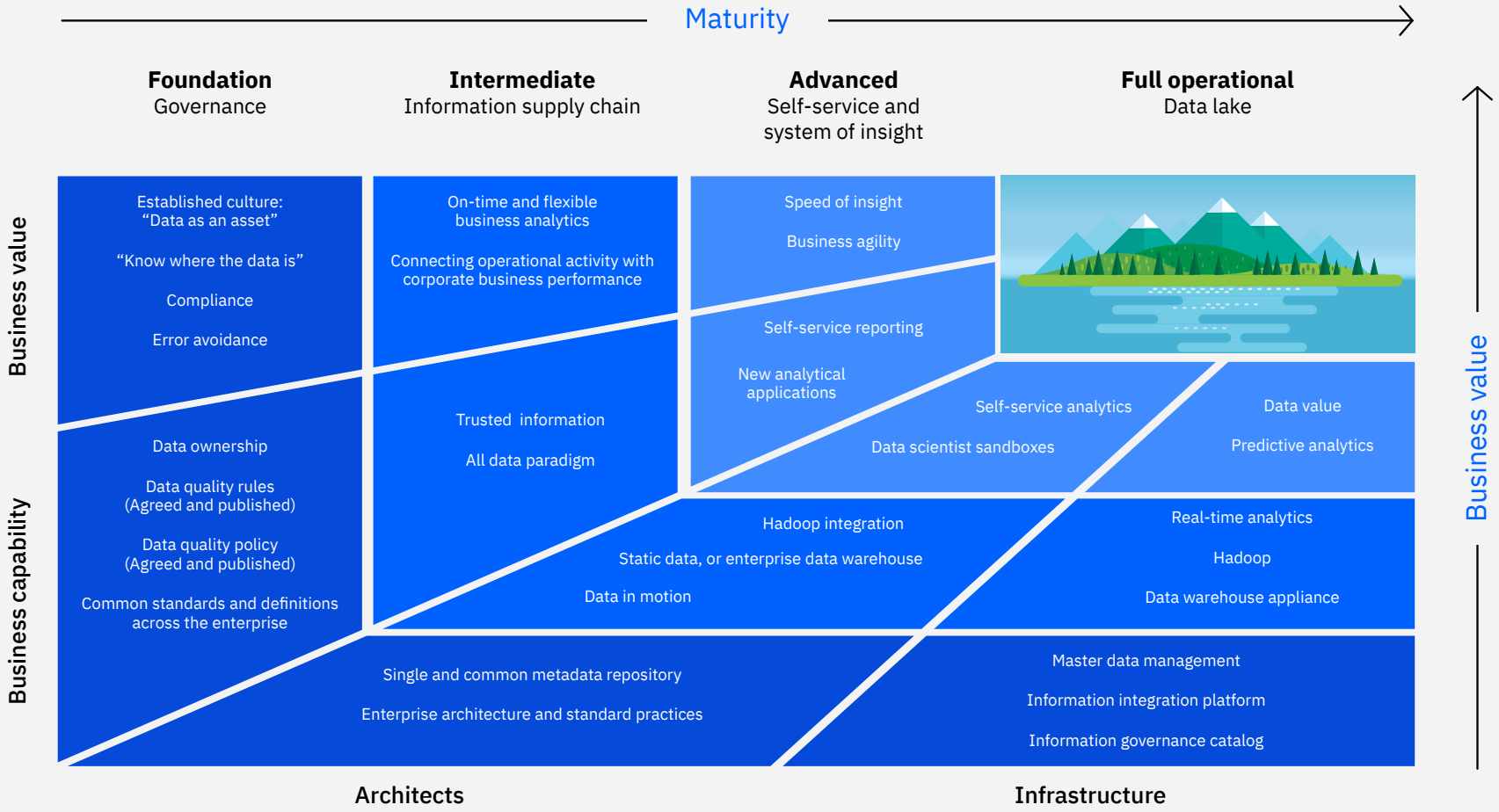
Governed data lake for business insights

Explore the key building blocks to effectively deliver trusted data

IBM **Cloud**

Governed data lakes add value

Data lakes are ideal solutions for organizations that prioritize data in their operations strategy. Secure data sharing is a crucial factor when multiple teams require access to enterprise data. To help manage that use, organizations can rely on a governed data lake that houses raw structured and unstructured data—trusted, secured and governed. For organizations that derive value from their data, including data about customers, employees, transactions and other assets, governed data lakes create opportunities to identify, understand, share and confidently act upon that information.



Architecture of a governed data lake

Key design decisions characterize the architecture of a governed data lake. Three main parts comprise a data reservoir. Data lake repositories provide platforms that store data and run analytics as close to the data as possible. Data lake services locate, access, prepare, transform, process and move data in and out of the data reservoir repositories. Finally, the information management and governance fabric helps govern and manage the data in the data lake.

Governance capabilities validate and enhance the data quality and are designed to protect the data from misuse. This measure ensures the data is refreshed, retained and eventually removed at appropriate points in its lifecycle.

Governance, the organization of your data and the ability to have confidence in its quality, is an important aspect of managing a data lake. While a data lake is designed to offer flexible access to data, you need a system of governance to ensure the data is security-rich, protected and continues to be useful. The governed data lake can be illustrated by its layers, as follows:

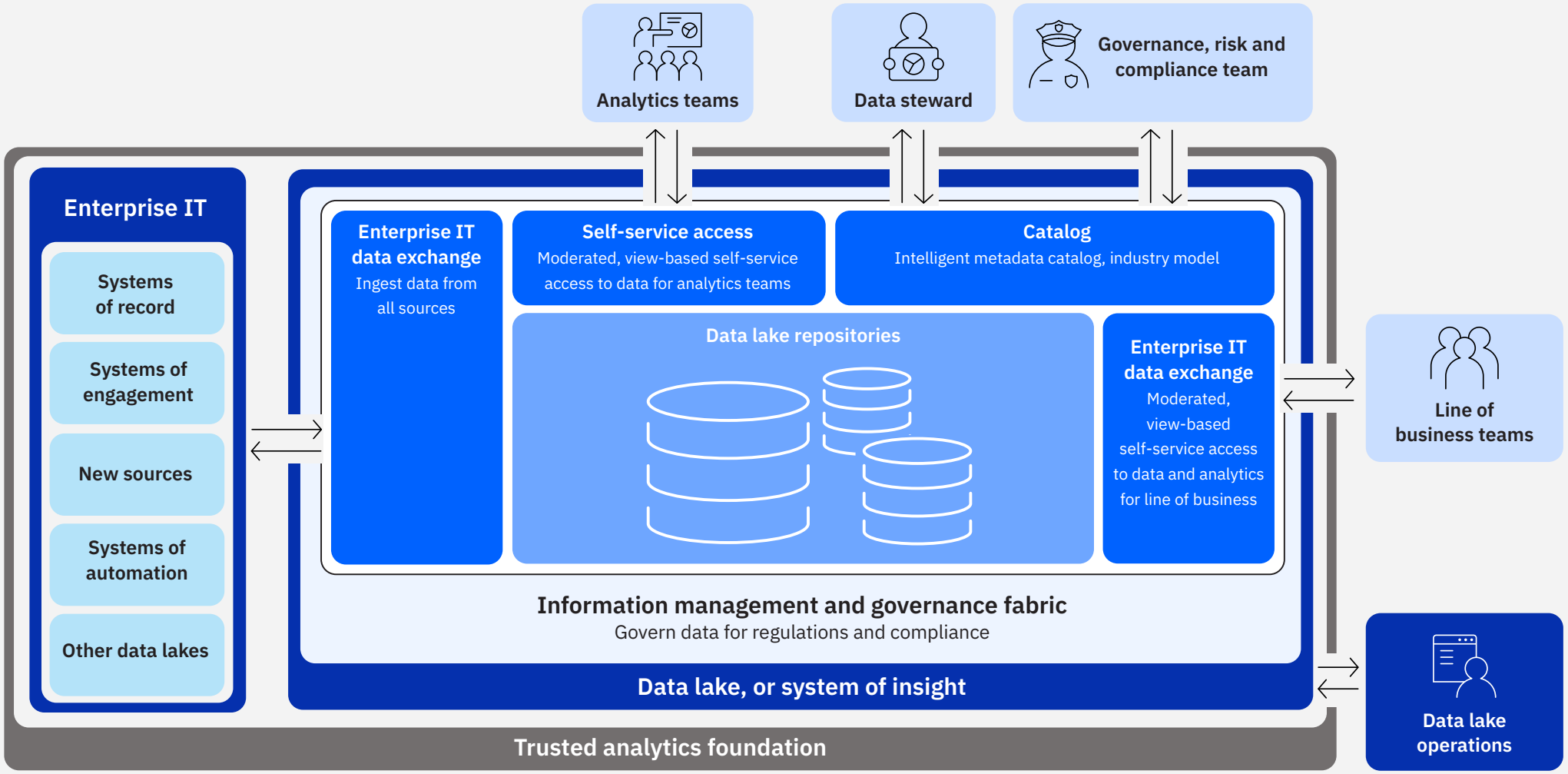
- Foundational, primarily based on data governance
- Intermediate, which expands the initial data lake repositories with new and additional data types and data behavior
- Advanced, which enables self-service analytics

Each layer holds specific value for different groups in the organization. Architects benefit from a published reference architecture, supported by a single and common repository of metadata. Data scientists benefit from a controlled area where they can deposit in-progress sandboxes.

The foundational benefits of a data lake are derived from governance. Governance drives a “data first” culture where business users take ownership of data and agree on rules and policies. Shared definitions create mutual understanding, which helps you avoid confusion among or between teams. With this common ground, you can access trusted data and accelerate insights from analytical applications. The business value shifts from awareness about data and its importance, to flexible analytics at any time.

A modular, scalable data lake consists of several elements that encourage self-service access throughout your organization.

Architecture of a governed data lake

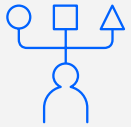


The four types of data consumers

Users who consume data from the data lake vary in key ways. Understanding the difference between their approaches to data is an important aspect of successful governance.

Analytics teams

- Data scientists who manage data and build models
- Analytics developers who turn models into applications
- Application developers who incorporate analytics applications into operational systems



Data steward

- Optimize data quality and prepare ETL jobs
- Catalog Data and perform metadata management
- Strike the balance between data protection and privacy

Governance, risk and compliance team

- Data governance specialists who build data governance and security policies
- Protect data to ensure privacy controls are enforced in all processes
- Compile retention, archival and disposal requirements and ensure that data is compliant with policy and regulations

Line of business teams

- Line-of-business (LOB) executives like CMOs, CFOs, or CHROs
- Chief data officers who are emerging as business owners of data
- LOB executives who implement systems for specific business outcomes or actionable insights

Building blocks of a governed data lake

A governed data lake is a reference architecture independent of specific technology that includes governance and management processes. It's not Hadoop or an enterprise data warehouse that you can buy or replace. A governed data lake is an on-premises or cloud-based solution for organizations that want to put data at the core of their operations. The building blocks of a governed data lake include the following elements:

Enterprise IT data exchange can extract, analyze, refine, transform and exchange data between data lakes and enterprise IT systems, and move it from data puddles to data lakes. It cleanses data and monitors data quality on an ongoing basis.

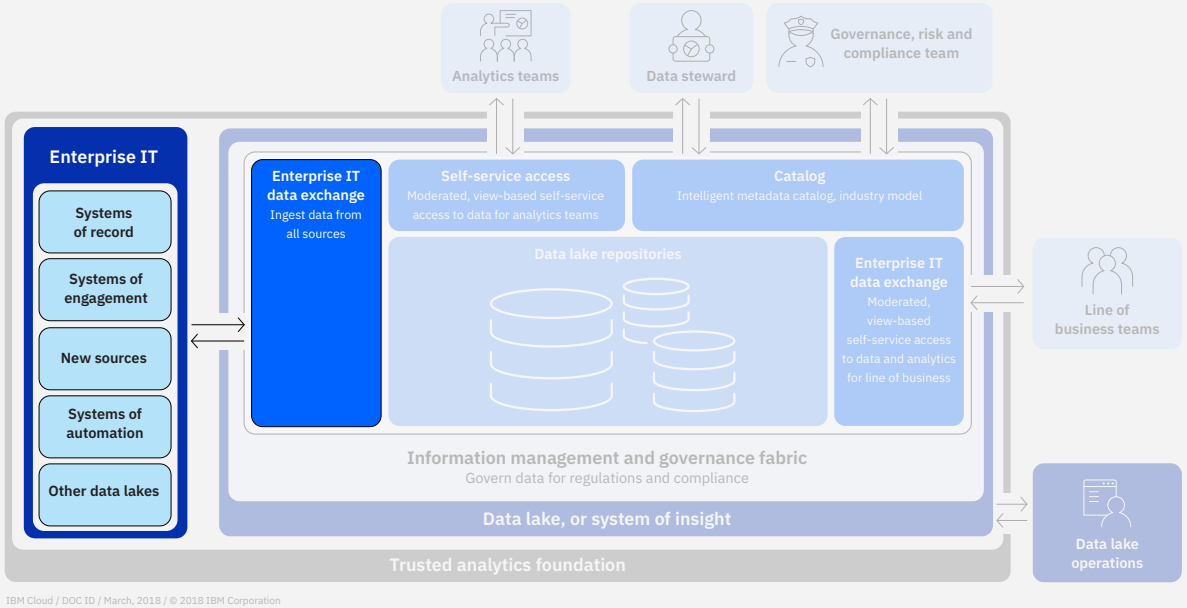
Catalog services describe the data in the data lake—what it means, how it's classified and the resulting governance requirements this places on the data.

Governance helps to govern the data in the data lake and apply appropriate policies, security, data quality and privacy to the data stored in the lake.

Self-service access consists of three sets of services that provide on-demand access to the data lake. Self-service access for analytics users allows access to raw data as it's stored. For LOB teams, the service provides normalized data in simplified data structures. For governance and risk and compliance teams, the service provides governed data for audits.

Data ingestion from various sources

Ingestion is the process of extracting, transforming, quality processing and exchanging data between the data lake, the enterprise IT systems and other existing data lakes. Much of the data in the data lake comes from the organization's IT systems. These data types can be structured, semi-structured or unstructured. Data sources can be systems operating the business, a website log, or other sources that monitor activity. IBM® offers the scalability on volume of data and the richness of transformation and replication.



When done and done right
– Flow data into the data lake without interruption
– Analyze data that's transformed, standardized and enriched
– Reduce storage costs even when data volumes increase
– Use a sandbox for exploratory analytics

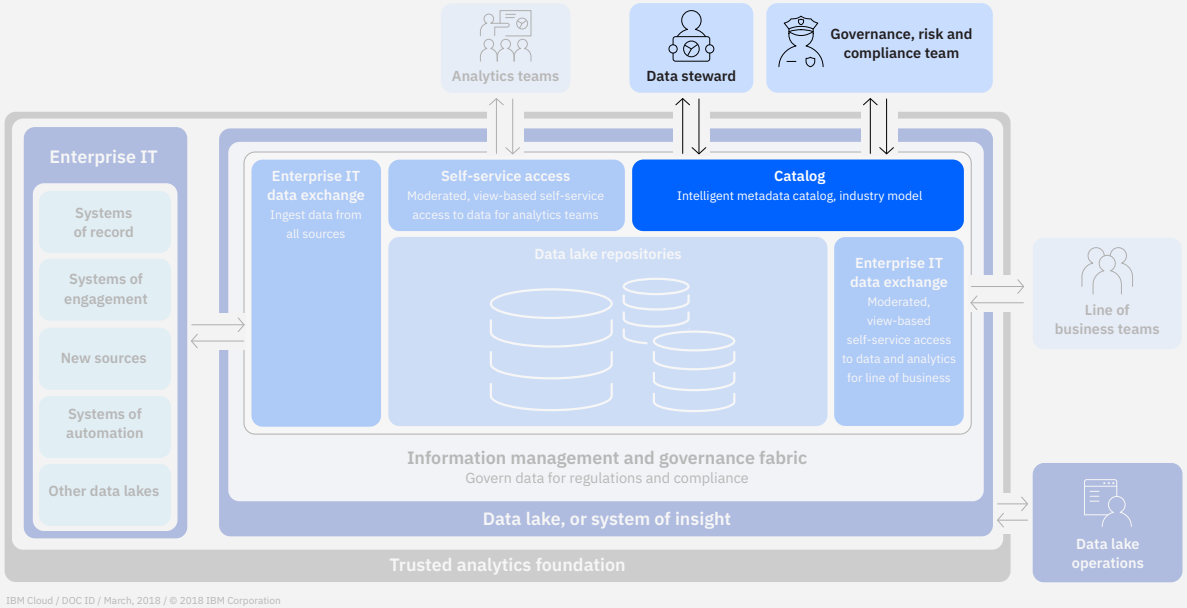
When not done or done incorrectly
– Struggle to keep data fresh as volumes increases
– Lose the ability to use unstructured information assets
– Pay higher storage costs
– Complicate data cleansing, which leads to higher data processing costs

Cataloging

Cataloging helps to tag the data in the data lake and create an inventory of information assets. The catalog interfaces provide data lake users with information about the data within its classification, lineage and how it's governed. IBM offerings in data lake cataloging include the following features:

- Allowance for unstructured information assets to be captured in the catalog
- Open ecosystem integration with virtually any information asset
  - One enterprise catalog for virtually all information assets in the organization
  - Industry-specific data and business terms enablers
  - Rating capability and social tagging as part of the metadata

Data brought into the governance pipeline must be understood, so technical data makes sense from a business point of view. For example, a nine-digit number might be a US social security number or an employee ID number, or both. The step of classification and business term assignment adds business meaning to technical data. Automation is a key attribute to make this process scale to meet the volume and variety of data in the lake. Then, curation workflows, quality assessment and data controls ensure data can move to cataloging, which makes this data available across the enterprise.



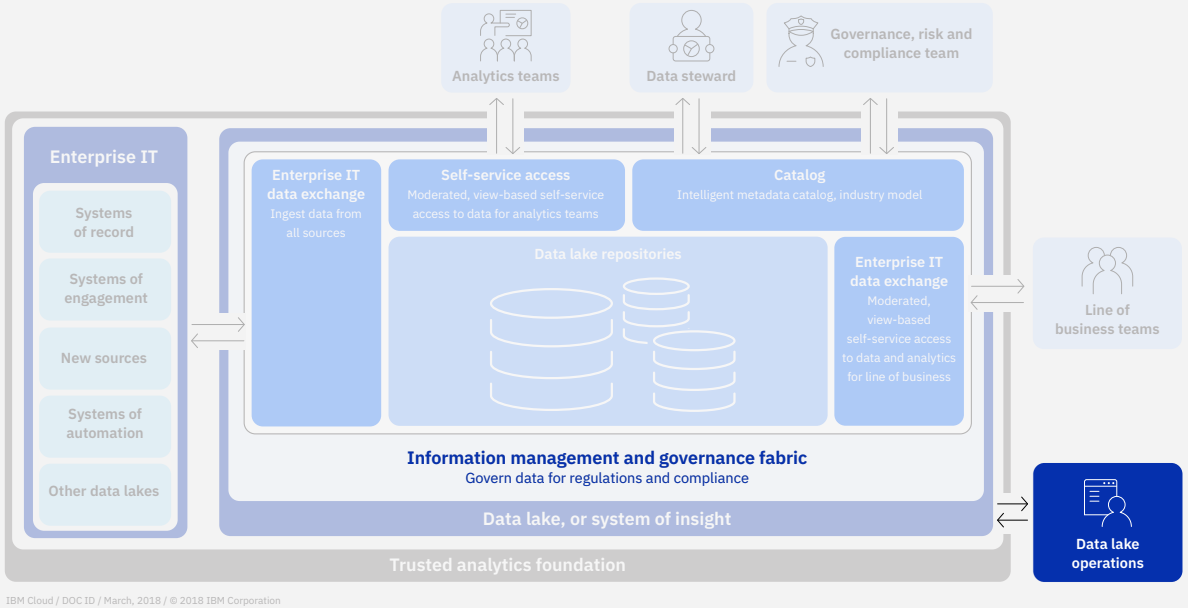
When done and done right
- Increase time to results and more time to analzye data
- Capture contextual asset knowledge and improve data usefulness
- Track data lineage and improve trust in your data
- Market information assets for broader consumption
- Assist with data compliance

When not done or done incorrectly
- Risk wasting time searching and tagging data
- Lose tribal knowledge when you locate data, but can't find colleagues who understand the data
- Lose knowledge of who has access to data
- Miss compliance and governance requirements

Govern and manage your data

Information integration and governance fabric provide the ability for the system to effectively track your data lake so that incoming information is understood and governing policies are automatically applied. The governance framework helps to document governance policies and enact rules to help you define how information should be structured, stored, transformed and moved.

The requirements for information governance are documented in the catalog as policies, rules and classifications. The key IBM differentiators are that unstructured assets can be part of the data lake and that volume, variety and velocity of data levels are maintained.



When done and done right

– Keep up with new data volume and still govern it
– Adhere to regulatory requirements using industry-specific compliance tools
– Accelerate master data adoption
– Improve insight accuracy with high quality data
– Quickly respond to compliance audits
- Increase your ability to protect data

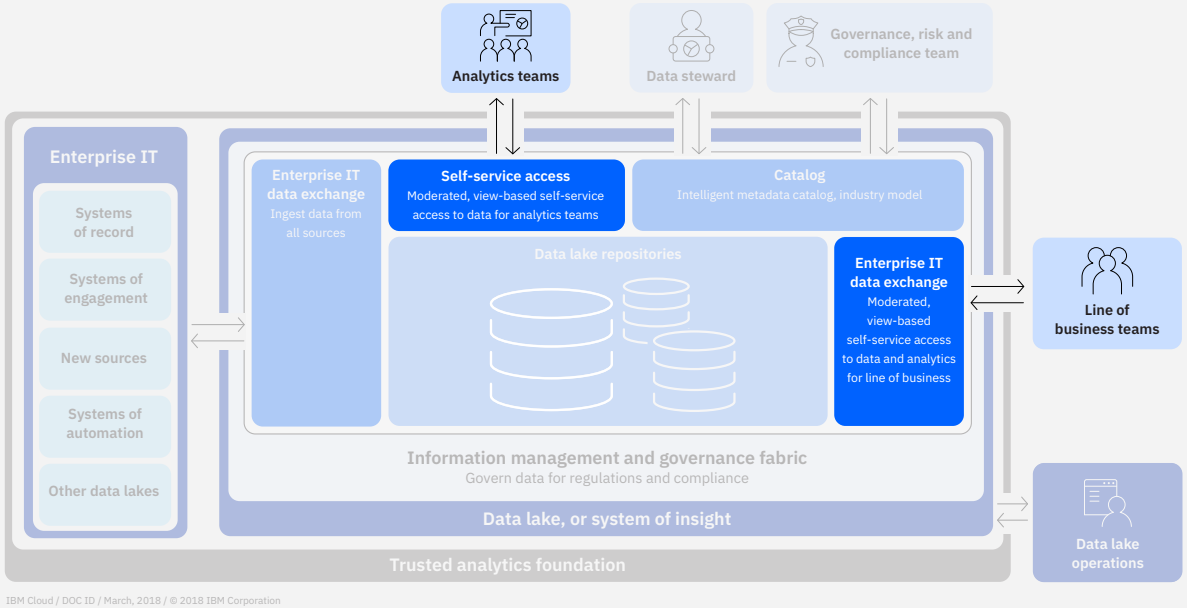
When not done or done incorrectly

– Lose the ability to manage growing data volumes from structured and unstructured sources
– Waste time finding data, which can impact audit readiness
– Miss opportunities to stay within compliance and governance requirements

Self-service or reporting

Self-service access helps you find relevant information from the data through simple search interfaces. It provides high quality, trusted data to self-sufficient builders who can use the data to build analytical models in their data science initiatives. It also enables non-technical users to transform data before building and deploying models.

Straightforward access to data helps IT builders in their data preparation and transformation efforts. This access helps the governance and compliance teams curate data for audit readiness. It also helps solution consumers create custom reports for their business requirements and have access to business-ready data so they can make quick decisions and derive meaningful business insights from their data.



When done and done right

–Empower data users to have access to contextual data
– Help data consumers trust data through tribal knowledge, social tagging and qualitative rating of information assets
– Watch data become an organizational asset accessible to all data consumers
– Get faster time to value
– Accelerate innovation
– Enable agile and iterative data exploration and analytics

When not done or done incorrectly

– Spend more time finding and preparing data than analyzing data
– Lose ability to find or access unstructured assets
– Make decisions slower due to lack of access to trusted data
– Experience impeded innovation

Why IBM

According to research conducted by Radiant Advisors, governance and security were identified by 72 percent of leaders as the key challenges, yet top success factors for their organization. Recognizing governance and information architecture as a priority is the first step. This will open up the conversation around the organization to clearly define what all data users require from their data. In a world where bad data in equals bad data out, every data user becomes part of the conversation.

Deploying an enterprise-wide single platform for data integration, data quality processing and data governance is essential to achieve success from your analytics initiatives. Doing so can give you the ability to ingest data, ensure it's of high quality and govern it to feed into your analytics processes. By approaching the challenges with a governed approach to your data lake, you build a foundation to deliver trusted data to be used for many uses.

No other vendor can match the breadth and depth of the IBM DataOps platform. Whether it's the scalability to manage immense volumes of data, industry-specific accelerators, capabilities to make structured, unstructured and semi structured data usable, or leading with machine learning and artificial intelligence expertise, IBM provides a comprehensive solution for you to build a trusted and governed data lake.

To find out more, visit ibm.com/governed-data-lake.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
August 2018

03018403-USEN-01