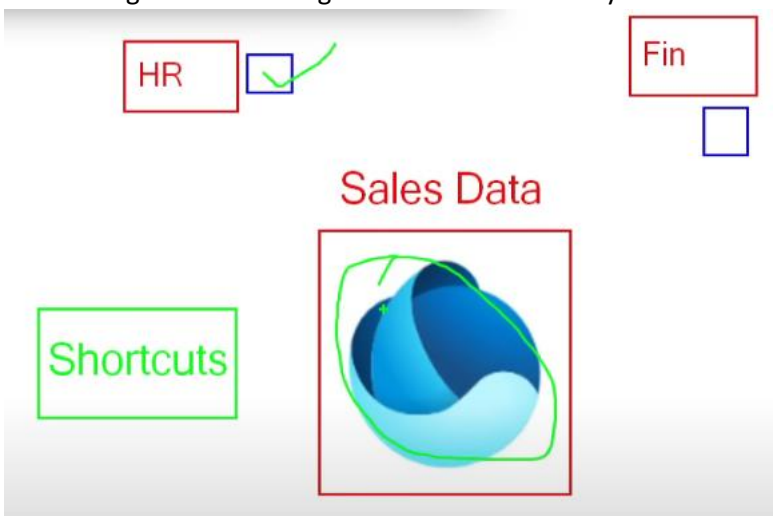
FABRIC DATA FACTORY

MF is End to End Unified Data Solution (from data ingestion to deployment)

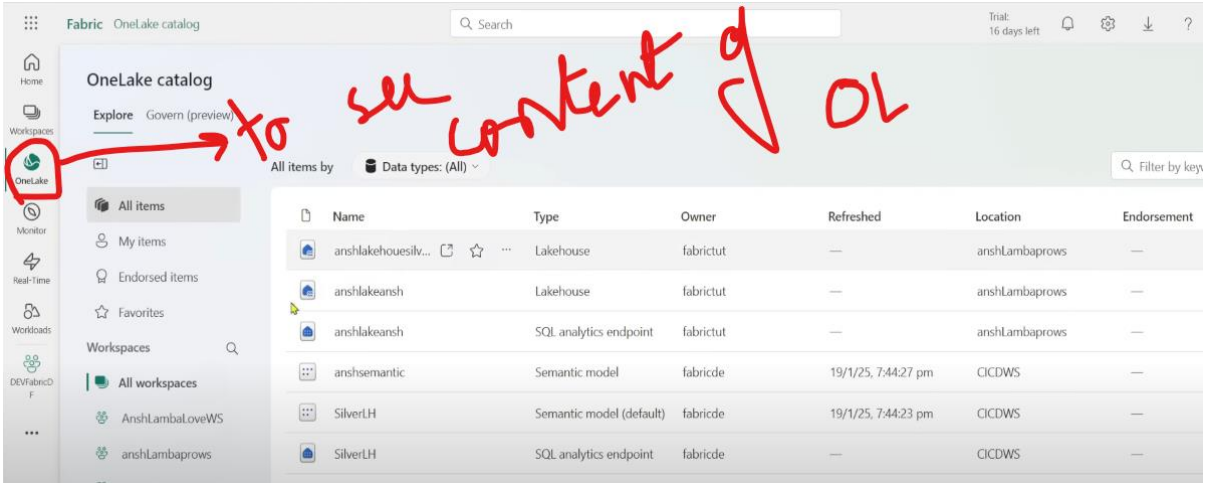
Best part of MS is that all the services are connected/integrated(we don't have to worry about connections)

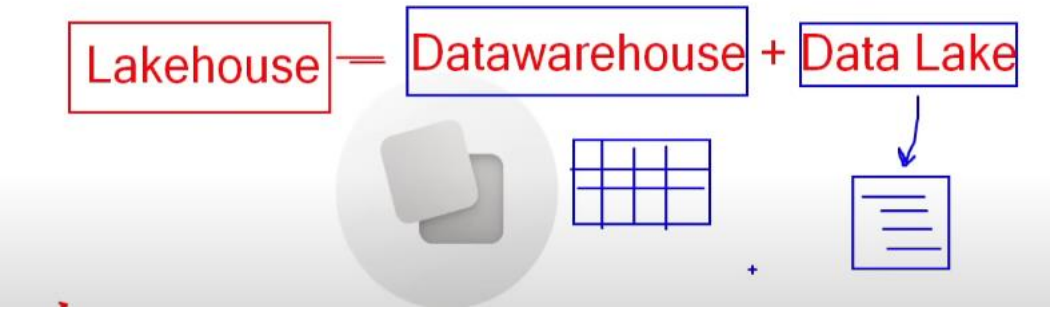
Fabric Data Factory → ETL/ELT solution of Fabric, helps in Data Movement, Transformation, Orchestration and Integration

OneLake is cheaper than ADLS Gen2 since we have only one copy of data, no need to manage networking between storage accounts since its only one lake.

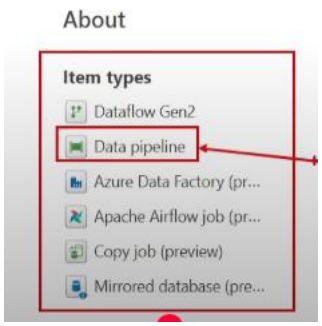


Let's say we have Sales data, now we have 2 workspaces HR and Finance, how the Sales data made available to both the workspaces? OL bts creates Shortcuts



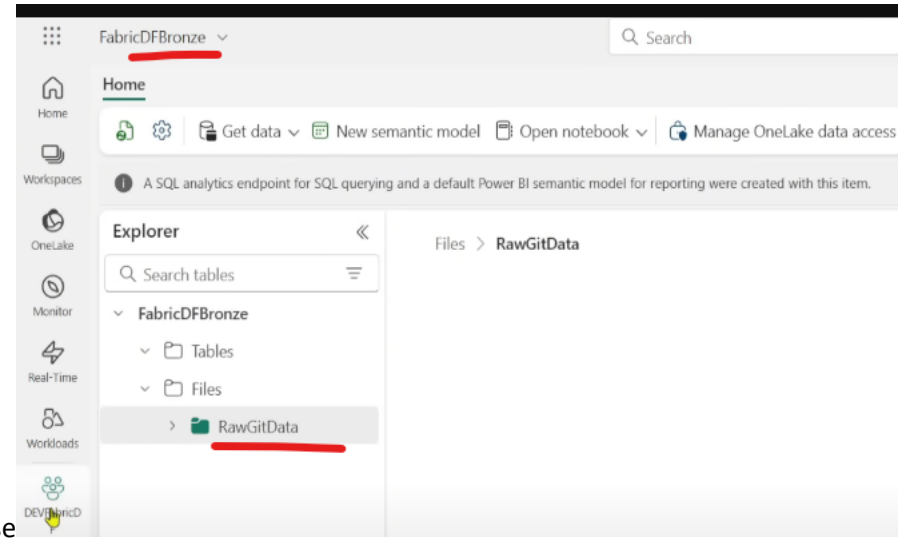


Simply put files in Data Lake, DWH is a logical DWH over here, it will create a metadata layer on top of files and it will behave like a table



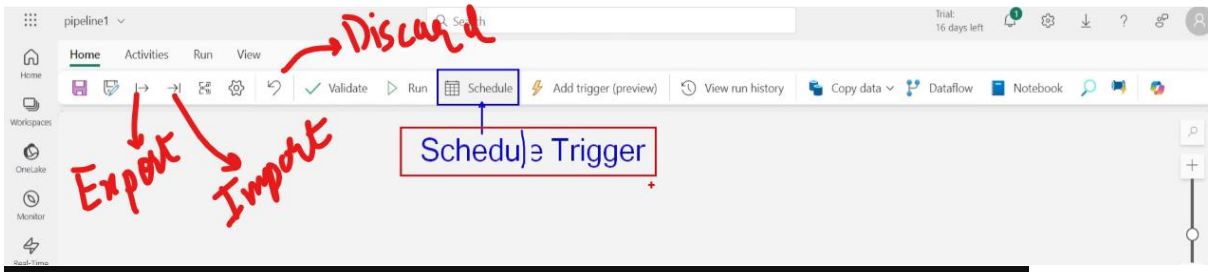
Go to app.fabric.microsoft.com

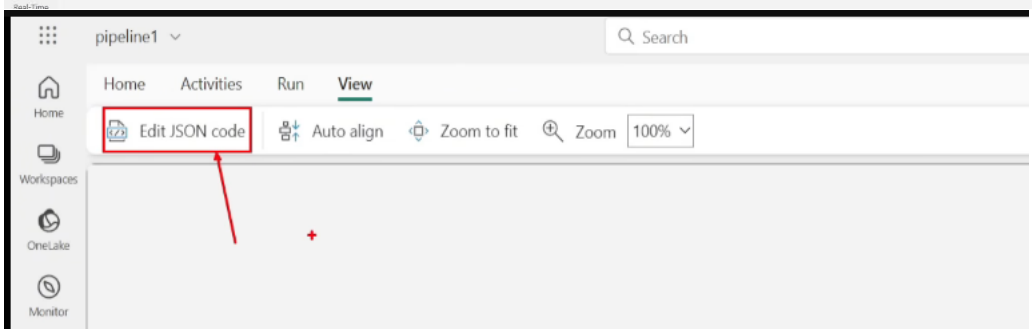
Create a Workspace



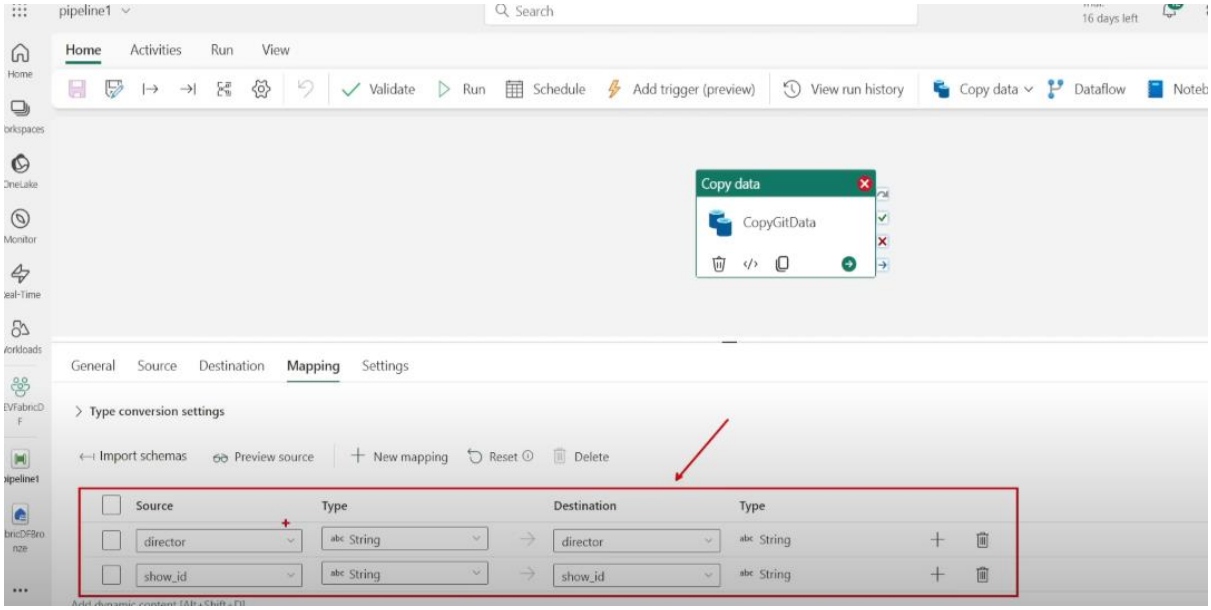
Create a Lakehouse

Create a Data Pipeline

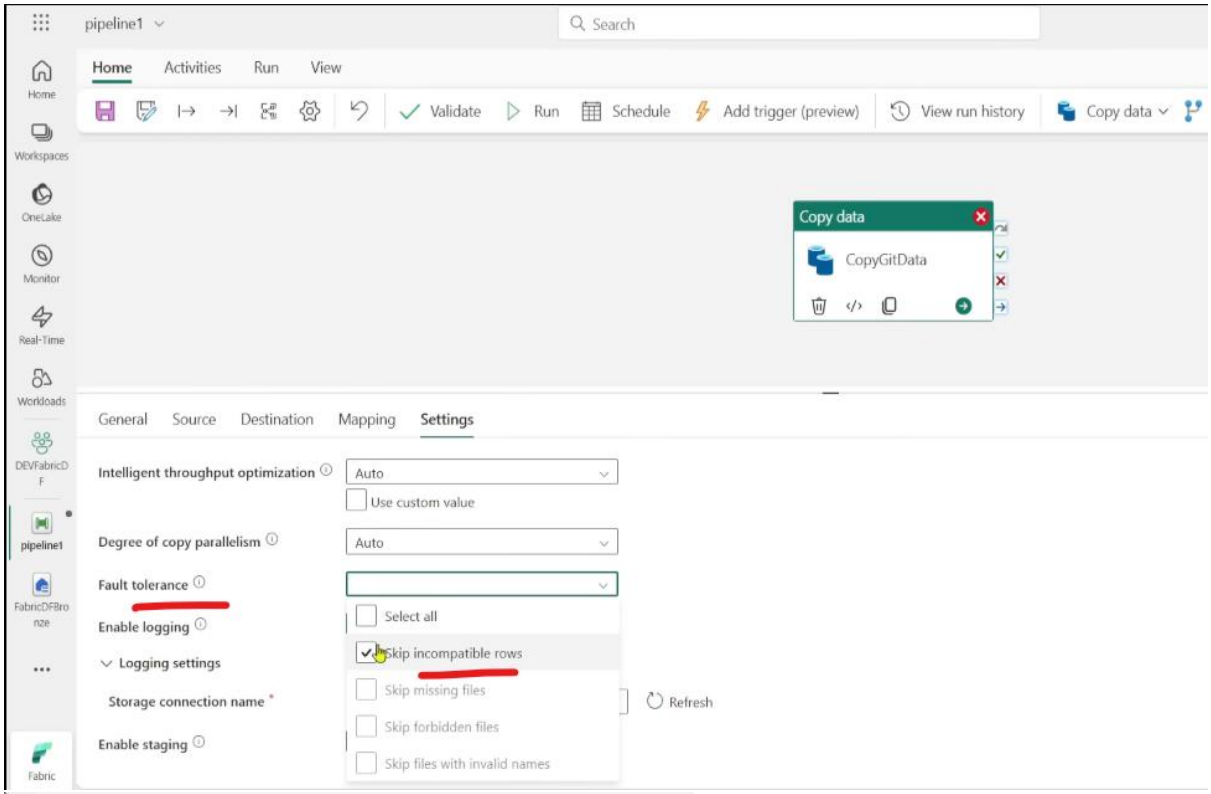


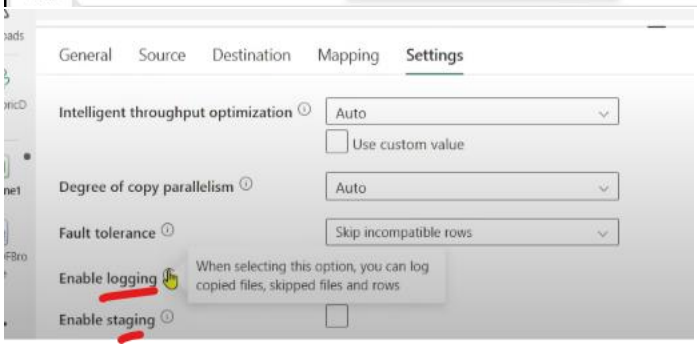


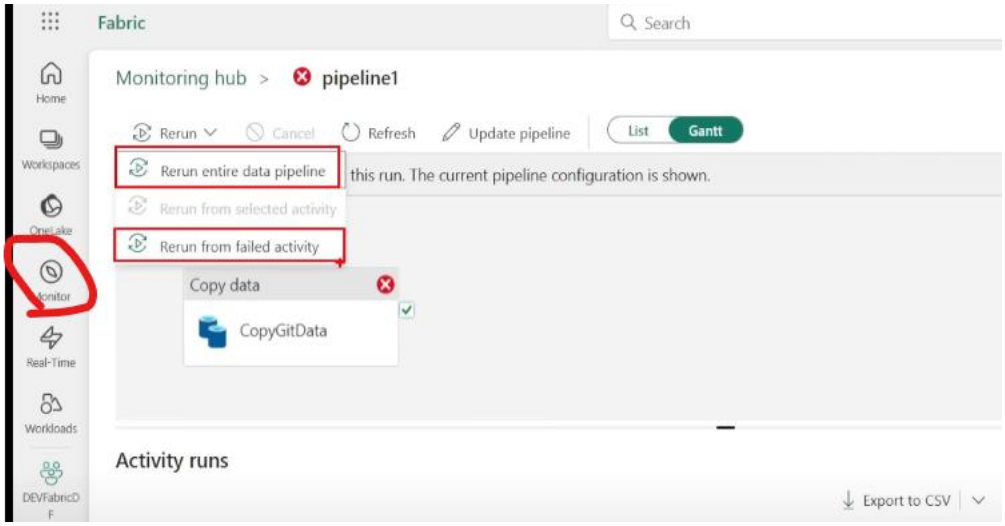
Get data from GitHub and store in LH



Mapping is required for Schema Enforcement

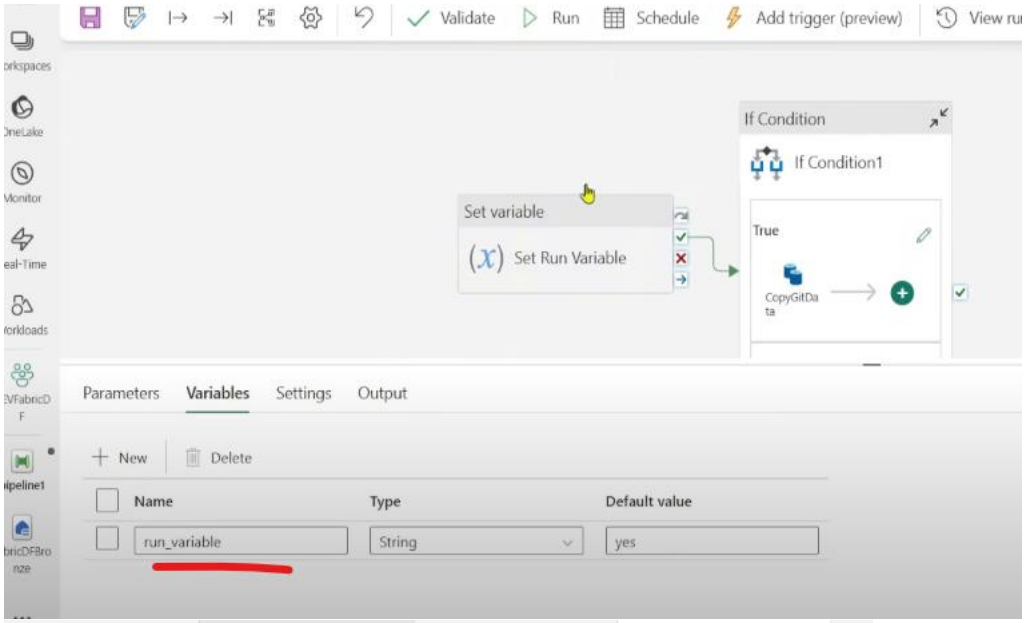


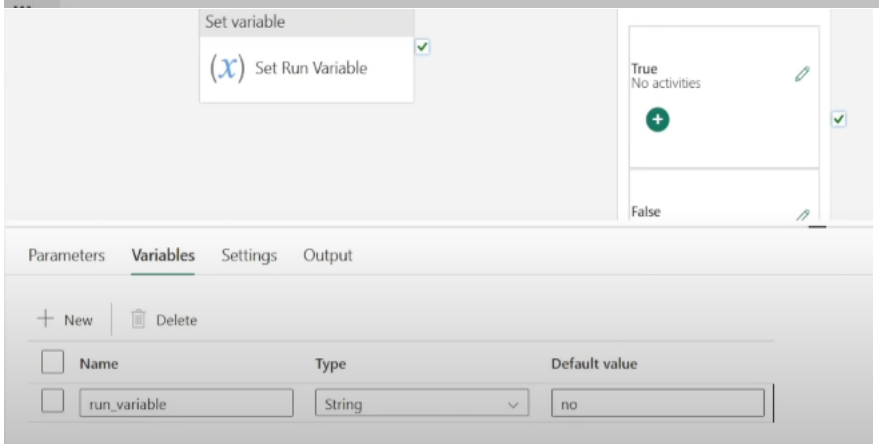


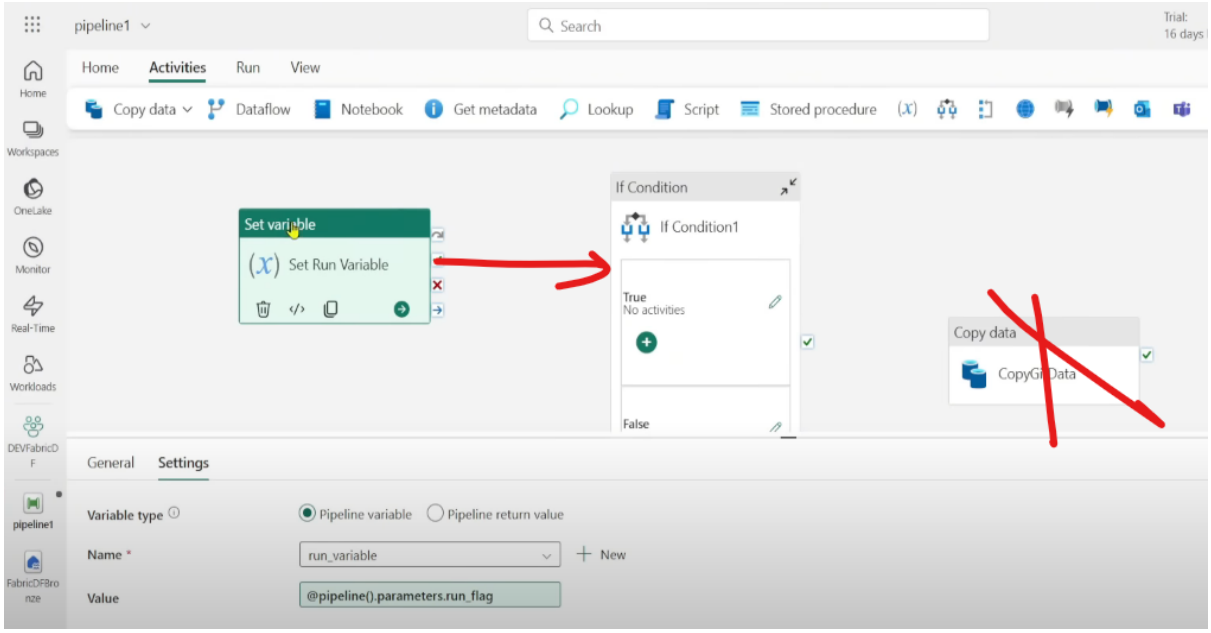


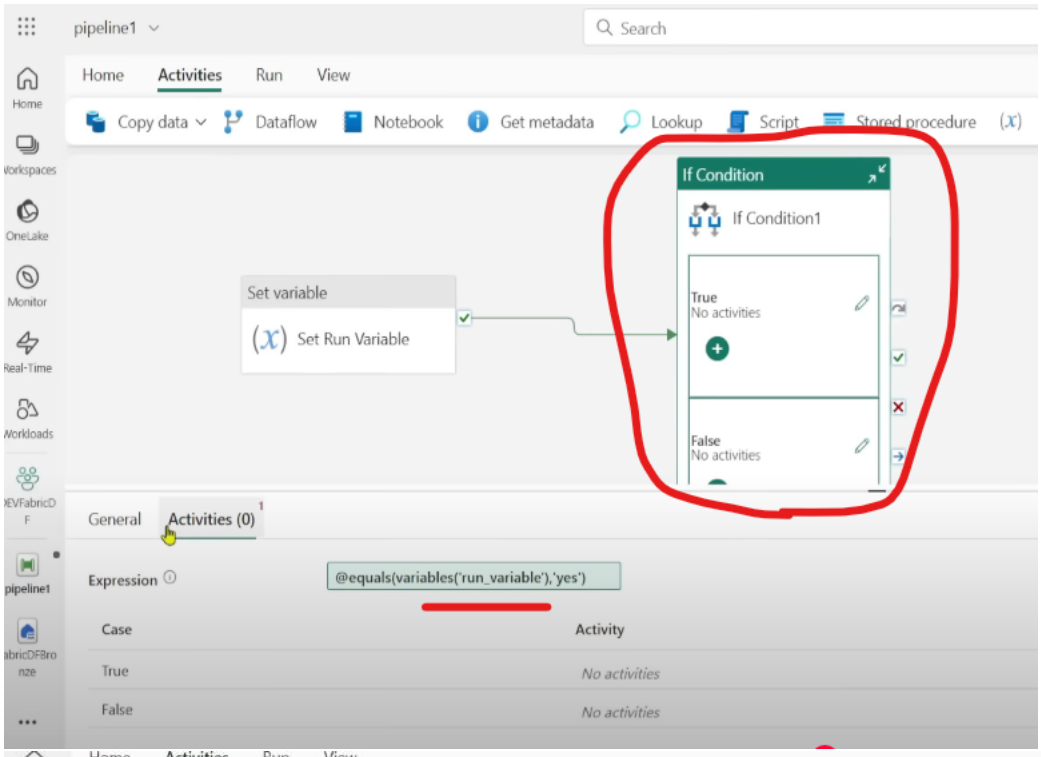
Problem statement1: Copy data from GitHub only if the pipeline parameter is yes

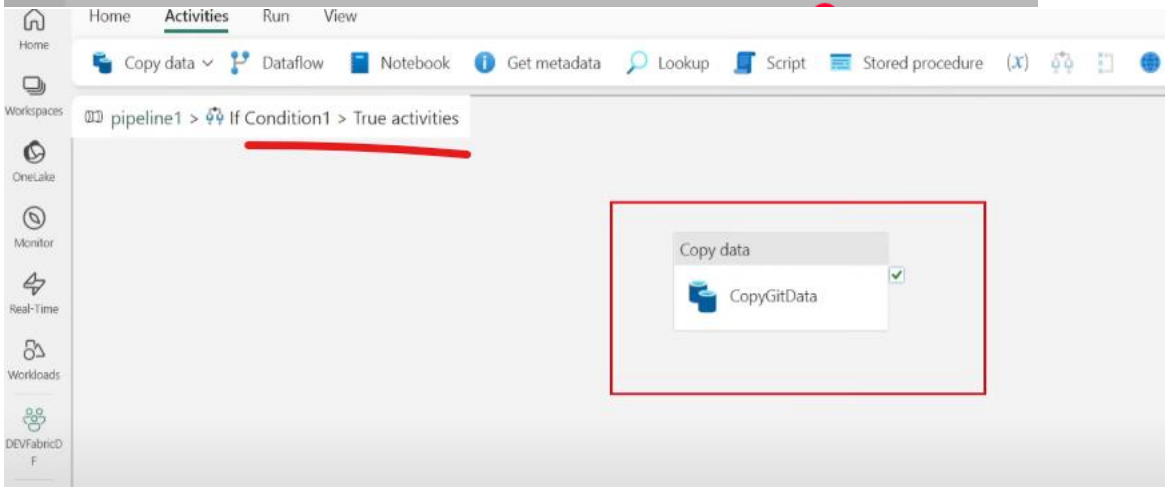
Create a pipeline parameter & a Variable

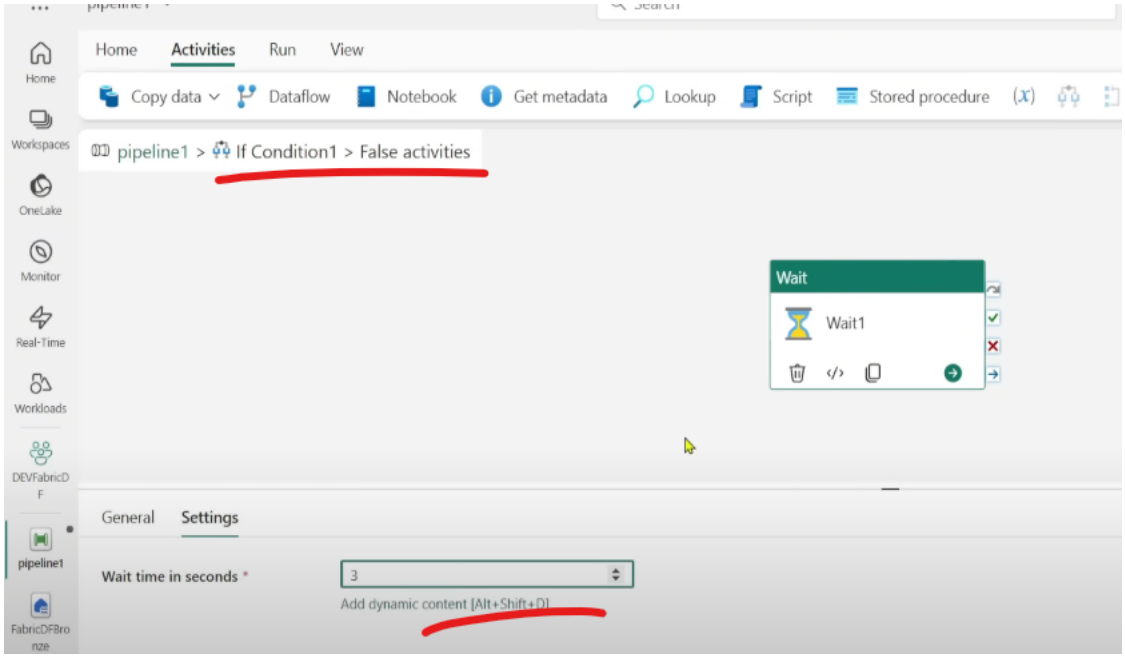





pipeline1
Search
Trial: 16 days
Home
Activities
Run
View
Copy data
Dataflow
Notebook
Get metadata
Lookup
Script
Stored procedure
Set variable
Set Run Variable
If Condition
If Condition1
True
No activities
False
Copy data
CopyGitData
General
Settings
Variable type
Pipeline variable
Pipeline return value
Name *
run_variable
New
Value
@pipeline().parameters.run_flag


pipeline1
Search
Home
Activities
Run
View
Copy data
Dataflow
Notebook
Get metadata
Lookup
Script
Stored procedure
Set variable
Set Run Variable
If Condition
If Condition1
True
No activities
False
No activities
General
Activities (0)
Expression
@equals(variables('run_variable'),'yes')
Case
Activity
True
No activities
False
No activities

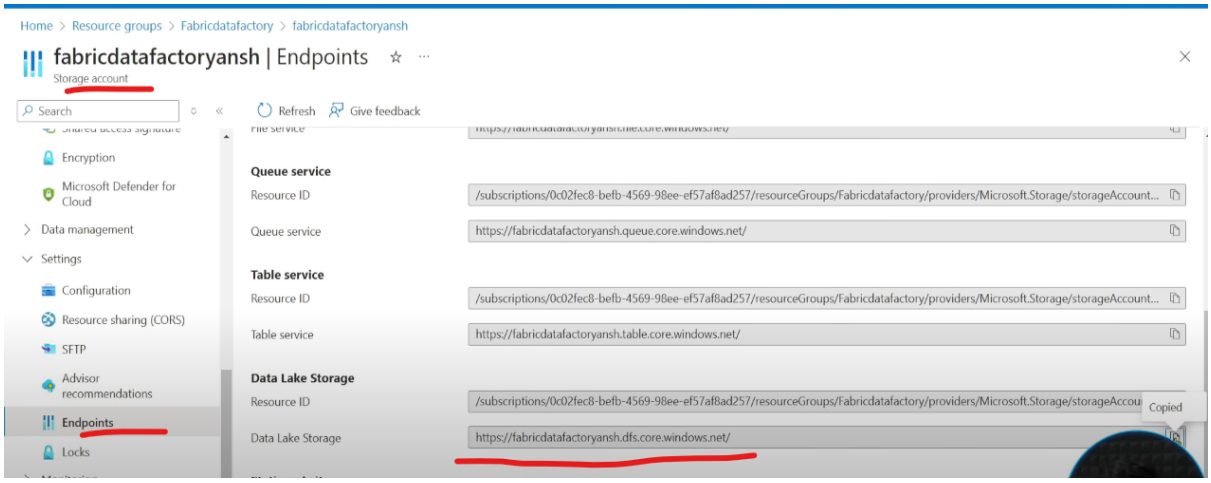

Home
Activities
Run
View
Copy data
Dataflow
Notebook
Get metadata
Lookup
Script
Stored procedure
pipeline1 > If Condition1 > True activities
Copy data
CopyGitData



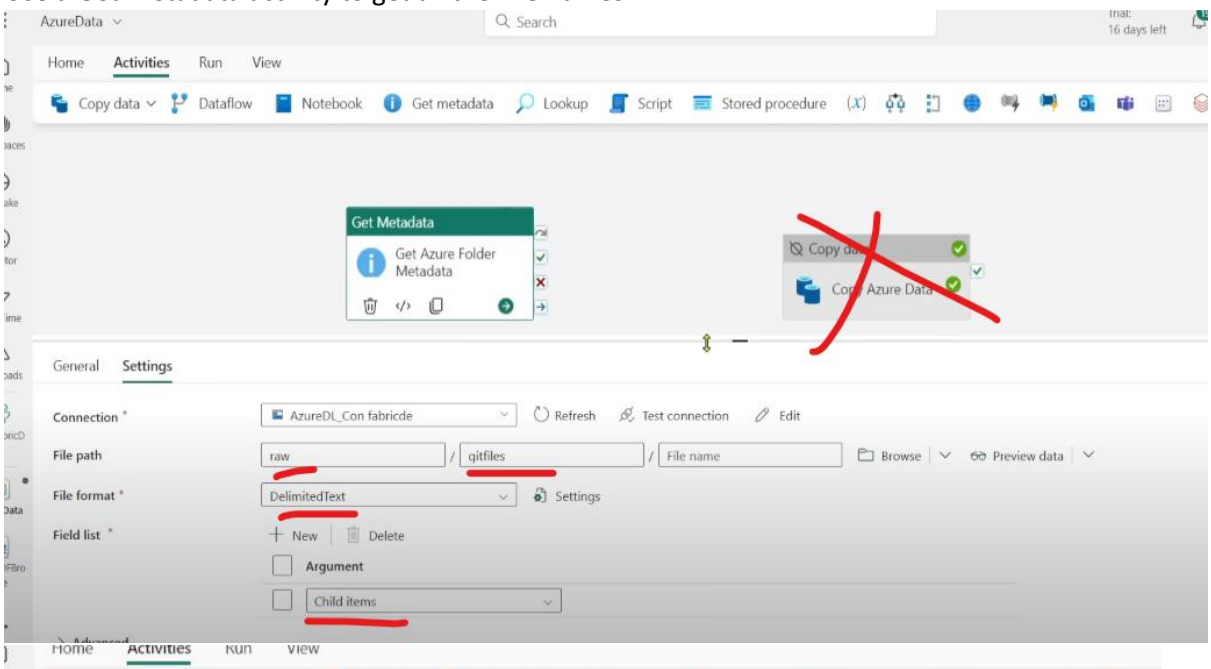
Note : We cant use For-Each inside If condition in Fabric (in ADF its supported)

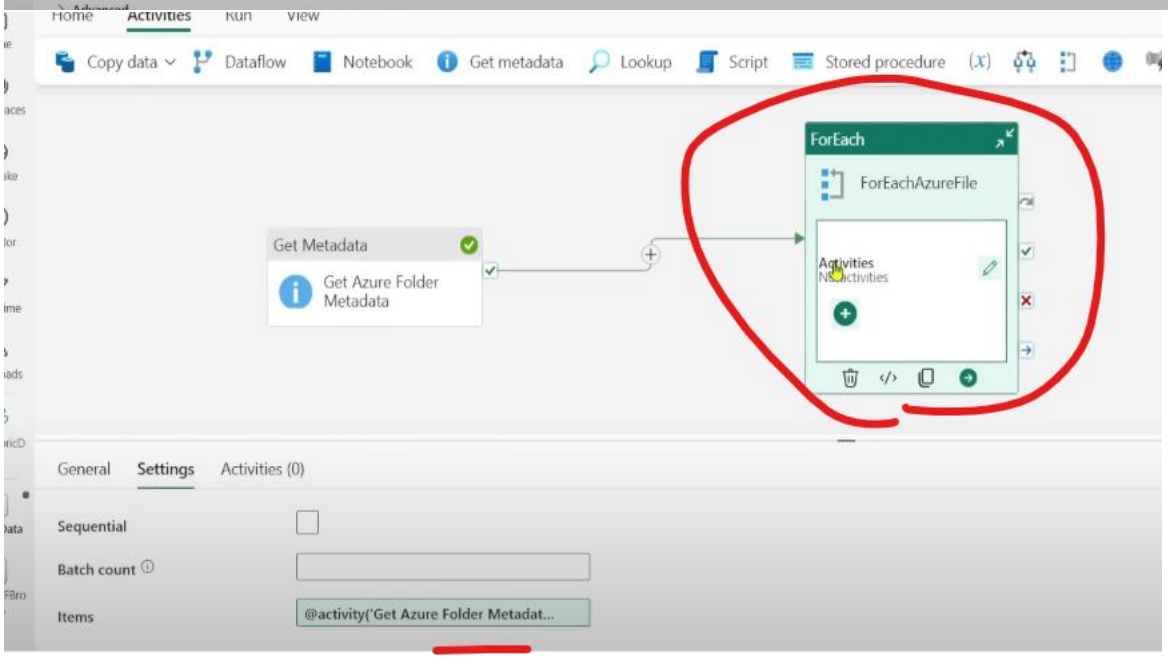
Problem statement2: Get a list of files from ADLS and keep it in Fabric LH

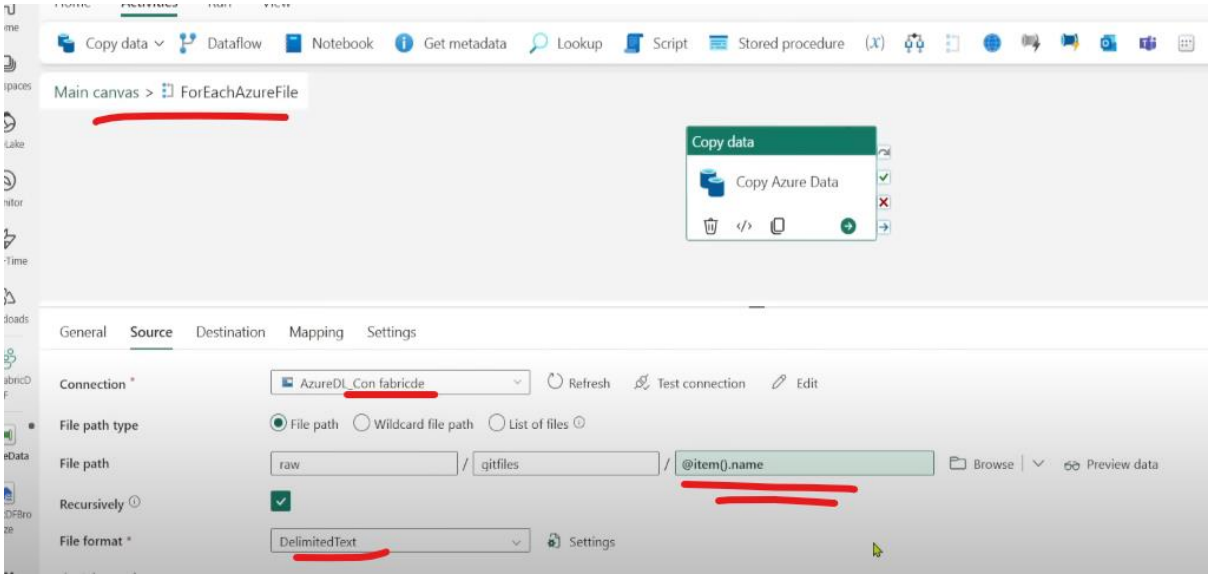
Give Storage Blob Data Contributor access of ADLS to Fabric



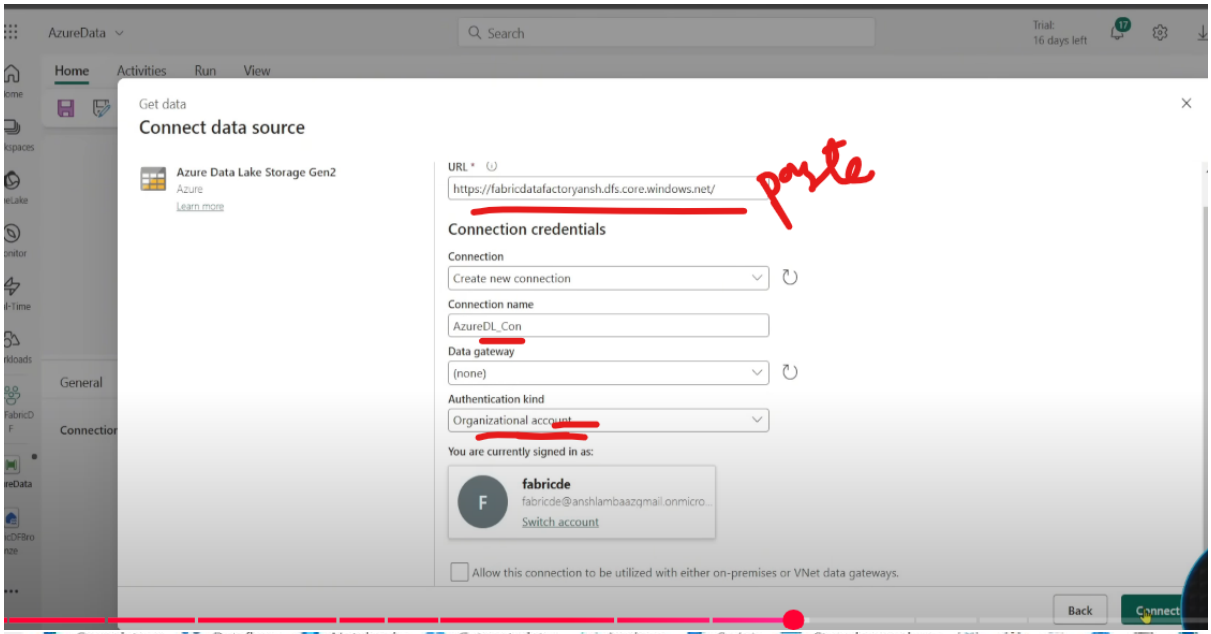
Use a Get Metadata activity to get all the File names

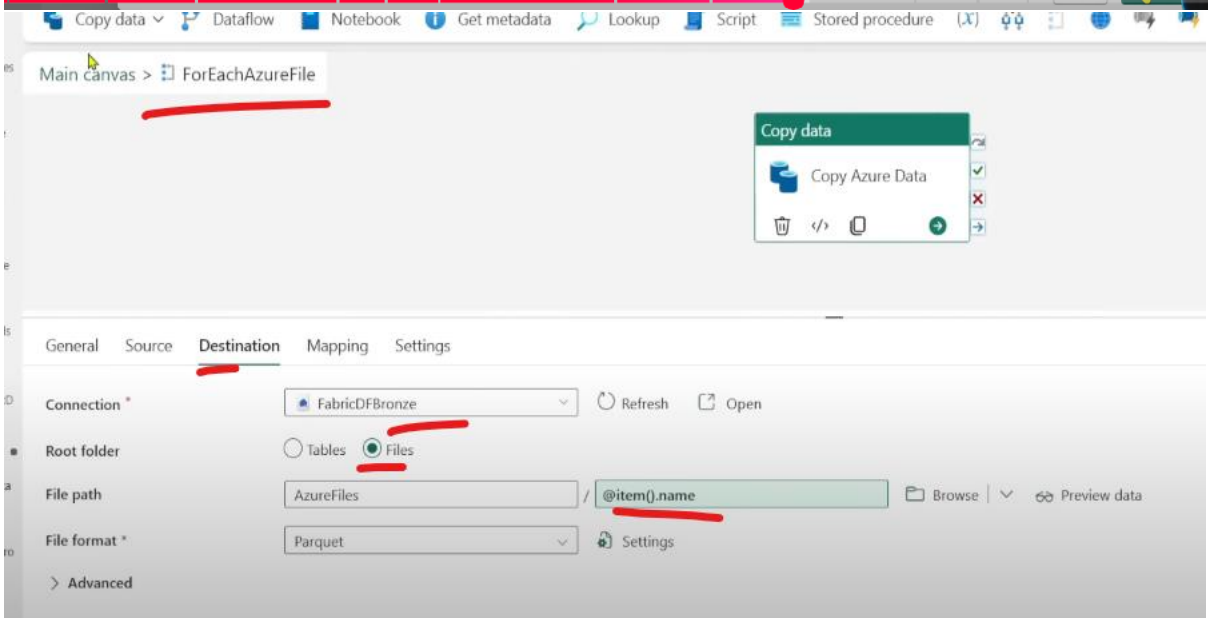




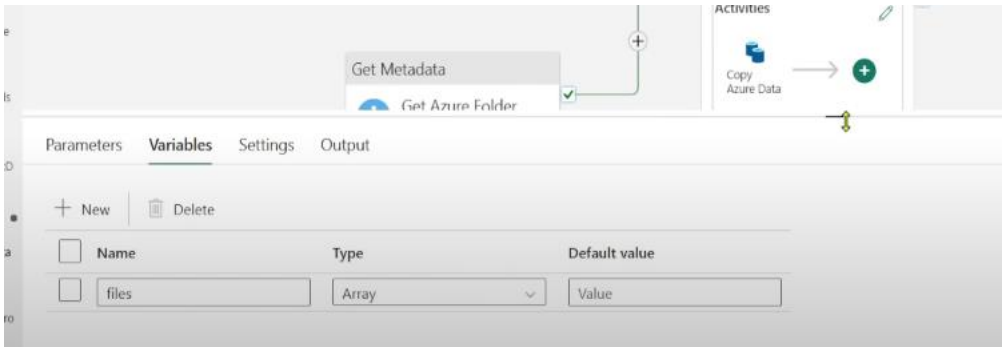


Copy activity → Source → connection for ADLS Gen2

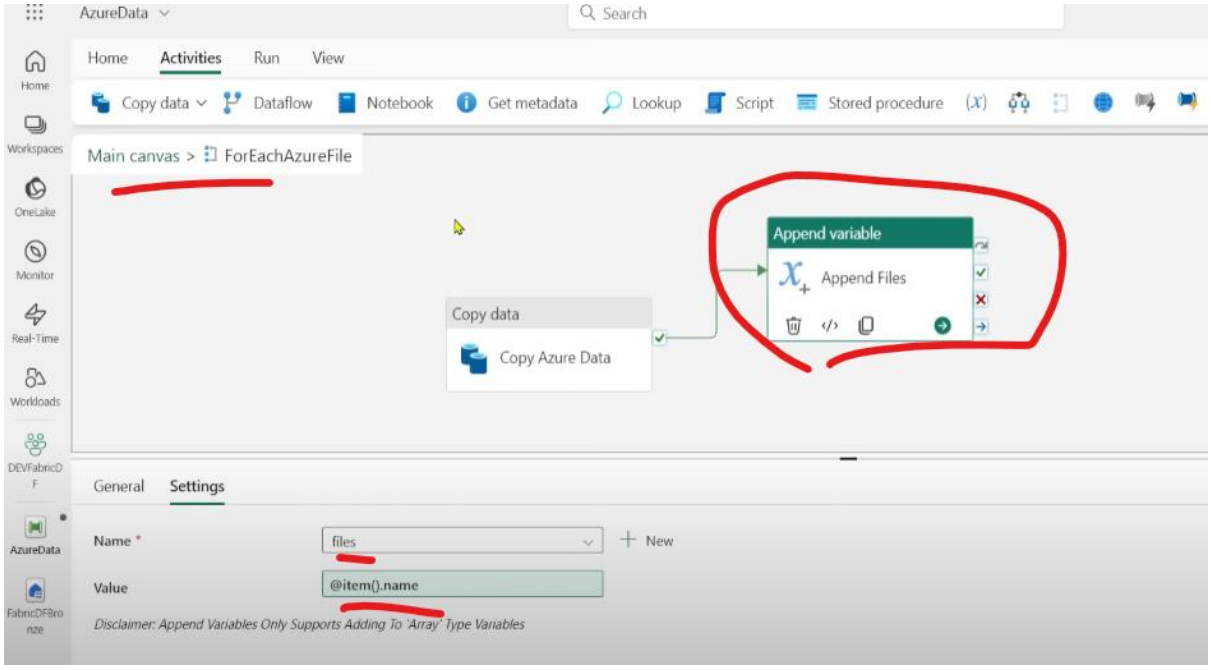




Lets enhance the pipeline by adding logging mechanism
Create a variable, everytime the files are getting copied, the information needs to be stored in this variable

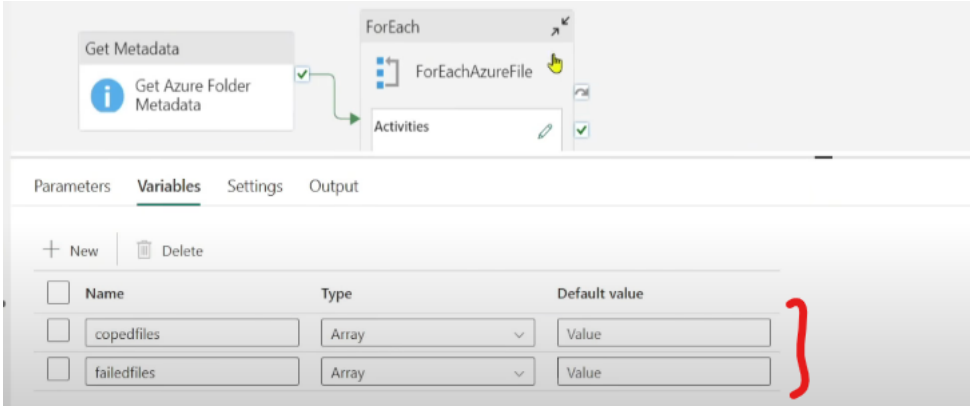


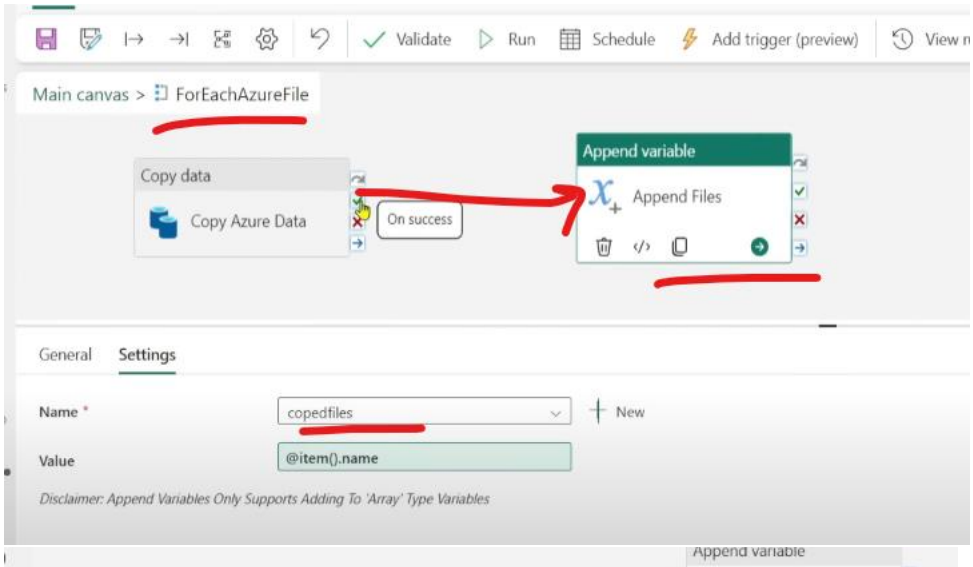
Add a Append variable activity after Copy to hold all the names

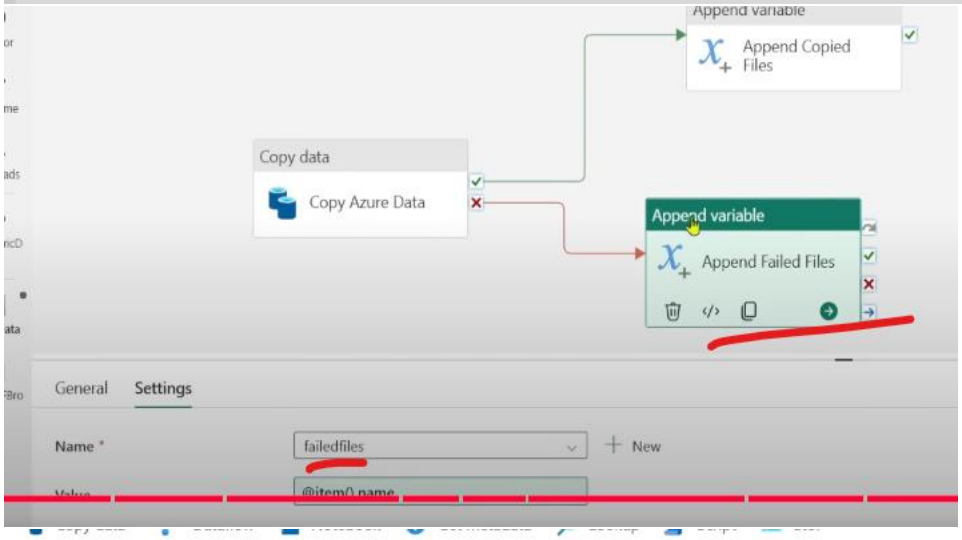


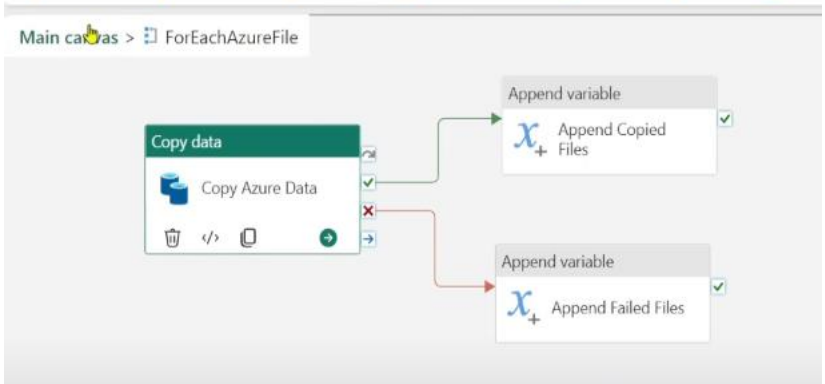
Lets enhance it more, need to track both successful and failed files

create 2 variables

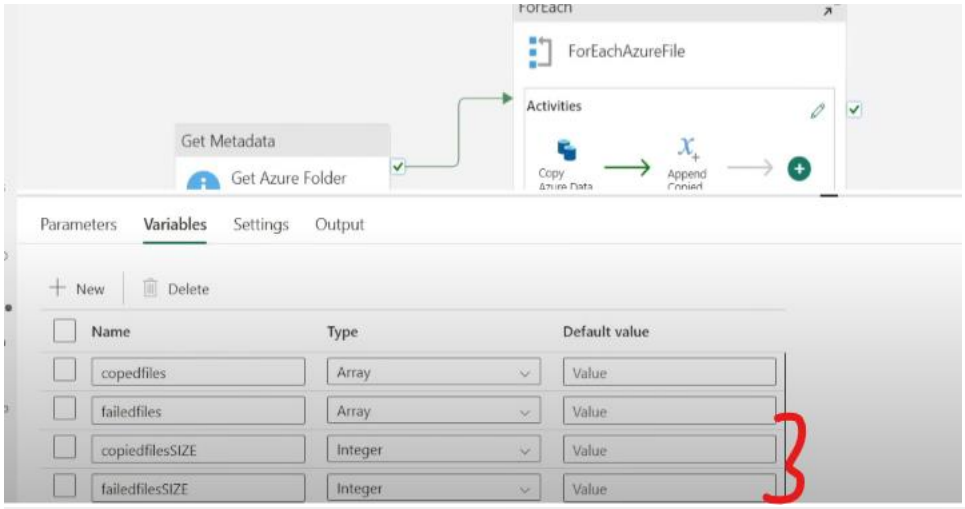


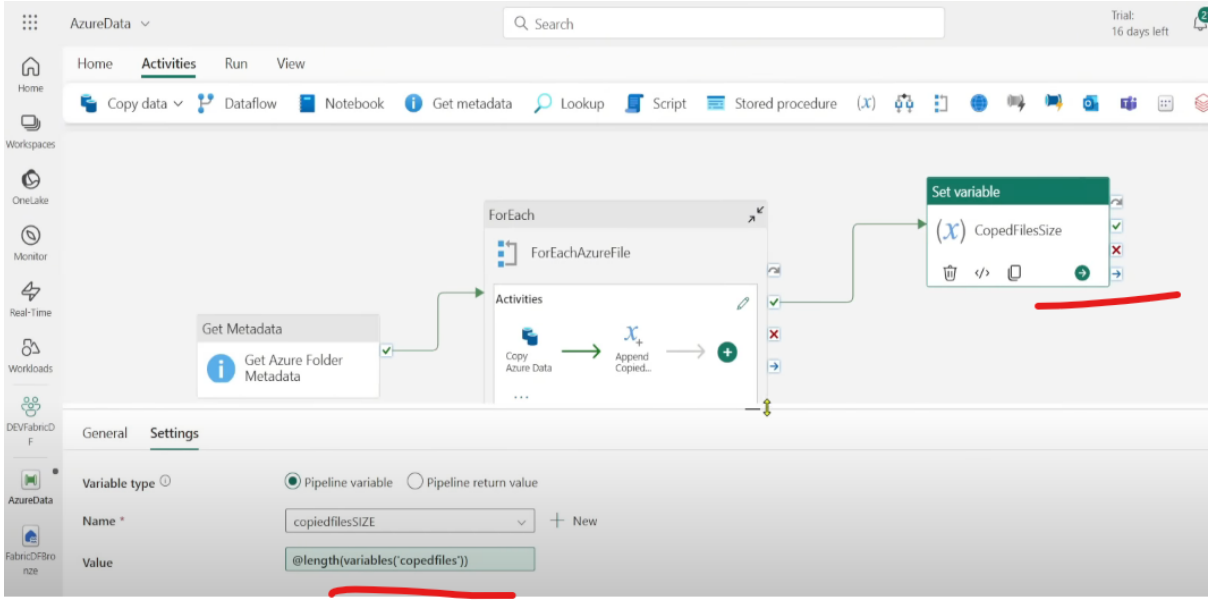


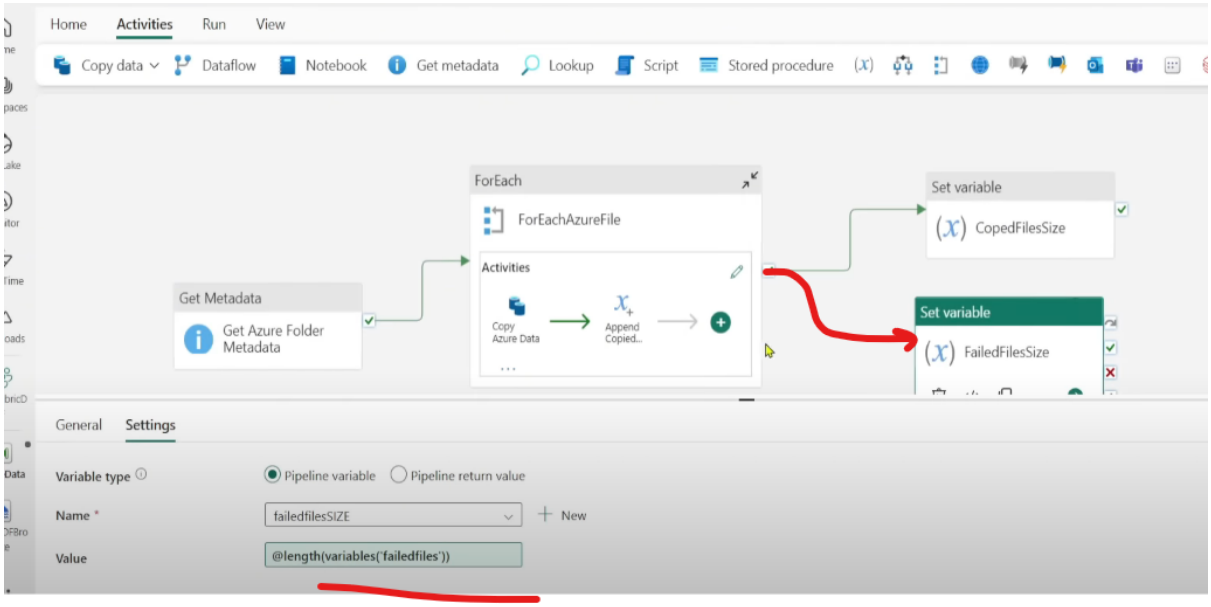




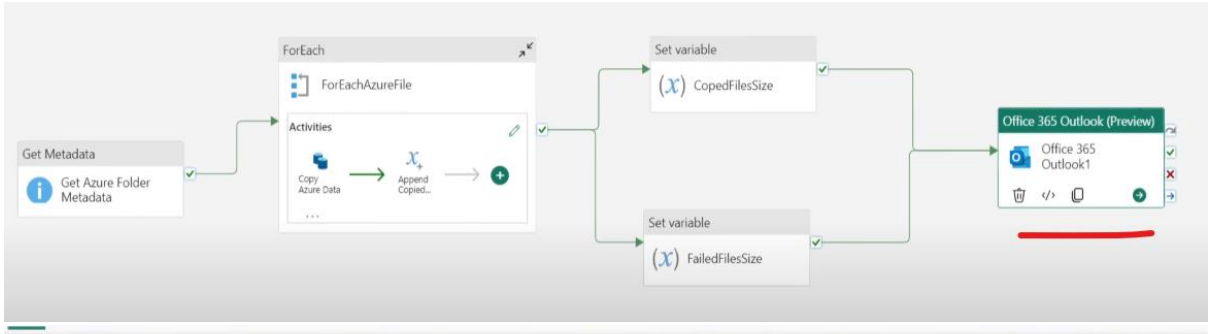
the data of 2 arrays inside the loop also needs to be stored, create 2 variables for that

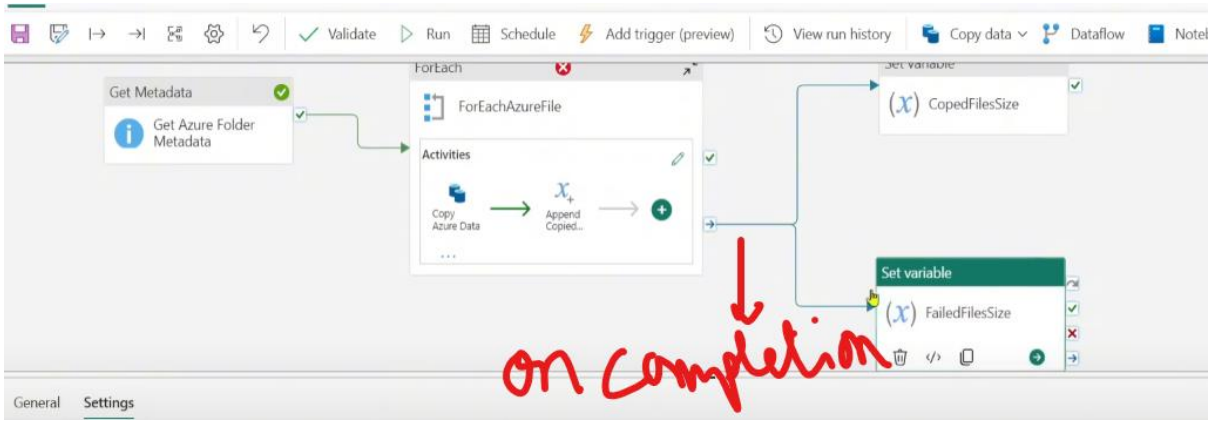






Now add Notifications for these activities

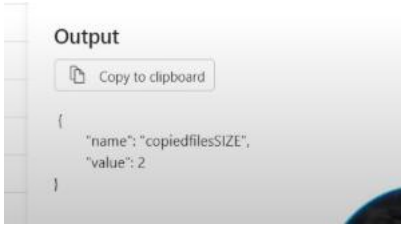


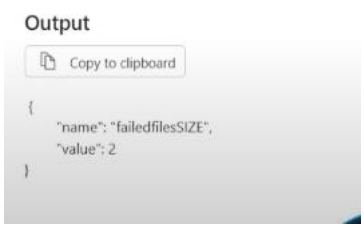


Just a change, link both the Set Variable activities with On completion with For_each (not with success)

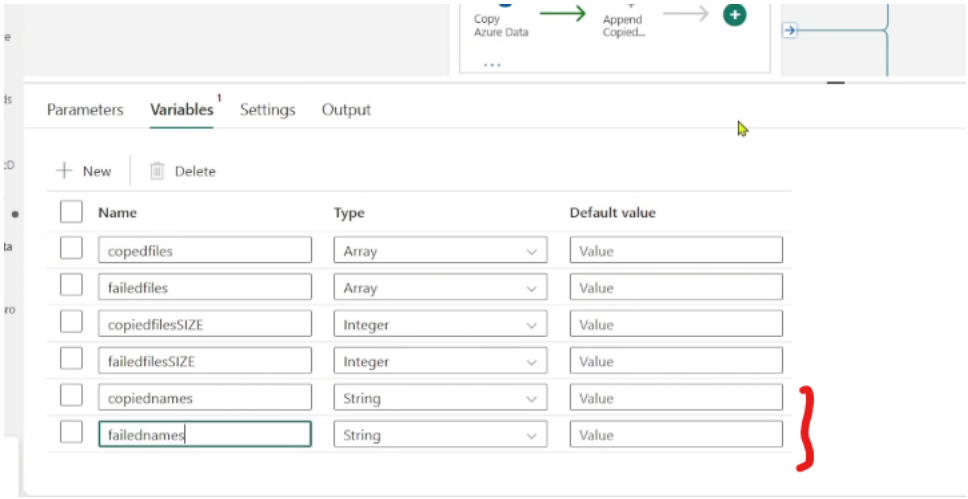
Delete the last activity and run the pipeline

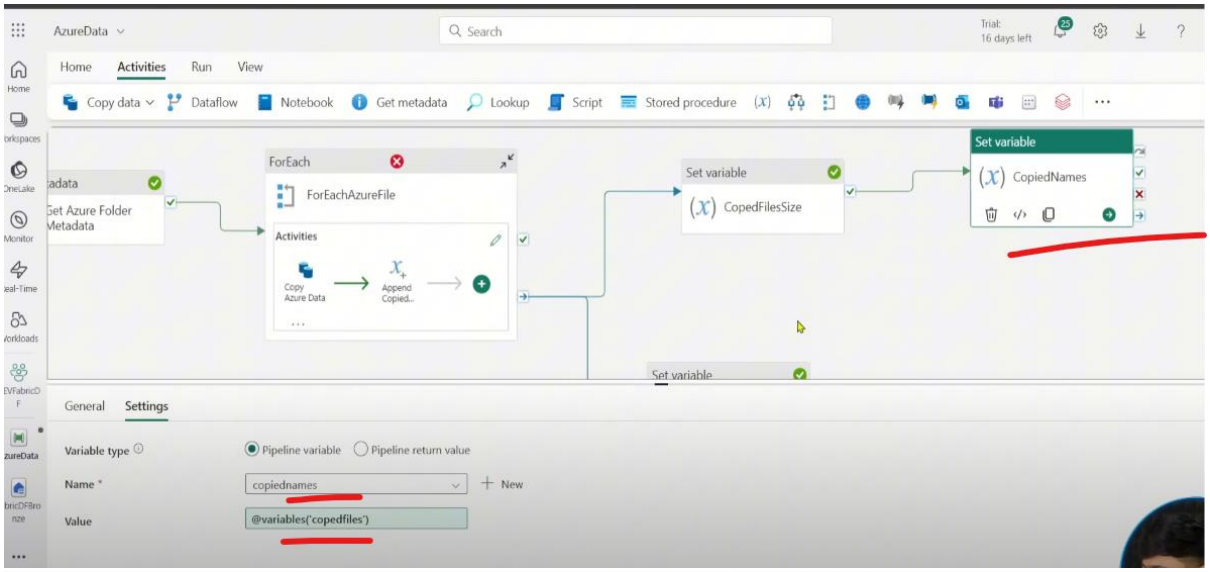
Because of bad data

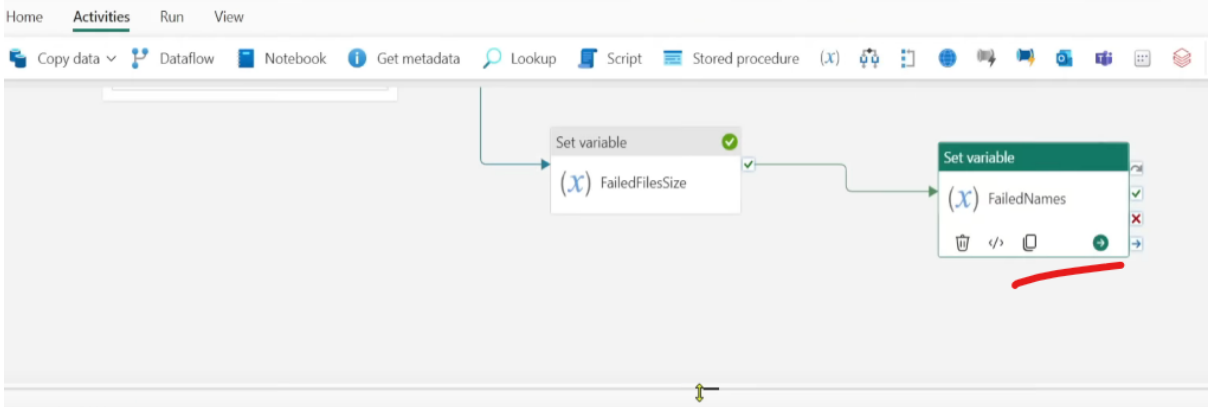


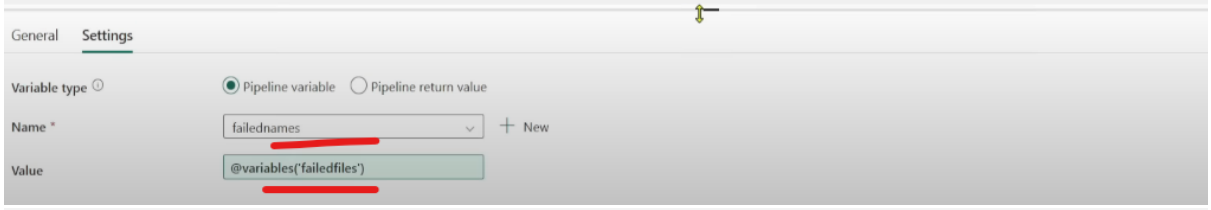


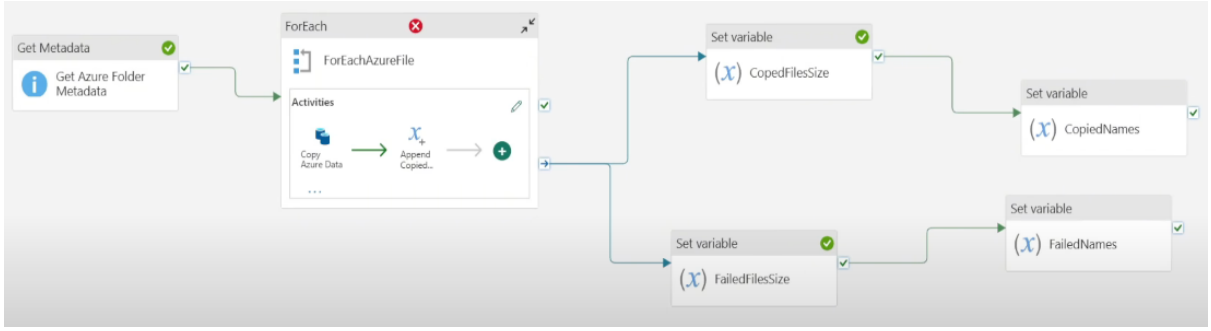
I also want to store the details of failed names, create 2 more variables

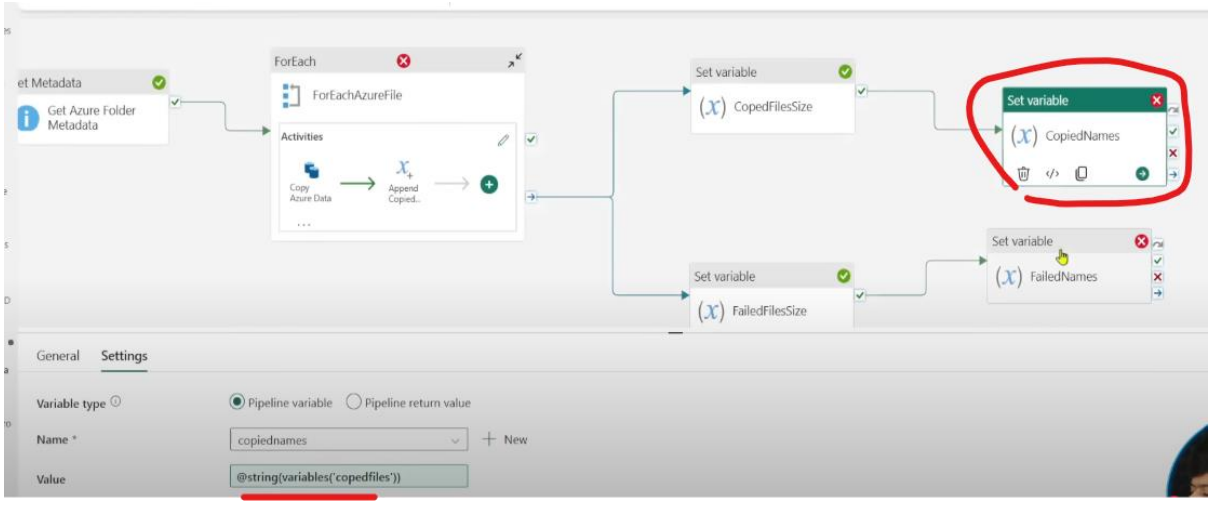



AzureData
Search
Trial: 16 days left
Home
Activities
Run
View
Copy data
Dataflow
Notebook
Get metadata
Lookup
Script
Stored procedure
ForEach
ForEachAzureFile
Activities
Copy Azure Data
Append Copied...
Get Azure Folder Metadata
Set variable
CopedFilesSize
Set variable
CopiedNames
General
Settings
Variable type
Pipeline variable
Pipeline return value
Name *
copiednames
New
Value
@variables('copedfiles')

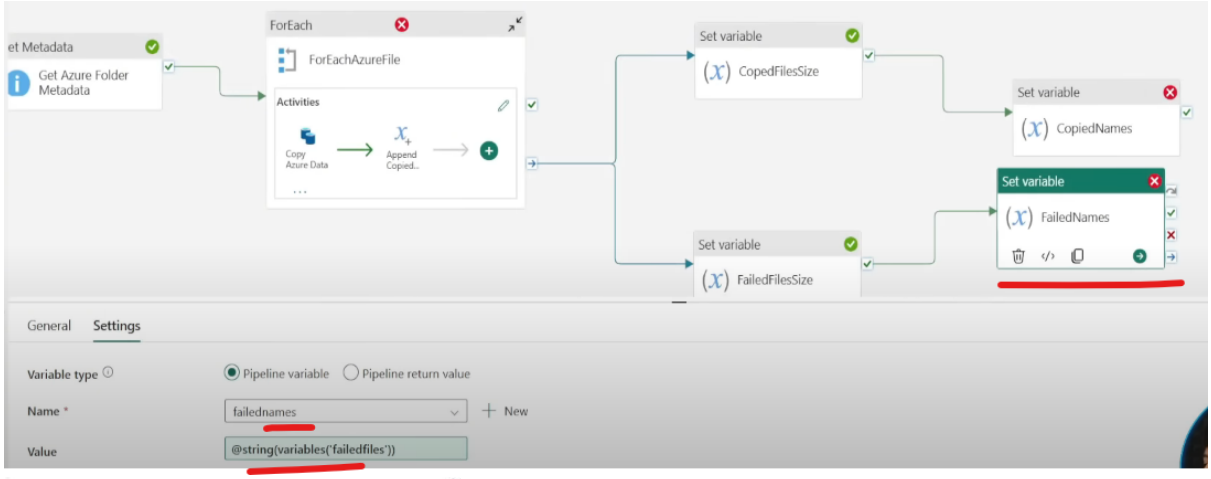

Home
Activities
Run
View
Copy data
Dataflow
Notebook
Get metadata
Lookup
Script
Stored procedure
Set variable
FailedFilesSize
Set variable
FailedNames

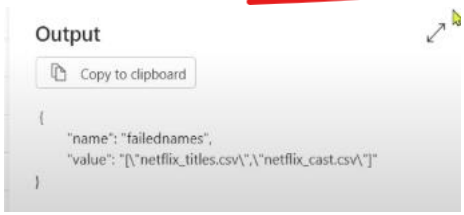

General
Settings
Variable type
Pipeline variable
Pipeline return value
Name *
failednames
New
Value
@variables('failedfiles')

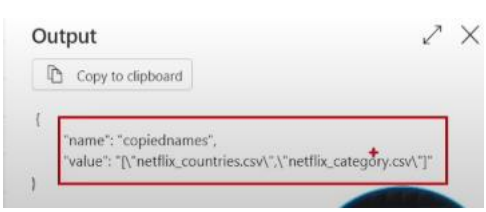

Get Metadata
Get Azure Folder Metadata
ForEach
ForEachAzureFile
Activities
Copy Azure Data
Append Copied...
Set variable
CopedFilesSize
Set variable
CopiedNames
Set variable
FailedFilesSize
Set variable
FailedNames



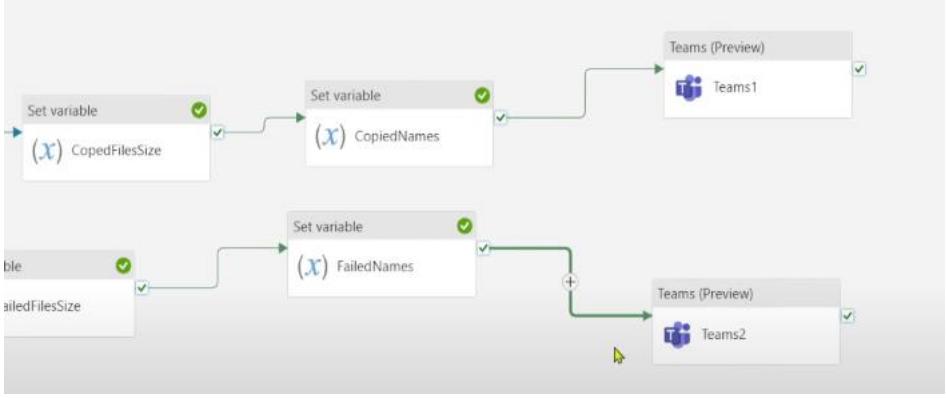
add string to hold array variable



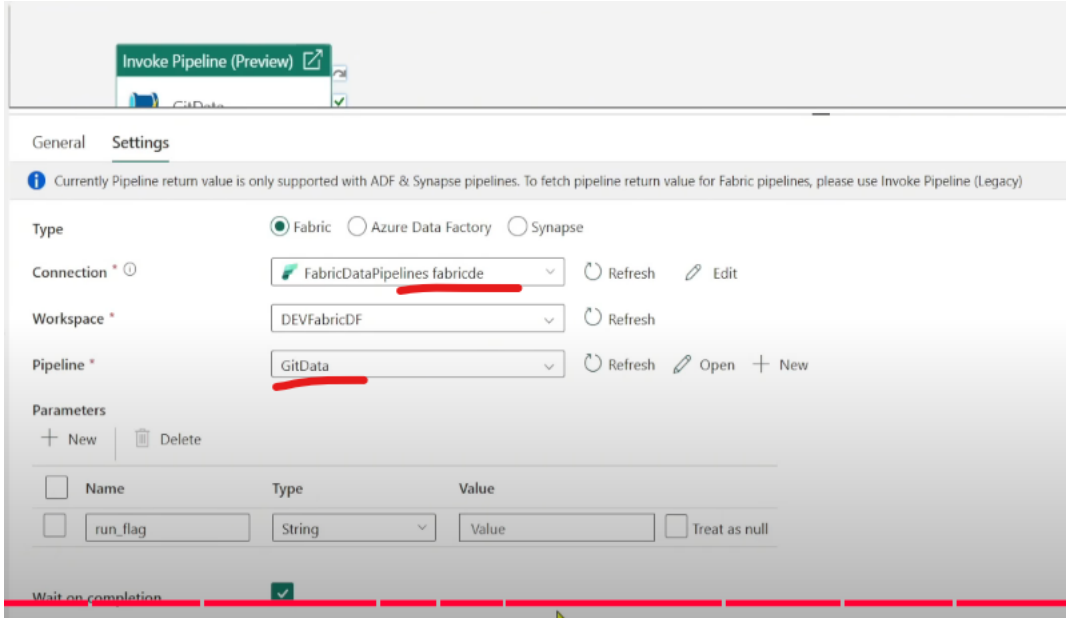




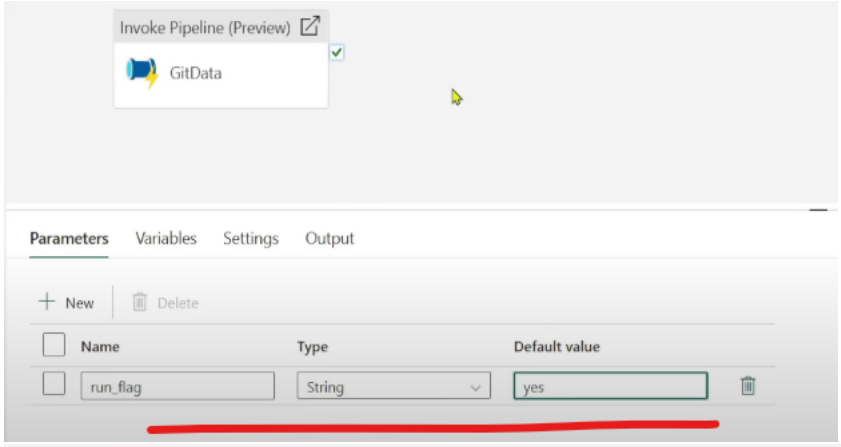
Lets add Teams notification activity at the end

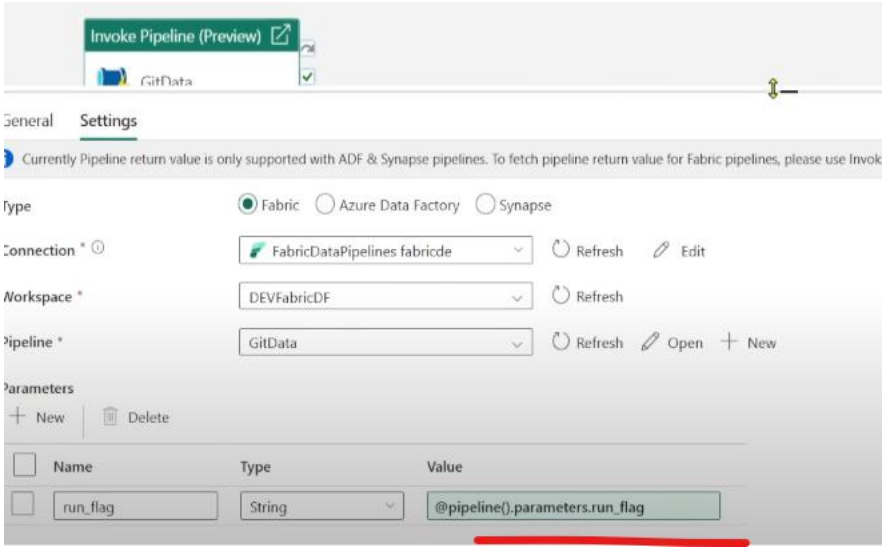


Create a parent pipeline → use Invoke pipeline activity



create a pipeline parameter and feed it to child pipeline

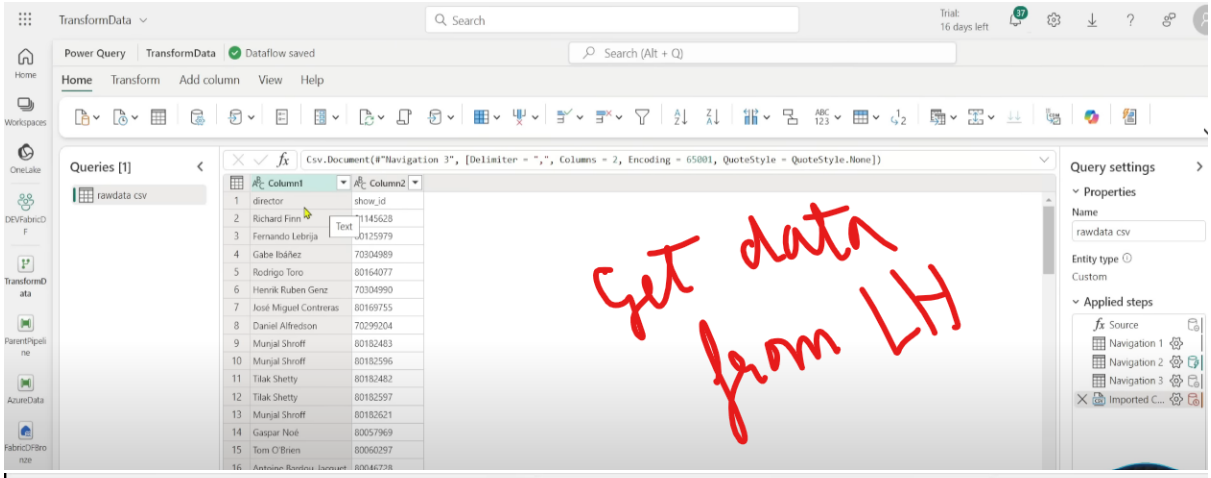


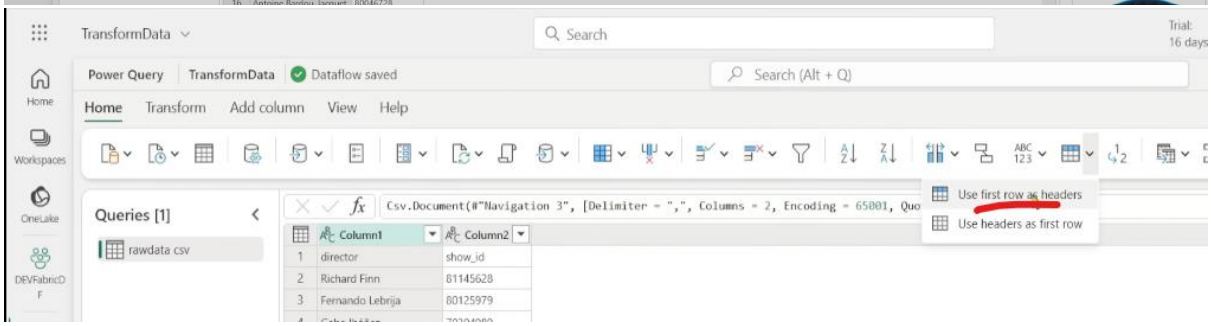


Transformation

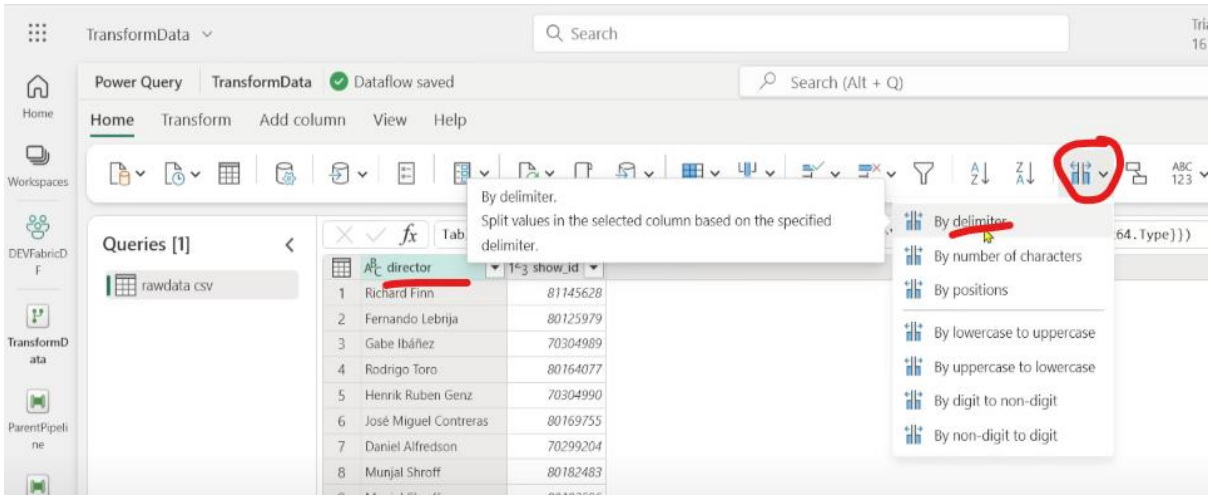
Data Flow Gen1 (backdated, use Gen2)

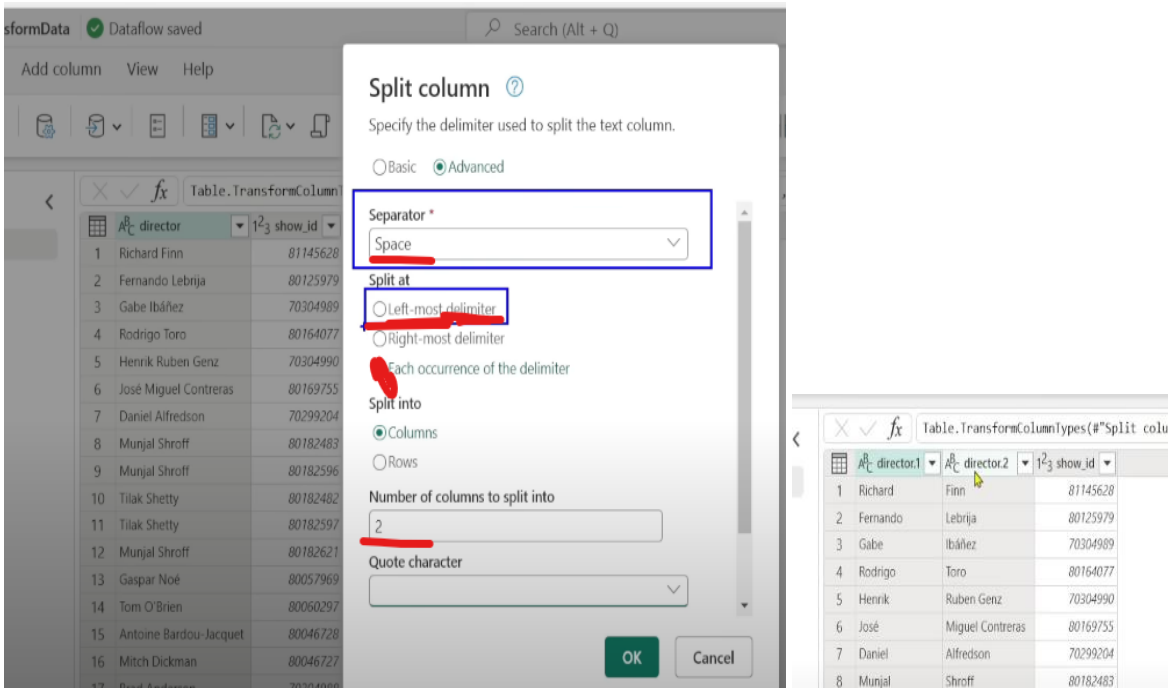
Create Data Flow Gen2 (Power Query online integrated with Fabric)→ Low code/no code manner
You can perform ETL in DFG2 but not ELT, but in Data Pipelines you can do ELT.



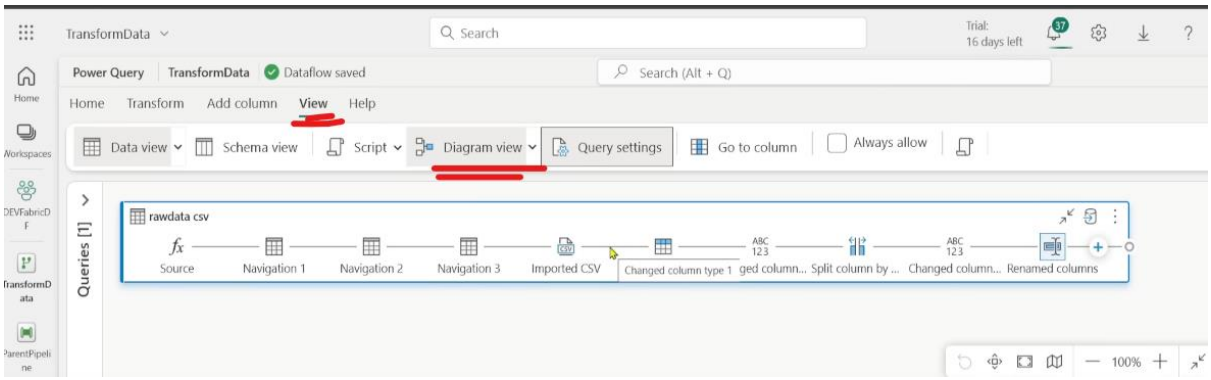


Wanna split the directors name

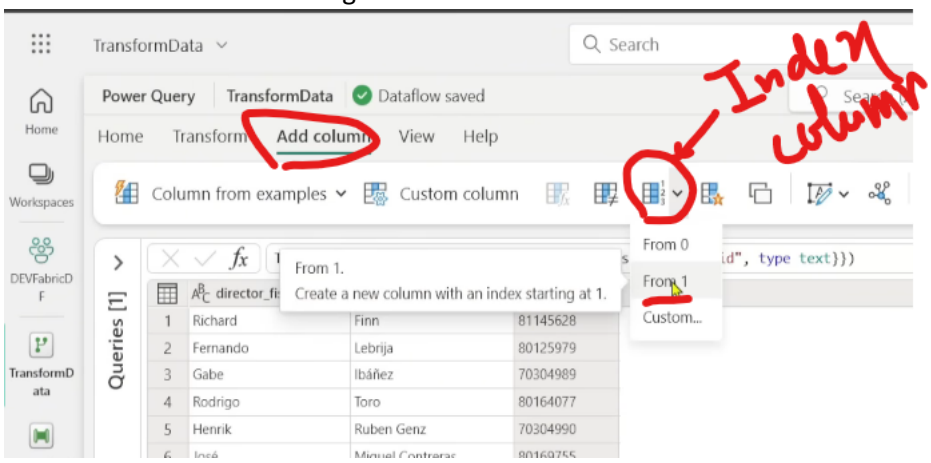


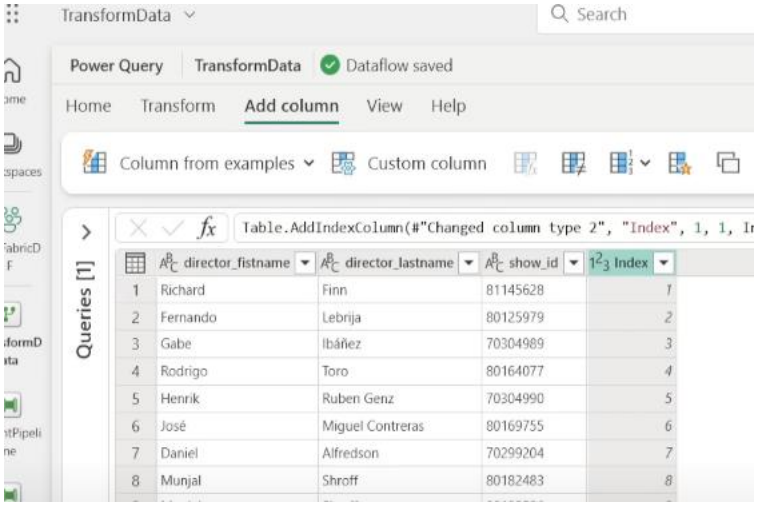


Rename the headers

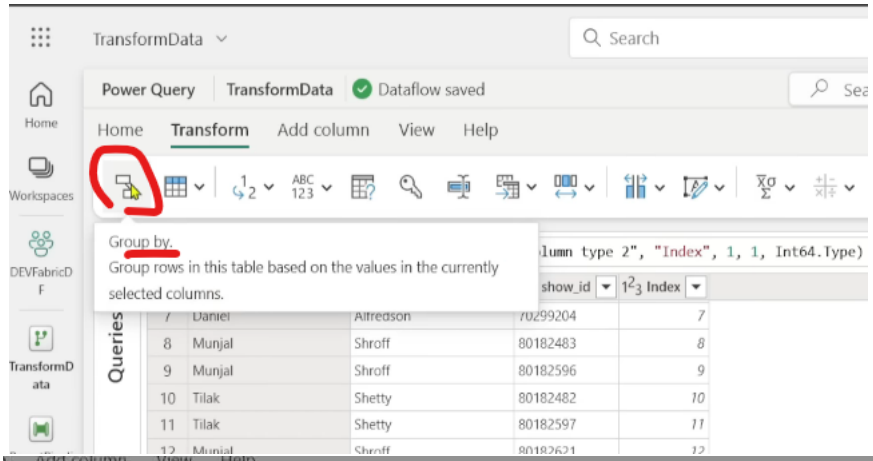


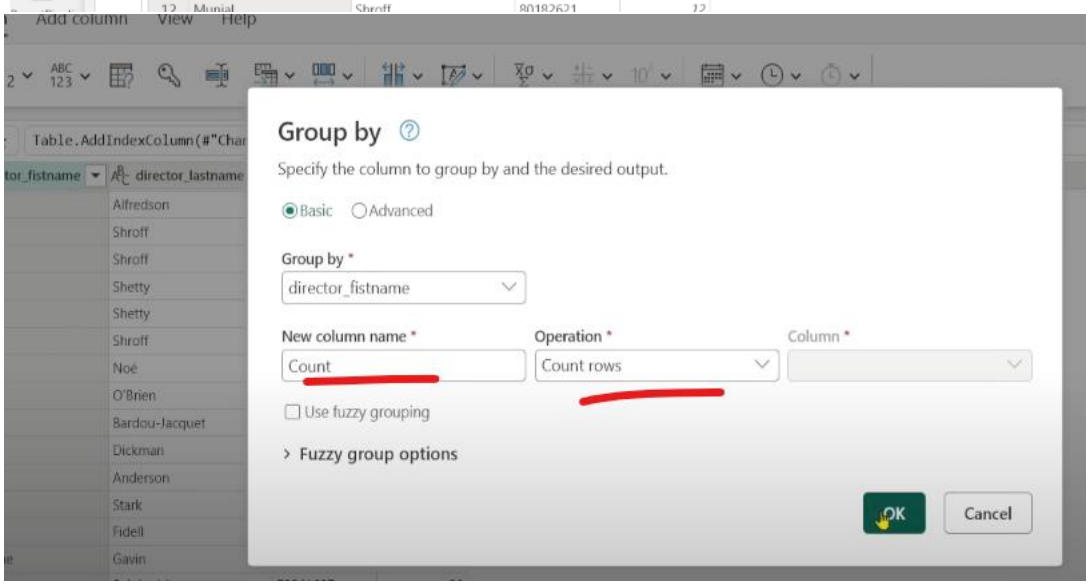
Add a new column containing row number

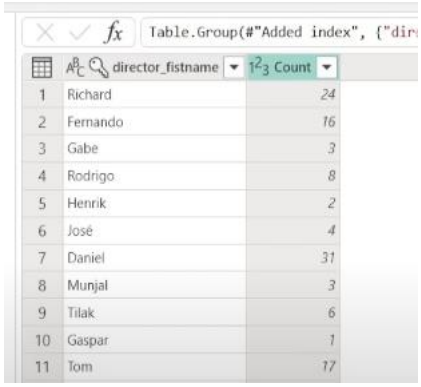




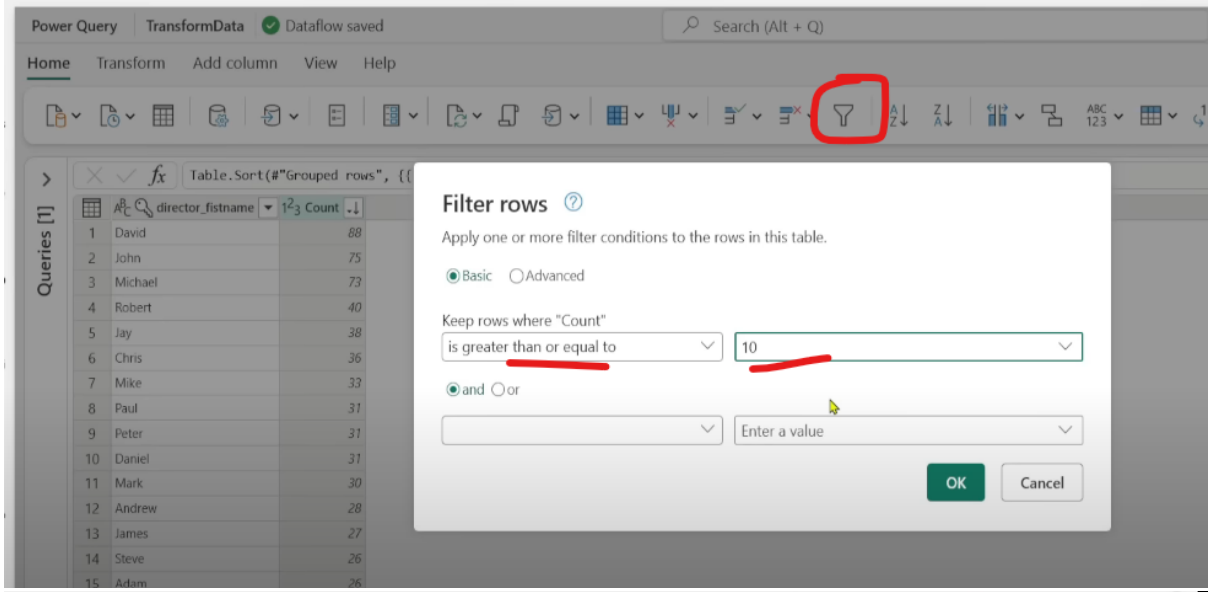
I want to see how many movies are done by the same director, Group by on directors

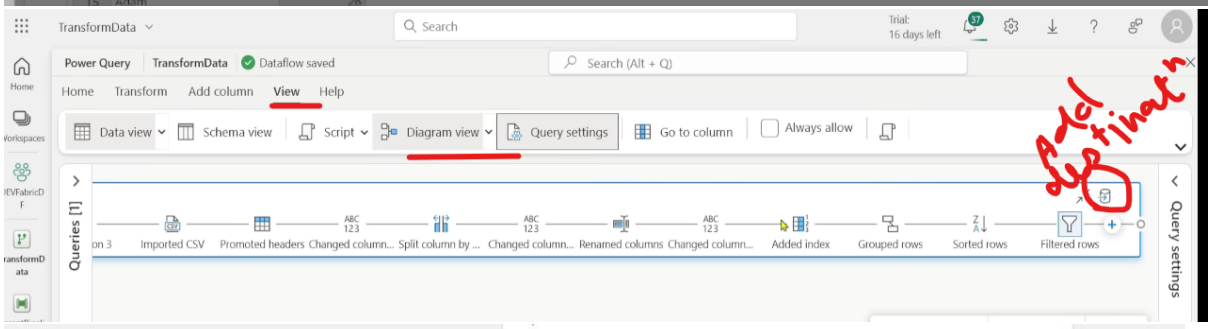


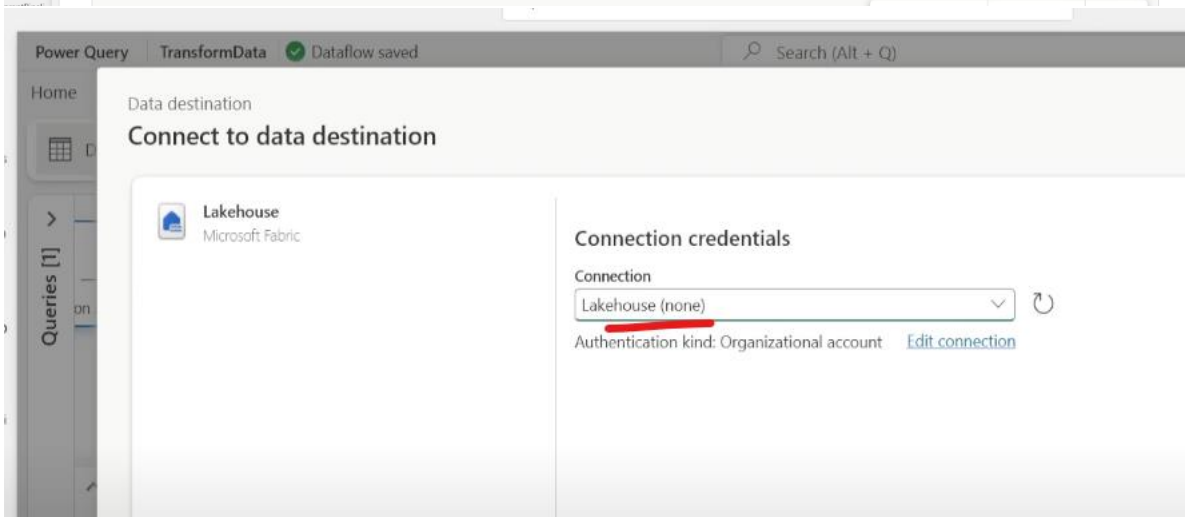


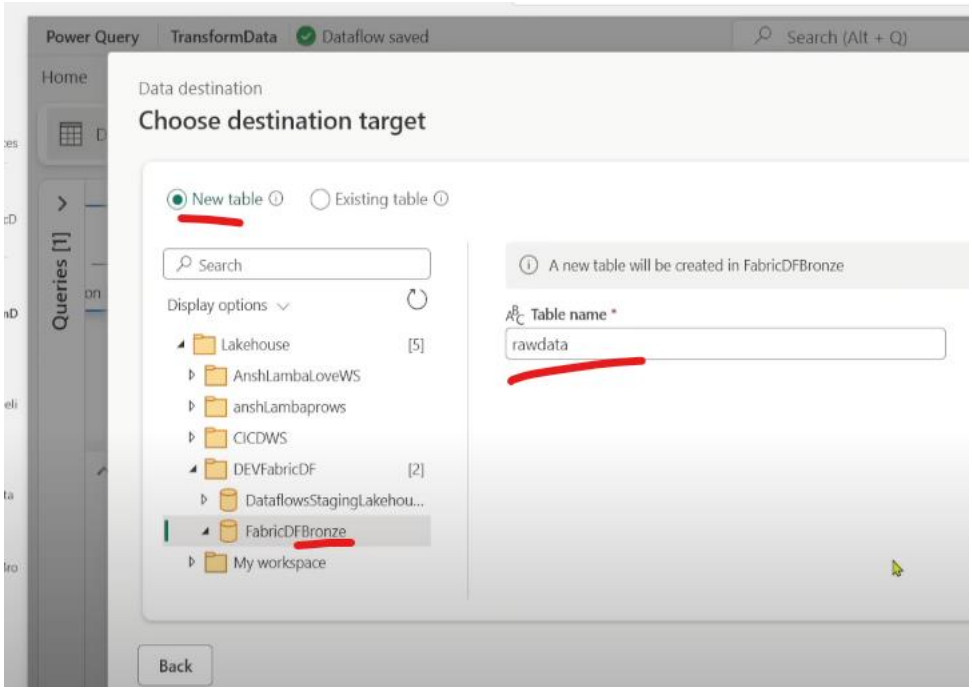


Filter only those if count more than 10



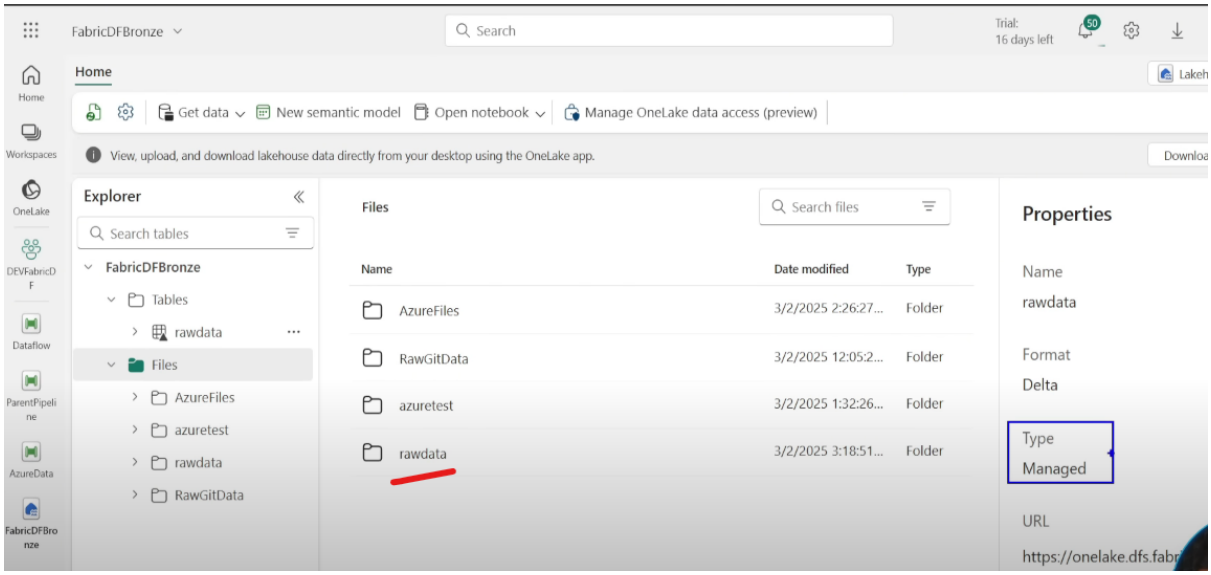


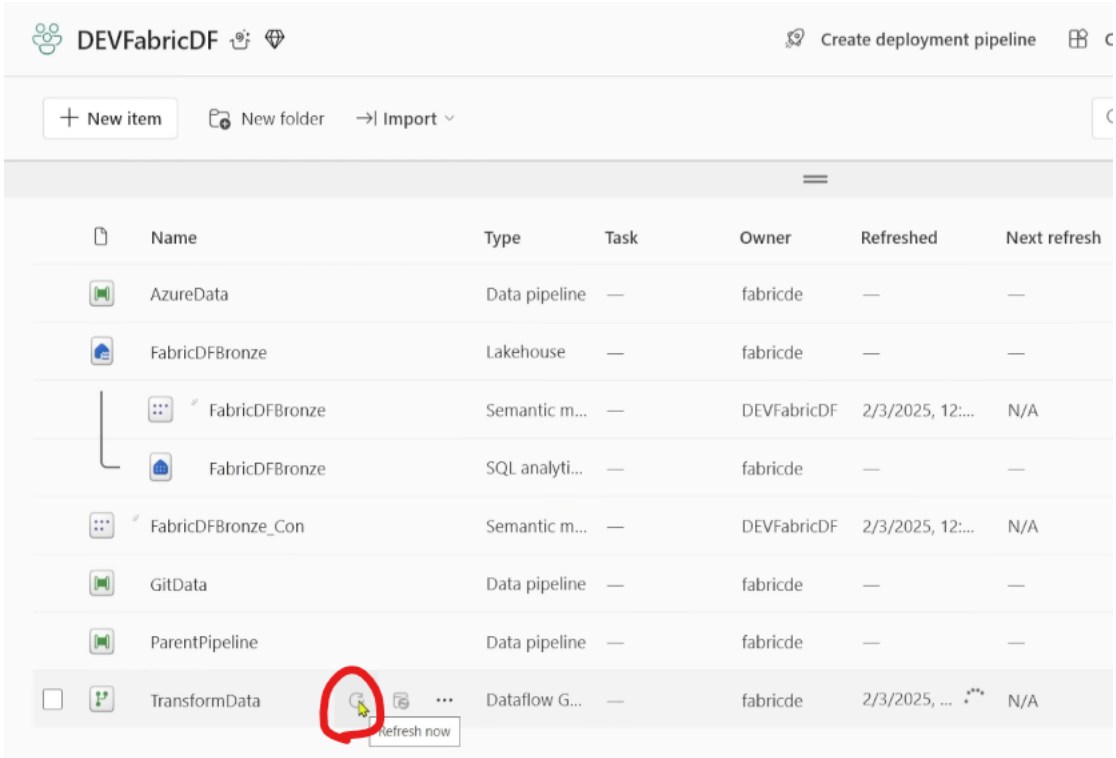




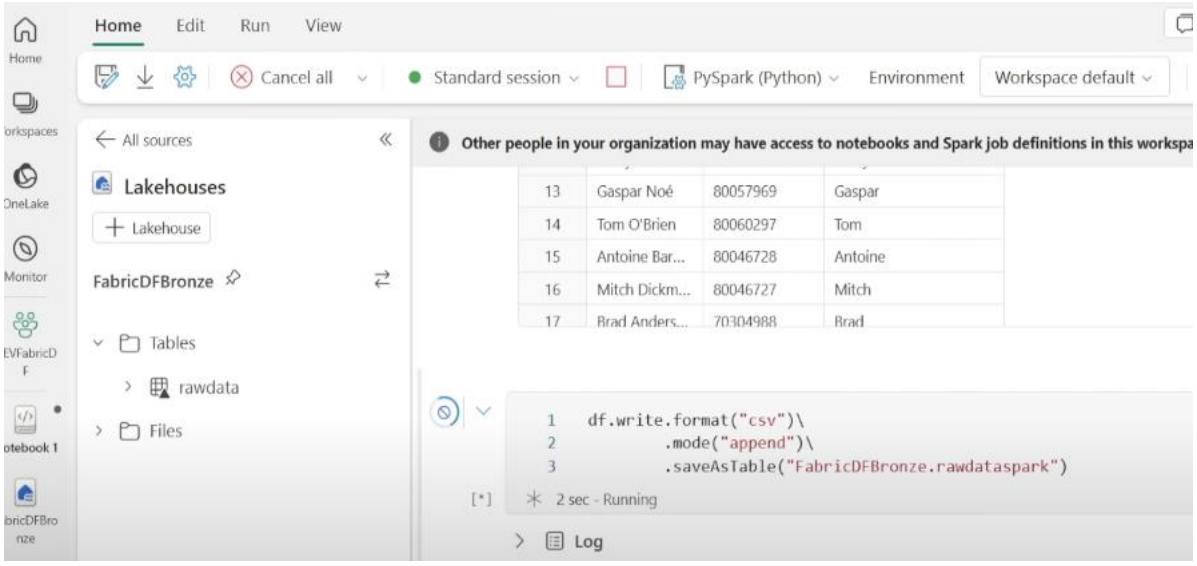
Publish the Data Flow, run it

We can run Data Flows in 2 ways: 1)manual 2)orchestrate with Data Pipeline



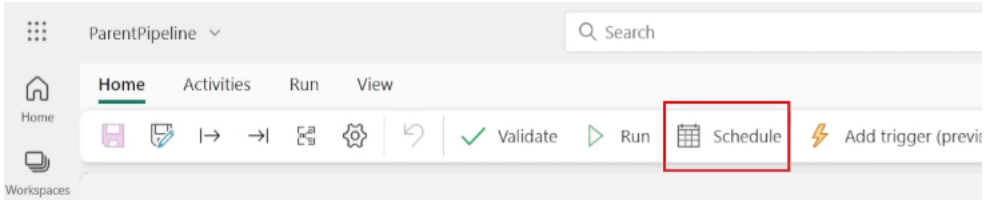


Data Transformation using Notebook

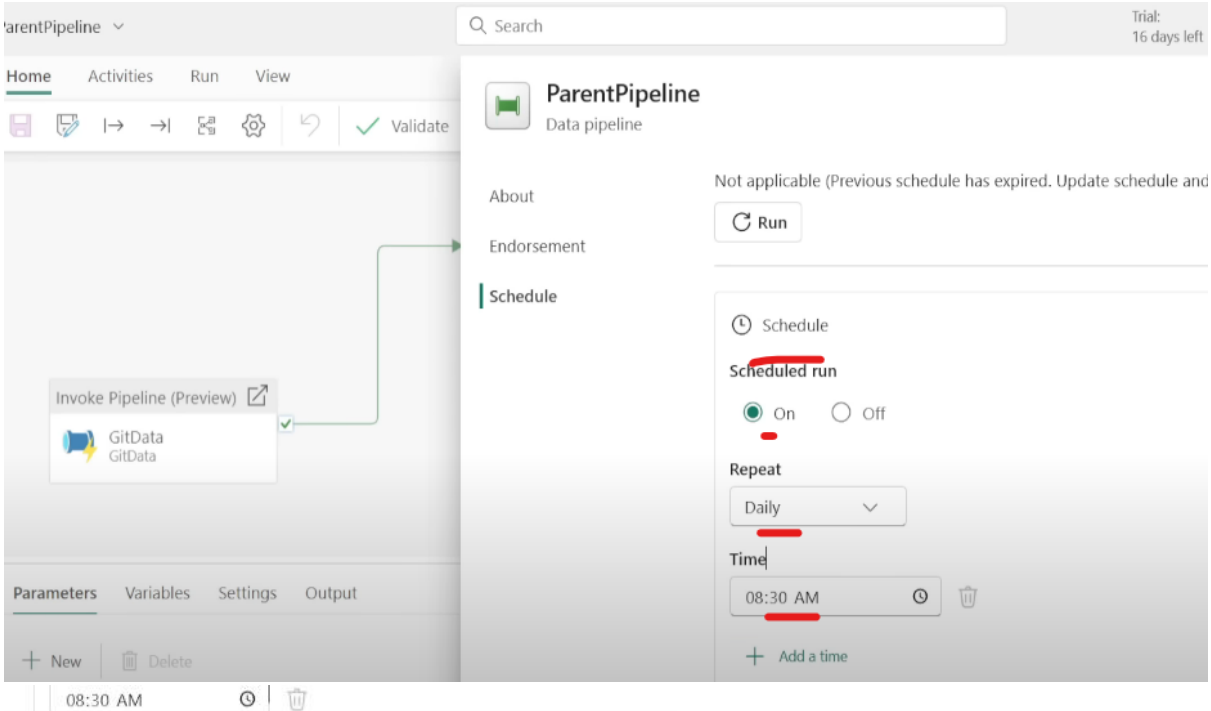


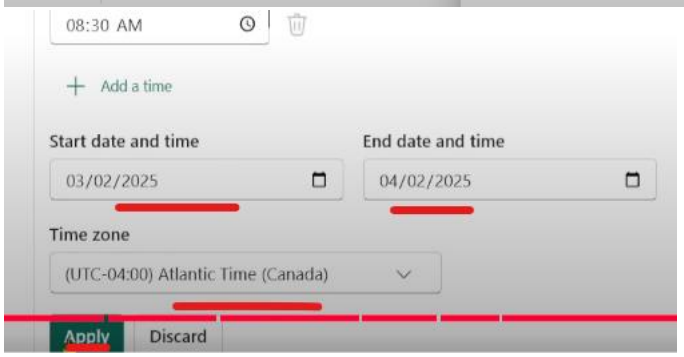
No triangle because it's a csv table managed

Use Notebook activity in Data Pipeline to orchestrate the flow



Add Schedule Trigger to automate the flow


arentPipeline
Search
Trial:
16 days left
Home
Activities
Run
View
Validate
ParentPipeline
Data pipeline
About
Endorsement
Schedule
Not applicable (Previous schedule has expired. Update schedule and
Run
Schedule
Scheduled run
On
Off
Repeat
Daily
Time
08:30 AM
Add a time
Invoke Pipeline (Preview)
GitData
GitData
Parameters
Variables
Settings
Output
New
Delete


08:30 AM
Add a time
Start date and time
03/02/2025
End date and time
04/02/2025
Time zone
(UTC-04:00) Atlantic Time (Canada)
Apply
Discard