

# PROJECT

## Data Integration & Transformation

- we have 3 tables customer, product, and sales. Here I choose retail domain you can take any domain like health care, and financial data etc.
- Customer.sql table in azure SQL database, product.json table in blob storage, and sales dbo table in an on-premises SQL server database.
- Here I choose data lake is a unified storage system. So we have to ingest data from these sources to a data lake and transform different data file formats to parquet file format.
- So we can easily analyze the data and generate insights from data.
- The tech stack I have used for this azure data factory, azure data lake, azure SQL database, SQL server.

**Azure data factory :-** Azure Data Factory is a cloud-based data integration and ETL (Extract, Transform, Load) service by Microsoft for orchestration and automation of data integration and transformation.

**Azure data lake:-** Azure Data Lake is a cloud-based big data storage and analytics service by Microsoft that allows storing, processing, and analyzing large amounts of structured, semi-structured, and unstructured data.

**Azure SQL database:-** Azure SQL Database is a fully managed relational database service by Microsoft that provides a scalable and secure platform for storing, processing, and managing structured data, based on the SQL Server engine.

**Azure blob storage:-** Azure Blob Storage is a cloud-based object storage service by Microsoft for unstructured data such as text and binary data, images, audio and video files, which can be accessed via HTTP/HTTPS from anywhere in the world.

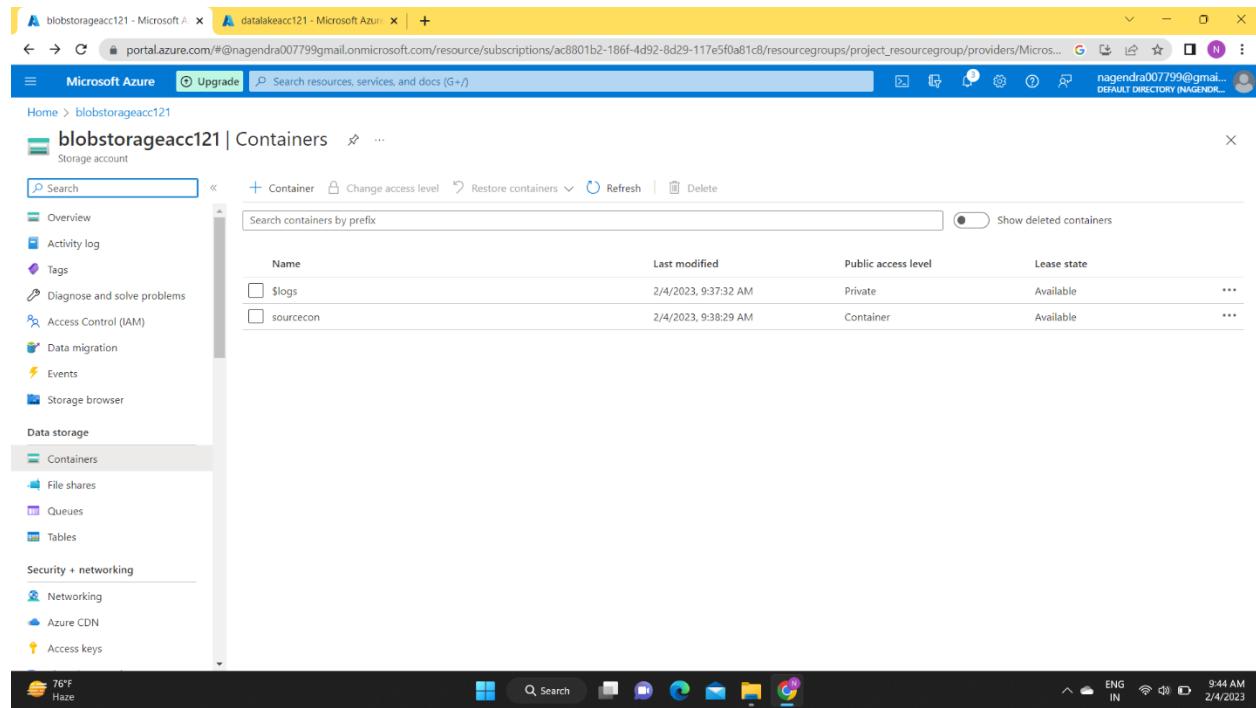
**Data integration:-** Data integration is the process of combining data from various sources into a unified, centralized repository for analysis and reporting.

**Data transformation:-** Data transformation is the process of converting data from one format or structure to another to make it usable for analysis or reporting.

**Linked service:-** A linked service in Azure Data Factory is a connection entity that defines the relationship between the data factory and an external data store, such as a database, file system, or cloud storage. It contains the connection details and authentication information needed to access the external data store.

**Data set:-** A dataset in Azure Data Factory represents the structure and metadata of data stored in a data store, such as a database table or file system, and is used as the basis for defining data transformations and movement activities in data pipelines.

- This is my blob storage “blobstorageacc121” and container name ‘sourcecon’.

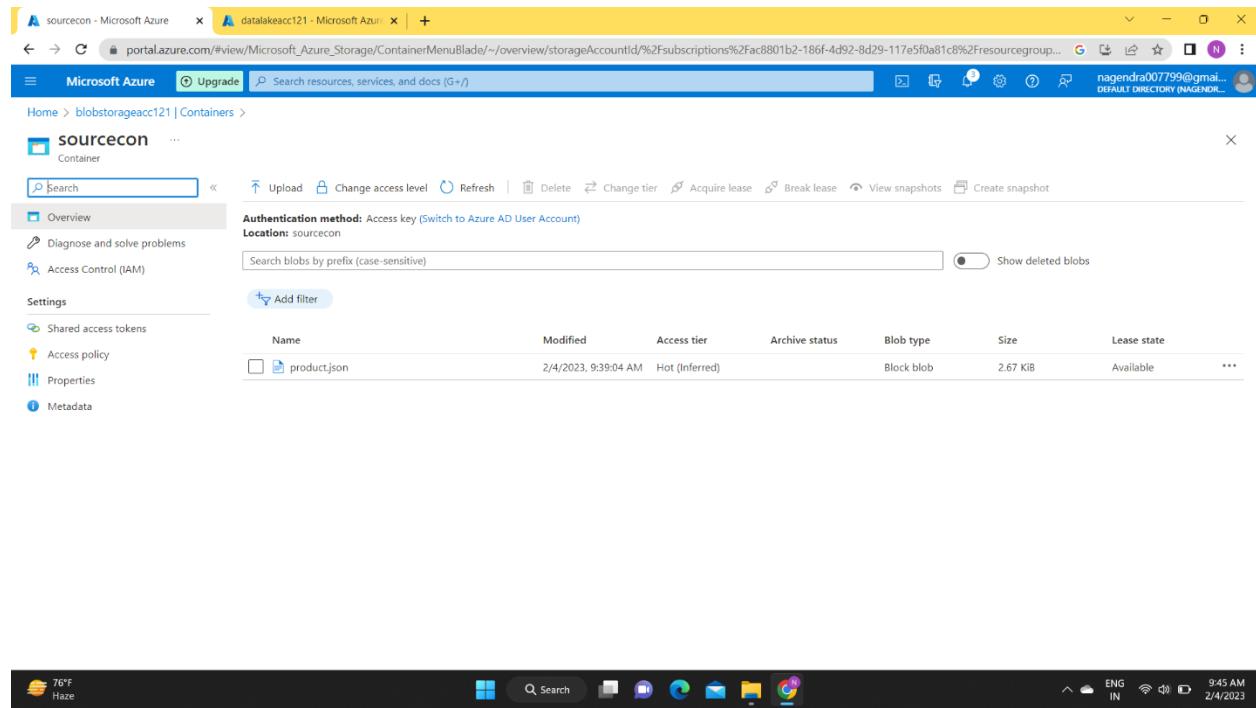


The screenshot shows the Microsoft Azure portal interface. The top navigation bar has two tabs: 'blobstorageacc121 - Microsoft Azure' and 'datalakeacc121 - Microsoft Azure'. The URL in the address bar is 'portal.azure.com/#/blobstorageacc121'. The left sidebar shows 'blobstorageacc121 | Containers' under 'Storage account'. The main content area displays a table of containers:

Name	Last modified	Public access level	Lease state
slogs	2/4/2023, 9:37:32 AM	Private	Available
sourcecon	2/4/2023, 9:38:29 AM	Container	Available

The bottom status bar shows the date and time as '2/4/2023' and '9:44 AM'.

- In this blob storage, sourcecon we have product table. The product table is in json file format.

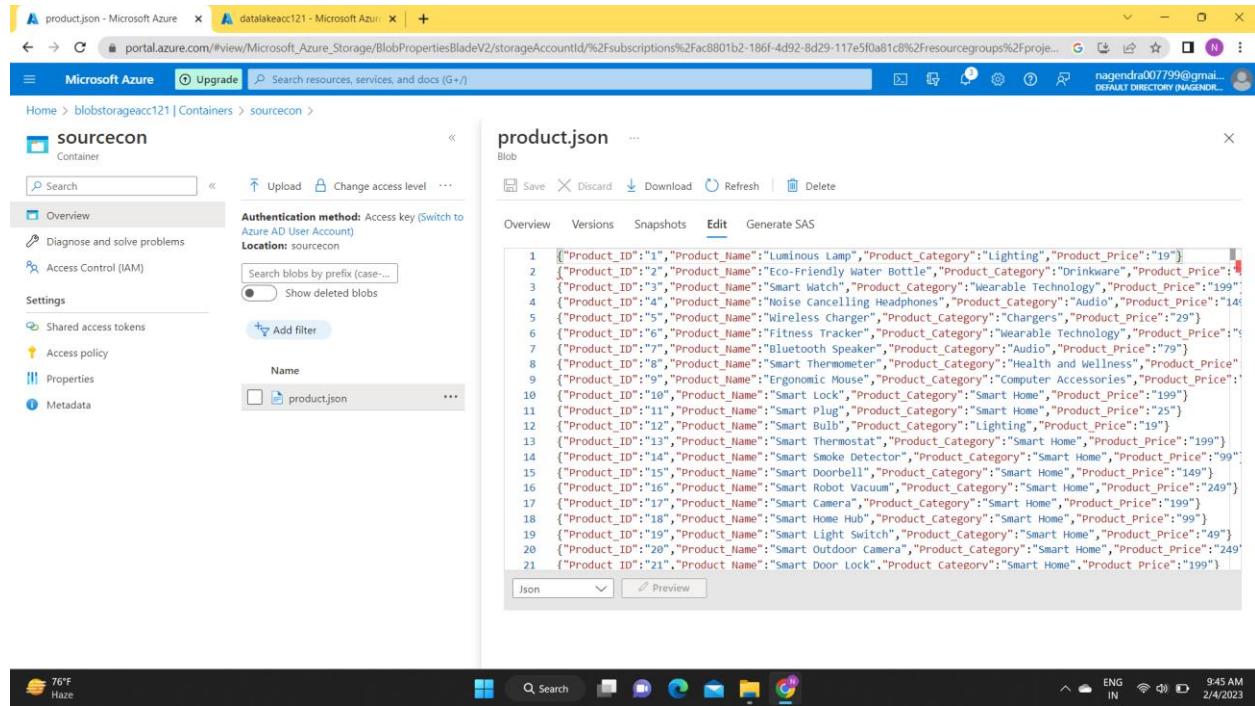


The screenshot shows the Microsoft Azure portal interface. The top navigation bar has two tabs: 'sourcecon - Microsoft Azure' and 'datalakeacc121 - Microsoft Azure'. The URL in the address bar is 'portal.azure.com/#/blobstorageacc121/sourcecon'. The left sidebar shows 'sourcecon' under 'Containers'. The main content area displays a table of blobs:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
product.json	2/4/2023, 9:39:04 AM	Hot (Inferred)		Block blob	2.67 KB	Available

The bottom status bar shows the date and time as '2/4/2023' and '9:45 AM'.

- Here we can see the data in the product table



The screenshot shows the Azure Storage Blob browser interface. On the left, the 'sourcecon' container is selected. In the center, the 'product.json' file is being viewed. The file content is a JSON array of 21 objects, each representing a product with fields like Product\_ID, Product\_Name, Product\_Category, and Product\_Price. The JSON is displayed in a code editor-like view with line numbers. Below the code, there are 'Json' and 'Preview' buttons. The browser has a standard Windows-style header with tabs for 'product.json' and 'datalakeacc121 - Microsoft Azure'. The status bar at the bottom shows system information like battery level, network, and date.

```

1  {"Product_ID": "1", "Product_Name": "Luminous Lamp", "Product_Category": "Lighting", "Product_Price": "199"}  

2  {"Product_ID": "2", "Product_Name": "Eco-Friendly Water Bottle", "Product_Category": "Drinkware", "Product_Price": "199"}  

3  {"Product_ID": "3", "Product_Name": "Smart Watch", "Product_Category": "Wearable Technology", "Product_Price": "199"}  

4  {"Product_ID": "4", "Product_Name": "Noise Cancelling Headphones", "Product_Category": "Audio", "Product_Price": "149"}  

5  {"Product_ID": "5", "Product_Name": "Wireless Charger", "Product_Category": "Chargers", "Product_Price": "29"}  

6  {"Product_ID": "6", "Product_Name": "Fitness Tracker", "Product_Category": "Wearable Technology", "Product_Price": "199"}  

7  {"Product_ID": "7", "Product_Name": "Bluetooth Speaker", "Product_Category": "Audio", "Product_Price": "79"}  

8  {"Product_ID": "8", "Product_Name": "Smart Thermometer", "Product_Category": "Health and Wellness", "Product_Price": "199"}  

9  {"Product_ID": "9", "Product_Name": "Ergonomic Mouse", "Product_Category": "Computer Accessories", "Product_Price": "199"}  

10 {"Product_ID": "10", "Product_Name": "Smart Lock", "Product_Category": "Smart Home", "Product_Price": "199"}  

11 {"Product_ID": "11", "Product_Name": "Smart Plug", "Product_Category": "Smart Home", "Product_Price": "29"}  

12 {"Product_ID": "12", "Product_Name": "Smart Bulb", "Product_Category": "Lighting", "Product_Price": "199"}  

13 {"Product_ID": "13", "Product_Name": "Smart Thermostat", "Product_Category": "Smart Home", "Product_Price": "199"}  

14 {"Product_ID": "14", "Product_Name": "Smart Smoke Detector", "Product_Category": "Smart Home", "Product_Price": "99"}  

15 {"Product_ID": "15", "Product_Name": "Smart Doorbell", "Product_Category": "Smart Home", "Product_Price": "149"}  

16 {"Product_ID": "16", "Product_Name": "Smart Robot Vacuum", "Product_Category": "Smart Home", "Product_Price": "249"}  

17 {"Product_ID": "17", "Product_Name": "Smart Camera", "Product_Category": "Smart Home", "Product_Price": "199"}  

18 {"Product_ID": "18", "Product_Name": "Smart Home Hub", "Product_Category": "Smart Home", "Product_Price": "99"}  

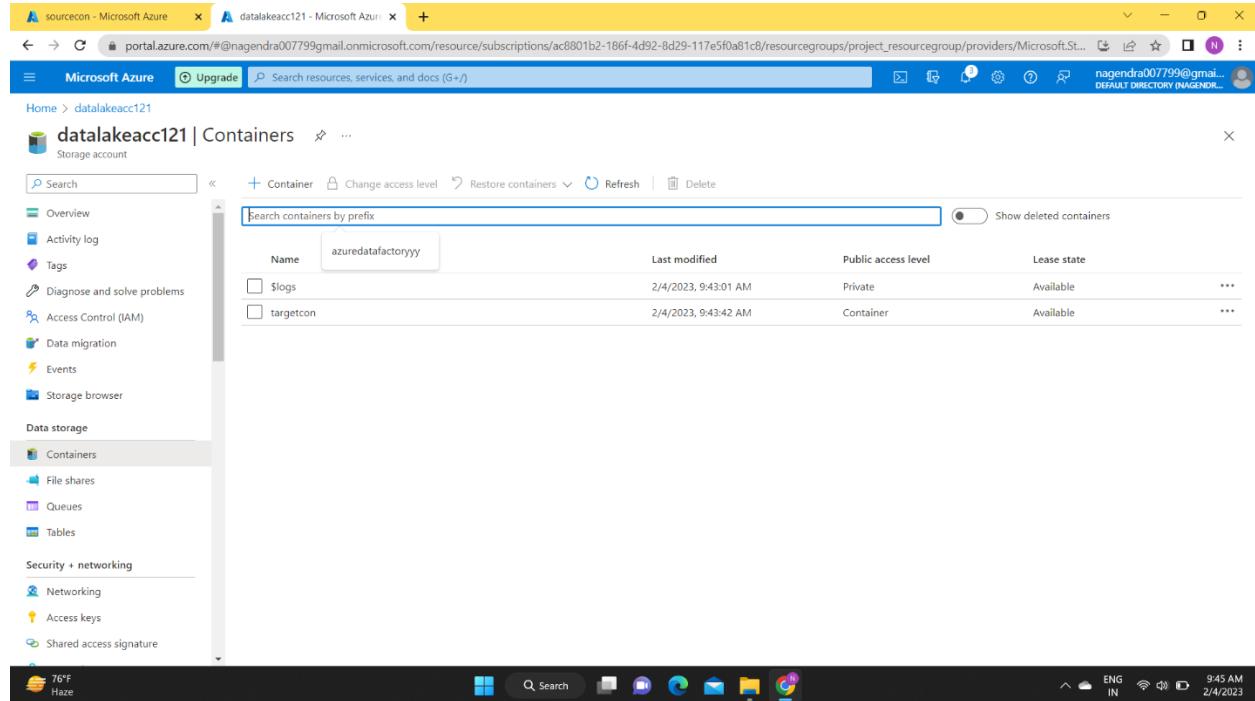
19 {"Product_ID": "19", "Product_Name": "Smart Light Switch", "Product_Category": "Smart Home", "Product_Price": "49"}  

20 {"Product_ID": "20", "Product_Name": "Smart Outdoor Camera", "Product_Category": "Smart Home", "Product_Price": "249"}  

21 {"Product_ID": "21", "Product_Name": "Smart Door Lock", "Product_Category": "Smart Home", "Product_Price": "199"}

```

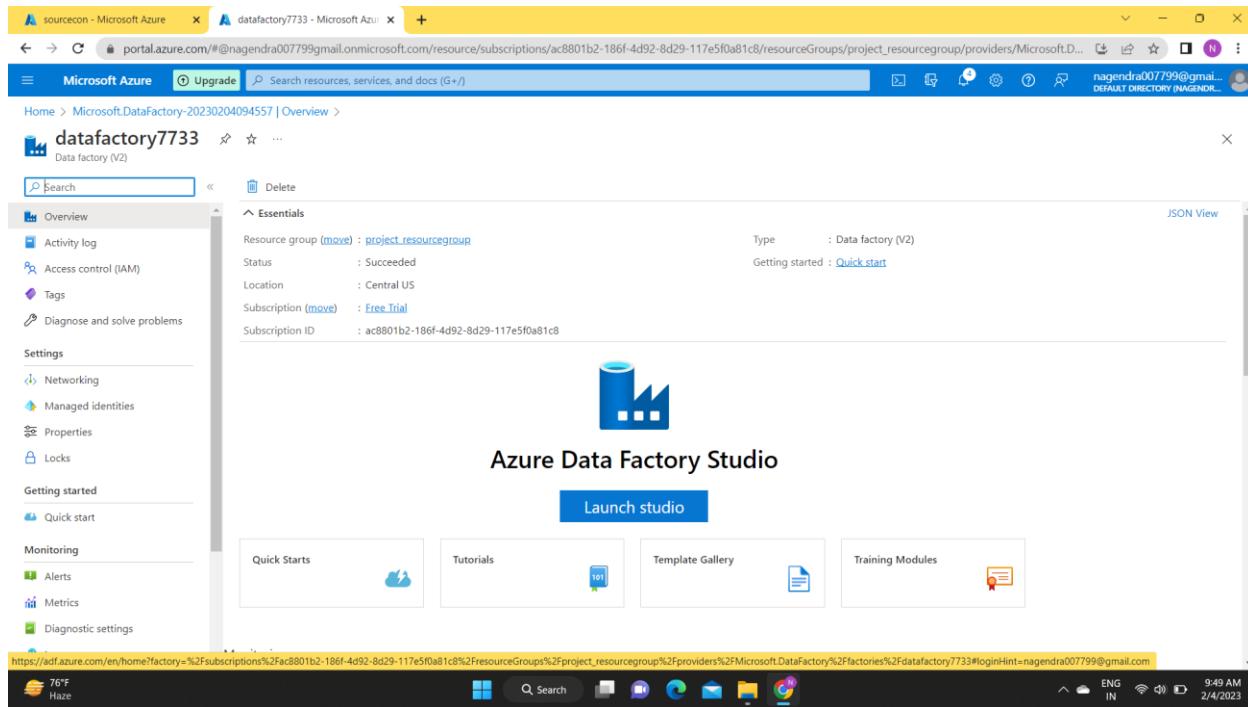
- Here data lake is our unified storage account
- Data lake name is "datalakeacc121" and container name is 'targetcon'



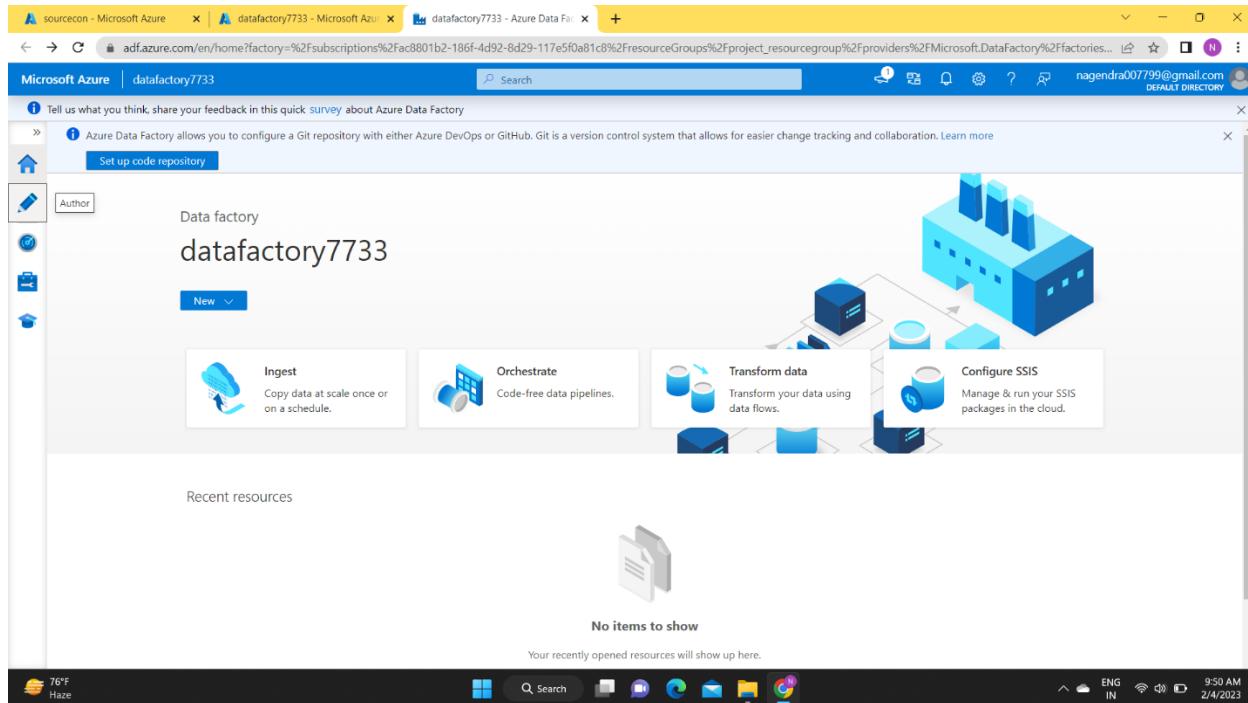
The screenshot shows the Azure Storage Blob browser interface. On the left, the 'datalakeacc121' container is selected. In the center, the 'Containers' section is displayed, showing a list of containers. One container, 'targetcon', is selected and highlighted. The table includes columns for Name, Last modified, Public access level, and Lease state. The status bar at the bottom shows system information like battery level, network, and date.

Name	Last modified	Public access level	Lease state
targetcon	2/4/2023, 9:43:42 AM	Container	Available

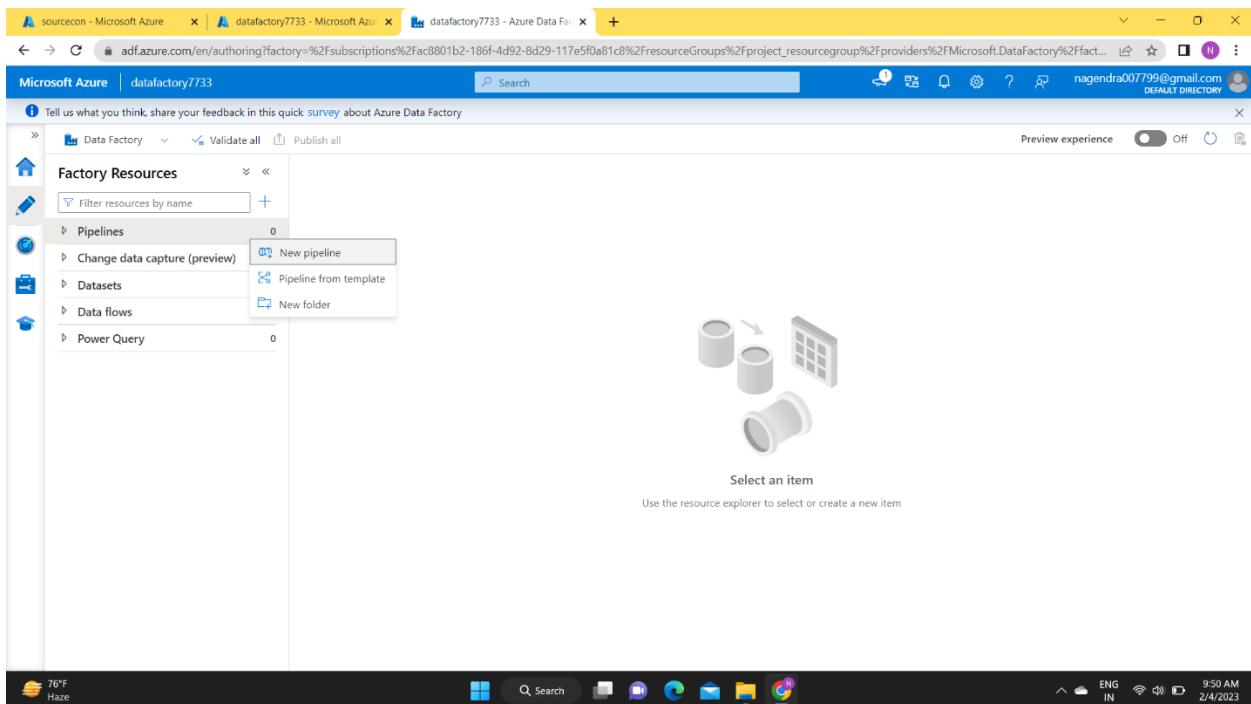
- Our azure data factory name is “datafactory7733”
- We are launching our datafactory by clicking “launch studio”



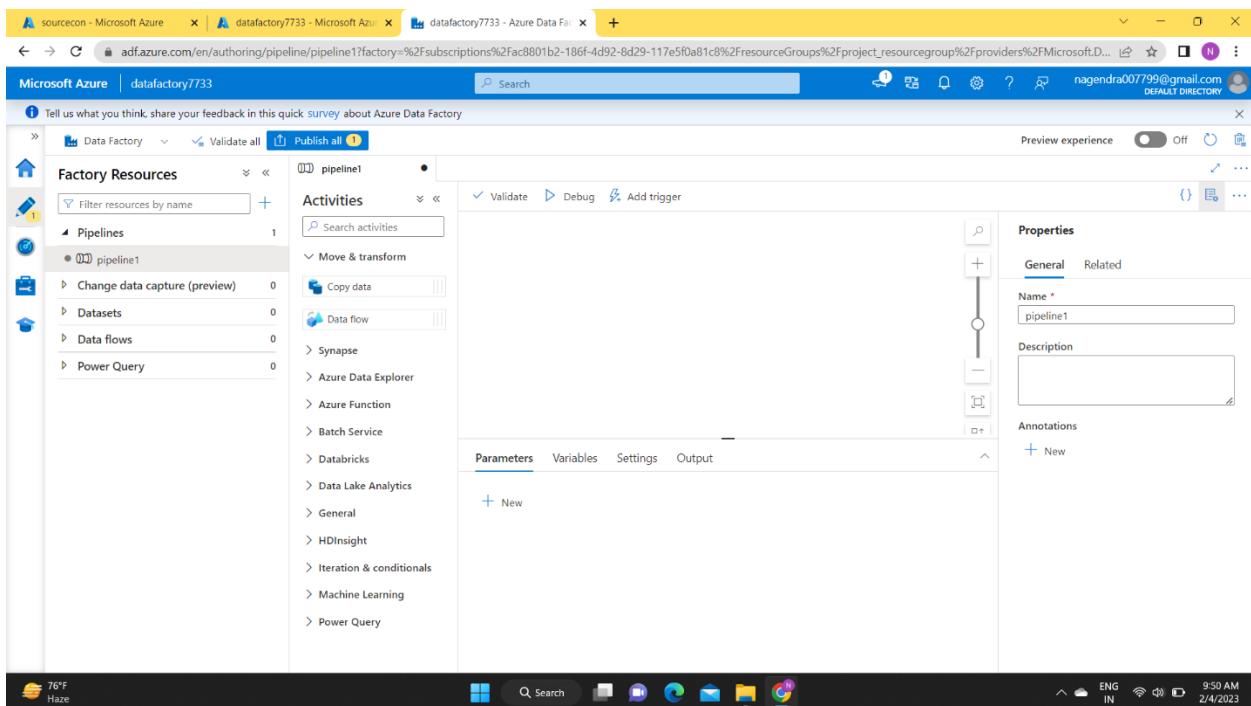
- Go to author tab



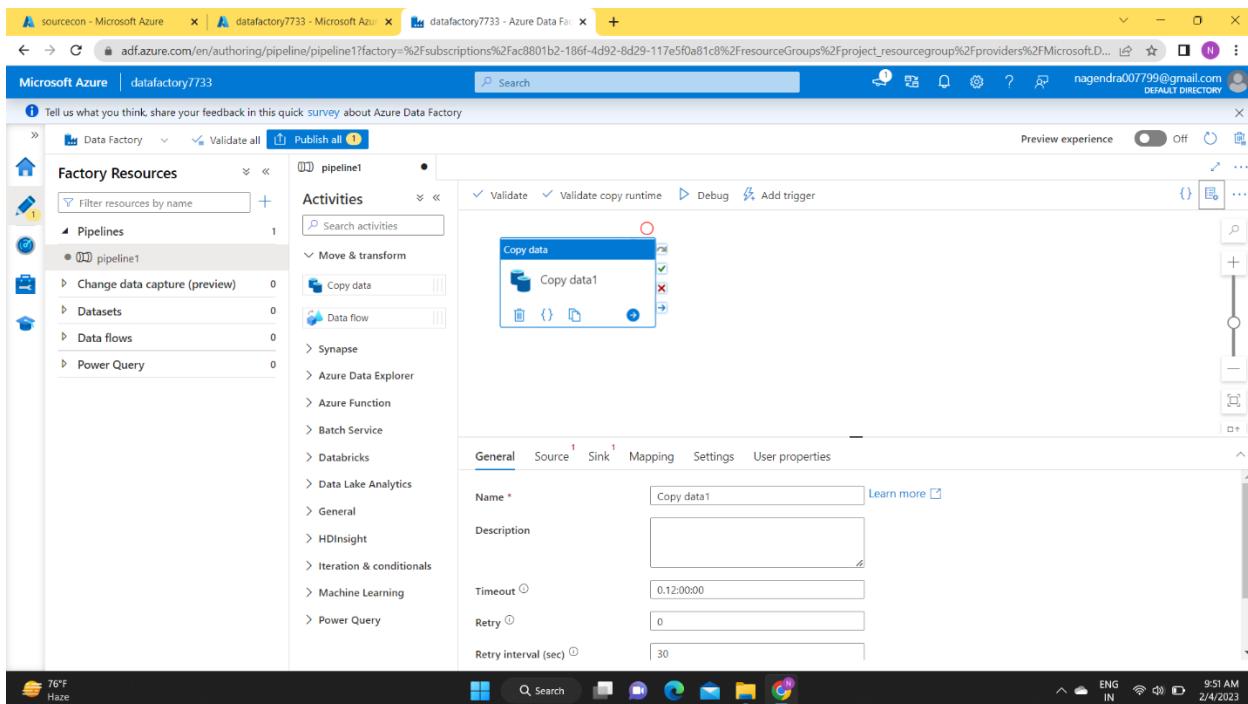
- In author go to pipeline and select new pipeline



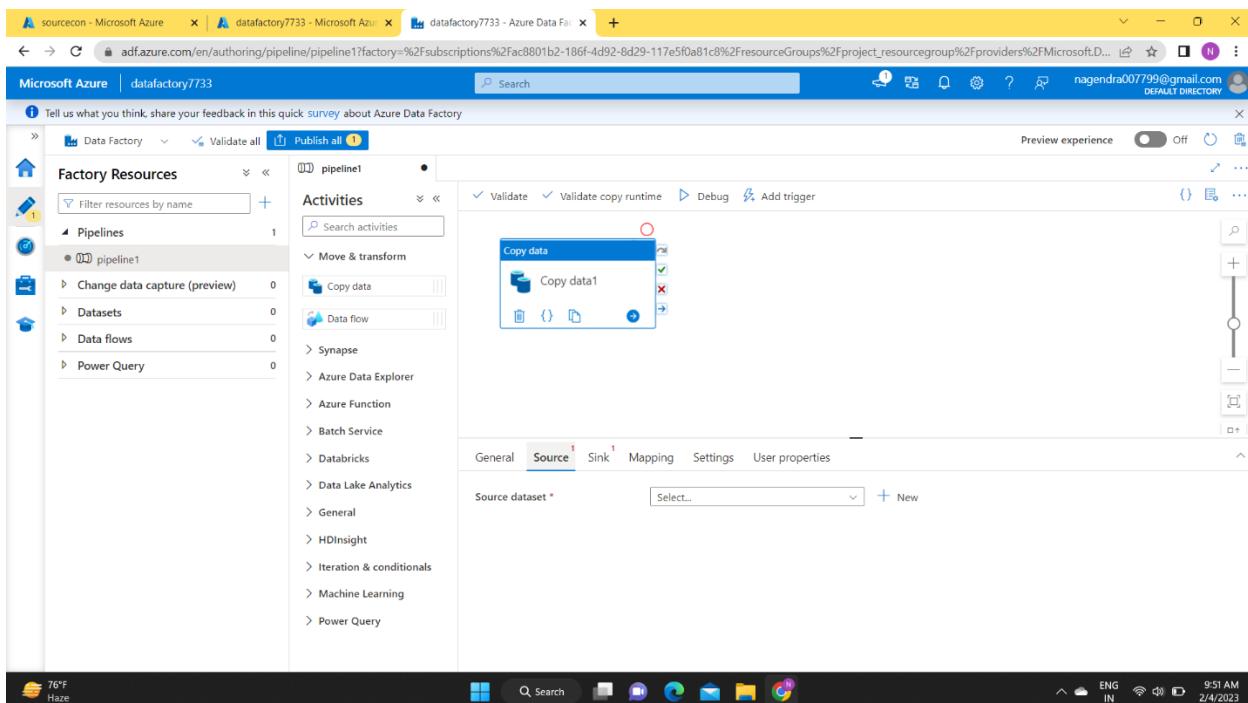
- Launch new pipeline
- Default pipeline name as “pipeline1”



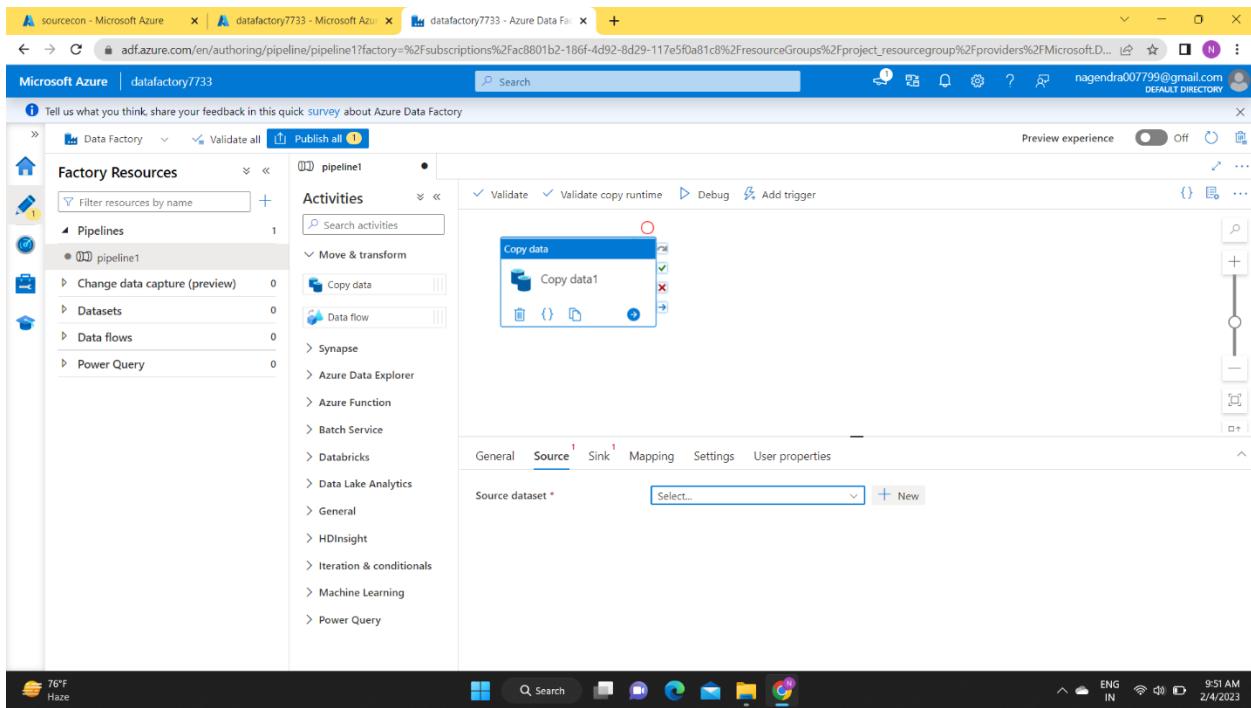
- In move and transform select copy data activity



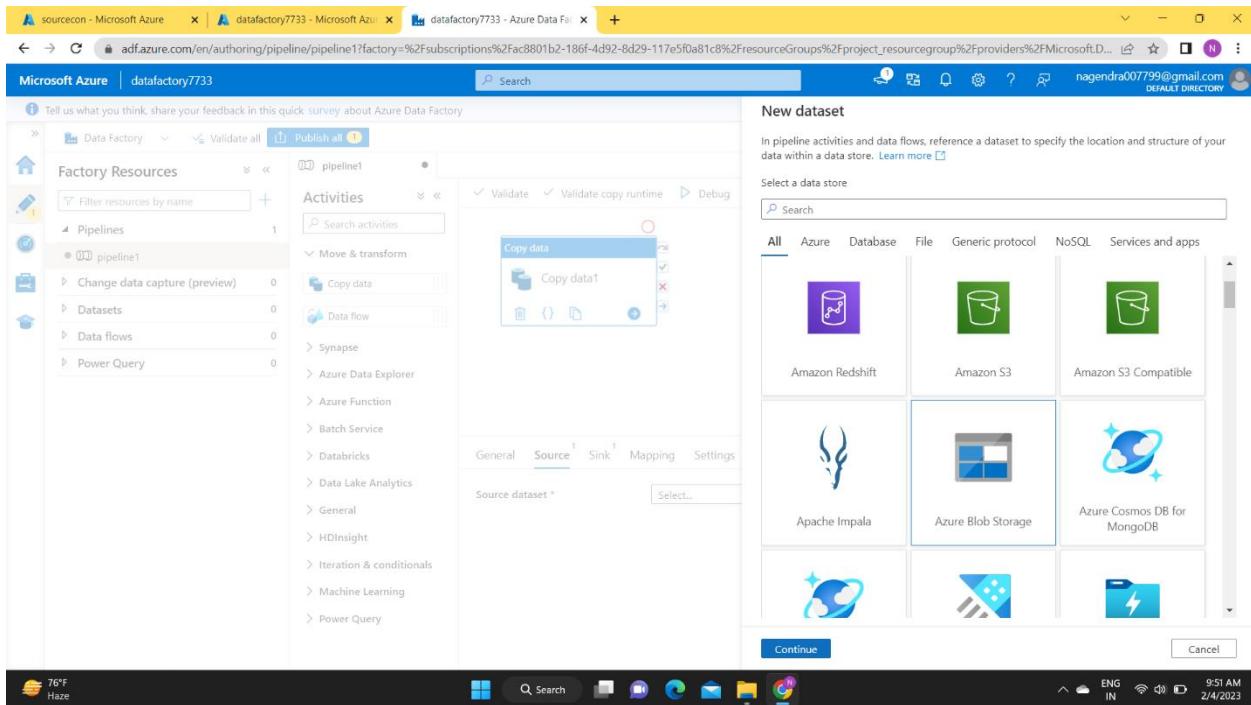
- Click on copy data and select source



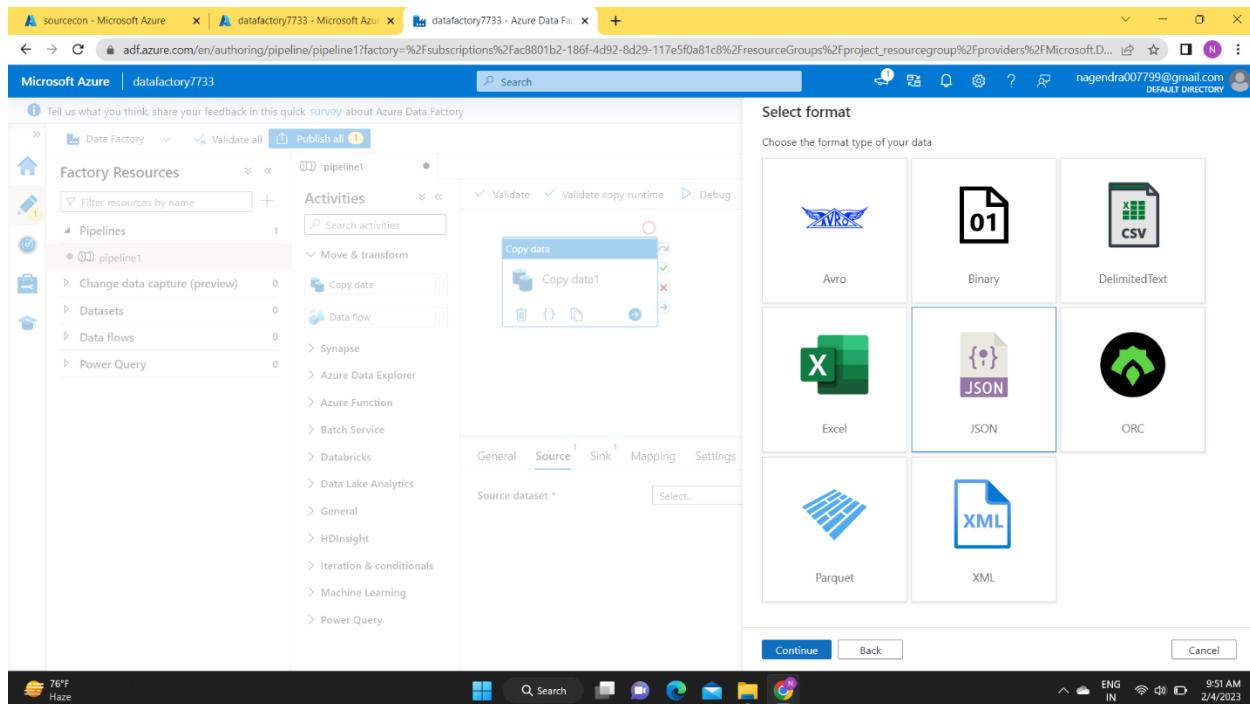
- In dataset box select +new



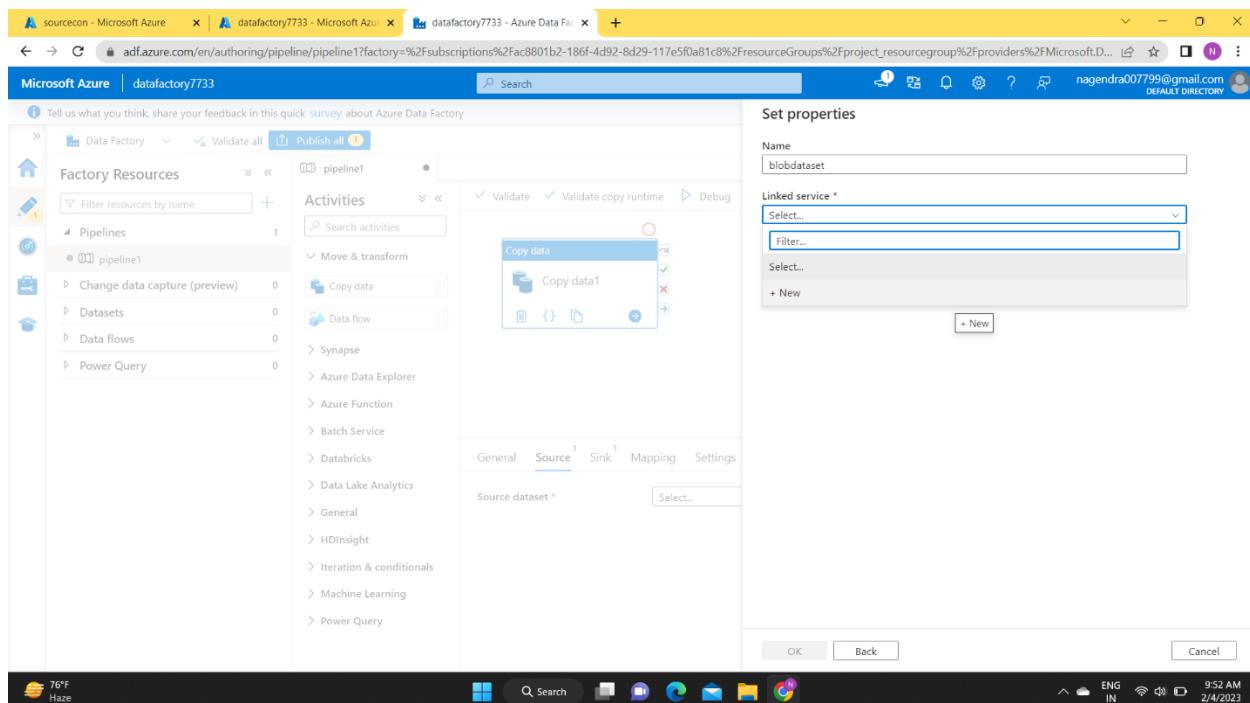
- Select the source storage type
- Here our source is blob storage and press continue.



- After the source storage type we have to choose the source file format type
- Our source file format type is json.

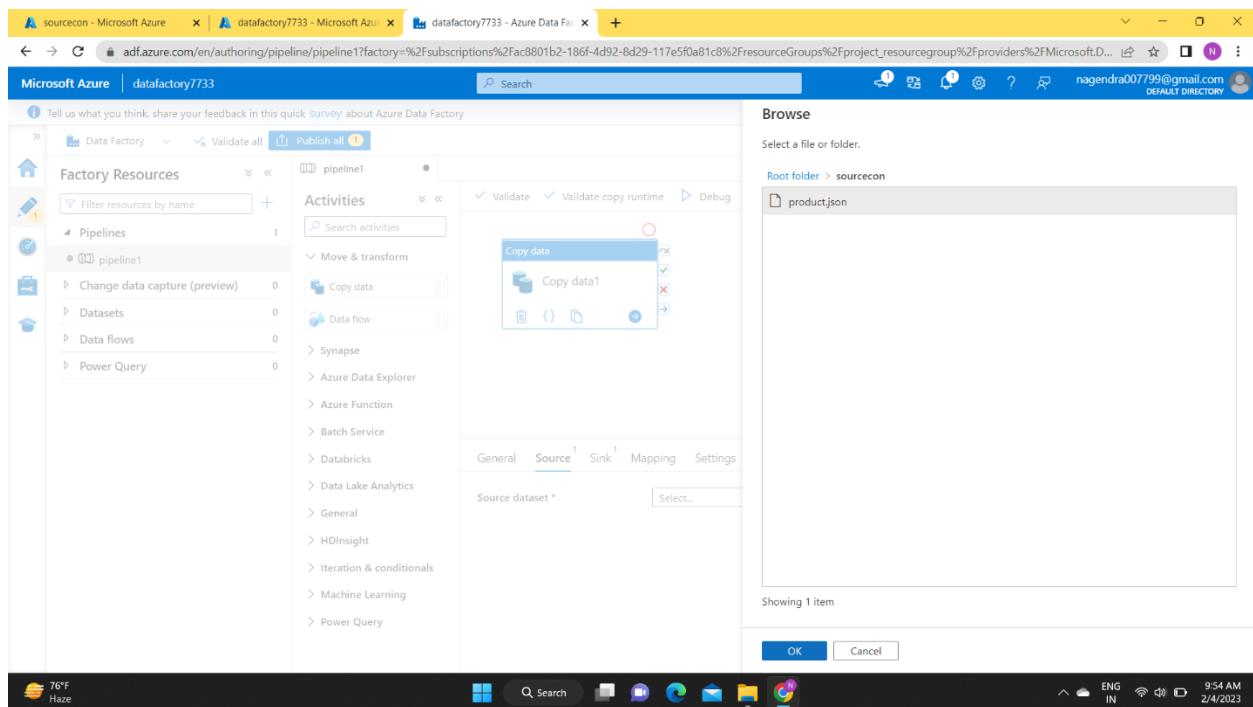
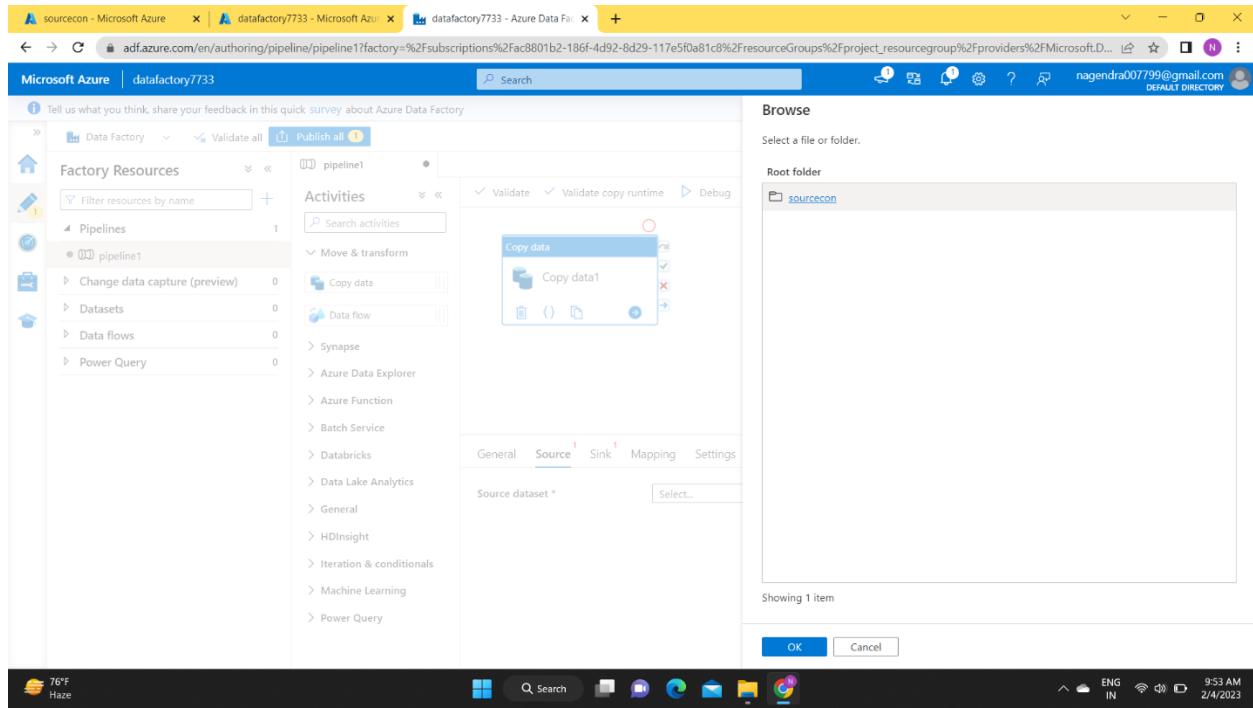


- In data set we have to give link service
- So we are creating new link service by clicking +new

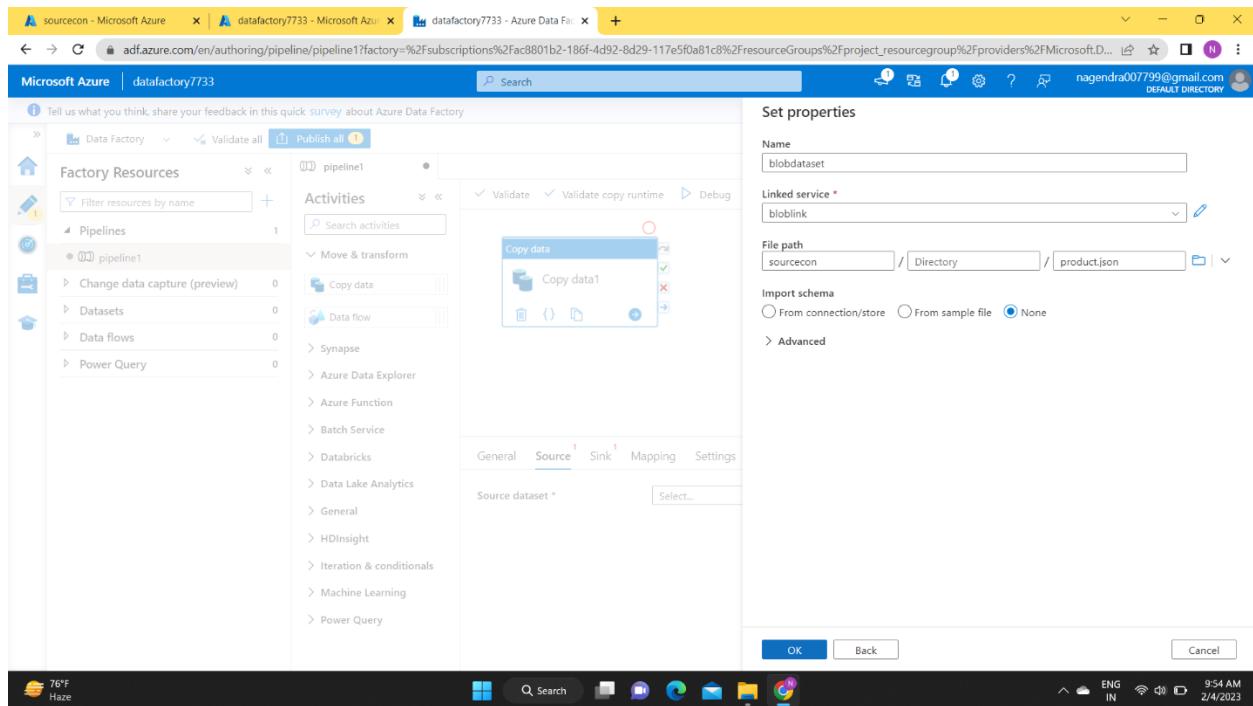


- Here you give name of your link service “bloblink”, and give azure subscription and storage account name.
- Must do the test connection.

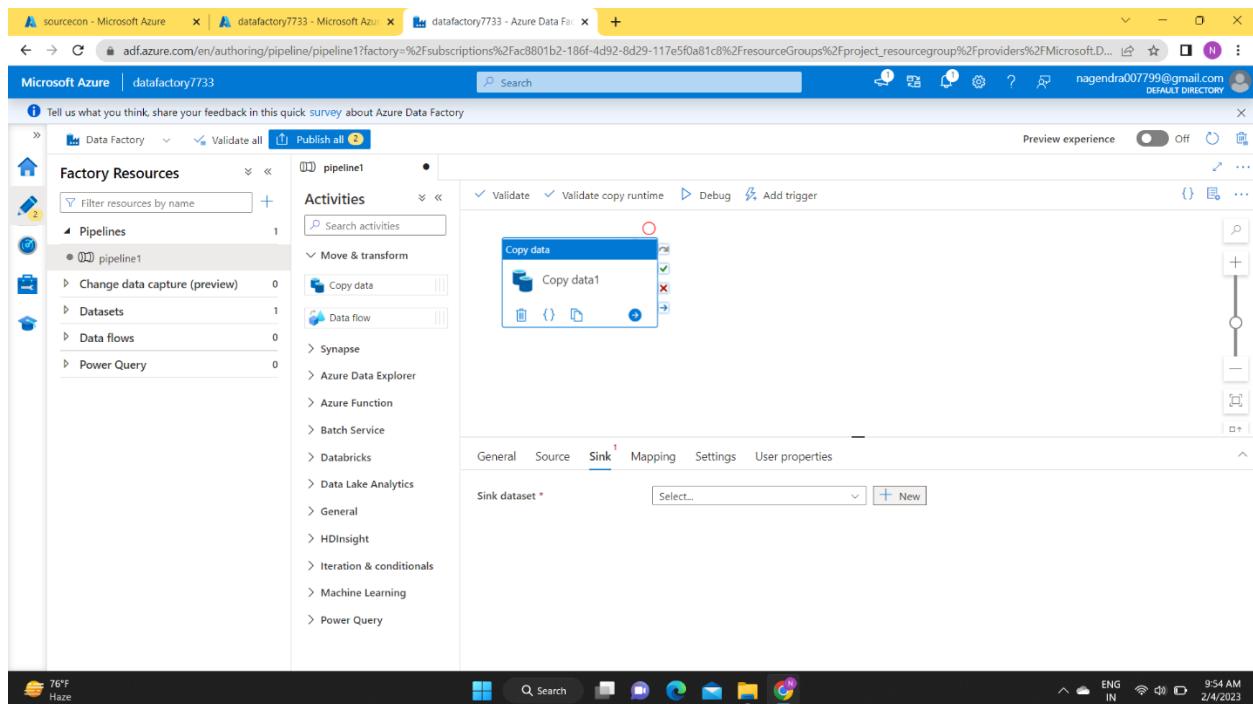
- After that select the source file location.



- Don't select "from connection" in "import schema" put it "none"
- Because we are transferring data from "json to parquet" file format.



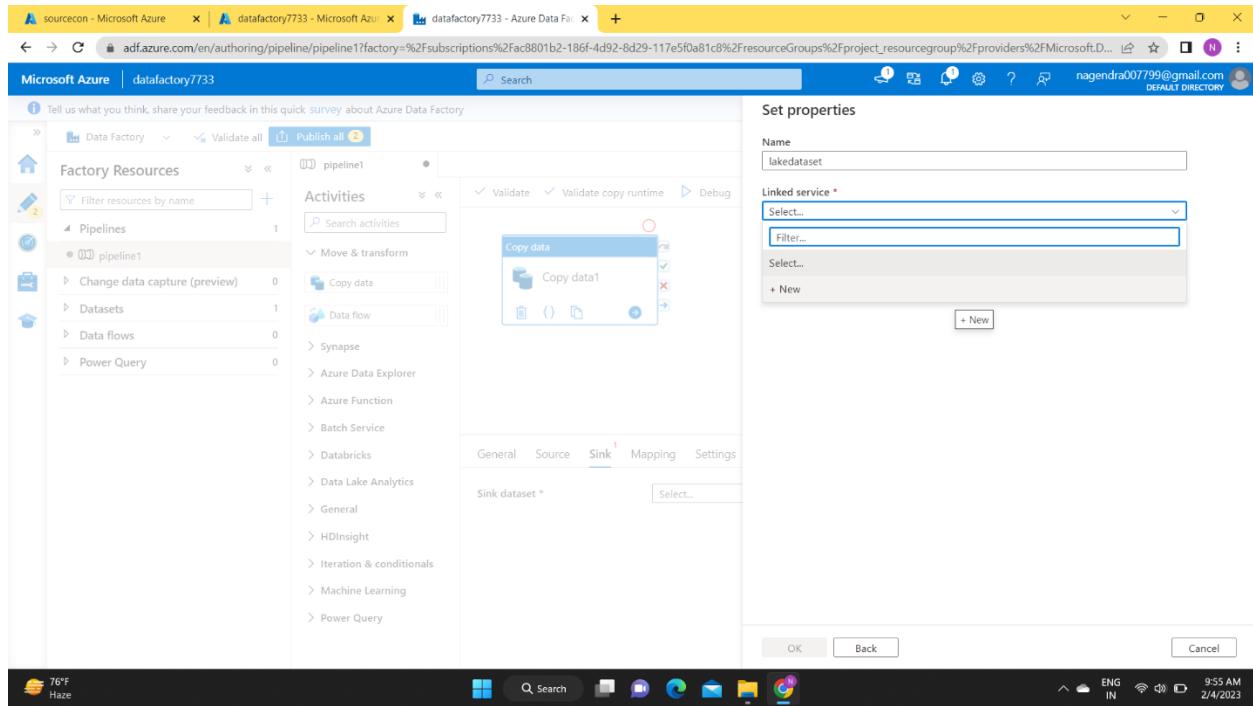
- Now configure sink (target or unified storage)



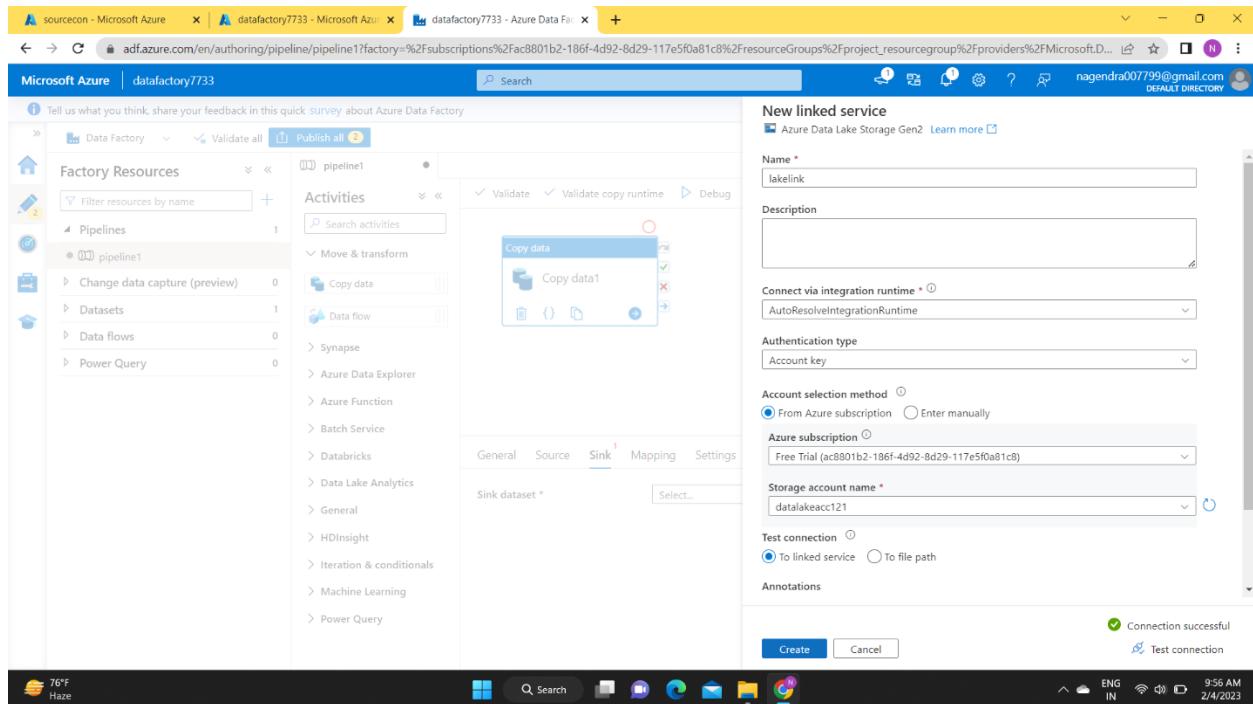
- Our target is datalake, so select datalake.

- Our final file format is parquet, so select parquet.

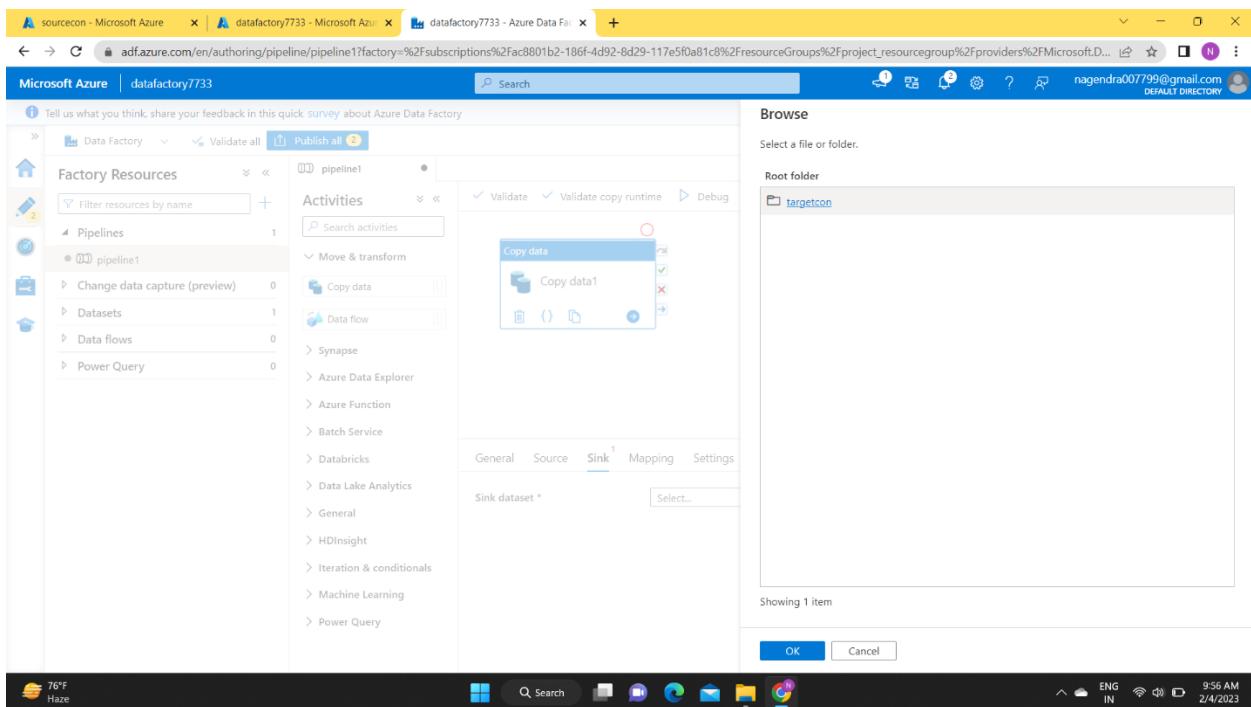
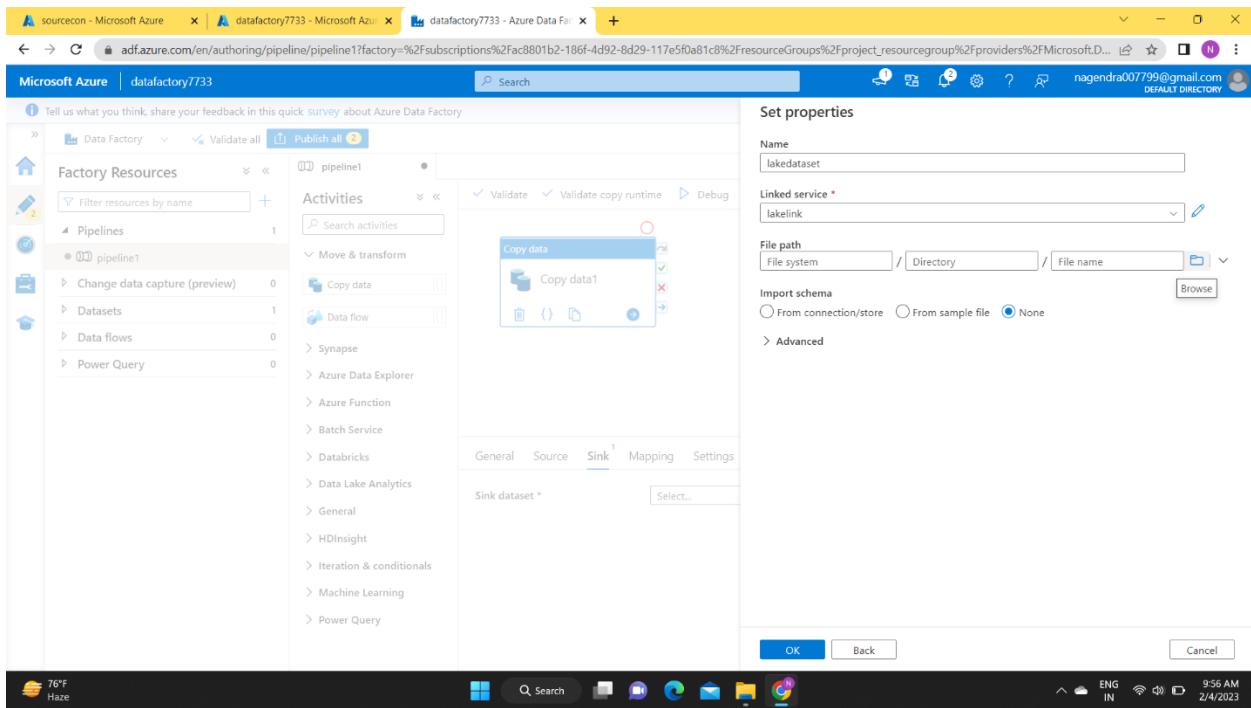
- Our dataset name is “lakedataset” and create a new link service for our target datalake.



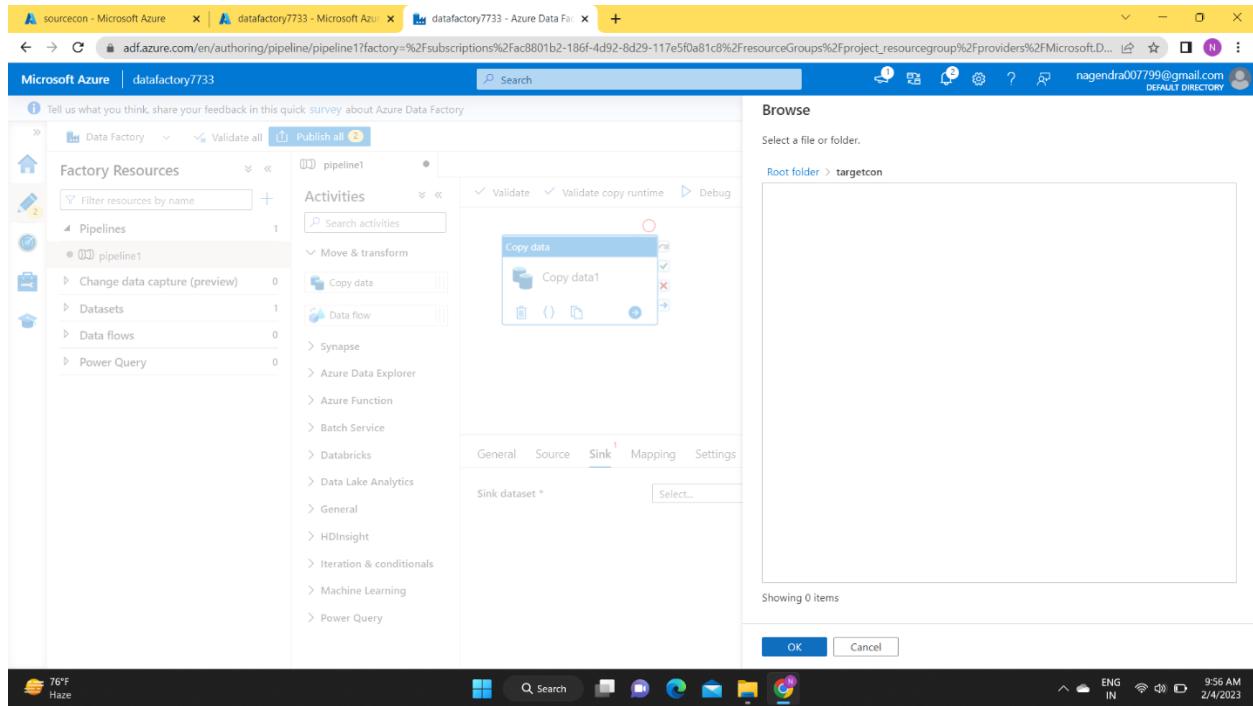
- Our target(data lake) link service name is “lakelink”, also select the azure subscription and storage account of our target.



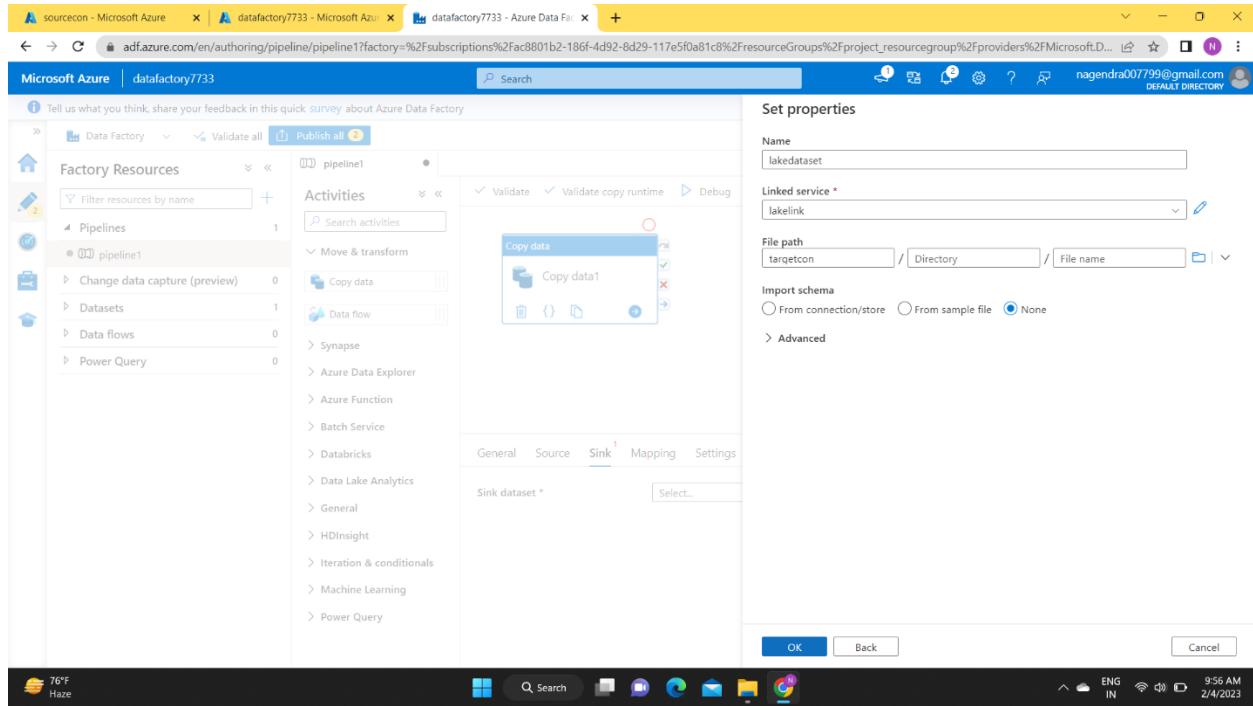
- We have to give target “file path” where we want to store our target file.



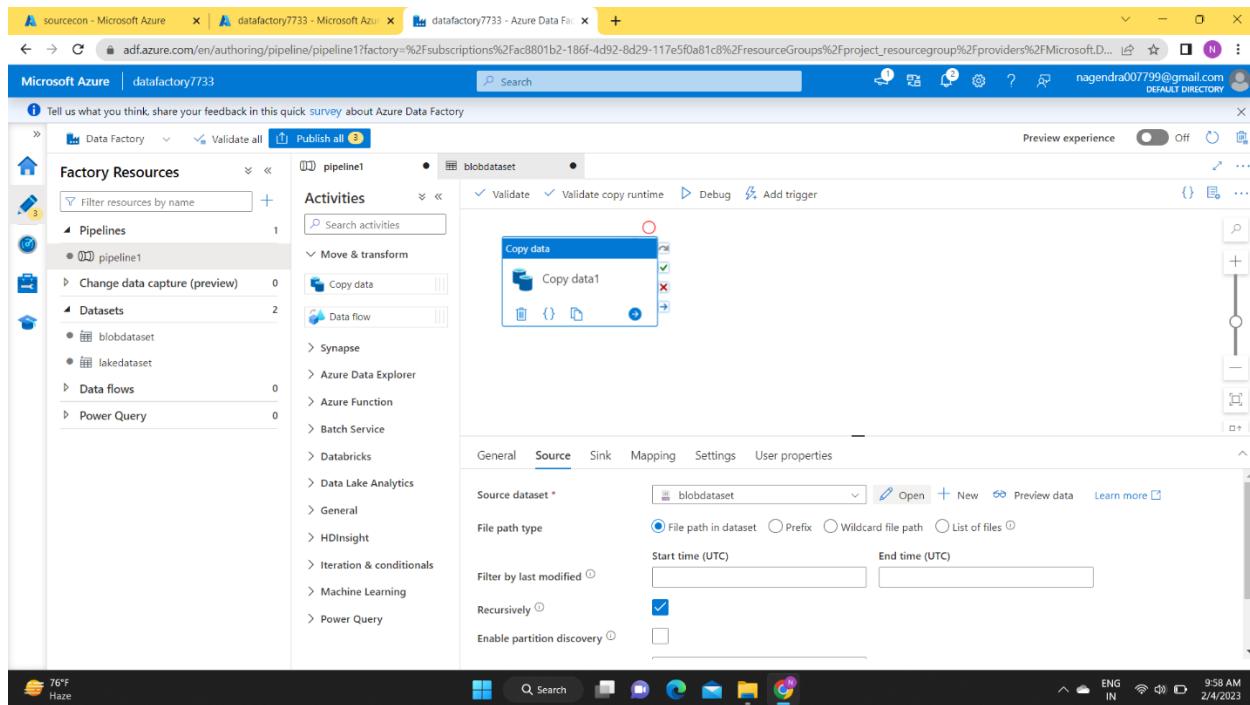
- Here no files in our target it's empty.



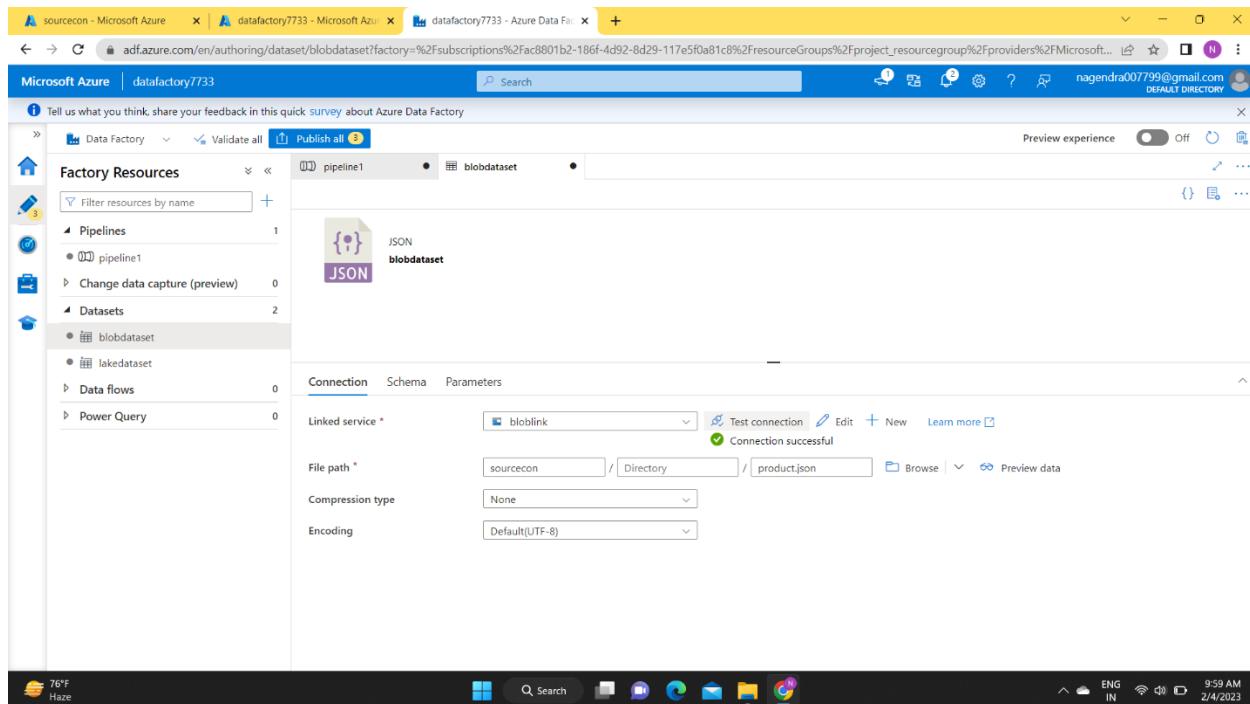
- We are not importing schema so we put none.



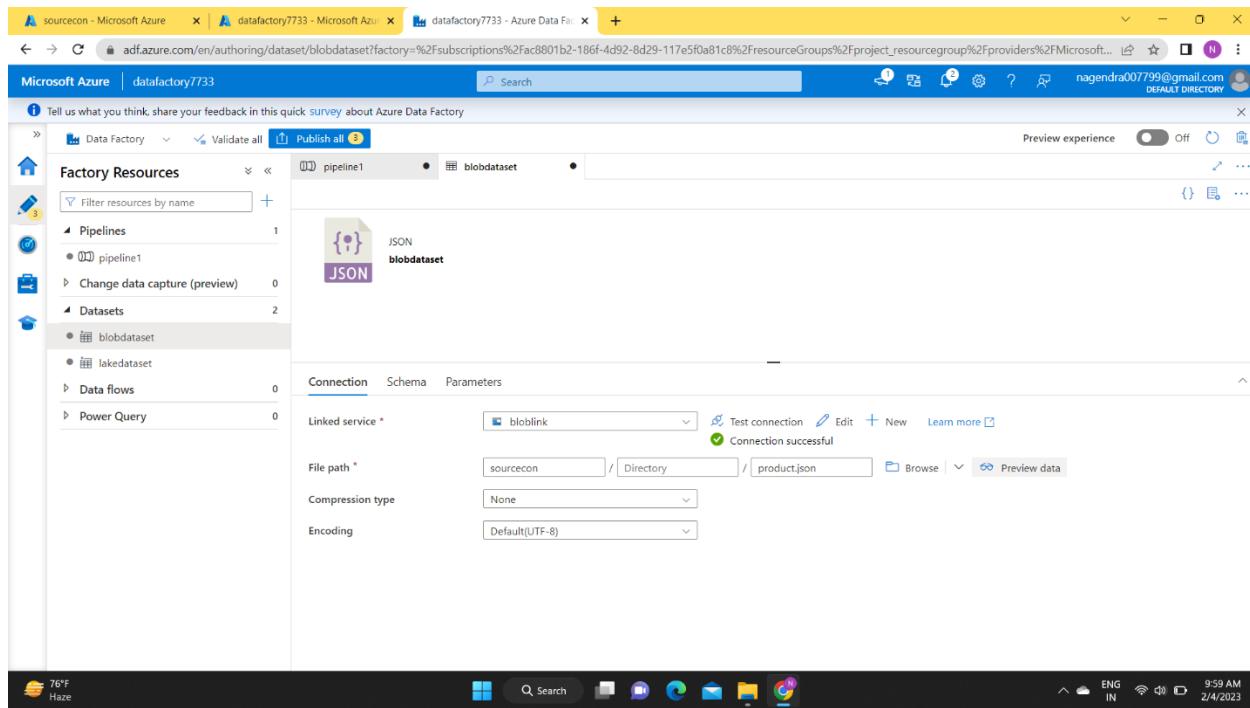
- In source option 2<sup>nd</sup> row “file path type” in that row select “file path in dataset”.
- We are transferring single file so we select “file path in dataset”
- If you want to transfer multiple files select “wild card file path”.
- Go to source there is an “open” option with pencil mark.



- Must do the test connection.



- Then click on preview data.



Microsoft Azure | datafactory7733

Preview experience: Off

Connection Schema Parameters

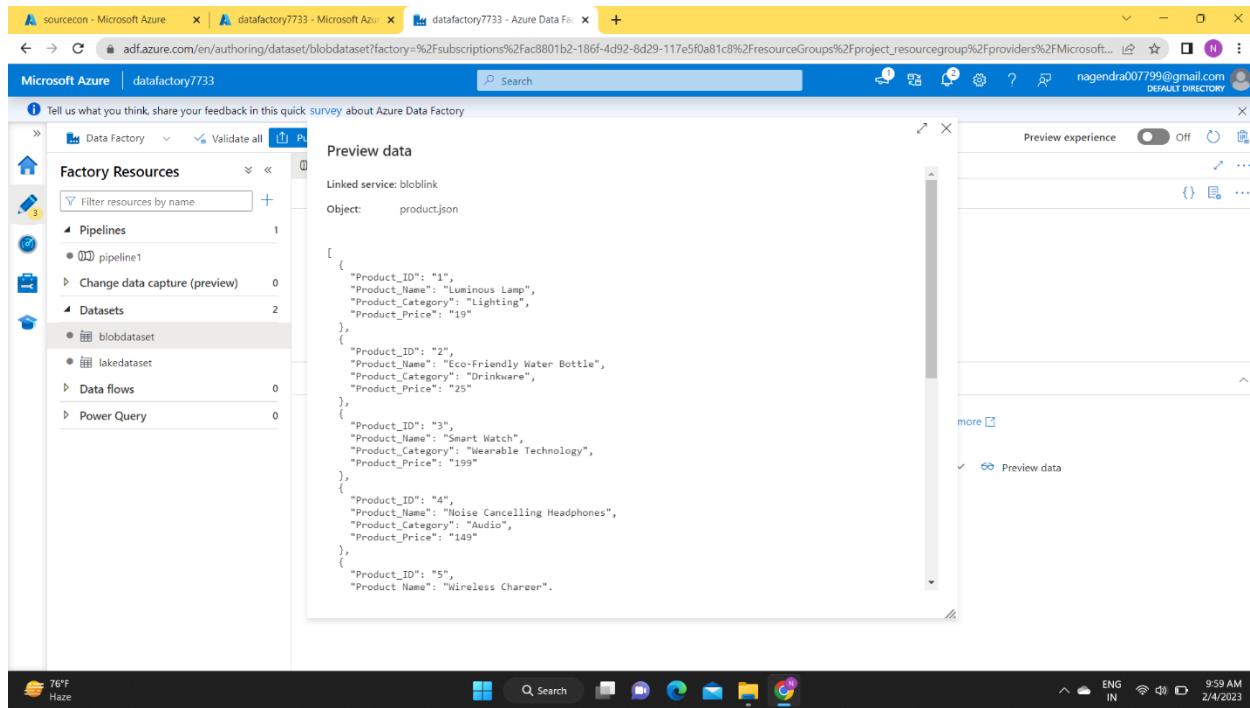
Linked service: bloblink (Connection successful)

File path: sourcecon / Directory / product.json

Compression type: None

Encoding: Default(UTF-8)

- This is our source json file format product data preview.



Preview data

Linked service: bloblink

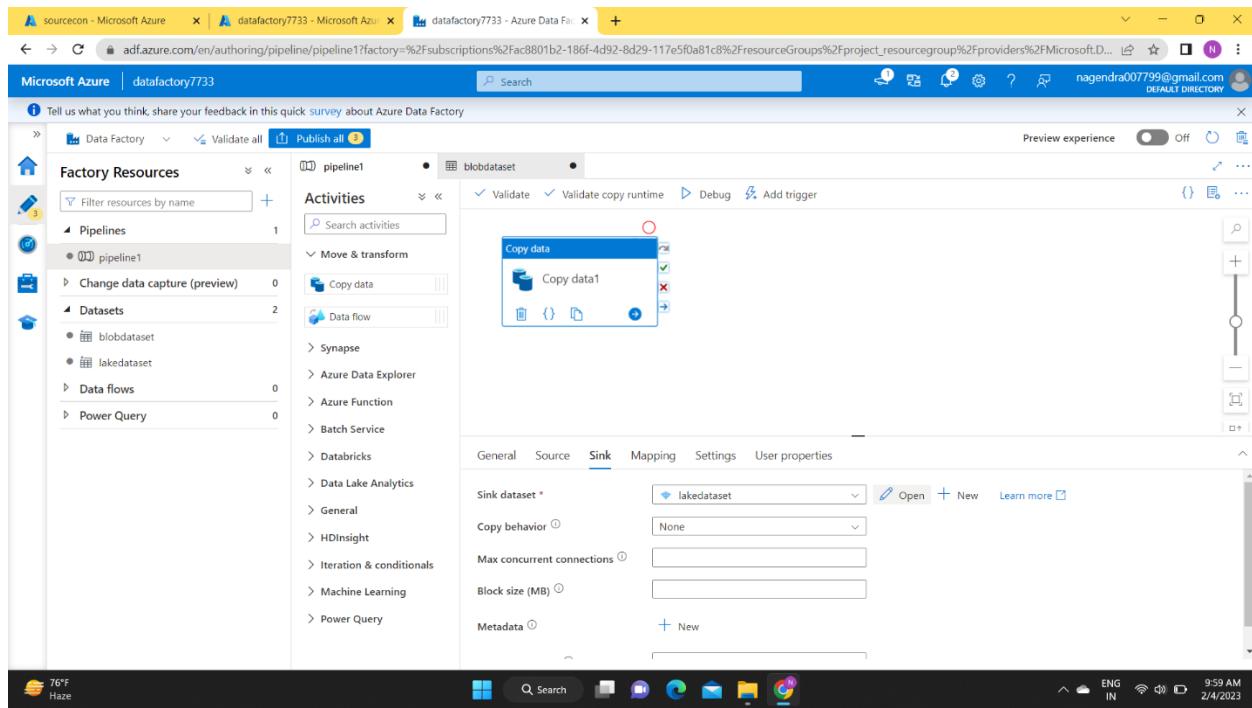
Object: product.json

```
[
  {
    "Product_ID": "1",
    "Product_Name": "Luminous Lamp",
    "Product_Category": "Lighting",
    "Product_Price": "19"
  },
  {
    "Product_ID": "2",
    "Product_Name": "Eco-Friendly Water Bottle",
    "Product_Category": "Drinkware",
    "Product_Price": "25"
  },
  {
    "Product_ID": "3",
    "Product_Name": "Smart Watch",
    "Product_Category": "Wearable Technology",
    "Product_Price": "199"
  },
  {
    "Product_ID": "4",
    "Product_Name": "Noise Cancelling Headphones",
    "Product_Category": "Audio",
    "Product_Price": "149"
  },
  {
    "Product_ID": "5",
    "Product_Name": "Wireless Charger",
    "Product_Category": "Accessories"
  }
]
```

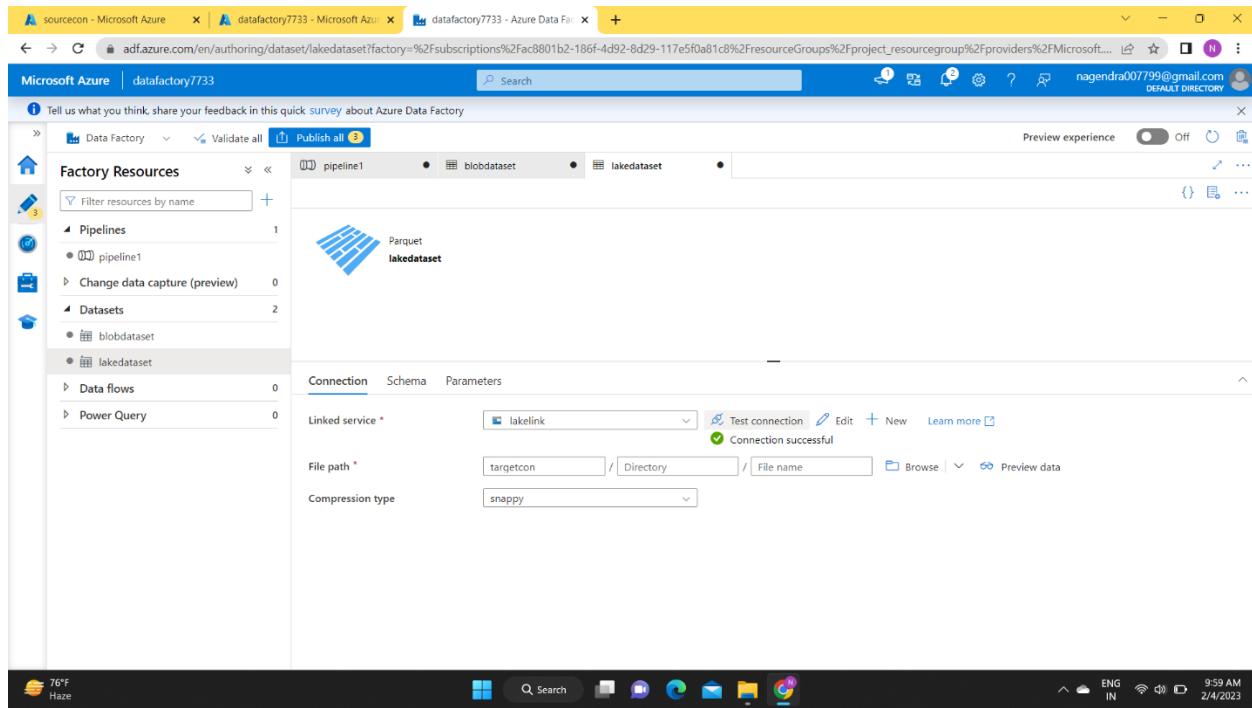
more ▾

✓ Preview data

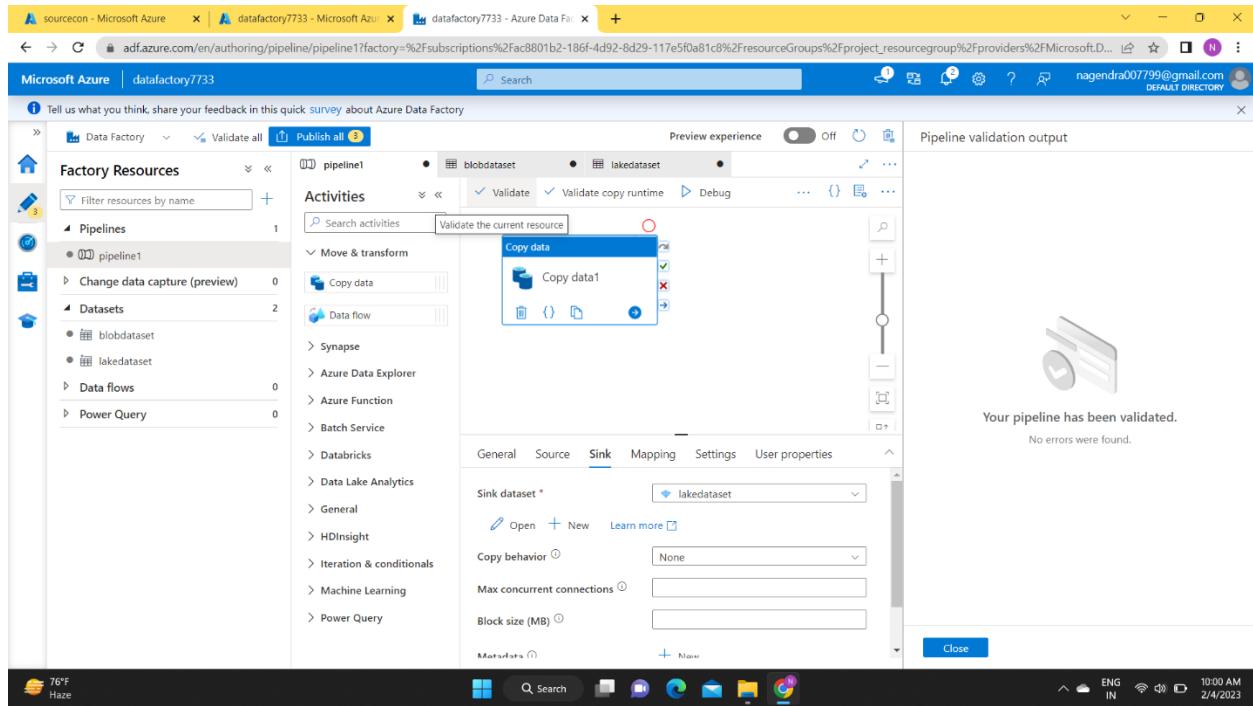
- Now go to sink there is “open” option.



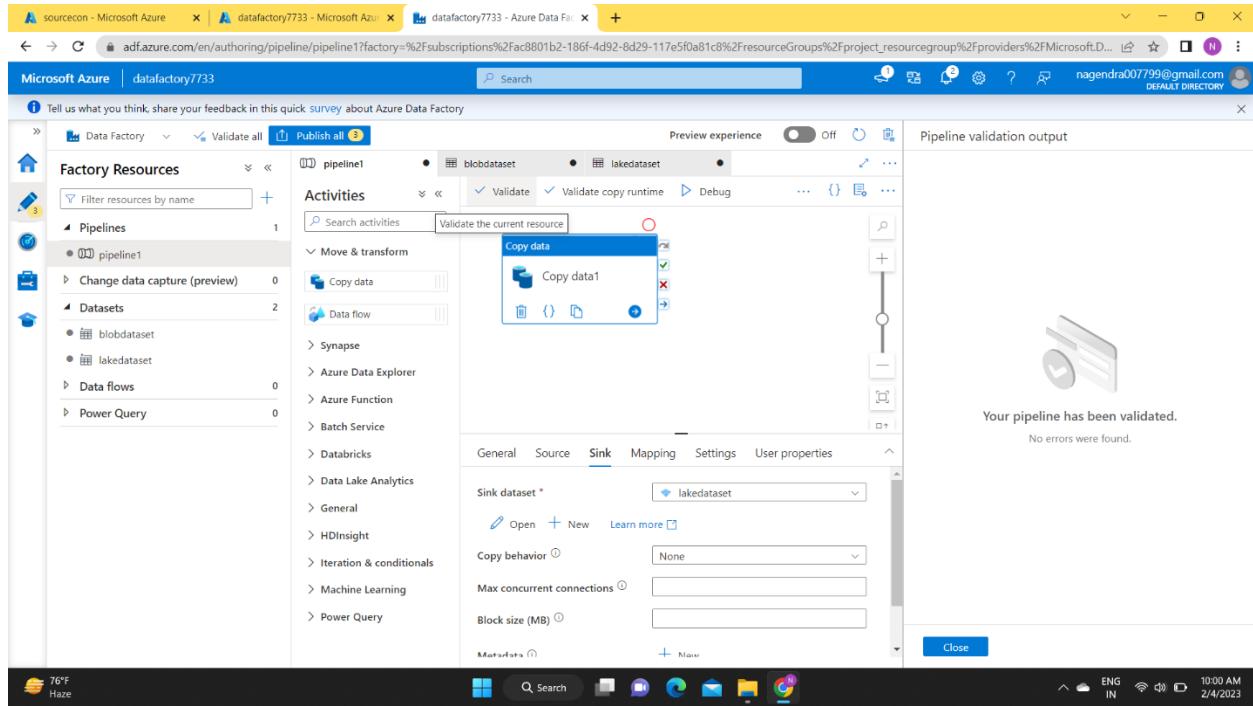
- check the test connection



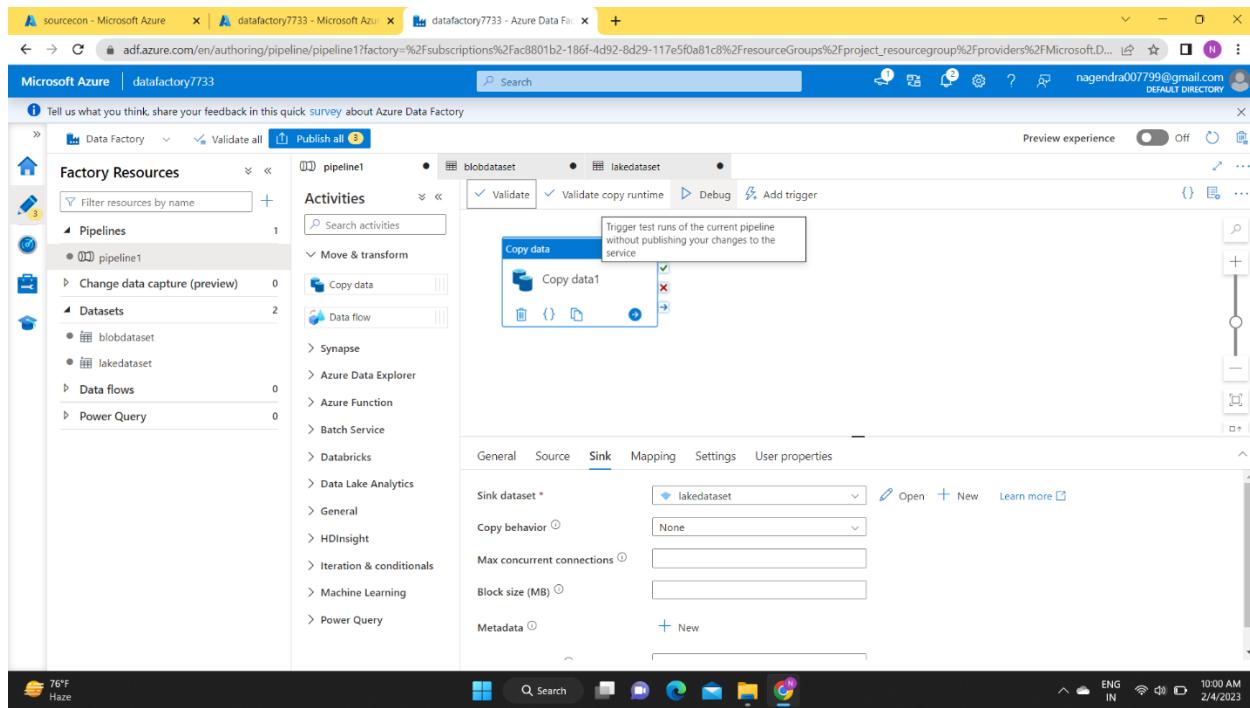
- now validate your pipeline it will show errors on right sidebar



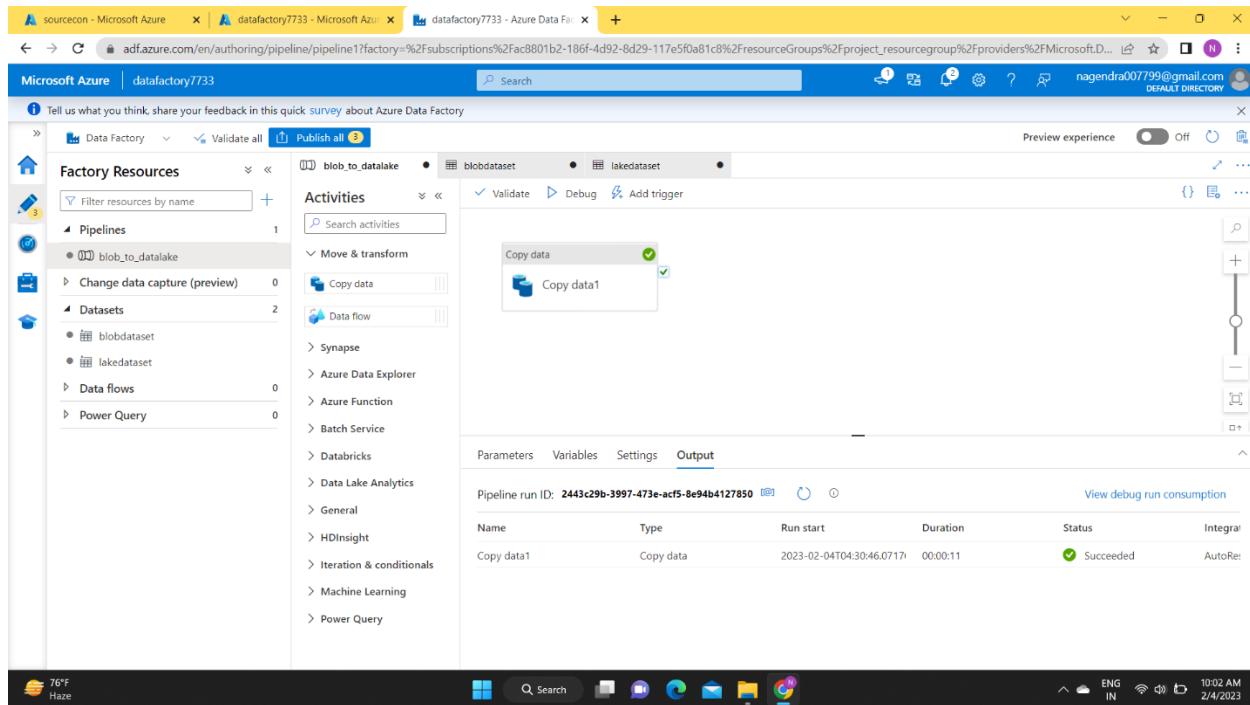
- our validation complete we don't have any errors.



- now do the debug



- our pipeline is succeeding, and you can see that in message box in pipeline.



- to check copy data we are going to data lake.
- In data lake, targetcon our parquet file should be present.

Name	Last modified	Public access level	Lease state
\$logs	2/4/2023, 9:43:01 AM	Private	Available
targetcon	2/4/2023, 9:43:42 AM	Container	Available

- Here is our parquet file in datalake

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
product.parquet	2/4/2023, 10:00:55 AM	Hot (Inferred)	Not yet archived	Block blob	1.46 KB	Available

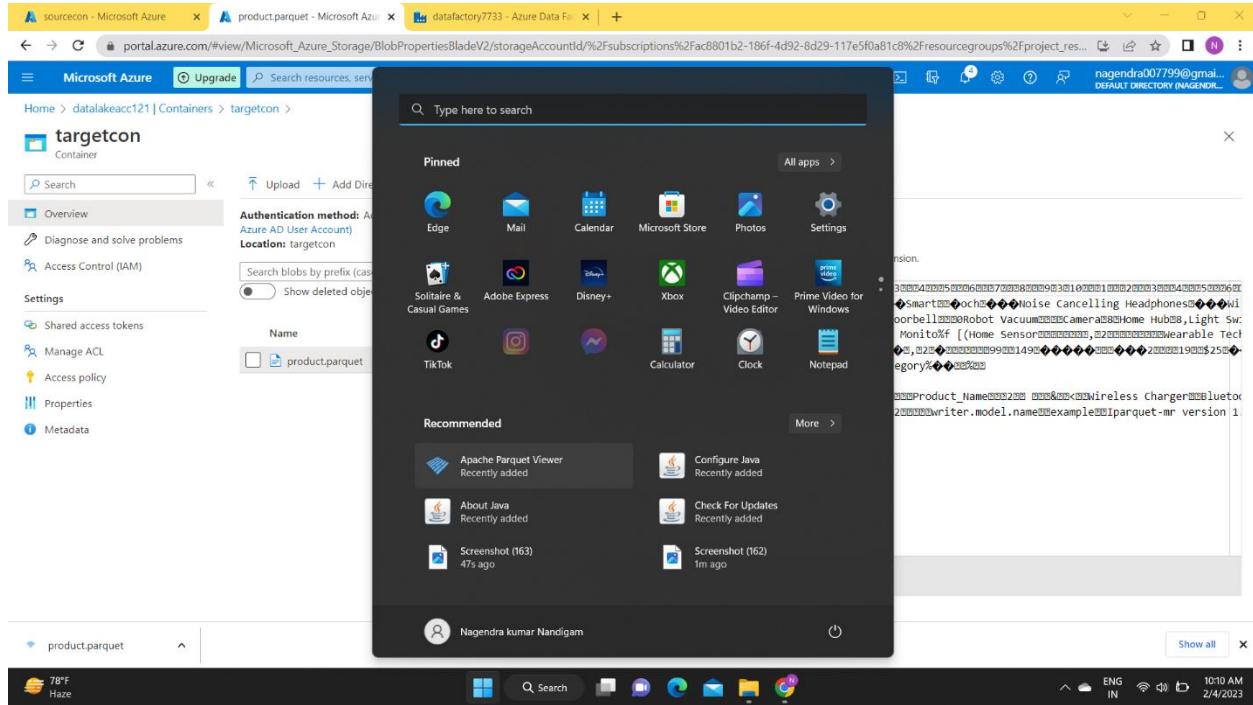
- Click on “edit” to see data present in file

The screenshot shows the Azure Storage Explorer interface. On the left, the 'targetcon' container is selected. In the center, the 'product.parquet' file is selected. The preview pane shows the raw binary data of the Parquet file, which is not rendered correctly due to its unrecognized extension. The status bar at the bottom shows the date and time as 2/4/2023 10:09 AM.

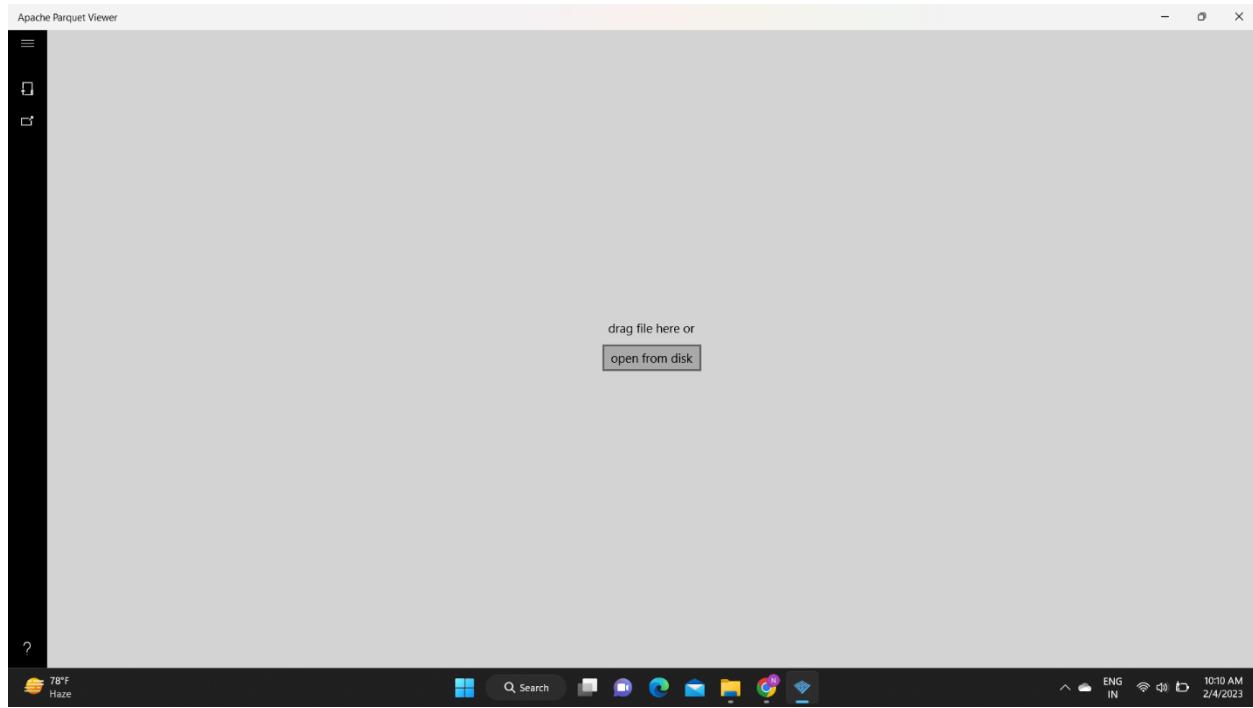
- We cannot understand the file so we have to download file.

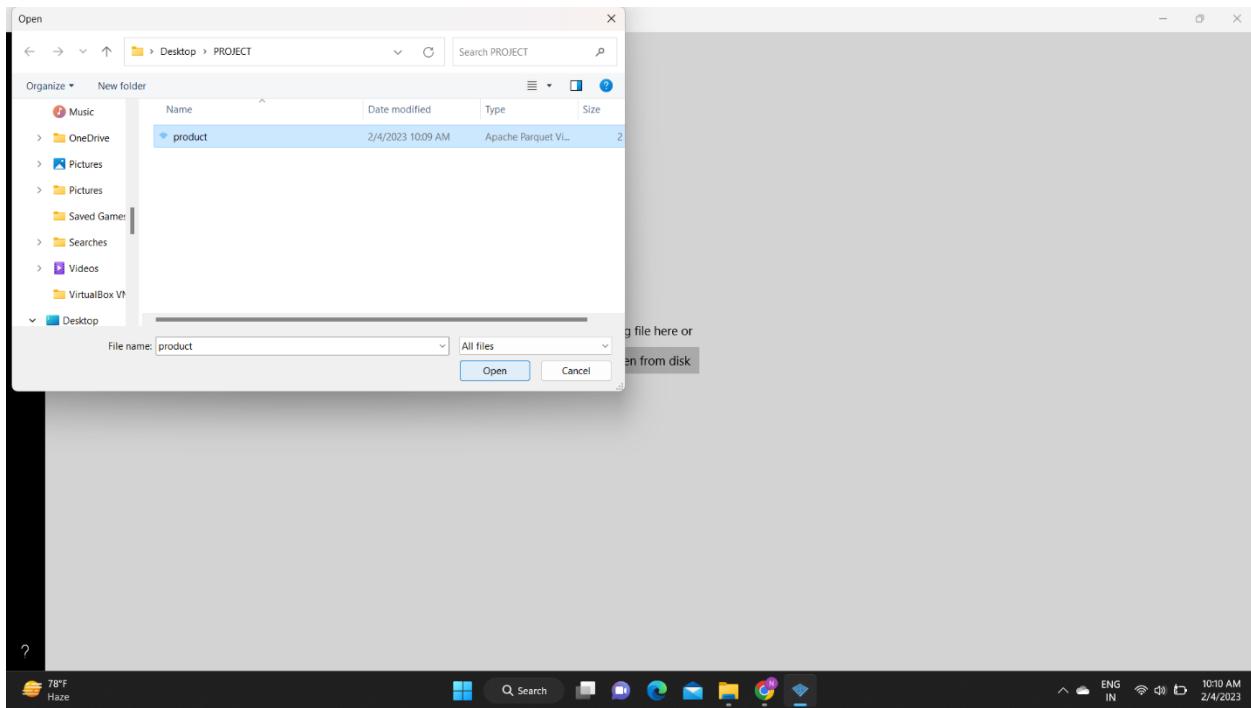
The screenshot shows the Azure Storage Explorer interface. The 'product.parquet' file is selected in the 'targetcon' container. The 'Download' button is highlighted, indicating the file is ready to be downloaded. The status bar at the bottom shows the date and time as 2/4/2023 10:09 AM.

- You have to download the advanced parquet viewer app to see data in parquet.



- Now open the advanced parquet viewer app drag the downloaded file from your pc.





- Now the product data is displayed.

Product_ID	Product_Name	Product_Category	Product_Price
1	Luminous Lamp	Lighting	19
2	Eco-Friendly Water Bottle	Drinkware	25
3	Smart Watch	Wearable Technology	199
4	Noise Cancelling Headphones	Audio	149
5	Wireless Charger	Chargers	29
6	Fitness Tracker	Wearable Technology	99
7	Bluetooth Speaker	Audio	79
8	Smart Thermometer	Health and Wellness	59
9	Ergonomic Mouse	Computer Accessories	39
10	Smart Lock	Smart Home	199
11	Smart Plug	Smart Home	25
12	Smart Bulb	Lighting	19
13	Smart Thermostat	Smart Home	199
14	Smart Smoke Detector	Smart Home	99
15	Smart Doorbell	Smart Home	149
16	Smart Robot Vacuum	Smart Home	249
17	Smart Camera	Smart Home	199
18	Smart Home Hub	Smart Home	99
19	Smart Light Switch	Smart Home	49
20	Smart Outdoor Camera	Smart Home	249
21	Smart Door Lock	Smart Home	199
22	Smart Home Security System	Smart Home	599
23	Smart Home Alarm	Smart Home	299

- Now we are set up the trigger pipeline by clicking “add trigger” and “new/edit”

Microsoft Azure | datafactory7733

Preview experience: Off

Activities: Trigger now, Copy data, New/Edit, Copy data1

Output: Pipeline run ID: 2443c29b-3997-473e-acf5-8e94b4127850

Name	Type	Run start	Duration	Status	Integral
Copy data1	Copy data	2023-02-04T04:30:46.071Z	00:00:11	Succeeded	AutoRe

Microsoft Azure | datafactory7733

Add triggers

Choose trigger...

Search

New

Close

- Here I am using “schedule trigger”

- Now I am attaching trigger successfully.

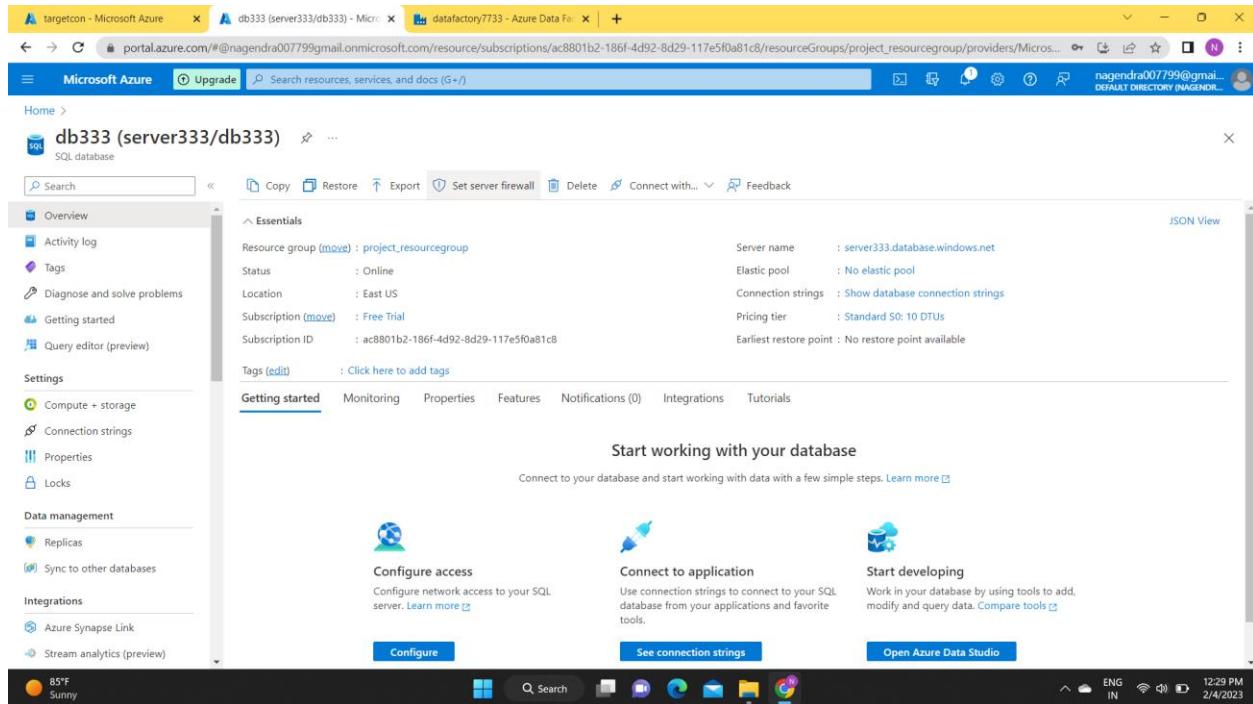
- Now I am publishing the pipeline with trigger.

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. The 'Activities' section is expanded, showing 'Copy data' and 'Copy data1' under 'Move & transform'. The 'Publishing' tab is selected. A 'Publish' dialog is open on the right, titled 'Publish all'. It displays 'Pending changes (4)' for the 'blob\_to\_datalake' pipeline, which includes a new 'blobdataset' dataset and a new 'lakedataset' dataset. It also shows a new 'trigger1' trigger. The 'Output' tab is selected in the dialog. At the bottom, there are 'Publish' and 'Cancel' buttons. The top navigation bar shows the URL 'datafactory7733 - Microsoft Azure' and the title 'datafactory7733 - Azure Data Factory'.

- This is the pipeline we are done.

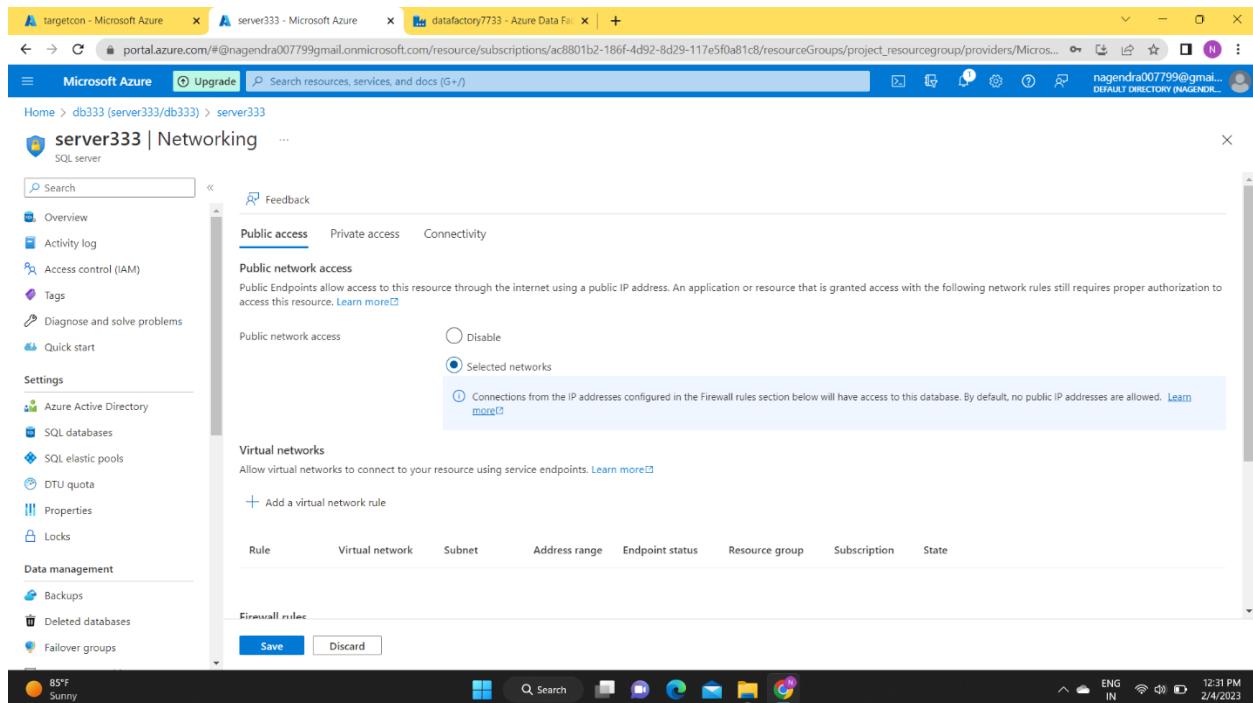
## Copy data from azure sql db to azure data lake

- This is our azure sql database “set server firewall” to give access to our azure resources to this sql db



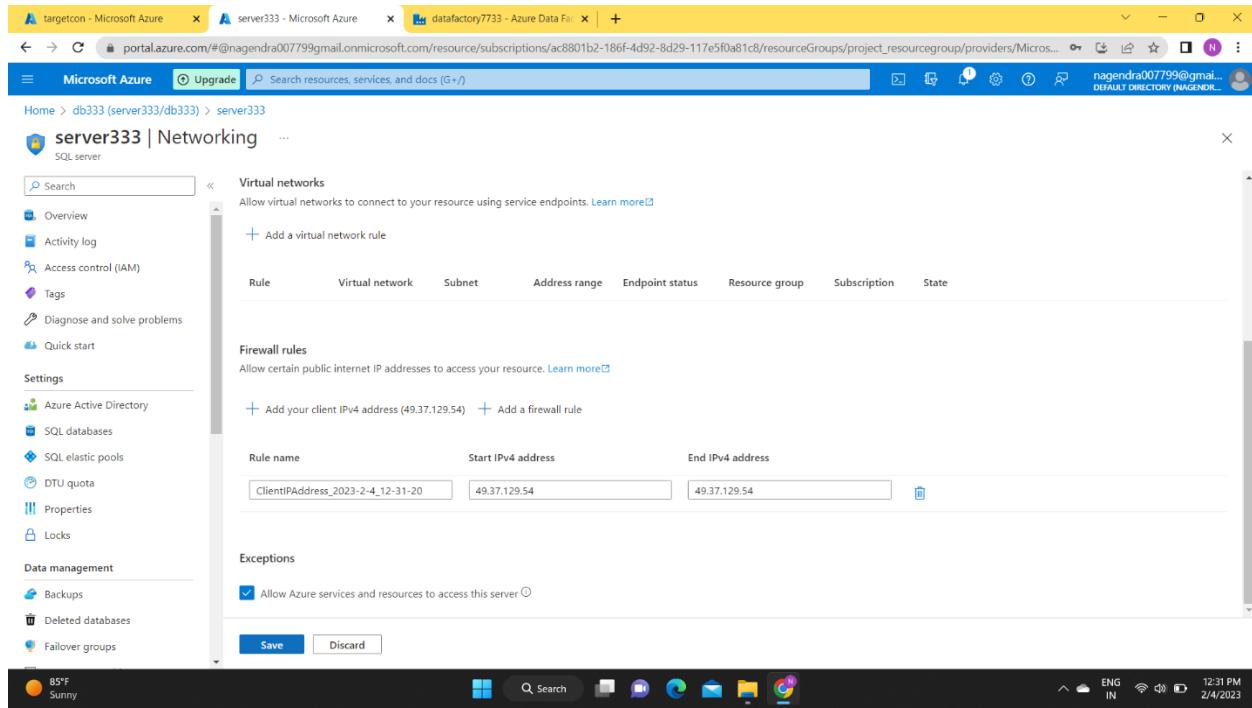
The screenshot shows the Azure portal interface with the following details:

- Top Navigation:** targetcon - Microsoft Azure, db333 (server333/db333) - Microsoft Azure, datafactory7733 - Azure Data Factory, portal.azure.com
- Page Title:** db333 (server333/db333) - Microsoft Azure
- Left Sidebar:** Overview, Activity log, Tags, Diagnose and solve problems, Getting started, Query editor (preview), Compute + storage, Connection strings, Properties, Locks, Data management (Replicas, Sync to other databases), Integrations (Azure Synapse Link, Stream analytics (preview)).
- Essentials Section:** Resource group (move) : project\_resourcegroup, Status : Online, Location : East US, Subscription (move) : Free Trial, Subscription ID : ac8801b2-186f-4d92-8d29-117e5f0a81c8, Tags (edit) : Click here to add tags, Server name : server333.database.windows.net, Elastic pool : No elastic pool, Connection strings : Show database connection strings, Pricing tier : Standard S0: 10 DTUs, Earliest restore point : No restore point available.
- Getting started:** Start working with your database, Configure access, Connect to application, Start developing.
- Bottom:** Weather (85°F, Sunny), Azure Data Studio icon, Save, Discard, and a toolbar with various icons.



The screenshot shows the Azure portal interface with the following details:

- Top Navigation:** targetcon - Microsoft Azure, server333 - Microsoft Azure, datafactory7733 - Azure Data Factory, portal.azure.com
- Page Title:** server333 | Networking - Microsoft Azure
- Left Sidebar:** Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Quick start, Compute + storage, SQL databases, SQL elastic pools, DTU quota, Properties, Locks, Data management (Backups, Deleted databases, Failover groups).
- Public access Section:** Public network access (Selected networks), Virtual networks (Add a virtual network rule), Firewall rules (Save, Discard).
- Bottom:** Weather (85°F, Sunny), Azure Data Studio icon, Save, Discard, and a toolbar with various icons.



server333 | Networking

Virtual networks

Rule	Virtual network	Subnet	Address range	Endpoint status	Resource group	Subscription	State

+ Add a virtual network rule

Firewall rules

Rule name	Start IPv4 address	End IPv4 address
ClientIPAddress_2023-2-4_12-31-20	49.37.129.54	49.37.129.54

+ Add your client IPv4 address (49.37.129.54) + Add a firewall rule

Rule name: ClientIPAddress\_2023-2-4\_12-31-20

Start IPv4 address: 49.37.129.54

End IPv4 address: 49.37.129.54

Exceptions

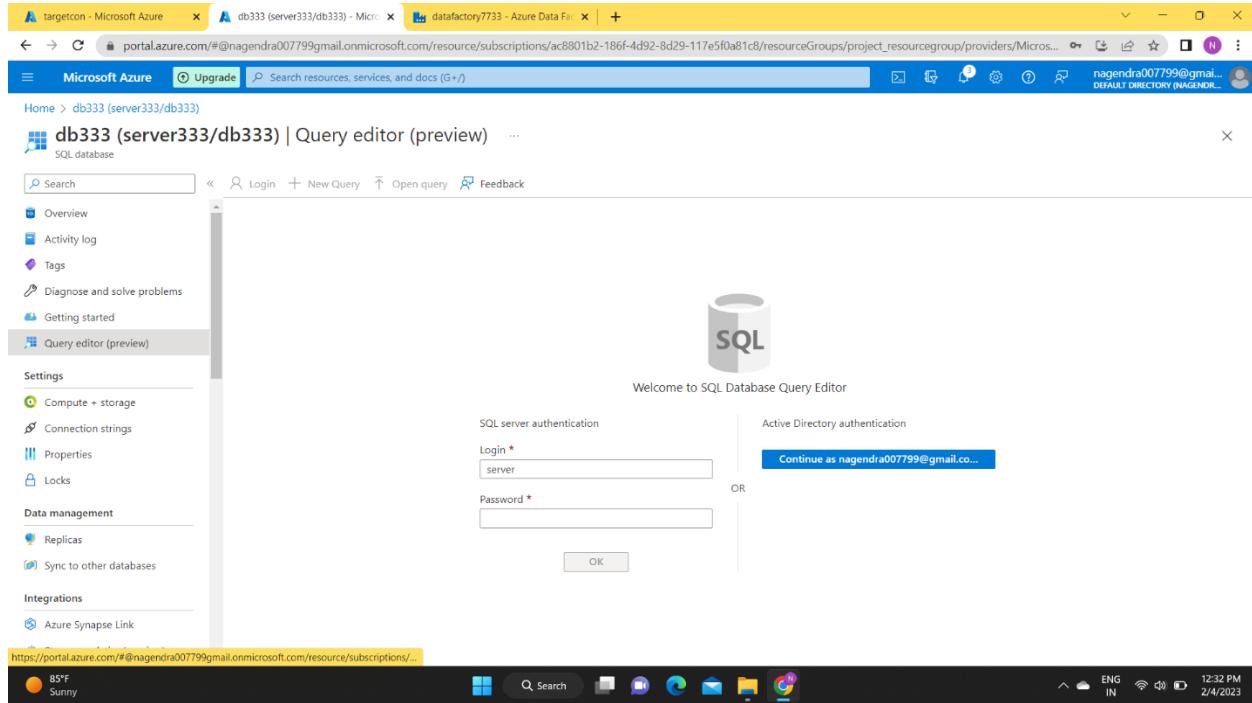
Allow Azure services and resources to access this server

Save Discard

85°F Sunny

ENG IN 12:31 PM 2/4/2023

- You can login azure sql db by using “query editor” option in left side of azure sql portal.



db333 (server333/db333) | Query editor (preview)

Welcome to SQL Database Query Editor

SQL server authentication

Login \* server

OR

Active Directory authentication

Continue as nagendra007799@gmail.com...

OK

Settings

Compute + storage

Connection strings

Properties

Locks

Data management

Replicas

Sync to other databases

Integrations

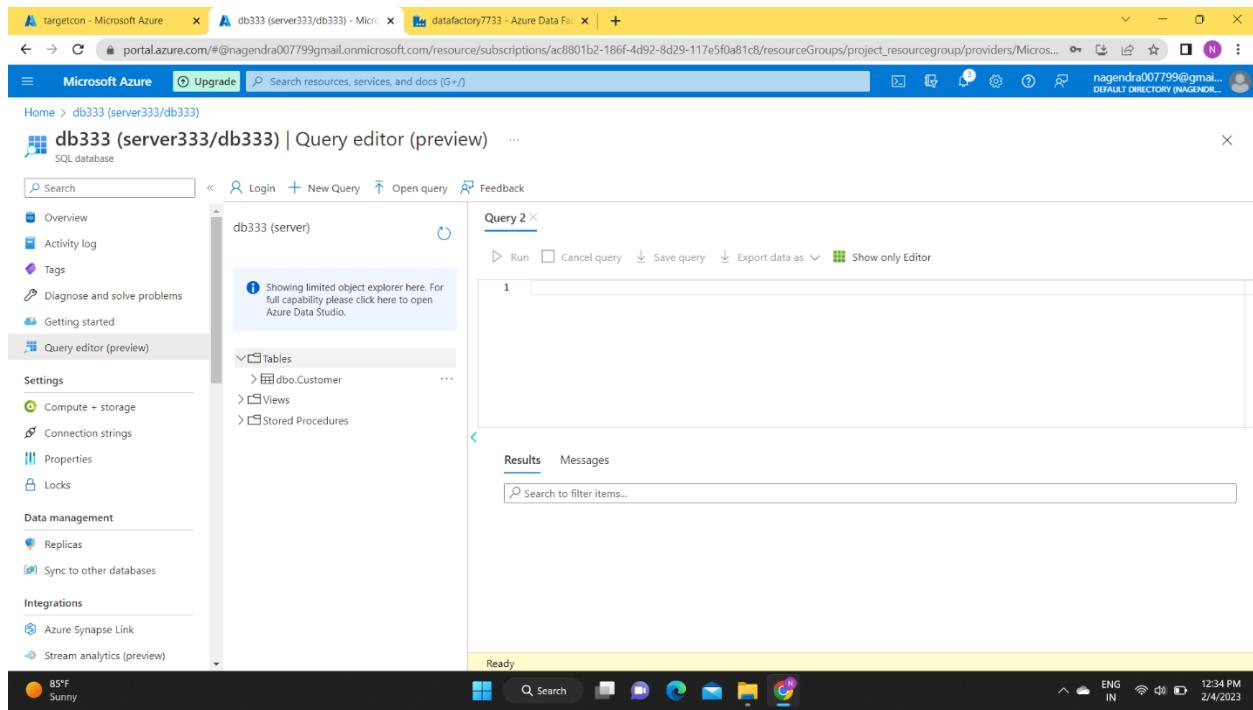
Azure Synapse Link

https://portal.azure.com/#@nagendra007799@gmail.com/resource/subscriptions/...

85°F Sunny

ENG IN 12:32 PM 2/4/2023

- Here you can see our customer table “dbo.customer”.



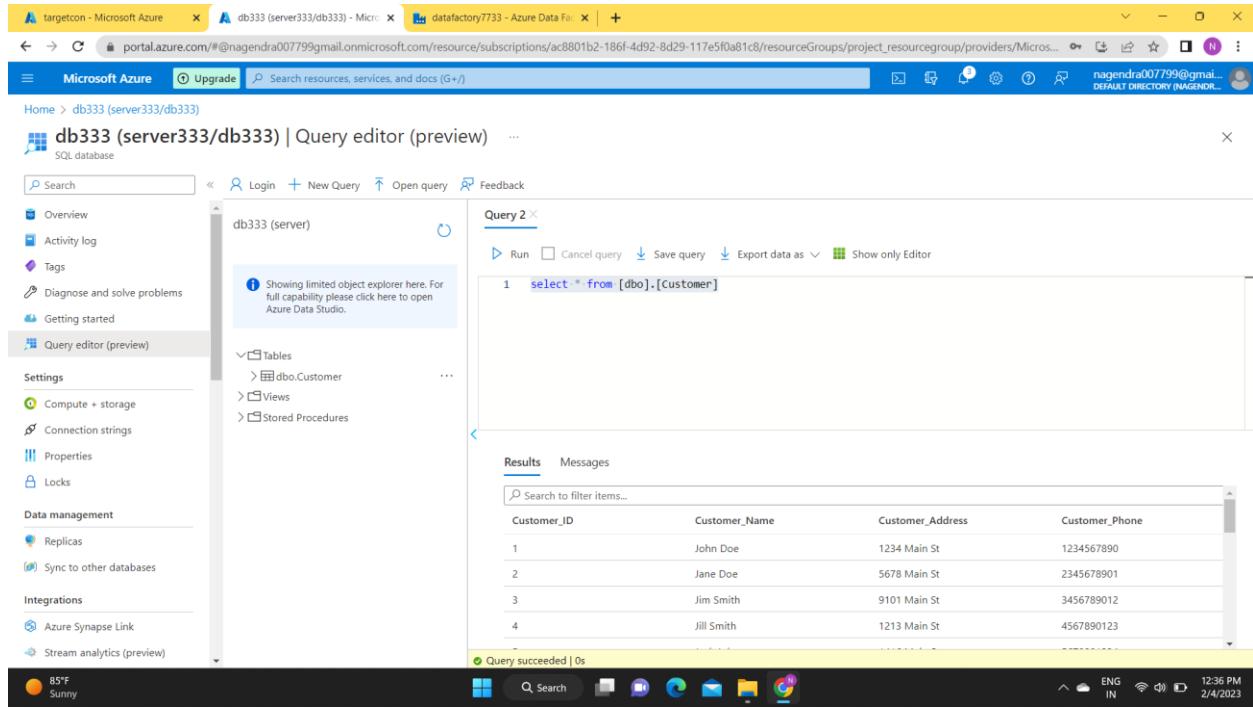
The screenshot shows the Azure Data Studio interface. The left sidebar is the 'Query editor (preview)' settings menu. The main area shows 'Query 2' with the following code:

```
1
```

The 'Tables' section in the sidebar shows:

- Tables: > dbo.Customer
- Views
- Stored Procedures

- You can see the data in customer table.



The screenshot shows the Azure Data Studio interface. The left sidebar is the 'Query editor (preview)' settings menu. The main area shows 'Query 2' with the following code:

```
1 select * from [dbo].[Customer]
```

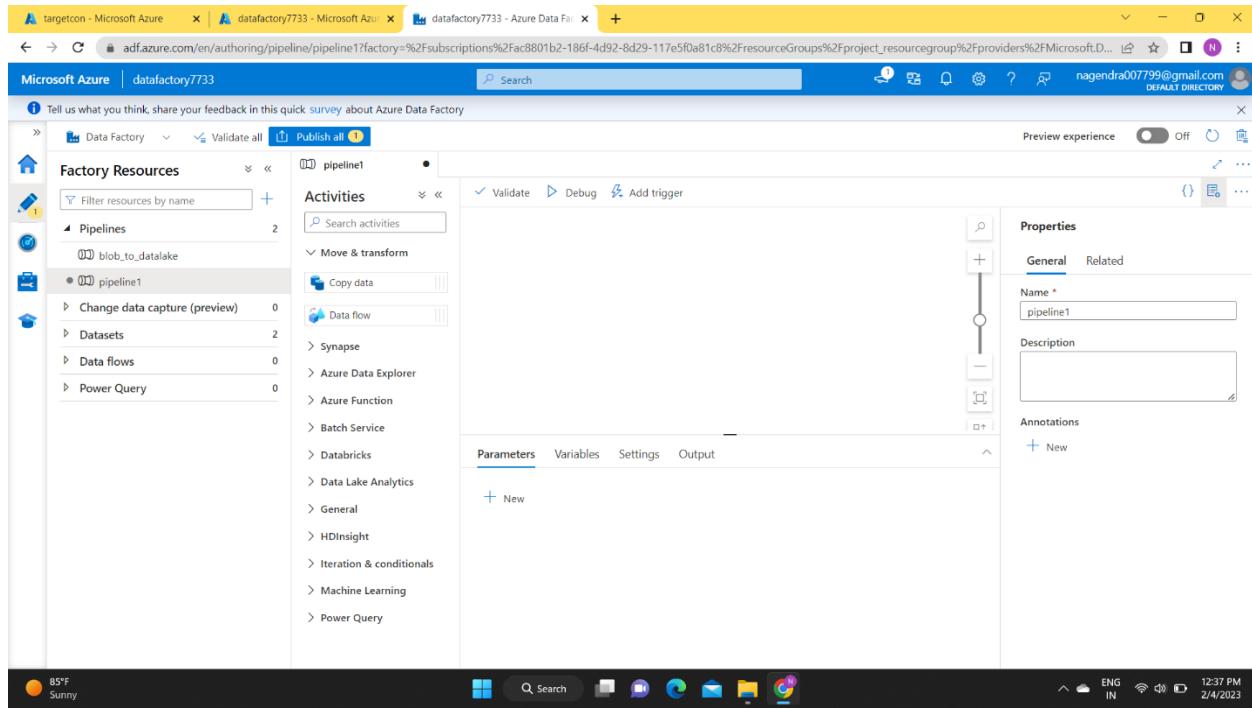
The 'Results' tab displays the data from the 'Customer' table:

Customer_ID	Customer_Name	Customer_Address	Customer_Phone
1	John Doe	1234 Main St	1234567890
2	Jane Doe	5678 Main St	2345678901
3	Jim Smith	9101 Main St	3456789012
4	Jill Smith	1213 Main St	4567890123

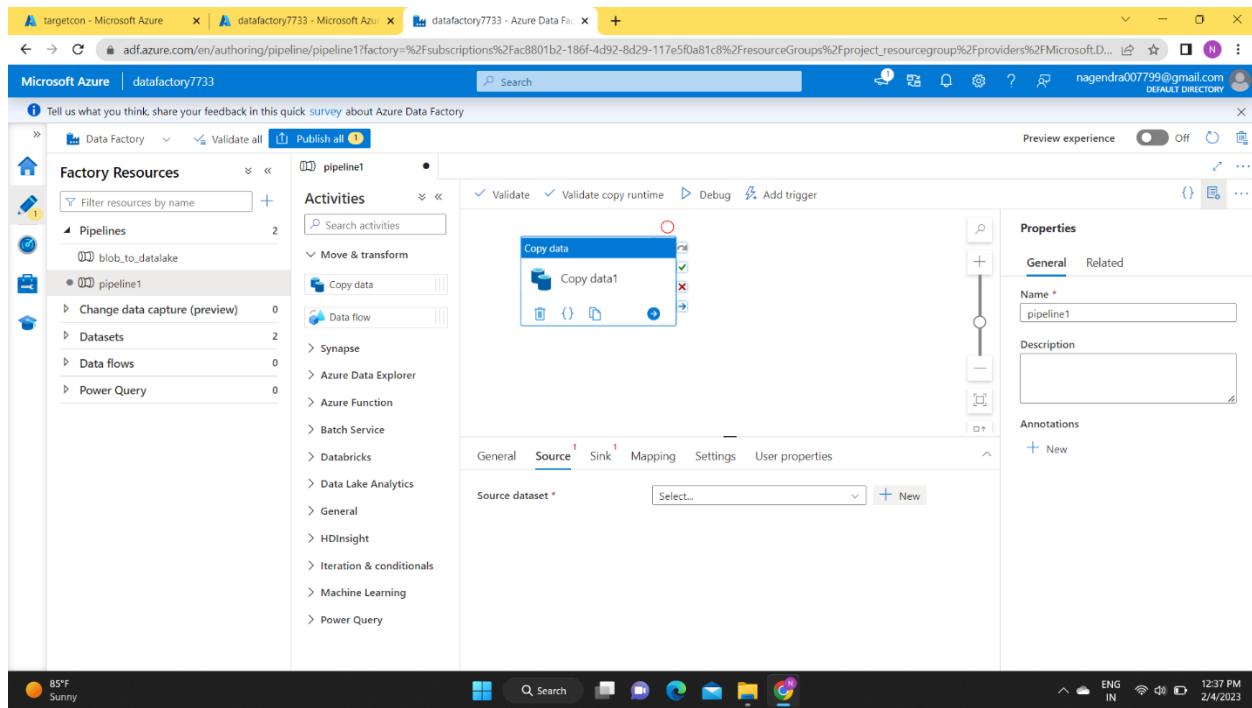
At the bottom, a message says 'Query succeeded | 0s'.

- Now we can launch azure data factory.

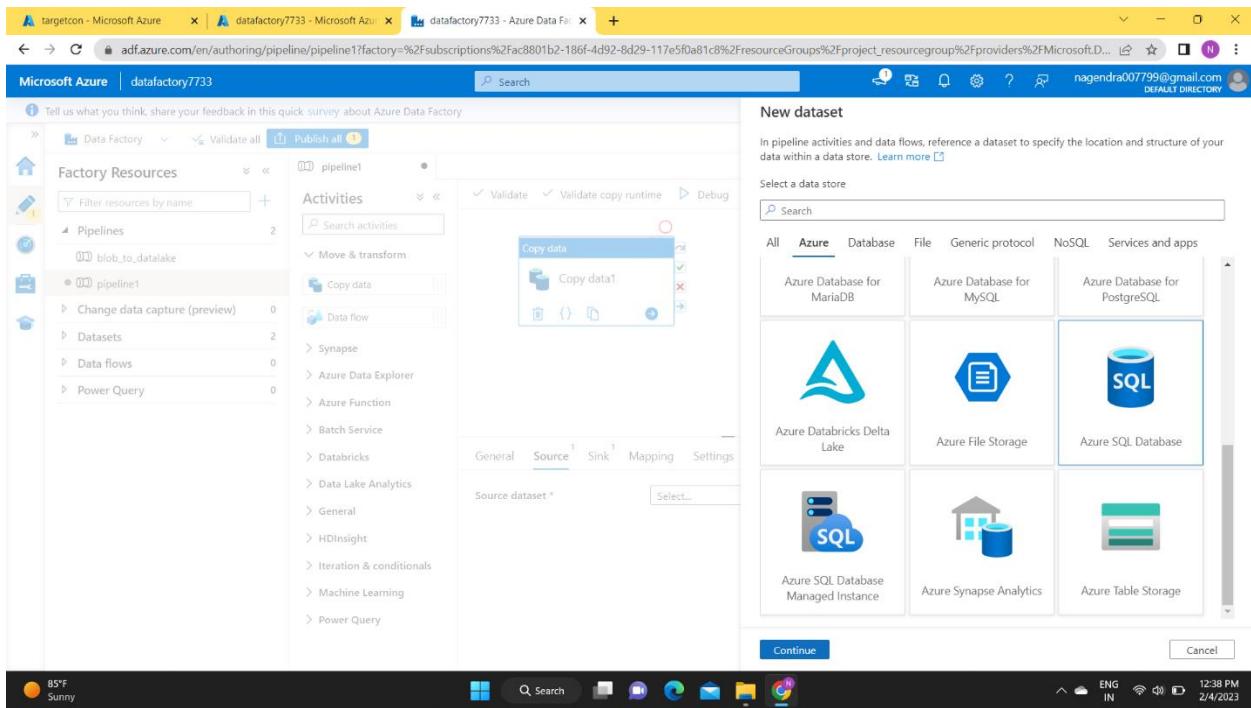
- Create new pipeline



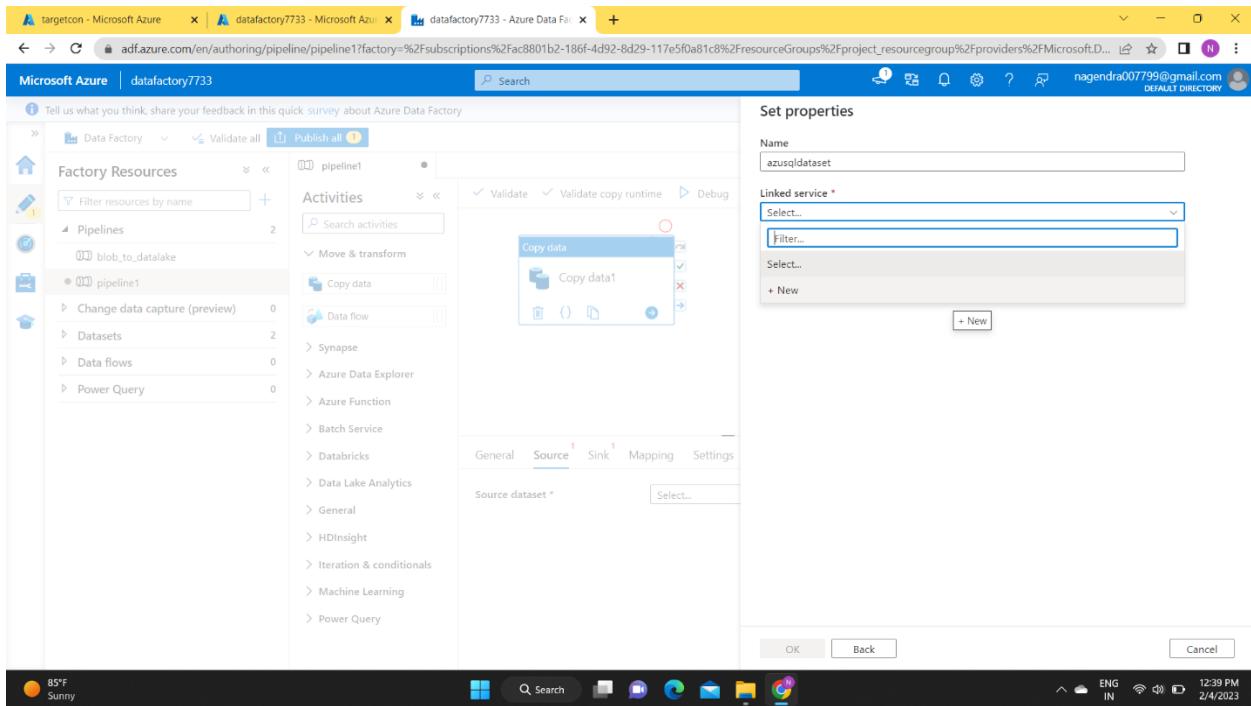
- Select copy data activity in adf.
- In source create new dataset for source azure sql db.



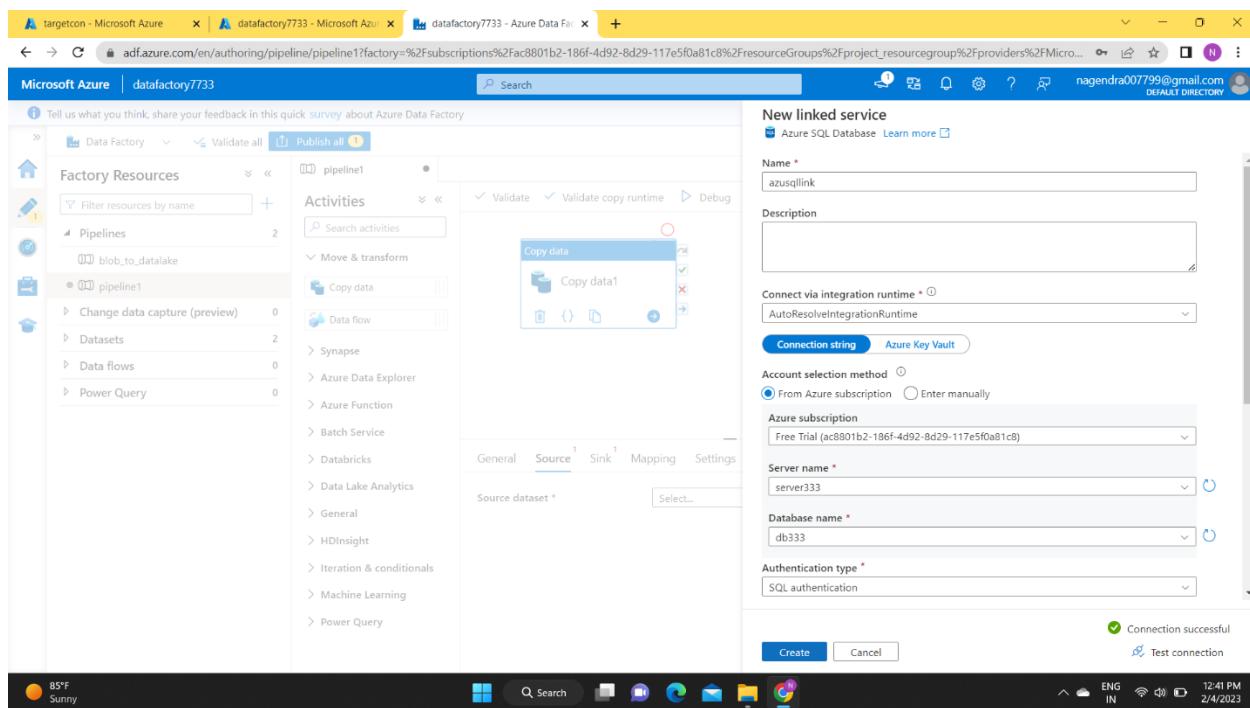
- In source you can choose azure sql db.



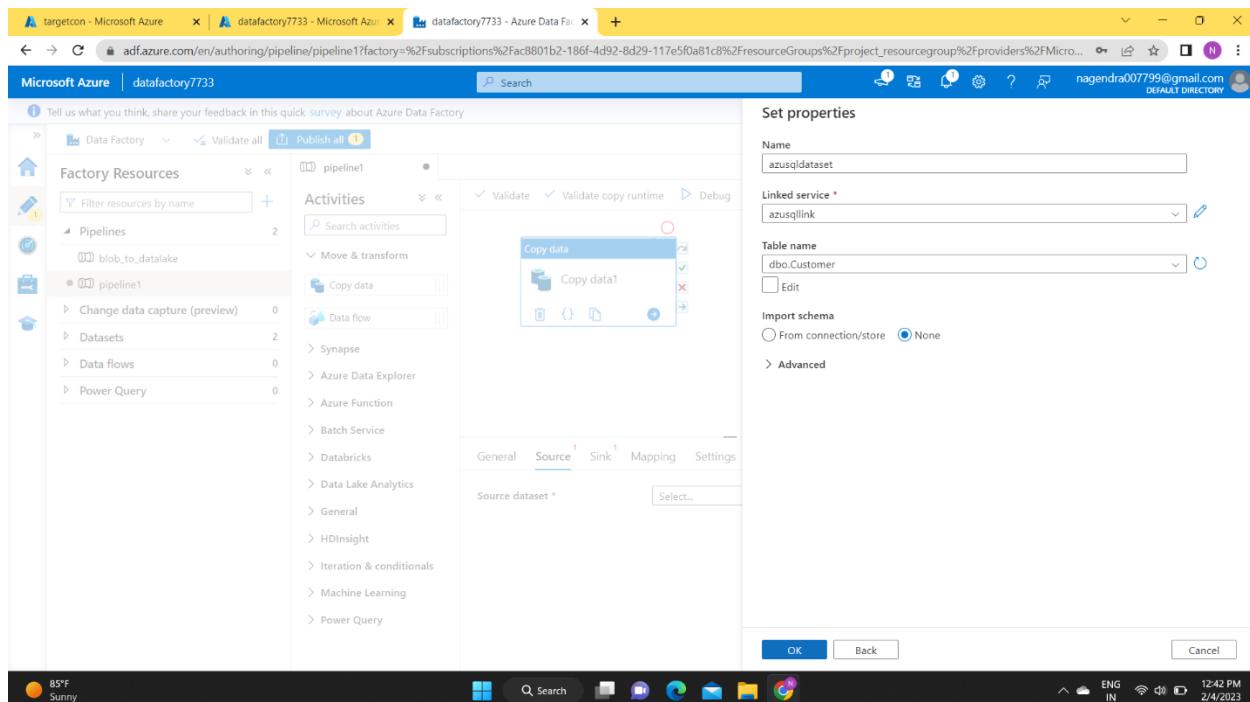
- Our dataset name is “azusqldataset”.



- Our azure sql dataset link service name is “azusqllink”
- Here we have to create sql server.
- The name of sql server is “server333”
- Test the connection.



- You select your table and put import schema should be none because we are importing data from .sql to parquet file format.



- Insource we are selecting “open” to edit our table name.

The screenshot shows the Azure Data Factory pipeline editor. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (blob\_to\_datalake, pipeline1), 'Datasets' (azusqldataset, blobdataset, lakedataset), and 'Data flows'. The main workspace shows a 'Copy data' activity within a 'pipeline1' pipeline. The 'Source dataset' is set to 'azusqldataset'. The 'Use query' section has 'Table' selected. The pipeline is named 'pipeline1'.

- Here I am changing file name dbo.customer to customer, and I am done the test connection.

The screenshot shows the Azure Data Factory dataset editor for 'azusqldataset'. The 'Connection' tab is selected, showing a successful test connection to 'azusqllink' with the table 'Customer' selected. The dataset is named 'azusqldataset'.

- Now I am going to see preview data

Microsoft Azure | datafactory7733

Preview experience: Off

Properties: General, Related

Name: azuresql\_to\_datalake

Description:

Annotations: + New

Source dataset: azusqldataset

Use query: Table

Query timeout (minutes): 120

Isolation level: None

- This is the preview of our customer table data

Microsoft Azure | datafactory7733

Preview data

Linked service: azusqllink

Object: Customer

	Customer_ID	Customer_Name	Customer_Address	Customer_Phone
1	1	John Doe	1234 Main St	1234567890
2	2	Jane Doe	5678 Main St	2345678901
3	3	Jim Smith	9101 Main St	3456789012
4	4	Jill Smith	1213 Main St	4567890123
5	5	Jack Johnson	1416 Main St	5678901234
6	6	Jenny Johnson	1719 Main St	6789012345
7	7	Jake Williams	2022 Main St	7890123456
8	8	Joan Williams	2325 Main St	8901234567
9	9	Joe Brown	2628 Main St	9012345678
10	10	Jane Brown	2931 Main St	0123456789

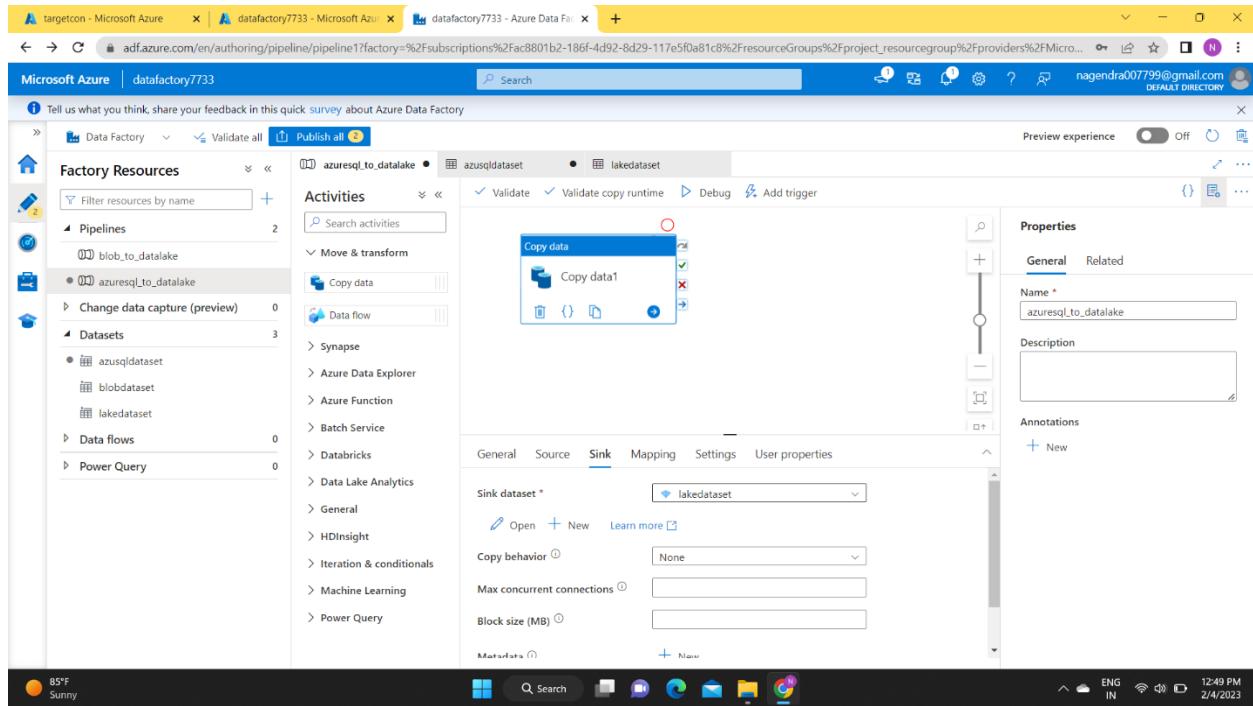
Properties: General, Related

Name: azuresql\_to\_datalake

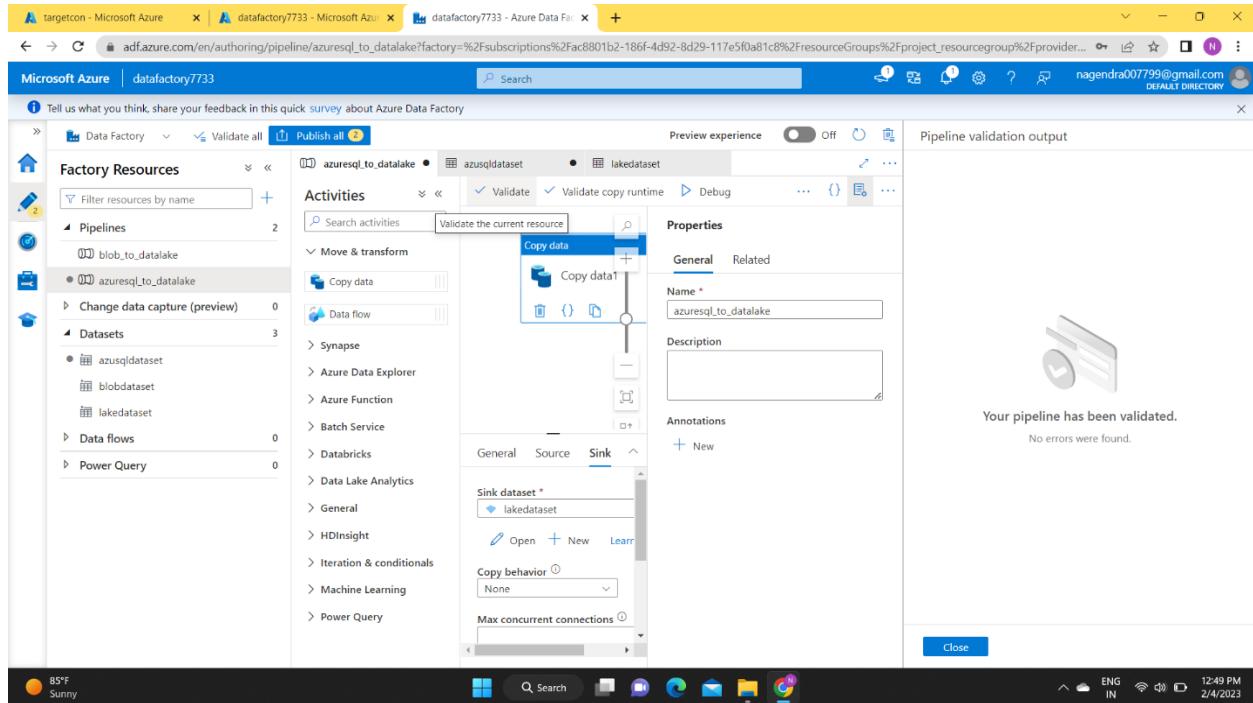
Description:

Annotations: + New

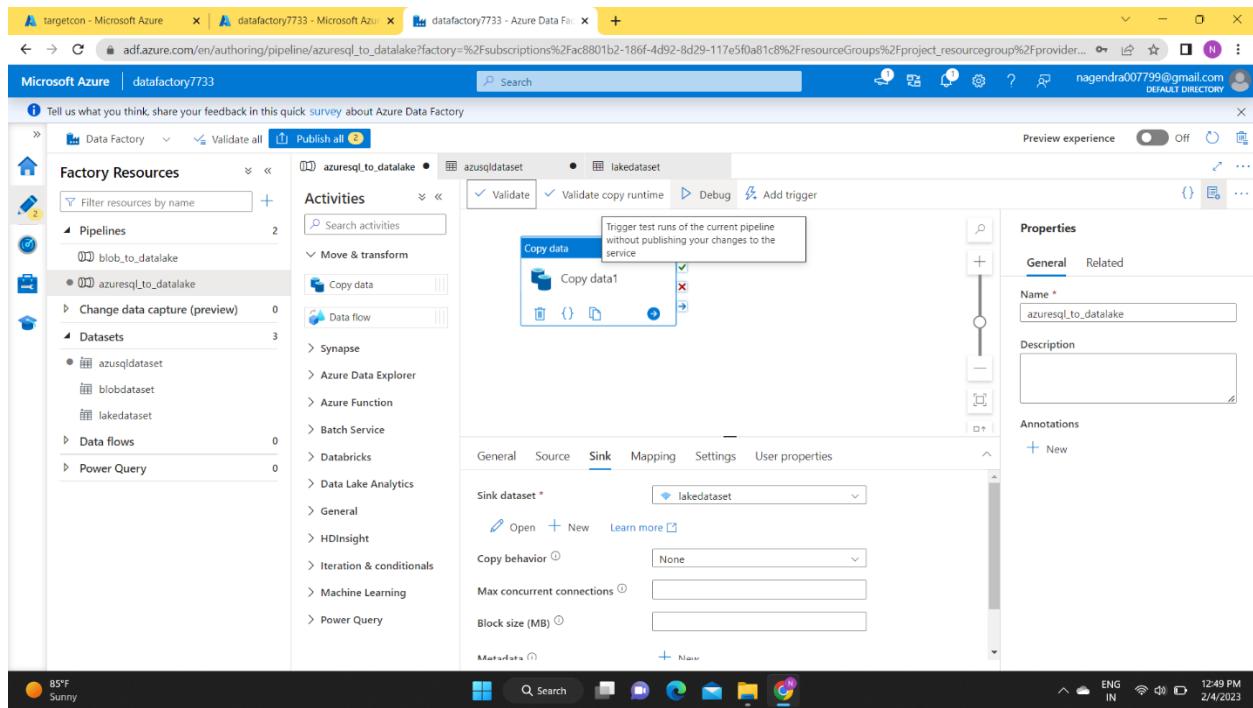
- we have data lake dataset and link service, so I am giving them.



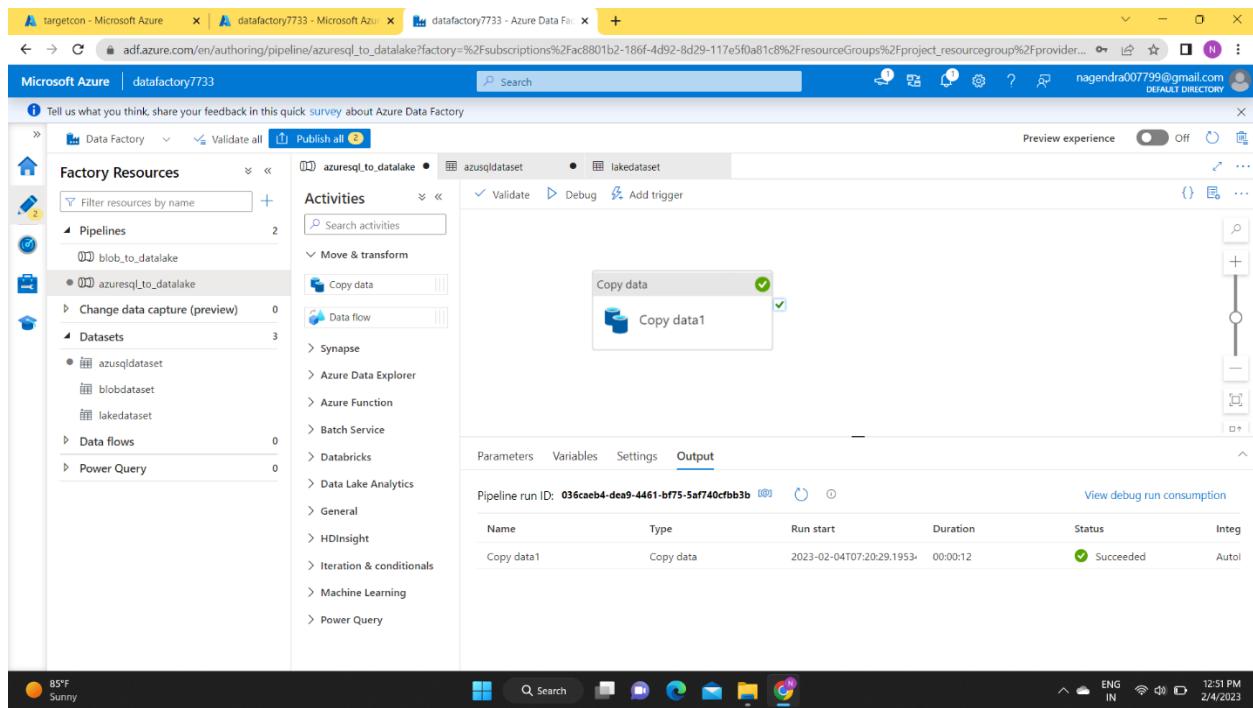
- now I am going to validate our pipeline



- now I am doing debug



- pipeline is successful.



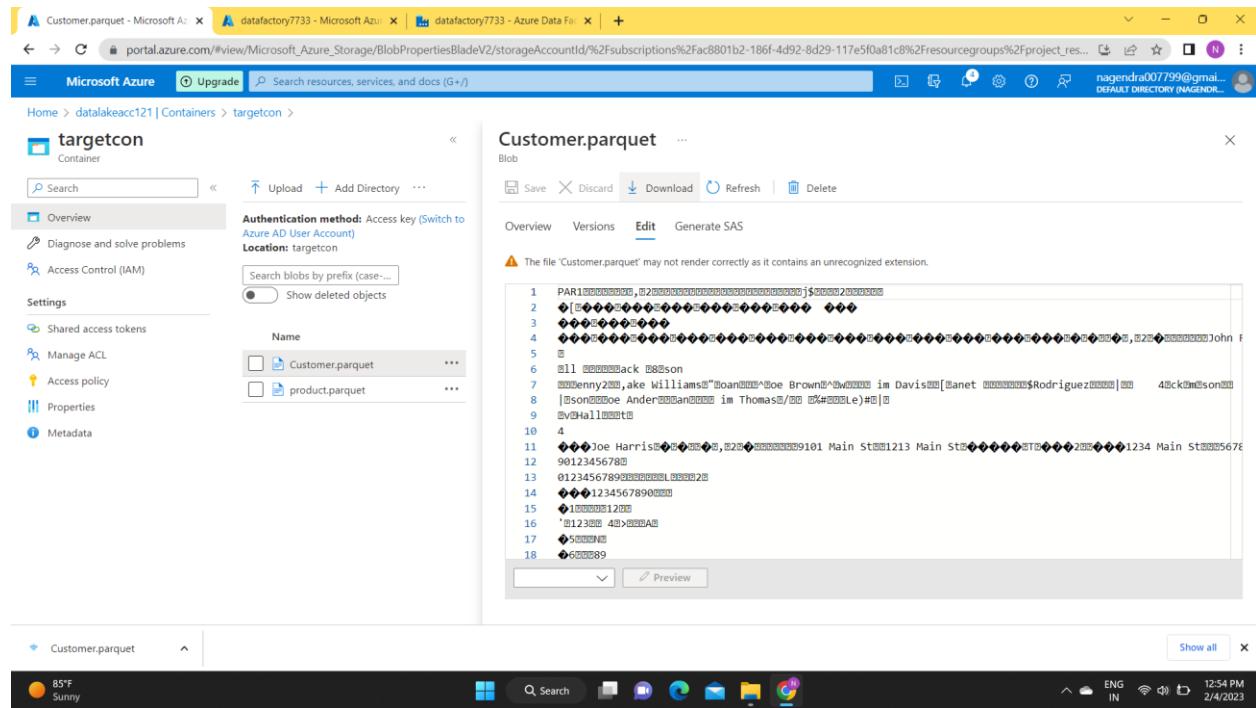
- Now I am checking customer data in target datakake

Name	Last modified	Public access level	Lease state
logs	2/4/2023, 9:43:01 AM	Private	Available
targetcon	2/4/2023, 9:43:42 AM	Container	Available

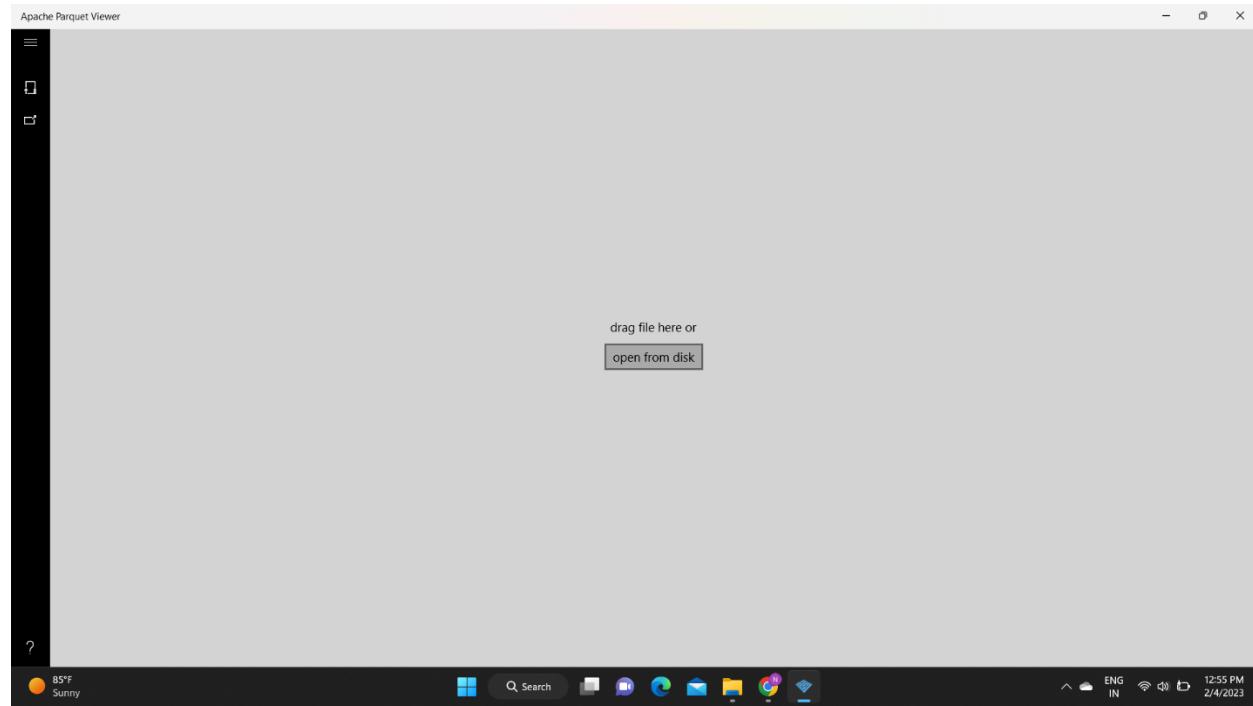
- Here is our customer file in parquet file format.

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
Customer.parquet	2/4/2023, 12:50:40 PM	Hot (Inferred)		Block blob	1.37 KIB	Available
product.parquet	2/4/2023, 12:36:13 PM	Hot (Inferred)		Block blob	1.46 KIB	Available

- We cannot understand data so download file.



- I am open the app “advanced parquet viewer” and upload your parquet file here.



- This is how our customer data looks like.

Apache Parquet Viewer

	Customer_ID	Customer_Name	Customer_Address	Customer_Phone
1	John Doe	1234 Main St	1234567890	
2	Jane Doe	5678 Main St	2345678901	
3	Jim Smith	9101 Main St	3456789012	
4	Jill Smith	1213 Main St	4567890123	
5	Jack Johnson	1416 Main St	5678901234	
6	Jenny Johnson	1719 Main St	6789012345	
7	Jake Williams	2022 Main St	7890123456	
8	Joan Williams	2325 Main St	8901234567	
9	Joe Brown	2628 Main St	9012345678	
10	Jane Brown	2931 Main St	0123456789	
11	Jim Davis	3234 Main St	1234567890	
12	Janet Davis	3537 Main St	2345678901	
13	John Rodriguez	3840 Main St	3456789012	
14	Jenny Rodriguez	4143 Main St	4567890123	
15	Jack Wilson	4446 Main St	5678901234	
16	Joan Wilson	4749 Main St	6789012345	
17	Joe Anderson	5052 Main St	7890123456	
18	Jane Anderson	5355 Main St	8901234567	
19	Jim Thomas	5658 Main St	9012345678	
20	Janet Thomas	5961 Main St	0123456789	
21	John Lee	6264 Main St	1234567890	
22	Jenny Lee	6567 Main St	2345678901	
23	Jack Hall	6870 Main St	3456789012	

showing first 25 records.

85°F Sunny

ENG IN 12:56 PM 2/4/2023

targetcon - Microsoft Azure | datafactory7733 - Microsoft Azure | datafactory7733 - Azure Data Factory +

adf.azure.com/en/authoring/pipeline/azuresql\_to\_datalake?factory=%2Fsubscriptions%2Fac8801b2-186f-4d92-8d29-117e5f0a81c8%2FresourceGroups%2Fproject\_resourcegroup%2Fprovider...

Microsoft Azure | datafactory7733

Tell us what you think, share your feedback in this quick [survey](#) about Azure Data Factory

Data Factory Validate all Publish all

Factory Resources

- Pipelines
  - blob\_to\_datalake
  - azuresql\_to\_datalake
  - Change data capture (preview)
- Datasets
  - azusqldataset
  - blobdataset
  - lakedataset
- Data flows
- Power Query

azuresql\_to\_datalake azusqldataset lakedataset

Activities

- Move & transform
  - Copy data
  - Data flow
- Synapse
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDInsight
- Iteration & conditionals
- Machine Learning
- Power Query

Trigger now

Copy data

Copy data1

Parameters Variables Settings Output

Pipeline run ID: 036caeab4-dea9-4461-bf75-5af740cfbb3b

Name	Type	Run start	Duration	Status	In
Copy data1	Copy data	2023-02-04T07:20:29.1953	00:00:12	Succeeded	Al

Show all

Customer.parquet

85°F Sunny

ENG IN 12:58 PM 2/4/2023

- Now I am adding triggers iam attaching previous trigger which is connected to blob\_to\_datalake pipeline.

Microsoft Azure | datafactory7733

Add triggers

Choose trigger...

Search

+ New

trigger1

trigger

Pipeline run ID: 036caeab4-dea9-4461-bf75-5af740cf0f1

Name	Type
Copy data1	Copy data

Close

Customer.parquet

85°F Sunny

12:58 PM 2/4/2023

Microsoft Azure | datafactory7733

Validate all resources and publish them to Data Factory

Preview experience: Off

Activities

Copy data1

Pipeline run ID: 036caeab4-dea9-4461-bf75-5af740cfbb3b

Name	Type	Run start	Duration	Status	In
Copy data1	Copy data	2023-02-04T07:20:29.195Z	00:00:12	Succeeded	4

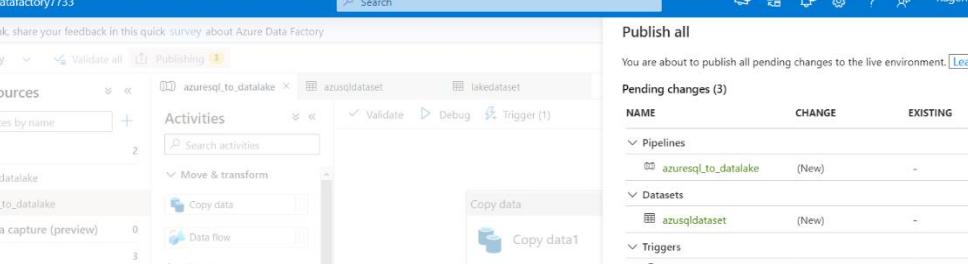
View debug run consumption

Customer.parquet

85°F Sunny

12:59 PM 2/4/2023

- Now I am publishing the pipeline.



The screenshot shows the Azure Data Factory 'Publishing' interface. A pipeline named 'azuresql\_to\_datalake' is selected. The pipeline has a single activity, 'Copy data1', which is configured to copy data from 'azusqldataset' to 'lakedataset'. The pipeline is currently in 'Live' mode and has a trigger scheduled for 12:00 AM. The 'Publish all' button is highlighted, indicating the pending changes that will be published to the live environment.

NAME	CHANGE	EXISTING
azuresql_to_datalake	(New)	-
azusqldataset	(New)	-
trigger	(Edited)	trigger

- This is our azure sql to azure datalake pipeline

## Copy data from on-premises sql server to the azure data lake

- If you want to fetch on-premises sql server to cloud in parquet file format, java jdk, jre packages must be needed
  - Search on google to java jdk download click on first link choose your os and download X64 installer
  - Link to download jdk <https://www.oracle.com/in/java/technologies/downloads/>

Java Downloads | Oracle India [oracle.com/in/java/technologies/downloads/#jdk19-windows](https://www.oracle.com/in/java/technologies/downloads/#jdk19-windows)

Java downloads Tools and resources Java archive

## Java SE Development Kit 19.0.2 downloads

Thank you for downloading this release of the Java™ Platform, Standard Edition Development Kit (JDK™). The JDK is a development environment for building applications and components using the Java programming language.

The JDK includes tools for developing and testing programs written in the Java programming language and running on the Java platform.

Product/file description	File size	Download
x64 Compressed Archive	179.15 MB	<a href="https://download.oracle.com/java/19/latest/jdk-19_windows-x64_bin.zip">https://download.oracle.com/java/19/latest/jdk-19_windows-x64_bin.zip</a> (sha256)
x64 Installer	158.91 MB	<a href="https://download.oracle.com/java/19/latest/jdk-19_windows-x64_bin.exe">https://download.oracle.com/java/19/latest/jdk-19_windows-x64_bin.exe</a> (sha256)
x64 MSI Installer	157.76 MB	<a href="https://download.oracle.com/java/19/latest/jdk-19_windows-x64_bin.msi">https://download.oracle.com/java/19/latest/jdk-19_windows-x64_bin.msi</a> (sha256)

**JDK Script-friendly URLs**

The URLs listed above will remain the same for JDK update releases to allow their use in scripts.

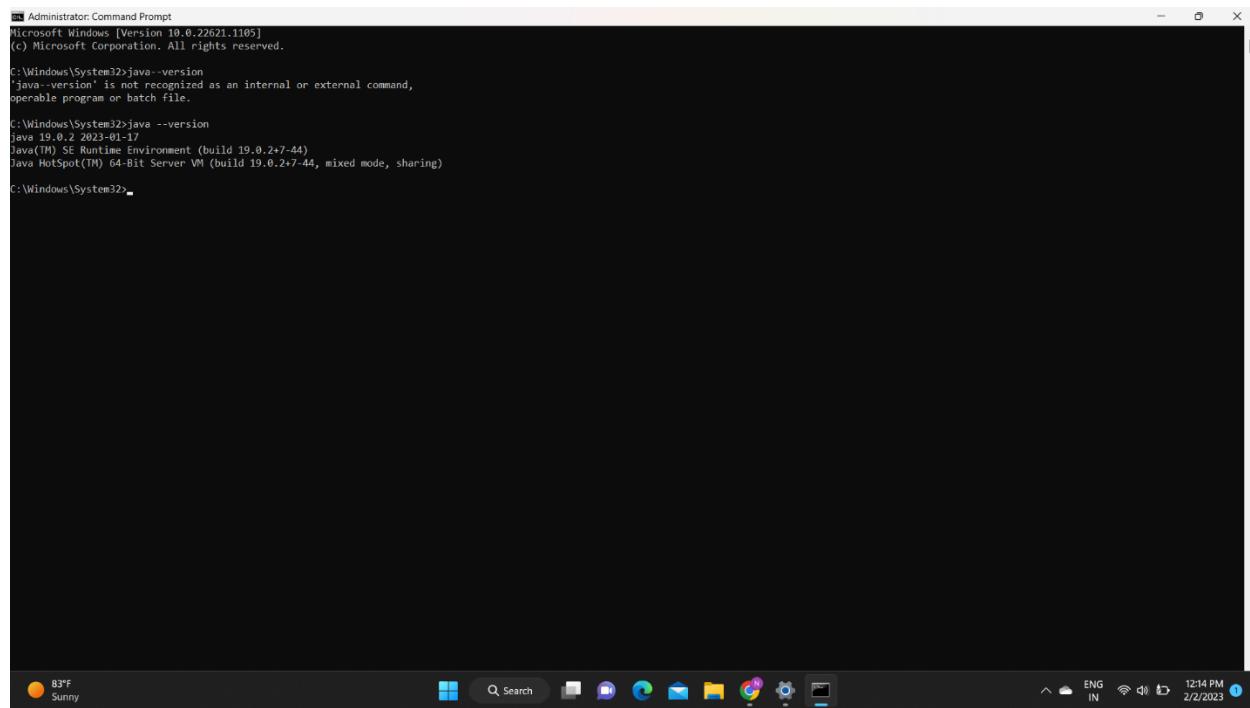
[Learn more about automating the downloads of JDK](#)

- After jdk download search on java jre download and download jre.
  - Link to download jre <https://www.java.com/en/>

A screenshot of a web browser displaying the Java website at java.com/en/. The page has a blue header with the Java logo, navigation links for 'Download', 'Developer Resources', and 'Help', and a search bar. The main content area features a sub-header 'Get Java for desktop applications', a 'Download Java' button, and links for 'What is Java?' and 'Uninstall help'. To the right is a photograph of a woman with curly hair smiling while using a laptop. The browser's address bar shows 'java.com/en/' and the title bar shows 'Java Downloads | Oracle India' and 'Java | Oracle'.

The image shows the Java.com website header. It features the Java logo (a steaming coffee cup) on the left. To its right is a search bar with the placeholder "Search Java.com". Below the search bar are two buttons: "OpenJDK Early Access Builds" and "Java SE Development Kit". The background is a dark grey gradient.

- You can check whether java packages are installed or not in window terminal



```

Administrator: Command Prompt
Microsoft Windows [Version 10.0.22621.1105]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>java --version
'java --version' is not recognized as an internal or external command,
operable program or batch file.

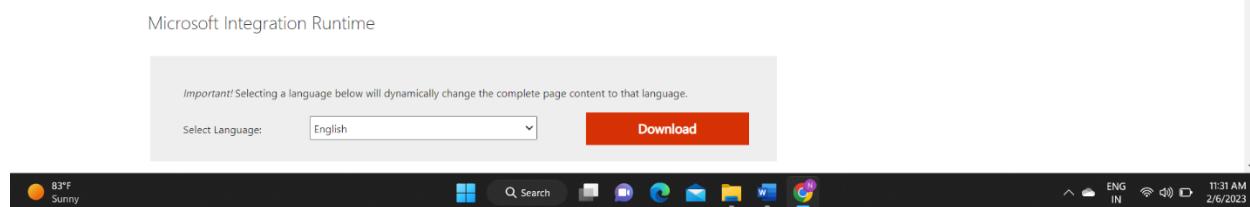
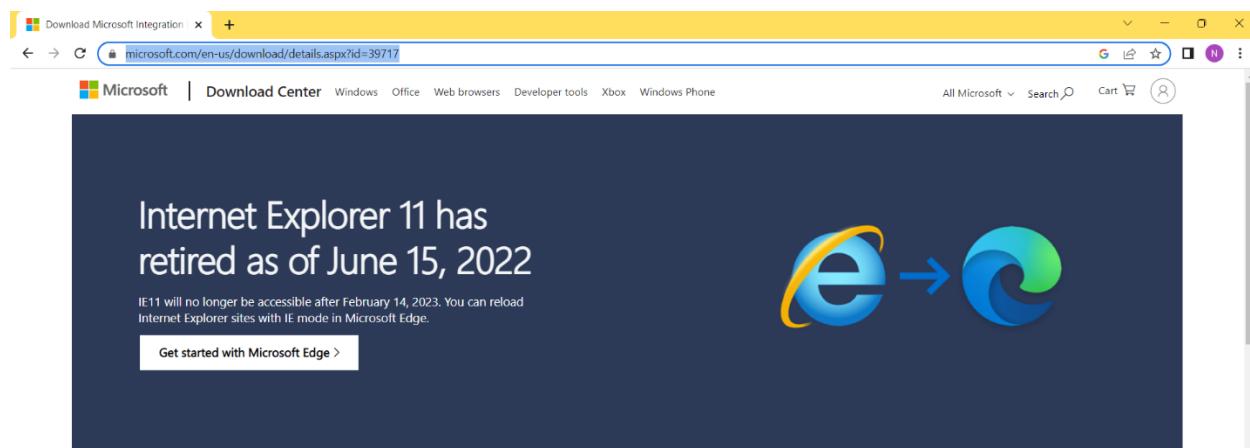
C:\Windows\System32>java --version
java 19.0.2 2023-01-17
Java(TM) SE Runtime Environment (build 19.0.2+7-44)
Java HotSpot(TM) 64-Bit Server VM (build 19.0.2+7-44, mixed mode, sharing)

C:\Windows\System32>

```

83°F Sunny 12:14 PM 2/2/2023

- There is a software Microsoft integration runtime we have to install in our on-premises server
- So it connect to azure integration runtime
- Link to download Microsoft integration runtime  
<https://www.microsoft.com/en-us/download/details.aspx?id=39717>



- Launch the azure data factory

datafactory7733 - Microsoft Azure

datafactory7733 - Data factory (V2)

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Settings

Networking

Managed identities

Properties

Locks

Getting started

Quick start

Monitoring

Alerts

Metrics

Diagnostic settings

Resource group (move) : project\_resourcegroup

Status : Succeeded

Location : Central US

Subscription (move) : Free Trial

Subscription ID : ac8801b2-186f-4d92-8d29-117e5f0a81c8

Type : Data factory (V2)

Getting started : Quick start

Azure Data Factory Studio

Launch studio

Quick Starts

Tutorials

Template Gallery

Training Modules

85°F Sunny

https://adf.azure.com/en/home?factory=%2Fsubscriptions%2Fac8801b2-186f-4d92-8d29-117e5f0a81c8%2FresourceGroups%2Fproject\_resourcegroup%2Fproviders%2FMicrosoft.DataFactory%2Ffactories%2Fdatafactory7733#loginHint=nagendra007799@gmail.com

- Go to integration runtime and one “auto resolve integration runtime” is running which is commonly used in azure integration.
- Click on +new symbol.

datafactory7733 - Microsoft Azure

datafactory7733 - Azure Data Factory

Microsoft Azure | datafactory7733

Tell us what you think, share your feedback in this quick [survey](#) about Azure Data Factory

Integration runtimes

The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environment. [Learn more](#)

+ New    Refresh

Filter by name

Showing 1 - 1 of 1 items

Name	Type	Sub-type	Status	Related	Region	Version
AutoResolveIntegrationR...	Azure	Public	Running	0	Auto Resolve	---

General

Factory settings

Connections

Linked services

Integration runtimes

Microsoft Purview

Source control

Git configuration

ARM template

Author

Triggers

Global parameters

Data flow libraries

Security

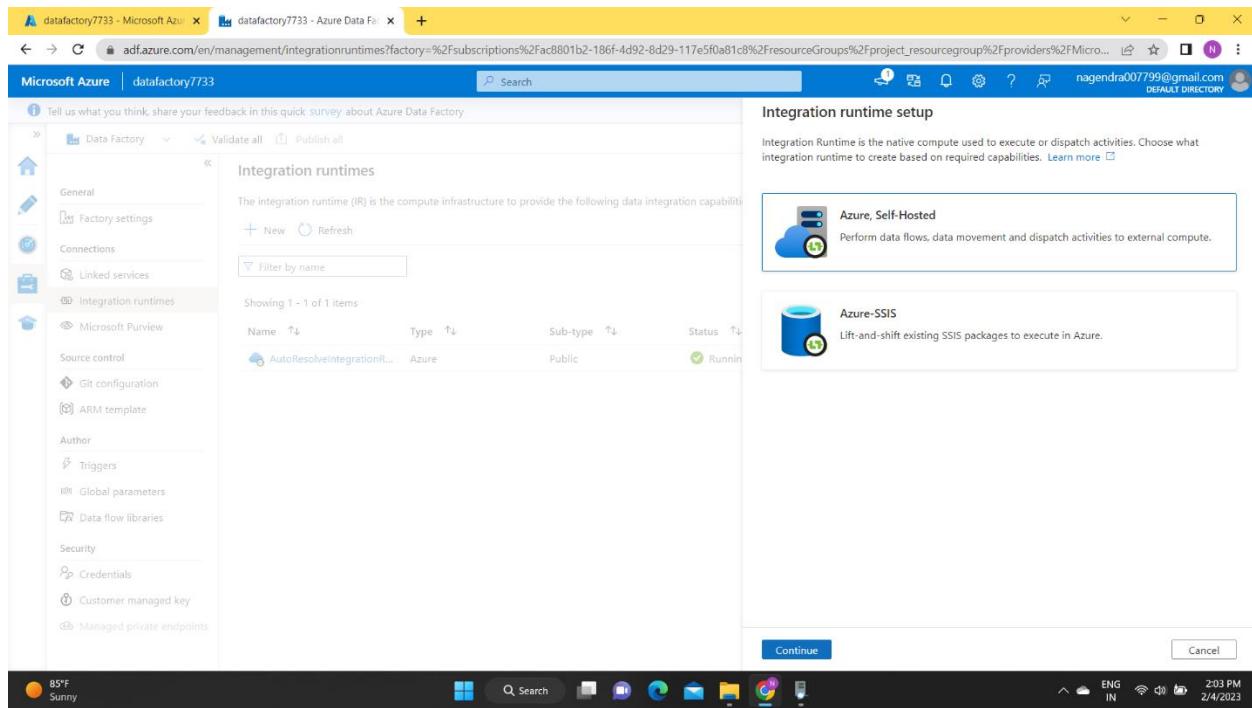
Credentials

Customer managed key

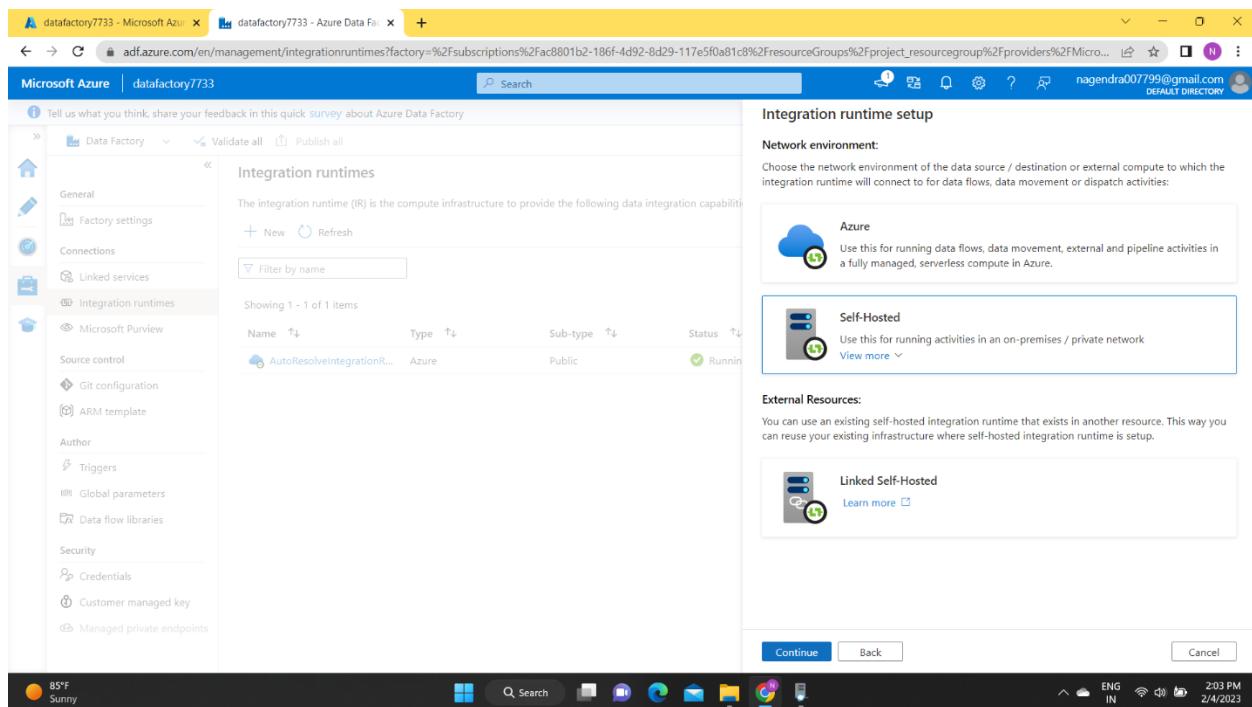
Managed private endpoints

85°F Sunny

- Select azure self hosted runtime.



The screenshot shows the 'Integration runtime setup' page in the Azure Data Factory portal. The 'Azure, Self-Hosted' option is selected, with a description: 'Perform data flows, data movement and dispatch activities to external compute.' Below it, the 'Azure-SSIS' option is also listed with its description: 'Lift-and-shift existing SSIS packages to execute in Azure.' At the bottom, there are 'Continue', 'Cancel', and 'Back' buttons.



The screenshot shows the 'Integration runtime setup' page in the Azure Data Factory portal. The 'Self-Hosted' option is selected, with a description: 'Use this for running activities in an on-premises / private network' and a 'View more' link. Below it, the 'Azure' option is listed with its description: 'Use this for running data flows, data movement, external and pipeline activities in a fully managed, serverless compute in Azure.' At the bottom, there are 'Continue', 'Back', and 'Cancel' buttons.

- Give the name to the integration runtime “ironpremsql”.

Integration runtime setup

Private network support is realized by installing integration runtime to machines in the same on-premises network/VNET as the resource the integration runtime is connecting to. Follow below steps to register and install integration runtime on your self-hosted machines.

Name \*

Description

Type

Create Back Cancel

- Copy any link from these two.

Integration runtime setup

Settings Nodes Auto update Sharing Links

Install integration runtime on Windows machine or add further nodes using the Authentication Key.

Name

Option 1: Express setup  
[Click here to launch the express setup for this computer](#)

Option 2: Manual setup

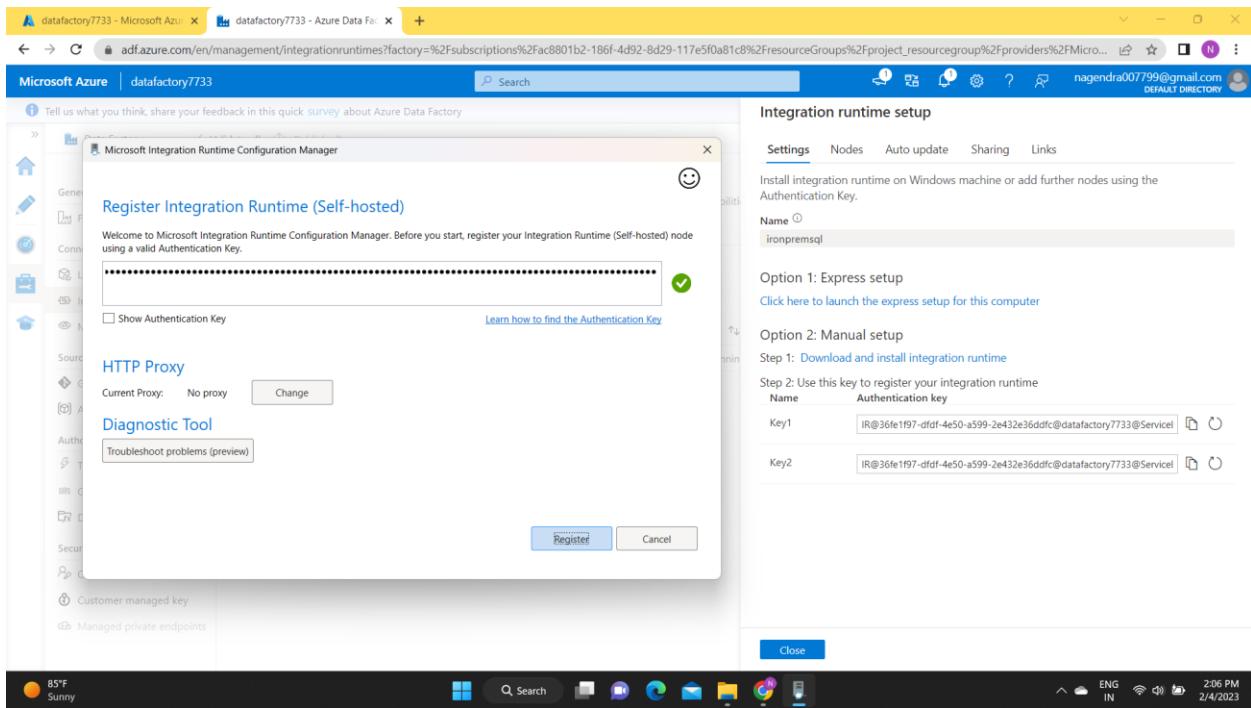
Step 1: [Download and install integration runtime](#)

Step 2: Use this key to register your integration runtime

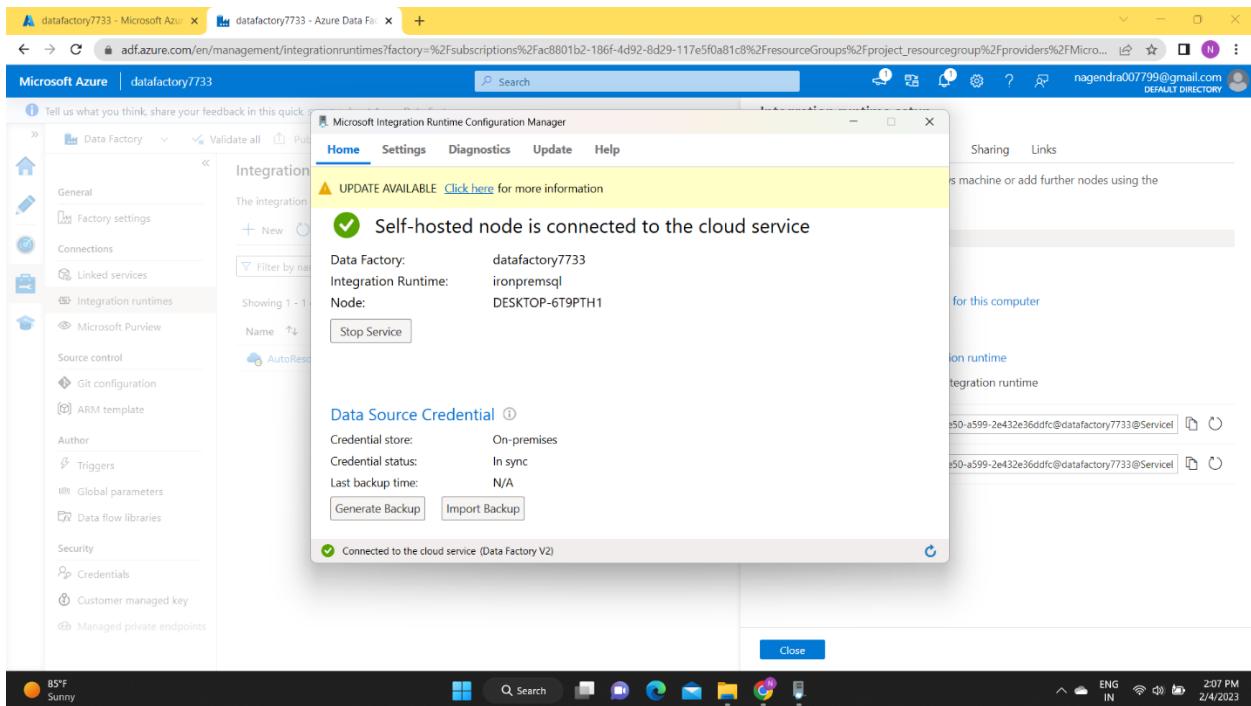
Name	Authentication key
Key1	IR@36fe1f97-dfdf-4e50-a599-2e432e36ddfc@datafactory7733@Service
Key2	IR@36fe1f97-dfdf-4e50-a599-2e432e36ddfc@datafactory7733@Service

Close

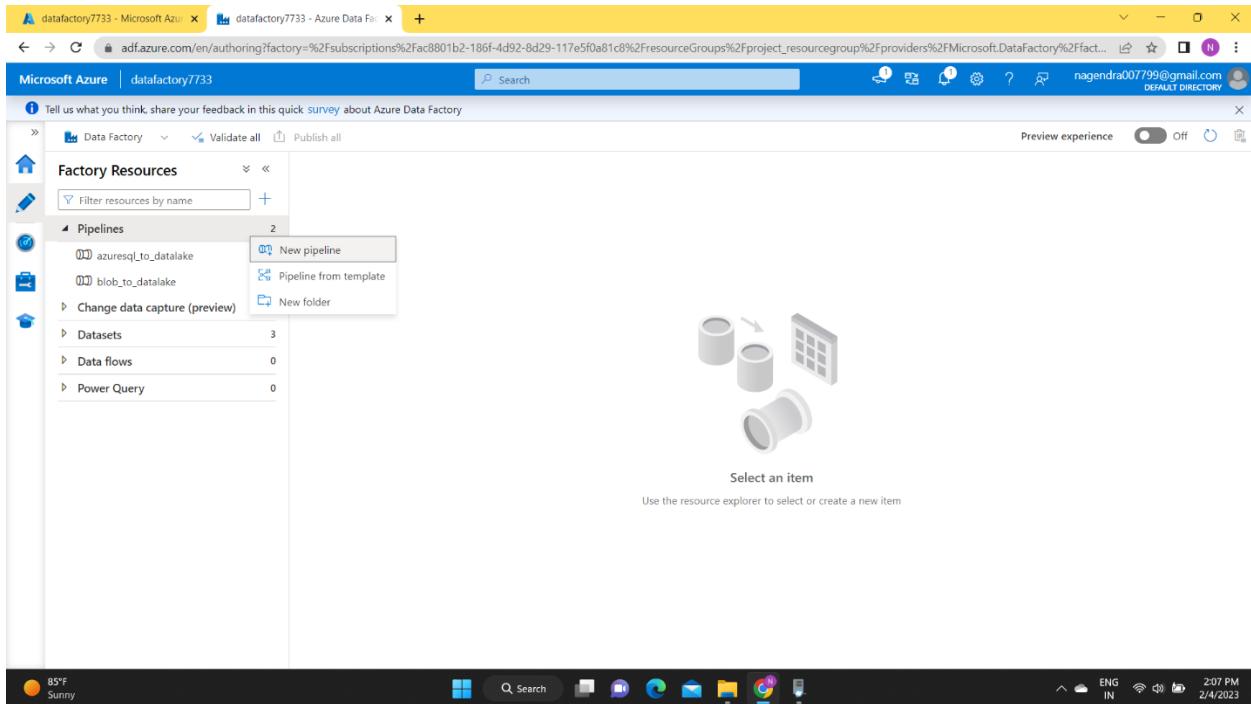
- Open windows integration runtime and paste the link in that box.



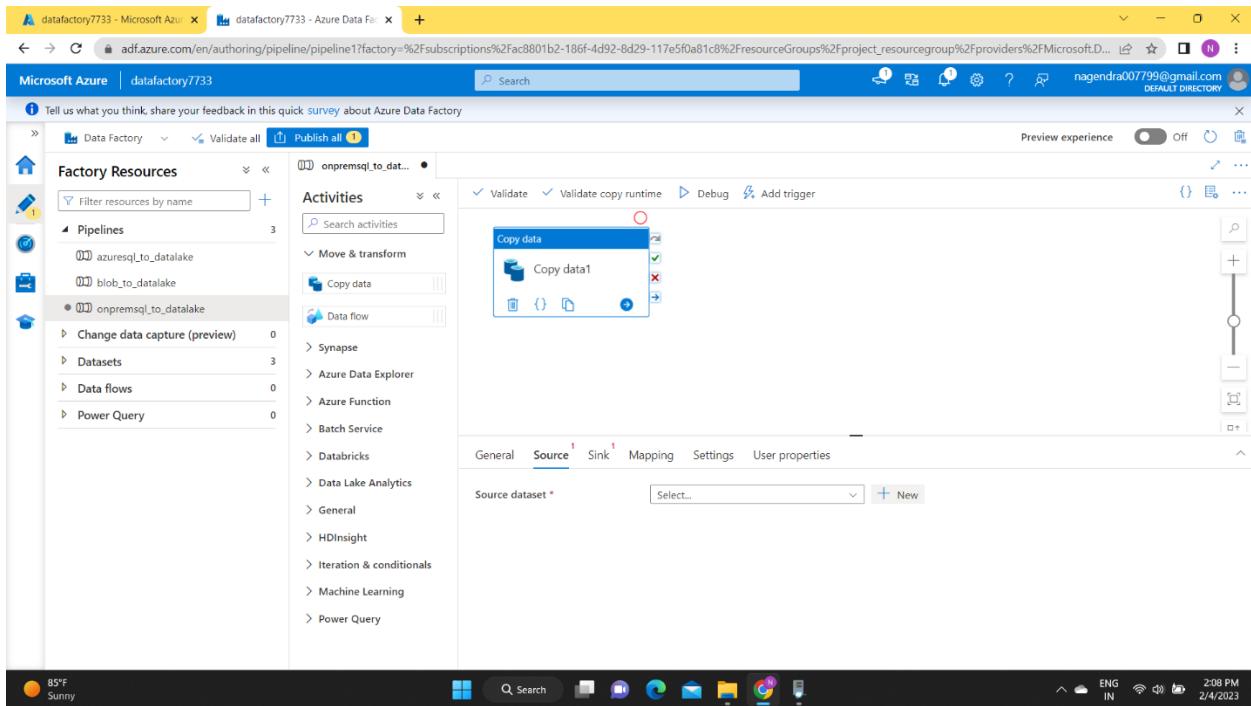
- Now our local windows server connected to azure cloud.



- Open the adf and create new pipe line “onpremsql\_to\_azuresqldb.



- Drag the copy activity from move and transform.



- We don't need to create sql server we are using on-premises sql server details .
- Select sql server for the source dataset.

The screenshot shows the Azure Data Factory pipeline editor. On the left, the 'Factory Resources' pane is open, showing a list of pipelines, datasets, data flows, and Power Query. In the center, a 'Copy data' activity is selected. The 'Source' tab is active, and the 'Select...' button is highlighted. On the right, a 'New dataset' dialog box is open, displaying a grid of data store icons. The 'SQL' icon is selected, indicating it as the source dataset for the copy activity. The status bar at the bottom shows the date and time as 2/4/2023, 2:09 PM.

- Click on create new link service to on-premises sql server.
- Sql server dataset name is “onpremsqldataset”

The screenshot shows the Azure Data Factory pipeline editor. On the left, the 'Factory Resources' pane is open, showing a list of pipelines, datasets, data flows, and Power Query. In the center, a 'Copy data' activity is selected. The 'Source' tab is active, and the 'Select...' button is highlighted. On the right, a 'Set properties' dialog box is open for the 'onpremsqldataset'. The 'Name' field is set to 'onpremsqldataset'. The 'Linked service' dropdown is open, showing a list of options. The 'Select...' button is highlighted. The 'OK' button is at the bottom left of the dialog box. The status bar at the bottom shows the date and time as 2/4/2023, 2:10 PM.

- Our SQL server link service name is “onpremssqllink”
- In connect via integration runtime box select our “ironpremssql”
- You can fill remaining details by using below pic

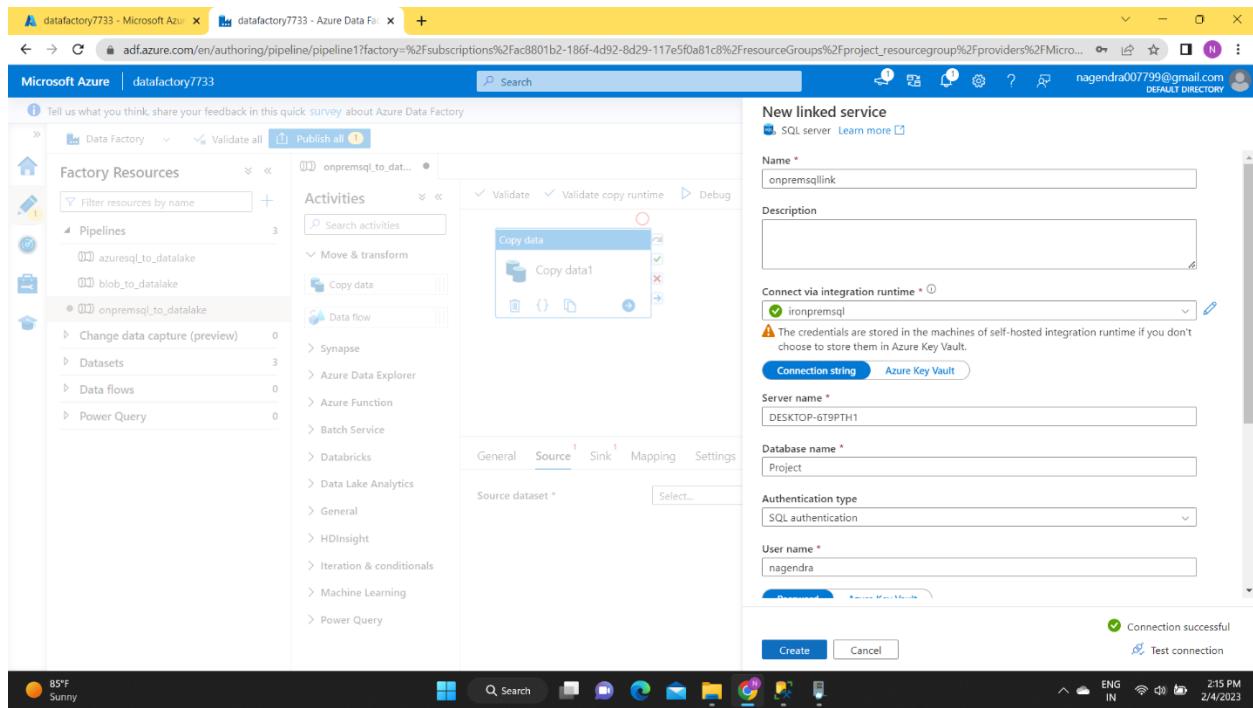
The screenshot shows the Azure Data Factory portal. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. A pipeline named 'onpremssql\_to\_datalake' is selected. In the main area, a 'Copy data' activity is highlighted. On the right, a 'New linked service' dialog is open. The 'Source' tab is selected, showing the 'onpremssql' connection. The 'Connect via integration runtime' dropdown is set to 'AutoResolveIntegrationRuntime', and the 'ironpremssql' runtime is selected. Other fields include 'Database name' (ironpremssql), 'Authentication type' (SQL authentication), 'User name' (nagendra), and 'Password' (a masked field). A 'Create' button is at the bottom.

The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface. The title bar says 'Microsoft SQL Server Management Studio'. The main area displays a 'Connect to Server' dialog for a 'SQL Server' instance. The dialog fields are as follows:
 

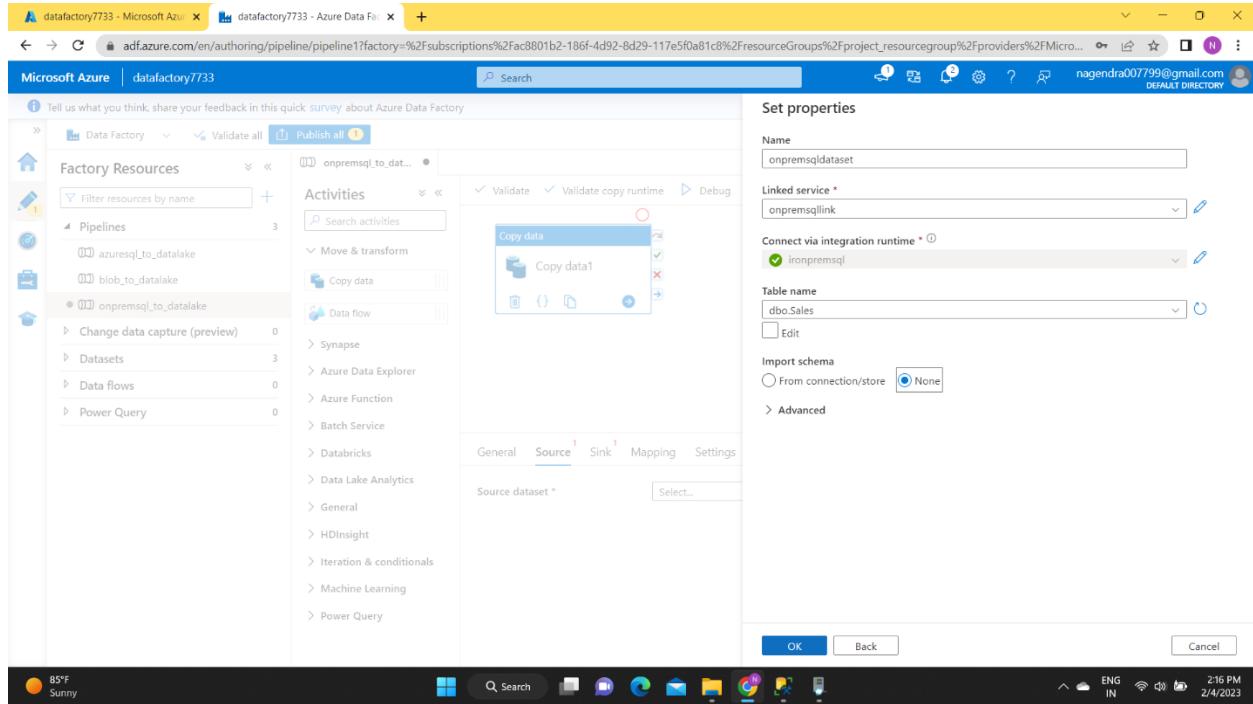
- Server type: Database Engine
- Server name: DESKTOP-619PT1H
- Authentication: SQL Server Authentication
- Login: nagendra
- Password: (masked)
- Remember password:

 At the bottom of the dialog are 'Connect', 'Cancel', 'Help', and 'Options >' buttons.

- Do the test connection.



- In dataset select table name and put import schema is none.



- Remove dbo.sales to sales. By clicking edit button.

Properties

General Related (1)

Name \* onpremsqldataset

Description

Annotations + New

Preview data

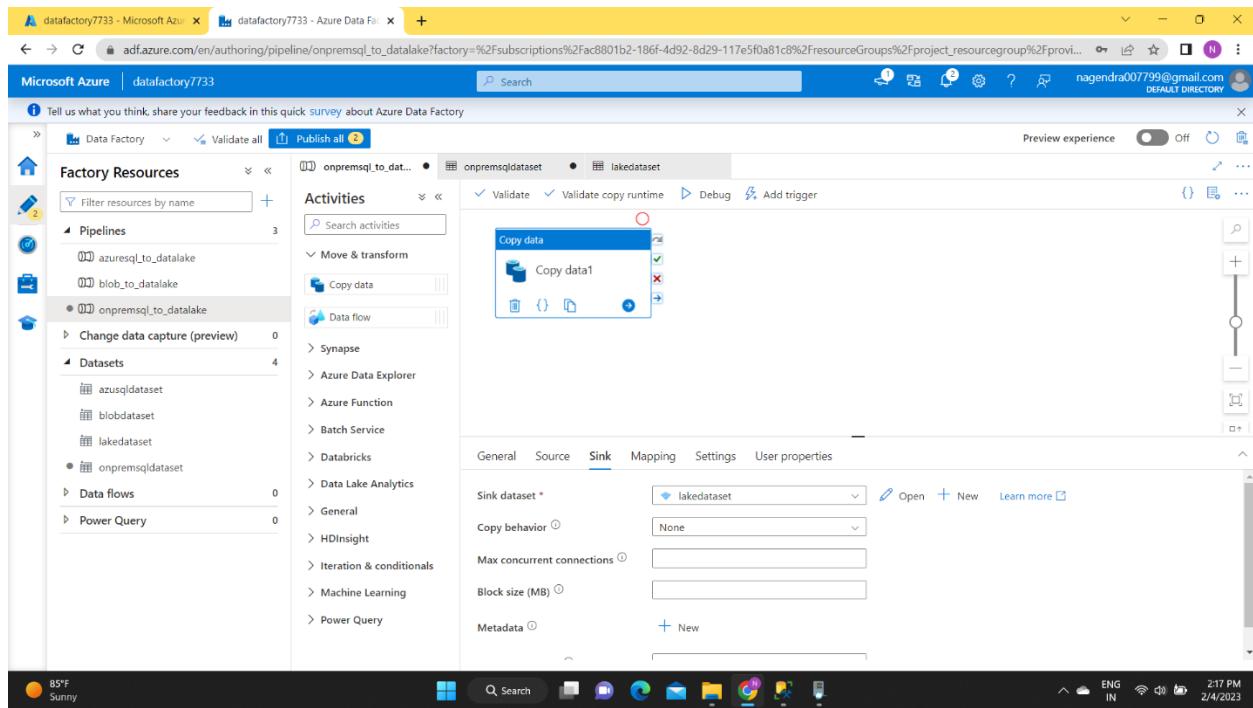
- This is the preview of our sales table.

Sales_ID	Product_ID	Customer_ID	Sales_Date	Sales_Quantity	Sales_Amount
1	1	1	01/01/2022	2	38.00
2	2	2	01/02/2022	1	25.00
3	3	3	01/03/2022	1	199.00
4	4	4	01/04/2022	2	298.00
5	5	5	01/05/2022	3	87.00
6	6	6	01/06/2022	1	99.00
7	7	7	01/07/2022	2	158.00
8	8	8	01/08/2022	1	59.00
9	9	9	01/09/2022	2	78.00
10	10	10	01/10/2022	1	199.00

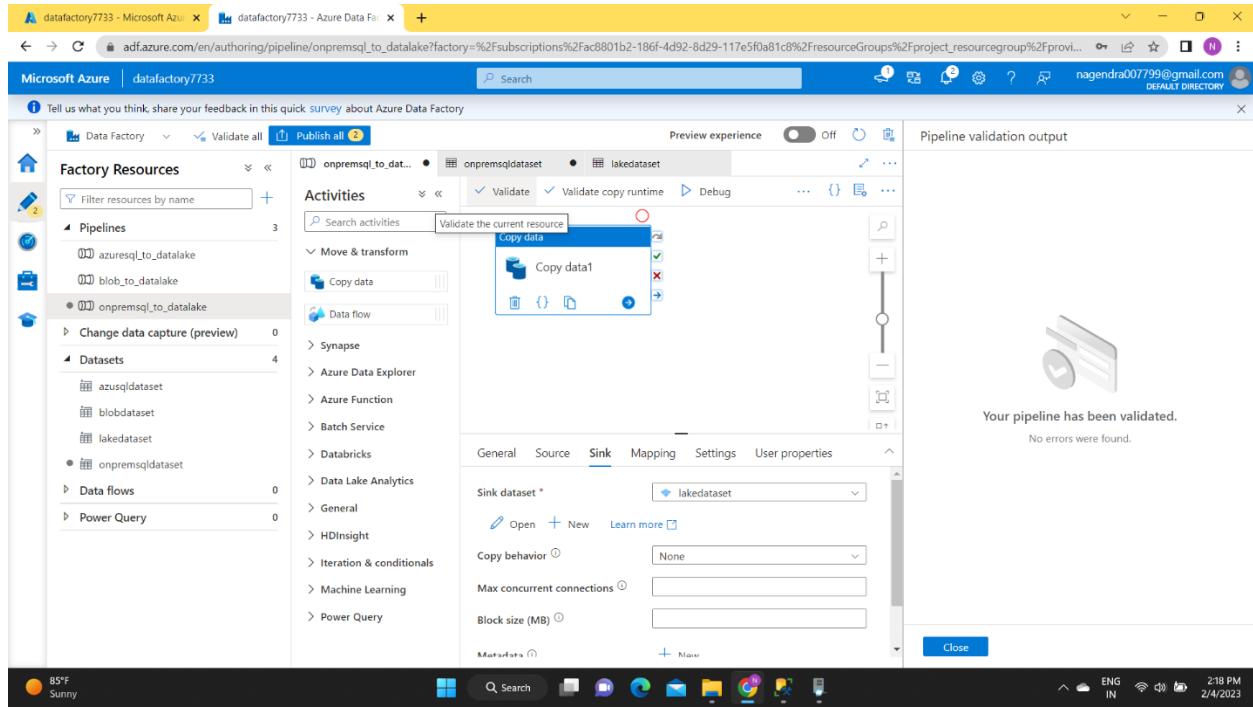
Partition option: None Physical partitions of table Dynamic range

Please preview data to validate the partition settings are correct before you trigger a run or publish the pipeline.

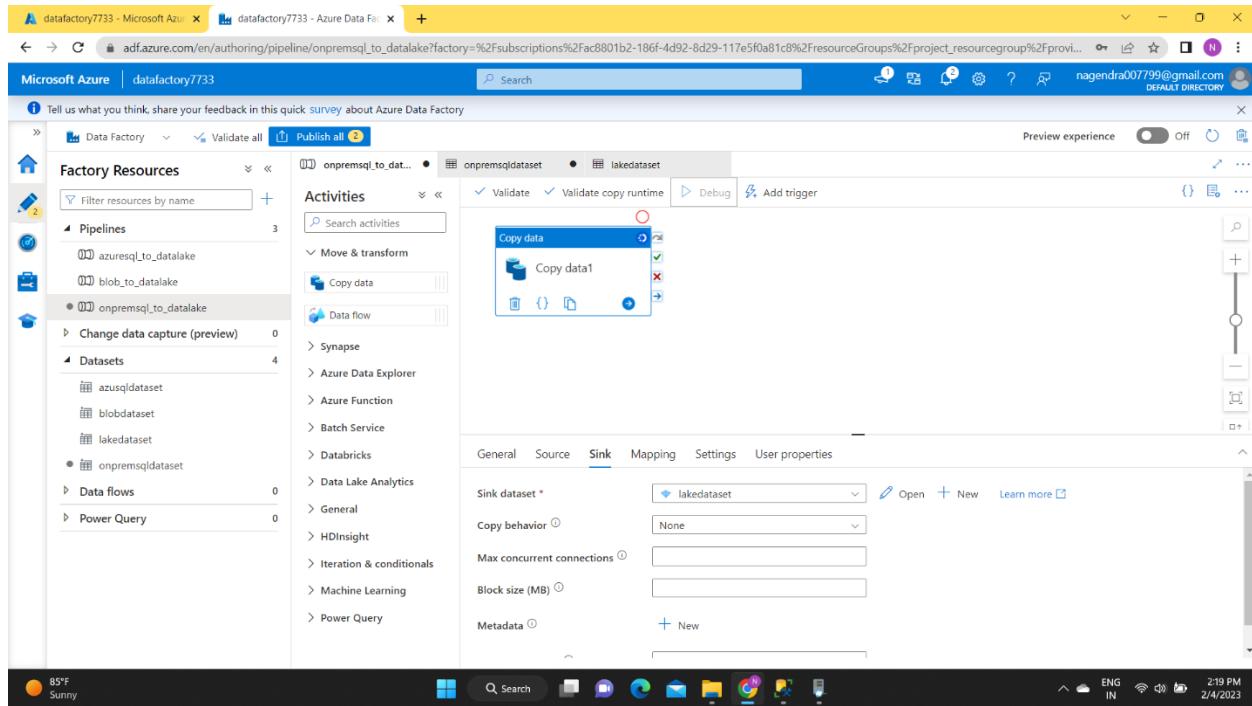
- In sink we have data lake dataset .



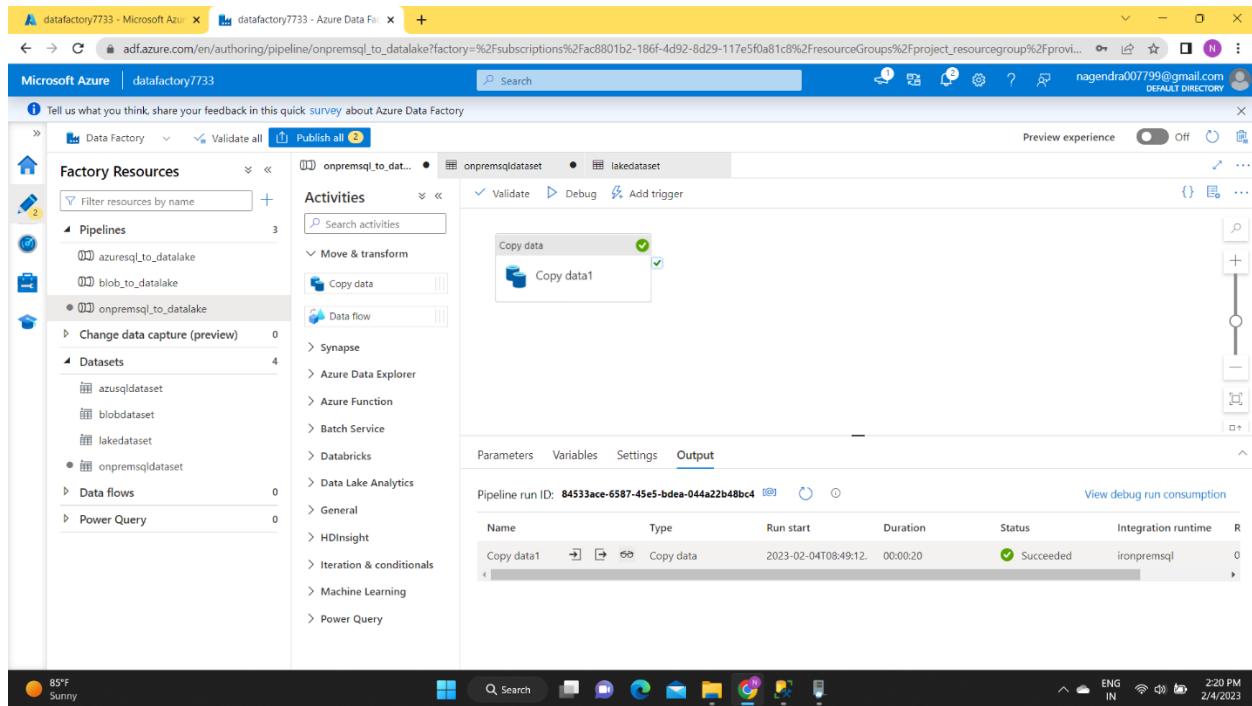
- Validate our pipeline .



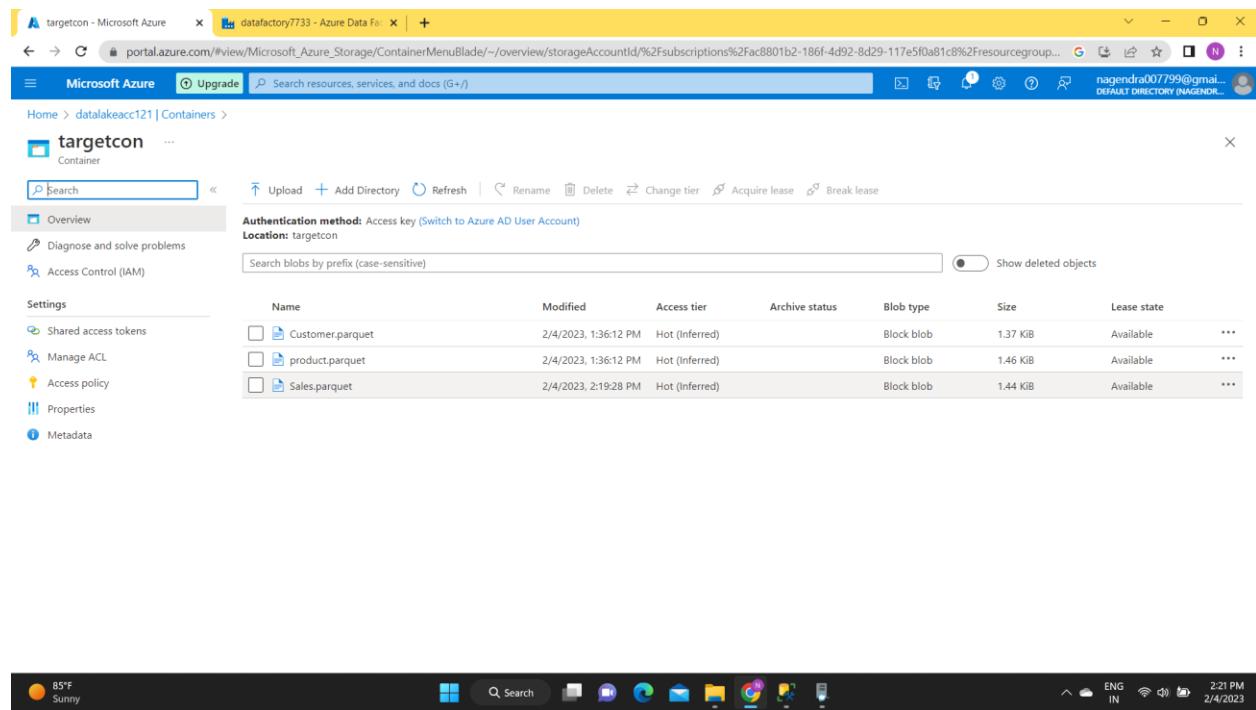
- Now do the debug.



- Our pipeline is succed

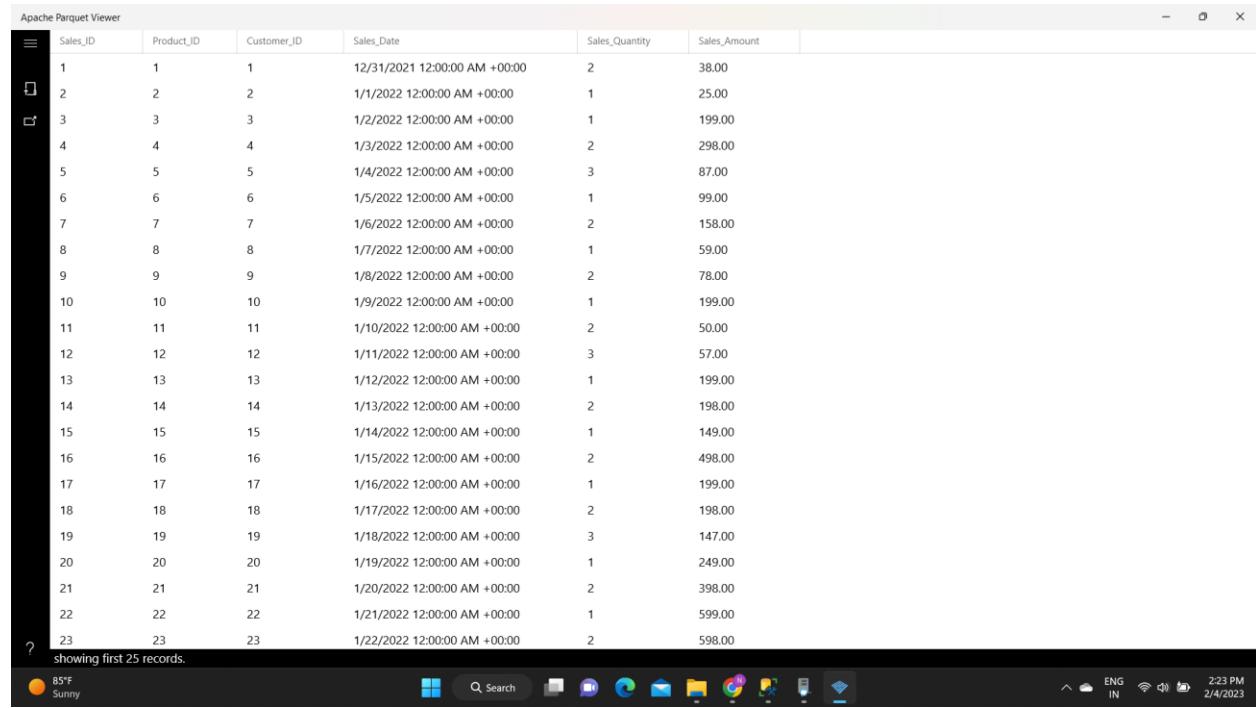


- This is our sales.parquet table in unified data lake.



The screenshot shows the Azure Data Factory portal with the URL [https://portal.azure.com/#view/Microsoft\\_Azure\\_Storage/ContainerMenuBlade/~/overview/storageAccountId/%2Fsubscriptions%2FfacB801b2-186f-4d92-8d29-117e5f0a81c8%2Fresourcegroup...](https://portal.azure.com/#view/Microsoft_Azure_Storage/ContainerMenuBlade/~/overview/storageAccountId/%2Fsubscriptions%2FfacB801b2-186f-4d92-8d29-117e5f0a81c8%2Fresourcegroup...). The page displays the contents of the 'targetcon' container, which contains three parquet files: Customer.parquet, product.parquet, and Sales.parquet. The table includes columns for Name, Modified, Access tier, Archive status, Blob type, Size, and Lease state. The interface also shows options for Upload, Add Directory, Refresh, Rename, Delete, Change tier, Acquire lease, and Break lease. The location is set to 'targetcon' and the authentication method is 'Access key (Switch to Azure AD User Account)'. The status bar at the bottom shows the date as 2/4/2023 and the time as 2:21 PM.

- This is our sales data view.



The screenshot shows the Apache Parquet Viewer application displaying the contents of the sales.parquet file. The data is presented in a table with columns: Sales\_ID, Product\_ID, Customer\_ID, Sales\_Date, Sales\_Quantity, and Sales\_Amount. The table contains 25 records, with the first few rows shown below. The application interface includes a toolbar with icons for file operations and a status bar at the bottom showing the date as 2/4/2023 and the time as 2:23 PM.

Sales_ID	Product_ID	Customer_ID	Sales_Date	Sales_Quantity	Sales_Amount
1	1	1	12/31/2021 12:00:00 AM +00:00	2	38.00
2	2	2	1/1/2022 12:00:00 AM +00:00	1	25.00
3	3	3	1/2/2022 12:00:00 AM +00:00	1	199.00
4	4	4	1/3/2022 12:00:00 AM +00:00	2	298.00
5	5	5	1/4/2022 12:00:00 AM +00:00	3	87.00
6	6	6	1/5/2022 12:00:00 AM +00:00	1	99.00
7	7	7	1/6/2022 12:00:00 AM +00:00	2	158.00
8	8	8	1/7/2022 12:00:00 AM +00:00	1	59.00
9	9	9	1/8/2022 12:00:00 AM +00:00	2	78.00
10	10	10	1/9/2022 12:00:00 AM +00:00	1	199.00
11	11	11	1/10/2022 12:00:00 AM +00:00	2	50.00
12	12	12	1/11/2022 12:00:00 AM +00:00	3	57.00
13	13	13	1/12/2022 12:00:00 AM +00:00	1	199.00
14	14	14	1/13/2022 12:00:00 AM +00:00	2	198.00
15	15	15	1/14/2022 12:00:00 AM +00:00	1	149.00
16	16	16	1/15/2022 12:00:00 AM +00:00	2	498.00
17	17	17	1/16/2022 12:00:00 AM +00:00	1	199.00
18	18	18	1/17/2022 12:00:00 AM +00:00	2	198.00
19	19	19	1/18/2022 12:00:00 AM +00:00	3	147.00
20	20	20	1/19/2022 12:00:00 AM +00:00	1	249.00
21	21	21	1/20/2022 12:00:00 AM +00:00	2	398.00
22	22	22	1/21/2022 12:00:00 AM +00:00	1	599.00
23	23	23	1/22/2022 12:00:00 AM +00:00	2	598.00

showing first 25 records.

- Now add trigger.

The screenshot shows the Azure Data Factory pipeline editor. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. The main workspace displays a pipeline with a single 'Copy data' activity named 'Copy data1'. The 'Output' tab is selected, showing the pipeline run ID: 84533ace-6587-45e5-bdea-044a22b48bc4, run start: 2023-02-04T08:49:12, duration: 00:00:20, status: Succeeded, and integration runtime: ironpremsql. The status bar at the bottom indicates the weather as 85°F and sunny.

- Here I am giving the previous trigger.

The screenshot shows the Azure Data Factory pipeline editor with the 'Add triggers' dialog open. The dialog has a search bar and a list of triggers: 'trigger1' and 'trigger'. The main workspace shows the same pipeline and activity as the previous screenshot. The status bar at the bottom indicates the weather as 85°F and sunny.

- Now I am going to publish the pipeline.

The screenshot shows the Azure Data Factory 'Publishing' interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, and Activities. In the center, a 'Copy data' activity is selected under 'Activities'. On the right, the 'Publish all' dialog shows pending changes:

NAME	CHANGE	EXISTING
Pipelines	onpremssql_to_datalake (New)	-
Datasets	onpremssqldataset (New)	-
Triggers	trigger (Edited)	trigger

At the bottom right of the dialog are 'Publish' and 'Cancel' buttons.

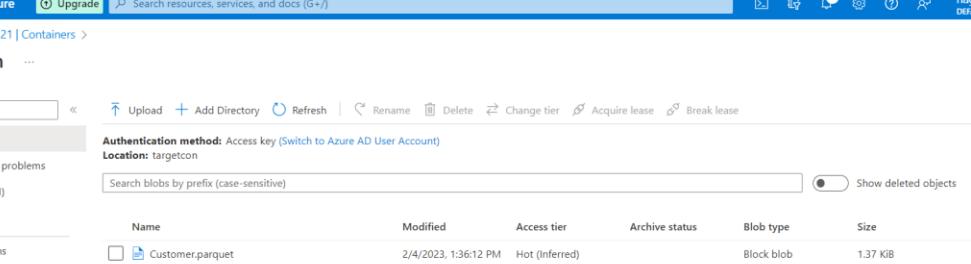
- This is our on-premises sql server\_ to azure data lake pipeline.

The screenshot shows the Azure Data Factory 'Details' view for a copy activity. The activity is named 'onpremssql\_to\_datalake'. The 'From' source is 'SQL server' and the 'To' sink is 'Azure Data Lake Storage Gen2' (Region: Central US). The activity status is 'Succeeded'. Key performance metrics are displayed:

Source (SQL server)	Sink (Azure Data Lake Storage Gen2)
Data read: 1.1 KB	Data written: 1.478 KB
Rows read: 25	Files written: 1
Peak connections: 1	Rows written: 25
Peak connections: 1	Peak connections: 1

Below the metrics, the 'Copy duration' is listed as 00:00:11 with a throughput of 275 bytes/s. The 'Transfer' details show a 'Time to first byte' of 00:00:02, 'Reading from source' of 00:00:00, and 'Writing to sink' of 00:00:03, totaling 00:00:06.

- THIS IS OUR UNIFIED STORAGE SYSTEM AND ALL 3 TABLES



targetcon - Microsoft Azure datafactory7733 - Azure Data Factory +

portal.azure.com/#view/Microsoft\_Azure\_Storage/ContainerMenuBlade/~/overview/storageAccountId/%2fsubscriptions%2fac8801b2-186f-4d92-8d29-117e5f0a81c8%2fresourcegroup...

Microsoft Azure Upgrade Search resources, services, and docs (G+)

Home > [datalakeacct121](#) | Containers > targetcon

**targetcon** Container

Search

Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease

Authentication method: Access key (Switch to Azure AD User Account)

Location: targetcon

Search blobs by prefix (case-sensitive)  Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	...
Customer.parquet	2/4/2023, 1:36:12 PM	Hot (Inferred)		Block blob	1.37 kB	Available	...
product.parquet	2/4/2023, 1:36:12 PM	Hot (Inferred)		Block blob	1.46 kB	Available	...
Sales.parquet	2/4/2023, 2:19:28 PM	Hot (Inferred)		Block blob	1.44 kB	Available	...

Sales.parquet Show all

85°F Sunny ENG IN 2:25 PM 2/4/2023