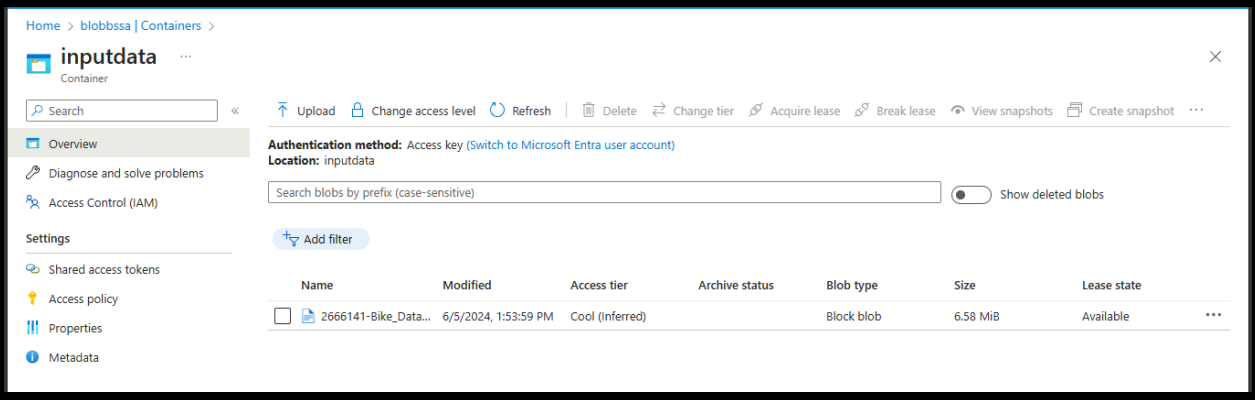
8th pipeline : Blob Storage → Azure Databricks → Azure SQL Database :

Services required :

Azure Blob Storage
Azure Databricks
Azure SQL Database
Azure key vault

Creation of azure blob storage :

→ The creation of azure blob storage is shown in the document below.
https://docs.google.com/document/d/1vZCMfM9ieAlTQm6Jdwl2JzxcMkwnyvkSJHCrry/edit2kVIo?usp=sharing
→ Create a container : inputData
→ Upload bike data in the inputData container



Creation of Azure SQL Database.
→ Creation of Azure SQL Database is shown in the document below.
https://docs.google.com/document/d/16iB1EsGKHc6-bcgTPSfqkK6BVf3n8_fpbat42uNOvXc/edit?usp=sharing

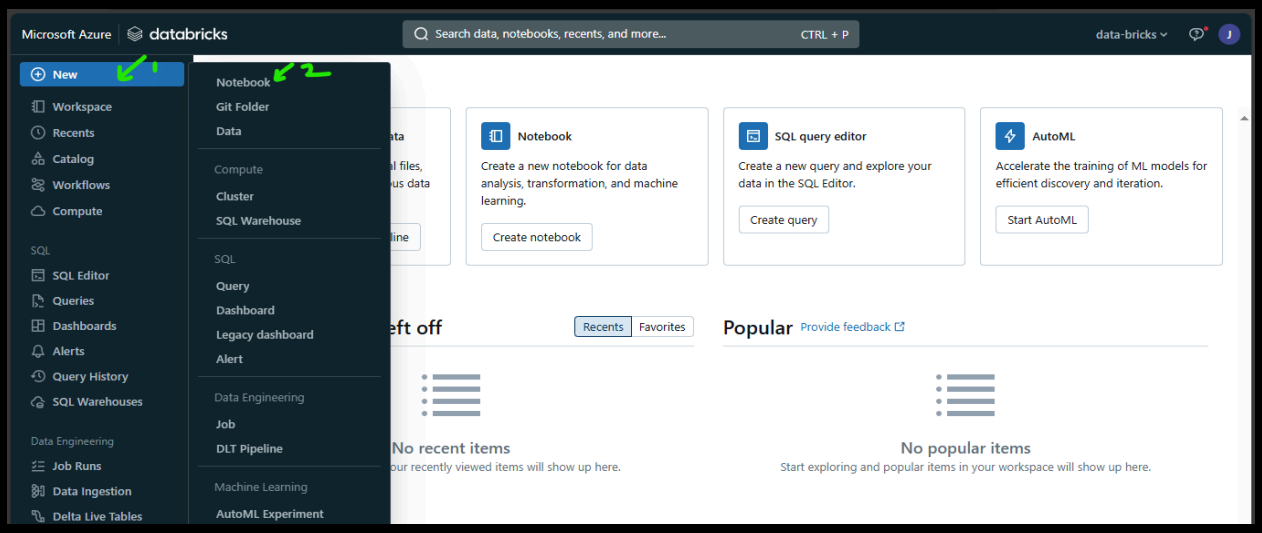
Creation of Azure Databricks
→ Creation of Azure SQL Databricks is shown in the document below.
https://docs.google.com/document/d/1s7Vqs4gcZbGMPRk7vEWGFBXIMC5AoaOvC-E57qILwRw/edit?usp=sharing

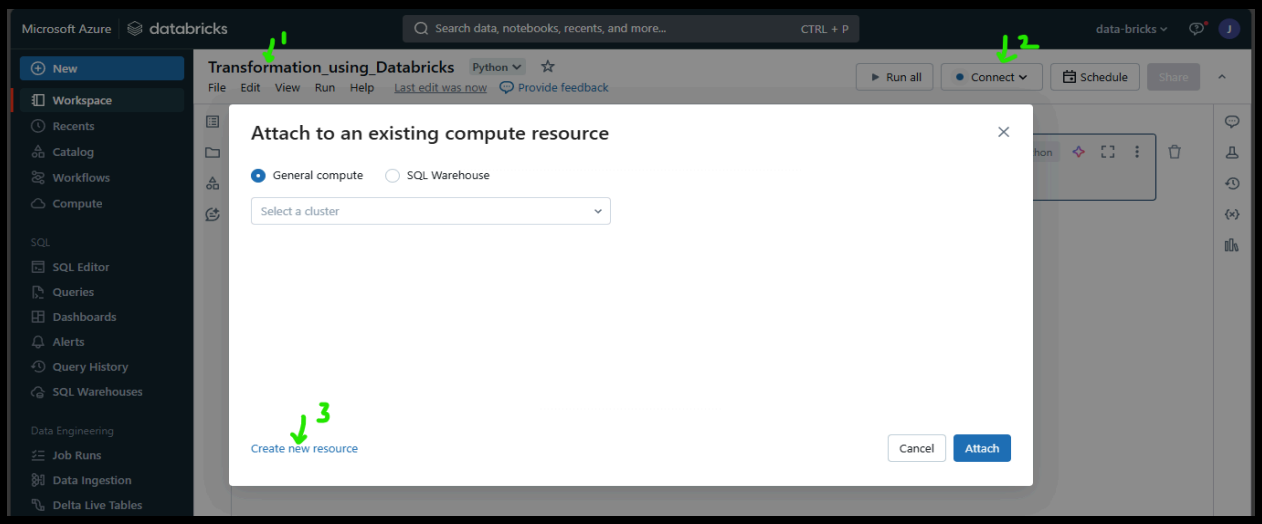
Creation of azure key vault :

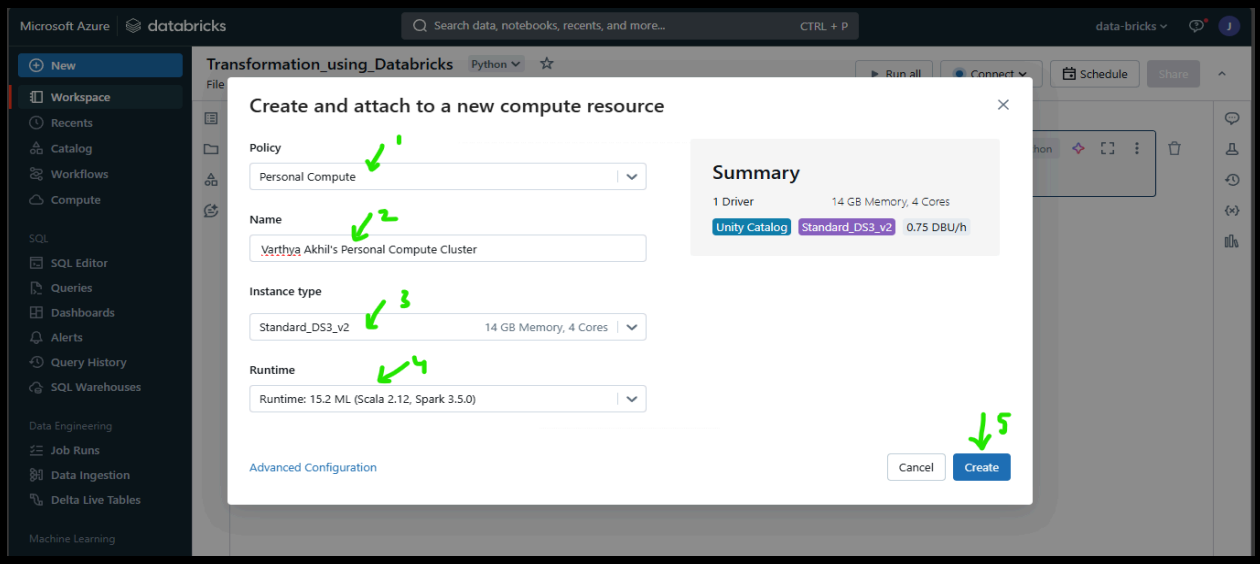
→ Creation of Azure Key Vault is shown in the document below.

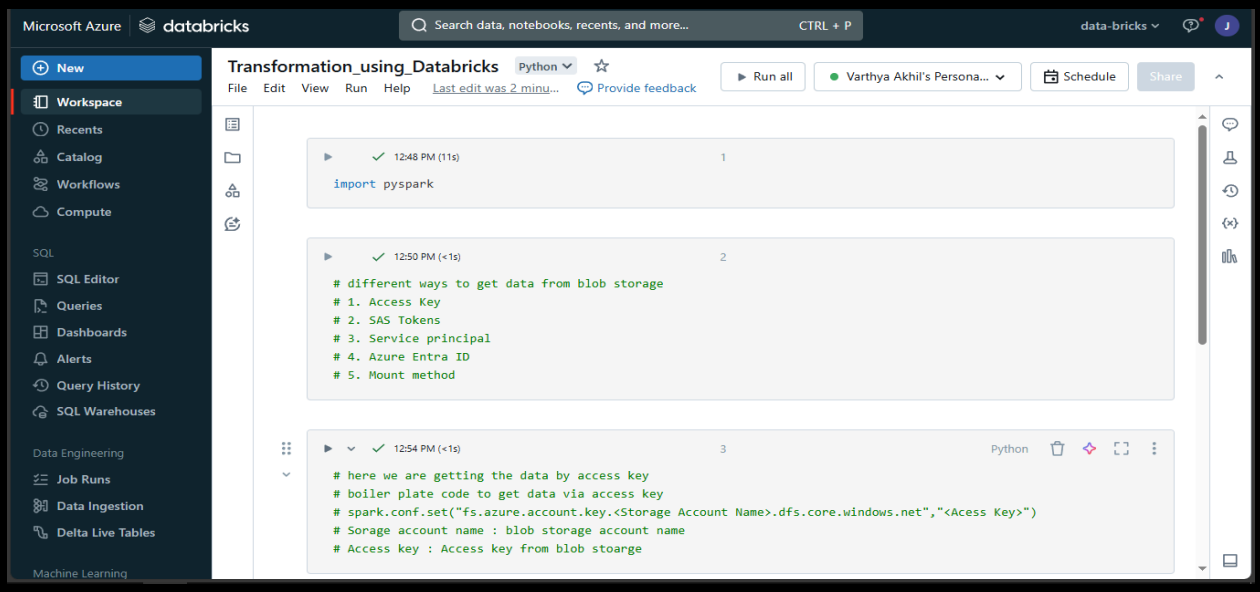
https://docs.google.com/document/d/1ScQ42B5c5ZuRnFjtLsdJpnV9067pLZkpcDf4WyKjaVw/edit?usp=sharing

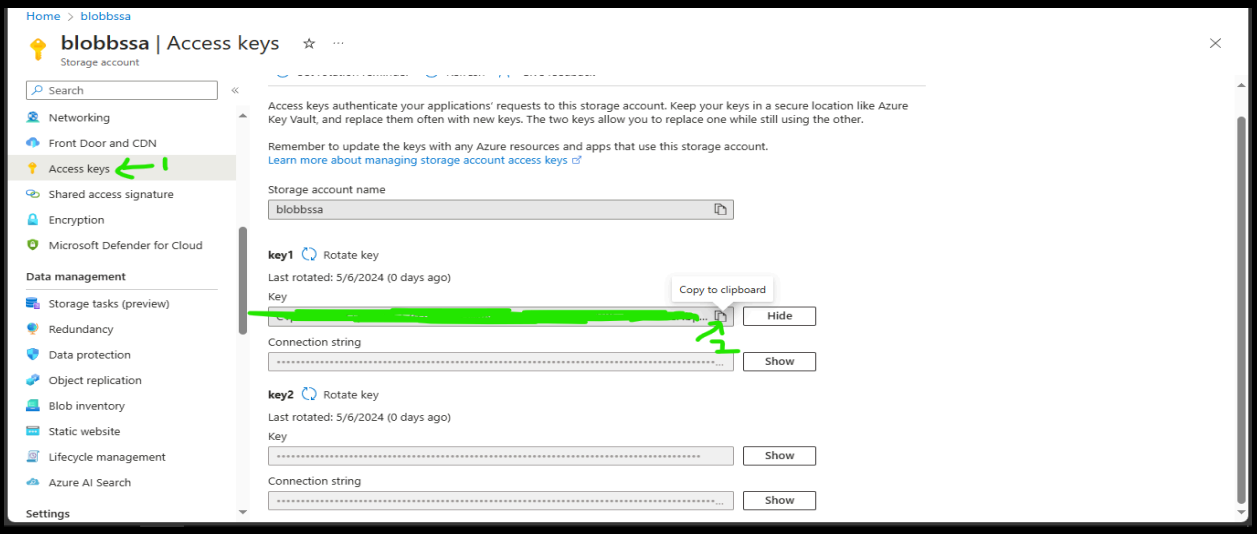
In Databricks workspace →

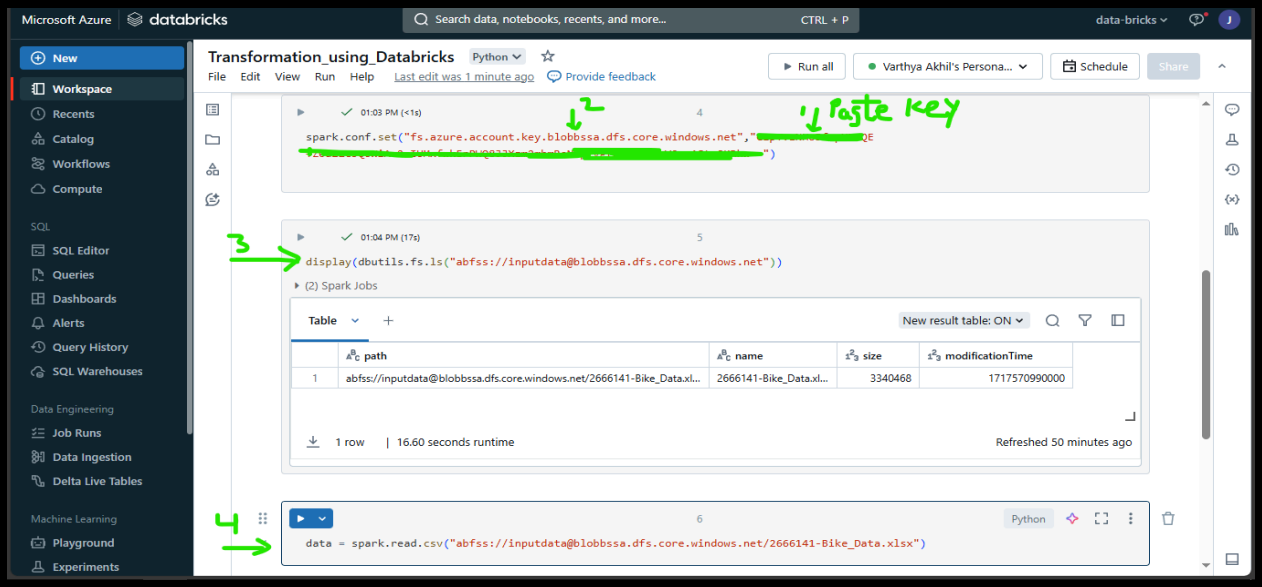




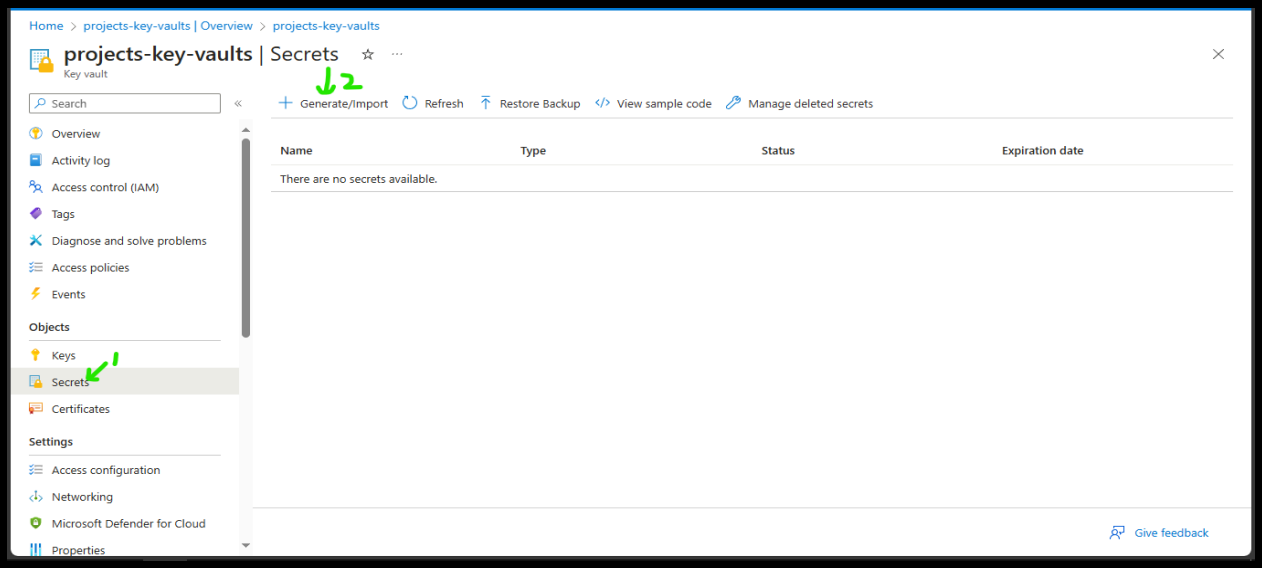

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
data-bricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Transformation_using_Databricks
Python
File
Run all
Connect
Schedule
Share
Create and attach to a new compute resource
Policy
Personal Compute
Name
Varthya Akhil's Personal Compute Cluster
Instance type
Standard_DS3_v2
14 GB Memory, 4 Cores
Runtime
Runtime: 15.2 ML (Scala 2.12, Spark 3.5.0)
Summary
1 Driver
14 GB Memory, 4 Cores
Unity Catalog
Standard_DS3_v2
0.75 DBU/h
Advanced Configuration
Cancel
Create
1
2
3
4
5

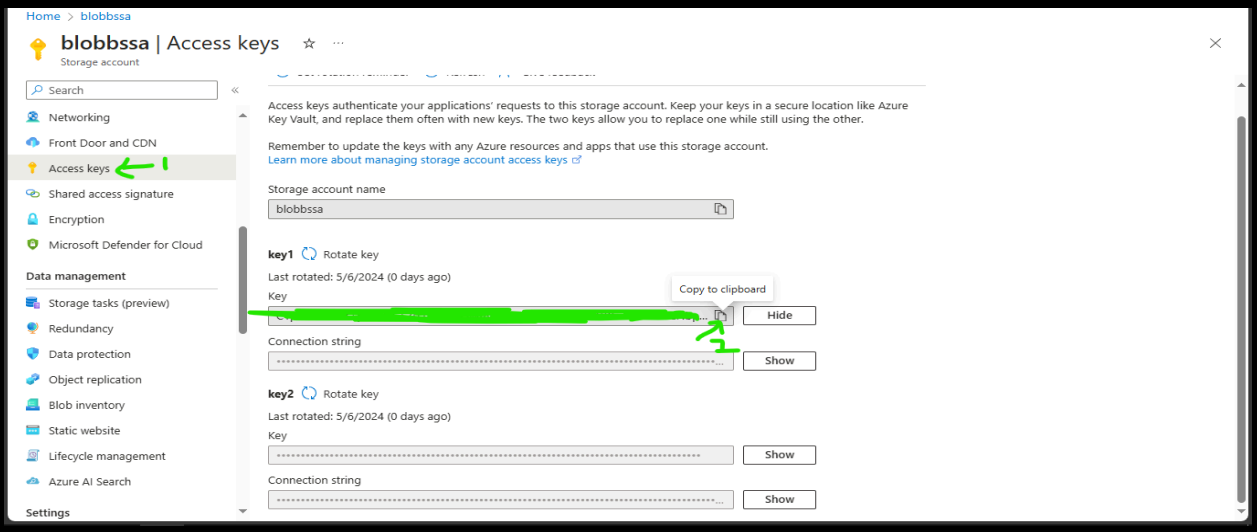

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
data-bricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Transformation_using_Databricks
Python
File
Edit
View
Run
Help
Last edit was 2 minu...
Provide feedback
Run all
Varthya Akhil's Persona...
Schedule
Share
12:48 PM (11s)
1
import pyspark
12:50 PM (<1s)
2
different ways to get data from blob storage
1. Access Key
2. SAS Tokens
3. Service principal
4. Azure Entra ID
5. Mount method
12:54 PM (<1s)
3
Python
here we are getting the data by access key
boiler plate code to get data via access key
spark.conf.set("fs.azure.account.key.<Storage Account Name>.dfs.core.windows.net","<Acess Key>")
Sorage account name : blob storage account name
Access key : Access key from blob stoarge


Home
blobbssa
blobbssa | Access keys
Storage account
Search
Networking
Front Door and CDN
Access keys
Shared access signature
Encryption
Microsoft Defender for Cloud
Data management
Storage tasks (preview)
Redundancy
Data protection
Object replication
Blob inventory
Static website
Lifecycle management
Azure AI Search
Settings
Access keys authenticate your applications' requests to this storage account. Keep your keys in a secure location like Azure Key Vault, and replace them often with new keys. The two keys allow you to replace one while still using the other.
Remember to update the keys with any Azure resources and apps that use this storage account.
Learn more about managing storage account access keys
Storage account name
blobbssa
key1
Rotate key
Last rotated: 5/6/2024 (0 days ago)
Key
Copy to clipboard
Hide
Connection string
Show
key2
Rotate key
Last rotated: 5/6/2024 (0 days ago)
Key
Show
Connection string
Show
1
2

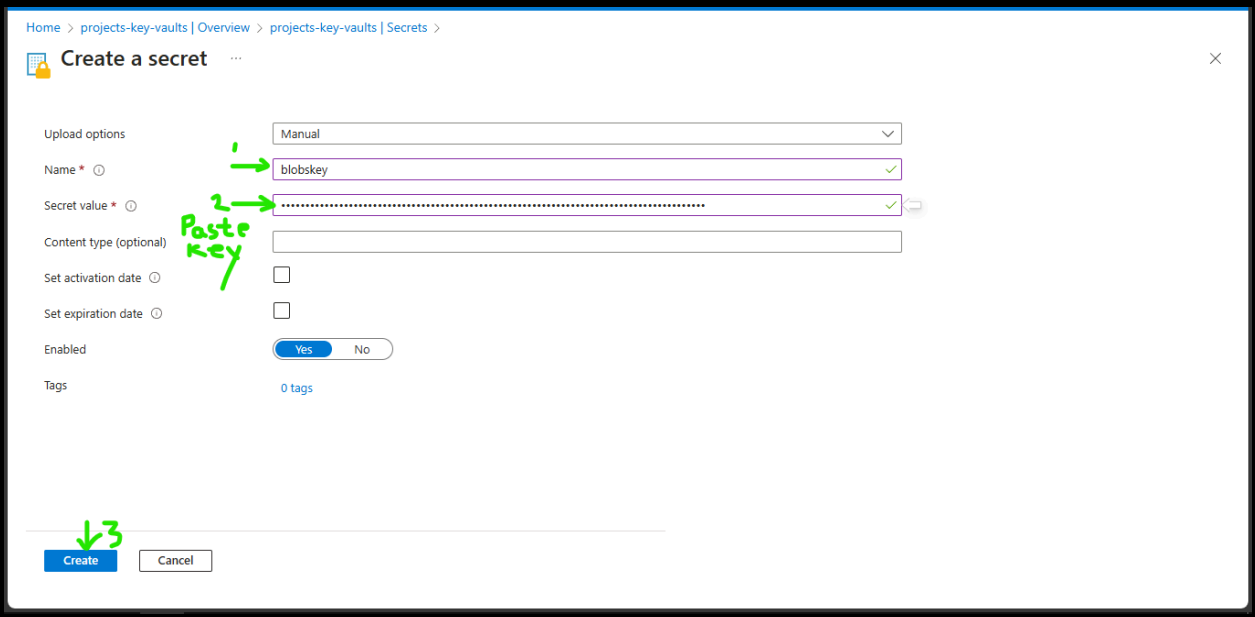


Here, we exposed the access key. It is not a good practice to expose access key. Instead we will import the access key via azure key vault and scope.





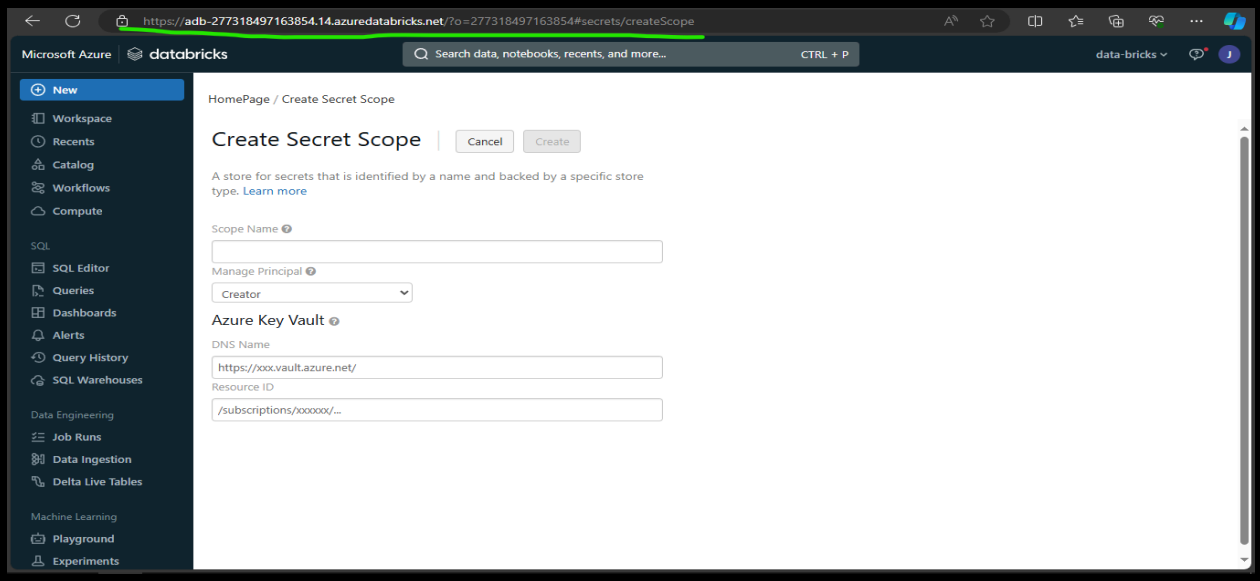
Copy the access key from blob storage and paste in the secret value.

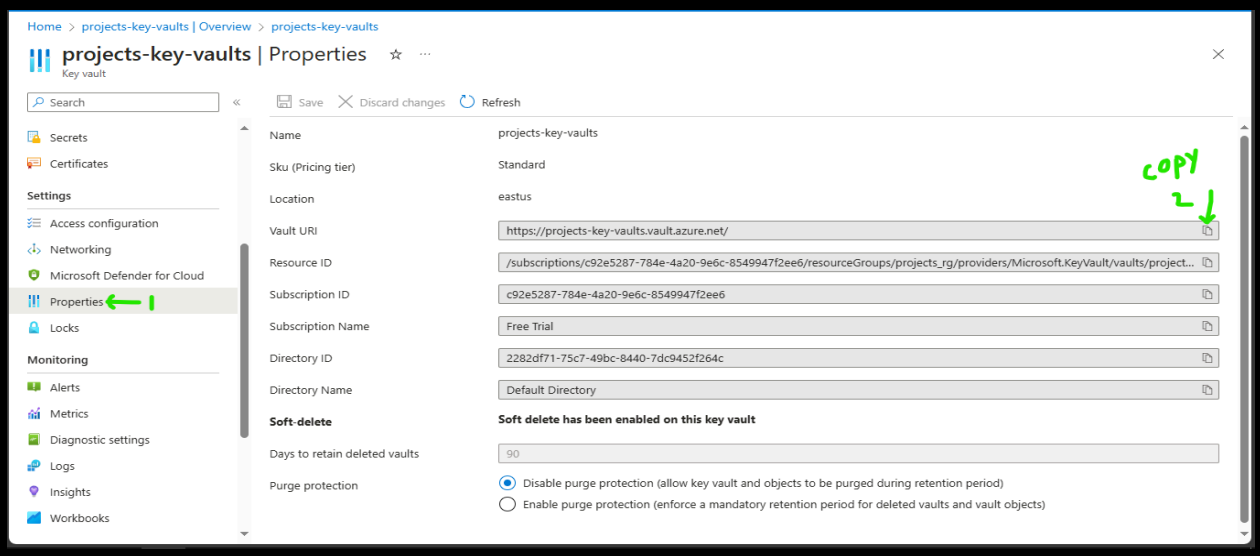


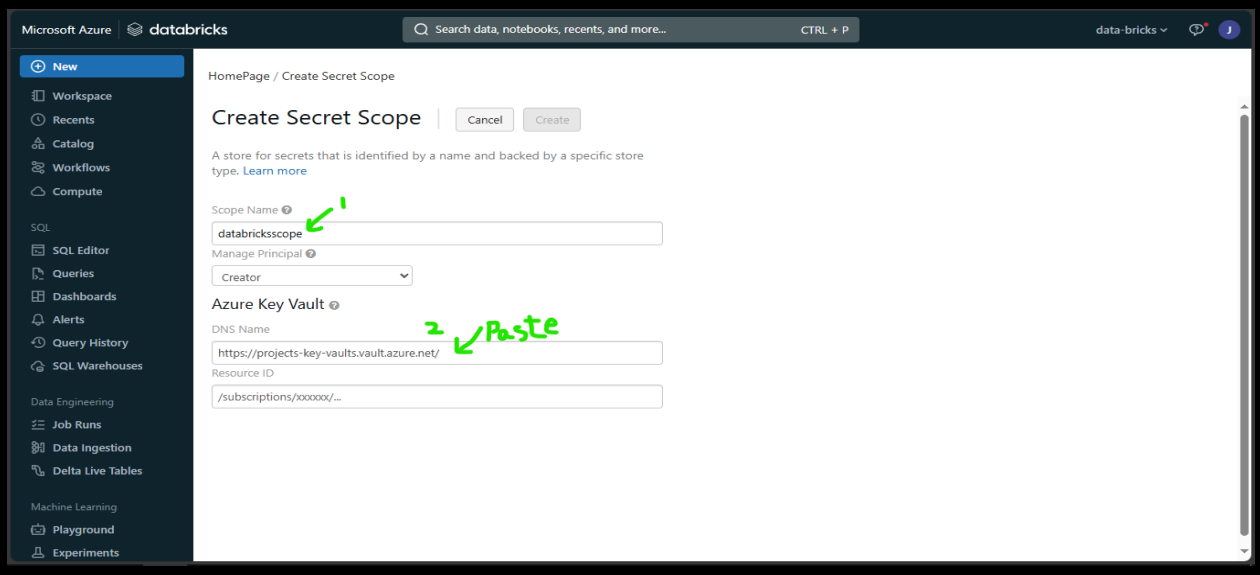
Let's create a scope in databricks.

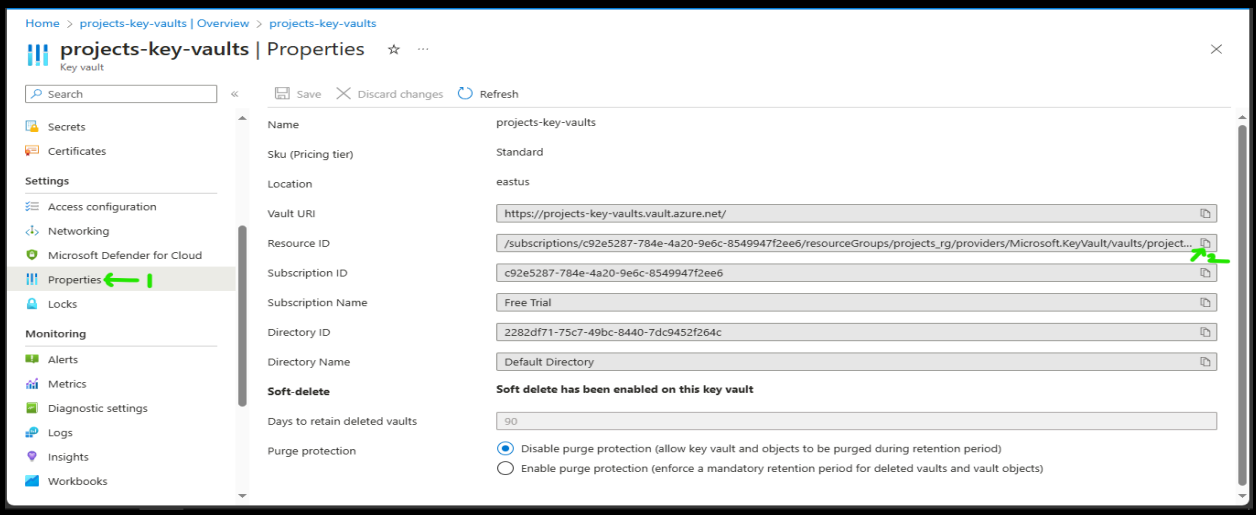
Using this scope we will import the key.

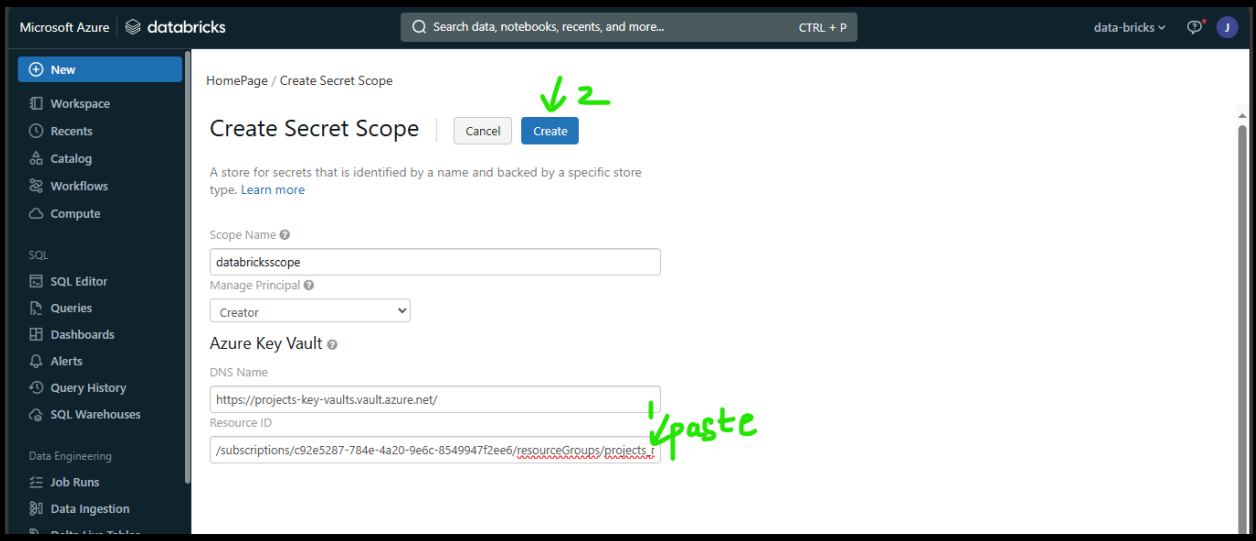
Copy the url, paste in the new tab and remove the characters after # and give secrets/createscope in the url.

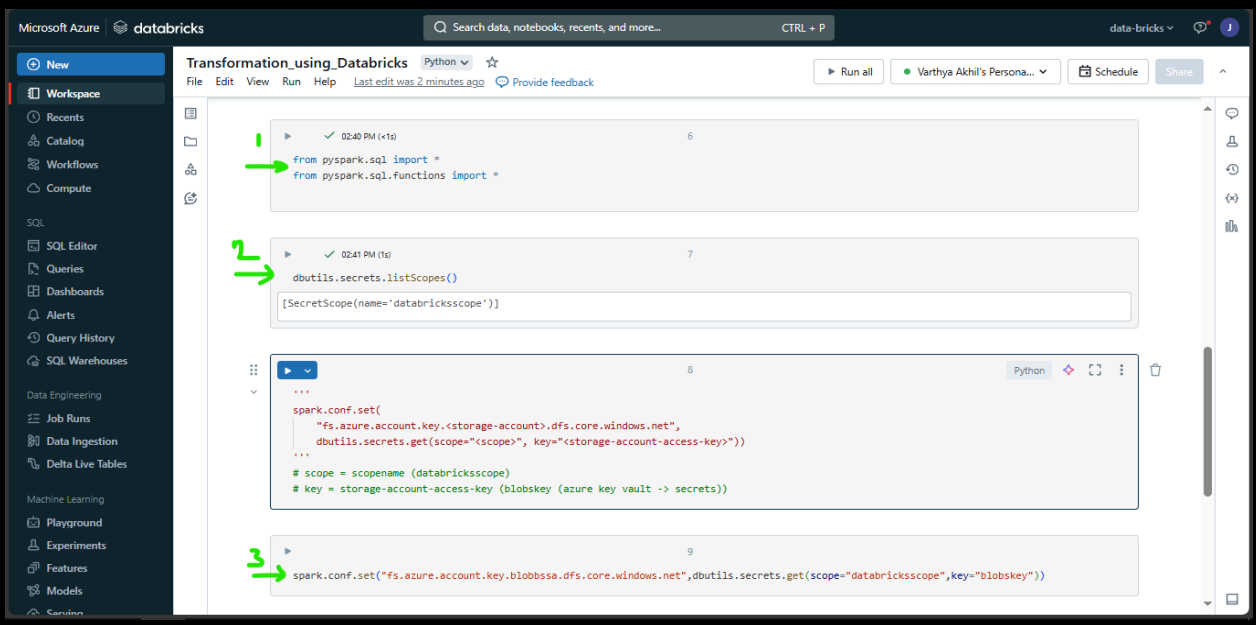

https://adb-2773184971638 54.14.azuredatabricks.net/?o=2773184971638 54#secrets/createScope
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
data-bricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
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
Manage Principal
Creator
Azure Key Vault
DNS Name
https://xxx.vault.azure.net/
Resource ID
/subscriptions/xxxxxx/...

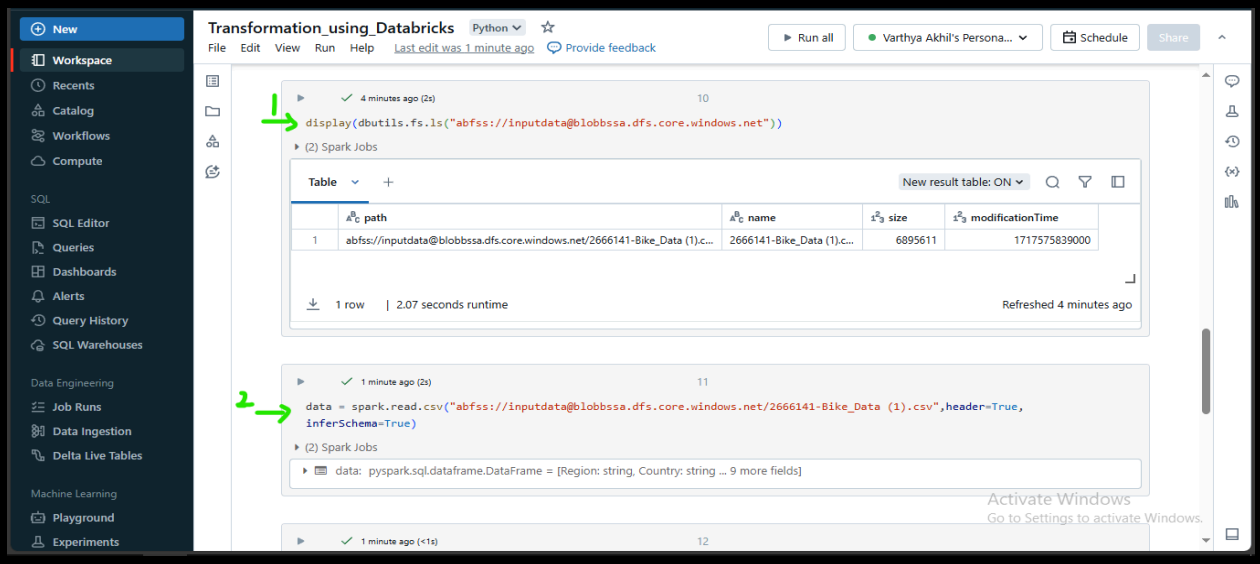

Home > projects-key-vaults | Overview > projects-key-vaults
projects-key-vaults | Properties
Key vault
Search
Save
Discard changes
Refresh
Secrets
Certificates
Settings
Access configuration
Networking
Microsoft Defender for Cloud
Properties
Locks
Monitoring
Alerts
Metrics
Diagnostic settings
Logs
Insights
Workbooks
Name
projects-key-vaults
Sku (Pricing tier)
Standard
Location
eastus
Vault URI
https://projects-key-vaults.vault.azure.net/
Resource ID
/subscriptions/c92e5287-784e-4a20-9e6c-8549947f2ee6/resourceGroups/projects_rg/providers/Microsoft.KeyVault/vaults/project...
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)
copy
2
1

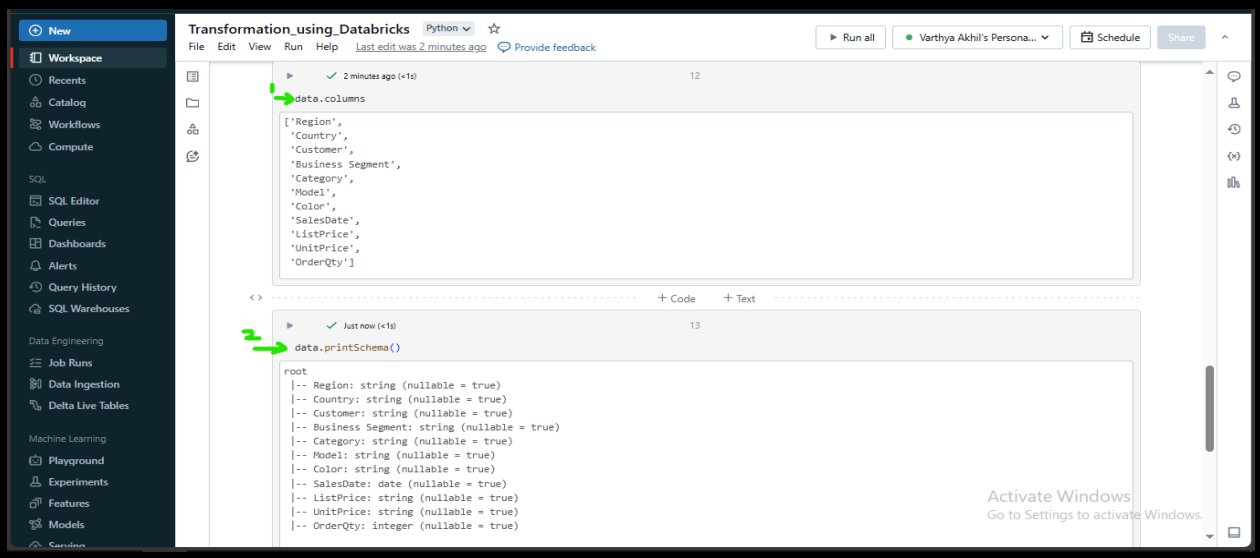

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
data-bricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
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
1
Manage Principal
Creator
Azure Key Vault
DNS Name
2
Paste
https://projects-key-vaults.vault.azure.net/
Resource ID
/subscriptions/xxxxxx/...

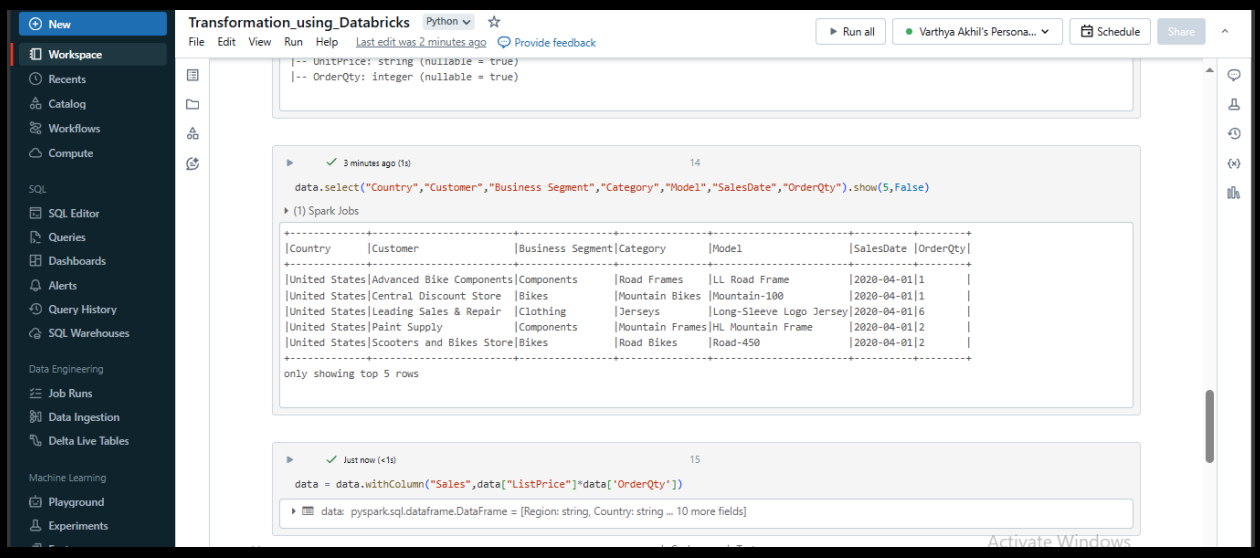

Home > projects-key-vaults | Overview > projects-key-vaults
projects-key-vaults | Properties
Key vault
Search
Save
Discard changes
Refresh
Secrets
Certificates
Settings
Access configuration
Networking
Microsoft Defender for Cloud
Properties
Locks
Monitoring
Alerts
Metrics
Diagnostic settings
Logs
Insights
Workbooks
Name
projects-key-vaults
Sku (Pricing tier)
Standard
Location
eastus
Vault URI
https://projects-key-vaults.vault.azure.net/
Resource ID
/subscriptions/c92e5287-784e-4a20-9e6c-8549947f2ee6/resourceGroups/projects_rg/providers/Microsoft.KeyVault/vaults/project...
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)

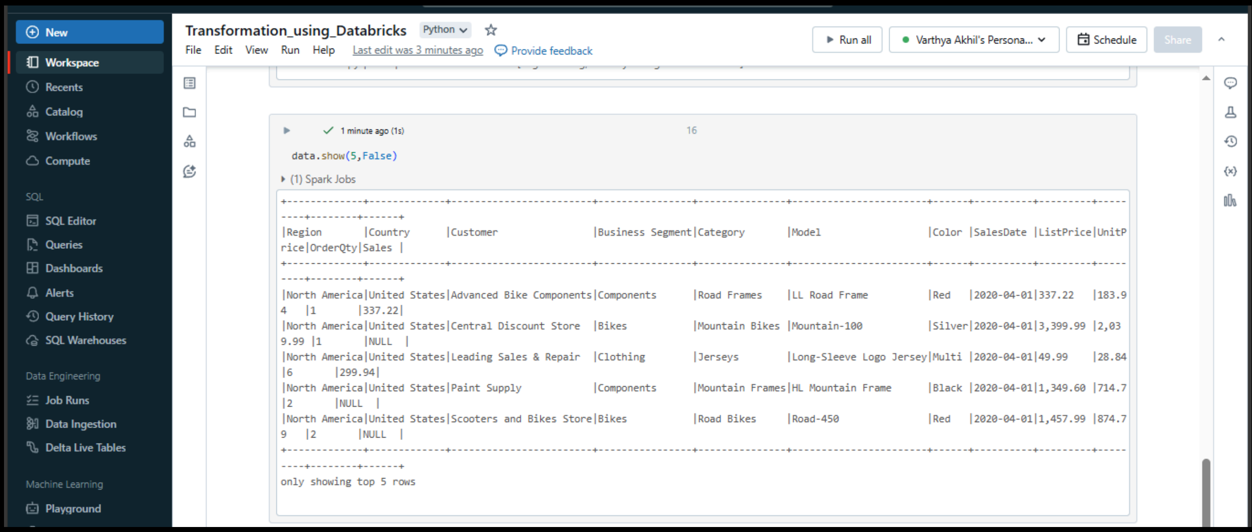

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
data-bricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
HomePage / Create Secret Scope
Create Secret Scope
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
https://projects-key-vaults.vault.azure.net/
Resource ID
/subscriptions/c92e5287-784e-4a20-9e6c-8549947f2ee6/resourceGroups/projects_r
paste

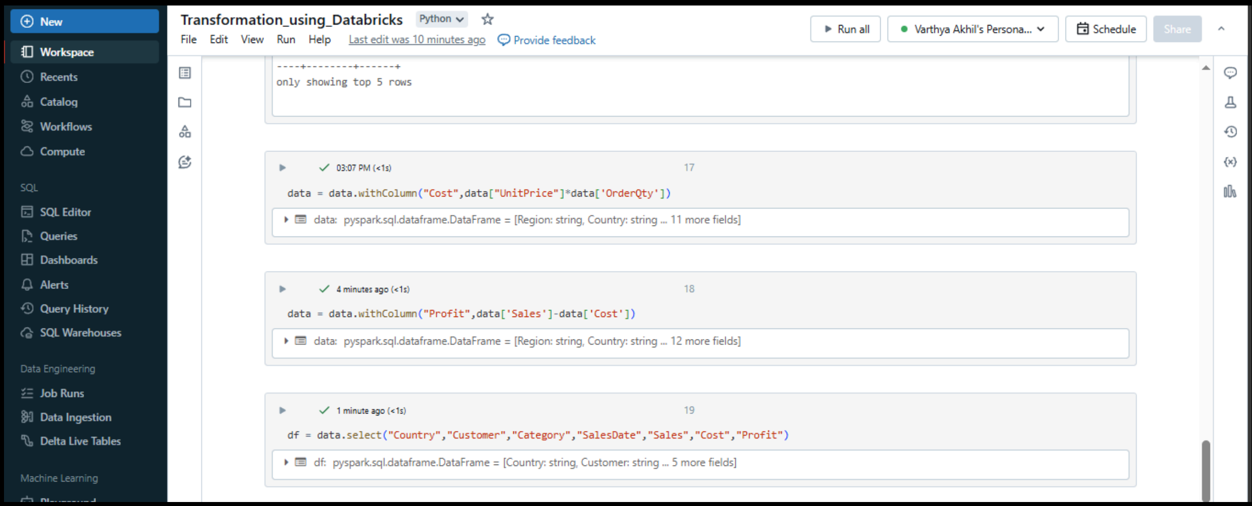

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
data-bricks
Transformation_using_Databricks
Python
File Edit View Run Help
Last edit was 2 minutes ago
Provide feedback
Run all
Varthya Akhil's Persona...
Schedule
Share
Workspace
Machine Learning
Playground
Experiments
Features
Models
from pyspark.sql import *
from pyspark.sql.functions import *
dbutils.secrets.listScopes()
[SecretScope(name='databricksscope')]
spark.conf.set(
"fs.azure.account.key.<storage-account>.dfs.core.windows.net",
dbutils.secrets.get(scope="<scope>", key="<storage-account-access-key>"))
# scope = scopename (databricksscope)
# key = storage-account-access-key (blobskey (azure key vault -> secrets))
spark.conf.set("fs.azure.account.key.blobbssa.dfs.core.windows.net",dbutils.secrets.get(scope="databricksscope",key="blobskey"))


Transformation_using_Databricks
Python
File Edit View Run Help
Last edit was 1 minute ago
Provide feedback
Run all
Varthya Akhil's Persona...
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
Experiments
4 minutes ago (2s)
10
display(dbutils.fs.ls("abfss://inputdata@blobbssa.dfs.core.windows.net"))
(2) Spark Jobs
Table
New result table: ON
path
name
size
modificationTime
1
abfss://inputdata@blobbssa.dfs.core.windows.net/2666141-Bike_Data (1).c...
2666141-Bike_Data (1).c...
6895611
1717575839000
1 row | 2.07 seconds runtime
Refreshed 4 minutes ago
1 minute ago (2s)
11
data = spark.read.csv("abfss://inputdata@blobbssa.dfs.core.windows.net/2666141-Bike_Data (1).csv",header=True, inferSchema=True)
(2) Spark Jobs
data: pyspark.sql.dataframe.DataFrame = [Region: string, Country: string ... 9 more fields]
1 minute ago (<1s)
12
Activate Windows
Go to Settings to activate Windows.


Transformation_using_Databricks
Python
File Edit View Run Help
Last edit was 2 minutes ago
Provide feedback
Run all
Varthya Akhil's Persona...
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
Experiments
Features
Models
2 minutes ago (<1s)
12
data.columns
['Region',
'Country',
'Customer',
'Business Segment',
'Category',
'Model',
'Color',
'SalesDate',
'ListPrice',
'UnitPrice',
'OrderQty']
+ Code
+ Text
Just now (<1s)
13
data.printSchema()
root
|-- Region: string (nullable = true)
|-- Country: string (nullable = true)
|-- Customer: string (nullable = true)
|-- Business Segment: string (nullable = true)
|-- Category: string (nullable = true)
|-- Model: string (nullable = true)
|-- Color: string (nullable = true)
|-- SalesDate: date (nullable = true)
|-- ListPrice: string (nullable = true)
|-- UnitPrice: string (nullable = true)
|-- OrderQty: integer (nullable = true)
Activate Windows
Go to Settings to activate Windows.


Transformation_using_Databricks
Python
File Edit View Run Help
Last edit was 2 minutes ago
Provide feedback
Run all
Varthya Akhil's Persona...
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
Experiments
|-- UnitPrice: string (nullable = true)
|-- OrderQty: integer (nullable = true)
3 minutes ago (1s)
14
data.select("Country","Customer","Business Segment","Category","Model","SalesDate","OrderQty").show(5,False)
(1) Spark Jobs
Country	Customer	Business Segment	Category	Model	SalesDate	OrderQty
United States	Advanced Bike Components	Components	Road Frames	LL Road Frame	2020-04-01	1
United States	Central Discount Store	Bikes	Mountain Bikes	Mountain-100	2020-04-01	1
United States	Leading Sales & Repair	Clothing	Jerseys	Long-Sleeve Logo Jersey	2020-04-01	6
United States	Paint Supply	Components	Mountain Frames	HL Mountain Frame	2020-04-01	2
United States	Scooters and Bikes Store	Bikes	Road Bikes	Road-450	2020-04-01	2
only showing top 5 rows
Just now (<1s)
15
data = data.withColumn("Sales",data["ListPrice"]*data['OrderQty'])
data: pyspark.sql.dataframe.DataFrame = [Region: string, Country: string ... 10 more fields]
Activate Windows


Transformation_using_Databricks
Python
File Edit View Run Help
Last edit was 3 minutes ago
Provide feedback
Run all
Varthya Akhil's Persona...
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
1 minute ago (1s)
16
data.show(5,False)
(1) Spark Jobs
Region	Country	Customer	Business Segment	Category	Model	Color	SalesDate	ListPrice	UnitPrice	OrderQty	Sales
North America	United States	Advanced Bike Components	Components	Road Frames	LL Road Frame	Red	2020-04-01	337.22	183.94	1	337.22
North America	United States	Central Discount Store	Bikes	Mountain Bikes	Mountain-100	Silver	2020-04-01	3,399.99	2,039.99	1	NULL
North America	United States	Leading Sales & Repair	Clothing	Jerseys	Long-Sleeve Logo Jersey	Multi	2020-04-01	49.99	28.84	6	299.94
North America	United States	Paint Supply	Components	Mountain Frames	HL Mountain Frame	Black	2020-04-01	1,349.60	714.7	2	NULL
North America	United States	Scooters and Bikes Store	Bikes	Road Bikes	Road-450	Red	2020-04-01	1,457.99	874.79	2	NULL
only showing top 5 rows

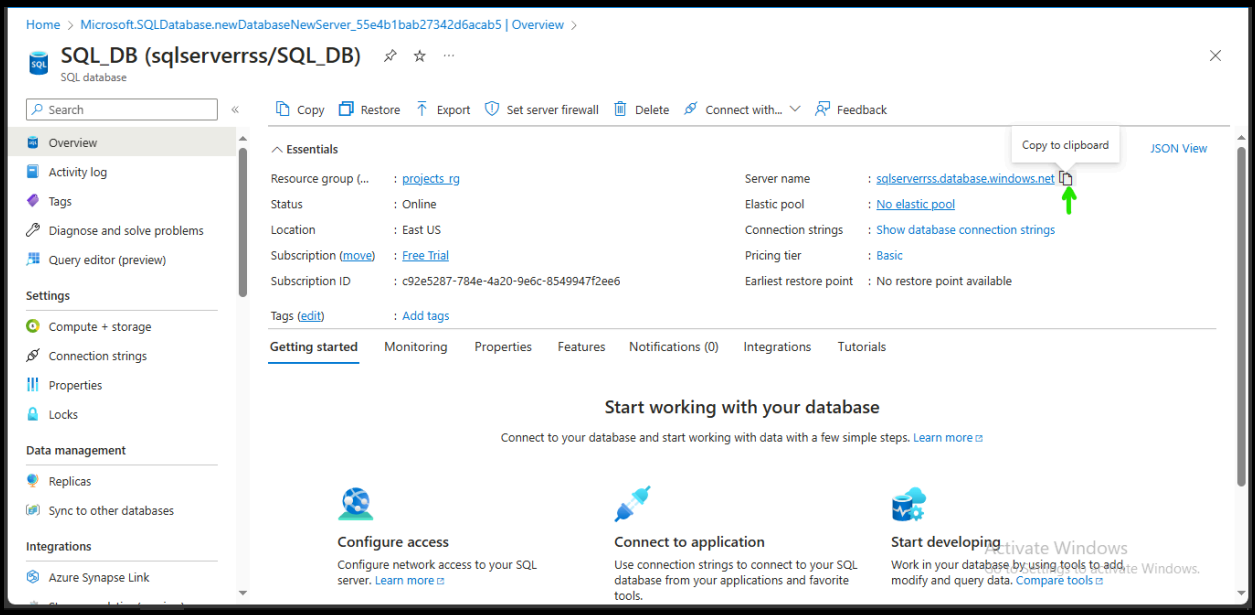

Transformation_using_Databricks
Python
File Edit View Run Help
Last edit was 10 minutes ago
Provide feedback
Run all
Varthya Akhil's Persona...
Schedule
Share
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
only showing top 5 rows
03:07 PM (<1s)
17
data = data.withColumn("Cost",data["UnitPrice"]*data['OrderQty'])
data: pyspark.sql.dataframe.DataFrame = [Region: string, Country: string ... 11 more fields]
4 minutes ago (<1s)
18
data = data.withColumn("Profit",data['Sales']-data['Cost'])
data: pyspark.sql.dataframe.DataFrame = [Region: string, Country: string ... 12 more fields]
1 minute ago (<1s)
19
df = data.select("Country","Customer","Category","SalesDate","Sales","Cost","Profit")
df: pyspark.sql.dataframe.DataFrame = [Country: string, Customer: string ... 5 more fields]

Loading dataframe into sql database.

Boilerplate code to connect sql database.

→ `jdbc_url = "jdbc:sqlserver://<your_server_name>.database.windows.net:1433;database=<your_database_name>"`

Copy the servername from azure sql database overview



Paste the server name in the boilerplate code.

```
21
#loading data into azure sql database
# boiler plate code to connect azure sql database
# jdbc_url = "jdbc:sqlserver://<your_server_name>.database.windows.net:1433;database=<your_database_name>"
```

```
1 minute ago (<1s)
22
server = "sqlserverrss.database.windows.net"
port = 1433
Database = "SQL_DB"
db_properties = {"user":"server_admin","password":"[redacted]"}
```

```
1 minute ago (<1s)
23
# jdbc_url = "jdbc:sqlserver://sqlserverrss.database.windows.net:1433;SQL_DB" or we can use below one as well
jdbc_url = "jdbc:sqlserver://{0}:{1};database={2}".format(server,port,Database)
```

Boilerplate code to writeback to sql server database

→ `df.write.jdbc(url=jdbc_url, table="your_table_name", mode="overwrite", properties=connection_properties)`

Now data has been successfully loaded into an azure sql database.

Lets see data in azure sql database by querying .