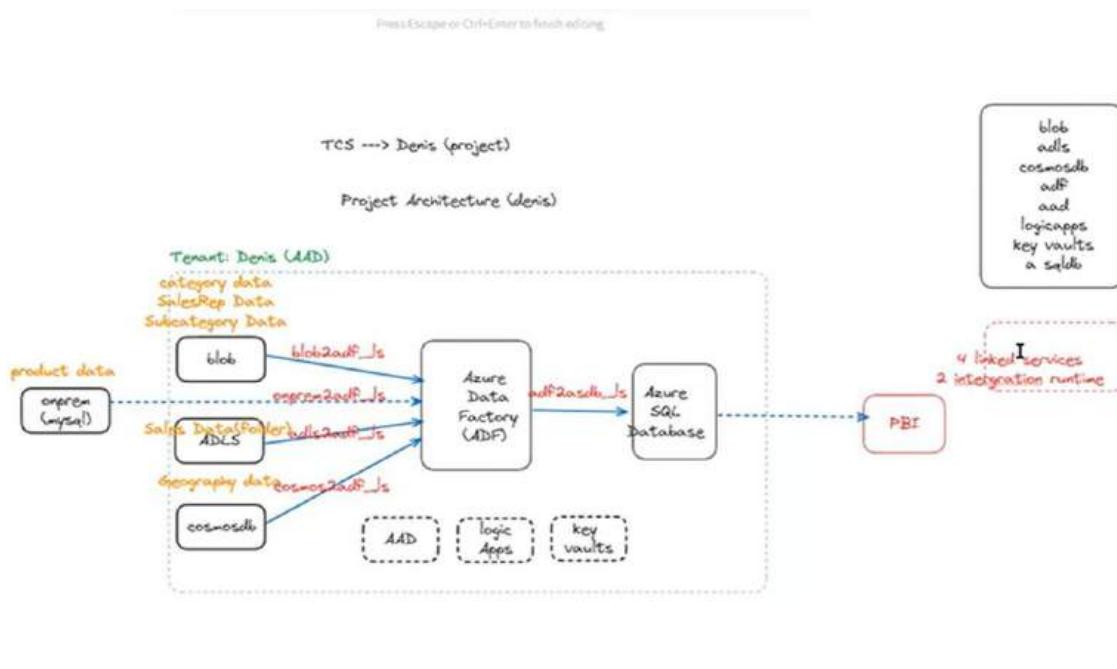


DENNIS PROJECT IN AZURE

REQUIRED SERVICES:

- AZURE BLOB STORAGE
- ADF
- AZURE SQL SERVER
- AAD
- LOGIC APPS
- KEY VALUATE
- ONPREM-MYSQL WITH INTEGRATION RUN TIME

ARCHITECTURE OF A PROJECT



NOW CREATE ALL THE SERVICE ,ALREADY YOU KNOW HOW TO CREATE THE SERVICES

NEEDED STORAGES:

1. BLOB STROAGE
2. ADLS
3. COSMOS DB
4. ONPREM DATA (FROM MYSQL) WITH INTREGRATIN RUN TIME

IMPORTANT NOTES:

While starting the project what ever data is available in onprem first we need to pull to our blob storage .

The more and more u touch the data base of onprem it will increase the cost

If team A & team b is working in project any one team take the data share it with your teams in your organization

Because this onprem data is maintained by some other onprem team by client.

Staging: In every business organization have a staging data means simple modification by raw data

So create a **staging data** container in blobs.

Link service: before starting the project create the link service and intregration run time because it will be easier to while creating the dataset in ADF

Team meet: when ever any meting happened just do the documents in IT field because it is proof what u have done what they done because of this document you don't get any blaming issues by other ,because we have a prof to show the document

Everything is done now

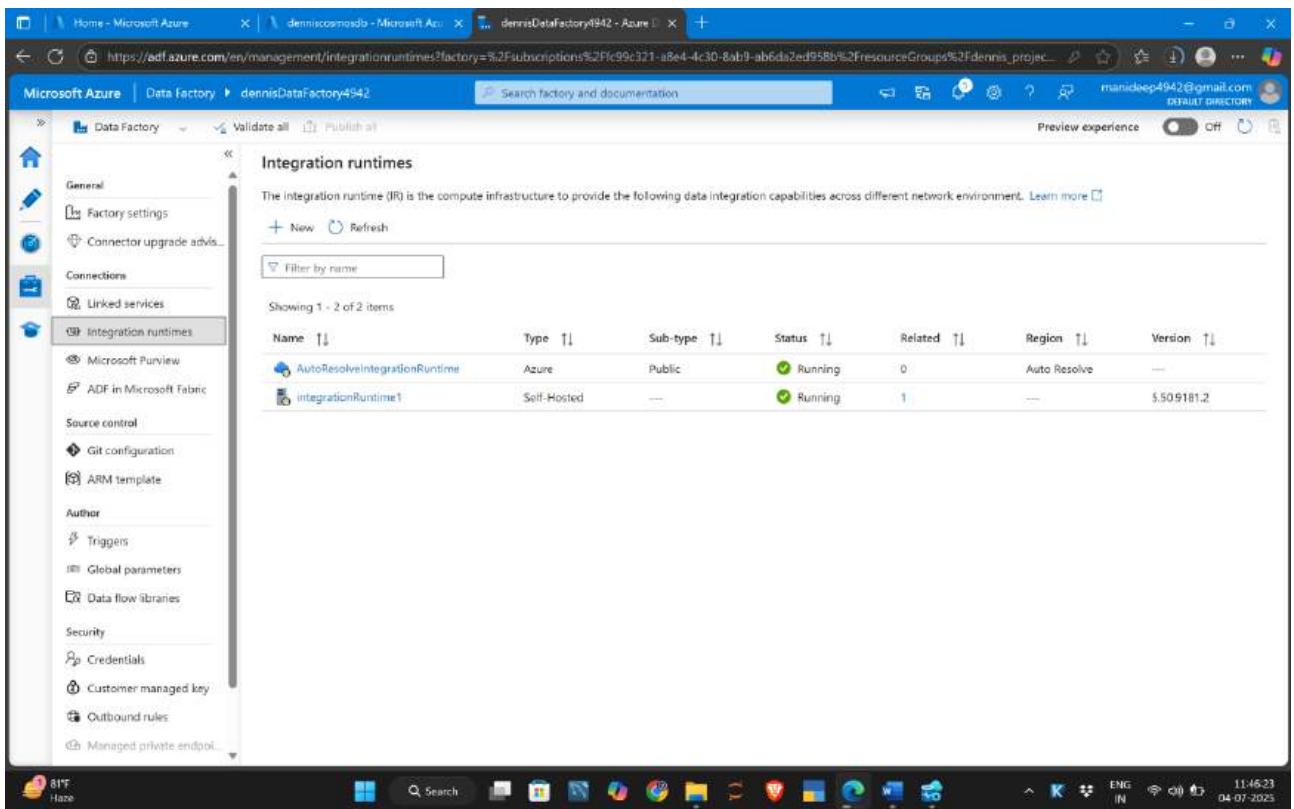
- 1.upload files subcategory,sales rep data, category file data push to blob storage
- 2.sales data files push to ADLS storage (with separate folder creation)
- 3.geographical data set push to the cosmos db
4. from onprem data push the product data.

Now in ADF create the link services to do project smoothly.

The screenshot shows the Microsoft Azure Data Factory interface. The left sidebar navigation includes General, Factory settings, Connector upgrade advisor, Connections, and the currently selected **Linked services**. The main content area is titled "Linked services" and contains a table listing four items:

Name	Type	Related	Annotations
Azureblob2Adf_ls	Azure Blob Storage	0	
AzureDataLake2Adf_ls	Azure Data Lake Storage Gen2	0	
CosmosDB/MongoDb2ADF_ls	Azure Cosmos DB for MongoDB	0	
MySql2ADF_ls	MySQL	0	

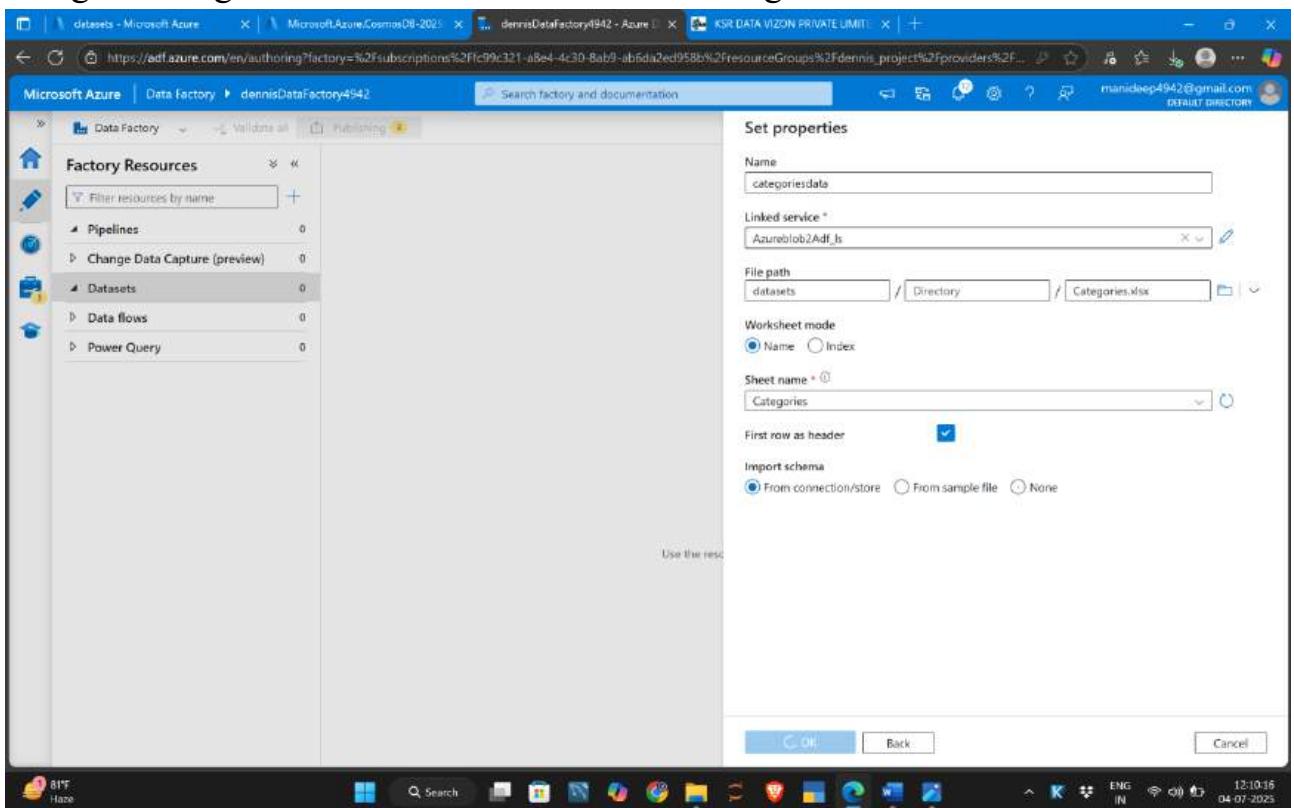
A success message box is visible in the top right corner: "Successfully created MySQL2ADF_ls (Linked service)." The status bar at the bottom shows system information like battery level (81%), network connection, and date/time (04-07-2023).



The screenshot shows the Microsoft Azure Data Factory Integration Runtimes page. The left sidebar navigation includes General, Factory settings, Connector upgrade advs..., Connections, Linked services, Integration runtimes (selected), Microsoft Purview, ADF in Microsoft Fabric, Source control (Git configuration, ARM template), Author (Triggers, Global parameters, Data flow libraries), Security (Credentials, Customer managed key, Outbound rules, Managed private endpoint). The main content area displays the 'Integration runtimes' section with two items:

Name	Type	Sub-type	Status	Related	Region	Version
AutoResolveIntegrationRuntime	Azure	Public	Running	0	Auto Resolve	---
IntegrationRuntime1	Self-Hosted	---	Running	1	---	\$509181.2

Bring the categories data set into ADF with creating in datasets



The screenshot shows the Microsoft Azure Data Factory Set properties dialog for a dataset named 'categoriesdata'. The dialog includes fields for Name (categoriesdata), Linked service (Azureblob2Adf_ls), File path (datasets / Categories.xlsx), Worksheet mode (Name selected), Sheet name (Categories), First row as header (checked), Import schema (From connection/store selected), and a 'Use the rest...' button. The left sidebar navigation shows Factory Resources (Pipelines, Change Data Capture (preview), Datasets, Data flows, Power Query) and a preview of the datasets list.

Screenshot of Microsoft Azure Data Factory preview interface showing a dataset named "categoriesdata".

Factory Resources:

- Pipelines
- Change Data Capture (preview)
- Datasets
- Data Flows
- Power Query

Preview data:

Linked service: Azureblob2Adf_ls
Object: Categories.xlsx

CategoryKey	Category
1	Special
2	General

Properties:

Name: categoriesdata

Description:

Annotations:

Screenshot of Microsoft Azure Data Factory publishing interface for the "categoriesdata" dataset.

Factory Resources:

- Pipelines
- Change Data Capture (preview)
- Datasets
- Data Flows
- Power Query

categoriesdata

Connection:

Linked service: Azureblob2Adf_ls

File path: datasets /

Compression type: No compression

Worksheet mode: Name (selected)

Sheet name: Categories

Range: e.g. A3:H5

Null value:

First row as header:

Publish all:

You are about to publish all pending changes to the live environment. [Learn more](#)

Pending changes (1):

NAME	CHANGE	EXISTING
categoriesdata	(New)	-

Buttons:

Publish Cancel

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. Under 'Datasets', 'SalesRepData' is selected. In the main area, a 'Preview data' window is open for the object 'SalesRep.xlsx'. The linked service is 'Azureblob2Adf_ls'. The data table has two columns: 'SalesRepID' and 'Sales Rep Name'. The rows are:

SalesRepID	Sales Rep Name
1	Jan Novotny
2	John White
3	Ellen Woody
4	Mark Spangler
5	Elie Gill
6	Bill Muray
7	El Bob

At the bottom of the preview window, there is a checkbox labeled 'First row as header' with a checked status.

Now we are pulling data from onprem form mysql her it is the data prieview

The screenshot shows the Microsoft Azure Data Factory interface. The 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. Under 'Datasets', 'ProductData' is selected. In the main area, a 'Preview data' window is open for the object 'product'. The linked service is 'MySQL2ADF_ls'. The data table has six columns: ProductID, Sub Category Key, Color, ProductName, RetailPrice, and StandardCost. The rows are:

ProductID	Sub Category Key	Color	ProductName	RetailPrice	StandardCost
1	3	Red	Alder	23.95	7.55
2	2	Blue	Linder	23.95	7.55
3	2	Green	Magnum	23.95	7.55
4	1	Red	Quad	43.95	13.75
5	1	Blue	Black Monk	43.95	13.75
6	4	Green	Quad	43.95	13.75
7	1	Red	Bing	26.95	8.25
8	3	Blue	Vanhelen	26.95	8.25
9	1	Green	Magnum	26.95	8.25
10	1	Florescent Pink	Carlota	29.95	9.15

To the right of the preview window, a 'Properties' panel is visible, showing the 'General' tab with 'Name' set to 'ProductData' and 'Description' empty. There is also an 'Annotations' section with a '+ New' button.

And just do the publish

You called as a dataengineer when u use cost of the pipelines is less

Now bring the onprem data in to blob storage first:

The screenshot shows the Azure Data Factory pipeline editor. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. A pipeline named 'Onpremysql_to_blob' is selected. The main workspace displays the 'Onpremysql_to_blob' pipeline with a single 'Copy data' activity. The 'Source' tab is selected, showing 'Source dataset' set to 'ProductData'. Below it, 'Use query' is set to 'Table'. The pipeline is currently in preview mode, indicated by a green checkmark icon.

I changed the product.txt into product.csv

The screenshot shows the 'Raw_productdata' dataset configuration in the Azure Data Factory dataset editor. The 'Connection' tab is selected. The 'Linked service' dropdown is set to 'Azureblob2Adf_ls'. The 'File path' field is set to 'datasets/Directory/Product.csv'. Other settings include 'Compression type: No compression', 'Column delimiter: Comma (,),' 'Row delimiter: Default (\r\n or \n\r)', 'Encoding: Default(UTF-8)', 'Quote character: Double quote ("), and 'Escape character: Backslash (\'). The 'First row as header' checkbox is checked. A red line highlights the 'File path' field.

Run pipeline again

The screenshot shows the Microsoft Azure Storage Container blade for the 'datasets' container. The left sidebar includes 'Overview', 'Diagnose and solve problems', 'Access Control (IAM)', and 'Settings'. The main area displays a table of blobs:

Name	Last modified	Access tier	Blob type	Size	Lease state
Categories.xlsx	4/7/2025, 11:23:04 am	Hot (Inferred)	Block blob	17.03 KiB	Available
SalesRep.xlsx	4/7/2025, 11:23:04 am	Hot (Inferred)	Block blob	17.62 KiB	Available
SubCategories.xlsx	4/7/2025, 11:23:04 am	Hot (Inferred)	Block blob	18.09 KiB	Available
product.txt	4/7/2025, 1:05:16 pm	Hot (Inferred)	Block blob	778 B	Available

And also pulled data from the cosmos her it

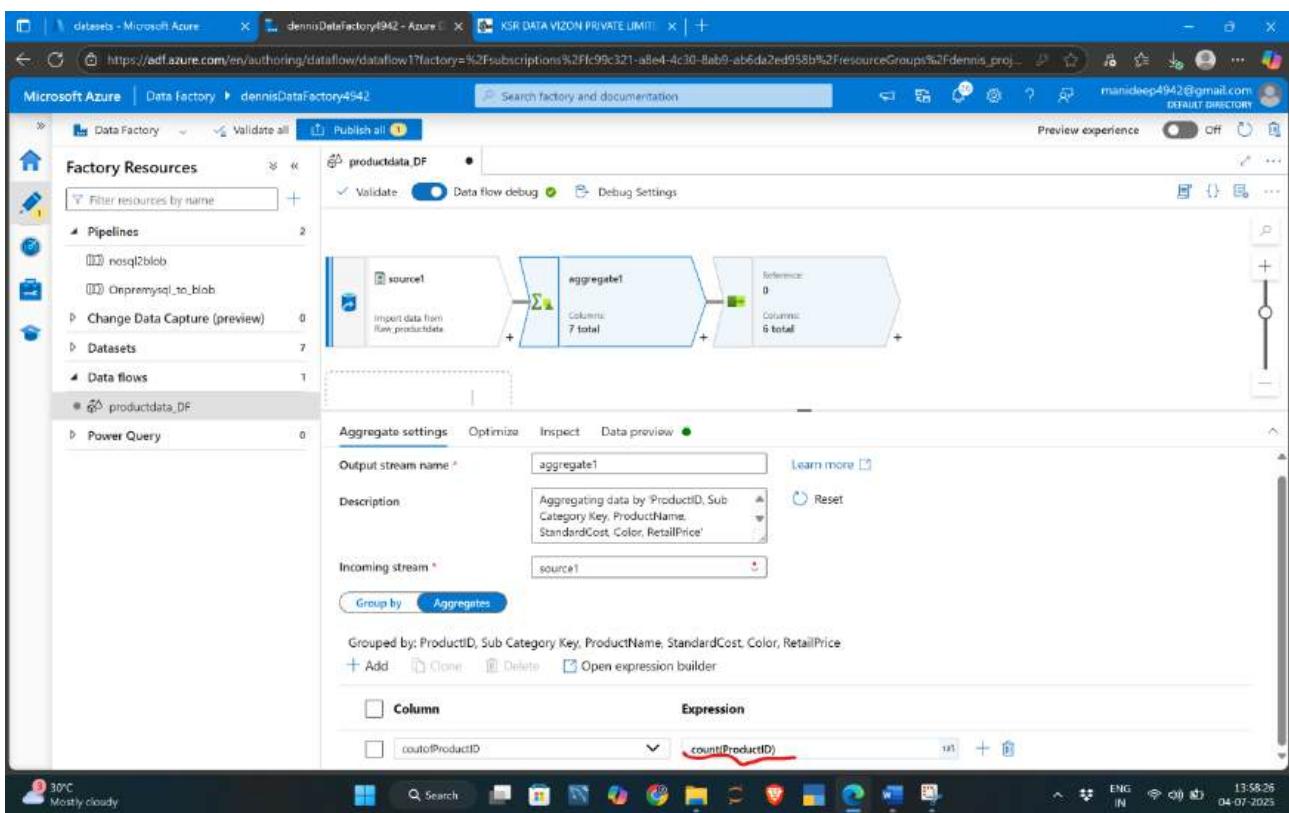
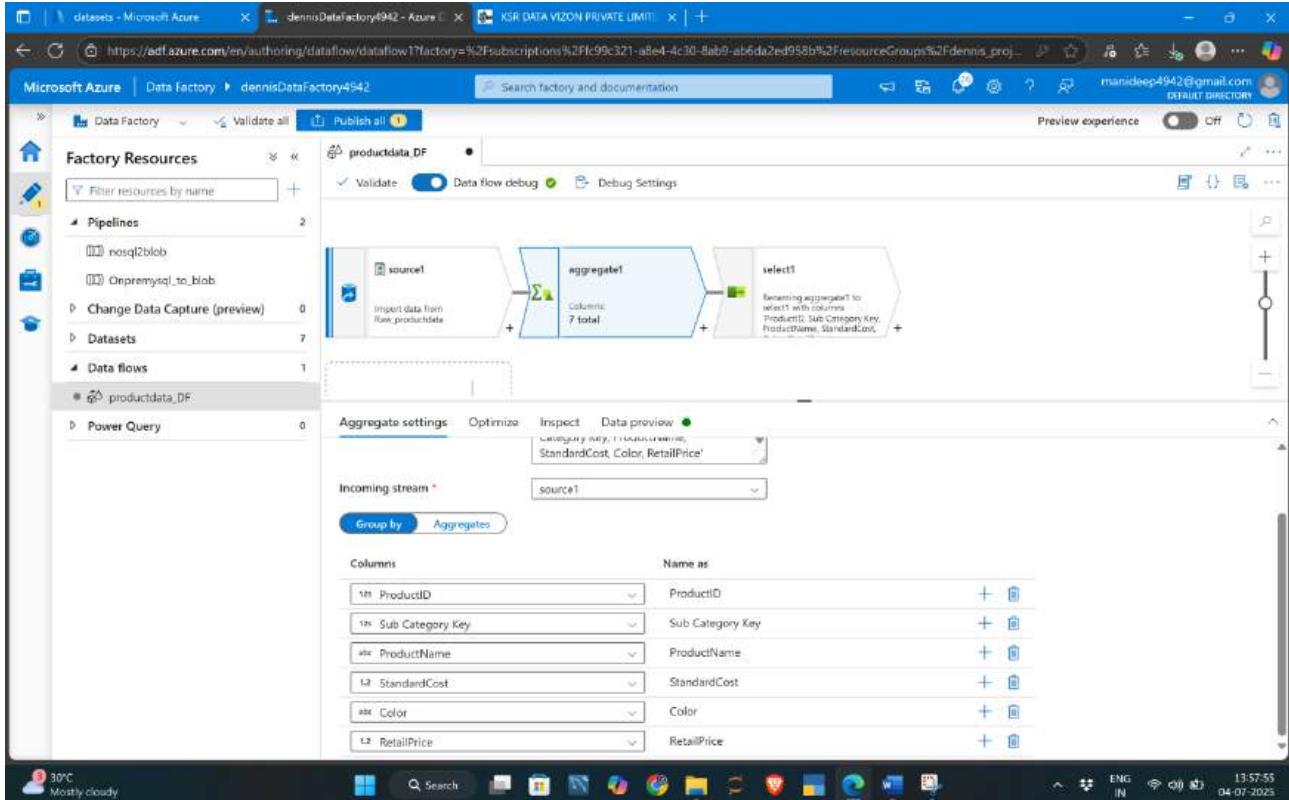
The screenshot shows the Microsoft Azure Storage Container blade for the 'datasets' container. The left sidebar includes 'Overview', 'Diagnose and solve problems', 'Access Control (IAM)', and 'Settings'. The main area displays a table of blobs, including newly transformed files:

Name	Last modified	Access tier	Blob type	Size	Lease state
Categories.xlsx	4/7/2025, 11:23:04 am	Hot (Inferred)	Block blob	17.03 KiB	Available
Geography.csv	4/7/2025, 1:22:27 pm	Hot (Inferred)	Block blob	430 B	Available
Product.csv	4/7/2025, 1:09:14 pm	Hot (Inferred)	Block blob	778 B	Available
SalesRep.xlsx	4/7/2025, 11:23:04 am	Hot (Inferred)	Block blob	17.62 KiB	Available
SubCategories.xlsx	4/7/2025, 11:23:04 am	Hot (Inferred)	Block blob	18.09 KiB	Available

For this 5 dataset we need to do some basic transformation and push to stagingdataset folder.

In first dataflow we are doing some small transaction

We are doing group by in aggregate function :



This will show the duplicates how many there in this product dataset

productdata_DF

source1

aggregate1

select1

Data preview

ProductID	Sub Category Key	ProductName	StandardCost	Color	RetailPrice	countofProductID
1	3	Alder	7.55	Red	23.95	1
2	2	Linder	7.55	Blue	23.95	1
3	2	Magnum	7.55	Green	23.95	1
4	1	Quad	13.75	Red	43.95	1
5	1	Black Monk	13.75	Blue	43.95	1
6	4	Quad	13.75	Green	43.95	1
7	1	Bing	8.25	Red	26.95	2
8	3	VanHelen	8.25	Blue	26.95	1
9	1	Magnum	8.25	Green	26.95	1
10	1	Carlota	9.15	Florescent Pink	29.95	2
11	4	Carlota	9.15	Florescent Blue	29.95	1

Now delete the count of product:

Select settings

aggregate1

Options

Input columns

aggregate1's column

Name as

ProductID

Sub Category Key

ProductName

StandardCost

Color

RetailPrice

countofProductID

7 mappings: All inputs mapped

BAJFINANCE 132%

The screenshot shows the Microsoft Azure Data Factory Data Flow blade. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, and Data flows. Under 'Data flows', 'productdata_DF' is selected. The main area displays the data flow structure: source1 (Import data from Raw.productdata) -> aggregate1 (Aggregating data by ProductID, SubCategoryKey, ProductName, StandardCost, Color, RetailPrice producing) -> select1 (Selecting aggregated rows with columns ProductID, SubCategoryKey, ProductName, StandardCost, Color, RetailPrice) -> sink1 (Sink). Below the diagram, the 'Data preview' tab is active, showing a table of data with 11 rows. A red vertical line highlights the first column (ProductID).

ProductID	Sub Category Key	ProductName	StandardCost	Color	RetailPrice
1	3	Alder	7.55	Red	23.95
2	2	Linder	7.55	Blue	23.95
3	2	Magnum	7.55	Green	23.95
4	1	Quad	13.75	Red	43.95
5	1	Black Monk	13.75	Blue	43.95
6	4	Quad	13.75	Green	43.95
7	1	Bing	8.25	Red	26.95
8	3	VanHelen	8.25	Blue	26.95
9	1	Magnum	8.25	Green	26.95
10	1	Carlota	9.15	Fluorescent Pink	29.95
11	4	Carlota	9.15	Fluorescent Blue	29.95

Now we are going to do sink the data.:)

The screenshot shows the Microsoft Azure Data Factory Data Flow blade with the 'sink1' configuration open. The 'Sink' tab is selected. The configuration includes:

- Output stream name:** sink1
- Description:** Export data to stagingproductdata
- Incoming stream:** select1
- Sink type:** Dataset (selected)
- Dataset:** stagingproductdata
- Options:** Allow schema drift (checked), Validate schema (unchecked)

Now do publish the dataflow

Now categories dataflow

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, and Data flows. Under Data flows, there is a section for 'categories.DF' which is currently selected. The main workspace displays the 'categories.DF' pipeline. It contains a single data flow step with one source named 'source1'. Below the pipeline diagram, the 'Data preview' tab is active, showing a table with the following data:

CategoryKey	Category
1	Special
2	General

In this we don't have any duplicate but we need to check.

Screenshot of Microsoft Azure Data Factory Data Flow blade for 'categories_DF'.

The Data Flow interface shows a pipeline with two main components:

- source1**: Import data from 'categorydata'.
- aggregate1**: An aggregate function with three output columns: 'Category', 'CategoryKey', and 'countofcategory'.

Aggregate settings section details:

- Output stream name**: aggregate1
- Description**: Aggregating data by 'Category'. 'CategoryKey' producing columns 'countofcategory'
- Incoming stream**: source1

Aggregates tab is selected, showing the grouped columns:

Columns	Name as
Category	Category
CategoryKey	CategoryKey

Screenshot of Microsoft Azure Data Factory Data Flow blade for 'categories_DF'.

The Data Flow interface shows a pipeline with two main components:

- source1**: Import data from 'categorydata'.
- aggregate1**: An aggregate function with three output columns: 'Category', 'CategoryKey', and 'countofcategory'.

Aggregate settings section details:

- Output stream name**: aggregate1
- Description**: Aggregating data by 'Category'. 'CategoryKey' producing columns 'countofcategory'
- Incoming stream**: source1

Aggregates tab is selected, showing the grouped columns and an expression builder:

Grouped by: Category, CategoryKey

Column	Expression
countofcategory	count(Category)

Factory Resources

Pipelines: nosql2blob, Onpremssql_to_blob

Change Data Capture (preview): 0

Datasets: 8

Data flows: 2

categories_DF

productdata_DF

Power Query: 0

Validate all

Search factory and documentation

Preview experience: Off

Data preview

	Category	CategoryKey	countofcategory
+	Special	1	1
+	General	2	1

so there is no duplicates

Factory Resources

Pipelines: nosql2blob, Onpremssql_to_blob

Change Data Capture (preview): 0

Datasets: 8

Data flows: 2

categories_DF

productdata_DF

Power Query: 0

Validate all

Search factory and documentation

Preview experience: Off

Select settings

Output stream name: select1

Description: Renaming aggregate1 to select1 with columns 'Category', 'CategoryKey', 'countofcategory'

Incoming stream: aggregate1

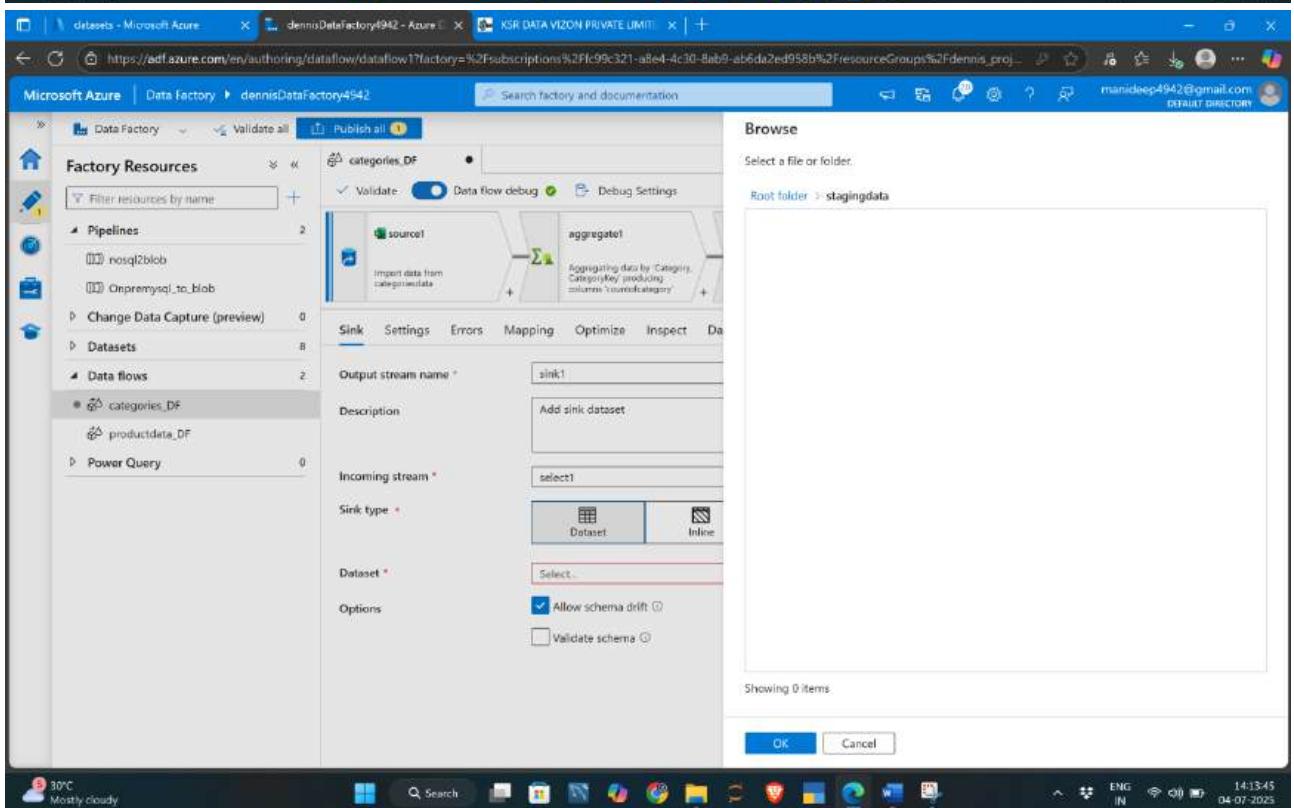
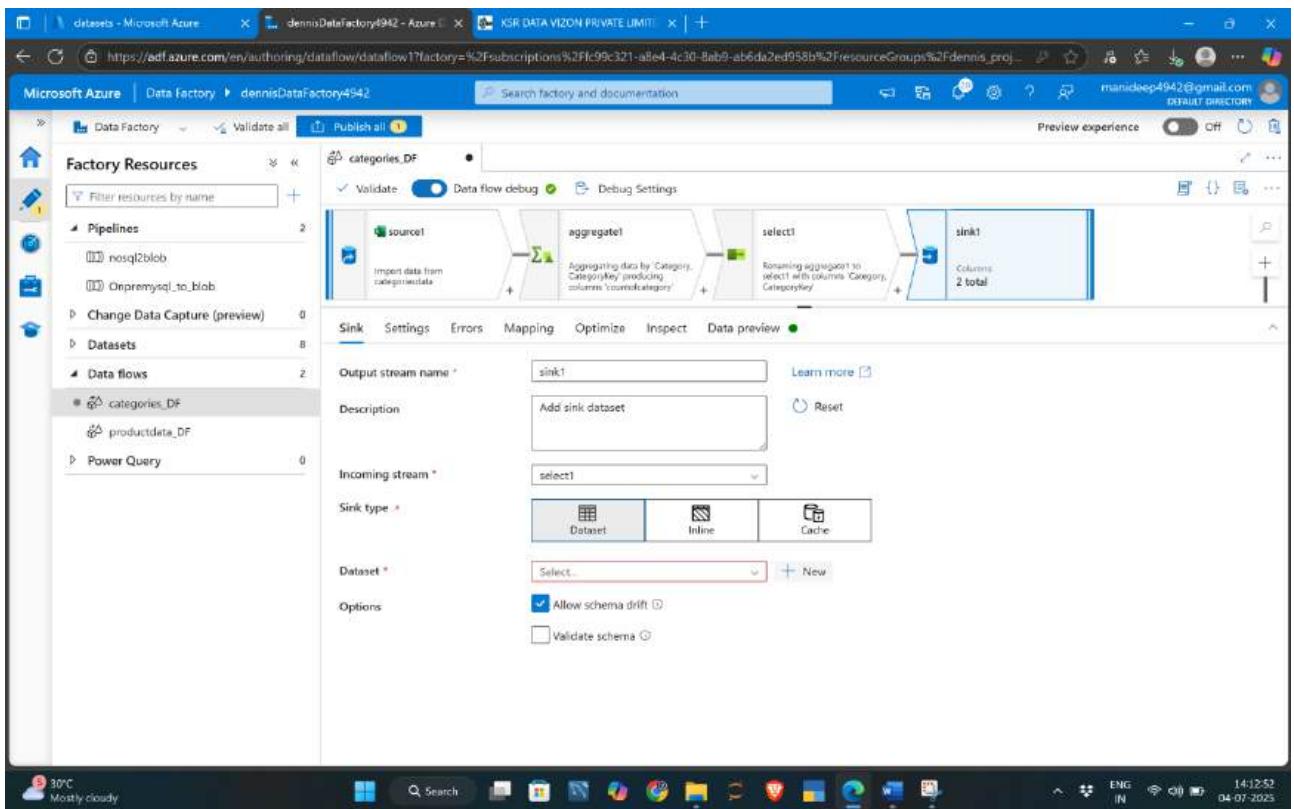
Options:

- Skip duplicate input columns
- Skip duplicate output columns

Input columns:

aggregate1's column	Name as
abc Category	Category
abc CategoryKey	CategoryKey
abc countofcategory	countofcategory

now remove the countofcategory



Microsoft Azure | Data Factory > dennisDataFactory4942

Validate all Publish all

Factory Resources

Pipelines: nosql2blob, Onpremssql_to_blob

Change Data Capture (preview): 0

Datasets: 8

Data flows: 2

categories_DF

Productdata_DF

Power Query: 0

Set properties

Name: stagingcategoriesdata

Linked service: Azureblob2Adf_ls

File path: stagingdata / Directory / File name

First row as header:

Import schema:

From connection/store From sample file None

Advanced

Output stream name: sink1

Description: Add sink dataset

Incoming stream: select1

Sink type: Dataset Inline

Dataset: Select...

Options: Allow schema drift Validate schema

OK Back Cancel

Microsoft Azure | Data Factory > dennisDataFactory4942

Validate all Publish all

Factory Resources

Pipelines: nosql2blob, Onpremssql_to_blob

Change Data Capture (preview): 0

Datasets: 9

Data flows: 2

categories_DF

Productdata_DF

Power Query: 0

Preview experience: Publishing Deployed 2 of 2 resources

categories_DF

source1

aggregate

select1

sink1

Add Source

Sink Settings Errors Mapping Optimize Inspect Data preview

Number of rows: INSERT N/A UPDATE N/A DELETE N/A UPSERT N/A LOOKUP N/A ERROR N/A TOTAL 2

Refresh Statistics Export to CSV

Category	CategoryKey
Special	1
General	2

Now publish it

Now do for SalesRep:

The screenshot shows the Microsoft Azure Data Factory Data Flow blade. On the left, the 'Factory Resources' sidebar lists Pipelines, Change Data Capture (preview), Datasets, Data flows, Power Query, and a preview of the salesrep_DF data flow. The main area displays the salesrep_DF data flow diagram. It starts with a 'source1' component connected to a 'derivedColumn1' component. The 'derivedColumn1' component has three output columns: SalesRepID, Sales Rep Name, and dimensionRefId. Below the diagram, the 'Source settings' tab is selected in the 'Data preview' section. Under 'Source type', 'Dataset' is chosen. The 'Dataset' dropdown shows 'SalesRepData' is selected. Other tabs in the 'Source settings' section include Source options, Projection, Optimize, Inspect, and Data preview.

The screenshot shows the Microsoft Azure Data Factory Data Flow blade. The interface is identical to the previous screenshot, but the 'Data preview' tab is now selected in the 'Source settings' section. This tab displays a preview of the data from the 'SalesRepData' dataset. The preview table has two columns: 'SalesRepID' and 'Sales Rep Name'. The data rows are:

SalesRepID	Sales Rep Name
ID - 6	Jan Novotny
ID - 7	John White
ID - 5	Ellen Woody
ID - 3	Mark Spancer
ID - 1	Ellie Gill
ID - 2	Bill Muray
ID - 4	EI Bob

In client doc they have give that we need to remove the ID – from this ID – 1 so it should be show only the number not the ID -

Now we are going to check with substring it will work are not?

The screenshot shows the Azure Data Factory Data Flow blade. On the left, the 'Factory Resources' sidebar lists Pipelines, Change Data Capture (preview), Datasets, and Data flows. Under Data flows, 'salesrep_DF' is selected. The main area displays a data flow diagram with a 'source1' component connected to a 'derivedColumn1' component. The 'derivedColumn1' component has three output columns: 'SalesRepID', 'Sales Rep Name', and 'derivedsales ref id'. Below the diagram, the 'Derived column's settings' tab is active, showing the expression for the 'derivedsales ref id' column: `substring(SalesRepID, 5, 6)`. A red arrow points to this expression.

The screenshot shows the same Azure Data Factory Data Flow blade, but now the 'Data preview' tab is active. The preview pane shows a table with the following data:

SalesRepID	Sales Rep Name	derivedsales ref id
ID - 6	Jai Novotny	6
ID - 7	John White	7
ID - 5	Elian Woody	5
ID - 3	Mark Spancer	3
ID - 1	Ellie Gill	1
ID - 2	Bil Muray	2
ID - 4	El Bob	4

It is looking like correct as of today but its not correct because . Now It is limited data

what if the data will increase by tomorrow . now it is in single digit ,then 2 digit,3 digit,4 digit it will get some issue

As a dataengineer we to give solution for future work also not only for now.

Lets try left and right expression

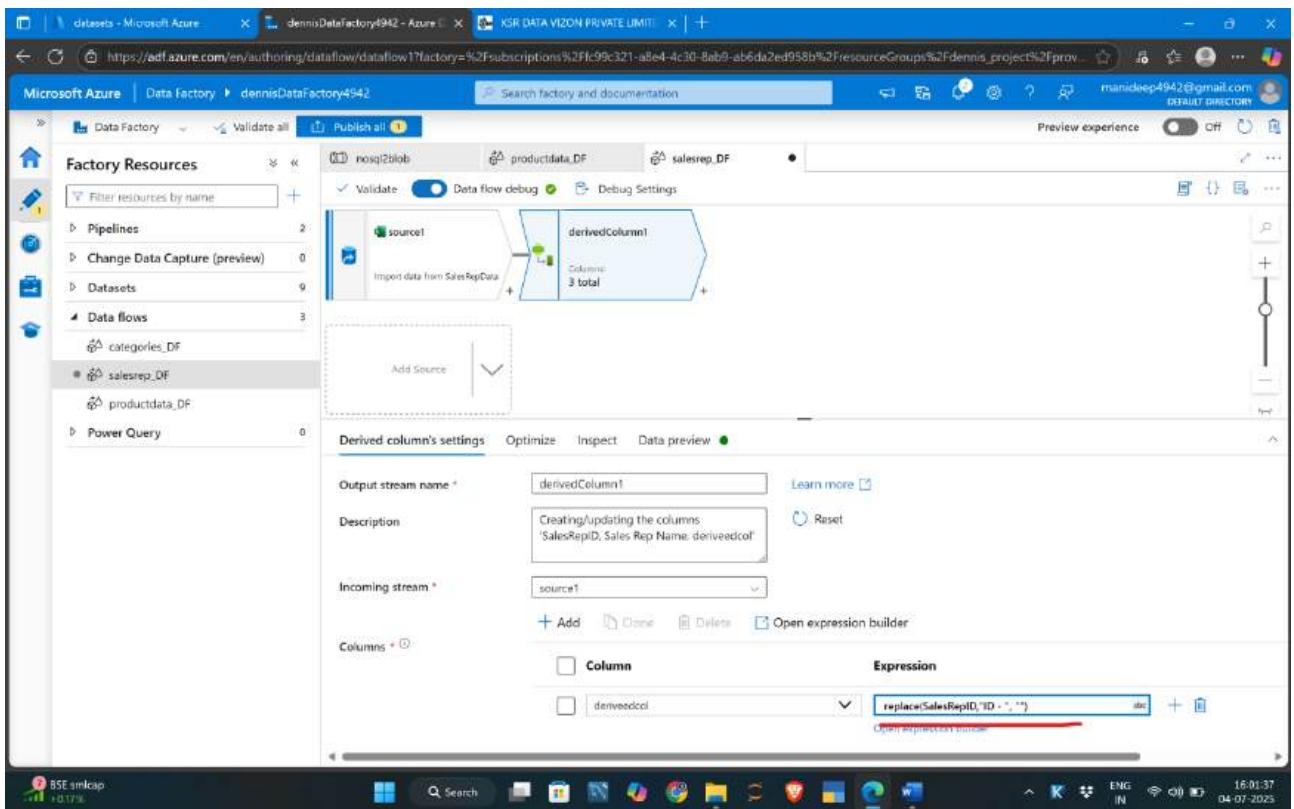
Left(SalesRepID,2) → it give ouput as ID

Right(SalesRepID,2)→it only 6 means it actually read from right

All are give right thing but we can use this every time .NO because today is 2 digit no 3 digit ,4 digit

Ok lets try with Replace expression

replace(SalesRepID,"ID","")



The screenshot shows the Microsoft Azure Data Factory Data Flow preview experience. On the left, the 'Factory Resources' sidebar lists Pipelines (2), Change Data Capture (preview) (0), Datasets (9), and Data flows (3). The 'salesrep_DF' data flow is selected. The main area displays a data flow diagram with a 'source1' (Import data from SalesRepData) connected to a 'derivedColumn1' (derivedColumn1). The 'derivedColumn1' step has a note indicating 'Columns 3 total'. Below the diagram, the 'Data preview' tab is active, showing a table with 7 rows. The columns are 'SalesRepID' (Typecasted to integer), 'Sales Rep Name' (Typecasted to string), and 'derivedcol' (Typecasted to integer). The data is as follows:

SalesRepID	Sales Rep Name	derivedcol
6	Jan Novotny	6
7	John White	7
5	Ellen Woody	5
3	Mark Spancer	3
1	Elie Gill	1
2	Bill Muray	2
4	El Bob	4

It is also the best way we just replace with null but what if we have data like

ID - 1

ID - 1

ID - 2

ID -2

Because we have disconnect spaces.....

The screenshot shows the Microsoft Azure Data Factory Data Flow blade. On the left, the 'Factory Resources' sidebar lists Pipelines, Change Data Capture (preview), Datasets, and Data flows. Under Data flows, 'salesrep_DF' is selected. The main area displays a data flow diagram with three main components: 'source1' (Import data from SalesRepData), 'derivedColumn1' (Creating/updating the columns SalesRepID, Sales Rep Name, derivedSalesRepID), and 'select1' (Renaming derivedColumn1 to select with columns SalesRepID). A tooltip for 'derivedColumn1' indicates it is renaming derivedColumn1 to select with columns SalesRepID. Below the diagram, the 'Derived column's settings' tab is active, showing the output stream name 'derivedColumn1', description 'Creating/updating the columns SalesRepID, Sales Rep Name, derivedSalesRepID', incoming stream 'source1', and a single column 'derivedSalesRepID' with the expression 'replace(SalesRepID, 'ID - ', '')'. The status bar at the bottom shows '29°C Haze' and the date '04-07-2025'.

now we use the replace for this the another data set we use the another function

Now lets sink the data into azure blob storage staging folder

The screenshot shows the Microsoft Azure Data Factory Data Flow blade. The 'Sink' tab is selected under the 'Settings' section. The 'Output stream name' is set to 'sink1', and the 'Description' is 'Add sink dataset'. The 'Incoming stream' is 'select1'. The 'Sink type' is set to 'Dataset'. The 'Dataset' dropdown is set to 'Select...'. Under 'Options', the 'Allow schema drift' checkbox is checked, and the 'Validate schema' checkbox is unchecked. To the right, a 'Select format' panel titled 'Choose the format type of your data' is open, displaying icons for Avro, DelimitedText, JSON, ORC, Parquet, and Binary. At the bottom of the screen, there are 'Continue', 'Back', and 'Cancel' buttons.

Now we are going to do subcategory data set

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. Under 'Data flows', 'subcategory_DF' is selected. The main workspace displays a data flow named 'subcategory_DF' with a single source component labeled 'source1'. Below the data flow, the 'Data preview' tab is active, showing a table with four rows of data. The columns are 'SubCategoryKey' and 'SubCategory Name'. The data is as follows:

SubCategoryKey	SubCategory Name
1	Extra
2	Regular
3	Micro
4	Super

See here the data what it is categorykey

ID - 1

ID - 1

ID - 2

ID - 2

First we use the replace(Subcategory,"ID - ","") expression

The screenshot shows the Microsoft Azure Data Factory Data Flow blade. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. Under 'Data flows', 'subcategory_DF' is selected. The main area displays a data flow diagram with a 'source1' component (Import data from SubCategoryData) connected to a 'derivedColumn1' component. The 'derivedColumn1' component has four columns: 'SubCategoryKey', 'CategoryKey', 'SubCategory Name', and 'derviesubcategory'. Below the diagram, the 'Derived column's settings' tab is active, showing the output stream name 'derivedColumn1', a description 'Creating/updating the columns 'SubCategoryKey, CategoryKey, SubCategory Name, derviesubcategory'', and the incoming stream 'source1'. The 'Expression' section shows the formula `replace(categoryKey, 'ID - ', '')`. The status bar at the bottom indicates '16:31:05 04-07-2023'.

The screenshot shows the 'Data preview' tab for the 'subcategory_DF' data flow. The preview table shows the following data:

	SubCategoryKey	CategoryKey	SubCategory Name	derviesubcategory
1	1	ID-1	Extra	ID-1
2	2	ID-2	Regular	2
3	3	ID-1	Micro	1
4	4	ID-2	Super	2

Now we use the `regexreplace()` expression

Isted of replace.....

We use `regexreplace(categoryKey"[^0-9]","")`

The screenshot shows the Microsoft Azure Data Factory Data Flow interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. The 'Data flows' section is expanded, showing four data flows: categories_DF, subcategory_DF, productdata_DF, and salesrep_DF. The 'subcategory_DF' data flow is selected and displayed in the main editor area. The data flow consists of a single step: 'source1' followed by a 'derivedColumn1' step. The 'derivedColumn1' step has one input column 'derviesubcategory' and one output column 'SubCategoryName'. The 'derivedColumn1' settings show the expression: `regexReplace(CategoryKey, '[^0-9]', '')`. The 'Data preview' tab is selected, showing the following data:

SubCategoryKey	CategoryKey	SubCategory Name	derviesubcategory
1	ID-1	Extra	1
2	ID-2	Regular	2
3	ID-1	Micro	1
4	ID-2	Super	2

This screenshot is identical to the one above, showing the Microsoft Azure Data Factory Data Flow interface for the 'subcategory_DF' data flow. The data preview tab displays the same data as before:

SubCategoryKey	CategoryKey	SubCategory Name	derviesubcategory
1	ID-1	Extra	1
2	ID-2	Regular	2
3	ID-1	Micro	1
4	ID-2	Super	2

Now it is worked right now....

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. Under 'Data flows', 'subcategory_DF' is selected. The main area displays the 'subcategory_DF' data flow diagram. The flow consists of a 'source1' (Import data from SubCategoryData) connected to a 'derivedColumn1' (Creating/updating the columns: SubCategoryKey, CategoryKey, SubCategoryName, derivedSubCategory) which then connects to a 'select1' (Columns: 2 total). Below the diagram, the 'Data preview' tab is active, showing the 'Select settings' panel with 'Output stream name' set to 'select1'. The 'Description' field contains the note: 'Renaming derivedColumn1 to select1 with columns 'CategoryKey, SubCategory Name''. The 'Incoming stream' is set to 'derivedColumn1'. Under 'Options', 'Skip duplicate input columns' and 'Skip duplicate output columns' are checked. The 'Input columns' section shows three unmapped input columns: 'derivedColumn1's column, 'CategoryKey', and 'SubCategory Name', each mapped to their respective output columns: 'CategoryKey' and 'SubCategory Name'. A note at the bottom right indicates '3 mappings: 2 column(s) from the inputs left unmapped'.

I replaced the new one with the original one

The screenshot shows the Microsoft Azure Data Factory interface, identical to the previous one but with a key difference: the 'derivedColumn1' component has been removed from the data flow diagram. The remaining components are 'source1', 'select1', and 'sink1'. The 'Data preview' tab is still active, showing the same 'Select settings' panel and unmapped input columns. The 'Data preview' table now displays the following data:

SubCategoryKey	CategoryKey	SubCategory Name
1	1	Extra
2	2	Regular
3	1	Micro
4	2	Super

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

All	Azure	Database	File	Generic protocol

[Continue](#) [Cancel](#)

Publish all

You are about to publish all pending changes to the live environment. [Learn more](#)

Pending changes (2)

NAME	CHANGE	EXISTING
stagingsubcategorydata	(New)	-
subcategory_DF	(New)	-

Number of rows: **INSERT** N/A **UPDATE** N/A

[Refresh](#) [Statistics](#) [Export to CSV](#)

SubCategoryKey	CategoryKey
1	1
2	2
3	1
4	2

[Publish](#) [Cancel](#)

Now geography data :

Microsoft Azure | Data Factory | dennisDataFactory4542 | Search factory and documentation

Preview experience: Off

Factory Resources

- Pipelines (2)
- Onpremises_to_blob
- Onpremises_to_blob
- Change Data Capture (preview) (0)
- Datasets (12)
- categoriesdata
- Geography
- ProductData
- Raw_productdata
- RawGeographydata
- SalesR RawGeographydata
- stagingcategoriesdata
- staginggeographydata
- stagingproductdata
- stagingsalesrepdata
- stagingsubcategorydata
- SubCategoryData
- Data flows (5)
- categories_DF

geographydata_DF

Validate: Data flow debug, Debug Settings

Source: Import data from RawGeographydata

select1: Renaming source1 to select with columns: 'Country', 'Town'

Sink: Columns: 2 total

Add Source

Sink Output stream name: sink1

Description: Export data to staginggeographydata

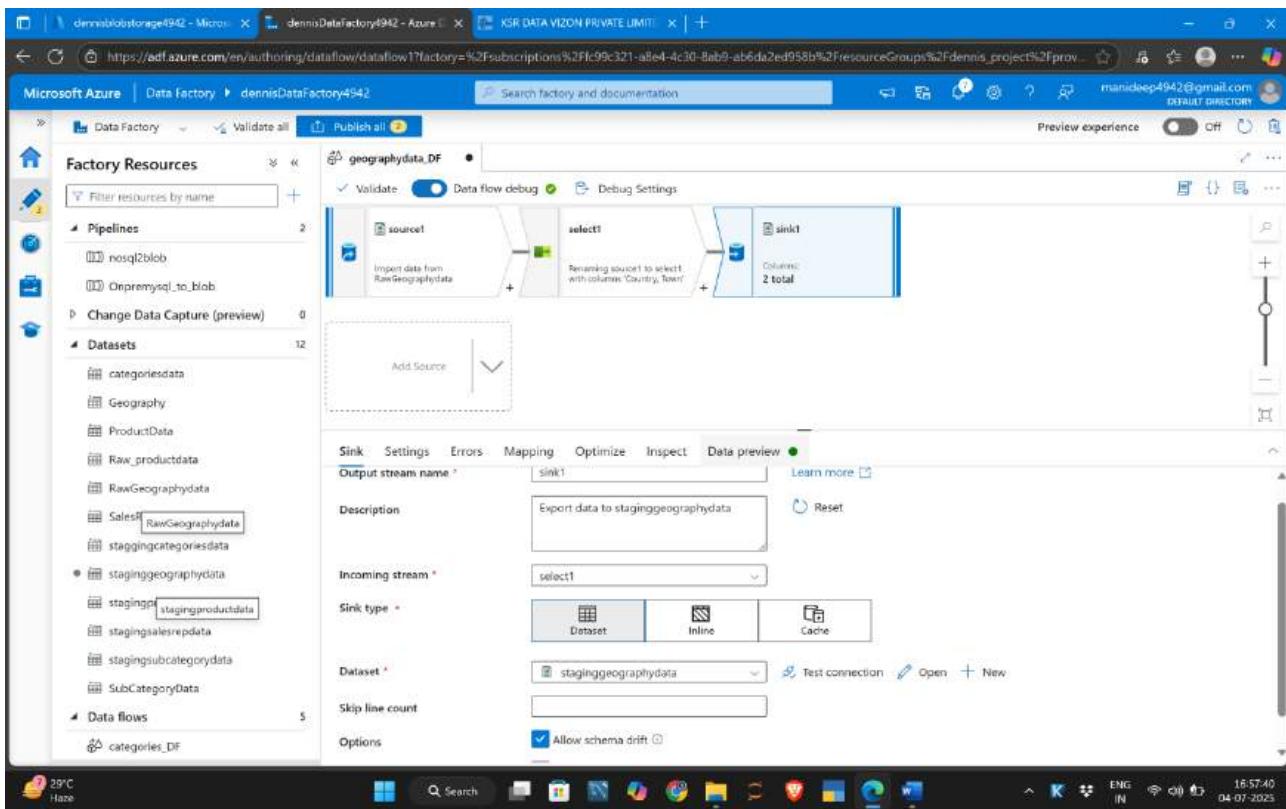
Incoming stream: select1

Sink type: Dataset

Dataset: staginggeographydata

Skip line count:

Options: Allow schema drift



Microsoft Azure | Data Factory | dennisDataFactory4542 | Search factory and documentation

Preview experience: Off

Factory Resources

- Pipelines (2)
- Onpremises_to_blob
- Onpremises_to_blob
- Change Data Capture (preview) (0)
- Datasets (12)
- categoriesdata
- Geography
- ProductData
- Raw_productdata
- RawGeographydata
- SalesR RawGeographydata
- stagingcategoriesdata
- staginggeographydata
- stagingproductdata
- stagingsalesrepdata
- stagingsubcategorydata
- SubCategoryData
- Data flows (5)
- categories_DF

geographydata_DF

Validate: Data flow debug, Debug Settings

Source: Import data from RawGeographydata

select1: Renaming source1 to select with columns: 'Country', 'Town'

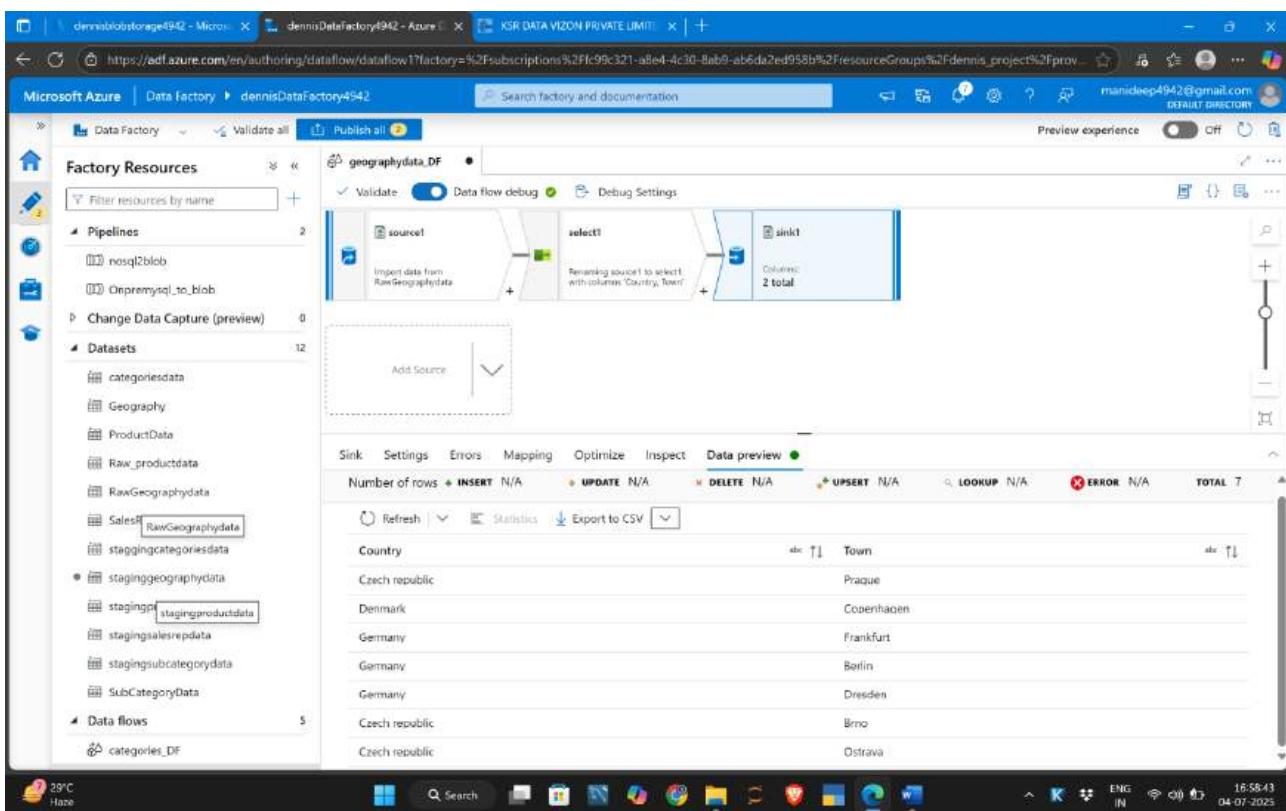
Sink: Columns: 2 total

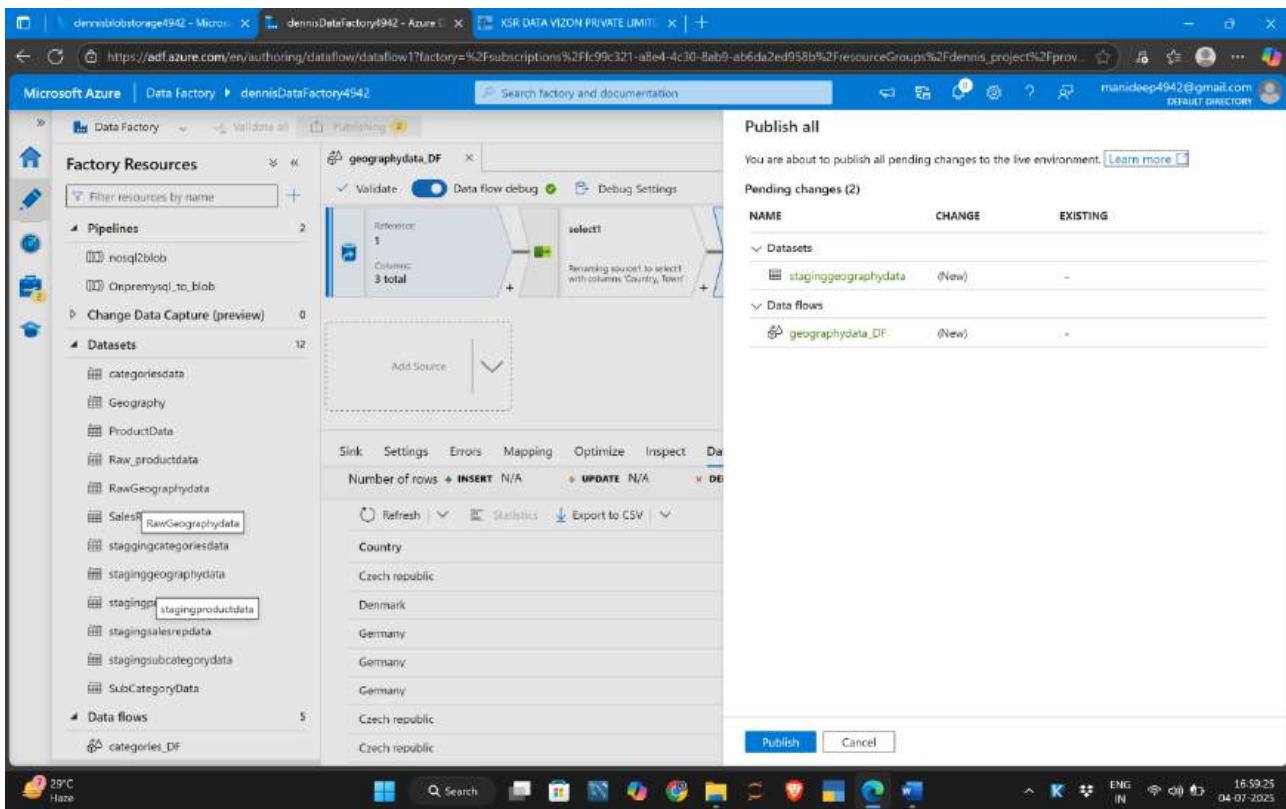
Add Source

Data preview

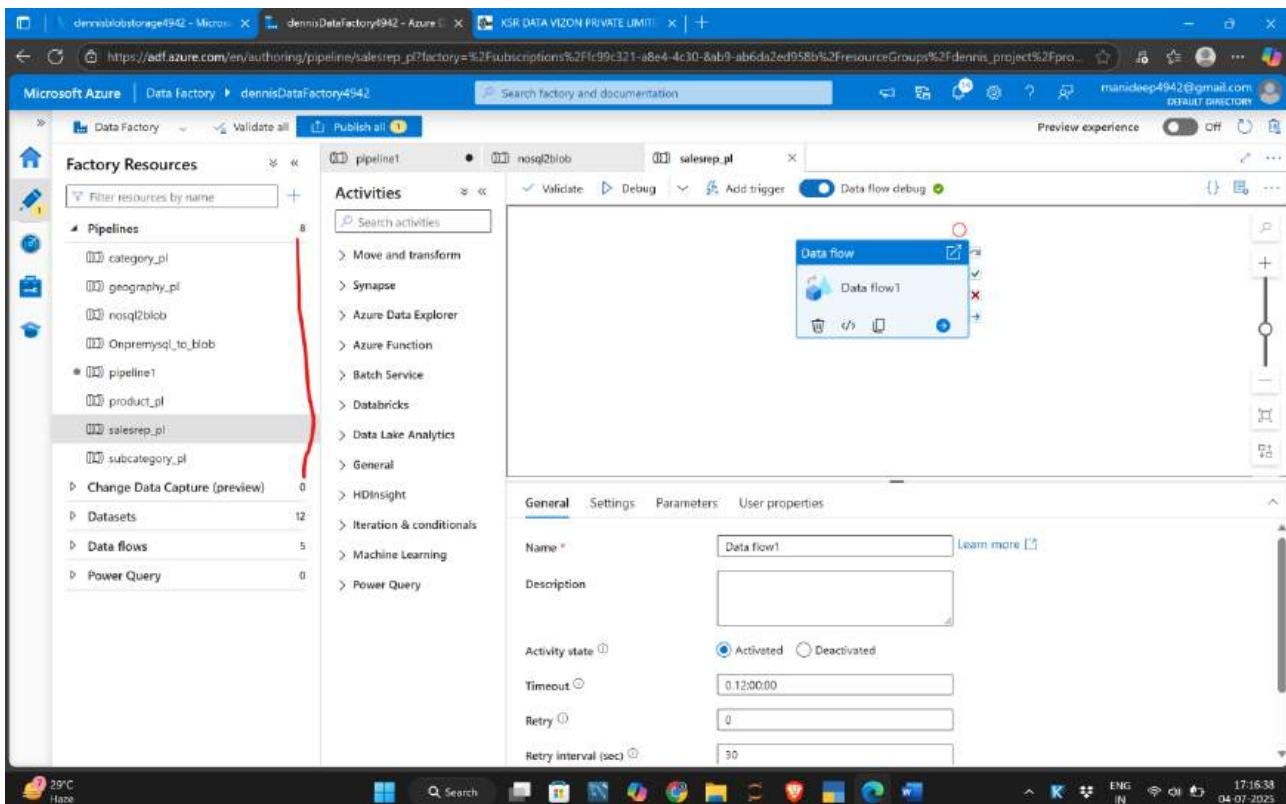
Number of Rows	INSERT	N/A	UPDATE	N/A	DELETE	N/A	UPSERT	N/A	LOOKUP	N/A	ERROR	N/A	TOTAL
7													
	Refresh	Statistics	Export to CSV										

Country	Town
Czech republic	Prague
Denmark	Copenhagen
Germany	Frankfurt
Germany	Berlin
Germany	Dresden
Czech republic	Brno
Czech republic	Ostrava



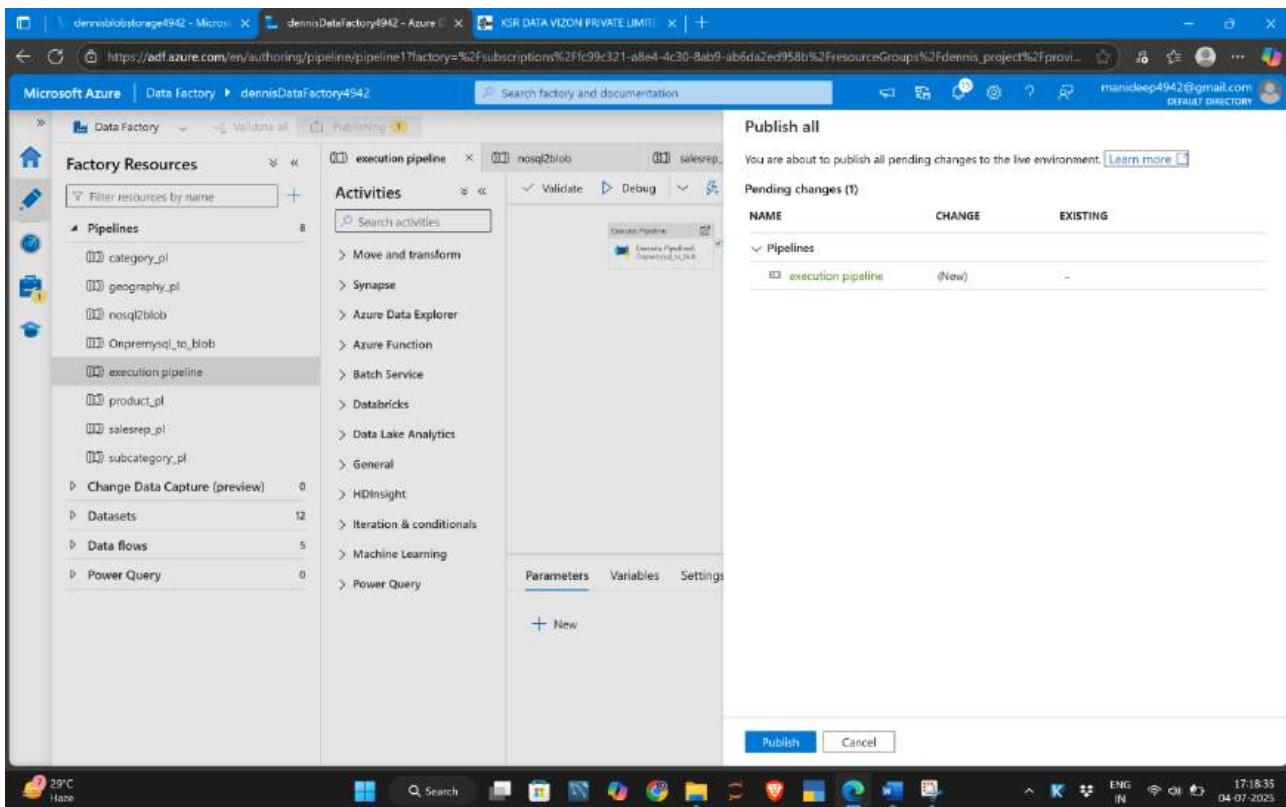


Now create the pipeline for every data flow what ever you have created here this is the photo

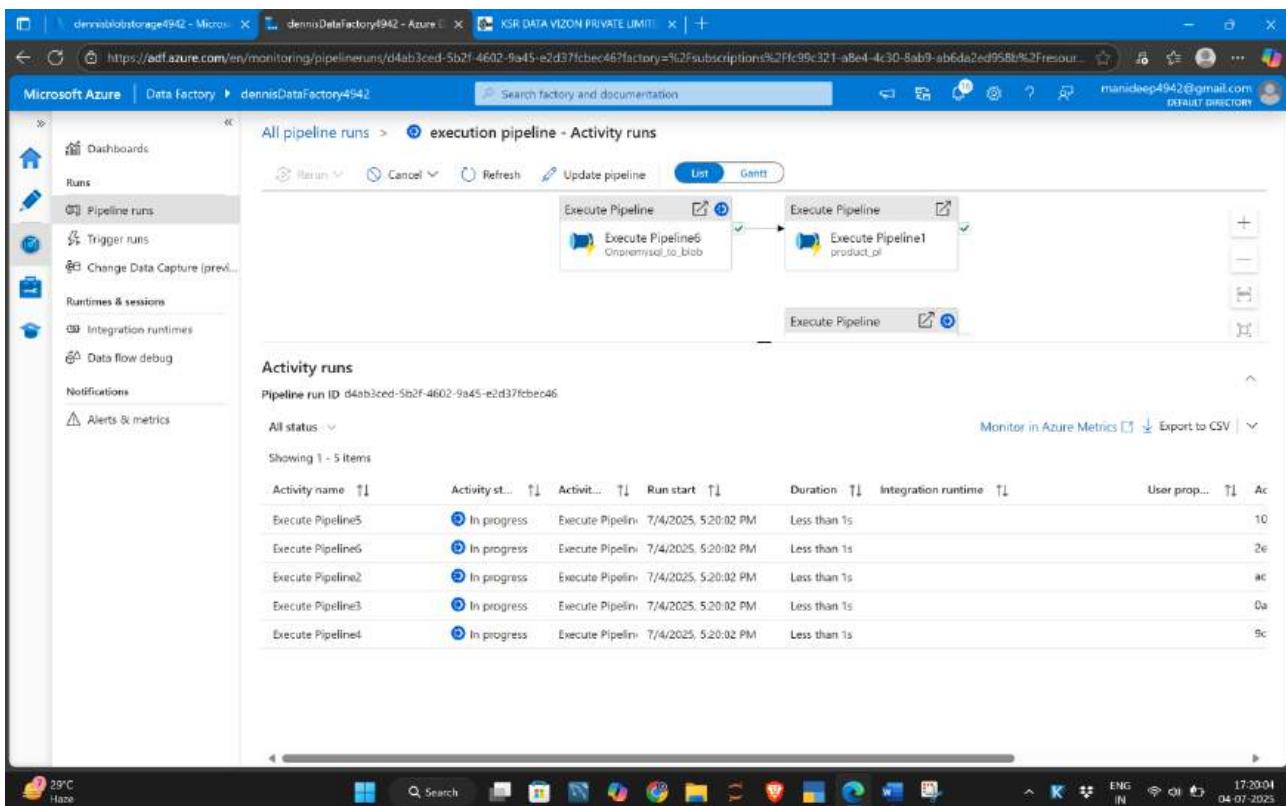


Use execution pipeline to run all the pipelines at a time .

Do the publish



Now trigger now the execution pipeline to execute all the pipelines.



Here the success pipeline

The screenshot shows the Microsoft Azure Data Factory interface. On the left, a sidebar lists navigation options: Dashboards, Runs, Pipeline runs (selected), Trigger runs, Change Data Capture (preview), Runtimes & sessions, Integration runtimes, Data flow debug, Notifications, and Alerts & metrics. The main content area is titled "All pipeline runs > execution pipeline - Activity runs". It displays a summary card for "Execute Pipeline4" with a status of "Succeeded". Below this, a table lists "Activity runs" with the following data:

Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime	User prop...	Ac
Execute Pipeline1	Succeeded	Execute Pipeline1	7/4/2025, 5:29:37 PM	3m 5s			c8
Execute Pipeline3	Succeeded	Execute Pipeline3	7/4/2025, 5:29:15 PM	3m 9s			ab
Execute Pipeline5	Succeeded	Execute Pipeline5	7/4/2025, 5:29:15 PM	3m 6s			81
Execute Pipeline4	Succeeded	Execute Pipeline4	7/4/2025, 5:29:15 PM	3m 11s			36
Execute Pipeline6	Succeeded	Execute Pipeline6	7/4/2025, 5:29:15 PM	22s			39
Execute Pipeline2	Succeeded	Execute Pipeline2	7/4/2025, 5:29:15 PM	3m 32s			a7

These are the files which have created in blob storage :

The screenshot shows the Microsoft Azure Storage Container overview for "stagingdata". The left sidebar includes "Overview", "Diagnose and solve problems", "Access Control (IAM)", and "Settings". The main area displays a table of blobs with the following data:

Name	Last modified	Access tier	Blob type	Size	Lease state
_SUCCESS	4/7/2025, 5:32:39 pm	Hot (Inferred)	Block blob	0	Available
par-00000-3bd98e8c-8559-48e5-b724-012...	4/7/2025, 5:23:10 pm	Hot (Inferred)	Block blob	146 B	Available
par-00000-4726e766d-beed-482f-ab62-727...	4/7/2025, 5:32:34 pm	Hot (Inferred)	Block blob	395 B	Available
par-00000-4d193f72-90f8-4a28-80f5-793d...	4/7/2025, 5:32:16 pm	Hot (Inferred)	Block blob	117 B	Available
par-00000-53bb5212-4c7a-41fb-b38f-c2d3...	4/7/2025, 5:23:07 pm	Hot (Inferred)	Block blob	117 B	Available
par-00000-593aa2ca-257b-435d-a314-031...	4/7/2025, 5:23:15 pm	Hot (Inferred)	Block blob	86 B	Available
part-00000-6c9cf080-03ad-4038-b4fa-4bf1...	4/7/2025, 5:32:39 pm	Hot (Inferred)	Block blob	41 B	Available
par-00000-81286d92-1ef1a-4cb9-81a0-ab9...	4/7/2025, 5:32:19 pm	Hot (Inferred)	Block blob	86 B	Available
par-00000-b91092fc-6120-460f-8480-9949...	4/7/2025, 5:23:04 pm	Hot (Inferred)	Block blob	41 B	Available
par-00000-cbb0589a-0919-474d-a5eb-765...	4/7/2025, 5:23:55 pm	Hot (Inferred)	Block blob	395 B	Available
par-00000-d56dd91a-fcc3-470a-a307-cb8...	4/7/2025, 5:32:14 pm	Hot (Inferred)	Block blob	146 B	Available

We do what should do next.

Ok now we are going to see the sales data from the ADLS storage

And also need to create the staging data folder to store the correction data.

The screenshot shows the Microsoft Azure Storage account interface for the 'dennisadlsstorage4942' account. The left sidebar is expanded to show 'Data storage' and 'Containers'. Under 'Containers', there are three items listed:

Name	Last modified	Anonymous access level	Lease state
Slogs	4/7/2025, 11:17:31 am	Private	Available
salesdatasets	4/7/2025, 11:23:36 am	Private	Available
stagingdata	5/7/2025, 10:16:28 am	Private	Available

Here u see we have created the stagingdata....folder.

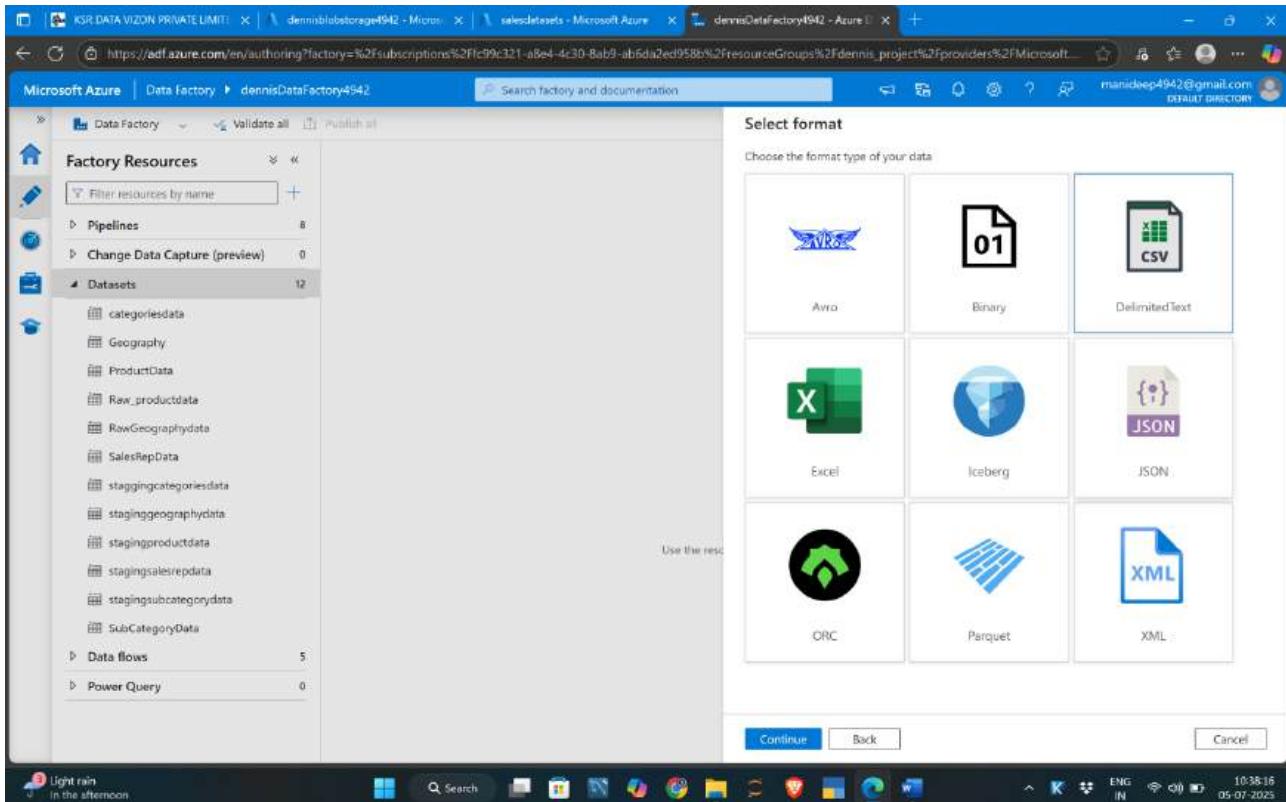
Now we need to create the dataset for azure data lake.

The screenshot shows the Microsoft Azure Data Factory interface under the 'dennisDataFactory4942' factory. On the left, the 'Factory Resources' sidebar is expanded to show 'Datasets' (12 items) and 'Data flows' (5 items). The main area is titled 'New dataset' and displays a grid of data store icons:

All	Azure	Database	File	Generic protocol
Azure AI Search	Azure Blob Storage	Azure Cosmos DB for MongoDB		
Azure Cosmos DB for NoSQL	Azure Data Explorer (Kusto)	Azure Data Lake Storage Gen2		
My	My	Azure		

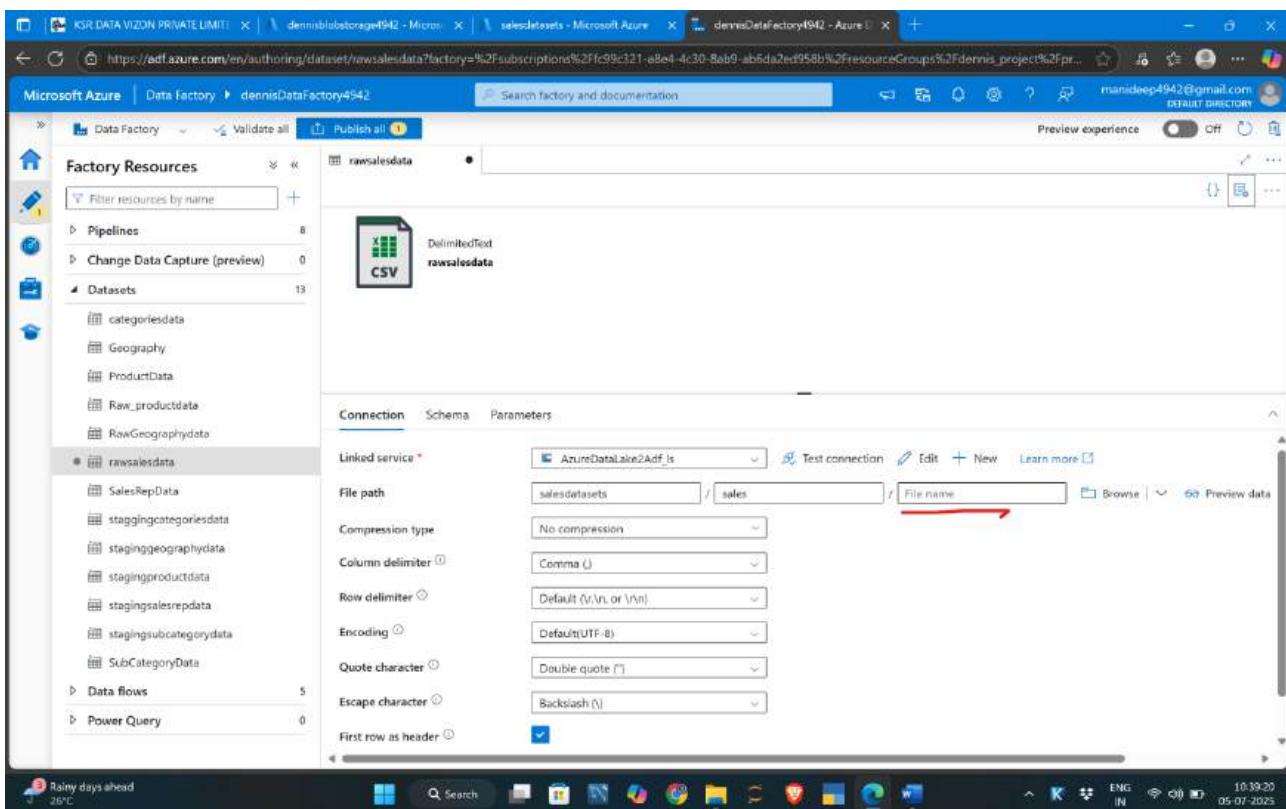
Below the grid are 'Continue' and 'Cancel' buttons.

All files are in csv format



The screenshot shows the 'Select format' dialog in Microsoft Azure Data Factory. On the left, the 'Factory Resources' sidebar lists various datasets and data flows. In the center, a grid of icons represents different file formats: Avro, Binary, DelimitedText, Excel, Iceberg, JSON, ORC, Parquet, and XML. Below the grid, there are 'Continue', 'Back', and 'Cancel' buttons. The status bar at the bottom indicates 'Light rain in the afternoon' and shows the date as 05-07-2023.

We are taking the all file because all file structures are same , and all columns and data as been that what we are going to change the data In all files at a time.



The screenshot shows the configuration for the 'rawsalesdata' dataset in Microsoft Azure Data Factory. The 'Connection' tab is selected, showing the 'Linked service' set to 'AzureDataLake2Adf_ls'. The 'File path' field contains 'salesdatasets / sales /' with a red arrow pointing to the 'File name' input field, which is currently empty. Other settings include 'No compression' for compression type, 'Comma (,) for column delimiter, 'Default (\r\n or \n\r)' for row delimiter, 'Default(UTF-8)' for encoding, 'Double quote (")' for quote character, 'Backslash (\)' for escape character, and 'First row as header' checked. The status bar at the bottom indicates 'Rainy days ahead' and shows the date as 05-07-2023.

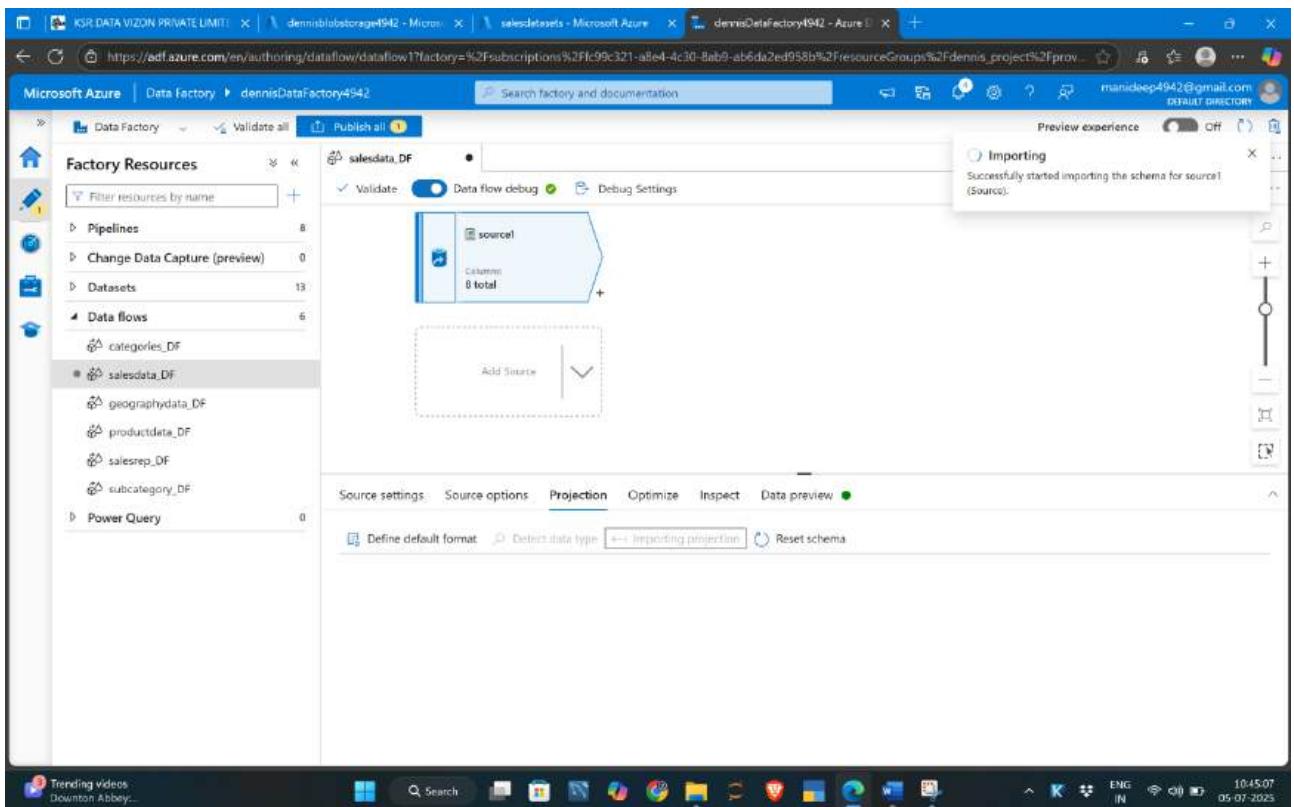
Here it is the previews data.

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists 'Pipelines', 'Change Data Capture (preview)', 'Datasets', and 'Data flows'. Under 'Datasets', 'rawsalesdata' is selected, and its preview is displayed in a large central window. The preview title is 'Linked service: AzureDataLake2Adl_ls'. The preview pane shows a table with columns: fSalesPrimaryKey, ProductID, SalesRepID, Location, Date, Units, and PercentOfStandardCo. The data consists of 10 rows of sales records from February 2014. The bottom of the preview pane has settings for 'Escape character' (Backslash \) and 'First row as header' (checked). The status bar at the bottom right shows the date as 05-07-2025.

Now create a dataflow to this sales data.

The screenshot shows the Microsoft Azure Data Factory interface. The 'Data flows' section in the sidebar has 5 items: categories_DF, geographydata_DF, productdata_DF, salesrep_DF, and subcategory_DF. A context menu is open over the first item, 'categories_DF', with options: 'New data flow' (selected), 'New flowlet', and 'New folder'. To the right of the menu, there is a placeholder area with a cylinder icon and the text 'Select an item' and 'Use the resource explorer to select or create a new item'. The status bar at the bottom right shows the date as 05-07-2025.

On debug mode tho see the data preview.



Import the projections

This screenshot shows the 'Projection' tab of the Data Flow blade for 'salesdata_DF'. The 'Projection' tab is active, indicated by a blue underline. Below it, the 'Source settings', 'Source options', 'Optimize', 'Inspect', and 'Data preview' tabs are visible. The 'Projection' section contains a table with columns for 'Column name', 'Type', and 'Format'. The columns listed are:

Column name	Type	Format
IsSalesPrimaryKey	integer	Specify format
ProductID	short	Specify format
SalesRepID	short	Specify format
Location	string	Specify format
Date	string	Specify format
Units	short	Specify format
PercentOfStandardCost	double	Specify format
RevenueDiscount	double	Specify format

The 'Import projection' button is located just above the table. The rest of the interface is identical to the previous screenshot, showing the 'source' component and the 'Importing' status message.

The screenshot shows the Microsoft Azure Data Factory Data Flow interface. On the left, the 'Factory Resources' sidebar lists various data flows, with 'salesdata_DF' selected. The main workspace displays a data flow diagram with a 'source' block followed by a 'derivedColumn1' block. The 'derivedColumn1' block has an output stream named 'derivedColumn1'. Below the diagram, a 'Data preview' section shows a table with 15 rows of data from the source. The columns are: ProductID, SalesRepID, Location, Date, Units, PercentOfStandardCost, and RevenueDiscount. The 'Location' column contains values like 'Germany;Frankfurt', 'Denmark;Copenhagen', 'Germany;Dresden', etc.

here it is the data preview in location col we have state and country so we need to separate it

So use derive column

The screenshot shows the 'Derived column's settings' configuration pane for the 'derivedColumn1' step. Under 'Output stream name', the value is 'derivedColumn1'. In the 'Description' field, it says 'Creating/updating the columns: [Sales.PrimaryKey, ProductID, SalesRepID, Location, Date, Units]'. Under 'Incoming stream', the value is 'source1'. The 'Columns' section is expanded, showing a table with two columns: 'Column' and 'Expression'. There are six rows, each defining a new column based on an incoming column: 'Sales.PrimaryKey' is mapped to 'Sales.PrimaryKey', 'ProductID' is mapped to 'ProductID', 'SalesRepID' is mapped to 'SalesRepID', 'Location' is mapped to 'Location', 'Date' is mapped to 'Date', and 'Units' is mapped to 'Units'.

Lets check this condition it is separating are not .

Dataflow expression builder

Column name: location

Expression: regexExtract(Location, ".*")

Expression elements:

- All
- Functions
- Input schema
- Parameters
- Cached lookup
- Data flow library functions
- Locals

Expression values:

- SalesRepID
- Location
- Date
- Units
- PercentOfStandardCost
- PercentOfStandardCost

Data preview:

tID	SalesRepID	Location	Date	Units	PercentOfStandardCost	RevenueDiscount	location
7	Germany:Frankfurt	1.10.2017	60	0.99	0.35		
2	Denmark:Copenhagen	26.4.2017	71	0.972	0.35		
6	Germany:Dresden	28.9.2017	68	0.968	0.35		
5	Germany:Berlin	16.4.2017	157	0.97	0.55		
2	Denmark:Copenhagen	13.2.2017	103	0.965	0.5		
1	Czech republic:Prague	3.12.2017	48	0.997	0.25		
2	Denmark:Copenhagen	26.8.2017	30	0.984	0.25		
7	Germany:Frankfurt	16.4.2017	57	0.97	0.35		
4	Germany:Frankfurt	3.11.2017	145	0.994	0.5		
5	Germany:Berlin	1.1.2017	31	0.954	0.25		

No its not giving anything

Factory Resources

salesdata_DF

Derived column's settings

tID	SalesRepID	Location	Date	Units	PercentOfStandardCost	RevenueDiscount	location
7	Germany:Frankfurt	1.10.2017	60	0.99	0.35		
2	Denmark:Copenhagen	26.4.2017	71	0.972	0.35		
6	Germany:Dresden	28.9.2017	68	0.968	0.35		
5	Germany:Berlin	16.4.2017	157	0.97	0.55		
2	Denmark:Copenhagen	13.2.2017	103	0.965	0.5		
1	Czech republic:Prague	3.12.2017	48	0.997	0.25		
2	Denmark:Copenhagen	26.8.2017	30	0.984	0.25		
7	Germany:Frankfurt	16.4.2017	57	0.97	0.35		
4	Germany:Frankfurt	3.11.2017	145	0.994	0.5		
5	Germany:Berlin	1.1.2017	31	0.954	0.25		

We try another expression.

Microsoft Azure | Data Factory > dennisDataFactory4942 - Azure | +

Search factory and documentation

Dataflow expression builder

derivedColumn1

Derived Columns

+ Create new

location

Column name *

location

Expression

regexSplit(Location, ";")

Save

Expression elements

All

Functions

Input schema

Parameters

Cached lookup

Data flow library functions

Locals

Expression values

Filter by keyword

SalesPrimarykey

ProductID

SalesRepID

Location

Date

Data preview

Refresh

Save and finish

Cancel

Clear contents

Getting some thing by this.

Microsoft Azure | Data Factory > dennisDataFactory4942 - Azure | +

Search factory and documentation

Validate all

Publish all

Preview experience

Factory Resources

Filter resources by name

Pipelines

Change Data Capture (preview)

Datasets

Data flows

categories_DF

salesdata_DF

geographydata_DF

productdata_DF

salesrep_DF

subcategory_DF

Power Query

salesdata_DF

source

derivedColumn1

Import data from rawtestdata

Derived column's settings

Optimize

Inspect

Data preview

UPDATE 0

DELETE 0

UPSERT 0

LOOKUP 0

ERROR 0

TOTAL: 1000

Merge

Map shifted

Statistics

Remove

Export to CSV

tID	SalesRepID	Location	Date	Units	PercentOfStandardCost	RevenueDiscount	location
7		Germany:Frankfurt	1.10.2017	60	0.99	0.35	Germany:Frankfurt
2		Denmark:Copenhagen	26.4.2017	71	0.972	0.25	Denmark:Copenhagen
6		Germany:Dresden	28.9.2017	68	0.988	0.35	Germany:Dresden
5		Germany:Berlin	16.4.2017	157	0.97	0.25	Germany:Berlin
2		Denmark:Copenhagen	13.2.2017	103	0.965	0.25	Denmark:Copenhagen
1		Czech republic:Prague	3.12.2017	48	0.997	0.25	Czech republic:Prague
2		Denmark:Copenhagen	26.8.2017	30	0.984	0.25	Denmark:Copenhagen
7		Germany:Frankfurt	16.4.2017	57	0.97	0.35	Germany:Frankfurt
4		Germany:Frankfurt	3.11.2017	145	0.994	0.5	Germany:Frankfurt
6		Germany:Berlin	1.1.2017	31	0.954	0.25	Germany:Berlin

We do some more modification in this data .

Dataflow expression builder

Column name: location

Expression: `regexSplit(Location, ";")[1]`

Expression elements	Expression values
All	<input type="text" value="Filter by keyword"/>
Functions	+ Create new
Input schema	↳ SalesPrimarykey
Parameters	↳ ProductID
Cached lookup	↳ SalesRepID
Data flow library functions	↳ Location
Locals	↳ Date

Data preview: Refresh

Save and finish Cancel Clear contents

Its worked :

Factory Resources

- Pipelines: 8
- Change Data Capture (preview): 0
- Datasets: 13
- Data flows: 6
 - categories_DF
 - salesdata_DF
 - geographydata_DF
 - productdata_DF
 - salesrep_DF
 - subcategory_DF
- Power Query: 0

salesdata_DF

Validate: ✓ Data flow debug: Debug Settings:

Derived column's settings: Optimize: Inspect: Data preview:

ID	SalesRepID	Location	Date	Units	PercentOfStandardCost	RevenueDiscount	location
7	Germany;Frankfurt	1.10.2017	60	0.99	0.35		Germany
2	Denmark;Copenhagen	26.4.2017	71	0.972	0.35		Denmark
6	Germany;Dresden	28.9.2017	68	0.988	0.35		Germany
5	Germany;Berlin	16.4.2017	157	0.97	0.55		Germany
2	Denmark;Copenhagen	13.2.2017	103	0.965	0.5		Denmark
1	Czech republic;Prague	3.12.2017	48	0.997	0.25		Czech republic
2	Denmark;Copenhagen	26.8.2017	30	0.984	0.25		Denmark
7	Germany;Frankfurt	16.4.2017	57	0.97	0.35		Germany
4	Germany;Frankfurt	3.11.2017	145	0.994	0.5		Germany
6	Germany;Berlin	1.1.2017	31	0.954	0.25		Germany

We name it as country now and also do the state separate now.

The screenshot shows the Microsoft Azure Data Factory Data Flow blade. On the left, the 'Factory Resources' sidebar lists Pipelines, Change Data Capture (preview), Datasets, and Data flows. Under Data flows, 'salesdata_DF' is selected. The main area displays a data flow diagram with a 'source' block connected to a 'derivedColumn1' block. The 'derivedColumn1' block has a tooltip indicating 'Columns: 10 total'. Below the diagram, the 'Derived column's settings' tab is active, showing the output stream name 'derivedColumn1', a description 'Creating/updating the columns /SalesPrimaryKey, ProductID, SalesRepID, Location, Date, Units', and the incoming stream 'source1'. The 'Columns' section contains two columns: 'Country' with the expression `regexSplit(Location, ',')[0]` and 'State' with the expression `regexSplit(Location, ',')[2]`. The status bar at the bottom shows '26°C Partly sunny' and the date '05-07-2025'.

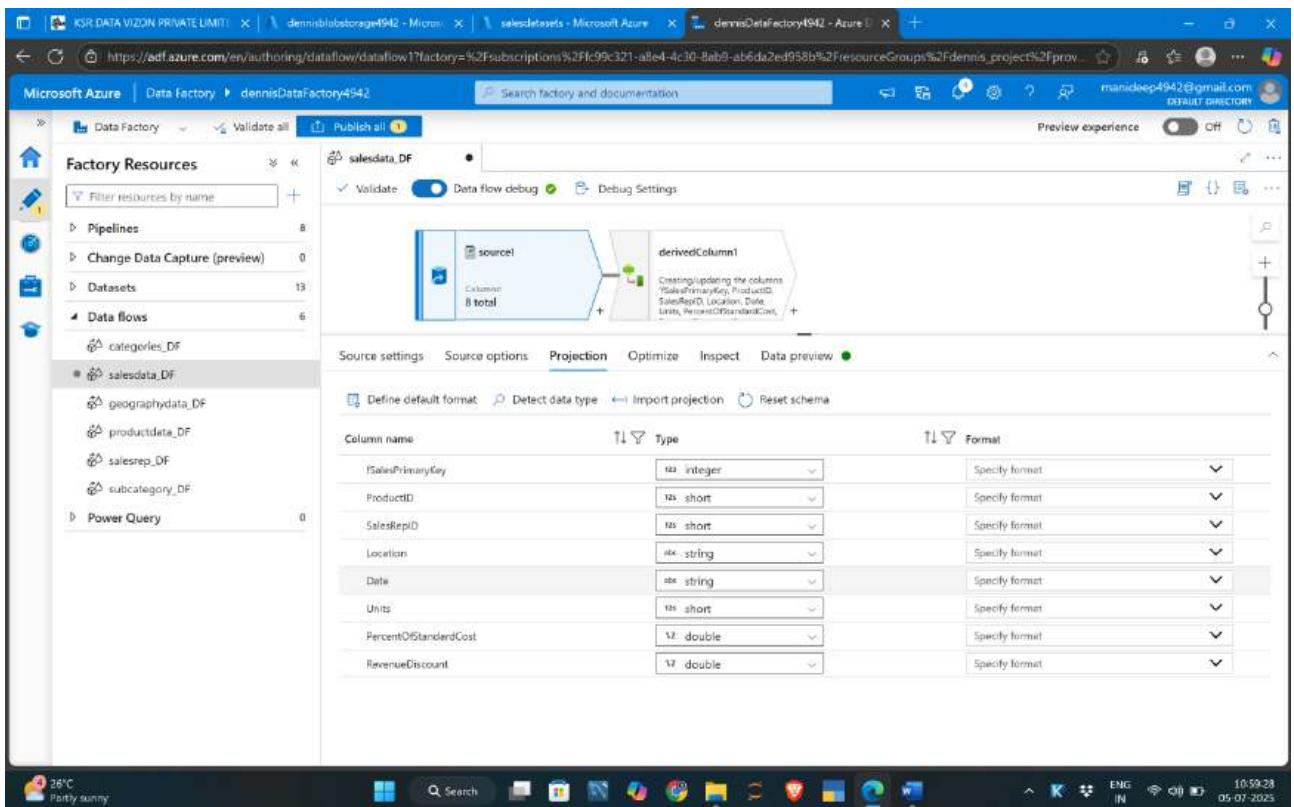
The screenshot shows the Microsoft Azure Data Factory Data Flow blade with the 'Data preview' tab selected. The preview pane displays a table with the following data:

SalesRepID	Location	Date	Units	PercentOfStandardCost	RevenueDiscount	Country	State
Germany;Frankfurt	1.10.2017	60	0.99	0.35	Germany	Frankfurt	
Germany;Dresden	28.9.2017	68	0.988	0.35	Germany	Dresden	
Germany;Berlin	16.4.2017	157	0.97	0.55	Germany	Berlin	
Germany;Frankfurt	16.4.2017	57	0.97	0.35	Germany	Frankfurt	
Germany;Frankfurt	3.11.2017	145	0.994	0.5	Germany	Frankfurt	
Germany;Berlin	1.1.2017	31	0.954	0.25	Germany	Berlin	
Germany;Dresden	4.3.2017	87	0.964	0.4	Germany	Dresden	
Germany;Berlin	7.6.2017	76	0.977	0.4	Germany	Berlin	
Germany;Berlin	29.11.2017	71	0.998	0.35	Germany	Berlin	
Germany;Berlin	13.9.2017	39	0.98	0.25	Germany	Berlin	

The status bar at the bottom shows '26°C Partly sunny' and the date '05-07-2025'.

Ok now we also have an date issue it is not in the right format we need to get it in write formate.

See actually the date is in the string formate.



We have tried these functions but not worked here it is

`formatDateTime(toTimeStamp(Date,"dd.m.yyyy"),"yyyy-mm-dd")`

`toTimeStamp(Data,"yyyy-mm-dd")` it is coming null

late we known that it is in string formate

`toDate(toTimeStamp(Date,"dd-m-yyyy"))`

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists various datasets and data flows, with 'salesdata_DF' selected. The main workspace displays a data flow diagram and its configuration details.

Data Flow Diagram:

```
graph LR; source1[salesdata_DF] -- "Import data from rawsalesdata" --> derivedColumn1[derivedColumn1]; derivedColumn1 -- "Columns: 11 total" --> output
```

Derived Column's Settings:

- Output stream name:** derivedColumn1
- Description:** Creating/updating the columns: 'SalesPrimarykey', 'ProductID', 'SalesRepID', 'Location', 'Date', 'Units'.
- Incoming stream:** source1
- Columns:**
 - column1: toDate(timestamp(Data,'dd-m-yyyy'))
 - Country: regexSplit(Location,":")[]
 - State: regexSplit(Location,":")[2]

It is not worked we have used another expressions

The screenshot shows the Microsoft Azure Data Factory interface. A Data Flow named "salesdata_DF" is selected. The flow consists of a "source" and a "derivedColumn1" transformation. The "derivedColumn1" transformation has three columns: "Country", "State", and "date". The expressions for these columns are:

- Country: `regexSplit(Location, ',')[1]`
- State: `regexSplit(Location, ',')[2]`
- date: `toString(toDate(Date, 'd.M.yyyy'), 'dd-MM-yyyy')`

The "Data preview" tab shows the resulting data:

Location	Date	Units	PercentOfStandardCost	RevenueDiscount	Country	State	date
Germany;Frankfurt	1.10.2017	60	0.99	0.35	Germany	Frankfurt	01-10-2017
Denmark;Copenhagen	26.4.2017	71	0.972	0.35	Denmark	Copenhagen	26-04-2017
Germany;Dresden	28.9.2017	68	0.988	0.35	Germany	Dresden	28-09-2017
Germany;Berlin	16.4.2017	157	0.97	0.55	Germany	Berlin	16-04-2017
Denmark;Copenhagen	13.2.2017	103	0.965	0.5	Denmark	Copenhagen	13-02-2017
Czech republic;Prague	3.12.2017	48	0.997	0.25	Czech republic	Prague	03-12-2017
Denmark;Copenhagen	26.8.2017	30	0.984	0.25	Denmark	Copenhagen	26-08-2017
Germany;Frankfurt	16.4.2017	57	0.97	0.35	Germany	Frankfurt	16-04-2017
Germany;Frankfurt	3.11.2017	145	0.994	0.5	Germany	Frankfurt	03-11-2017
Germany;Berlin	1.1.2017	11	0.954	0.25	Germany	Berlin	01-01-2017

`toString(toDate(Date, 'd.M.yyyy'), 'dd-MM-yyyy')`

The screenshot shows the Microsoft Azure Data Factory interface. The same Data Flow "salesdata_DF" is selected. The "Data preview" tab shows the resulting data, which is currently empty (0 rows). This indicates that the transformation has been successfully applied.

Ok now we have done the modified with derived columns, now we need to delete the original col using select .

Screenshot of Microsoft Azure Data Factory Data Flow interface showing the mapping of columns from a source dataset to a select operation.

The Data Flow pipeline consists of a source, a derivedColumn1 transformation, and a select1 transformation.

Input columns:

- fSalesPrimaryKey
- ProductID
- SalesRepID
- Location
- source1@Date
- Units
- PercentOfStandardCost
- RevenueDiscount
- Country
- State
- derivedColumn1@date

Output columns (Name as):

- fSalesPrimaryKey
- ProductID
- SalesRepID
- Location
- Date
- Units
- PercentOfStandardCost
- RevenueDiscount
- Country
- State
- date

Mapping Summary: 11 mappings: All inputs mapped.

Screenshot of Microsoft Azure Data Factory Data Flow interface showing the mapping of columns from a source dataset to a select operation.

The Data Flow pipeline consists of a source, a derivedColumn1 transformation, and a select1 transformation.

Input columns:

- fSalesPrimaryKey
- ProductID
- SalesRepID
- Country
- State
- derivedColumn1@date

Output columns (Name as):

- fSalesPrimaryKey
- ProductID
- SalesRepID
- Country
- State
- Date
- Units
- PercentOfStandardCost
- RevenueDiscount

Mapping Summary: 9 mappings: 2 column(s) from the inputs left unmapped.

The screenshot shows the Microsoft Azure Data Factory Data Flow preview interface. A data flow named "salesdata_DF" is selected. The flow consists of three main stages: "source", "derivedColumn1", and "select1". The "derivedColumn1" stage is currently active, showing a "Creating/Updating" status. Below the flow diagram is a "Data preview" pane. The preview shows a table with the following columns and data:

	ProductID	SalesRepID	Country	State	Date	Units	PercentOfStandardCost	RevenueDiscount
10.	7	Germany	Frankfurt	01-10-2017	60	0.99	0.35	
8	2	Denmark	Copenhagen	26-04-2017	71	0.972	0.35	
9	6	Germany	Dresden	28-09-2017	68	0.988	0.35	
2	5	Germany	Berlin	16-04-2017	157	0.97	0.55	
8	2	Denmark	Copenhagen	13-02-2017	103	0.965	0.5	
8	1	Czech republic	Prague	03-12-2017	48	0.997	0.25	
4	2	Denmark	Copenhagen	26-08-2017	30	0.984	0.25	
4	7	Germany	Frankfurt	16-04-2017	57	0.97	0.35	
11	4	Germany	Frankfurt	03-11-2017	145	0.994	0.5	
5	5	Germany	Berlin	01-01-2017	31	0.954	0.25	
9	6	Germany	Dresden	04-03-2017	87	0.964	0.4	
9	3	Czech republic	Ostrava	23-04-2017	50	0.973	0.35	
3	1	Czech republic	Prague	21-01-2017	72	0.964	0.35	

Now we need to sink the data in azure lake storage in staging data folder.

Create output data set as stagingsalesdata

The screenshot shows the Microsoft Azure Data Factory Data Flow sink configuration. A sink named "sink1" is selected. The configuration includes the following fields:

- Output stream name:** sink1
- Description:** Add sink dataset
- Incoming stream:** select1
- Sink type:** Dataset (selected)
- Dataset:** Select... (dropdown menu)
- Options:**
 - Allow schema drift
 - Validate schema

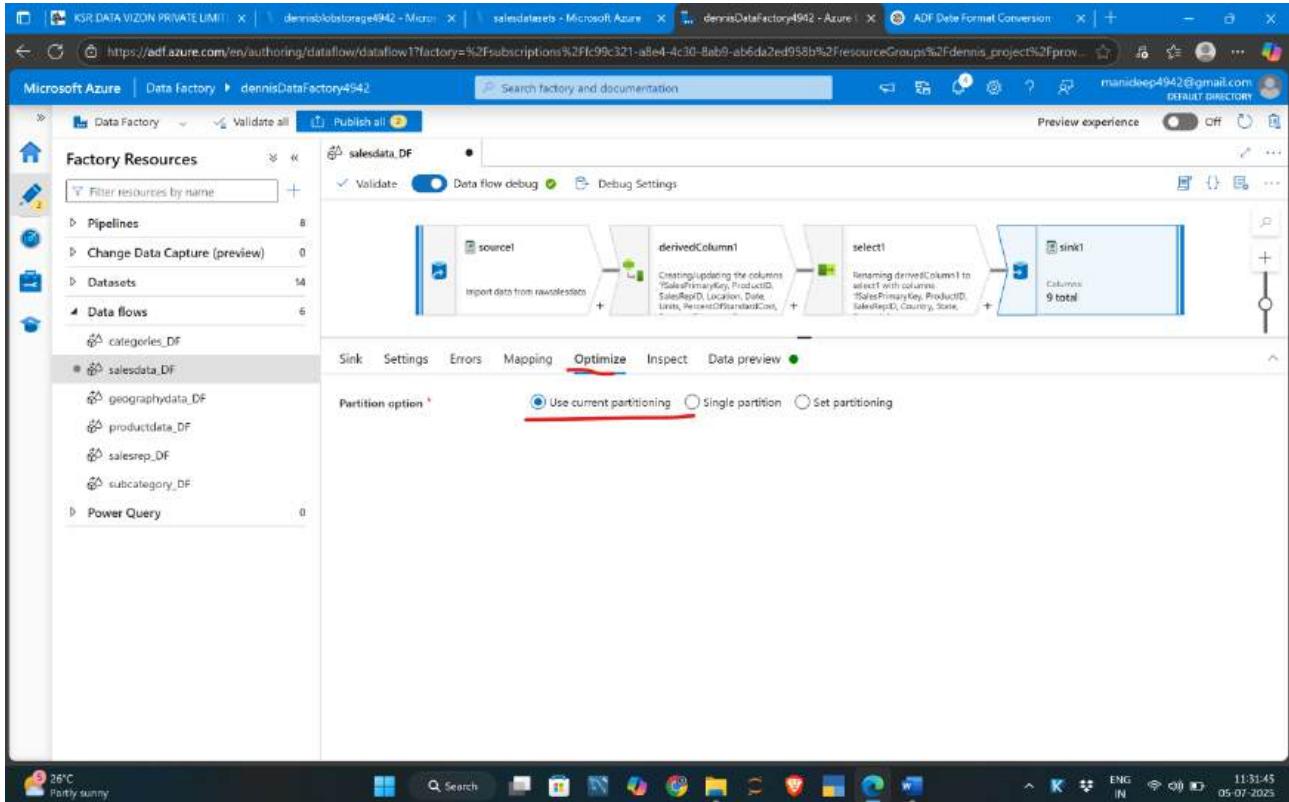
On the right side of the interface, a "New dataset" dialog is open, listing various Azure data stores:

- All
- Azure
- Database
- File
- Generic protocol

Azure Data Explorer (Kusto)	Azure Data Lake Storage Gen2	Azure Data Lake Storage Gen2
Azure Database for PostgreSQL	Azure SQL Database	Azure SQL Database Managed Instance

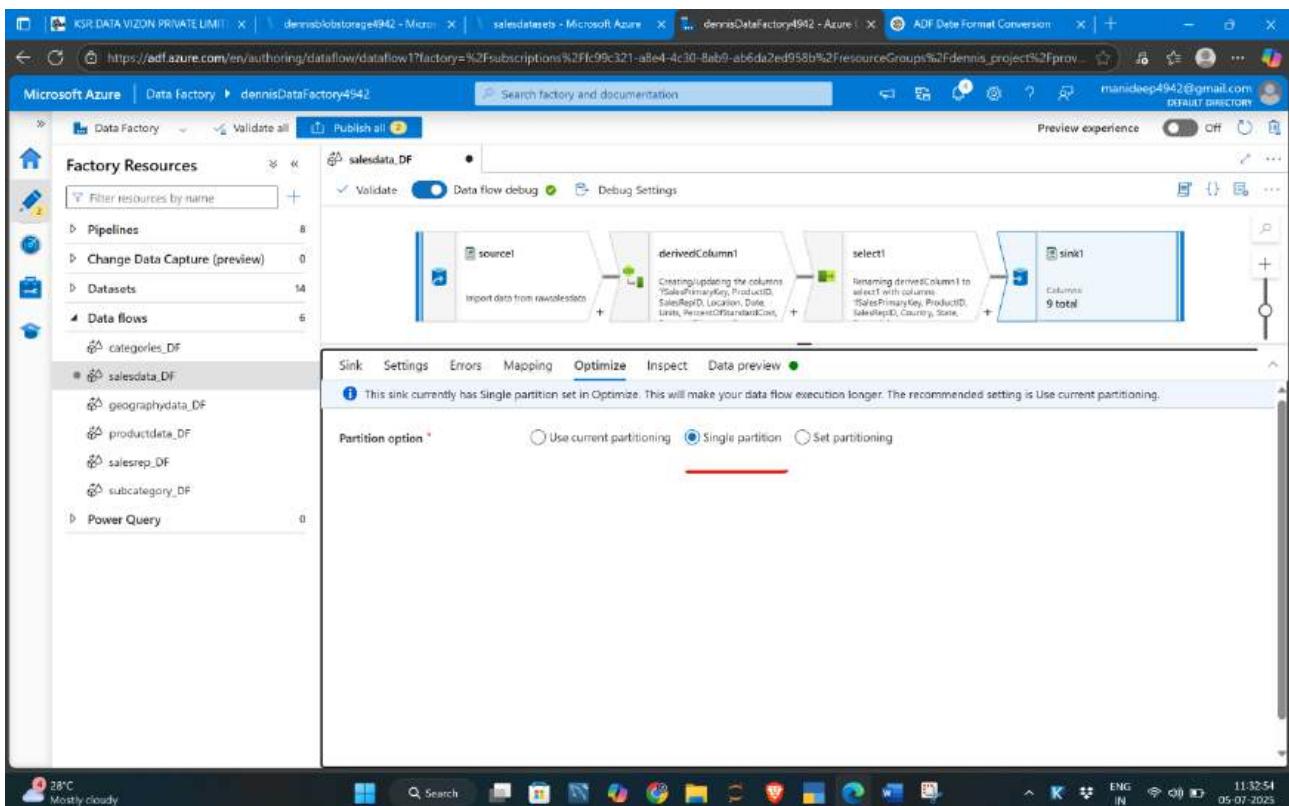
Ok now here is the main step we have clearer we have runned the pipelines

It creates multiple files whatever we have derived in dataflow that files it has been created because of we selected this option in dataflow.



We need to select the single partition if we select this we get a single file now

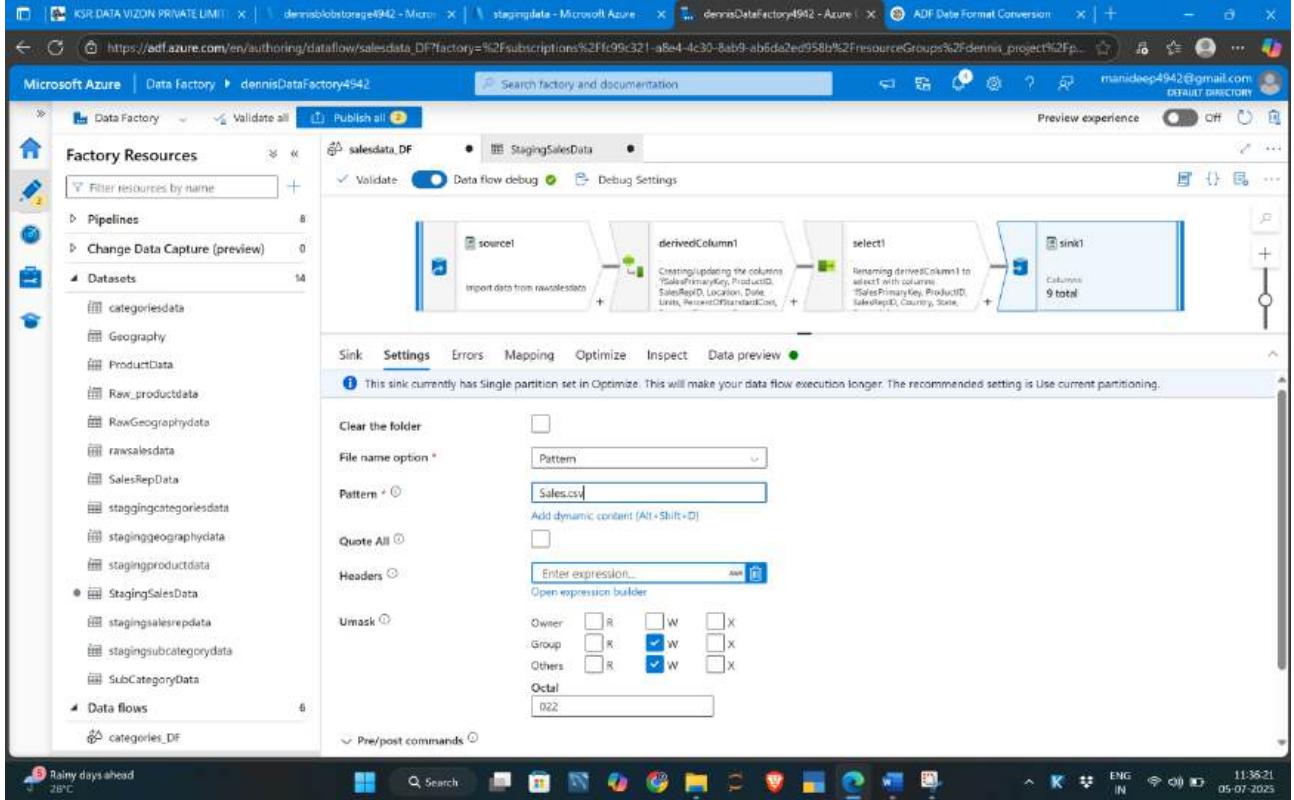
It takes all 5 sales files and give the single sales file .



And also one thing is that it take the random file name by the azure while getting output to us

So if you need a needed file name we have do one more setting right now

Go to setting → Filenameoption → select pattern → give pattern name (Sales.csv)



Now do publish

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists various datasets and data flows. The main workspace displays a data flow named 'salesdata_DF' with the following components:

- source:** Import data from rawsalesdata
- derivedColumn:** Creating 1 primary key SalesRepID, Logins, PercentOf
- sink:** File name option: Pattern (Sales.csv)

The 'Pending changes' table shows two items:

NAME	CHANGE	EXISTING
Datasets	StagingSalesData (New)	-
Data flows	salesdata_DF (New)	-

At the bottom, there are 'Publish' and 'Cancel' buttons.

So now aging we do the delete the yesterday file generated by azure as single file

We go and on the pattern again

See we have changed here same as data set name

The screenshot shows the Microsoft Azure Data Factory interface. The 'Factory Resources' sidebar lists various datasets and data flows. The main workspace displays a data flow named 'categories_DF' with the following components:

- source:** Import data from categoriesdata
- aggregate:** Aggregating data by Category, CategoryKey producing columns: countofCategory
- select:** Renaming aggregate to select with columns: Category, CategoryKey
- sink:** Columns 2 total

The 'Sink' settings show a 'File name option' set to 'Pattern' (stagingcategoriesdata.csv).

Microsoft Azure | Data Factory | dennisDataFactory4542

Search factory and documentation

Preview experience: Off

Factory Resources

- categoriesdata
- Geography
- ProductData
- Raw_productdata
- RawGeographydata
- rawsalesdata
- SalesRepData
- stagingcategoriesdata
- staginggeographydata
- stagingproductdata
- StagingSalesData
- stagingsalesrepdata
- stagingsubcategorydata
- SubCategoryData

Data flows (6)

- categories_DF
- salesdata_DF
- geographydata_DF
- productdata_DF

geographydata_DF

Validate: Data flow debug, Debug Settings

Source: Import data from RawGeographydata

Transformation: select1 (Renaming source1 to select1 with column: 'Country', 'Town')

Sink: Column: 2 total

Sink Settings Errors Mapping Optimize Inspect Data preview

This sink currently has Single partition set in Optimize. This will make your data flow execution longer. The recommended setting is Use current partitioning.

Clear the folder

File name option: Pattern

Quote All

Headers Enter expression...

```
graph LR; source[Import data from RawGeographydata] --> select1[select1]; select1 --> sink1[Column: 2 total];
```

Microsoft Azure | Data Factory | dennisDataFactory4542

Search factory and documentation

Preview experience: Off

Factory Resources

- categoriesdata
- Geography
- ProductData
- Raw_productdata
- RawGeographydata
- rawsalesdata
- SalesRepData
- stagingcategoriesdata
- staginggeographydata
- stagingproductdata
- StagingSalesData
- stagingsalesrepdata
- stagingsubcategorydata
- SubCategoryData

Data flows (6)

- categories_DF
- salesdata_DF
- geographydata_DF
- productdata_DF

productdata_DF

Validate: Data flow debug, Debug Settings

Source: Import data from Raw_productdata

Transformation: aggregate (Aggregating data by ProductID, Sub Category Key, ProductName, StandardCost, Color, RetailPrice, producing)

Transformation: select1 (Renaming aggregate1 to select1 with columns: ProductID, Sub Category Key, ProductName, StandardCost, Color, RetailPrice, producing)

Sink: Column: 6 total

Sink Settings Errors Mapping Optimize Inspect Data preview

This sink currently has Single partition set in Optimize. This will make your data flow execution longer. The recommended setting is Use current partitioning.

Clear the folder

File name option: Pattern

Quote All

Headers Enter expression...

```
graph LR; source[Import data from Raw_productdata] --> aggregate[aggregate]; aggregate --> select1[select1]; select1 --> sink1[Column: 6 total];
```

Microsoft Azure | Data Factory | dennisDataFactory4542

Validate all | Publish all

Factory Resources

- categories_DF
- geographydata_DF
- productdata_DF
- salesrep_DF**
- Geography
- ProductData
- Raw_productdata
- RawGeographydata
- rawsalesdata
- SalesRepData
- stagingcategoriesdata
- staginggeographydata
- stagingproductdata
- StagingSalesData
- stagingsalesrepdata**
- stagingsubcategorydata
- SubCategoryData

Data flows

- categories_DF
- salesdata_DF**
- geographydata_DF
- productdata_DF

28°C Mostly cloudy 11:43:46 05-07-2023

Microsoft Azure | Data Factory | dennisDataFactory4542

Validate all | Publish all

Factory Resources

- Raw_productdata
- RawGeographydata
- rawsalesdata
- SalesRepData
- stagingcategoriesdata
- staginggeographydata
- stagingproductdata
- StagingSalesData**
- stagingsalesrepdata
- stagingsubcategorydata
- SubCategoryData

Data flows

- salesdata_DF**
- categories_DF
- geographydata_DF
- productdata_DF
- salesrep_DF
- subcategory_DF

28°C Mostly cloudy 11:45:02 05-07-2023

Do publish all

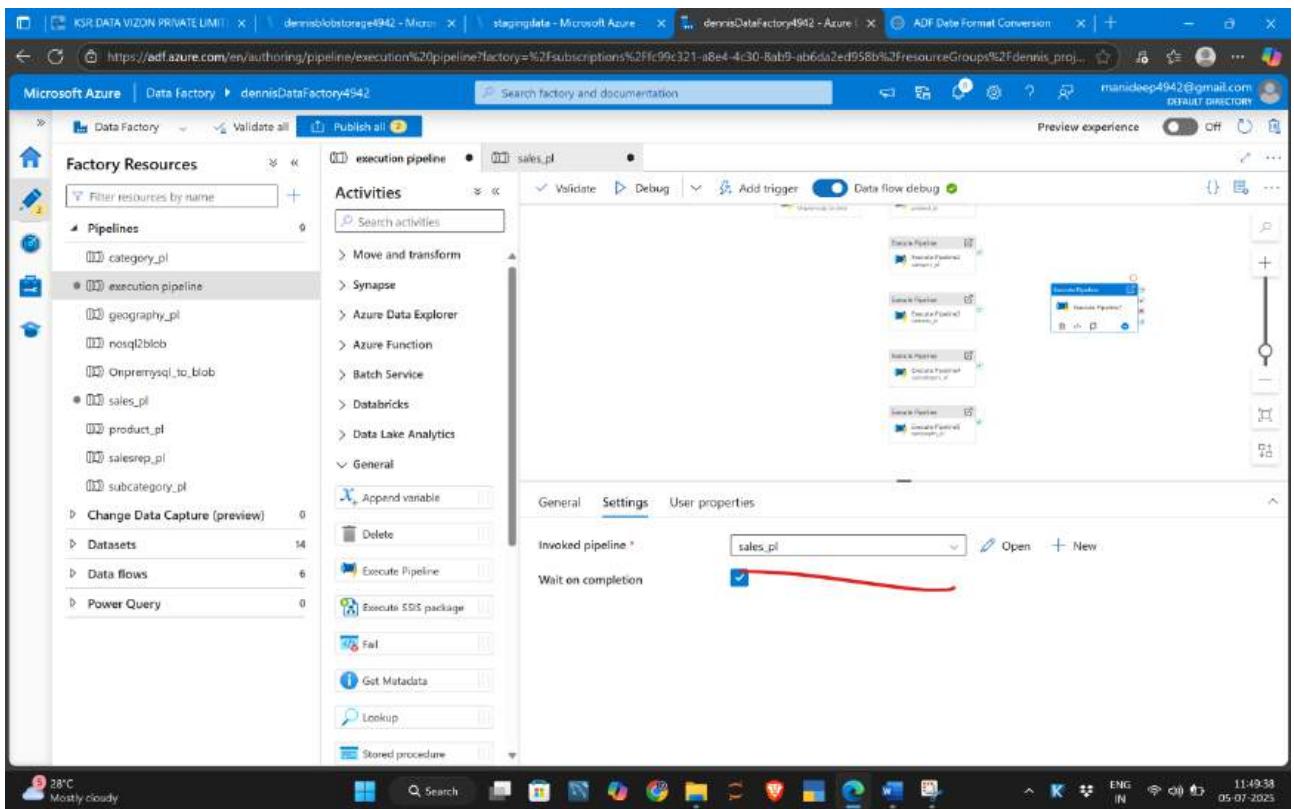
The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists various datasets and data flows. The 'Data flows' section is expanded, showing several data flows including 'categories_DF', 'salesdata_DF', 'geographydata_DF', 'productdata_DF', 'salesrep_DF', and 'subcategory_DF'. The 'salesdata_DF' data flow is currently selected. The main workspace displays the data flow structure: a 'source1' (rawsalesdata) connected to a 'sink' (StagingSalesData.csv). The sink settings include a 'File name option' set to 'Pattern' with value 'StagingSalesData.csv'. The 'Settings' tab indicates that the sink has a 'Single partition' set in 'Optimize'. At the bottom right, there is a 'Publish' button.

Now we need to go and run the pipeline

Create sales pipeline first

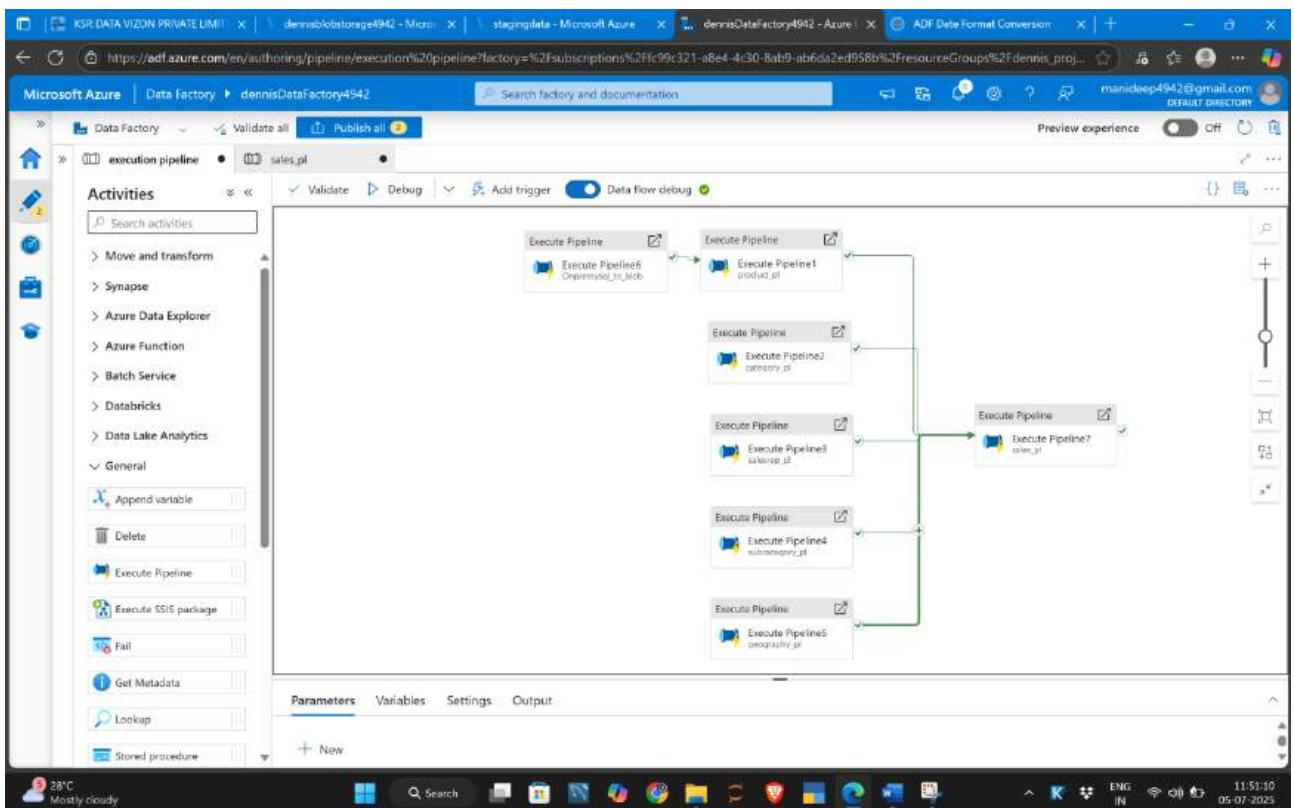
The screenshot shows the Microsoft Azure Data Factory interface. The 'Factory Resources' sidebar lists pipelines, datasets, data flows, and power query resources. A new pipeline named 'pipeline1' is being created. The pipeline activities pane shows a single activity named 'Data flow'. The pipeline properties pane on the right shows the general settings for 'pipeline1', including the name 'pipeline1' and a description field. The 'General' tab is selected, showing the data flow is 'salesdata_DF'. Other tabs include 'Settings', 'Parameters', and 'User properties'. The pipeline is currently in preview experience mode, indicated by a switch at the top right.

Go to → execute pipeline → add one more execute pipe for salesDF



Now connect all pipelines to the sales pipeline to execute all

Like this



Now we go delete all file from the blob stagingdata

The screenshot shows the Microsoft Azure Storage Container Overview page for the 'stagingdata' container. The table lists 11 blobs, all of which have checkboxes checked under the 'Name' column. The columns include Name, Last modified, Access tier, Blob type, Size, and Lease state. A red box highlights the 'Delete' button at the top of the table.

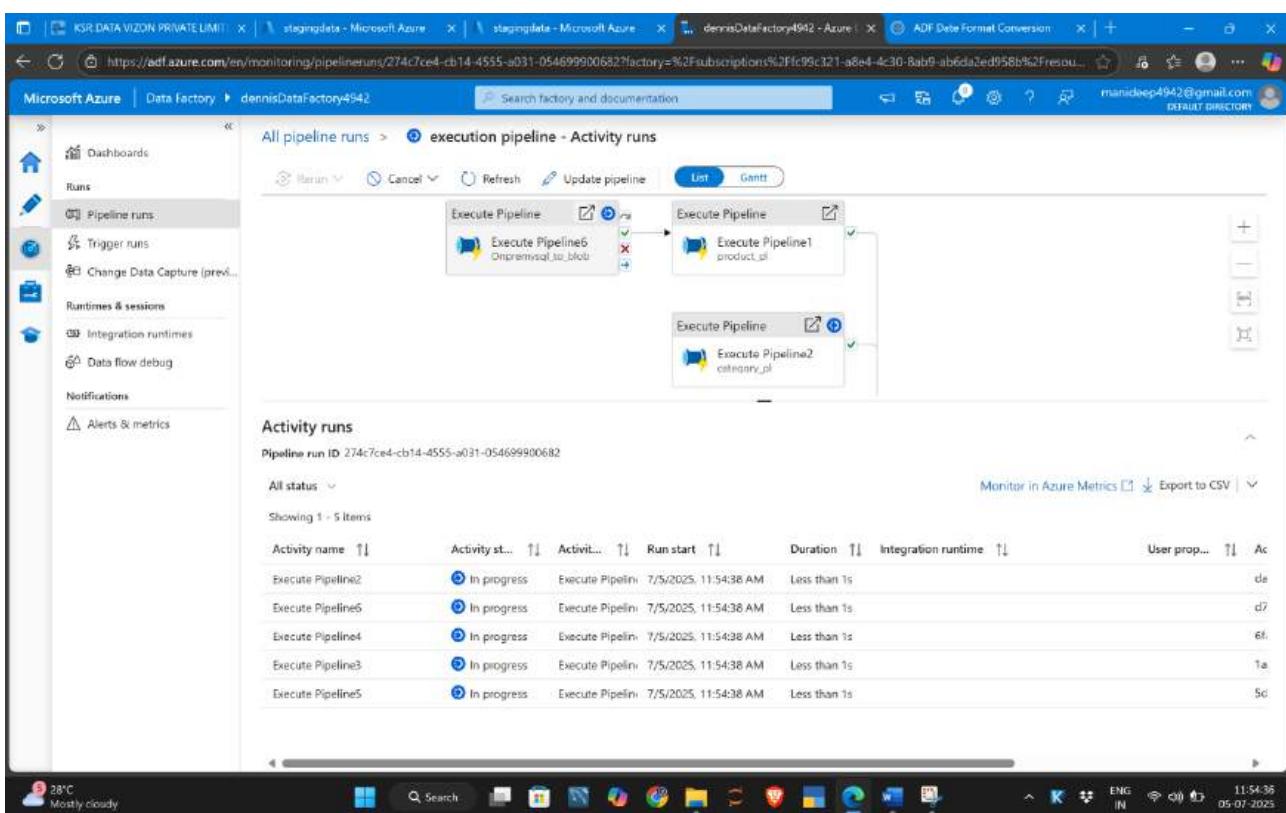
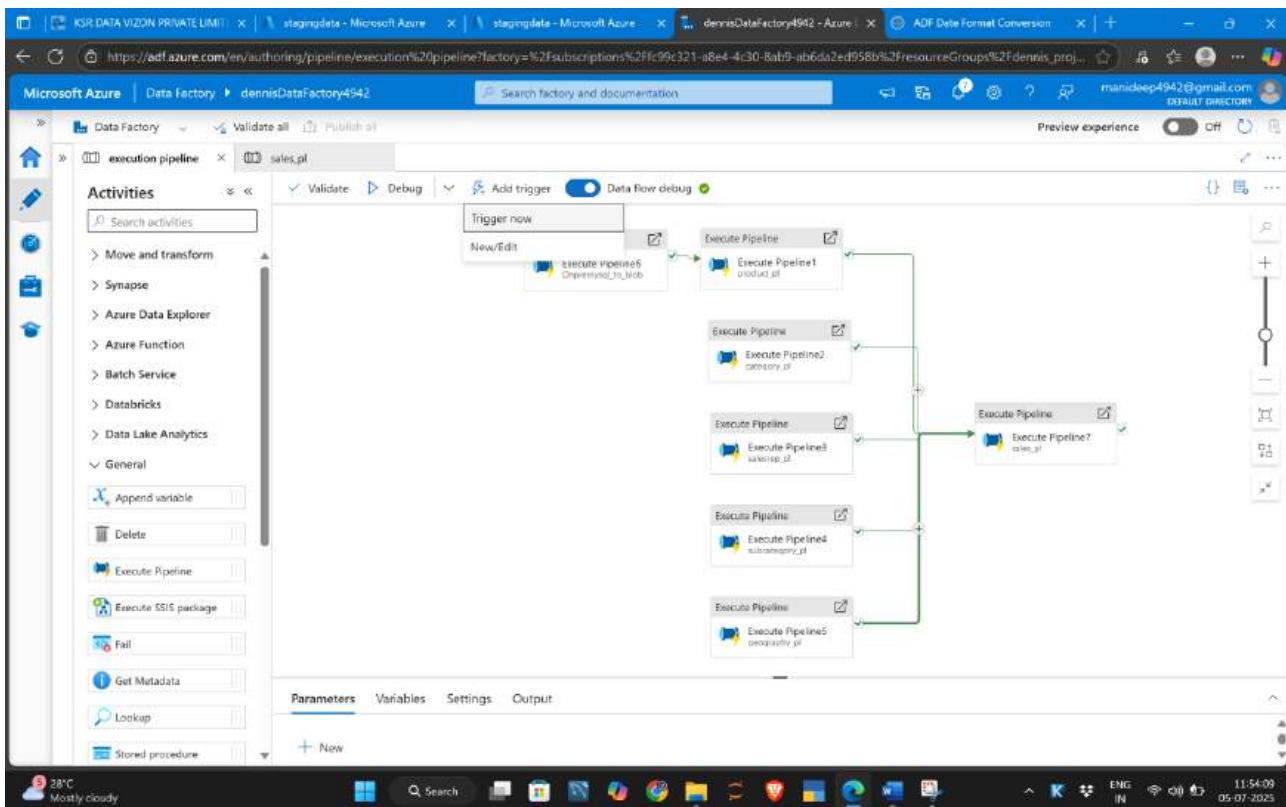
Name	Last modified	Access tier	Blob type	Size	Lease state
_SUCCESS	4/7/2025, 5:32:39 pm	Hot (Inferred)	Block blob	0	Available
par-00000-3bd98e8c-8559-48e5-b724-012...	4/7/2025, 5:23:10 pm	Hot (Inferred)	Block blob	146 B	Available
par-00000-476e766d-bedd-482f-abb2-727...	4/7/2025, 5:32:34 pm	Hot (Inferred)	Block blob	395 B	Available
par-00000-4d193f72-90f8-4a28-8055-793d...	4/7/2025, 5:32:16 pm	Hot (Inferred)	Block blob	117 B	Available
par-00000-53bb5212-4c7a-41fb-b38f-c2d3...	4/7/2025, 5:23:07 pm	Hot (Inferred)	Block blob	117 B	Available
par-00000-593aa2ca-257b-435d-a314-031...	4/7/2025, 5:23:15 pm	Hot (Inferred)	Block blob	86 B	Available
par-00000-6c9cf080-03ad-4038-b4fa-4bf1...	4/7/2025, 5:32:39 pm	Hot (Inferred)	Block blob	41 B	Available
par-00000-8128dd62-1e9a-4cb9-83a0-ab9...	4/7/2025, 5:32:19 pm	Hot (Inferred)	Block blob	86 B	Available
par-00000-b91092fc-6f20-46bf-8480-9949...	4/7/2025, 5:23:04 pm	Hot (Inferred)	Block blob	41 B	Available
par-00000-cb00589a-0919-474d-a5eb-765...	4/7/2025, 5:23:55 pm	Hot (Inferred)	Block blob	395 B	Available
par-00000-d56dd91a-fcc3-470a-a307-cb8...	4/7/2025, 5:32:14 pm	Hot (Inferred)	Block blob	146 B	Available

All have been deleted

The screenshot shows the Microsoft Azure Storage Container Overview page for the 'stagingdata' container. A green success message box appears in the top right corner, stating 'Successfully deleted blobs and directories' and 'Successfully deleted 10 blobs and directories.' The table below shows 0 items, indicating the blobs have been deleted.

Name	Last modified	Access tier	Blob type	Size	Lease state
No items found					

Lets run the pipeline we see the files



Successfully runnne pipeline

The screenshot shows the Microsoft Azure Data Factory interface. On the left, a sidebar lists navigation options: Dashboards, Runs, Pipeline runs (selected), Trigger runs, Change Data Capture (preview), Runtimes & sessions, Integration runtimes, Data flow debug, Notifications, Alerts & metrics, and Help & support. The main content area is titled "All pipeline runs > execution pipeline - Activity runs". It displays a Gantt chart with two tasks: "Execute Pipeline6" (status: Succeeded) and "Execute Pipeline1" (status: Succeeded). Below the chart is a table titled "Activity runs" showing seven rows of data:

Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime	User prop...	Ac...
Execute Pipeline7	Succeeded	Execute Pipeli...	7/5/2025, 11:58:05 AM	31s			d3
Execute Pipeline1	Succeeded	Execute Pipeli...	7/5/2025, 11:55:01 AM	3m 4s			3d
Execute Pipeline2	Succeeded	Execute Pipeli...	7/5/2025, 11:54:38 AM	3m 10s			da
Execute Pipeline6	Succeeded	Execute Pipeli...	7/5/2025, 11:54:38 AM	23s			d7
Execute Pipeline4	Succeeded	Execute Pipeli...	7/5/2025, 11:54:38 AM	3m 5s			6d
Execute Pipeline3	Succeeded	Execute Pipeli...	7/5/2025, 11:54:38 AM	3m 6s			1a
Execute Pipeline5	Succeeded	Execute Pipeli...	7/5/2025, 11:54:38 AM	3m 9s			5c

See here the data as a single file

Blob storage :

The screenshot shows the Microsoft Azure Storage Container Overview page for the "stagingdata" container. The top navigation bar includes links for Home, dennisblobstorage-942, Containers, and a search bar. The main content area shows the container's properties: Name (stagingdata), Type (Container), Status (Active), and Last modified (5/7/2025). Below this is a table titled "Showing all 5 items" listing five blob files:

Name	Last modified	Access tier	Block type	Size	Lease state
stagingsalesrepdata.csv	5/7/2025, 12:05:13 pm	Hot (Inferred)	Block blob	117 B	Available
stagingcategorydata.csv	5/7/2025, 12:05:21 pm	Hot (Inferred)	Block blob	41 B	Available
staginggeographydata.csv	5/7/2025, 12:05:09 pm	Hot (Inferred)	Block blob	146 B	Available
stagingproductdata.csv	5/7/2025, 12:05:35 pm	Hot (Inferred)	Block blob	395 B	Available
stagingsubcategorydata.csv	5/7/2025, 12:05:18 pm	Hot (Inferred)	Block blob	86 B	Available

ADLS

The screenshot shows the Microsoft Azure Storage Container blade for the 'stagingdata' container. The container overview shows one item: 'StagingSalesData.csv'. The blob details show it was last modified on 5/7/2025 at 12:06:08 pm, has a hot inferred access tier, is a block blob, and is 2.74 MB in size. The status is available.

Now lets do the final transformations:

Create one dataflow named as Dennis_DF

The screenshot shows the Microsoft Data Factory Data Flow blade. The 'Factory Resources' sidebar shows several data flows, including 'Dennis_DF', which is highlighted with a red arrow. The main workspace displays a data flow named 'Dennis_DF' with two main components: 'categorydata' and 'subcategorydata'. The 'categorydata' component has an output named '6 total'. The 'subcategorydata' component imports data from 'categorydata' and has an output named 'subcategorydata'. The 'subcategorydata' component is highlighted with a red arrow. The data preview pane shows a table with 31 rows of product data.

ProductID	Sub Category Key	ProductName	StandardCost	Color	RetailPrice
11	4	Carlota	9.15	Fluorescent Blue	29.95
9	1	Magnum	8.25	Green	26.95
4	1	Quad	13.75	Red	43.95
7	1	Bind	8.25	Red	26.95
5	1	Black Mark	13.75	Blue	43.95
3	2	Magnum	7.55	Green	23.95
1	3	Alder	7.55	Red	23.95

now we going to join the category data and subcategory data using join...

category data preview

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists various datasets and data flows. Under 'Data flows', 'Dennis_DF' is selected. The main workspace displays a data flow diagram titled 'Dennis_DF'. It consists of two sources: 'source1' (Raw_productdata) and 'source2' (Import data from stagingcategorydata). A 'join' operation is performed between them. The 'Data preview' tab is active, showing the resulting data. The preview table has two rows:

Category	CategoryKey
Special	1
General	2

Subcategory data.

The screenshot shows the Microsoft Azure Data Factory interface, similar to the previous one but with different data. The 'Factory Resources' sidebar shows 'Dennis_DF' is selected. The data flow diagram 'Dennis_DF' shows 'source1' (Import data from stagingsubcategorydata) and 'source2' (Raw_productdata). A 'join' operation is performed. The 'Data preview' tab is active, showing the resulting data. The preview table has 4 rows:

SubCategoryKey	CategoryKey	SubCategory Name
1	1	Extra
2	2	Regular
3	1	Micro
4	2	Super

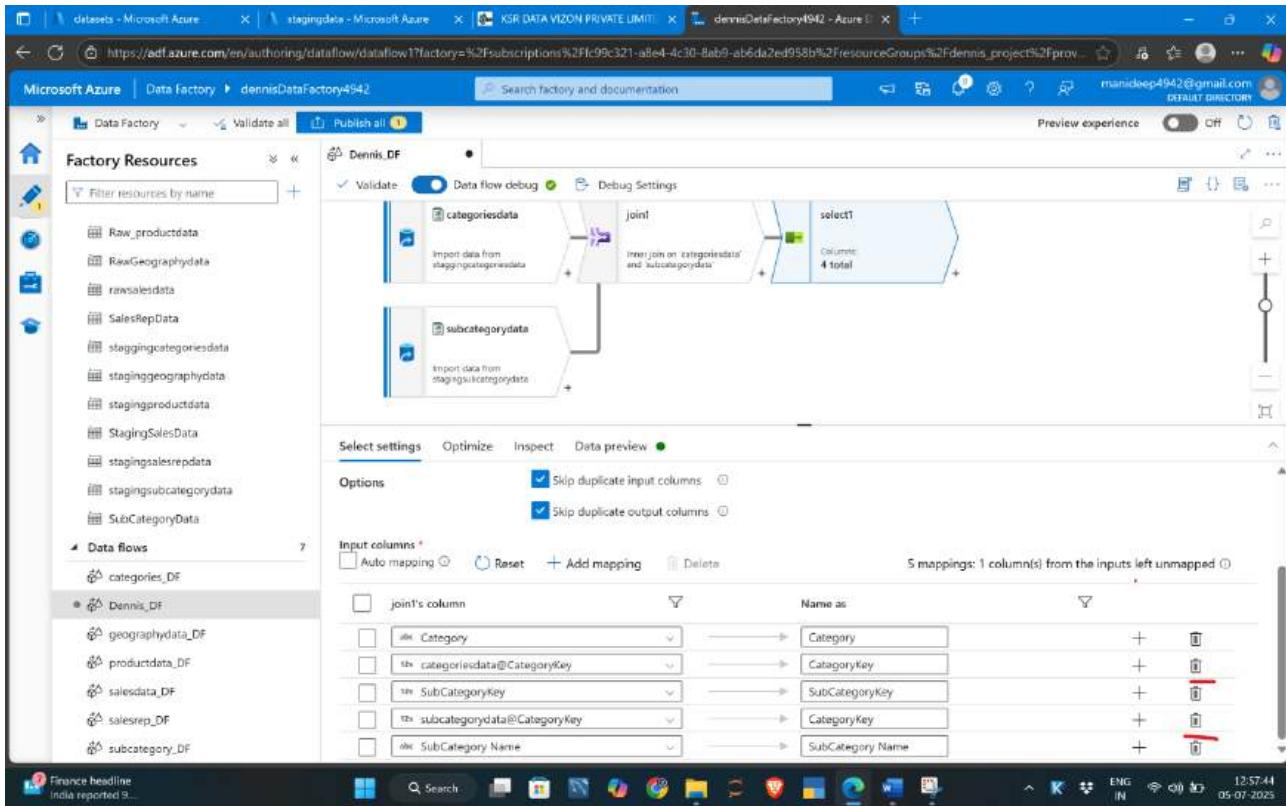
Now both tables we are joining table with inner join

The screenshot shows the Microsoft Azure Data Factory Data Flow interface. On the left, the 'Factory Resources' sidebar lists various datasets and data flows. Under 'Data flows', 'Dennis_DF' is selected, which contains a single data flow named 'join1'. The data flow diagram shows two inputs: 'categoriesdata' (imported from 'staggegeographedata') and 'subcategorydata' (imported from 'staggegeographodata'). These two streams are joined together. The 'Join settings' tab is active, showing the following configuration:

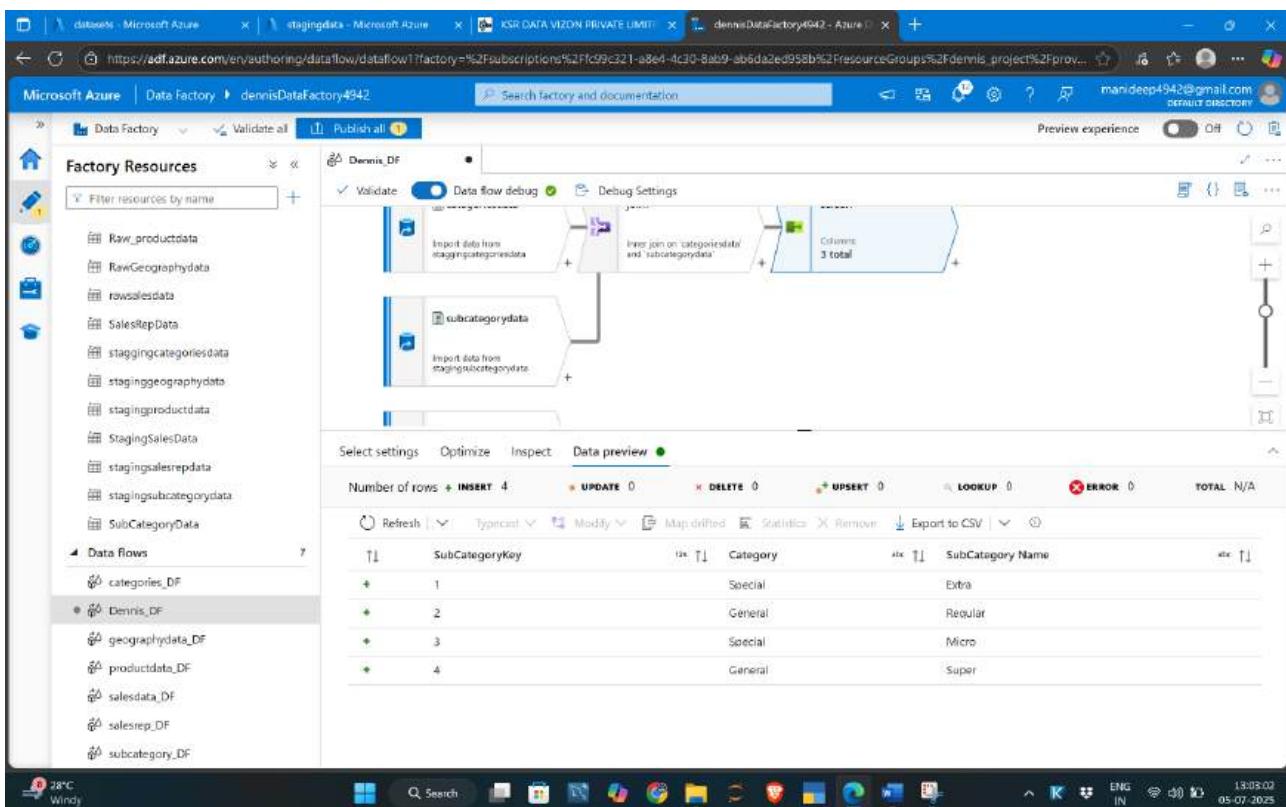
- Description:** Inner join on 'categoriesdata' and 'subcategorydata'
- Left stream:** categoriesdata
- Right stream:** subcategorydata
- Join type:** Inner (selected)
- Use fuzzy matching:**

The second screenshot shows the same interface after saving changes. The 'Output stream name' field now contains 'join1'. Additionally, under 'Join conditions', there is a condition: 'Left: categoriesdata's column CategoryKey = Right: subcategorydata's column CategoryKey'. The rest of the configuration remains the same.

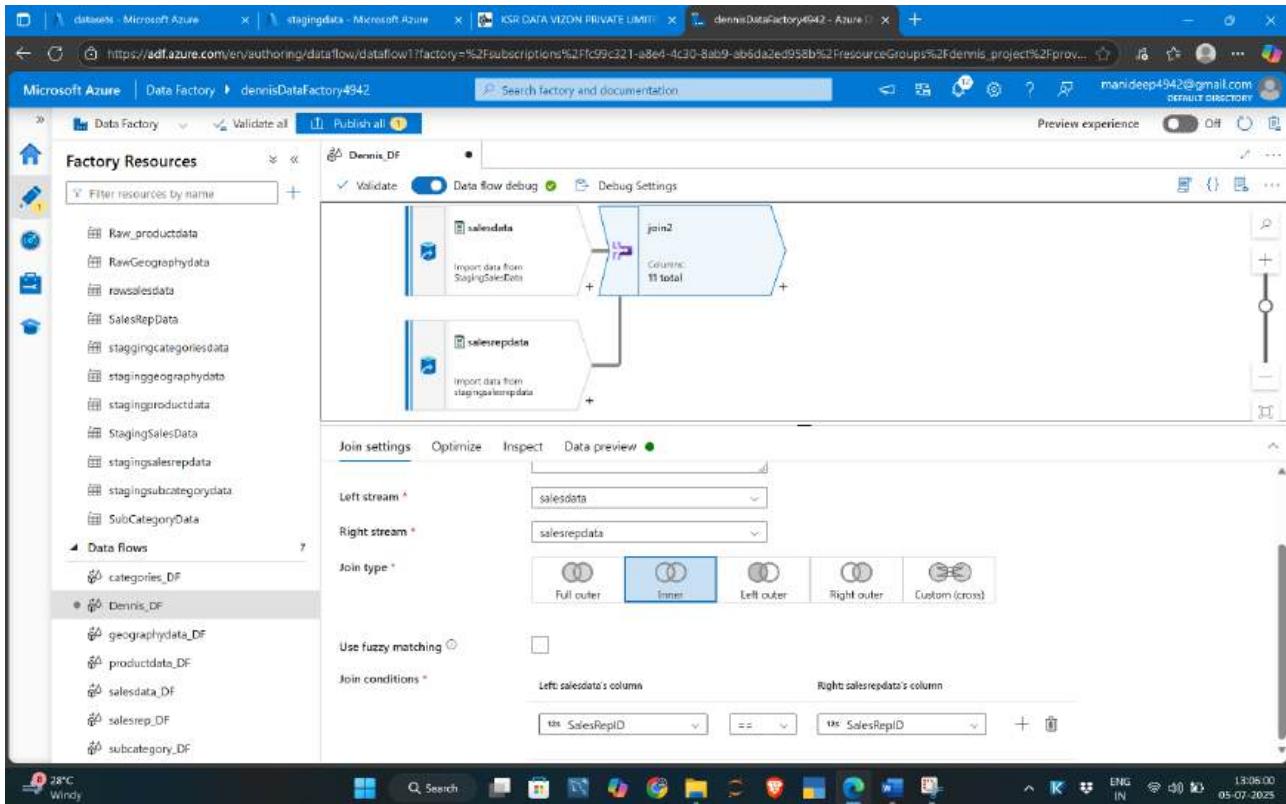
We have deleted the category key in one table category key in another table.



It is been ok now here it how it looks



Now do for sales data:



now we see the preview of data

The screenshot shows the Microsoft Azure Data Factory Data Flow interface with the 'Data preview' tab selected. The data flow diagram remains the same, showing the join operation between 'salesdata' and 'salesrepdata'. The preview pane displays a table with 8 rows of sample data. The columns are SalesRepID, Country, City, date, Units, PercentOfStandardCost, RevenueDiscount, and Sales Rep Name. All rows show 'Denmark' as the country, 'Copenhagen' as the city, and 'Bill Muray' as the Sales Rep Name.

SalesRepID	Country	City	date	Units	PercentOfStandardCost	RevenueDiscount	Sales Rep Name
Denmark	Copenhagen	2017-04-26	71	0.972	0.35	Bill Muray	
Denmark	Copenhagen	2017-02-13	103	0.965	0.5	Bill Muray	
Denmark	Copenhagen	2017-08-26	30	0.964	0.25	Bill Muray	
Denmark	Copenhagen	2017-05-18	138	0.977	0.5	Bill Muray	
Denmark	Copenhagen	2017-09-05	85	0.988	0.25	Bill Muray	
Denmark	Copenhagen	2017-08-24	83	0.987	0.4	Bill Muray	
Denmark	Copenhagen	2017-01-14	77	0.955	0.4	Bill Muray	
Denmark	Copenhagen	2017-05-01	86	0.971	0.4	Bill Muray	

Now we are removing the sales rep key

The screenshot shows the Microsoft Azure Data Factory Data Flow blade. On the left, the 'Factory Resources' sidebar lists Pipelines, Change Data Capture (preview), Datasets, Data flows, Power Query, and Dennis_DF. Under Dennis_DF, there are datasets: geographydata_DF, productdata_DF, salesdata_DF, salerep_DF, and subcategory_DF. The main area displays a data flow named 'Dennis_DF'. It consists of two inputs: 'Import data from StagingSalesData' and 'getsalesrepdata'. A 'Join' component is used to join these two inputs. The output of the join has 10 total columns. Below the data flow diagram, the 'Data preview' tab is selected, showing a list of input columns and their corresponding names as they appear in the output. One column, 'SalesRepID', is highlighted with a red border.

Her we go

This screenshot is identical to the one above, showing the Microsoft Azure Data Factory Data Flow blade. The 'Data preview' tab is still selected, and the 'SalesRepID' column is highlighted with a red border. A red box highlights the 'Remove mapping' button in the bottom right corner of the mapping table, indicating the action being performed.

Here the sales rep name is added

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists various datasets and data flows. The main workspace displays a data flow named 'Dennis_DF'. This flow starts with an 'Import data from stagingdata' source, followed by a 'join2' operation, and ends with a target dataset 'salesdata'. The 'Data preview' tab is selected, showing a preview of the data with 100 rows. The columns in the preview table are: SalesPrimaryKey, ProductID, Sales Rep Name, Country, City, date, Units, and PercentOfStandardC. The preview table actually contains 11 rows of data.

Now we are going to do for the product data →with subcategorydata(finaloutput in df)

This is product data

The screenshot shows the Microsoft Azure Data Factory interface. The 'Factory Resources' sidebar lists datasets and data flows. The main workspace displays a data flow named 'Dennis_DF'. This flow starts with an 'Import data from stagingdata' source, followed by a 'join3' operation, and ends with a 'select1' operation. The 'Data preview' tab is selected, showing a preview of the data with 11 rows. The columns in the preview table are: ProductID, Sub Category Key, ProductName, StandardCost, Color, and RetailPrice.

This is sub category data

The screenshot shows the Microsoft Azure Data Factory Data Flow blade. On the left, the 'Factory Resources' sidebar lists various datasets and data flows. In the center, a data flow named 'Dennis_DF' is displayed. The data flow consists of two main stages: 'stagingproductdata' and 'categorysubdata'. A green arrow points from 'stagingproductdata' to 'categorysubdata'. Below the stages, the 'Data preview' tab is selected, showing a table with four rows. The columns are 'SubCategoryKey', 'Category', and 'SubCategory Name'. The data is as follows:

SubCategoryKey	Category	SubCategory Name
1	Special	Extra
2	General	Regular
3	Special	Micro
4	General	Super

Now we need to join both the tables with matching

The screenshot shows the 'Join conditions' section in the Data Flow blade. It includes a checkbox for 'Use fuzzy matching' and a 'Join conditions' table. The table has two columns: 'Left: getproductdata's column' and 'Right: categorysubdata's column'. A join condition is defined between 'Sub Category Key' and 'SubCategoryKey' using an equals operator (=). The condition is highlighted with a red border.

Left: getproductdata's column	Right: categorysubdata's column
12s Sub Category Key	= 12s SubCategoryKey

Preview data

The screenshot shows the Microsoft Azure Data Factory Data Flow preview interface. A join operation is being performed between two datasets: 'stagingproductdata' and 'categorysubcategory'. The preview pane displays 11 rows of data from the 'stagingproductdata' dataset, which includes columns such as ProductID, Sub Category Key, ProductName, StandardCost, Color, RetailPrice, and SubCategoryKey. The 'categorysubcategory' dataset is shown as a join condition. The interface includes standard data manipulation controls like Refresh, Typecast, Modify, Map drifted, Statistics, Remove, and Export to CSV.

Now we are removing this columns in this data & rearranging

The screenshot shows the Microsoft Azure Data Factory Data Flow mapping settings. The 'join3' stream is selected as the incoming stream. Under 'Options', 'Skip duplicate input columns' and 'Skip duplicate output columns' are checked. The 'Input columns' section lists the columns from the joined stream. The 'Mappings' section shows 9 mappings: ProductID to ProductID, Sub Category Key to Sub Category Key, ProductName to ProductName, StandardCost to StandardCost, Color to Color, RetailPrice to RetailPrice, SubCategoryKey to SubCategoryKey, Category to Category, and SubCategory Name to SubCategory Name. A red box highlights the first mapping for ProductID.

Don't delete the product id

This is all about product data

Now we are joining the product data with sales data

Using joins



This is what matching with productid in product data & product id in sales data set

Microsoft Azure | Data Factory | Dennis_DF | Dennis_DF | dennisDataFactory4542 | https://adf.azure.com/en/authoring/dataflow/Dennis_DF?factory=%2fsubscriptions%2fc99c321-a8e4-4c30-8ab9-ab6d92ed958b%2fresourceGroups%2fdennis_project%2fproviders... | manideep4942@gmail.com | DETAIL DIRECTORY | Preview experience | OFF

Factory Resources

- geography_pl
- nosql2blob
- Onpremssql_to_blob
- dennim_pl**
- product_pl
- sales_pl
- salesrep_pl
- subcategory_pl

Change Data Capture (preview) 0

Datasets 15

Data flows 7

categories_DF

Dennis_DF

- geographydata_DF
- productdata_DF
- salesdata_DF**
- salesrep_DF
- subcategory_DF

Power Query 0

Join settings

Output stream name: join4

Description: Inner join on 'salesdata' and 'productdata'

Left stream: salesdata

Right stream: productdata

Join type: Inner

Use fuzzy matching

Join conditions:

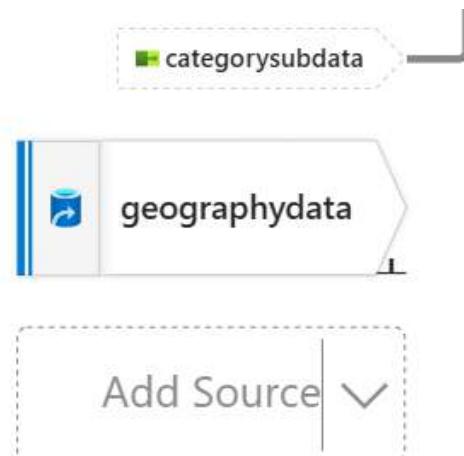
Left: salesdata's column ProductID == Right: productdata's column ProductID

After this join we are going to use select and removes the product id key in this list and rest of the list here.

The screenshot shows the Microsoft Azure Data Factory interface. On the left, there's a sidebar with 'Factory Resources' containing datasets like 'geography_pl', 'nosql2blob', 'Onpremssql_to_blob', 'dennim_pl', 'product_pl', 'sales_pl', 'salesrep_pl', 'subcategory_pl', and others. Under 'Data flows', 'Dennis_DF' is selected. The main area shows a data flow configuration for 'dennim_pl'. It has tabs for 'Select settings', 'Optimize', 'Inspect', 'Data preview', and 'Options'. Under 'Options', 'Skip duplicate output columns' is checked. The 'Input columns' section lists 14 mappings from 'join4's column to various output columns such as fSalesPrimaryKey, Country, City, date, Units, PercentOfStandardCost, RevenueDiscount, Sales Rep Name, ProductName, SubCategory Name, StandardCost, Color, RetailPrice, and Category.

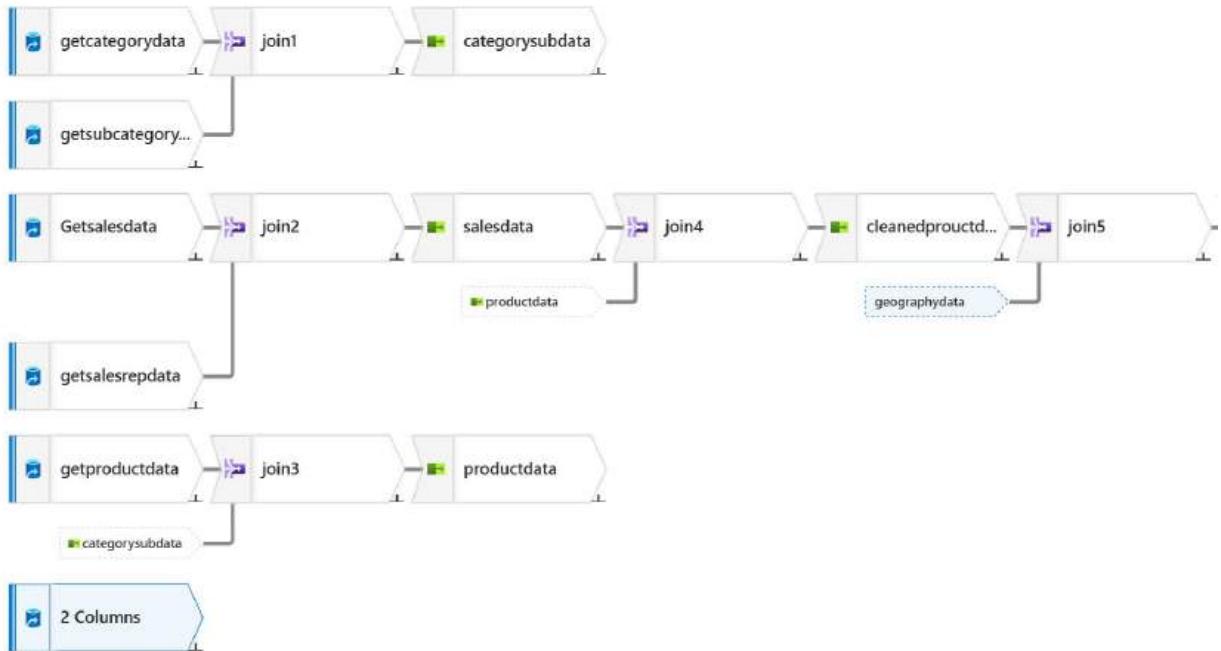
Now we have only one data set that is geography data set

So it is also take it as a consideration in this now we check the geography data



now we use the last join with final date what we have joined product and sales data .

Here it for reference



Now we add the columns what the client as been asked we do that face in this dataflow.

3. Additional column calculations

Task 3.1:

Calculate **Total Revenue** in Sales table, using the Product's Retail Price, and multiplying it by the Units.

Task 3.2:

Calculate **Total Cost** in Sales table, using the Product's Standard Cost, and multiplying it by the Units.

Task 3.3:

Calculate **Gross Profit** in Sales: Total Revenue – Total Cost

Task 3.4:

Calculate a measure for **AVG sales per day** – this is the average sum of **Total Revenue** per day based on the Dates of actual Sales.

Task 3.5:

- Breakdown Analysis by **Product (drop or increase)**

We have added total revenue, totalcost, gross profit

The screenshot shows the Microsoft Azure Data Factory Data Flow interface. A data flow named 'denim_pl' is being edited. The 'join5' stream is selected, and the 'derivedColumn1' settings are visible. The output stream name is 'derivedColumn1'. The description is 'Creating/Updating the columns: SalesPrimarykey, Country, City, date, Units, PercentOfStandardCost.' The incoming stream is 'join5'. The 'Columns' section contains two columns: 'TotalRevenue' with expression 'RetailPrice*Units' and 'TotalCost' with expression 'StandardCost*Units'.

Add another derived columns to grossprofit

The screenshot shows the Microsoft Azure Data Factory Data Flow interface. A data flow named 'denim_pl' is being edited. The 'join5' stream is selected, and the 'derivedColumn2' settings are visible. The output stream name is 'derivedColumn2'. The description is 'Creating/Updating the columns: SalesPrimarykey, Country, City, date, Units, PercentOfStandardCost, GrossProfit.' The incoming stream is 'derivedColumn1'. The 'Columns' section contains one column: 'GrossProfit' with expression 'TotalRevenue-TotalCost'.

Now add groupby and aggretate what client asked

The screenshot shows the Microsoft Azure Data Factory Data Flow designer interface. The pipeline is named 'denim_pl' and is part of the 'Dennis_DF' data factory. The pipeline consists of several stages:

- Stage 1:** 'getcategorydata' followed by a 'join' operation. The 'join' stage has three columns.
- Stage 2:** 'getsubcategorydata' followed by another 'join' operation. This stage also has three columns.
- Stage 3:** 'GetSalesData' followed by a 'join' operation. This stage has four columns.
- Stage 4:** 'join4' followed by a 'join' operation. This stage has five columns.
- Stage 5:** 'cleanedproductdata' followed by a 'jines' operation. This stage has six columns.
- Stage 6:** 'derivedColumn1' followed by 'derivedColumn2'. Both have two columns.

Aggregate settings:

- Output stream name:** AVGsales
- Description:** Aggregating data by 'date' producing columns 'DailyRevenue'
- Incoming stream:** derivedColumn2

Group by: date

The screenshot shows the Microsoft Azure Data Factory Data Flow designer interface, similar to the previous one but with the 'Aggregates' tab selected in the 'Group by' section.

Grouped by: date

Add **Clone** **Delete** **Open expression builder**

Column	Expression
DailyRevenue	sum(TotalRevenue)

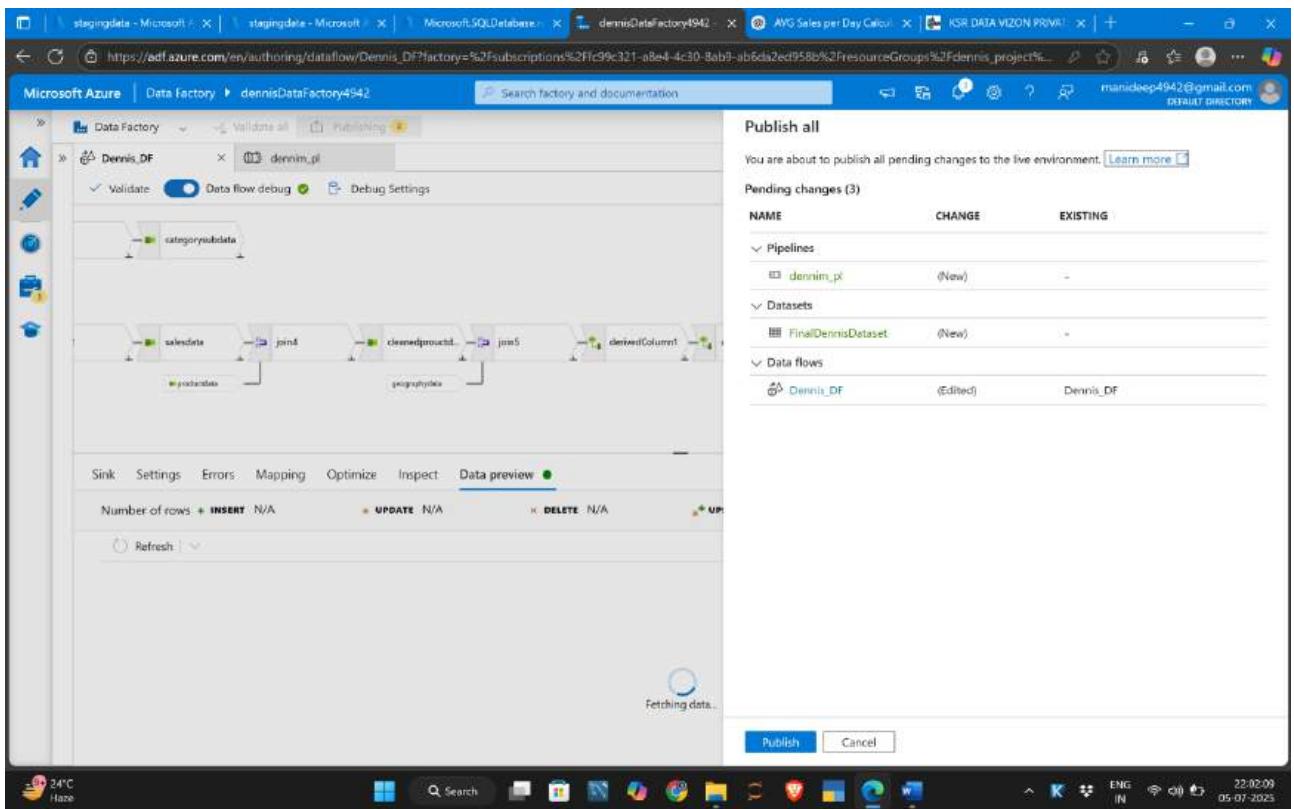
The screenshot shows the Microsoft Azure Data Factory Data Flow preview interface. A complex data pipeline is displayed, consisting of several stages: 'getcategorydata', 'join', 'categorysubcategory', 'getsubcategory...', 'GetSalesdata', 'join2', 'salesdata', 'join4', 'cleanedproducts...', 'join5', 'join6', 'derivedColumn1', 'derivedColumn2', and '2 Columns'. The 'Data preview' tab is selected, showing a table with two columns of data. The first column contains dates from 2017-06-07 to 2017-06-23, and the second column contains numerical values ranging from 14953.99999999998 to 8527.05.

Now the final we have to move themodified date into sql data base.

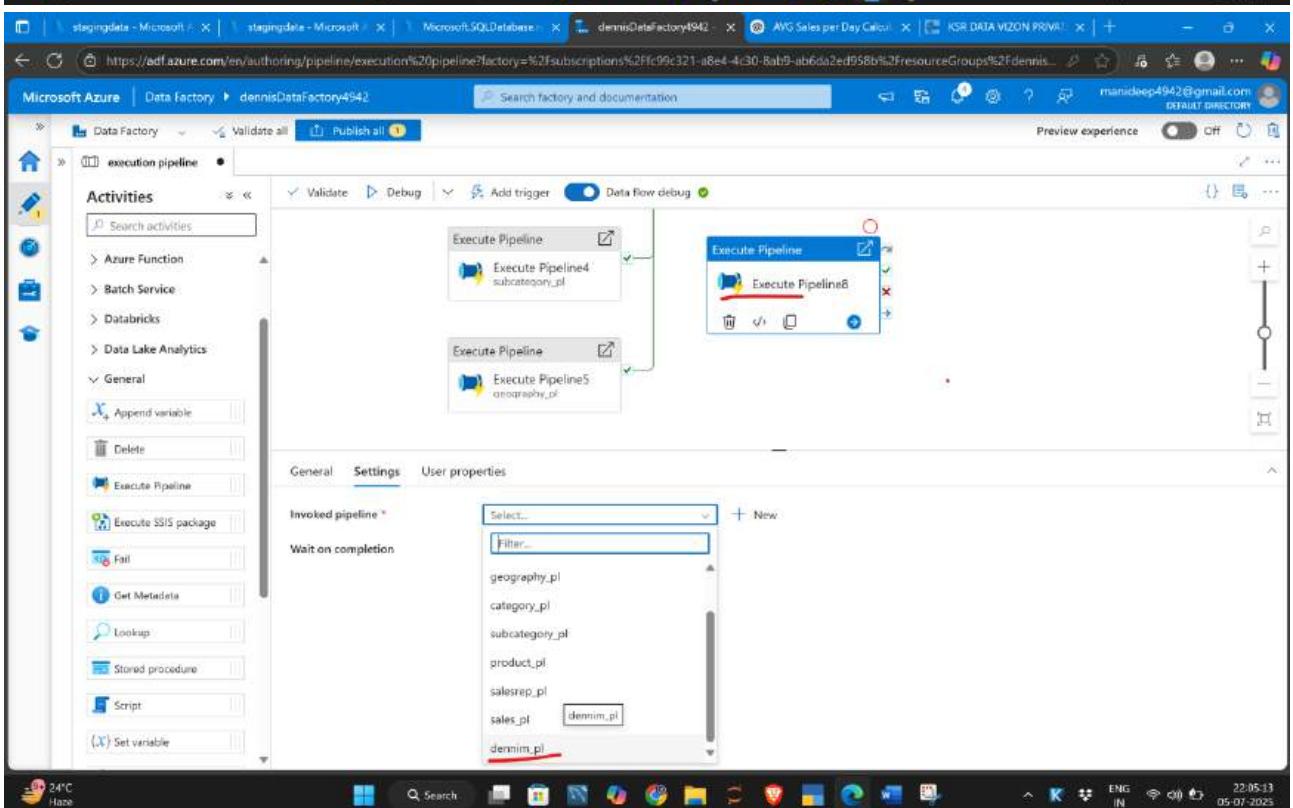
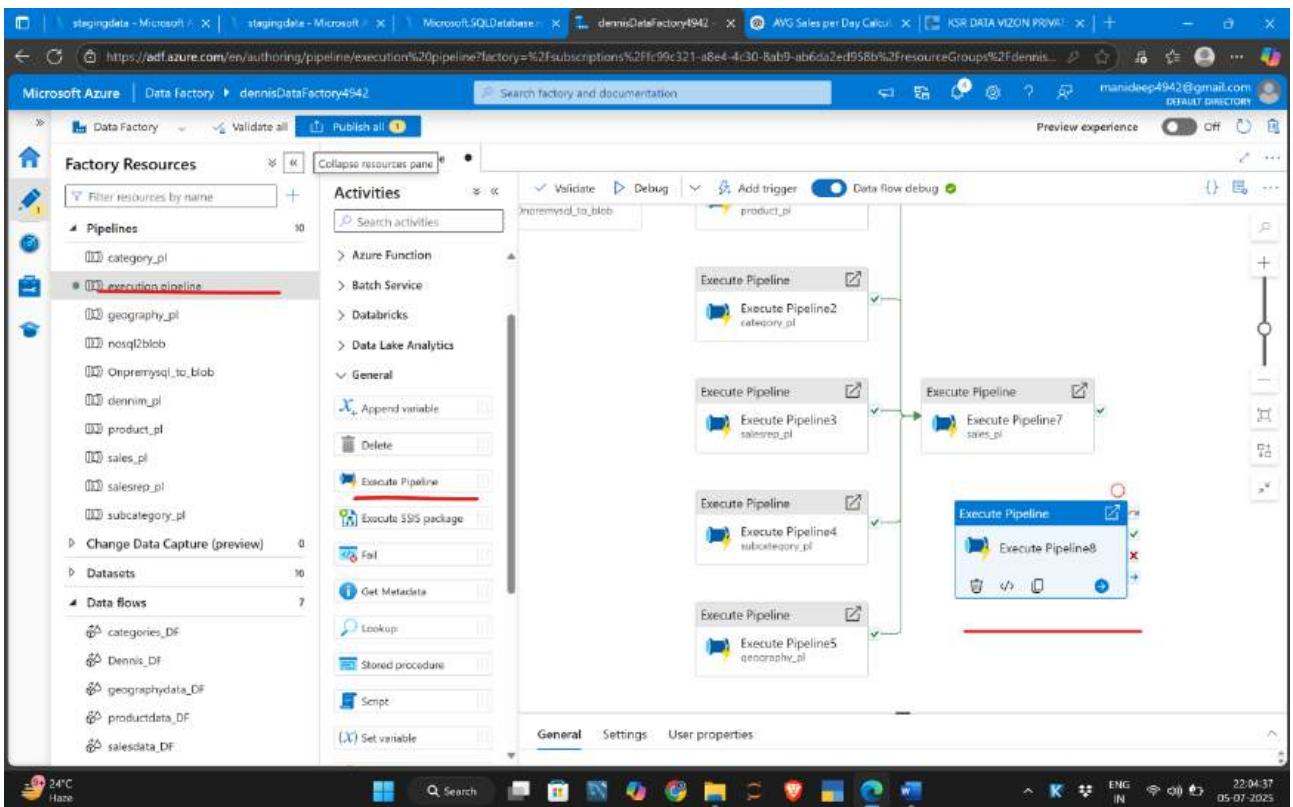
For this we have created the service

The screenshot shows the Microsoft Azure Data Factory Data Flow sink configuration. The 'Sink' tab is selected. The 'Description' field contains 'Add sink dataset.' The 'Incoming stream' dropdown is set to 'aggregate1'. Under 'Sink type', the 'Dataset' option is selected. In the 'Dataset' dropdown, 'Select...' is chosen. The 'Options' section includes checked checkboxes for 'Allow schema drift' and 'Validate schema'. On the right side of the screen, a 'New linked service' dialog is open for 'Azure SQL Database'. It shows the connection details: 'Server name' is 'maniserver4942', 'Database name' is 'manidatabase', 'Authentication type' is 'SQL authentication', 'User name' is 'maniadmin', and 'Password' is '*****'. The 'Create' button at the bottom of the dialog is highlighted. A status message at the bottom right says 'Connection successful'.

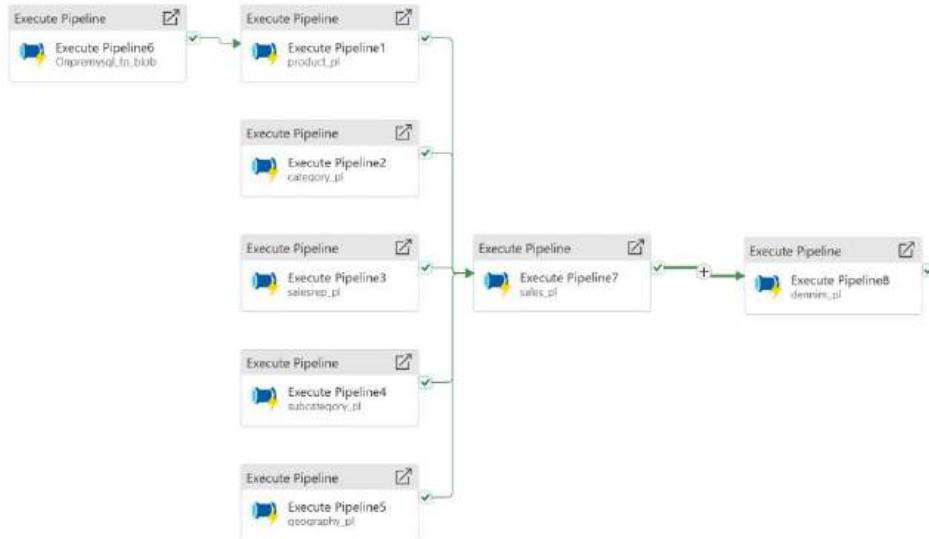
Do publish



Now go and add the dataflow to the execution pipeline run all the pipeline again



Do the formate pipelines in this formate to execute one by one



No do publish the and trigger the pipelines

The screenshot shows the Azure Data Factory pipeline editor interface. On the left, the pipeline structure is visible with various Execute Pipeline tasks. On the right, a 'Publish all' dialog box is open, displaying 'Pending changes (1)'. The table lists a single change for the 'execution pipeline'.

NAME	CHANGE	EXISTING
execution pipeline	(Edited)	execution pipeline

At the bottom of the dialog, there are 'Publish' and 'Cancel' buttons. The status bar at the bottom of the screen shows system information like weather (24°C Haze), date (05-07-2023), and time (22:06:35).

Microsoft Azure | Data Factory | dennisDataFactory4942 | Search factory and documentation

Pipeline run

Trigger pipeline now using last p Publishing Deploying changes to the factory

Parameters

Name	Type	Value
No records found		

OK Cancel

Validate Debug Add trigger Data flow debug

24°C Haze 22.07.08 05.07.2023

Microsoft Azure | Data Factory | dennisDataFactory4942 | Search factory and documentation

All pipeline runs > execution pipeline - Activity runs

Return Cancel Refresh Update pipeline List Gantt

Activity name	Activity start	Activity end	Run start	Duration	Integration runtime	User properties	Activity ID
Execute Pipeline4	In progress	Execute Pipeline	7/5/2025, 10:07:30 PM	Less than 1s		bc	
Execute Pipeline3	In progress	Execute Pipeline	7/5/2025, 10:07:30 PM	Less than 1s		80	
Execute Pipeline2	In progress	Execute Pipeline	7/5/2025, 10:07:30 PM	Less than 1s		35	
Execute Pipeline6	In progress	Execute Pipeline	7/5/2025, 10:07:30 PM	Less than 1s		78	
Execute Pipeline5	In progress	Execute Pipeline	7/5/2025, 10:07:30 PM	Less than 1s		8e	

Monitor in Azure Metrics Export to CSV

24°C Haze 22.07.30 05.07.2023

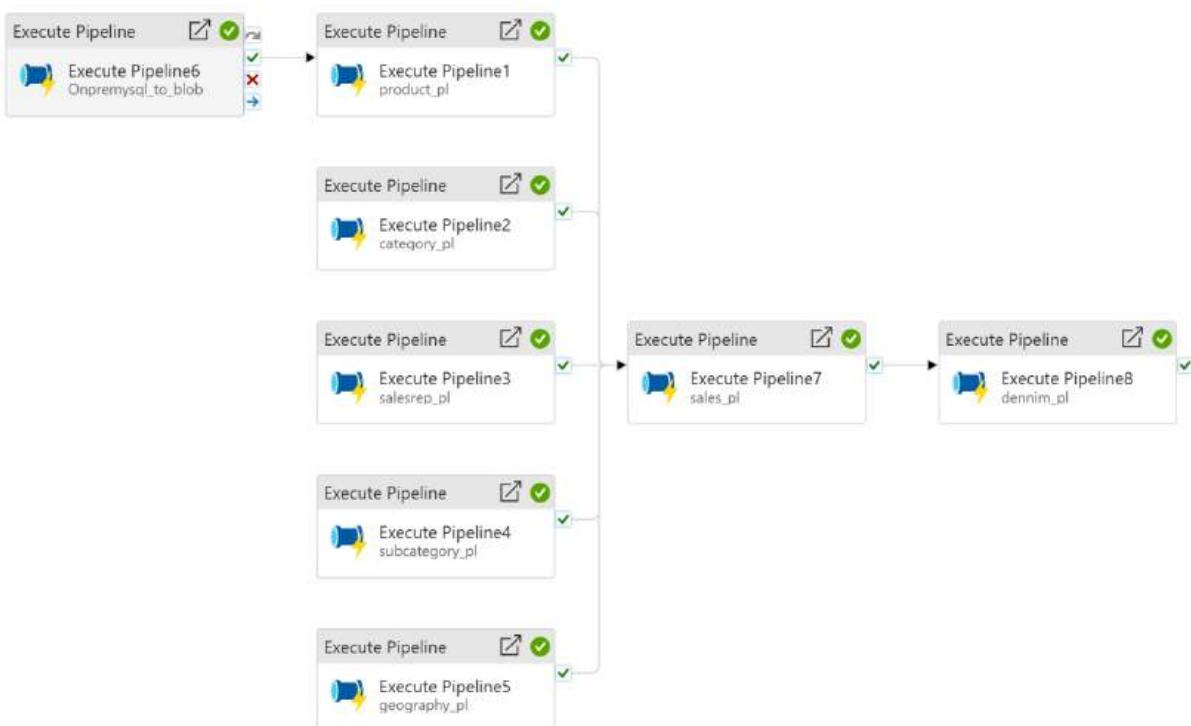
Screenshot of the Microsoft Azure Data Factory execution pipeline activity runs page.

The URL is: https://adf.azure.com/en/monitoring/pipelineruns/08c14766-4a1f-41a3-943d-08c302344877#factory=%2Fsubscriptions%2Fc99c321-a8e4-4c30-8eb9-ab6da2ed958b%2FresourceGroups%2Fadfdemo%2Fpipelines%2FexecutePipeline

The page shows a summary of pipeline runs and a detailed table of activity runs.

Activity runs:

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Actions
Execute Pipeline7	Succeeded	Execute Pipeline	7/5/2025, 10:11:02 PM	34s			5c
Execute Pipeline1	Succeeded	Execute Pipeline	7/5/2025, 10:11:54 PM	3m 8s			e0
Execute Pipeline4	Succeeded	Execute Pipeline	7/5/2025, 10:07:30 PM	3m 14s			bc
Execute Pipeline3	Succeeded	Execute Pipeline	7/5/2025, 10:07:30 PM	3m 8s			80
Execute Pipeline2	Succeeded	Execute Pipeline	7/5/2025, 10:07:30 PM	3m 12s			35
Execute Pipeline6	Succeeded	Execute Pipeline	7/5/2025, 10:07:30 PM	24s			78
Execute Pipeline5	Succeeded	Execute Pipeline	7/5/2025, 10:07:30 PM	3m 11s			8e



See here we have moved data into the sql server

The screenshot shows the Microsoft Azure portal interface. The left sidebar is open, showing various database management options like Overview, Activity log, Tags, Diagnose and solve problems, and the selected 'Query editor (preview)'. The main area displays a 'Query 1' window with a message: 'Showing limited object explorer here. For full capability please click here to open Azure Data Studio.' Below this, there's a tree view of database objects under 'Tables', showing 'dbo.Dennis_Tb'. The bottom right of the main area has tabs for 'Results' and 'Messages', and a search bar. The status bar at the bottom shows the date and time as 05-07-2025.

This screenshot shows the same Azure portal setup as the first one, but with two queries open: 'Query 1' and 'Query 2'. The 'Query 2' window contains the following SQL code:

```
1 SELECT TOP (1000) * FROM [dbo].[Dennis_Tb]
```

The results tab shows a table with the following data:

Color	RetailPrice	Category	Town	TotalRevenue	TotalCost	GrossProfit
Fluorescent Pink	29.95	Special	Frankfurt	1797	549	1248
Blue	26.95	Special	Copenhagen	1913.45	585.75	1327.7
Green	26.95	Special	Dresden	1832.6	561	1271.6
Blue	23.95	General	Berlin	3760.15	1185.35	2574.8
Blue	26.95	General	Amsterdam	2275.85	870.75	1402.1

Here we go completed project very successfully

Thankyou