



Loading data from Pipelines – Storage Account

In Azure Synapse Analytics, pipelines are a fundamental concept used in the development and orchestration of data workflows. A pipeline is a logical grouping of activities that perform a series of tasks to move, transform, or analyze data. Azure Synapse provides a robust set of tools for building, scheduling, and monitoring pipelines, allowing organizations to automate and streamline their data integration, ETL (Extract, Transform, Load), and analytics processes.

Here are key components and features of pipelines in Azure Synapse:

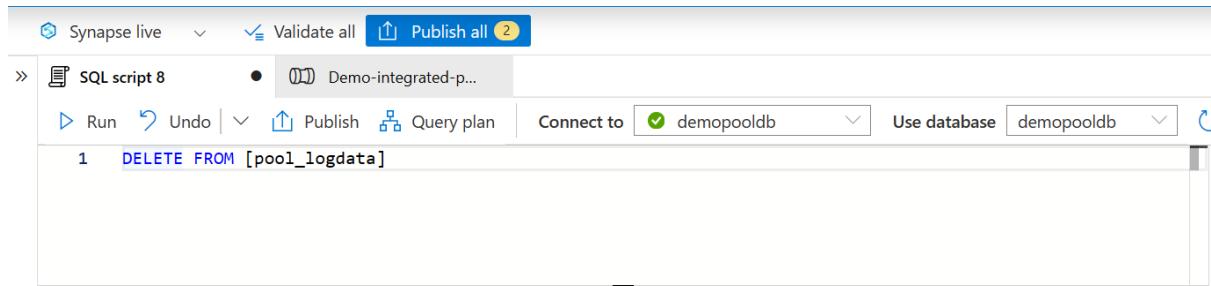
1. **Activities:** Activities are the building blocks of pipelines. Each activity represents a task or operation that performs a specific action on data, such as copying data, transforming data using SQL or Spark, running a machine learning model, or executing a custom script.
2. **Connectors:** Azure Synapse provides a wide range of connectors to interact with various data sources and destinations, including Azure Blob Storage, Azure Data Lake Storage, Azure SQL Database, Azure Cosmos DB, and more. Connectors simplify data integration by providing pre-built connectivity to different data platforms.
3. **Data Movement and Transformation:** Pipelines enable data movement and transformation operations to extract data from source systems, transform it using SQL or Spark transformations, and load it into target systems. This enables organizations to implement ETL processes for data integration and data preparation tasks.
4. **Monitoring and Logging:** Azure Synapse provides monitoring and logging capabilities to track the execution of pipelines in real-time. You can monitor pipeline runs, view execution statistics, and diagnose issues using built-in monitoring tools and logs, enabling proactive management of data workflows.
5. **Triggers:** Triggers define when a pipeline should be executed. Azure Synapse supports various trigger types, including schedule-based triggers, event-based triggers, and manual triggers. Triggers allow you to automate the execution of pipelines based on predefined schedules or events, ensuring timely data processing and analysis.
6. **Parameters and Variables:** Pipelines can utilize parameters and variables to dynamically configure and control pipeline behavior at runtime. Parameters enable you to pass inputs to pipelines, while variables allow you to store and manipulate data within pipelines, enhancing flexibility and reusability.
7. **Integration with Azure Services:** Azure Synapse integrates seamlessly with other Azure services such as Azure Data Factory, Azure Databricks, Azure Machine Learning, and Azure Monitor. This enables end-to-end data workflows, advanced analytics, and operational insights across the Azure ecosystem.



To begin with the Lab:

1. In this lab we are going to see how we copy data from our Data Lake Gen 2 storage account onto Azure Synapse using the pipeline feature that's available in Synapse itself.

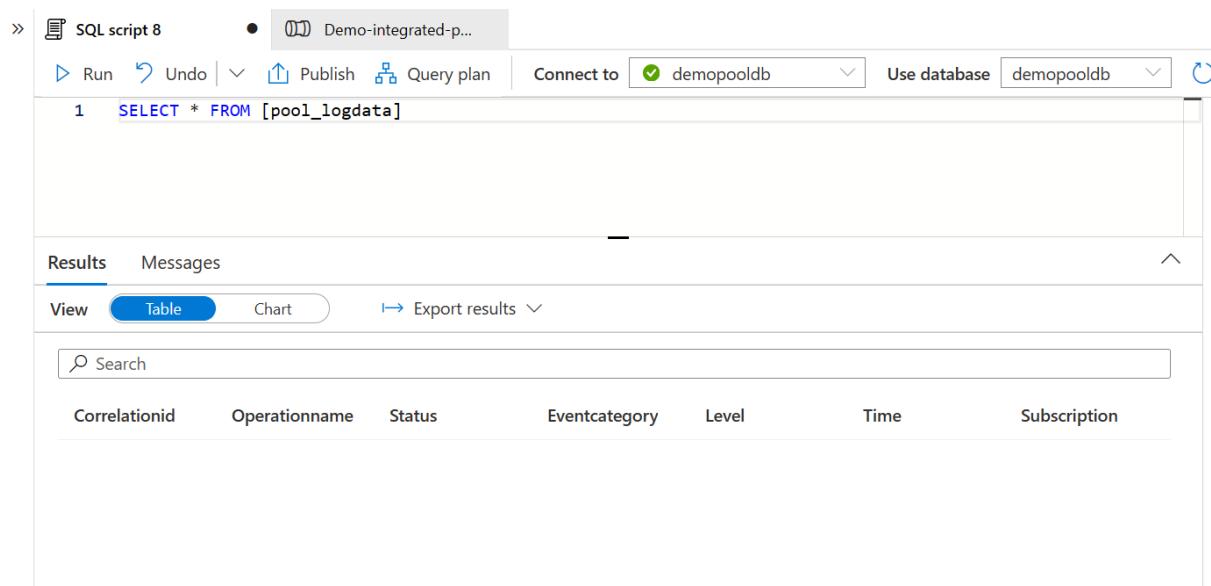
2. First you have to create a new SQL Script.
3. Then you are going to delete the data in your dedicated SQL Pool.



The screenshot shows the Azure Synapse Studio interface. At the top, there are buttons for 'Synapse live', 'Validate all', and 'Publish all'. Below that, a tab bar shows 'SQL script 8' and 'Demo-integrated-p...'. The main area contains a code editor with the following content:

```
1  DELETE FROM [pool_logdata]
```

4. Below you can see that there is no data available.



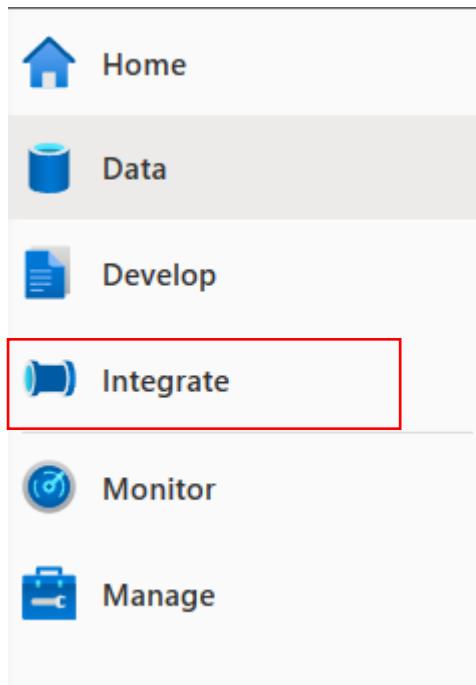
The screenshot shows the Azure Synapse Studio interface. At the top, there are buttons for 'Run', 'Undo', 'Publish', 'Query plan', 'Connect to', and 'Use database'. Below that, a tab bar shows 'SQL script 8' and 'Demo-integrated-p...'. The main area contains a code editor with the following content:

```
1  SELECT * FROM [pool_logdata]
```

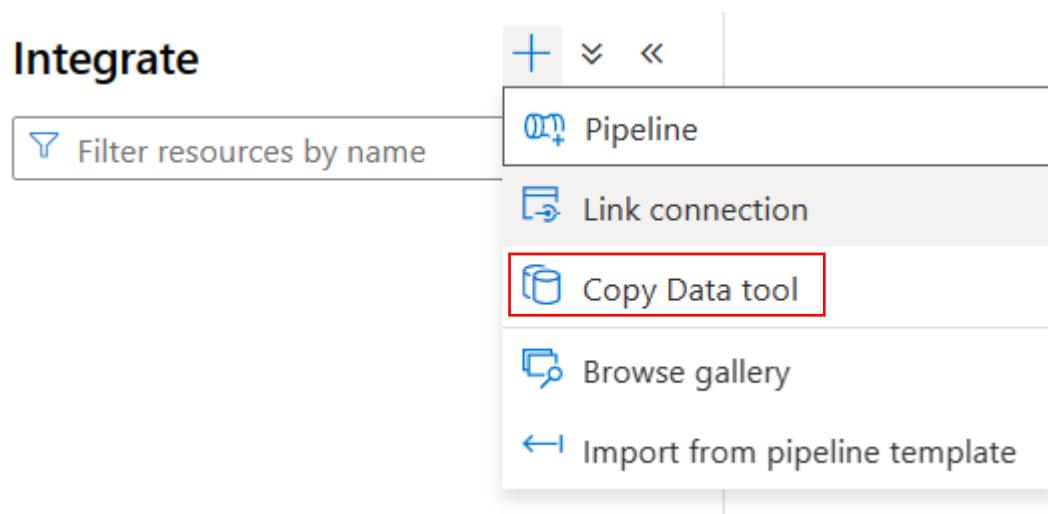
Below the code editor is a results pane. It has tabs for 'Results' and 'Messages', with 'Results' selected. It also has a 'View' dropdown set to 'Table' and a search bar. The results table has the following columns:

Correlationid	Operationname	Status	Eventcategory	Level	Time	Subscription

5. Now from the left pane you can see the Integrate section, navigate towards it.



- Now in Integrate if you click on the plus icon then you will have some options, from these options you must choose the Copy Data tool.

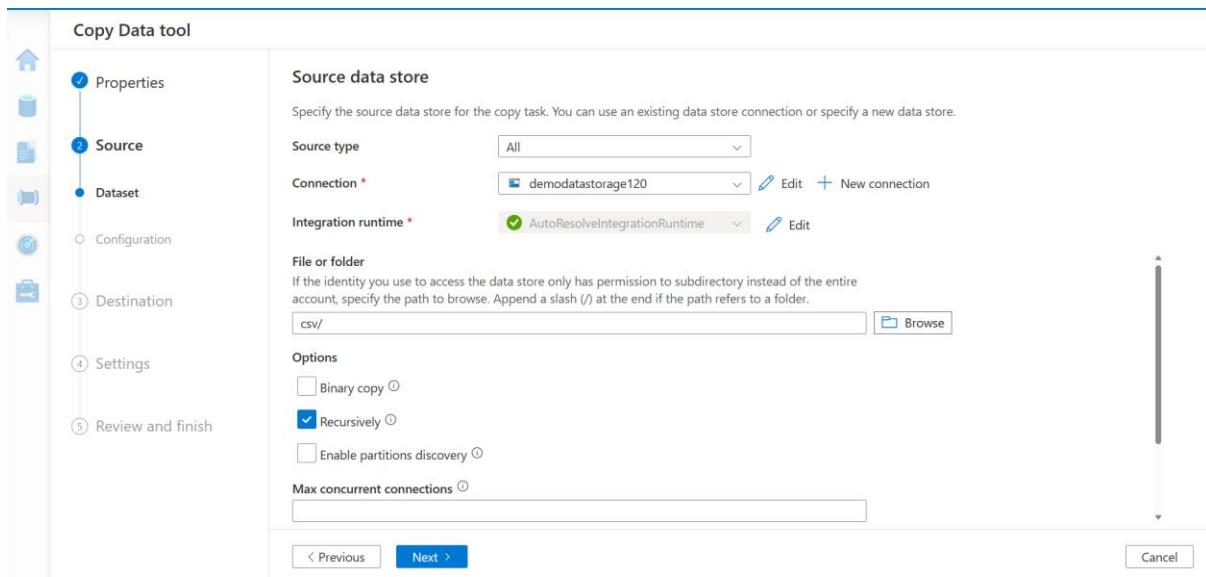


- Once you have clicked on it you will be on this kind of wizard.

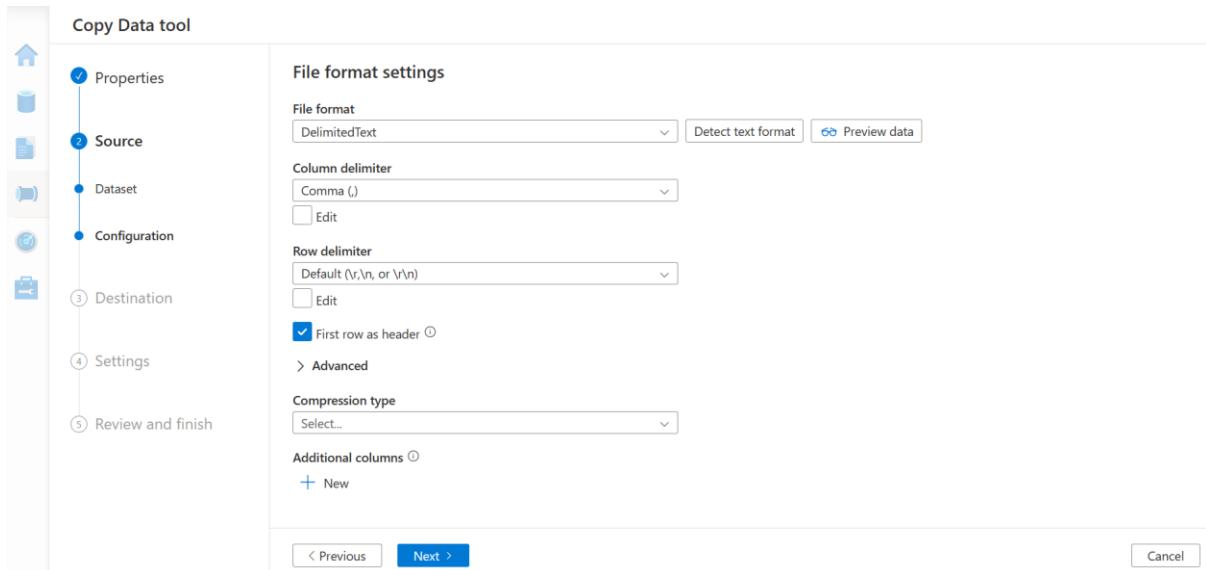
- Now you must click on next and move to step 2.

The screenshot shows the 'Copy Data tool' wizard. On the left is a navigation sidebar with five steps: 'Properties' (selected), 'Source', 'Destination', 'Settings', and 'Review and finish'. The main area is titled 'Properties' and contains instructions: 'Use Copy Data Tool to perform a one-time or scheduled data load from 90+ data sources. Follow the wizard experience to specify your data loading settings, and let the Copy Data Tool generate the artifacts for you, including pipelines, datasets, and linked services.' Below this is a 'Task type' section with two options: 'Built-in copy task' (selected) and 'Metadata-driven copy task'. The 'Built-in copy task' section includes a note: 'You will get single pipeline to quickly copy objects from data source store to destination in a very intuitive manner.' Underneath is a 'Task cadence or task schedule' section with a radio button for 'Run once now' (selected).

- Here in the connection, you must choose your storage account and then in the file or folder option choose the CSV container. Then just click on next.



10. Keep everything to default on the next page and move forward.



11. Now on the next page you must choose the destination. For that choose your connection as your Pool database and then for file storage click on use existing.

12. So, here you have to choose that existing table in which you just deleted all of your data at the start of this lab. Then click on next.

Copy Data tool

The screenshot shows the 'Destination' step in the Copy Data tool. On the left, a vertical navigation bar lists steps: Properties (checkmark), Source (checkmark), Destination (highlighted with a blue circle), Dataset (blue dot), Configuration (radio button), Settings (radio button), and Review and finish (radio button). The main panel is titled 'Destination data store' with the sub-instruction: 'Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.' It shows 'Destination type' set to 'All' and 'Connection *' set to 'demopooldb'. Below this, a table maps 'Source' (AzureBlobFSFile) to 'Target' (AzureBlobFSFile). A note '(auto-create)' is shown next to the target table. A link 'Use existing table' is also present.

13. On the next page it is doing the table mapping click on next.

Copy Data tool

The screenshot shows the 'Column mapping' step in the Copy Data tool. The left navigation bar shows the 'Source' step is also checked. The main panel is titled 'Column mapping' with the sub-instruction: 'Choose how source and destination columns are mapped'. It shows 'Table mappings (1)' with a single entry: 'Source' (Azure Data Lake Storage Gen2 file) and 'Target' (dbo.pool_logdata_parquet). Below this is a table titled 'Column mappings' with the header 'Type conversion settings'. It includes buttons for '+ New mapping', 'Clear', 'Reset', and 'Delete'. The table lists seven column pairs:

Source	Type	Destination
Correlation id	abc String	Correlationid
Operation name	abc String	Operationname
Status	abc String	Status
Event category	abc String	Eventcategory
Level	abc String	Level
Time	abc String	Time
Subscription	abc String	Subscription

At the bottom are navigation buttons '< Previous' and 'Next >'.

14. Now you must give it the name of your choice and then in the copy method choose Bulk insert.

15. After moving to the review page click on finish.

Copy Data tool

<ul style="list-style-type: none"><input checked="" type="checkbox"/> Properties<input checked="" type="checkbox"/> Source<input checked="" type="checkbox"/> Destination<input checked="" type="checkbox"/> Settings<input type="checkbox"/> Review and finish	<p>Settings</p> <p>Enter name and description for the copy data task, more options for data movement</p> <p>Task name * <input type="text" value="Demo-integrated-pipeline"/></p> <p>Task description <input type="text"/></p> <p>Fault tolerance <input type="text"/></p> <p>Enable logging <input type="checkbox"/></p> <p>Enable staging <input type="checkbox"/></p> <p>Advanced</p> <p>Copy method <input type="radio"/> Copy command <input type="radio"/> PolyBase <input checked="" type="radio"/> Bulk insert <input type="radio"/> Upsert</p> <p>Bulk insert table lock <input type="radio"/> Yes <input checked="" type="radio"/> No</p>
---	--

16. Below you can see that it has started the deployment.

<ul style="list-style-type: none"><input checked="" type="checkbox"/> Properties<input checked="" type="checkbox"/> Source<input checked="" type="checkbox"/> Destination<input checked="" type="checkbox"/> Settings<input checked="" type="checkbox"/> Review and finish<input type="checkbox"/> Review<input type="checkbox"/> Deployment	<p>Copy Data tool</p> <p>Azure Data Lake Storage Gen2 → Azure Synapse dedicated SQL pool</p> <p>Deploying ...</p> <table border="1"><thead><tr><th>Deployment step</th><th>Status</th></tr></thead><tbody><tr><td>Validating copy runtime environment</td><td>✓ Succeeded</td></tr><tr><td>> Creating datasets</td><td>⌚ In progress</td></tr><tr><td>> Creating pipelines</td><td>⌚ Pending</td></tr><tr><td>> Running pipelines</td><td>⌚ Pending</td></tr></tbody></table>	Deployment step	Status	Validating copy runtime environment	✓ Succeeded	> Creating datasets	⌚ In progress	> Creating pipelines	⌚ Pending	> Running pipelines	⌚ Pending
Deployment step	Status										
Validating copy runtime environment	✓ Succeeded										
> Creating datasets	⌚ In progress										
> Creating pipelines	⌚ Pending										
> Running pipelines	⌚ Pending										

17. Once your deployment is done click on finish. Now in the integrate section, you can see your pipeline.

Copy Data tool

The screenshot shows the 'Copy Data tool' interface. On the left, a vertical navigation bar lists steps: Properties, Source, Destination, Settings, Review and finish, Review, and Deployment. The 'Destination' step is checked. In the center, a diagram shows 'Azure Data Lake Storage Gen2' connected by an arrow to 'Azure Synapse dedicated SQL pool'. Below this, the text 'Deployment complete' is displayed. A table titled 'Deployment step' shows four rows: 'Validating copy runtime environment' (Status: Succeeded), 'Creating datasets' (Status: Succeeded), 'Creating pipelines' (Status: Succeeded), and 'Running pipelines' (Status: Succeeded). At the bottom, a message states: 'Datasets and pipelines have been created. You can now monitor and edit the copy pipelines or click finish to close Copy Data Tool.' Three buttons at the bottom are 'Finish' (highlighted in blue), 'Edit pipeline', and 'Monitor'.

The screenshot shows the Azure Synapse Analytics workspace. The left sidebar has icons for Home, Data, Develop, Integrate (which is selected and highlighted in grey), Monitor, and Manage. The main area is titled 'Synapse live' and 'Integrate'. It features a search bar 'Filter resources by name' and a section titled 'Pipelines' with one item: 'Demo-integrated-pipeline'. There are also 'Validate all' and 'Upload' buttons at the top right.

18. Now if you navigate to the Monitor section from the left pane and go to pipeline runs. Here you can see that your pipeline run was successful.

The screenshot shows the Azure Data Factory 'Monitor' interface. On the left, there's a navigation sidebar with icons for Home, Data, Develop, Integrate, Monitor (which is highlighted with a red box), and Manage. Under the 'Integrate' section, there's a 'Pipeline runs' item also highlighted with a red box. The main area is titled 'All pipeline runs > Demo-integrated-pipeline - Activity runs'. It shows a list of activity runs for a 'Copy data' activity named 'Copy_3t8'. The run status is 'Succeeded'. There are buttons for Rerun, Cancel, Refresh, Update pipeline, List (selected), and Gantt.

19. Now come back to your script and run the select command to check whether the data has been copied or not.

20. Below you can see that we have the data.

The screenshot shows the Azure Data Studio SQL script editor. The script tab contains a single query: 'SELECT * FROM [pool logdata]'. The results tab displays a table of log data with columns: Correlation id, Operation name, Status, Event category, Level, Time, and Subscription. The results show several entries, all of which are 'Succeeded'. At the bottom, a message indicates '00:00:04 Query executed successfully.'

Correlation id	Operation name	Status	Event category	Level	Time	Subscription
84fb2d66-c063...	'auditIfNotExist...	Succeeded	Policy	Informational	2023-04-17T15:...	6912d7a0-bc28...
2127d1c8-ba1f...	'audit' Policy ac...	Succeeded	Policy	Warning	2023-04-17T15:...	6912d7a0-bc28...
84fb2d66-c063...	'auditIfNotExist...	Started	Policy	Informational	2023-04-17T15:...	6912d7a0-bc28...
647accd7-c9f0-...	Delete Network...	Started	Administrative	Informational	2023-04-17T14:...	6912d7a0-bc28...
3536525e-4881...	'auditIfNotExist...	Succeeded	Policy	Informational	2023-04-17T14:...	6912d7a0-bc28...
3536525e-4881	'auditIfNotExist...	Succeeded	Policy	Informational	2023-04-17T14:...	6912d7a0-bc28...