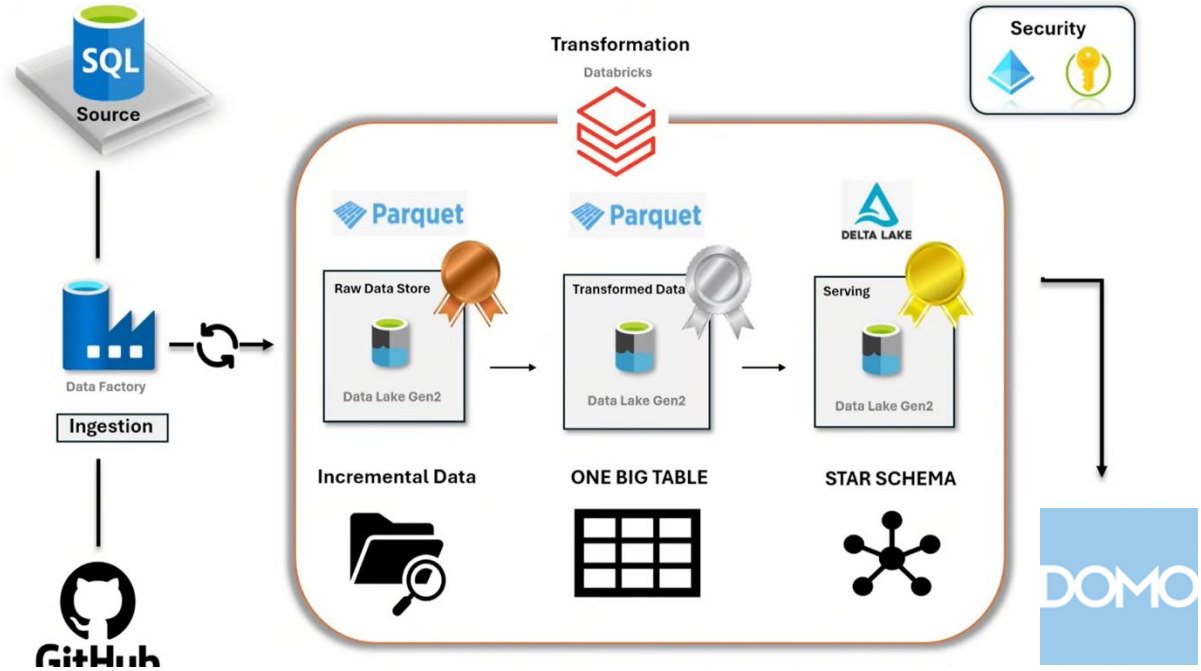
Data Bricks Project Work

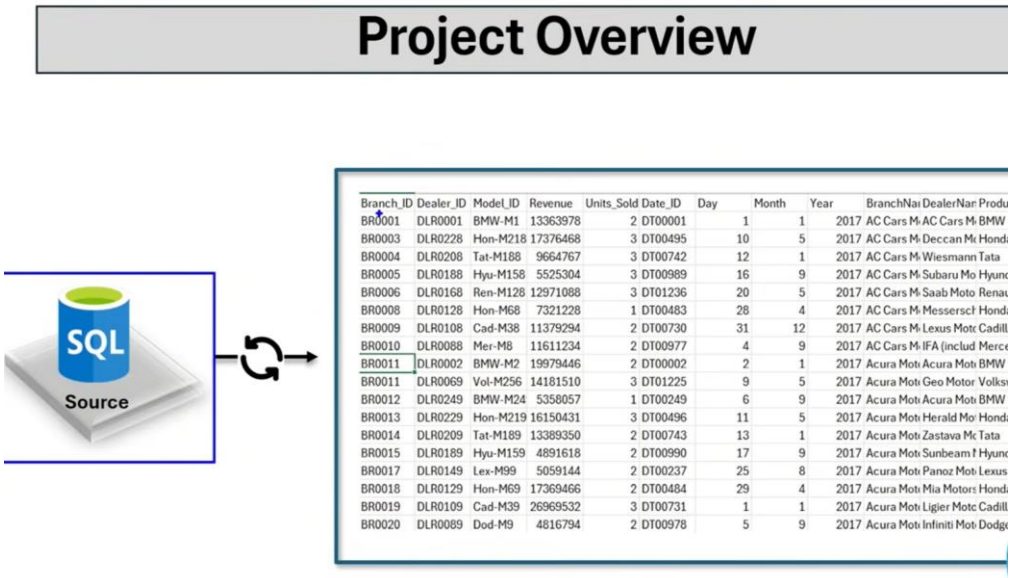
1) Architecture



1) Data Source – Azure SQL Data Source
2) Git hub Repository -Pull the data from Github Repository to Azure SQL Data Source
3) ETL Tools -Azure Data Factory → After Azure SQL data source to connect pipeline to Azure data factory. It is advance Pipeline
4) Bronze layer→ Data Factory to add data in bronze layer with increment load data only. That is why is advance data pipeline create CDC(Change data capture) ,lookup files and store procedure
5) Parquet →it is file format
6) After land bronze layer then use Databricks. The databricks transform the data in silver layer and add data in Delta Lake format or gold layer
7) Unity Catalog → it is unity is the unified Solution govern of the data to apply data quality check and more. **It is hint**

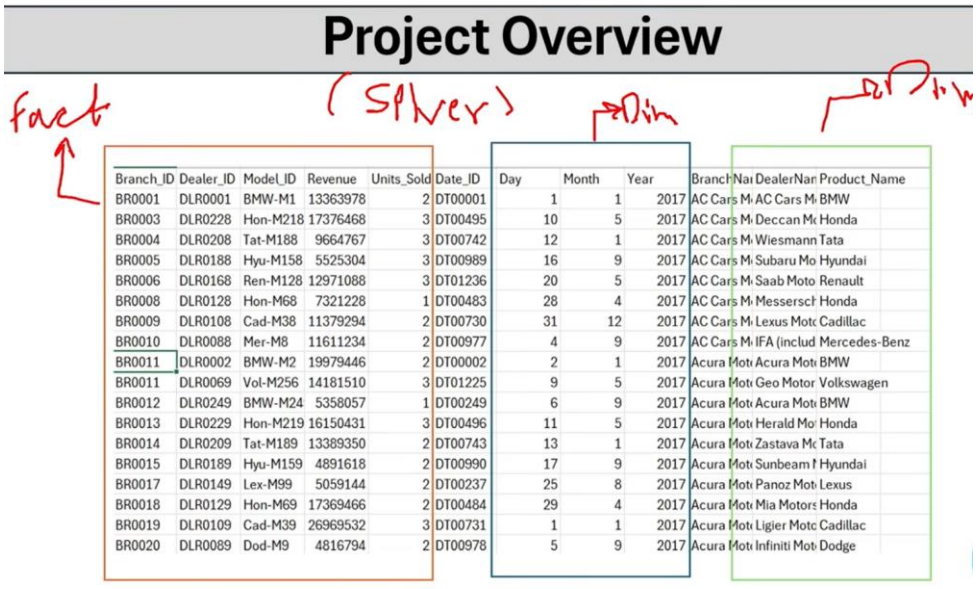
Project Overview:

1) This is fact table data. which is present in Bozen layer



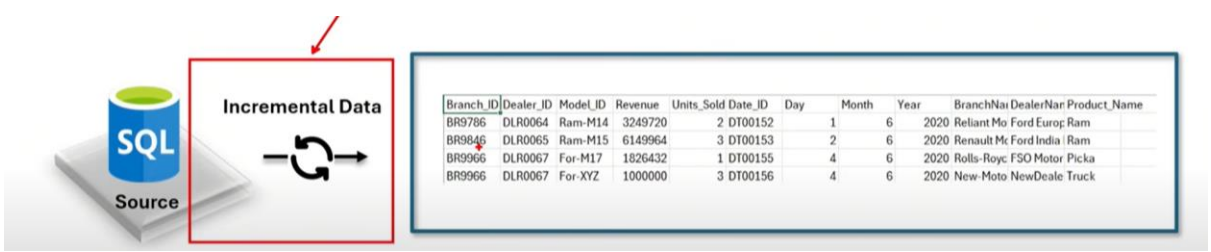
Branch_ID	Dealer_ID	Model_ID	Revenue	Units_Sold	Date_ID	Day	Month	Year	BranchNai	DealerNar	Produ
BR0001	DLR0001	BMW-M1	13363978	2	DT00001	1	1	2017	AC Cars M	AC Cars M	BMW
BR0003	DLR0228	Hon-M218	17376468	3	DT00495	10	5	2017	AC Cars M	Deccan M	Hond
BR0004	DLR0208	Tat-M188	9664767	3	DT00742	12	1	2017	AC Cars M	Wiesmann	Tata
BR0005	DLR0188	Hyu-M158	5525304	3	DT00989	16	9	2017	AC Cars M	Subaru Mo	Hyun
BR0006	DLR0168	Ren-M128	12971088	3	DT01236	20	5	2017	AC Cars M	Saab Moto	Renau
BR0008	DLR0128	Hon-M68	7321228	1	DT00483	28	4	2017	AC Cars M	Messersch	Hond
BR0009	DLR0108	Cad-M38	11379294	2	DT00730	31	12	2017	AC Cars M	Lexus Mot	Cadill
BR0010	DLR0088	Mer-M8	11611234	2	DT00977	4	9	2017	AC Cars M	IFA (includ	Merce
BR0011	DLR0002	BMW-M2	19979446	2	DT00002	2	1	2017	Acura Mot	Acura Mot	BMW
BR0011	DLR0069	Vol-M256	14181510	3	DT01225	9	5	2017	Acura Mot	Geo Motor	Volks
BR0012	DLR0249	BMW-M24	5358057	1	DT00249	6	9	2017	Acura Mot	Acura Mot	BMW
BR0013	DLR0229	Hon-M219	16150431	3	DT00496	11	5	2017	Acura Mot	Herald Mo	Hond
BR0014	DLR0209	Tat-M189	13389350	2	DT00743	13	1	2017	Acura Mot	Zastava Mc	Tata
BR0015	DLR0189	Hyu-M159	4891618	2	DT00990	17	9	2017	Acura Mot	Sunbeam M	Hyun
BR0017	DLR0149	Lex-M99	5059144	2	DT00237	25	8	2017	Acura Mot	Panoz Mot	Lexus
BR0018	DLR0129	Hon-M69	17369466	2	DT00484	29	4	2017	Acura Mot	Mia Motors	Hond
BR0019	DLR0109	Cad-M39	26969532	3	DT00731	1	1	2017	Acura Mot	Ligier Mot	Cadill
BR0020	DLR0089	Dod-M9	4816794	2	DT00978	5	9	2017	Acura Mot	Infiniti Mot	Dodge

2) Silver layer –fact table into fact and dimension table and final connect each dimension table in one fact table



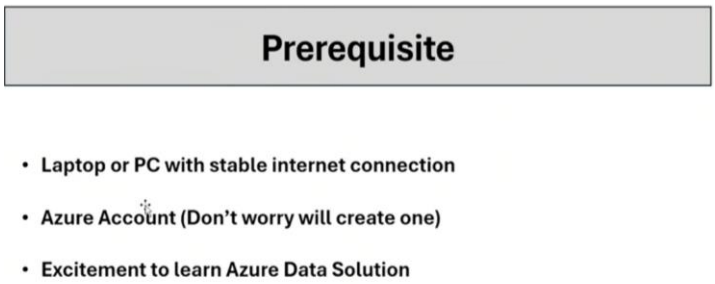
Branch_ID	Dealer_ID	Model_ID	Revenue	Units_Sold	Date_ID	Day	Month	Year	BranchNai	DealerNar	Product_Name
BR0001	DLR0001	BMW-M1	13363978	2	DT00001	1	1	2017	AC Cars M	AC Cars M	BMW
BR0003	DLR0228	Hon-M218	17376468	3	DT00495	10	5	2017	AC Cars M	Deccan M	Honda
BR0004	DLR0208	Tat-M188	9664767	3	DT00742	12	1	2017	AC Cars M	Wiesmann	Tata
BR0005	DLR0188	Hyu-M158	5525304	3	DT00989	16	9	2017	AC Cars M	Subaru Mo	Hyundai
BR0006	DLR0168	Ren-M128	12971088	3	DT01236	20	5	2017	AC Cars M	Saab Moto	Renault
BR0008	DLR0128	Hon-M68	7321228	1	DT00483	28	4	2017	AC Cars M	Messersch	Honda
BR0009	DLR0108	Cad-M38	11379294	2	DT00730	31	12	2017	AC Cars M	Lexus Mot	Cadillac
BR0010	DLR0088	Mer-M8	11611234	2	DT00977	4	9	2017	AC Cars M	IFA (includ	Mercedes-Benz
BR0011	DLR0002	BMW-M2	19979446	2	DT00002	2	1	2017	Acura Mot	Acura Mot	BMW
BR0011	DLR0069	Vol-M256	14181510	3	DT01225	9	5	2017	Acura Mot	Geo Motor	Volkswagen
BR0012	DLR0249	BMW-M24	5358057	1	DT00249	6	9	2017	Acura Mot	Acura Mot	BMW
BR0013	DLR0229	Hon-M219	16150431	3	DT00496	11	5	2017	Acura Mot	Herald Mo	Honda
BR0014	DLR0209	Tat-M189	13389350	2	DT00743	13	1	2017	Acura Mot	Zastava Mc	Tata
BR0015	DLR0189	Hyu-M159	4891618	2	DT00990	17	9	2017	Acura Mot	Sunbeam M	Hyundai
BR0017	DLR0149	Lex-M99	5059144	2	DT00237	25	8	2017	Acura Mot	Panoz Mot	Lexus
BR0018	DLR0129	Hon-M69	17369466	2	DT00484	29	4	2017	Acura Mot	Mia Motors	Honda
BR0019	DLR0109	Cad-M39	26969532	3	DT00731	1	1	2017	Acura Mot	Ligier Mot	Cadillac
BR0020	DLR0089	Dod-M9	4816794	2	DT00978	5	9	2017	Acura Mot	Infiniti Mot	Dodge

3) Data will load in Incremental load. Because past data cannot load again and again

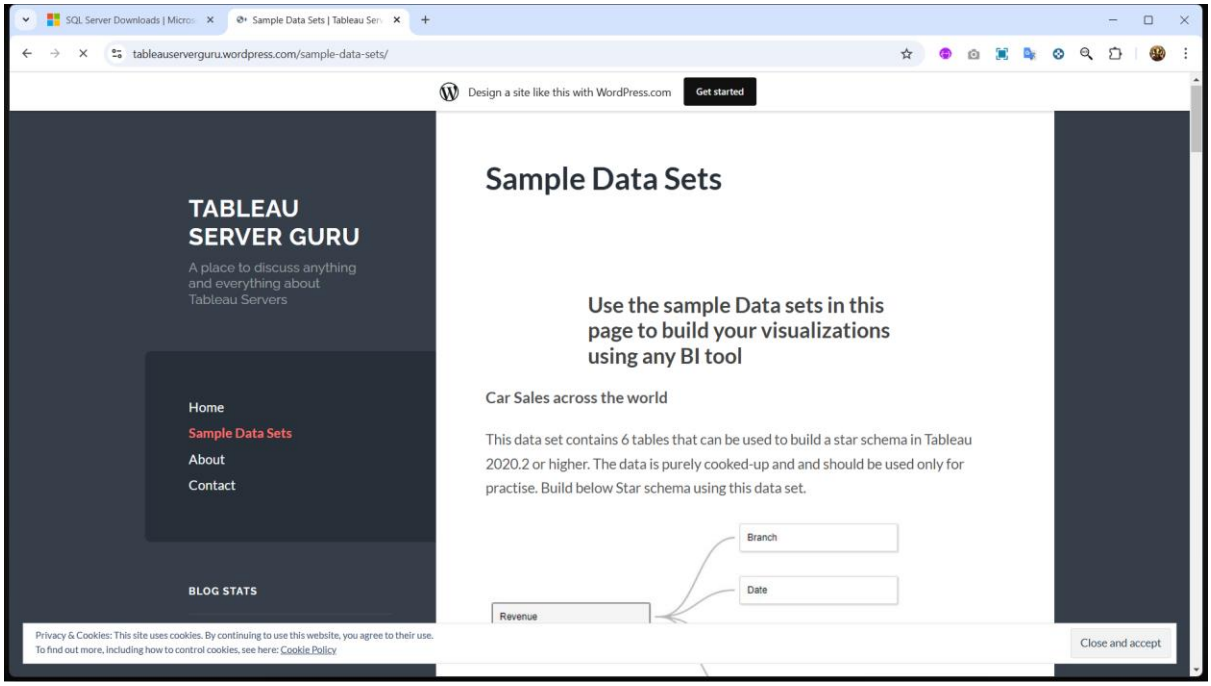


Branch_ID	Dealer_ID	Model_ID	Revenue	Units_Sold	Date_ID	Day	Month	Year	BranchNai	DealerNar	Product_Name
BR9786	DLR0064	Ram-M14	3249720	2	DT00152	1	6	2020	Reliant Mo	Ford Europ	Ram
BR9846	DLR0065	Ram-M15	6149964	3	DT00153	2	6	2020	Renault Mc	Ford India	Ram
BR9966	DLR0067	For-M17	1826432	1	DT00155	4	6	2020	Rolls-Royc	FSO Motor	Picka
BR9966	DLR0067	For-XYZ	1000000	3	DT00156	4	6	2020	New Moto	NewDeale	Truck

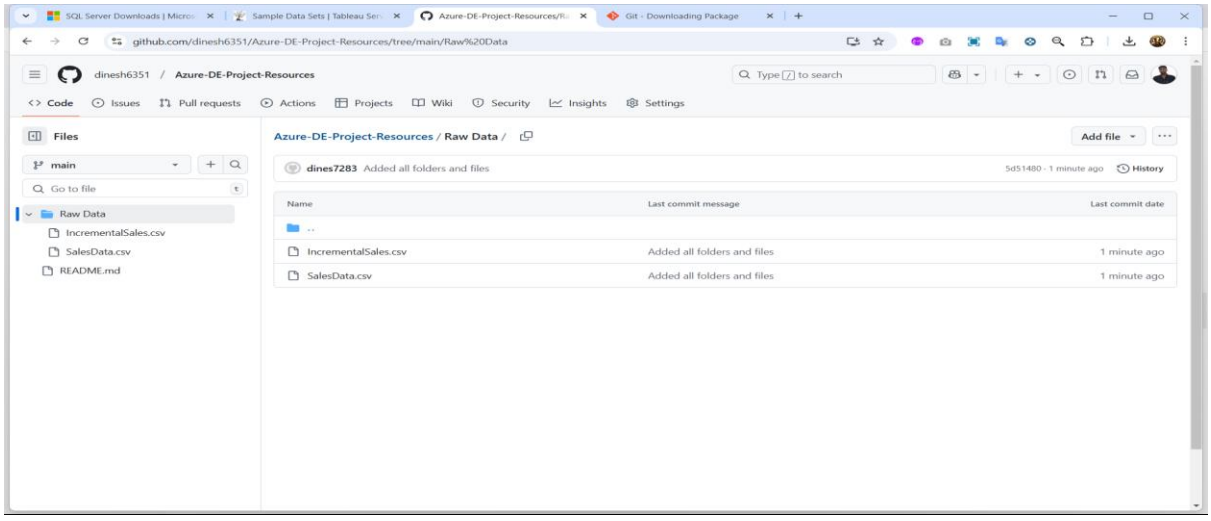
Prerequisite:



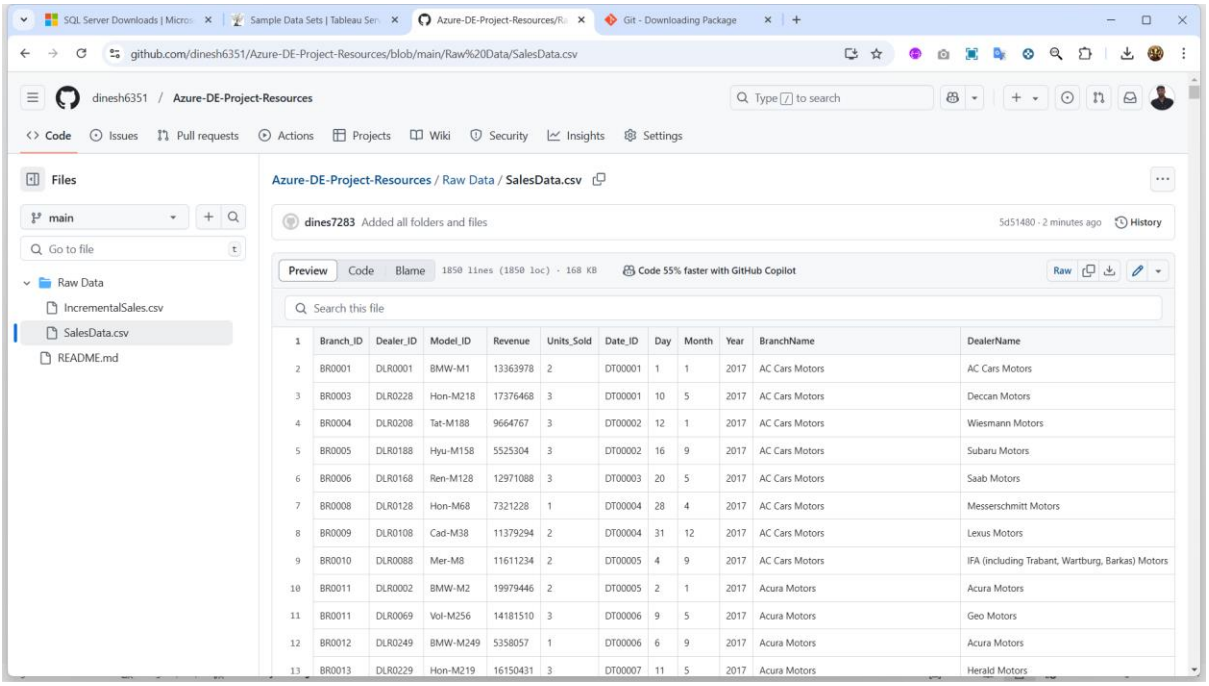
Sample Dataset Website-



--Github Dataset



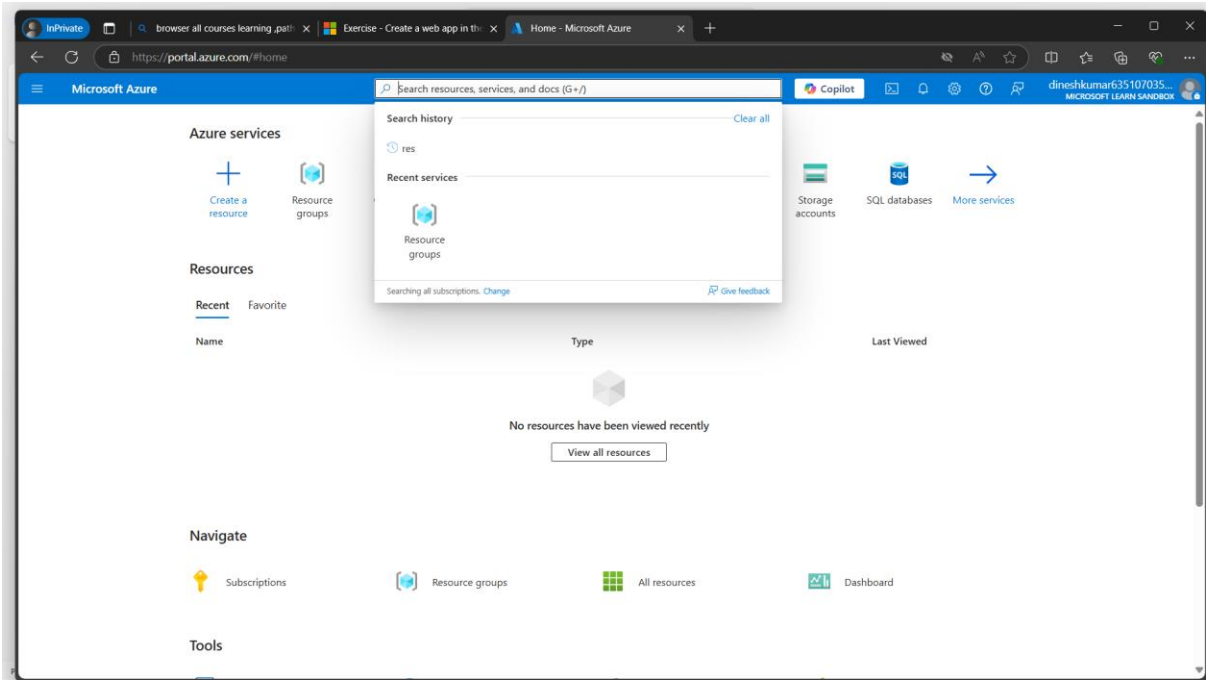
➔ Sales Data have present all Transactions records
➔ Sample



	Branch_ID	Dealer_ID	Model_ID	Revenue	Units_Sold	Date_ID	Day	Month	Year	BranchName	DealerName
2	BR0001	DLR0001	BMW-M1	13363978	2	DT00001	1	1	2017	AC Cars Motors	AC Cars Motors
3	BR0003	DLR0228	Hon-M218	17376468	3	DT00001	10	5	2017	AC Cars Motors	Deccan Motors
4	BR0004	DLR0208	Tat-M188	9664767	3	DT00002	12	1	2017	AC Cars Motors	Wiesmann Motors
5	BR0005	DLR0188	Hyu-M158	5525304	3	DT00002	16	9	2017	AC Cars Motors	Subaru Motors
6	BR0006	DLR0168	Ren-M128	12971088	3	DT00003	20	5	2017	AC Cars Motors	Saab Motors
7	BR0008	DLR0128	Hon-M68	7321228	1	DT00004	28	4	2017	AC Cars Motors	Messerschmitt Motors
8	BR0009	DLR0108	Cad-M38	11379294	2	DT00004	31	12	2017	AC Cars Motors	Lexus Motors
9	BR0010	DLR0088	Mer-M8	11611234	2	DT00005	4	9	2017	AC Cars Motors	IFA (including Trabant, Wartburg, Barkas) Motors
10	BR0011	DLR0002	BMW-M2	19979446	2	DT00005	2	1	2017	Acura Motors	Acura Motors
11	BR0011	DLR0069	Vol-M256	14181510	3	DT00006	9	5	2017	Acura Motors	Geo Motors
12	BR0012	DLR0249	BMW-M249	5358057	1	DT00006	6	9	2017	Acura Motors	Acura Motors
13	BR0013	DLR0229	Hon-M219	16150431	3	DT00007	11	5	2017	Acura Motors	Herald Motors

Azure part

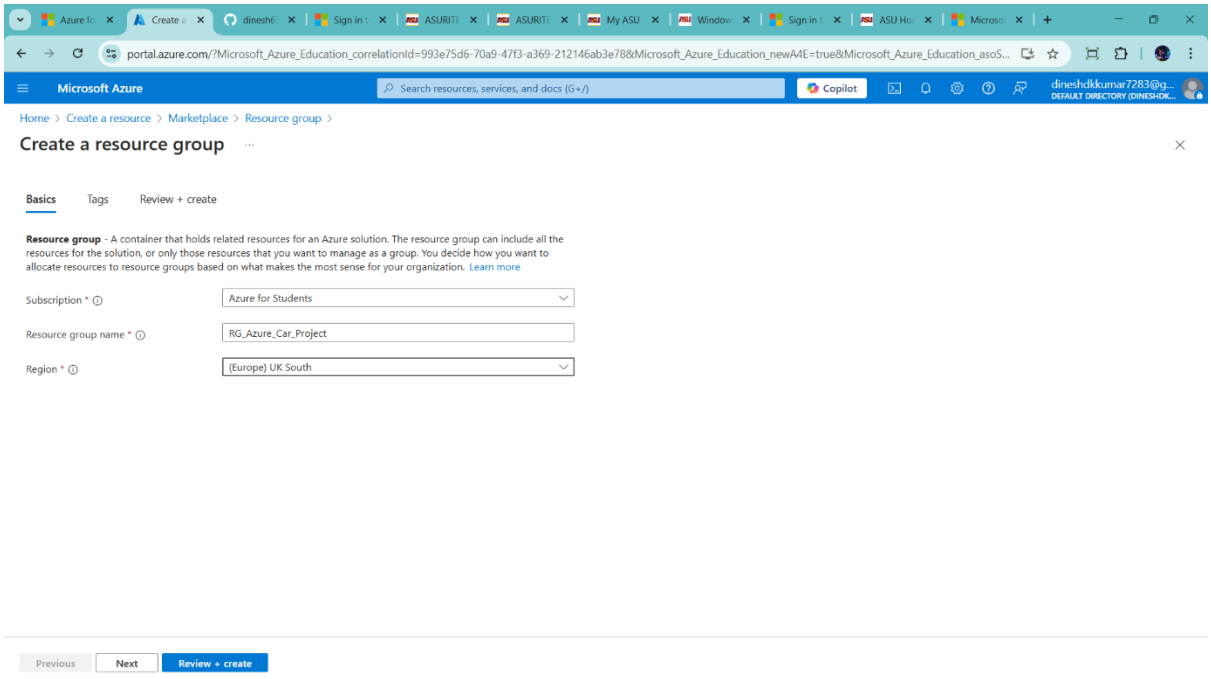
--Resource group – store all the information like data lake, catelog , unitity and more

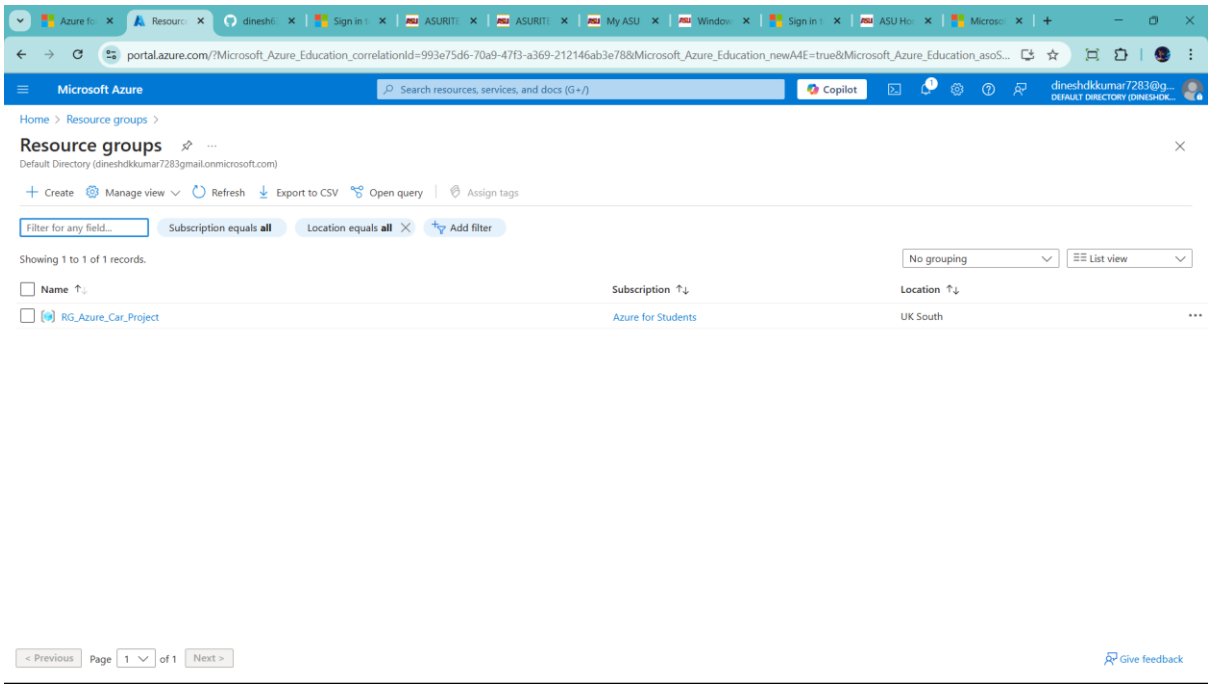


Create resource group:

Home➔ search resource group

By create resource group

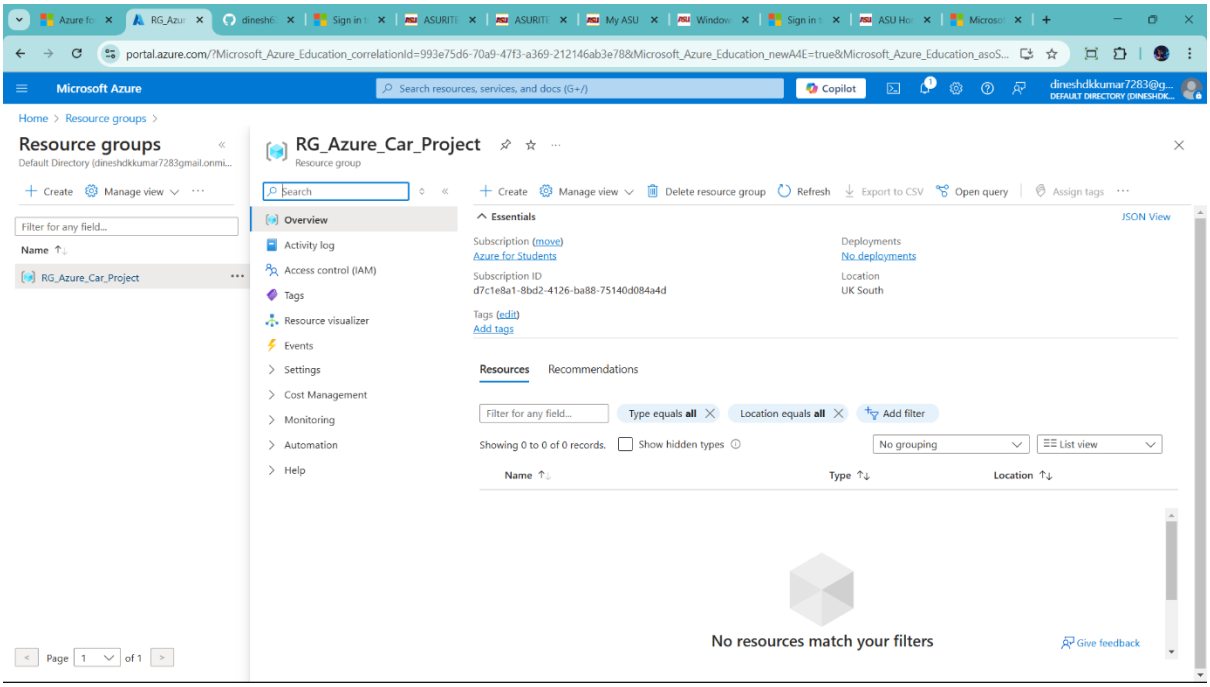


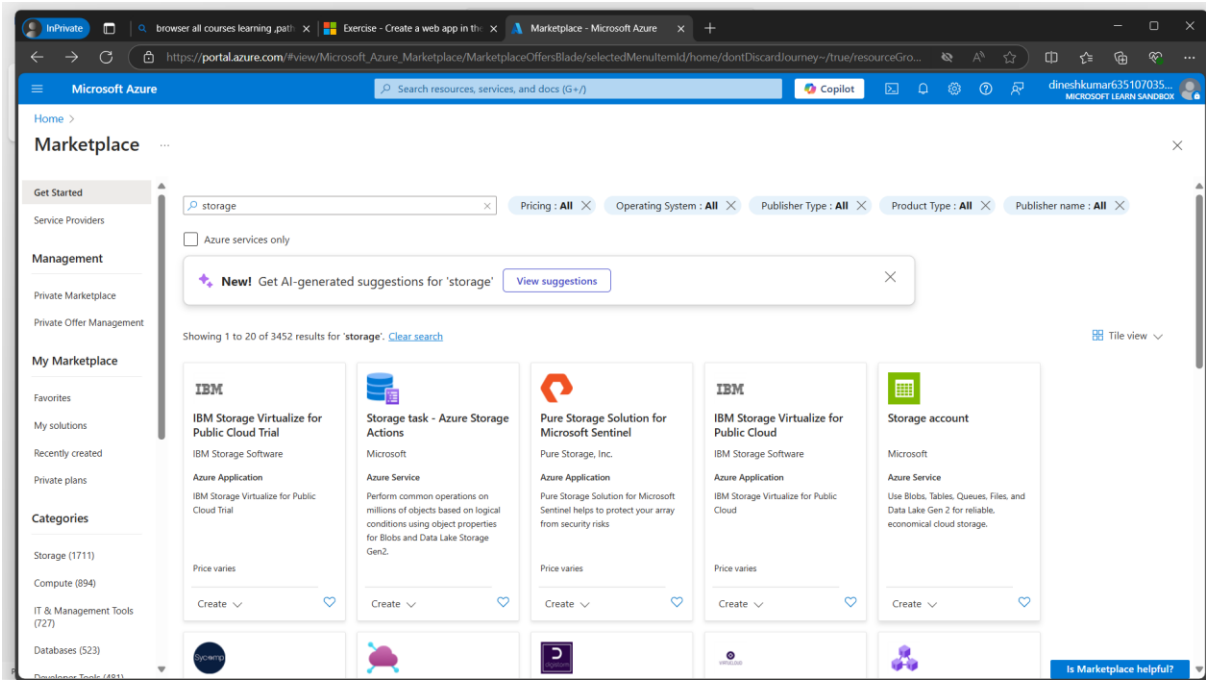


1) Create Data Lake
2) Not source → because souce going to sql database
3) After create Data Lake → to create two resource data factory and sql Databases

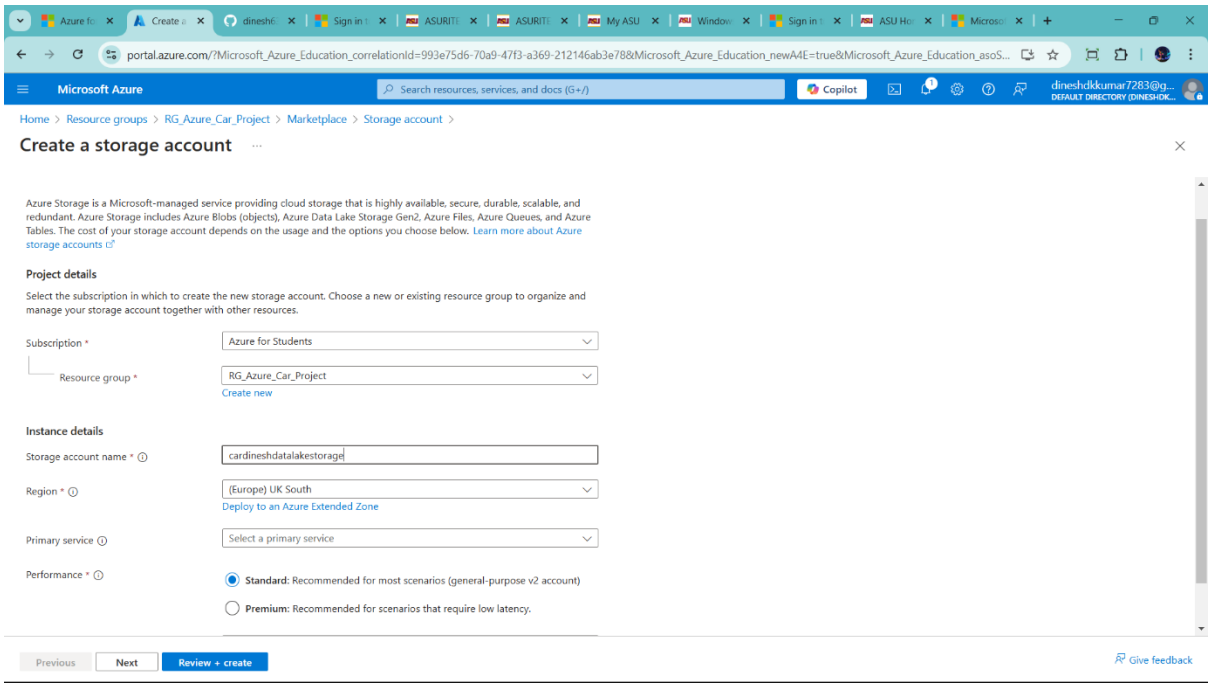
--Let Start

1) click create → search storage account

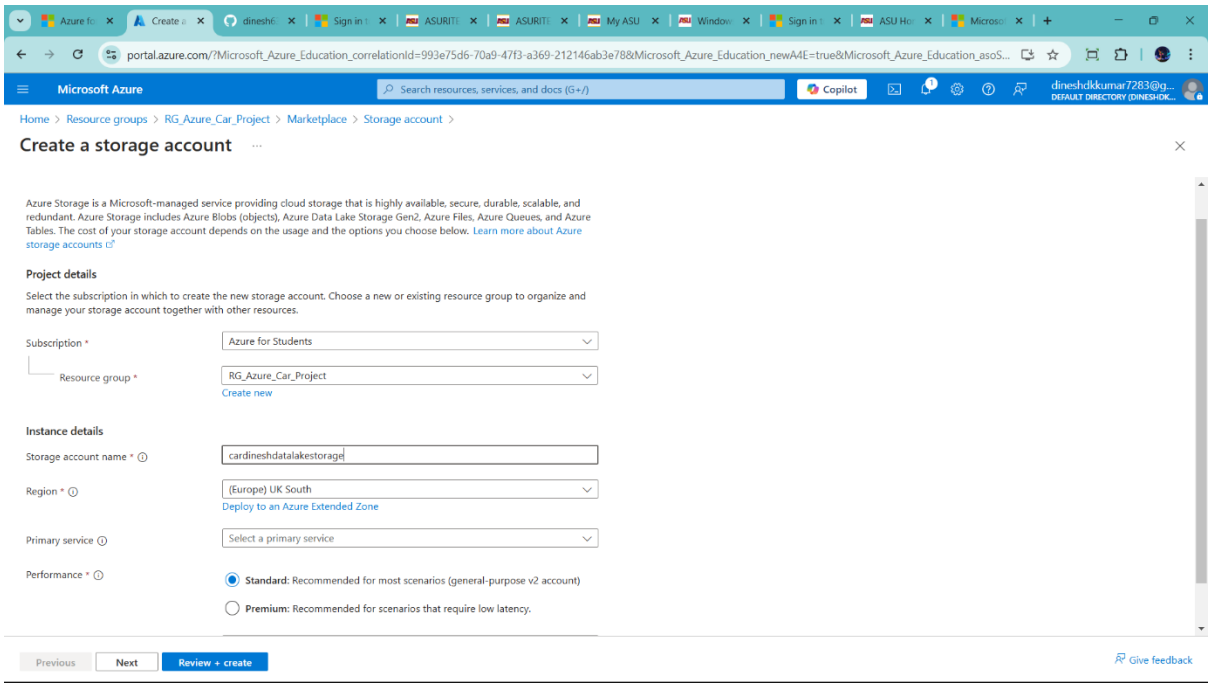




Create storage account and convert into data lake



I use LRS for reduce cost. Because data store in same data centre



Click next

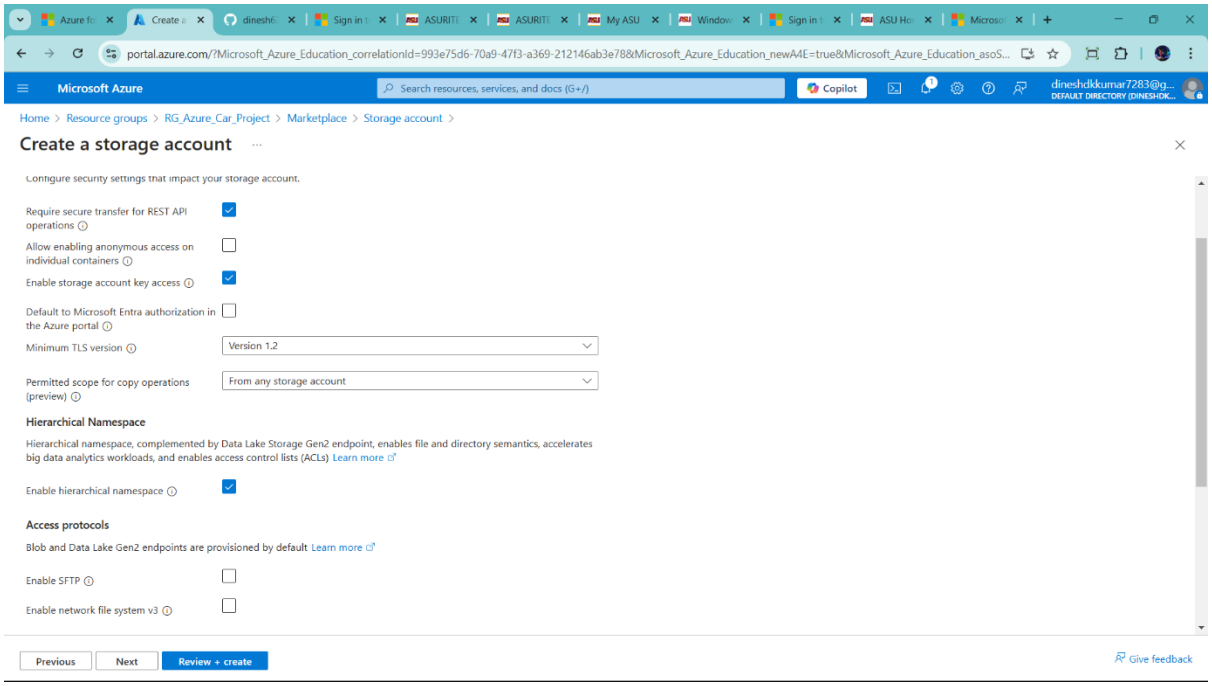
And **!important** click Enable hierarchincal namespace → it means create data lake instance of block storage

--cold check point enable → it is cost very very low it store one month only

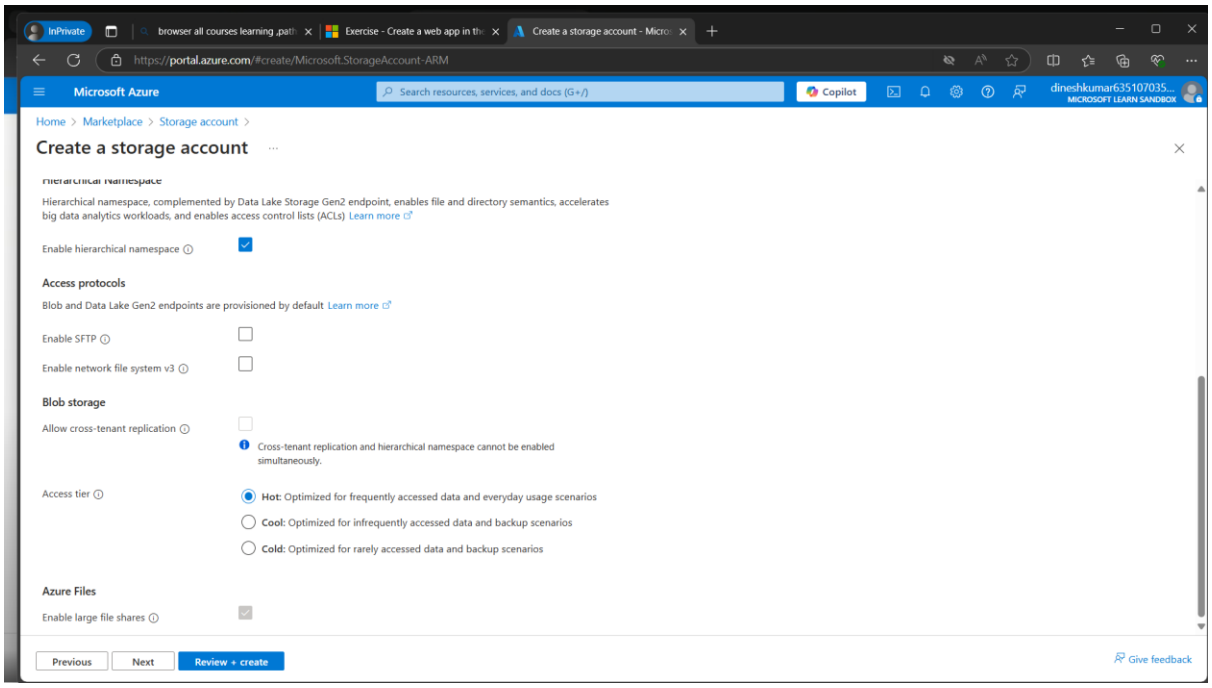
→HOT- acces the data every day

→Cool → once a month

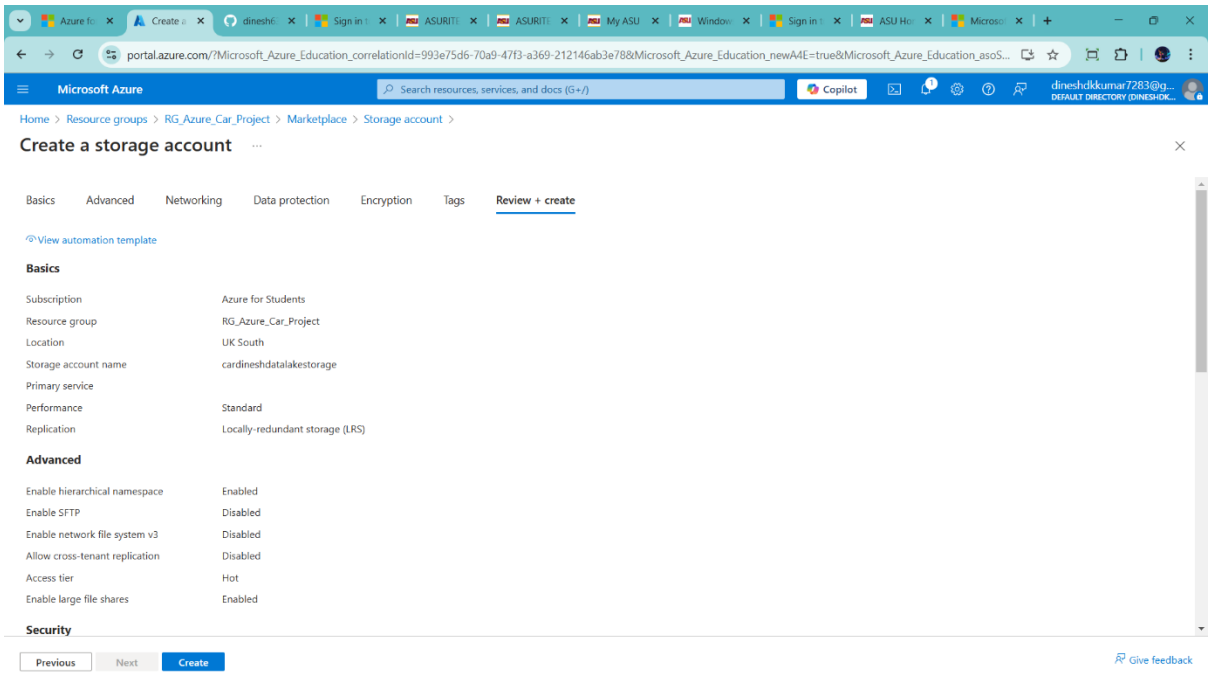
→Cold →6 month once



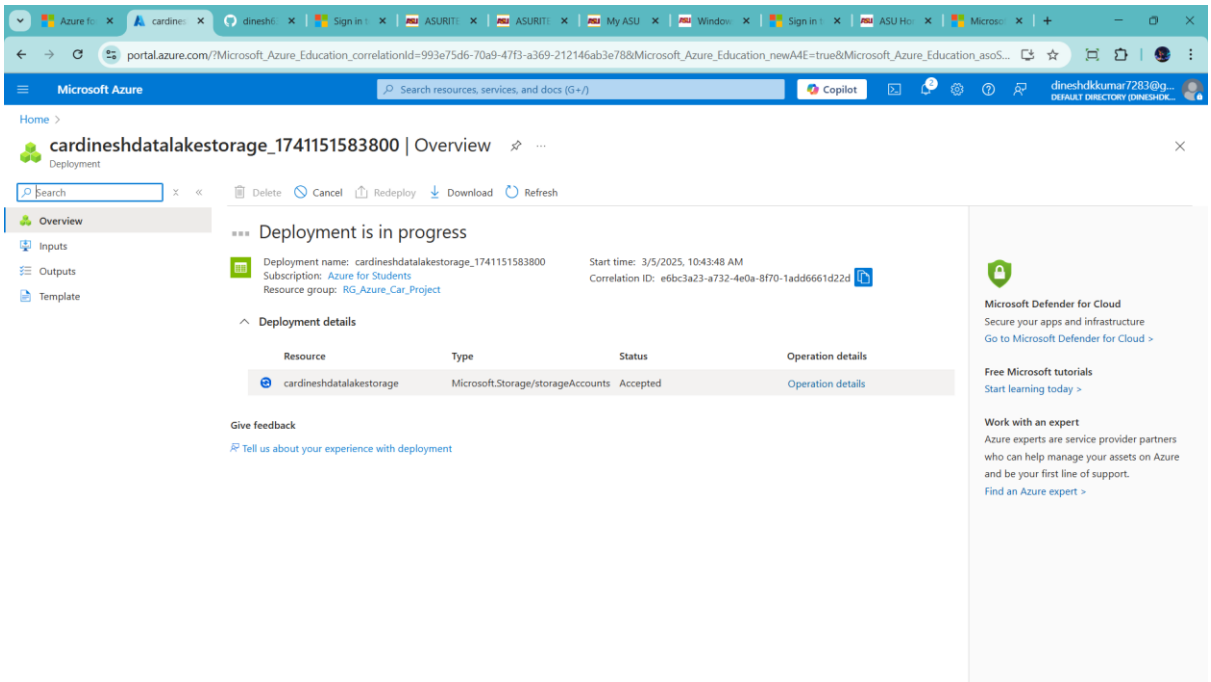
We use HOT – now for learning

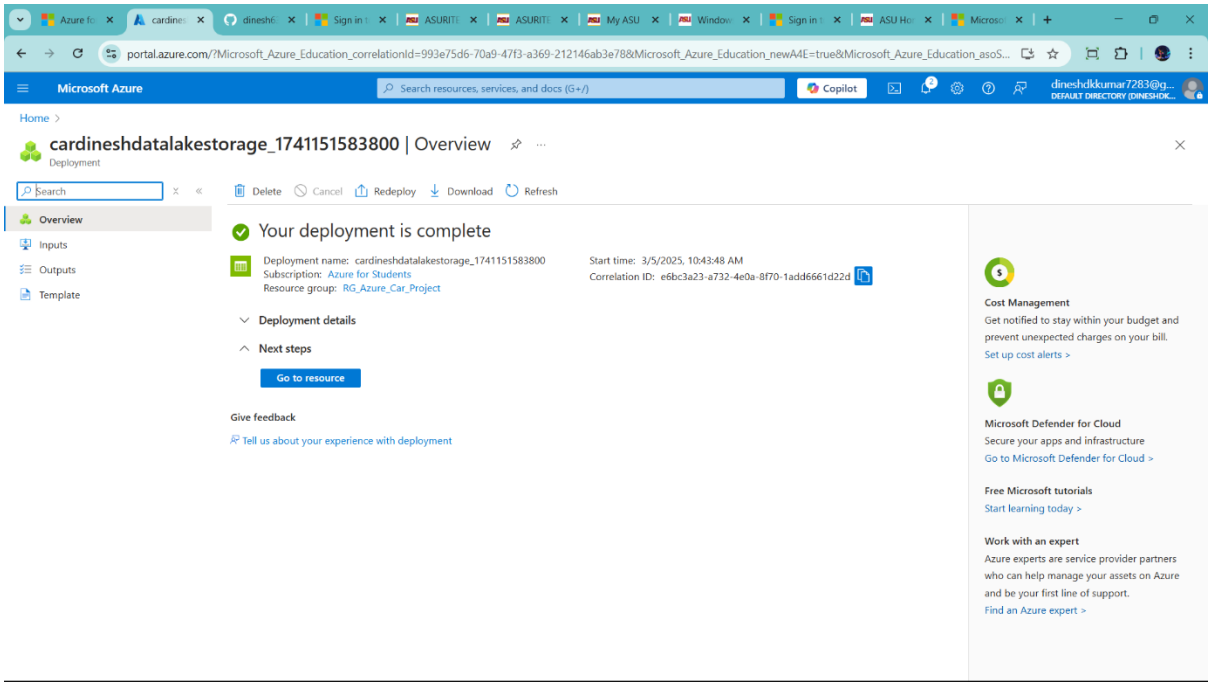


Finally review + create

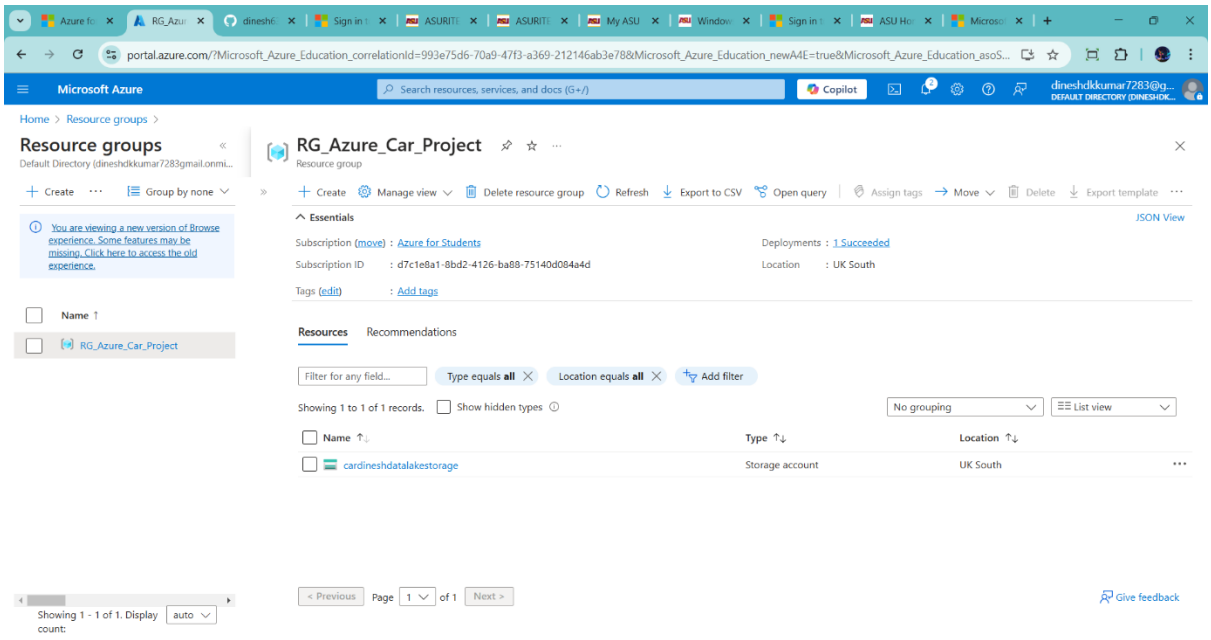


Finally click create –it is data lake

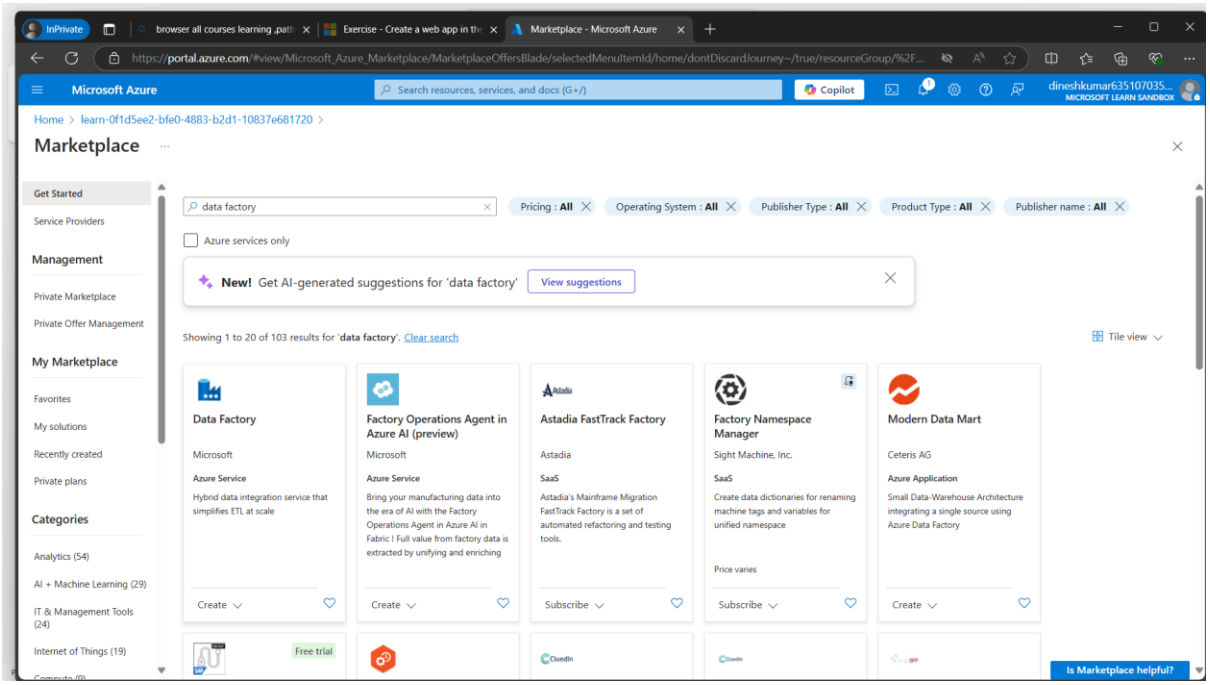


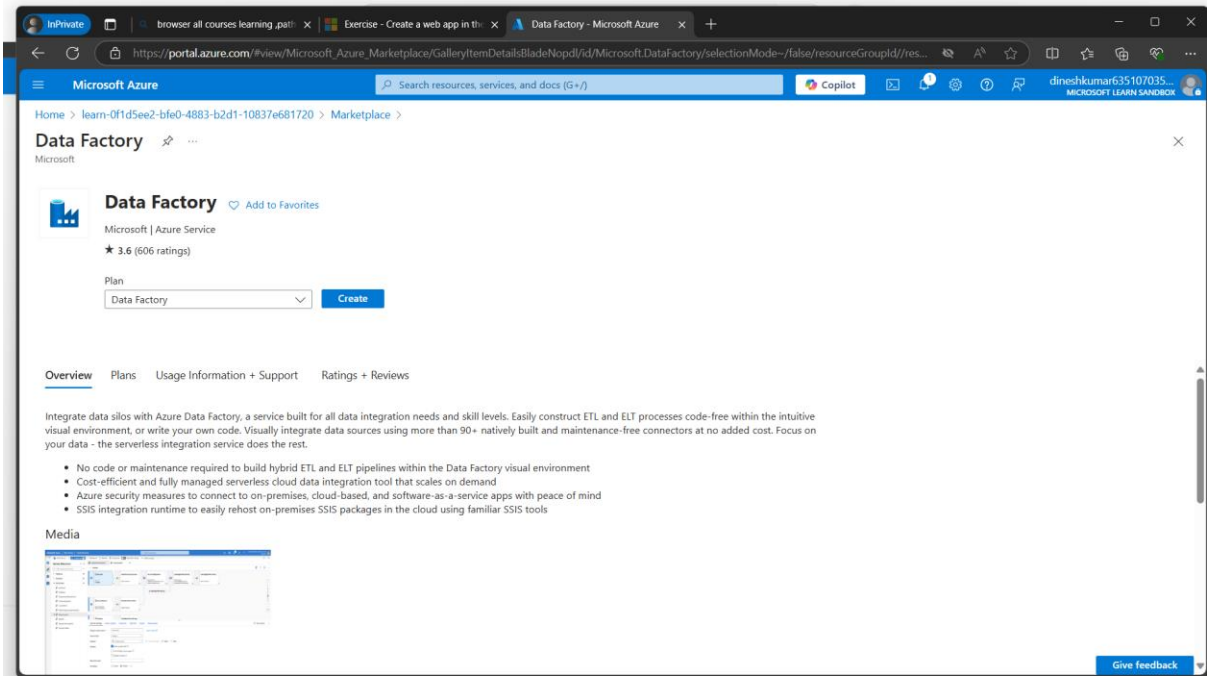


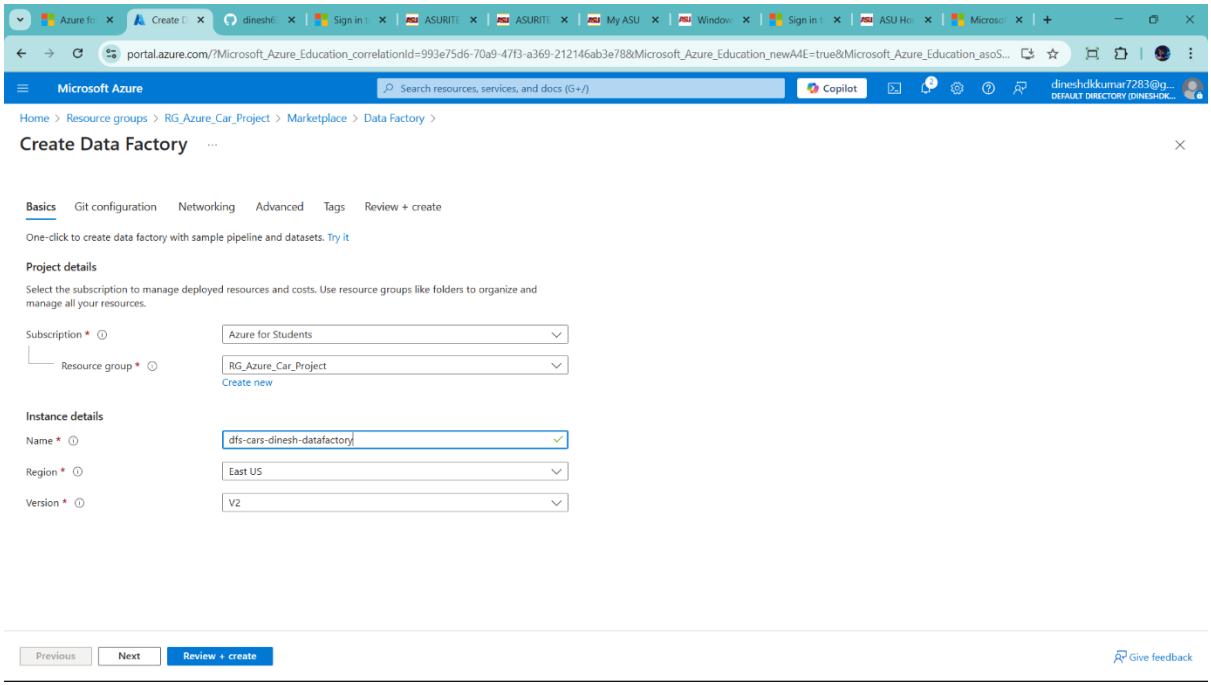
Successfully create data lake



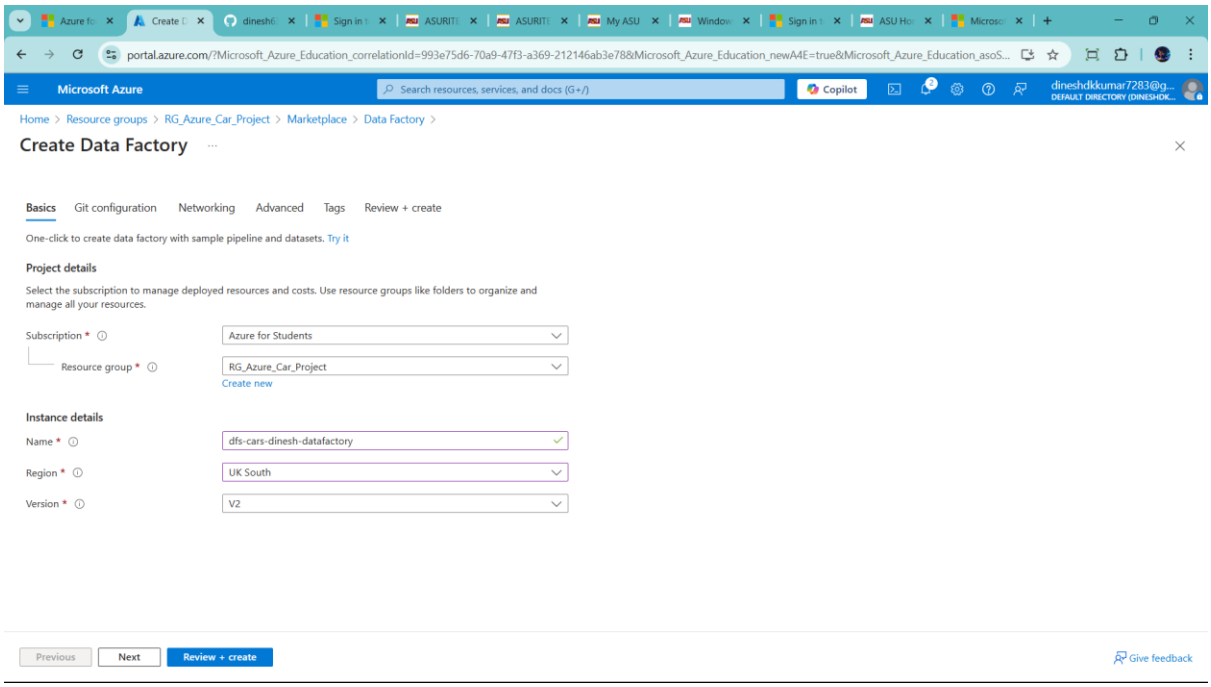
Now create Data factory

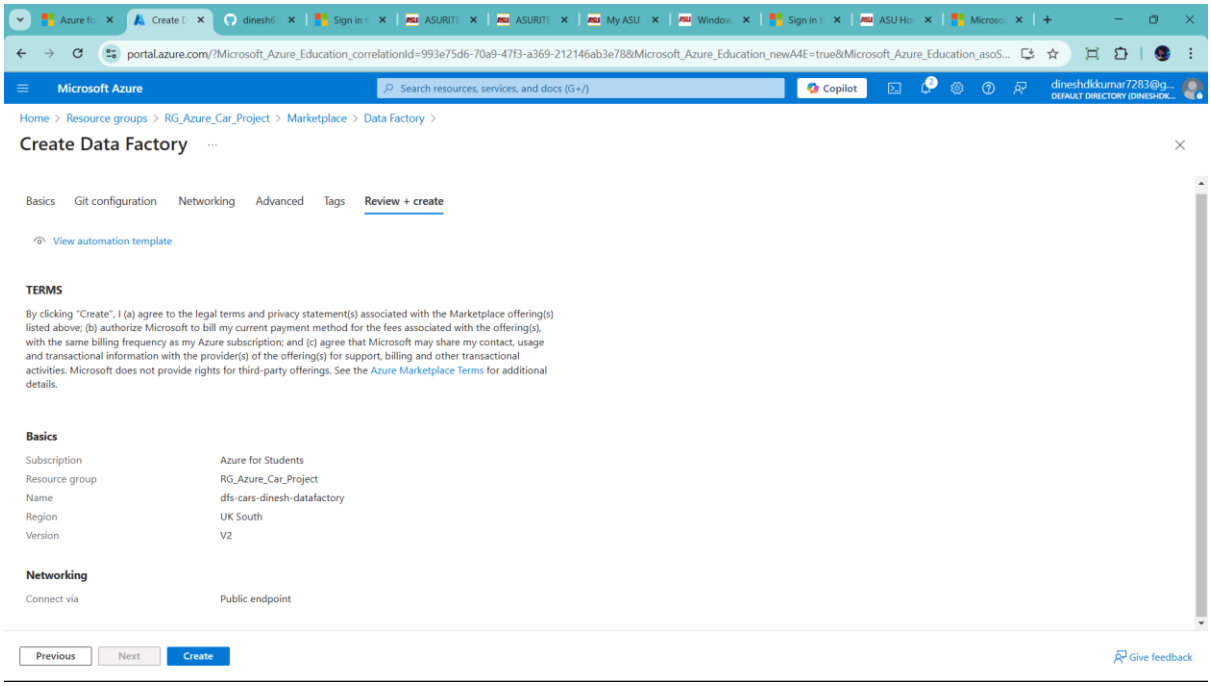




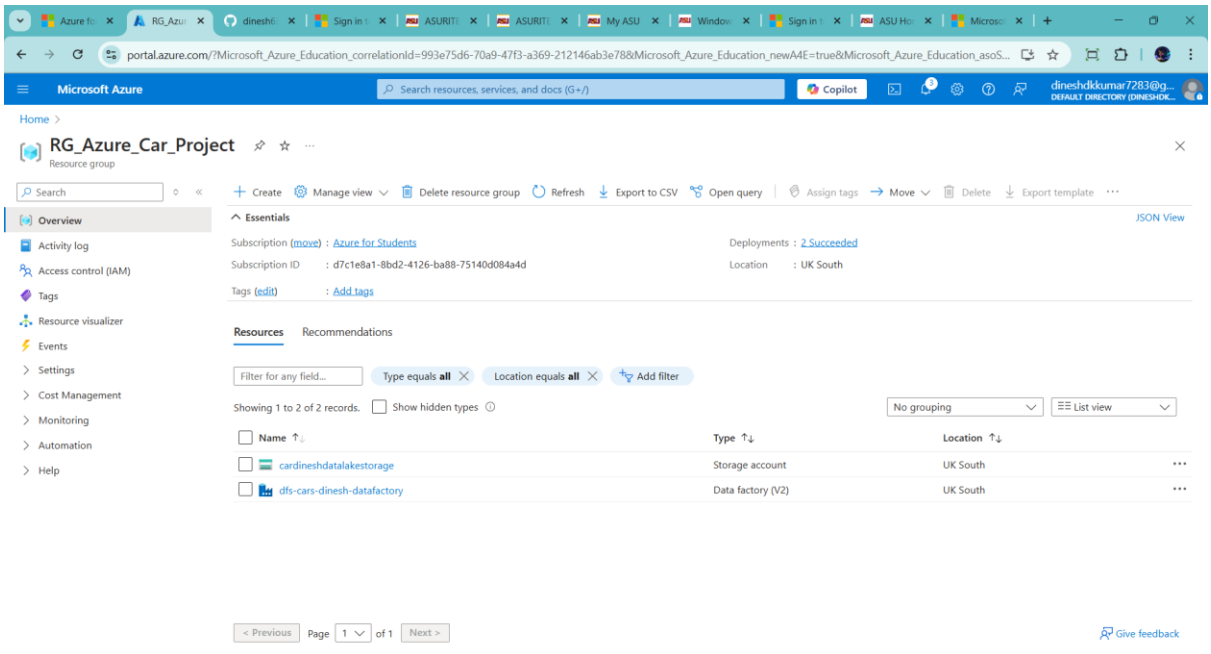


Click review + create

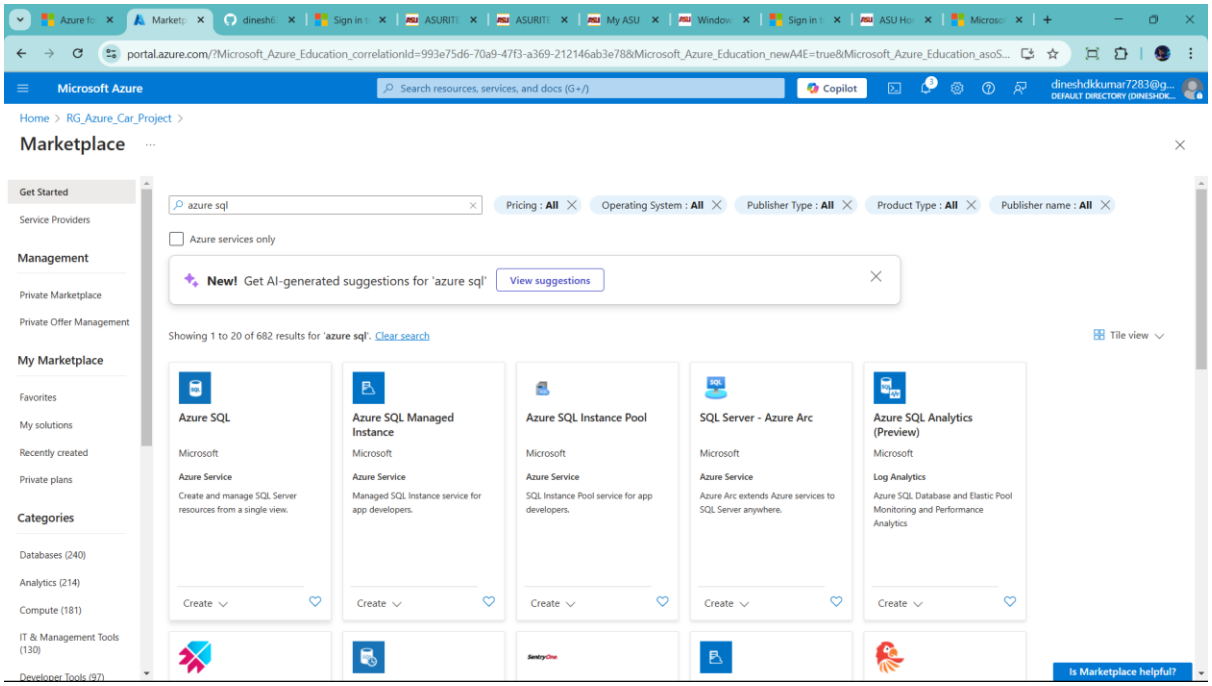


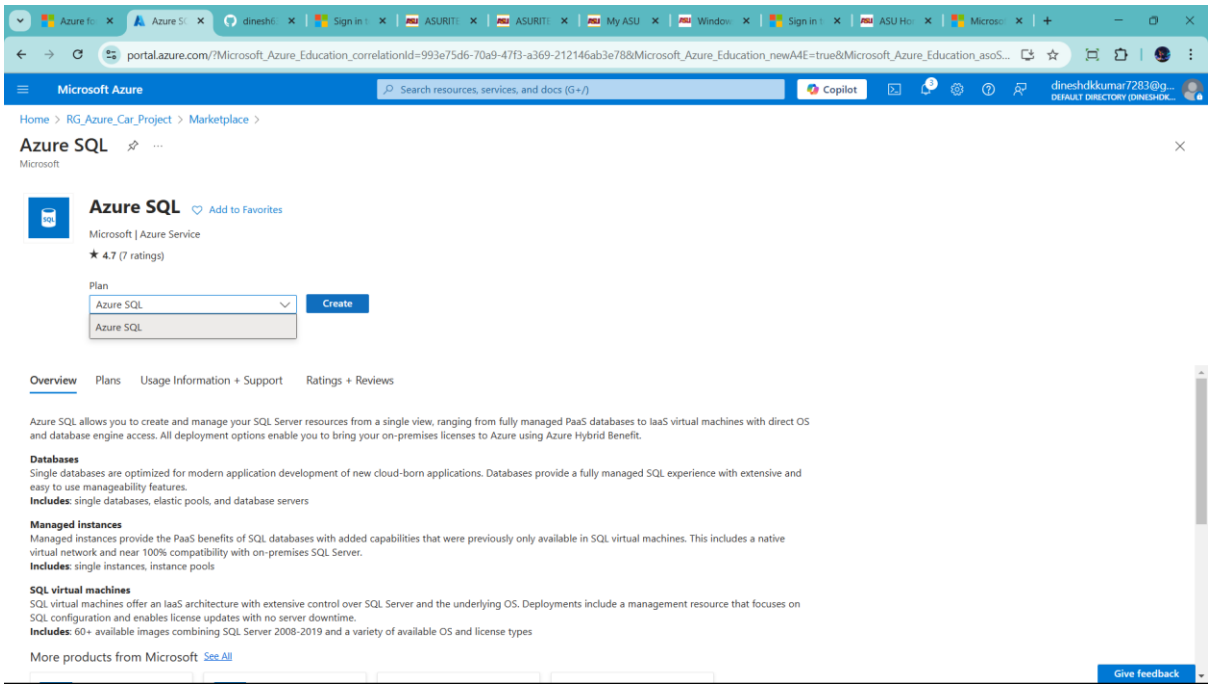


Finally click create to success full generate data factory

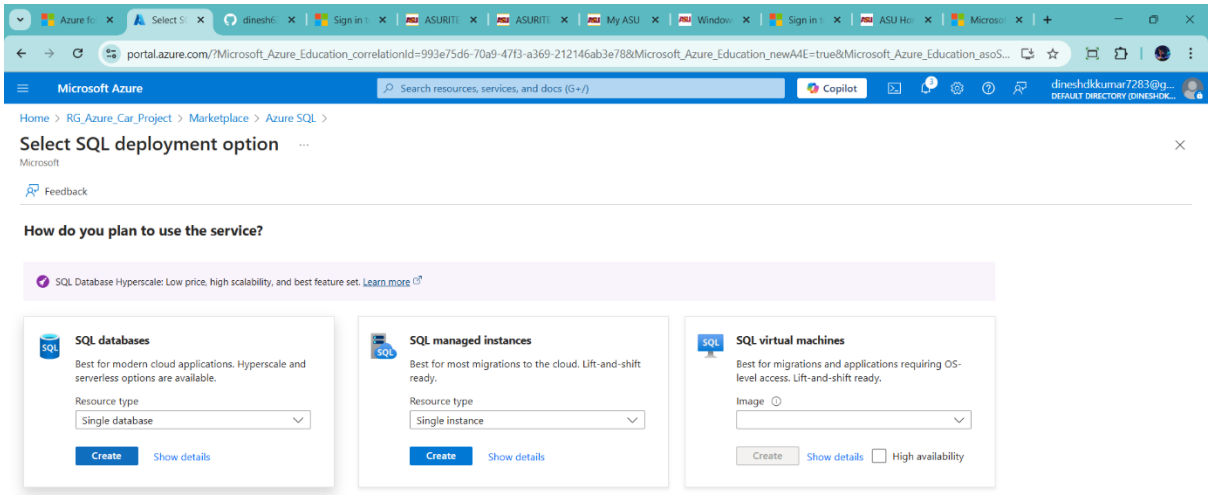


Now we Create third Resource Azure SQL DB



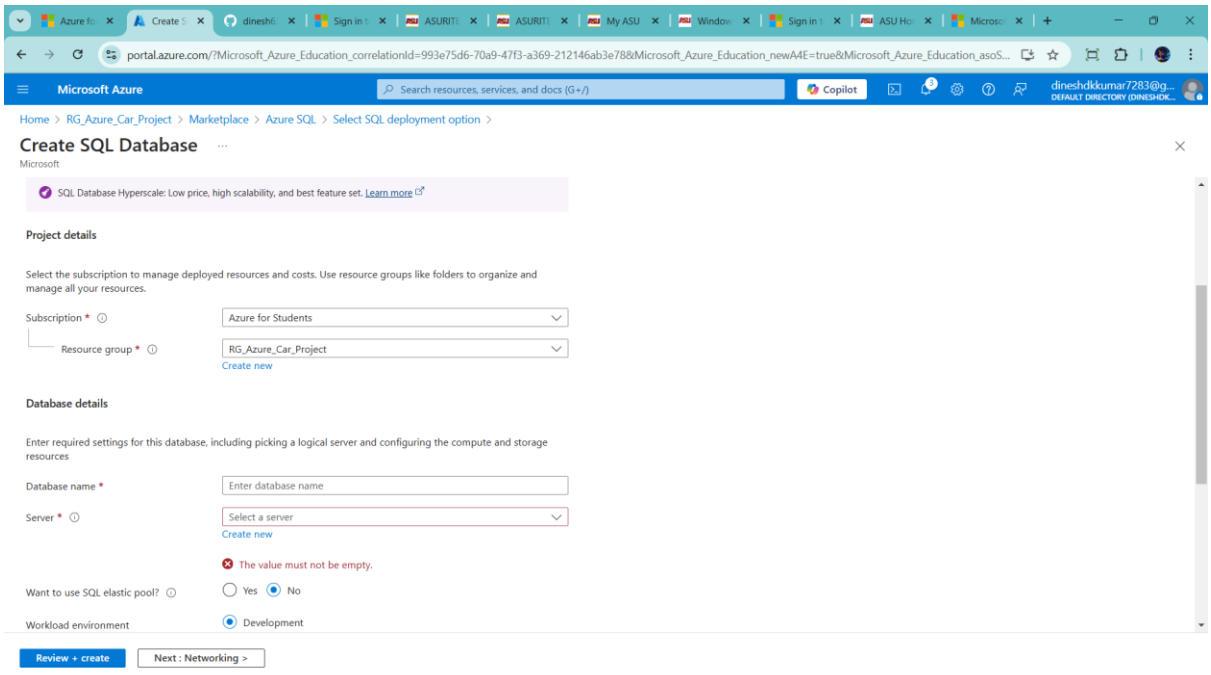


Click create



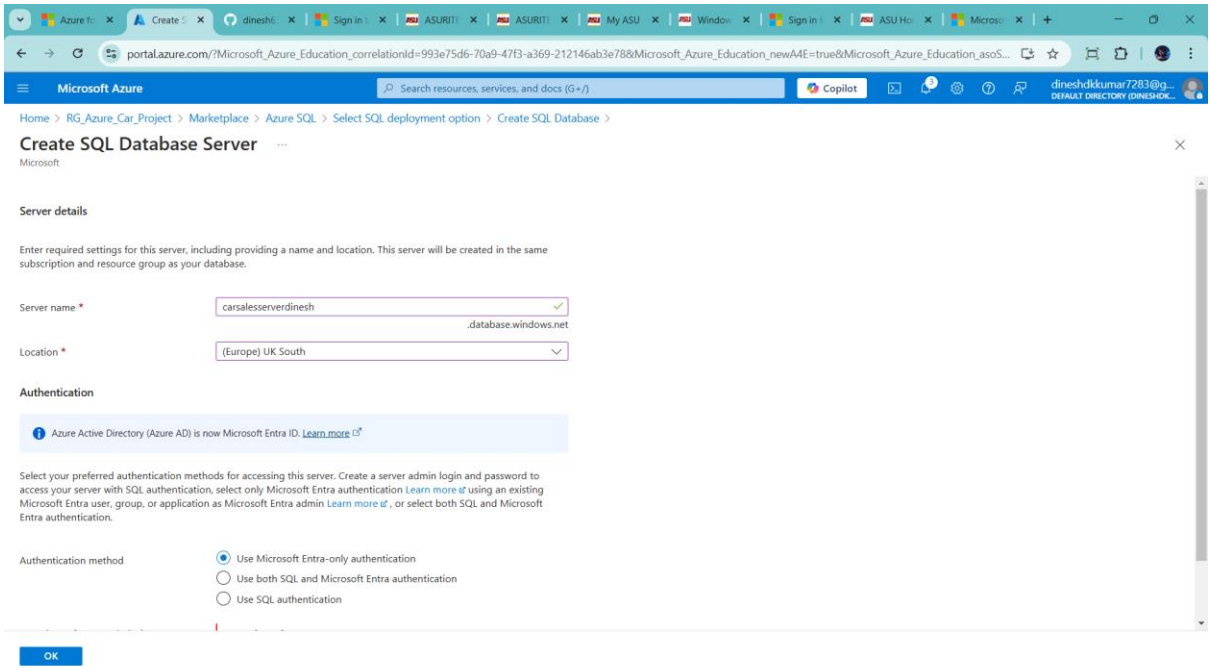
Click sql datbases for storage purpose only.

Click create

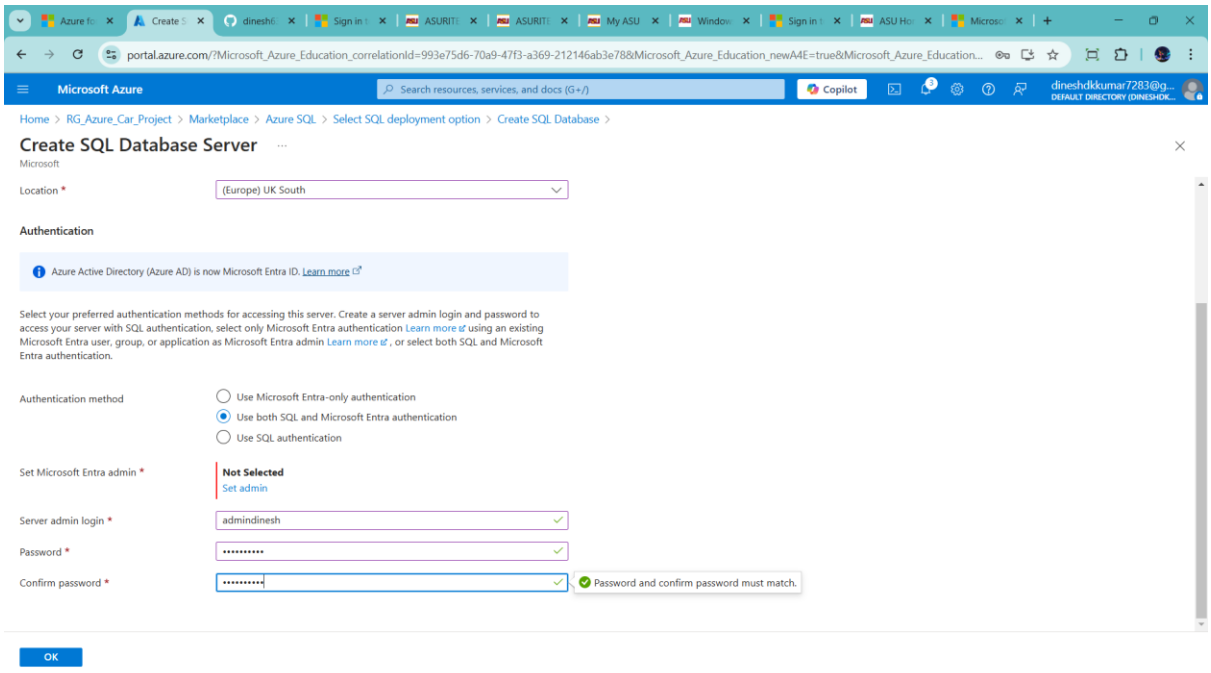


Fill database name and server

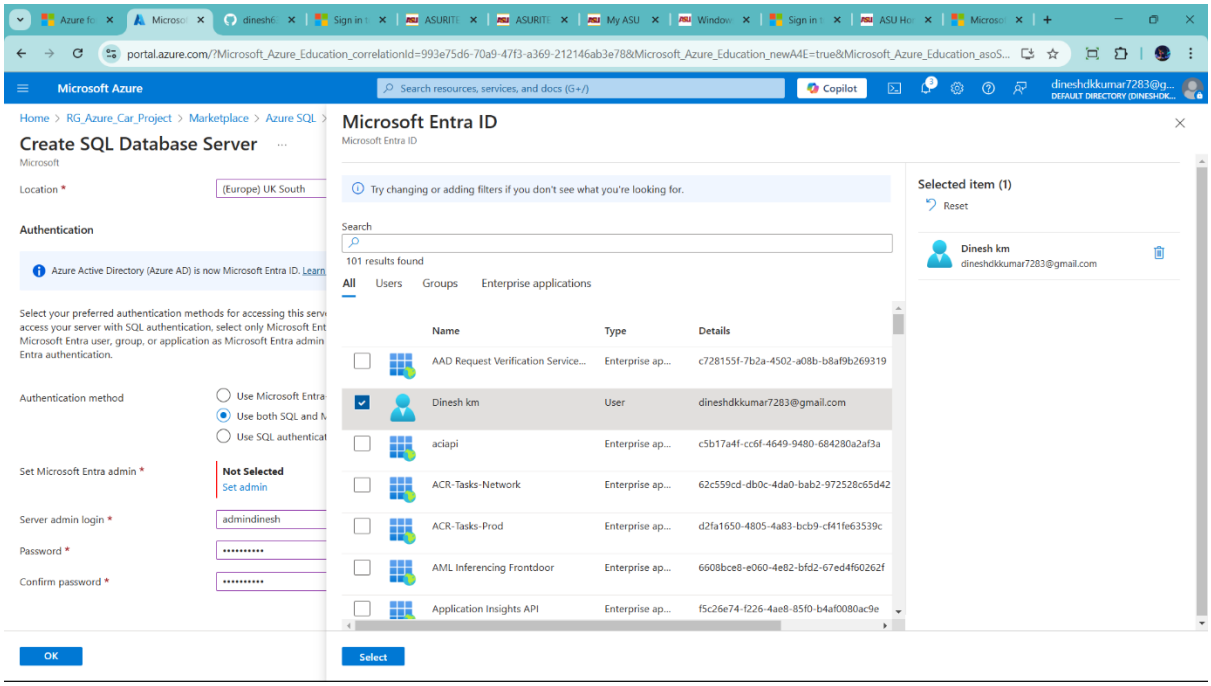
We donot have server. So create server So click new



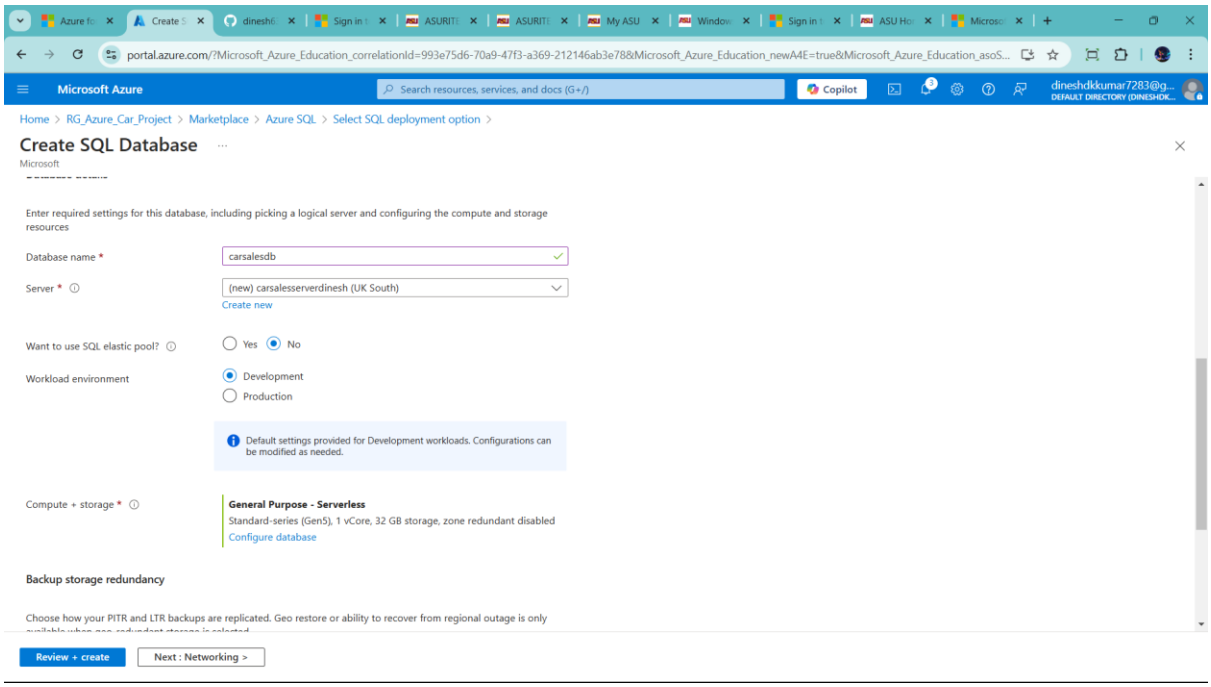
➔ Use both sql and Microsoft entra autenication in authentication method
➔ Set admin name and password
➔ Name→ admindinesh password →DIN123@kum



Than click set admin →



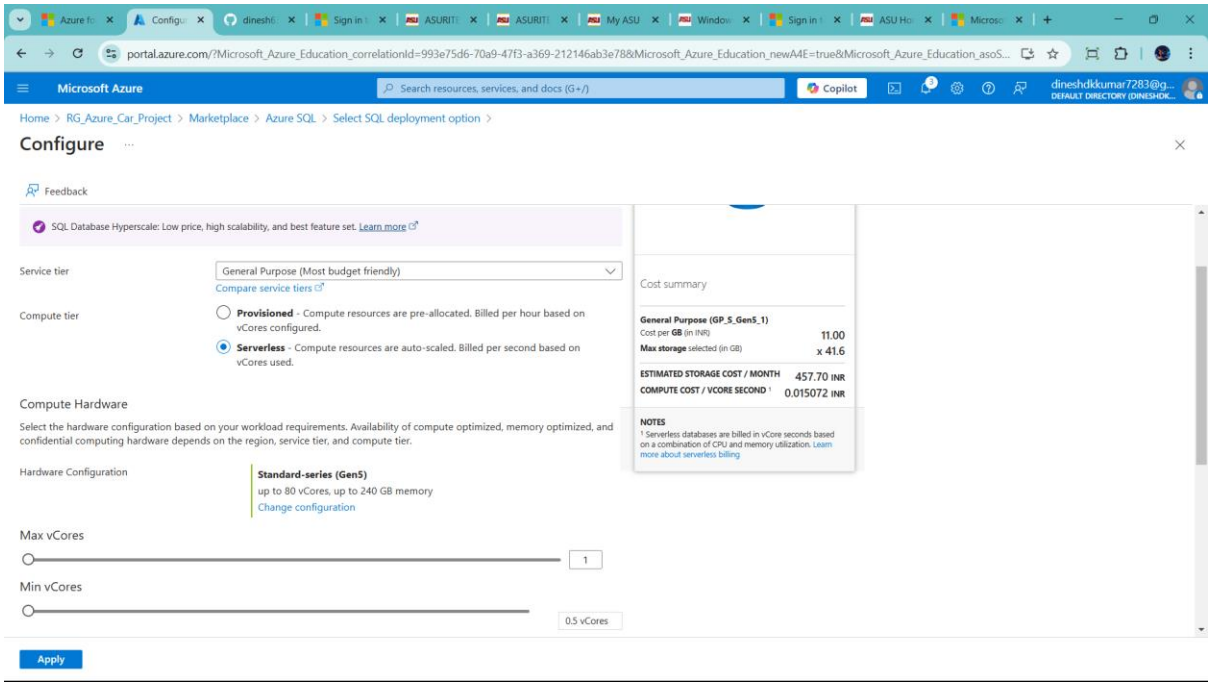
Final click ok



When set up your perfernce

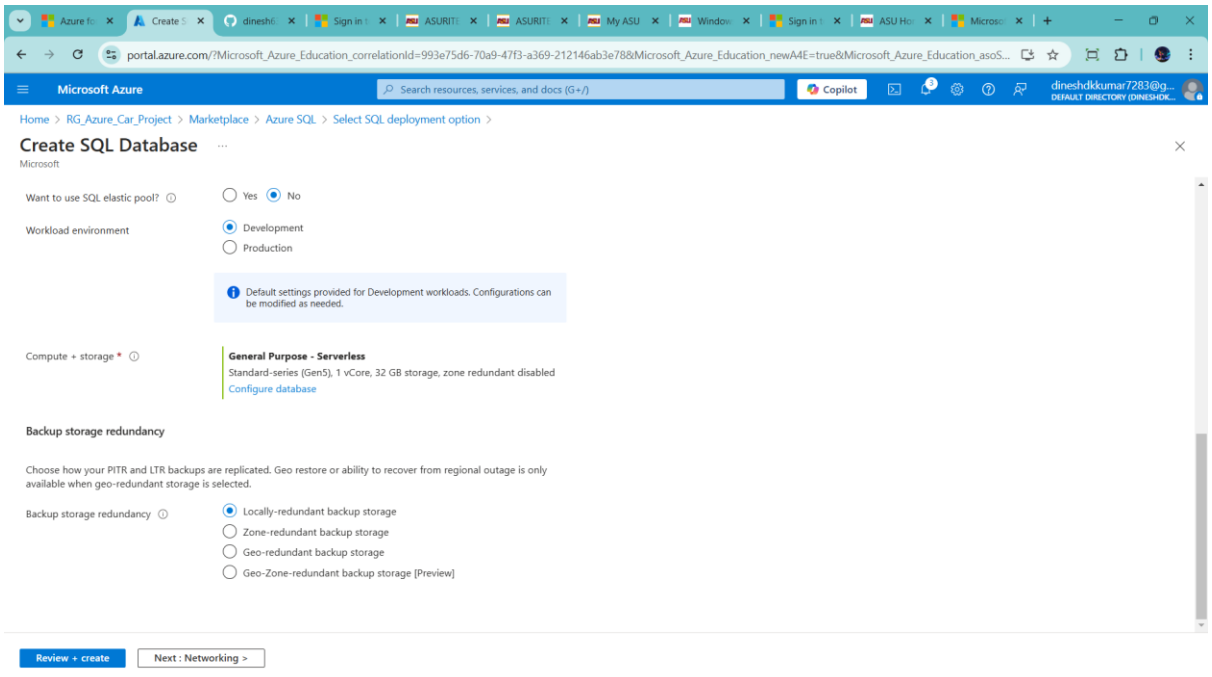
1) workload environment → Development

2) compute + storage → click configure database (for less storage)



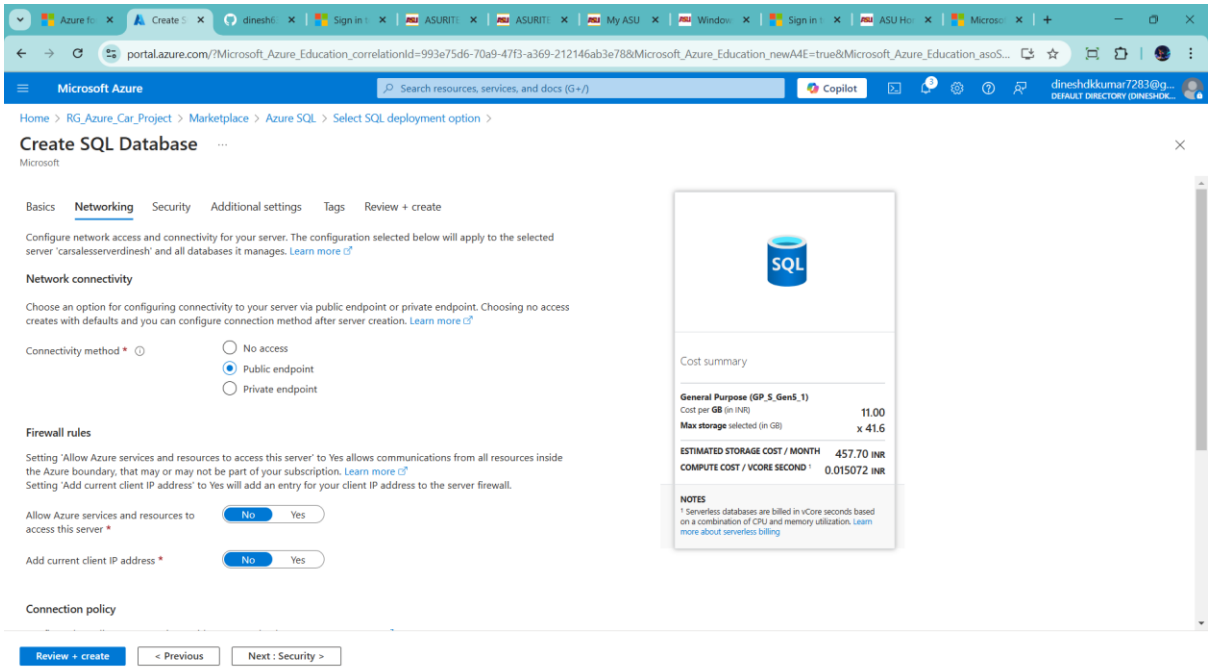
Than click apply

Backup storage redundancy → check to Local- redundant backup storage(for less)

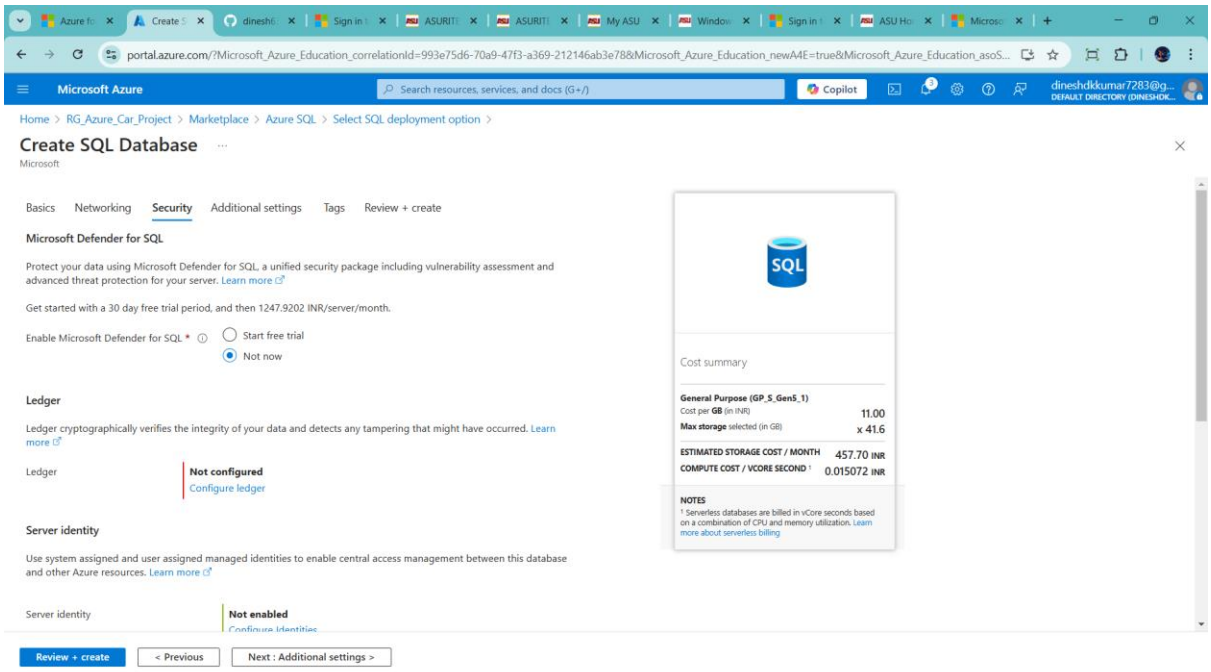


Click Next:Networking

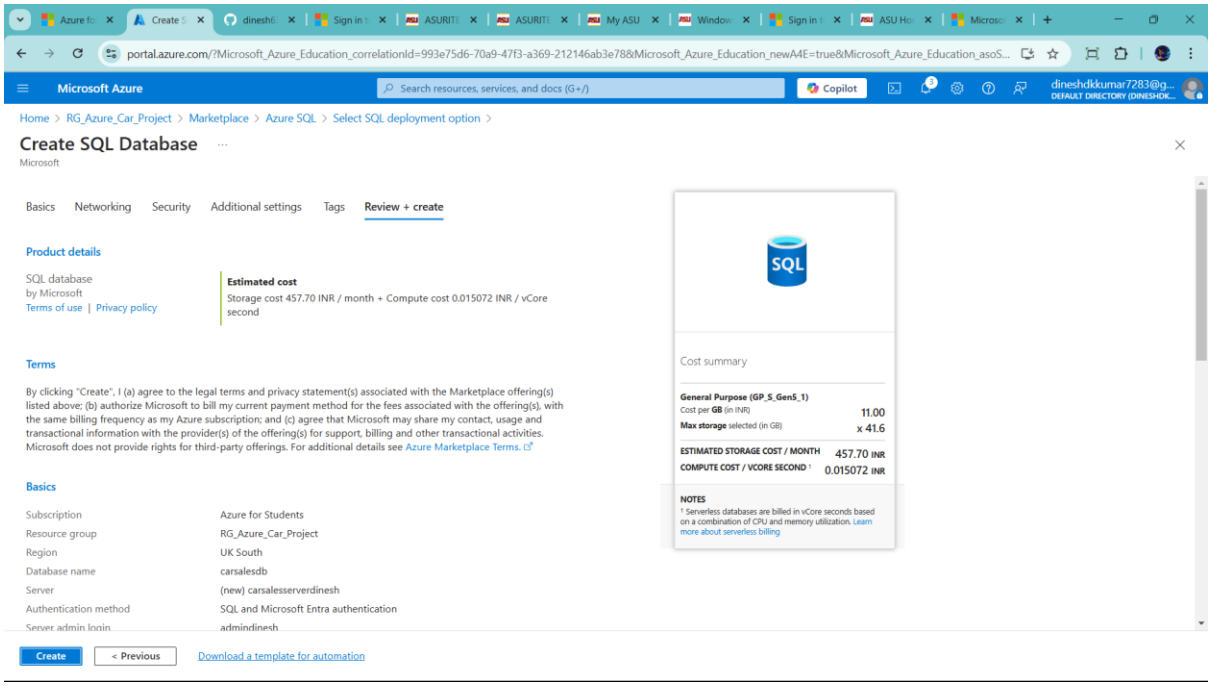
Change connectivity method →public endpoint



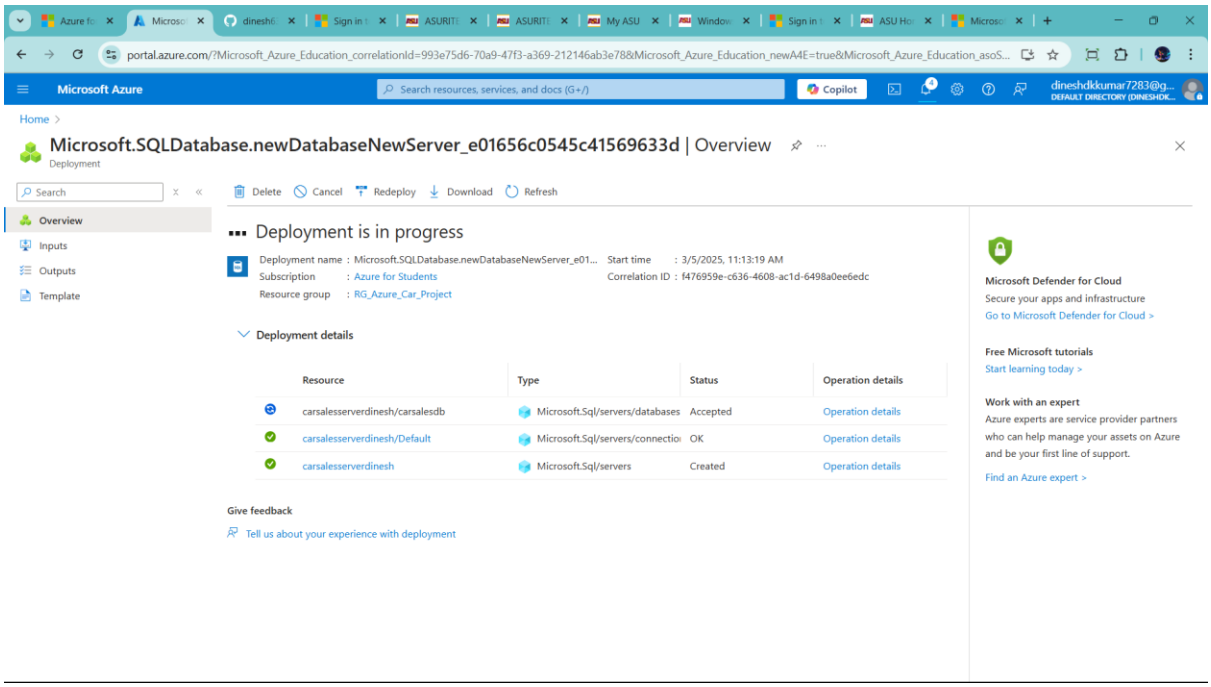
Click Next: Securtiy



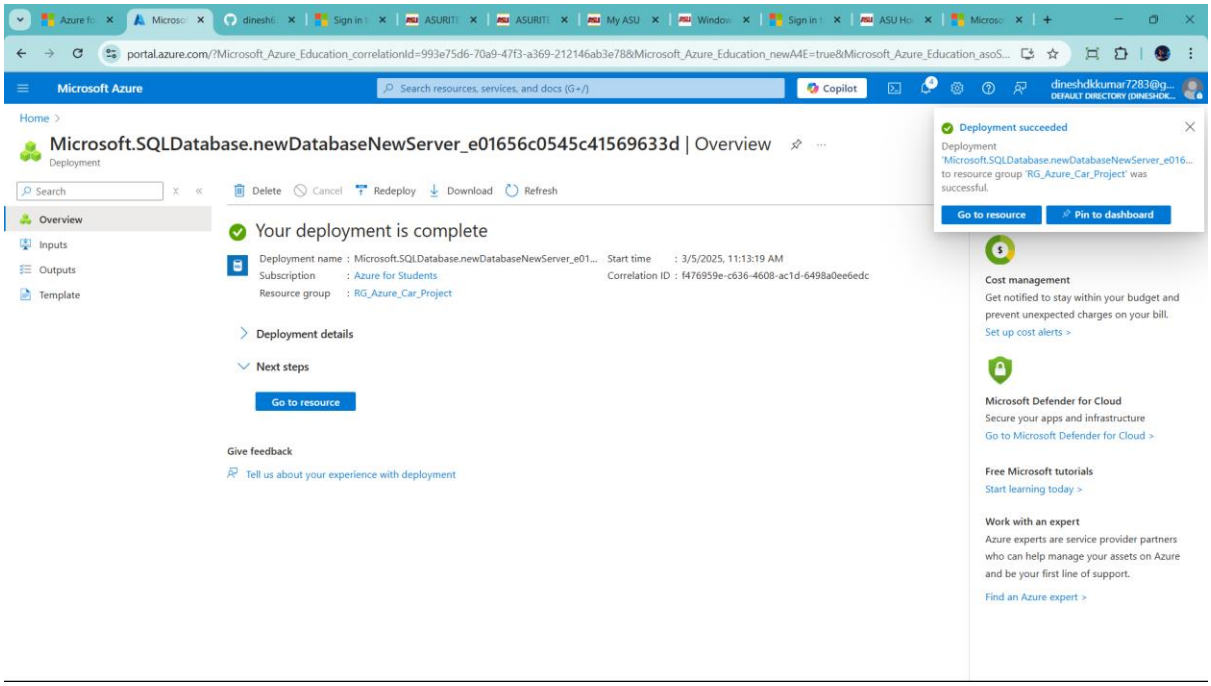
Click review+create

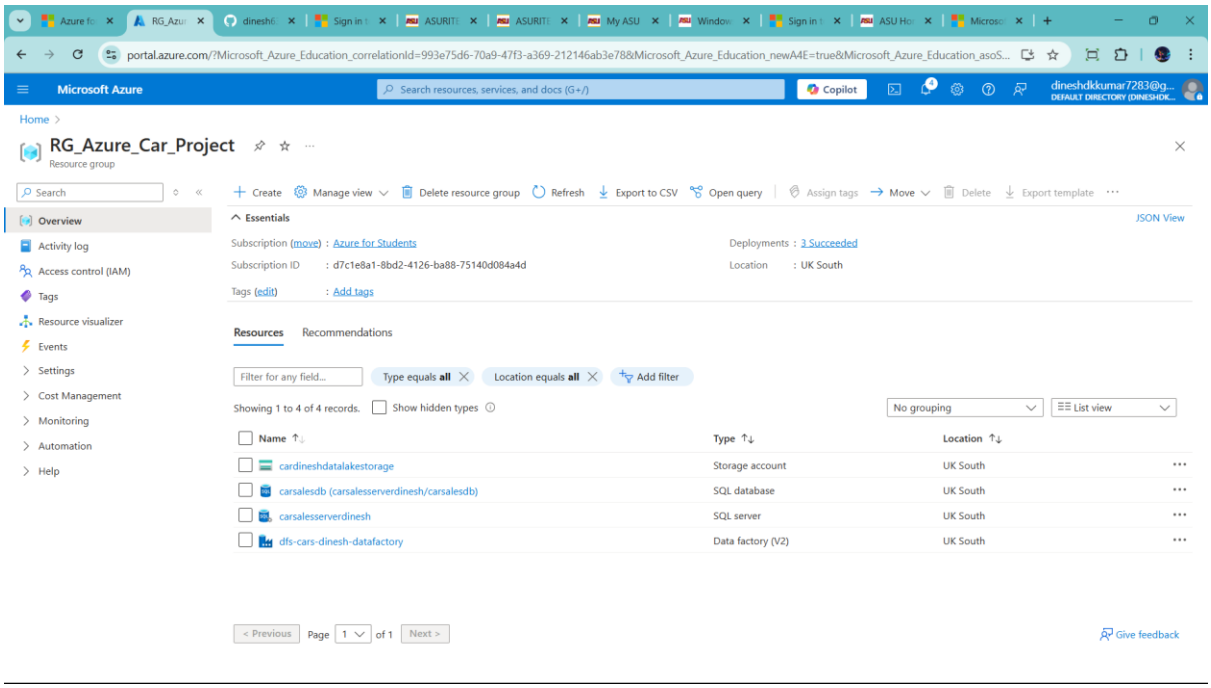


Click create



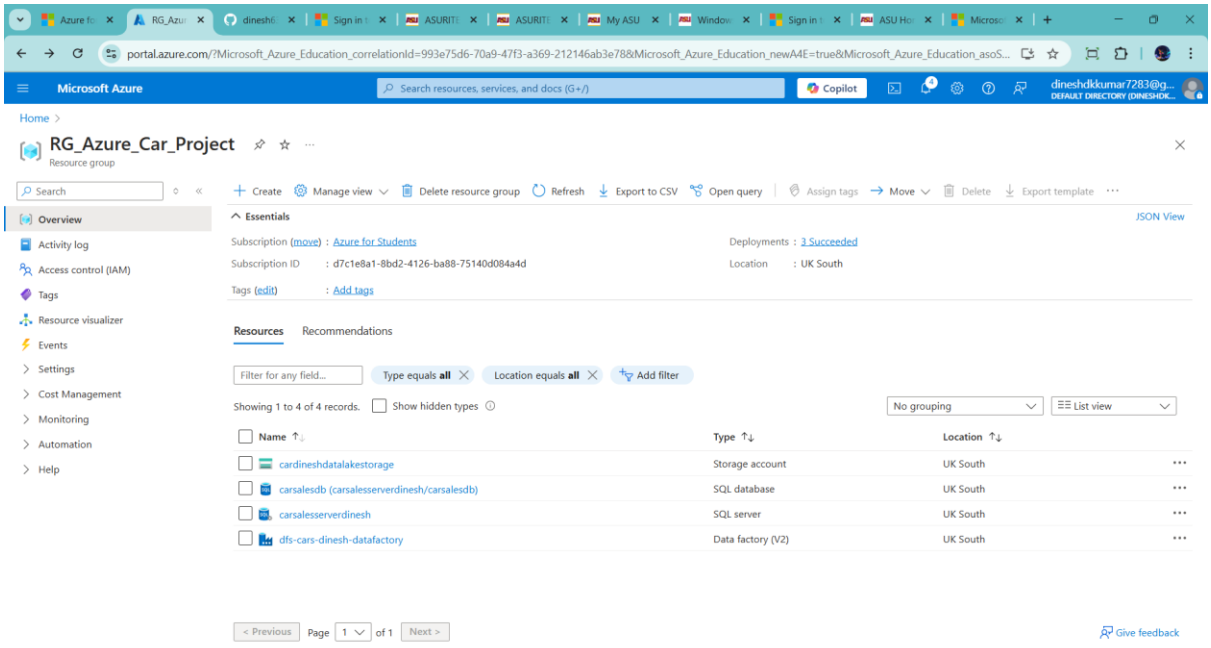
Finally complete sql setup

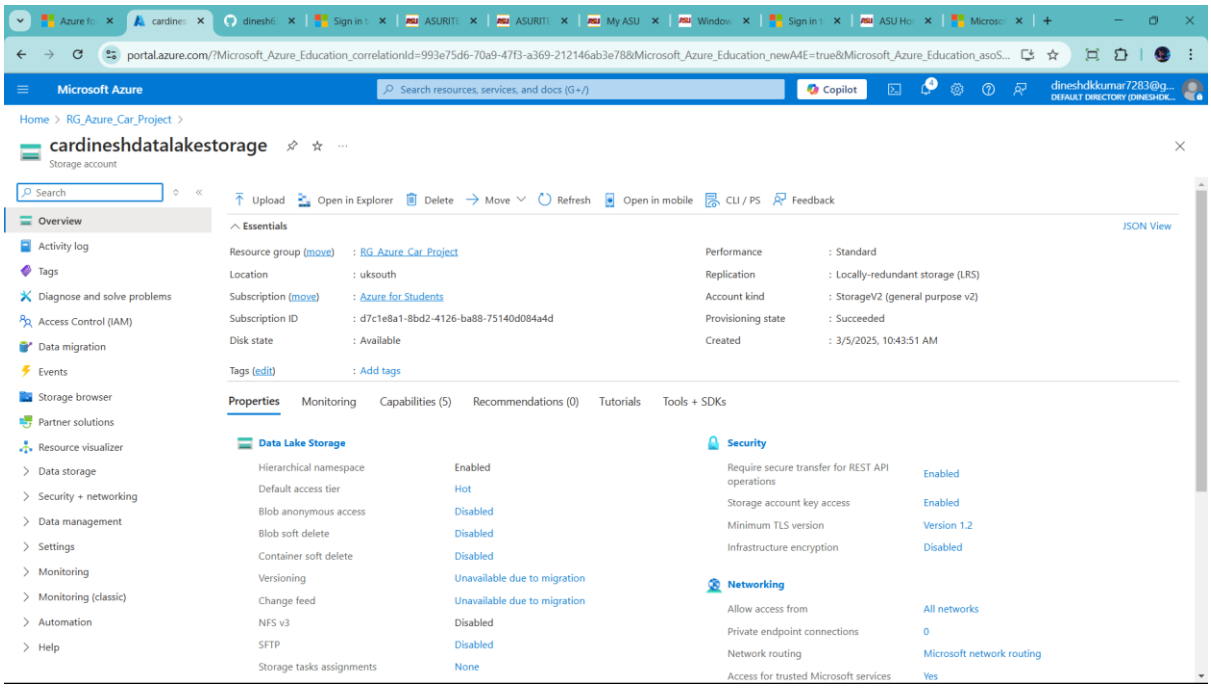

Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home
Microsoft.SQLDatabase.newDatabaseNewServer_e01656c0545c41569633d | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Your deployment is complete
Deployment name : Microsoft.SQLDatabase.newDatabaseNewServer_e01...
Subscription : Azure for Students
Resource group : RG_Azure_Car_Project
Start time : 3/5/2025, 11:13:19 AM
Correlation ID : f476959e-c636-4608-ac1d-6498a0ee6edc
Deployment details
Next steps
Go to resource
Give feedback
Tell us about your experience with deployment
Deployment succeeded
Deployment 'Microsoft.SQLDatabase.newDatabaseNewServer_e016... to resource group 'RG_Azure_Car_Project' was successful.
Go to resource
Pin to dashboard
Cost management
Get notified to stay within your budget and prevent unexpected charges on your bill.
Set up cost alerts >
Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >
Free Microsoft tutorials
Start learning today >
Work with an expert
Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support.
Find an Azure expert >


Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home
RG_Azure_Car_Project
Resource group
Search
Overview
Activity log
Access control (IAM)
Tags
Resource visualizer
Events
Settings
Cost Management
Monitoring
Automation
Help
Create
Manage view
Delete resource group
Refresh
Export to CSV
Open query
Assign tags
Move
Delete
Export template
Essentials
JSON View
Subscription (move) : Azure for Students
Subscription ID : d7c1e8a1-8bd2-4126-ba88-75140d084a4d
Tags (edit) : Add tags
Deployments : 3 Succeeded
Location : UK South
Resources
Recommendations
Filter for any field...
Type equals all
Location equals all
Add filter
Showing 1 to 4 of 4 records.
Show hidden types
No grouping
List view
Name
Type
Location
cardineshdatalakestorage
Storage account
UK South
carsalesdb (carsalesserverdinesh/carsalesdb)
SQL database
UK South
carsalesserverdinesh
SQL server
UK South
dfs-cars-dinesh-datafactory
Data factory (V2)
UK South
< Previous
Page 1 of 1
Next >
Give feedback

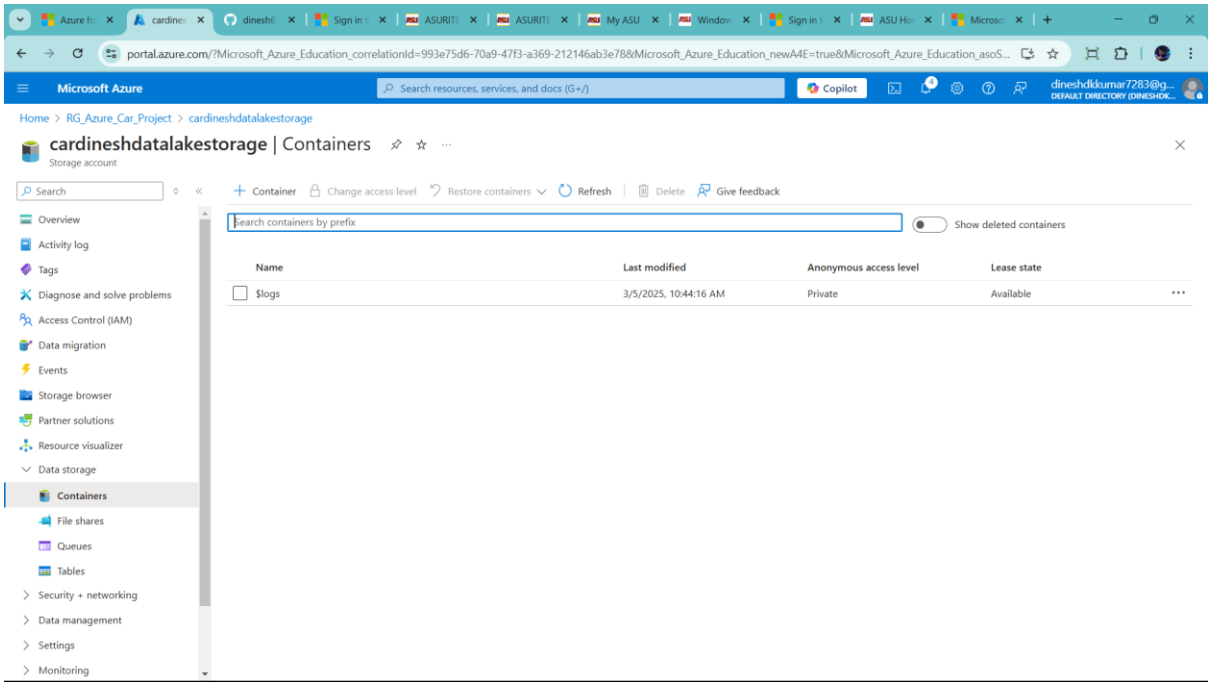
Now create datalake in brozen,silver and Gold

Click resource manager → cardineshdatalakesstorage

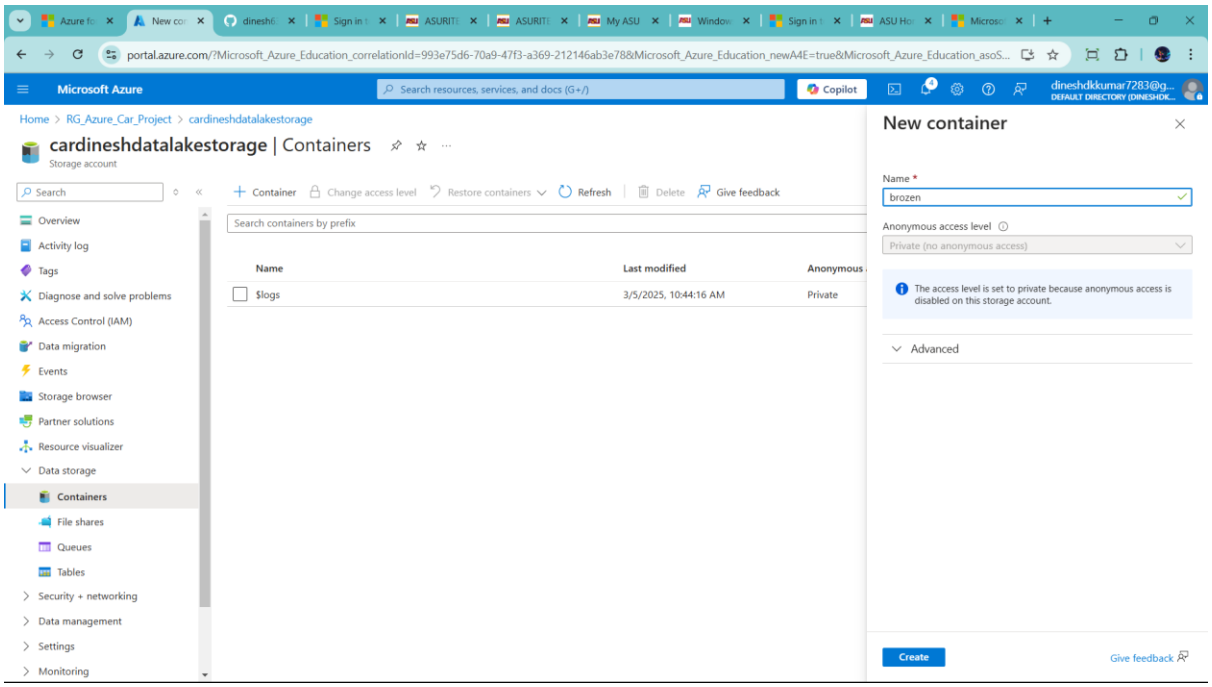


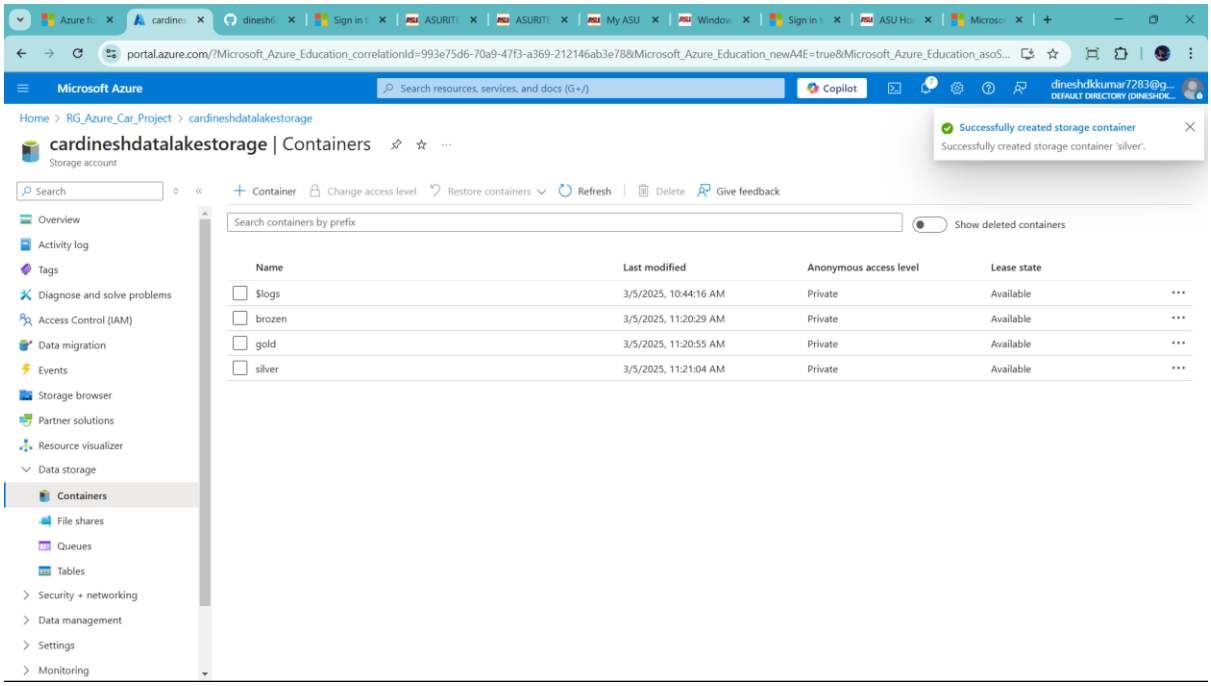


Click container prepare data layers brozen,silver,gold

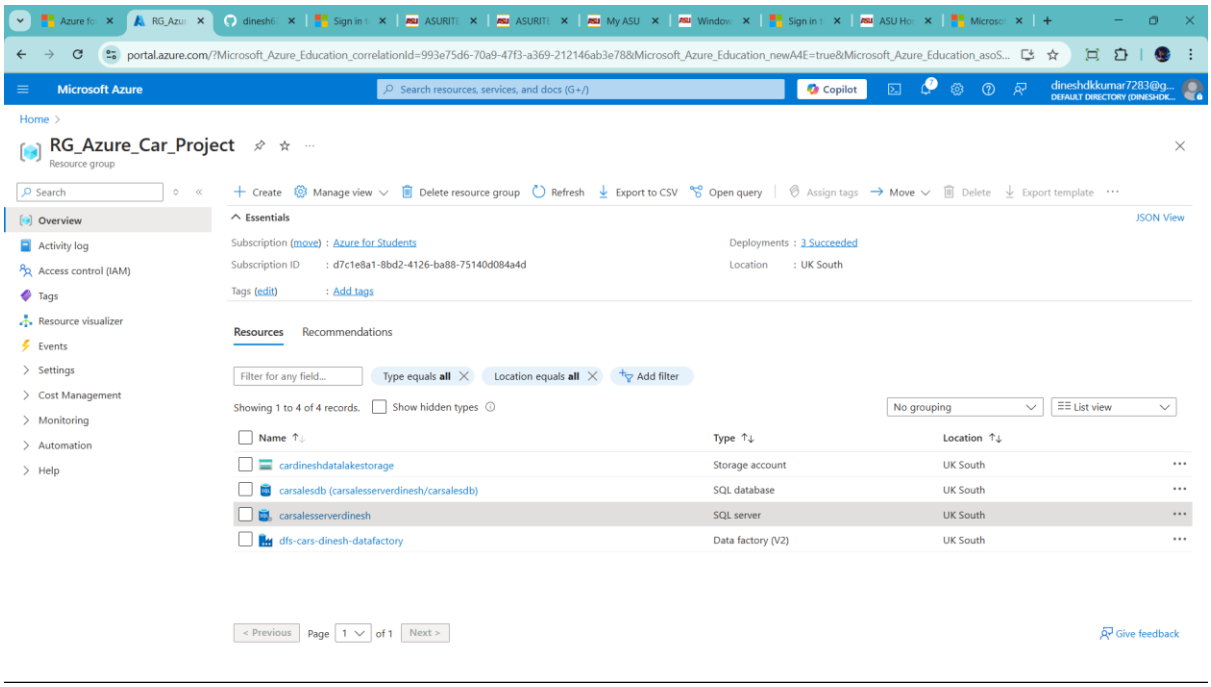


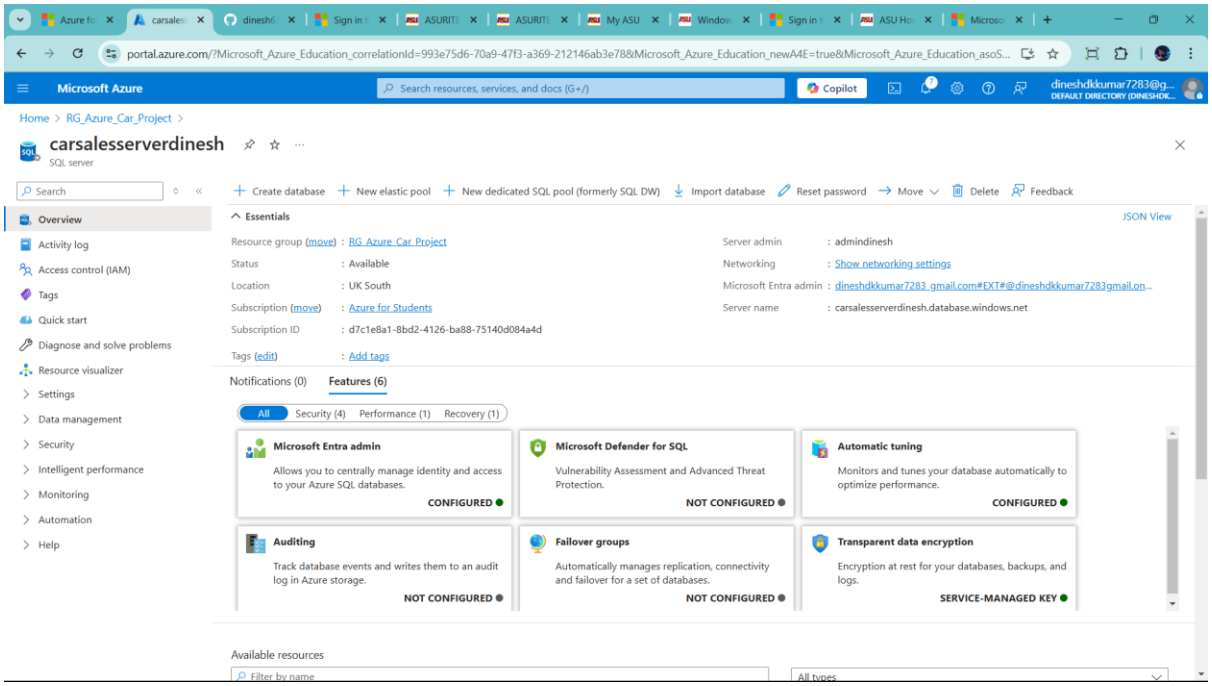
Click container → write brozen





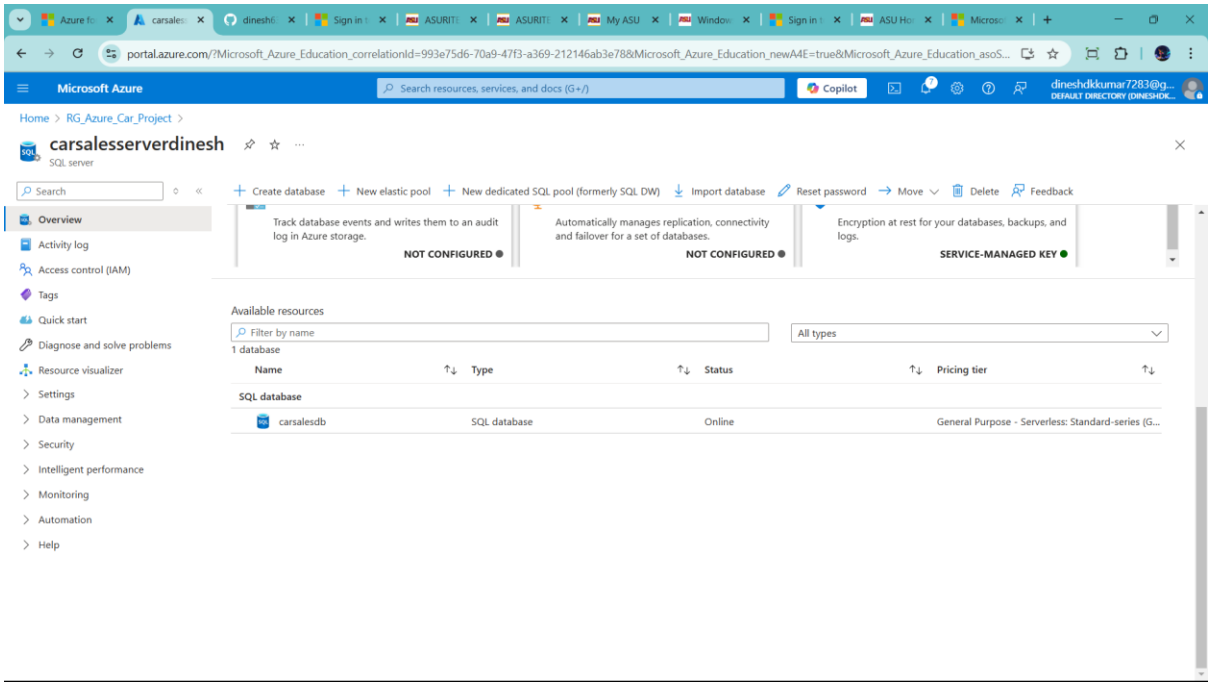
Click carsalesserverdinesh (SQL server)



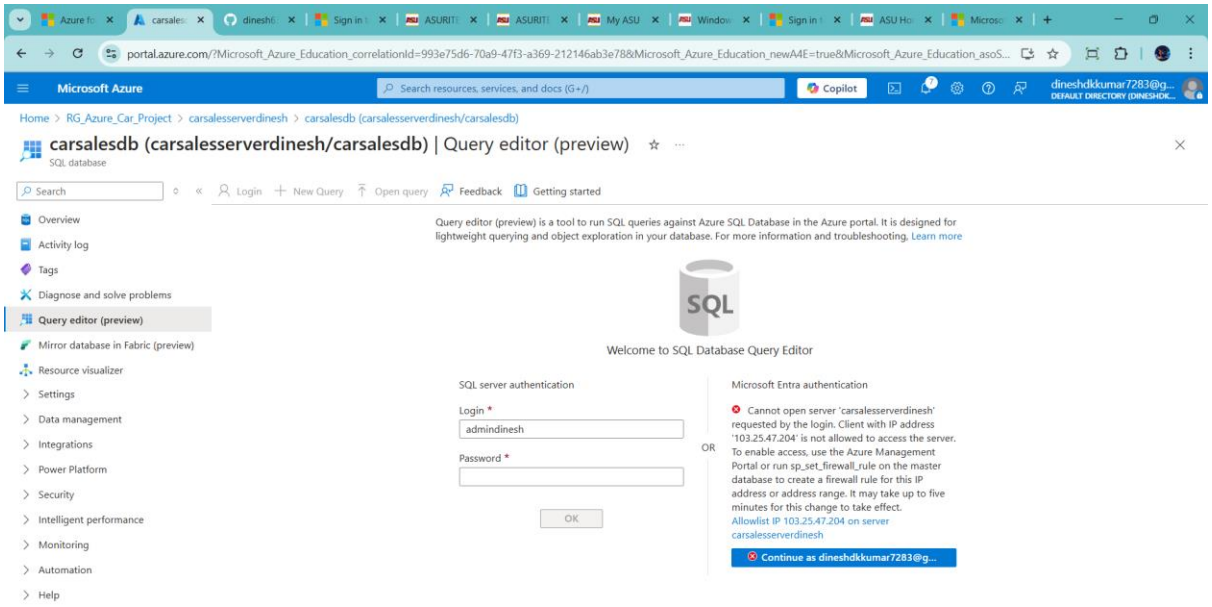


this method add more data bases as you preference

This is databased already create previse

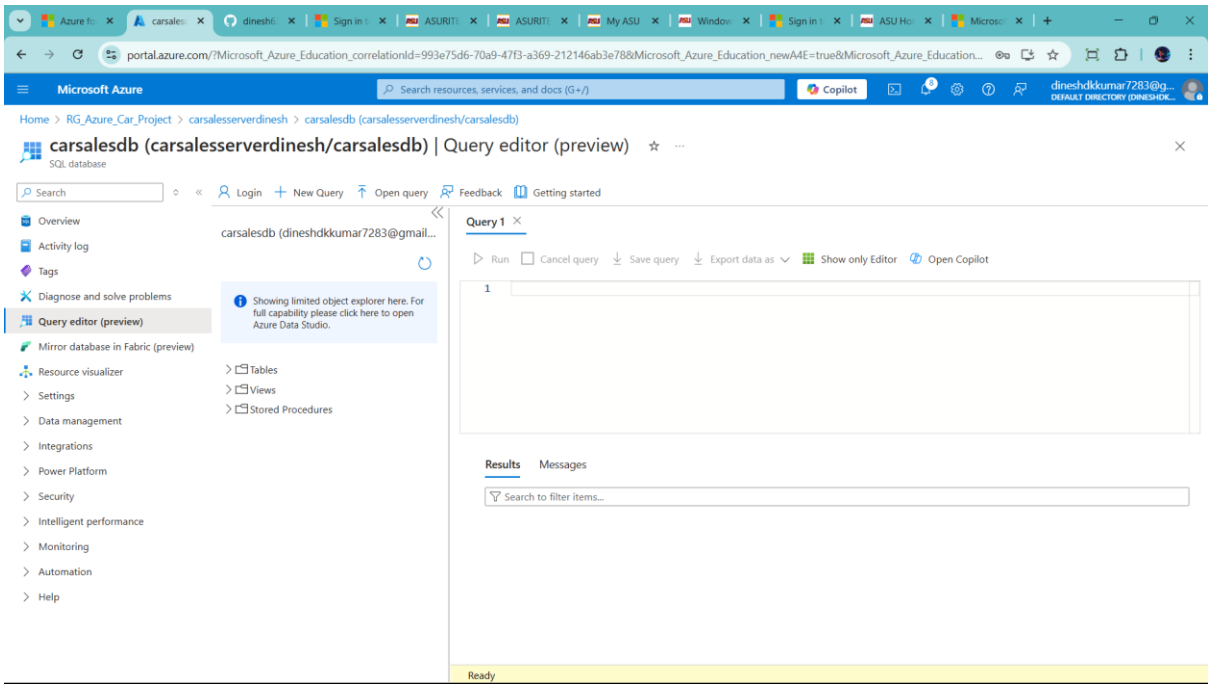


Click carsalesdb →then click query editor(preview)

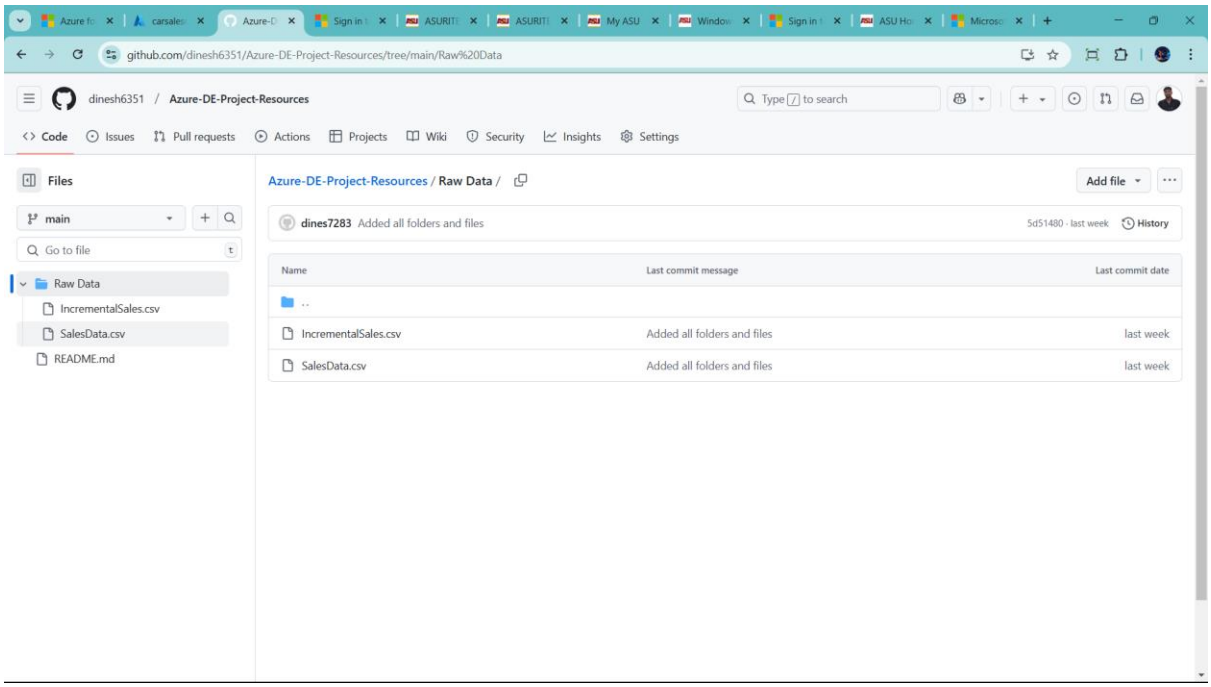


Click allowlist IP (server less card)

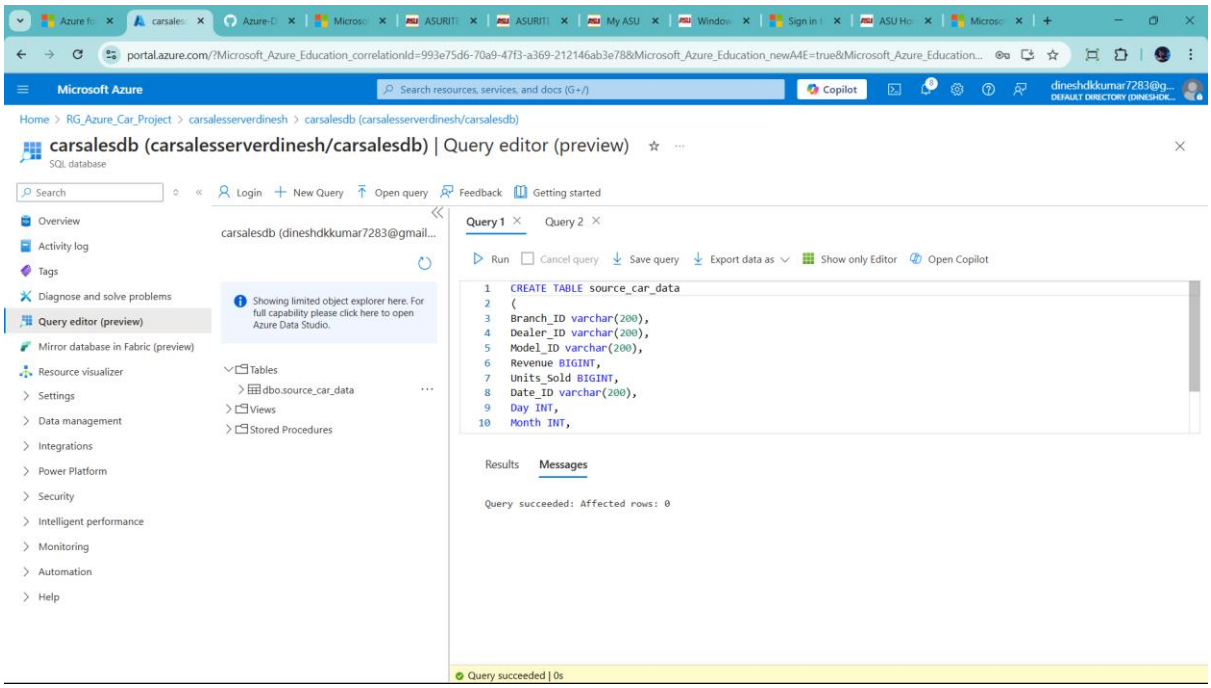
After remove error than add admin and password

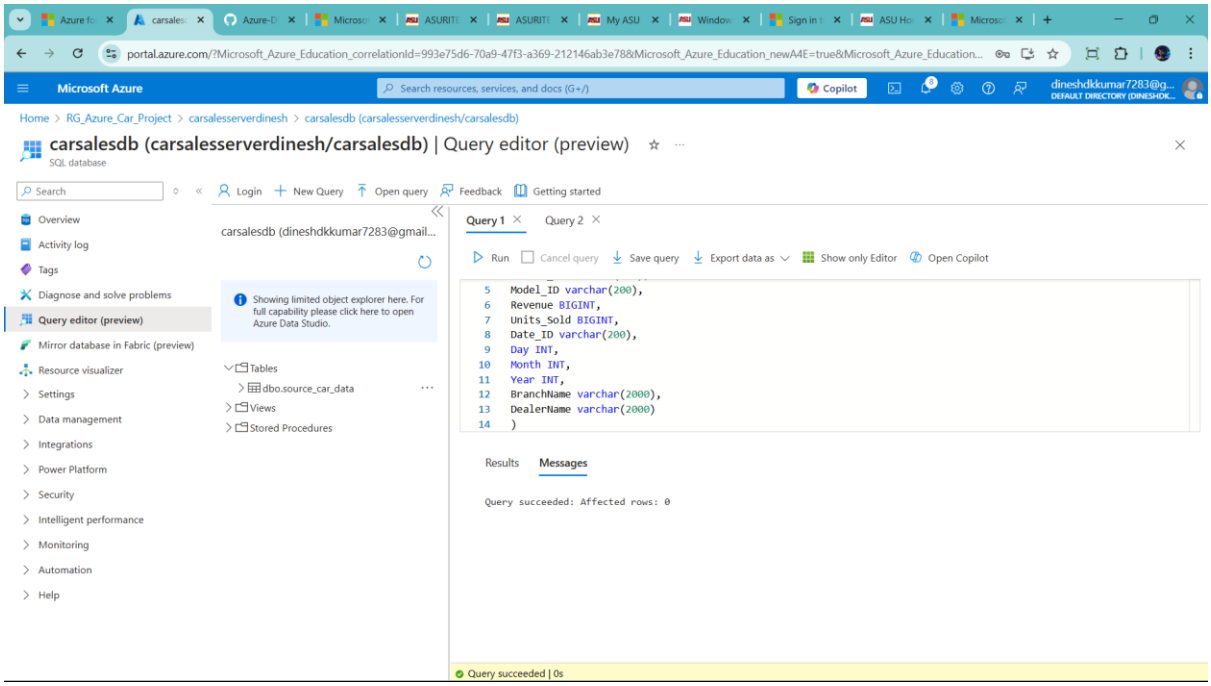


Lets See Source data in github

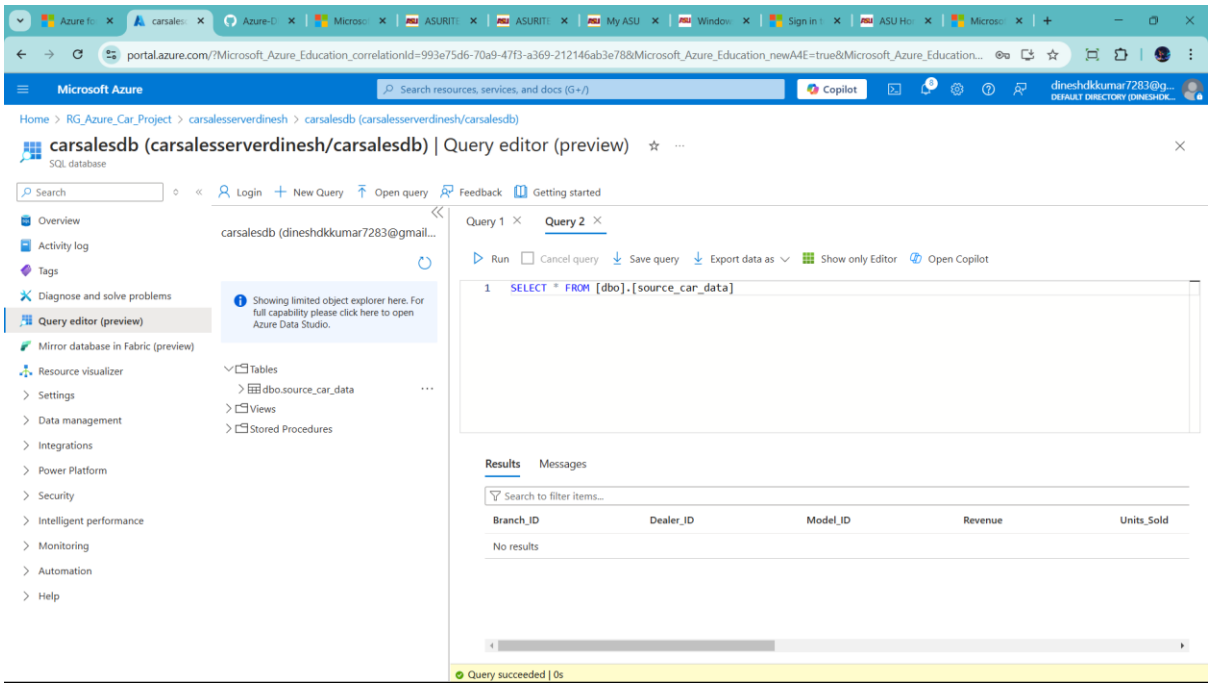


Let start create source table in Azure SQL





Check table create or not



CREATE TABLE source_car_data
(
Branch_ID varchar(200),
Dealer_ID varchar(200),
Model_ID varchar(200),
Revenue BIGINT,
Units_Sold BIGINT,
Date_ID varchar(200),
Day INT,
Month INT,
Year INT,
BranchName varchar(2000),
DealerName varchar(2000)
)

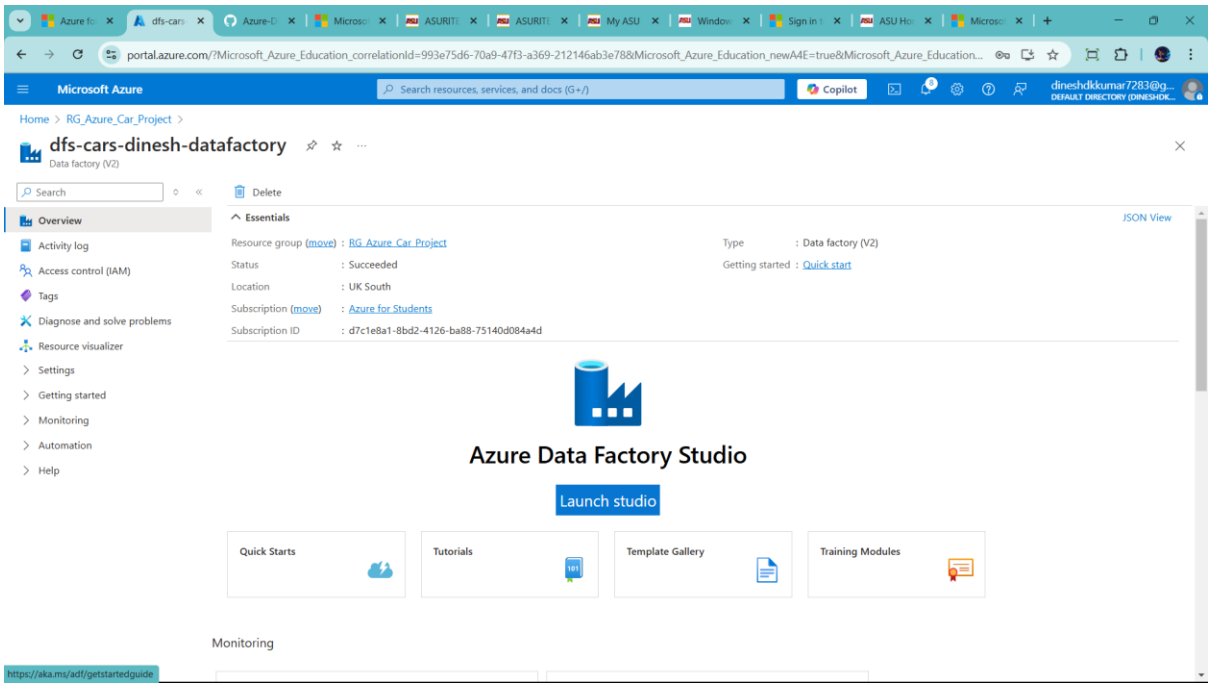
→data lake ready

→sql server ready

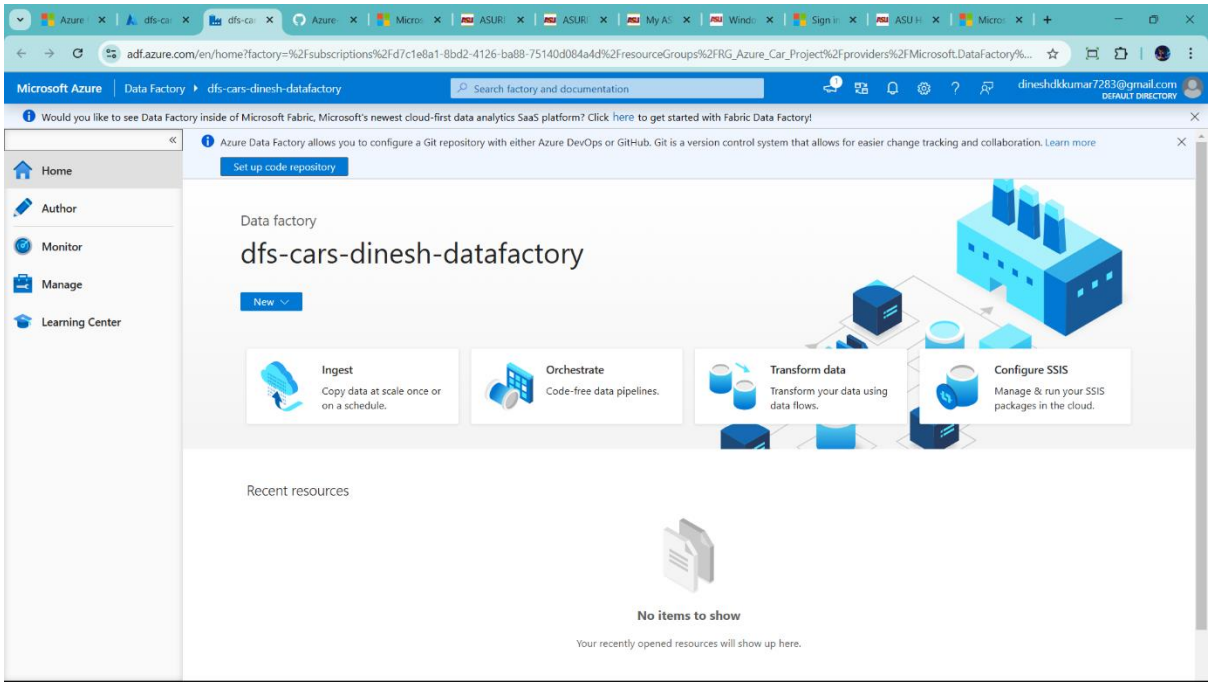
Now Strat build Data factory

→pull the data from github to azure

→let's start



→click launch studio



1) Author →use for craft the pipe lines →main one

2) Monitor → Monitor the pipline(all logs)

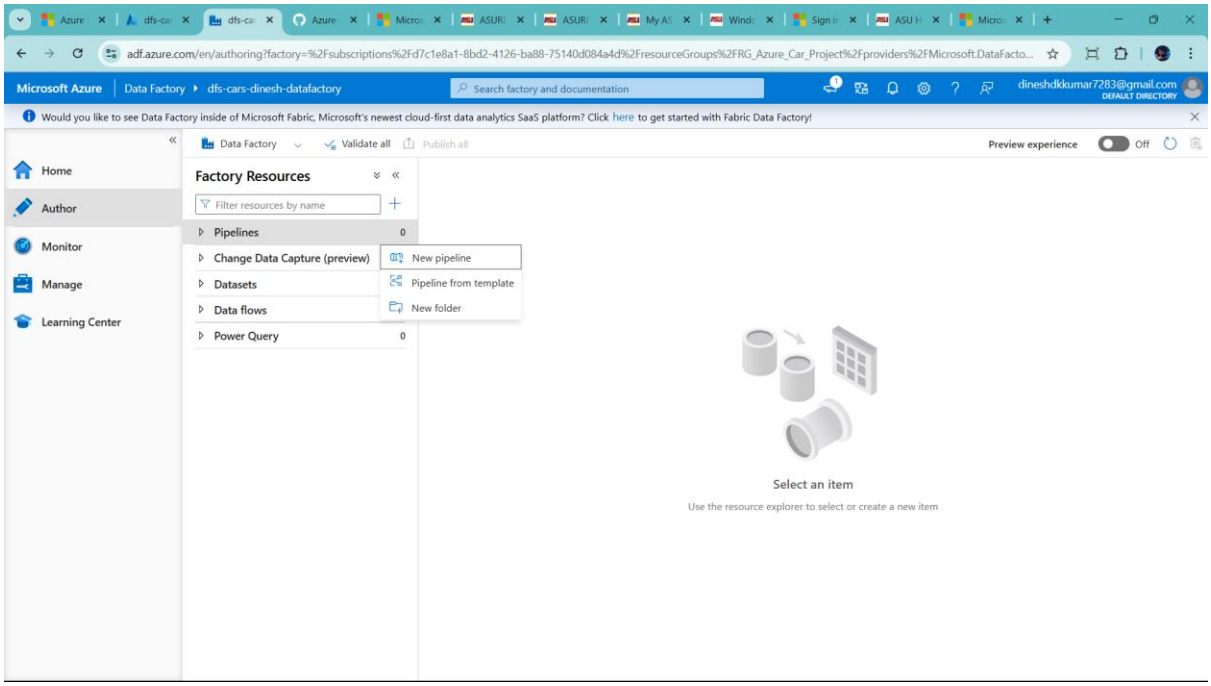
3) Manage → manage all the tools like github repostory,connection, all the trigger and more

4) finally learning tap

Now Start with Author tap and start build piple

Step1) click author

Step2) click pipeline and add new pipeline

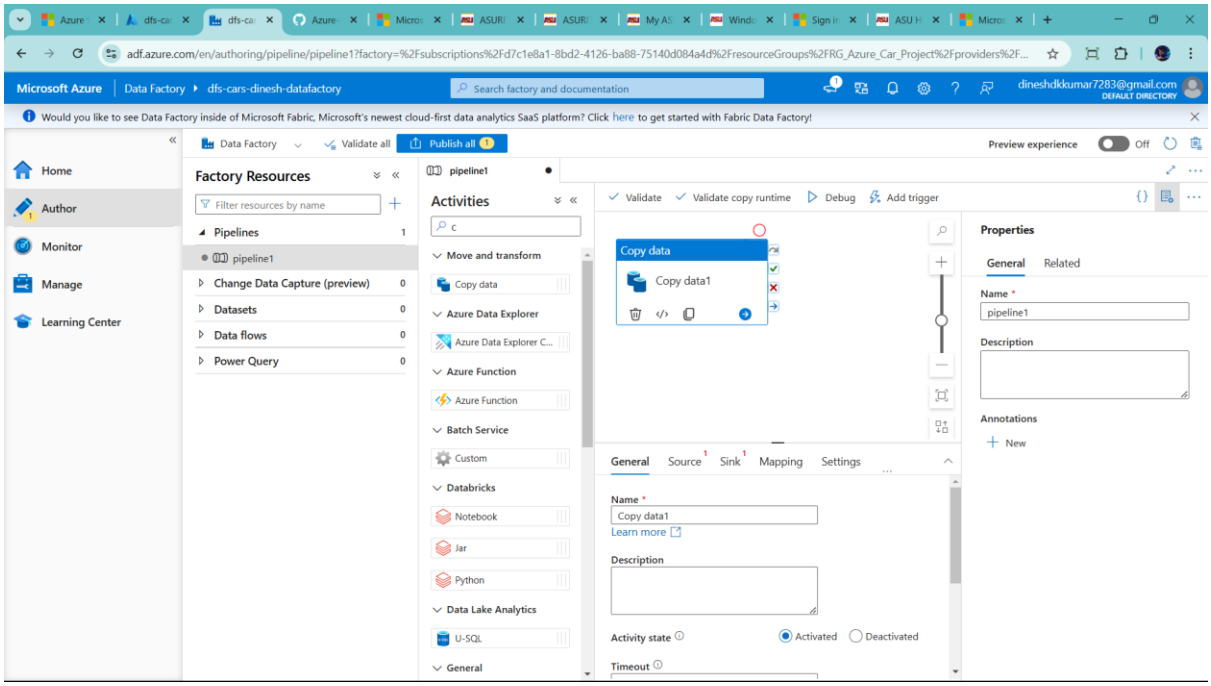


How to connection work

Github -------→ Data Factory ----------→ SQL

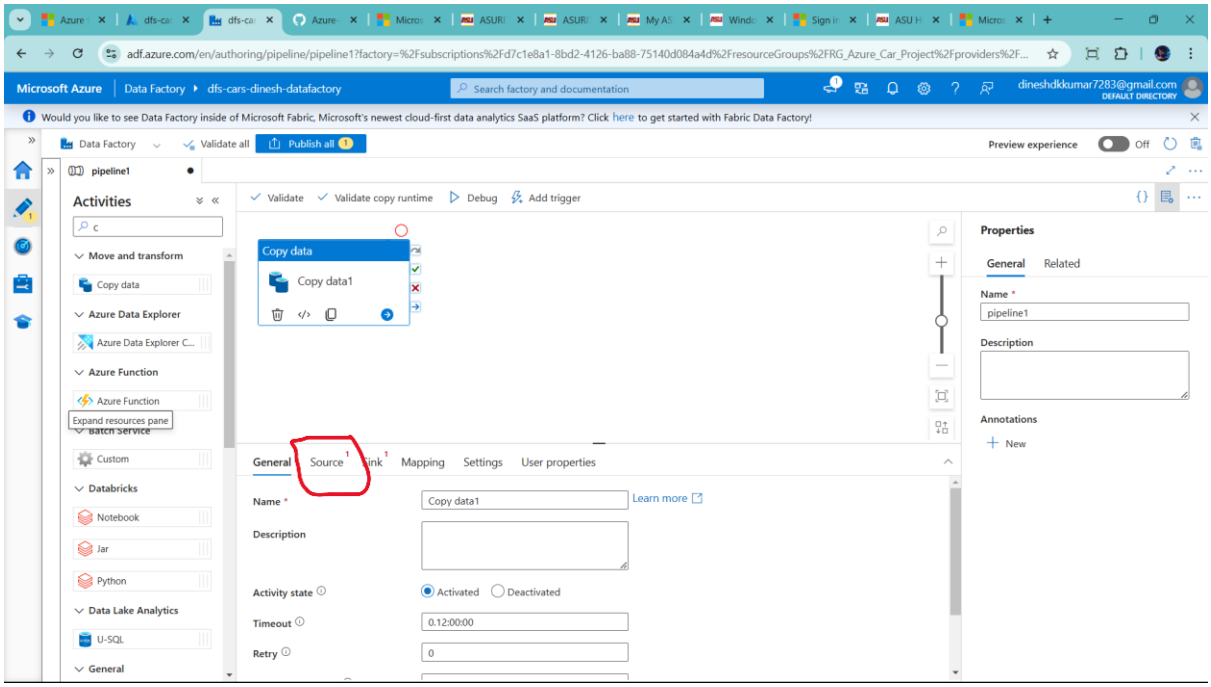
Load increment data

1) copy activity in data factory → migrate data one location another

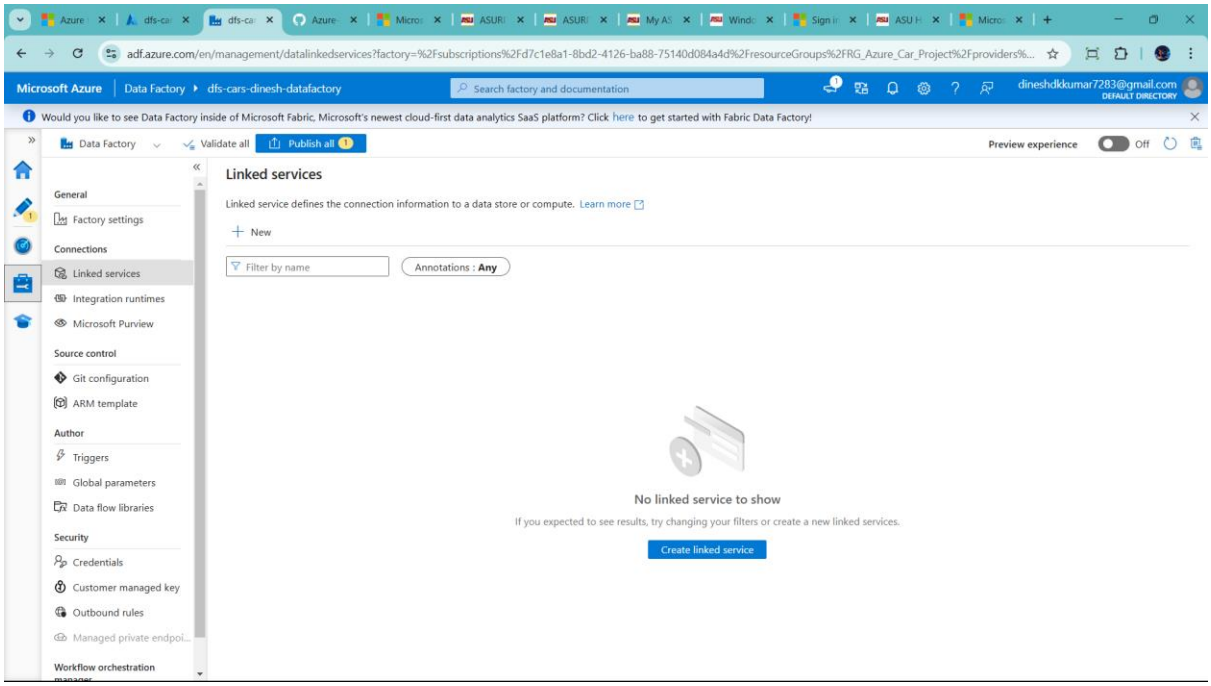


It is connection with source

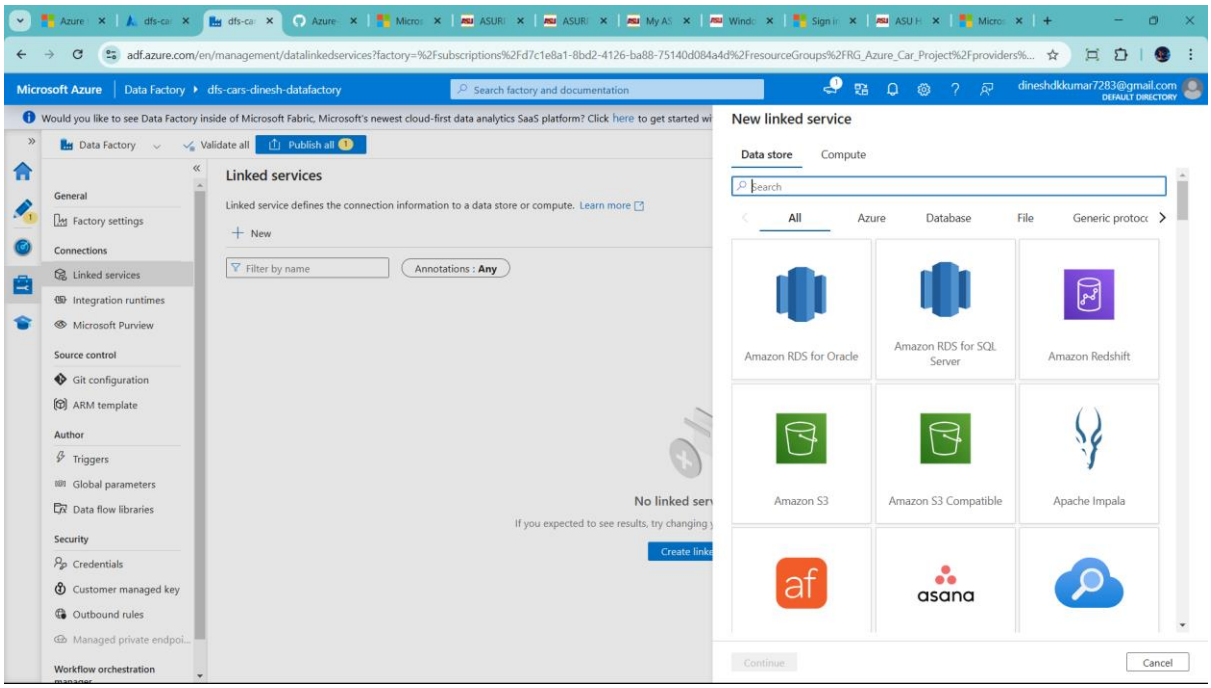
Two side to manage the copy data activity one is below source and another is manage tab’



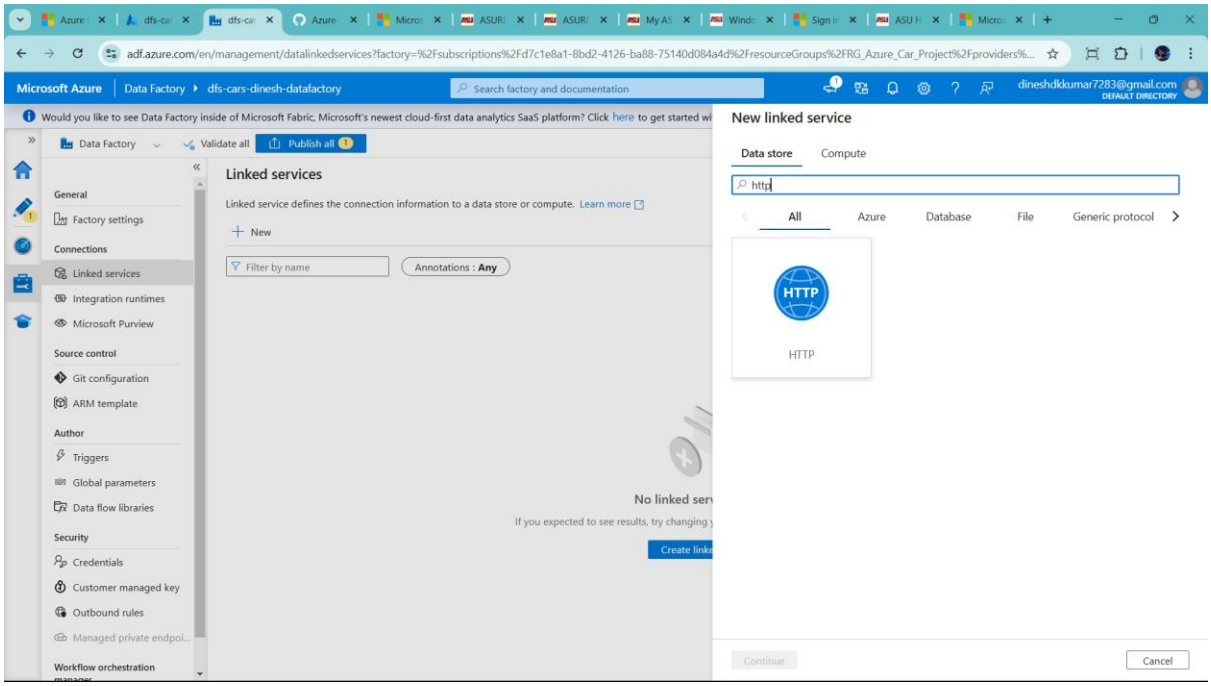
Another one



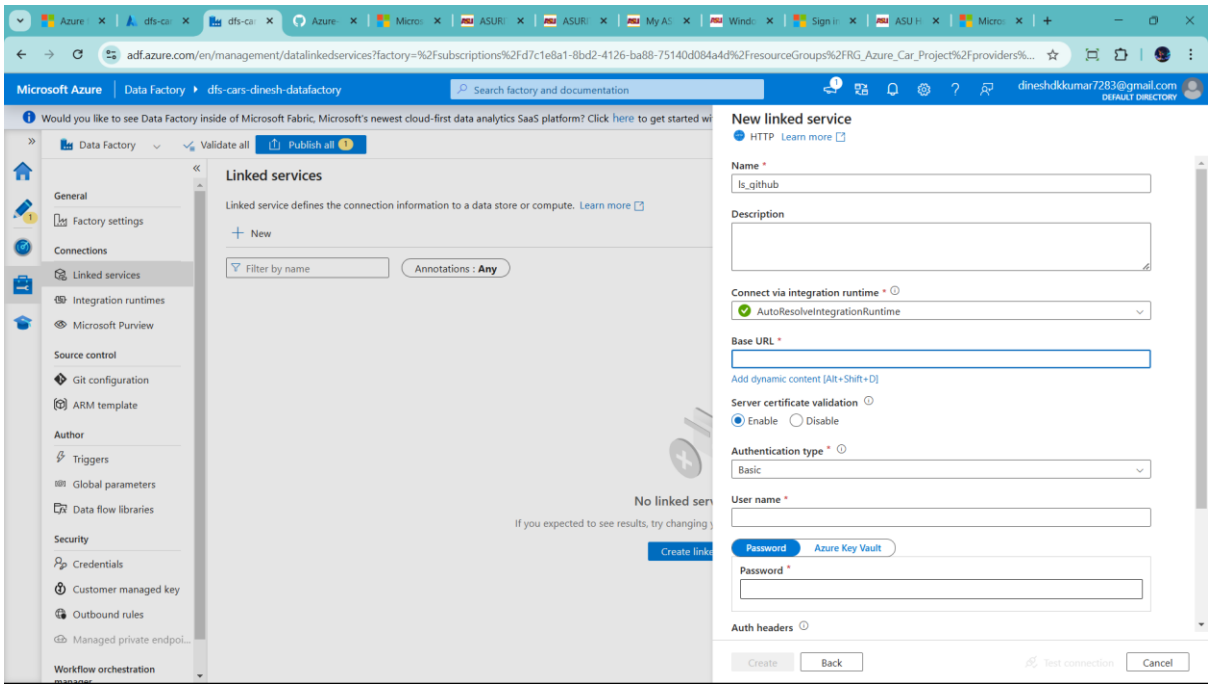
Go to manage tab →click linked service → click + new



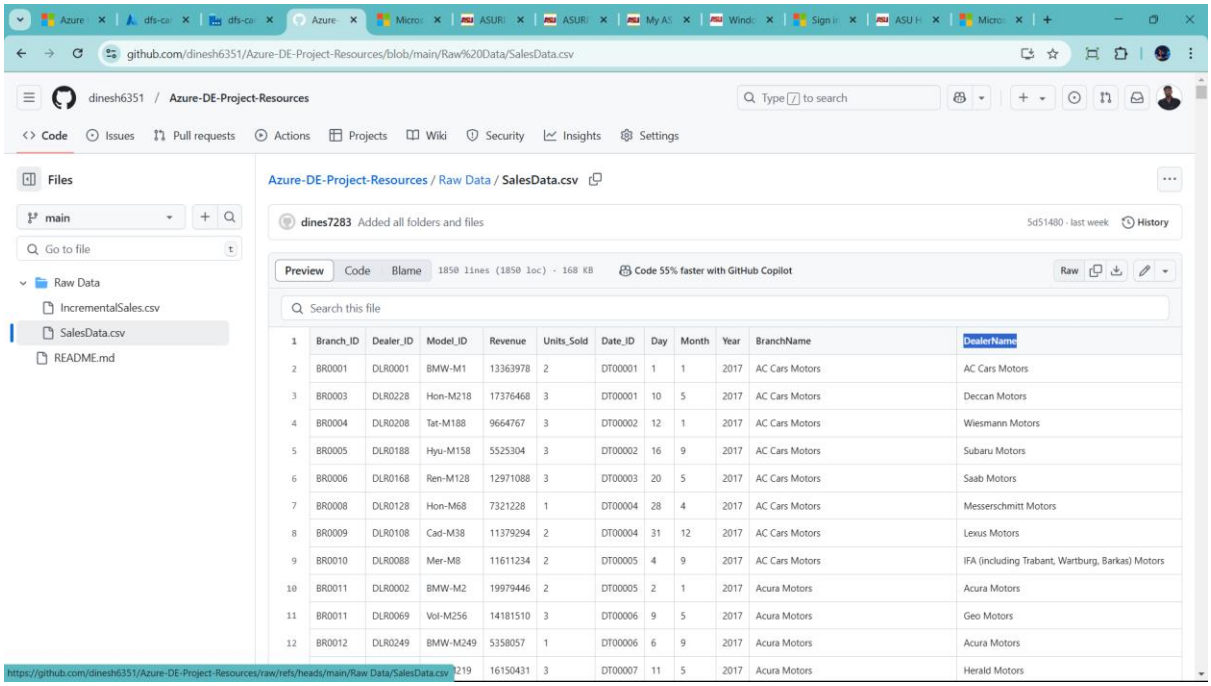
Search http →for github resporitory



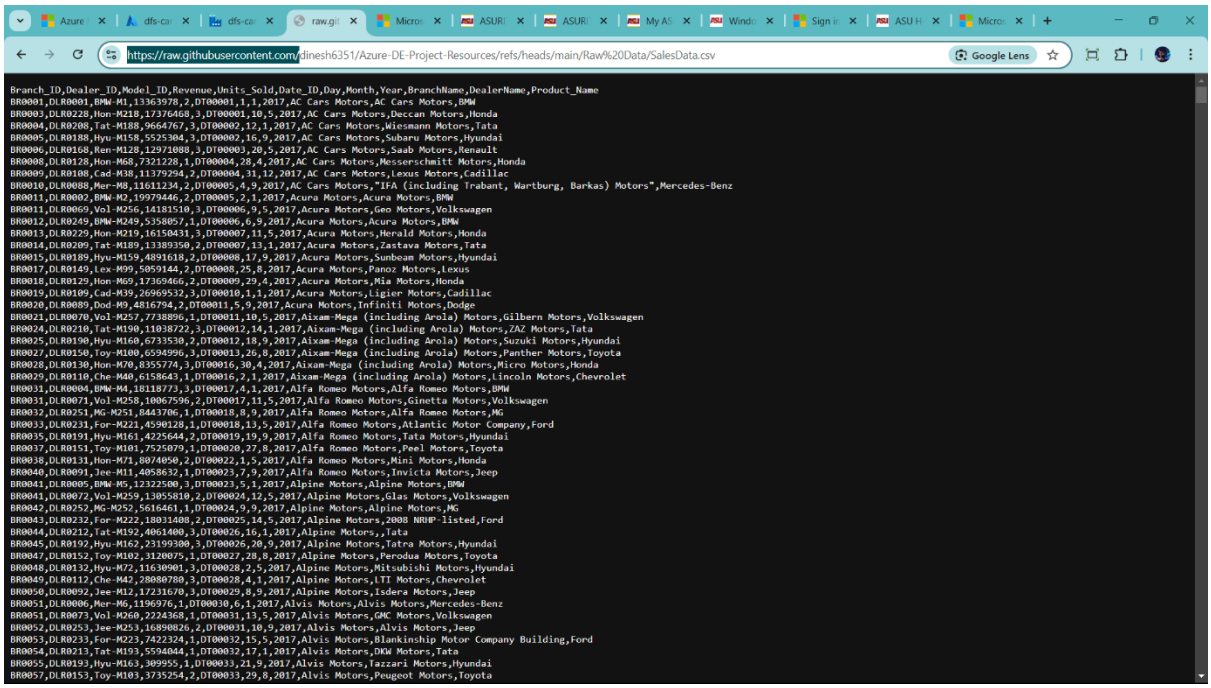
Click http → add name ls_github → add base url (find raw url in github)



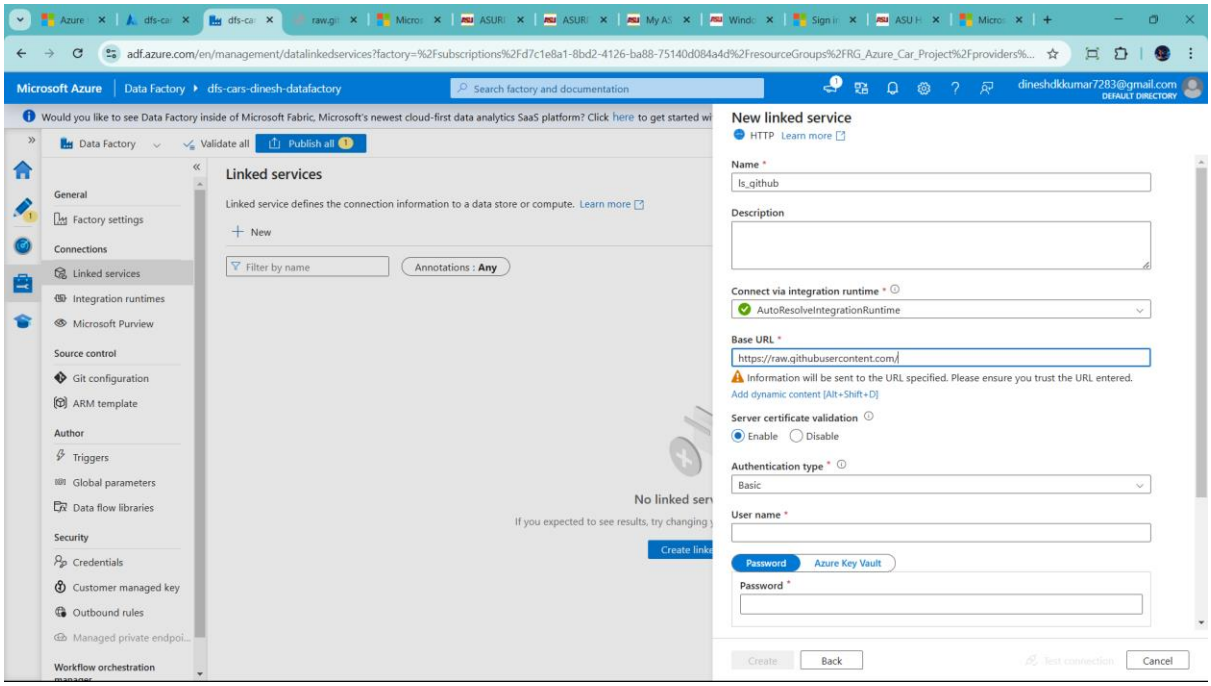
In github



Click raw

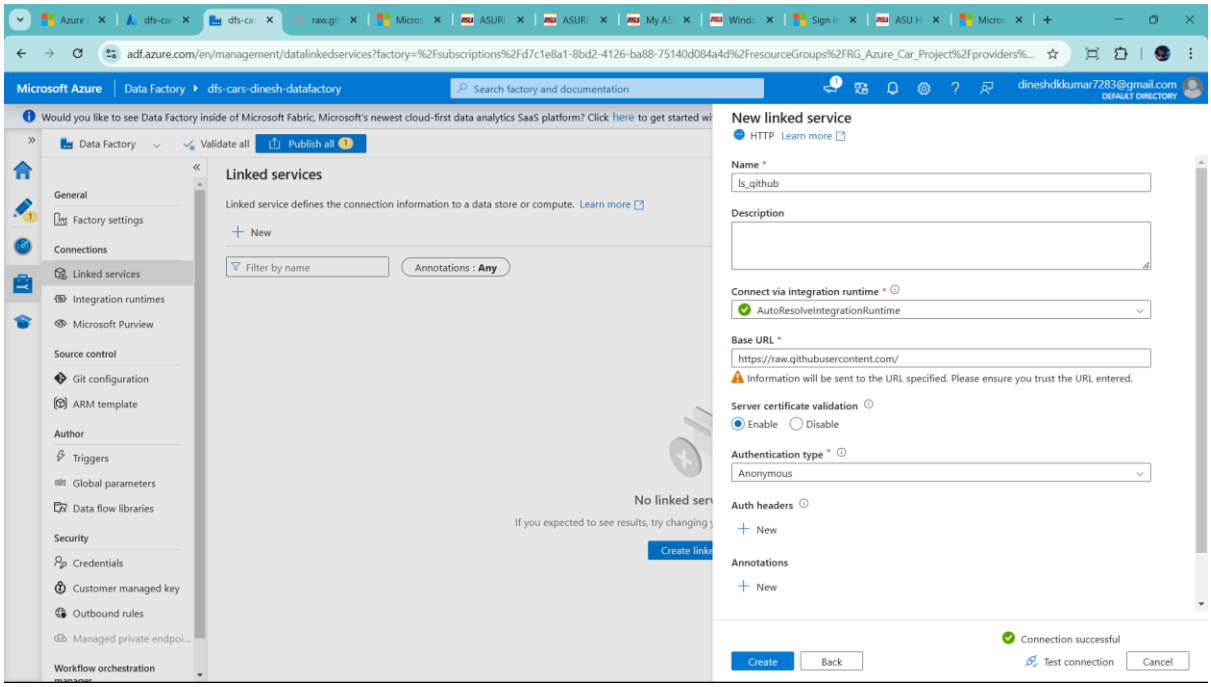


Than copy that url add in bases url in add link service

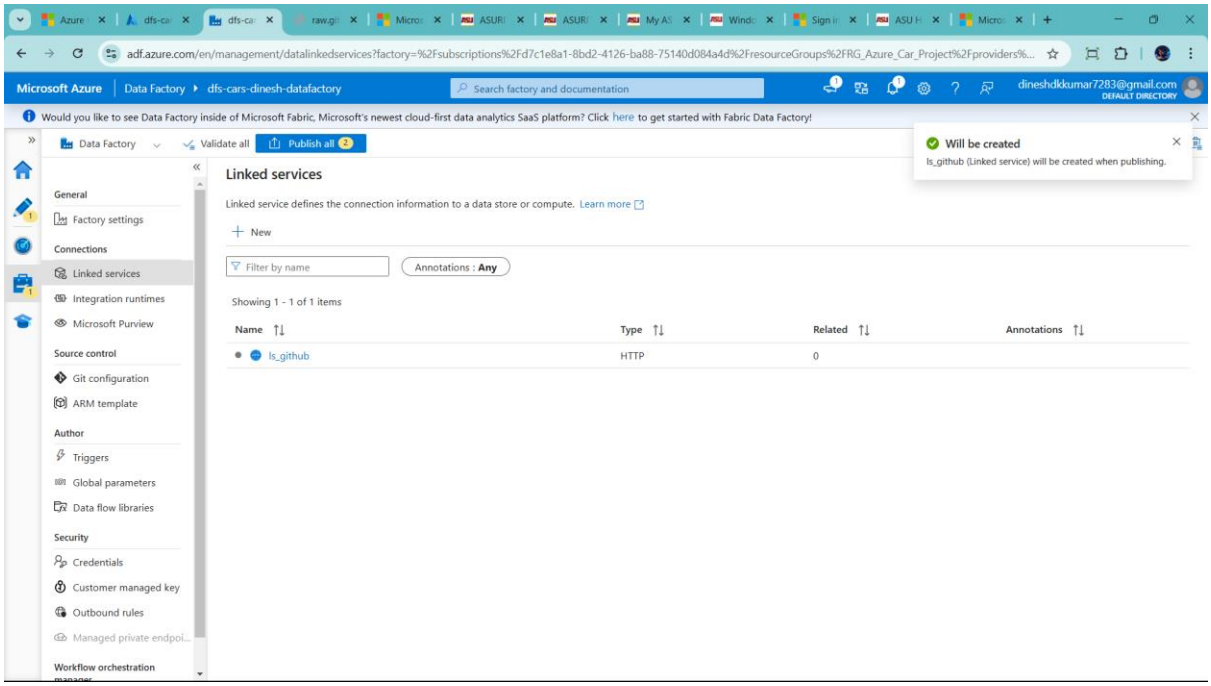


Authentication type chage Anonymos

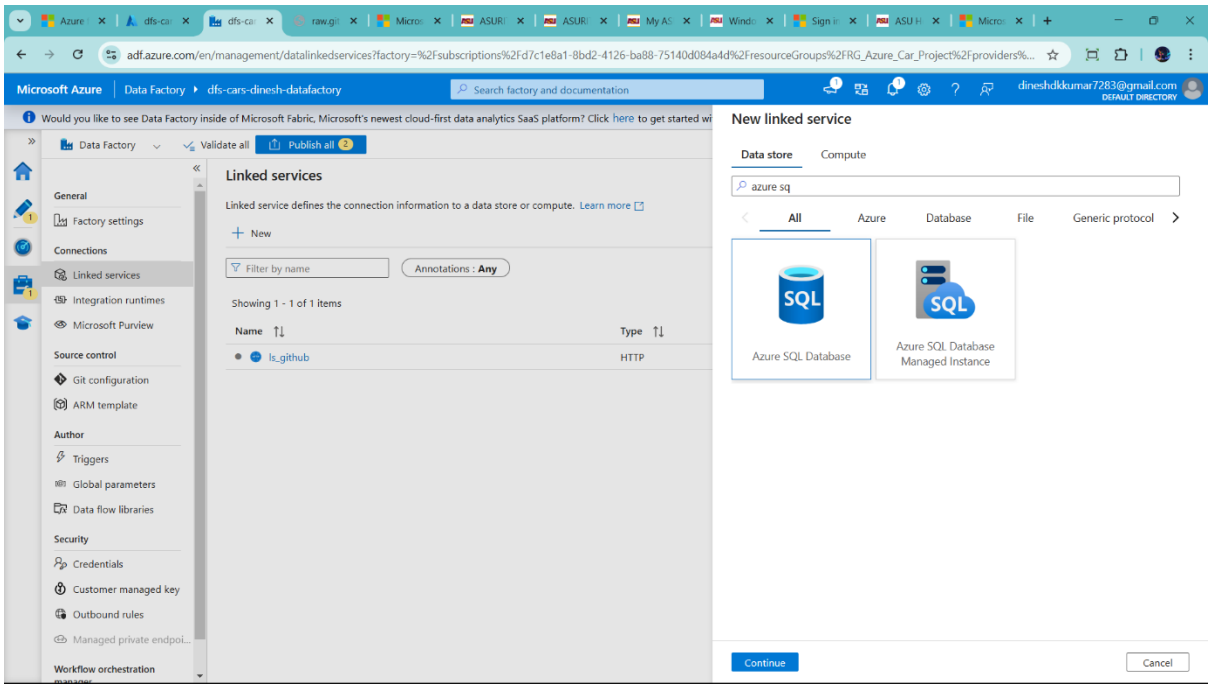
And test the connection



Click Create

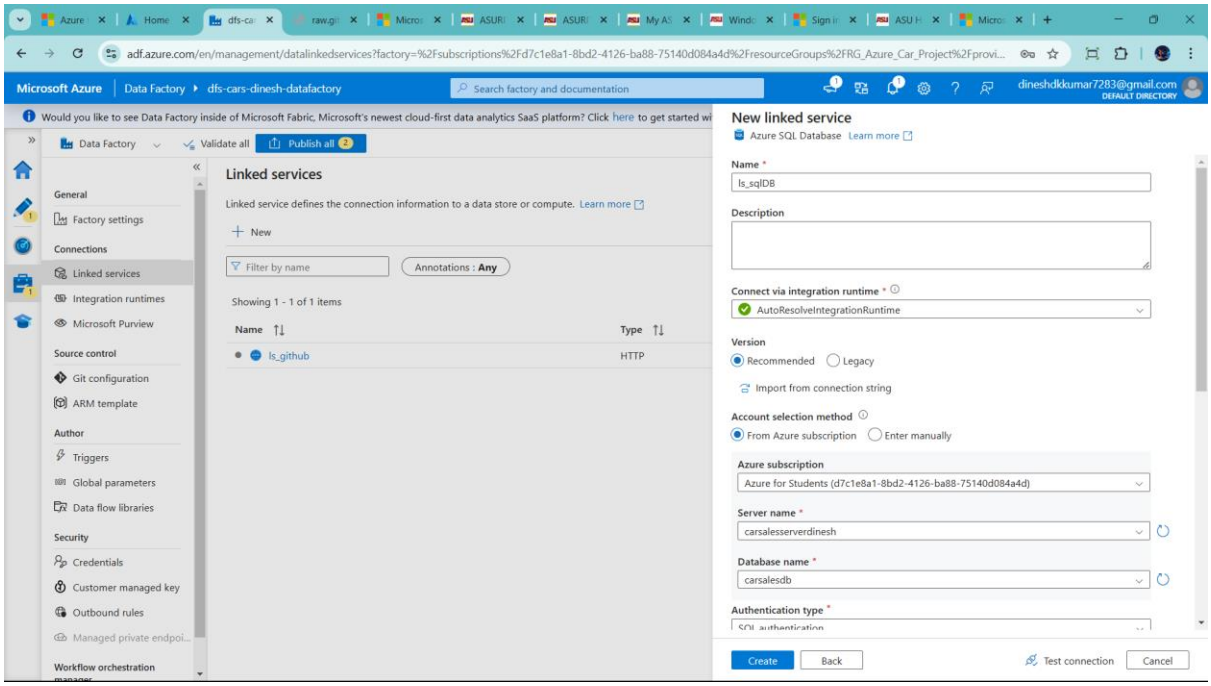


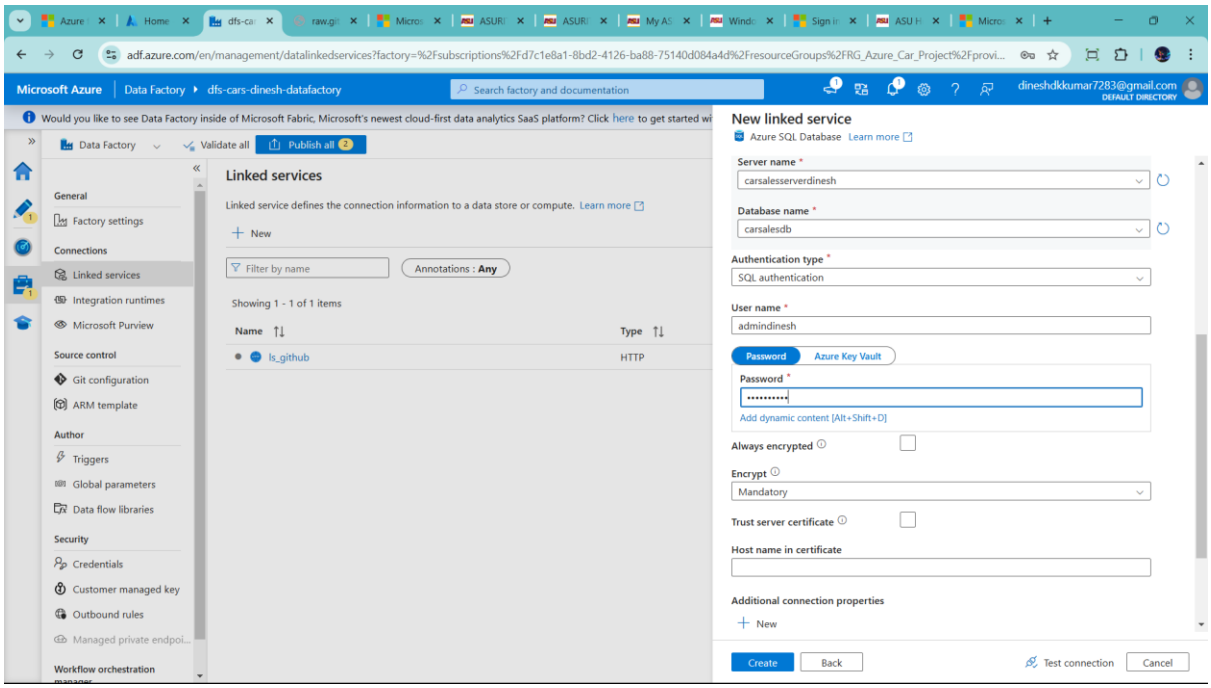
Now create another connection to push data in azure sql db



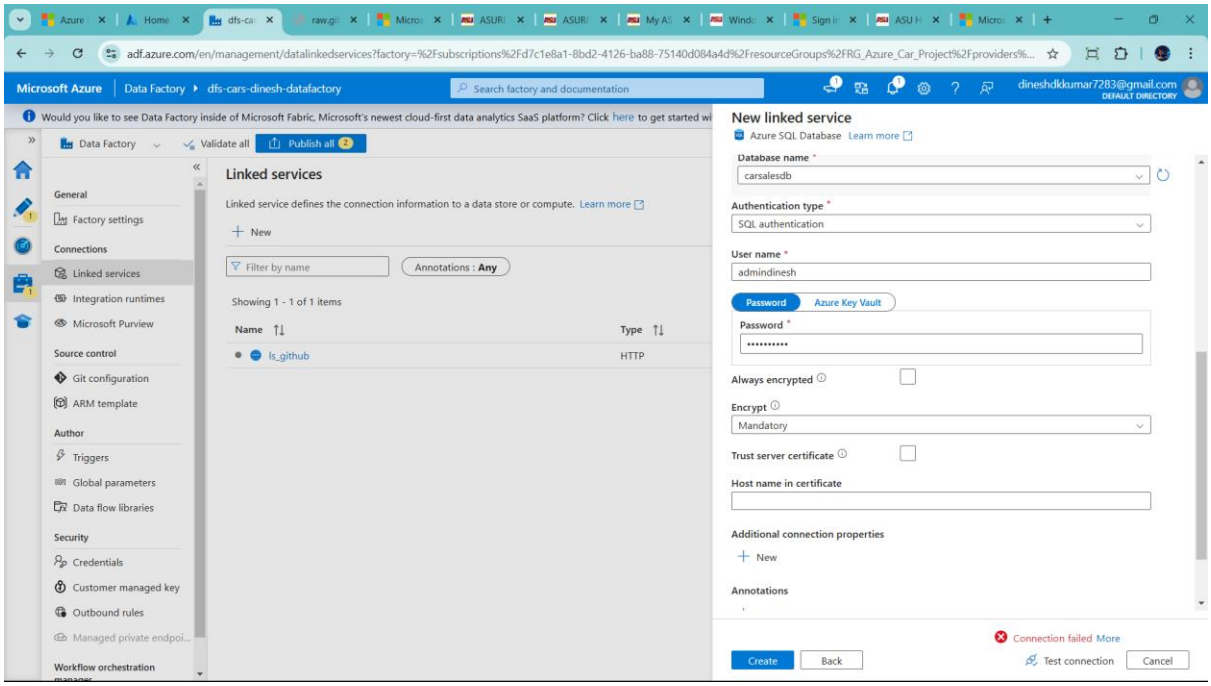
Add name ls_sqlDB → change azure subscription →azure of student (like) →add server name →

Add databases name and password (already created)


adf.azure.com/en/management/datalinkedservices?factory=%2Fsubscriptions%2Fd7c1e8a1-8bd2-4126-ba88-75140d084a4d%2FresourceGroups%2FRG_Azure_Car_Project%2Fprovi...
Microsoft Azure
Data Factory
dfs-cars-dinesh-datafactory
Search factory and documentation
dineshdkkumar7283@gmail.com
DEFAULT DIRECTORY
Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get started wi
Data Factory
Validate all
Publish all 2
General
Factory settings
Connections
Linked services
Integration runtimes
Microsoft Purview
Source control
Git configuration
ARM template
Author
Triggers
Global parameters
Data flow libraries
Security
Credentials
Customer managed key
Outbound rules
Managed private endpoi...
Workflow orchestration
Linked services
Linked service defines the connection information to a data store or compute. Learn more
New
Filter by name
Annotations : Any
Showing 1 - 1 of 1 items
Name
Type
ls_github
HTTP
New linked service
Azure SQL Database Learn more
Name *
ls_sqlDB
Description
Connect via integration runtime *
AutoResolveIntegrationRuntime
Version
Recommended
Legacy
Import from connection string
Account selection method
From Azure subscription
Enter manually
Azure subscription
Azure for Students (d7c1e8a1-8bd2-4126-ba88-75140d084a4d)
Server name *
carsalesserverdinesh
Database name *
carsalesdb
Authentication type *
Create
Back
Test connection
Cancel

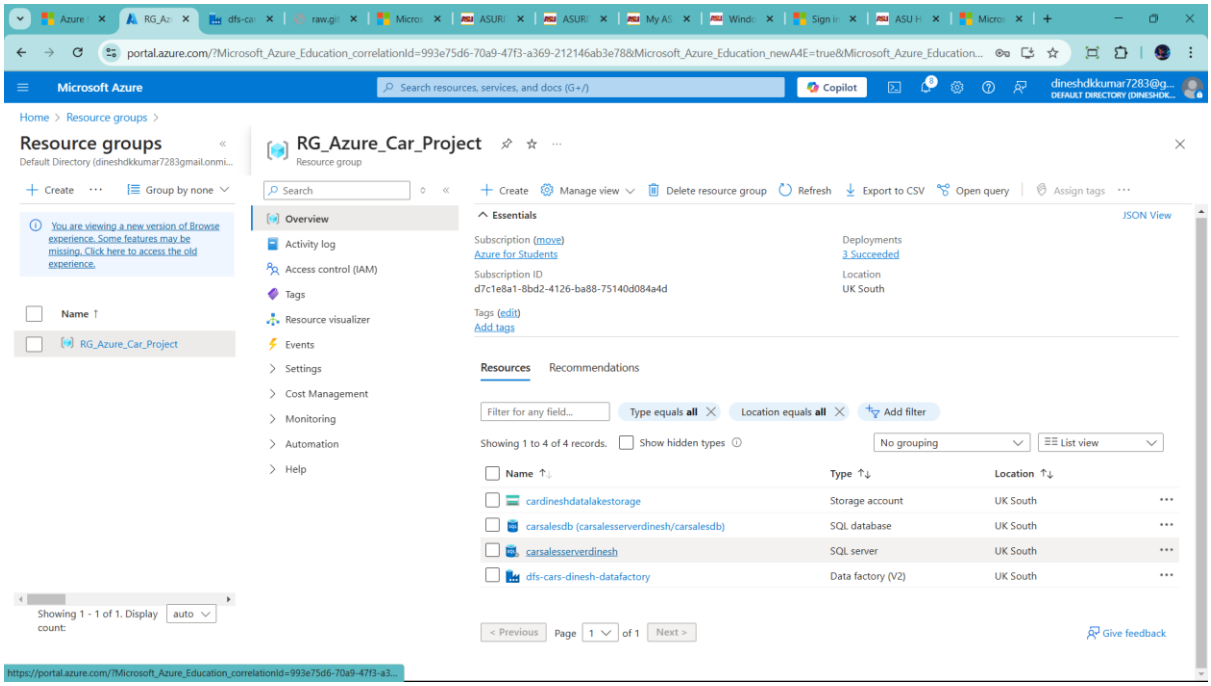

adf.azure.com/en/management/datalinkedservices?factory=%2Fsubscriptions%2Fd7c1e8a1-8bd2-4126-ba88-75140d084a4d%2FresourceGroups%2FRG_Azure_Car_Project%2Fprovi...
Microsoft Azure
Data Factory
dfs-cars-dinesh-datafactory
Search factory and documentation
dineshdkkumar7283@gmail.com
DEFAULT DIRECTORY
Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get started wi
Data Factory
Validate all
Publish all 2
General
Factory settings
Connections
Linked services
Integration runtimes
Microsoft Purview
Source control
Git configuration
ARM template
Author
Triggers
Global parameters
Data flow libraries
Security
Credentials
Customer managed key
Outbound rules
Managed private endpoi...
Workflow orchestration
Linked services
Linked service defines the connection information to a data store or compute. Learn more
New
Filter by name
Annotations : Any
Showing 1 - 1 of 1 items
Name
Type
ls_github
HTTP
New linked service
Azure SQL Database Learn more
Server name *
carsalesserverdinesh
Database name *
carsalesdb
Authentication type *
SQL authentication
User name *
admindinesh
Password
Azure Key Vault
Password *
Add dynamic content [Alt+Shift+D]
Always encrypted
Encrypt
Mandatory
Trust server certificate
Host name in certificate
Additional connection properties
New
Create
Back
Test connection
Cancel

Test the connect show error → the give some security add in sqlserver

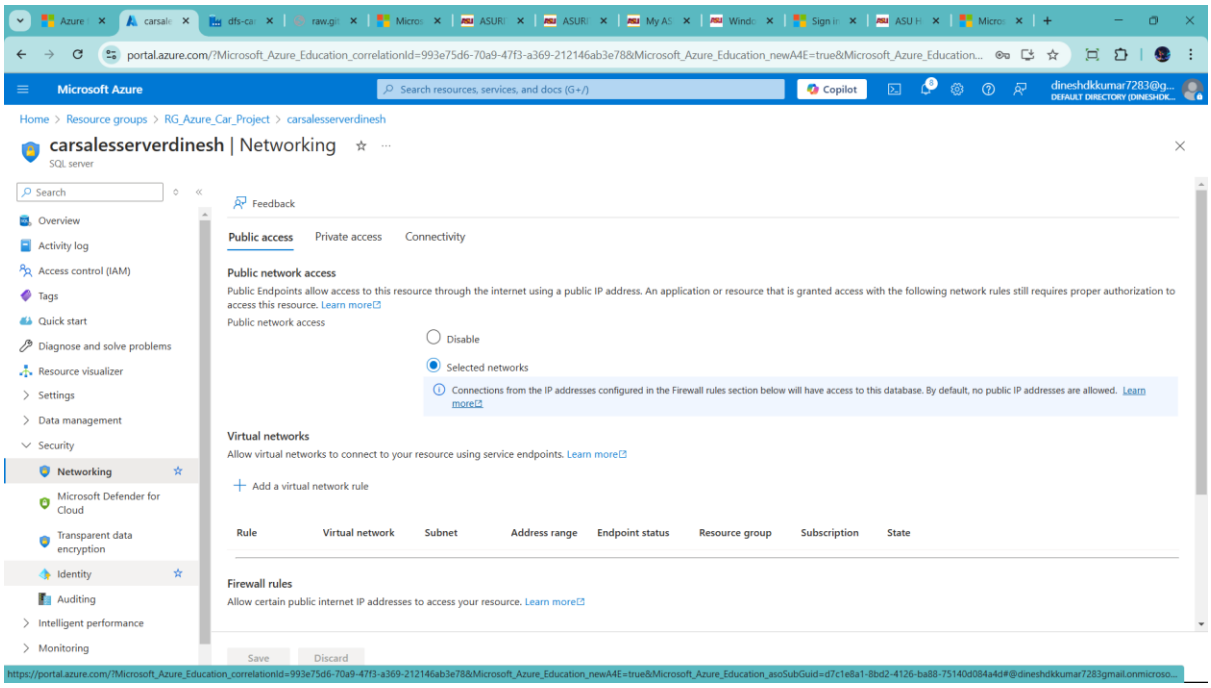


Let change it

Click server

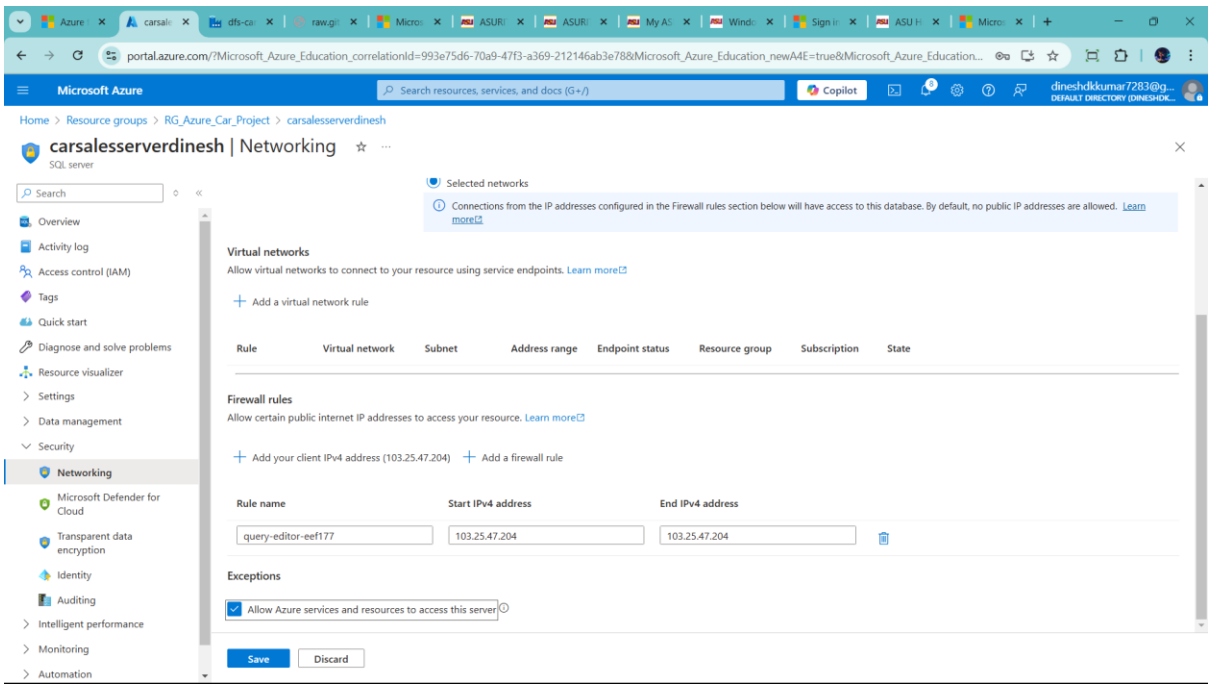


They show error because it is only selected networks

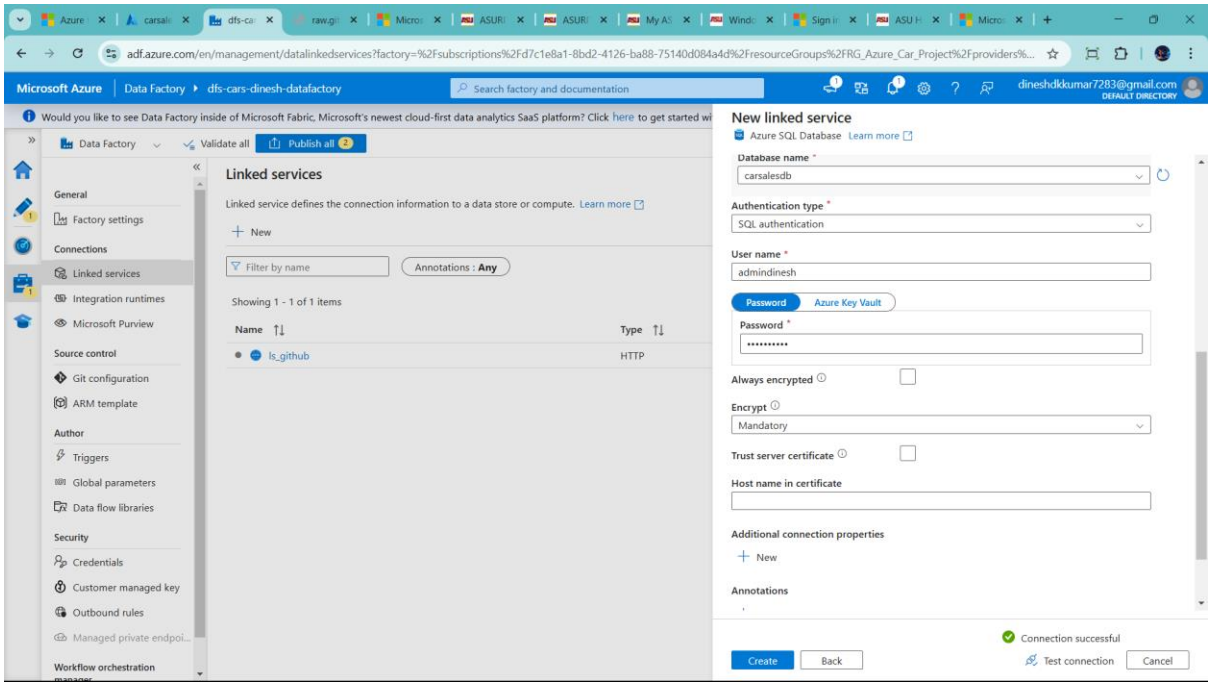


So when click allow all the network ip address. So check the box in exception

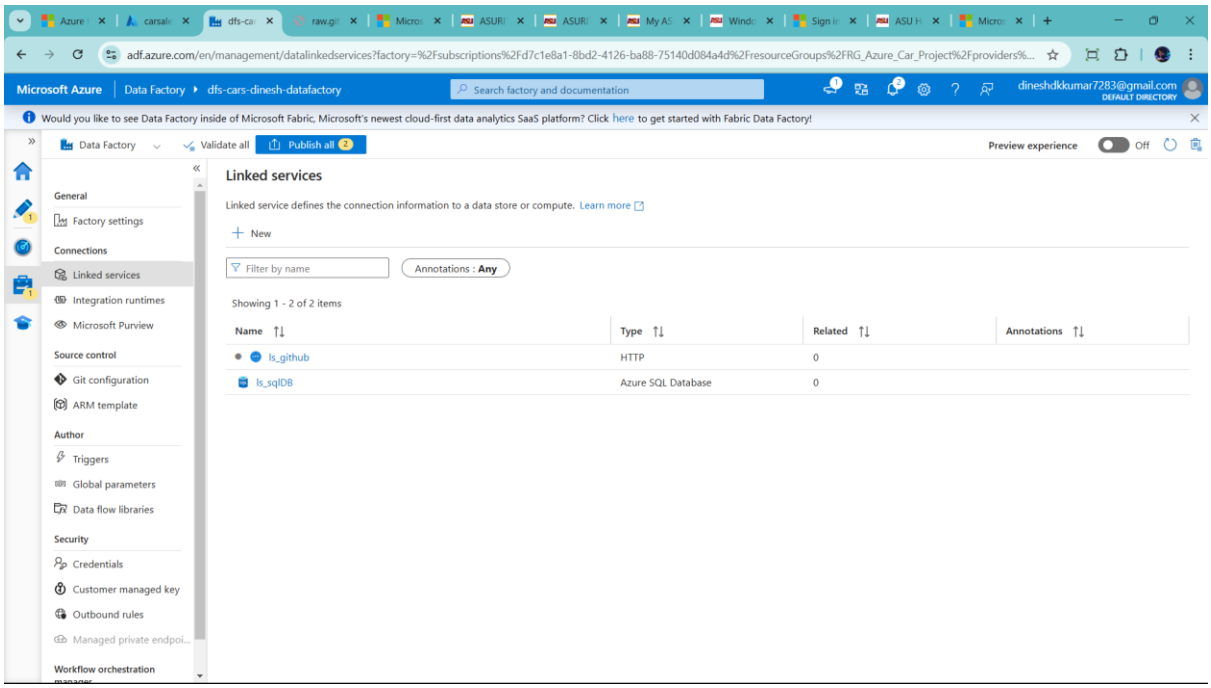
And save it



Now success full



Now create two activity in this pipline

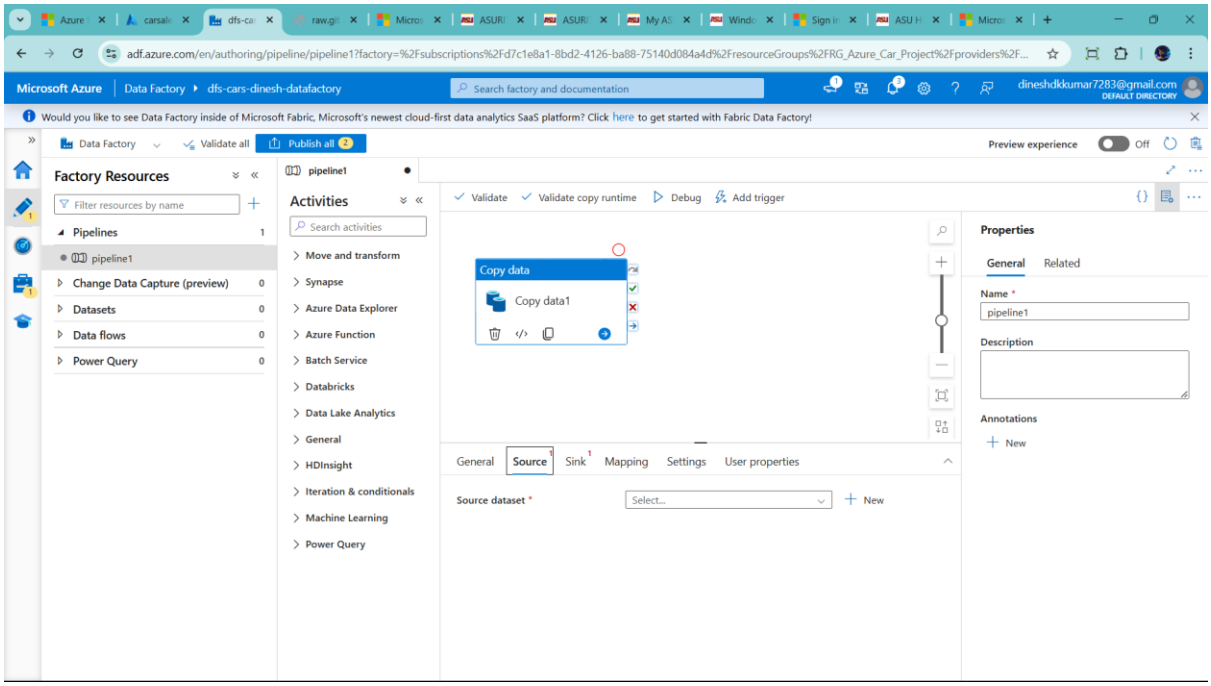


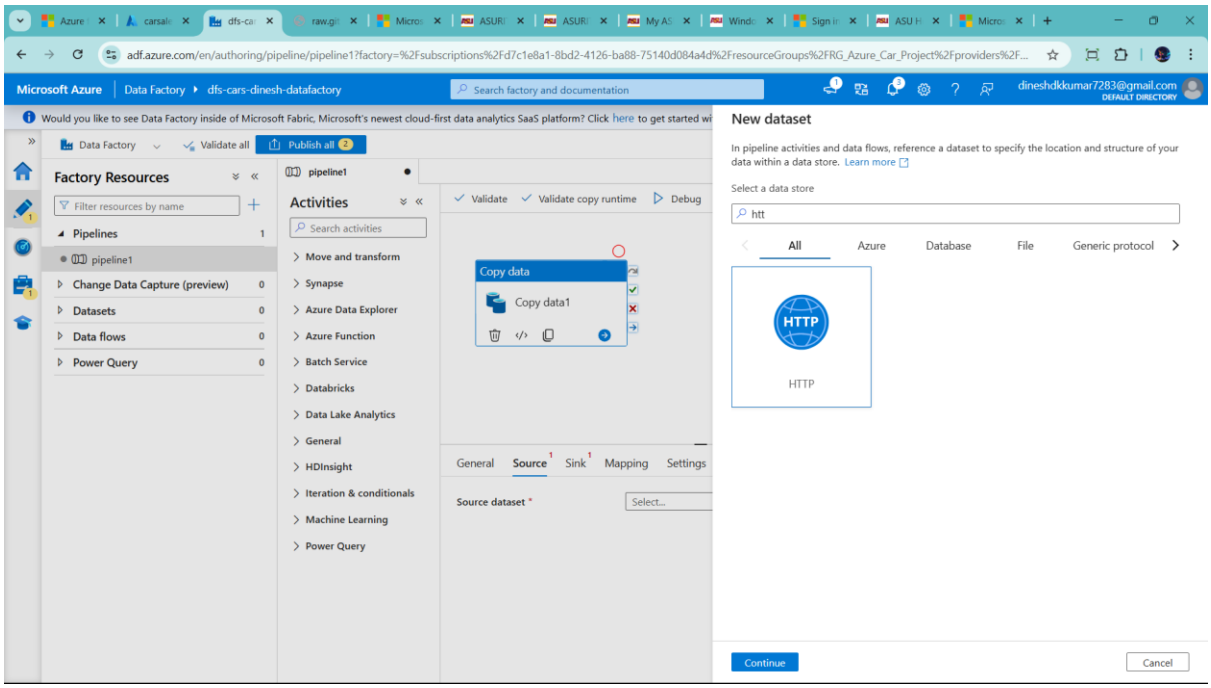
Now click Author to create pipeline

First click the copy data → then click source

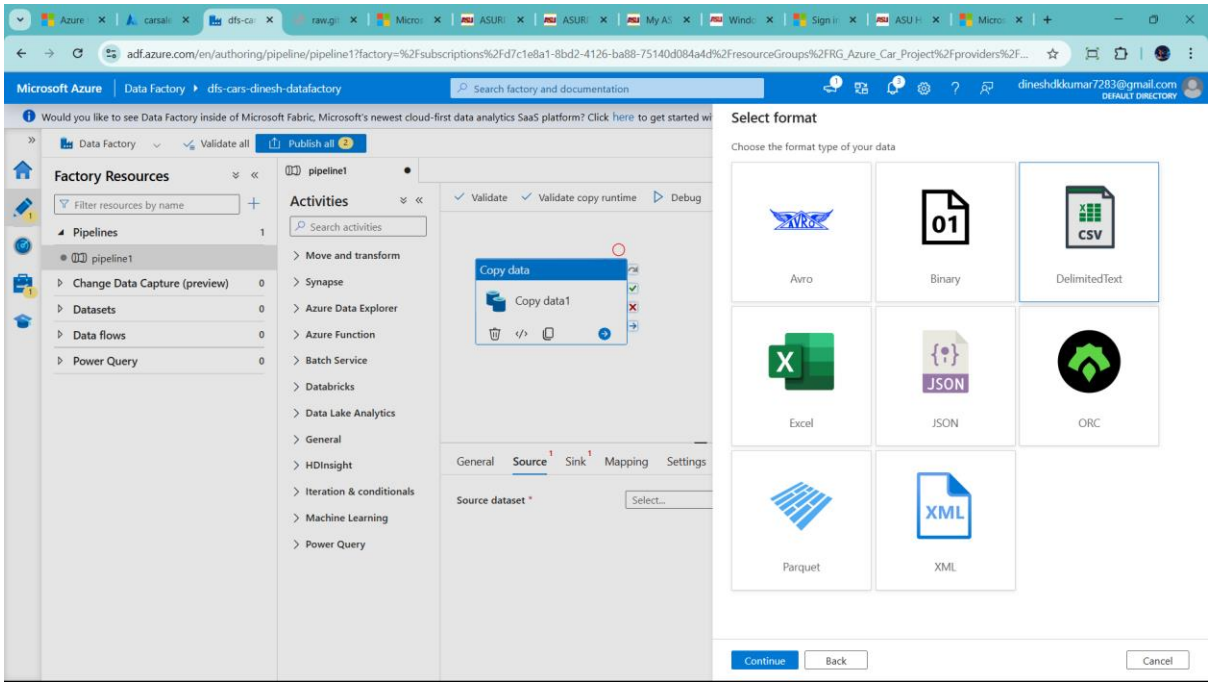
Let create parameter dataset because it is dynmanic dataset

So click +new


Microsoft Azure
Data Factory
dfs-cars-dinesh-datafactory
Search factory and documentation
dineshdkkumar7283@gmail.com
DEFAULT DIRECTORY
Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get started with Fabric Data Factory!
Data Factory
Validate all
Publish all
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines
pipeline1
Change Data Capture (preview)
Datasets
Data flows
Power Query
pipeline1
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Copy data
Copy data1
Properties
General
Related
Name *
pipeline1
Description
Annotations
New
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
Select...
New

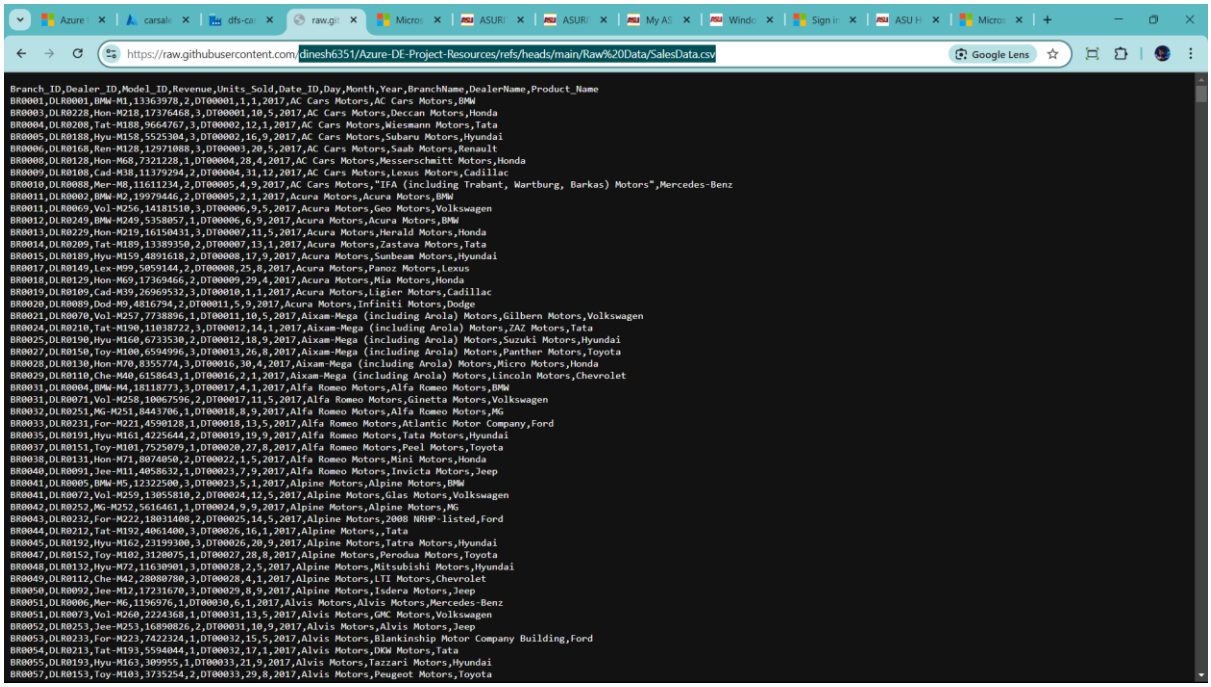

Microsoft Azure
Data Factory
dfs-cars-dinesh-datafactory
Search factory and documentation
dineshdkkumar7283@gmail.com
DEFAULT DIRECTORY
Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get started
Data Factory
Validate all
Publish all
Factory Resources
Filter resources by name
Pipelines
pipeline1
Change Data Capture (preview)
Datasets
Data flows
Power Query
pipeline1
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Copy data
Copy data1
General
Source
Sink
Mapping
Settings
Source dataset *
Select...
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
htt
All
Azure
Database
File
Generic protocol
HTTP
HTTP
Continue
Cancel

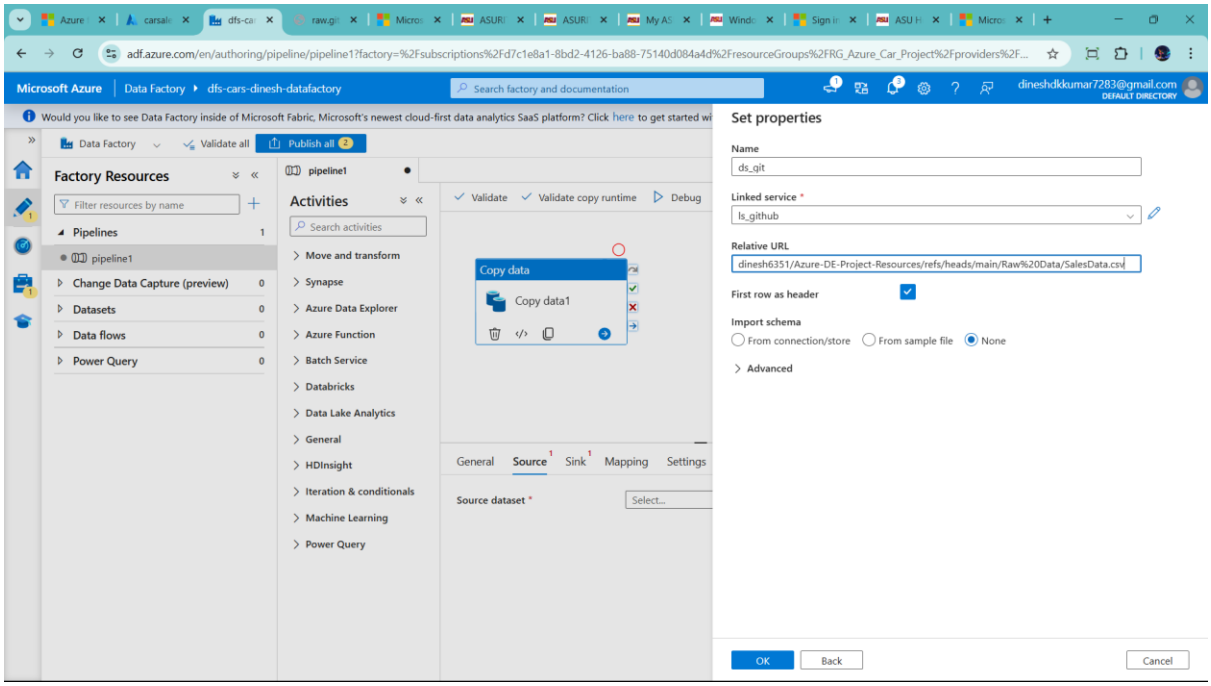
Csv



Now add name, add linked service,and releative url

This releative url

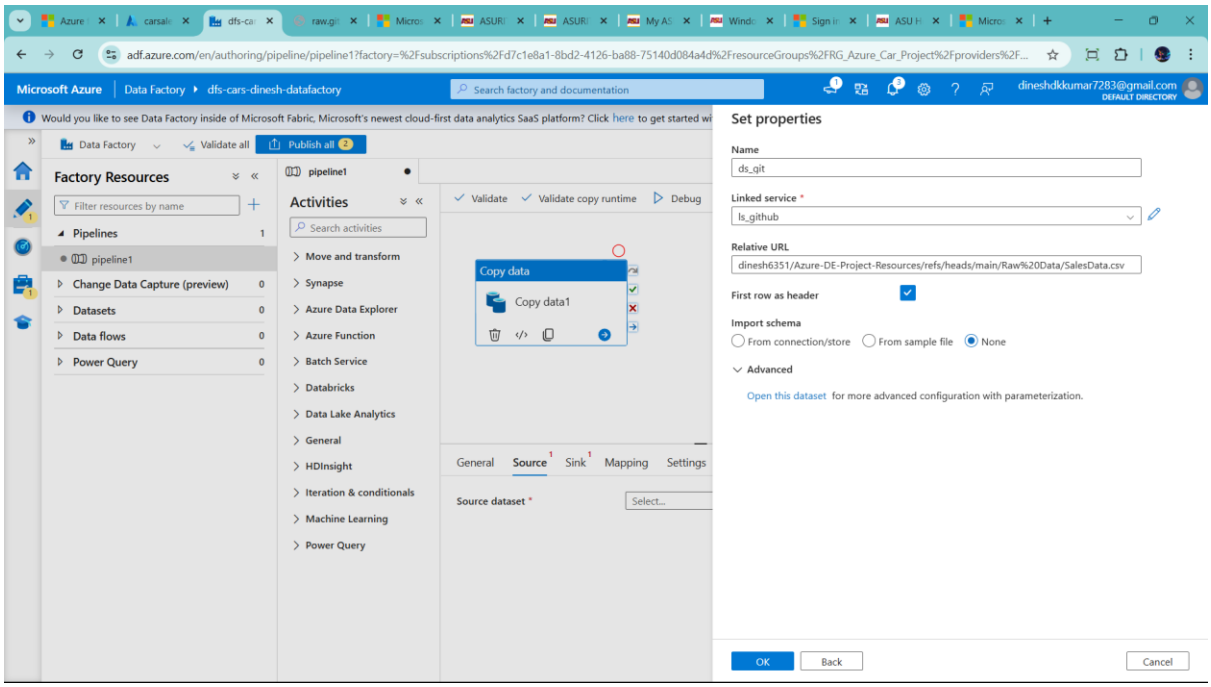




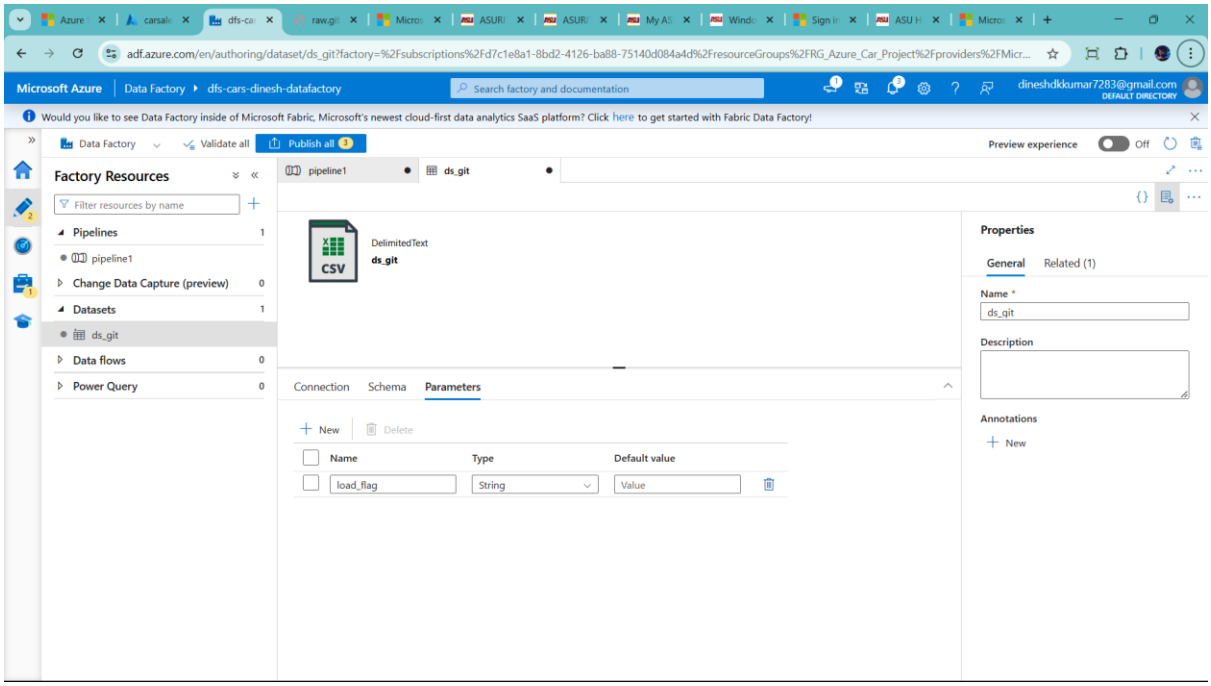
Because create dynamic to add csv file as incremental load

Now add parameter in this dataset

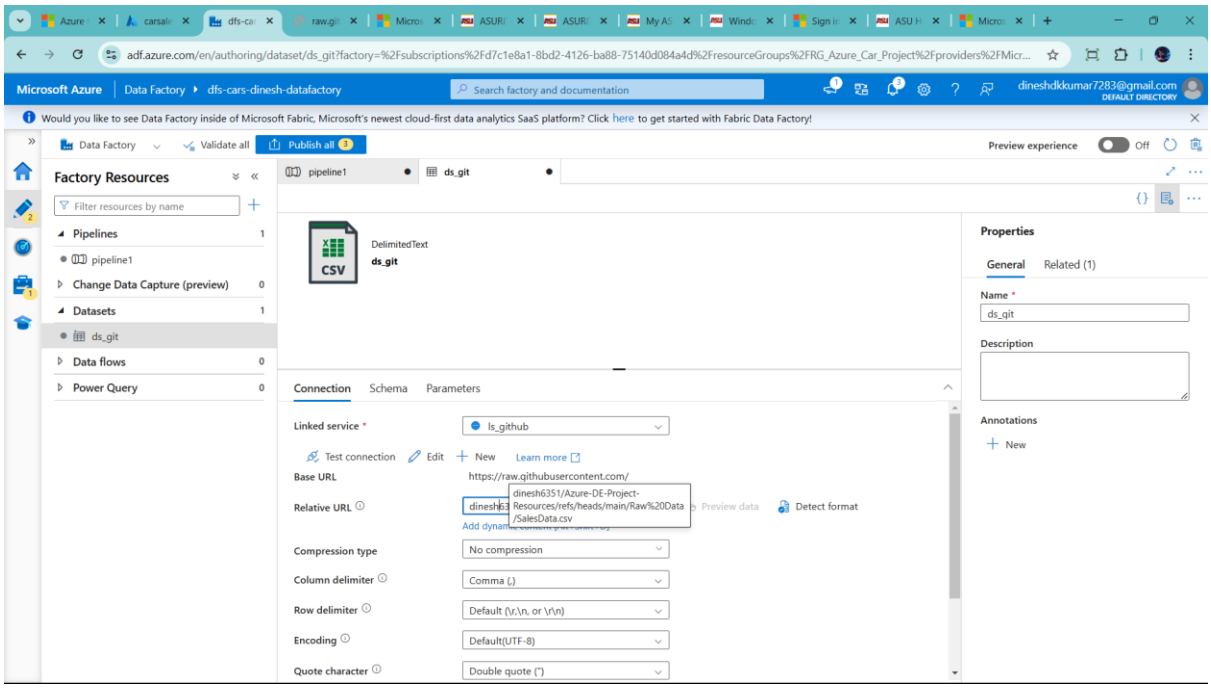
So ,click advance option and open this dataset

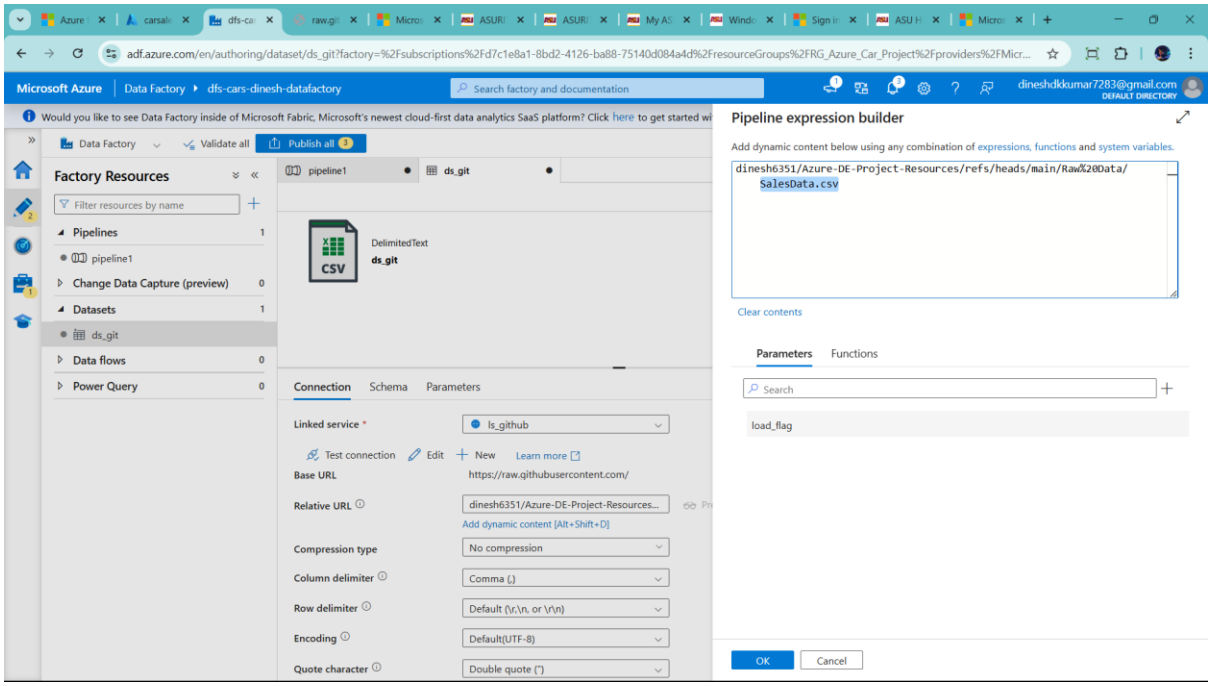


Than click parameter tap and click new

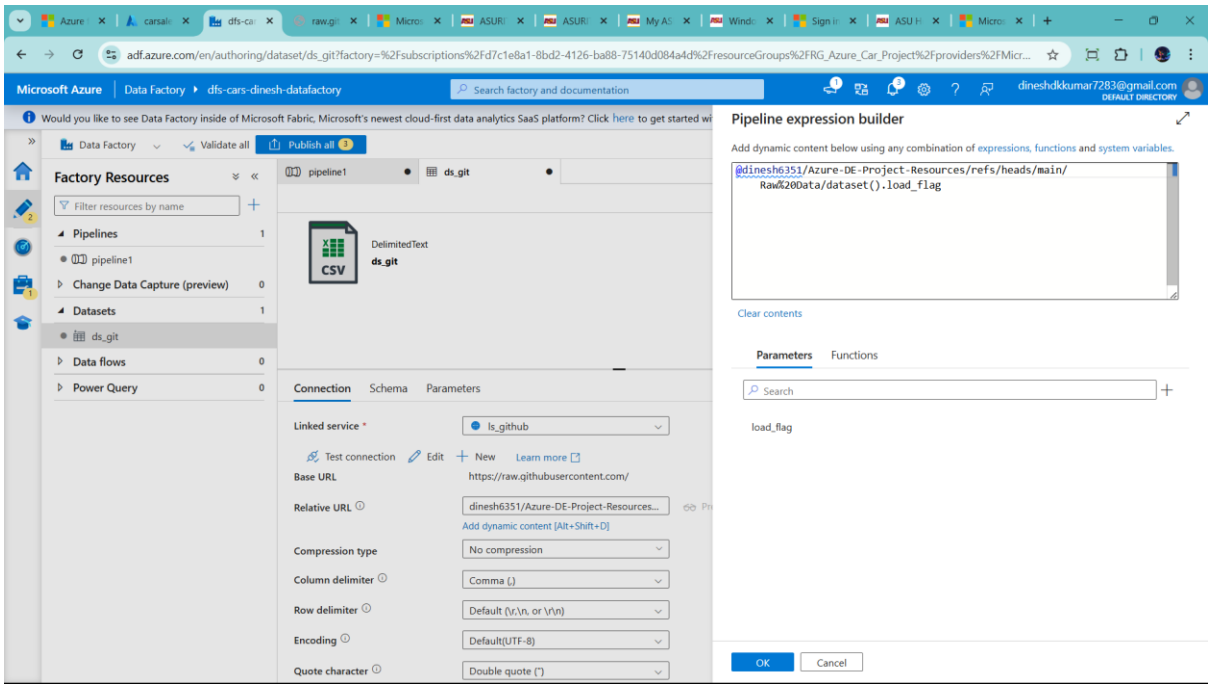


Go back to connection and click on releative url And clicl dynamic content show below

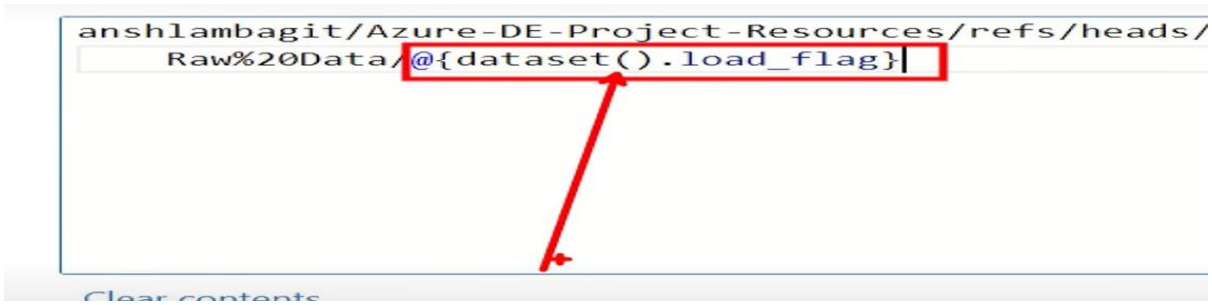


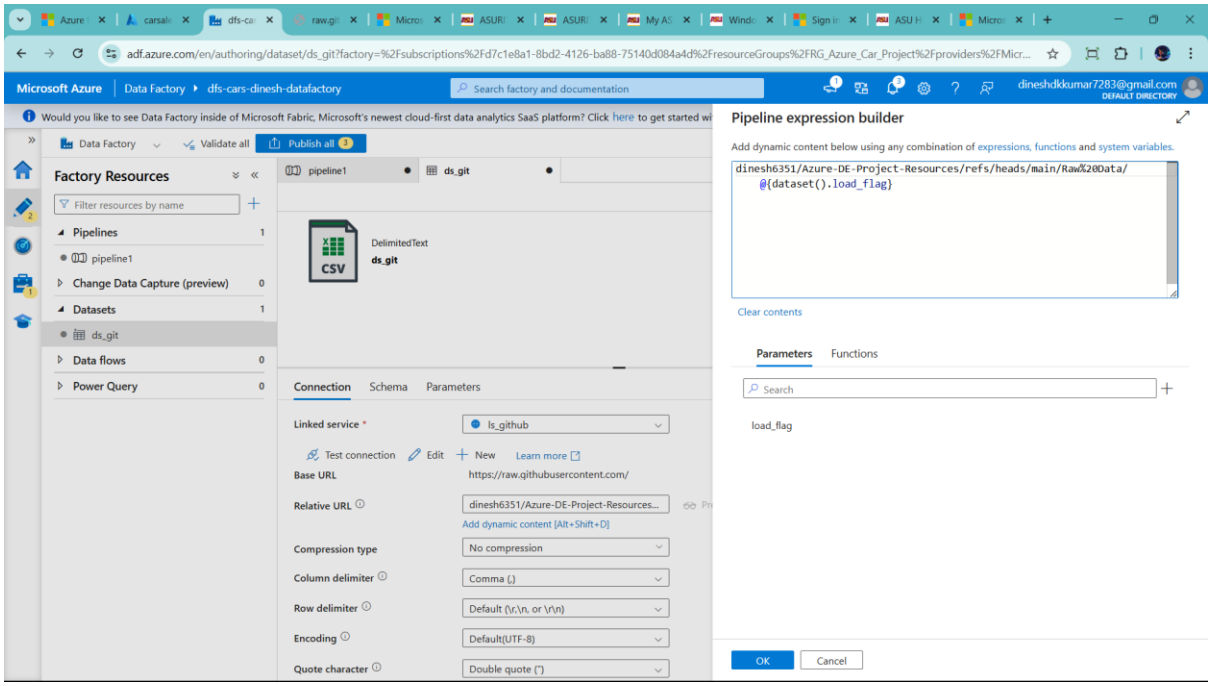


Select SalesData.csv =→click load_flag →

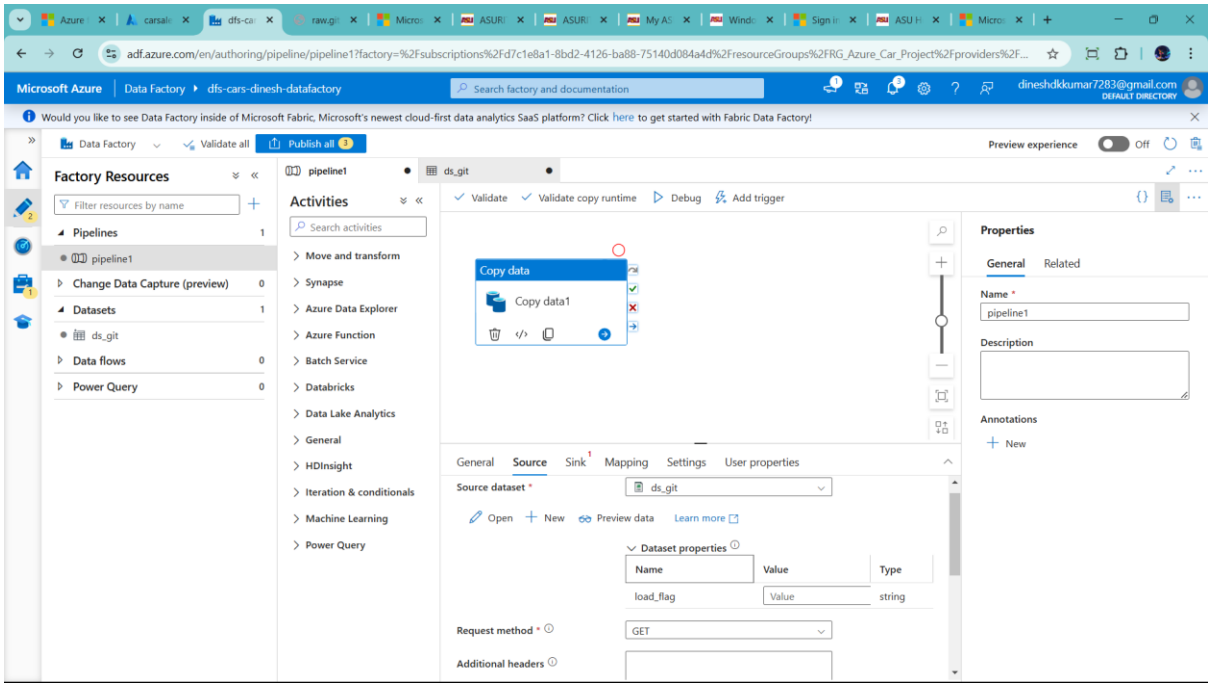


Now change string to varable



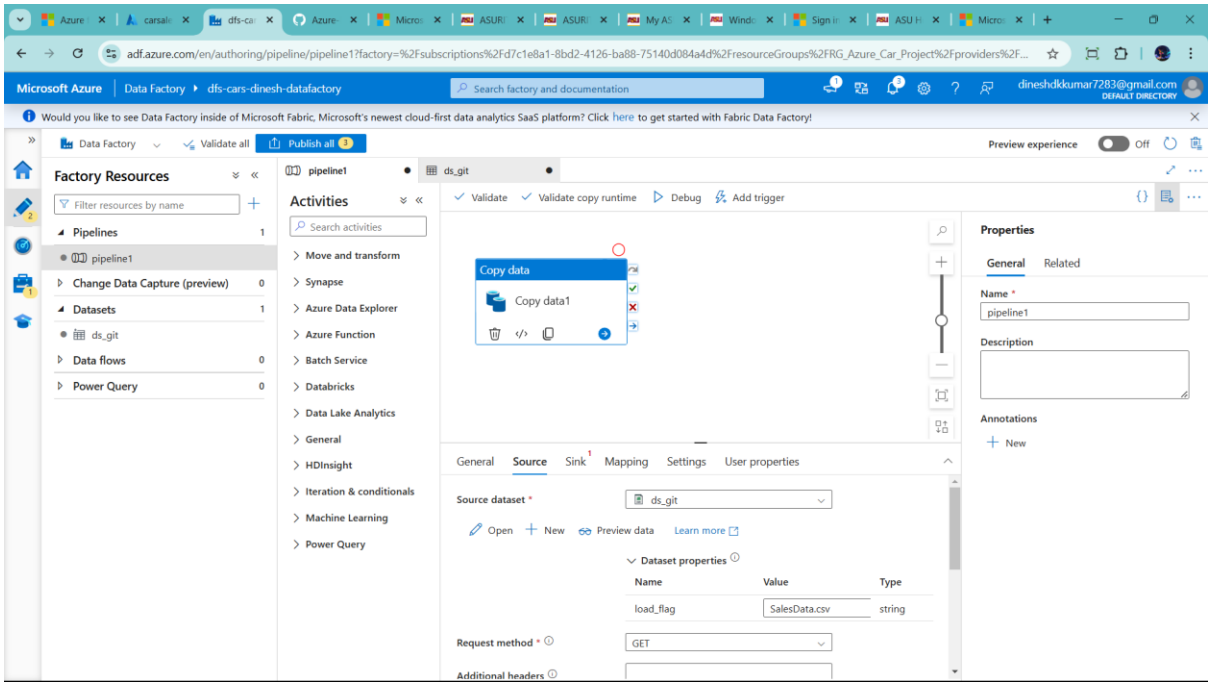


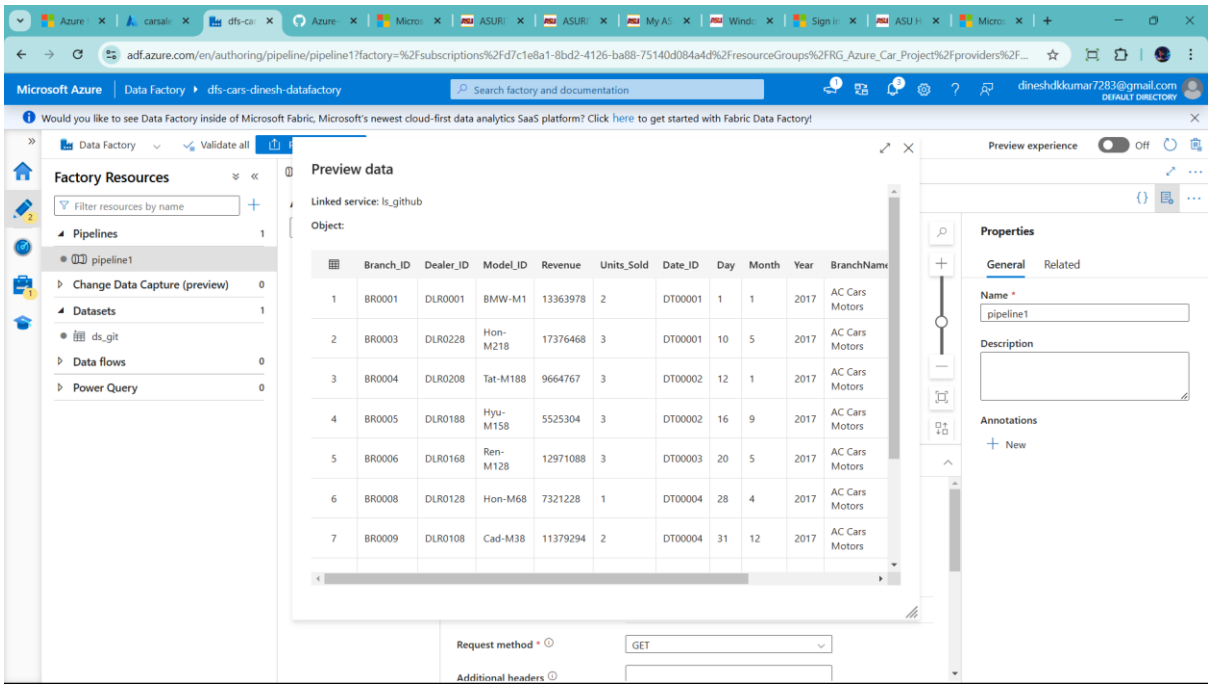
Now click ok and go to pipeline1/source_prep



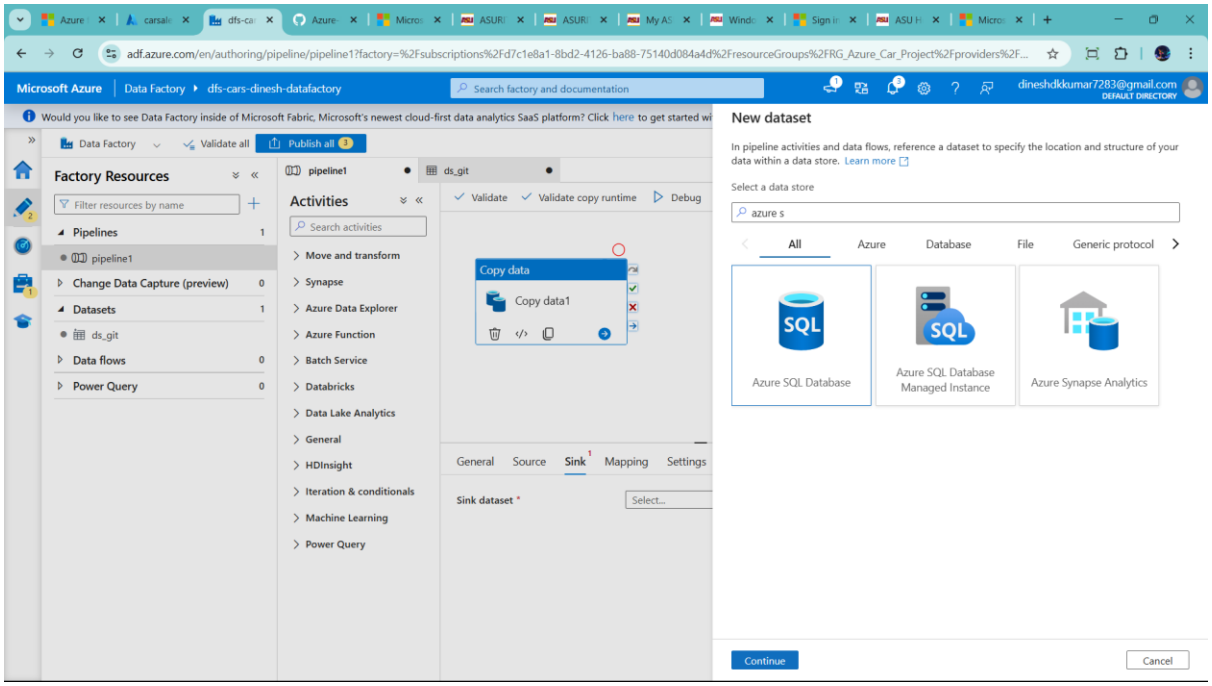
Ask to add value

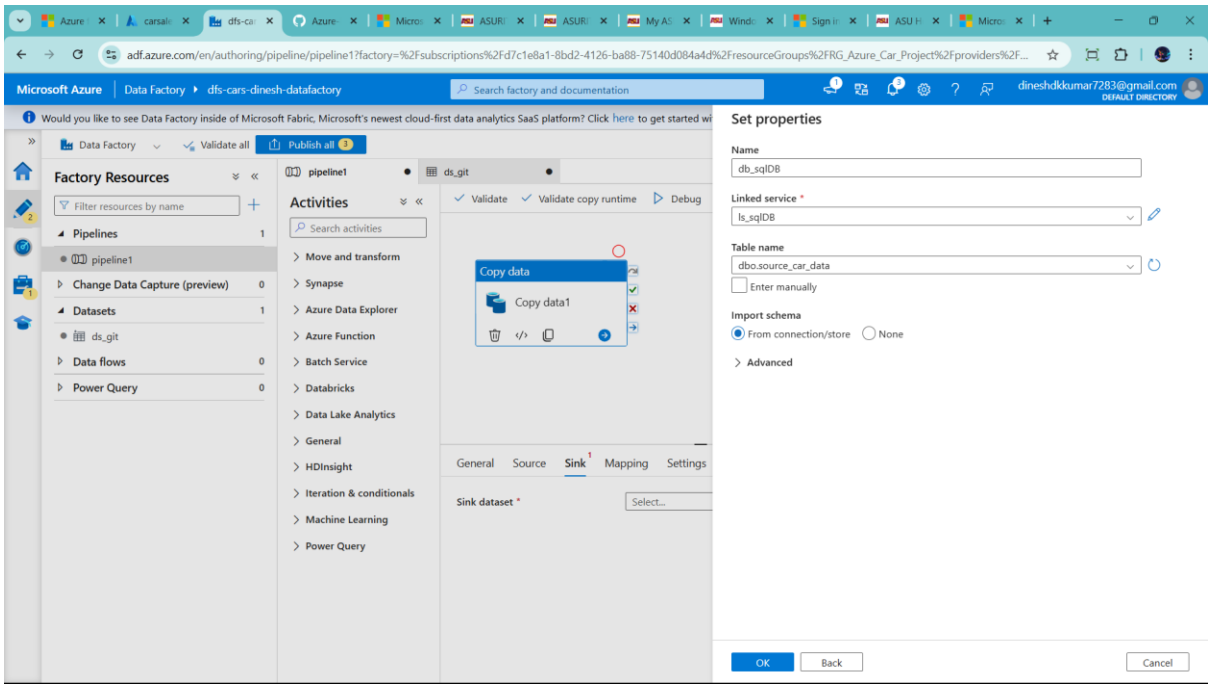
So I add SalesData.csv and click preview data



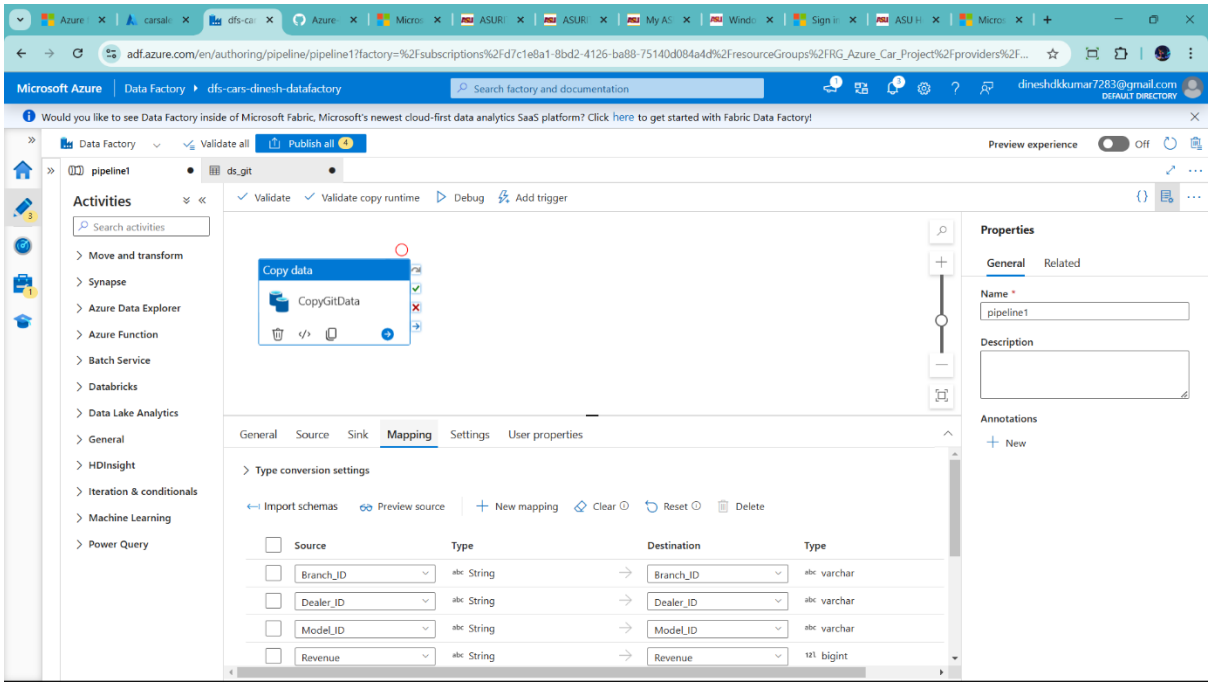


Now Click sink → click new →type azure sqldatabase →click ok

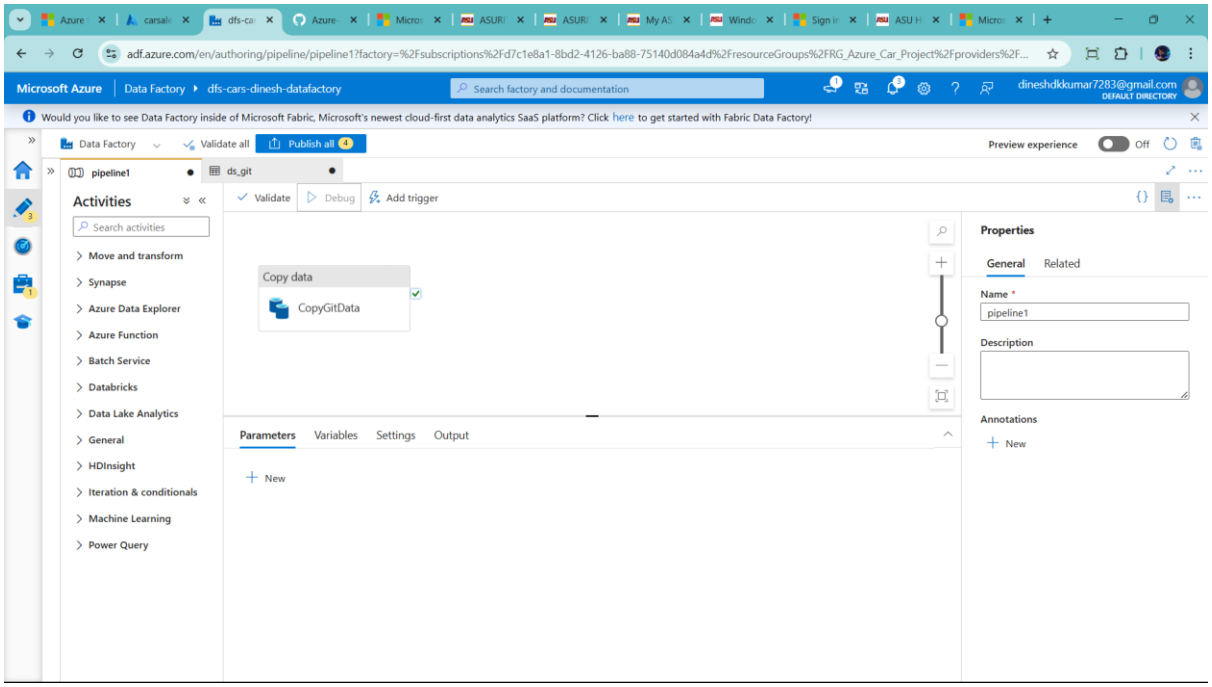

Microsoft Azure
Data Factory
dfs-cars-dinesh-datafactory
Search factory and documentation
dineshdkkumar7283@gmail.com
DEFAULT DIRECTORY
Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get started wi
Data Factory
Validate all
Publish all 3
Factory Resources
Filter resources by name
Pipelines 1
pipeline1
Change Data Capture (preview) 0
Datasets 1
ds_git
Data flows 0
Power Query 0
pipeline1
ds_git
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Copy data
Copy data1
General
Source
Sink 1
Mapping
Settings
Sink dataset *
Select...
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
azure s
All
Azure
Database
File
Generic protocol
Azure SQL Database
Azure SQL Database Managed Instance
Azure Synapse Analytics
Continue
Cancel


Microsoft Azure
Data Factory
dfs-cars-dinesh-datafactory
Search factory and documentation
dineshdkkumar7283@gmail.com
DEFAULT DIRECTORY
Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get started wi
Data Factory
Validate all
Publish all 3
Factory Resources
Filter resources by name
Pipelines 1
pipeline1
Change Data Capture (preview) 0
Datasets 1
ds_git
Data flows 0
Power Query 0
pipeline1
ds_git
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Copy data
Copy data1
General
Source
Sink 1
Mapping
Settings
Sink dataset *
Select...
Set properties
Name
db_sqlDB
Linked service *
ls_sqlDB
Table name
dbo.source_car_data
Enter manually
Import schema
From connection/store
None
Advanced
OK
Back
Cancel

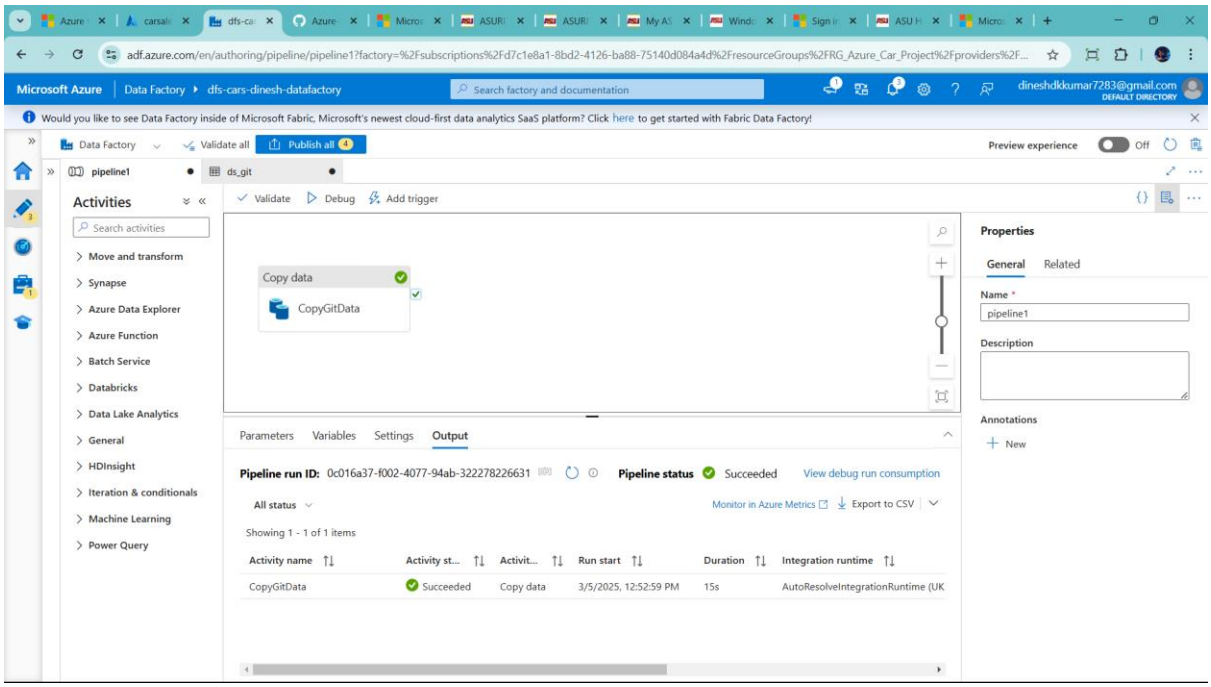
Now click mapping and import schema



Now load the data in sql db , so click Debug



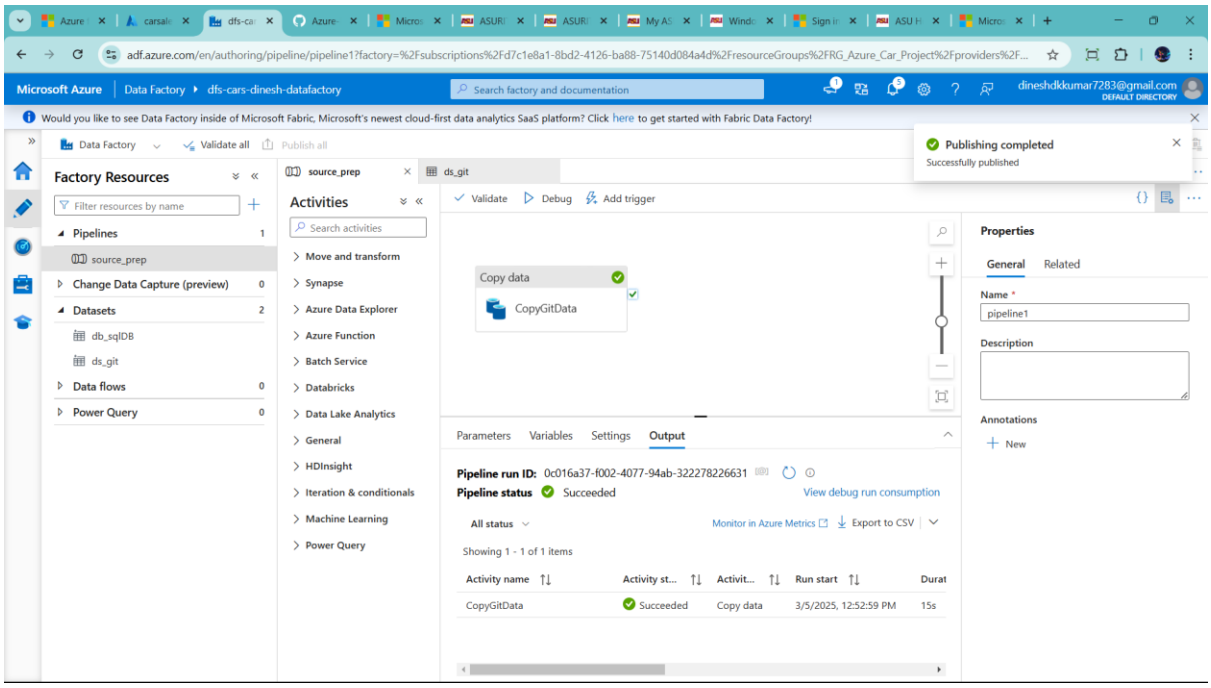
Success


Microsoft Azure
Data Factory
dfs-cars-dinesh-datafactory
Search factory and documentation
dineshdkkumar7283@gmail.com
DEFAULT DIRECTORY
Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get started with Fabric Data Factory!
Data Factory
Validate all
Publish all
Preview experience
Off
pipeline1
ds_git
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Debug
Add trigger
Copy data
CopyGitData
Properties
General
Related
Name *
pipeline1
Description
Annotations
New
Parameters
Variables
Settings
Output
Pipeline run ID: 0c016a37-f002-4077-94ab-322278226631
Pipeline status
Succeeded
View debug run consumption
All status
Monitor in Azure Metrics
Export to CSV
Showing 1 - 1 of 1 items
Activity name
Activity st...
Activit...
Run start
Duration
Integration runtime
CopyGitData
Succeeded
Copy data
3/5/2025, 12:52:59 PM
15s
AutoResolveIntegrationRuntime (UK

Dynamic ETL pipeline in Azure Data factory

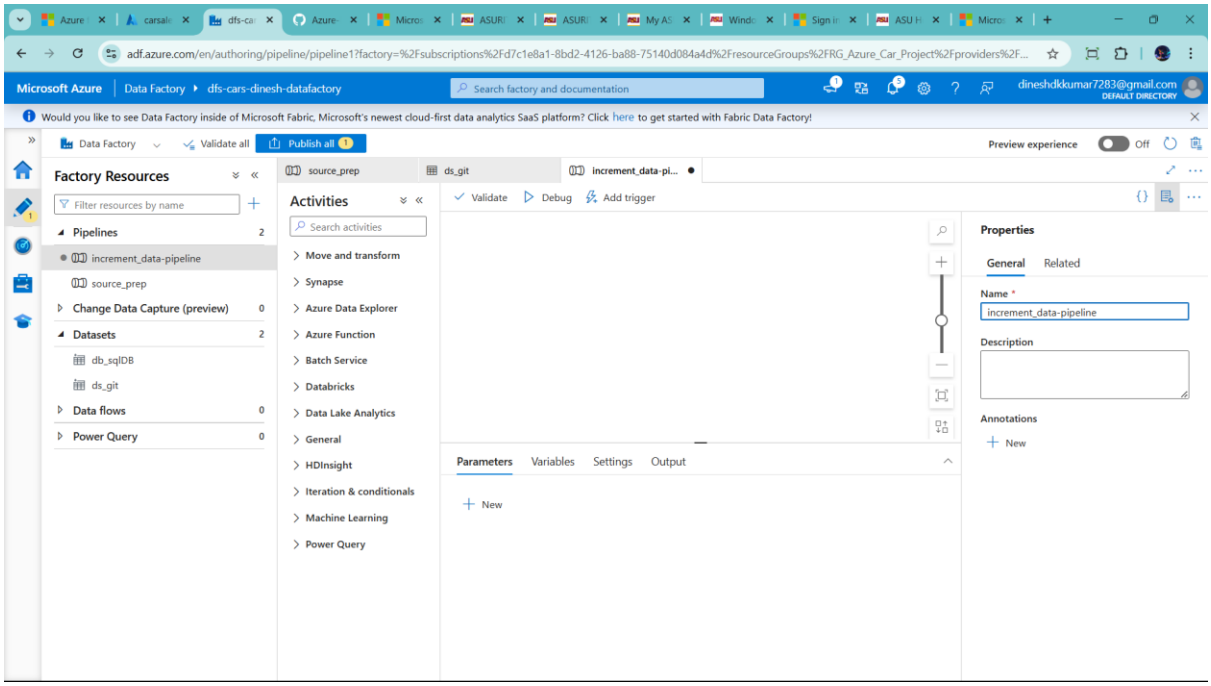
Now the public the data , so click publishing and publish

→it source-prep pipline publish success



Now create new pipeline for incremental load

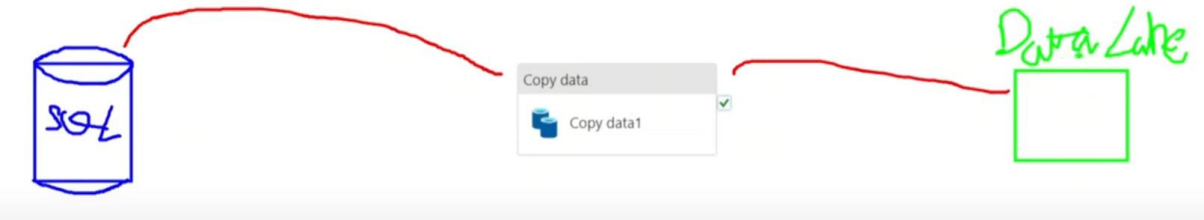
Click pipeline→new pipline → name: incremental_data_pipeline



Incremental loading in Azure data factory

In this create two pipline in incremental load

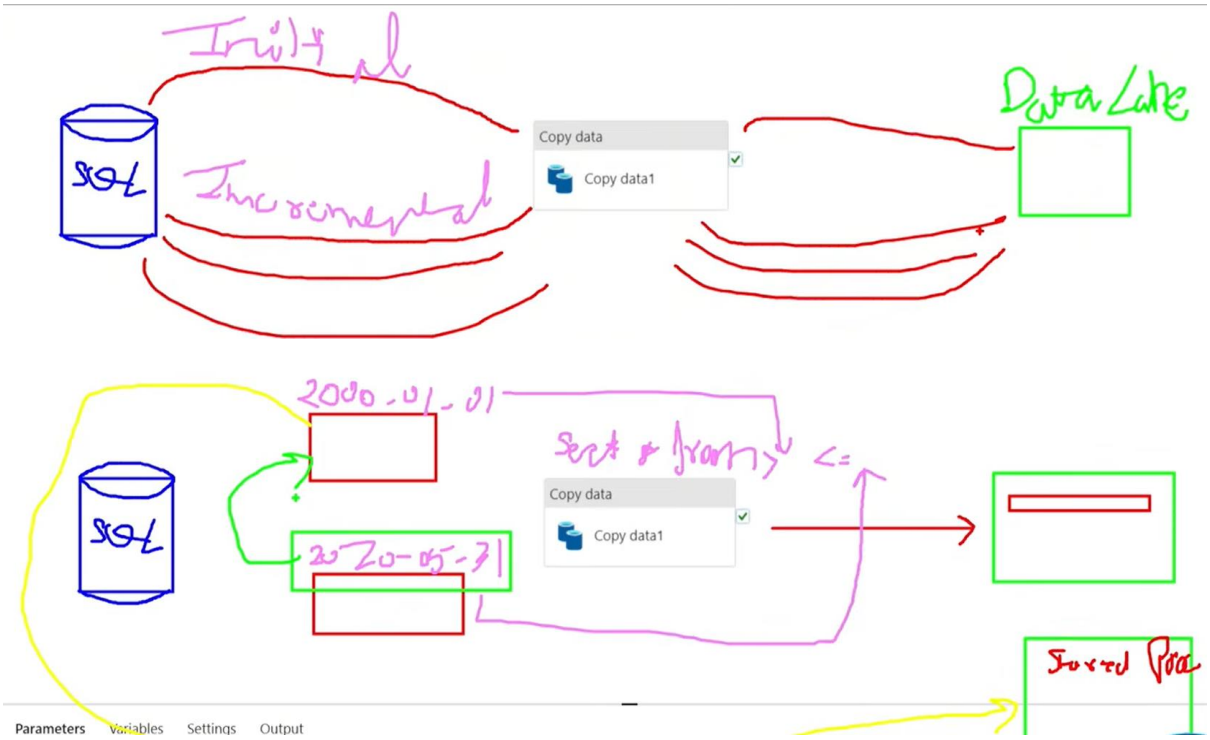
1) initial load pipeline



This is initial load . it load all the data to store in datalake

First time consider in instial load pipeline and after first time increment load pipeline use

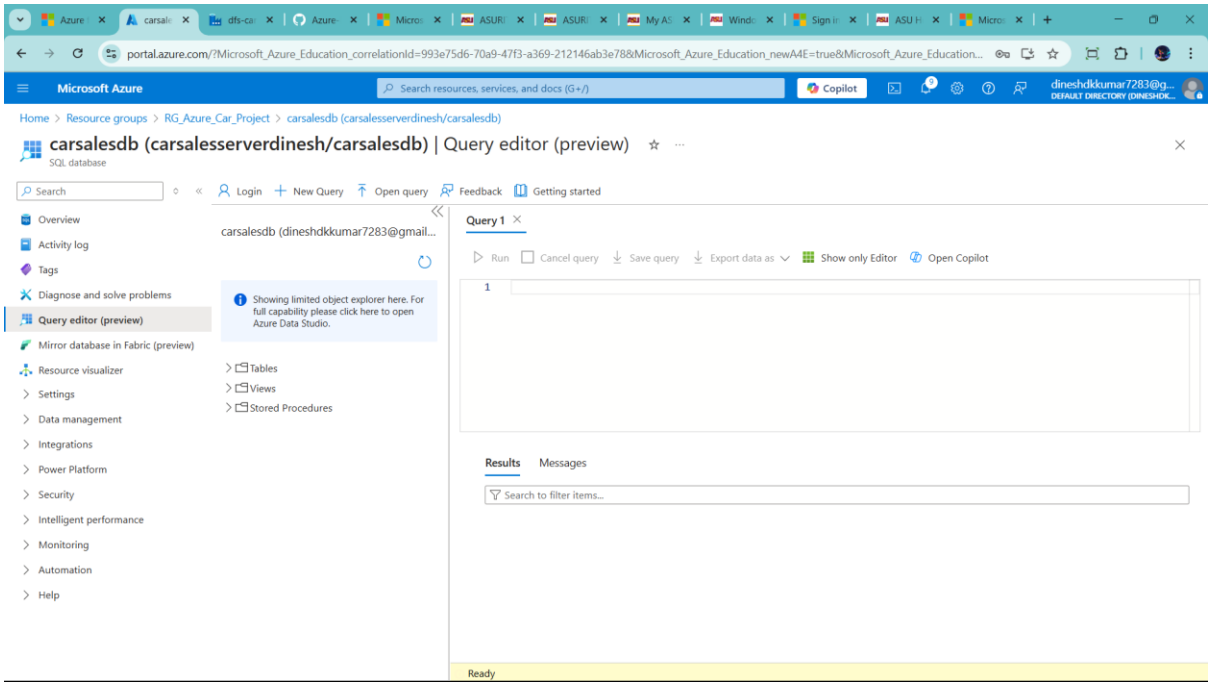
---



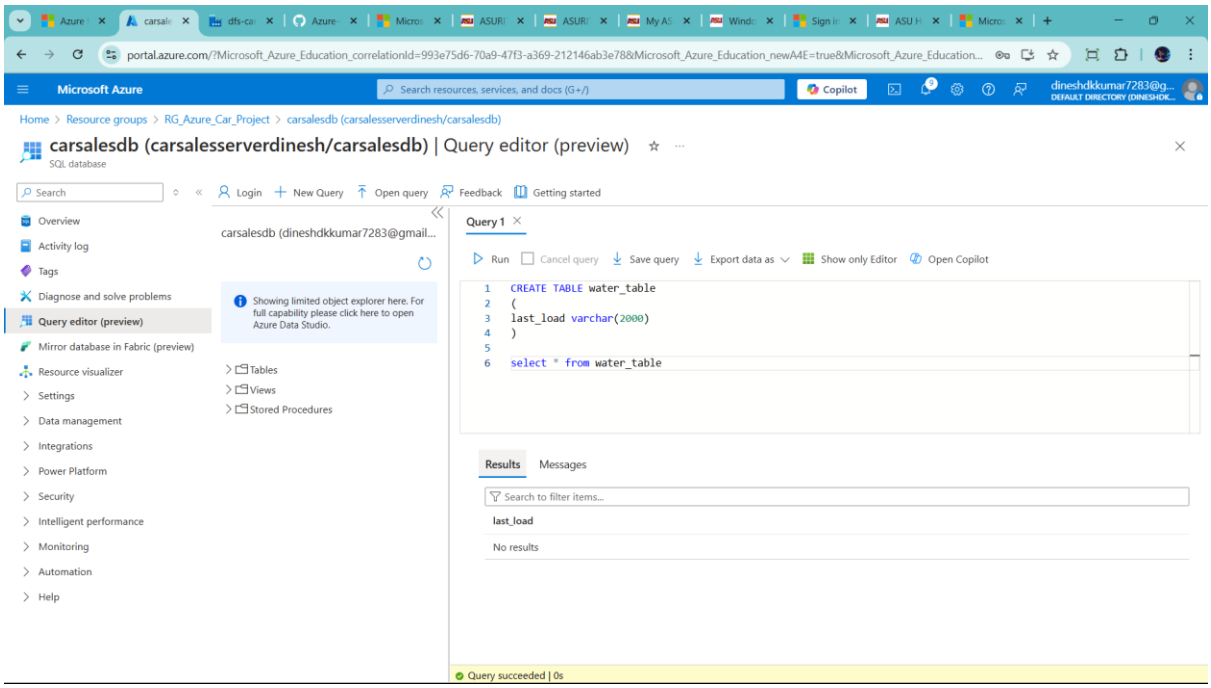
Last date add in store procedure to use for incremental load

Let's create watermark table and store procedure

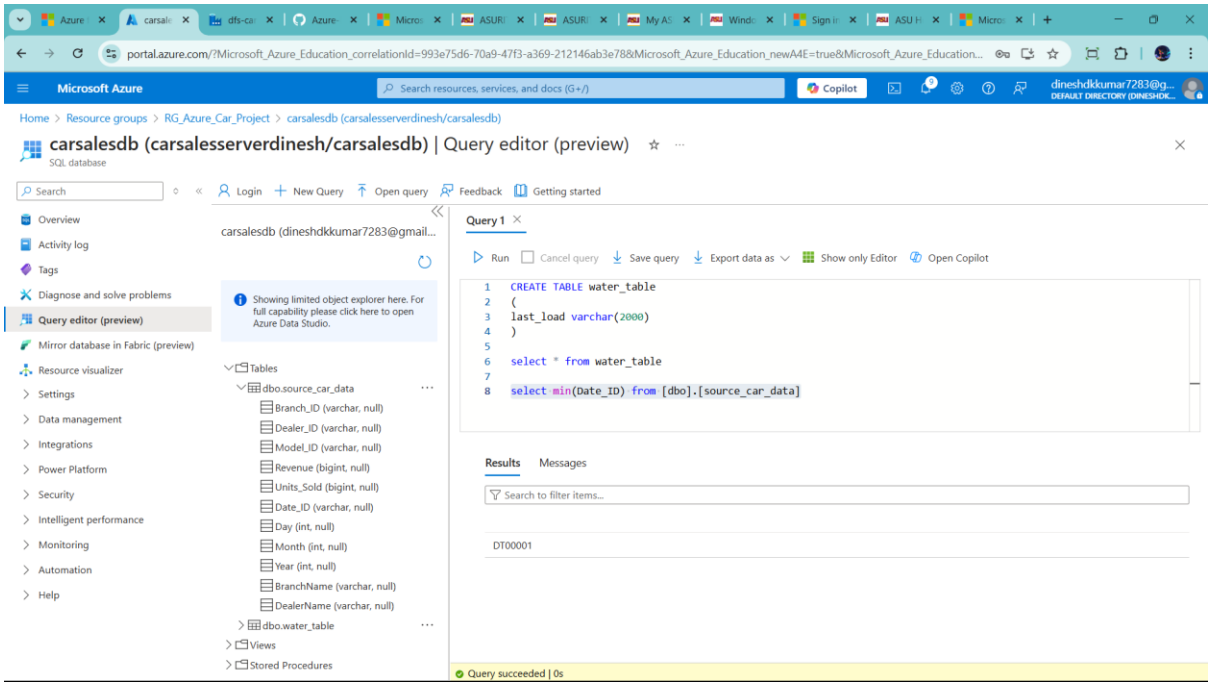
Click query section

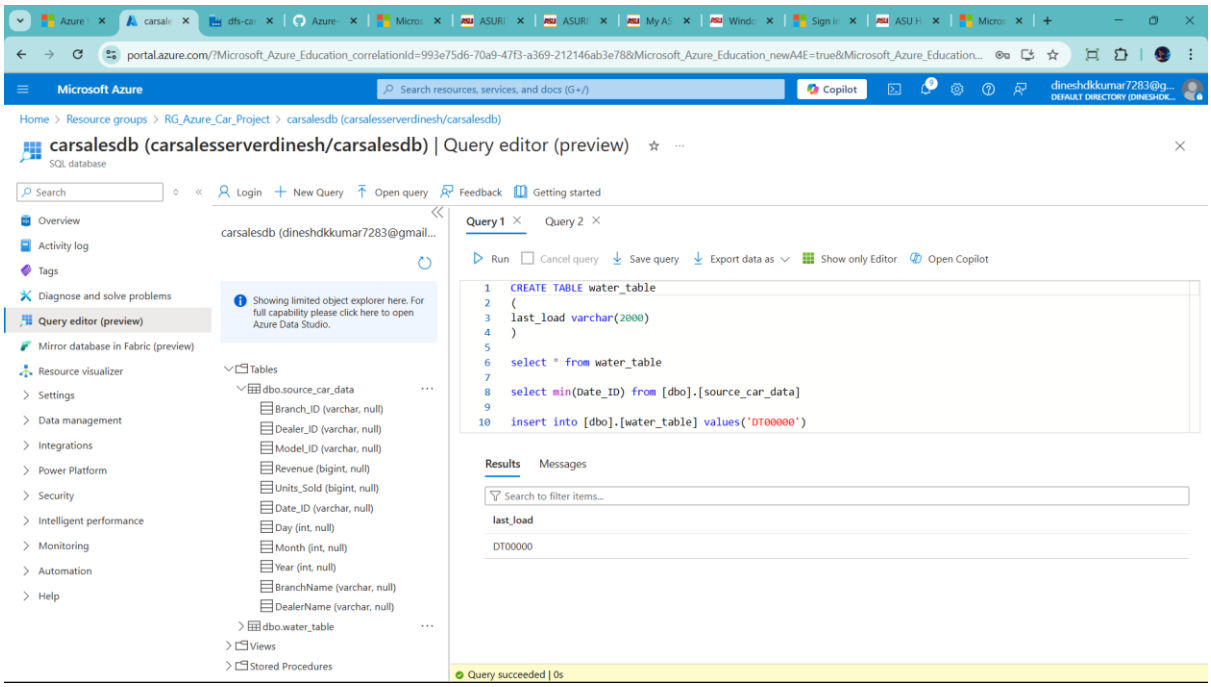


Create water mark table for store last datetime of excucte for store in this

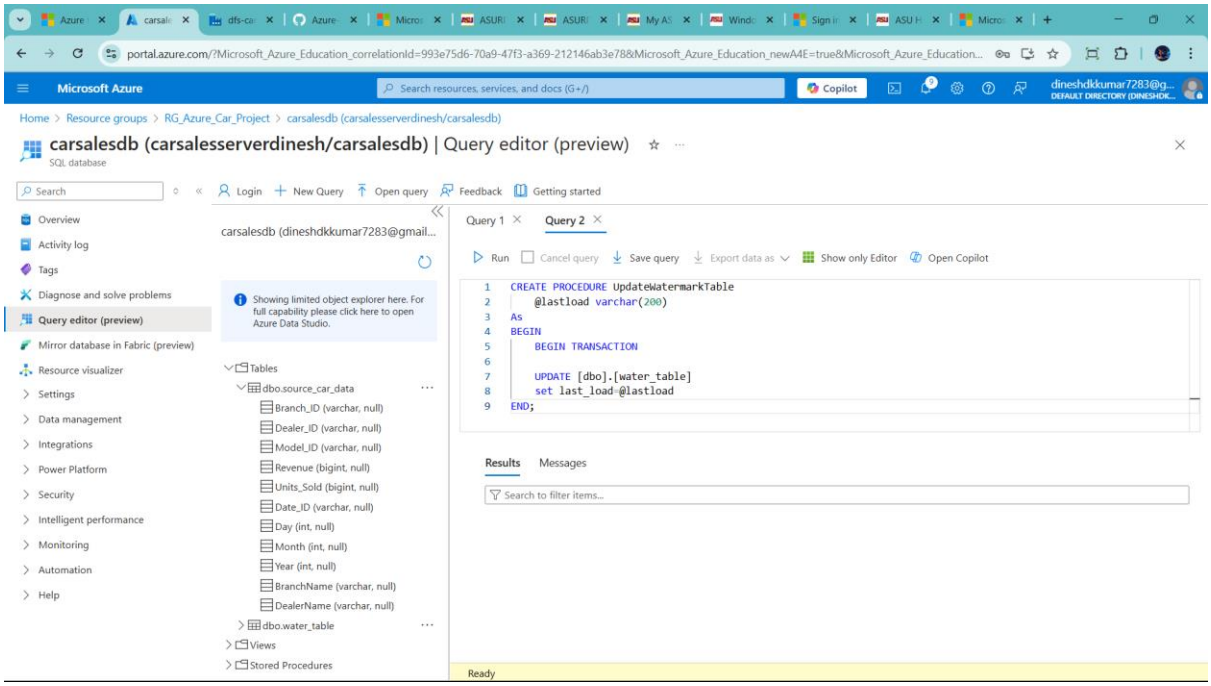


Let check min of date id in source sql table

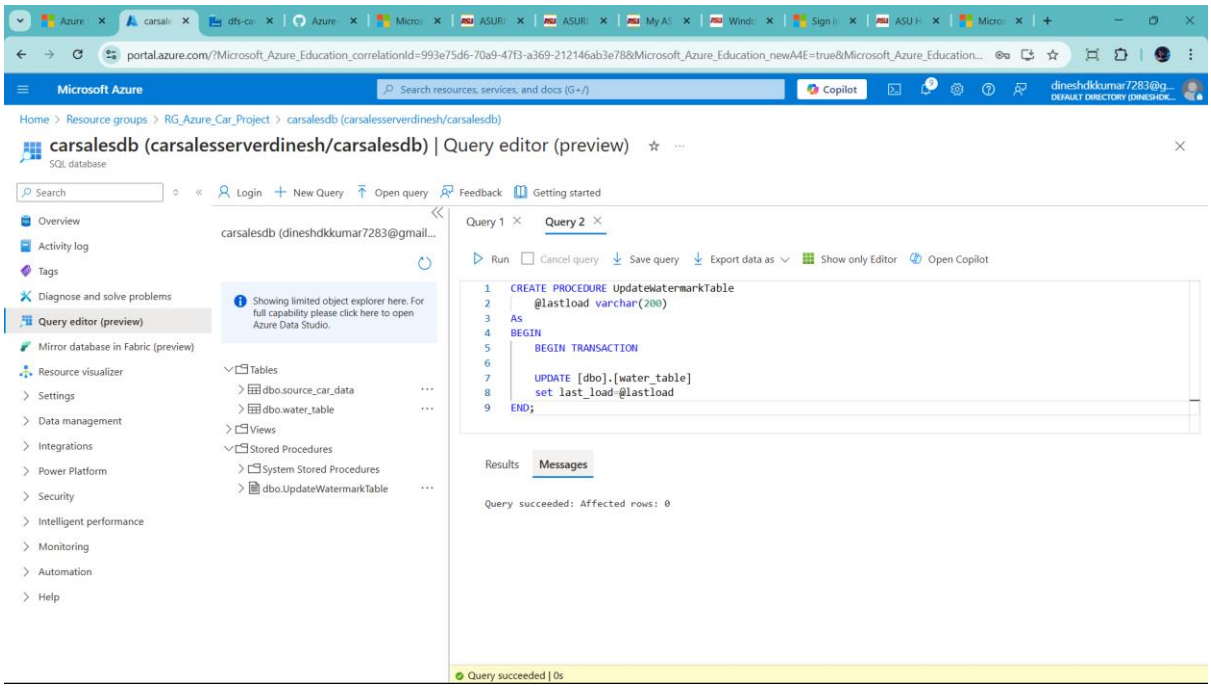




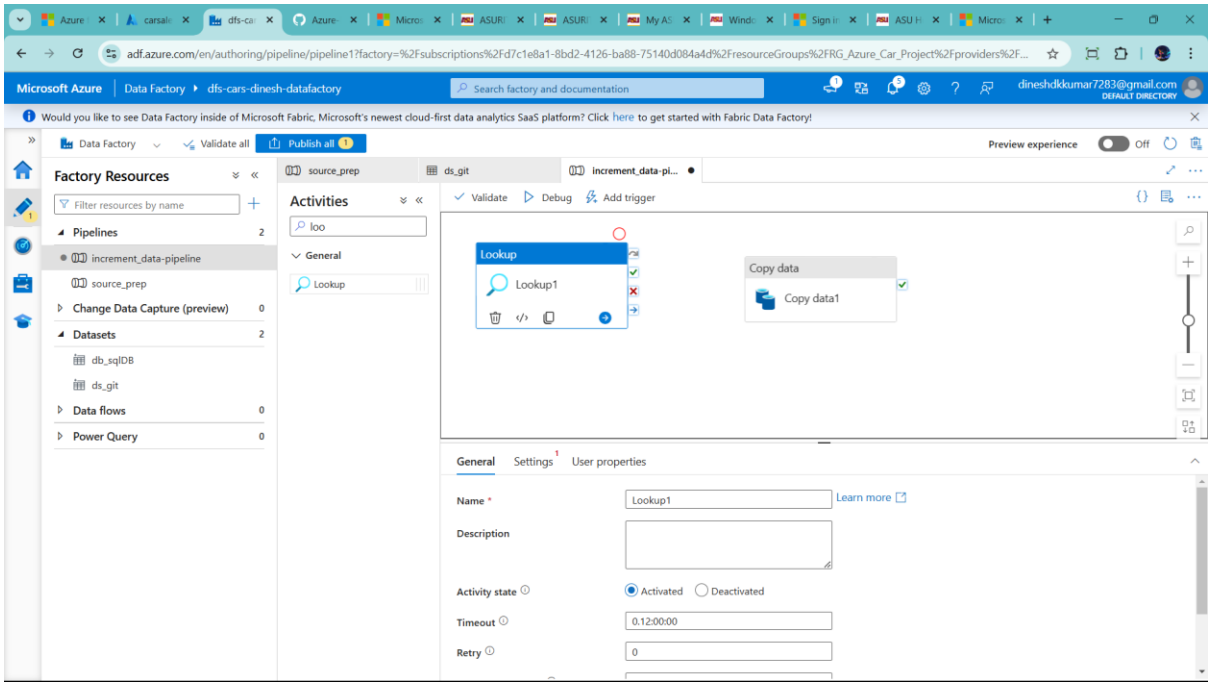
Create store procedure to add last date of data


Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home > Resource groups > RG_Azure_Car_Project > carsalesdb (carsalesserverdinesh/carsalesdb)
carsalesdb (carsalesserverdinesh/carsalesdb) | Query editor (preview)
SQL database
Login
New Query
Open query
Feedback
Getting started
Overview
Activity log
Tags
Diagnose and solve problems
Query editor (preview)
Mirror database in Fabric (preview)
Resource visualizer
Settings
Data management
Integrations
Power Platform
Security
Intelligent performance
Monitoring
Automation
Help
carsalesdb (dineshdkkumar7283@gmail...
Showing limited object explorer here. For full capability please click here to open Azure Data Studio.
Tables
dbo.source_car_data
Branch_ID (varchar, null)
Dealer_ID (varchar, null)
Model_ID (varchar, null)
Revenue (bigint, null)
Units_Sold (bigint, null)
Date_ID (varchar, null)
Day (int, null)
Month (int, null)
Year (int, null)
BranchName (varchar, null)
DealerName (varchar, null)
dbo.water_table
Views
Stored Procedures
Query 1
Query 2
Run
Cancel query
Save query
Export data as
Show only Editor
Open Copilot
1 CREATE PROCEDURE UpdateWatermarkTable
2 @lastload varchar(200)
3 As
4 BEGIN
5 BEGIN TRANSACTION
6
7 UPDATE [dbo].[water_table]
8 set last_load=@lastload
9 END;
Results
Messages
Search to filter items...
Ready

And runit

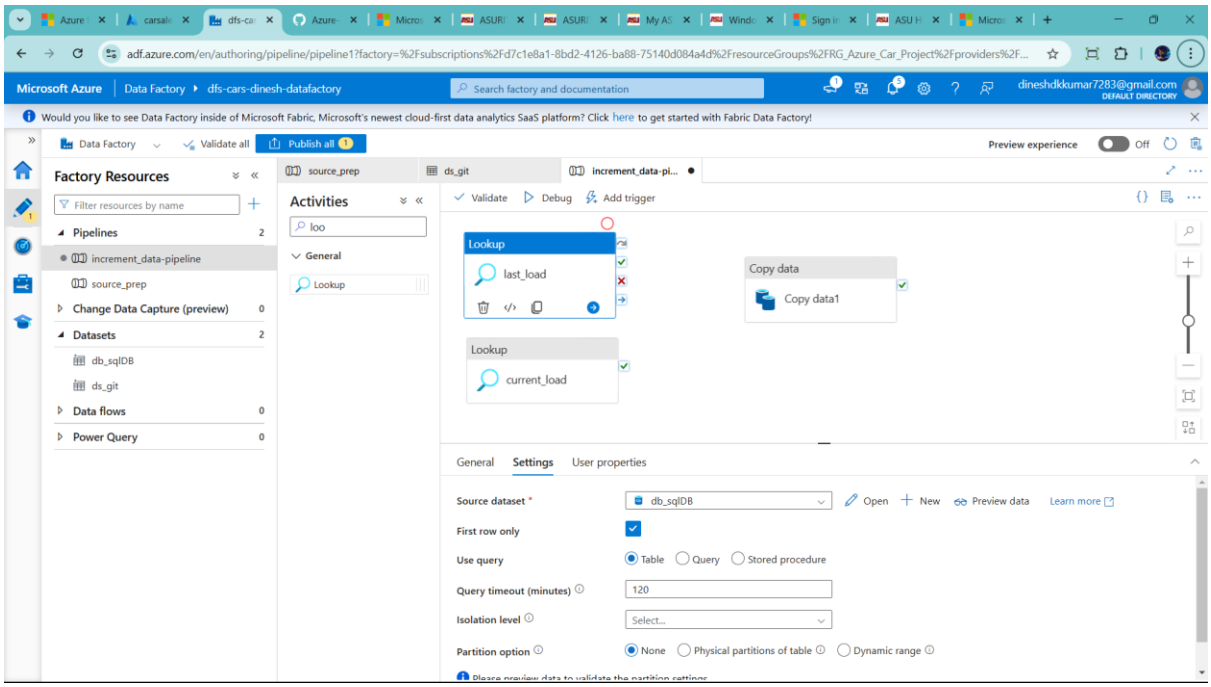

Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home > Resource groups > RG_Azure_Car_Project > carsalesdb (carsalesserverdinesh/carsalesdb)
carsalesdb (carsalesserverdinesh/carsalesdb) | Query editor (preview)
SQL database
Login
New Query
Open query
Feedback
Getting started
Overview
Activity log
Tags
Diagnose and solve problems
Query editor (preview)
Mirror database in Fabric (preview)
Resource visualizer
Settings
Data management
Integrations
Power Platform
Security
Intelligent performance
Monitoring
Automation
Help
carsalesdb (dineshdkkumar7283@gmail...
Showing limited object explorer here. For full capability please click here to open Azure Data Studio.
Tables
dbo.source_car_data
dbo.water_table
Views
Stored Procedures
System Stored Procedures
dbo.UpdateWatermarkTable
Query 1
Query 2
Run
Cancel query
Save query
Export data as
Show only Editor
Open Copilot
1 CREATE PROCEDURE UpdateWatermarkTable
2 @lastload varchar(200)
3 As
4 BEGIN
5 BEGIN TRANSACTION
6
7 UPDATE [dbo].[water_table]
8 set last_load=@lastload
9 END;
Results
Messages
Query succeeded: Affected rows: 0
Query succeeded | 0s

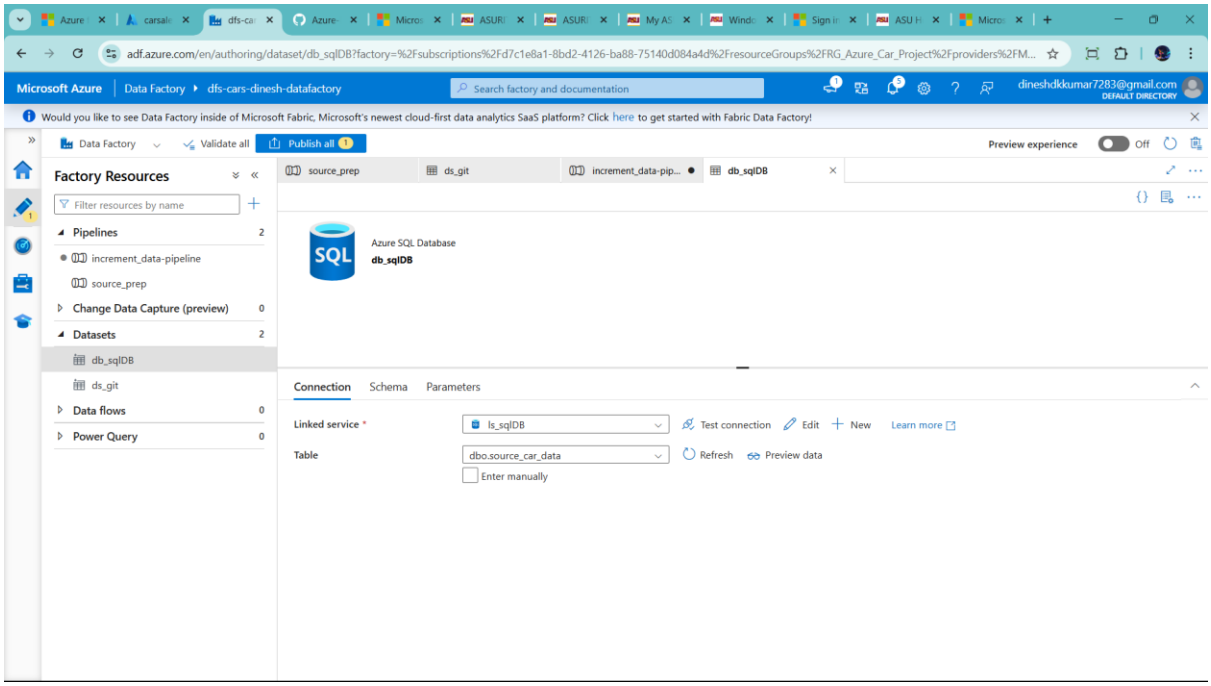
Go to data factory →click increm_data_pipeline (when you create it) →click activity use lookup



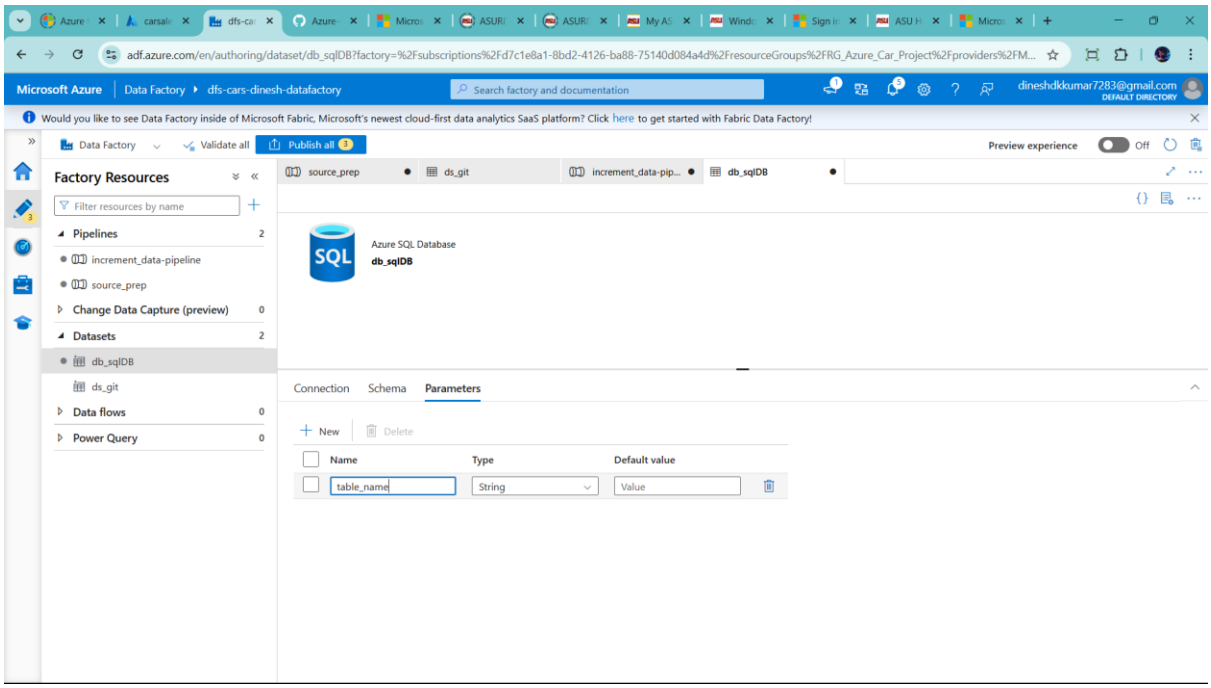
Use two lookup one is lastload and current load

Last_load →use db_sqlDB → it is source dataset

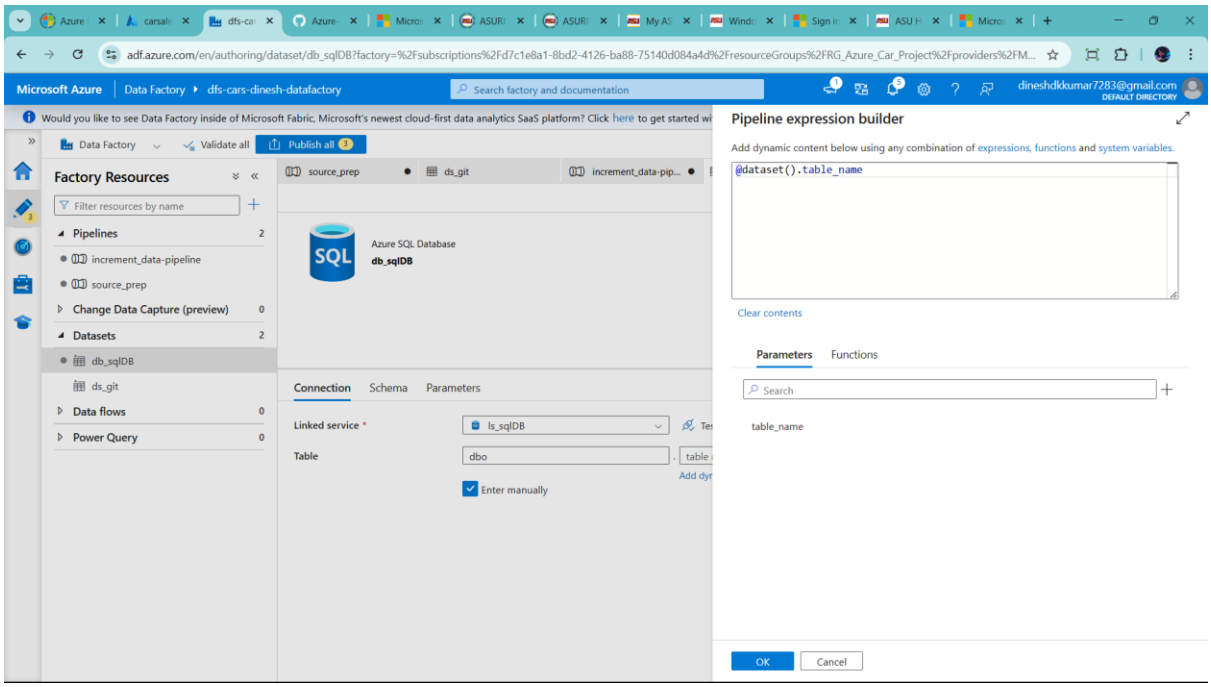


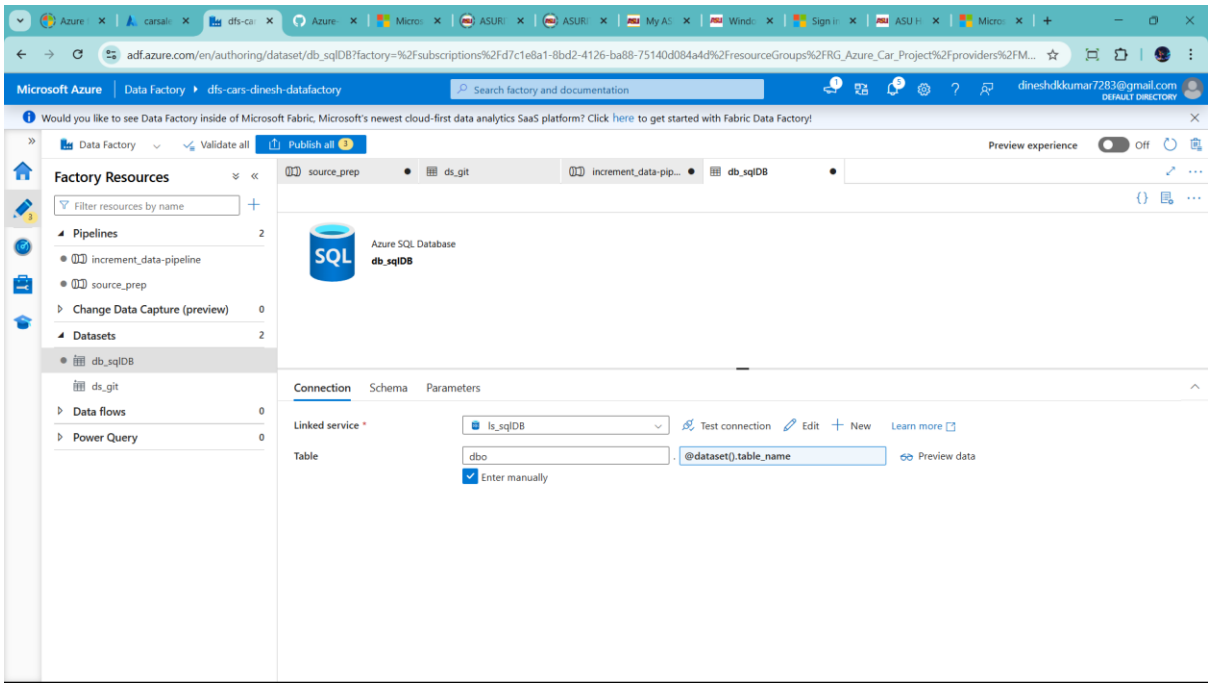


Go to parameter →click new →give name as table name

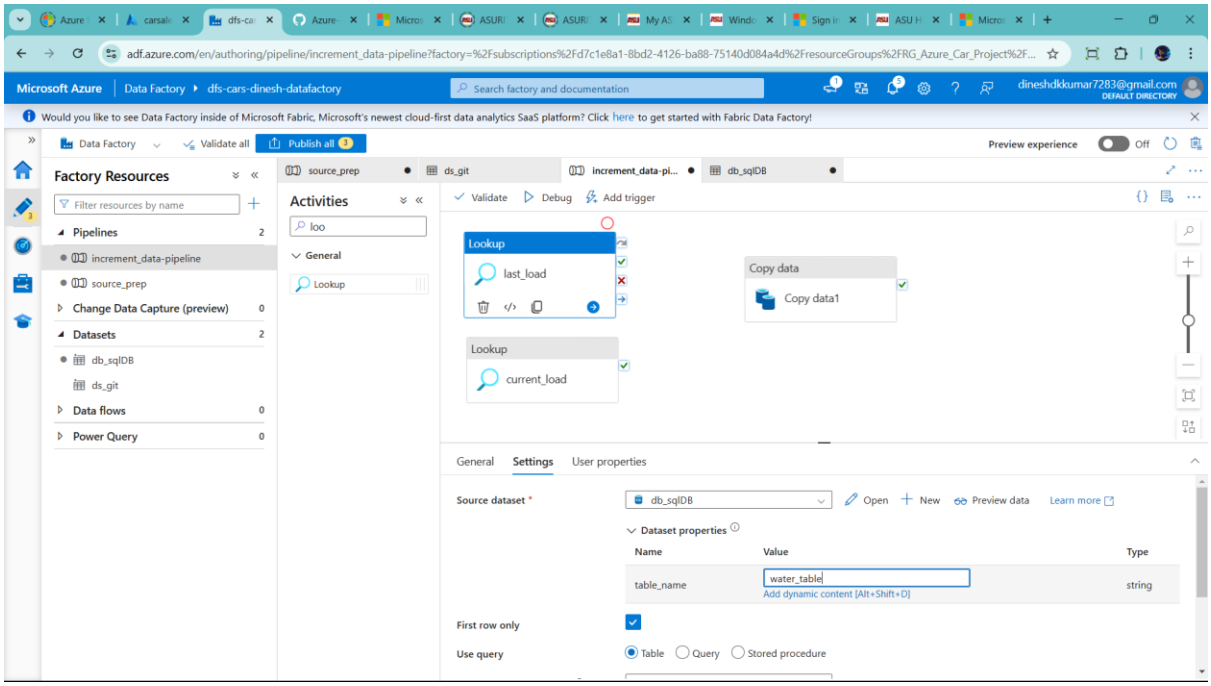


Go to connection

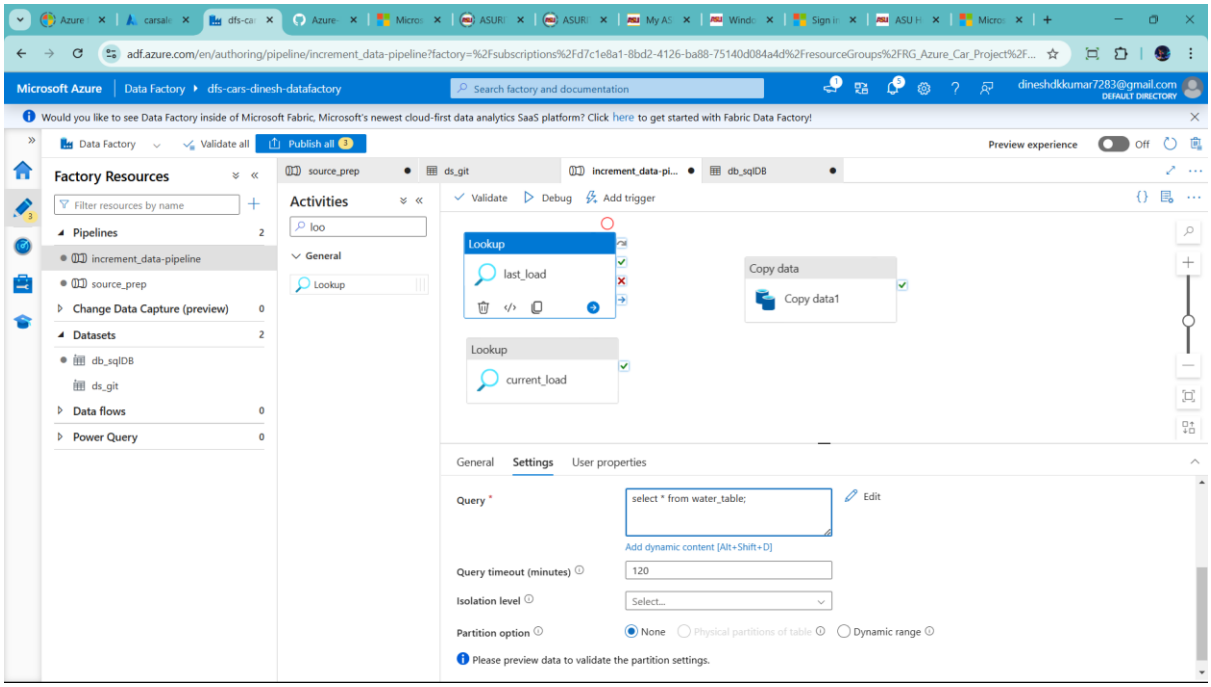

Microsoft Azure
Data Factory
dfs-cars-dinesh-datafactory
Search factory and documentation
dineshdkkumar7283@gmail.com
DEFAULT DIRECTORY
Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get started wi
Data Factory
Validate all
Publish all
Factory Resources
Filter resources by name
Pipelines
increment_data-pipeline
source_prep
Change Data Capture (preview)
Datasets
db_sqlDB
ds_git
Data flows
Power Query
source_prep
ds_git
increment_data-pip...
Azure SQL Database
db_sqlDB
Connection
Schema
Parameters
Linked service *
ls_sqlDB
Table
dbo
Enter manually
Pipeline expression builder
Add dynamic content below using any combination of expressions, functions and system variables.
@dataset().table_name
Clear contents
Parameters
Functions
Search
table_name
OK
Cancel


Microsoft Azure
Data Factory
dfs-cars-dinesh-datafactory
Search factory and documentation
dineshdkkumar7283@gmail.com
DEFAULT DIRECTORY
Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get started with Fabric Data Factory!
Data Factory
Validate all
Publish all
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines
increment_data-pipeline
source_prep
Change Data Capture (preview)
Datasets
db_sqlDB
ds_git
Data flows
Power Query
source_prep
ds_git
increment_data-pip...
db_sqlDB
Azure SQL Database
db_sqlDB
Connection
Schema
Parameters
Linked service *
ls_sqlDB
Test connection
Edit
New
Learn more
Table
dbo
@dataset().table_name
Preview data
Enter manually

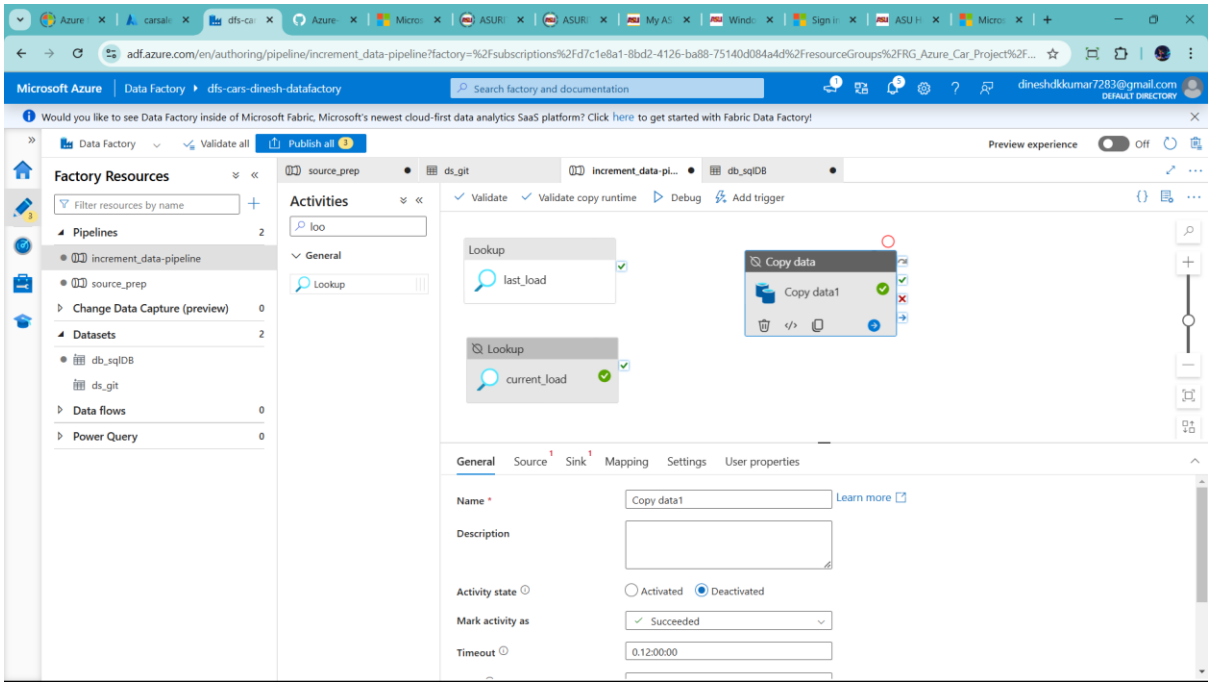
Now go to increment_data_load → click last load flow →add table name water table

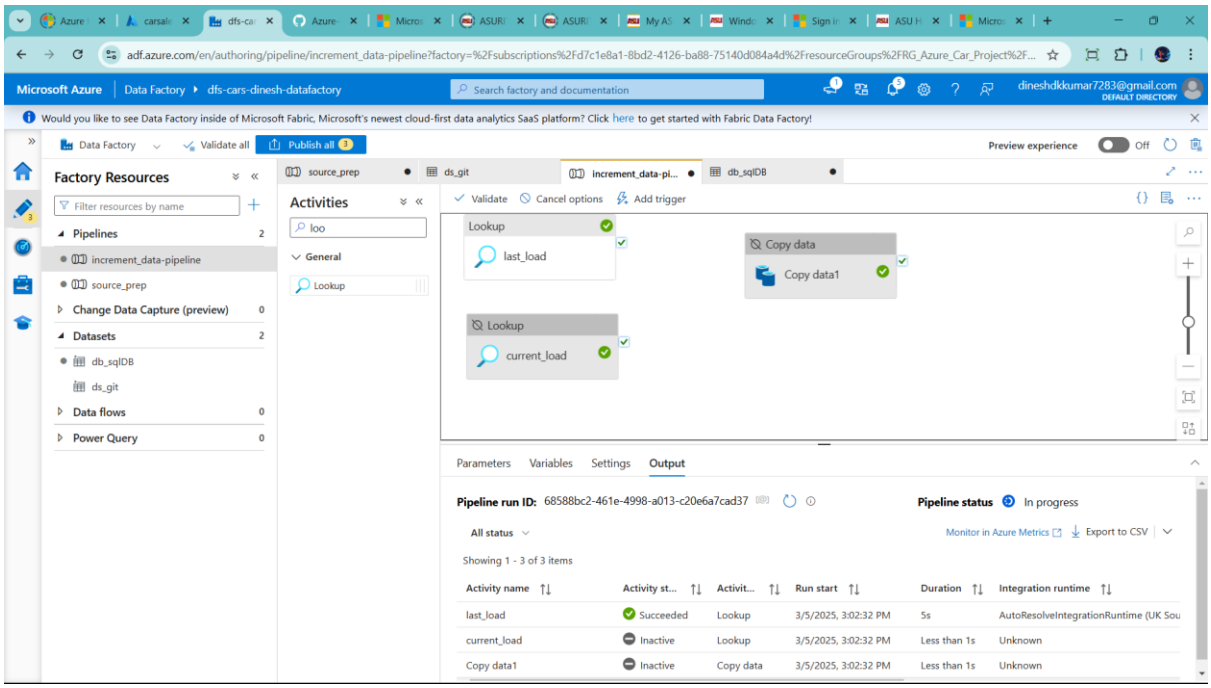


Uncheck first row only → click query →select * from water_table

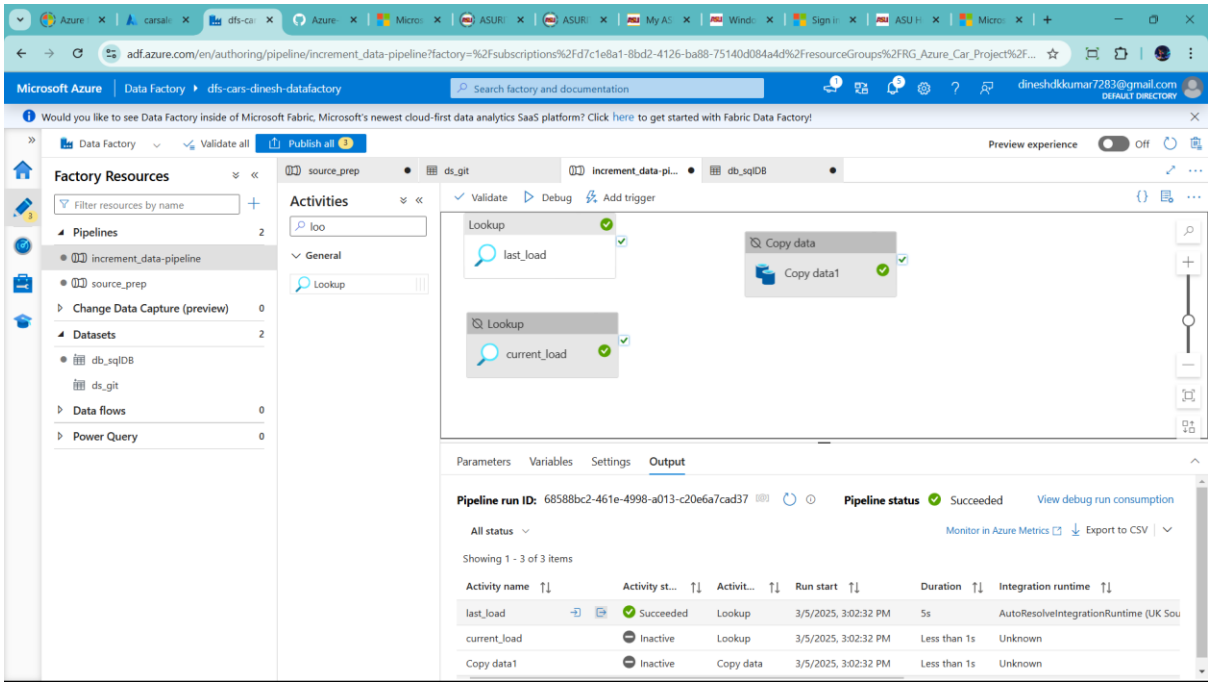


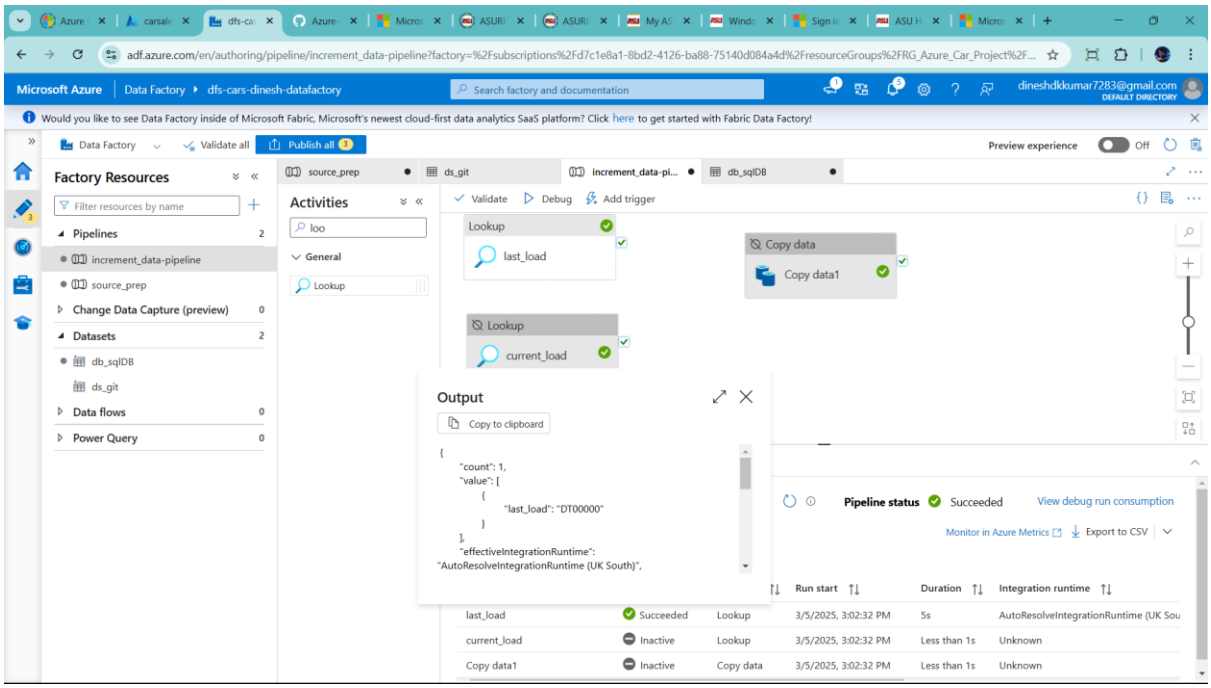
Deactive both the debug it



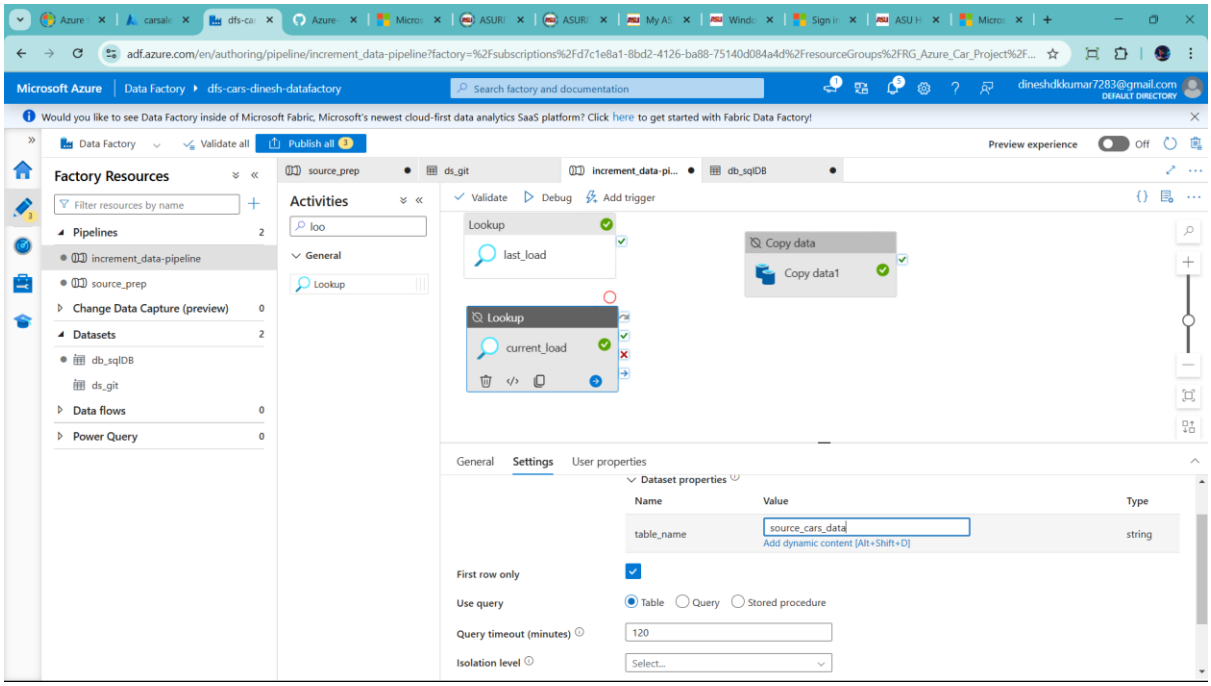


Click output for success

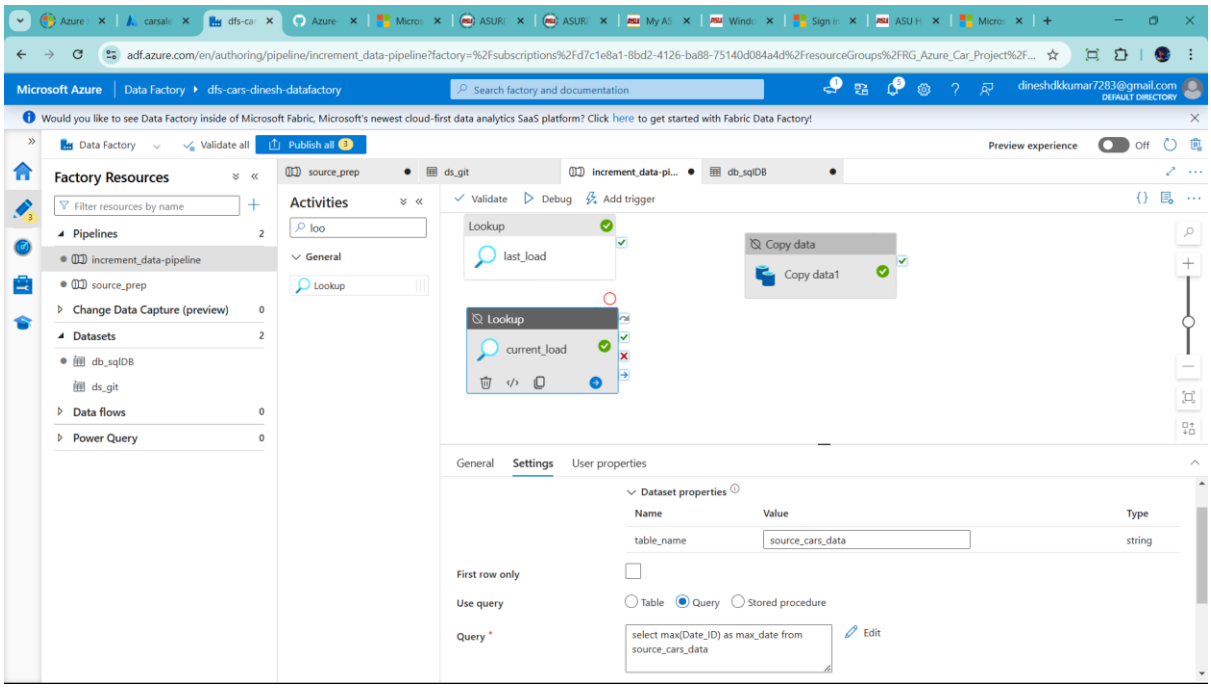




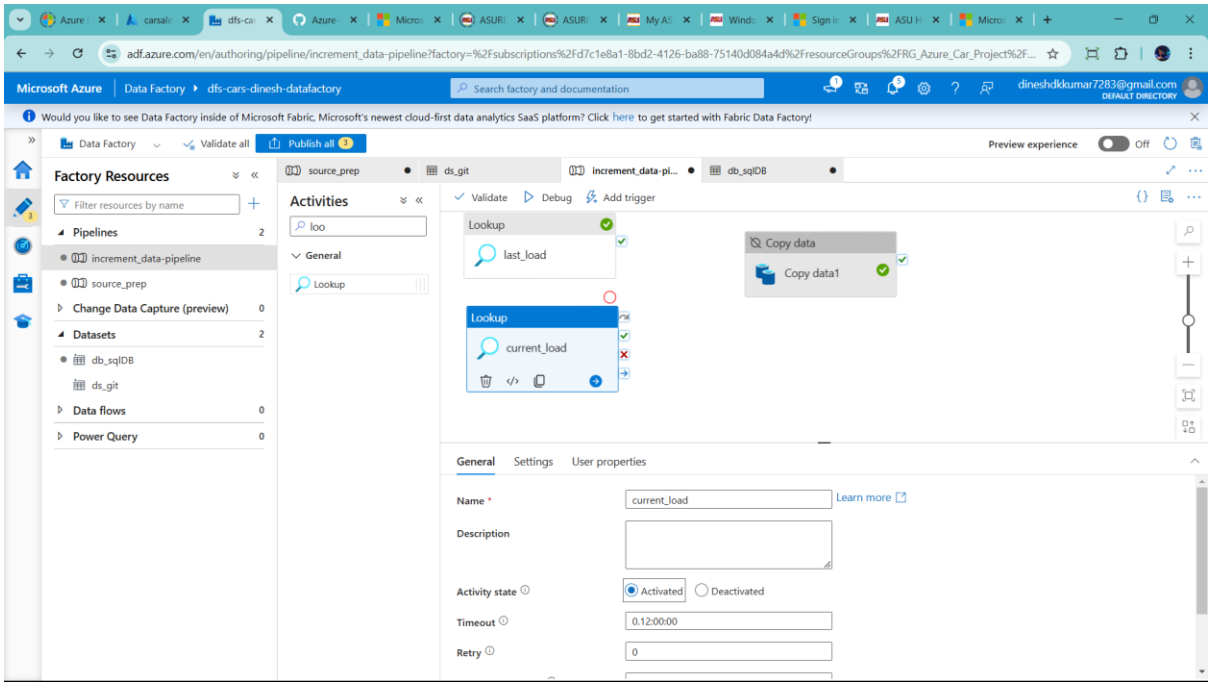
Now change the name loopup card for current load→ add parameter to pass name table name source_cars_data

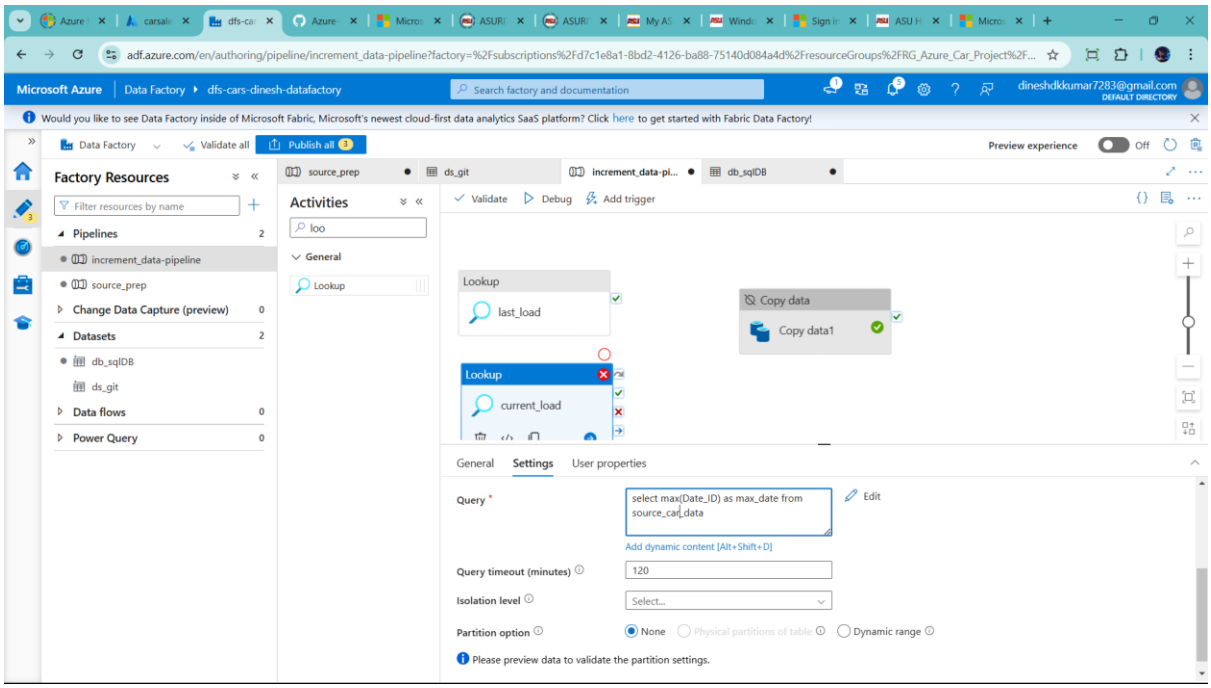


Uncheck first row only → use query -check Query



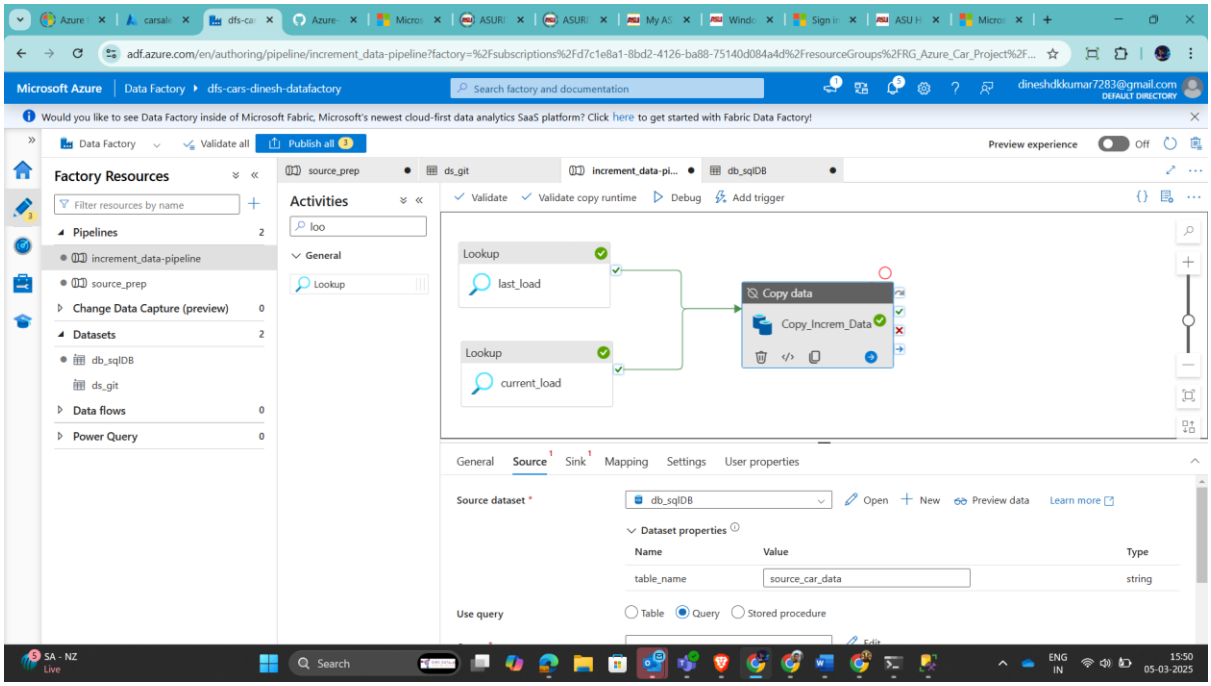
Final activate

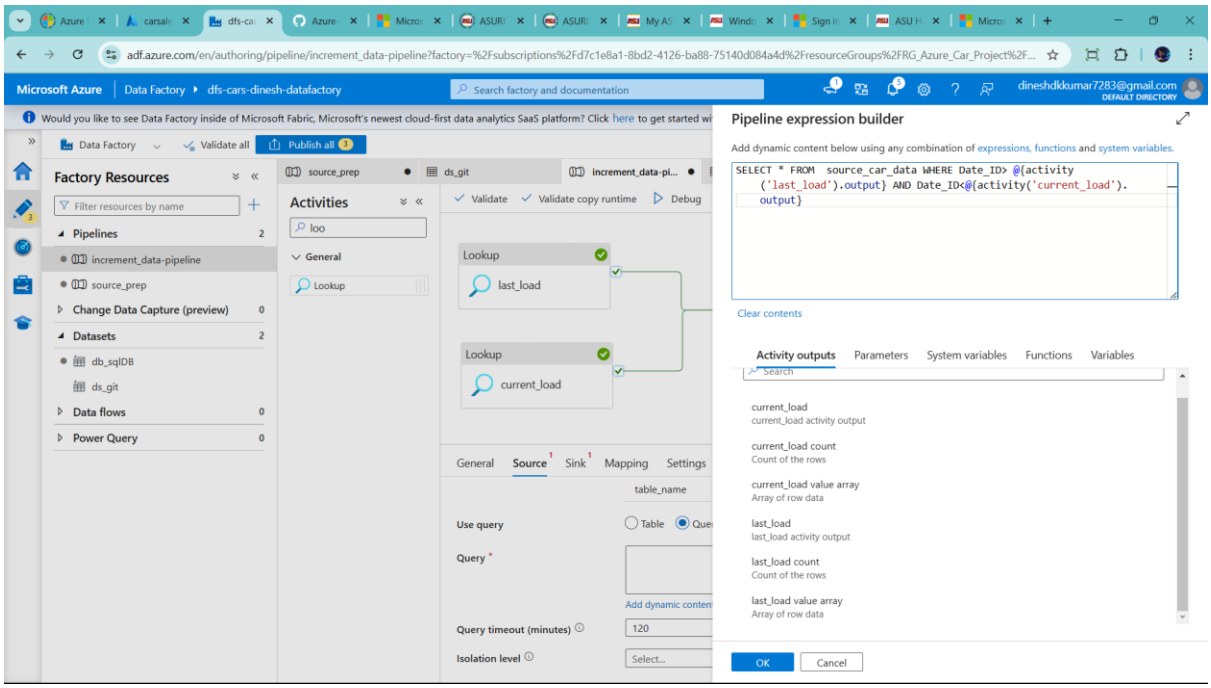


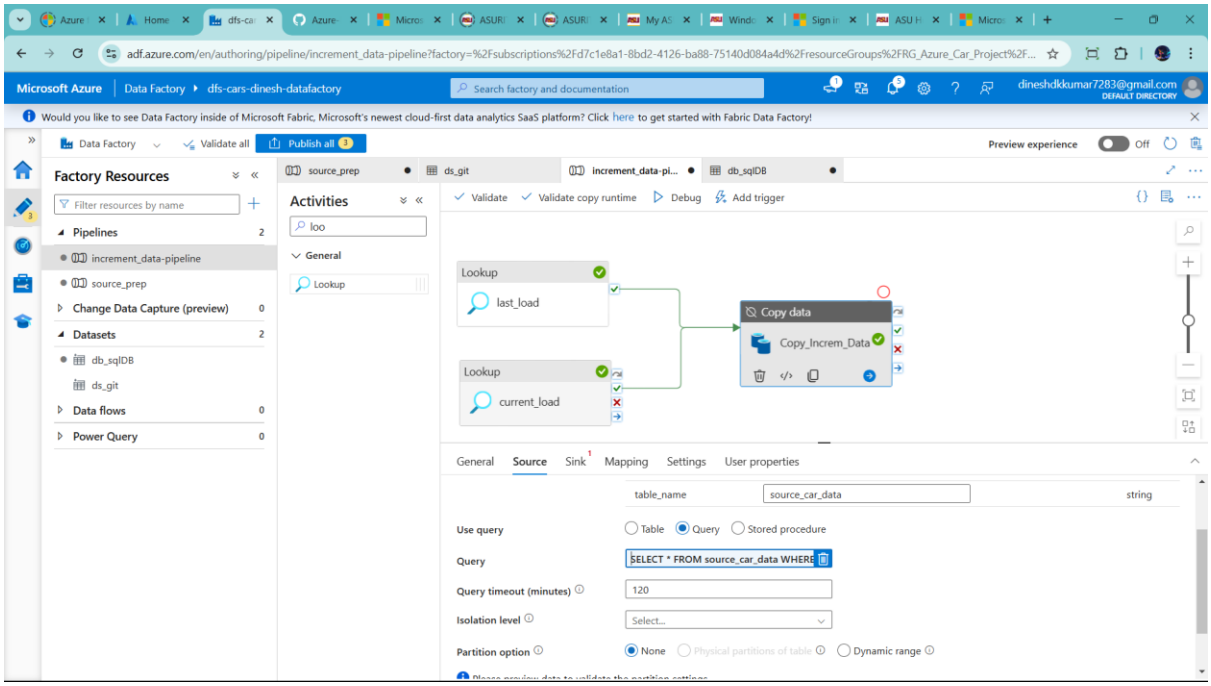


Not change as well Copy Data1 to rename →Copy_Increm_Data→click source and dynamic name- : source_car_data → and click Query

→before write query →connect with two lookup to copy data/ sql


Microsoft Azure
Data Factory
dfs-cars-dinesh-datafactory
Factory Resources
Pipelines
increment_data-pipeline
source_prep
Change Data Capture (preview)
Datasets
db_sqlDB
ds_git
Data flows
Power Query
Activities
General
Lookup
last_load
current_load
Copy data
Copy_Increm_Data
General
Source
Sink
Mapping
Settings
User properties
Source dataset
db_sqlDB
Dataset properties
Name
Value
Type
table_name
source_car_data
string
Use query
Table
Query
Stored procedure

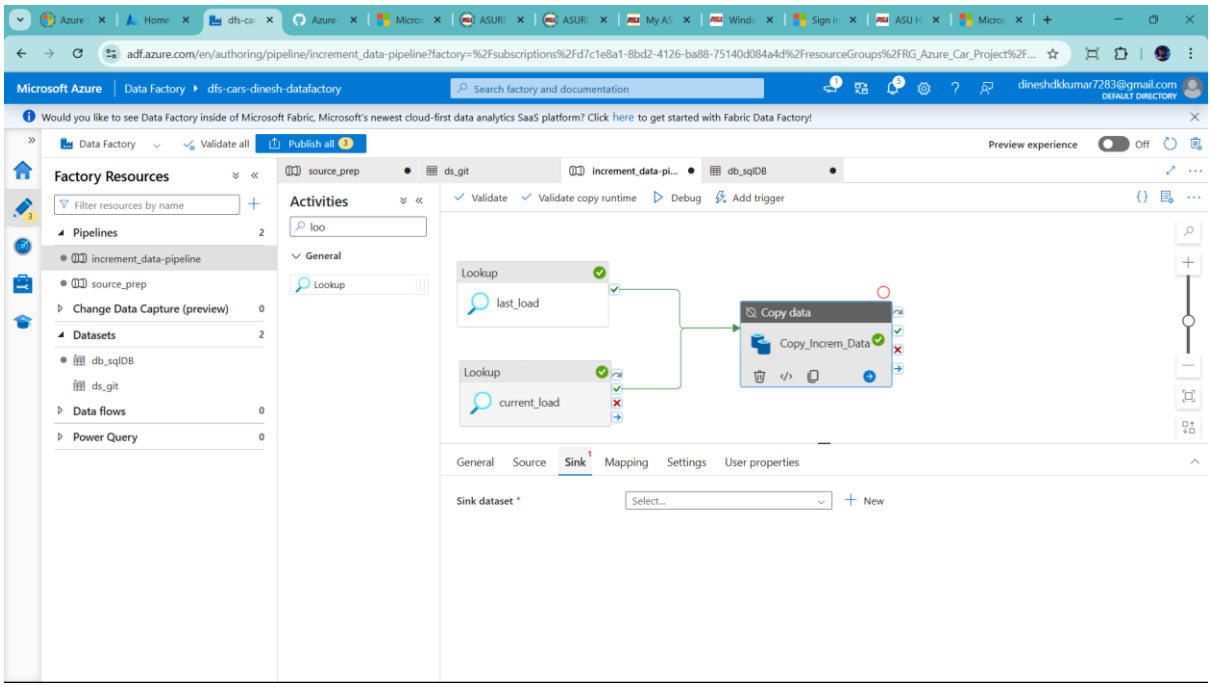

Pipeline expression builder
Add dynamic content below using any combination of expressions, functions and system variables.
SELECT * FROM source_car_data WHERE Date_ID> @{activity
('last_load').output} AND Date_ID<@{activity('current_load').
output}
Clear contents
Activity outputs
Parameters
System variables
Functions
Variables
current_load
current_load activity output
current_load count
Count of the rows
current_load value array
Array of row data
last_load
last_load activity output
last_load count
Count of the rows
last_load value array
Array of row data
OK
Cancel



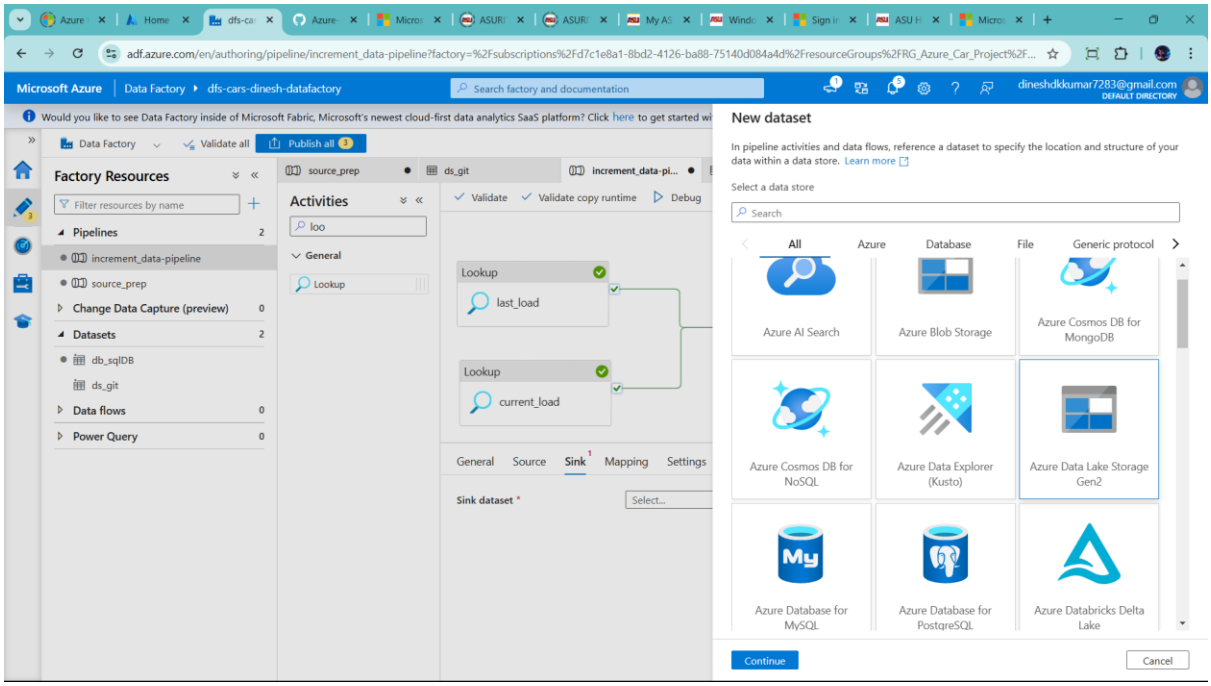
Let understood

SQL -→ Data Factory → Datalake (Brozen Container)

You can create data lake → based on sink /manager tap

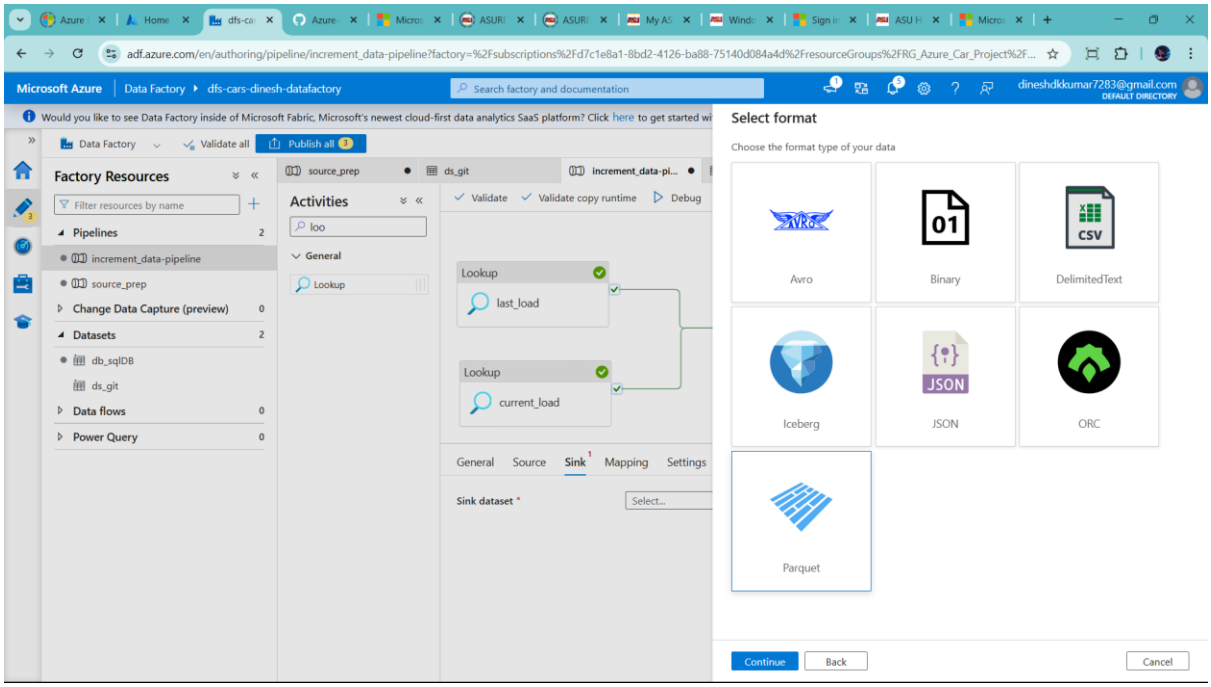


--in sink → click new →azure data lake gen 2



--click continue

→parquet formate to store data

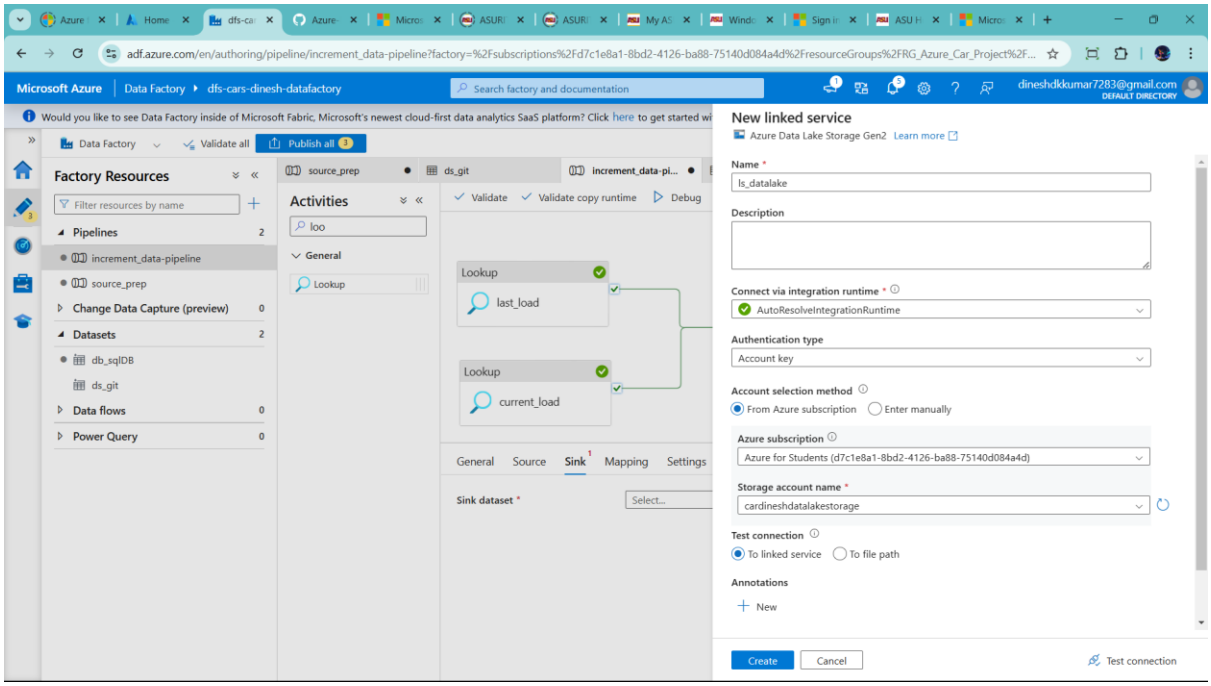


Click continue

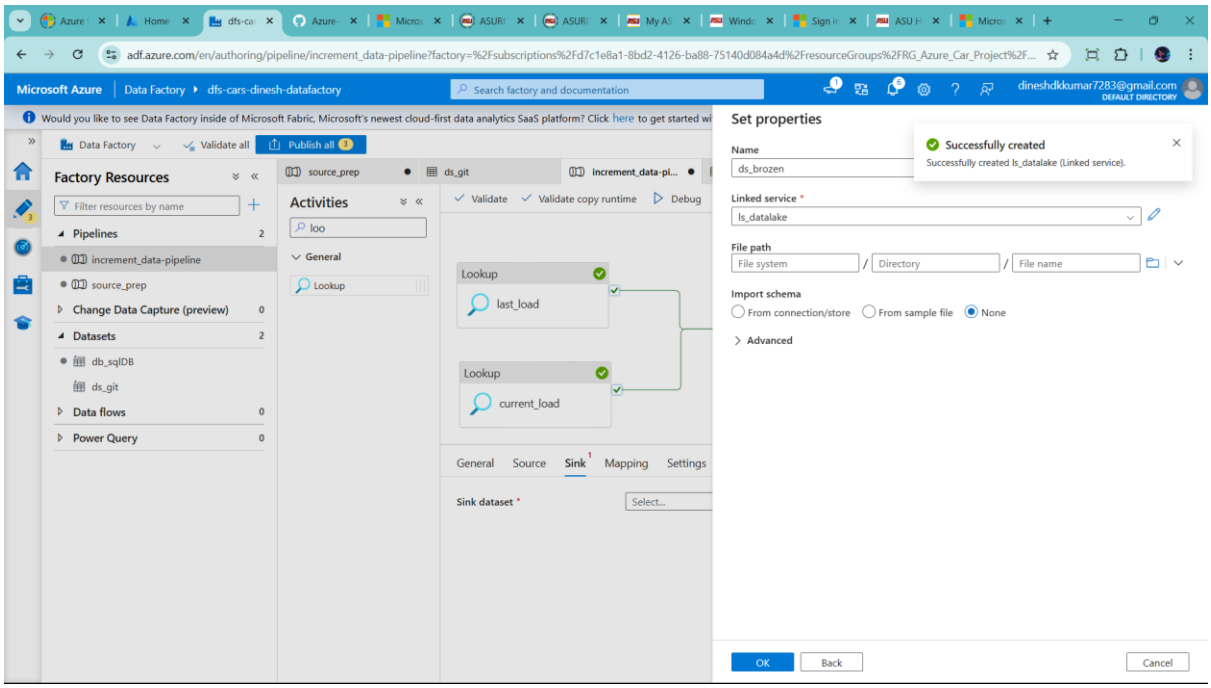
→name as ds_brozen →linked service to ls_datalake

→Azure subriscation chage to azure for student

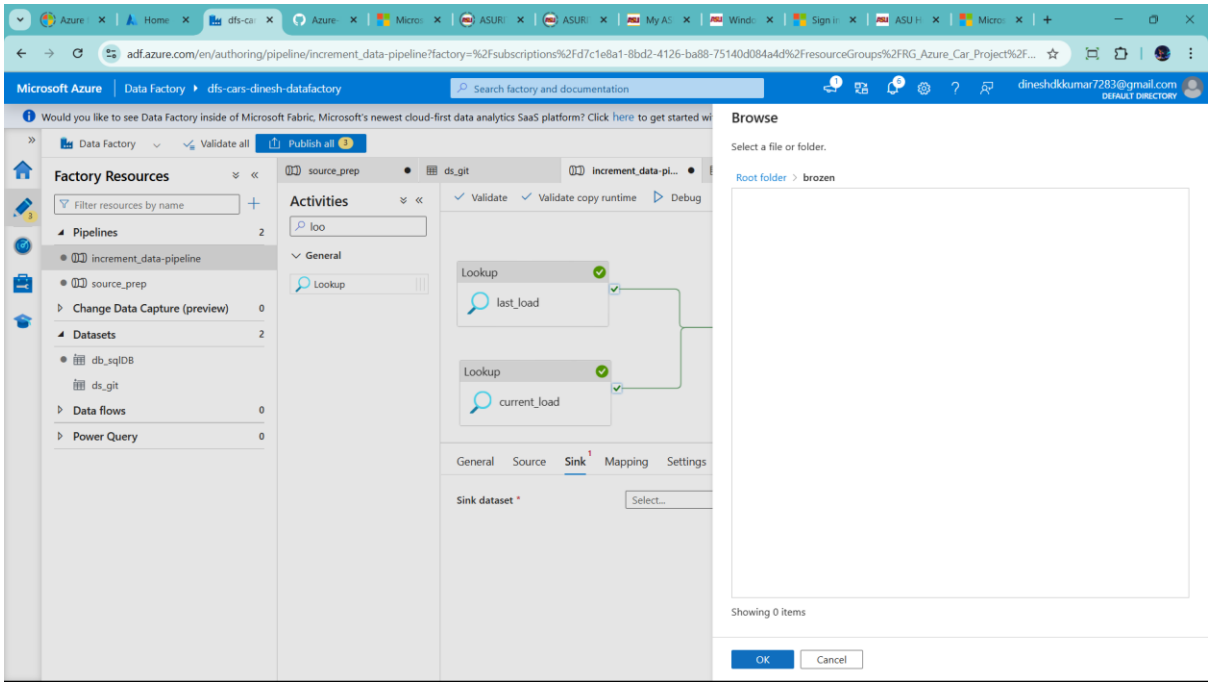
→storage account name



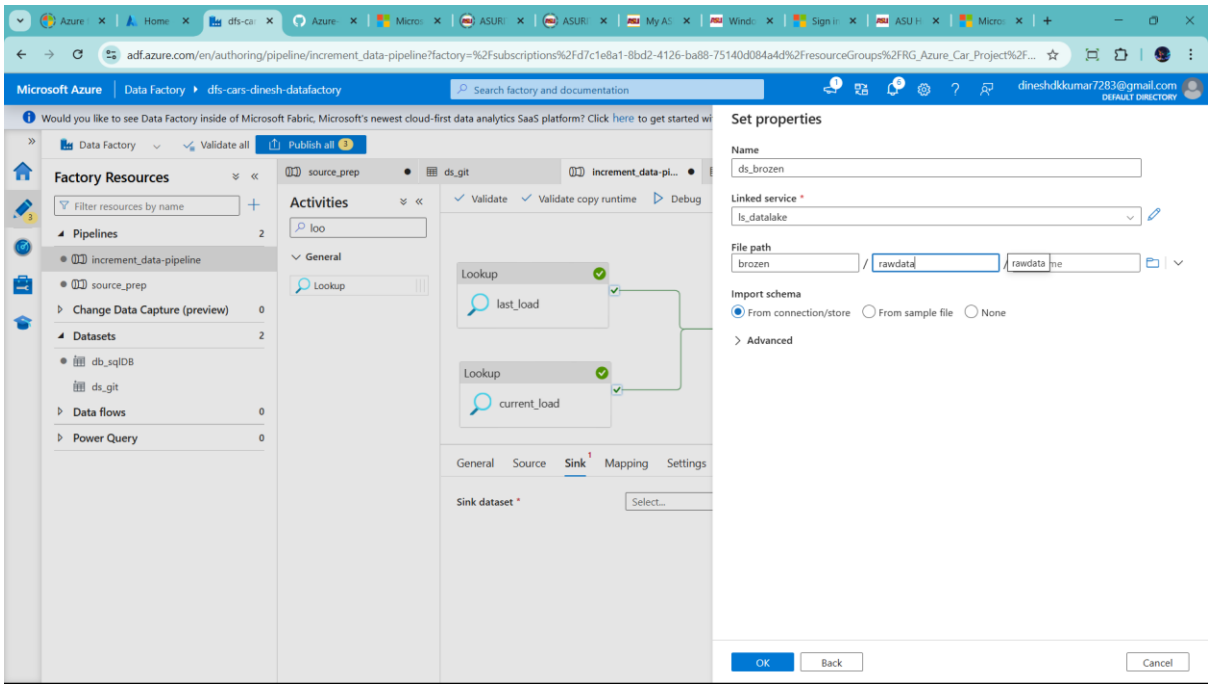
→click create



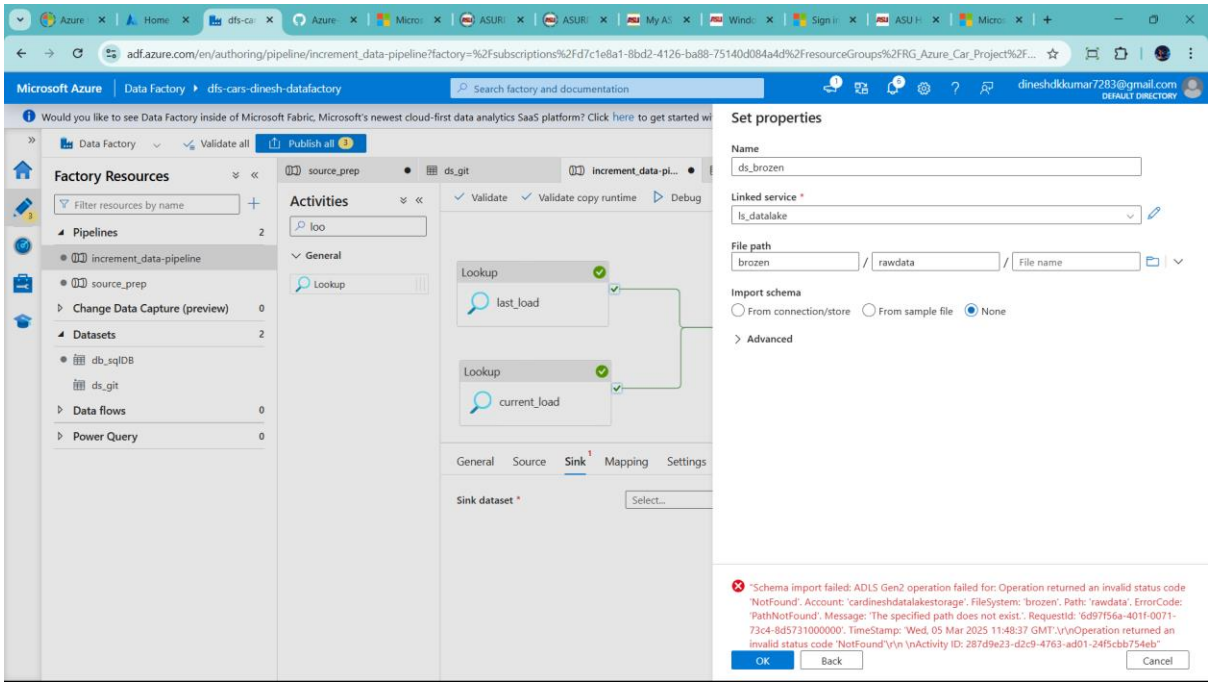
Click brozen file path →click ok



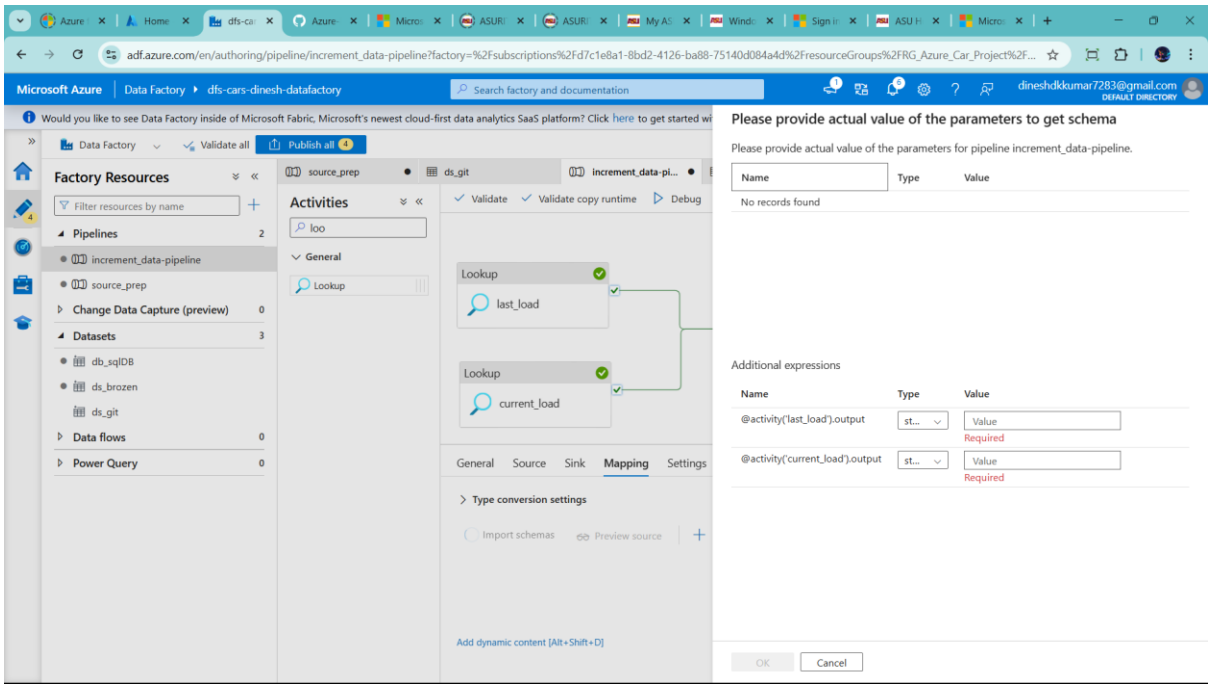
Add directory →raw data

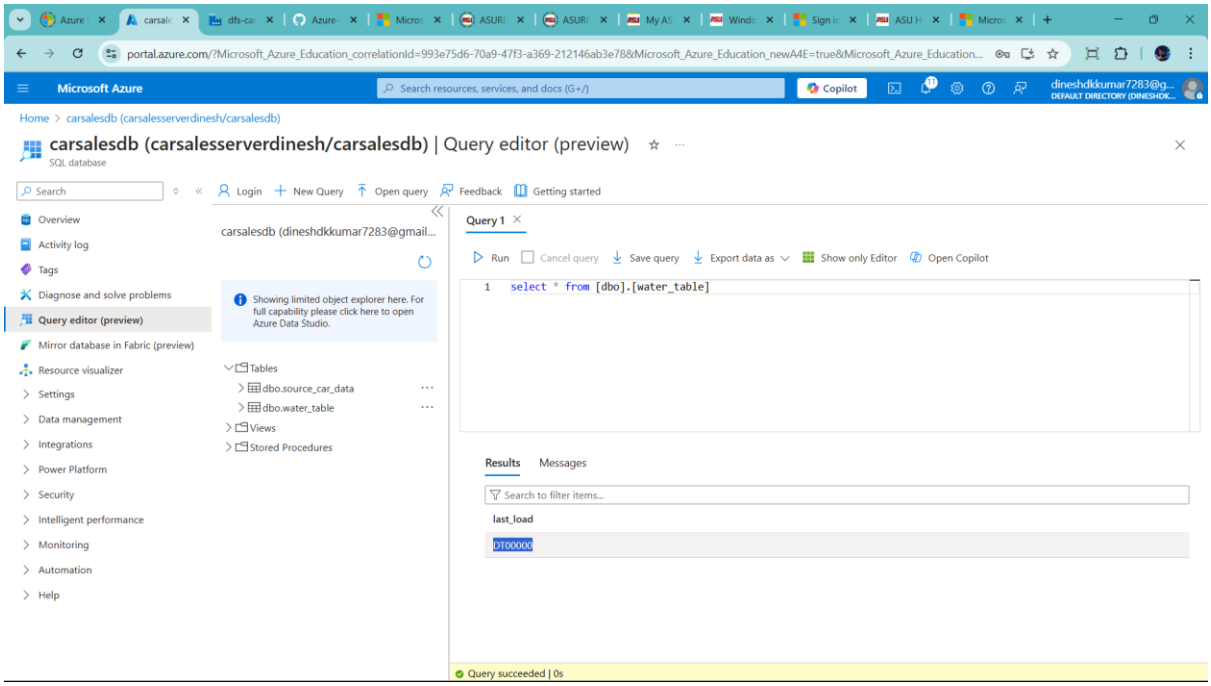


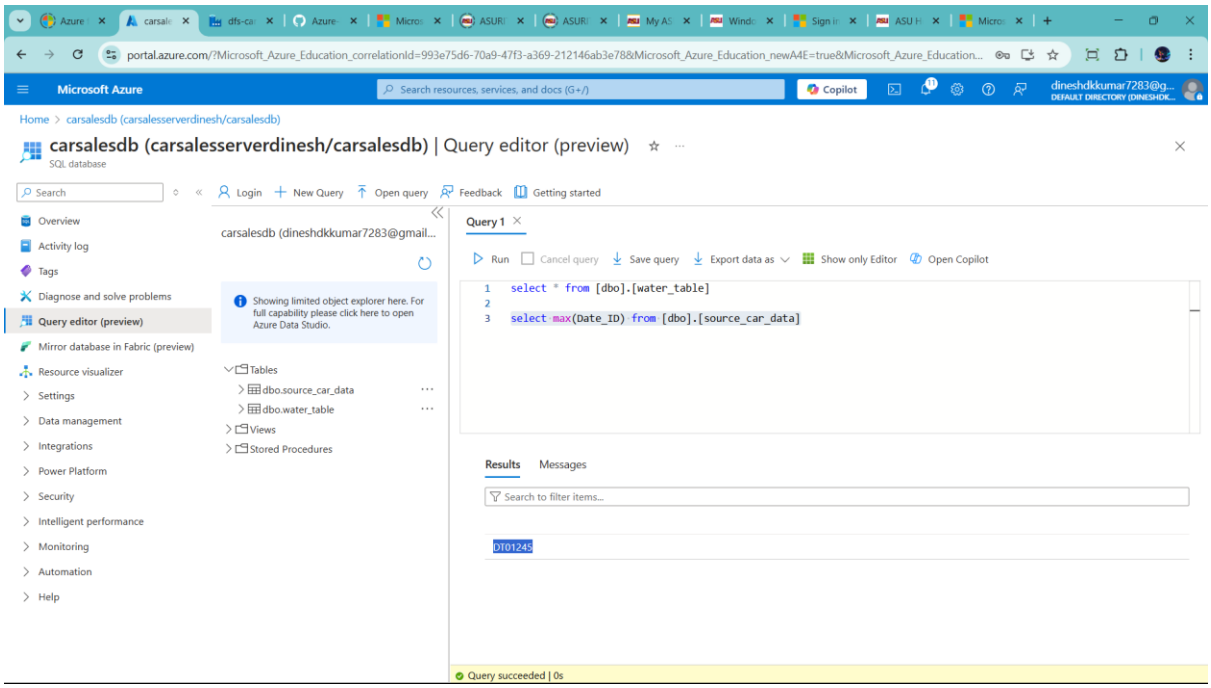
Click inport schema none

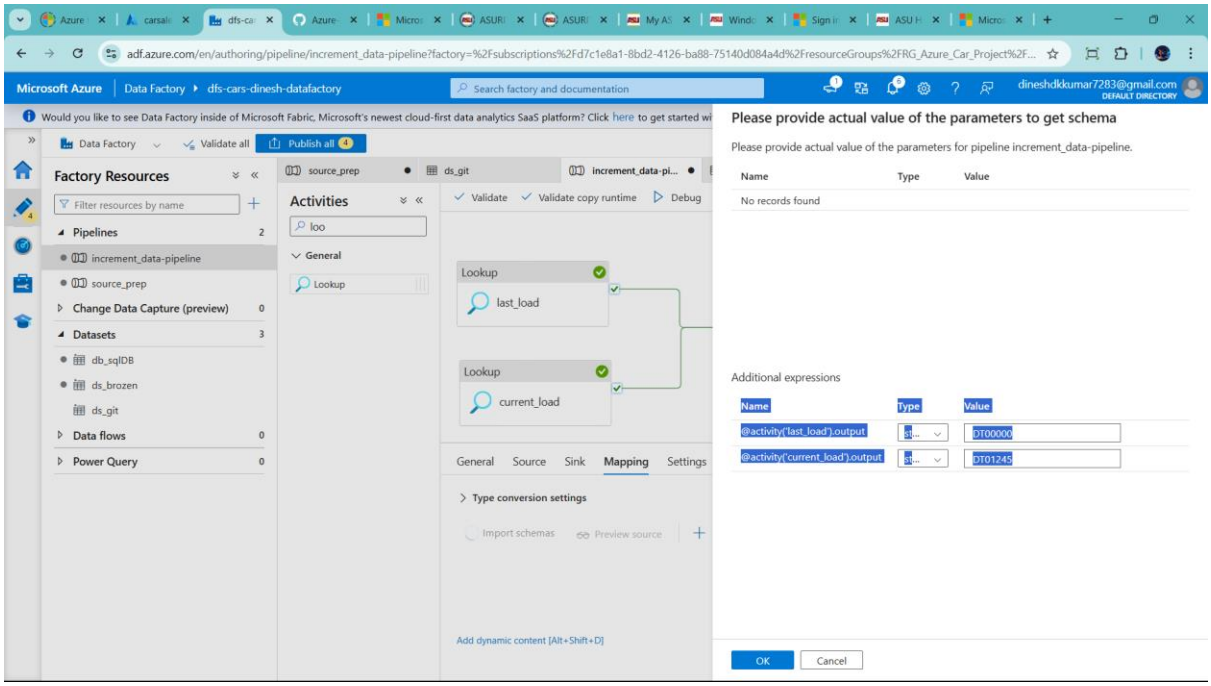


→now go mapping→ click import schemas →add values based on query



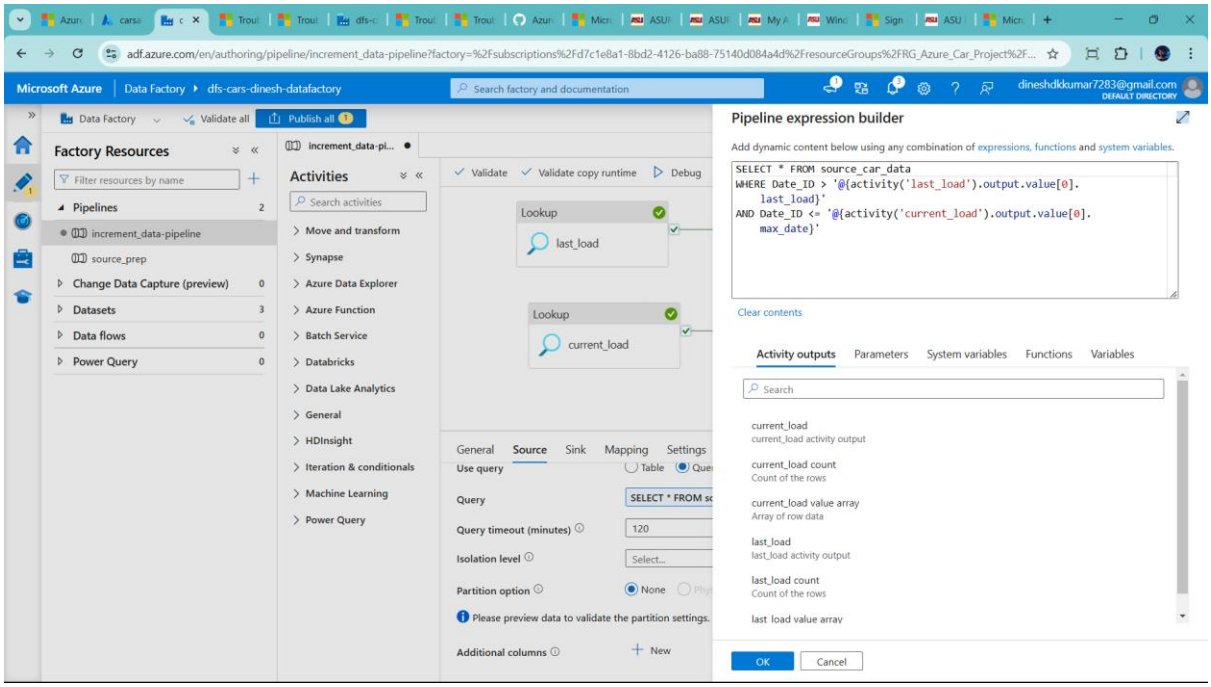

Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home > carsalesdb (carsalesserverdinesh/carsalesdb)
carsalesdb (carsalesserverdinesh/carsalesdb) | Query editor (preview)
SQL database
Search
Login
New Query
Open query
Feedback
Getting started
Overview
Activity log
Tags
Diagnose and solve problems
Query editor (preview)
Mirror database in Fabric (preview)
Resource visualizer
Settings
Data management
Integrations
Power Platform
Security
Intelligent performance
Monitoring
Automation
Help
carsalesdb (dineshdkkumar7283@gmail...
Showing limited object explorer here. For full capability please click here to open Azure Data Studio.
Tables
dbo.source_car_data
dbo.water_table
Views
Stored Procedures
Query 1
Run
Cancel query
Save query
Export data as
Show only Editor
Open Copilot
1 select * from [dbo].[water_table]
Results
Messages
Search to filter items...
last_load
DT00000
Query succeeded | 0s


Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
Home > carsalesdb (carsalesserverdinesh/carsalesdb)
carsalesdb (carsalesserverdinesh/carsalesdb) | Query editor (preview)
SQL database
Search
Login
New Query
Open query
Feedback
Getting started
Overview
Activity log
Tags
Diagnose and solve problems
Query editor (preview)
Mirror database in Fabric (preview)
Resource visualizer
Settings
Data management
Integrations
Power Platform
Security
Intelligent performance
Monitoring
Automation
Help
carsalesdb (dineshdkkumar7283@gmail...
Showing limited object explorer here. For full capability please click here to open Azure Data Studio.
Tables
dbo.source_car_data
dbo.water_table
Views
Stored Procedures
Query 1
Run
Cancel query
Save query
Export data as
Show only Editor
Open Copilot
1 select * from [dbo].[water_table]
2
3 select max(Date_ID) from [dbo].[source_car_data]
Results
Messages
Search to filter items...
DT01245
Query succeeded | 0s

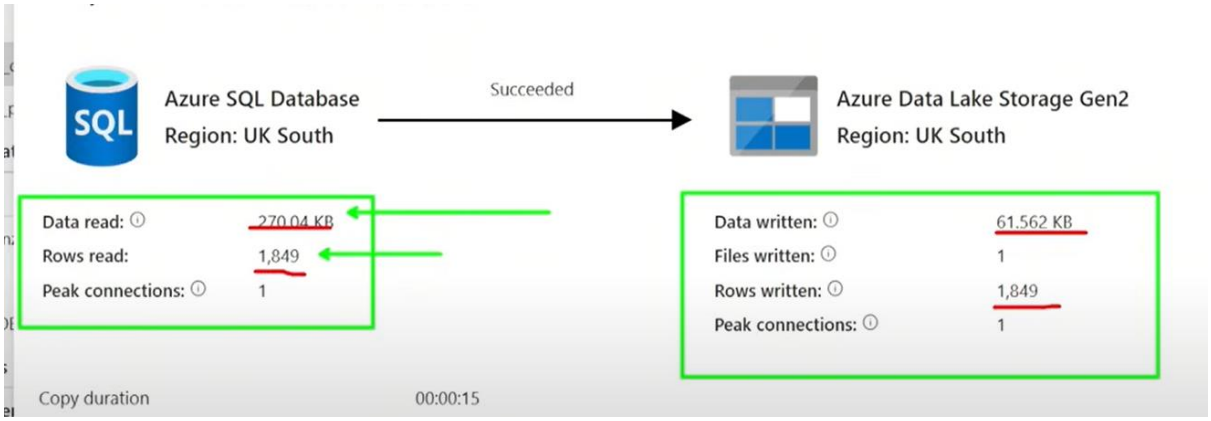


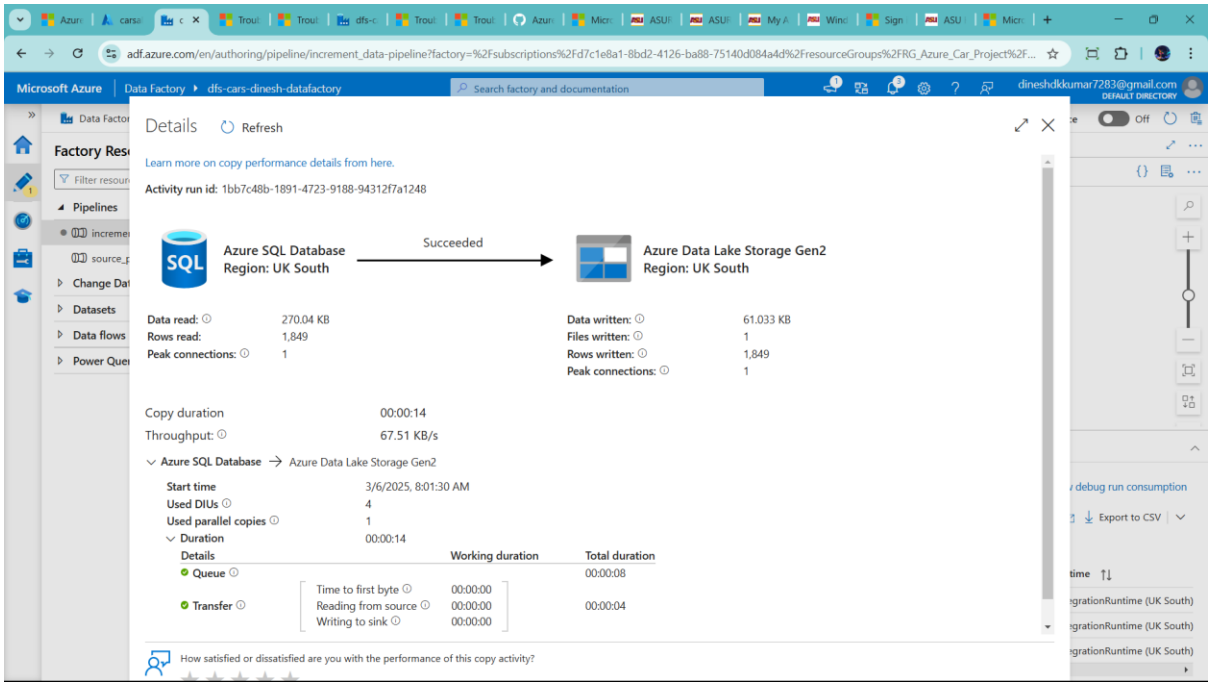
--click ok →they show error change sql query

--chage query as per my respectivity

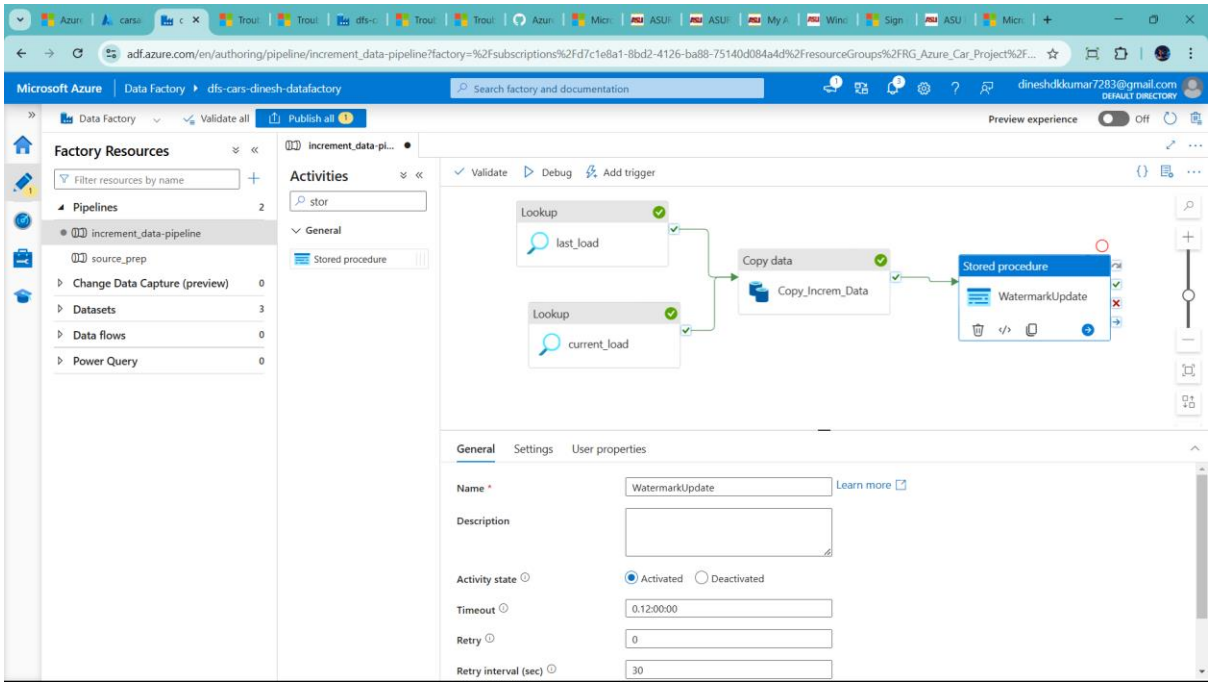


Debug it all activity



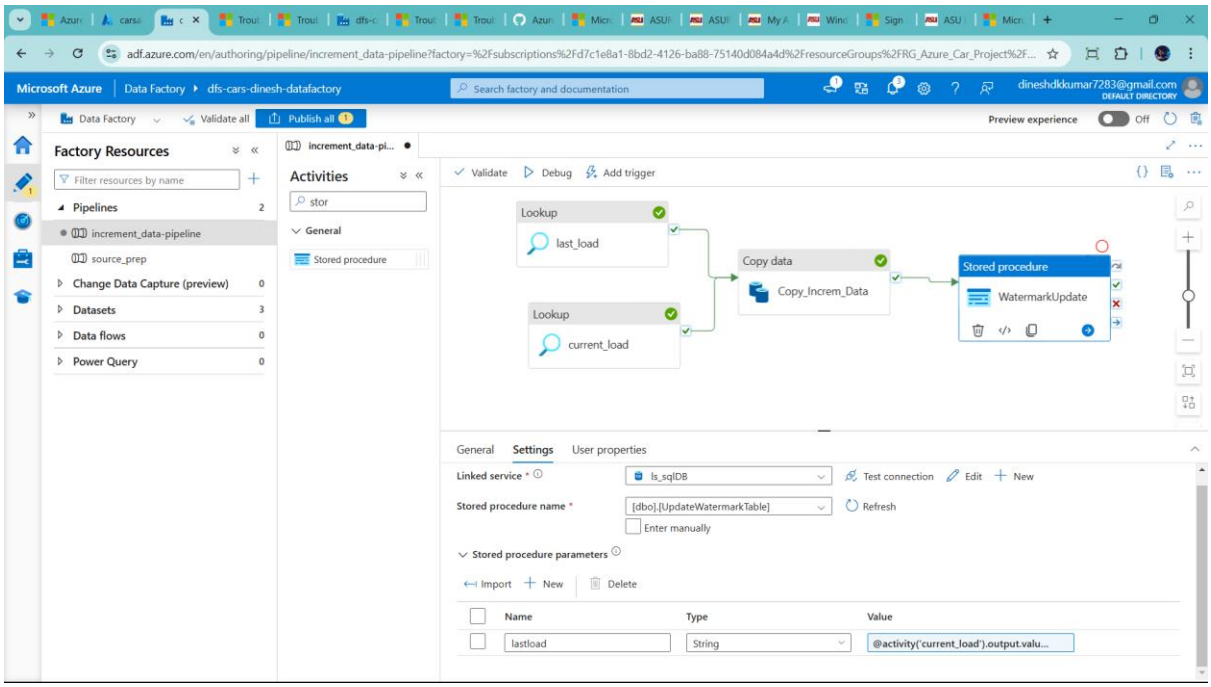


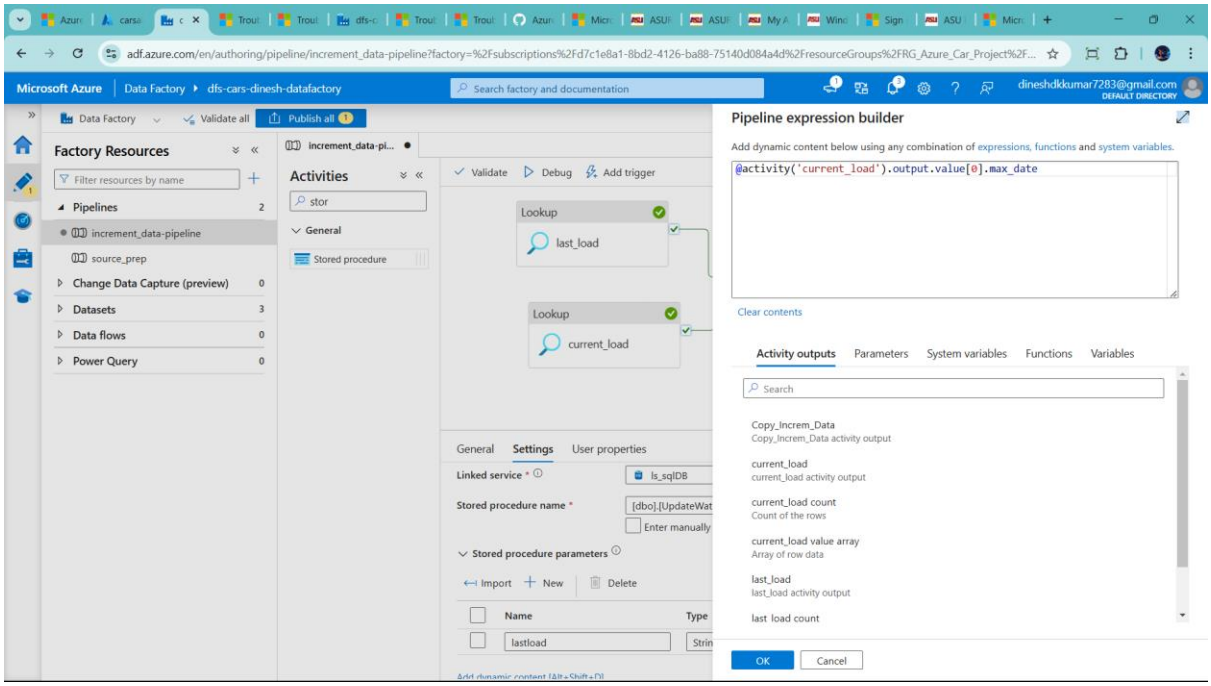
Add store procedure

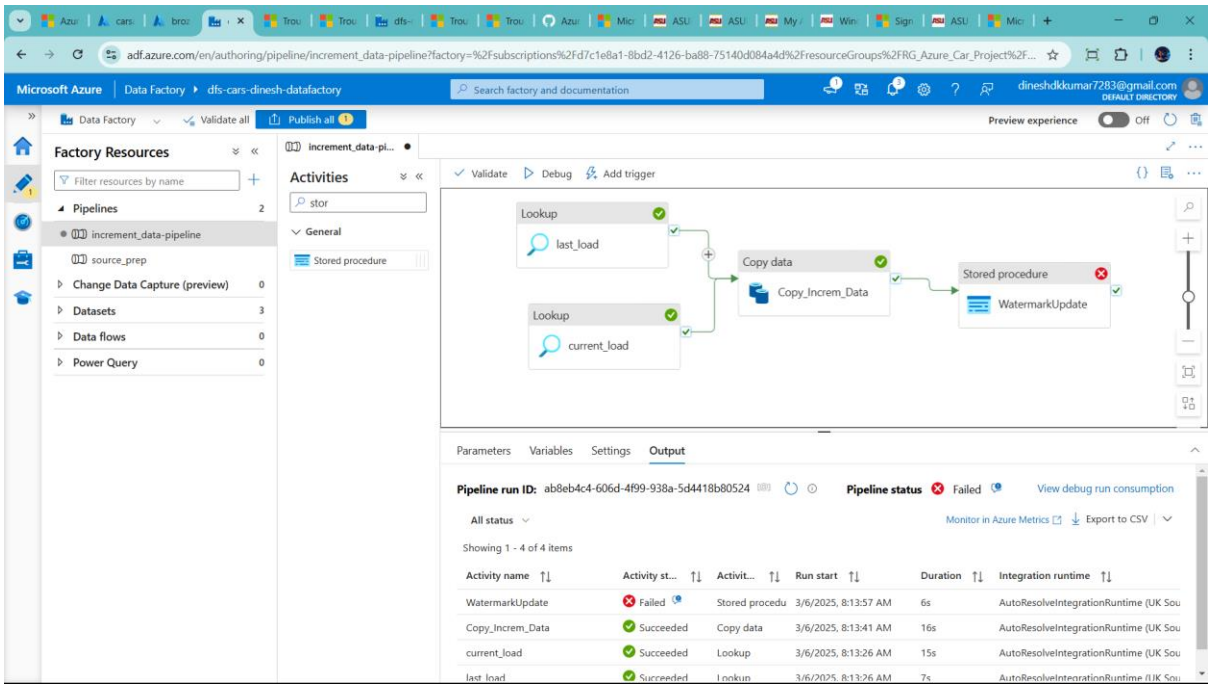


Add in setting section

Linked service, storeprocedure name, click import parameter



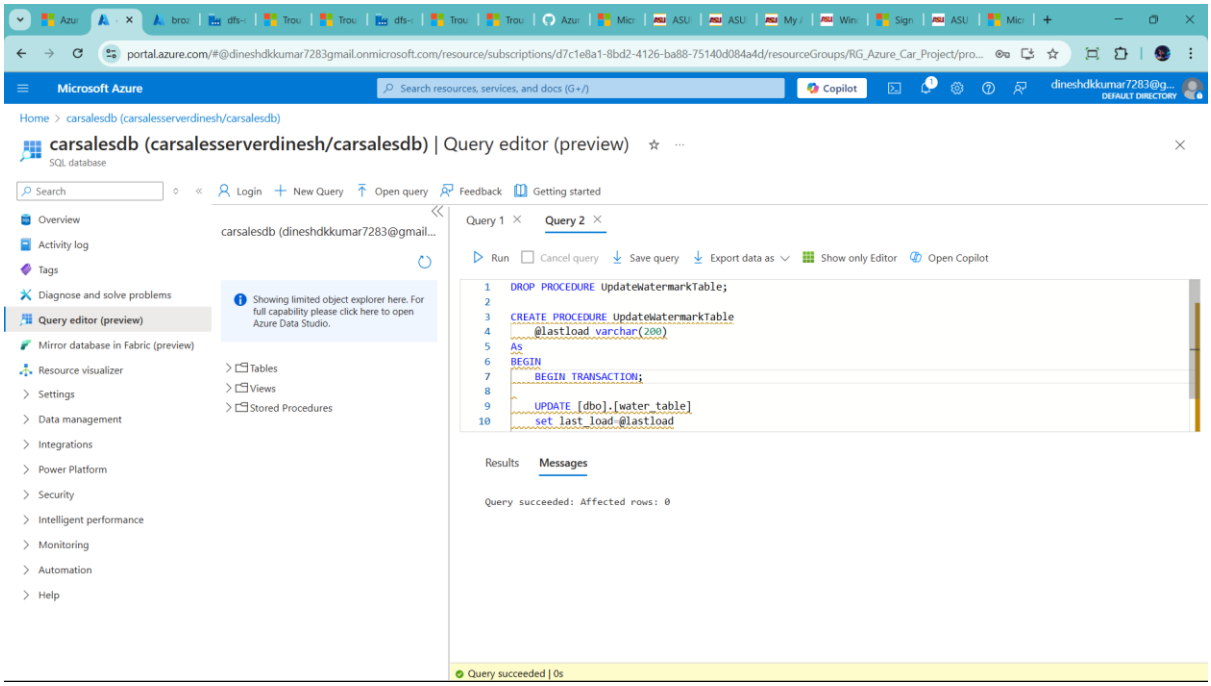




Handle this error

Go edit store procedure

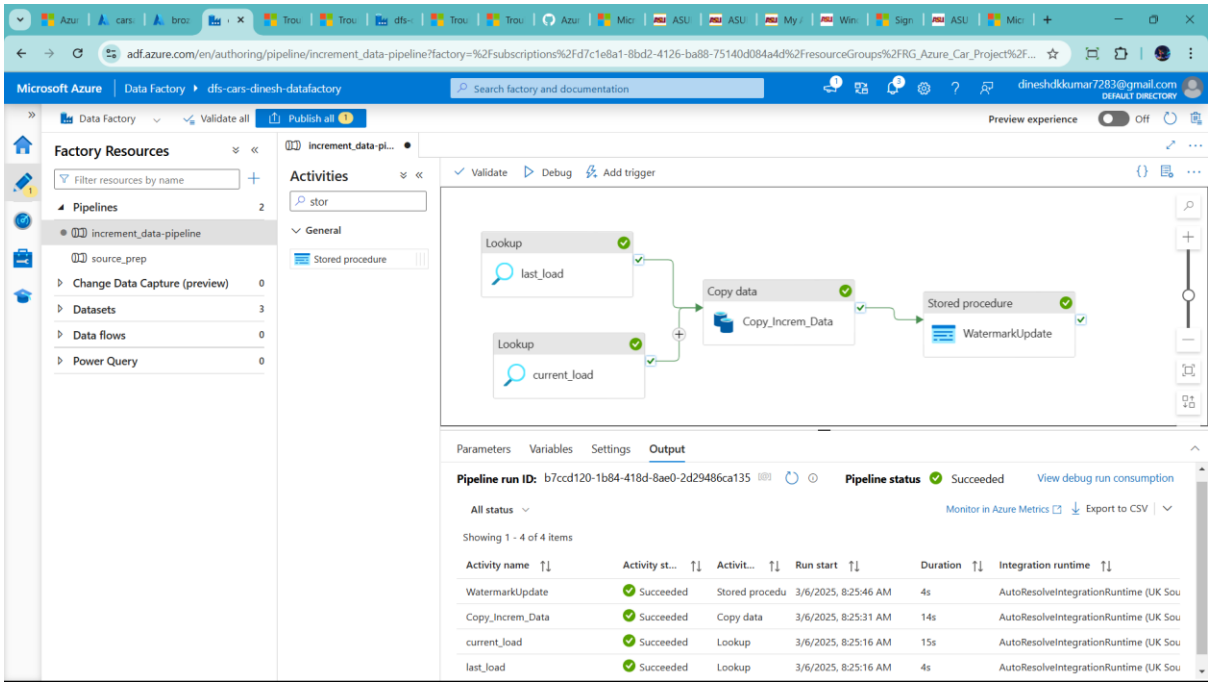
DROP PROCEDURE AND RECREATE



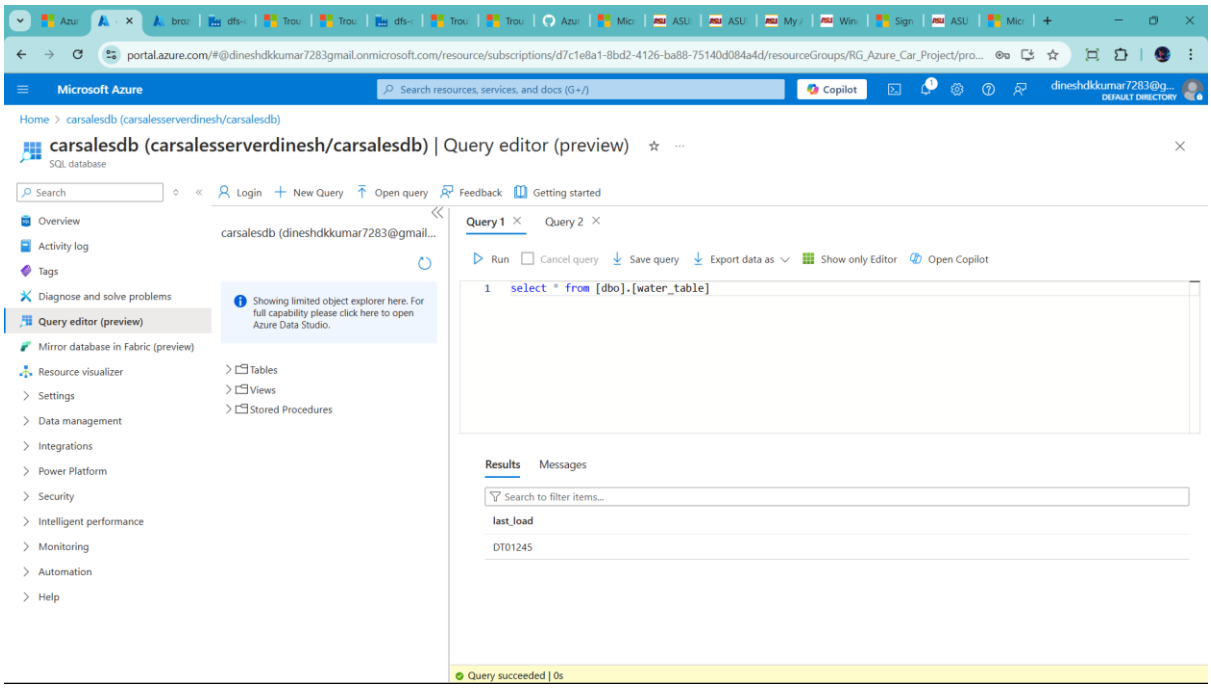

```
DROP PROCEDURE UpdateWatermarkTable;


CREATE PROCEDURE UpdateWatermarkTable
  @lastload varchar(200)
As
BEGIN
  BEGIN TRANSACTION;


  UPDATE [dbo].[water_table]
  set last_load=@lastload


  commit transaction;
END;
```

--delete data lake rawdata folder

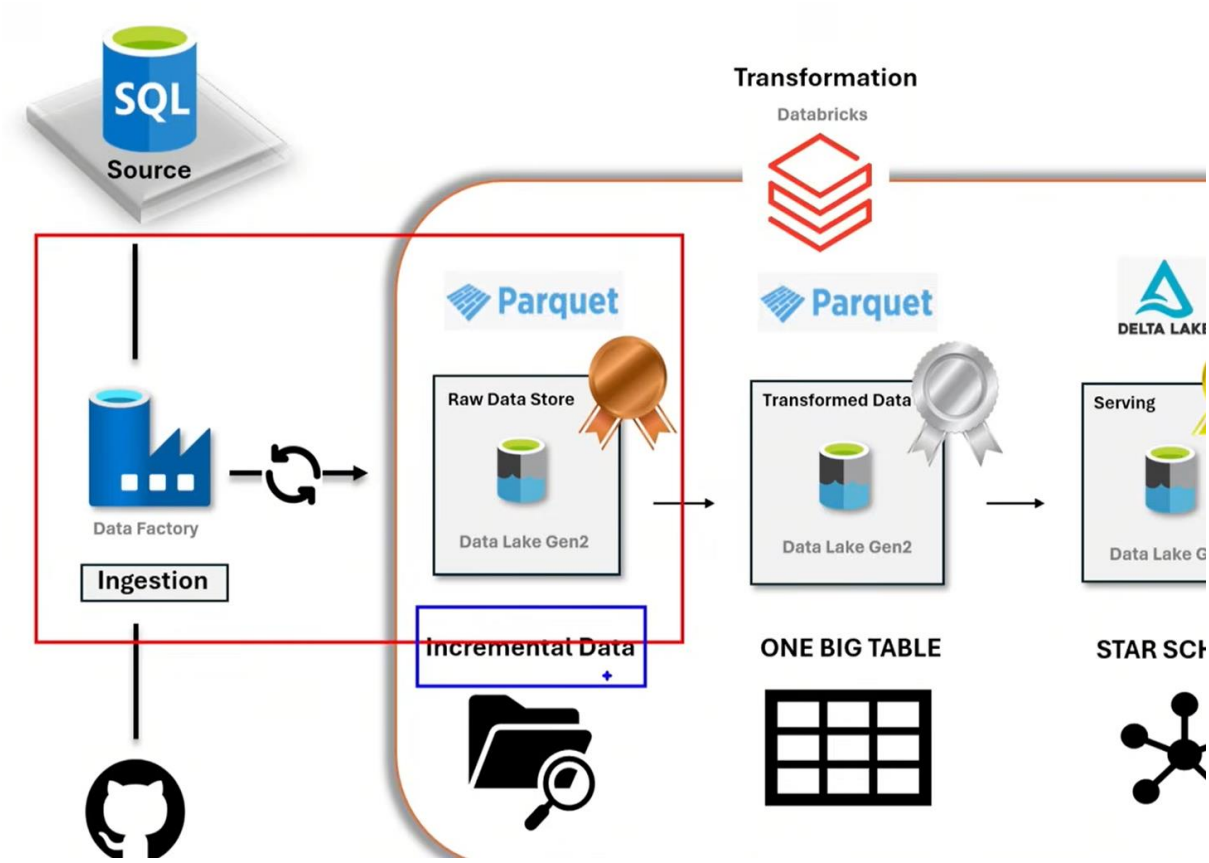
And rerun pipeline



Successful updated watermark table

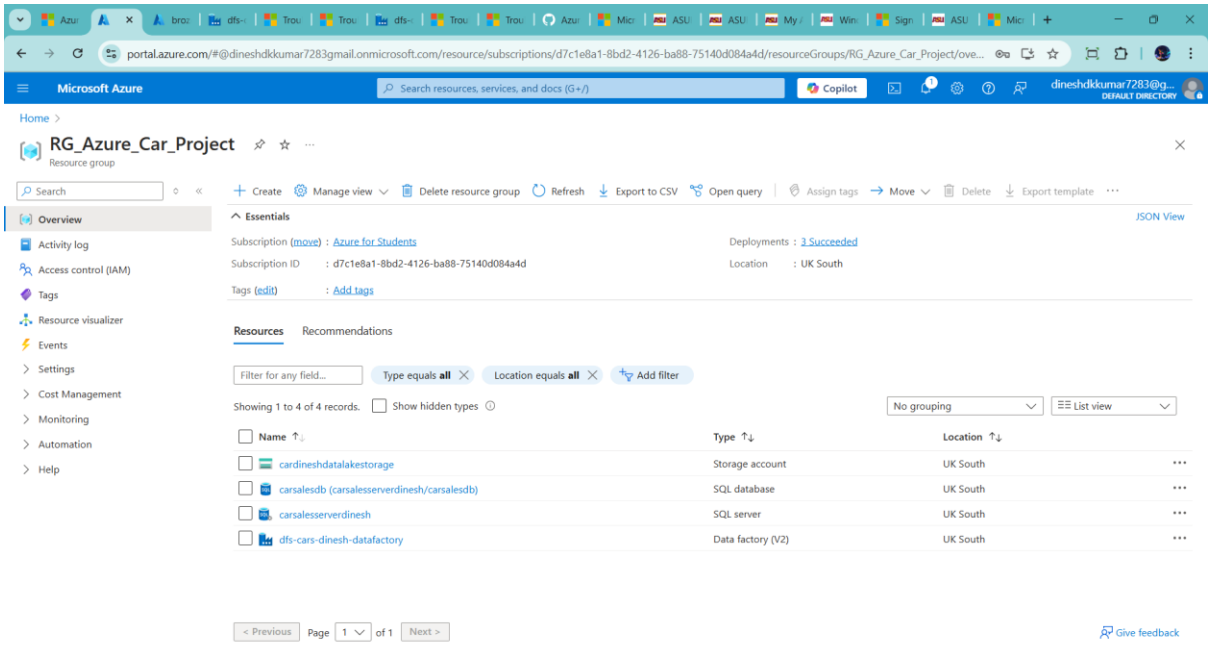


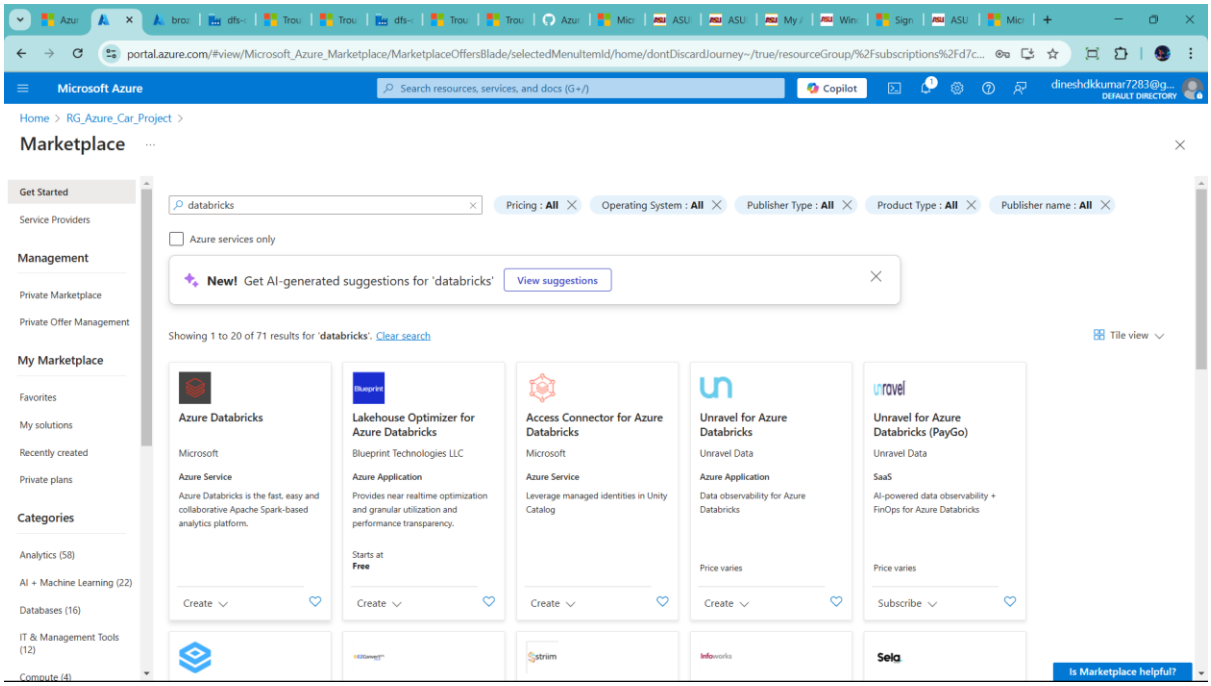
Now Success full Complete First part


SQL
Source
Transformation
Databricks
Parquet
Parquet
DELTA LAKE
Raw Data Store
Data Lake Gen2
Transformed Data
Data Lake Gen2
Serving
Data Lake G
Data Factory
Ingestion
Incremental Data
ONE BIG TABLE
STAR SCH

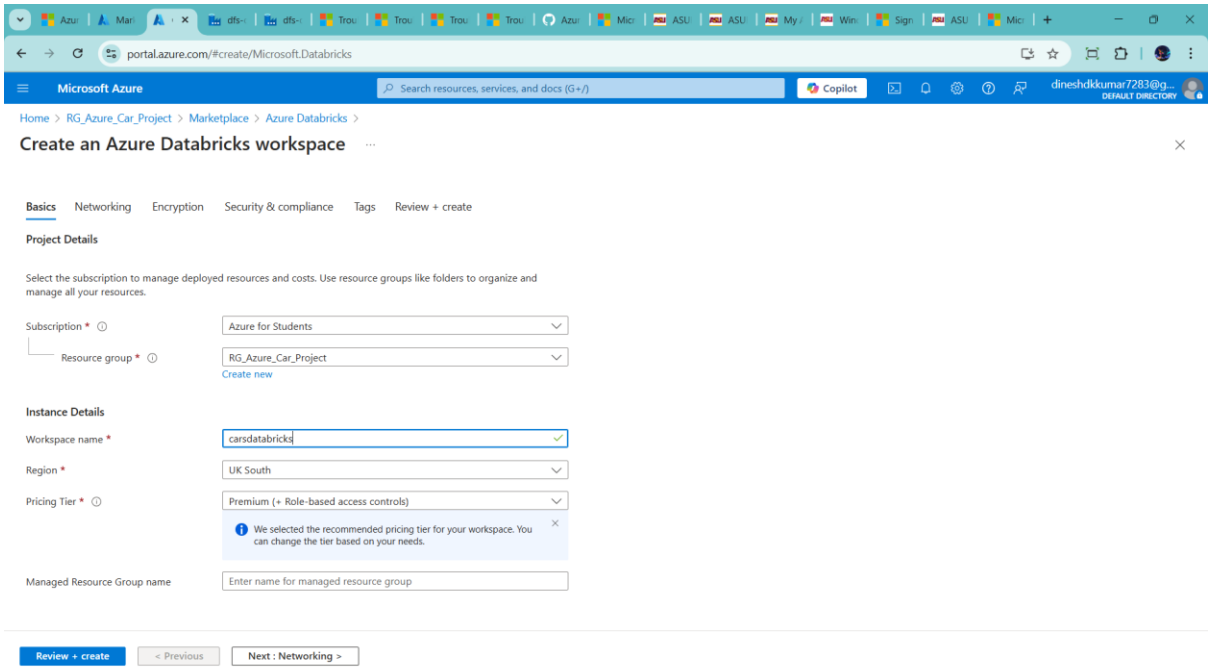
Let Start Create Databricks Resource

Go resource → click create

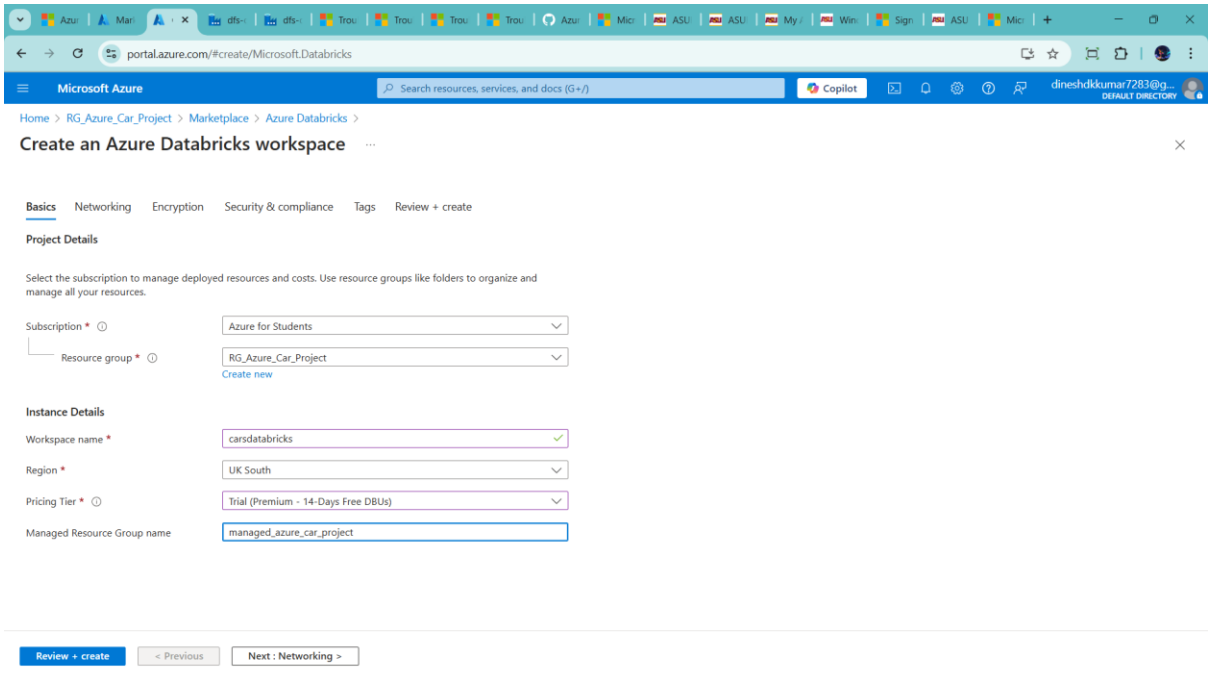




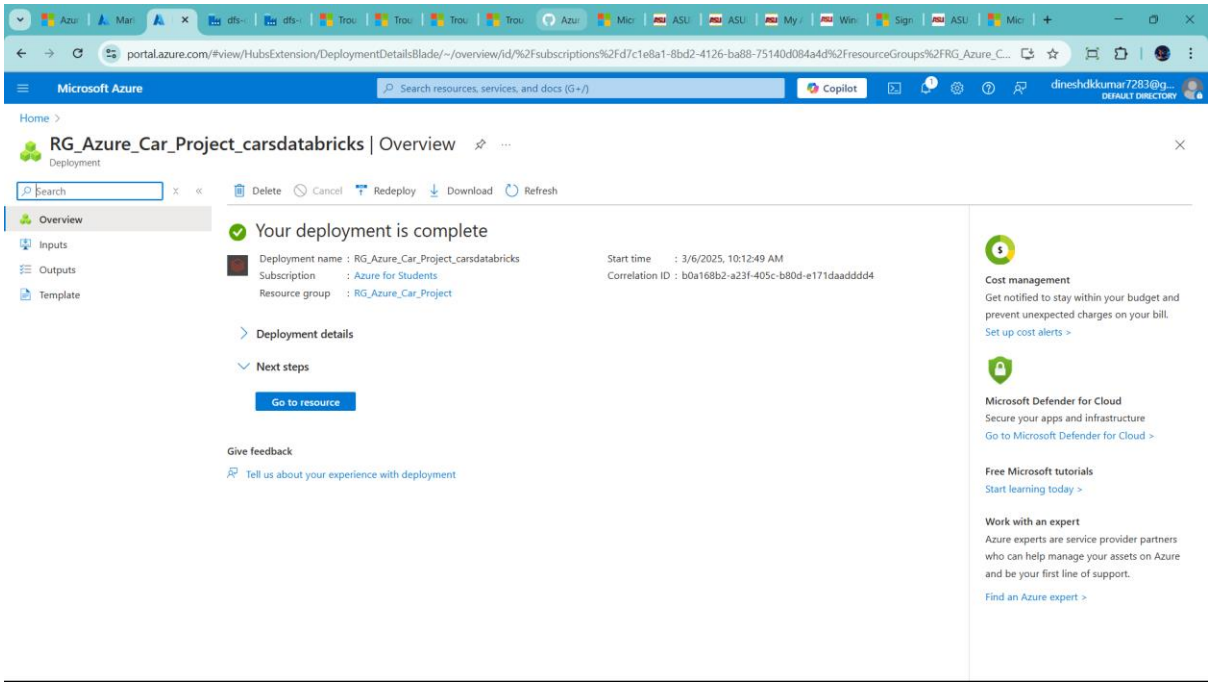
➔ Cars databricks



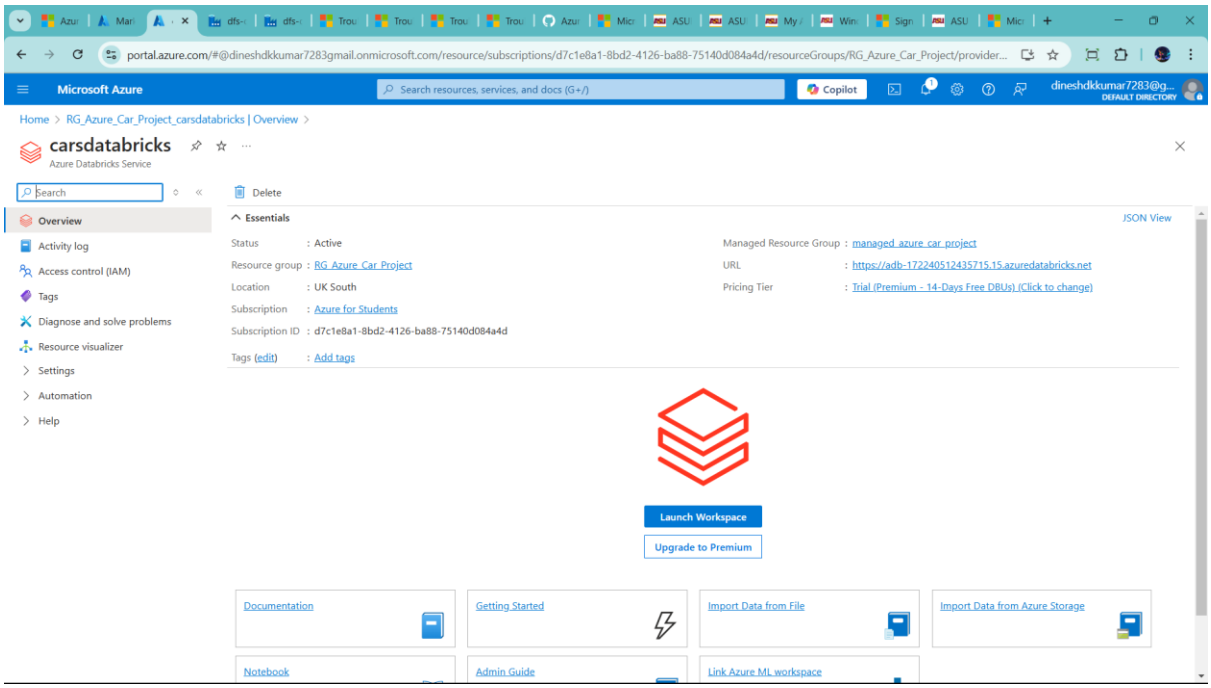
➔ Put workspace name
➔ Add change price Tier → trail 14 days
➔ Add manager resource group →(use for catlog,etc)
➔ Click review+create



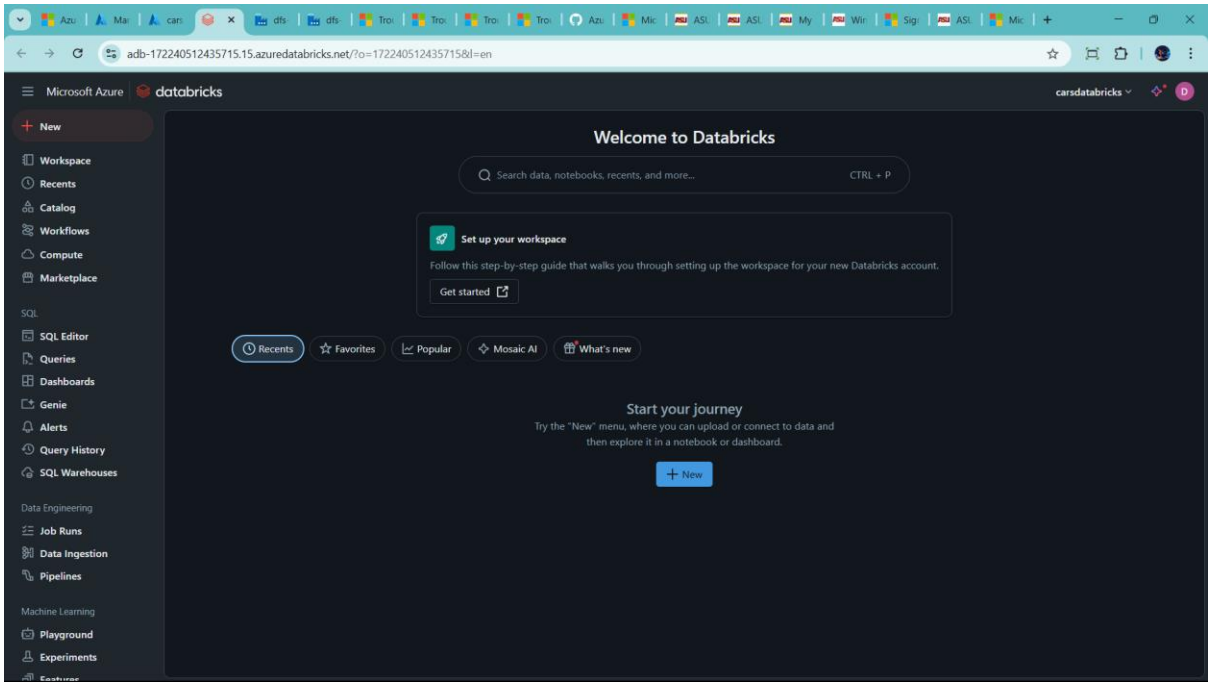
→click create and completed



Click go to workspace



Click launch workspace



→workspace →manage folder,notebook and maintain hierchery

→Recents → view all the recent items →notebook,delta table,live table and more in recent tap

→Catalog → manage a Databases/ table,schema and volumn

→Workflows → it like ETL frameworks ,build pipeline low/no code,

→compute →create a clusters

→SQL → it is section use for data analysis

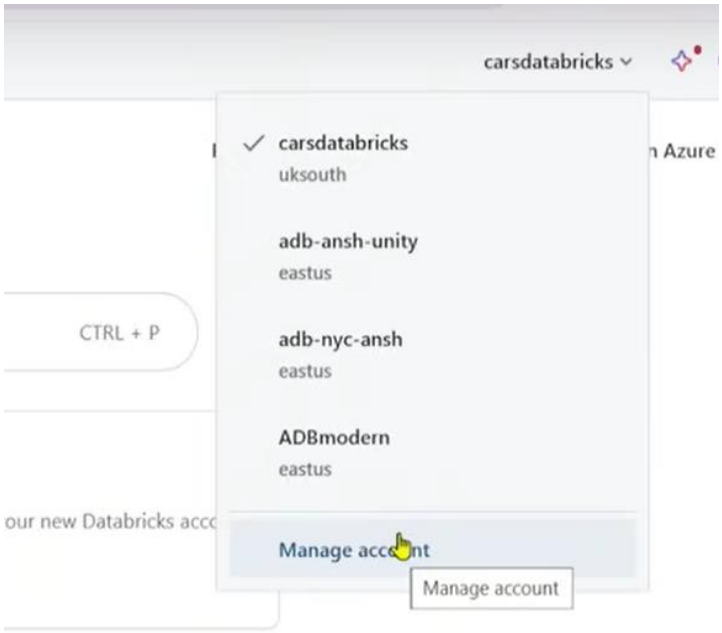
→data engineering → have 3 frameworkd

- ➔ One is pipelines/Delta live tables →it is ETL flow
- ➔ Machine learning→ build models
- ➔ Marketplace →use service for 3rd party like azure,more

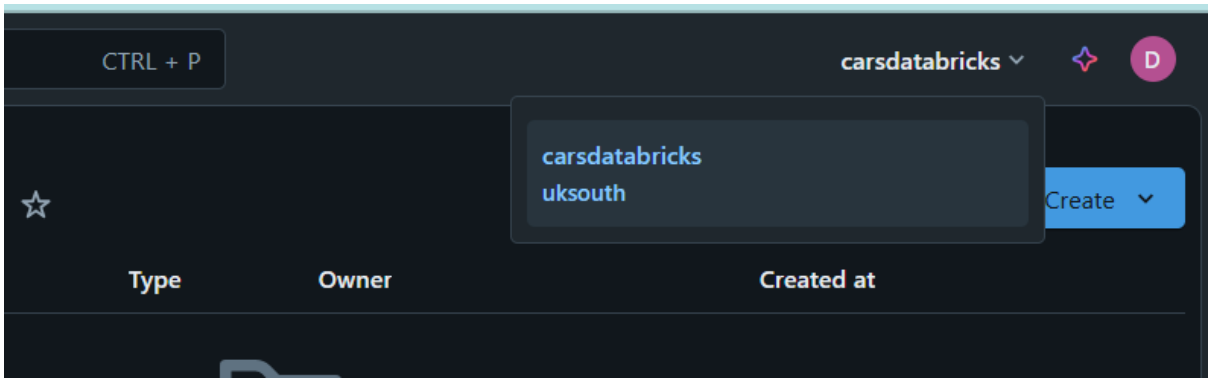
First Create Unity Metastore

→but in databricks not able to create unity metastore

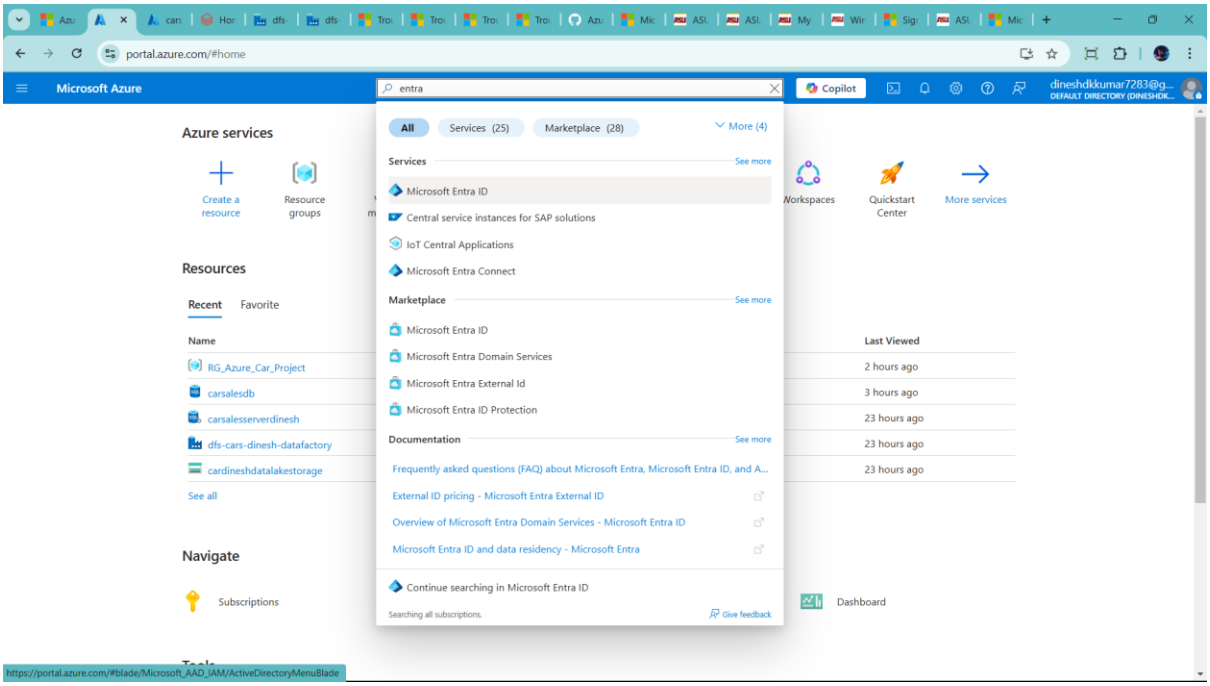
→ first change in Admin in databricks

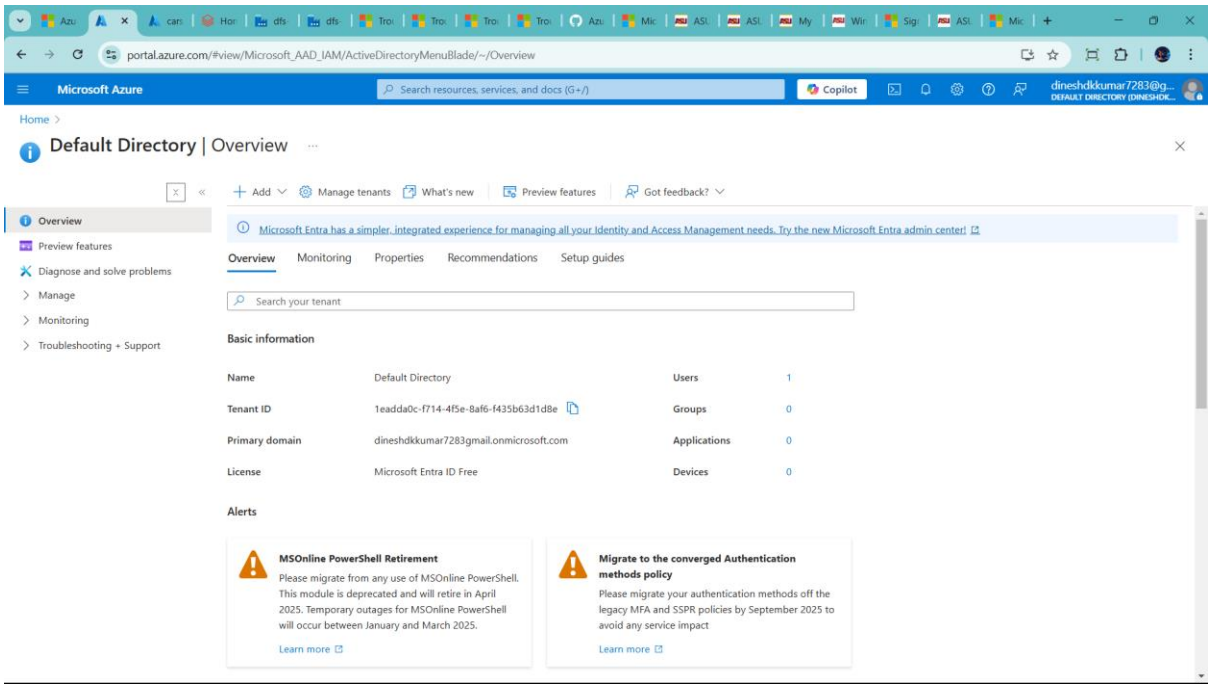


- **But over databricks not show manage account option**
- **Let how to enable**

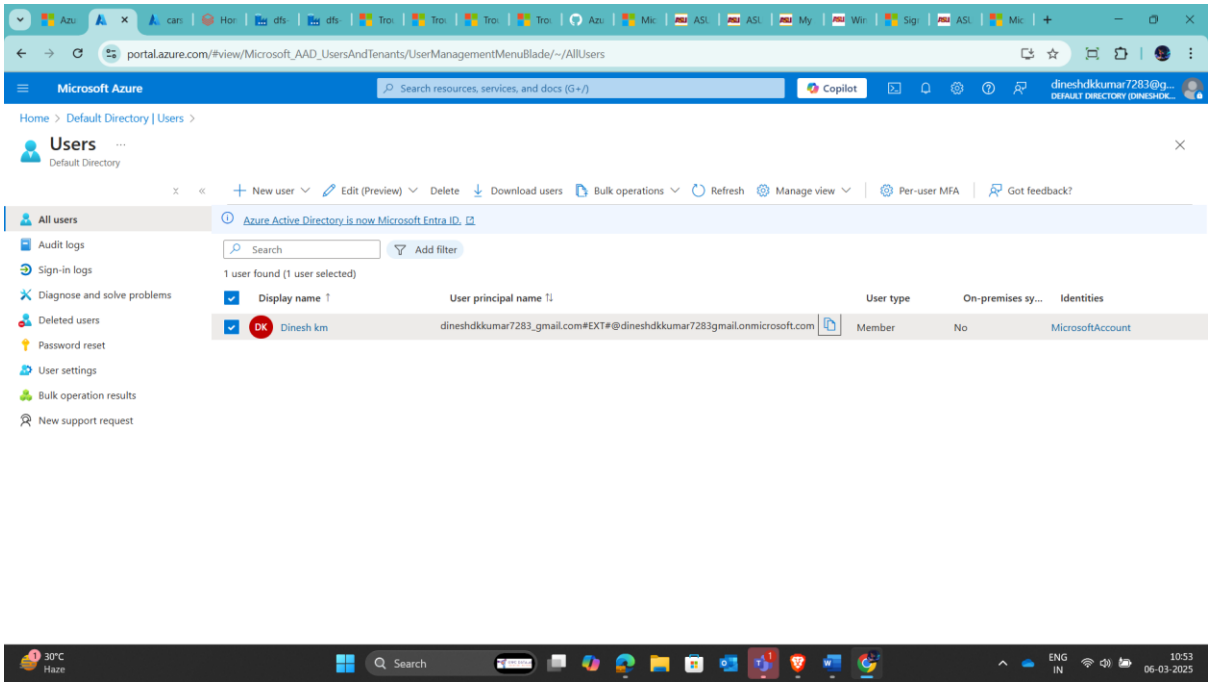


- First go to azure account → click Microsoft entra id





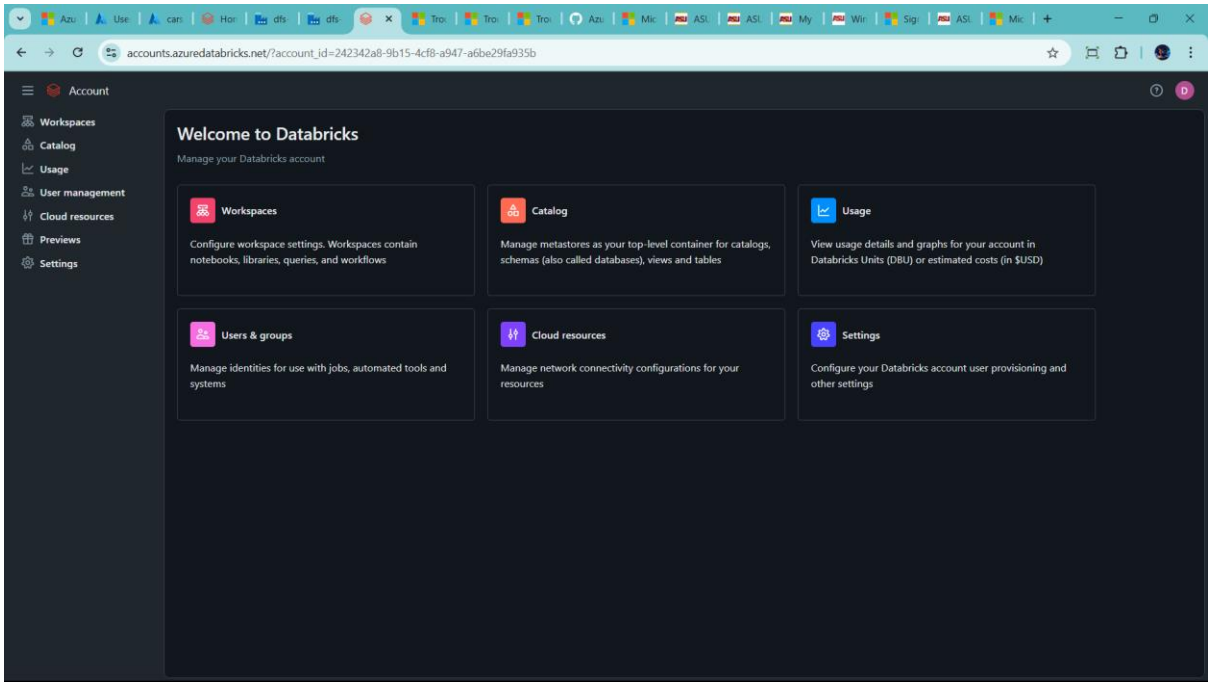
→go to manage →user →copy this user mail id with #tag



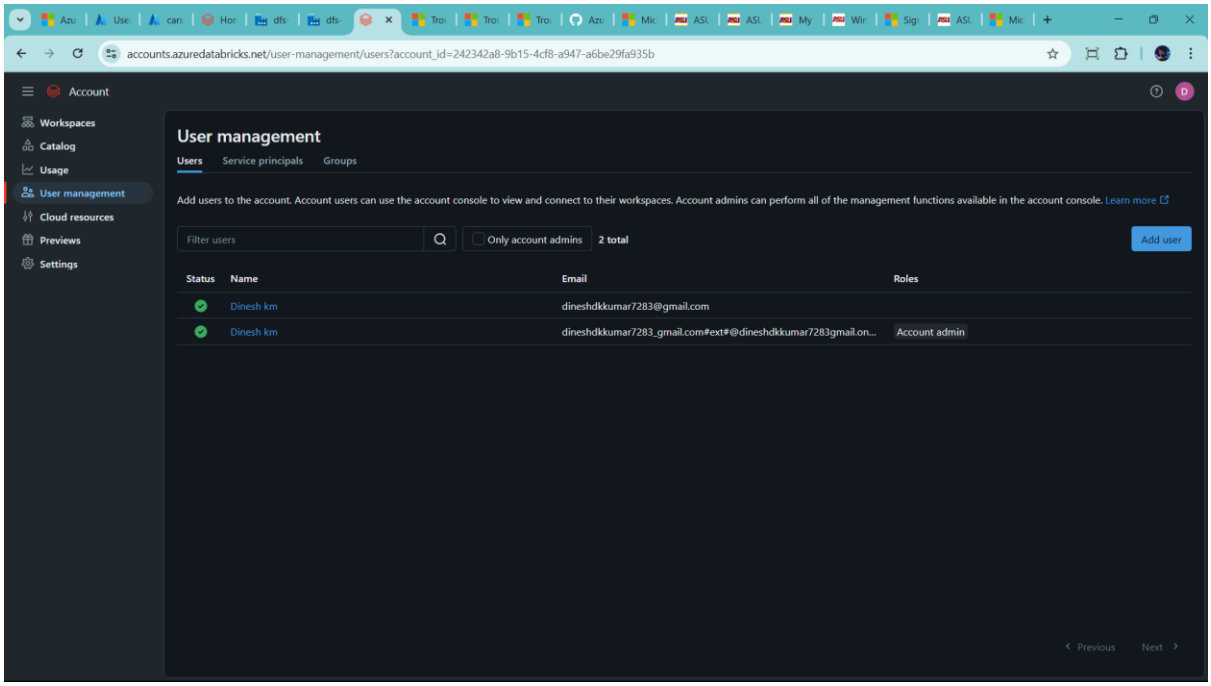
Login in another portal →accounts.azuredatabricks.net

→first reset the password → now login in the #tag mail id

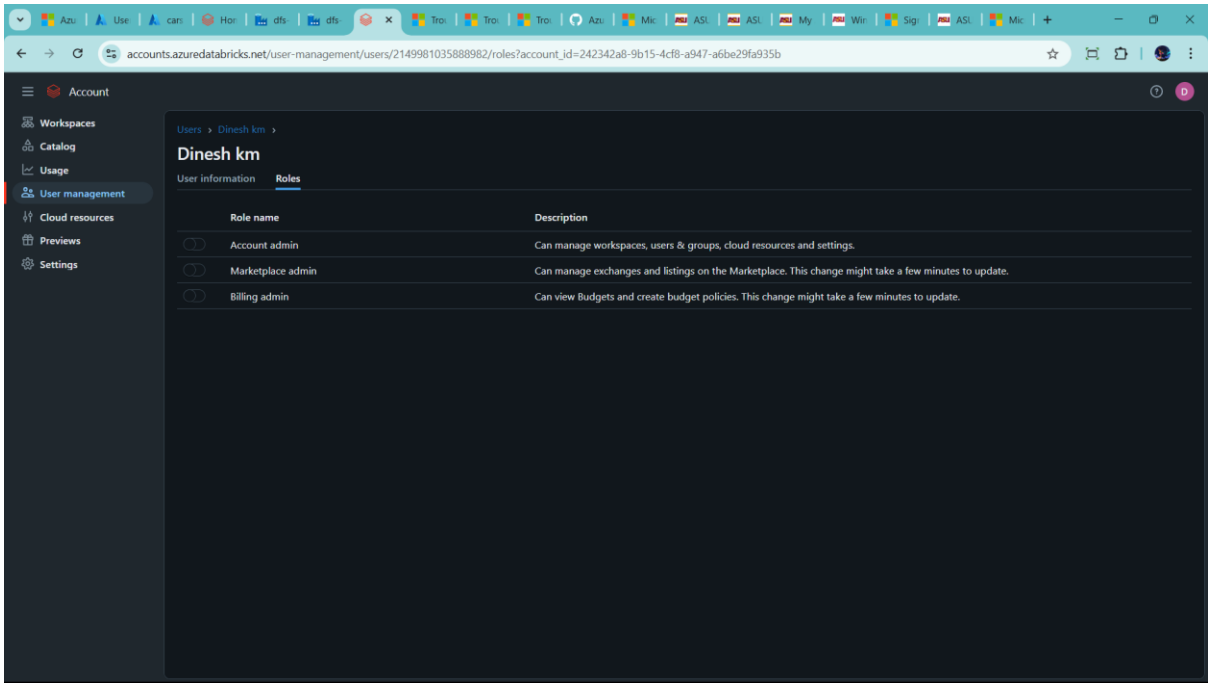
→success open Admin account



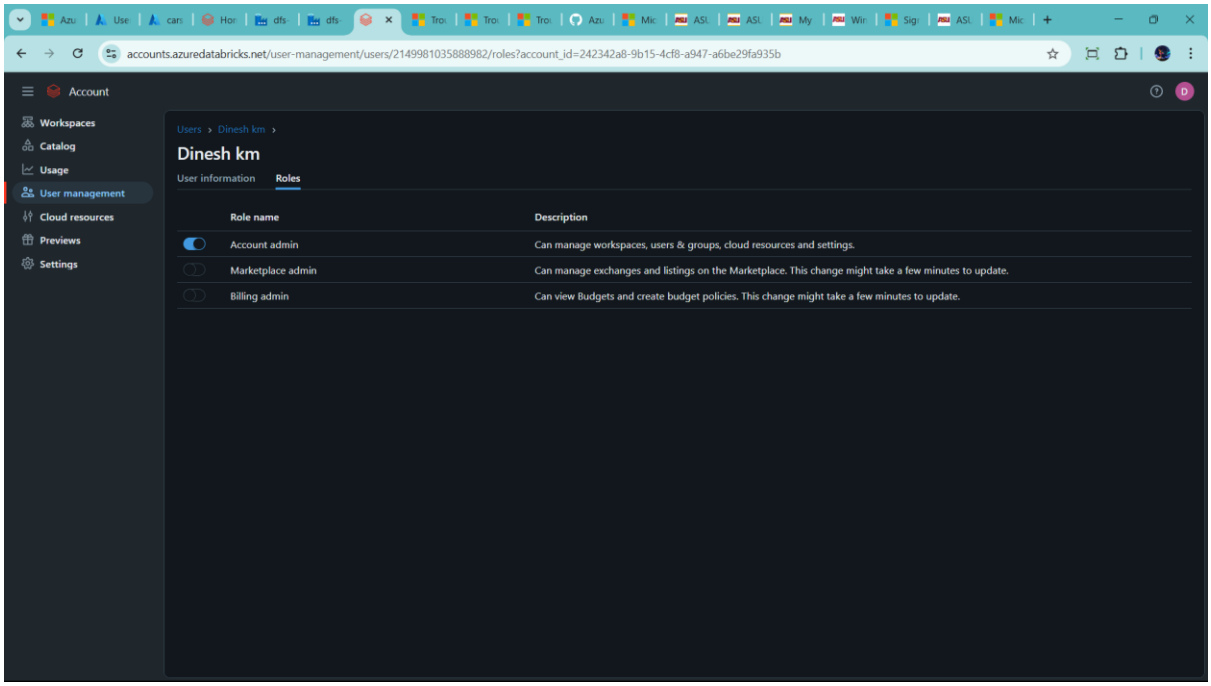
→go to user management section

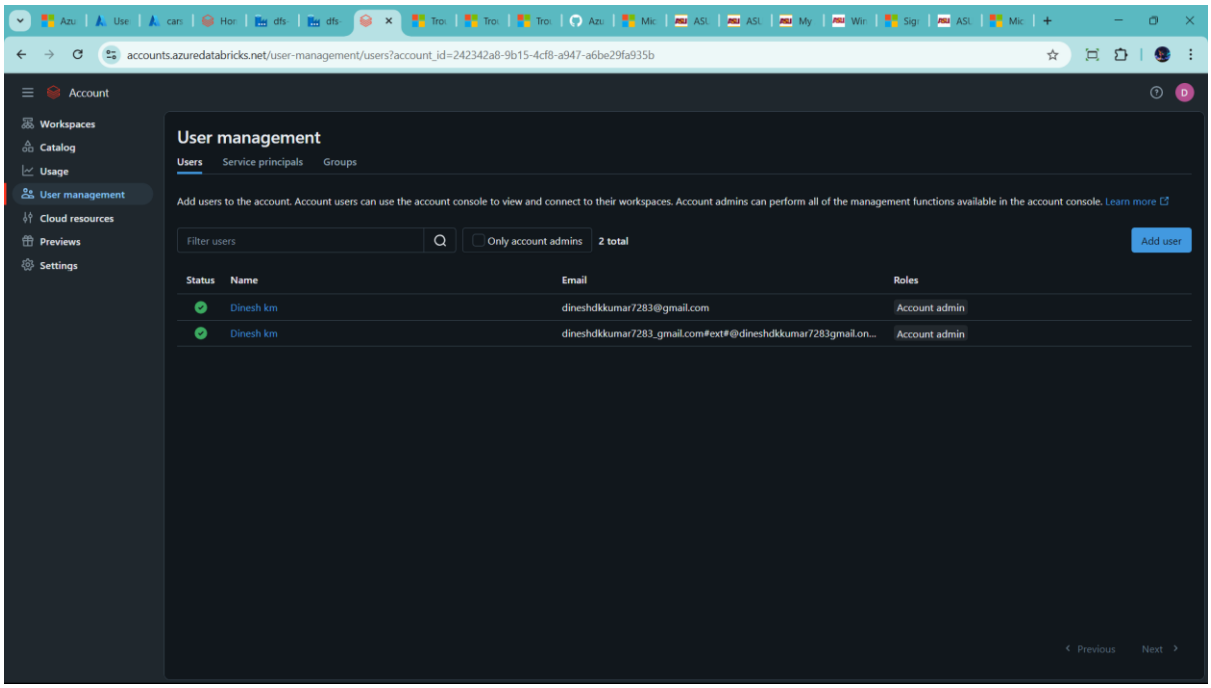


→add new user or already present → click the name

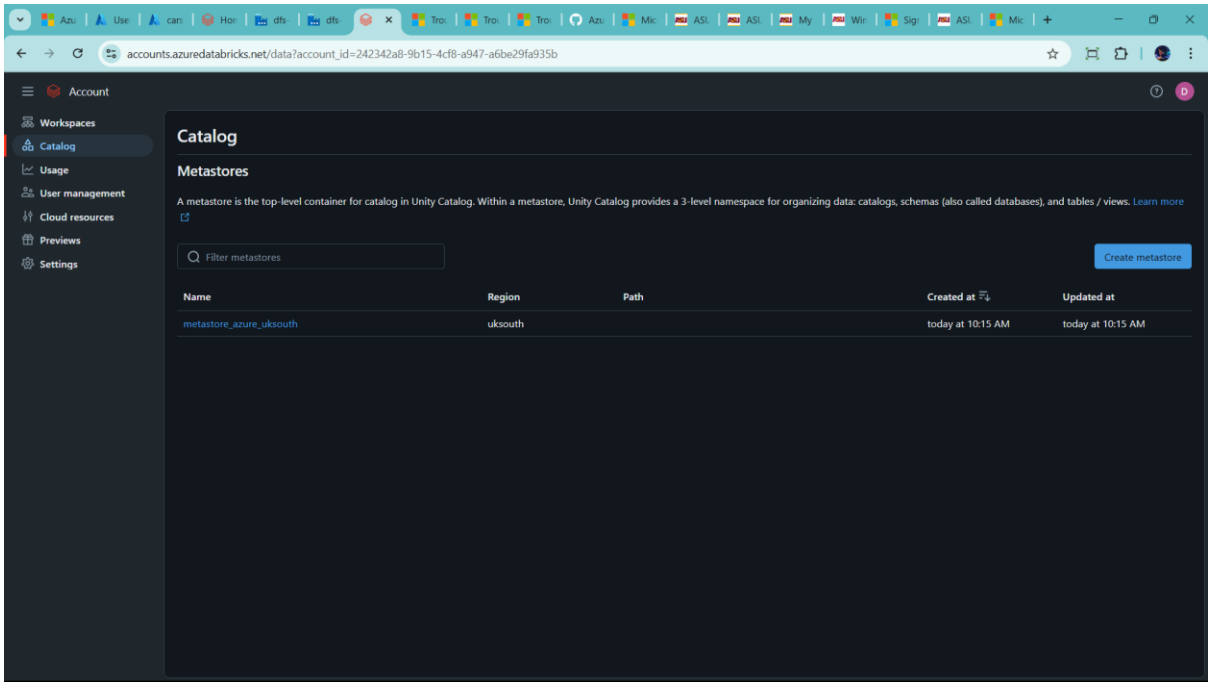


→click Admin account active





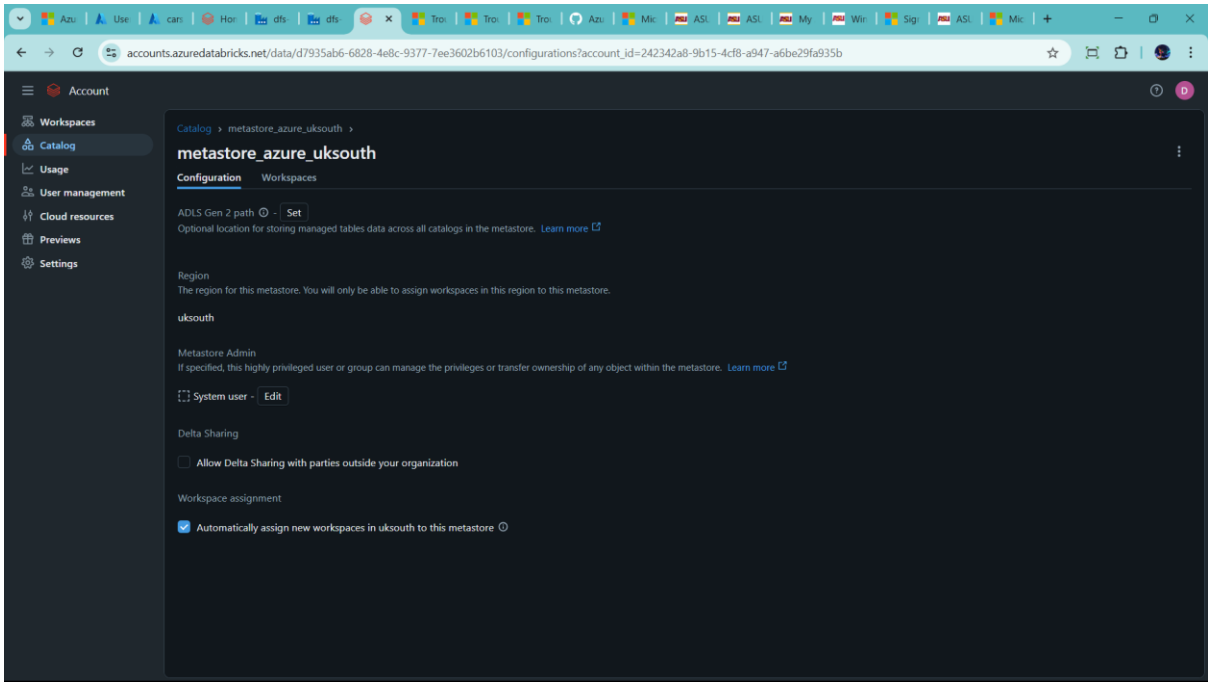
Now check catelog in admin console →the already create unity catelog



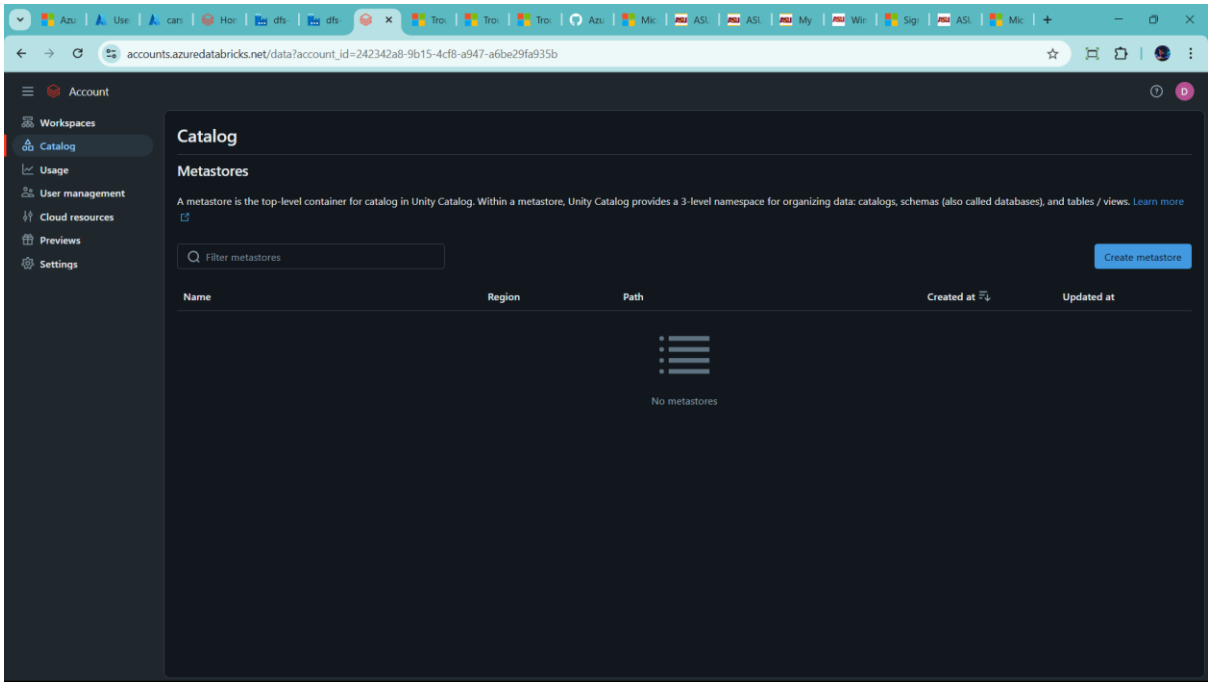
→now I deleted my unity catalog

-→only one region same unity catelog possible to create

→first click the name present in catalog

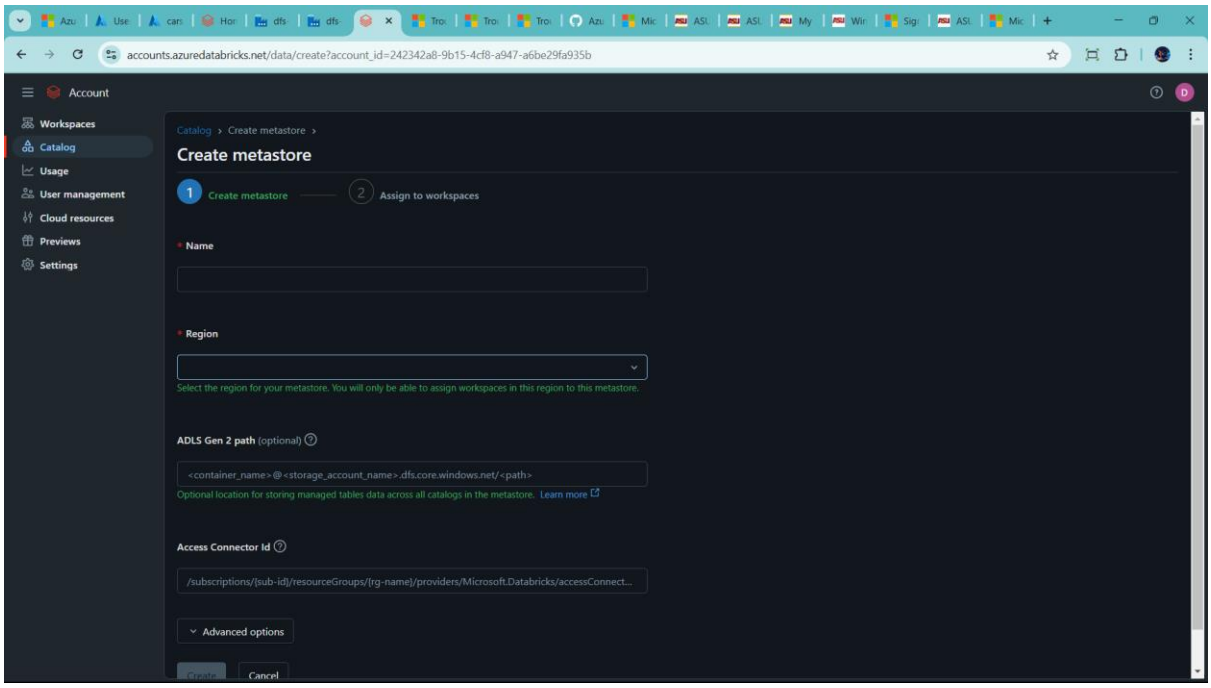


Click 3 dots → delete



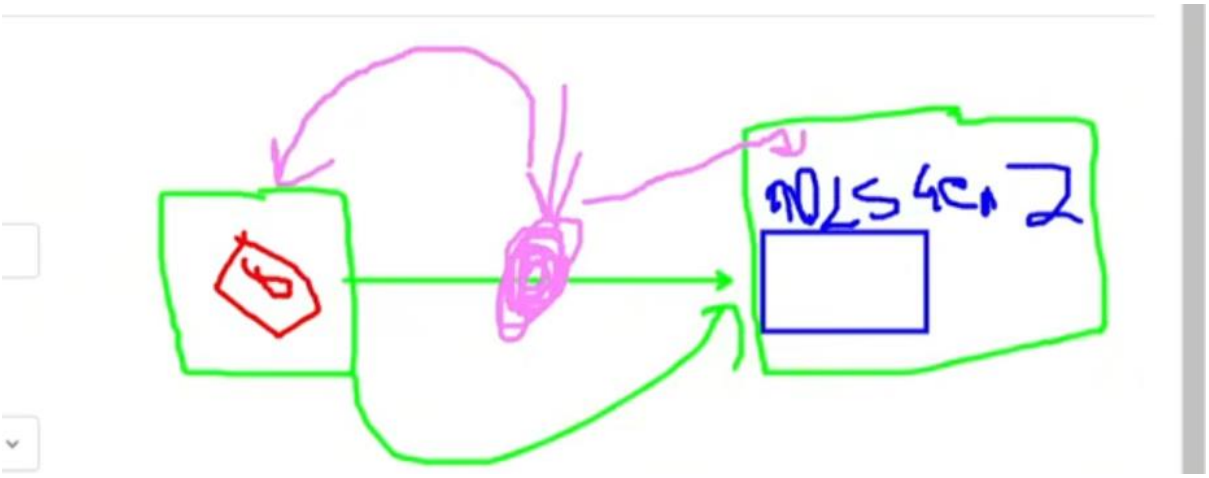
➔ Now create manually

➔ Click create metasote



→fill it

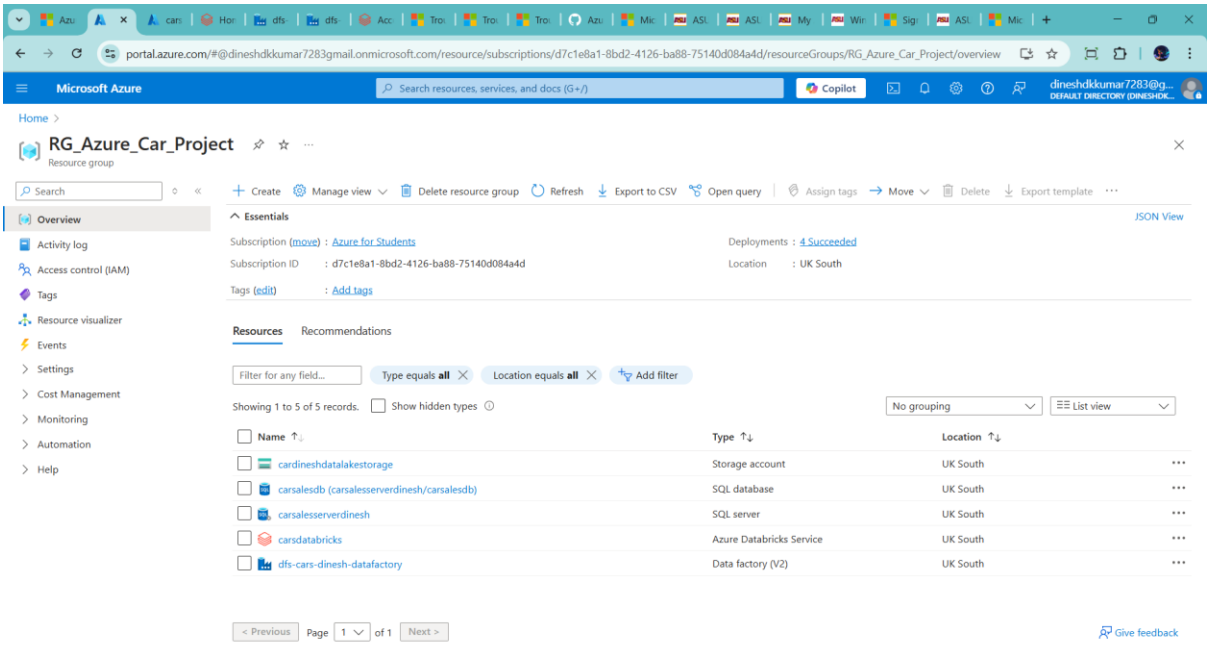
→before databricks → connector → ADLS Gen 2



➔ Databricks have permission to use it datalake
➔ That use access connector id in account admin

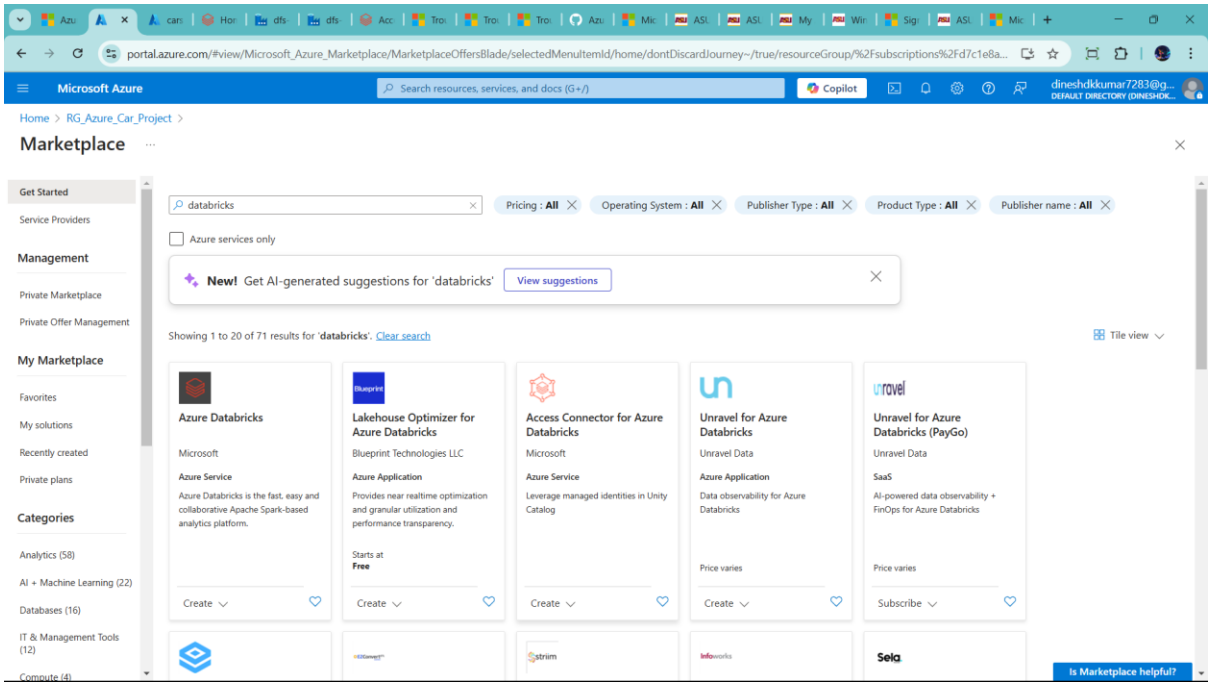
→Let create Access connector in azure account

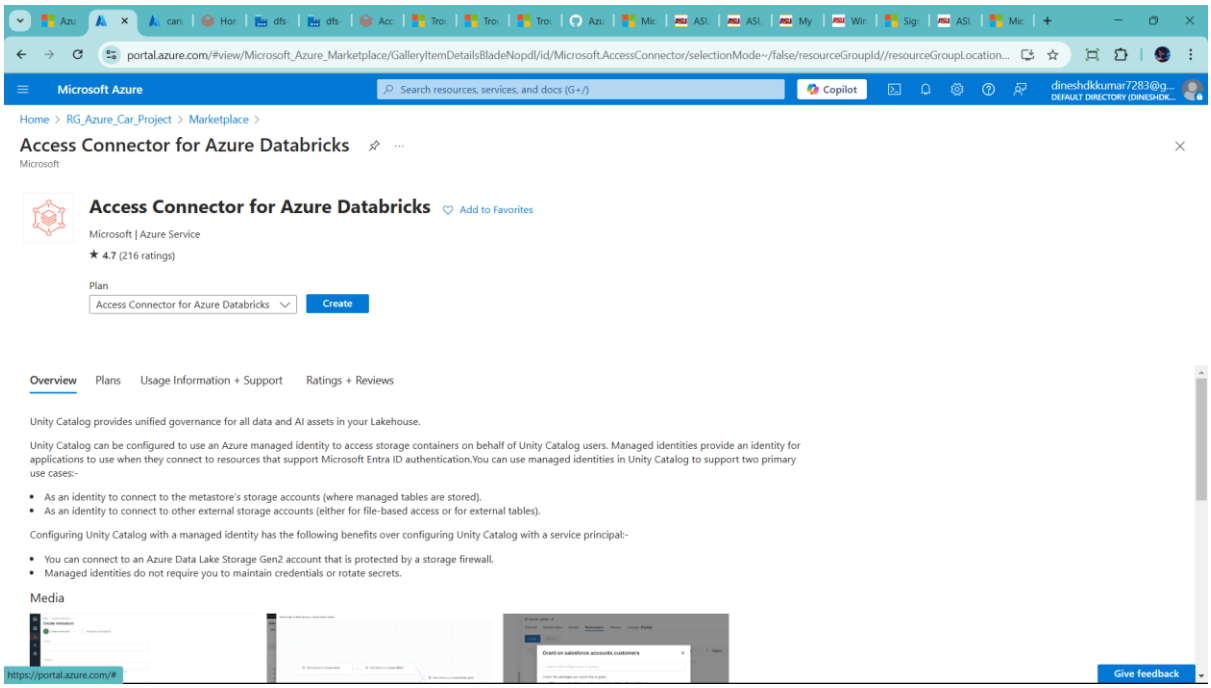
→go to azure resource group

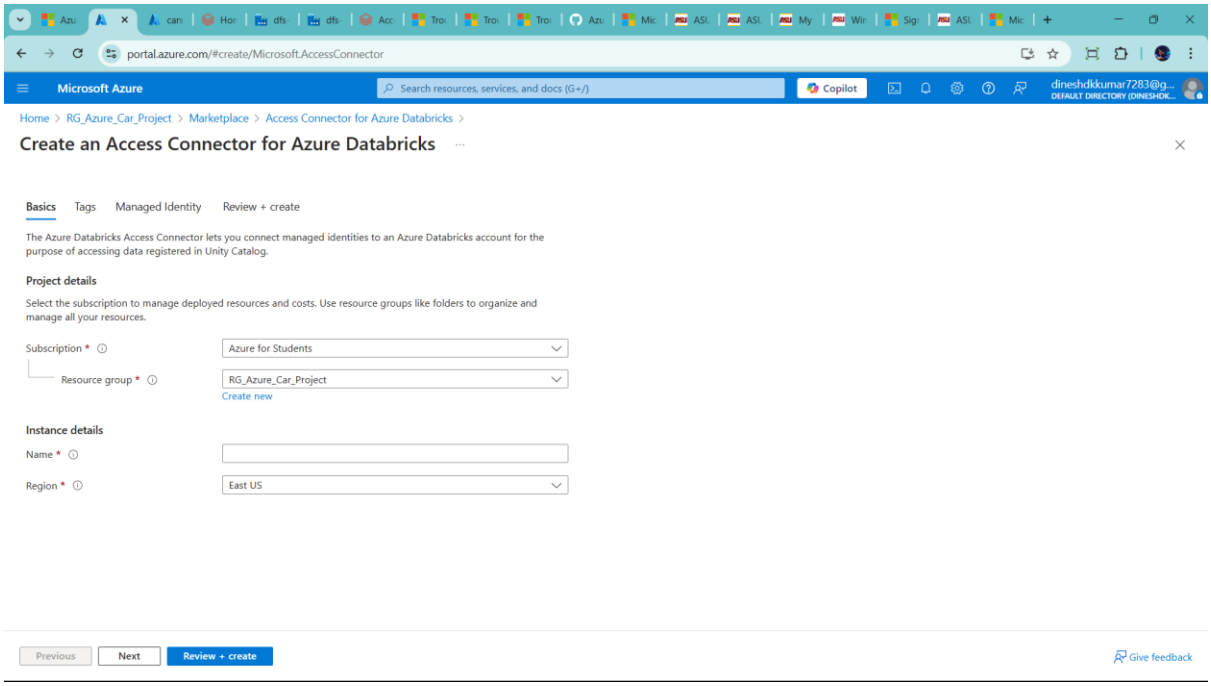


--click create →search databricks → find access connector for azure databricks

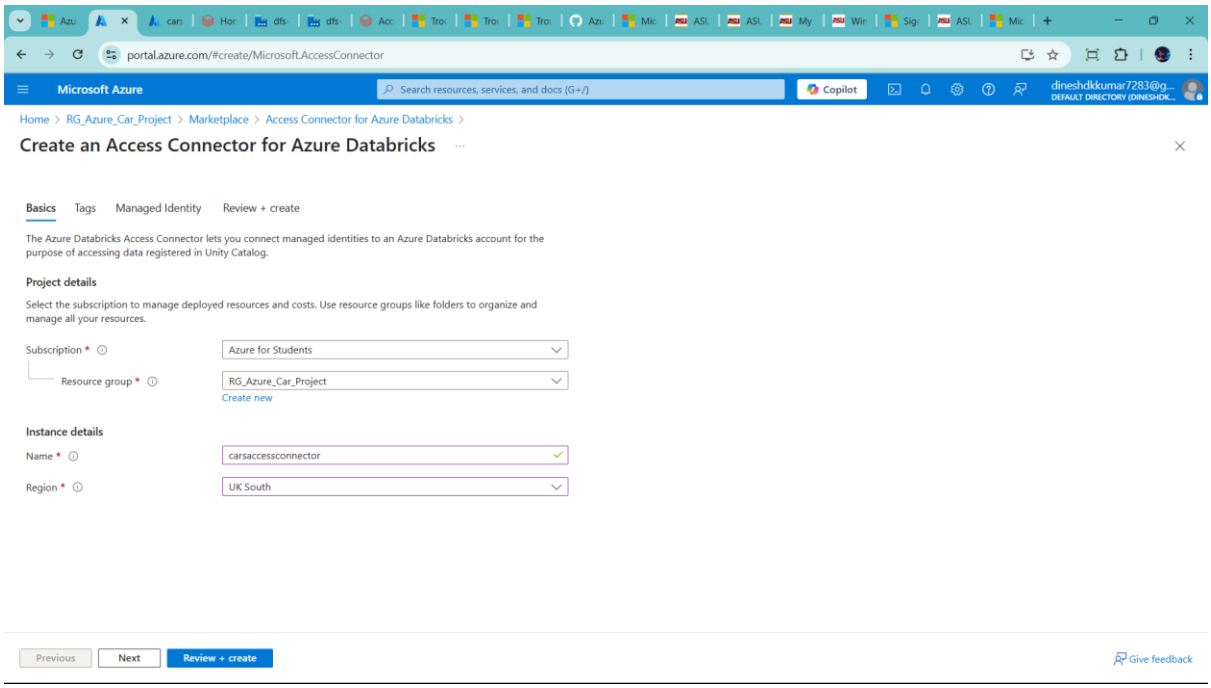
→click create


Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
dineshdkkumar7283@g...
DEFAULT DIRECTORY (DINESHDK...
Home > RG_Azure_Car_Project >
Marketplace
Get Started
Service Providers
Management
Private Marketplace
Private Offer Management
My Marketplace
Favorites
My solutions
Recently created
Private plans
Categories
Analytics (58)
AI + Machine Learning (22)
Databases (16)
IT & Management Tools (12)
Compute (4)
databricks
Pricing : All
Operating System : All
Publisher Type : All
Product Type : All
Publisher name : All
Azure services only
New! Get AI-generated suggestions for 'databricks'
View suggestions
Showing 1 to 20 of 71 results for 'databricks'. Clear search
Tile view
Azure Databricks
Microsoft
Azure Service
Azure Databricks is the fast, easy and collaborative Apache Spark-based analytics platform.
Create
Lakehouse Optimizer for Azure Databricks
Blueprint Technologies LLC
Azure Application
Provides near realtime optimization and granular utilization and performance transparency.
Starts at
Free
Create
Access Connector for Azure Databricks
Microsoft
Azure Service
Leverage managed identities in Unity Catalog
Create
Unravel for Azure Databricks
Unravel Data
Azure Application
Data observability for Azure Databricks
Price varies
Create
Unravel for Azure Databricks (PayGo)
Unravel Data
SaaS
AI-powered data observability + FinOps for Azure Databricks
Price varies
Subscribe
Is Marketplace helpful?

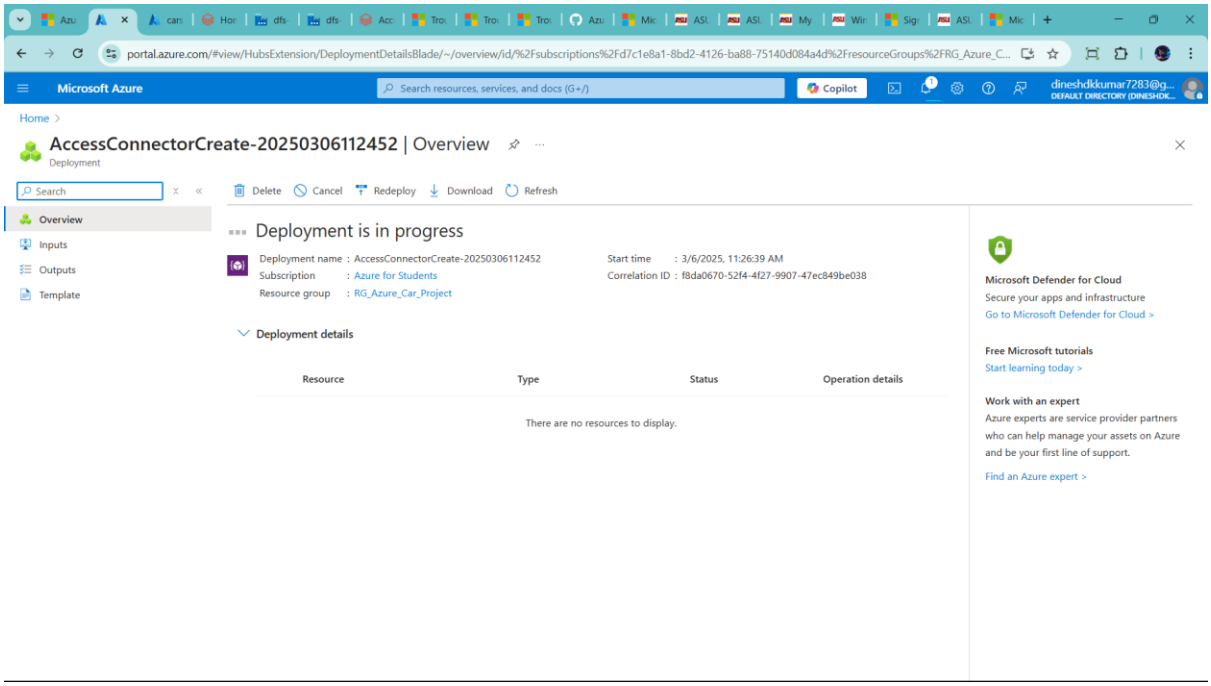

Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
dineshdkkumar7283@g...
DEFAULT DIRECTORY (DINESHDK...
Home > RG_Azure_Car_Project > Marketplace >
Access Connector for Azure Databricks
Microsoft
Access Connector for Azure Databricks
Add to Favorites
Microsoft | Azure Service
4.7 (216 ratings)
Plan
Access Connector for Azure Databricks
Create
Overview
Plans
Usage Information + Support
Ratings + Reviews
Unity Catalog provides unified governance for all data and AI assets in your Lakehouse.
Unity Catalog can be configured to use an Azure managed identity to access storage containers on behalf of Unity Catalog users. Managed identities provide an identity for applications to use when they connect to resources that support Microsoft Entra ID authentication.You can use managed identities in Unity Catalog to support two primary use cases:-
As an identity to connect to the metastore's storage accounts (where managed tables are stored).
As an identity to connect to other external storage accounts (either for file-based access or for external tables).
Configuring Unity Catalog with a managed identity has the following benefits over configuring Unity Catalog with a service principal:-
You can connect to an Azure Data Lake Storage Gen2 account that is protected by a storage firewall.
Managed identities do not require you to maintain credentials or rotate secrets.
Media
https://portal.azure.com/#
Give feedback

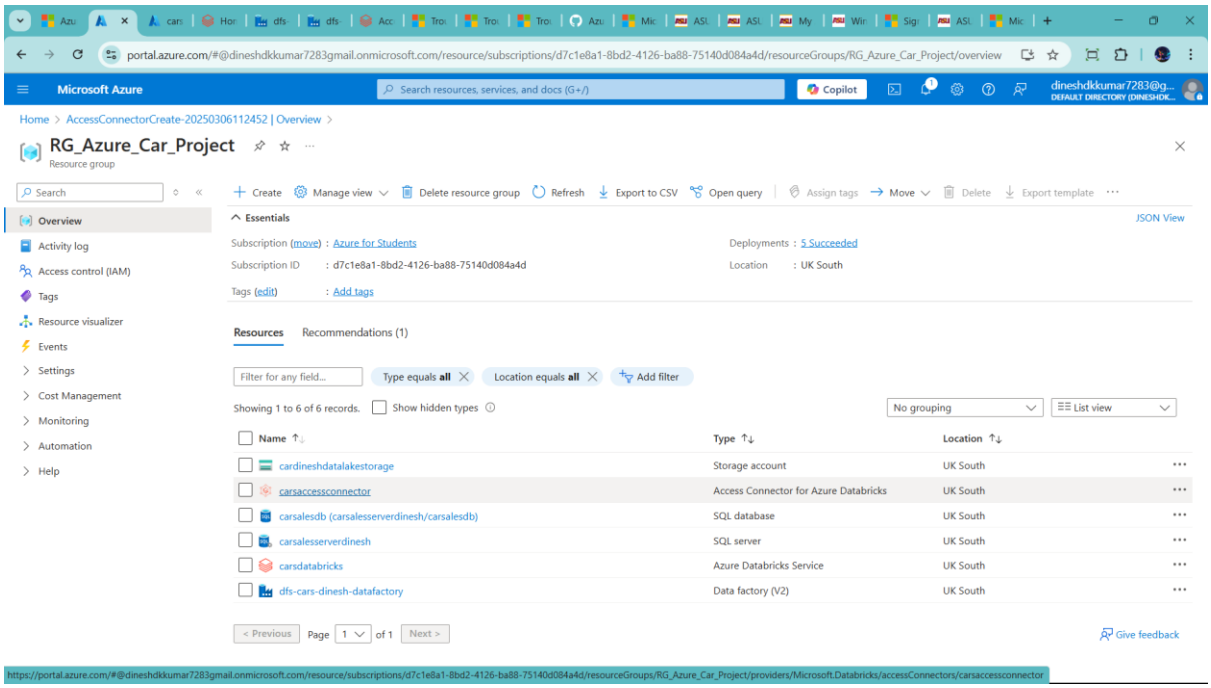


Fill it



--review+create and click create

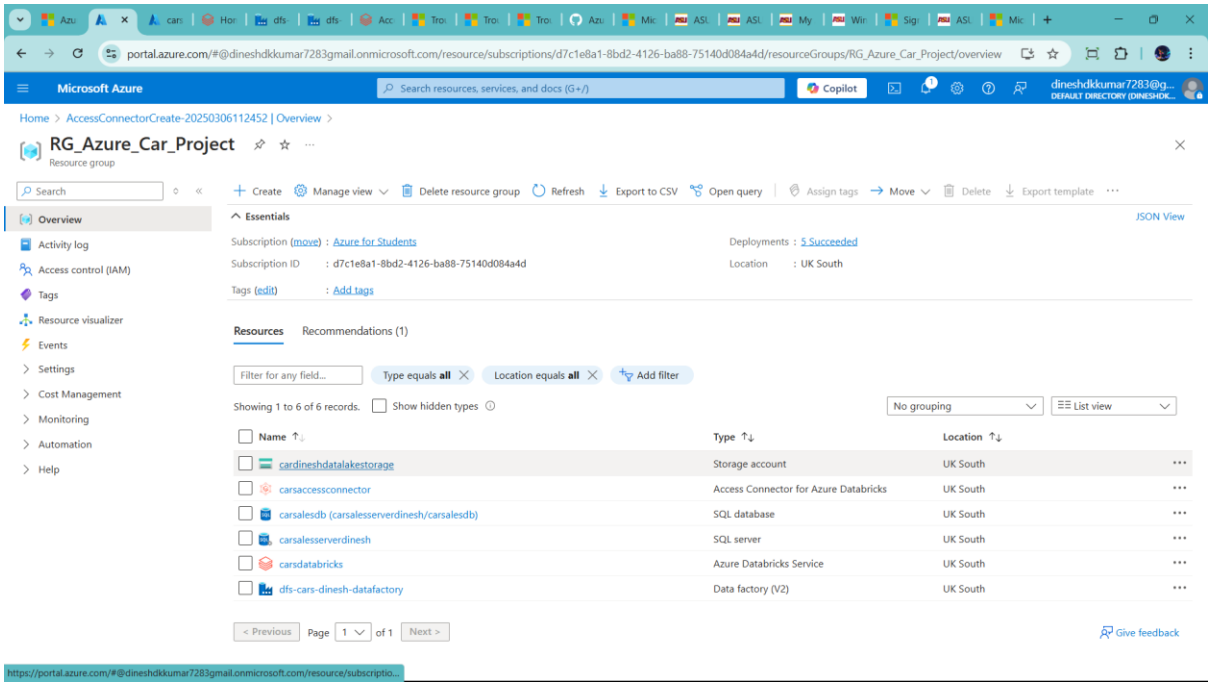




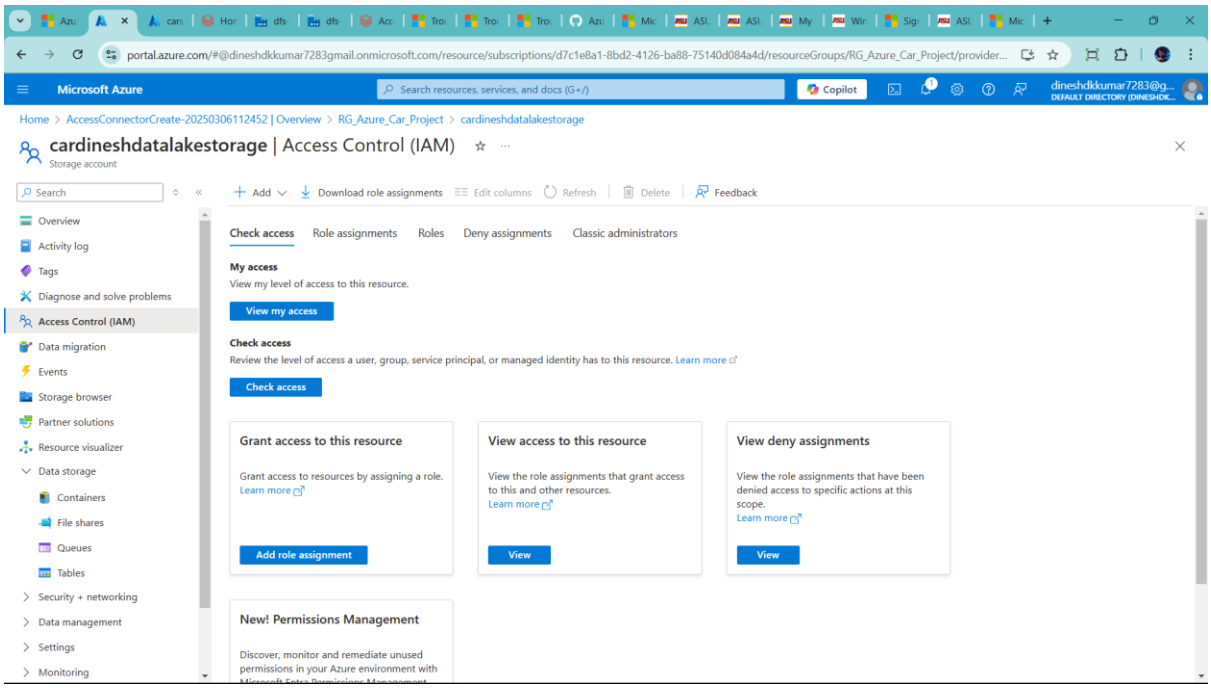
→access connect use to storage whole contributer → it means contribute to cardineshdatalakestorage

→it is indirect link and more data security

→lets go to datalake

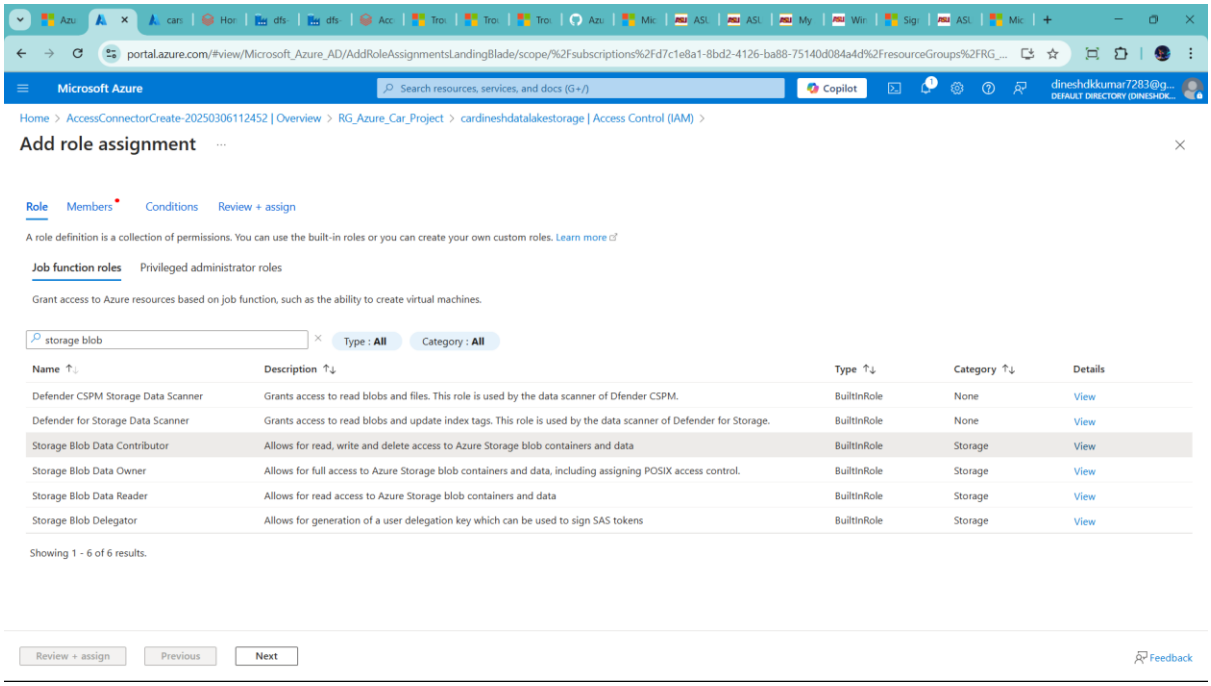


→go to acess control IAM

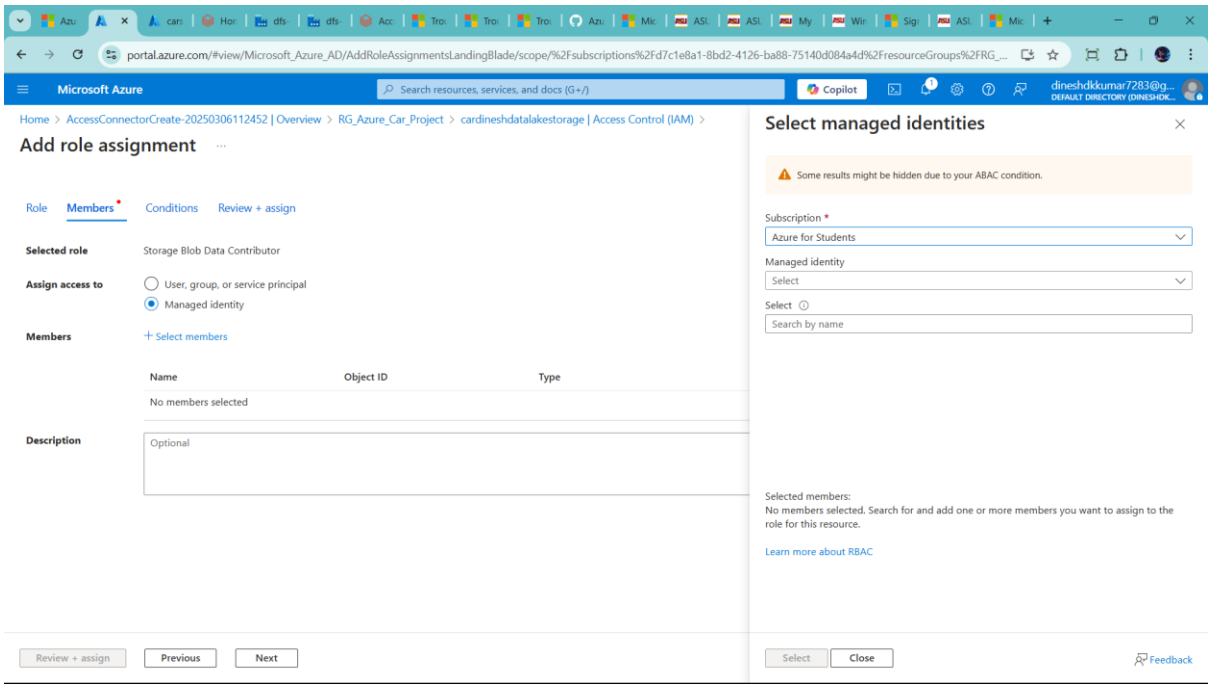


→click +Add →Add role assignment

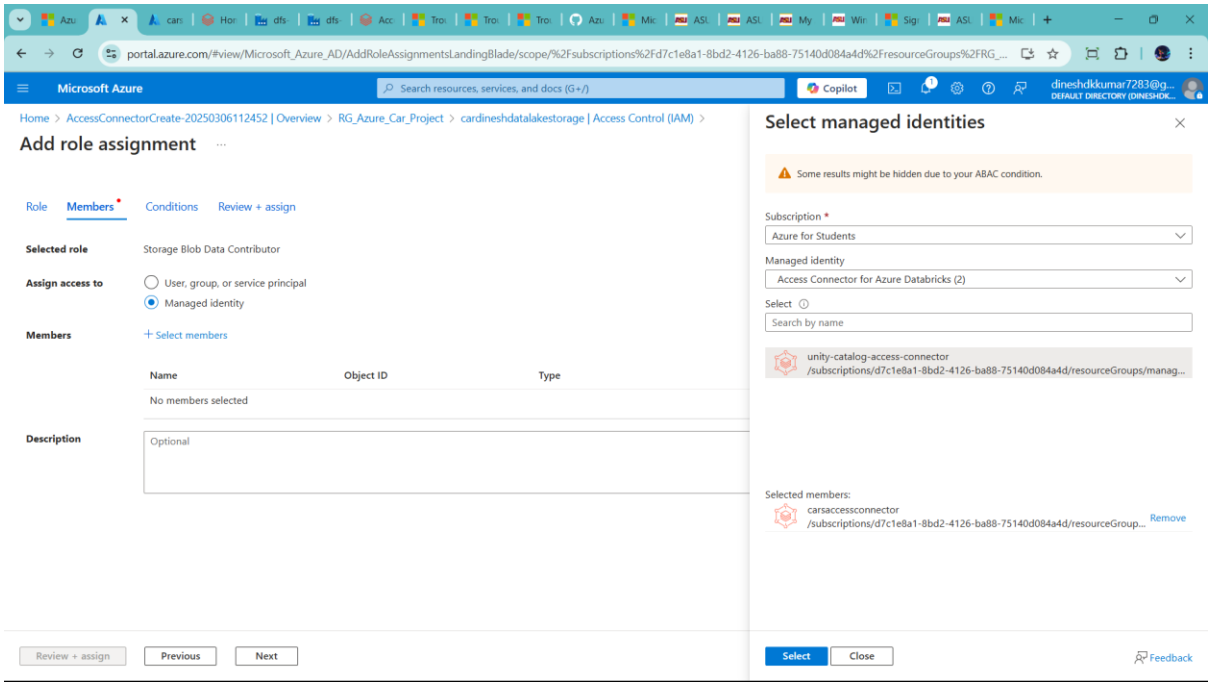
→search storage bolb data contributor→ click storage bolb data contributor →click next



→click manage identity → now select members

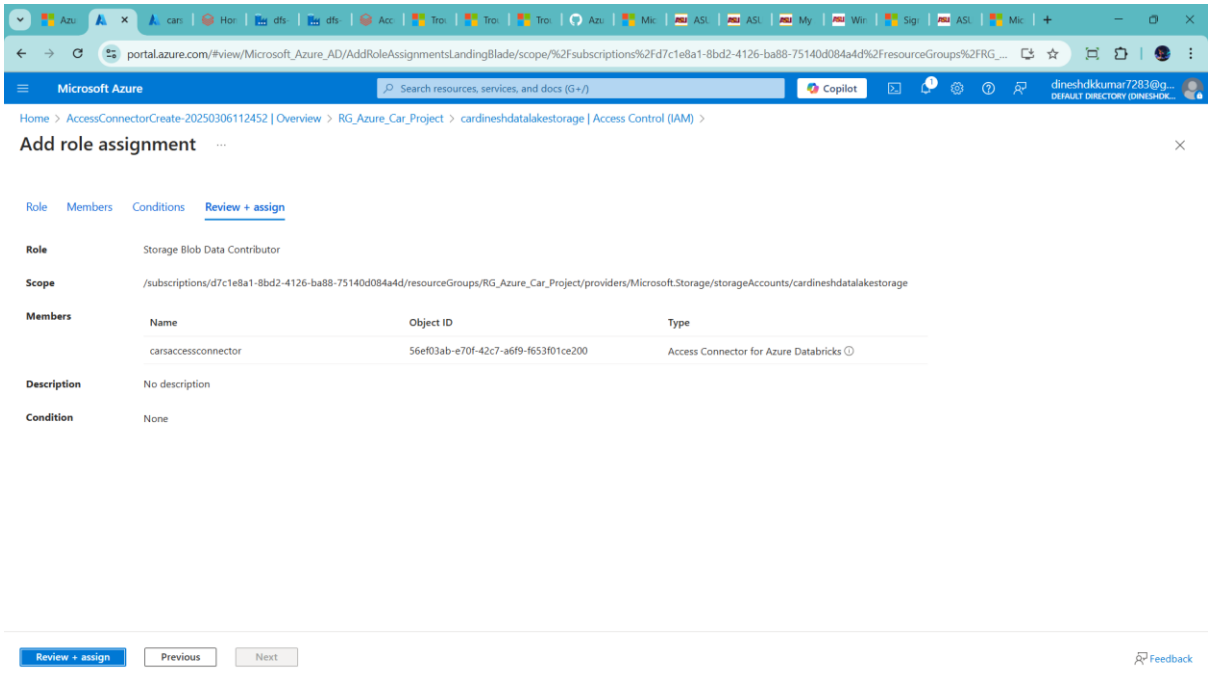


→add like this



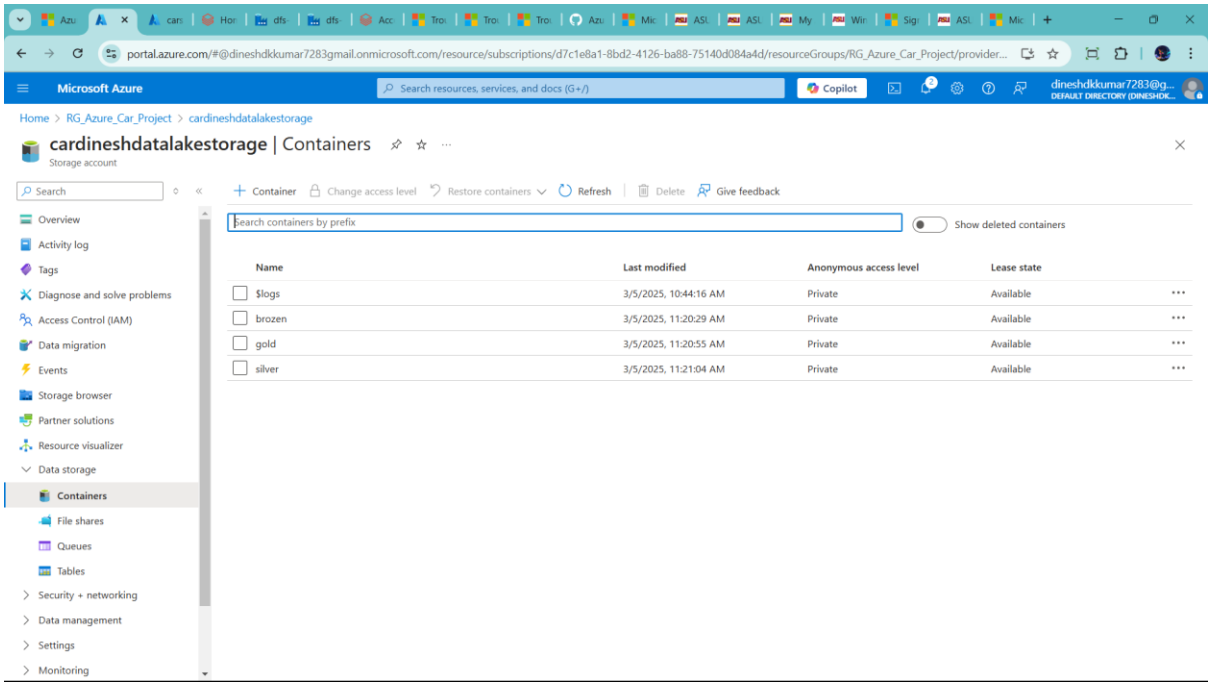
→click select

→review+assign

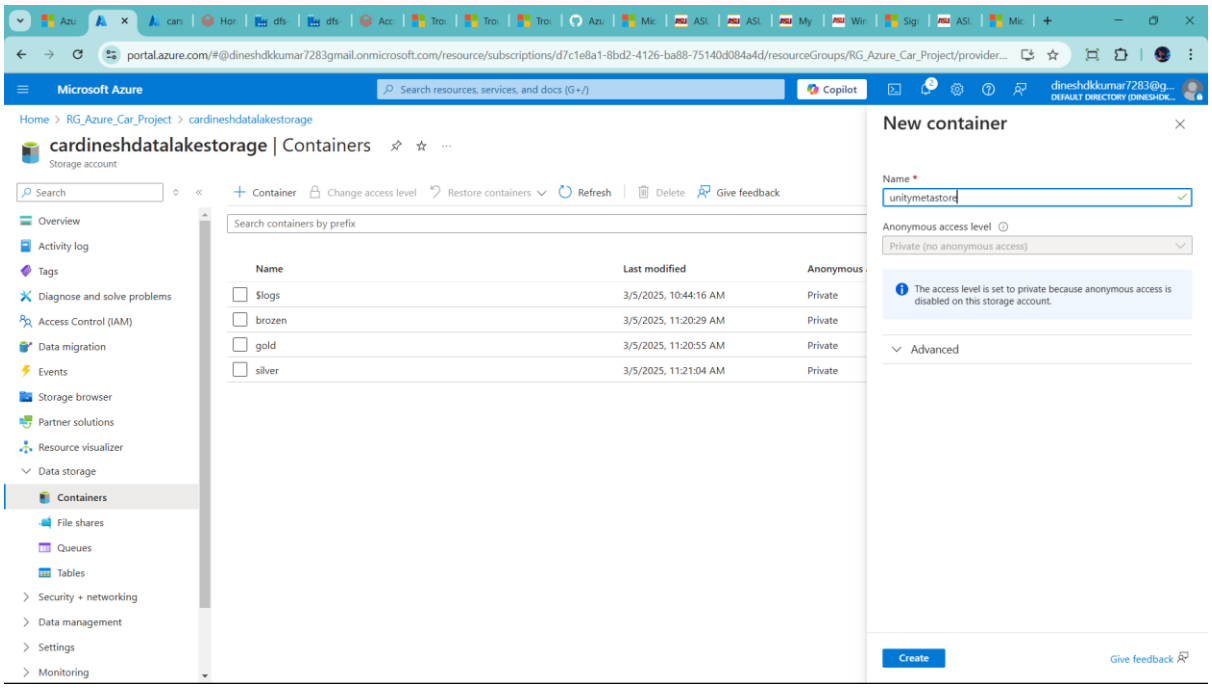


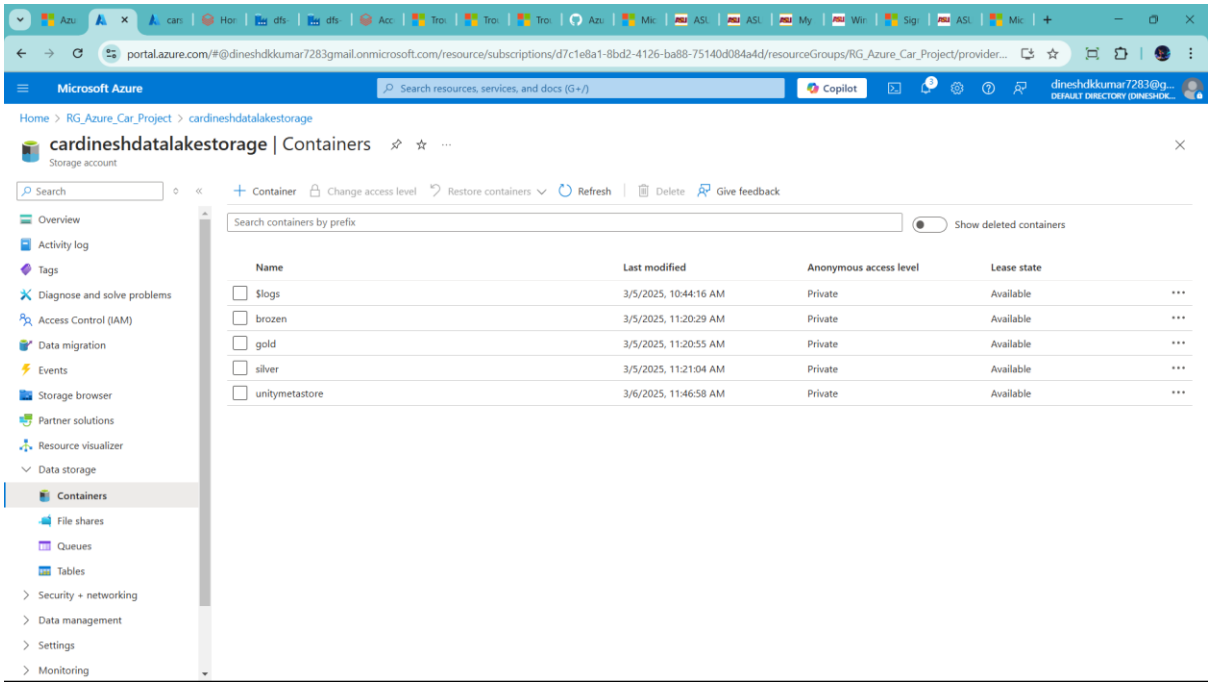
→click one more time →add it

→one more step add container as well

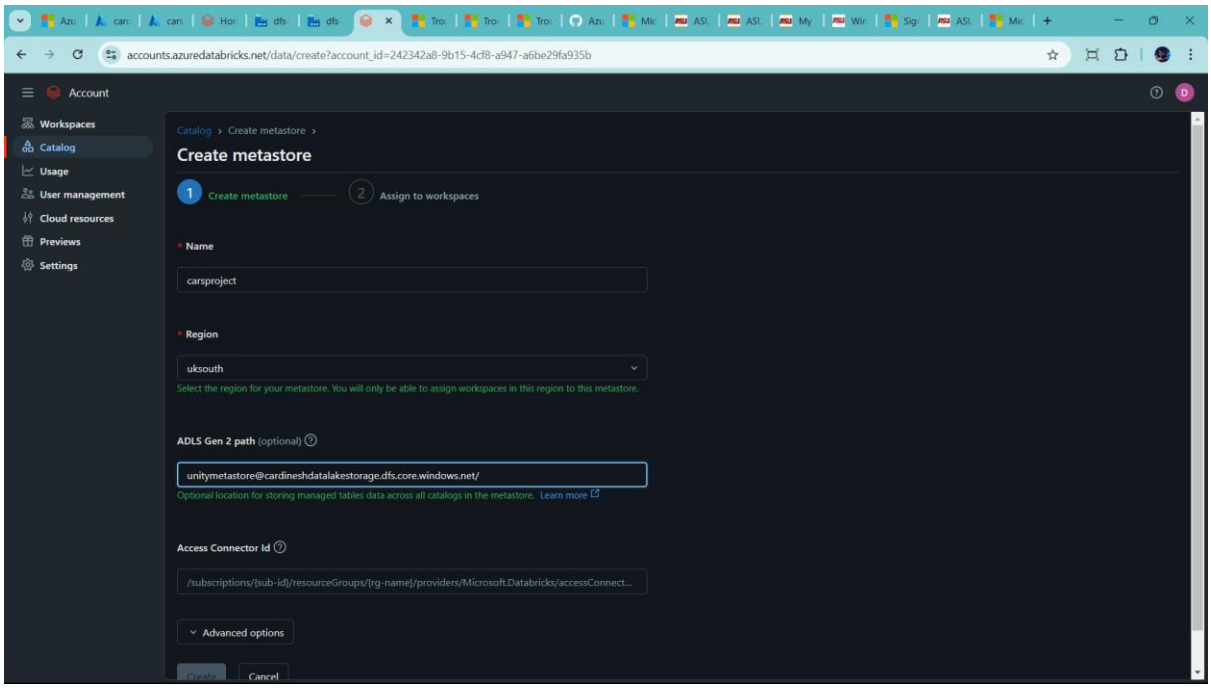


→click + container →add name →create



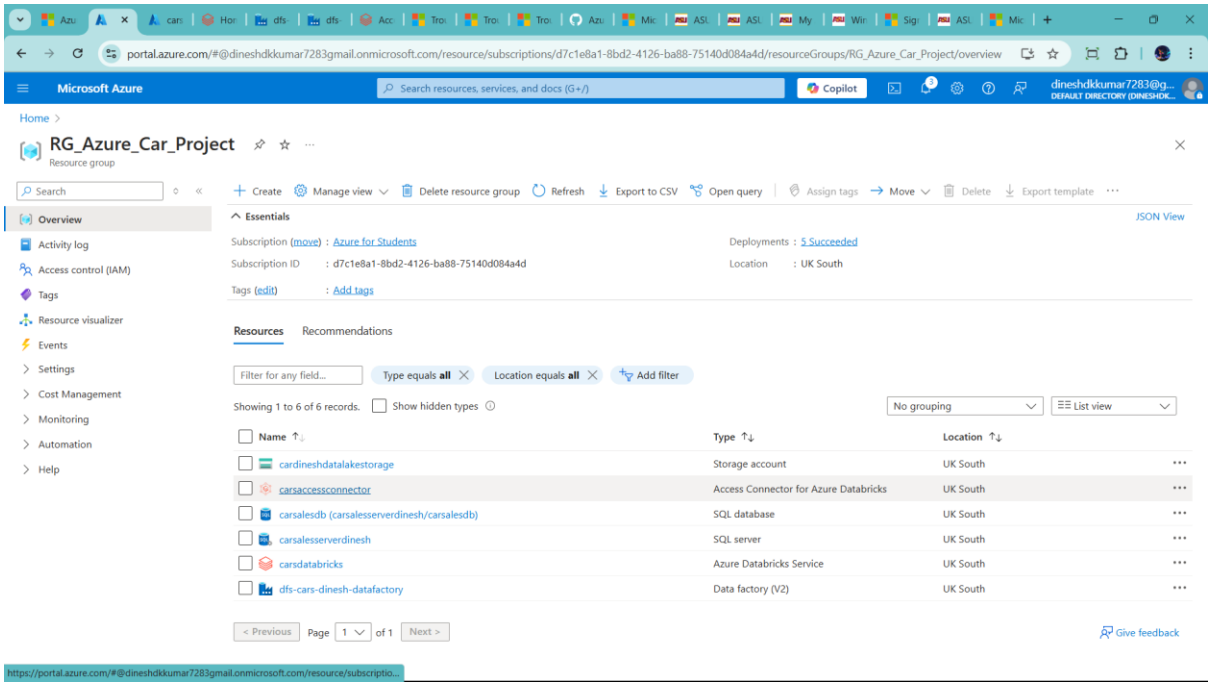


→go add Databricks Admin side

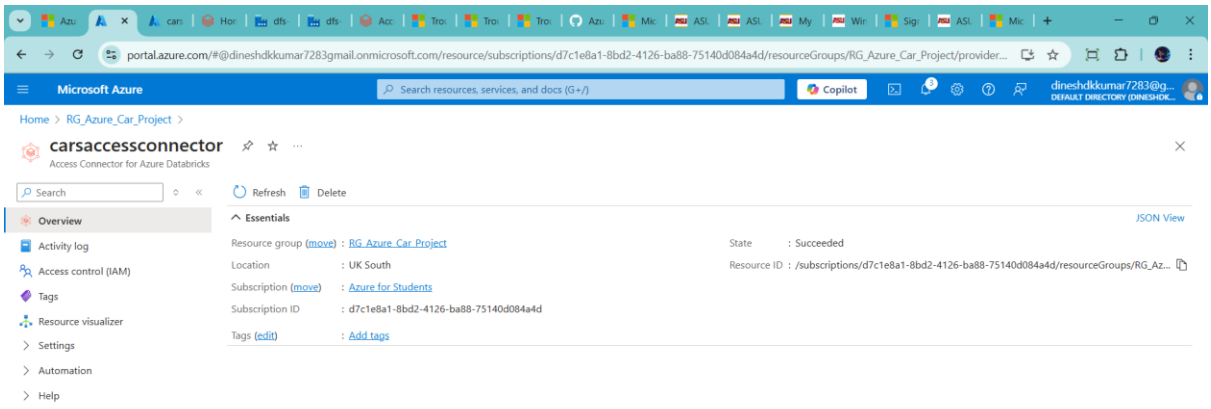


→acess connector id as well

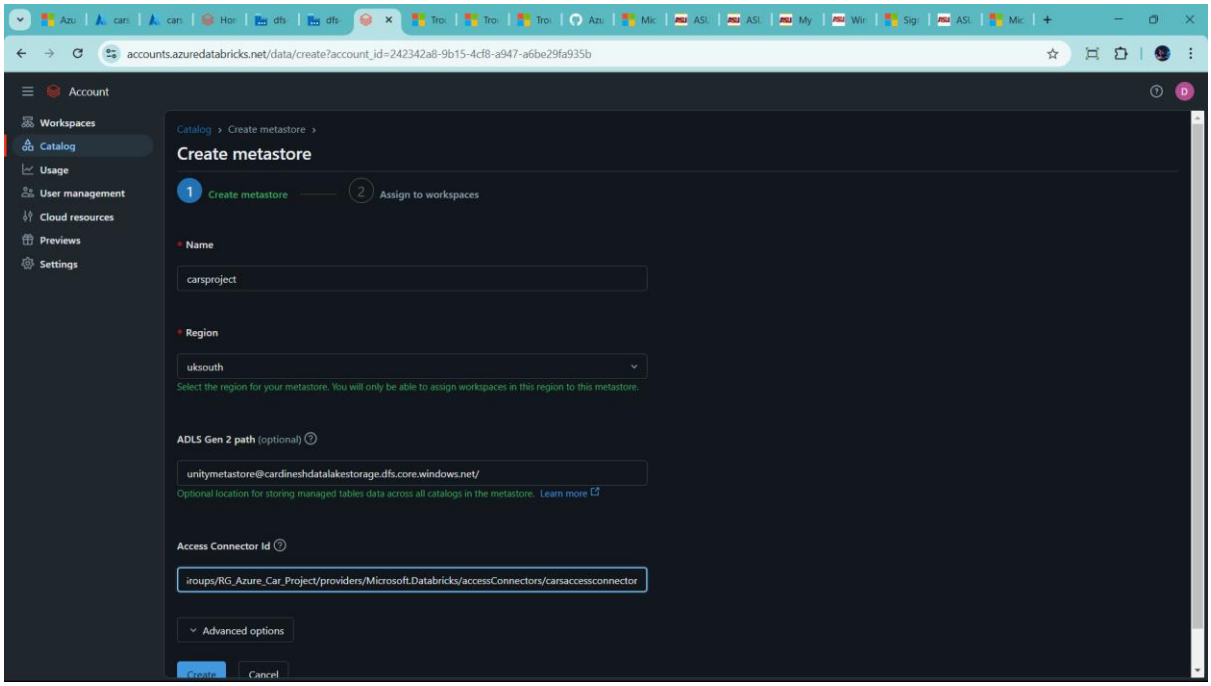
→go to azure→resouce→



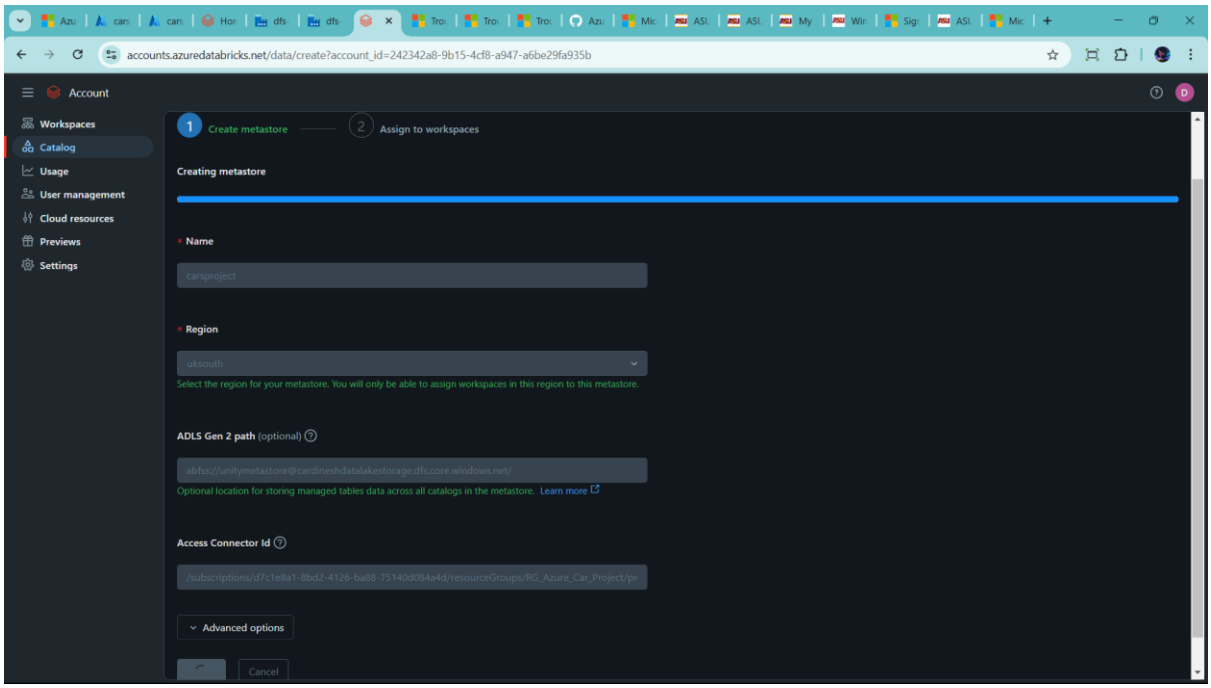
→copy the resource id



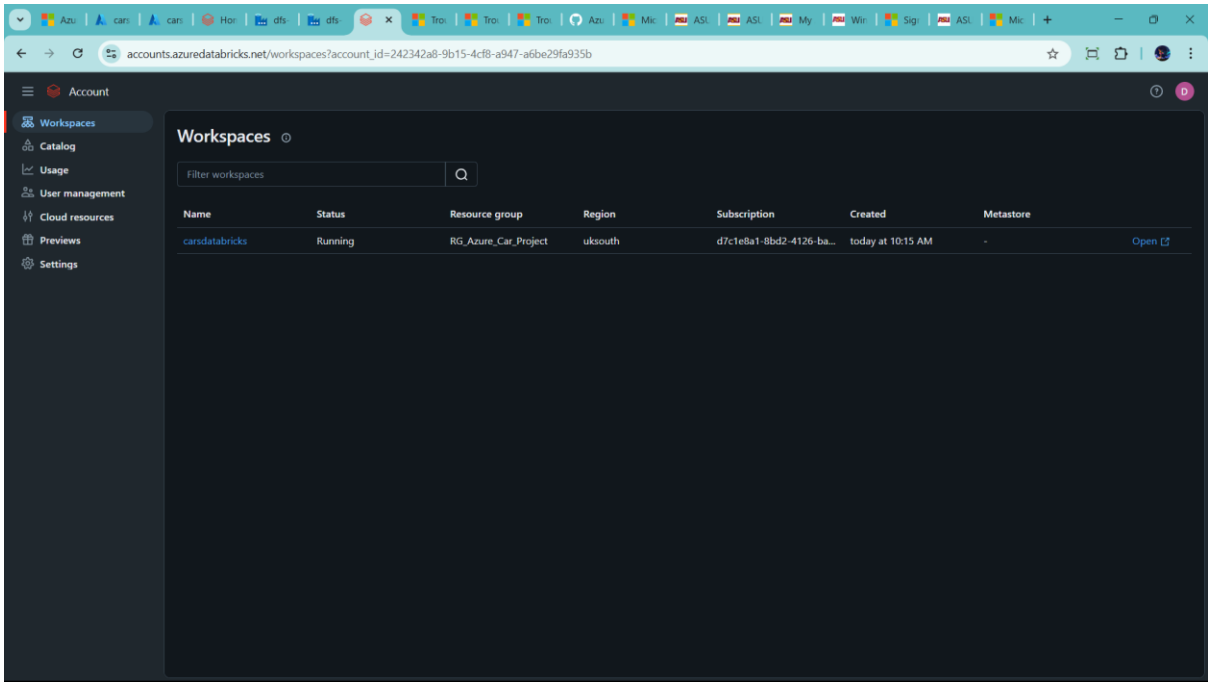
→paste in databricks admin account



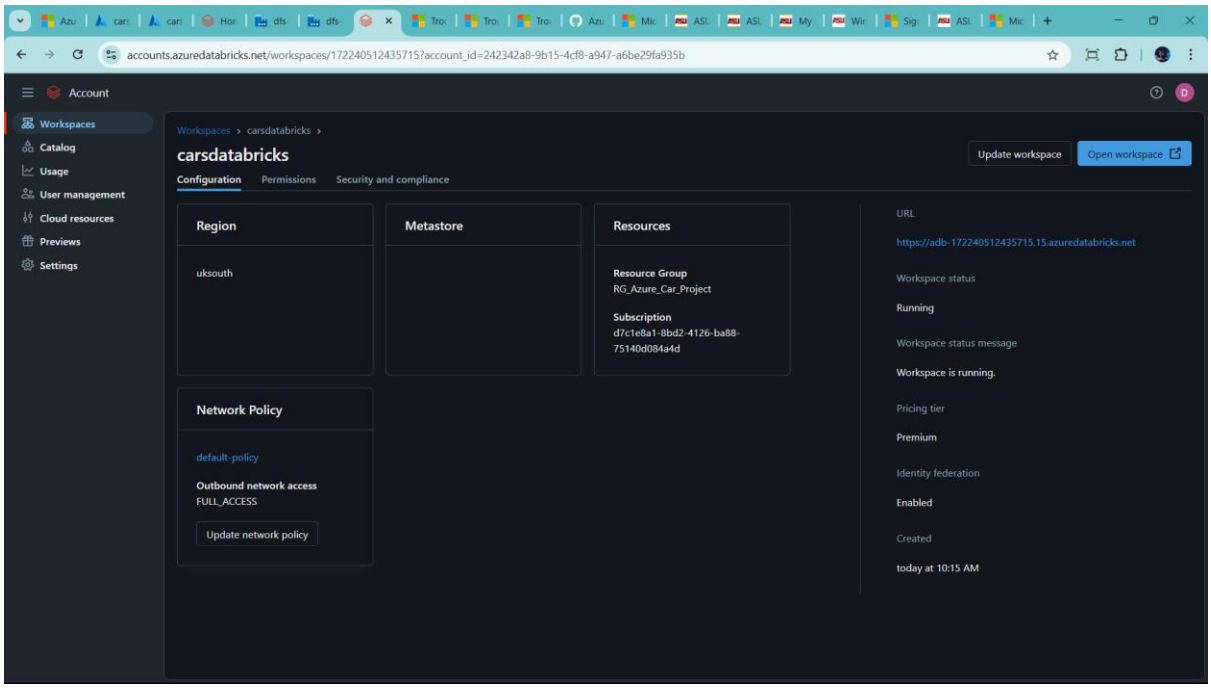
--Click create



→go to workspaces

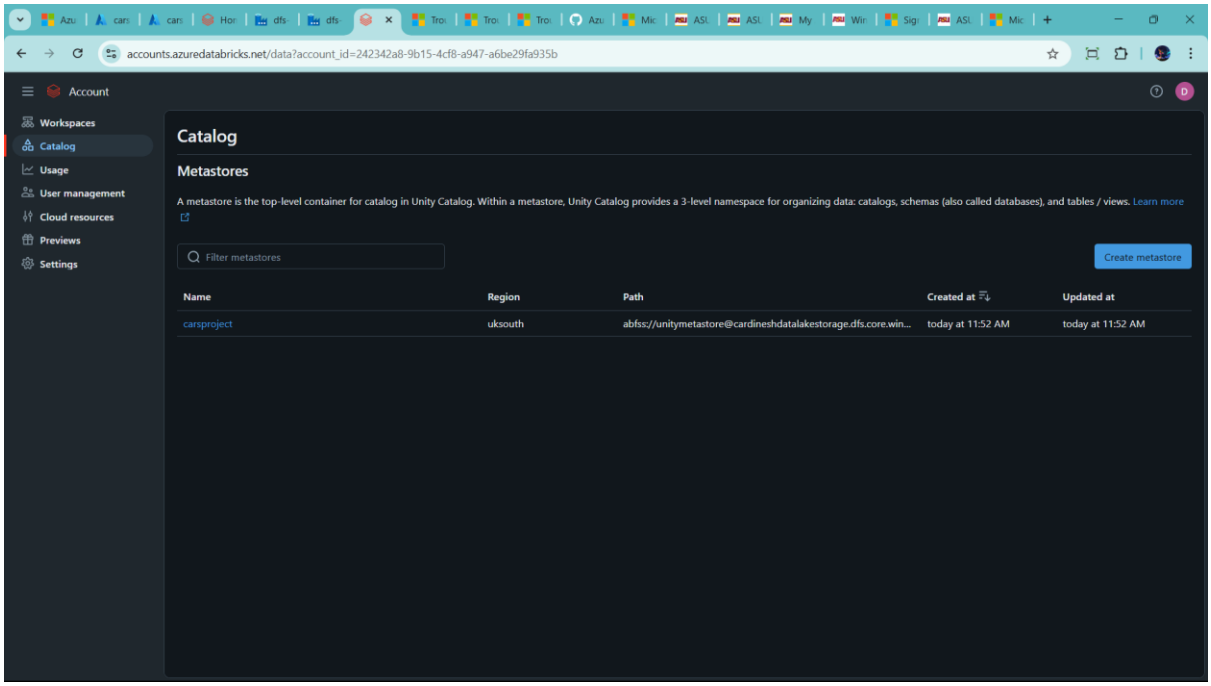


→click name

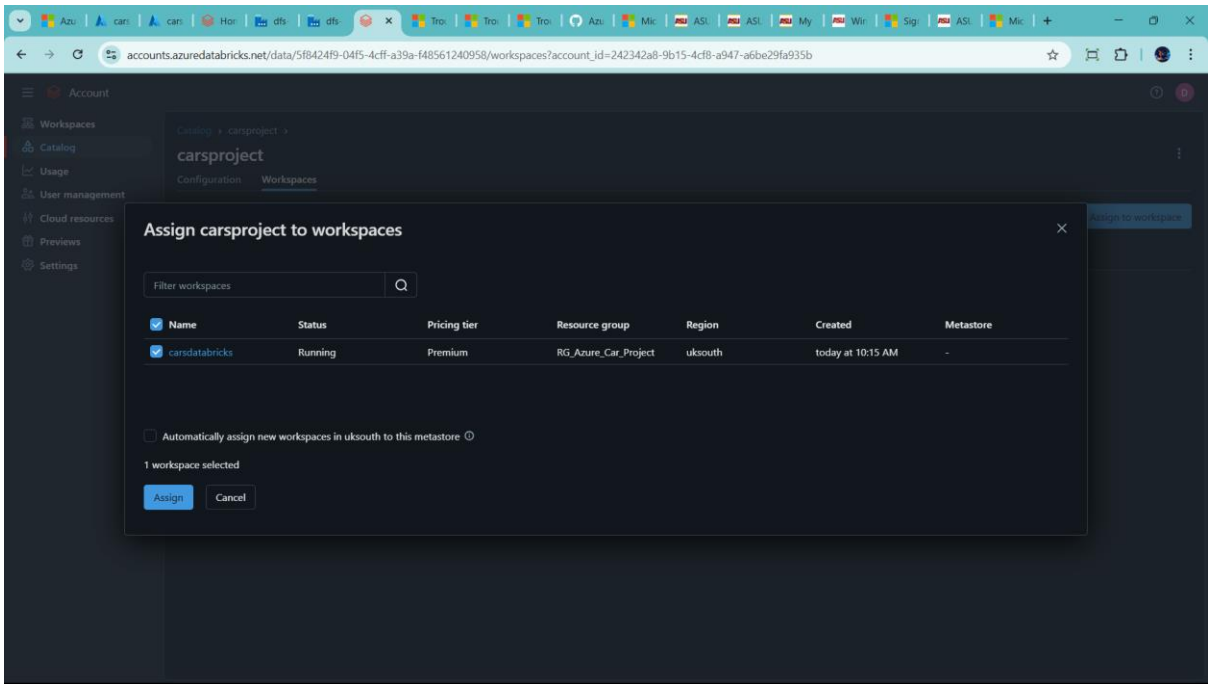


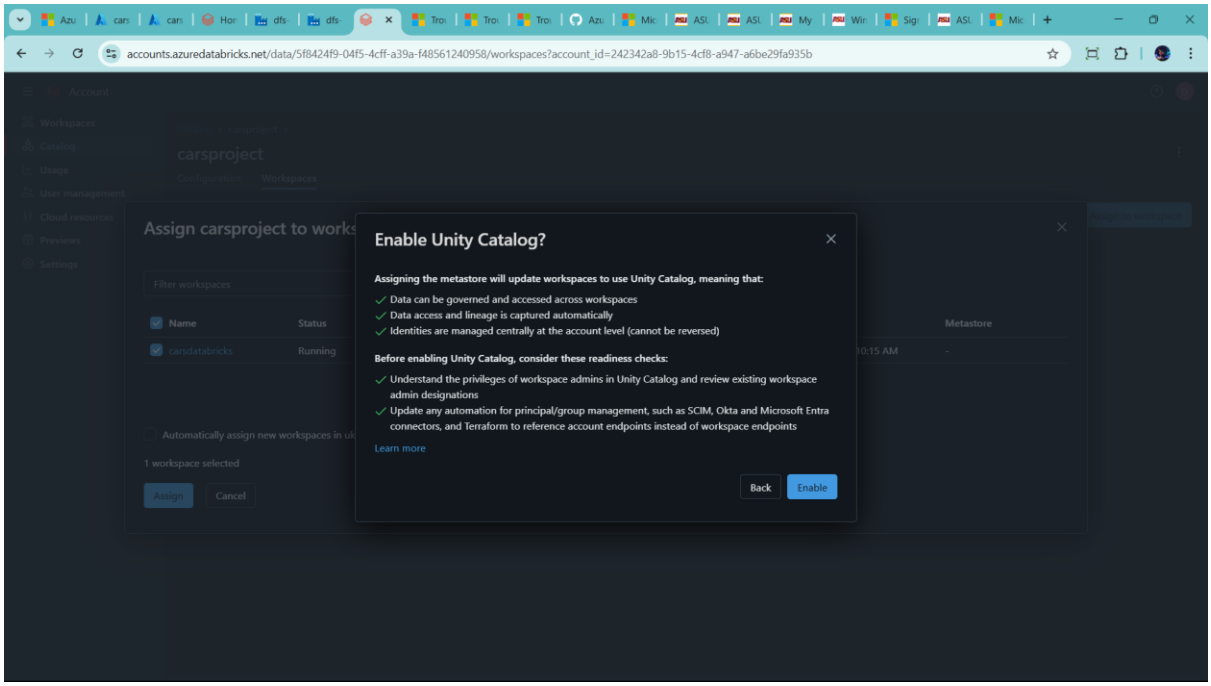
→now assign metastore

→go to catalog →click name →



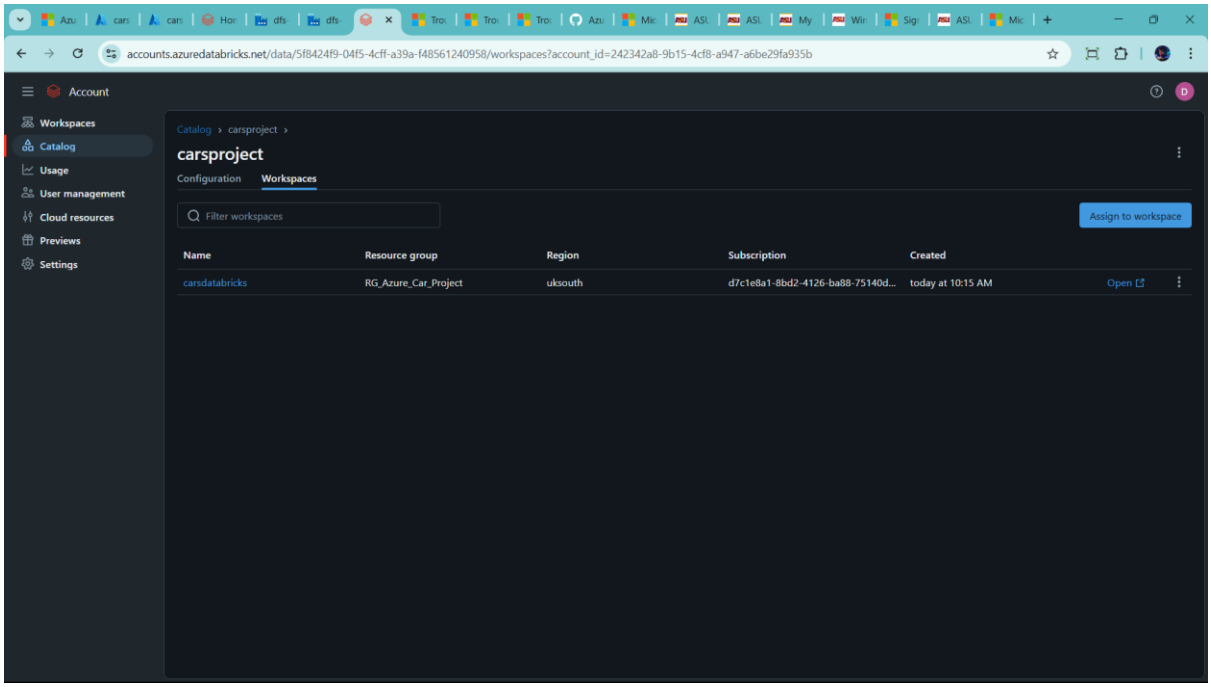
→go to workspaces in catalog → assign workspace



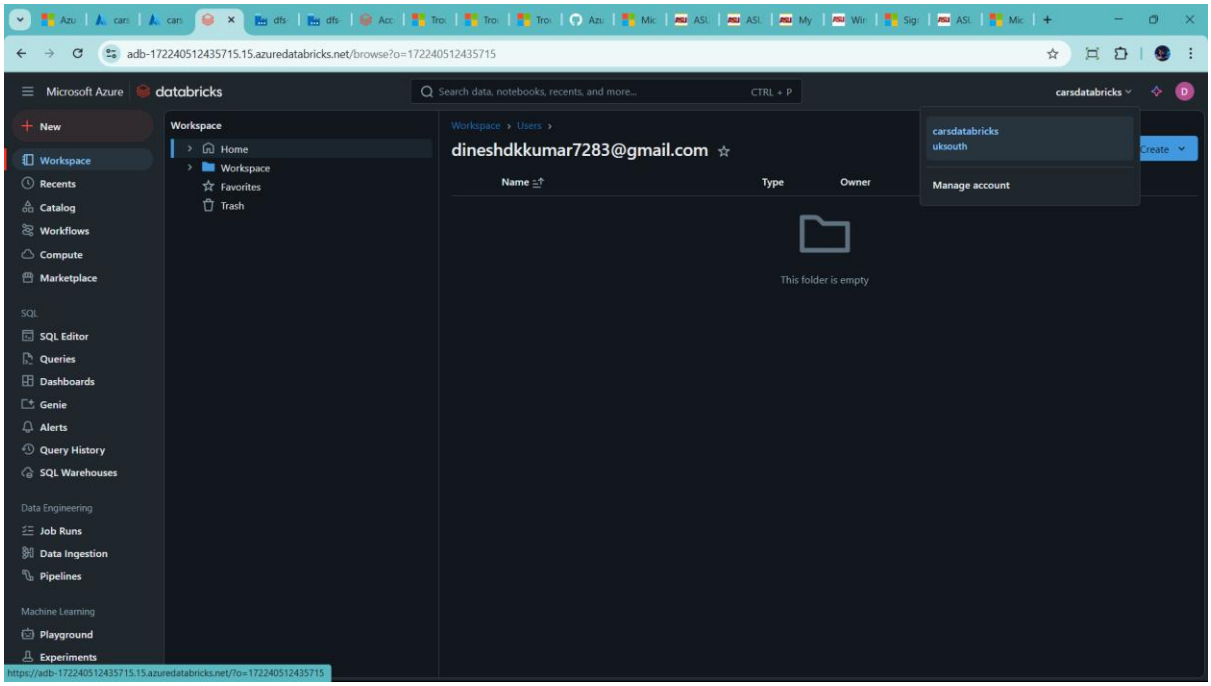


→unity catalog only use this workspace only in databricks

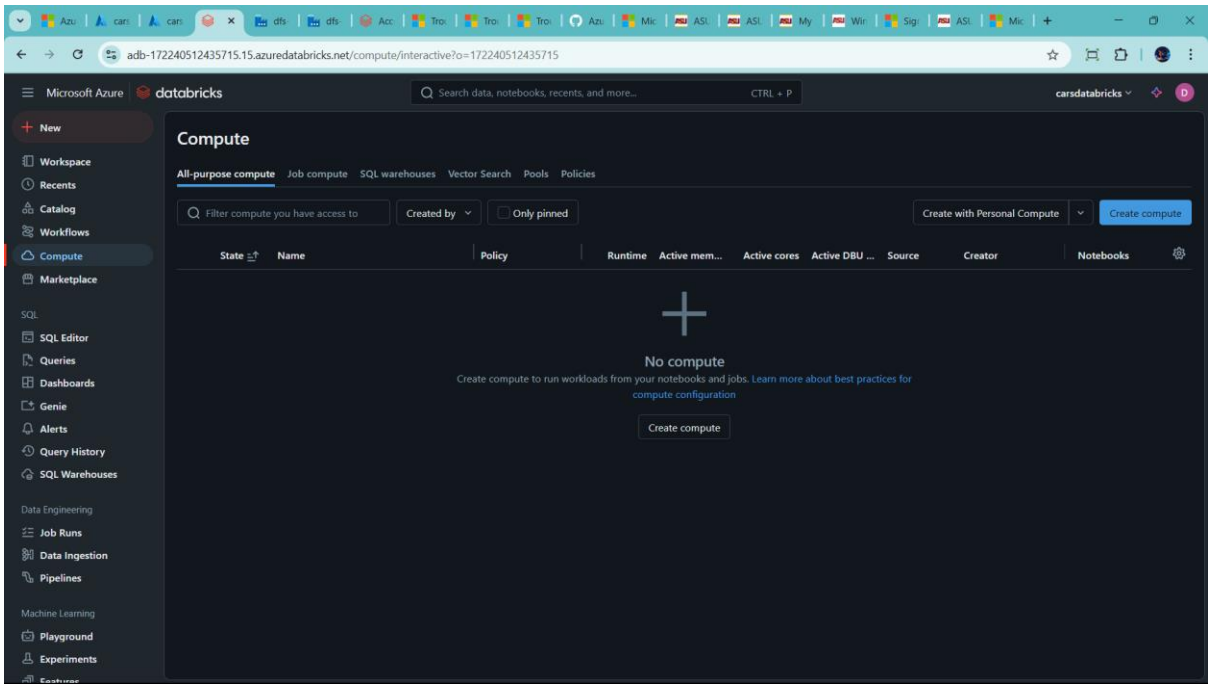
→click enable

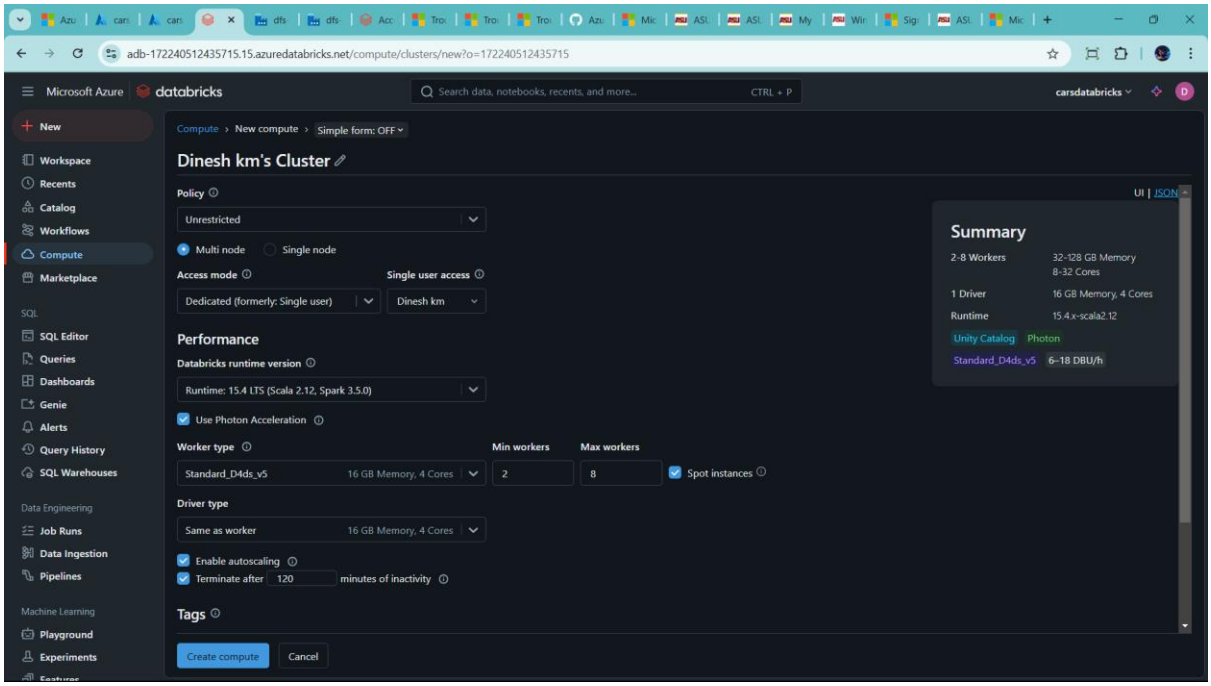


→Now move forward azure databricks and refresh →check it is enable it



→now click compute →create compute





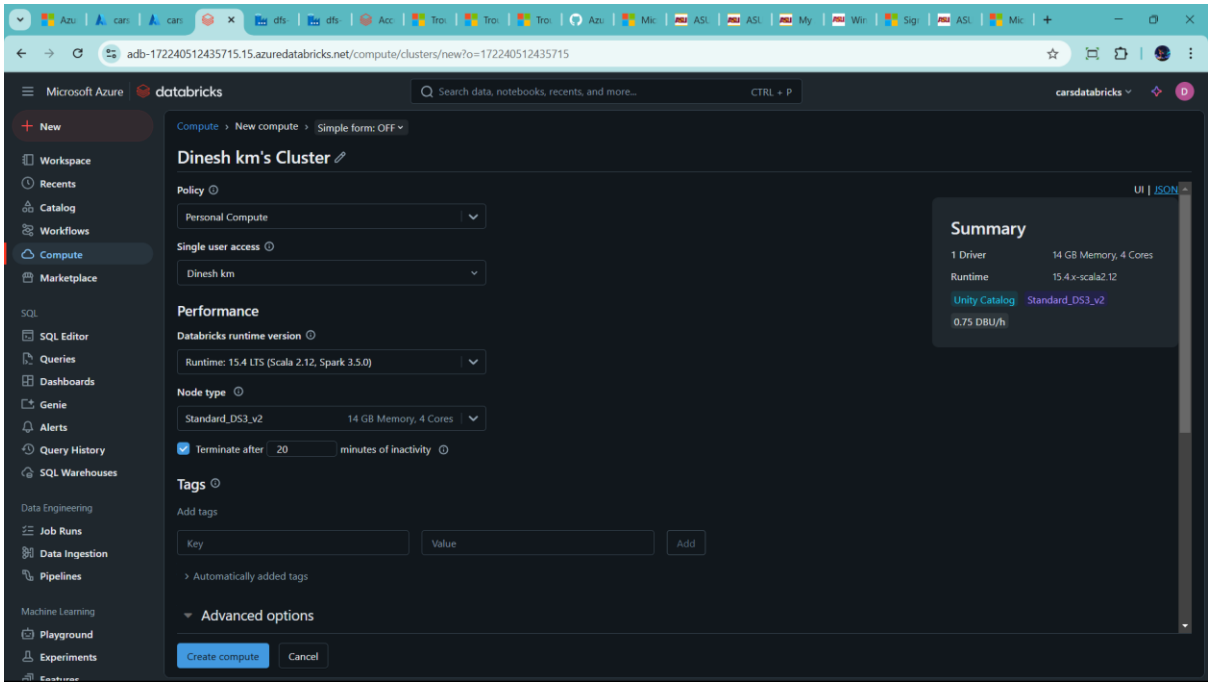
→change it as your preference

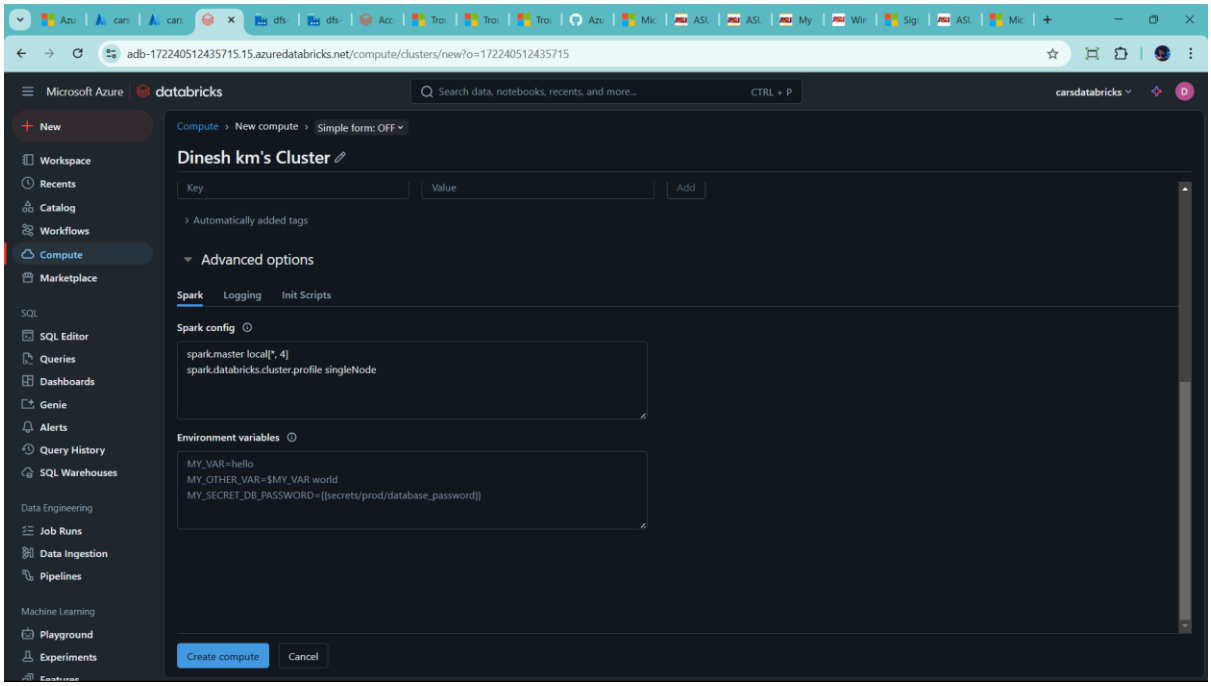
→first policy →change to personal computer →it is single node

→change runtime version

→change termiter time

→go to advance option →no cange





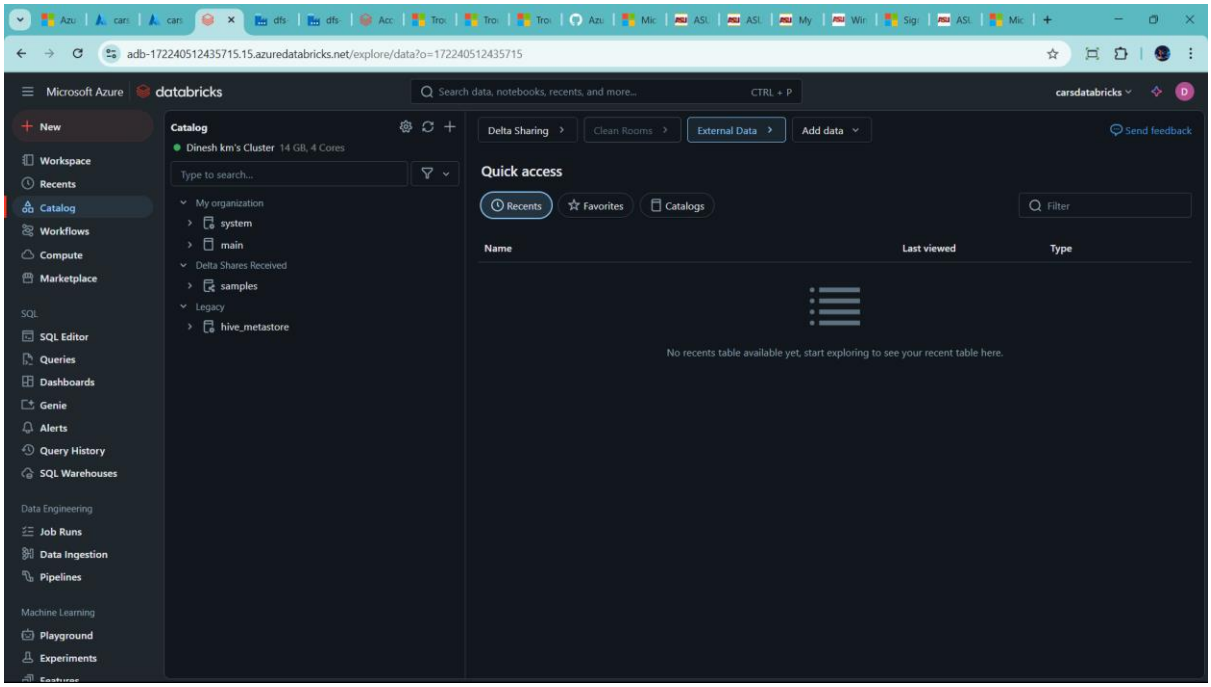
→click create compute

→why use create compute? → it act as a virtual machine

→how to connect databricks and data lake container (brozen,silver,gold folder)

Ans) it is storage credential →it use access connector . to have access permission. To read the data at any container

→lets go databicks →click catalog →click extranal data



→two ways create External location

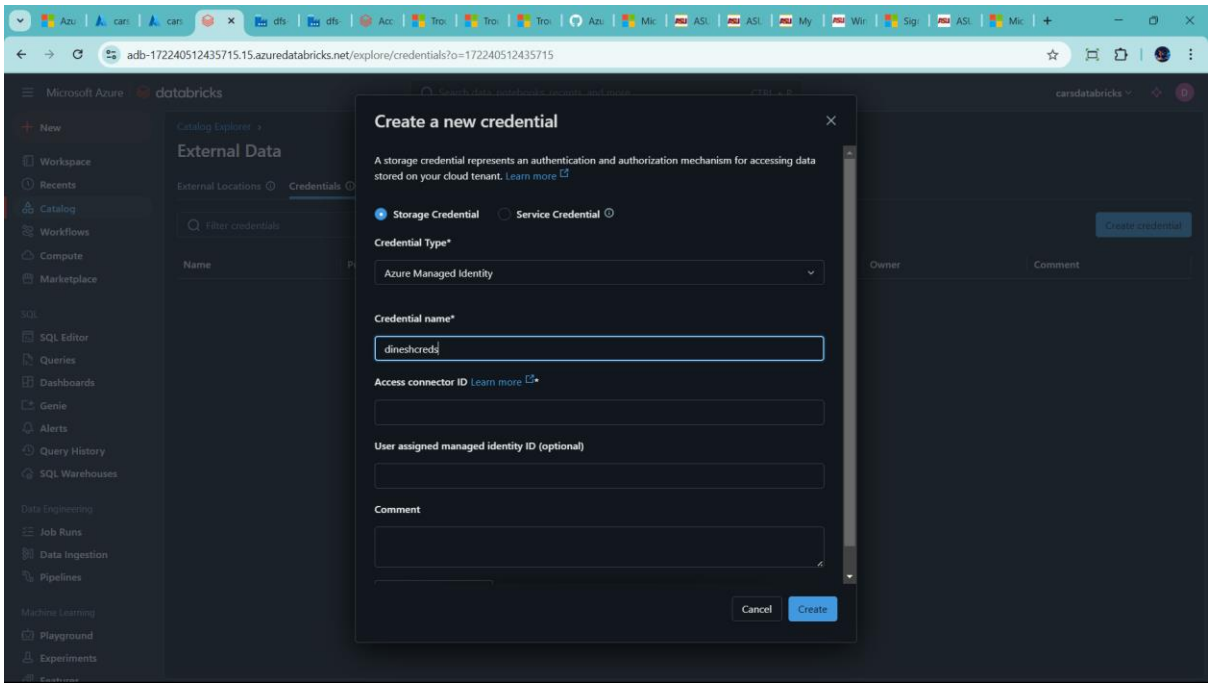
1) UI

2) Notebooks

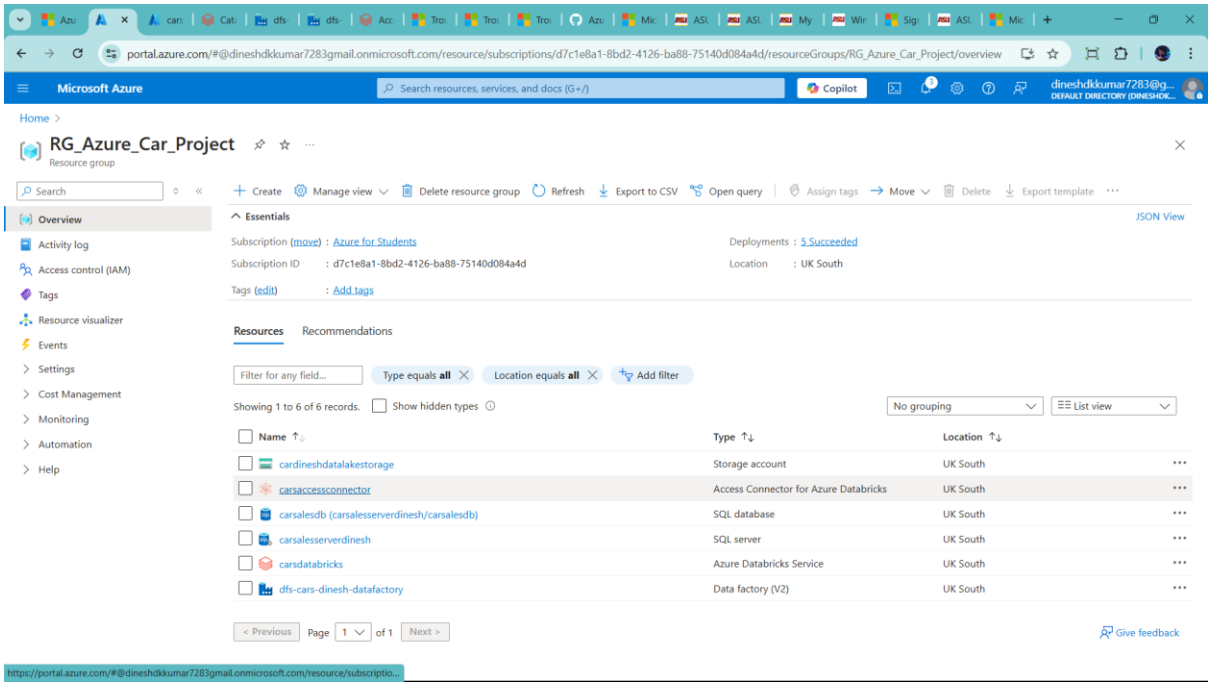
→perfer UI

→click on credentials →create credentials

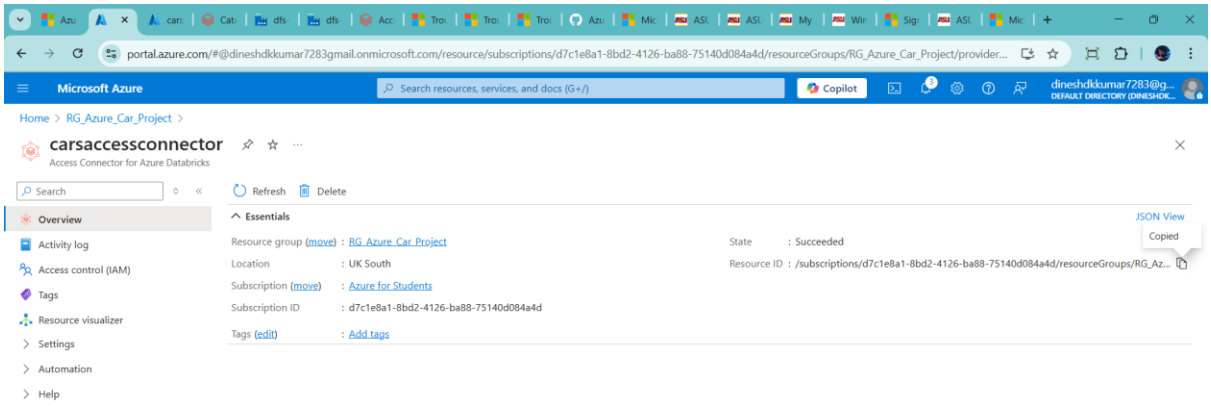
→add name



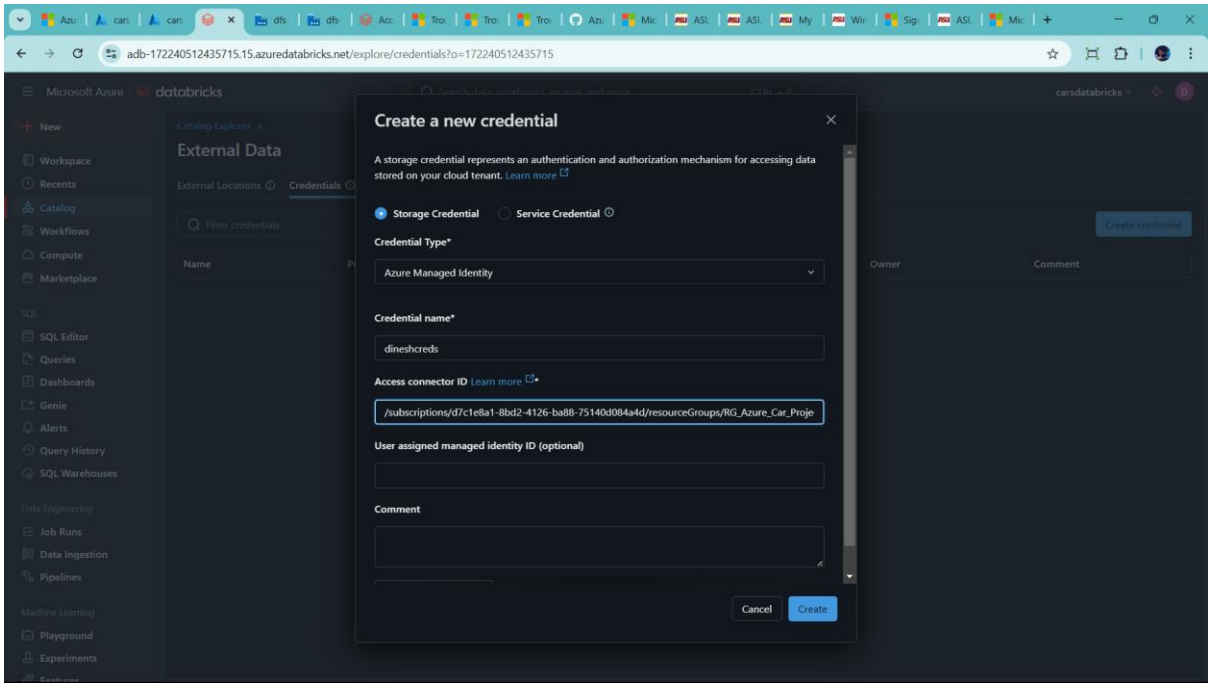
→add Access Connector ID → id present in Azure connector



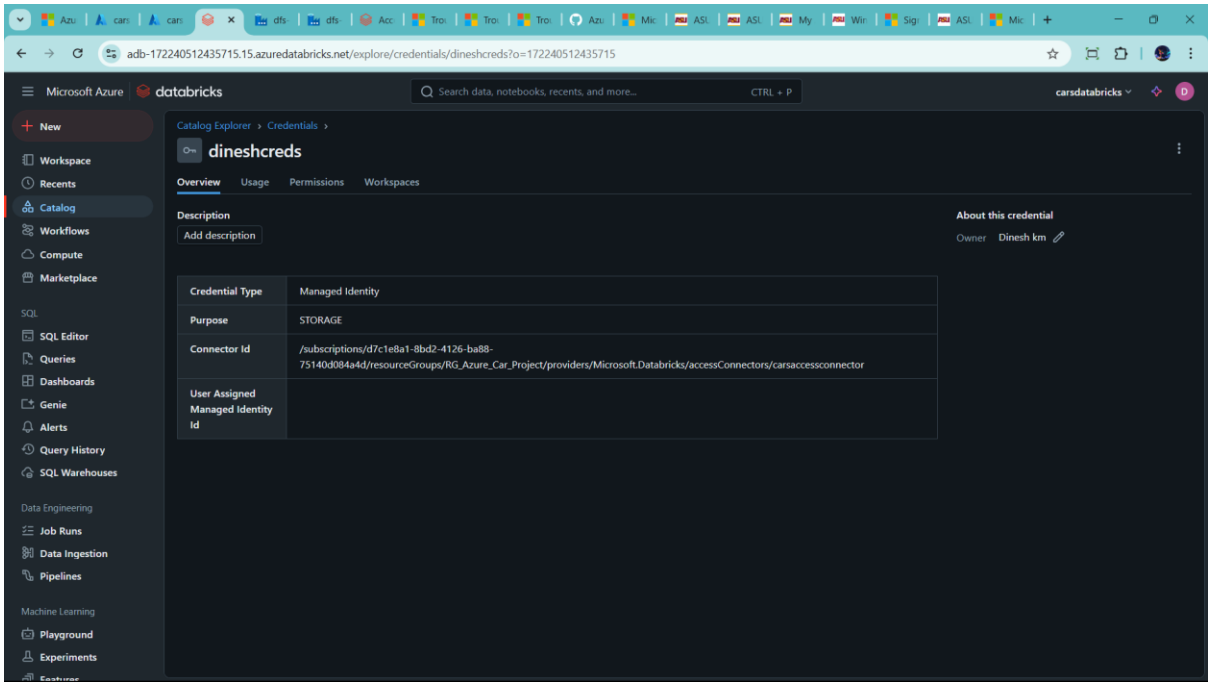
→copy the Resouce id



→past in databricks create new credential → access connecter id

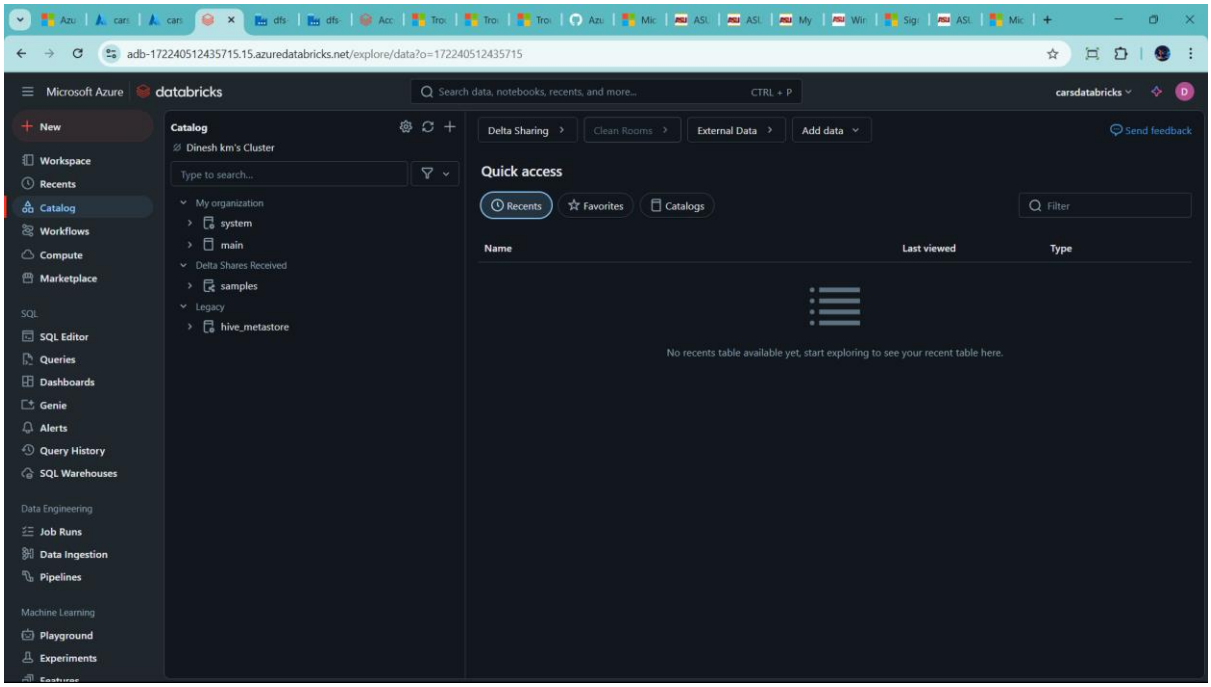


➔ Click create

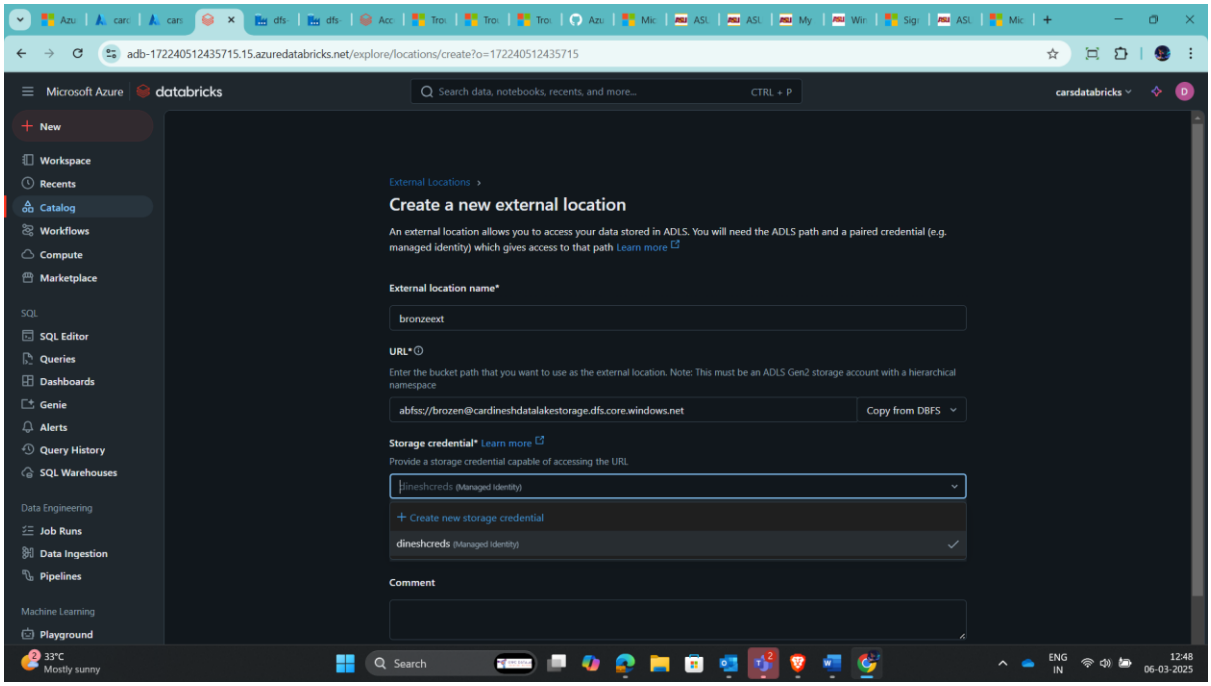


→Now create same UI external creditals

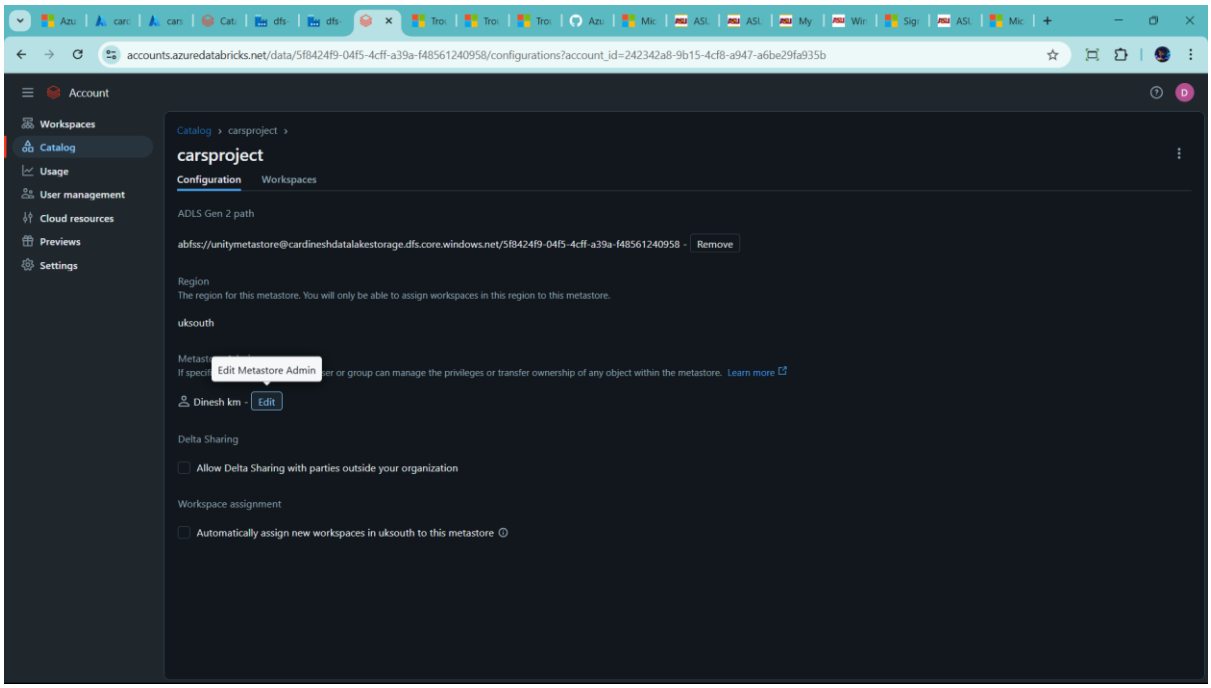
→go catalog →click external data →

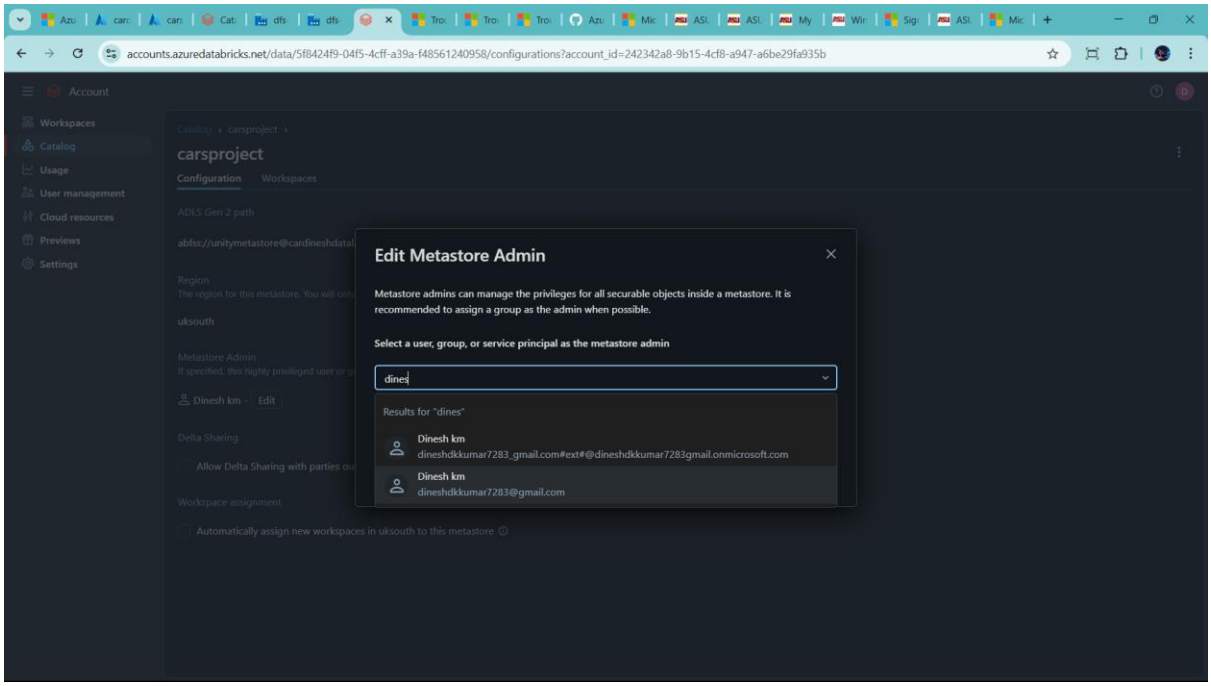


→create external location

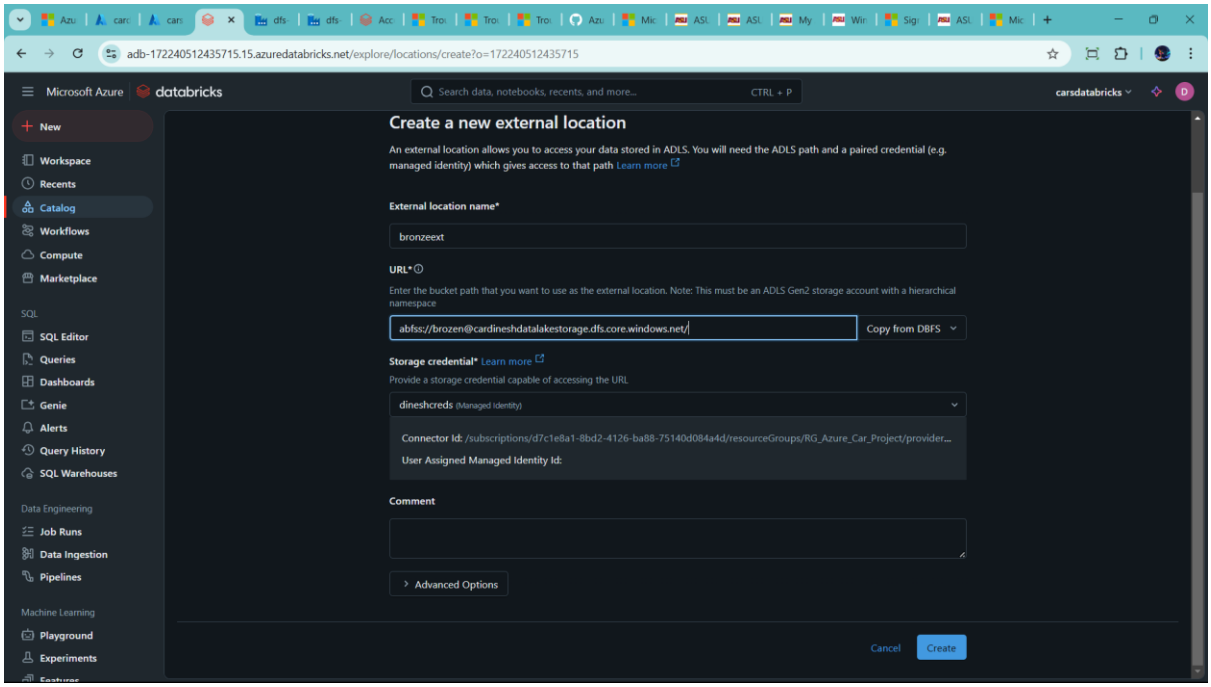


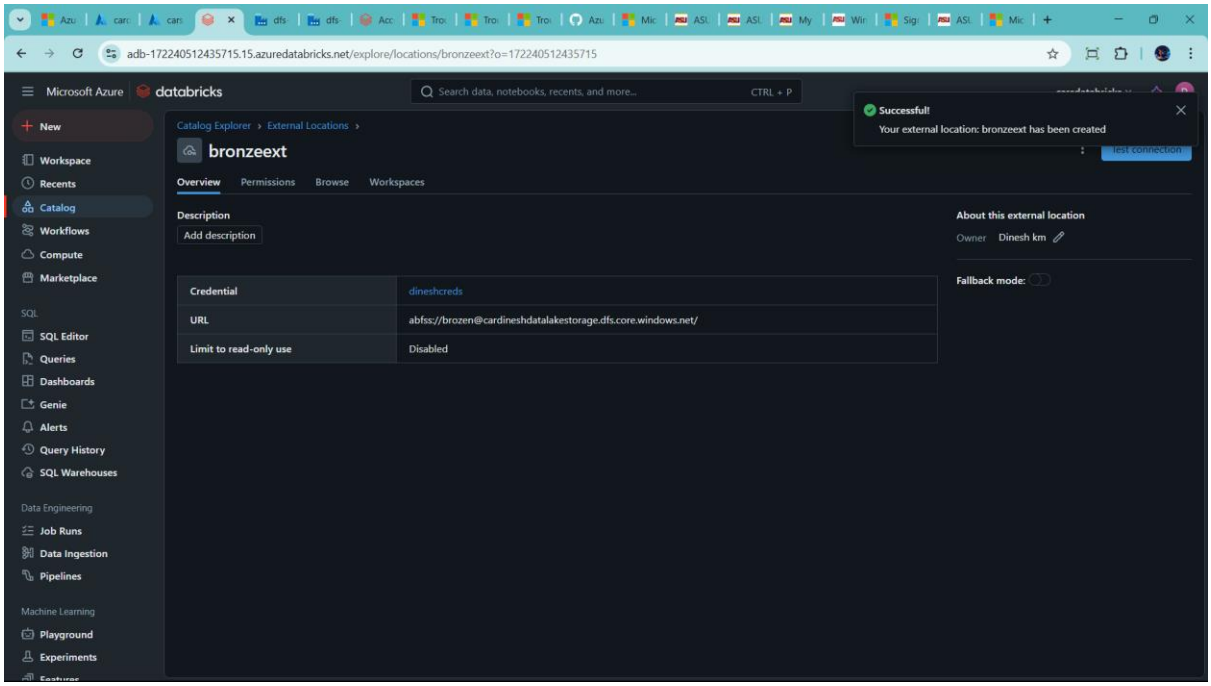
--handle error →#tag account to normal account →save it



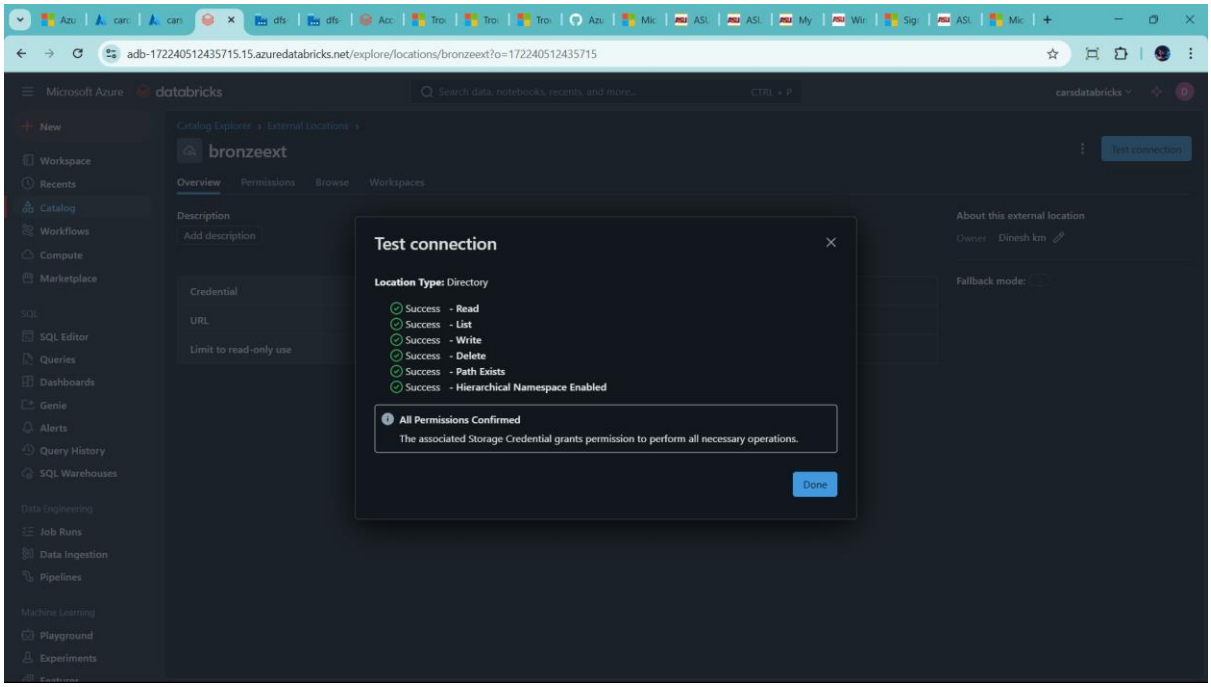

accounts.azuredatabricks.net/data/5f8424f9-04f5-4cff-a39a-f48561240958/configurations?account_id=242342a8-9b15-4cf8-a947-a6be29fa935b
Account
Workspaces
Catalog
Usage
User management
Cloud resources
Previews
Settings
carsproject
Configuration
Workspaces
ADLS Gen 2 path
Region
uksouth
Metastore Admin
Dinesh km
Edit
Delta Sharing
Allow Delta Sharing with parties
Workspace assignment
Automatically assign new workspaces in uksouth to this metastore
Edit Metastore Admin
Metastore admins can manage the privileges for all securable objects inside a metastore. It is recommended to assign a group as the admin when possible.
Select a user, group, or service principal as the metastore admin
dines
Results for "dines"
Dinesh km
dineshdkkumar7283_gmail.com#ext#@dineshdkkumar7283gmail.onmicrosoft.com
Dinesh km
dineshdkkumar7283@gmail.com

→now create

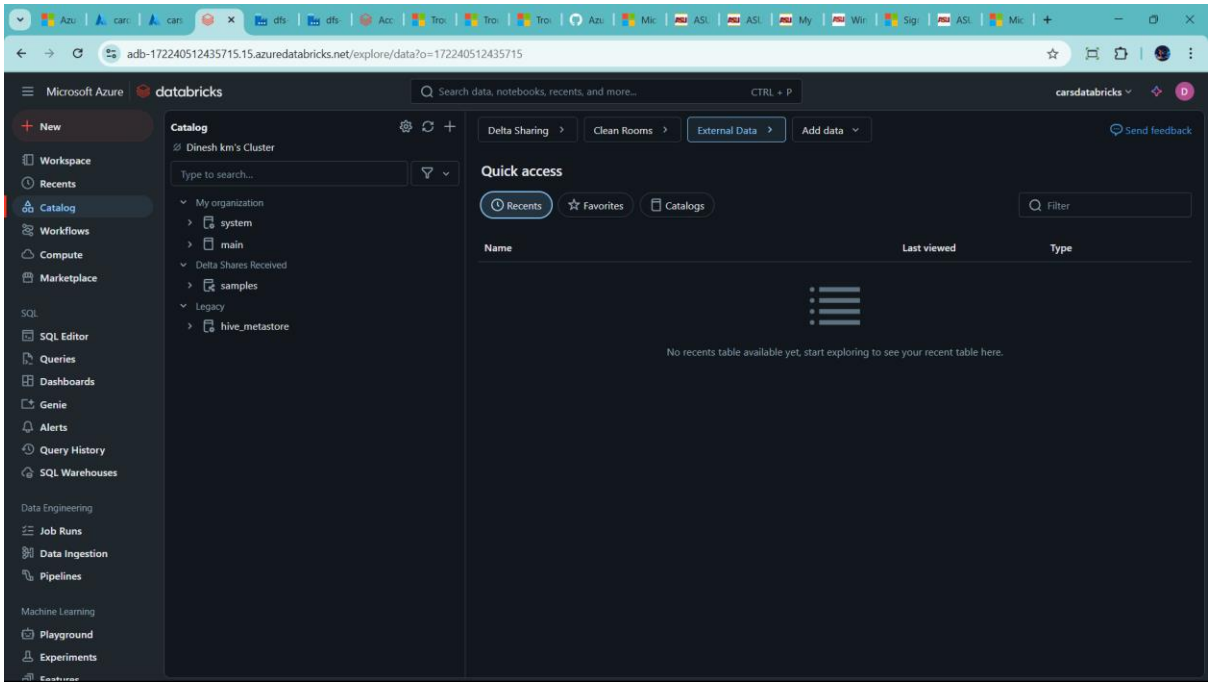

adb-1722405124357 15.15.azuredatabricks.net/explore/locations/create?o=1722405124357 15
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Create a new external location
An external location allows you to access your data stored in ADLS. You will need the ADLS path and a paired credential (e.g. managed identity) which gives access to that path Learn more
External location name*
bronzeext
URL*
Enter the bucket path that you want to use as the external location. Note: This must be an ADLS Gen2 storage account with a hierarchical namespace
abfss://brozen@cardineshdatalakestorage.dfs.core.windows.net/
Copy from DBFS
Storage credential* Learn more
Provide a storage credential capable of accessing the URL
dineshcreds (Managed Identity)
Connector Id: /subscriptions/d7c1e8a1-8bd2-4126-ba88-75140d084a4d/resourceGroups/RG_Azure_Car_Project/provider...
User Assigned Managed Identity Id:
Comment
Advanced Options
Cancel
Create

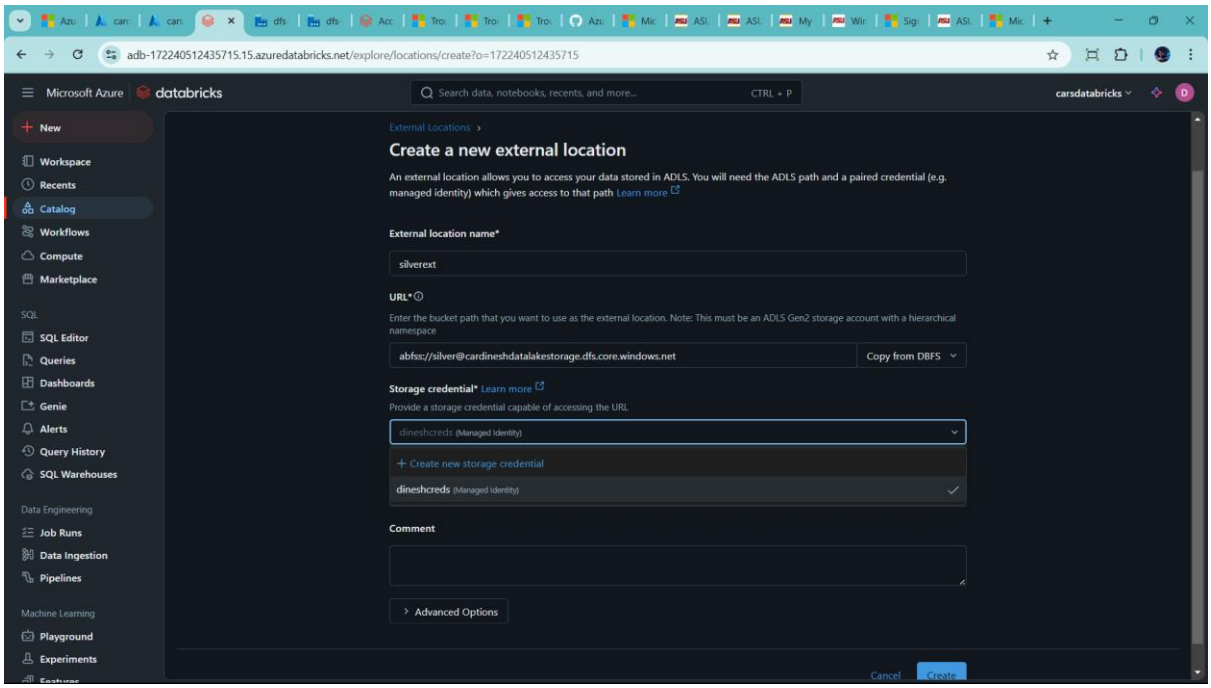


Test the connection

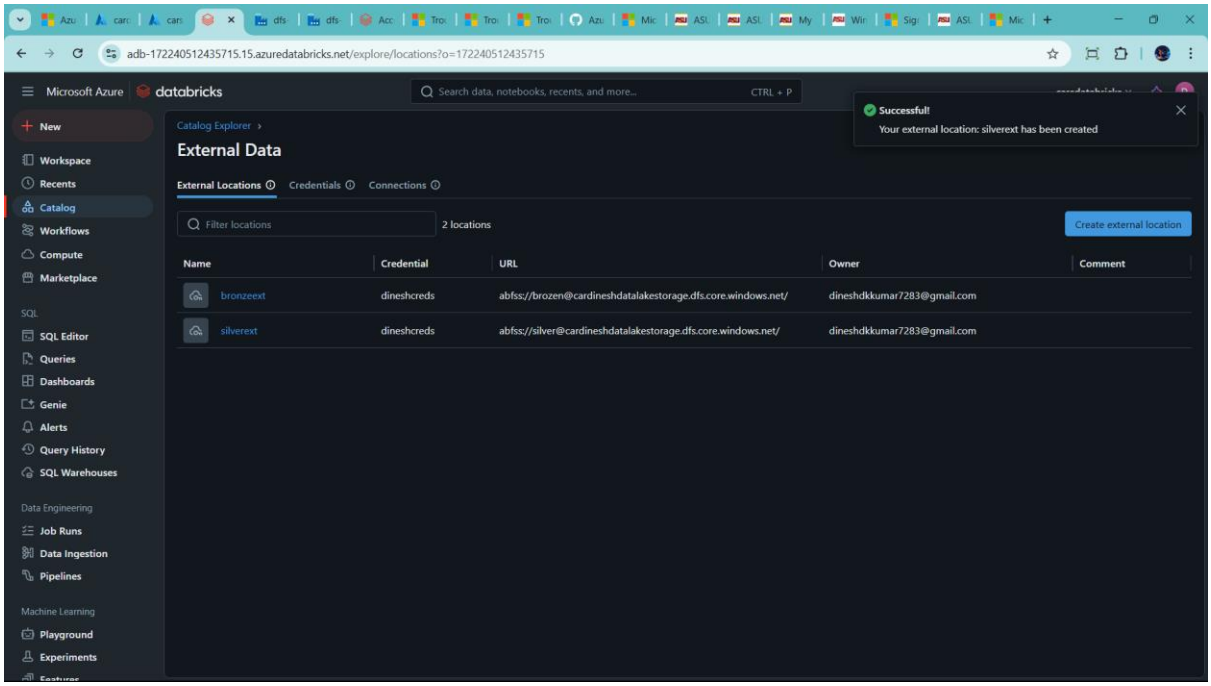


Create same like silver and gold

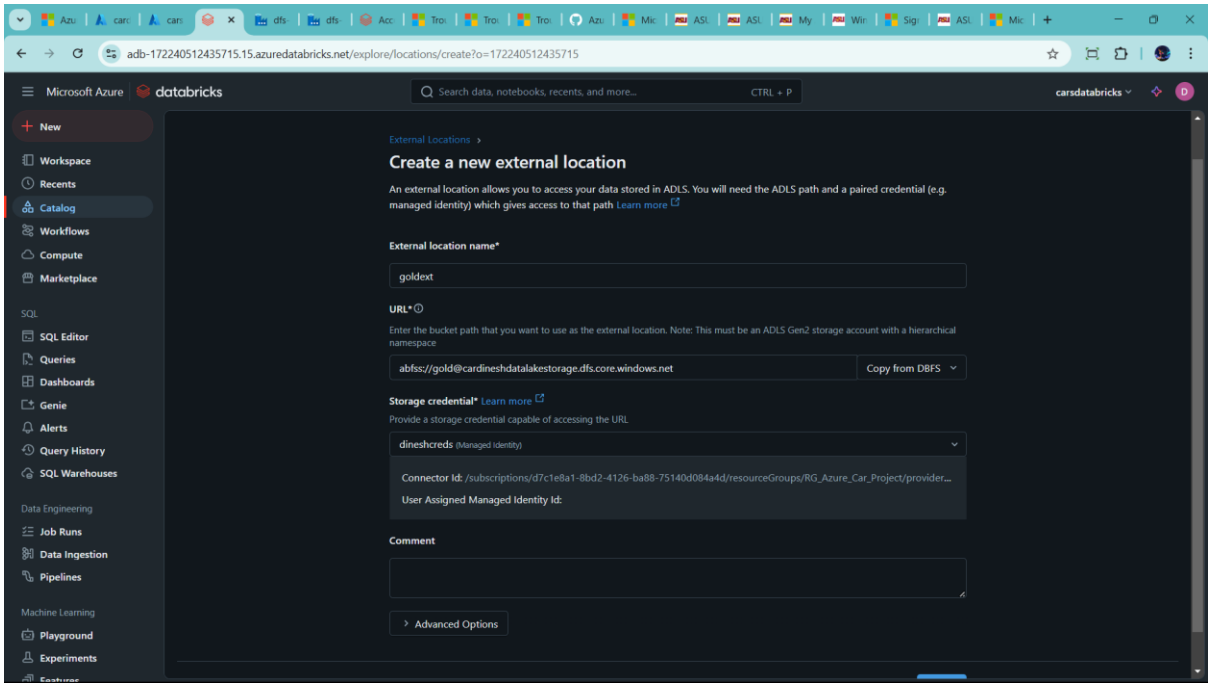


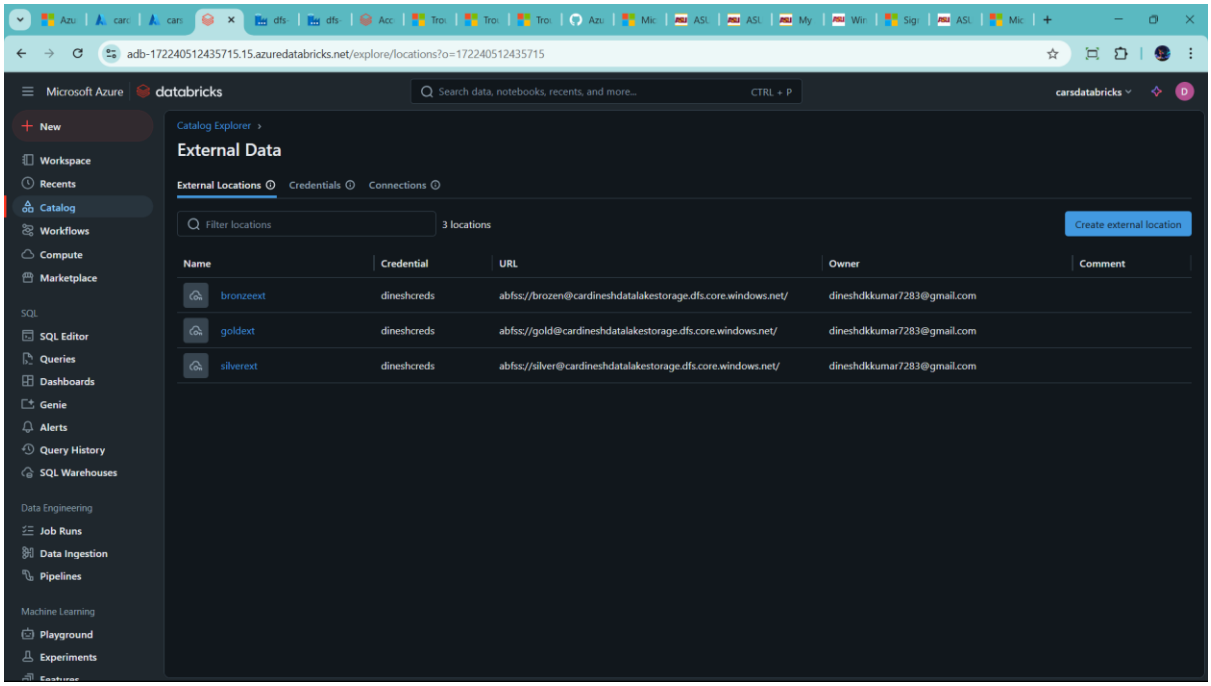


→click create


adb-1722405124357715.15.azuredatabricks.net/explore/locations?o=1722405124357715
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
Successful!
Your external location: silverext has been created
Catalog Explorer
External Data
External Locations
Credentials
Connections
Filter locations
2 locations
Create external location
Name
Credential
URL
Owner
Comment
bronzeext
dineshcreds
abfss://brozen@cardineshdatalakestorage.dfs.core.windows.net/
dineshdkkumar7283@gmail.com
silverext
dineshcreds
abfss://silver@cardineshdatalakestorage.dfs.core.windows.net/
dineshdkkumar7283@gmail.com
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments

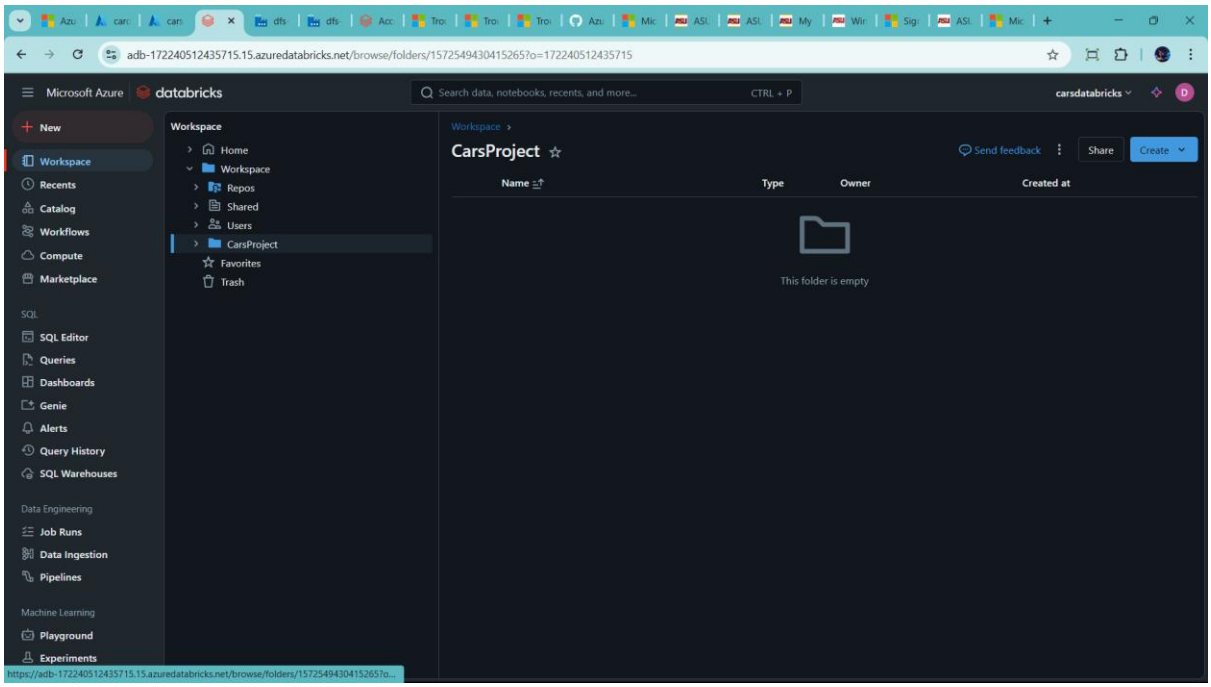
→same as gold


adb-1722405124357715.15.azuredatabricks.net/explore/locations/create?o=1722405124357715
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
External Locations
Create a new external location
An external location allows you to access your data stored in ADLS. You will need the ADLS path and a paired credential (e.g. managed identity) which gives access to that path Learn more
External location name*
goldext
URL*
Enter the bucket path that you want to use as the external location. Note: This must be an ADLS Gen2 storage account with a hierarchical namespace
abfss://gold@cardineshdatalakestorage.dfs.core.windows.net
Copy from DBFS
Storage credential* Learn more
Provide a storage credential capable of accessing the URL
dineshcreds (Managed Identity)
Connector Id: /subscriptions/d7c1e8a1-8bd2-4126-ba88-75140d084a4d/resourceGroups/RG_Azure_Car_Project/provider...
User Assigned Managed Identity Id:
Comment
Advanced Options
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments

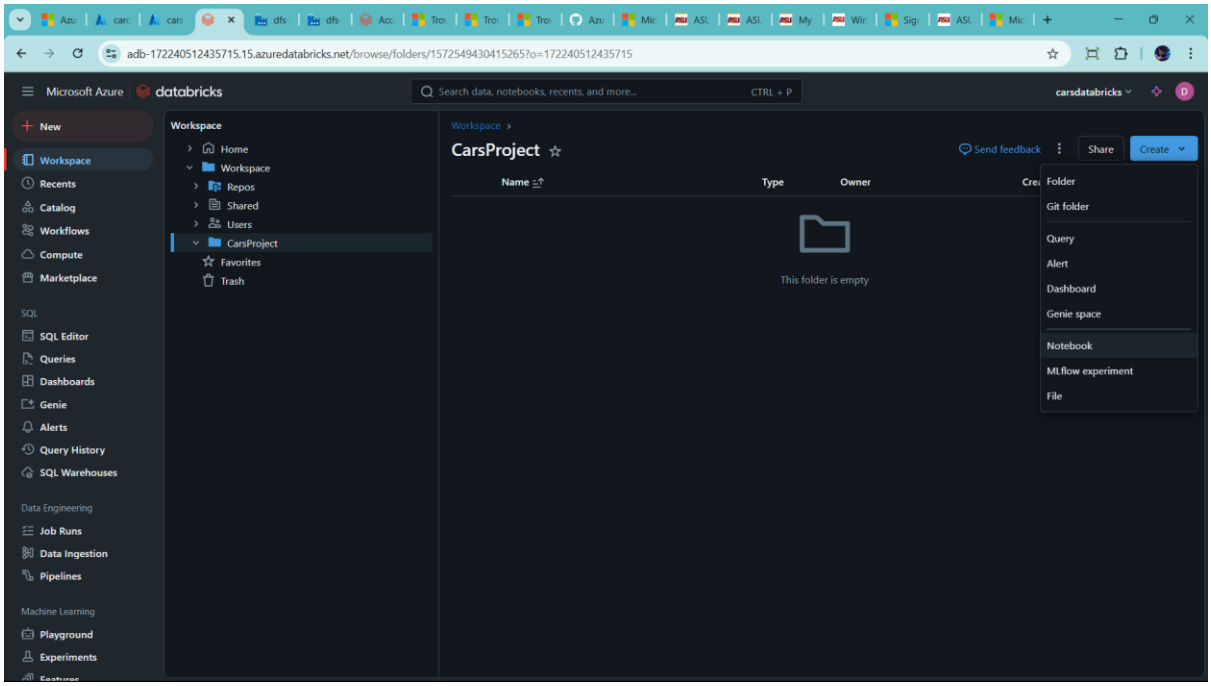


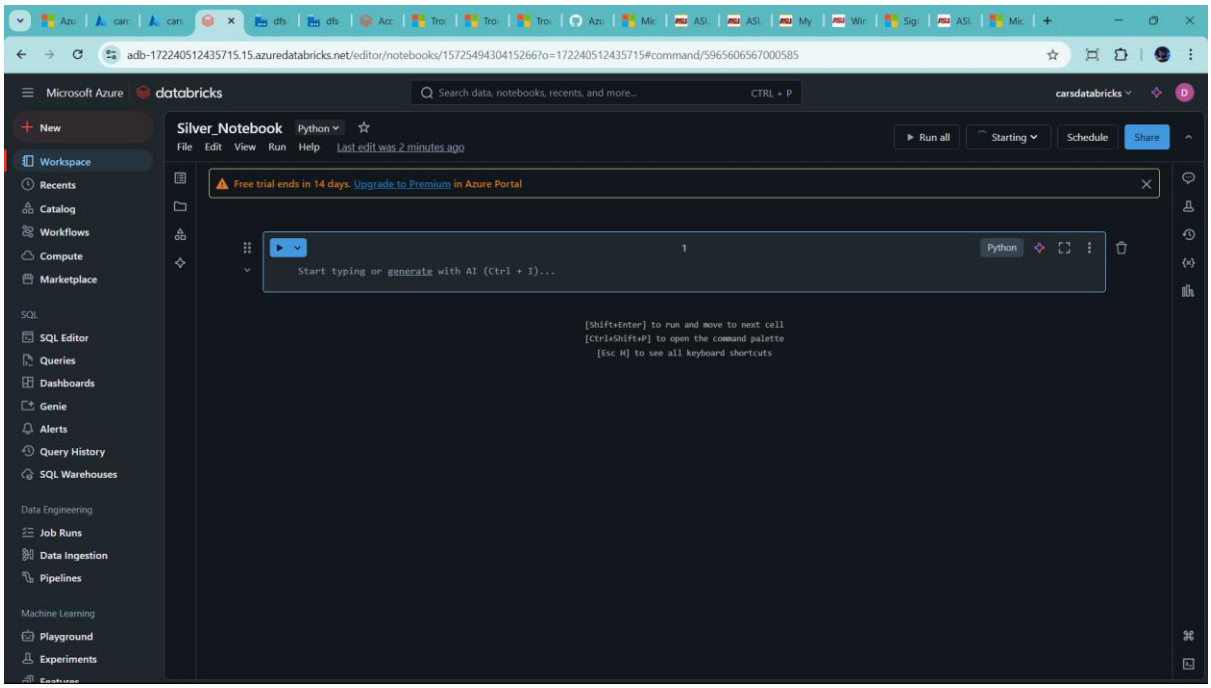
Now Start Create WorkSpaces

→click workspaces →add folder name

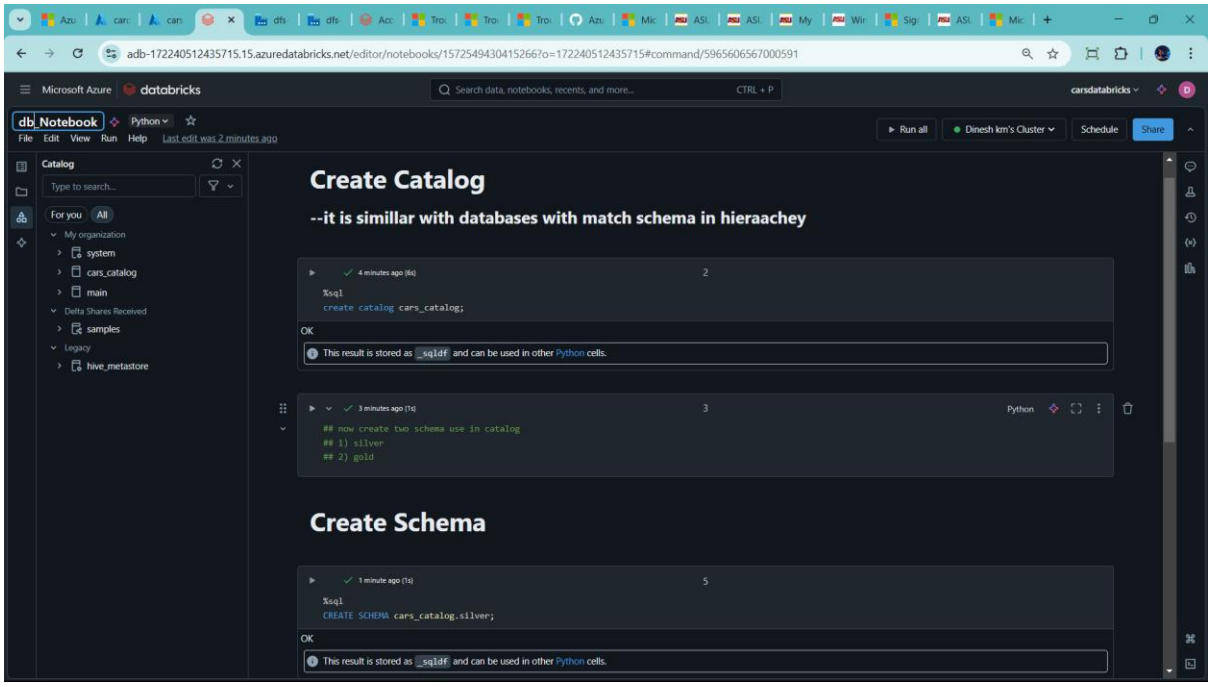


→click Carsproject → now create notebook

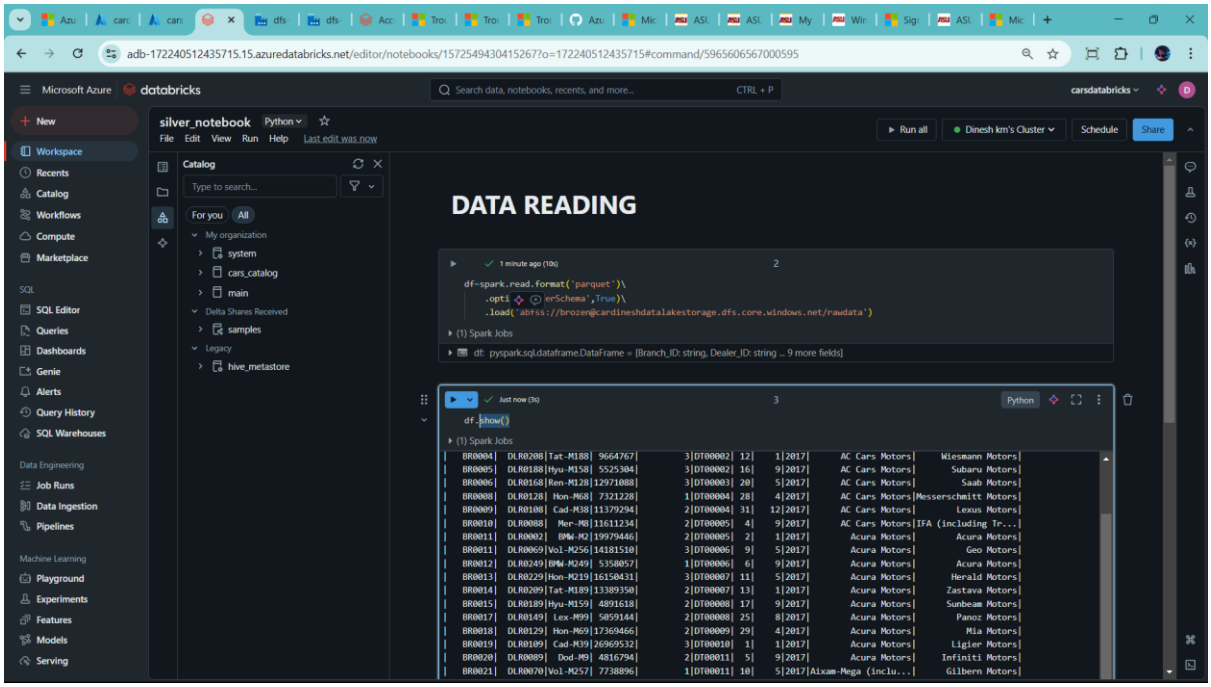

adb-1722405124357 15.15.azuredatabricks.net/browse/folders/1572549430415265?o=1722405124357 15
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Workspace
Home
Workspace
Repos
Shared
Users
CarsProject
Favorites
Trash
Workspace >
CarsProject
Send feedback
Share
Create
Name
Type
Owner
This folder is empty
Folder
Git folder
Query
Alert
Dashboard
Genie space
Notebook
MLflow experiment
File

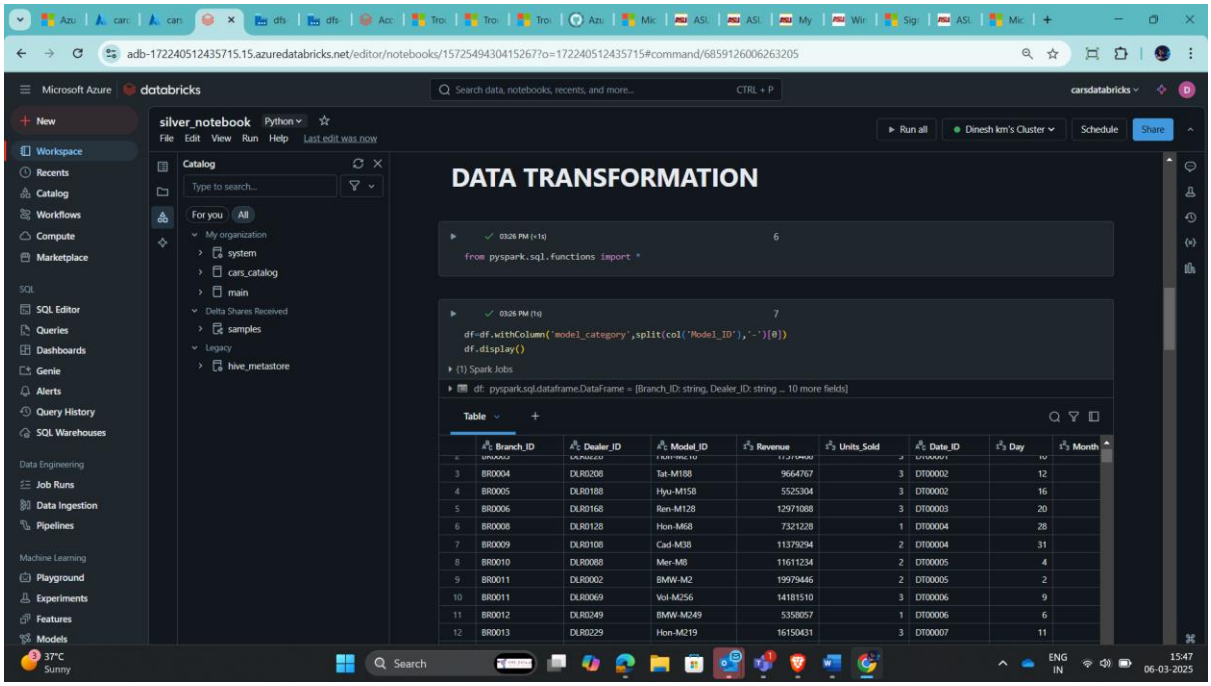

adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/1572549430415266?o=1722405124357 15#command/5965606567000585
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Silver_Notebook
Python
File Edit View Run Help
Last edit was 2 minutes ago
Run all
Starting
Schedule
Share
Free trial ends in 14 days. Upgrade to Premium in Azure Portal
1
Python
Start typing or generate with AI (Ctrl + I)...
[Shift+Enter] to run and move to next cell
[Ctrl+Shift+P] to open the command palette
[Esc H] to see all keyboard shortcuts

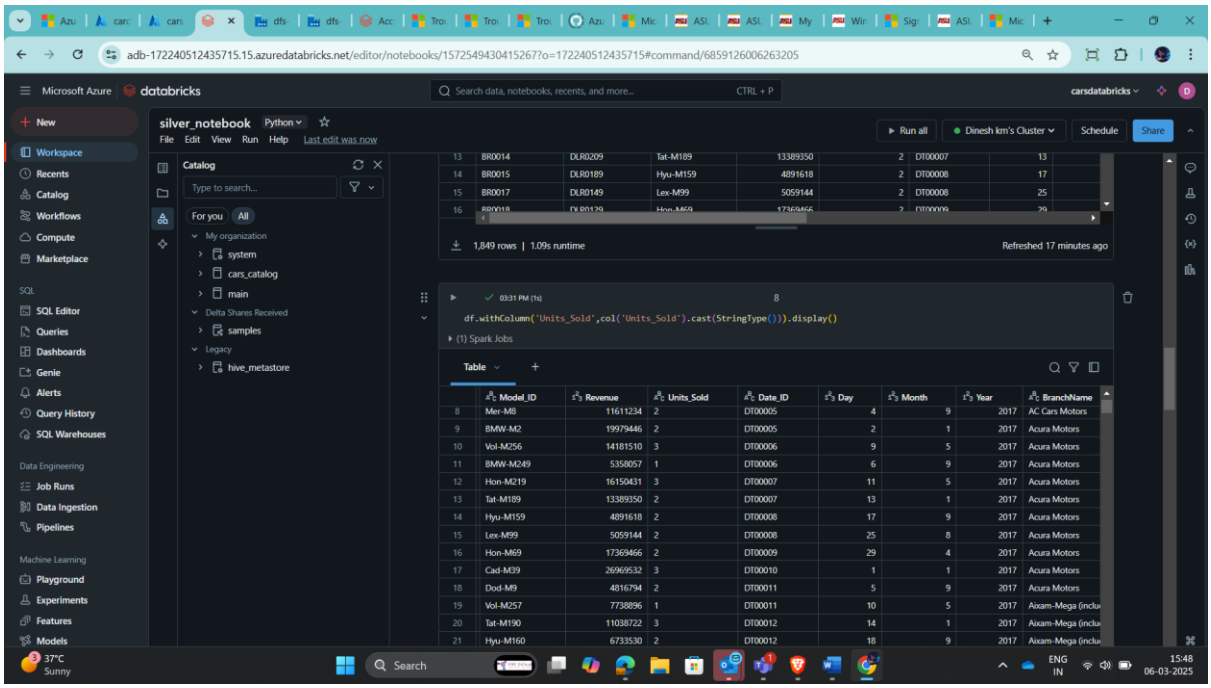
→rename db_Notebook

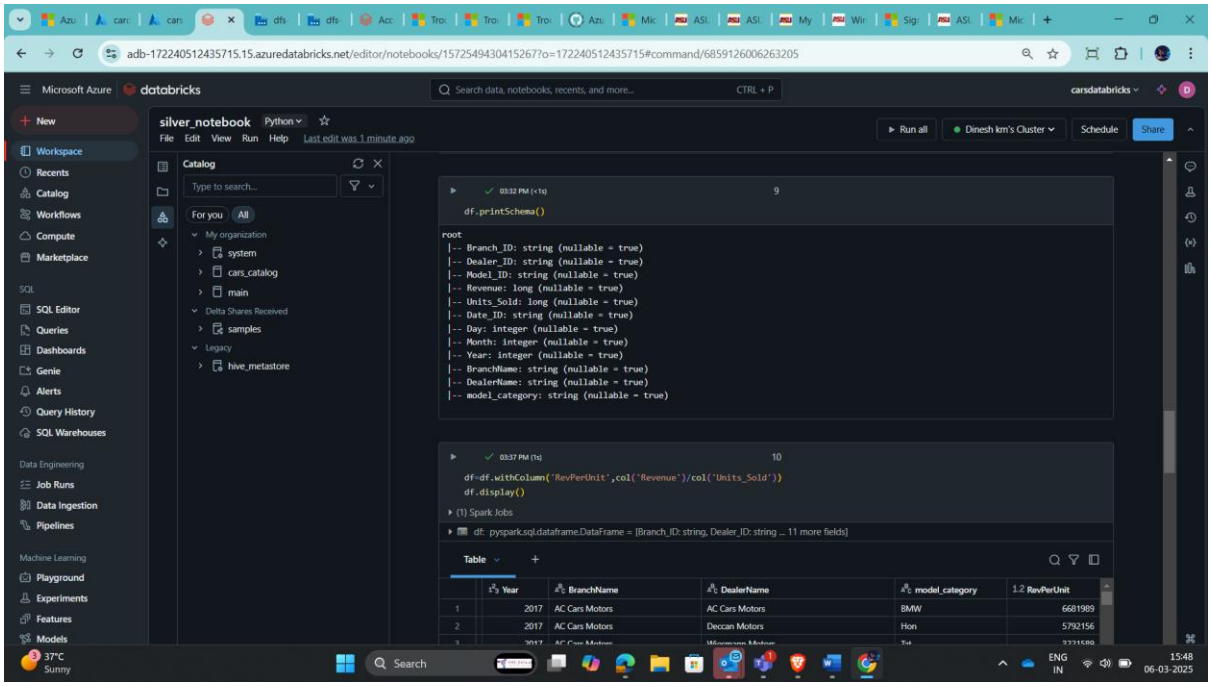


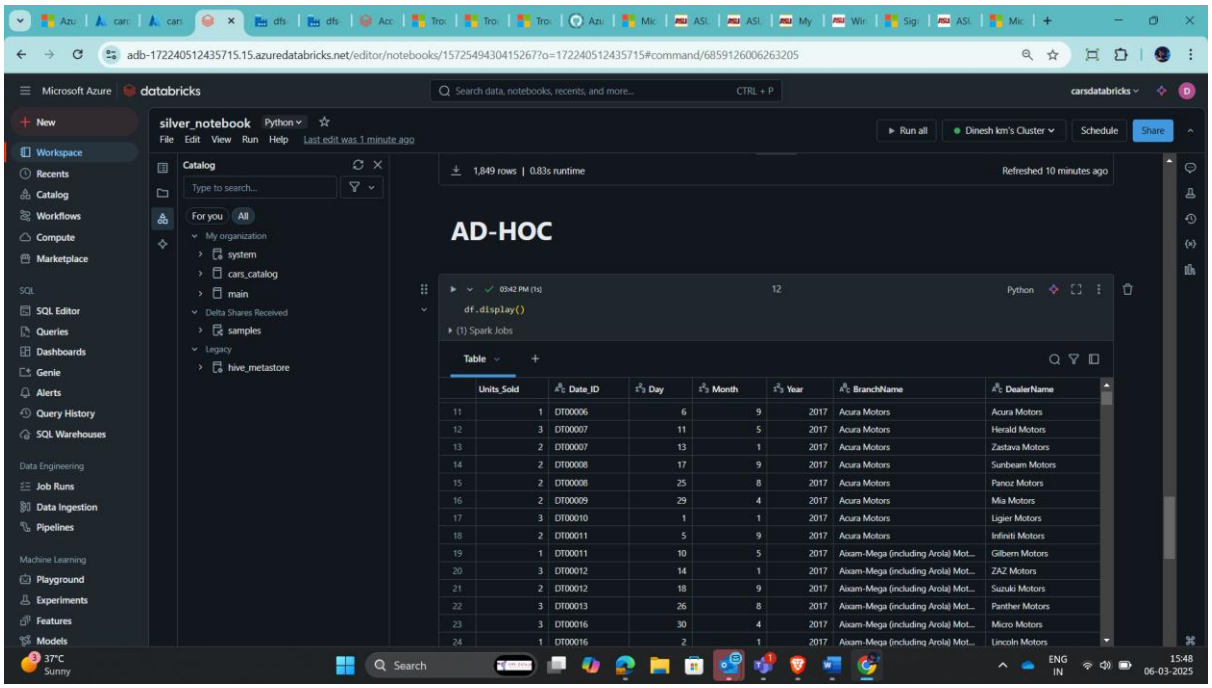
→create another notebook for reading data

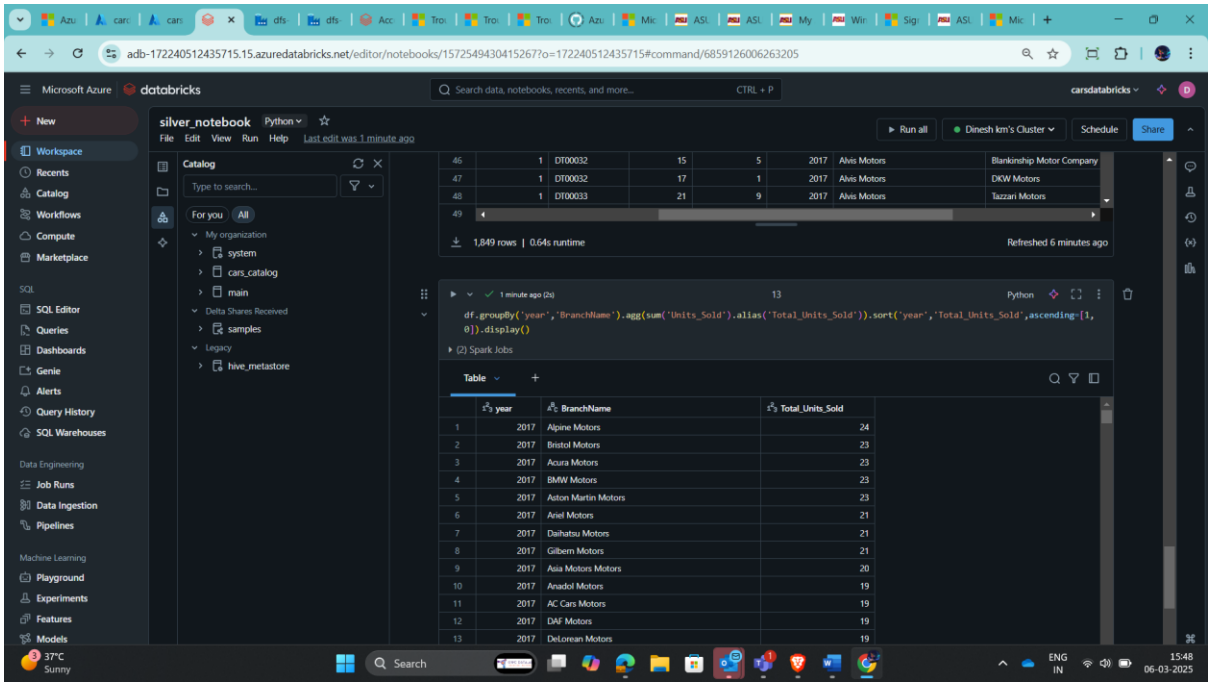


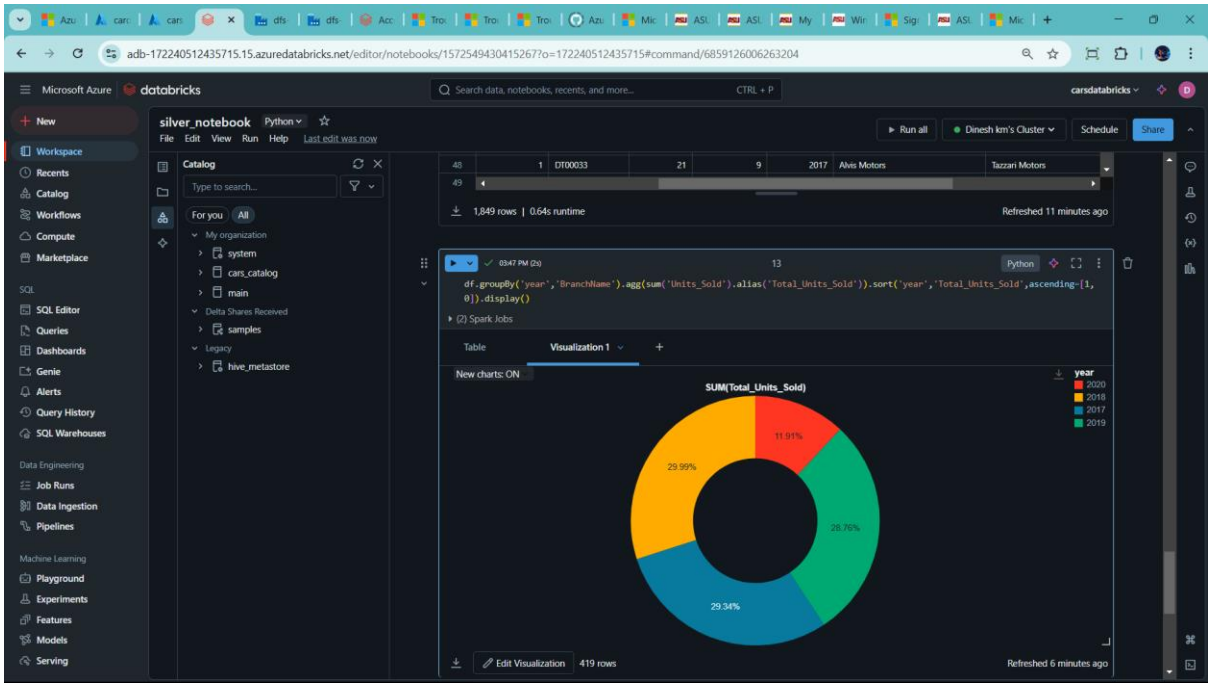

adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/1572549430415267?o=1722405124357 15#command/6859126006263205
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
silver_notebook
Python
File Edit View Run Help Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
Catalog
Type to search...
For you All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
DATA TRANSFORMATION
from pyspark.sql.functions import *
df=df.withColumn('model_category',split(col('Model_ID'),'-')[0])
df.display()
(1) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [Branch_ID: string, Dealer_ID: string ... 10 more fields]
Table
Branch_ID
Dealer_ID
Model_ID
Revenue
Units_Sold
Date_ID
Day
Month
BR0004 DLR0208 Tat-M188 9664767 3 DT00002 12
BR0005 DLR0188 Hyu-M158 5525304 3 DT00002 16
BR0006 DLR0168 Ren-M128 12971088 3 DT00003 20
BR0008 DLR0128 Hon-M68 7321228 1 DT00004 28
BR0009 DLR0108 Cad-M38 11379294 2 DT00004 31
BR0010 DLR0088 Mer-M8 11611234 2 DT00005 4
BR0011 DLR0002 BMW-M2 19979446 2 DT00005 2
BR0011 DLR0069 Vol-M256 14181510 3 DT00006 9
BR0012 DLR0249 BMW-M249 5358057 1 DT00006 6
BR0013 DLR0229 Hon-M219 16150431 3 DT00007 11
37°C
Sunny
Search
ENG
IN
15:47
06-03-2025

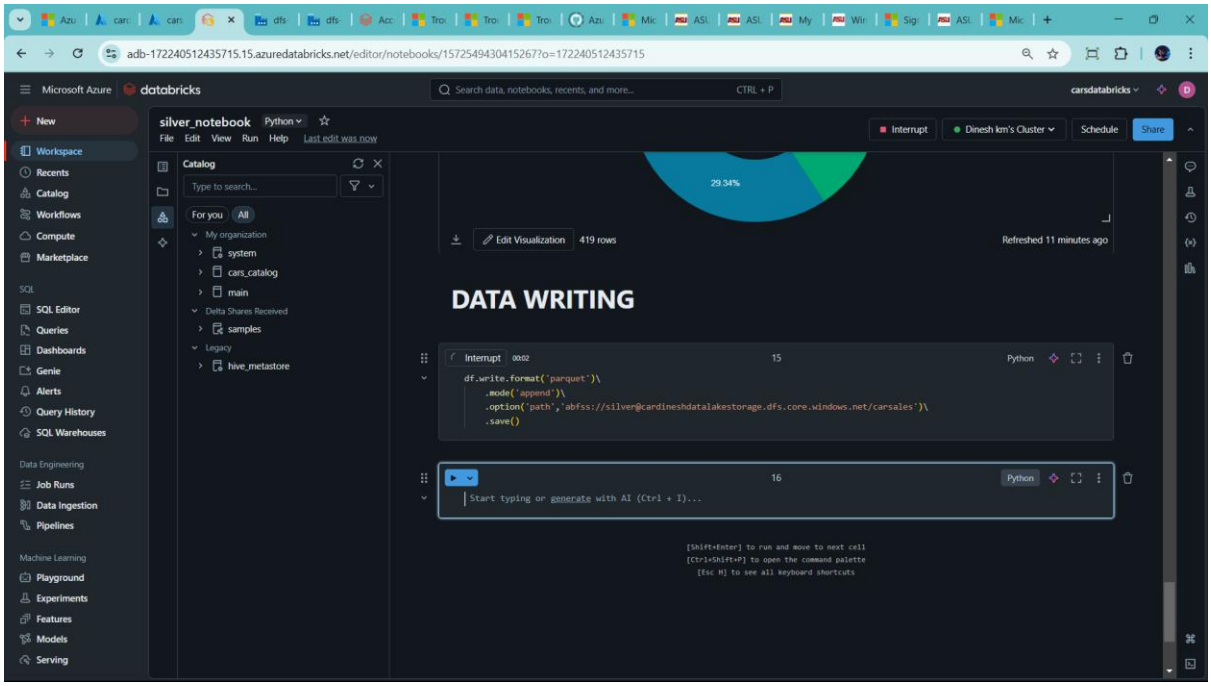

adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/1572549430415267?o=1722405124357 15#command/6859126006263205
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
silver_notebook
Python
File Edit View Run Help Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
Catalog
Type to search...
For you All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
BR0014 DLR0209 Tat-M189 13389350 2 DT00007 13
BR0015 DLR0189 Hyu-M159 4891618 2 DT00008 17
BR0017 DLR0149 Lex-M99 5059144 2 DT00008 25
1,849 rows | 1.09s runtime
Refreshed 17 minutes ago
df.withColumn('Units_Sold',col('Units_Sold').cast(StringType())).display()
(1) Spark Jobs
Table
Model_ID
Revenue
Units_Sold
Date_ID
Day
Month
Year
BranchName
Mer-M8 11611234 2 DT00005 4 9 2017 AC Cars Motors
BMW-M2 19979446 2 DT00005 2 1 2017 Acura Motors
Vol-M256 14181510 3 DT00006 9 5 2017 Acura Motors
BMW-M249 5358057 1 DT00006 6 9 2017 Acura Motors
Hon-M219 16150431 3 DT00007 11 5 2017 Acura Motors
Tat-M189 13389350 2 DT00007 13 1 2017 Acura Motors
Hyu-M159 4891618 2 DT00008 17 9 2017 Acura Motors
Lex-M99 5059144 2 DT00008 25 8 2017 Acura Motors
Hon-M69 17369466 2 DT00009 29 4 2017 Acura Motors
Cad-M39 26969532 3 DT00010 1 1 2017 Acura Motors
Dod-M9 4816794 2 DT00011 5 9 2017 Acura Motors
Vol-M257 7738896 1 DT00011 10 5 2017 Aixam-Mega (inclu
Tat-M190 11038722 3 DT00012 14 1 2017 Aixam-Mega (inclu
Hyu-M160 6733530 2 DT00012 18 9 2017 Aixam-Mega (inclu
37°C
Sunny
Search
ENG
IN
15:48
06-03-2025


adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/1572549430415267?o=1722405124357 15#command/6859126006263205
Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
carsdatabricks
silver_notebook
Python
File Edit View Run Help Last edit was 1 minute ago
Run all
Dinesh km's Cluster
Schedule
Share
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Features
Models
Catalog
Type to search...
For you All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
df.printSchema()
root
|-- Branch_ID: string (nullable = true)
|-- Dealer_ID: string (nullable = true)
|-- Model_ID: string (nullable = true)
|-- Revenue: long (nullable = true)
|-- Units_Sold: long (nullable = true)
|-- Date_ID: string (nullable = true)
|-- Day: integer (nullable = true)
|-- Month: integer (nullable = true)
|-- Year: integer (nullable = true)
|-- BranchName: string (nullable = true)
|-- DealerName: string (nullable = true)
|-- model_category: string (nullable = true)
df=df.withColumn('RevPerUnit',col('Revenue')/col('Units_Sold'))
df.display()
(1) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [Branch_ID: string, Dealer_ID: string ... 11 more fields]
Table
Year
BranchName
DealerName
model_category
RevPerUnit
1 2017 AC Cars Motors AC Cars Motors BMW 6681989
2 2017 AC Cars Motors Deccan Motors Hon 5792156
37°C Sunny
Search
ENG IN
15:48 06-03-2025

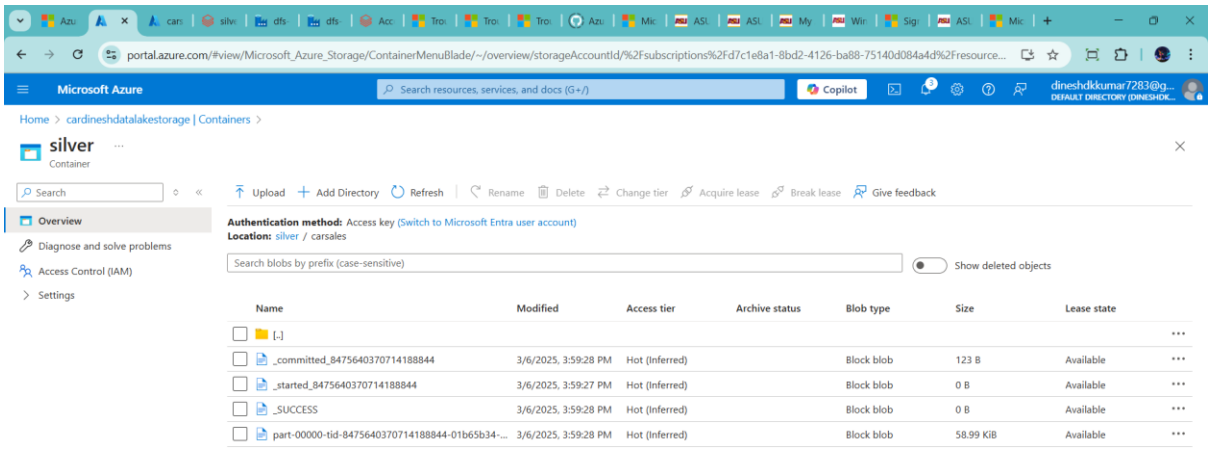

adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/1572549430415267?o=1722405124357 15#command/6859126006263205
Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
carsdatabricks
silver_notebook
Python
File Edit View Run Help Last edit was 1 minute ago
Run all
Dinesh km's Cluster
Schedule
Share
1,849 rows | 0.83s runtime
Refreshed 10 minutes ago
AD-HOC
df.display()
(1) Spark Jobs
Table
Units_Sold
Date_ID
Day
Month
Year
BranchName
DealerName
11 1 DT00006 6 9 2017 Acura Motors Acura Motors
12 3 DT00007 11 5 2017 Acura Motors Herald Motors
13 2 DT00007 13 1 2017 Acura Motors Zastava Motors
14 2 DT00008 17 9 2017 Acura Motors Sunbeam Motors
15 2 DT00008 25 8 2017 Acura Motors Panoz Motors
16 2 DT00009 29 4 2017 Acura Motors Mia Motors
17 3 DT00010 1 1 2017 Acura Motors Ligier Motors
18 2 DT00011 5 9 2017 Acura Motors Infiniti Motors
19 1 DT00011 10 5 2017 Aixam-Mega (including Arola) Mot... Gilbern Motors
20 3 DT00012 14 1 2017 Aixam-Mega (including Arola) Mot... ZAZ Motors
21 2 DT00012 18 9 2017 Aixam-Mega (including Arola) Mot... Suzuki Motors
22 3 DT00013 26 8 2017 Aixam-Mega (including Arola) Mot... Panther Motors
23 3 DT00016 30 4 2017 Aixam-Mega (including Arola) Mot... Micro Motors
24 1 DT00016 2 1 2017 Aixam-Mega (including Arola) Mot... Lincoln Motors
37°C Sunny
Search
ENG IN
15:48 06-03-2025

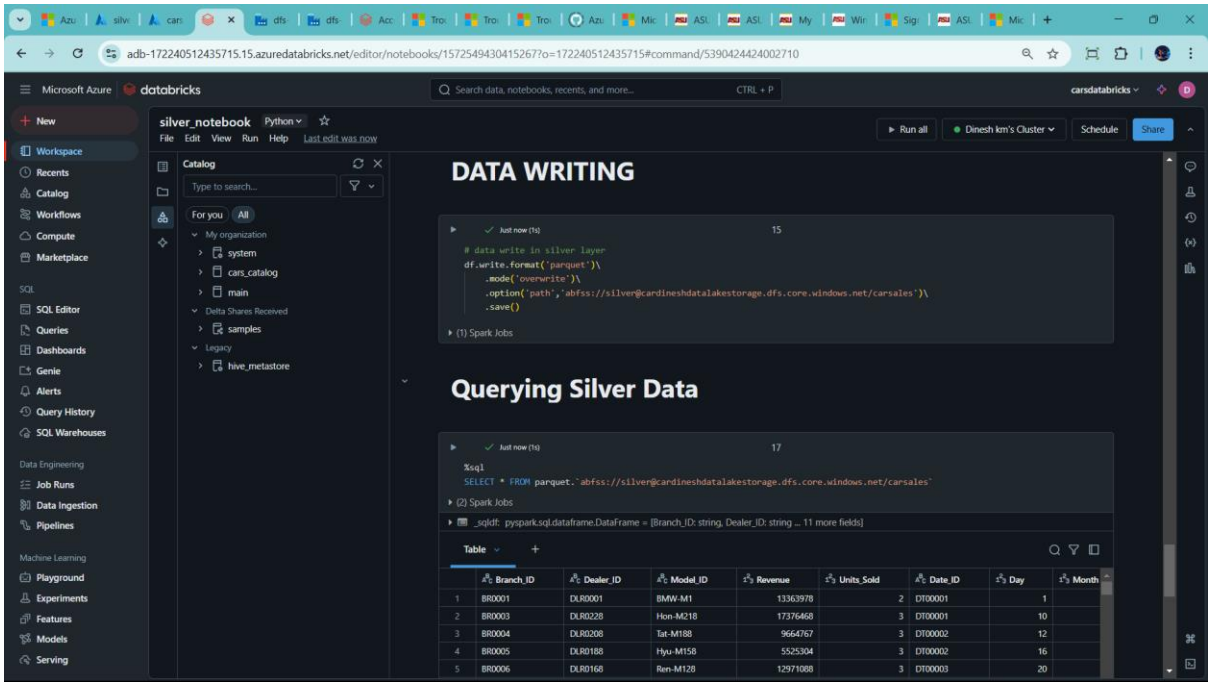

adb-1722405512435715.15.azuredatabricks.net/editor/notebooks/1572549430415267?o=1722405512435715#command/6859126006263205
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
silver_notebook
Python
File Edit View Run Help
Last edit was 1 minute ago
Run all
Dinesh km's Cluster
Schedule
Share
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
46 1 DT00032 15 5 2017 Alvis Motors Blankinship Motor Company
47 1 DT00032 17 1 2017 Alvis Motors DKW Motors
48 1 DT00033 21 9 2017 Alvis Motors Tazzari Motors
49
1,849 rows | 0.64s runtime
Refreshed 6 minutes ago
1 minute ago (2s)
13
Python
df.groupBy('year','BranchName').agg(sum('Units_Sold').alias('Total_Units_Sold')).sort('year','Total_Units_Sold',ascending=[1,0]).display()
(2) Spark Jobs
Table
year
BranchName
Total_Units_Sold
1 2017 Alpine Motors 24
2 2017 Bristol Motors 23
3 2017 Acura Motors 23
4 2017 BMW Motors 23
5 2017 Aston Martin Motors 23
6 2017 Ariel Motors 21
7 2017 Daihatsu Motors 21
8 2017 Gilbern Motors 21
9 2017 Asia Motors Motors 20
10 2017 Anadol Motors 19
11 2017 AC Cars Motors 19
12 2017 DAF Motors 19
13 2017 DeLorean Motors 19
37°C
Sunny
Search
ENG
IN
15:48
06-03-2025


adb-1722405512435715.15.azuredatabricks.net/editor/notebooks/1572549430415267?o=1722405512435715#command/6859126006263204
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
silver_notebook
Python
File Edit View Run Help
Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
48 1 DT00033 21 9 2017 Alvis Motors Tazzari Motors
49
1,849 rows | 0.64s runtime
Refreshed 11 minutes ago
03:47 PM (2s)
13
Python
df.groupBy('year','BranchName').agg(sum('Units_Sold').alias('Total_Units_Sold')).sort('year','Total_Units_Sold',ascending=[1,0]).display()
(2) Spark Jobs
Table
Visualization 1
New charts: ON
SUM(Total_Units_Sold)
year
2020
2018
2017
2019
11.91%
29.99%
28.76%
29.34%
Edit Visualization
419 rows
Refreshed 6 minutes ago



Folder and file success added

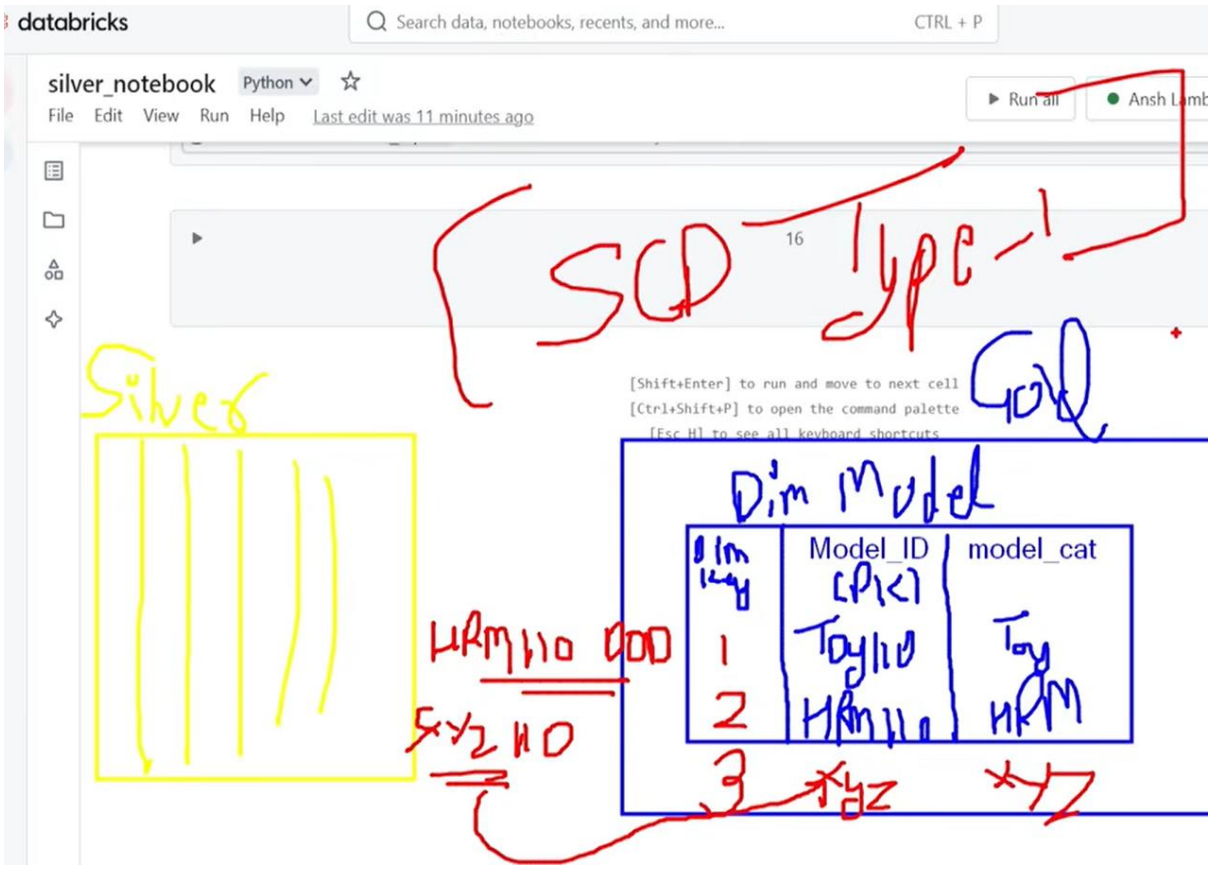




-→ Now Create Data Modelling

→Star Schema →so Create New Notebook → for gold Layer

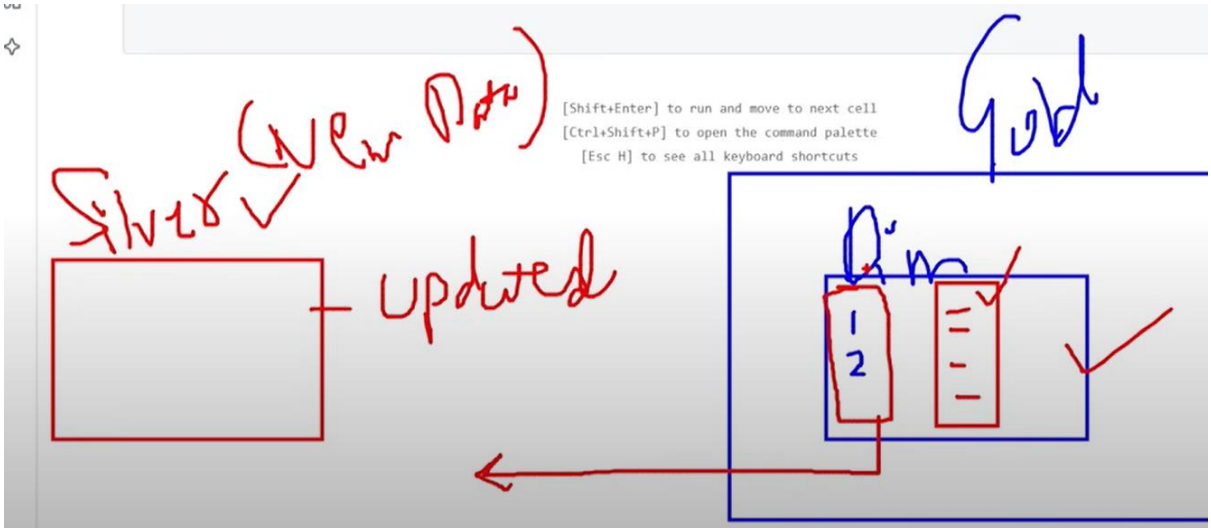
→slow change dimension



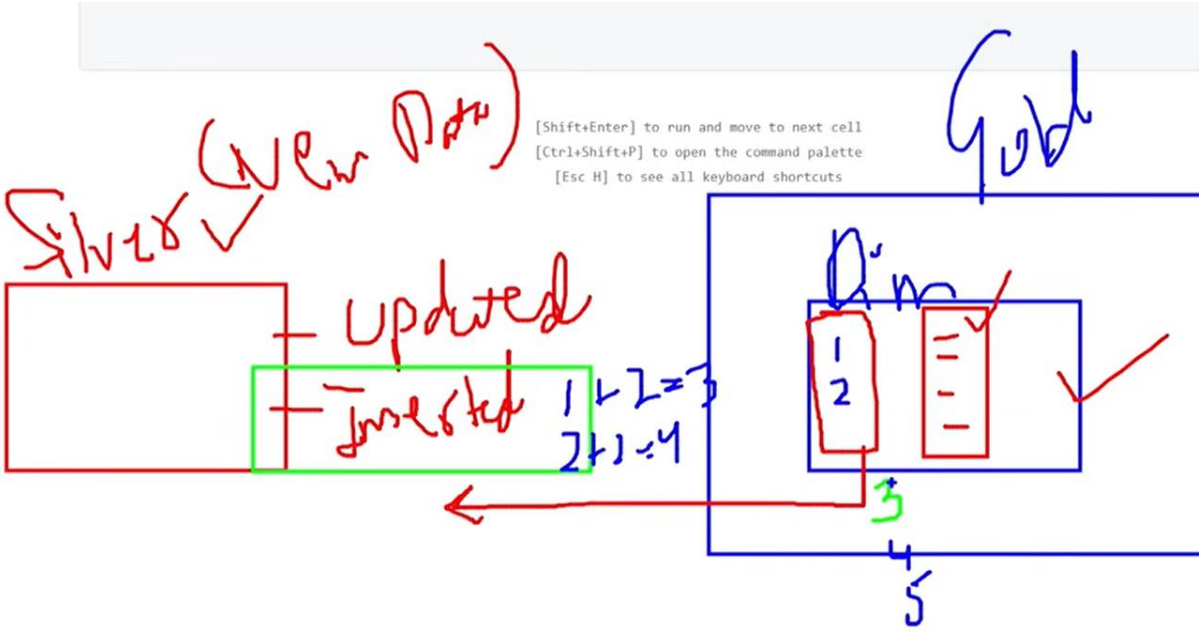
→gold table have dim id, id,name

→the silver table have new data →let check gold id but dim id, because dim id not present in silver table

➔ Then id of gold table join with silver table. If left join both table new records in silver but not in gold table→ it means dim id return null to inform insert new records present



→dim id is surrogate key



>

→ first time run gold table use left join because gold layer not available table

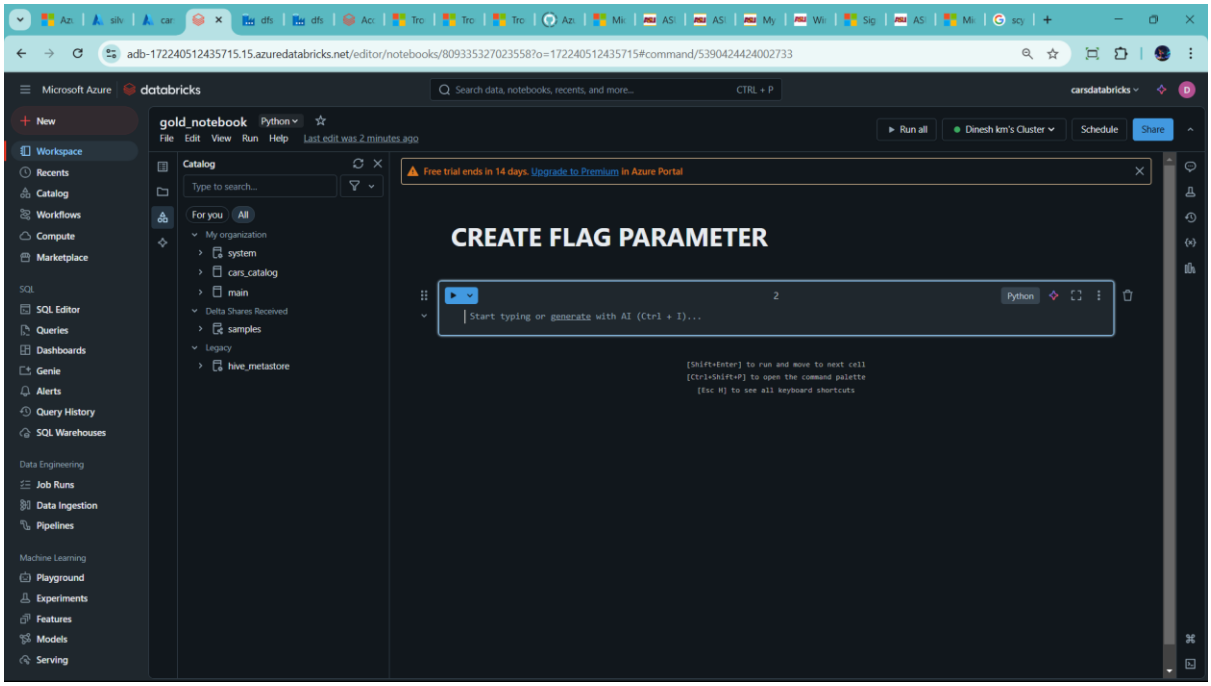
TYPE – 1 (UPSERT)

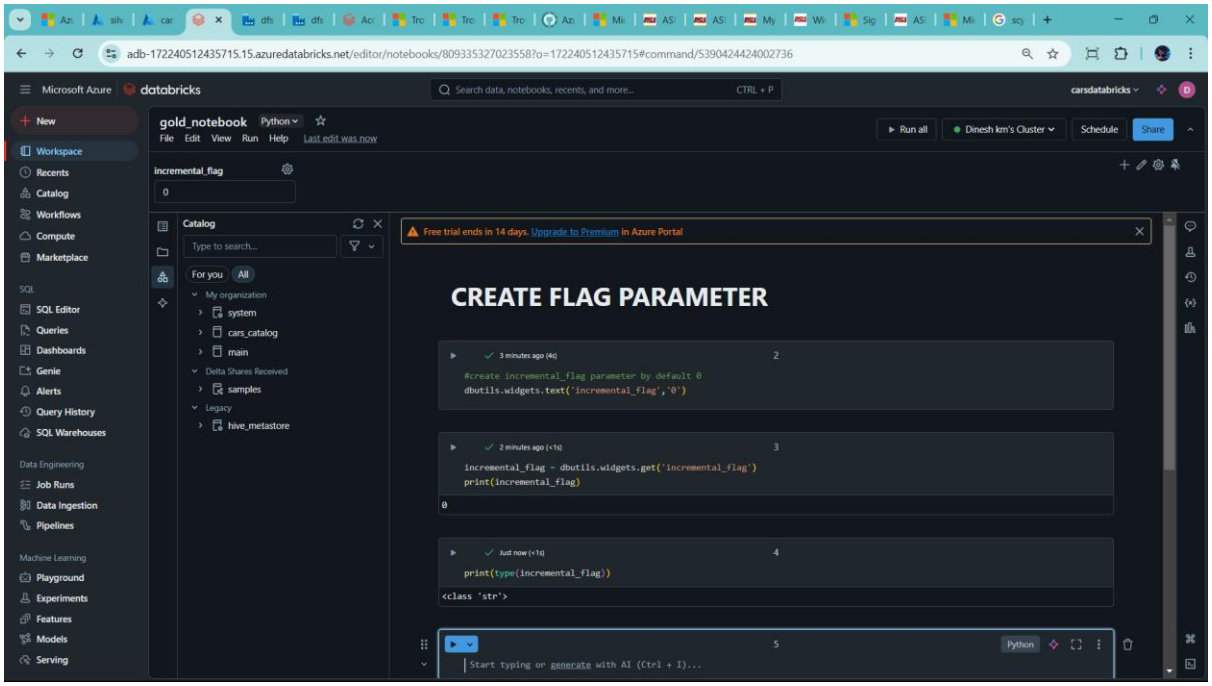
Product_ID	Name	Prod_Cat
1	Honey	Food
2	Shirt	Clothing
3	Comb	Clothing

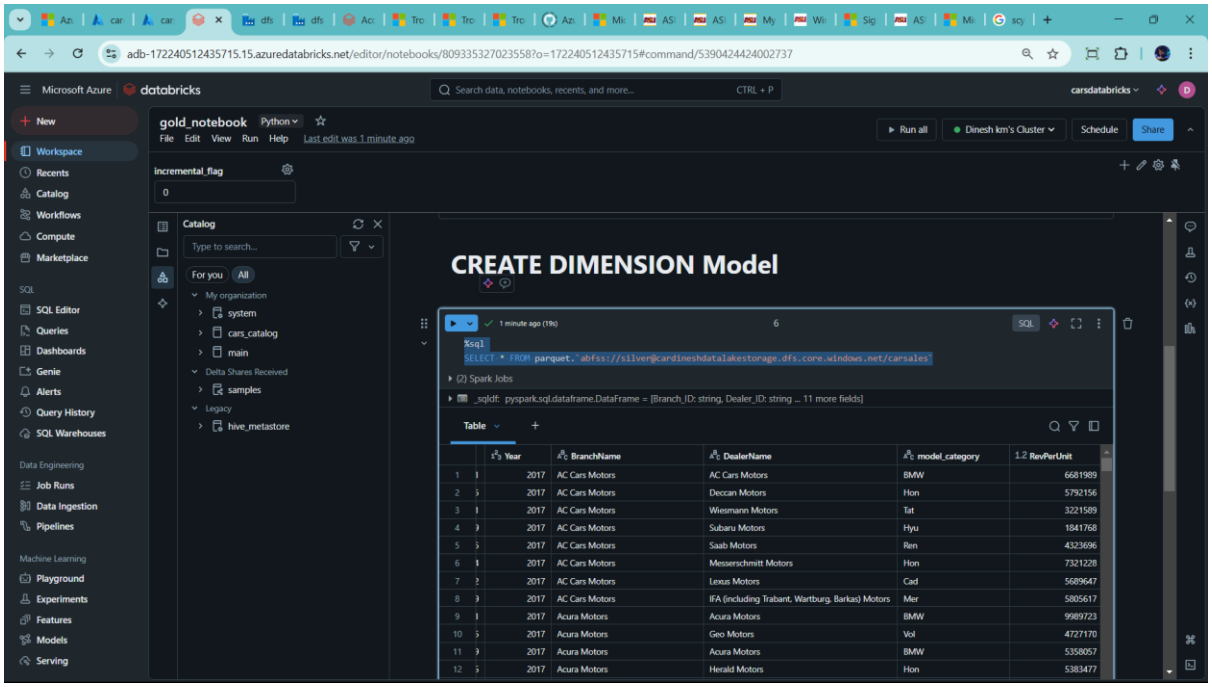
Before

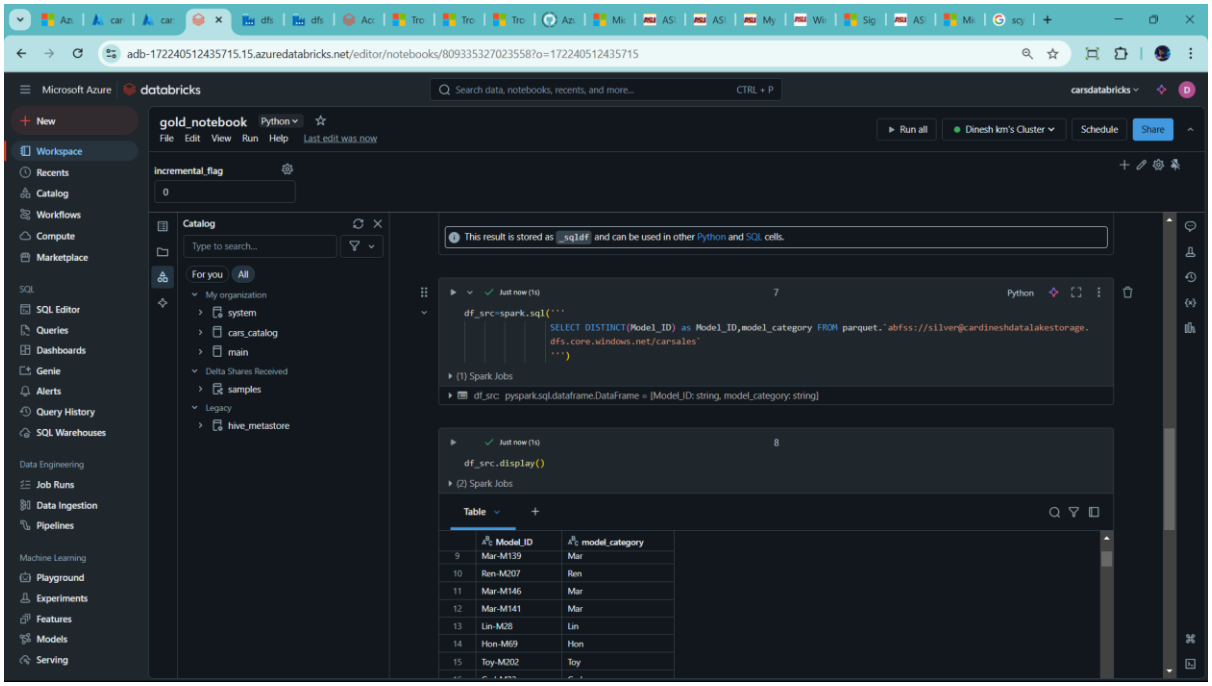
Product_ID	Name	Prod_Cat
1	Honey	Food
2	Shirt	Clothing
3	Comb	Hair

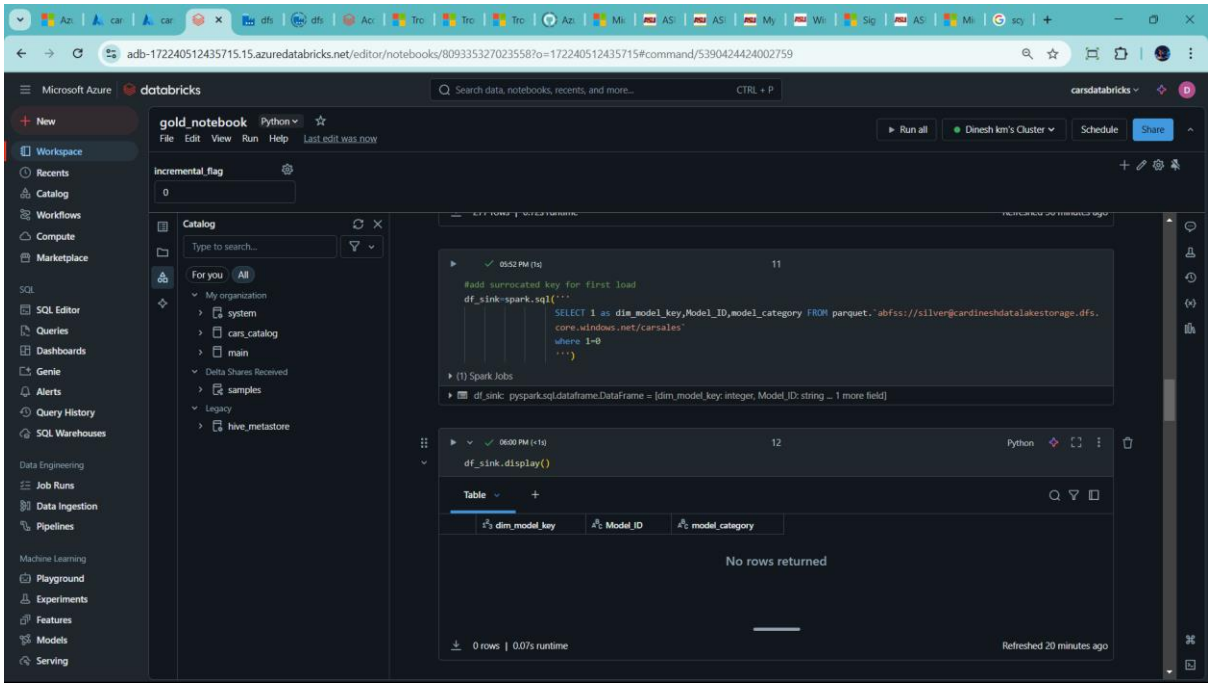
AFTER

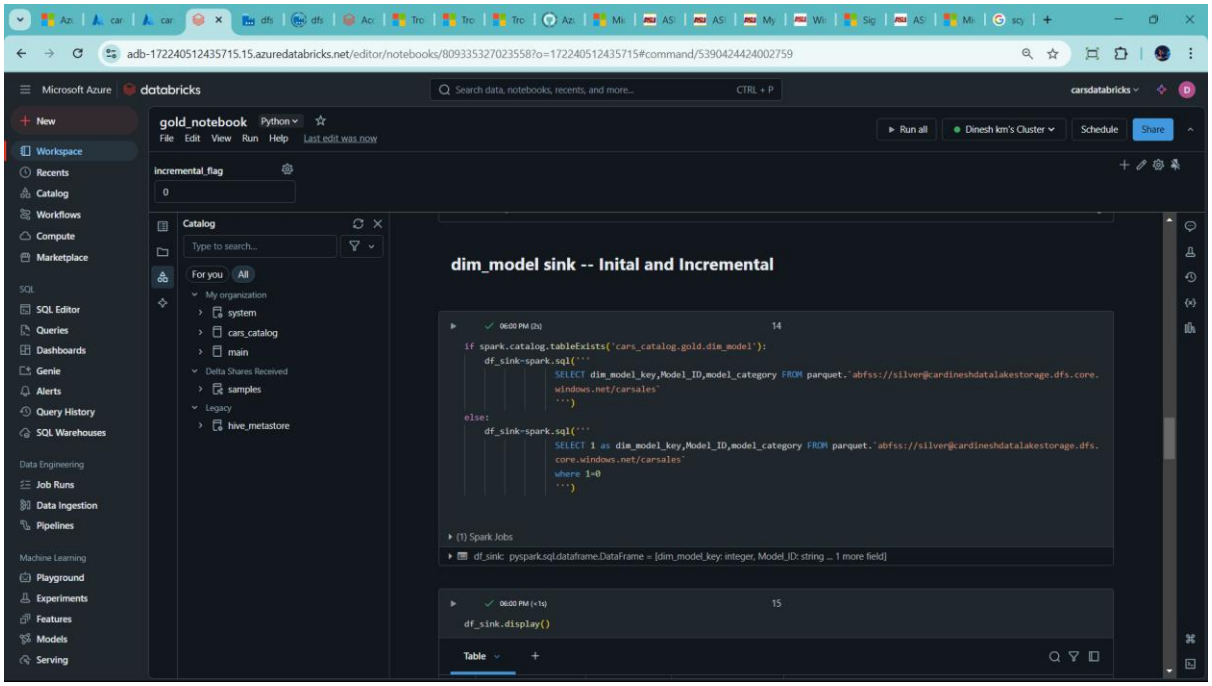


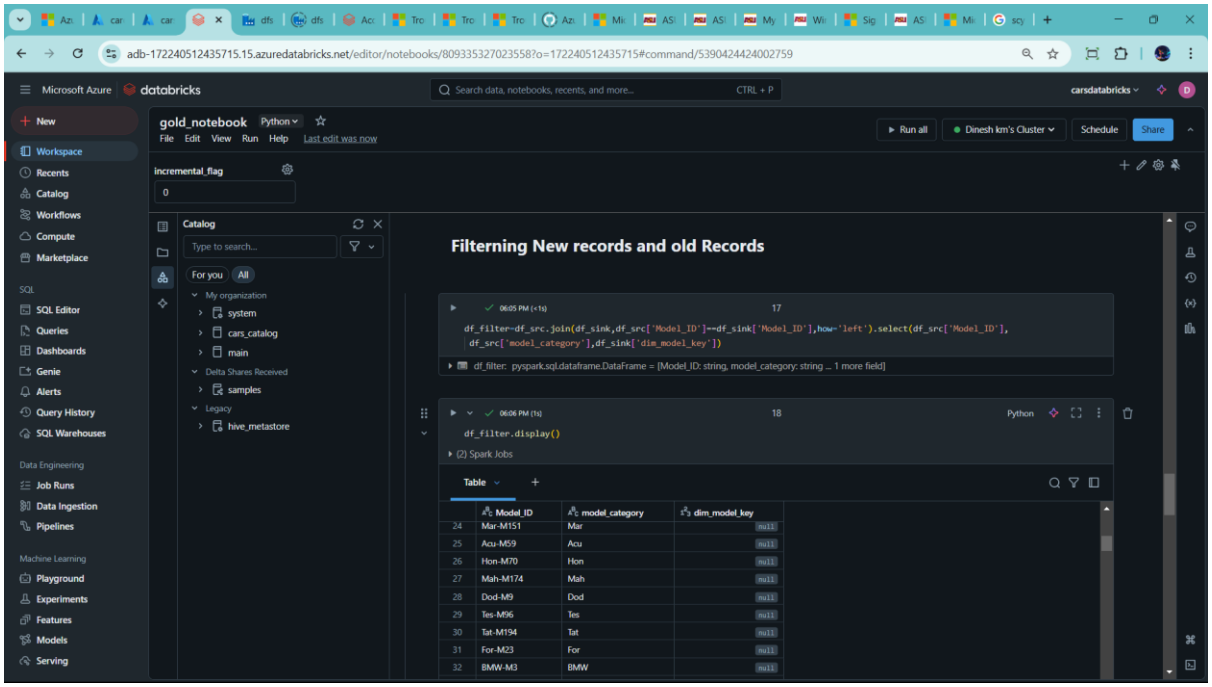

adb-1722405512435715.15.azuredatabricks.net/editor/notebooks/809335327023558?o=1722405512435715#command/5390424424002736
Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
carsdatabricks
gold_notebook
Python
File Edit View Run Help
Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
Free trial ends in 14 days. Upgrade to Premium in Azure Portal
CREATE FLAG PARAMETER
3 minutes ago (4s)
2
#create incremental_flag parameter by default 0
dbutils.widgets.text('incremental_flag','0')
2 minutes ago (<1s)
3
incremental_flag = dbutils.widgets.get('incremental_flag')
print(incremental_flag)
0
Just now (<1s)
4
print(type(incremental_flag))
<class 'str'>
5
Python
Start typing or generate with AI (Ctrl + I)...
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Features
Models
Serving

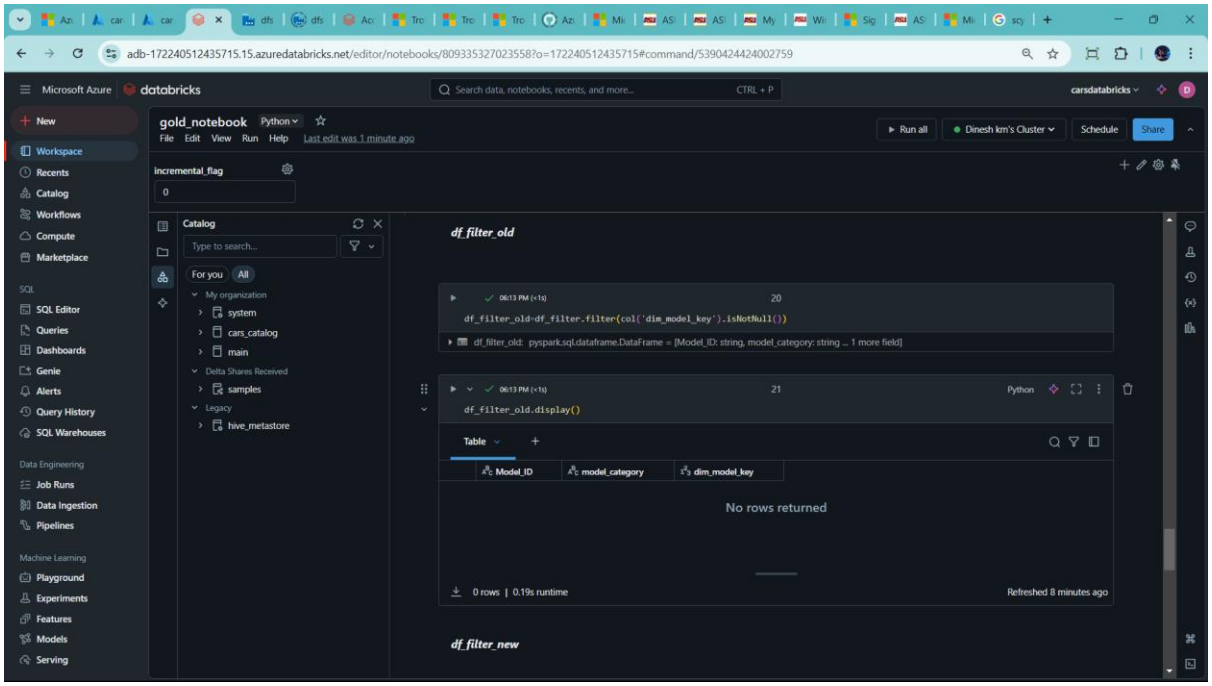

adb-1722405512435715.15.azuredatabricks.net/editor/notebooks/809335327023558?o=1722405512435715#command/5390424424002737
Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
carsdatabricks
gold_notebook
Python
File Edit View Run Help
Last edit was 1 minute ago
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
CREATE DIMENSION Model
1 minute ago (19s)
6
SQL
%sql
SELECT * FROM parquet.`abfss://silver@cardineshdatalakestorage.dfs.core.windows.net/carsales`
(2) Spark Jobs
_sqldf: pyspark.sql.dataframe.DataFrame = [Branch_ID: string, Dealer_ID: string ... 11 more fields]
Table
Year
BranchName
DealerName
model_category
RevPerUnit
1 2017 AC Cars Motors AC Cars Motors BMW 6681989
2 2017 AC Cars Motors Deccan Motors Hon 5792156
3 2017 AC Cars Motors Wiesmann Motors Tat 3221589
4 2017 AC Cars Motors Subaru Motors Hyu 1841768
5 2017 AC Cars Motors Saab Motors Ren 4323696
6 2017 AC Cars Motors Messerschmitt Motors Hon 7321228
7 2017 AC Cars Motors Lexus Motors Cad 5689647
8 2017 AC Cars Motors IFA (including Trabant, Wartburg, Barkas) Motors Mer 5805617
9 2017 Acura Motors Acura Motors BMW 9989723
10 2017 Acura Motors Geo Motors Vol 4727170
11 2017 Acura Motors Acura Motors BMW 5358057
12 2017 Acura Motors Herald Motors Hon 5383477
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Features
Models
Serving

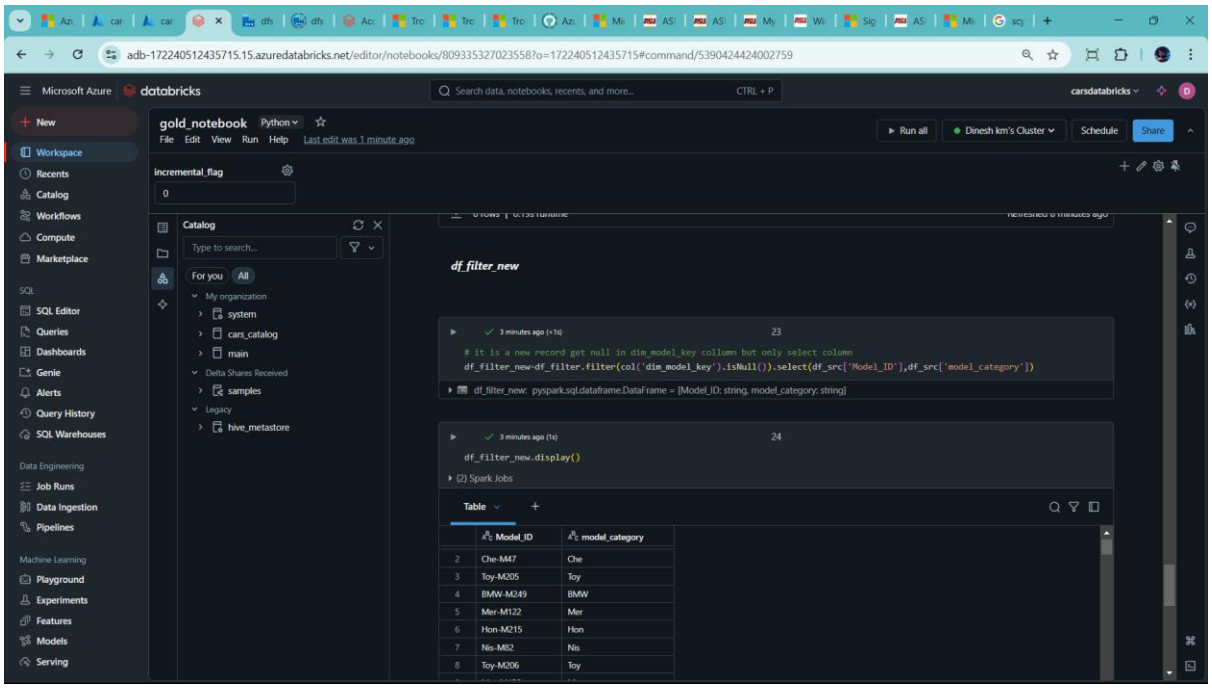

adb-1722405512435715.15.azuredatabricks.net/editor/notebooks/8093353270235587o=1722405124357l5
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
gold_notebook
Python
File Edit View Run Help
Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
This result is stored as _sqldf and can be used in other Python and SQL cells.
Just now (1s)
7
Python
df_src=spark.sql('''
SELECT DISTINCT(Model_ID) as Model_ID,model_category FROM parquet.`abfss://silver@cardineshdatalakestorage.dfs.core.windows.net/carsales`
''')
(1) Spark Jobs
df_src: pyspark.sql.dataframe.DataFrame = [Model_ID: string, model_category: string]
Just now (1s)
8
df_src.display()
(2) Spark Jobs
Table
Model_ID
model_category
9 Mar-M139 Mar
10 Ren-M207 Ren
11 Mar-M146 Mar
12 Mar-M141 Mar
13 Lin-M28 Lin
14 Hon-M69 Hon
15 Toy-M202 Toy
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Features
Models
Serving

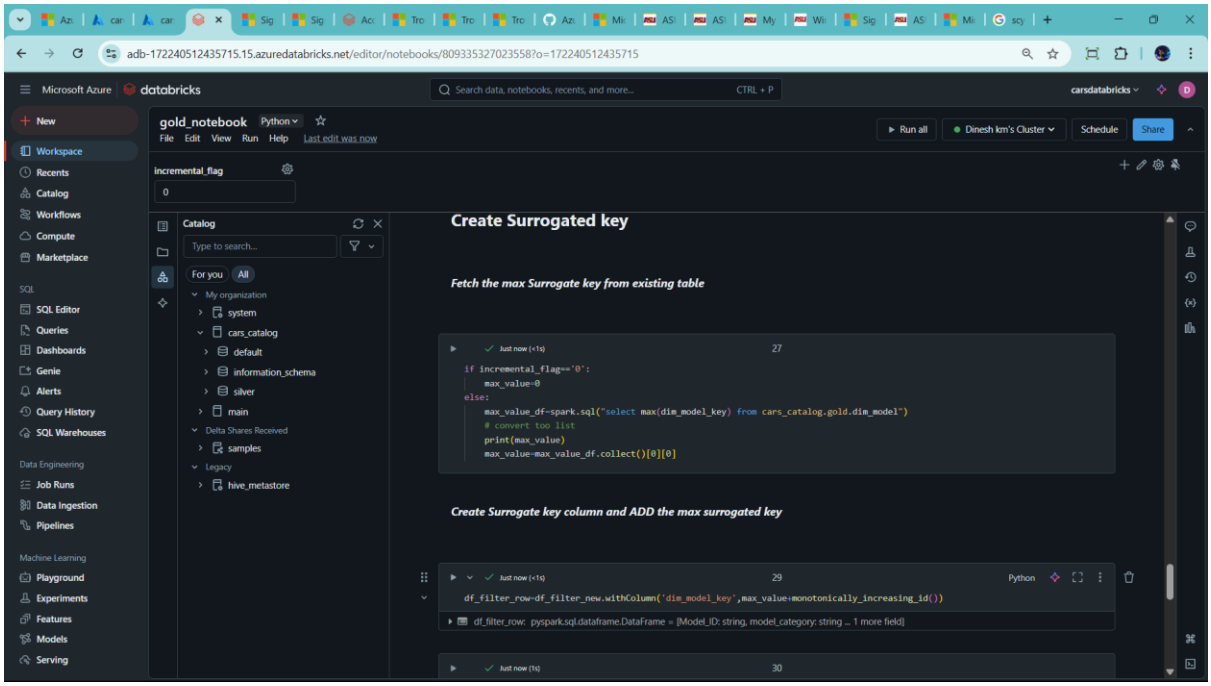

adb-1722405512435715.15.azuredatabricks.net/editor/notebooks/8093353270235587o=1722405124357l5#command/5390424424002759
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
gold_notebook
Python
File Edit View Run Help
Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
05:52 PM (1s)
11
#add surrocated key for first load
df_sink=spark.sql('''
SELECT 1 as dim_model_key,Model_ID,model_category FROM parquet.`abfss://silver@cardineshdatalakestorage.dfs.core.windows.net/carsales`
where 1=0
''')
(1) Spark Jobs
df_sink: pyspark.sql.dataframe.DataFrame = [dim_model_key: integer, Model_ID: string ... 1 more field]
06:00 PM (<1s)
12
Python
df_sink.display()
Table
dim_model_key
Model_ID
model_category
No rows returned
0 rows | 0.07s runtime
Refreshed 20 minutes ago
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Features
Models
Serving

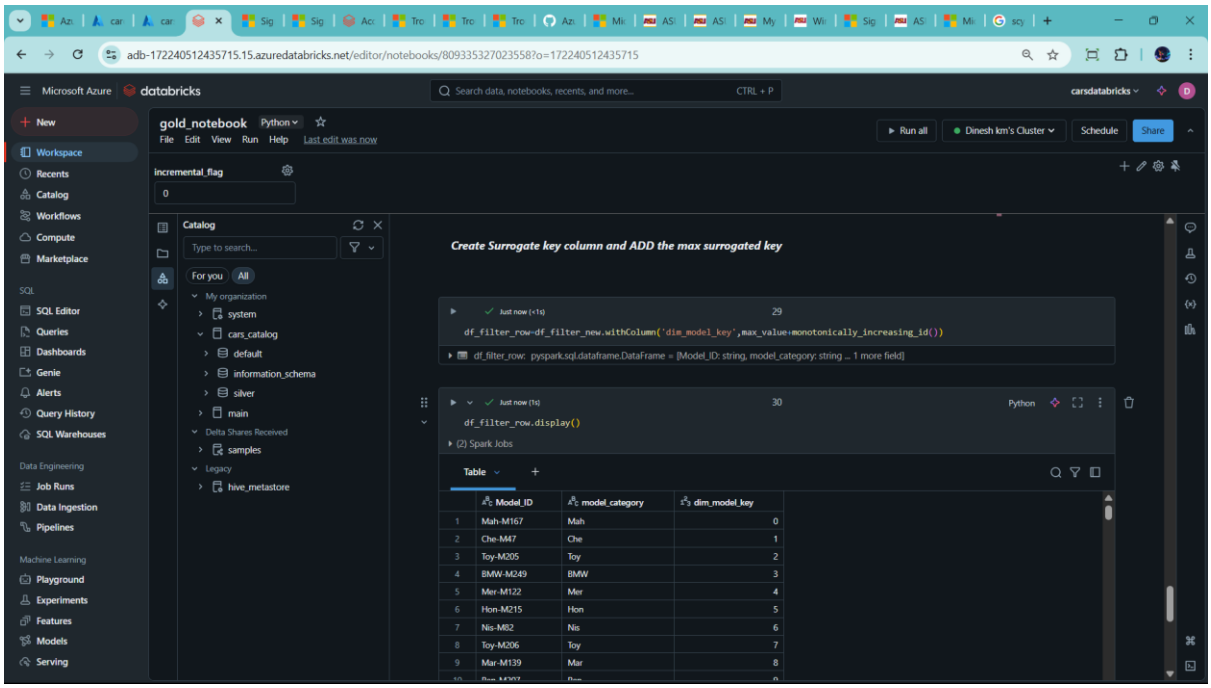

adb-1722405512435715.15.azuredatabricks.net/editor/notebooks/8093353270235587o=1722405124357157#command/5390424424002759
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
gold_notebook
Python
File Edit View Run Help Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Features
Models
Serving
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
dim_model sink -- Inital and Incremental
06:00 PM (2s)
14
if spark.catalog.tableExists('cars_catalog.gold.dim_model'):
df_sink=spark.sql('''
SELECT dim_model_key,Model_ID,model_category FROM parquet.`abfss://silver@cardineshdatalakestorage.dfs.core.windows.net/carsales`
''')
else:
df_sink=spark.sql('''
SELECT 1 as dim_model_key,Model_ID,model_category FROM parquet.`abfss://silver@cardineshdatalakestorage.dfs.core.windows.net/carsales`
where 1=0
''')
(1) Spark Jobs
df_sink: pyspark.sql.dataframe.DataFrame = [dim_model_key: integer, Model_ID: string ... 1 more field]
06:00 PM (<1s)
15
df_sink.display()
Table


adb-1722405512435715.15.azuredatabricks.net/editor/notebooks/8093353270235587o=1722405124357157#command/5390424424002759
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
gold_notebook
Python
File Edit View Run Help Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
Filterning New records and old Records
06:05 PM (<1s)
17
df_filter=df_src.join(df_sink,df_src['Model_ID']==df_sink['Model_ID'],how='left').select(df_src['Model_ID'],
df_src['model_category'],df_sink['dim_model_key'])
df_filter: pyspark.sql.dataframe.DataFrame = [Model_ID: string, model_category: string ... 1 more field]
06:06 PM (1s)
18
Python
df_filter.display()
(2) Spark Jobs
Table
Model_ID
model_category
dim_model_key
24 Mar-M151 Mar null
25 Acu-M59 Acu null
26 Hon-M70 Hon null
27 Mah-M174 Mah null
28 Dod-M9 Dod null
29 Tes-M96 Tes null
30 Tat-M194 Tat null
31 For-M23 For null
32 BMW-M3 BMW null

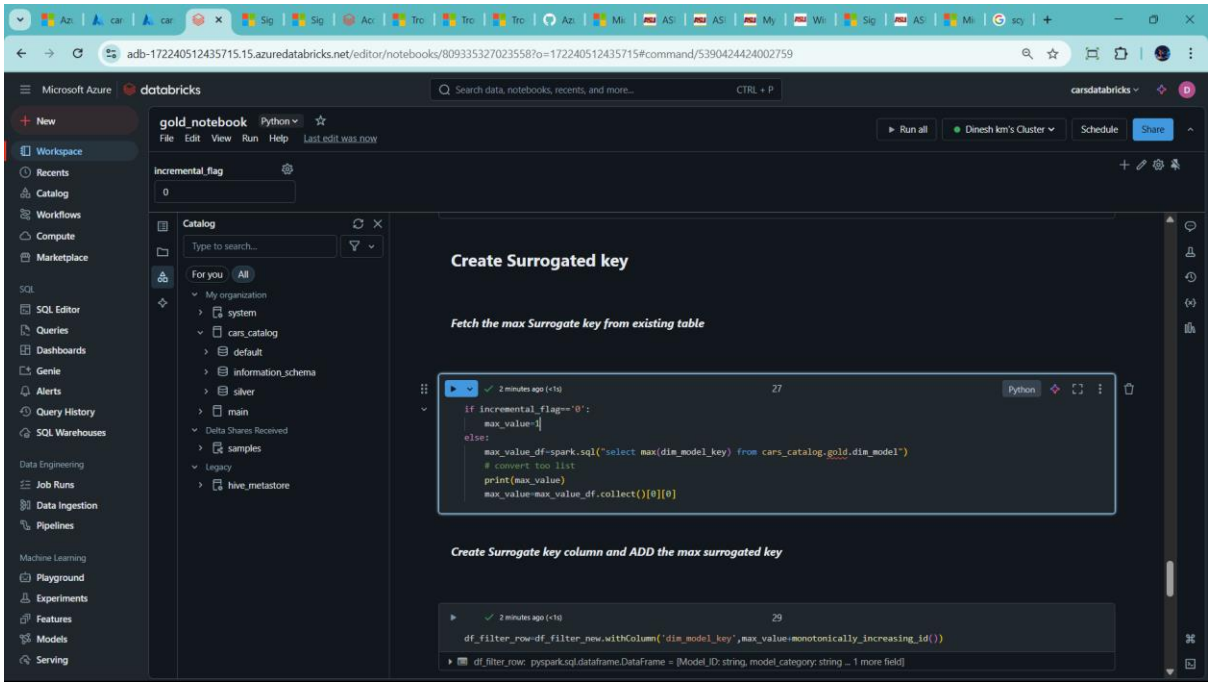

adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/809335327023558?o=1722405124357 15#command/5390424424002759
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
gold_notebook
Python
File Edit View Run Help
Last edit was 1 minute ago
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
df_filter_old
20
df_filter_old=df_filter.filter(col('dim_model_key').isNotNull())
df_filter_old: pyspark.sql.dataframe.DataFrame = [Model_ID: string, model_category: string ... 1 more field]
21
df_filter_old.display()
Table
Model_ID
model_category
dim_model_key
No rows returned
0 rows | 0.19s runtime
Refreshed 8 minutes ago
df_filter_new
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Features
Models
Serving

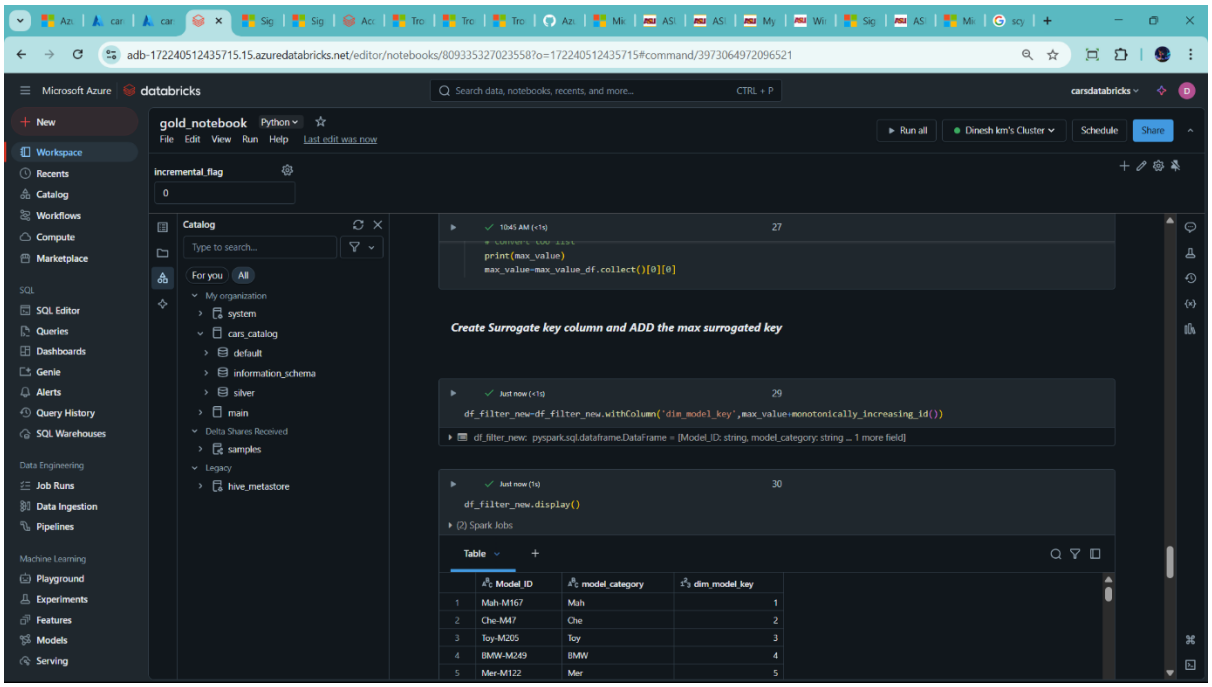

adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/809335327023558?o=1722405124357 15#command/5390424424002759
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
gold_notebook
Python
File Edit View Run Help
Last edit was 1 minute ago
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
df_filter_new
3 minutes ago (<1s)
23
# it is a new record get null in dim_model_key collumn but only select column
df_filter_new=df_filter.filter(col('dim_model_key').isNull()).select(df_src['Model_ID'],df_src['model_category'])
df_filter_new: pyspark.sql.dataframe.DataFrame = [Model_ID: string, model_category: string]
3 minutes ago (1s)
24
df_filter_new.display()
(2) Spark Jobs
Table
Model_ID
model_category
2 Che-M47 Che
3 Toy-M205 Toy
4 BMW-M249 BMW
5 Mer-M122 Mer
6 Hon-M215 Hon
7 Nis-M82 Nis
8 Toy-M206 Toy
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Features
Models
Serving

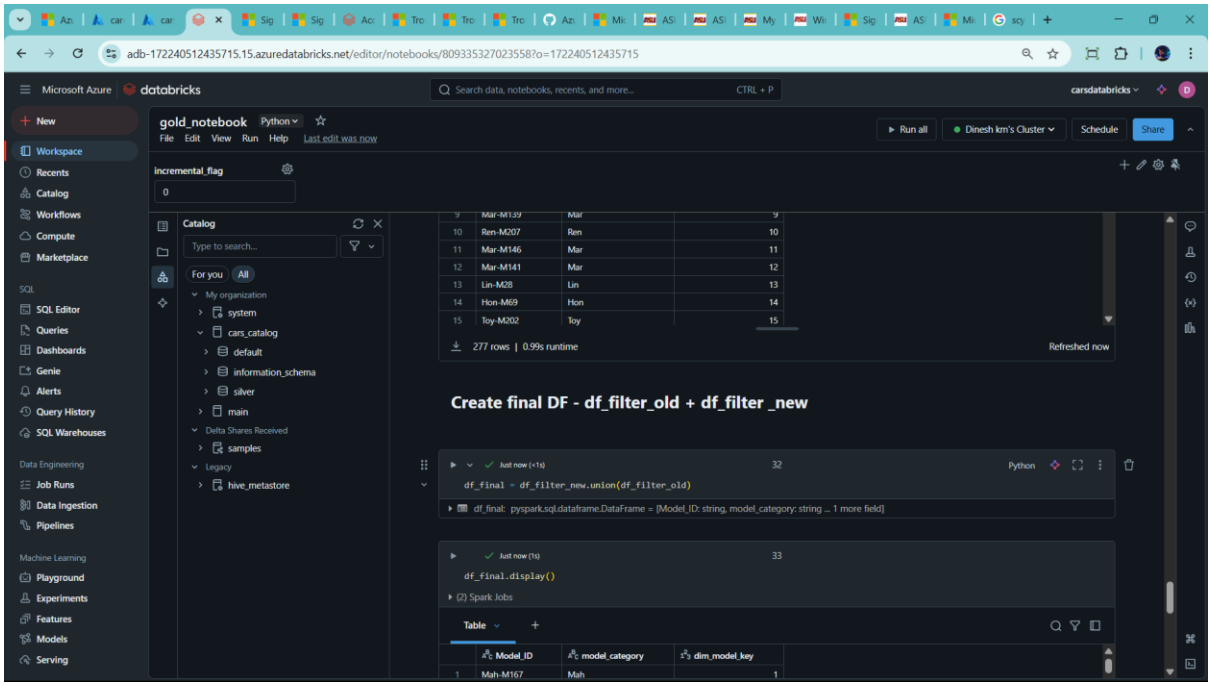


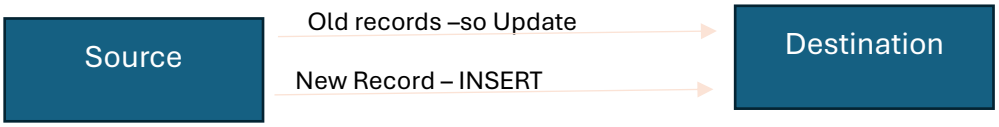


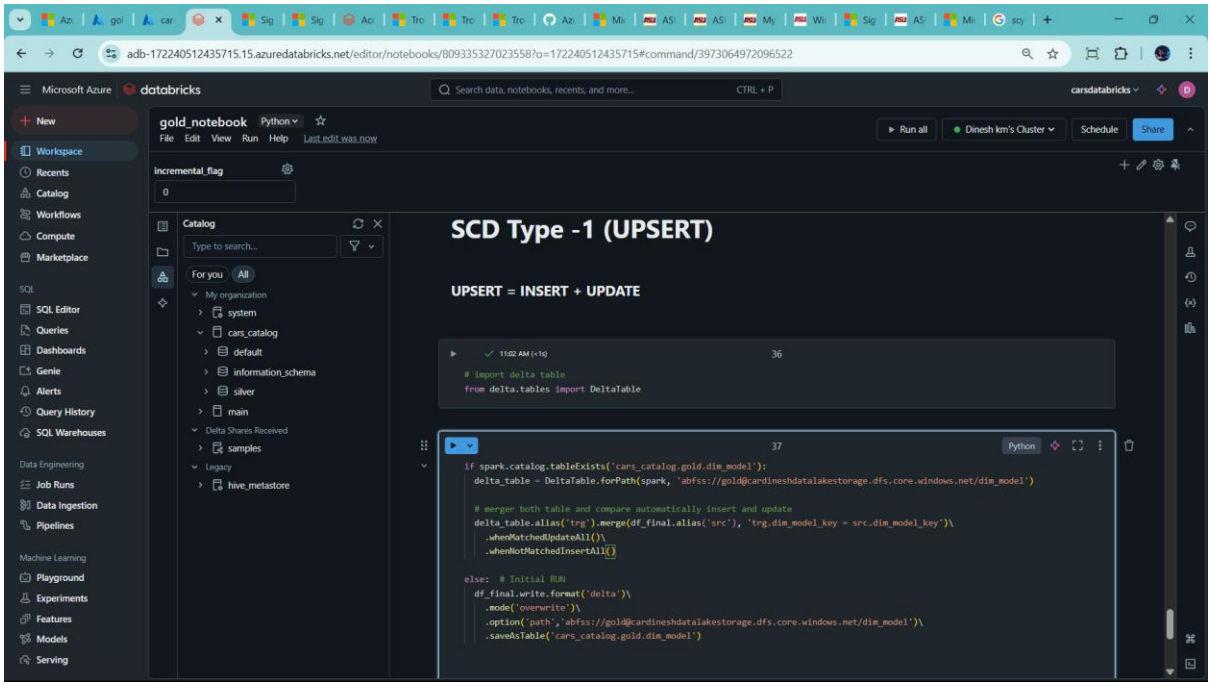
--change to max_value= 1 because surrogate key start 0

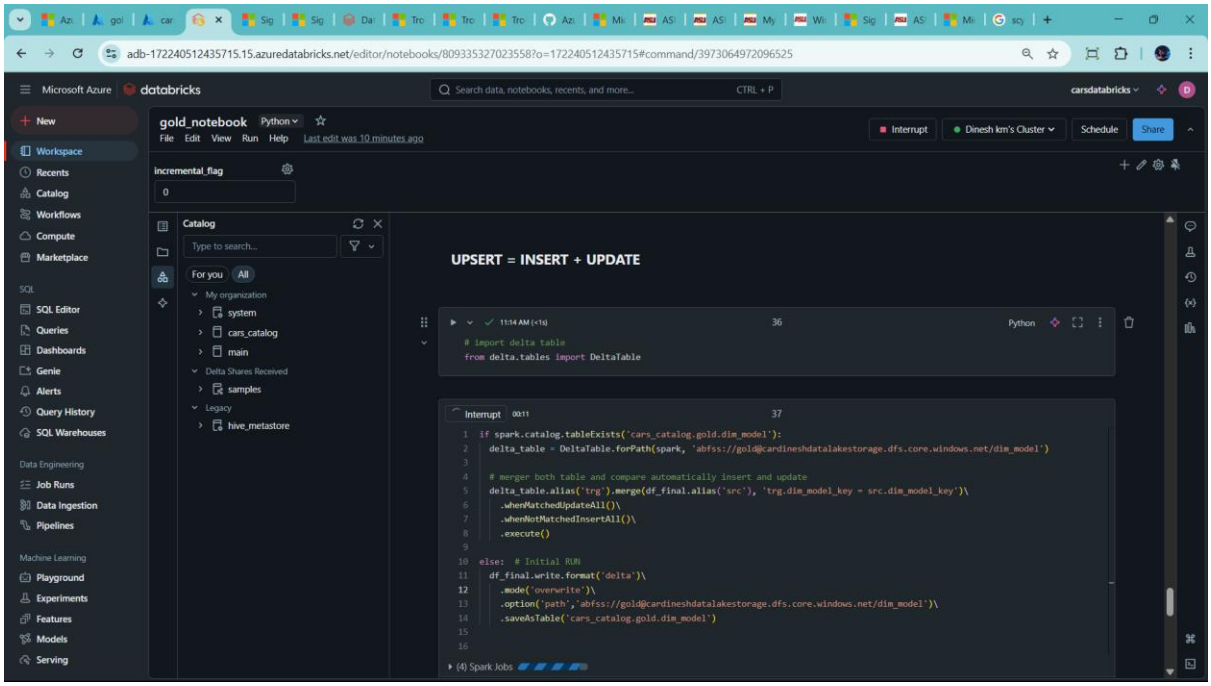

adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/8093353270235587o=1722405124357 15#command/5390424424002759
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
gold_notebook
Python
File Edit View Run Help
Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
default
information_schema
silver
main
Delta Shares Received
samples
Legacy
hive_metastore
Create Surrogated key
Fetch the max Surrogate key from existing table
2 minutes ago (<1s)
27
Python
if incremental_flag='0':
max_value=1
else:
max_value_df=spark.sql("select max(dim_model_key) from cars_catalog.gold.dim_model")
# convert too list
print(max_value)
max_value=max_value_df.collect()[0][0]
Create Surrogate key column and ADD the max surrogated key
2 minutes ago (<1s)
29
df_filter_row=df_filter_new.withColumn('dim_model_key',max_value+monotonically_increasing_id())
df_filter_row: pyspark.sql.dataframe.DataFrame = [Model_ID: string, model_category: string ... 1 more field]
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Features
Models
Serving
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
New


adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/8093353270235587o=1722405124357 15#command/3973064972096521
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
gold_notebook
Python
File Edit View Run Help
Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
default
information_schema
silver
main
Delta Shares Received
samples
Legacy
hive_metastore
10:45 AM (<1s)
27
print(max_value)
max_value=max_value_df.collect()[0][0]
Create Surrogate key column and ADD the max surrogated key
Just now (<1s)
29
df_filter_new=df_filter_new.withColumn('dim_model_key',max_value+monotonically_increasing_id())
df_filter_new: pyspark.sql.dataframe.DataFrame = [Model_ID: string, model_category: string ... 1 more field]
Just now (1s)
30
df_filter_new.display()
(2) Spark Jobs
Table
Model_ID
model_category
dim_model_key
1 Mah-M167 Mah 1
2 Che-M47 Che 2
3 Toy-M205 Toy 3
4 BMW-M249 BMW 4
5 Mer-M122 Mer 5
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Features
Models
Serving
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
New

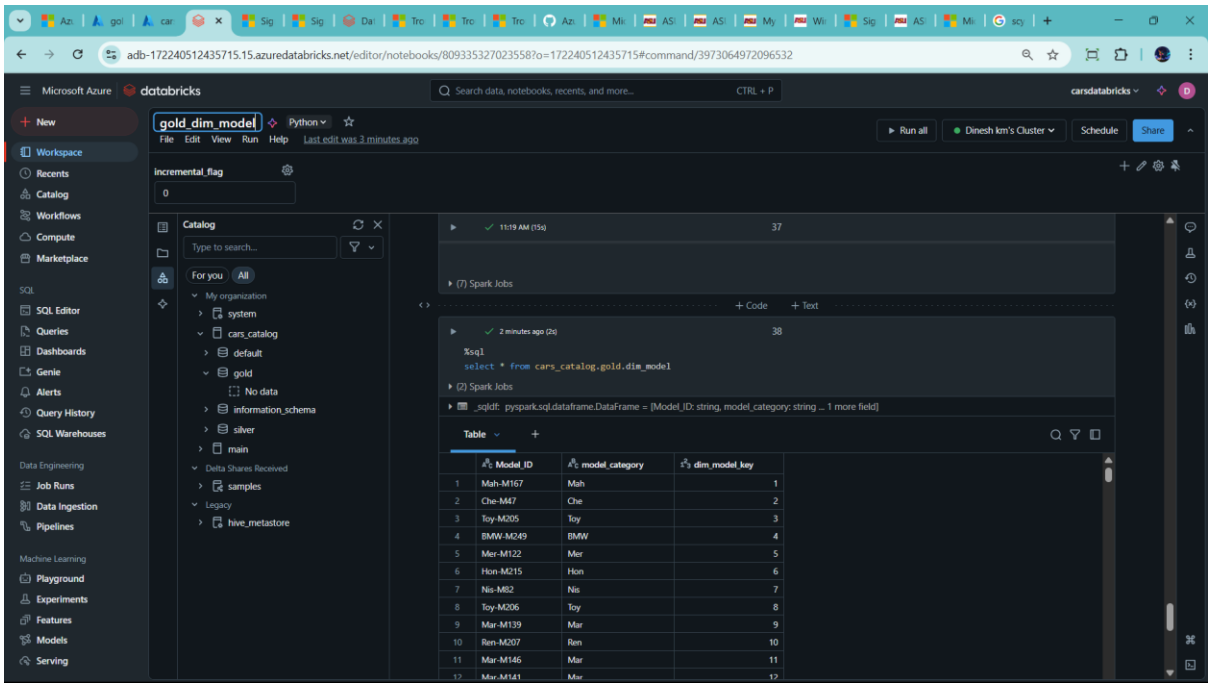

adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/809335327023558?o=1722405124357 15
Microsoft Azure
databricks
gold_notebook
Python
File Edit View Run Help Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you All
My organization
system
cars_catalog
default
information_schema
silver
main
Delta Shares Received
samples
Legacy
hive_metastore
277 rows | 0.99s runtime
Refreshed now
Create final DF - df_filter_old + df_filter _new
df_final = df_filter_new.union(df_filter_old)
df_final.display()
(2) Spark Jobs
Table
Model_ID
model_category
dim_model_key
Mah-M167
Mah
1


Source
Old records –so Update
New Record – INSERT
Destination

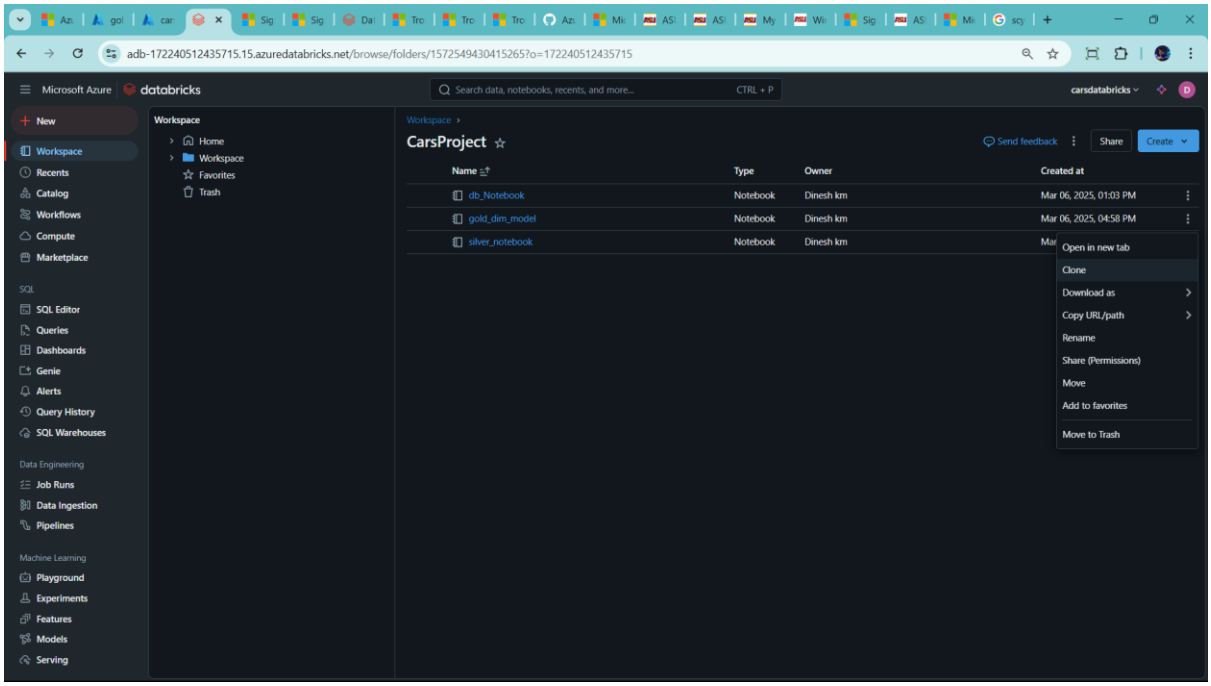

adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/809335327023558?o=1722405124357 15#command/3973064972096522
Microsoft Azure
databricks
gold_notebook
Python
File Edit View Run Help Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you All
My organization
system
cars_catalog
default
information_schema
silver
main
Delta Shares Received
samples
Legacy
hive_metastore
SCD Type -1 (UPSERT)
UPSERT = INSERT + UPDATE
# import delta table
from delta.tables import DeltaTable
if spark.catalog.tableExists('cars_catalog.gold.dim_model'):
delta_table = DeltaTable.forPath(spark, 'abfss://gold@cardineshdatalakestorage.dfs.core.windows.net/dim_model')
# merger both table and compare automatically insert and update
delta_table.alias('trg').merge(df_final.alias('src'), 'trg.dim_model_key = src.dim_model_key')\
.whenMatchedUpdateAll()\
.whenNotMatchedInsertAll()
else: # Initial RUN
df_final.write.format('delta')\
.mode('overwrite')\
.option('path','abfss://gold@cardineshdatalakestorage.dfs.core.windows.net/dim_model')\
.saveAsTable('cars_catalog.gold.dim_model')



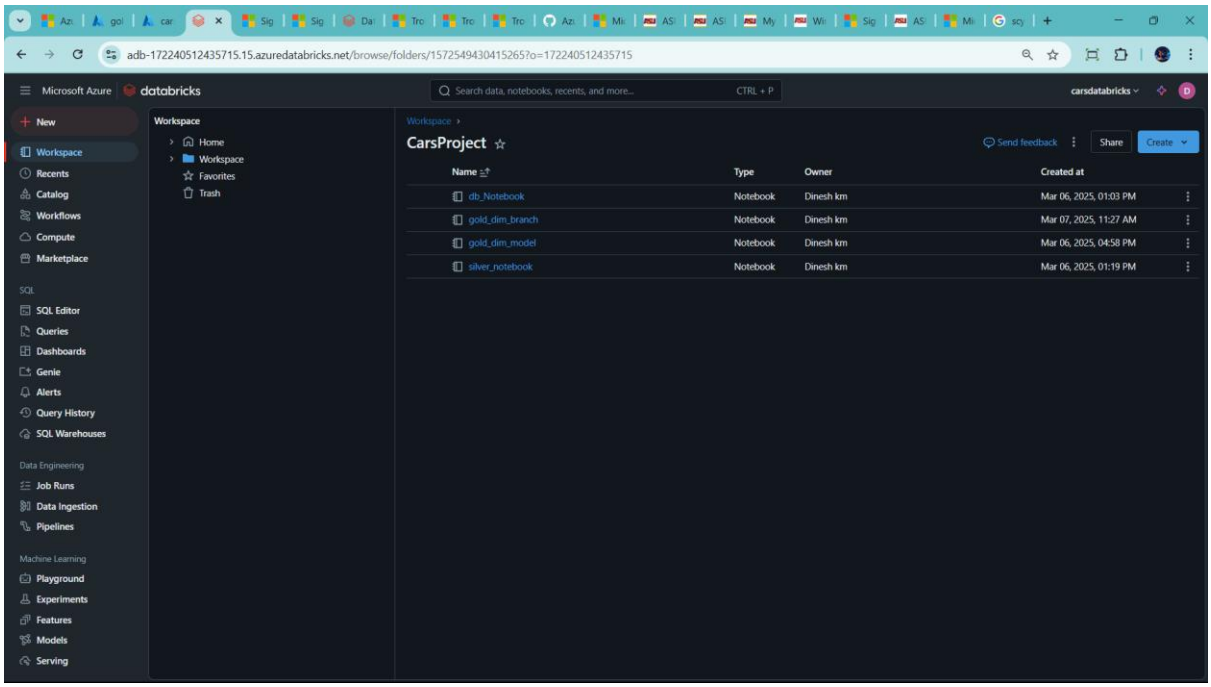
--change name gold_dim_model



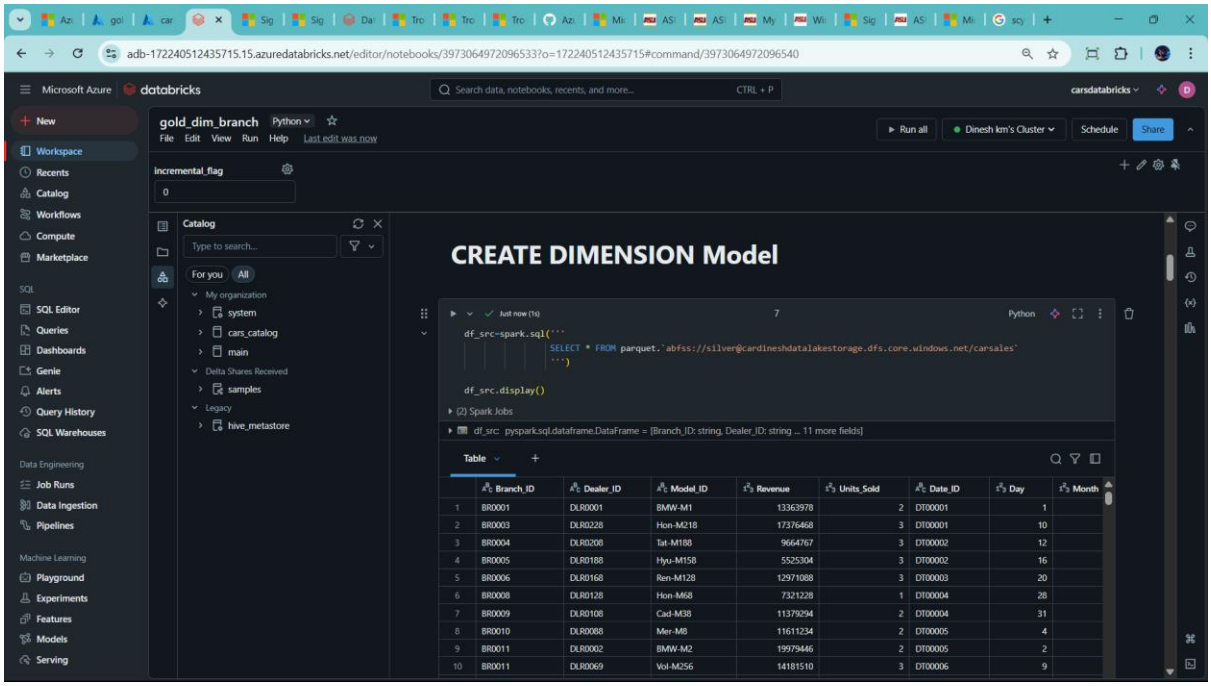
-→ copy this gold_dim_model to gold_dim_branch in clone method

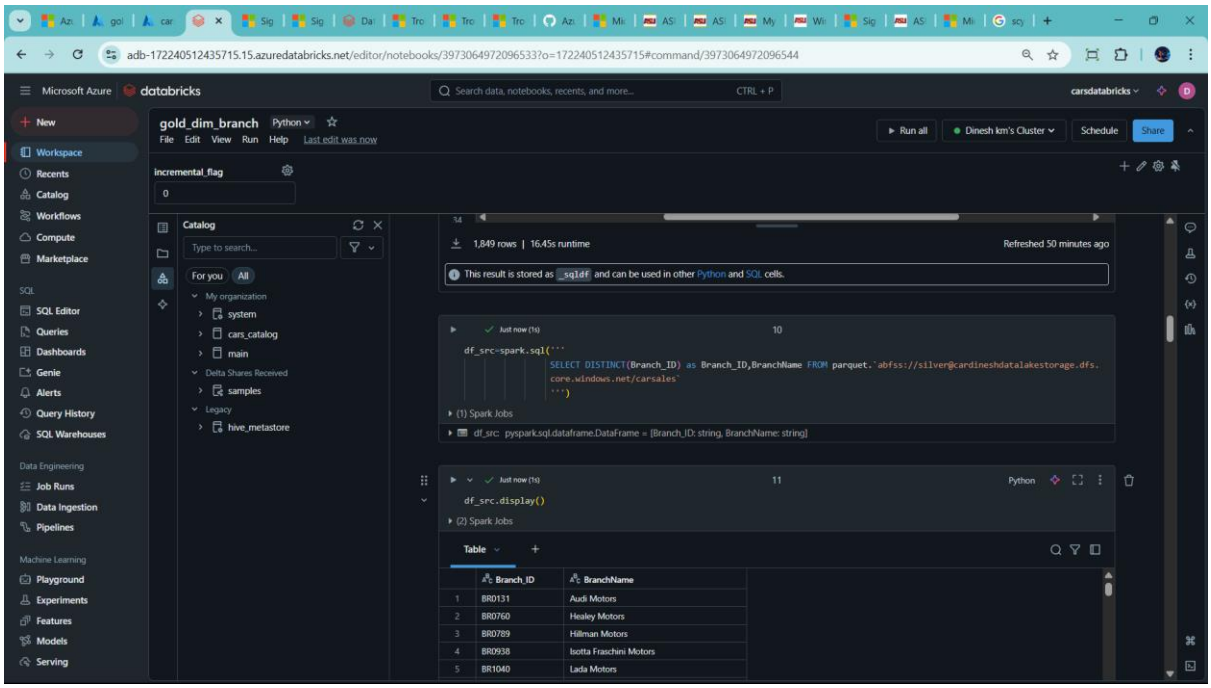


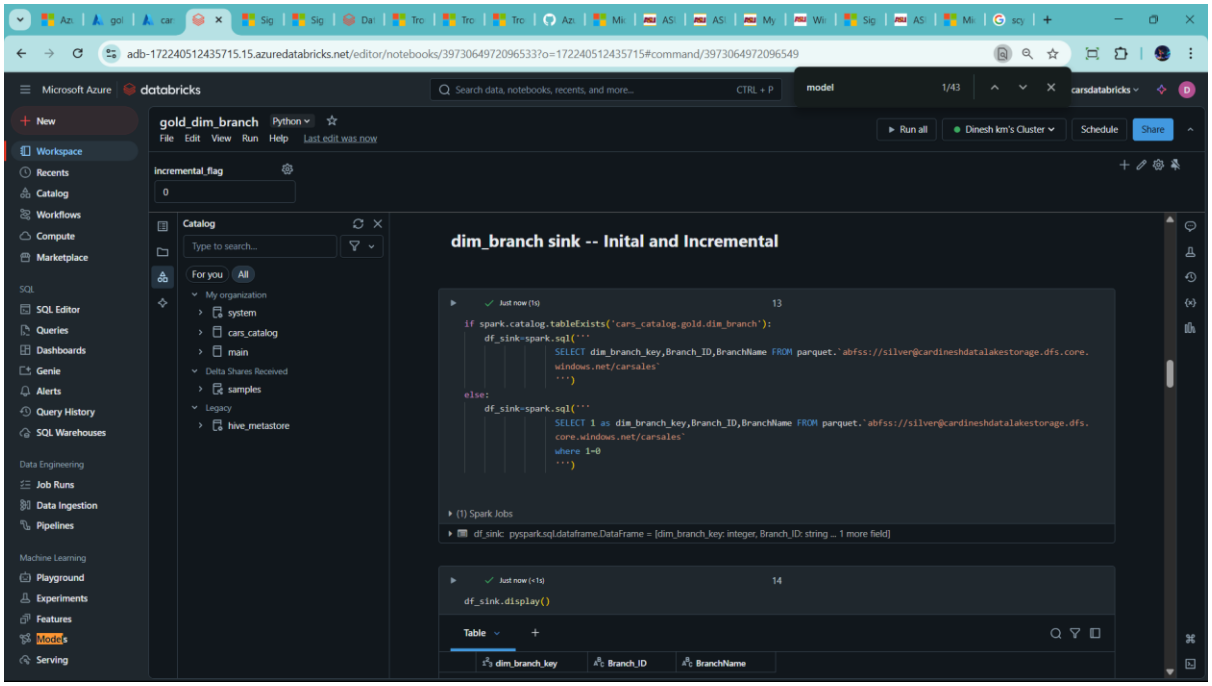
→

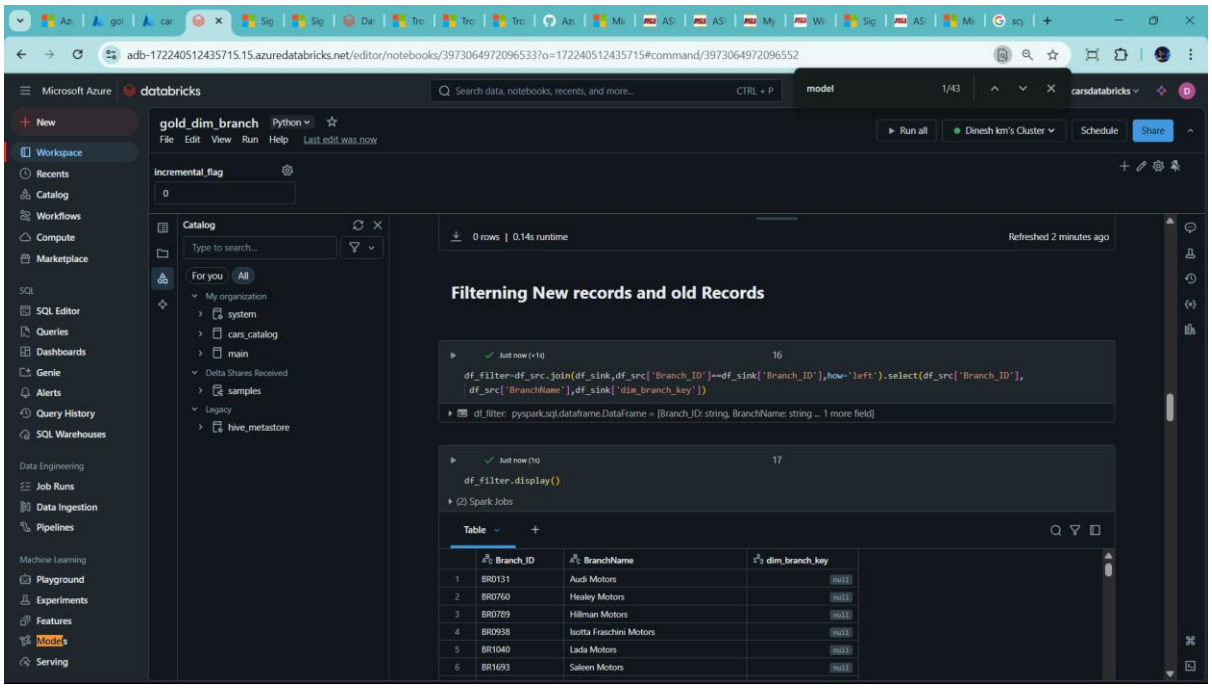


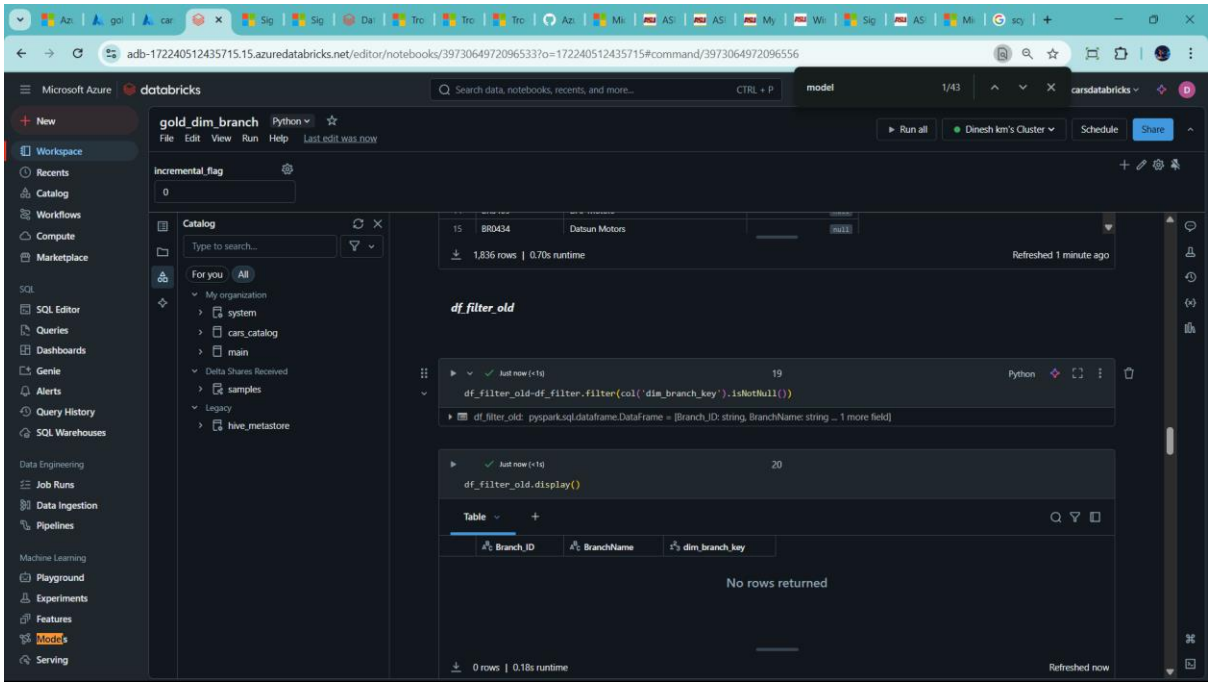
→change accouding

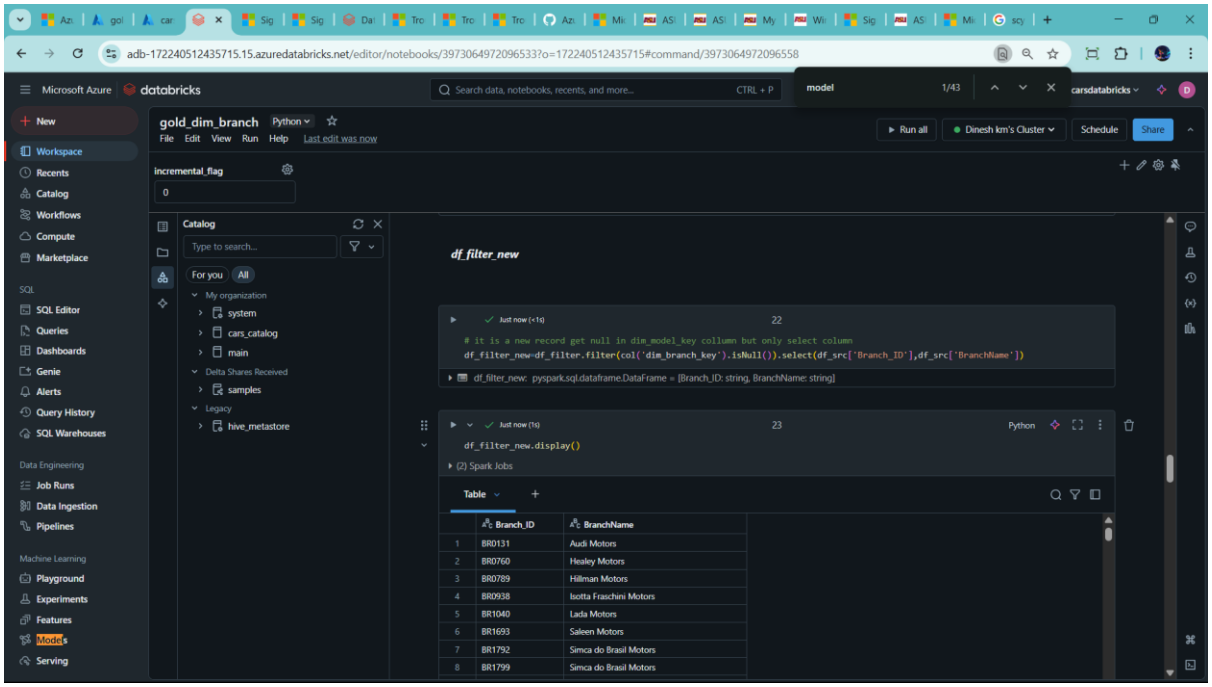

adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/3973064972096533?o=1722405124357 15#command/3973064972096540
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
gold_dim_branch
Python
File Edit View Run Help Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
CREATE DIMENSION Model
Just now (1s)
7
Python
df_src=spark.sql('''
SELECT * FROM parquet.`abfss://silver@cardineshdatalakestorage.dfs.core.windows.net/carsales`
''')
df_src.display()
(2) Spark Jobs
df_src: pyspark.sql.dataframe.DataFrame = [Branch_ID: string, Dealer_ID: string ... 11 more fields]
Table
Branch_ID Dealer_ID Model_ID Revenue Units_Sold Date_ID Day Month
1 BR0001 DLR0001 BMW-M1 13363978 2 DT00001 1
2 BR0003 DLR0228 Hon-M218 17376468 3 DT00001 10
3 BR0004 DLR0208 Tat-M188 9664767 3 DT00002 12
4 BR0005 DLR0188 Hyu-M158 5525304 3 DT00002 16
5 BR0006 DLR0168 Ren-M128 12971088 3 DT00003 20
6 BR0008 DLR0128 Hon-M68 7321228 1 DT00004 28
7 BR0009 DLR0108 Cad-M38 11379294 2 DT00004 31
8 BR0010 DLR0088 Mer-M8 11611234 2 DT00005 4
9 BR0011 DLR0002 BMW-M2 19979446 2 DT00005 2
10 BR0011 DLR0069 Vol-M256 14181510 3 DT00006 9

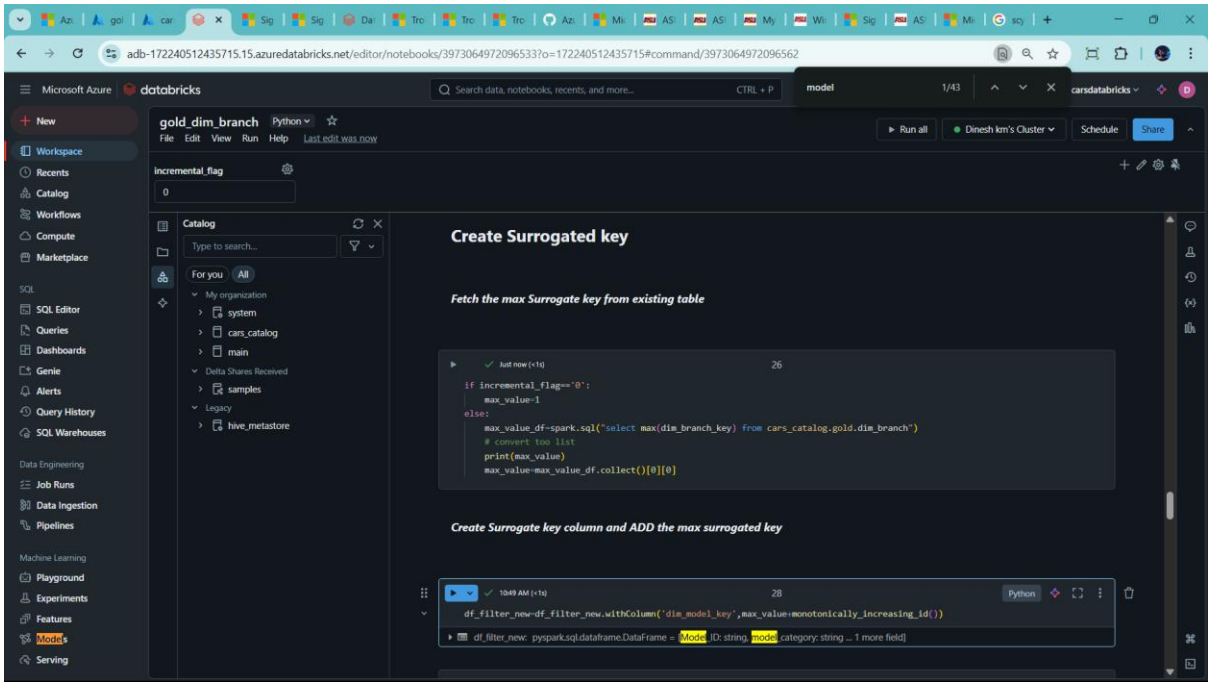

adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/3973064972096533?o=1722405124357 15#command/3973064972096544
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
gold_dim_branch
Python
File Edit View Run Help Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
1,849 rows | 16.45s runtime
Refreshed 50 minutes ago
This result is stored as _sqldf and can be used in other Python and SQL cells.
Just now (1s)
10
df_src=spark.sql('''
SELECT DISTINCT(Branch_ID) as Branch_ID,BranchName FROM parquet.`abfss://silver@cardineshdatalakestorage.dfs.core.windows.net/carsales`
''')
(1) Spark Jobs
df_src: pyspark.sql.dataframe.DataFrame = [Branch_ID: string, BranchName: string]
Just now (1s)
11
Python
df_src.display()
(2) Spark Jobs
Table
Branch_ID BranchName
1 BR0131 Audi Motors
2 BR0760 Healey Motors
3 BR0789 Hillman Motors
4 BR0938 Isotta Fraschini Motors
5 BR1040 Lada Motors

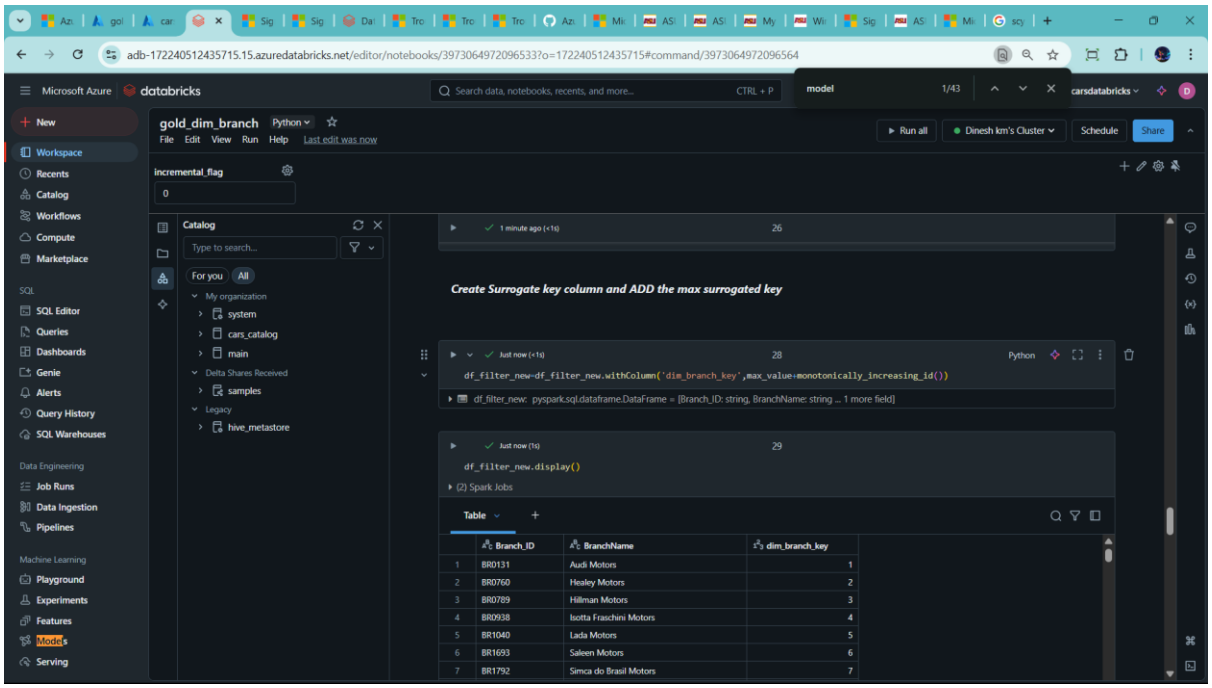

adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/3973064972096533?o=1722405124357 15#command/3973064972096549
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
model
1/43
carsdatabricks
gold_dim_branch
Python
File Edit View Run Help Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
dim_branch sink -- Inital and Incremental
13
Just now (1s)
if spark.catalog.tableExists('cars_catalog.gold.dim_branch'):
df_sink=spark.sql('''
SELECT dim_branch_key,Branch_ID,BranchName FROM parquet.`abfss://silver@cardineshdatalakestorage.dfs.core.windows.net/carsales`
''')
else:
df_sink=spark.sql('''
SELECT 1 as dim_branch_key,Branch_ID,BranchName FROM parquet.`abfss://silver@cardineshdatalakestorage.dfs.core.windows.net/carsales`
where 1=0
''')
(1) Spark Jobs
df_sink: pyspark.sql.dataframe.DataFrame = [dim_branch_key: integer, Branch_ID: string ... 1 more field]
14
Just now (<1s)
df_sink.display()
Table
dim_branch_key
Branch_ID
BranchName
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Features
Models
Serving

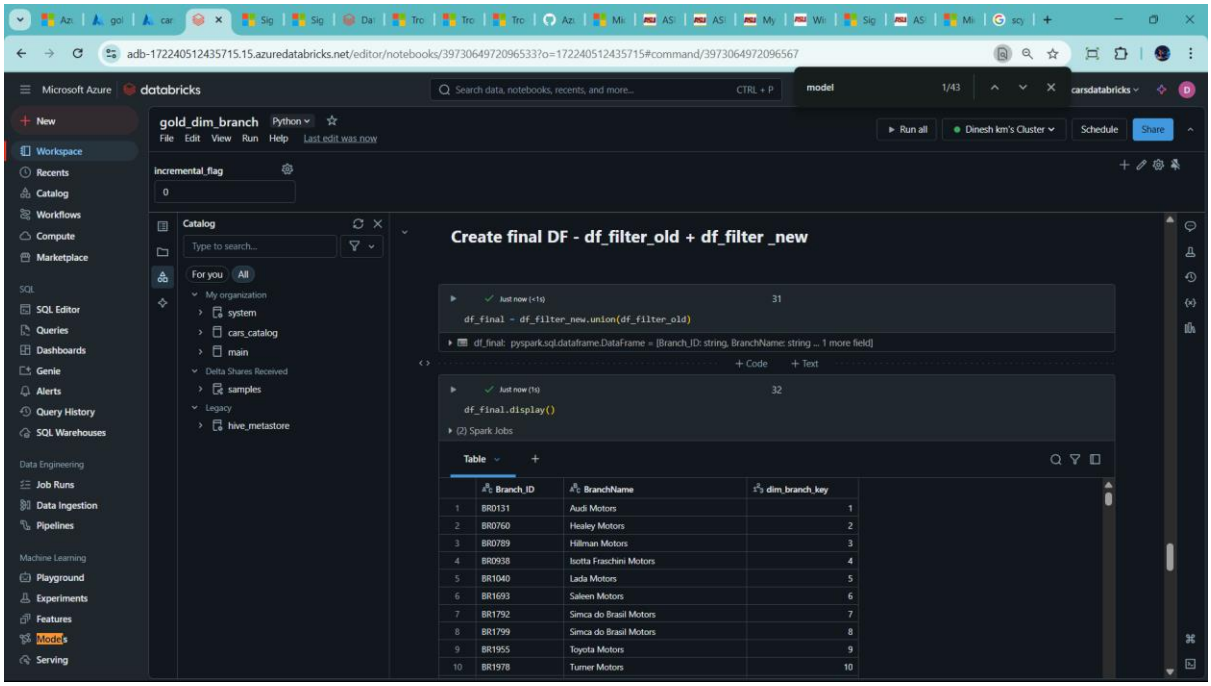

adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/3973064972096533?o=1722405124357 15#command/3973064972096552
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
model
1/43
carsdatabricks
gold_dim_branch
Python
File Edit View Run Help Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
0 rows | 0.14s runtime
Refreshed 2 minutes ago
Filterning New records and old Records
16
Just now (<1s)
df_filter=df_src.join(df_sink,df_src['Branch_ID']==df_sink['Branch_ID'],how='left').select(df_src['Branch_ID'],
df_src['BranchName'],df_sink['dim_branch_key'])
df_filter: pyspark.sql.dataframe.DataFrame = [Branch_ID: string, BranchName: string ... 1 more field]
17
Just now (1s)
df_filter.display()
(2) Spark Jobs
Table
Branch_ID
BranchName
dim_branch_key
1 BR0131 Audi Motors null
2 BR0760 Healey Motors null
3 BR0789 Hillman Motors null
4 BR0938 Isotta Fraschini Motors null
5 BR1040 Lada Motors null
6 BR1693 Saleen Motors null
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Features
Models
Serving

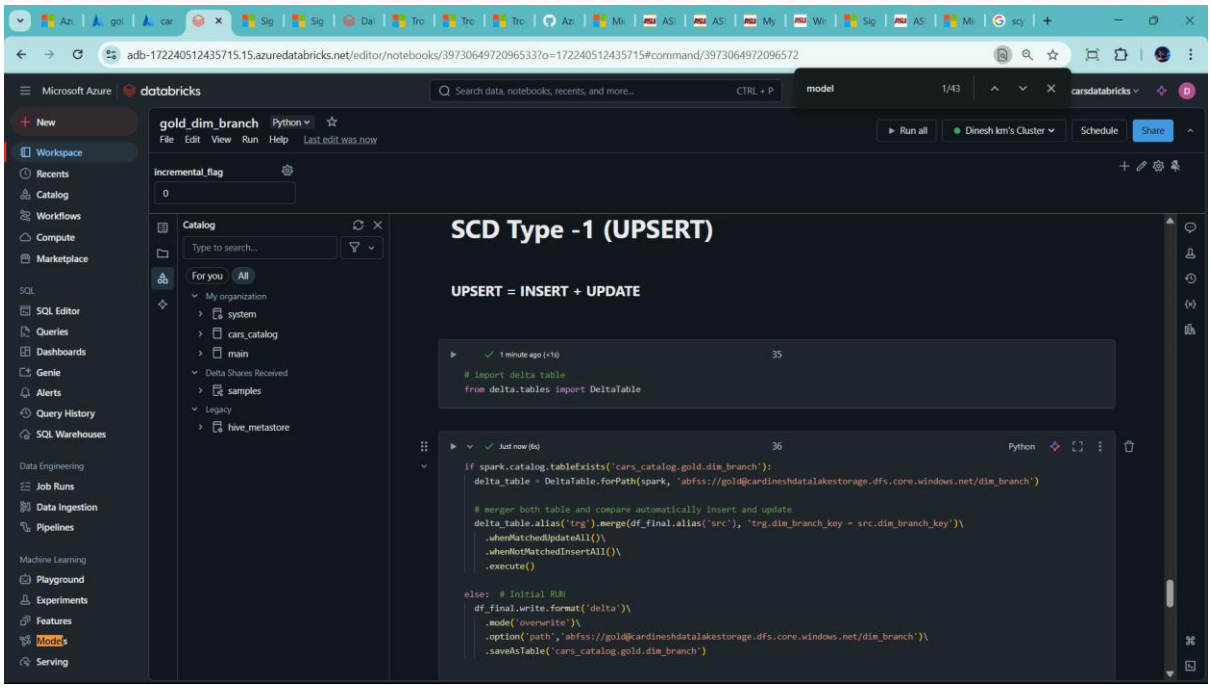

adb-1722405512435715.15.azuredatabricks.net/editor/notebooks/3973064972096533?o=1722405512435715#command/3973064972096556
Microsoft Azure
databricks
gold_dim_branch
Python
File Edit View Run Help Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
df_filter_old
15 BR0434 Datsun Motors null
1,836 rows | 0.70s runtime
Refreshed 1 minute ago
df_filter_old=df_filter.filter(col('dim_branch_key').isNotNull())
df_filter_old: pyspark.sql.dataframe.DataFrame = [Branch_ID: string, BranchName: string ... 1 more field]
df_filter_old.display()
Table
Branch_ID
BranchName
dim_branch_key
No rows returned
0 rows | 0.18s runtime
Refreshed now


adb-1722405512435715.15.azuredatabricks.net/editor/notebooks/3973064972096533?o=1722405512435715#command/3973064972096558
Microsoft Azure
databricks
gold_dim_branch
Python
File Edit View Run Help Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
df_filter_new
# it is a new record get null in dim_model_key collumn but only select column
df_filter_new=df_filter.filter(col('dim_branch_key').isNull()).select(df_src['Branch_ID'],df_src['BranchName'])
df_filter_new: pyspark.sql.dataframe.DataFrame = [Branch_ID: string, BranchName: string]
df_filter_new.display()
(2) Spark Jobs
Table
Branch_ID
BranchName
1 BR0131 Audi Motors
2 BR0760 Healey Motors
3 BR0789 Hillman Motors
4 BR0938 Isotta Fraschini Motors
5 BR1040 Lada Motors
6 BR1693 Saleen Motors
7 BR1792 Simca do Brasil Motors
8 BR1799 Simca do Brasil Motors


adb-1722405512435715.15.azuredatabricks.net/editor/notebooks/3973064972096533?o=1722405512435715#command/3973064972096562
Microsoft Azure
databricks
gold_dim_branch
Python
File Edit View Run Help Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Create Surrogated key
Fetch the max Surrogate key from existing table
if incremental_flag=='0':
max_value=1
else:
max_value_df=spark.sql("select max(dim_branch_key) from cars_catalog.gold.dim_branch")
# convert too list
print(max_value)
max_value=max_value_df.collect()[0][0]
Create Surrogate key column and ADD the max surrogated key
df_filter_new=df_filter_new.withColumn('dim_model_key',max_value+monotonically_increasing_id())
df_filter_new: pyspark.sql.dataframe.DataFrame = [Model_ID: string, model_category: string ... 1 more field]

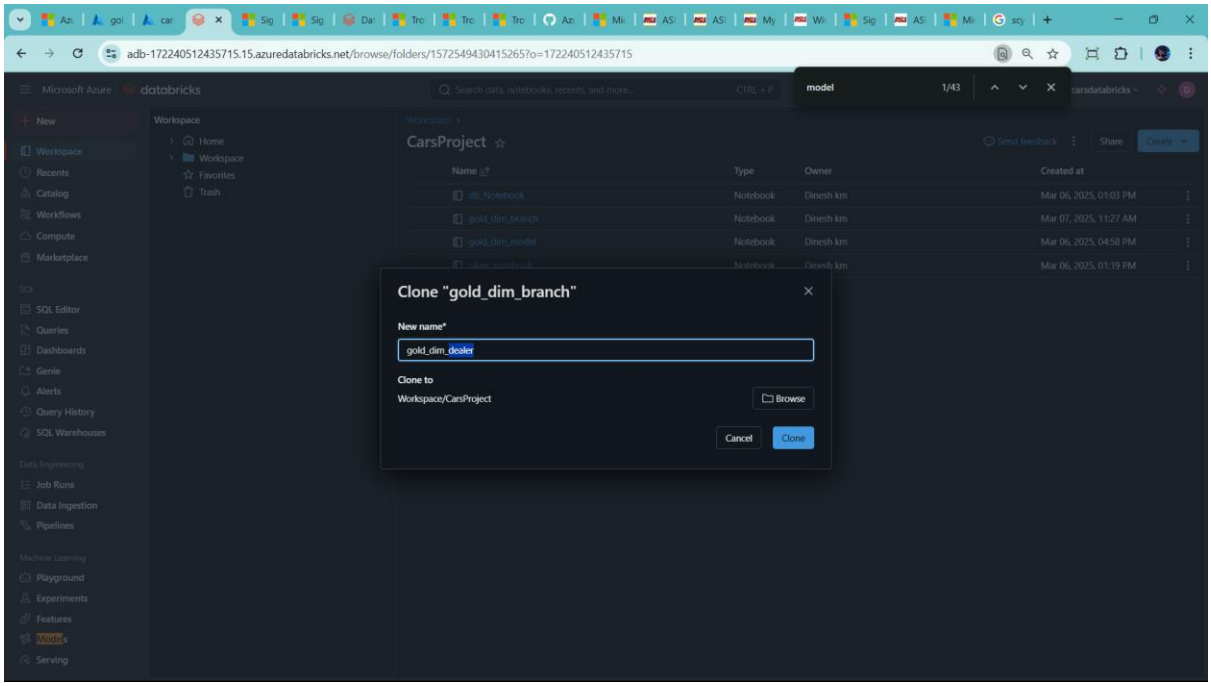

adb-1722405512435715.15.azuredatabricks.net/editor/notebooks/3973064972096533?o=1722405512435715#command/3973064972096564
Microsoft Azure
databricks
gold_dim_branch
Python
File Edit View Run Help Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
incremental_flag
0
Catalog
Create Surrogate key column and ADD the max surrogated key
df_filter_new=df_filter_new.withColumn('dim_branch_key',max_value+monotonically_increasing_id())
df_filter_new: pyspark.sql.dataframe.DataFrame = [Branch_ID: string, BranchName: string ... 1 more field]
df_filter_new.display()
(2) Spark Jobs
Table
Branch_ID
BranchName
dim_branch_key
BR0131 Audi Motors 1
BR0760 Healey Motors 2
BR0789 Hillman Motors 3
BR0938 Isotta Fraschini Motors 4
BR1040 Lada Motors 5
BR1693 Saleen Motors 6
BR1792 Simca do Brasil Motors 7

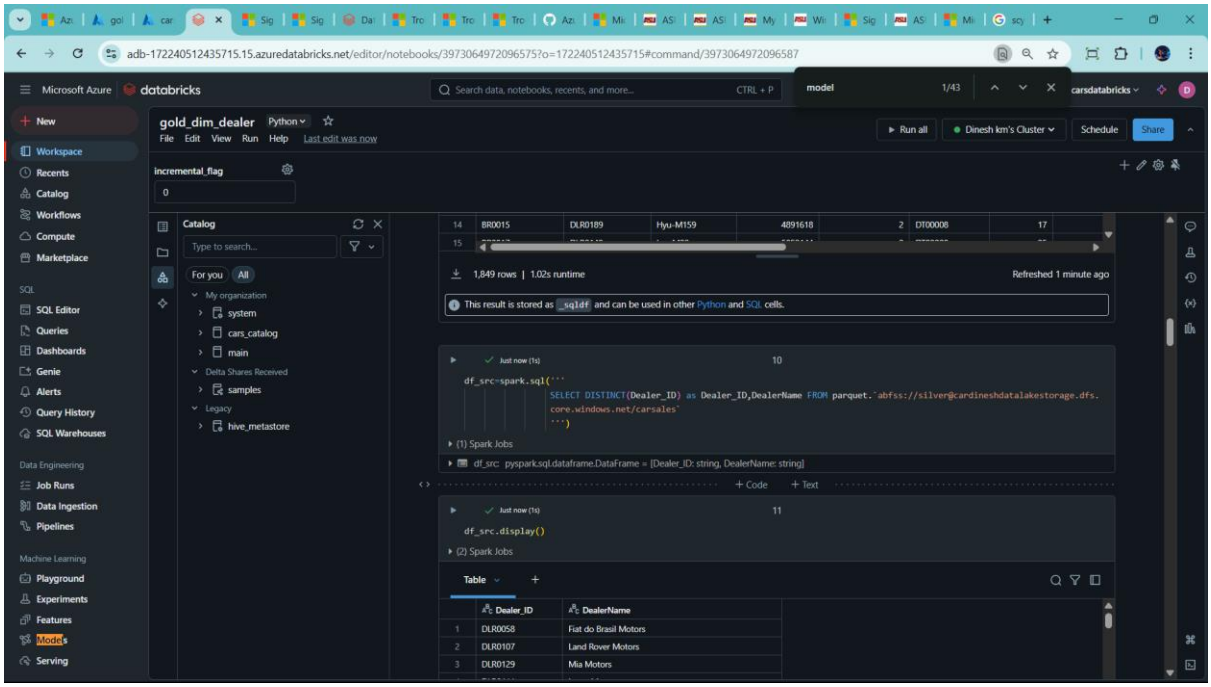




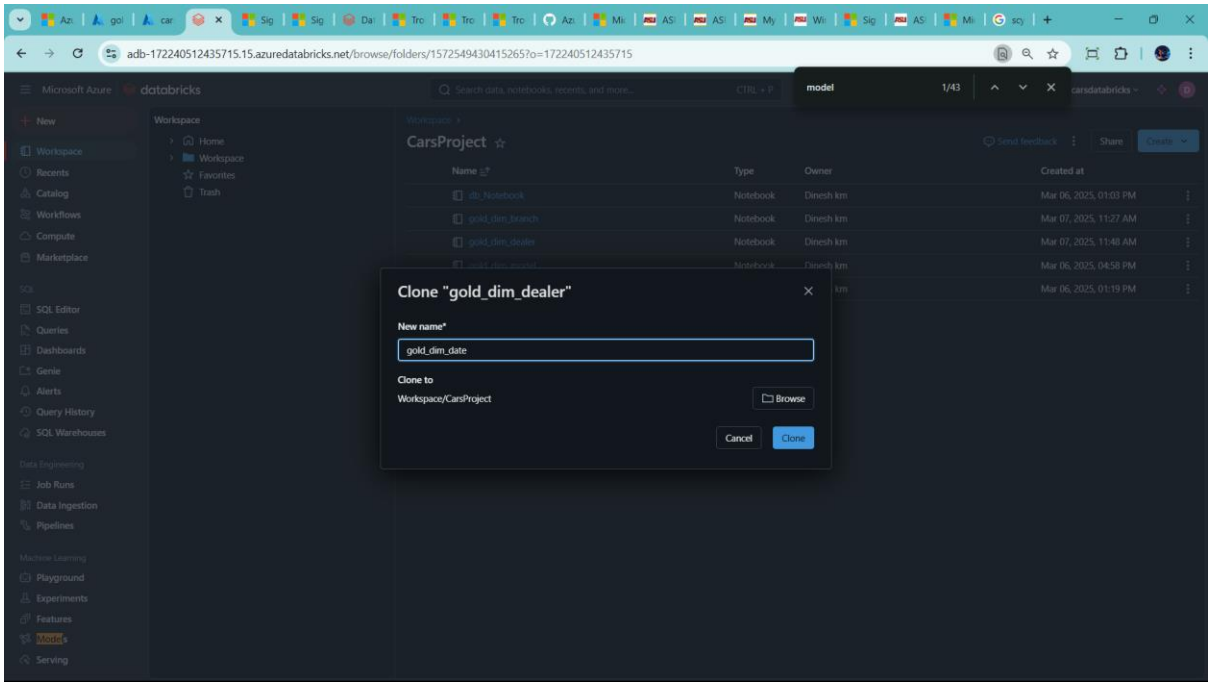
-→ Now Create Another Dimension table – dim_dealer

→copy the clone the another gold_dim_branch

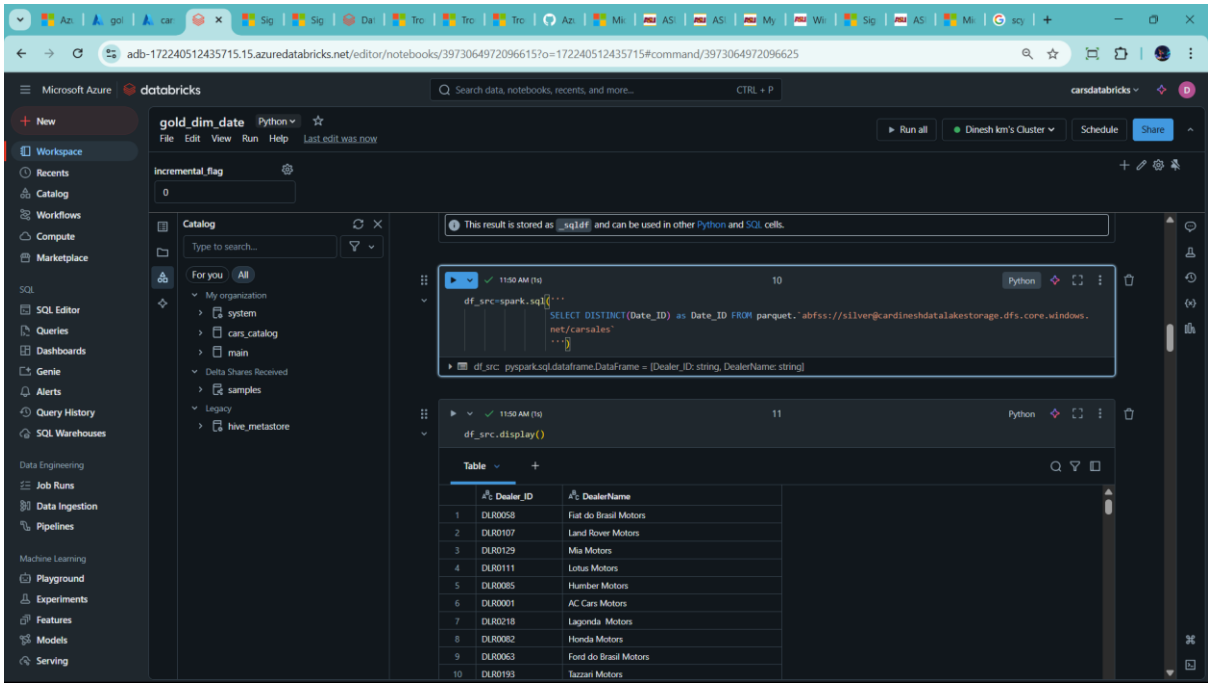


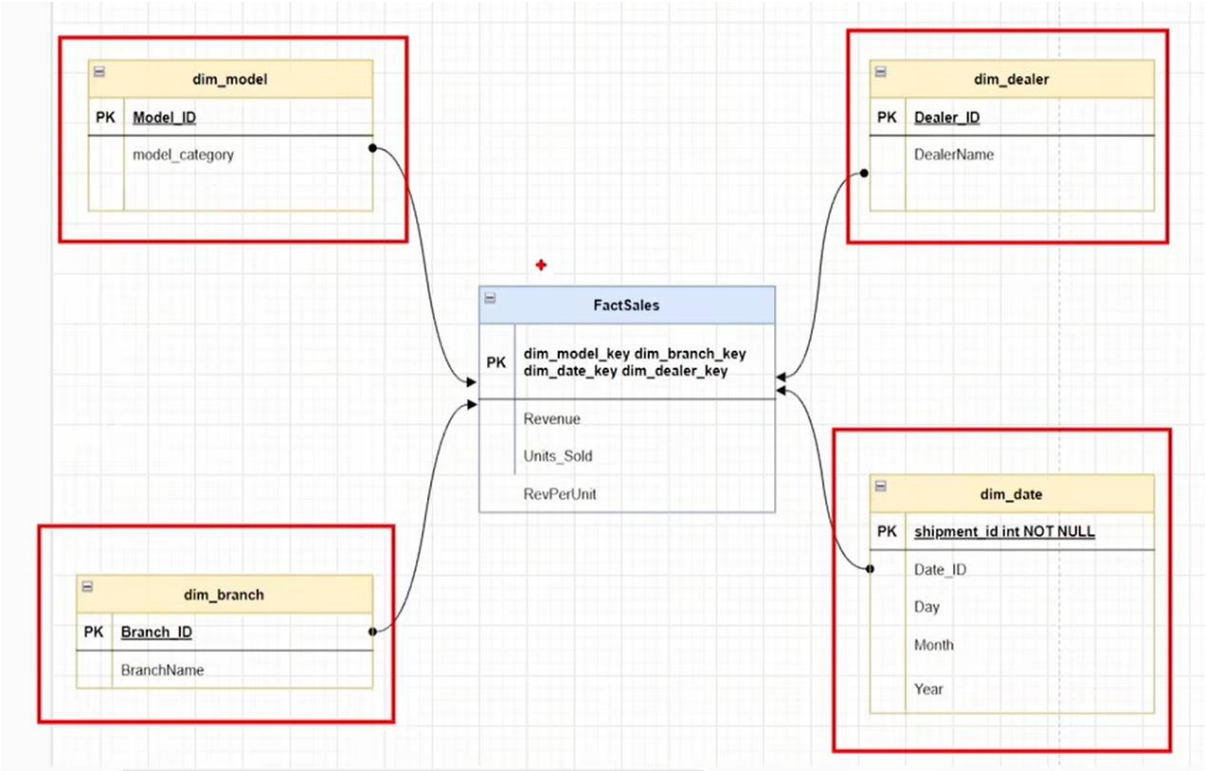


--Change dim_date as well create another clone file for dim_dealer



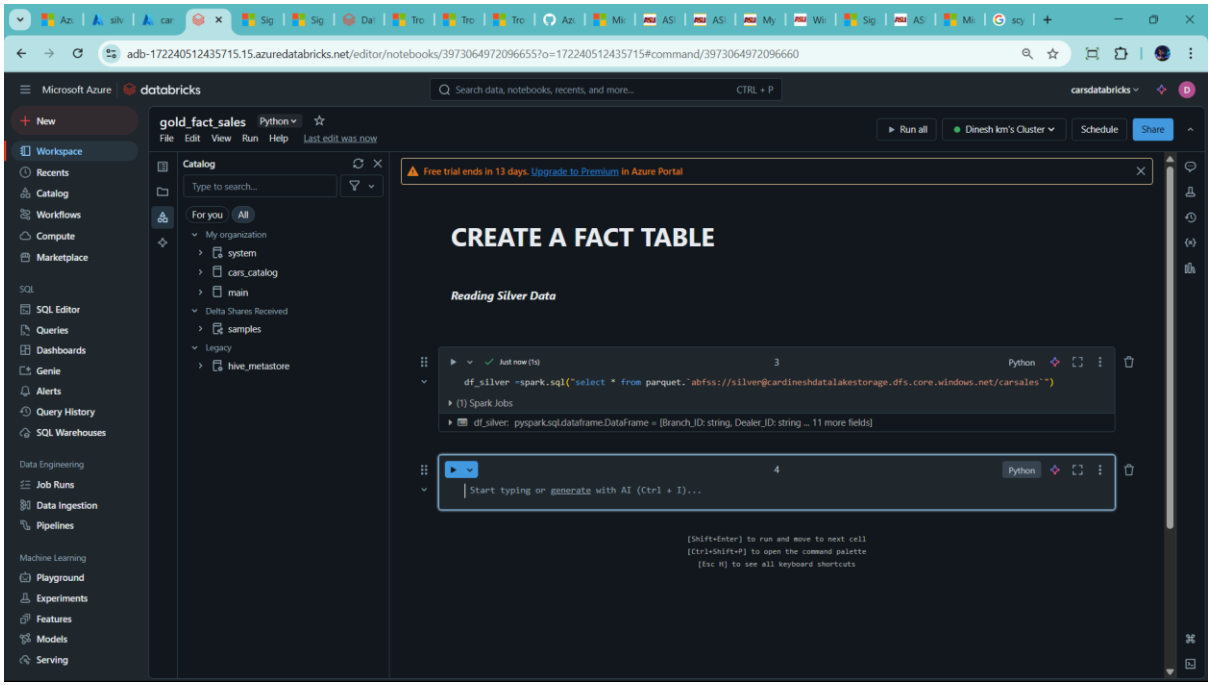
--use only one column date id



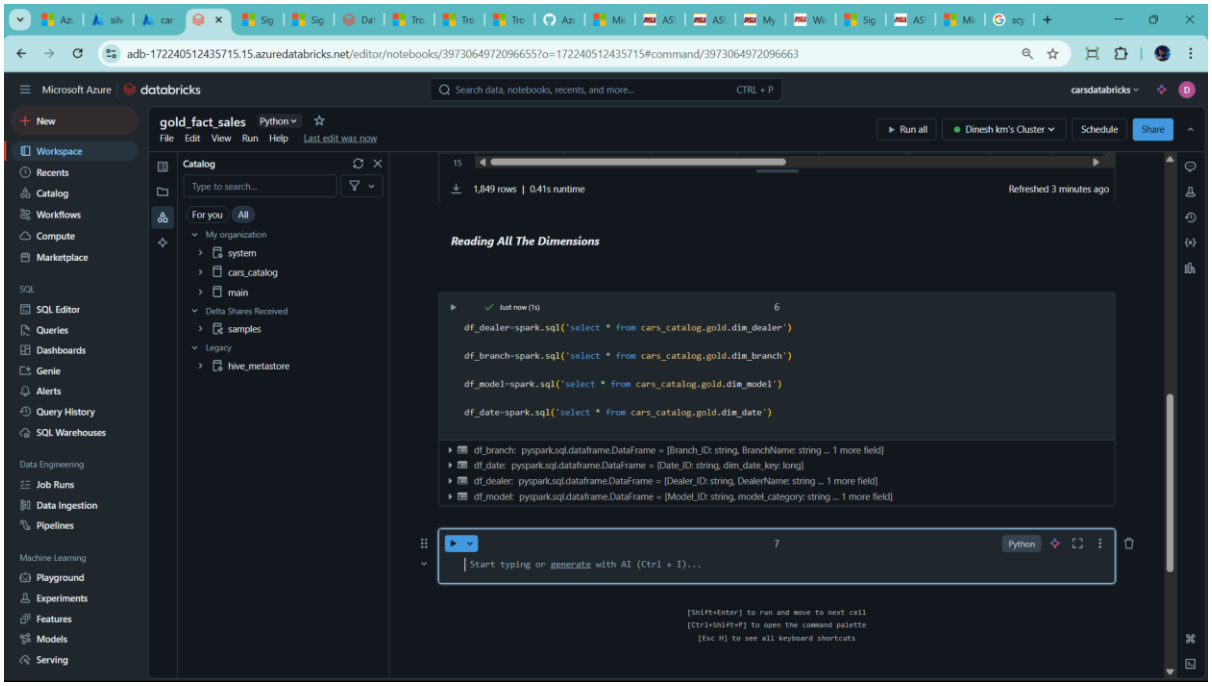


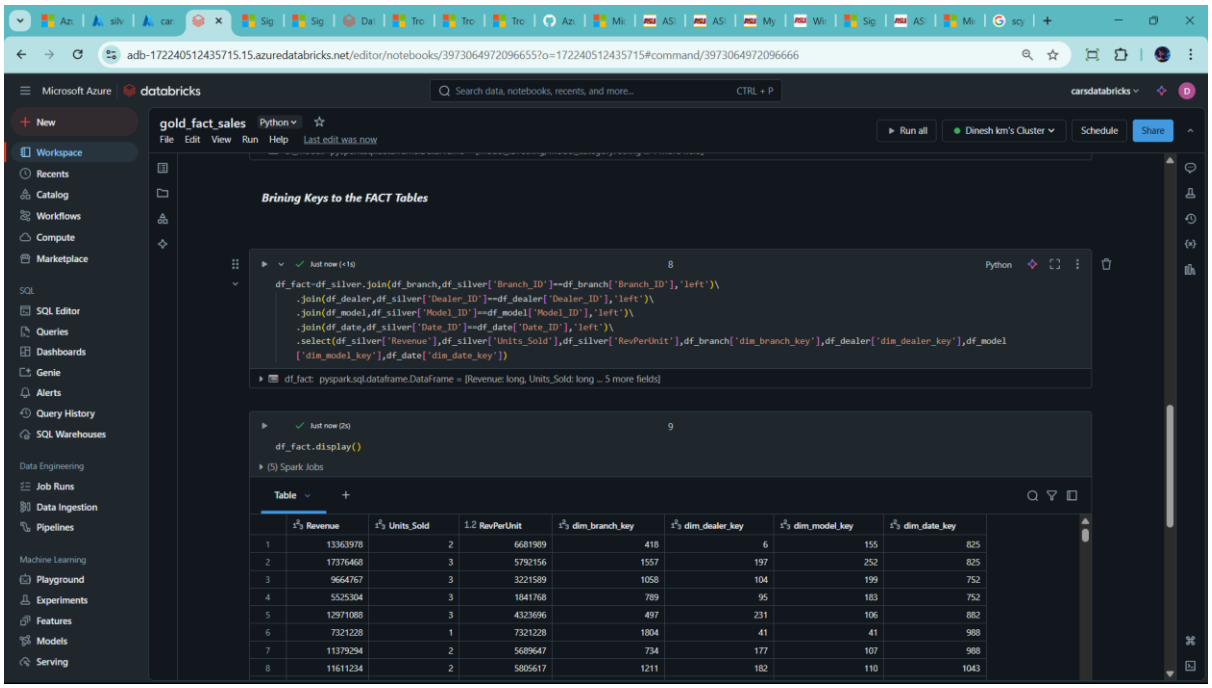
--Create all dim table as above → now create Fact table

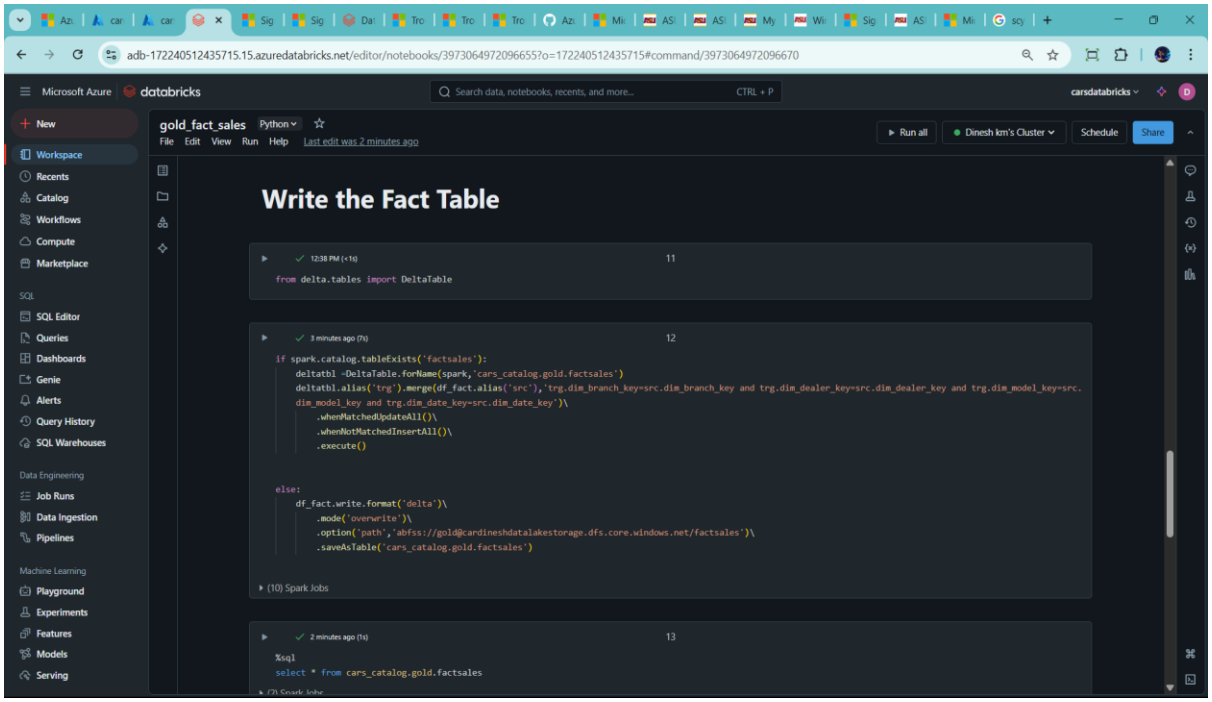
→ create new notebook → name as gold_fact_sales

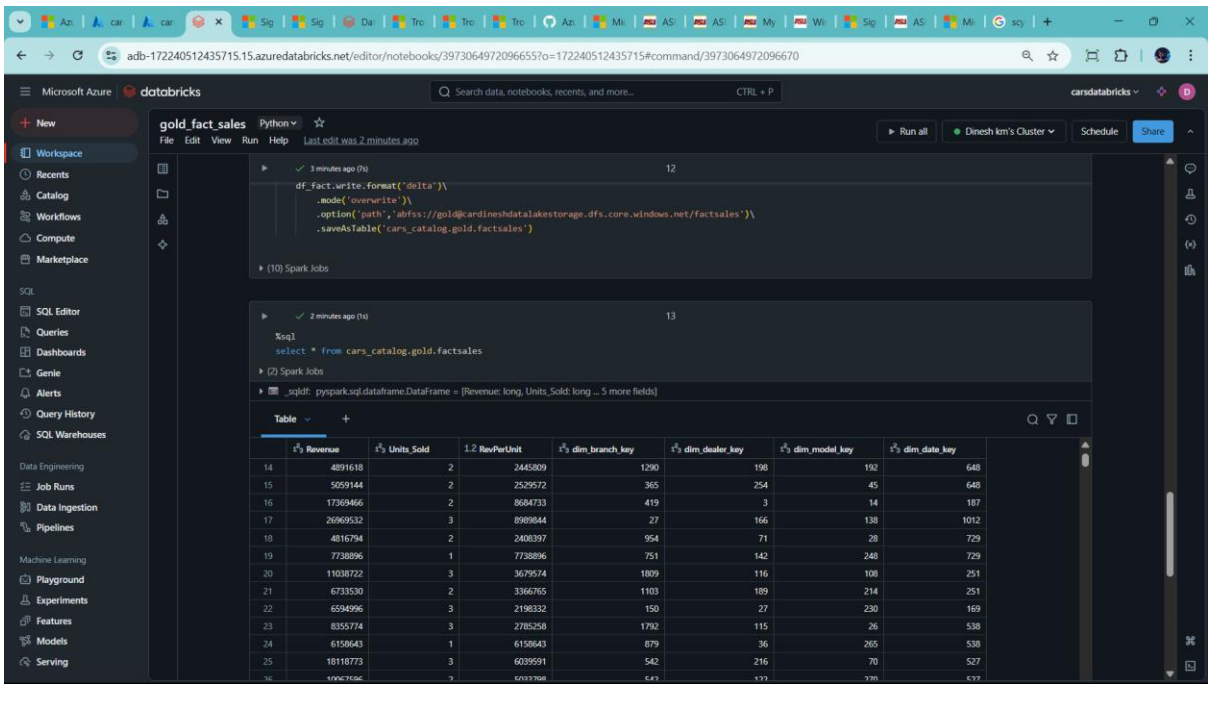


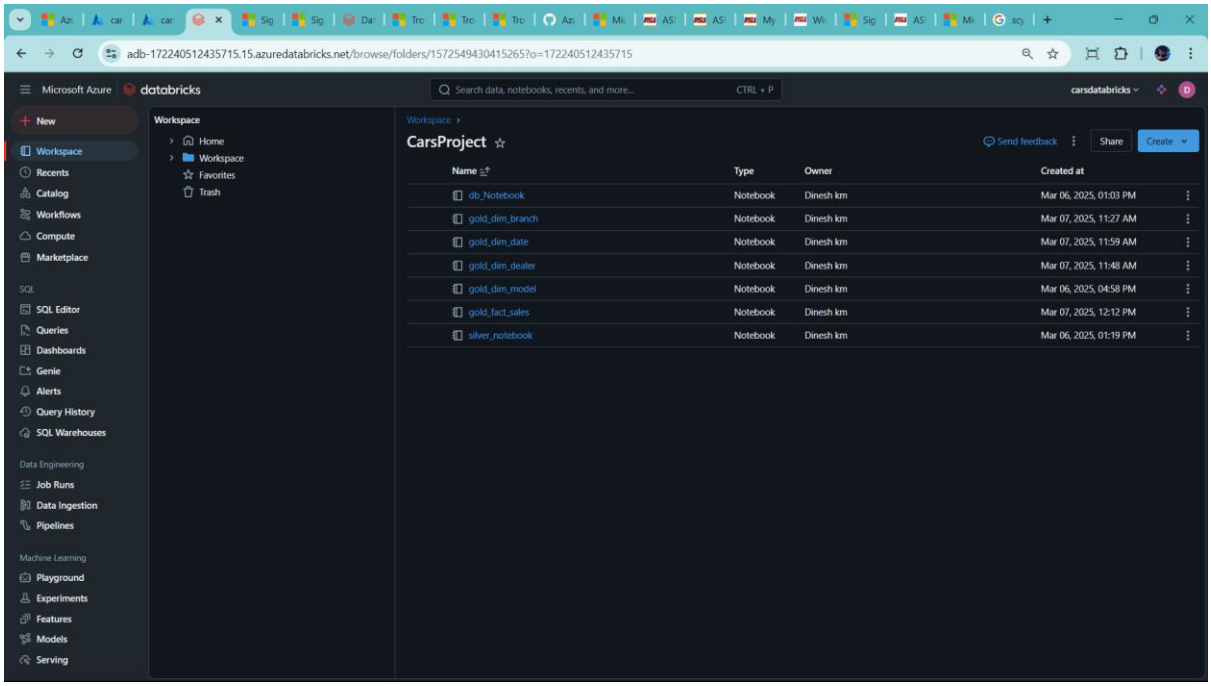
→retrvie the source data/silver


adb-1722405512435715.15.azuredatabricks.net/editor/notebooks/3973064972096655?o=1722405512435715#command/3973064972096663
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
gold_fact_sales
Python
File Edit View Run Help
Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Features
Models
Serving
Catalog
Type to search...
For you
All
My organization
system
cars_catalog
main
Delta Shares Received
samples
Legacy
hive_metastore
1,849 rows | 0.41s runtime
Refreshed 3 minutes ago
Reading All The Dimensions
Just now (1s)
6
df_dealer=spark.sql('select * from cars_catalog.gold.dim_dealer')
df_branch=spark.sql('select * from cars_catalog.gold.dim_branch')
df_model=spark.sql('select * from cars_catalog.gold.dim_model')
df_date=spark.sql('select * from cars_catalog.gold.dim_date')
df_branch: pyspark.sql.dataframe.DataFrame = [Branch_ID: string, BranchName: string ... 1 more field]
df_date: pyspark.sql.dataframe.DataFrame = [Date_ID: string, dim_date_key: long]
df_dealer: pyspark.sql.dataframe.DataFrame = [Dealer_ID: string, DealerName: string ... 1 more field]
df_model: pyspark.sql.dataframe.DataFrame = [Model_ID: string, model_category: string ... 1 more field]
7
Python
Start typing or generate with AI (Ctrl + I)...
[Shift+Enter] to run and move to next cell
[Ctrl+Shift+P] to open the command palette
[Esc H] to see all keyboard shortcuts

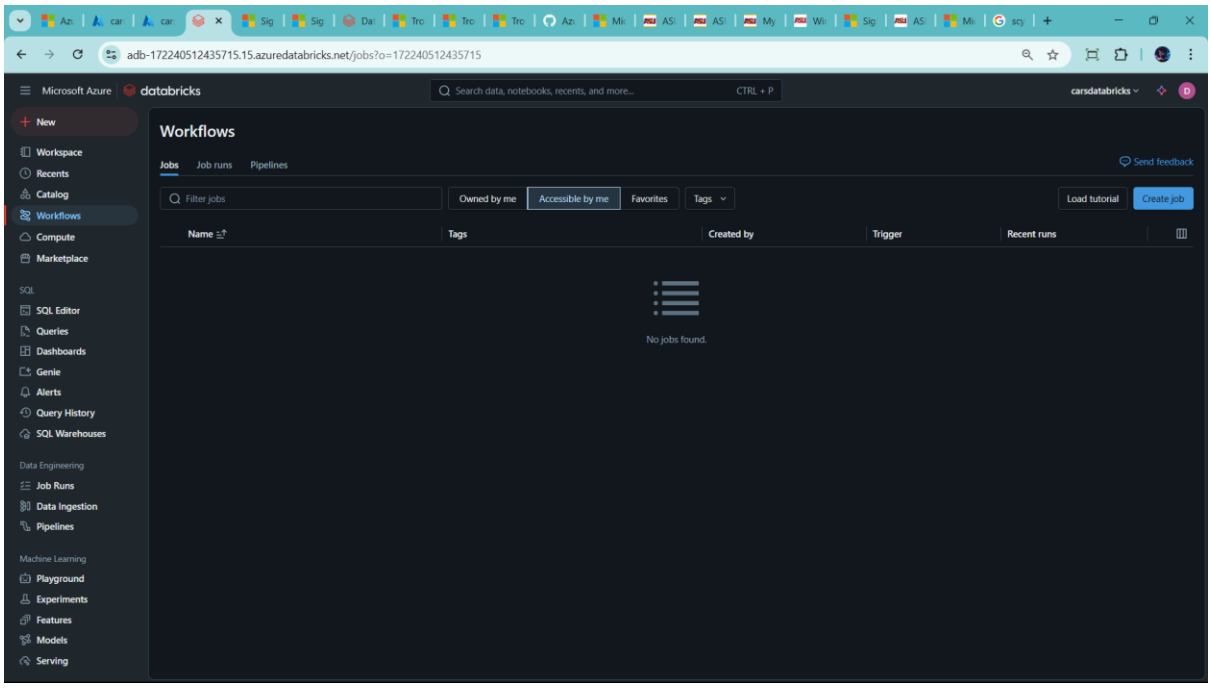

adb-1722405512435715.15.azuredatabricks.net/editor/notebooks/3973064972096655?o=1722405512435715#command/3973064972096666
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
gold_fact_sales
Python
File Edit View Run Help
Last edit was now
Run all
Dinesh km's Cluster
Schedule
Share
Brining Keys to the FACT Tables
Just now (<1s)
8
Python
df_fact=df_silver.join(df_branch,df_silver['Branch_ID']==df_branch['Branch_ID'],'left')\
.join(df_dealer,df_silver['Dealer_ID']==df_dealer['Dealer_ID'],'left')\
.join(df_model,df_silver['Model_ID']==df_model['Model_ID'],'left')\
.join(df_date,df_silver['Date_ID']==df_date['Date_ID'],'left')\
.select(df_silver['Revenue'],df_silver['Units_Sold'],df_silver['RevPerUnit'],df_branch['dim_branch_key'],df_dealer['dim_dealer_key'],df_model['dim_model_key'],df_date['dim_date_key'])
df_fact: pyspark.sql.dataframe.DataFrame = [Revenue: long, Units_Sold: long ... 5 more fields]
Just now (2s)
9
df_fact.display()
(5) Spark Jobs
Table
Revenue
Units_Sold
RevPerUnit
dim_branch_key
dim_dealer_key
dim_model_key
dim_date_key
1 13363978 2 6681989 418 6 155 825
2 17376468 3 5792156 1557 197 252 825
3 9664767 3 3221589 1058 104 199 752
4 5525304 3 1841768 789 95 183 752
5 12971088 3 4323696 497 231 106 882
6 7321228 1 7321228 1804 41 41 988
7 11379294 2 5689647 734 177 107 988
8 11611234 2 5805617 1211 182 110 1043


adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/3973064972096655?o=1722405124357 15#command/3973064972096670
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Machine Learning
Playground
Experiments
Features
Models
Serving
gold_fact_sales
Python
File Edit View Run Help
Last edit was 2 minutes ago
Run all
Dinesh km's Cluster
Schedule
Share
Write the Fact Table
11
12:38 PM (<1s)
from delta.tables import DeltaTable
12
3 minutes ago (7s)
if spark.catalog.tableExists('factsales'):
deltatbl = DeltaTable.forName(spark,'cars_catalog.gold.factsales')
deltatbl.alias('trg').merge(df_fact.alias('src'),'trg.dim_branch_key=src.dim_branch_key and trg.dim_dealer_key=src.dim_dealer_key and trg.dim_model_key=src.dim_model_key and trg.dim_date_key=src.dim_date_key')\
.whenMatchedUpdateAll()\
.whenNotMatchedInsertAll()\
.execute()
else:
df_fact.write.format('delta')\
.mode('overwrite')\
.option('path','abfss://gold@cardineshdatalakestorage.dfs.core.windows.net/factsales')\
.saveAsTable('cars_catalog.gold.factsales')
(10) Spark Jobs
13
2 minutes ago (1s)
%sql
select * from cars_catalog.gold.factsales

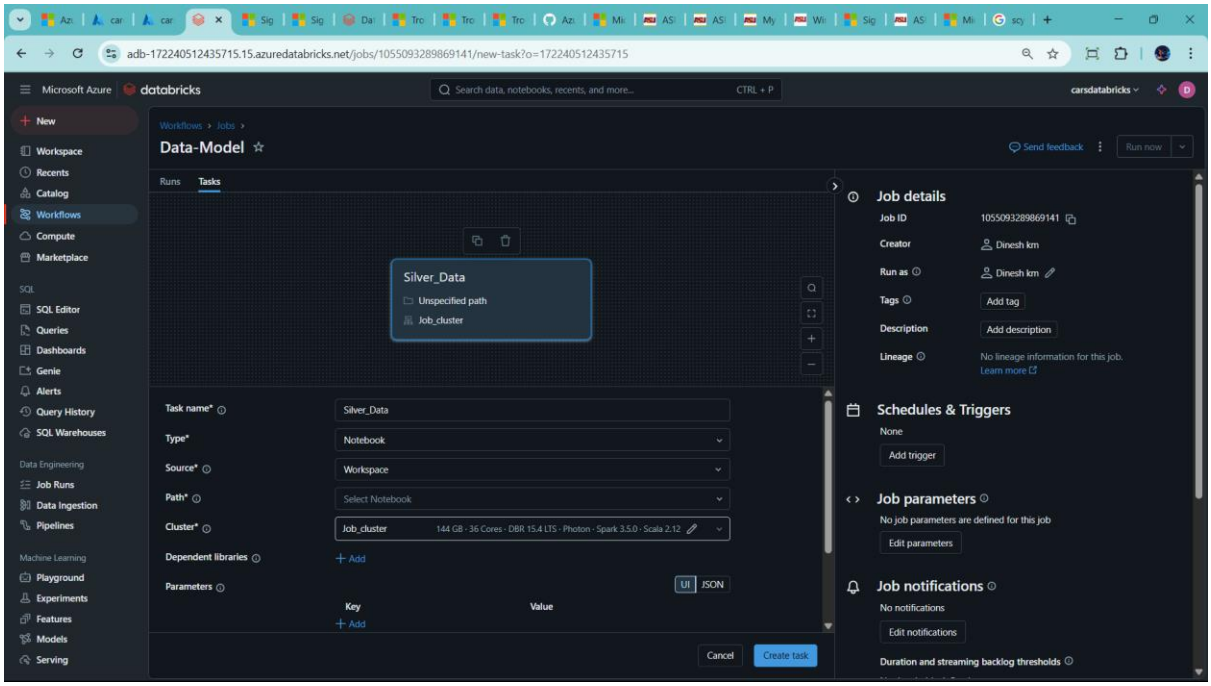

adb-1722405124357 15.15.azuredatabricks.net/editor/notebooks/3973064972096655?o=1722405124357 15#command/3973064972096670
Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
carsdatabricks
gold_fact_sales
Python
File Edit View Run Help
Last edit was 2 minutes ago
Run all
Dinesh km's Cluster
Schedule
Share
12
3 minutes ago (7s)
df_fact.write.format('delta')\
.mode('overwrite')\
.option('path','abfss://gold@cardineshdatalakestorage.dfs.core.windows.net/factsales')\
.saveAsTable('cars_catalog.gold.factsales')
(10) Spark Jobs
13
2 minutes ago (1s)
%sql
select * from cars_catalog.gold.factsales
(2) Spark Jobs
_sqldf: pyspark.sql.dataframe.DataFrame = [Revenue: long, Units_Sold: long ... 5 more fields]
Table
Revenue
Units_Sold
RevPerUnit
dim_branch_key
dim_dealer_key
dim_model_key
dim_date_key
14 4891618 2 2445809 1290 198 192 648
15 5059144 2 2529572 365 254 45 648
16 17369466 2 8684733 419 3 14 187
17 26969532 3 8989844 27 166 138 1012
18 4816794 2 2408397 954 71 28 729
19 7738896 1 7738896 751 142 248 729
20 11038722 3 3679574 1809 116 108 251
21 6733530 2 3366765 1103 189 214 251
22 6594996 3 2198332 150 27 230 169
23 8355774 3 2785258 1792 115 26 538
24 6158643 1 6158643 879 36 265 538
25 18118773 3 6039591 542 216 70 527



→Now Create WorkFlow in DataBricks

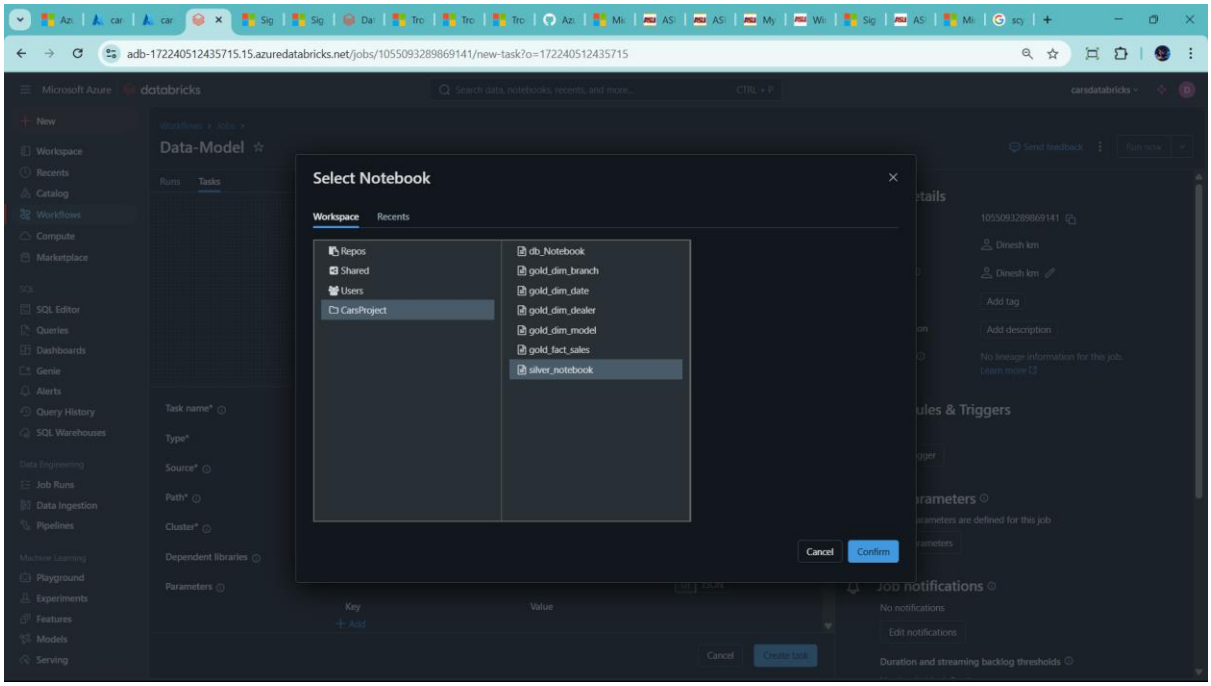


➔ Create job



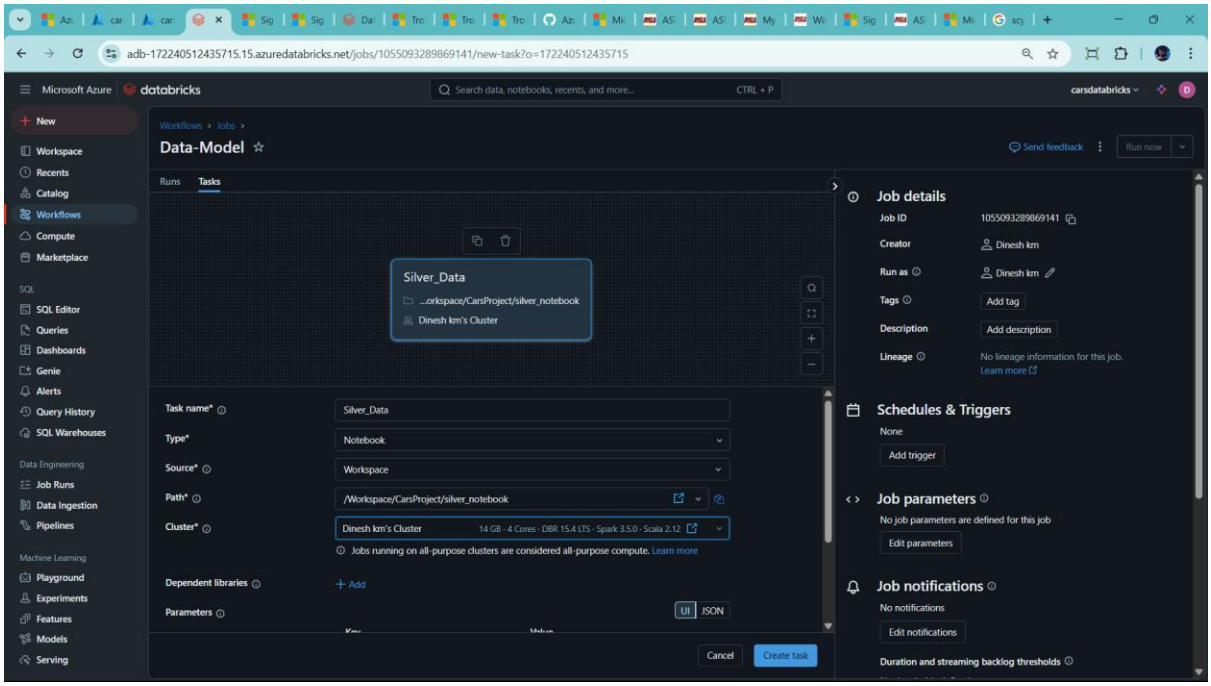
→name as Data-model

- ➔ Task name – Silver_data
- ➔ Click path – choose silver_notebook

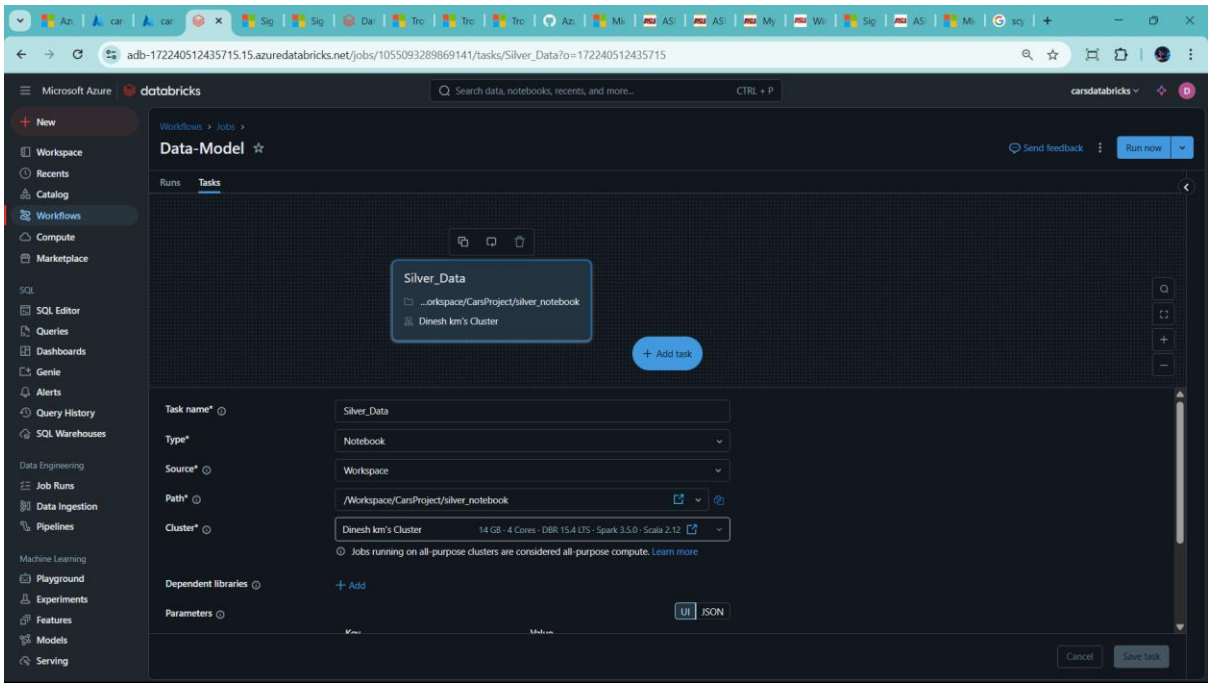


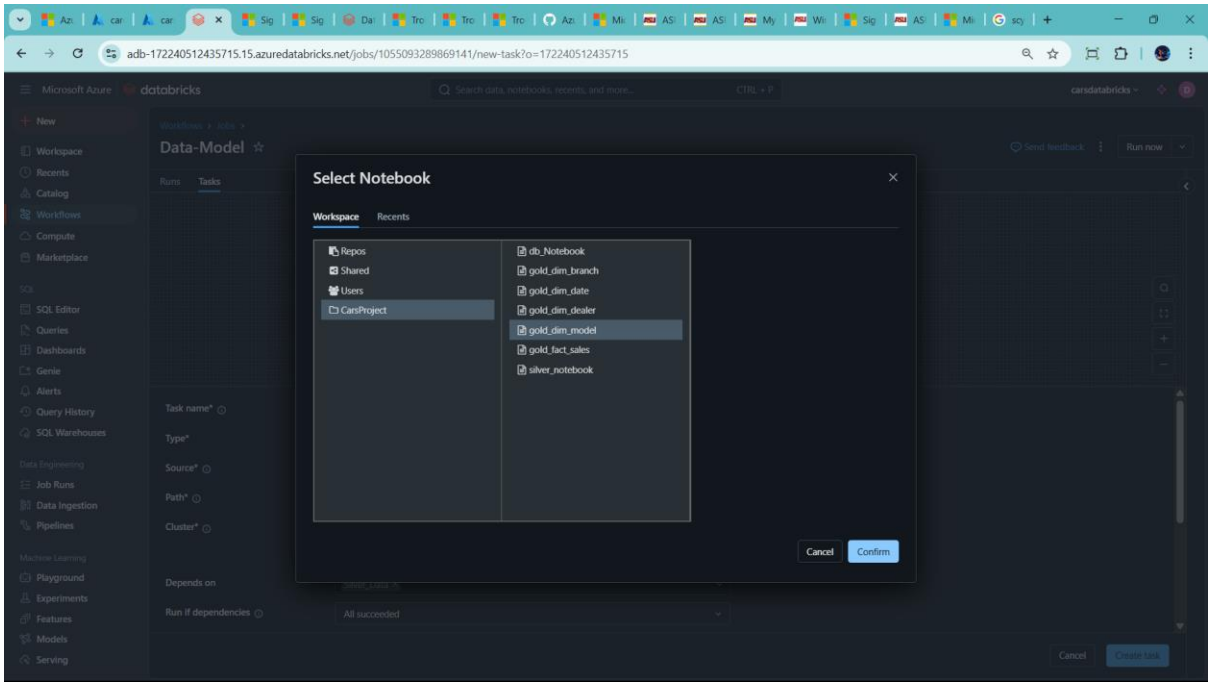
→chose cluse as well

→click create task

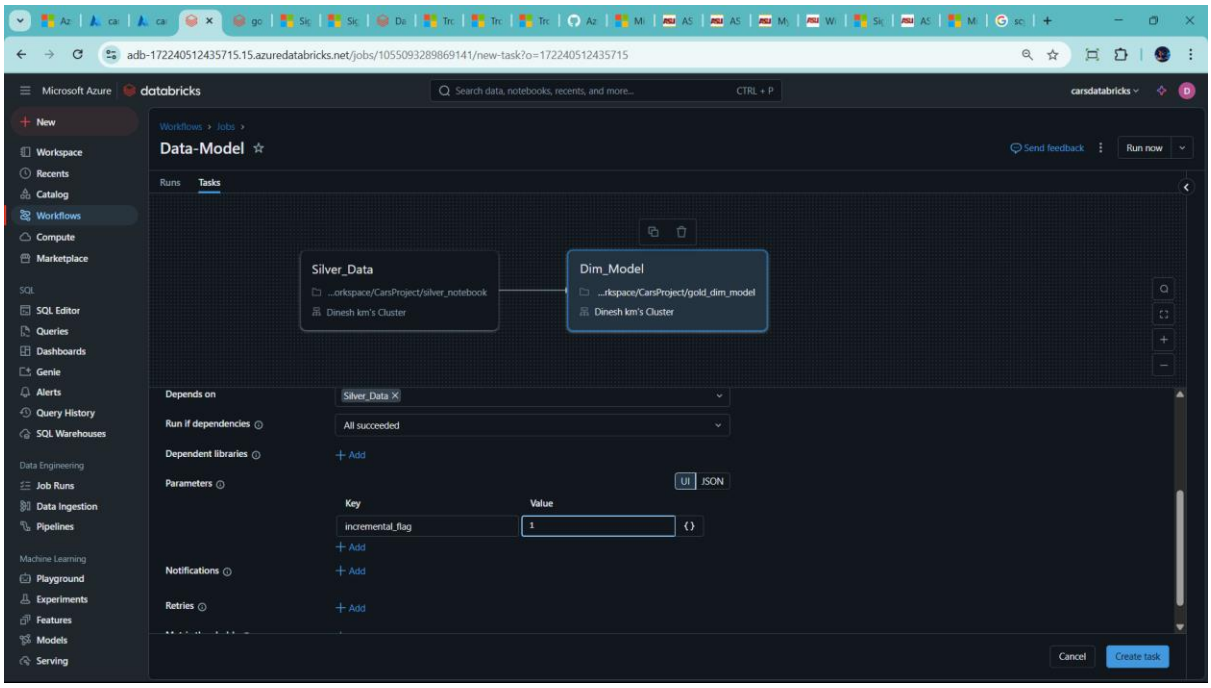


--click add task dim model



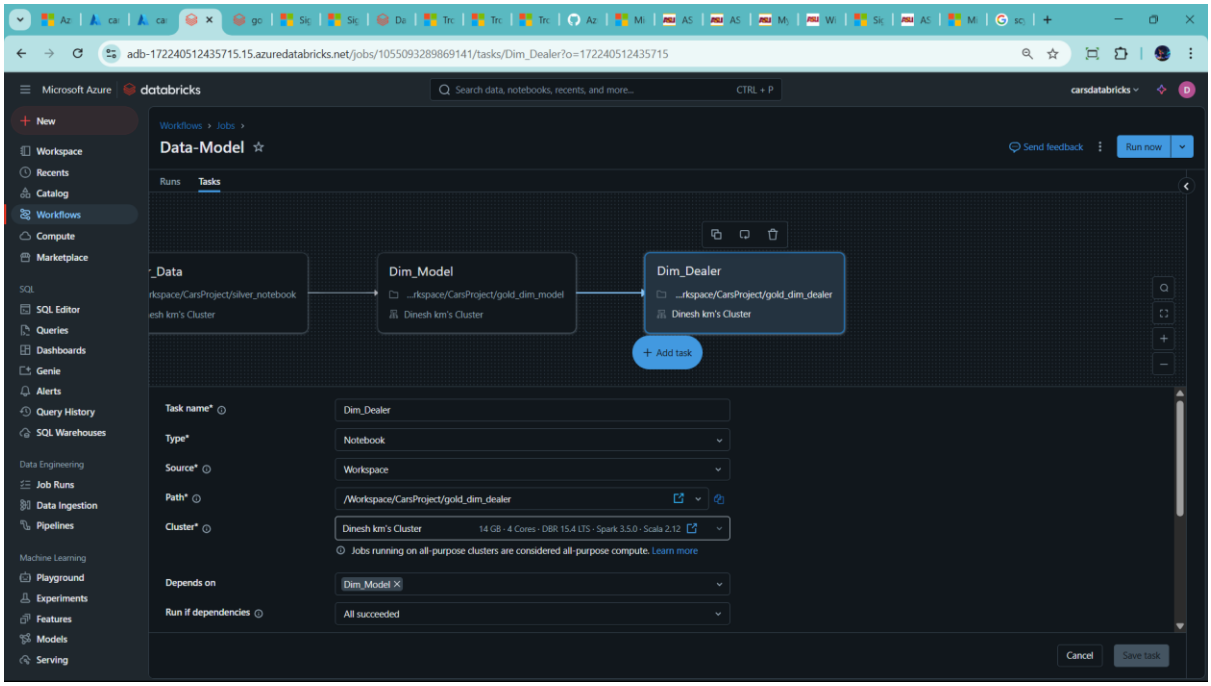


→add parameter →incremental_flag

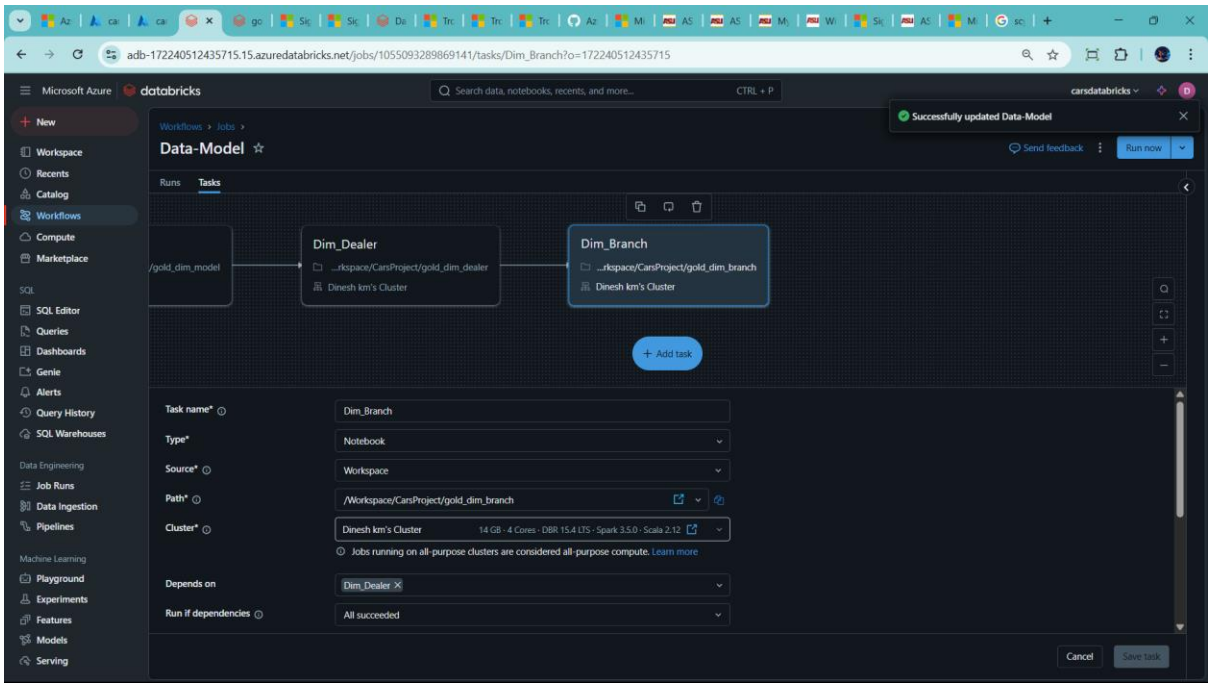


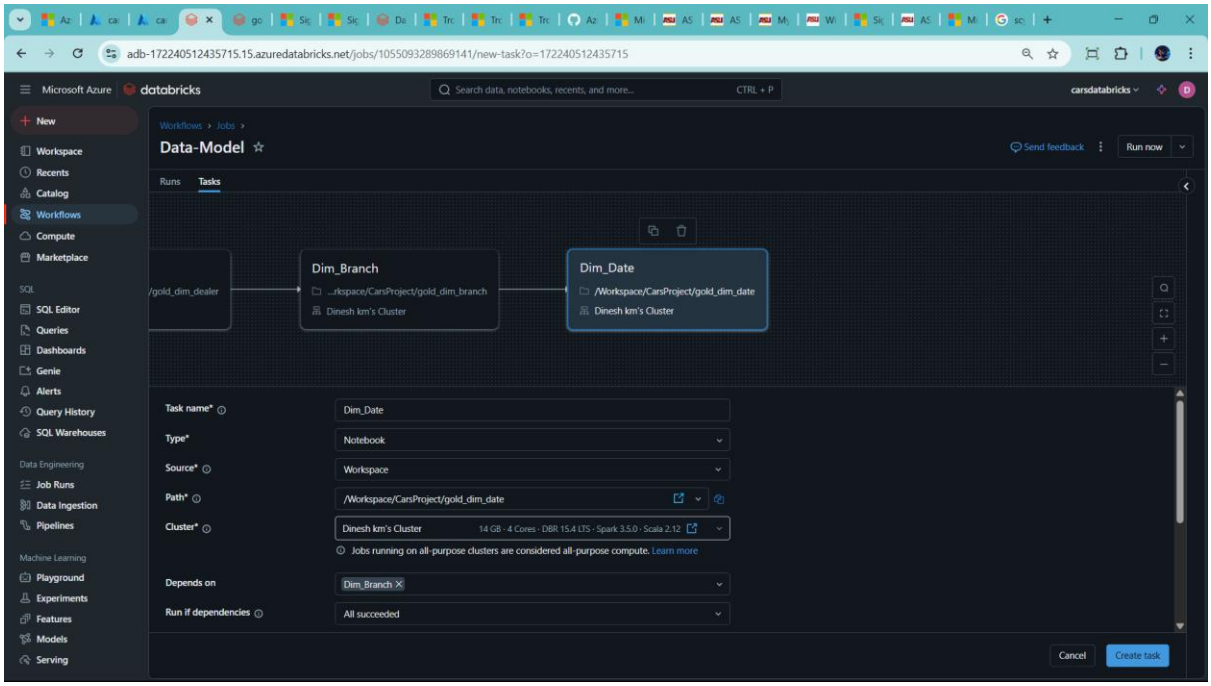
--click create

→same as other Dim assign task

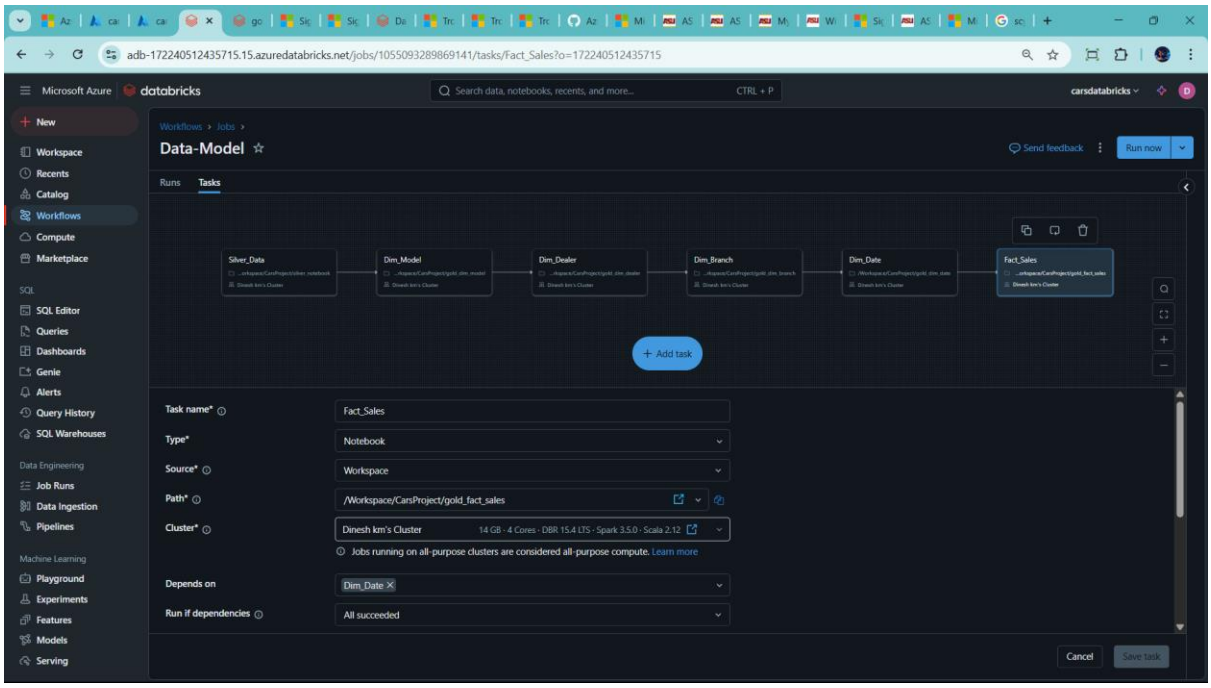


--create another task Dim_Branch

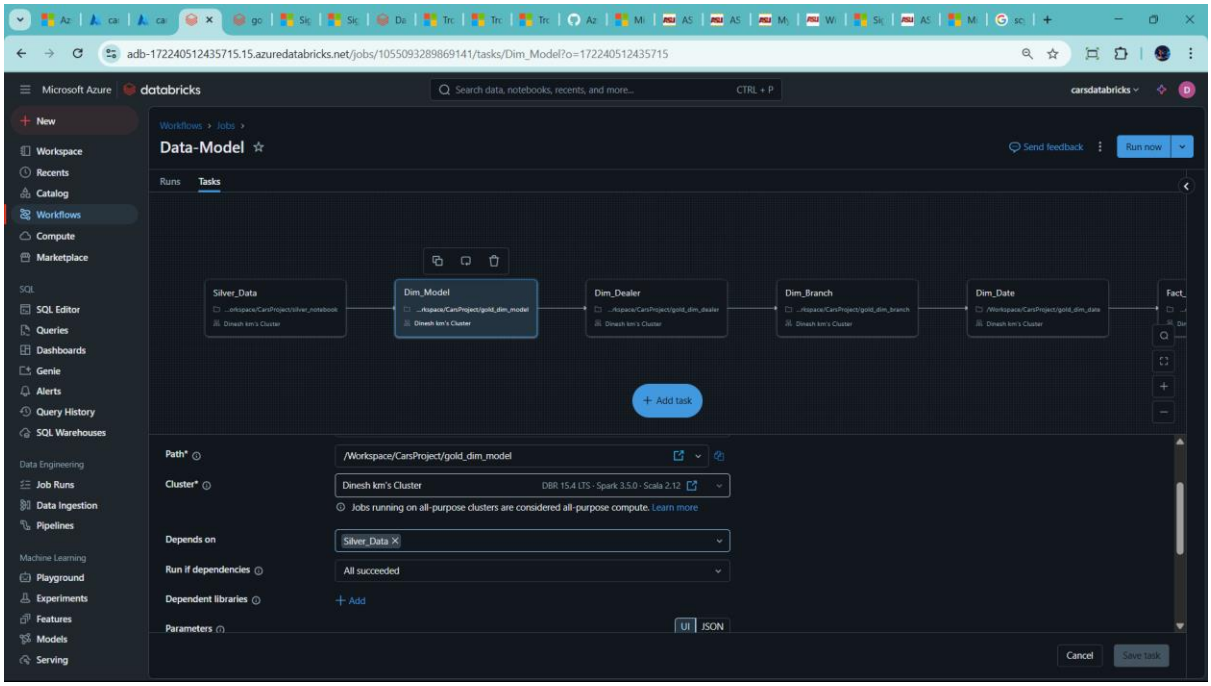




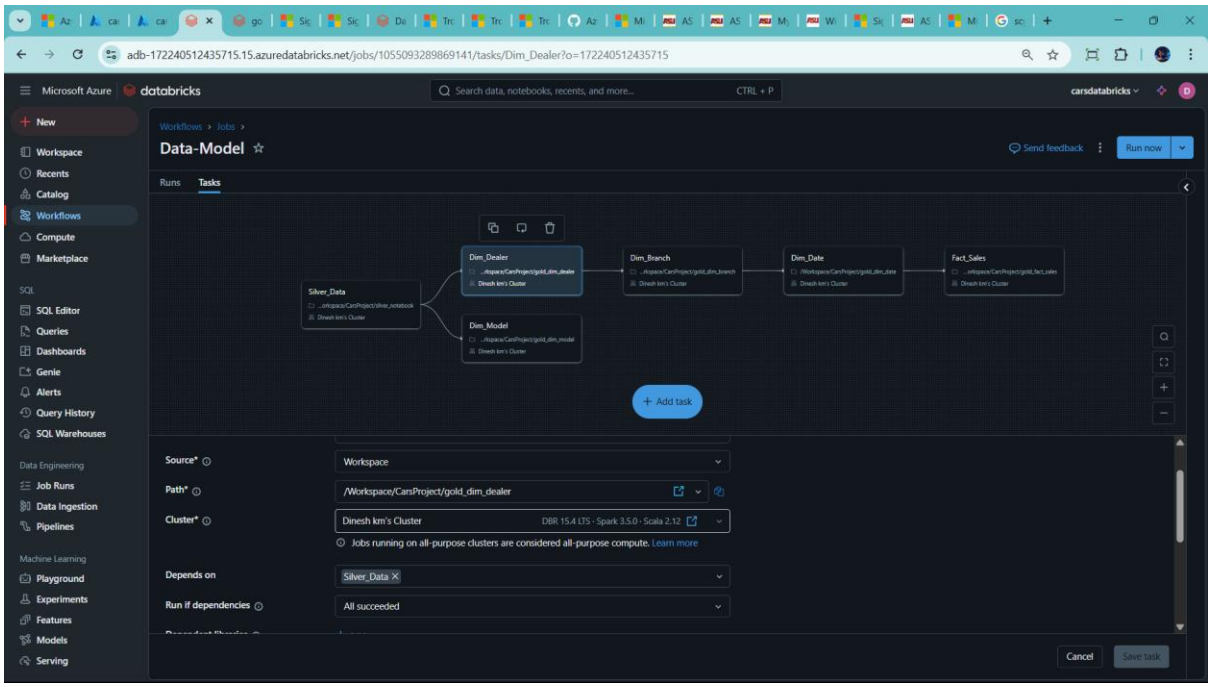
--Create task as fact_sales notebook as well

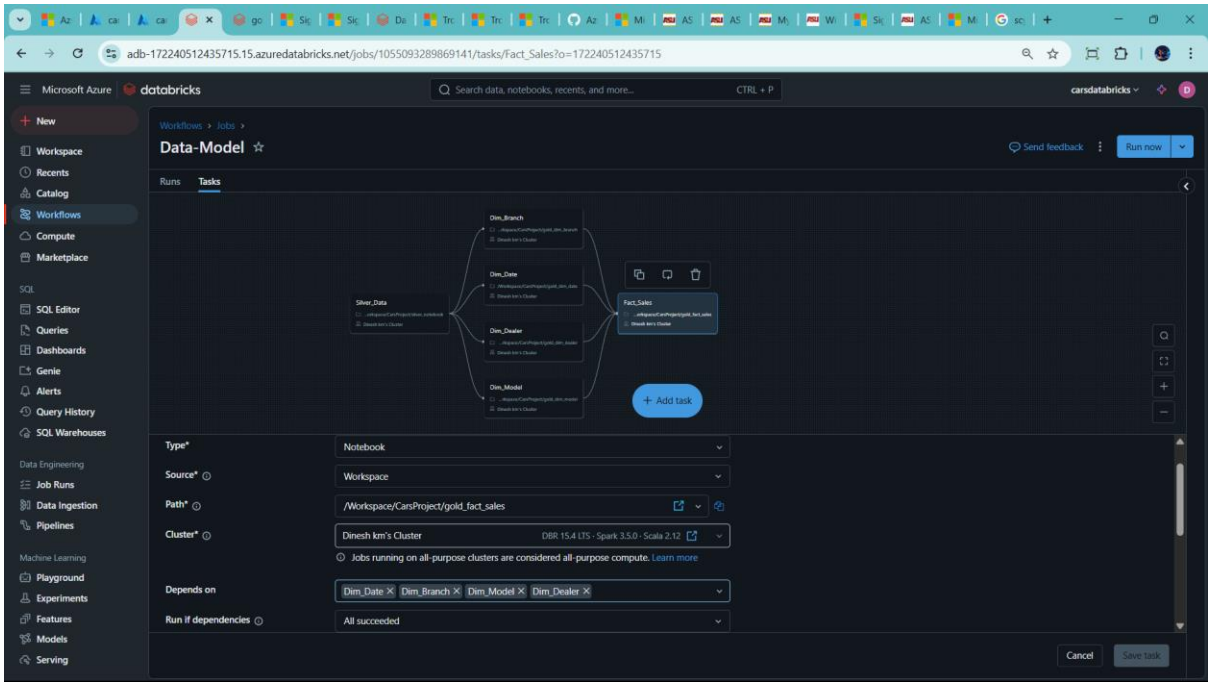


→I want some flow run in parallel



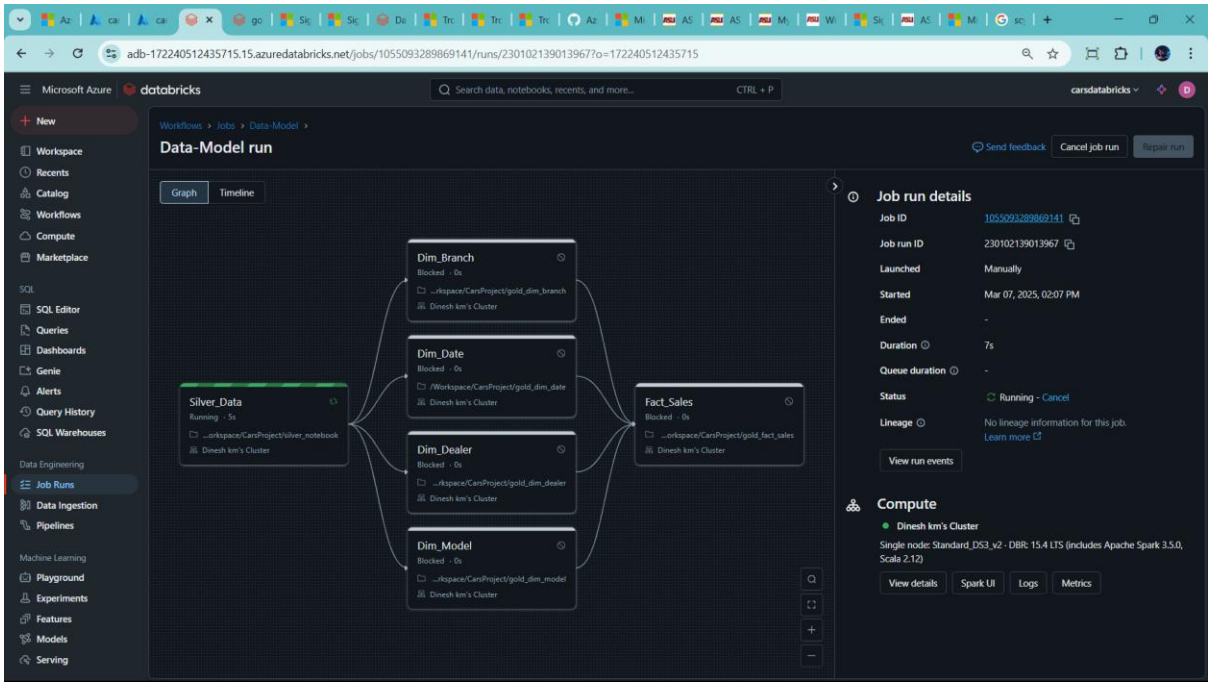
--click Depends on → choose silver data





→merge all dim task in fact task

--finally click run

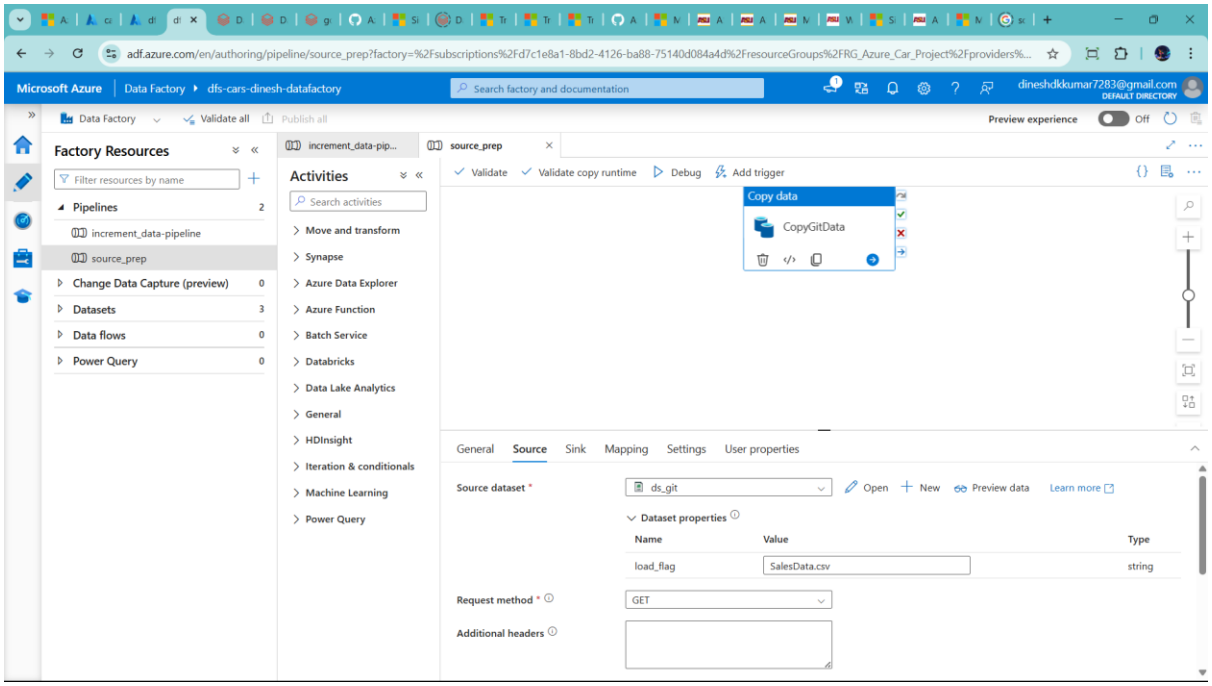


→go to SQL query editor check table

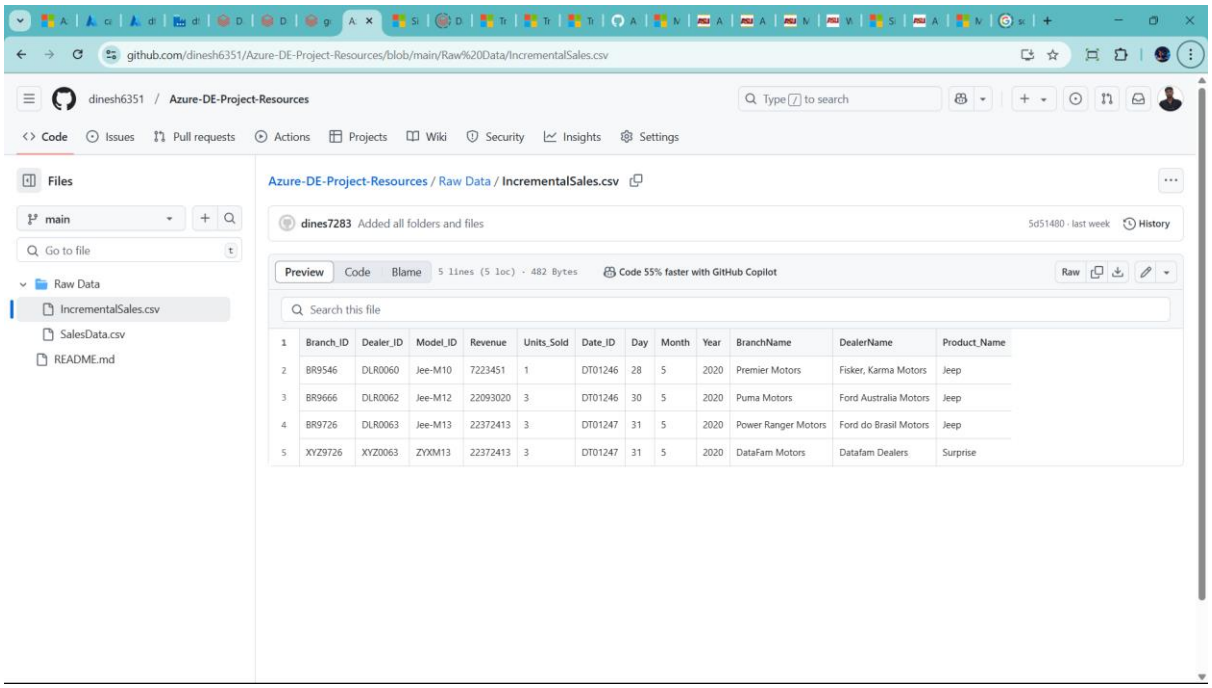
→Success

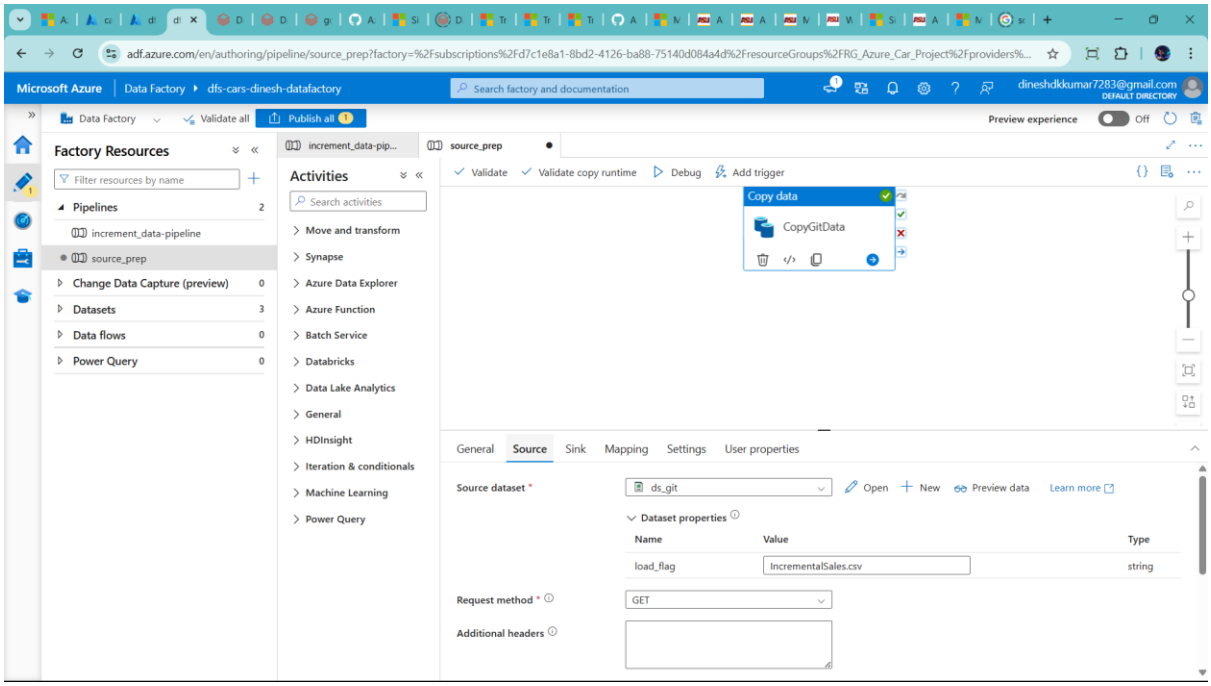
→Try incremental load method

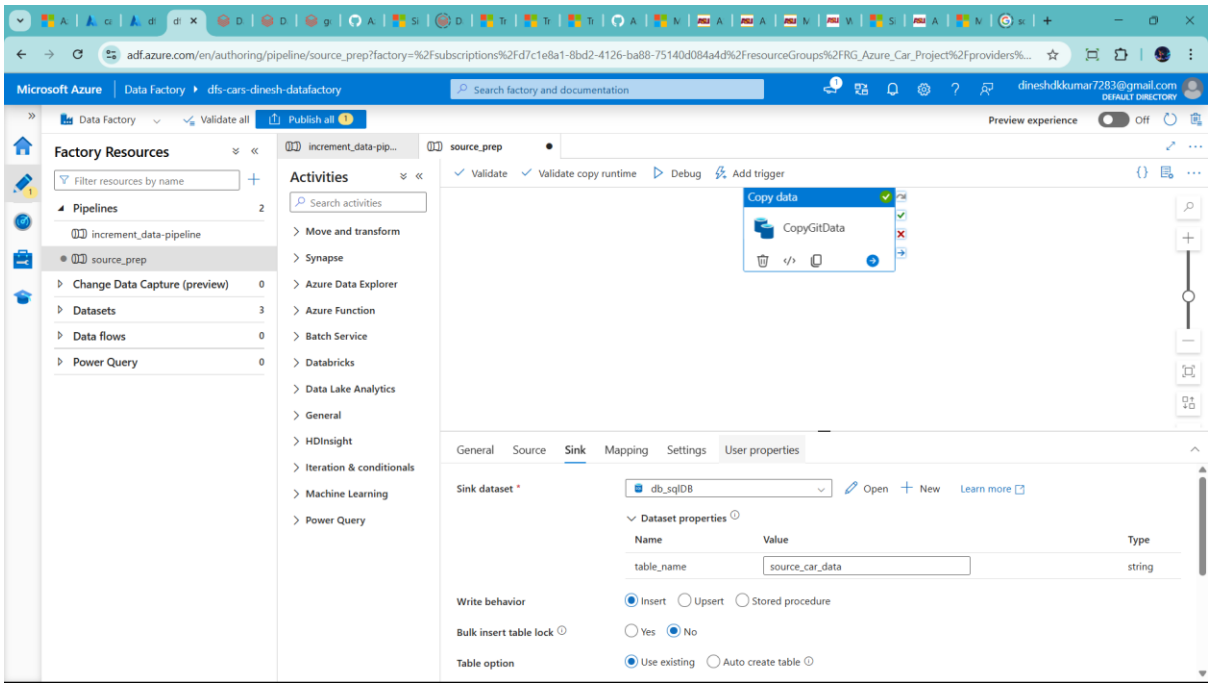
Go data factory



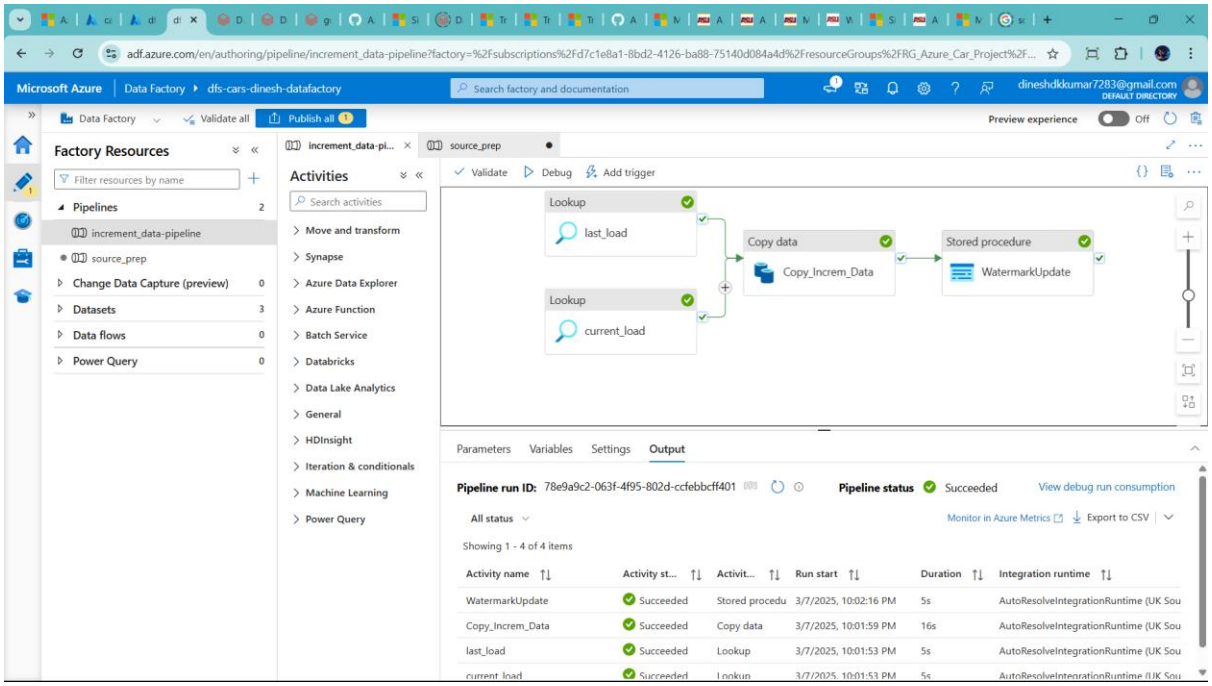
They have new data and old data with updated

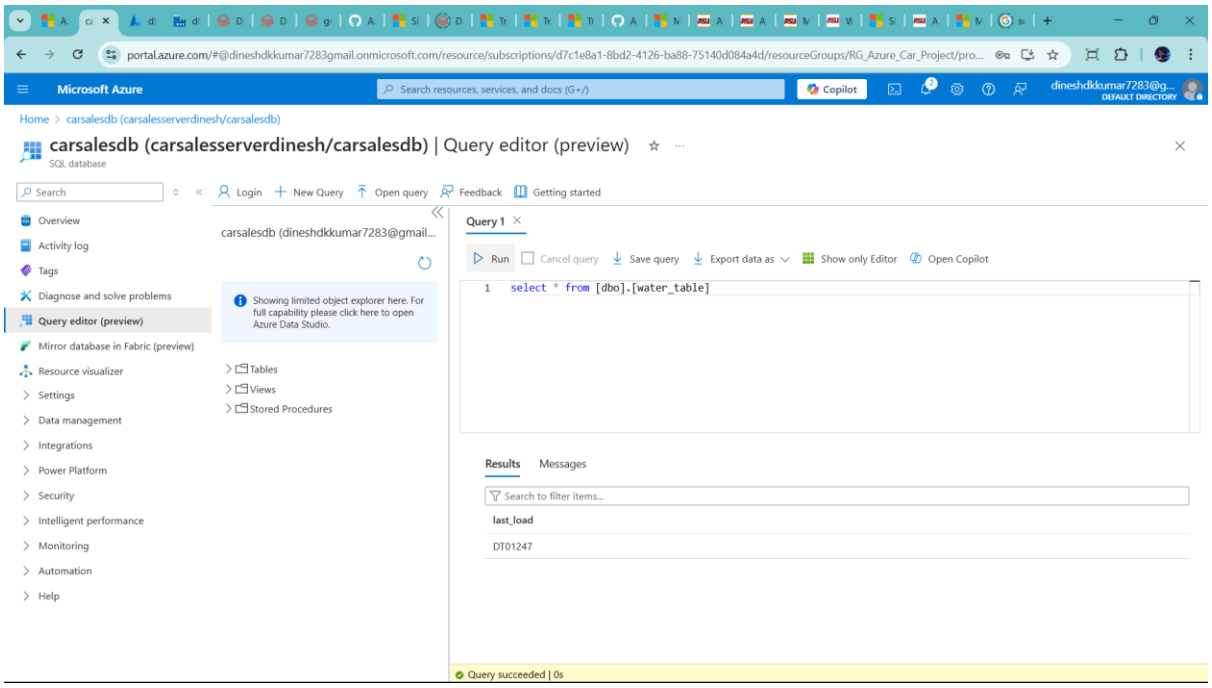



Microsoft Azure
Data Factory
dfs-cars-dinesh-datafactory
Search factory and documentation
dineshdkkumar7283@gmail.com
DEFAULT DIRECTORY
Validate all
Publish all
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines
increment_data-pipeline
source_prep
Change Data Capture (preview)
Datasets
Data flows
Power Query
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Copy data
CopyGitData
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
ds_git
Open
New
Preview data
Learn more
Dataset properties
Name
Value
Type
load_flag
IncrementalSales.csv
string
Request method *
GET
Additional headers


Microsoft Azure
Data Factory
dfs-cars-dinesh-datafactory
Search factory and documentation
dineshdkkumar7283@gmail.com
DEFAULT DIRECTORY
Validate all
Publish all
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines
increment_data-pipeline
source_prep
Change Data Capture (preview)
Datasets
Data flows
Power Query
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Copy data
CopyGitData
General
Source
Sink
Mapping
Settings
User properties
Sink dataset *
db_sqlDB
Open
New
Learn more
Dataset properties
Name
Value
Type
table_name
source_car_data
string
Write behavior
Insert
Upsert
Stored procedure
Bulk insert table lock
Yes
No
Table option
Use existing
Auto create table

Now run

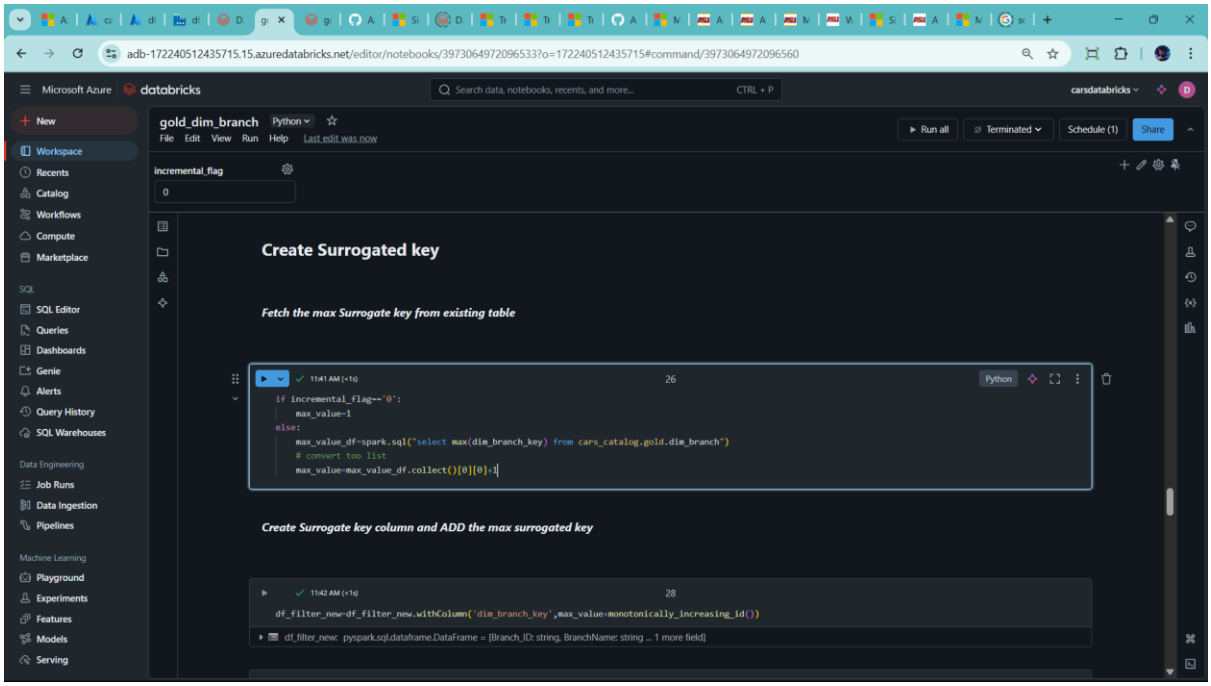


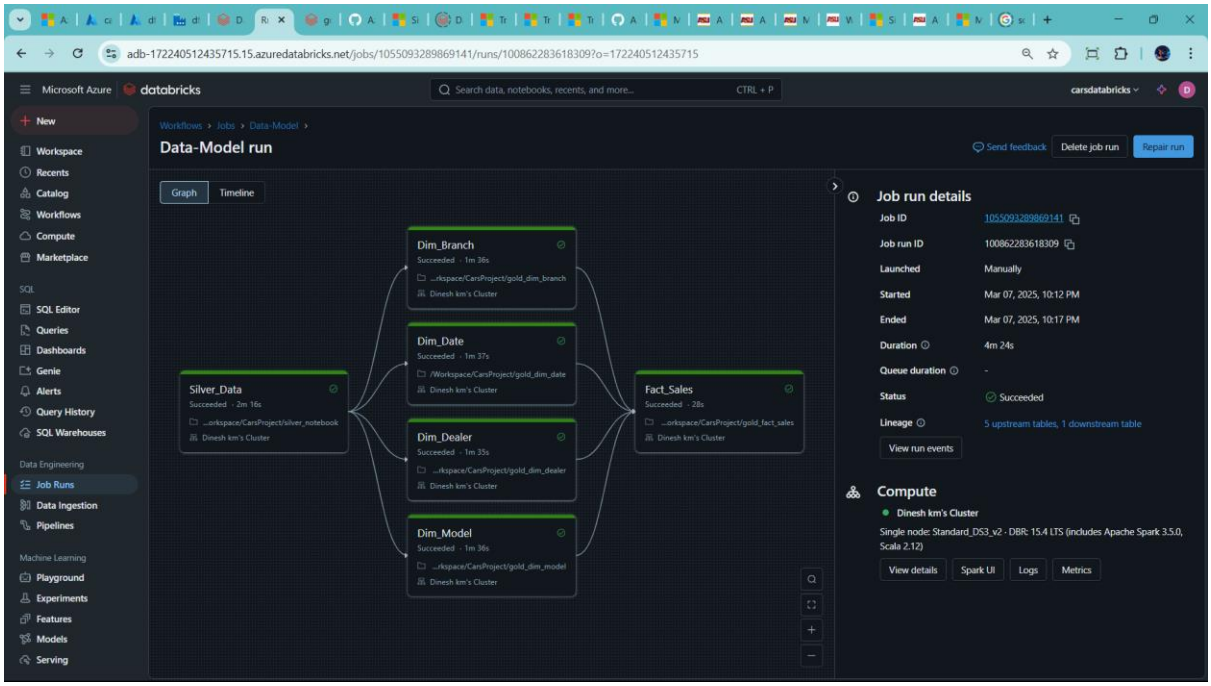


➔ Run work flow in databricks

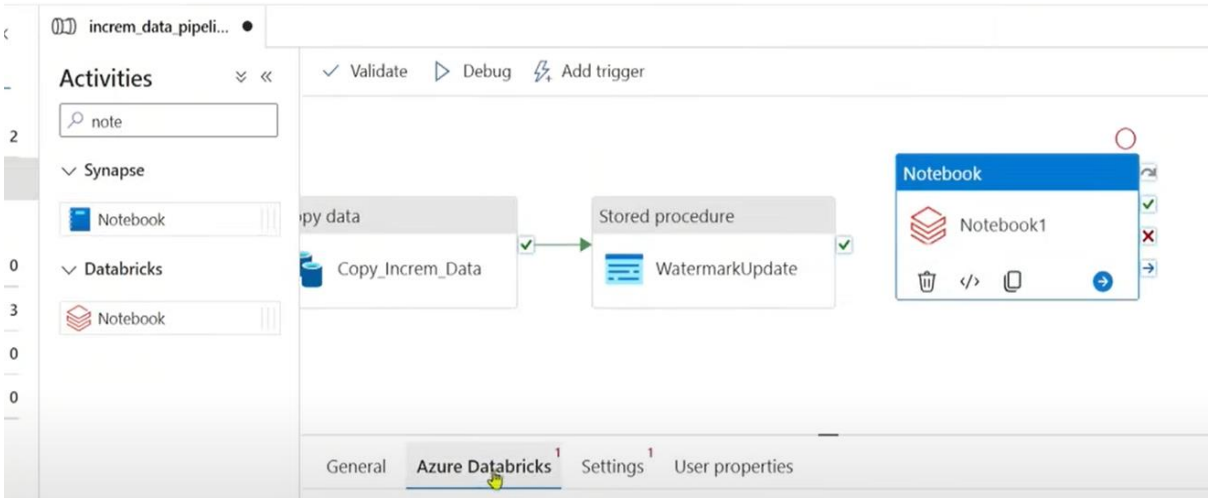
➔ Small mistake to miss add +1 in else as well

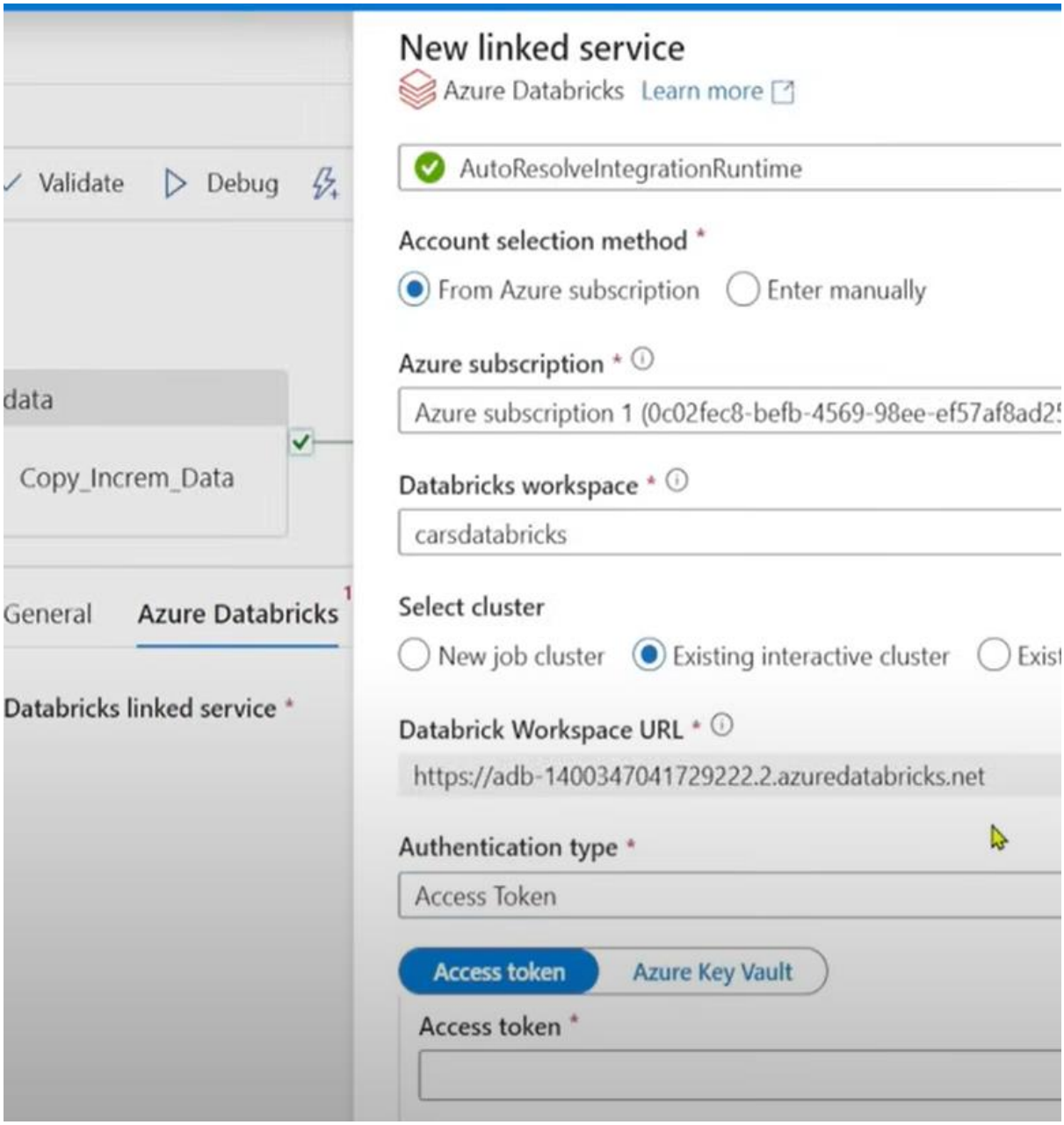
➔ Apply all dim table



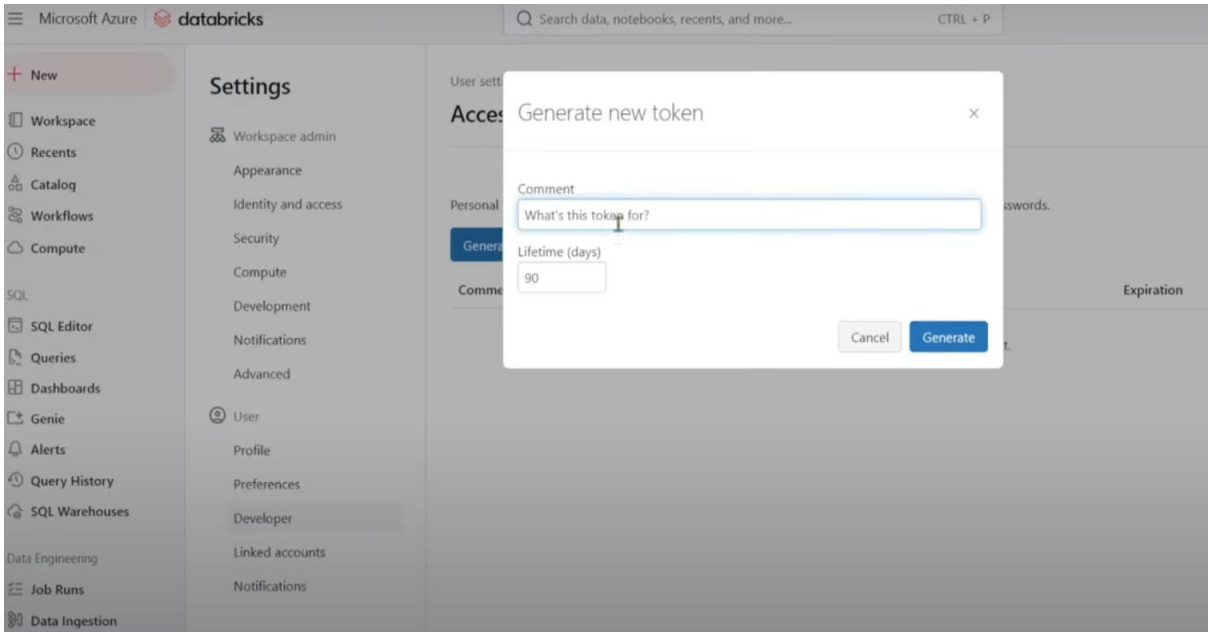


➔Also use notebook in Data factory as well

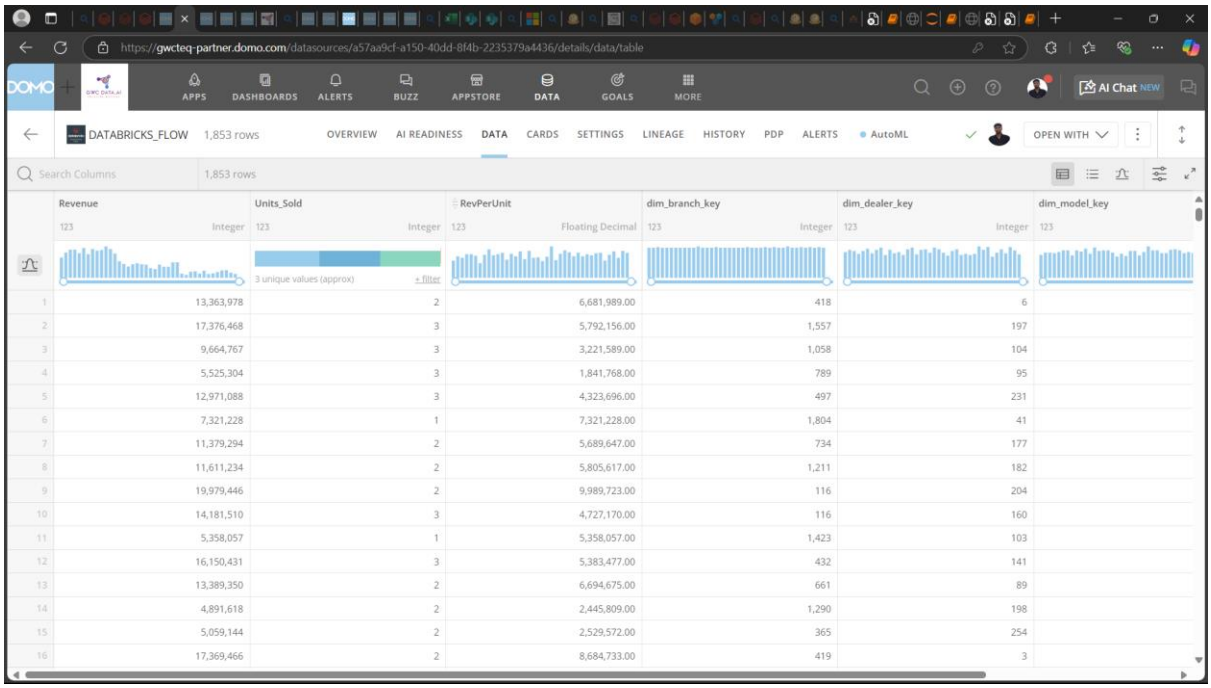




Find access token



Successful connect



	Revenue	Units_Sold	RevPerUnit	dim_branch_key	dim_dealer_key	dim_model_key
	123 Integer	123 Integer	123 Floating Decimal	123 Integer	123 Integer	123
1	13,363,978	2	6,681,989.00	418	6	
2	17,376,468	3	5,792,156.00	1,557	197	
3	9,664,767	3	3,221,589.00	1,058	104	
4	5,525,304	3	1,841,768.00	789	95	
5	12,971,088	3	4,323,696.00	497	231	
6	7,321,228	1	7,321,228.00	1,804	41	
7	11,379,294	2	5,689,647.00	734	177	
8	11,611,234	2	5,805,617.00	1,211	182	
9	19,979,446	2	9,989,723.00	116	204	
10	14,181,510	3	4,727,170.00	116	160	
11	5,358,057	1	5,358,057.00	1,423	103	
12	16,150,431	3	5,383,477.00	432	141	
13	13,389,350	2	6,694,675.00	661	89	
14	4,891,618	2	2,445,809.00	1,290	198	
15	5,059,144	2	2,529,572.00	365	254	
16	17,369,466	2	8,684,733.00	419	3	

→Follow this step connect databrick in DOMO

1. Enable Databricks SQL Warehouse

Ensure that your **Databricks SQL Warehouse** (formerly SQL Endpoint) is enabled:

- Go to **Databricks Workspace** > **SQL** > **SQL Warehouses**.
- Start or create a new SQL Warehouse.
- Copy the **Server Hostname**, **Port**, and **HTTP Path** from the **Connection Details**.

---

2. Get Authentication Credentials

- You need an **Access Token** from Databricks.
- In **Databricks**, go to **User Settings** > **Access Tokens** > **Generate New Token**.
- Copy and save the token securely.

---

3. Connect Databricks to Domo

1. **Go to Domo** > **Data Center** > **Connectors**.
2. Search for **Databricks Connector** and select it.
3. Click **Get the Data**.
4. Fill in the connection details:
    - **Host**: <Databricks Server Hostname>
    - **Port**: 443
    - **HTTP Path**: <Databricks HTTP Path>
    - **Token**: <Databricks Access Token>
5. Click **Connect**.

---

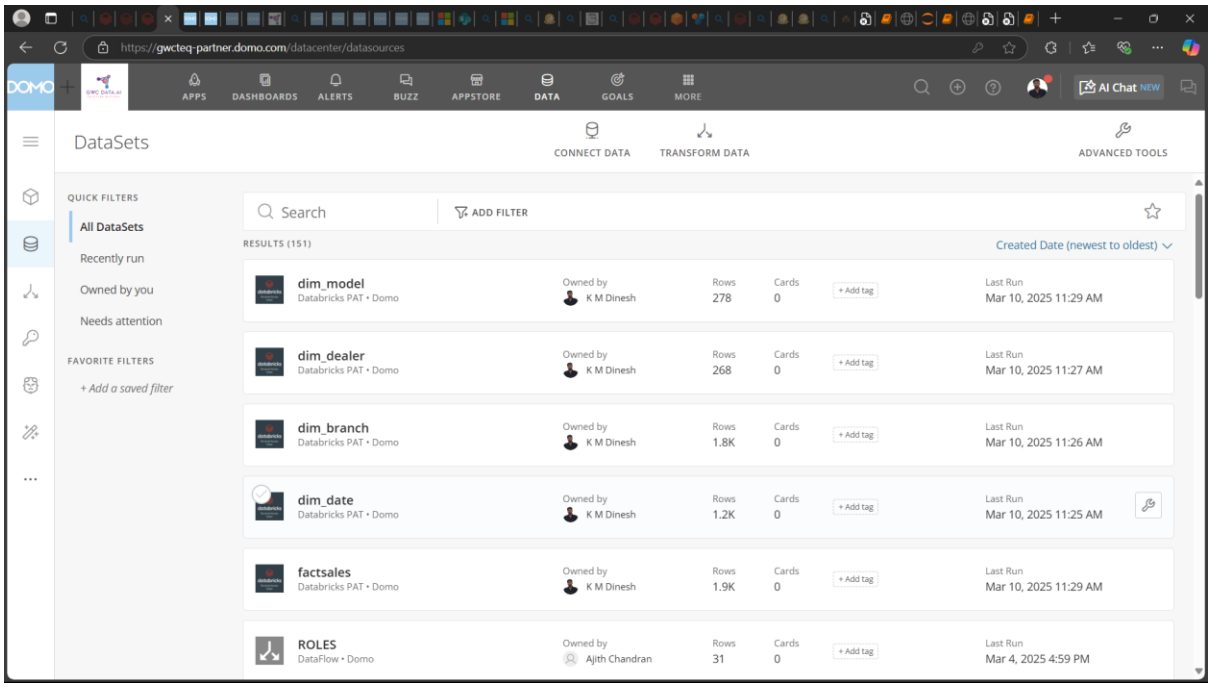
4. Configure Data Import

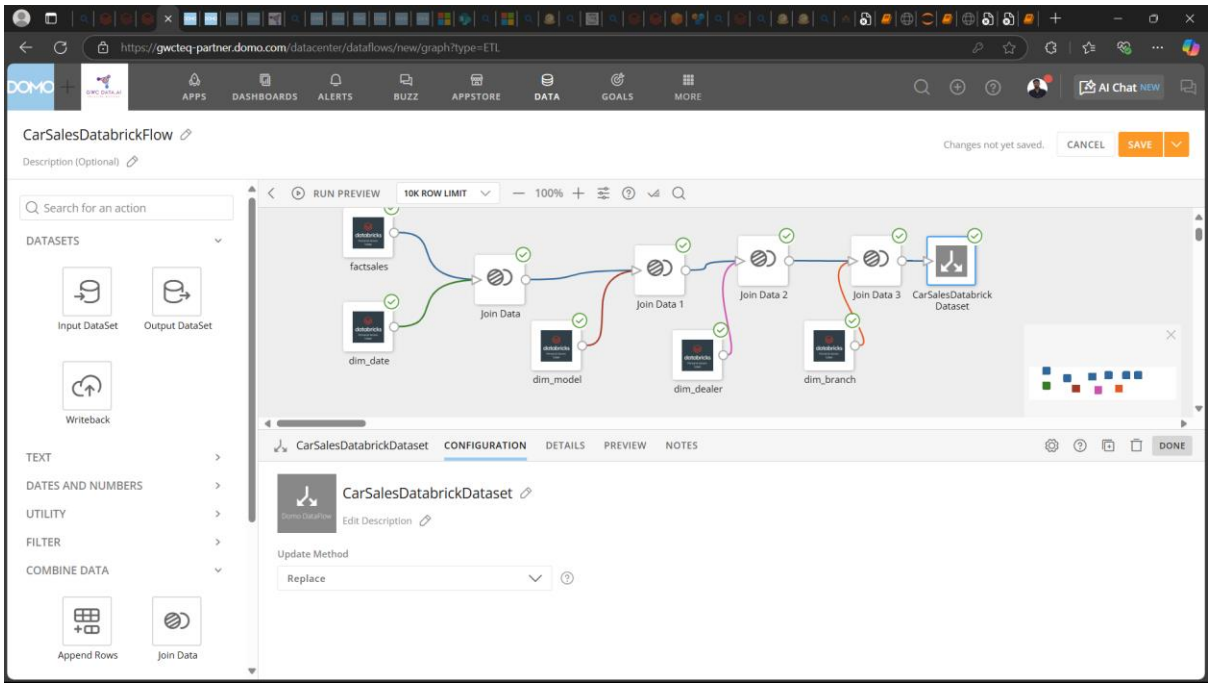
- Choose the **Databricks database** and **table** you want to pull into Domo.
- Set up **Query Mode** (Predefined Query or Custom SQL).
- Define a **Schedule** for data updates.
- Click **Save & Run**.

---

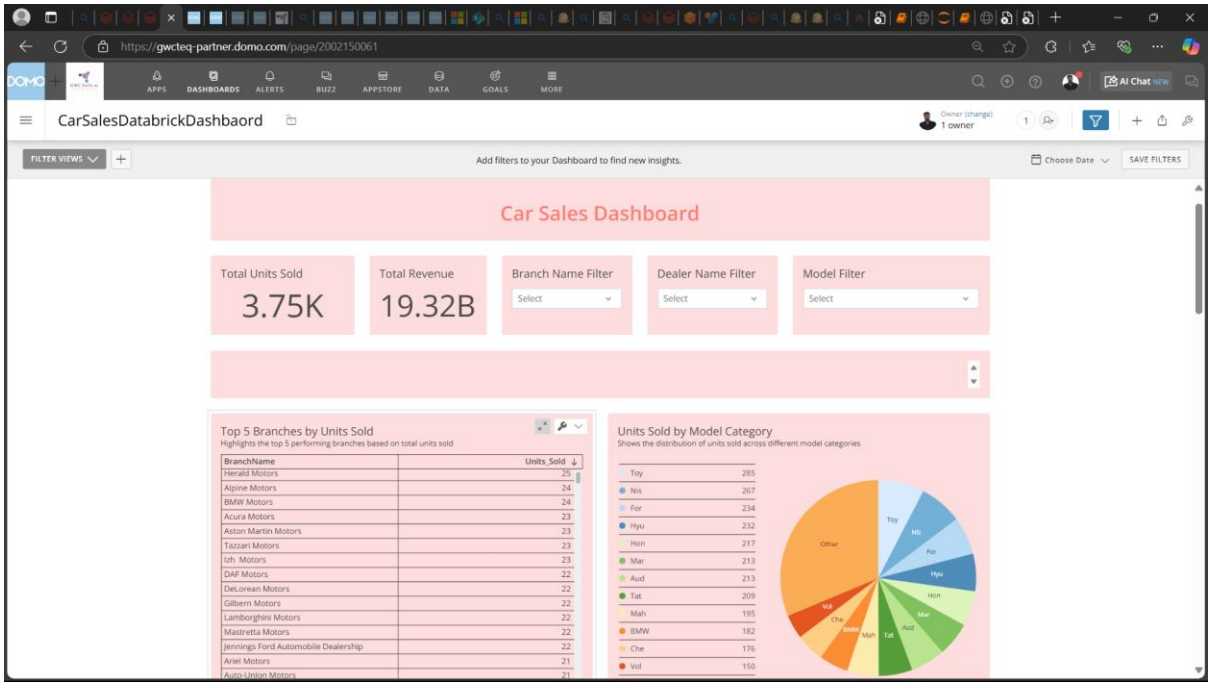
5. Verify and Use the Data

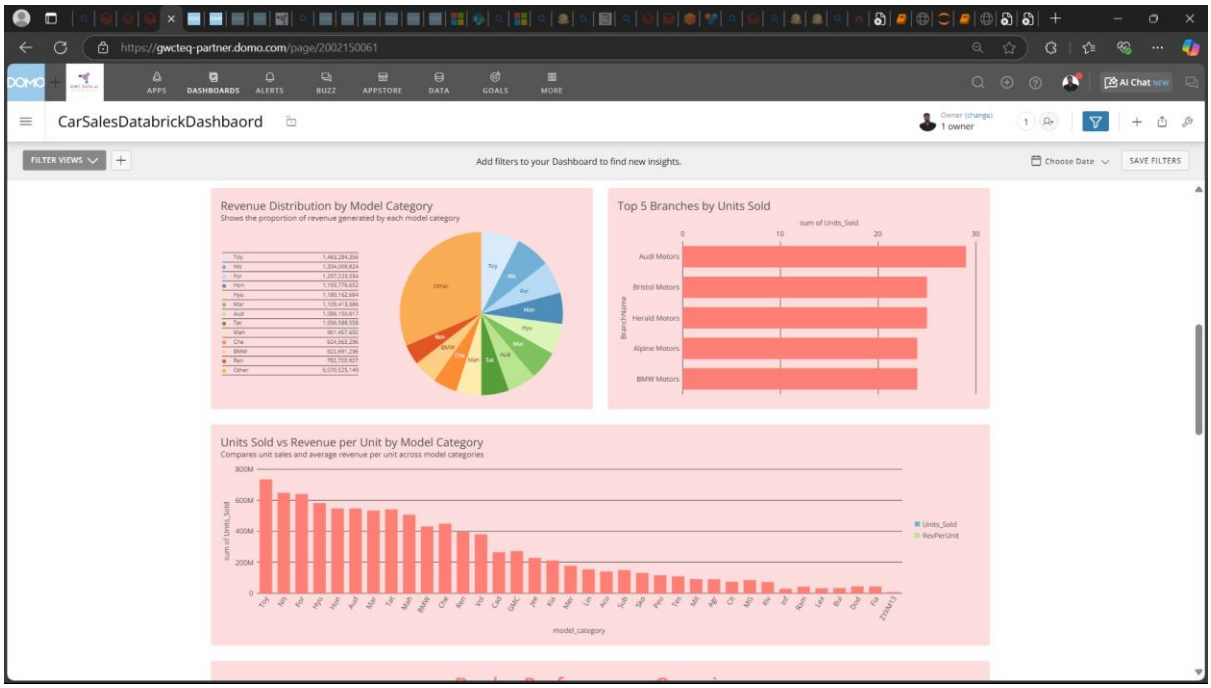
- Once the data is imported, it appears as a **Domo dataset**.
- Use it in **Domo Analyzer, Cards, or Dashboards**.

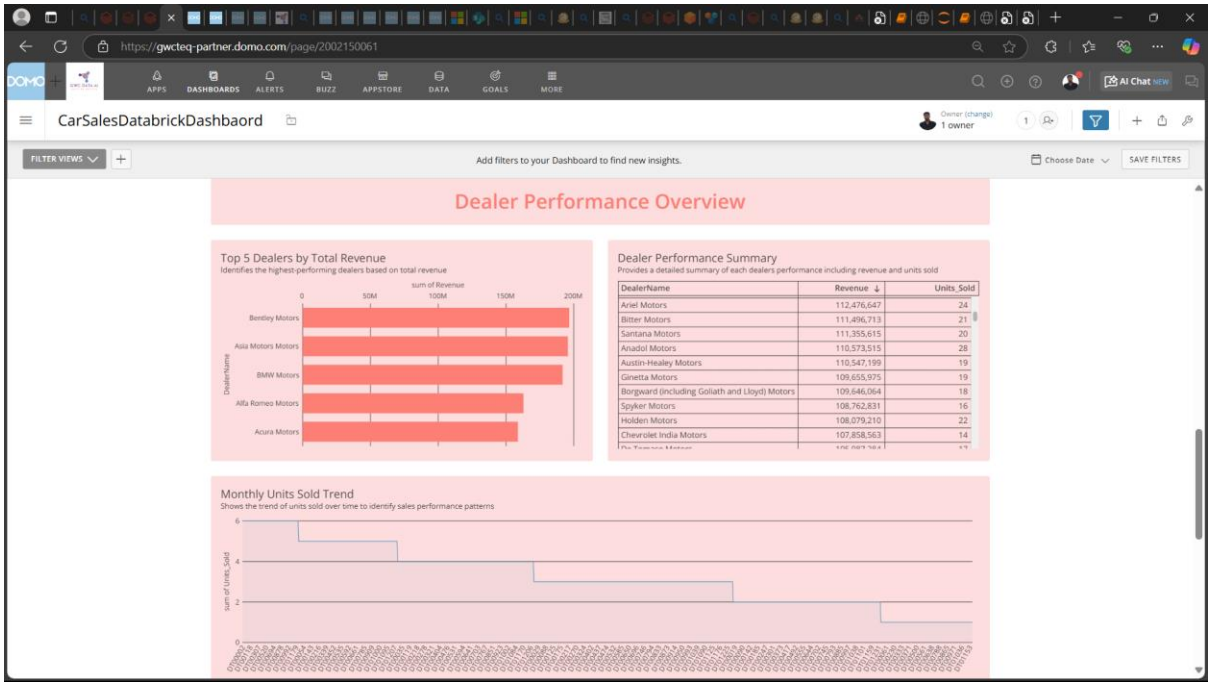

https://gwcteq-partner.domo.com/datacenter/datasources
APPS
DASHBOARDS
ALERTS
BUZZ
APPSTORE
DATA
GOALS
MORE
AI Chat NEW
DataSets
CONNECT DATA
TRANSFORM DATA
ADVANCED TOOLS
QUICK FILTERS
All DataSets
Recently run
Owned by you
Needs attention
FAVORITE FILTERS
+ Add a saved filter
Search
ADD FILTER
RESULTS (151)
Created Date (newest to oldest)
dim_model
Databricks PAT • Domo
Owned by
K M Dinesh
Rows
278
Cards
0
+ Add tag
Last Run
Mar 10, 2025 11:29 AM
dim_dealer
Databricks PAT • Domo
Owned by
K M Dinesh
Rows
268
Cards
0
+ Add tag
Last Run
Mar 10, 2025 11:27 AM
dim_branch
Databricks PAT • Domo
Owned by
K M Dinesh
Rows
1.8K
Cards
0
+ Add tag
Last Run
Mar 10, 2025 11:26 AM
dim_date
Databricks PAT • Domo
Owned by
K M Dinesh
Rows
1.2K
Cards
0
+ Add tag
Last Run
Mar 10, 2025 11:25 AM
factsales
Databricks PAT • Domo
Owned by
K M Dinesh
Rows
1.9K
Cards
0
+ Add tag
Last Run
Mar 10, 2025 11:29 AM
ROLES
DataFlow • Domo
Owned by
Ajith Chandran
Rows
31
Cards
0
+ Add tag
Last Run
Mar 4, 2025 4:59 PM

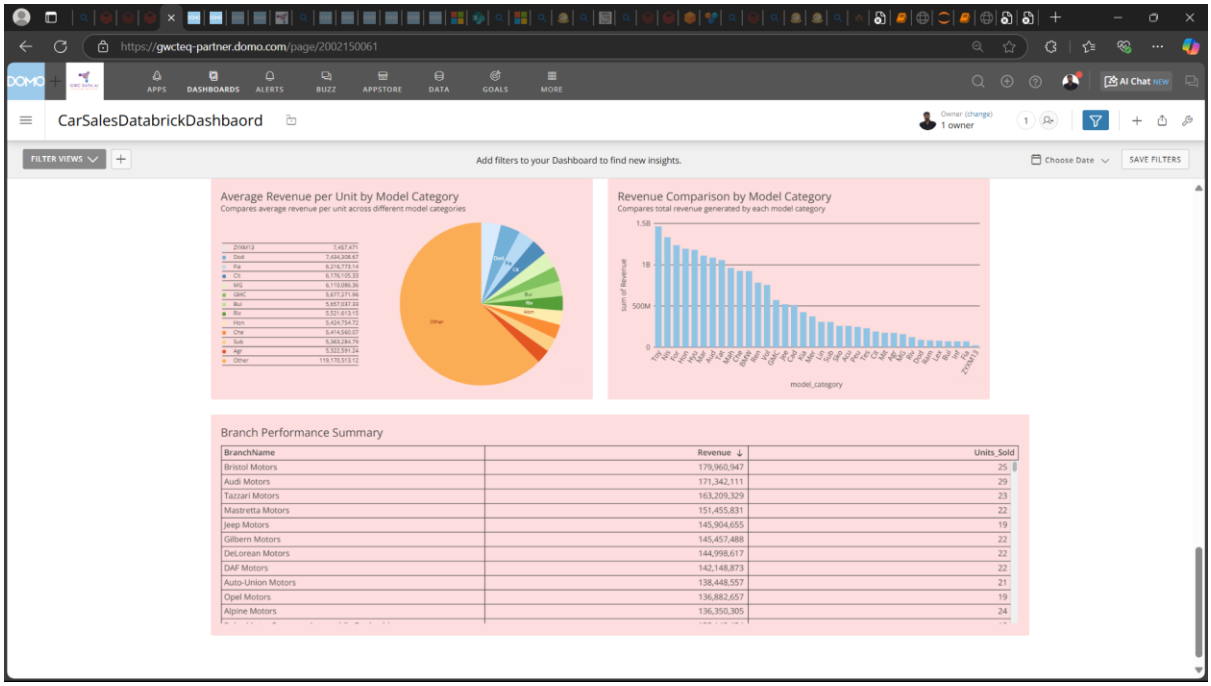

https://gwcteq-partner.domo.com/datacenter/dataflows/new/graph?type=ETL
APPS
DASHBOARDS
ALERTS
BUZZ
APPSTORE
DATA
GOALS
MORE
AI Chat NEW
CarSalesDatabrickFlow
Description (Optional)
Changes not yet saved.
CANCEL
SAVE
Search for an action
DATASETS
Input DataSet
Output DataSet
Writeback
TEXT
DATES AND NUMBERS
UTILITY
FILTER
COMBINE DATA
Append Rows
Join Data
RUN PREVIEW
10K ROW LIMIT
100%
factsales
dim_date
Join Data
dim_model
Join Data 1
dim_dealer
Join Data 2
dim_branch
Join Data 3
CarSalesDatabrick Dataset
CarSalesDatabrickDataset
CONFIGURATION
DETAILS
PREVIEW
NOTES
DONE
CarSalesDatabrickDataset
Edit Description
Update Method
Replace

Dashboard


https://gwcteq-partner.domo.com/page/2002150061
APPS
DASHBOARDS
ALERTS
BUZZ
APPSTORE
DATA
GOALS
MORE
AI Chat
CarSalesDatabrickDashbaord
Owner (change)
1 owner
FILTER VIEWS
Add filters to your Dashboard to find new insights.
Choose Date
SAVE FILTERS
Car Sales Dashboard
Total Units Sold
3.75K
Total Revenue
19.32B
Branch Name Filter
Select
Dealer Name Filter
Select
Model Filter
Select
Top 5 Branches by Units Sold
Highlights the top 5 performing branches based on total units sold
BranchName
Units_Sold
Herald Motors 25
Alpine Motors 24
BMW Motors 24
Acura Motors 23
Aston Martin Motors 23
Tazzari Motors 23
Izh Motors 23
DAF Motors 22
DeLorean Motors 22
Gilbern Motors 22
Lamborghini Motors 22
Mastretta Motors 22
Jennings Ford Automobile Dealership 22
Ariel Motors 21
Auto-Union Motors 21
Units Sold by Model Category
Shows the distribution of units sold across different model categories
Toy 285
Nis 267
For 234
Hyu 232
Hon 217
Mar 213
Aud 213
Tat 209
Mah 195
BMW 182
Che 176
Vol 150
Other


https://gwcteq-partner.domo.com/page/2002150061
APPS
DASHBOARDS
ALERTS
BUZZ
APPSTORE
DATA
GOALS
MORE
AI Chat
CarSalesDatabrickDashbaord
Owner (change)
1 owner
FILTER VIEWS
Add filters to your Dashboard to find new insights.
Choose Date
SAVE FILTERS
Revenue Distribution by Model Category
Shows the proportion of revenue generated by each model category
Top 5 Branches by Units Sold
sum of Units_Sold
Audi Motors
Bristol Motors
Herald Motors
Alpine Motors
BMW Motors
BranchName
Units Sold vs Revenue per Unit by Model Category
Compares unit sales and average revenue per unit across model categories
sum of Units_Sold
model_category
Units_Sold
RevPerUnit


https://gwcteq-partner.domo.com/page/2002150061
APPS
DASHBOARDS
ALERTS
BUZZ
APPSTORE
DATA
GOALS
MORE
AI Chat NEW
CarSalesDatabrickDashbaord
Owner (change)
1 owner
FILTER VIEWS
Add filters to your Dashboard to find new insights.
Choose Date
SAVE FILTERS
Dealer Performance Overview
Top 5 Dealers by Total Revenue
Identifies the highest-performing dealers based on total revenue
sum of Revenue
0
50M
100M
150M
200M
DealerName
Bentley Motors
Asia Motors Motors
BMW Motors
Alfa Romeo Motors
Acura Motors
Dealer Performance Summary
Provides a detailed summary of each dealers performance including revenue and units sold
DealerName
Revenue
Units_Sold
Ariel Motors 112,476,647 24
Bitter Motors 111,496,713 21
Santana Motors 111,355,615 20
Anadol Motors 110,573,515 28
Austin-Healey Motors 110,547,199 19
Ginetta Motors 109,655,975 19
Borgward (including Goliath and Lloyd) Motors 109,646,064 18
Spyker Motors 108,762,831 16
Holden Motors 108,079,210 22
Chevrolet India Motors 107,858,563 14
Monthly Units Sold Trend
Shows the trend of units sold over time to identify sales performance patterns
sum of Units_Sold
6
4
2
0


https://gwcteq-partner.domo.com/page/2002150061
APPS
DASHBOARDS
ALERTS
BUZZ
APPSTORE
DATA
GOALS
MORE
AI Chat NEW
CarSalesDatabrickDashbaord
Owner (change)
1 owner
FILTER VIEWS
Add filters to your Dashboard to find new insights.
Choose Date
SAVE FILTERS
Average Revenue per Unit by Model Category
Compares average revenue per unit across different model categories
ZYAM13 7,437,471
Dod 7,434,308.67
Fia 6,218,773.14
Cit 6,176,105.33
MG 6,110,086.36
GMC 5,877,271.96
Bui 5,657,037.33
Riv 5,521,613.15
Hon 5,424,754.72
Che 5,414,560.07
Sub 5,363,284.79
Agr 5,322,591.24
Other 119,170,513.12
Other
Revenue Comparison by Model Category
Compares total revenue generated by each model category
sum of Revenue
1.5B
1B
500M
0
model_category
Branch Performance Summary
BranchName
Revenue
Units_Sold
Bristol Motors 179,960,947 25
Audi Motors 171,342,111 29
Tazzari Motors 163,209,329 23
Mastretta Motors 151,455,831 22
Jeep Motors 145,904,655 19
Gilbern Motors 145,457,488 22
DeLorean Motors 144,998,617 22
DAF Motors 142,148,873 22
Auto-Union Motors 138,448,557 21
Opel Motors 136,882,657 19
Alpine Motors 136,350,305 24