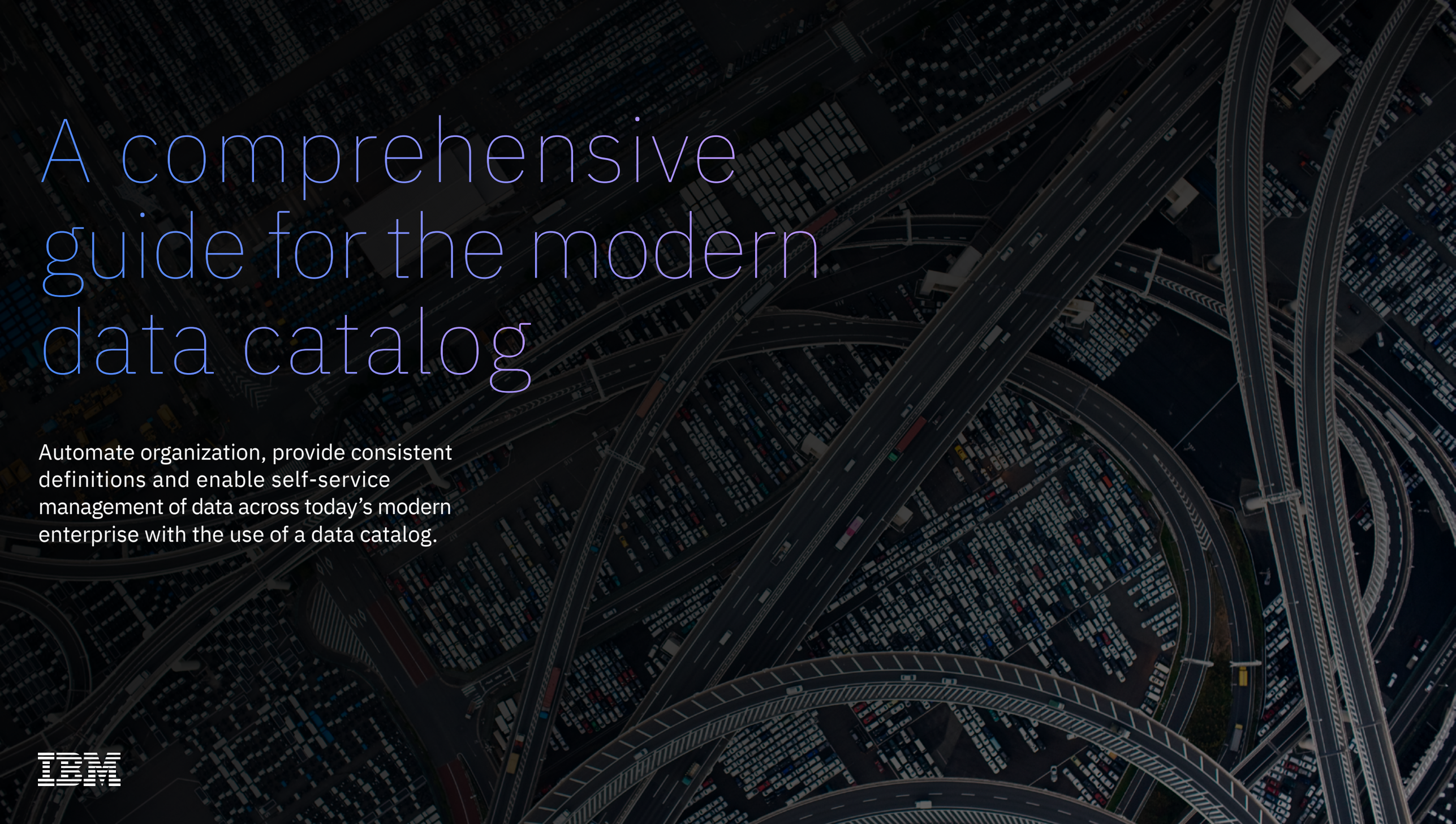


A comprehensive guide for the modern data catalog

Automate organization, provide consistent definitions and enable self-service management of data across today's modern enterprise with the use of a data catalog.

IBM

Table of contents

3
Introduction

5
The importance of a modern data catalog

9
Setting up a business taxonomy

12
Data catalog use cases

– Enhance use of data
– Improve regulatory compliance
– Automate data governance for DataOps
– Support a governed data lake
– Enable AI governance

22
Get started with IBM Watson Knowledge Catalog

Introduction

Businesses need to maximize the value of their data to drive monetization and increase what McKinsey & Company refers to as “the insights to value chain.”[1] In many cases this includes the leveraging of artificial intelligence (AI) that can fuel predictive insights and proactive outcomes. However, growing volume of data spread across multiple deployments as well as internal obstacles of traditional manual processes and data stewardship roles remains a challenge. Leaders are discovering their current data processes don’t efficiently scale to tackle today’s needs, nor ones they will face in the future, and yet the importance of being able to find a solution is absolutely imperative.

Gartner estimates that by 2021, AI augmentation–a human-centered partnership model between people and AI technologies working together–will create a business value of $2.9 trillion and 6.2 billion hours of heightened worker productivity worldwide.[2]

As a solution, many organizations have begun to implement DataOps (data operations) practices to deliver continuous enterprise data that is high-quality and trustworthy. DataOps orchestrates people, process, and technology to solve the challenges associated with inefficiencies in accessing, preparing, and integrating data. This enables collaboration across an organization to drive agility, speed and new initiatives at scale.

At the heart of an effective DataOps practice is a data catalog, a metadata management tool designed to help organizations find and manage large amounts of data. It puts trusted data in the hands of a business by automating the organization of a common and known business vocabulary, self-service management of data and on-boarding of data content. This ebook focuses on the importance of a modern data catalog and the benefits a business can reap from its use when it’s implemented correctly. From supporting multicloud adoption and integration, to accelerating an organization’s journey to AI, the data catalog is at the foundation.

Discover DataOps with an interactive guide →



Gartner estimates that by 2021, AI augmentation–a human-centered partnership model between people and AI technologies working together–will create a business value of $2.9 trillion and 6.2 billion hours of heightened worker productivity worldwide.



The importance of a modern data catalog

Gartner originally defined a data catalog as a tool that “creates and maintains an inventory of data assets through the discovery, description and organization of distributed data sets.” As the quantity of data available to organizations has grown exponentially over the last several years, data catalogs have grown in importance and their definition and scope have grown as well. Delivering business-ready data to feed analytics and AI projects begins with a data catalog that can automate organization, provide consistent definitions and enable self-service management of enterprise data.

A modern data catalog allows data analysts to find all the data available in each database or application maintained by their organization. This can include both relational data and unstructured data which can be found in word documents or spreadsheets, whereas analytic assets will include Jupyter Notebooks, trained models and dashboards. Because data catalogs make data sources more discoverable and manageable, they help organizations make more informed decisions about how to use their data. How to access the data, the data format, the classification of the asset, the asset lineage and the list of collaborators that have access to certain kinds of data is the kind of information that should be embedded inside data assets.



Benefits of a data catalog:

01

Index and enrich assets
When looking for a data catalog, it is essential for the catalog to have a metadata repository that acts as an index for data and other assets, making it easier to understand what kind of data and analytic assets are in your catalog.

Here's how a data catalog can smoothly ensure the addition of assets:

- Leaves your data where it is. Whether it is in the cloud or on premises, just add the connection information into your data catalog to access it.
- Automatically discovers and adds all tables from a connection to a relational data source as assets in the catalog.
- Uploads files to the dedicated encrypted cloud object storage bucket that's associated with the catalog.
- Includes an object storage instance to store assets that are copied into the catalog.

After adding data assets to a catalog, they can be profiled to add generated metadata about the data assets' contents, and in addition, you can enrich assets by having catalog collaborators add ratings and reviews. Catalog collaborators can also create tags that describe different assets while making sure data classes accurately define the type of data stored within the assets—all while having set business terms that help describe data in a standard way for your enterprise.

02

Control access to data policies
Policies should apply to all catalogs within an enterprise and the corresponding policy tools should only be available to users who have special permissions within the catalog.

Policy tools should allow you to:

- Create business terms that describe your data to use in policies
- Write policies to deny access and protect sensitive data assets
- Write policies to mask data values in columns that contain sensitive data
- Monitor trends in policy enforcement over time

03

Benefit from data discovery
Data catalogs must have a record of collaborators who need access to certain assets and corresponding information in data sets from across an entire organization, without needing separate credentials for every source. This creates a single platform where any member in an enterprise can locate their data. To ensure security, the data catalog assigns the correct roles to its users based on their needs and will place the necessary restrictions on what the user can and can't do inside the catalog.

Types of collaborators and their functions:

- Authors: Subject-matter experts who will pull and draft the appropriate information into the catalog
- Approvers: Once authors have completed their draft, approvers can review, comment, approve, or deny the delivered information
- Publishers: Authorized to publish the approved information and make the new business terms and data assets available to anyone with access to the business glossary

04

Expedite data preparation
In order to help transform large amounts of raw data into consumable, quality information that's ready for analysis, a data catalog should have self-service preparation features to support any data preparation solution your company already has in place.

Make sure the following features are included in your catalog to make it easy to explore, prepare and deliver data that can be trusted and used across your business.

- Powerful operations that clean, organize, fix and validate your data
- Scripting support for the efficient and flexible manipulation of data
- Scheduling and monitoring of data preparation flows
- Profiles for validating your data
- Visualizations for gaining insight into your data
- Policies that mask data are enforced
- Support for unstructured data

05

Collaborate across governed assets
A catalog helps alleviate manual processes and dependencies with advanced discovery capabilities typically driven by machine learning and semantic context. This makes it easier to find relevant assets quickly and at scale.

Ways in which a catalog enables data discovery include:

- Search keywords and filters based on subject tags and other asset properties
- Preview capabilities to ensure that you are selecting the correct data asset
- Reviews about assets created by collaborators within the catalog to help identify the best assets to pull from
- Asset recommendations that are automatically compiled based on your usage history, similar assets and other factors

Dive deeper into the benefits of cataloging →



What it looks like when your business has a data catalog and it is implemented correctly:

- Decrease time to results with more time to analyze data and put it to use
- Capture contextual asset knowledge and improve data's utility
- Track data lineage and improve trust in your data quality
- Market information assets for broader consumption
- Assist with data governance and compliance



What it looks like when your business doesn't have a data catalog or is implemented incorrectly:

- Risk wasting time searching and tagging your data
- Lose crucial knowledge when you locate data but can't find colleagues who understand the data
- Lose knowledge of who has access to data
- Failure to meet compliance and governance requirements

Setting up a business taxonomy

Understanding the benefits of a modern data catalog is just beginning. It’s equally important to understand how to start integrating it into your business to realize value faster. When the goal of your organization is to increase efficiency and collaboration across stakeholders, the first place to focus your improvements on should be the company’s taxonomy. This will become the foundation for content categorization, data relationships, and provide a guideline that improves that speed at which data can be found, accessed or reused.

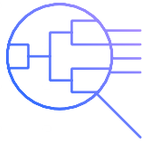


Best practices when establishing a robust business taxonomy include:



Step one: Focus on a single high-value information area

As opposed to trying to organize all of your assets at once, it is far more efficient to focus on a particular segment of the business that will drive the greatest impact. For instance, if compliance and regulatory processes, such as for GDPR and CCPA, are high priority for your organization, begin with establishing terms and classifying assets related to personally identifiable information.



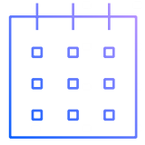
Step two: Concentrate on the meaning of business definitions

Use the language of your industry in the form of logical or business intelligence models to power existing terms and standards already set in place. Take time to understand how certain concepts and definitions are currently being applied throughout your organization, then build your catalog specific to these key components, data types and common uses of data.



Step three: Establish benefit and gain interest

Though adoption of a business taxonomy might not happen overnight, it is critical for your organization to understand the advantage of having a single place where all information is stored. Within a specific sector of your business, champion the idea of selecting a focused area to start integrating a data catalog with an established business taxonomy, so the organization's data can be consolidated in one place.



Step four: Develop and commit to milestones

The final step is to establish official milestones that your organization will commit to for implementation of the business categories, business terms, and correct assignment of user roles—and moreover the data catalog process. Whether you have a mature DataOps culture in place or this is your first step, it is important to remember that each organization has unique needs where stakeholders in and out of IT need to add value to drive success of data projects.

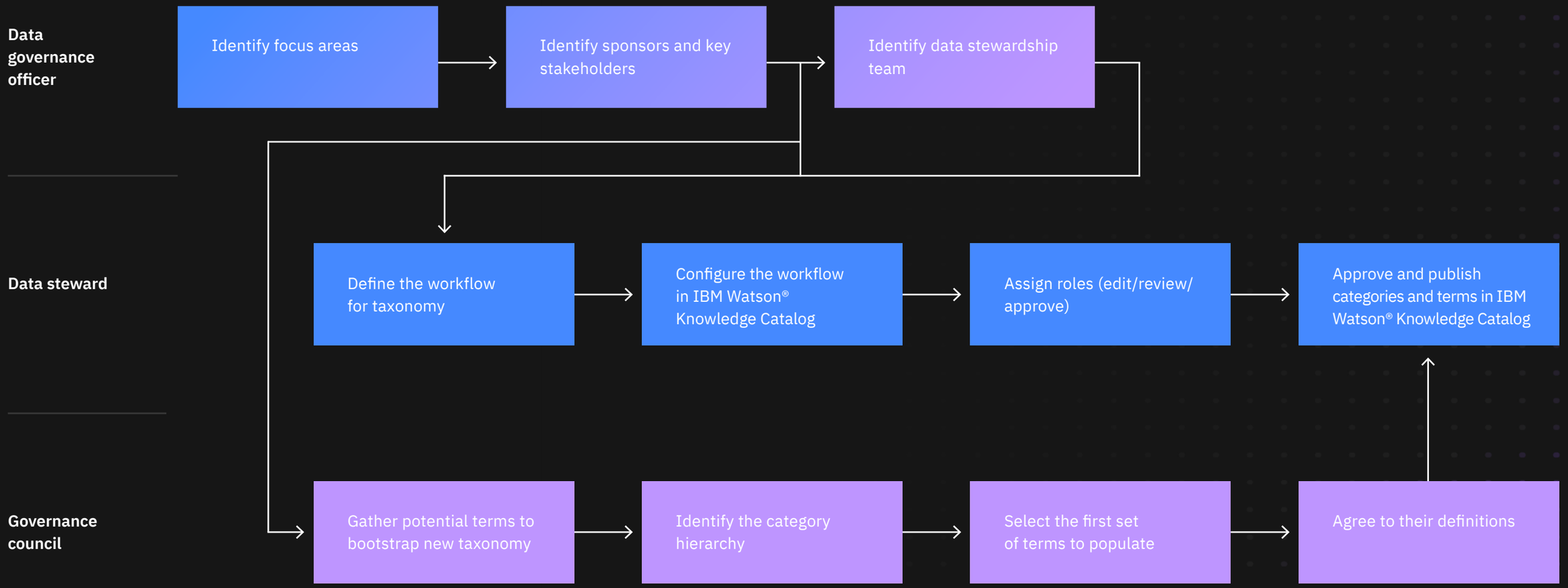


Figure 1: Data citizens must work together to build business taxonomy that benefits their organization as a whole.

Data catalog use cases

An organization can leverage a data catalog to accomplish the levels of success that enterprise data leaders are experiencing today. From ensuring that your enterprise can meet compliance regulations, facilitating data lake governance, or cutting down on the time consuming labor that it takes to govern your data, the following stories share the data struggles five different companies were able to overcome by implementing their own data catalog.



Enhance use of data

A data catalog offers a single place for data analysts to view and easily find all data assets across different departments. This consolidated view enables team members to share insights that can improve the business. For example, team members might discover cross-sell and up-sell opportunities that can generate new revenue streams.

How Credito Valtellinese used cognitive analysis to find hidden opportunities

Seeking growth through customer-centric banking, Credito Valtellinese needed to reposition itself. In order to do so, the organization launched a plan that was predicted to increase revenue per customer by optimizing its cross-selling and up-selling marketing campaigns. However, the bank always encountered the same roadblock—internal systems were not centered around their client relationships, making it near impossible to market to existing customers.

Credito Valtellinese had to create an analytical foundation, inclusive of a data catalog, in order to understand its customers' behaviors and needs on new level of depth and granularity, and by adapting cascading styling sheets (CSS) their organization was able to create just that.

Their comprehensive system and management solution delivered precisely targeted promotions to those which were most likely to convert, therefore increasing outbound marketing campaign conversion rates by 10%.



Improve regulatory compliance

Ungoverned sensitive data may lead to regulatory penalties. For instance, if a business does not rectify any of their violations against the California Consumer Privacy Act, an attorney general could impose a civil penalty of anywhere from $2,500 to $7,500 per violation,[3] and when it comes to the GPDR, financial penalties could go as high as 20 million euros or 4% of worldwide annual turnover.[4] Therefore, as organizations face growing data privacy regulations, they must look more holistically at how they store and use data.

A data catalog can automate the classification and profiling of data assets and automatically enforce data protection rules established to anonymize and restrict access to sensitive information. More importantly, if something goes wrong, controls allow the organization to rapidly respond to an issue, whether that means flagging sensitive data, identifying and remediating issues, or collecting information in response to an audit.

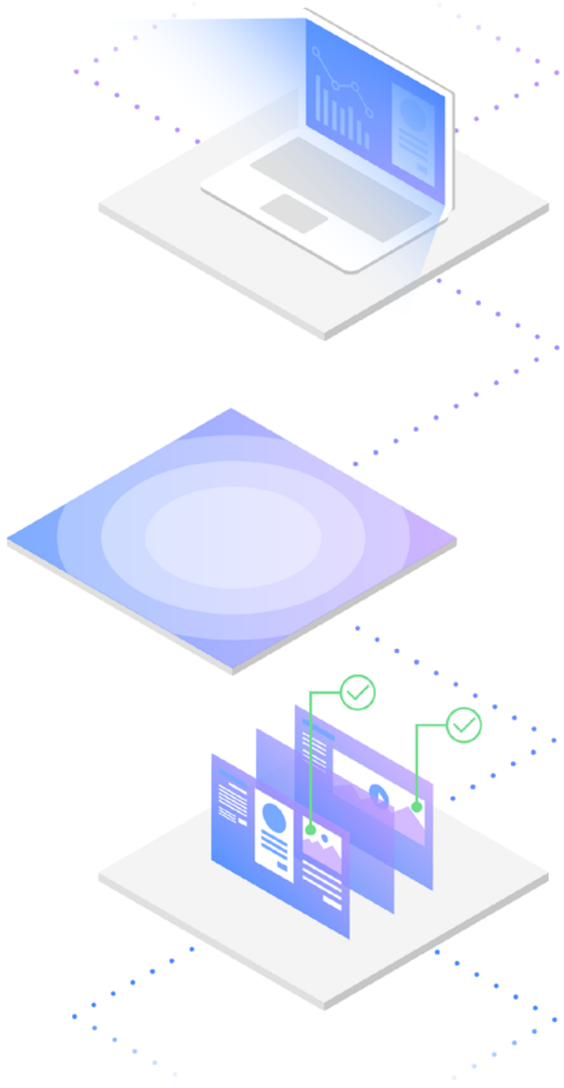
The IBM Global Chief Data Office helps analyze and visualize business risks round sensitive data

Due to GDPR readiness, companies in possession of personal data from European Union data subjects are legally obligated to understand the types of data they store, where the data lives and its associated levels of risk.

For a company as large as IBM, which operates in more than 170 countries, it can be a daunting task to refresh an organization's privacy practices and ensure that the GDPR guidelines are met—all while enhancing products and services that will ultimately benefit all of its clients. To undergo this task, the Global Chief Data Office (GCDO) created a global program, among numerous work streams, to address the GDPR requirement and more comprehensively understand the type of personal data IBM controls.

The results of this effort were collected in a central data privacy catalog as a key first step in the journey to readiness, but it was still uncertain how to identify, evaluate and share the discovered information of data that needed to be in compliance with the GDPR. As a result, IBM used their own cataloging technologies and created a central store for their privacy data. To compliment the catalog, IBM Data Risk Manager was also implemented to provide a data risk control center for executives and their teams to easily view the updated information from the privacy catalog in a central dashboard and ensure that ongoing requirements to meet data privacy regulations are met.

Learn More: Forrester names IBM a leader in Machine Learning Catalogs →



“The job involved examining more than 6,500 application across the company, about 3,400 of which are critical from a GDPR perspective.”

Neera Mathur, GCDO Senior Technical Staff Member in the Global Chief Data Office



Automate data governance for DataOps

An integrated quality and governance platform helps manage data and protects it from misuse. For effective governance, an enterprise data catalog must be in place. You can't effectively apply governance if you don't have organized data with proper metadata tags and lineage. Data organization includes detailing each data object: documenting data properties, ownership, business context, origin, and structure; evaluating data quality; and properly classifying data so it can automatically be used to define and refine an organization's DataOps practice.

How Integra LifeSciences adopted an integrated approach to manage all parts of their business

When implementing various new systems and processes into their organization, Integra LifeSciences, a surgical and medical instrument manufacturing company, found that governance in their organization was not a simple feat. The quantity of data they needed to keep track of was quickly multiplying, and they were losing track of where the data was located and how they could effectively use it to benefit their business. By turning to an integrated approach that collected, defined and managed their data all in one platform, Integra was able to cut 50% of business systems, reduce their complex management of systems and data, and cut operational costs in order to maximize the organization's growth benefits.

Integra LifeSciences worked with IBM to implement IBM data cataloging technology that creates consistent definitions of its business data and helps them better understand what their data could do for them.

To learn more about what IBM Watson Knowledge Catalog can do for your business take a guided tour to see how business users can quickly discover, curate, categorize and share data assets across a whole organization.



“Integra has substantially reduced operational costs as a proportion of revenue–and we predict the solutions will unlock greater financial benefits as we move towards our USD 1 billion revenue goal.”

William Compton, Chief Information Officer,
Integra LifeSciences



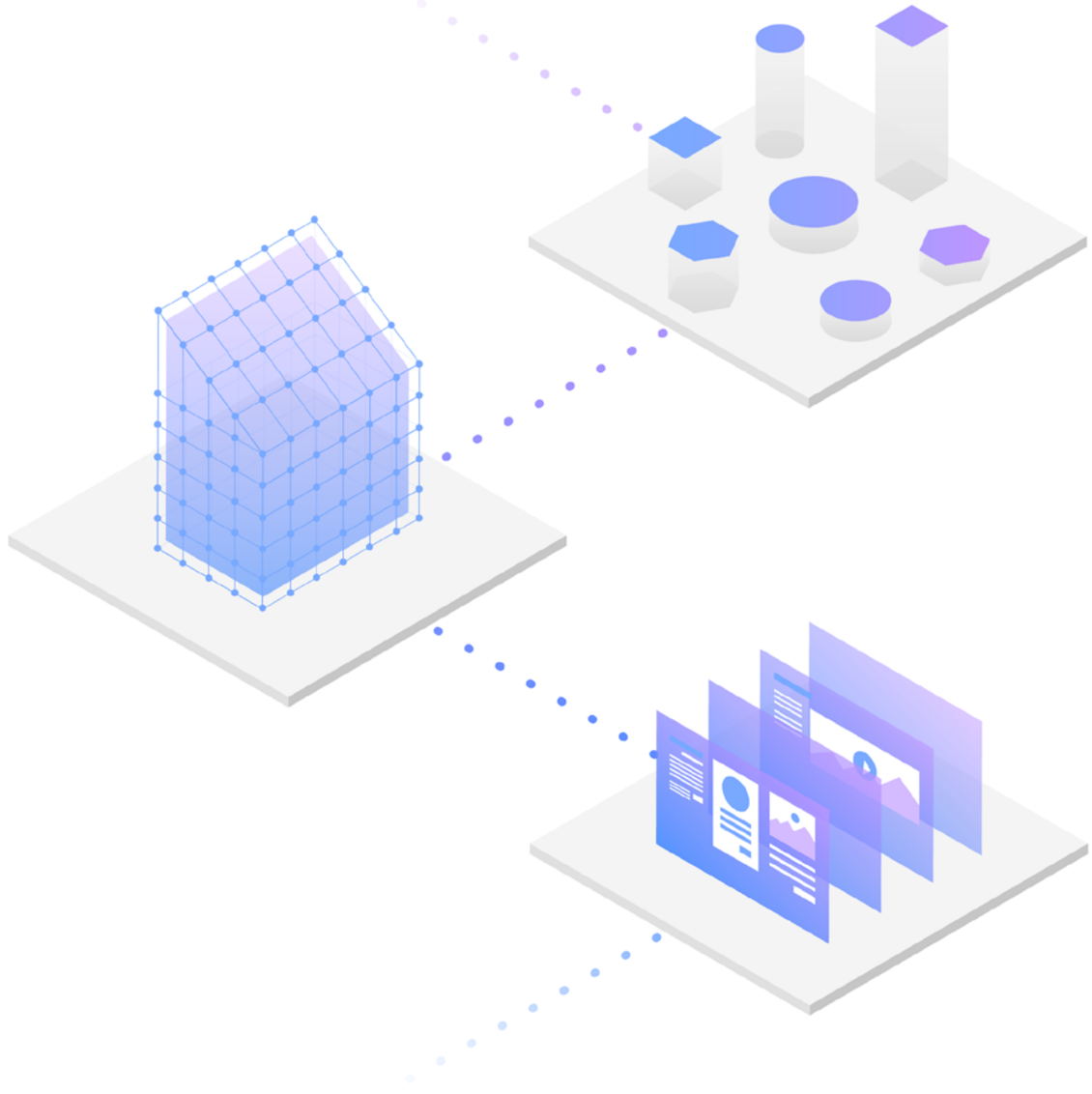
Support a governed data lake

Data lake governance takes discipline, good policy and collaboration between the people who manage data access and the people who access the data. Cataloging helps to tag the data in the data lake and create an inventory of information assets. The catalog interface provides data lake users with information about the data within its classification, lineage and how it's governed. The catalog can serve multiple stakeholders in the organization, eliminating inefficiencies associated with "lost in translation" issues.

Deliver clean, reliable data with data lake governance →

Enable AI governance

A data catalog can help the enterprise governance program grow to support the maturing demands of AI governance. As AI takes root, you'll need an organizational approach toward developing policies which lets you create a framework to effectively design, deploy, and monitor AI-powered models and algorithms with a focus on fairness, accountability, transparency, safety, and privacy, ensuring fair outcomes.



Get started with IBM Watson Knowledge Catalog

The modern data catalog goes way beyond that of the legacy metadata repository businesses have been using for decades. They surpass the concept of metadata capture and management by including automation and discovery techniques such as visual recognition, natural language classification and machine learning. With these capabilities, a data catalog can organize data in near real-time with the added benefit of eliminating the inefficient manual processes required by older repositories.

The new wave of intelligent data catalogs is not only changing the way business is run via virtualization and multicloud deployment, but how organizations are carving new business models and preparing for the future of AI.

Therefore, as businesses continue to digitally transform themselves to build and incorporate AI into their overall business strategies, the value of data catalogs integrated with a data quality and governance platform becomes more essential.

IBM Watson Knowledge Catalog is an open and intelligent data catalog for managing enterprise data and AI model governance, quality and collaboration. By providing an end-to-end experience rooted in metadata and active policy management, it helps data citizens quickly discover, curate, categorize, and share data assets, data sets, analytical models, and their relationships with other members of your organization.



There's a reason our customers named Watson Knowledge Catalog a 2020 Gartner Customer Choice Award Winner. Test drive the product to see why.

Talk to an expert to learn more about Watson Knowledge Catalog and explore its seamless integration with IBM DataOps services for IBM Cloud Pak® for Data.

© Copyright IBM Corporation 2020

IBM Corporation
Route 100
Somers, NY 10589

Produced in the United States of America
July 2020

IBM, the IBM logo, ibm.com, IBM Cloud Pak and Watson are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

01 Holger Hürtgen and Niko Mohr. “Achieving business impact with data”, Microsoft Report, April 2018.

02 “AI Augmentation Will Create $2.9 Trillion of Business Value in 2021”, Gartner, August 2019.

03 Nicholas Schmidt. “Top 5 Operational Impacts of CCPA: Part 5 - Penalties and enforcement mechanisms”, International Association of Privacy Professionals (IAPP), August 2018.

04 “IBM Pathways for GDPR readiness”, IBM White Paper, September 2017.

EWDPJZDQ