This enables data ownership to be centralized to one single team or decentralized to many global teams. When ownership is distributed, each department or the analytics team maintain their business glossaries so that data consumers know that they are using the right data.

Amazon DataZone works with other Amazon applications, streamlining the ability to curate, understand, and protect your data. For example, to curate data, Amazon DataZone is built to enable data users to launch analytical services—such as Amazon Athena, Amazon QuickSight, Amazon Redshift, and Amazon SageMaker—so they can analyze data and determine which new data products to create.

It's also important to reduce your time to insights by finding the right data in the right context. Using the right data requires understanding the data context. Amazon DataZone is built on top of the technical catalog in AWS Glue, which includes structured and semistructured datasets on Amazon Simple Storage Service (S3) and Redshift. The data catalog is a centralized metadata repository for data assets across various data sources. It enables users to store and query information about data formats, schemas, and sources.

Additionally, Amazon DataZone extends the technical data catalog from AWS Glue into a business data catalog. The business data catalog adds business context and includes business glossaries. As a result, Amazon DataZone automates adding business descriptions and names to data, which helps users easily understand context and avoid dealing with cryptic technical names. The addition of this metadata also makes data more discoverable. This automation is powered by large language models (LLMs), which increase accuracy and consistency. Data quality results are automatically presented so that users understand the health of their data and know whether they can trust it.

Lastly, Amazon DataZone protects data centrally to ensure that the data can only be accessed by the right people and in the right context. In order to protect data, access controls need to be provided in the context of the purpose in which data will be used, and data owners should approve that access based on the purpose and user role. Amazon DataZone accomplishes this by providing access to data via data subscriptions to a project. Groups of users can also collaborate on various business use cases that involve publishing, discovering, subscribing to, and consuming data assets. When a user requests access to a dataset, the data owner can approve that data access. Once approval has been granted, Amazon DataZone automatically fulfills and manages permissions on top of AWS Glue databases and tables and Amazon Redshift tables.

While previously we talked about how Amazon DataZone brings these experiences together, we will discuss how individual services help customers address key data governance challenges in the following section.

Curating Your Data

AWS aims to help you curate data at scale to limit data sprawl. Curating your data at scale means identifying and managing your most valuable data sources, including databases, data lakes, and data warehouses, so you can limit the proliferation and transformation of critical data assets.

Data Integration

AWS provides users with flexible and purpose-built solutions for data integration challenges. For example, if a user is looking for a serverless data integration service, they can use AWS Glue. AWS Glue makes it easier to discover, prepare, move, and integrate data from multiple sources for analytics, ML, and application development.

With AWS Glue, you can use the appropriate engine for any workload based on the characteristics of your workload and the preferences of your developers and analysts. AWS Glue supports multiple data integration engines, including Apache Spark, Python, and Ray. Recently, AWS Glue began using generative AI so you do not need to be an Apache Spark or SQL expert to build data integration pipelines. Instead, you can build data integration pipelines using natural language. Describe your intent through a chat interface, and AWS Glue will generate a complete job.

For customers that want more flexibility for tuning their workloads, they can build their data integration on top of engines such as EMR Serverless for Spark, Amazon EMR on EC2 for Apache Spark and SQL (Trino), or Amazon EMR on EKS. With Amazon EMR, you can build applications using the latest open source frameworks. Each provides different tuning options and infrastructure control to maximize the performance of your data integration workload.

AWS also provides a set of zero-ETL capabilities. AWS's capabilities in zero-ETL cover federated query across data sources without having to worry about data movement. Athena's federated query enables you to analyze data or build applications from an S3 data lake and 30 data sources, including on-premises data sources or other cloud systems, using SQL or Python. In addition, AWS provides capabilities in streaming ingestion. A zero-ETL streaming integration with a data warehouse lets you ingest data from multiple such streams and present it for analytics almost instantly.

Regardless of the engine of choice for your data integration workload, customers can centrally develop these data pipelines with AWS Glue and then orchestrate, manage, and monitor their data integration pipelines with managed workflow for Apache airflow.

Data Quality

Data lakes may become data swamps without proper oversight. Setting up data quality checks is time-consuming, tedious, and error prone. You must manually create data quality rules, write code to monitor data pipelines, and alert data consumers when data quality deteriorates. Data quality inside of AWS Glue reduces these manual quality measures. It will automatically compute statistics, recommend quality rules, monitor performance, surface data quality results to Amazon DataZone, and alert you when it detects issues. It uses these statistics to recommend a set of quality rules that checks for freshness, accuracy, integrity, and even hard-to-find issues. You can adjust recommended rules, discard rules, or add new rules as needed. If it detects quality issues, AWS Glue also alerts you so that you can act and remediate them.

For hidden and hard-to-find issues, AWS Glue uses ML algorithms. The combined power of a rule-based and ML approach, along with the serverless, scalable, and open solution, enables users to deliver high-quality data to make confident business decisions. Additionally, AWS Glue is serverless, so you can scale without having to manage infrastructure.

Master Data Management

You probably have MDM solutions for specific types of data, like customer data in a customer relationship management system. Your IoT devices are probably managed in an IoT-specific solution. Your health care and life science data are managed differently so that you can store, query, and analyze genomic, transcriptomic, and other genomics data. AWS has services to help solve various data management challenges.

For example, to help users manage customer profiles, AWS provides a purpose-built service with Amazon Connect. Amazon Connect provides customer service agents automated experiences with real-time access to up-to-date customer info in order to personalize each customer interaction. Users can simplify how they provide contact

centers with customer information by automatically syncing data from disparate applications (e.g., CRMs) and databases across their enterprise into a unified profile. Then, they can accurately surface the right profile by matching customer identifiers (e.g., phone number, email address) in real time to improve automated IVR and agent experiences. Using generative AI, Amazon Connect automatically maps your data from disparate sources using 80+ prebuilt connectors from third-party applications such as Salesforce, ServiceNow, Zendesk, and homegrown applications.

With a complete view of relevant customer information in a single place, using generative AI to combine contact history information (e.g., transcripts, customer sentiment) with customer information (e.g., current product orders, mobile app interactions), customer service agents can provide more personalized customer service, addressing issues faster and improving customer satisfaction. For example, you can add drag-and-drop flows to access customer information and then automate interactions and contact routing. As a customer call comes in, Amazon Connect automatically scans and matches the customer phone number or customer ID against customer information from connected applications. It then surfaces a unified profile to the agent or IVR, directly in Amazon Connect, using the drag-and-drop workflow designer.

Amazon Connect uses AWS Entity Resolution to detect similar profiles based on similar name, email address, mailing address, and phone number and consolidates them into a unified profile. AWS Entity Resolution helps users match and link related records without building custom solutions. ML and rule-based techniques optimize record matching based on business needs. AWS Entity Resolution helps companies easily match, link, and enhance related records across multiple applications, channels, and data stores to improve the quality of their data so that they can better understand and engage their customers. Amazon Connect uses AWS Entity Resolution to match and deduplicate customer records to build a unified view of each customer, leveraging data from data providers such as LiveRamp and Transunion.

AWS Entity Resolution is not limited to just customer profile information. AWS Entity Resolution helps users match, link, and enhance related, product, business, or health care records stored across multiple applications, channels, and data stores. It's possible to use flexible and configurable rules, ML, or data service provider

matching techniques to optimize records based on business needs. For example, users can match information such as SKUs, UPCs, or proprietary product IDs into a unified product record to improve tracking across stores and supply chains.

Understanding Your Data

AWS aims to help you discover and understand your data in context to accelerate data-driven decisions. Understanding your data in context means that all users can discover and comprehend the meaning of their data so they can use it confidently to drive business value. With a centralized data catalog, data can be found easily, access can be requested, and data can be used to make business decisions.

Data Catalog

To reduce your time searching for data, you need a data catalog as part of your metadata management capabilities. For a business data catalog, you can use Amazon DataZone. For a technical catalog, you can use AWS Glue. The technical catalog in AWS Glue is a catalog of structured and semistructured datasets on S3 and other supported relational and NoSQL data stores. AWS Glue serves critical features to search data sources, databases, tables, Sagemaker Feature Store, and statistics. AWS Glue also connects disparate data sources by loading metadata or federating data source's catalogs to a central place to be searched. Finally, you can automate capabilities to improve query engines performance and provide a framework to build permissions that are based on schemas.

After data is populated and users access that data, you must maintain schemas, optimize data structures, and identify changes in data. The changes could be schema changes, new sensitive data, data value changes, and data velocity. These changes need to be monitored and automated to ensure your users have a seamless experience when using data to create business value. AWS Glue also manages changes to your data structures such as schemas, partitions, and indexes. You can use a crawler to populate the technical catalog in AWS Glue with tables.

Data Profiling by Analyzing Data Context

Now that you've established a robust data catalog to organize and manage your data assets, it's equally important to delve into the content and connections within your data to derive valuable insights.

To do so, AWS Glue offers automated column statistics that provide Athena and Amazon Redshift the necessary information to optimize query plans and provide users an understanding of the data profiles.

AWS offers a variety of capabilities that derive contextual meaning from data. For example, AWS Glue sensitive data detection has transforms that identify PII in your data. The Detect PII transform provides the ability to detect, mask, or remove entities that you define or that are predefined by AWS. This enables you to increase compliance and reduce liability. For example, you may want to ensure that no PII exists in your data that can be read, and you may want to mask social security numbers with a fixed string (such as xxx-xx-xxxx), phone numbers, or addresses.

Amazon Comprehend also offers trust and safety features that help organizations moderate text content. Amazon Comprehend toxicity classifies text across seven categories, including sexual harassment, hate speech, threat, abuse, profanity, insult, and graphic. Amazon Comprehend prompt safety classifier enables moderation of generative AI input prompts to prevent inappropriate use of generative AI applications. Lastly, Amazon Comprehend PII detect API can prevent PII data leak by redacting all personal information from generative AI output. Amazon Comprehend PII detect can mask 22 universal PII entities, like address, age, credit card number, etc., and up to 14 country-specific entities like US social security number, CA health number, passport number, etc.

In addition to Amazon Comprehend, AWS provides a set of AI services that help with data classification of unstructured data such as Amazon Transcribe, Amazon Textract, and Amazon Rekognition. These AI services provide support to understand the context within unstructured data to help you identify inappropriate or unwanted content inside of PDFs, videos, audio, images and other data types.

Protecting Your Data

AWS aims to help you protect and securely share your data with control and confidence. Protecting your data means being able to strike the right balance between data privacy, security, and access. It's essential to be able to govern data access across organizational boundaries, with tools that are intuitive for both business and engineering users.

Permissions Management

To avoid data duplication, you want to add permissions on your tabular data using database-like features, which enable access control at the database, table, column, and row level of the data. You also want to be able to manage the access request workflows to enable data producers to easily inspect and approve or reject data access requests. Finally, you need to track data interactions by role and user and provide comprehensive data access auditing to ensure that the right data was accessed by the right users at the right time. Amazon DataZone has robust permissions management capabilities across multiple data sources.

Data Lifecycle Management

Lastly, all of the AWS Analytics services work in concert to help customers support data lifecycle management. For most customers, their data lake data is stored on S3. S3 provides a variety of data lifecycle management capabilities to store data in the right storage tier. S3 supports the eleven nines of durability to ensure your data is protected. The Amazon S3 Intelligent-Tiering storage class is designed to optimize storage costs by automatically moving data to the most cost-effective access tier when access patterns change. Each of these work with to support transactional data lakes to meet customer's data lifecycle needs.

Apache Iceberg is a distributed, community-driven, Apache 2.0-licensed, 100% open-source data table format that helps simplify data processing on large datasets stored in data lakes. Apache Iceberg supports transactional data lakes. A *transactional data lake* is a type of data lake that not only stores data at scale but also supports transactional operations, ensures that data is accurate and

consistent, and allows you to track how data and data structure changes over time.

Governed Data Collaboration

To securely collaborate with partners on collective datasets without sharing or copying one another's underlying data, you can use AWS Clean Rooms. AWS Clean Rooms enables customers to define custom, mutually agreed-upon queries that will run on their collective datasets. AWS Clean Rooms provides differential privacy features to protect the privacy of customers. Differential privacy is a leading privacy-enhancing technique that works by adding a controlled amount of randomized "noise" to aggregate query results. The noise protects the data so the users can directly understand the population of the individual records without exposing individuals.

Governed Data Sharing

Centralized teams govern data sharing, but that centralized work can lead to slow processes, data access sprawl, and reduced time to value. Amazon DataZone makes it easy to support a data mesh for data lake and data warehouse data across your organization.

ML Governance

When you need a purpose-built governance tool to help you implement ML responsibly, you can use SageMaker. Your administrators can define minimum permissions in minutes. When you need to capture, retrieve, and share essential model information, such as intended uses, risk ratings, and training details, you can use SageMaker Model Cards. When you need to stay on top of model behavior in production, you can use SageMaker Model Dashboard.

You can use ML governance tools in SageMaker to generate customized roles that allow ML practitioners to work with SageMaker faster; streamline model documentation and provide visibility into key assumptions, characteristics, and artifacts; quickly audit and troubleshoot performance for all models, endpoints, and model monitoring jobs; and track deviations from expected model behavior.

When you need to make FMs available through an API so you can choose the FM with the best fit, you can use Amazon Bedrock. You can also privately customize FMs using your organization's data and use familiar AWS capabilities to secure generative AI applications.

Conclusion

Choosing data governance tools is a critical aspect of building a successful data-driven organization. With a comprehensive data governance strategy that embraces the three pillars—curating your data at scale, understanding your data, and protecting your data—you can ensure the integrity and accessibility of your data assets.

How do you know which capabilities you need? Work backwards from the capabilities required by your funded business initiatives. Build expertise on those capabilities over time and leverage the tools and services that can grow along with you. Embracing data governance is not only a strategic advantage but also an ethical responsibility as organizations handle the data that shapes the future of businesses and society alike.

About the Authors

Kevin Lewis is a global practice principal for AWS Professional Services. In this role, Kevin enables a consulting capability that guides customers through the entire data analytics journey, from planning and organizing to building and deploying. He brings nearly three decades of expertise in managing large-scale data initiatives. Prior to joining AWS, Kevin was director of data strategy and governance at Teradata. His career began at Publix Super Markets, where he started as a bagger in high school and, several steps later, led Publix's data management program, spearheading the planning and development of an enterprise-wide data analytics ecosystem.

Jason Berkowitz is the head of product for AWS Lake Formation, AWS Glue Data Catalog, and AWS Glue Crawlers. As part of this role, he holds a leadership position in defining AWS's data governance strategy, partnering with product leadership from Analytics, Amazon DataZone, AWS Glue, and the AI and ML teams. Prior to leading products, Jason delivered analytics and ML technology and strategy consulting services to Fortune 500 organizations when heading up the Global Data Analytics Practice in AWS Professional Services.

Ina Felsheim is the head of product marketing for data governance and Amazon DataZone. She brings nearly three decades of expertise in defining successful data management initiatives. She previously ran data management initiatives at SAP, including data governance, analytics, cloud databases, data quality, metadata management, and more.

Joseph D. Stec studied English Literature at Miami University. He has worked on the American Stock Exchange in New York City and for DataRobot. He currently lives in Poland and writes books and articles related to history, data, and AI, including *Lost Legends of the Black Sea* (2018) and *Why External Data Needs to Be Part of Your Data and Analytics Strategy* (O'Reilly, 2022).