Designing a modern data catalog at Microsoft to enable business insights

In this digital age, data is a company’s lifeblood. Growing recognition of data’s true value has spawned entirely new industries such as online advertising and ecommerce. Other traditional enterprises are adding value to their products by using data to gain insights and offer services that their customers need. What does this mean for technical decision makers and their leadership? If you are committed to digitally transforming your company, treat your data as a strategic asset.

For many enterprises, this is more easily said than done. The larger the company, the more barriers data professionals encounter as they attempt to access needed data: Where is the data in each department? Who owns it? Which tables’ contents are curated? Is the right data being used? The promise of using new technologies such as machine learning (ML) to deliver best-in-class user experiences falls short when the underlying data—the fuel that drives the AI modeling engine—lacks quality. As the saying goes, garbage in generates garbage out.

This inability to discover and manage data holistically in the enterprise stems from how departments, infrastructure, teams, and data are traditionally siloed. This silo structure makes it extremely difficult for people to locate the data they need to answer their business questions. Any query that requires data stored outside a single silo involves tracking down the appropriate group, requesting and gaining authorization from the data owner, looking up data in unfamiliar physical tables, and attempting to manually join information from different sources in a manner that produces the correct output. Most people consider this process so onerous that many important business questions are simply not answered.

This article describes how we at Microsoft view our data as a strategic asset. Here, we discuss the approach that Microsoft Core Services Engineering and Operations (CSEO) is taking to create a connected data catalog that reduces the time required to discover and use trustworthy data.

Treat your data as a strategic asset

Today, companies can no longer afford to limit their data usage to single siloed sources. Technical decision makers need connected data that spans departments to help the company do business better. Data is the core driver for your success and your competitive advantage, but only when you use it to its full potential. Connecting trusted, curated data from multiple sources will help you better understand your customers and their needs. Access to such information might confer a competitive advantage today but extracting more meaning from your data will be critical to your company’s survival in the near future. Enterprises must gain insights into customers’ business challenges, and how well a company addresses those needs will soon define its success or failure.

As a technical decision maker, how can you help move your organization’s data from siloed, unintuitive tables to an easily discoverable, connected data pool that enables the enterprise to gain strategic insights that drive its digital transformation? Furthermore, how do you give people access to this wide array of data without violating regulations such as the General Data Protection Regulation (GDPR), the Sarbanes-Oxley Act of 2002 (SOX), or the Health Insurance Portability and Accountability Act (HIPAA)?

At Microsoft, a key aspect of how we’re addressing the challenge of making trusted, connected data more readily available to users is by creating a modern data catalog. Moreover, our strategy is to develop a data catalog that not only enables discovery of the right connected data, but also enhances data governance.
The modern data catalog: enabling discoverability and governance

At a minimum, a data catalog is an information map that lists the data residing in one or many data sources. However, it can be so much more. As part of our mission to modernize our data assets, the CSEO team has designed a common, cross-department data catalog that connects data sets across the company, reduces the time to discover and use trustworthy data, and helps enforce proper data governance. In the following sections, we discuss some of the key discoverability and governance features we’re building into our data catalog.

Making data more accessible and discoverable

To enable CSEO to help leadership use our data to ask the big questions—the ones whose answers lead us to strategic insights—we needed to run data searches that can connect data from multiple departments. A data catalog is instrumental in performing such cross-department searches. In addition to listing many disparate data sources, our new data catalog offers an intuitive interface that simplifies the entire data discovery process. Our data catalog:

- Spans multiple departments, providing a holistic view into virtually any aspect of the company’s operations.
- Provides “recipes” with instructions that vastly simplify how to utilize connected datasets for key business scenarios.
- Displays intuitive logical names (such as Customer) instead of physical table names (such as “CUST09”) to improve discovery.
- Highlights quality level, certification, and source-of-truth indicators for each data asset and the corresponding scenarios where it is accessed to help users assess its appropriateness for use.

The graphic below illustrates the strategy that we’re employing with our data catalog: we’re expanding its scope of included data and enhancing its functionality so that anyone can use it to ask business questions and gain valuable insights.

Figure 1. Our data catalog development strategy involves a multiphase approach that ultimately will enable business users to obtain the answers to their data questions with the click of a button

- Phase 0 is siloed data. Where each department’s data is locked into its own silo, preventing discovery of and access to data.
- Phase 1 is connected data for developers. This is where developers work with a data catalog to identify and access cross-departmental data to develop data-sharing structures.
- Phase 2 is curated data for analysts and data scientists. This is where analysts and data scientists use the data-sharing structures to access quality data and follow instructions in recipes to define query objects that answer business questions.
• Phase 3 is ready-made data objects for everyone. This is where business users select the appropriate object to get the answer to the question they have. The object automatically forms the correct query and accesses the connected data.

What this means for you
• **Break the silos.** Developing a data catalog might not be a priority if you’re limiting your data search to a single set of siloed data, but these simpler queries won’t answer the big questions or support predictive analytics. Using such advanced techniques requires working across departments and incorporating different groups’ data—and that necessitates building a data catalog. Remember: data is the key ingredient that enables people to ask the big questions that ultimately produce the deepest insights. As you work across departments, include as many different data types as possible in your catalog, because you can’t predict which data will produce the next big discovery. At Microsoft, we’re taking a holistic approach, building out a data catalog that provides our data scientists with visibility into our entire rich assortment of data. This deep, high-quality information is precisely what our AI models need to identify important trends.

• **Follow the recipe.** Instead of having your data catalog present technical lists and unfamiliar data structures that are difficult to decipher, consider designing a more intuitive interface. For example, we’ve built logical models that map to the data assets, as well as “recipes” that provide easy-to-follow instructions for how to generate the right connected dataset for a given business scenario. Building these recipes is a collaborative effort that involves technical teams who design the framework/interface, and business process experts who can identify the appropriate data sources and provide instructions for how these “ingredients” are best used for each scenario.

Applying data governance without stifling discovery
A good data catalog does more than ease discovery. It also helps drive data governance by building controls into the catalog’s data discovery process to ensure that only people with appropriate need and authority can access sensitive data. With the right mechanisms in place, a data catalog can even require the requestor to fill out additional forms before user access is granted, to promote compliance with various government regulations. Data catalogs can also incorporate traceability, tracking who is accessing what, how often the data is accessed, and how it’s being used. Our data catalog:
• Gives management a clear understanding of the data we have and what those data represent, which is key to applying appropriate policies to ensure compliance.
• Includes certification attributes that flag vetted data, then prioritizes data from these certified tables as the recommended source to be used to build connected data sets. This helps us utilize the highest-quality data while bypassing other tables that reside in the catalog whose information might be less accurate (such as test data).
• Spreads data ownership responsibility across the teams and individuals who understand its business value.
• Incorporates traceability to monitor usage of data assets and enforce protection.

Technical decision makers can use our data catalog’s dashboards to quickly scan all the organization’s data, review health information, then drill into details to identify potential issues that can impact governance and subsequently take appropriate action for remediation.
Figure 2. We’re designing an intuitive interface that technical decision makers and IT leadership can use to scan the data sources, review health and other conditions, identify potential issues, and quickly remediate.

1. Clicking the **Customer** hot spot in the data map displays summary information about the health of the Customer data.

2. This health table highlights that non-certified asset consumption is higher than desired. Click **Consumption of non-certified assets** to view consumption details.

3. You can easily identify which non-certified data is used the most, and then take appropriate action.

**What this means for you**

- **Incorporate governance into your data discovery processes.** As you develop your data catalog’s cross-departmental capabilities, you should consider inserting mechanisms that enhance governance. You can design your data catalog to become the primary tool that helps you identify how your data is being used in the enterprise, and then use that information to enforce security policies that help maintain compliance. At Microsoft, we’re incorporating the following governance capabilities into our data catalog:
  
  - Sending automatic access requests for restricted data to the data owner as soon as the need for that data is identified.
  
  - Providing links to forms to fill out if additional authorization for access is required.
  
  - Tracking who is using sensitive data, how frequently that data is being pulled from the data catalog, and where the data is going.
  
  - Setting retention policies based on how frequently data needs to be accessed, such as setting an auto-expire to every six months.
  
  - Using the data catalog’s telemetry to identify areas of improvement for data management processes, such as redundant data sets.

- **Align each data set’s curation process with its related business process.** This entails both a cultural change and a process change, because to manage the data catalog successfully, you need to forge partnerships between the developers who define the data catalog structures, and the business people who have a good understanding of the value the data and how each data element should be used. Employees who become the data owners must understand the value of supplying and maintaining their data. You also need to streamline the data management process as much as possible. Don’t assume that you can just create a form and expect people to fill it out and continue to manage their data on an ongoing basis. Instead, invest time to review the business, and how business users are actually orchestrating their data. Then, give users tools to move the information into the catalog that align with that process.

The goal is to simplify data entry and management so that people are encouraged to maintain their data. For example, we’ve created auto scanners that monitor our data lake and automatically register a data file on behalf of the developer the moment it’s added to the data lake. The appropriate business person—the data owner—is then notified and asked to fill out details of the new data table files that have been placed into the data lake for them. We’ve also created a Microsoft Excel export capability to simplify how people add metadata to the catalog.
• **Certify and promote your best data.** Find the right balance between including only the best curated data (which would be extremely limiting) and adding everything to your catalog, including noncurated and test tables which can add large amounts of poor-quality data. At Microsoft, we’ve taken a best-of-both-worlds approach. We allow most data to be catalogued, but we also apply several attributes to each data source that describe the nature of the data, including whether it’s certified. Certified data has passed a validation process to ensure that it’s properly curated. Furthermore, both certified data and other higher quality data are surfaced in the discovery experience (such as search results) so that data seekers always use the highest-quality data in the data catalog.

**Conclusions**

Technical decision makers are under immense pressure to help the business transform into a digital company—and the foundation of any digital company is its data. Companies must use their data more holistically, promoting a means for people and AI to synthesize information across the enterprise to gain deeper, more accurate insights that unlock new business capabilities. However, the data that is commonly siloed in different departments in the enterprise is a major stumbling block; you can’t use what you can't discover.

Creating a modern data catalog is foundational aspect of finding and connecting your data. If it includes data from all the enterprise’s departments, a data catalog can disrupt your old business model and deliver insights into the state of your business, what your customers need, and more. Furthermore, when designed properly, a data catalog can dispel the notion that data discoverability conflicts with governance. Data catalogs that incorporate traceability and other governance features make it easier for leadership to understand what data they have, how it's being used, and whether the regulatory mechanisms they have in place are effective.

At Microsoft, our strategy is to build a holistic data catalog that incorporates both discoverability and traceability. Our data catalog connects information from many different areas and ensures that information is curated to help consumers leverage it appropriately. This ensures that we can ask the big questions, that the right conversations are happening, and that the best AI models are being built. Our connected data catalog is helping us discover new ways of adding value to our products and services, aligning our capabilities to benefit our customers now and in the future.

**Our next steps**

We continue to enhance the scope of the data sets in the catalog, adding more meta-data to simplify the query process. We began with key business operations data and “recipes” for data engineers who need connected data sets to support key business domain questions and create connected experiences. Next, we plan to add additional capabilities that offer guidance for data scientists to find the data (including non-curated) that they need to build new AI models. We’ll also add recipes for data analysts to help them answer questions for their line of business.

As we continue to add data to our catalog, we envision empowering every employee to leverage, utilize, and share the data catalog. Our data catalog will be key to helping people find the connected data they need quickly, taking advantage of its functionality to gain insights and create data-driven, intelligent user experiences that continue to propel us along our digital transformation journey.

**For more information**

**Microsoft IT Showcase**

Driving effective data governance for improved quality and analytics

Microsoft tames the 'wild west' of big data with modern data management

Transforming business through the power of data

[microsoft.com/itshowcase](microsoft.com/itshowcase)