Azure Data Catalog

Customer Fact Sheet

# Introduction

Microsoft Azure Data Catalog is a fully managed cloud service that serves as a system of registration and system of discovery for enterprise data assets. Azure Data Catalog provides capabilities that enable any user – from analysts to data scientists to developers – to register, discover, understand, and consume data assets.

# Why Azure Data Catalog?

Traditionally, discovering enterprise data sources and related information assets has been an organic process based on undocumented, tribal knowledge. This presents numerous challenges for companies wanting to get the most value from their information assets.

* Users are not aware that data assets exist unless they come into contact with it as part of another process; there is no central location where data assets are registered.
* Unless a user knows the location of a data asset, he cannot connect to the data using a client application; data consumption experiences require users to know the connection string or path.
* Unless a user knows the location of a data asset's documentation, he cannot understand the intended uses of the data; data and documentation live in different places and are consumed through different experiences.
* If a user has questions about an information asset, he must locate the expert or team responsible for the data and engage those experts offline; there is no explicit connection between data and those with expert perspectives on its use.
* Unless a user understands the process for requesting access to the data source, discovering the data asset and its documentation still does not enable him to access the data he requires.

While these challenges face data consumers, users responsible for producing and maintaining information assets face challenges of their own.

* Making the data they produce discoverable by the people and teams who should use it can be an ongoing challenge without a central location for data asset registration and discovery.
* Annotating data assets with descriptive metadata is often a lost effort; client applications typically ignore descriptions stored in the data source.
* Creating documentation for data assets is often a lost effort; keeping documentation in sync with the data asset data asset is an ongoing responsibility, and users lack trust in documentation as it is often perceived as being out of date.
* Restricting access to the data source, and ensuring that data consumers know how to request access is an ongoing challenge.

Creating and maintaining documentation for a data asset data asset is complex and time-consuming. The challenge of making that documentation readily available to everyone who uses the data asset data asset is often even more so.

When combined, these challenges present a significant barrier for companies who want to encourage and promote the use and understanding of enterprise data assets.

# Service Description and Capabilities

Data Catalog is designed to address these problems and to enable enterprises to get the most value from their existing information assets, by making it easy for users to discover the data sources they need, understand the data sources they discover, and consume the right data sources in their tools of choice. Data Catalog also enables users to contribute their own knowledge into the catalog, through tagging and annotating existing data assets, or through registering additional data assets to make the catalog more complete.

Extend

Figure - Azure Data Catalog Capabilities

As illustrated in Figure 1, Data Catalog provides a cloud-based metadata repository into which data assets can be registered. The data itself remains in in its existing location, but a copy of the metadata, along with a reference to the data asset location, is added to Data Catalog. This metadata is also indexed to make each data asset easily discoverable via search, and understandable to users who discover it.

Once a data asset has been registered, its metadata can then be enriched, either by the user who performed the registration, or by other users in the enterprise. Any user can annotate a data asset by providing descriptions, tags, or other metadata, such as documentation and processes for requesting data asset access. This descriptive metadata supplements the structural metadata (such as column names and data types) registered from the data source, to make discovering and understanding it easier.

## Register data assets

Data asset registration is performed using the Data Catalog data asset registration tool. This application is launched from the Data Catalog portal.

The registration process involves three basic steps:

1. Connect to a data asset - the user specifies the data asset location and the credentials to connect to the data source, such as a SQL Server instance.
2. Select objects to register - the user selects the objects in the specified location that should be registered with Data Catalog. This may be the full set of tables in all databases on the server, or a specifically selected subset of tables and views.
3. Provide annotations – the user can optionally provide tags and experts to be associated with the objects being registered, and can choose to include additional metadata from the data source, such as previews and data profiles.
4. Complete registration - the user completes the process, and the data asset registration tool extracts metadata from the data source, and sends that metadata to the Data Catalog cloud service.

Users can also register data assets with Data Catalog using the manual entry capabilities in the Data Catalog portal, and by using the Data Catalog REST API.

## Explore, discover, and understand data assets

The goal of registering data assets in Data Catalog is so that they can be discovered, understood, and used by users across the enterprise. Data Catalog allows users to discover via search, and to browse and filter through the catalog contents as well.

Each data asset returned will be displayed in a tile view or list view, based on the user’s preferences, with basic metadata included. By selecting a data source, users can view its complete metadata, including column-level metadata where available.

## Enrich data asset metadata

Users can also provide their own metadata for the data assets they discover. The ability to annotate data assets is not limited to the users who register the assets; each Data Catalog user can add his own tags and descriptions for registered data assets and their attributes. Data Catalog will track the value and source of each annotation and will display the user who added each annotation. This crowdsourcing approach to metadata ensures that every user with a perspective on the data and its use can share their opinions and resources with the user community at large.

This user-supplied metadata delivers significant additional benefits:

* The registered data sources are more easily discoverable. The user-provided metadata is added to the Data Catalog search index. This allows users to discover the data by using terms and concepts that may not be present in the original data source. For example, if a database table that contains customer data is named "tbl\_c45", providing a friendly name of "Customer" will make it more easily discoverable by users looking for customer data. Similarly, providing a description that includes the names of reports, dashboards, or processes that use the data will make the data asset easier to find by users who use those downstream artifacts as their search terms.
* The registered data sources are more easily understood once discovered. The user-provided metadata is presented to any Data Catalog user who views the annotated data source, which helps provide additional context and information. Most data sources typically do not include meaningful descriptions or documentation, and those that do are often focused on the technical Database Administrator or database developer audiences. By enriching data sources in Data Catalog with audience-appropriate descriptions and tags, users can help ensure that those who discover the data can understand its details and intended use.
* Each registered data asset can include request access information, so that users can easily understand and follow existing processes to request access to the data asset and its data.

## Remove data asset metadata

After a data asset has been registered, it can sometimes be necessary to remove the data asset reference from Data Catalog. This may be due to changing business requirements, or to the source system being retired. Regardless of the reason, Data Catalog makes it easy to remove data sources by simply selecting to delete so that they can no longer be discovered and consumed.

Deleting a data asset from Data Catalog only deletes the metadata stored in the Data Catalog service. The original data asset is not affected in any way.

## Managing registered data assets

Users can take ownership of data assets by selecting the “Take Ownership” option in the Data Catalog portal. No special permissions are required to take ownership of an unowned data asset; any user can take ownership of an unowned data asset. If a data asset is already owned, users cannot simply take ownership – they must be added as co-owners by an existing owner. Any owner can add additional users or security groups as co-owners.

Data asset owners can control the visibility of the data assets they own. To restrict visibility from the default – where all Data Catalog users can discover and view the data asset – the asset owner can toggle the visibility setting from “Everyone” to “Owners & These Users” in the properties for the asset. Owners can then add specific users and security groups.

## Consume data sources

The ultimate goal of data discovery is to find the data that you need, and to use it in the data tool of your choice. The data consumption experience in Data Catalog enables this capability in multiple ways.

1. For client applications that are directly supported by Data Catalog, users can click on the *Open In* menu in the data asset tile found in the portal. The client application will then launch with a connection to the selected data source.
2. For all client applications, users can use the connection information displayed in the properties pane for a selected data source. This information includes all details (such as server name, database name, and object name) required to connect to the data, and can be copied into the client tool's connection experience. If request access details have been provided for a data source, this information will be displayed next to the connection details.
3. With a single click of a button, users can add ADO.NET, ODBC, OLEDB and JDBC connection strings for a selected data source to the clipboard, for easy pasting into the connection editor in their tool of choice

# Common Scenarios

Data Catalog’s capabilities can help organizations get more value from their data sources and broader information assets in many different situations. Some common scenarios faced by many organizations include discovering and understanding high-value data sources, supporting self-service business intelligence and analysis, and capturing tribal knowledge that would otherwise not be used to support the business.

## Registration of central data sources

Organizations often have a number of high-value data sources. These data sources include line of business OLTP systems, data warehouses, and business intelligence / analytics databases. Often the number of systems, and the overlap between systems, grows over time as the needs of the business evolve, and as the business itself evolves through mergers and acquisitions.

Often it is difficult for users to know where to locate the data within these data sources. Questions like these are all too common:

* Of the three HR systems used within the company, which should I use to create this type of report?
* Where should I go to get the certified sales numbers for the fiscal year that just ended?
* Who should I ask, or what is the process I should use to get access to the data warehouse?
* I don’t know if these numbers are right – who can I ask for insight on how this data is supposed to be used before I share this dashboard with my team?

In this scenario, Data Catalog can help. The central, high-value, IT-managed data sources that are used across the organization are often the logical starting point for populating the catalog. Although any user can register a data source, having the catalog kick-started with the data sources that are most likely to provide value to the largest number of users will help drive adoption and use of the system. For customers getting started with Data Catalog, identifying and registering the key data sources used by many different teams of data consumers can be the first step to success.

This scenario also presents an opportunity to annotate the high-value data sources to make them easier to understand and access. One key aspect of this effort is to include information on how users can request access to the data source. Data Catalog allows users to provide the email address of the user or team responsible for controlling data asset access, links to existing tools or documentation, or free text that describes the access request process. With this information included in the catalog, users who discover registered data sources but who do not yet have permissions to access the data can easily request access using the processes defined and controlled by the data asset owners.

## Self-service business intelligence

Although traditional corporate business intelligence solutions continue to be an invaluable part of many organizations’ data landscapes, the changing pace of business has made self-service BI more and more important. Self-service BI allows information workers and analysts to create their own reports, workbooks, and dashboards without relying on a central IT team – or being restricted by that IT team’s schedule and availability.

In self-service BI scenarios, it is common for users to combine data from multiple sources, many of which may not have previously been used for BI and analysis. Although some of these data sources may already be known, there is frequently a process to discover what must take place to locate and evaluate potential data sources for a given task.

Traditionally, this discovery process is a manual one: analysts will use their peer network connections to identify other people who work with the data being sought. Once a data asset is found it can be used, but the process repeats itself again for each subsequent self-service BI effort, with multiple users performing the same redundant manual process of discovery.

With Data Catalog, users can break this cycle of redundant effort. Once a data asset has been discovered through traditional means, an analyst can register the data asset to make it more easily discoverable in the future. Although the user can add more value by annotating the registered data assets, this does not need to take place at the same time as registration. Users can contribute over time, as their schedules permit, gradually adding value to the data sources registered in the catalog.

This organic growth of the catalog content is a natural complement to the up-front registration of central data sources. Pre-populating the catalog with data that many users will need can be a motivator for initial use and discovery. Enabling users to register and annotate additional sources can be a way to keep them, and their peers, engaged.

It’s also worth noting that although this scenario focuses specifically on self-service BI, the same patterns and challenges apply to large-scale corporate BI projects as well. Any effort that involves a manual process of data asset discovery is an effort that can add value to the organization through the use of Data Catalog.

## Capturing tribal knowledge

How do you know what data you need to do your job, and where to find that data?

If you’ve been in your job for a while, you probably just know. You’ve gone through a gradual learning process, and over time have learned about the data sources that are key to your day to day work.

When a new employee joins your team, how does he know what data he needs to do his job, and where to find it?

Odds are, he asks you.

This ongoing transfer of tribal knowledge is part of the data asset discovery process in organizations large and small. More senior and experienced team members have built up knowledge over the years, and newer team members have learned to ask them when they have questions. The most vital information often exists only in the heads of a few key people, and when those people are on vacation or leave the team, the organization suffers.

Sometimes these data experts will make the effort to document their knowledge, sharing it via email, or in Word documents on a team SharePoint site. Although this can be valuable, it introduces a new discovery problem – how do people know what documentation exists, and where to find it…

Data Catalog provides a location for sharing this tribal knowledge, and for making it easily discoverable. Data experts can directly annotate data assets, and can also include links to existing documentation. Not only does this capture the knowledge itself, but it also puts the knowledge in the same experience that is used for data asset discovery. When someone uses the catalog to discover a data asset not only will he find the source itself, he will also find the expert’s knowledge that previously existed only in the mind of the expert himself.

Discovering and understanding data sources and their use is the primary purpose of registering the sources. When enterprise users need data for their efforts (which could be business intelligence, application development, data science, or any other task where the right data is required) they can use the Data Catalog discovery experience to quickly find data that matches their needs, understand the data to evaluate its fitness for purpose, and consume that data by opening the data asset in their tool of choice.

# Get Started with Azure Data Catalog

Data Catalog provides capabilities for discovering and understanding enterprise data sources, to help organizations to get more value from their existing data. Data Catalog empowers users to spend less time searching for data sources and more time working with the data they need.

Get started with Data Catalog today. Visit [www.azuredatacatalog.com](http://www.azuredatacatalog.com) to create a catalog for your organization.