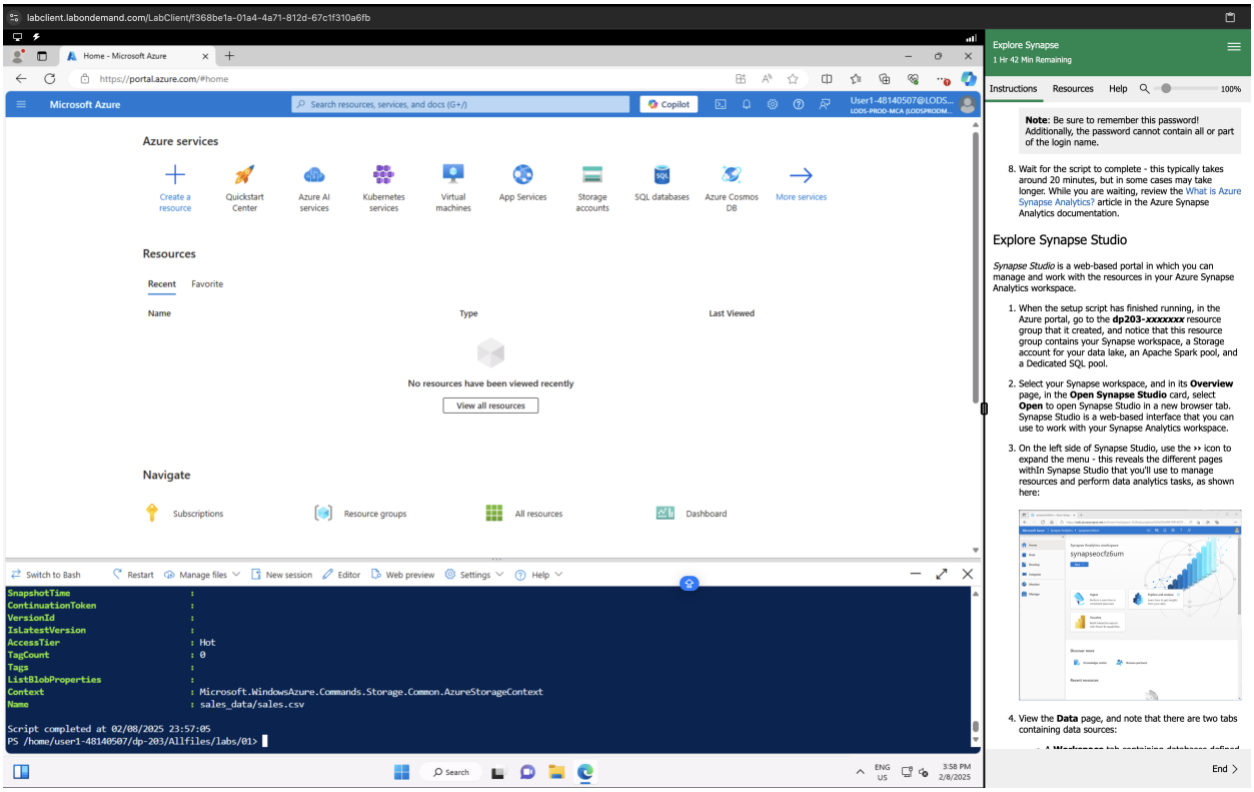
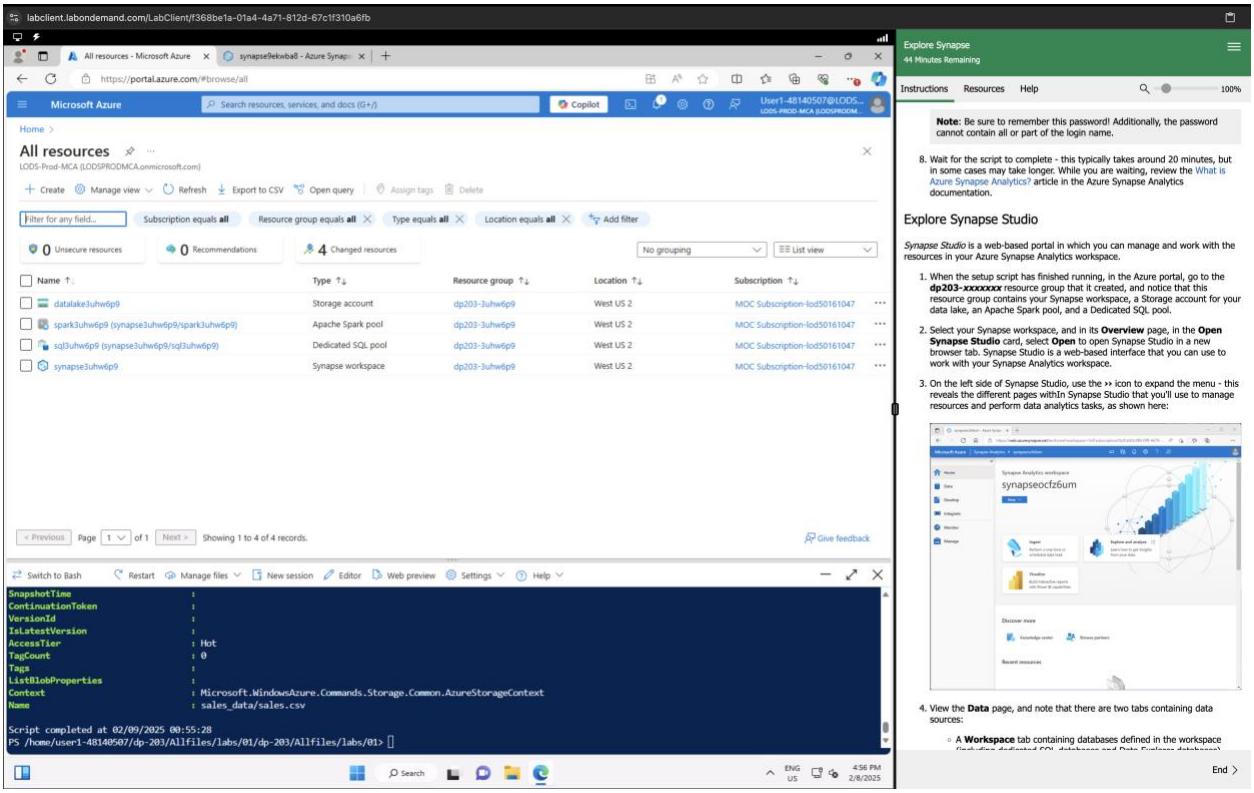
Lab1-Explore Synapse
BUAN 6390.001 – Analytics Practicum

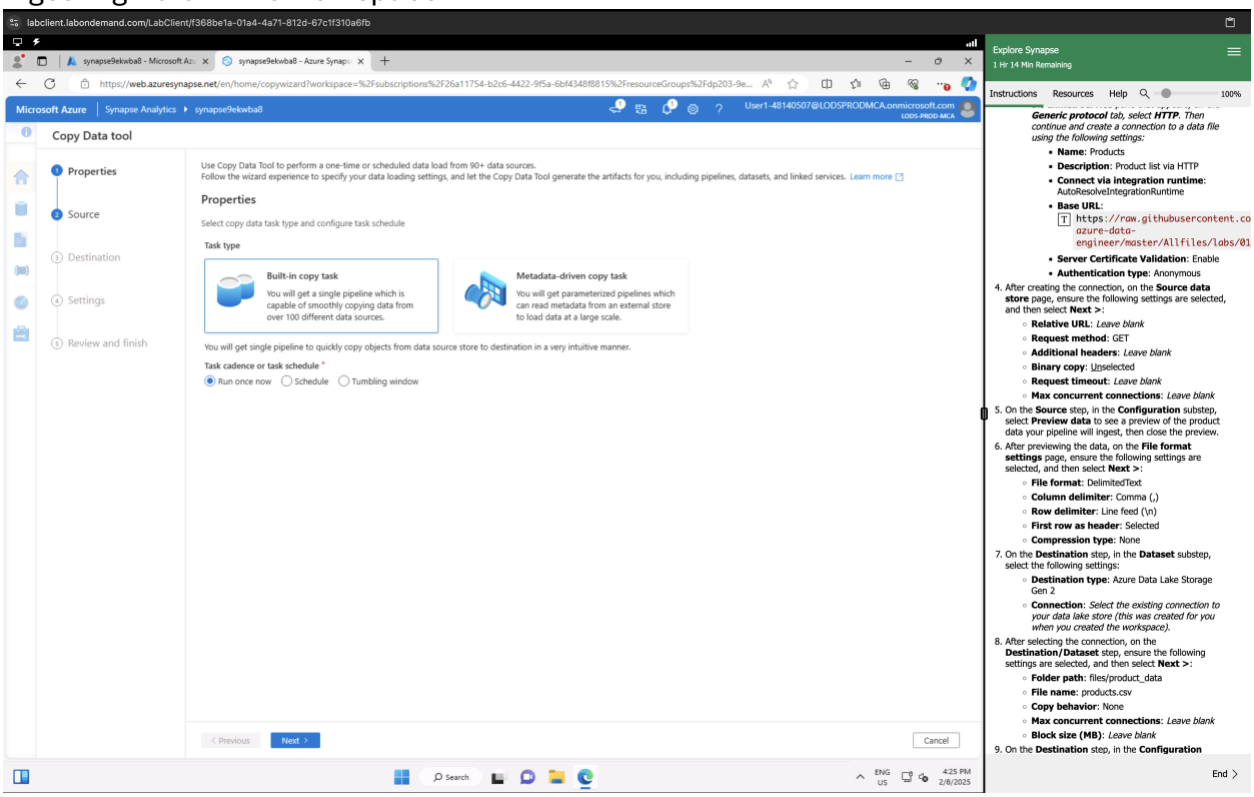
Name: Swastik Bhatnagar
NetID: sxb220210

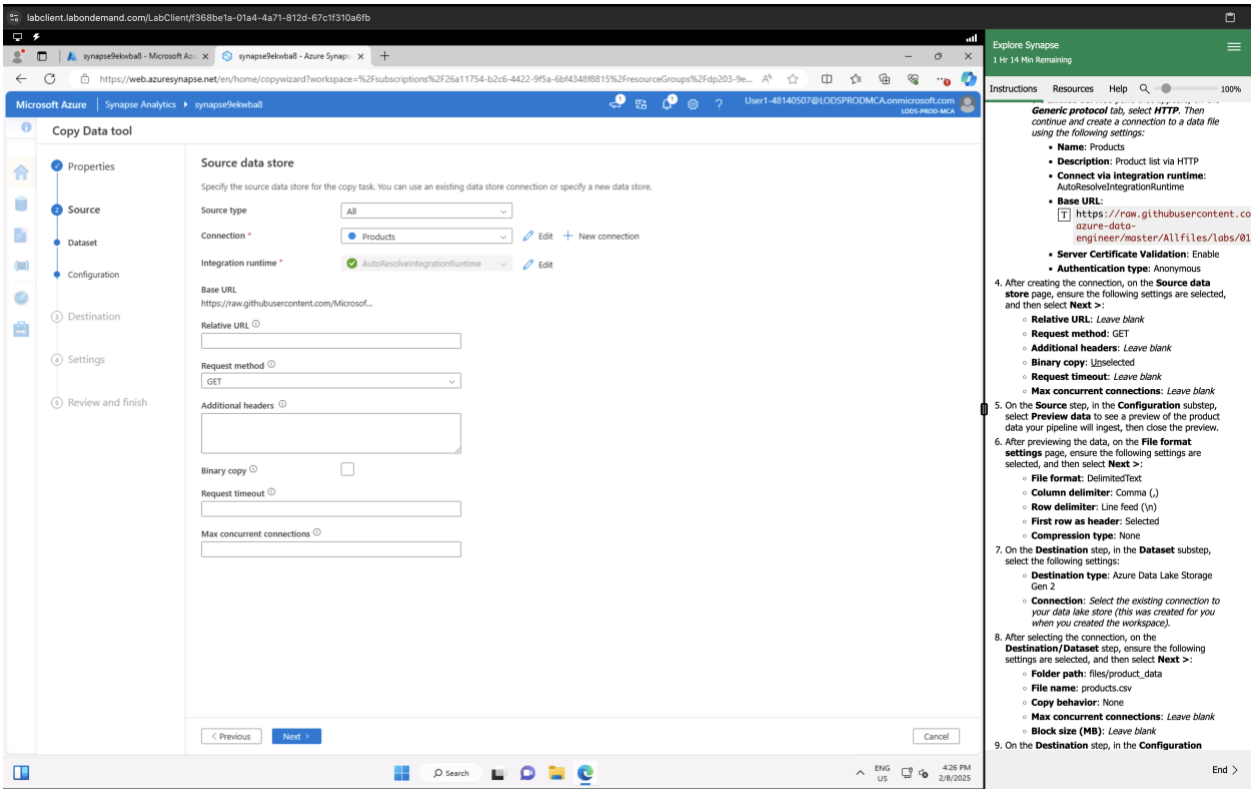
Connecting to PowerShell and Creating a resource group





Ingesting Data in the workspace:

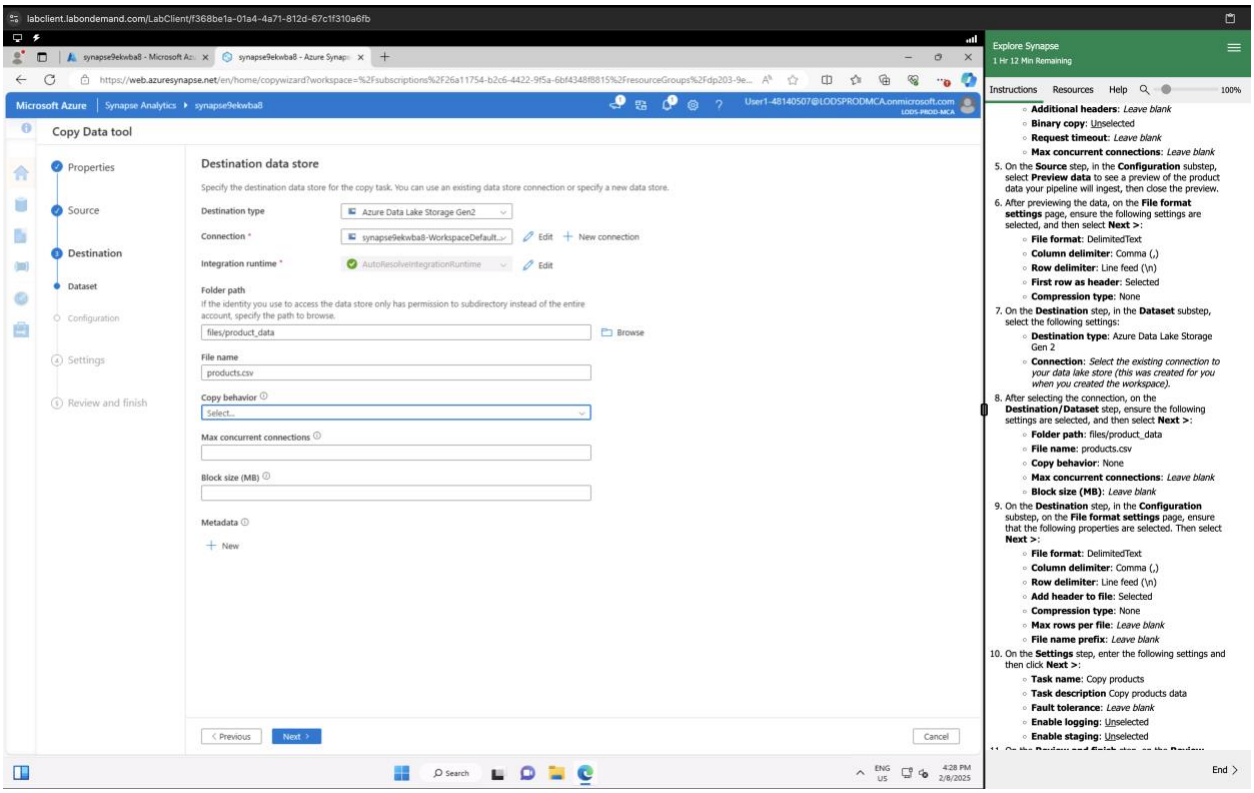




***Generic protocol** tab, select **HTTP**. Then continue and create a connection to a data file using the following settings:*

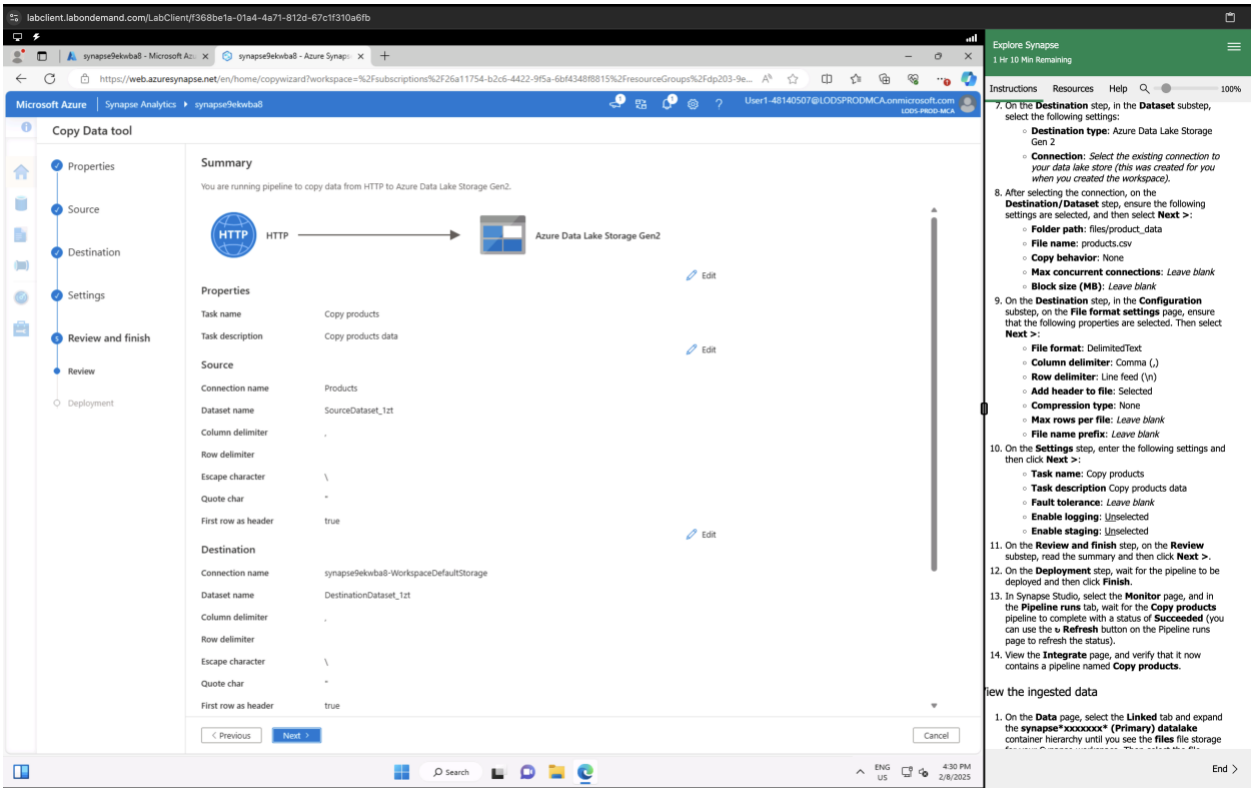
- **Name**: Products
- **Description**: Product list via HTTP
- **Connect via integration runtime**: AutoResolveIntegrationRuntime
- **Base URL**: https://raw.githubusercontent.com/MicrosoftLearning/azure-data-engineer/master/Allfiles/labs/01
- **Server Certificate Validation**: Enable
- **Authentication type**: Anonymous

4. After creating the connection, on the **Source data store** page, ensure the following settings are selected, and then select **Next >**:
   - **Relative URL**: *Leave blank*
   - **Request method**: GET
   - **Additional headers**: *Leave blank*
   - **Binary copy**: Unselected
   - **Request timeout**: *Leave blank*
   - **Max concurrent connections**: *Leave blank*
5. On the **Source** step, in the **Configuration** substep, select **Preview data** to see a preview of the product data your pipeline will ingest, then close the preview.
6. After previewing the data, on the **File format settings** page, ensure the following settings are selected, and then select **Next >**:
   - **File format**: DelimitedText
   - **Column delimiter**: Comma (,)
   - **Row delimiter**: Line feed (\n)
   - **First row as header**: Selected
   - **Compression type**: None
7. On the **Destination** step, in the **Dataset** substep, select the following settings:
   - **Destination type**: Azure Data Lake Storage Gen 2
   - **Connection**: *Select the existing connection to your data lake store (this was created for you when you created the workspace).*
8. After selecting the connection, on the **Destination/Dataset** step, ensure the following settings are selected, and then select **Next >**:
   - **Folder path**: files/product_data
   - **File name**: products.csv
   - **Copy behavior**: None
   - **Max concurrent connections**: *Leave blank*
   - **Block size (MB)**: *Leave blank*
9. On the **Destination** step, in the **Configuration**



- **Additional headers**: *Leave blank*
- **Binary copy**: Unselected
- **Request timeout**: *Leave blank*
- **Max concurrent connections**: *Leave blank*

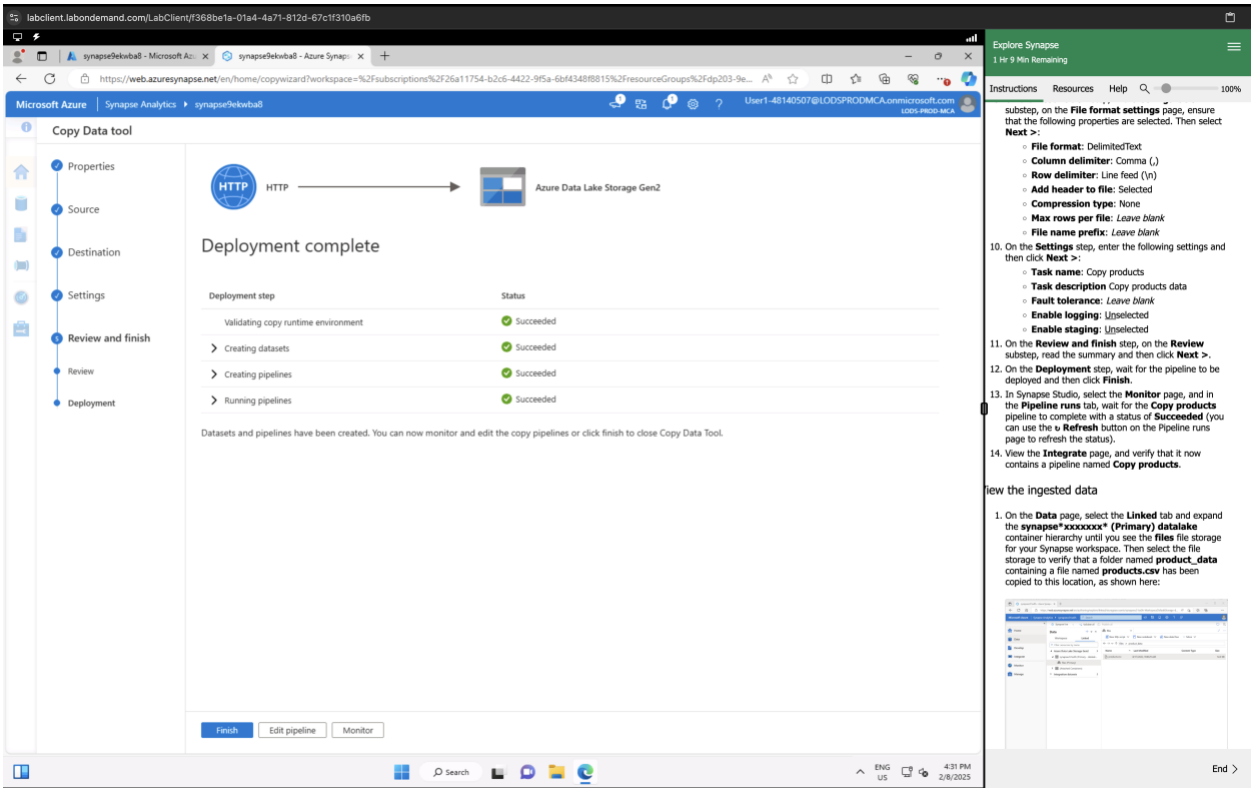
5. On the **Source** step, in the **Configuration** substep, select **Preview data** to see a preview of the product data your pipeline will ingest, then close the preview.
6. After previewing the data, on the **File format settings** page, ensure the following settings are selected, and then select **Next >**:
   - **File format**: DelimitedText
   - **Column delimiter**: Comma (,)
   - **Row delimiter**: Line feed (\n)
   - **First row as header**: Selected
   - **Compression type**: None
7. On the **Destination** step, in the **Dataset** substep, select the following settings:
   - **Destination type**: Azure Data Lake Storage Gen 2
   - **Connection**: *Select the existing connection to your data lake store (this was created for you when you created the workspace).*
8. After selecting the connection, on the **Destination/Dataset** step, ensure the following settings are selected, and then select **Next >**:
   - **Folder path**: files/product_data
   - **File name**: products.csv
   - **Copy behavior**: None
   - **Max concurrent connections**: *Leave blank*
   - **Block size (MB)**: *Leave blank*
9. On the **Destination** step, in the **Configuration** substep, on the **File format settings** page, ensure that the following properties are selected. Then select **Next >**:
   - **File format**: DelimitedText
   - **Column delimiter**: Comma (,)
   - **Row delimiter**: Line feed (\n)
   - **Add header to file**: Selected
   - **Compression type**: None
   - **Max rows per file**: *Leave blank*
   - **File name prefix**: *Leave blank*
10. On the **Settings** step, enter the following settings and then click **Next >**:
    - **Task name**: Copy products
    - **Task description** Copy products data
    - **Fault tolerance**: *Leave blank*
    - **Enable logging**: Unselected
    - **Enable staging**: Unselected



7. On the **Destination** step, in the **Dataset** substep, select the following settings:
   - **Destination type**: Azure Data Lake Storage Gen 2
   - **Connection**: *Select the existing connection to your data lake store (this was created for you when you created the workspace).*
8. After selecting the connection, on the **Destination/Dataset** step, ensure the following settings are selected, and then select **Next >**:
   - **Folder path**: files/product_data
   - **File name**: products.csv
   - **Copy behavior**: None
   - **Max concurrent connections**: *Leave blank*
   - **Block size (MB)**: *Leave blank*
9. On the **Destination** step, in the **Configuration** substep, on the **File format settings** page, ensure that the following properties are selected. Then select **Next >**:
   - **File format**: DelimitedText
   - **Column delimiter**: Comma (,)
   - **Row delimiter**: Line feed (\n)
   - **Add header to file**: Selected
   - **Compression type**: None
   - **Max rows per file**: *Leave blank*
   - **File name prefix**: *Leave blank*
10. On the **Settings** step, enter the following settings and then click **Next >**:
    - **Task name**: Copy products
    - **Task description** Copy products data
    - **Fault tolerance**: *Leave blank*
    - **Enable logging**: Unselected
    - **Enable staging**: Unselected
11. On the **Review and finish** step, on the **Review** substep, read the summary and then click **Next >**.
12. On the **Deployment** step, wait for the pipeline to be deployed and then click **Finish**.
13. In Synapse Studio, select the **Monitor** page, and in the **Pipeline runs** tab, wait for the **Copy products** pipeline to complete with a status of **Succeeded** (you can use the ↻ **Refresh** button on the Pipeline runs page to refresh the status).
14. View the **Integrate** page, and verify that it now contains a pipeline named **Copy products**.

iew the ingested data

1. On the **Data** page, select the **Linked** tab and expand the **synapse*xxxxxxx* (Primary) datalake** container hierarchy until you see the **files** file storage

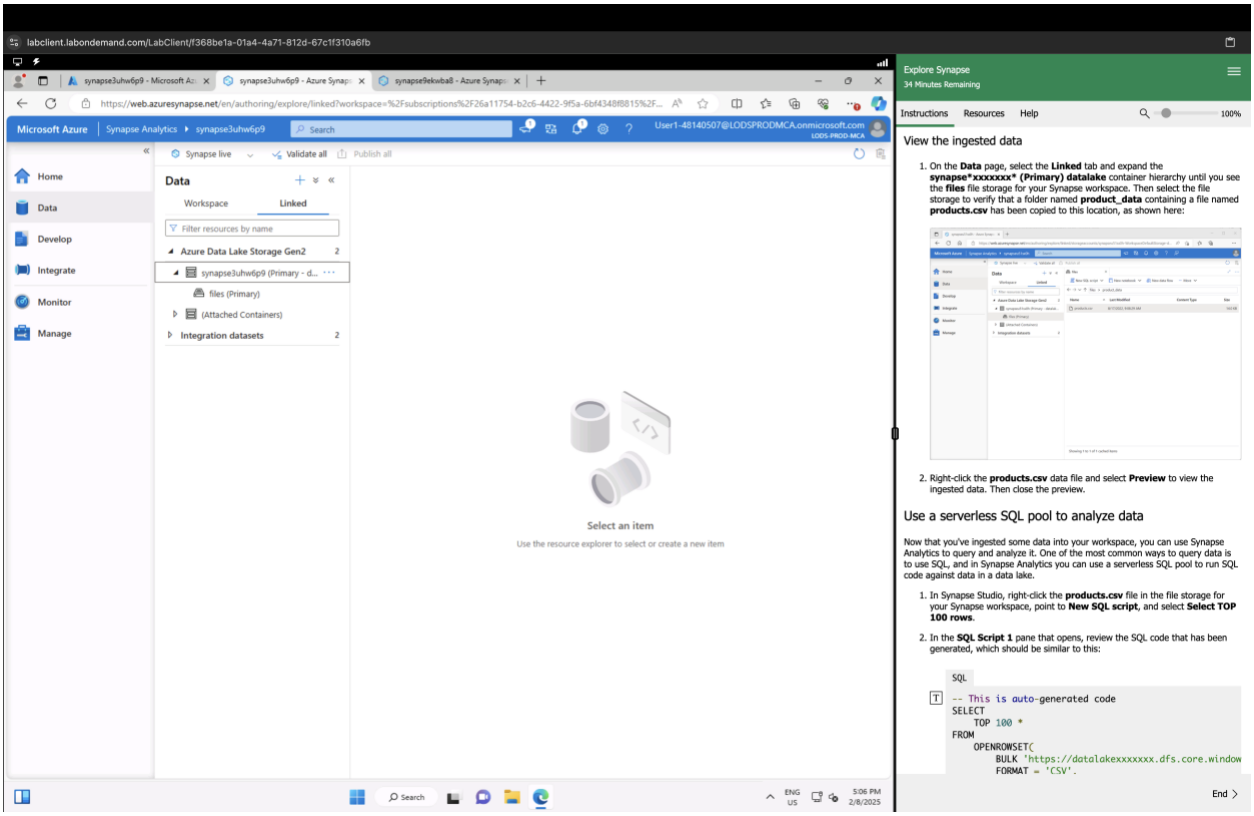


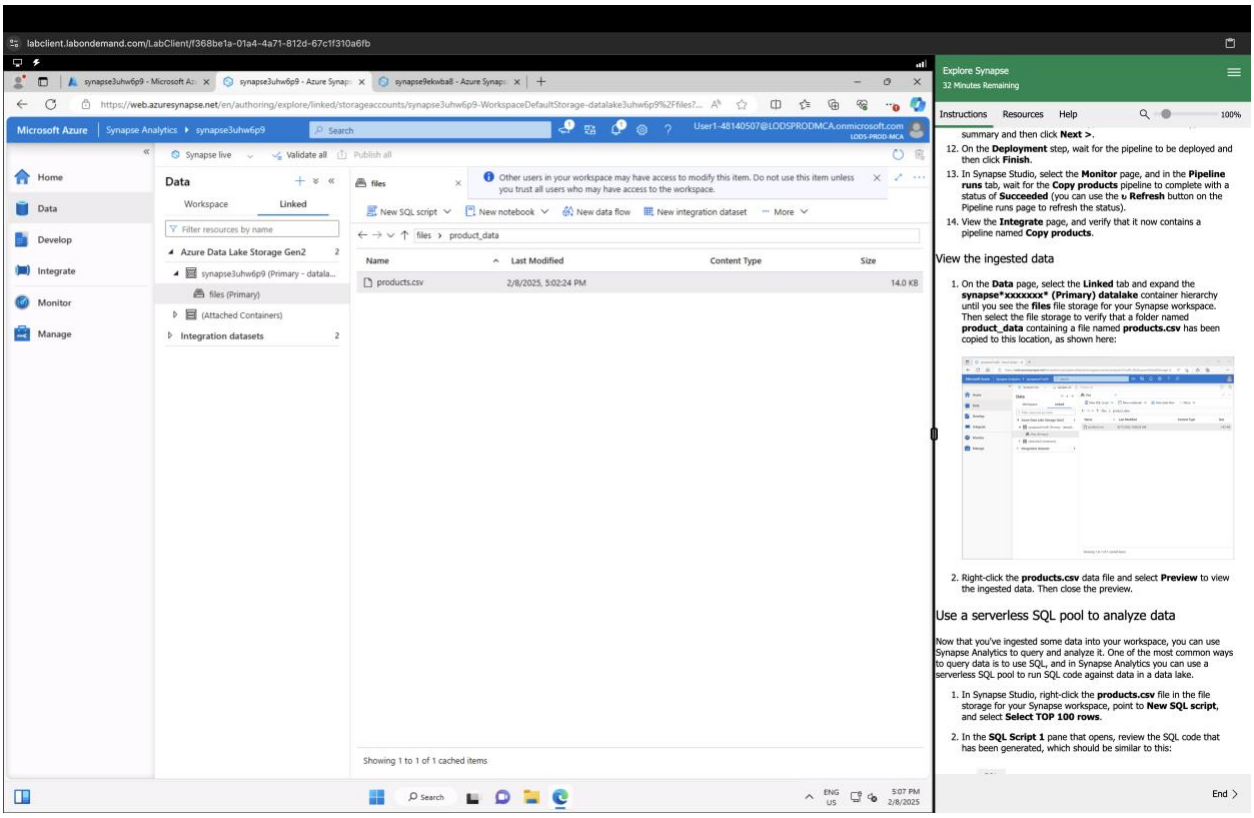
substep, on the **File format settings** page, ensure that the following properties are selected. Then select **Next >**:
   - **File format**: DelimitedText
   - **Column delimiter**: Comma (,)
   - **Row delimiter**: Line feed (\n)
   - **Add header to file**: Selected
   - **Compression type**: None
   - **Max rows per file**: *Leave blank*
   - **File name prefix**: *Leave blank*
10. On the **Settings** step, enter the following settings and then click **Next >**:
    - **Task name**: Copy products
    - **Task description** Copy products data
    - **Fault tolerance**: *Leave blank*
    - **Enable logging**: Unselected
    - **Enable staging**: Unselected
11. On the **Review and finish** step, on the **Review** substep, read the summary and then click **Next >**.
12. On the **Deployment** step, wait for the pipeline to be deployed and then click **Finish**.
13. In Synapse Studio, select the **Monitor** page, and in the **Pipeline runs** tab, wait for the **Copy products** pipeline to complete with a status of **Succeeded** (you can use the ↻ **Refresh** button on the Pipeline runs page to refresh the status).
14. View the **Integrate** page, and verify that it now contains a pipeline named **Copy products**.

iew the ingested data

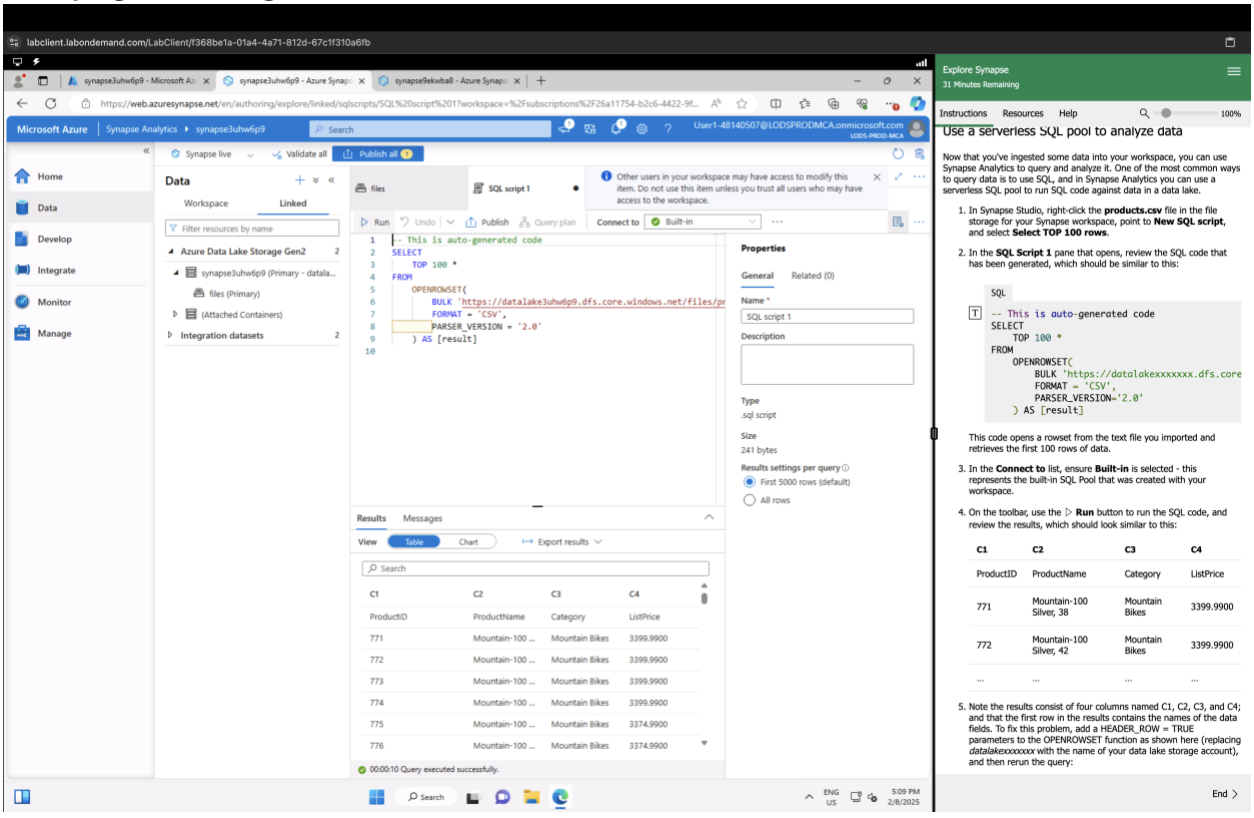
1. On the **Data** page, select the **Linked** tab and expand the **synapse*xxxxxxx* (Primary) datalake** container hierarchy until you see the **files** file storage for your Synapse workspace. Then select the file storage to verify that a folder named **product_data** containing a file named **products.csv** has been copied to this location, as shown here:

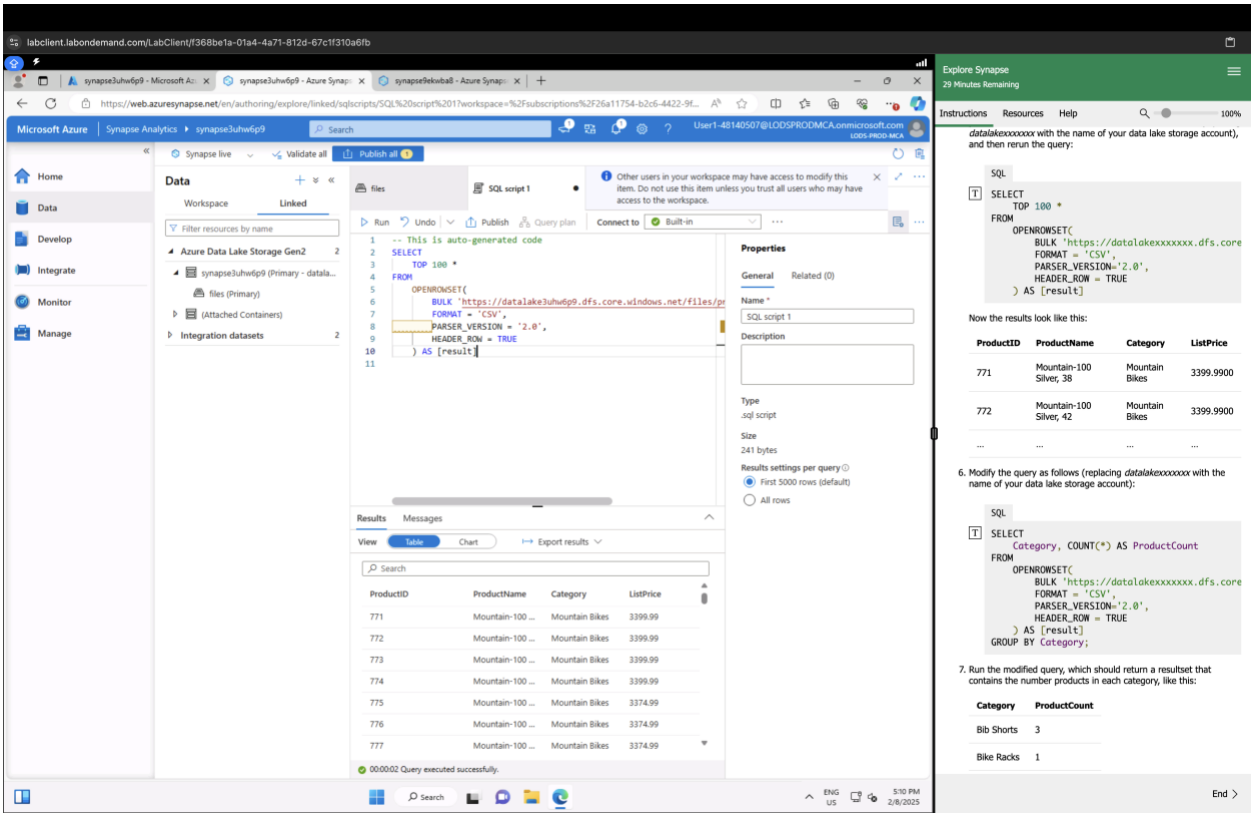
Viewing the ingested data:

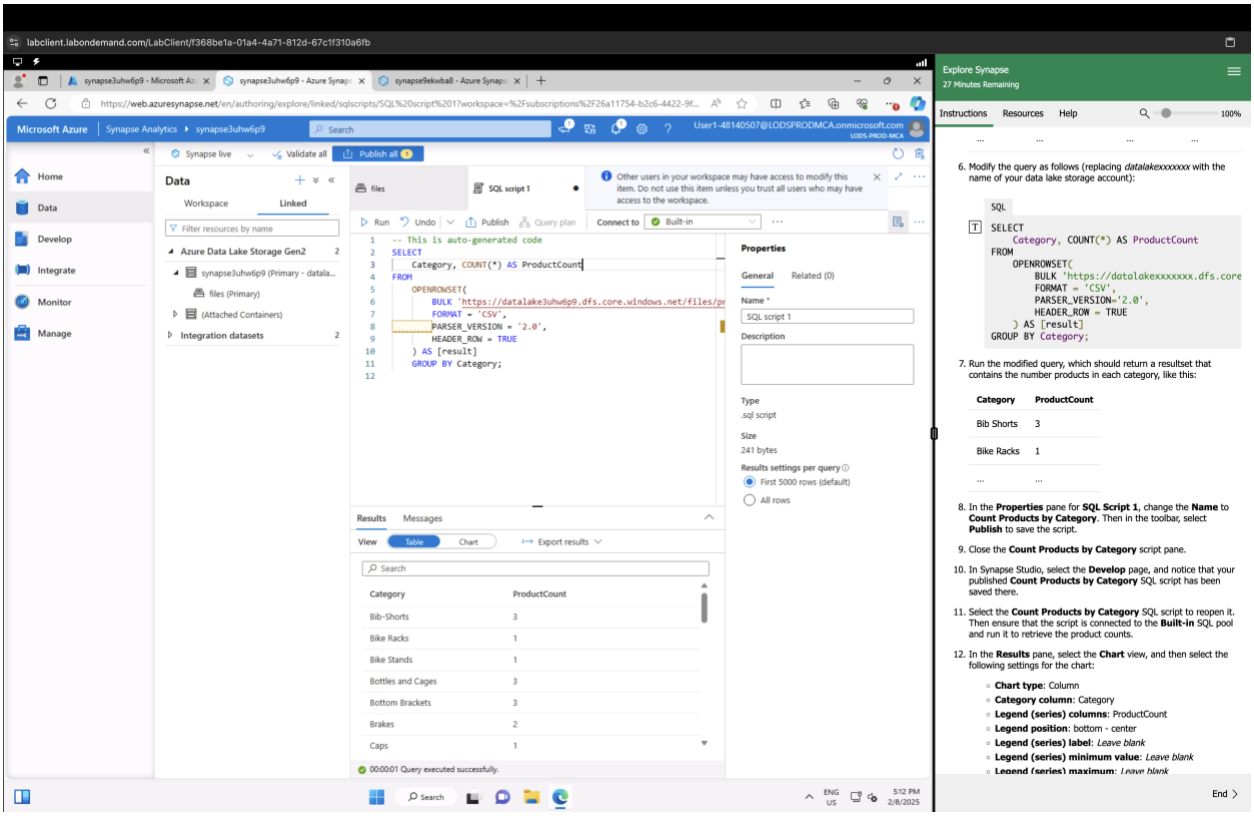




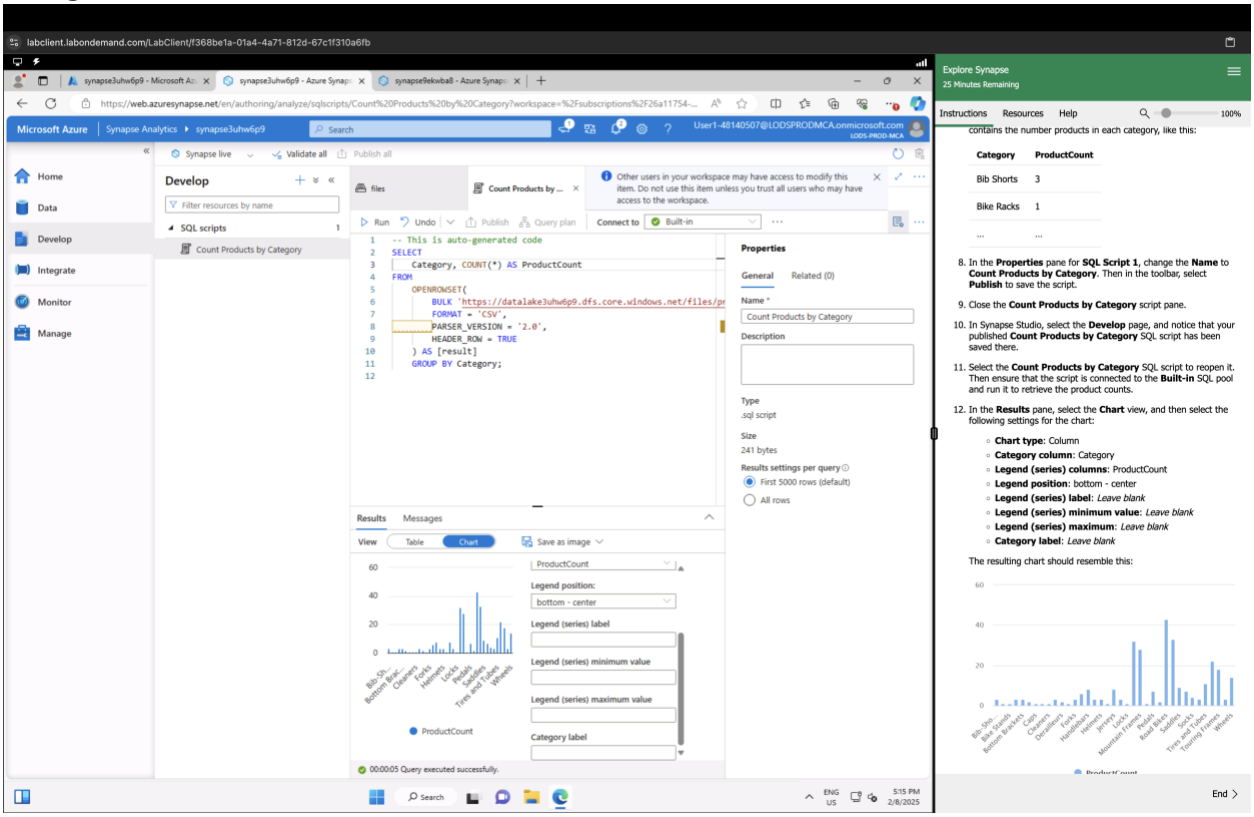
Querying data using SQL:



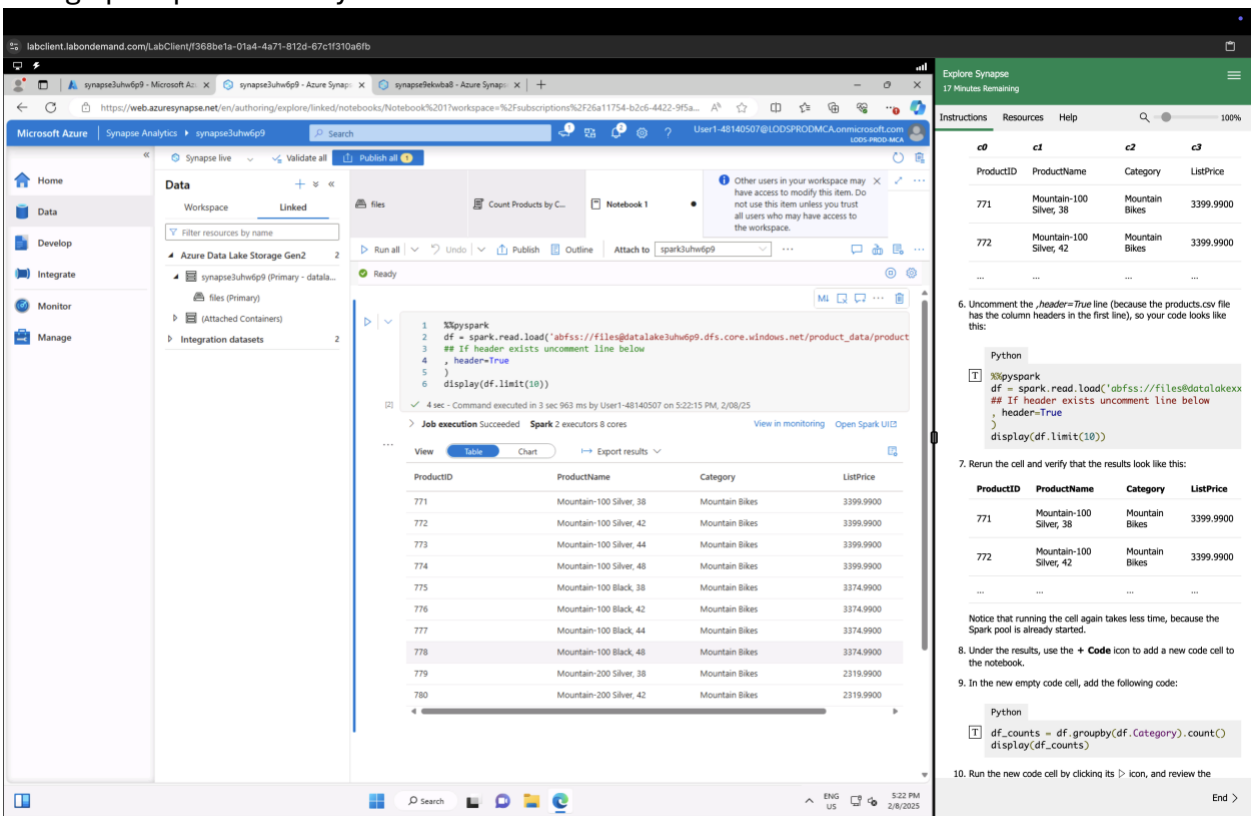

labclient.labondemand.com/LabClient/f368be1a-01a4-4a71-812d-67c1f310a6fb
Explore Synapse
29 Minutes Remaining
Instructions
Resources
Help
100%
datalakexxxxxxx with the name of your data lake storage account), and then rerun the query:
SQL
SELECT
TOP 100 *
FROM
OPENROWSET(
BULK 'https://datalakexxxxxxx.dfs.core
FORMAT = 'CSV',
PARSER_VERSION='2.0',
HEADER_ROW = TRUE
) AS [result]
Now the results look like this:
ProductID
ProductName
Category
ListPrice
771
Mountain-100 Silver, 38
Mountain Bikes
3399.9900
772
Mountain-100 Silver, 42
Mountain Bikes
3399.9900
6. Modify the query as follows (replacing datalakexxxxxxx with the name of your data lake storage account):
SQL
SELECT
Category, COUNT(*) AS ProductCount
FROM
OPENROWSET(
BULK 'https://datalakexxxxxxx.dfs.core
FORMAT = 'CSV',
PARSER_VERSION='2.0',
HEADER_ROW = TRUE
) AS [result]
GROUP BY Category;
7. Run the modified query, which should return a resultset that contains the number products in each category, like this:
Category
ProductCount
Bib Shorts
3
Bike Racks
1
End


labclient.labondemand.com/LabClient/f368be1a-01a4-4a71-812d-67c1f310a6fb
Explore Synapse
27 Minutes Remaining
Instructions
Resources
Help
100%
6. Modify the query as follows (replacing datalakexxxxxxx with the name of your data lake storage account):
SQL
SELECT
Category, COUNT(*) AS ProductCount
FROM
OPENROWSET(
BULK 'https://datalakexxxxxxx.dfs.core
FORMAT = 'CSV',
PARSER_VERSION='2.0',
HEADER_ROW = TRUE
) AS [result]
GROUP BY Category;
7. Run the modified query, which should return a resultset that contains the number products in each category, like this:
Category
ProductCount
Bib Shorts
3
Bike Racks
1
8. In the Properties pane for SQL Script 1, change the Name to Count Products by Category. Then in the toolbar, select Publish to save the script.
9. Close the Count Products by Category script pane.
10. In Synapse Studio, select the Develop page, and notice that your published Count Products by Category SQL script has been saved there.
11. Select the Count Products by Category SQL script to reopen it. Then ensure that the script is connected to the Built-in SQL pool and run it to retrieve the product counts.
12. In the Results pane, select the Chart view, and then select the following settings for the chart:
Chart type: Column
Category column: Category
Legend (series) columns: ProductCount
Legend position: bottom - center
Legend (series) label: Leave blank
Legend (series) minimum value: Leave blank
End

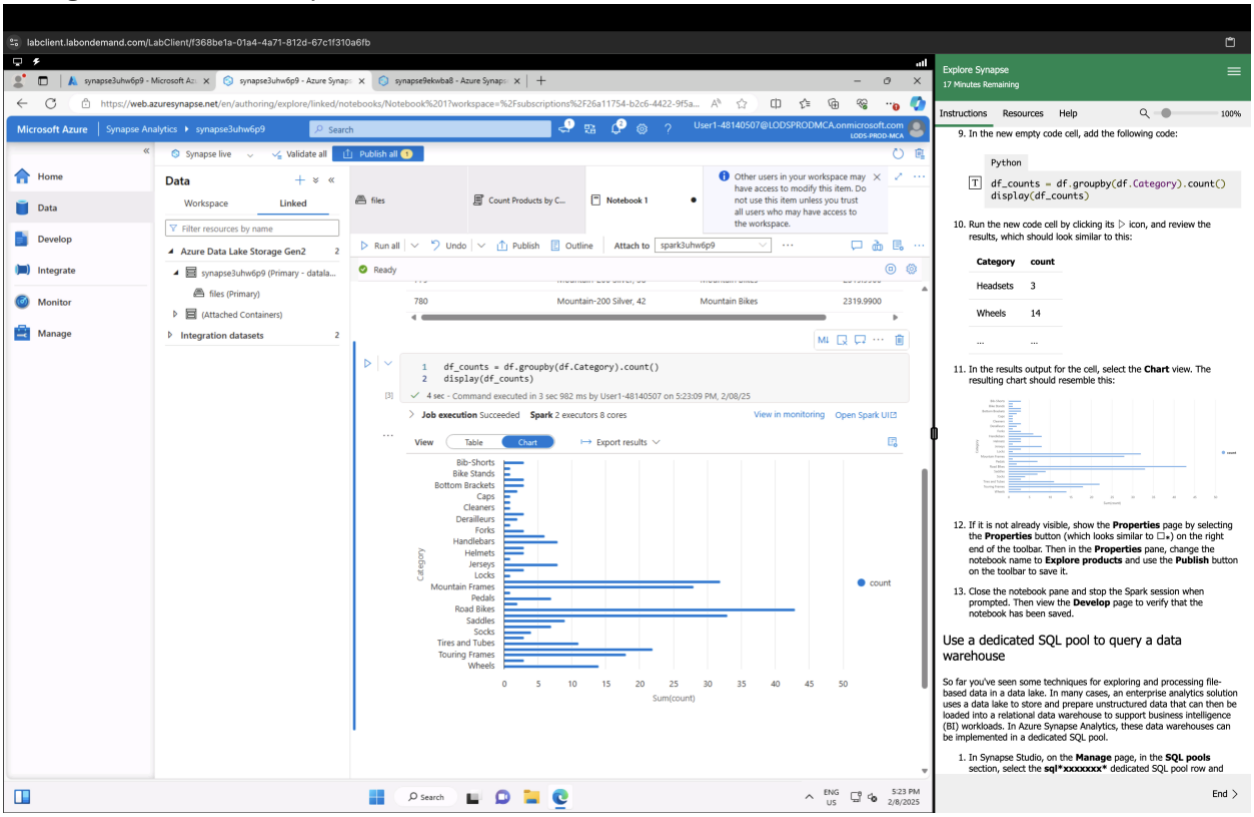
Using the Chart view:

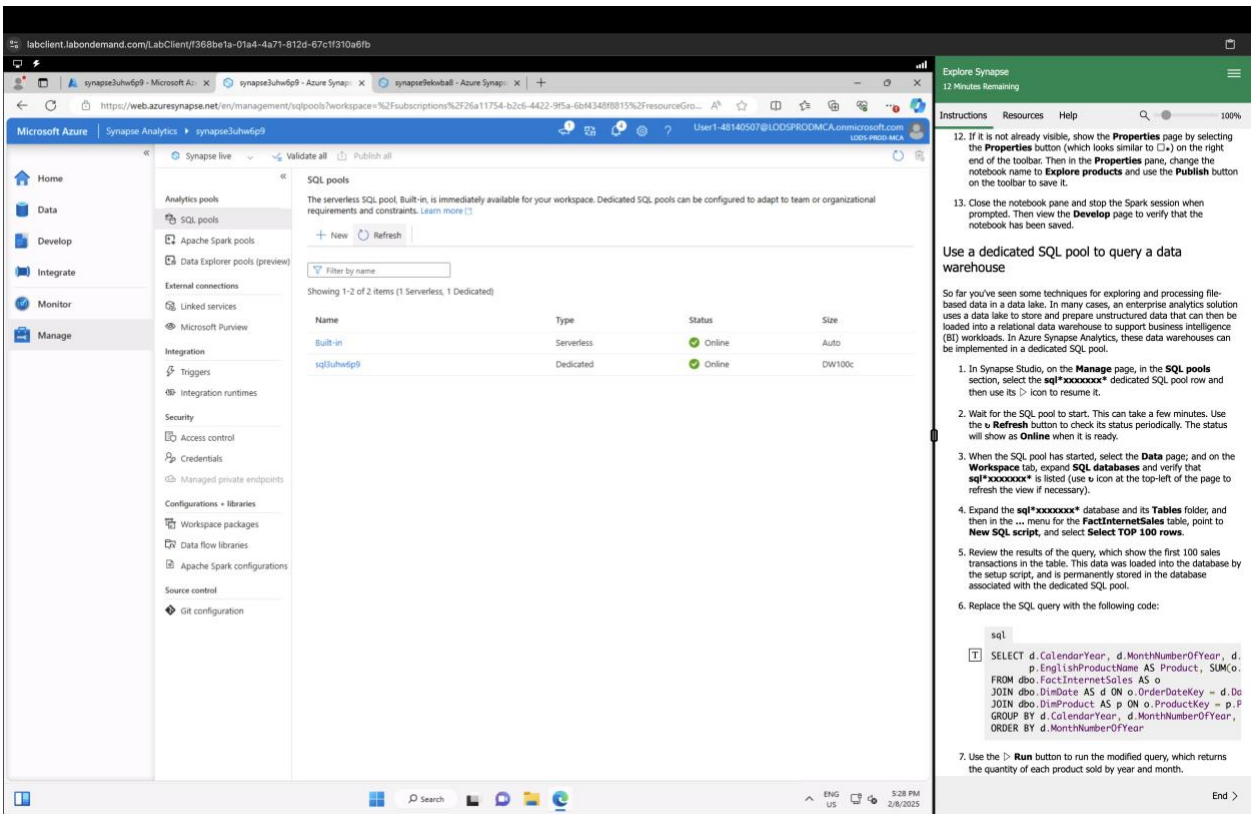


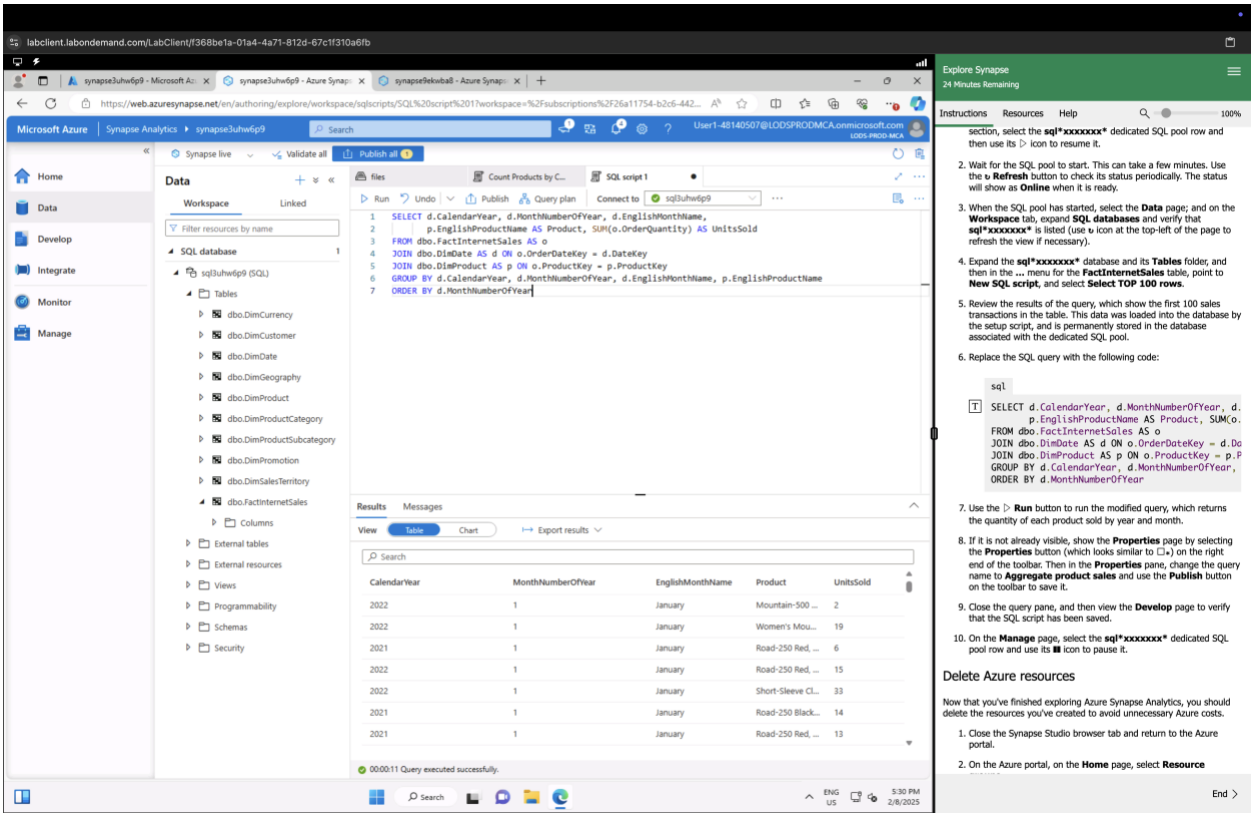
Using Spark pool to analyze data

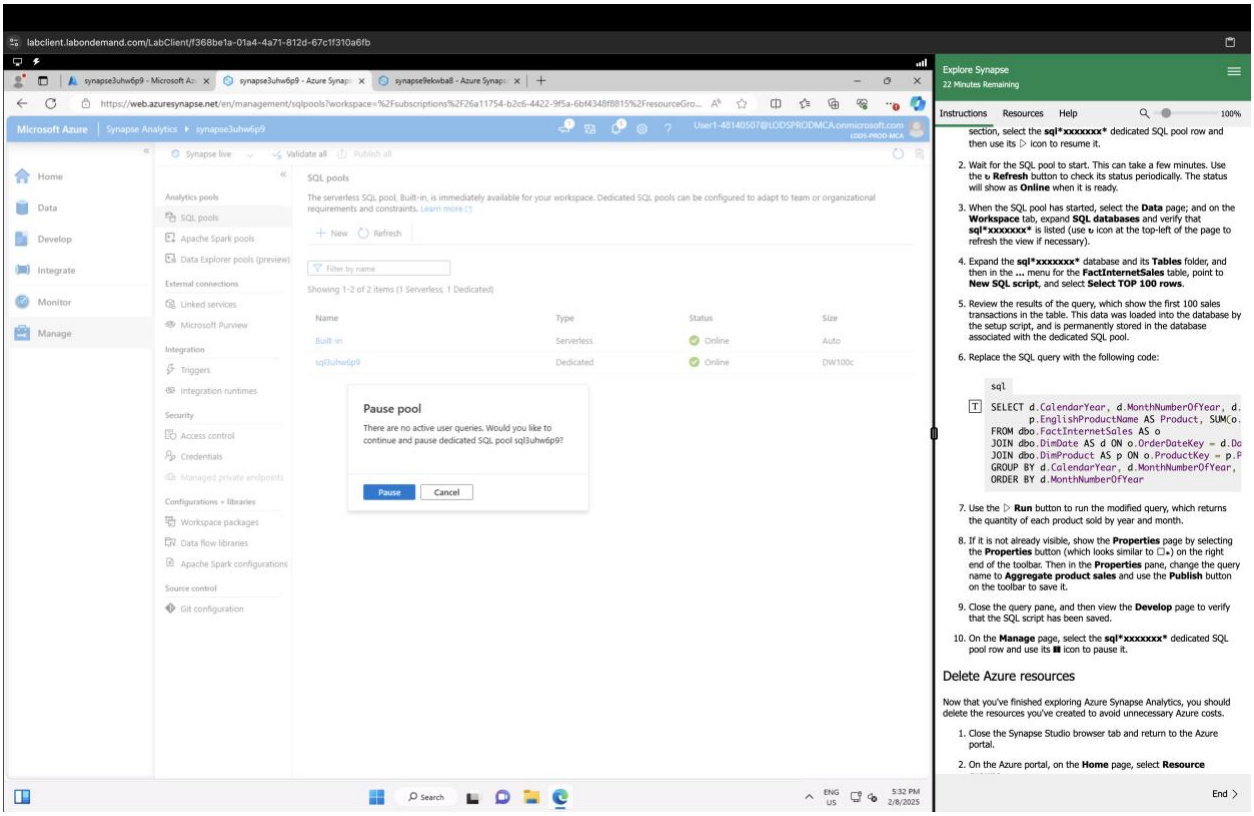


Using the Chart view option

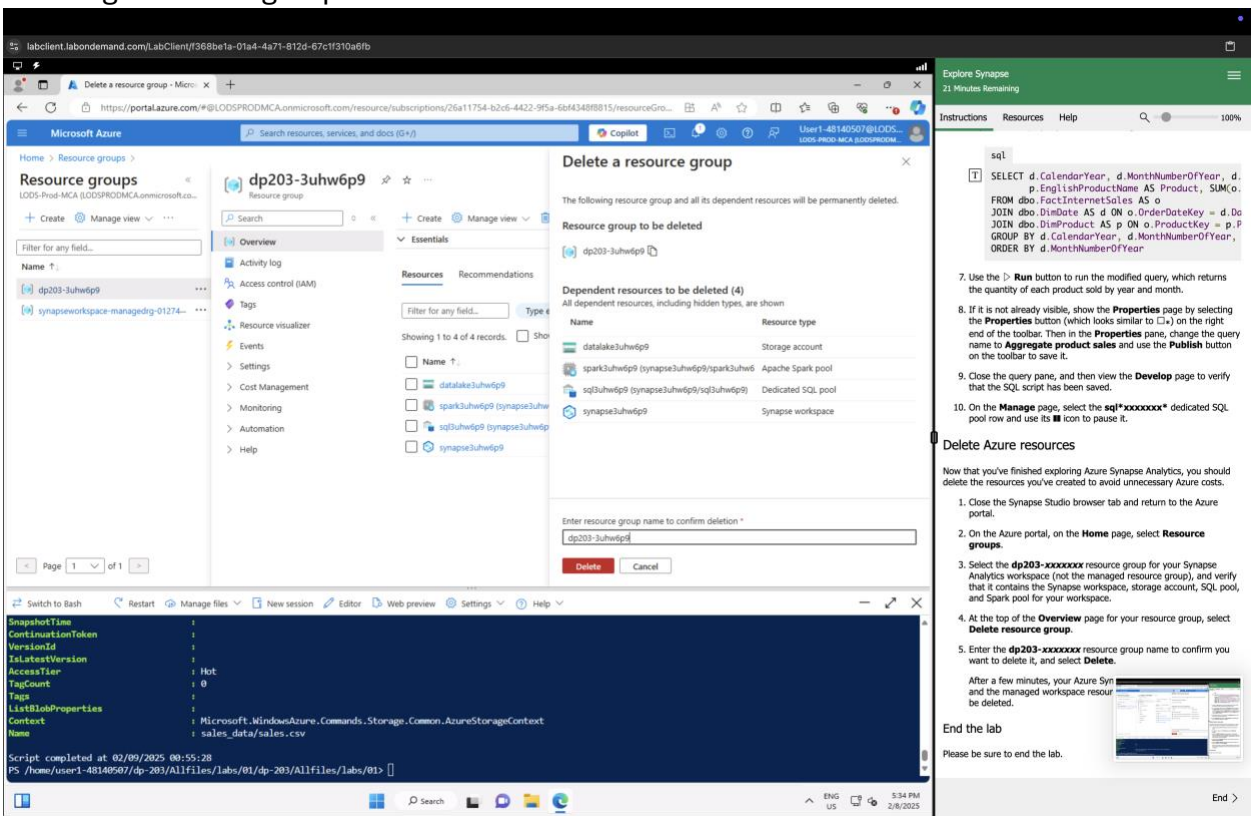





labclient.labondemand.com/LabClient/f368be1a-01a4-4a71-812d-67c1f310a6fb
Explore Synapse
24 Minutes Remaining
Instructions
Resources
Help
100%
section, select the sql*xxxxxxx* dedicated SQL pool row and then use its ▷ icon to resume it.
2. Wait for the SQL pool to start. This can take a few minutes. Use the ↻ Refresh button to check its status periodically. The status will show as Online when it is ready.
3. When the SQL pool has started, select the Data page; and on the Workspace tab, expand SQL databases and verify that sql*xxxxxxx* is listed (use ↻ icon at the top-left of the page to refresh the view if necessary).
4. Expand the sql*xxxxxxx* database and its Tables folder, and then in the ... menu for the FactInternetSales table, point to New SQL script, and select Select TOP 100 rows.
5. Review the results of the query, which show the first 100 sales transactions in the table. This data was loaded into the database by the setup script, and is permanently stored in the database associated with the dedicated SQL pool.
6. Replace the SQL query with the following code:
sql
SELECT d.CalendarYear, d.MonthNumberOfYear, d.
p.EnglishProductName AS Product, SUM(o.
FROM dbo.FactInternetSales AS o
JOIN dbo.DimDate AS d ON o.OrderDateKey = d.Da
JOIN dbo.DimProduct AS p ON o.ProductKey = p.P
GROUP BY d.CalendarYear, d.MonthNumberOfYear,
ORDER BY d.MonthNumberOfYear
7. Use the ▷ Run button to run the modified query, which returns the quantity of each product sold by year and month.
8. If it is not already visible, show the Properties page by selecting the Properties button (which looks similar to □•) on the right end of the toolbar. Then in the Properties pane, change the query name to Aggregate product sales and use the Publish button on the toolbar to save it.
9. Close the query pane, and then view the Develop page to verify that the SQL script has been saved.
10. On the Manage page, select the sql*xxxxxxx* dedicated SQL pool row and use its ■ icon to pause it.
Delete Azure resources
Now that 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
End


labclient.labondemand.com/LabClient/f368be1a-01a4-4a71-812d-67c1f310a6fb
Explore Synapse
22 Minutes Remaining
Instructions
Resources
Help
100%
Pause pool
There are no active user queries. Would you like to continue and pause dedicated SQL pool sql3uhw6p9?
Pause
Cancel
section, select the sql*xxxxxxx* dedicated SQL pool row and then use its ▷ icon to resume it.
2. Wait for the SQL pool to start. This can take a few minutes. Use the ↻ Refresh button to check its status periodically. The status will show as Online when it is ready.
3. When the SQL pool has started, select the Data page; and on the Workspace tab, expand SQL databases and verify that sql*xxxxxxx* is listed (use ↻ icon at the top-left of the page to refresh the view if necessary).
4. Expand the sql*xxxxxxx* database and its Tables folder, and then in the ... menu for the FactInternetSales table, point to New SQL script, and select Select TOP 100 rows.
5. Review the results of the query, which show the first 100 sales transactions in the table. This data was loaded into the database by the setup script, and is permanently stored in the database associated with the dedicated SQL pool.
6. Replace the SQL query with the following code:
sql
SELECT d.CalendarYear, d.MonthNumberOfYear, d.
p.EnglishProductName AS Product, SUM(o.
FROM dbo.FactInternetSales AS o
JOIN dbo.DimDate AS d ON o.OrderDateKey = d.Da
JOIN dbo.DimProduct AS p ON o.ProductKey = p.P
GROUP BY d.CalendarYear, d.MonthNumberOfYear,
ORDER BY d.MonthNumberOfYear
7. Use the ▷ Run button to run the modified query, which returns the quantity of each product sold by year and month.
8. If it is not already visible, show the Properties page by selecting the Properties button (which looks similar to □•) on the right end of the toolbar. Then in the Properties pane, change the query name to Aggregate product sales and use the Publish button on the toolbar to save it.
9. Close the query pane, and then view the Develop page to verify that the SQL script has been saved.
10. On the Manage page, select the sql*xxxxxxx* dedicated SQL pool row and use its ■ icon to pause it.
Delete Azure resources
Now that 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
End

Deleting Resource group:



Conclusion: In this lab, I got hands-on experience with Microsoft Azure Synapse and its data analytics features. I started by connecting to PowerShell and setting up a resource group. Then, I ingested data into the Synapse workspace and explored it using SQL queries. I also used the Chart view to visualize insights and worked with Spark pools to analyze large datasets more efficiently. Finally, I wrapped up by deleting the resource group, giving me a full cycle understanding of how to manage and analyze data in Synapse. This lab provided hands-on experience with Azure Synapse's data ingestion, querying, and visualization capabilities.