Data Catalogs

For dummies®
A Wiley Brand

- Learn what a data catalog is
- Easily find and understand your data
- Deliver trusted data insights

Collibra Special Edition

Bineesh Babu

Ben Moser

Chandra Papudesu

Luis Romero

About 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. The Global 2000 relies on Collibra to create the critical alignment that accelerates workflows and delivers better results faster. We have a diverse global footprint, with offices in the U.S., Belgium, Australia, Czech Republic, France, Poland, and the U.K. To learn more, visit collibra.com, follow @Collibra on Twitter, or follow us on LinkedIn.

Data
Catalogs

for dummies

A Wiley Brand

Data Catalogs

Collibra Special Edition

by Bineesh Babu, Ben Moser,
Chandra Papudesu, and
Luis Romero

for
dummies
A Wiley Brand

Data Catalogs For Dummies®, Collibra Special Edition

Publisher's Acknowledgments

Table of Contents

Introduction

Are you drowning in data? It seems like most organizations are, in this data-driven era.

Organizations know they need to become more data-driven to maintain profitability, support innovation, and grow the business. The most successful organizations in the world — from governments to global corporations — have made data a strategic asset. In fact, a January 2021 report from Forrester, "Chief Data Officers: Evolve Your Teams To Accelerate Impact From Data Insights," states that 63 percent of organizations have appointed a Chief Data Officer.

But there are many headwinds in the form of challenges and risks that stand in their way. Chief among these challenges is the scale and complexity of the data ecosystem. As more sources of data come online, the problem only becomes worse. Then there are the threats — both internal and external — to security and reputation.

What's the solution? For many organizations, it's implementing a *data catalog*. Data catalogs pull everything together from across your entire digital environment and help you make sense of it, at both the macro and micro levels.

If that sounds interesting, and perhaps a good fit for your organization, you've come to the right place!

About the Book

Data Catalogs For Dummies explains why organizations are increasingly turning to data catalogs, and how they can address today's data challenges. You'll learn about:

>> Taking the first steps to control your data landscape

>> Identifying the right data catalog for your organization

>> Maximizing the value of your data catalog

>> Leveraging your data catalog for data governance and privacy

>> Turning your data catalog into a full-featured data intelligence platform

>> Defining and measuring success metrics and outcomes

Foolish Assumptions

We've tried to make this book as accessible and readable as possible for everyone. We've designed the content of this book for anyone that wants to learn how the right data catalog can help them discover, understand, and extract insight across their organization's data.

Icons Used In This Book

To help understand data catalogs a little better, we use these icons to point out information we want you to pay attention to:

REMEMBER

We use this icon when there's something important you should commit to memory before you progress on with the chapter.

TIP

You find these icons dotted throughout the book giving you help-ful suggestions and bite-sized nuggets of useful information.

WARNING

We don't just know what to do. We also know what not to do. Look out for these icons to help you avoid some common pitfalls.

TECHNICAL STUFF

This is the jargon buster icon. When we have to get technical, we'll give a little plain English explanation of what we mean.

Beyond the Book

Want to find out how you can take the next stop along your data intelligence journey? You can visit `www.collibra.com/data-catalog` to learn more and sign up for a tour of Collibra Data Catalog.

Chapter **1**

Overcoming Today's Data Challenges

The status quo of the last decade — in other words, "the way we've always done it" — isn't serving today's business data needs very well, and many businesses are feeling the pain. In this chapter we review some of the challenges businesses are facing, and then take a first look at how data catalogs can help.

Identifying Common Data Management Challenges

Today's organizations are beset on all sides with more data, more demands to derive value from it, and less certainty about how to more efficiently protect and access this data. This section takes a look at some of these challenges to quantify what you're up against.

Dealing with increasing volume, variety, and velocity

The amount of data in today's world is almost unimaginably large. In the Worldwide IDC Global DataSphere Forecast, 2022–2026: Enterprise Organizations Driving Most of the Data Growth, IDC predicts that 221 exabytes of data will be generated in 2026, more than double the volume generated in 2022, and IDC expects enterprise organizations in 2026 will generate 70% of that data volume.

REMEMBER Today's organizations are dealing with a volume, variety, and velocity of data that pose significant risks to the business. Enterprises find themselves in an unenviable dilemma: Data has become the most important business asset for decision-making, but it's siloed across disparate systems.

Enterprise IT continues to experience challenges, including:

» Fragmented tools and manual processes across siloed data make it difficult to create ownership and accessibility

» Wasted time wrangling information across thousands of applications and systems to define data, owners, and who should have access

» Duplicate data sets across the org that makes data unreliable leading to inaccurate reporting and costly inefficiencies

» Tools that are not integrated so you never get a clear picture of all your data — and your business

All this can lead to slower responses to changing market conditions and decreased value to the business. The potential for revenue loss thanks to poor decision-making and lower productivity is real.

Most don't trust the ability of their organization to create a data-driven culture. A 2020 Forrester Consulting study commissioned by Collibra shows that professionals believe that a data practice is 56 percent less likely to positively contribute to optimal business decisions.

Meeting ever-more-stringent privacy rules

It's not just the data deluge that creates challenges. Today's regulations require organizations that do business with the public to abide by stringent privacy rules, such as the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA). These regulations require visibility into what data you have, why you have it, and how it's used.

And while your business complies with current regulations, you must also keep up with new regulations.

Compliance is costly and involves a range of risks:

>> Shifting regulations make it difficult when data resides and is replicated across multiple systems, cloud platforms, and geographies

>> Time is wasted on manual steps to identify data and manage privacy

>> Slow responses to data subject access requests (DSAR)

>> Potential compliance penalties, brand damage, and loss of customer trust

Spending valuable time on bad data

It's true. Too much time is spent looking for data that — when it's found — can't be trusted.

A 2020 study from Forrester Consulting on behalf of Collibra showed that business analysts waste 76 percent of their time finding, accessing, and validating data. How much time are your analysts wasting?

Why is so much time wasted? Common causes include:

>> Large volumes of data coming from thousands of sources across the organization

>> Business analysts lack visibility about where data is stored, what the data is, or whether they can use it

>> When analysts do find data it often requires significant manual intervention and cleansing to get it into a usable format

>> Too much time spent on the non-value effort of searching for and validating data instead of focusing on deriving insights from the data

Data scientists spend much of their time on manual, ad hoc data discovery and preparation instead of creating and deploying analytic models.

Lack of data management negatively impacts the business in many ways including:

>> Decreased productivity leading to delayed responses and less time for impactful work such as report and model development

>> Increased costs due to new report requests and reporting redundancies/reconciliations

>> Higher business risk from non-compliant use of data

>> Inconsistent or incorrect analytics and reporting, leading to lack of trust in insights generated

>> Potential revenue loss from slow, inaccurate decision making

>> Lack of trust in models because data isn't trustworthy

>> Errors in data due to manual formatting

Struggling with data pipeline management

As a result of challenges around data management, organizations are experiencing difficulty in creating and maintaining data pipelines.

These challenges are widespread and include:

>> Creating data pipelines to feed data visualization tools and machine learning (ML) models, and address the demand for immediately consumable data

>> Finding and accessing data sets, data elements, and ML models for use in pipelines from different data sources, data domains, and data owners

>> Trusting sources, when deciding what data should be moved from an application to a data lake or from a data lake to an analytics database

>> Determining whether to reuse current data pipelines or create new ones

These questions are difficult to answer when there are such large volumes and variety of data sources and formats.

All this is compounded by a lack of governance, lack of data quality, manual lineage, and frequent updates to data sources.

The current situation creates:

>> Compliance gaps as data sets are merged from different sources

>> Difficulty to collaborate with data stewards and data scientists

>> Too much time spent on updating, maintaining, and assuring data quality of data pipelines

>> Long lead times to move from raw data to insight for the business

>> Limited reusability and productivity

Given so many challenges, how can your organization become more data driven? It all starts with a data catalog.

Data Catalogs to the Rescue

Why the surge in interest in data catalogs? You don't have far to look to find an answer. You're knee-deep in a tsunami of data. And the tsunami isn't slowing down or getting any smaller. The world is becoming more and more digital. More and more data is created. And more and more data flows through hybrid cloud ecosystems.

You're fighting a multi-front battle to keep innovation moving and your teams developing, to manage your technology stack in a rapidly evolving market and to protect your data from breaches.

But what exactly is a data catalog? And what purpose does it serve in an enterprise?

Understanding the purpose of a data catalog

A *data catalog* creates and maintains a well-organized inventory of an enterprise's data assets across its entire digital environment.

Implementing a data catalog is the first step on the journey to data intelligence. (If you want to skip ahead, we talk more about data intelligence in Chapter 4.) An effective data intelligence platform connects all data citizens, empowering them to collaborate, shop for trusted data, and attain optimal business outcomes.

Data catalogs deliver their benefits by leveraging *metadata*, which is data that describes the data.

Metadata is often subdivided into three categories:

» **Technical metadata** provides information on the format and structure of the data, such as the data models, storage schemas, file layouts, and APIs.

» **Business metadata** provides business knowledge in plain language to help understand the business context.

» **Operational metadata** describes when and how the assets were created and accessed. When was this object refreshed? Which ETL (Extract, Transform, Load) job created it? How many times has a table been accessed, and by which users?

At the end of the day, the overarching role of the data catalog is to leverage all this metadata to make digital assets easier to find, use, and trust. When you can find, use, and trust data, you're in a position to drive business decision-making informed by data.

AN ANALOGY: THE DEWEY DECIMAL SYSTEM

Before the digital age, public libraries used to have to solve a similar problem: If a library has upwards of a couple hundred thousand books, how do you make it easy for a reader to walk in and quickly find the book of interest? The Dewey Decimal System was the answer!

Libraries used the Dewey Decimal System to classify books based on their genre, author, edition, publisher, age appropriateness, and most importantly, location within the library.

Although an oversimplification, an enterprise data catalog serves a similar function as the Dewey Decimal System to help catalog digital assets distributed across an enterprise and provide an avenue to data discovery.

Why are there so many data catalogs and providers?

Digital transformation continues to be the global economy's primary driver. As a result, organizations are moving large volumes of data across a spaghetti of hybrid cloud environments, data lakes, and data warehouses. And this drives a ferocious market for data products to try to extract more value out of all this data.

Savvy data leaders are using data catalogs as an essential component of modern enterprise data architecture. They know how difficult it is for modern enterprises to leverage distributed data in a safe, trustworthy, and reliable manner.

Given these pervasive challenges, a wide variety of data catalog providers have emerged across the past couple decades to try to help solve this growing data problem.

But not all catalogs are created equal.

Why you need an enterprise data catalog

Every data professional responsible for navigating their organization's journey to digital maturity faces complexity that is often challenging to manage. That's why it may be challenging to understand the data catalog market, which is made even more complex because not all data catalogs have the same capabilities and functionality.

To put a fine point on it, enterprise data catalogs support enterprise metadata management and data governance across all data sources.

On the other hand, embedded and tactical data catalogs are shipped within other products to offer data catalog functionality only for a particular product or data source.

TIP

Some business intelligence (BI) tools include embedded data catalogs to manage the metadata that the BI tool itself needs. But it is a narrow focus.

Embedded catalogs serve their purpose in targeted use cases. But they are limited in that they don't provide end-to-end metadata management.

To get end-to-end metadata management across your entire stack, you'll want to consider an *enterprise* data catalog. (You can find out more about what to expect from a data catalog depending on where your organization is in your data intelligence journey in Chapter 2.)

REMEMBER

It's important to find an enterprise data catalog that has the functionality to serve as your foundation from which to build more data intelligence capabilities.

Considering the users

It's also important to consider who will be using your data catalog. When data catalogs first emerged several decades ago, they were almost exclusively used by highly technical users within IT organizations and were tightly linked with data integration tools. As a result, the catalog was geared towards this tech-savvy user.

Today, users are not limited to IT, but come from across the business — from Finance to Product Management, Marketing to Sales.

To get broad adoption, it's important to find a data catalog that can serve all of these roles equally well.

Chapter **2**

Determining the Right Data Catalog for Your Organization

With data volumes growing ever larger, and data becoming more integral to business success, data professionals have more pressure to manage data effectively. Data catalogs help you discover, understand, and classify the data that matters to generate insights that drive business value.

You know this. That's why you're reading this book: to understand how a data catalog can help your organization get more intelligent about how data is managed.

You're on the path to transforming how your organization understands and uses data. And now you're focused on implementing a data catalog that works for your organization.

The challenge will always be simplifying complexity. As you learn what data catalogs offer and how you can leverage them, you need to evaluate what type of solution you need. It's a great opportunity to take stock of your organization's progress toward creating a data-driven decision-making culture.

In this chapter, we focus on the goals that are going to drive your data catalog journey and detail three stages on your data catalog journey:

» Crawl: Some things all data catalogs should have

» Walk: Taking it to the next level

» Run: Making it work for the enterprise

Setting Your Goals

You start your data catalog journey by identifying the goals that are relevant to your role, your business, and your current situation.

Here are a few common scenarios that might resonate with you:

» You're a Data Steward trying to automate governance activities and establish a common understanding around data

» You're a VP of Data Engineering trying to enable self-service data access to increase developer efficiency

» You're a Data Privacy Officer building out your policy management and enforcement regime

» You're a Data Scientist who needs access to all the trusted data and models across the organization in a centralized catalog

» You're a Head of Analytics looking for an easier way to build, trust, and use BI reports and dashboards

Even if you don't see your specific role and challenge in this brief list, you can start to see the value of identifying your specific goals.

The importance of identifying your specific goals is relevant in creating the key performance indicators (KPIs) that keep your data catalog project on track with your team and other key stakeholders.

Data catalogs have many different applicable use cases. It's up to you to understand what is right for your team and your organization, and then define your requirements based on your goals.

Identifying the right metrics

A good place to start is by defining the success metrics (or KPIs) that are most relevant to measuring your progress. (You can read more on KPIs in Chapter 5.)

For example, if your data science team's efficiency is critical to driving new initiatives, then measure average time-to-production for machine learning algorithms.

A data catalog is designed to drive faster data access and faster understanding across your organization. That way your data scientists aren't wasting time looking for data, IT isn't overwhelmed servicing the data infrastructure, and managers get the productivity and insights the business needs.

Other success metrics that may be relevant include:

>> Self-service data access requests

>> Data policy enforcements

>> Report/BI/analytics catalogs

>> Data governance and stewardship

>> Cloud migration

These are only a few examples of the many reasons why organizations look to data catalogs to drive more value from data. Here are some other example success metrics that may be relevant:

>> **Goal**: Enable self-service access to data

Metric: Activity metrics such as number of data access requests

>> **Goal:** Increase robustness of data operations

Metric: Time to resolution (TTR) of data issues

>> **Goal:** Increase efficiency of data analysts

Metric: Number of reports created per analyst per month

>> **Goal:** Data policy enforcement

> **Metric:** Proportion of cataloged data assets with well-defined access policies

» **General Use Metric:** Number of assets ingested

» **General Use Metric:** Number of daily active users

REMEMBER
Define your success metrics based on the goals you've set for your data catalog implementation. It's an important investment in the health of your business. Before deciding which catalog is right for you, you need to understand your own goals and motivation. What is your goal? And how are you going to measure success?

Identifying the Three Stages of a Data Catalog

A data catalog begins with documenting and defining the data (and metadata) that you have. In fact, the business glossary is a data catalog's most basic function. But a data catalog can do much more than that.

We like to compare the stages of a data catalog journey to the progression a child makes from just learning to crawl, to walking, and finally to running. You don't start out running. The same can be said of building your data catalog competency.

As you progress on your data catalog journey, you may want to move very quickly from the Crawl to the Run stage. But we recommend taking it one step at a time.

TIP
Carefully building your data catalog roadmap is important. Make sure to create realistic, achievable goals based on an accurate understanding of your current strengths.

REMEMBER
No matter where you are on your data catalog journey, you're about to see a big jump in the data maturity of your organization. At each stage, you get a clear set of expectations about what you can expect from your data catalog.

Crawl: Ensuring you have the basics covered

While there are many different applications and use cases for a data catalog, your most basic data catalog should offer:

- » Basic connectors to your most widely and actively used data sources

- » Simple tagging, grouping, and documentation

- » Collaboration tools to allow users to communicate and interact with each other within the context of their data

TIP

For example, if you're on a data team looking to build a catalog to decrease time-to-market for your ML algorithms, don't try to run before you can walk. Focusing your attention on specific key projects early can set you up for success down the road.

Walk: Taking it to the next level

For those moving toward a more mature data management model, you'll want to establish ownership and governance of your evolving data infrastructure as you migrate to cloud-first data storage.

You'll want to look for a data catalog solution that offers semantic and business context capabilities such as business glossaries and logical data model integrations. You'll also be sure to seek out automatic profiling and classification so your data analysts and data citizens can understand the data more easily.

WARNING

Be sure you understand any data analysis techniques — from the algorithms used to where the analysis was carried out. Any data going outside of your firewall can open you up to privacy issues.

Other important features to consider include:

- » Wide-ranging connectivity to a number of different data and analytic sources

- » Integration with lineage and ETL solutions to understand the movement and transformation of data within and across sources

- » Workflow capabilities that enable you to create repeatable, automated workflows driven by events in the system

Run: Making your data catalog work for the enterprise

Making a data catalog work for the enterprise can drive significant benefits. If you're a CIO spearheading your organization's

initiative to better understand your customers and operations and drive new revenue opportunities, then you'll need an enterprise data catalog solution.

To drive an enterprise–grade data catalog initiative, you'll want to make sure your data catalog offers semantic and business context capabilities such as:

» Business glossaries

» Logical data model integrations (across multiple languages and regions)

» Deep data quality to drive trust and understanding of information

TECHNICAL STUFF

Be sure to do a deep dive on any data quality capabilities. For deep data quality and observability, you should have the following metrics:

» Point-in-time

» Time-series quality metrics

» Manual and automated rule engines for fine-grained results

Other important features for the enterprise include:

» Data enrichment capabilities that allow for the creation of an active metadata graph

» Exhaustive connectivity that can be customized based on specific needs

» Embedded data governance and privacy by design to drive trust and compliance based on worldwide data privacy regulations

» Solutions that are architected for secure, enterprise-wide adoption, including compliance with major security regulations

» Integration with chat platforms for flexible engagement across multiple platforms

» Interactive query interfaces so users can use familiar languages (SQL/Python) to investigate their data and popular queries that people have run

Chapter **3**

Maximizing Your Data Catalog's Value

You've got your data catalog in place. Great! But now what? How do you ensure you and your teams maximize the value of this powerful new lever to drive more intelligent, efficient, effective use of data?

You can do several things to get the most out of your catalog so that you can increase the value from your data, cut costs, and make it easier for data consumers to find and understand your data.

In this chapter, you discover some ways you can make your organization's data work for you by improving your data catalog, but you also learn an important truth: You don't need to do everything all at once to make a measurable difference with a data catalog.

Setting Priorities

You've got your data catalog up and running — and this is a really big milestone. The next step is to take stock of where you are as an organization. That'll inform your data catalog roadmap. Your roadmap not only details where you are and where you want to

go, but also a good roadmap gives all your stakeholders — your team, your organization, your customers — confidence and a set of expectations to rally around.

Of course, every organization is unique. Your company is on a unique journey to becoming a mature data culture. If it's like most organizations, it has a variety of data sources, databases, data types, and requirements across the enterprise. Getting every database into the data catalog at once can definitely feel over-whelming. But the good news is you absolutely don't need to get everything in your catalog all at once.

REMEMBER So what should you do first? Focus on the high-demand, high-value data your users need. It doesn't make sense to use resources to pull in data that won't be utilized. Data that is years old or has rarely been accessed is likely to be of low value. Data that is new, relevant, and often used or requested by users offers higher value.

You can do a lot of things to make things easier for your data stewards — that is, the people across your organization who are building the data catalog. But what should you do first?

» Make sure you know what you currently have.

» Do an audit of your existing database landscape and get an understanding of what is business critical and what isn't.

» Focus on high-demand, high-value data. That way you can ensure you're supporting the business and prioritizing as you create your data catalog roadmap.

Once you've identified the core data sources to pull into the cata-log, make sure you set aside time to set up the connections to your data sources.

TIP Choose a data catalog with native connectivity that can discover and register data sources, and ingest metadata using out-of-the-box integrations.

It's a good opportunity to win over your partners — the data stewards, scientists, engineers, business analysts, and privacy managers across your organization who will be leveraging the data catalog.

TAKE ADVANTAGE OF MACHINE LEARNING (ML)

With today's volume and variety of data, it's difficult to scale a data catalog initiative with a purely human-based approach to classifying and building business context into ingested data.

However, this presents a great opportunity to take advantage of machine learning. ML offers these key benefits:

- Automatic data class proposals
- Continuous learning and improved accuracy
- Increased profile and classification accuracy

A data steward can provide oversight by approving or rejecting the ML-proposed data classifications. The ML algorithm learns from the feedback and provides more accurate recognition each subsequent time.

Doing so saves a ton of time for your colleagues who would otherwise be requesting help from your in-demand IT team. By setting up the connections between your data sources and your data catalog, you make it much easier for data owners to choose the databases to register and start the ingestion process.

Leveraging Your Data Catalog for Data Governance and Privacy Initiatives

Your data governance and privacy programs are critical to protecting your business from a range of costly threats. Like many businesses, you may have inherited long-standing privacy risks. The truth is that the biggest data governance headaches come from simply not knowing where your risks are.

So how do you get maximum value out of your new data catalog? Use your data catalog to drive data governance and privacy!

Applying your data privacy, retention, and security policy terms directly to your catalog can make your compliance pain points visible in context, embedding transparency into your organization's DNA. Armed with context, you can create a burn-down list, tackling compliance risks with a plan aligned with organizational priorities. This feedback loop can be your secret weapon for leveling up awareness of the value of a savvy data culture across your entire organization.

Once you get the ball rolling, you're likely to see an organizational shift in the desire to apply governance, security, and privacy principles right from the beginning, which provides a big efficiency gain, too. That means less retrofitting and more thoughtful data design that can preemptively guide your organization toward compliance.

Curating Your Data Collections

Every data source has a life story. Someone (or several someones) originally spec'd the data source. Someone else may have engineered it. Yet another person might be responsible for its day-to-day maintenance. And most importantly, the data source may or may not be a relevant source for analysis.

Treating all data sources equally can result in misleading conclusions. Ask yourself: Where is each data source in its trajectory? Every garden needs a gardener who can reliably answer the question, "How does your garden grow?"

» An abandoned, unwatered data source that stops growing and fossilizes may still be a useful reference, but will lose its value for continuing insights.

» An untended but still-growing data source may be filled with great latent insights, but, relapsed to its wild state of brambles, serpents, and sour fruit, can no longer be trusted.

Assign ownership to every data asset. Deprecate and retire data assets for which no one will take responsibility. Once every data source has a clear line of accountability, you can ensure the right

people are involved and alerted throughout the various governance and privacy workflows. That way a data source with a problem can efficiently become a data owner with a task list.

With a data catalog, you can make asking the resident expert a breeze. You also make it easier for users to avoid low-quality data sources in the first place.

Ensuring Trustworthy Data

It's easy to forklift a ton of metadata into your new data catalog. But if users can't trust the quality and provenance of the data, what's the point? Garbage in, garbage out, as the saying goes. Not to worry.

TIP Remember those owners you assigned? One task of ownership is quality control. Your task as the data leader is to delegate workflow checks that verify each owner, provide descriptions, define and certify metrics, and link lineage and quality information. This could include an assessment of data protection risk points.

The completeness of these checks give you and your business users a multivariate score, an at-a-glance tool for gauging their on-going level of trust in the data source.

TIP You should also enable data consumers to rate the data and provide additional details, feedback, and context. Why? Your consumers are a community of real people with relevant experience, domain expertise, and wisdom you can put to work.

Sure, you've got machine learning (ML), but ML alone only goes so far. In fact, ML works best when paired with real human curation. Your human community of data consumers can put their collective heads together to create meta-knowledge about the data in your catalog, sharing ratings, workflows, business term glossaries, and even contributing subject matter references.

Embracing the social angle of a data catalog informs *other* users in choosing the high-value data *they* need for *their own* purposes.

Creating an Integrated Experience

Business users are dealing with too many disparate applications today. No one wants to constantly move between multiple applications to get the data they need. Now is the time to bring your data together.

The point of an enterprise data catalog is to pull your sprawling data galaxy into a searchable, transparent, contextual, and comprehensive place where everything a business analyst or data engineer needs to build a report or create a new pipeline is together at their fingertips.

Chapter **4**

Powering a Data Intelligence Platform

t's said that knowledge is power. But why do so few of us have it at our fingertips? It's because organizations everywhere are drowning in a data-shaped ocean of chaotic, siloed, and untrustworthy information. The solution? Data intelligence.

It begins with a data catalog. As we cover in Chapter 1, a *catalog* can be defined as an informative list of data assets. *Cataloging* is the art of creating a list of all relevant data assets, and then sharing this list with the world. Data catalogs help your organization take back control of your data. Instead of the chaos of siloed data and data stewards, the data catalog brings your data together.

But a data catalog can do a lot more. A robust catalog can connect not just to data sets but also to analytical assets such as reports and dashboards. If the catalog is well designed, it can also connect with other departmental and tactical catalogs, becoming a "catalog of catalogs." Once cataloged, the data assets can be enriched further to enable multiple use cases ranging from governance and privacy to operations, and all the way to data observability.

The best enterprise data catalogs are integrated into a true data intelligence platform. To reach your full potential, you must embrace active metadata management and transform into a data-driven enterprise. A data intelligence platform should connect all data citizens, empowering them to collaborate, shop for trusted data, and attain optimal business outcomes.

This chapter shows how to do just that. You find out how to activate the metadata graph, enrich your data landscape with data lineage and data quality, extend governance and privacy, and operationalize with workflows.

Creating a Data Intelligence Strategy

Data intelligence is a strategy for leveraging intelligence about data, which then empowers you to use it to drive better business outcomes.

REMEMBER

Data intelligence enables you to understand the creation, movement, and transformation of data throughout the enterprise.

By treating data as a product, you can unlock opportunities for:

>> Cost savings, such as report rationalization, duplicate data spend, and savings from cloud usage

>> Compliance, such as accessing lineage for financial regulatory reporting

>> Opportunities for increasing revenue, such as marketing campaigns and new product development

But data intelligence delivers a lot more than that. With an enterprise-grade platform, you can provide a single pane of glass to discover all data sets within the "four walls" of the organization.

You'll reduce risk by:

>> Proactively identifying data sets not used and remove them to avoid accidental breaches.

>> Proactively understanding impact of any technology or data changes and respond accordingly.

>> Proactively identifying sensitive data and ensuring it is protected and used appropriately.

And you'll monetize data by treating data as a product, unlocking opportunities for monetization with both internal and external stakeholders.

Activating Your Metadata Graph

There are a few ways to enhance your catalog — and they all begin with enhancing your metadata graph.

A *metadata graph* spans diverse information asset types and connects metadata about these assets to give rich content and context to data so that data can be trusted and acted upon.

The metadata graph provides a complete picture of how data is interconnected. The graph connects data and information about the data from different sources. This information can be technical metadata, business metadata, or privacy information, and can include many different types of data (customer, product, employee, for example) across every department. An active metadata graph, for example the one shown in Figure 4-1, helps:

>> Generate greater visibility into the data landscape

>> Evaluate the right data for your needs

>> Deliver faster insights

Your data catalog can be enriched with data lineage. Data lineage creates relationships between data assets and allows you to trace the movement of data through your enterprise, from source to destination. This improves trust in your data assets, which in turn enhances adoption rates.

Data Domain
Customer
Supplier
Employee
Product

Metadata
Governance → Business Metadata
Catalog → Technical Metadata
Privacy → Privacy Metadata
Lineage → Relationships
Quality → Enrichment

Department
Finance, HR, Marketing, etc.

| ● A Data Asset, Insight, Algorithm or User that exists in an organization but is not identified or managed | ○ A Data Asset, Insight, Algorithm or User that exists and is managed by the enterprise data catalog |

FIGURE 4-1: An active metadata graph.

With this knowledge, your business users and data stewards who want to understand the origins of data sets can trust the data in the catalog. And your data engineers are able to do impact analysis of any schema changes on downstream systems.

Policies and workflows make your data catalog truly operational, allowing for optimized effectiveness.

Enriching your metadata graph with policies

To truly operationalize your enterprise data catalog, you need to enrich it with governance and privacy policies.

Policies provide significant value to an enterprise catalog. Policies are driven by regulatory requirements, internal governance standards, and acceptable usage. They can dictate who should have access to what data, enforce data sharing/usage agreements, and provide guardrails that make the catalog truly operational and allow for maximal utility.

As an enterprise catalog becomes the single source of truth for your policies, it's critical that it's designed to meet regulatory requirements, internal governance standards, and acceptable usage guidelines. The data catalog can also provide a single point of entry for data sharing and usage agreements between parties.

When data is governed appropriately, with policies applied intrinsically as part of the governance, it gives everyone the peace of mind to use data offensively to drive revenue, spur innovation, and optimize efficiency. This is where the real value lies.

Operationalizing your data catalog with workflows

Workflows bring people, processes, and tools together by designating specific roles within the organization where processes can be initiated through a single click. Workflows can automate tasks, and through this automation provide the needed context to all stakeholders so that they can more easily and quickly complete their tasks.

REMEMBER

To truly operationalize an enterprie catalog, you can design workflows to support multiple use cases that automate processes, enable collaboration, and speed up the completion of tasks. Here are few examples:

» Shopping for data

» Asking the expert

» Logging a data issue

» Initiating a data quality check

These use cases are just the beginning — your data catalog should include workflows that are configurable and customizable in order to focus on exactly what your business needs.

Chapter **5**

Evaluating Your Data Catalog's Success

S omeone wise once said that anything worth doing is worth doing well. So how do you know how well your data catalog is helping your organization?

Measuring success doesn't have to be complicated. Help yourself by defining clear success metrics, otherwise known as key performance indicators (KPIs). These pesky, persistent metrics make it clear to everyone — your team, your managers, and your broader organization — that you're making progress against your goals for your data catalog and your business.

In this chapter, you find out how KPIs help you evaluate whether and to what extent an endeavor like data catalog implementation is successful. You also find out about three types of KPIs — enablement, adoption, and business-value metrics — and specific types of KPIs you can leverage right away in your organization.

Understanding Why KPIs Are Important

A *data catalog* is a collection of services that help organizations find, understand, trust, and access a repository of technical and business metadata. And the success of any data catalog depends on the value provided to the end users.

The end goal of a data catalog is to become the primary source for data discovery. So, it's important to have a range of KPIs to truly get a good measure of success. You want to have a detailed picture of how your data catalog is performing. Otherwise, how do you know whether you're achieving your goals as a team and an organization?

The good news is you can design your approach to understanding success by simply asking the questions that address how well you're moving toward accomplishing your team and organizational goals.

To begin measuring success via a KPI-driven approach, you can start by creating a list of questions. Here are a few examples of the type of high-level questions that can be the basis for your KPI dashboard:

» Is your data catalog adding value to the organization?

» Is your data catalog a reliable source for all data-related questions?

» Is your data catalog helping users make data-driven decisions?

» Are you improving end user adoption?

» Does your data catalog present a single pane for data discovery?

Of course, it's easy to ask questions. The more challenging part of this equation is collecting and evaluating answers.

Categorizing the Types of KPIs

To know when you're successful, you need to measure the results of your data catalog use. To do so, you can categorize KPIs into three types of metrics:

- >> Enablement metrics: Deployment and education
- >> Adoption metrics: Usage
- >> Business value metrics: Quantifiable outcomes

Enablement metrics: Deployment and education

Your data catalog is implemented and being used. But you're not sure how much it's making an impact on the organization. You've introduced the new data catalog to your data analysts and others. But you really want to know how far and wide the benefits of the data catalog are being felt.

Enablement metrics detail your progress in terms of the number of sources ingested and the extent that the information in your data catalog is complete and has ownership assigned.

Enablement metrics also help you understand how well your end users are trained on the data catalog, and how soon you can implement a use case and provide first value.

REMEMBER Teaching your users how to tackle specific use cases is a powerful method for driving deployment and adoption.

Your enablement metrics could extend to a very wide range of signals. But to give you a solid starting point, here are a few essential ones that can get you off to a great start:

- >> **Number and percentage of terms and data sets with assigned stewardship**

  The value: These KPIs help identify how well you're doing with assigning ownership.

- >> **Number of terms and metrics identified for definition**

  The value: These KPIs help identify how well you're doing with completeness.

- >> **Number of assets identified as personally identifiable information (PII)**

  The value: This KPI helps support compliance and regulatory requirements.

>> **Number of new data sources onboarded**

The value: This KPI helps you understand your data catalog's ability to ingest and display metadata from various sources. It also helps answer the question, "Were improvement areas identified?"

REMEMBER As your catalog matures, integration with disparate data sources will become more simplified.

>> **Percentage of data sets with logical data dictionary; percentage of data sets with physical dictionary**

The value: This KPI helps you know how well you're doing with data lineage.

>> **Number and percentage of users trained**

The value: These KPIs help you know how familiar users are with your data catalog.

>> **Number of data sources connected**

The value: This KPI helps you understand your platform's degree of openness.

>> **Cycle time to deploy a feature**

The value: This KPI indicates the time required to implement a use case. By the way, it should be minimal.

TIP When you have a higher number of data sources integrated, you understand more about your data catalog's openness and ease of integration. The number of data sources connected answers two crucial questions:

>> Does the data catalog serve as a single source of truth for data-related questions?

>> Was cost-efficiency achieved in managing the data catalog?

Adoption metrics: Usage

You're integrating your data catalog, adding a glossary, and ingesting data sources. You have a functioning data catalog. But the problem — and it's a common one — is that the people in your organization aren't using it.

EDUCATION AND DATA CATALOG SUCCESS

It's important to teach your team and relevant data professionals across your organization how to use your data catalog.

There is a smart way to get an educational initiative off the ground, and it includes a Training Needs Assessment.

An assessment helps identify the tasks each user must complete as well as assesses the current ability to complete those tasks. You can then identify the training required to bridge that gap in your team's data catalog skills.

Additional ways to advance the skills of your data catalog users include:

- Role-based education: Including different aspects of data governance and data catalog along with regular training/coaching sessions

- Internal "Data Catalog University": Create an education portal with a content and training path for various data governance roles in the organization

- Data governance onboarding: Create a data governance standard as part of new staff onboarding

- Knowledge repository: Create a central place to store all the knowledge gathered during the program, such as the data catalog terminology, documentation, business and technical requirements, solution architecture, use cases, roles, relevant and updated roadmap
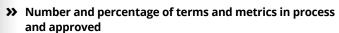
Of course, it's no good if you roll out a new data catalog and no one uses it. *Adoption metrics* help you understand the extent that data catalog adoption is progressing across your organization. A clear understanding of your adoption metrics help you define how you can improve and increase the use of your data catalog. Then you can use these metrics to benchmark your progress.

The following metrics can help in identifying whether your data catalog is achieving your adoption goals, including:

REMEMBER

» **Number of data set access requests; percentage approved**

The value: These KPIs help you understand how well your data catalog and, specifically, your data shopping (or data sharing) feature is being leveraged.

Access requests are a key adoption indicator you want to track.

» **Number and percentage of terms and metrics in process and approved**

The value: These KPIs help you understand your metadata maturity.

» **Number of unique logins; percentage of users logging in**

The value: These KPIs signify the maturity of your data catalog. A large number of logins or a high percentage of users logging in helps indicate the comfort level that data professionals feel in using the data catalog for data discovery requirements.

Alternatively, you can measure the number of users on-boarded across different user personas. For example, you can look at how many data stewards, data owners, analysts, data scientists, and data engineers are using your data catalog. Differing levels of adoption can be a good indicator of where to focus your education efforts.

» **Number of users added weekly/monthly**

The value: This KPI sheds light on the pace of your data catalog adoption. A weekly increase in this number means that existing platform users have given good feedback to users who have yet to onboard. A faster, more seamless user onboarding likely shows a higher number and answers the question: "Did I achieve cost-efficiency in driving adoption of the data catalog?"

» **Number of search queries per day**

The value: Search is a core capability of a data catalog. This KPI shows how often users come to the data catalog for their data discovery needs. The question answered here is: "Does your data catalog act as single source of truth for all data-related questions?"

>> **Number of asset pages viewed**

The value: This KPI measures whether your users are becoming more active on the data catalog. If the value remains the same or declines month-to-month, adoption is slowing down. An increase in the value of this metric helps answer the question: "Are you improving end user adoptions?"

TIP

To encourage adoption, it's important to choose a data catalog that's user friendly. You can make your data catalog UI more user friendly and recognizable by:

>> Updating your data catalog's colors to match your corporate style

>> Changing default icons and logos

>> Creating a welcome dashboard

>> Customizing your email templates

Business value metrics: Quantifiable outcomes

Is your data catalog delivering value to the business? That is the bottom line. Is it helping you achieve your goals? Are you seeing the outcomes — to productivity, efficiency, and much more — that you need to keep your business on track?

Business value metrics indicate how well your data catalog is driving business value to end users and the organization.

These are quantifiable outcomes, and the related KPIs can include:

>> **Increased productivity**

The value: The KPIs for productivity demonstrates whether your catalog increased the productivity of your data teams.

It answers questions such as: How many hours do users spend searching for the right data and requesting access to data? This KPI helps answer the question, "Does your data catalog help users make data-driven decisions?"

>> **Daily active users (DAUs)**

The value: The more active users of your data catalog, the more you can hypothesize that it is trusted and data

professionals are actively using the data catalog as their primary source of data discovery needs. This KPI helps answer the question: "Is your data catalog platform a single source of truth for all data-related questions?"

>> **Number of unique data assets pages viewed**

The value: This KPI helps understand the baseline value of your data catalog. As your data catalog matures, new data assets are added, governed, and curated. Therefore, if users view many unique data assets pages, you can infer your catalog's value proposition of being the single source of truth for data discovery needs is true.

REMEMBER It's important to control your progress, measure your success, and revise your roadmap.

Following Three Essential Strategies

Now that you know the value of KPIs and the three main KPI categories you'll want to understand the success of your data catalog. Focusing on these key strategies can help you sustain the success of your data catalog initiative.

>> **Manage stakeholders' feedback.** Identify existing and potential stakeholders (decision-makers) for each use case. Stay in contact with them and keep them updated on progress so you can closely manage their feedback and concerns.

>> **Define and refine your KPIs.** Continue to define and refine the KPIs that are most critical for your organization, the data catalog program, and your use cases. Set up a method to monitor progress and communicate to people about key milestones.

>> **Revise the roadmap regularly.** By defining and following a process of new use case identification and prioritization, you ensure that your data catalog roadmap continues to make a relevant impact on your organization.

REMEMBER Always focus on the business value your use cases offer.

IN THIS CHAPTER

» **Support the entire enterprise**

» **Connect to all your data**

» **Enable automation at scale**

» **Gain visibility with data lineage**

» **Drive trust with collaboration**

» **Embed data governance and privacy**

» **Establish a data marketplace**

Chapter **6**

Seven Steps to Data Intelligence

C hapter 4 covers data intelligence — a strategy for leveraging intelligence about data, which them empowers you to use it to drive better business outcomes. Sounds good, right? So in this chapter, we leave you with seven steps to making sure your organization achieves full data intelligence with a data catalog.

Support the Entire Enterprise

So where should you start? Start with a data catalog that's strategically deployed to support the entire enterprise. If you have multiple tactical and technical catalogs by department or data source, you only continue to create additional silos.

To maximize the value of an enterprise data catalog, make it the single source of truth by cataloging all sources across the organization.

The data catalog should be part of a comprehensive platform that is flexible to meet the needs of organizations early in their data

intelligence journey, but also can grow with the needs and maturity of the business including enabling governance, security, privacy, and quality use cases.

Focusing your attention on specific key projects early can set you up for success down the road.

Connect to All Your Data

It might sound obvious, but to reach data intelligence, you need to know your data. Without the ability to connect to *all* your data sources, you can't get the visibility you need.

You must be able to connect to a wide range of data sources, including databases, data warehouses, data lakes, cloud platforms, ETL tools, BI tools, ERP, and CRM systems. With today's digital transformation initiatives, strong integration with cloud platforms will streamline cloud migration.

Enable Automation at Scale

With the explosive increase in data volume and variety, your organization can't continue to take a manual, human-based approach to data intelligence. It's important to invest in automation for data discovery, classification, stewardship, and policy enforcement. Take advantage of machine learning-based automation to minimize manual tasks.

You can get faster time-to-value from your data catalog with automatic classification of data. Plus, you can leverage configurable workflows that notify data owners of issues and enable them to take action.

Gain Visibility with Data Lineage

You need to understand how data flows through the organization with end-to-end lineage to ensure it can be trusted. Map relationships between systems, applications, and reports to provide a context-rich view across the enterprise.

- » Provide lineage from key source systems, SQL dialects, ETL tools, and BI tools.

- » Enable the business to view lineage diagrams to trace data flows from source to destination.

- » Use technical lineage to visualize the flow of data and assist with use cases such as impact and root cause analysis.

71 percent of business analysts say that data lineage makes it easier to see where data has come from and how it has changed—allowing your teams to spend more time on strategic initiatives (2020 commissioned study conducted by Forrester Consulting on behalf of Collibra).

Drive Trust with Collaboration

You can drive even more trust in your organization's data by collaborating with others about the data. Better collaboration:

- » Enables employees to easily share, reuse, and collaborate with data.

- » Allows users to leave feedback including ratings, reviews, and comments to accelerate data usage and adoption.

- » Breaks down organizational silos by enabling the sharing of knowledge across the organization.

- » Automates processes such as data access requests and data certification right within the data catalog without the need to use email.

Embed Data Governance and Privacy

An enterprise data catalog provides the organization a one-stop shop to find, understand, and access data across the enterprise. But how the data is accessed and by who is just as important as finding the data. Enabling access to just a few users defeats the purpose of the catalog, but providing unabated access to all data by all users poses risks.

Most organizations need to comply with corporate, regional, or global regulations to protect access to data for security and

privacy purposes. Without governance embedded into the catalog, you can't properly protect your data, nor can you ensure that you can fully trust the data.

Your goals should be to:

» Establish standardized governance processes across all data, along with a centralized business glossary

» Provide the ability to certify data sets

» Identify sensitive data such as PII that needs to be protected

» Enforce governance and privacy policies and data access controls

Policies ensure only the right people can access the right data for the right purposes.

Establish a Data Marketplace

Perhaps most importantly, data intelligence can enable both technical and business colleagues to use data more intelligently and get more value from it.

REMEMBER

64 percent of business analysts say shopping for data in a central location enables effective outcomes (2020 commissioned study conducted by Forrester Consulting on behalf of Collibra). A data catalog should help democratize data and eliminate the need to rely on IT to request access to trusted data.

To support digital transformation, you need to provide a self-service experience via an intuitive, easy-to-use interface for users to request access to their data. This access request process should have governance embedded to ensure access is compliant.

Implement a data catalog in this digital-driven world

Every modern organization is drowning in data. It's increasingly becoming a valuable asset that needs to be visible, understood, and trusted in order to drive your organization's profitability, innovation, and growth. Data catalogs are your solution to the ever-growing data challenge. They pull everything together across your data landscape and help you make sense of it. With *Data Catalogs For Dummies* in your hands, you learn how you can comprehensively foster data intelligence to drive data-driven decisions.

Inside…

- Take steps to control your data
- Identify the right data catalog
- Maximize the value of your data catalog
- Use your catalog for governance and privacy
- Build a foundation for data intelligence
- Measure success against metrics

Collibra

Bineesh Babu is the data catalog Center of Excellence Lead at Collibra. **Ben Moser** is the Product Manager focused on Collibra Data Catalog. **Chandra Papudesu** leads Product Management for Collibra Integrations and Data Lineage. **Luis Romero** leads Product Marketing for Collibra Data Catalog, Integrations and Data Lineage.

Go to Dummies.com™
for videos, step-by-step photos, how-to articles, or to shop!

for dummies
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.