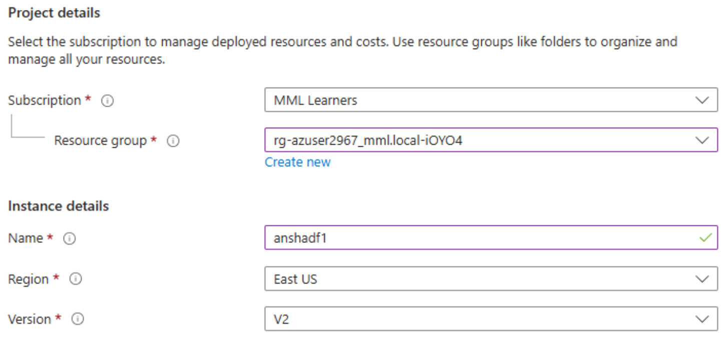
Ansh Ranjan
Azure Data Factory
All Exercises

Exercise 1: Setting Up Azure Databricks & Spark Basics

TASK 1: Create a new Azure Databricks workspace.

1. Go to Data Factories > Create > Enter details



Project details

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

Subscription * : MML Learners

Resource group * : rg-azuser2967_mml.local-iOYO4

Create new

Instance details

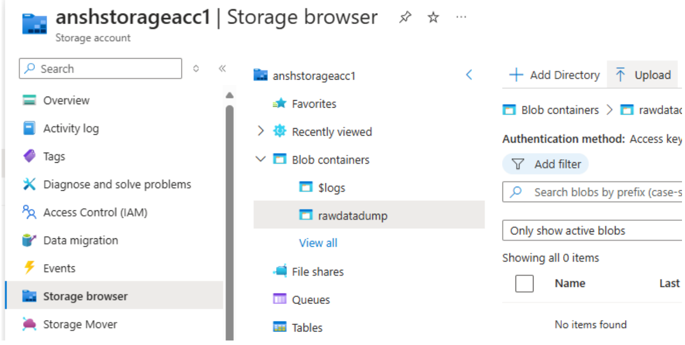
Name * : anshadf1

Region * : East US

Version * : V2

TASK 2 & 3: Download the dataset and Upload to Blob Storage

1. Go to your storage account > Storage Browser side tab > Blob Containers > select a blob container to upload your dataset > click on Upload



anshstorageacc1 | Storage browser

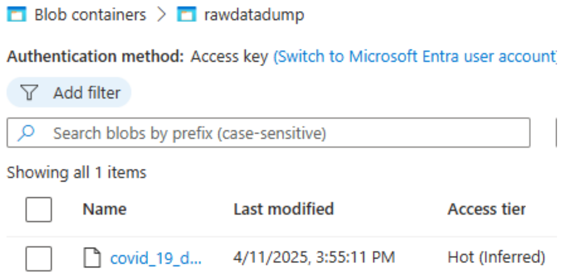
Storage account

Authentication method: Access key

Showing all 0 items

No items found

2. Browse for your dataset csv file and hit Upload



Blob containers > rawdatadump

Authentication method: Access key (Switch to Microsoft Entra user account

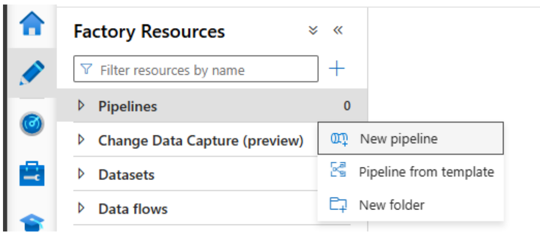
Showing all 1 items

□	Name	Last modified	Access tier
□	covid_19_d...	4/11/2025, 3:55:11 PM	Hot (Inferred)

Exercise 2: Building Pipelines for Covid-19 Data Ingestion

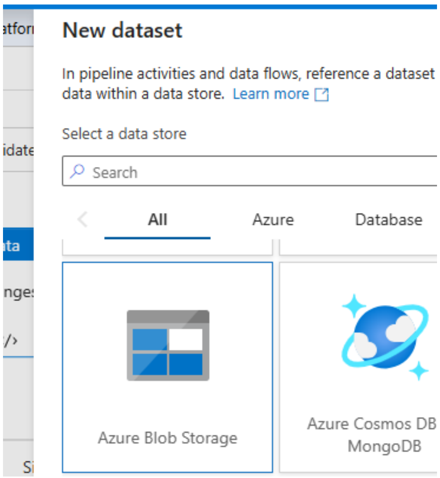
TASK 1: Create an ADF pipeline to copy data

1. Open your Data Factory and go to Author tab > Pipeline



2. Drag a Copy Data activity onto stage

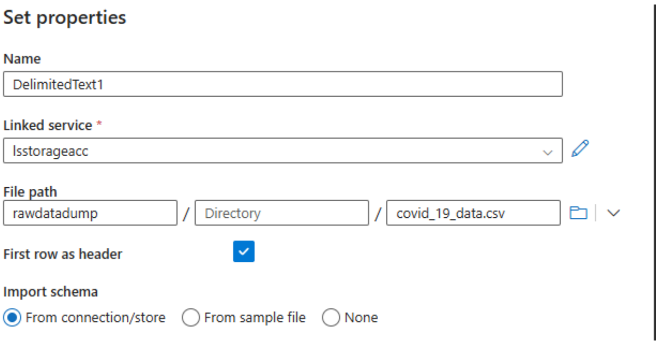
3. Go to source > New Dataset > Azure Blob Storage > Select CSV format


New dataset

In pipeline activities and data flows, reference a dataset data within a data store. Learn more

Select a data store

4. Create a new linked service to your storage account. And select your path of csv file


Set properties

Name: DelimitedText1

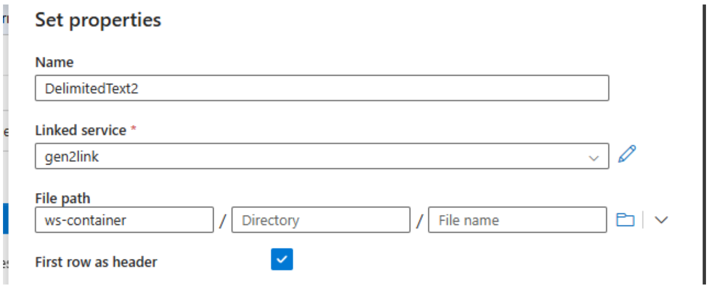
Linked service *: lsstorageacc

File path: rawdatadump / Directory / covid_19_data.csv

First row as header ☑

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

6. In Sink tab, select your destination. Click on New > Azure Gen 2 Data lake


Set properties

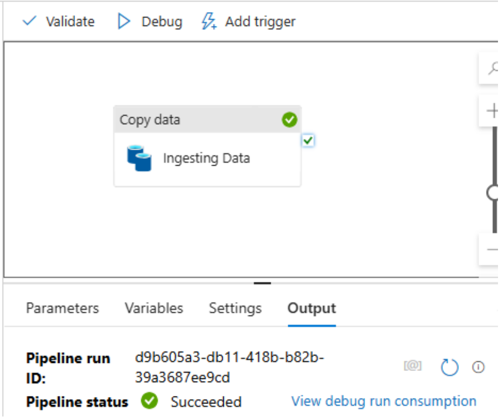
Name: DelimitedText2

Linked service *: gen2link

File path: ws-container / Directory / File name

First row as header ☑

7. Click on Debug button to test your pipeline

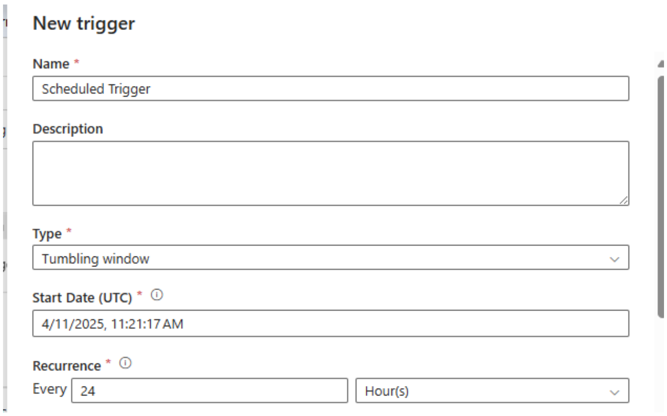


Pipeline run ID: d9b605a3-db11-418b-b82b-39a3687ee9cd

Pipeline status ✓ Succeeded View debug run consumption

TASK 2: Configure pipeline triggers for scheduled runs.

1. Click on Trigger > New/Edit > New > select type as Tumbling Runs > define start dates and reccurance time (going with 24 hrs)



New trigger

Name *
Scheduled Trigger

Description

Type *
Tumbling window

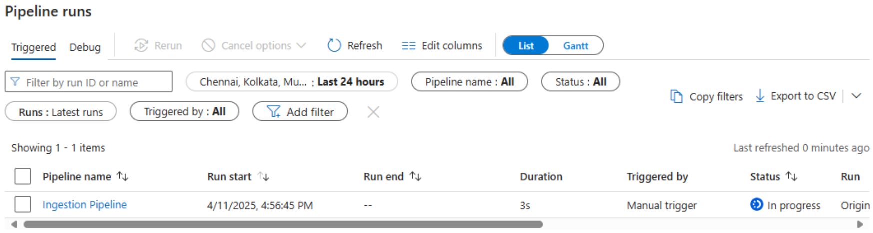
Start Date (UTC) *
4/11/2025, 11:21:17 AM

Recurrence *
Every 24 Hour(s)

2. Publish your pipelines and Triggers.

TASK 3: Monitor and troubleshoot the pipeline execution

1. Trigger your pipeline once for execution. You can view the execution in Monitor tab

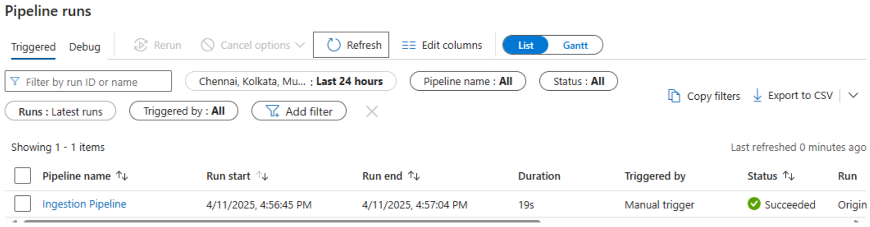


Showing 1 - 1 items

Last refreshed 0 minutes ago

	Pipeline name ↑↓	Run start ↑↓	Run end ↑↓	Duration	Triggered by	Status ↑↓	Run
□	Ingestion Pipeline	4/11/2025, 4:56:45 PM	--	3s	Manual trigger	In progress	Origin

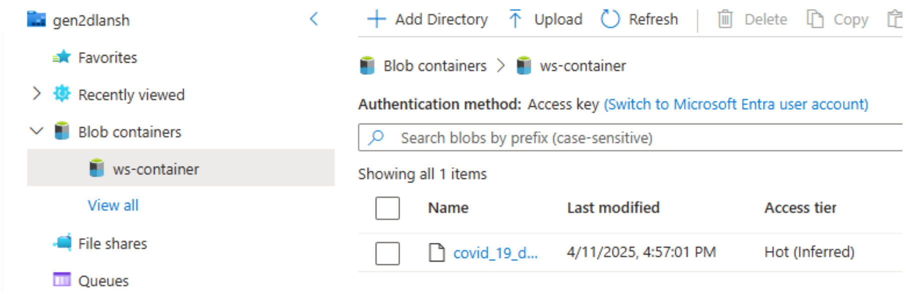
2. The pipeline has successfully executed



Showing 1 - 1 items

Last refreshed 0 minutes ago

	Pipeline name ↑↓	Run start ↑↓	Run end ↑↓	Duration	Triggered by	Status ↑↓	Run
□	Ingestion Pipeline	4/11/2025, 4:56:45 PM	4/11/2025, 4:57:04 PM	19s	Manual trigger	✓ Succeeded	Origin



Authentication method: Access key (Switch to Microsoft Entra user account)

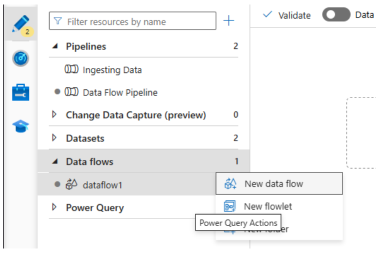
Showing all 1 items

□	Name	Last modified	Access tier
□	covid_19_d...	4/11/2025, 4:57:01 PM	Hot (Inferred)

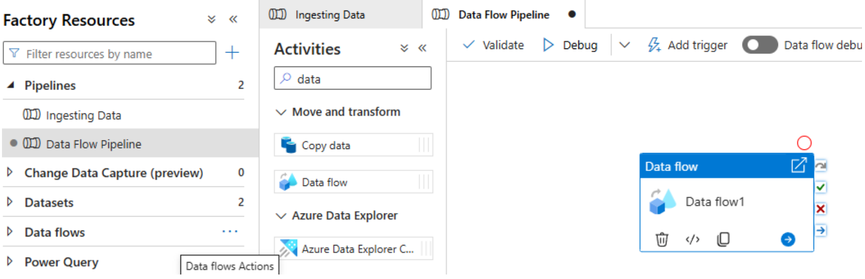
Exercise 3: Data Transformation & Loading for Covid-19 Analysis

TASK 1: Use ADF's Mapping Data Flow

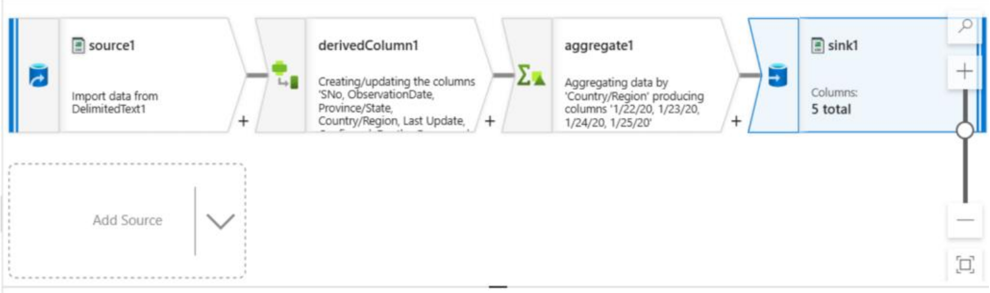
1. Go to Author tab > Data Flow > New Data Flow



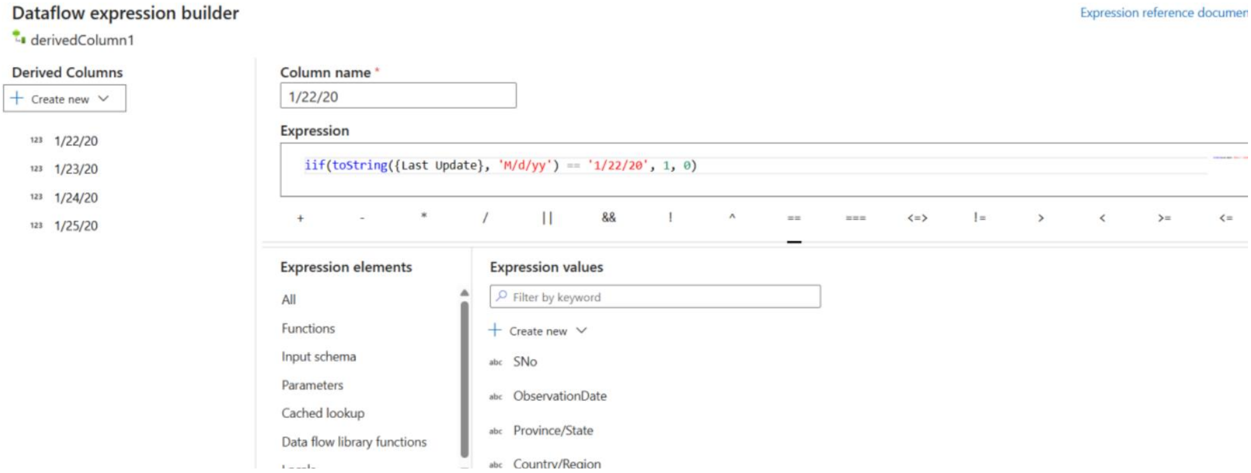
2. Go to Author tab > Pipeline > New Pipeline > and grab the Data Flow activity



3. Designing the Data Flow. To use adf’s data flow, for aggregate purpose first we must use derived columns, then aggregate and at last sink in storage which is accessible in databricks notebooks.


source1
Import data from DelimitedText1
derivedColumn1
Creating/updating the columns 'SNo, ObservationDate, Province/State, Country/Region, Last Update,
aggregate1
Aggregating data by 'Country/Region' producing columns '1/22/20, 1/23/20, 1/24/20, 1/25/20'
sink1
Columns:
5 total
Add Source

4. In derived columns, make columns for date name for which you are going to do analytical process.



Dataflow expression builder

derivedColumn1

Derived Columns

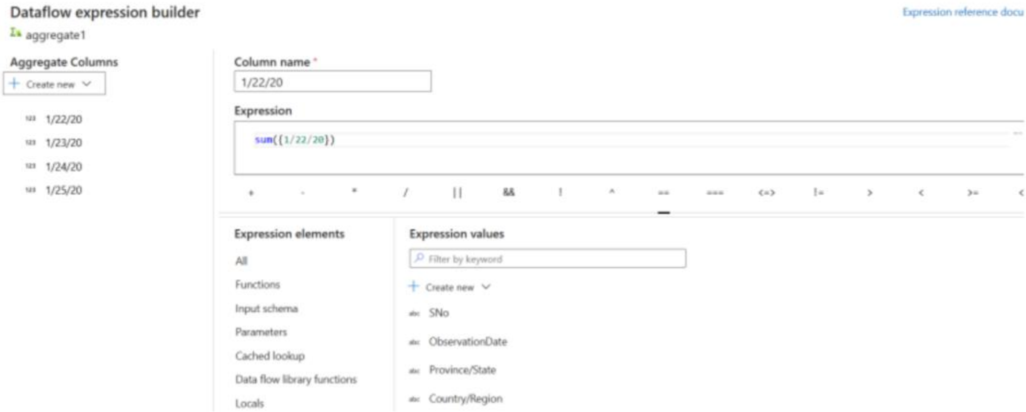
1/22/20
1/23/20
1/24/20
1/25/20

Column name *: 1/22/20

Expression:

```
iif(toString({Last Update}, 'M/d/yy') == '1/22/20', 1, 0)
```

5. After creating columns count the no of cases by date and grouping by each region.



Dataflow expression builder

aggregate1

Aggregate Columns

1/22/20
1/23/20
1/24/20
1/25/20

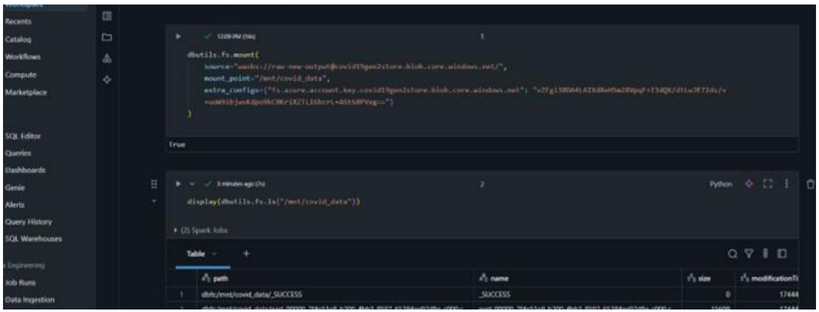
Column name *: 1/22/20

Expression:

```
sum({1/22/20})
```

TASK 2: Load the transformed data into Azure Databricks for further analysis

1. Launch a Databricks workspace and Launch it
2. Create a new notebook



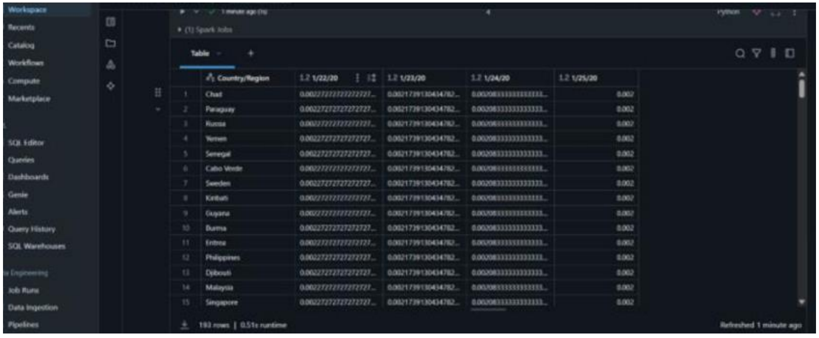
```
dbutils.fs.mount(
    source="wasbs://raw-new-output@covid19gen2store.blob.core.windows.net/",
    mount_point="/mnt/covid_data",
    extra_configs={"fs.azure.account.key.covid19gen2store.blob.core.windows.net": "vZFgl3BDA4tAIXdBaHSm28Vpqf+I3dQK/dtLwJE72dv/v
    +uoW9ibjwvKdpo9kC0KriXZTL16bcrL+ASt58PVxg=="}
)
```

```
True
```

```
display(dbutils.fs.ls("/mnt/covid_data"))
```

(2) Spark Jobs

	path	name	size	modificationTi
1	dbfs:/mnt/covid_data/_SUCCESS	_SUCCESS	0	17444



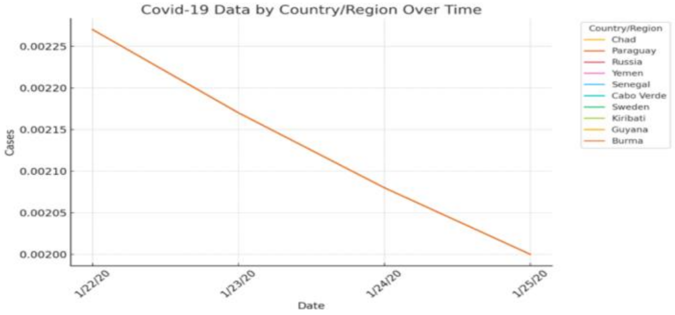
(1) Spark Jobs

	Country/Region	1/22/20	1/23/20	1/24/20	1/25/20
1	Chad	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002
2	Paraguay	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002
3	Russia	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002
4	Yemen	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002
5	Senegal	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002
6	Cabo Verde	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002
7	Sweden	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002
8	Kiribati	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002
9	Guyana	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002
10	Burma	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002
11	Eritrea	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002
12	Philippines	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002
13	Djibouti	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002
14	Malaysia	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002
15	Singapore	0.0022727272727272727...	0.0021739130434782...	0.00208333333333333...	0.002

193 rows | 0.51s runtime

Refreshed 1 minute ago

3. Creating a visualization based on number of cases per day grouping

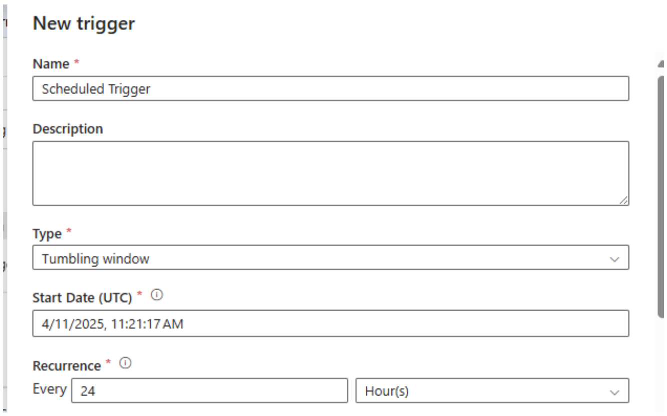


Exercise 4: Orchestrating Covid-19 Data Updates with ADF

TASK 1: Schedule the pipeline from Exercise 2 to run daily

We already accomplish this task while building the said pipeline earlier

1. Click on Trigger > New/Edit > New > select type as Tumbling Runs > define start dates and reccurance time (going with 24 hrs)



New trigger

Name *: Scheduled Trigger

Description:

Type *: Tumbling window

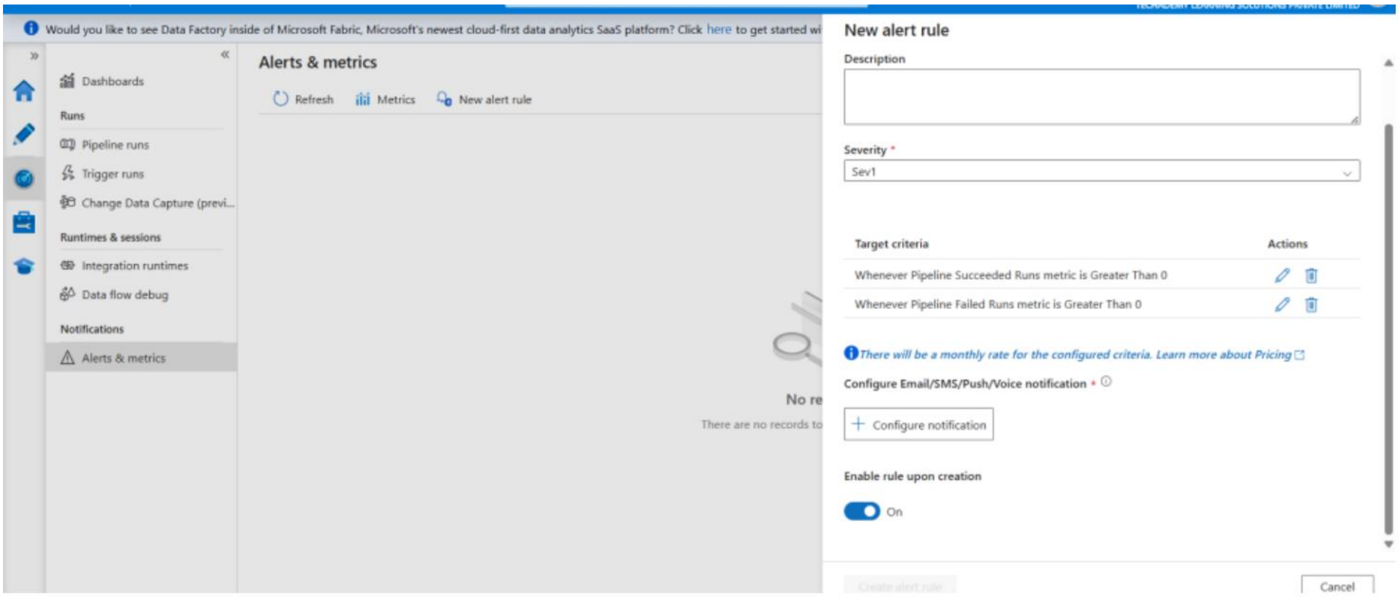
Start Date (UTC) *: 4/11/2025, 11:21:17 AM

Recurrence *: Every 24 Hour(s)

2. Publish your pipelines and Triggers.

TASK 2: Implement a notification system for pipeline success/failure

1. For notification system > Go to alerts and metrics > Declare new alert rule > In target criteria configure, when a pipeline is succeeded or failed you will get a notification. > You can choose the method of notification by configuring settings



New alert rule

Description:

Severity *: Sev1

Target criteria	Actions
Whenever Pipeline Succeeded Runs metric is Greater Than 0	
Whenever Pipeline Failed Runs metric is Greater Than 0	

There will be a monthly rate for the configured criteria. Learn more about Pricing

Configure Email/SMS/Push/Voice notification *

Enable rule upon creation: On