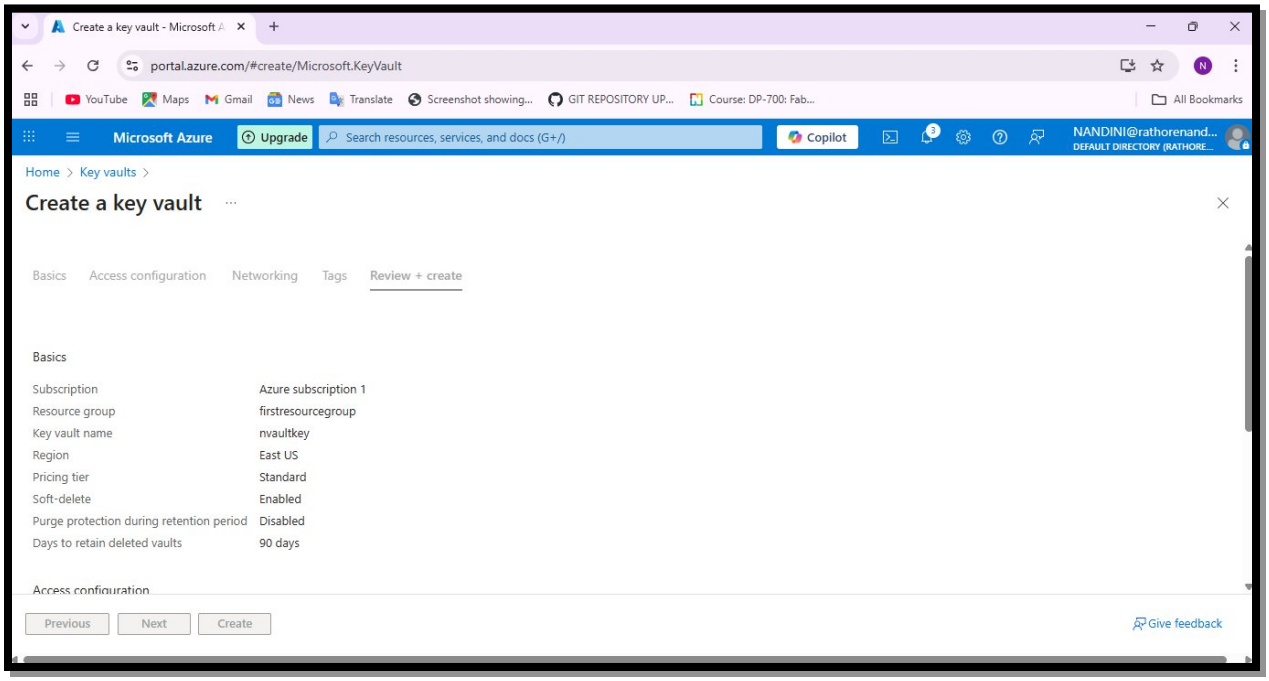
PROJECT-07

End -to-End Pipeline Design:SCD TYPE 1 with ADLSGen2,Delta Lake ,Databricks & ADF.

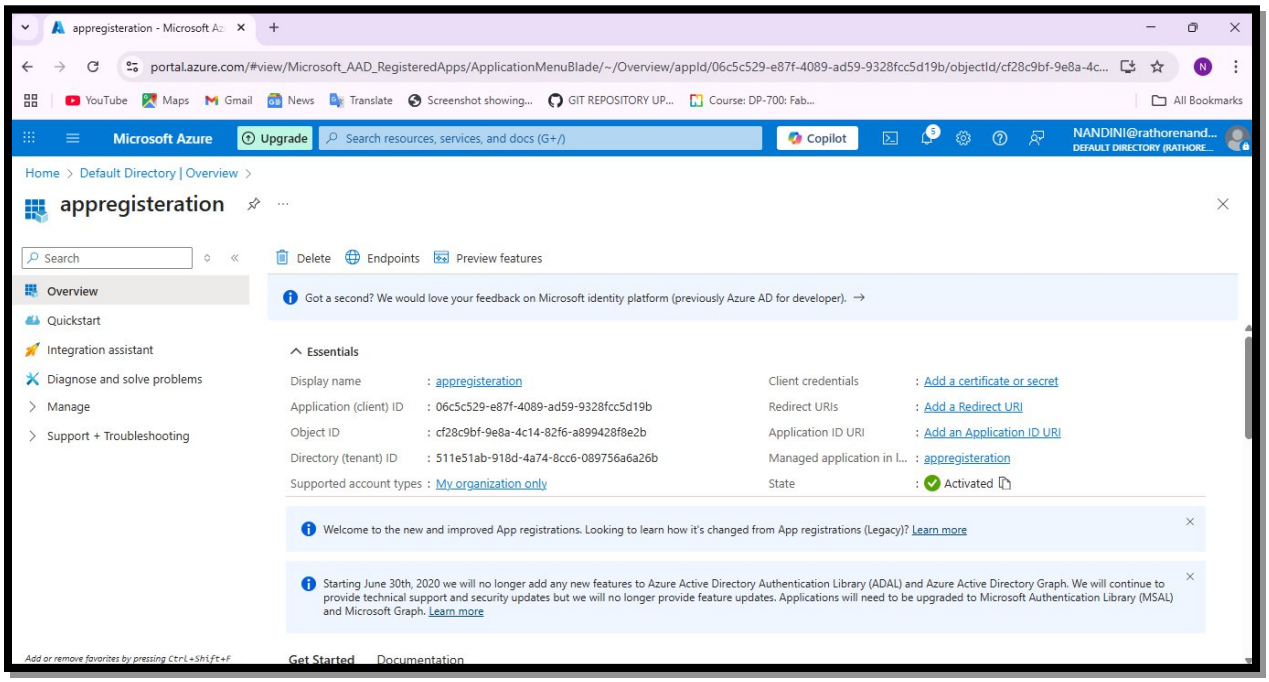
Read data from ADLS GEN2

- Apply slowly changing dimension (SCD) Type 1logic.
- Store the data in Delta Lake format as an external table.
- Parameterize and trigger it from Azure Data Factory (ADF) using dynamic parameters and Databricks notebook activity.

NANDINI RATHORE



1.creating key vault.



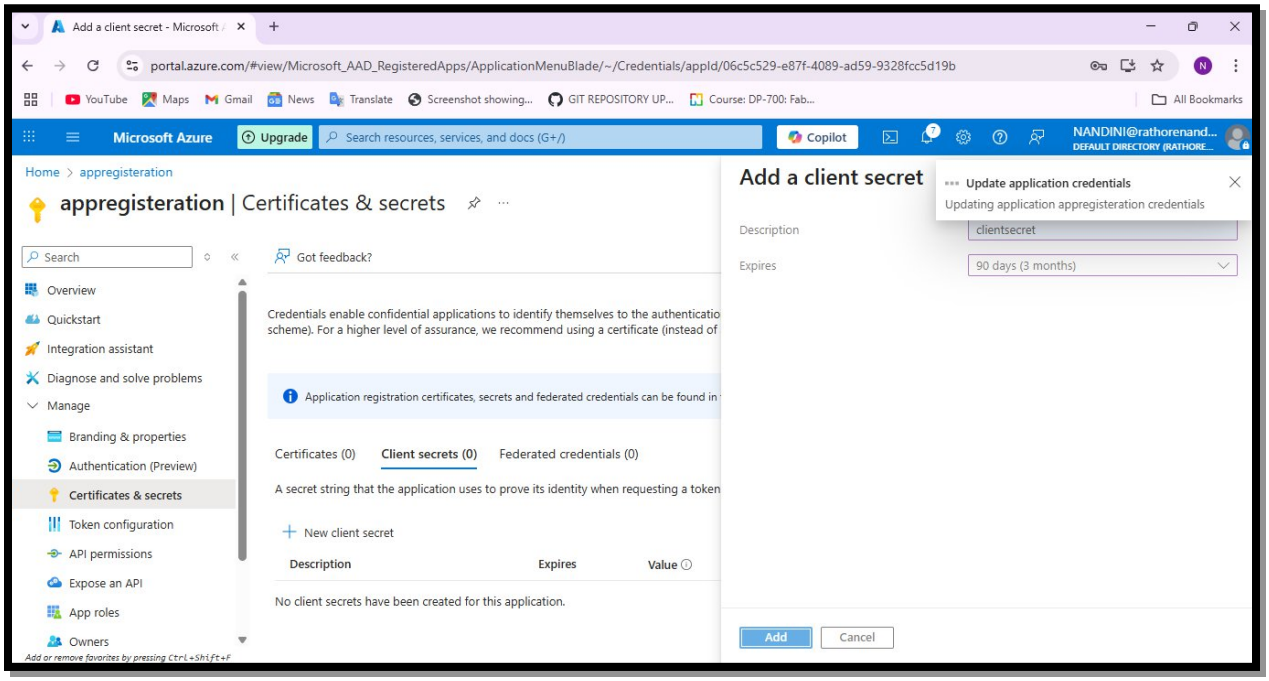
2.creating app registration.

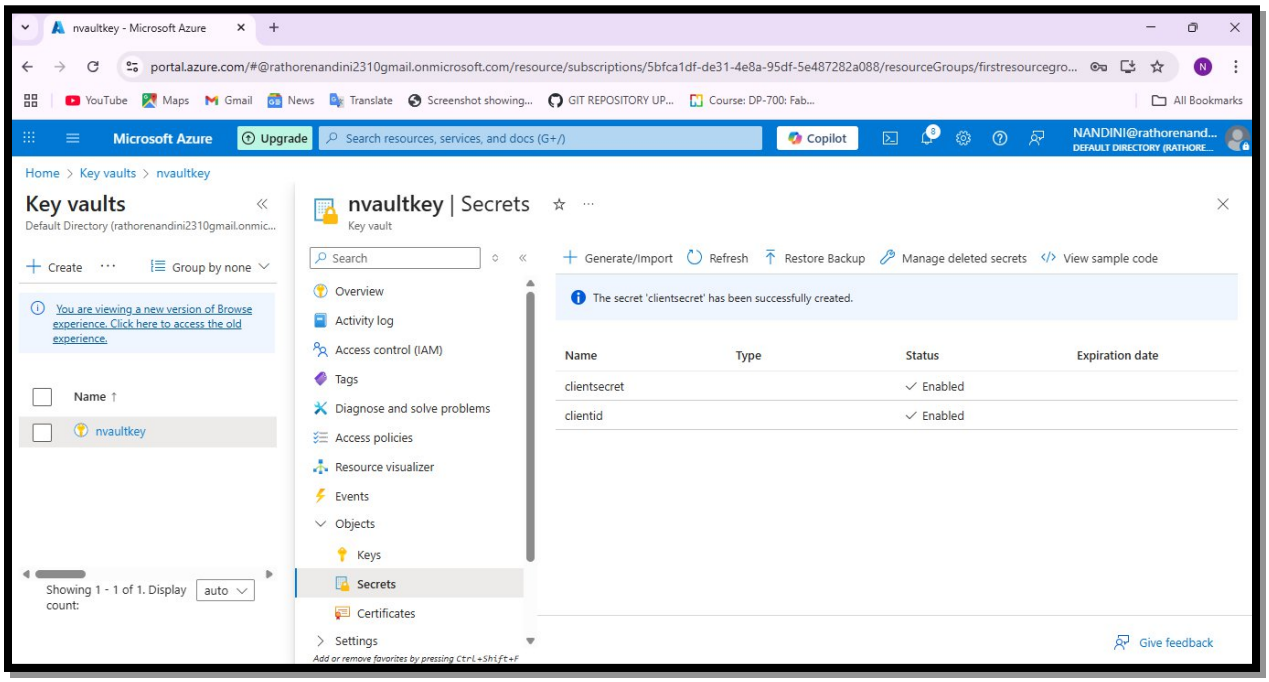
06c5c529-e87f-4089-ad59-9328fcc5d19b client id

NANDINI RATHORE

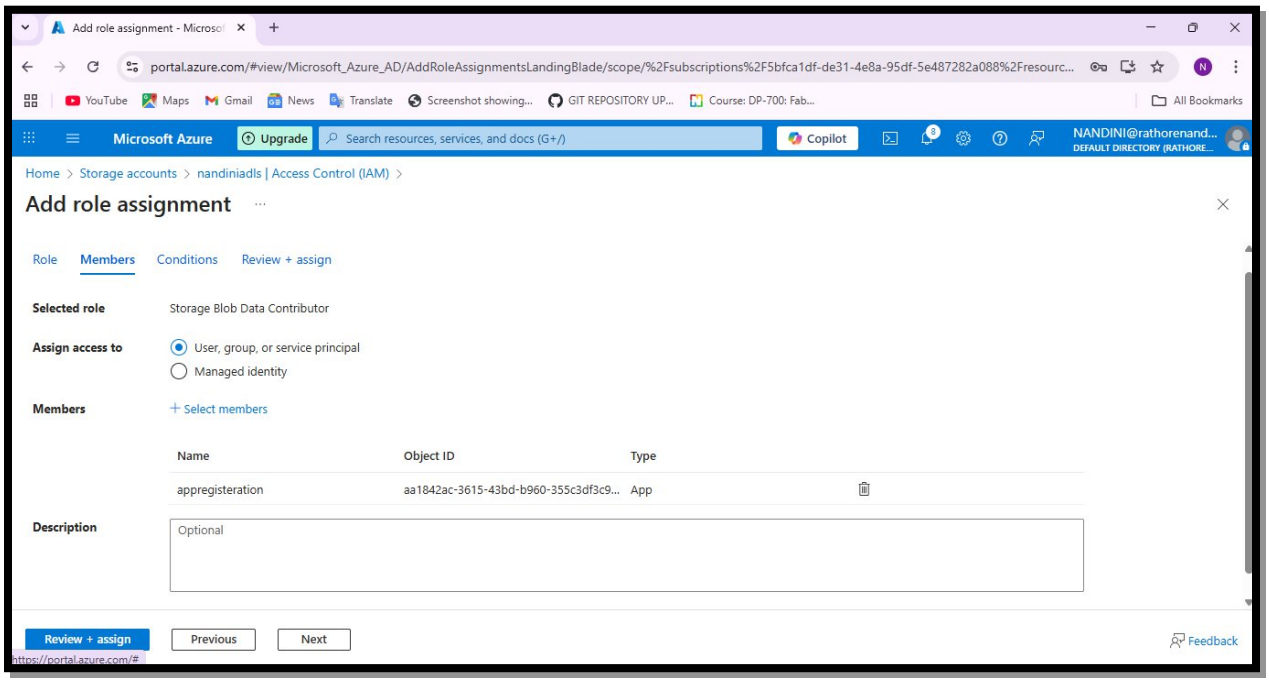
Client secret Jxt8Q~ubtFOSer39kU4SCGy2_S4afWOioHKFkbBV

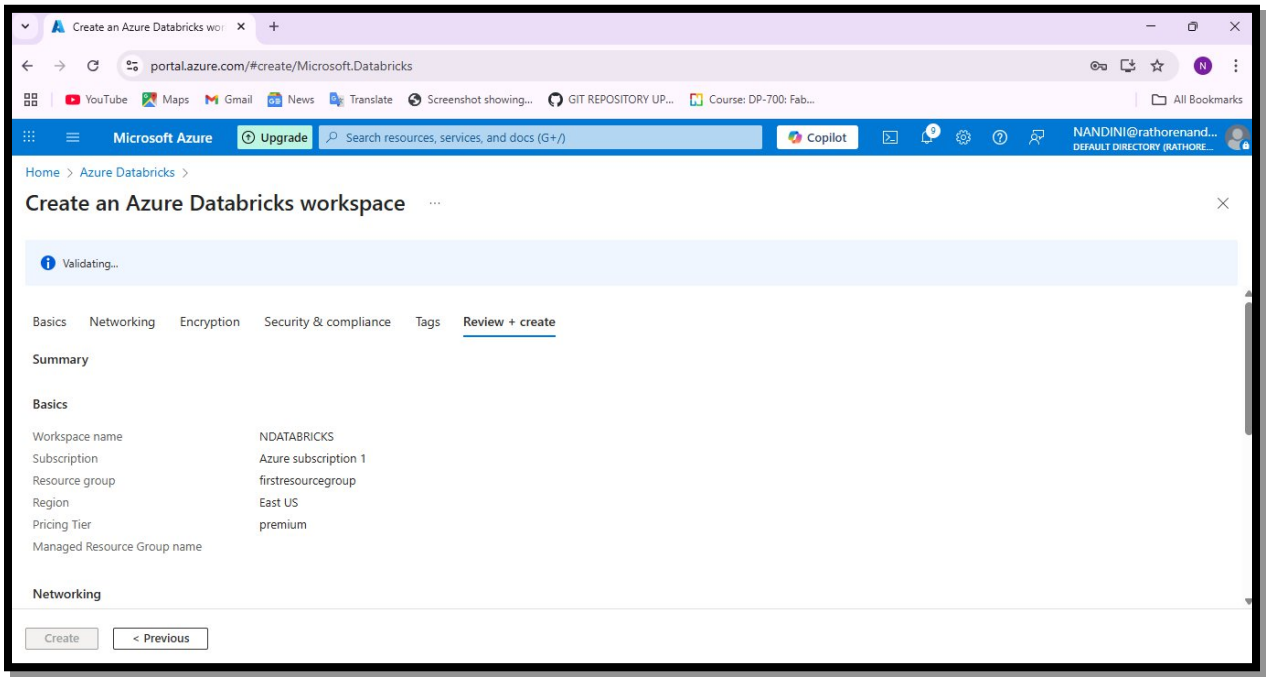
511e51ab-918d-4a74-8cc6-089756a6a26b tenant id.



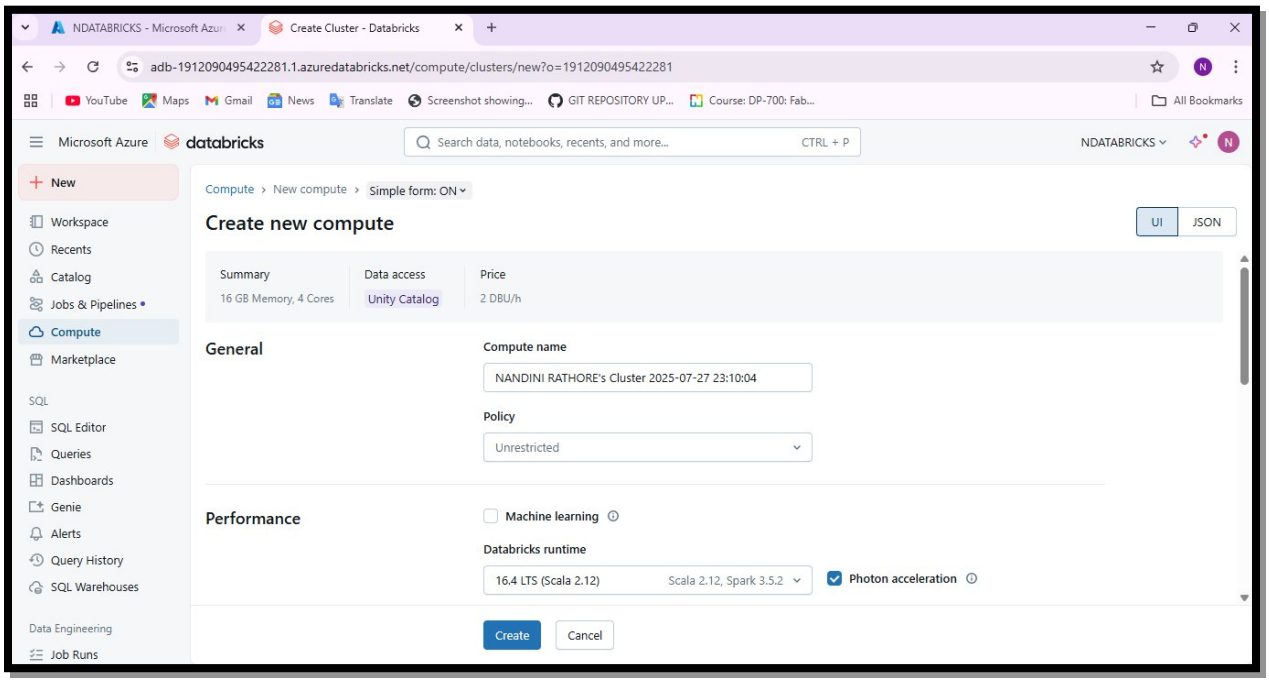


Vault uri:- https://nvaultkey.vault.azure.net/


Add role assignment - Microsof
portal.azure.com/#view/Microsoft_Azure_AD/AddRoleAssignmentsLandingBlade/scope/%2Fsubscriptions%2F5bfca1df-de31-4e8a-95df-5e487282a088%2Fresourc...
YouTube
Maps
Gmail
News
Translate
Screenshot showing...
GIT REPOSITORY UP...
Course: DP-700: Fab...
All Bookmarks
Microsoft Azure
Upgrade
Search resources, services, and docs (G+/)
Copilot
NANDINI@rathorenand...
DEFAULT DIRECTORY (RATHORE...
Home > Storage accounts > nandiniadls | Access Control (IAM) >
Add role assignment
Role
Members
Conditions
Review + assign
Selected role
Storage Blob Data Contributor
Assign access to
User, group, or service principal
Managed identity
Members
+ Select members
Name
Object ID
Type
appregisteration
aa1842ac-3615-43bd-b960-355c3df3c9...
App
Description
Optional
Review + assign
Previous
Next
Feedback
https://portal.azure.com/#

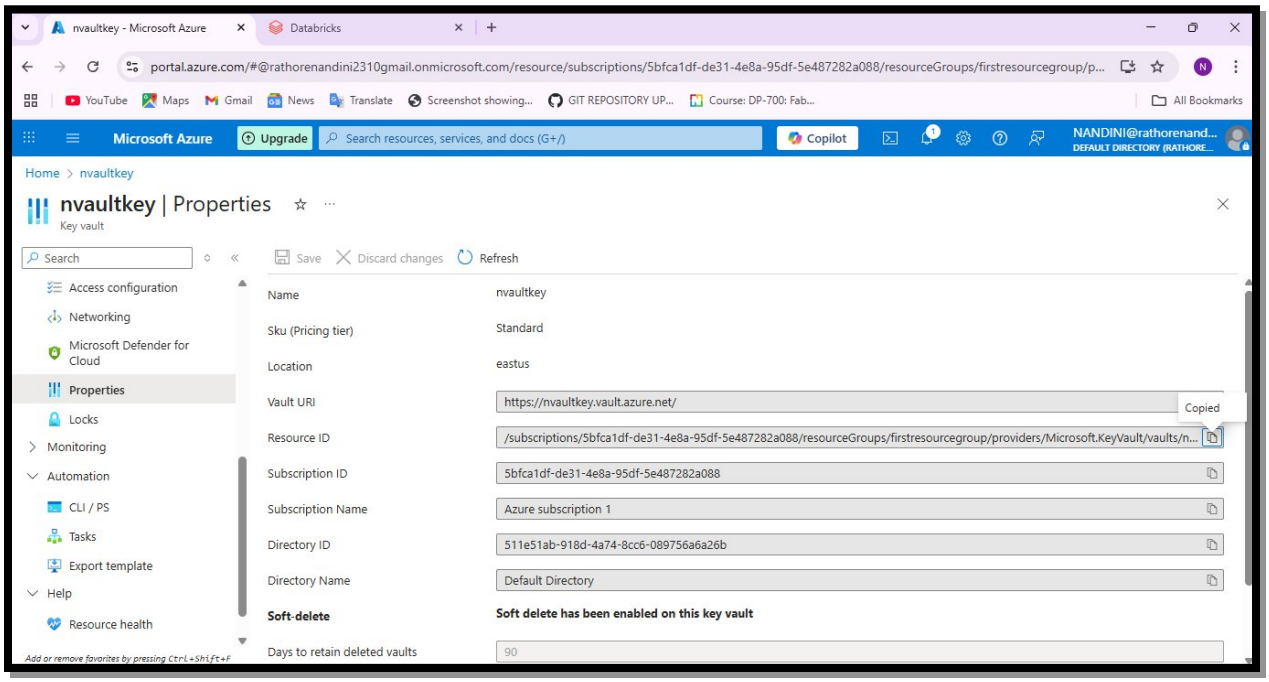

Create an Azure Databricks wor
portal.azure.com/#create/Microsoft.Databricks
YouTube
Maps
Gmail
News
Translate
Screenshot showing...
GIT REPOSITORY UP...
Course: DP-700: Fab...
All Bookmarks
Microsoft Azure
Upgrade
Search resources, services, and docs (G+/)
Copilot
NANDINI@rathorenand...
DEFAULT DIRECTORY (RATHORE...
Home > Azure Databricks >
Create an Azure Databricks workspace
Validating...
Basics
Networking
Encryption
Security & compliance
Tags
Review + create
Summary
Basics
Workspace name
NDATABRICKS
Subscription
Azure subscription 1
Resource group
firstresourcegroup
Region
East US
Pricing Tier
premium
Managed Resource Group name
Networking
Create
< Previous

NANDINI RATHORE

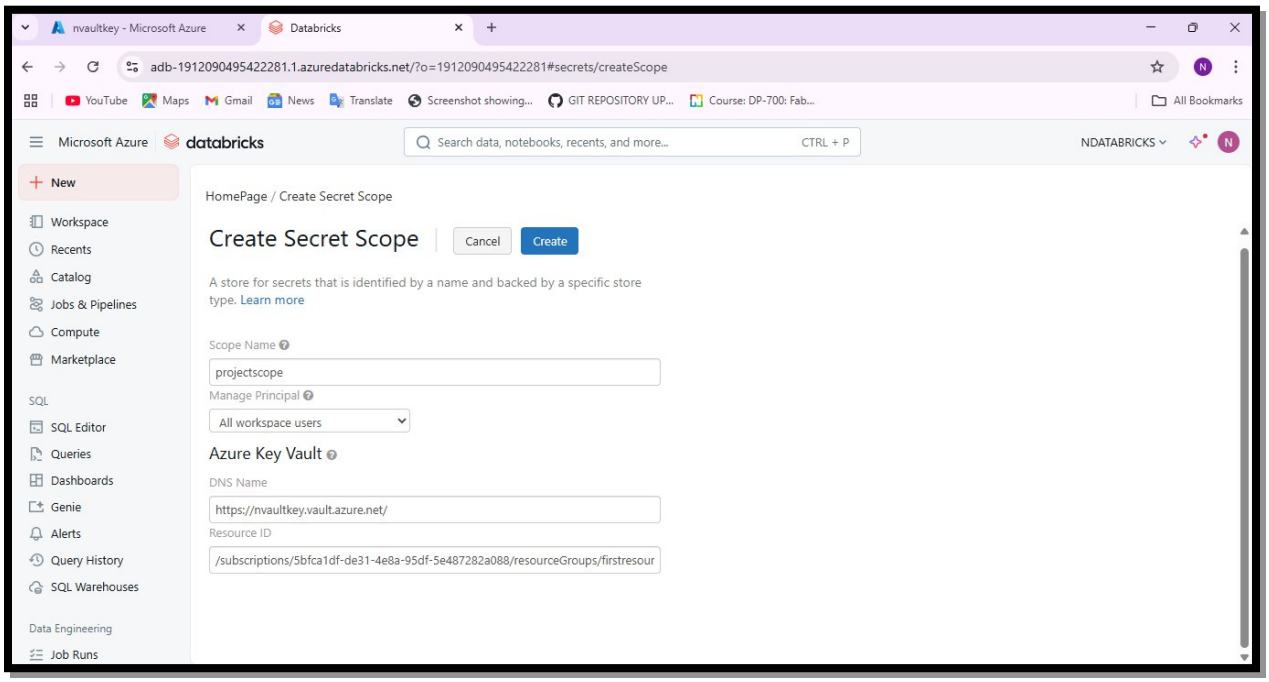


For creating scope:-

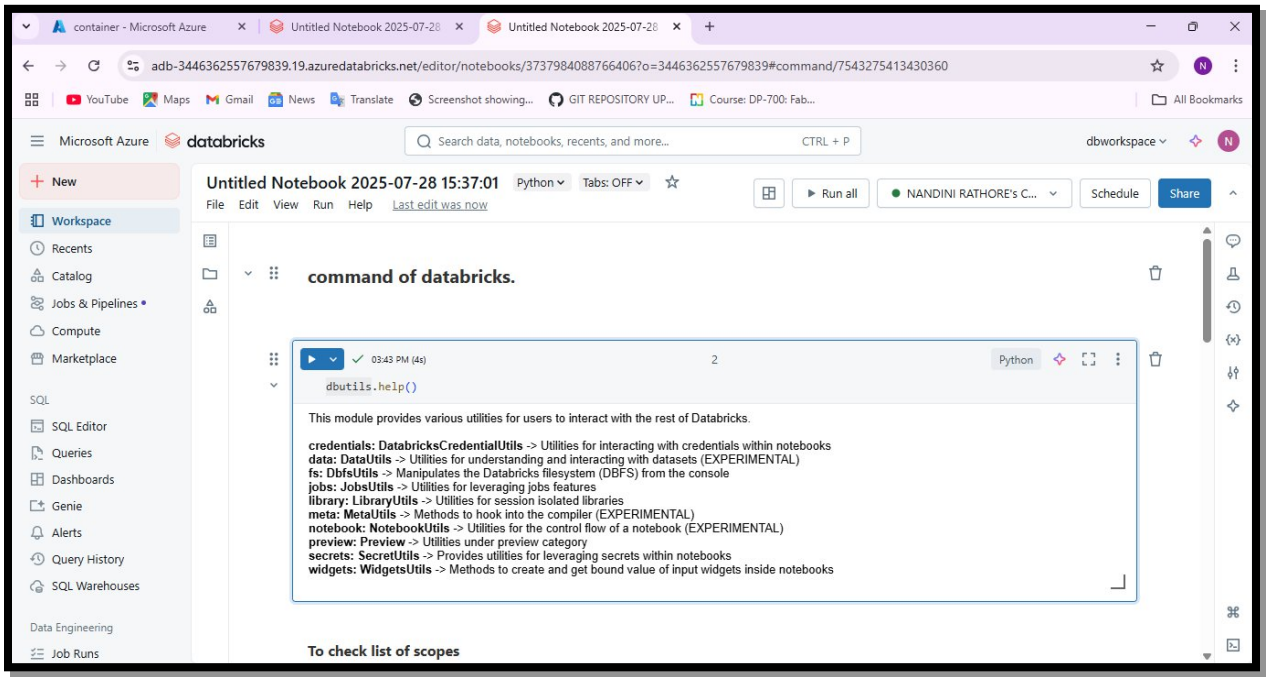
https://adb-1912090495422281.1.azuredatabricks.net/?o=1912090495422281#secrets/createScope



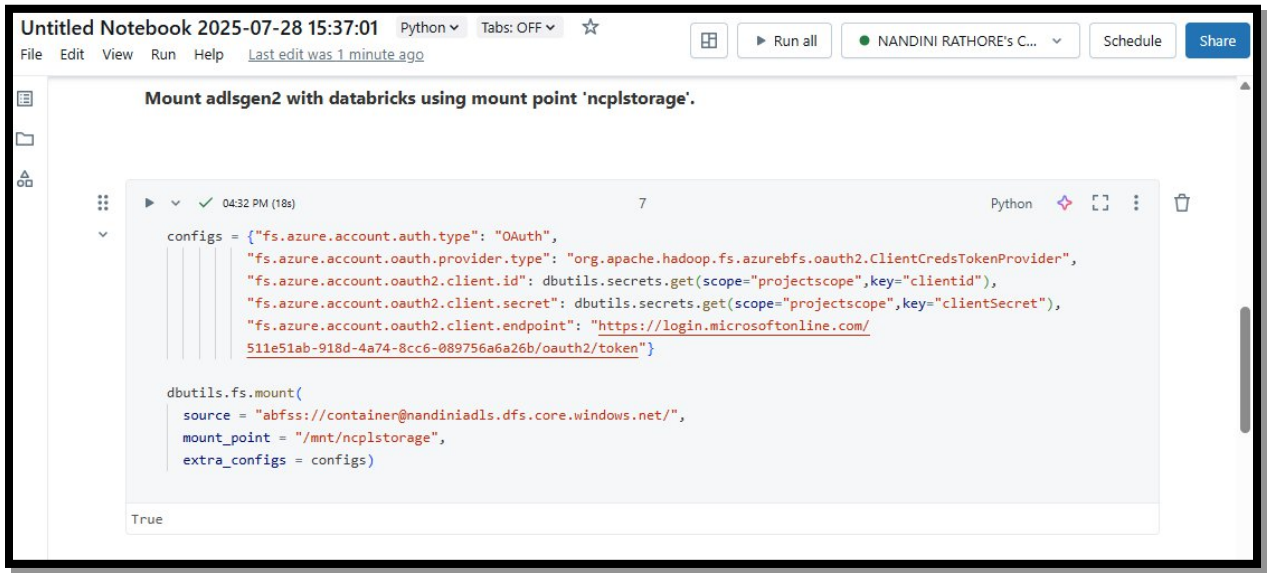
Copy resource ID .



Paste here resource ID and azure vault URI. It is the **unique URL** used to access a specific **Azure Key Vault** in your Azure environment.

- 

Dbutils.help() command is used in Databricks notebooks to display a list of available utility commands provided by the dbutils module.




```
configs = {"fs.azure.account.auth.type": "OAuth",
       "fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",
       "fs.azure.account.oauth2.client.id": dbutils.secrets.get(scope="projectscope",key="clientid"),
       "fs.azure.account.oauth2.client.secret": dbutils.secrets.get(scope="projectscope",key="clientSecret"),
       "fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/511e51ab-918d-4a74-8cc6-
089756a6a26b/oauth2/token"}

dbutils.fs.mount(
 source = "abfss://container@nandiniadls.dfs.core.windows.net/",
 mount_point = "/mnt/ncplstorage",
 extra_configs = configs)
```

To check the mount.

2 minutes ago (<1s) 9 Python

```
dbutils.fs.mounts()
```

```
[MountInfo(mountPoint='/databricks-datasets', source='databricks-datasets', encryptionType=''),
 MountInfo(mountPoint='/Volumes', source='UnityCatalogVolumes', encryptionType=''),
 MountInfo(mountPoint='/databricks/mlflow-tracking', source='databricks/mlflow-tracking', encryptionType=''),
 MountInfo(mountPoint='/databricks-results', source='databricks-results', encryptionType=''),
 MountInfo(mountPoint='/databricks/mlflow-registry', source='databricks/mlflow-registry', encryptionType=''),
 MountInfo(mountPoint='/Volume', source='DbfsReserved', encryptionType=''),
 MountInfo(mountPoint='/volumes', source='DbfsReserved', encryptionType=''),
 MountInfo(mountPoint='/', source='DatabricksRoot', encryptionType=''),
 MountInfo(mountPoint='/mnt/ncplstorage', source='abfss://container@nandiniadls.dfs.core.windows.net/', encryptionType=''),
 MountInfo(mountPoint='/volume', source='DbfsReserved', encryptionType='')]
```

to check number of files in adls using mount point.

4 minutes ago (1s) 9

```
dbutils.fs.ls("/mnt/ncplstorage")
```

```
[FileInfo(path='dbfs:/mnt/ncplstorage/Azurecapacities.csv', name='Azurecapacities.csv', size=49, modificationTime=1753294870000),
 FileInfo(path='dbfs:/mnt/ncplstorage/firstfile.csv', name='firstfile.csv', size=98, modificationTime=1753663867000),
 FileInfo(path='dbfs:/mnt/ncplstorage/folder1/', name='folder1/', size=0, modificationTime=1753292131000),
 FileInfo(path='dbfs:/mnt/ncplstorage/secondfile.csv', name='secondfile.csv', size=92, modificationTime=1753663884000)]
```

Read a file from adlsgen2.

1 minute ago (2s) 14

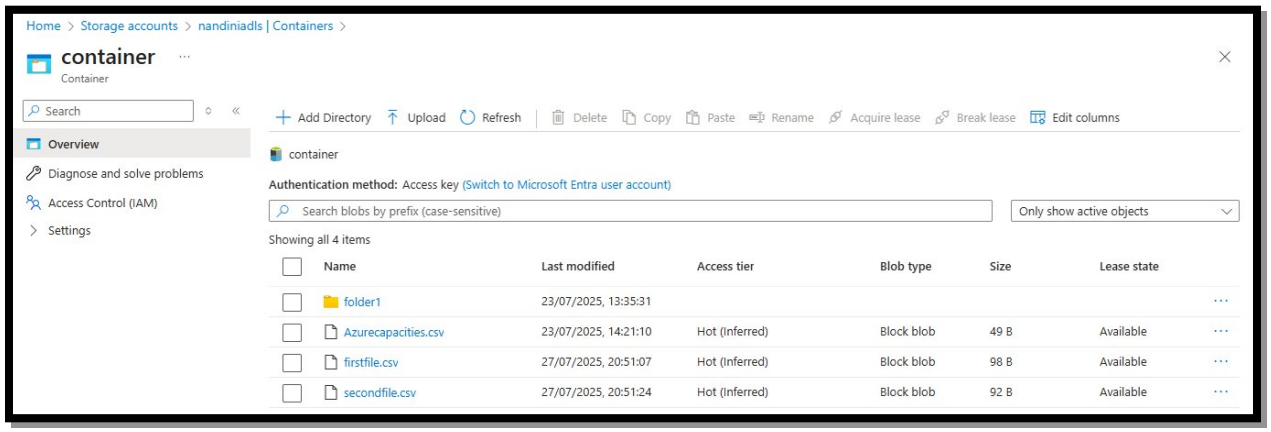
```
df=spark.read.format("csv").option("header",True).option("inferschema",True).load("/mnt/ncplstorage/firstfile.csv")
display(df)
```

(3) Spark Jobs

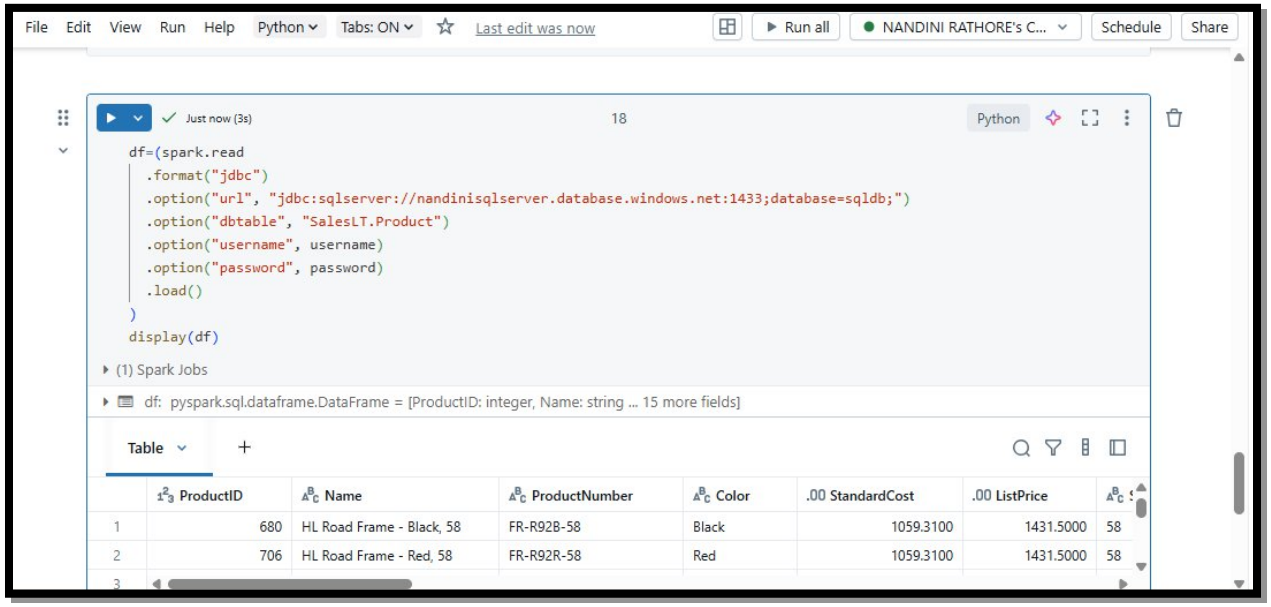
df: pyspark.sql.dataframe.DataFrame = [id: integer, name: string ... 2 more fields]

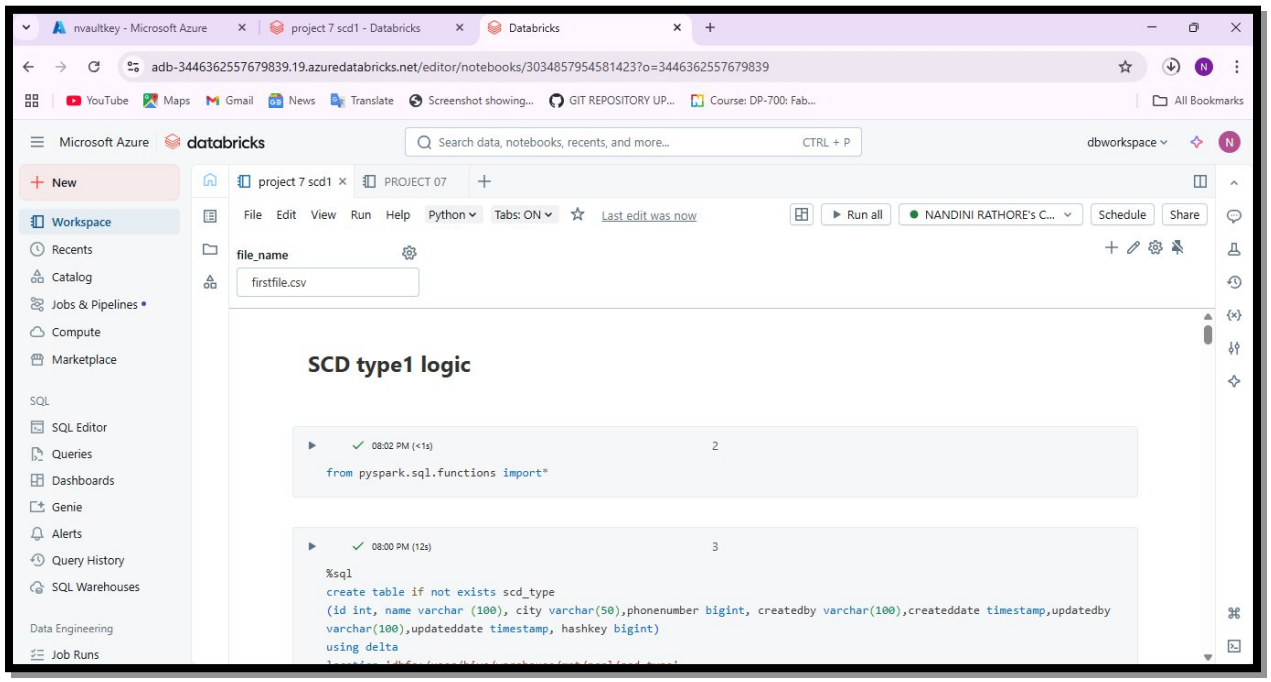
Table +

	id	name	city	phonenumber
1	1	nandini	ptl	814653852
2	2	lakhwinder	ptl	754123694
3	3	dimple	hry	888742012

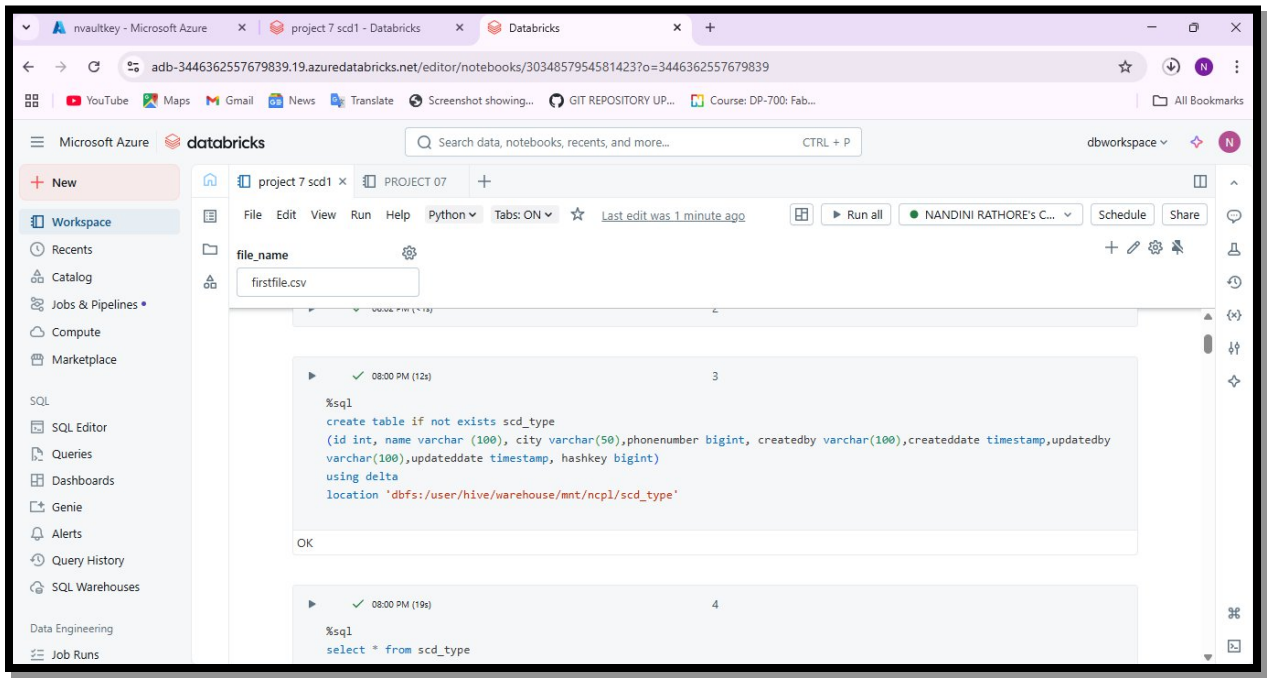

Home > Storage accounts > nandiniadls | Containers >
container
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Add Directory
Upload
Refresh
Delete
Copy
Paste
Rename
Acquire lease
Break lease
Edit columns
container
Authentication method: Access key (Switch to Microsoft Entra user account)
Search blobs by prefix (case-sensitive)
Only show active objects
Showing all 4 items
Name
Last modified
Access tier
Blob type
Size
Lease state
folder1
23/07/2025, 13:35:31
Azurecapacities.csv
23/07/2025, 14:21:10
Hot (Inferred)
Block blob
49 B
Available
firstfile.csv
27/07/2025, 20:51:07
Hot (Inferred)
Block blob
98 B
Available
secondfile.csv
27/07/2025, 20:51:24
Hot (Inferred)
Block blob
92 B
Available


Connect Databricks with SQLDB.
05:22 PM (2s)
16
dbutils.secrets.list("projectscope")
[SecretMetadata(key='clientid'),
SecretMetadata(key='clientsecret'),
SecretMetadata(key='password'),
SecretMetadata(key='username')]
05:23 PM (2s)
17
Python
username=dbutils.secrets.get(scope="projectscope",key="username")
password=dbutils.secrets.get(scope="projectscope",key="password")

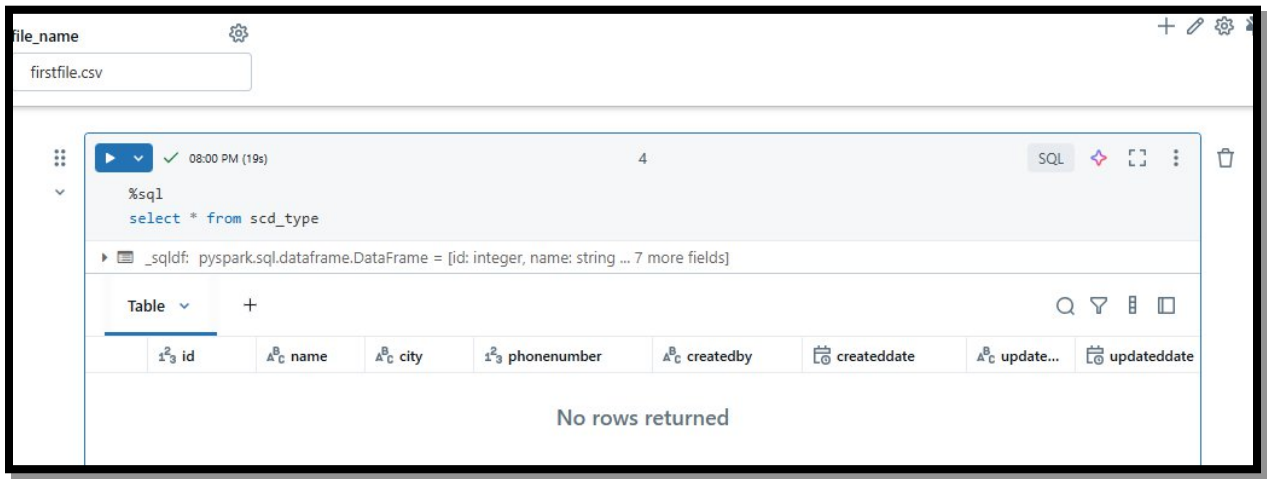

File
Edit
View
Run
Help
Python
Tabs: ON
Last edit was now
Run all
NANDINI RATHORE's C...
Schedule
Share
Just now (3s)
18
Python
df=(spark.read
.format("jdbc")
.option("url", "jdbc:sqlserver://nandinisqlserver.database.windows.net:1433;database=sqldb;")
.option("dbtable", "SalesLT.Product")
.option("username", username)
.option("password", password)
.load()
)
display(df)
(1) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [ProductID: integer, Name: string ... 15 more fields]
Table
ProductID
Name
ProductNumber
Color
StandardCost
ListPrice
1
680
HL Road Frame - Black, 58
FR-R92B-58
Black
1059.3100
1431.5000
58
2
706
HL Road Frame - Red, 58
FR-R92R-58
Red
1059.3100
1431.5000
58



Creating a target table in databricks.



Creating delta table because while creating I gave location for the table and in delta table if we delete table its data is not deleted it remain in storage.



to check type of the table we use this command.

	col_name	data_type
1	id	int
2	name	varchar(100)
3	city	varchar(50)
4	phonenumber	bigint
5	createdby	varchar(100)

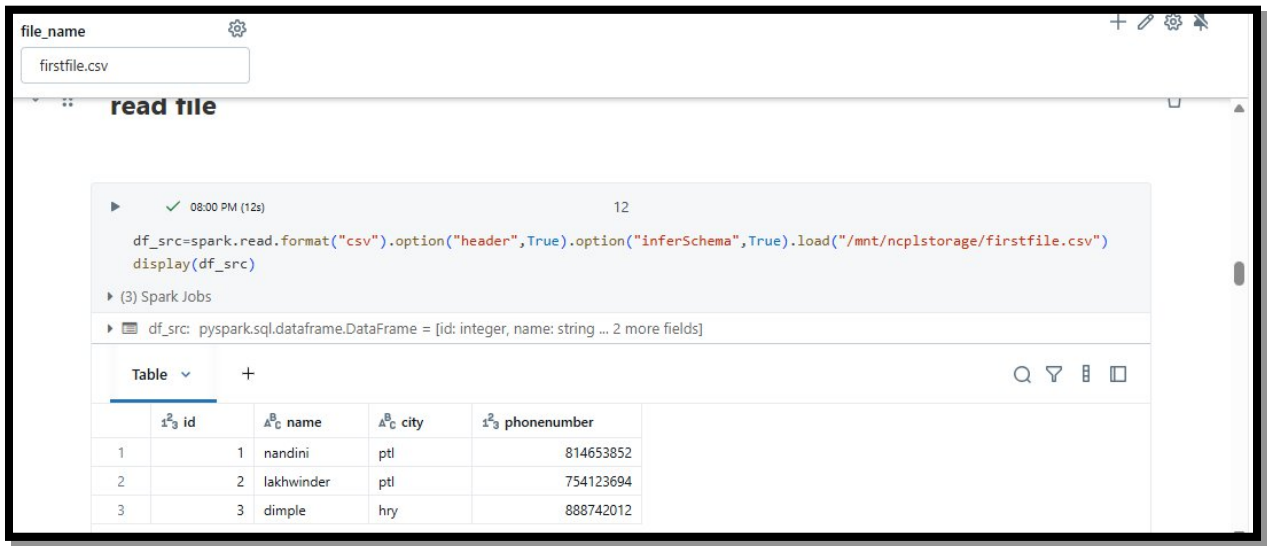
10		
11	# Detailed Table Information	
12	Catalog	spark_catalog
13	Database	default
14	Table	scd_type
15	Created Time	Mon Jul 28 23:37:32 UTC 2025
16	Last Access	UNKNOWN
17	Created By	Spark 3.5.0
18	Type	EXTERNAL
19	Location	dbfs:/user/hive/warehouse/mnt/ncpl/scd_type
20	Provider	delta
21	Owner	root
22	Table Properties	[delta.enableDeletionVectors=true,delta.feature.deletionVectors=supported,delta.minReaderVersion=3,delta.minWriterVersion=7]

Type and location of the table.

dbutils.widgets are used to **create input controls** in your Databricks notebook, so we can make code dynamic, interactive, and reusable.

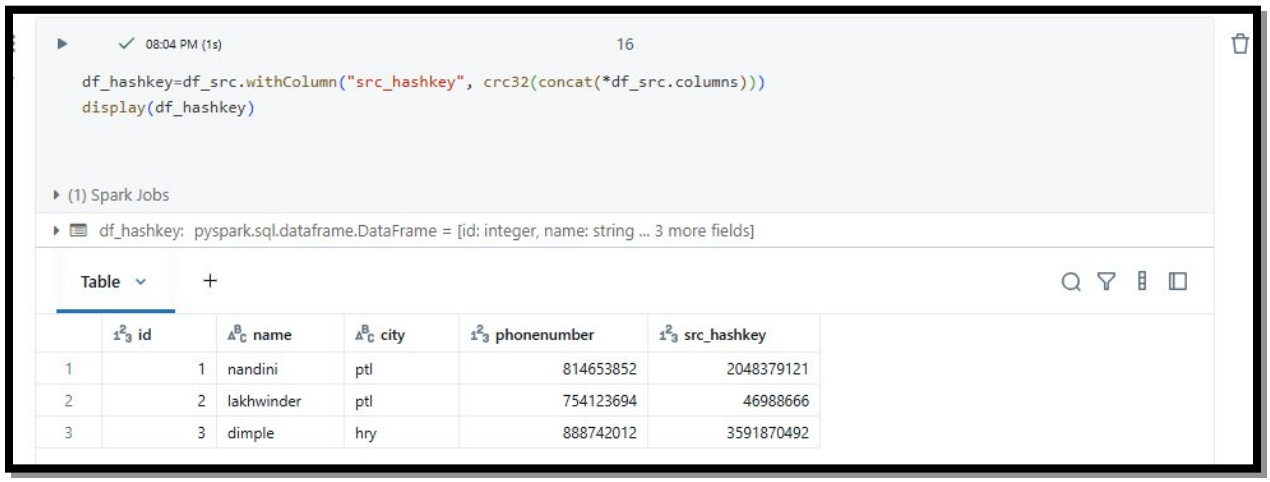
Pass different values (like filenames, dates) without changing the code.

Run the same notebook with different inputs.



Read the csvfile.

At first target data is empty because no new data is added.



```
df_hashkey=df_src.withColumn("src_hashkey", crc32(concat(*df_src.columns)))
display(df_hashkey)
```

08:04 PM (1s) 16

(1) Spark Jobs

df_hashkey: pyspark.sql.dataframe.DataFrame = [id: integer, name: string ... 3 more fields]

Table +

	id	name	city	phonenumber	src_hashkey
1	1	nandini	ptl	814653852	2048379121
2	2	lakhwinder	ptl	754123694	46988666
3	3	dimple	hry	888742012	3591870492

Generating hashkey

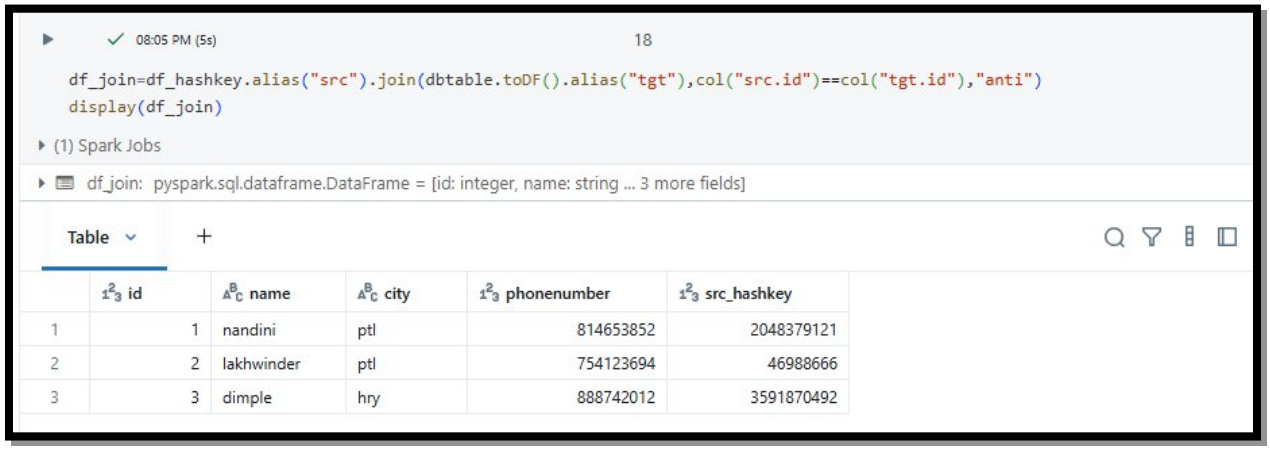
JOIN

```
df_join=df_hashkey.alias("src").join(dbtable.toDF().alias("tgt"),col("src.id")==col("tgt.id"),"anti")
display(df_join)
```

08:05 PM (5s) 18

(1) Spark Jobs

df_join: pyspark.sql.dataframe.DataFrame = [id: integer, name: string ... 3 more fields]



```
df_join=df_hashkey.alias("src").join(dbtable.toDF().alias("tgt"),col("src.id")==col("tgt.id"),"anti")
display(df_join)
```

08:05 PM (5s) 18

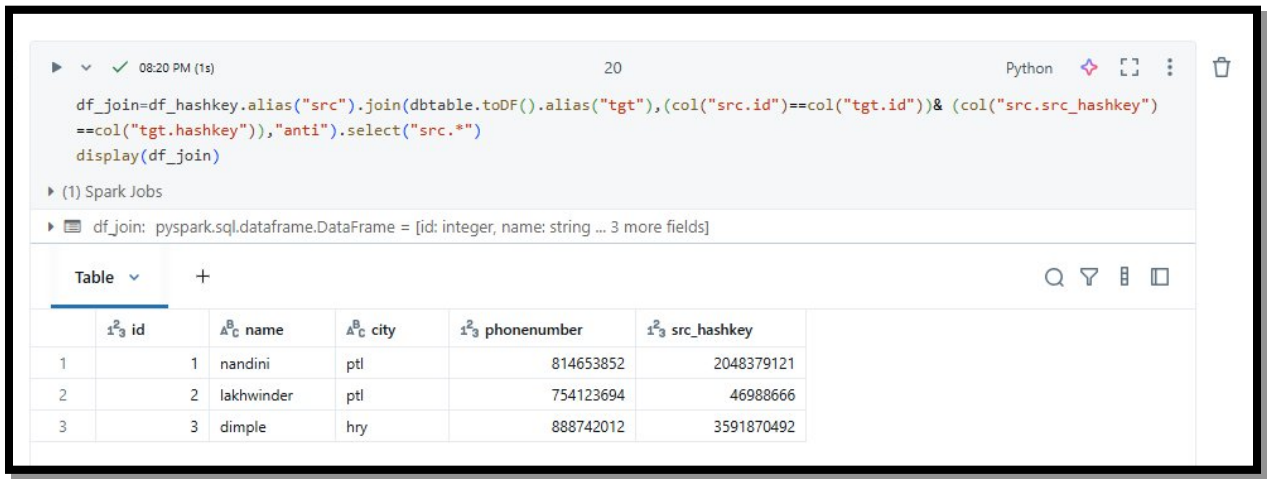
(1) Spark Jobs

df_join: pyspark.sql.dataframe.DataFrame = [id: integer, name: string ... 3 more fields]

Table +

	id	name	city	phonenumber	src_hashkey
1	1	nandini	ptl	814653852	2048379121
2	2	lakhwinder	ptl	754123694	46988666
3	3	dimple	hry	888742012	3591870492

"anti" ☾ Left anti join: returns only the rows from src (left side) that do not have a matching row in tgt based on the condition.



08:20 PM (1s) 20 Python

```
df_join=df_hashkey.alias("src").join(dbtable.toDF().alias("tgt"),(col("src.id")==col("tgt.id"))& (col("src.src_hashkey")
==col("tgt.hashkey")),"anti").select("src.*")
display(df_join)
```

(1) Spark Jobs

df_join: pyspark.sql.dataframe.DataFrame = [id: integer, name: string ... 3 more fields]

Table +

	id	name	city	phonenumber	src_hashkey
1	1	nandini	ptl	814653852	2048379121
2	2	lakhwinder	ptl	754123694	46988666
3	3	dimple	hry	888742012	3591870492

Join src table with target table,on the basis of id and hashkey.

Src_id==tgt_id

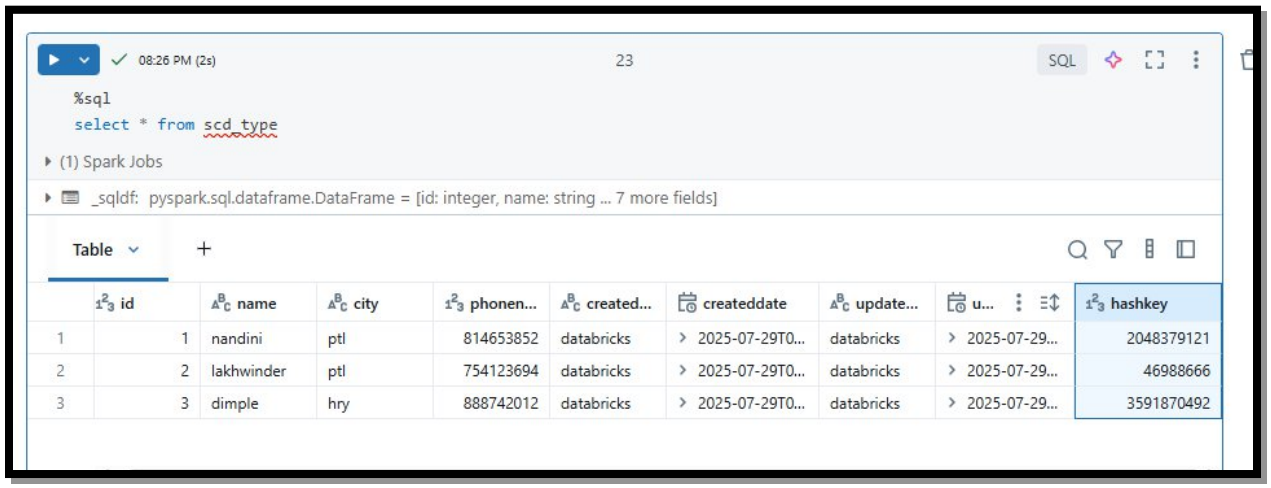
Src_hashkey==tgt_hashkey.

final merge

08:24 PM (23s) 22

```
dbtable.alias("tgt").merge(df_join.alias("src"),"src.id==tgt.id")\
  .whenMatchedUpdate(set={"tgt.id":"src.id",
    "tgt.name":"src.name","tgt.city":"src.city","tgt.phonenumber":"src.phonenumber","tgt.updateddate":"current_timestamp()
    ","tgt.updatedby":lit("databricks-updated"),"tgt.hashkey":"src.src_hashkey"
  })\
    .whenNotMatchedInsert(values={"tgt.id":"src.id",
    "tgt.name":"src.name","tgt.city":"src.city","tgt.phonenumber":"src.phonenumber","tgt.updateddate":"current_timestamp()
    ","tgt.updatedby":lit("databricks"),"tgt.createdby":lit("databricks"),"tgt.createddate":"current_timestamp()","tgt.
    hashkey":"src.src_hashkey"
    })\
      .execute()
```

(6) Spark Jobs

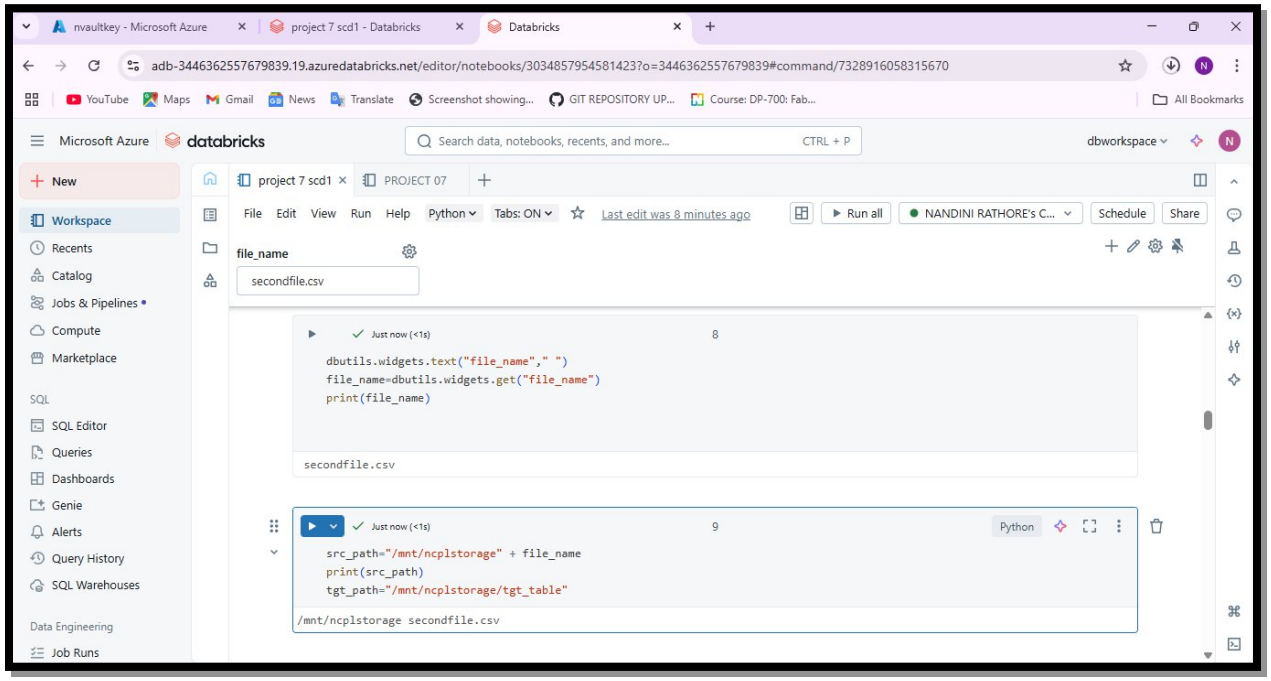


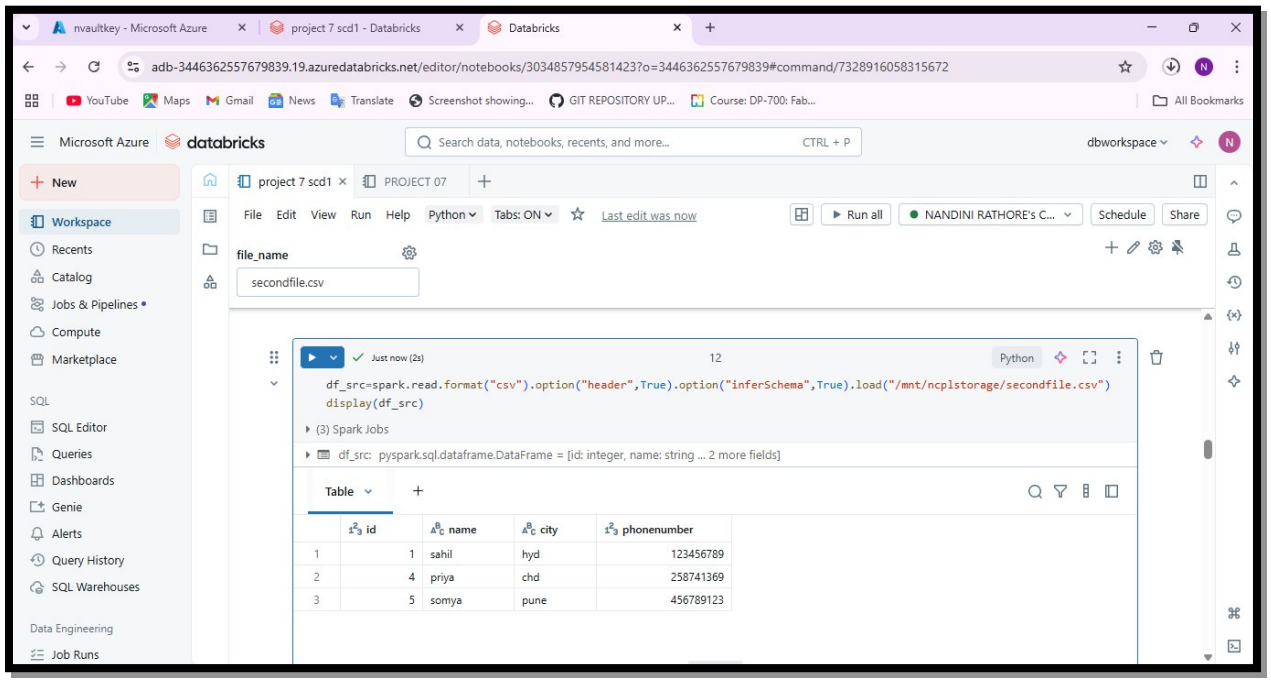
	id	name	city	phonen...	created...	createddate	update...	u...	hashkey
1	1	nandini	ptl	814653852	databricks	2025-07-29T0...	databricks	2025-07-29...	2048379121
2	2	lakhwinder	ptl	754123694	databricks	2025-07-29T0...	databricks	2025-07-29...	46988666
3	3	dimple	hry	888742012	databricks	2025-07-29T0...	databricks	2025-07-29...	3591870492

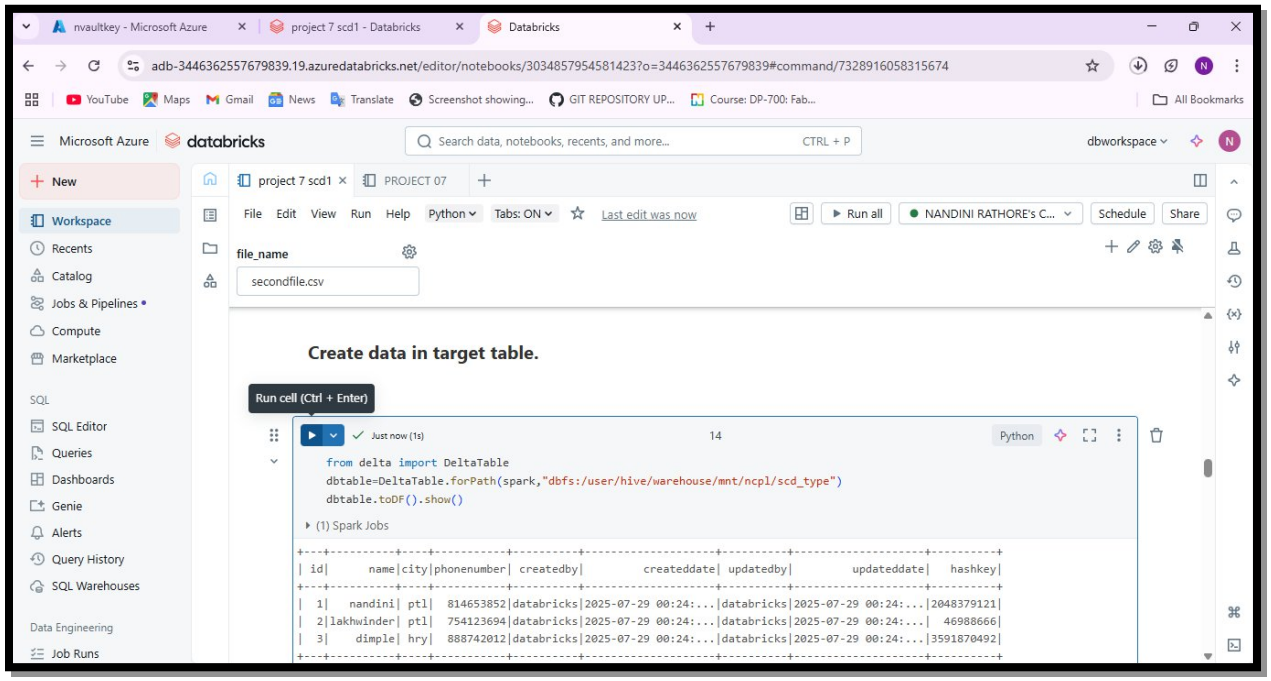
Target table gives data of new inserted values.

I used another file from adls storage account and update,insert values.

Now add new values and update 1 row.

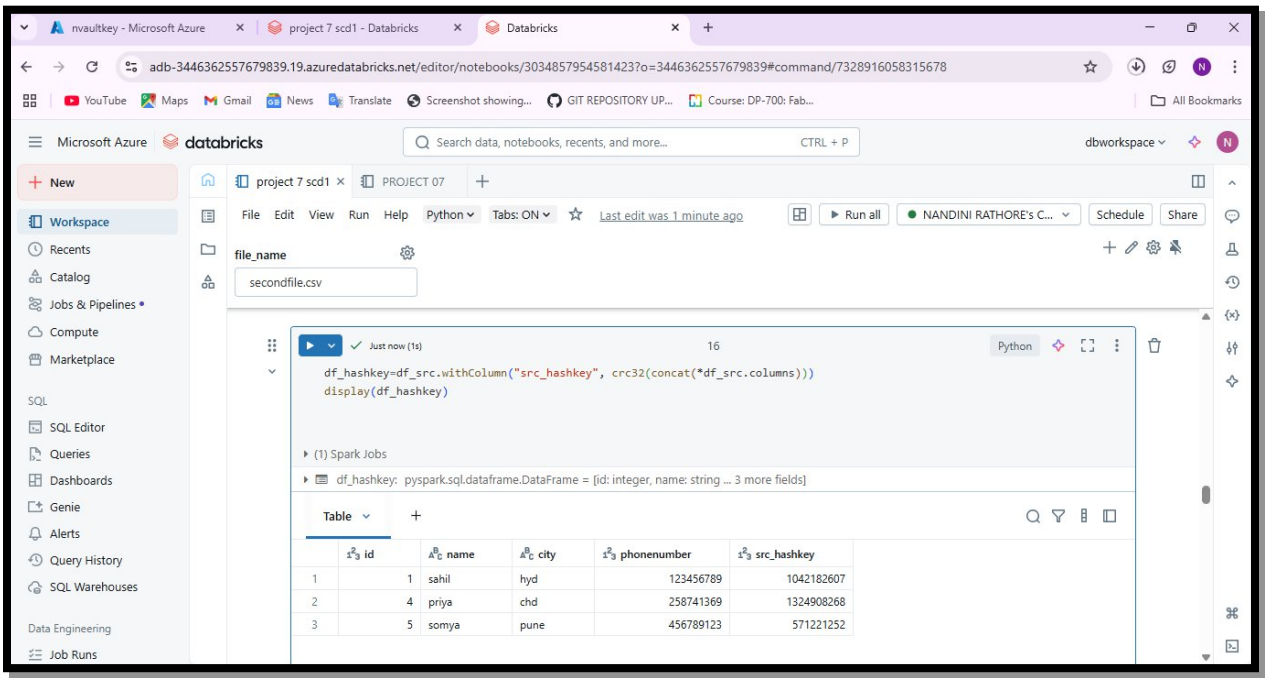


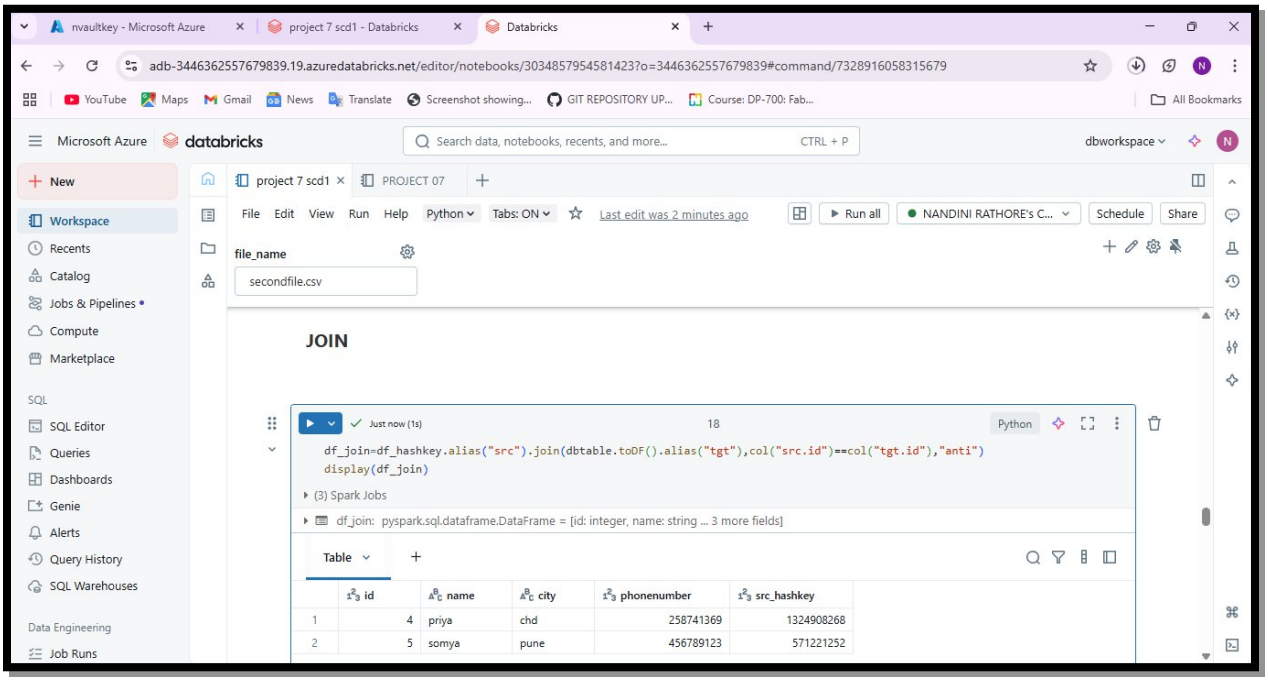


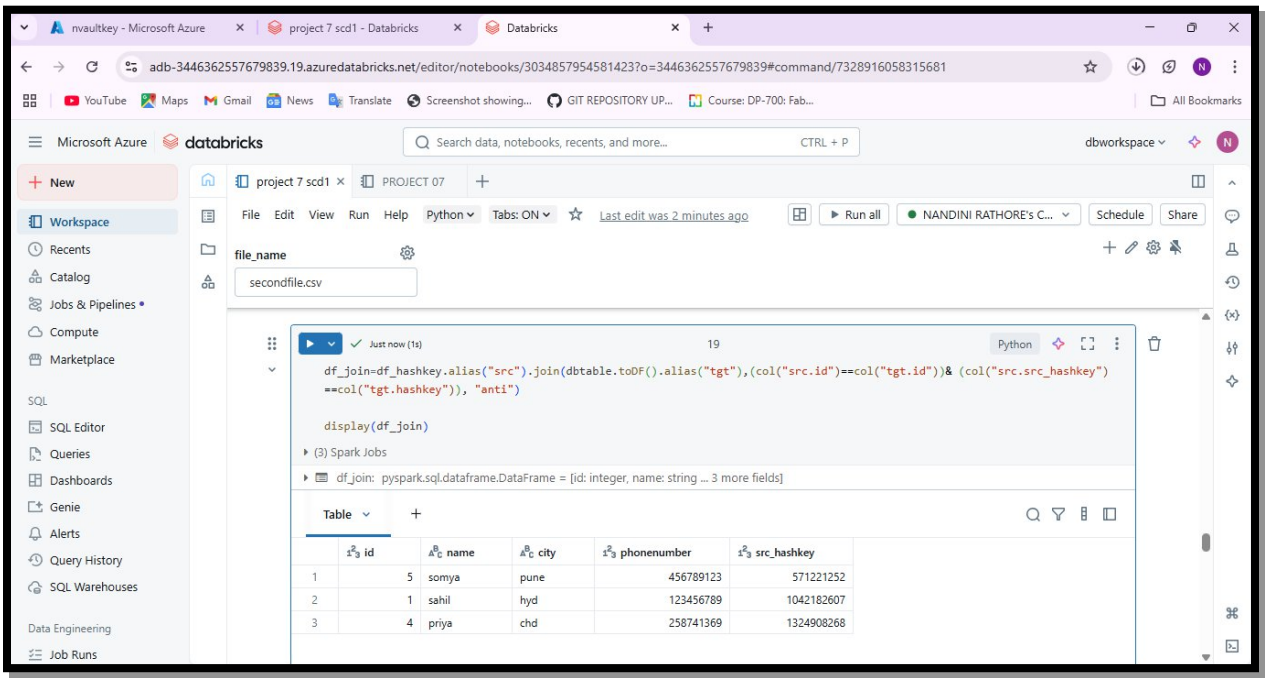


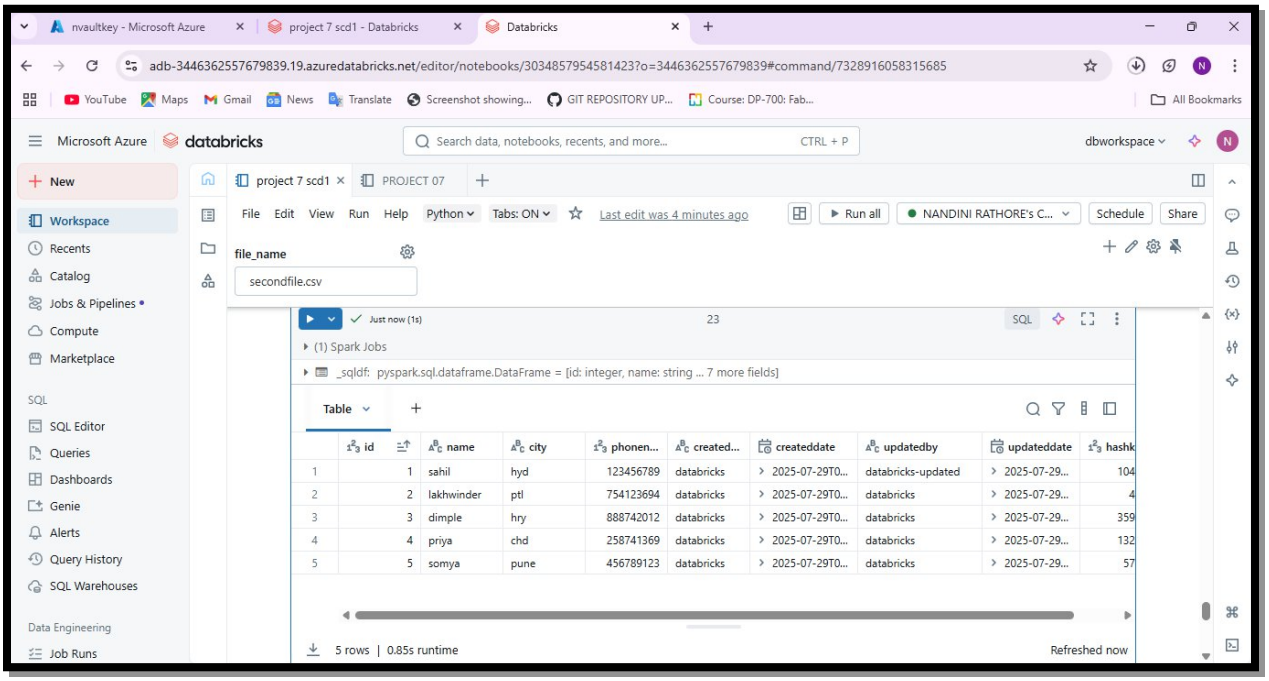
Now it is showing data in target table.

NANDINI RATHORE

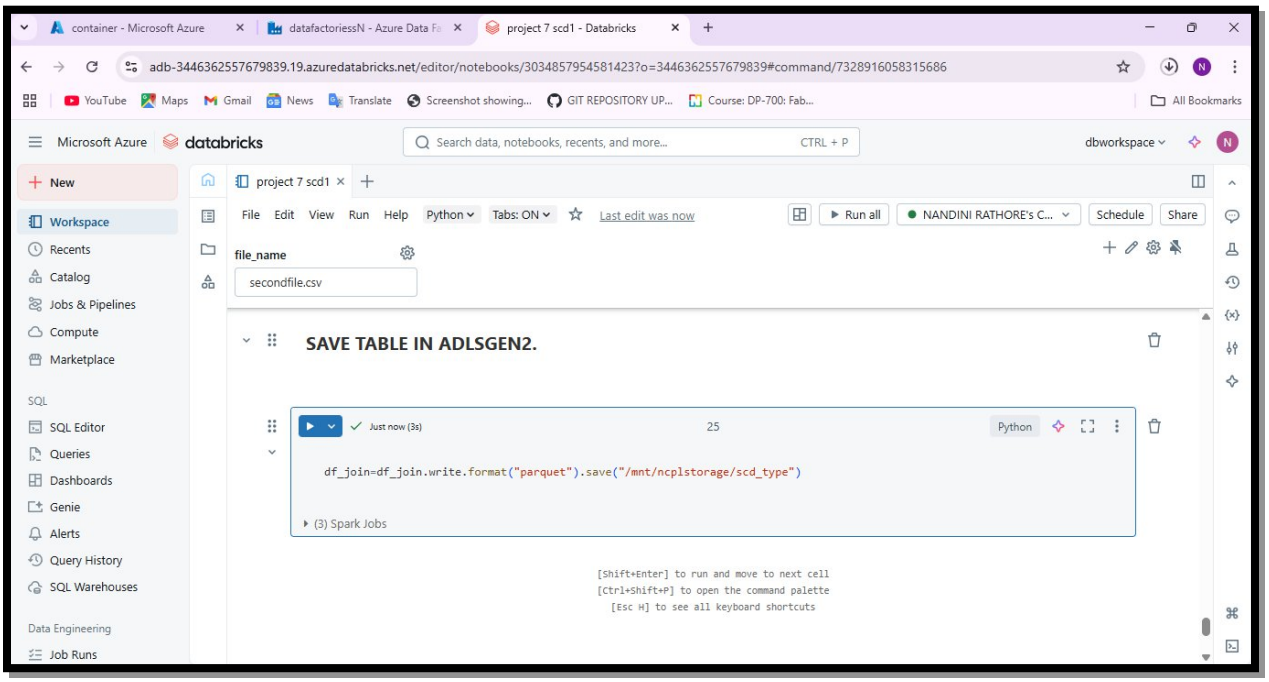


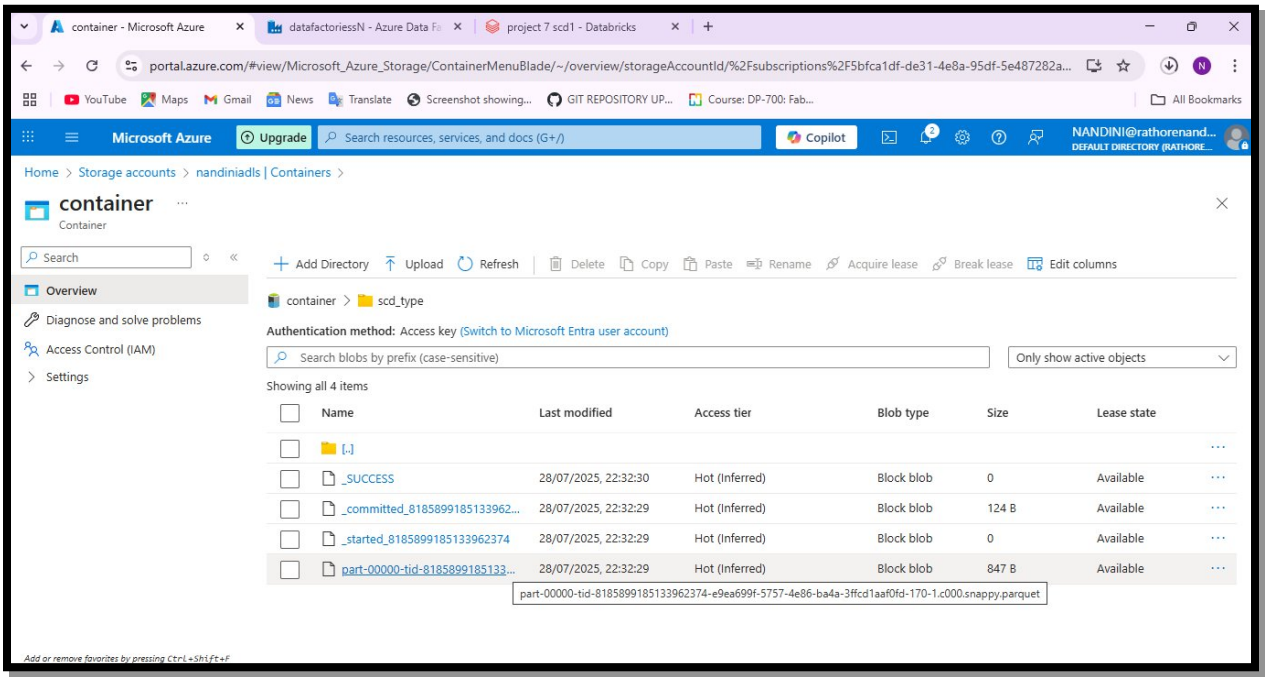


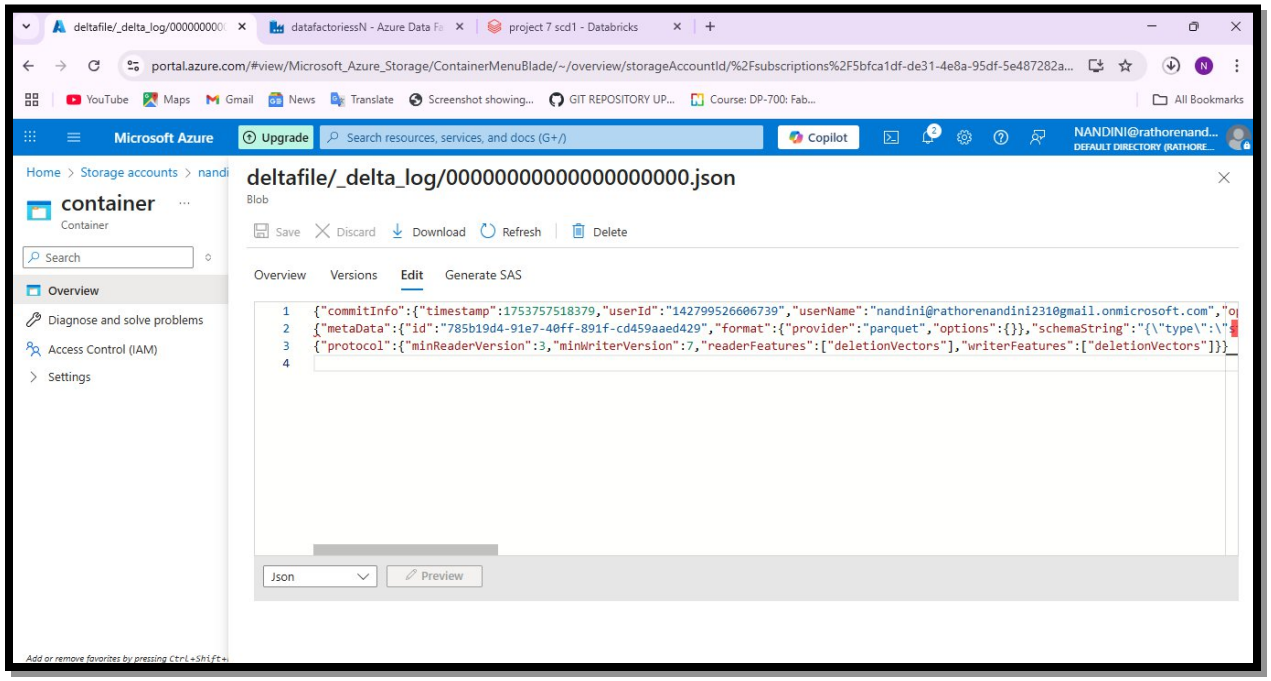




1 row is updated as it get “ databricks-updated” in updatedby column And row 4,5 are new inserted values. In scd typ 1 it keeps the record of new inserted and updated rows, didn’t keep historical data.

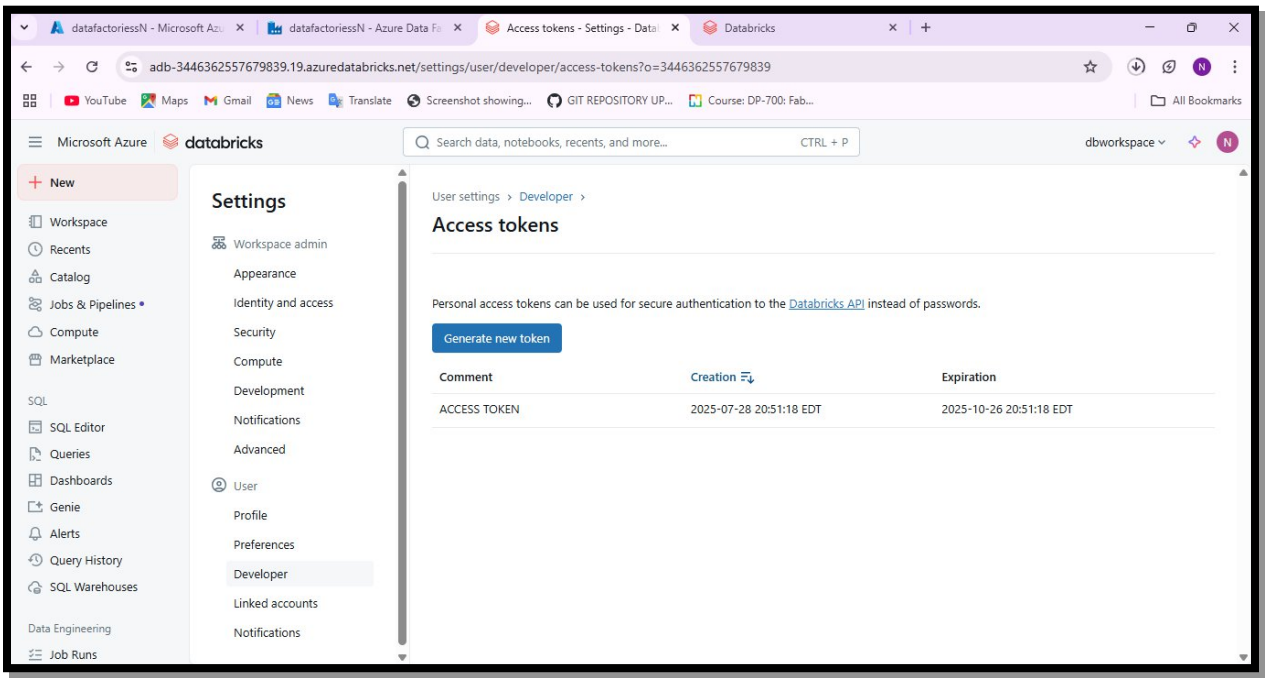

container - Microsoft Azure
datafactoriessN - Azure Data Fa
project 7 scd1 - Databricks
adb-3446362557679839.19.azuredatabricks.net/editor/notebooks/3034857954581423?o=3446362557679839#command/7328916058315686
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
dbworkspace
New
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
project 7 scd1
File Edit View Run Help Python Tabs: ON
Last edit was now
Run all
NANDINI RATHORE's C...
Schedule
Share
file_name
secondfile.csv
SAVE TABLE IN ADLSGEN2.
Just now (3s)
25
Python
df_join=df_join.write.format("parquet").save("/mnt/ncplstorage/scd_type")
(3) Spark Jobs
[Shift+Enter] to run and move to next cell
[Ctrl+Shift+P] to open the command palette
[Esc H] to see all keyboard shortcuts


container - Microsoft Azure
datafactoriessN - Azure Data Fa
project 7 scd1 - Databricks
portal.azure.com/#view/Microsoft_Azure_Storage/ContainerMenuBlade/~/overview/storageAccountId/%2Fsubscriptions%2F5bfca1df-de31-4e8a-95df-5e487282a...
Microsoft Azure
Upgrade
Search resources, services, and docs (G+/)
Copilot
NANDINI@rathorenand...
DEFAULT DIRECTORY (RATHORE...
Home > Storage accounts > nandiniadls | Containers >
container
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Add Directory
Upload
Refresh
Delete
Copy
Paste
Rename
Acquire lease
Break lease
Edit columns
container > scd_type
Authentication method: Access key (Switch to Microsoft Entra user account)
Search blobs by prefix (case-sensitive)
Only show active objects
Showing all 4 items
Name
Last modified
Access tier
Blob type
Size
Lease state
[..]
_SUCCESS
28/07/2025, 22:32:30
Hot (Inferred)
Block blob
0
Available
_committed_8185899185133962...
28/07/2025, 22:32:29
Hot (Inferred)
Block blob
124 B
Available
_started_8185899185133962374
28/07/2025, 22:32:29
Hot (Inferred)
Block blob
0
Available
part-00000-tid-8185899185133...
28/07/2025, 22:32:29
Hot (Inferred)
Block blob
847 B
Available
part-00000-tid-8185899185133962374-e9ea699f-5757-4e86-ba4a-3ffcd1aaf0fd-170-1.c000.snappy.parquet
Add or remove favorites by pressing Ctrl+Shift+F



The _delta_log/ folder and its .json files:

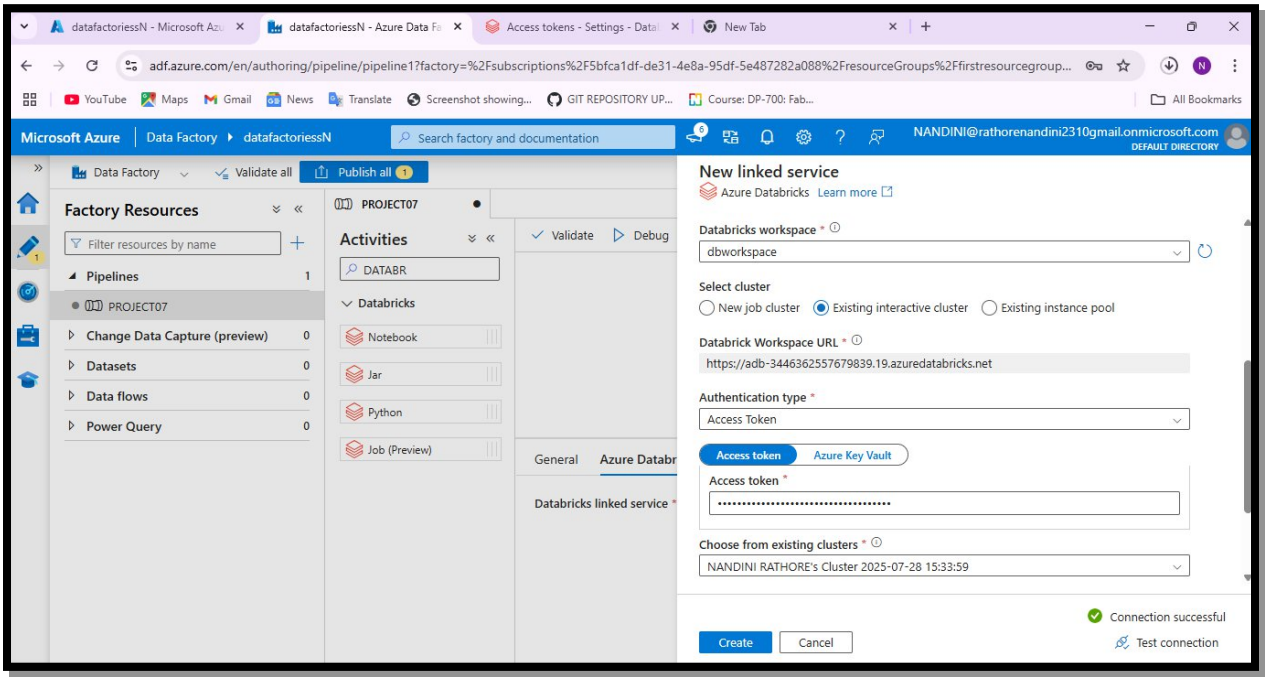
- Are required for Delta Lake to function
- Store metadata and transactional history
- Make Delta format more powerful than plain Parquet.
- Delta lake are formed on the top of parquet format supports ACID transactions,schema evolution, time travel, efficient reads.



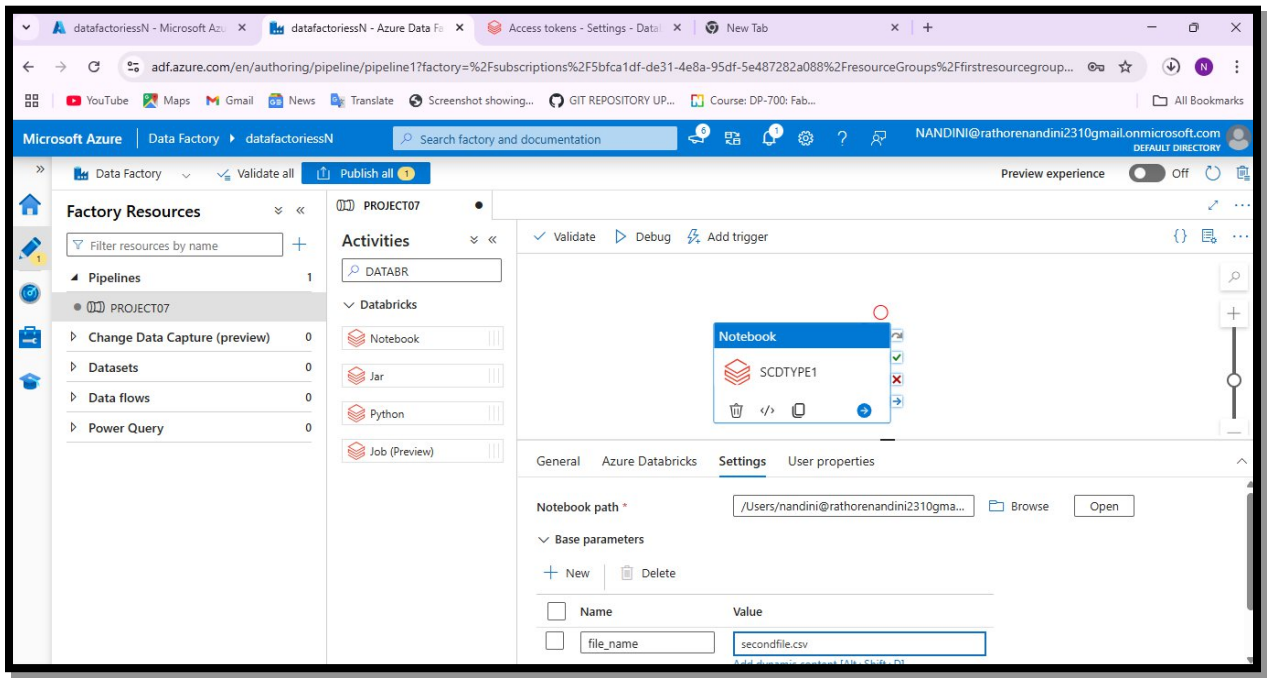
For run pipeline in adf we need access token .

Profileàsettingsàusersàdevelopersàmanageàaccess token.

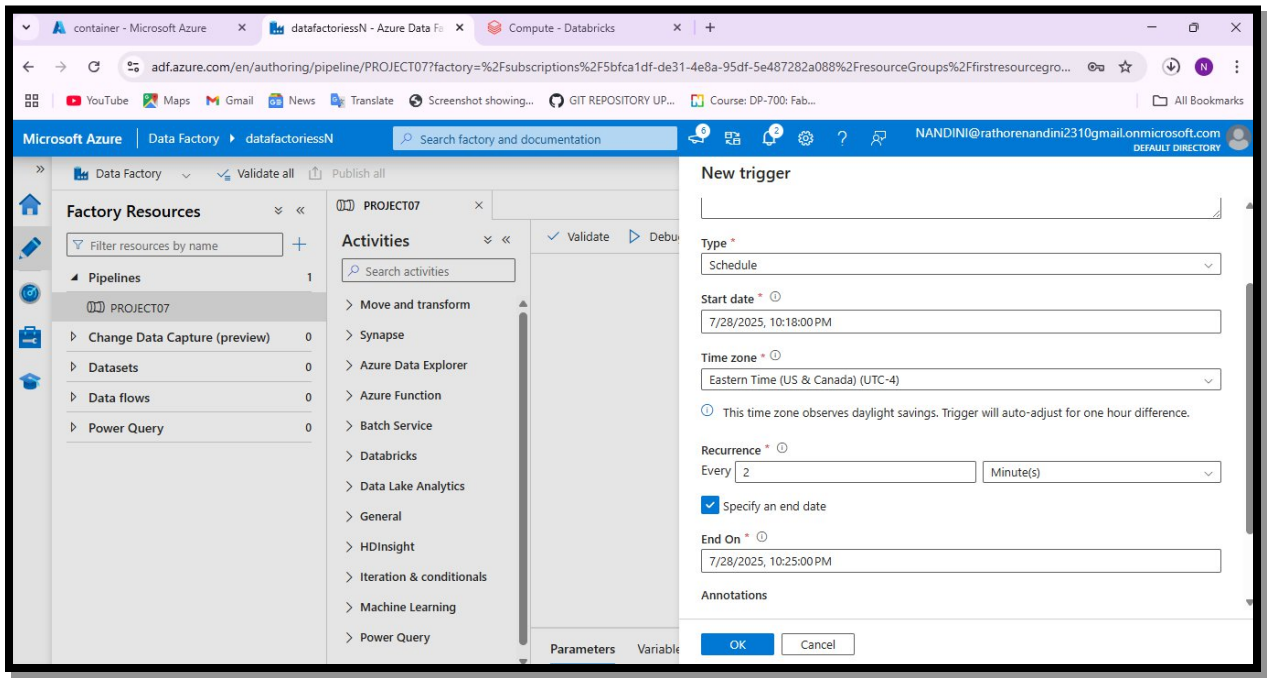
dapi936b47130cd59c9732f9ac8ecc82752f

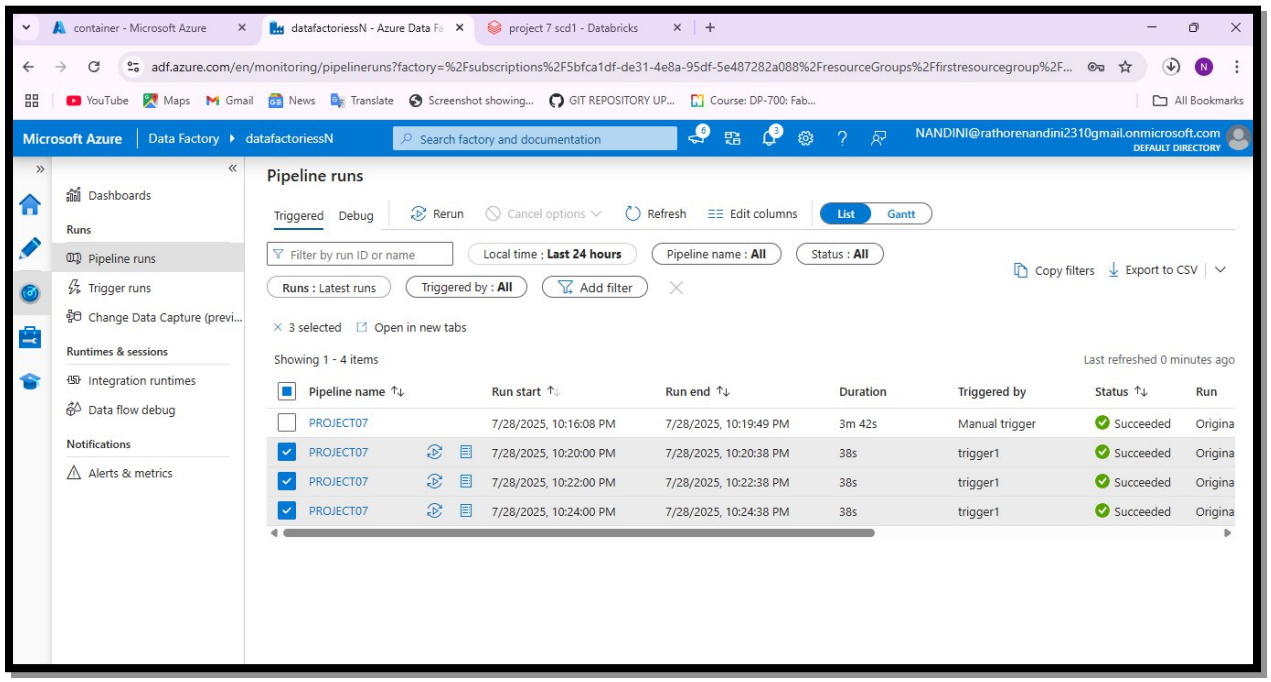


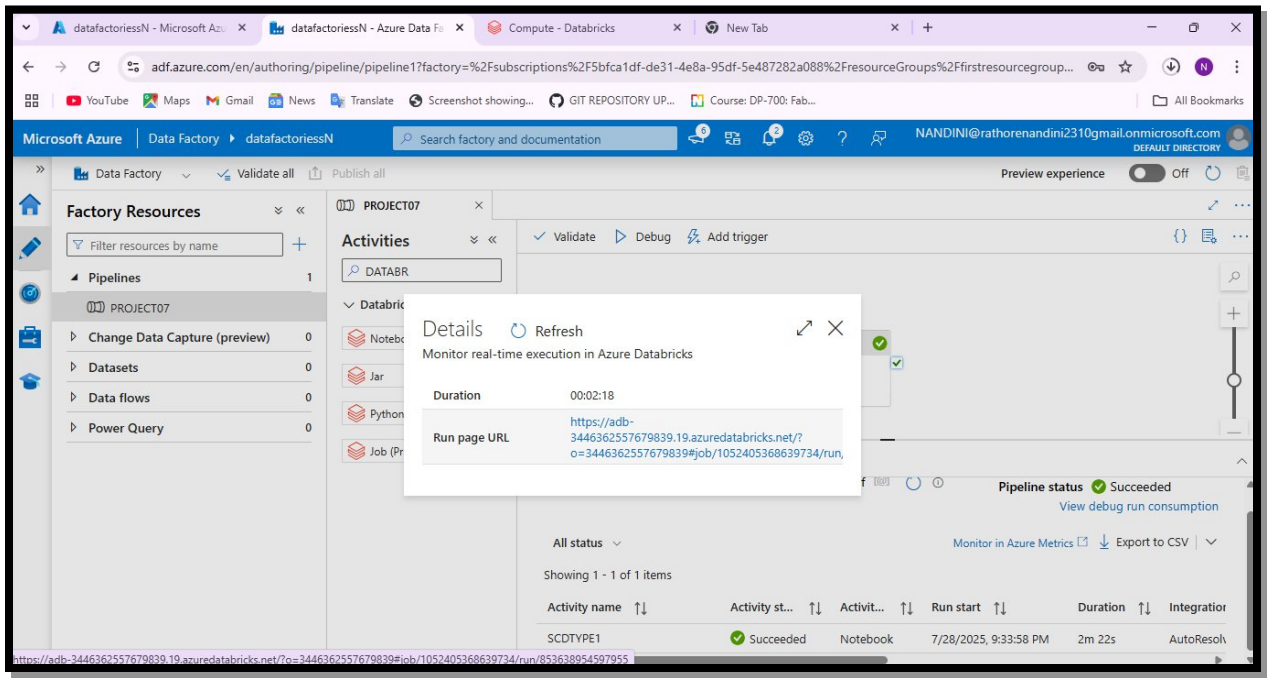
Creating linked service in azure data factory.



Select the notebook path and give parameters which is also used in notebooks.

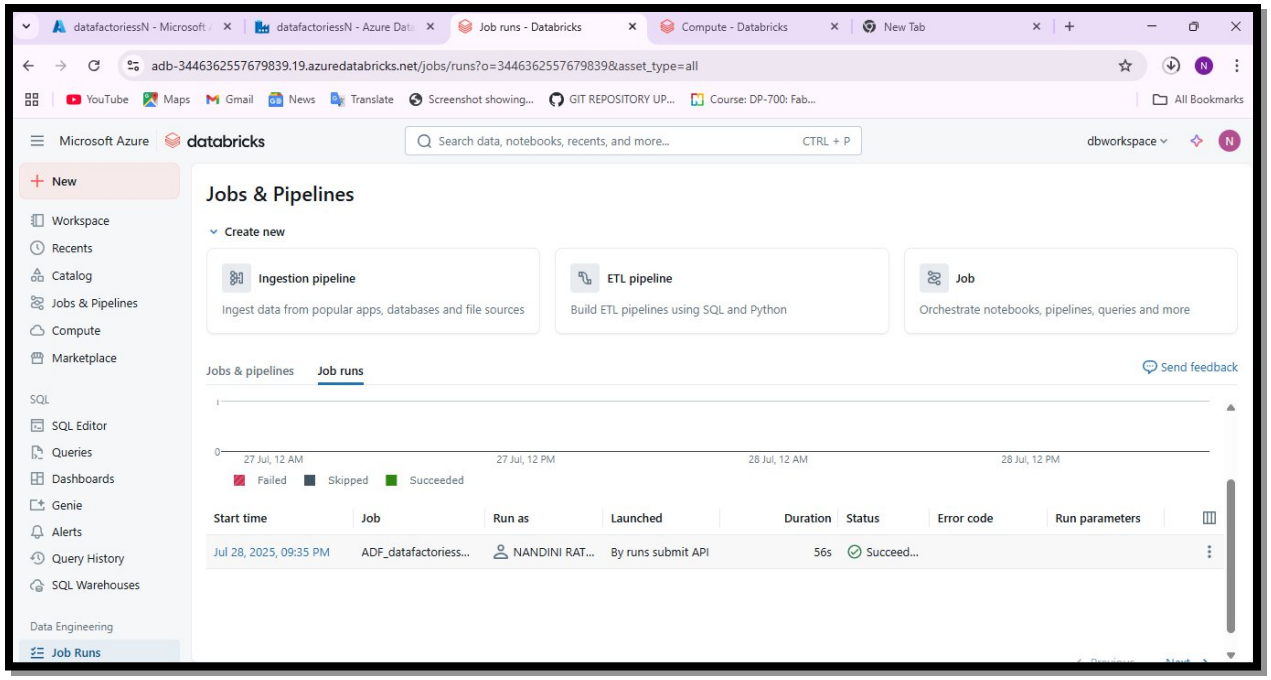


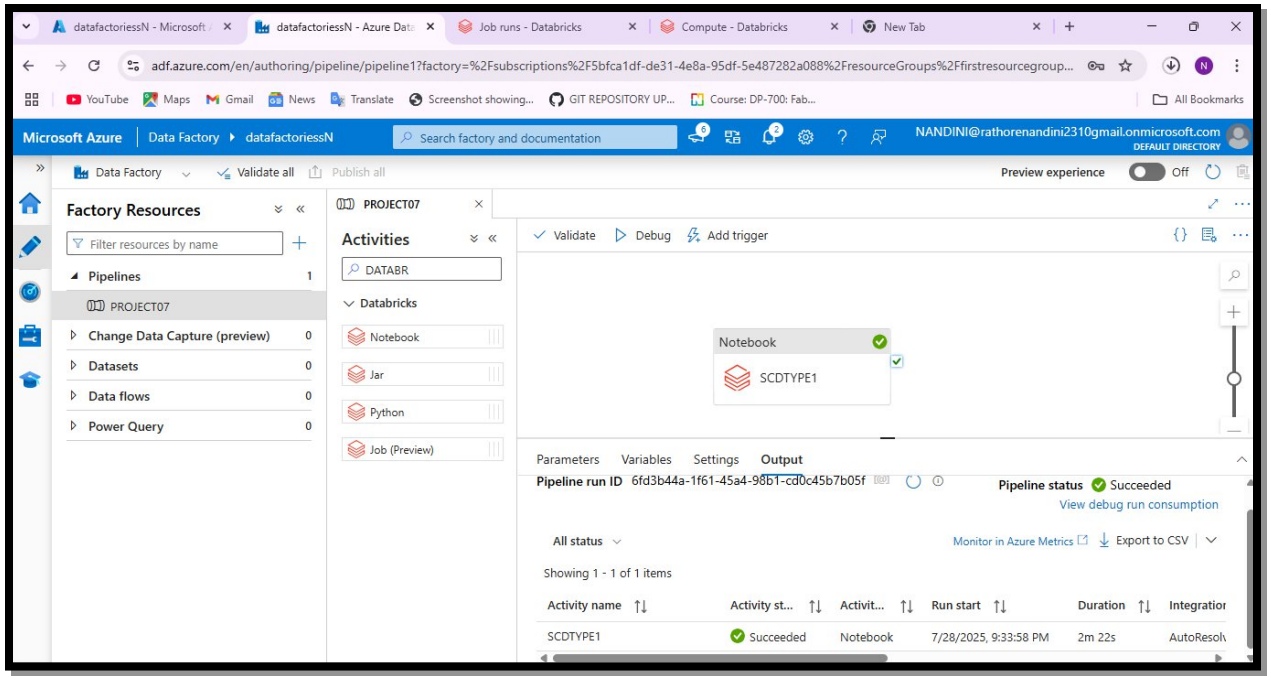




Trigger pipeline and by clicking on this URL, to open the notebook manually to check the code or debug.

Jobs run in databricks.





In this project we use Azure Date Factory , because it is good at moving data, but Databricks Notebooks are better at transforming **or** cleaning the data using Python or Spark.

CONCLUSION:->

- dbutils.widgets are used to **create input controls** in your Databricks notebook, so we can make code dynamic, interactive, and reusable.
- Pass different values (like filenames, dates) without changing the code.
- Run the same notebook with different inputs.

- In this project we use Azure Date Factory , because it is good at moving data, but Databricks Notebooks are better at transforming **or** cleaning the data using Python or Spark.
- We use ADF to run notebooks so we can automate and combine: Data copying (by ADF)
- Data cleaning/processing (by notebook),In one pipeline.
- dbutils.fs.mount() doesn't work with Unity Catalog:

Unity Catalog is a new system in Databricks that helps manage data access securely and centrally.

But It blocks dbutils.fs.mount() because:

- Mounted storage can’t be tracked or controlled by Unity Catalog.
- Unity Catalog wants full control and security over data access.
- Mounts could bypass security rules, so they are not allowed.