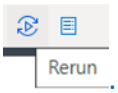
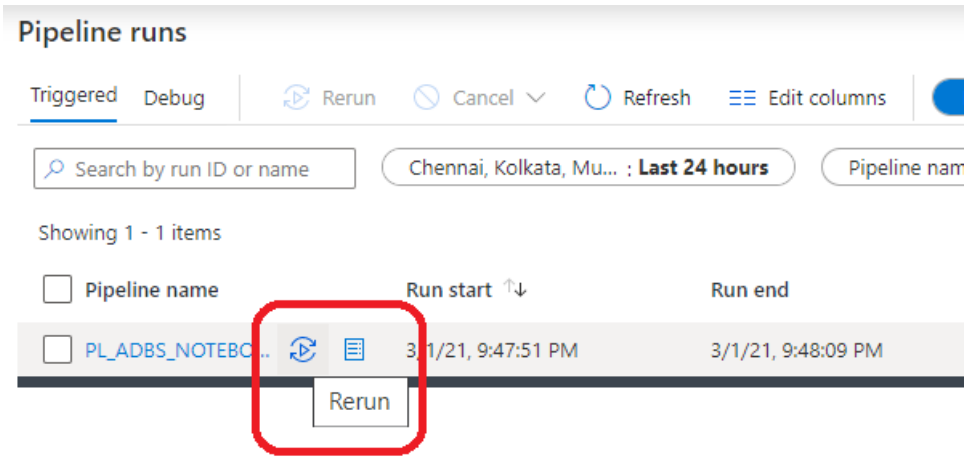
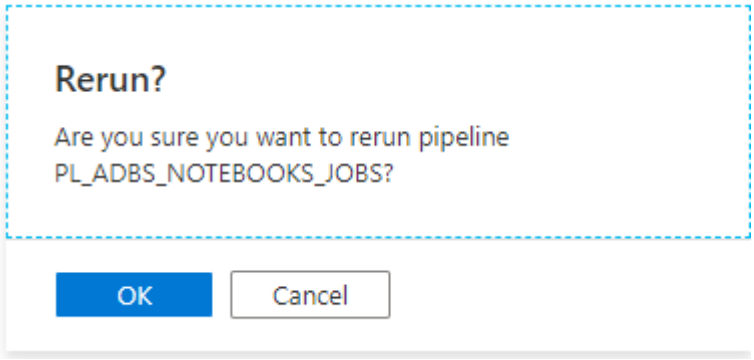
Azure Data Factory Advanced Interview Questions And Answers Youtube: TechLake

1) How to rerun a pipe line from Data Factory Monitor.

Ans: Simply navigate to the 'Monitor' => Pipeline Runs => Triggered => Click below red circle rerun icon

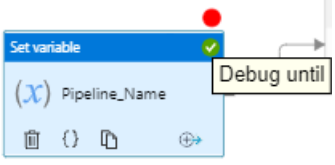






2) If you have 15 activities in pipeline, if we want to debug only first 10 activities how do we do that?

Ans: Every Activity top red Circle will be there. If you select that "Debug Until" it will debug until that activity.

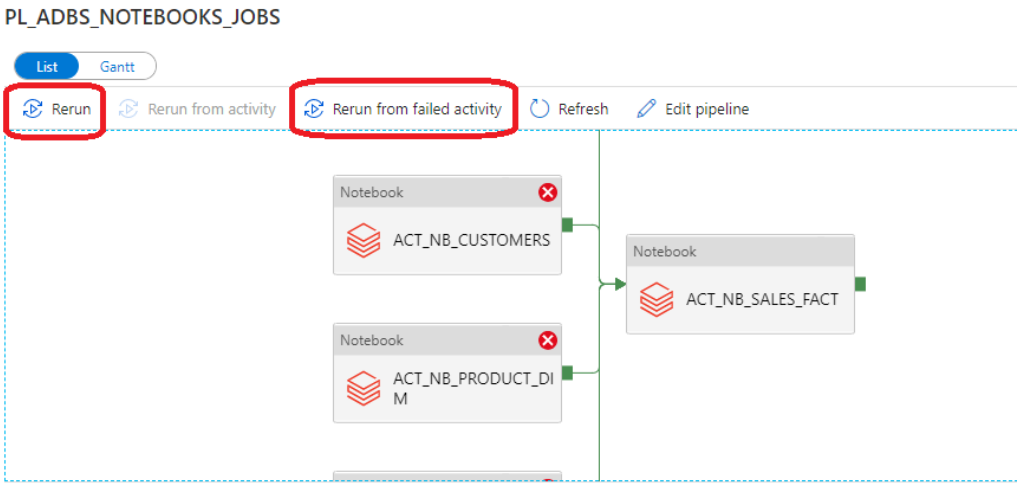


3) How to restart failed pipe line jobs in Azure data factory.

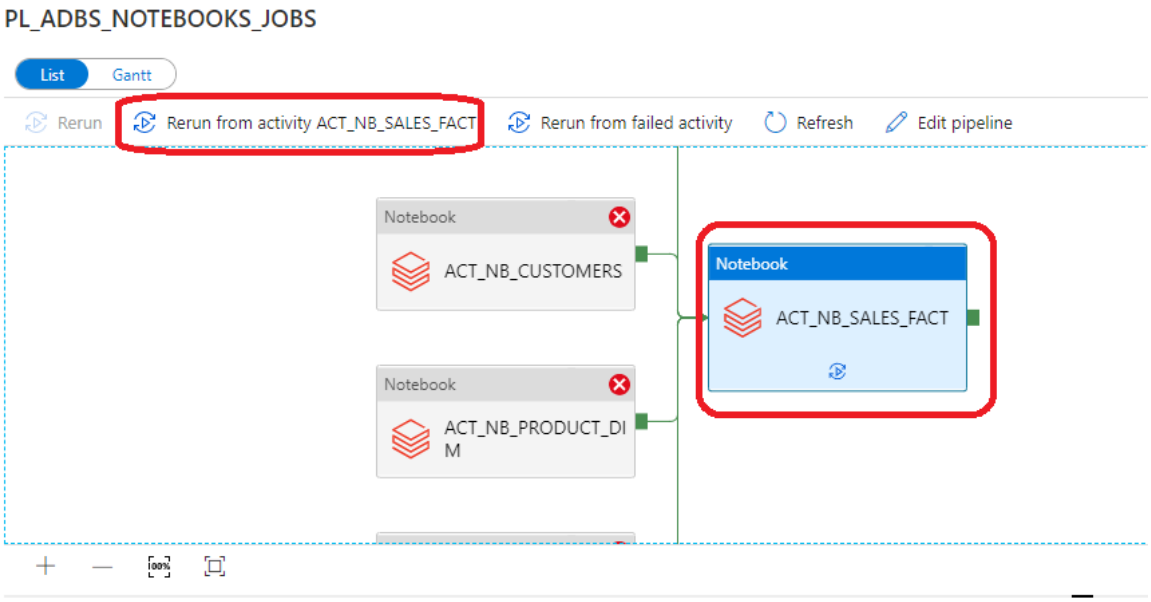
Filed pipelines will have multiple options.

i) **Rerun**: if you select rerun option it will return or restart from first step. If some steps are succeeded, those will be retrigger. So before triggering rerun option verify succeeded steps will be retriggered. There are some changes of reloading data again if its having copy activities.

ii) **Rerun From Failed Activity**: If an activity fails, times out, or is canceled, you can rerun the pipeline from that failed activity by selecting Rerun from failed activity.



iii) **Rerun from Activity**: If you wish to rerun starting at a specific point, you can do so from the activity runs view. Select the activity you wish to start from and select Rerun from activity.



4) How to make activity dependency in Pipeline Multiple Activities and those types.

Activity Dependency defines how subsequent activities depend on previous activities, determining the condition of whether to continue executing the next task. An activity can depend on one or multiple previous activities with different dependency conditions.

The different dependency conditions are: **Succeeded**, **Failed**, **Skipped**, and **Completed**.

For example, if a pipeline has Activity A -> Activity B, the different scenarios that can happen are:

- **succeeded** : Activity B has dependency condition on Activity A with **succeeded**: Activity B only runs if Activity A has a final status of succeeded
- **failed** : Activity B has dependency condition on Activity A with **failed**: Activity B only runs if Activity A has a final status of failed
- **completed** : Activity B has dependency condition on Activity A with **completed**: Activity B runs if Activity A has a final status of succeeded or failed
- **skipped**: Activity B has a dependency condition on Activity A with **skipped**: Activity B runs if Activity A has a final status of skipped. Skipped occurs in the scenario of Activity X -> Activity Y -> Activity Z, where each activity runs only if the previous activity succeeds. If Activity X fails, then Activity Y has a status of "Skipped" because it never executes. Similarly, Activity Z has a status of "Skipped" as well.

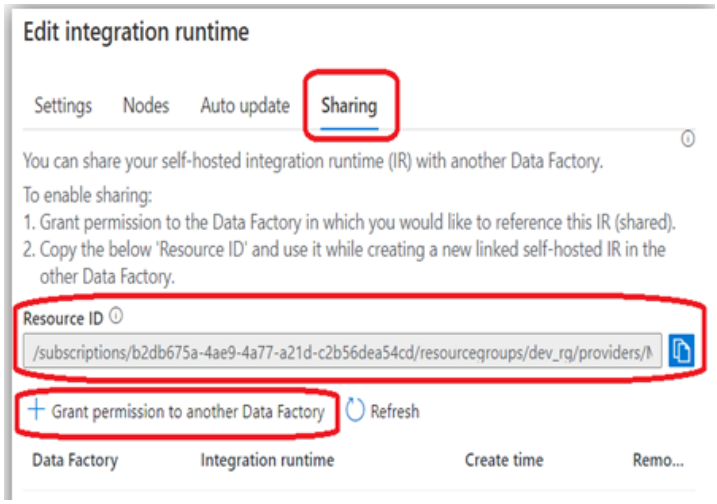
5) What is shared self-hosted Integration Runtime or How to Share Integration Runtime from One Data Factory to another Data Factory?

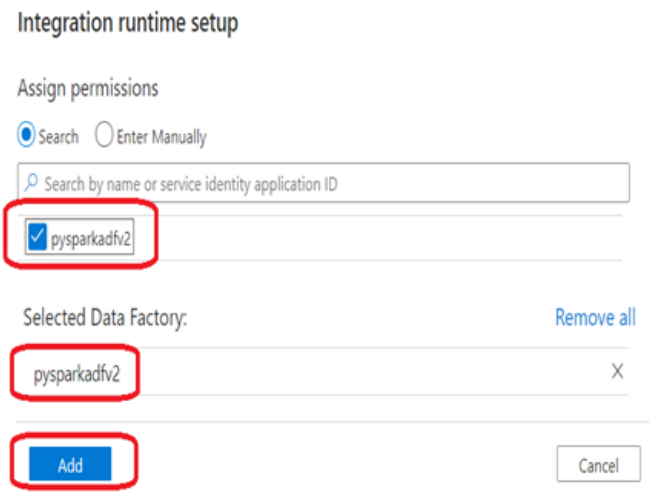
Shared IR: An original self-hosted IR that runs on a physical infrastructure.

Linked IR: An IR that references another shared IR. The linked IR is a logical IR and uses the infrastructure of another shared self-hosted IR.

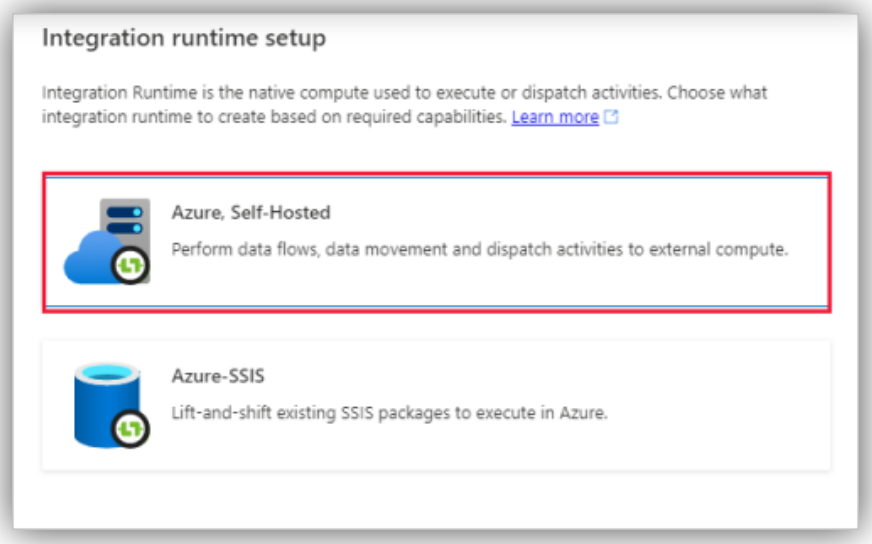
While Creating Self-Hosted Integration Runtime we will have separate option to Create Shared

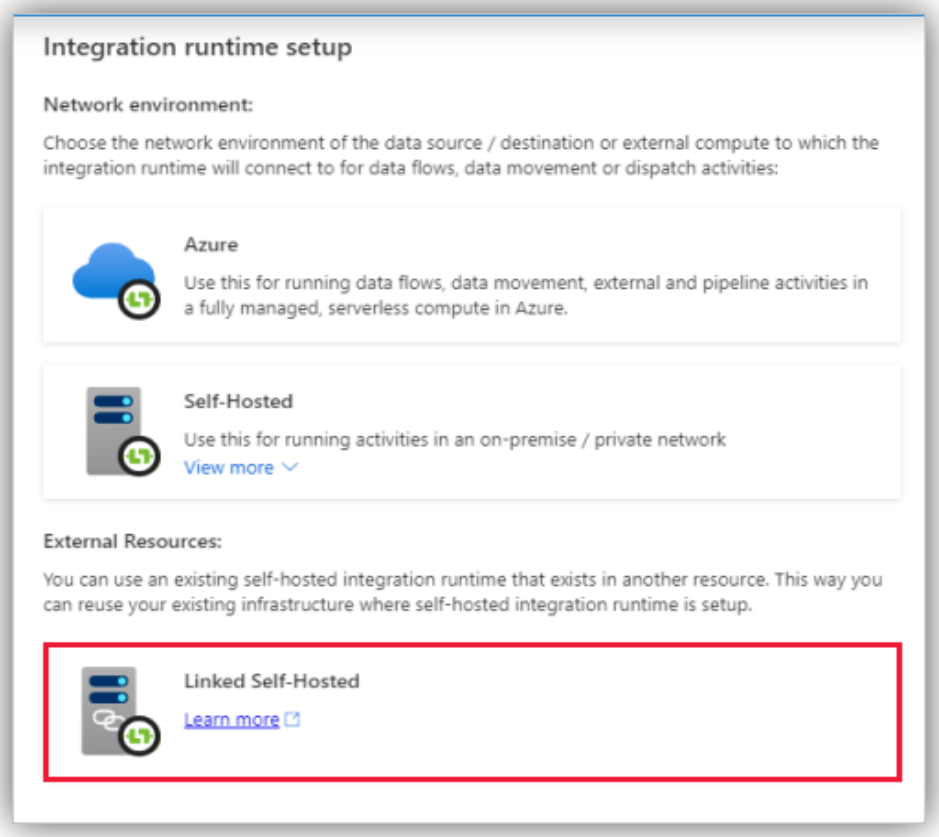
In the self-hosted IR to be shared, select **Grant permission to another Data factory** and in the "Integration runtime setup" page, select the Data factory in which you want to create the linked IR.



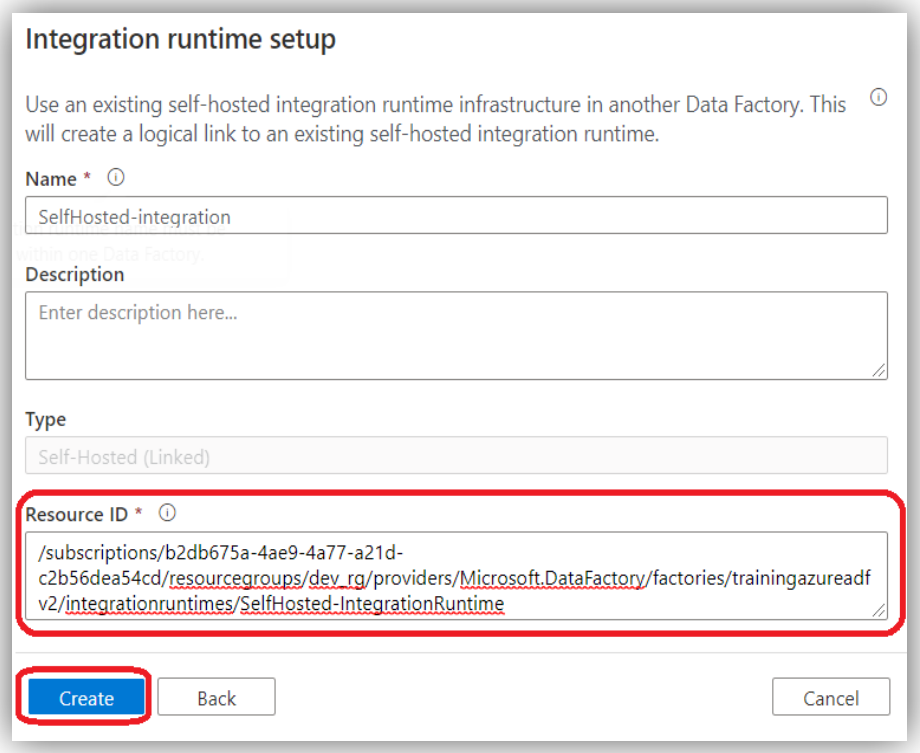


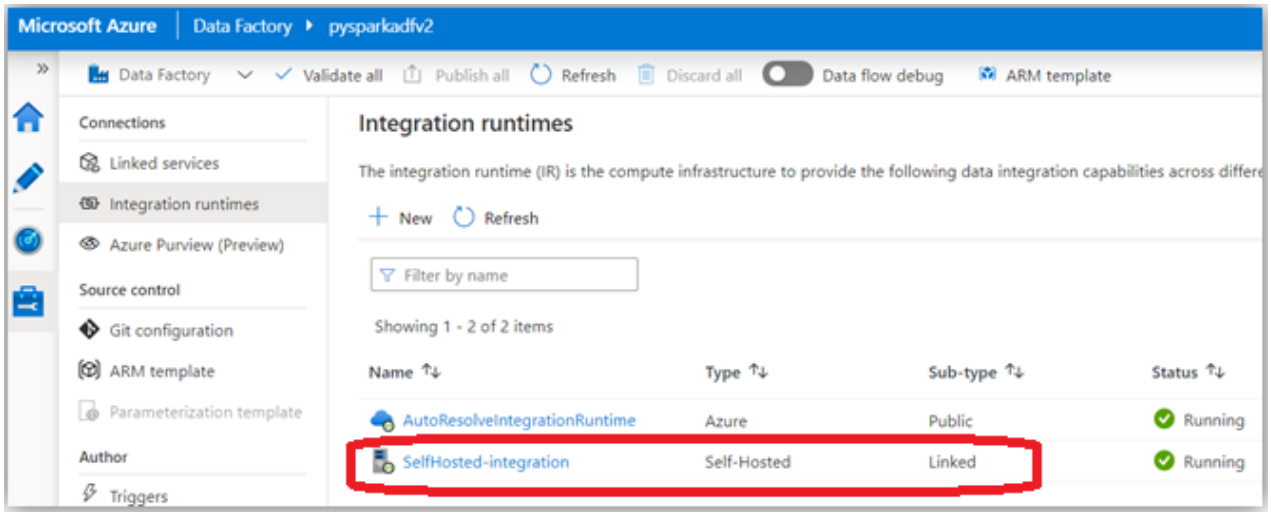
In the data factory to which the permissions were granted, create a new self-hosted IR (linked) and enter the resource ID.





Rename Integration runtime name as " Self-Hosted-Integration " And Paste Resource ID in down window. Then Create. It will be created new Share self-hosted Integration runtime in another Data factory

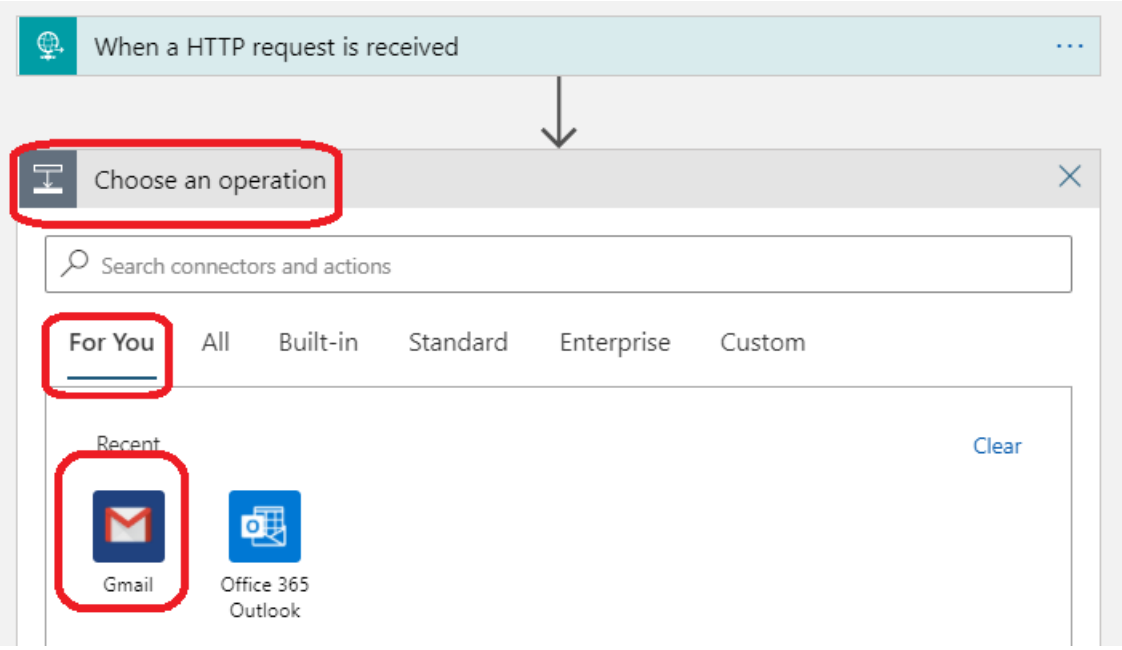


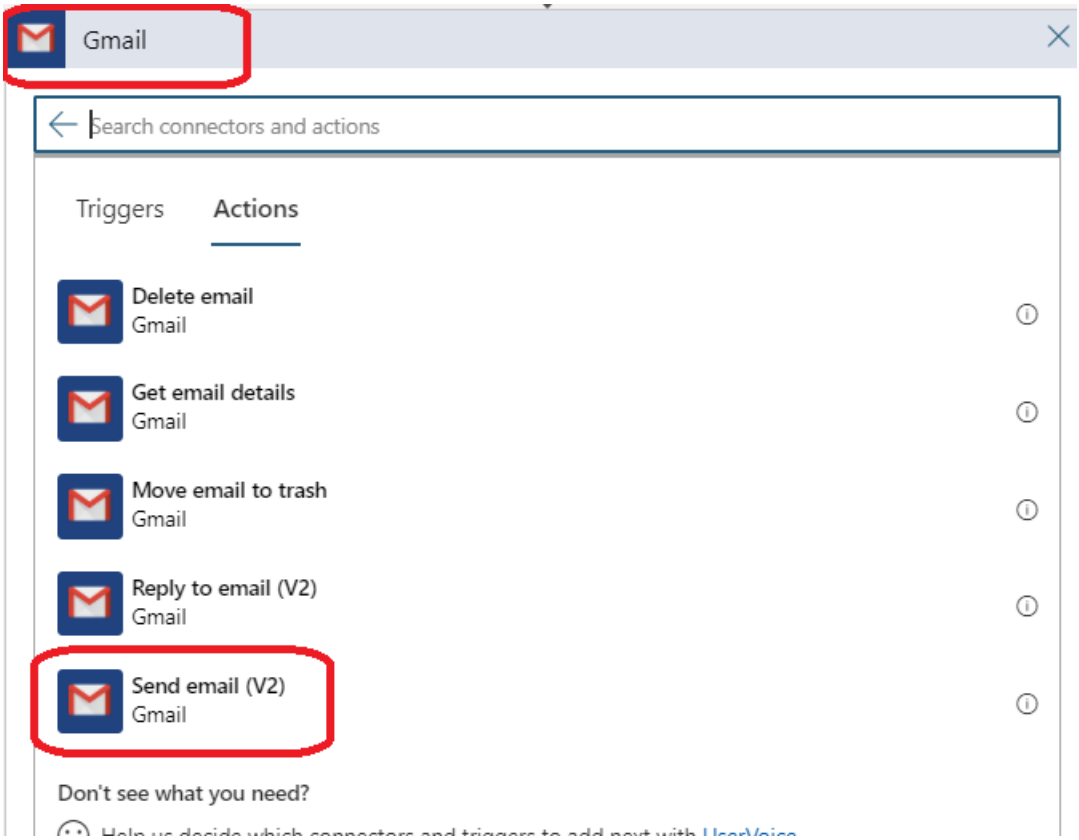


6) What is Logic App? Or How to send an email notification in Azure Data Factory?

Or what is WEB Activity and when we can use this activity?

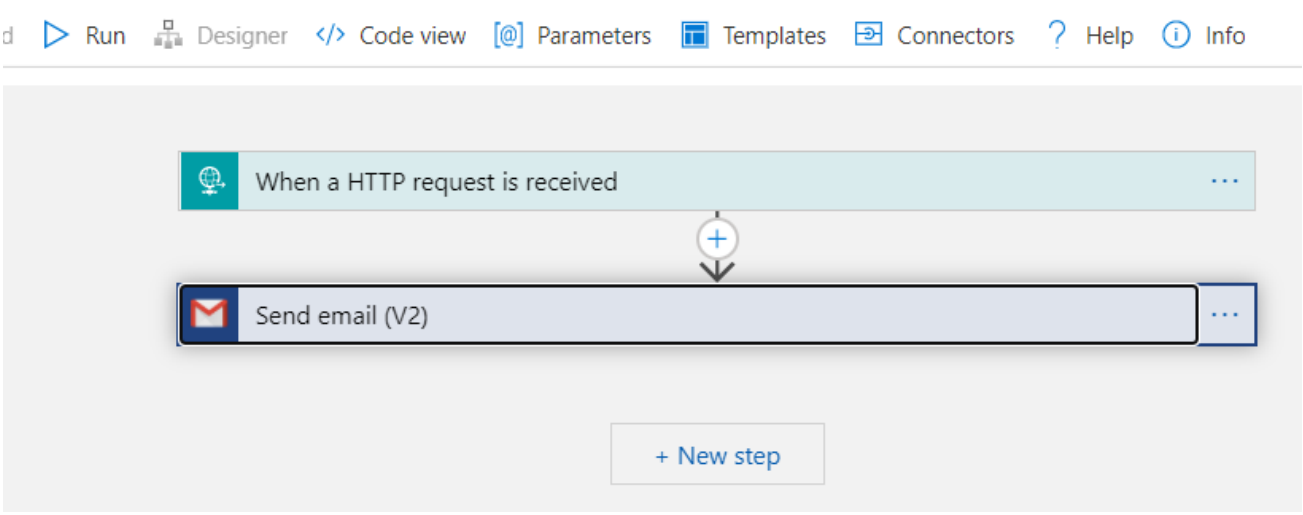
WE can use this for multiple scenarios. Logic App (email) and ADF (error handling). The communication between these two Azure parts is done with a JSON message via an HTTP request (post). The JSON message contains the name of the Data Factory and the pipeline that failed, an error message and an email address. You could of course hardcode the email address in Logic Apps, but now you can reuse the Logic App for various pipelines or data factories and notify different people.



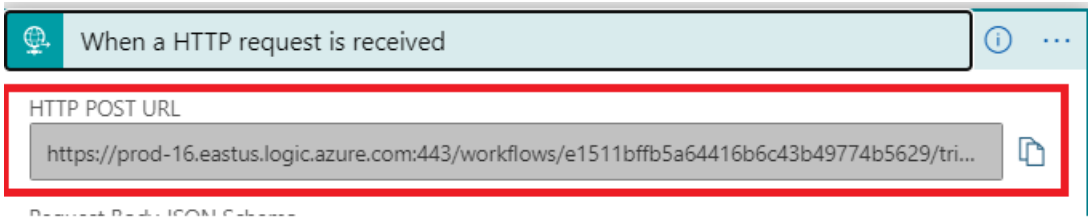


Next, we will add a new step to our Logic App, called “Send an email”. I will use gmail but if you want to use another email provider pick that one.

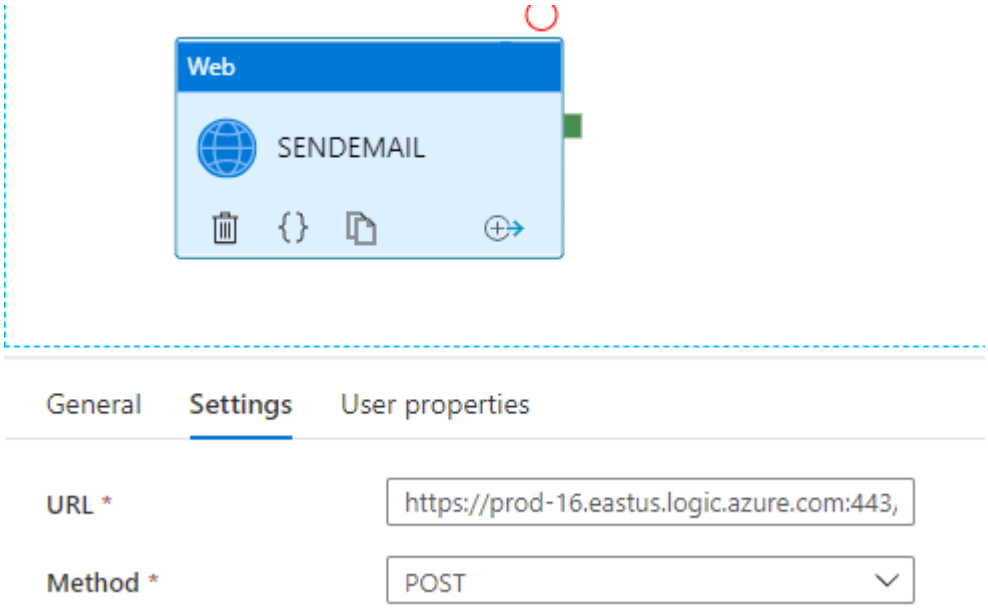
It’s the first time you connect Gmail account on Azure? Then you need to connect your Gmail account to Azure by signing in. (Note: allow pop-ups in your browser.)



After creation of Azure Logic App and saving the Logic App, Azure created an endpoint URL for our Logic Apps, you'll find in the first step. Copy this URL to a notepad, we'll need this later.



Now add an Web activity to the pipeline and rename it.



7) What is Get Metadata Activity? When we can use this?

OR How to get folders and filenames at dynamically?

You can use the Get Metadata activity to retrieve the metadata of any data in Azure Data Factory. You can use this activity in the following scenarios:

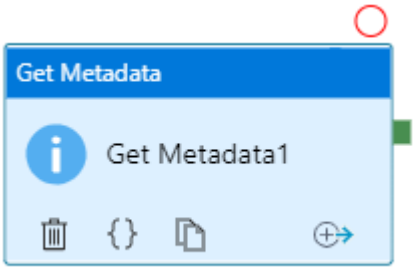
Validate the metadata of any data.

Trigger a pipeline when data is ready/available.

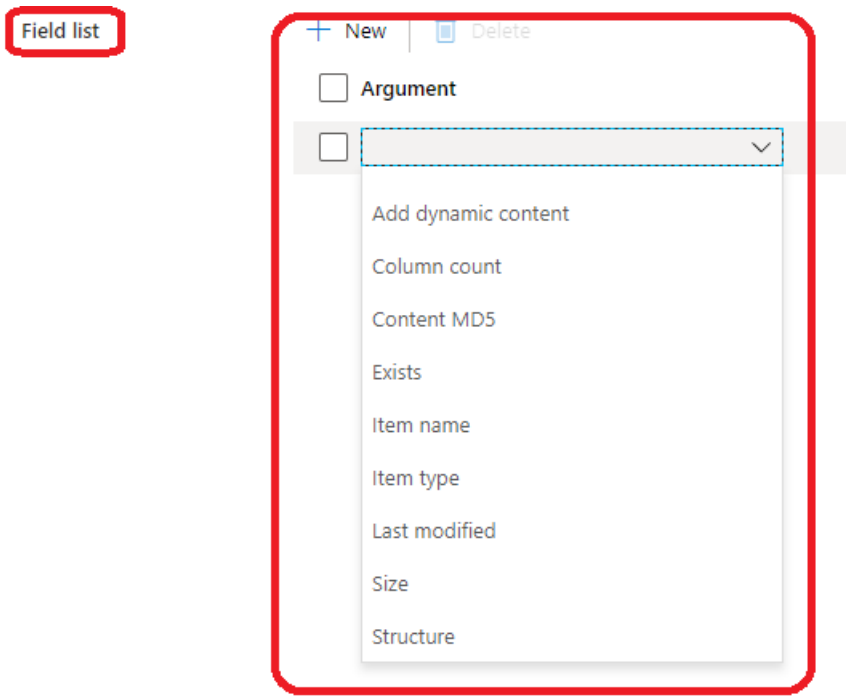
The following functionality is available in the control flow:

You can use the output from the Get Metadata activity in conditional expressions to perform validation.

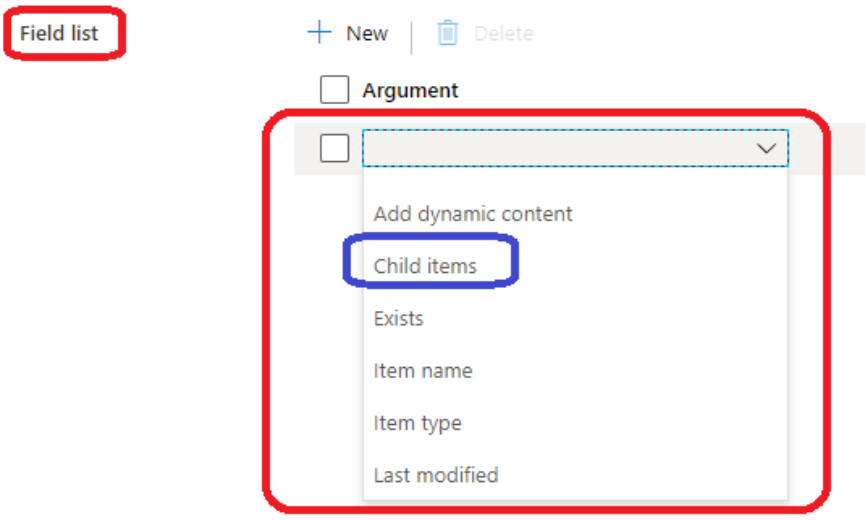
You can trigger a pipeline when a condition is satisfied via Do Until looping.



The Get Metadata activity takes a dataset as an input and returns metadata information as output. Currently, the following connectors and corresponding retrievable metadata are supported. The maximum size of returned metadata is around 4 MB.



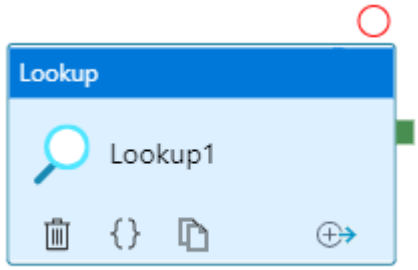
child Items List of subfolders and files in the given folder. Applicable only to folders. Returned value is a list of the name and type of each child item.



8) What is Lookup Activity and when we can use this activity?

Lookup activity can retrieve a dataset from any of the Azure Data Factory-supported data sources. Use it in the following scenario:

Dynamically determine which objects to operate on in a subsequent activity, instead of hard coding the object name. Some object examples are files and tables.



Use the Lookup activity result

The lookup result is returned in the output section of the activity run result.

When firstRowOnly is set to true (default), the output format is as shown in the following code. The lookup result is under a fixed firstRow key. To use the result in subsequent activity, use the pattern of @{activity('LookupActivity').output.firstRow.table}.

9) If you are running more no of pipelines and its taking longer time to execute.

How to resolve this type of issues?

We can go with splitting pipelines batches wise and create multiple integration runtimes. Then those loads will be shared by multiple integration runtimes and we can improve the load performance of more no of pipelines.

10)What is auto resolve integration runtime in azure data factory?

AutoResolveIntegrationRuntime. This is the default integration runtime, and the region is set to auto-resolve. That means that Azure Data Factory decides the physical location of where to execute activities based on the source, sink, or activity type.

11)Data Factory supports three types of triggers. Mention these types briefly

- The **Schedule trigger** that is used to execute the ADF pipeline on a wall-clock schedule
- The **Tumbling window** trigger that is used to execute the ADF pipeline on a periodic interval, and retains the pipeline state
- The **Event-based trigger** that responds to a blob related event, such as adding or deleting a blob from an Azure storage account

Pipelines and triggers have a **many-to-many relationship** (except for the tumbling window trigger).**Multiple triggers can kick off a single pipeline**, or **a single trigger can kick off multiple pipelines**.

12)Any Data Factory pipeline can be executed using three methods. Mention these methods

- Under **Debug mode**
- Manual execution using **Trigger now**
- Using an added **scheduled**, **tumbling window** or **event trigger**

13)How to load data whenever we receive a file in azure data factory? Or

How to run a pipeline if we receive a file or if we delete a file?

Using Event-based trigger we can solve above requirement.

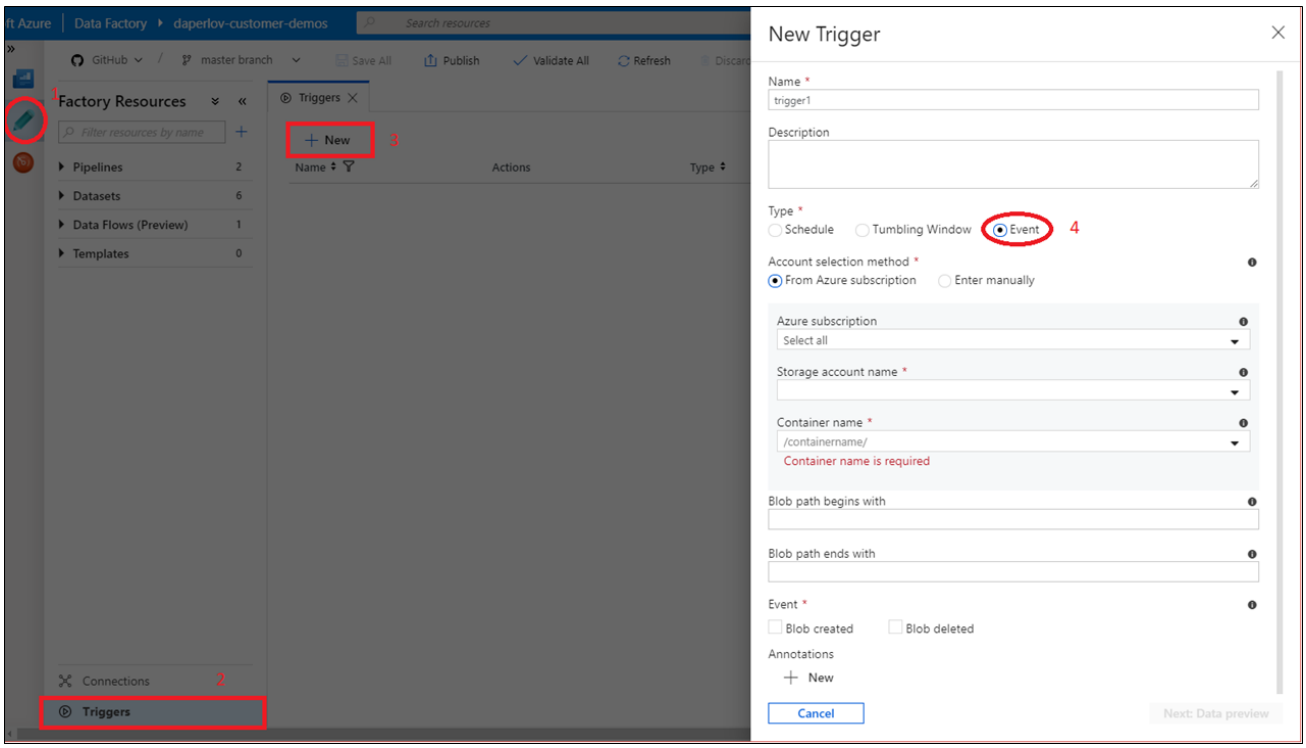
An event-based trigger runs pipelines in response to an event, such as the arrival of a file, or the deletion of a file, in Azure Blob Storage.

Steps to create Event Based Trigger:

In the bottom-left corner, click on the Triggers button

Click + New which will open up the create trigger side nav

Select trigger type Event

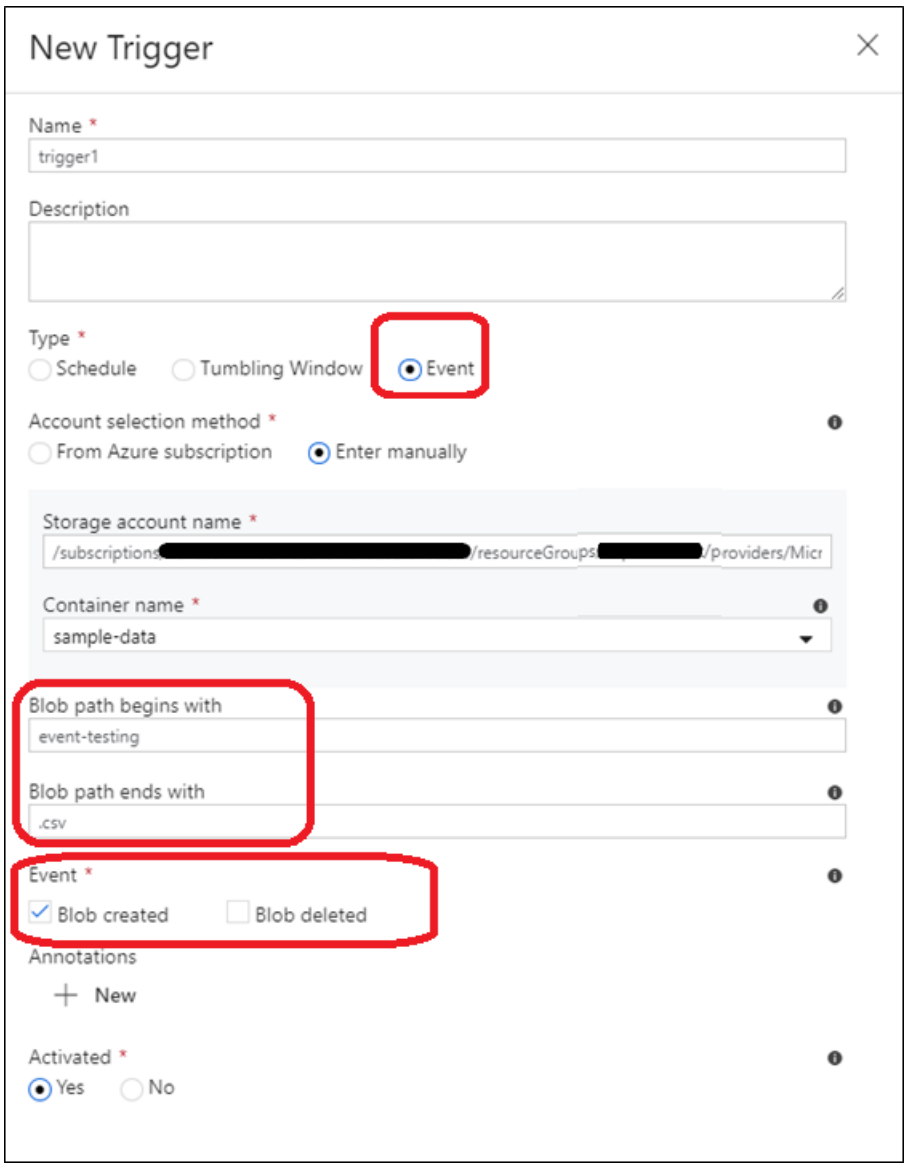


Blob path begins with: The blob path must start with a folder path. Valid values include `2018/` and `2018/april/shoes.csv`. This field can't be selected if a container isn't selected.
Blob path ends with: The blob path must end with a file name or extension. Valid values include `shoes.csv` and `.csv`. Container and folder name are optional but, when specified, they must be separated by a `/`

`blobs/` segment. For example, a container named 'orders' can have a value of `/orders/blobs/2018/april/shoes.csv`. To specify a folder in any container, omit the leading '/' character. For example, `april/shoes.csv` will trigger an event on any file named `shoes.csv` in folder a called 'april' in any container.

Note: Blob path **begins with** and **ends with** are the only pattern matching allowed in Event Trigger. Other types of wildcard matching aren't supported for the trigger type.

Select whether your trigger will respond to a **Blob created** event, **Blob deleted** event, or both. In your specified storage location, each event will trigger the Data Factory pipelines associated with the trigger



14)Difference between Scheduled Trigger and Tumbling window trigger?

The tumbling window trigger and the schedule trigger both operate on time heartbeats. How are they different?

The **tumbling window trigger run** waits for the **triggered pipeline run** to finish. Its run state reflects the state of the triggered pipeline run. For example, if a triggered pipeline run is cancelled, the corresponding tumbling window trigger run is marked cancelled. This is different from the "fire and forget" behavior of the **schedule trigger**, which is marked successful as long as a pipeline run started.

Backfill scenarios

Tumbling Window: Supported. Pipeline runs can be scheduled for windows in the past.

Scheduled Trigger: Not supported. Pipeline runs can be executed only on time periods from the current time and the future.

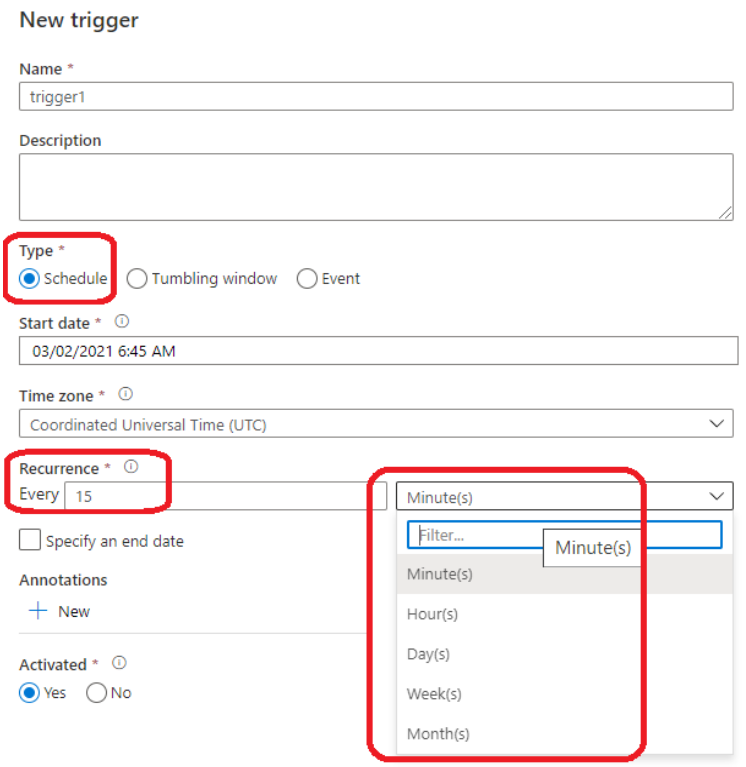
Pipeline-to-trigger relationship

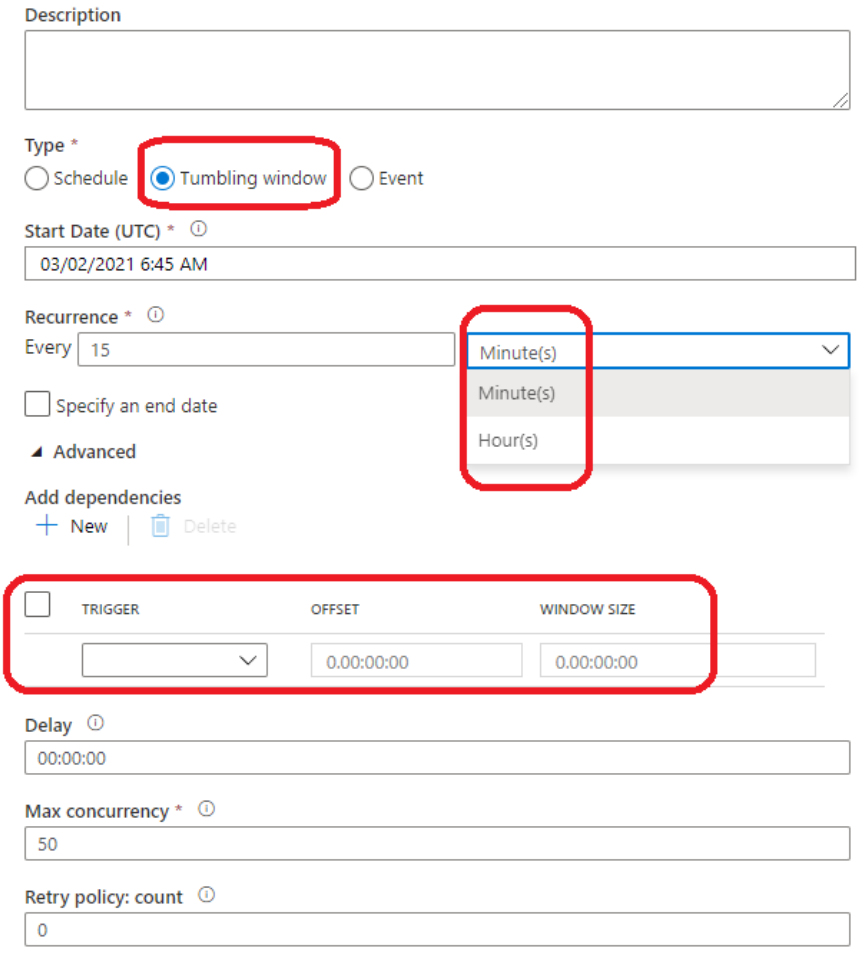
Tumbling Window: Supports a one-to-one relationship. Only one pipeline can be triggered.

Scheduled Trigger: Supports many-to-many relationships. Multiple triggers can kick off a single pipeline. A single trigger can kick off multiple pipelines.

Note:

In order to build a dependency chain and make sure that a trigger is executed only after the successful execution of another trigger in the data factory,


New trigger
Name *
trigger1
Description
Type *
Schedule
Tumbling window
Event
Start date *
03/02/2021 6:45 AM
Time zone *
Coordinated Universal Time (UTC)
Recurrence *
Every
15
Minute(s)
Specify an end date
Annotations
New
Activated *
Yes
No
Filter...
Minute(s)
Hour(s)
Day(s)
Week(s)
Month(s)


Description
Type *
Schedule
Tumbling window
Event
Start Date (UTC) *
03/02/2021 6:45 AM
Recurrence *
Every
15
Minute(s)
Minute(s)
Hour(s)
Specify an end date
Advanced
Add dependencies
New
Delete
TRIGGER
OFFSET
WINDOW SIZE
0.00:00:00
0.00:00:00
Delay
00:00:00
Max concurrency *
50
Retry policy: count
0

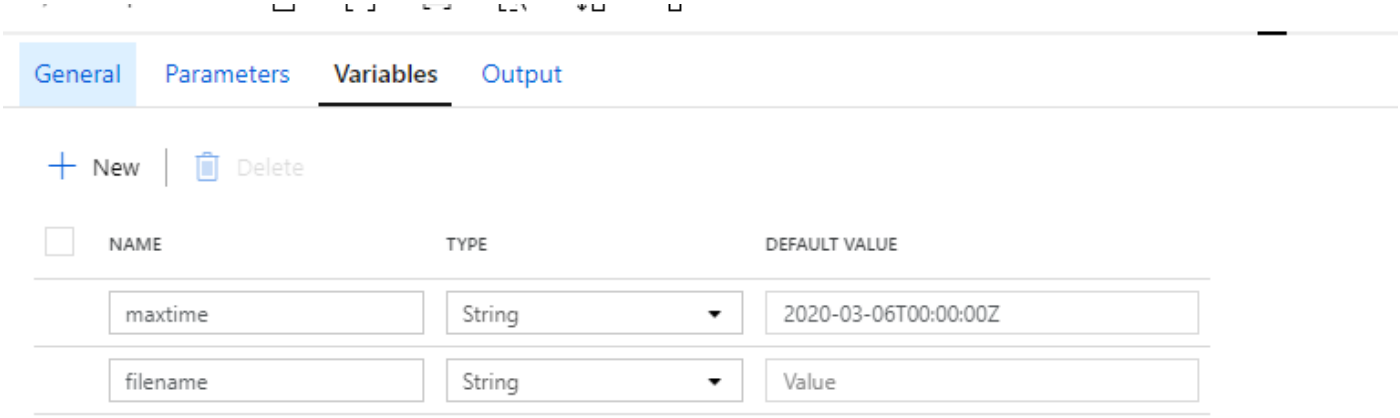
get the latest added file in a folder [Azure Data Factory]

You could set modifiedDatetimeStart and modifiedDatetimeEnd to filter the files in the folder when you use ADLS connector in copy activity.
Maybe it has two situations:

1.The data was pushed by external source **in the schedule**,you are suppose to know the schedule time to configure.
2.The frequency is **random**,then maybe you have to log the pushing data time in another residence,then pass the time as parameter into copy activity pipeline before you execute it.

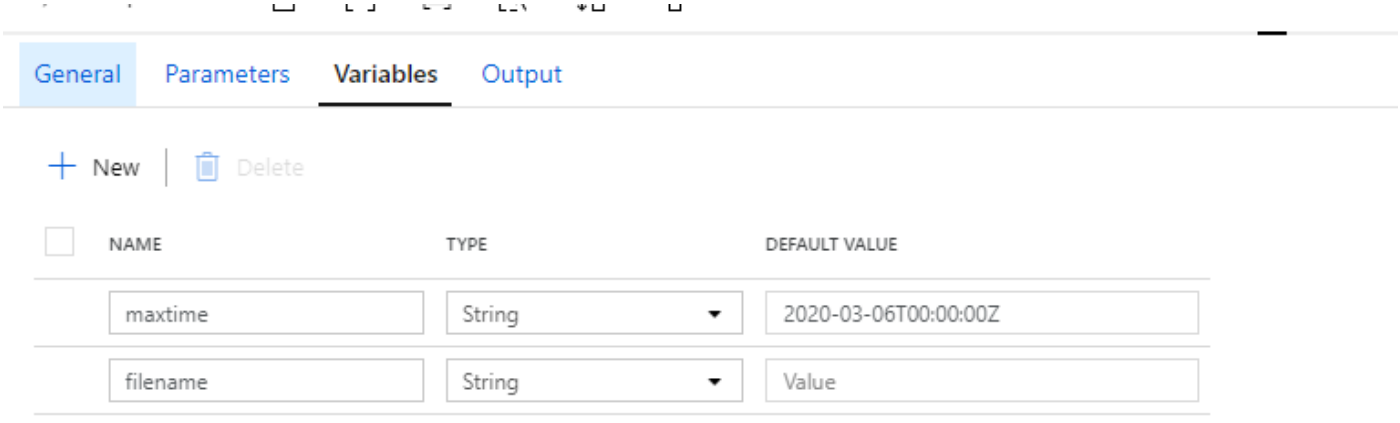
I try to provide a flow for you in ADF pipelines as below:

My sample files in same folder:

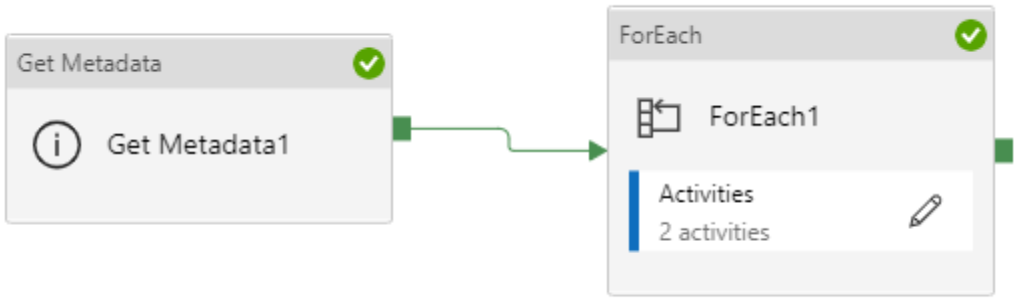


Step1,create two variables, maxtime and filename:

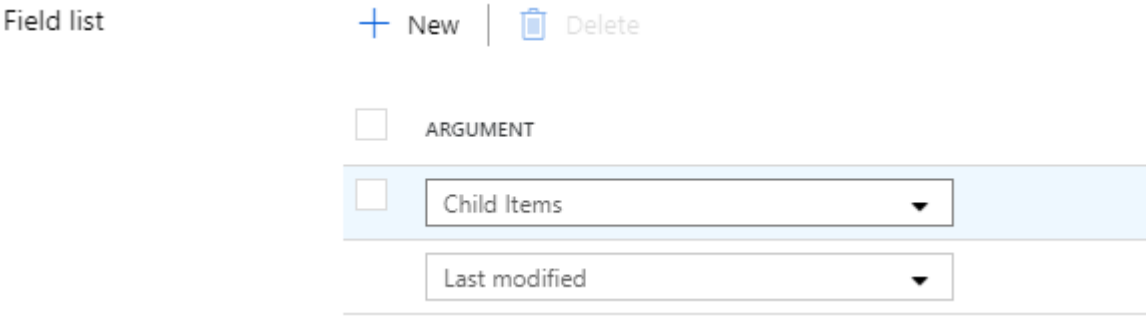
maxtime is the critical datetime of specific date, filename is empty string.



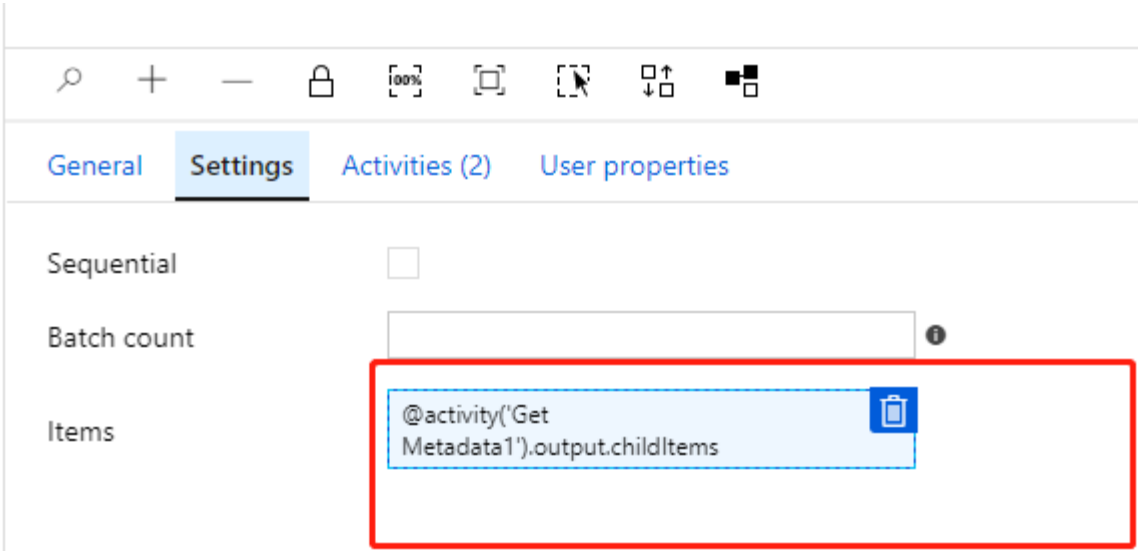
Step2, use GetMetadata Activity and ForEach Activity to get the files under folder.



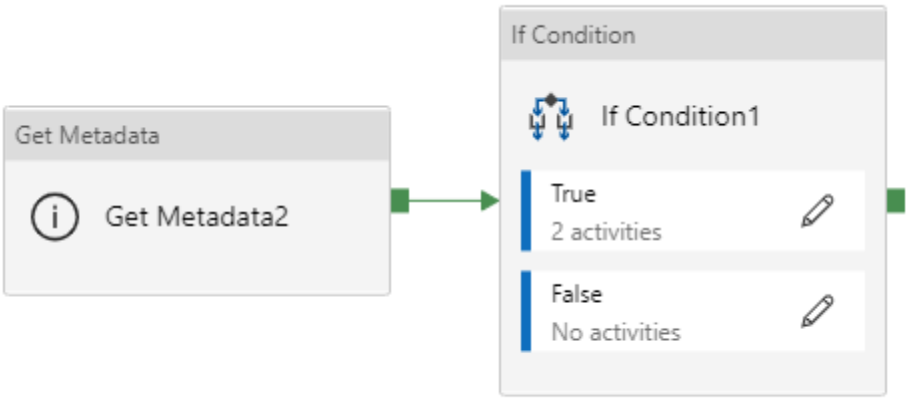
GetMetadata 1 configuration:



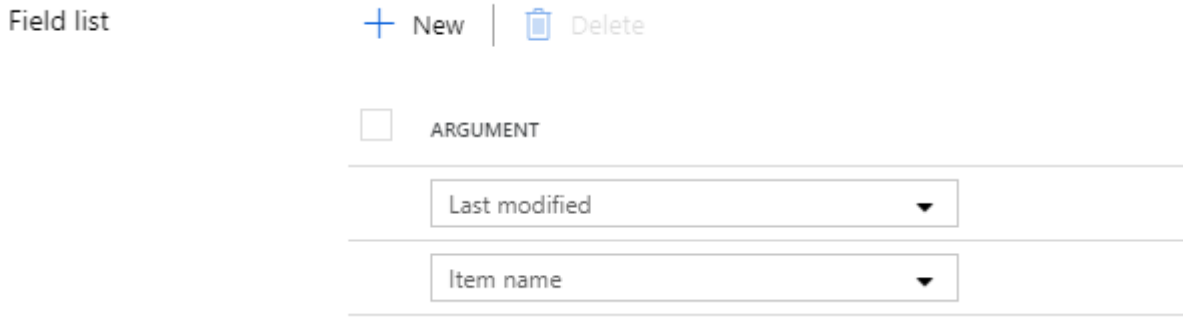
ForEach Activity configuration:



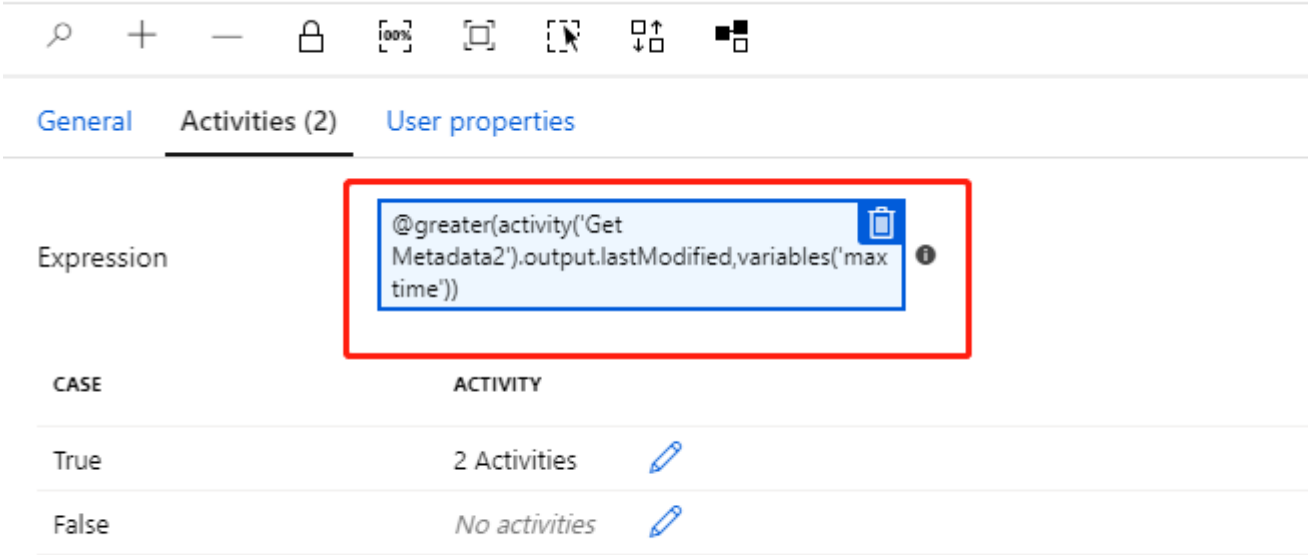
Step3: Inside ForEach Activity,use GetMetadata and If-Condition, the structure as below:



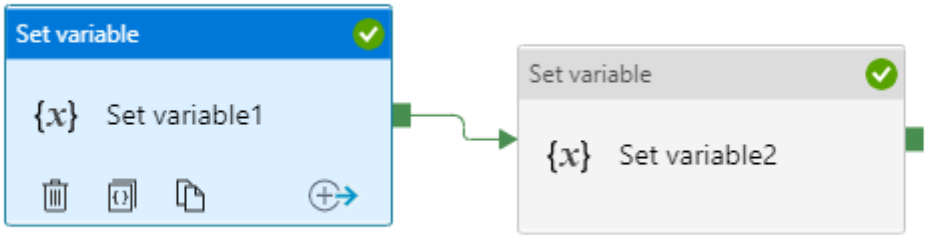
GetMetadata 2 configuration:



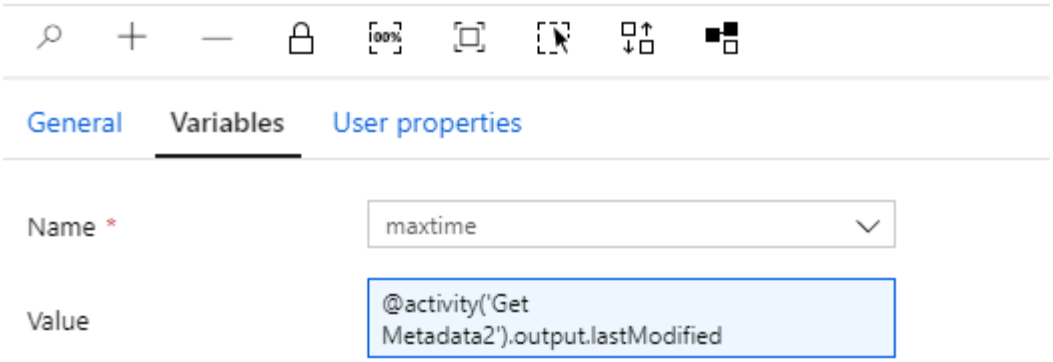
If-Condition Activity configuration:



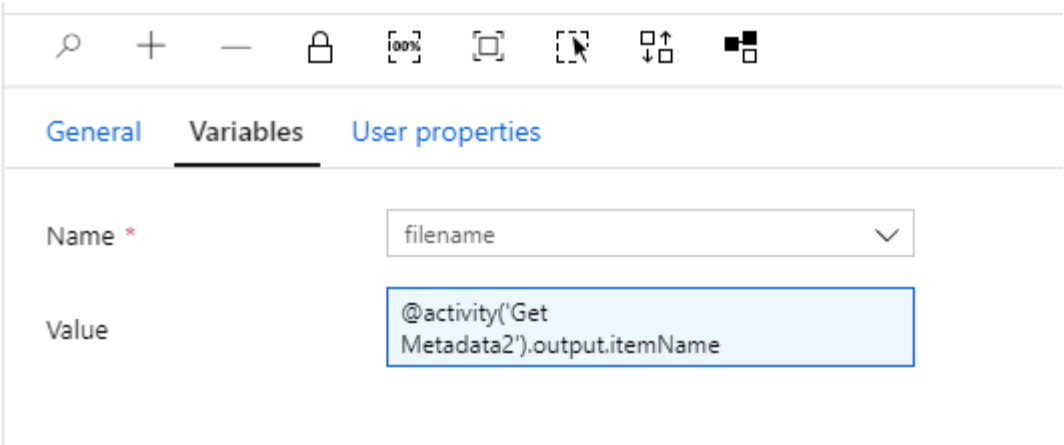
Step4: Inside If-Condition True branch,use Set Variable Activity:



Set variable1 configuration:

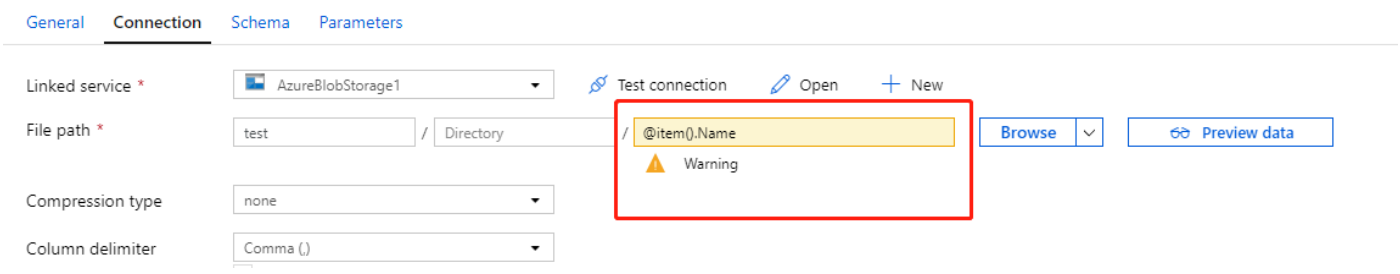


Set variable2 configuration:



All of above steps aim to finding the latest fileName, the variable fileName is exactly target.

Addition for another new dataset in GetMetadata 2



How To trigger pipeline from one ADF to Another ADF.

Using WEB Activity we can achieve this calling one ADF pipeline in another ADF.

Use below API to trigger in Web Activity.

API To Trigger Pipeline.

https://management.azure.com/subscriptions/{{sub_id}}/resourceGroups/{{resource_group_name}}/providers/Microsoft.DataFactory/factories/{{factory_name}}/pipelines/{{pipeline_name}}/createRun?api-version=2018-06-01

Caller ADF: batch25demoadfv2

Callee ADF: adfv2batch25

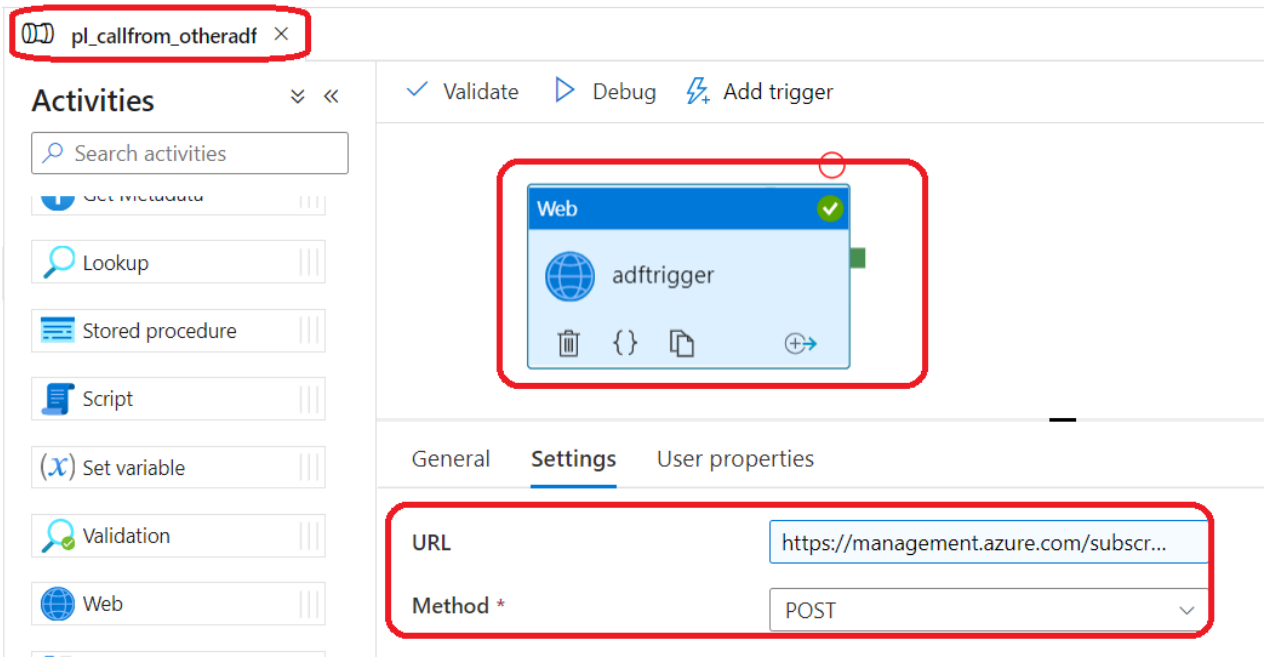
Create new pipeline in caller adf as “pl_callfrom_otheradf

And create one web activity.

URL : modified url based on above API.

https://management.azure.com/subscriptions/e0bf1971-b822-46ca-8e4a-72555f0a959e/resourceGroups/rg-dev-data-engineering-batch25/providers/Microsoft.DataFactory/factories/adfv2batch25/pipelines/pl_migrate_adlsgen1_to_adlsgen2/createRun?api-version=2018-06-01

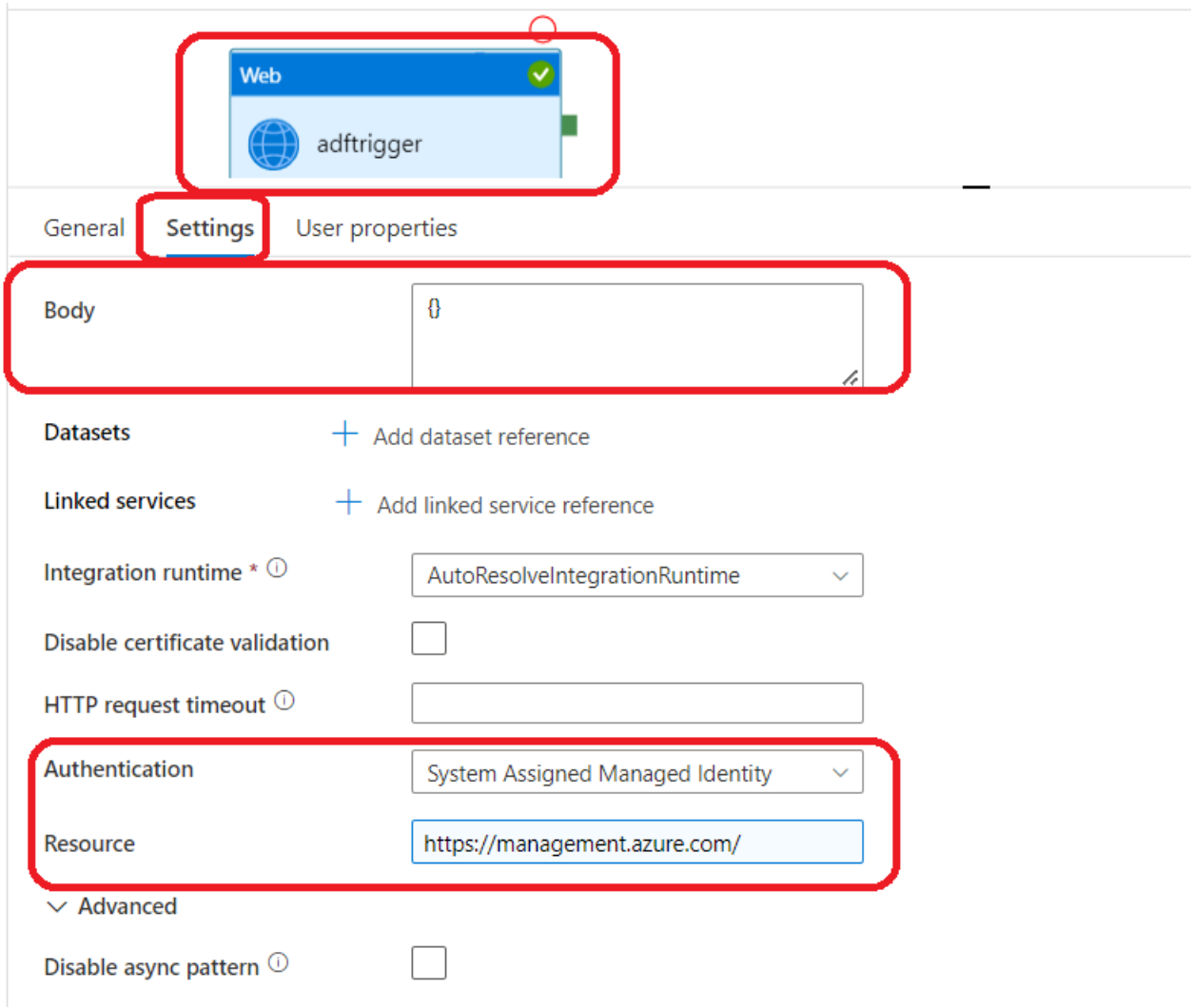
Method: POST



Body : {} (empty {})

Authentication : System Assigned Managed Identity

Resource: https://management.azure.com

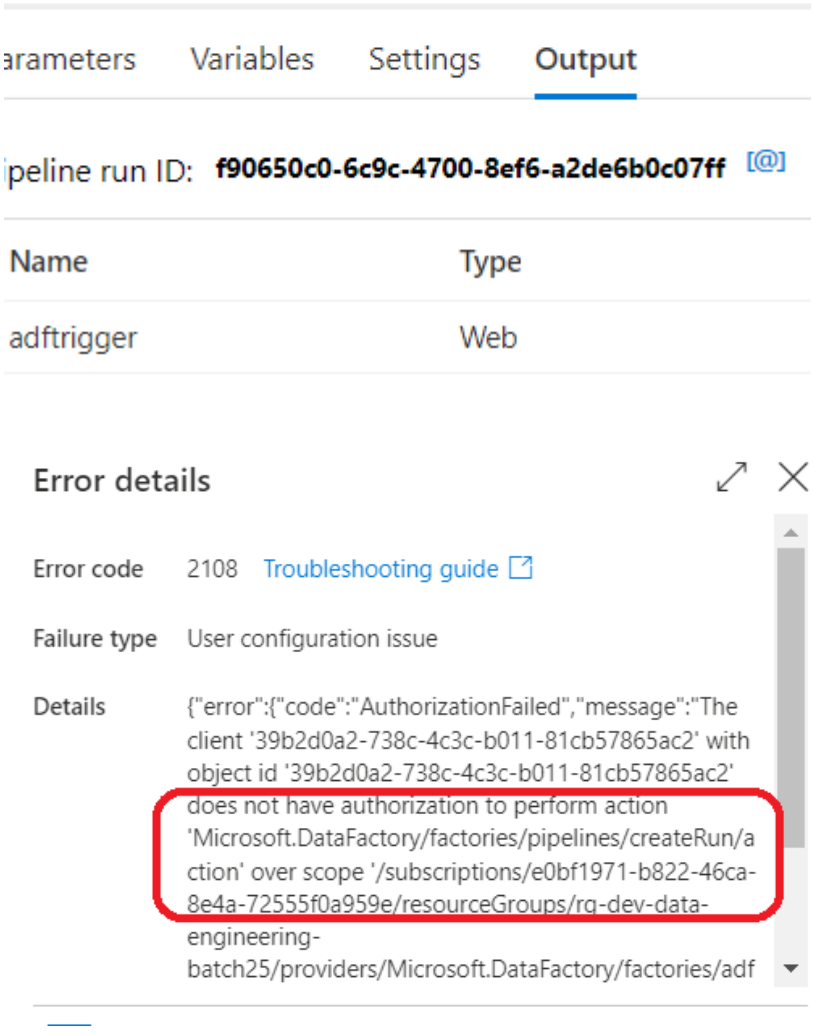


Save and Debug. Now it will raise error due to privileges issue.

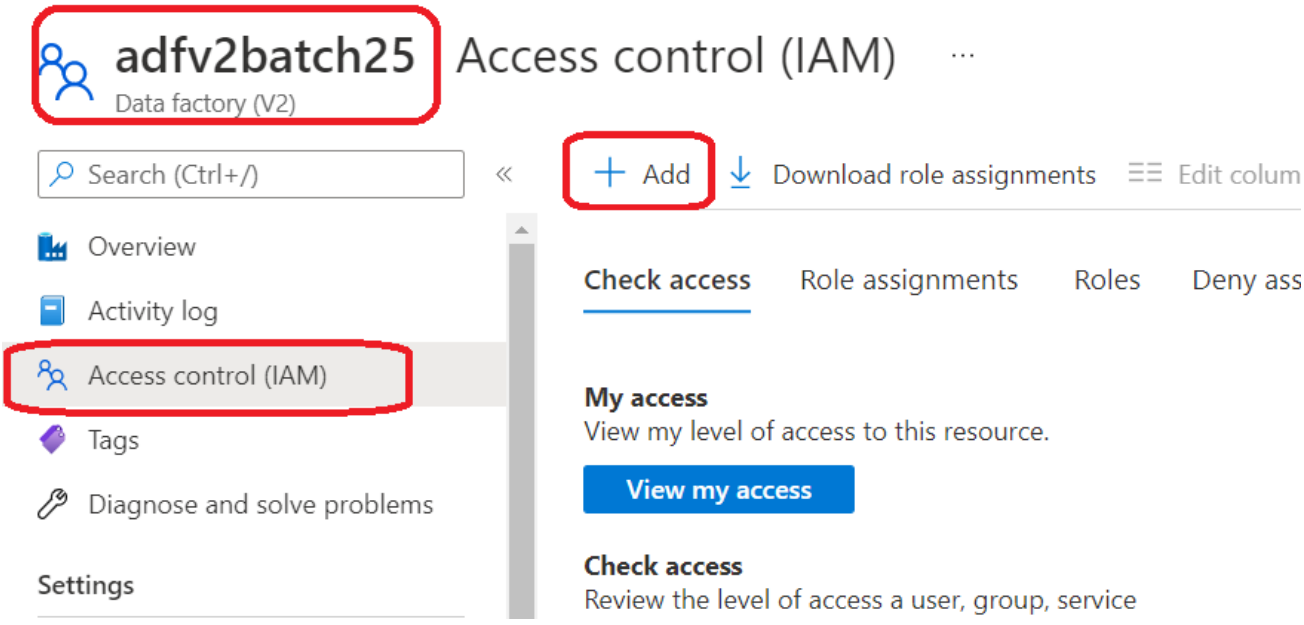
We need to grant access as Contributor role.

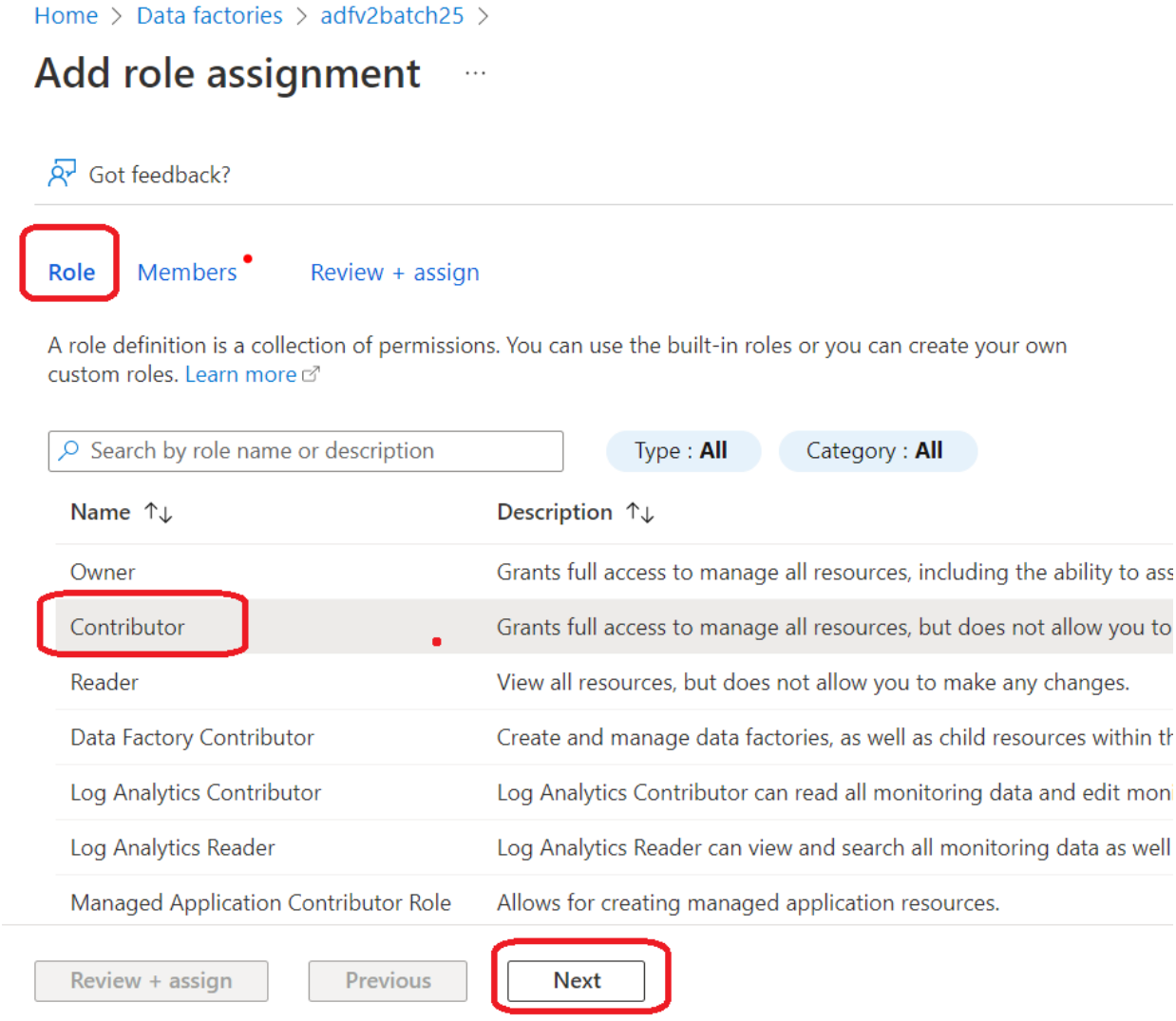
Caller ADF: batch25demoadfv2

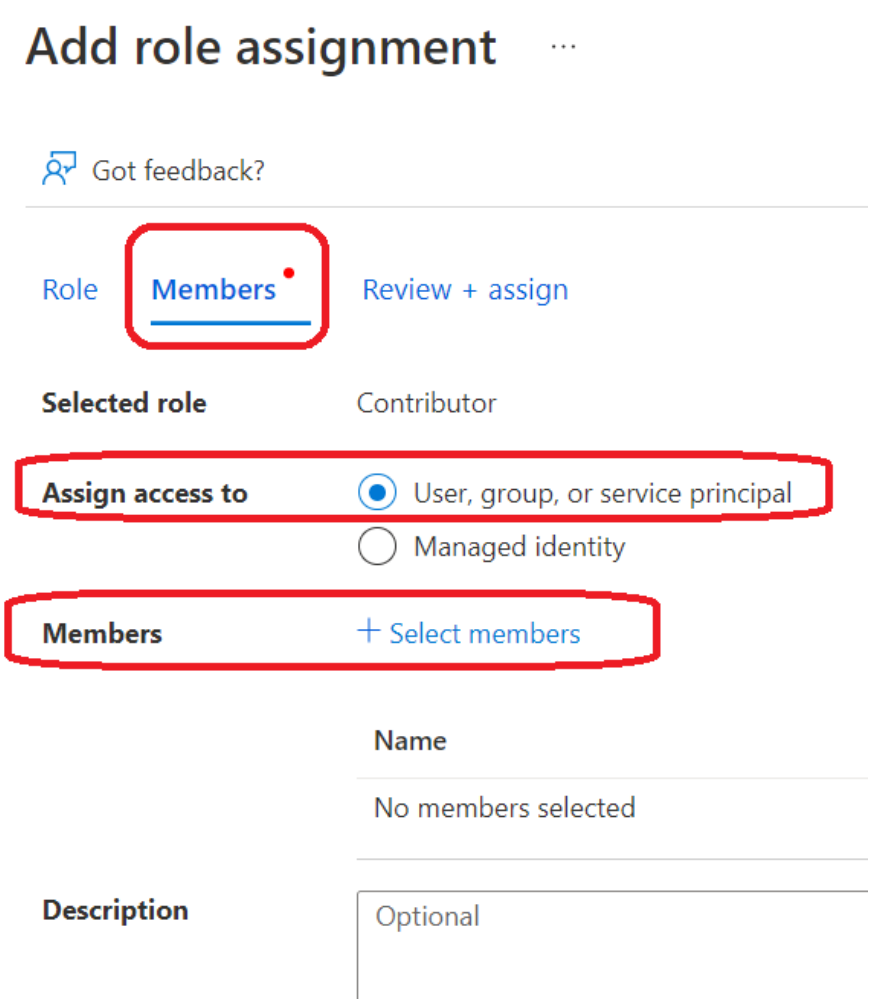
Callee ADF: adfv2batch25 (give contributor role to batch25demoadfv2)

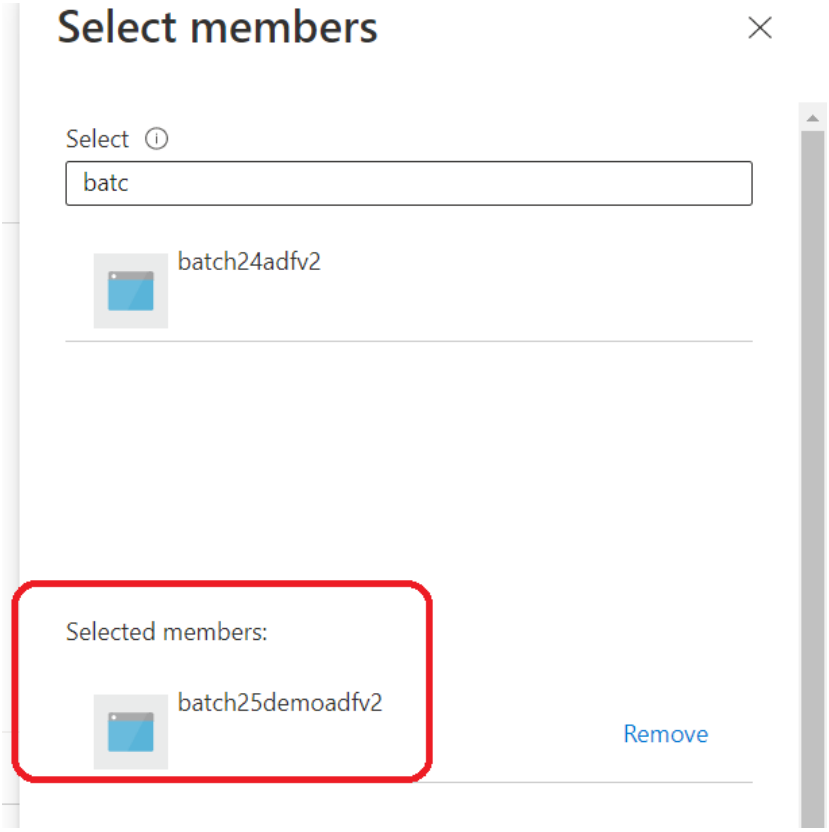


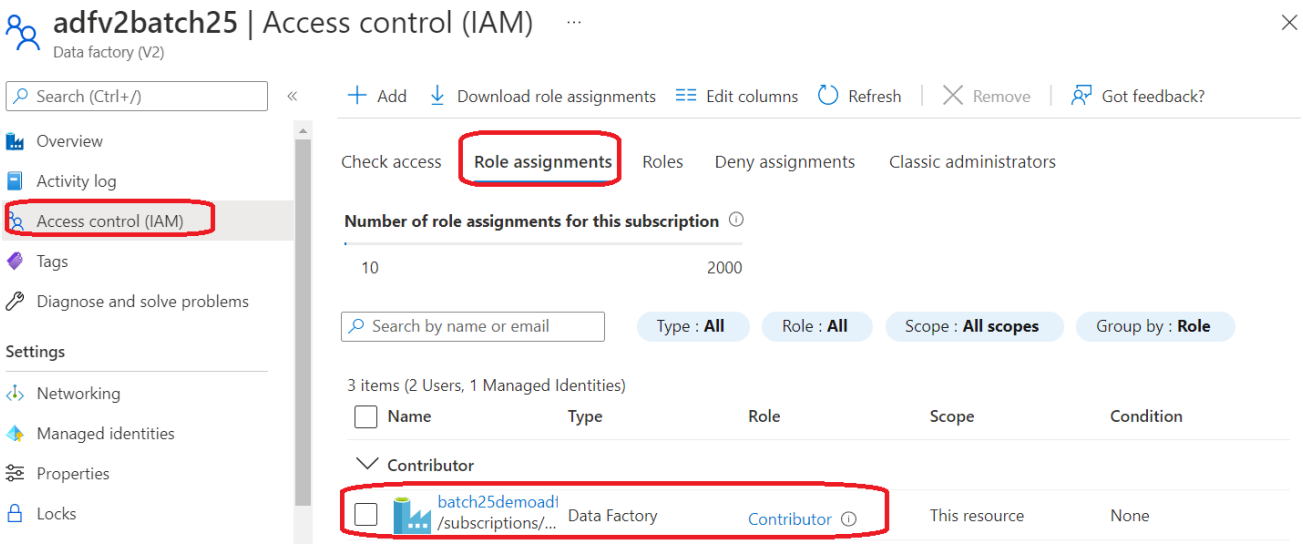
Grant Contributor role to batch25demoadfv2 from adfv2batch25 adf.


adfv2batch25
Data factory (V2)
Access control (IAM)
Search (Ctrl+/)
Add
Download role assignments
Edit colum
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Settings
Check access
Role assignments
Roles
Deny ass
My access
View my level of access to this resource.
View my access
Check access
Review the level of access a user, group, service

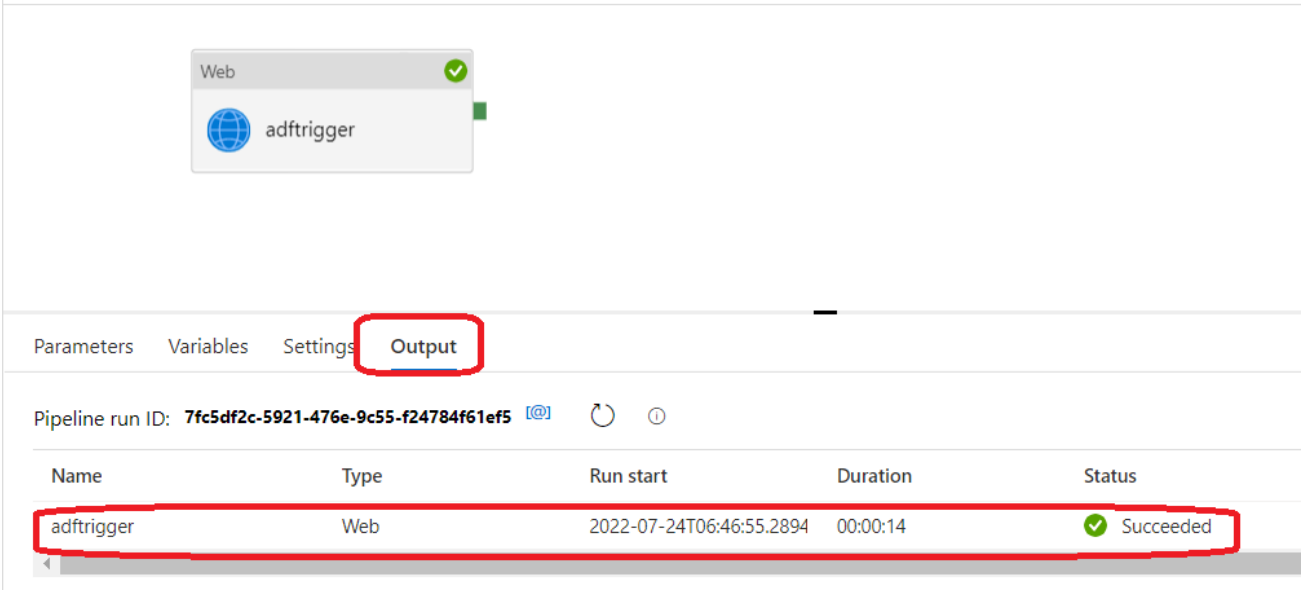

Home > Data factories > adfv2batch25 >
Add role assignment
Got feedback?
Role
Members
Review + assign
A role definition is a collection of permissions. You can use the built-in roles or you can create your own custom roles. Learn more
Search by role name or description
Type : All
Category : All
Name
Description
Owner
Grants full access to manage all resources, including the ability to ass
Contributor
Grants full access to manage all resources, but does not allow you to
Reader
View all resources, but does not allow you to make any changes.
Data Factory Contributor
Create and manage data factories, as well as child resources within th
Log Analytics Contributor
Log Analytics Contributor can read all monitoring data and edit mon
Log Analytics Reader
Log Analytics Reader can view and search all monitoring data as well
Managed Application Contributor Role
Allows for creating managed application resources.
Review + assign
Previous
Next

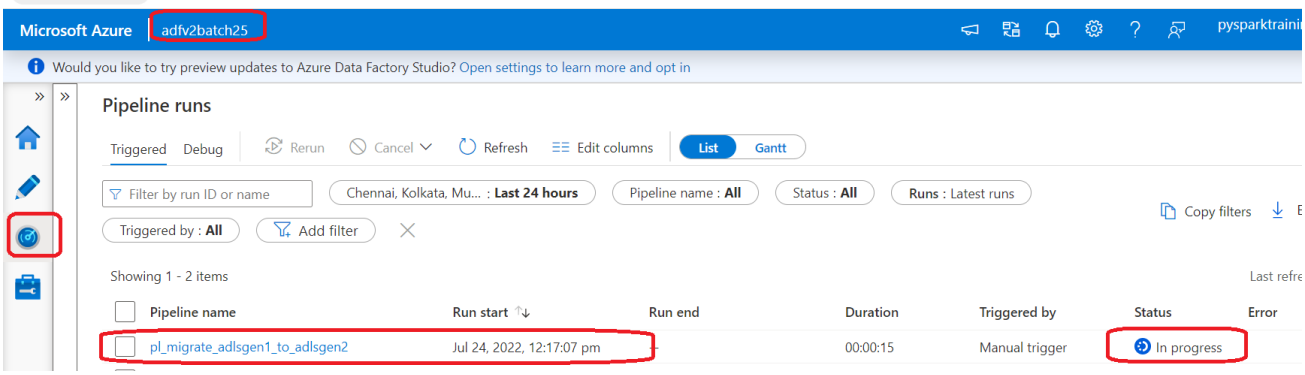

Add role assignment
Got feedback?
Role
Members
Review + assign
Selected role
Contributor
Assign access to
User, group, or service principal
Managed identity
Members
+ Select members
Name
No members selected
Description
Optional


Select members
Select
batc
batch24adfv2
Selected members:
batch25demoadfv2
Remove



Then Run Pipeline in batch25demoadfv2. We can verify in **adfv2batch25 monitor.**





How to automate Resume or Pause Dedicated SQL Pool in azure data factory?

Using below Two api's we can Pause/Resume SQL DWH.

Pause compute

https://management.azure.com/subscriptions/{subscription-id}/resourceGroups/{resource-group-name}/providers/Microsoft.Sql/servers/{server-name}/databases/{database-name}/pause?api-version=2014-04-01-preview

Resume compute

https://management.azure.com/subscriptions/{subscription-id}/resourceGroups/{resource-group-name}/providers/Microsoft.Sql/servers/{server-name}/databases/{database-name}/resume?api-version=2014-04-01-preview

Grant SQL DB Contributor role to Azure Data Factory to access SQL DB.

Add a Web activity to your pipeline and give it a suitable name

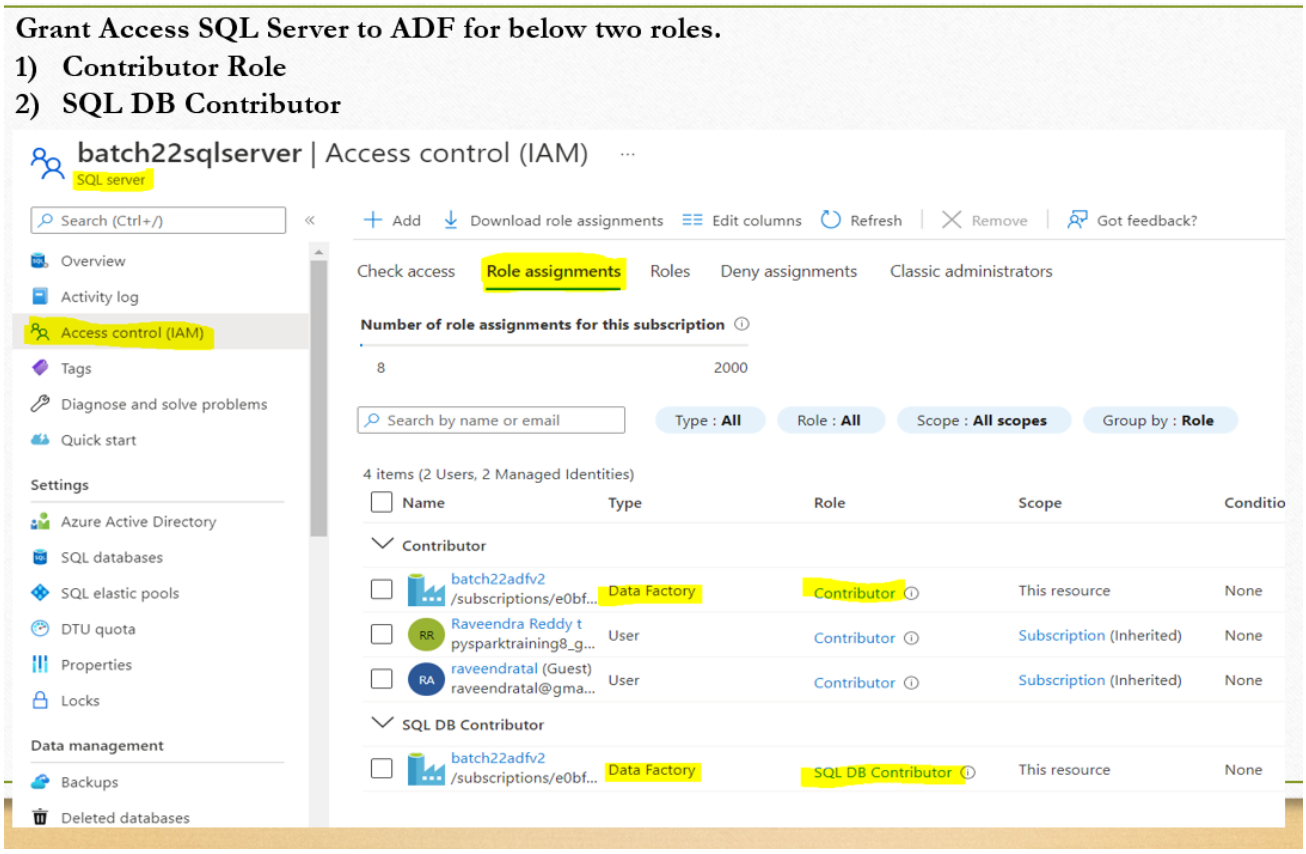
Go to the Settings tab and use the URL from the previous step in the URL property

Choose POST as method

Fill in {} as body (we don't need it, but it is required)

Authentication : Managed Identity

Resource : https://management.azure.com/



Use Web Activity to first verify status of Dedicated SQL pool.

Based on status we can use if condition if it is paused we can use resume API

If it is online we can use paused API.

Verify Status:

https://management.azure.com/subscriptions/e0bf1971-46ca-8e4a-72555f0a959e/resourceGroups/rg-dev-batch22-de/providers/Microsoft.Sql/servers/batch22sqlserver/databases/synapsedwh/?api-version=2014-04-01-preview

Resume:

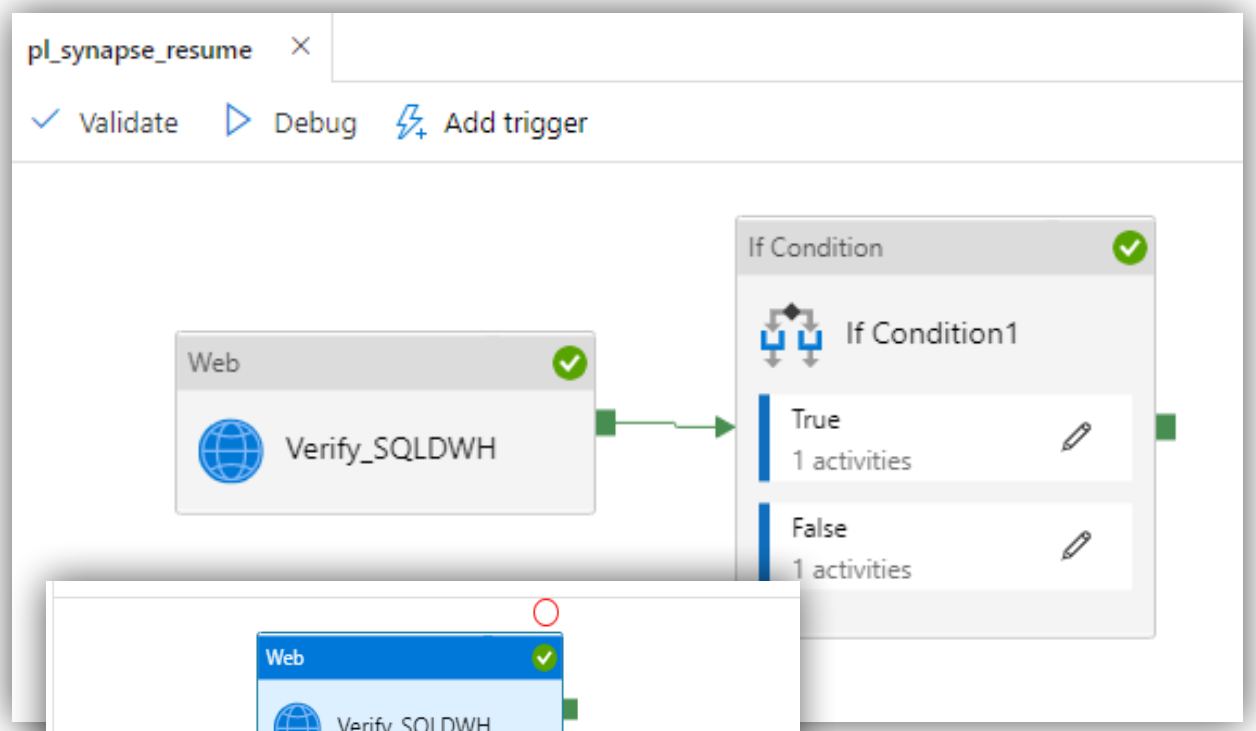
https://management.azure.com/subscriptions/e0bf1971-46ca-8e4a-72555f0a959e/resourceGroups/rg-dev-batch22-de/providers/Microsoft.Sql/servers/batch22sqlserver/databases/synapsedwh/resume?api-version=2014-04-01-preview

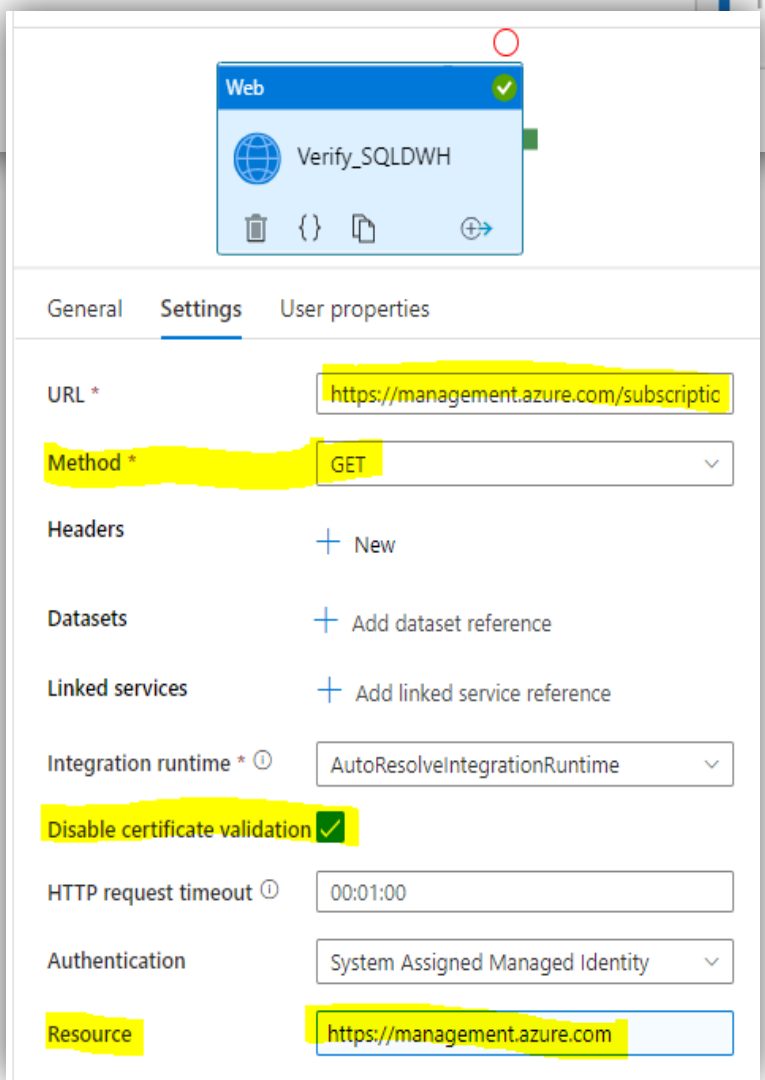
Azure Data Factory Advanced Interview Questions And Answers **Youtube:** [TechLake]

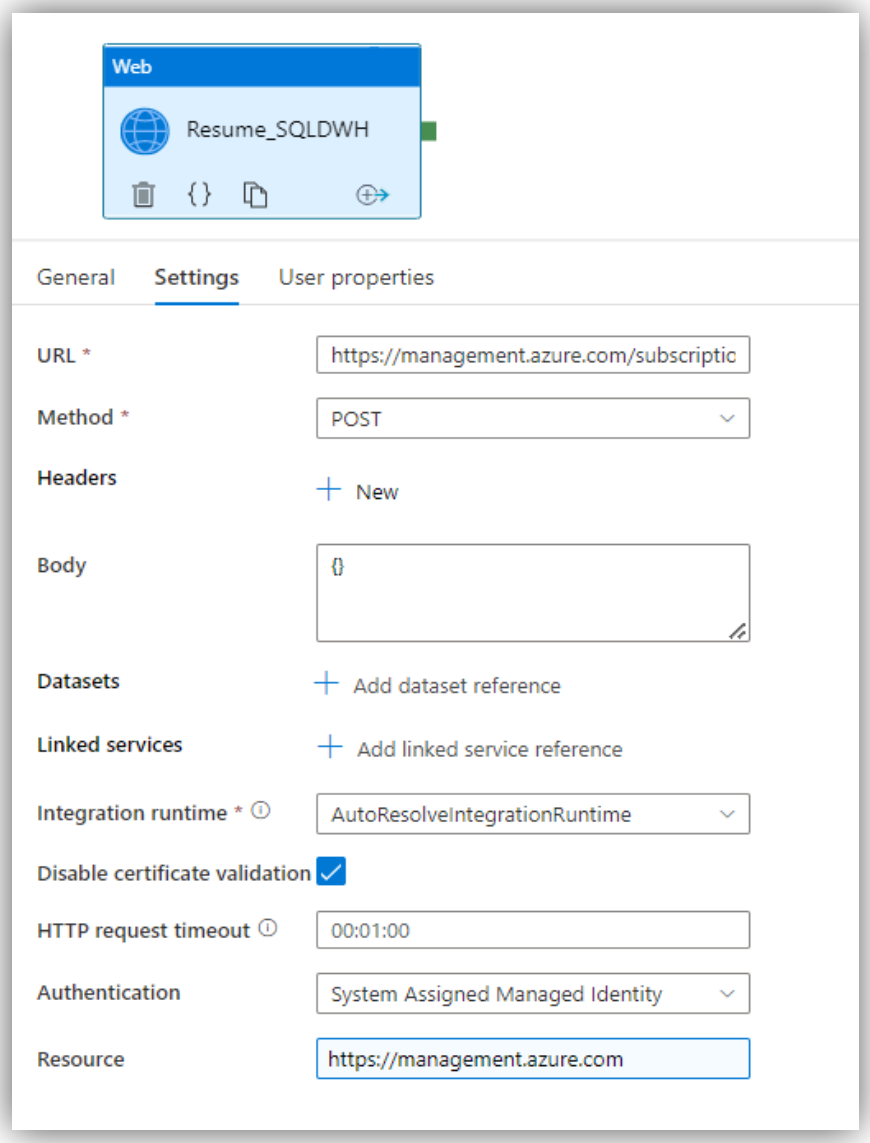
Pause:

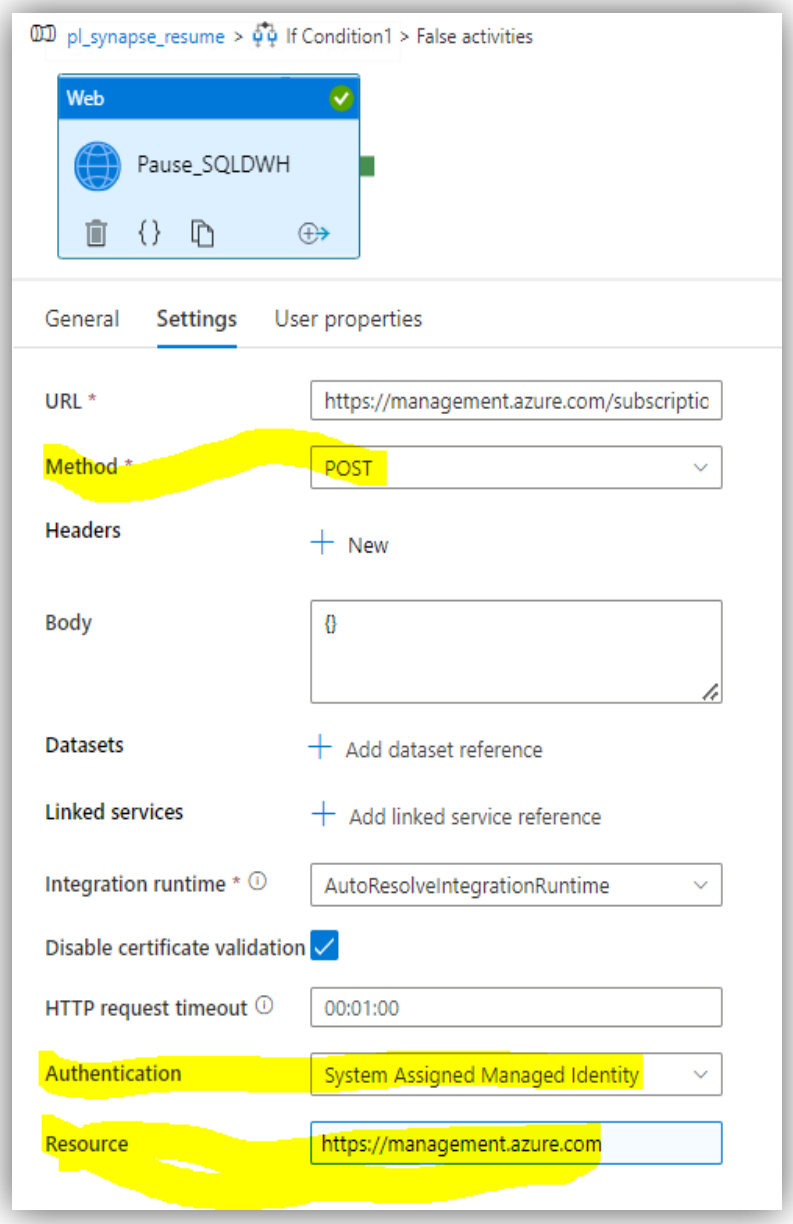
https://management.azure.com/subscriptions/e0bf1971-46ca-8e4a-72555f0a959e/resourceGroups/rg-dev-batch22-de/providers/Microsoft.Sql/servers/batch22sqlserver/databases/synapsedwh/pause?api-version=2014-04-01-preview

first activity we will get status using web activity. Based on status using if condition we call resume or pause we activities.






Web
Resume_SQLDWH
General
Settings
User properties
URL *
https://management.azure.com/subscriptic
Method *
POST
Headers
New
Body
{}
Datasets
Add dataset reference
Linked services
Add linked service reference
Integration runtime *
AutoResolveIntegrationRuntime
Disable certificate validation
HTTP request timeout
00:01:00
Authentication
System Assigned Managed Identity
Resource
https://management.azure.com


pl_synapse_resume > If Condition1 > False activities
Web
Pause_SQLDWH
General
Settings
User properties
URL *
https://management.azure.com/subscriptic
Method *
POST
Headers
New
Body
{}
Datasets
Add dataset reference
Linked services
Add linked service reference
Integration runtime *
AutoResolveIntegrationRuntime
Disable certificate validation
HTTP request timeout
00:01:00
Authentication
System Assigned Managed Identity
Resource
https://management.azure.com