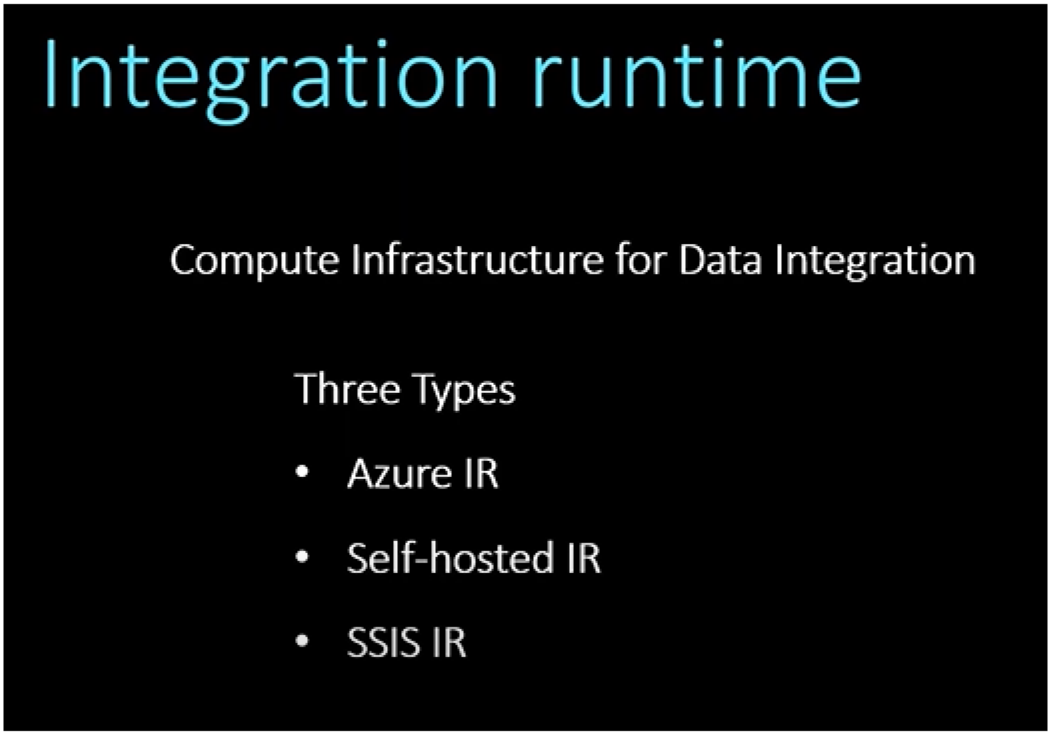
Integration Runtime, Linked Service and Datasets in ADF

1. SO what is integration runtime? It is a compute infrastructure for data integration.. We know that ADF is mainly used for data integration from diff sources…so it might need some compute power to do this… this compute power can be obtained from integration runtime
2. There are three types



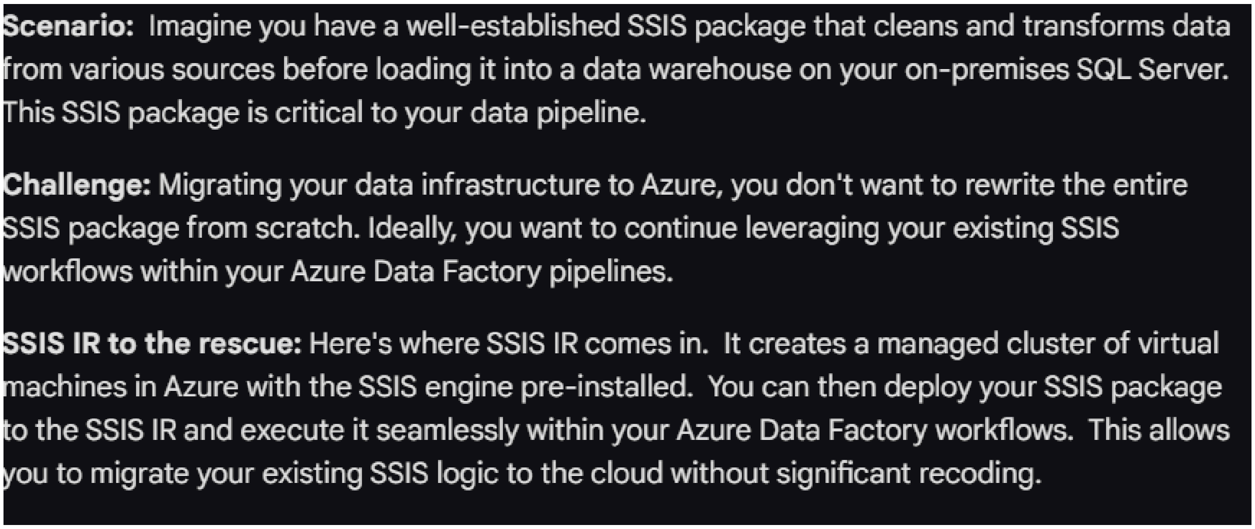
Integration runtime

Compute Infrastructure for Data Integration

Three Types

- Azure IR
- Self-hosted IR
- SSIS IR

3. Azure IR is mainly used to connect to cloud based data sources…like ADF to ADL etc
4. Self hosted IR is mainly used to connect onPerm SQL DB…here we need to install a package on onperm System…then azure can connect with this with the help of package
5. SSIS IR…

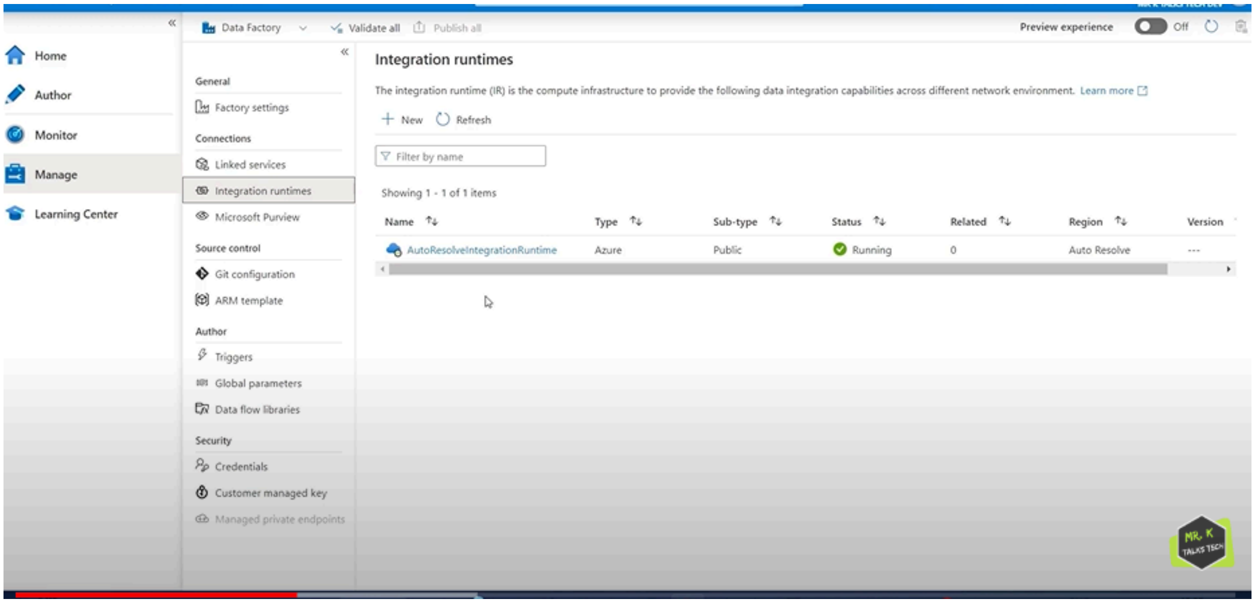


Scenario: Imagine you have a well-established SSIS package that cleans and transforms data from various sources before loading it into a data warehouse on your on-premises SQL Server. This SSIS package is critical to your data pipeline.

Challenge: Migrating your data infrastructure to Azure, you don't want to rewrite the entire SSIS package from scratch. Ideally, you want to continue leveraging your existing SSIS workflows within your Azure Data Factory pipelines.

SSIS IR to the rescue: Here's where SSIS IR comes in. It creates a managed cluster of virtual machines in Azure with the SSIS engine pre-installed. You can then deploy your SSIS package to the SSIS IR and execute it seamlessly within your Azure Data Factory workflows. This allows you to migrate your existing SSIS logic to the cloud without significant recoding.

6. By default here we’ll be having autoResovleIntegration runtime in our ADF



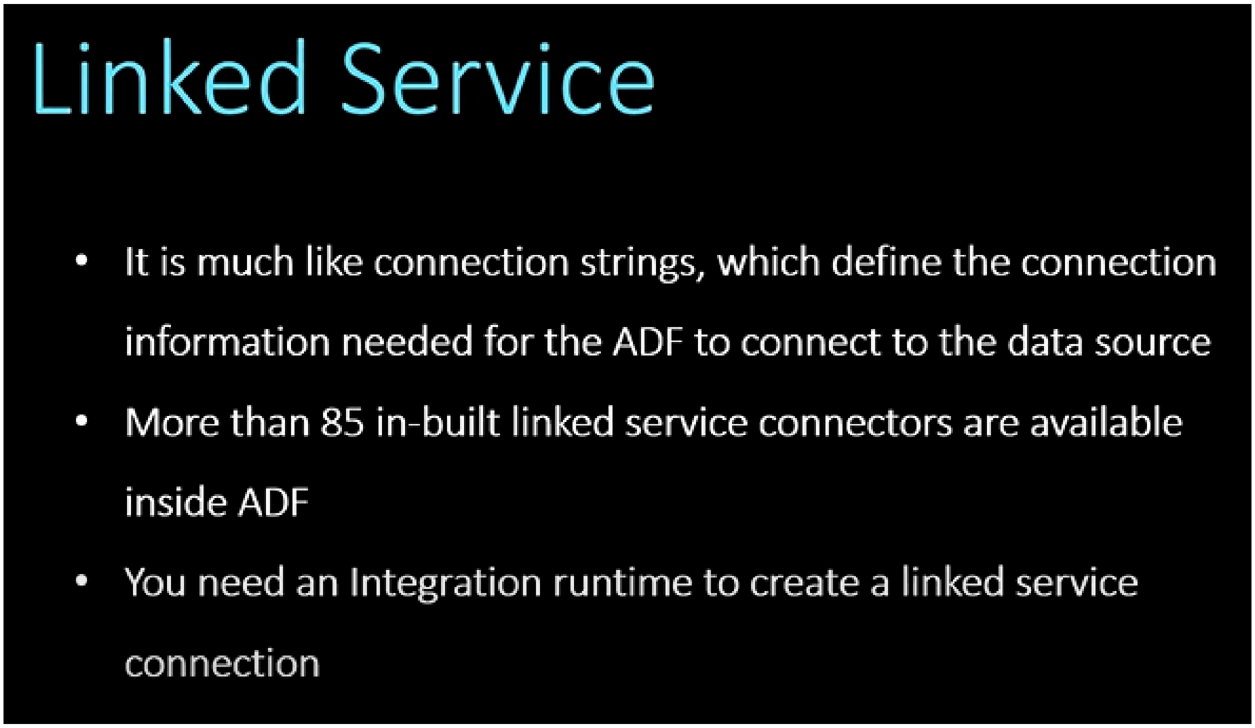
Integration runtimes

The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environment. Learn more

Showing 1 - 1 of 1 items

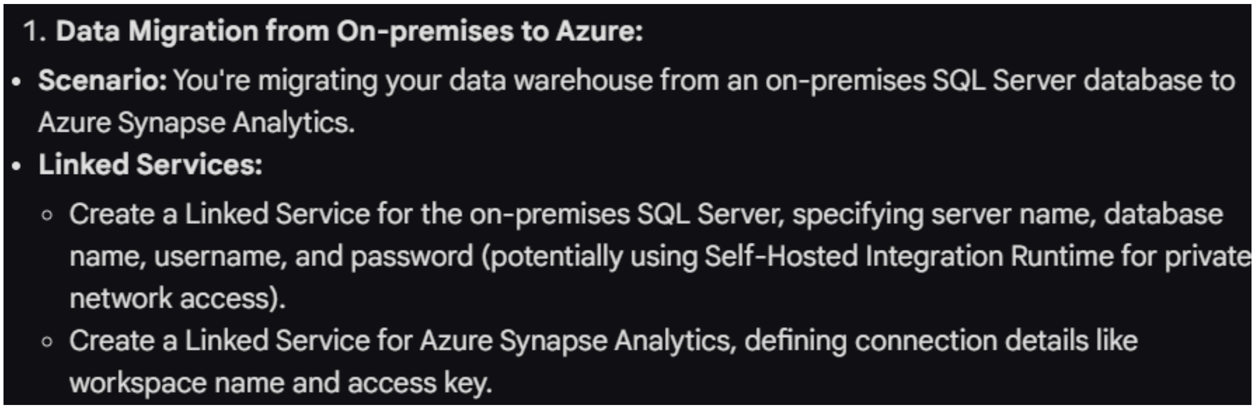
Name	Type	Sub-type	Status	Related	Region	Version
AutoResolveIntegrationRuntime	Azure	Public	✓ Running	0	Auto Resolve	---

7. To create our own Integration runtime ..click in new
8. Next topic is Linked Service

9. 

Linked Service

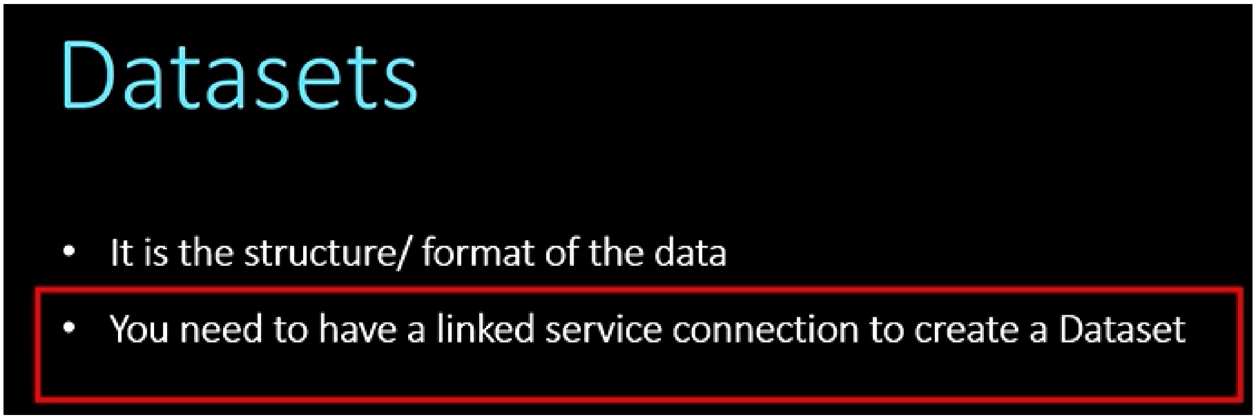
- It is much like connection strings, which define the connection information needed for the ADF to connect to the data source
- More than 85 in-built linked service connectors are available inside ADF
- You need an Integration runtime to create a linked service connection



1. **Data Migration from On-premises to Azure:**

- **Scenario:** You're migrating your data warehouse from an on-premises SQL Server database to Azure Synapse Analytics.
- **Linked Services:**
  - Create a Linked Service for the on-premises SQL Server, specifying server name, database name, username, and password (potentially using Self-Hosted Integration Runtime for private network access).
  - Create a Linked Service for Azure Synapse Analytics, defining connection details like workspace name and access key.

10. Datasets

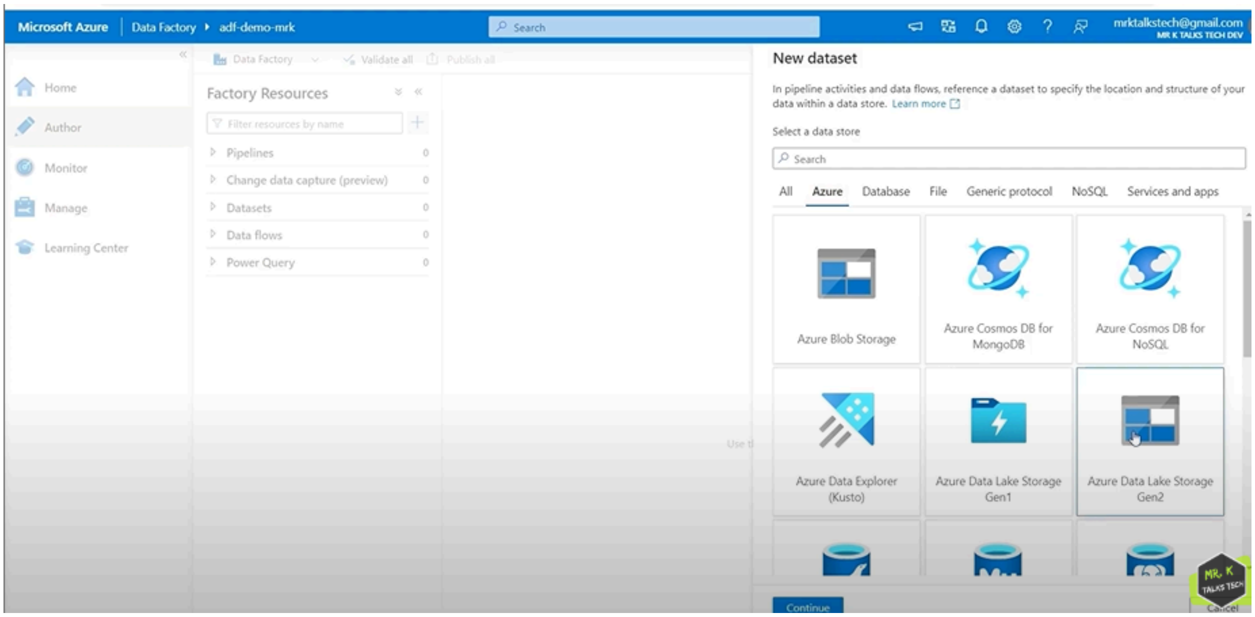
11. 

Datasets

- It is the structure/ format of the data
- You need to have a linked service connection to create a Dataset

12. Lets see how we can create datasets in azure

13. To create a dataset ..go to author tab

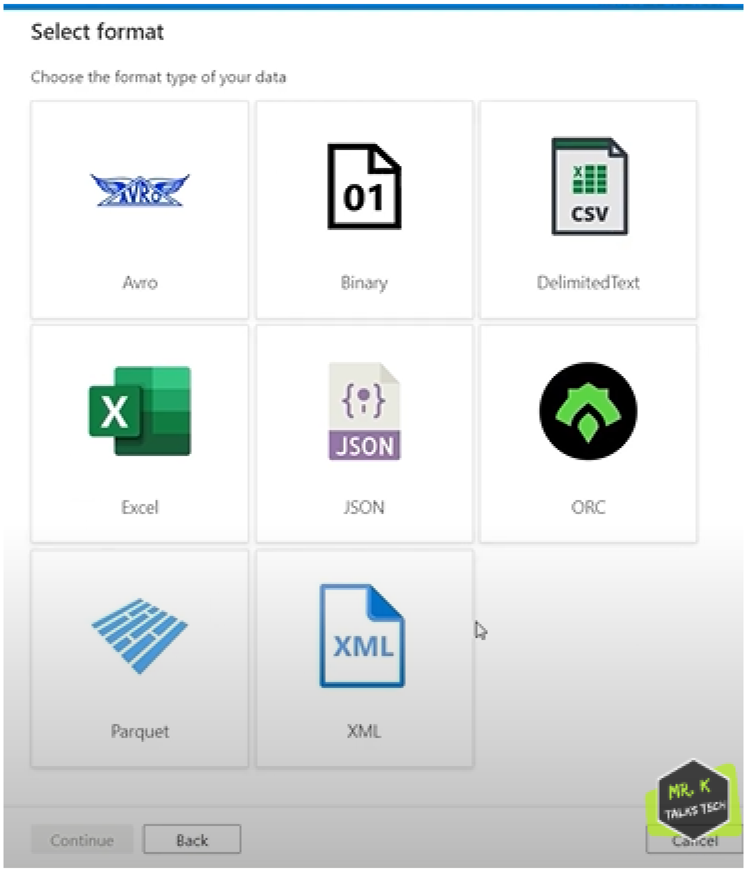


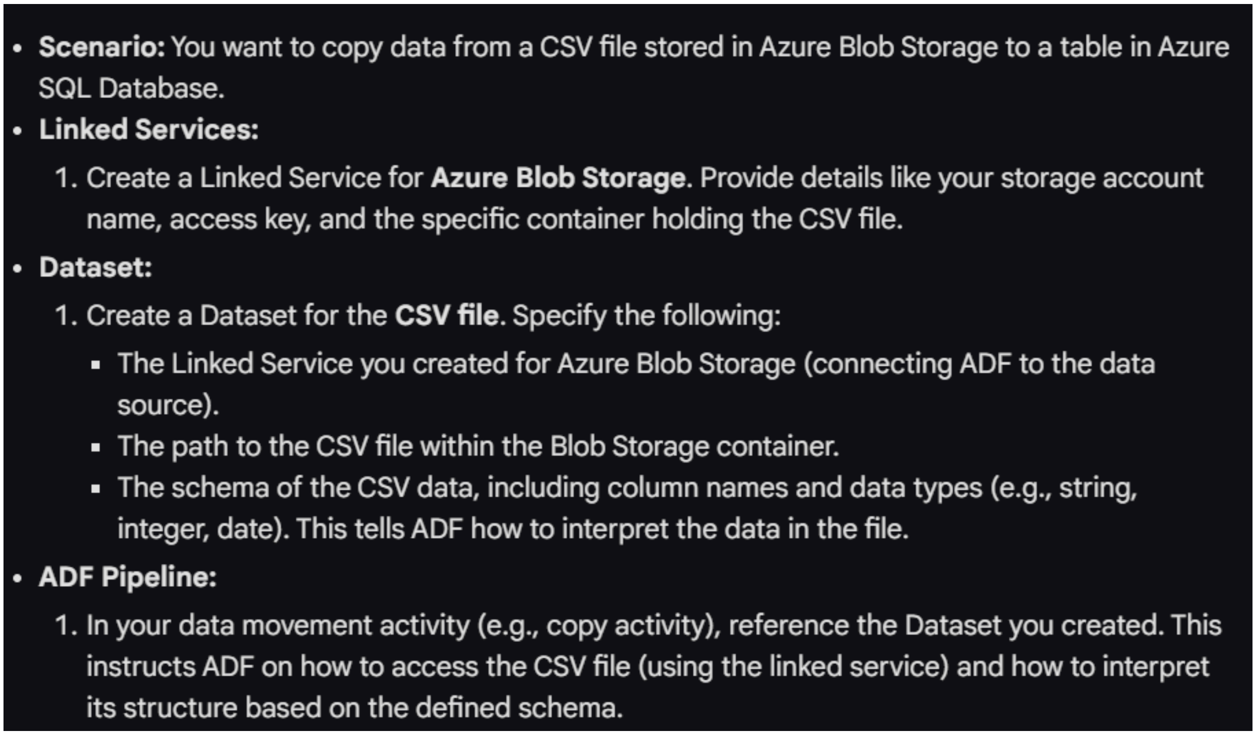
New dataset

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

first we need to specify data source…next we will choose the formar





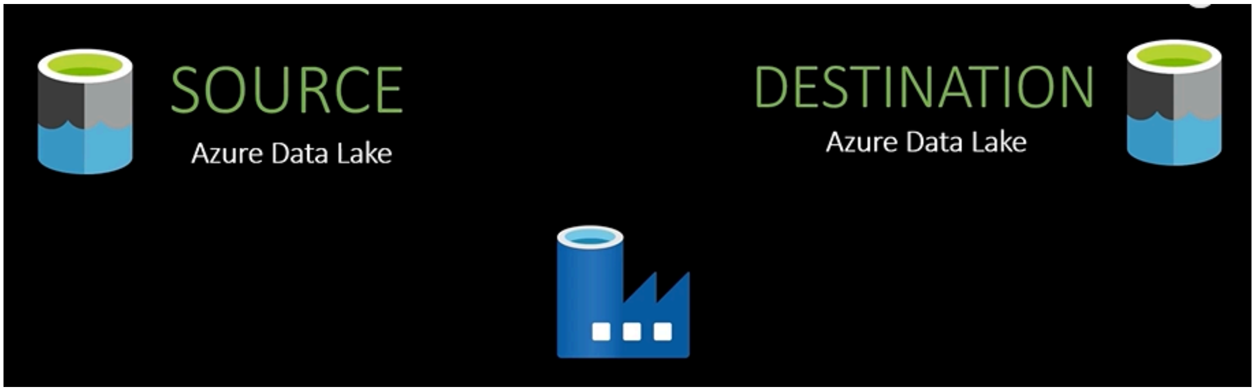
- **Scenario:** You want to copy data from a CSV file stored in Azure Blob Storage to a table in Azure SQL Database.
- **Linked Services:**
  1. Create a Linked Service for **Azure Blob Storage**. Provide details like your storage account name, access key, and the specific container holding the CSV file.
- **Dataset:**
  1. Create a Dataset for the **CSV file**. Specify the following:
     - The Linked Service you created for Azure Blob Storage (connecting ADF to the data source).
     - The path to the CSV file within the Blob Storage container.
     - The schema of the CSV data, including column names and data types (e.g., string, integer, date). This tells ADF how to interpret the data in the file.
- **ADF Pipeline:**
  1. In your data movement activity (e.g., copy activity), reference the Dataset you created. This instructs ADF on how to access the CSV file (using the linked service) and how to interpret its structure based on the defined schema.

14. Lets understand this with a better example

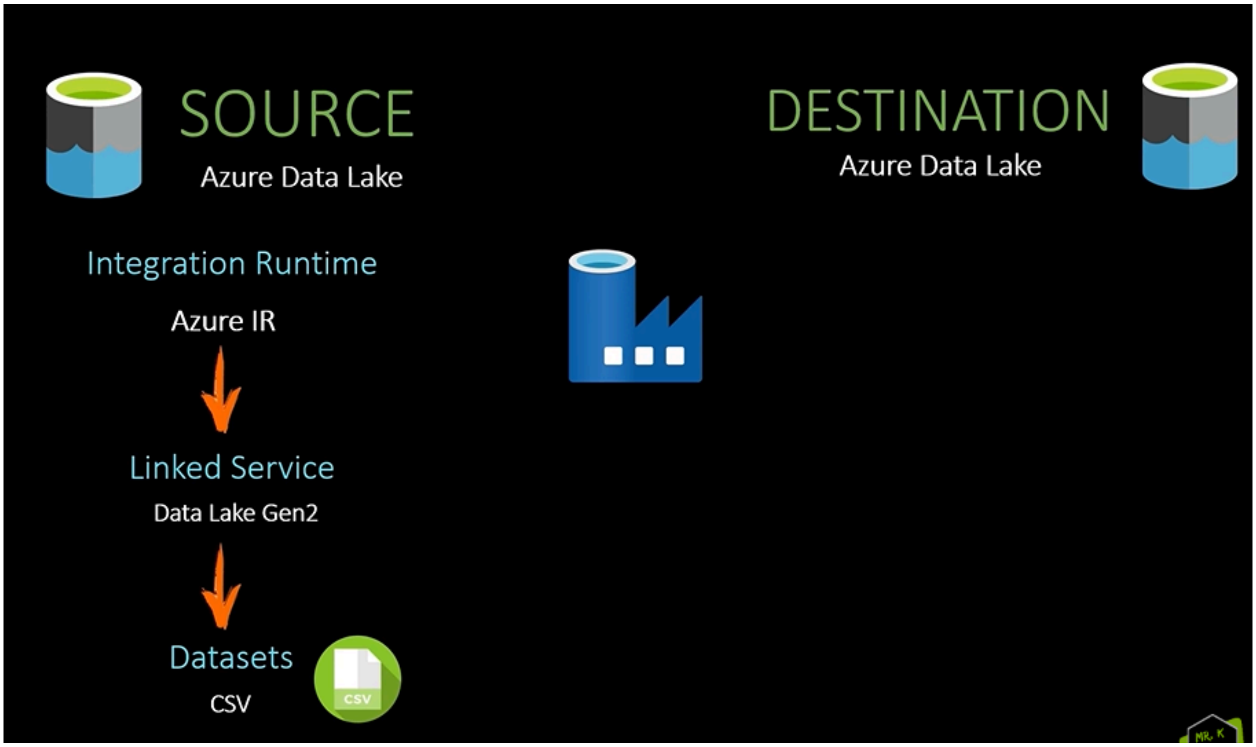
15. Consider we have source and destination

SOURCE
Azure Data Lake
DESTINATION
Azure Data Lake

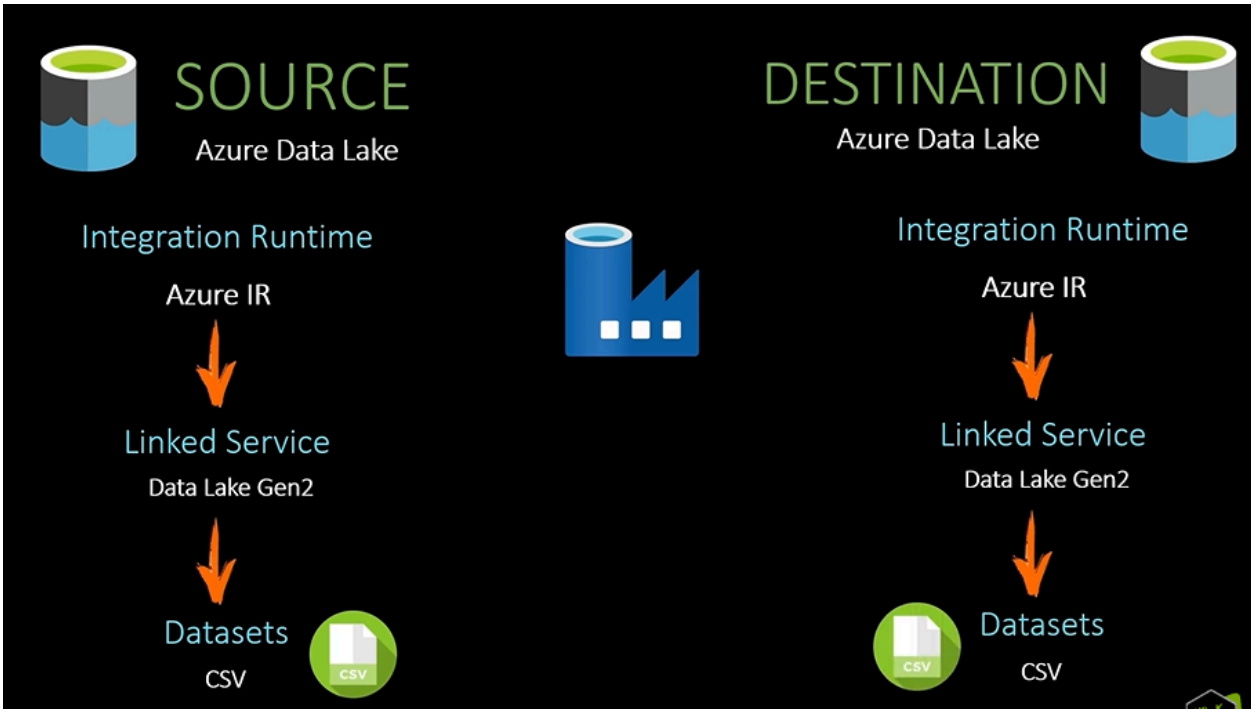
16. To perform this we’ll use ADF to connect the source and destination



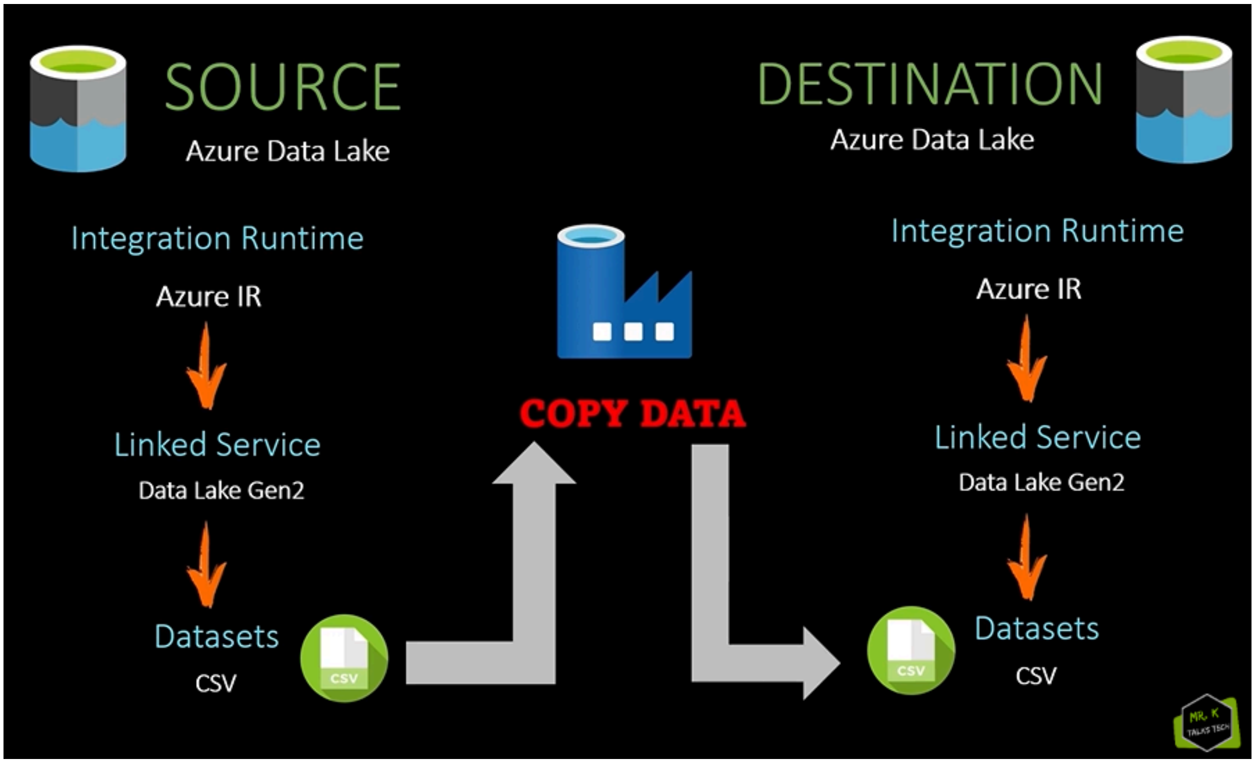
17. Here first our ADF will get data from the source
18. And first our ADF must connect to the source..to make that connection…we create Azure IR…and then next we create linked servie which is data lake gen2 …then next we create datasets..and here we need to specify the data format of Source



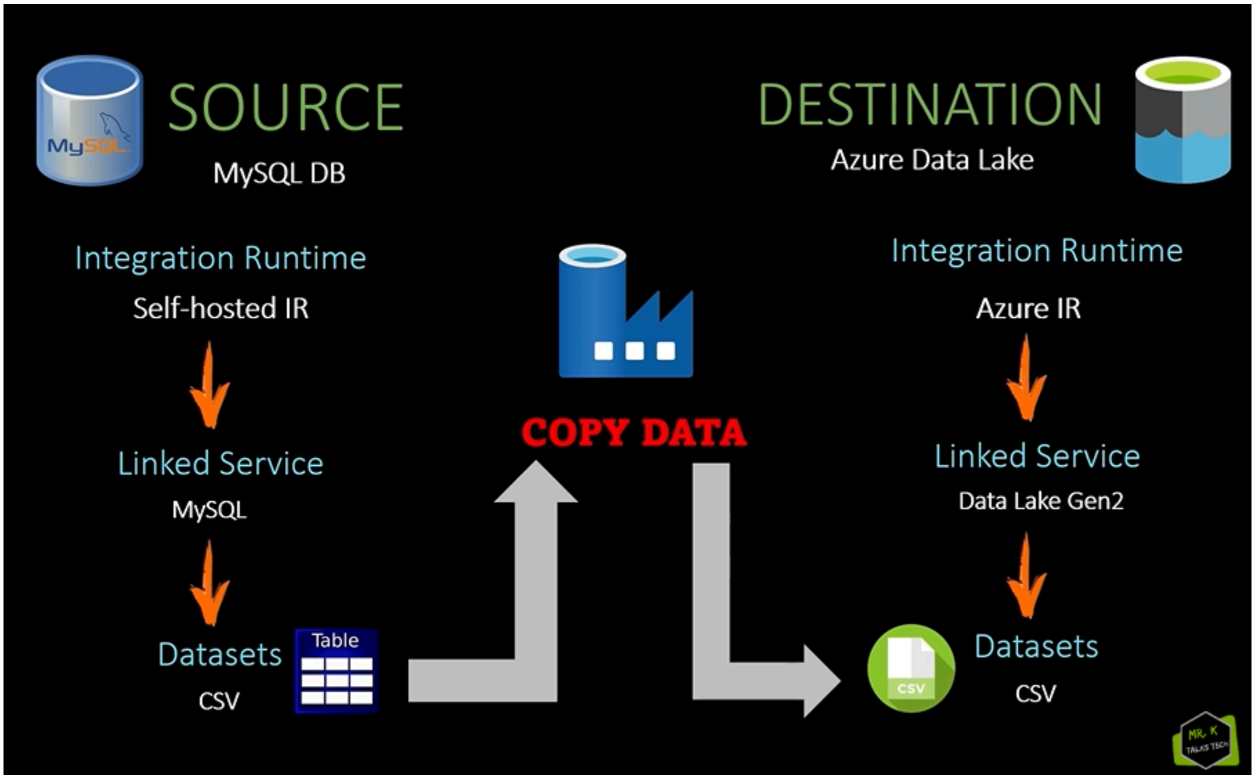
19. Next we have to connect ADF with destination ..and then we can execute copy activity in ADF



20. Now in copy activity we mention source and destination…and we execute this pipeline

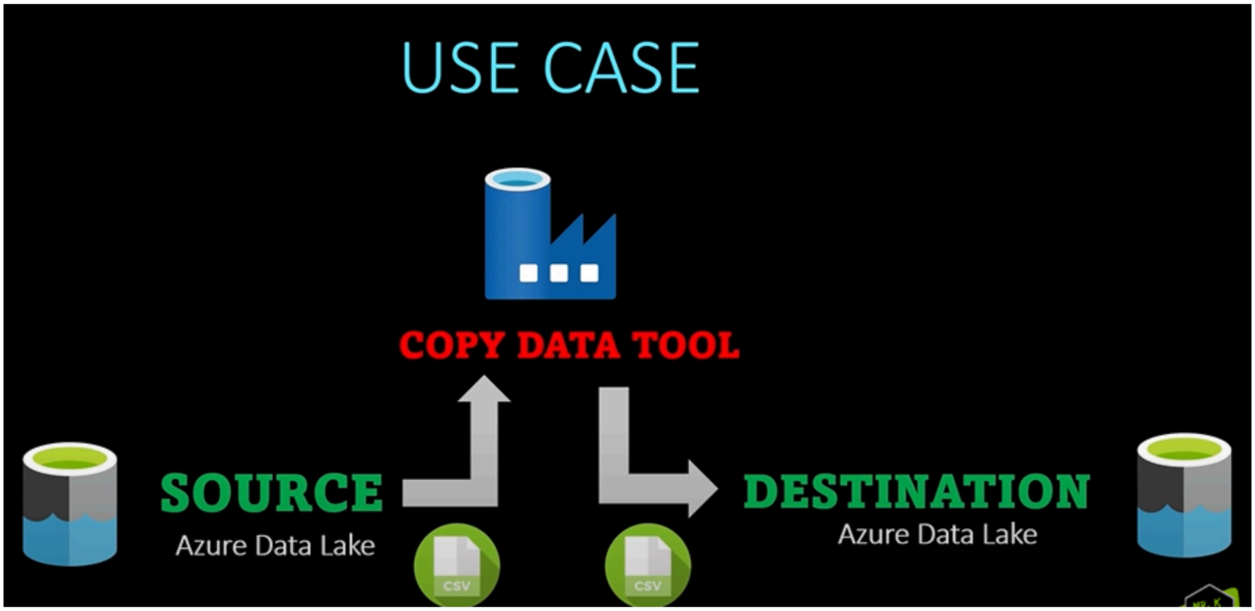


21. If source is onperm SQL DB



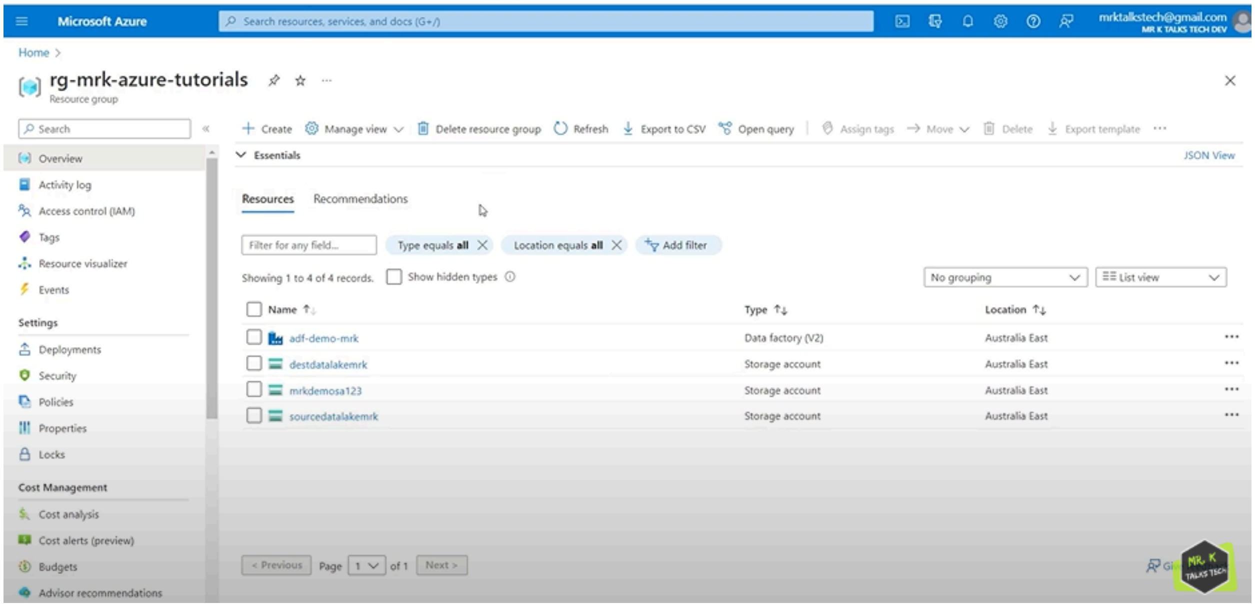
Copy data using ADF copy data tool

1. Here what we'll do is



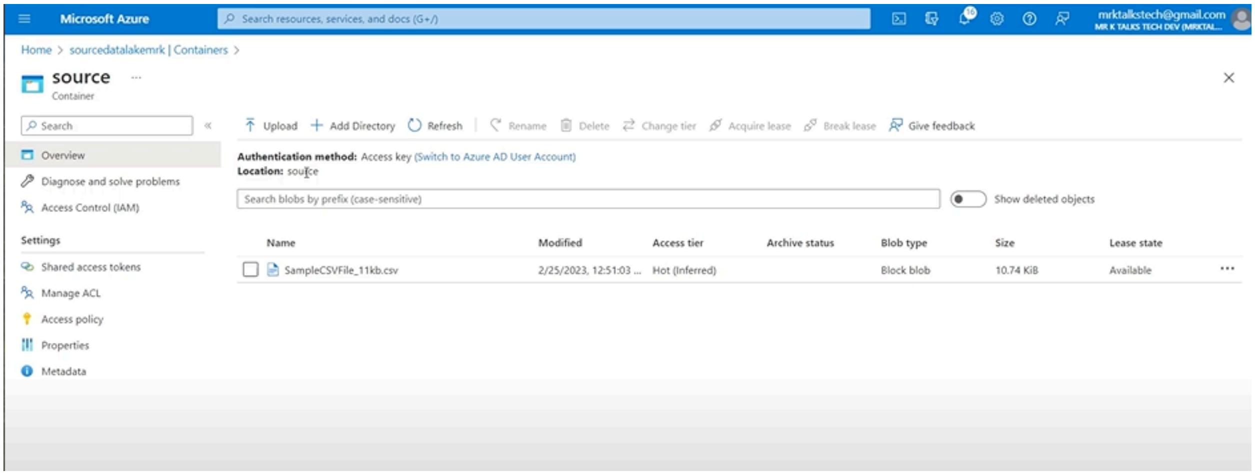
copy data from one ADL and paste it in 2nd ADL

2. Here in our resource group..we have 2 data lakes



Name	Type	Location
adf-demo-mrk	Data factory (V2)	Australia East
destdatalakemrk	Storage account	Australia East
mrkdemosa123	Storage account	Australia East
sourcedatalakemrk	Storage account	Australia East

3. And in the source data lake we have sample csv file

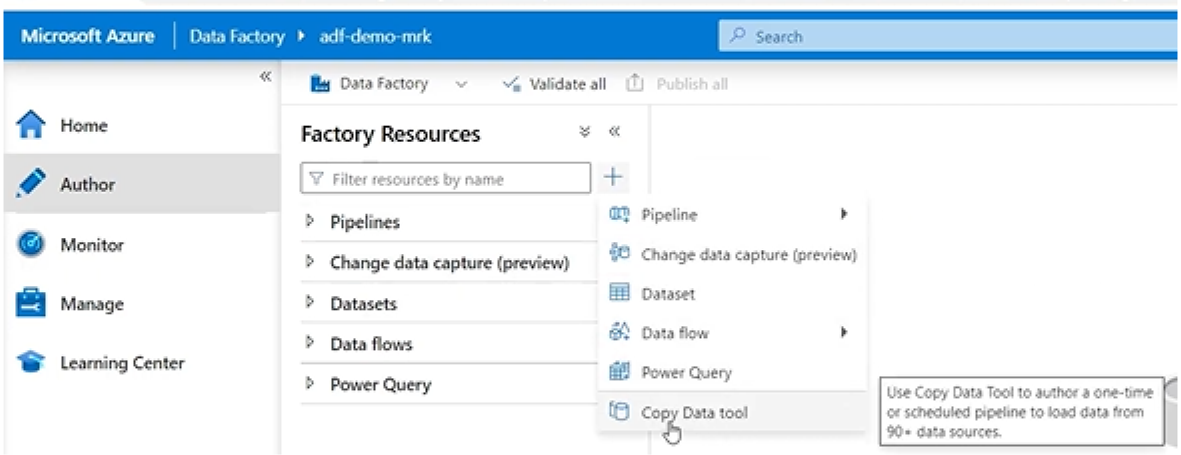


Authentication method: Access key (Switch to Azure AD User Account)
Location: source

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
SampleCSVFile_11kb.csv	2/25/2023, 12:51:03 ...	Hot (Inferred)		Block blob	10.74 KiB	Available

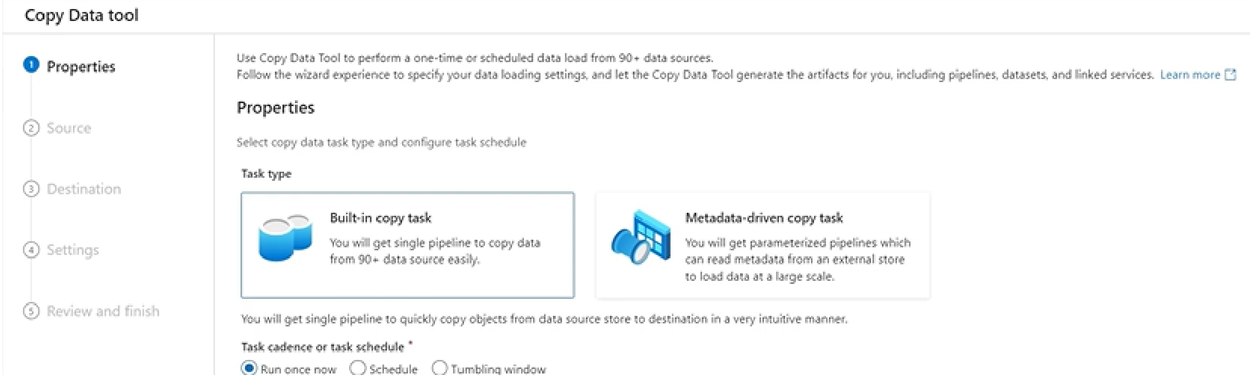
4. And in dest datalake we dont have any files
5. Lets go to ADF

6. Now click on copy data tool in ADF



Use Copy Data Tool to author a one-time or scheduled pipeline to load data from 90+ data sources.

7. In ADF the most commonly used activity is copy data..to transfer the data from one loc to another loc



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. Learn more

Properties

Select copy data task type and configure task schedule

Task type

Built-in copy task: You will get single pipeline to copy data from 90+ data source easily.

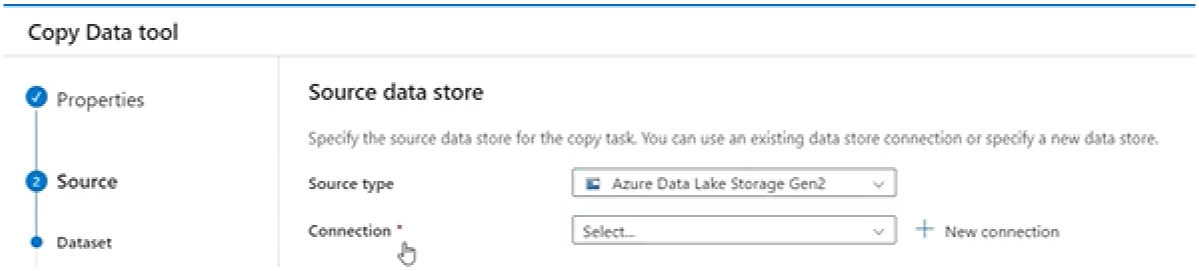
Metadata-driven copy task: You will get parameterized pipelines which can read metadata from an external store to load data at a large scale.

You will get single pipeline to quickly copy objects from data source store to destination in a very intuitive manner.

Task cadence or task schedule *
● Run once now ○ Schedule ○ Tumbling window

here we'll be choosing built-in copy task and run the pipeline once

8. Now next we have to specify the data source



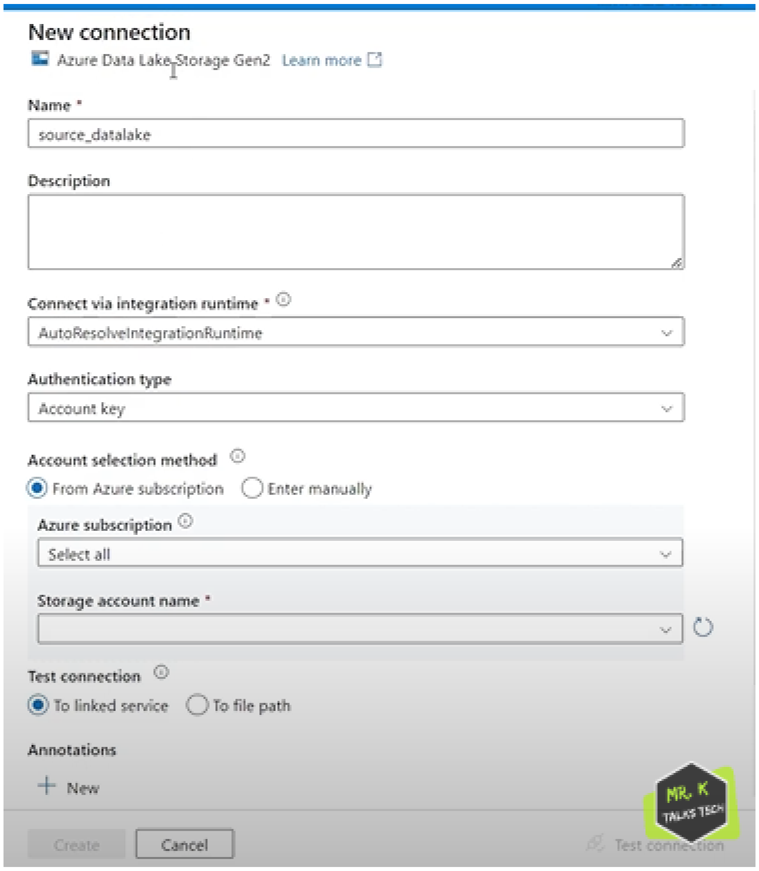
Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store.

Source type: Azure Data Lake Storage Gen2

Connection *: Select...

and in the connection..we have to specify Azure IR



New connection

Azure Data Lake Storage Gen2 Learn more

Name *

source_datalake

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Authentication type

Account key

Account selection method

● From Azure subscription ○ Enter manually

Azure subscription

Select all

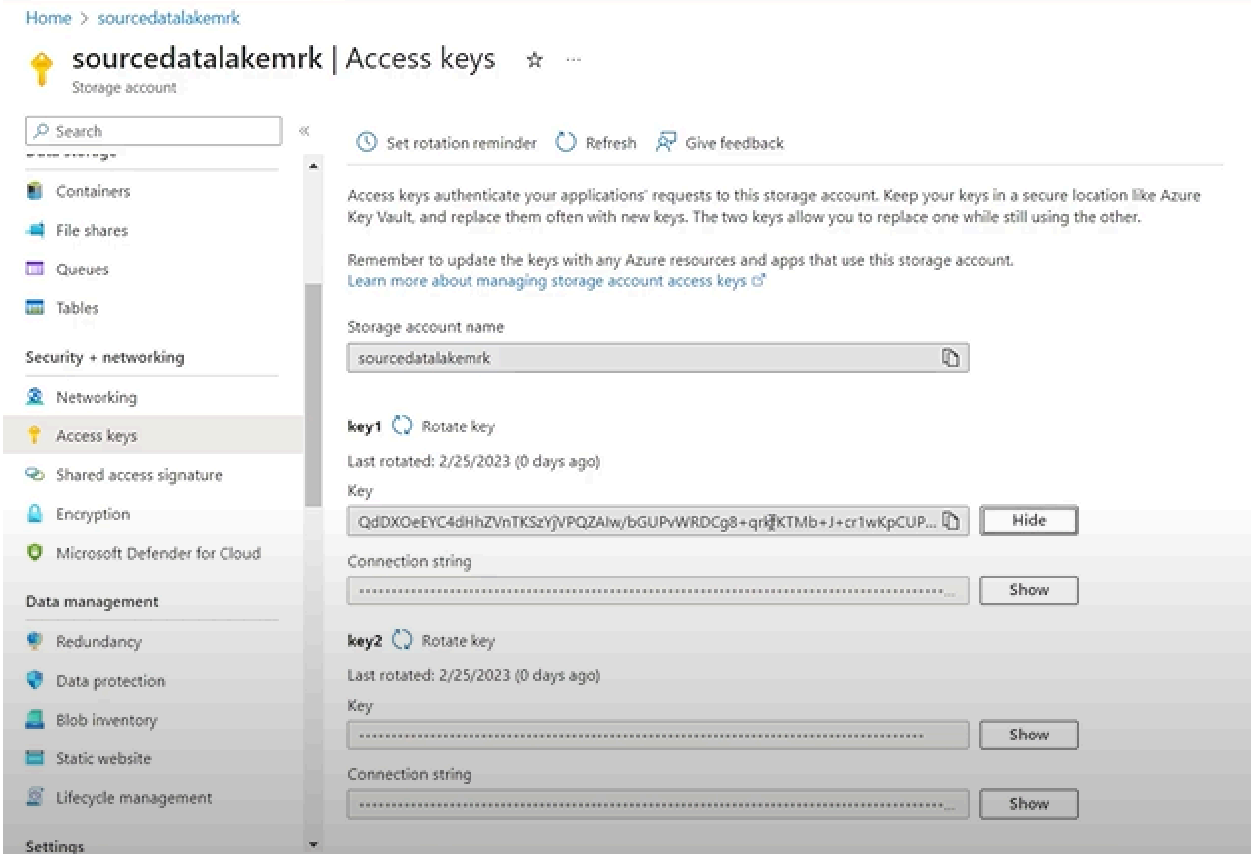
Storage account name *

Test connection

● To linked service ○ To file path

Annotations

9. In the authentication type we choose account key(here every ADL has a key



Home > sourcedatalakemrk

sourcedatalakemrk | Access keys ☆ ···

Storage account

Access keys authenticate your applications' requests to this storage account. Keep your keys in a secure location like Azure Key Vault, and replace them often with new keys. The two keys allow you to replace one while still using the other.

Remember to update the keys with any Azure resources and apps that use this storage account.
Learn more about managing storage account access keys

Storage account name: sourcedatalakemrk

key1 Rotate key

Last rotated: 2/25/2023 (0 days ago)

Key: QdDXOeEYC4dHhZVnTKSzYjVPQZAIw/bGUPvWRDCg8+qr[illegible]KTMb+J+cr1wKpCUP... Hide

Connection string: ••••••••••••••••••••••••••••••... Show

key2 Rotate key

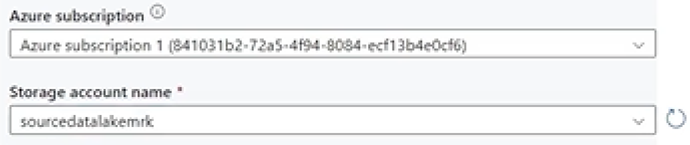
Last rotated: 2/25/2023 (0 days ago)

Key: •••••••••••••••••••••••••••••• Show

Connection string: ••••••••••••••••••••••••••••••... Show

)..now our ADF will connect to ADL using these keys

10. Next we need to specify this

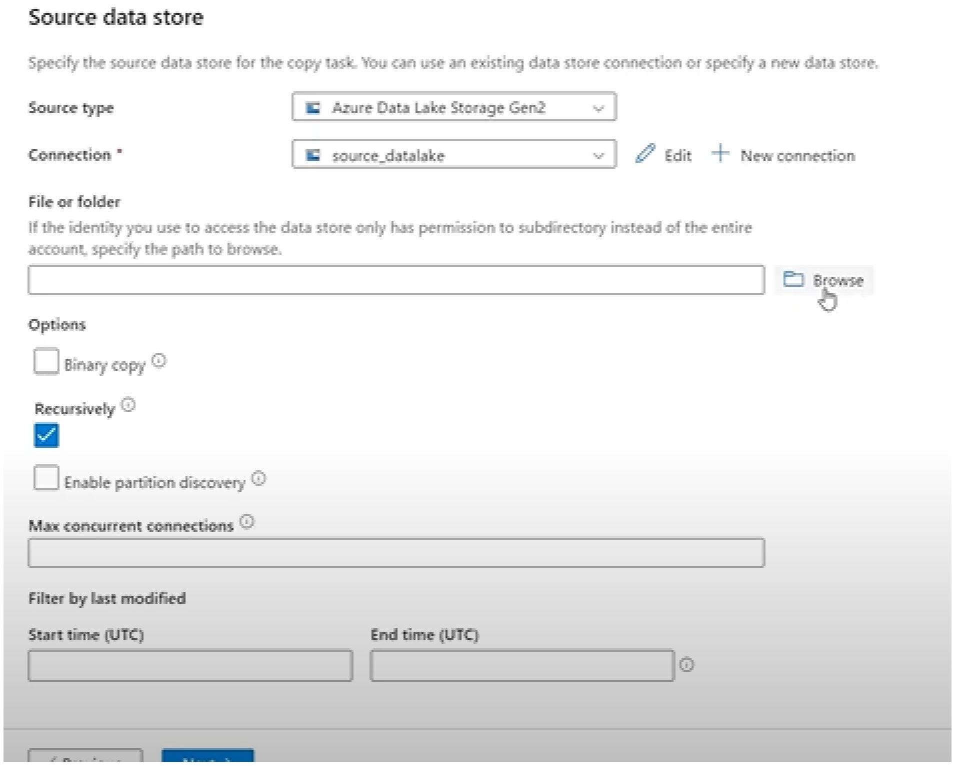


Azure subscription: Azure subscription 1 (841031b2-72a5-4f94-8084-ecf13b4e0cf6)

Storage account name *: sourcedatalakemrk

11. Next we test our connection and create this connection
12. Here we have create Azure IR, and linked service…next we have to create a dataset

13. For that first step would be selecting the csv file



Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store.

Source type: Azure Data Lake Storage Gen2

Connection *: source_datalake

File or folder

If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to browse.

Options

□ Binary copy

Recursively

☑

□ Enable partition discovery

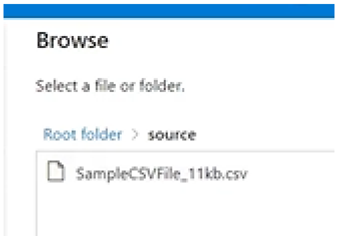
Max concurrent connections

Filter by last modified

Start time (UTC)

End time (UTC)

click on



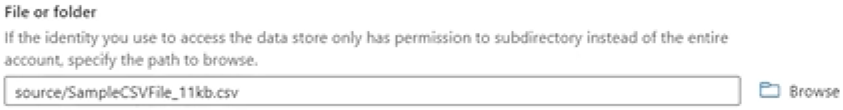
Browse

Select a file or folder.

Root folder > source

SampleCSVFile_11kb.csv

browse ..go to source and click on our file

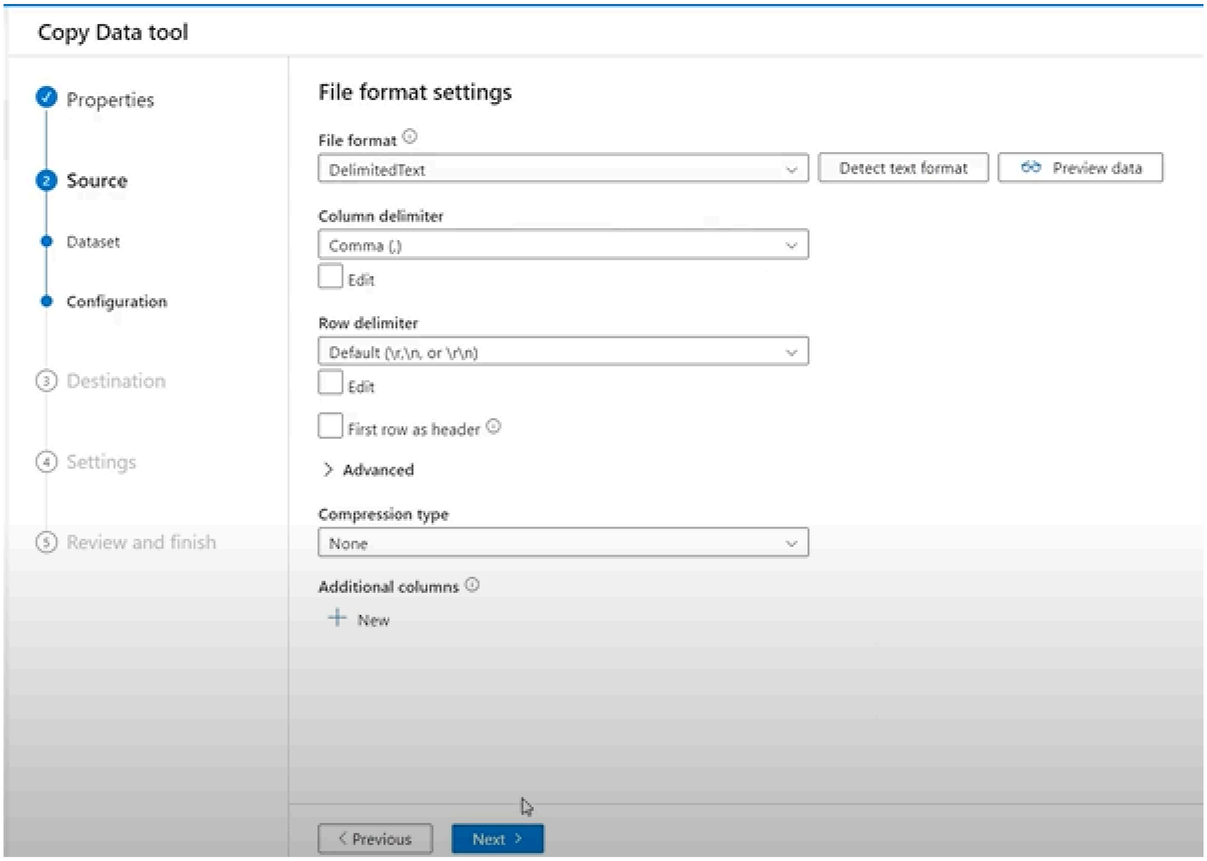


File or folder

If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to browse.

source/SampleCSVFile_11kb.csv

14. Next we need to specify the file format of source


Copy Data tool

File format settings

File format: DelimitedText

Column delimiter: Comma (,)

□ Edit

Row delimiter: Default (\r,\n, or \r\n)

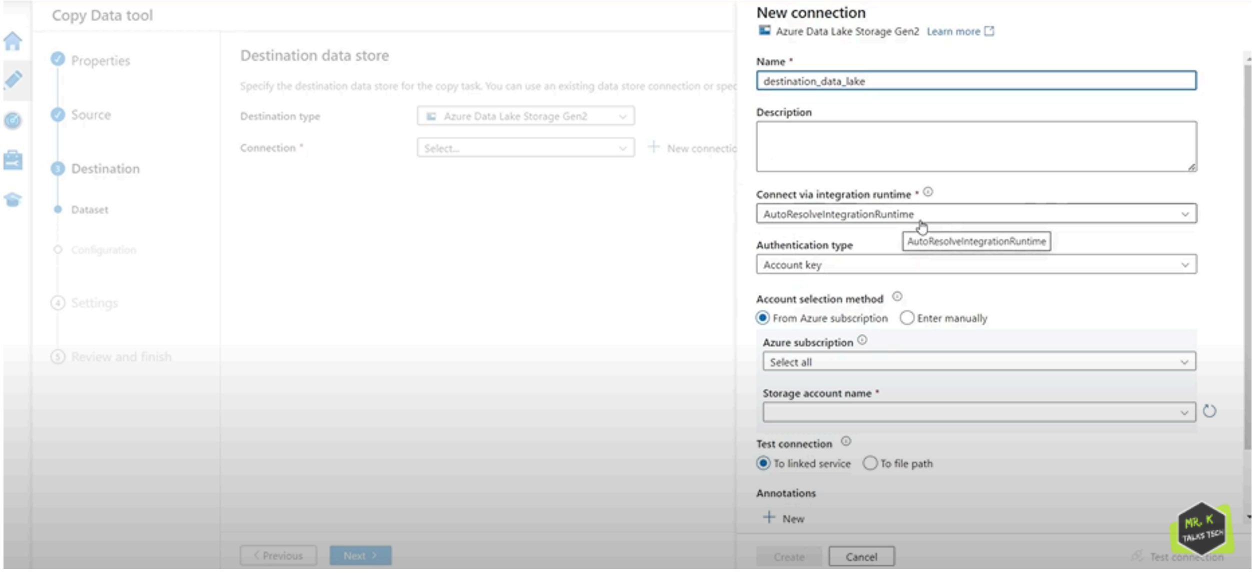
□ Edit

□ First row as header

Compression type: None

Additional columns

15. Next we setup the destination



Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or spec

Destination type: Azure Data Lake Storage Gen2

Connection *: Select... + New connectio

New connection

Azure Data Lake Storage Gen2 Learn more

Name *: destination_data_lake

Description:

Connect via integration runtime *: AutoResolveIntegrationRuntime

Authentication type: Account key

Account selection method: ● From Azure subscription ○ Enter manually

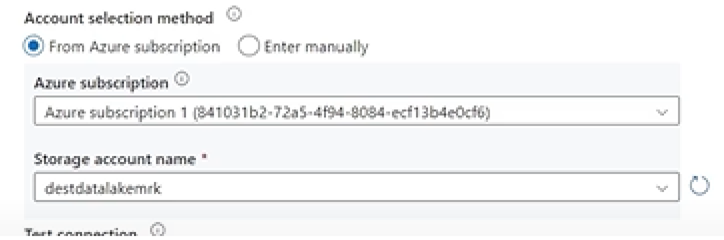
Azure subscription: Select all

Storage account name *:

Test connection: ● To linked service ○ To file path

Annotations

+ New



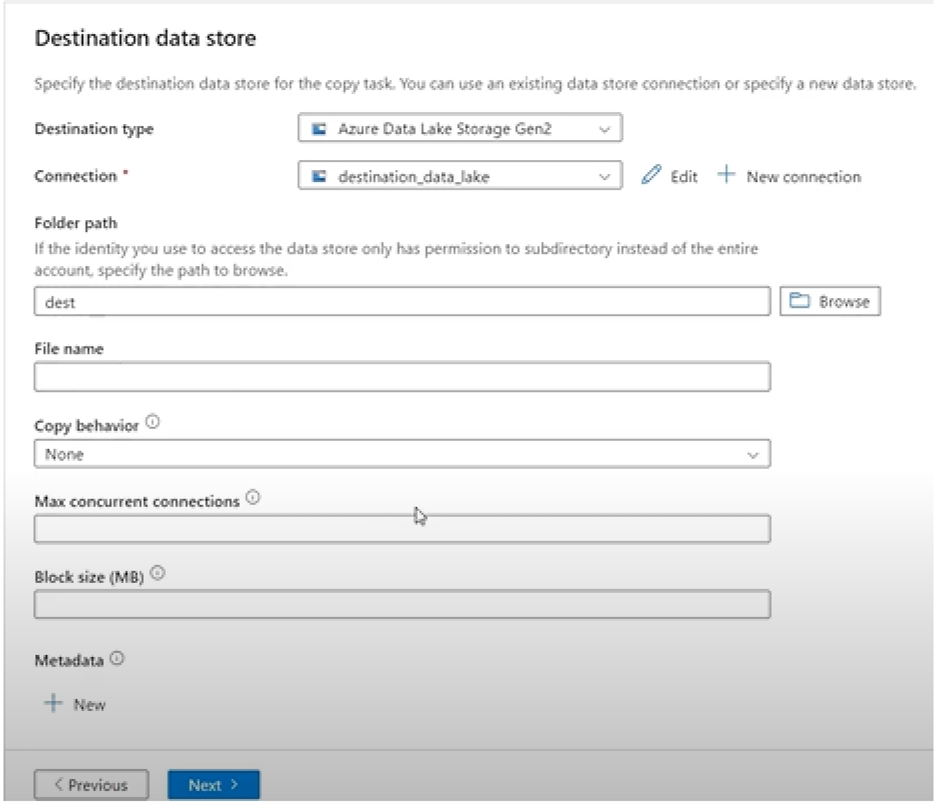
Account selection method: ● From Azure subscription ○ Enter manually

Azure subscription: Azure subscription 1 (841031b2-72a5-4f94-8084-ecf13b4e0cf6)

Storage account name *: destdatalakemrk

Test connection

16. Next we'll specify the folder path ..to where our data gets dumper



Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.

Destination type: Azure Data Lake Storage Gen2

Connection *: destination_data_lake

Folder path

If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to browse.

dest

File name:

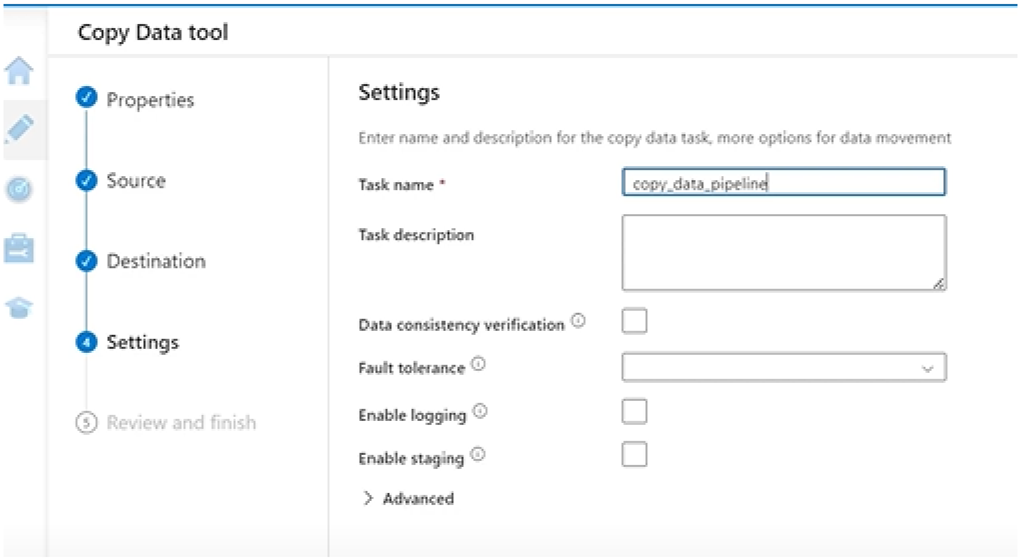
Copy behavior: None

Max concurrent connections:

Block size (MB):

Metadata

17. Next we configure file format ..which is csv
18. Now we give a name to our task



Copy Data tool

Settings

Enter name and description for the copy data task, more options for data movement

Task name * copy_data_pipeline

Task description

Data consistency verification □

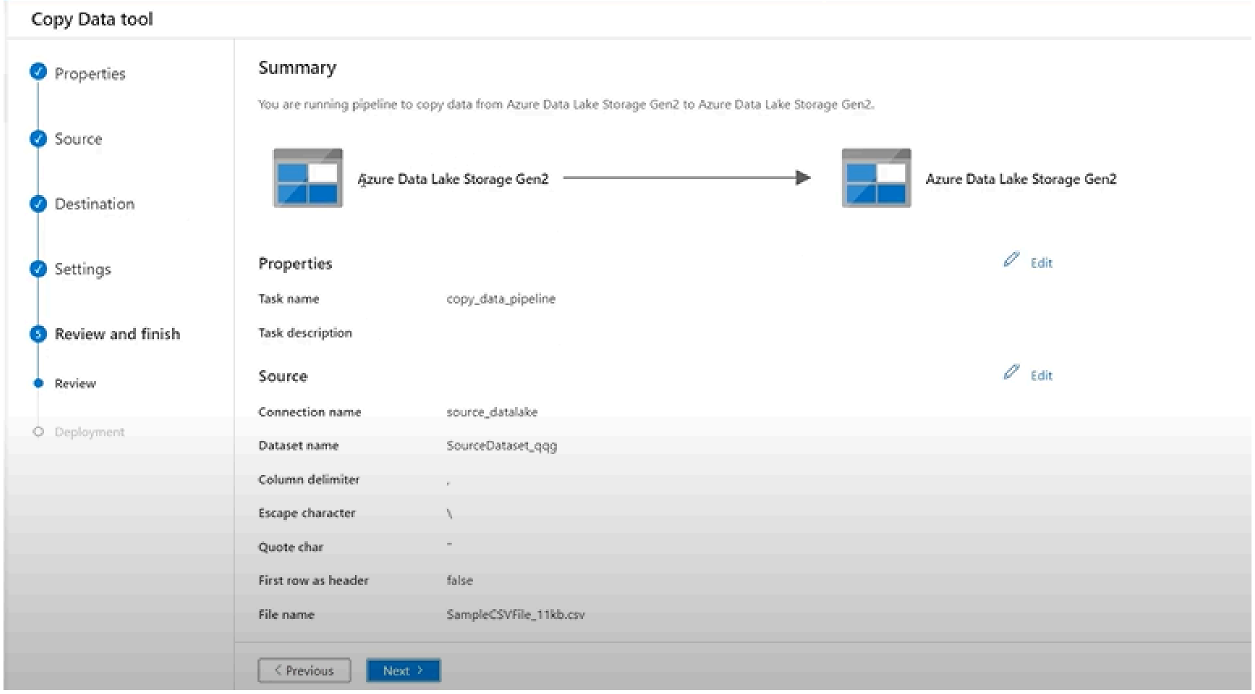
Fault tolerance

Enable logging □

Enable staging □

> Advanced

19. Summary of what we have created



Copy Data tool

Summary

You are running pipeline to copy data from Azure Data Lake Storage Gen2 to Azure Data Lake Storage Gen2.

Properties

Task name	copy_data_pipeline
Task description	

Source

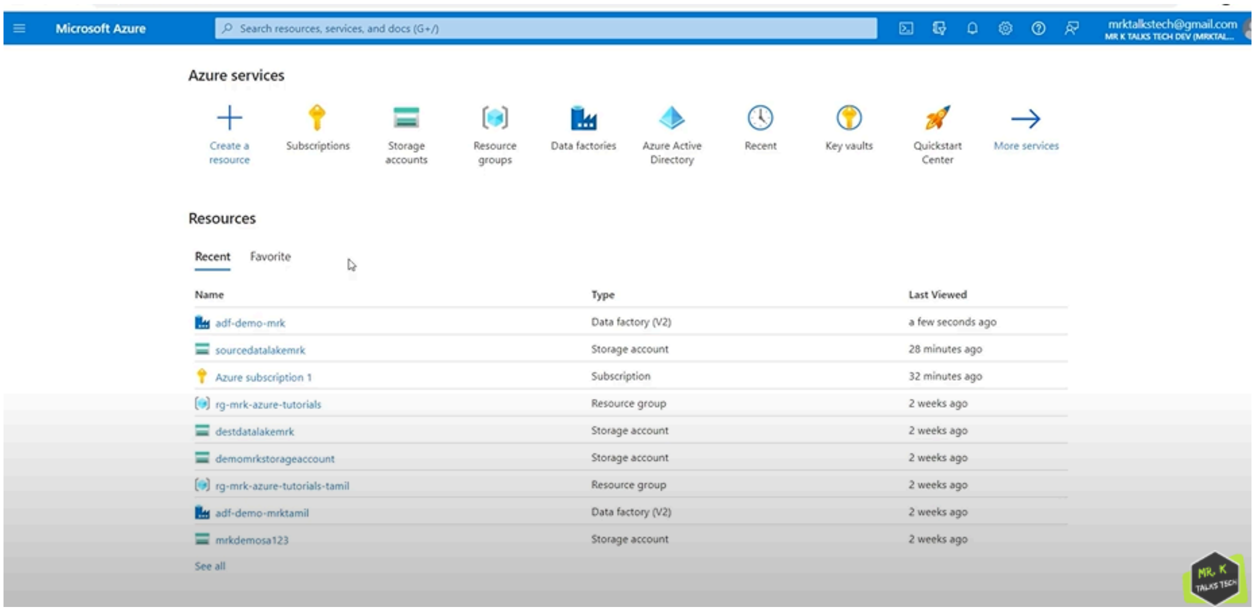
Connection name	source_datalake
Dataset name	SourceDataset_qqg
Column delimiter	,
Escape character	\
Quote char	"
First row as header	false
File name	SampleCSVFile_11kb.csv

20. Now if we deploy our copy_data_pipeline..then we can our file in dest
21. Next we'll learn how we can do this copy data from scratch

Create a copy data pipeline from scratch

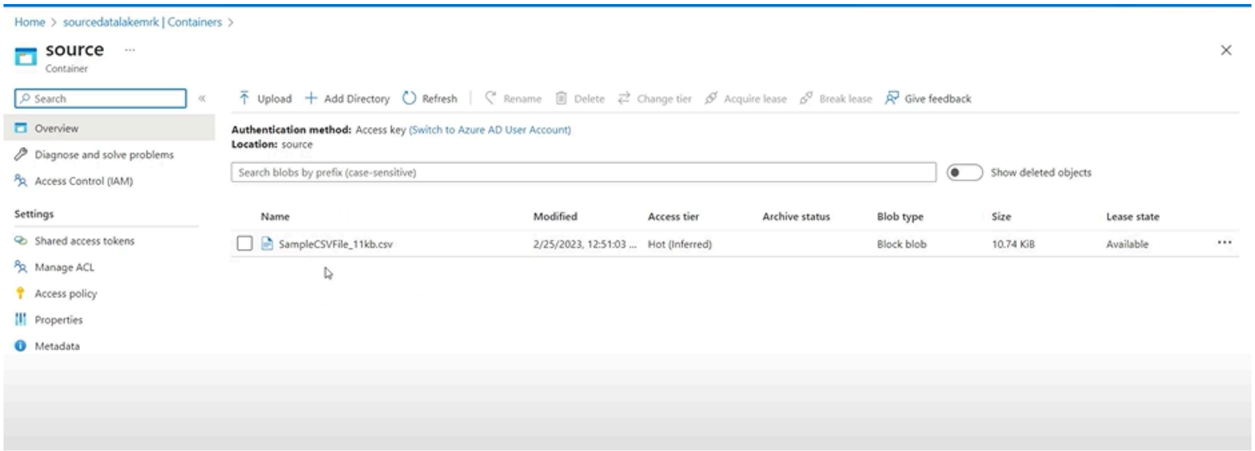
1. Here we'll create a datapipeline from scratch

2. We'll be using the same use case



Name	Type	Last Viewed
adf-demo-mrk	Data factory (V2)	a few seconds ago
sourcedatalakemrk	Storage account	28 minutes ago
Azure subscription 1	Subscription	32 minutes ago
rg-mrk-azure-tutorials	Resource group	2 weeks ago
destdatalakemrk	Storage account	2 weeks ago
demomrkstorageaccount	Storage account	2 weeks ago
rg-mrk-azure-tutorials-tamil	Resource group	2 weeks ago
adf-demo-mrktamil	Data factory (V2)	2 weeks ago
mrkdemosa123	Storage account	2 weeks ago

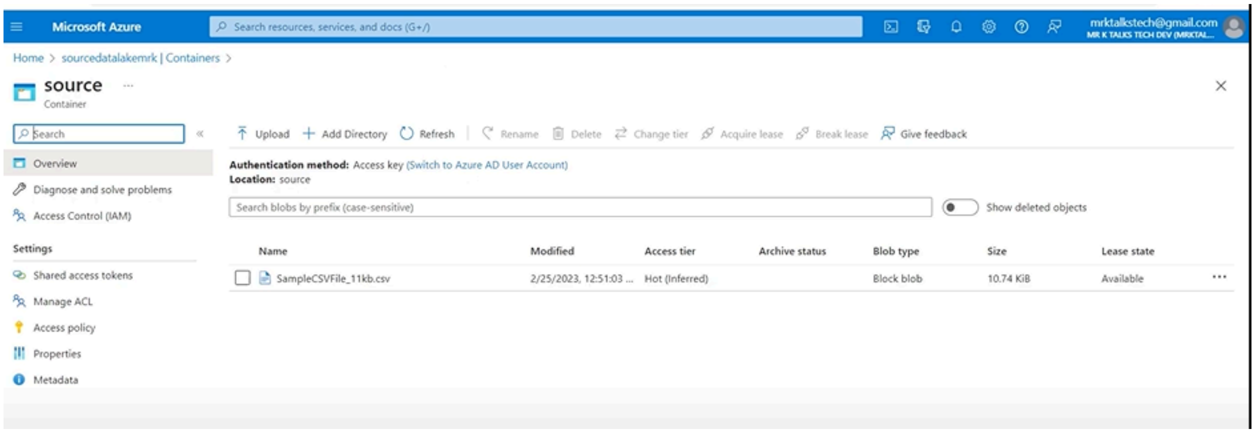
3. We have sample file in our source ADL container



Authentication method: Access key (Switch to Azure AD User Account)
Location: source

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
SampleCSVFile_11kb.csv	2/25/2023, 12:51:03 ...	Hot (Inferred)		Block blob	10.74 KiB	Available

4. We'll copy this file and dump it in dest container in dest DL

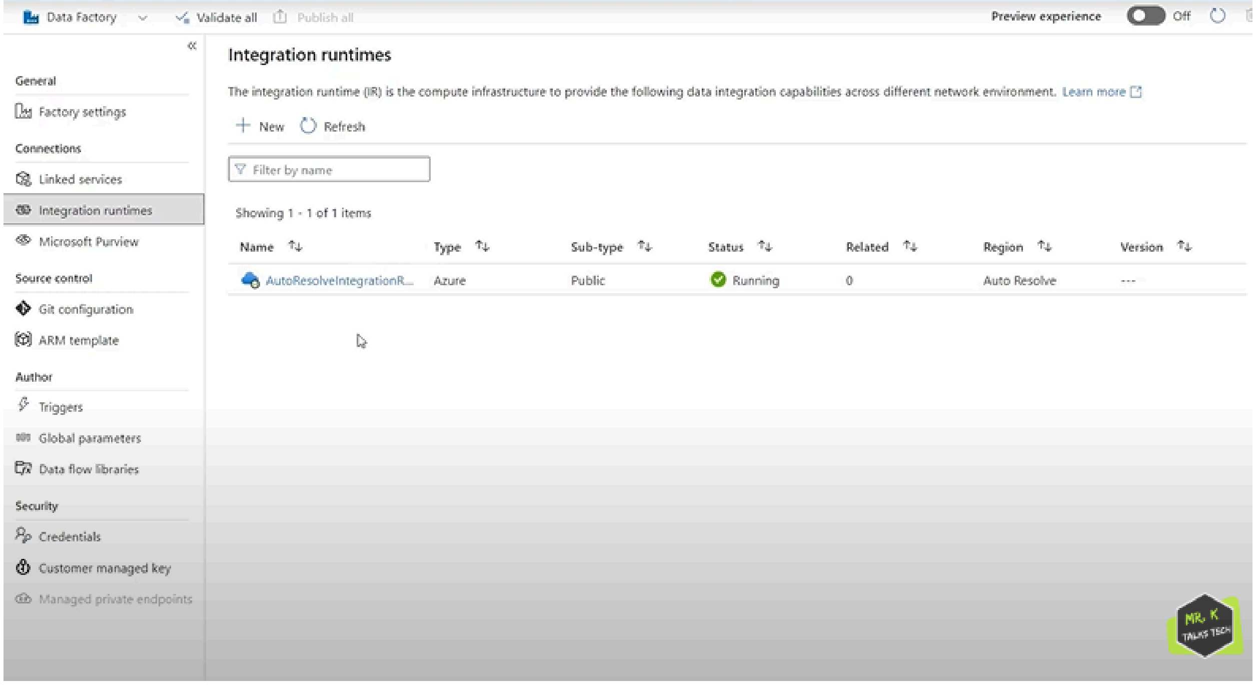


Authentication method: Access key (Switch to Azure AD User Account)
Location: source

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
SampleCSVFile_11kb.csv	2/25/2023, 12:51:03 ...	Hot (Inferred)		Block blob	10.74 KiB	Available

5. Here we already have one copy data pipeline in our ADF..which has been created by copy data activity

6. Lets create this pipeline from scratch
7. Now we'll use AZURE IR to create a runtime



Integration runtimes

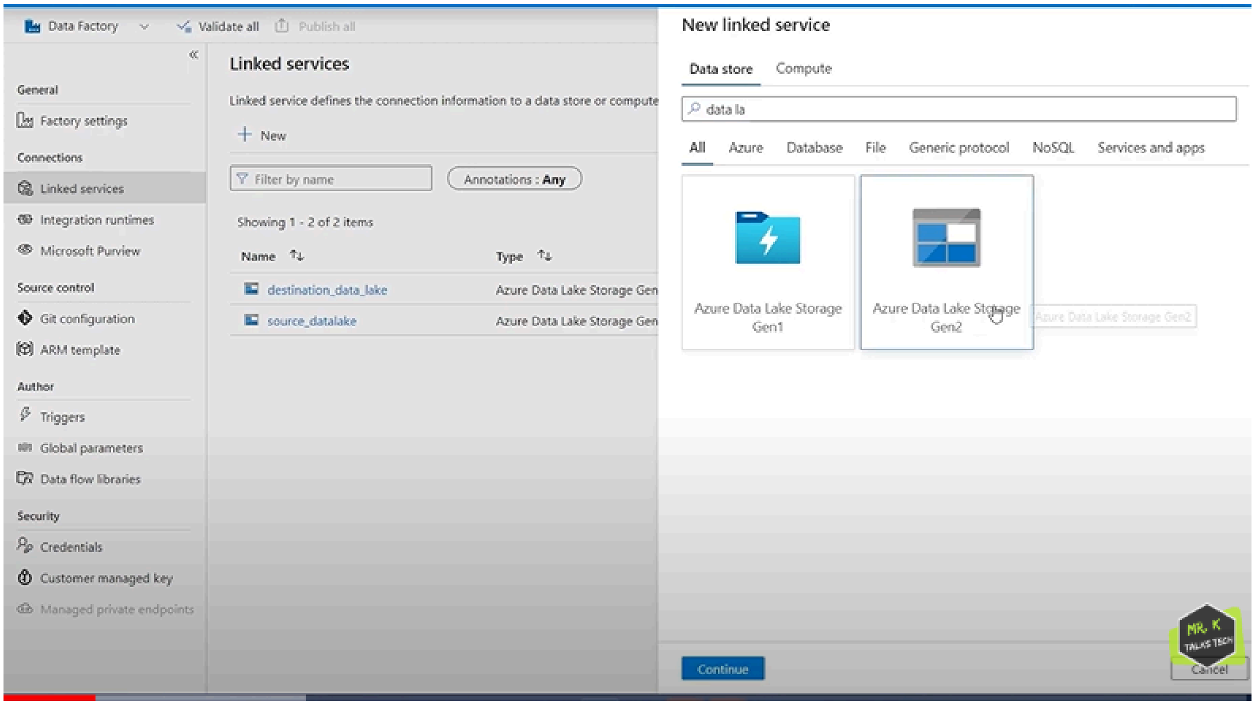
The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environment. Learn more

Showing 1 - 1 of 1 items

Name	Type	Sub-type	Status	Related	Region	Version
AutoResolveIntegrationR...	Azure	Public	Running	0	Auto Resolve	---

so here we use the default one

8. Next we need to create a linked service



Linked services

Linked service defines the connection information to a data store or compute

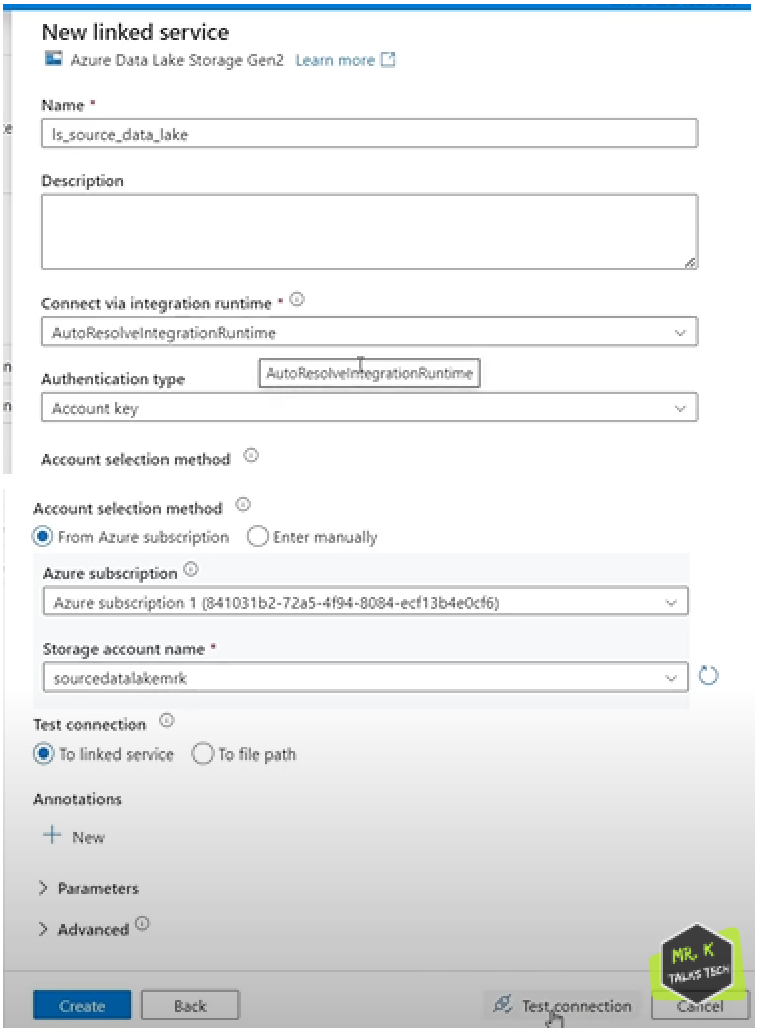
Showing 1 - 2 of 2 items

Name	Type
destination_data_lake	Azure Data Lake Storage Gen
source_datalake	Azure Data Lake Storage Gen

New linked service

data la

here we choose ADL service..next



New linked service

Azure Data Lake Storage Gen2 Learn more

Name *

ls_source_data_lake

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Authentication type

Account key

Account selection method

Account selection method

● From Azure subscription ○ Enter manually

Azure subscription

Azure subscription 1 (841031b2-72a5-4f94-8084-ecf13b4e0cf6)

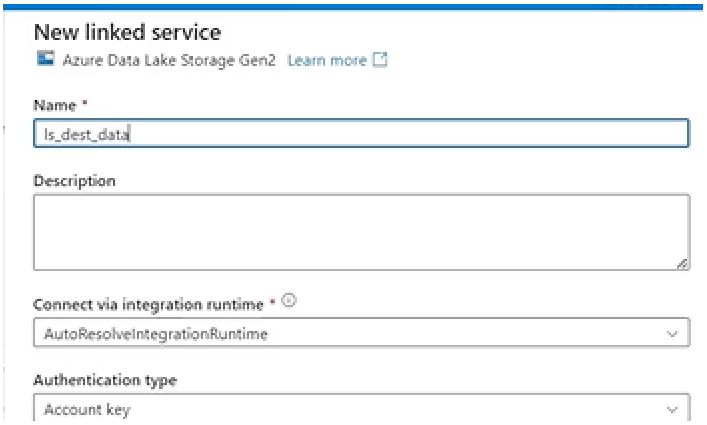
Storage account name *

sourcedatalakemrk

Test connection

● To linked service ○ To file path

9. Now test the connection bw ADL and ADF using test connection..then next click on create
10. Next we create another linkedservice for dest ADL as well



New linked service

Azure Data Lake Storage Gen2 Learn more

Name *

ls_dest_data

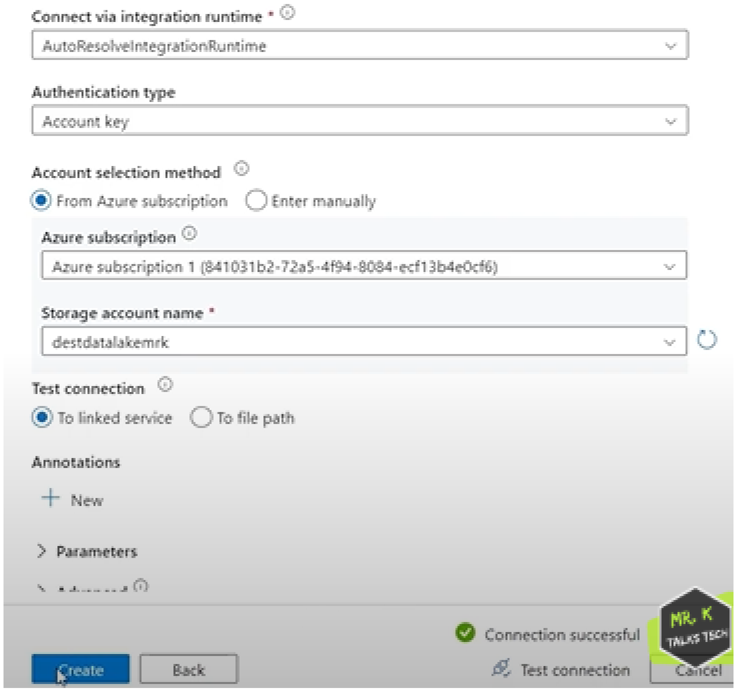
Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Authentication type

Account key



Connect via integration runtime *
AutoResolveIntegrationRuntime

Authentication type
Account key

Account selection method
● From Azure subscription ○ Enter manually

Azure subscription
Azure subscription 1 (841031b2-72a5-4f94-8084-ecf13b4e0cf6)

Storage account name *
destdatalakemrk

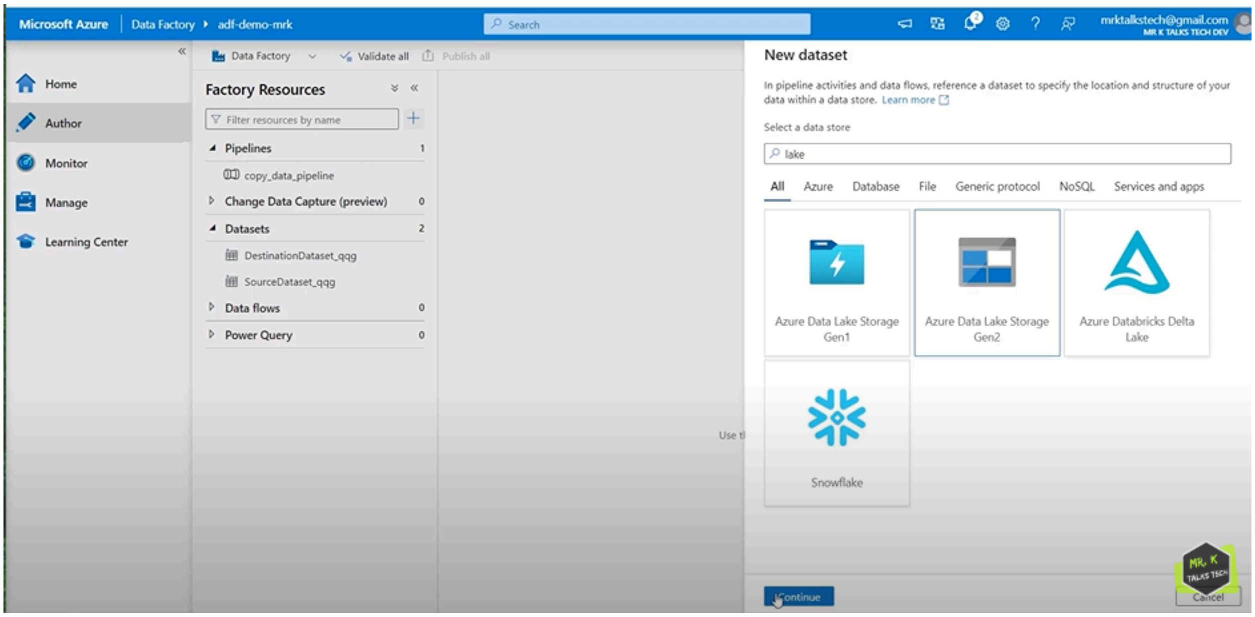
Test connection
● To linked service ○ To file path

Annotations

Connection successful

our connection is successful for ADF and dest ADL

11. Next we need to create the datasets



Pipelines 1
copy_data_pipeline
Change Data Capture (preview) 0
Datasets 2
DestinationDataset_qqg
SourceDataset_qqg
Data flows 0
Power Query 0

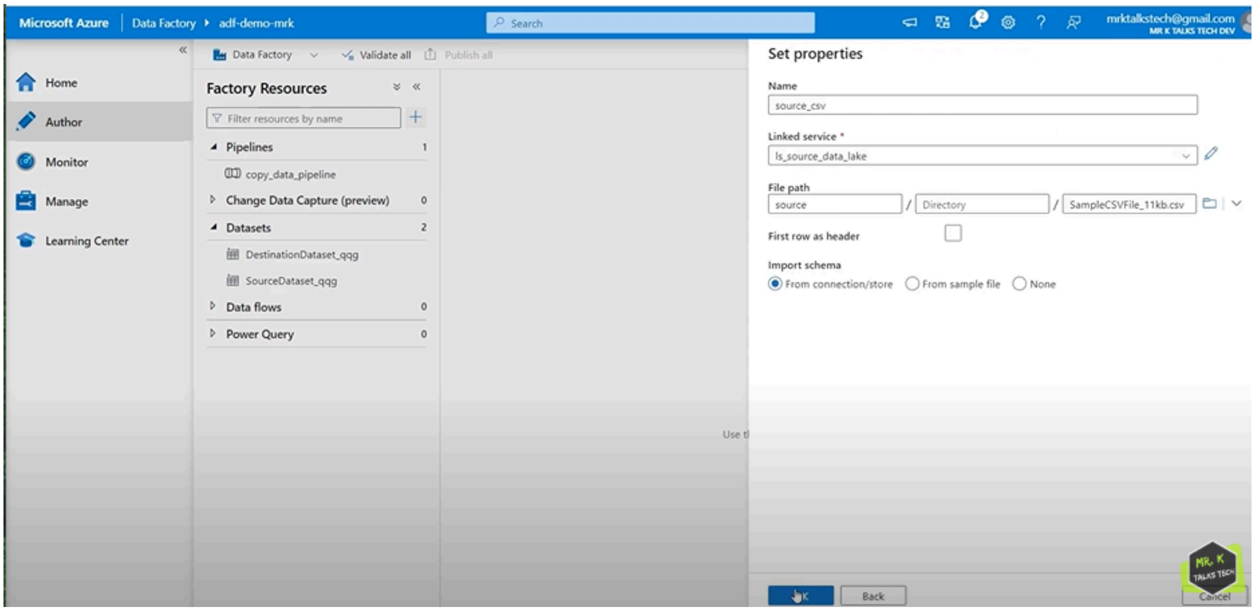
New dataset

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
lake

Here we'll choose ADL as our data source…and select the format

12. Next we set properties to our data set and select linked service and file path



Set properties

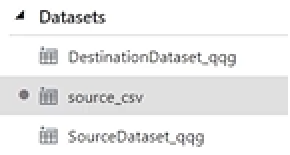
Name: source_csv

Linked service *: ls_source_data_lake

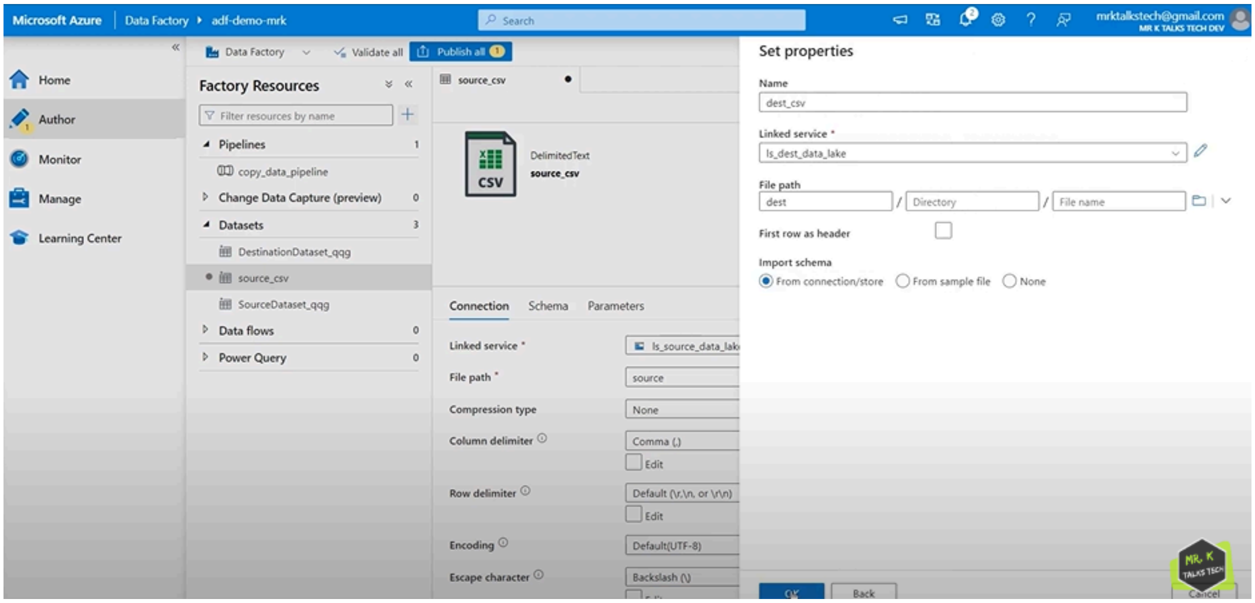
File path: source / Directory / SampleCSVFile_11kb.csv

First row as header □

Import schema: ● From connection/store ○ From sample file ○ None



Here we have created dataset for our source ...
Next we'll do the same for dest



Linked service *: ls_source_data_lak

File path *: source

Compression type: None

Column delimiter: Comma (,) □ Edit

Row delimiter: Default (\r,\n, or \r\n) □ Edit

Encoding: Default(UTF-8)

Escape character: Backslash (\)

Set properties

Name: dest_csv

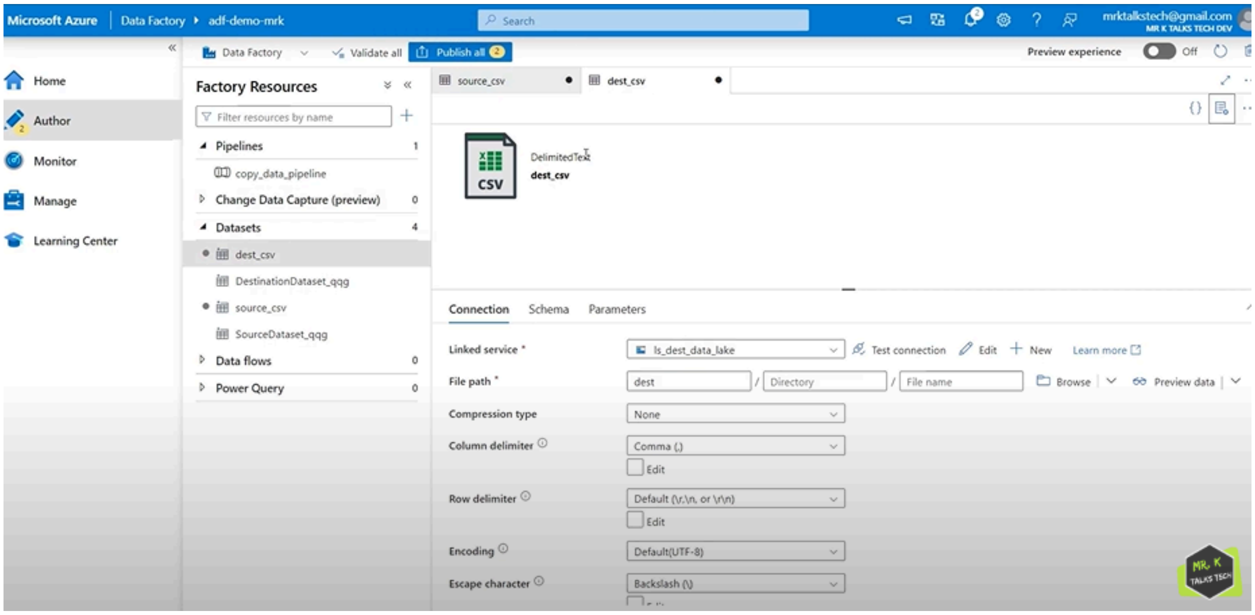
Linked service *: ls_dest_data_lake

File path: dest / Directory / File name

First row as header □

Import schema: ● From connection/store ○ From sample file ○ None

13. To save all our work ..click on publish all



Linked service * ls_dest_data_lake

File path * dest / Directory / File name

Compression type None

Column delimiter Comma (,)

☐ Edit

Row delimiter Default (\r,\n, or \r\n)

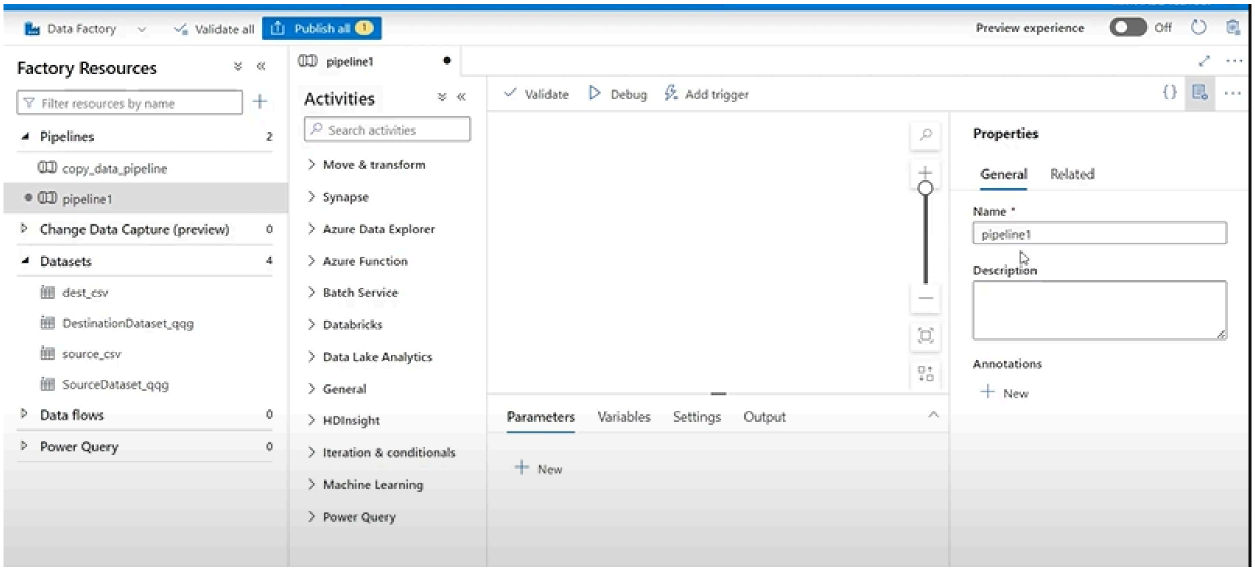
☐ Edit

Encoding Default(UTF-8)

Escape character Backslash (\)

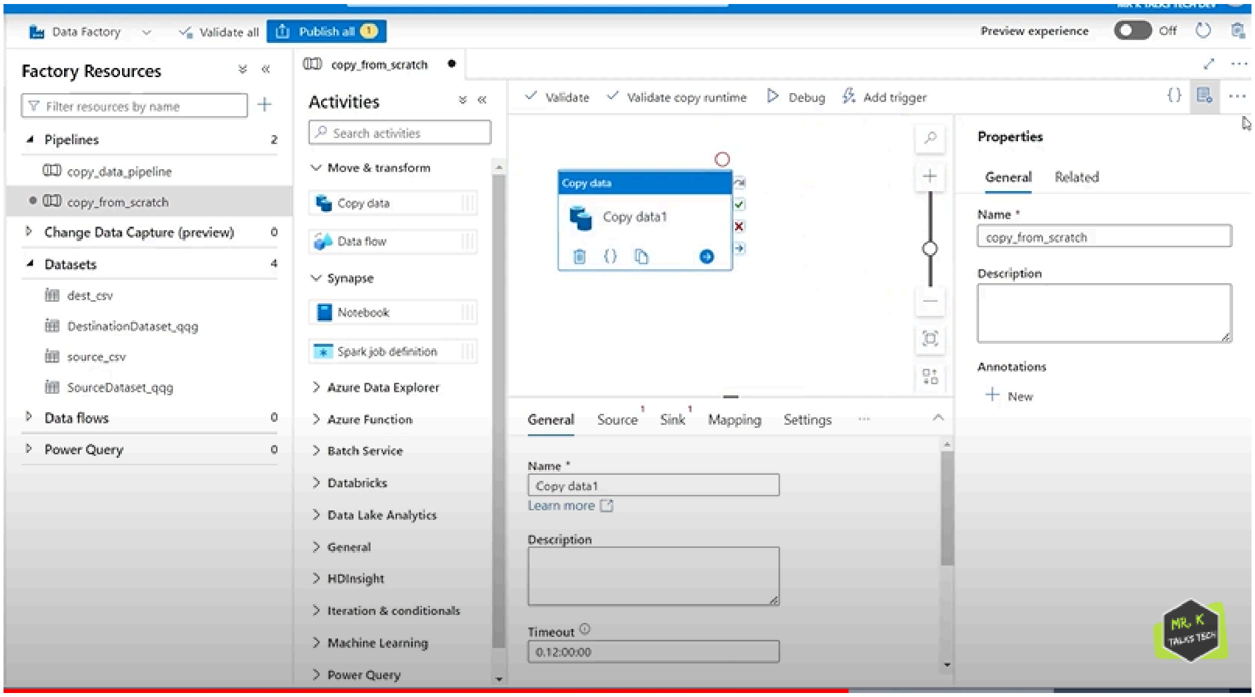
14. Now we have everything to create our pipeline

15. NExt to create a pipeline..go to author tab and click create pipeline



Name * pipeline1

16. Now we'll drag the copy activity in our pipeline



Name * copy_from_scratch

Description

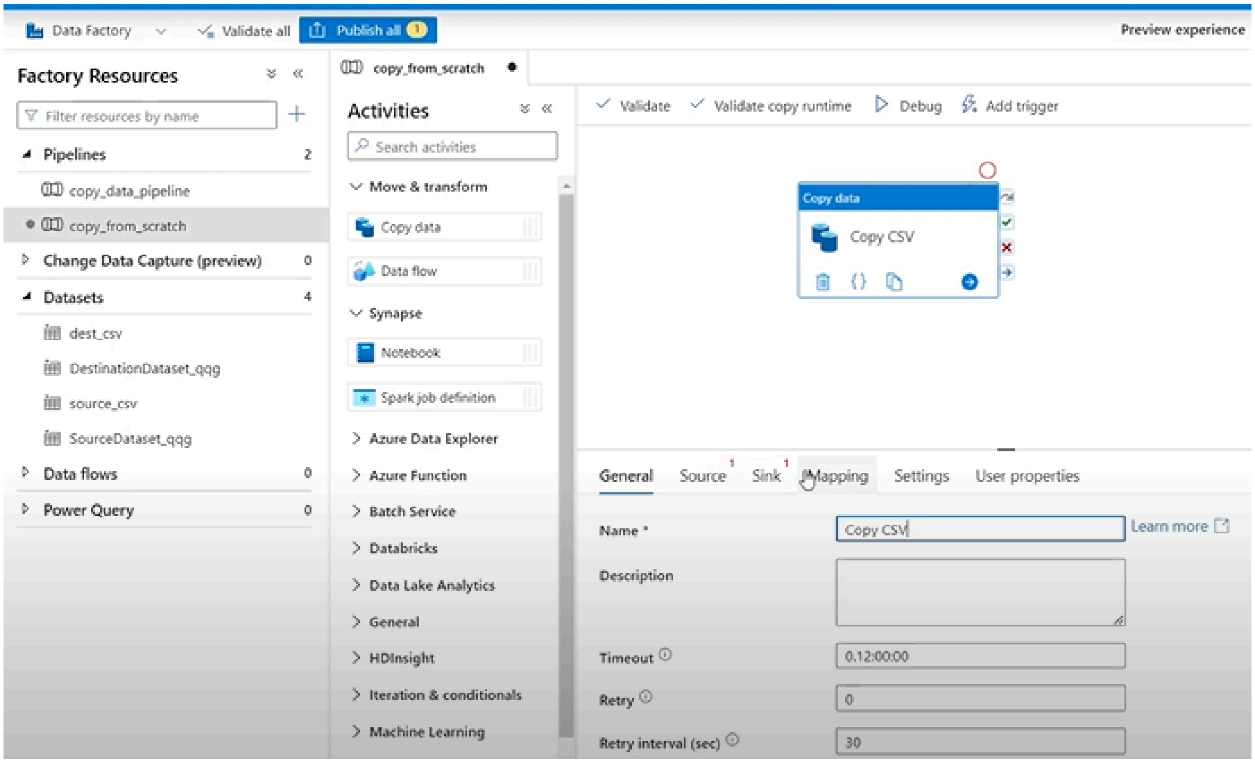
Name * Copy data1

Learn more

Description

Timeout 0.12:00:00

17. Next here we have to config the source,sink



Name * Copy CSV

Learn more

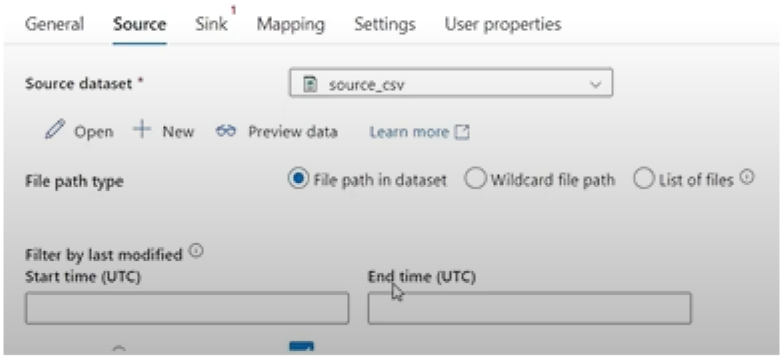
Description

Timeout 0.12:00:00

Retry 0

Retry interval (sec) 30

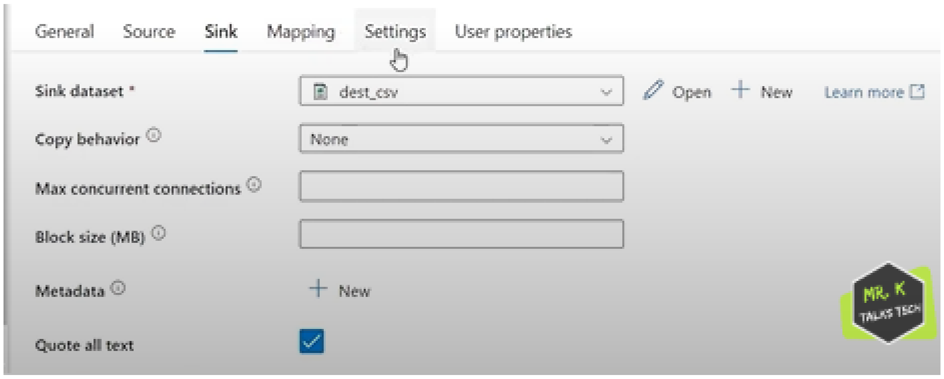
18. Click on source and select our source dataset



Source dataset * source_csv

File path type ● File path in dataset ○ Wildcard file path ○ List of files

Filter by last modified
Start time (UTC)
End time (UTC)



Sink dataset * dest_csv

Copy behavior None

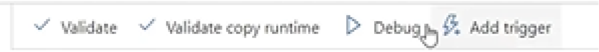
Max concurrent connections

Block size (MB)

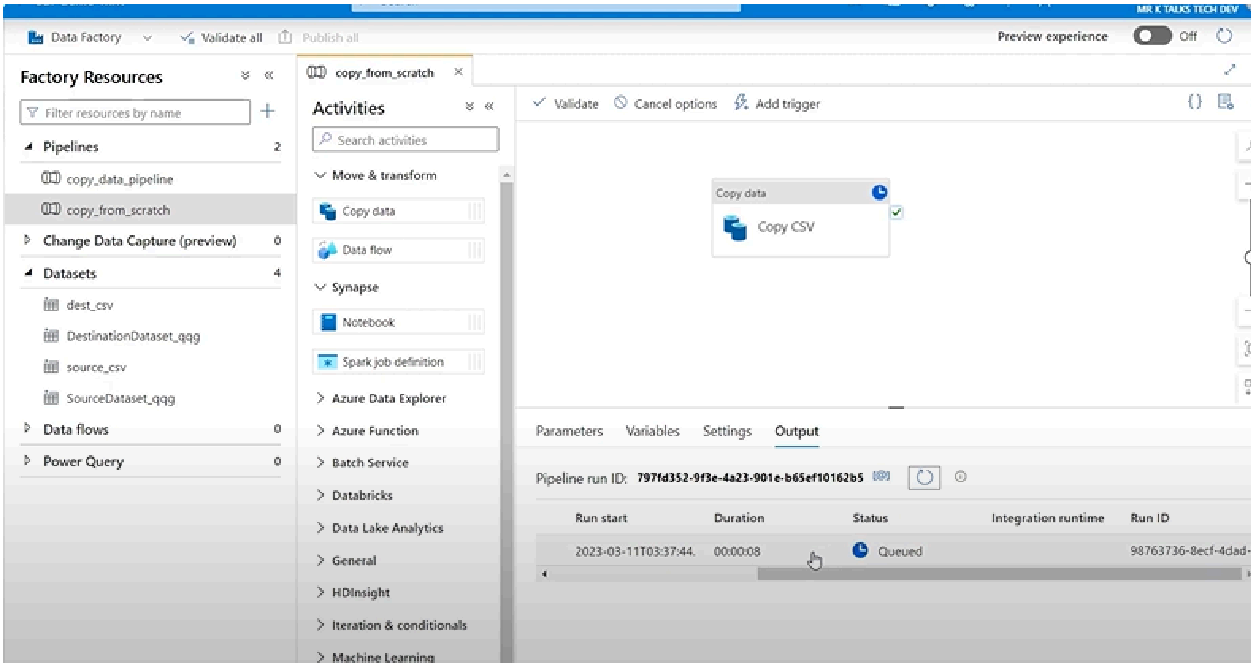
Metadata + New

Quote all text ☑

19. Next sink
20. Next click on publish to save our work
21. Now we’ll use DEBUG mode and run our pipeline



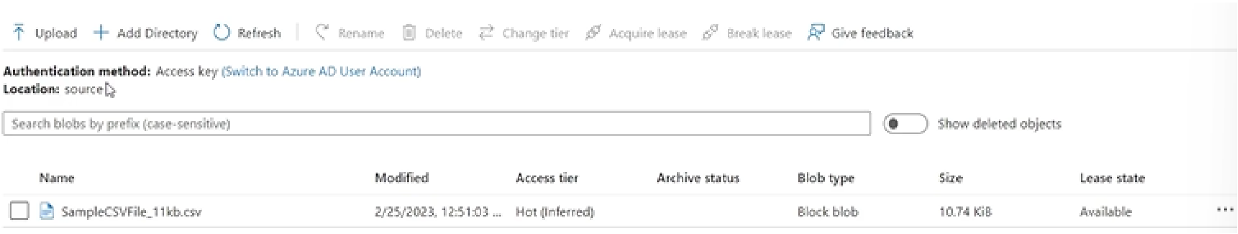
22. Our pipeline is currently running



Pipeline run ID: 797fd352-9f3e-4a23-901e-b65ef10162b5

Run start	Duration	Status	Integration runtime	Run ID
2023-03-11T03:37:44.	00:00:08	Queued		98763736-8ecf-4dad-

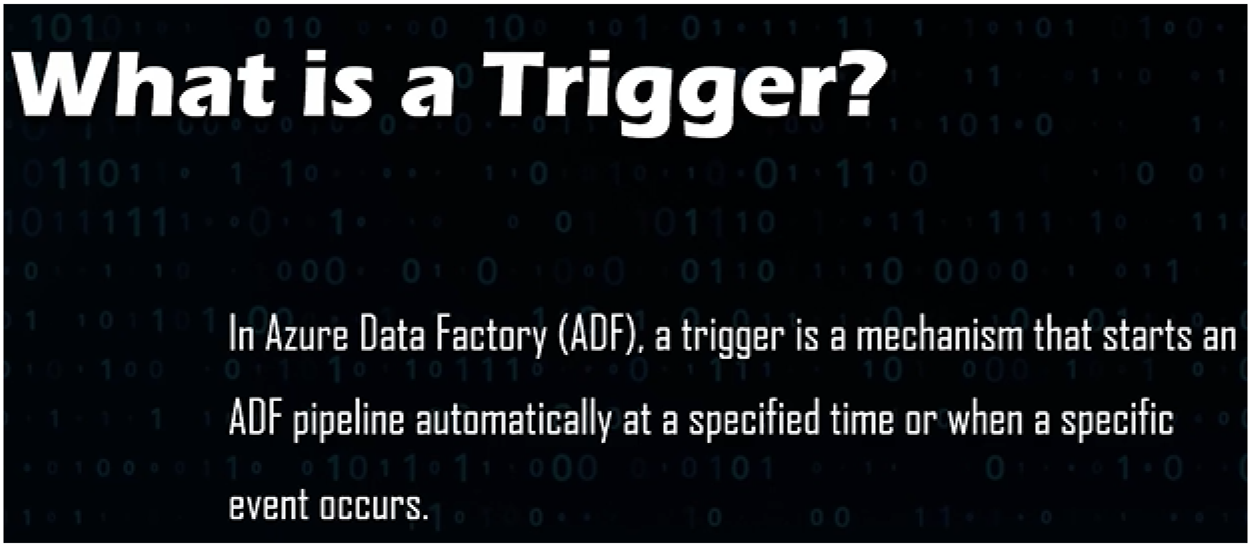
23. Now we can see our files in the dest Data lake



Authentication method: Access key (Switch to Azure AD User Account)
Location: source

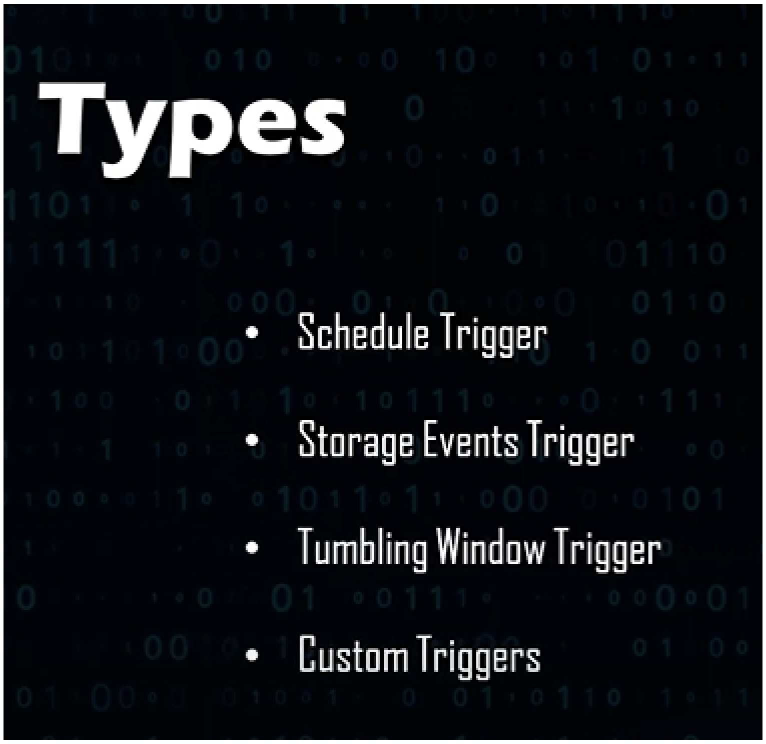
	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
□	SampleCSVFile_11kb.csv	2/25/2023, 12:51:03 ...	Hot (Inferred)		Block blob	10.74 KiB	Available

Triggers in DF and setting up Scheduled Trigger in ADF

1. 

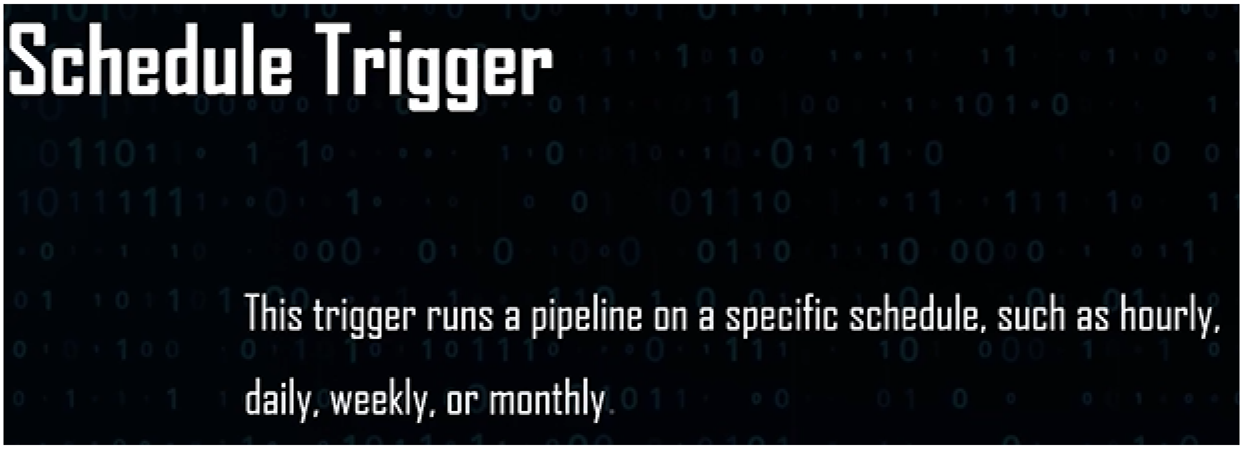
What is a Trigger?

In Azure Data Factory (ADF), a trigger is a mechanism that starts an ADF pipeline automatically at a specified time or when a specific event occurs.

2. Types 

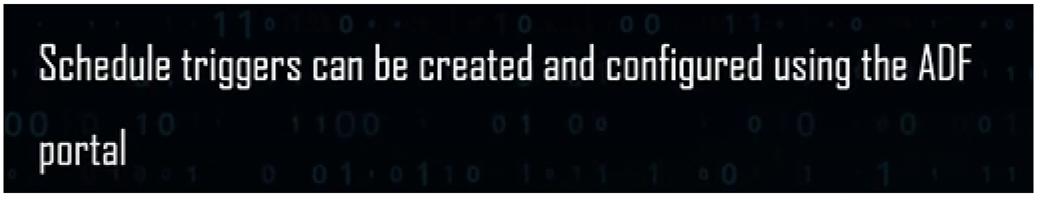
   ## Types

   - Schedule Trigger
   - Storage Events Trigger
   - Tumbling Window Trigger
   - Custom Triggers

3. Here we'll learn the Schedule trigger
4. 

   ## Schedule Trigger

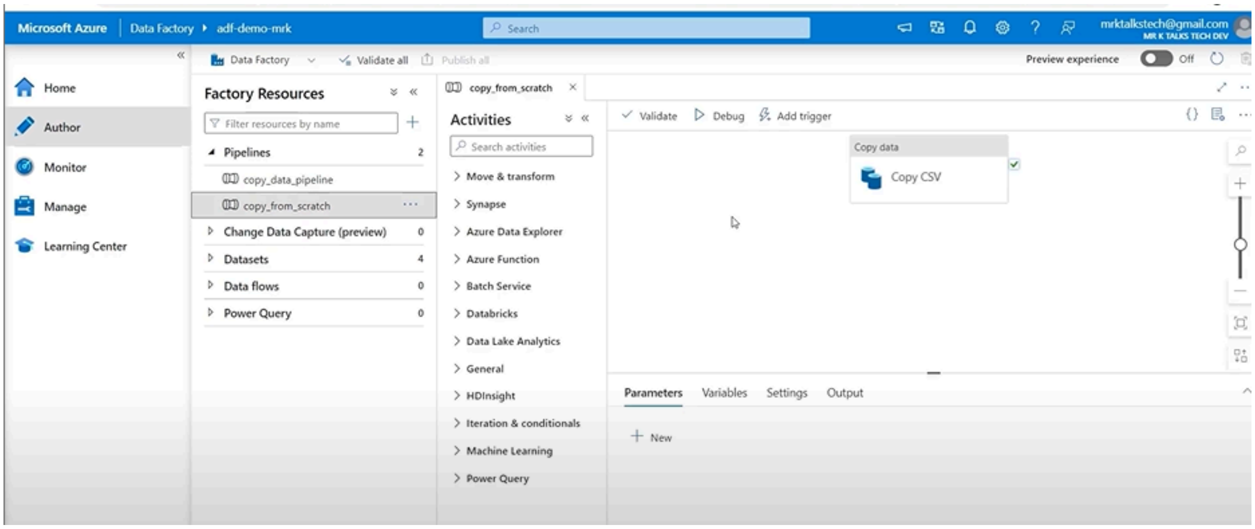
   This trigger runs a pipeline on a specific schedule, such as hourly, daily, weekly, or monthly.

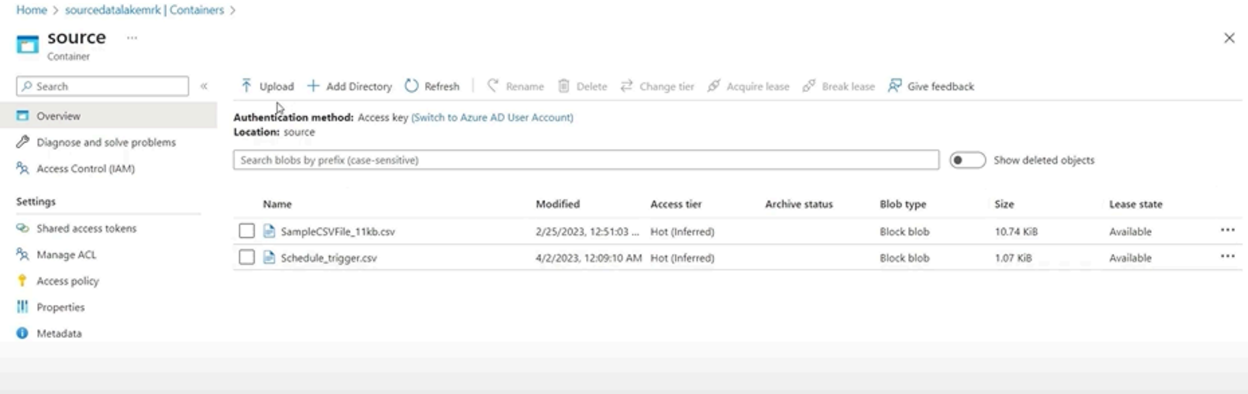
   Schedule triggers can be created and configured using the ADF portal

5. Lets see how we can schedule a trigger

6. Previously we have created a pipeline which copies the data from source DL to Dest DL



7. Now we'll add a file to our source DL and test our pipeline

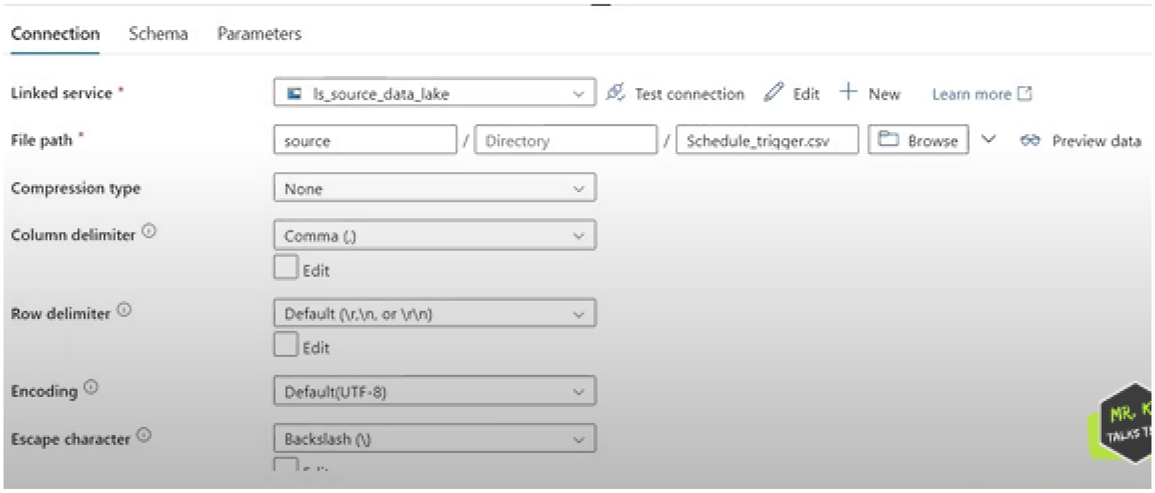


Authentication method: Access key (Switch to Azure AD User Account)
Location: source

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
SampleCSVFile_11kb.csv	2/25/2023, 12:51:03 ...	Hot (Inferred)		Block blob	10.74 KiB	Available
Schedule_trigger.csv	4/2/2023, 12:09:10 AM	Hot (Inferred)		Block blob	1.07 KiB	Available

here we have uploaded our file

8. And in our pipeline..we'll change the source path ..to the new file in the ADL



Linked service * ls_source_data_lake

File path * source / Directory / Schedule_trigger.csv

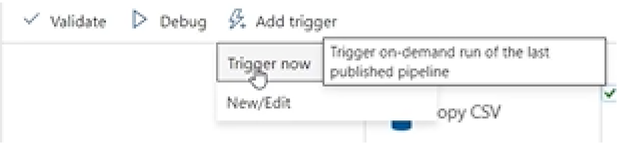
Compression type None

Column delimiter Comma (,)
□ Edit

Row delimiter Default (\r,\n, or \r\n)
□ Edit

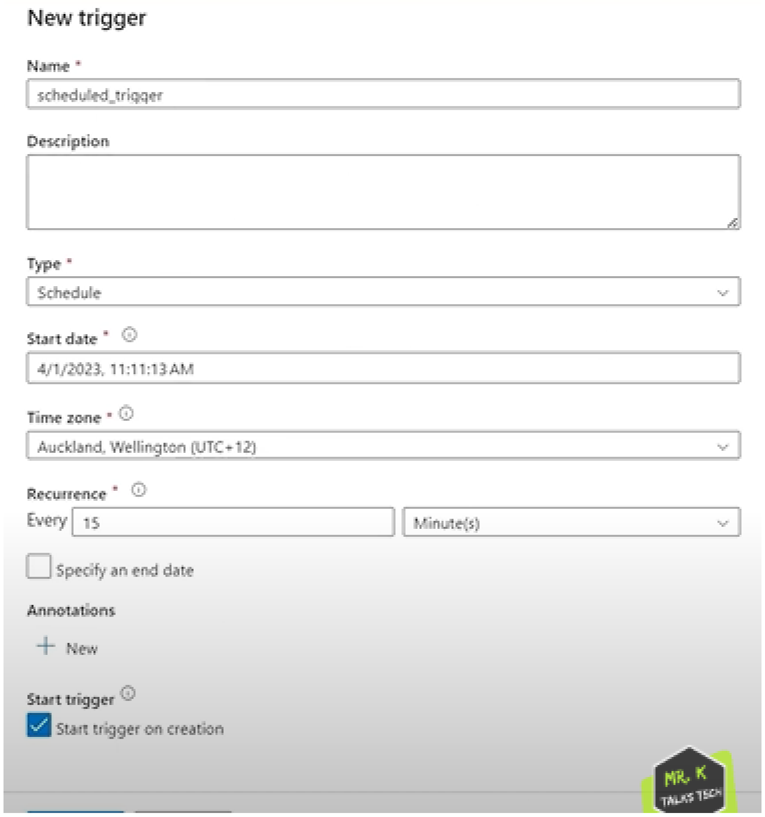
Encoding Default(UTF-8)

Escape character Backslash (\)



Trigger on-demand run of the last published pipeline

9. Here Trigger now is same as debug
10. Click on new/edit to add triggers..click bew trigger
11. Next we give properties for our trigger



New trigger

Name *
scheduled_trigger

Description

Type *
Schedule

Start date *
4/1/2023, 11:11:13 AM

Time zone *
Auckland, Wellington (UTC+12)

Recurrence *
Every 15 Minute(s)

☐ Specify an end date

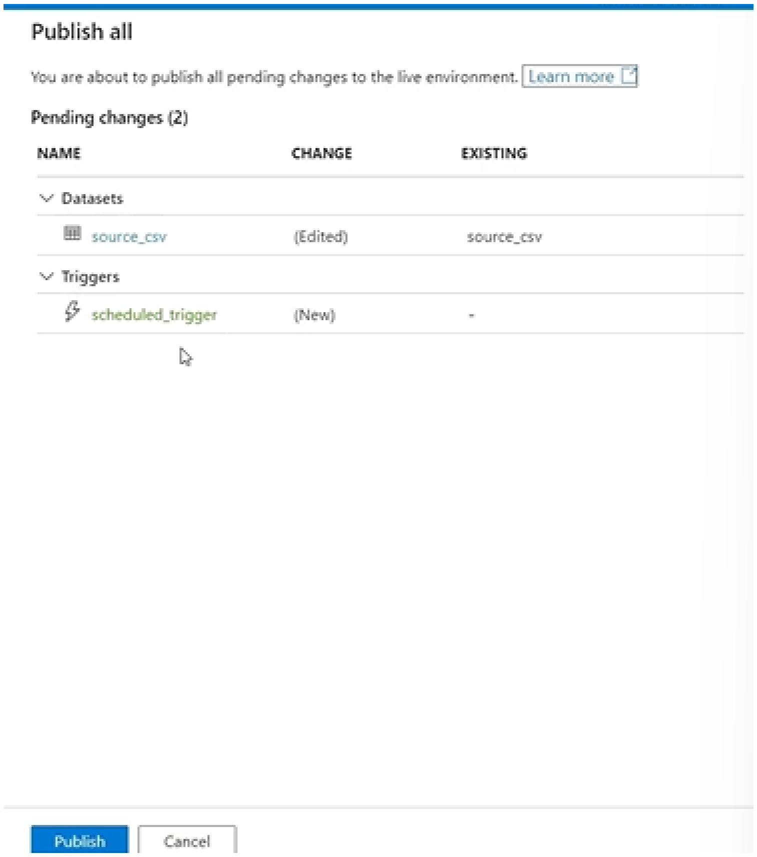
Annotations

+ New

Start trigger
☑ Start trigger on creation

we can also specify an end date to stop our trigger

12. Create the trigger and publish the changes



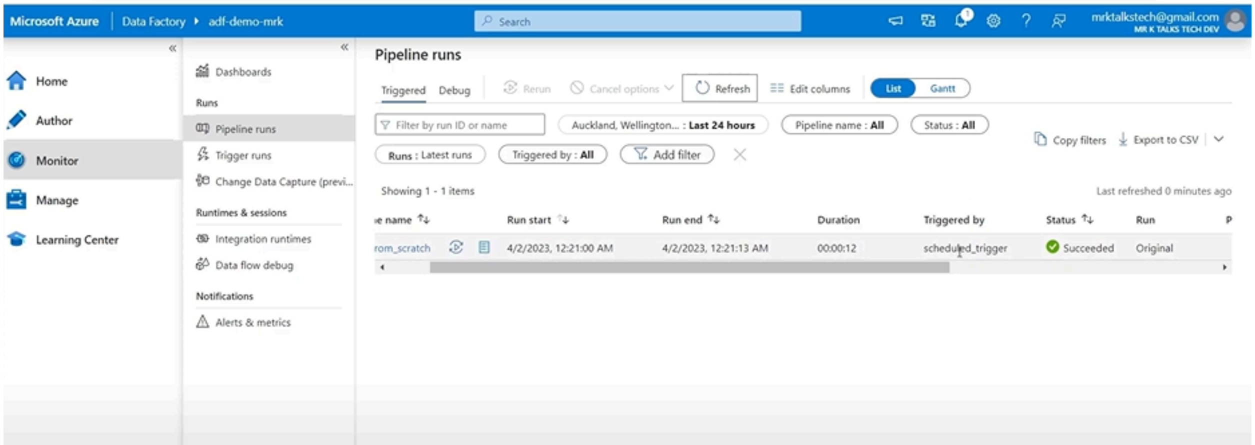
Publish all

You are about to publish all pending changes to the live environment. Learn more

Pending changes (2)

NAME	CHANGE	EXISTING
Datasets		
source_csv	(Edited)	source_csv
Triggers		
scheduled_trigger	(New)	-

13. Here we can our pipeline has succeeded



Pipeline runs

Auckland, Wellington... : Last 24 hours

Pipeline name : All

Status : All

Runs : Latest runs

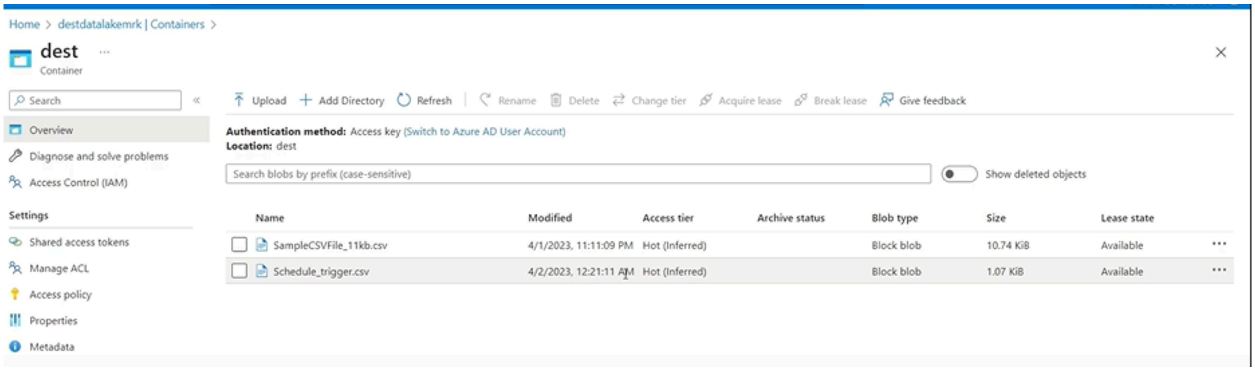
Triggered by : All

Showing 1 - 1 items

Last refreshed 0 minutes ago

ie name	Run start	Run end	Duration	Triggered by	Status	Run	P
rom_scratch	4/2/2023, 12:21:00 AM	4/2/2023, 12:21:13 AM	00:00:12	scheduled_trigger	Succeeded	Original	

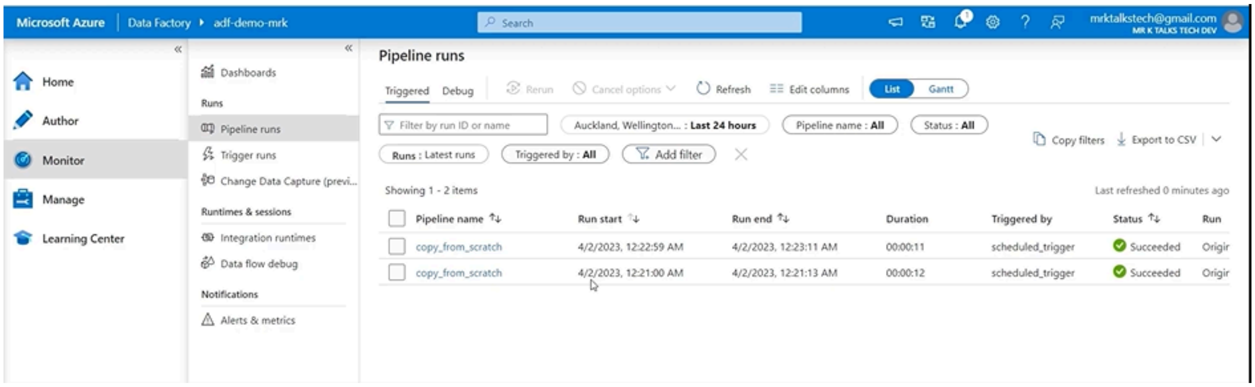
14. Here we can see our pipeline ran successfully and copies the file



Authentication method: Access key (Switch to Azure AD User Account)
Location: dest

	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
□	SampleCSVFile_11kb.csv	4/1/2023, 11:11:09 PM	Hot (Inferred)		Block blob	10.74 KiB	Available	···
□	Schedule_trigger.csv	4/2/2023, 12:21:11 AM	Hot (Inferred)		Block blob	1.07 KiB	Available	···

15. This pipeline will be run every two minutes



Auckland, Wellington... : **Last 24 hours** | Pipeline name : **All** | Status : **All** | **Runs** : Latest runs | Triggered by : **All**

Showing 1 - 2 items

Last refreshed 0 minutes ago

□	Pipeline name	Run start	Run end	Duration	Triggered by	Status	Run
□	copy_from_scratch	4/2/2023, 12:22:59 AM	4/2/2023, 12:23:11 AM	00:00:11	scheduled_trigger	✓ Succeeded	Origir
□	copy_from_scratch	4/2/2023, 12:21:00 AM	4/2/2023, 12:21:13 AM	00:00:12	scheduled_trigger	✓ Succeeded	Origir