

Copying Data Using Azure Data Factory

Exercise 0 – Setup Azure Data Lake Gen2 Account

1. Complete Lab 1 – Working with Azure Data Lake Gen2 account

Exercise 1 – Setup Azure Data Factory

1. Go to Azure portal (portal.azure.com)
2. In the search bar, search for Data Factories. And select it
3. Click on Create New
4. Fill up the properties to create account
 - a. [Basics Tab]
 - i. Select subscription
 - ii. Select resource group
 - iii. Provide a unique name
 - iv. Select region of your choice (example – East US 2)
 - v. Select version as V2
 - vi. Click Review + Create

Create Data Factory

Basics Git configuration Networking Advanced Tags Review + create

One-click to create data factory with sample pipeline and datasets. [Try it](#)

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription *

Resource group * [Create new](#)

Instance details

Name * ✓

Region *

Version *

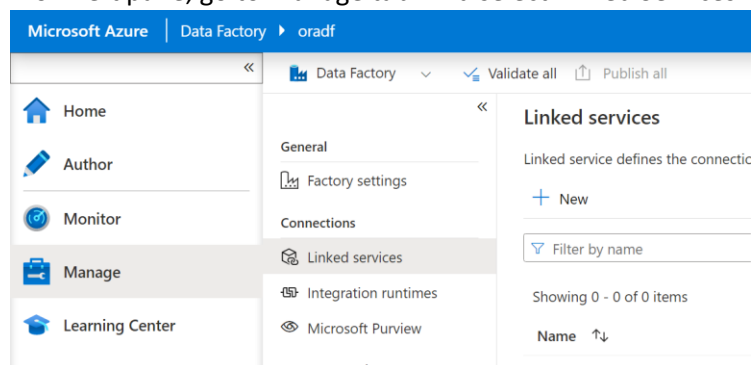
- b. Click Create

Exercise 2 – Copy File from One Data Lake Folder to Another Using ADF

1. Open Azure Data Factory instance created in the previous step
2. Click on Launch Studio, to open ADF UI



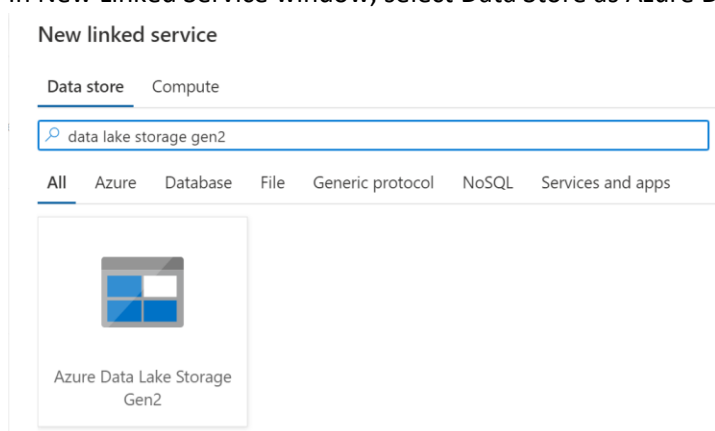
3. From left pane, go to Manage tab. And select Linked Services



4. Create new Linked Service for Azure Data Lake account

a. Click New

b. In New Linked Service window, select Data Store as Azure Data Lake Gen2



c. Fill up the properties:

- i. Name: MyStorageLinkedService
- ii. Authentication type: Account key (this is the access key of storage account)
- iii. Storage account name: Select name of Data Lake account
- iv. Click Test connection
- v. If Test connection is successful, click Create

New linked service
 Azure Data Lake Storage Gen2 [Learn more](#)

Name *
 MyStorageLinkedService

Description

Connect via integration runtime *
 AutoResolveIntegrationRuntime

Authentication type
 Account key

Account selection method
☒ From Azure subscription ☐ Enter manually

Azure subscription
 Select all

Storage account name *
 oreillydatalake

Test connection
☒ To linked service ☐ To file path

Annotations

Connection successful

Create Back Test connection Cancel

5. Create a dataset for source file – TaxiZones.csv

a. From left pane, go to Author tab

b. Add a new dataset

c. In New Dataset window, select Data Store as Azure Data Lake Gen2

New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store

data lake storage gen2

All Azure Database File Generic protocol NoSQL Services and apps

Azure Data Lake Storage Gen2

d. Select format for source file – which is DelimitedText for TaxiZones.csv

e. Fill up the properties:

- Name: TaxiZonesCsvDataset
- Linked service: MyStorageLinkedService
- In file path, select browse. And select TaxiZones.csv file
- Select checkbox: First row as header
- Import schema: From connection/store
- Click OK to create

Set properties

Name
TaxiZonesCsvDataset

Linked service *
MyStorageLinkedService

File path
taxidata / Raw / TaxiZones.csv

First row as header ☒

Import schema
☒ From connection/store
 ☐ From sample file
 ☐ None

- f. Once source dataset is created, click on Preview data to verify there are 4 columns

Preview data

Linked service: MyStorageLinkedService
Object: TaxiZones.csv

	LocationID	Borough	Zone	service_zone
1	1	EWB	Newark Airport	EWB
2	2	Queens	Jamaica Bay	Boro Zone
3	3	Bronx	Allerton/Pelham Gardens	Boro Zone
4	4	Manhattan	Alphabet City	Yellow Zone
5	5	Staten Island	Arden Heights	Boro Zone
6	6	Staten Island	Arrochar/Fort Wadsworth	Boro Zone
7	7	Queens	Astoria	Boro Zone
8	8	Queens	Astoria Park	Boro Zone
9	9	Queens	Auburndale	Boro Zone
10	10	Queens	Baisley Park	Boro Zone

[View more](#) [Preview data](#)

6. Create a dataset for sink file – TaxiZones.json (this file doesn't exist)

- a. Add a new dataset

- b. In New Dataset window, select Data Store as Azure Data Lake Gen2


New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store

data lake storage gen2

All Azure Database File Generic protocol NoSQL Services and apps


Azure Data Lake Storage Gen2

- c. Select format for sink file – select JSON to store output data in JSON format

- d. Fill up the properties:

- Name: TaxiZonesJsonDataset
- Linked service: MyStorageLinkedService
- In file path, select browse. And select folder location to store output file.
- Manually make changes to path to add folder & file names (see image below)
- Select checkbox: First row as header
- Import schema: None (since file doesn't exist, its schema can't be imported)
- Click OK to create

Set properties

Name
TaxiZonesJsonDataset

Linked service *
MyStorageLinkedService

File path
taxidata / Output / TaxiZones.json

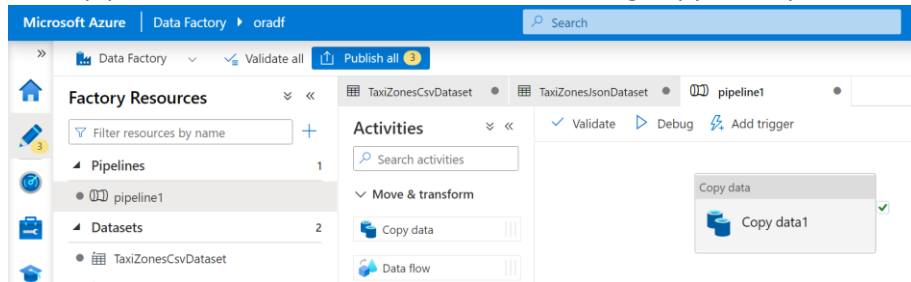
Import schema
☐ From connection/store ☐ From sample file ☒ None

e. After creation, you cannot preview the data since the file doesn't exist

7. Create a pipeline to copy data

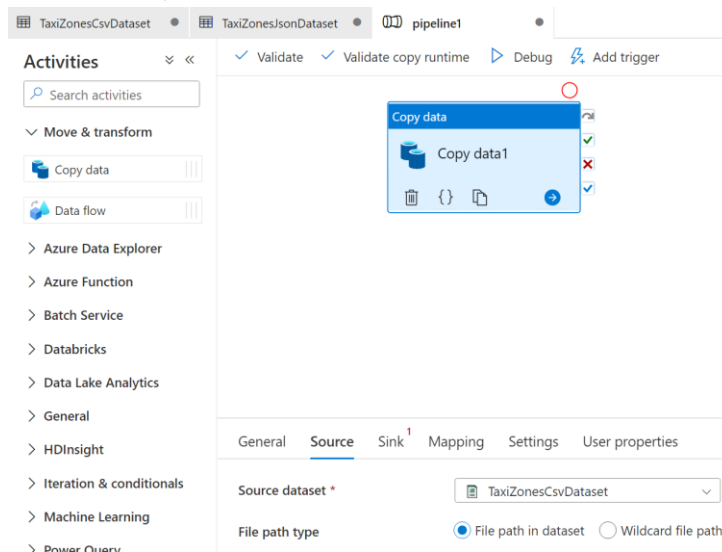
a. In Author tab, create new Pipeline

b. In the pipeline, from Move & transform section, drag Copy activity on canvas

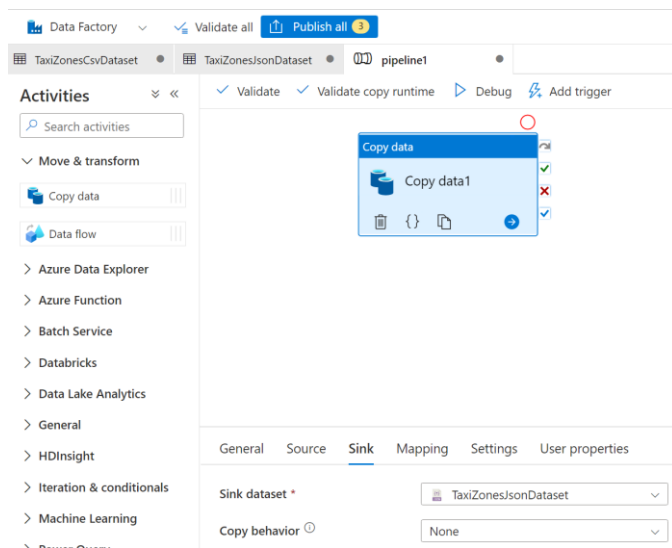


c. Click on Copy activity

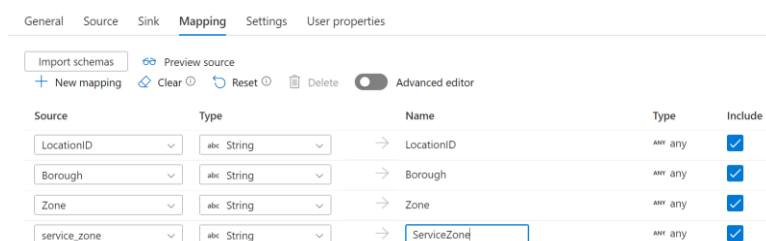
d. In source tab, select source dataset → TaxiZonesCsvDataset



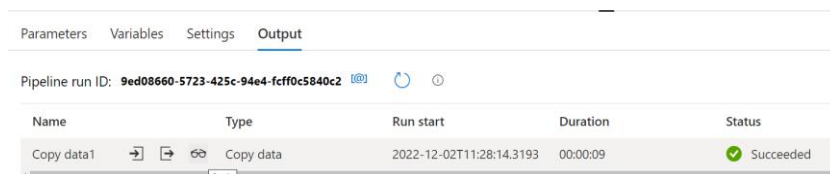
e. In sink tab, select sink dataset → TaxiZonesJsonDataset



- f. In Mapping tab, click Import schemas
- g. You can remove or rename sink columns (as shown in image below)



8. Once configured, click on Debug to execute the Pipeline
9. Click anywhere in canvas (but not on CopyData activity). Notice the glasses icon in the pipeline run details. Click on it to see the statistics



10. Go back to Data Lake account to check the creation of new file, TaxiZones.json
11. Click on Publish All to save the changes

Exercise 3 – Setup Azure SQL Database

1. Go to Azure portal (portal.azure.com)
2. In the search bar, search for SQL Databases. And select it
3. Click on Create New
4. Fill up the properties to create Azure SQL database
 - a. [Basics Tab]
 - i. Select subscription

- ii. Select resource group
- iii. Database name: DB1
- iv. Server: Select – Create new
 - 1. Server name: Provide a unique name
 - 2. Region: Select a region of your choice
 - 3. Authentication method: Use SQL authentication
 - 4. Provide admin username and password
 - 5. Click create

Home > SQL databases > Create SQL Database >

Create SQL Database Server

Microsoft

Server details

Enter required settings for this server, including providing a name and location. This server will be created in the same subscription and resource group as your database.

Server name * ✓
 .database.windows.net

Location * ✓

Authentication

Select your preferred authentication methods for accessing this server. Create a server admin login and password to access your server with SQL authentication, select only Azure AD authentication [Learn more](#) or using an existing Azure AD user, group, or application as Azure AD admin [Learn more](#), or select both SQL and Azure AD authentication.

Authentication method

☐ Use only Azure Active Directory (Azure AD) authentication

☐ Use both SQL and Azure AD authentication

☒ Use SQL authentication

Server admin login * ✓

Password * ✓

Confirm password * ✓

- v. Workload environment: Development
- vi. Compute + Storage: Click configure database
 - 1. Service tier: Basic
 - 2. Apply

Home > SQL databases > Create SQL Database >

Configure

[Feedback](#)

Service and compute tier

Select from the available tiers based on the needs of your workload. The vCore model provides a wide range of configuration controls and offers Hyperscale and Serverless to automatically scale your database based on your workload needs. Alternately, the DTU model provides set price/performance packages to choose from for easy configuration. [Learn more](#)

Service tier ✓
[Compare service tiers](#)

DTUs [Compare DTU options](#)

5 (Basic)

Data max size (GB)

- vii. See final configuration in image below:

Create SQL Database

Microsoft

⚠ Changing Basic options may reset selections you have made. Review all options prior to creating the resource.

Basics Networking Security Additional settings Tags Review + create

Create a SQL database with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize. [Learn more](#)

Did you know that new users in Azure can create a free Azure SQL Database and use it for 12 months using Azure free account? [Learn more](#)

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription *
Resource group *
[Create new](#)

Database details

Enter required settings for this database, including picking a logical server and configuring the compute and storage resources

Database name *
Server *
[Create new](#)

Want to use SQL elastic pool? ☐ Yes ☒ No

Workload environment
☒ Development
☐ Production

Default settings provided for Development workloads. Configurations can be modified as needed.

Compute + storage * **Basic**
2 GB storage
[Configure database](#)

- b. [Additional settings Tab]
 - i. Use existing data: Sample
 - ii. Click Review + Create

- c. Click Create

5. Once created, open Azure SQL database

6. From top menu, select – Set server Firewall



DB1 (orsqlsrvr/DB1)

SQL database

Search

[Copy](#) [Restore](#) [Export](#) [Set server firewall](#)

Overview

Activity log

This database was just created. Do you need any help [getting started](#)?

Essentials

7. In Networking window:
 - a. Select Public access
 - b. Click on selected networks
 - c. In Firewall rules, click on Add your client IPv4 address. This will create a rule as shown below:
 - d. In Exceptions, select “Allow Azure services and resources to access this server”

Networking

Feedback

Public access Private access Connectivity

Public network access
Public Endpoints allow access to this resource through the internet using a public IP address. An application or resource that is granted access with the following

Public network access ☐ Disable ☒ Selected networks

Connections from the IP addresses configured in the Firewall rules section below will have access to this database. By i

Virtual networks
Allow virtual networks to connect to your resource using service endpoints. [Learn more](#)

+ Add a virtual network rule

Rule	Virtual network	Subnet	Address range	Endpoint status	Resource group	Subscription	State
------	-----------------	--------	---------------	-----------------	----------------	--------------	-------

Firewall rules
Allow certain public internet IP addresses to access your resource. [Learn more](#)

+ Add your client IPv4 address + Add a firewall rule

Rule name	Start IPv4 address	End IPv4 address	
ClientIPAddress_2022-12-2_17-20-23	<input type="text"/>	<input type="text"/>	<input type="button" value="Delete"/>

Exceptions

☒ Allow Azure services and resources to access this server ☐

- e. Save
- f. Close Networking window

8. From left pane of SQL database, go to Query Editor

9. Add your Azure SQL admin username and password that you defined in previous step, and login

Home > Microsoft.SqlDatabase.newDatabaseNewServer_e96ce96bd96b4fcfb9d4b | Overview > DB1 (orsqlsrvr/DB1)

DB1 (orsqlsrvr/DB1) | Query editor (preview)

Search < Login + New Query ↑ Open query Feedback

Overview
Activity log
Tags
Diagnose and solve problems
Getting started
Query editor (preview)

Settings

Compute + storage
Connection strings
Properties
Locks

Data management

Replicas
Sync to other databases

Integrations

Welcome to SQL [

SQL server authentication

Login *

Password * ✓

OK

10. Once logged in, run the following script to create a watermark table:

```
CREATE Table Watermark
(
    TableName VARCHAR(255),
    WatermarkValue DATETIME,
);
```

```
INSERT INTO Watermark
VALUES ('SalesLT.Customer', '2000-01-01');
```

```
SELECT * FROM Watermark;
```

11. Remove previous script and run following script to create a stored procedure to update watermark value:

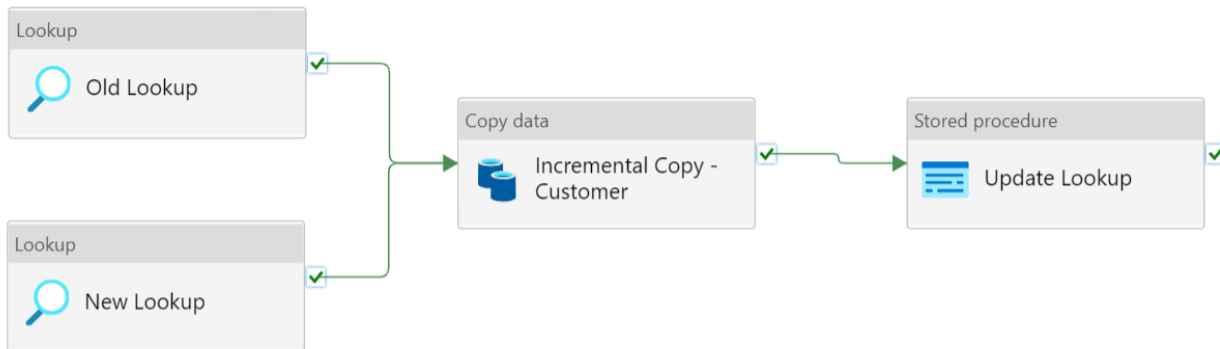
```
CREATE PROCEDURE UpdateWatermark @LastModifiedtime DATETIME, @TableName varchar(50)
```

```

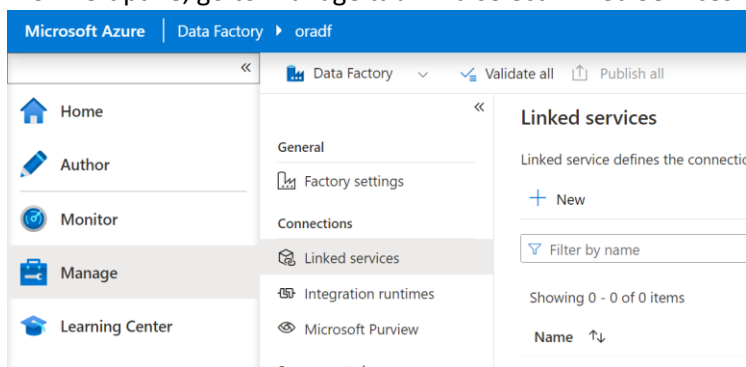
AS
BEGIN
    UPDATE Watermark
    SET WatermarkValue = @LastModifiedtime
    WHERE TableName = @TableName
END;

```

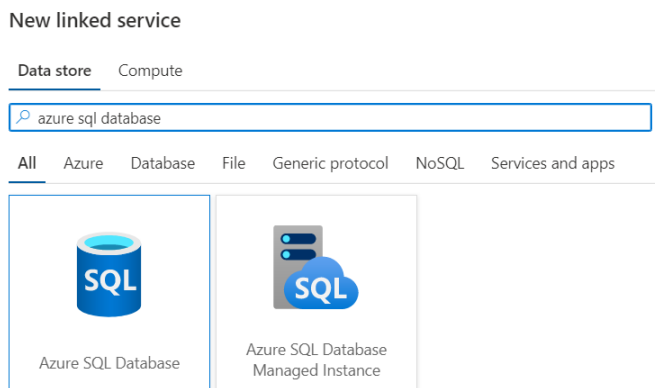
Exercise 4 – Copy Data Incrementally from Azure SQL Database to Data Lake



1. Go to Data Factory
2. From left pane, go to Manage tab. And select Linked Services




3. Create new Linked Service for Azure SQL database
 - a. Click New
 - b. In New Linked Service window, select Azure SQL Database



- d. Fill up the properties:
 - i. Name: AzureSqlLinkedService

- ii. Select your server name
- iii. Select your database name
- iv. Authentication type: SQL authentication
- v. Provide user name & password
- vi. Click Test connection
- vii. If Test connection is successful, click Create

New linked service

 Azure SQL Database [Learn more](#)

Name *

Description

Connect via integration runtime * ⓘ

Connection string

Azure Key Vault

Account selection method ⓘ

☒ From Azure subscription
☐ Enter manually

Azure subscription

Server name *

Database name *

Authentication type *

User name *

Password

Azure Key Vault

Password *

Always encrypted ⓘ
☐


Additional connection properties


+ New

Annotations

Create

Back

 Connection successful

 Test connection

Cancel

12. Create a dataset for source table – SalesLT.Customer

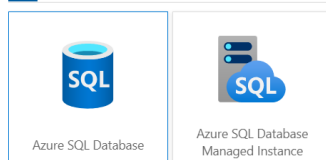
- a. From left pane, go to Author tab
- b. Add a new dataset
- c. In New Dataset window, select Data Store as Azure SQL Database

New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store

All Azure Database File Generic protocol NoSQL Services and apps



- d. Fill up the properties:
 - i. Name: CustomerSqlDataset
 - ii. Linked service: AzureSqlLinkedService
 - iii. Select table name: SalesLT.Customer
 - iv. Import schema: From connection/store
 - v. Click OK to create

Set properties

Name

Linked service *

Table name

☐ Edit

Import schema
☒ From connection/store ☐ None

- e. Once source dataset is created, click on Preview data to verify table data

CustomerSqlDataset

SQL Azure SQL Database
CustomerSqlDataset

Connection Schema Parameters

Linked service * Test connection Edit +

Preview data

Linked service: AzureSqlLinkedService
 Object: SalesLT.Customer

	CustomerID	NameStyle	Title	FirstName	MiddleName	LastName	Suffix	CompanyName
1	1	false	Mr.	Orlando	N.	Gee		A Bike Store
2	2	false	Mr.	Keith		Harris		Progressive Sports
3	3	false	Ms.	Donna	F.	Carreras		Advanced Bike Components

13. Create a dataset for sink file – Customer.json (this file doesn't exist)


- a. Add a new dataset
- b. In New Dataset window, select Data Store as Azure Data Lake Gen2

New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store

All Azure Database File Generic protocol NoSQL Services and apps



Azure Data Lake Storage Gen2

- c. Select format for sink file – select JSON to store output data in JSON format
- d. Fill up the properties:
 - i. Name: CustomerJsonDataset
 - ii. Linked service: MyStorageLinkedService
 - iii. In file path, select browse. And select folder location to store output file.
 - iv. Manually make changes to path to add folder. Don't add any file name (see image below)
 - v. Import schema: None (since file doesn't exist, its schema can't be imported)
 - vi. Click OK to create

Set properties

Name
CustomerJsonDataset

Linked service *
MyStorageLinkedService

File path
taxidata / Output / File name

Import schema
☐ From connection/store
 ☐ From sample file
 ☒ None

- e. After creation, click in File name textbox
- f. Click on Add dynamic content, and add the expression to dynamically generate file name:

`@CONCAT('Customers-', pipeline().RunId, '.json')`

14. Create a dataset for watermark

- a. Add a new dataset
- b. In New Dataset window, select Data Store as Azure SQL Database


New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)


Select a data store

azure sql

All Azure Database File Generic protocol NoSQL Services and apps



Azure SQL Database



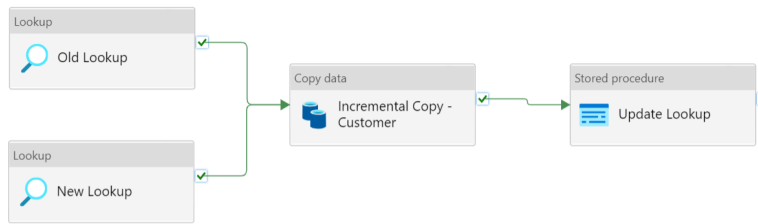
Azure SQL Database Managed Instance

- c. Fill up the properties:
 - i. Name: WatermarkSqlDataset
 - ii. Linked service: AzureSqlLinkedService
 - iii. Select table name: Watermark
 - iv. Import schema: From connection/store
 - v. Click OK to create
- d. Once watermark dataset is created, click on Preview data to verify table data

15. Create a pipeline to copy data

- a. In Author tab, create new Pipeline

16. In the pipeline, drag two Lookup, one Copy Data and one Stored procedure activities on canvas (see image below). Name them as shown below:



17. Select Old Lookup activity

- Go to settings
- Configure the properties:
 - Source dataset: WatermarkSqlDataset
 - Use query: Query
 - Query: `SELECT * FROM Watermark`

General Settings User properties

Source dataset * WatermarkSqlDataset [Open](#) [+ New](#) [Preview data](#) [Learn more](#)

First row only ☒

Use query ☐ Table ☒ Query ☐ Stored procedure

Query

`SELECT * FROM Watermark` [Add dynamic content \[Alt+Shift+D\]](#)

Query timeout (minutes)

Isolation level

Partition option ☒ None ☐ Physical partitions of table ☐ Dynamic range

Please preview data to validate the partition settings are correct before you trigger a run or publish the pipeline.

18. Select New Lookup activity

- Go to settings
- Configure the properties:
 - Source dataset: CustomerSqlDataset
 - Use query: Query
 - Query: `SELECT MAX(ModifiedDate) AS NewWatermarkValue FROM SalesLT.Customer`

General Settings User properties

Source dataset * WatermarkSqlDataset [Open](#) [+ New](#) [Preview data](#) [Learn more](#)

First row only ☒

Use query ☐ Table ☒ Query ☐ Stored procedure

Query

`SELECT * FROM Watermark` [Add dynamic content \[Alt+Shift+D\]](#)

Query timeout (minutes)

Isolation level

Partition option ☒ None ☐ Physical partitions of table ☐ Dynamic range

Please preview data to validate the partition settings are correct before you trigger a run or publish the pipeline.

19. Select Copy Data activity

- In source tab, select source dataset → CustomerSqlDataset
- Use query: Query
- Query: (Add dynamic content)

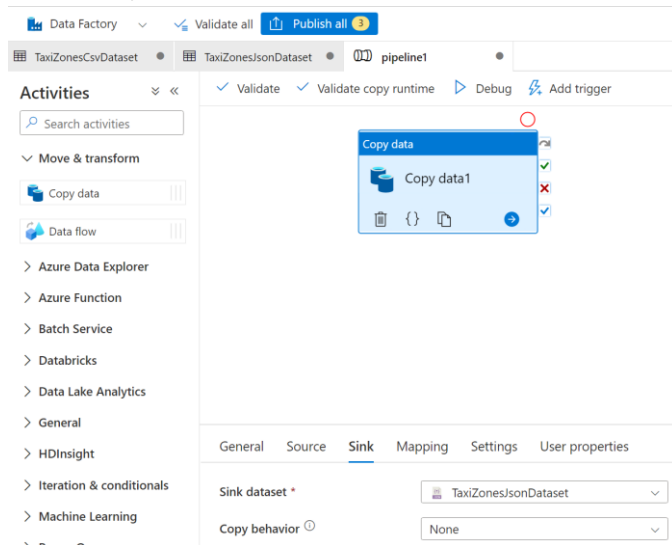
```
SELECT * FROM SalesLT.Customer
WHERE ModifiedDate > '@{activity('Old Lookup').output.firstRow.WatermarkValue}'
AND ModifiedDate <= '@{activity('New Lookup').output.firstRow.NewWatermarkValue}'
```

Pipeline expression builder

Add dynamic content below using any combination of [expressions](#), [functions](#) and [system variables](#).

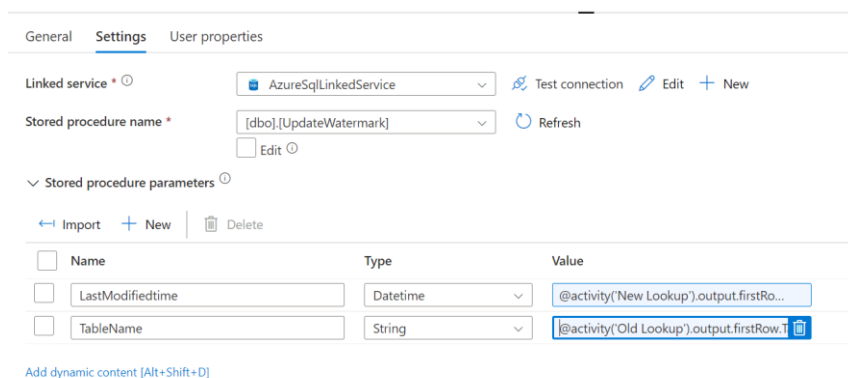
```
SELECT * FROM SalesLT.Customer
WHERE ModifiedDate > '@{activity('Old Lookup').output.firstRow.WatermarkValue}'
AND ModifiedDate <= '@{activity('New Lookup').output.firstRow.NewWatermarkValue}'
```

d. In sink tab, select sink dataset → CustomerJsonDataset



20. Select Update Lookup activity

- Go to settings
- Configure the properties:
 - Linked service: AzureSqlLinkedService
 - SP Name: UpdateWatermark
- Add two stored procedure parameters (as shown in image below):
 - LastModifiedtime → @activity('Old Lookup').output.firstRow.TableName
 - TableName → @activity('Old Lookup').output.firstRow.TableName









[Add dynamic content \[Alt+Shift+D\]](#)

21. Once configured, click on Debug to execute the Pipeline

22. Click anywhere in canvas (but not on any activity). Monitor the details

Parameters Variables Settings **Output**

Pipeline run ID: **abdb4da9-9edf-4f83-8dfd-2fb1595d40de**  

Name	Type	Run start	Duration	Status
Update Lookup	Stored procedure	2022-12-02T13:27:01.8462Z	00:00:03	 Succeeded
Incremental Copy - Customer	Copy data	2022-12-02T13:26:51.0218Z	00:00:10	 Succeeded
New Lookup	Lookup	2022-12-02T13:26:46.6649Z	00:00:04	 Succeeded
Old Lookup	Lookup	2022-12-02T13:26:46.6336Z	00:00:04	 Succeeded

23. Go back to Azure portal → SQL Database → Query Editor

24. Run following command to insert two new records in the table:

```
INSERT INTO SalesLT.Customer (NameStyle, FirstName, LastName, PasswordHash, PasswordSalt, rowguid, ModifiedDate)
```

```
VALUES ('False', 'Mohit', 'Batra', 'xxx', 'xxx', NEWID(), GETDATE());
```

```
INSERT INTO SalesLT.Customer (NameStyle, FirstName, LastName, PasswordHash, PasswordSalt, rowguid, ModifiedDate)
```

```
VALUES ('False', 'Andrew', 'Smith', 'yyy', 'yyy', NEWID(), GETDATE());
```

25. In Data Factory, debug the pipeline again

26. Monitor in Data factory that only 2 records have moved

27. Check new file in Data Lake and see that it only has 2 records