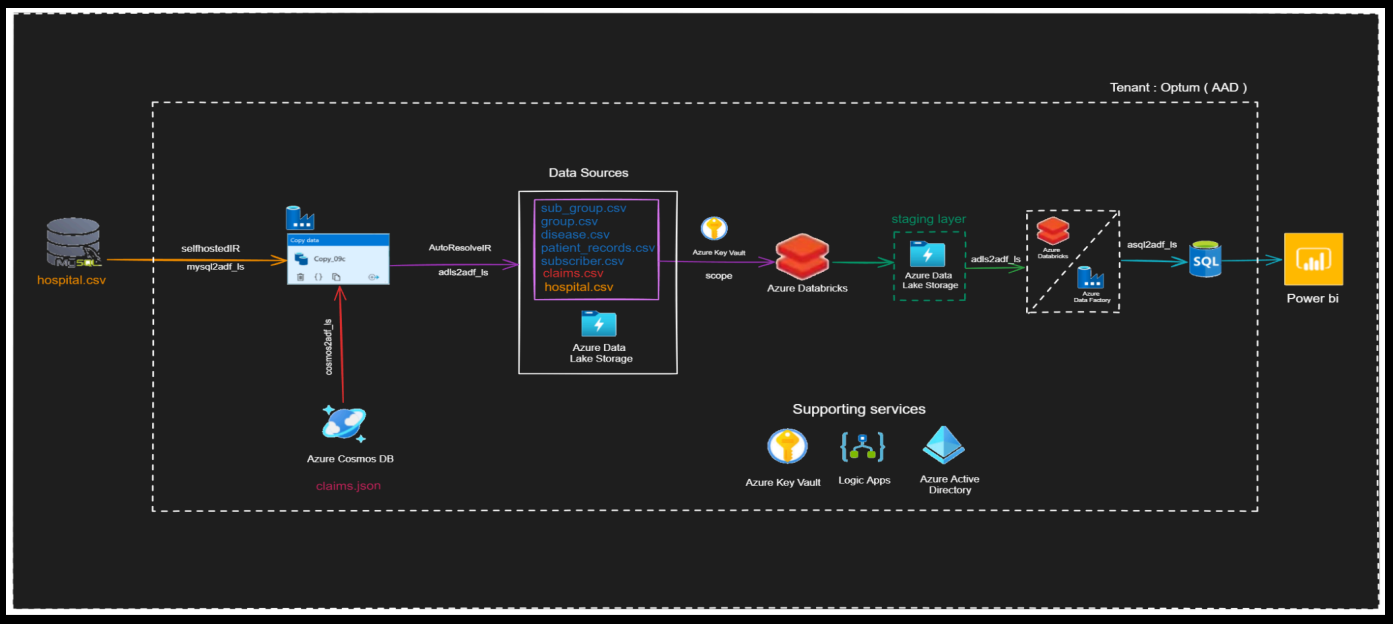
Project Architecture :

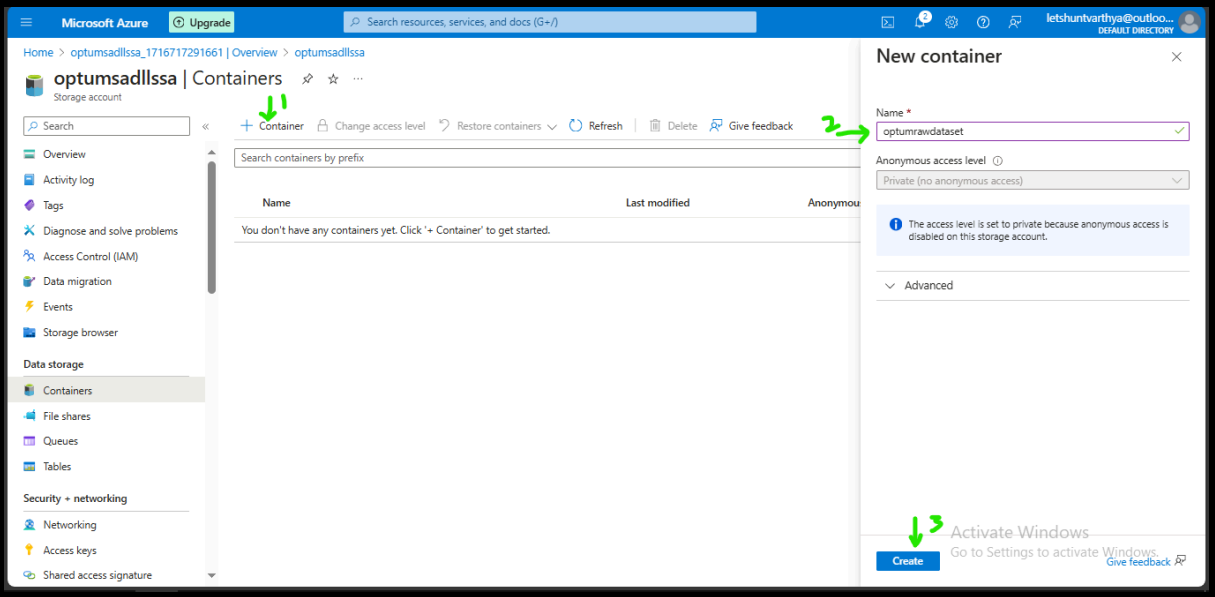


Services Required :

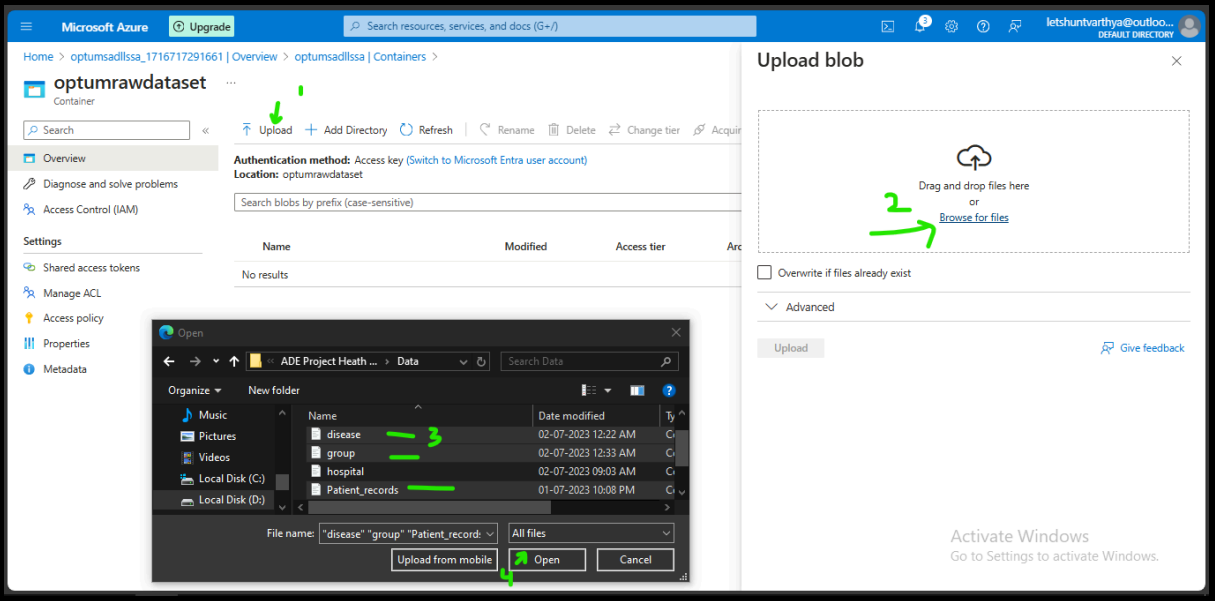
1. Azure Data Lake Storage.
2. Azure cosmos DB for mongoDB.
3. Azure SQL Database.
4. Azure Data Factory.
5. Azure Databricks
6. Azure Key Vault.
7. On prem mysql

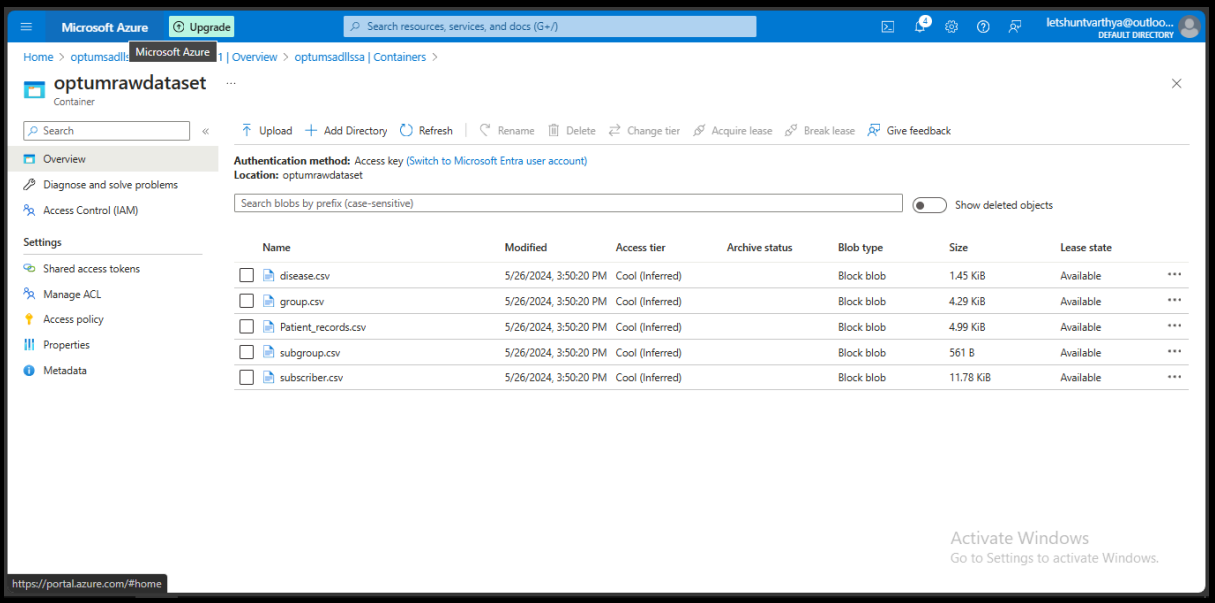
Azure Data Lake Storage creation :

- The Azure data lake storage creation is shown in the below document.
- https://docs.google.com/document/d/1Gyz7yN9HDF7d_i6wM_i9ph0Z0PNiKP_FLgPfRJshs_A/edit?usp=sharing
- Parameters used while creating this service.
  Storage account name : optumsadllssa
- Create a container called 'optumrawdata' and click on it.

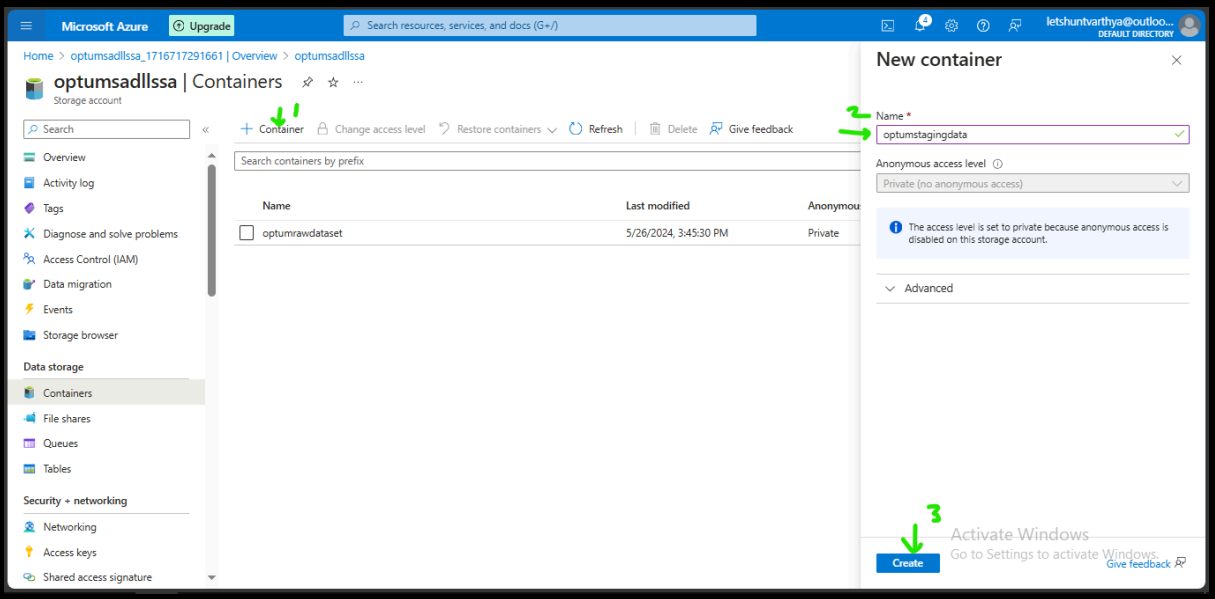


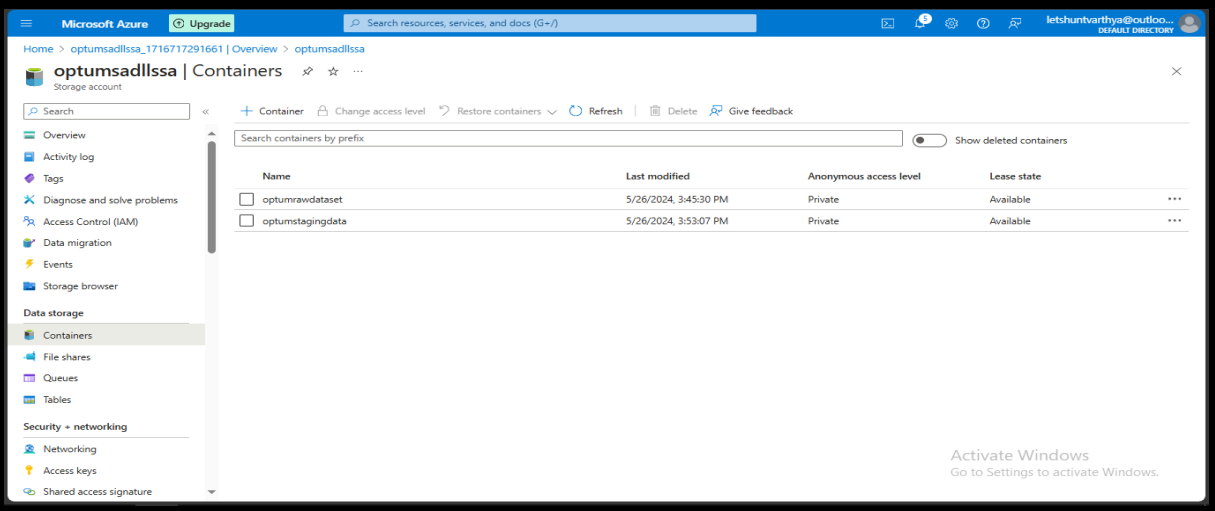
- Upload csv files (disease,group,patient_records, subgroup and subscriber in the optumrawdata and click on the 'upload' button.





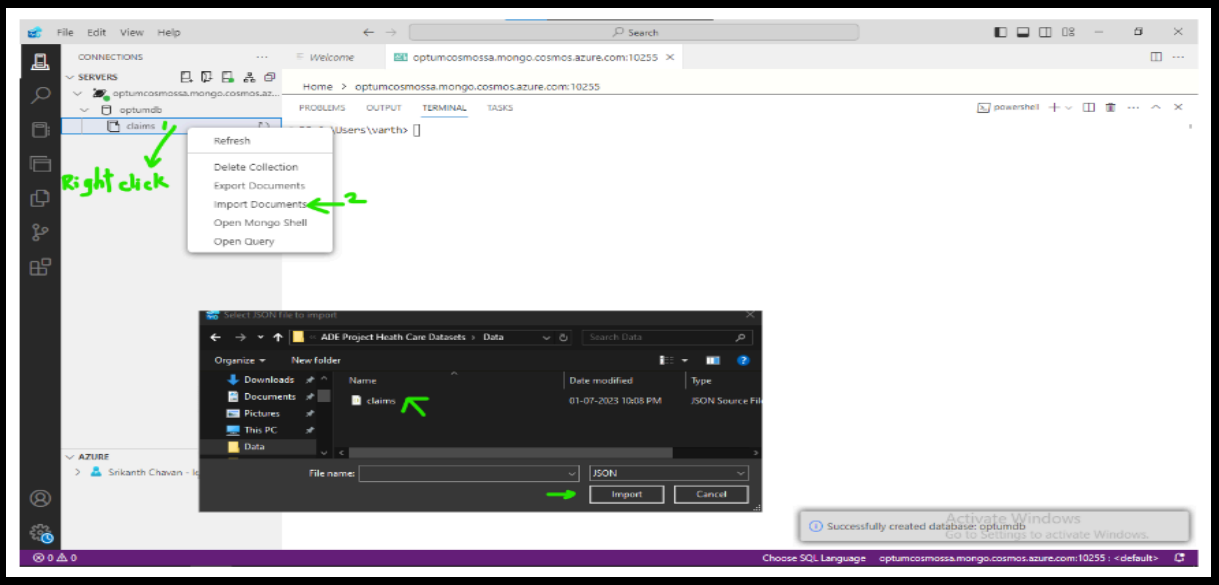
- Create a container called 'optumstagingdata in the optumsadllssa.



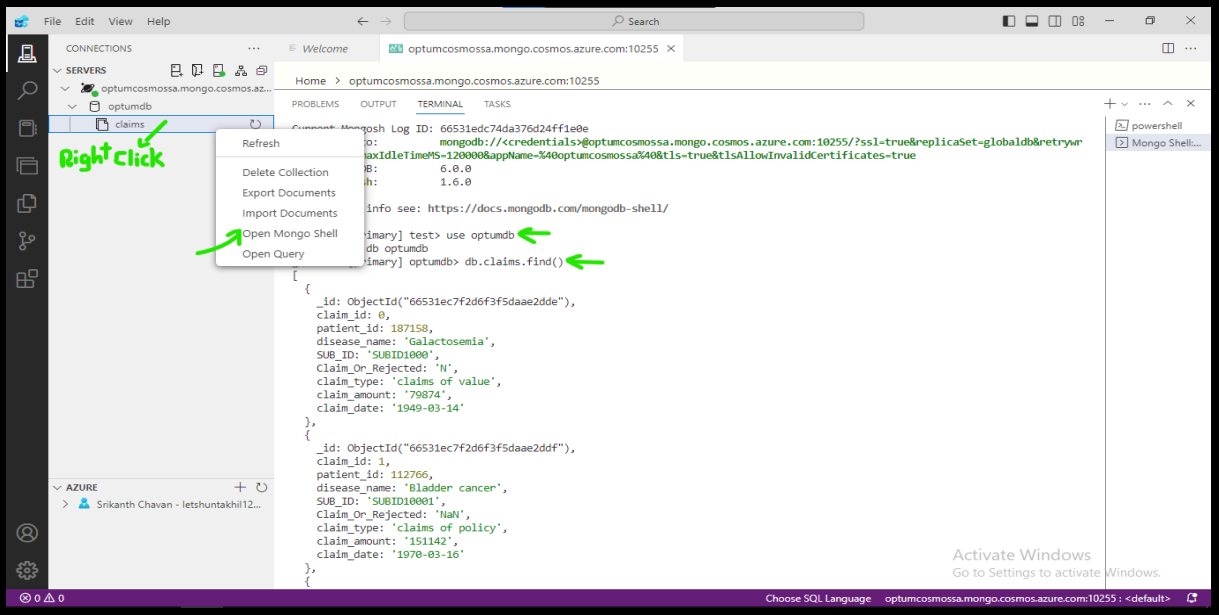


Azure Cosmos DB creation :

- 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:
  1. Create a database with the name : optumdb
  2. Create a collection with the name : claims



- Query the data.



Azure SQL Database Creation :

- The azure 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 : optumnosql_db
- 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/1IpvA7XumJjbIP0wPWWUlfd_gdvh6jTn_0a12HWv0jcM/edit?usp=sharing
- Parameters used while creating this service.
  Name : projects-datafactory

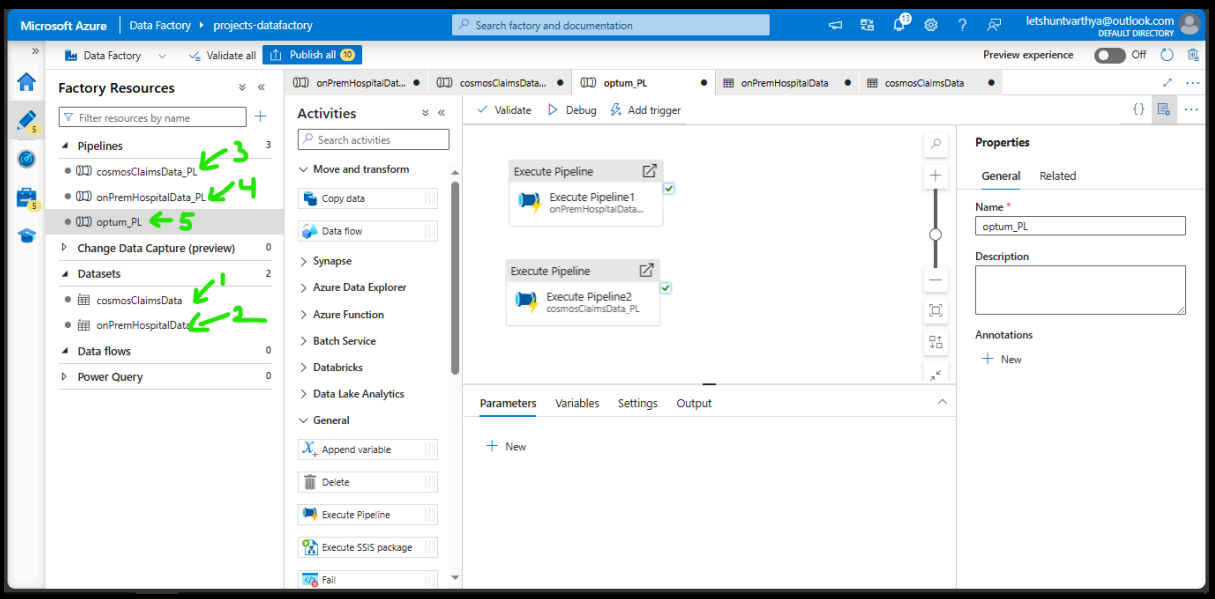
Azure databricks creation :

- The azure databricks creation is shown in the document below.
- https://docs.google.com/document/d/1s7Vqs4gcZbGMPRk7vEWGFBXIMC5AoaOvC-E57qILwRw/edit?usp=sharing
- Parameters used while creating this service.
- Workspace name : optum_ws

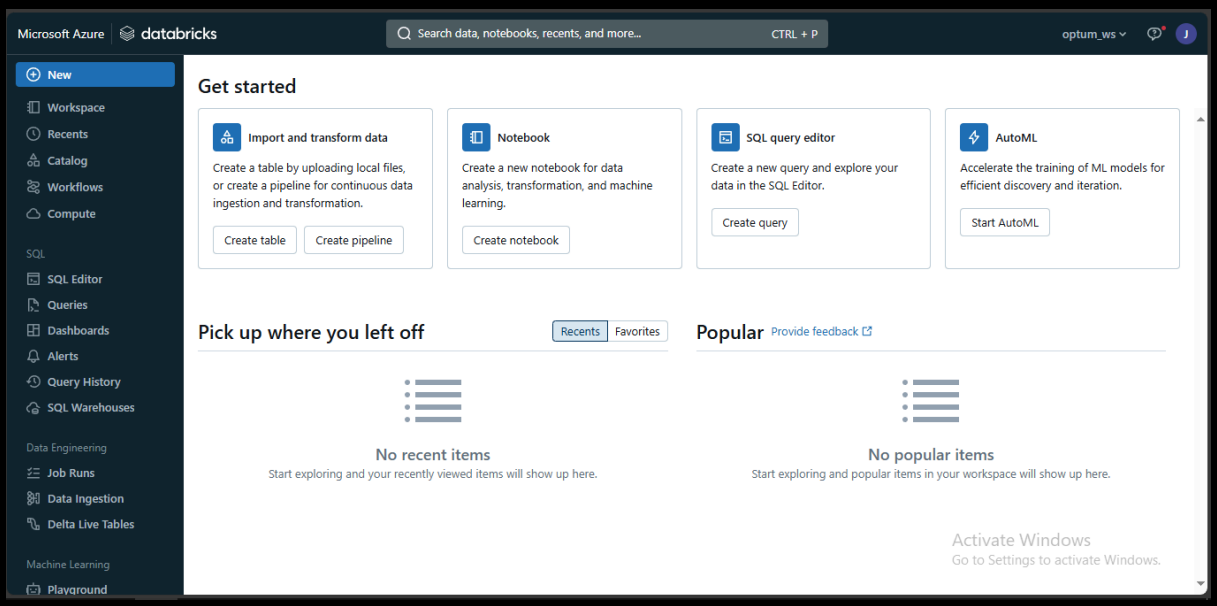
Azure keyvault creation :

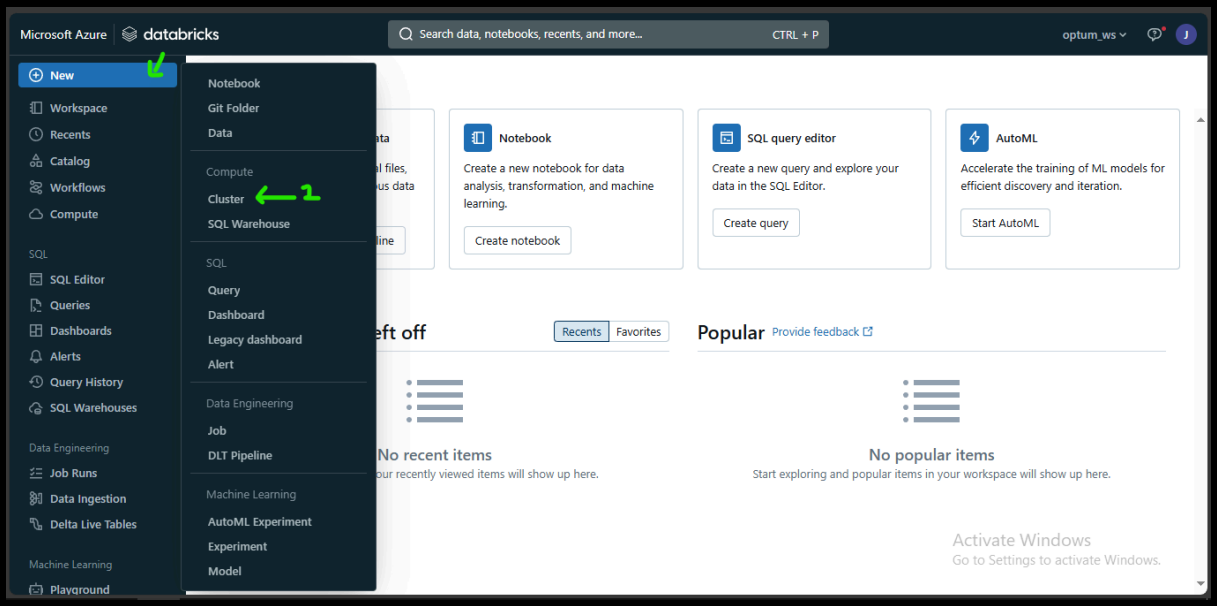
- The azure azure keyvault creation is shown in the document below.
- https://docs.google.com/document/d/1ScQ42B5c5ZuRnFjtLsdJpnV9067pLZkpcDf4WyKjaVw/edit?usp=sharing
- Parameters used while creating this service.
- Keyvault name : optums-keyvault

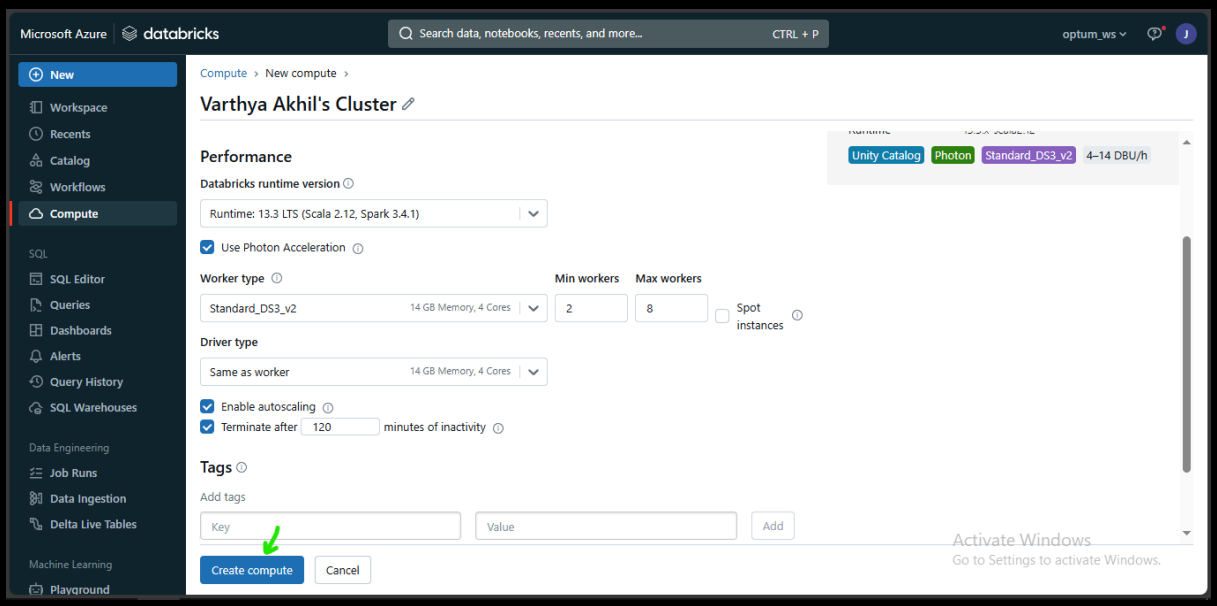
- Let's bring the hospital data (onprem mysql), Claims(optumcosmossa(Azure Cosmos db for mongodb)) to adls (optumrawdatafactory) via datafactory.
- Assuming hospital data is already there in onprem mysql.
- Create dataset and pipeline for hospitaldata
- Ceate dataset and piepeline for claimsdata
- Create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

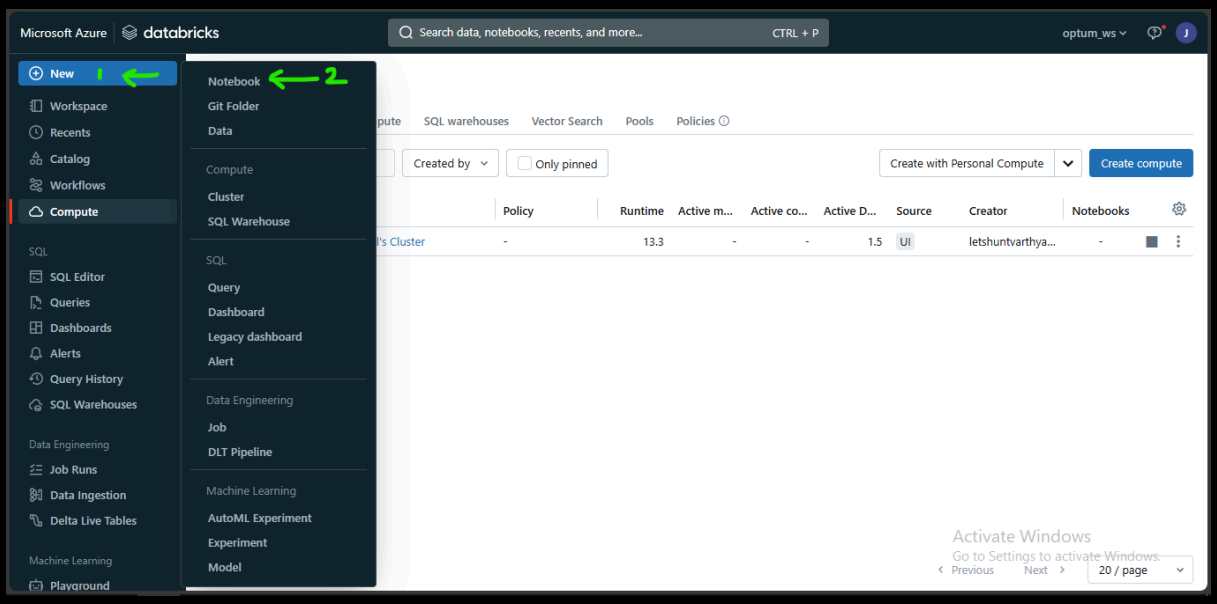


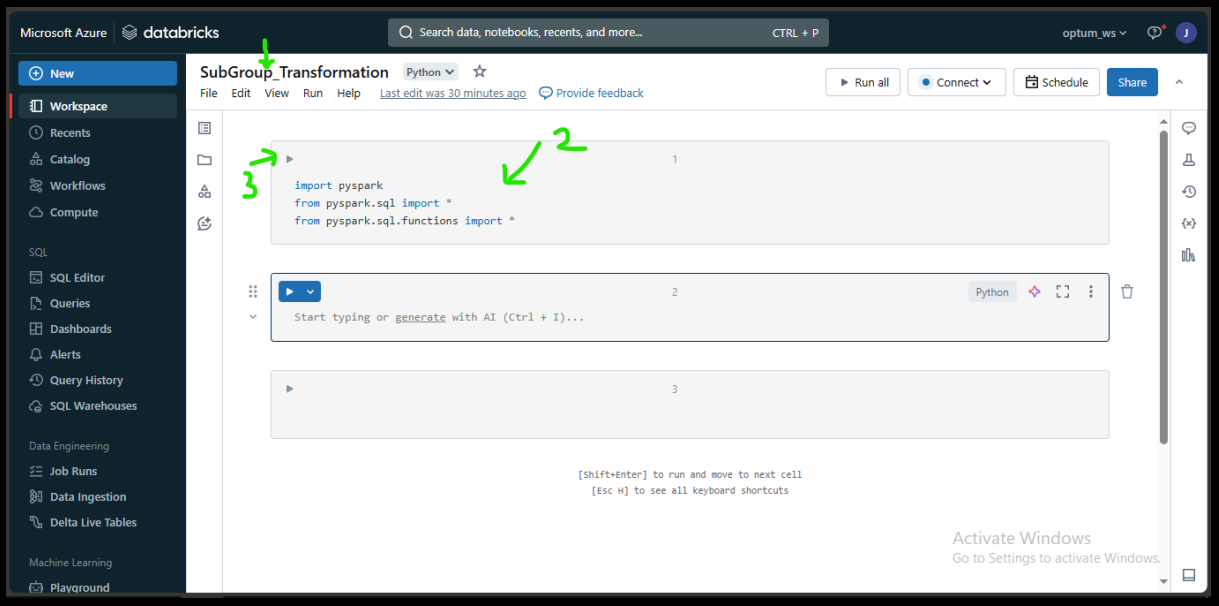
- Publish all → publish
- Add Trigger → Trigger Now
- Let's do some basic transformations on top of this data in databricks(optum_db).




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
Notebook
Create a new notebook for data analysis, transformation, and machine learning.
Create notebook
SQL query editor
Create a new query and explore your data in the SQL Editor.
Create query
AutoML
Accelerate the training of ML models for efficient discovery and iteration.
Start AutoML
Recents
Favorites
Popular
Provide feedback
No recent items
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
Varthya Akhil's Cluster
Performance
Unity Catalog
Photon
Standard_DS3_v2
4–14 DBU/h
Databricks runtime version
Runtime: 13.3 LTS (Scala 2.12, Spark 3.4.1)
Use Photon Acceleration
Worker type
Min workers
Max workers
Standard_DS3_v2
14 GB Memory, 4 Cores
2
8
Spot instances
Driver type
Same as worker
14 GB Memory, 4 Cores
Enable autoscaling
Terminate after
120
minutes of inactivity
Tags
Add tags
Key
Value
Add
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
1
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
SQL warehouses
Vector Search
Pools
Policies
Created by
Only pinned
Create with Personal Compute
Create compute
Policy
Runtime
Active m...
Active co...
Active D...
Source
Creator
Notebooks
's Cluster
13.3
1.5
UI
letshuntvarthya...
Activate Windows
Go to Settings to activate Windows.
Previous
Next
20 / page



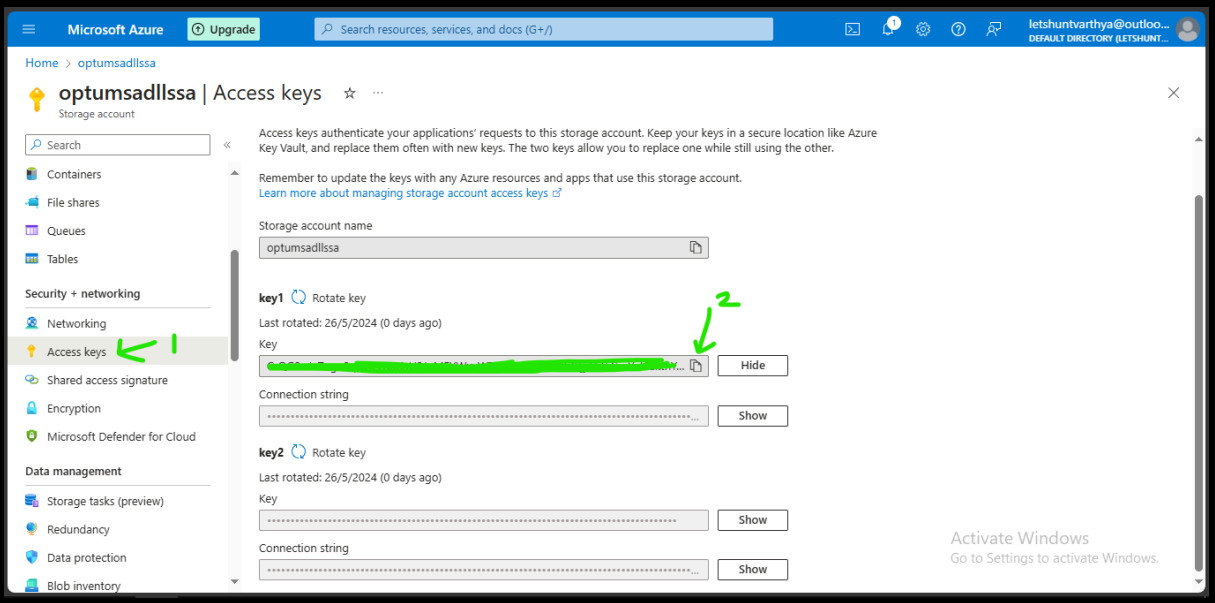
- Boilerplate code to import the data from adls to databricks via keyvault and access key

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

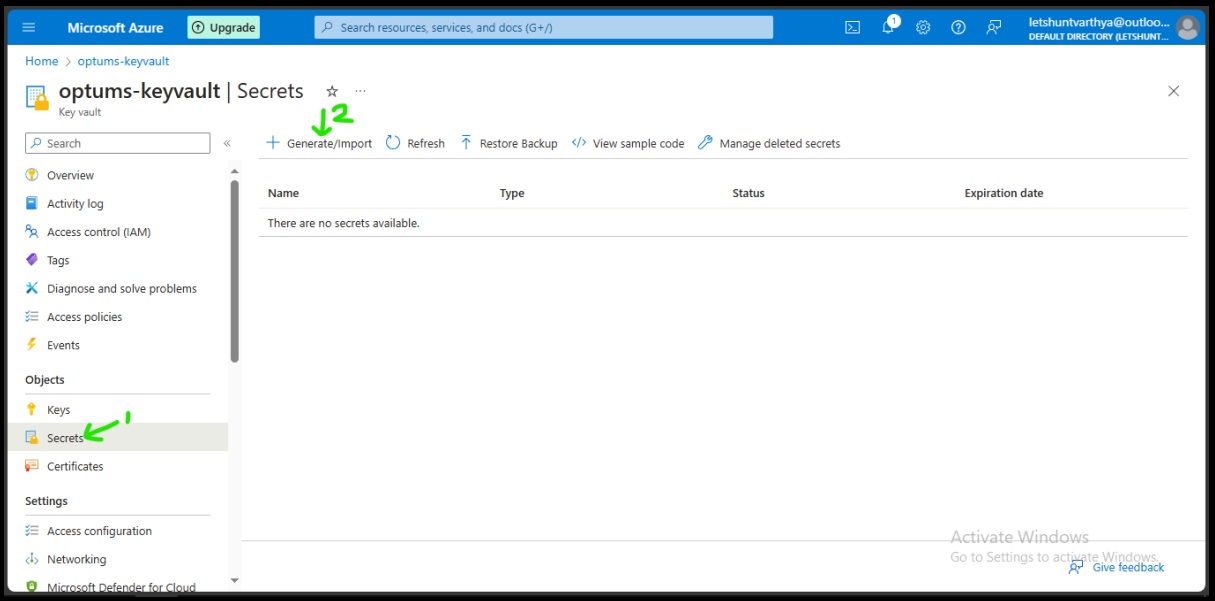
- Replace <storage-account> with adls account name (optumsadllsa) and <storage-account-acces-key> with access key from 'optumsadllsa' access key
- Here we cannot directly put the access key, due to security reasons. But we can import the access key via keyvault and scope.
- Optumsadllsa key is stored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

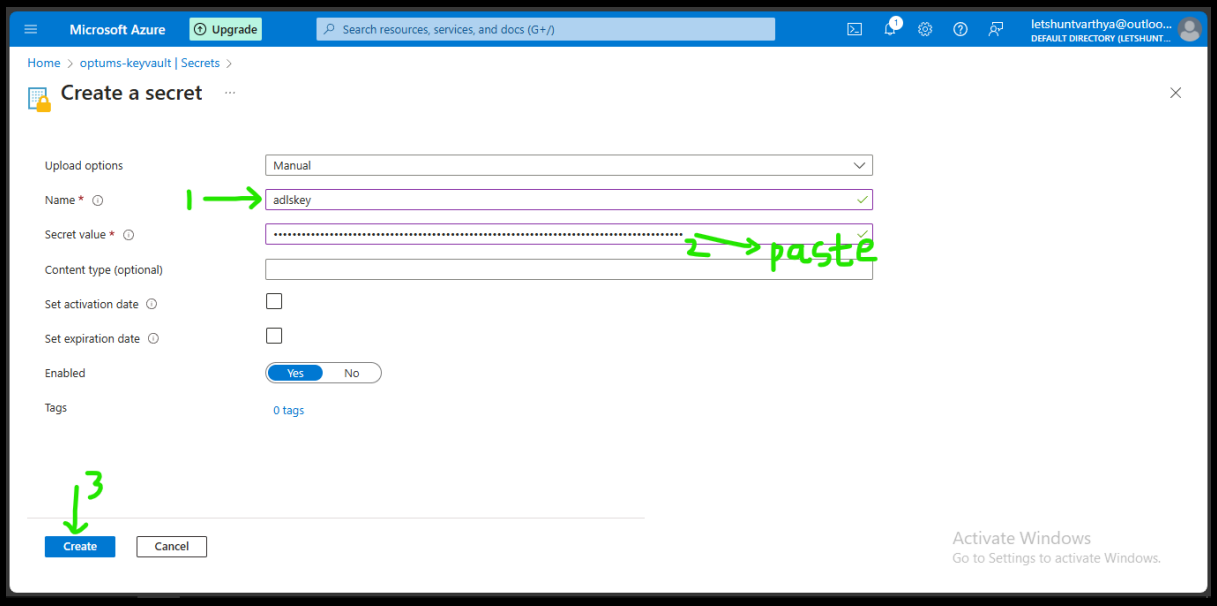
- Steps to import the data from adls to databricks via keyvault and access key.

1. Copy access key from optumsadllsa.



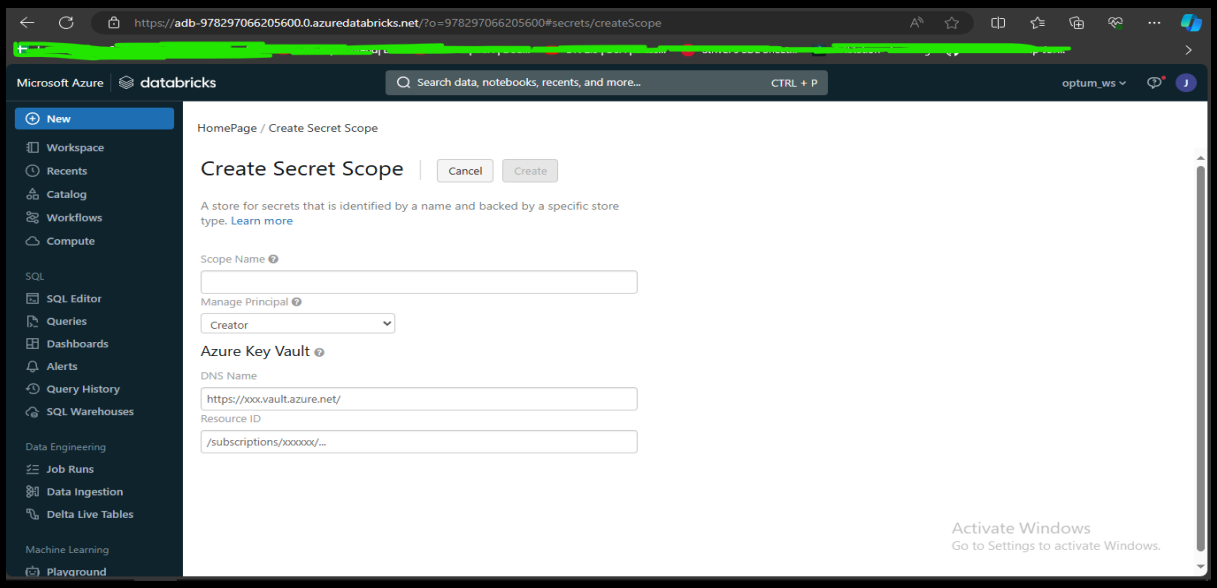
2. Create a secret for optumsadllsa.

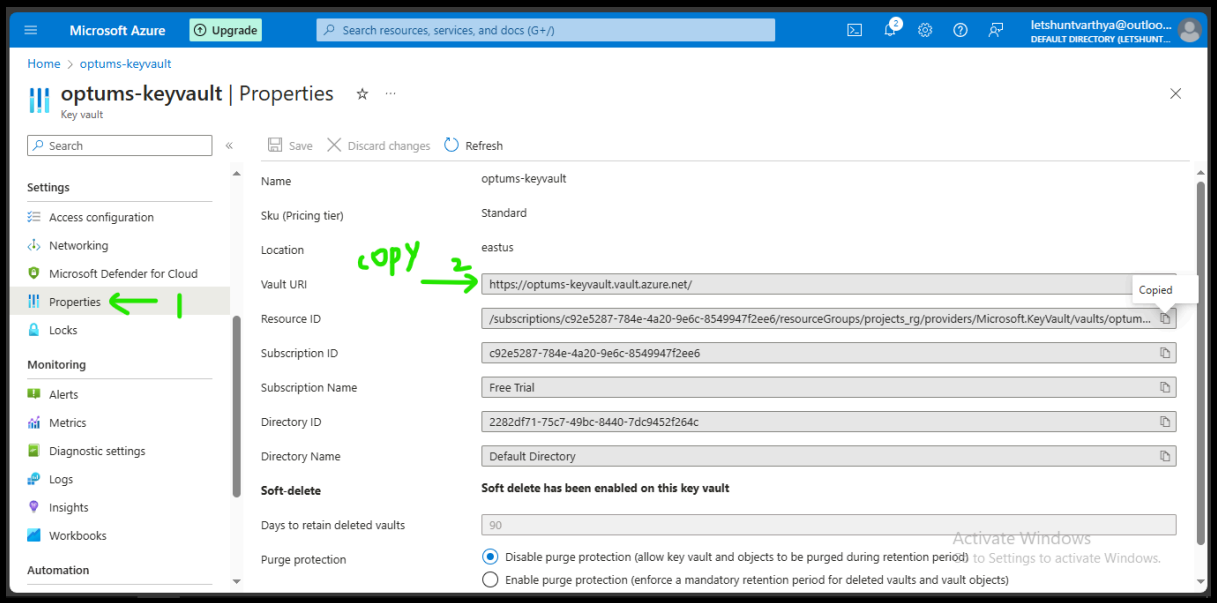


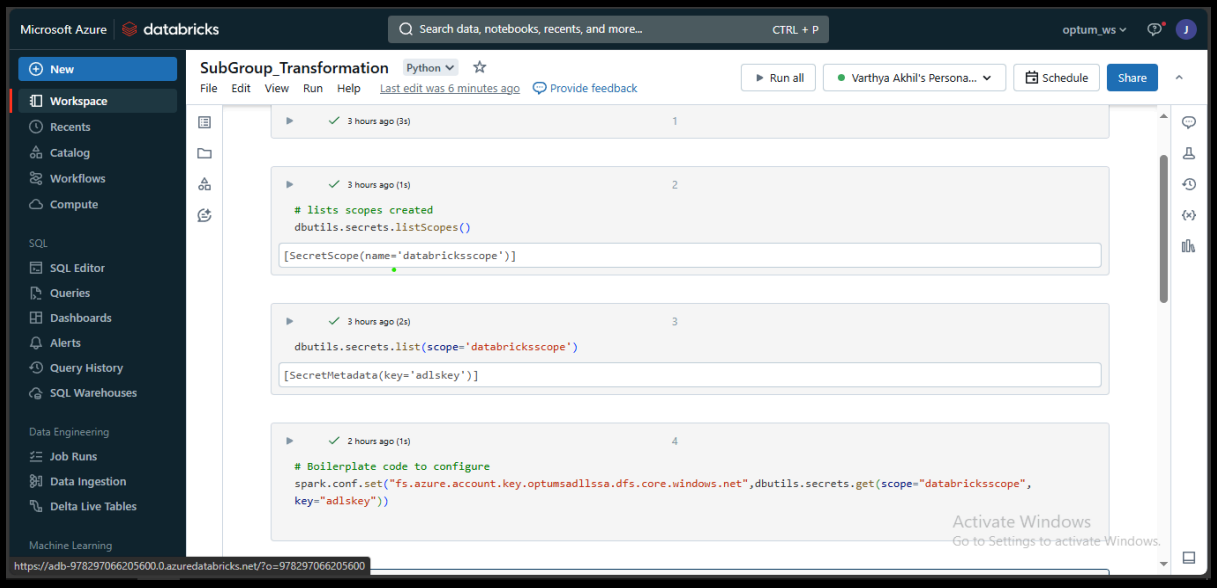


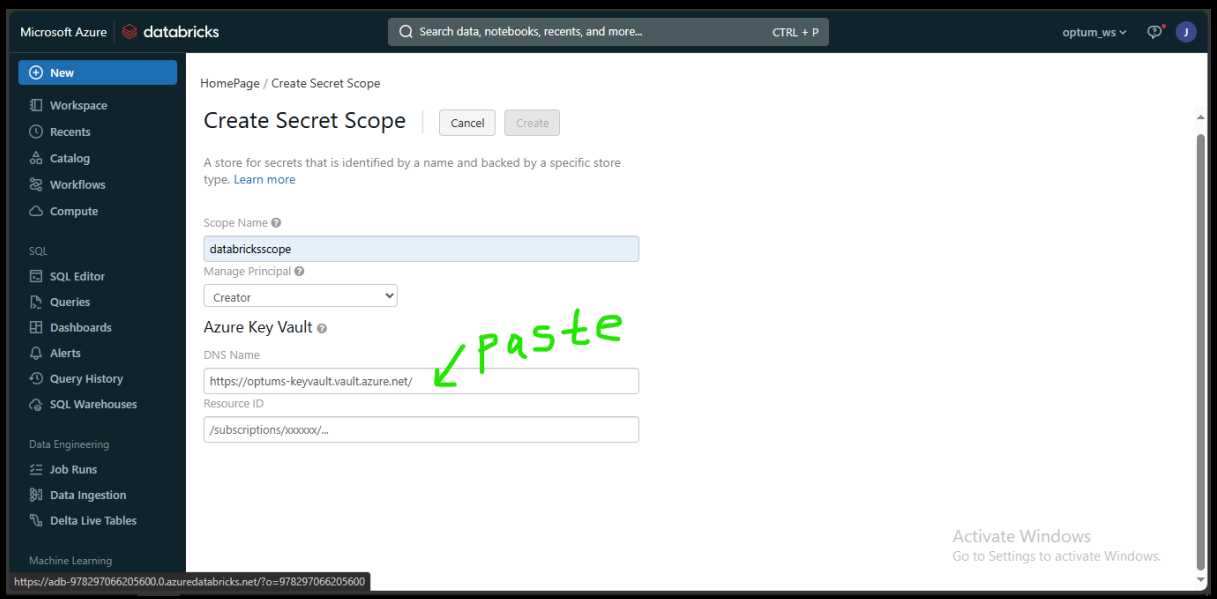
3. Creating a scope in databricks.

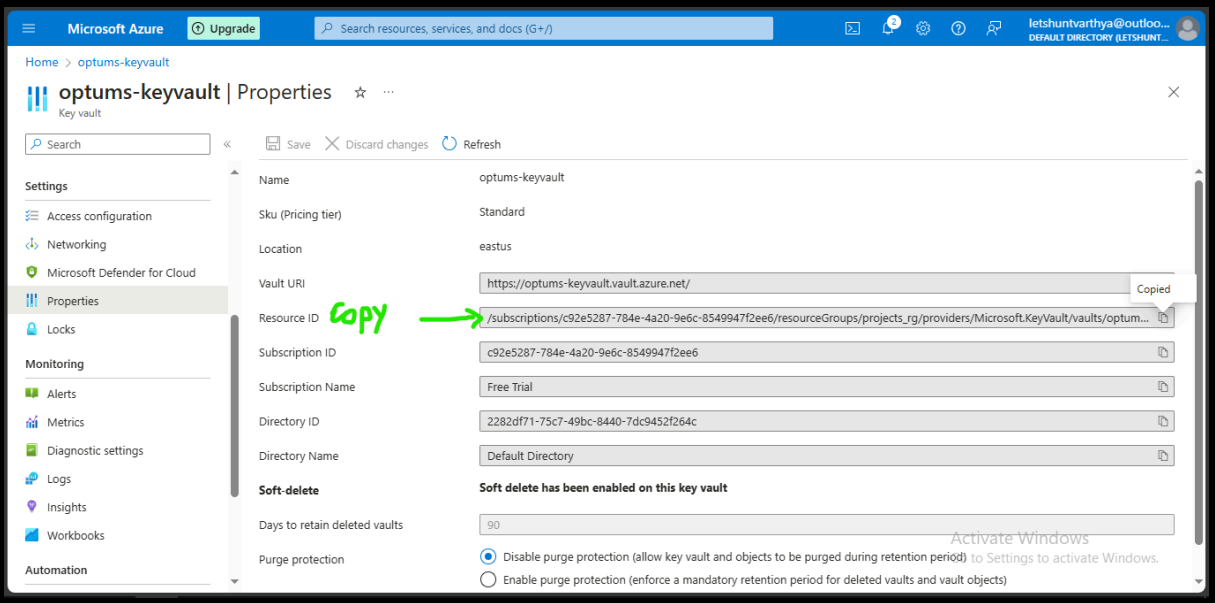
- Copy the url, paste in a new tab and remove the characters after the # and put secrets/createScope (url#secrets/createScope).

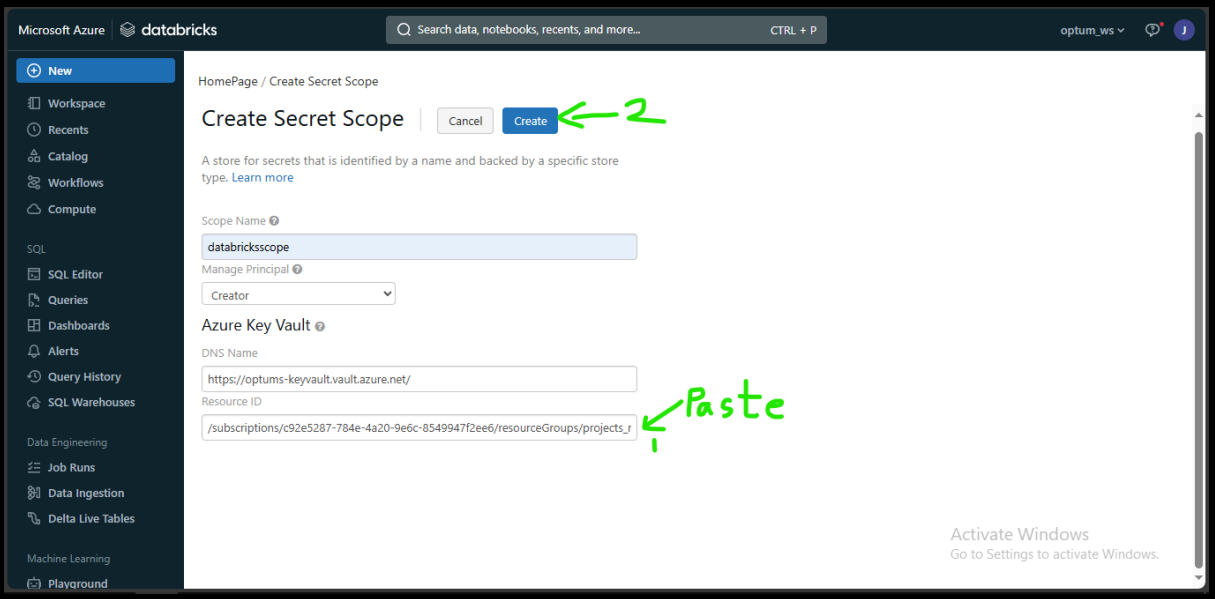


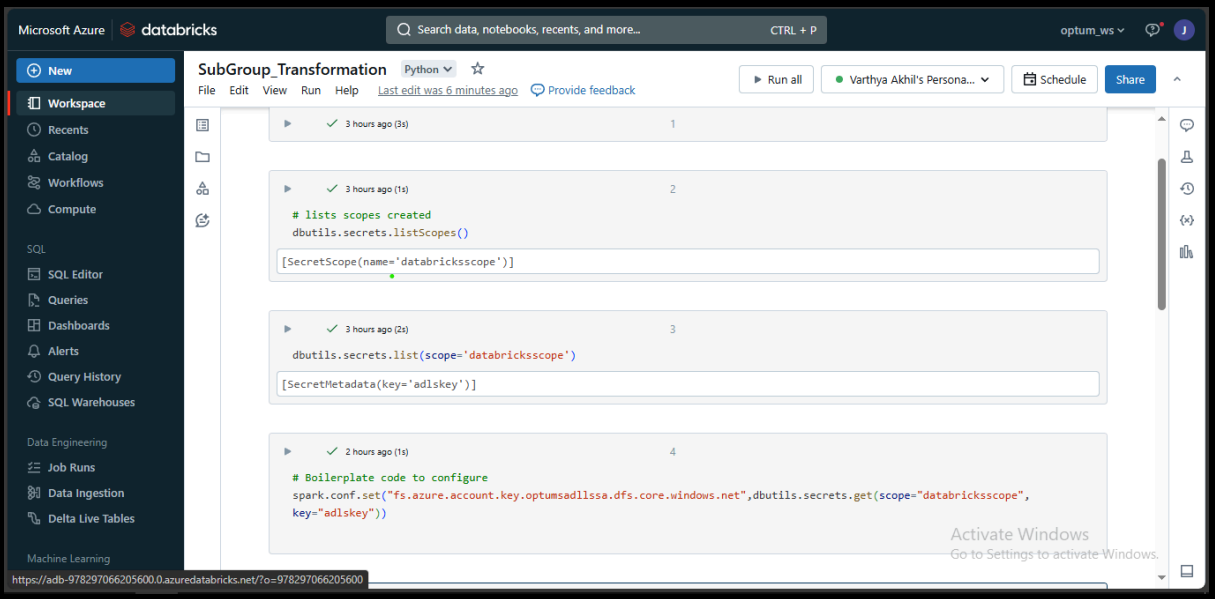

Microsoft Azure
Upgrade
Search resources, services, and docs (G+/)
letshuntvarthya@outloo...
DEFAULT DIRECTORY (LETSHUNT...
Home > optums-keyvault
optums-keyvault | Properties
Key vault
Search
Save
Discard changes
Refresh
Settings
Access configuration
Networking
Microsoft Defender for Cloud
Properties
Locks
Monitoring
Alerts
Metrics
Diagnostic settings
Logs
Insights
Workbooks
Automation
Name
optums-keyvault
Sku (Pricing tier)
Standard
Location
eastus
copy
Vault URI
https://optums-keyvault.vault.azure.net/
Copied
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)
Activate Windows
Go to Settings to activate Windows.

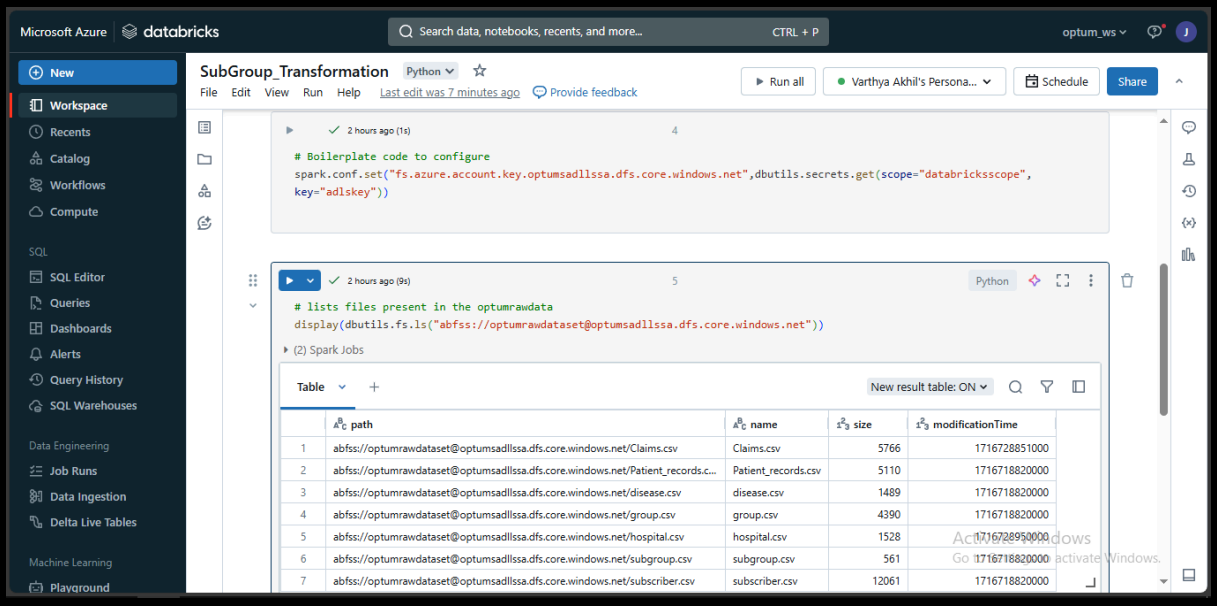

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
optum_ws
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
SubGroup_Transformation
Python
File Edit View Run Help
Last edit was 6 minutes ago
Provide feedback
Run all
Varthya Akhil's Persona...
Schedule
Share
3 hours ago (3s)
1
3 hours ago (1s)
2
# lists scopes created
dbutils.secrets.listScopes()
[SecretScope(name='databricksscope')]
3 hours ago (2s)
3
dbutils.secrets.list(scope='databricksscope')
[SecretMetadata(key='adlskey')]
2 hours ago (1s)
4
Boilerplate code to configure
spark.conf.set("fs.azure.account.key.optumsadllssa.dfs.core.windows.net",dbutils.secrets.get(scope="databricksscope",
key="adlskey"))
Activate Windows
Go to Settings to activate Windows.
https://adb-978297066205600.0.azuredatabricks.net/?o=978297066205600


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
HomePage / Create Secret Scope
Create Secret Scope
Cancel
Create
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
paste
https://optums-keyvault.vault.azure.net/
Resource ID
/subscriptions/xxxxxx/...
Activate Windows
Go to Settings to activate Windows.
https://adb-978297066205600.0.azuredatabricks.net/?o=978297066205600


Microsoft Azure
Upgrade
Search resources, services, and docs (G+/)
letshuntvarthya@outloo...
DEFAULT DIRECTORY (LETSHUNT...
Home > 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/
Copied
Resource ID
Copy
/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)
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
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_r
Paste
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
SubGroup_Transformation
Python
File Edit View Run Help
Last edit was 6 minutes ago
Provide feedback
Run all
Varthya Akhil's Persona...
Schedule
Share
3 hours ago (3s)
1
3 hours ago (1s)
2
# lists scopes created
dbutils.secrets.listScopes()
[SecretScope(name='databricksscope')]
3 hours ago (2s)
3
dbutils.secrets.list(scope='databricksscope')
[SecretMetadata(key='adlskey')]
2 hours ago (1s)
4
Boilerplate code to configure
spark.conf.set("fs.azure.account.key.optumsadllssa.dfs.core.windows.net",dbutils.secrets.get(scope="databricksscope",
key="adlskey"))
Activate Windows
Go to Settings to activate Windows.
https://adb-978297066205600.0.azuredatabricks.net/?o=978297066205600



12:39 AM (5s) 6 Python

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

(1) Spark Jobs

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

12:40 AM (1s) 7

```
# 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
```

12:41 AM (2s) 8

```
# Number of rows
data.count()
```

(2) Spark Jobs

```
10
```

12:41 AM (<1s) 9

```
# Removing duplicate Rows
data = data.dropDuplicates()
```

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

12:41 AM (1s) 10

```
data.count()
```

(3) Spark Jobs

```
10
```

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
```

12:47 AM (1s) 15 Python

```
# display of null values in all columns
data.select([count(when(isnan(c)|col(c).isNull(),c)).alias(c) for c in data.columns]).show(5,False)
```

(3) Spark Jobs

```
+---------+-----------+---------------+---------+
|subgrp_sk|subgrp_name|monthly_premium|subgrp_id|
+---------+-----------+---------------+---------+
|0        |0          |0              |0        |
+---------+-----------+---------------+---------+
```

12:50 AM (2s) 16 Python

```
# finding duplicates
data.groupby(data.columns).count().where("count>1").show(5,False)
```

(3) Spark Jobs

```
+---------+-----------+---------------+---------+-----+
|subgrp_sk|subgrp_name|monthly_premium|subgrp_id|count|
+---------+-----------+---------------+---------+-----+
+---------+-----------+---------------+---------+-----+
```


01:08 AM (2s) 17

```
# boiler plate to push code from databricks to adls
output_container_path = "abfss://optumstagingdata@optumsadllssa.dfs.core.windows.net"
output_blob_folder ="optumstagingdata"
data.coalesce(1).write.mode("overwrite").option("header","true").format("com.databricks.spark.csv").save(output_blob_folder)
files = dbutils.fs.ls(output_blob_folder)
output_file = [x for x in files if x.name.startswith("part-")]
dbutils.fs.mv(output_file[0].path,"%s/subgroupstagingdata.csv"%output_container_path)
```

(2) Spark Jobs

```
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_eiaFfPuUUtDX3d9Hsub9392DCVL/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_eiaFfPuUUtDX3d9Hsub9392DCVL/view?usp=sharing

- Transformation of patient data is shown in the document below.
    - Create a new notebook (Patient_Transformation).

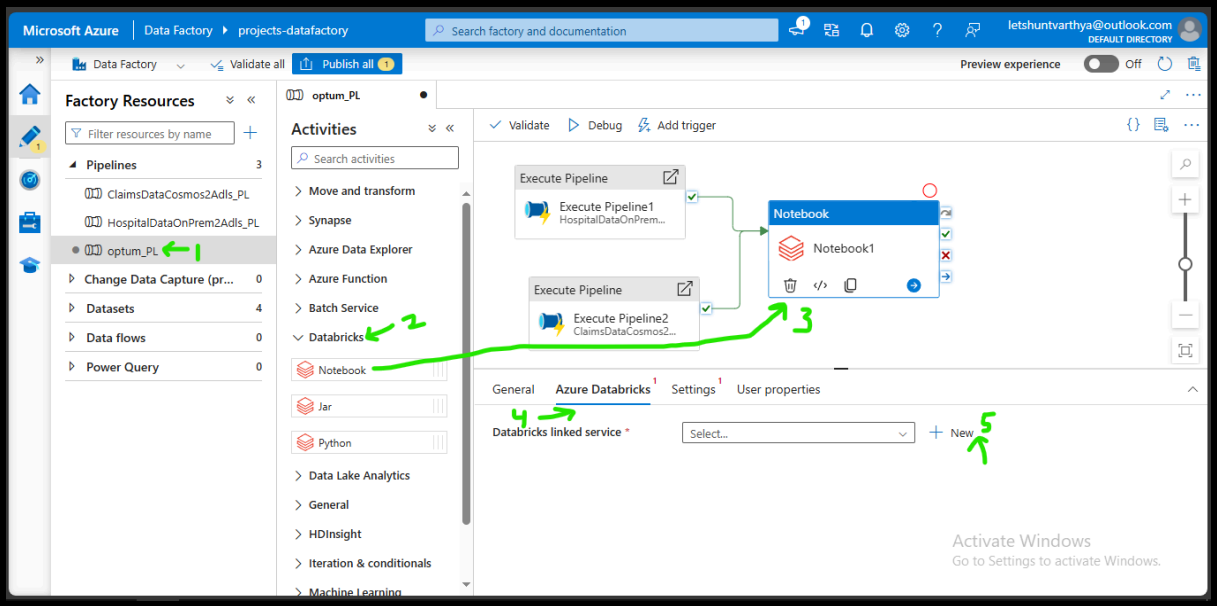
https://drive.google.com/file/d/1jEHawUK1OruV1ZEamZpxxtj4vliRhYUC/view?usp=sharing

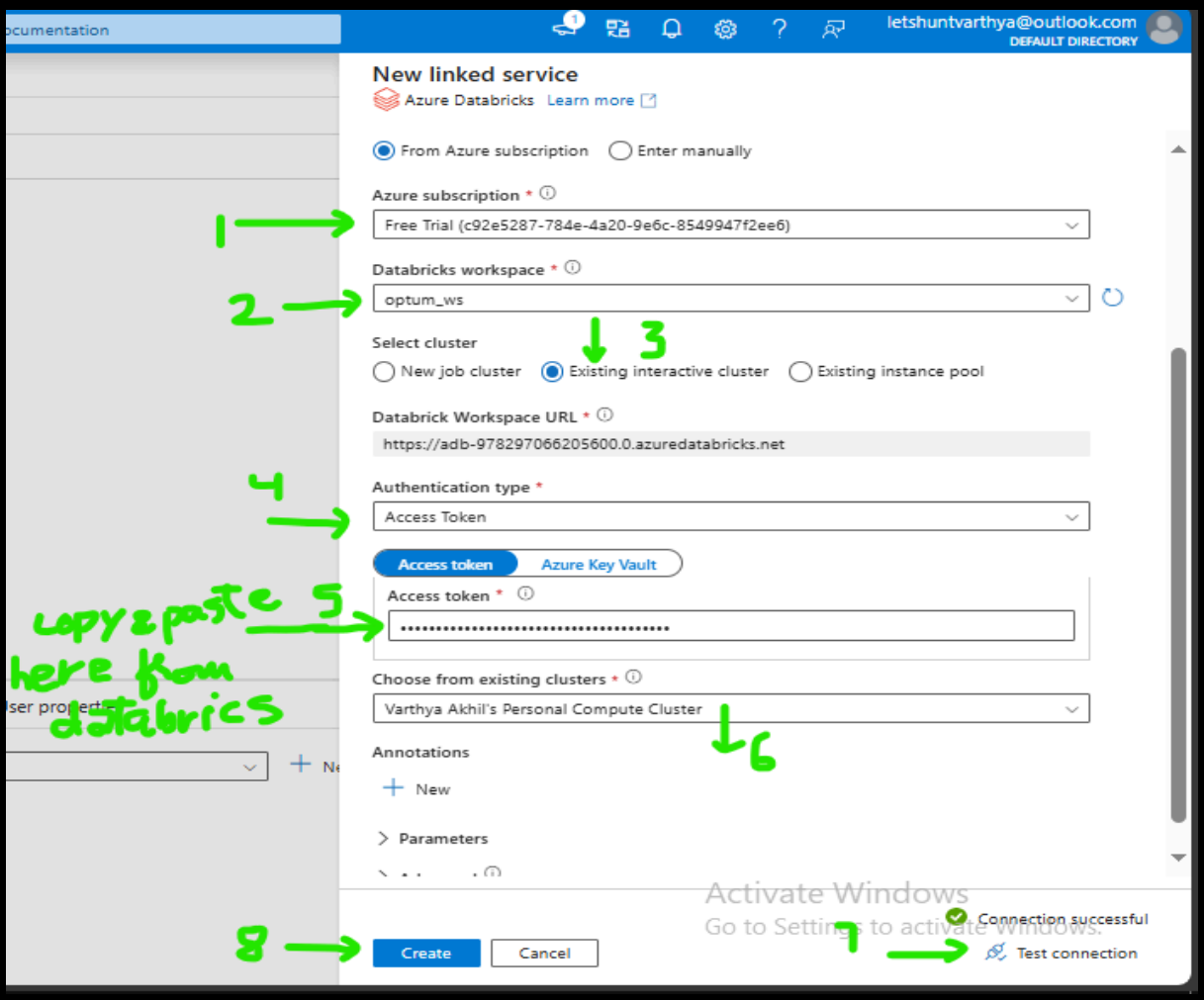
- Transformation of subscriber data is shown in the document below.
    - Create a new notebook (Subscriber_Transformation).
    https://drive.google.com/file/d/1reQ7WyybUi5TUqRrJ49SqCGHqKul2h2O/view?usp=sharing

- Transformation of claims data is shown in the document below.
    - Create a new notebook (Claims_Transformation).
    https://drive.google.com/file/d/1tJkE5jjdtaWFZ4osR4A1CDNaOu57rv1s/view?usp=sharing

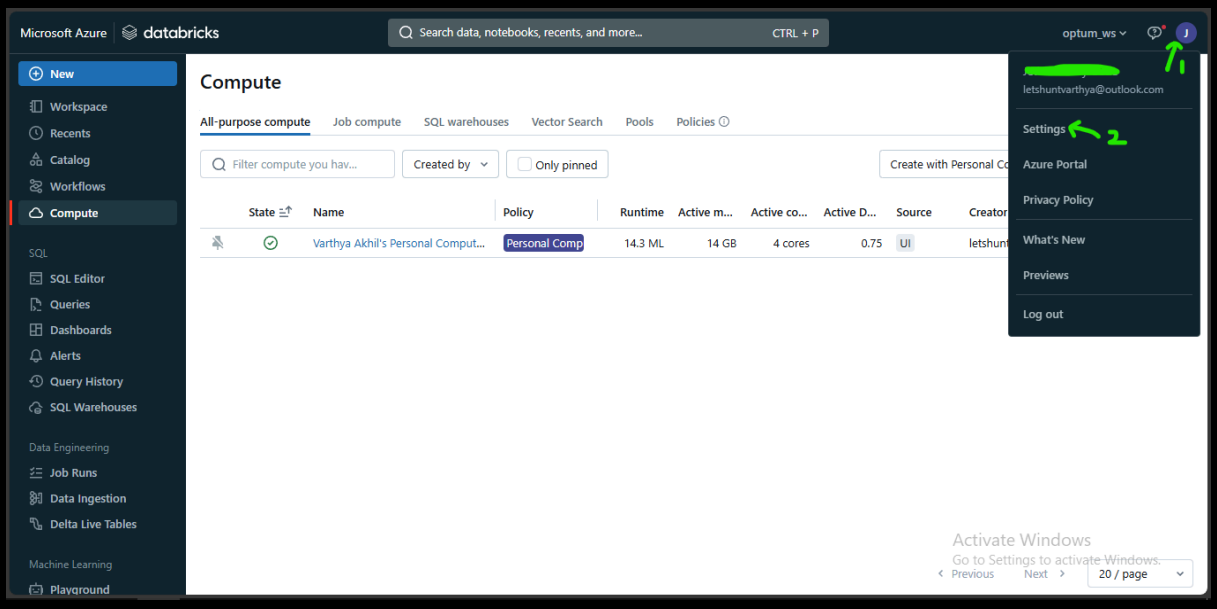
- We have executed this in the notebook itself, instead we can do it in the datafactory as well.

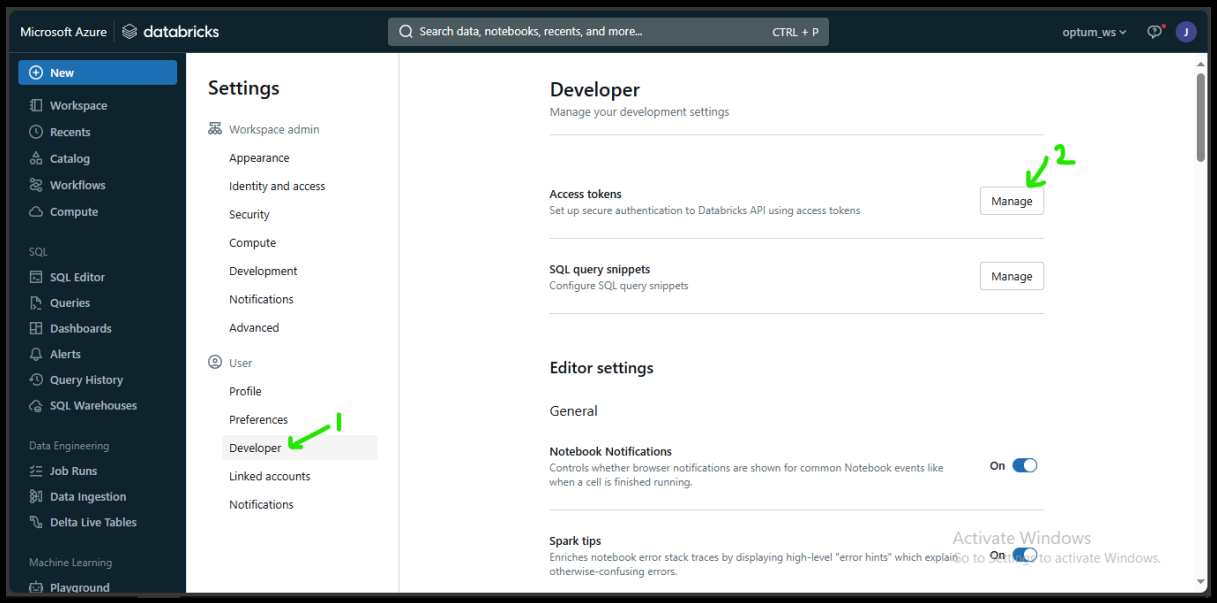
- Executing notebooks in the datafactory via access tokens.
- Drag and drop notebook activity in the optum_PL field.

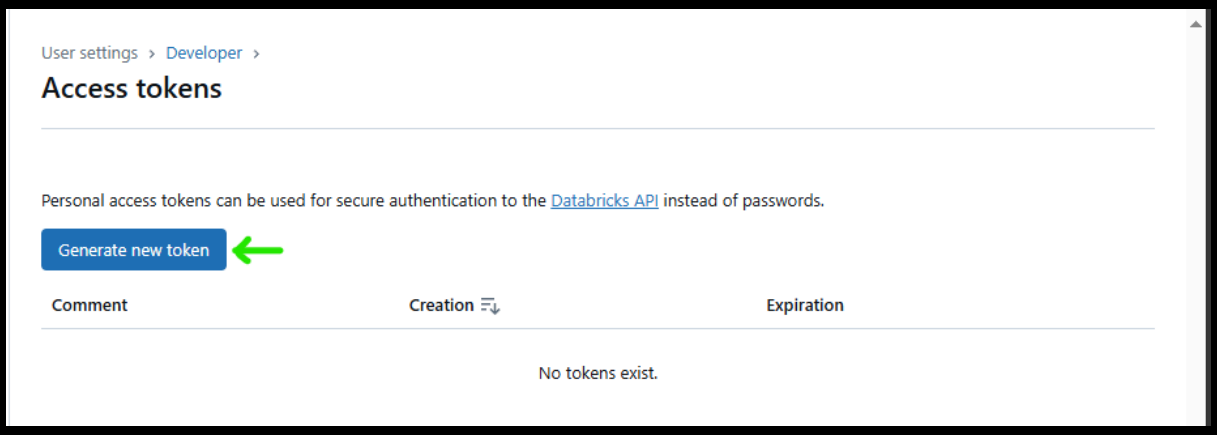


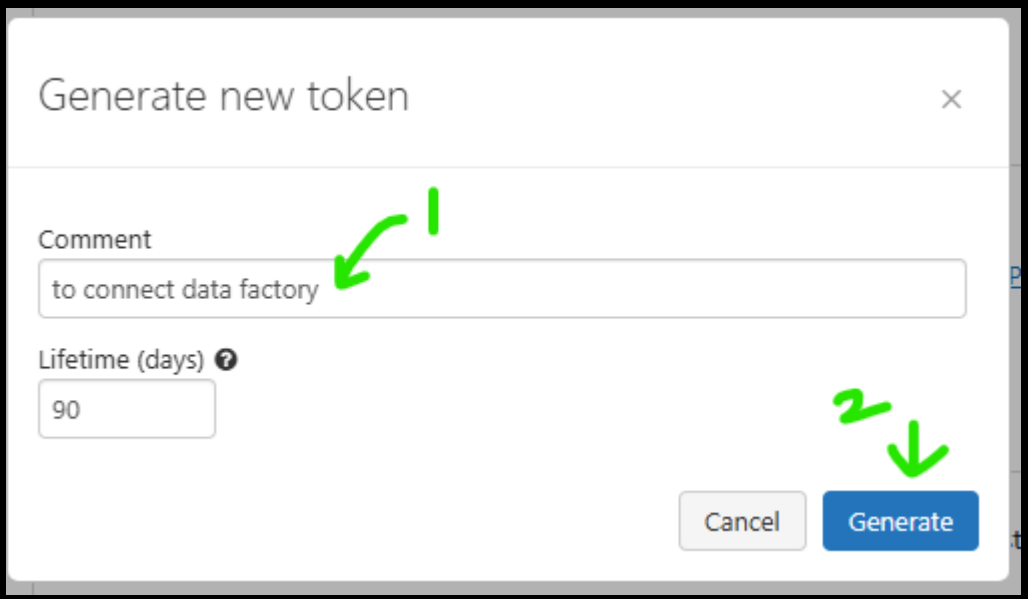


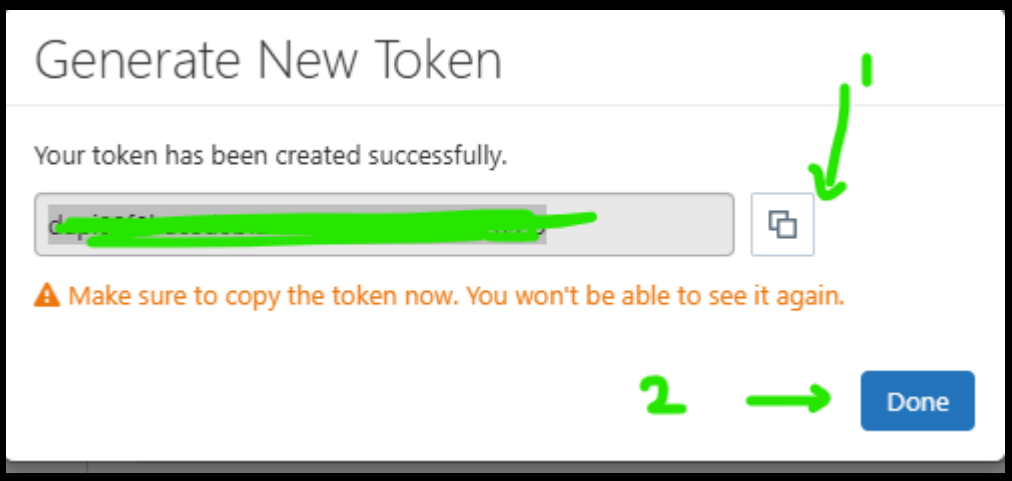
- While creating linked service access token is copied from databricks and steps are shown below.
- Connecting datafactory with databricks via access token.

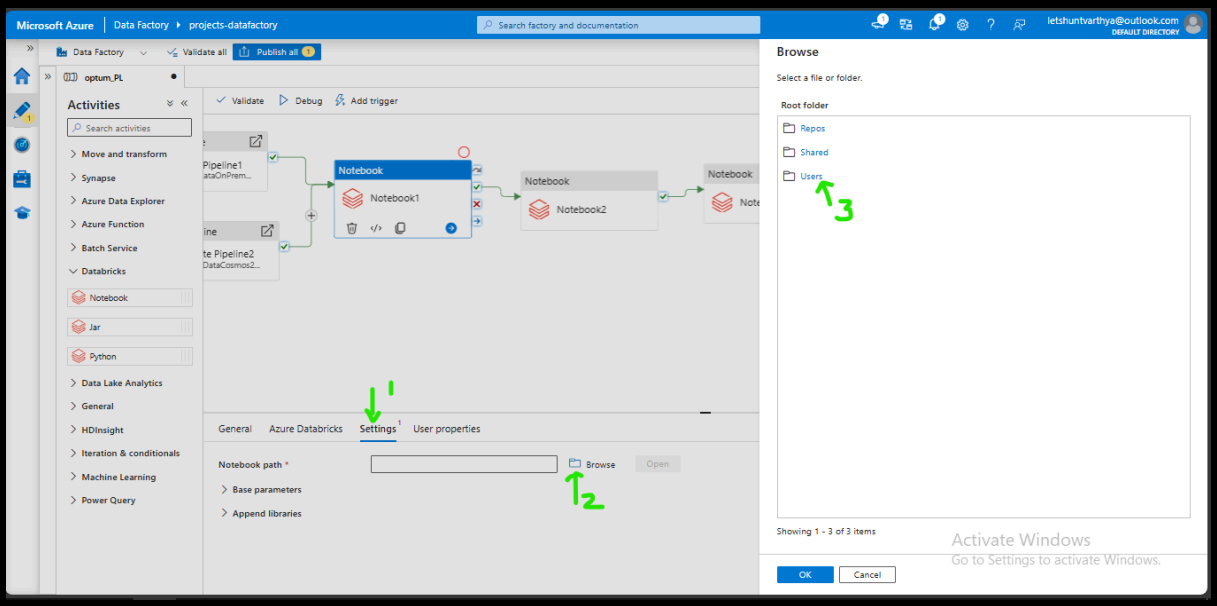


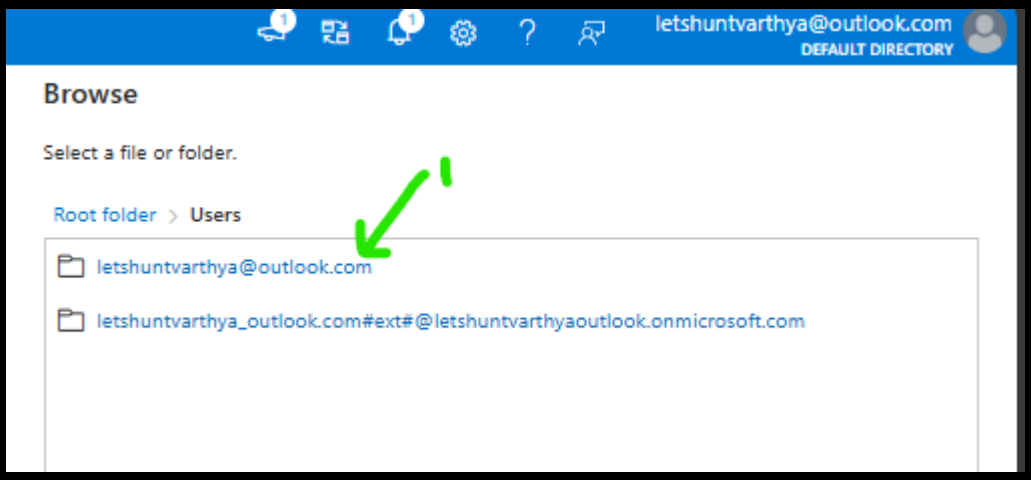

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
optum_ws
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Settings
Workspace admin
Appearance
Identity and access
Security
Compute
Development
Notifications
Advanced
User
Profile
Preferences
Developer
Linked accounts
Notifications
Developer
Manage your development settings
Access tokens
Set up secure authentication to Databricks API using access tokens
Manage
SQL query snippets
Configure SQL query snippets
Manage
Editor settings
General
Notebook Notifications
Controls whether browser notifications are shown for common Notebook events like when a cell is finished running.
On
Spark tips
Enriches notebook error stack traces by displaying high-level "error hints" which explain otherwise-confusing errors.
On
Activate Windows
1
2

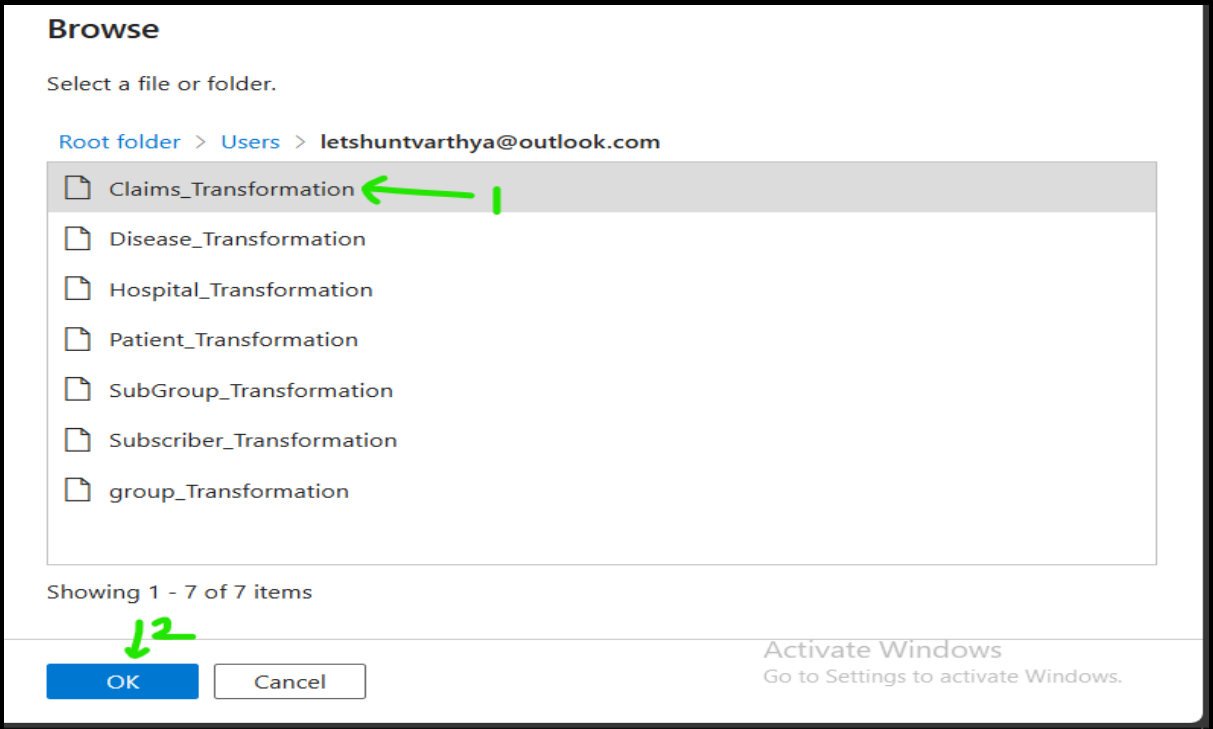

User settings › Developer ›
Access tokens
Personal access tokens can be used for secure authentication to the Databricks API instead of passwords.
Generate new token
Comment
Creation
Expiration
No tokens exist.


Generate new token
Comment
to connect data factory
Lifetime (days)
90
Cancel
Generate
1
2


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.
1
2
Done

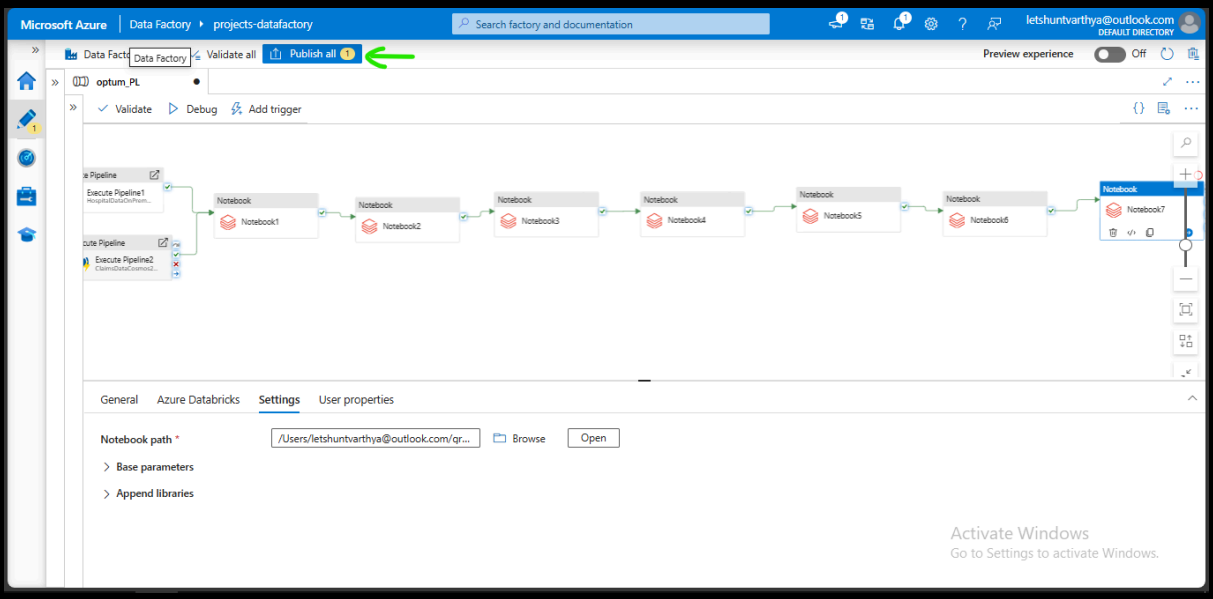

Microsoft Azure
Data Factory
projects-datafactory
Search factory and documentation
letshuntvarthya@outlook.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all
optum_PL
Validate
Debug
Add trigger
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Notebook
Jar
Python
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Pipeline1
Notebook
Notebook1
Notebook2
General
Azure Databricks
Settings
User properties
Notebook path *
Browse
Open
Base parameters
Append libraries
Browse
Select a file or folder.
Root folder
Repos
Shared
Users
Showing 1 - 3 of 3 items
OK
Cancel
Activate Windows
Go to Settings to activate Windows.
1
2
3

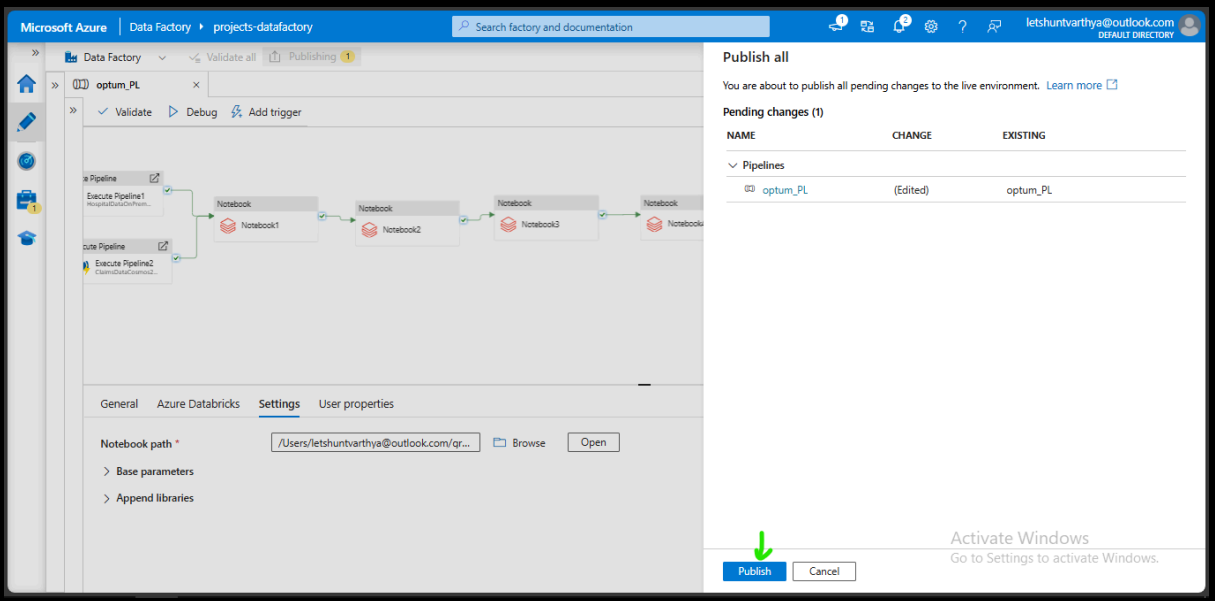

letshuntvarthya@outlook.com
DEFAULT DIRECTORY
Browse
Select a file or folder.
Root folder
Users
letshuntvarthya@outlook.com
letshuntvarthya_outlook.com#ext#@letshuntvarthyaoutlook.onmicrosoft.com
1

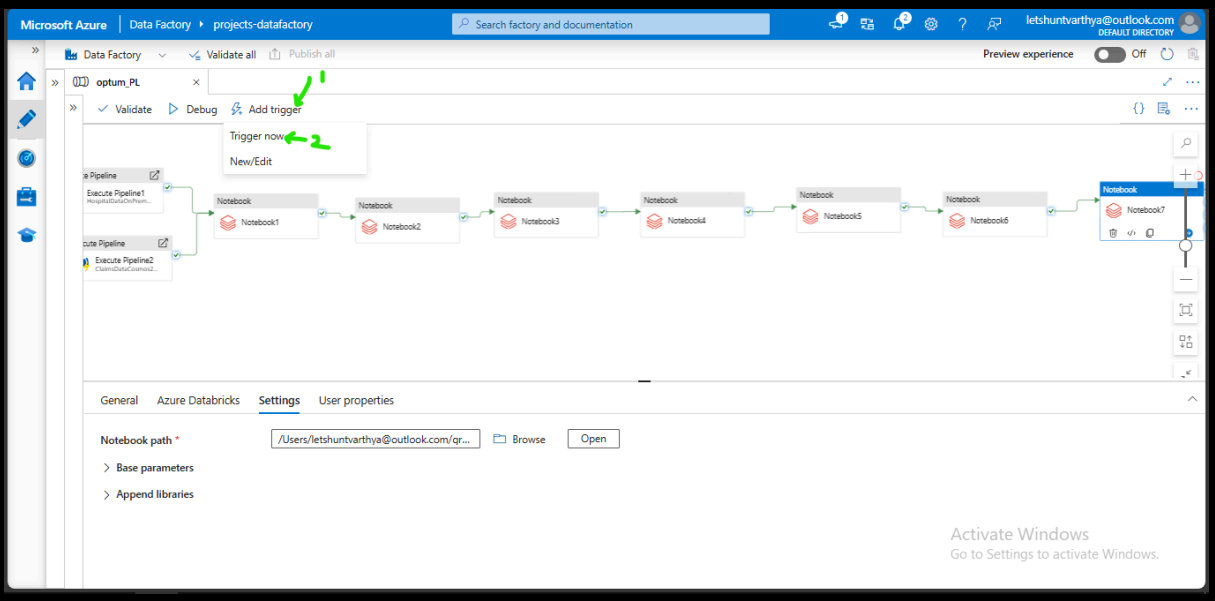


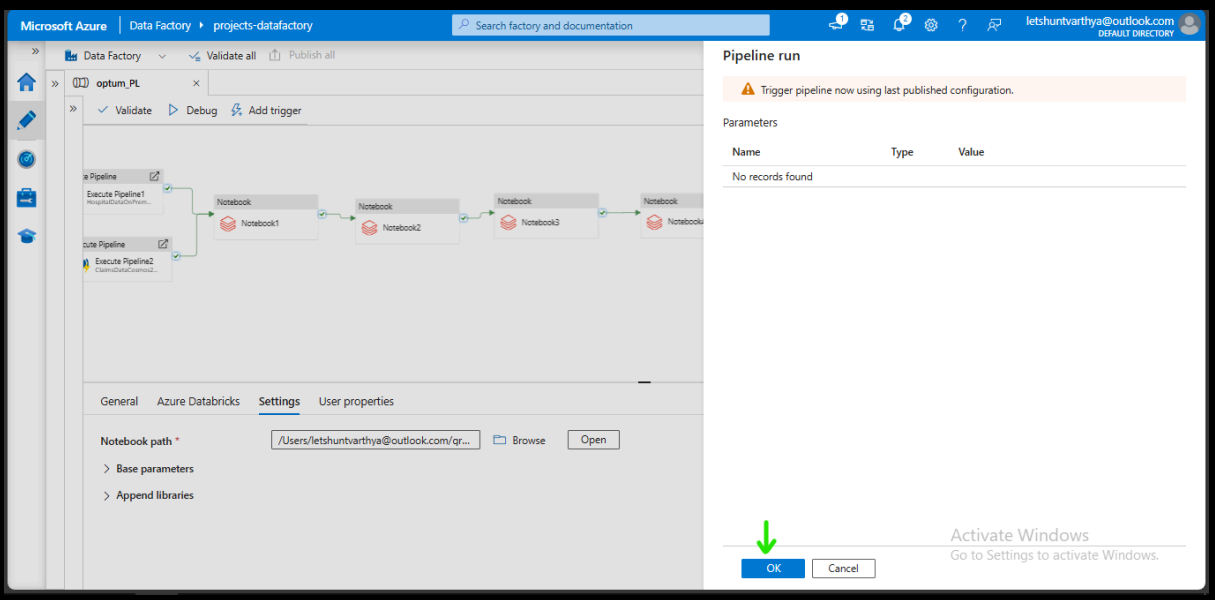
- Do this process for other notebooks (Disease_Transformation, Hospital_Transformation, Patient_Transformation, SubGroup_Transformation, Subscriber_Transformations and Group_Transformations) as well.

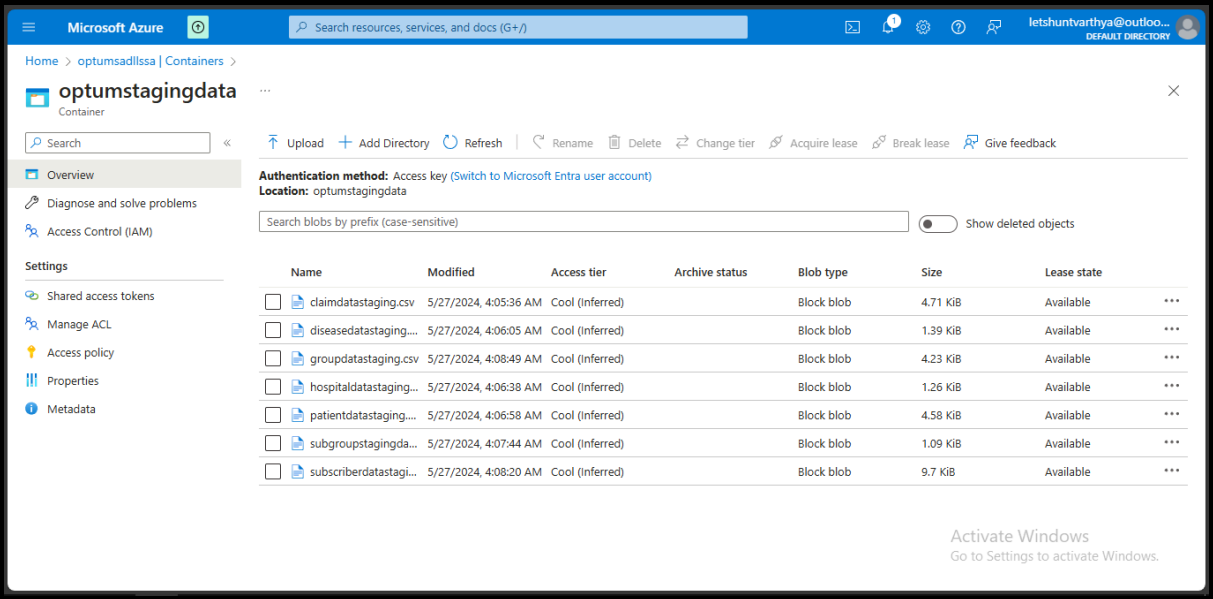
1. Drag and drop notebook activity to the field.
2. Select the linked service that we created just now.
3. Select the notebook



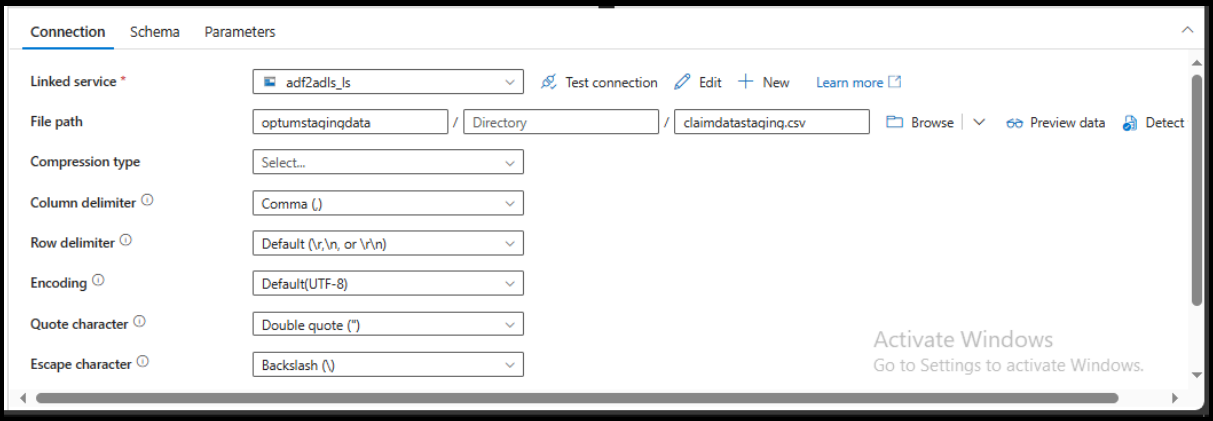

Microsoft Azure
Data Factory
projects-datafactory
Search factory and documentation
letshuntvarthya@outlook.com
DEFAULT DIRECTORY
Validate all
Publishing 1
optum_PL
Validate
Debug
Add trigger
Publish all
You are about to publish all pending changes to the live environment. Learn more
Pending changes (1)
NAME
CHANGE
EXISTING
Pipelines
optum_PL
(Edited)
optum_PL
General
Azure Databricks
Settings
User properties
Notebook path *
/Users/letshuntvarthya@outlook.com/qr...
Browse
Open
Base parameters
Append libraries
Publish
Cancel
Activate Windows
Go to Settings to activate Windows.

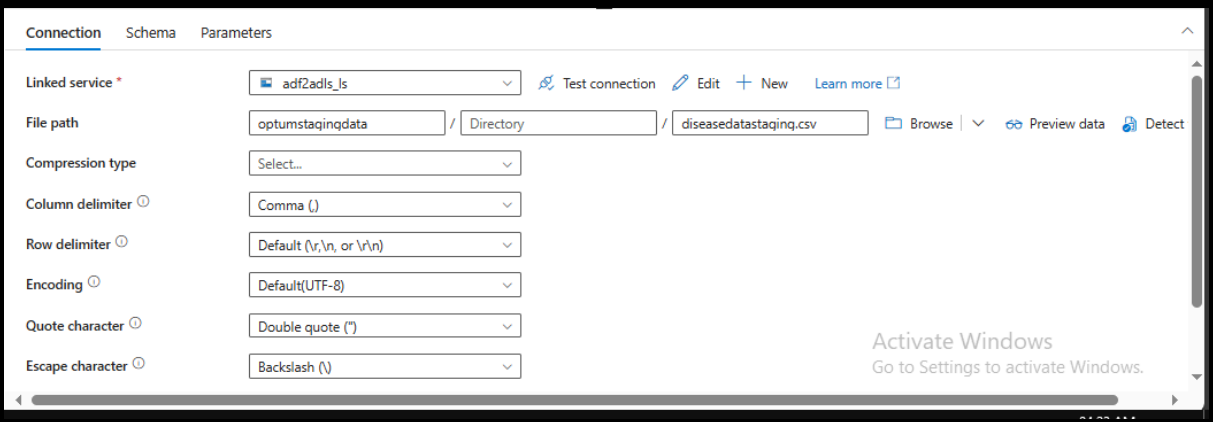

Microsoft Azure
Data Factory
projects-datafactory
Search factory and documentation
letshuntvarthya@outlook.com
DEFAULT DIRECTORY
Validate all
Publish all
Preview experience
Off
optum_PL
Validate
Debug
Add trigger
1
Trigger now
2
New/Edit
Notebook1
Notebook2
Notebook3
Notebook4
Notebook5
Notebook6
Notebook7
General
Azure Databricks
Settings
User properties
Notebook path *
/Users/letshuntvarthya@outlook.com/qr...
Browse
Open
Base parameters
Append libraries
Activate Windows
Go to Settings to activate Windows.

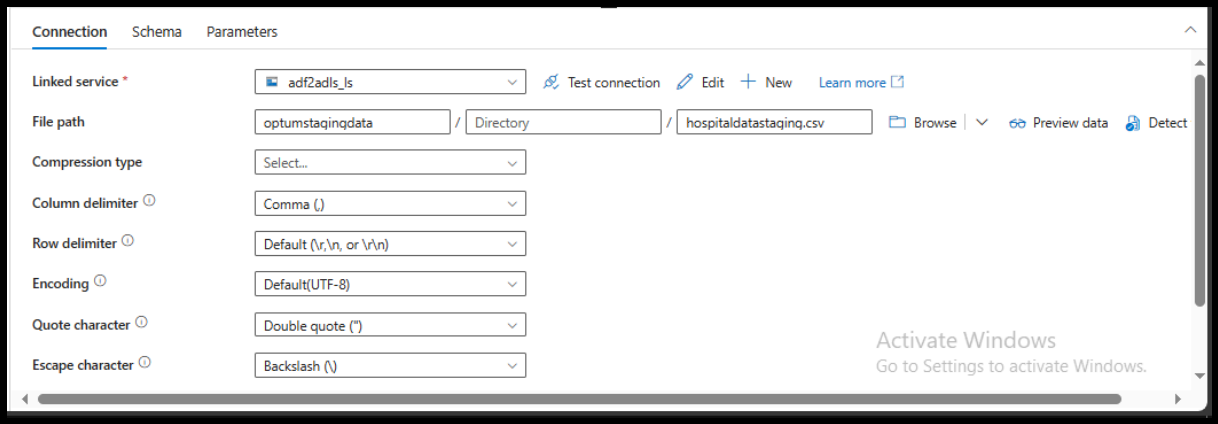

Microsoft Azure
Data Factory
projects-datafactory
Search factory and documentation
letshuntvarthya@outlook.com
DEFAULT DIRECTORY
Validate all
Publish all
optum_PL
Validate
Debug
Add trigger
Pipeline run
Trigger pipeline now using last published configuration.
Parameters
Name
Type
Value
No records found
General
Azure Databricks
Settings
User properties
Notebook path *
/Users/letshuntvarthya@outlook.com/qr...
Browse
Open
Base parameters
Append libraries
OK
Cancel
Activate Windows
Go to Settings to activate Windows.

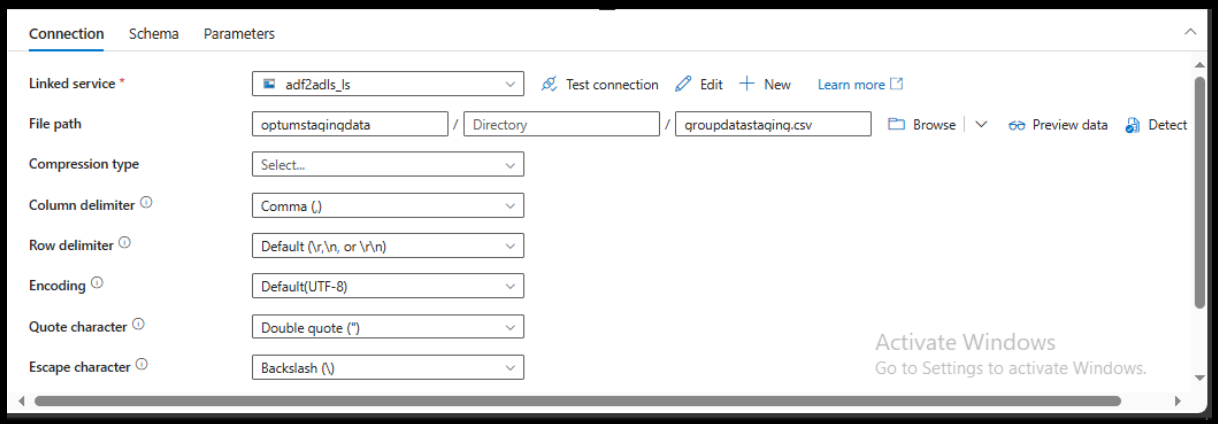


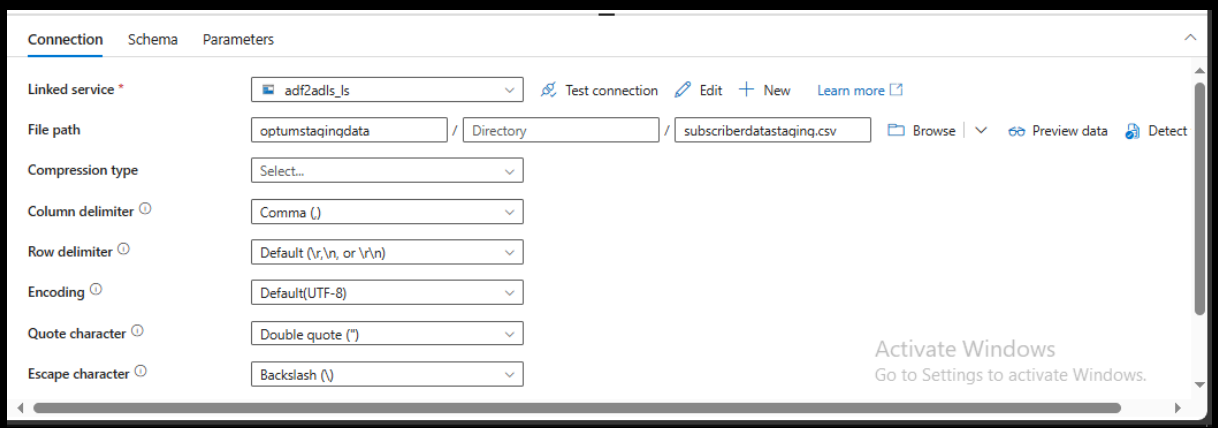
- The basic transformed data has been brought to optumstagingdata.
- Now let's do Final transformations using datafactory.
- Create datasets for this stagingdata.

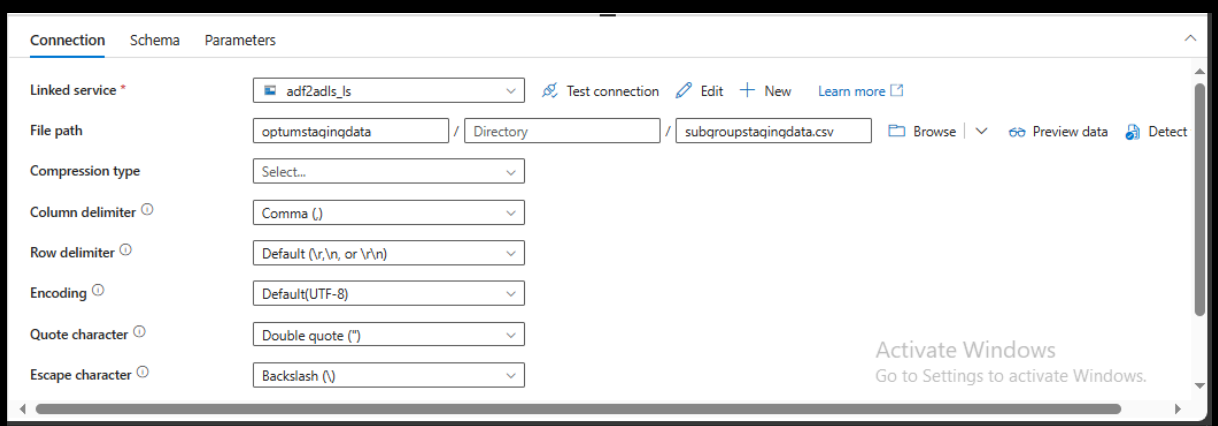


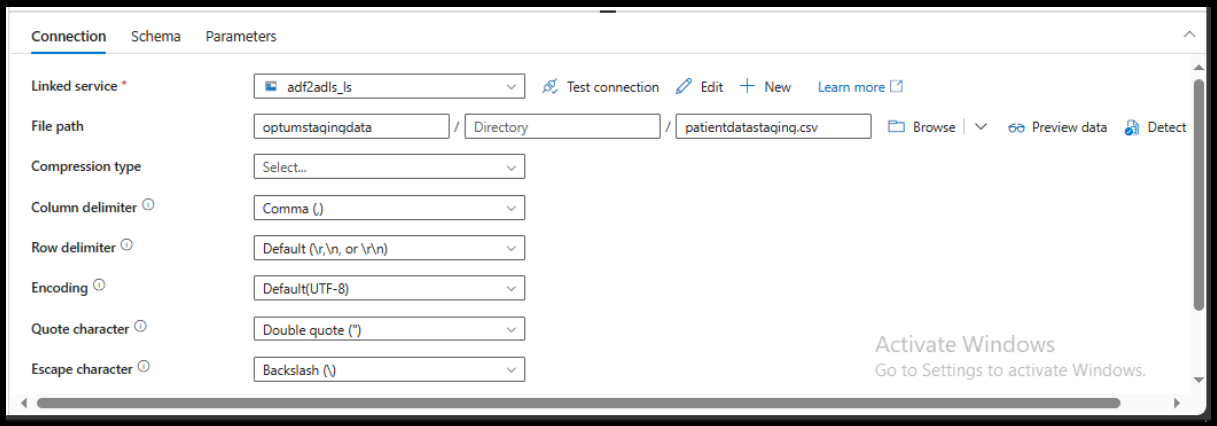



Connection
Schema
Parameters
Linked service *
adf2adls_ls
Test connection
Edit
New
Learn more
File path
optumstaqingdata
Directory
hospitaldatastaqing.csv
Browse
Preview data
Detect
Compression type
Select...
Column delimiter
Comma (,)
Row delimiter
Default (\r,\n, or \r\n)
Encoding
Default(UTF-8)
Quote character
Double quote (")
Escape character
Backslash (\)
Activate Windows
Go to Settings to activate Windows.

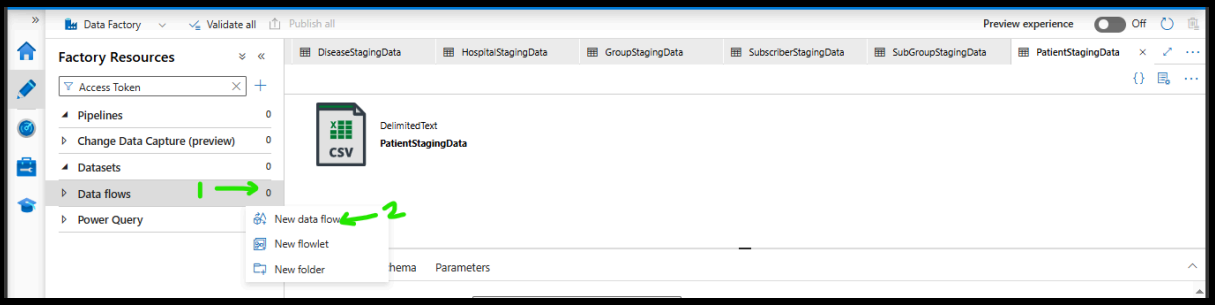

Connection
Schema
Parameters
Linked service *
adf2adls_ls
Test connection
Edit
New
Learn more
File path
optumstaqingdata
Directory
groupdatastaqing.csv
Browse
Preview data
Detect
Compression type
Select...
Column delimiter
Comma (,)
Row delimiter
Default (\r,\n, or \r\n)
Encoding
Default(UTF-8)
Quote character
Double quote (")
Escape character
Backslash (\)
Activate Windows
Go to Settings to activate Windows.

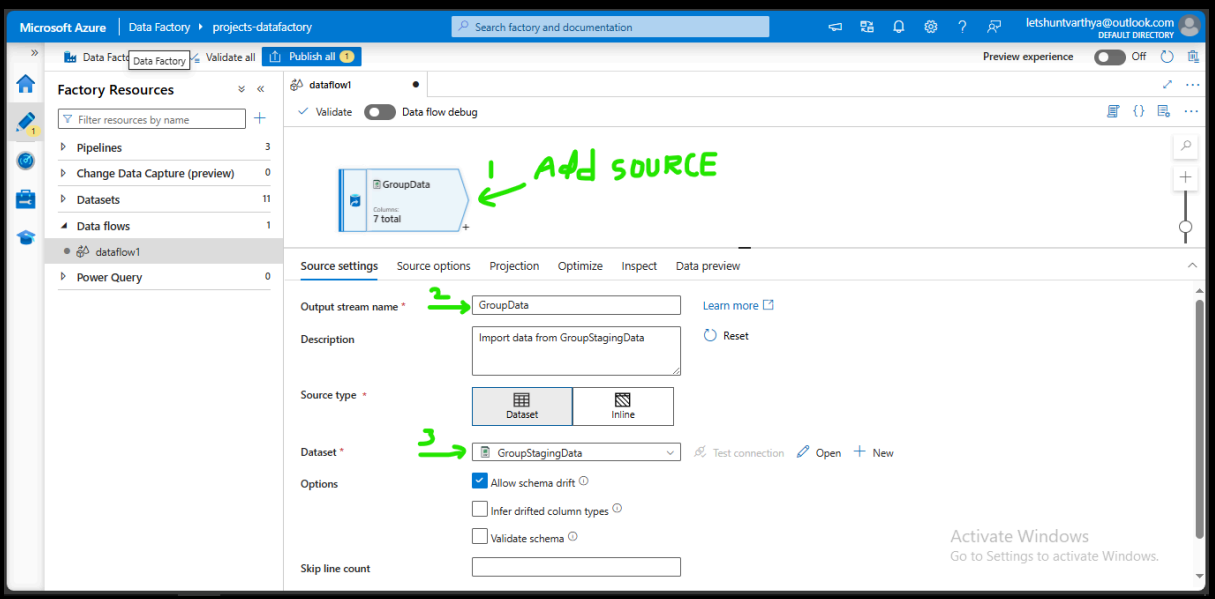

Connection
Schema
Parameters
Linked service *
adf2adls_ls
Test connection
Edit
New
Learn more
File path
optumstaqingdata
Directory
subscriberdatastaqing.csv
Browse
Preview data
Detect
Compression type
Select...
Column delimiter
Comma (,)
Row delimiter
Default (\r,\n, or \r\n)
Encoding
Default(UTF-8)
Quote character
Double quote (")
Escape character
Backslash (\)
Activate Windows
Go to Settings to activate Windows.


Connection
Schema
Parameters
Linked service *
adf2adls_ls
Test connection
Edit
New
Learn more
File path
optumstaqingdata
Directory
subgroupstaqingdata.csv
Browse
Preview data
Detect
Compression type
Select...
Column delimiter
Comma (,)
Row delimiter
Default (\r,\n, or \r\n)
Encoding
Default(UTF-8)
Quote character
Double quote (")
Escape character
Backslash (\)
Activate Windows
Go to Settings to activate Windows.

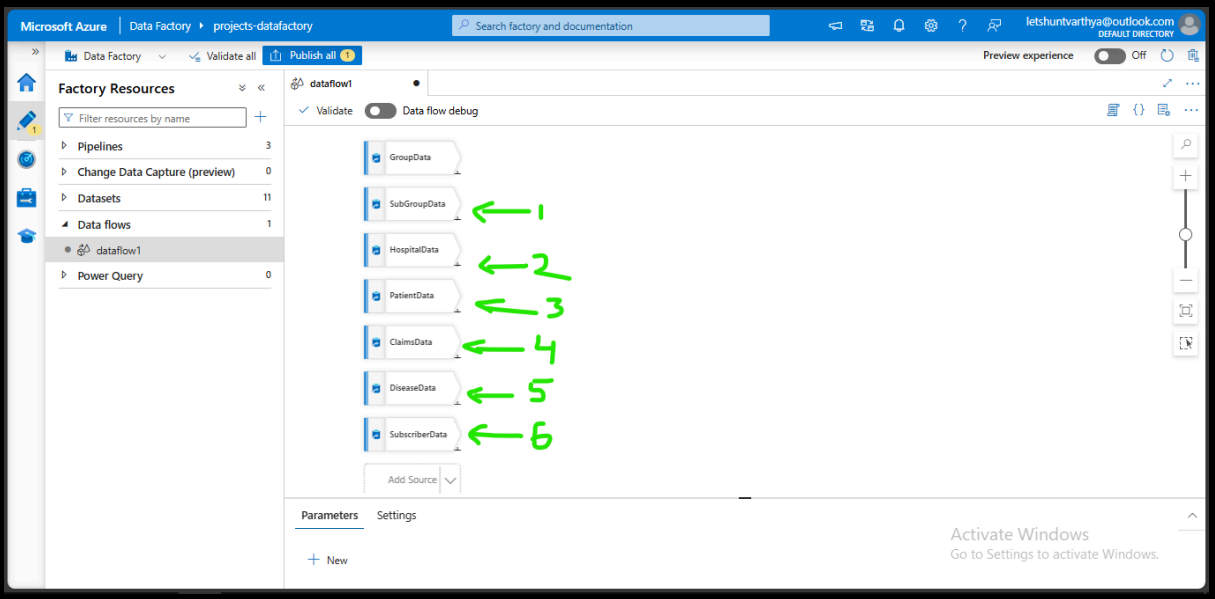


Create dataflow to do transformations.

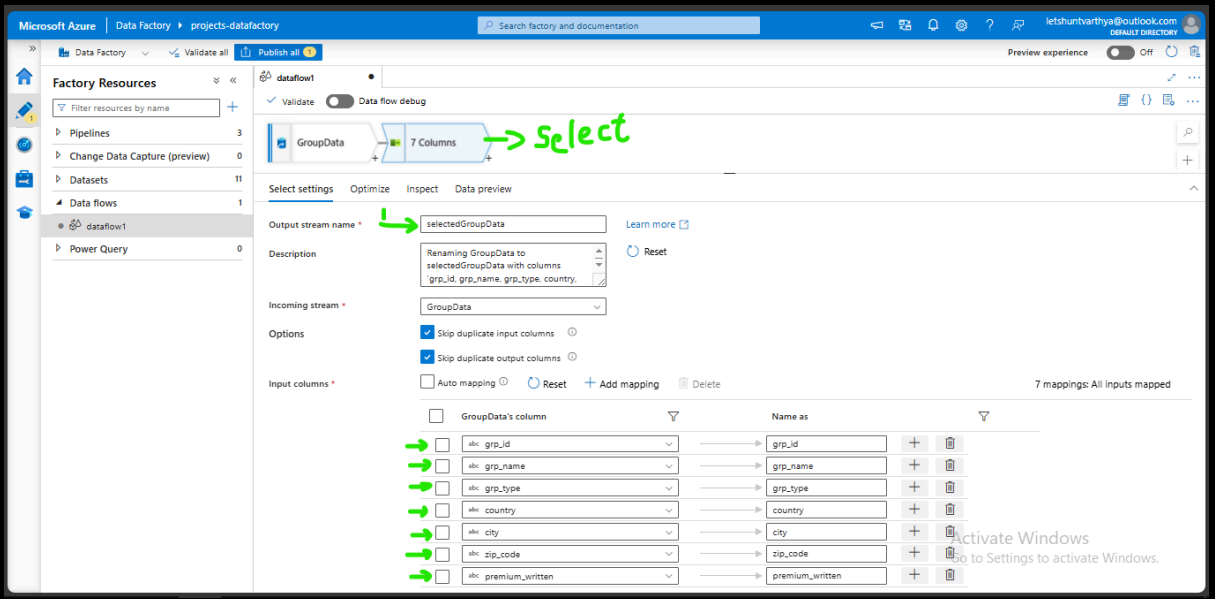




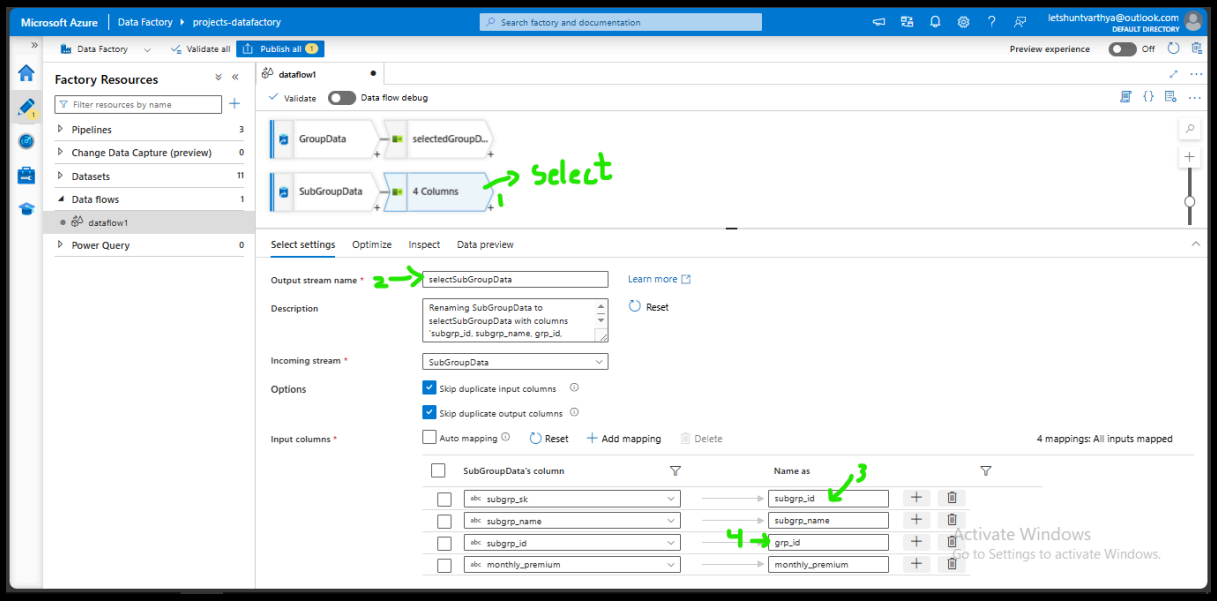
- Add sources like subgroup_data, hospital_data, patient_data, claims_data, disease_data and subscriber_data.



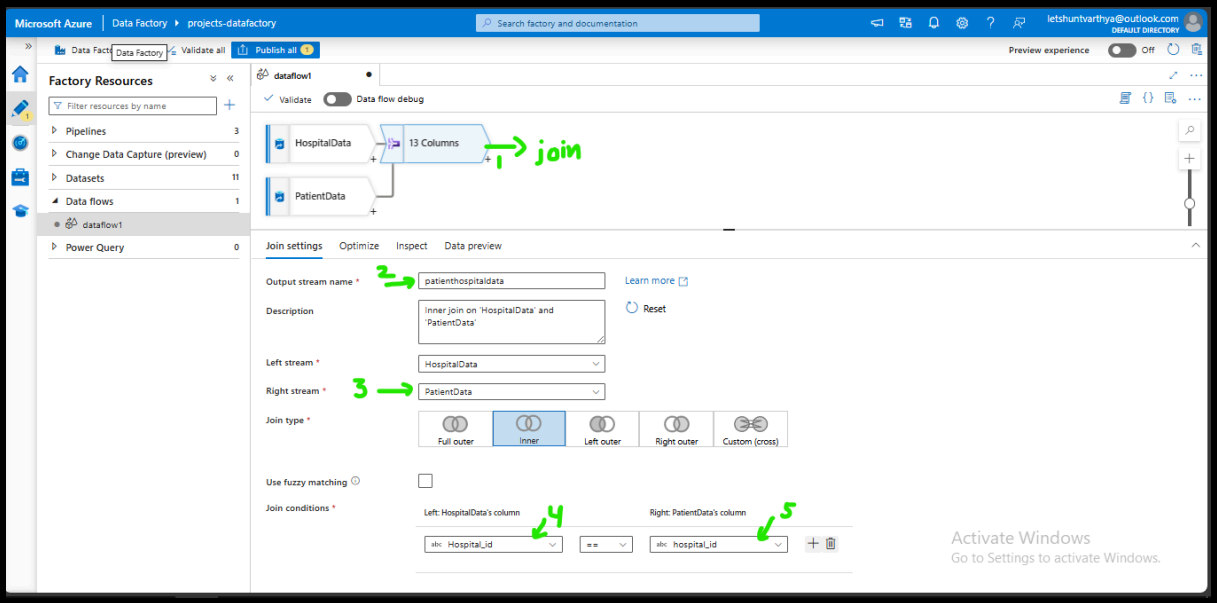
- Add select after group_data and rearrange properly.

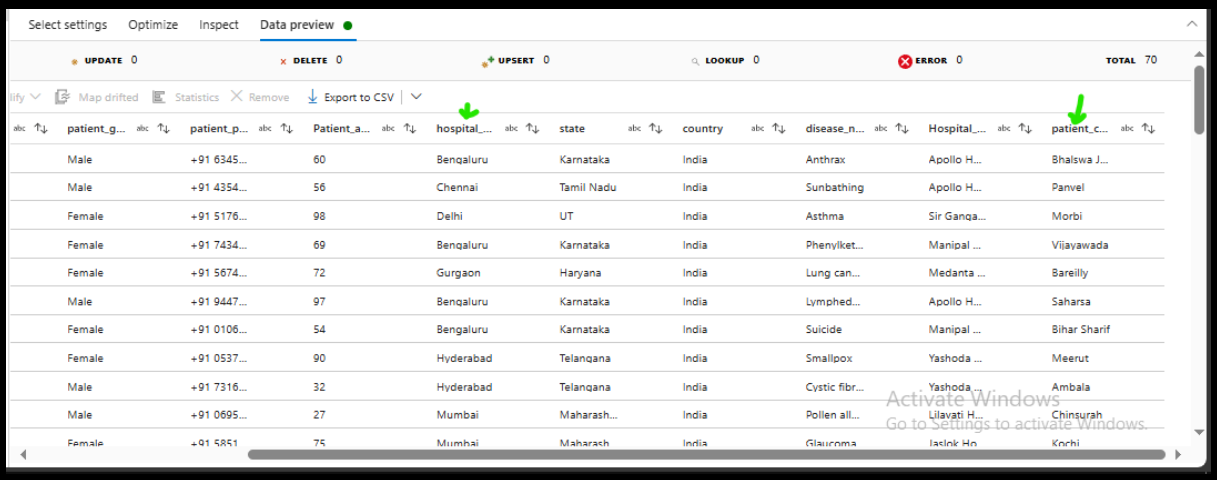


- Add select after subgroup_data, rename subgrp_sk to subgrp_id ,subgrp_id to grp_id and rearrange properly.



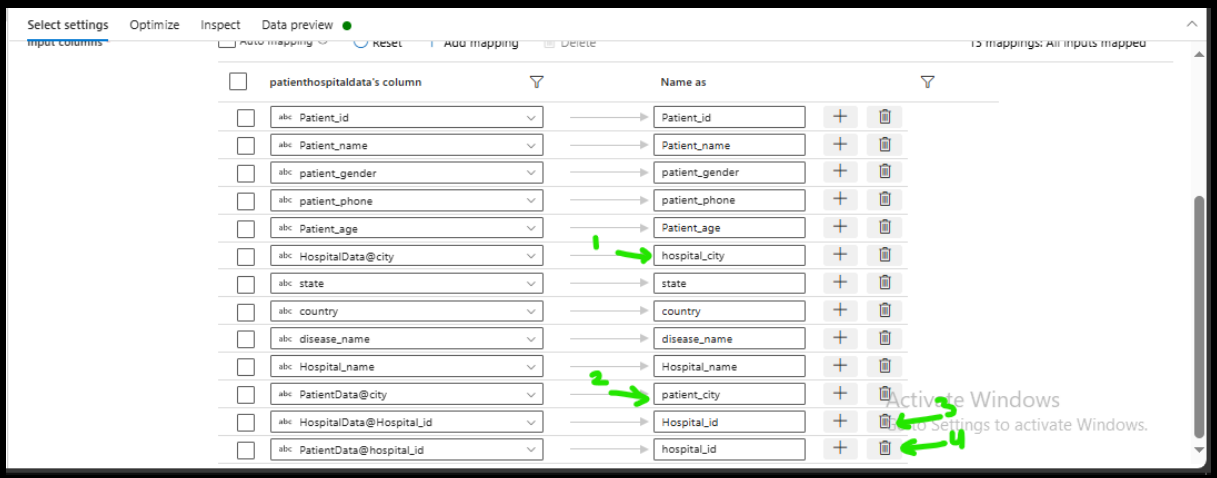
- Joining hospital_data with patient_data.



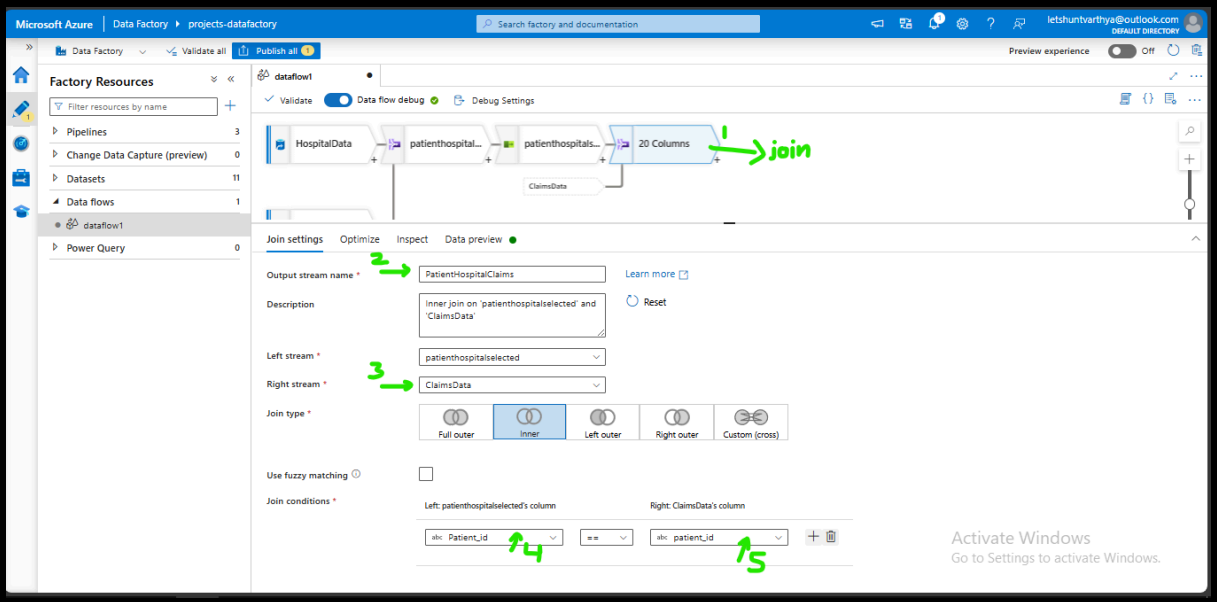


- Add select after this join activity.
- Output stream name : patienthospitalselected

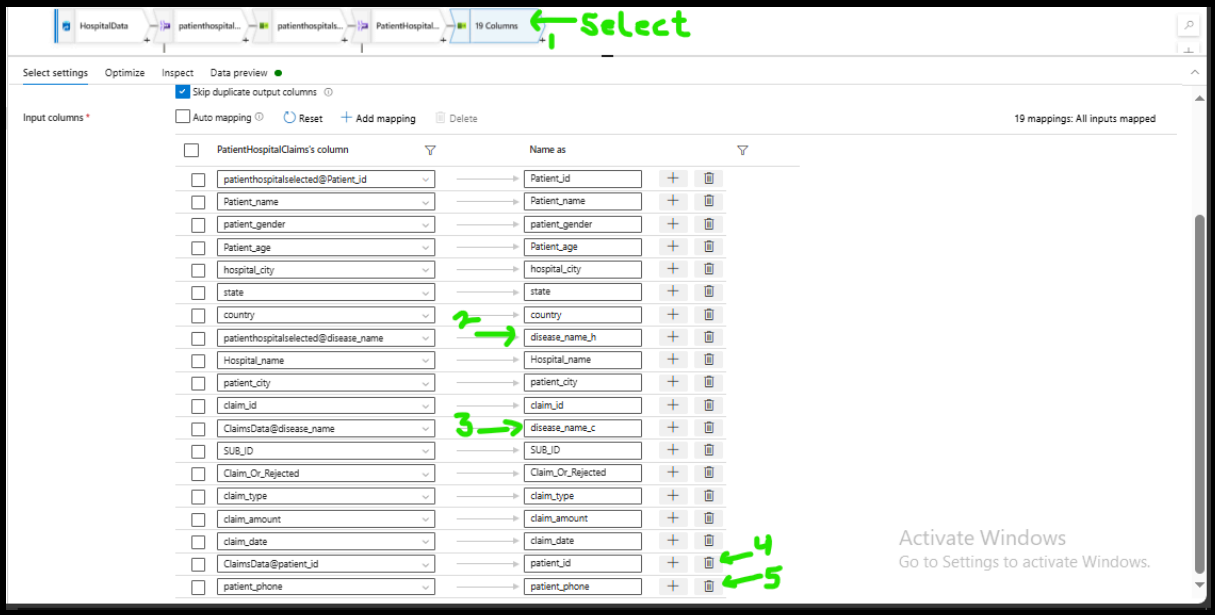
- Delete HospitalData@Hospital_id column
- Delete HospitalData@Hospital_id column
- Rename city (hopspitaldata) to hospital_city
- Rename city (patientdata) to patient_city
- Rearrange columns properly



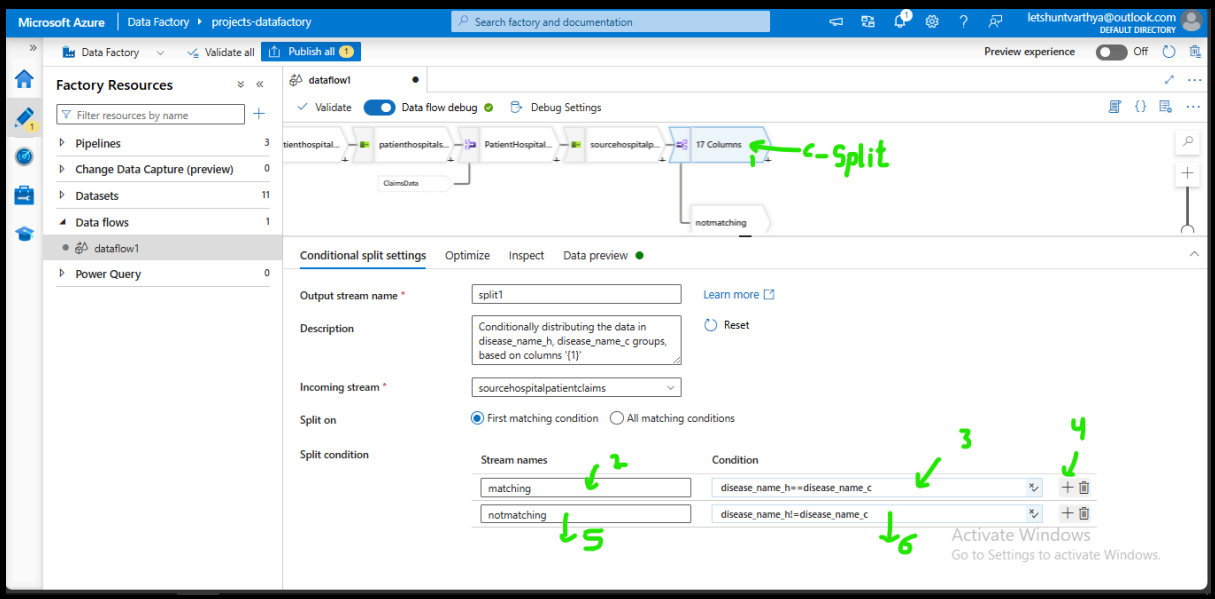
- Joining patienthospitalselected with claims_data

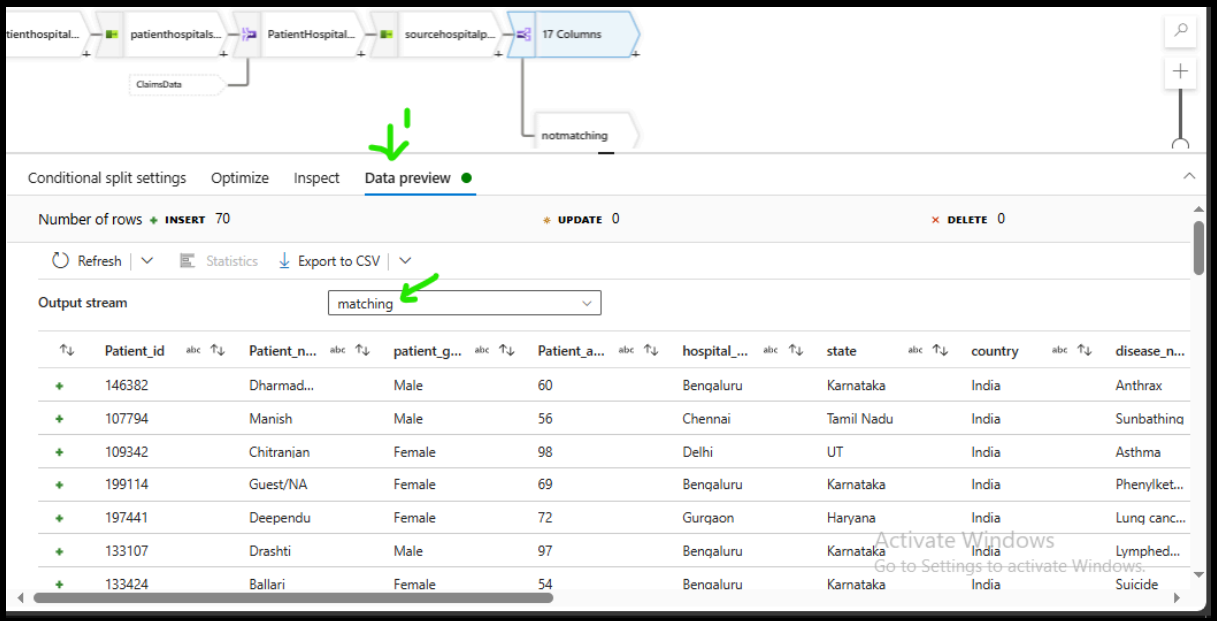


- Add select after this join activity (PatientHospitalClaims).
- Output stream name : sourcehospitalpatientclaims
- Delete patientdata@patient_phone column
- Delete ClaimsData@patient_id column
- Checking hospitalselected@disease_name and patienthospitalselected@disease_name are by renaming it to diseasename_h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

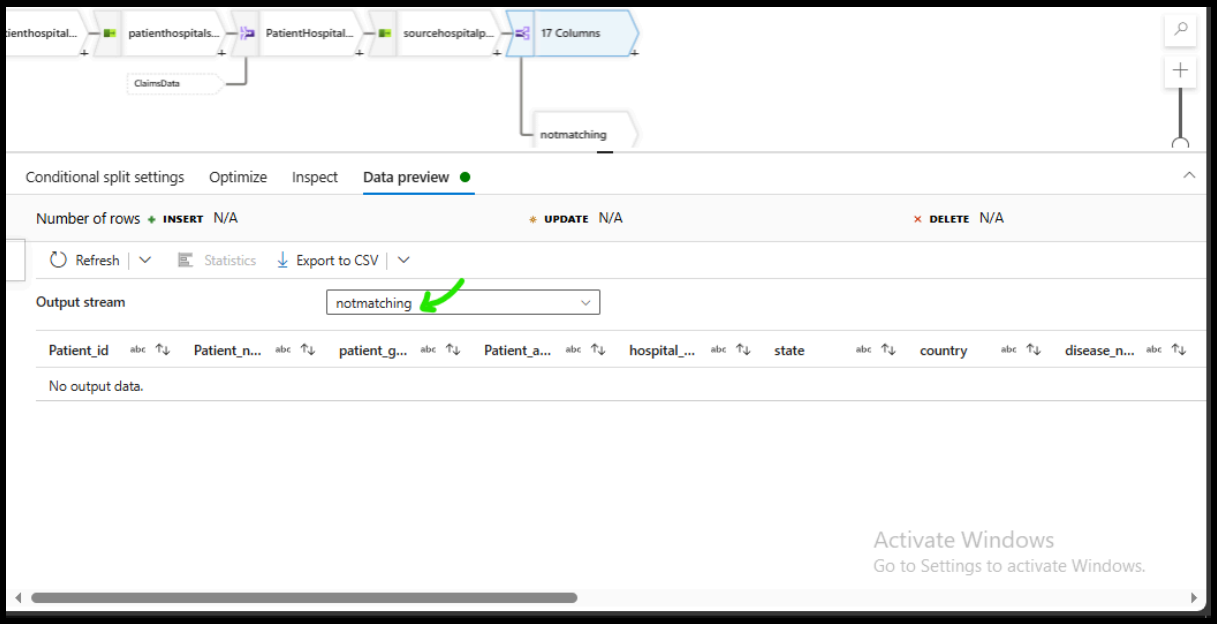


- Let's check whether diseasename_h and diseasename_p are the same by adding conditional split activity after this sourcehospitalpatientclaims.

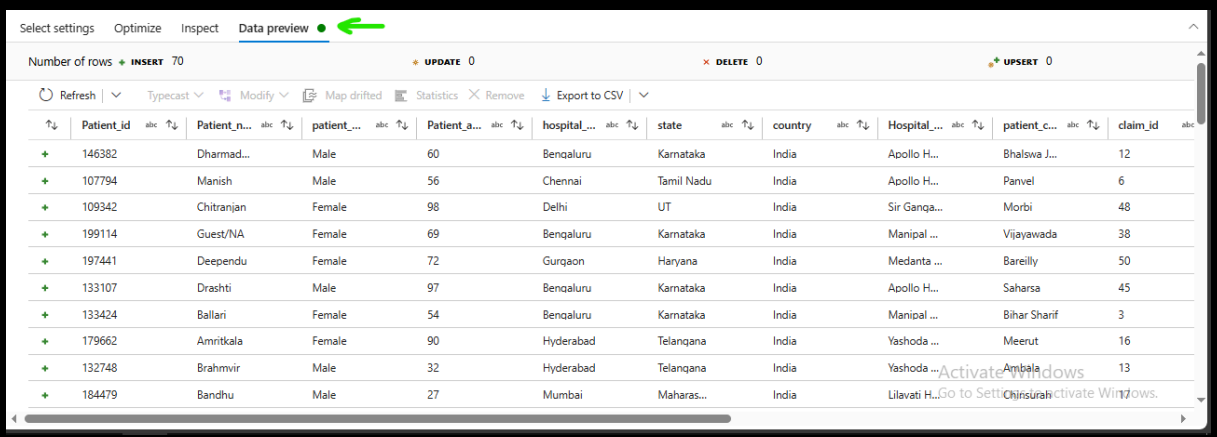




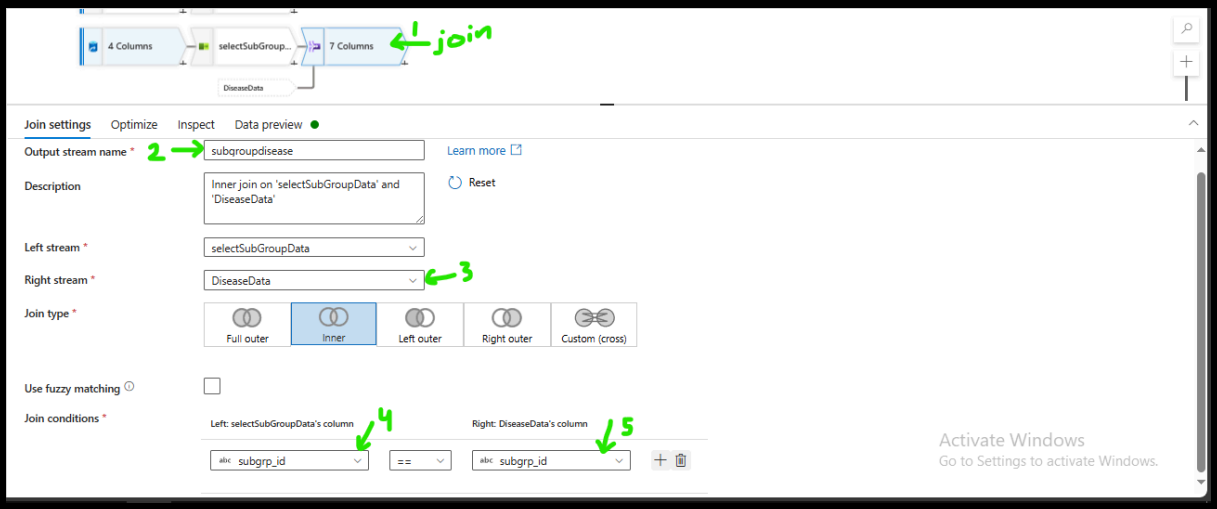
- From below image we can see that there is no non-matching records



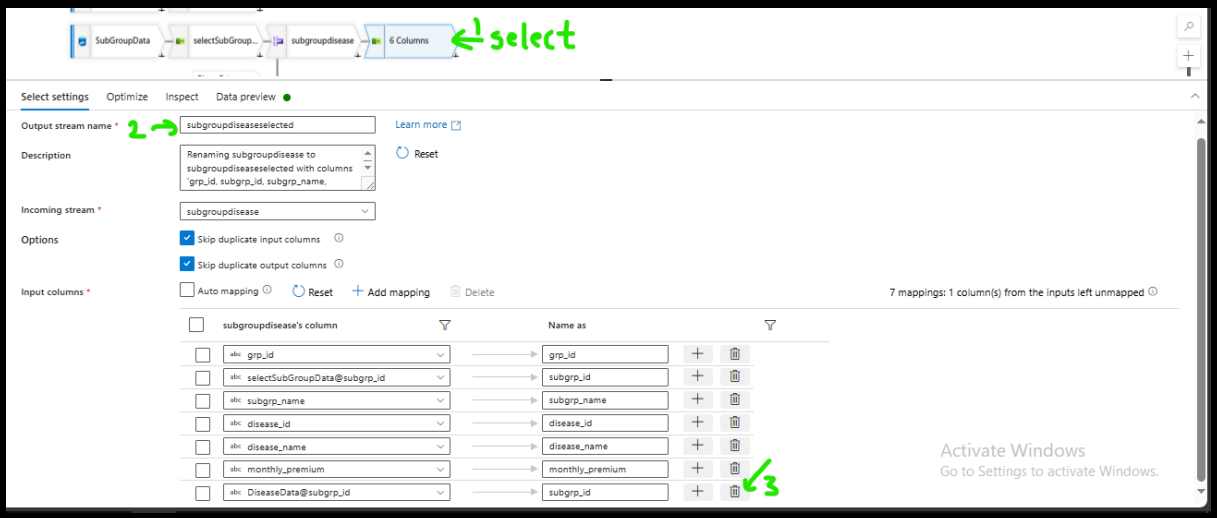
- Delete conditional split activity.
- Delete disease_name_h column and rename disease_name_c to disease_name.

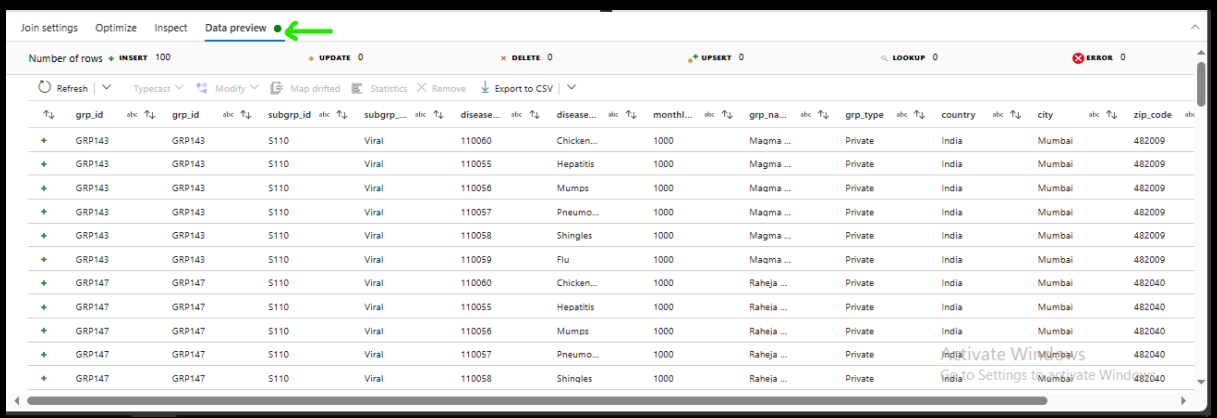


- Joining subGroupData with disease_data.
- Output stream name : subgroupdisease



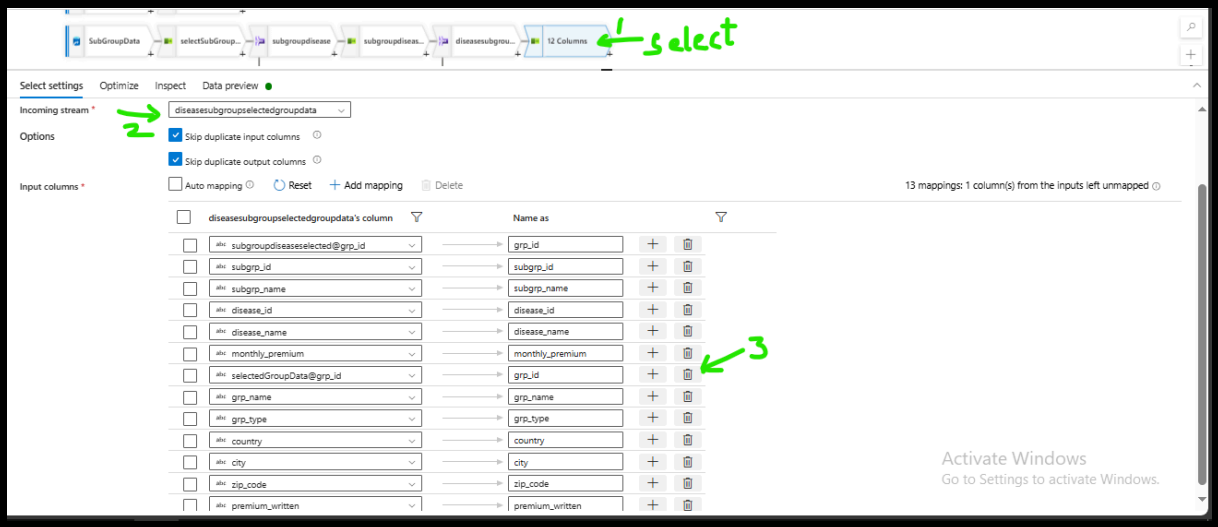
- Add select after this join and delete DiseaseData@subgrp_id column



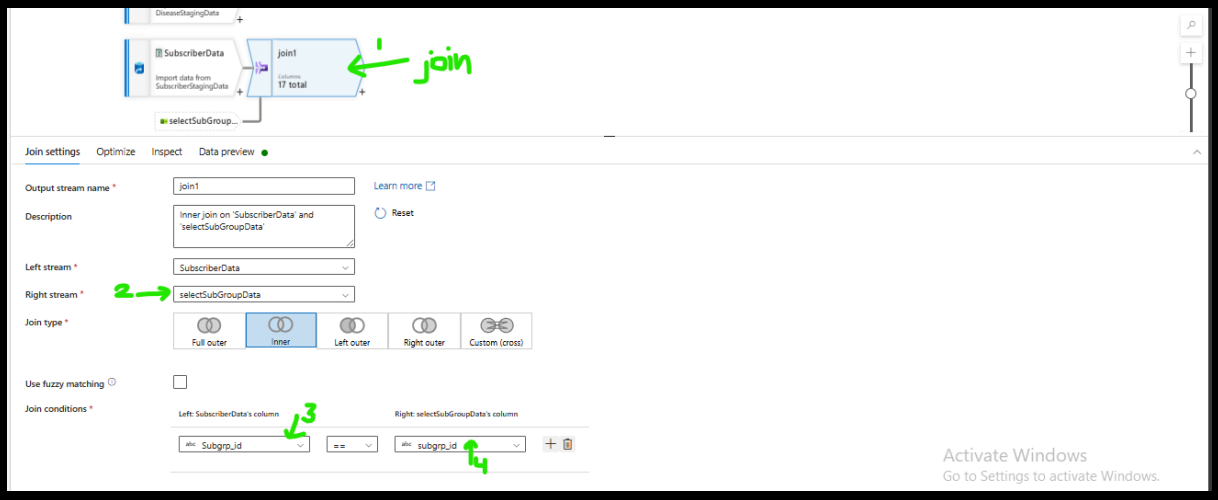


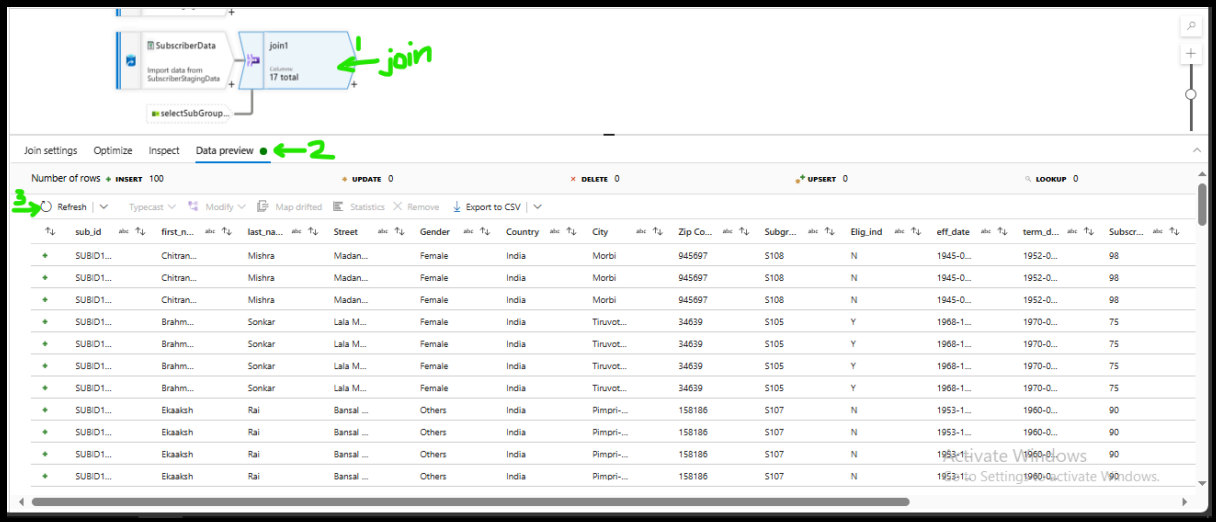
- Joining subgroupdiseaseselected with selectedgroupdata.

- Add select activity after diseasesubgroupselectedgroupdata.

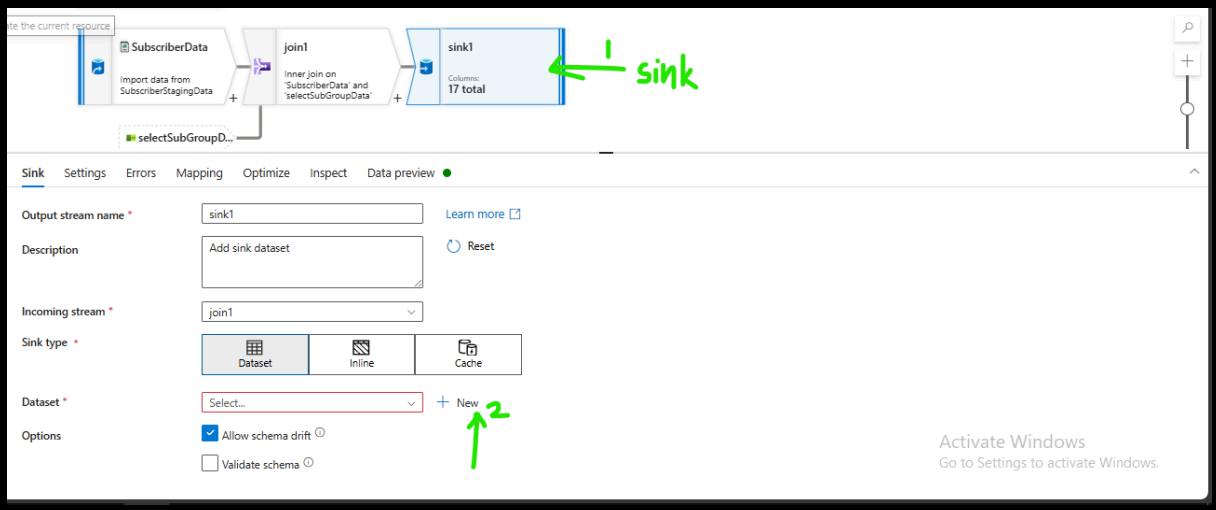


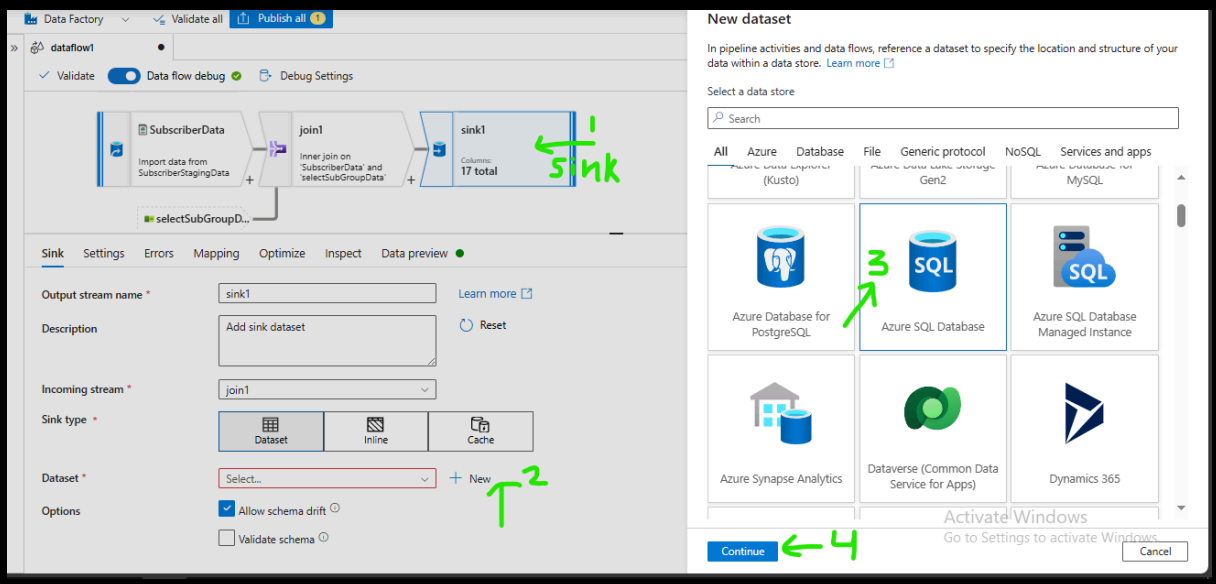
- Joining subscriber_data with selectedSubGroupData

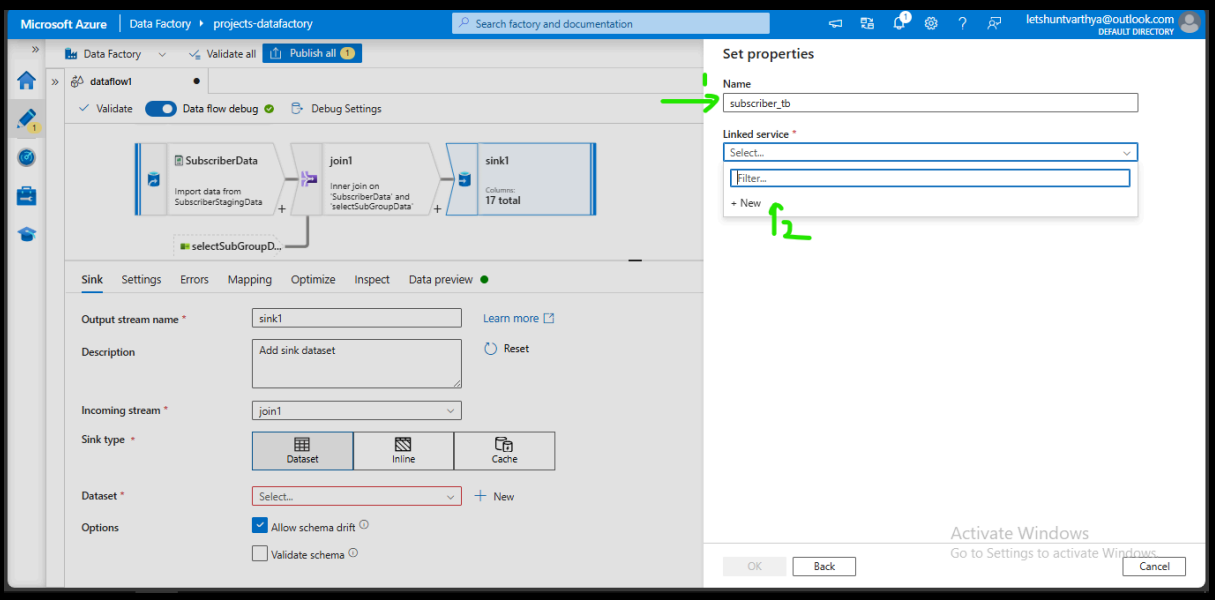


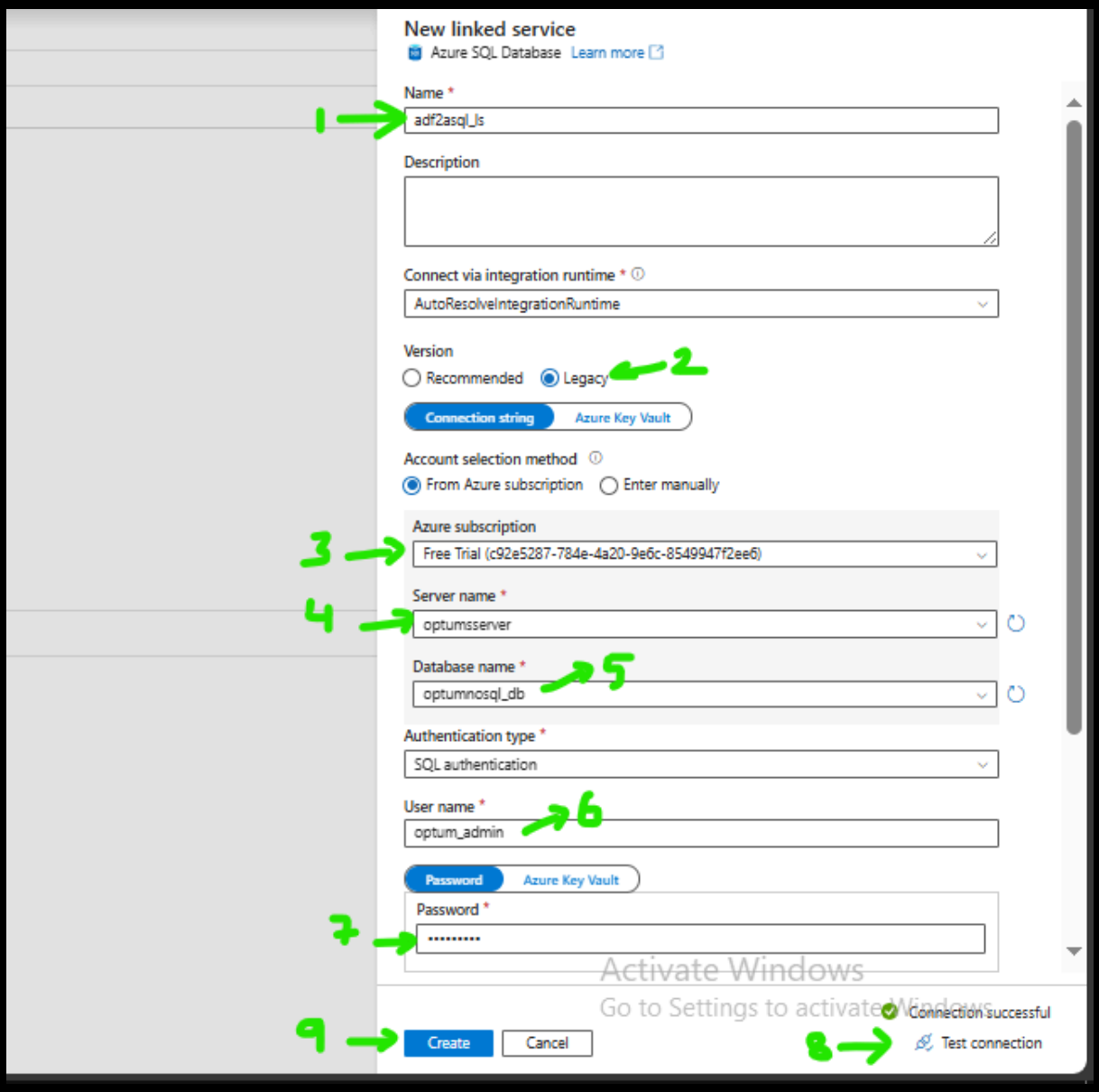


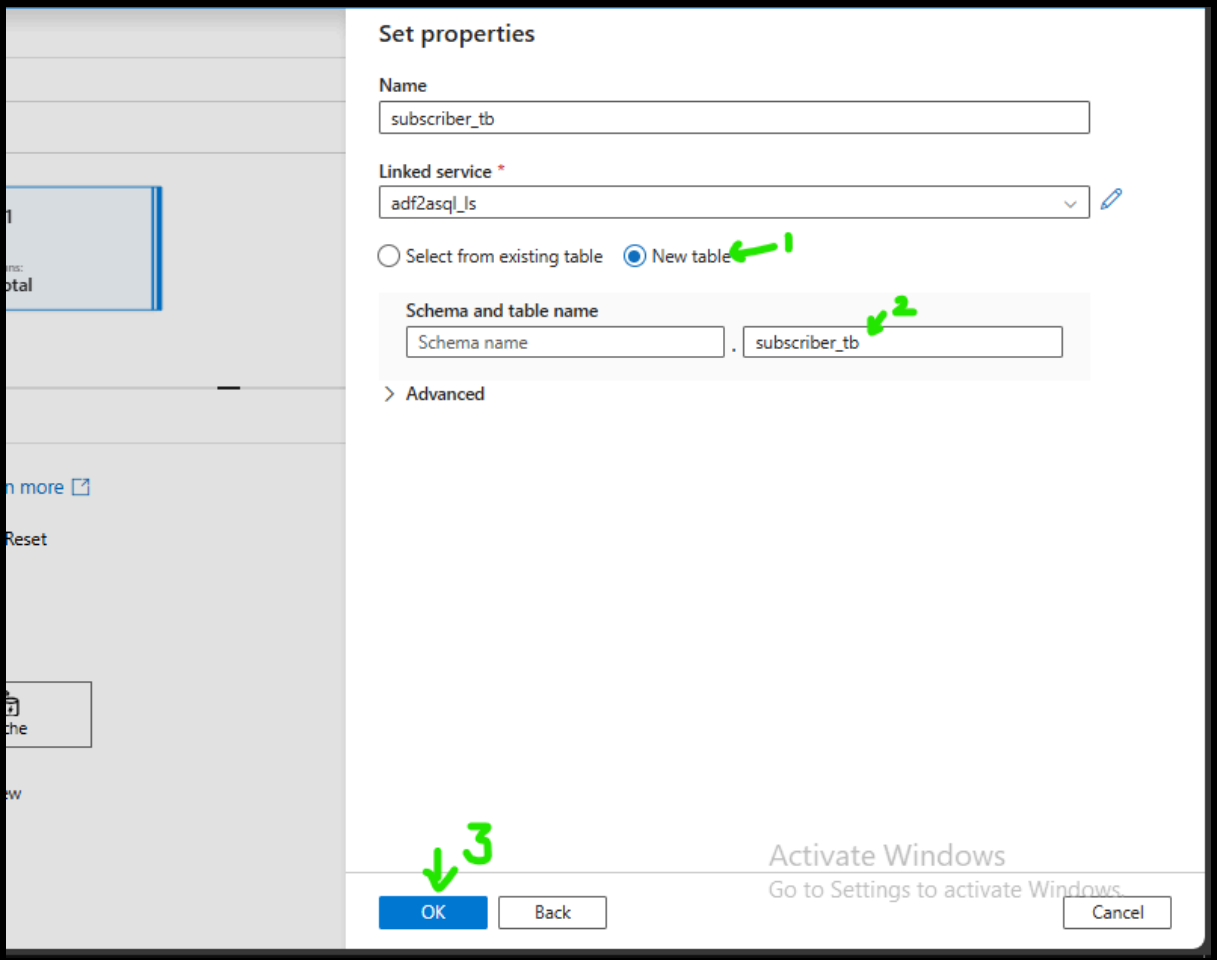
- Add sink after this join activity.



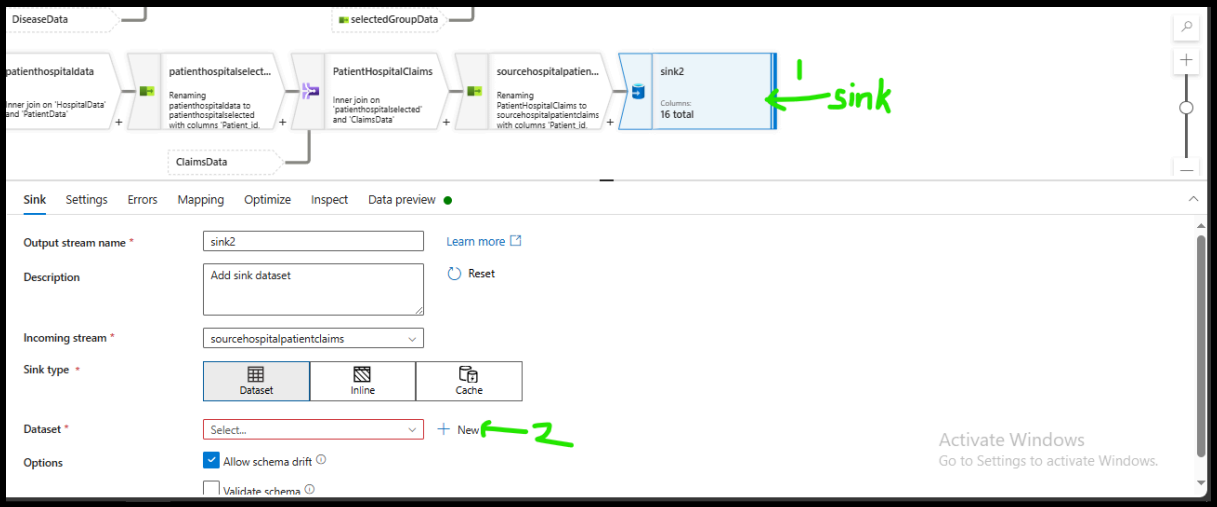


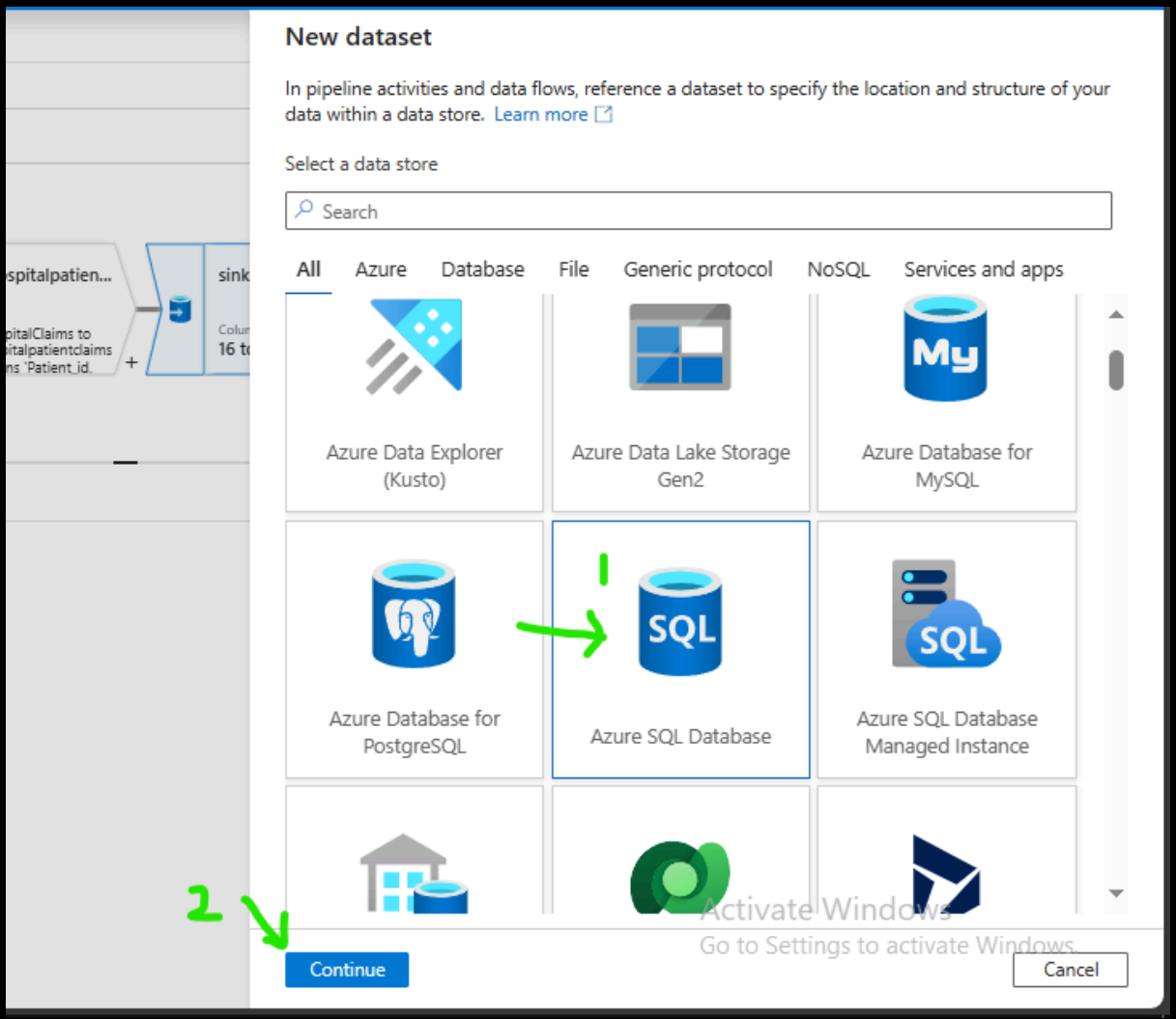

Microsoft Azure
Data Factory
projects-datafactory
Search factory and documentation
letshuntvarthya@outlook.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all
dataflow1
Validate
Data flow debug
Debug Settings
SubscriberData
Import data from SubscriberStagingData
join1
Inner join on 'SubscriberData' 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
2
Activate Windows
Go to Settings to activate Windows.
OK
Back
Cancel

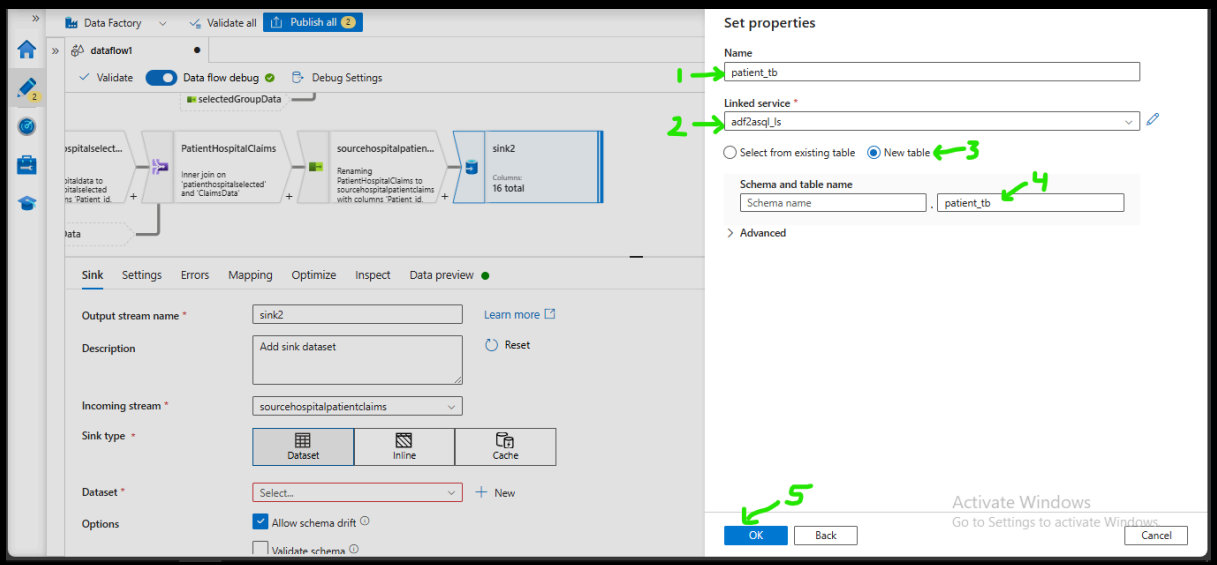

New linked service
Azure SQL Database
Learn more
Name *
1
adf2asql_ls
Description
Connect via integration runtime *
AutoResolveIntegrationRuntime
Version
Recommended
Legacy
2
Connection string
Azure Key Vault
Account selection method
From Azure subscription
Enter manually
Azure subscription
3
Free Trial (c92e5287-784e-4a20-9e6c-8549947f2ee6)
Server name *
4
optumsserver
Database name *
5
optumnosql_db
Authentication type *
SQL authentication
User name *
6
optum_admin
Password
Azure Key Vault
Password *
7
Activate Windows
Go to Settings to activate Windows.
Connection successful
9
Create
Cancel
8
Test connection

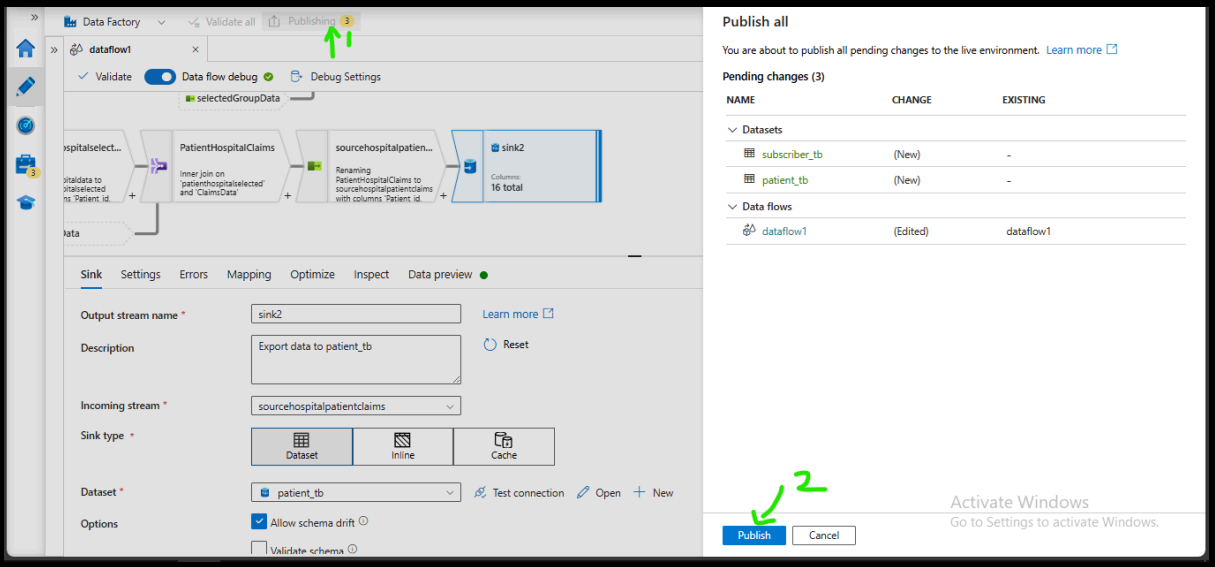


- Add sink after sourcehospitalpatientclaims.

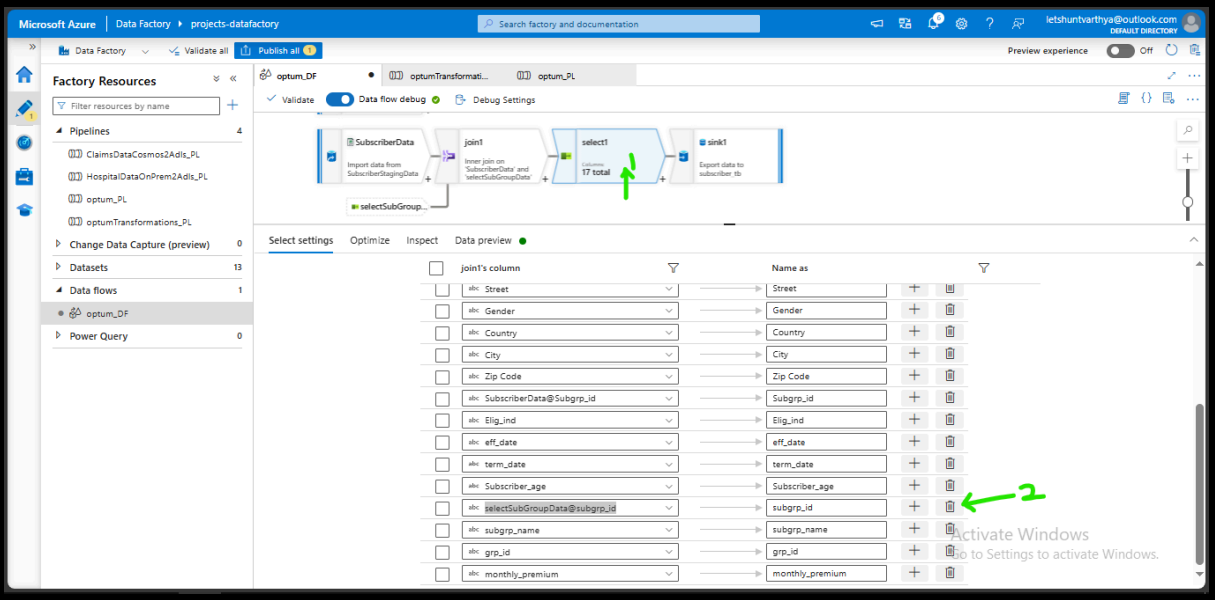



New dataset
In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. Learn more
Select a data store
Search
All
Azure
Database
File
Generic protocol
NoSQL
Services and apps
Azure Data Explorer (Kusto)
Azure Data Lake Storage Gen2
Azure Database for MySQL
Azure Database for PostgreSQL
Azure SQL Database
Azure SQL Database Managed Instance
1
2
Continue
Cancel
Activate Windows
Go to Settings to activate Windows.

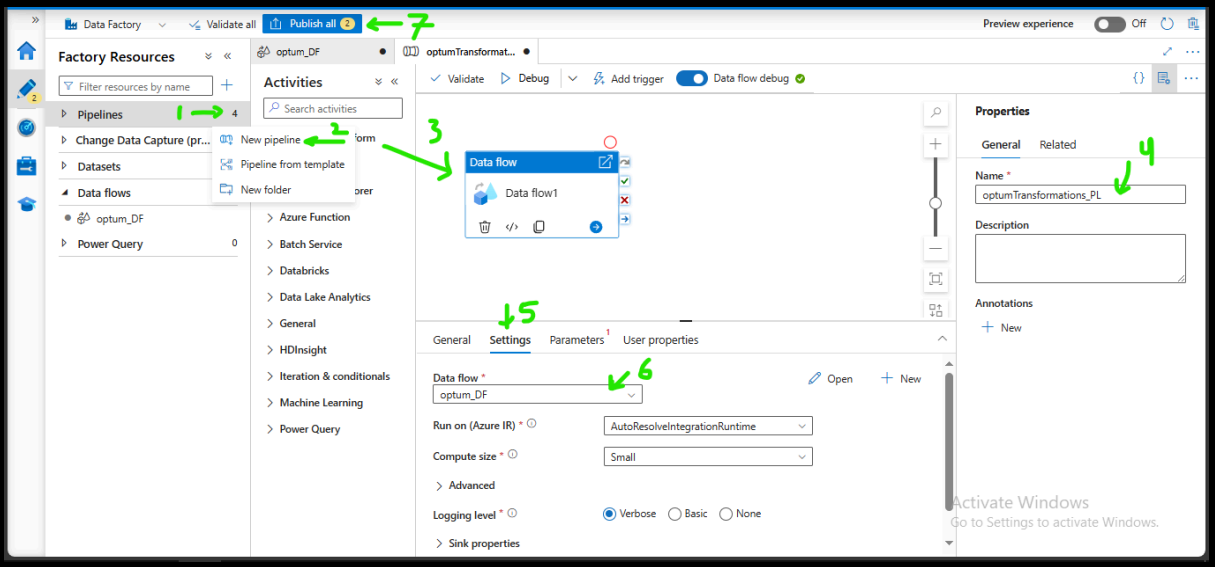

Data Factory
Validate all
Publish all
dataflow1
Validate
Data flow debug
Debug Settings
selectedGroupData
PatientHospitalClaims
sourcehospitalpatien...
sink2
16 total
Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Output stream name *
sink2
Learn more
Description
Add sink dataset
Reset
Incoming stream *
sourcehospitalpatientclaims
Sink type *
Dataset
Inline
Cache
Dataset *
Select...
New
Options
Allow schema drift
Validate schema
Set properties
Name
patient_tb
Linked service *
adf2asql_ls
Select from existing table
New table
3
Schema and table name
Schema name
patient_tb
4
Advanced
5
OK
Back
Cancel
Activate Windows
Go to Settings to activate Windows.

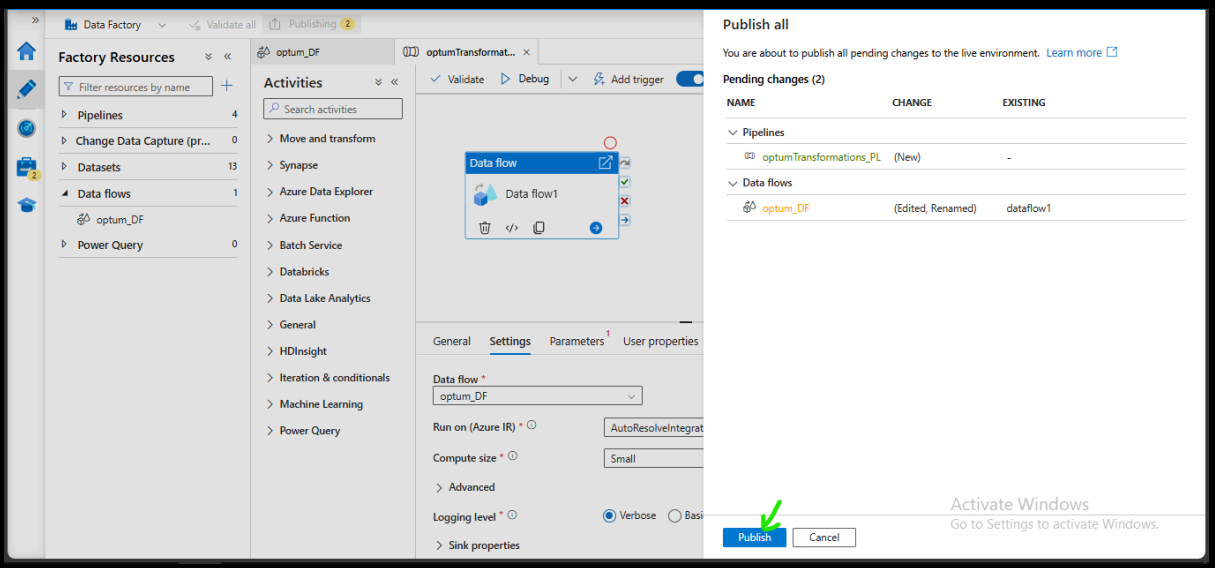


- In optum_DF add select after joining of subscriber and selectSubGroupData.
- Delete selectSubGroupData@subgrp_id column.

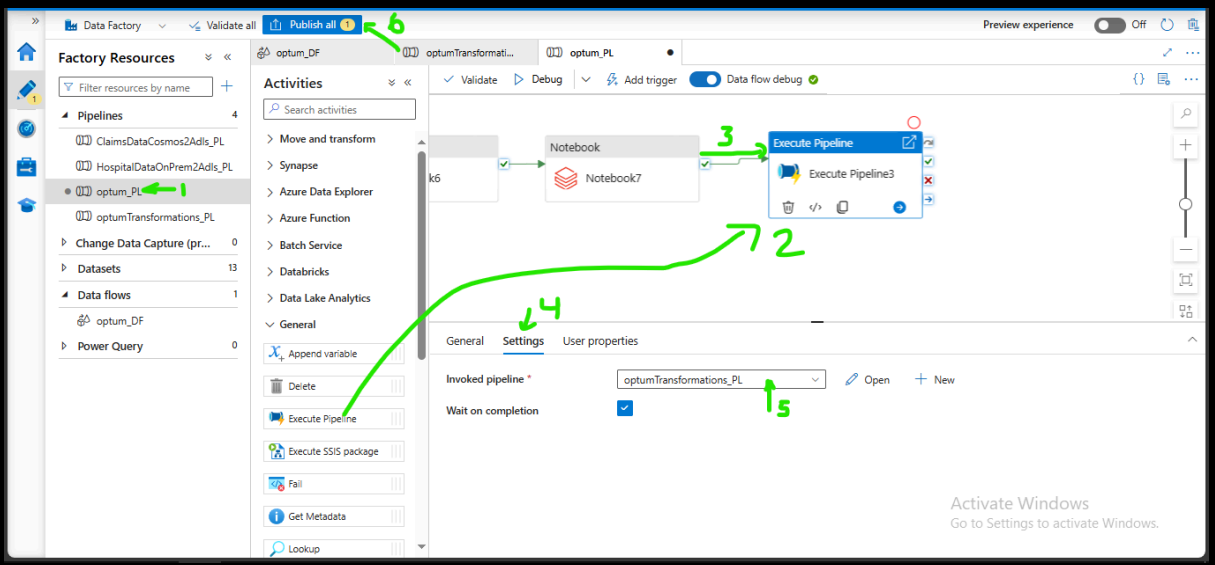


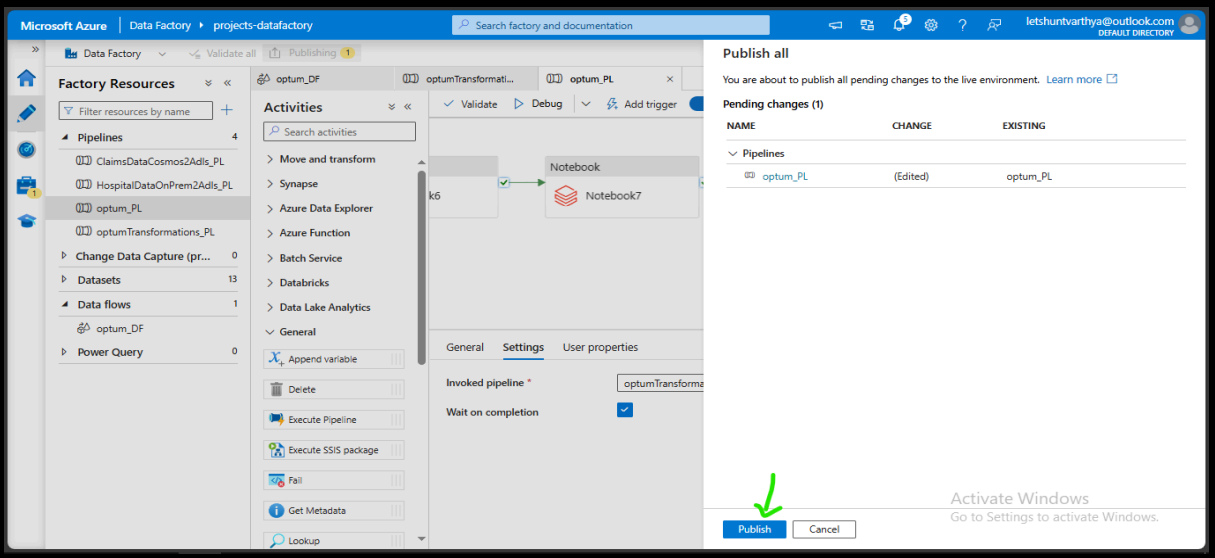
- Create a new pipeline to run optum_DF.
- Pipeline name : optumTransformation_PL
- Drag and drop data flow activity from move and transform section to the field.
- Publish all → publish



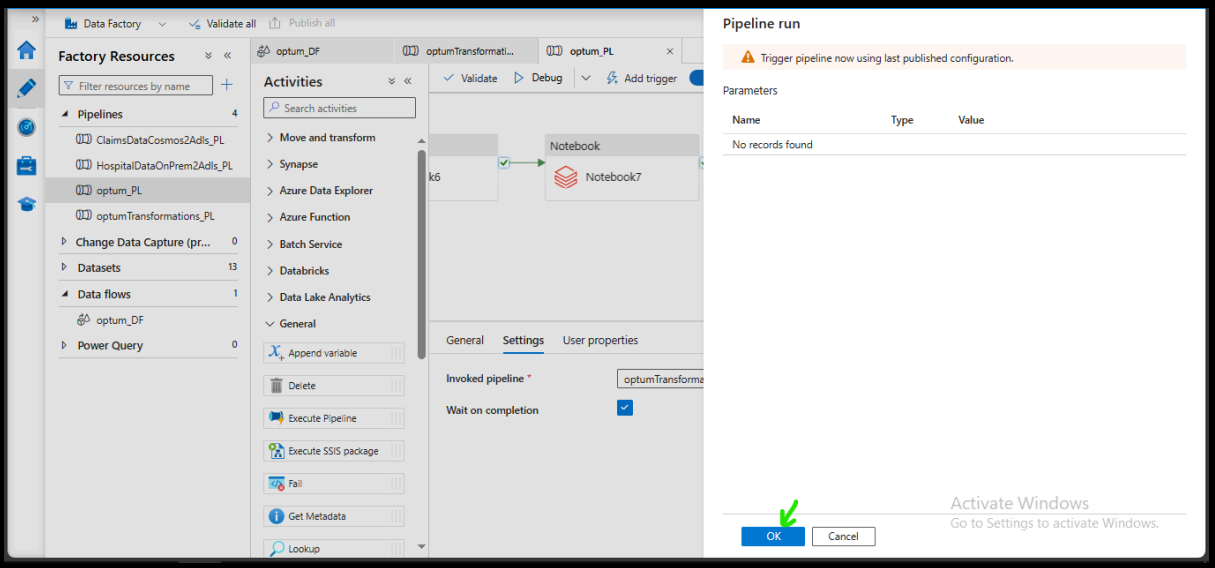


- In the optum_PL, drag and drop execute pipeline activity to the field and select optumTransformation_PL in the settings.
- Publish all → publish

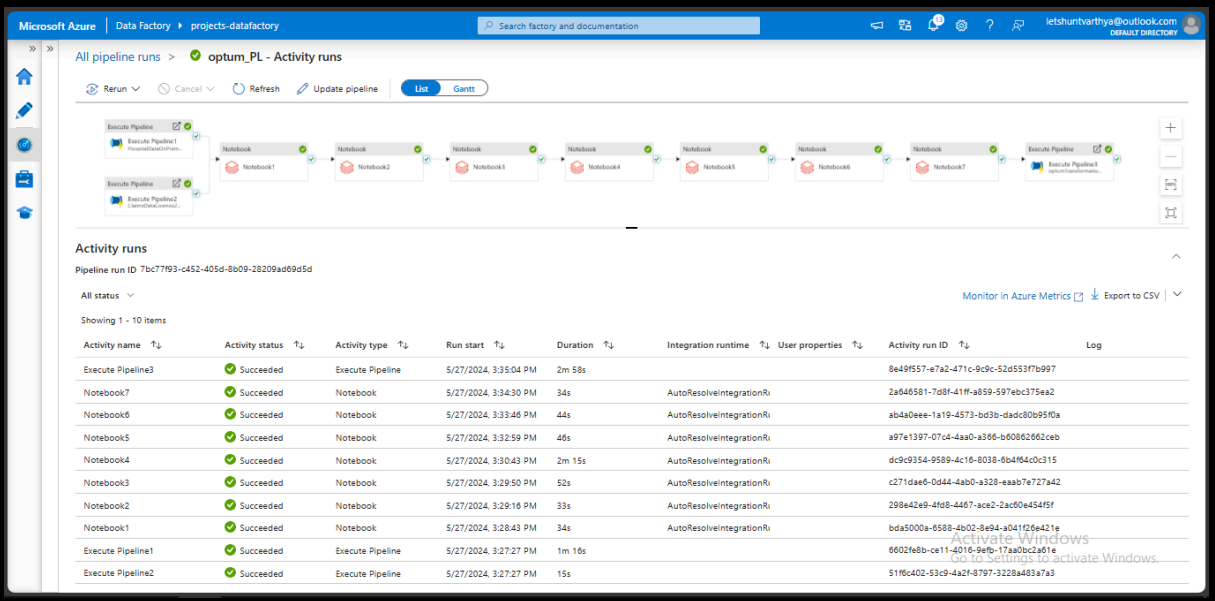




- Add trigger → trigger now → ok



- Successful execution of pipeline.



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