

Azure Data Factory LAB

Prerequisites

- Azure SQL Data Warehouse: The data warehouse holds the data that's copied over from the SQL database. You created your Azure SQL Data Warehouse in previous lab.
- Azure SQL Database: This tutorial copies data from an Azure SQL database with Adventure Works LT sample data. We have created this DB for you however should it be unavailable, follow this link to create it in your environment: [Create an Azure SQL database](#).

SQL DB name: AcademySQLDatabase

SQL Server name: academysqlserver

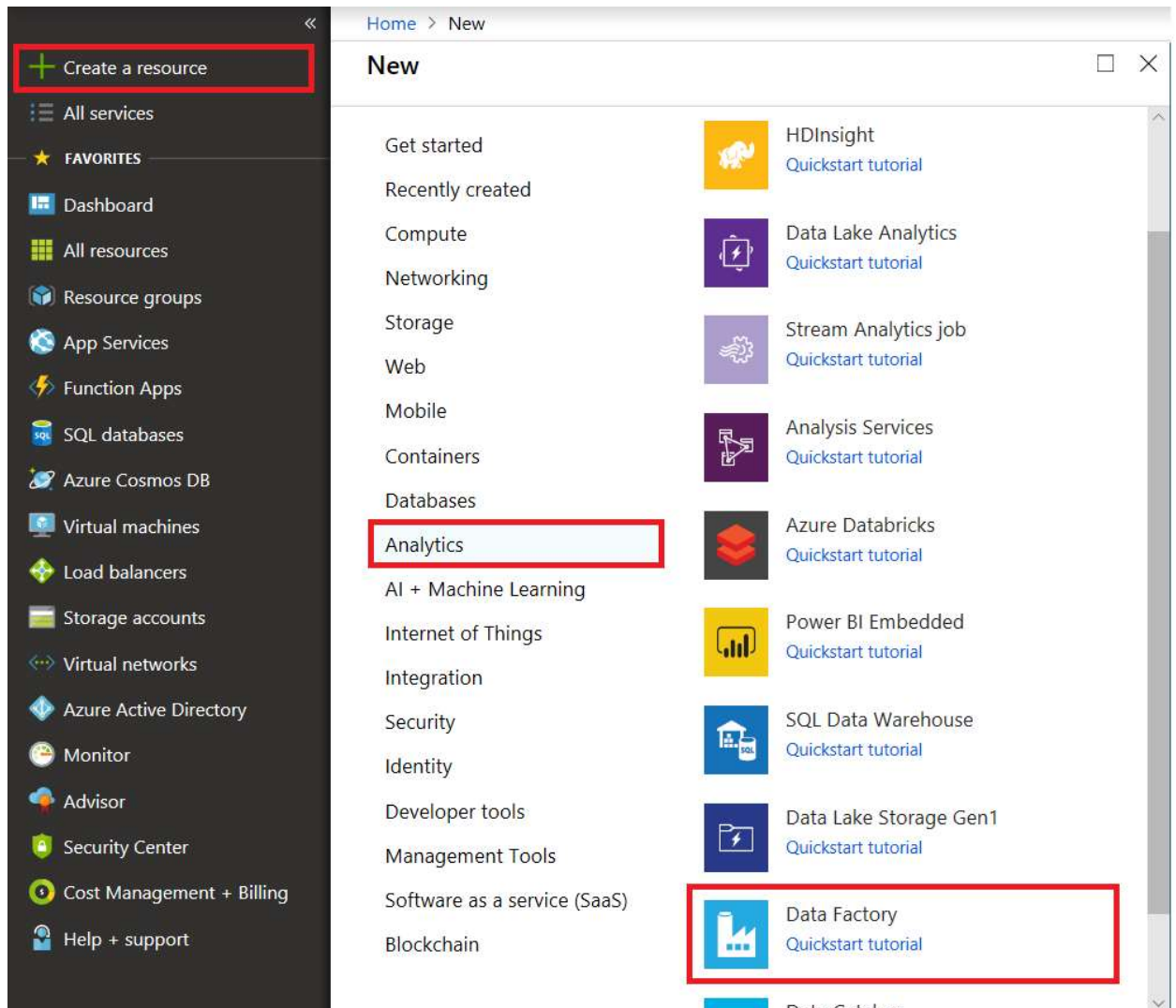
User name: academyuser

Password: Azure4Academy2019

- Azure storage account: Azure Storage is used as the *staging* blob in the bulk copy operation. If you don't have an Azure storage account, see the instructions in [Create a storage account](#).

Create a data factory

1. On the left menu, select **Create a resource > Data + Analytics > Data Factory**:



2. In the **New data factory** page, provide values for the fields that are shown in the following image:

New data factory ☐



* Name ⓘ
AzureAcademy-ADF ✓

* Subscription
Microsoft Azure Internal Consumption ▼

* Resource Group ⓘ
☐ Create new ☒ Use existing
Azure-Academy-2019 ▼

Version ⓘ
V2 ▼

* Location ⓘ
West Europe ▼

 Integrate with GIT source control (Azure DevOps GIT or GitHub) to do collaboration, source control, change tracking, change difference, continuous integration and deployment etc 

☐ Enable GIT ⓘ

Create Automation options

- **Name:** Enter a globally unique name for your Azure data factory. If you receive the error "Data factory name "AzureAcademy-ADF" is not available," enter a different name for the data factory. For example, you could use the name ***yournameADFTutorialDataFactory***. Try creating the data factory again. For the naming rules for Data Factory artifacts, see [Data Factory naming rules](#).
 - **Subscription:** Select your Azure subscription in which to create the data factory.
 - **Resource Group:** Select an existing resource group from the drop-down list, or select the **Create new** option and enter the name of a resource group. To learn about resource groups, see [Using resource groups to manage your Azure resources](#).
 - **Version:** Select **V2**.
 - **Location:** Select the location for the data factory. Only supported locations are displayed in the drop-down list. The data stores that are used by data factory can be in other locations and regions. These data stores include Azure Data Lake Store, Azure Storage, Azure SQL Database, and so on.
3. Select **Create**.
 4. After creation is complete, go to your data factory. You see the **Data Factory** home page as shown in the following image:



[Download PDF](#)
[Download PDF](#)

Data factory (V2)

Getting started

Quick start

West Europe

[Full Screen](#) [Print](#) [Download](#) [Share](#)

1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 2680, 26



Documentation

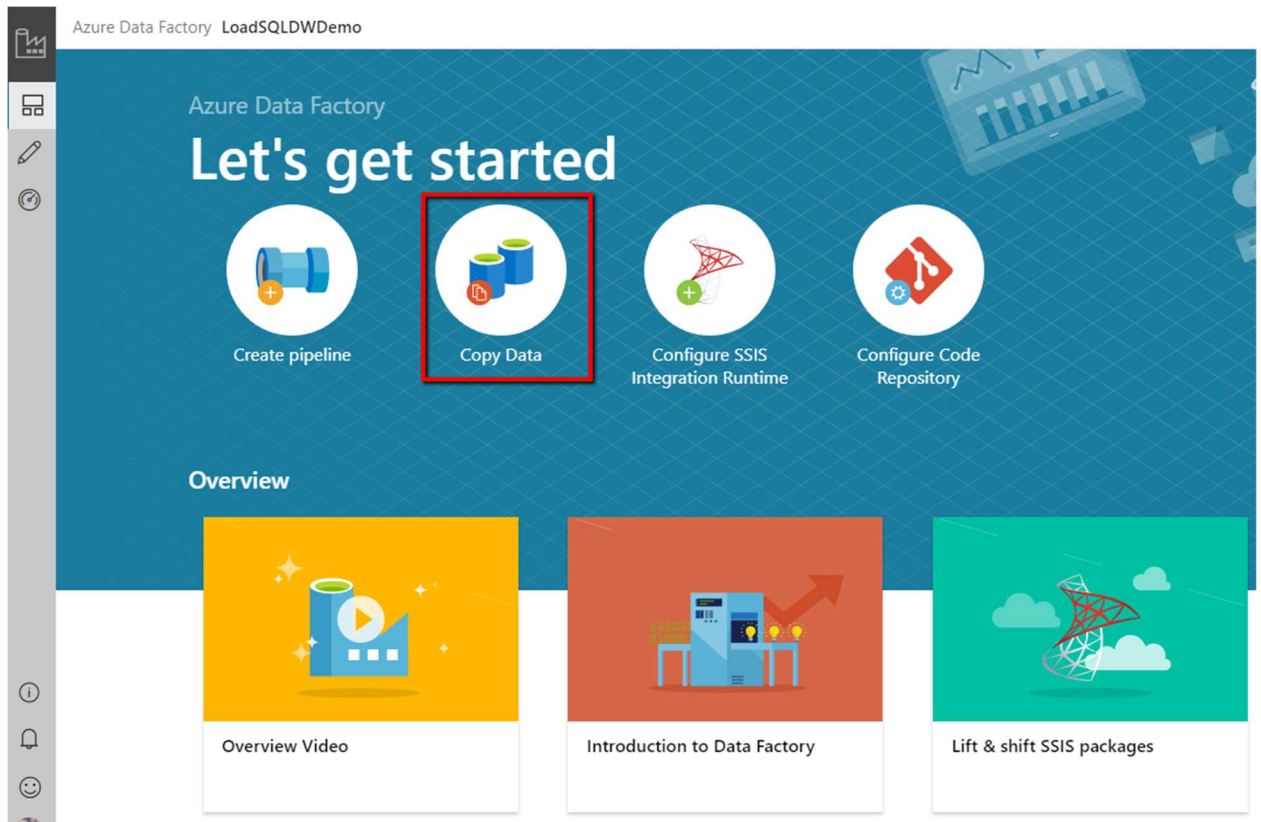


Author &
Monitor

DinelineRune

data into Azure SQL Data Warehouse

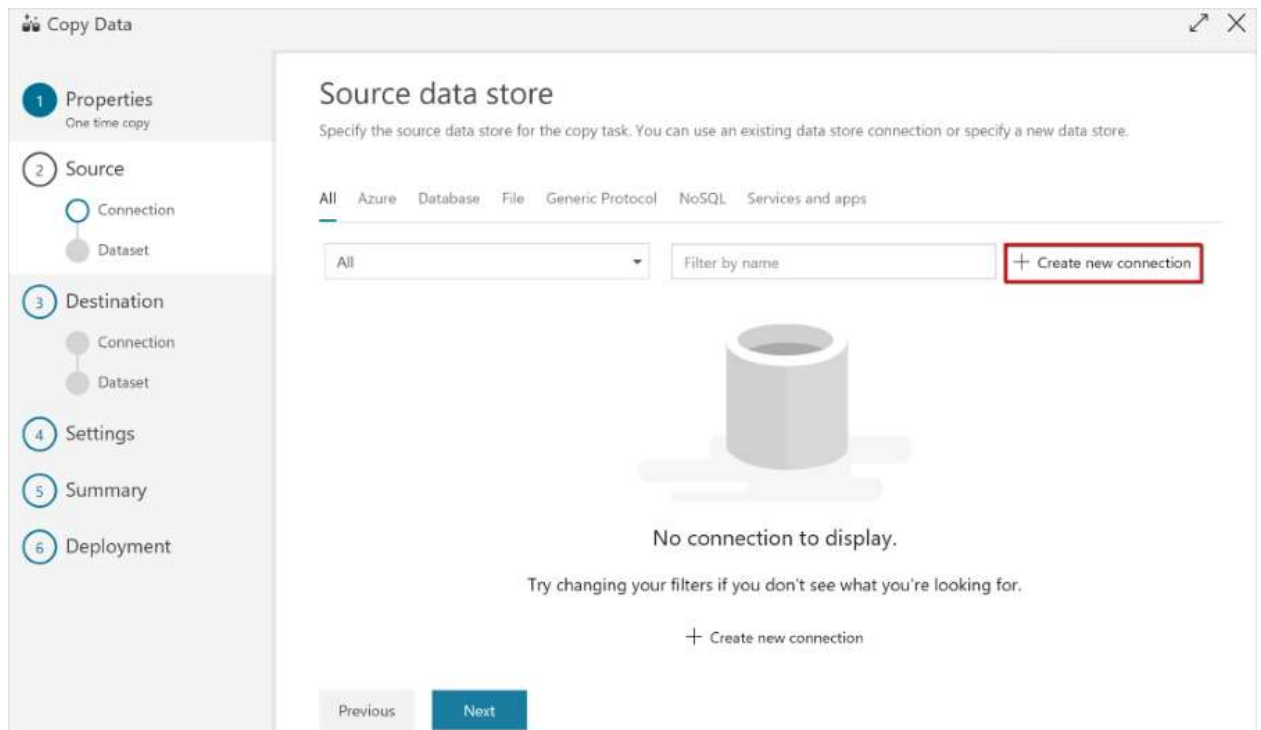
1. In the **Get started** page, select the **Copy Data** tile to launch the Copy Data tool:



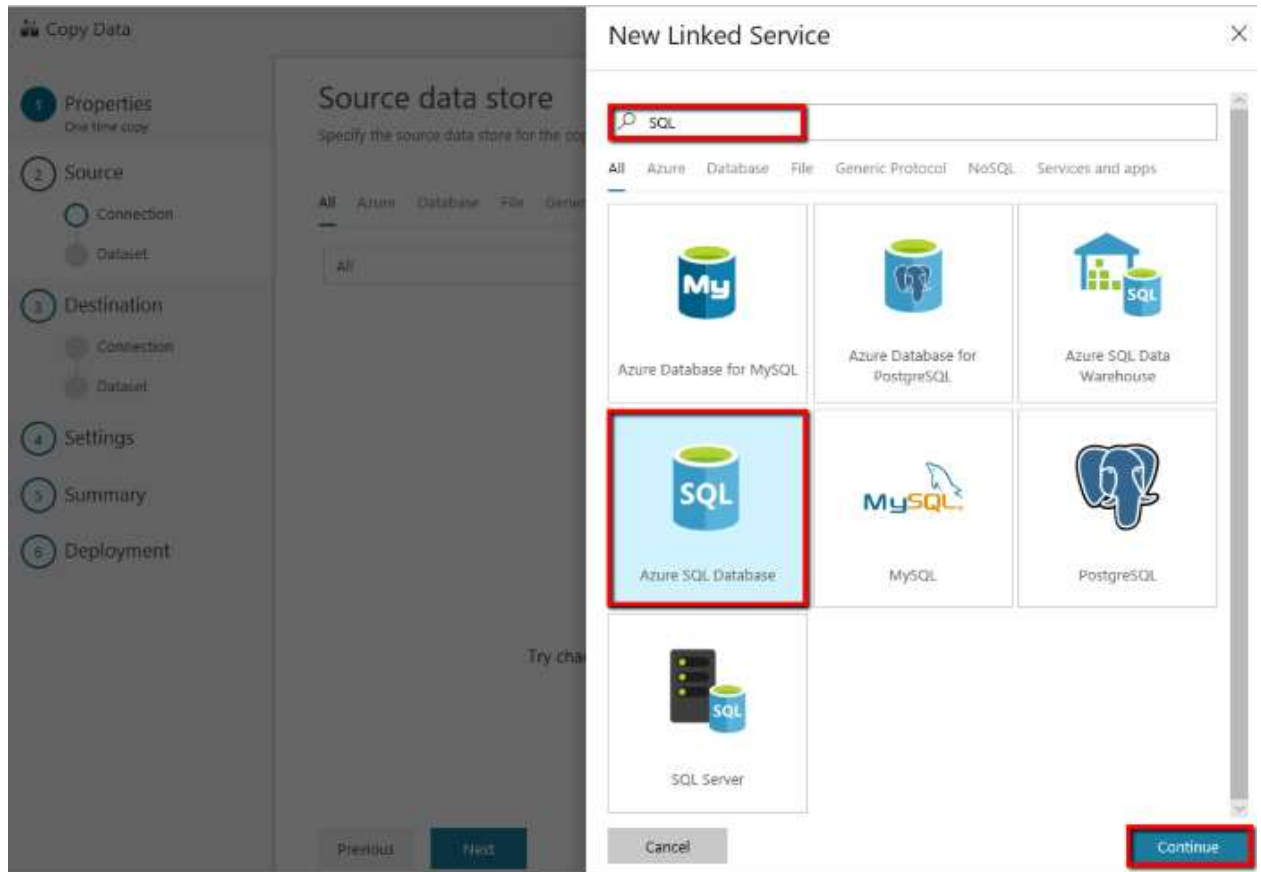
2. In the **Properties** page, specify **CopyFromSQLToSQLDW** for the **Task name** field, and select **Next**:

The screenshot shows the 'Copy Data' wizard interface. On the left, a sidebar contains a list of steps: 1 Properties, 2 Source, 3 Destination, 4 Settings, 5 Summary, and 6 Deployment. Step 2, 'Source', is currently selected. Within the 'Source' step, there are two sub-steps: 'Connection' and 'Dataset'. The main panel displays the 'Properties' configuration for the task. It includes a 'Task name' field with the value 'CopyFromSQLToSQLDW', a 'Task description' text area, and a 'Task cadence or Task schedule' section with two radio buttons: 'Run once now' (selected) and 'Run regularly on schedule'. At the bottom of the main panel, there are 'Previous' and 'Next' buttons. The 'Next' button is highlighted with a red rectangular border.

3. In the **Source data store** page, complete the following steps:
 - a. click + **Create new connection**:



b. Select **Azure SQL Database** from the gallery and select **Continue**. You can type "SQL" in the search box to filter the connectors.



c. In the **New Linked Service** page, select your server name and DB name from the dropdown list, and specify the username and password. If you don't have your own SQL Database, please connect to ours with credentials mentioned in the beginning of this document.

Click **Test connection** to validate the settings, then select **Finish**.

← New Linked Service (Azure SQL Database)

Name *

SQLDB_Academy

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Connection String

Azure Key Vault

Account selection method

☒ From Azure subscription

☐ Enter manually

Azure subscription

Microsoft Azure Internal Consumption (7f8d70a2-4d8b-41e2-b24c-a08046e098a9)

Server name *

academysqlserver

Database name *

AcademySQLDatabase

Authentication type *

SQL Authentication

User name *

academyuser

Password

Azure Key Vault

Password *

.....

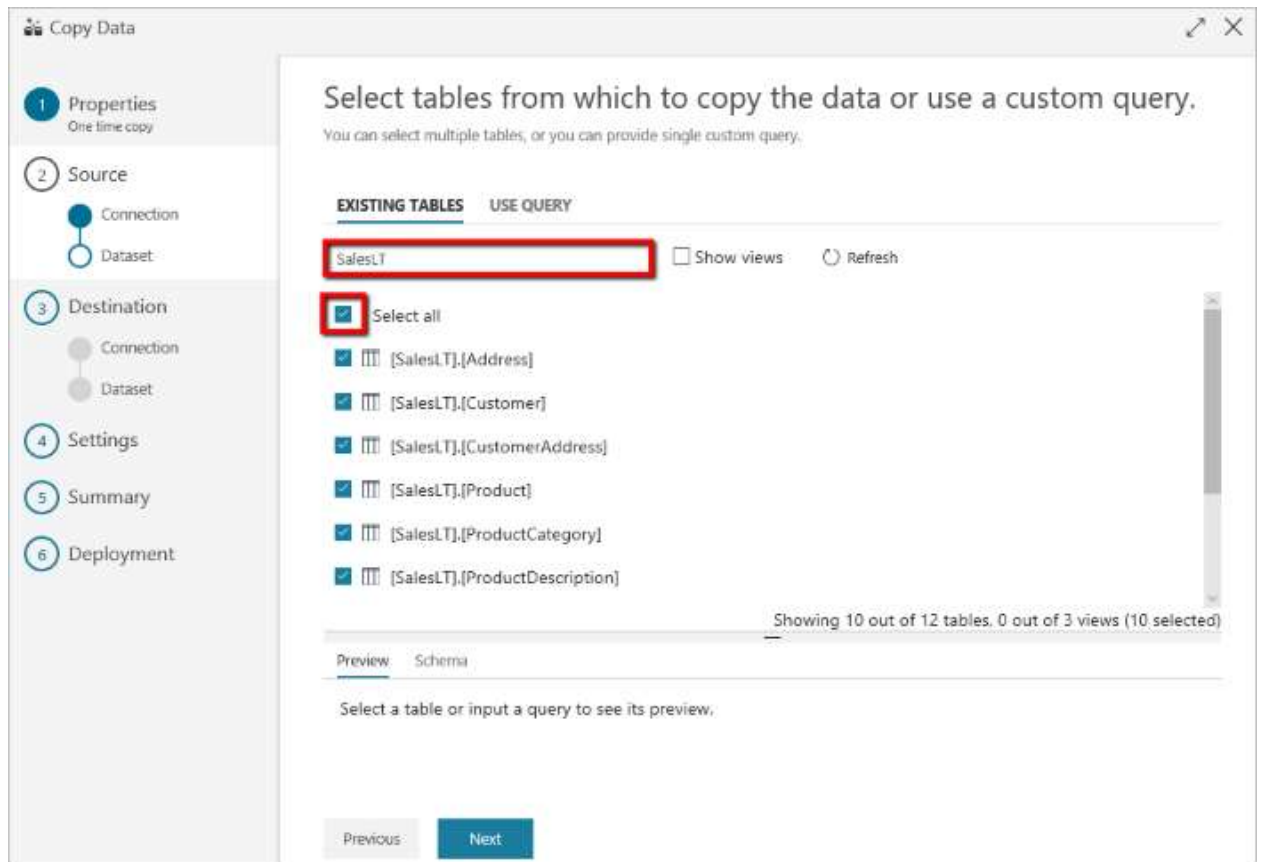
Additional connection properties

Your screen could look like this.

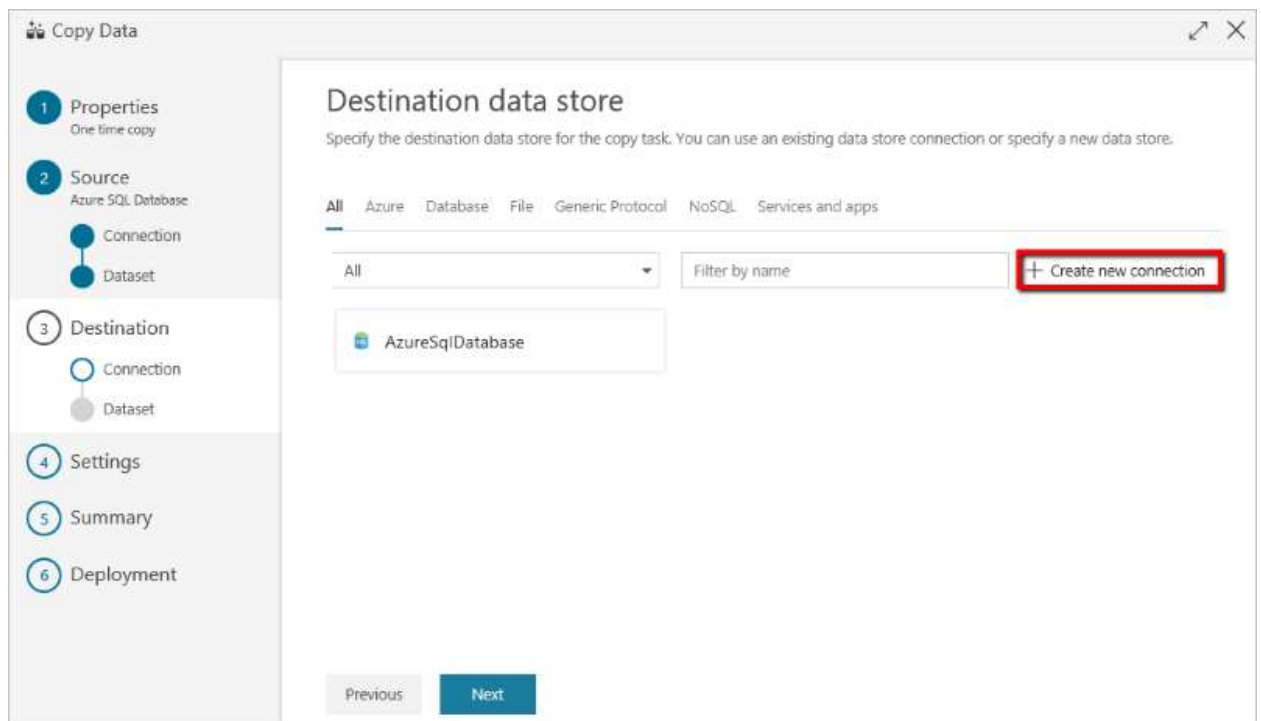
d. Select the newly created linked service as source, then click **Next**.

The screenshot shows the 'Copy Data' wizard in the Azure Data Studio interface. The left sidebar contains a navigation pane with six steps: 1 Properties (One time copy), 2 Source, 3 Destination, 4 Settings, 5 Summary, and 6 Deployment. Step 2, 'Source', is currently selected. Under 'Source', there are two options: 'Connection' (selected with a blue circle) and 'Dataset' (unselected with a grey circle). The main panel is titled 'Source data store' and contains the instruction: 'Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store.' Below this, there are tabs for 'All', 'Azure', 'Database', 'File', 'Generic Protocol', 'NoSQL', and 'Services and apps'. The 'All' tab is active. A dropdown menu shows 'All' and a search box labeled 'Filter by name' is next to it. A '+ Create new connection' button is on the right. A list of connections is displayed, with 'AzureSqlDatabase' highlighted in a blue box. At the bottom, there are 'Previous' and 'Next' buttons. The 'Next' button is highlighted with a red rectangle.

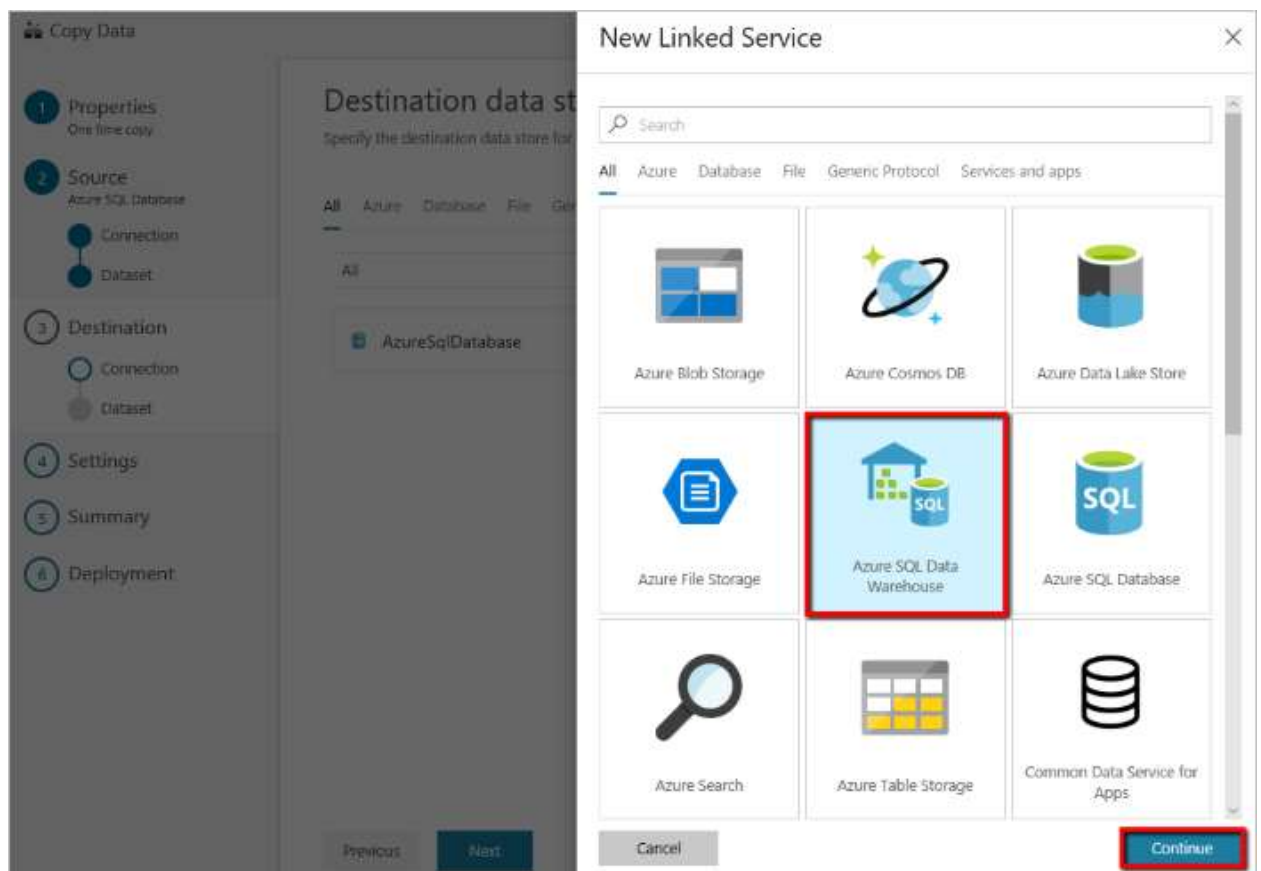
4. In the **Select tables from which to copy the data or use a custom query** page, enter **SalesLT** to filter the tables. Choose the **(Select all)** box to use all of the tables for the copy, and then select **Next**:



5. In the **Destination data store** page, complete the following steps:
 - a. Click + **Create new connection** to add a connection



b. Select **Azure SQL Data Warehouse** from the gallery, and select **Next**.



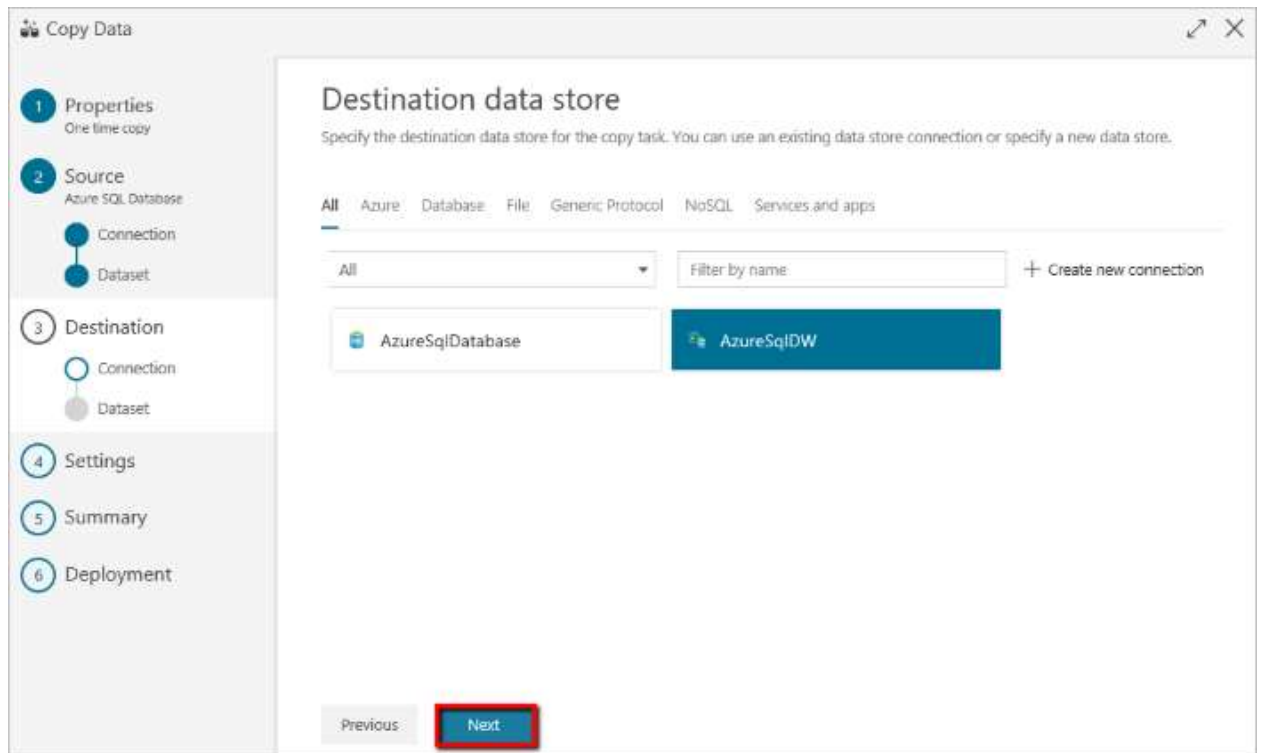
c. In the **New Linked Service** page, select your server name and DB name from the dropdown list, and specify the username and password. Click **Test connection** to validate the settings, then select **Finish**.

The screenshot shows the 'New Linked Service (Azure SQL Data Warehouse)' dialog box. The background is the 'Copy Data' tool's 'Destination data store' selection screen, which is dimmed. The dialog box has a title bar with a back arrow, the title 'New Linked Service (Azure SQL Data Warehouse)', and a close button. The fields and options are as follows:

- Name ***: Text input field containing 'AzureSqlDW'.
- Description**: Text input field.
- Connect via integration runtime ***: Dropdown menu showing 'AutoResolveIntegrationRuntime'.
- Connection String**: Tabbed interface with 'Connection String' (selected) and 'Azure Key Vault' tabs.
- Account selection method**: Dropdown menu showing 'From Azure subscription'.
- Azure subscription**: Dropdown menu showing 'Select all'.
- Server name ***: Dropdown menu showing '<your server name>'.
- Database name ***: Dropdown menu showing '<your database name>'.
- Authentication type ***: Dropdown menu showing 'SQL Authentication'.
- User name ***: Text input field showing '<your user name>'.
- Password ***: Password input field showing masked characters.

At the bottom right of the dialog, there is a green checkmark icon and the text 'Connection successful'. Below this, there are three buttons: 'Cancel', 'Test connection', and 'Finish'.

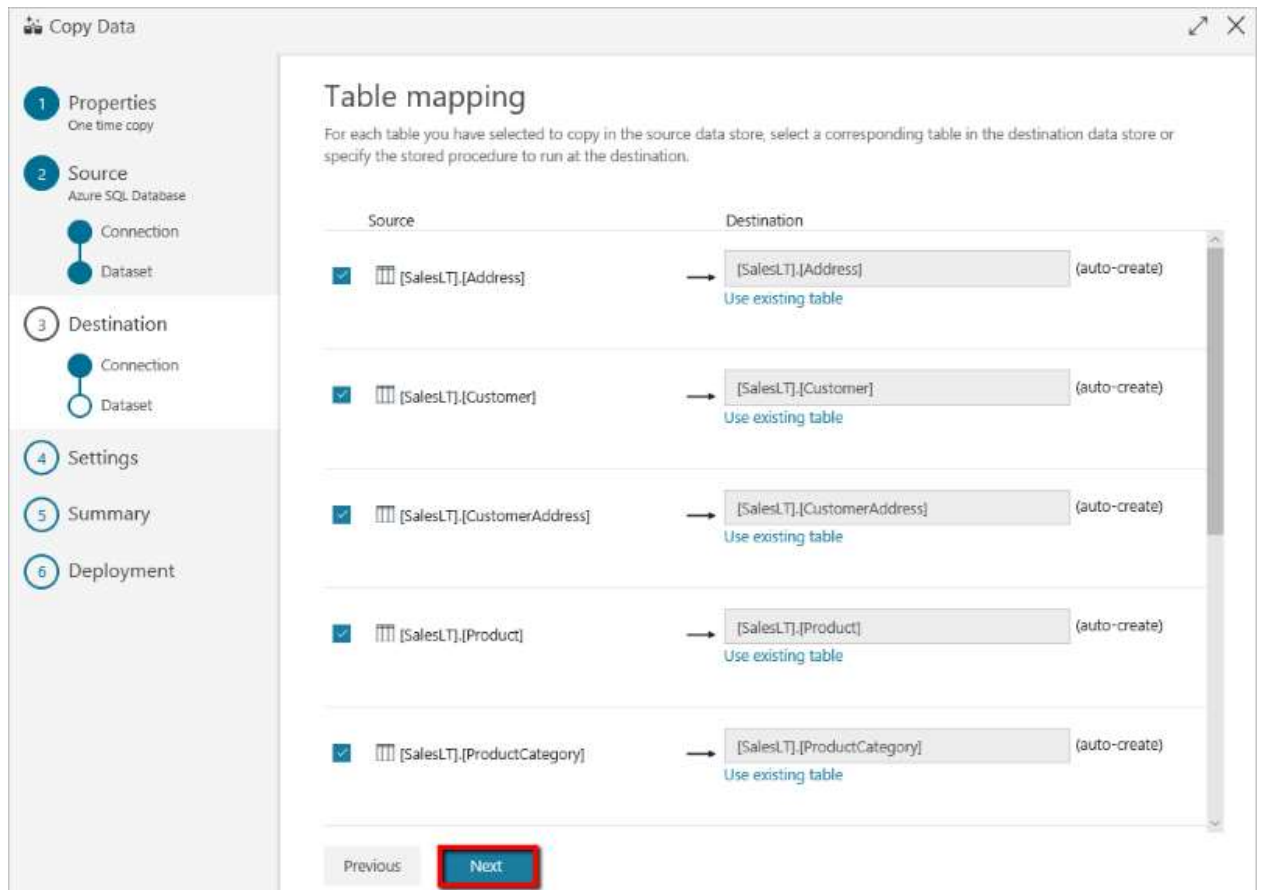
d. Select the newly created linked service as sink, then click **Next**.



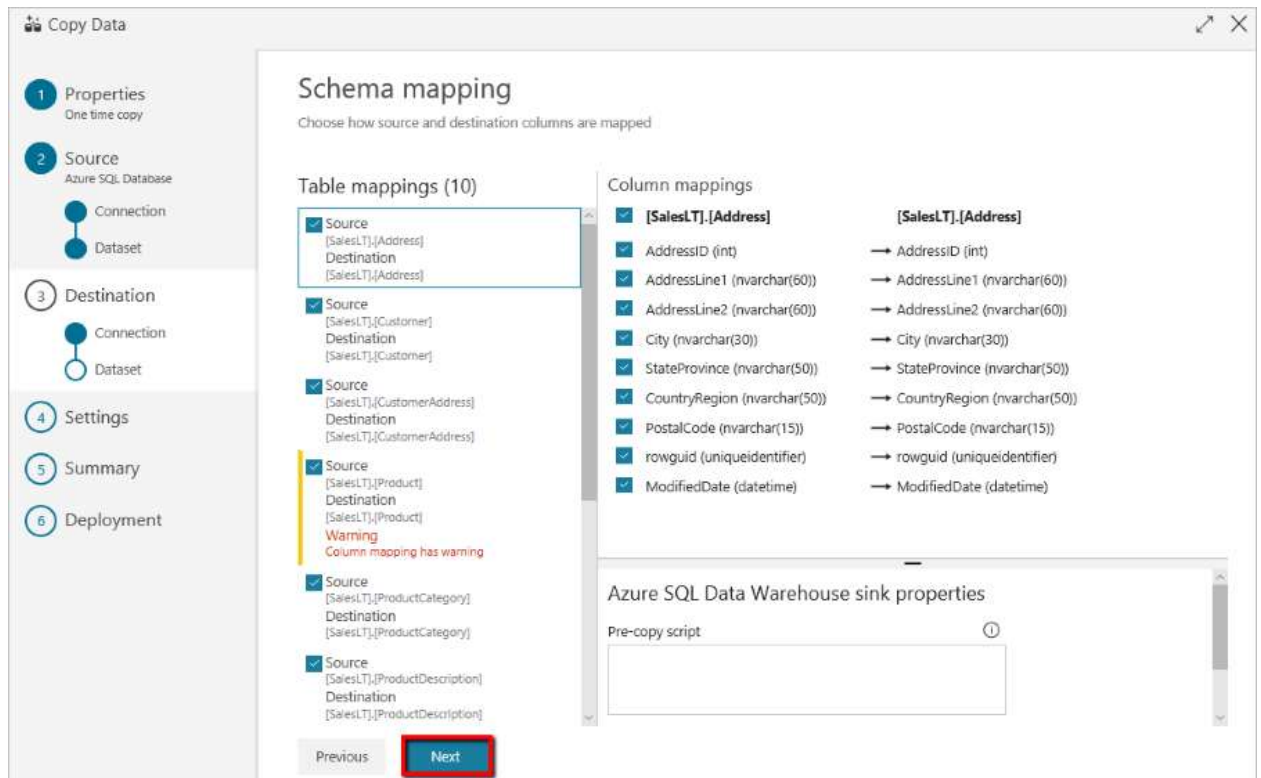
6. In the **Table mapping** page, review the content, and select **Next**. An intelligent table mapping displays. The source tables are mapped to the destination tables based on the table names. If a source table doesn't exist in the destination, Azure Data Factory creates a destination table with the same name by default. You can also map a source table to an existing destination table.

Note

Automatic table creation for the SQL Data Warehouse sink applies when SQL Server or Azure SQL Database is the source. If you copy data from another source data store, you need to pre-create the schema in the sink Azure SQL Data Warehouse before executing the data copy.



7. In the **Schema mapping** page, review the content, and select **Next**. The intelligent table mapping is based on the column name. If you let Data Factory automatically create the tables, data type conversion can occur when there are incompatibilities between the source and destination stores. If there's an unsupported data type conversion between the source and destination column, you see an error message next to the corresponding table.



8. In the **Settings** page, complete the following steps:

a. In **Staging settings** section, click **+ New** to new a staging storage. The storage is used for staging the data before it loads into SQL Data Warehouse by using PolyBase. After the copy is complete, the interim data in Azure Storage is automatically cleaned up.

Copy Data

1 Properties
One time copy

2 Source
Azure SQL Database
Connection
Dataset

3 Destination
Azure SQL Data Warehouse
Connection
Dataset

4 **Settings**

5 Summary

6 Deployment

Settings

More options for data movement

Fault tolerance settings

Fault tolerance: Abort activity on first incompatible row ⓘ

Performance settings

☒ Enable Staging ⓘ

Staging Settings

Staging Account Linked Service: Select... ⓘ **+ New**

Storage Path: ⓘ **Browse**

☐ Enable Compression ⓘ

Advanced settings

☒ Allow polybase ⓘ

Reject type: Value

Reject value: 0

☒ Use Type default

Previous **Next**

b. In the **New Linked Service** page, select your storage account, and select **Finish**.

The screenshot shows the 'New Linked Service (Azure Storage)' dialog box. The left sidebar contains a navigation pane with steps: 1 Properties (One time copy), 2 Source (Azure SQL Database), 3 Destination (Azure SQL Data Warehouse), 4 Settings, 5 Summary, and 6 Deployment. The 'Settings' step is selected. The main area shows configuration options for the linked service. The 'Name' field is 'AzureStorage'. The 'Description' field is empty. The 'Connect via integration runtime' dropdown is set to 'AutoResolveIntegrationRuntime'. The 'Authentication method' dropdown is set to 'Use account key'. Below this, there are two tabs: 'Connection String' (selected) and 'Azure Key Vault'. The 'Account selection method' dropdown is set to 'From Azure subscription'. The 'Azure subscription' dropdown is set to 'Select all'. The 'Storage account name' field is highlighted with a red box and contains the placeholder text '<your storage name>'. Below this, there is an 'Advanced' section with a plus icon and a question mark. At the bottom, there are three buttons: 'Cancel', 'Test connection', and 'Finish' (highlighted with a red box).

Copy Data

1 Properties
One time copy

2 Source
Azure SQL Database

3 Destination
Azure SQL Data Warehouse

4 Settings

5 Summary

6 Deployment

Settings

More options

Fault tolerance

Performance

Enable staging account

Staging account

Storage pattern

Enable staging account

Advanced settings

Allow parallel copy

Reject type

Reject value

Use type default

Previous

New Linked Service (Azure Storage)

Name *

AzureStorage

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Authentication method

Use account key

Connection String

Azure Key Vault

Account selection method ⓘ

From Azure subscription

Azure subscription ⓘ

Select all

Storage account name *

<your storage name>

Advanced ⓘ

Cancel

Test connection

Finish

c. In the **Advanced settings** section, deselect the **Use type default** option, then select **Next**.

Copy Data

1 Properties
One time copy

2 Source
Azure SQL Database

3 Destination
Azure SQL Data Warehouse

4 Settings

5 Summary

6 Deployment

Settings

More options for data movement

Fault tolerance settings

Fault tolerance: Abort activity on first incompatible row

Performance settings

☒ Enable Staging

Staging Settings

Staging Account Linked Service: AzureStorage

Storage Path:

☐ Enable Compression

Advanced settings

☒ Allow polybase

Reject type: Value

Reject value: 0

☐ Use Type default

Previous Next

9. In the **Summary** page, review the settings, and select **Next**:

Copy Data

1 Properties
One time copy

2 Source
Azure SQL Database

3 Destination
Azure SQL Data Warehouse

4 Settings

5 Summary

6 Deployment

Summary

You are running pipeline to copy data from Azure SQL Database to Azure SQL Data Warehouse.

SQL Azure SQL Database Copy run time region: East US Azure Blob Storage Region: Unknown Copy run time region: East US Azure SQL Data Warehouse Region: Unknown

Properties

Task name: CopyPipeline_kd8

Task description:

Source

Existing connection: AzureSqlDatabase

Dataset name: SourceDataset_kd8

Table name: items

Destination

Existing connection: AzureSqlDW

Dataset name: DestinationDataset_kd8

Table name: items

Copy settings

Timeout: 7:00:00:00

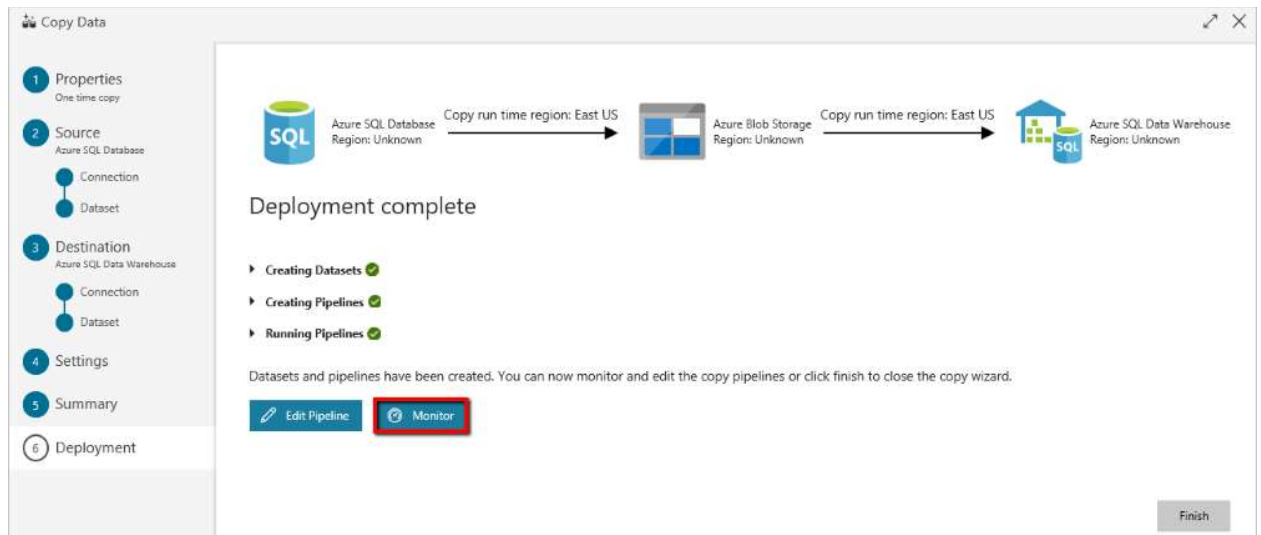
Retry: 0

Retry interval: 30

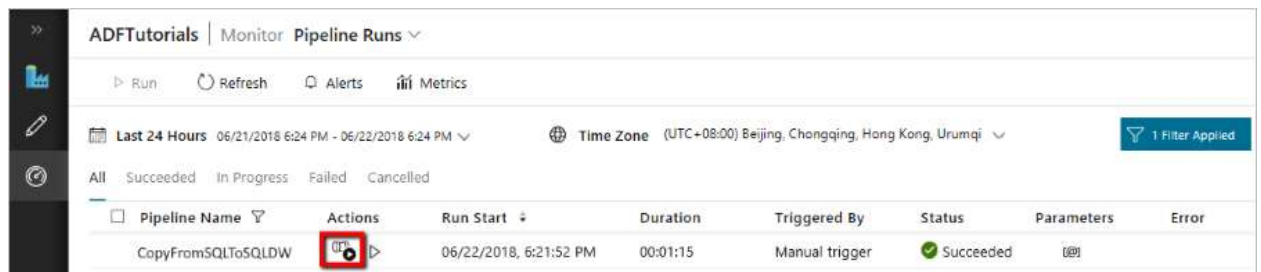
Secure Output: false

Previous Next

10. In the **Deployment** page, select **Monitor** to monitor the pipeline (task):






11. Notice that the **Monitor** tab on the left is automatically selected. The **Actions** column includes links to view activity run details and to rerun the pipeline:







12. To view activity runs that are associated with the pipeline run, select the **View Activity Runs** link in the **Actions** column. To switch back to the pipeline runs view, select the **Pipelines** link at the top. Select **Refresh** to refresh the list.


13. To monitor the execution details for each copy activity, select the **Details** link under **Actions** in the activity monitoring view. You can monitor details like the volume of data copied from the source to the sink, data throughput, execution steps with corresponding duration, and used configurations:







All Pipeline Runs / CopyFromSQLToSQLDW - Activity Runs

 Rerun  Rerun from activity  Refresh

ForEach 












 ForEach_xeg

Activity Runs

Pipeline Run ID **4f6eb0f4-4d5d-47ba-a6cb-67b4d96981f0**

All Succeeded In Progress Timed Out Failed Cancelled

ACTIVITY NAME	ACTIVITY TYPE	ACTIONS	SOURCE	DESTINATION
ForEach_xeg	ForEach	 		
Copy_xeg	Copy	  	[SalesLT].[ProductCate...	[SalesLT].[ProductCate...]
Copy_xeg	Copy	  	[SalesLT].[SalesOrderD...	[SalesLT].[SalesOrderD...]
Copy_xeg	Copy	  	[SalesLT].[ProductMod...	[SalesLT].[ProductMod...]

You should see the following details.

Details

[Learn more on copy performance details from here.](#)



Execution details:

▲ Azure SQL Database → Azure Blob Storage Queue 00:00:04 | Time to first byte 00:00:00 | Transfer 00:00:01

Start time	01/17/2018, 11:14:03 PM
Duration	00:00:06
Used DMUs	4
Used parallel copies	1

▲ Azure Blob Storage → Azure SQL Data Warehouse Queue 00:00:01 | Transfer 00:00:05

Start time	01/17/2018, 11:14:10 PM
Duration	00:00:06
Used DMUs	4
Used parallel copies	1
SQL DW Polybase	true

Review what has been created by the wizard in Azure Data Factory.

»

Data Factory

▼

Publish All

✓ Validate All

Refresh

Discard All

Data Flow

Factory Resources

Filter resources by name

Pipelines

1

CopyFromSQLToSQLDW

Datasets

2

DestinationDataset_xeg

2

SourceDataset_xeg

Data Flows (Preview)

0

CopyFromSQL...

Activities

Search Activities

Move & Transform

Batch Service

Databricks

Data Lake Analytics

General

HDInsight

Iteration & Conditionals

Machine Learning

Save as template

Validate

▶

ForEach

ForEach_xeg

+

−

🔒

100%

🔍

🔄

📏

🗑️

General

Settings

Activities (1)

Name *

ForEach_xeg

Description

Create ADF pipeline using Mapping Data Flows

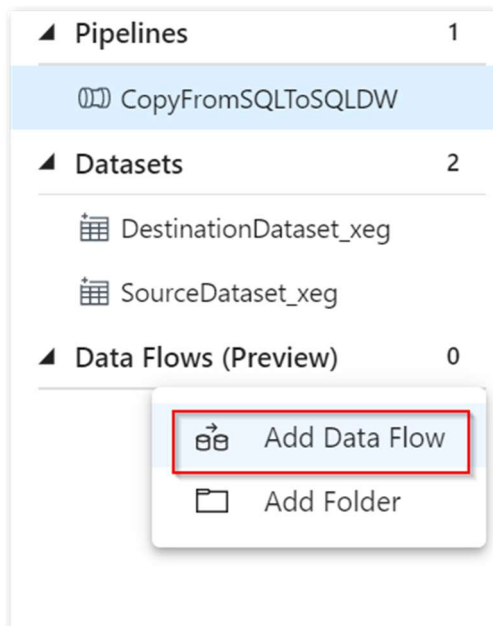
ADF Mapping data flows is a data transformation part of ADF which helps to visually define the data transformation using the Azure Databricks technology in the background.

These defined data flows could be then used in Azure Data Factory pipeline as any other object available in ADF.

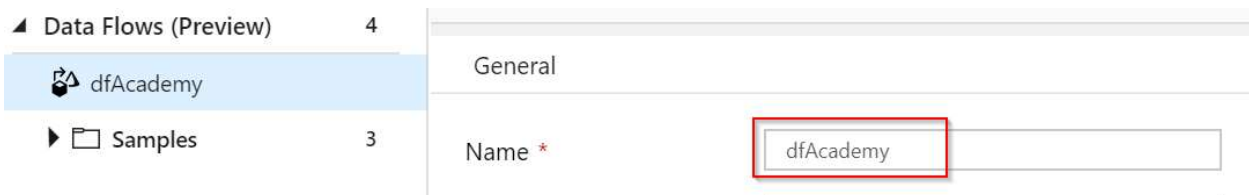
Azure Databricks cluster could be pre-provisioned in your environment or is created upon request in the background by ADF and destroyed after the pipeline finishes.

Create Mapping Data Flow

In the left panel, create new blank Data Flow



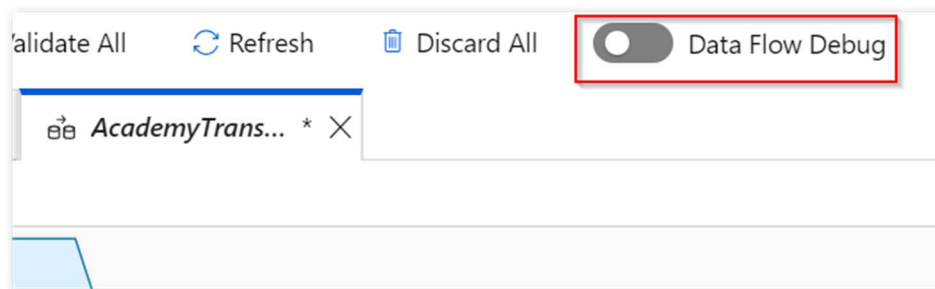
Give your new Data Flow the proper name (for example **dfAcademy**).



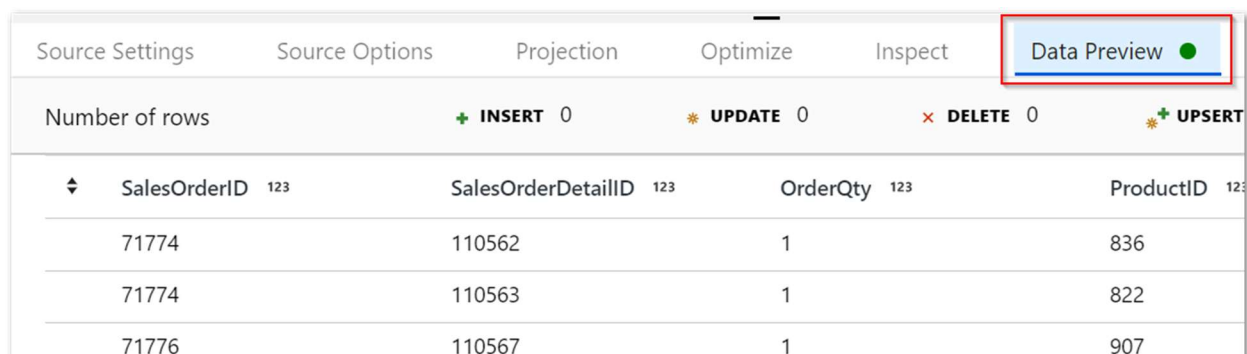
For the use of this demo, we will get some of the tables from the source SQL Database and denormalize them into one table suitable for (say) reporting.

We will create one table with order details along with data about customer and its address.

Click on the **Data Flow Debug** switch to enable seeing your data live without actually invasively touching it. This takes a few minutes as Data Factory runs small cluster under the covers and needs to start it.



After the switch is activated, you can preview all your changes and data on the **Data Preview** tab located on the right side of properties window.



Source Settings	Source Options	Projection	Optimize	Inspect	Data Preview ●
Number of rows	+ INSERT 0	✳ UPDATE 0	✗ DELETE 0	✳ UPSERT 0	
⬇ SalesOrderID 123	SalesOrderDetailID 123	OrderQty 123		ProductID 123	
71774	110562	1		836	
71774	110563	1		822	
71776	110567	1		907	

Create source datasets

First, create the following source datasets representing the source tables from SQL database: *SalesOrderHeader*, *Customer*, *CustomerAddress*, *Address*. These are obviously fact table and 3 dimensions.

- Click on Add source and in the Properties window give your source dataset a name (**srcSQLSalesOrderHeader**)
- After you give it a name, click on the **+New** button to add connection to the actual table

Source Settings
Source Options
Projection
Optimize
Inspect

Output stream name *
srcSQLTableName

Source dataset *
Select...

+ New

- Choose Azure SQL Database and fill in the appropriate values on the Properties pop-up window

← Set Properties

Name
srcSQLSalesOrderHead 1

Linked service *
SQLDB_Academy 2

Edit Connection

Table
[SalesLT].[SalesOrderHeader] 3

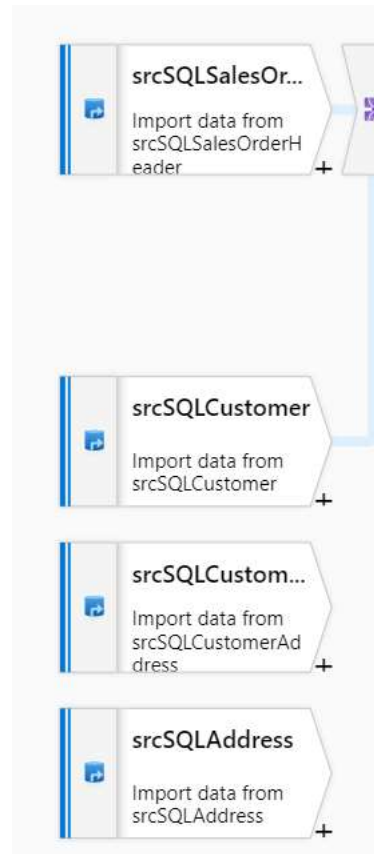
☐ Edit

Import schema

☒ From connection/store
☐ None

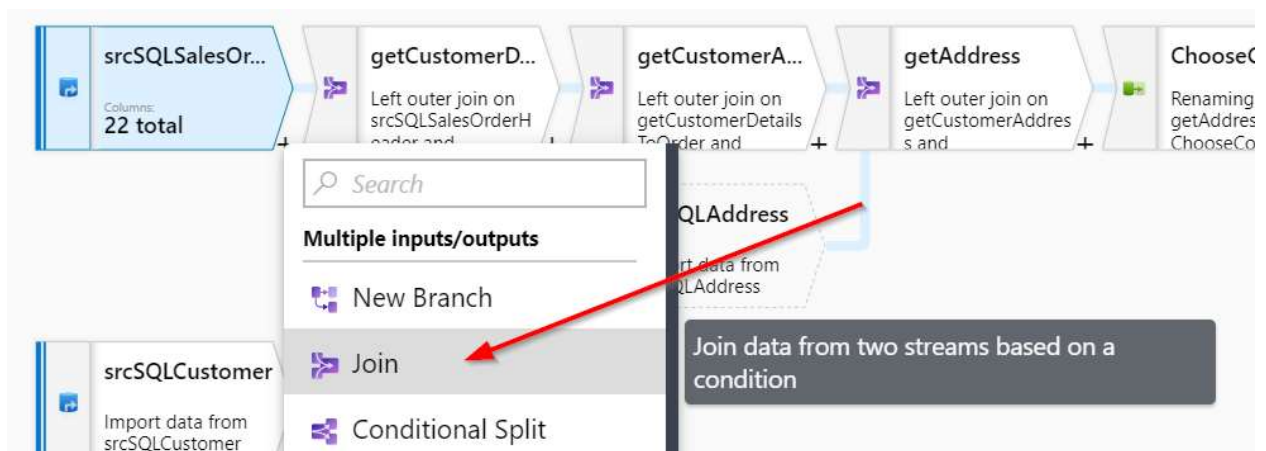
Iterate this process for all the tables – your datasets list and canvas should now look similar to this:

▲ Datasets	22
▲ Academy	9
DestinationDataset_req	
srcSQLSalesOrderHeader	
srcSQLCustomerAddress	
srcSQLCustomer	
srcSQLProduct	
srcSQLSalesOrderHeader	



Join tables

From the first data source, click on the + sign, review possible activities and choose **Join** as the desired activity:



- Give your join activity proper name (**getCustomerDetailsToOrder**)
- Define the right stream – with what the join should be executed (**srcSQLCustomer**)
- On the left stream choose **CustomerID** from orders table and on the right side the **CustomerID** from the Customer table
- Regarding the join type, choose **Left outer** join in order to bring customers to their orders.

Join Settings Optimize Inspect Data Preview

Output stream name * 1

Left stream *

Right stream * 2

Join type * Full outer Inner Left outer 5 Right outer Cross

Join conditions * **Left: srcSQLSalesOrderHeader's column** 3 == **Right: srcSQLCustomer's column** 4

Join additional tables

Follow the same procedure as above but create joins for **CustomerAddress** and **Address** tables in addition.

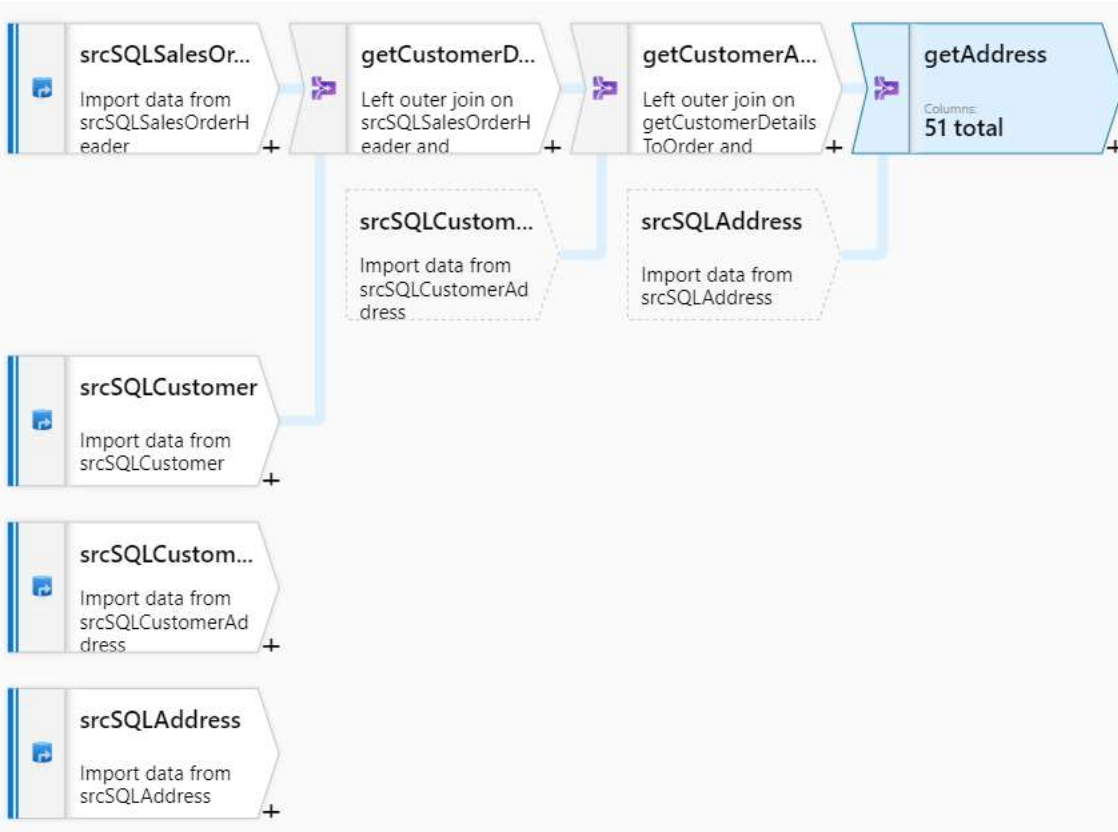
You can always check our data within the flow on the **Data Preview** tab.

Join Settings Optimize Inspect **Data Preview**

When hovering over the data flow element you see how many columns are currently there in the outgoing data set.



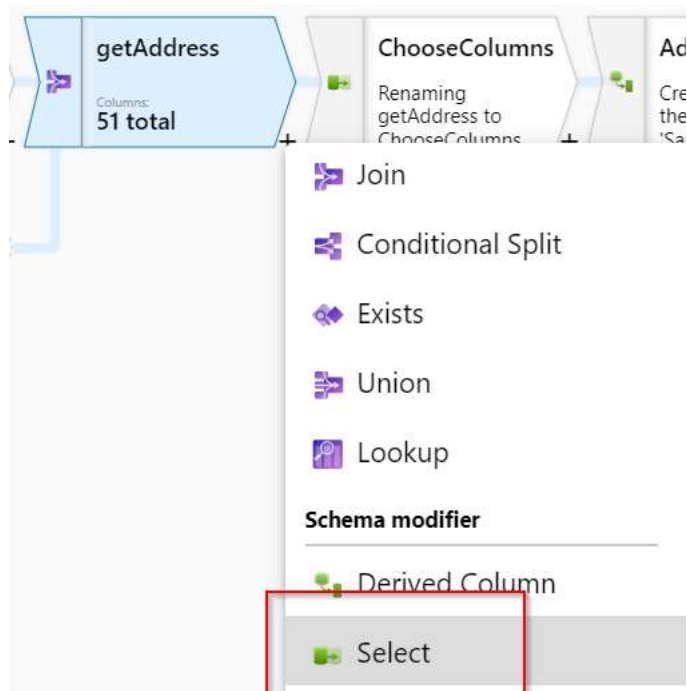
Your canvas should currently look something similar to this:



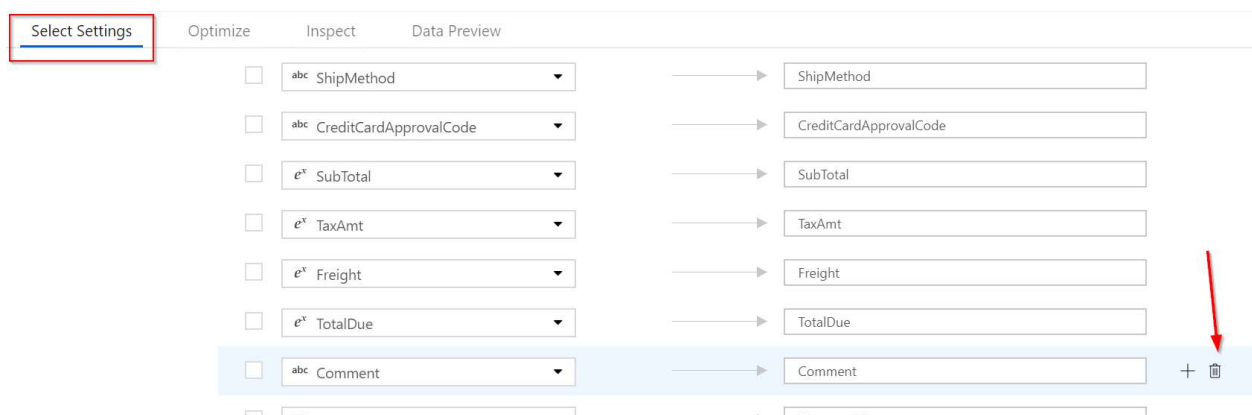
Limit columns in output

As you can imagine it is quite rough to have all the columns within the output table after joining fact table with 3 dimensions.

After the last flow, click on the + sign and choose **Select** activity.



There are some redundant columns with the same names from within multiple tables, please clean your data choosing and deleting columns you do not want in your output.



You can also change the name of the output columns if needed.

Add timestamp columns to control loading

In this simple demo, we would like to show how to create custom column in our data – in this case we will create the timestamp column to see when the load into the Warehouse happened.

- After the last flow, click on the + sign and choose **Derived Column** as the activity

- In the properties window, give your flow proper name, define the name of your new column and click on the Expression text box to define, how your new column should look like

- Review the **Visual Expression Builder** – test, what functions are available and how the editor behaves
- For the contents of this column, choose **currentTimestamp()** function and Save the window

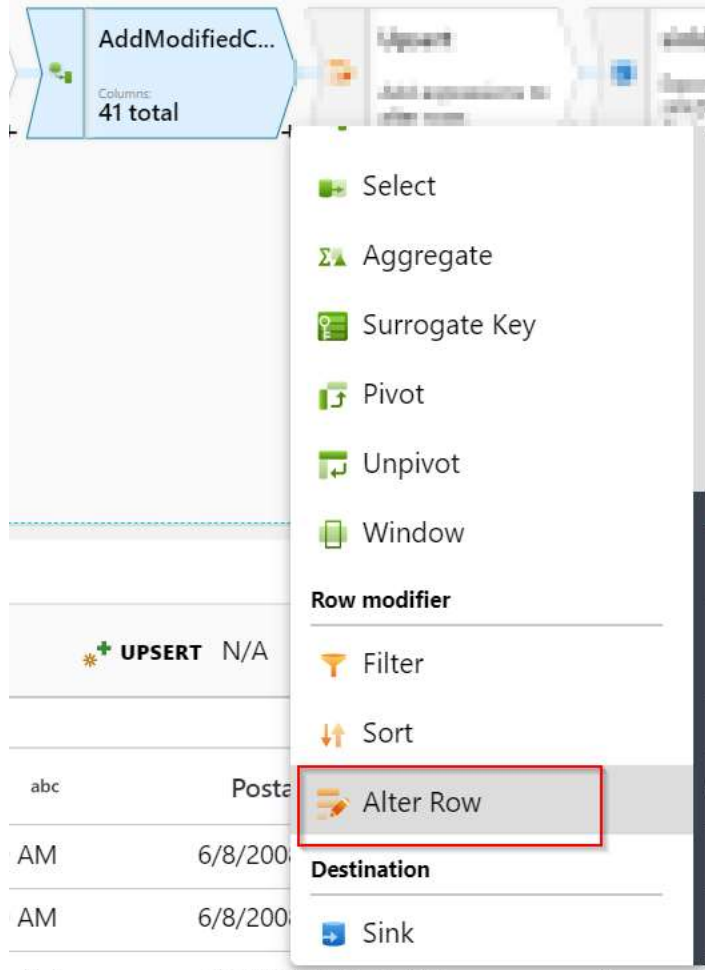
- Preview the data clicking on **Data Preview**

For the **upserts** being setup in the following chapter, this column gives you an idea, what records have been inserted and updated and when.

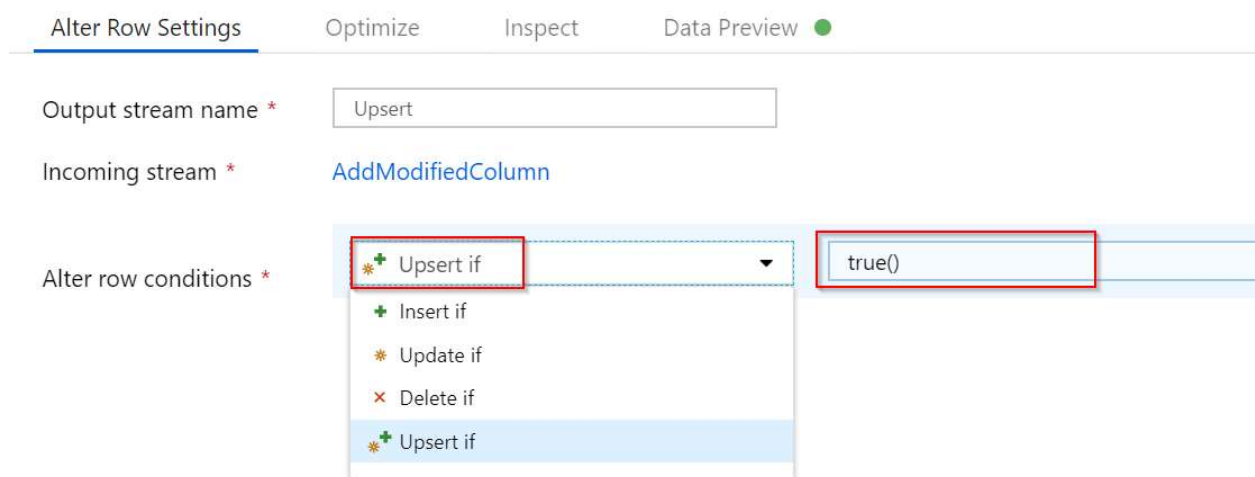
Alter row upserts conditions

Although not needed for this case, you can define much more complex conditions for your inserts, deletes, updates and upserts.

In order to achieve this, add the **Alter Row** activity



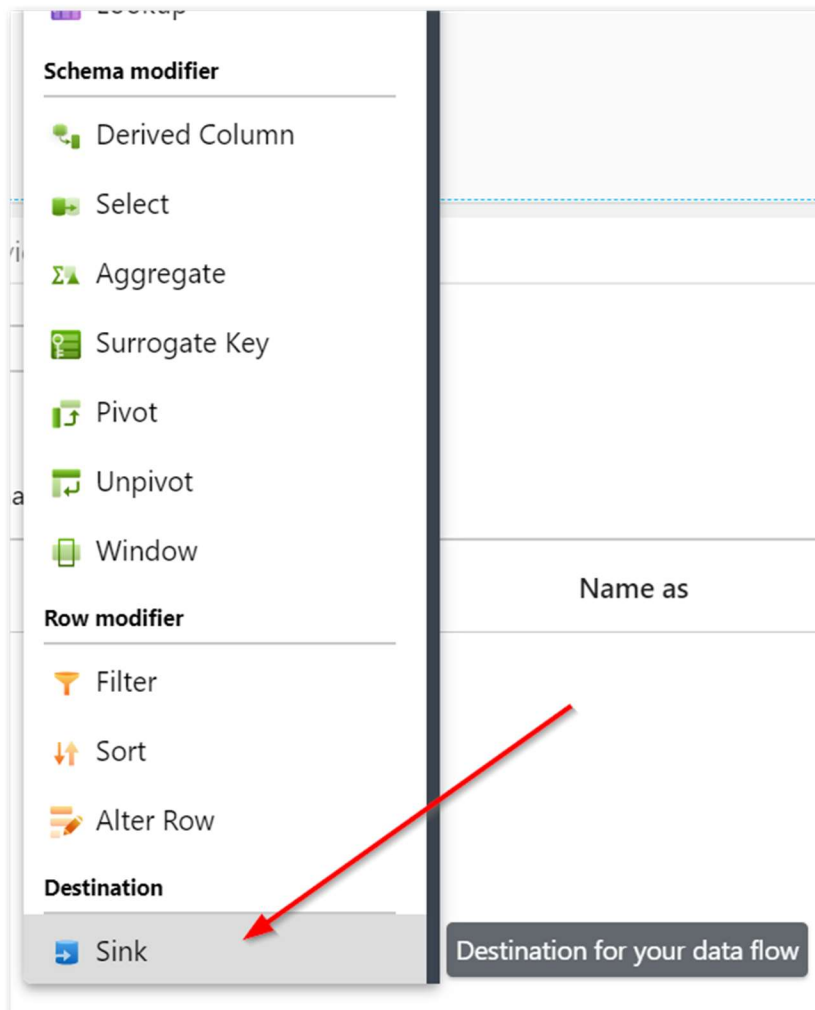
After you give your step a proper name, you can choose from the conditions (on the left) and the functions adding the business logic returning **true()** or **false()**.



For the sake of this demo, we put **true()** function (which is obviously always true) and thus does nothing – but you get the picture.

Create Azure Data Warehouse as a sink

Continuing in your flow, press + and choose Sink from the context menu.



With the properties window, add the Azure Data Warehouse table as the sink for your flow. Create the new table and name it **SalesLT.CustomerOrders**.

General


Connection

Schema

Parameters

Linked service *

 SQLDWH_Academy ▼

 Test connection

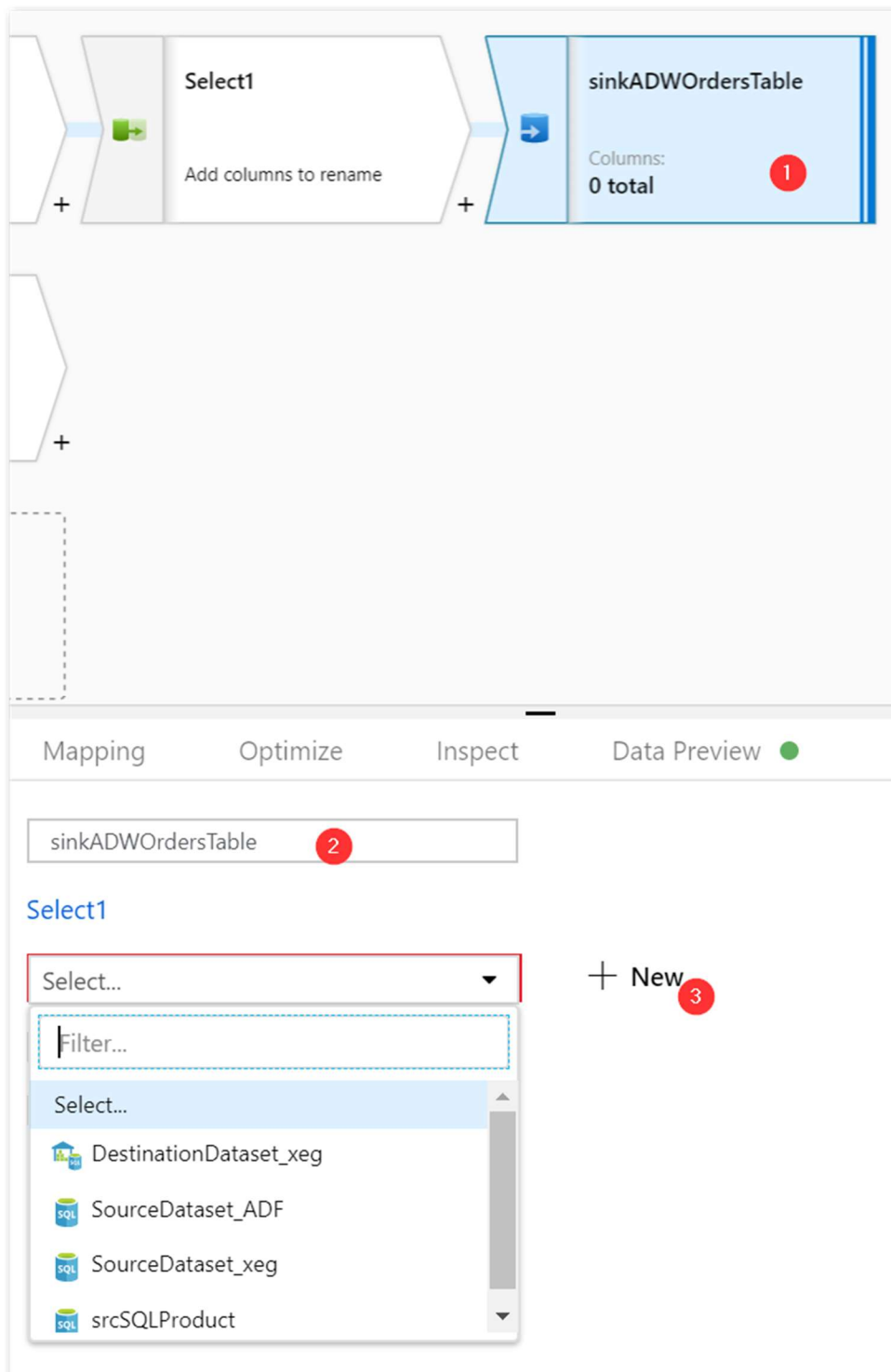


Table

SalesLT.CustomerOrders

 Preview data

 Edit



Following the “sink” settings, please make sure you allow **upserts**, choose the **SalesOrderID** as the mapping ID column and Enable staging for faster loads.

Sink	Settings	Mapping	Optimize	Inspect	Data
Update method	<input type="checkbox"/> Allow insert				
	<input type="checkbox"/> Allow delete				
	<input checked="" type="checkbox"/> Allow upsert				
	<input type="checkbox"/> Allow update				
Key columns	123 SalesOrderID				
Table action	<input checked="" type="radio"/> None	<input type="radio"/> Recreate table	<input type="radio"/> Truncate table		
Enable staging	<input checked="" type="checkbox"/>				

The settings above are available only when the sources and sinks supporting them are available (SQL DB & SQL DW).

Create the pipeline to execute your data flow

ADF Mapping Data Flows are the ETL concept backed by Azure Databricks capabilities. As such, you define individual data flows and although you can debug them, you need the **Pipeline** to include the ADF Mapping Data Flow into some process. To be able to activate it upon trigger and monitor it.

- Create new pipeline and name it **CustomerOrdersSQLtoDW**.
- Choose Data Flow from the **Move & Transform** section.

Activities



Move & Transform

Copy Data

Data Flow (Preview)

▶ Batch Service

▶ Databricks

▶ Data Lake Analytics

▶ General

▶ HDInsight

▶ Iteration & Conditionals

▶ Machine Learning

- Choose your data flow within the popup window
- Click on the data flow to be able to set its “running” parameters
- On the **Settings** page, choose the proper Blob Storage created for this exercise and define the container where the data should be storage while staging

General

Settings

User Properties

Data Flow *

dfAcademy



Edit

+ New

Run on *

AutoResolveIntegrationRuntime



▶ PolyBase

Staging linked service



AcademyStorage

Staging storage folder

adfstagedpolybasetempdata



Connection successful



Test connection

Browse

You can click Debug to run the Pipeline including your data flow or you can add trigger to activate your pipeline as usual.

