# The governed data lake approach

*Expand self-service data access to accelerate analytics and actionable insights*

## Highlights

- Provide users across the organization with fast, self-service access to a full range of data
- Accelerate analytics and improve the accuracy of insights
- Store massive volumes of data—in its native format—within a cost-efficient, scalable environment
- Maintain data quality, security and governance

Organizations today collect a tremendous amount of data and are bolstering their analytics capabilities to generate new, data-driven insights from this expanding resource. To make the most of growing data volumes, they need to provide rapid access to data across the enterprise. At the same time, they need efficient and workable ways to store and manage data over the long term.

A governed data lake approach offers an opportunity to manage these challenges. A data lake is a shared data environment that comprises multiple repositories and capitalizes on big data technologies. Organizations are increasingly exploring the data lake approach to address demands for an agile yet secure and well-governed data environment that supports both structured and unstructured data.

Unlike a data warehouse, a data lake uses a flat architecture that keeps data in its native format until it's needed. It allows rapid landing and storage of data, and it provides ready, unfettered self-service access to data for analysis. Comprehensive governance capabilities help ensure data can be easily found, understood and stored without duplication (Figure 1).

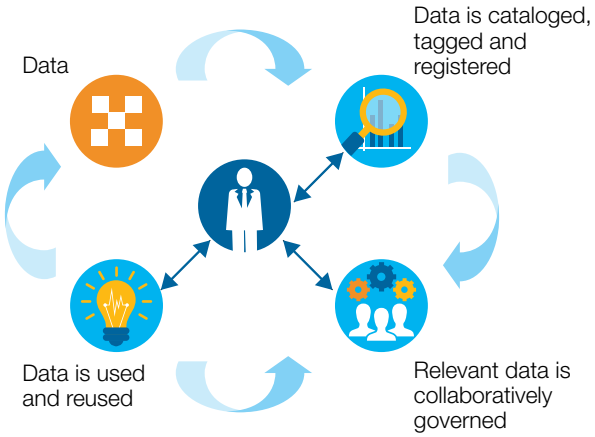


*Figure 1*. The components of a governed data lake efficiently process and discover data to make it easier for users to use it as the foundation for analytics, reporting and decision-making.

## Address key challenges with a data lake

A data lake can play a critical role in helping organizations address several strategic business and IT challenges:

- Enabling timely access to a full range of data in a timely fashion for business users, analysts, data scientists and developers so they can generate accurate, meaningful insights.
- Accelerating analytics and data preparation to keep up with the speed of business.
- Facilitating collaboration between knowledge workers, data scientists (when deeper analytic capabilities are required) and data engineers (when it's time to deploy data lake–based applications to line-of-business users).
- Ensuring data quality, security and governance to provide users with trusted, understandable data while protecting privacy and maintaining compliance with regulations.
- Accommodating a rapidly growing collection of data with a scalable approach to storage that supports fast-growing data volumes cost-effectively while maximizing existing resources.

## Avoid a data swamp

The most effective data lake incorporates a robust set of components that actively ingest, govern, protect and manage data. Planning is critical for optimizing the data lake and avoiding a "data swamp." A data swamp is just what it sounds like—murky, and lacking the organization, visibility and governance required to fully capitalize on all available data.

In a data swamp, users are not sure of the origin of data or its purity. They lack the ability to find the data they need or even know if it is present in the repositories. Additionally, users don't know whether data is adequately protected. Furthermore, they would typically not have the business context for data.

### What a data lake is and isn't

**A data lake is:**

- An environment where users can access vast amounts of raw data
- An environment for developing and proving an analytics model, and then moving it into production
- An analytics sandbox for exploring data to gain insight
- An enterprise-wide catalog that helps users find data and link business terms with technical metadata
- An environment for enabling reuse of data transformations and queries

**A data lake is not:**

- A data warehouse or data mart for housing all of the data in an enterprise
- A replacement operational data store (ODS)
- A high-performance production environment
- A production reporting application
- A purpose-built system to solve a specific problem (though a purpose-built data mart could be fed from a data lake)

## Implement best practices to optimize the data lake

Taking the following steps to incorporate best practices can help organizations optimize their data lake and avoid the pitfalls of a data swamp.

### Build data repositories

The repositories provide structured and unstructured data to the organization. Each repository either supports unique workload capabilities or offers a unique perspective on a collection of data. Multiple repositories can hold different types of data. Organizations should be able to easily add new repositories and remove obsolete ones as necessary.

### Implement data lake services

Data lake services control and support access to the data lake repositories by analysts, data scientists, developers and business users. These services also keep copies of data in sync. By including a catalog of data, data lake services enable people to locate the data they need and verify that it is suitable for their work. The data catalog also provides data lineage for validating the source of the data, thereby bolstering the confidence users have in the data.

### Create a comprehensive information management and governance model

Data lake services should be supported by specialized middleware that provides information management and governance capabilities. The middleware should include:

- Provisioning engines for moving and transforming data
- A workflow engine to enable collaboration among individuals working with the data
- Monitoring, access control and auditing functions

## Realize a wide range of benefits with a governed data lake approach

When implemented correctly, a data lake can help organizations efficiently extract real business value from their data environment. Examples of benefits from creating a data lake include:

- **Easier data access to a broad range of data across the organization:** With a governed data lake, users can access structured and unstructured data residing both on premises and in the cloud. They can access what they need, when they need it, without sending time-consuming requests to IT.
- **Faster data preparation:** A data lake can accelerate data preparation in several ways. For example, implementing a catalog for the data helps increase knowledge and understanding of the data, which in turn accelerates data preparation. In addition, building a data lake with a hybrid cloud infrastructure gives organizations the ability to store data on the platform most appropriate for its use. When data is placed in its ideal location, it takes less time to access and locate it, thereby speeding up data preparation and reuse efforts.
- **Enhanced agility:** Faster data preparation lets users explore more. Components of the data lake can be employed as a sandbox that enables users to build and test analytics models with greater agility. They can experiment with analytics and—in some cases—"fail fast" to move on to the most productive avenues more quickly.
- **More accurate insights, stronger decisions:** By providing access to more data, accelerating data preparation and letting users experiment with data, a data lake can help organizations generate more accurate insights. A well-constructed data lake can also track data lineage to help ensure data is trustworthy. All of these capabilities help organizations make better business decisions.

At the same time, the governed data lake can provide some important IT benefits. For example, a data lake approach can help IT groups prepare for continued data growth. Incorporating Hadoop into the data lake architecture can give IT a highly scalable, affordable environment that is capable of analysis at scale (though possibly not at speed). In addition to supplying the resources to answer pressing business questions, the data lake can also provide a query-able archive for low-touch data at an attractive cost. By building a data lake on a hybrid cloud environment, organizations can add resources quickly while reducing capital expenditures.

In many cases, a data lake can also free up expensive enterprise data warehouse (EDW) resources so the EDW can better perform other duties, such as supporting business analysts in monitoring and analyzing historical performance data. The data lake enables self-service analytics by fulfilling data requests without affecting the EDW's service-level agreements.

## Enable fast analytics and nimble business decisions

A governed data lake offers a powerful approach to capitalizing on the massive influx of data available today. By following some key best practices for constructing a governed data lake, organizations can provide ready access to a wide range of data across the enterprise while helping ensure data is trustworthy and secure. When optimized for an organization's particular needs, a governed data lake can play a central role in enhancing business agility and improving decision-making.

## For more information

To learn more about the benefits and opportunities associated with successful data lakes, visit: ibm.biz/data_lake



IBM Analytics
Route 100
Somers, NY 10589

Produced in the United States of America
July 2016



IME14033-USEN-00