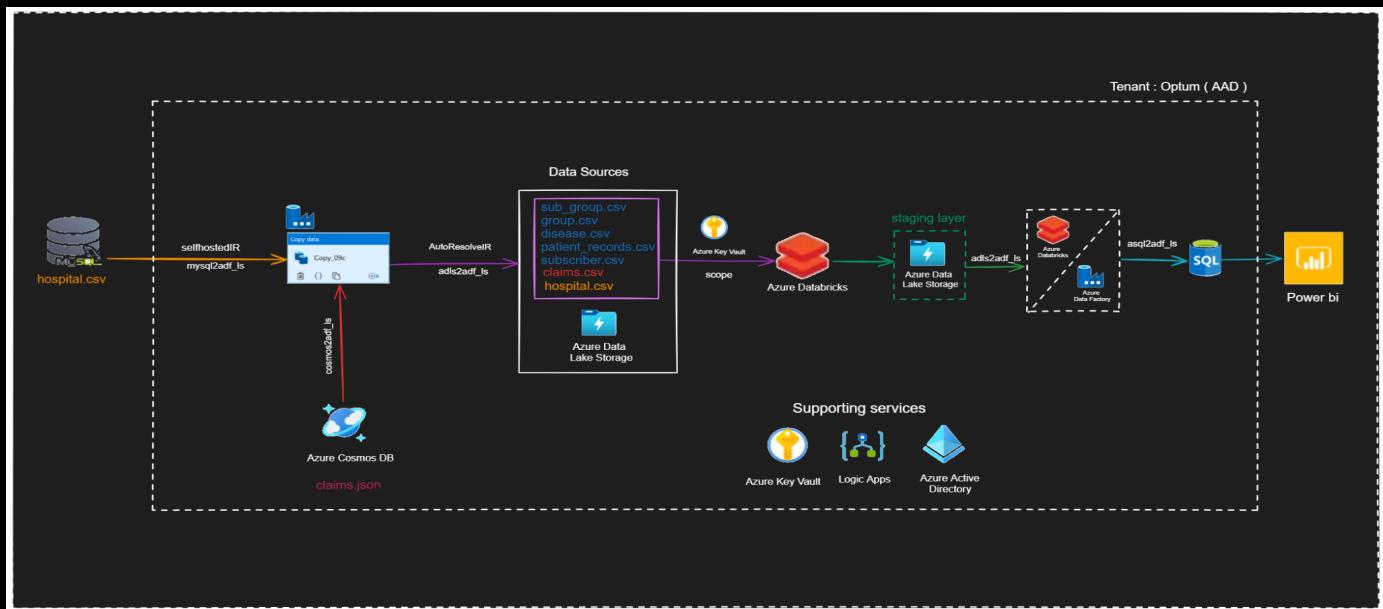


Project Architecture:



Services Required:

1. Azure Data Lake Storage.
2. Azure cosmos DB for mongoDB.
3. Azure SQL Database.
4. Azure Data Factory.
5. Azure Databricks
6. Azure Key Vault.
7. On prem mysql

Azure Data Lake Storage creation:

- The Azure data lake storage creation is shown in the below document.
- https://docs.google.com/document/d/1Gyz7yN9HDF7d_i6wM_i9ph0Z0PNiKPFLgPfRJshsA/edit?usp=sharing
- Parameters used while creating this service.
Storage account name : optumsadllssa
- Create a container called '**optumrawdata**' and click on it.

This screenshot shows the Microsoft Azure Storage Containers page for the 'optumsadllssa' storage account. On the left, there's a sidebar with options like Overview, Activity log, Tags, and Data storage (Containers, File shares, Queues, Tables). The 'Containers' option is selected. In the main pane, there's a table with columns 'Name', 'Last modified', and 'Anonymous'. A message says 'You don't have any containers yet. Click '+ Container' to get started.' To the right, a 'New container' dialog is open with a 'Name' field containing 'optumrawdataset'. There are dropdowns for 'Anonymous access level' (set to 'Private') and 'Advanced' settings. At the bottom right of the dialog is a blue 'Create' button with a green arrow pointing to it.

- Upload csv files (**disease,group,patient_records, subgroup** and **subscriber**) in the **optumrawdata** and click on the '**upload**' button.

This screenshot shows the 'optumrawdataset' container page in Microsoft Azure. The sidebar includes 'Overview', 'Diagnose and solve problems', 'Access Control (IAM)', 'Settings' (Shared access tokens, Manage ACL, Access policy, Properties, Metadata), and a file browser window. The main pane shows a table with columns 'Name', 'Modified', 'Access tier', and 'Arc'. Below the table is a message 'No results'. To the right, an 'Upload blob' dialog is open. It has a 'Drag and drop files here or Browse for files' area with a green arrow pointing to it. There's also a checkbox 'Overwrite if files already exist' and an 'Advanced' section. At the bottom right is a blue 'Upload' button with a green arrow pointing to it. A local file browser window is overlaid on the page, showing a folder structure with files: 'disease', 'group', 'hospital', and 'Patient_records'. The 'Patient_records' file is highlighted with a green box and a number '3' above it. The 'Upload from mobile' button at the bottom of the file browser is also highlighted with a green box and a number '4' above it.

optumrawdataset Container

Search blobs by prefix (case-sensitive)

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
disease.csv	5/26/2024, 3:50:20 PM	Cool (Inferred)		Block blob	1.45 KiB	Available
group.csv	5/26/2024, 3:50:20 PM	Cool (Inferred)		Block blob	4.29 KiB	Available
Patient_records.csv	5/26/2024, 3:50:20 PM	Cool (Inferred)		Block blob	4.99 KiB	Available
subgroup.csv	5/26/2024, 3:50:20 PM	Cool (Inferred)		Block blob	561 B	Available
subscriber.csv	5/26/2024, 3:50:20 PM	Cool (Inferred)		Block blob	11.78 kB	Available

- Create a container called '**optumstagingdata**' in the **optumsadllssa**.

New container

Name * **optumstagingdata**

Anonymous access level **Private (no anonymous access)**

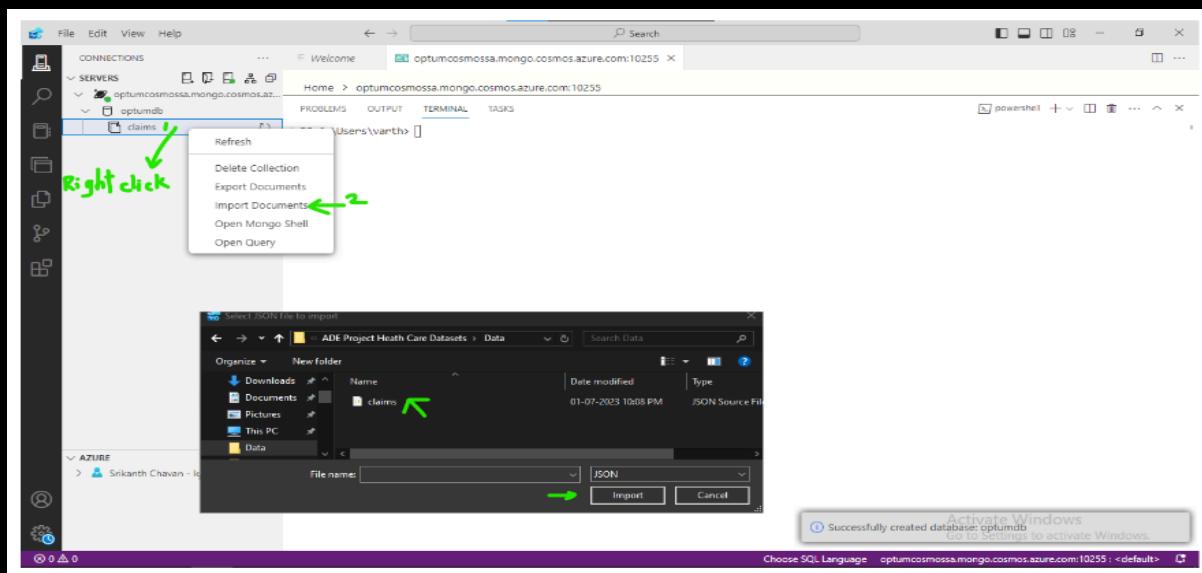
Create

Name	Last modified	Anonymous
optumrawdataset	5/26/2024, 3:45:30 PM	Private
optumstagingdata	5/26/2024, 3:53:07 PM	Private

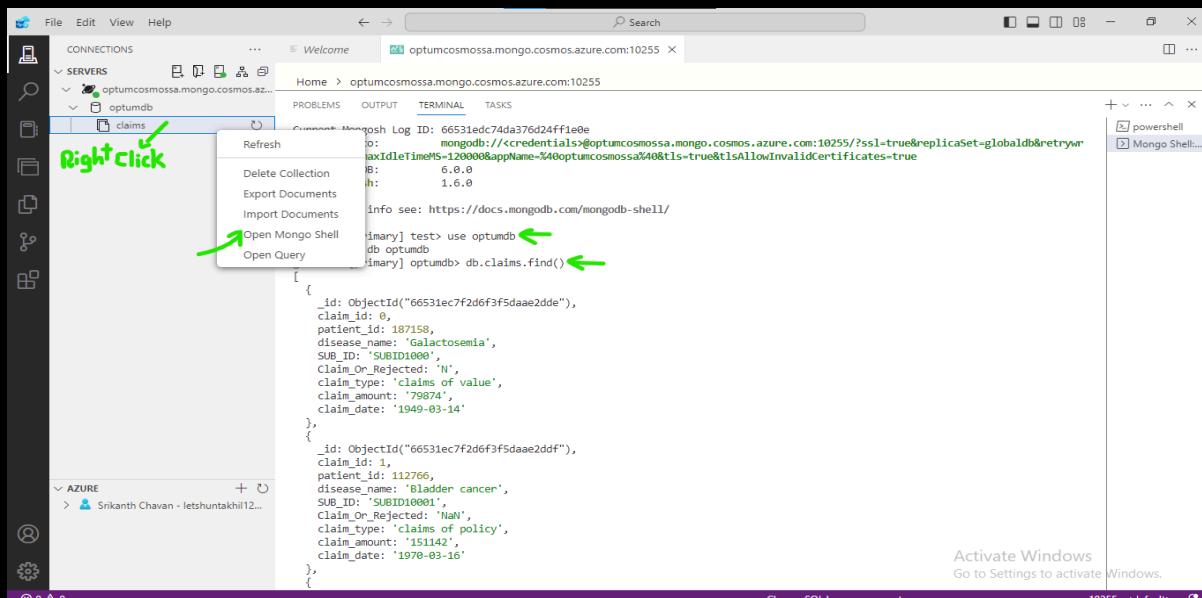
Name	Last modified	Anonymous access level	Lease state
optumrawdataset	5/26/2024, 3:45:30 PM	Private	Available
optumstagingdata	5/26/2024, 3:53:07 PM	Private	Available

Azure Cosmos DB creation:

- The Azure Cosmos Db creation is shown in the document below.
- <https://docs.google.com/document/d/1FCrfSEN3mewFedhw4MkKUyFDYPaG0wQGKi4pk3EoQGY/edit?usp=sharing>
- Parameters used while creating this service.
Azure Cosmos account name : **optumcosmossa**
After creation of Azure cosmosDB service, import '**claims.json**' document into the cosmos db.
- Steps to import document:
 - Create a database with the name : **optumdb**
 - Create a collection with the name : **claims**



- Query the data.



Azure SQL Database Creation :

- The azure sql database creation is shown in the document below.
- https://docs.google.com/document/d/16iB1EsGKHc6-bcgTPSfqkK6BVf3n8_fpbat42uNOvXc/edit?usp=sharing
- Parameters used while creating this service.
- Database name : **optumnosqlDb**
- Server : **optumsserver**
- Server admin login : **optum_admin**

Azure Data factory creation :

- The azure data factory creation is shown in the document below.
- https://docs.google.com/document/d/1lpvA7XumJjbIP0wPWWUlf_d_gdvh6jTn_0a12HWv0jcM/edit?usp=sharing
- Parameters used while creating this service.
Name : **projects-datafactory**

Azure databricks creation :

- The azure databricks creation is shown in the document below.
- <https://docs.google.com/document/d/1s7Vqs4gcZbGMPRk7vEWGFBXIMC5AoaOvCE57qILwRw/edit?usp=sharing>
- Parameters used while creating this service.
- Workspace name : **optum_ws**

Azure keyvault creation :

- The azure keyvault creation is shown in the document below.
- <https://docs.google.com/document/d/1ScQ42B5c5ZuRnFjtLsdJpnV9067pLZkpcDf4WyKjaVw/edit?usp=sharing>
- Parameters used while creating this service.
- Keyvault name : **optums-keyvault**

- Let's bring the hospital data (onprem mysql), Claims(Azure Cosmos db for mongodb) to adls (optumrawdatafactory) via datafactory.
- Assuming hospital data is already there in onprem mysql.
- Create dataset and pipeline for hospitaldata
- Create dataset and pipeline for claimsdata
- Create a new pipeline and execute this hospital and claims data pipelines.

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (cosmosClaimsData_PL, onPremHospitalData_PL, optum_PL), 'Datasets' (cosmosClaimsData, onPremHospitalData), and 'Data flows'. The main workspace displays two 'Execute Pipeline' activities: 'Execute Pipeline1 onPremHospitalData...' and 'Execute Pipeline2 cosmosClaimsData_PL'. The 'Properties' panel on the right is set for the 'optum_PL' pipeline, with 'Name' set to 'optum_PL' and 'Description' empty. Green numbered arrows (1 through 5) point to specific items in the sidebar and workspace: 1 points to 'cosmosClaimsData'; 2 points to 'onPremHospitalData'; 3 points to 'optum_PL' in the Pipelines list; 4 points to 'onPremHospitalData_PL' in the Pipelines list; and 5 points to 'optum_PL' in the Pipelines list.

- Publish all → publish
- Add Trigger → Trigger Now
- Let's do some basic transformations on top of this data in databricks(optum_db).

The screenshot shows the Microsoft Azure Databricks workspace. The left sidebar includes 'New' (Workspace, Recents, Catalog, Workflows, Compute), 'SQL' (SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses), 'Data Engineering' (Job Runs, Data Ingestion, Delta Live Tables), 'Machine Learning' (Playground), and 'Popular' sections for 'Import and transform data', 'Notebook', 'SQL query editor', and 'AutoML'. The main area features 'Get started' with options for 'Import and transform data', 'Notebook', 'SQL query editor', and 'AutoML'. Below it are 'Pick up where you left off' (No recent items) and 'Popular' (No popular items). A message at the bottom right says 'Activate Windows Go to Settings to activate Windows.'

Microsoft Azure | databricks

Search data, notebooks, recents, and more... CTRL + P

optum_ws

New

- Workspace
- Recents
- Catalog
- Workflows
- Compute
- SQL
- SQL Editor
- Queries
- Dashboards
- Alerts
- Query History
- SQL Warehouses
- Data Engineering
- Job Runs
- Data Ingestion
- Delta Live Tables
- Machine Learning
- Playground

Notebook

Git Folder

Data

Compute

Cluster ← 2

SQL Warehouse

SQL

Query

Dashboard

Legacy dashboard

Alert

Data Engineering

Job

DLT Pipeline

Machine Learning

AutoML Experiment

Experiment

Model

Search results for "left off":

- Recent
- Favorites

Popular Provide feedback ↗

No recent items

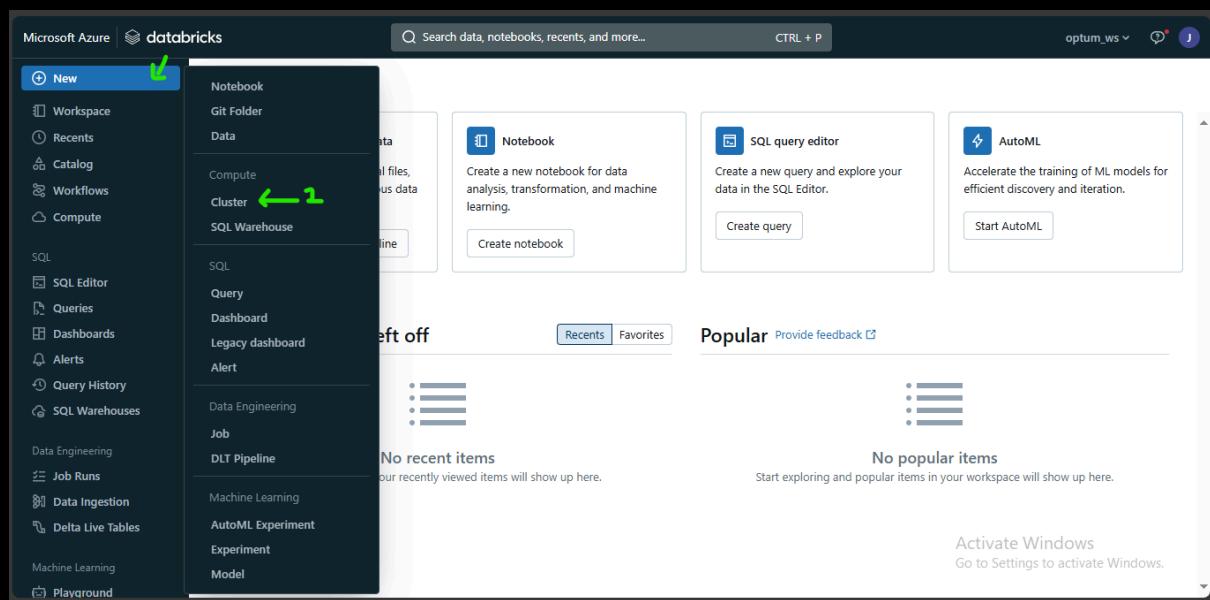
Your recently viewed items will show up here.

No popular items

Start exploring and popular items in your workspace will show up here.

Activate Windows

Go to Settings to activate Windows.



Microsoft Azure | databricks

Search data, notebooks, recents, and more... CTRL + P

optum_ws

New

- Workspace
- Recents
- Catalog
- Workflows
- Compute
- SQL
- SQL Editor
- Queries
- Dashboards
- Alerts
- Query History
- SQL Warehouses
- Data Engineering
- Job Runs
- Data Ingestion
- Delta Live Tables
- Machine Learning
- Playground

Compute > New compute >

Varthy Akhil's Cluster

Performance

Databricks runtime version

Runtime: 13.3 LTS (Scala 2.12, Spark 3.4.1)

Use Photon Acceleration

Worker type

Standard_DS3_v2 14 GB Memory, 4 Cores

Min workers: 2 Max workers: 8

Spot instances

Driver type

Same as worker 14 GB Memory, 4 Cores

Enable autoscaling

Terminate after 120 minutes of inactivity

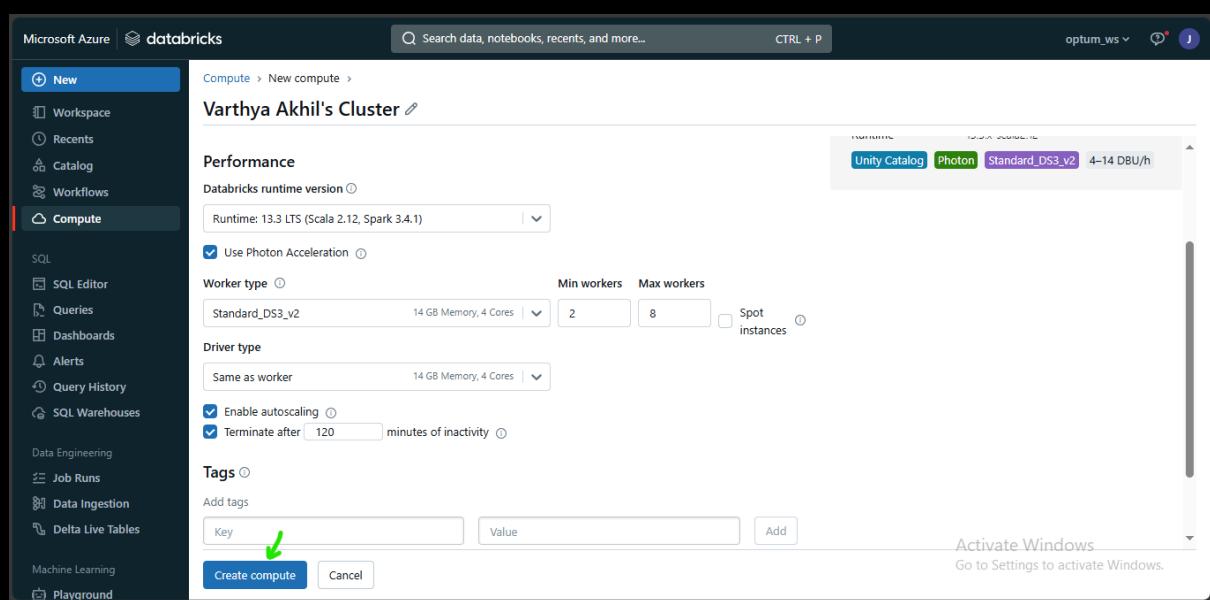
Tags

Add tags

Create compute Cancel

Activate Windows

Go to Settings to activate Windows.



Microsoft Azure | databricks

Search data, notebooks, recents, and more... CTRL + P

optum_ws

New

- Workspace
- Recents
- Catalog
- Workflows
- Compute
- SQL
- SQL Editor
- Queries
- Dashboards
- Alerts
- Query History
- SQL Warehouses
- Data Engineering
- Job Runs
- Data Ingestion
- Delta Live Tables
- Machine Learning
- Playground

Notebook ← 2

Git Folder

Data

Compute

Cluster

SQL Warehouse

SQL

Query

Dashboard

Legacy dashboard

Alert

Data Engineering

Job

DLT Pipeline

Machine Learning

AutoML Experiment

Experiment

Model

Created by Only pinned

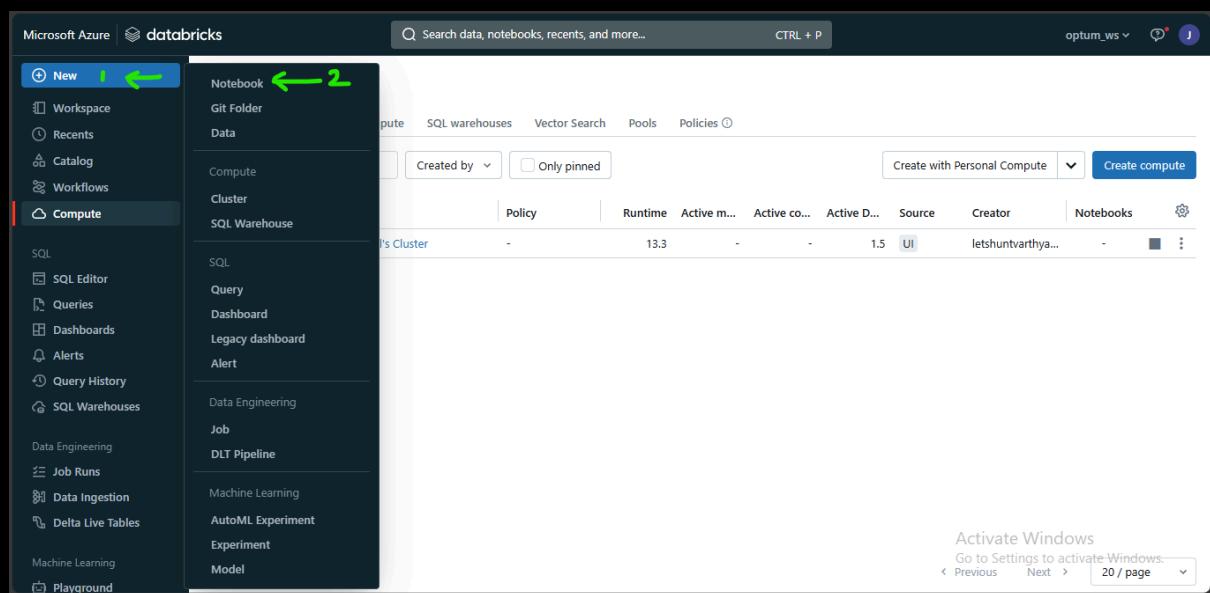
Create with Personal Compute Create compute

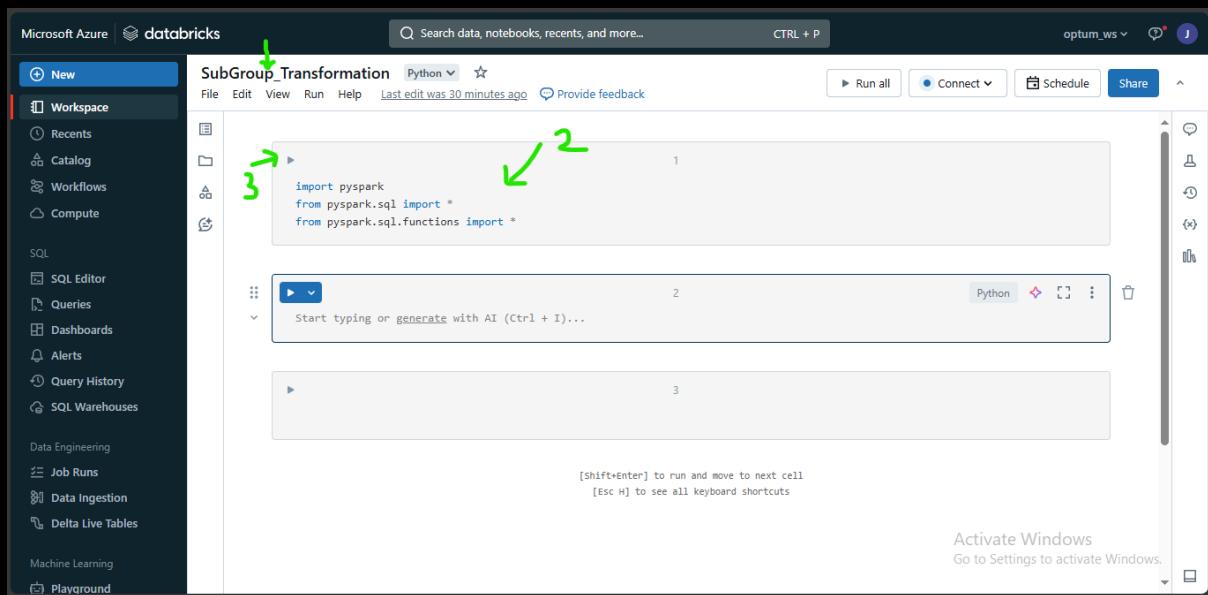
Policy	Runtime	Active m...	Active co...	Active D...	Source	Creator	Notebooks	⋮
Varthy's Cluster	-	13.3	-	-	1.5 UI	letshuntvarthy...	-	⋮

Activate Windows

Go to Settings to activate Windows.

Previous Next 20 / page



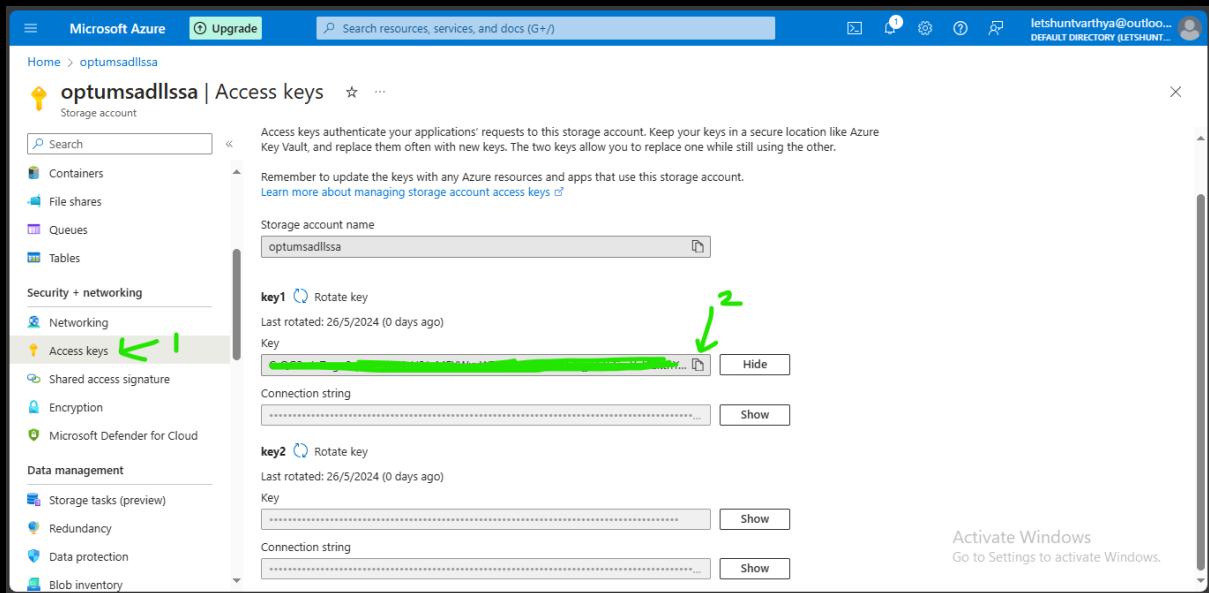


- Boilerplate code to import the data from `adls` to `databricks` via `keyvault` and `access key`

```
spark.conf.set(
    "fs.azure.account.key.<storage-account>.dfs.core.windows.net", dbutils.secrets.get(scope = "<scope>", key = "<storage-account-access-key>"))
```

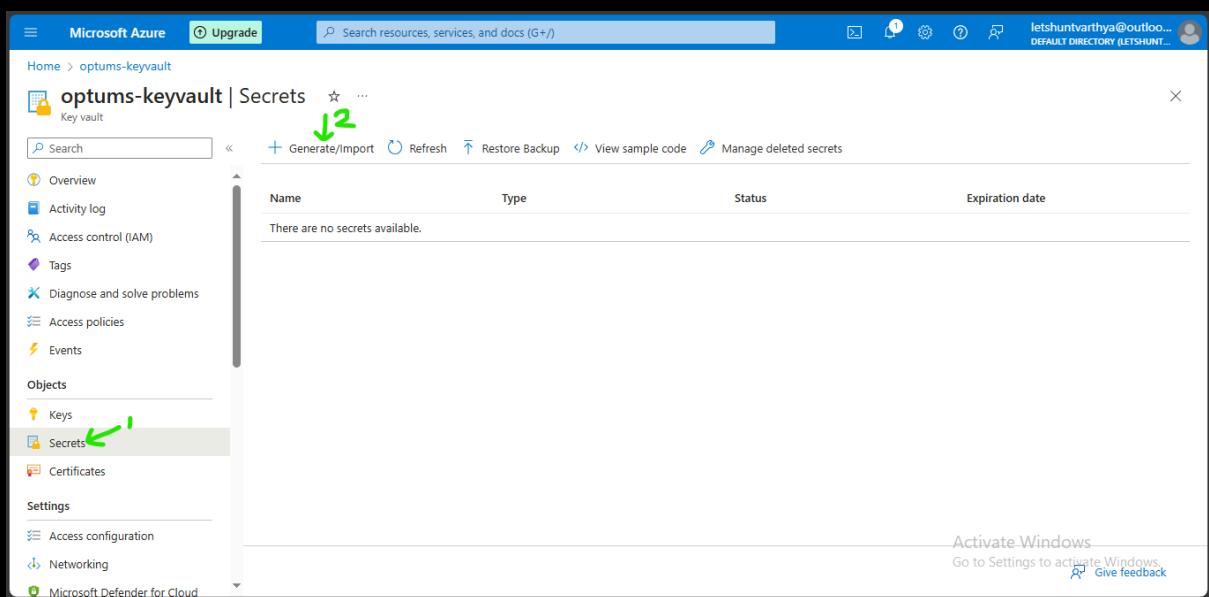
- Replace `<storage-account>` with adls account name (`optumsadllsa`) and `<storage-account-acces-key>` with access key from '`optumsadllsa`' access key
- Here we cannot directly put the access key, due to security reasons. But we can import the access key via keyvault and scope.
- `Optumsadllsa` key is stored in keyvault as a secret.
- In databricks we will make a connection to keyvault by creating a scope.
- Steps to import the data from adls to databricks via keyvault and `access key`.

1. Copy access key from `optumsadllsa`.

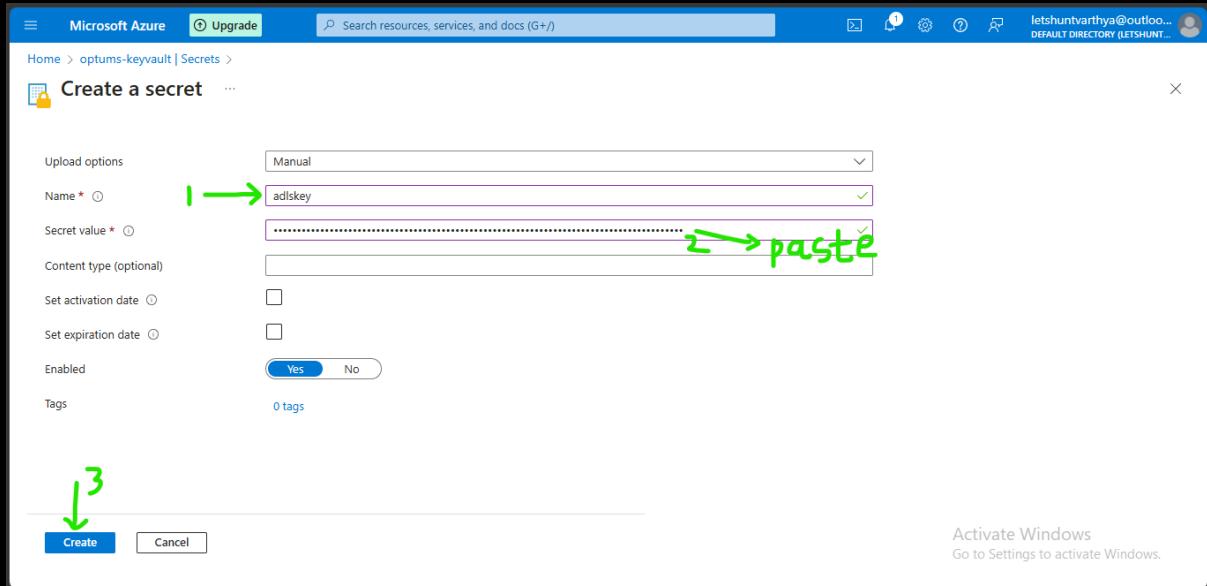


The screenshot shows the 'Access keys' section of the Microsoft Azure Storage account settings for 'optumsadllsa'. The left sidebar lists 'Security + networking' (with 'Access keys' highlighted by a green arrow), 'Data management', and other storage options. The main area displays two sets of access keys: 'key1' and 'key2'. Each key has a 'Rotate key' button, a 'Last rotated' timestamp (26/5/2024), and a 'Key' field containing a long redacted string. A 'Show' button is next to each key field. Below the keys are 'Connection string' fields. A green arrow points to the 'key1' key field, and another green arrow points to the 'key1' connection string field.

2. Create a secret for `optumsadllsa`.

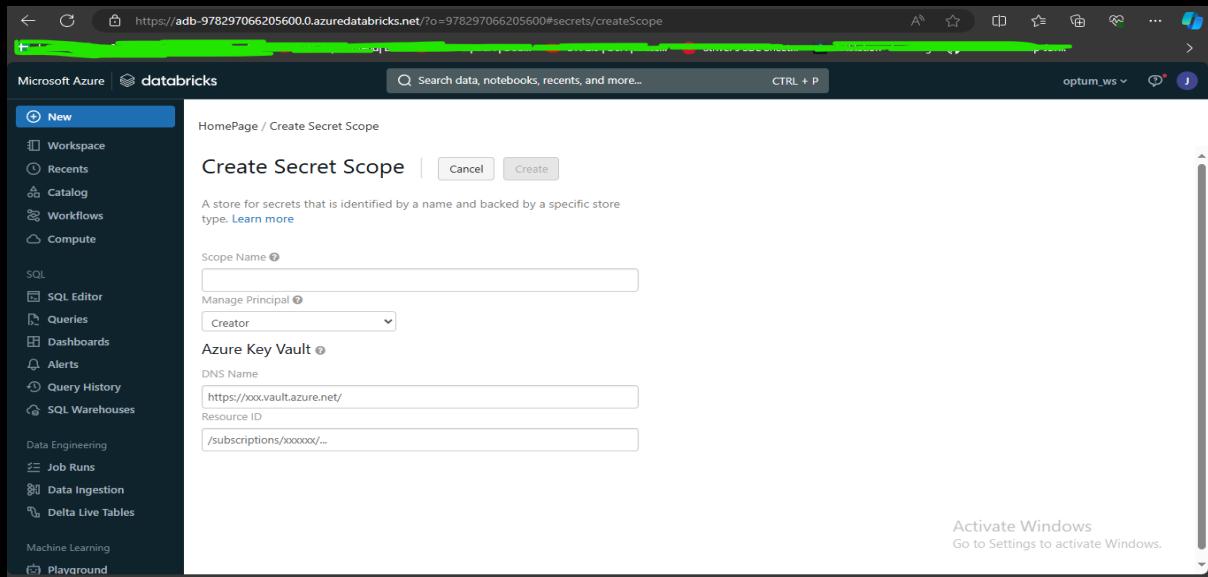


The screenshot shows the 'Secrets' section of the Microsoft Azure Key Vault for 'optums-keyvault'. The left sidebar lists 'Overview', 'Activity log', 'Access control (IAM)', 'Tags', 'Diagnose and solve problems', 'Access policies', 'Events', 'Objects' (with 'Secrets' highlighted by a green arrow), 'Certificates', and 'Settings'. The main area shows a table with columns: Name, Type, Status, and Expiration date. A message at the top states 'There are no secrets available.' A green arrow points to the 'Generate/Import' button at the top of the page.



3. Creating a scope in databricks.

- Copy the url, paste in a new tab and remove the characters after the # and put secrets/createScope (url#secrets/createScope).



The screenshot shows the 'Properties' section of the 'optums-keyvault' key vault in the Azure portal. A green arrow labeled 'copy' points to the 'Vault URI' field, which contains the value `https://optums-keyvault.vault.azure.net/`. A tooltip 'Copied' is visible next to the field.

The screenshot shows a Databricks notebook titled 'SubGroup_Transformation'. The code cell contains the following Python code:

```

# lists scopes created
dbutils.secrets.listScopes()

[SecretScope(name='databricksscope')]

```

The screenshot shows a Databricks notebook titled 'HomePage / Create Secret Scope'. The code cell contains the following Python code:

```

# Boilerplate code to configure
spark.conf.set("fs.azure.account.key.optumsadlissa.dfs.core.windows.net",dbutils.secrets.get(scope="databricksscope",key="adlskey"))

```

A green arrow labeled 'paste' points to the 'Azure Key Vault' field, which contains the value `https://optums-keyvault.vault.azure.net/`.

Microsoft Azure | optums-keyvault

optums-keyvault | Properties

Key vault

Search Save Discard changes Refresh

Settings

- Access configuration
- Networking
- Microsoft Defender for Cloud
- Properties**
- Locks
- Monitoring
- Alerts
- Metrics
- Diagnostic settings
- Logs
- Insights
- Workbooks
- Automation

Name: optums-keyvault
Sku (Pricing tier): Standard
Location: eastus
Vault URI: https://optums-keyvault.vault.azure.net/
Resource ID: /subscriptions/c92e5287-784e-4a20-9e6c-8549947f2ee6/resourceGroups/projects_rg/providers/Microsoft.KeyVault/vaults/optum...
Subscription ID: c92e5287-784e-4a20-9e6c-8549947f2ee6
Subscription Name: Free Trial
Directory ID: 2282df71-75c7-49bc-8440-7dc9452f264c
Directory Name: Default Directory
Soft-delete: Soft delete has been enabled on this key vault
Days to retain deleted vaults: 90
Purge protection: Disable purge protection (allow key vault and objects to be purged during retention period) Enable purge protection (enforce a mandatory retention period for deleted vaults and vault objects)

Copied

Microsoft Azure | databricks

HomePage / Create Secret Scope

Create Secret Scope | Cancel | **Create** ← 2

A store for secrets that is identified by a name and backed by a specific store type. [Learn more](#)

Scope Name: databricksscope
Manage Principal: Creator

Azure Key Vault

DNS Name: https://optums-keyvault.vault.azure.net/
Resource ID: /subscriptions/c92e5287-784e-4a20-9e6c-8549947f2ee6/resourceGroups/projects_rg/providers/Microsoft.KeyVault/vaults/optum...

Paste

Microsoft Azure | databricks

SubGroup_Transformation | Python | **New**

File Edit View Run Help Last edit was 6 minutes ago Provide feedback

Run all | Varthya Akhil's Persona... | Schedule | Share

Recent code runs:

- 3 hours ago (3s)

```
# lists scopes created
dbutils.secrets.listScopes()
```
- 3 hours ago (1s)

```
[SecretScope(name='databricksscope')]
```
- 3 hours ago (2s)

```
dbutils.secrets.list(scope='databricksscope')
```
- 2 hours ago (1s)

```
[SecretMetadata(key='adlskey')]
```
- 2 hours ago (1s)

```
# Boilerplate code to configure
spark.conf.set("fs.azure.account.key.optumsadlissadfs.core.windows.net",dbutils.secrets.get(scope="databricksscope",
key="adlskey"))
```

Activate Windows
Go to Settings to activate Windows.

Microsoft Azure | databricks

SubGroup_Transformation Python

File Edit View Run Help Last edit was 7 minutes ago Provide feedback

Run all Schedule Share

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Playground

Boilerplate code to configure
spark.conf.set("fs.azure.account.key.optumsadlssa.dfs.core.windows.net",dbutils.secrets.get(scope="databricksscope", key="adlskey"))

lists files present in the optumrawdata
display(dbutils.fs.ls("abfss://optumrawdataset@optumsadlssa.dfs.core.windows.net"))

(2) Spark Jobs

Table + New result table: ON

#	path	name	size	modificationTime
1	abfss://optumrawdataset@optumsadlssa.dfs.core.windows.net/Claims.csv	Claims.csv	5766	1716728851000
2	abfss://optumrawdataset@optumsadlssa.dfs.core.windows.net/Patient_records.c...	Patient_records.csv	5110	1716718820000
3	abfss://optumrawdataset@optumsadlssa.dfs.core.windows.net/disease.csv	disease.csv	1489	1716718820000
4	abfss://optumrawdataset@optumsadlssa.dfs.core.windows.net/group.csv	group.csv	4390	1716718820000
5	abfss://optumrawdataset@optumsadlssa.dfs.core.windows.net/hospital.csv	hospital.csv	1528	1716728950000
6	abfss://optumrawdataset@optumsadlssa.dfs.core.windows.net/subgroup.csv	subgroup.csv	561	1716718820000
7	abfss://optumrawdataset@optumsadlssa.dfs.core.windows.net/subscriber.csv	subscriber.csv	12061	1716718820000

Activate Windows Go to Settings to activate Windows

12:39 AM (5s)

loading the subgroup data in variable called data
data = spark.read.csv('abfss://optumrawdataset@optumsadlssa.dfs.core.windows.net/subgroup.csv',header=True)

(1) Spark Jobs

data: pyspark.sql.DataFrame = [subgrp_sk: string, subgrp_name: string ... 2 more fields]

12:40 AM (1s)

display of data
data.show(5,False)

(1) Spark Jobs

subgrp_sk	subgrp_name	monthly_premium	subgrp_id
S101	Deficiency Diseases	3000	[GRP101,GRP105]
S102	Accident	1000	[GRP110,GRP150,GRP136]
S103	Physiology	2000	[GRP122,GRP108,GRP138,GRP148]
S104	Therapy	1500	[GRP103,GRP113,GRP123,GRP133,GRP143]
S105	Allergies	2300	[GRP153,GRP104,GRP114,GRP124]

only showing top 5 rows

Activate Windows Go to Settings to activate Windows

12:41 AM (2s)

Number of rows
data.count()

(2) Spark Jobs

10

12:41 AM (<1s)

Removing duplicate Rows
data = data.dropDuplicates()

(2) Spark Jobs

data: pyspark.sql.DataFrame = [subgrp_sk: string, subgrp_name: string ... 2 more fields]

12:41 AM (1s)

data.count()

(3) Spark Jobs

10

Activate Windows Go to Settings to activate Windows

▶ ✓ 12:43 AM (<1s) 11

```
# splitting a string column into an array of substrings based on a delimiter
data = data.withColumn("subgrp_id",split(data['subgrp_id'],","))
```

▶ [data: pyspark.sql.dataframe.DataFrame = [subgrp_sk: string, subgrp_name: string ... 2 more fields]]

▶ ✓ 12:45 AM (<1s) 12

```
data.show(5)
```

▶ (2) Spark Jobs

subgrp_sk	subgrp_name	monthly_premium	subgrp_id
S110	Viral	1000	GRP143
S110	Viral	1000	GRP147
S110	Viral	1000	GRP126
S107	Cancer	3200	GRP151
S107	Cancer	3200	GRP131

only showing top 5 rows

▶ ✓ 12:44 AM (<1s) 13

```
# "explode" or "flatten" arrays or maps into separate rows.
data = data.withColumn("subgrp_id",explode(data['subgrp_id']))
```

▶ [data: pyspark.sql.dataframe.DataFrame = [subgrp_sk: string, subgrp_name: string ... 2 more fields]]

▶ ✓ 12:45 AM (<1s) 14

```
data.show(5)
```

▶ (2) Spark Jobs

subgrp_sk	subgrp_name	monthly_premium	subgrp_id
S110	Viral	1000	GRP143
S110	Viral	1000	GRP147
S110	Viral	1000	GRP126
S107	Cancer	3200	GRP151
S107	Cancer	3200	GRP131

only showing top 5 rows

The image consists of three vertically stacked screenshots of a Jupyter Notebook interface, each showing a code cell, its execution time, a numerical ID, and the resulting Spark DataFrame output.

- Screenshot 15:** Shows code to display null values across all columns. The output shows a DataFrame with four columns: subgrp_sk, subgrp_name, monthly_premium, and subgrp_id, all containing zero values.
- Screenshot 16:** Shows code to find duplicates. The output shows a DataFrame with four columns: subgrp_sk, subgrp_name, monthly_premium, and count, where count is greater than 1.
- Screenshot 17:** Shows code to write data to Azure Blob Storage. The output is a single word "True".

- Transformation of **Hospital data** is shown in the document below.
 - Create a new notebook (**Hospital_Transformation**).

https://drive.google.com/file/d/1WFp2Bcw5Xozcz2ZRts_PXXJfuX8LCfFJ/view?usp=sharing
- Transformation of **group data** is shown in the document below.
 - Create a new notebook (**group_Transformation**).

https://drive.google.com/file/d/1B2Nv_eioFfPuUUtDX3d9Hsub9392DCVL/view?usp=sharing
- Transformation of **disease data** is shown in the document below.
 - Create a new notebook (**Disease_Transformation**).

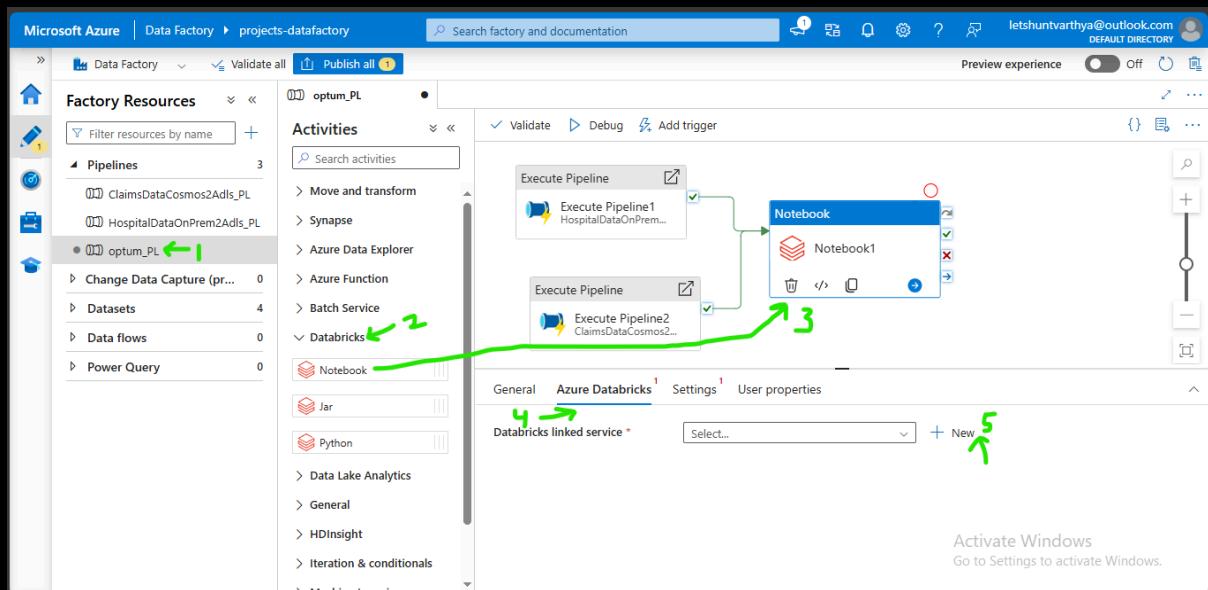
https://drive.google.com/file/d/1B2Nv_eioFfPuUUtDX3d9Hsub9392DCVL/view?usp=sharing
- Transformation of **patient data** is shown in the document below.
 - Create a new notebook (**Patient_Transformation**).

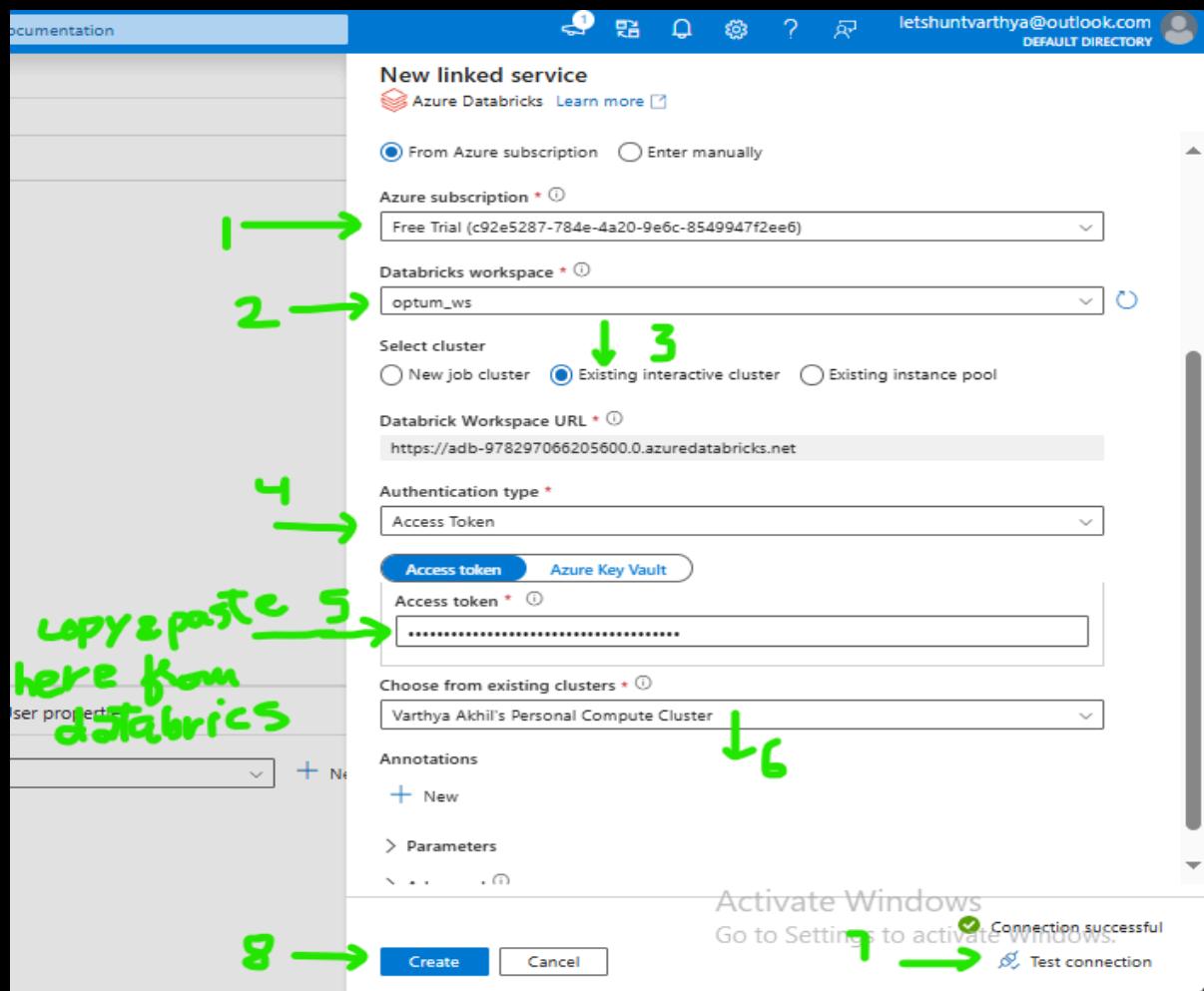
<https://drive.google.com/file/d/1jEHawUK1OruV1ZEamZpxxtj4vliRhYUC/view?usp=sharing>

- Transformation of **subscriber data** is shown in the document below.
 - Create a new notebook (**Subscriber_Transformation**).

<https://drive.google.com/file/d/1reQ7WyybUi5TUqRrJ49SqCGHqKul2h2O/view?usp=sharing>
- Transformation of **claims data** is shown in the document below.
 - Create a new notebook (**Claims_Transformation**).

<https://drive.google.com/file/d/1tJkE5jdtWFZ4osR4A1CDNaOu57rv1s/view?usp=sharing>
- We have executed this in the notebook itself, instead we can do it in the datafactory as well.
- Executing **notebooks** in the **datafactory** via access tokens.
- Drag and drop **notebook activity** in the **optum_PL** field.





- While creating linked service access token is copied from databricks and steps are shown below.
- Connecting datafactory with databricks via access token.

Microsoft Azure | databricks

Search data, notebooks, recents, and more... CTRL + P

optum_ws

Compute

All-purpose compute Job compute SQL warehouses Vector Search Pools Policies

Filter compute you have... Created by Only pinned Create with Personal Co...

State	Name	Policy	Runtime	Active m...	Active co...	Active D...	Source	Creator
	Varthya Akhil's Personal Comput...	Personal Comp	14.3 ML	14 GB	4 cores	0.75 UI	letshun...	letshun...

Settings

Azure Portal
Privacy Policy
What's New
Previews
Log out

Activate Windows
Go to Settings to activate Windows.

copy & paste here from databricks

The screenshot shows the Microsoft Azure Databricks interface. On the left, there's a sidebar with various navigation options like Workspace, Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, Delta Live Tables, Machine Learning, and Playground. A green arrow labeled '1' points to the 'Developer' option under the User section. The main area is titled 'Developer' and contains sections for 'Access tokens', 'SQL query snippets', and 'Editor settings'. Under 'Access tokens', there's a 'Manage' button with a green arrow labeled '2' pointing to it.

This screenshot shows the 'Access tokens' sub-page. It includes fields for 'Comment', 'Creation' (with a dropdown arrow), and 'Expiration'. Below these fields, it says 'No tokens exist.' A green arrow labeled '1' points to the 'Generate new token' button at the top left. Another green arrow labeled '2' points to the 'On' toggle switch for 'Notebook Notifications'.

This screenshot shows the 'Generate new token' dialog box. It has fields for 'Comment' (containing 'to connect data factory') and 'Lifetime (days)' (set to 90). At the bottom are 'Cancel' and 'Generate' buttons. A green arrow labeled '1' points to the 'Comment' input field, and another green arrow labeled '2' points to the 'Generate' button.

Generate New Token

Your token has been created successfully.

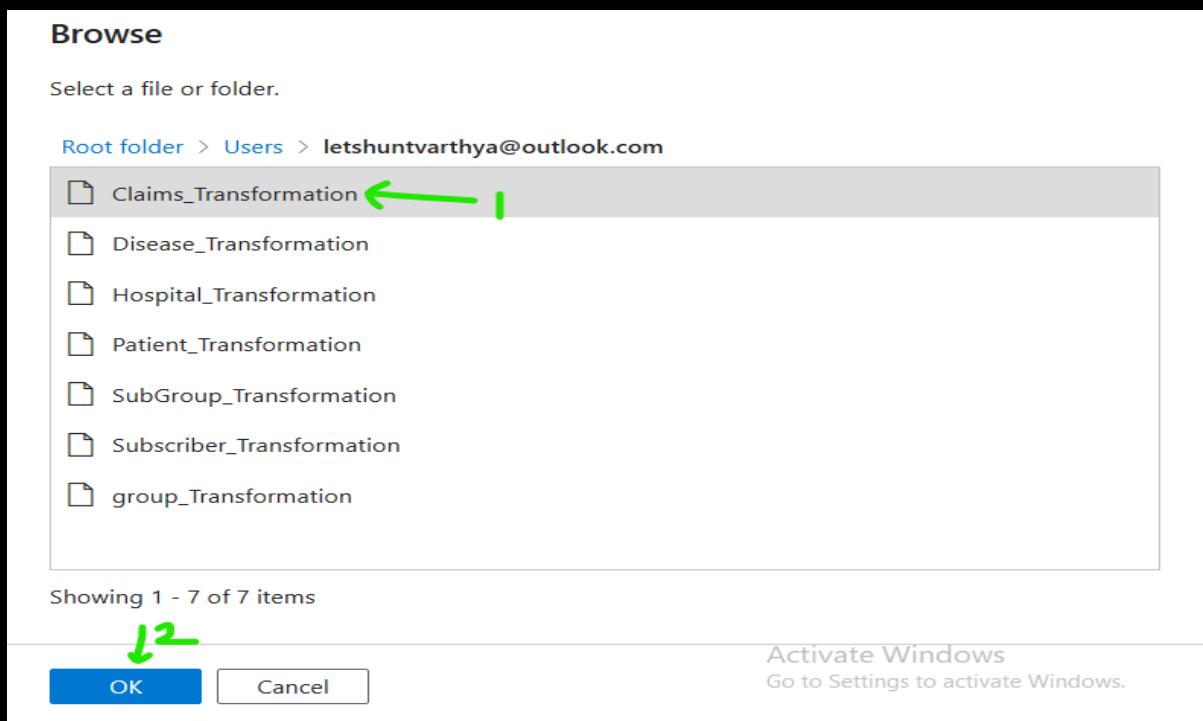


⚠ Make sure to copy the token now. You won't be able to see it again.

2 → Done

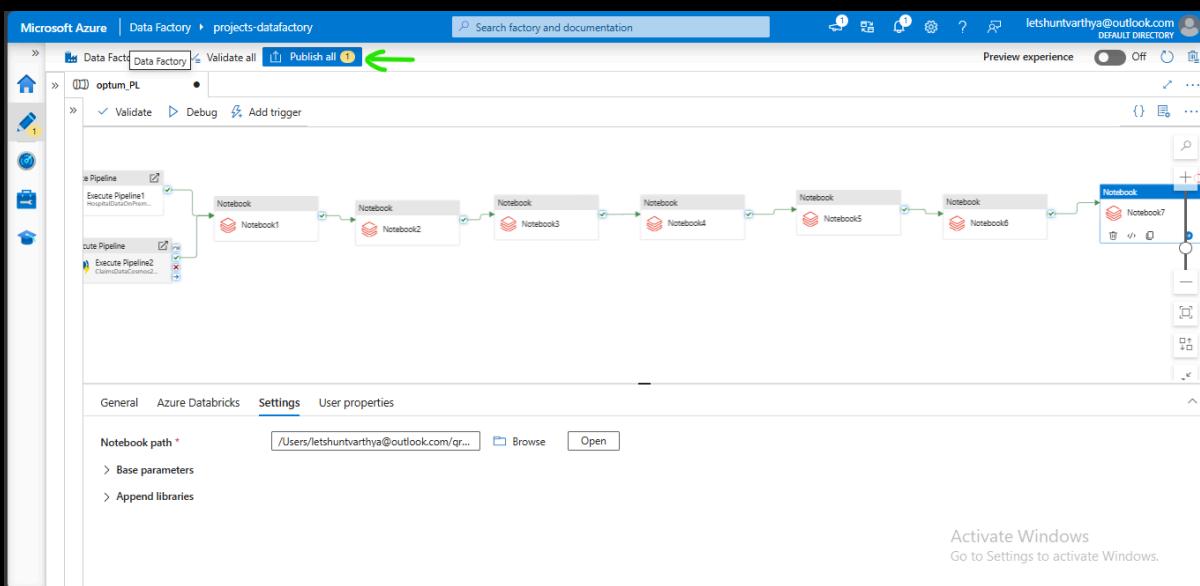
A screenshot of the Microsoft Azure Data Factory pipeline editor. It shows a pipeline named 'Pipeline1' with two notebook activities: 'Notebook1' and 'Notebook2'. The pipeline is currently in 'Validate' mode. On the left, there's a sidebar for 'Activities' including 'Move and transform', 'Synapse', 'Azure Data Explorer', 'Azure Function', 'Batch Service', and 'Databricks'. Under 'Databricks', there are options for 'Notebook', 'Jar', and 'Python'. Below these are sections for 'Data Lake Analytics', 'General', 'HDInsight', 'Iteration & conditionals', 'Machine Learning', and 'Power Query'. At the bottom of the pipeline editor, there are tabs for 'General', 'Azure Databricks', and 'Settings'. A green arrow labeled '1' points from the top towards the pipeline. A green arrow labeled '2' points from the bottom towards the 'Settings' tab. A green arrow labeled '3' points from the right towards the 'Users' folder in the 'Browse' sidebar.

A screenshot of a 'Browse' dialog box. It asks 'Select a file or folder.' and shows a tree view under 'Root folder > Users'. Two items are listed: 'letshunvarthya@outlook.com' and 'letshunvarthya_outlook.com#ext#@letshunvarthyaoutlook.onmicrosoft.com'. A green arrow labeled '1' points from the top towards the 'Root folder' path. A green arrow labeled '2' points from the bottom towards the first item in the list.



- Do this process for other notebooks (**Disease_Transformation**, **Hospital_Transformation**, **Patient_Transformation**, **SubGroup_Transformation**, **Subscriber_Transformations** and **Group_Transformations**) as well.

 1. Drag and drop notebook activity to the field.
 2. Select the linked service that we created just now.
 3. Select the notebook



Microsoft Azure | Data Factory > projects-datafactory

Data Factory > Validate all Publishing 1

Validate | Debug | Add trigger

Pipeline

```
graph LR; A[Execute Pipeline1<br/>HospitalDataSources1] --> B[Notebook<br/>Notebook1]; B --> C[Notebook<br/>Notebook2]; C --> D[Notebook<br/>Notebook3]; D --> E[Notebook<br/>Notebook4]; E --> F[Notebook<br/>Notebook5]; F --> G[Notebook<br/>Notebook6]; G --> H[Notebook<br/>Notebook7]
```

Publish all

You are about to publish all pending changes to the live environment. Learn more

Pending changes (1)

NAME	CHANGE	EXISTING
optum_PL	(Edited)	optum_PL

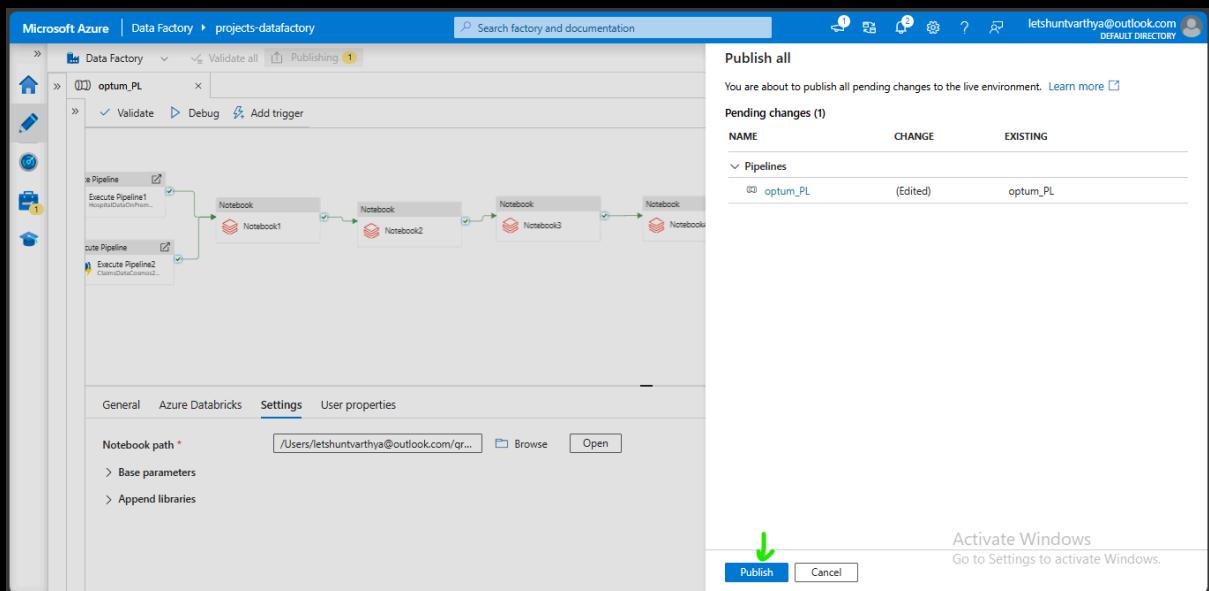
General Azure Databricks Settings User properties

Notebook path * /Users/letshuntvarthy@outlook.com/qr... Browse Open

Base parameters Append libraries

Activate Windows Go to Settings to activate Windows.

Publish Cancel



Microsoft Azure | Data Factory > projects-datafactory

Data Factory > Validate all Publishing 1

Validate | Debug | Add trigger

Trigger now New/Edit

Pipeline

```
graph LR; A[Execute Pipeline1<br/>HospitalDataSources1] --> B[Notebook<br/>Notebook1]; B --> C[Notebook<br/>Notebook2]; C --> D[Notebook<br/>Notebook3]; D --> E[Notebook<br/>Notebook4]; E --> F[Notebook<br/>Notebook5]; F --> G[Notebook<br/>Notebook6]; G --> H[Notebook<br/>Notebook7]
```

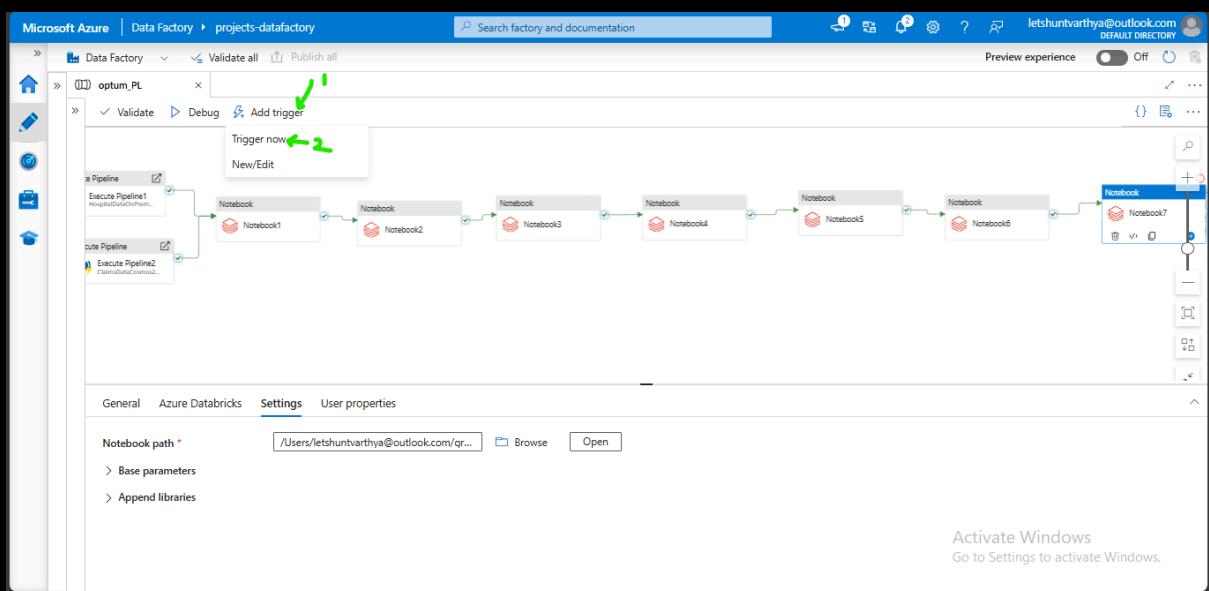
Preview experience Off

General Azure Databricks Settings User properties

Notebook path * /Users/letshuntvarthy@outlook.com/qr... Browse Open

Base parameters Append libraries

Activate Windows Go to Settings to activate Windows.



Microsoft Azure | Data Factory > projects-datafactory

Data Factory > Validate all Publishing 1

Validate | Debug | Add trigger

Pipeline

```
graph LR; A[Execute Pipeline1<br/>HospitalDataSources1] --> B[Notebook<br/>Notebook1]; B --> C[Notebook<br/>Notebook2]; C --> D[Notebook<br/>Notebook3]; D --> E[Notebook<br/>Notebook4]; E --> F[Notebook<br/>Notebook5]; F --> G[Notebook<br/>Notebook6]; G --> H[Notebook<br/>Notebook7]
```

Pipeline run

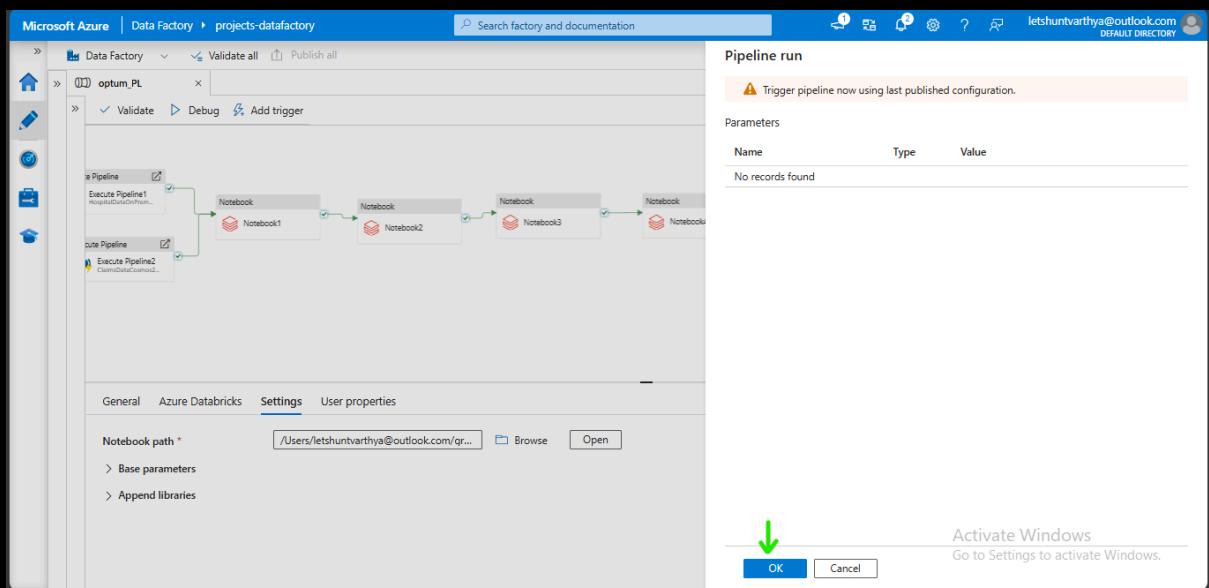
Trigger pipeline now using last published configuration.

Parameters

Name	Type	Value
No records found		

OK Cancel

Activate Windows Go to Settings to activate Windows.



optumstagingdata

Authentication method: Access key ([Switch to Microsoft Entra user account](#))
Location: optumstagingdata

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
claimdatastaging.csv	5/27/2024, 4:05:36 AM	Cool (Inferred)		Block blob	4.71 KiB	Available
diseasedatastaging...	5/27/2024, 4:06:05 AM	Cool (Inferred)		Block blob	1.39 KiB	Available
groupdatastaging.csv	5/27/2024, 4:08:49 AM	Cool (Inferred)		Block blob	4.23 KiB	Available
hospitaldatastaging...	5/27/2024, 4:06:38 AM	Cool (Inferred)		Block blob	1.26 KiB	Available
patientdatastaging...	5/27/2024, 4:06:58 AM	Cool (Inferred)		Block blob	4.58 KiB	Available
subgroupstagingda...	5/27/2024, 4:07:44 AM	Cool (Inferred)		Block blob	1.09 KiB	Available
subscriberdatastagi...	5/27/2024, 4:08:20 AM	Cool (Inferred)		Block blob	9.7 KiB	Available

Activate Windows
Go to Settings to activate Windows.

- The basic transformed data has been brought to **optumstagingdata**.
- Now let's do Final transformations using datafactory.
- Create datasets for this **stagingdata**.

Connection Schema Parameters

Linked service * adf2adls_ls

File path optumstagingdata / claimdatastaging.csv

Compression type Select...

Column delimiter Comma (,)

Row delimiter Default (\r\n, or \n)

Encoding Default(UTF-8)

Quote character Double quote ("")

Escape character Backslash (\)

Activate Windows
Go to Settings to activate Windows.

Connection Schema Parameters

Linked service * adf2adls_ls

File path optumstagingdata / diseasedatastaging.csv

Compression type Select...

Column delimiter Comma (,)

Row delimiter Default (\r\n, or \n)

Encoding Default(UTF-8)

Quote character Double quote ("")

Escape character Backslash (\)

Activate Windows
Go to Settings to activate Windows.

Connection Schema Parameters

Linked service * Test connection Edit + New Learn more

File path / / Browse Preview data Detect

Compression type

Column delimiter

Row delimiter

Encoding

Quote character

Escape character

Activate Windows
Go to Settings to activate Windows.

Connection Schema Parameters

Linked service * Test connection Edit + New Learn more

File path / / Browse Preview data Detect

Compression type

Column delimiter

Row delimiter

Encoding

Quote character

Escape character

Activate Windows
Go to Settings to activate Windows.

Connection Schema Parameters

Linked service * Test connection Edit + New Learn more

File path / / Browse Preview data Detect

Compression type

Column delimiter

Row delimiter

Encoding

Quote character

Escape character

Activate Windows
Go to Settings to activate Windows.

Connection Schema Parameters

Linked service * Test connection Edit + New Learn more

File path / / Browse Preview data Detect

Compression type

Column delimiter

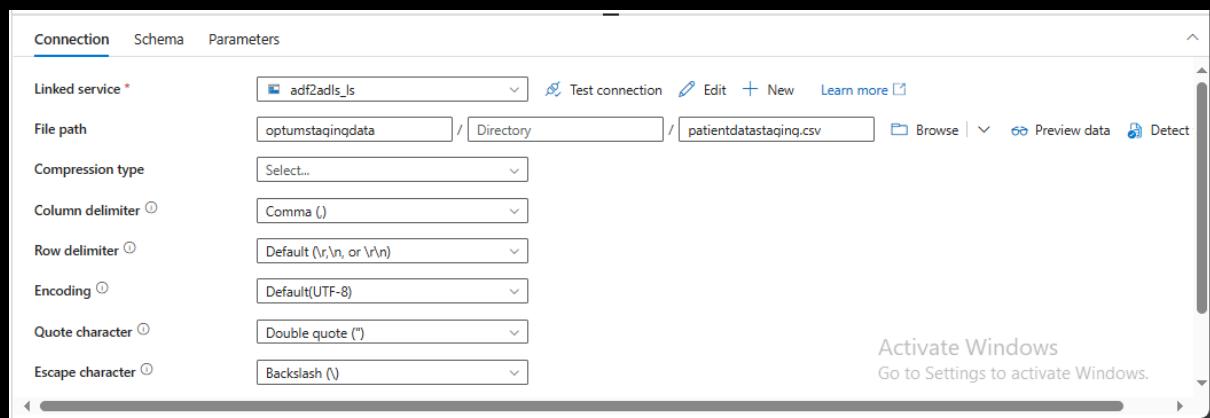
Row delimiter

Encoding

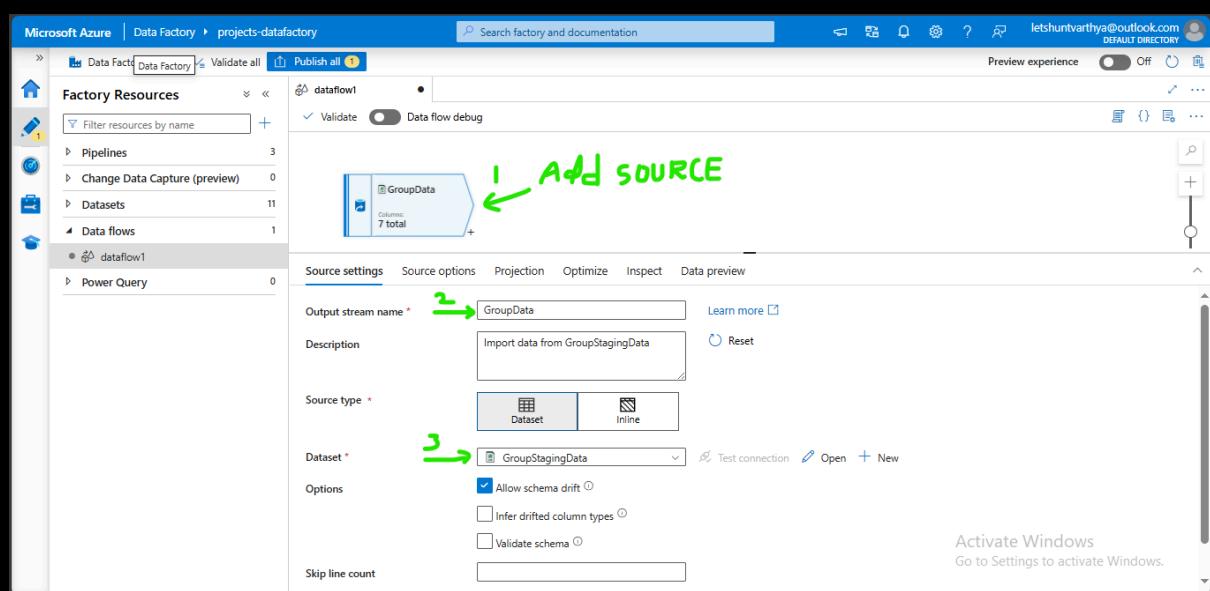
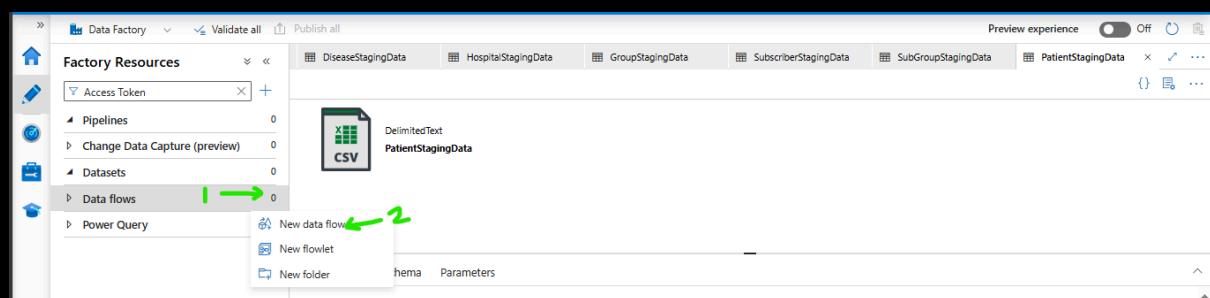
Quote character

Escape character

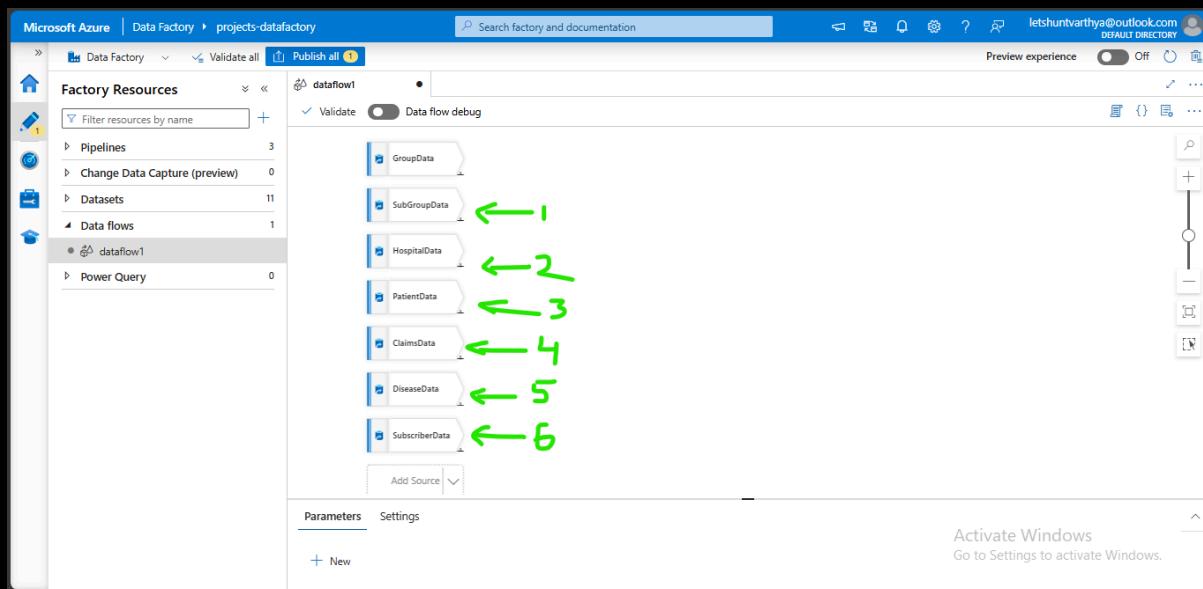
Activate Windows
Go to Settings to activate Windows.



Create **dataflow** to do **transformations**.



- Add sources like `subgroup_data`, `hospital_data`, `patient_data`, `claims_data`, `disease_data` and `subscriber_data`.



- Add `select` after `group_data` and rearrange properly.

The screenshot shows the Microsoft Azure Data Factory Data Flow blade. The 'selectedGroupData' component is selected. The 'Output stream name' field is set to 'selectedGroupData'. The 'Description' section shows 'Renaming GroupData to selectedGroupData with columns grp_id, grp_name, grp_type, country'. The 'Incoming stream' dropdown is set to 'GroupData'. The 'Input columns' section shows seven mappings:

GroupData's column	Name as
abc.grp_id	grp_id
abc.grp_name	grp_name
abc.grp_type	grp_type
abc.country	country
abc.city	city
abc.zip_code	zip_code
abc.premium_written	premium_written

A green arrow points from the word 'Select' to the 'Output stream name' field.

- Add `select` after `subgroup_data`, rename `subgrp_sk` to `subgrp_id`, `subgrp_id` to `grp_id` and rearrange properly.

- Joining `hospital_data` with `patient_data`.

Data preview																		Total 70
Select settings		Optimize		Inspect		Data preview		Update 0		Delete 0		Upsert 0		Lookup 0		Error 0		Total 70
abc ↑	patient_g... ↑	abc ↑	patient_p... ↑	abc ↑	Patient_a... ↑	abc ↑	hospital_... ↑	abc ↑	state ↑	abc ↑	country ↑	abc ↑	disease_n... ↑	abc ↑	Hospital_... ↑	abc ↑	patient_c... ↑	abc ↑
Male	+91 6345...		60		Bengaluru		Karnataka		India		Anthrax		Apollo H...		Bhalswa J...			
Male	+91 4354...		56		Chennai		Tamil Nadu		India		Sunbathing		Apollo H...		Panvel			
Female	+91 5176...		98		Delhi		UT		India		Asthma		Sir Ganqa...		Morbi			
Female	+91 7434...		69		Bengaluru		Karnataka		India		Phenylket...		Manipal ...		Viiayawada			
Female	+91 5674...		72		Gurgaon		Haryana		India		Lung can...		Medanta ...		Bareilly			
Male	+91 9447...		97		Bengaluru		Karnataka		India		Lymphed...		Apollo H...		Saharsa			
Female	+91 0106...		54		Bengaluru		Karnataka		India		Suicide		Manipal ...		Bihar Sharif			
Female	+91 0537...		90		Hyderabad		Telangana		India		Smallpox		Yashoda ...		Meerut			
Male	+91 7316...		32		Hyderabad		Telangana		India		Cystic fibr...		Yashoda ...		Ambala			
Male	+91 0695...		27		Mumbai		Maharash...		India		Pollen ill...		Lilavati H...		Chinsurah			
Female	+91 5R51		75		Mumbai		Maharash		India		Glaucoma		Jaslok Ho...		Krishi			

- Add `Select` after this `join` activity.
- Output stream name : `patienthospitalselected`

- Delete HospitalData@Hospital_id column
- Delete HospitalData@Hospital_id column
- Rename city (hopspitaldata) to hospital_city
- Rename city (patientdata) to patient_city
- Rearrange columns properly

Select settings Optimize Inspect Data preview ●

Input columns Name as

patienthospitaldata's column	Name as
abc Patient_id	Patient_id
abc Patient_name	Patient_name
abc patient_gender	patient_gender
abc patient_phone	patient_phone
abc Patient_age	Patient_age
abc HospitalData@city	hospitalCity
abc state	state
abc country	country
abc disease_name	disease_name
abc Hospital_name	Hospital_name
abc PatientData@city	patientCity
abc HospitalData@Hospital_id	Hospital_id
abc PatientData@hospital_id	hospital_id

- Joining patienthospitalselected with claims_data

Microsoft Azure | Data Factory > projects-datafactory

Factory Resources

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

dataflow1

Validate Data flow debug Debug Settings

Join settings Optimize Inspect Data preview ●

Output stream name * 1 PatientHospitalClaims

Description Inner join on 'patienthospitalselected' and 'ClaimsData'

Left stream * 2 patienthospitalselected

Right stream * 3 ClaimsData

Join type * 4 Inner

Use fuzzy matching

Join conditions * 5

Left: patienthospitalselected's column Right: ClaimsData's column

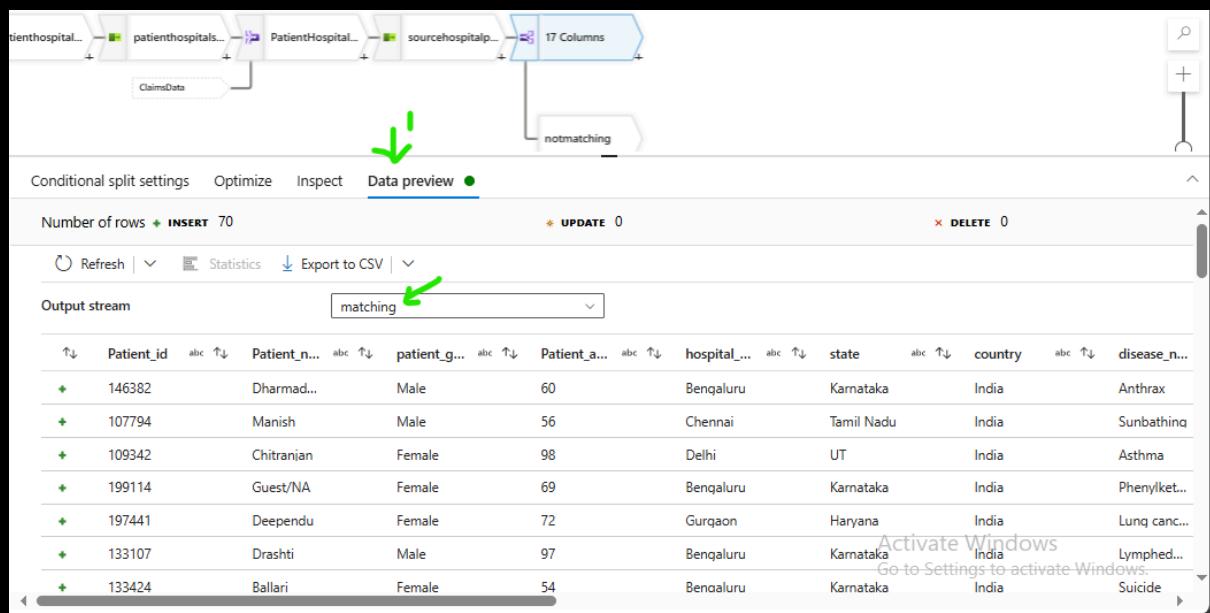
abc Patient_id == abc patient_id

- Add select after this join activity (PatientHospitalClaims).
- Output stream name : sourcehospitalpatientclaims
- Delete patientdata@patient_phone column
- Delete ClaimsData@patient_id column
- Checking hospitalselected@disease_name and patienthospitalselected@disease_name are by renaming it to diseaseName_h and diseaseName_p respectively.

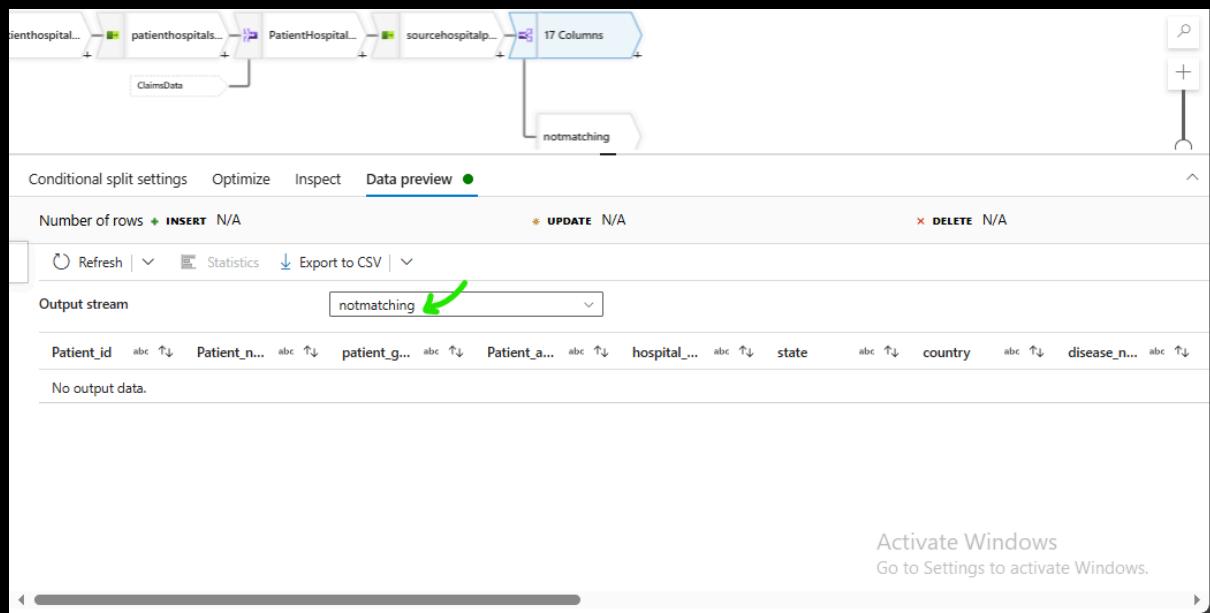
Screenshot of the Microsoft Power Query Editor showing a mapping step. A green arrow labeled "Select" points to the "PatientHospitalClaims" column. Another green arrow labeled "3" points to the "disease_name_h" column. Green arrows labeled "4" and "5" point to the "patient_id" and "patient_phone" columns respectively.

- Let's check whether `diseasename_h` and `diseasename_p` are the same by adding conditional split activity after this `sourcehospitalpatientclaims`.

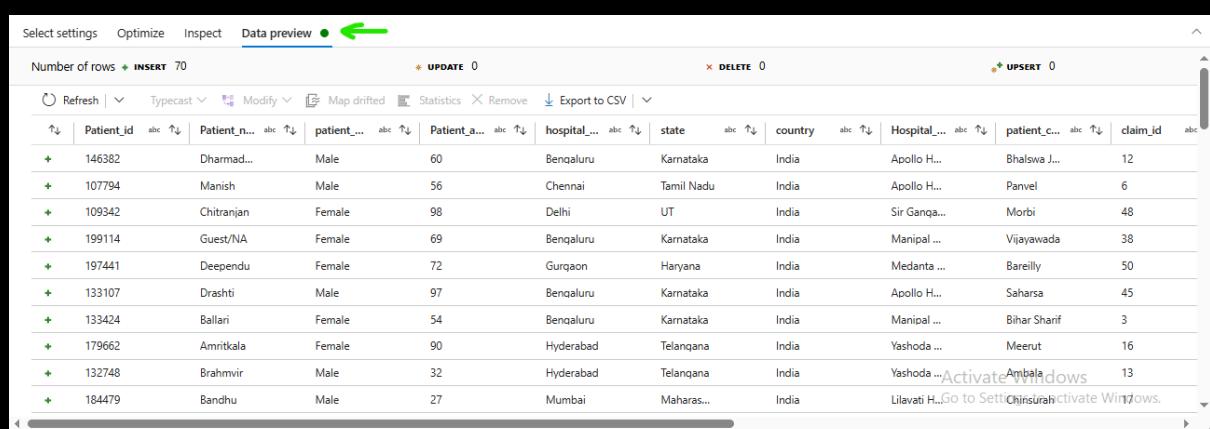
Screenshot of the Microsoft Azure Data Factory portal showing a data flow named "dataflow1". A green arrow labeled "c-Split" points to the conditional split activity. Green arrows labeled "2", "3", "4", "5", and "6" point to the "matching", "notmatching", "disease_name_h=disease_name_c", "disease_name_h!=disease_name_c", and "Stream names" fields respectively.



- From below image we can see that there is no **non-matching** records



- Delete conditional split activity.
- Delete **disease_name_h** column and rename **disease_name_c** to **disease_name**.



- Joining `subGroupData` with `disease_data`.
- Output stream name : `subgroupdisease`

Join settings

Output stream name * 2 → subgroupdisease

Description Inner join on 'selectSubGroupData' and 'DiseaseData'

Left stream * selectSubGroupData

Right stream * DiseaseData 3

Join type * Full outer Inner Left outer Right outer Custom (cross)

Use fuzzy matching

Join conditions * Left: selectSubGroupData's column 4 Right: DiseaseData's column 5

abc subgrp_id == abc subgrp_id

Activate Windows Go to Settings to activate Windows.

- Add `select` after this `join` and delete `DiseaseData@subgrp_id` column

Select settings

Output stream name * 2 → subgroupdiselected

Description Renaming subgroupdisease to subgroupdiselected with columns grp_id, subgrp_id, subgrp_name.

Incoming stream * subgroupdisease

Options Skip duplicate input columns Skip duplicate output columns

Input columns * Auto mapping + Add mapping Delete

7 mappings: 1 column(s) from the inputs left unmapped

subgroupdisease's column	Name as
abc grp_id	grp_id
abc selectSubGroupData@subgrp_id	subgrp_id
abc subgrp_name	subgrp_name
abc disease_id	disease_id
abc disease_name	disease_name
abc monthly_premium	monthly_premium
abc DiseaseData@subgrp_id	subgrp_id

Activate Windows Go to Settings to activate Windows.

Number of rows **INSERT** 100 **UPDATE** 0 **DELETE** 0 **UPSERT** 0 **LOOKUP** 0 **ERROR** 0

Refresh Typecast Modify Map drifted Statistics Remove Export to CSV

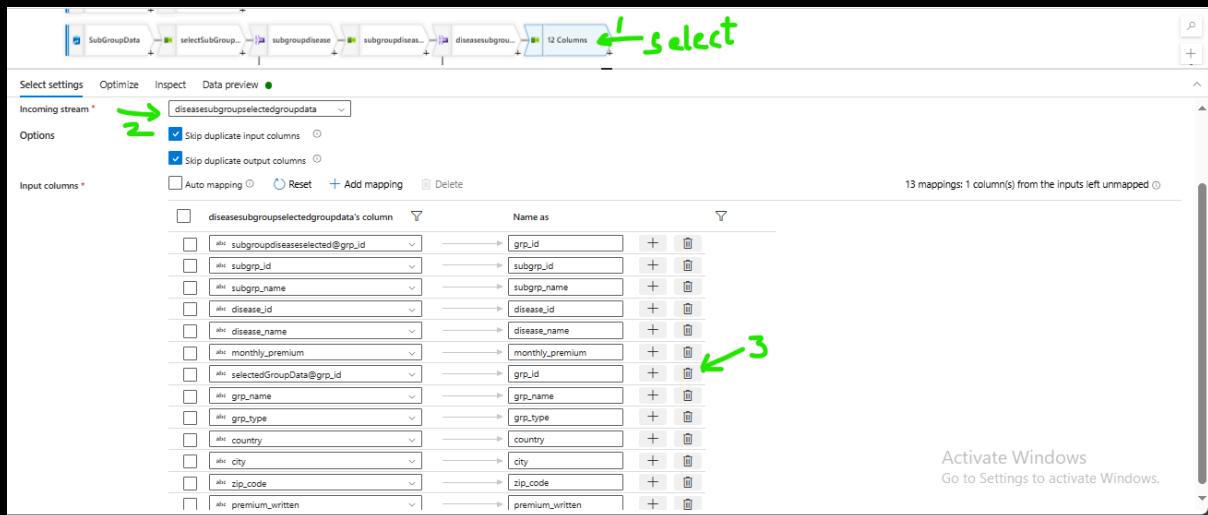
grp_id	subgrp_id	subgrp_name	disease_id	disease_name	monthly_premium	grp_na...	grp_type	country	city	zip_code
GRP143	S110	Viral	110060	Chicken...	1000	Magma ...	Private	India	Mumbai	482009
GRP143	S110	Viral	110055	Hepatitis	1000	Magma ...	Private	India	Mumbai	482009
GRP143	S110	Viral	110056	Mumps	1000	Magma ...	Private	India	Mumbai	482009
GRP143	S110	Viral	110057	Pneumo...	1000	Magma ...	Private	India	Mumbai	482009
GRP143	S110	Viral	110058	Shingles	1000	Magma ...	Private	India	Mumbai	482009
GRP143	S110	Viral	110059	Flu	1000	Magma ...	Private	India	Mumbai	482009
GRP147	S110	Viral	110060	Chicken...	1000	Raheja ...	Private	India	Mumbai	482040
GRP147	S110	Viral	110055	Hepatitis	1000	Raheja ...	Private	India	Mumbai	482040
GRP147	S110	Viral	110056	Mumps	1000	Raheja ...	Private	India	Mumbai	482040
GRP147	S110	Viral	110057	Pneumo...	1000	Raheja ...	Private	India	Mumbai	482040
GRP147	S110	Viral	110058	Shingles	1000	Raheja ...	Private	India	Mumbai	482040

Activate Windows Go to Settings to activate Windows.

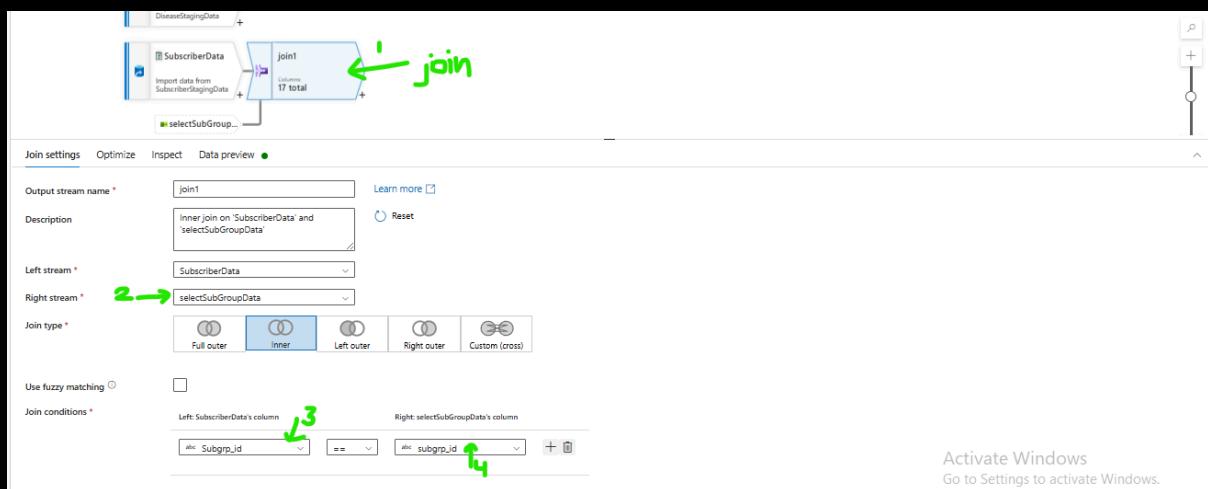
- Joining `subgroupdiseaseselected` with `selectedgroupdata`.

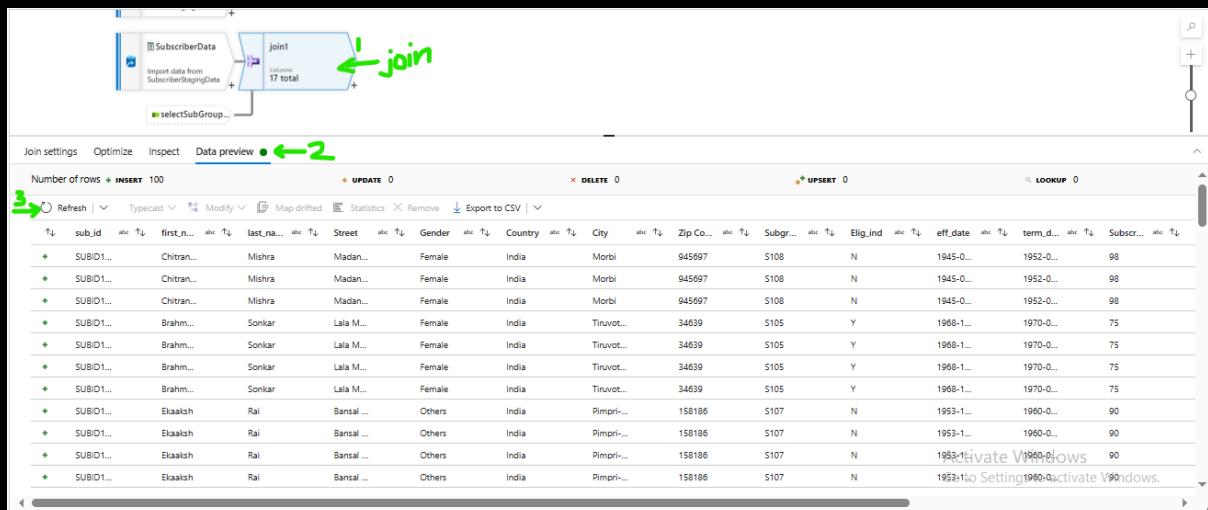


- Add `select activity` after `diseasesubgroupselectedgroupdata`.



- Joining `subscriber_data` with `selectedSubGroupData`





- Add sink after this join activity.

Output stream name * **sink1**

Description Add sink dataset

Incoming stream * **join1**

Sink type * **Dataset**

Dataset * **Select...** + New **2**

Options Allow schema drift Validate schema

New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store

All	Azure	Database	File	Generic protocol	NoSQL	Services and apps
Azure Data Lake Storage (Kusto)		Gen2				MySQL
	Azure Database for PostgreSQL	Azure SQL Database 3			Azure SQL Database Managed Instance	
			Azure Synapse Analytics	Dataverse (Common Data Service for Apps)	Dynamics 365	

Output stream name * **sink1**

Description Add sink dataset

Incoming stream * **join1**

Sink type * **Dataset**

Dataset * **Select...** + New **2**

Options Allow schema drift Validate schema

Continue **4**

Microsoft Azure | Data Factory > projects-datafactory

Validate all Publish all 1

dataflow1

Validate Data flow debug Debug Settings

SubscriberData Import data from SubscriberStagingData join1 Inner join on SubscriberData and selectSubGroupData sink1 Columns: 17 total

selectSubGroupD...

Sink Settings Errors Mapping Optimize Inspect Data preview •

Output stream name * sink1 Learn more

Description Add sink dataset Reset

Incoming stream * join1

Sink type * Dataset Inline Cache

Dataset * Select... + New

Options Allow schema drift Validate schema

Set properties

Name subscriber_tb

Linked service * Select... Filter... + New f2

Activate Windows Go to Settings to activate Windows.

OK Back Cancel

New linked service

Azure SQL Database [Learn more](#)

Name * adf2asql_ls

Description

Connect via integration runtime * [\(i\)](#) AutoResolveIntegrationRuntime

Version

Recommended Legacy 2

[Connection string](#) [Azure Key Vault](#)

Account selection method [\(i\)](#)

From Azure subscription Enter manually

Azure subscription

Free Trial (c92e5287-784e-4a20-9e6c-8549947f2ee6) 3

Server name *

optumsserver 4

Database name *

optumnosql_db 5

Authentication type *

SQL authentication

User name *

optum_admin 6

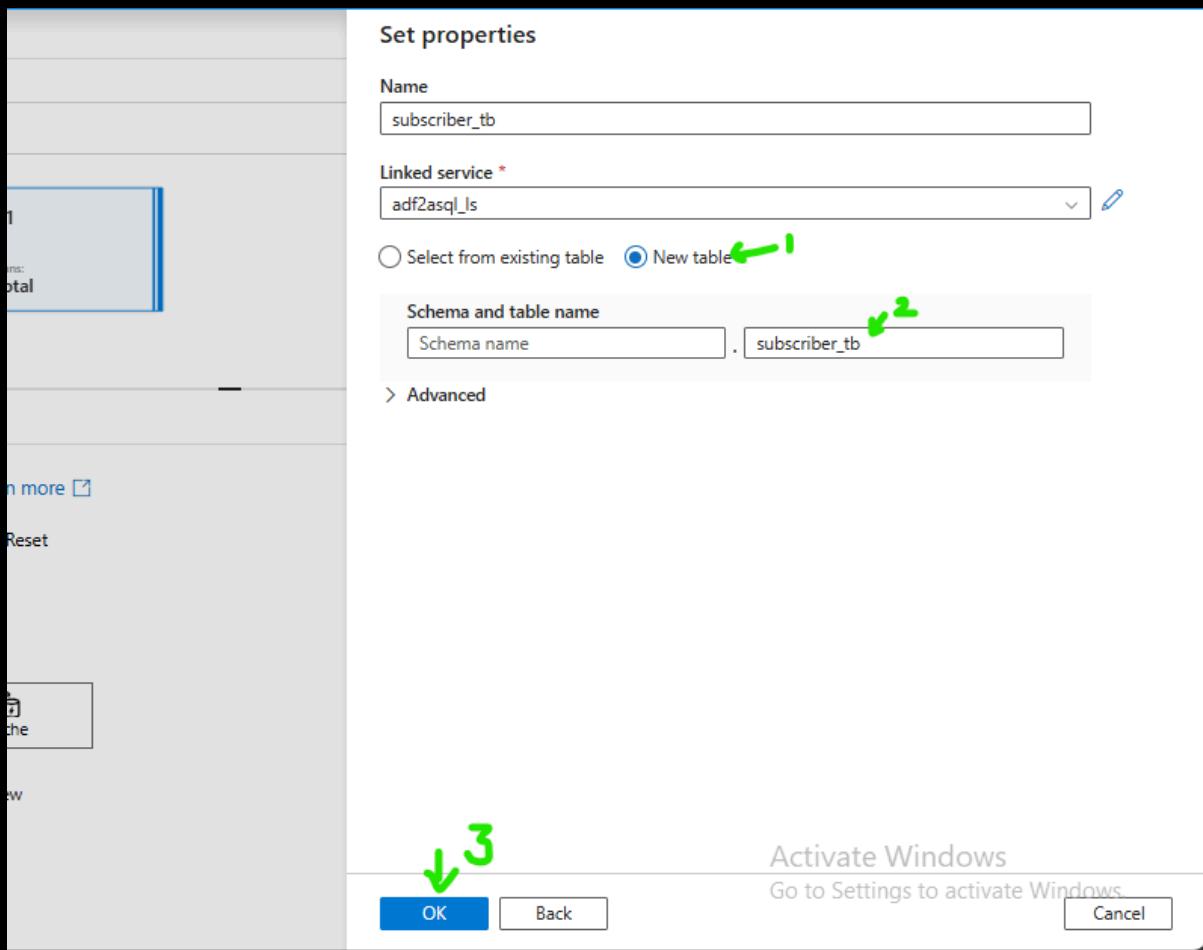
[Password](#) [Azure Key Vault](#)

Password *

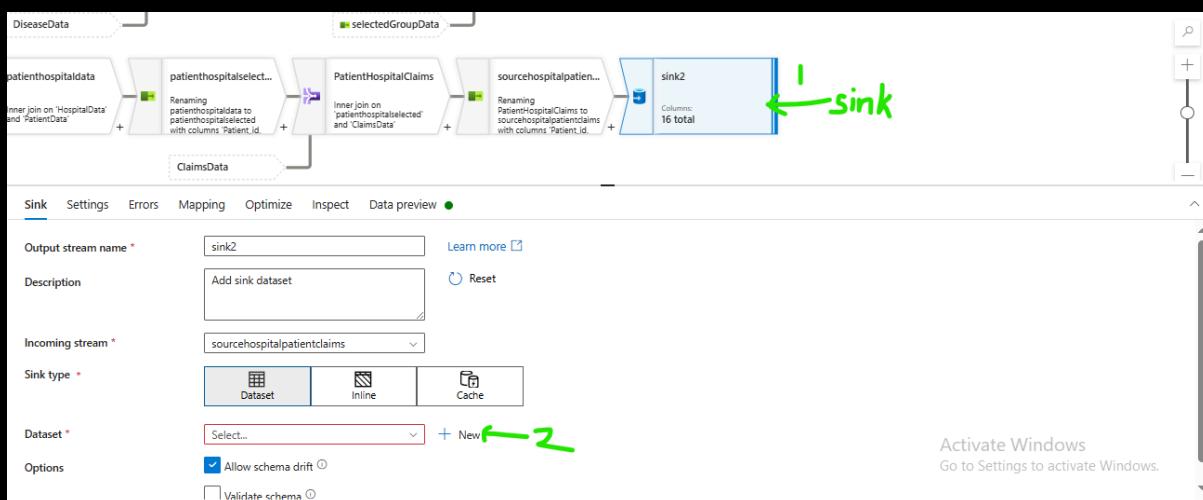
..... 7

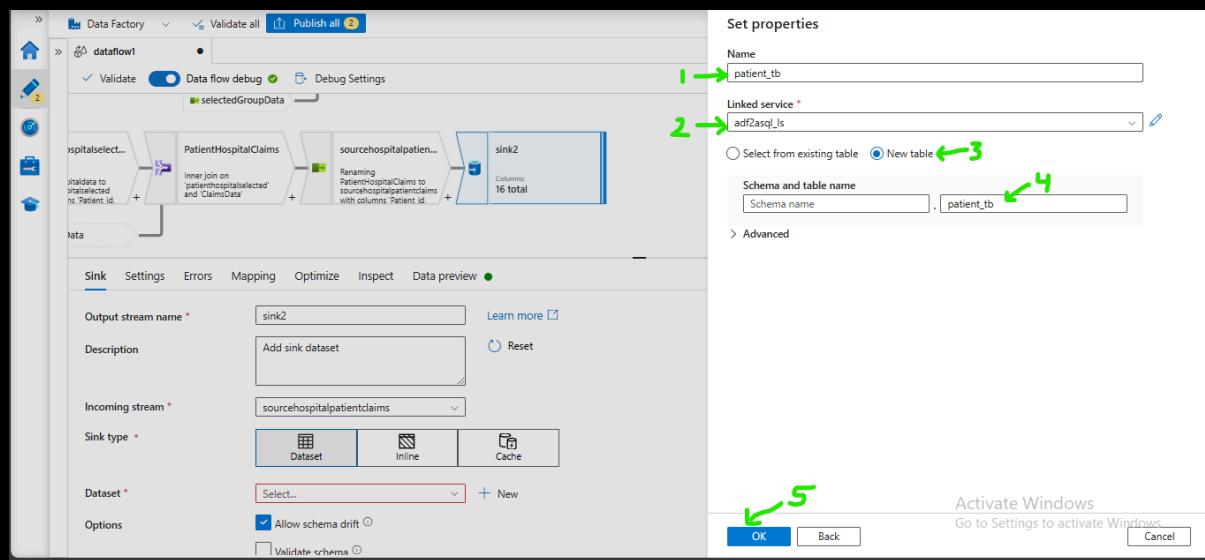
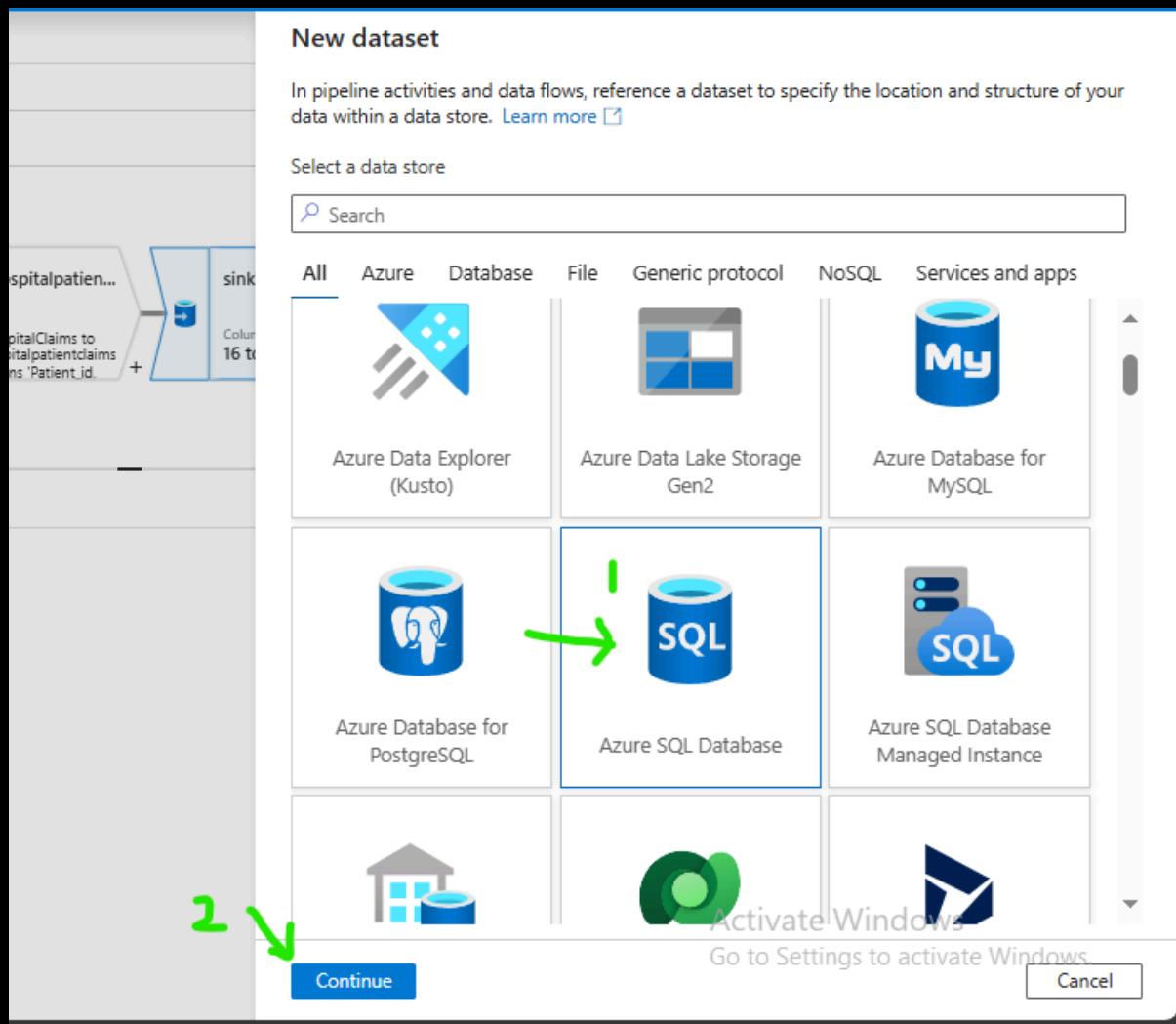
Activate Windows
Go to Settings to activate Windows. Connection successful

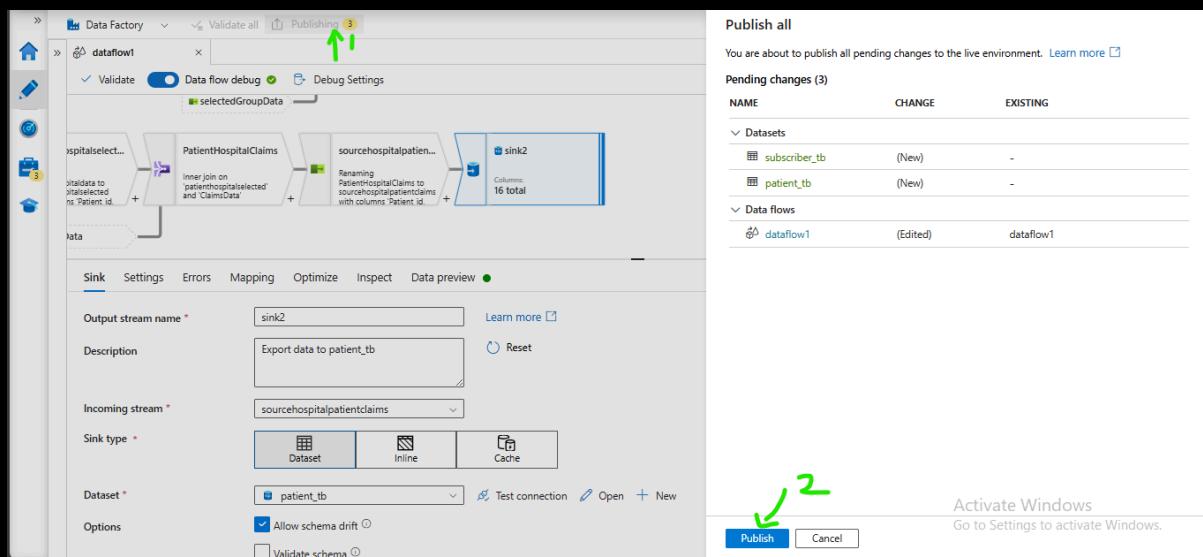
Create Cancel 9 Test connection 8



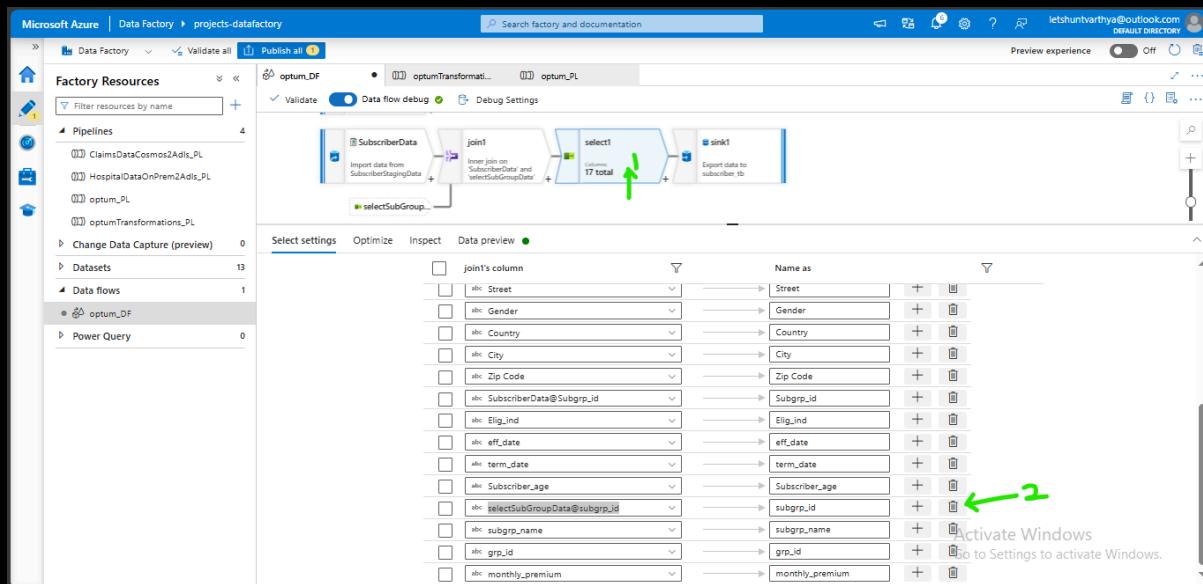
- Add sink after sourcehospitalpatientclaims.



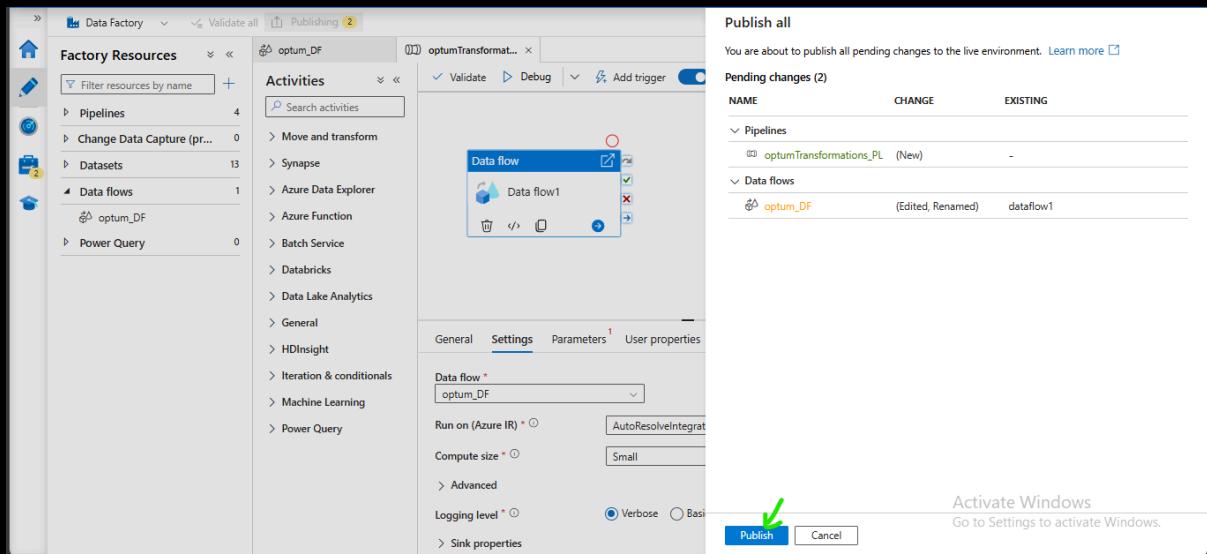
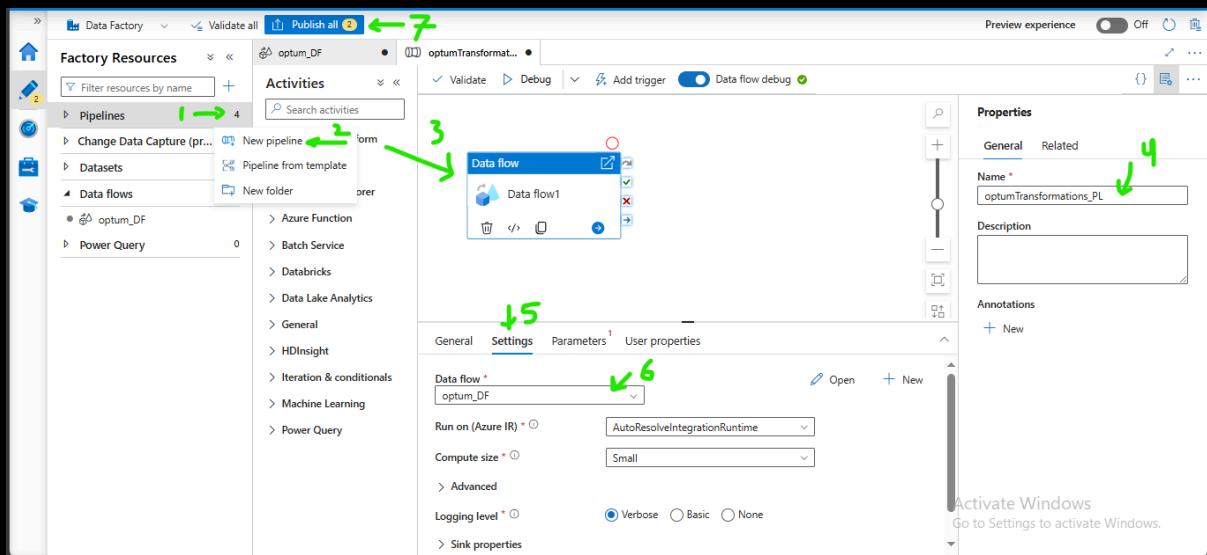




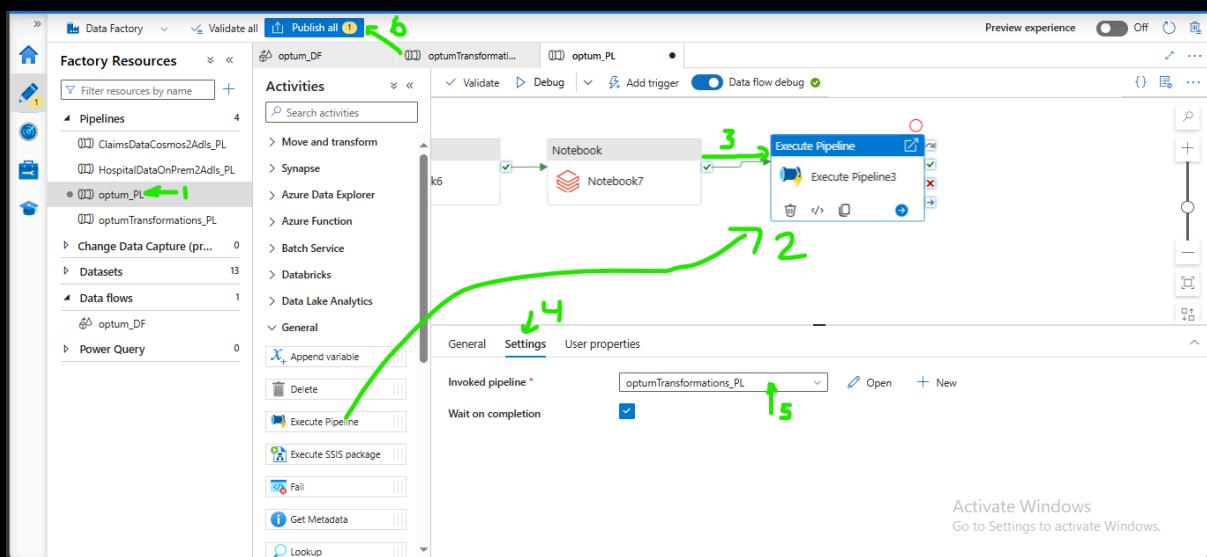
- In **optum_DF** add **select** after joining of **subscriber** and **selectSubGroupData**.
- Delete **selectSubGroupData@subgrp_id** column.

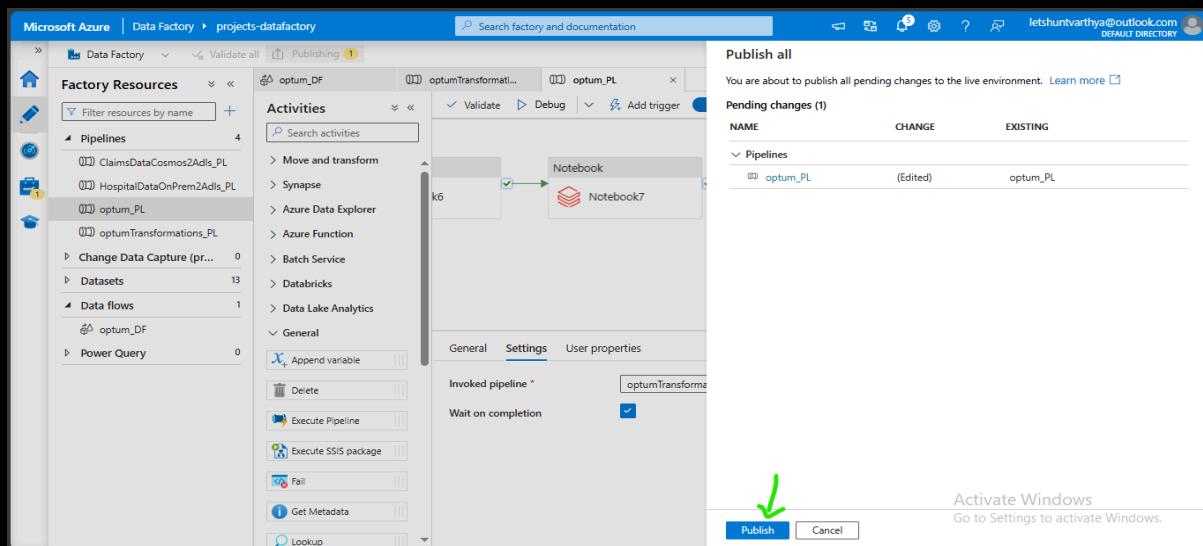


- Create a new **pipeline** to run **optum_DF**.
- Pipeline name: **optumTransformation_PL**
- Drag and drop **data flow activity** from move and transform section to the field.
- Publish all → publish**

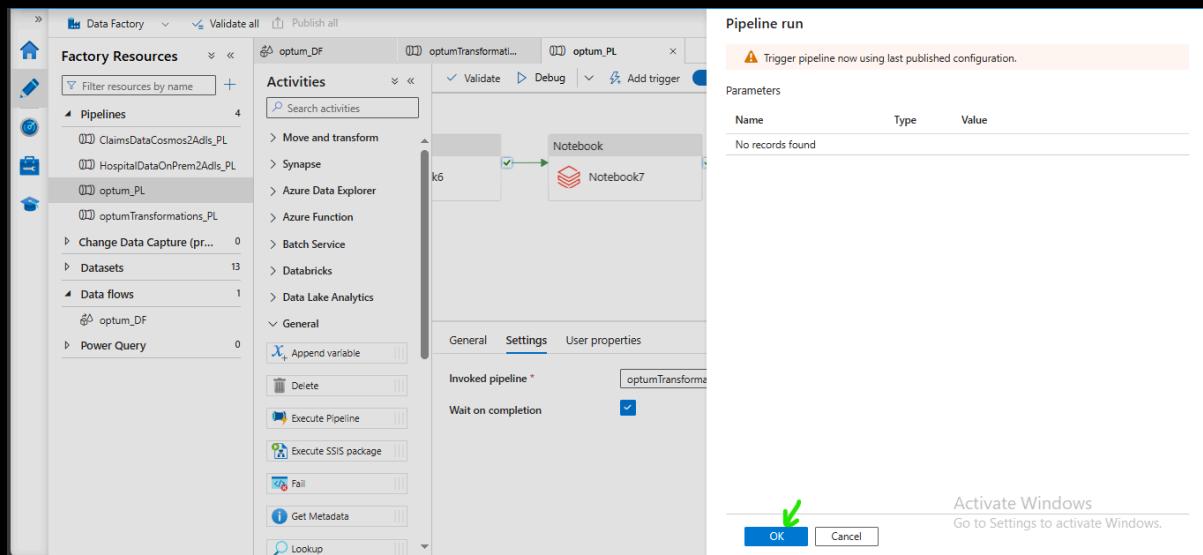


- In the **optum_PL**, drag and drop **execute pipeline activity** to the **field** and select **optumTransformation_PL** in the **settings**.
- Publish all → publish**

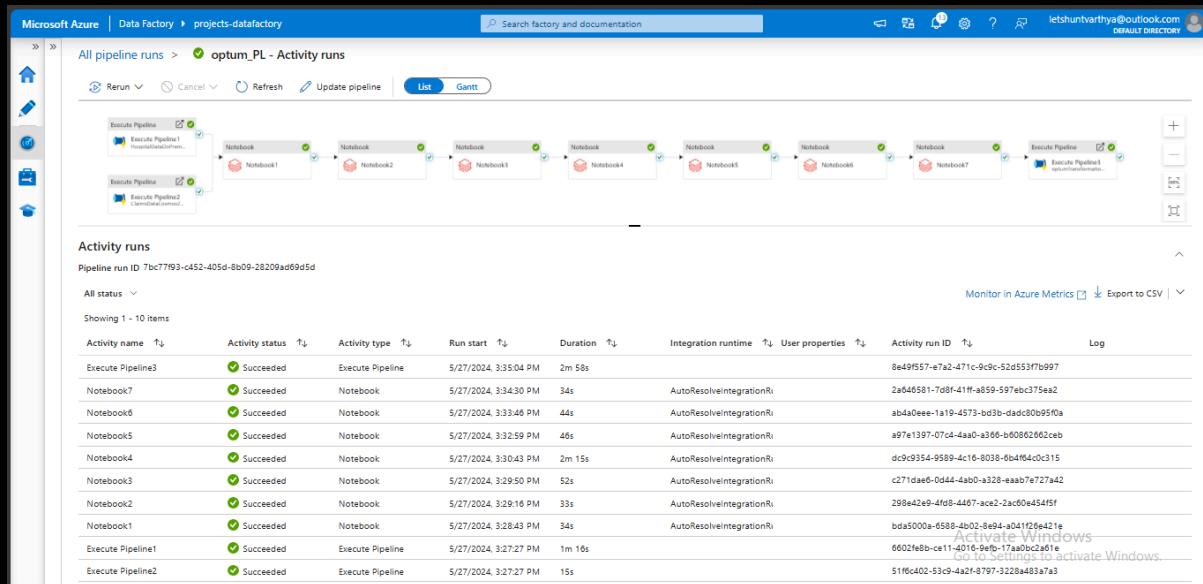




- Add trigger → trigger now → ok



- Successful execution of pipeline.



- After successful execution of pipelines. We can see the data in the **azure sql database**.

Screenshot 1: Azure SQL Database Query Editor (preview) - Step 1

Welcome to SQL Database Query Editor

SQL server authentication

Login * optum_admin

Password *

OK ← 4

Microsoft Entra authentication

Continue as letsrunvarthy@outlook.c...

OR

Activate Windows
Go to Settings to activate Windows.

Screenshot 2: Azure SQL Database Query Editor (preview) - Step 2

Showing limited object explorer here. For full capability please click here to open Azure Data Studio.

Tables

1 SELECT TOP (1000) * FROM [dbo].[patient_tb] → 2

Results Messages

Patient_Id	Patient_name	patient_gender	Patient_Age	hospital_city	state	country	Hospital_name	patient_city	claim_id	disease
146382	Dharmadas	Male	60	Bengaluru	Karnataka	India	Apollo Hospitals - Bannerghatta Road	Bhalwa Jahangir Pur	12	Anhra
107794	Manish	Male	56	Chennai	Tamil Nadu	India	Apollo Hospital - Chennai	Panvel	6	Sunbat
109242	Chitrangan	Female	98	Delhi	UT	India	Sir Ganga Ram Hospital	Morbi	48	Asthm
199114	Guest/NA	Female	69	Bengaluru	Karnataka	India	Manipal Hospitals	Vijayawada	38	Phenyl
197441	Deependu	Female	72	Gurgaon	Haryana	India	Medanta The Medicity	Bareilly	50	Lung c

Activate Windows
Go to Settings to activate Windows.

Screenshot 3: Azure SQL Database Query Editor (preview) - Step 3

Showing limited object explorer here. For full capability please click here to open Azure Data Studio.

Tables

1 SELECT TOP (1000) * FROM [dbo].[subscriber_tb] → 2

Results Messages

sub_id	first_name	last_name	Street	Gender	Country	City	Zip_Code	Subgrp_id	Elig_ind	eff
SUBID10048	Chitrangan	Mishra	Madan Nagar	Female	India	Morbi	945697	S108	N	194
SUBID10048	Chitrangan	Mishra	Madan Nagar	Female	India	Morbi	945697	S108	N	194
SUBID10048	Chitrangan	Mishra	Madan Nagar	Female	India	Morbi	945697	S108	N	194
SUBID10001	Brahmdev	Sonkar	Lala Marg	Female	India	Tiruvottiyur	34639	S105	Y	196
SUBID10001	Brahmdev	Sonkar	Lala Marg	Female	India	Tiruvottiyur	34639	S105	Y	196

Activate Windows
Go to Settings to activate Windows.