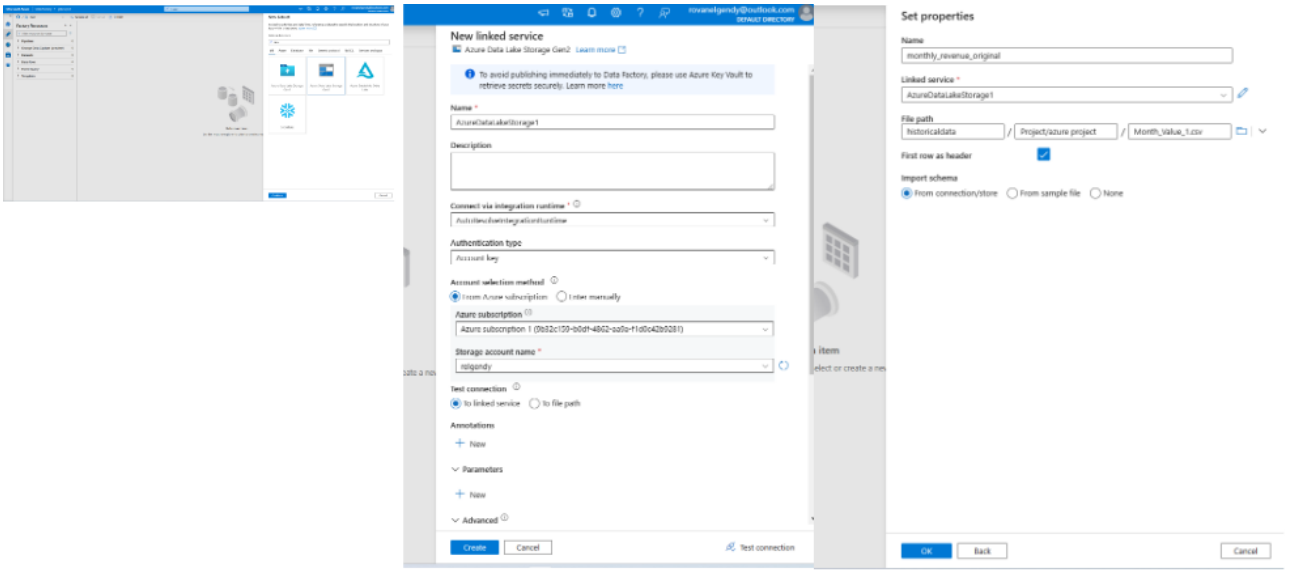
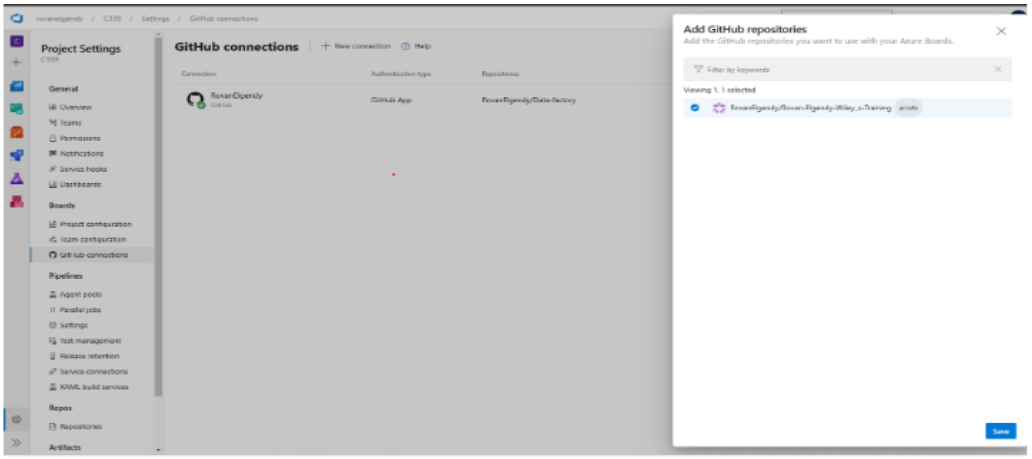
Azure Project Steps/Demo

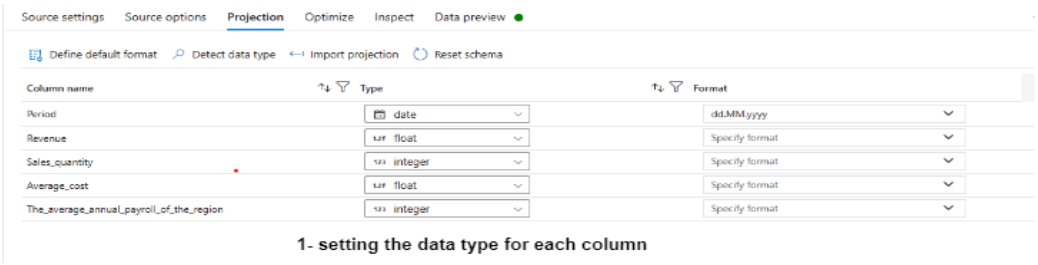
Obj: Use Data Flow to move data to Azure Data Lake Storage or Azure Blob Storage. Use Databricks to read and write data from and to these data stores, and also perform advanced analytics and machine learning tasks.



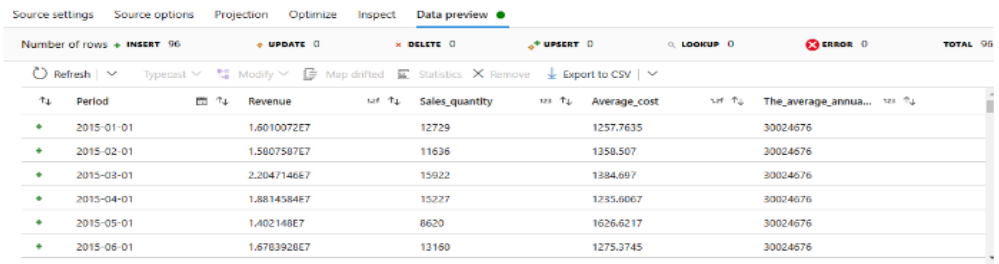
Setting up the Linked Services



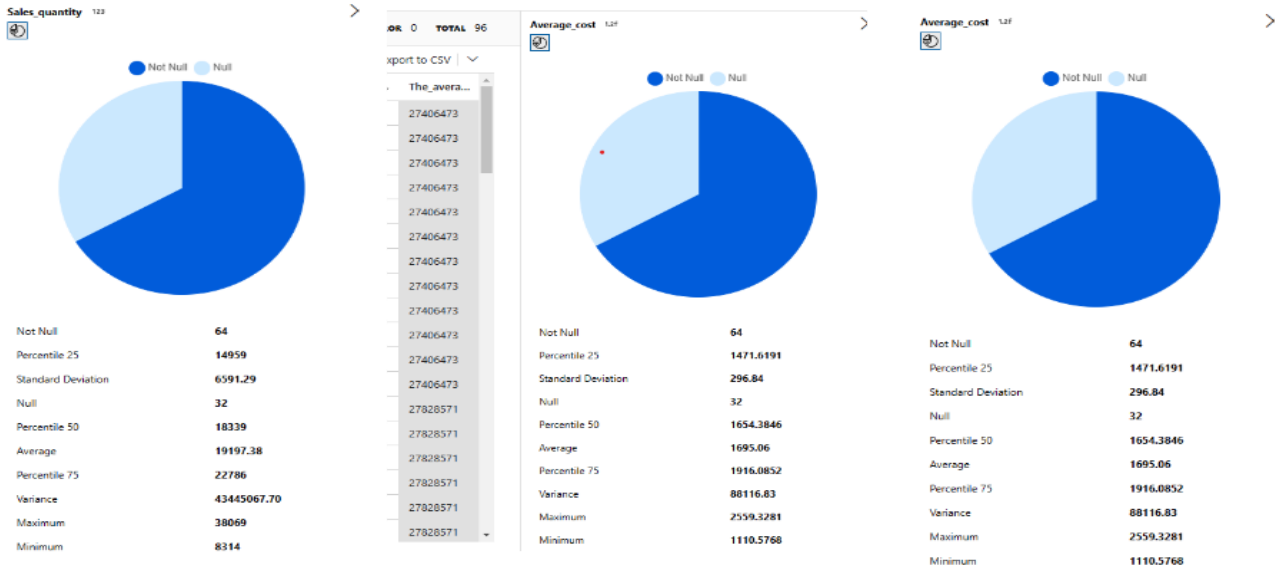
Git configuration/integration to Git Repo
from organisation settings
after adding new Repo in Git



1- setting the data type for each column

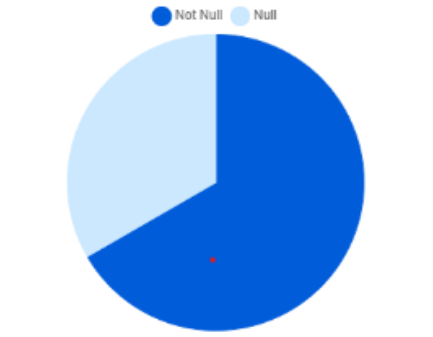


2- Data preview



3- Dataset Description/Statistics

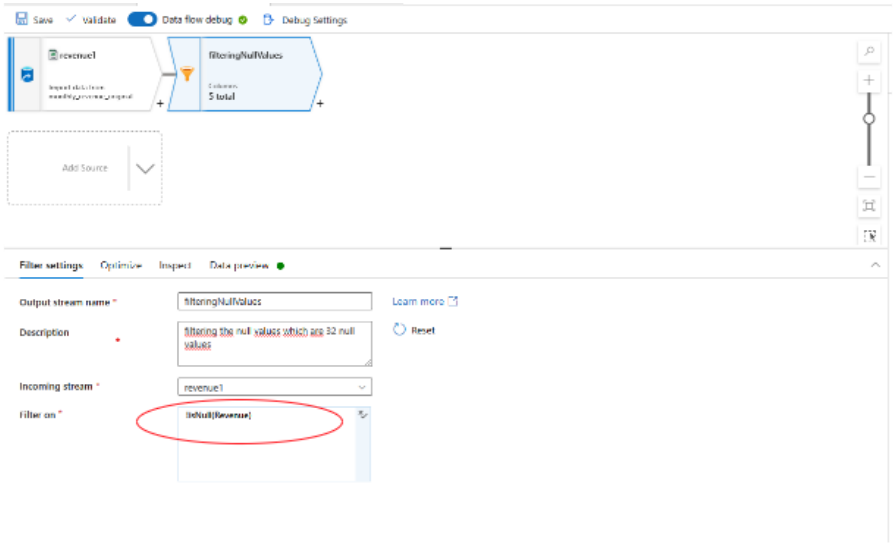
Revenue 1.2f



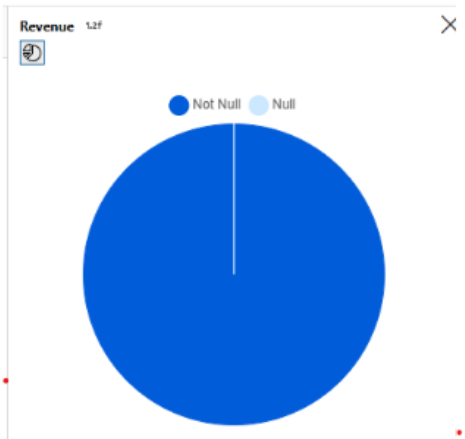
Not Null	**64**
Percentile 25	**2.2367074E7**
Standard Deviation	**11641498.36**
Null	**32**
Percentile 50	**3.1650092E7**
Average	**32360452.23**
Percentile 75	**3.9831564E7**
Variance	**135524484023262.06**
Maximum	**58756472**
Minimum	**14021480**

Data cleaning

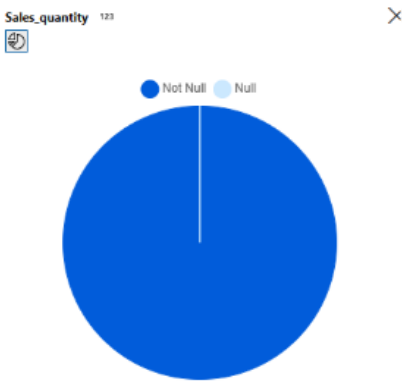
- Check the nan values from each column using the statistics above
- Using Filter data flow using the expression: !isNull(Revenue)



1- using Filter dataflow to remove the nan values using builder expression

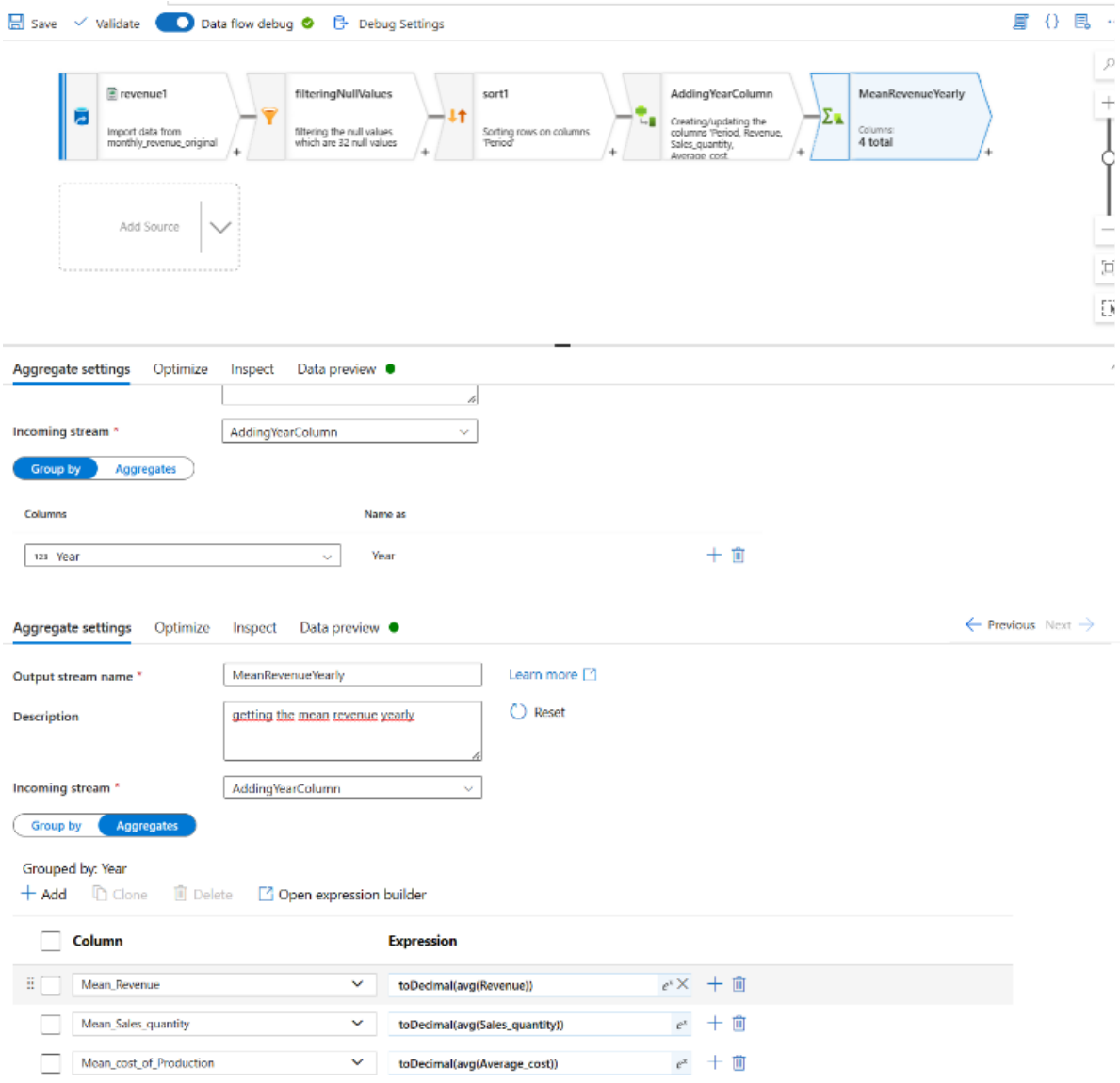


Not Null	64
Percentile 25	2.2367074E7
Standard Deviation	11641498.36
Null	0
Percentile 50	3.1650092E7
Average	32360452.23
Percentile 75	3.9831564E7
Variance	135524484023262.06
Maximum	58756472
Minimum	14021480

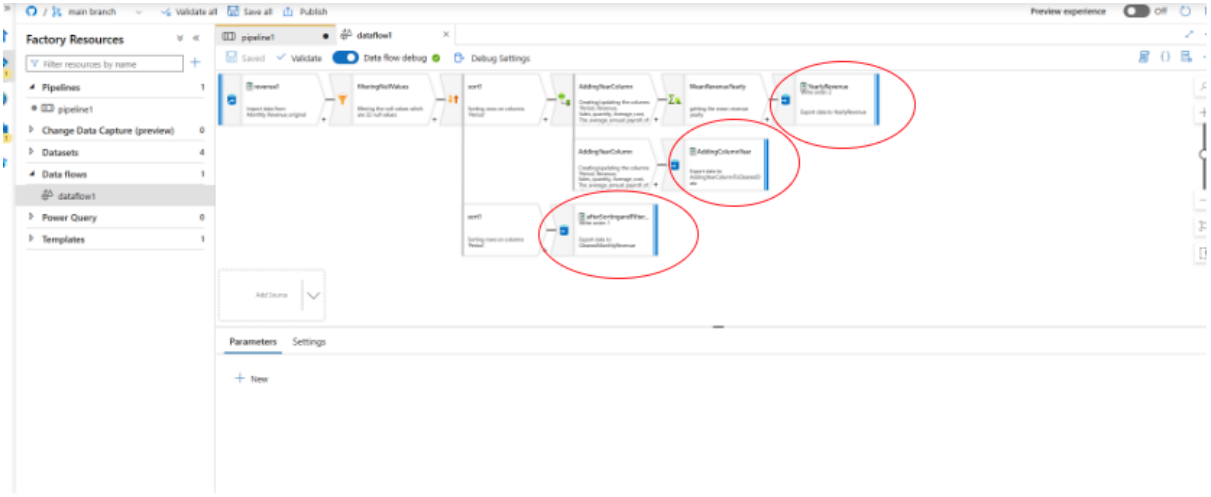


Not Null	64
Percentile 25	14959
Standard Deviation	6591.29
Null	0
Percentile 50	18339
Average	19197.38
Percentile 75	22786
Variance	43445067.70
Maximum	38069
Minimum	8314

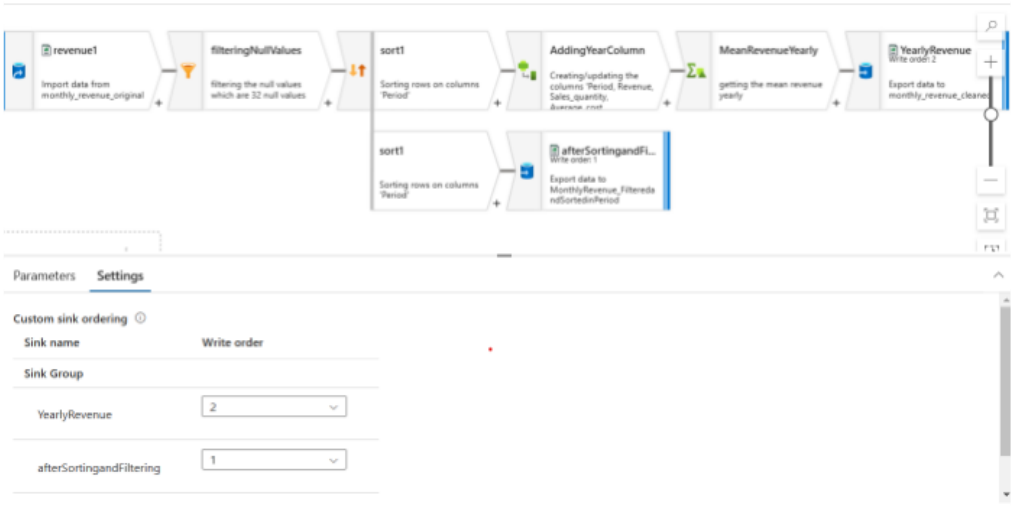
2- Columns after filtering has zero Null values



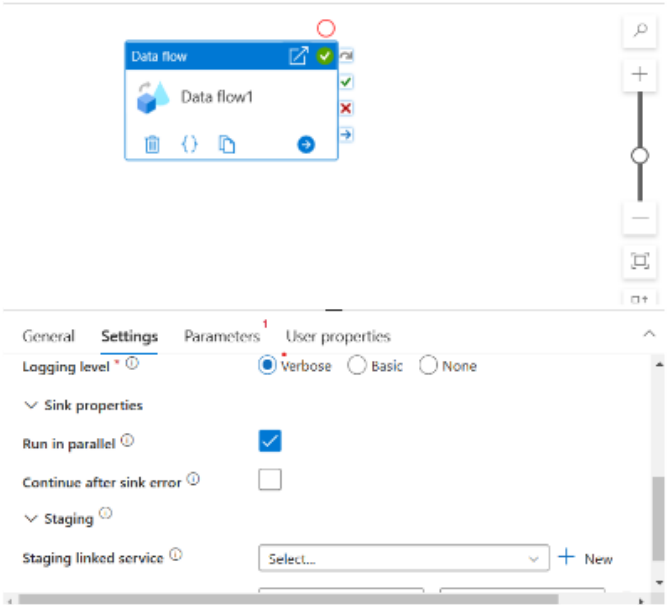
Getting the Avg of all columns per year using the aggregate transformation



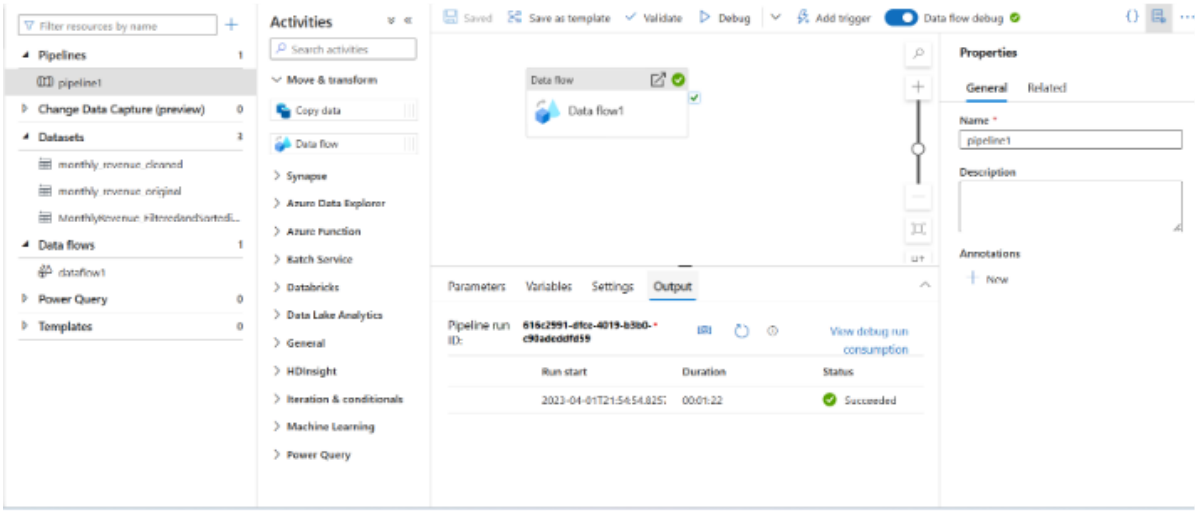
three sinks are added:
afterSortingandFiltering
AddingColumnYear
yearlyAvgRevenue



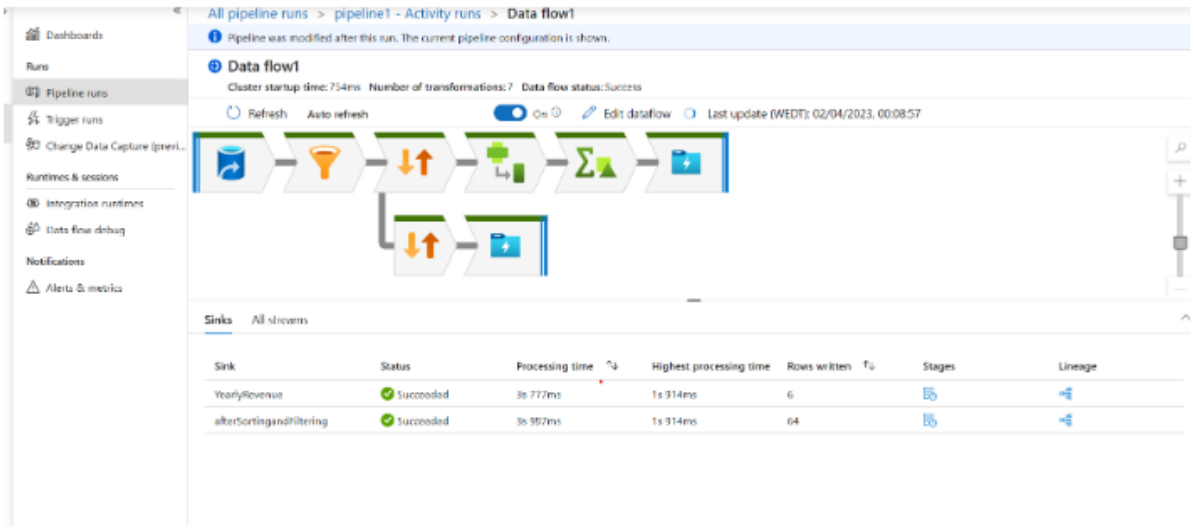
Writing the order of the three sinks



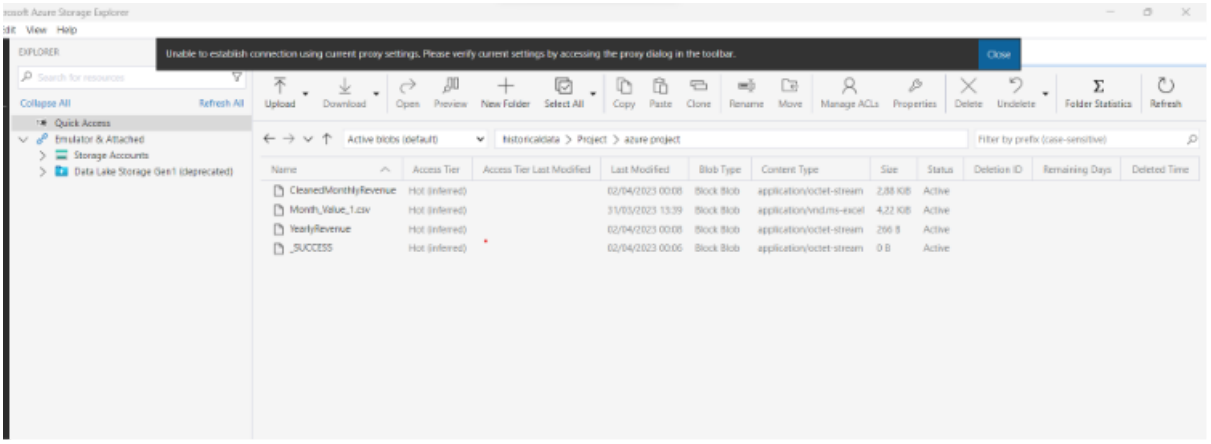
Creating new pipeline with our dataflow setting the sinks to run in parallel



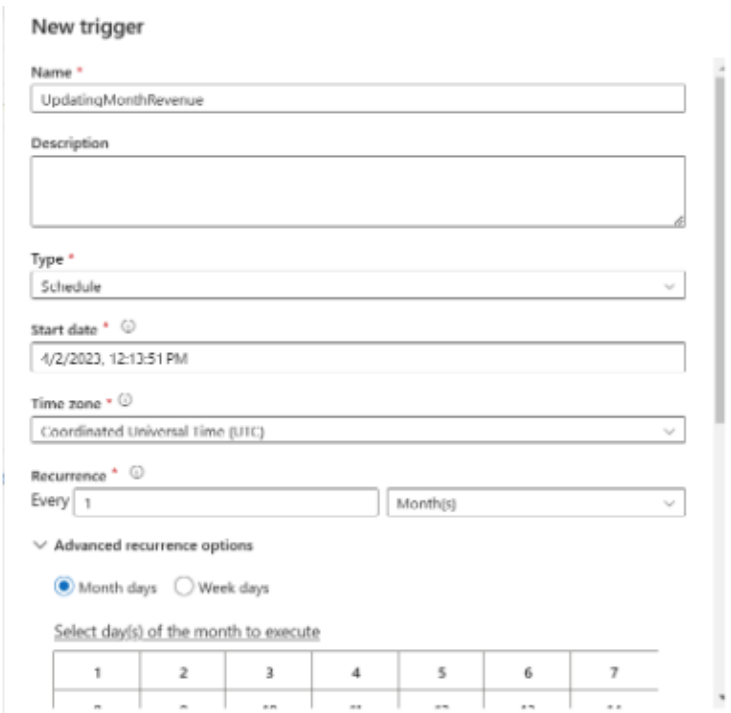
Debugging the pipeline



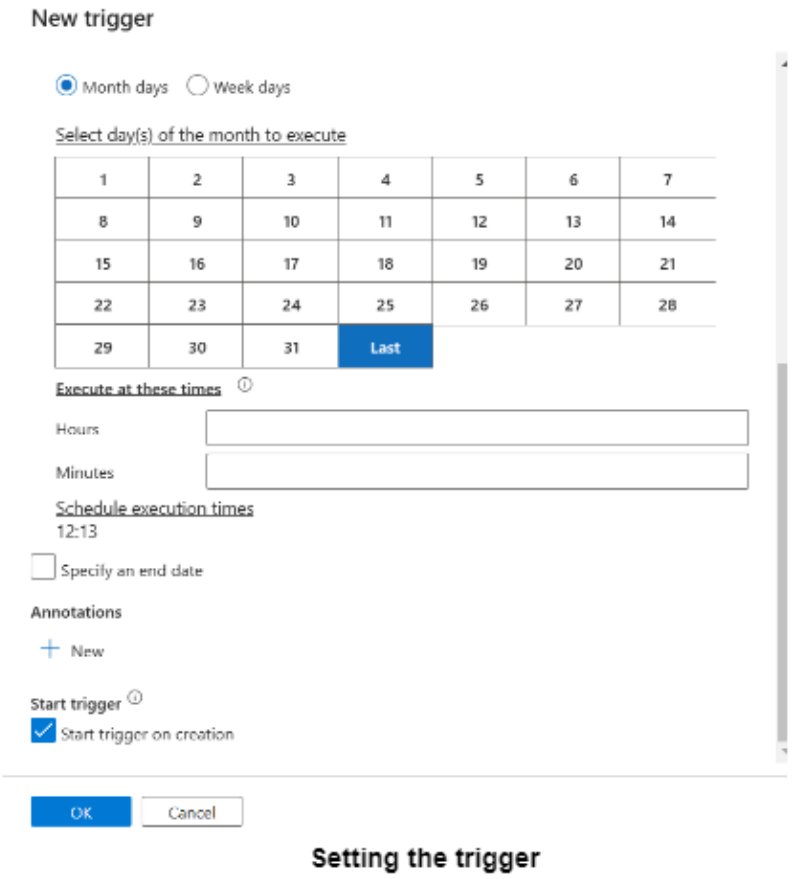
Dataflow Debugging Dashboard after enabling the debug



the 3 Sinks have been added to the azure storage after debugging the pipeline

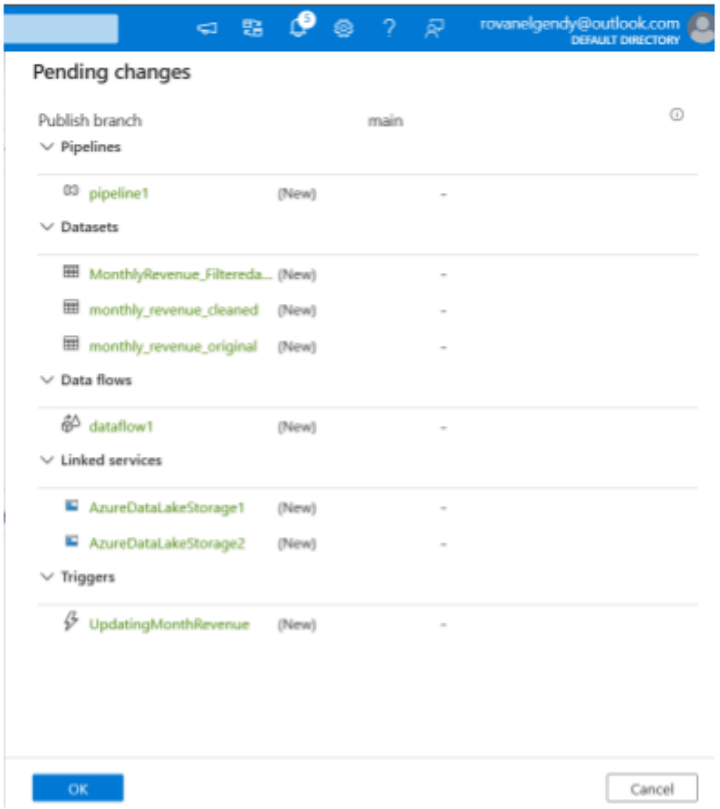


adding trigger to be triggered monthly

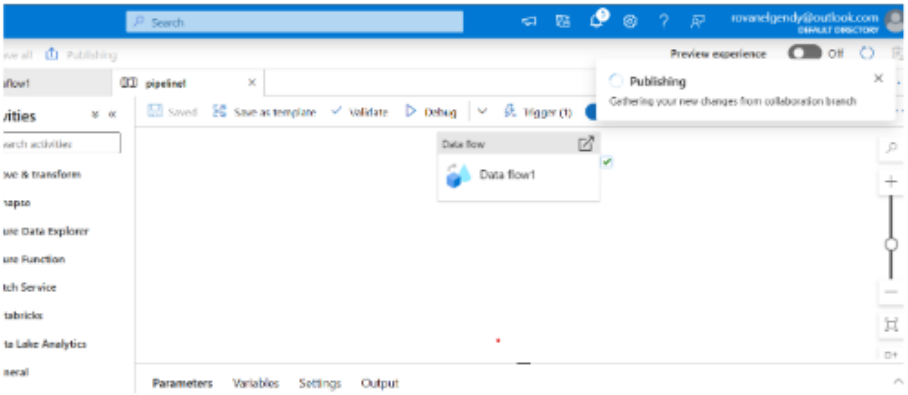


Setting the trigger

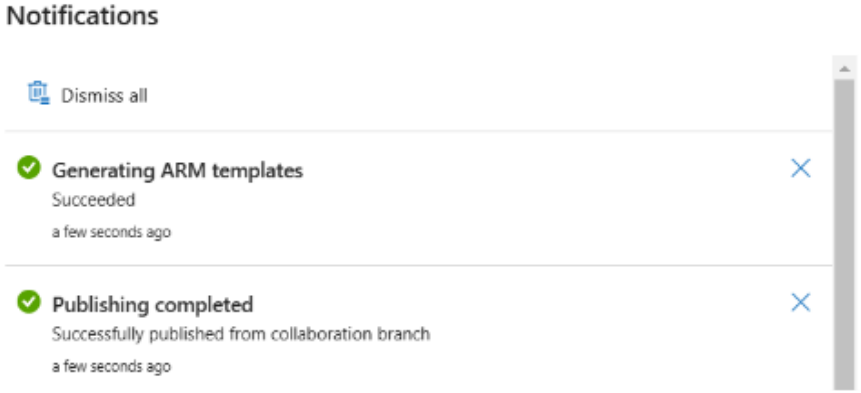
Setting the trigger

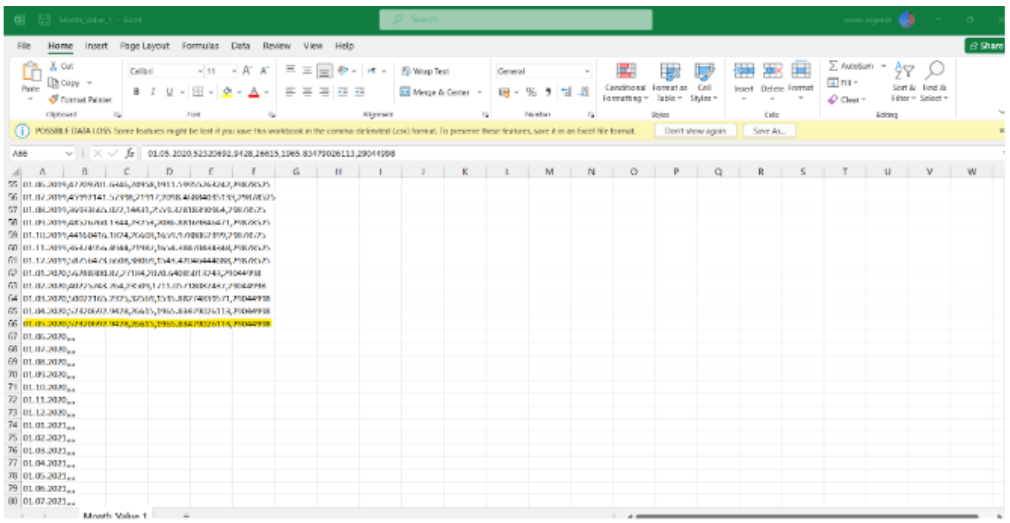


Saved and to be published

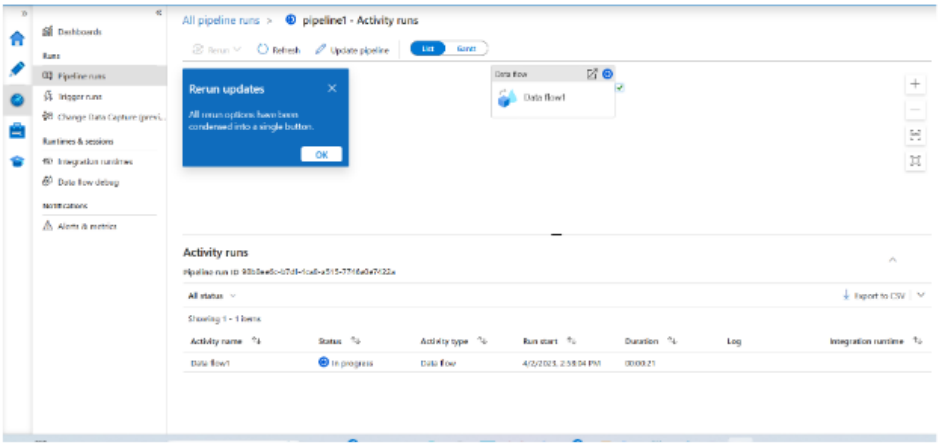


published





Updating the original dataset by one row



pushing the trigger we made earlier manually

Notifications

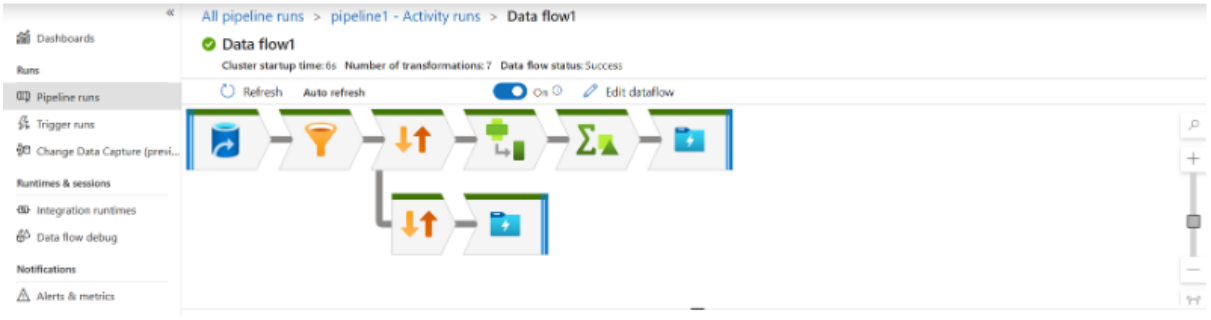
Dismiss all

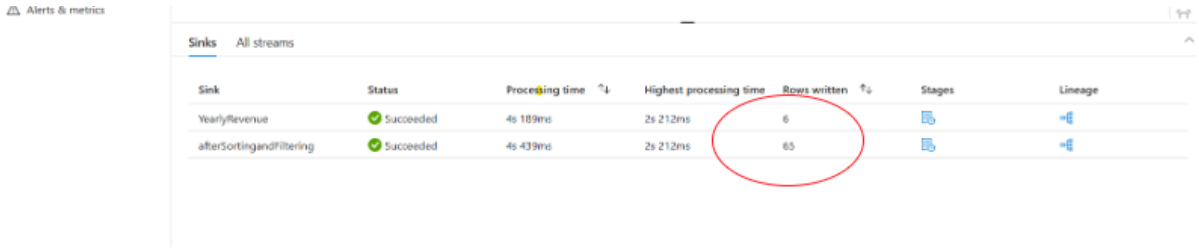
UpdatingMonthRevenue

UpdatingMonthRevenue has been saved.

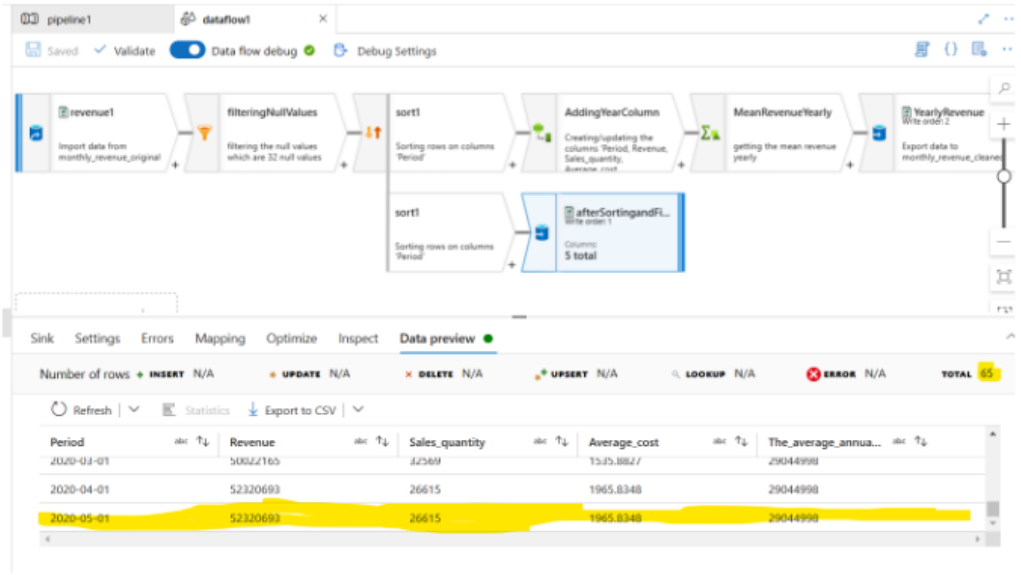
a few seconds ago

Trigger success notification

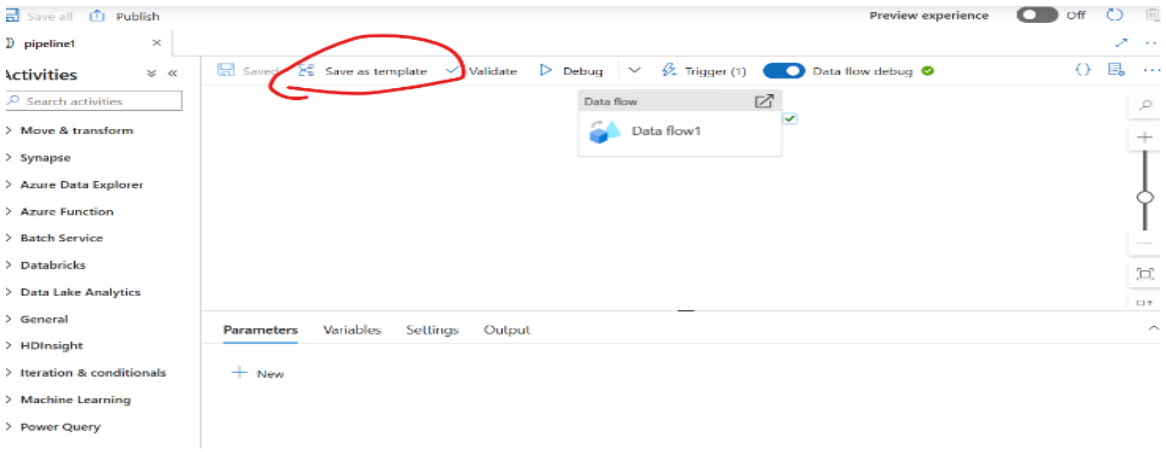




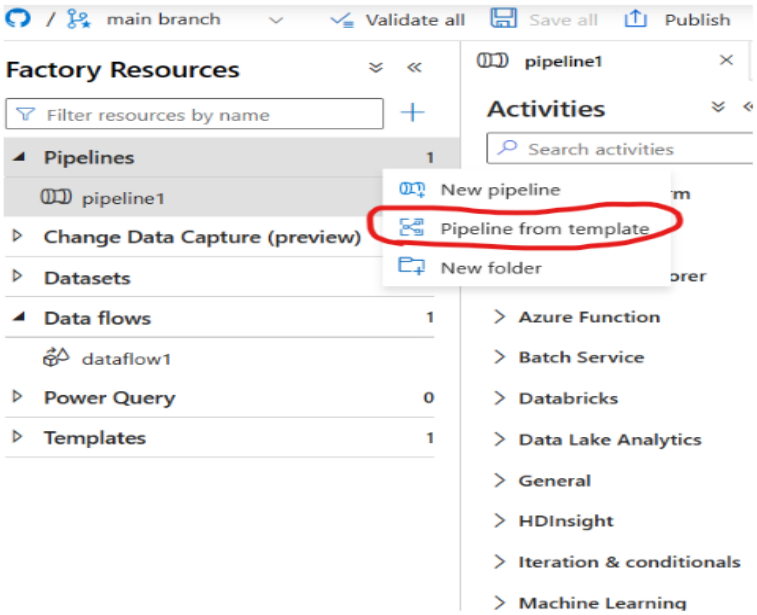
here we can see after trigger the rows read is updated to 65 rows instead of 64 rows



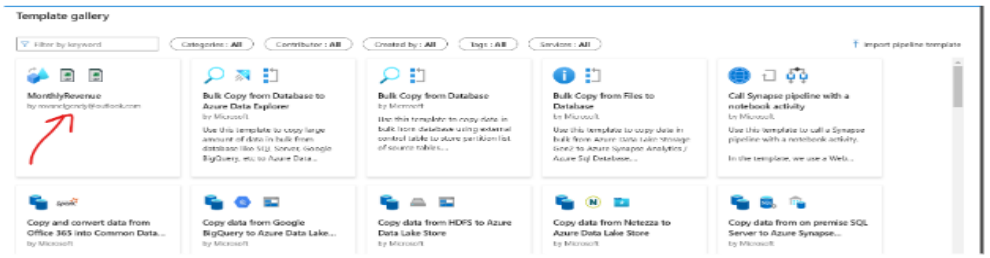
the updated row in the data preview



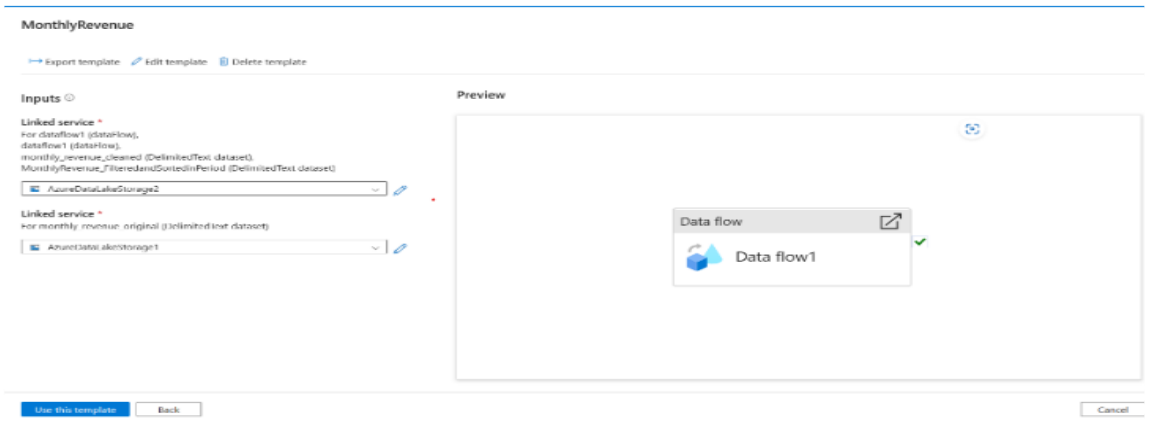
You can save the pipeline as template to import later



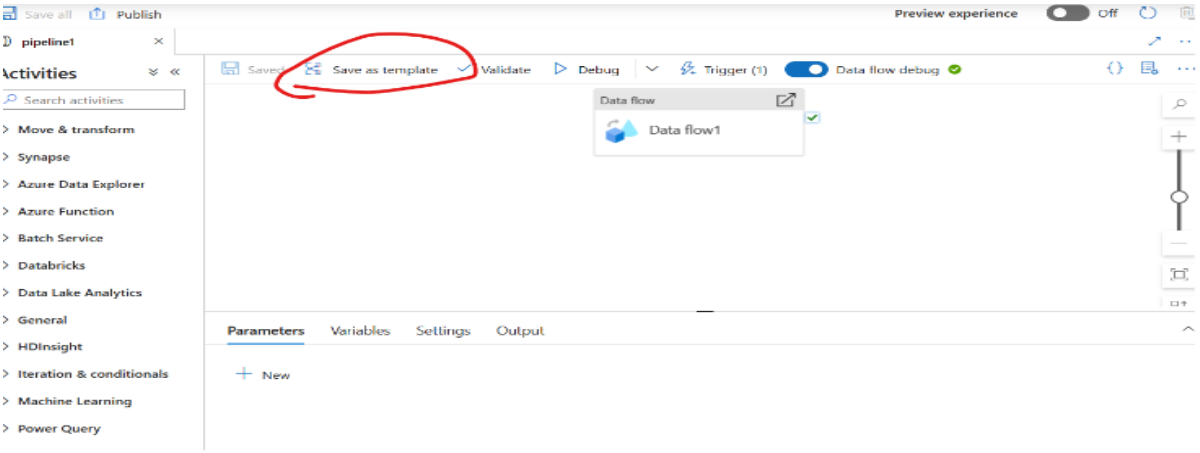
importing the tempelate



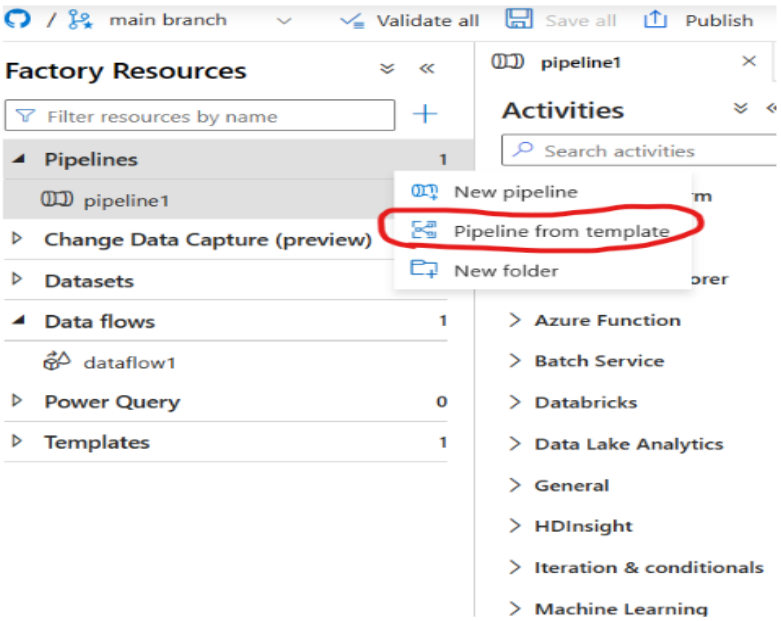
the tempelate gallery



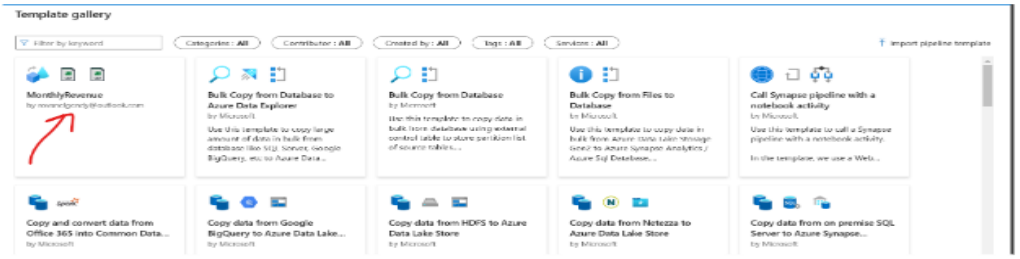
last step to click on use this tempelate



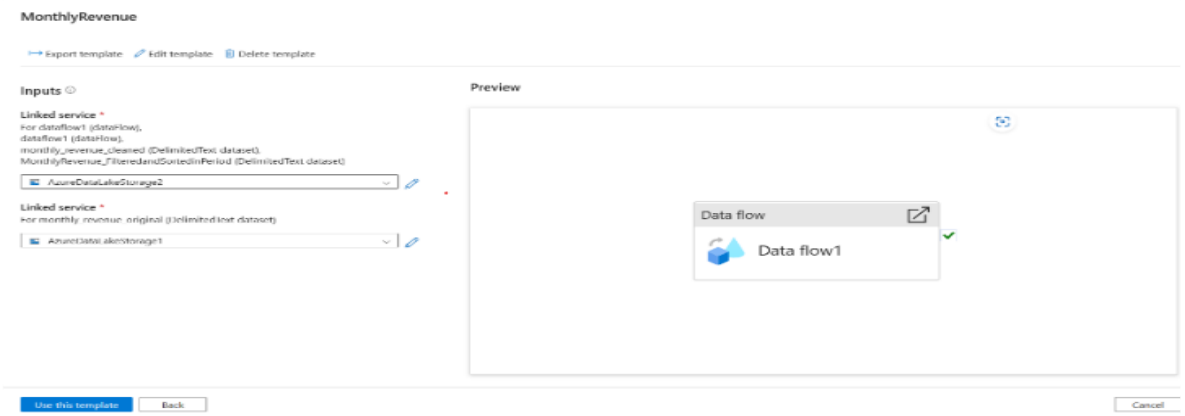
You can save the pipeline as template to import later



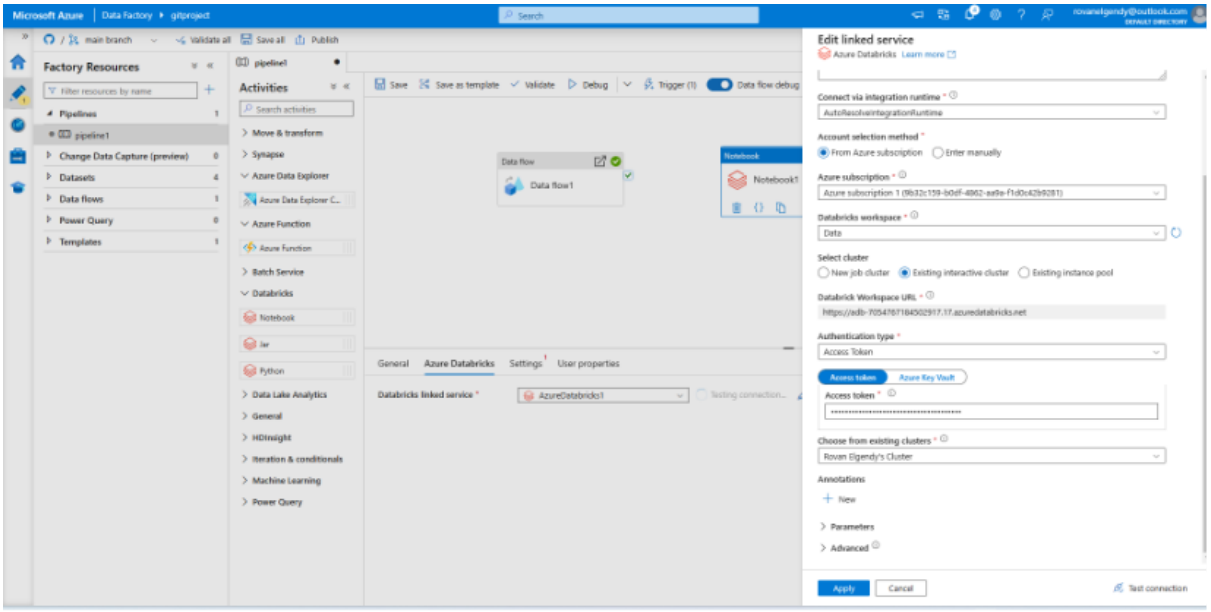
importing the tempelate



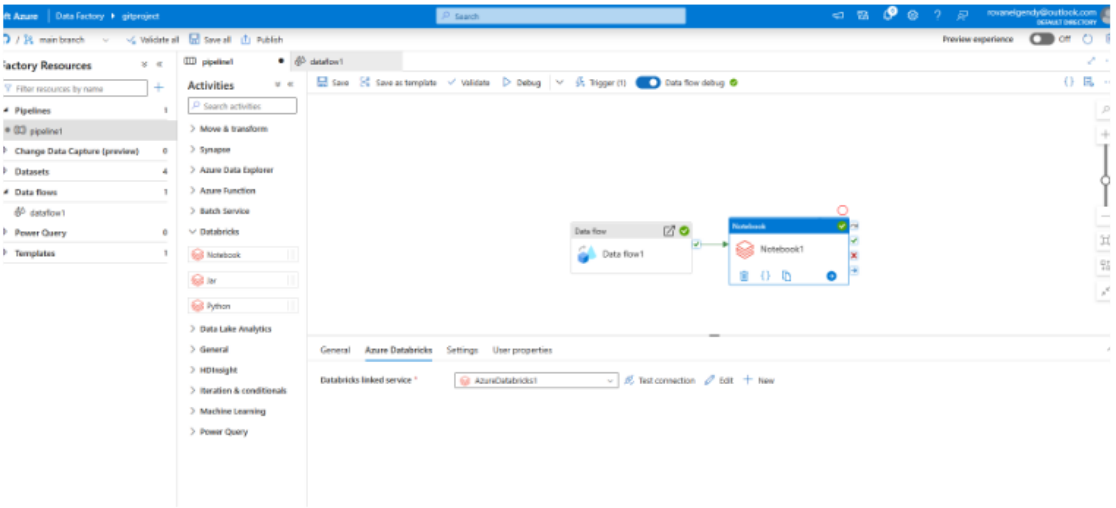
the tempelate gallery



last step to click on use this tempelate



Adding Databricks Service using existing cluster and the access token



Adding databricks notebook -which is already made in different file- to our dataflow