

# 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.

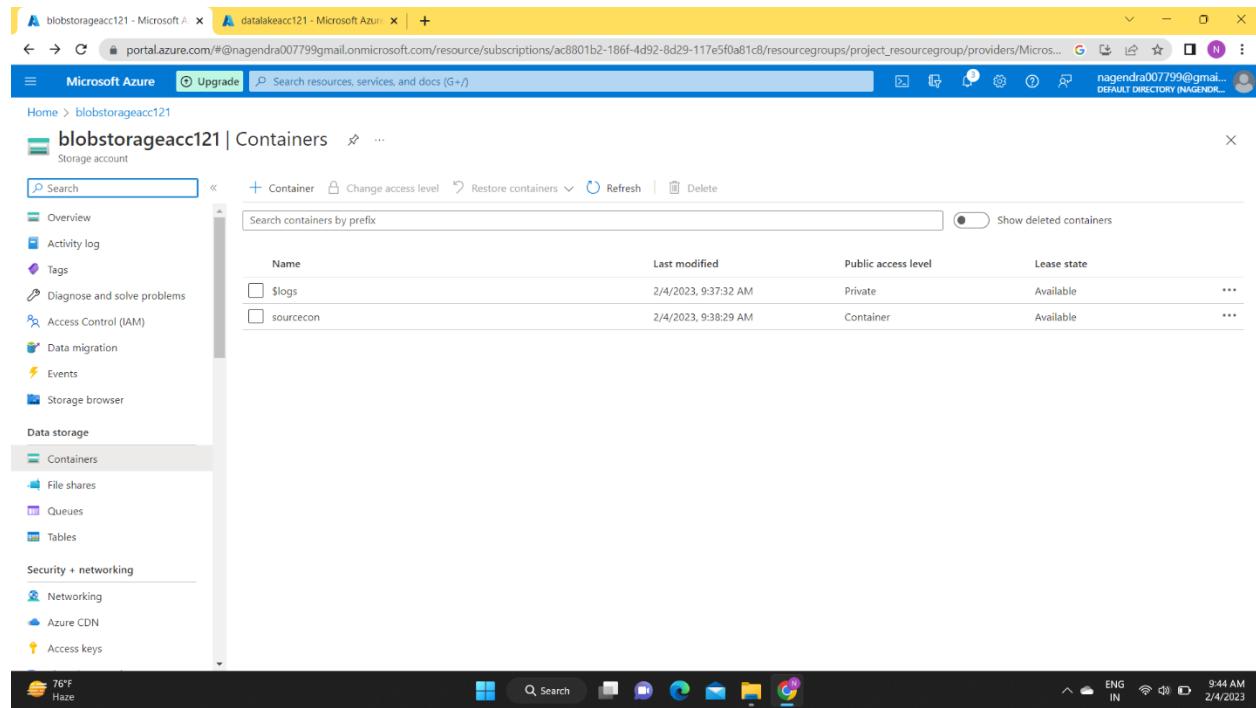
**Data integration:-** Data integration is the process of combining data from various sources into a unified, centralized repository for analysis and reporting.

**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.

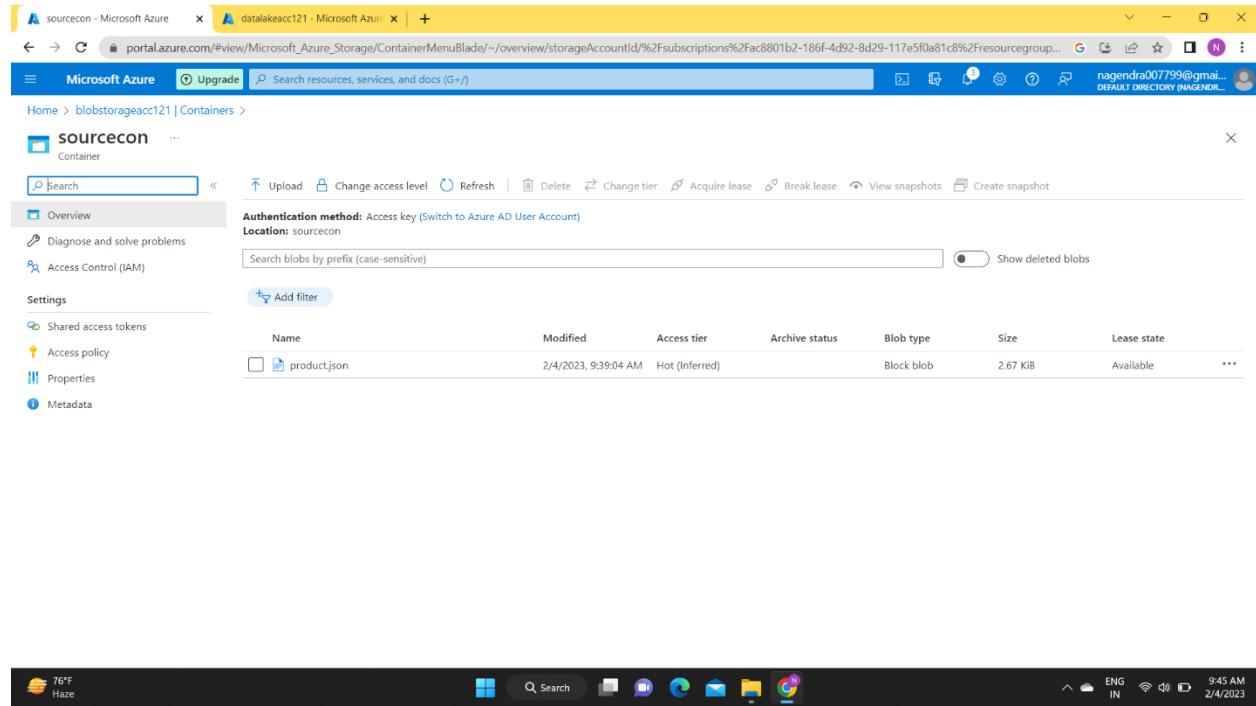
- This is my blob storage “blobstorageacc121” and container name ‘sourcecon’.



The screenshot shows the Microsoft Azure portal interface. The left sidebar is for the storage account 'blobstorageacc121'. The main content area shows the 'Containers' section with a table of containers. One container, 'sourcecon', is selected. The table data is as follows:

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

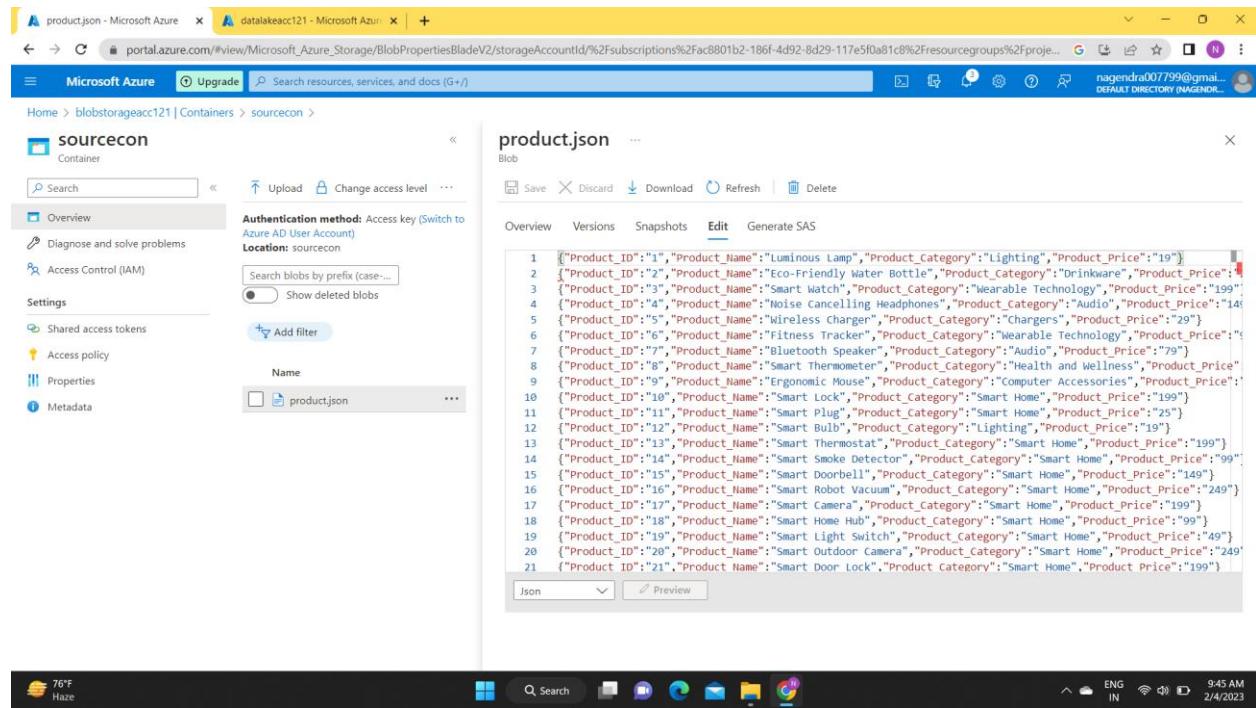
- In this blob storage, sourcecon we have product table. The product table is in json file format.



The screenshot shows the Microsoft Azure portal interface, specifically the 'sourcecon' container within the 'blobstorageacc121' storage account. The main content area shows the 'Blobs' section with a table of blobs. One blob, 'product.json', is listed. The table data is as follows:

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

- Here we can see the data in the product table



The screenshot shows the Azure Storage Blob browser interface. The left sidebar shows the container 'sourcecon' with various settings like Overview, Access Control, and Shared access tokens. The main area displays the 'product.json' file content, which is a JSON array of 21 product objects. The JSON data is as follows:

```

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"}  

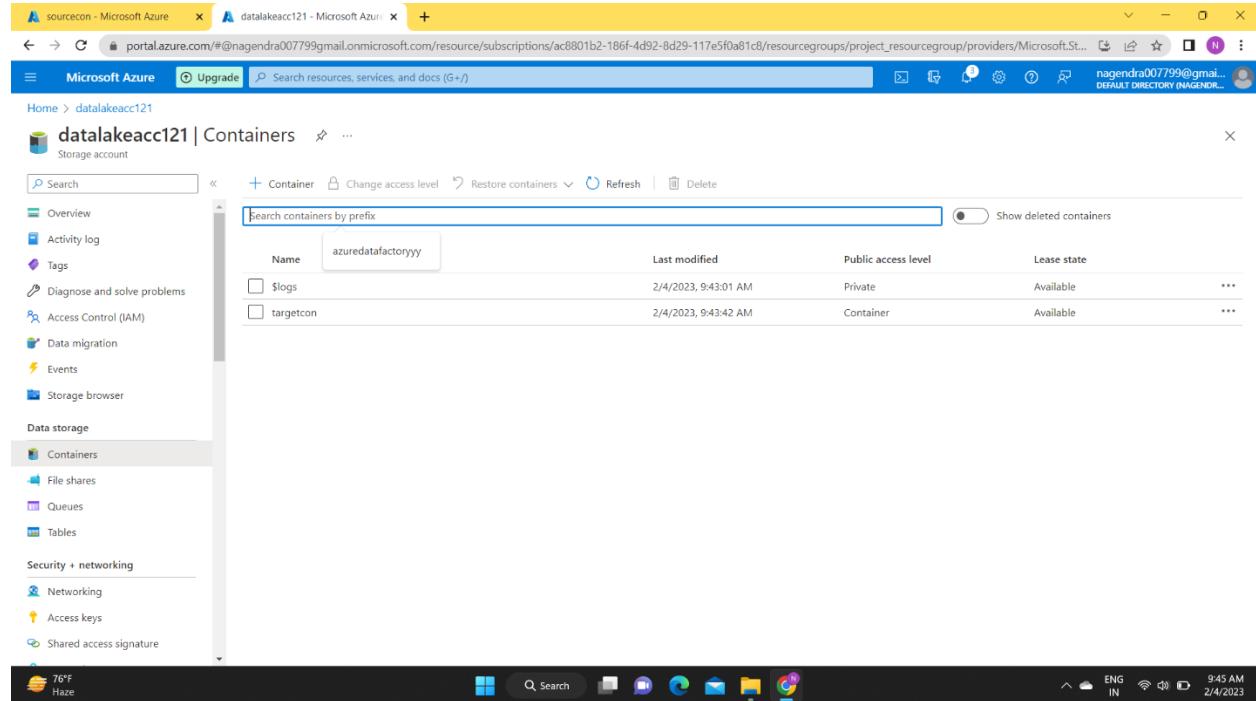
20 {"Product_ID": "20", "Product_Name": "Smart Outdoor Camera", "Product_Category": "Smart Home", "Product_Price": "249"}  

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

```

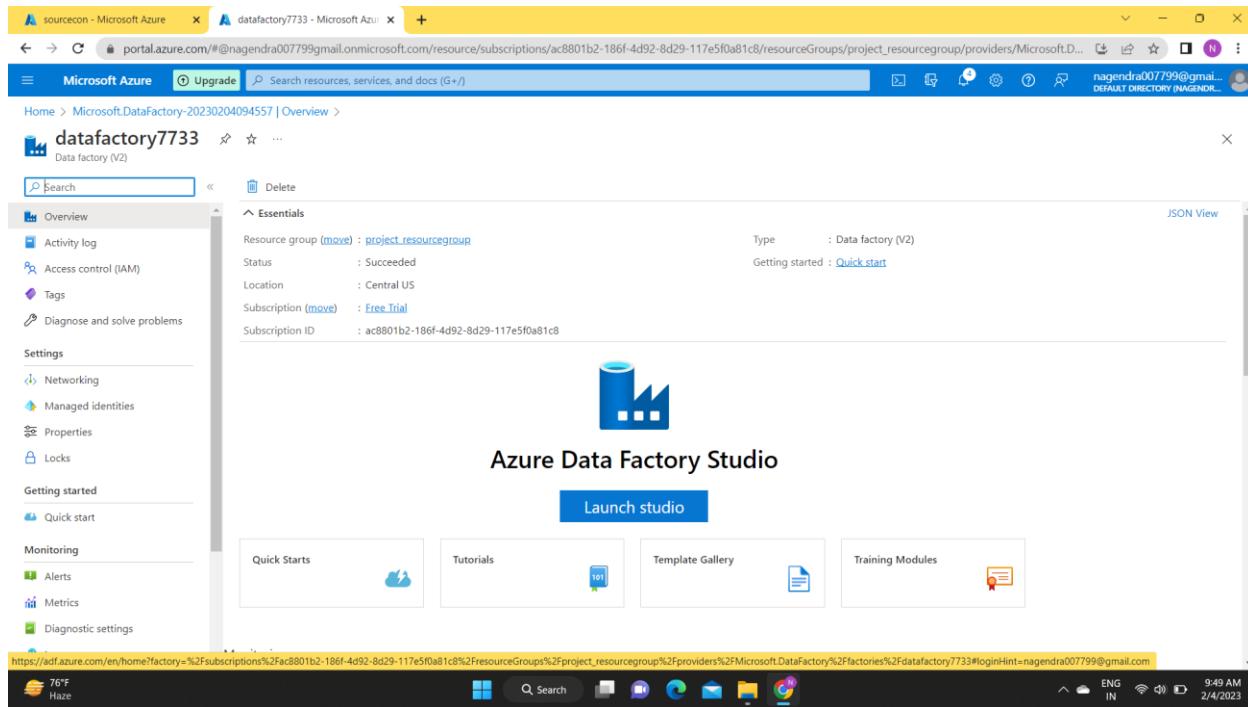
The interface includes tabs for Overview, Versions, Snapshots, Edit, and Generate SAS. The JSON tab is selected. The preview tab shows a sample of the JSON data.

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

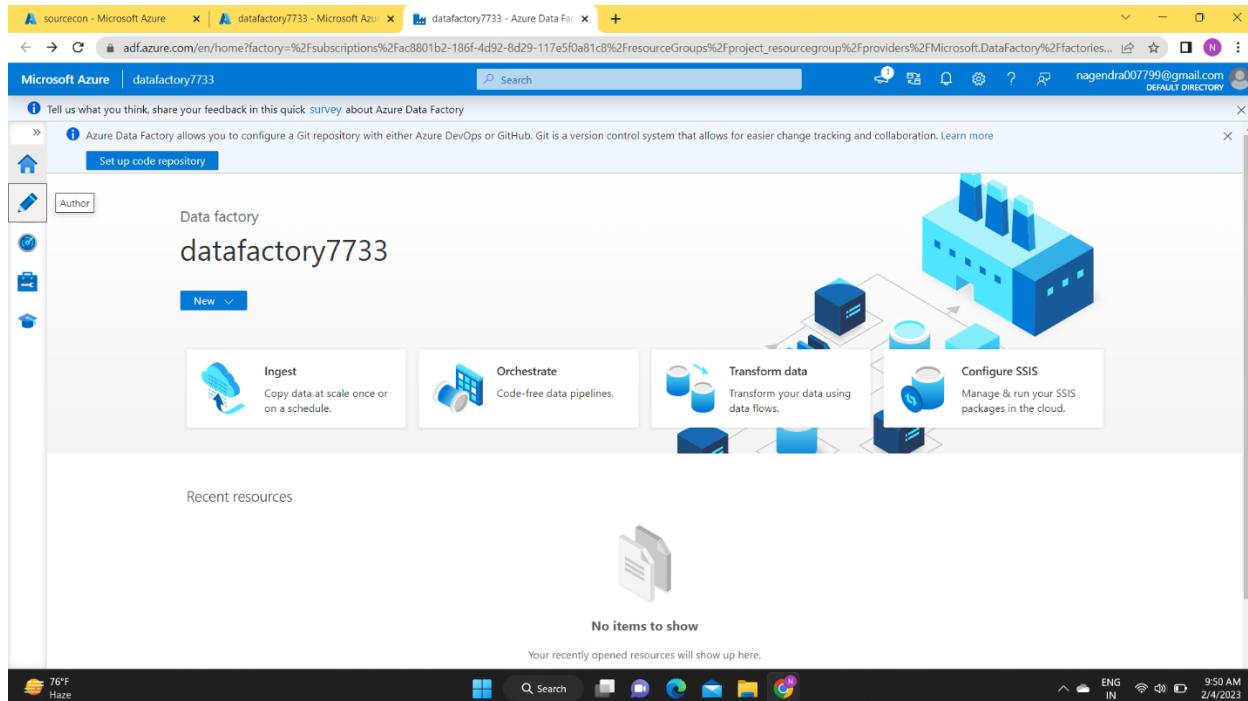


The screenshot shows the Azure Storage Container browser interface. The left sidebar shows the storage account 'datalakeacc121' with various settings like Overview, Activity log, and Data migration. The main area displays the 'targetcon' container content, which is an empty container. The interface includes tabs for Container, Change access level, Restore containers, Refresh, and Delete. The JSON tab is selected. The preview tab shows a sample of the JSON data.

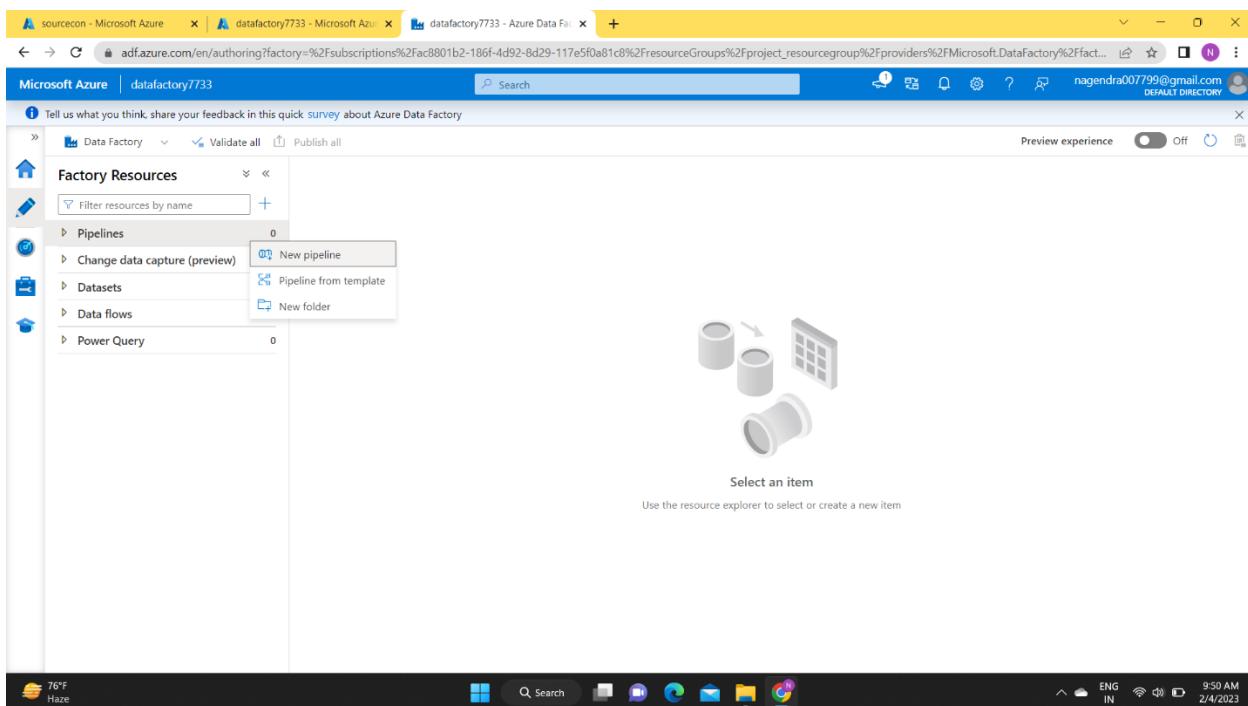
- Our azure data factory name is “datafactory7733”
- We are launching our datafactory by clicking “launch studio”



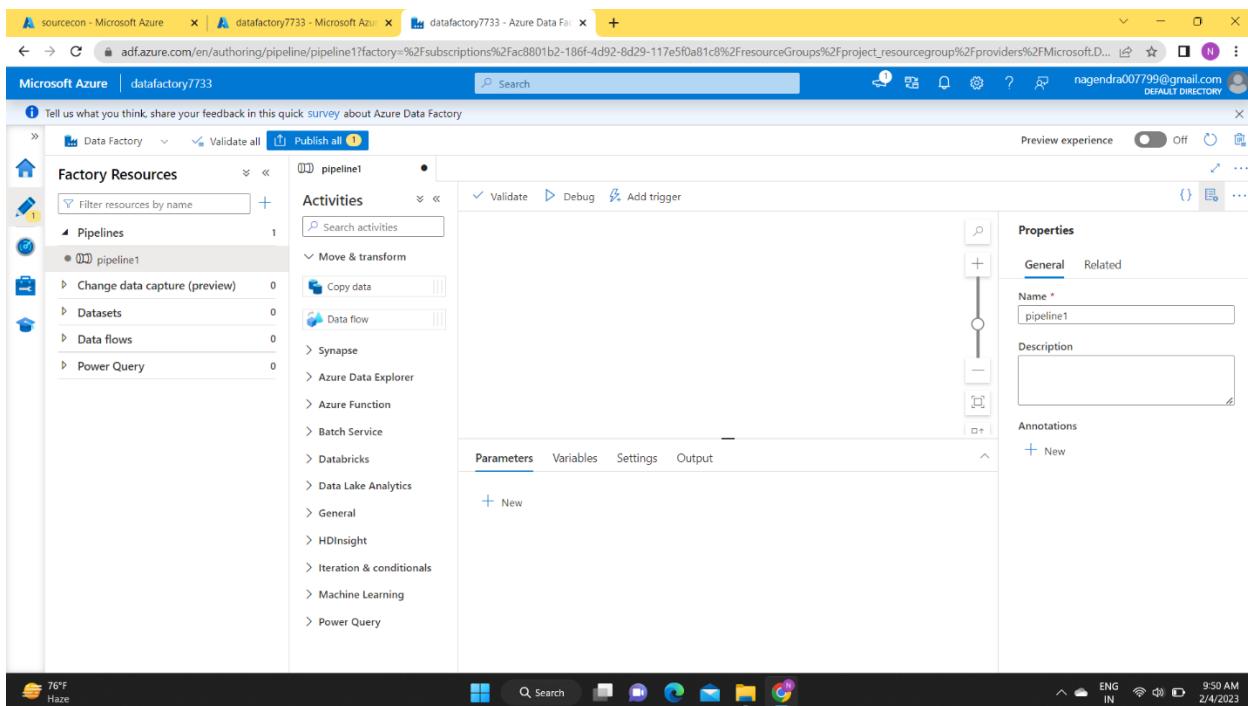
- Go to author tab



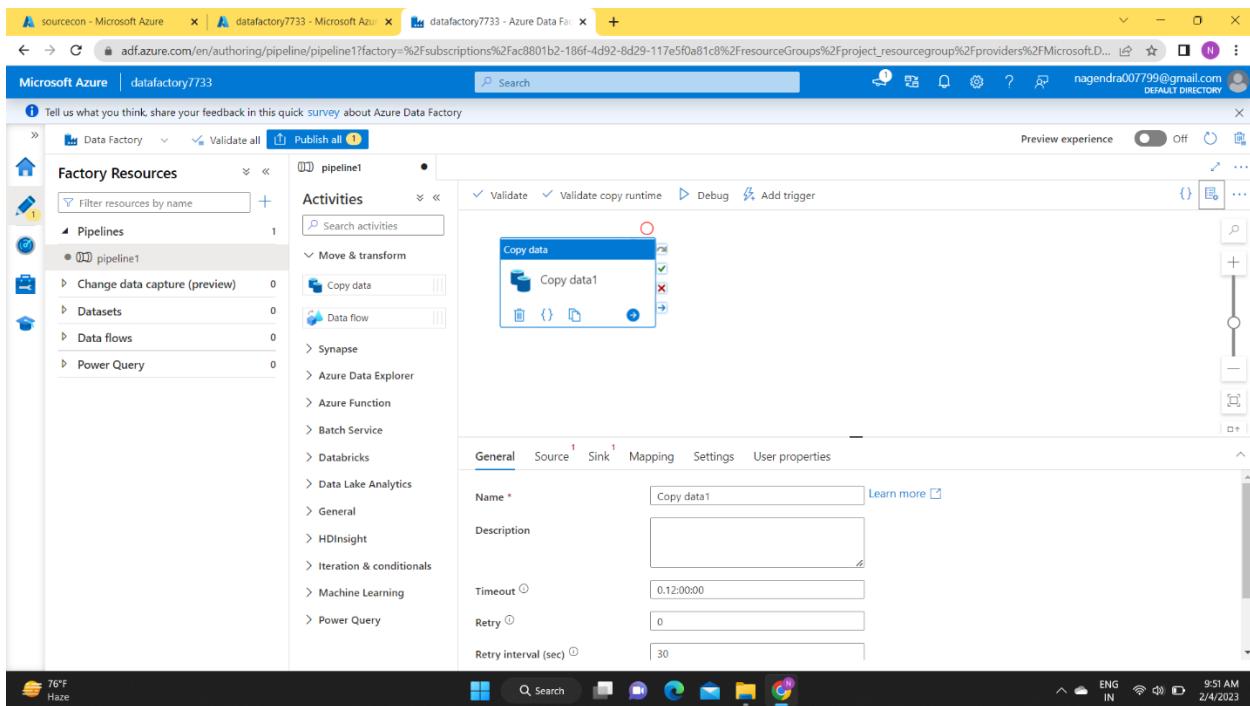
- In author go to pipeline and select new pipeline



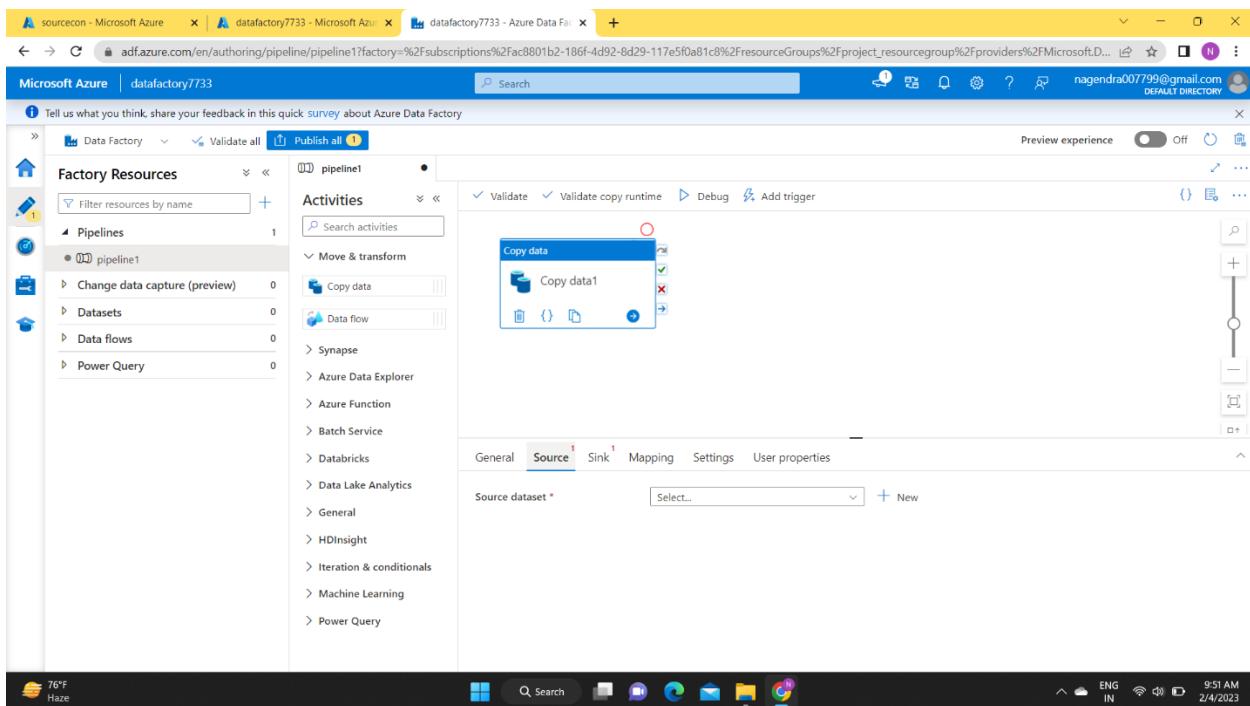
- Launch new pipeline
- Default pipeline name as “pipeline1”



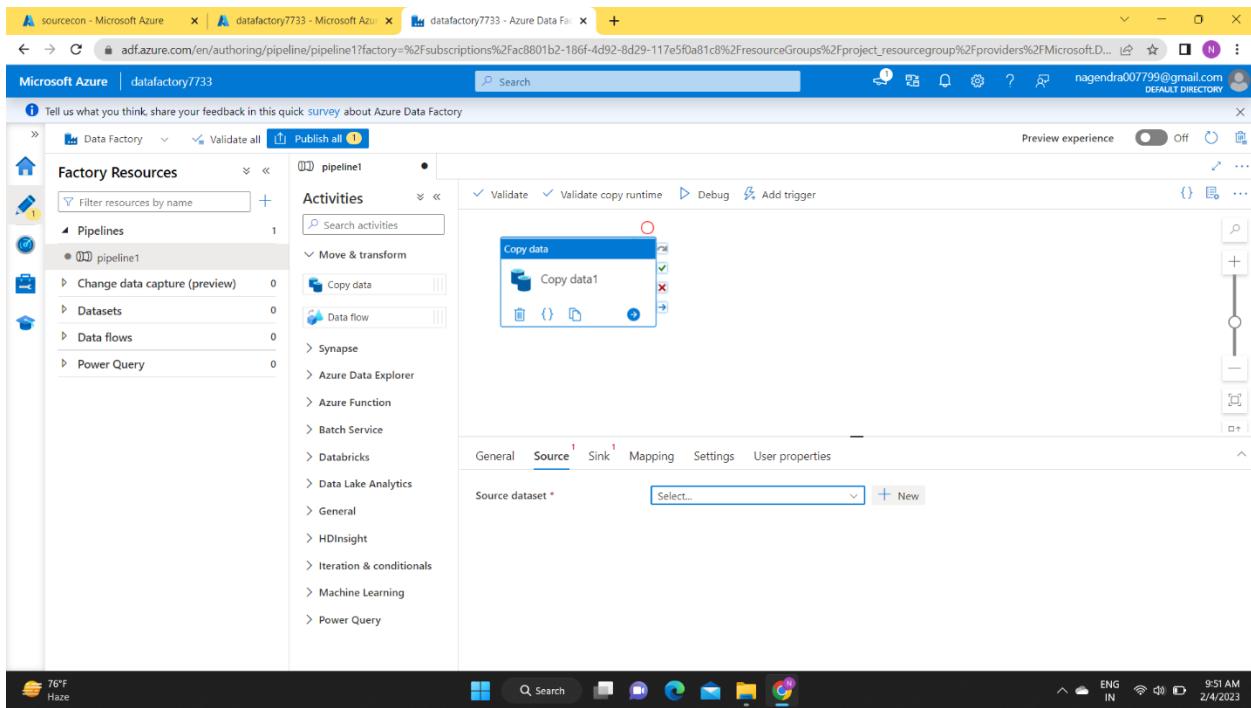
- In move and transform select copy data activity



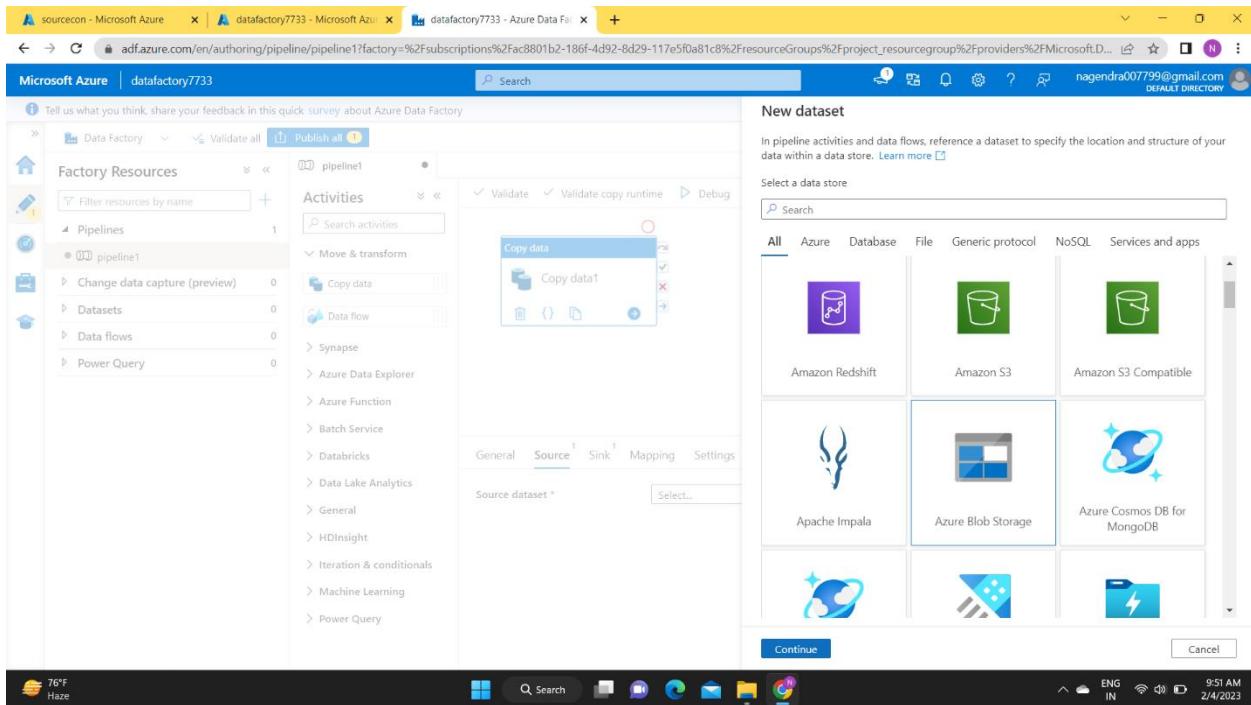
- Click on copy data and select source



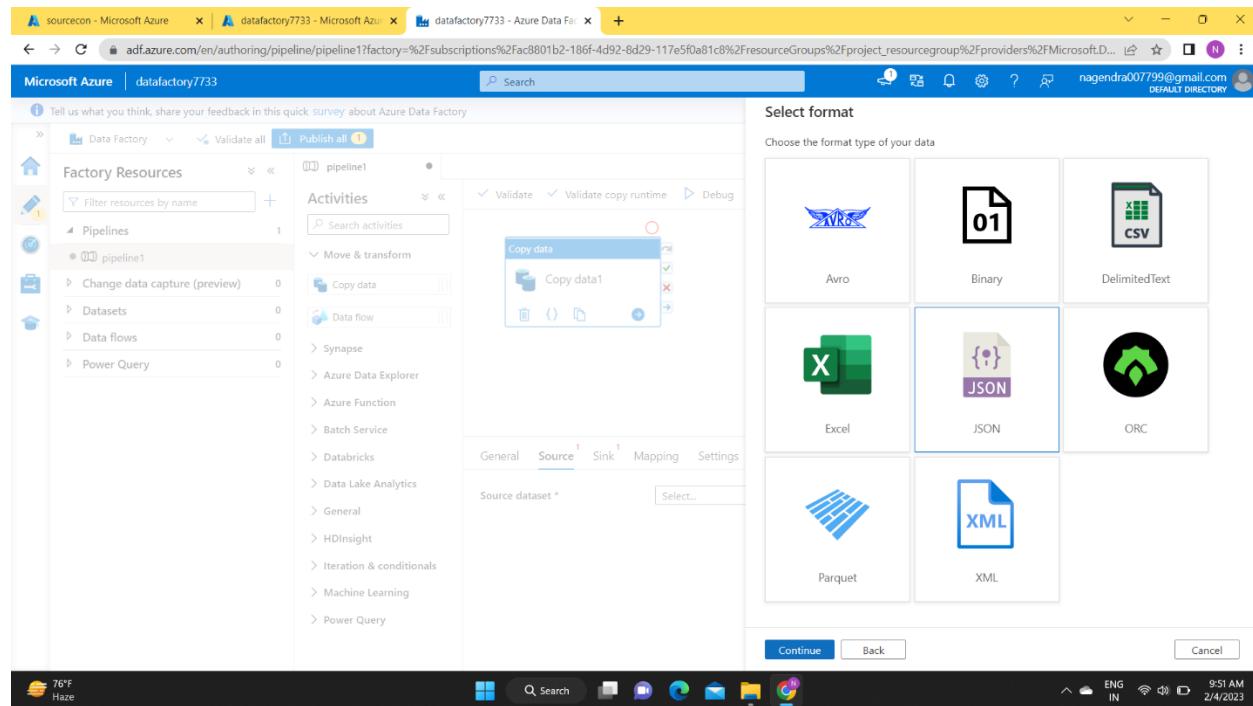
- In dataset box select +new



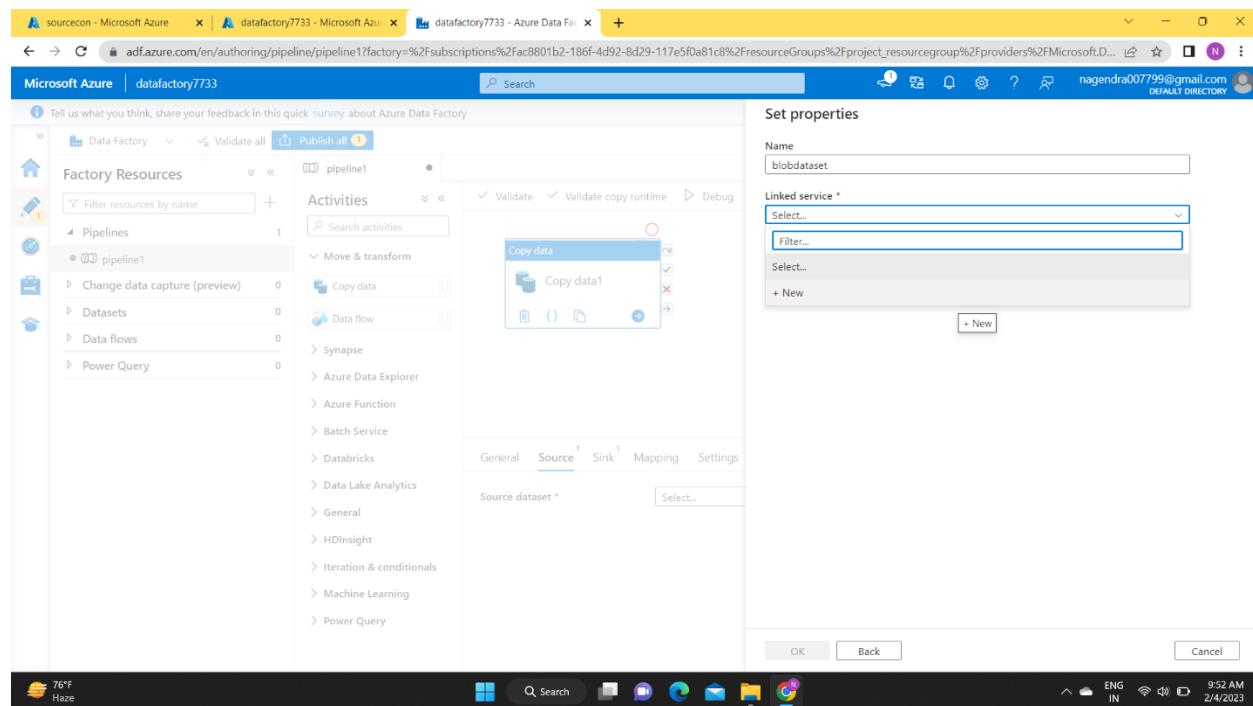
- Select the source storage type
- Here our source is blob storage and press continue.



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

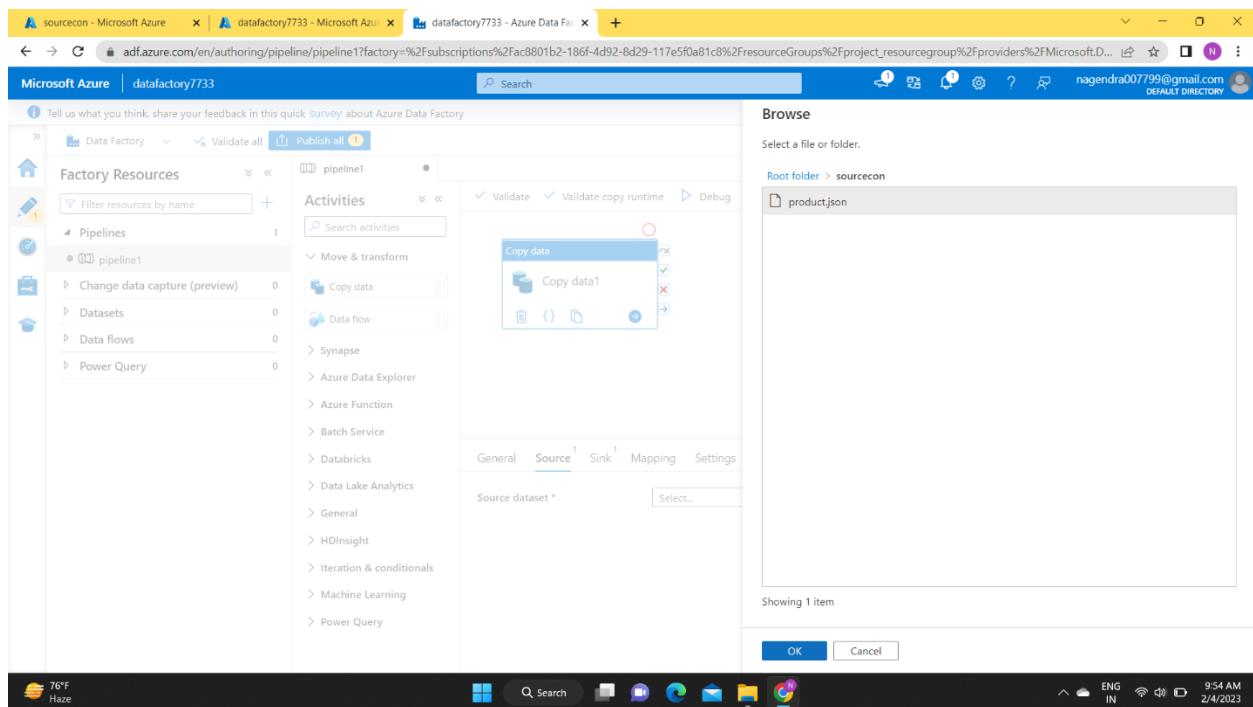
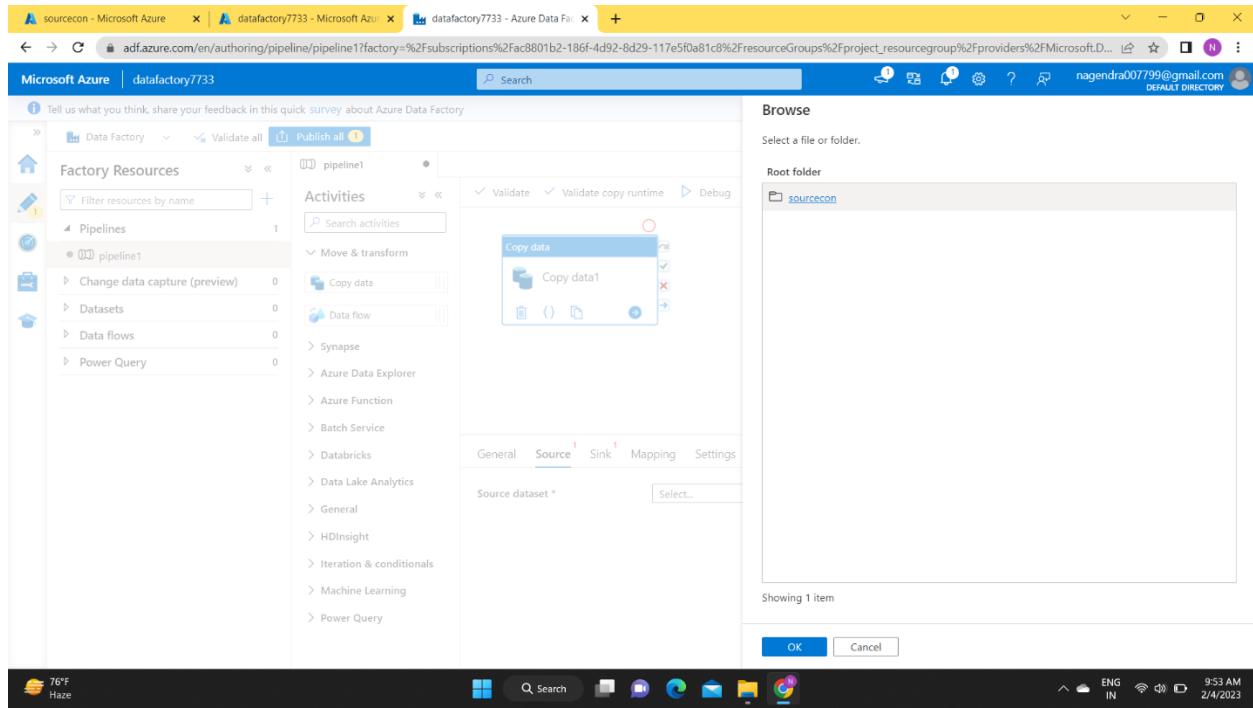


- In data set we have to give link service
- So we are creating new link service by clicking +new

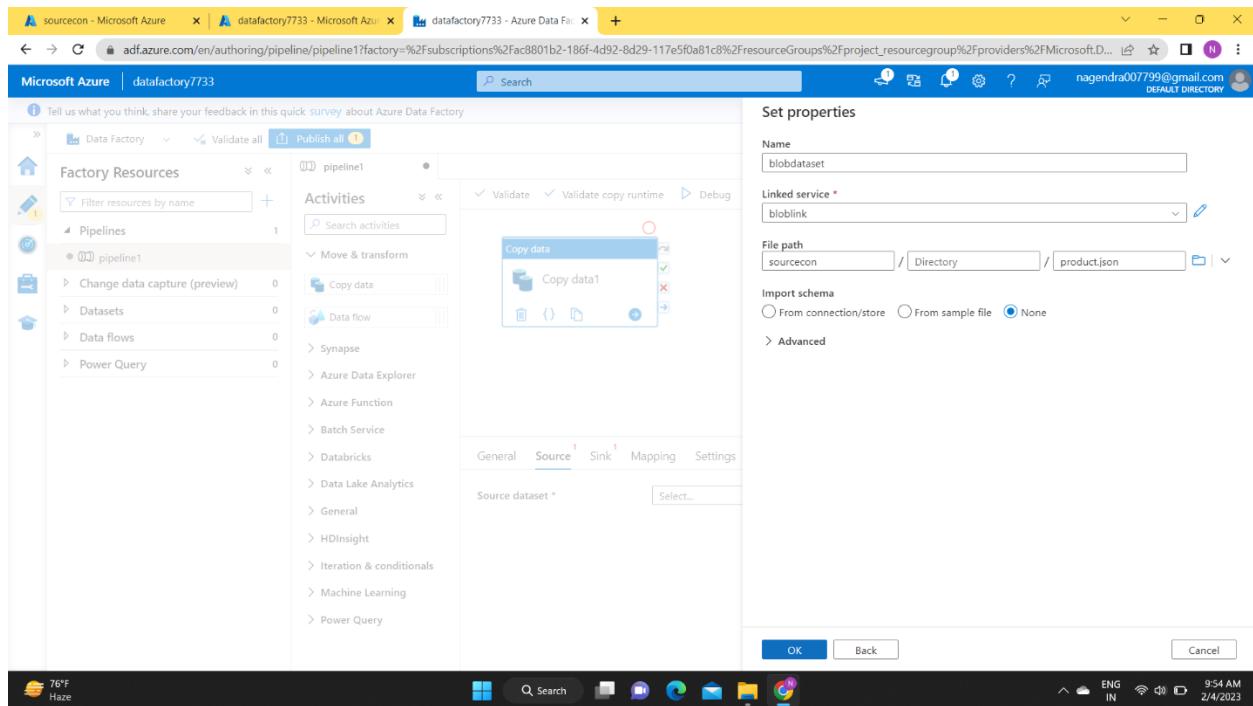


- Here you give name of your link service “bloblink”, and give azure subscription and storage account name.
- Must do the test connection.

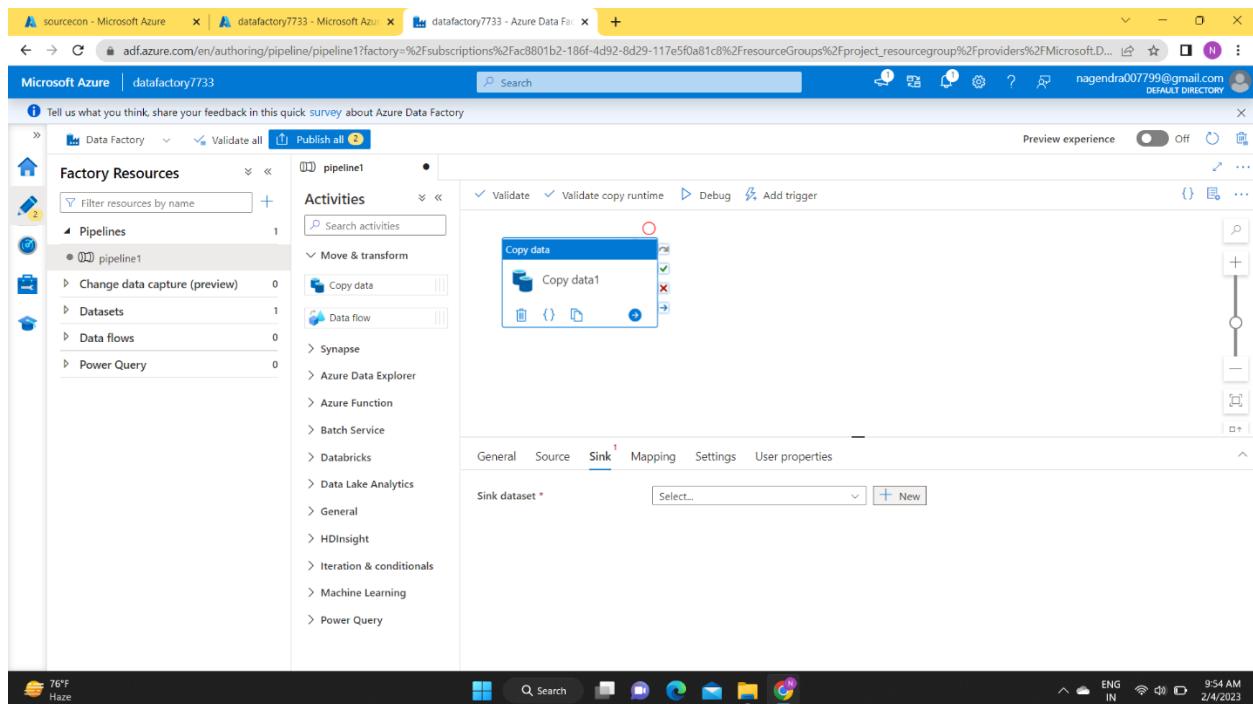
- After that select the source file location.



- 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)



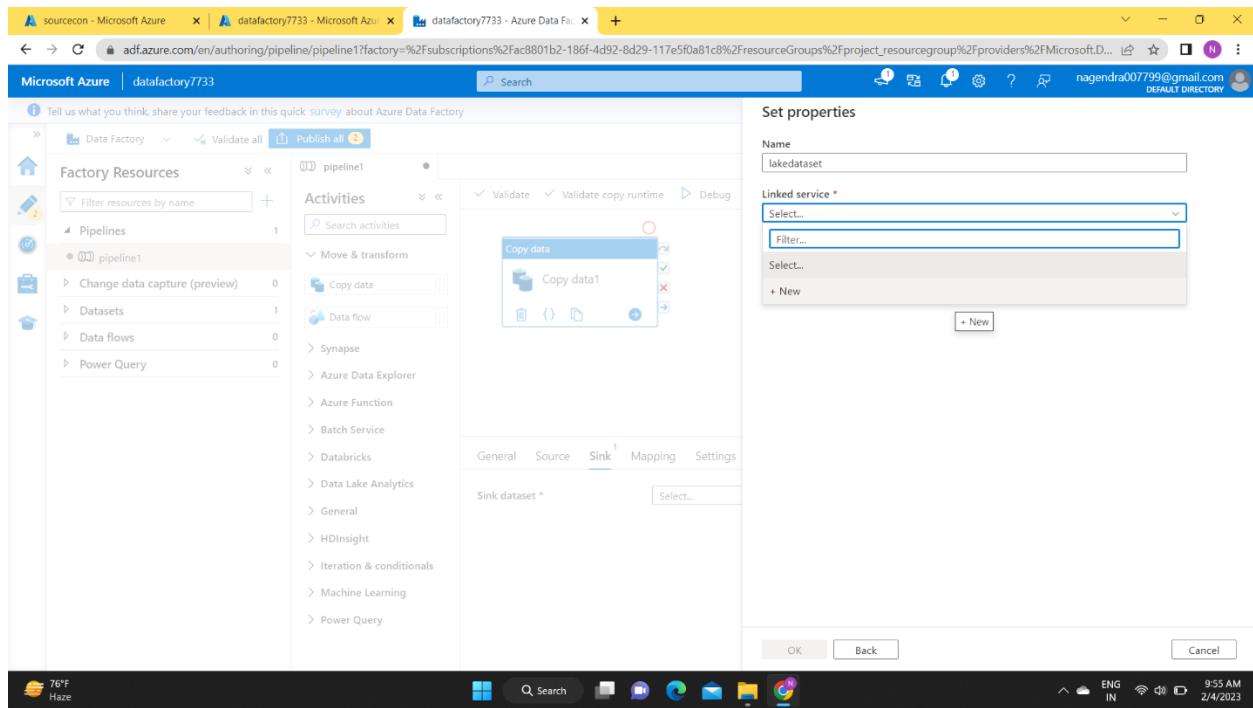
- Our target is datalake, so select datalake.

The screenshot shows the Azure Data Factory pipeline editor. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (1), 'Datasets' (1), 'Data flows' (0), and 'Power Query' (0). The main area shows a 'pipeline1' pipeline with a 'Copy data' activity selected. The 'Sink' tab is active, and the 'Select...' button is highlighted. A 'New dataset' dialog box is open on the right, showing a grid of data store options. The 'Azure Data Lake Storage Gen1' option is selected. The status bar at the bottom shows '76°F Haze' and the date '2/4/2023'.

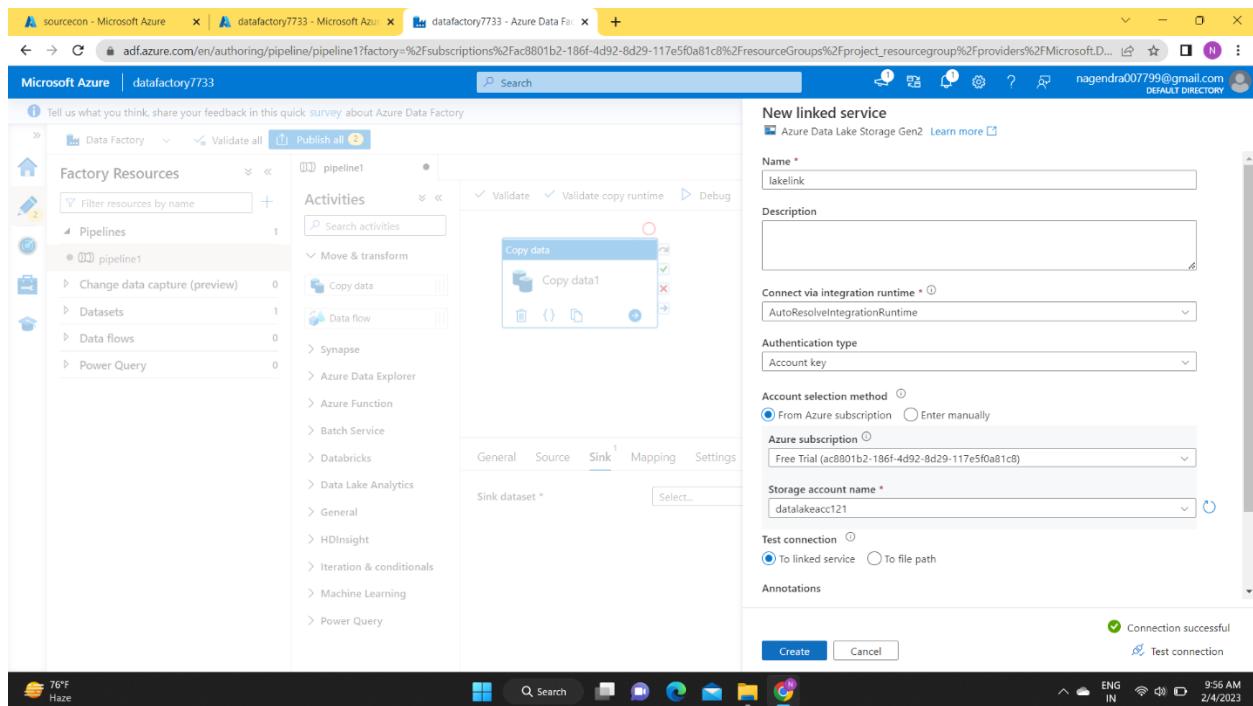
- Our final file format is parquet, so select parquet.

The screenshot shows the Azure Data Factory pipeline editor. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (1), 'Datasets' (1), 'Data flows' (0), and 'Power Query' (0). The main area shows a 'pipeline1' pipeline with a 'Copy data' activity selected. The 'Sink' tab is active, and the 'Select...' button is highlighted. A 'Select format' dialog box is open on the right, showing a grid of file format options. The 'Parquet' option is selected. The status bar at the bottom shows '76°F Haze' and the date '2/4/2023'.

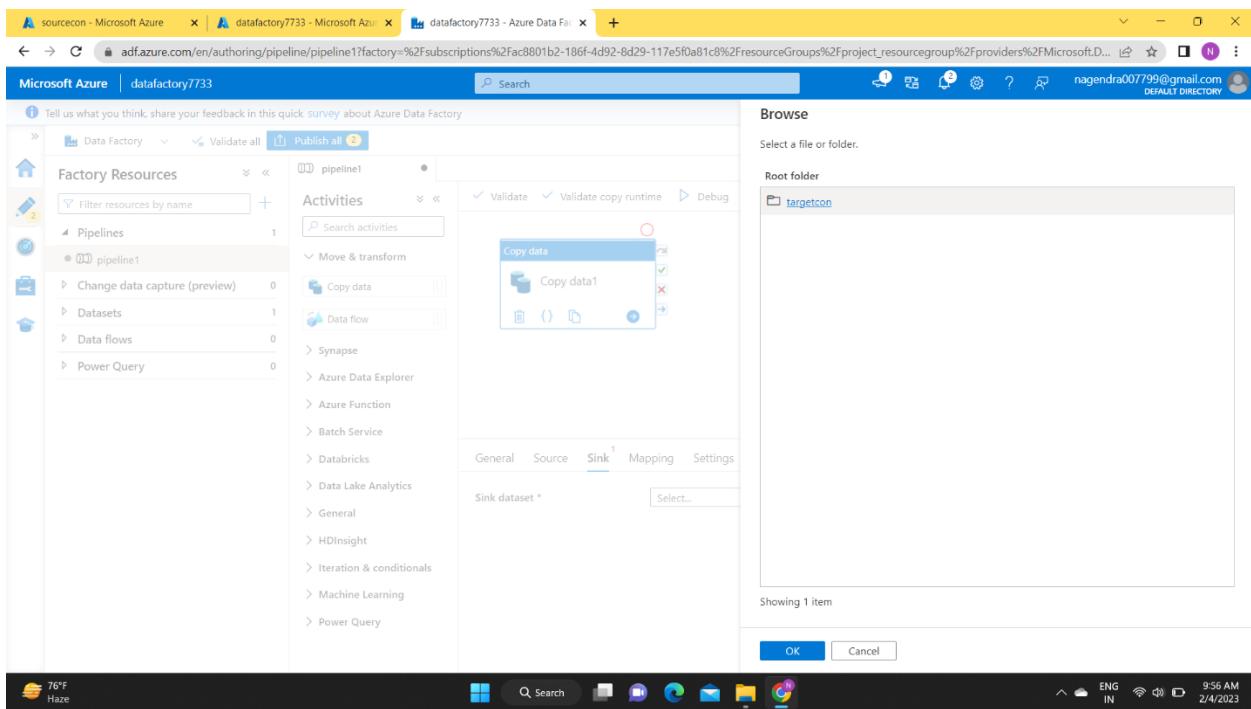
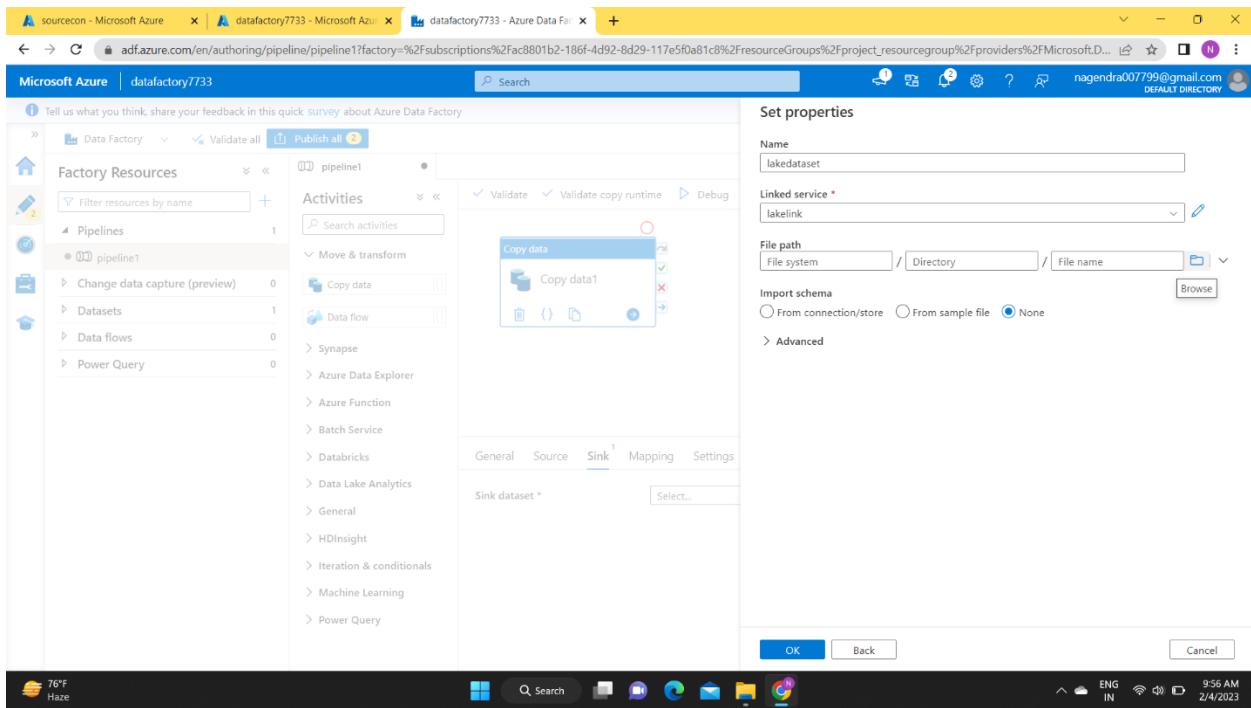
- Our dataset name is “lakedataset” and create a new link service for our target datalake.



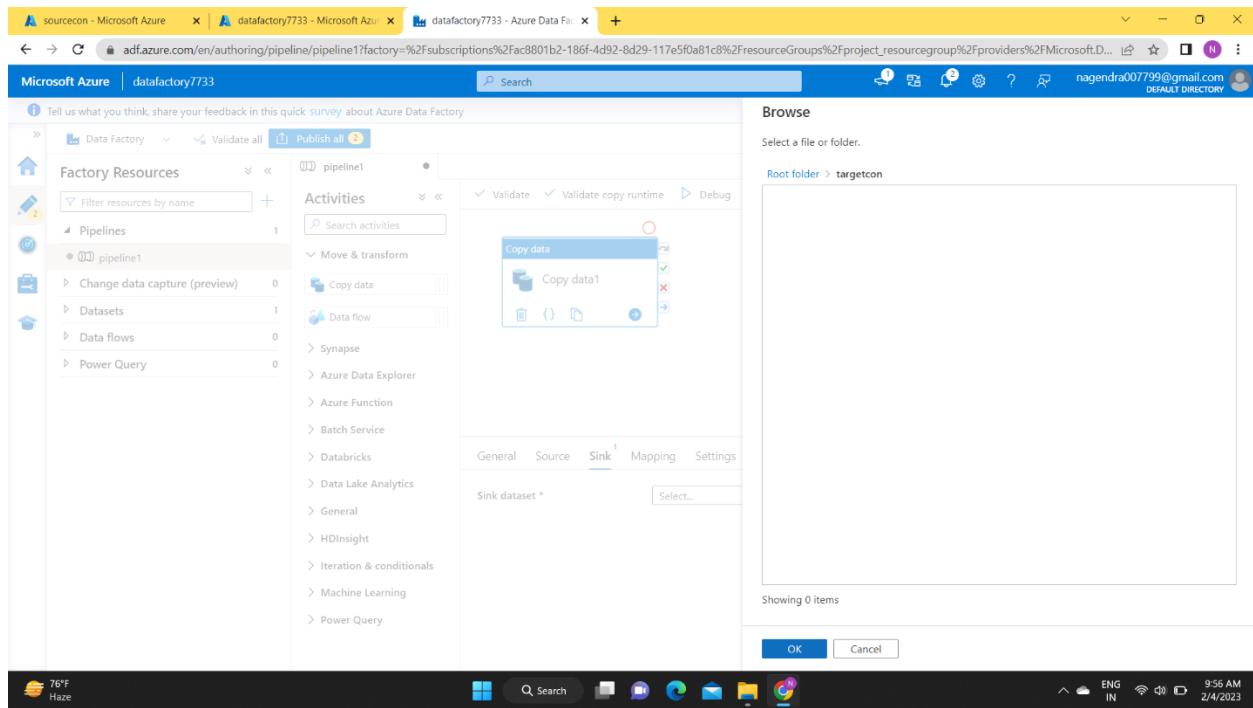
- Our target(data lake) link service name is “lakelink”, also select the azure subscription and storage account of our target.



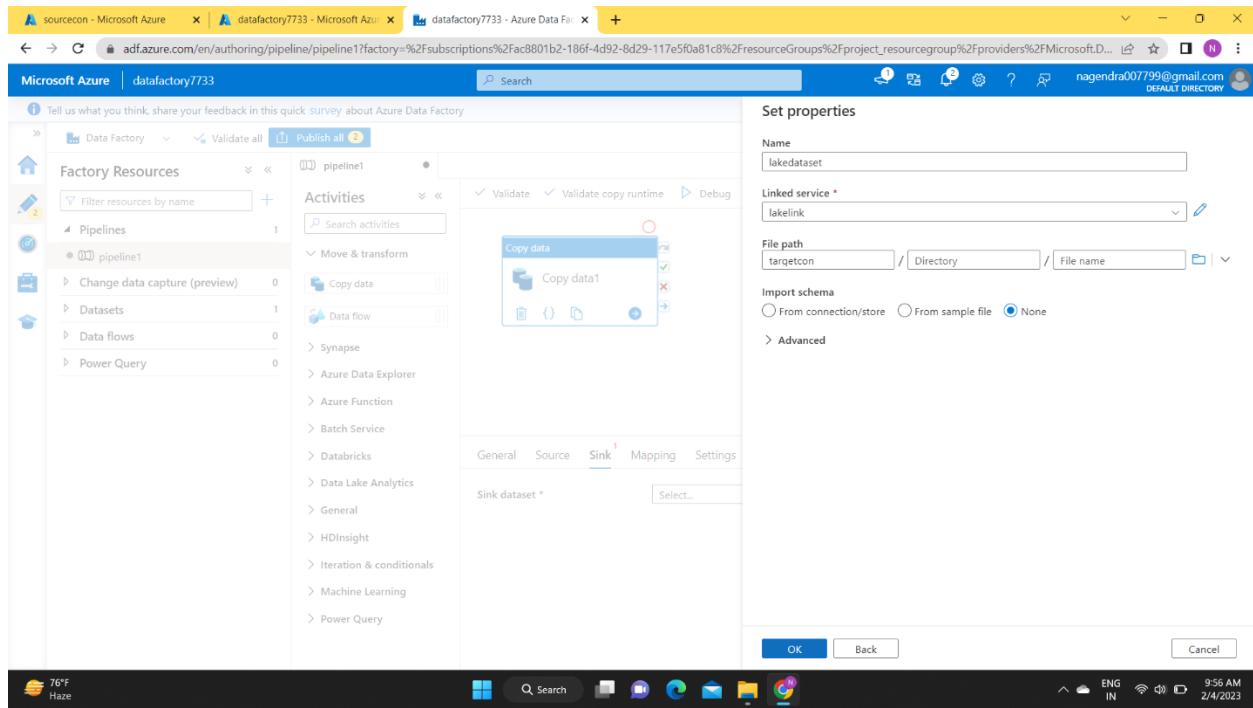
- We have to give target “file path” where we want to store our target file.



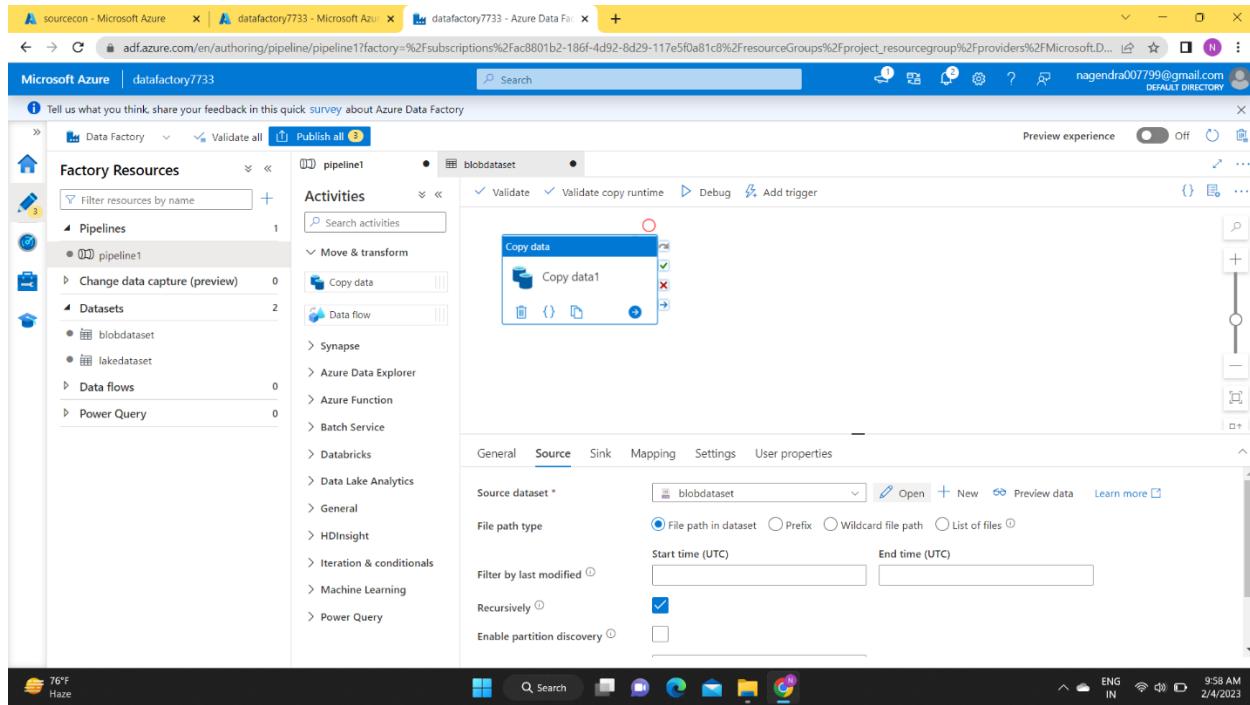
- Here no files in our target it's empty.



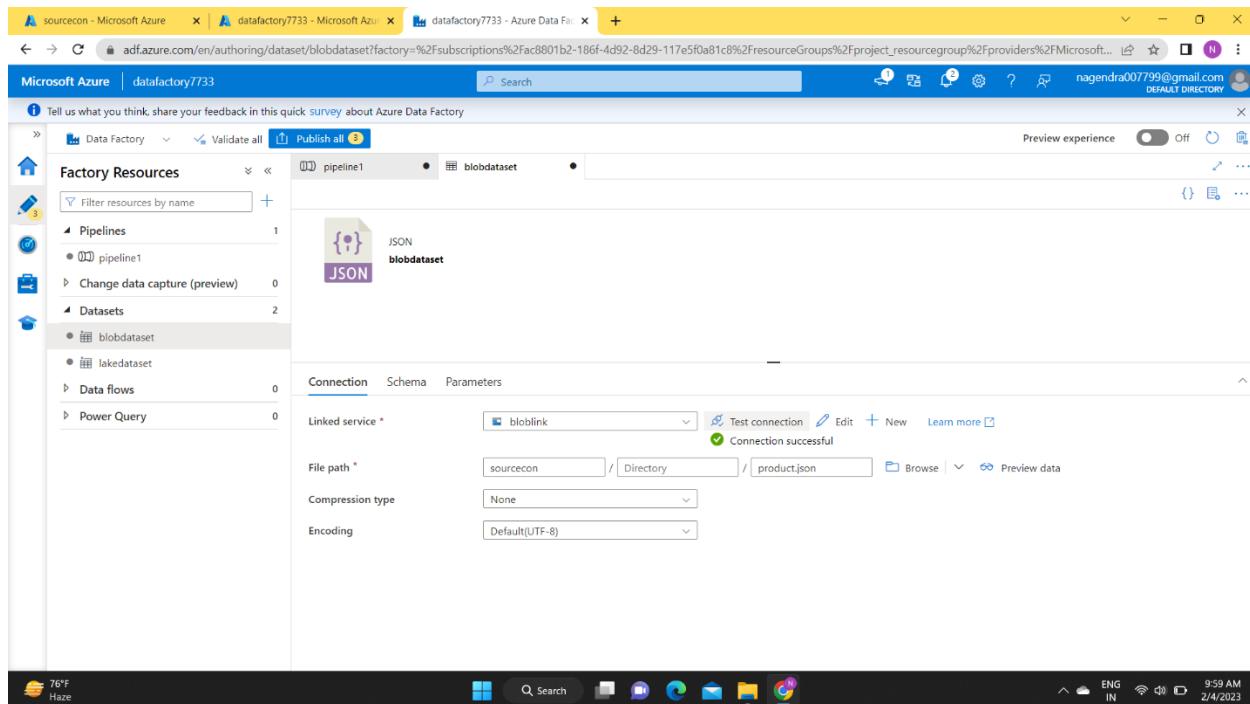
- We are not importing schema so we put none.



- 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

Factory Resources

JSON blobdataset

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

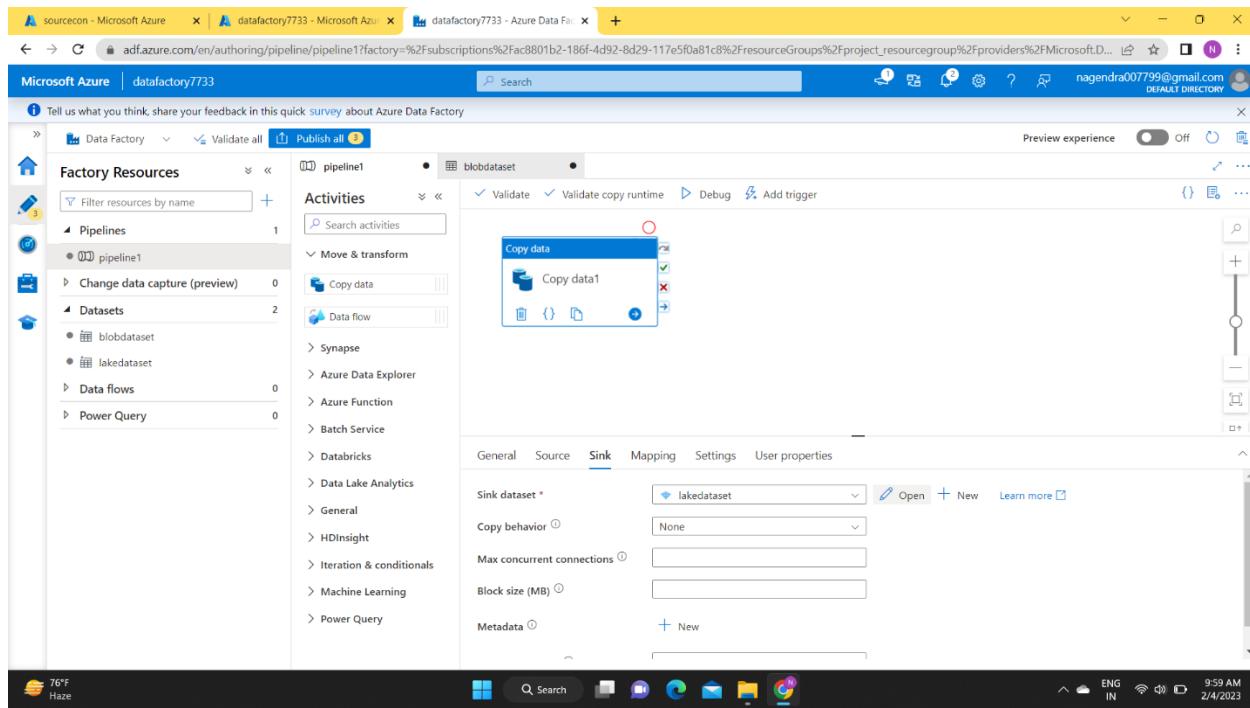
Object: product.json

```
[
  {
    "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": "Charging",
    "Product_Price": "25"
  }
]
```

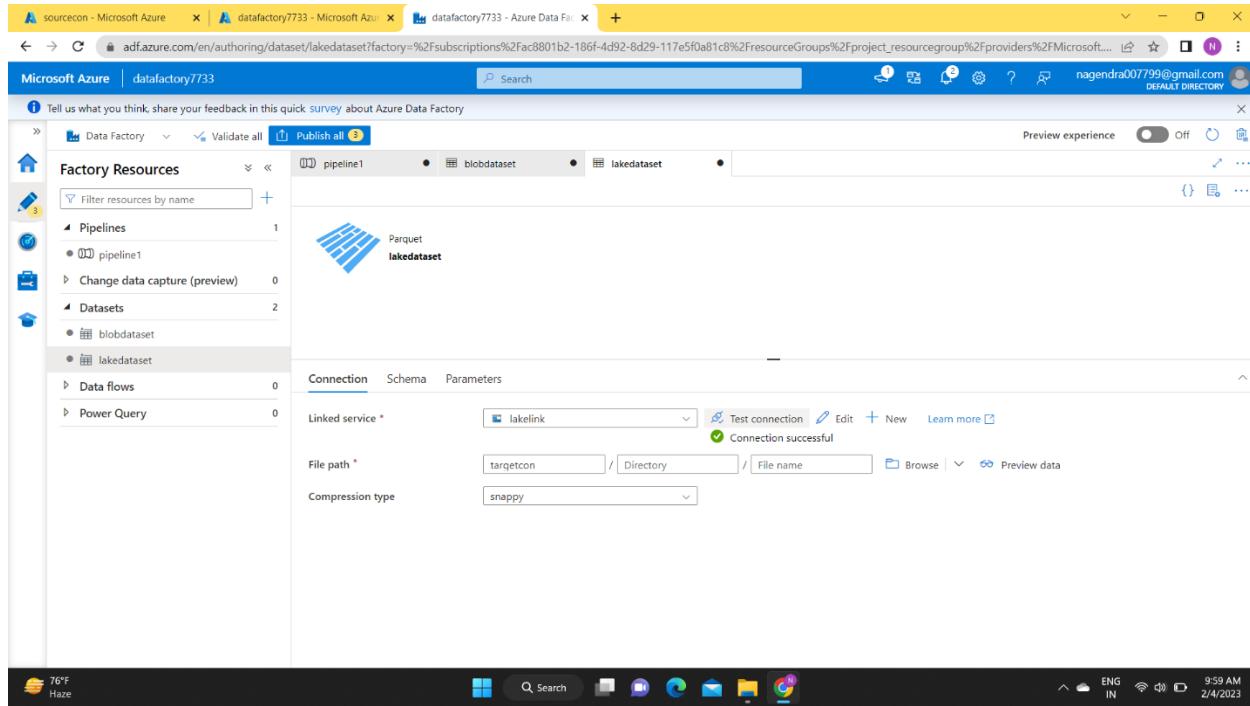
more

Preview data

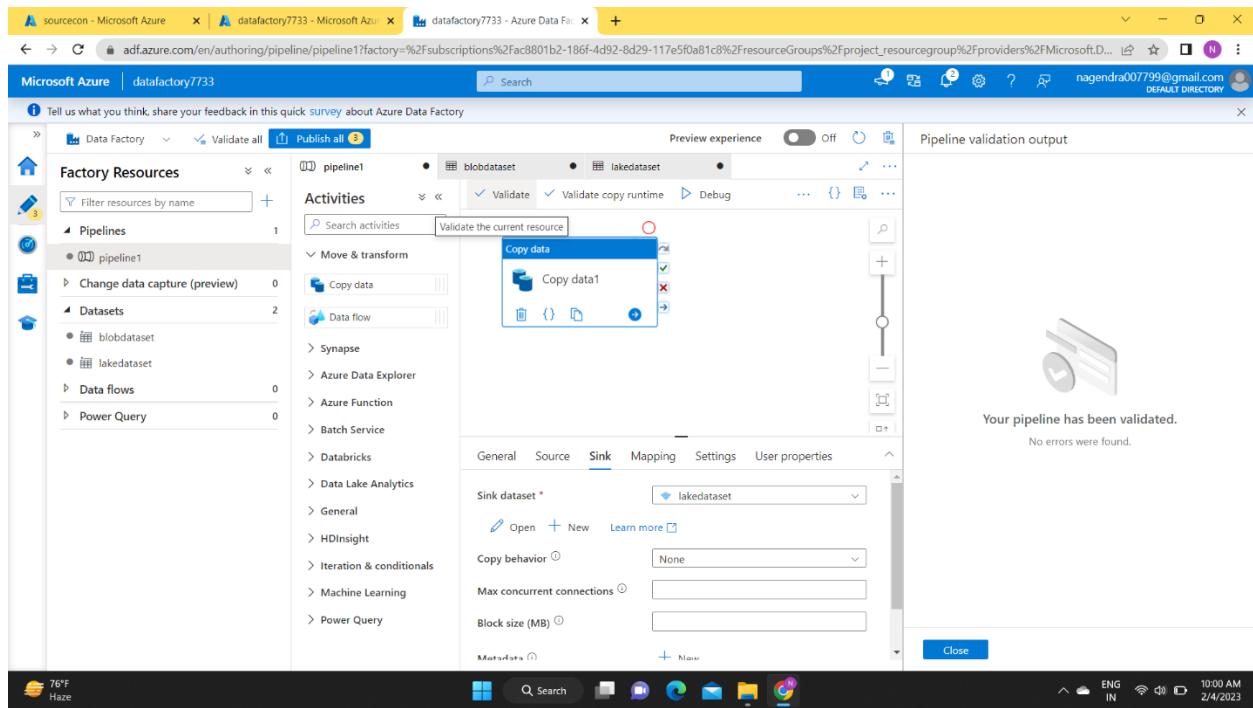
- Now go to sink there is “open” option.



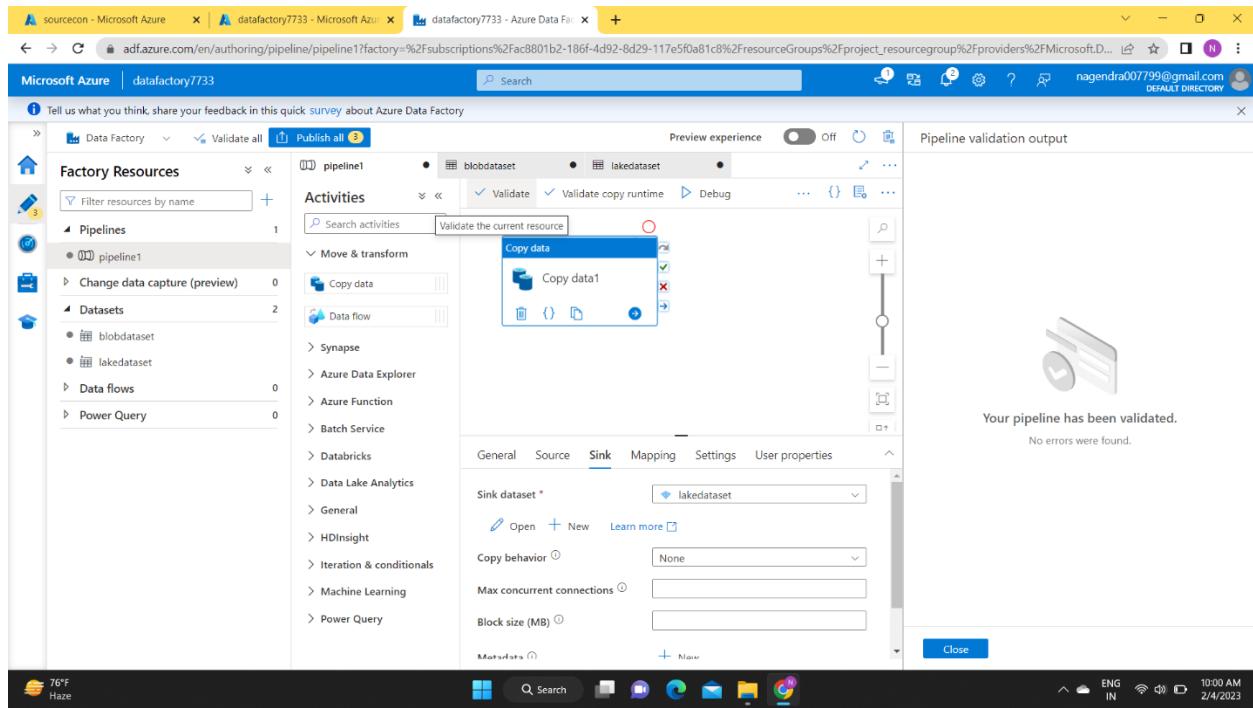
- check the test connection



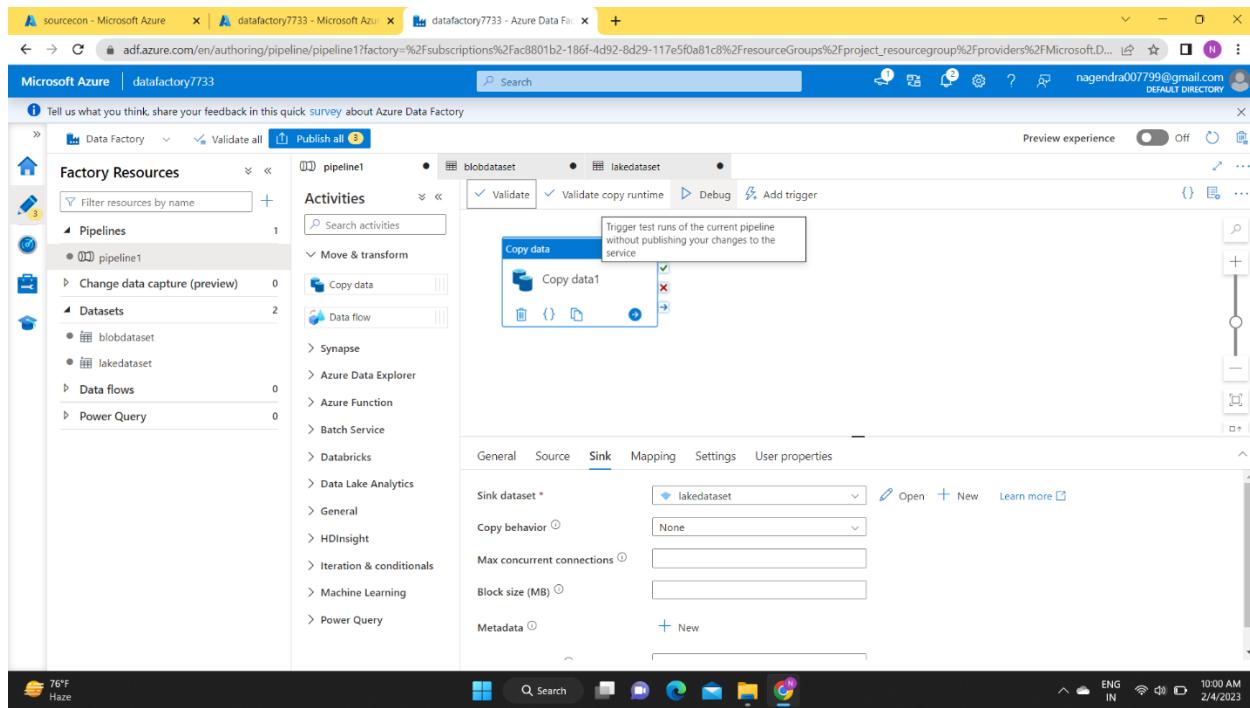
- now validate your pipeline it will show errors on right sidebar



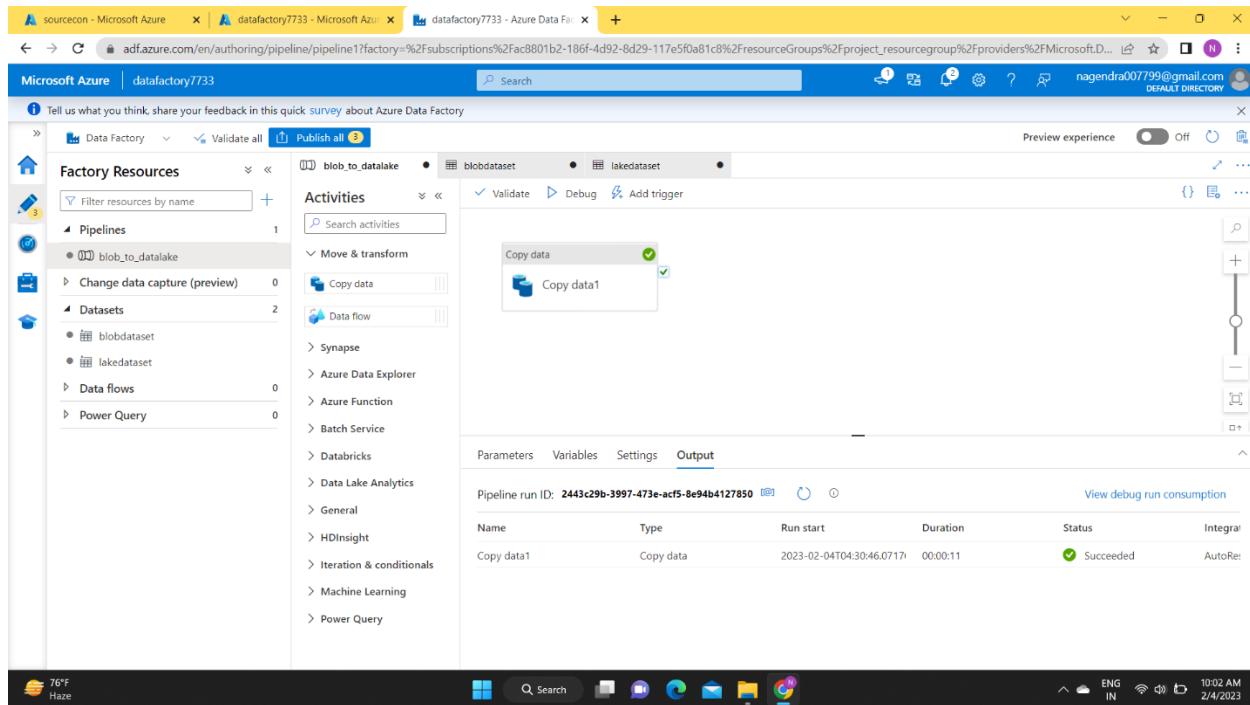
- our validation complete we don't have any errors.



- 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

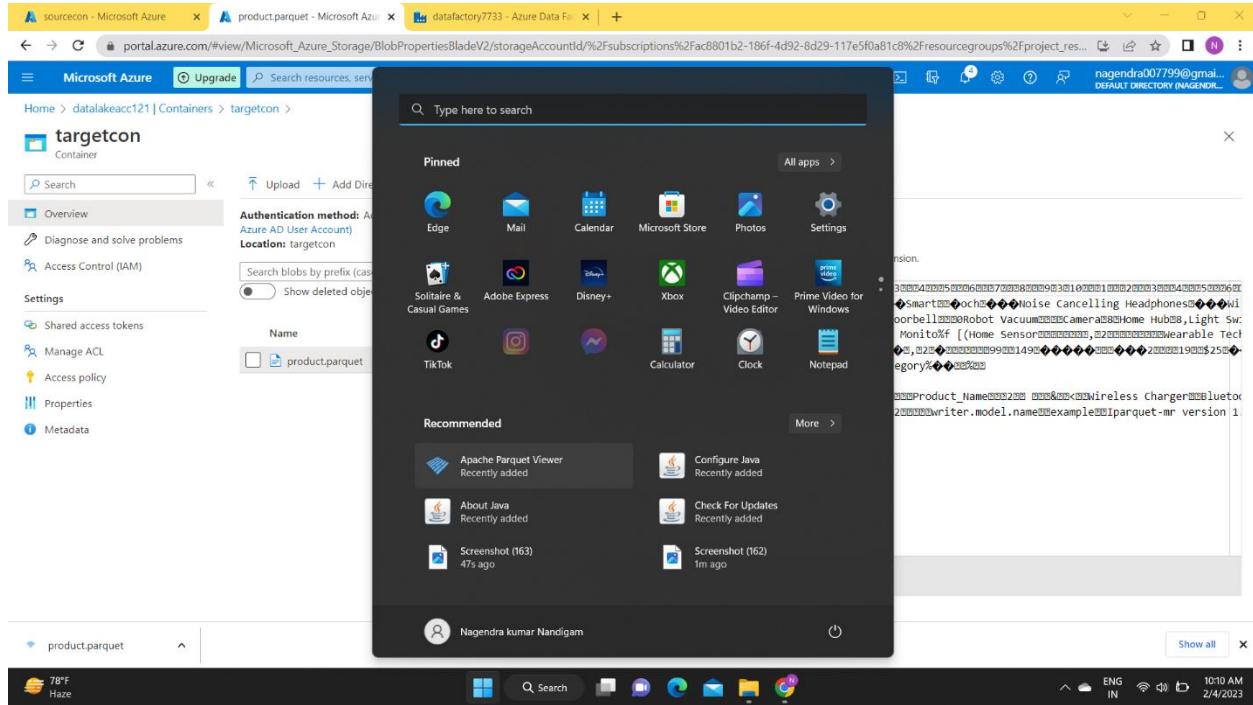
- Here is our parquet file in datalake

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

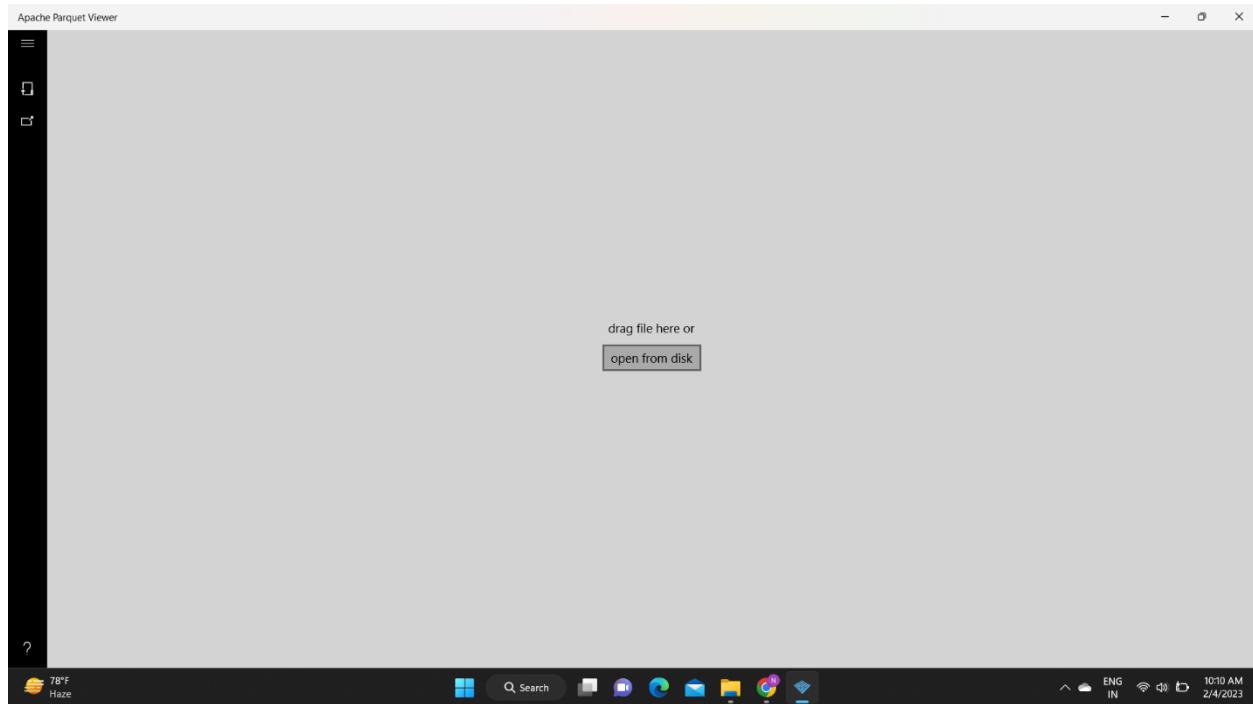
- Click on “edit” to see data present in file

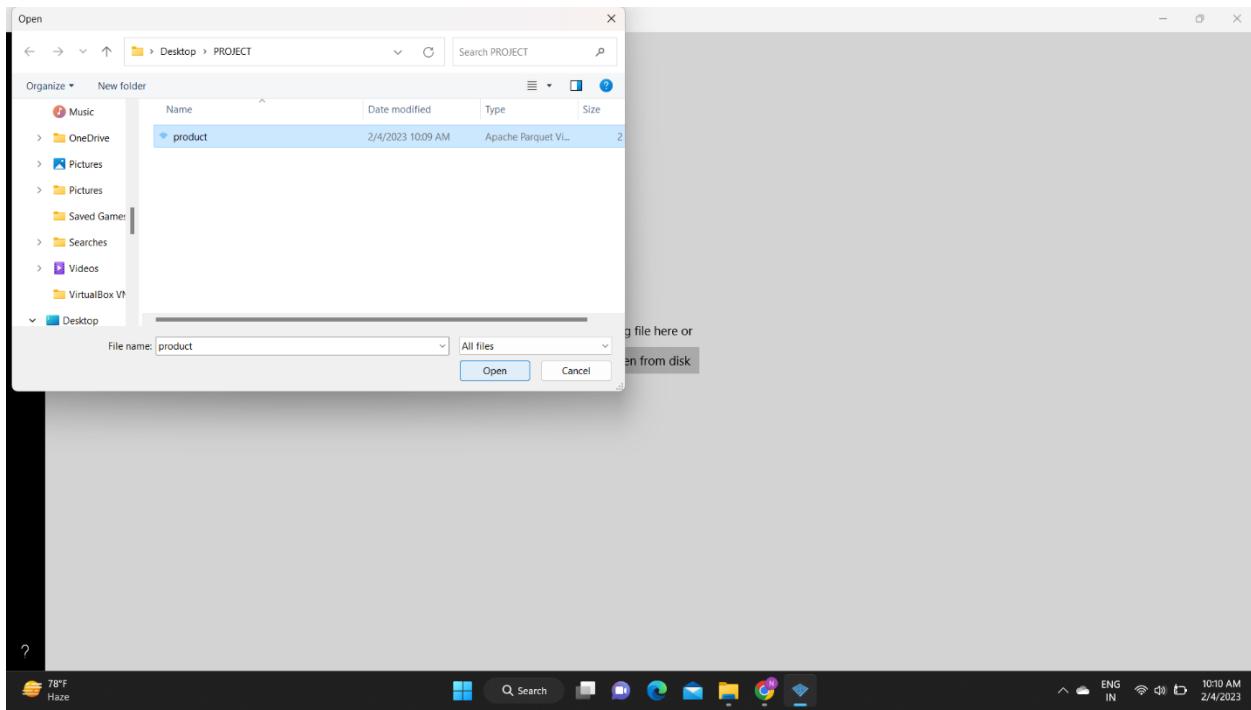
- We cannot understand the file so we have to download file.

- 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 data1

Parameters Variables Settings 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	0:00:11	Succeeded	AutoRe

Microsoft Azure | datafactory7733

Add triggers

Choose trigger...

Search

+ New

Close

- Here I am using “schedule trigger”

New trigger

Type: Schedule

Start date: 2/4/2023, 9:00:00 AM

Time zone: Coordinated Universal Time (UTC)

Recurrence: Every 1 Day(s)

Execute at these times: Hours: 10, Minutes: 0

Specify an end date: End On: 2/9/2023, 4:33:47 AM

Pipeline run ID: 2443c29b-3997-473e-acf5-8e94b4127828

Copy data1

- Now I am attaching trigger successfully.

Preview experience: Off

Trigger (1)

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

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

- Now I am publishing the pipeline with trigger.

Publish all

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

Pending changes (4)

NAME	CHANGE	EXISTING
blob_to_datalake	(New)	-
blobdataset	(New)	-
lakedataset	(New)	-
trigger1	(New)	-

Pipeline run ID: 2443c29b-3997-473e-acf5-8e94b41278

Name Type

Copy data1 Copy data

Publish Cancel

- This is the pipeline we are done.

Details Refresh

Learn more on copy performance details from here.

Activity run id: 13bc3195-6912-4b53-8adc-c3729499e40a

Azure Blob Storage → Azure Data Lake Storage Gen2

Succeeded

Azure Blob Storage Region: East US      Azure Data Lake Storage Gen2 Region: Central US

Data read: 2.735 KB      Data written: 1.495 KB  
 Files read: 1      Files written: 1  
 Rows read: 25      Rows written: 25  
 Peak connections: 1      Peak connections: 1

Copy duration: 00:00:11      Throughput: 547 bytes/s

Azure Blob Storage → Azure Data Lake Storage Gen2

Start time: 2/4/2023, 8:56:24 AM      Used DIUs: 4  
 Used parallel copies: 1      Duration: 00:00:11  
 Details Working duration Total duration

How satisfied or dissatisfied are you with the performance of this copy activity?

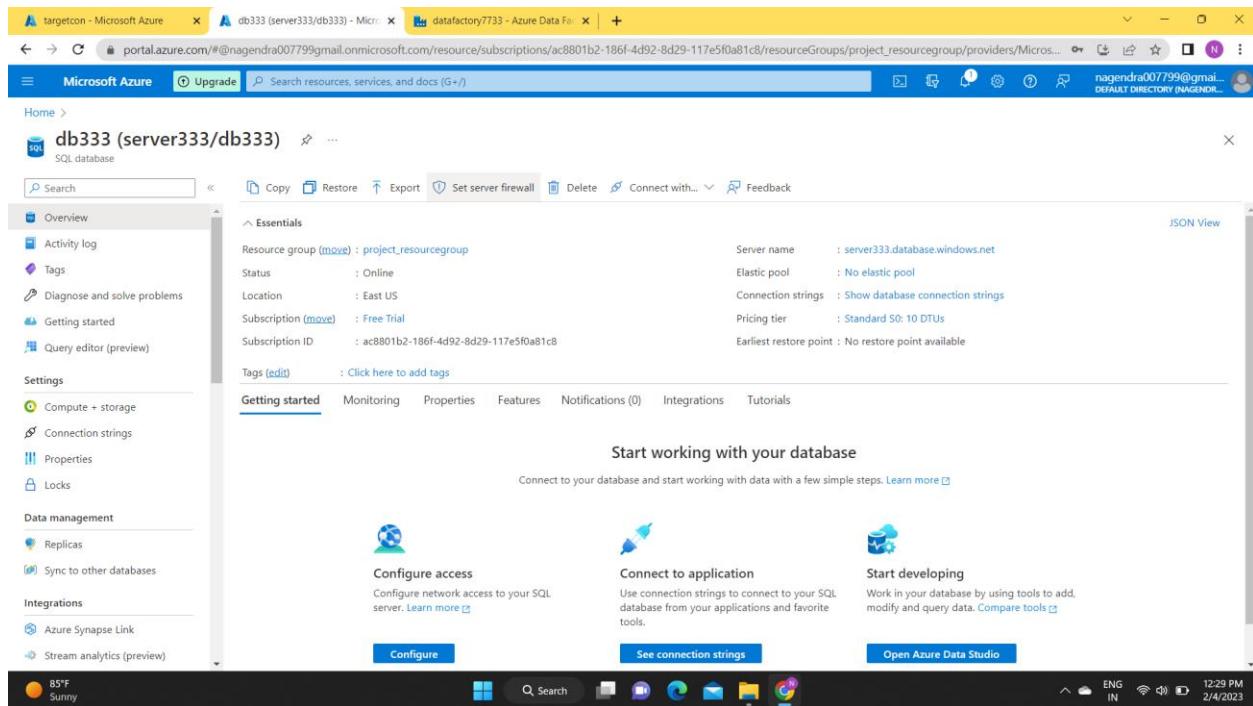
Sales.parquet

Show all

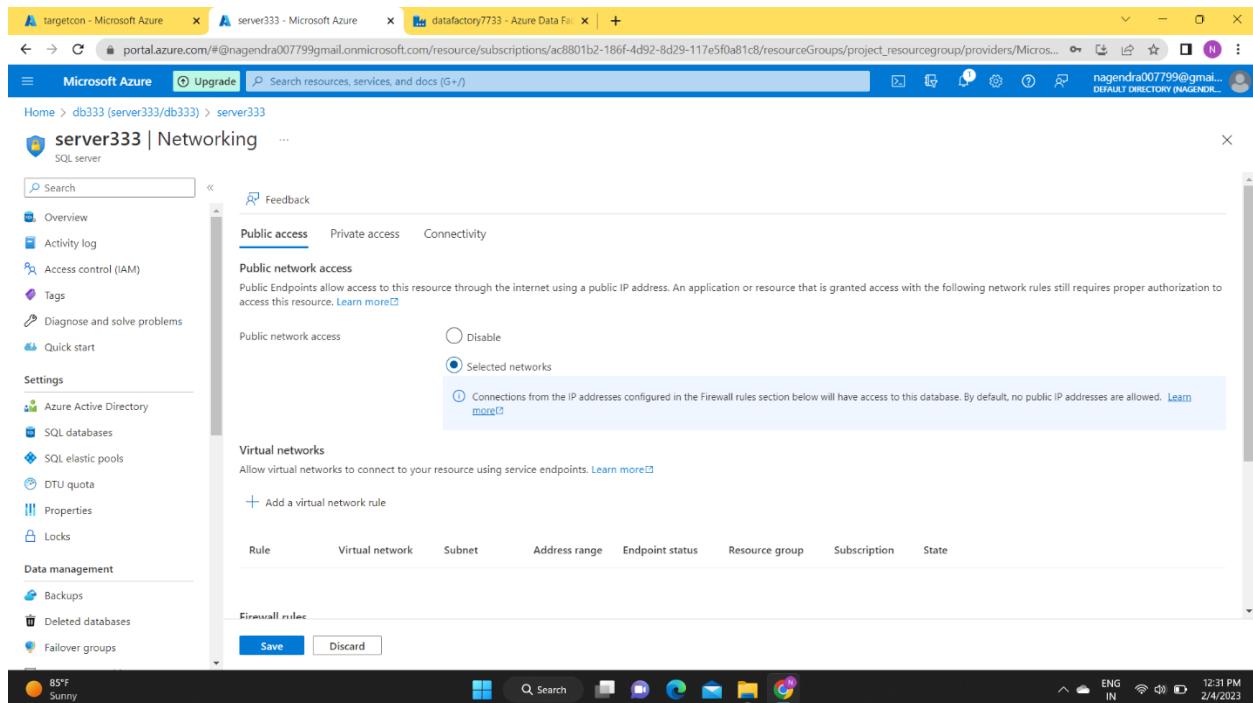
85°F Sunny

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

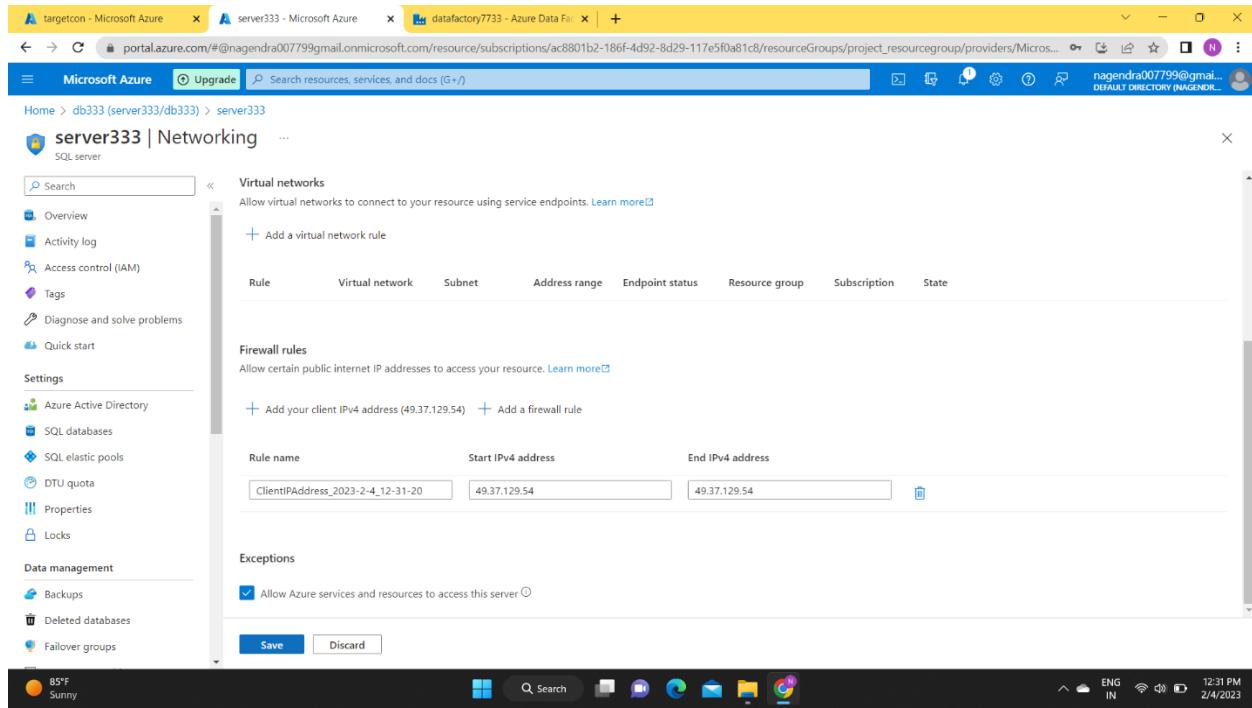
- This is our azure sql database “set server firewall” to give access to our azure resources to this sql db



The screenshot shows the Azure portal interface for a SQL database named 'db333 (server333/db333)'. The top navigation bar includes tabs for 'targetcon - Microsoft Azure', 'db333 (server333/db333) - Microsoft Azure', and 'datafactory7733 - Azure Data Factory'. The 'Set server firewall' button is located in the top right of the navigation bar. The main page displays the database's 'Essentials' section with details like Resource group (move), Status (Online), Location (East US), and Subscription (Free Trial). Below this, there are sections for 'Getting started', 'Start working with your database', and links to 'Configure access', 'Connect to application', and 'Start developing'. The left sidebar contains navigation links for Overview, Activity log, Tags, Diagnose and solve problems, Getting started, Query editor (preview), Settings, Compute + storage, Connection strings, Properties, Locks, Data management (Replicas, Sync to other databases), and Integrations (Azure Synapse Link, Stream analytics (preview)). The bottom of the screen shows the Windows taskbar with various pinned icons and the system tray.

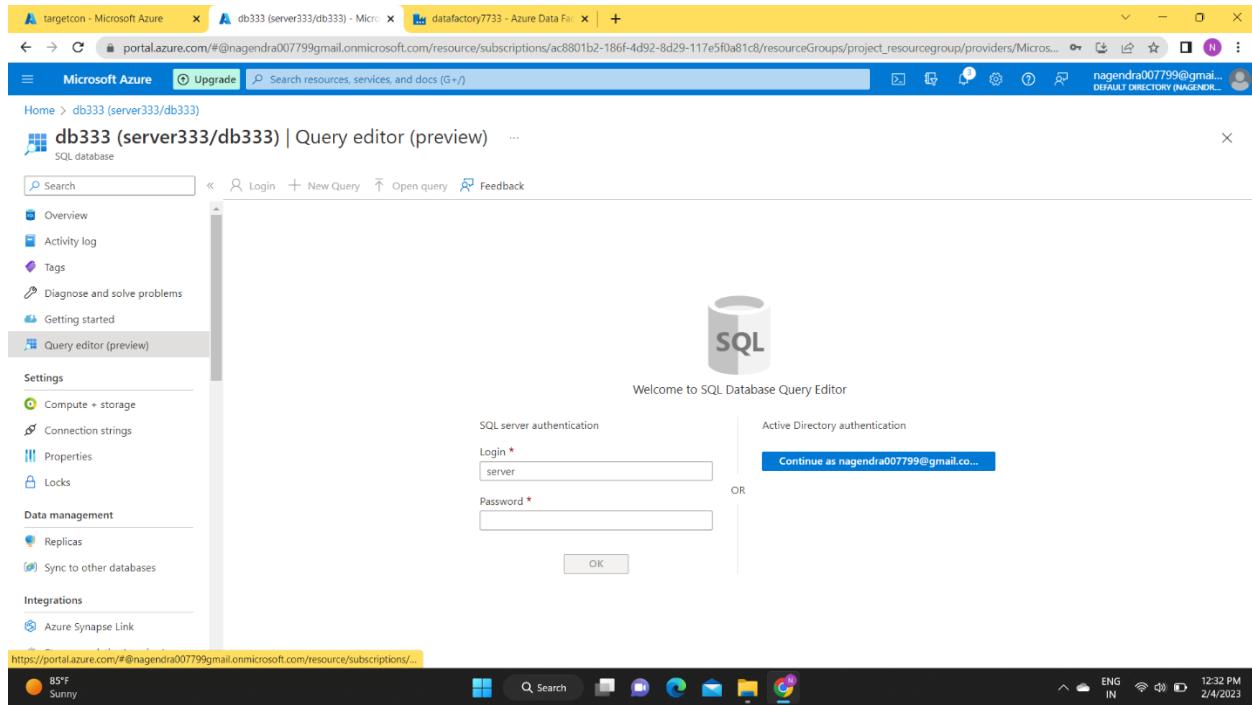


The screenshot shows the 'Networking' page for the 'server333' SQL server. The 'Public access' tab is selected, showing options for 'Disable' or 'Selected networks'. The 'Selected networks' option is chosen, with a note that connections from the IP addresses configured in the Firewall rules section will have access to the database. Below this, there is a 'Virtual networks' section with a 'Add a virtual network rule' button. The left sidebar contains navigation links for Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Quick start, Settings (Azure Active Directory, SQL databases, SQL elastic pools, DTU quota, Properties, Locks), and Data management (Backups, Deleted databases, Failover groups). The bottom of the screen shows the Windows taskbar with various pinned icons and the system tray.



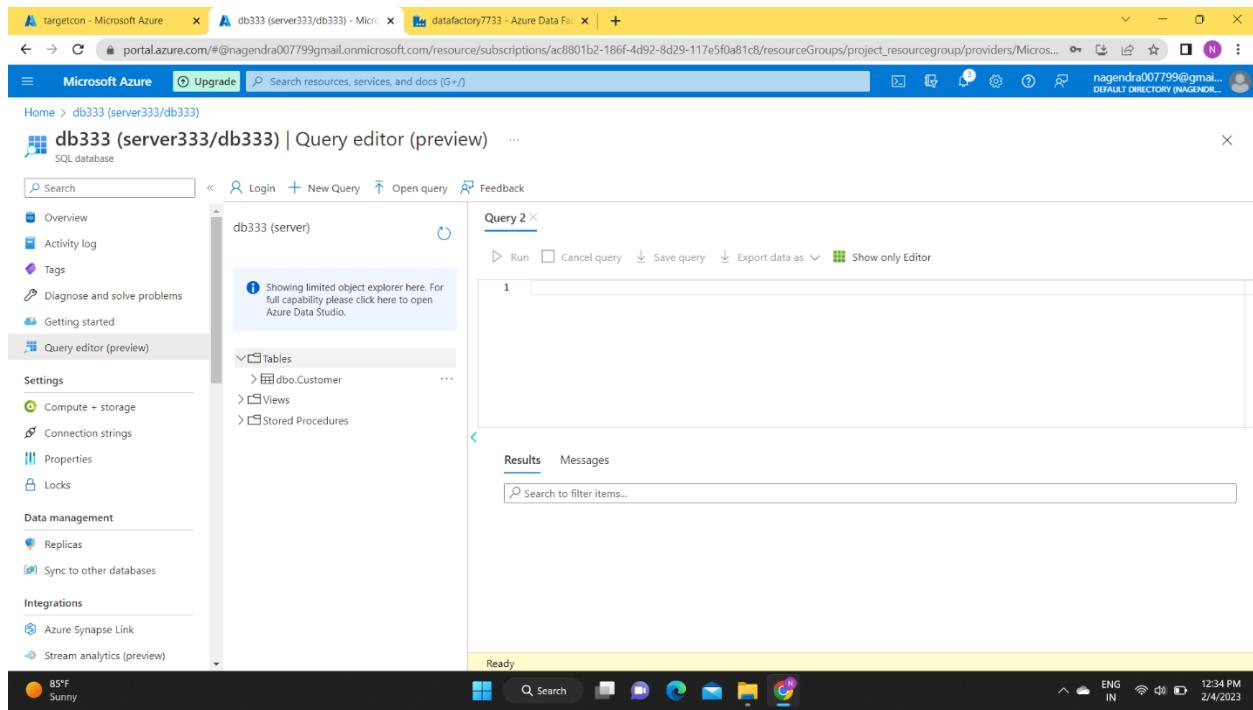
The screenshot shows the Azure portal interface for a SQL server named 'server333'. The left sidebar contains navigation links for Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Quick start, Settings, Azure Active Directory, SQL databases, SQL elastic pools, DTU quota, Properties, Locks, Data management, Backups, Deleted databases, and Failover groups. The main content area is titled 'server333 | Networking' and shows the 'Virtual networks' and 'Firewall rules' sections. Under 'Virtual networks', there is a table with columns: Rule, Virtual network, Subnet, Address range, Endpoint status, Resource group, Subscription, and State. Under 'Firewall rules', there is a table with columns: Rule name, Start IPv4 address, and End IPv4 address. A row is selected with the rule name 'ClientIPAddress\_2023-2-4\_12-31-20', start address '49.37.129.54', and end address '49.37.129.54'. The 'Exceptions' section has a checked checkbox for 'Allow Azure services and resources to access this server'. At the bottom are 'Save' and 'Discard' buttons. The status bar at the bottom right shows the date and time as 2/4/2023, 12:31 PM, and the weather as 85°F, Sunny.

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



The screenshot shows the Azure portal interface for a database named 'db333'. The left sidebar contains navigation links for Overview, Activity log, Tags, Diagnose and solve problems, Getting started, and Query editor (preview). The main content area is titled 'db333 (server333/db333) | Query editor (preview)' and shows the 'Welcome to SQL Database Query Editor' screen. It features two authentication options: 'SQL server authentication' (Login: 'server', Password: ' ') and 'Active Directory authentication' (button: 'Continue as nagendra007799@gmail.com...'). Below these options is an 'OK' button. The status bar at the bottom right shows the date and time as 2/4/2023, 12:32 PM, and the weather as 85°F, Sunny.

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

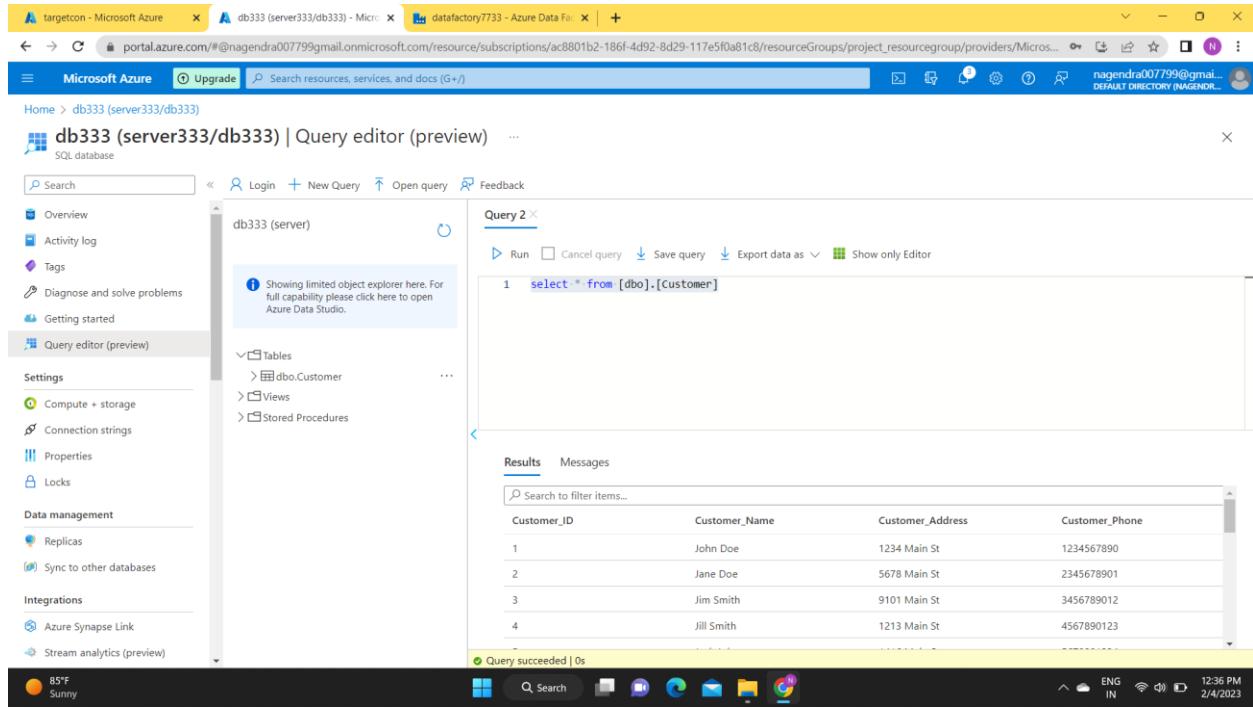


The screenshot shows the Azure Data Studio interface. The left sidebar is the Object Explorer, showing 'db333 (server)' with 'Tables' expanded, showing 'dbo.Customer'. The main area is the 'Query 2' editor with the following content:

```
1
```

The results pane is empty, showing a search bar: 'Search to filter items...'.

- You can see the data in customer table.



The screenshot shows the Azure Data Studio interface with a query executed:

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

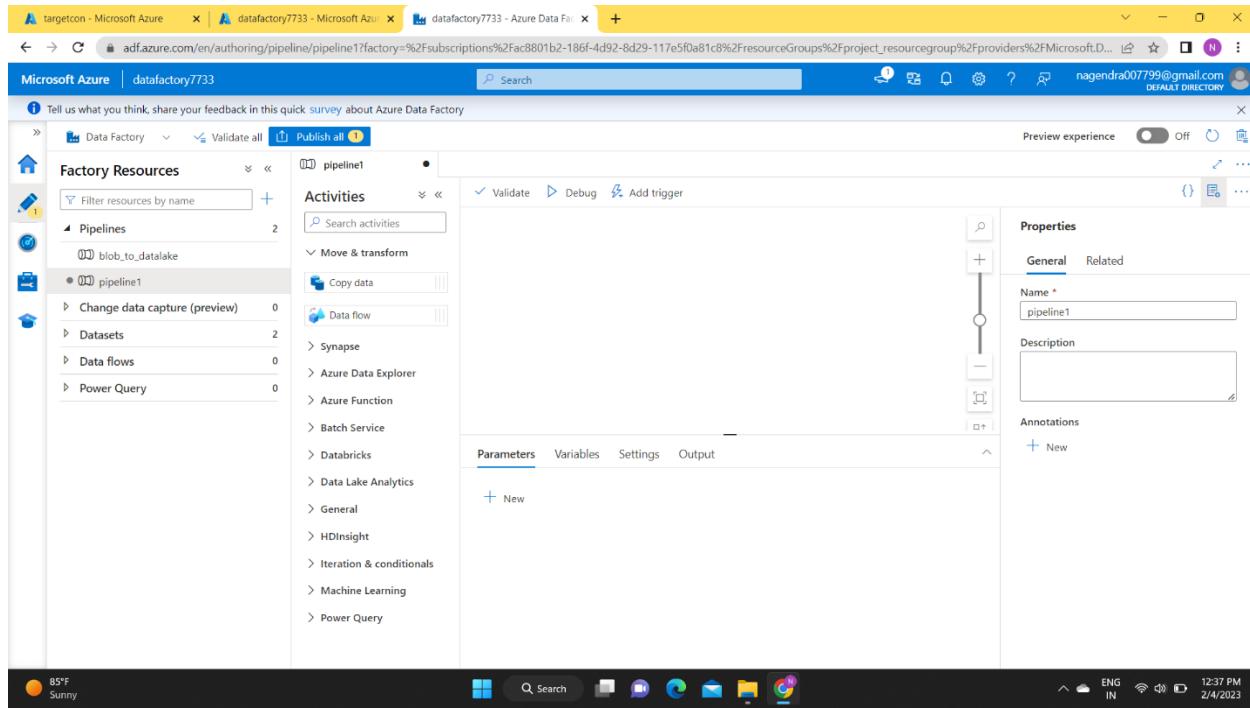
The results pane displays the data from the 'Customer' table:

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

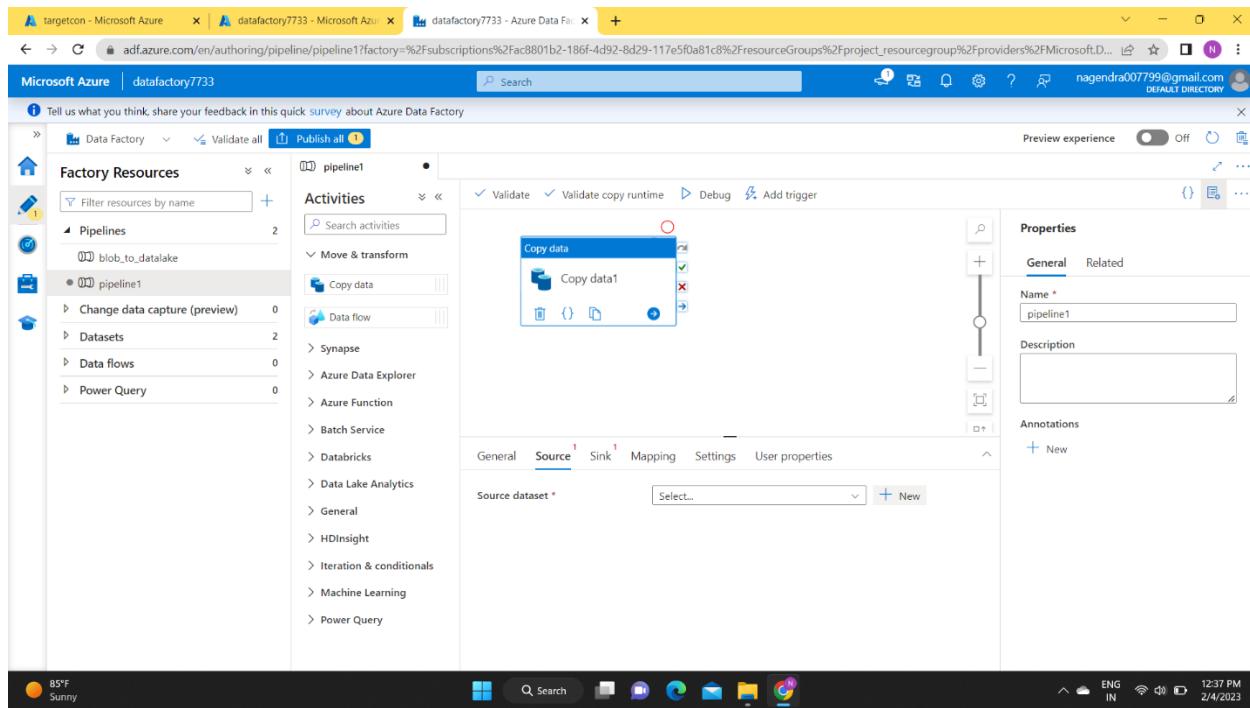
The status bar at the bottom indicates 'Query succeeded | 0s'.

- Now we can launch azure data factory.

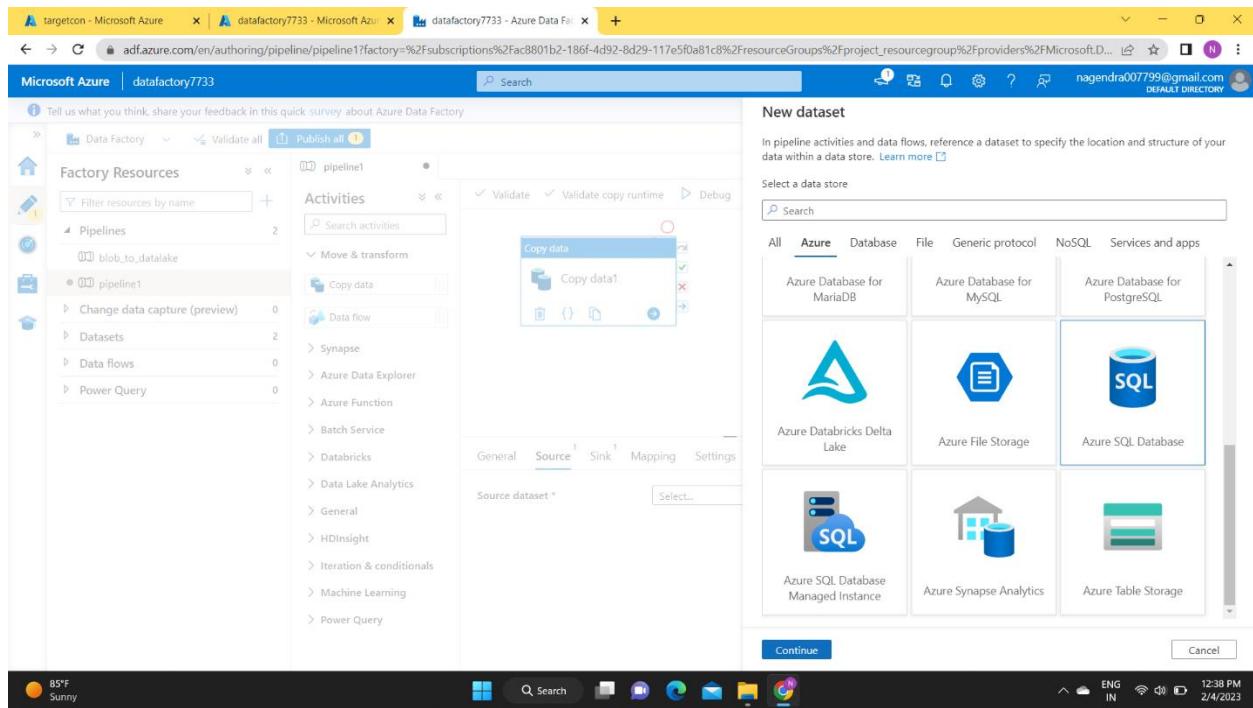
- Create new pipeline



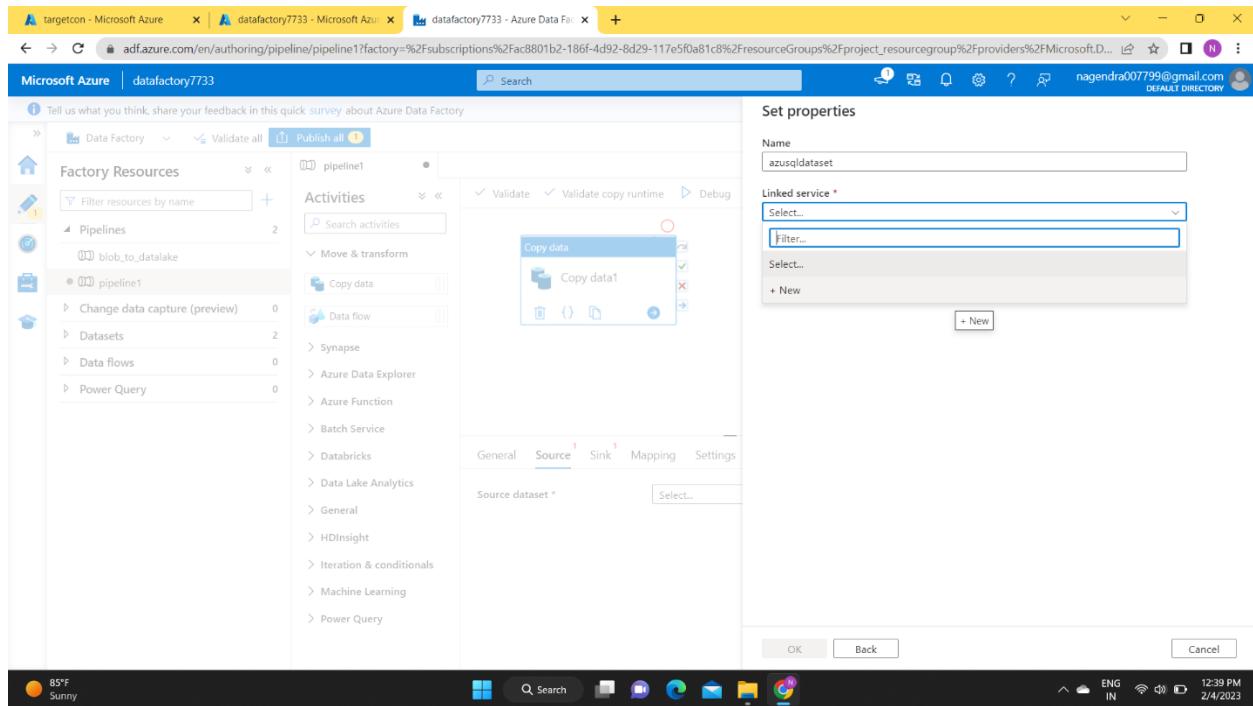
- Select copy data activity in adf.
- In source create new dataset for source azure sql db.



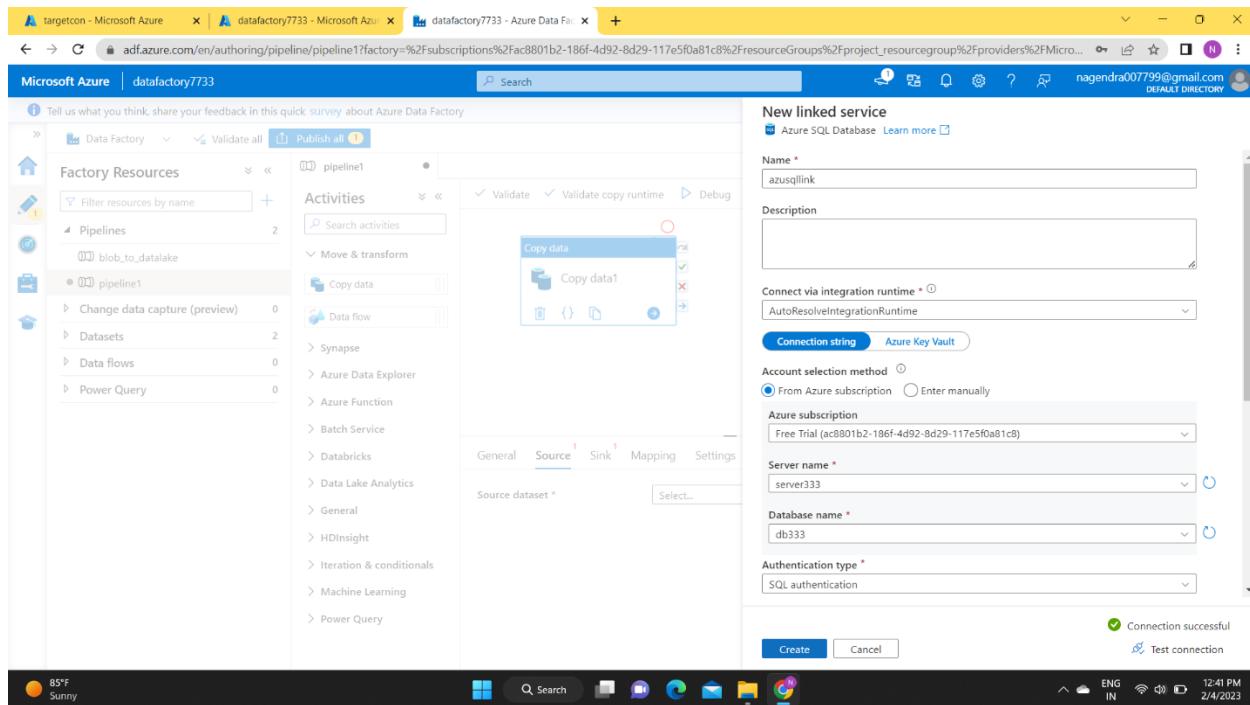
- In source you can choose azure sql db.



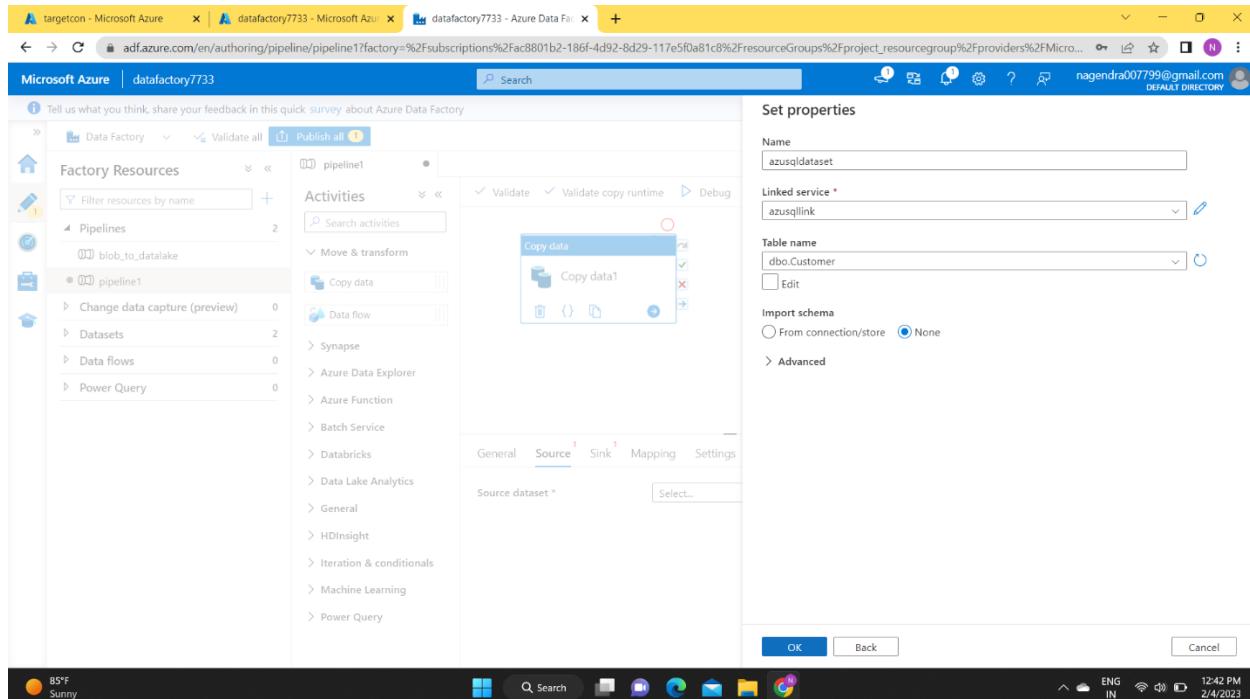
- Our dataset name is “azusqldataset”.



- 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'. 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

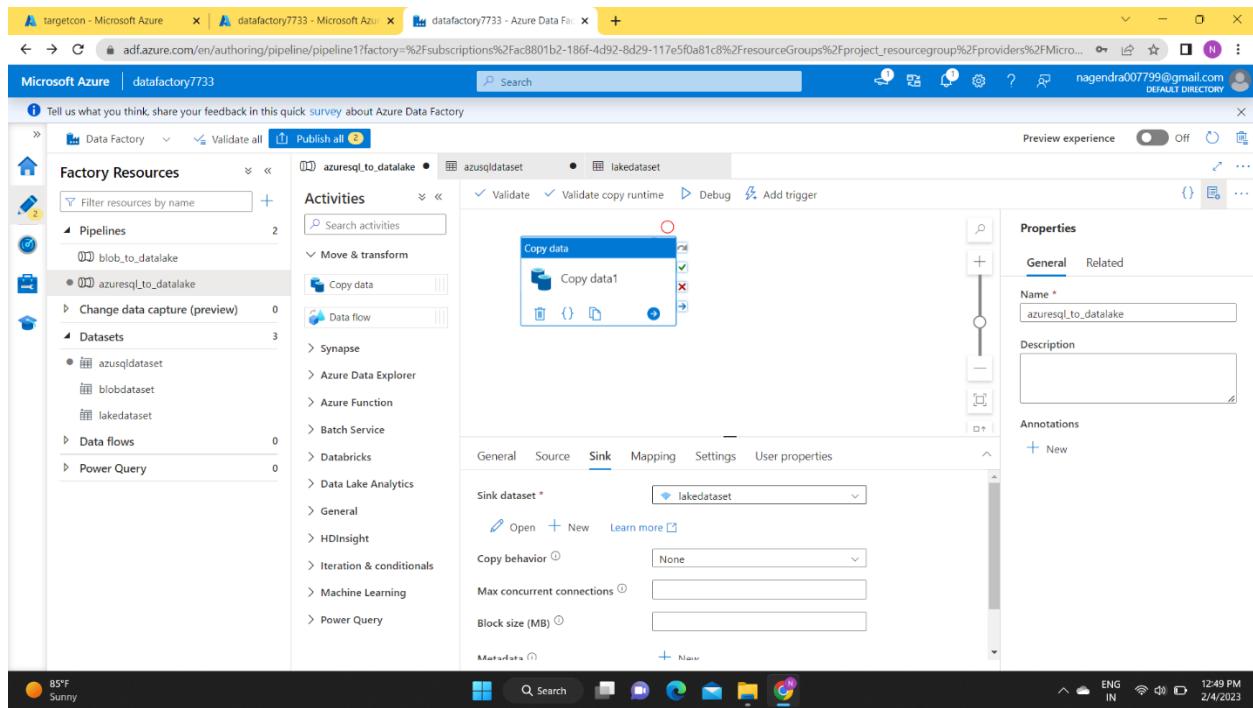
Properties: General, Related

Name: azuresql\_to\_datalake

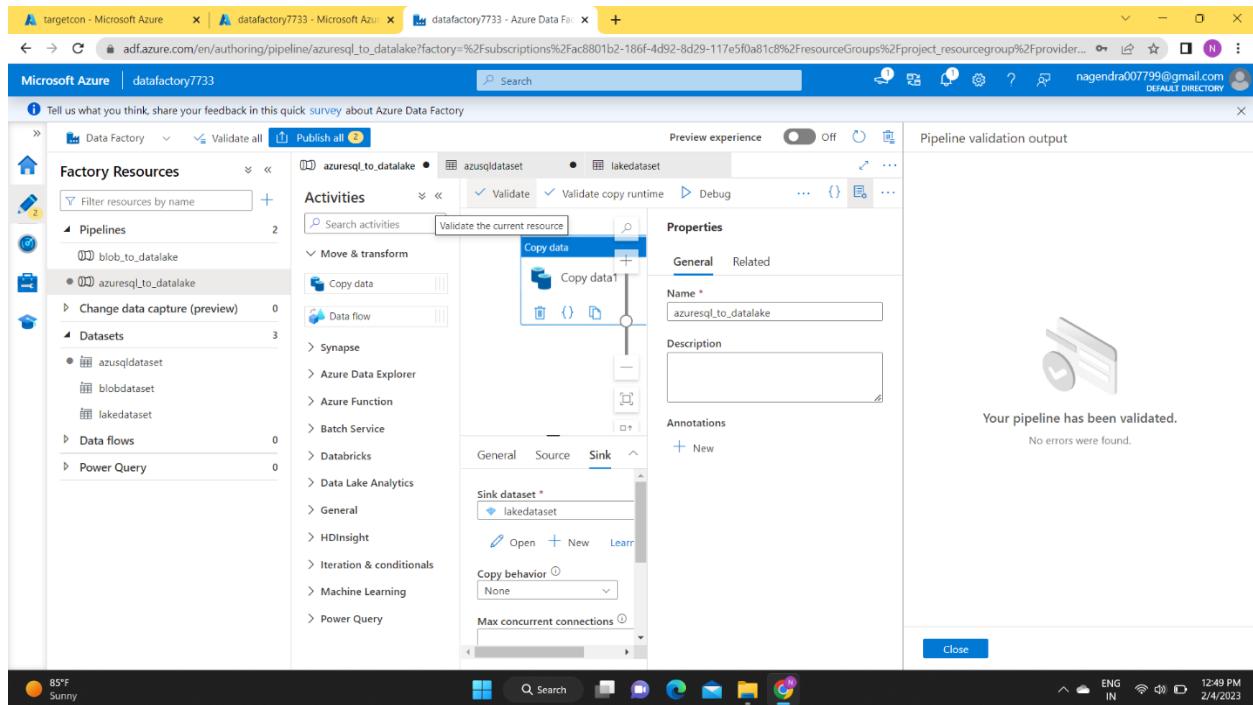
Description:

Annotations: + New

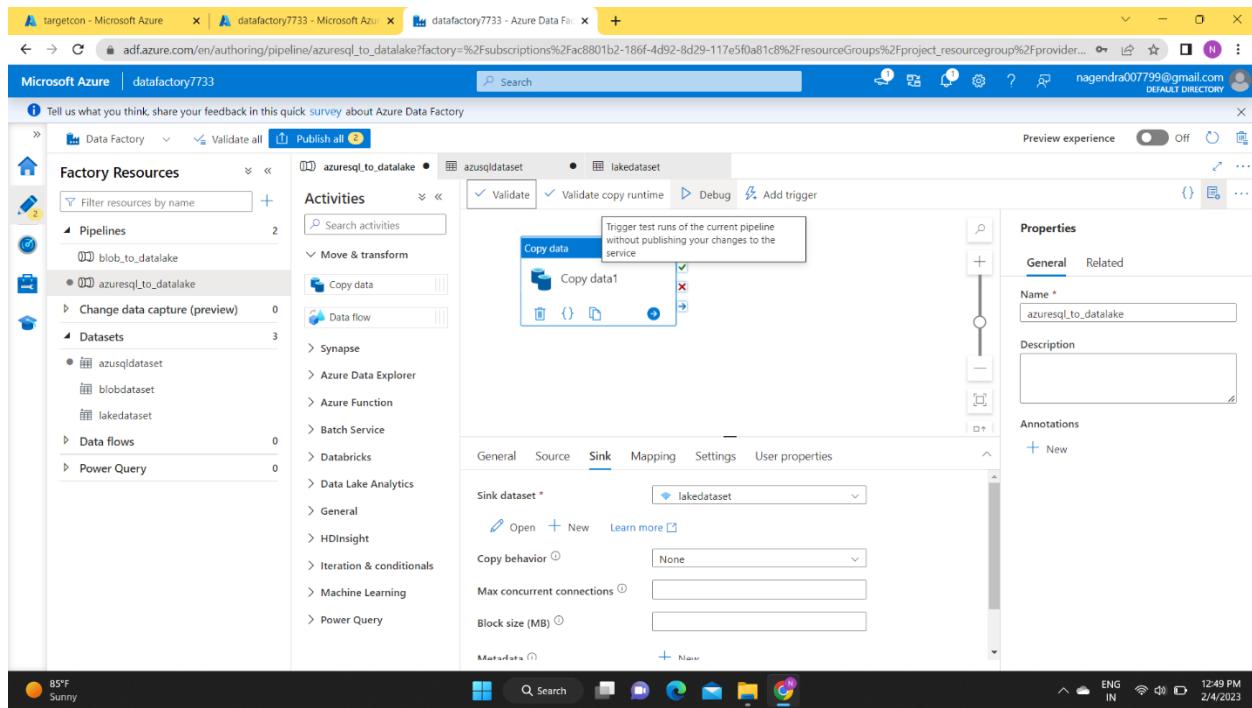
- we have data lake dataset and link service, so I am giving them.



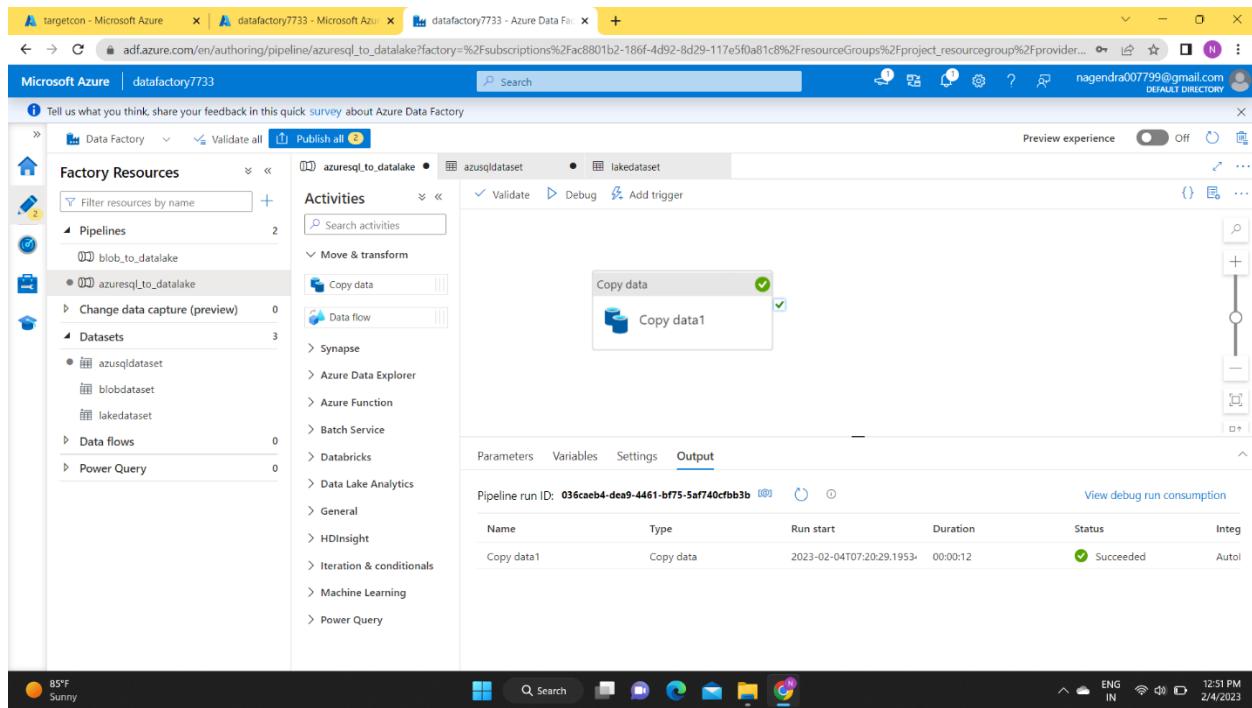
- now I am going to validate our pipeline



- now I am doing debug



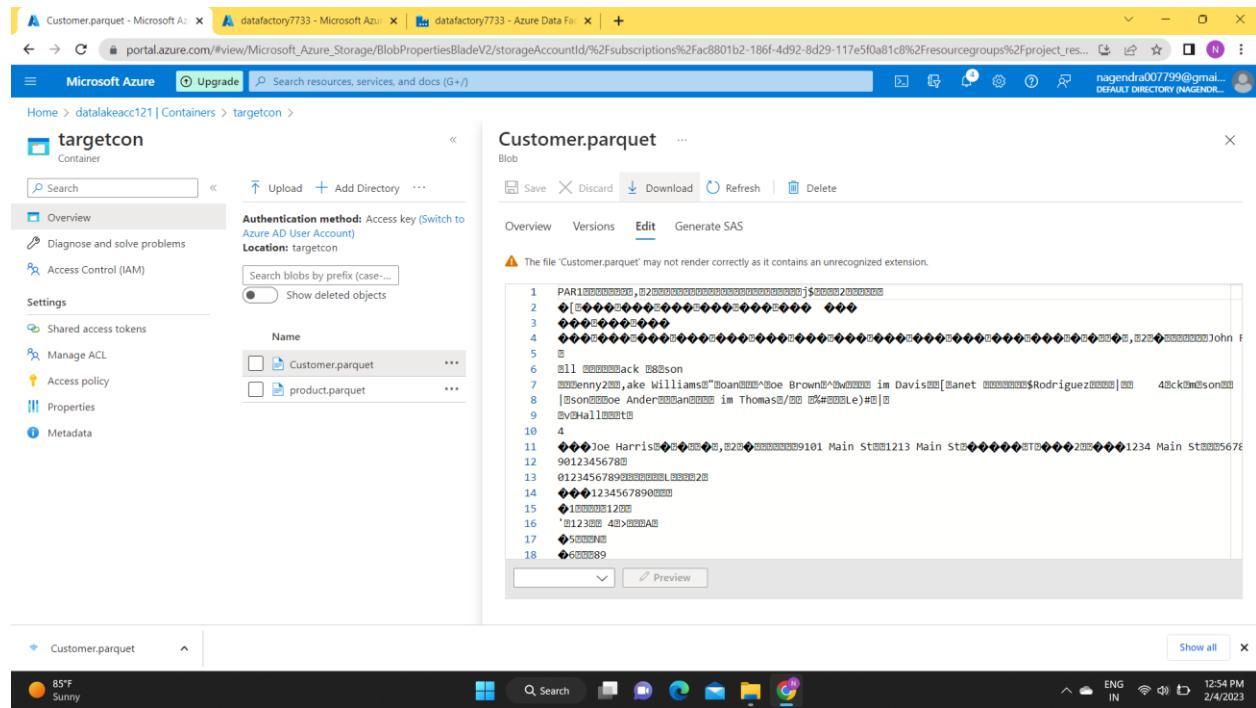
- pipeline is successful.



- Now I am checking customer data in target datakake

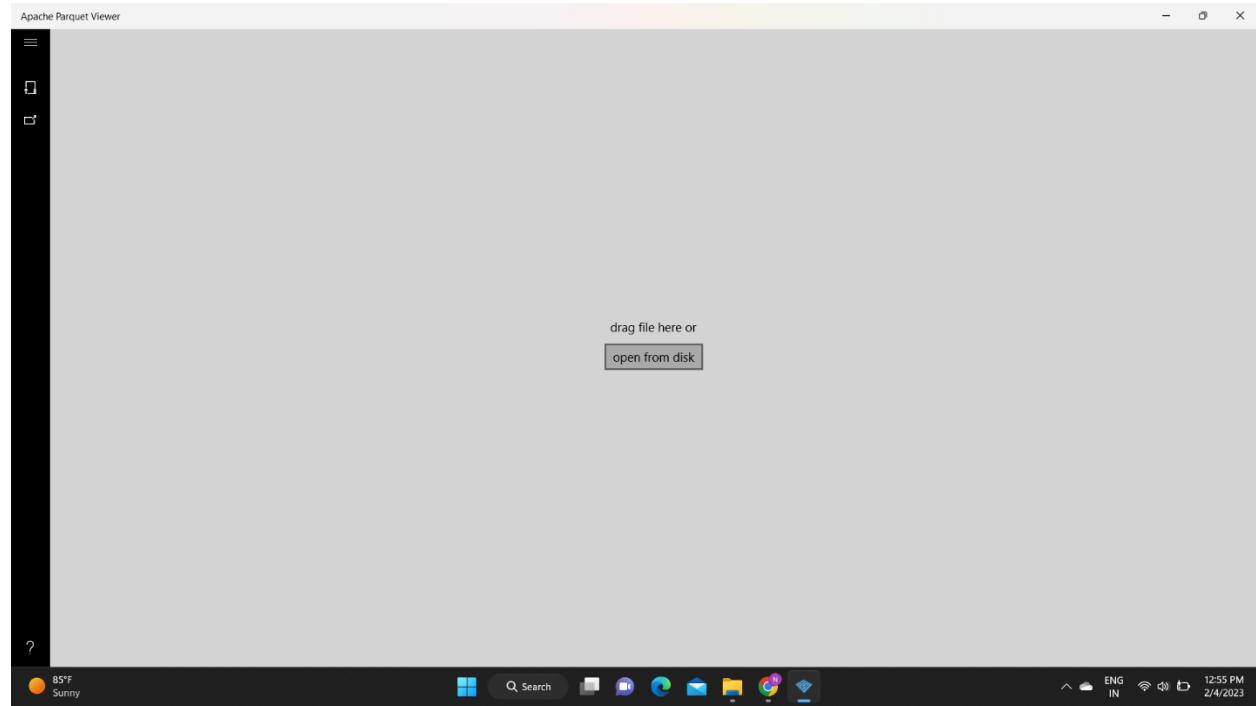
- Here is our customer file in parquet file format.

- We cannot understand data so download file.



The screenshot shows the Microsoft Azure Storage Blob Properties page for a file named 'Customer.parquet'. The file is located in a container named 'targetcon'. The content of the file is displayed as a large block of binary data, with a warning message: 'The file 'Customer.parquet' may not render correctly as it contains an unrecognized extension.' The file size is 1.2 MB and it was last modified on 2/4/2023 at 12:54 PM.

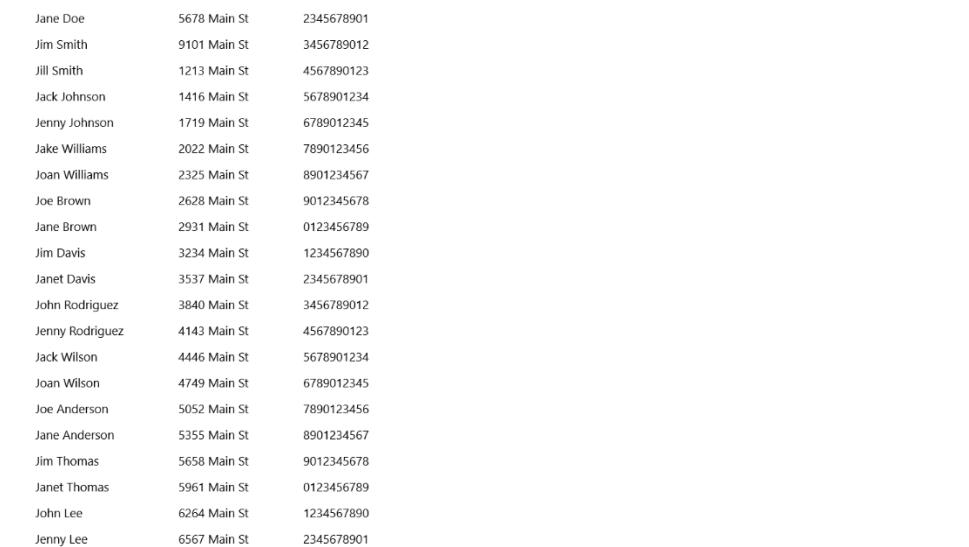
- I am open the app “advanced parquet viewer” and upload your parquet file here.



The screenshot shows the Apache Parquet Viewer application window. The interface is dark-themed with a light gray header. In the center, there is a large text input field with the placeholder 'drag file here or' and a button labeled 'open from disk'. The application window has a title bar 'Apache Parquet Viewer' and a standard Windows taskbar at the bottom.

- This is how our customer data looks like.

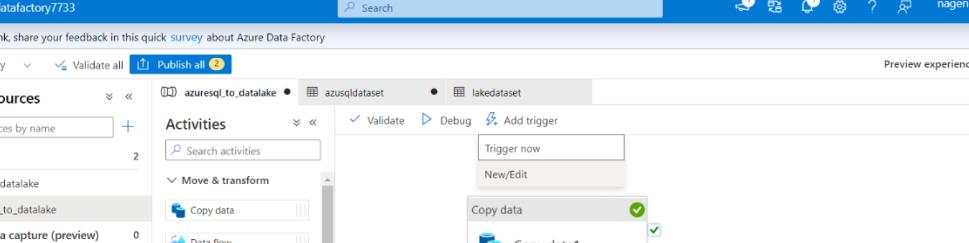
Apache Parquet Viewer



The screenshot shows a Windows application window titled "Apache Parquet Viewer". The main content is a table with four columns: "Customer\_ID", "Customer\_Name", "Customer\_Address", and "Customer\_Phone". The table contains 25 rows of data. The first few rows are as follows:

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

At the bottom left, there is a "Showing first 25 records." message. The bottom right shows system status icons for battery, signal, and date/time (12:56 PM, 2/4/2023). The bottom center features a Windows taskbar with icons for File Explorer, Task View, Mail, Edge, and File Explorer.



The screenshot shows the Microsoft Azure Data Factory interface. The left sidebar lists 'Factory Resources' including Pipelines, Datasets, Data flows, and Power Query. The main area shows a pipeline named 'azuresql\_to\_datalake' with a 'Copy data' activity selected. The pipeline run ID is 036cae84-dea9-4461-bf75-5ef740cfbb3b. The pipeline run table shows one run named 'Copy data1' of type 'Copy data' that has succeeded.

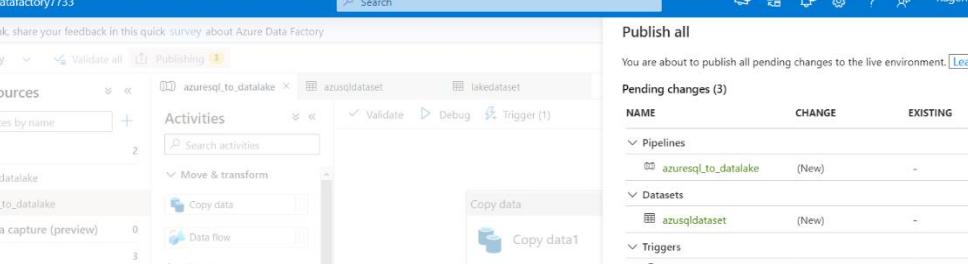
Name	Type	Run start	Duration	Status
Copy data1	Copy data	2023-02-04T07:20:29.195Z	00:00:12	Succeeded

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

The screenshot shows the Azure Data Factory pipeline editor. The left sidebar shows 'Factory Resources' with a pipeline named 'blob\_to\_datalake'. The main area shows an 'Activities' list with 'Copy data' and 'Copy data1' selected. A 'trigger1' trigger is listed in the 'Add triggers' dialog. The pipeline run ID is 036caeab4-dea9-4461-bf75-5af740cf0f1. The 'Output' tab is selected for the 'Copy data1' activity.

The screenshot shows the Azure Data Factory pipeline editor. The 'Trigger (1)' button is highlighted. The pipeline run ID is 036caeab4-dea9-4461-bf75-5af740cfbb3b. The 'Output' tab is selected for the 'Copy data1' activity.

- Now I am publishing the pipeline.



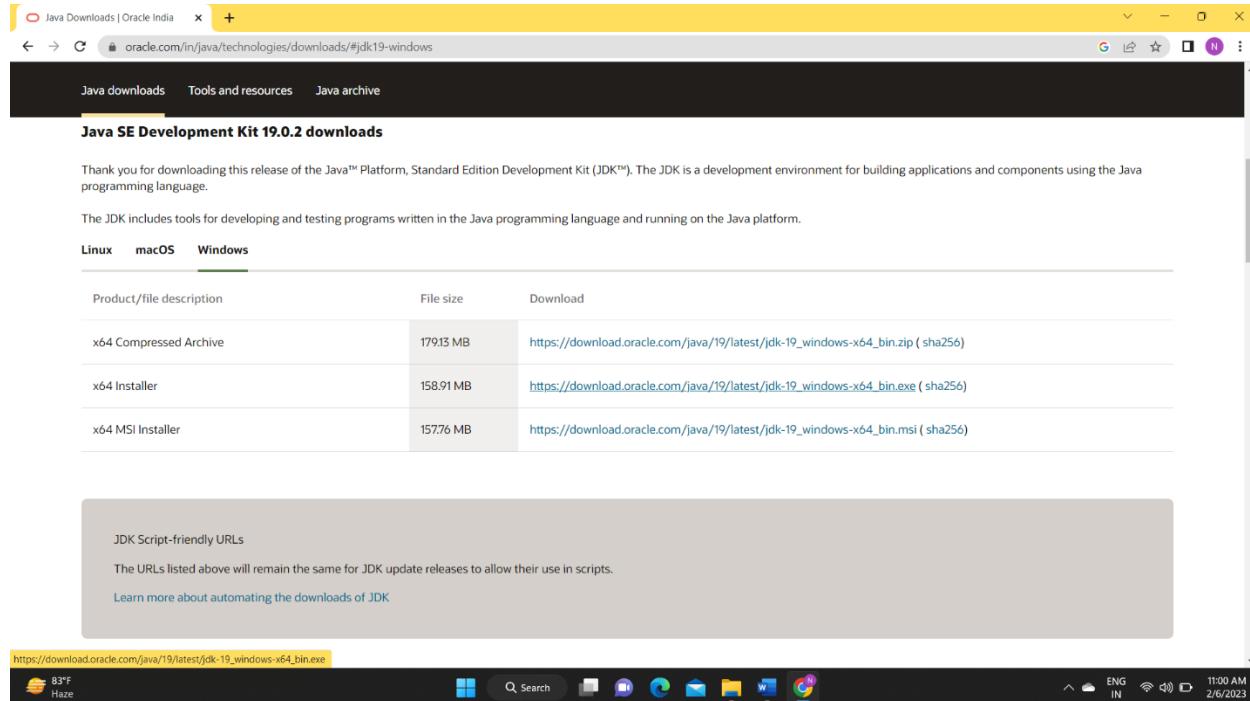
The screenshot shows the Azure Data Factory interface with a 'Publish all' dialog box open. The dialog box is titled 'Publish all' and contains a message: 'You are about to publish all pending changes to the live environment.' It includes a 'Learn more' link. Below this, a table lists 'Pending changes (3)' under three categories: Pipelines, Datasets, and Triggers. The 'azuresql\_to\_datalake' pipeline is listed under Pipelines as a new item. The 'azusqldataset' dataset is listed under Datasets as a new item. The 'trigger' trigger is listed under Triggers as an edited item. At the bottom of the dialog box are 'Publish' and 'Cancel' buttons. The background shows the Data Factory dashboard with a pipeline named 'azuresql\_to\_datalake' and a dataset named 'azusqldataset'.

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/>

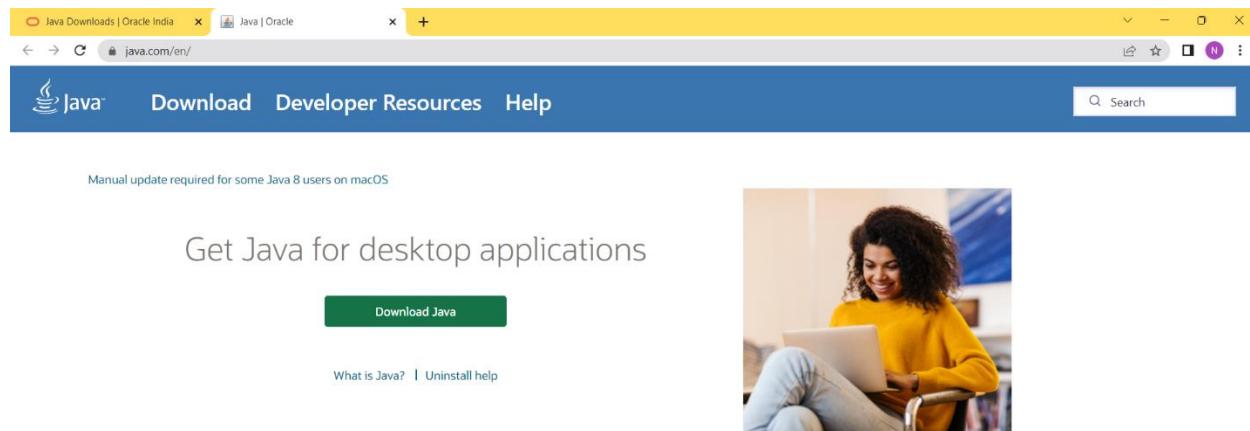


The screenshot shows the Oracle Java Downloads page for Java SE Development Kit 19.0.2. The 'Windows' tab is selected. A table lists four download options:

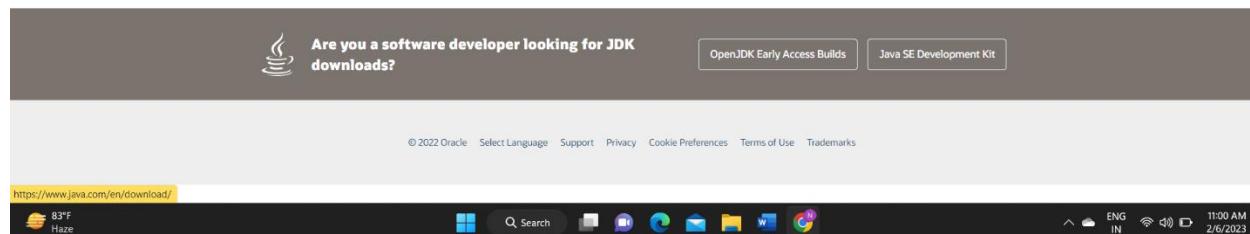
Product/file description	File size	Download
x64 Compressed Archive	179.13 MB	<a href="https://download.oracle.com/java/19/latest/jdk-19_windows-x64_bin.zip">https://download.oracle.com/java/19/latest/jdk-19_windows-x64_bin.zip</a> ( sha256 )
x64 Installer	158.91 MB	<a href="https://download.oracle.com/java/19/latest/jdk-19_windows-x64_bin.exe">https://download.oracle.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 )

Below the table, a box contains 'JDK Script-friendly URLs' and a note: 'The URLs listed above will remain the same for JDK update releases to allow their use in scripts.' It also links to '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/>

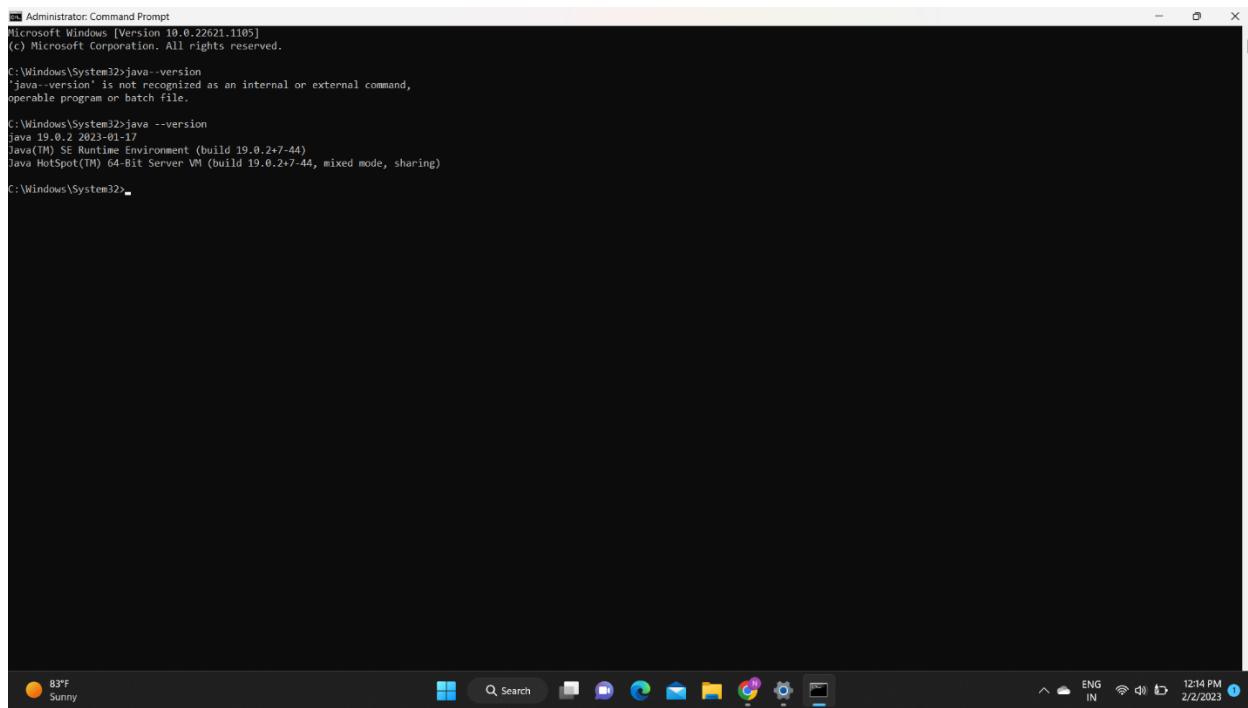


The screenshot shows the Java.com website. The 'Download' tab is selected. A message at the top says 'Manual update required for some Java 8 users on macOS'. Below it, a heading 'Get Java for desktop applications' has a 'Download Java' button. A sidebar on the right features a photo of a woman using a laptop.



The screenshot shows the Java.com download page. It features a 'Are you a software developer looking for JDK downloads?' section with 'OpenJDK Early Access Builds' and 'Java SE Development Kit' buttons. At the bottom, there are links for '© 2022 Oracle' and 'Select Language' through 'Trademarks'.

- 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.

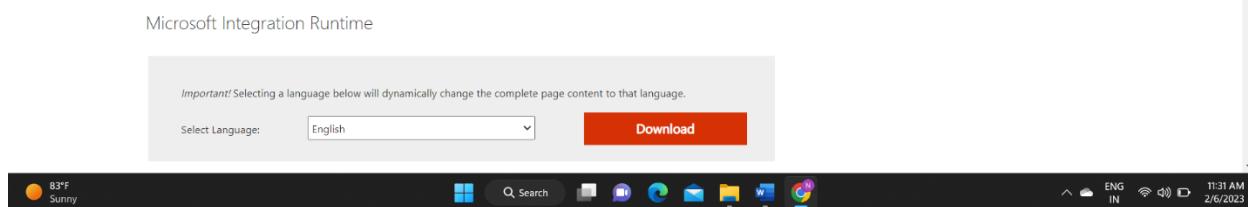
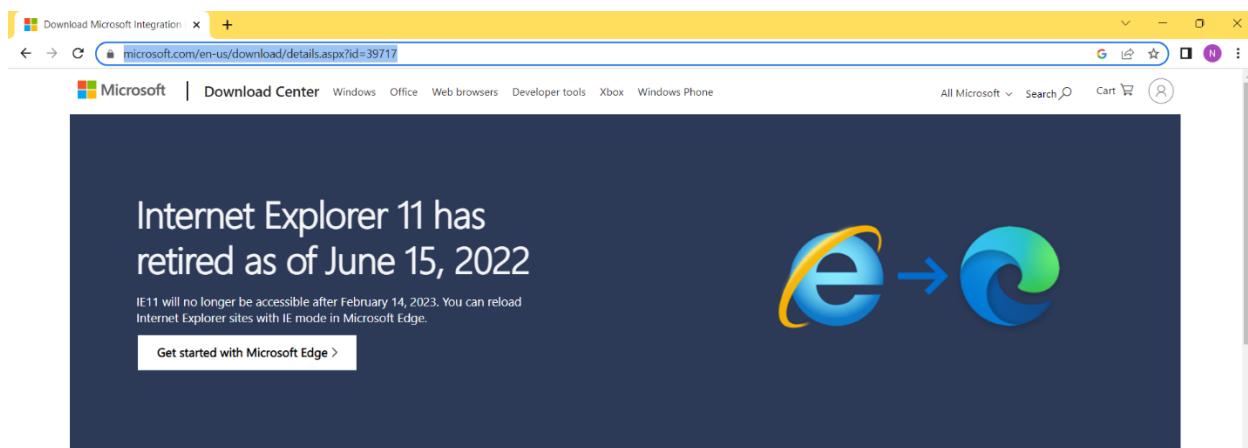
C:\Windows\System32>java --version
java 19.0.2 2023-01-17
Java(TM) SE Runtime Environment (build 19.0.2+7-44)
Java HotSpot(TM) 64-Bit Server VM (build 19.0.2+7-44, mixed mode, sharing)

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 - Microsoft Azure

Microsoft Azure

datafactory7733

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

85°F Sunny

Launch studio

Quick Starts

Tutorials

Template Gallery

Training Modules

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	---

Preview experience

85°F Sunny

- Select azure self hosted runtime.

Integration runtime setup

Integration Runtime is the native compute used to execute or dispatch activities. Choose what integration runtime to create based on required capabilities. [Learn more](#)

**Azure, Self-Hosted**

Perform data flows, data movement and dispatch activities to external compute.

**Azure-SSIS**

Lift-and-shift existing SSIS packages to execute in Azure.

**Continue** **Cancel**

Integration runtime setup

**Network environment:**

Choose the network environment of the data source / destination or external compute to which the integration runtime will connect to for data flows, data movement or dispatch activities:

**Azure**

Use this for running data flows, data movement, external and pipeline activities in a fully managed, serverless compute in Azure.

**Self-Hosted**

Use this for running activities in an on-premises / private network

**External Resources:**

You can use an existing self-hosted integration runtime that exists in another resource. This way you can reuse your existing infrastructure where self-hosted integration runtime is setup.

**Linked Self-Hosted**

[Learn more](#)

**Continue** **Back** **Cancel**

- 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

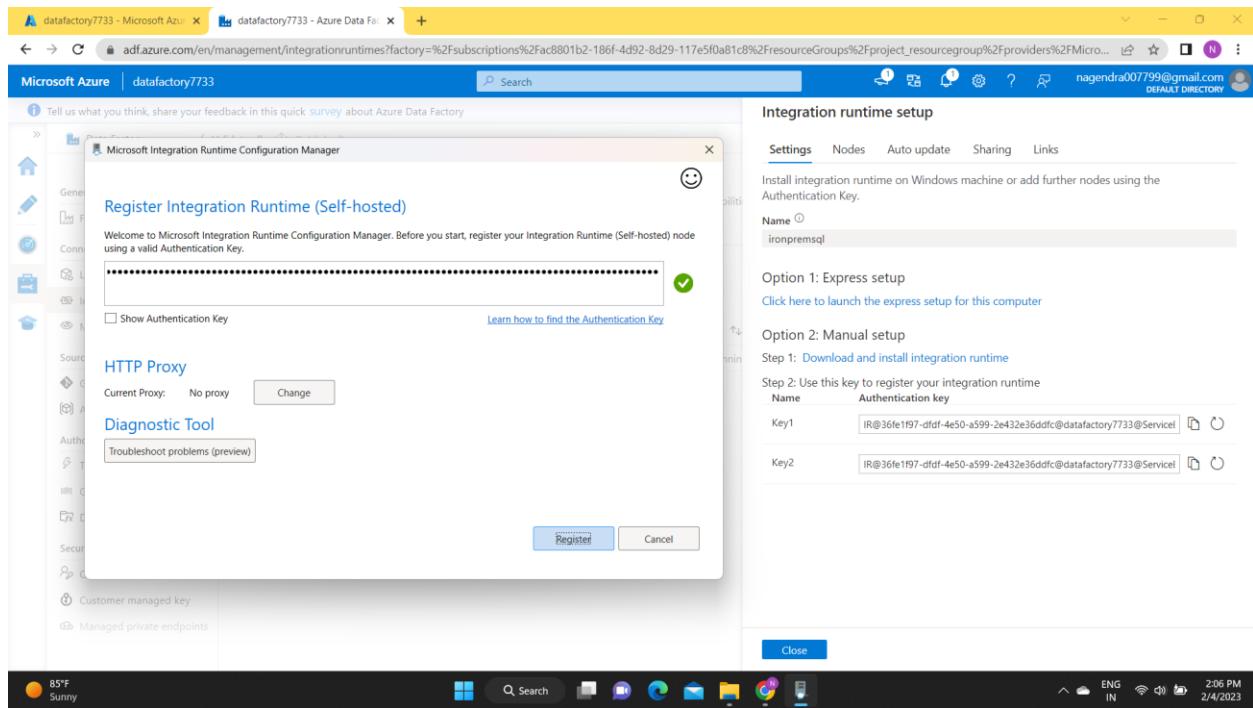
Step 1: [Download and install integration runtime](#)

Step 2: Use this key to register your integration runtime

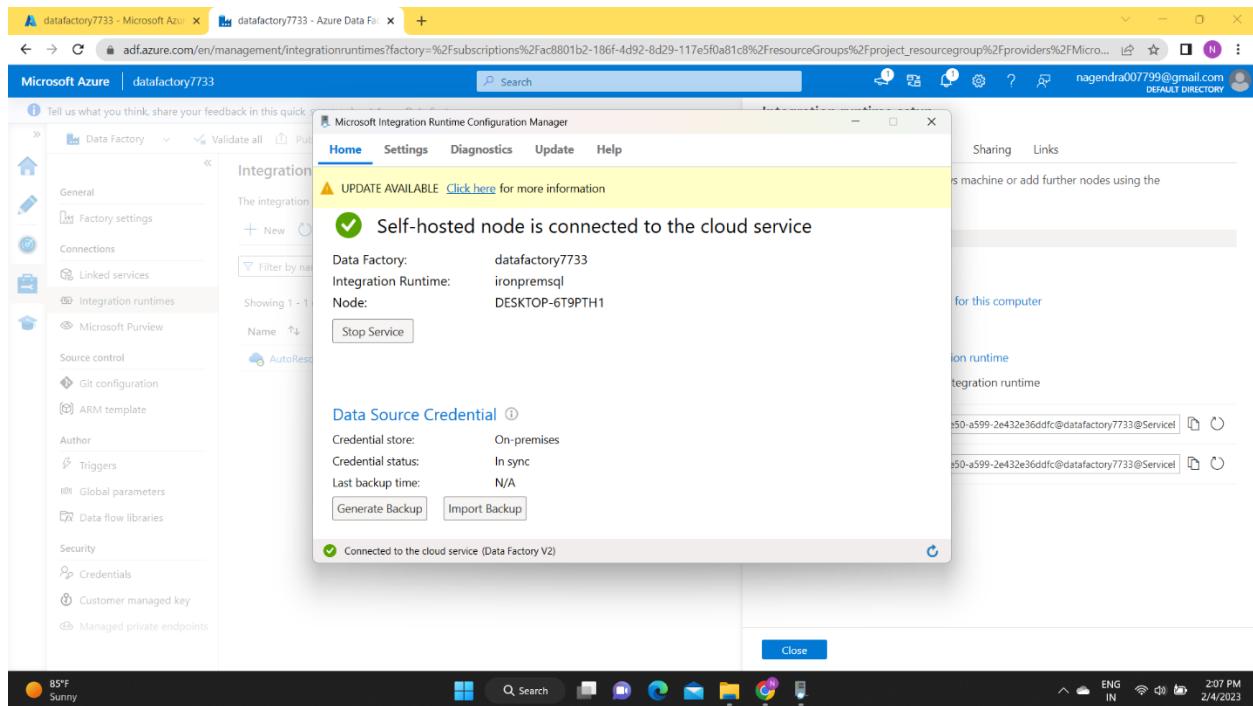
Name	Authentication key
Key1	IR@36fe1f97-dfdf-4e50-a599-2e432e36ddfc@datafactory7733@Service
Key2	IR@36fe1f97-dfdf-4e50-a599-2e432e36ddfc@datafactory7733@Service

Close

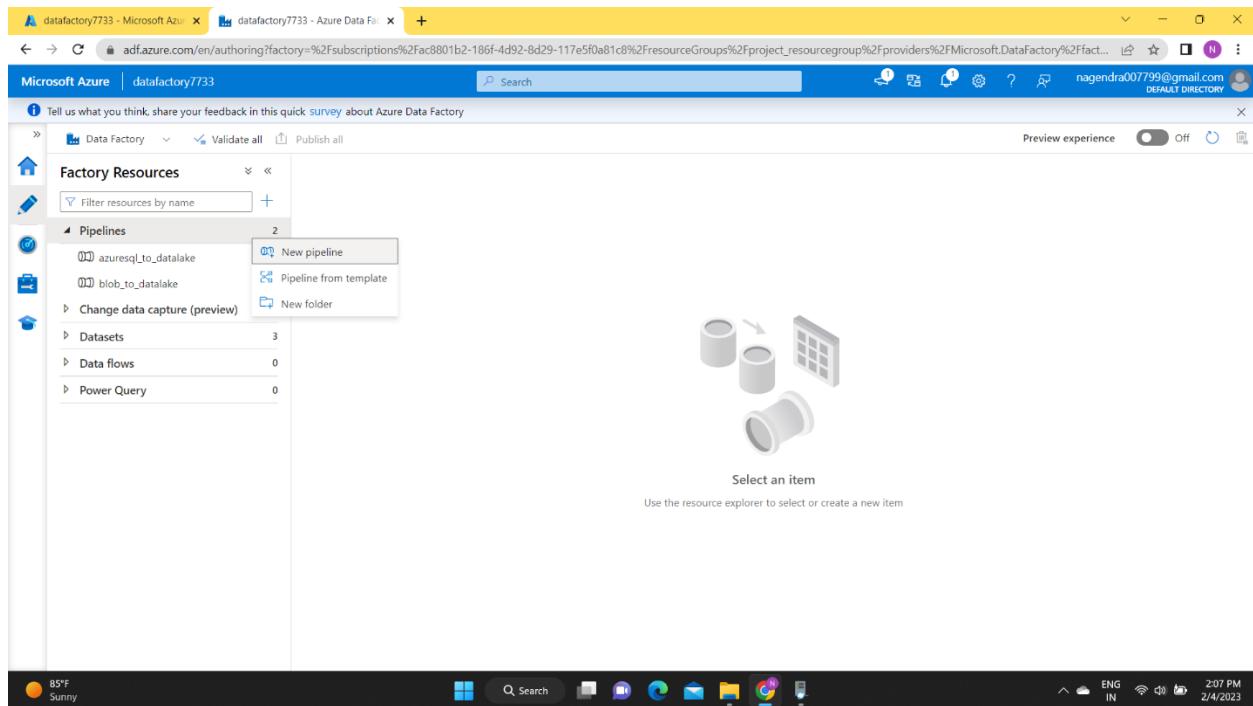
- Open windows integration runtime and paste the link in that box.



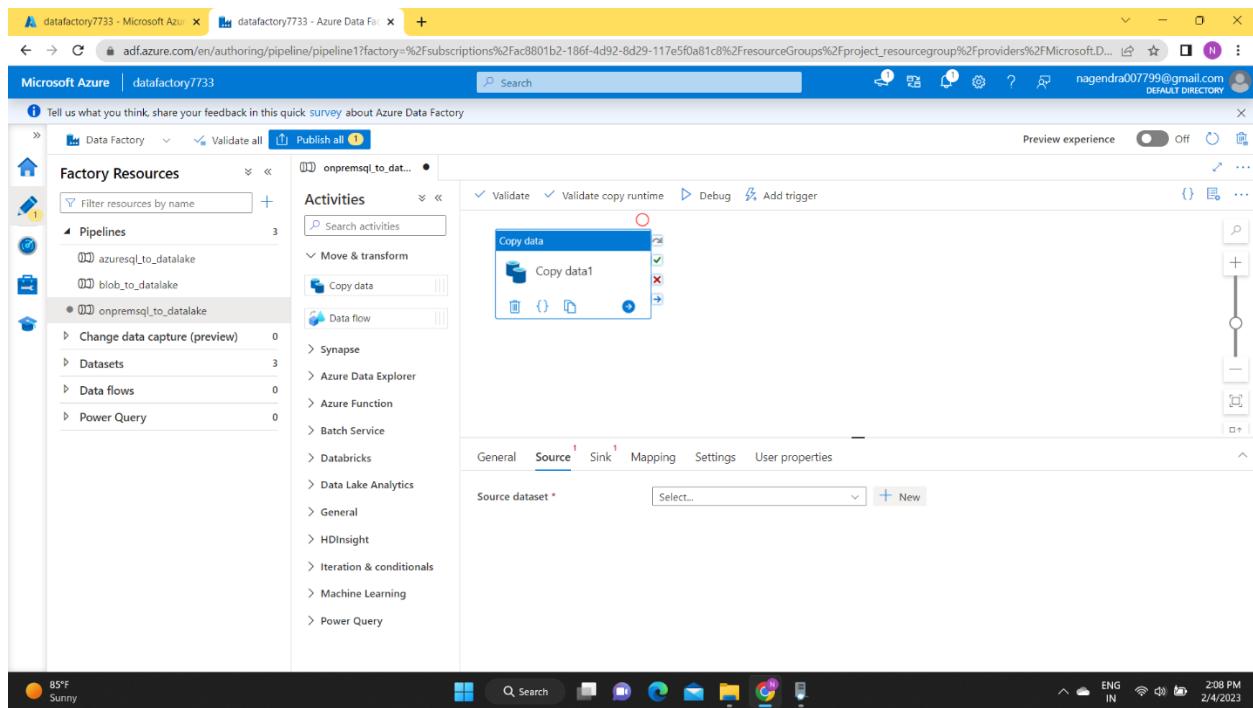
- Now our local windows server connected to azure cloud.



- 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. On the right, a 'New dataset' dialog is open, displaying a grid of data store icons. The 'SQL server' icon is highlighted, indicating it is the selected source dataset for the activity.

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

The screenshot shows the 'Set properties' dialog for a dataset named 'onpremsqldataset'. The 'Linked service' dropdown is open, showing a list of options: 'Select...', 'Filter...', 'Select...', and '+ New'. The 'Name' field is filled with 'onpremsqldataset'. The dialog has 'OK' and 'Cancel' buttons at the bottom.

- 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

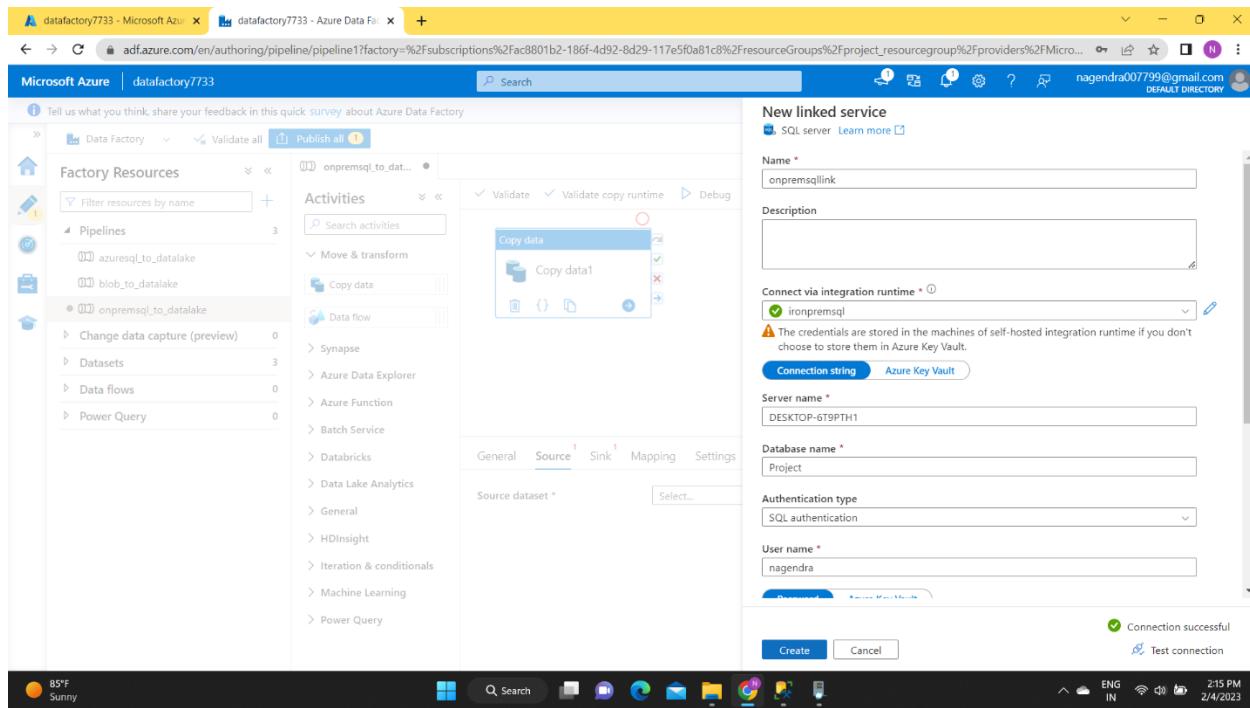
The screenshot shows the Azure Data Factory portal. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. In the center, a pipeline named 'onpremssql\_to\_datalake' is selected. A 'Copy data' activity is highlighted. On the right, a 'New linked service' dialog is open, with the 'Name' field set to 'onpremssqllink'. The 'Connect via integration runtime' dropdown is set to 'AutoResolveIntegrationRuntime', and the 'Database name' dropdown is set to 'ironpremssql'. The 'Source' tab of the 'Copy data1' activity is selected. The status bar at the bottom shows the date and time as 2/4/2023.

The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface. The 'File' menu is open, showing options like 'File', 'Edit', 'View', 'Tools', 'Window', and 'Help'. A 'Connect to Server' dialog is open in the foreground, titled 'SQL Server'. It shows the following connection details:
 

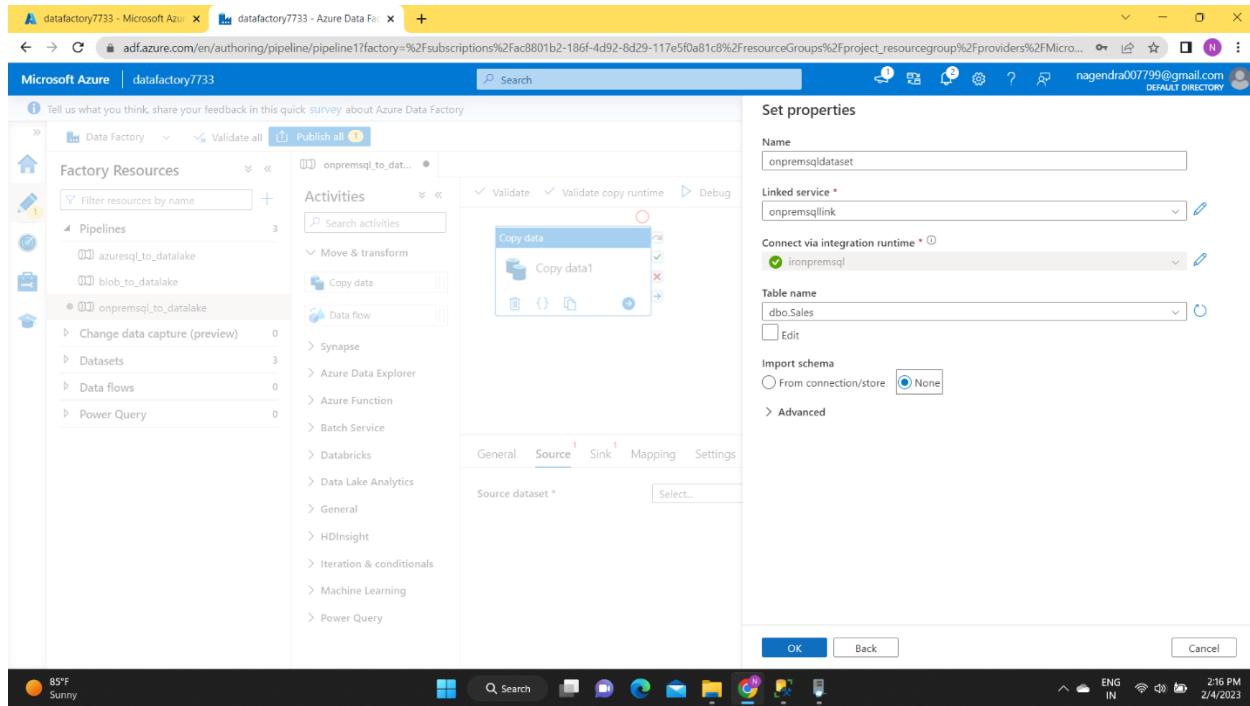
- Server type: Database Engine
- Server name: DESKTOP-618PT1H
- Authentication: SQL Server Authentication
- Login: nagendra
- Password: \*\*\*\*\*
- Remember password: checked

 The 'Connect' button is visible at the bottom of the dialog. The SSMS interface includes a toolbar with various icons and a status bar at the bottom showing 'Ready', 'AH43 / NH44 /...', 'Construction', and the date and time as 2/4/2023.

- 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

Preview experience Off

- This is the preview of our sales table.

Preview data

Linked service: onpremsqllink

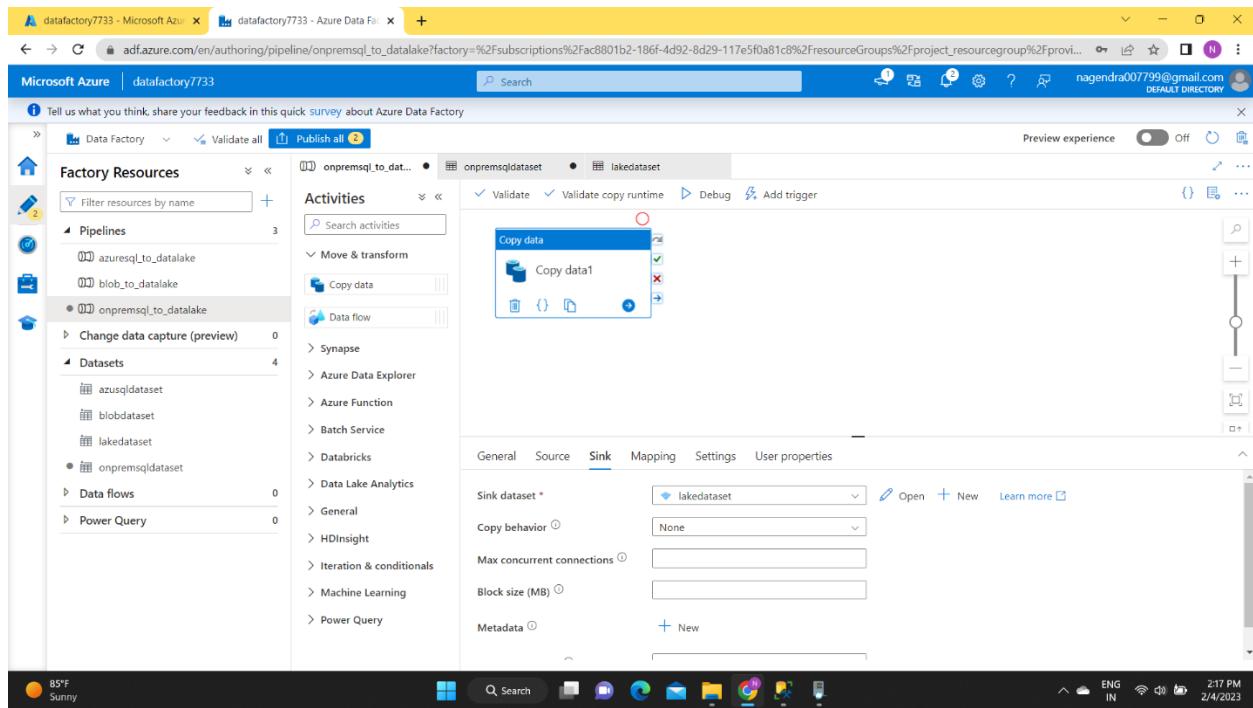
Object: Sales

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

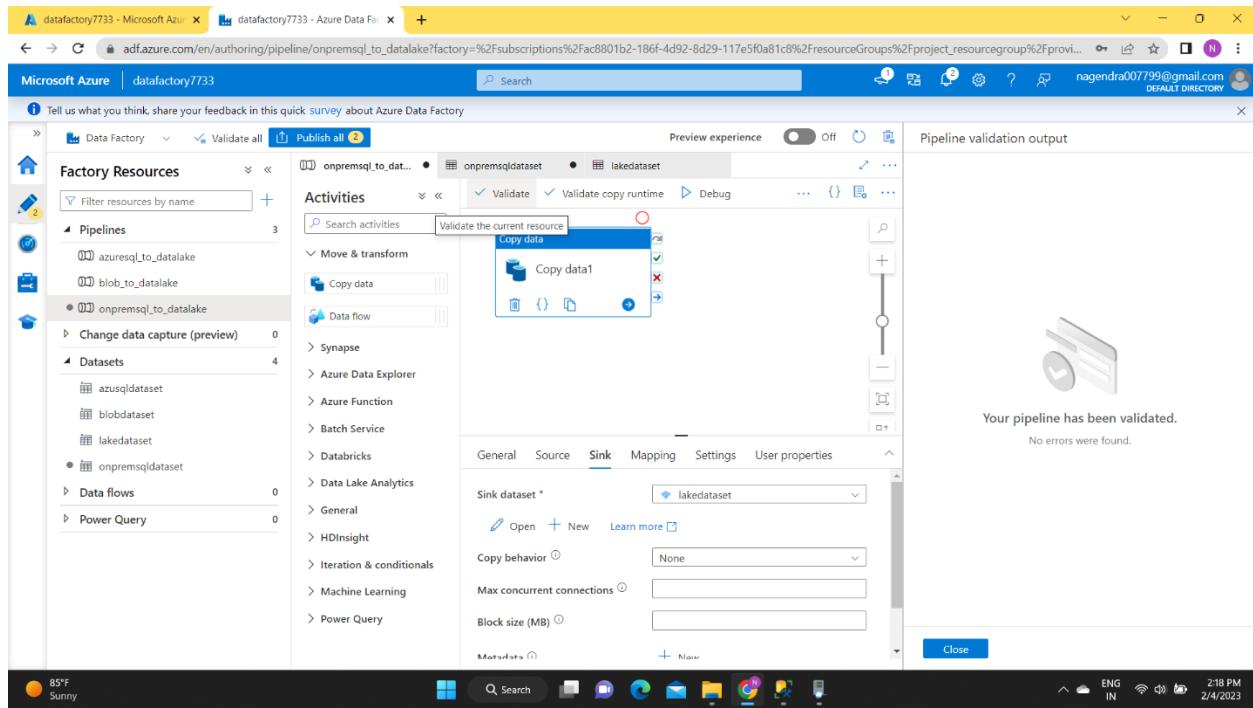
Partition option None Physical partitions of table Dynamic range

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

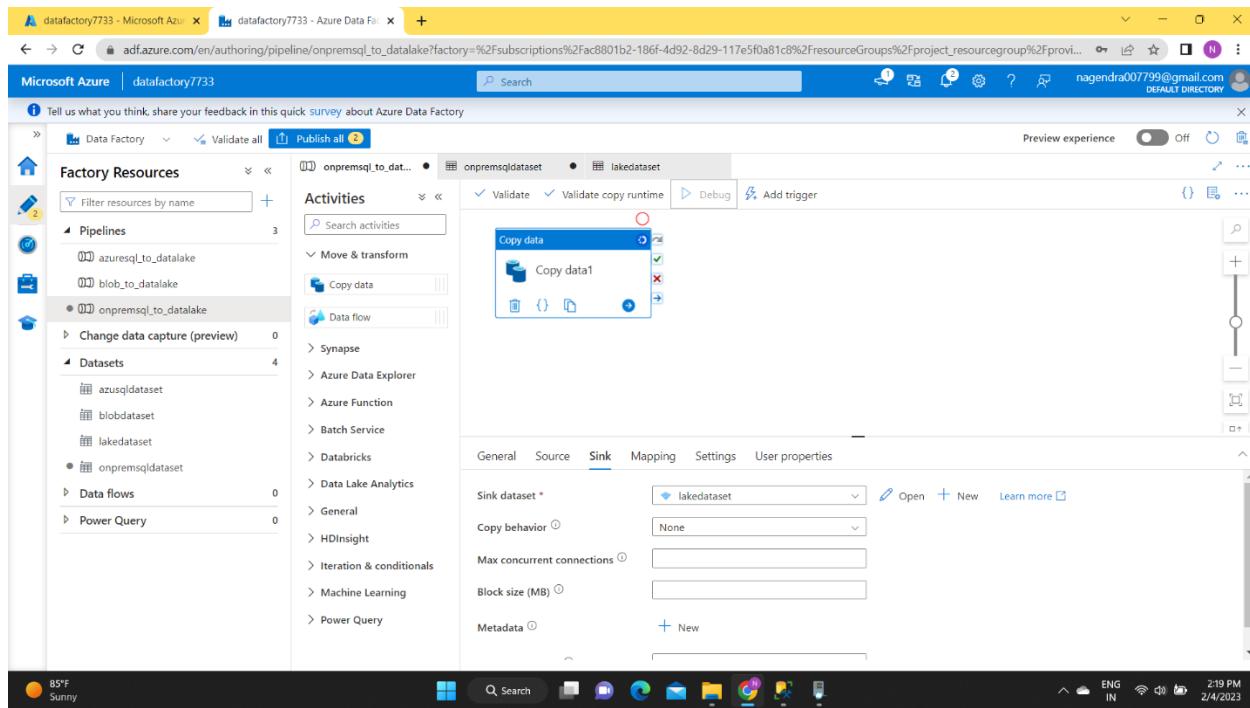
- In sink we have data lake dataset .



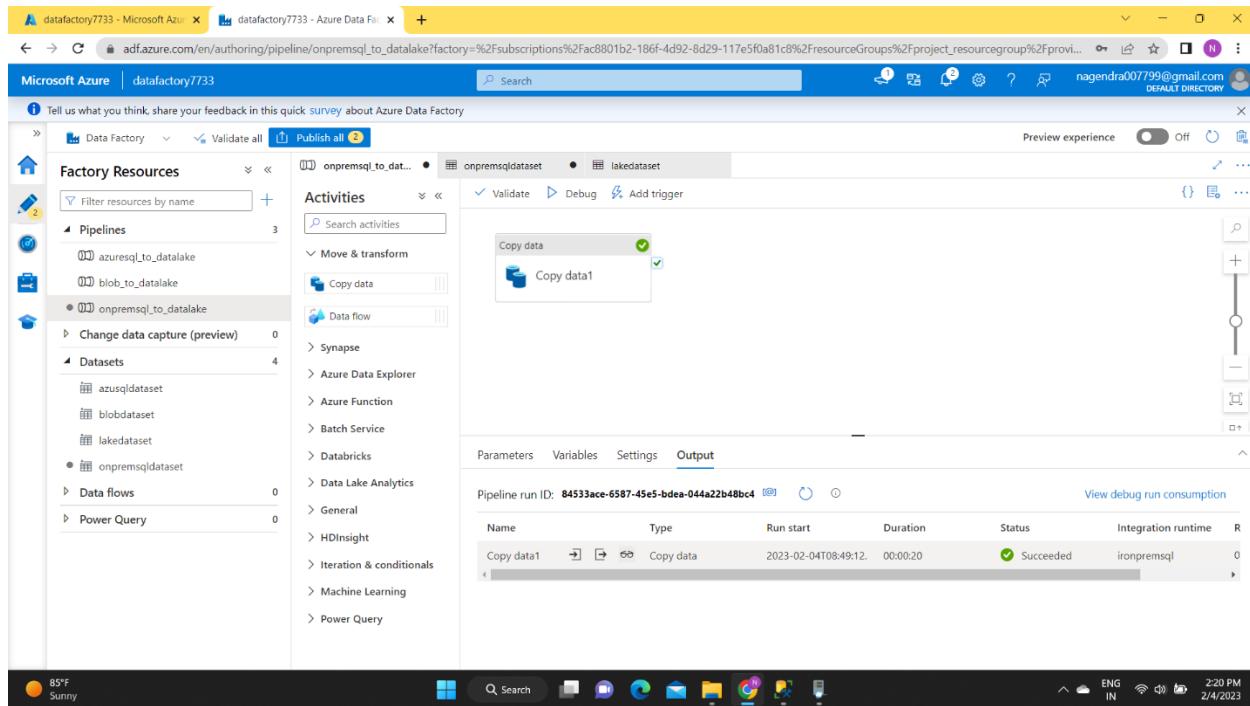
- Validate our pipeline .



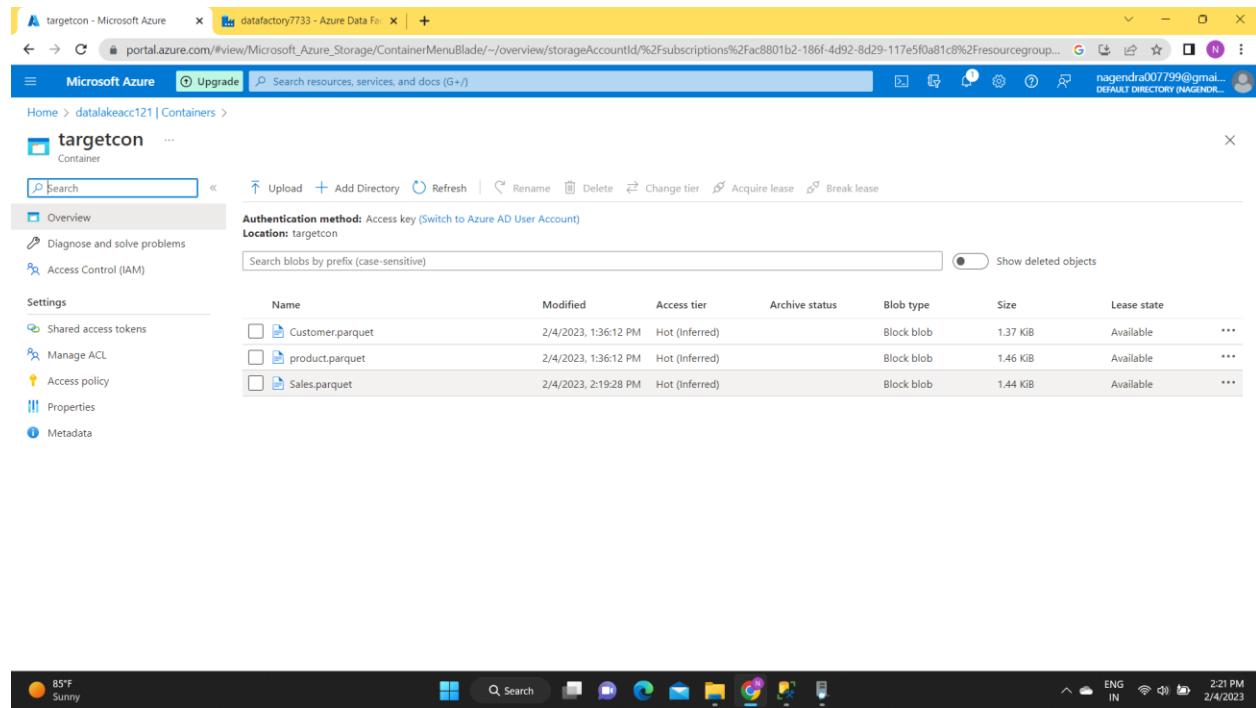
- Now do the debug.



- Our pipeline is succed

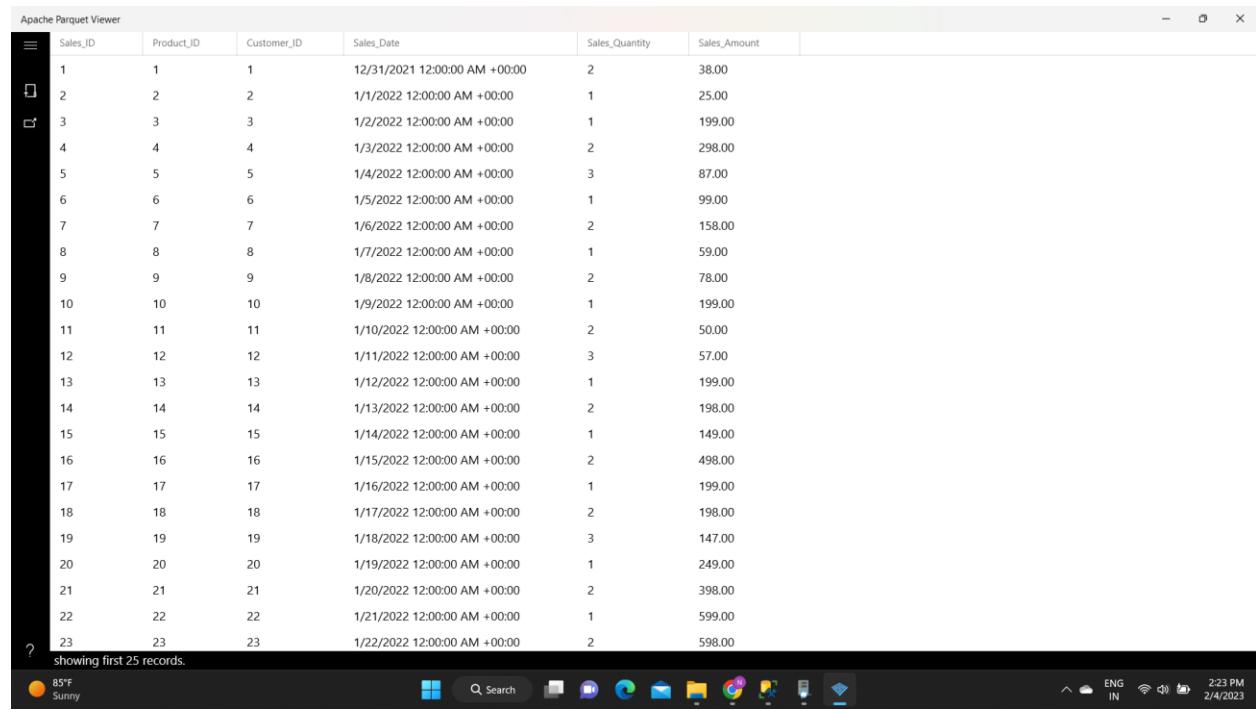


- This is our sales.parquet table in unified data lake.



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.

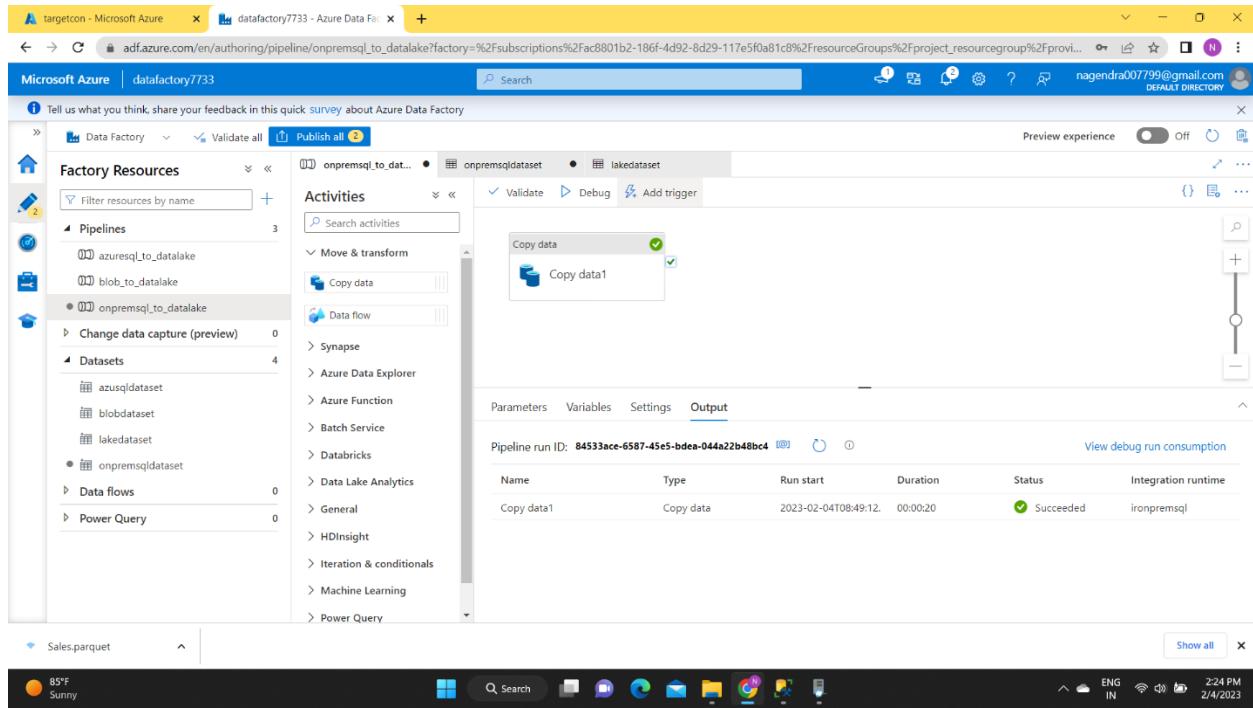


The screenshot shows the Apache Parquet Viewer application displaying the contents of the sales.parquet file. The data is presented in a table with columns: Sales\_ID, Product\_ID, Customer\_ID, Sales\_Date, Sales\_Quantity, and Sales\_Amount. The table contains 25 records, with the first few rows shown 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.

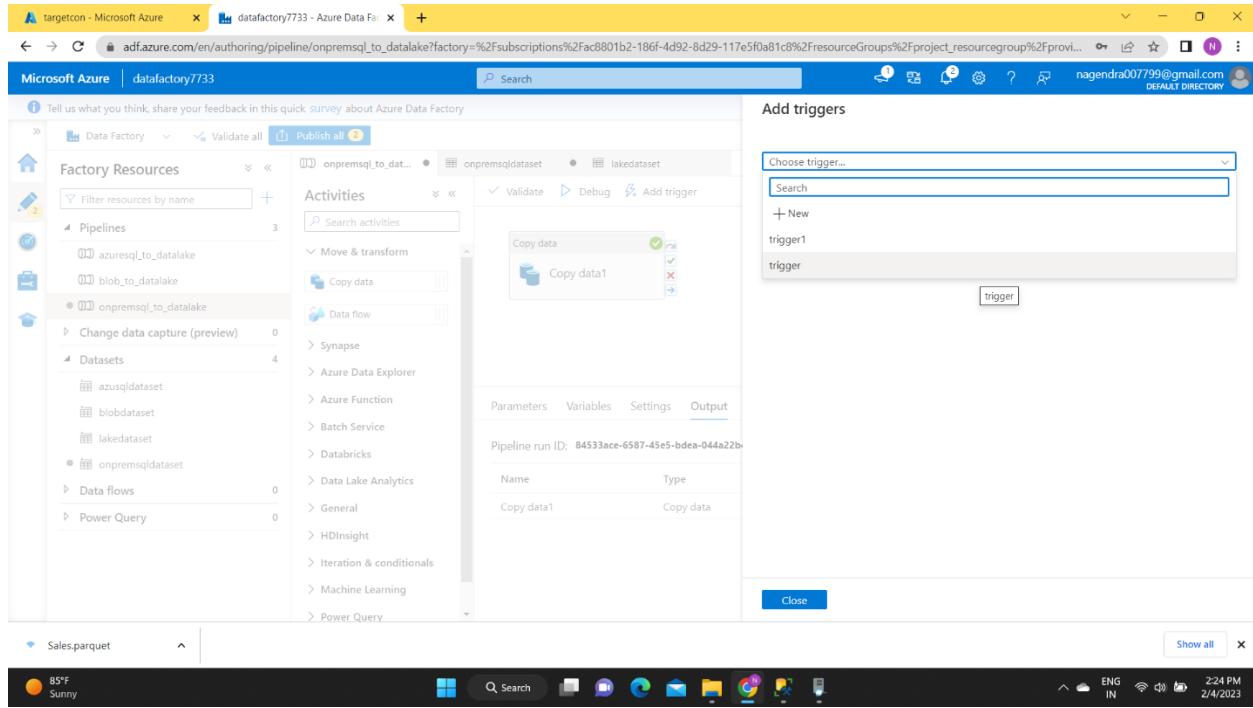
Sales_ID	Product_ID	Customer_ID	Sales_Date	Sales_Quantity	Sales_Amount
1	1	1	12/31/2021 12:00:00 AM +00:00	2	38.00
2	2	2	1/1/2022 12:00:00 AM +00:00	1	25.00
3	3	3	1/2/2022 12:00:00 AM +00:00	1	199.00
4	4	4	1/3/2022 12:00:00 AM +00:00	2	298.00
5	5	5	1/4/2022 12:00:00 AM +00:00	3	87.00
6	6	6	1/5/2022 12:00:00 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.



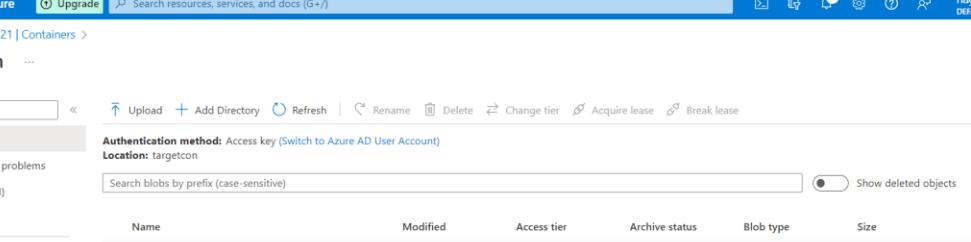
- Here I am giving the previous trigger.



- Now I am going to publish the pipeline.

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

- THIS IS OUR UNIFIED STORAGE SYSTEM AND ALL 3 TABLES



targetcon - Microsoft Azure datafactory7733 - Azure Data Fair +

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 > [datalakeacc121](#) | 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