Turning analytic assets into actionable ideas

How a better understanding of your organization's information can help you do more with AI

Highlights

- Unites both internal and external data sets—along with other analytic assets, like machine learning models—into a single, metadata-rich catalog, with AI-powered search-and-suggest capabilities to help users find what they need for self-service analytics

- Allows data stewards to track usage, prioritize integration projects and lessen the time needed for preparing data and enforcing governance policies

- Provides choice and flexibility for data location, allowing highly regulated industries to realize benefits from their analytic assets, without requiring them to move their information from wherever it's currently stored

Think for a second or two about all the things you know: the quickest route to your office, how to make a phone call, where to find a good cup of coffee, your mother's birthday. The list could go on for a very long time. But because it's all stored in your own brain, you can usually find what you're looking for in milliseconds.

Now think about all the knowledge that resides within your entire organization—not just data, but all the models, analytic assets and other things that help make it all useful. Once again, the list goes on and on. And it's not just the information itself. It's that the information is fluid. Consider, for example, all the stakeholders who can access it, change it and use it for their own projects. When it's your job to help harness all that knowledge and take advantage of it to help make better decisions, it can be difficult to know where to begin—or how long it will take.

So much information, so little order

It can take a lot more time for data scientists, business analysts and knowledge workers to find and prepare the information they need than it does for them to use it and extract value from it.

Why? Because as the amount of information within any organization increases, so does the challenge of zeroing in on the right information. For one thing, information is rarely saved in a single place. Instead, it's likely to be distributed across an array of shared locations, company networks, content management systems, file systems, cloud apps, email servers and even personal hard drives.

The process can seem like walking into a six-story library filled with hundreds of thousands of books, periodicals, videos and recordings, all randomly distributed on shelves and in boxes—with no index or catalog to help you find what you need. It's an exercise in wasted time and missed opportunities. And it will put your company behind competitors with a better handle on their data.

Shopping for answers

Clearly, having a catalog to help you find the information you need would be a big help. But a catalog that simply lists what's available and where to find it is only the beginning—especially in an environment containing millions of entries. For example, let's say you want to purchase a blue T-shirt. Type "blue T-shirt" into a search engine query box and you'll find more than 3 million results. And even if you narrow it down by changing your request to "shop for blue T-shirt," you're still left with more than 160 million results. Yes, it's roughly half as many, but still not much help.

How can you improve on that? If you think of a search engine as a somewhat unrefined catalog, you might find a specific shopping site to be a better resource. For example, a well-designed shopping site would let you chose what kind of "blue" you'd like, whether it's for a man, woman or child, the size you need, and so on. It would also offer options for shipping and reviews to help you determine the quality of both the T-shirts the site offers and the service it provides. And you could browse as many shopping sites as you like—until you locate exactly the right blue T-shirt for you.

IBM Watson® Knowledge Catalog is an intelligent, cloud-based tool that works in much the same way as that shopping site, cataloging content in its entirety, with context included—instead of offering just a title and the name of whoever created it. It's available as a software as a service solution on the IBM CloudTM. In addition, it's a built-in component of IBM Watson Studio, an integrated, cloud-based environment designed to make it easy for data scientists, developers and business analysts to develop, train and manage models and deploy AI-powered applications. Either way, it's a far more valuable resource for locating the content you select, along with rich metadata that tells you:

- Where it's been
- Who else has accessed it
- How it's been used
- What others have to say about it

That means it can offer data scientists and analysts both more context for the information they're researching and the potential to drive collaboration. Plus, it can save everyone involved a lot of time.

How Watson Knowledge Catalog changes the game

At the heart of Watson Knowledge Catalog is a powerful cataloging engine that indexes all the available data sets and analytic assets your business can access, regardless of whether they live in a data warehouse or transactional system, or even in a set of spreadsheets. They can be structured or unstructured—thanks to Watson Natural Language Understanding—and stored on premises or hosted in the cloud. What's more, the catalog can include external data sets and sources, such as proprietary data services to which your company subscribes, or open data APIs.

In addition to providing a single source of truth about your data sets, the catalog also offers a single point of access. AI-powered search-and-suggest capabilities allow Watson to recommend relevant data and analytic assets in the catalog, based on its understanding of relationships between assets and how they're being used and socialized among users in existing projects. The available metadata can help you understand what you've found and determine whether it's useful, so you can quickly discover and apply the most relevant assets needed for model training and creation. Plus, embedded self-service data preparation tools can cut the time it takes to get the data ready for productive use in analytics and AI applications.

With Watson Knowledge Catalog, you can also see how data and other assets are being used across the organization. So, if someone has already combined data set X with data set Y, you could use that work as a foundation for your own project. Or you could collaborate with them on their work.

To improve findability even further, the catalog lets users tag and comment on data sets and analytic assets, enriching the metadata and adding extra context to help coworkers find what they need. Users can also take advantage of built-in data discovery algorithms that use machine learning to auto-classify the contents of each data set—so authors don't need to annotate the data manually.
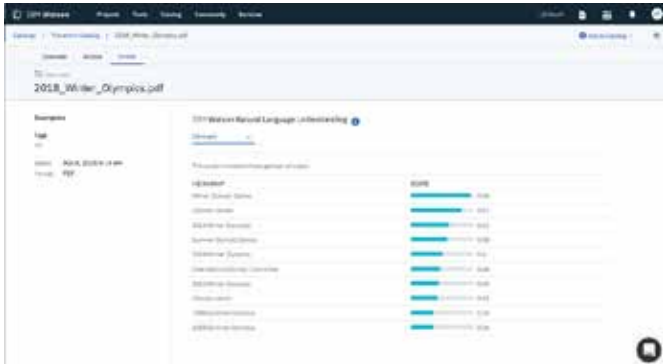
Figure 1. Watson Knowledge Catalog offers a security-rich, enterprise-level environment that allows users to curate and shape analytic assets, including machine learning models and both structured and unstructured data—taking advantage of Watson Natural Language Understanding—to fuel data science and as well as AI, machine learning and deep learning.

Figure 2. Take advantage of automatic, dynamic masking of sensitive data elements to unlock information previously unavailable to AI application development and power collaborative development—making information a trusted enterprise asset while protecting it from misuse.
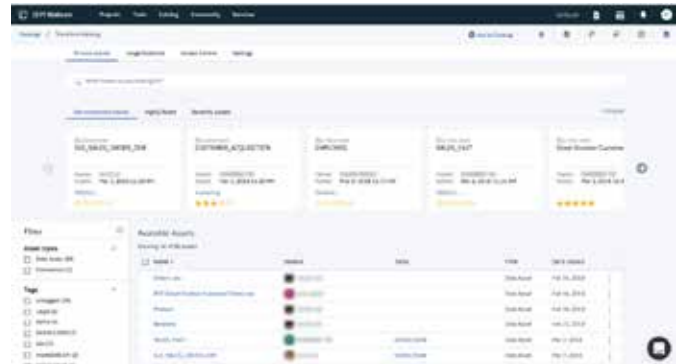
That intelligent metadata model of your assets offers an industry-leading approach to automatically enforcing data governance for these assets, as well. With a built-in governance policy manager and engine, Watson Knowledge Catalog ensures that the policies and rules describing how your data should be used are enforced each time a data asset is accessed. And as a result, it helps data governance become an actionable way to proactively protect your business data.

The catalog also helps data stewards in the CDO's office by tagging and classifying data sets and automatically tracing their lineage and usage. It even leverages a built-in business glossary to standardize business terminology across your data. As a result, it's easier for stewards to understand what each data set contains, where your sensitive or personal identifiable information resides and who should be allowed to access it.

Take advantage of automatic, dynamic masking of sensitive data elements to unlock information previously unavailable to AI application development and power collaborative development—making information a trusted enterprise asset while protecting it from misuse.

Self-service access to analytic assets

Watson Knowledge Catalog offers data scientists, developers and business analysts self-service access to a great deal more than just data. That includes:

- **AI-powered search-and-suggest tools** that can guide you to the most relevant analytic assets in the catalog—based on an understanding of the relationships between assets, usage of those assets and social connections among the users of those assets
- **Enhanced support for unstructured data**—including text—which can automatically profile and classify text assets
- **Policy-driven data masking,** which allows for sharing a wealth of information, while protecting sensitive data from misuse with on-the-fly masking of sensitive data elements—unlocking data that's previously been unusable for data science and business analytics while supporting compliance with industry and enterprise policies
- **Integrated data preparation** that can help data professionals easily handle and prepare data found in the catalog—and improve productivity—with IBM Data Refinery, which is fully integrated into the catalog
- **Integration with Information Governance Catalog,** which can help you leverage your existing curated data sets

Why IBM?

IBM Watson is a cognitive system enabling a new partnership between people and computers. Watson continuously learns from previous interactions, gaining value and knowledge over time. With Watson, organizations are harnessing the power of AI and cognitive computing to transform industries, help professionals make better decisions, scale expertise, and solve complex challenges.

For more information

To learn more about how IBM Watson Knowledge Catalog can help your organization's data scientists and analysts gain easy and direct access all the data sets and analytic assets your business can access, please contact your IBM representative or IBM Business Partner, or get started here: **ibm.com**/cloud/watson-knowledge-catalog