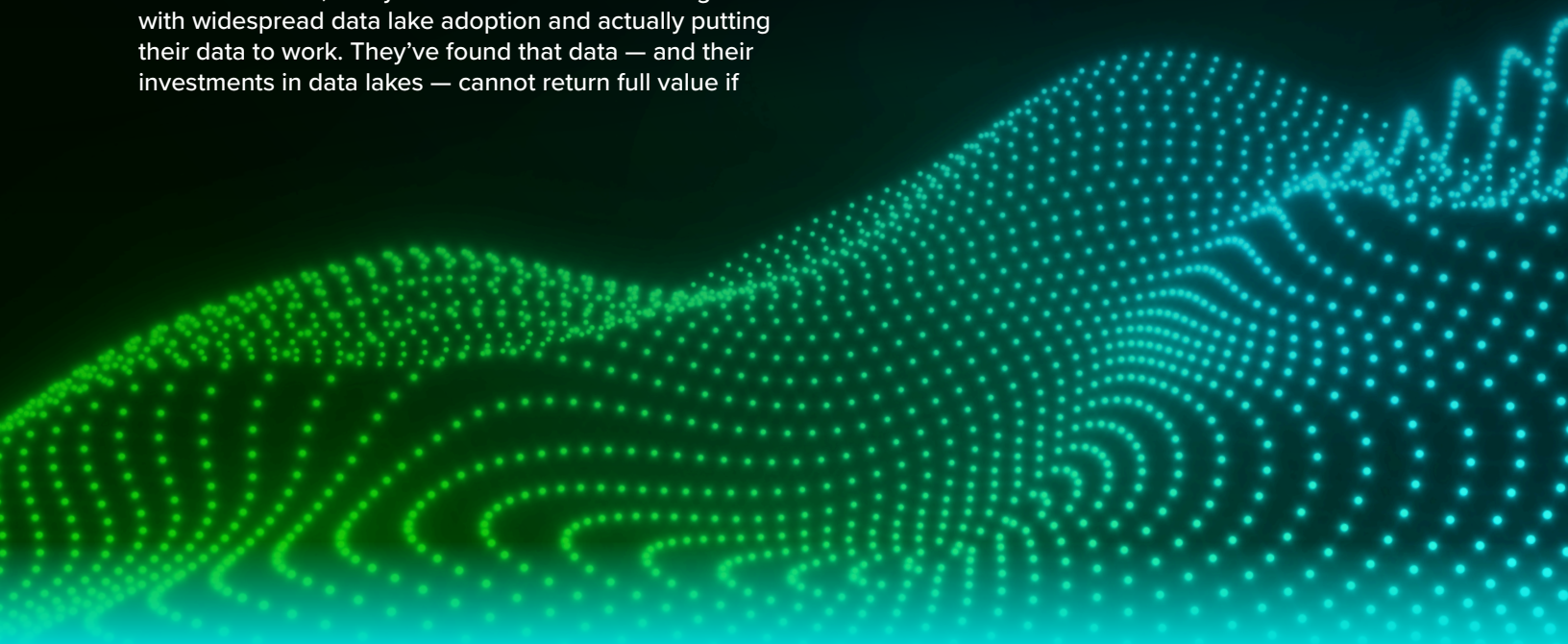
Why Government Agencies Need Data Lake Governance

Government agencies are under mounting pressure to extract value from the ever-expanding troves of data they generate, collect and store. When properly used, data enables better decision-making and reporting; seamless, timely and personalized constituent experiences; and intelligent automation.

Moreover, in light of the Evidence-Based Policymaking Act,[1] Congress proved how much the government could benefit from using data evidence to create policies and inform programs. To ensure adherence to the Act, agencies will need to establish a comprehensive data management strategy enabling them to easily account for all the data assets and metadata (data about data) created by, collected by, under the control of, and maintained by each agency. This will allow better intra- and cross-agency collaboration (sharing of data) and increase data literacy (understanding of data).

While agencies have invested in data lakes as a step toward storing, processing, analyzing and sharing large amounts of data, many have encountered challenges with widespread data lake adoption and actually putting their data to work. They've found that data — and their investments in data lakes — cannot return full value if users cannot easily find and understand the data they need or ensure the data is trustworthy and of high quality. In a highly regulated environment like government, data lake adoption requires rigorous checks and balances to ensure data is easily findable, traceable and that it meets the highest data quality thresholds to derive value from it in a timely manner.

Data lake management and adaptive governance unlocks the value of data by making data ecosystems accessible, transparent and trustworthy for all users – including data producers, data scientists and data consumers – which allows them to connect data to various use cases and make information actionable. This in turn fosters greater data lake adoption across departments, agencies and jurisdictions.



Data lake management and governance makes data more consumable and trustworthy by enabling organizations to protect data more systematically and build processes that define who owns what data, how it can be used and more.

Stuck in the Mire

The following challenges are driving agencies to pursue a governed data lake solution.

- **Data swamps.** Data lakes are a relatively inexpensive way to store data, making them ideal repositories for both structured and unstructured data. However, as the volume and types of information grow, a data lake can very quickly turn into a data swamp without rigorous data management practices in place.

For instance, users often complain of not being able to quickly find the data they need or understand key attributes about the data – including who owns the data, where it originated and whether it's from a reliable source. Users may not know whom to contact if corrections are needed or what data-sharing agreements exist. Data integrity issues such as poor-quality inputs are another deterrent to more widespread adoption. Often, data that resides in data lakes does not meet adequate data quality thresholds. It may be incomplete, incorrect, outdated or redundant. Leading industry analysts such as Gartner predict that poor quality data can cost an organization an average of $15 million.[2] This does not account for the cost or losses associated with decisions that are based on inaccurate information.

With an ever-growing emphasis on transparency, auditability and traceability of data as well as reducing waste while optimizing constituent experiences, agencies need to implement a rigorous data management and adaptive governance strategy to ensure they can derive maximum returns from their investments in data lakes.

- **Ineffective collaboration.** Agencies and jurisdictions within a single discipline (e.g., health and human services, law enforcement or transportation) can gain efficiencies and achieve better outcomes for their communities by collaborating and sharing data. However, each entity typically maintains its own data sets. This siloed approach leads to duplications of effort, data redundancies, inconsistencies and a lack of standardized definitions across entities — all of which prevent effective collaboration.

- **Inconsistent policies and enforcement.** As the number of data sets grows and multiple groups create their own data sets, as well as internal user and data protection policies, there are likely to be inconsistencies. These inconsistencies increase the risk of breaches and non-compliance with both internal policies and government regulations for data privacy and protection.

- **Low adoption rate.** Even when agencies encourage departments, groups and individuals to move their data to a data lake, they often run into issues with users actually using it. Data citizens (i.e. IT, data stewards, data governance leads, data scientists and data consumers) have concerns about compliance breaches and control over their data. For instance, they may find it difficult to locate the right data, understand its chain of custody, know whether they can trust the source and add new data to data sets.

Data Lake Governance: Clearing the Way to Actionable Insights

Data lake management and governance make data more consumable and trustworthy by enabling organizations to protect data more systematically and build processes that define who owns what data, how it can be used and more. With a transparent view into this information, users can be confident they have high-quality data for internal collaboration, data sharing across ecosystems and artificial intelligence-fueled decision-making.

For example, a state department of transportation can pull real-time and historical data from multiple regions to make transportation safer and more efficient. Each region can onboard its current data catalog into the shared data lake while still owning, controlling and managing their data. The individual region also defines what its data sets mean. For instance, a region can define its ridership for public busing and share what it's doing when it comes to ridership. Data consumers — for example, a county health and human services (HHS) agency — can access and use that data alongside their own data sets within the data lake to see how they can help HHS clients more easily access public transportation. Similarly, the county HHS agency can share and access data across other ecosystems to promote a unified "whole person" approach to health, housing, childcare, transportation and food benefits.

Foundational Components of Data Lake Governance

The following capabilities are foundational to data lake governance.

- **Native connectivity.** Native integration and connectivity allow organizations to register commonly used data sources and external source systems so they can quickly and easily ingest metadata from them into the data catalog.

- **Embedded governance and privacy policies.** Embedded policies help ensure compliant use of data by governing access to the data lake and providing transparency around user roles, data ownership and data workflows. They provide a standardized way to define data, establish ownership, implement and enforce policies, observe behavior and ensure data integrity as data moves from one system to another. Under this framework, users know the data that moves from the original source system into the data lake has gone through a well-documented process that ensures the right security controls are in place, that reports are generated from the correct sources and more. This knowledge drives trust and encourages data lake adoption by producers and consumers of data.

- **Data catalog.** The data catalog is a repository of your metadata enabling you to build a single source of truth for all your data. It provides visibility into metadata in the data lake. The metadata includes rich business and technical context enabling data citizens (data producers and consumers) to obtain granular details about the data.

- **Automated data classification and lineage.** These processes automatically classify physical data assets by type and sensitivity so organizations can consistently apply and enforce user policies and more easily comply with regulatory requirements. They also help data users understand the technical and business lineage of data by providing visibility into how data transforms and flows from source to destination. This capability is important for evaluating whether the data is trustworthy and for analyzing and resolving the root causes of any data issues.

- **Federated approach.** A federated approach to cataloging and governing data enables states, cross-agency collaborators and others to set up their own area within the data governance platform for loading their catalogs and technical metadata. Under this approach, groups or individuals within HHS have a compartment for their own metadata, as do groups within housing, HR, finance and more. This allows agencies to use their own business terminology and policies for data access and data sharing. Besides providing structure and access control, this approach also enables organizations to leverage their roles, responsibilities and workflows in a way that fits with how they operate.

- **Automated workflows.** Automated workflows allow agencies to more easily assign and track data steward roles and responsibilities; accurately document, manage and report on data and controls; remediate and enrich metadata; and more. Increasing data transparency in these ways makes it easier for all data citizens to find, understand, trust and contribute to data sets so they can spend more time working with data to drive business decisions.

Benefits of a Well-Governed Data Lake:

- **Builds transparency and trust** in data by providing clear visibility into data lineage, stewardship, quality and meaning, as well as enabling proactive remediation of data issues

- **Improves user experience,** productivity and decision-making by helping data scientists, data governance leads, data stewards and citizen scientists find and examine the right data faster

- **Provides a framework for collaboration** by allowing organizations to communicate about their data at the metadata level and minimizing the need to reveal details about the content itself

- **Mitigates compliance and cybersecurity risks** by applying data policies and controls consistently across all data sets and simplifying reporting

- **Streamlines compliant access** to data by aligning requests with appropriate privacy controls

The Infrastructure Investment and Jobs Act and other federal funding are adding impetus to agencies' plans to invest in data lakes and data lake governance so they can use data to plan for and report on projects, improve compliance, respond with greater agility to disruption and better prepare for the future.

If a citizen scientist doesn't see a term, data source or policy they're looking for in the data set, they can easily propose an addition by filling out predefined fields. They don't need to know the name of a technical column they want to add to or who to contact to onboard the addition. The automated workflow routes the proposed addition to the appropriate data set owner, and the end user can track the process. Users can also rate data sets or add comments, which provides an engagement path for both professional data scientists and non-technical users.

Best Practices for Deriving Value from Your Data Lake Investments

The following best practices will help ensure the success of a data lake governance initiative.

- **Zero in on the business case.** Be explicit about the use case, the problem(s) to be solved and the desired outcomes. Identify priorities for leveraging data to reach goals. For example, the business case could be using public ridership data to plan where to expand or reduce service or where to locate satellite public health clinics.

- **Clarify organizational structure and roles and responsibilities.** Identify subject matter experts, data scientists and other individuals or groups that own data, know about data, or have a data-related role such as report writer or report consumer. Understand and document their role around data and where they fit in the organization. This task can be done separately from any identified use case and helps shift the culture around data's importance and each individual's stake in it.

- **Start with a small, concrete process.** It's usually best to start by focusing on a small set of data sources rather than the entire data lake. One initial process could be creating a glossary that defines commonly used terms in a housing program application process (e.g., "adult," "child," "household" or "income") and establishing a searchable repository of data sets needed to approve the applicant. Once completed, other teams can leverage this information for their agencies' processes. HHS can document differences in their definitions and add their own data sets needed in their processes. Delivering specific business value or mission value quickly helps create buy-in and builds momentum for further development.

- **Align on key performance indicators (KPIs) and metrics.** Work with stakeholders to identify what KPIs the agencies want to influence by using data, and then identify metrics to measure progress toward KPI goals. Track and report progress, refine processes as needed and then expand from there.

Ready to Launch

The Infrastructure Investment and Jobs Act (IIJA) and other federal funding are adding impetus to agencies' plans to invest in data lakes and data lake governance so they can use data to plan for and report on projects, improve compliance, respond with greater agility to disruption and better prepare for the future. Data lake governance is a nonlinear journey. Whether an organization first addresses data lineage, compliance, quality or another aspect of data governance, the key is to get started. The sooner agencies do so, the sooner they will realize the full value of their data.

This piece was written and produced by the Center for Digital Government Content Studio, with information and input from Collibra.

1. https://www.congress.gov/bill/115th-congress/house-bill/4174
2. https://www.forbes.com/sites/forbestechcouncil/2021/10/14/flying-blind-how-bad-data-undermines-business/?sh=5aebc04029e8

Produced by: gt CENTER FOR DIGITAL GOVERNMENT

The Center for Digital Government, a division of e.Republic, is a national research and advisory institute on information technology policies and best practices in state and local government. Through its diverse and dynamic programs and services, the Center provides public and private sector leaders with decision support, knowledge and opportunities to help them effectively incorporate new technologies in the 21st century. **www.centerdigitalgov.com.**

Sponsored by: Collibra

Since 2008, Collibra has been uniting organizations by delivering trusted data for every use, for every user, and across every source. Our Data Intelligence Cloud brings flexible governance, continuous quality, and built-in privacy to all types of data **www.collibra.com.**