Microsoft Purview enables you to catalog data assets across your data estate and track the flow of data as it is transferred from one data source to another - a key element of a comprehensive data governance solution.

This exercise should take approximately ****40**** minutes to complete.

## Before you start

You’ll need an [Azure subscription](https://azure.microsoft.com/free) in which you have administrative-level privileges and exclusive access to the tenant in which the subscription is defined.

## Provision Azure resources

1. In the PowerShell pane, enter the following commands to clone this repo:

rm -r dp-203 -f

git clone https://github.com/MicrosoftLearning/dp-203-azure-data-engineer dp-203

1. After the repo has been cloned, enter the following commands to change to the folder for this lab and run the ****setup.ps1**** script it contains:

cd dp-203/Allfiles/labs/22

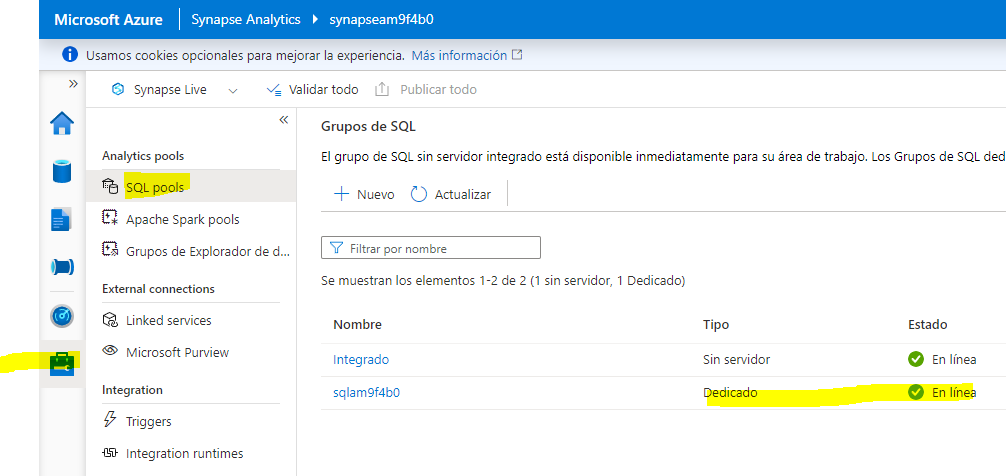
./setup.ps1

## Create lake database

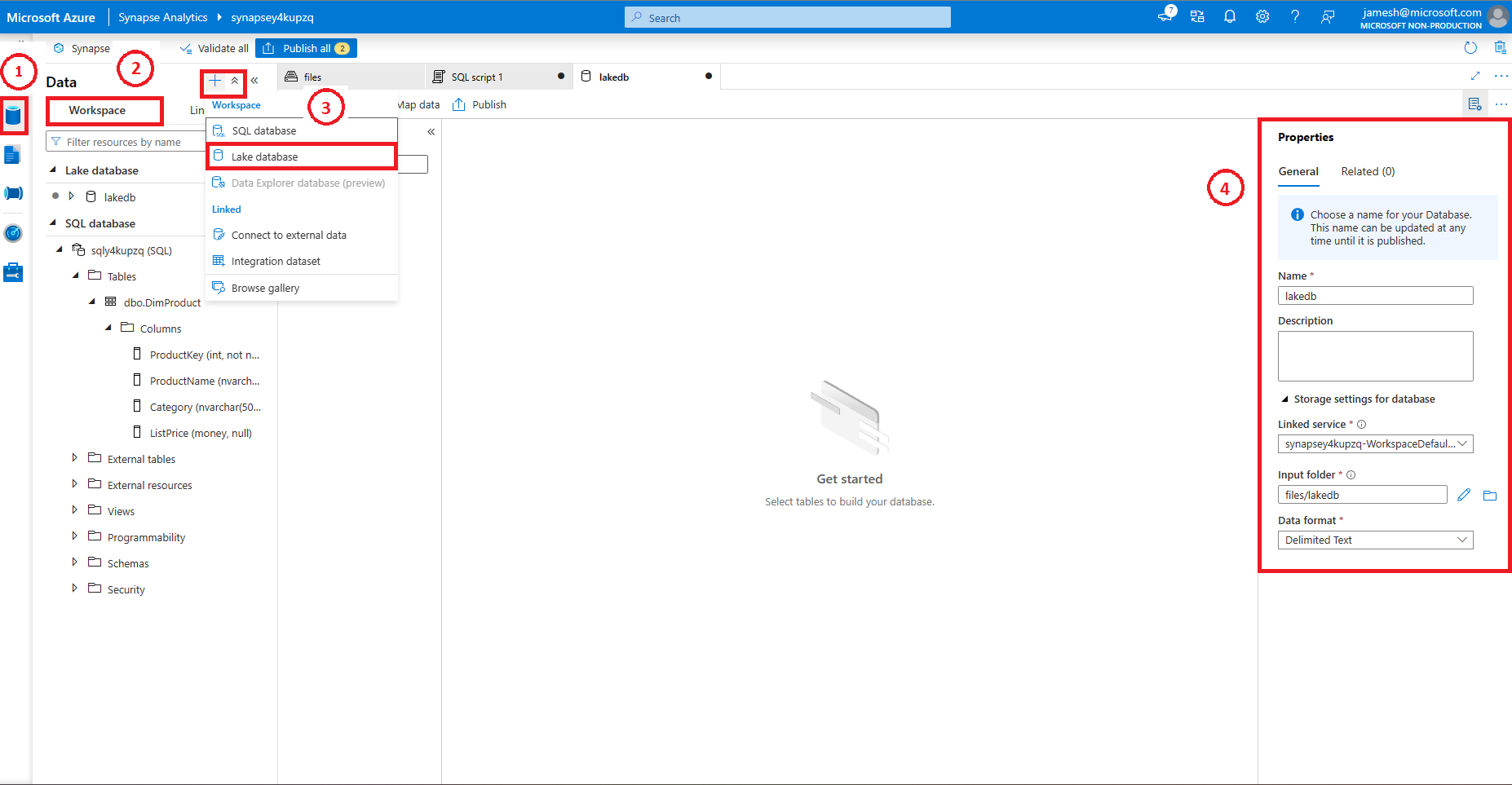
Lake databases store data in a data lake on Azure Storage. You can use Parquet, Delta or CSV formats and different settings to optimize storage. Each lake database has a linked service to define the root data folder.

Lake databases are accessible in Synapse SQL serverless SQL pool and Apache Spark, letting users separate storage from compute. The metadata of the lake database makes it easy for different engines to provide an integrated experience and use extra information (like relationships) that wasn’t supported on the data lake.

1. To create a lake database let’s first open the synapsexxxxxxx from the proper resource group and then click on the **open** link on the ****Open Synapse Studio****.
2. Next, we’ll click on the toolbox with the wrench in it, which is also the Manage section of the Synapse Analytics workspace and make sure our Dedicated pool is running. This could take a few minutes to start.



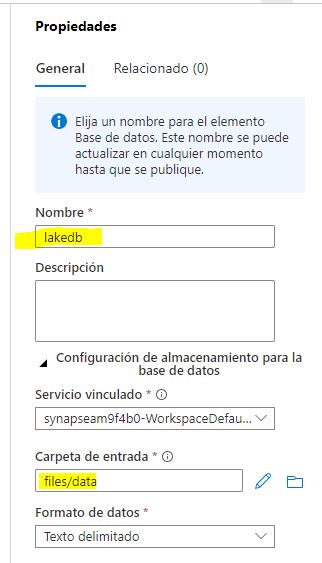
1. From here, we’ll click on the database symbol which looks like a barrel and has a label of data.
2. Once in the Data panel, click on the + symbol to the right of the word ****Data**** and select **Lake database**

[](https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/images/lakedb-configure.png)

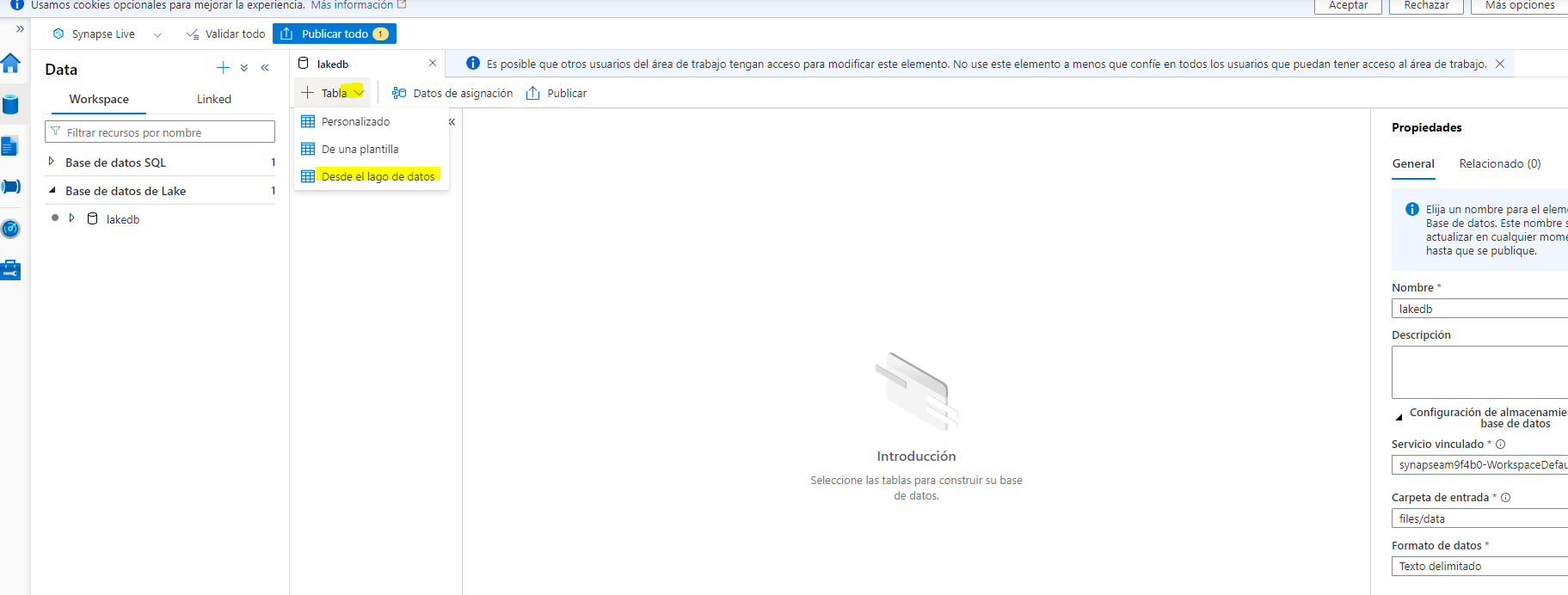
****Note****: You will receive a prompt ****Azure Synapse Database Template Terms of Use**** which you should read and understand prior to clicking the ****OK**** button.

1. You’ll note on the far right-hand side that there is a Properties window.
   1. Type ****lakedb**** in the name field.
   2. under ****Input folder**** select the folder and browse to root/files/data, then press ****OK****.

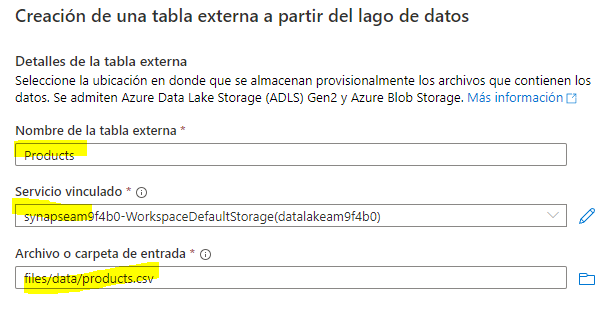
****Note****: You will likely see an error when opening the ****Input folder****, just double-click on the root folder and work your way down to data before clicking ****OK**** if that’s the case.



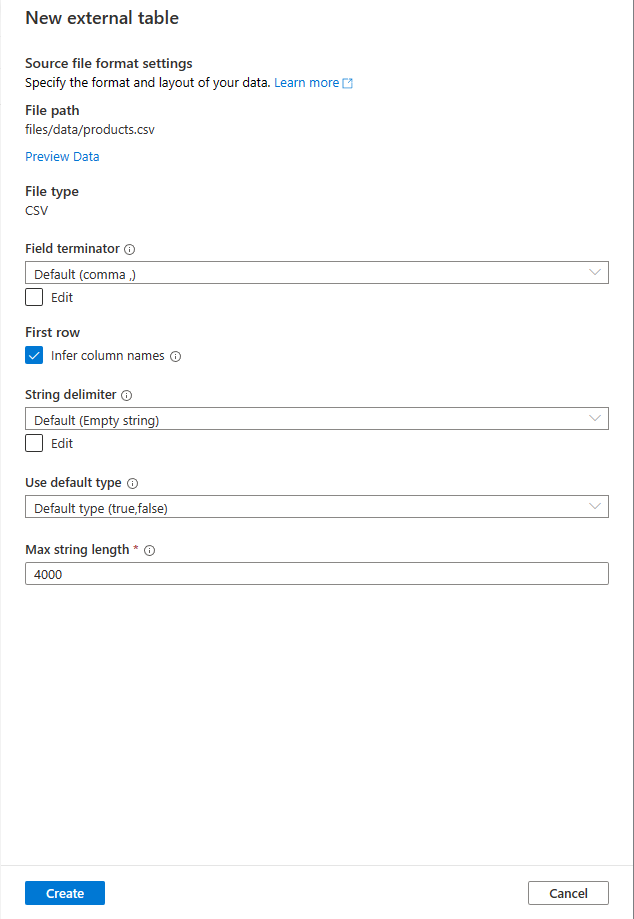
1. To the left of this screen you’ll see a pillar with the name of your ****lakedb**** with a ****+Table**** below it, click on this, and select **From data lake**



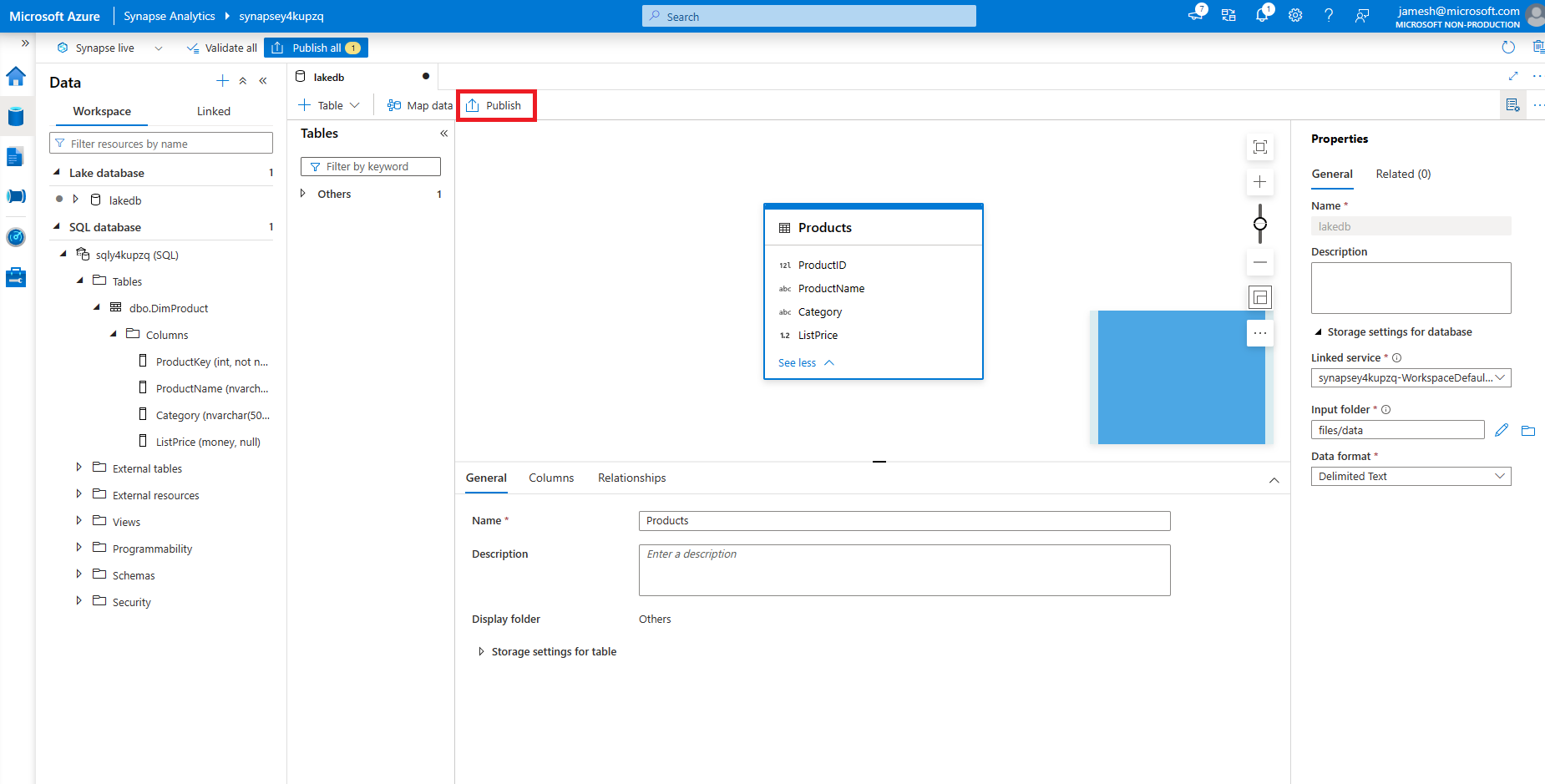
1. Under the **External table name** type: ****Products****.
2. Under **Linked service**, select the default option.
3. Within the **Input file or folder** click on the file folder on the far right and browse to ****root > files > data >**** and select the **products.csv** and click ****OK****, then click ****Continue****.



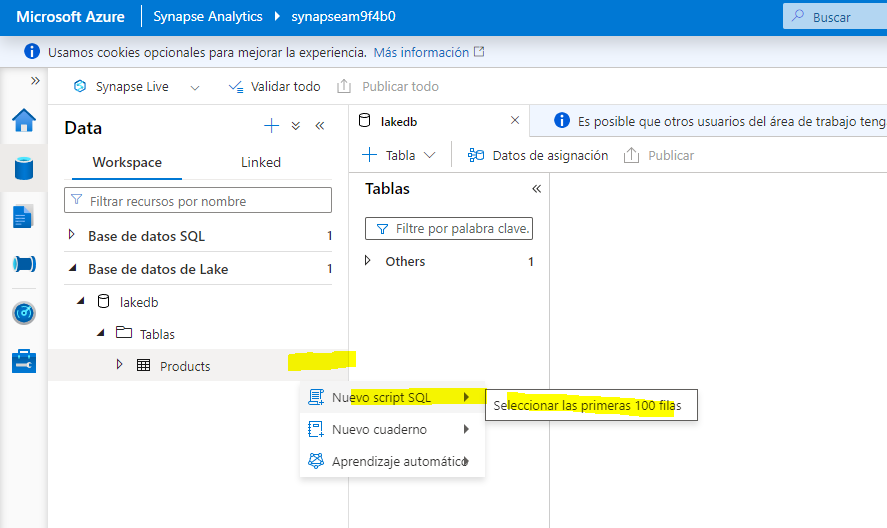
1. On the ****New external table**** pane, select the First Row option to **infer column names** then click ****create****.

[](https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/images/lakedb-external-table-source.png)

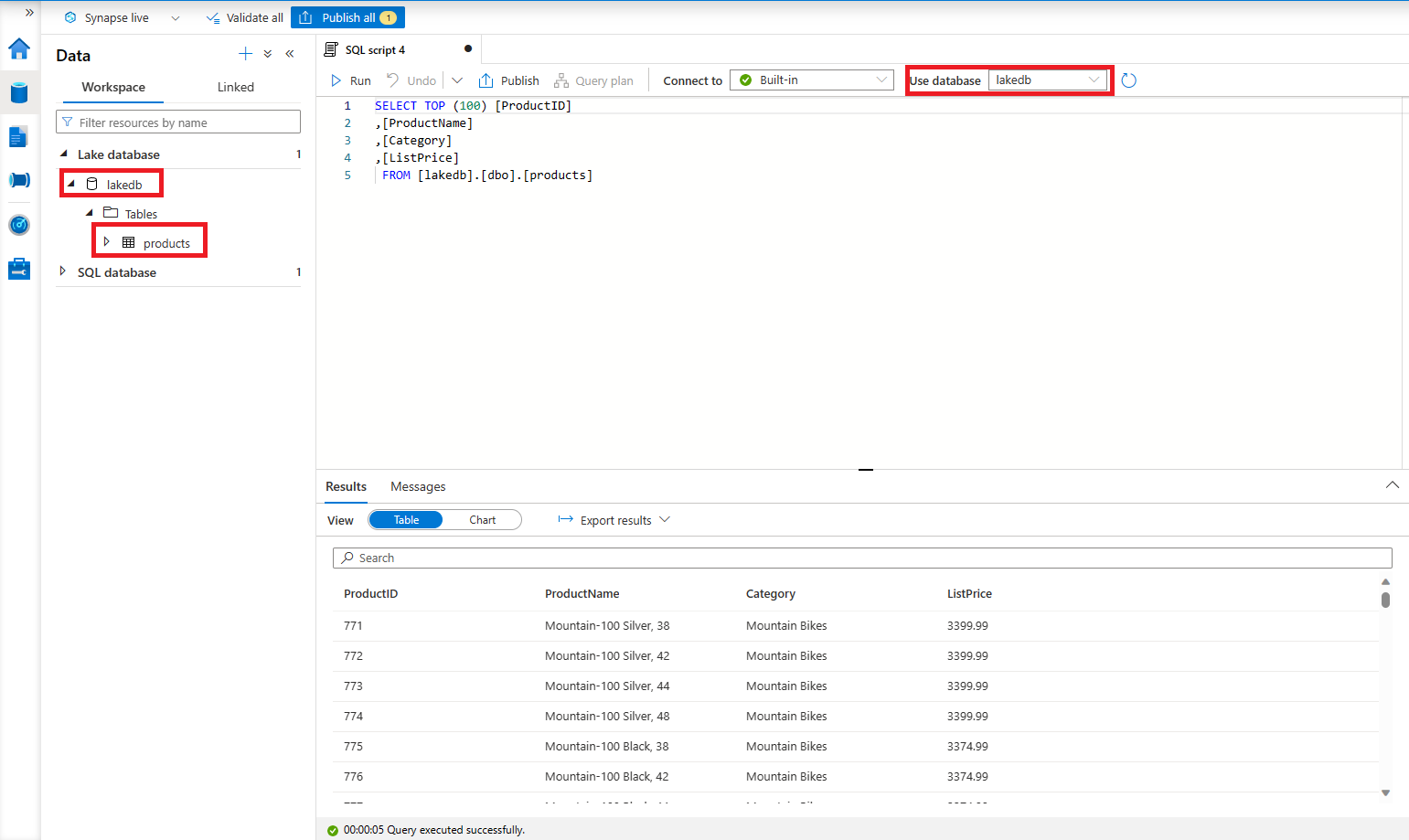
1. Press the ****Publish**** on the top portion of the query design window.

[](https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/images/lakedb-external-table-definition.png)

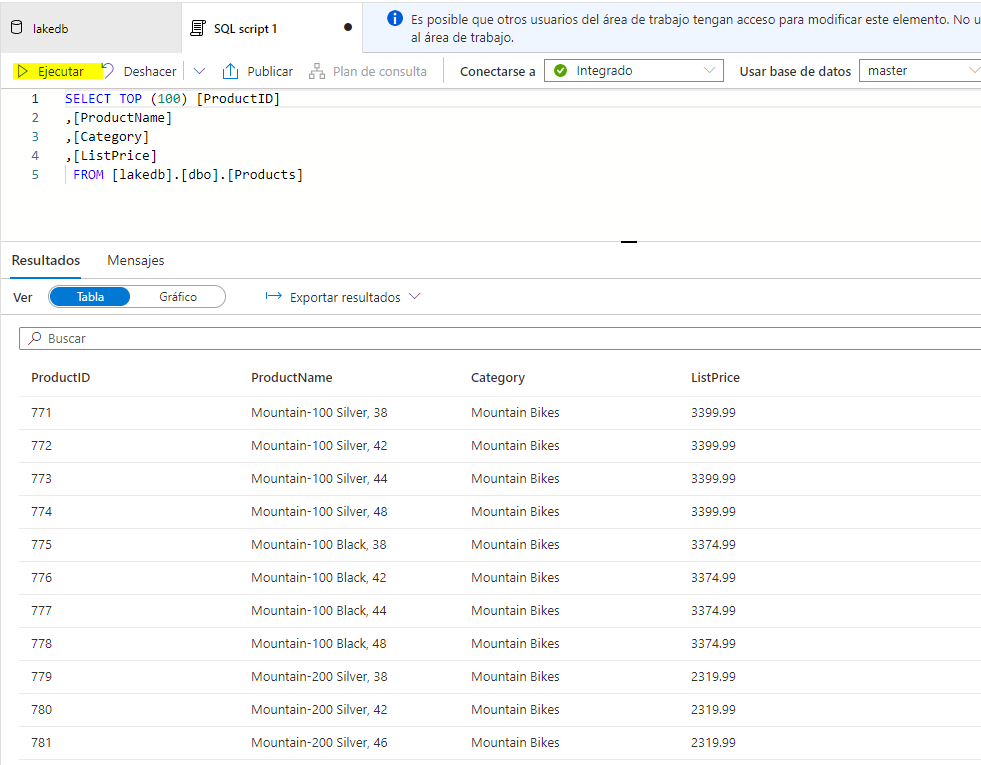
1. Under ****Data**** make sure you’re in the ****Workspace**** area on the left hand side and expand the ****Lake database**** section, the expand ****lakedb****, then **mouse-over** the right side of the ****Products**** table and choose the **Top 100 rows**.



****Note****: You want to make sure that the ****Connect to**** is listed as ****Built-in**** and you can leave the ****master**** database selected or click the refresh button to the right and select ****lakedb**** database. As it uses a 3 part naming convention [database].[schema].[table] both will work.

[](https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/images/lakedb-first-external-query.png)

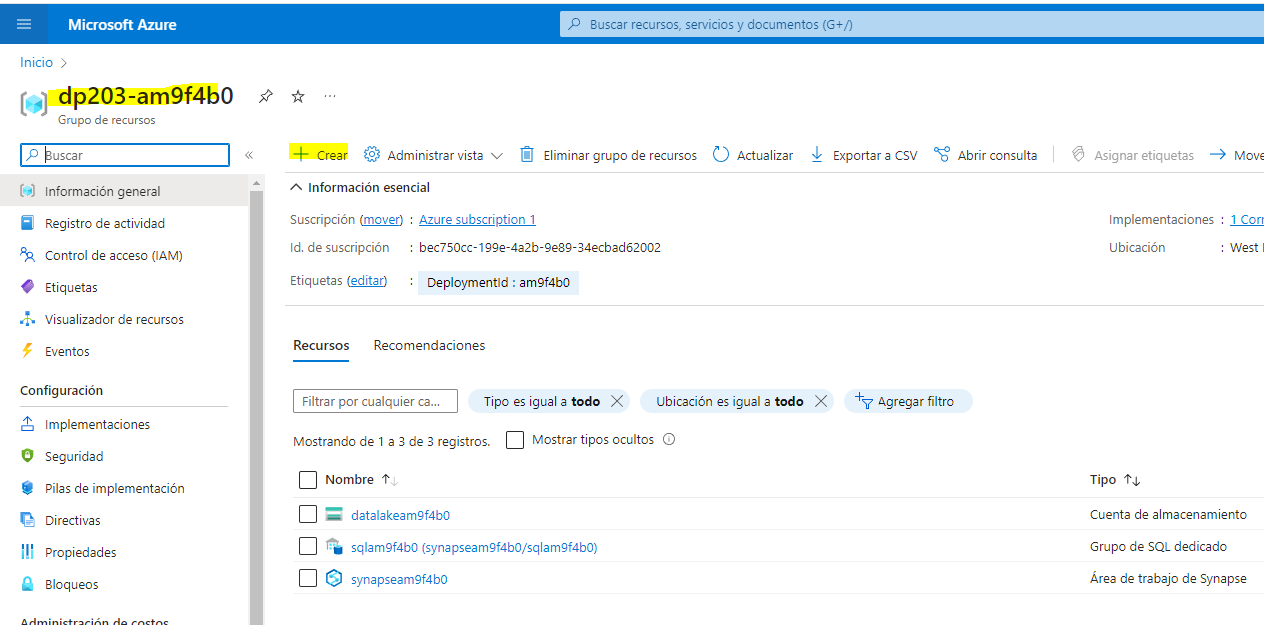
1. Press the ****Run**** button to view the data within the lake database table.



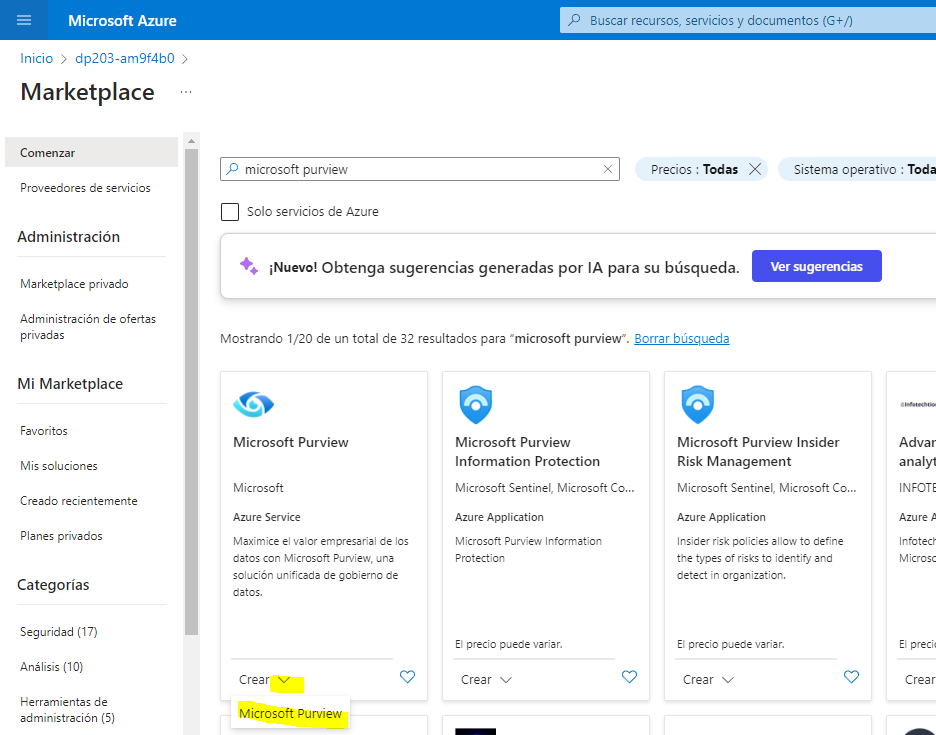
## Add Microsoft Purview Service to Account

Microsoft Purview is a comprehensive portfolio of products spanning data governance, information protection, risk management, and compliance solutions. It helps you govern, protect, and manage your entire data estate across your on-premises, multi-cloud, and software as a service (SaaS) data

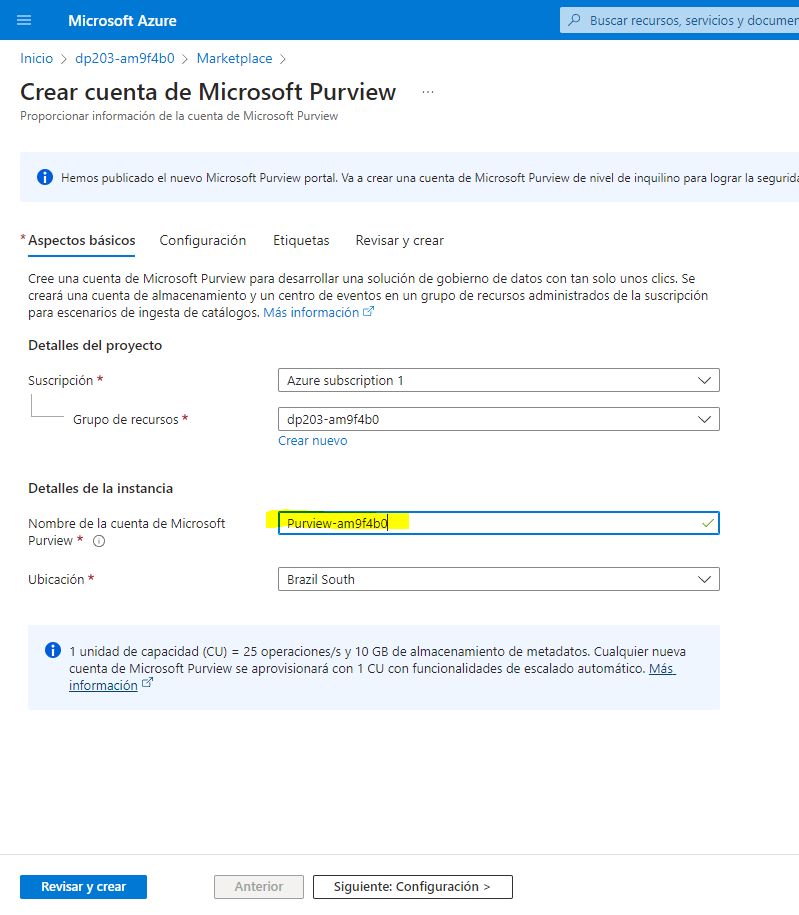
To set it up, we will first return to our main resource group which will be named dp203-xxxxxxx based upon the random number you were assigned. Once you are within the ****resource group**** click on the **+ Create** button to add a new service.



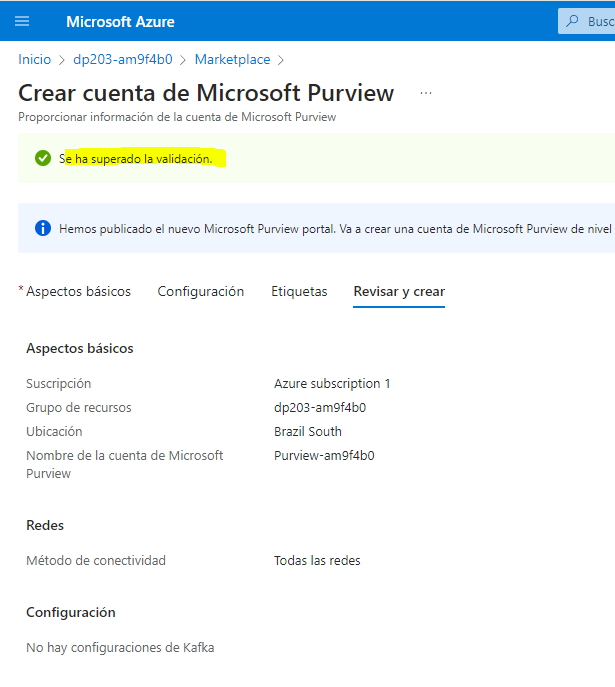
1. Select the ****Microsoft Purview service**** and then click on the ****Create**** button.



1. During the creation process, since you started in the appropriate resource group, it should already be selected. Next, we’ll give ****Purview**** a name using your randomly assigned number. Next, select the best region for your instance.



1. Click on the ****Review & Create**** button and wait for **Validation** prior to moving forward.



1. Once the validation has passed, press the ****Create**** button.

****Note****: You may have to try a few regions in order to be able to pass validation with Purview.

## 

## Catalog Azure Synapse Analytics data assets in Microsoft Purview

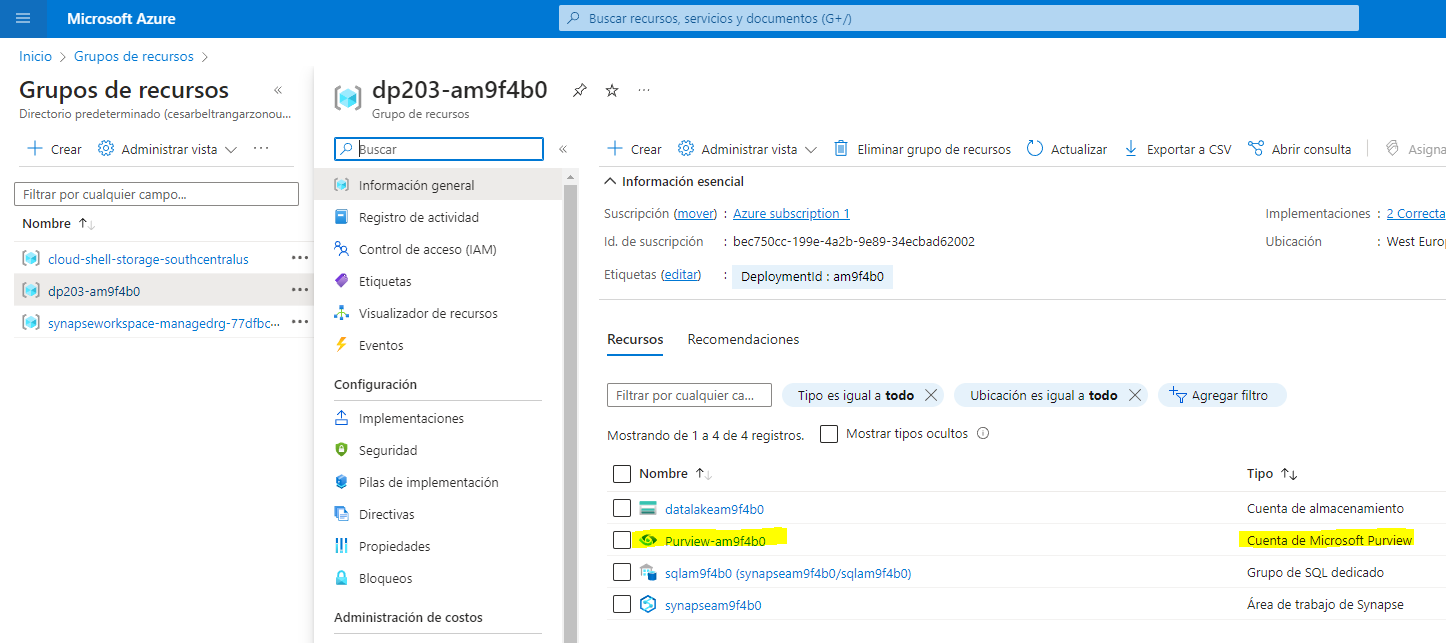
With Microsoft Purview, you can catalog data assets across your data estate - including data sources in an Azure Synapse Workspace. The workspace you just deployed includes a data lake (in an Azure Data Lake Storage Gen2 account), a serverless database, and a data warehouse in a dedicated SQL pool.

### Configure role-based access for Microsoft Purview

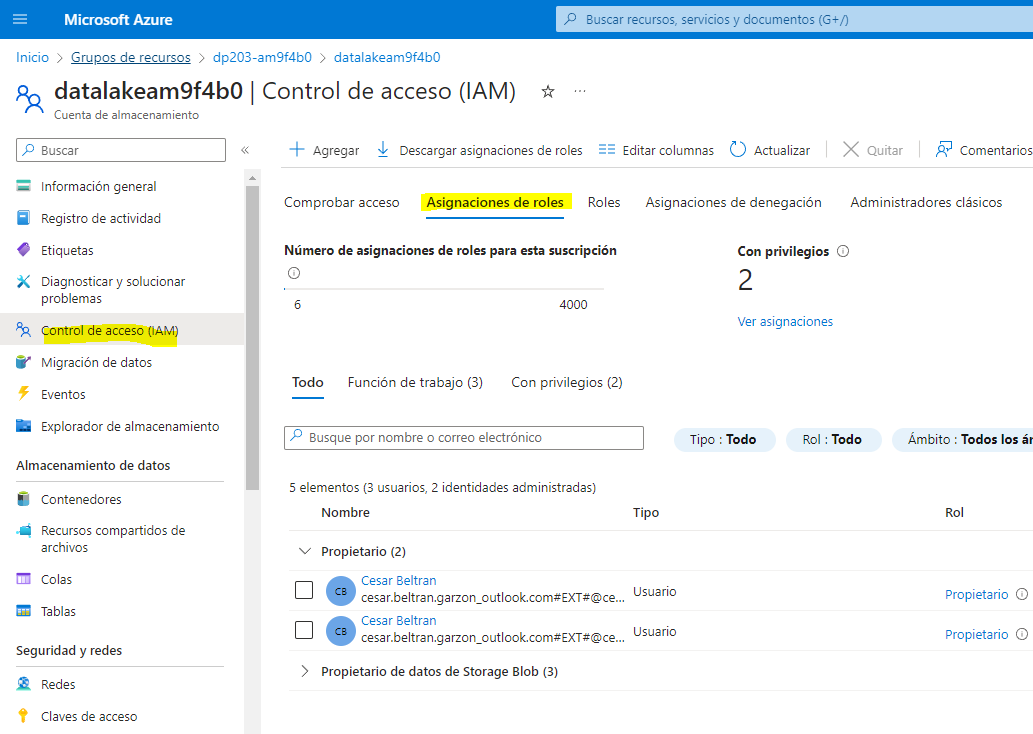
Microsoft Purview is configured to use a managed identity. In order to catalog data assets, this managed identity account must have access to the Azure Synapse Analytics workspace and the storage account for its data lake store.

1. In the [Azure portal](https://portal.azure.com/), browse to the ****dp203-**xxxxxxx** resource group that was created by the setup script and view the resources that it created. These include:

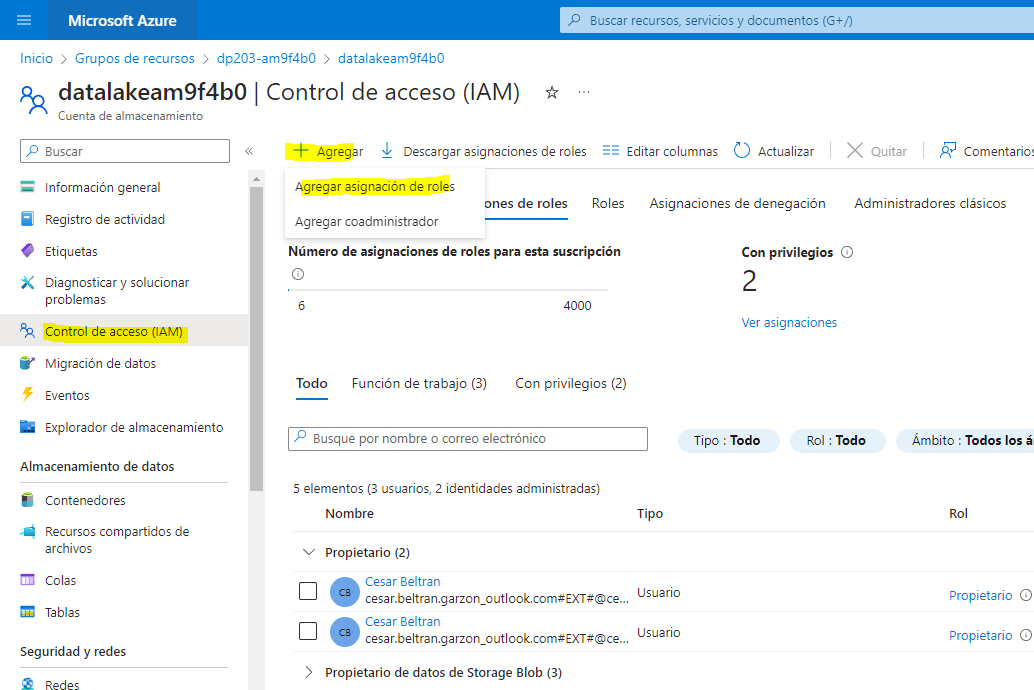
* A storage account with a name similar to ****datalake**xxxxxxx**.
* A Microsoft Purview account with a name similar to ****purview**xxxxxxx**.
* A dedicated SQL pool with a name similar to ****sql**xxxxxxx**.
* A Synapse workspace with a name similar to ****synapse**xxxxxxx**.



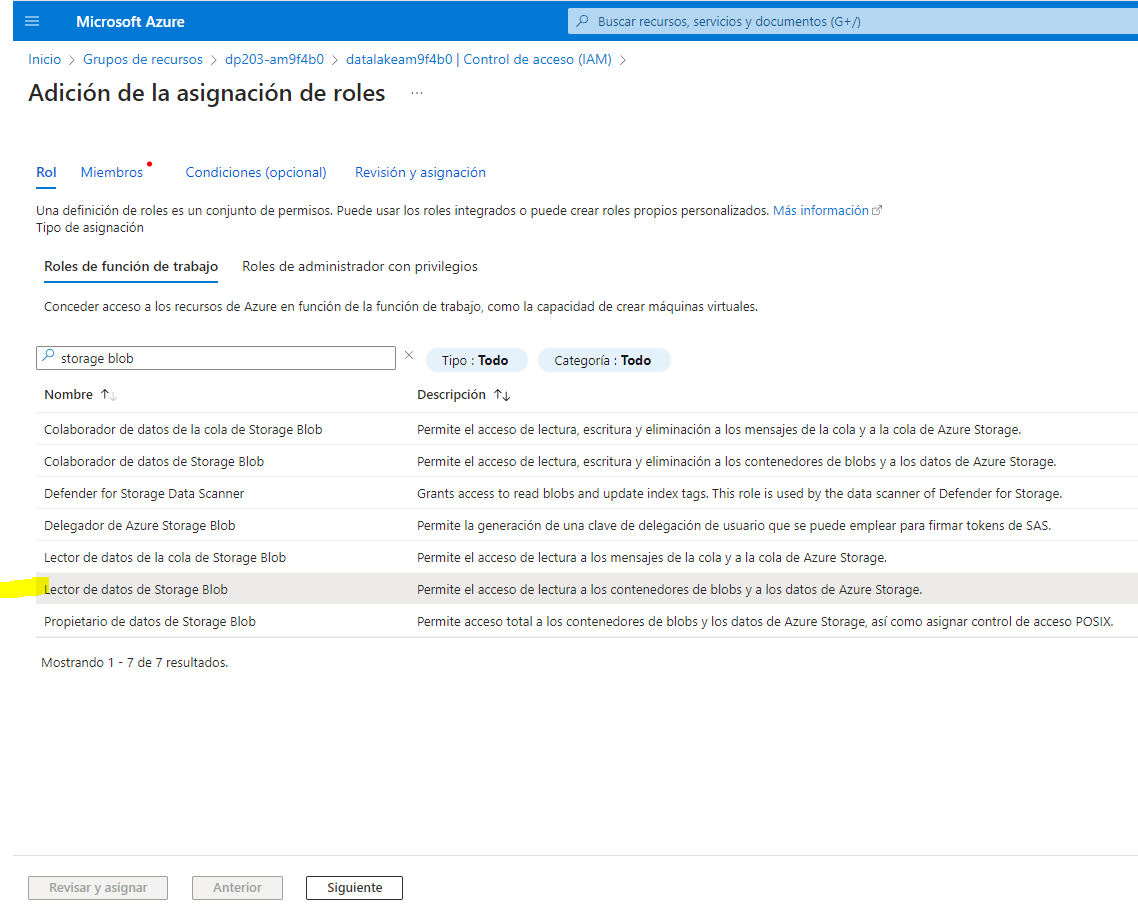
1. Open the ****datalake**xxxxxxx** storage account, and on its ****Access Control (IAM)**** page, view the ****Role Assignments**** tab; as shown here:



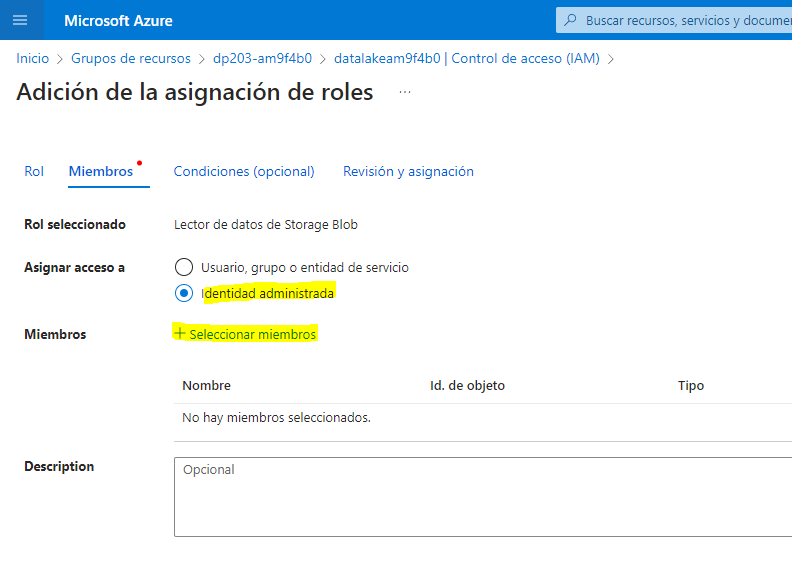
1. On the ****+ Add**** button menu, select ****Add role assignment****.



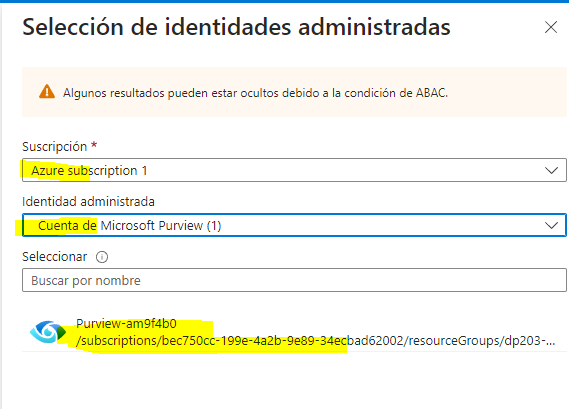
1. In the ****Add role assigmnent**** page, on the ****Role**** tab, search for “storage blob” and select the ****Storage Blob Data Reader**** role. Then select ****Next**** to move to the ****Members**** tab:



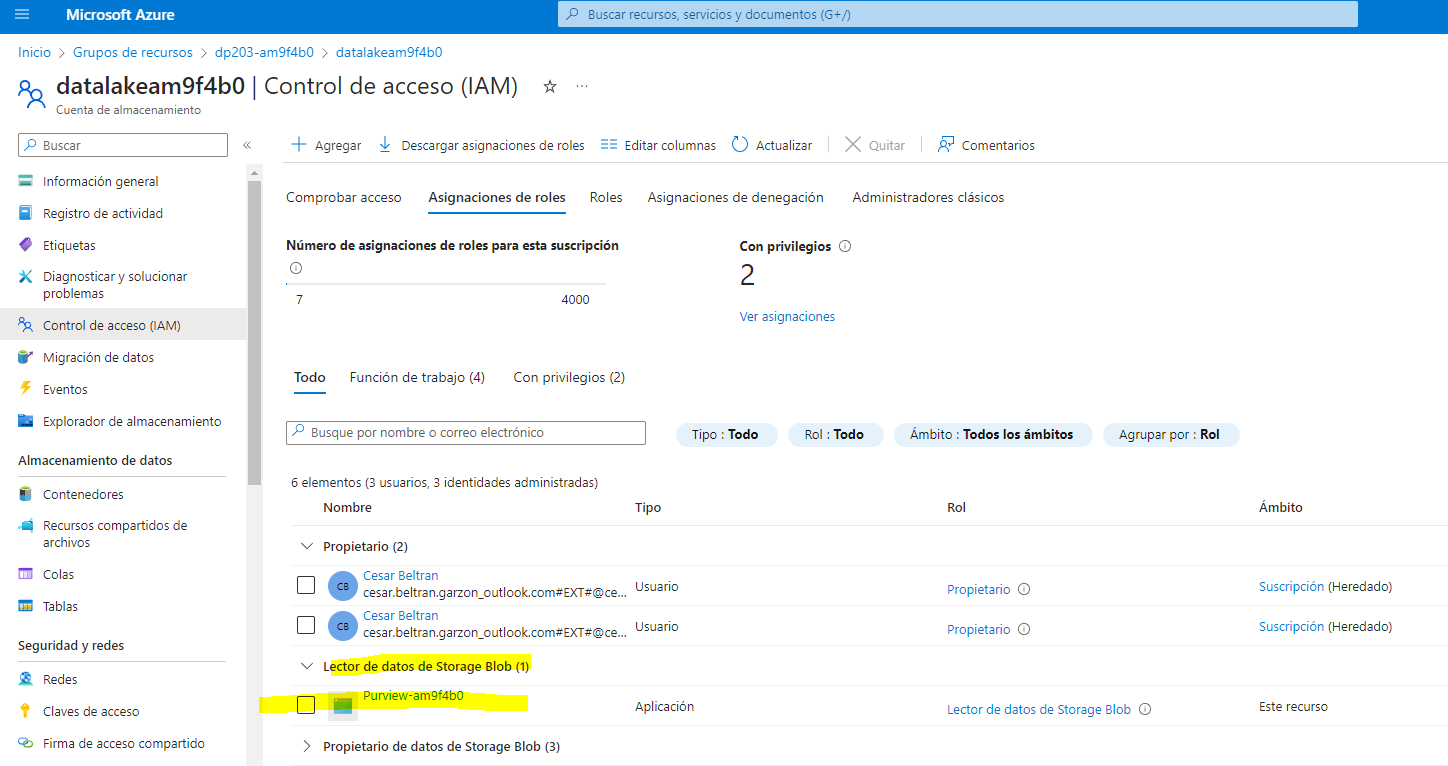
1. On the ****Members**** tab, in the ****Assign access to**** list, select ****Managed identity****; and then, under ****Members****, select ****Select members****.



1. In the ****Select managed identities**** pane, in the ****Managed identity**** list, select ****Microsoft Purview account (**n**)**** and select your Microsoft Purview account, which should have a name similar to ****purview**xxxxxxx**; and use the ****Select**** button to add this account to the role assignment:



1. Use the ****Review + Assign**** button to complete the role assignment, which makes the ****purview**xxxxxxx** account used by the managed identity for your Microsoft Purview resource a member of the ****Storage Blob Data Reader**** role for your storage account.



1. In the Azure portal, return to the ****dp203-**xxxxxxx** resource group and open the ****synapse**xxxxxxx** Synapse Analytics workspace. Then, on its ****Access Control (IAM)**** page, add a role assignment to make the ****purview**xxxxxxx** managed identity account a member of the ****Reader**** role in the workspace.

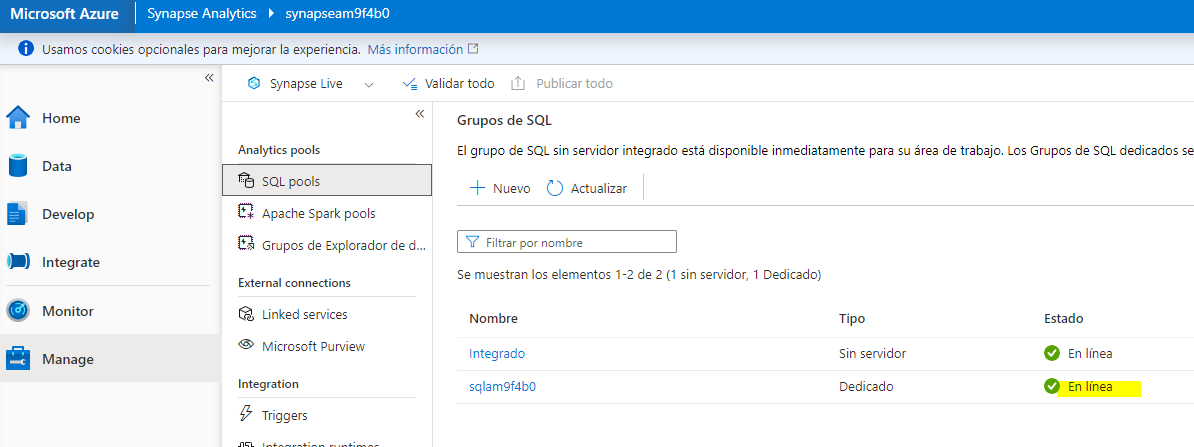
### Configure database permissions for Microsoft Purview

Your Azure Synapse Analytics workspace includes databases in both serverless and dedicated SQL pools, to which the managed identity used by Microsoft Purview requires access.

1. In the Azure portal, on the page for your Synapse Analytics workspace, view the ****Overview**** tab. Then in the ****Open Synapse Studio**** tile, use the link to open Azure Synapse Studio in a new browser tab - signing in if prompted.

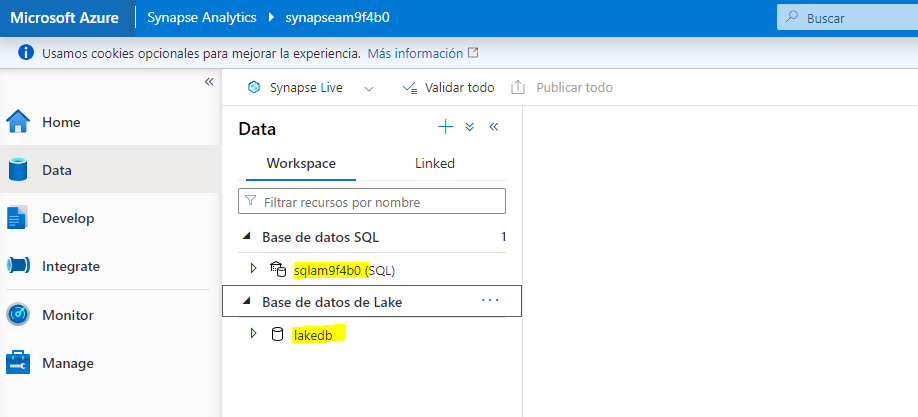
****Tip****: Alternatively, you can open Azure Synapse Studio by browsing directly to https://web.azuresynapse.net in a new browser tab.

1. On the left side of Synapse Studio, use the ****››**** icon to expand the menu - this reveals the different pages within Synapse Studio.
2. On the ****Manage**** page, on the ****SQL pools**** tab, select the row for the ****sql**xxxxxxx** dedicated SQL pool and use its ****▷**** icon to start it; confirming that you want to resume it when prompted.
3. Wait for the SQL pool to resume. This can take a few minutes. You can use the ****↻ Refresh**** button to check its status periodically. The status will show as ****Online**** when it is ready.

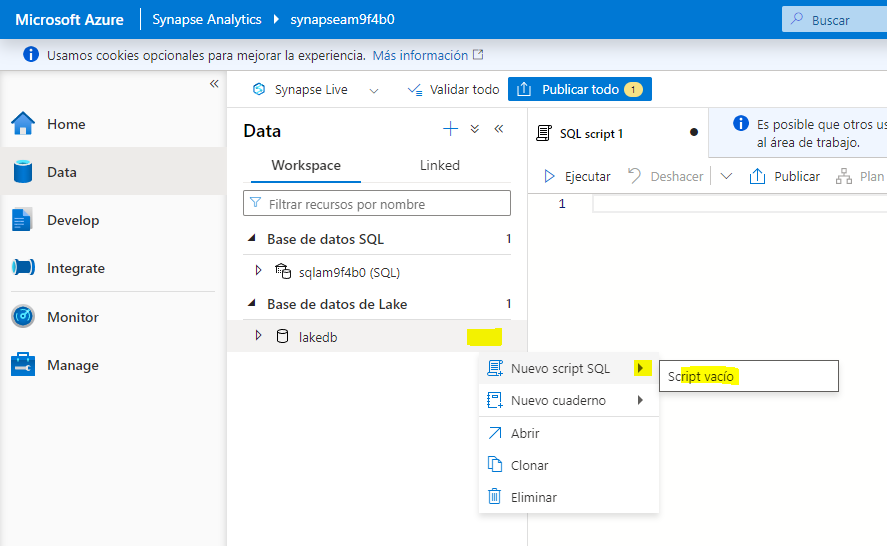


1. In Azure Synapse Studio, view the ****Data**** page, and in the ****Workspace**** tab, expand ****SQL database**** to see the databases in your workspace. These should include:

* A serverless SQL pool database named ****lakedb****.
* A dedicated SQL pool database named ****sql**xxxxxxx**.



1. Select the ****lakedb**** database, and then in its ****…**** menu, select ****New SQL script**** > ****Empty script**** to open a new ****SQL script 1**** pane. You can use the ****Properties**** button (which looks similar to ****🗏\*****) on the right end of the toolbar to hide the ****Properties**** pane and see the script pane more easily.



1. In the ****SQL script 1**** pane, enter the following SQL code, replacing all instances of **purviewxxxxxxx** with the managed identity name for your Microsoft Purview account:

CREATE LOGIN purviewxxxxxxx FROM EXTERNAL PROVIDER;

GO

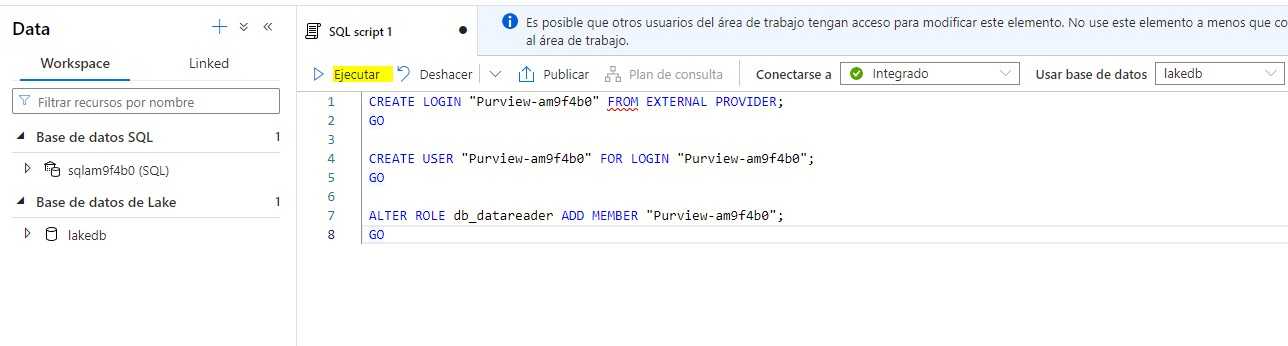
CREATE USER purviewxxxxxxx FOR LOGIN purviewxxxxxxx;

GO

ALTER ROLE db\_datareader ADD MEMBER purviewxxxxxxx;

GO

1. Use the ****▷ Run**** button to run the script, which creates a login in the serverless pool and a user in the ****lakedb**** user for the managed identity used by Microsoft Purview, and adds the user to the ****db\_datareader**** role in the ****lakedb**** database.



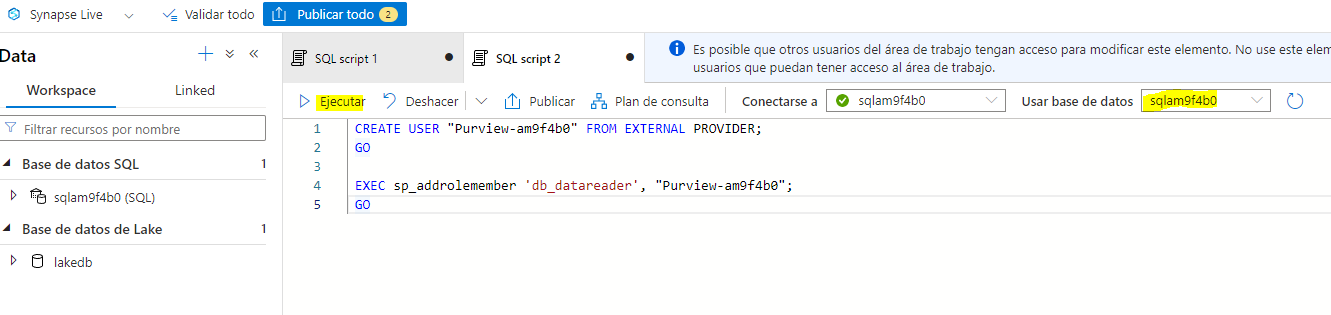
1. Create a new empty script for the ****sql**xxxxxxx** decicated SQL pool database, and use it to run the following SQL code (replacing **purviewxxxxxxx** with the managed identity name for your Microsoft Purview account); which creates a user in the dedicated SQL pool for the managed identity used by Microsoft Purview, and adds it to the ****db\_datareader**** role in the ****sql**xxxxxxx** database.

CREATE USER purviewxxxxxxx FROM EXTERNAL PROVIDER;

GO

EXEC sp\_addrolemember 'db\_datareader', purviewxxxxxxx;

GO

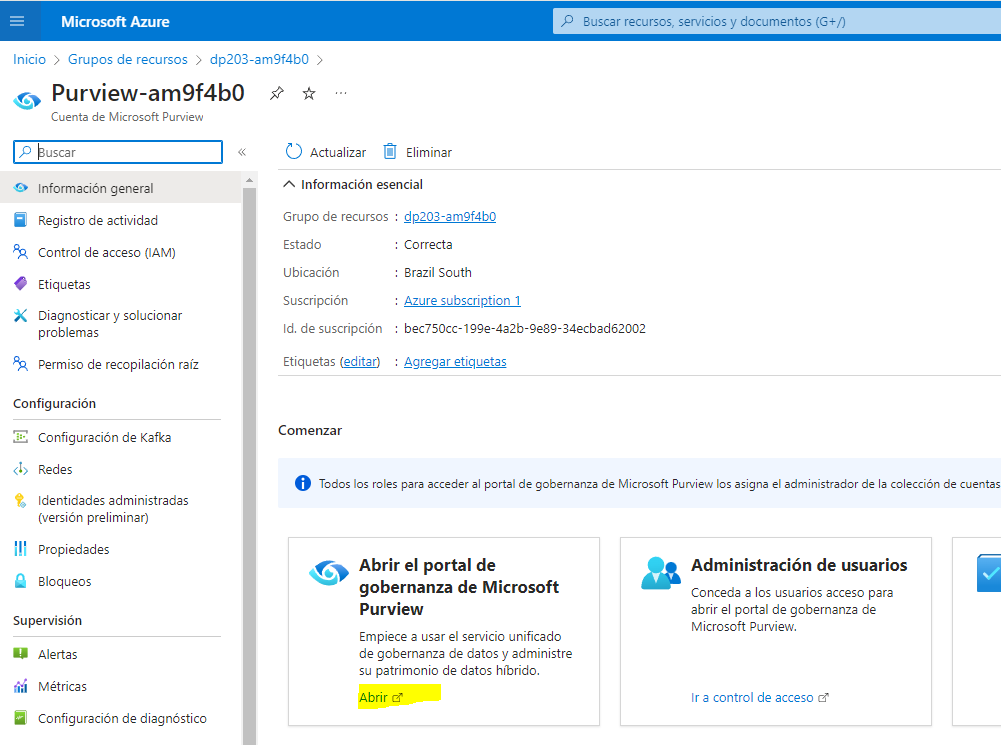


### Register sources in the Microsoft Purview catalog

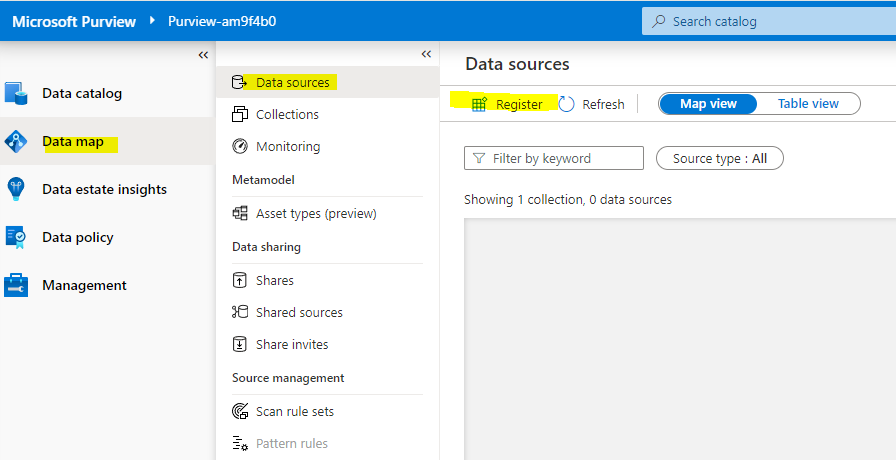
Now that you’ve configured the required access for Microsoft Purview to scan the data sources used by your Azure Synapse Analytics workspace, you can register them in your Microsoft Purview catalog.

1. Switch back to the browser tab containing the Azure portal, and view the page for the ****dp203-**xxxxxxx** resource group.
2. Open the ****purview**xxxxxxx** Microsoft Purview account, and on its ****Overview**** page, use the link to open the ****Microsoft Purview Governance Portal**** in a new browser tab - signing in if prompted.

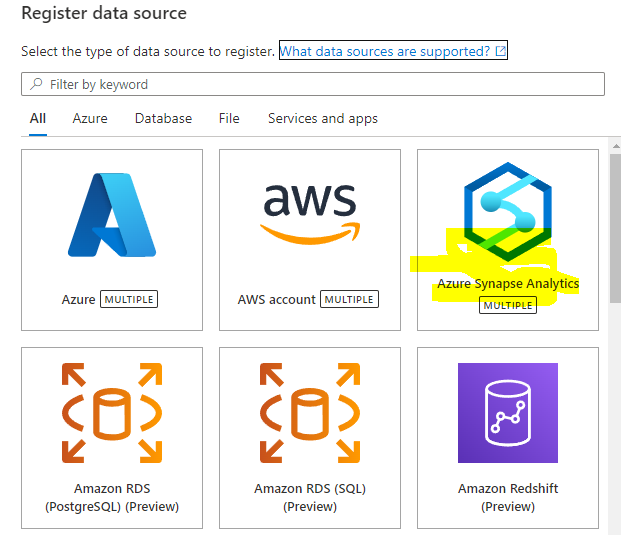
****Tip****: Alternatively, you can browse directly to https://web.purview.azure.com in a new browser tab.



1. On the left side of the Azure Purview Governance Portal, use the ****››**** icon to expand the menu - this reveals the different pages within the portal.
2. On the ****Data map**** page, on the ****Sources**** sub-page, select ****Register****:

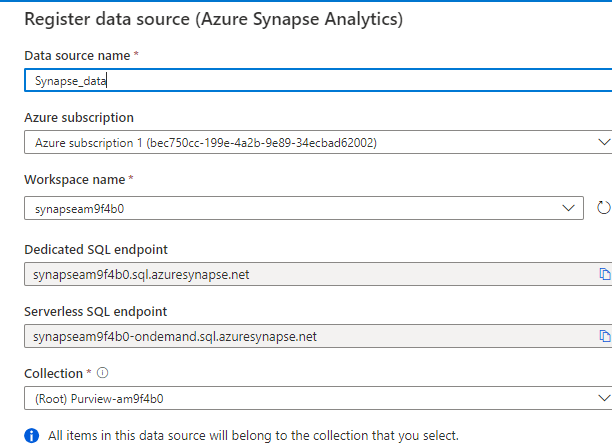


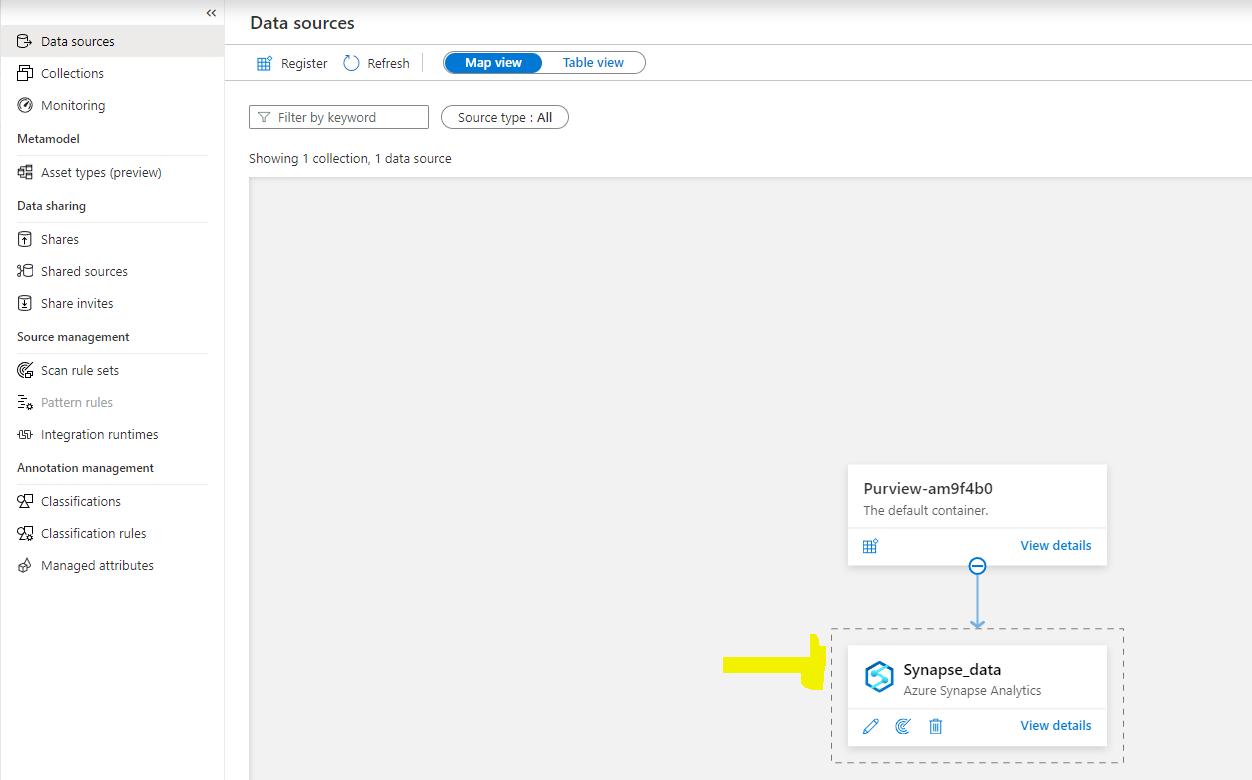
1. In the ****Register sources**** tab that appears, select ****Azure Synapse Analytics****, and continue to register a source with the following settings:



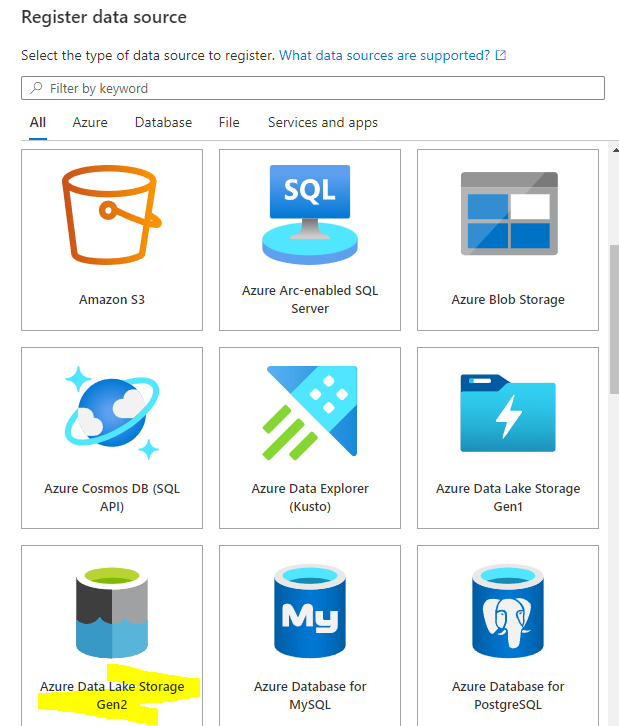
* ****Name****: Synapse\_data
* ****Azure subscription****: Select your Azure subscription
* ****Workspace name****: Select your ****synapsexxxxxxx**** workspace
* ****Dedicated SQL endpoint****: sqlxxxxxxx.sql.azuresynapse.net
* ****Serverless SQL endpoint****: sqlxxxxxxx-ondemand.sql.azuresynapse.net
* ****Select a collection****: Root (purviewxxxxxxx)

This data source includes the SQL databases in your Azure Synapse Analytics workspace.

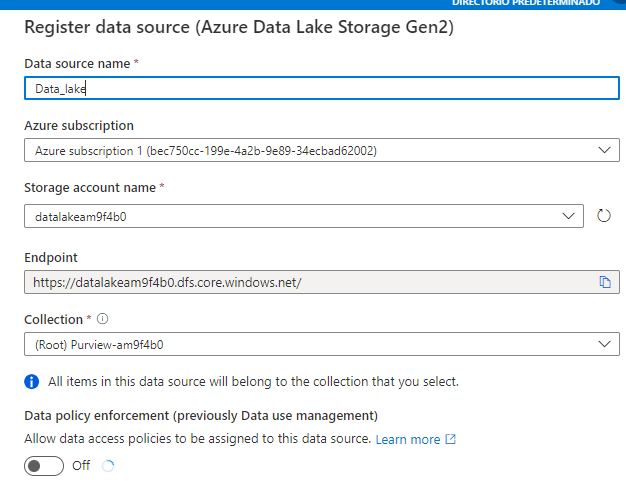




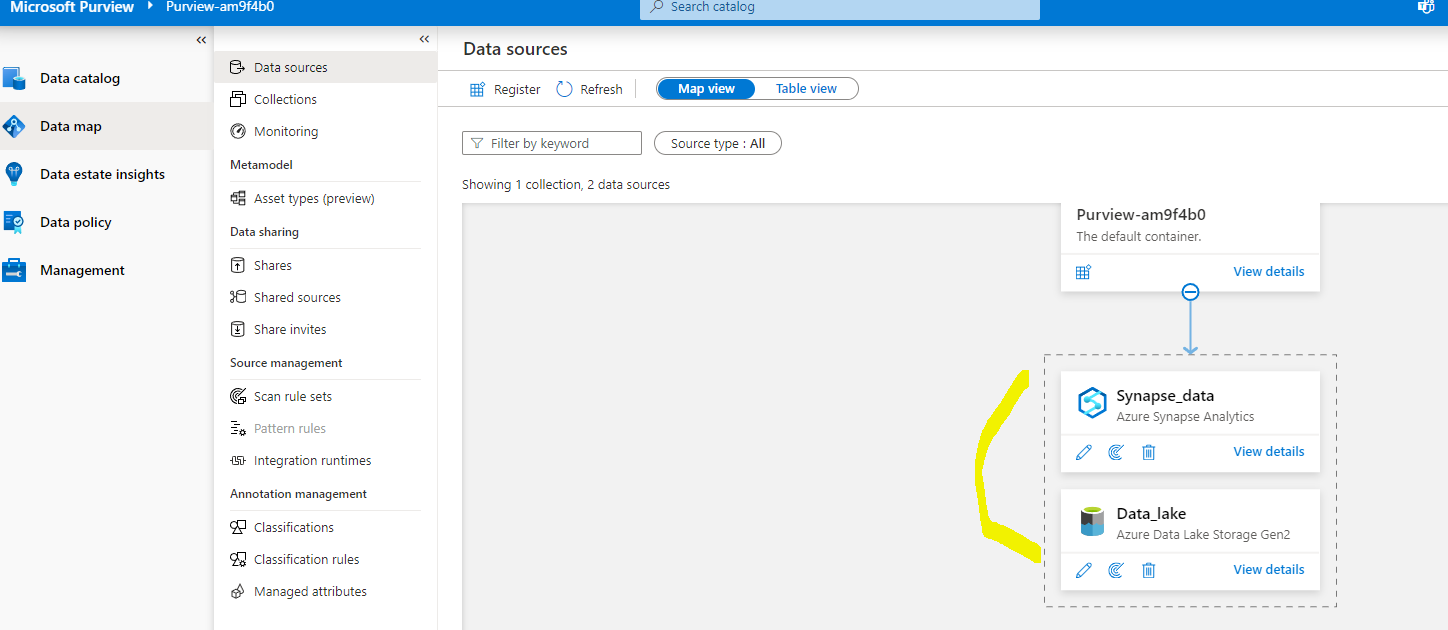
1. After registering the ****Synapse\_data**** source, select ****Register**** again, and register a second source for the data lake storage used by your Azure Synapse workspace. Select ****Azure Data Lake Storage Gen2****, and specify the following settings:



* ****Name****: Data\_lake
* ****Azure subscription****: Select your Azure subscription
* ****Workspace name****: Select your ****datalakexxxxxxx**** storage account
* ****Endpoint****: https:/ /datalakexxxxxxx.dfs.core.windows.net/
* ****Select a collection****: Root (purviewxxxxxxx)
* ****Data use management****: Disabled

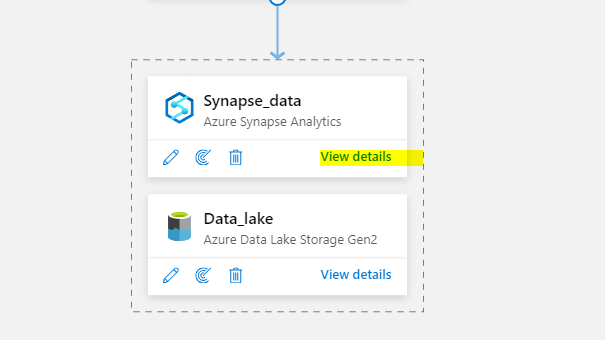


After registering both the ****Synapse\_data**** and ****Data\_lake**** sources, they should both be displayed under the ****purview**xxxxxxx** root collection in the data map as shown here:

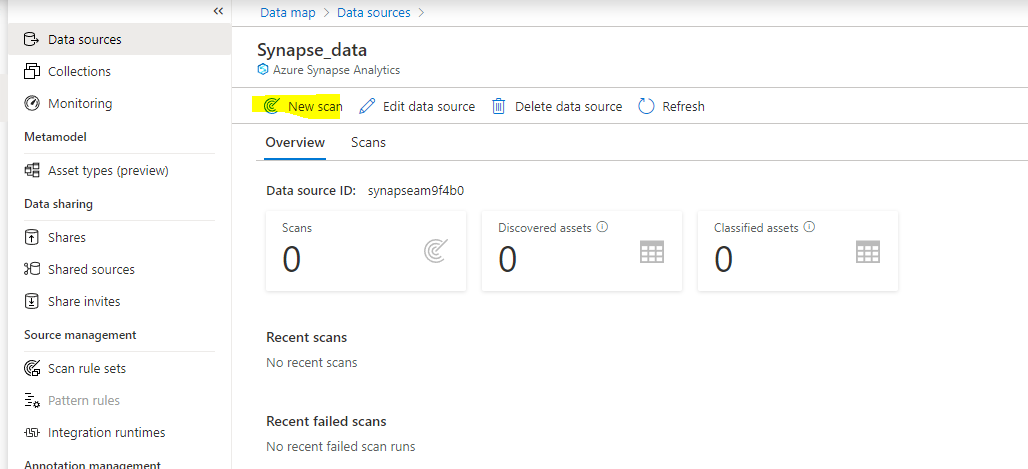


### Scan registered sources

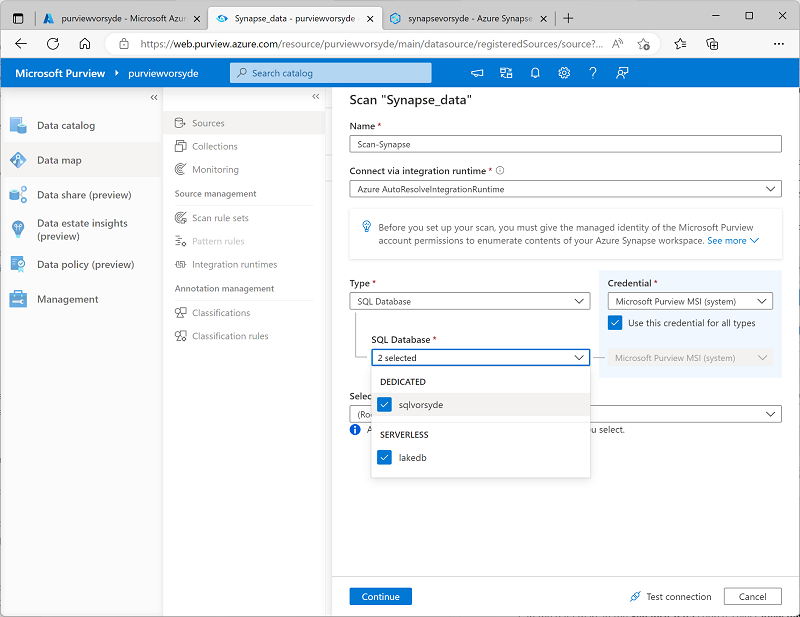
1. In the data map, in the ****Synapse\_data**** source, select ****View details****; and observe that the source has no assets cataloged. You will need to scan the source to find the data assets it contains.



1. In the ****Synapse\_data**** details page, select ****New scan****, and then configure a scan with the following settings:



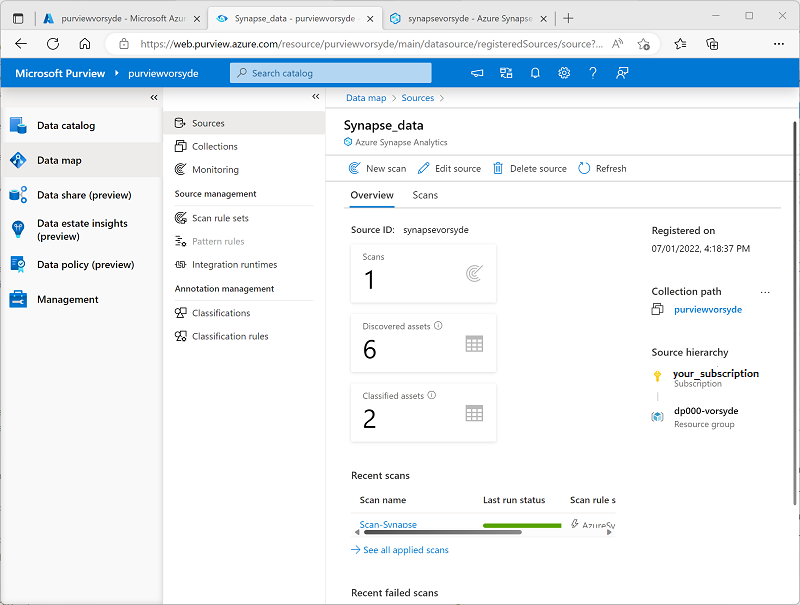
* ****Name****: Scan-Synapse
* ****Connect to integration runtime****: Azure AutoresolveIntegrationRuntime
* ****Type****: SQL Database
* ****Credential****: Microsoft Purview MSI (system)
* ****SQL Database****: Select both the ****sqlxxxxxxx**** dedicated database and the ****lakedb**** serverless database.
* ****Select a collection****: Root (purviewxxxxxxx)

[](https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/images/purview-scan.png)

1. Continue to the ****Select a scan rule set**** page, on which the default ****AzureSynapseSQL**** rule set should be selected.
2. Continue to the ****Set a scan trigger**** page, and select ****Once**** to run the scan one-time.
3. Continue to the ****Review your scan**** page, and then save and run the scan.
4. While the ****Synapse\_data**** scan is running, return to the ****Sources**** page to view the data map, and in the ****Data\_lake**** source use the ****New scan**** icon to start a scan of the data lake with the following settings:

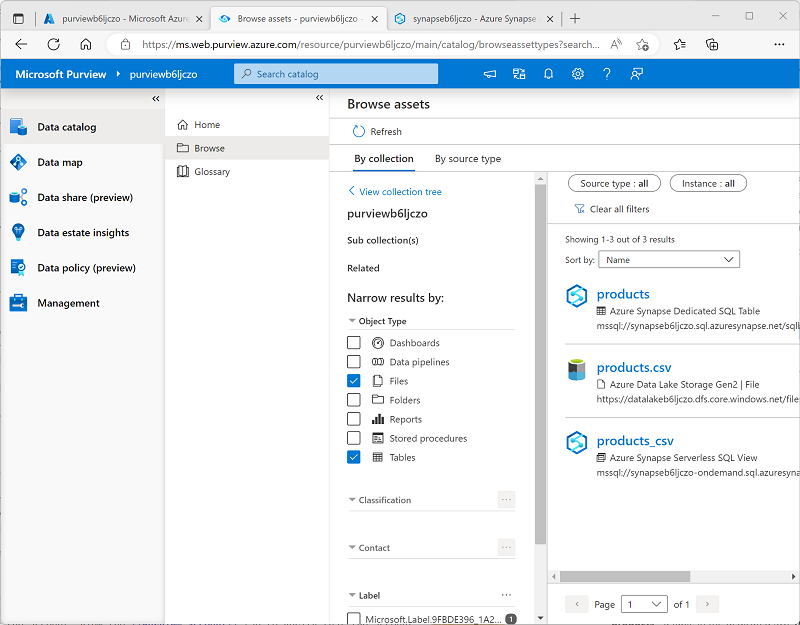
* ****Name****: Scan-Data-Lake
* ****Connect to integration runtime****: Azure AutoresolveIntegrationRuntime
* ****Credential****: Microsoft Purview MSI (system)
* ****Select a collection****: Root (purviewxxxxxxx)
* ****Scope your scan****: Select ****Data\_lake**** and all sub-assets
* ****Select a scan rule set****: AdlsGen2
* ****Set a scan trigger****: Once
* ****Review your scan**** Save and run

1. Wait for both scans to complete - this may take several minutes. You can view the details page for each of the sources to see the ****Last run status****, as shown below (you can use the ****↻ Refresh**** button to update the status). You can also view the ****Monitoring**** page (though it may take some time for the scans to appear there):

[](https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/images/purview-scan-complete.png)

### View the scanned assets

1. On the ****Data catalog**** page, on the ****Browse**** sub-page, select the ****purview**xxxxxxx** collection. Here you can see the data assets that were cataloged in your Azure Synapse Workspace and data lake storage, including the Azure Synapse Analytics workspace, the Azure Storage account for the data lake, the two SQL pool databases in Azure Synapse Analytics, the ****dbo**** schema in each database, the tables and views in the databases, and the folders and files in the data lake.
2. To filter the results, in the ****Narrow results by**** list of object types, select ****Files**** and ****Tables**** so that only the files, tables, and views that were cataloged by the scan are listed:

[](https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/images/purview-collection.png)

Note that the data assets include:

* ****products**** - a table in the dedicated SQL pool for product data.
* ****products.csv**** - a file in the data lake.
* ****products\_csv**** - a view in the serverless SQL pool that reads product data from the products.csv file.

1. Explore the assets that were found by selecting them and viewing their properties and schema. You can edit the properties of the assets (including individual fields) to add metadata, categorizations, contact details for subject-matter experts, and other useful details so that data analysts can find a lot of information about the data assets in your data estate by exploring them in the Microsoft Purview data catalog.

So far, you’ve used Microsoft Purview to catalog data assets in your Azure Synapse Analytics workspace. You can register multiple kinds of data source in a Microsoft Purview catalog, enabling you to create a central, consolidated view of data assets.

Now let’s explore some other ways to integrate Azure Synapse Analytics and Microsoft Purview.

## Integrate Microsoft Purview with Azure Synapse Analytics

Azure Synapse Analytics supports integration with Microsoft Purview to make data assets discoverable and to track data lineage through ingestion pipelines that transfer data from one source to another.

### Enable Microsoft Purview integration in Azure Synapse Analytics

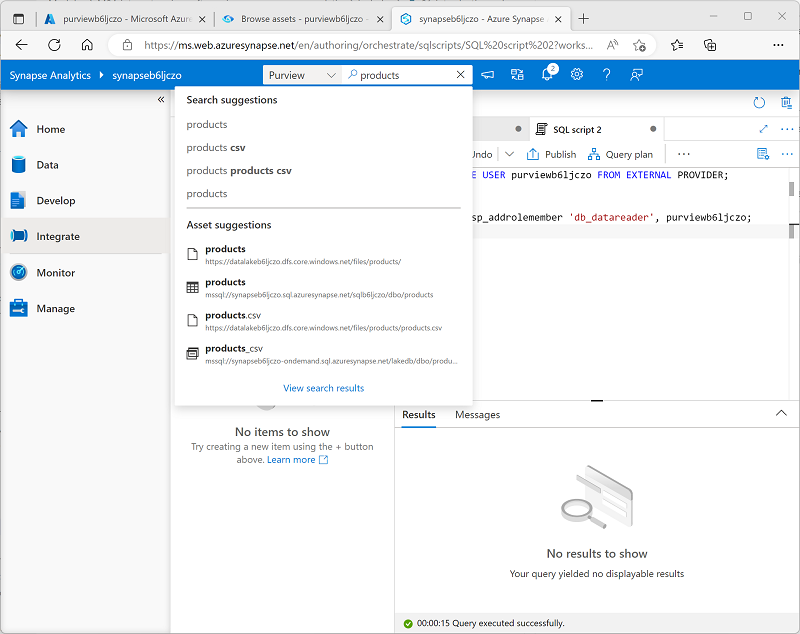
1. Switch back to the browser tab containing Synapse Studio, and on the ****Manage**** page, select the ****Microsoft Purview**** tab, and then use the ****Connect to a Purview account**** button to connect the ****purview**xxxxxxx** account in your subscription to the workspace.
2. After connecting the account, view the ****Purview account**** tab to verify that the account is has a ****Data Lineage - Synapse Pipeline**** status of ****Connected****:

### 

### Search the Purview catalog in Synapse Studio

Now that you’ve connected your Microsoft Purview account to your Azure Synapse Analytics workspace, you can search the catalog from Synapse Studio, enabling you to discover data assets across your data estate.

1. In Synapse Studio, view the ****Integrate**** page.
2. At the top of the page, use the ****Search**** box at the top to search the ****Purview**** source for the term “products”, as shown here:

[](https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/images/synapse-search-purview.png)

1. In the results, select ****products.csv**** to view its details from the Purview catalog.

By integrating the Purview catalog into the Synapse Studio interface, data analysts and engineers can find and examine registered data assets from across the entire data estate (not just within the Azure Synapse Studio workspace).

### Create and run a pipeline

The ****products\_csv**** view in the ****lakedb**** database is based on a text file in the data lake that contains product data. The ****products**** table in the ****sql**xxxxxxx** dedicated SQL database is currently empty. Let’s use a Synapse pipeline to load data from the data lake into the table.

1. In Synapse Studio, on the ****Integrate**** page, in the ****+**** menu, select ****Copy Data tool****.
2. In the Copy Data tool, select ****Built-in copy task****, and ****Run once now****, and then select ****Next****.
3. On the ****Source data store**** page, in the ****Connection**** list, select the ****synapse**xxxxxxx**-WorkspaceDefaultStorage**** connection (which references the data lake for the workspace), and for the ****File or folder****, browse to the ****files/products/products.csv**** file. Then select ****Next****.
4. On the ****File format settings**** page, select ****Detect text format****. Then ensure the following settings are specified before selecting ****Next****:

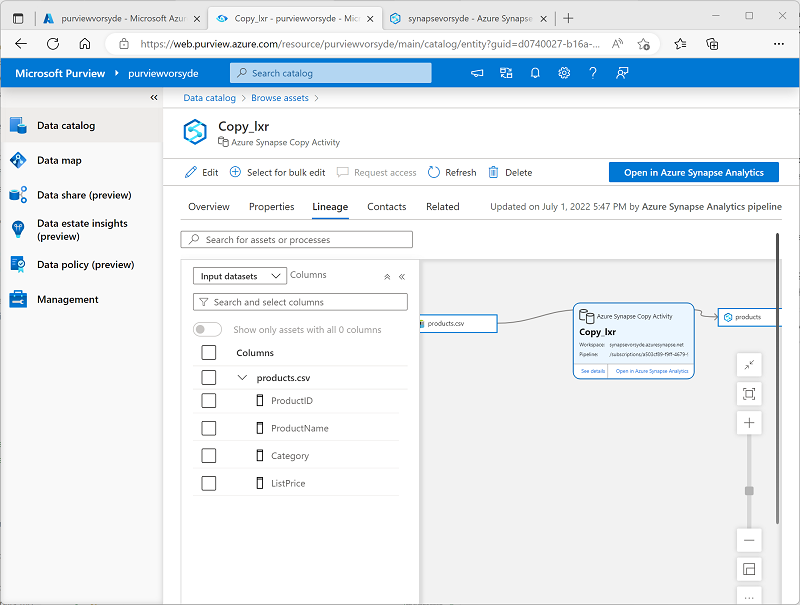
* ****File format****: DelimitedText
* ****Column delimiter****: Comma (,)
* ****Row delimiter****: Line feed (\n)
* ****First row as header****: Selected
* ****Compression type****: None

1. On the ****Destination data store**** page, in the ****Connection**** list, select ****sql**xxxxxxx** (the connection to your dedicated SQL pool). Then set the target to the existing ****dbo.products**** table, and select ****Next****.
2. On the ****Column mapping**** page, review the default column mappings and then select ****Next****.
3. On the ****Settings**** page, set the ****Task name**** to ****Load\_Product\_Data****. Then select the ****Bulk insert**** copy method and select ****Next****.
4. On the ****Summary**** page, select ****Next****.
5. Wait for the pipeline to be deployed, and then select ****Finish****.
6. In Synapse Studio, view the ****Monitor**** page. Then on the ****Pipeline runs**** tab, observe the status of the ****Load\_Product\_Data**** pipeline. It may take a few minutes for the status to change to ****Succeeded****.
7. When the pipeline run has completed successfully, select its name (****Load\_Product\_Data****) to view details of the activities in the pipeline; and observe that the pipeline includes a ****Copy data**** task with an automatically-derived name similar to ****Copy\_**xxx**. This activity copied the data from the text file in the data lake into the ****products**** table in the ****sql**xxxxxxx** database.

### View data lineage in Microsoft Purview

You’ve used a Synapse pipeline to load data into a database. Let’s verify that this activity has been tracked in Microsoft Purview.

1. Switch to the browser tab containing the Microsoft Purview Governance Portal.
2. On the ****Data catalog**** page, on the ****Browse**** sub-page, select the ****purview**xxxxxxx** collection.
3. Filter the assets to show only ****Data pipelines****, ****Files****, and ****Tables****. The list of assets should include the ****products.csv**** file, the ****Copy\_**xxx** pipeline activity, and the ****products**** table.
4. Select the ****Copy\_**xxx** asset to view its details, noting the ****Updated time**** reflects the recent pipeline run.
5. On the ****Lineage**** tab for the ****Copy\_**xxx** asset, view the diagram showing the data flow from the ****products.csv**** file to the ****products**** table:

[](https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/images/purview-lineage.png)

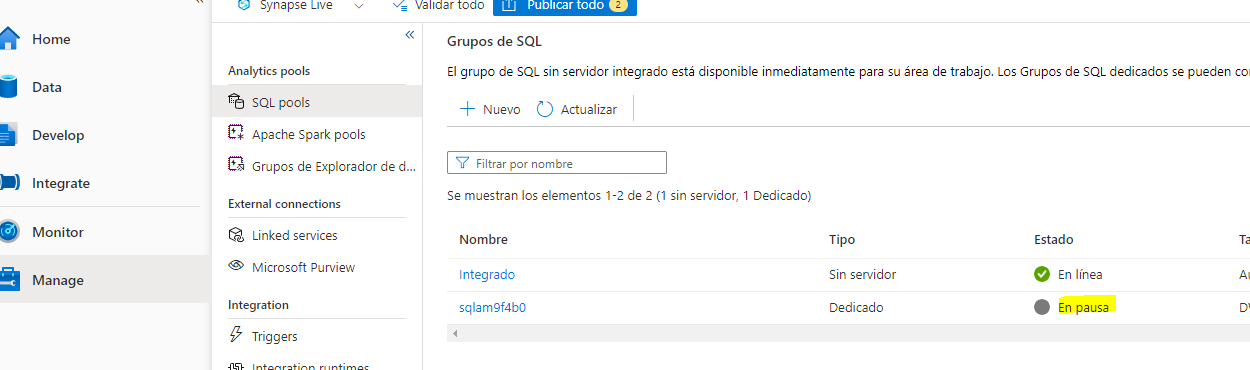
1. In the ****Copy\_**xxx** lineage diagram, select the ****products.csv**** file and use its ****Switch to asset**** link to view details of the source file.
2. In the ****products.csv**** lineage diagram, select the ****products**** table and use its ****Switch to asset**** link to view details of the table (you may need to use the ****↻ Refresh**** button to see the table lineage diagram).

The lineage tracking capability enabled by integrating Azure Synapse Analytics with Microsoft Purview enables you to determine how and when the data in your data stores was loaded, and where it came from.

****Tip****: In this exercise, you viewed the lineage information for in the Microsoft Purview Governance portal; but remember that the same assets can also be viewed in Synapse Studio through the search integration feature.

### Pause the dedicated SQL pool

1. Switch back to the Synapse Studio tab, and on the ****Manage**** page, pause the ****sql**xxxxxxx** dedicated SQL pool.



## Delete Azure resources

If you’ve finished exploring Azure Synapse Analytics, you should delete the resources you’ve created to avoid unnecessary Azure costs.

1. Close the Synapse Studio browser tab and return to the Azure portal.
2. On the Azure portal, on the ****Home**** page, select ****Resource groups****.
3. Select the ****dp203-**xxxxxxx** resource group for your Synapse Analytics workspace (not the managed resource group), and verify that it contains the Synapse workspace, storage account, and dedicated SQL pool for your workspace.
4. At the top of the ****Overview**** page for your resource group, select ****Delete resource group****.
5. Enter the ****dp203-**xxxxxxx** resource group name to confirm you want to delete it, and select ****Delete****.

After a few minutes, your Azure Synapse workspace resource group and the managed workspace resource group associated with it will be deleted.