

## CHAPTER 15



# Azure Data Catalog

Throughout this book, you have learned that data can come from many different sources and many different formats. Specifically speaking, the source of the data for this book has come from a number of devices and sensors. The different sources and formats are two of the three Vs mentioned at the beginning of this book: variety and volume.

You have also learned that this data can be stored in a number of different data stores in Microsoft Azure, including Azure Data Storage, Azure Data Lake Store, Azure SQL Database, and more. But these data stores are just a few of the data store options in the cloud and within your organization.

Looking at this through multiple lenses of a DBA and an end user, how do you know what data is in your organization or company and the value it has? How do you figure out and understand all the different sources of data and, more importantly, even remotely begin to gain insight into the data?

Assume for a moment that you are starting a new job today as a DBA or someone who needs access to critical data via reports or other means. Figuring out all the different data sources and what data those data sources contain is a challenge. Typically, users (whether a DBA or end user) have no idea a data source exists until they either discover it by accident or via another process.

The solution to all of these challenges is a central location where users can discover data sources and understand the data that is in those data sources easily so that they can get more value from existing information. This is where Azure Data Catalog comes in.

## What Is Azure Data Catalog?

Data discovery can be looked at from two sides; those producing the data and those consuming the data. Data producers have the challenge of securing the data: admitting those who need to have access and restricting those who don't need access. They also have the challenge of documenting and annotating data. Today, if I want to know what data is in a folder, I look at the folder name, which is not the best documentation.

Data consumers also have challenges, and a few were mentioned earlier. Data discovery challenges for those looking for data are typically in the class of not knowing data exists or where to go looking for it. Once found, the consumer may or may not have access to the data and, if they do, they spend more time browsing the data to understand it and its intended use. The frustration doesn't stop there because if they have a question about the data, the new challenge becomes one of finding the owner of the data.

Azure Data Catalog was created to address every single one of these challenges and more. As a fully-managed cloud service (just like all of the other Azure services), the intent and goal of Azure Data Catalog is to make data discovery simple, help users easily understand the discovered data, and to get the best value possible out of said discovered data.

Azure Data Catalog doesn't move any data around. Instead, it keeps the data in its existing location but tags the data with metadata, which is in turn stored in Azure Data Catalog for easy discoverability. Azure Data Catalog also allows users to contribute to the catalog by tagging and annotating data sources that have

already been registered to further enrich the discovery capabilities, and to register new data sources for discovering. At its core, Azure Data Catalog is an API like the other services in Azure, which provide many benefits for working and integrating with Azure Data Catalog.

## Scenarios

The most basic scenario is that of plain and simple data discovery: finding what data is out there, what data do you need to do your job, and who owns it. Beyond this scenario there is one other scenario that you probably would not think of, and this of self-service business intelligence.

Self-service BI lets users create their own reports and dashboards without waiting or relying on different teams or organizations to do the development. In BI scenarios, it is common when building reports and dashboard for data to be pulled in from multiple sources. The majority of these sources are known, but at times some of these data sources are not known.

Azure Data Catalog helps in the effort of not only the data discovery of the multiple data sources but to allow users to build their reports and dashboards from the discovered data sources easily. From there, the users can contribute to the growth of the catalog and add value to the existing data sources.

So, with this background, the rest of this chapter will build a data catalog, register a couple of datasets, and then connect to them.

## Working with Azure Data Catalog

The following sections will take a detailed look at working with Azure Data Catalog. As you learned earlier, Azure Data Catalog makes it easy to discover different data sources. For the purposes of the example in this chapter, I created a couple of data sources for you to register. As shown in Figure 15-1, I created an Azure SQL Database and an Azure Data Lake Store account. The Data Lake Store account has a single file that contains the sensor data. The Azure SQL Database contains a single table that contains supporting data. Very simple. The AdventureWorks sample database can be downloaded from Codeplex (<http://msftdbprodsamples.codeplex.com/>).

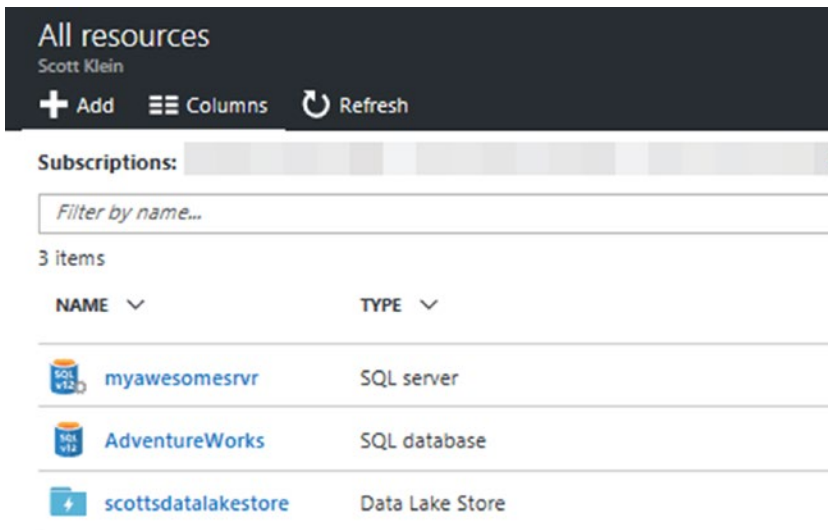


Figure 15-1. Data sources

Let's get started. First, you'll create the Data Catalog and register the data sources. The following sections will walk through that.

## Provision Azure Data Catalog

This section will be pretty short, and I'll explain why. There are a couple of details that I will point out that will save you a lot of time. First, each Azure Data Catalog requires an organizational account. Not a Microsoft account such as a Live ID, but an account that is tied to your organization with Active Directory. Second, only a single Azure Data Catalog can be created per organization. Note that I said "organization," not Azure subscription.

The reason for this restriction is that Azure Data Catalog is intended to be a system of records for all data sources across the enterprise. OK, that makes sense, but at this point you are asking yourself "How can I play around with and test Azure Data Catalog at home?" Great question, and luckily there is an answer.

SQL Server MVP Melissa Coates wrote a blog post that walks through how to do this using a Microsoft account (for example, a Live ID). It essentially entails creating an Azure Active Directory Account, allowing that account to be co-administrator in your Azure subscription, and then signing into the Azure Data Catalog portal using the new Azure Active Directory account. So please read Melissa's awesome blog post:

[www.sqlchick.com/entries/2016/4/20/how-to-create-a-demo-test-environment-for-azure-data-catalog](http://www.sqlchick.com/entries/2016/4/20/how-to-create-a-demo-test-environment-for-azure-data-catalog)

Follow the blog post step by step and you will have an Azure Data Catalog up and running in just a few minutes. Now, I did try this in the new portal since Azure Active Directory is supported in the new portal but I think there must have been a hiccup or something because it let me create an Azure Active Directory but then it didn't show up in my list. So, feel free to try it in the new portal. I haven't reached out to the PMs yet on this issue but I will do so to see if it is just me or if this is a bug. However, it works in the old portal so if you run into issues, hop over to the old portal and have at it.

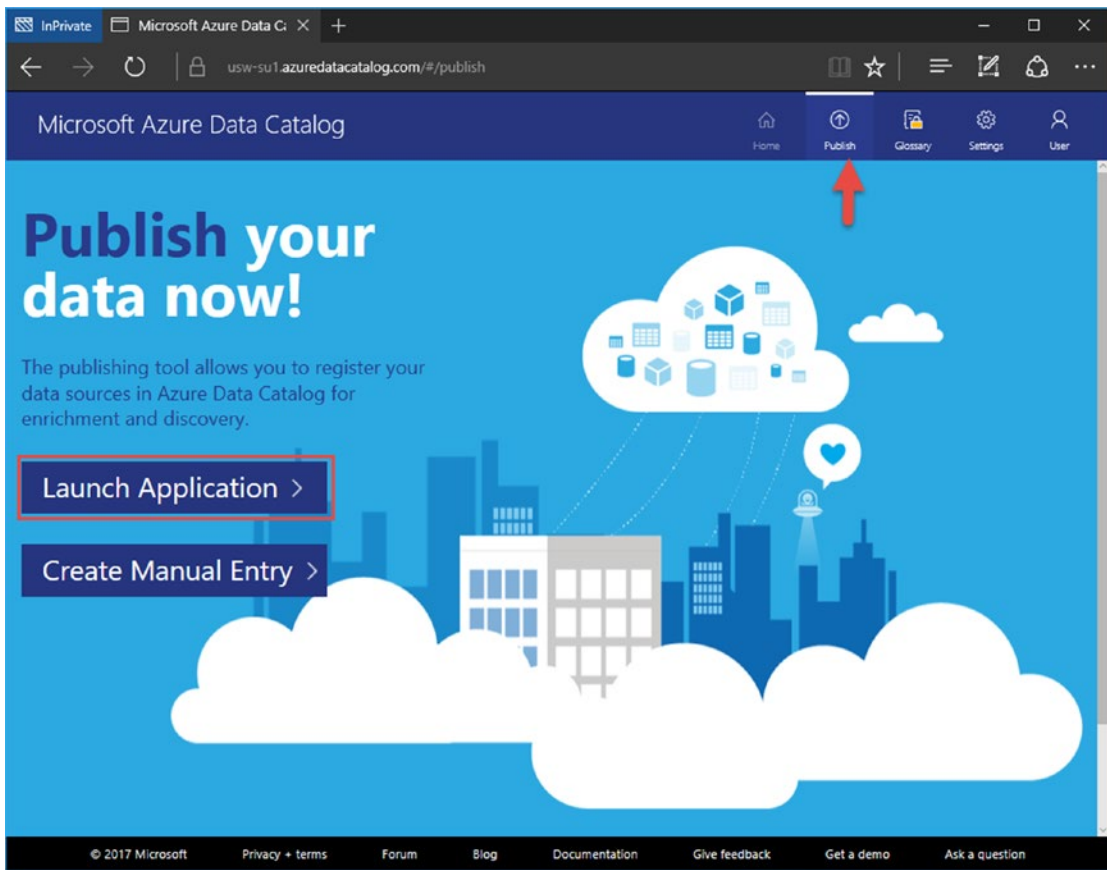
When you are done, be sure to thank Melissa for the awesome blog post and then pop back over to this chapter to pick up where you left off, which is the next section on registering data sources in your new Data Catalog.

## Registering Data Sources

So you have your Azure Data Catalog created. The next step is to register some data assets, which is the process of pointing Azure Data Catalog at a data source and extracting key metadata from the data source including names, location, and other vital information. Part of the registration process copies that metadata into the catalog but the data assets remain in their original location. Copying over the metadata into the Data Catalog is which makes the data sources more discoverable and easier to understand.

To register a data source, open your favorite browser and navigate to the Azure Data Catalog home page at <https://azuredatacatalog.com>. The home page of the Azure Data Catalog includes several links along the top to publish data and configure settings for the Data Catalog. The home page also shows any registered and pinned assets as well as any saved searches for quick access.

Since you have no registered assets, your first step is to defining a data source and register assets. To do this, you can either click the big blue Publish Data button in the middle of the home page or click the Publish button near the top right. Clicking either will take you to the Publish page, shown in Figure 15-2.

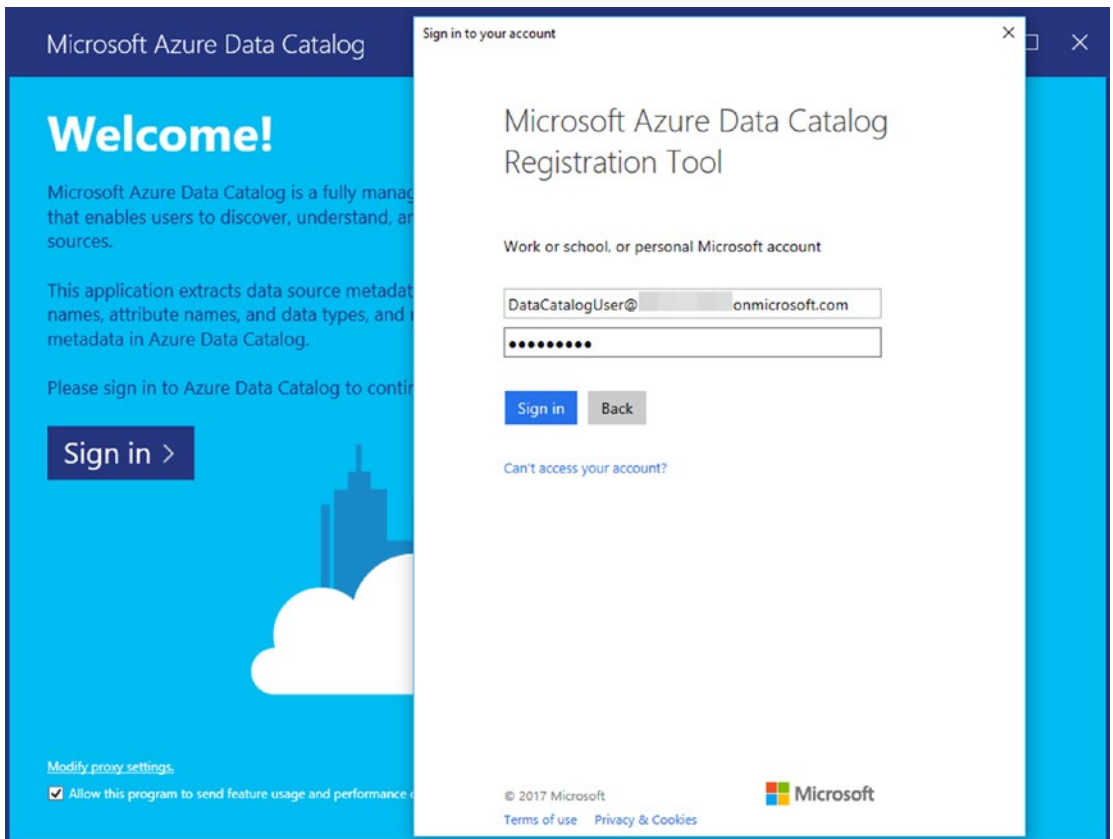


**Figure 15-2.** Launching the registration tool

On this page you can create a manual entry, which I won't walk through here, but the easier choice is to click the big blue Launch Application button. The application is a simple Windows application that runs on the local computer and allows you to register a data asset. So, go ahead and click the Launch Application button.

The registration tool will download and install. The download will take a minute or two and then install. You may need to click Accept as part of the install. Once installed, it will automatically run and you will be presented with the Welcome page.

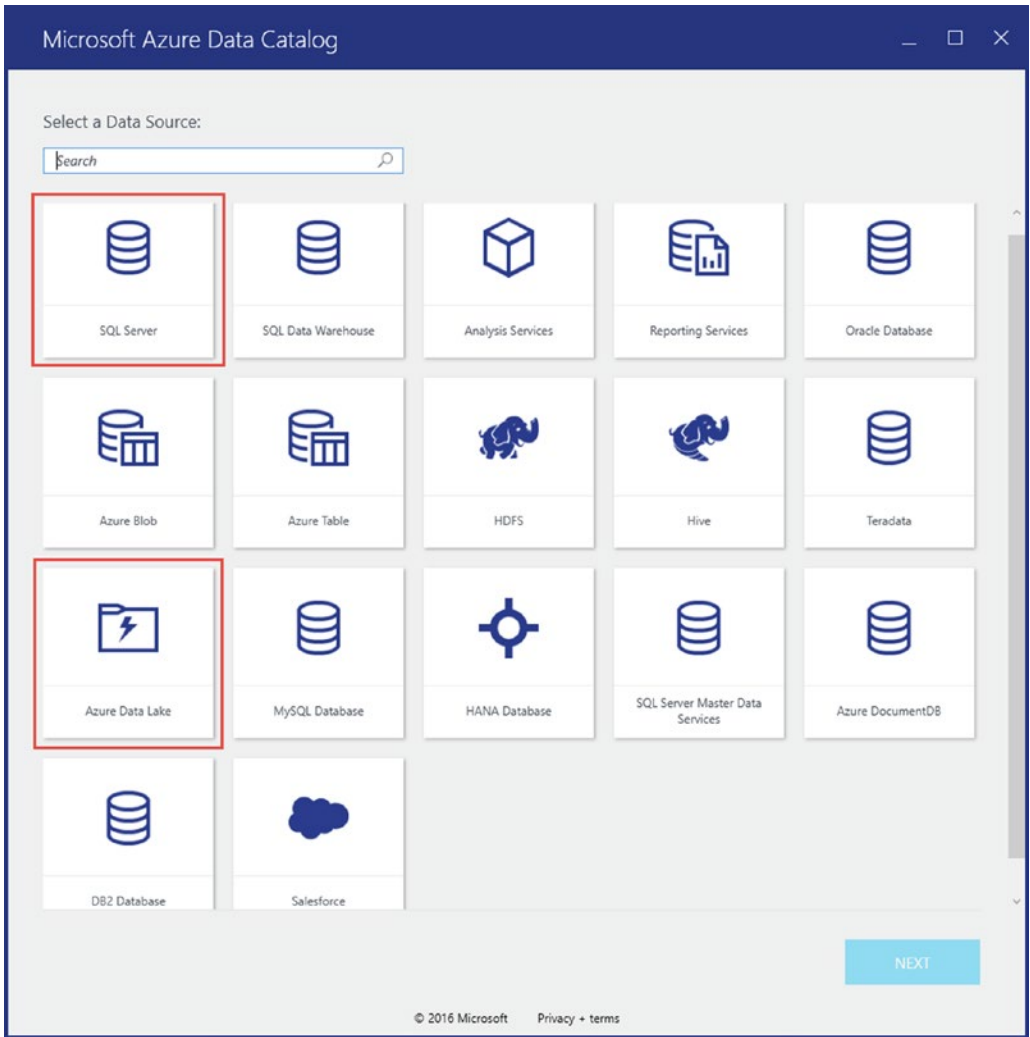
On the Welcome page, click the Sign In button, which will pop up the Sign in dialog, shown in Figure 15-3. In this dialog, use the Active Directory user you created as part of the Azure Data Catalog creation. Since you are registering data assets with Azure Data Catalog, you need to authenticate via the registration tool with Azure Data Catalog.



**Figure 15-3.** Signing in to Data Catalog

Once authenticated, you will be presented with a list of data sources to connect to and from which to register data assets. As you can see from Figure 15-4, there are currently 17 options that range from relational to non-relational data sources, including SQL Server, SQL Data Warehouse, Hadoop, Azure DocumentDB, Azure Table and Blob Storage, and more. Users of Azure Data Catalog can work with other supported data sources through APIs, manual entry, or through the registration tool. I won't list them all here, but the following URL shows all the data sources supported today and through which method they can be accessed and published.

<https://docs.microsoft.com/en-us/azure/data-catalog/data-catalog-dsr>




**Figure 15-4.** Azure Data Catalog data sources

For this example, select SQL Server (which also allows you to connect to Azure SQL Database) and Azure Data Lake Store. I will walk you through registering data assets from Azure SQL Database, and your homework assignment will be to do the same for Azure Data Lake.

In the Select a Data Source page, shown in Figure 15-4, click SQL Server and then click Next. In the Server Connection page, shown in Figure 15-5, enter the server information and click Connect. For this example, I created an Azure SQL Database and associated server. The server is called myawesomeserver, and when I created the database, instead of creating a blank database, I elected to use the sample AdventureWorksLT database.

## Microsoft Azure Data Catalog

 SQL Server

Enter the location of the data source to be registered:

Server Name:

Authentication Type:

User Name:

Password:

Database (required for Azure SQL Database):

Encrypt Connection

These credentials will be used to connect to the data source and to extract the metadata for the objects you select.

**Figure 15-5.** Connecting to Azure SQL Database

Also, since I am using Azure SQL Database in this example, I am using SQL Server Authentication instead of Windows Authentication (since my Azure SQL Database is not part of an Active Directory). You can read more about connecting to Azure SQL Database or SQL Data Warehouse via Azure Active Directory at <https://docs.microsoft.com/en-us/azure/sql-database/sql-database-aad-authentication>.

Once the information is filled out, click Connect.

After clicking Connect, the next page will show the list of objects from which you can register assets, as shown in Figure 15-6.

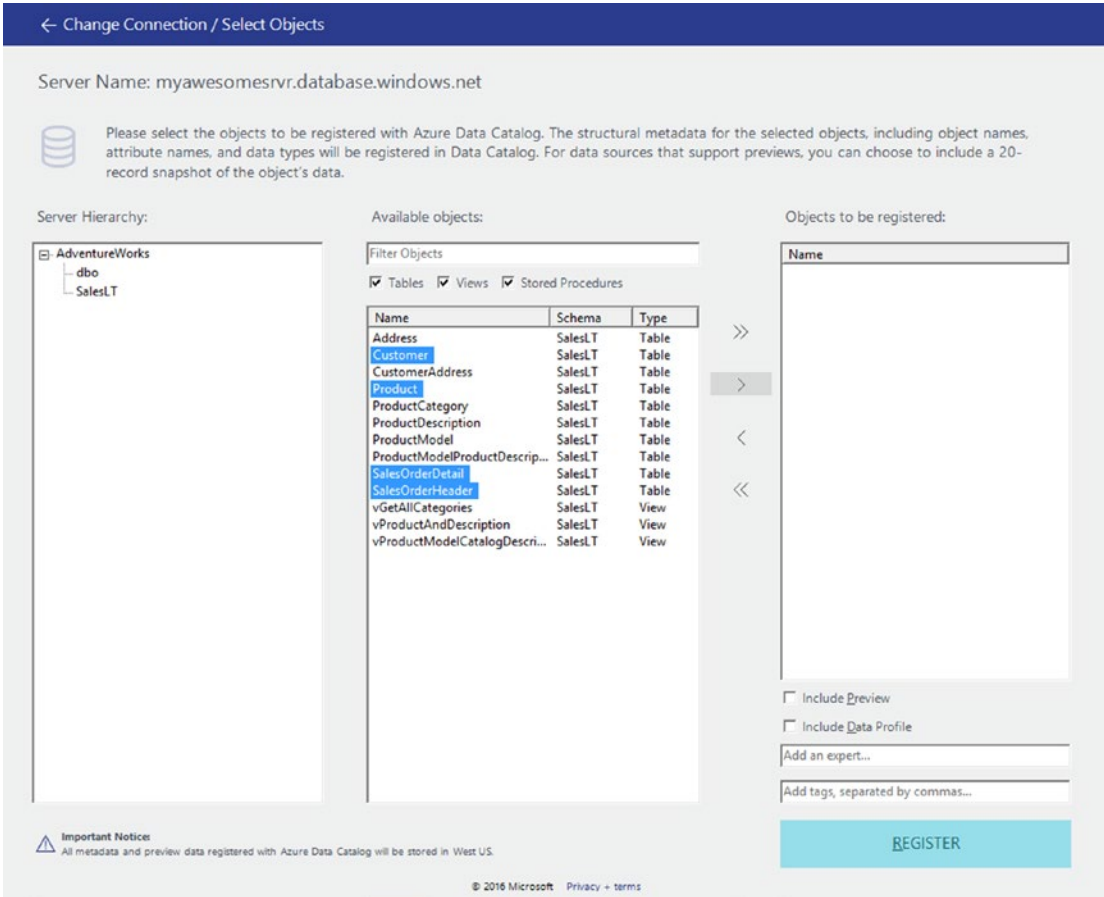
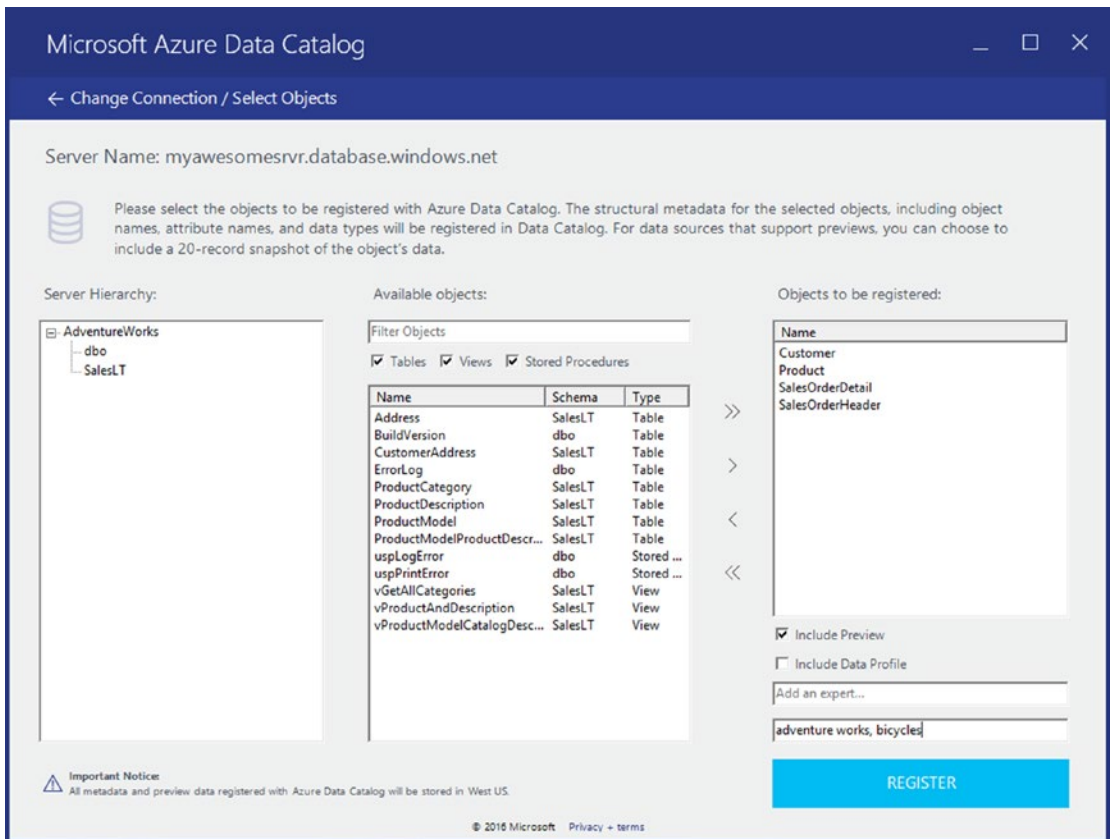


Figure 15-6. Selecting the objects to register in the Data Catalog

On the left is the server hierarchy, which shows a list of schemas in the database. In the middle is a list of available objects that pertain to the selected schema. These objects, as you can see in Figure 15-6, include tables, views, and stored procedures. You can remove or include from the list by unchecking or checking the appropriate checkbox above the list. The right section will list the objects you wish to register.

To register the metadata of the assets (objects), Ctrl+click and select any number of objects from the list of objects. In this example, I have selected the Customer, Product, SalesOrderDetail, and SalesOrderHeader tables. Once selected, click the selected arrow to move those objects to the “Objects to be registered” list, as shown in Figure 15-7.



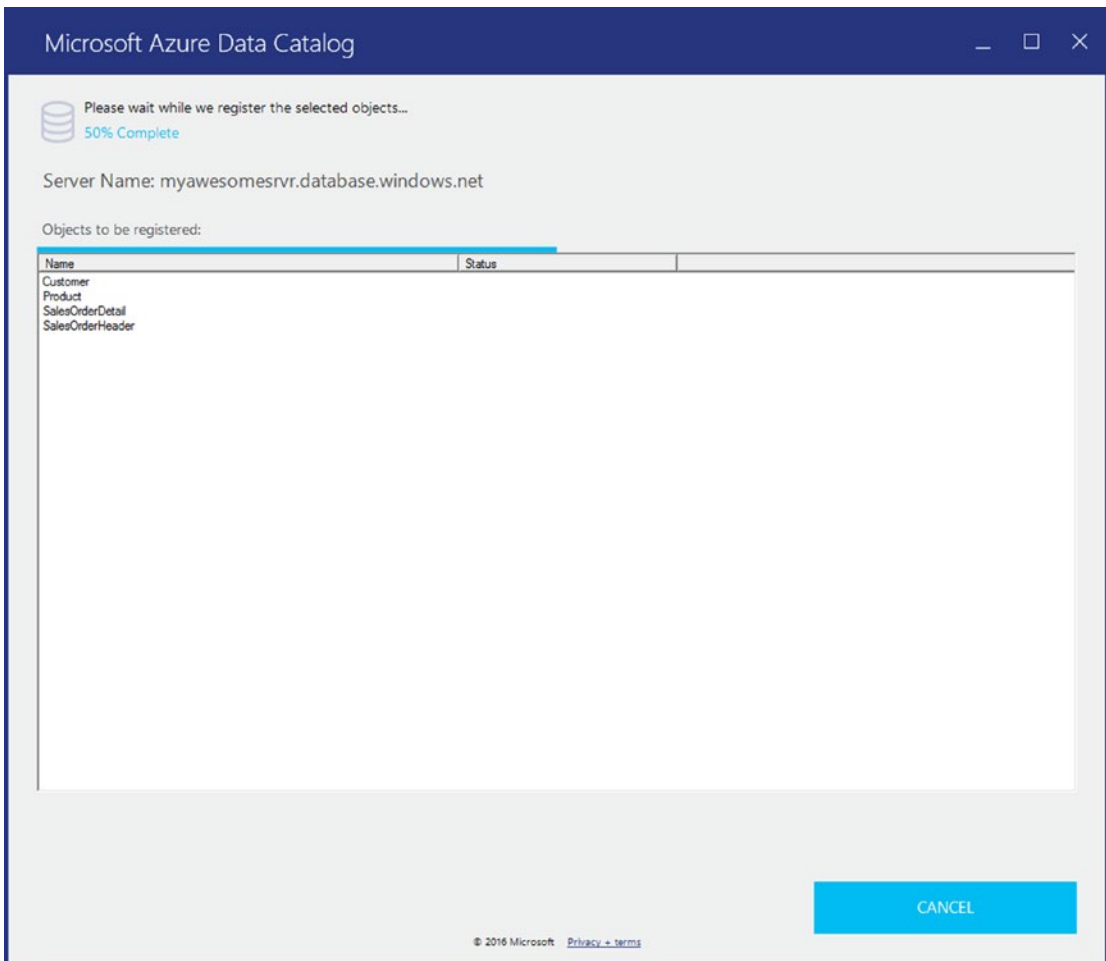


**Figure 15-7.** Adding tags to selected objects

Here is where the fun begins. Before clicking the Register button, now is the time to add tags. In the Add tags text box, enter a few tags that will help users find the data source. For example, I entered *bicycles* and *adventure works*. Adding tags adds search tags for the selected assets in the list. The tags should be descriptive and make it easy for users to find the data source and identify the type of data in the data source.

Also, be sure to check the Include Preview checkbox. This will include a snapshot preview of the data when registering the data source. Checking this checkbox will copy up to 20 records from each table selected and copy them to the data catalog, which will allow users to get a quick visual as to the type of data in the asset. The Include Data Profile option will include a snapshot of the statistics for each object. For this example, I left that option unchecked.

Once you are finished on this page, click Register. Azure Data Catalog will then go through the object selected and register that asset. Figure 15-8 shows the registration process. Depending on the number of assets selected, the registration process could take a while. Luckily, there is a percent completion on the page so you know roughly how long it will take and the status of the process.

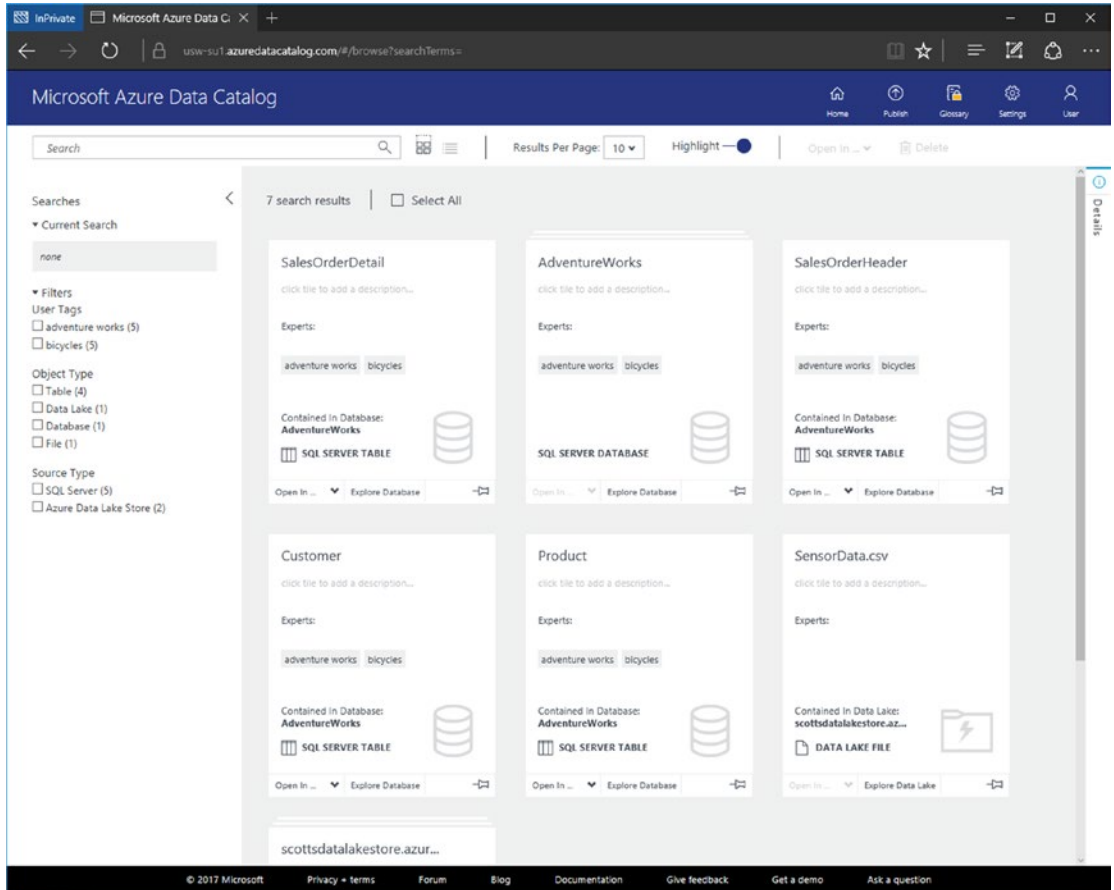


**Figure 15-8.** Registering the objects

As I mentioned earlier, your homework assignment is to go through the registration process again and this time select the Azure Data Lake data source and pick the file sitting the Data Lake Store. The registration process is similar so you should have no problems.

When the registration process is complete, you can register more objects (click this to go back and register the Azure Data Lake data source) or click View Portal (click this when you are done registering all the assets). Clicking the View Portal button opens up the Azure Data Catalog portal, displaying the registered assets, as shown in Figure 15-9.

Figure 15-9 shows the four tables from the Azure SQL Database, the AdventureWorks database itself, and the SensorData.csv file from Azure Data Lake Store.



**Figure 15-9.** Registered objects in Azure Data Catalog

Before moving on, let's spend a few minutes on this page. Along the left are filters you can apply based on the registered assets. You can filter by tag, object type, and source type. Very cool. Along the top is the option of viewing the assets in grid view, as shown in Figure 15-9, or list view, as shown in Figure 15-10.

<input type="checkbox"/>	NAME	DES...	E...	TAGS	CONTAINED IN	SOURCE TYPE	OBJECT TYPE	LAS
<input type="checkbox"/>	SensorData.csv				scottsdatalakestor...	Azure Data Lake :	File	1/11
<input type="checkbox"/>	scottsdatalakestor...					Azure Data Lake :	Data Lake	1/11
<input type="checkbox"/>	SalesOrderDetail			adventure work...	AdventureWorks	SQL Server	Table	1/11
<input type="checkbox"/>	AdventureWorks			adventure work...		SQL Server	Database	1/11
<input type="checkbox"/>	SalesOrderHeader			adventure work...	AdventureWorks	SQL Server	Table	1/11
<input type="checkbox"/>	Customer			adventure work...	AdventureWorks	SQL Server	Table	1/11
<input type="checkbox"/>	Product			adventure work...	AdventureWorks	SQL Server	Table	1/11

**Figure 15-10.** Registered objects in grid view

Go ahead and click the List View option. This view gives you an interesting perspective into the registered assets. You get a verbal description and icons, and the information is easier to view, such as the Last Updated and Last Registered columns.

Along the top, you will also notice the ability to turn highlighting on and off. You will see this in action shortly but I highly recommend you leave highlighting on. To the right of the highlighting option is the ability to open the selected asset in a different program in order to view the data. Until you select an asset, this option, as well as the Delete button, are greyed out.

Along the right of the screen you will see the Details blade. I will discuss this blade in more detail in the next section but this panel has a few tabs that display information regarding the object selected in the list.

One question that is frequently asked is about changing data sources and the impact this has on the registered data asset. For example, if a database table is altered to remove or add a column, are those changes reflected in Azure Data Catalog? The short answer is no. To update the metadata, the data source needs to be reregistered, at which point the metadata will be updated in the catalog and any annotations will be maintained.

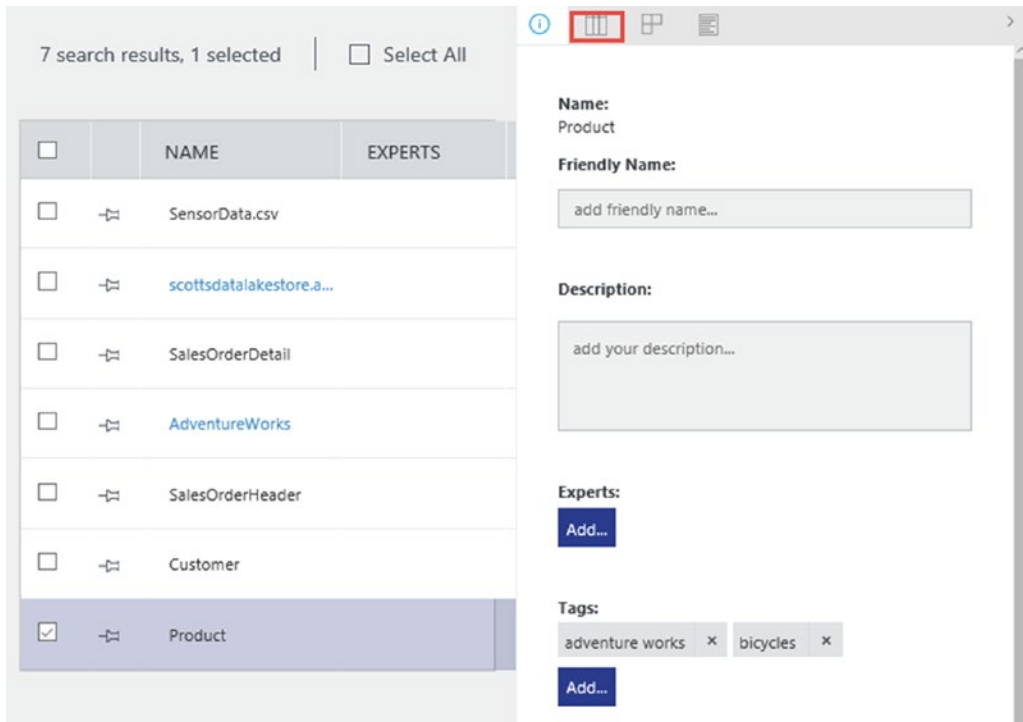
OK, time to move on. With the assets registered, it is time to go discover these assets.

## Discover Data Sources

Asset discovery is all about searching the catalog using search terms. The results will be the assets that have a match on any property in the asset. When the assets were registered, you applied tags to the assets and these tags are used to filter the search. You will see this momentarily but first let's explore the Details blade.

In the list of assets, click the Product asset. Automatically, the Details blade will slide out, showing the Properties tab. There are four tabs on the Details blade: Properties, Preview, Columns, and Documentation.

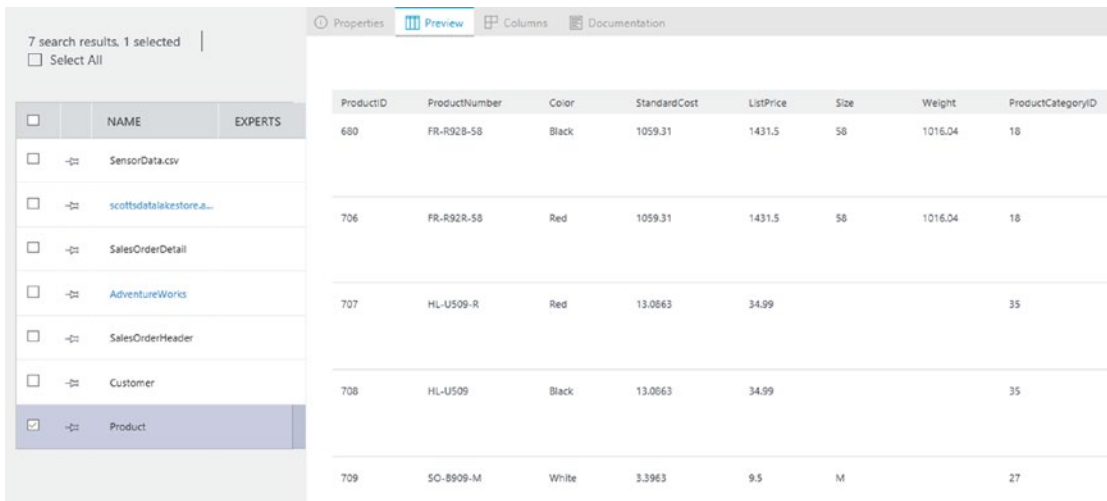
As shown in Figure 15-11, there is a lot you can do on the Properties tab, including adding a friendly name or a description of the asset, adding an expert (a subject matter expert for the selected asset, such as the owner of the data), and even adding additional tags. If you add a friendly name, this is the name that the user will see, so be careful how you use this field.



**Figure 15-11.** The Properties tab of the Details blade

As mentioned, this is where additional tags can be added. As users preview the data and discover the data source, they can add tags based on their needs for data discovery. Each user of Azure Data Catalog can add multiple tags for each asset, but only the user who created the tag can edit their own tags. Admins and asset owners can delete tags but not edit them.

Clicking the tab shows a preview of the data in the selected asset. As shown in Figure 15-12, the Preview tab shows a preview of the data in the Products table.



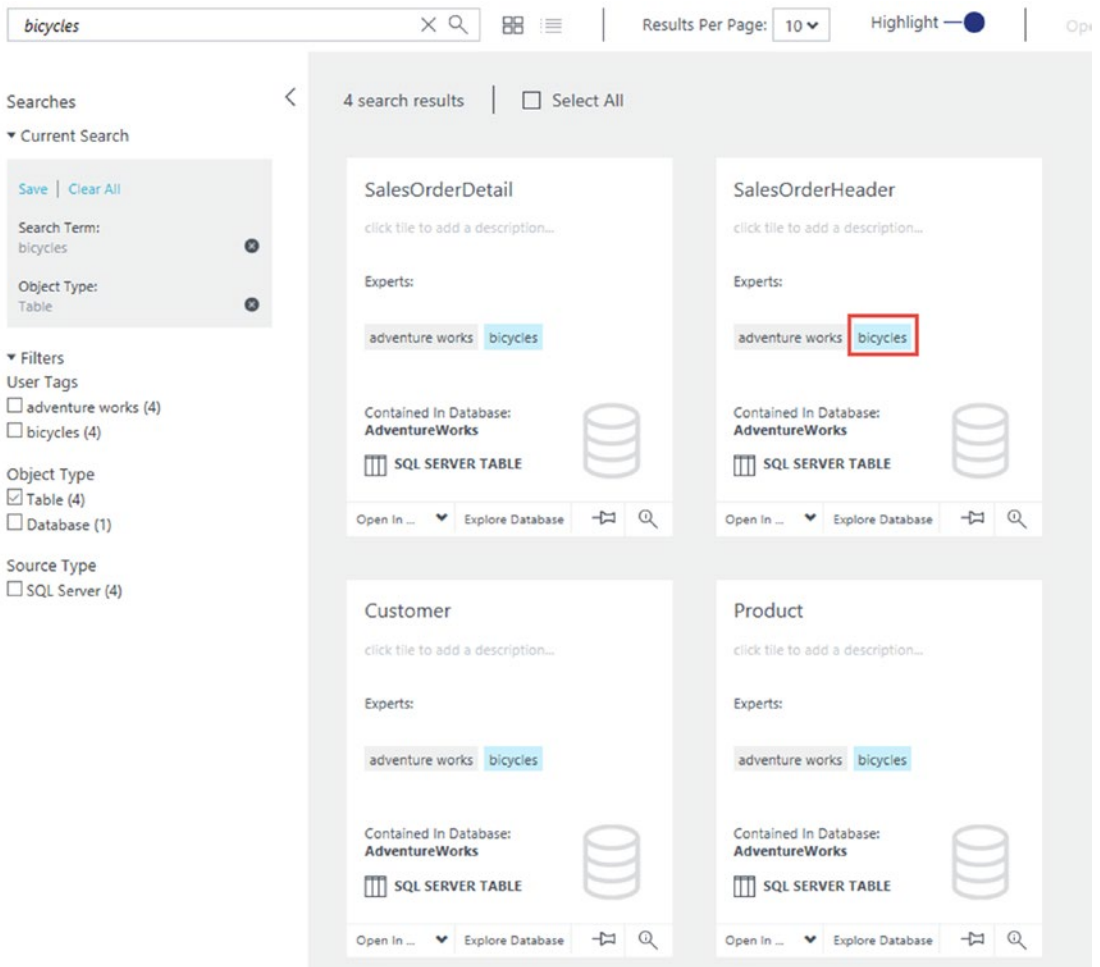
**Figure 15-12.** *Previewing data*

Going back for a moment, the preview of the data is available due to the fact that during the asset registration process you checked the Include Preview checkbox (see Figure 15-7). Remember also that checking that checkbox will copy up to 20 records from each table selected into the data catalog. Thus, it is recommended that you check this checkbox.

The Columns tab shows details about the columns in the table such as name and data type. The Documentation tab allows you to add a description and overall documentation regarding this selected asset.

## Data Discovery

Let’s talk about data discovery, which is accomplished via searches. In the search box, which is in the upper left, type in *bicycles*, as shown in Figure 15-13. Press Enter.



**Figure 15-13.** Discovering data via search

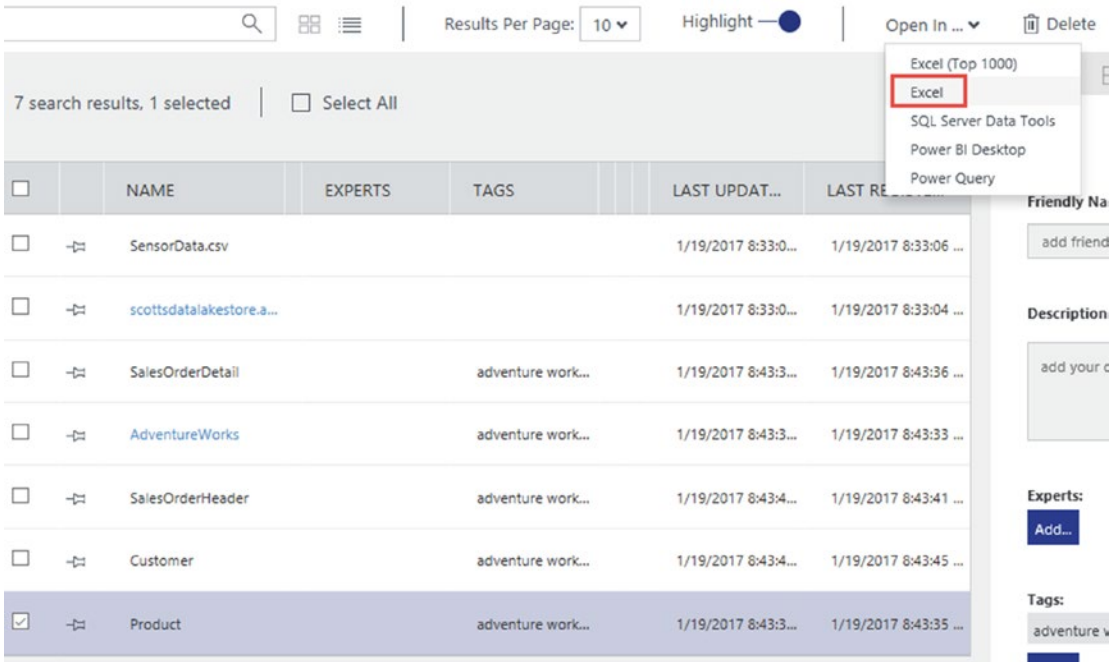
You will notice a couple of things, both of which are shown in Figure 15-13. First, because highlighting is turned on, the search keyword is highlighted in the search results. Second, only the assets that match the search keyword are returned. In this example, the four tables from the AdventureWorks database are returned and the keyword is highlighted.

This was a very simple search example, but there are other ways you can search. For example, you can discover assets via property scoping in the search. For example, you can type `tags:bicycles`. You can also do a Boolean search, such as `tags:bicycles AND objecttype:table`.

What is cool is that any and all executed searches can be saved. As you can see in Figure 15-13, in the Searches blade you can save a search by clicking the Save link. Again, very cool, especially when you have a more complicated search such as a query that has logical isolation via parentheses, such as `name:Customer AND (tags:bicycles AND objecttype:table)`.

## Connect to Data Sources

OK, you have registered data sources and discovered them. This section will show how to connect to them through existing tools, such as Excel. So, with the Product asset selected, the Open In option will now be enabled. Click the drop-down and select Excel, as shown in Figure 15-14.

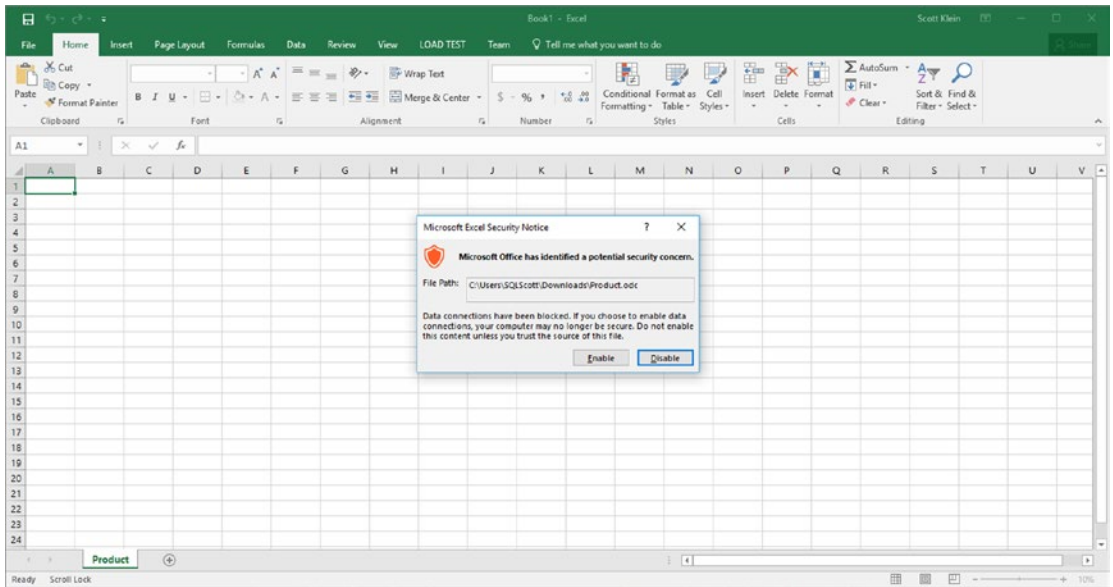


**Figure 15-14.** Connecting to a Dataset via Excel

Notice all of the different applications you can connect to that are integrated into Azure Data Catalog. It's a very nice list, including SQL Server Data Tools and Power BI.

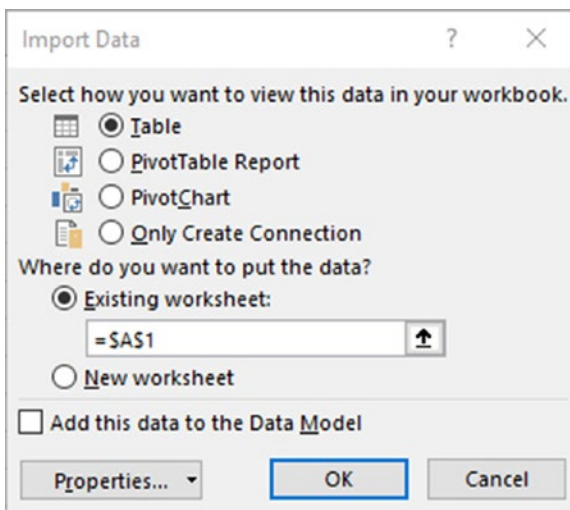
OK, back to the example. Clicking on the Excel option will download a file called Product.odc. An .odc file is an Office Data Connection file and it provides the ability to connect to a data source via, in this case, Excel. Depending on the browser, you can either open it or save it (see Figure 15-15.) Open the file, which will open up Excel with a nice security notice. Click Enable on the notice.





**Figure 15-15.** Opening Excel

Your next dialog will be the Import Data dialog, shown in Figure 15-16. Leave everything as is and click OK.



**Figure 15-16.** Importing the data

Since the data is coming from Azure SQL Database, you will be prompted to enter credentials to authenticate. Once authenticated, Excel will pull the data in and display it, as shown in Figure 15-17.

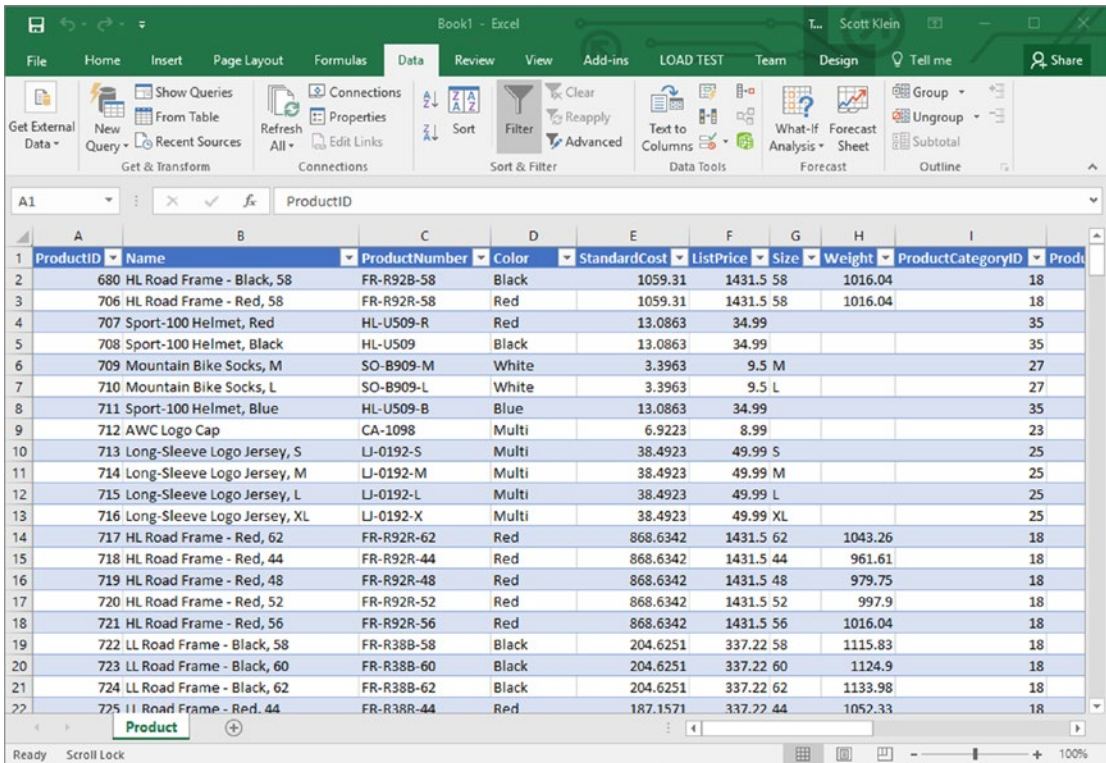


Figure 15-17. Viewing the data in Excel

**Note** Checking with engineering on one item about opening excel first without going through the ADC portal. Will update during AR.

Time for a quick review. Think about where and how Azure Data Catalog fits into the overall picture. Data is all over your organization, and in respect to this book, it could be coming from myriad different data sources including devices and sensors. How do your users discover this data easily? Azure Data Catalog makes it easy to register, discover, and consume that data.

Currently, it is not possible to add data source types, but in the future Azure Data Catalog will allow third parties to add new data source types through an extensibility API.

## Summary

In this chapter, you learned about discovering data. Data is all over within your organization and it should be easily discoverable and understandable. You began by creating the Azure Data Catalog and then you registered different data assets from different data sources. You then discovered those data sources through searches and the different ways to search. Lastly, you looked at how to connect to the data sources and use different applications to consume the data.