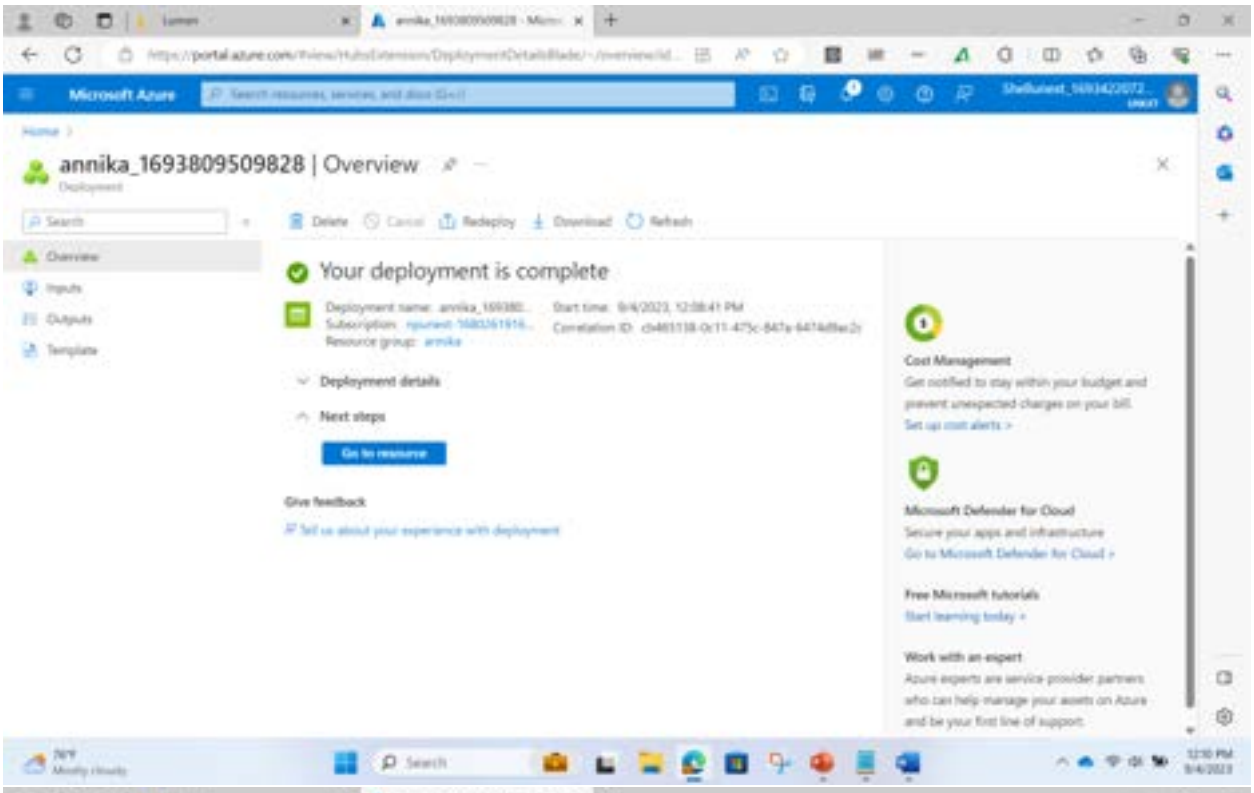
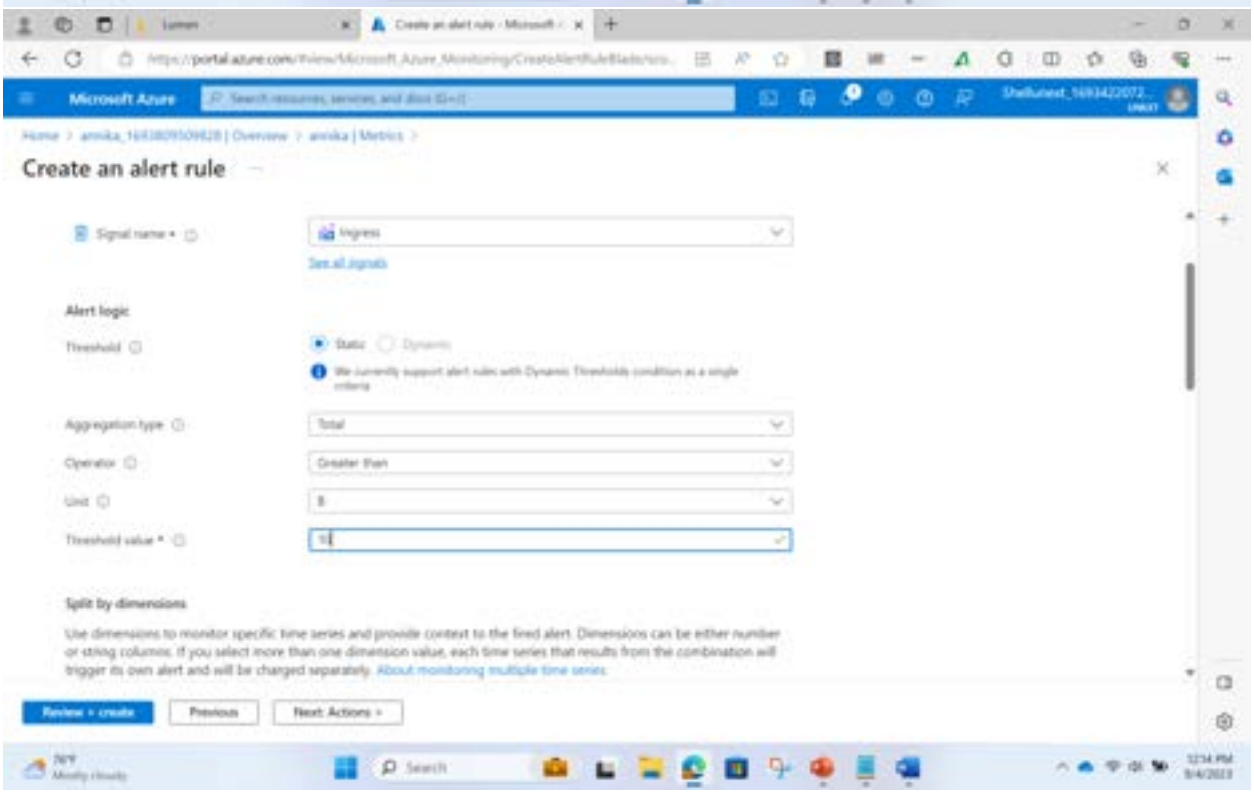
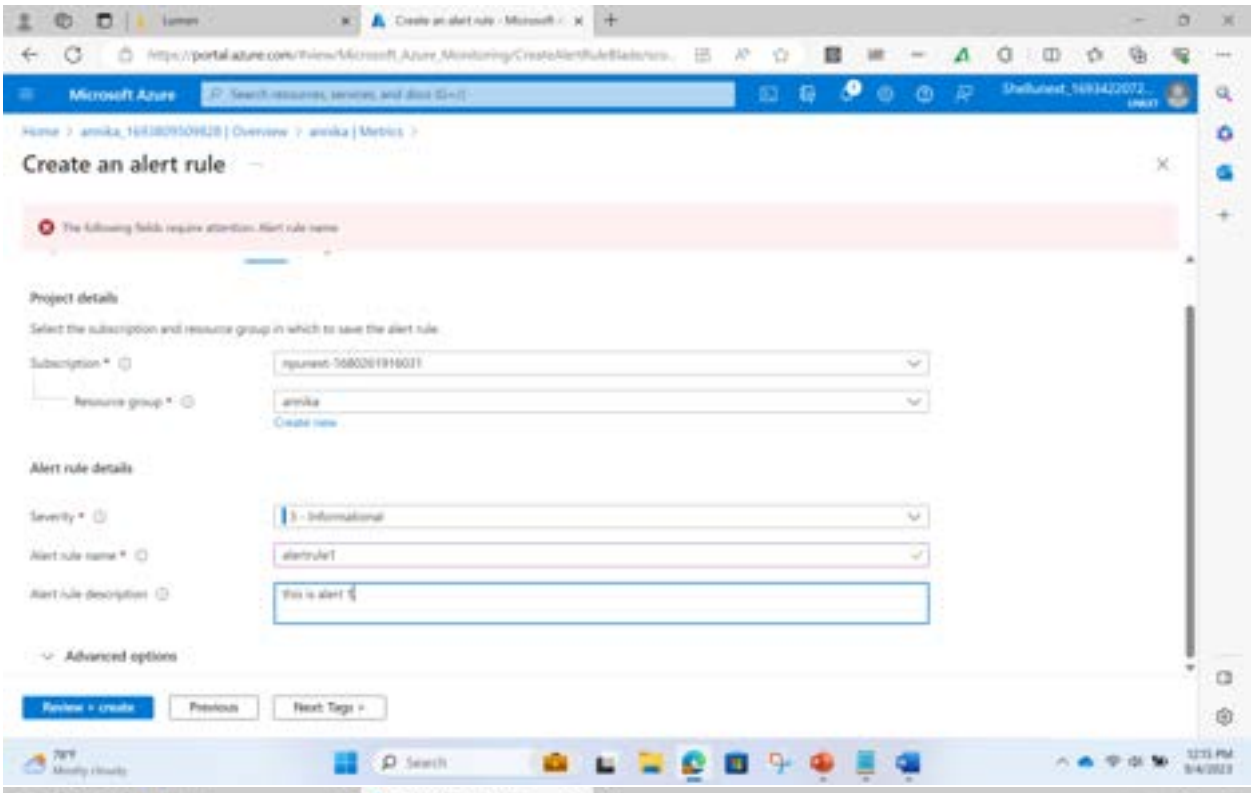
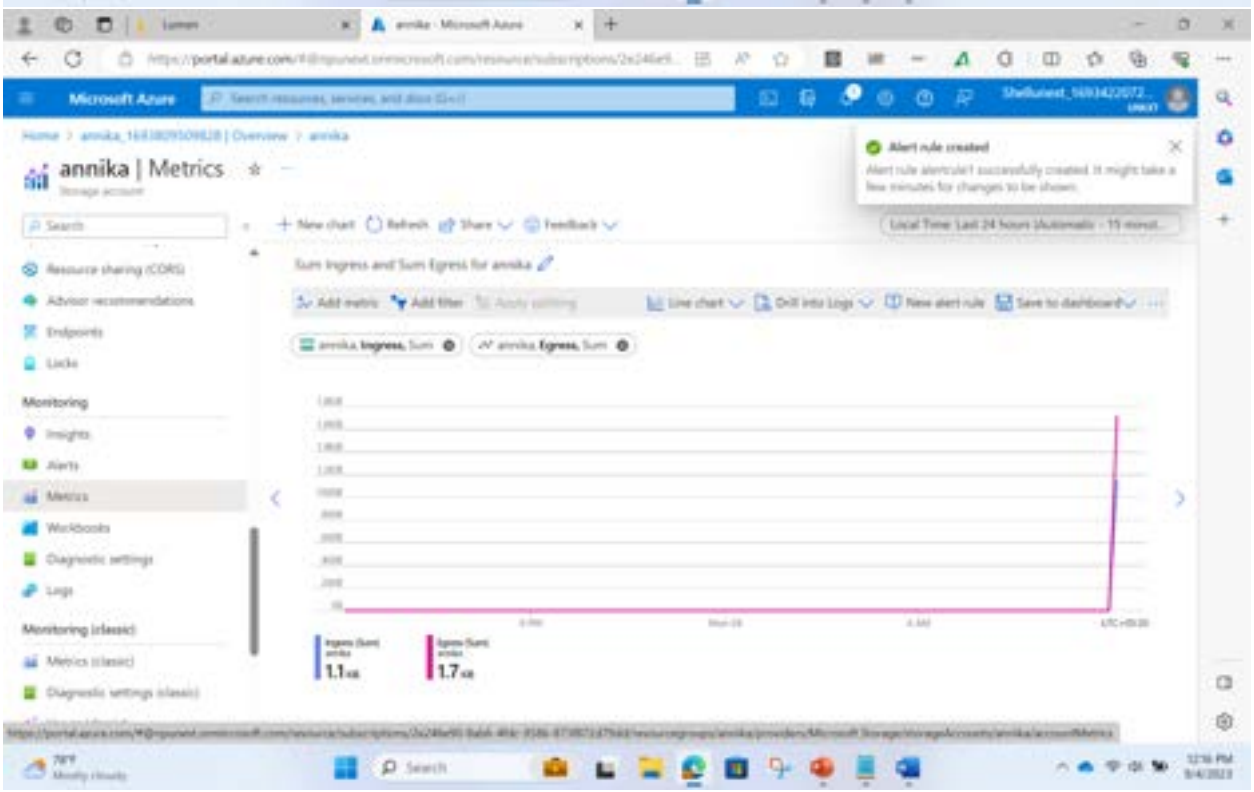
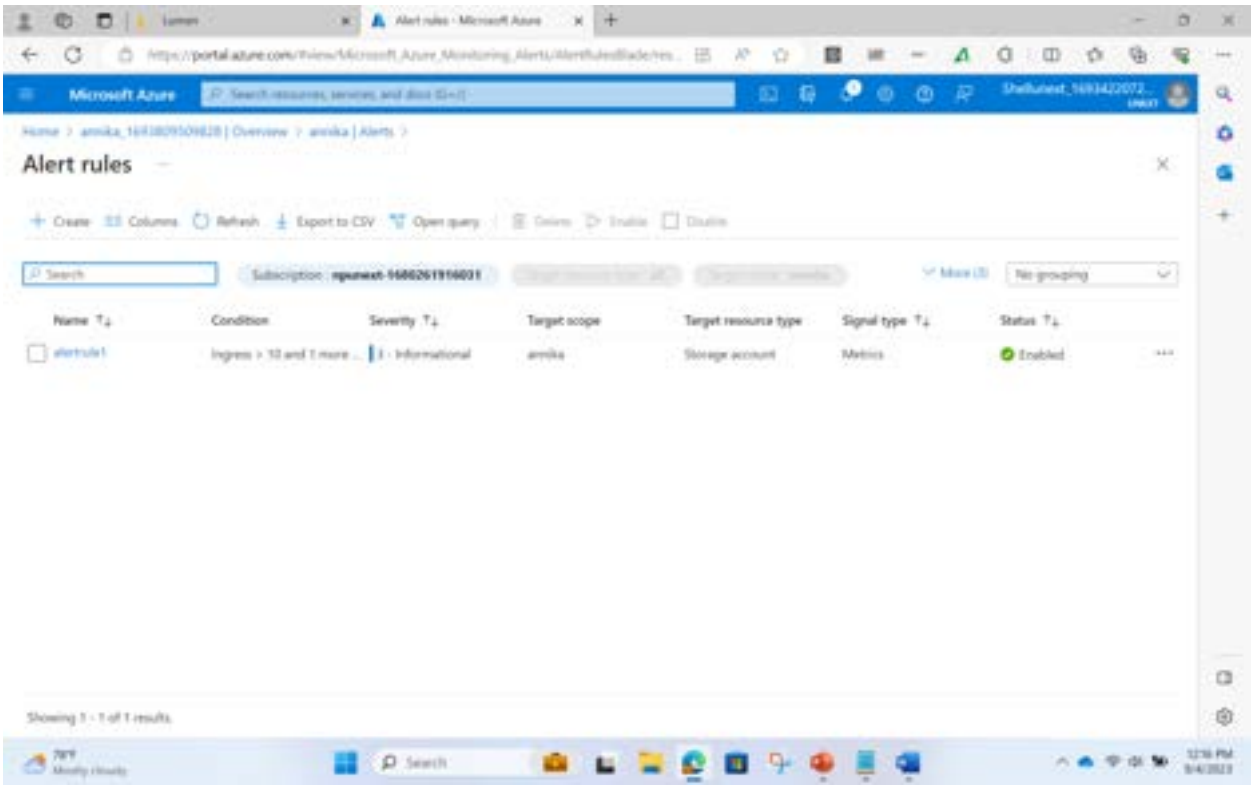
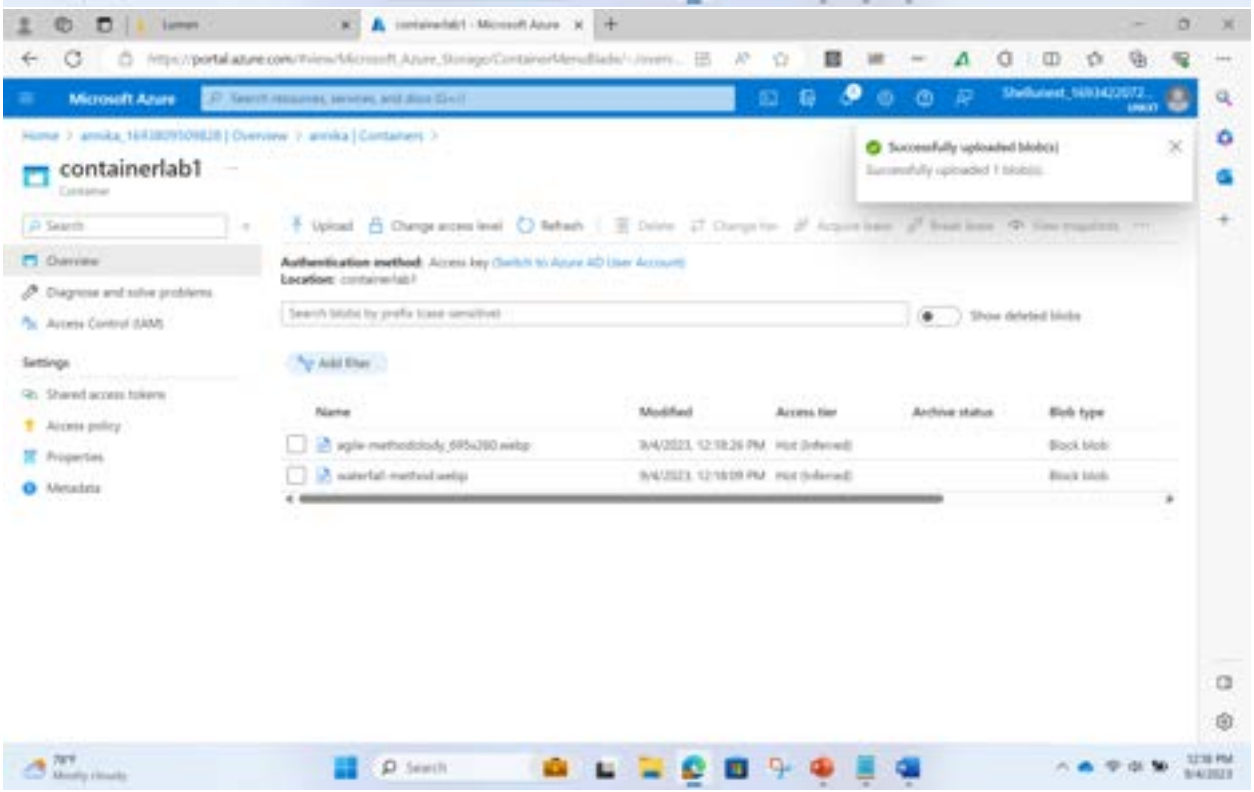

Microsoft Azure
annika_1693809509828 | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Your deployment is complete
Deployment name: annika_169380...
Subscription: ngunest-16803619...
Resource group: annika
Start time: 9/4/2023, 12:08:41 PM
Correlation ID: cb481138-0c11-475c-847a-6474d8ec2c
Deployment details
Next steps
Go to resource
Give feedback
Tell us about your experience with deployment
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.

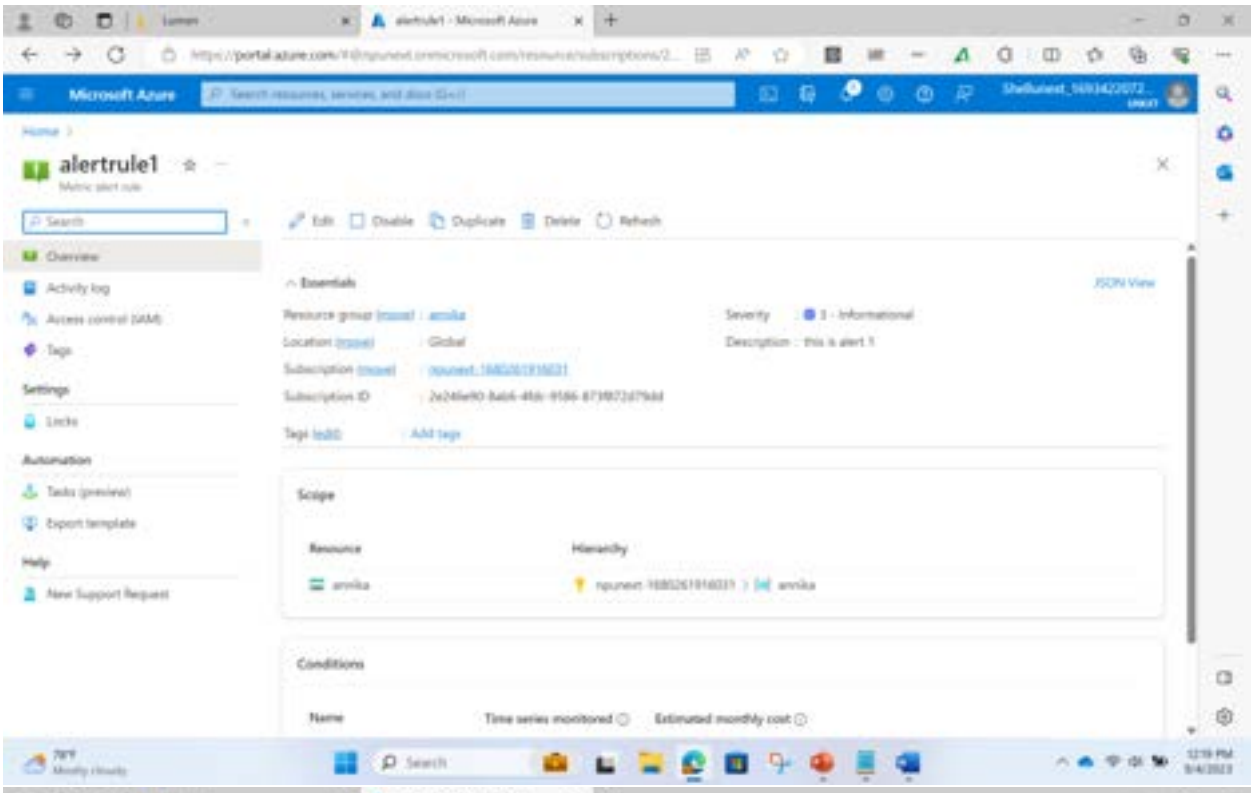

Microsoft Azure
Home > annika_1693809509828 | Overview > annika | Metrics >
Create an alert rule
Signal name *
Ingress
See all signals
Alert logic
Threshold
Static
Dynamic
Aggregation type
Total
Operator
Greater than
Unit
1
Threshold value *
10
Split by dimensions
Use dimensions to monitor specific time series and provide context to the fired alert. Dimensions can be either number or string columns. If you select more than one dimension value, each time series that results from the combination will trigger its own alert and will be charged separately. About monitoring multiple time series
Review + create
Previous
Next: Actions >

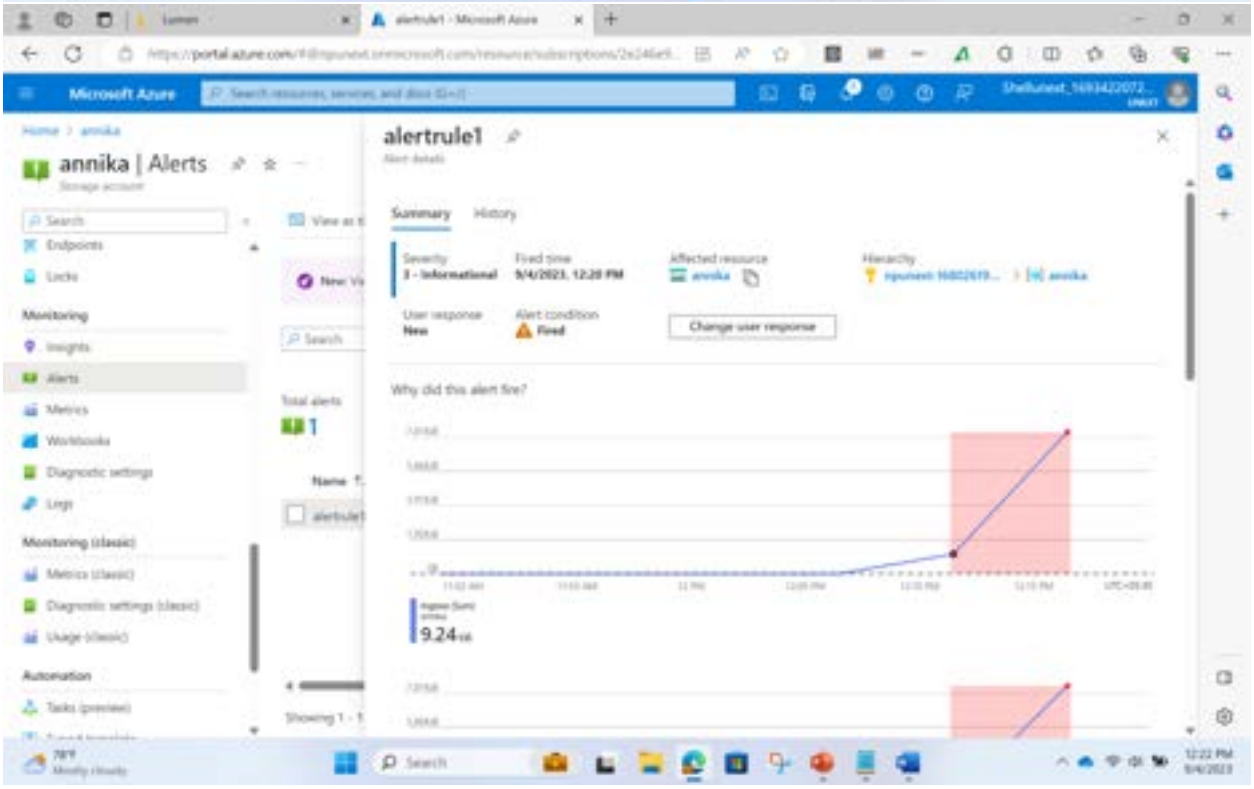

Microsoft Azure
Create an alert rule
Project details
Select the subscription and resource group in which to save the alert rule.
Subscription *
Resource group *
annika
Create new
Alert rule details
Severity *
3 - Informational
Alert rule name *
alertrule1
Alert rule description
This is alert 1
Advanced options
Review + create
Previous
Next: Tags >

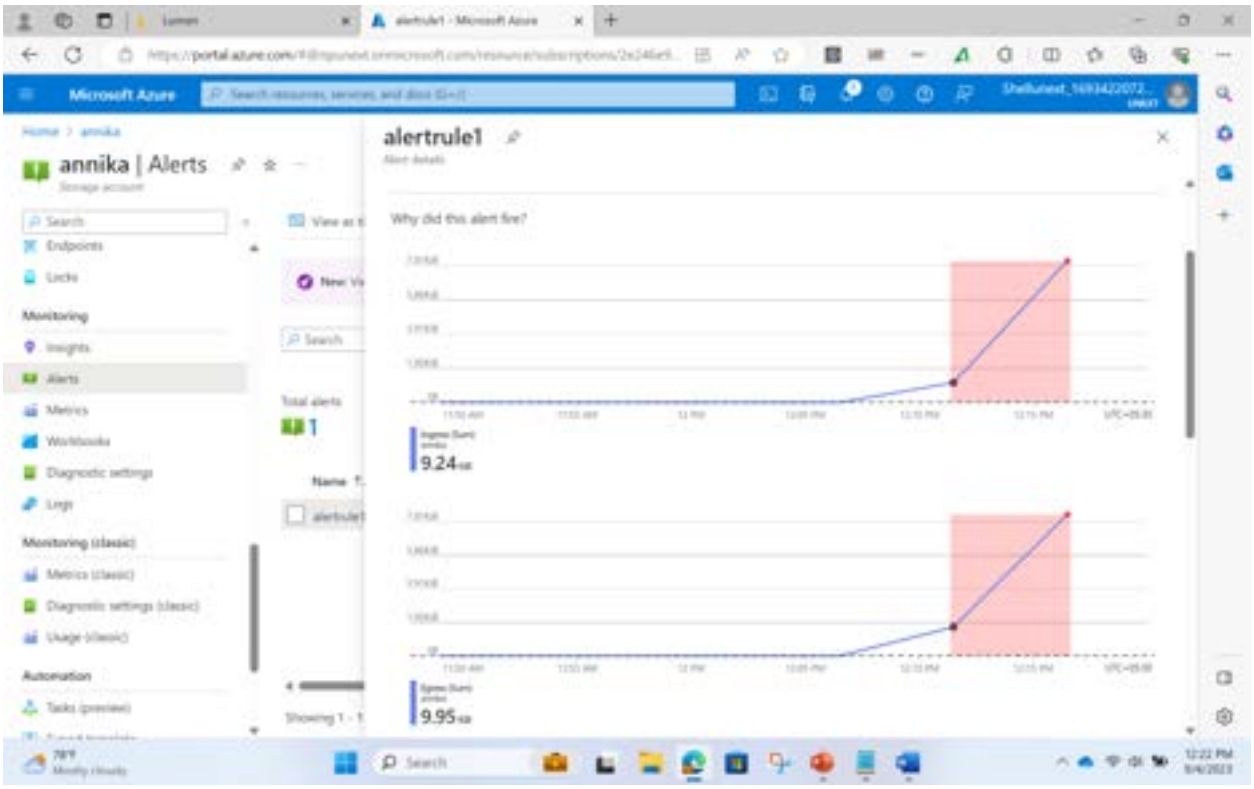

Microsoft Azure
annika | Metrics
Storage account
Alert rule created
Alert rule alertrule1 successfully created. It might take a few minutes for changes to be shown.
New chart
Refresh
Share
Feedback
Sum Ingress and Sum Egress for annika
Add metric
Add filter
Line chart
Drill into Logs
New alert rule
Save to dashboard
annika, Ingress, Sum
annika, Egress, Sum
Resource sharing (CORS)
Advisor recommendations
Endpoints
Locks
Monitoring
Insights
Alerts
Metrics
Workbooks
Diagnostic settings
Logs
Monitoring (classic)
Metrics (classic)
Diagnostic settings (classic)
1.1 KB
1.7 KB

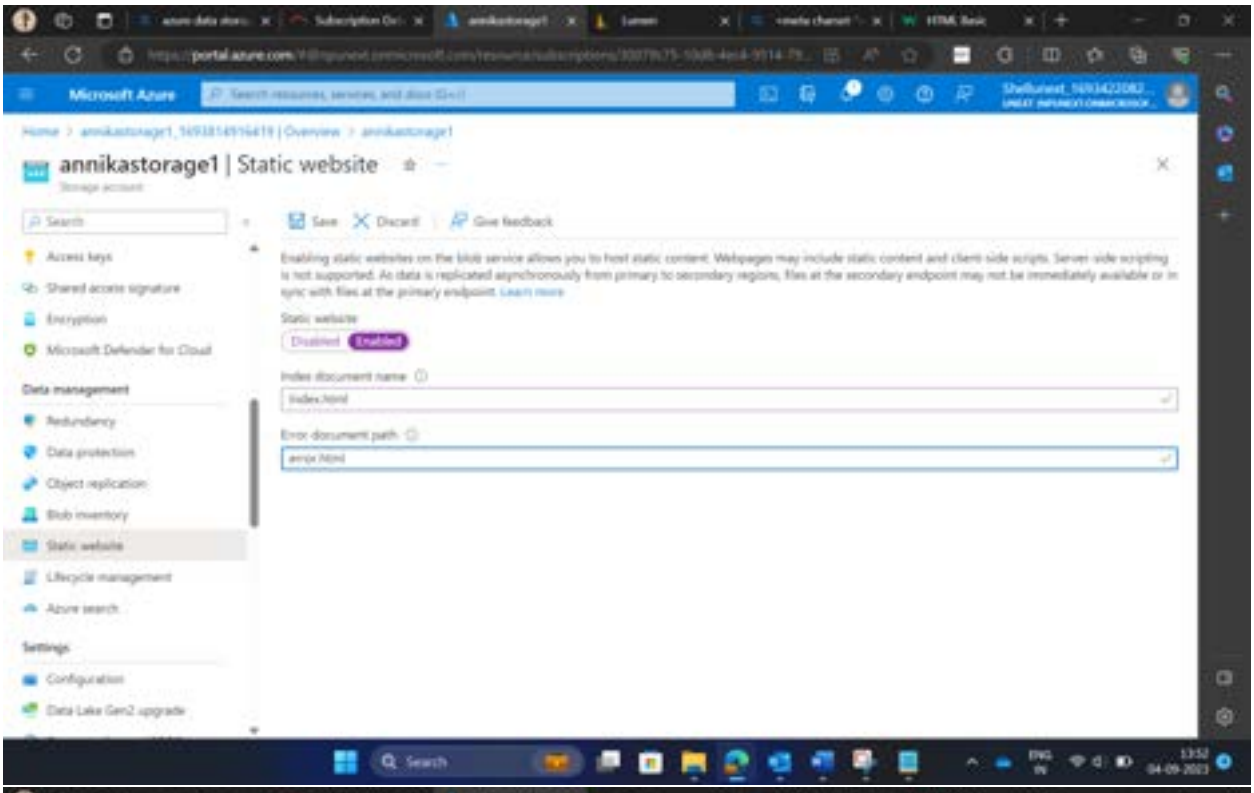

Microsoft Azure
Alert rules
Create
Columns
Refresh
Export to CSV
Open query
Name
Condition
Severity
Target scope
Target resource type
Signal type
Status
alertrule1
Enabled

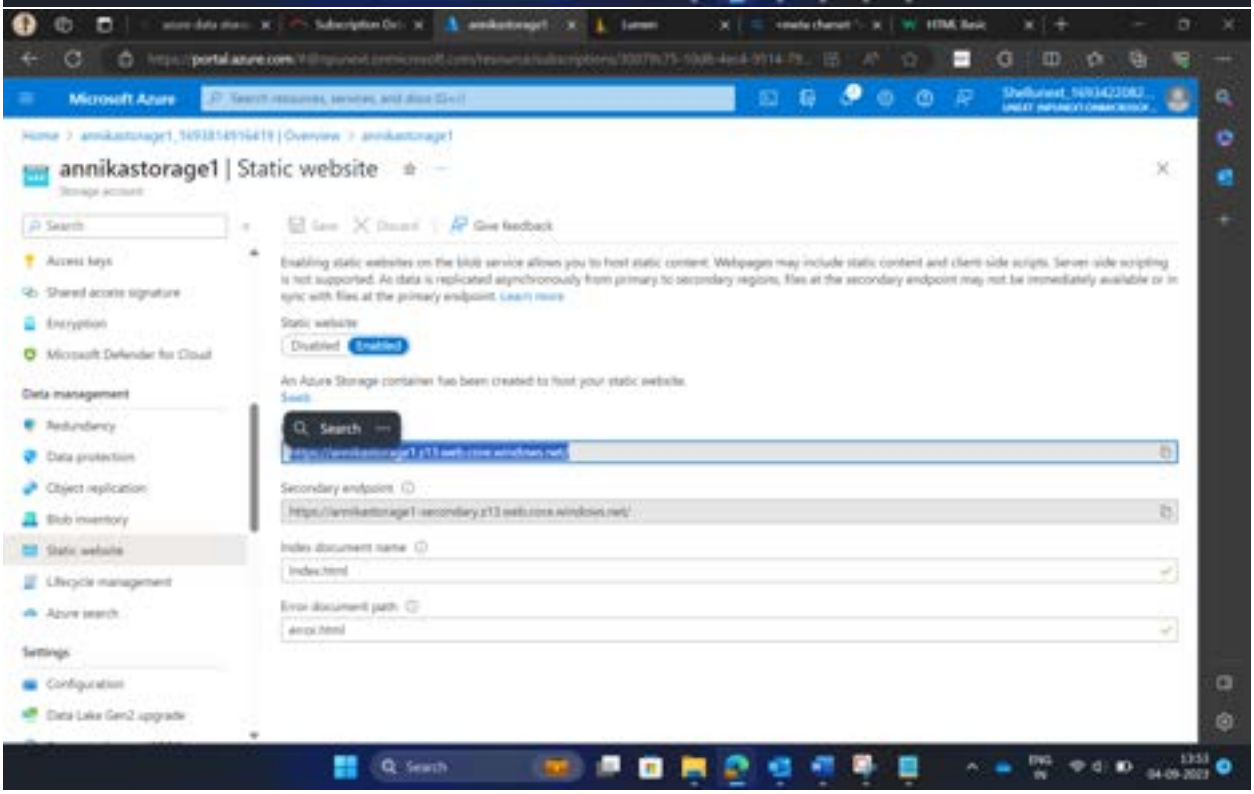

Microsoft Azure
containerlab1
Container
Successfully uploaded blob(s)
Upload
Change access level
Refresh
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Access policy
Properties
Metadata
Name
Modified
Access tier
Archive status
Blob type


Microsoft Azure
alertrule1
Edit
Disable
Duplicate
Delete
Refresh
Overview
Activity log
Access control (IAM)
Tags
Settings
Locks
Automation
Tasks (preview)
Export template
Help
New Support Request
Essentials
JSON View
Severity : 3 - Informational
Description : this is alert 1
Scope
Resource
Hierarchy
annika
Conditions
Name
Time series monitored
Estimated monthly cost

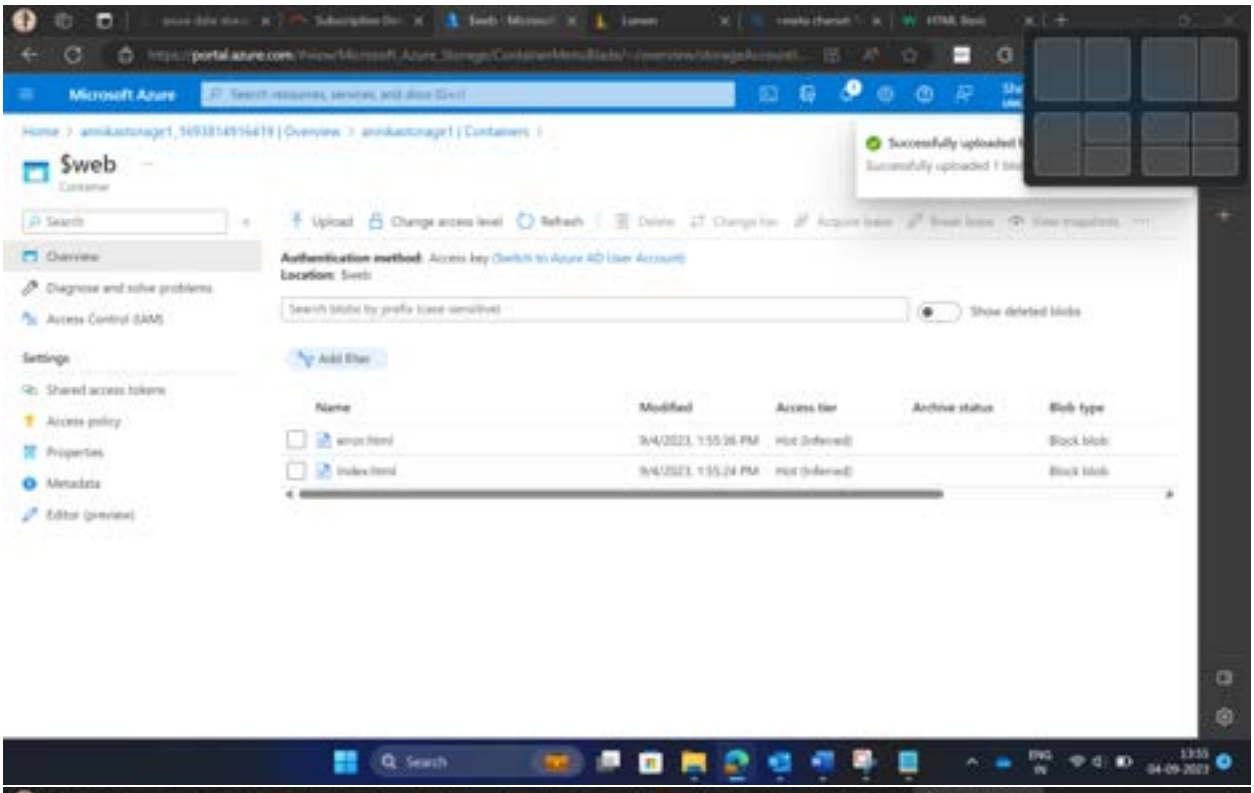

Microsoft Azure
annika | Alerts
alertrule1
Summary
History
Severity
3 - Informational
Fired time
9/4/2023, 12:20 PM
Affected resource
annika
Hierarchy
User response
New
Alert condition
Fired
Change user response
Why did this alert fire?
Monitoring
Insights
Alerts
Metrics
Workbooks
Diagnostic settings
Logs
Monitoring (classic)
Automation
Total alerts
1
9.24

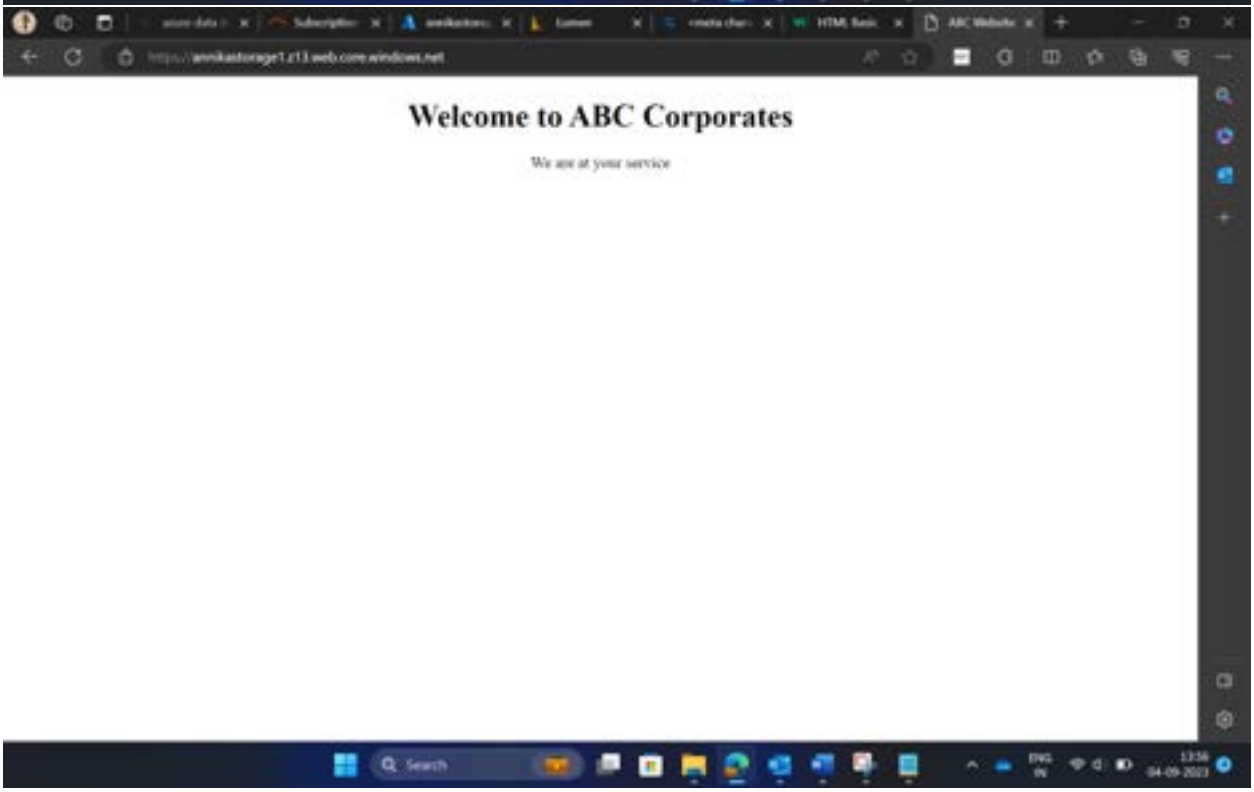

Microsoft Azure
annika | Alerts
Storage account
Search
Endpoints
Locks
Monitoring
Insights
Alerts
Metrics
Workbooks
Diagnostic settings
Logs
Monitoring (classic)
Metrics (classic)
Diagnostic settings (classic)
Usage (classic)
Automation
Tasks (preview)
Total alerts
1
alertrule1
Alert details
Why did this alert fire?
9.24
9.95

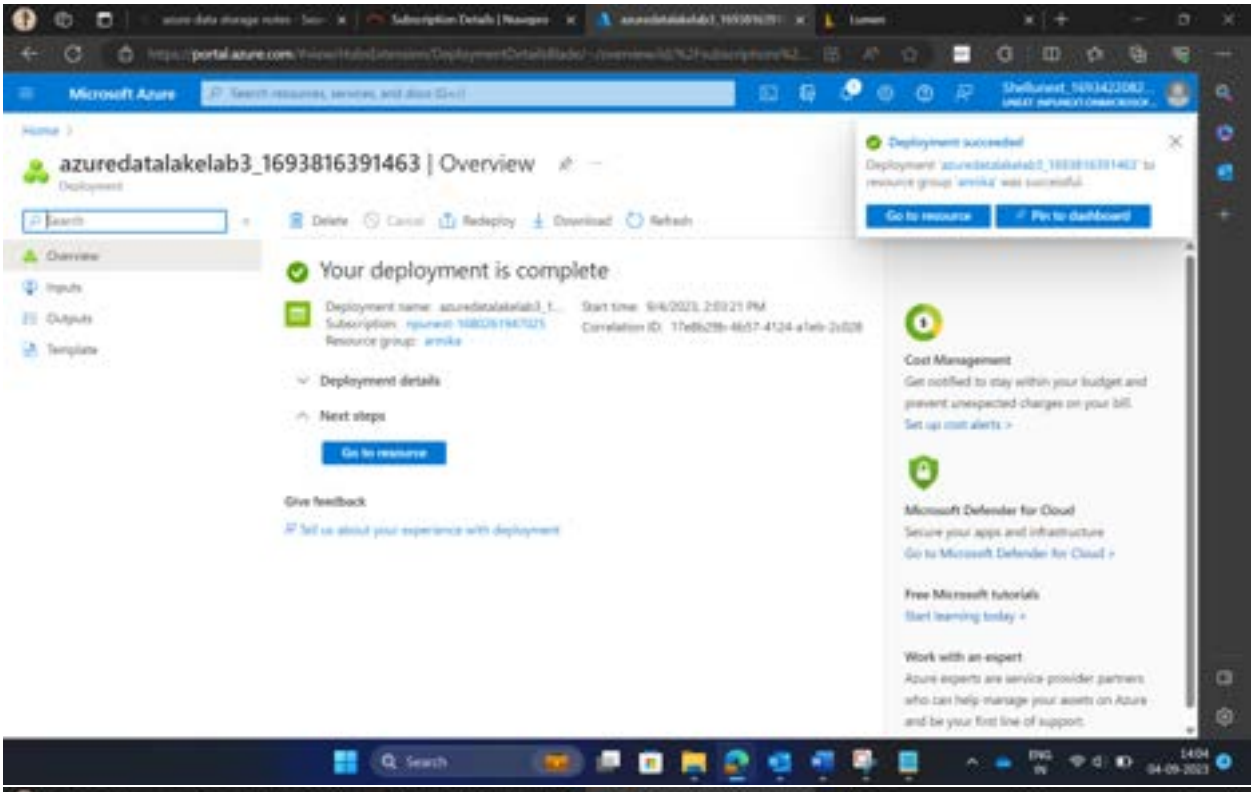


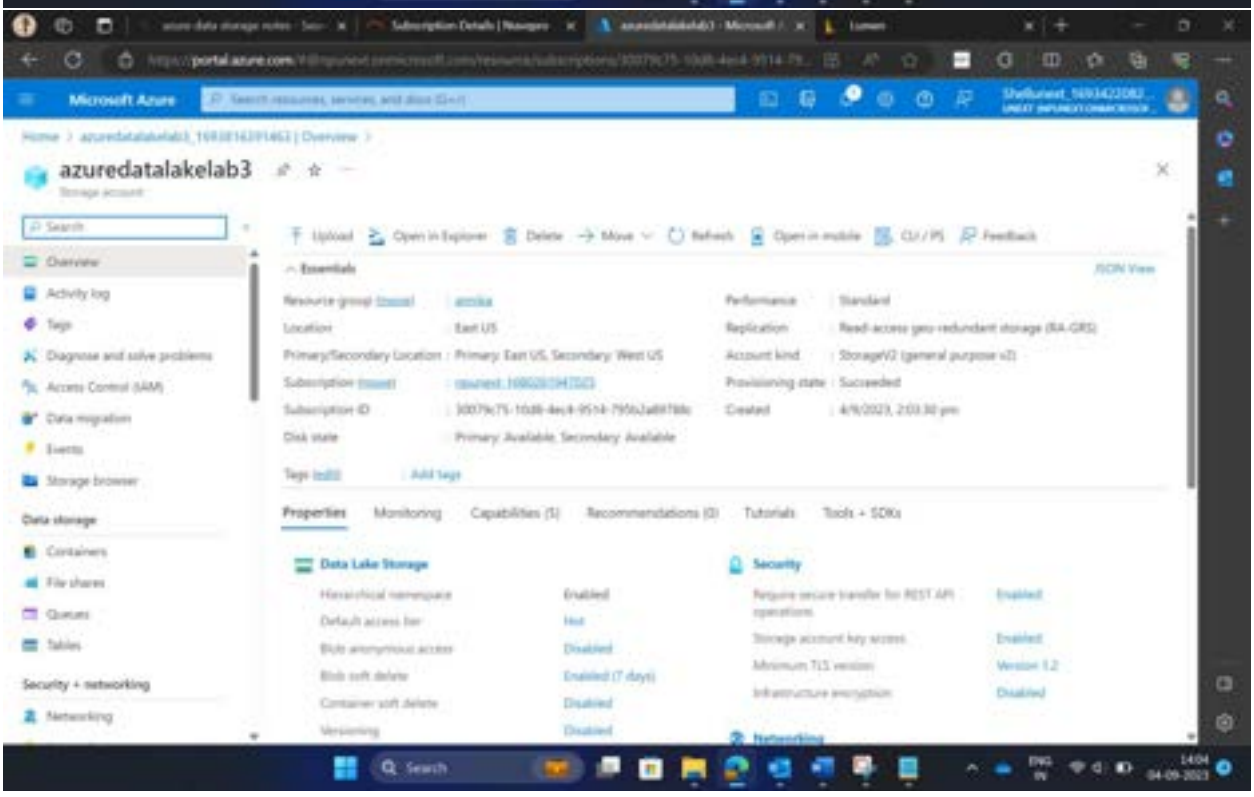


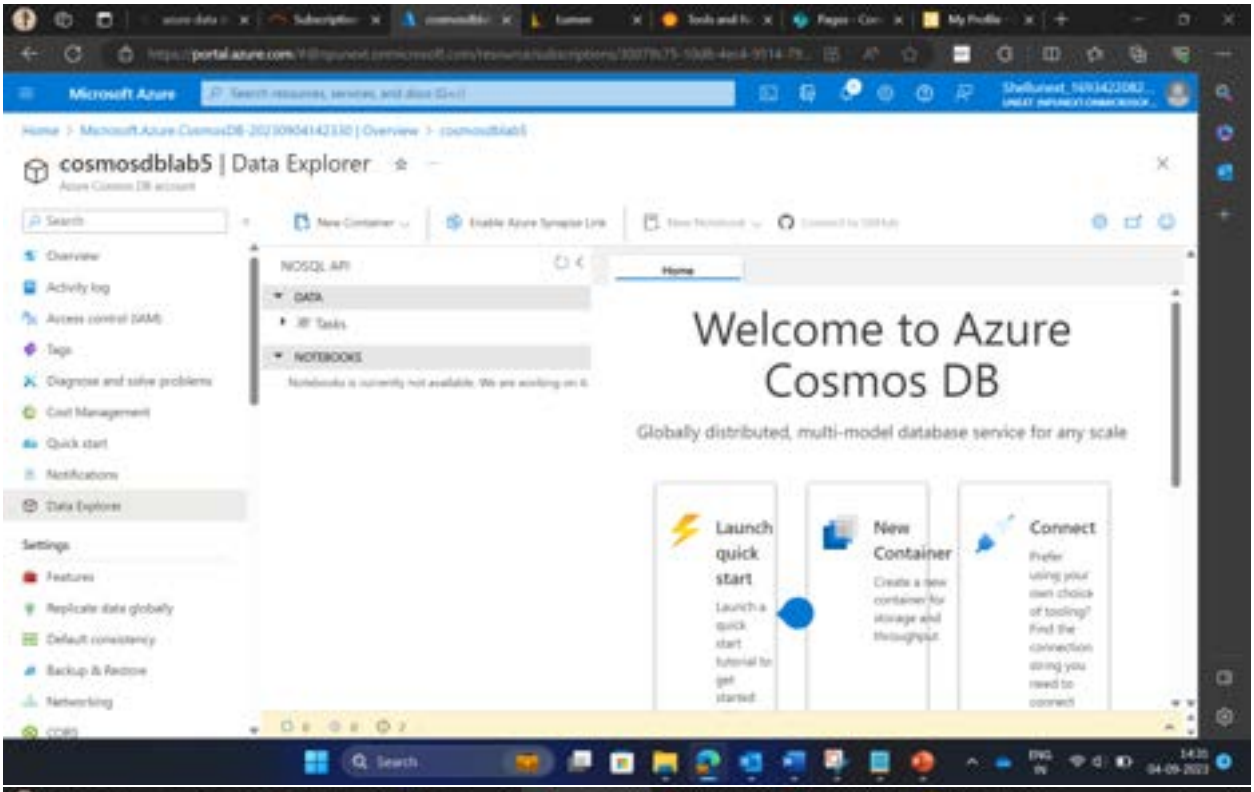
https://annikastorage1.z13.web.core.windows.net/

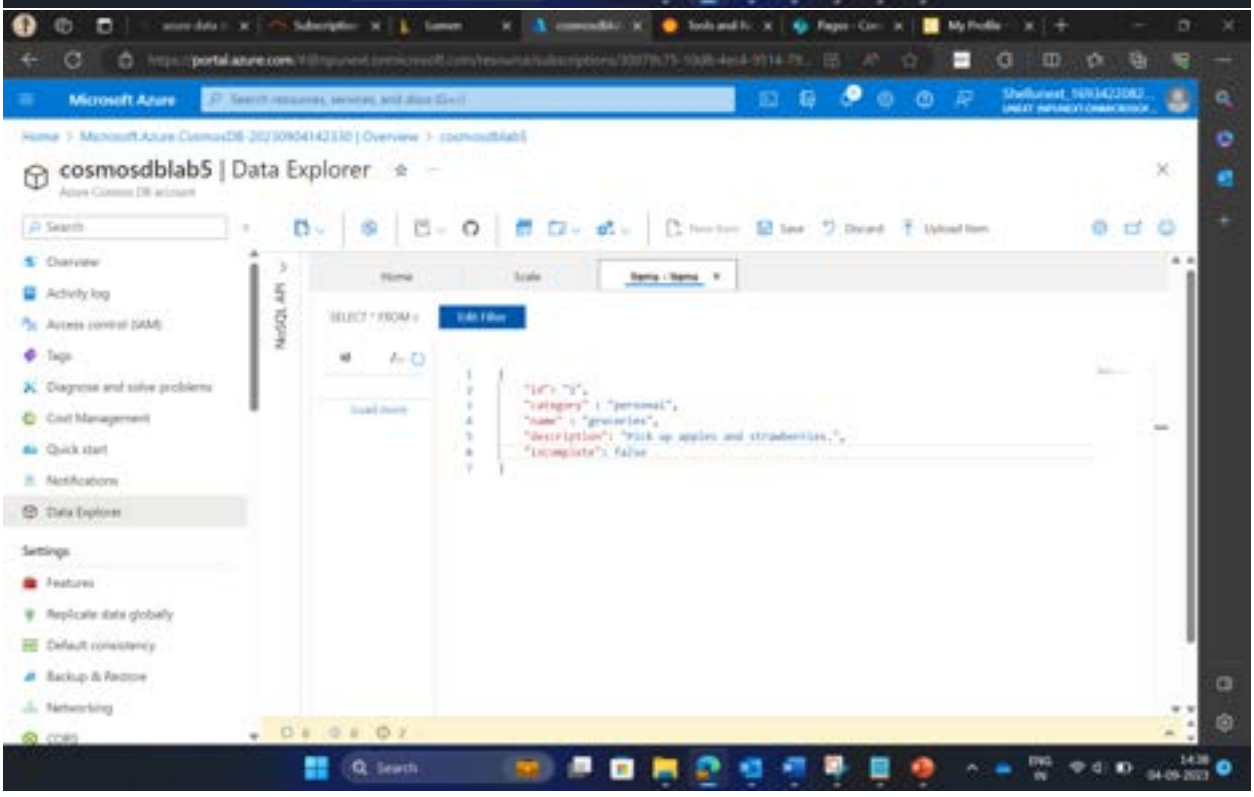

Microsoft Azure
$web
Upload
Change access level
Refresh
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Access policy
Properties
Metadata
Editor (preview)
Name
Modified
Access tier
Archive status
Blob type

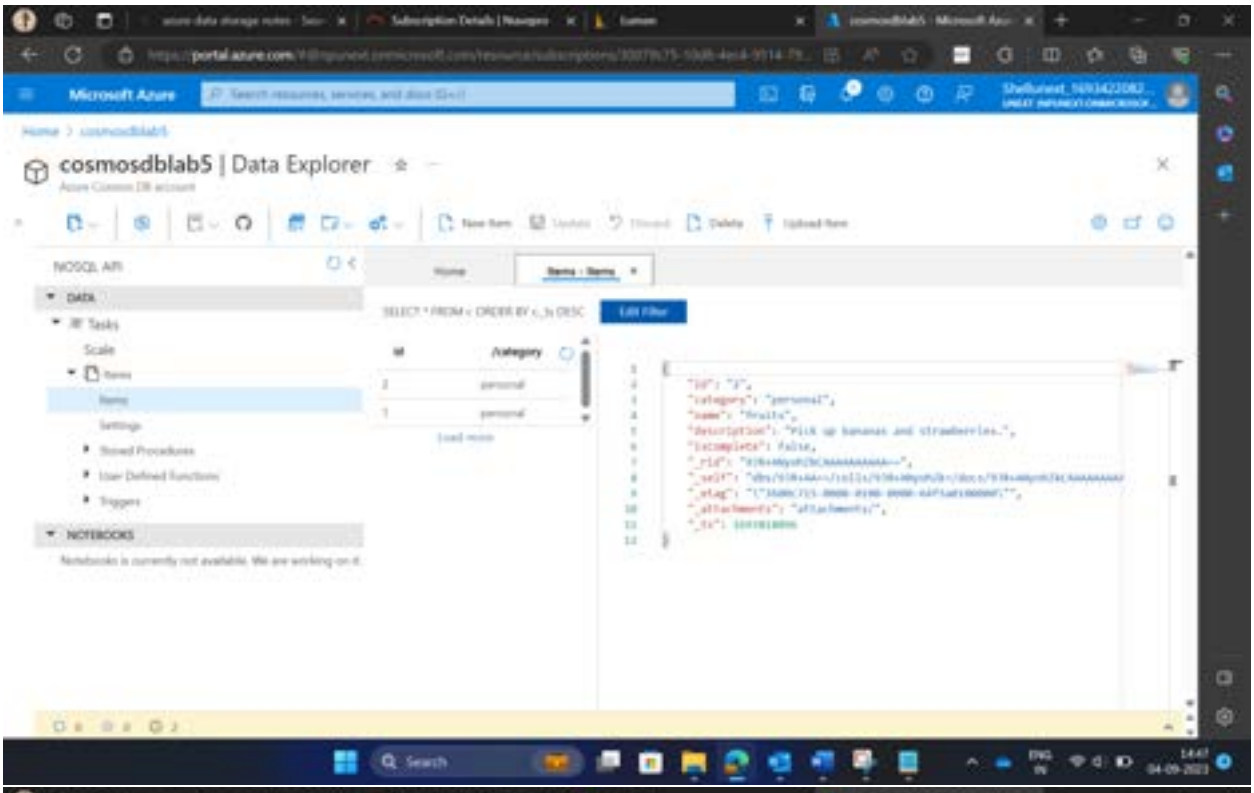

Welcome to ABC Corporates
We are at your service

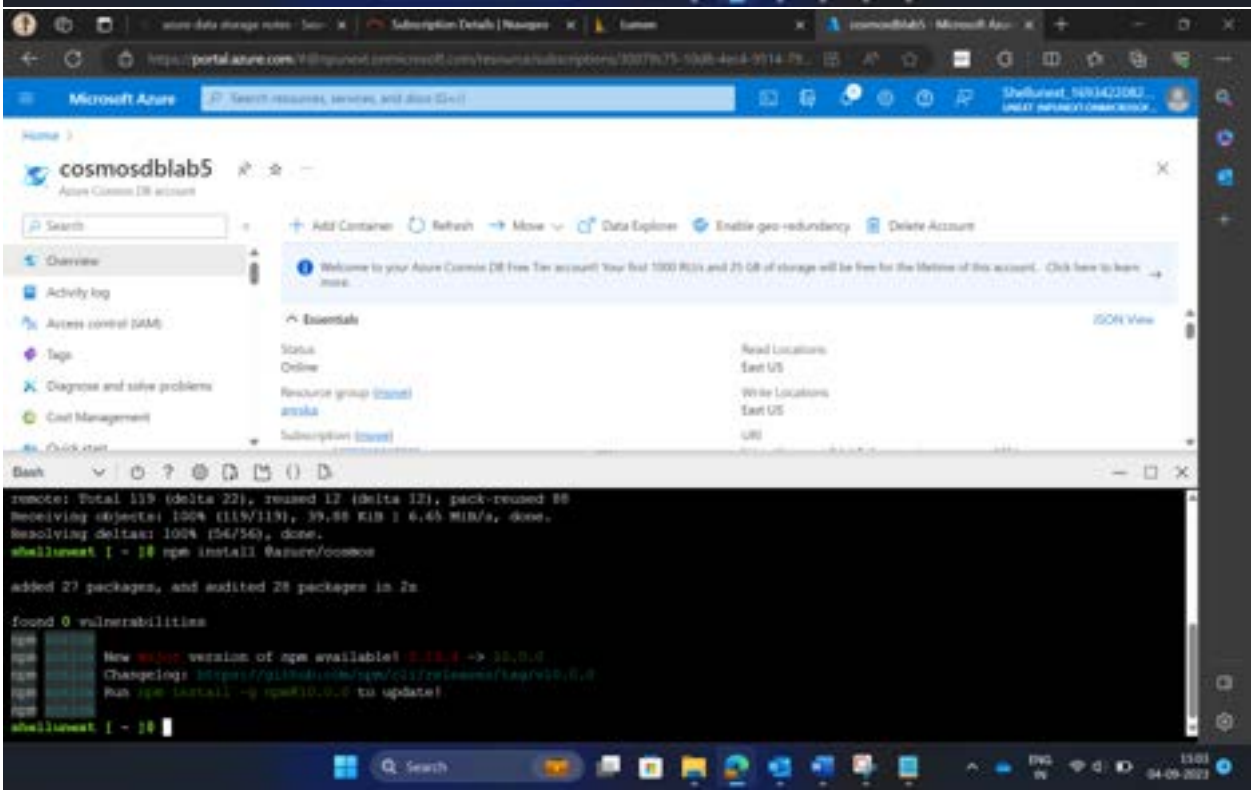

Microsoft Azure
azuredatalakelab3_1693816391463 | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Deployment succeeded
Go to resource
Pin to dashboard
Your deployment is complete
Deployment name: azuredatalakelab3_1...
Subscription: nguneet-1680261907025
Resource group: amrita
Start time: 9/4/2023, 2:03:21 PM
Correlation ID: 17e8b2f8-4b57-4124-a1eb-2c028
Deployment details
Next steps
Go to resource
Give feedback
Tell us about your experience with deployment
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.

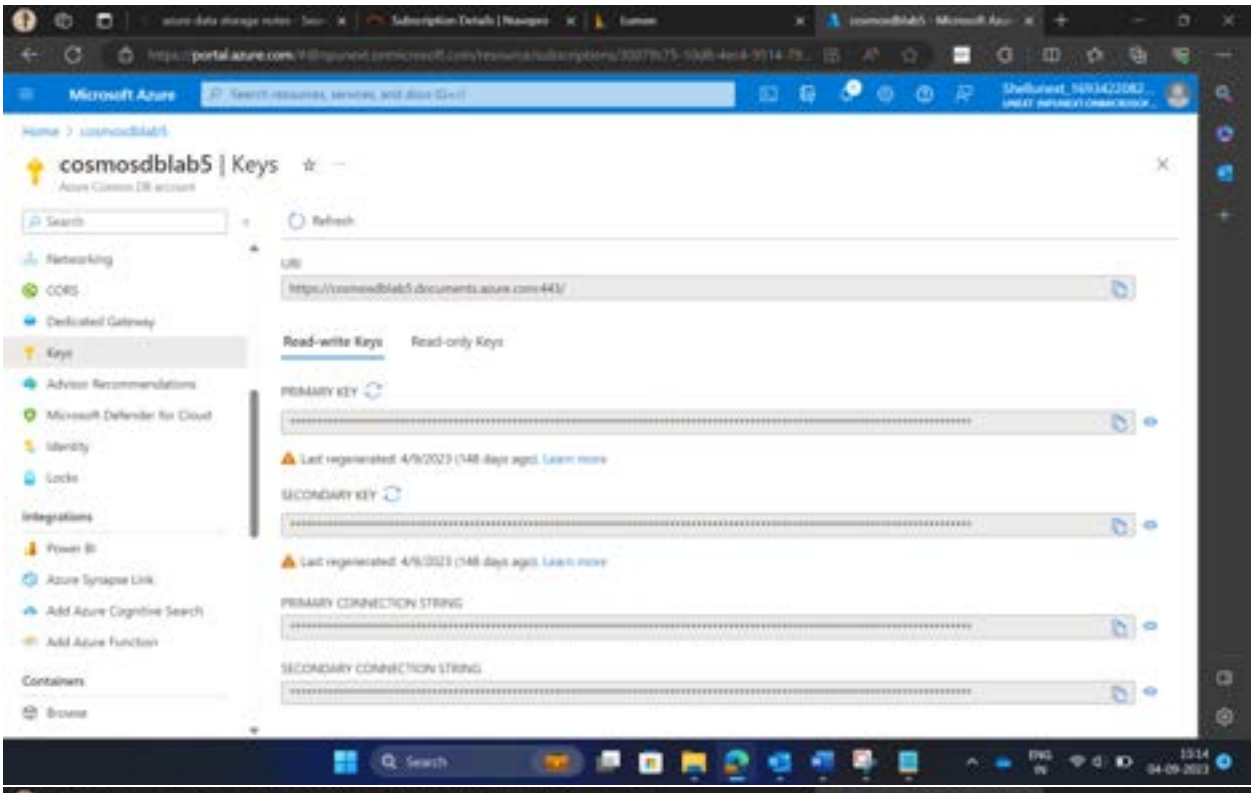

Microsoft Azure
azuredatalakelab3
Storage account
Search
Overview
Activity log
Tags
Diagnose and solve problems
Access Control (IAM)
Data migration
Events
Storage browser
Data storage
Containers
File shares
Queues
Tables
Security + networking
Networking
Upload
Open in Explorer
Delete
Move
Refresh
Open in mobile
CLI / PS
Feedback
Essentials
JSON View
Resource group (move) : amrita
Location : East US
Primary/Secondary Location : Primary: East US, Secondary: West US
Subscription (move) : nguneet-1680261907025
Subscription ID : 30079c75-10d8-4ec4-9514-7956a2a897bb
Disk state : Primary: Available, Secondary: Available
Performance : Standard
Replication : Read-access geo-redundant storage (RA-GRS)
Account kind : StorageV2 (general purpose v2)
Provisioning state : Succeeded
Created : 4/9/2023, 2:03:30 pm
Tags (edit) : Add tags
Properties
Monitoring
Capabilities (5)
Recommendations (0)
Tutorials
Tools + SDKs
Data Lake Storage
Hierarchical namespace : Enabled
Default access tier : Hot
Blob anonymous access : Disabled
Blob soft delete : Enabled (7 days)
Container soft delete : Disabled
Versioning : Disabled
Security
Require secure transfer for REST API operations : Enabled
Storage account key access : Enabled
Minimum TLS version : Version 1.2
Infrastructure encryption : Disabled
Networking

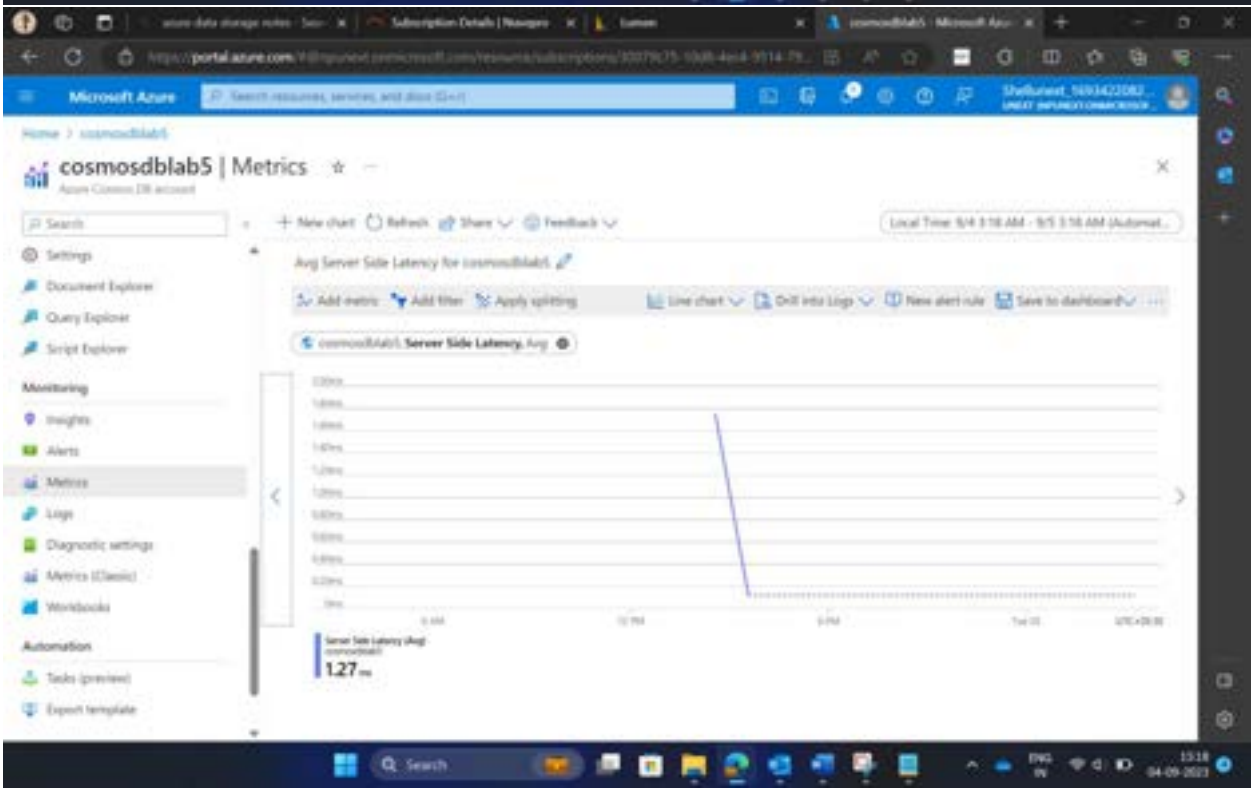

Microsoft Azure
cosmosdblab5 | Data Explorer
Azure Cosmos DB account
Search
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Cost Management
Quick start
Notifications
Data Explorer
Settings
Features
Replicate data globally
Default consistency
Backup & Restore
Networking
New Container
Enable Azure Synapse Link
NOSQL API
DATA
NOTEBOOKS
Home
Welcome to Azure Cosmos DB
Globally distributed, multi-model database service for any scale
Launch quick start
New Container
Connect


Microsoft Azure
cosmosdblab5 | Data Explorer
Azure Cosmos DB account
Search
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Cost Management
Quick start
Notifications
Data Explorer
Settings
Features
Replicate data globally
Default consistency
Backup & Restore
Networking
NoSQL API
Home
Scale
Items
Save
Discard
Upload Item
Edit Filter
{
"id": "1",
"category" : "personal",
"name" : "groceries",
"description": "Pick up apples and strawberries.",
"isComplete": false
}

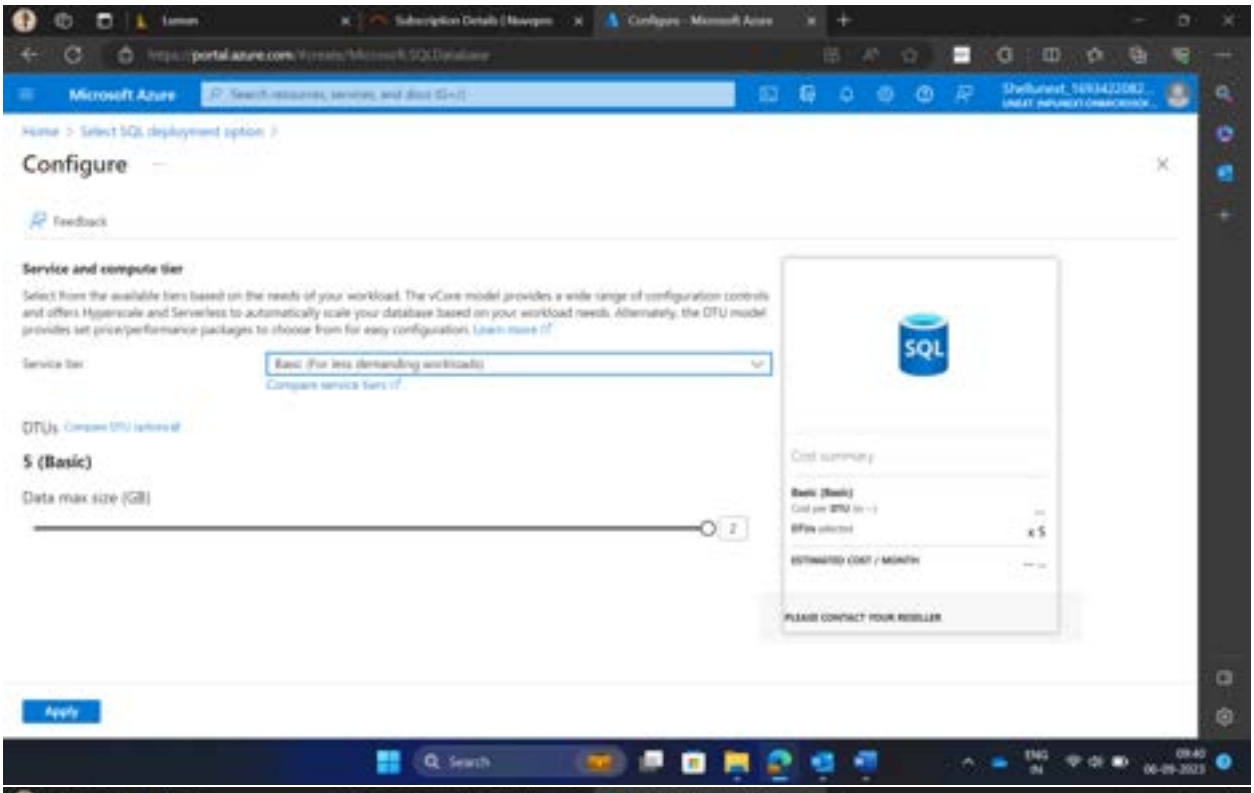

Microsoft Azure
cosmosdblab5 | Data Explorer
NOSQL API
DATA
Tasks
Scale
Items
Settings
Stored Procedures
User Defined Functions
Triggers
NOTEBOOKS
Notebooks is currently not available. We are working on it.
Home
Items
New Item
Update
Discard
Delete
Upload Item
Edit Filter
id
/category
personal
personal
Load more
"id": "3",
"category": "personal",
"name": "Fruits",
"description": "Pick up bananas and strawberries.",
"iscomplete": false,

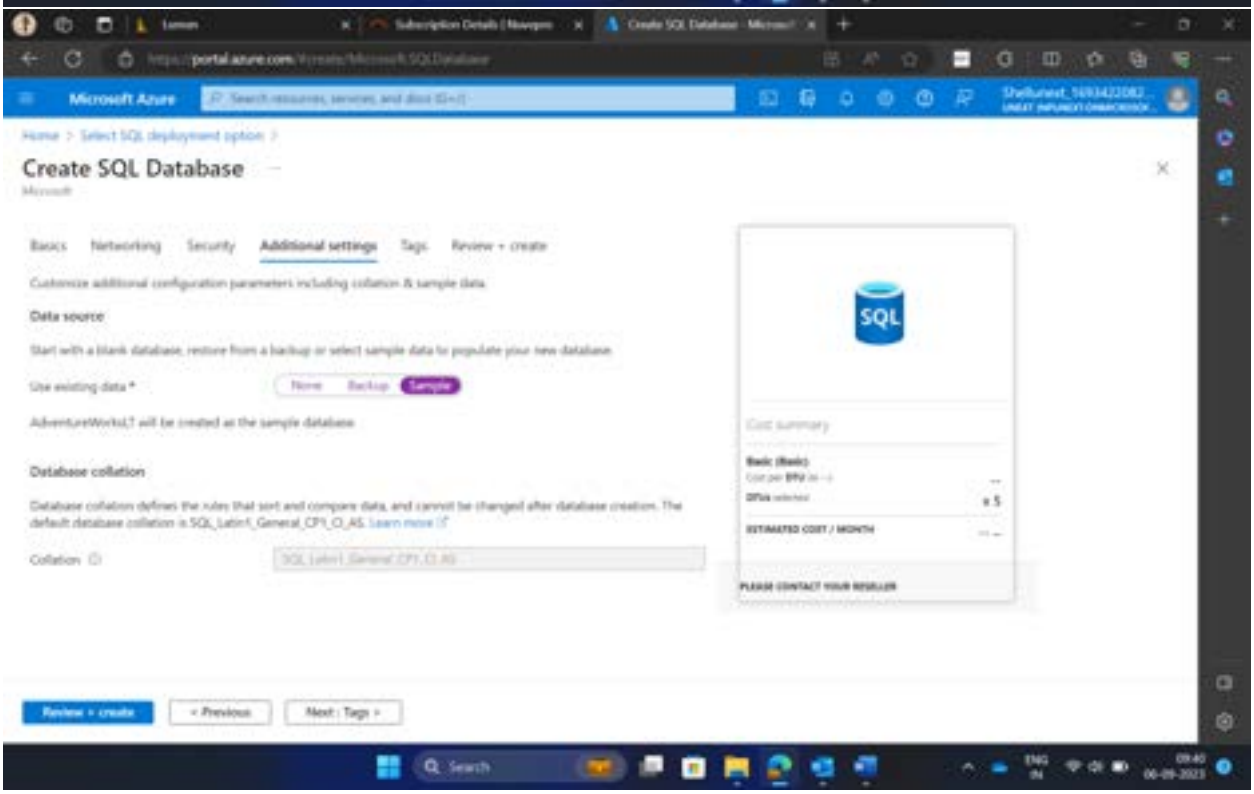

Microsoft Azure
cosmosdblab5
Add Container
Refresh
Move
Data Explorer
Enable geo-redundancy
Delete Account
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Cost Management
Essentials
JSON View
Status
Online
Read Locations
East US
Write Locations
East US
Bash
remote: Total 119 (delta 22), reused 12 (delta 12), pack-reused 88
Receiving objects: 100% (119/119), 39.88 KiB | 6.65 MiB/s, done.
Resolving deltas: 100% (56/56), done.
shelluser4 [~]$ npm install @azure/cosmos
added 27 packages, and audited 28 packages in 2s
found 0 vulnerabilities
shelluser4 [~]$

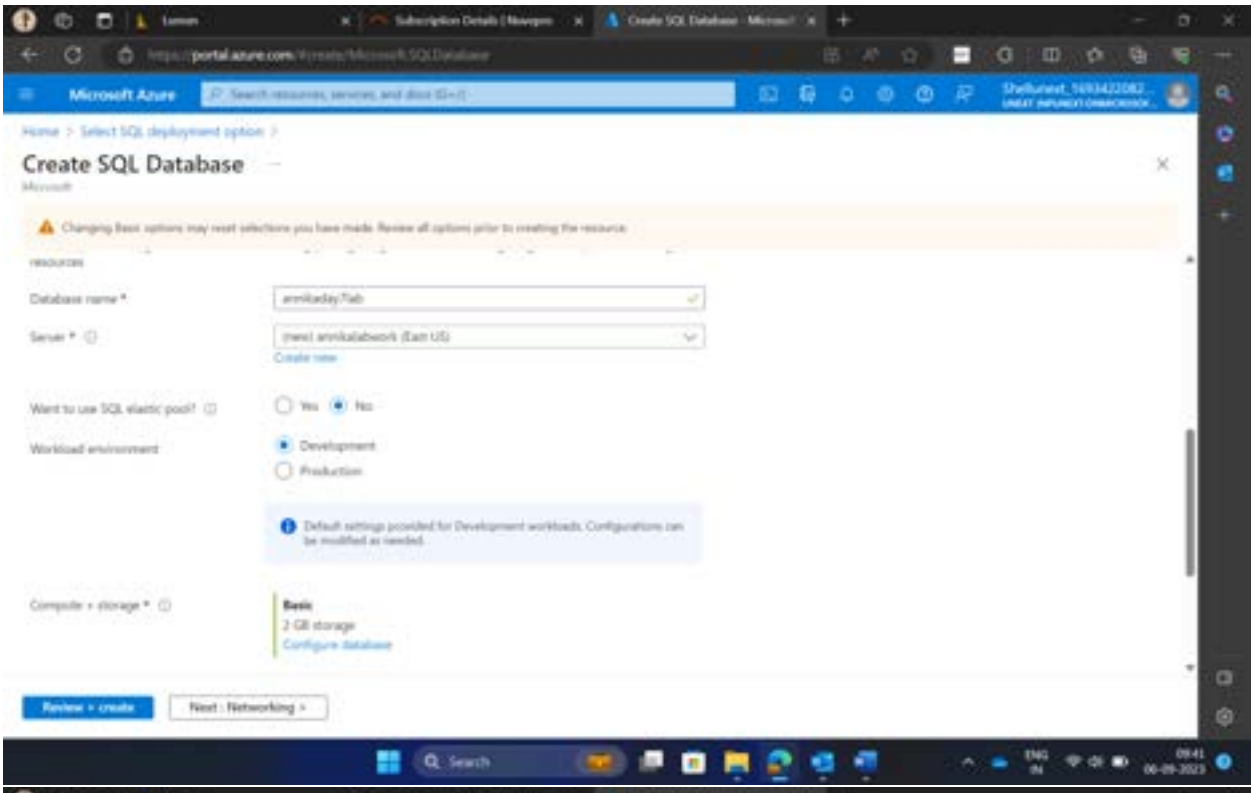

Microsoft Azure
cosmosdblab5 | Keys
Search
Networking
CORS
Dedicated Gateway
Keys
Advisor Recommendations
Microsoft Defender for Cloud
Identity
Locks
Integrations
Power BI
Azure Synapse Link
Add Azure Cognitive Search
Add Azure Function
Containers
Browse
Refresh
URI
https://cosmosdblab5.documents.azure.com:443/
Read-write Keys
Read-only Keys
PRIMARY KEY
Last regenerated: 4/9/2023 (148 days ago). Learn more
SECONDARY KEY
Last regenerated: 4/9/2023 (148 days ago). Learn more
PRIMARY CONNECTION STRING
SECONDARY CONNECTION STRING

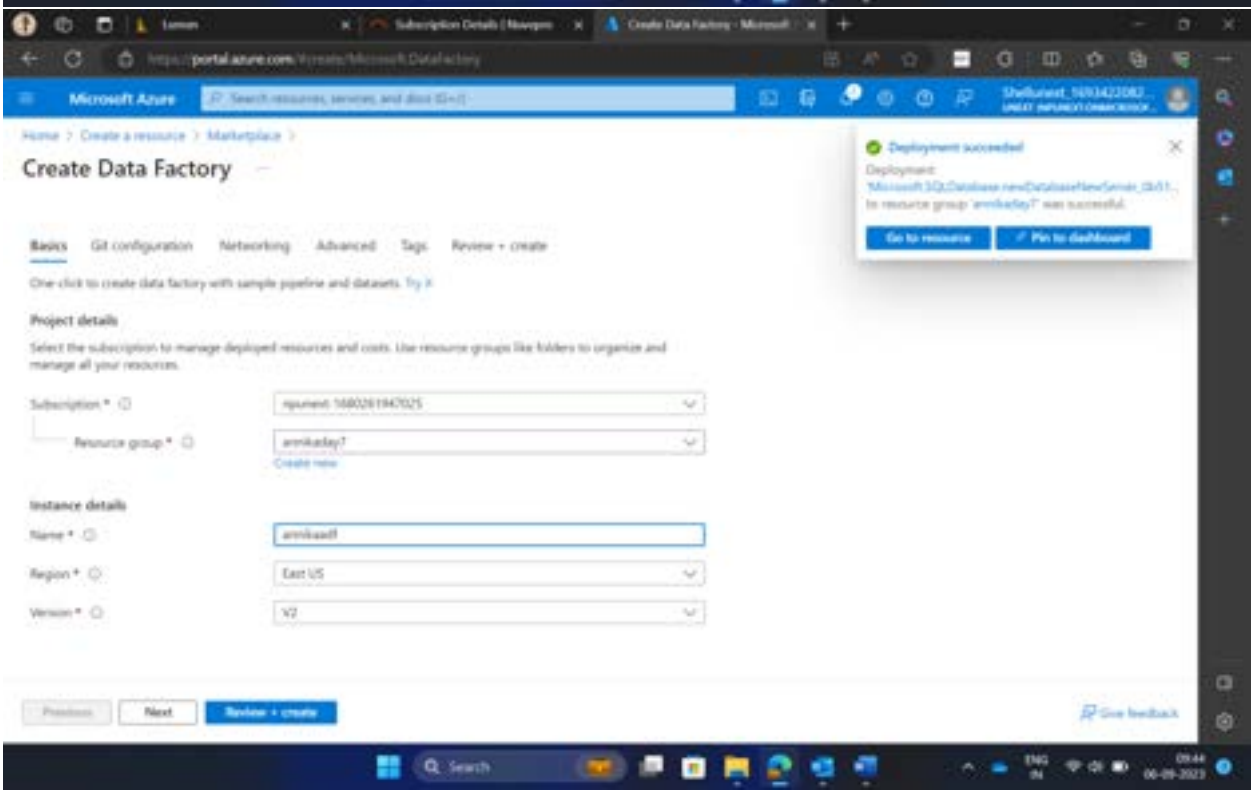

Microsoft Azure
cosmosdblab5 | Metrics
Search
Settings
Document Explorer
Query Explorer
Script Explorer
Monitoring
Insights
Alerts
Metrics
Logs
Diagnostic settings
Metrics (Classic)
Workbooks
Automation
Tasks (preview)
Export template
New chart
Refresh
Share
Feedback
Avg Server Side Latency for cosmosdblab5
Add metric
Add filter
Apply splitting
Line chart
Drill into Logs
New alert rule
Save to dashboard
cosmosdblab5 Server Side Latency Avg
1.27 ms

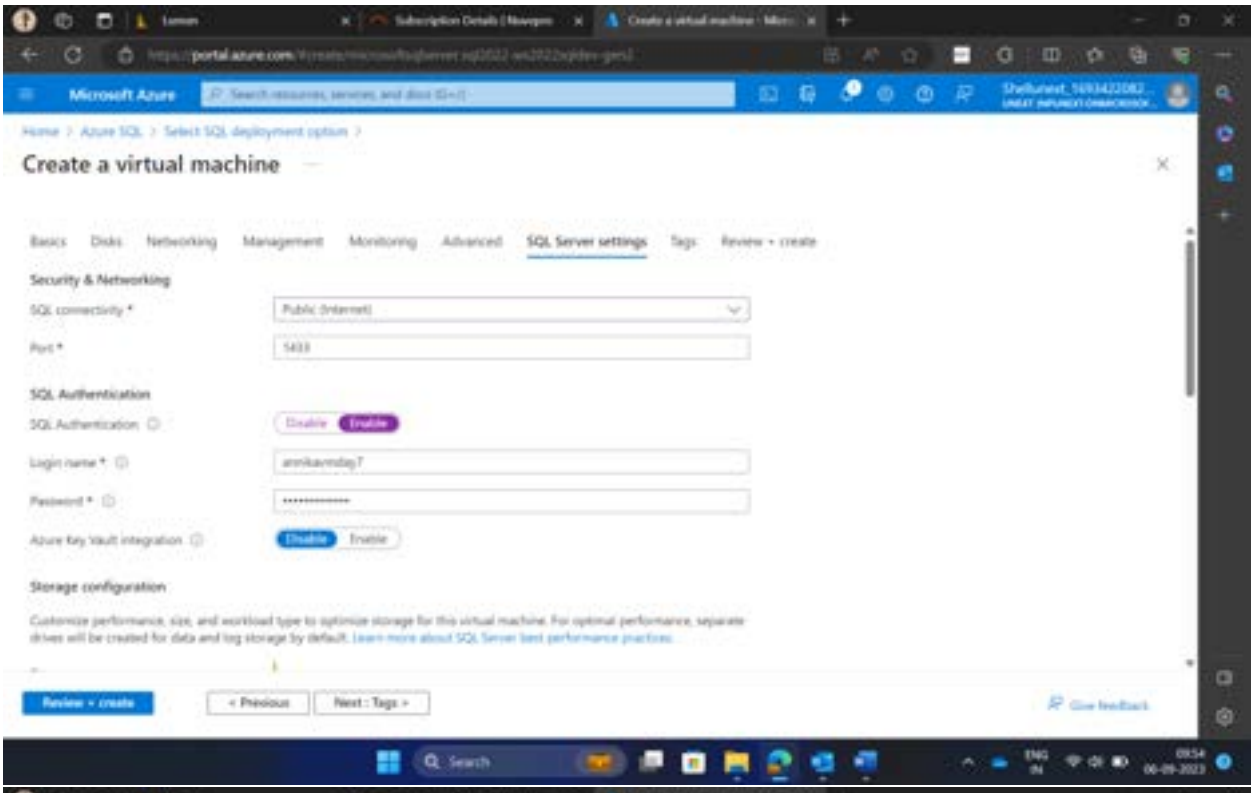
DAY 7

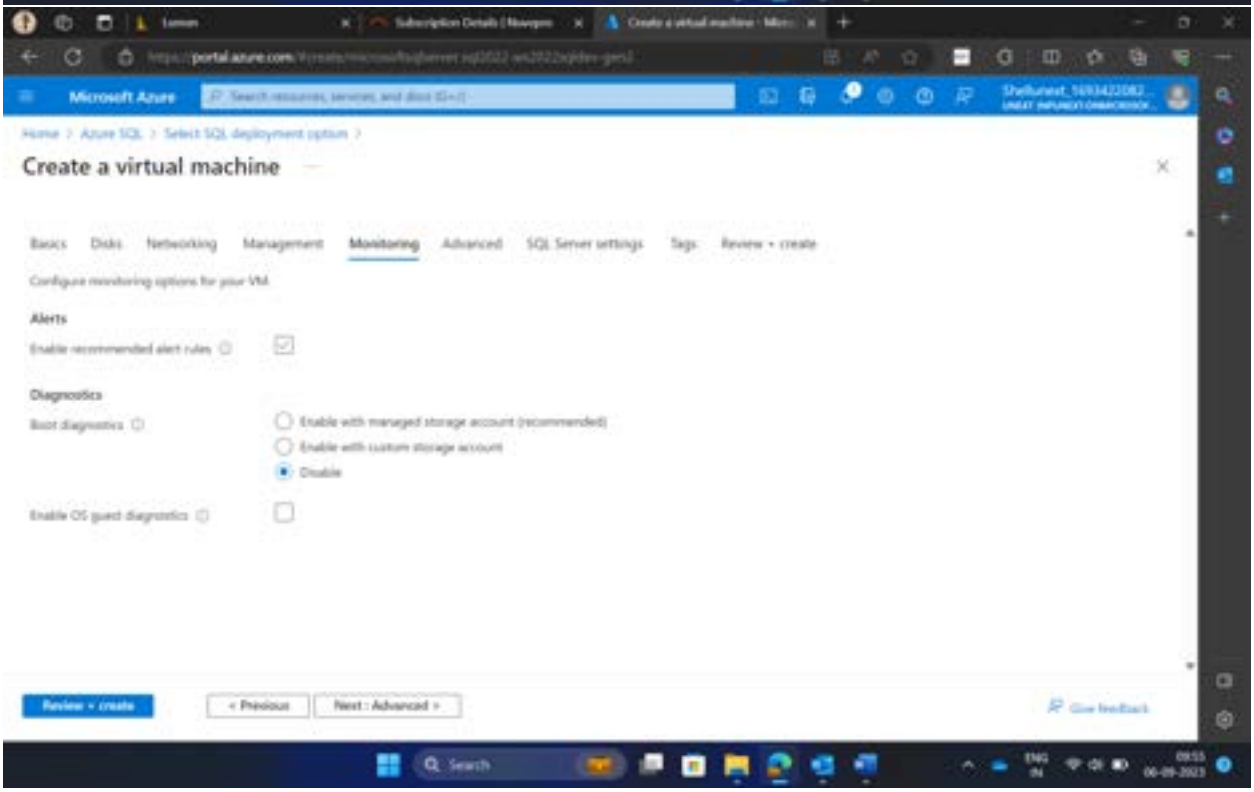

Microsoft Azure
Configure
Feedback
Service and compute tier
Select from the available tiers based on the needs of your workload. The vCore model provides a wide range of configuration controls and offers Hyperscale and Serverless to automatically scale your database based on your workload needs. Alternately, the DTU model provides set price/performance packages to choose from for easy configuration. Learn more
Service tier
Basic (For less demanding workloads)
Compare service tiers
DTUs
5 (Basic)
Data max size (GB)
2
Cost summary
Basic (Basic)
x 5
ESTIMATED COST / MONTH
PLEASE CONTACT YOUR RESELLER
Apply

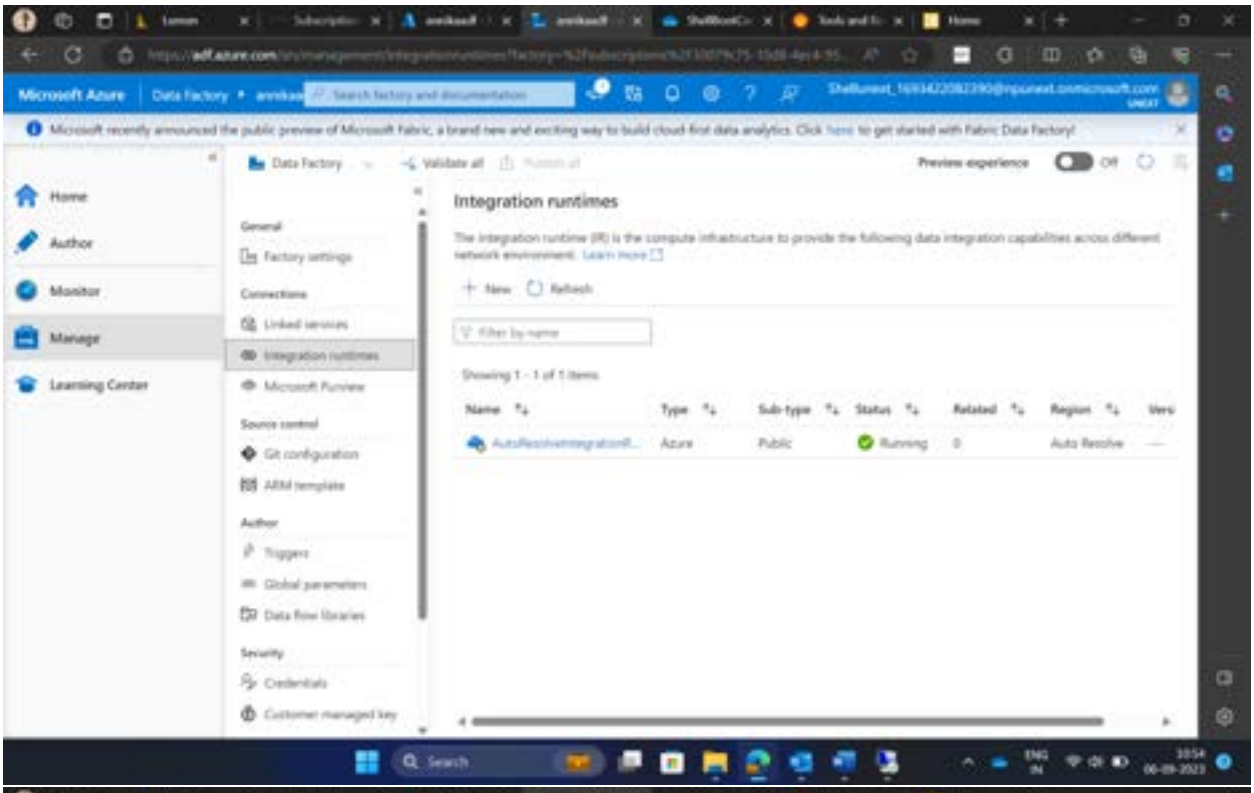

Microsoft Azure
Home > Select SQL deployment option >
Create SQL Database
Basics
Networking
Security
Additional settings
Tags
Review + create
Customize additional configuration parameters including collation & sample data.
Data source
Start with a blank database, restore from a backup or select sample data to populate your new database.
Use existing data *
None
Backup
Sample
AdventureWorksLT will be created as the sample database.
Database collation
Database collation defines the rules that sort and compare data, and cannot be changed after database creation. The default database collation is SQL_Latin1_General_CP1_CI_AS. Learn more
Collation
Cost summary
Basic (Basic)
x 5
ESTIMATED COST / MONTH
PLEASE CONTACT YOUR RESELLER
Review + create
< Previous
Next : Tags >

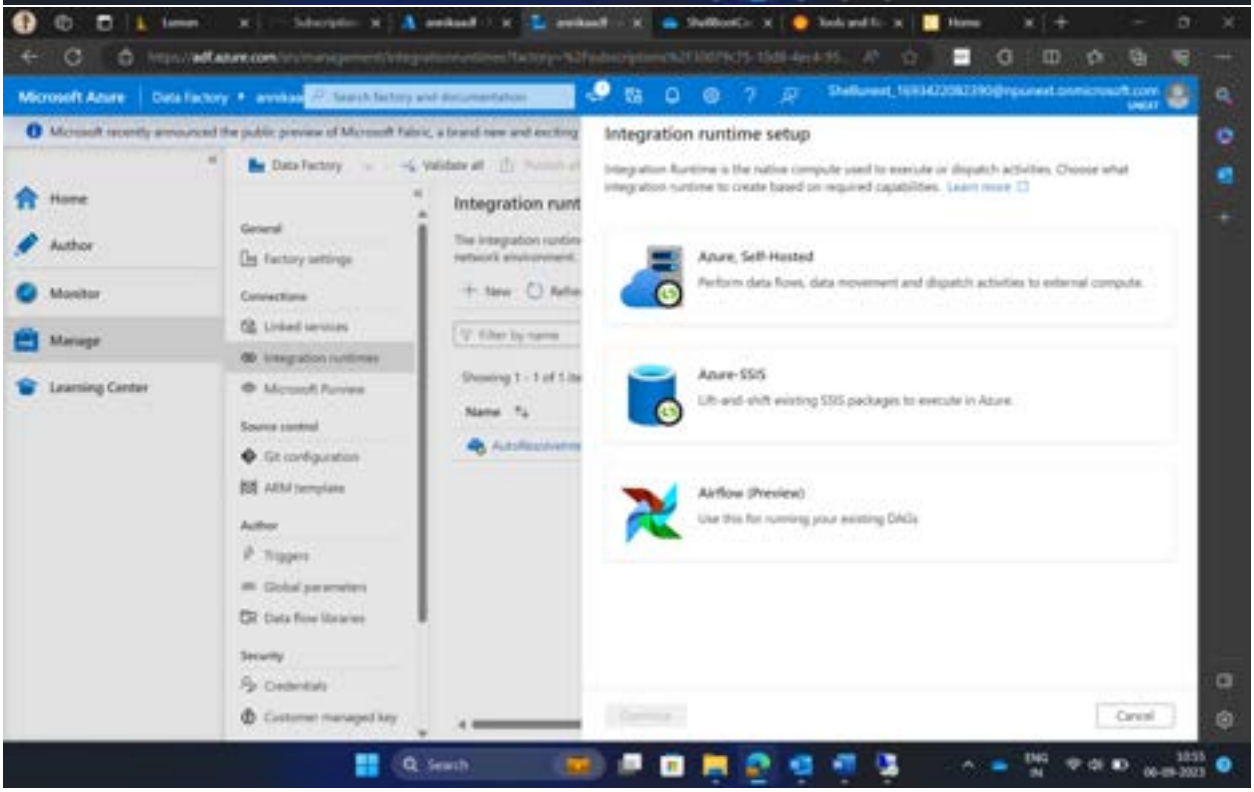

Microsoft Azure
Create SQL Database
Database name *
Server *
Want to use SQL elastic pool?
Workload environment
Development
Production
Compute + storage *
Basic
2 GB storage
Review + create
Next : Networking >

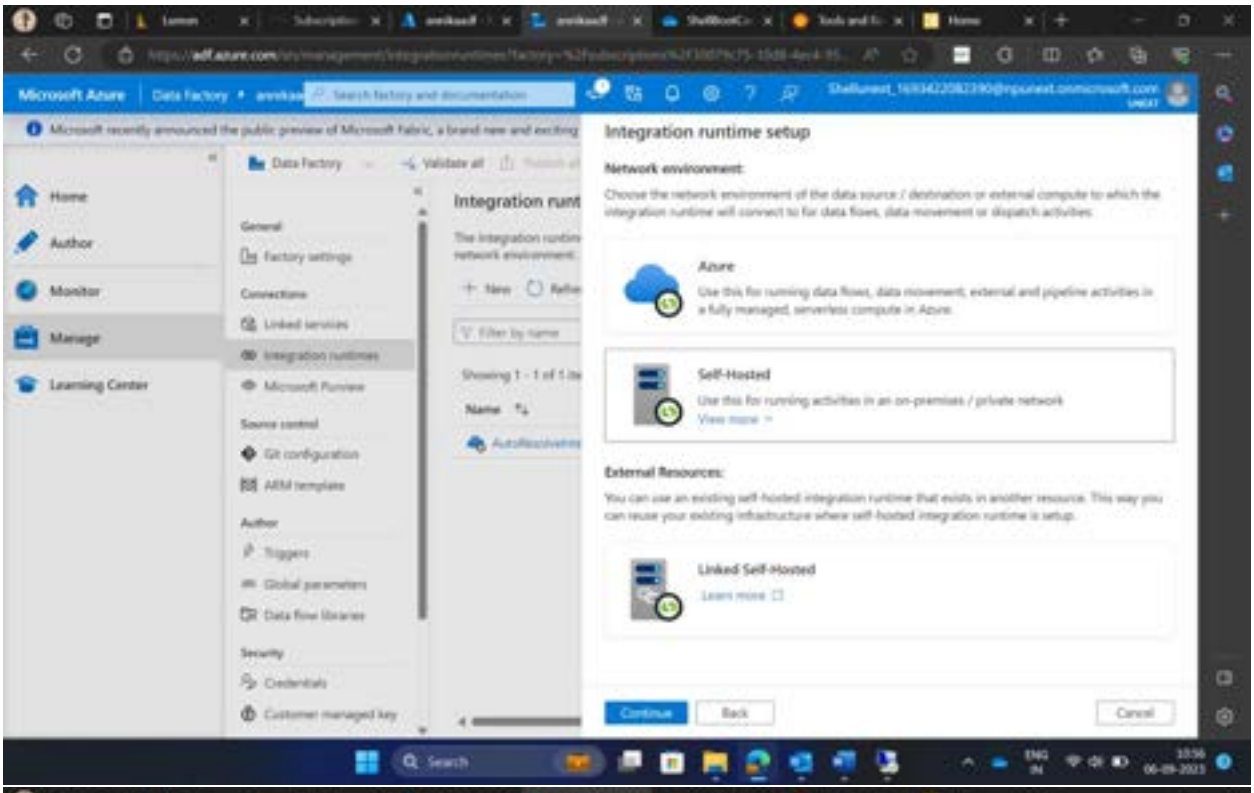

Microsoft Azure
Create Data Factory
Deployment succeeded
Go to resource
Pin to dashboard
Basics
Git configuration
Networking
Advanced
Tags
Review + create
Project details
Subscription *
Resource group *
Instance details
Name *
Region *
East US
Version *
V2
Next
Review + create

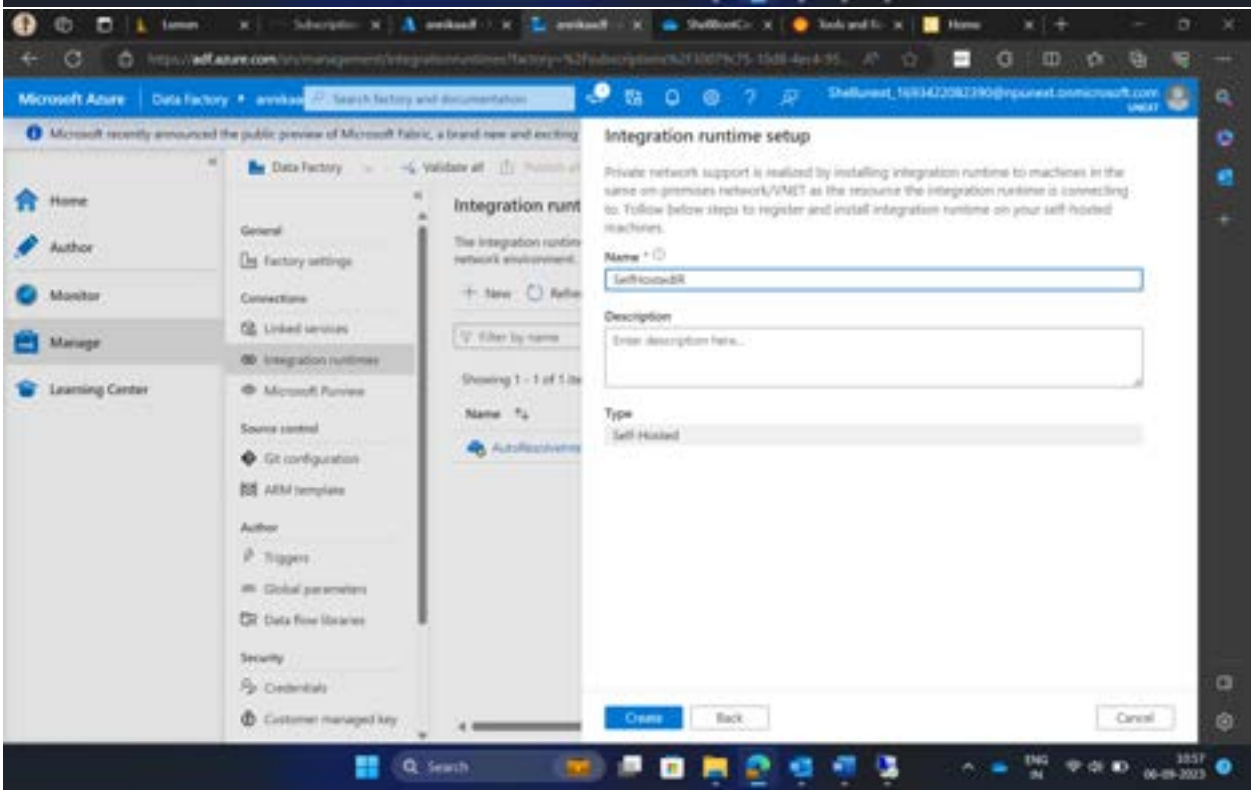

Microsoft Azure
Home > Azure SQL > Select SQL deployment option >
Create a virtual machine
Basics Disks Networking Management Monitoring Advanced SQL Server settings Tags Review + create
Security & Networking
SQL connectivity *
Public (Internet)
Port *
1433
SQL Authentication
SQL Authentication
Disable Enable
Login name *
Password *
Azure Key Vault integration
Disable Enable
Storage configuration
Customize performance, size, and workload type to optimize storage for this virtual machine. For optimal performance, separate drives will be created for data and log storage by default. Learn more about SQL Server best performance practices
Review + create
< Previous
Next : Tags >
Give feedback

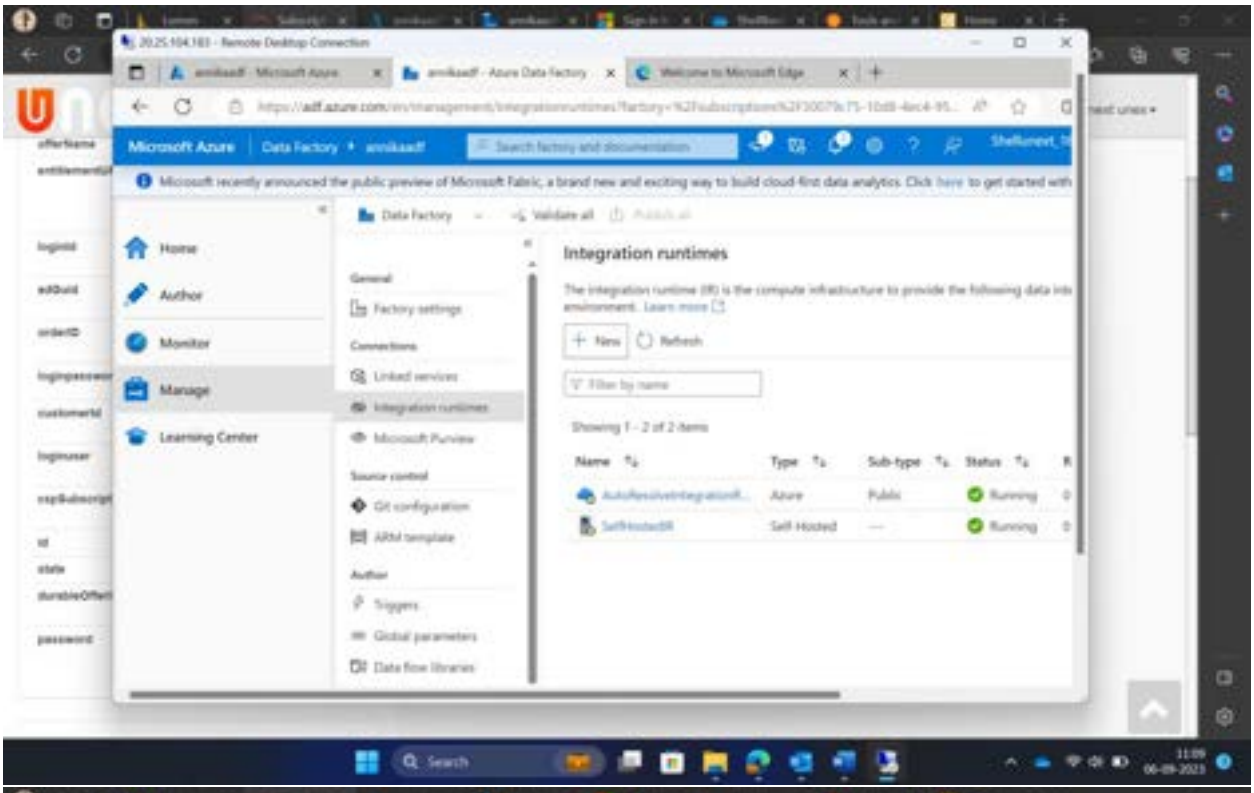

Microsoft Azure
Home > Azure SQL > Select SQL deployment option >
Create a virtual machine
Basics Disks Networking Management Monitoring Advanced SQL Server settings Tags Review + create
Configure monitoring options for your VM.
Alerts
Enable recommended alert rules
Diagnostics
Boot diagnostics
Enable with managed storage account (recommended)
Enable with custom storage account
Disable
Enable OS guest diagnostics
Review + create
< Previous
Next : Advanced >
Give feedback

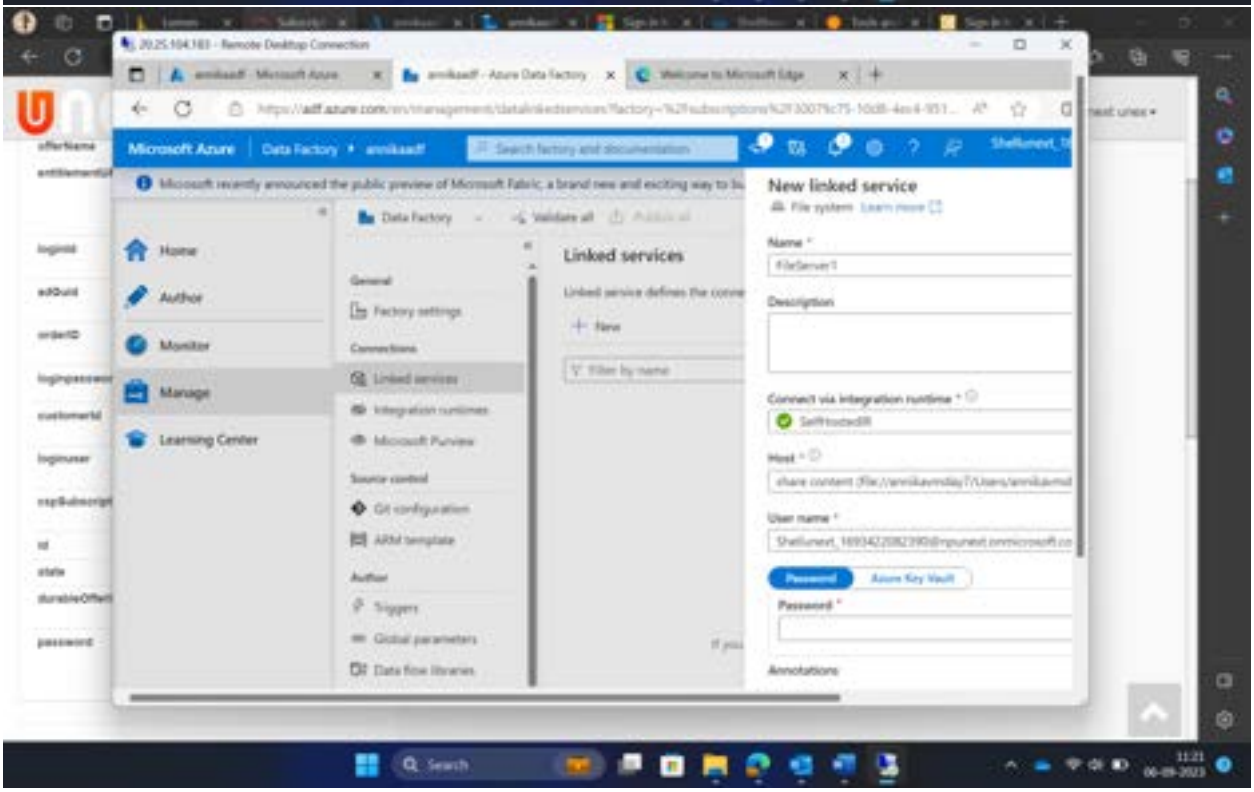

Microsoft Azure
Data Factory
Home
Author
Monitor
Manage
Learning Center
Integration runtimes
New
Refresh
Name
Type
Sub-type
Status
Related
Region
Azure
Public
Running
Auto Resolve

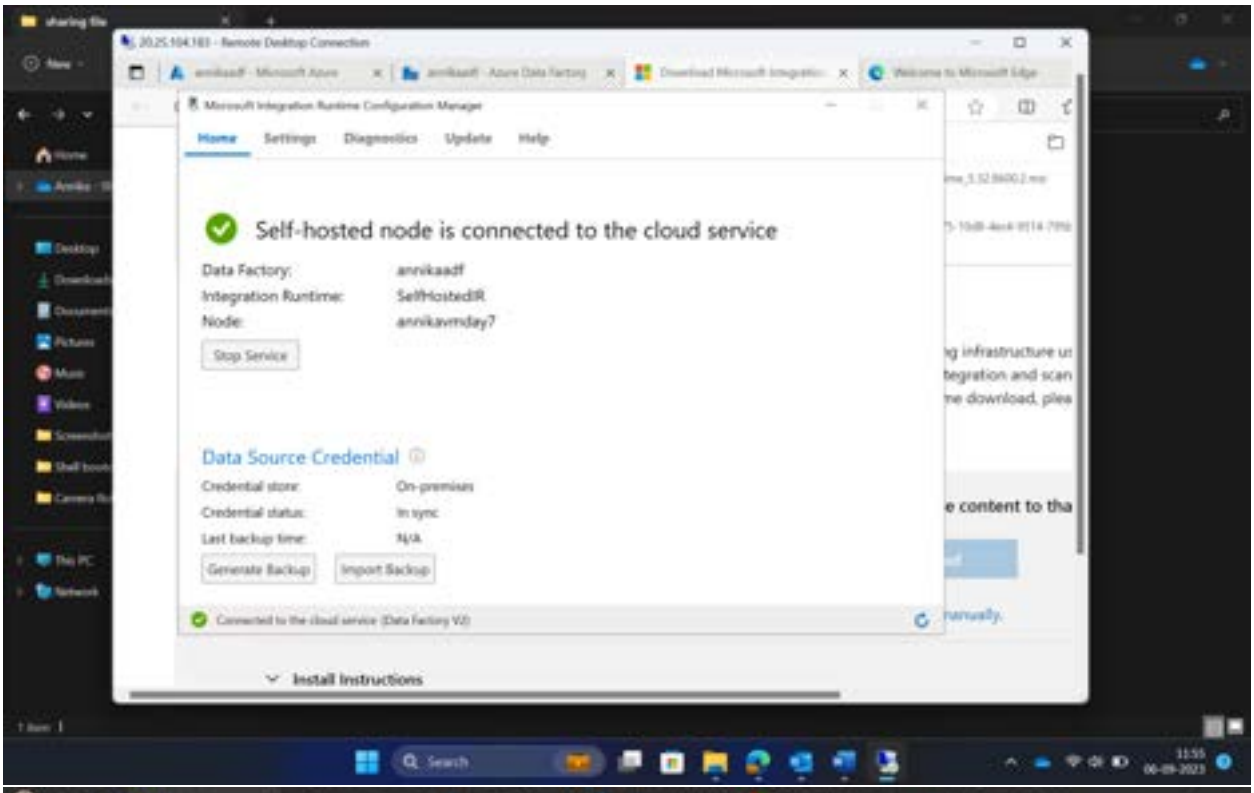

Microsoft Azure
Data Factory
Home
Author
Monitor
Manage
Learning Center
Integration runtime setup
Azure, Self-Hosted
Perform data flows, data movement and dispatch activities to external compute.
Azure-SSIS
Lift-and-shift existing SSIS packages to execute in Azure.
Airflow (Preview)
Use this for running your existing DAGs
Cancel

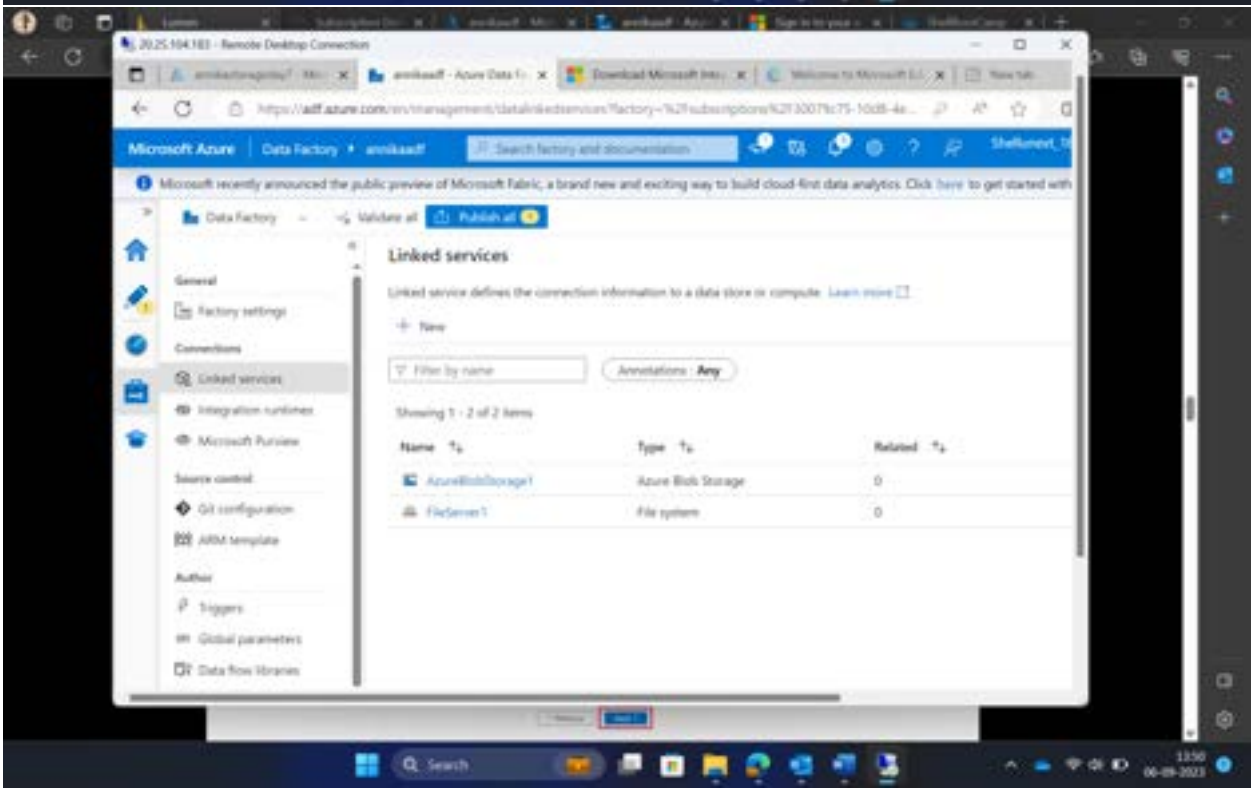

Microsoft Azure
Data Factory
Integration runtime setup
Network environment
Choose the network environment of the data source / destination or external compute to which the integration runtime will connect to for data flows, data movement or dispatch activities
Azure
Use this for running data flows, data movement, external and pipeline activities in a fully managed, serverless compute in Azure.
Self-Hosted
Use this for running activities in an on-premises / private network
View more
External Resources
You can use an existing self-hosted integration runtime that exists in another resource. This way you can reuse your existing infrastructure where self-hosted integration runtime is setup.
Linked Self-Hosted
Learn more
Continue
Back
Cancel
Home
Author
Monitor
Manage
Learning Center
Integration runt
Name
Search

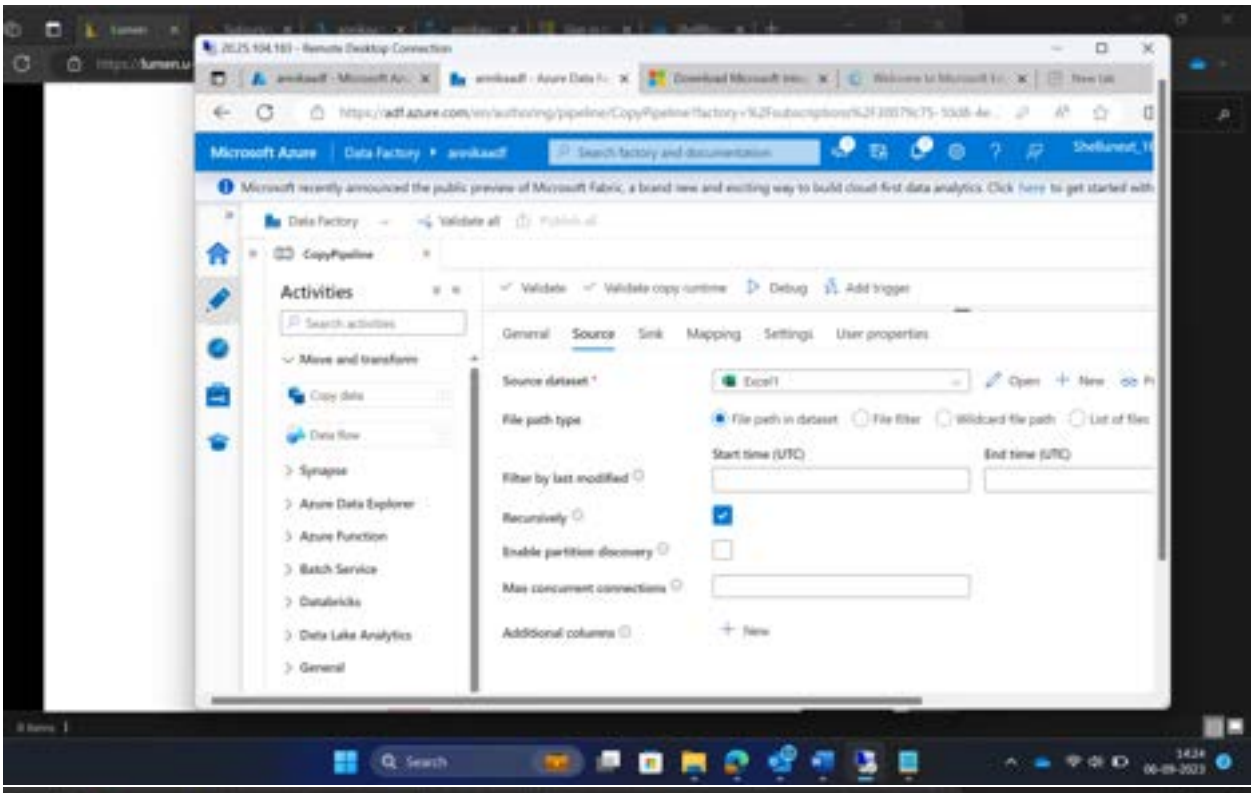

Microsoft Azure
Data Factory
Integration runtime setup
Private network support is realized by installing integration runtime to machines in the same on-premises network/VNET as the resource the integration runtime is connecting to. Follow below steps to register and install integration runtime on your self-hosted machines.
Name
SelfHostedIR
Description
Type
Self-Hosted
Create
Back
Cancel
Home
Author
Monitor
Manage
Learning Center
Integration runt
Name
Search

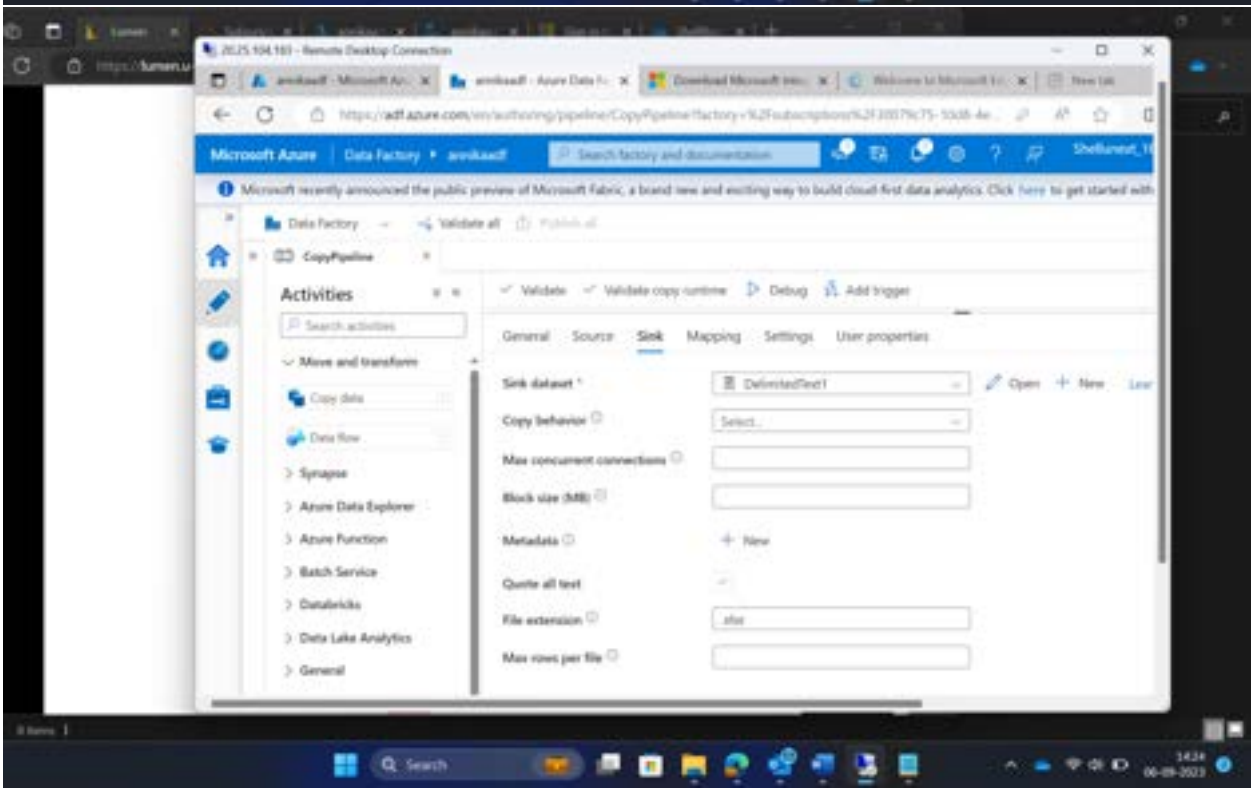

Microsoft Azure
Data Factory
Home
Author
Monitor
Manage
Learning Center
Integration runtimes
New
Refresh
Name
Type
Sub-type
Status

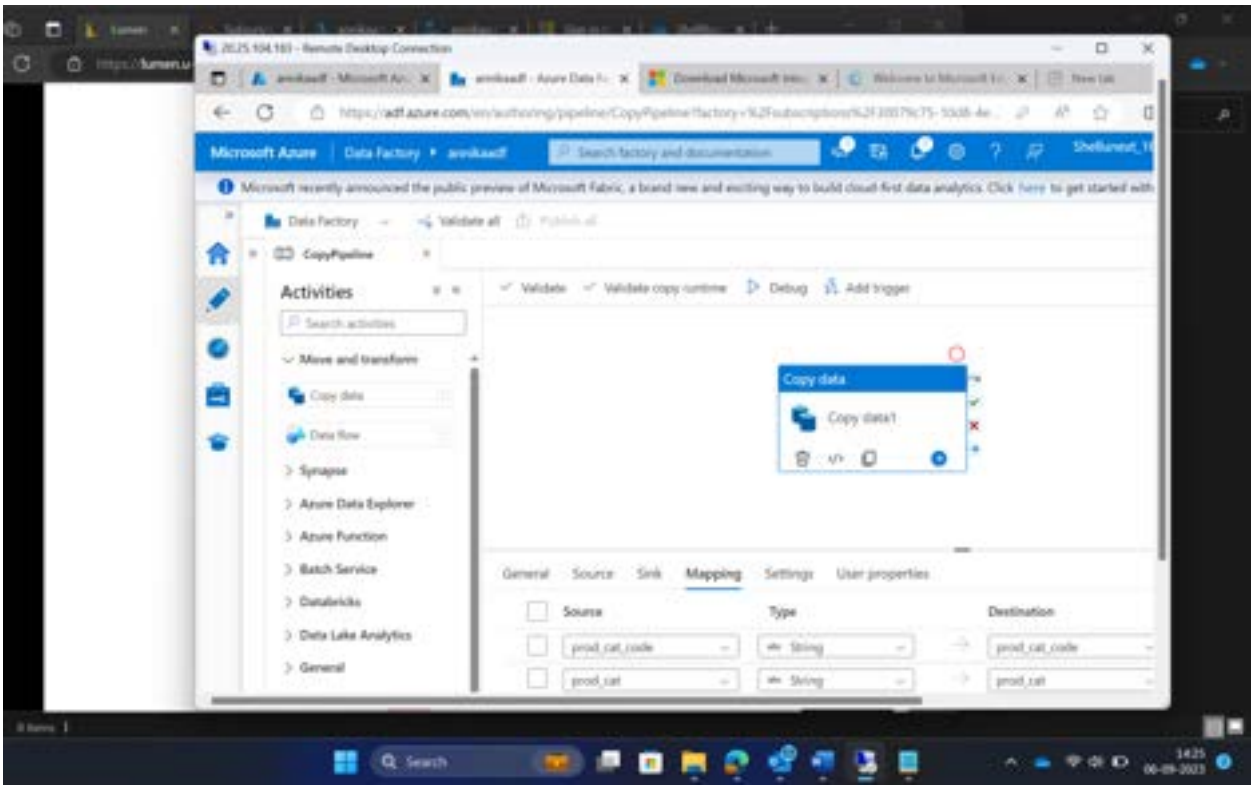

Microsoft Azure
Data Factory
Home
Author
Monitor
Manage
Learning Center
Linked services
New linked service
File system
Name
Description
Connect via integration runtime
Host
User name
Password
Annotations


Microsoft Integration Runtime Configuration Manager
Home
Settings
Diagnostics
Update
Help
Self-hosted node is connected to the cloud service
Data Factory: annikaadf
Integration Runtime: SelfHostedIR
Node: annikavmday7
Stop Service
Data Source Credential
Credential store: On-premises
Credential status: In sync
Last backup time: N/A
Generate Backup
Import Backup
Connected to the cloud service (Data Factory V2)
Install Instructions

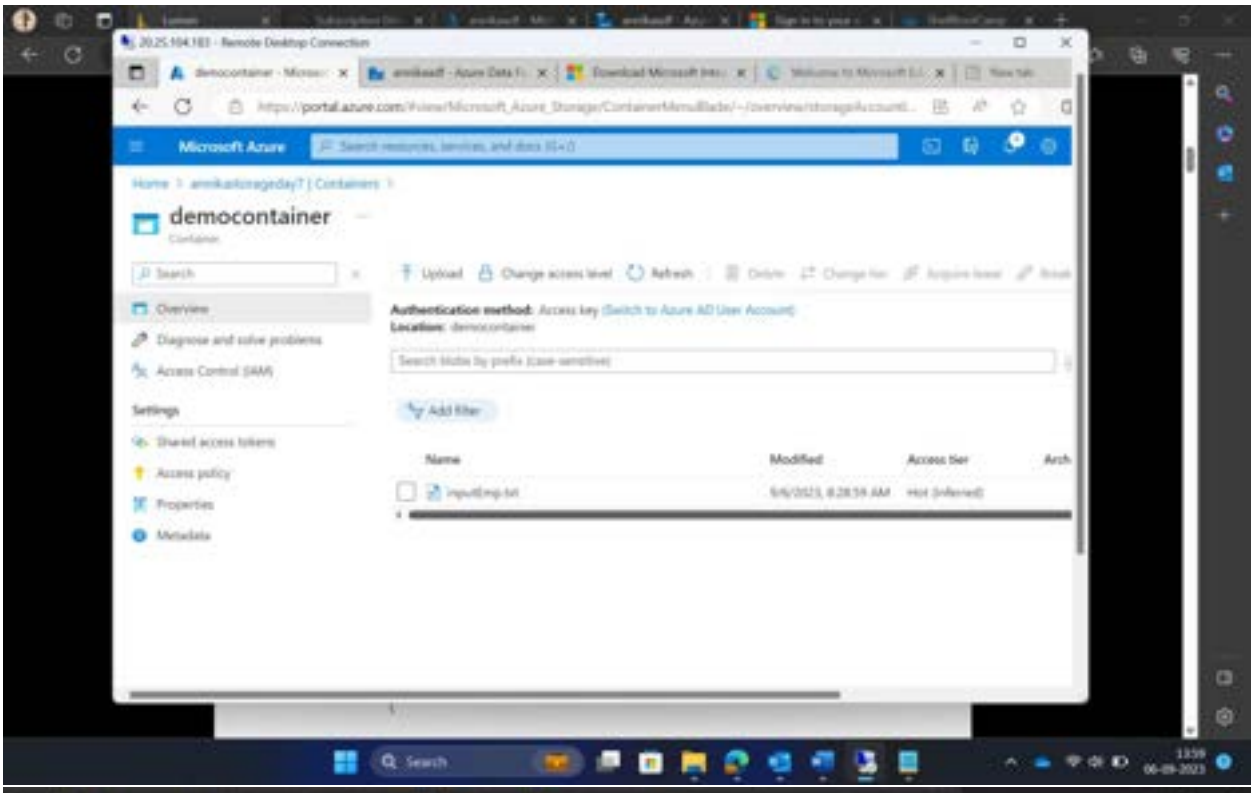

Microsoft Azure
Data Factory
annikaadf
Linked services
Linked service defines the connection information to a data store or compute.
New
Name
Type
Related
Azure Blob Storage
File system

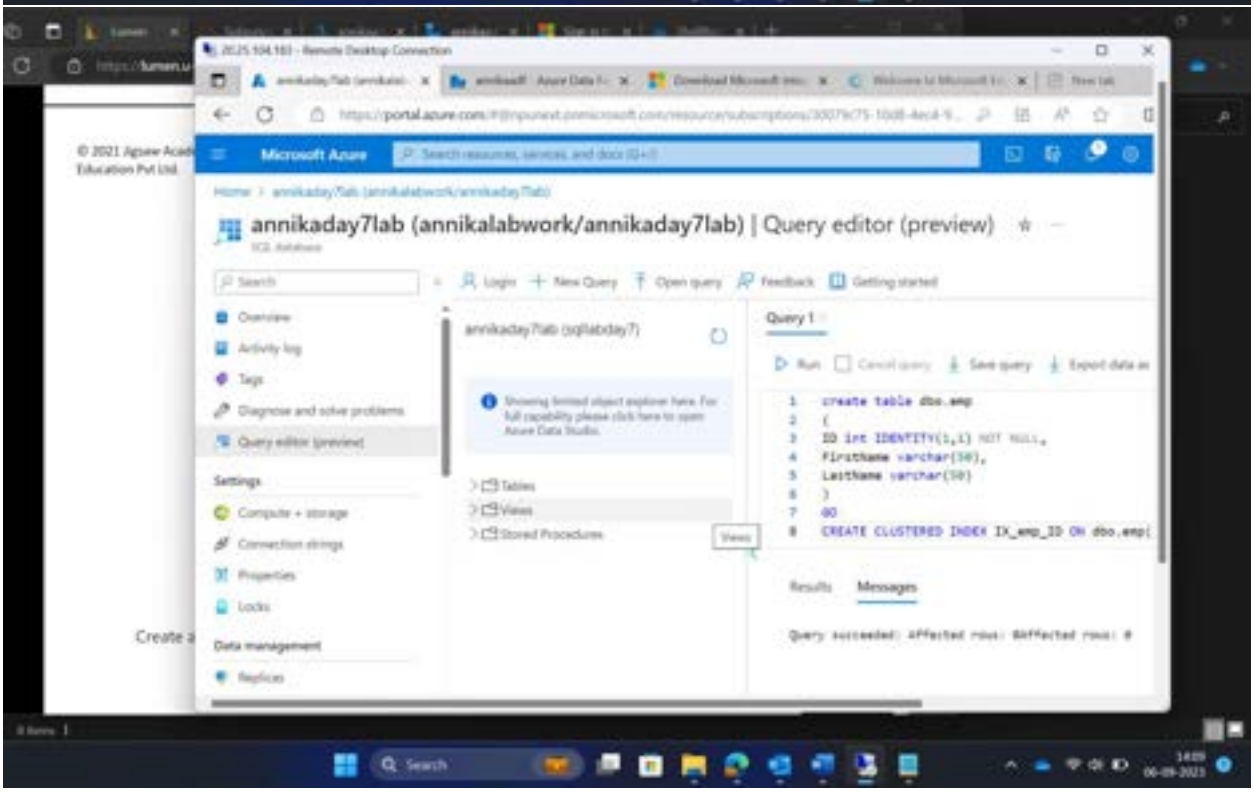

Microsoft Azure
Data Factory
CopyPipeline
Activities
Validate
Validate copy runtime
Debug
Add trigger
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
File path type
Filter by last modified
Recursively
Enable partition discovery
Max concurrent connections
Additional columns
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General

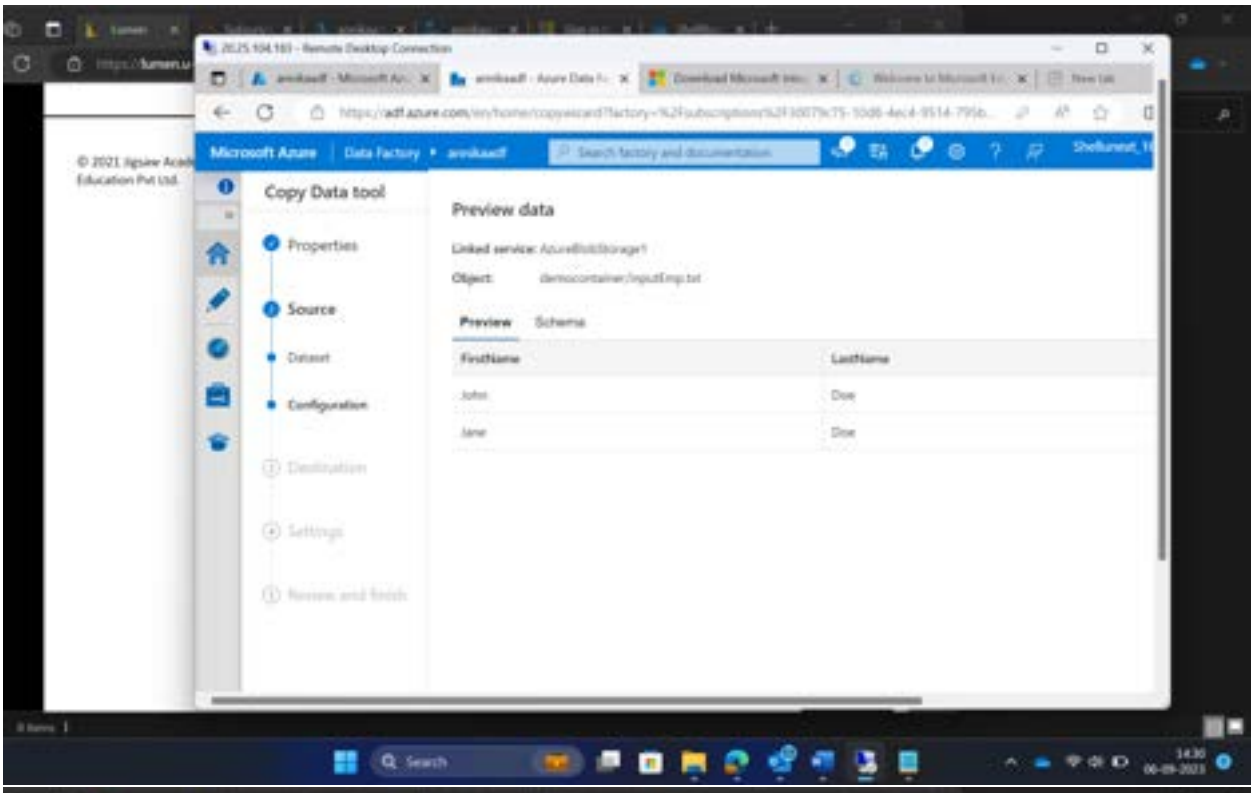

Microsoft Azure
Data Factory
CopyPipeline
Activities
Validate
Validate copy runtime
Debug
Add trigger
General
Source
Sink
Mapping
Settings
User properties
Sink dataset *
Copy behavior
Max concurrent connections
Block size (MB)
Metadata
Quote all text
File extension
Max rows per file
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General

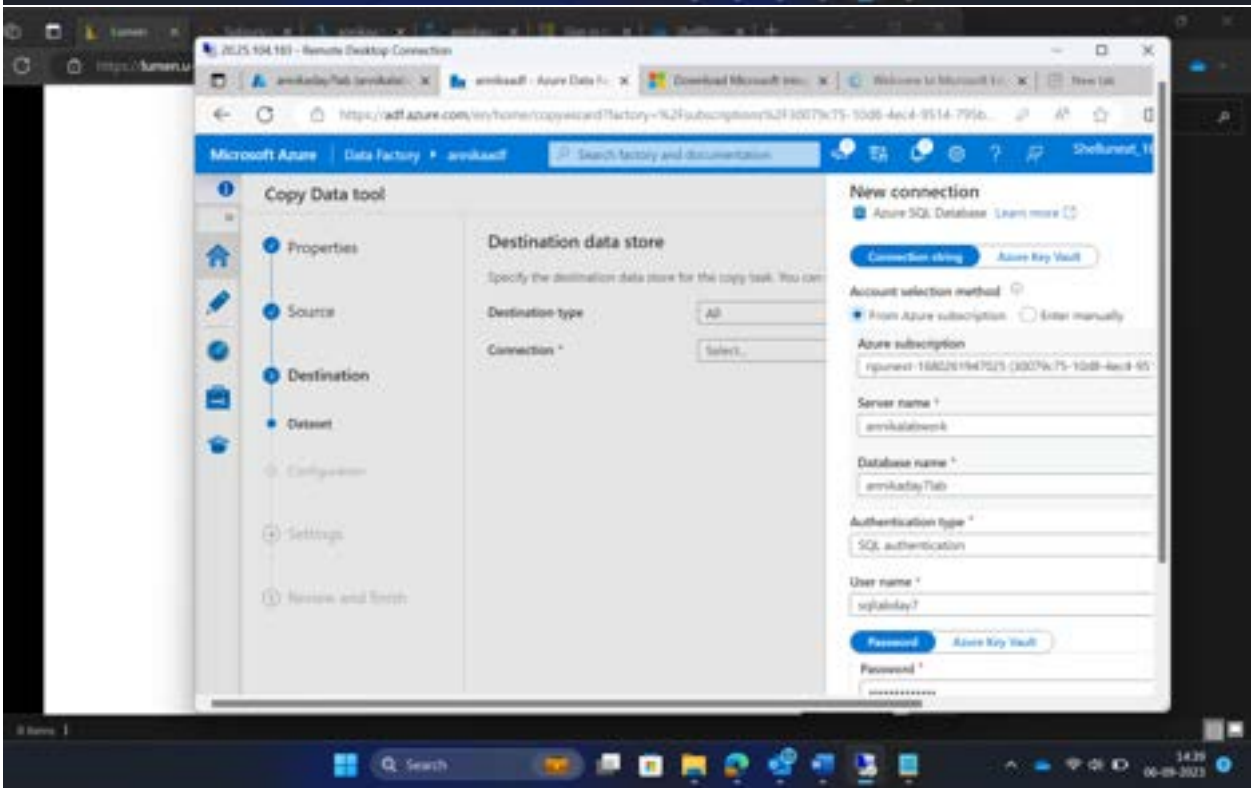


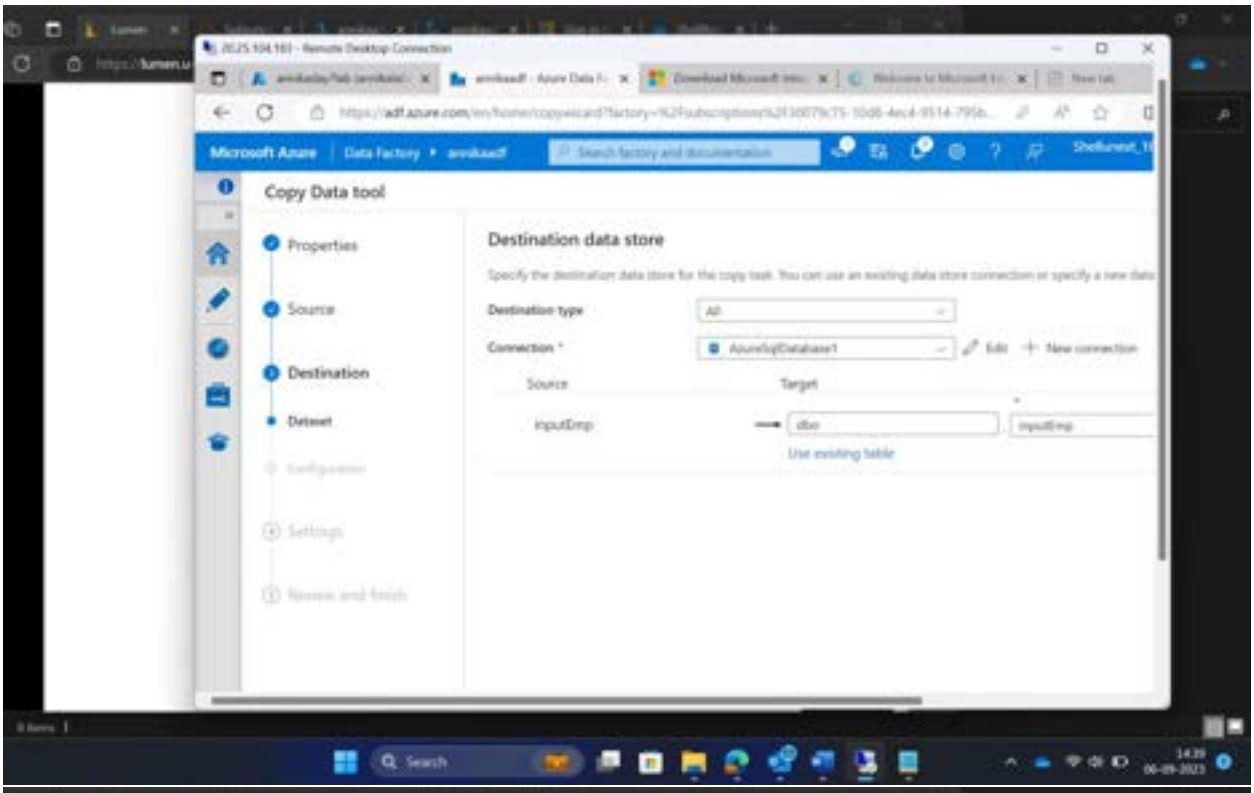
Lab 1 Azure Data Factory

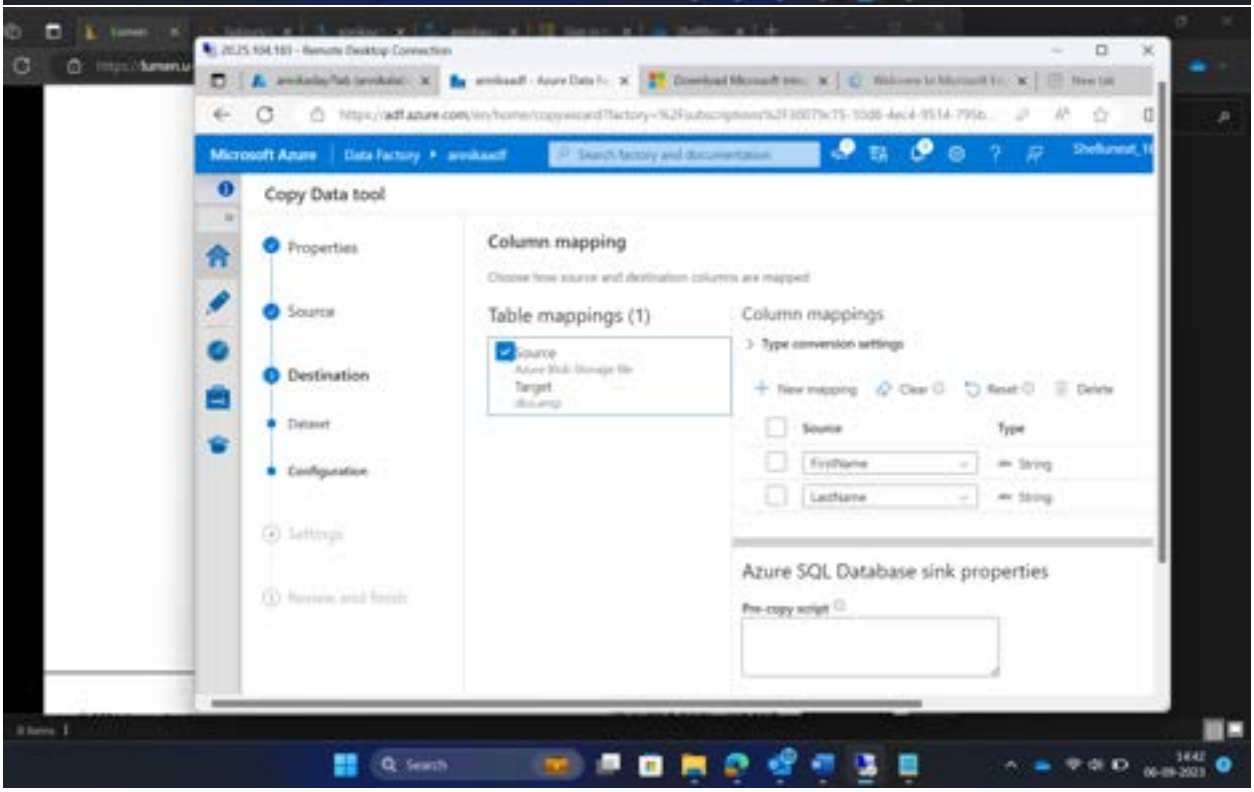

Microsoft Azure
democontainer
Container
Upload
Change access level
Refresh
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Access policy
Properties
Metadata
Authentication method: Access key (Switch to Azure AD User Account)
Location: democontainer
Add filter
Name
Modified
Access tier
Search

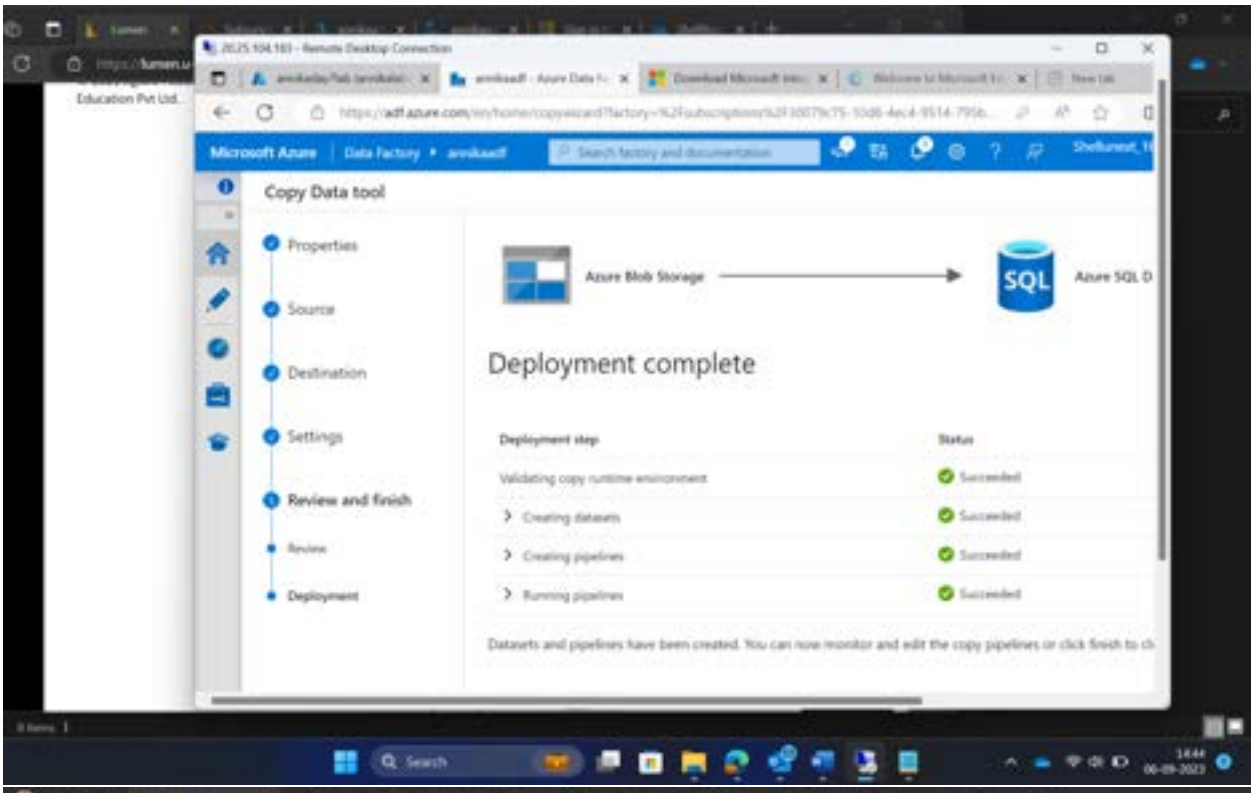

Microsoft Azure
annikaday7lab (annikalabwork/annikaday7lab) | Query editor (preview)
SQL database
Login
New Query
Open query
Feedback
Getting started
Overview
Activity log
Tags
Diagnose and solve problems
Query editor (preview)
Settings
Compute + storage
Connection strings
Properties
Locks
Data management
Replicas
annikaday7lab (sqllabday7)
Tables
Views
Stored Procedures
Query 1
Run
Cancel query
Save query
Export data as
1 create table dbo.emp
2 (
3 ID int IDENTITY(1,1) NOT NULL,
4 Firstname varchar(50),
5 Lastname varchar(50)
6)
7 GO
8 CREATE CLUSTERED INDEX IX_emp_ID ON dbo.emp(
Results
Messages
Search

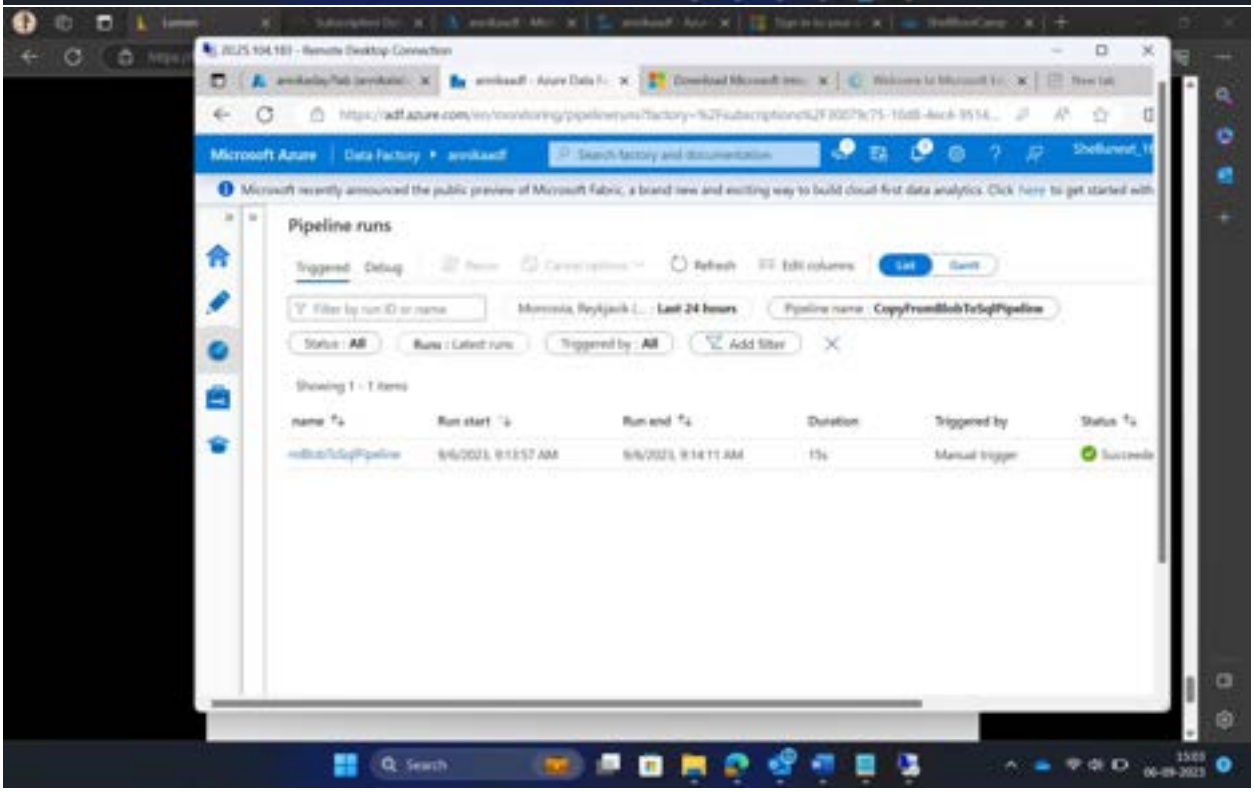

Microsoft Azure
Data Factory
Copy Data tool
Properties
Source
Dataset
Configuration
Destination
Settings
Review and finish
Preview data
Linked service: AzureBlobStorage1
Object:
Preview
Schema
FirstName
LastName
John
Doe
Jane
Doe

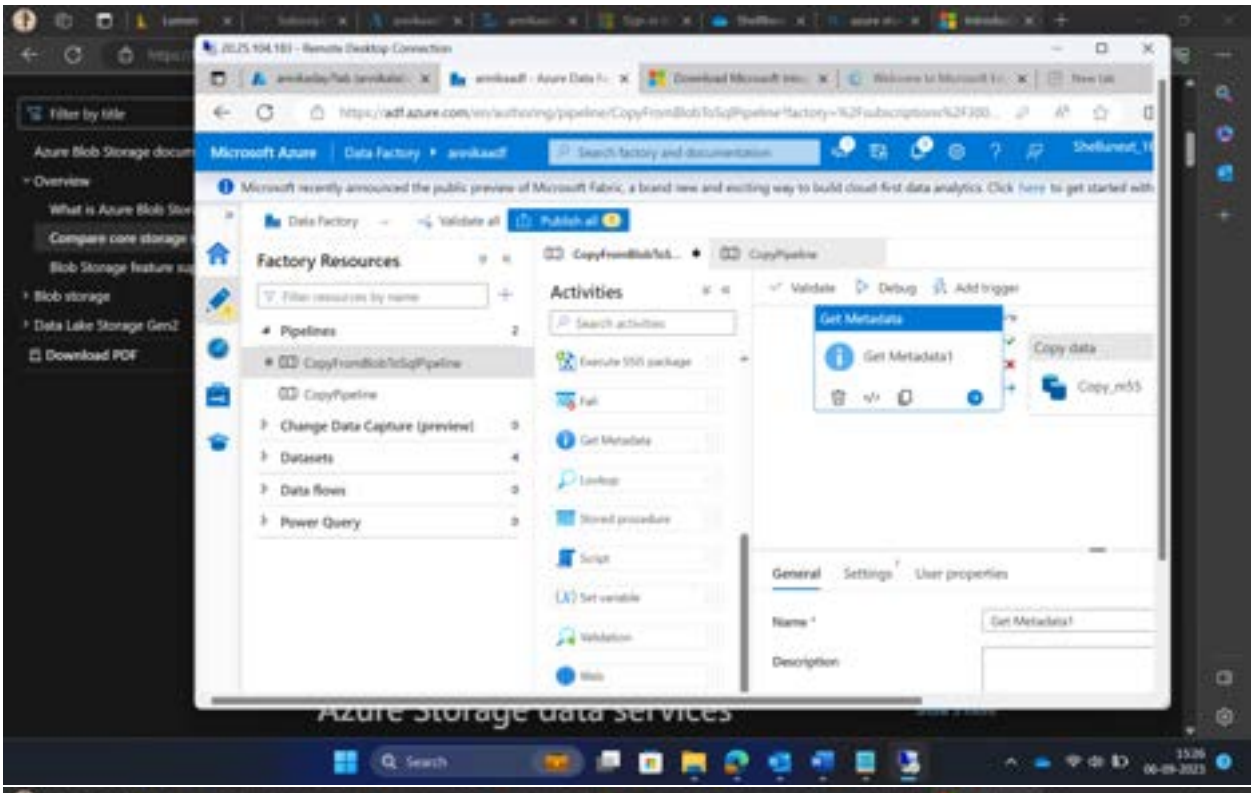

Microsoft Azure
Data Factory
Copy Data tool
Properties
Source
Destination
Dataset
Settings
Review and finish
Destination data store
Destination type
Connection
New connection
Azure SQL Database
Account selection method
From Azure subscription
Enter manually
Azure subscription
Server name
Database name
Authentication type
SQL authentication
User name
Password

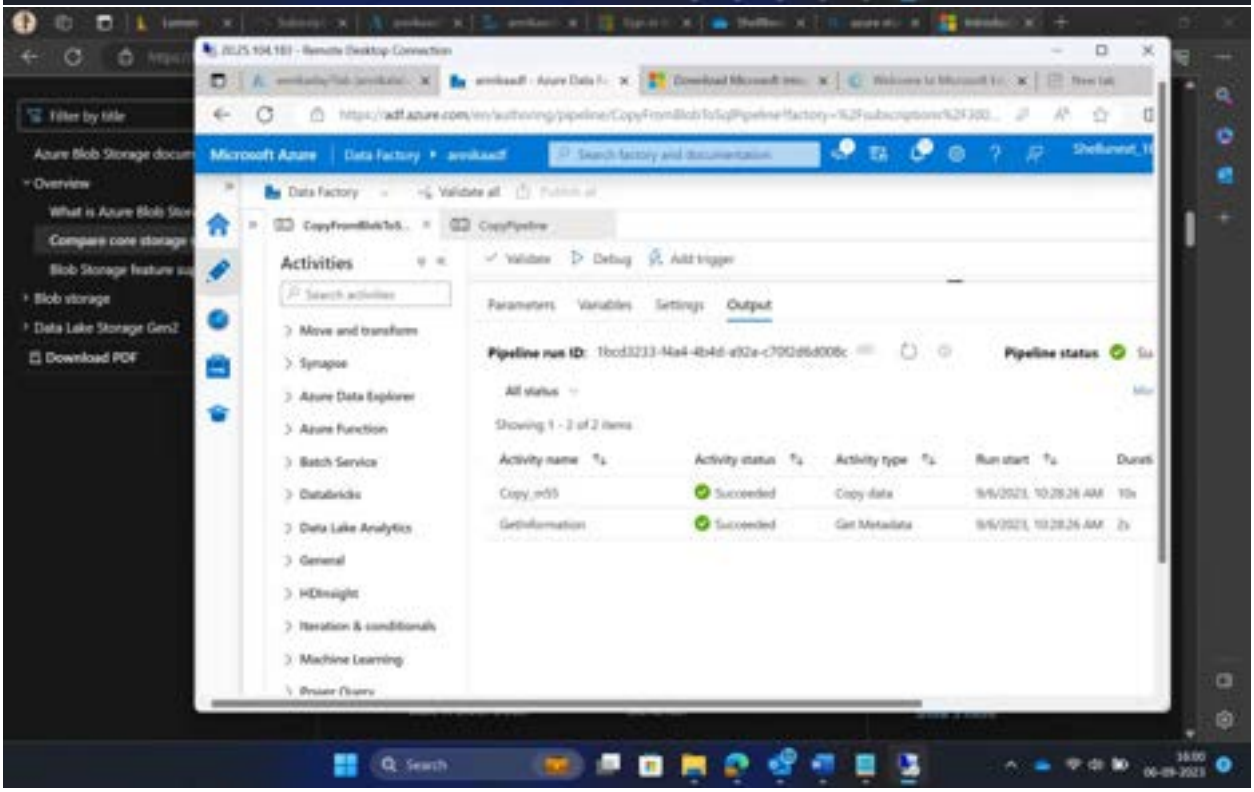

Microsoft Azure
Data Factory
Copy Data tool
Properties
Source
Destination
Dataset
Settings
Review and finish
Destination data store
Destination type
Connection *
Source
Target
inputEmp
dbo
Use existing table

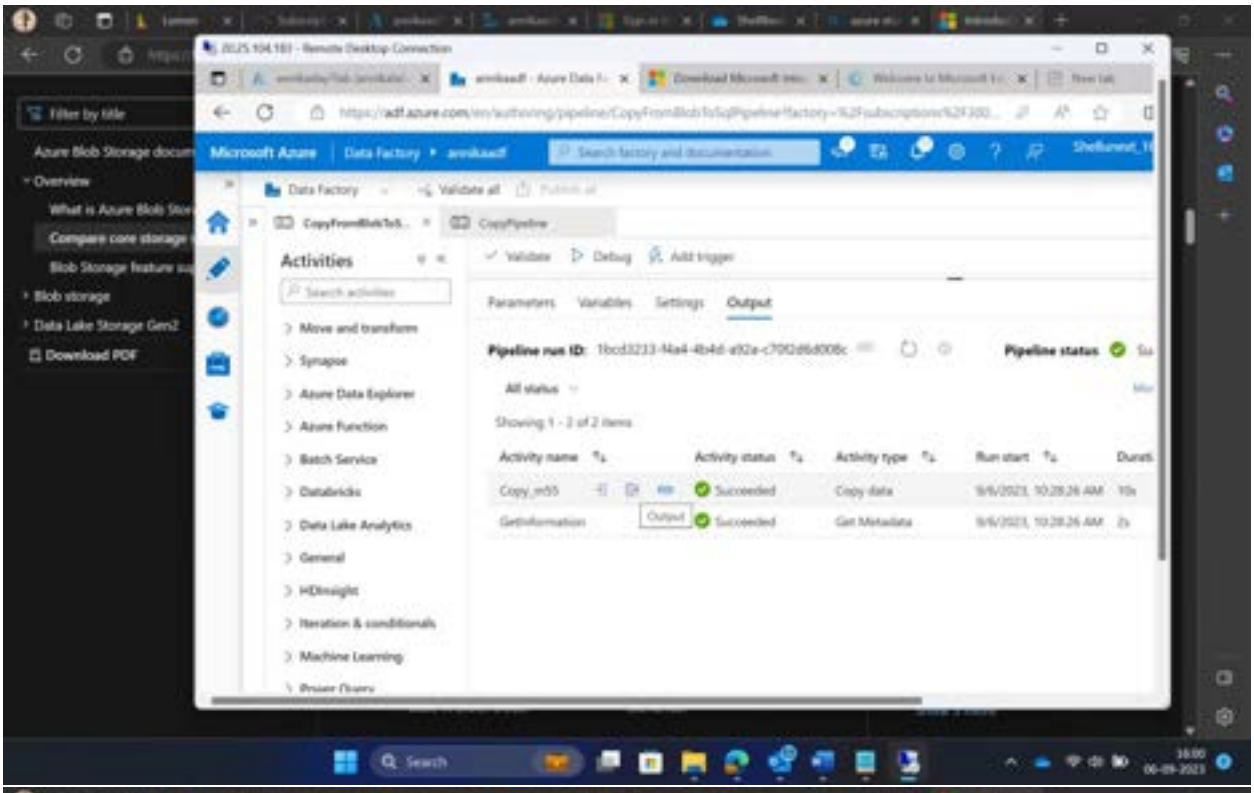

Microsoft Azure
Data Factory
Copy Data tool
Properties
Source
Destination
Dataset
Configuration
Settings
Review and finish
Column mapping
Table mappings (1)
Column mappings
Type conversion settings
New mapping
Clear
Reset
Delete
Source
Type
FirstName
LastName
String
Azure SQL Database sink properties
Pre-copy script

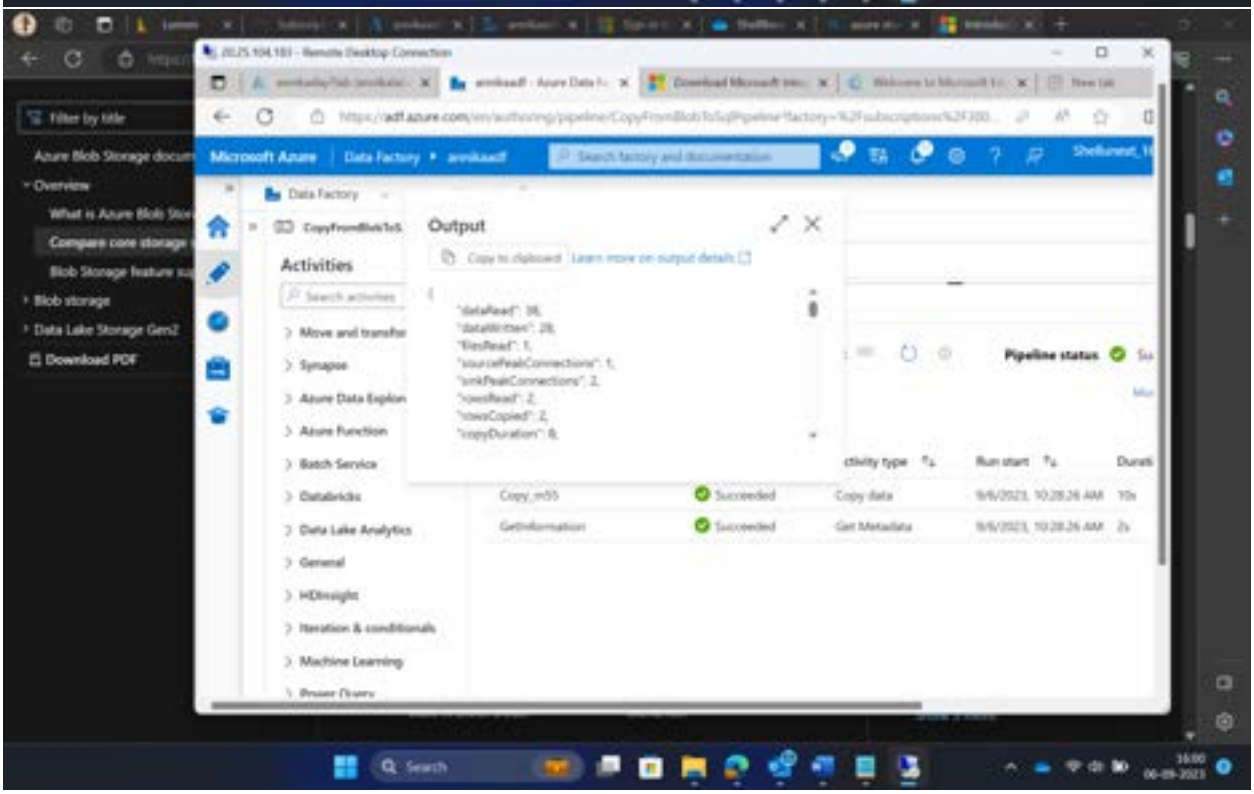

Microsoft Azure
Data Factory
Copy Data tool
Properties
Source
Destination
Settings
Review and finish
Review
Deployment
Azure Blob Storage
Azure SQL D
Deployment complete
Deployment step
Status
Validating copy runtime environment
Succeeded
Creating datasets
Succeeded
Creating pipelines
Succeeded
Running pipelines
Succeeded
Datasets and pipelines have been created. You can now monitor and edit the copy pipelines or click finish to cl


Microsoft Azure
Data Factory
Pipeline runs
Triggered
Debug
Refresh
Edit columns
Last 24 hours
Add filter
Showing 1 - 1 items
name
Run start
Run end
Duration
Triggered by
Status
Manual trigger


Microsoft Azure
Data Factory
Factory Resources
Pipelines
CopyFromBlobToSqlPipeline
CopyPipeline
Change Data Capture (preview)
Datasets
Data flows
Power Query
Activities
Get Metadata
Get Metadata1
Copy data
Copy_m55
General
Settings
User properties
Name *
Description

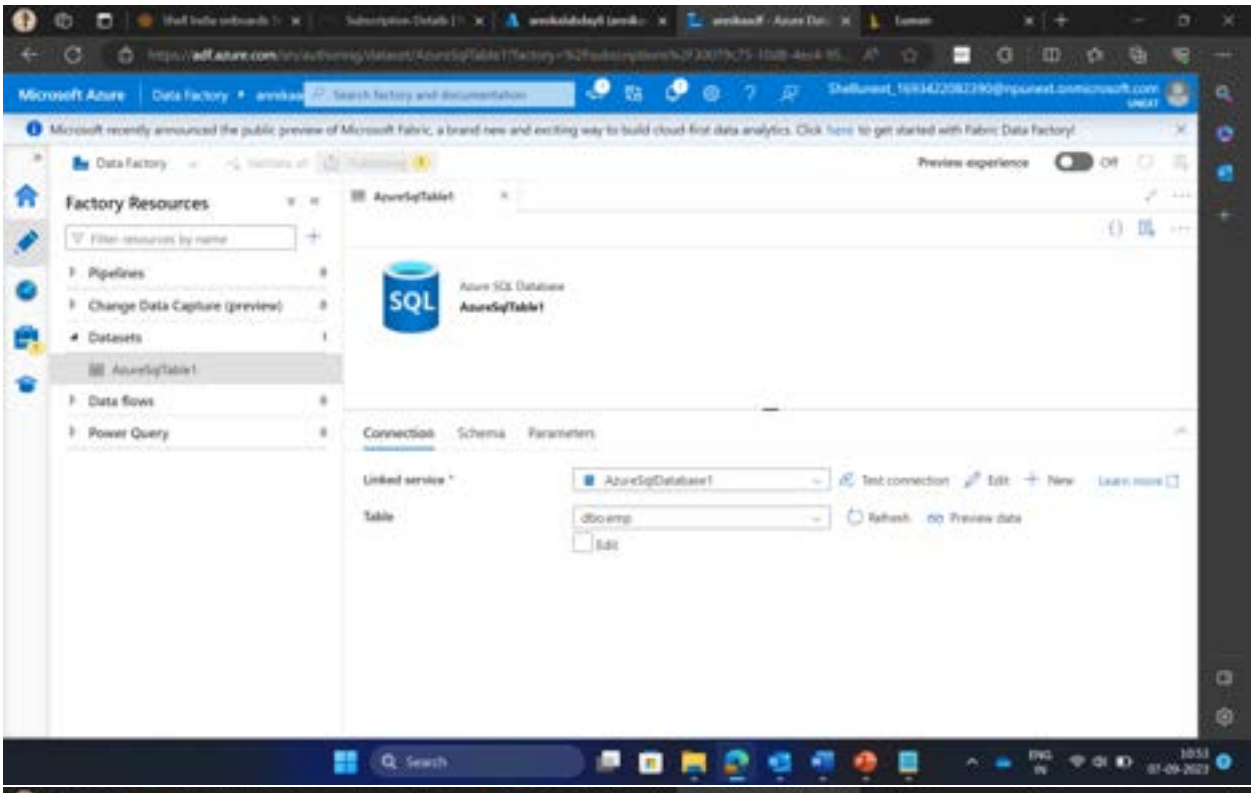

Microsoft Azure
Data Factory
Activities
Parameters
Variables
Settings
Output
Pipeline run ID:
Pipeline status
Activity name
Activity status
Activity type
Run start
Copy_m55
Succeeded
Copy data
GetInformation
Succeeded
Get Metadata

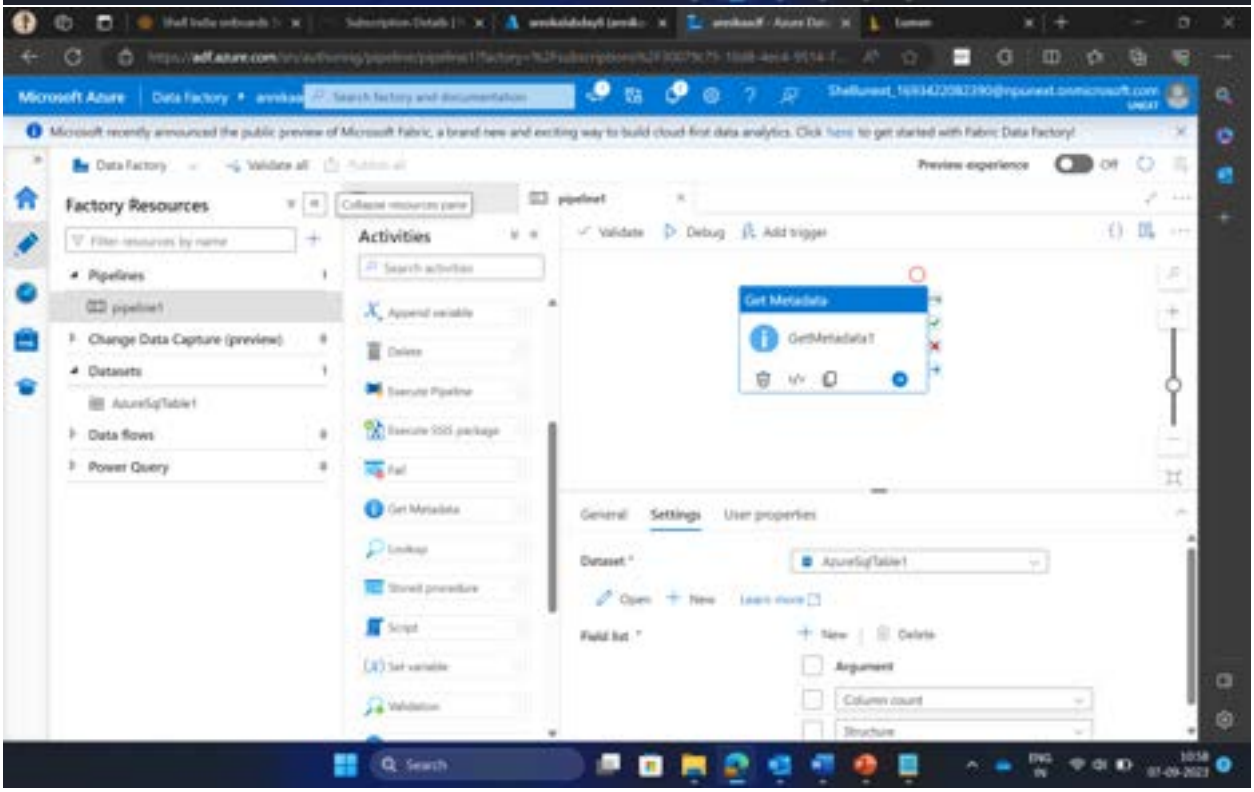

Microsoft Azure
Data Factory
Activities
Parameters
Variables
Settings
Output
Pipeline run ID:
Pipeline status
Activity name
Activity status
Activity type
Run start
Copy_m55
Succeeded
Copy data
GetInformation
Succeeded
Get Metadata

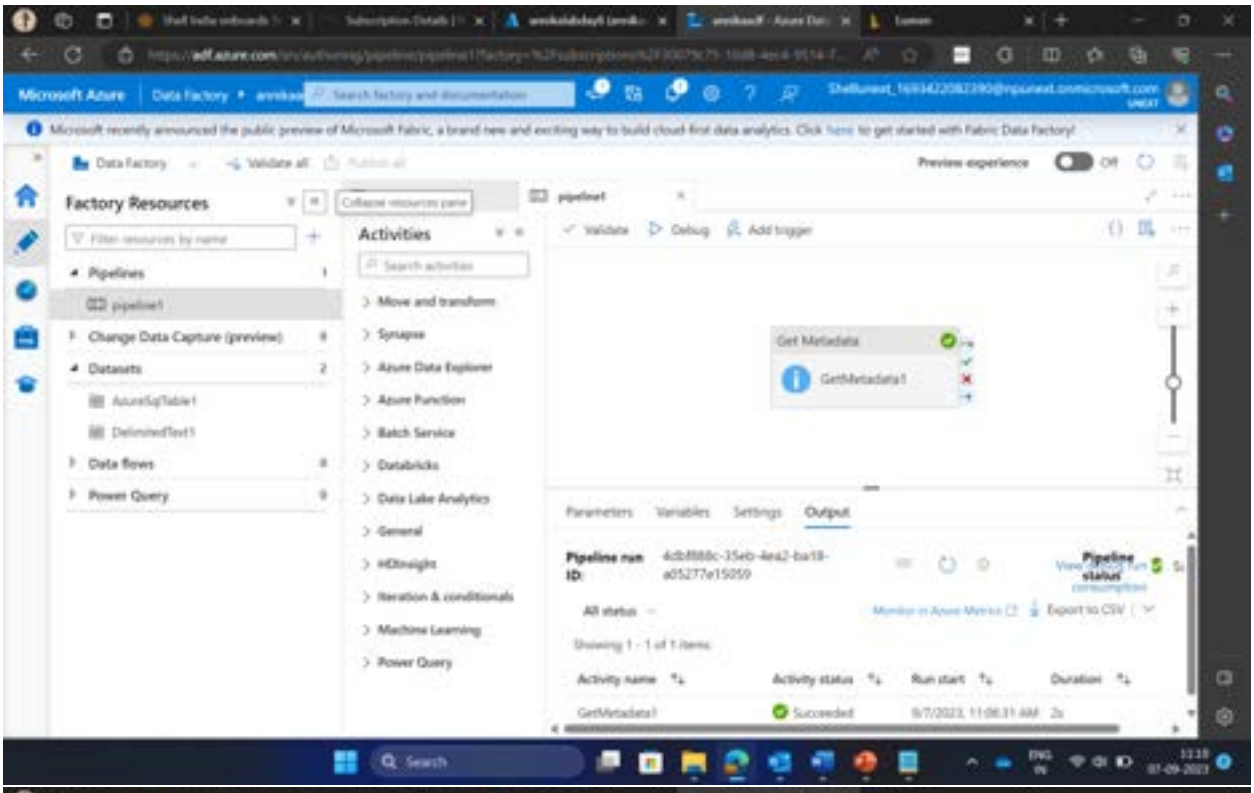

Microsoft Azure
Data Factory
Output
Activities
"dataRead": 38,
"dataWritten": 28,
"filesRead": 1,
"sourcePeakConnections": 1,
"sinkPeakConnections": 2,
"rowsRead": 2,
"rowsCopied": 2,
"copyDuration": 8,
Pipeline status
Copy_m55
Succeeded
Copy data
GetInformation
Succeeded
Get Metadata

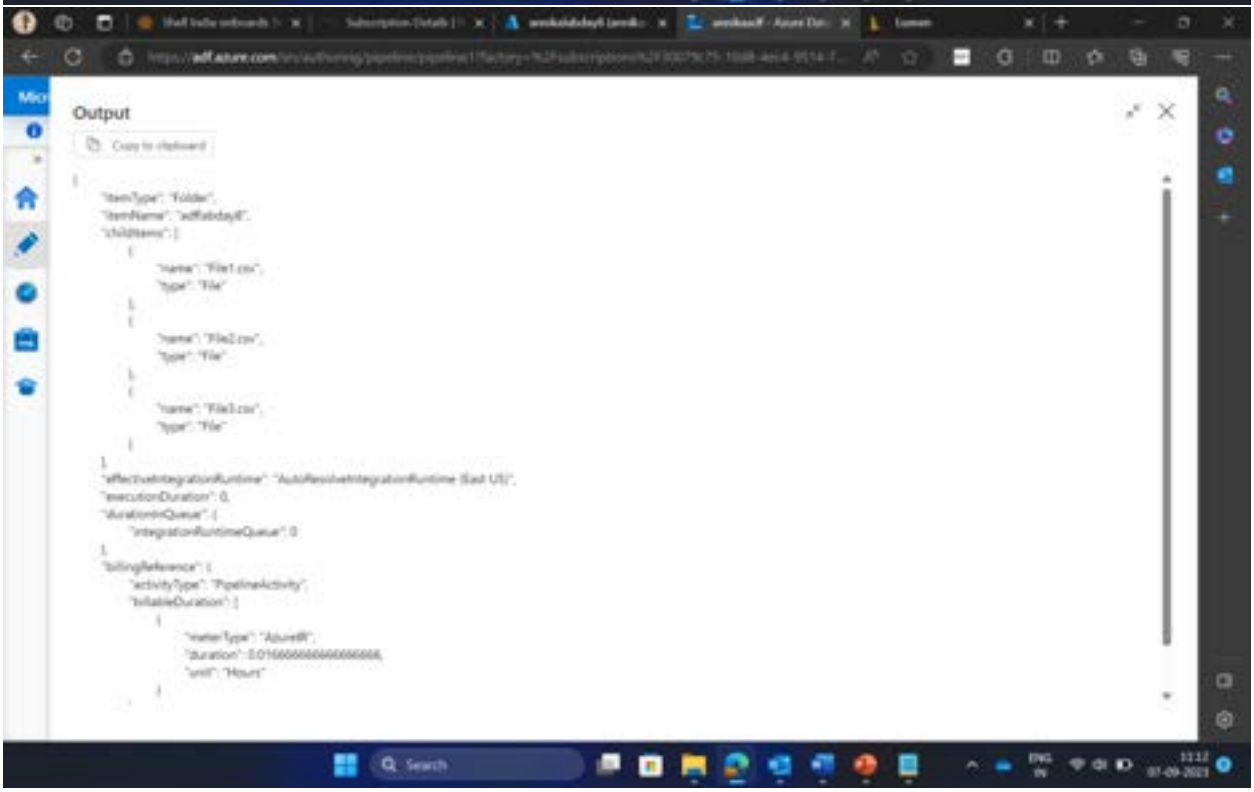
Day 8

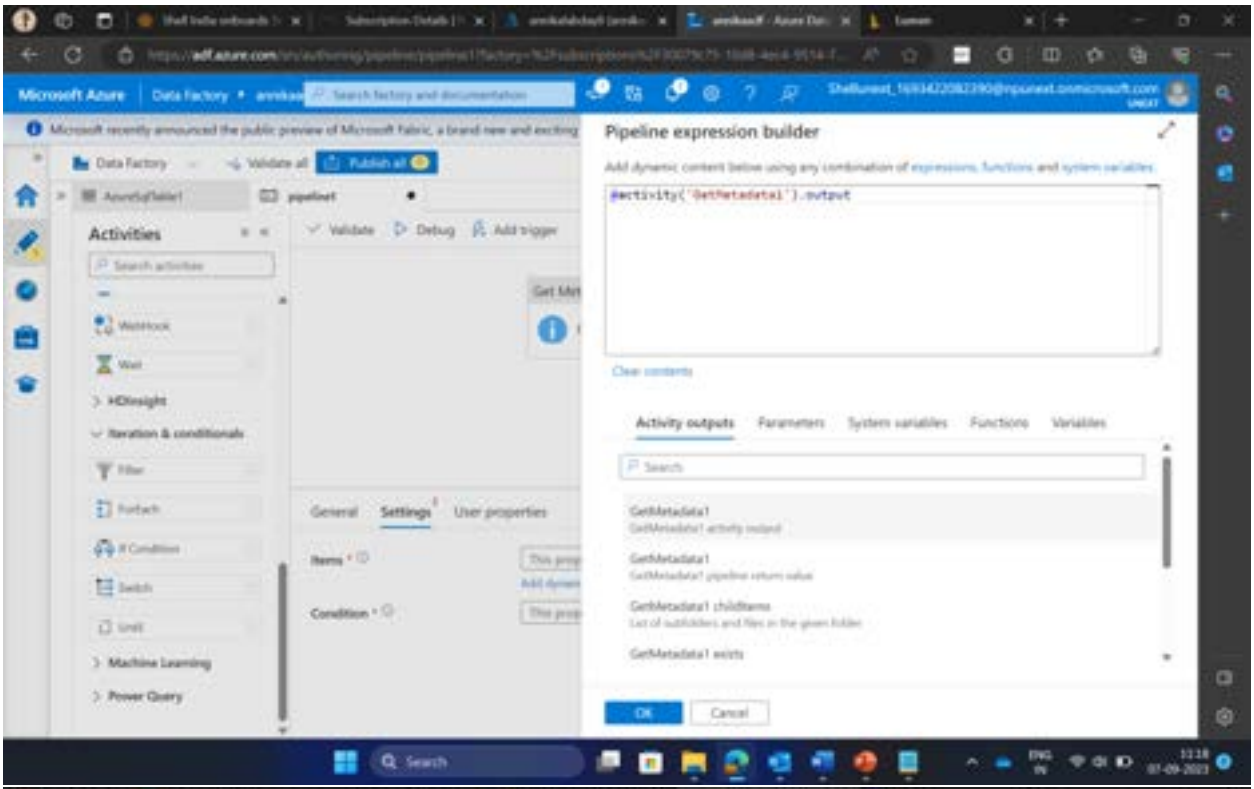
```sql
create table emp
(
id int,
name varchar(20),
city varchar(20)
)
```

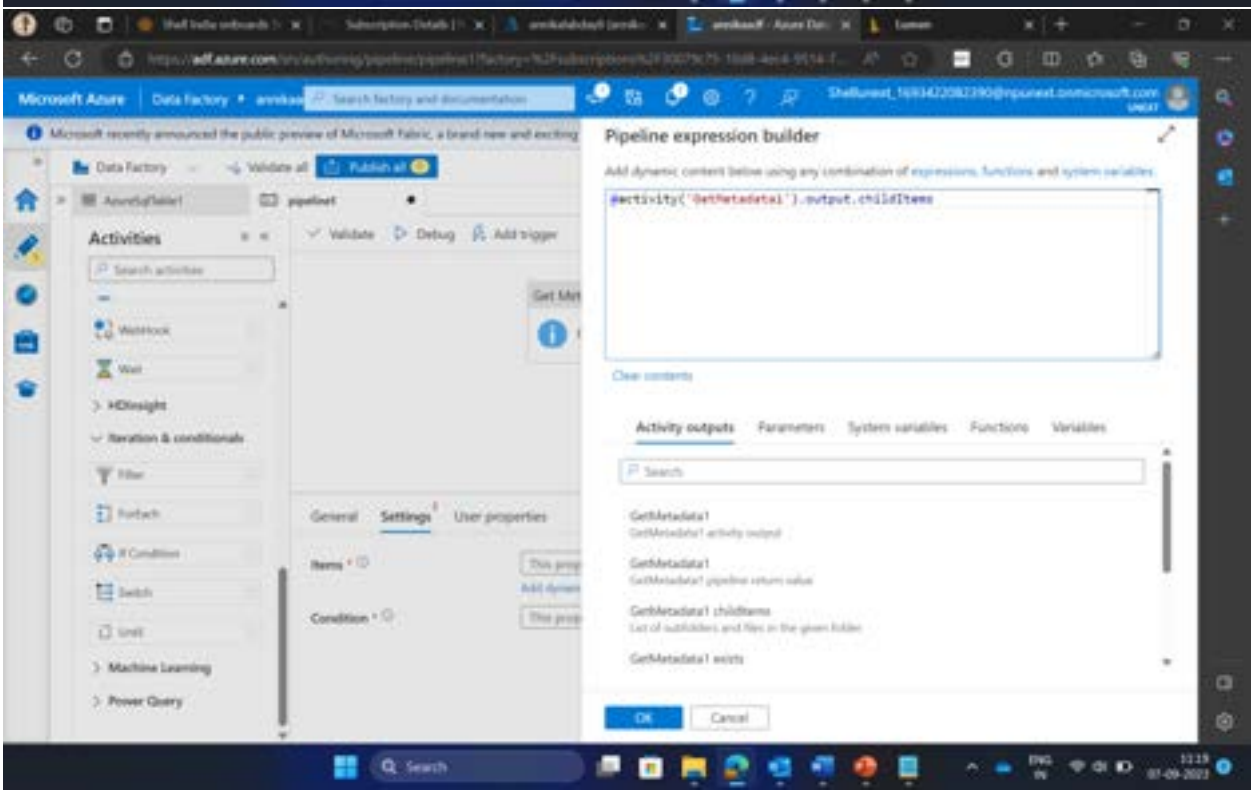

Microsoft Azure
Data Factory
Factory Resources
Pipelines
Change Data Capture (preview)
Datasets
AzureSqlTable1
Data flows
Power Query
Preview experience
Azure SQL Database
AzureSqlTable1
Connection
Schema
Parameters
Linked service *
AzureSqlDatabase1
Test connection
Edit
New
Table
dbo.emp
Refresh
Preview data

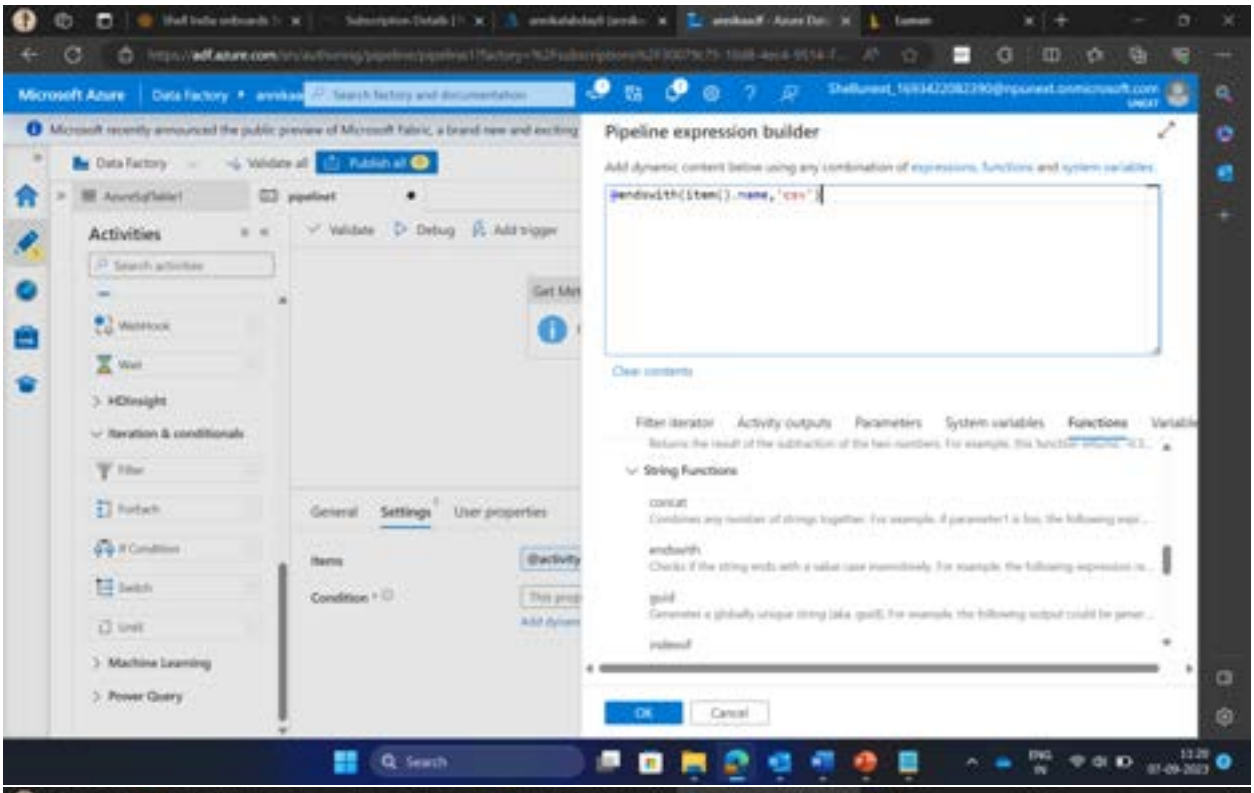

Microsoft Azure
Data Factory
Validate all
Factory Resources
Pipelines
pipeline1
Change Data Capture (preview)
Datasets
AzureSqlTable1
Data flows
Power Query
Activities
Validate
Debug
Add trigger
Get Metadata
GetMetadata1
General
Settings
User properties
Dataset *
AzureSqlTable1
Open
New
Field list *
Argument
Column count
Structure

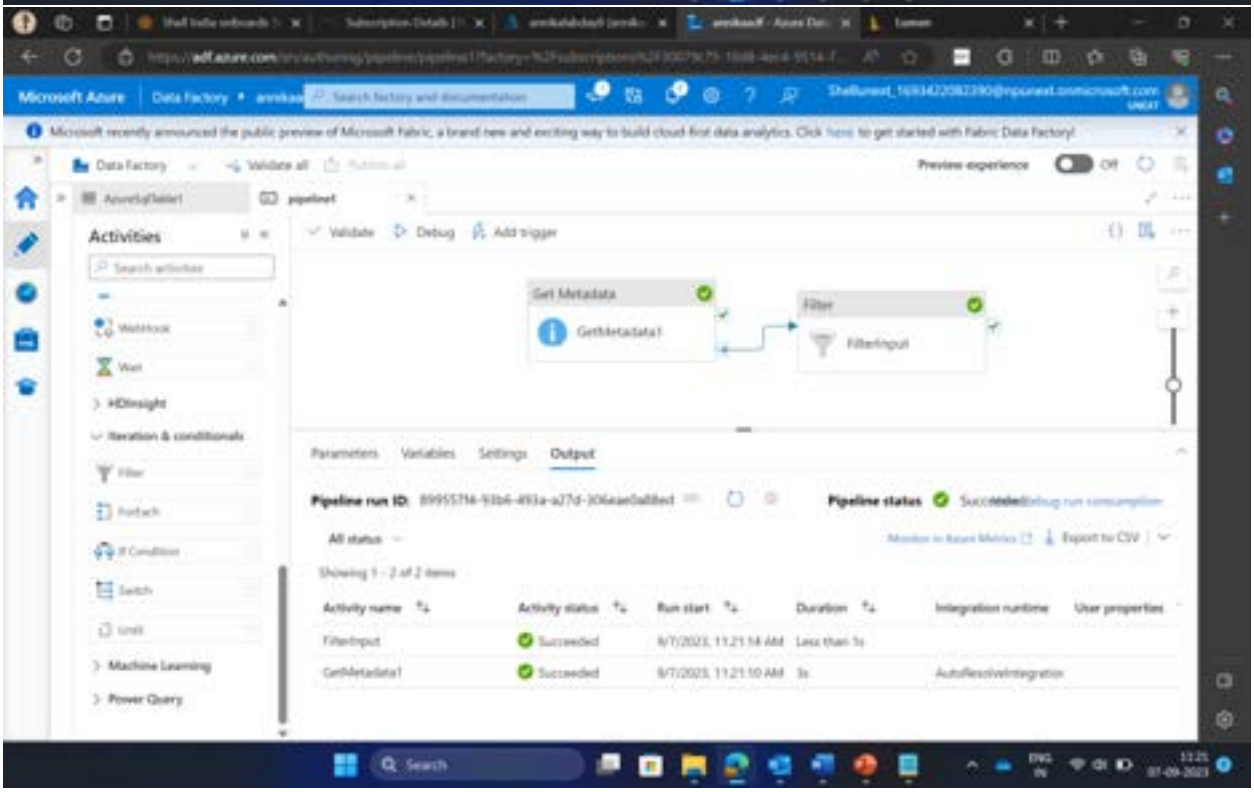

Microsoft Azure
Data Factory
Factory Resources
Pipelines
pipeline1
Change Data Capture (preview)
Datasets
AzureSqlTable1
DelimitedText1
Data flows
Power Query
Activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Debug
Add trigger
Get Metadata
GetMetadata1
Parameters
Variables
Settings
Output
Pipeline run ID:
4dbf88dc-35eb-4ea2-ba18-a05277e15059
Activity name
Activity status
Run start
Duration
GetMetadata1
Succeeded

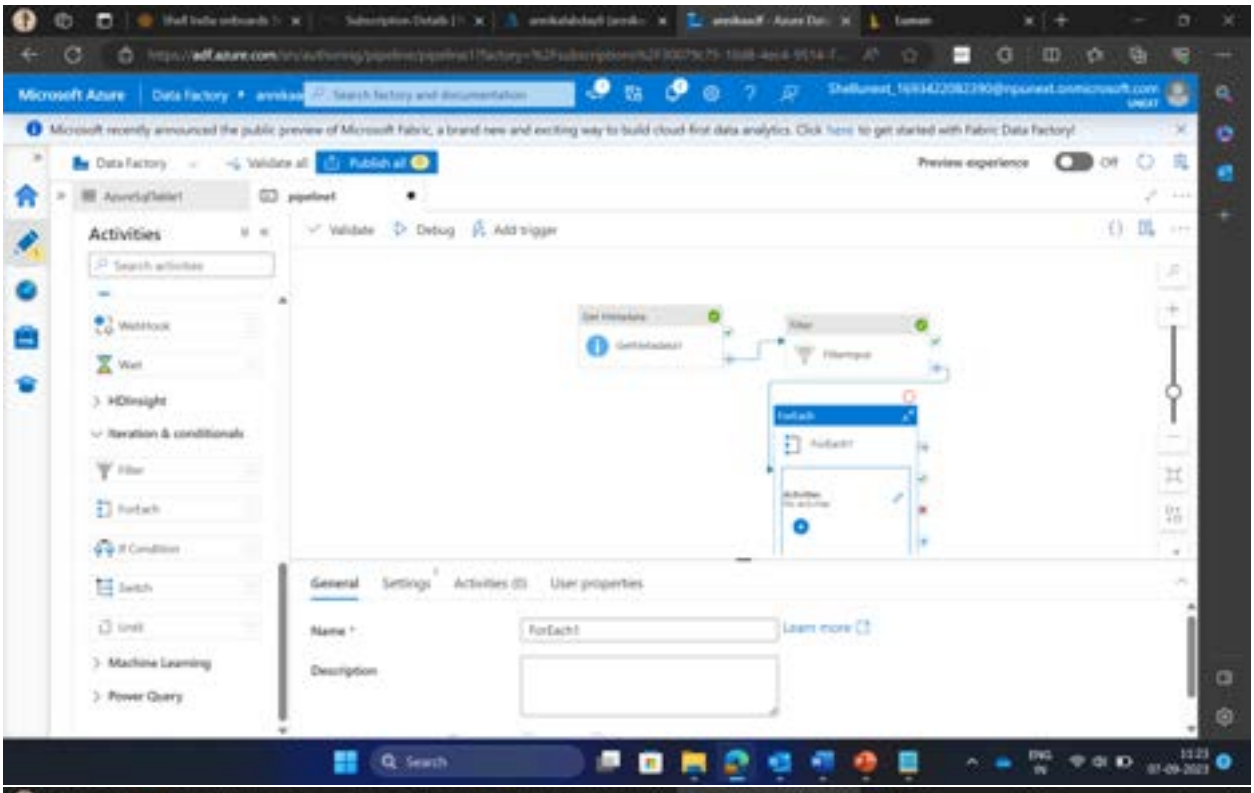

Output
{
"itemType": "Folder",
"itemName": "adfdatadf",
"childItems": [
{
"name": "File1.csv",
"type": "File"
},
{
"name": "File2.csv",
"type": "File"
},
{
"name": "File3.csv",
"type": "File"
}
],
"effectiveIntegrationRuntime": "AutoResolveIntegrationRuntime (East US)",
"executionDuration": 0,
"durationInQueue": {
"integrationRuntimeQueue": 0
},
"billingReference": {
"activityType": "PipelineActivity",
"billableDuration": [
{
"meterType": "AzureIR",
"duration": 0.016666666666666668,
"unit": "Hours"
}

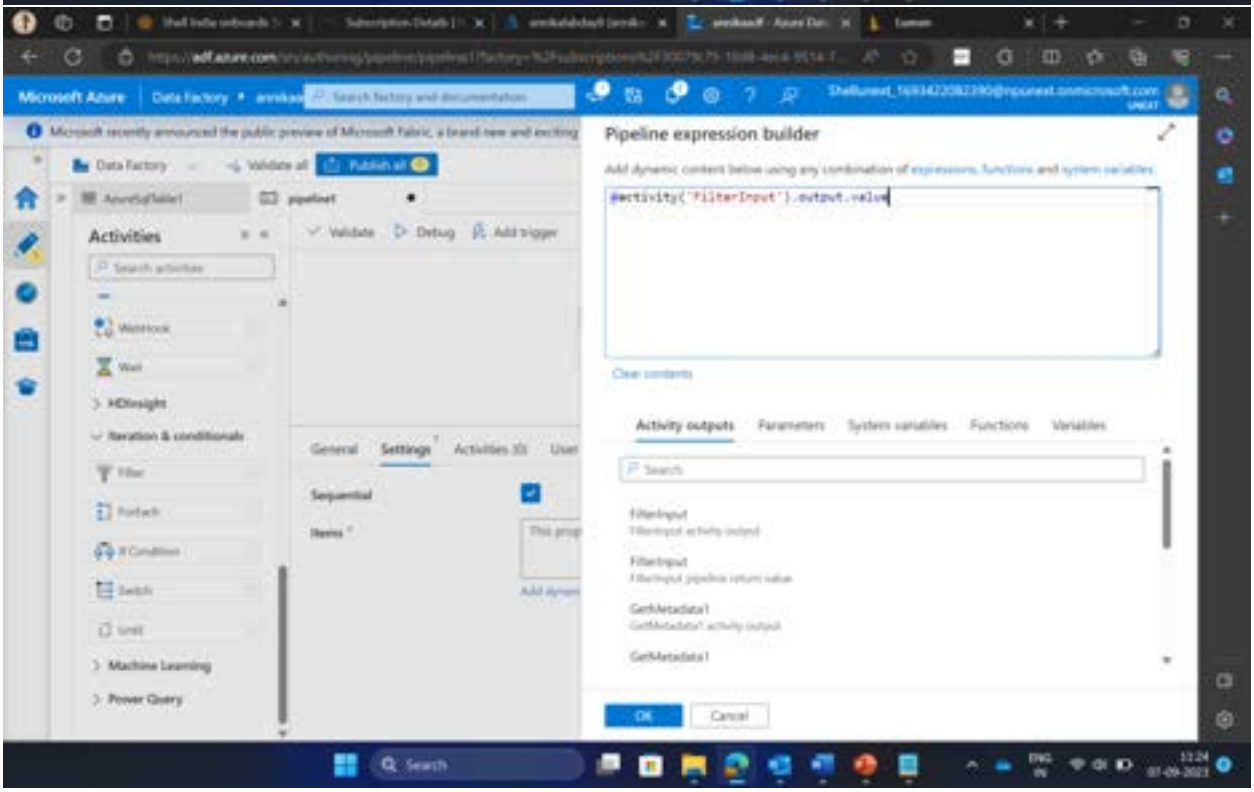

Pipeline expression builder
@activity('GetMetadata1').output
Activity outputs
Parameters
System variables
Functions
Variables
GetMetadata1
GetMetadata1 childItems
GetMetadata1 exists
OK
Cancel

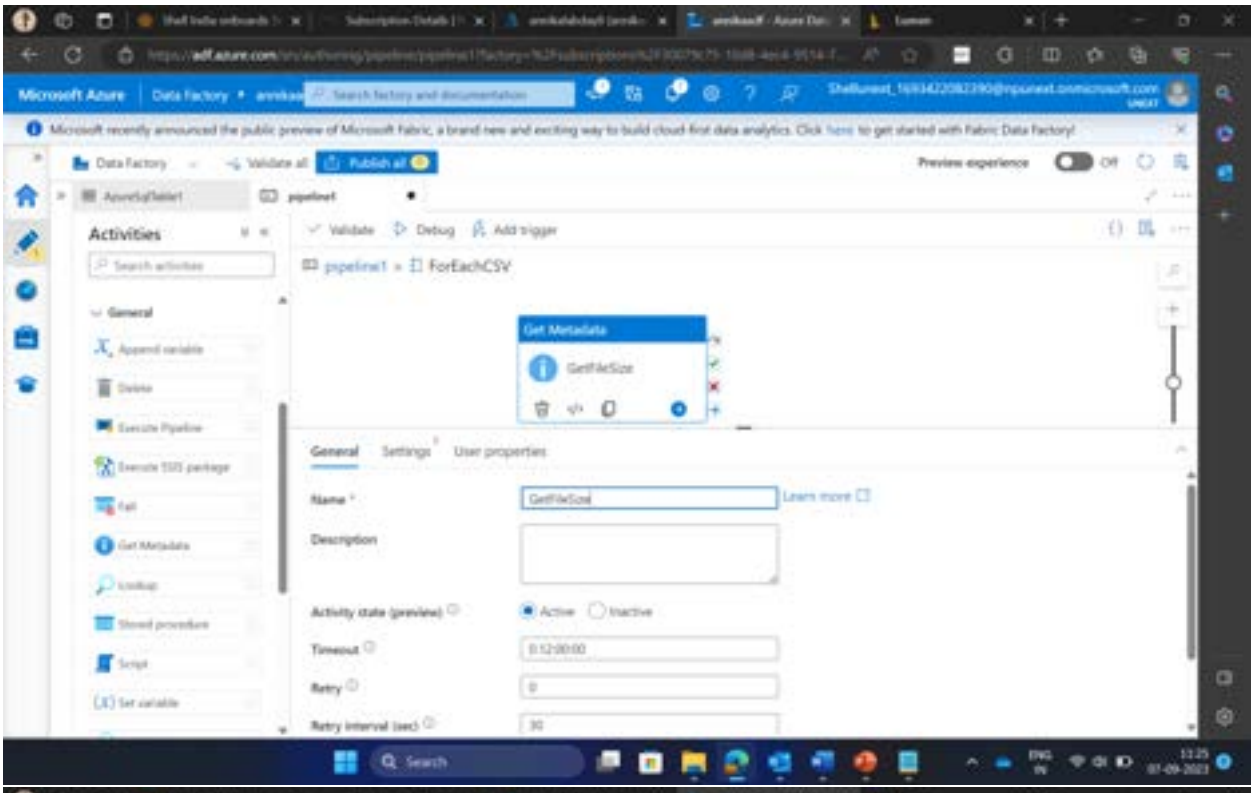

Pipeline expression builder
@activity('GetMetadata1').output.childItems
Activity outputs
Parameters
System variables
Functions
Variables
GetMetadata1
GetMetadata1 childItems
GetMetadata1 exists
OK
Cancel

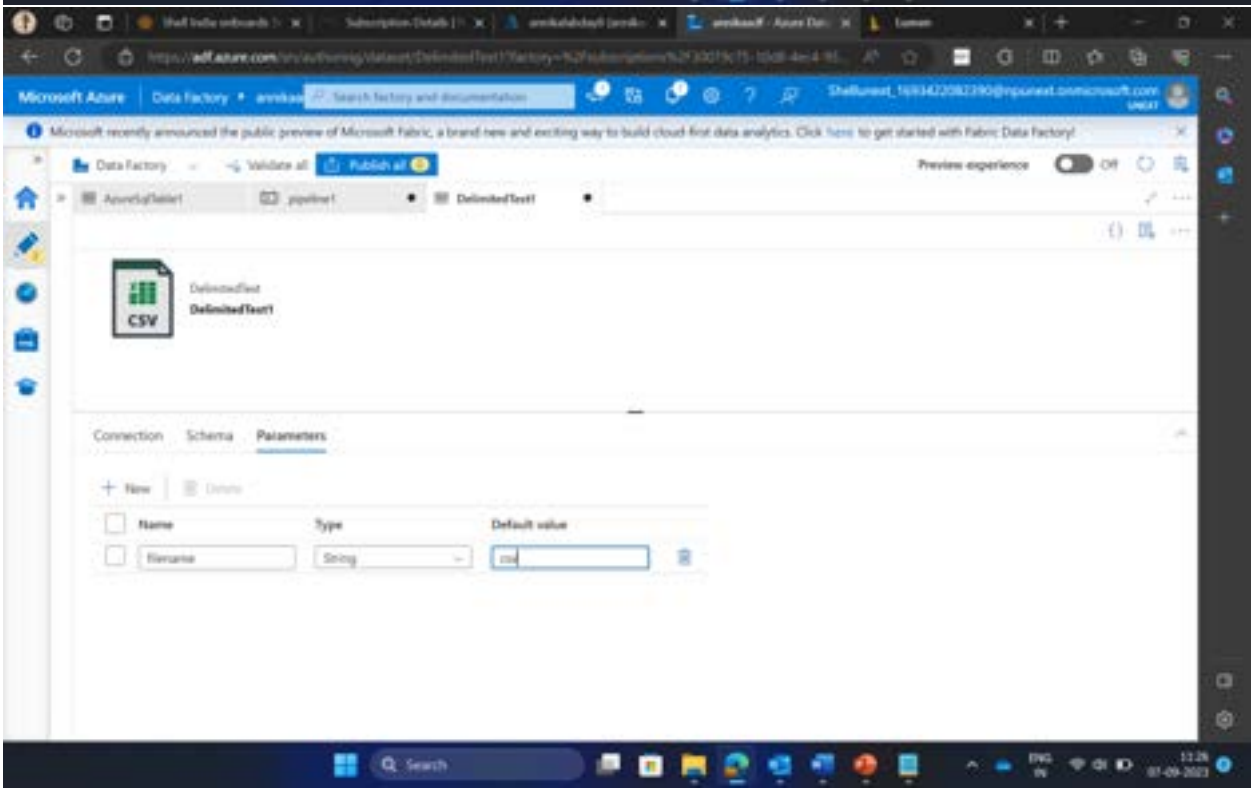

Pipeline expression builder
@endswith(item().name,'csv')
Activities
Validate
Debug
Add trigger
General
Settings
User properties
Items
Condition
String Functions
concat
endswith
guid
OK
Cancel

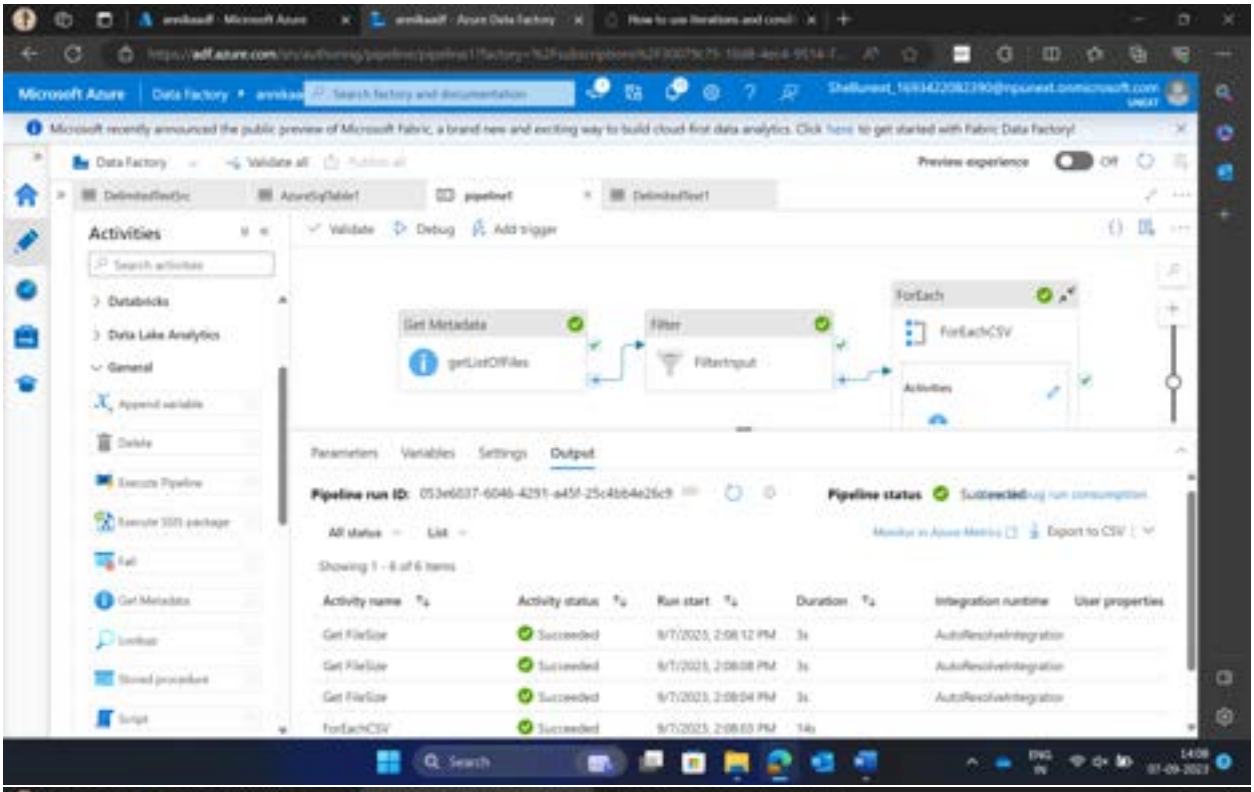

Activities
Validate
Debug
Add trigger
Get Metadata
GetMetadata1
Filter
FilterInput
Parameters
Variables
Settings
Output
Pipeline run ID:
Pipeline status
Activity name
Activity status
Run start
Duration
Integration runtime
User properties
FilterInput
GetMetadata1
Succeeded

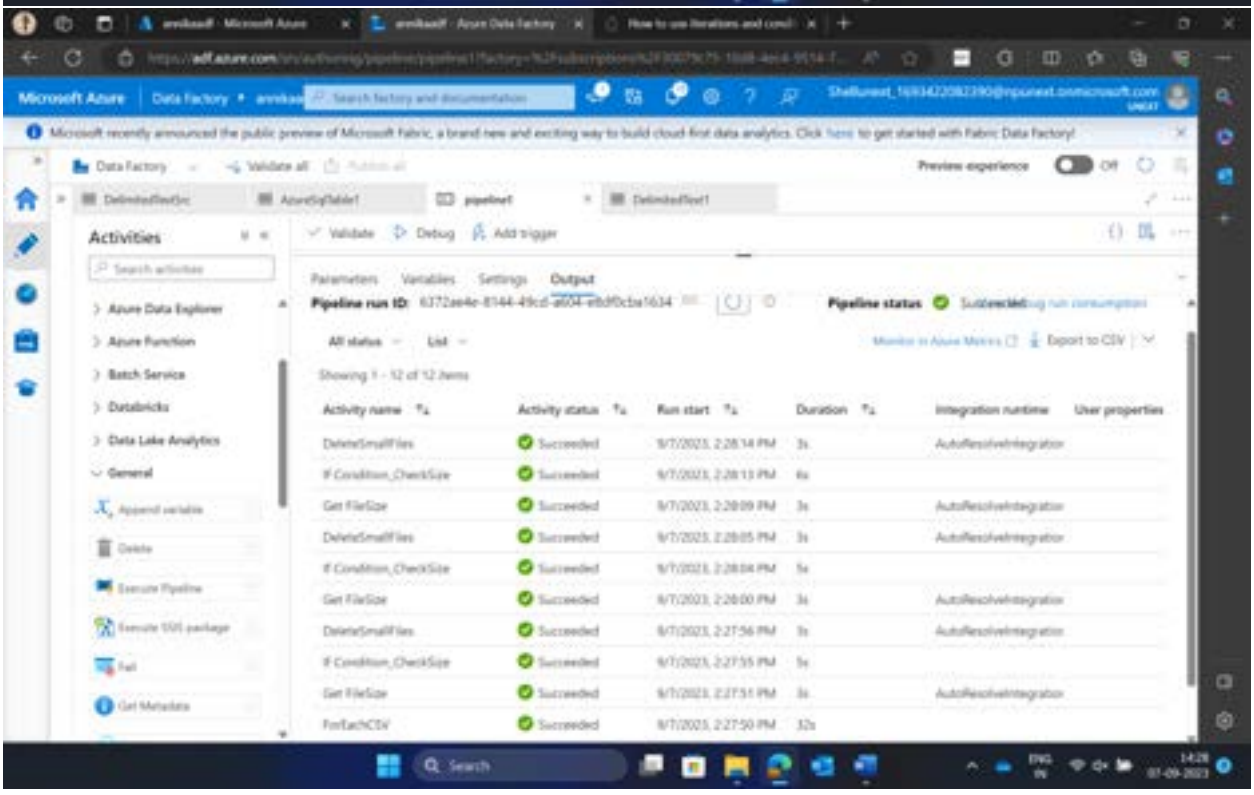

Microsoft Azure
Data Factory
Activities
pipeline1
Validate
Debug
Add trigger
General
Settings
Activities (0)
User properties
Name *
ForEach1
Description
Learn more
Machine Learning
Power Query

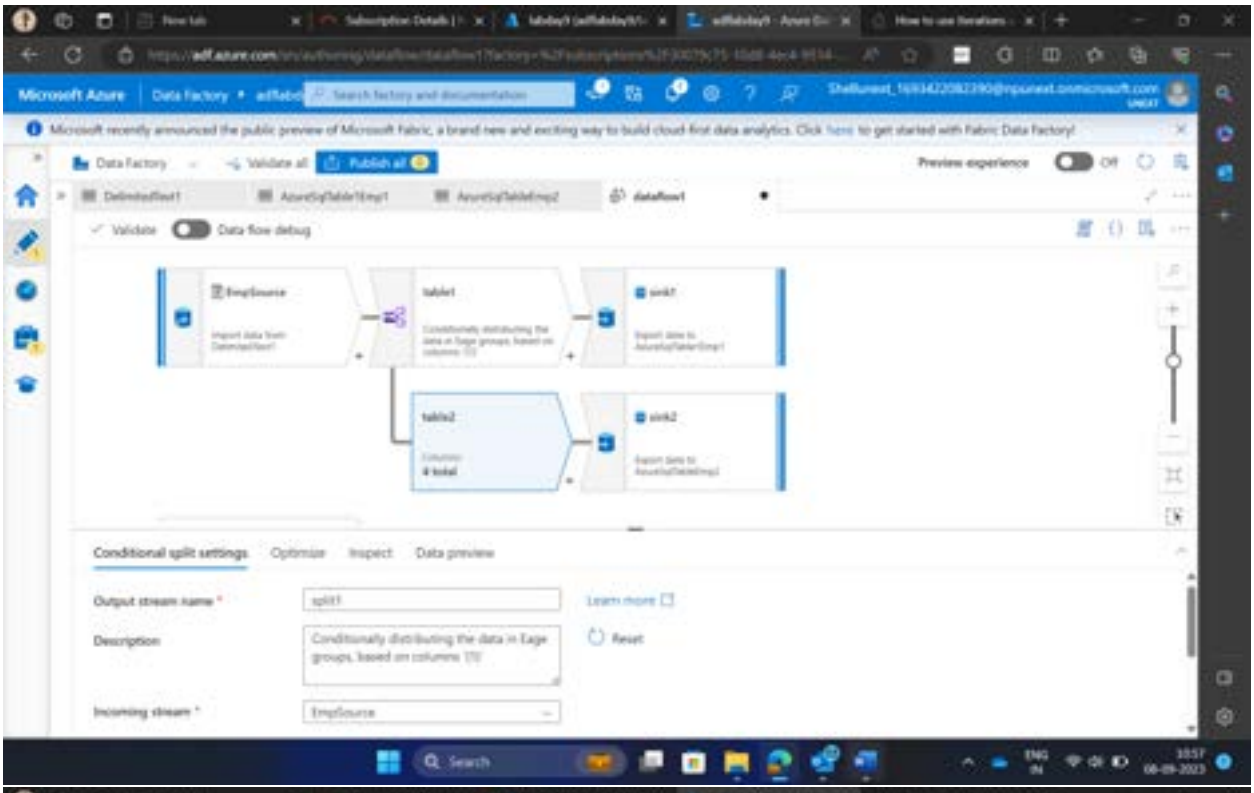

Microsoft Azure
Data Factory
Activities
pipeline1
Pipeline expression builder
Add dynamic content below using any combination of expressions, functions and system variables.
@activity('FilterInput').output.value
Clear contents
Activity outputs
Parameters
System variables
Functions
Variables
Search
FilterInput
GetMetadata1
OK
Cancel
Settings
Sequential
Items *
Machine Learning
Power Query

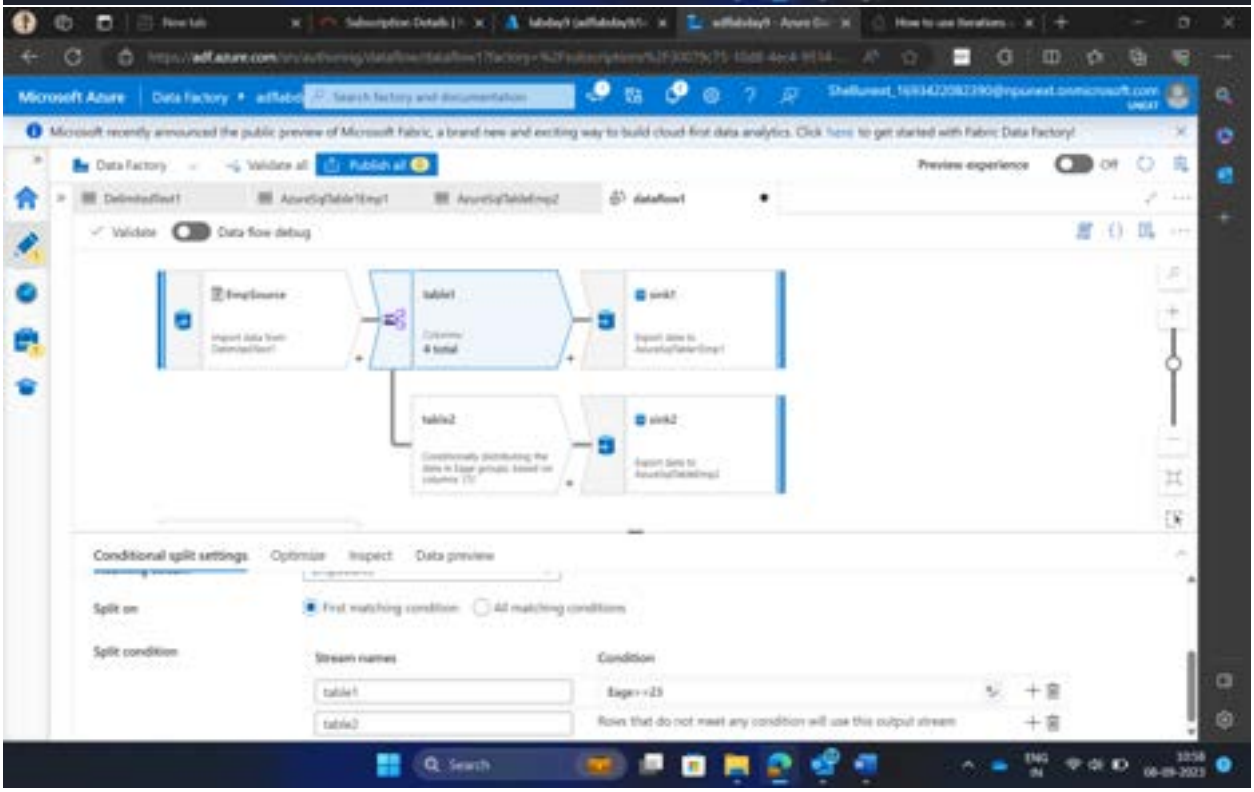

Microsoft Azure
Data Factory
Activities
pipeline1
ForEachCSV
Get Metadata
GetFileSize
General
Settings
User properties
Name
Description
Activity state (preview)
Timeout
Retry
Retry interval (sec)

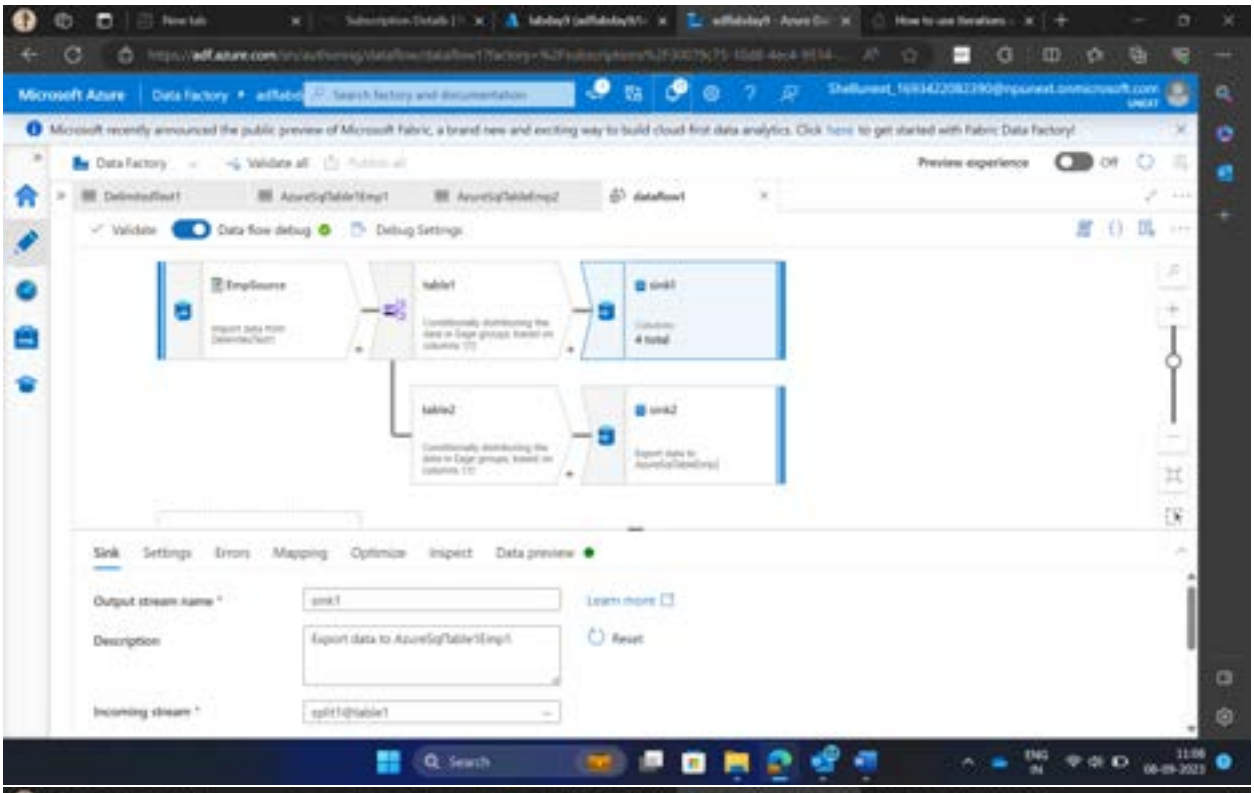

Microsoft Azure
Data Factory
DelimitedText1
CSV
Connection
Schema
Parameters
New
Name
Type
Default value
Filename
String

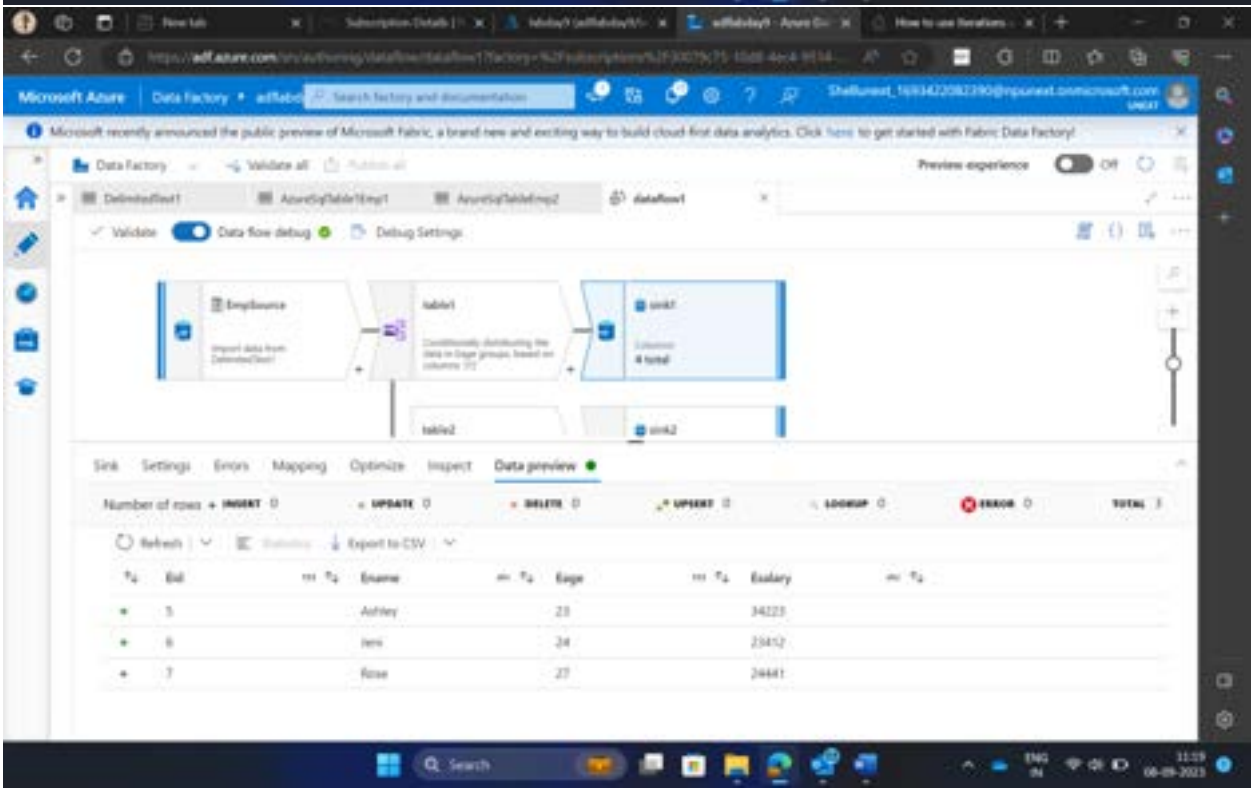

Microsoft Azure
Data Factory
Activities
Get Metadata
Filter
ForEach
ForEachCSV
Parameters
Variables
Settings
Output
Pipeline run ID:
Pipeline status
Activity name
Activity status
Run start
Duration
Integration runtime
User properties

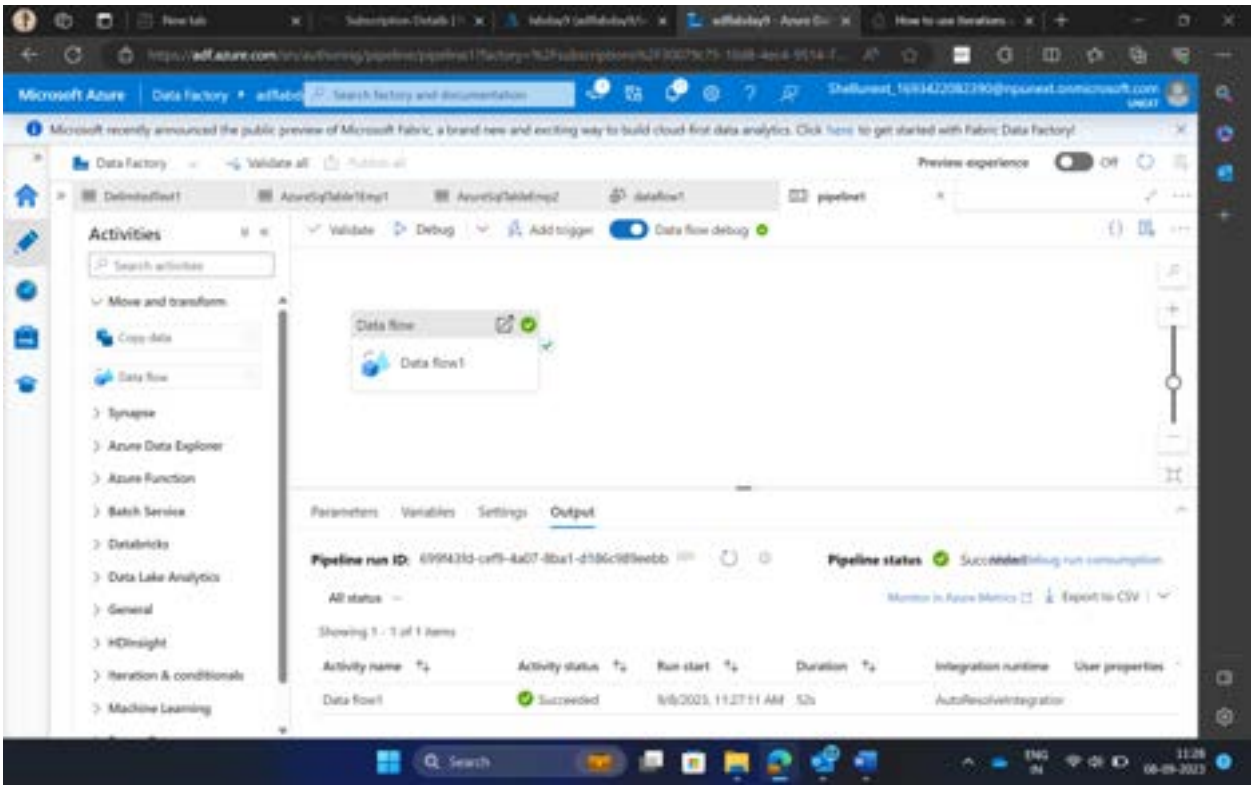

Microsoft Azure
Data Factory
Activities
Parameters
Variables
Settings
Output
Pipeline run ID:
Pipeline status
Activity name
Activity status
Run start
Duration
Integration runtime
User properties

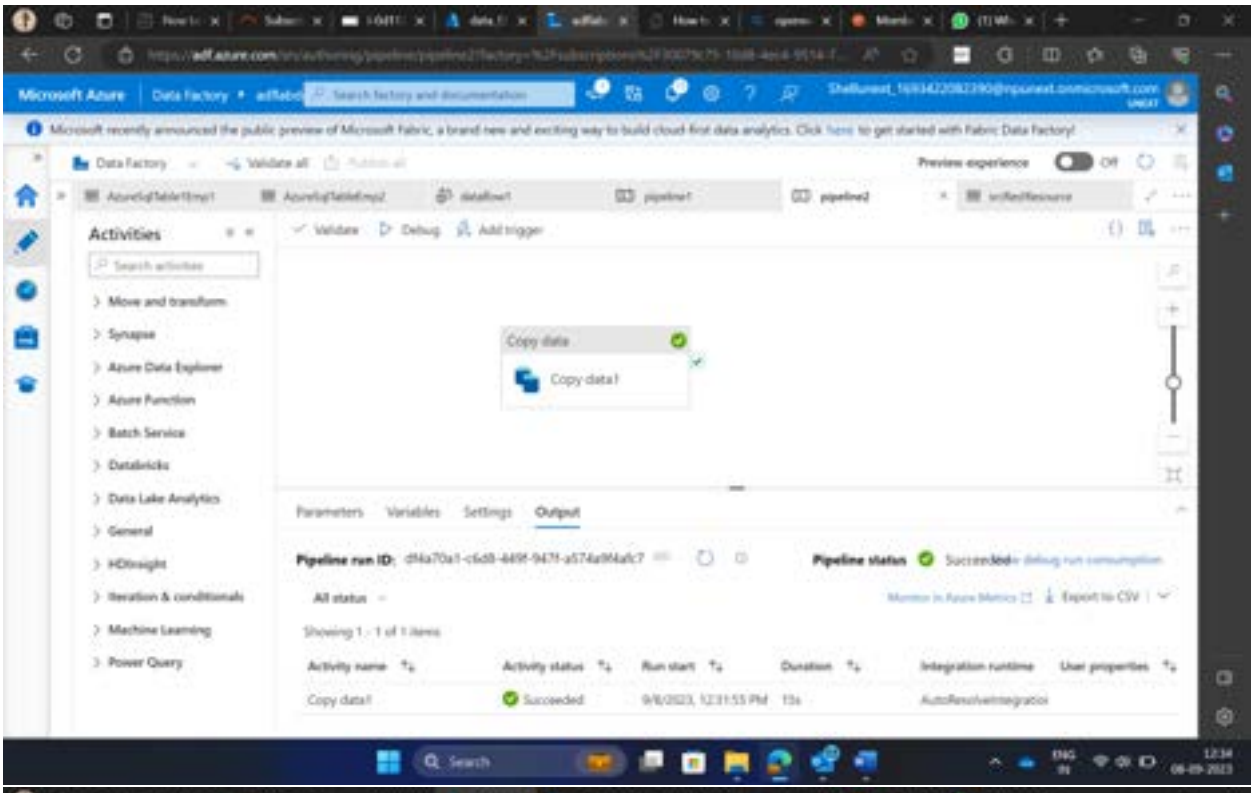

Microsoft Azure
Data Factory
Publish all
Preview experience
Data flow debug
EmpSource
table1
table2
sink1
sink2
Conditional split settings
Optimize
Inspect
Data preview
Output stream name *
split1
Learn more
Description
Conditionally distributing the data in Eage groups, based on columns
Reset
Incoming stream *
EmpSource

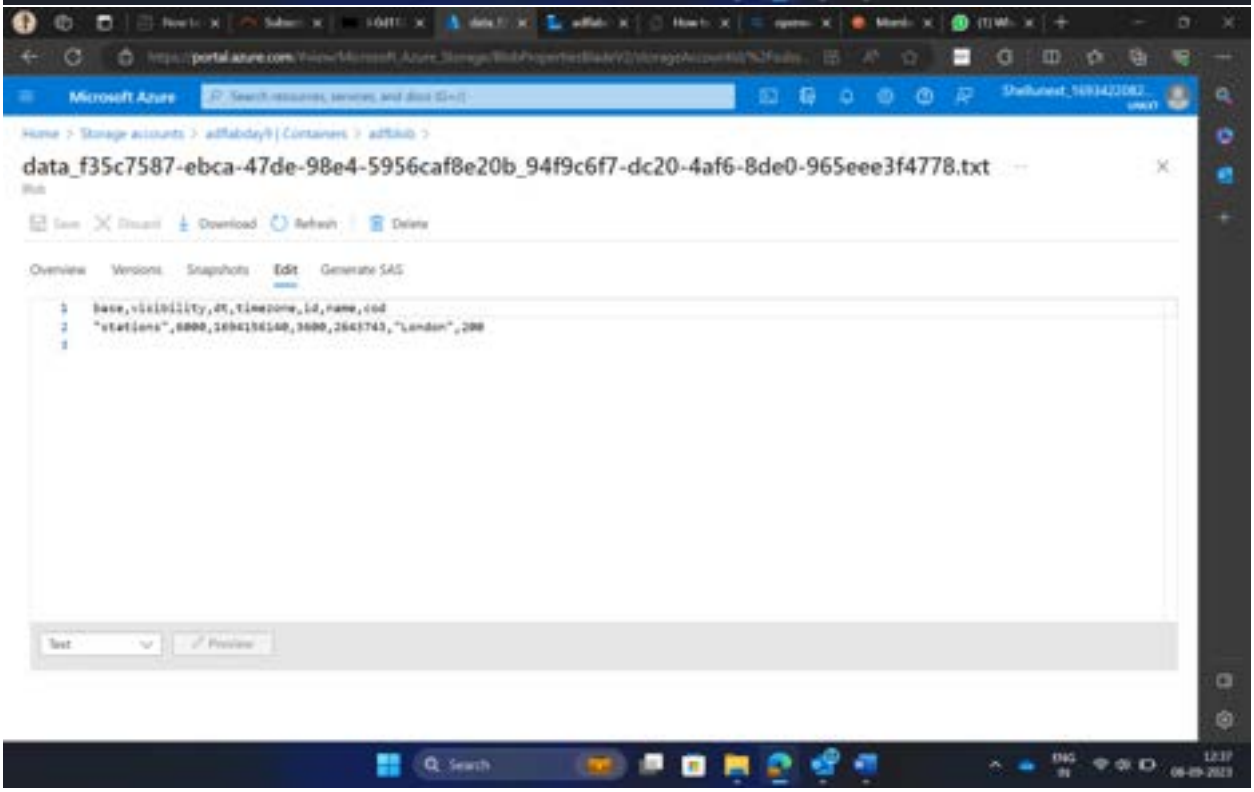

Microsoft Azure
Data Factory
Publish all
Preview experience
Data flow debug
EmpSource
table1
4 total
table2
sink1
sink2
Conditional split settings
Optimize
Inspect
Data preview
Split on
First matching condition
All matching conditions
Split condition
Stream names
Condition
table1
Eage<=25
table2
Rows that do not meet any condition will use this output stream


Microsoft Azure
Data Factory
Data Factory
Validate all
Preview experience
Off
Validate
Data flow debug
Debug Settings
EmpSource
table1
sink1
4 total
table2
sink2
Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Output stream name *
sink1
Learn more
Description
Export data to AzureSqlTable1Emp1
Reset
Incoming stream *
split1@table1

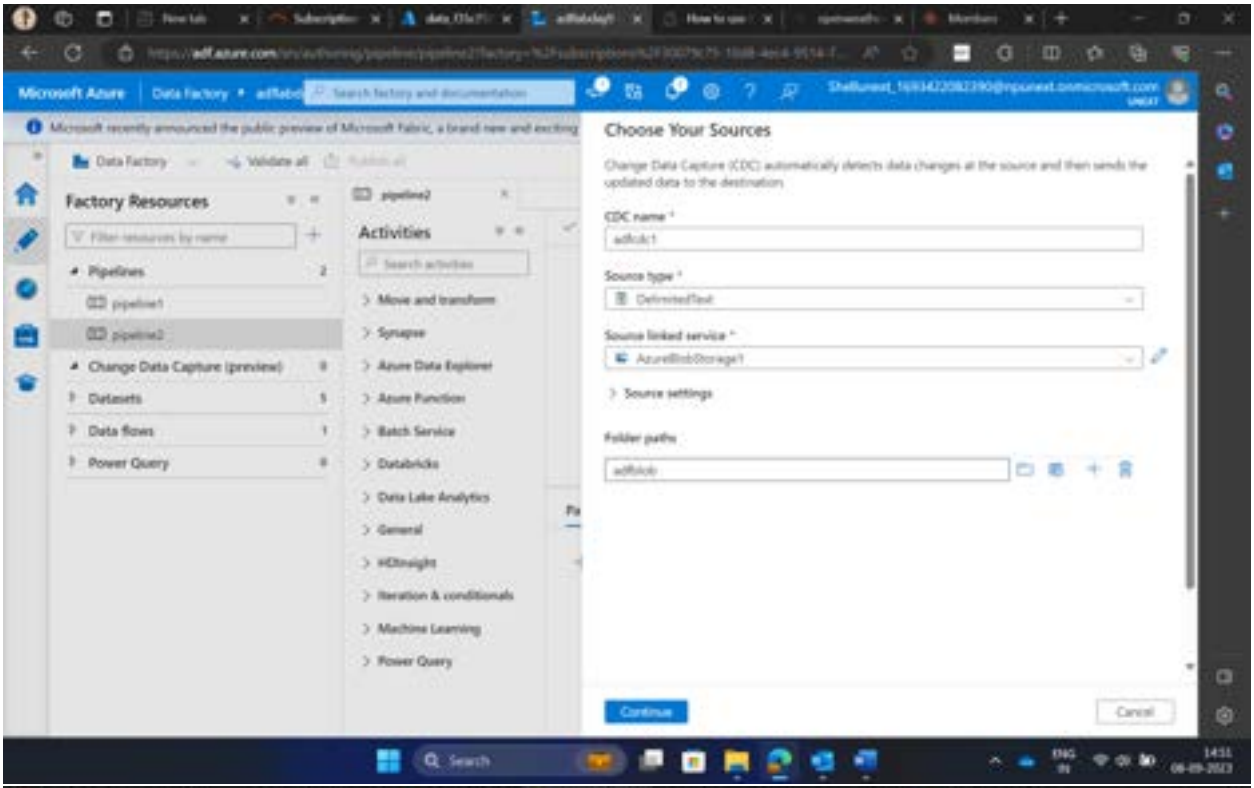

Microsoft Azure
Data Factory
Data Factory
Validate all
Preview experience
Off
Validate
Data flow debug
Debug Settings
EmpSource
table1
sink1
4 total
table2
sink2
Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Number of rows
INSERT 0
UPDATE 0
DELETE 0
UPSERT 0
LOOKUP 0
ERROR 0
TOTAL 3
Refresh
Export to CSV
Eid
Ename
Eage
Esalary
5
Ashley
23
34223
6
Jeni
24
23452
7
Rose
27
24441

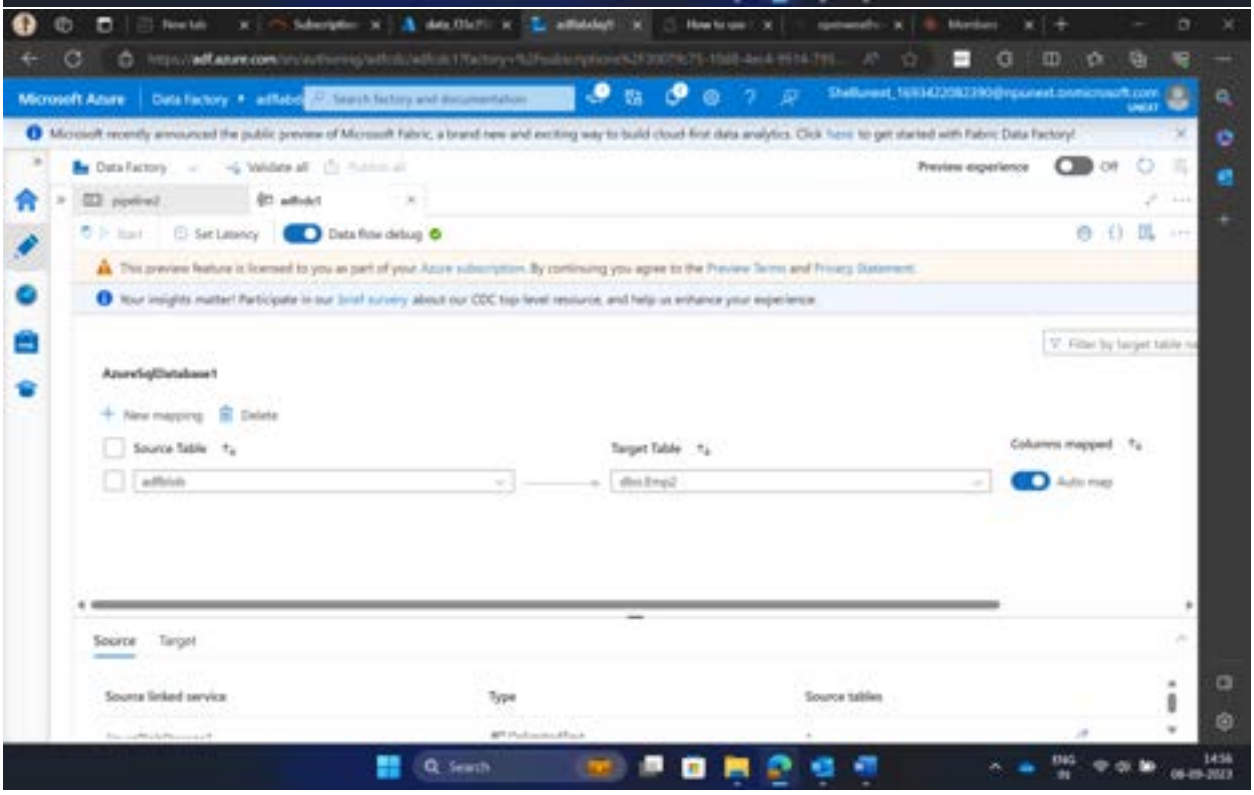

Microsoft Azure
Data Factory
Activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Validate
Debug
Add trigger
Data flow debug
Data flow
Data flow1
Parameters
Variables
Settings
Output
Pipeline run ID:
Pipeline status
Activity name
Activity status
Run start
Duration
Integration runtime
User properties
Data flow1
Succeeded

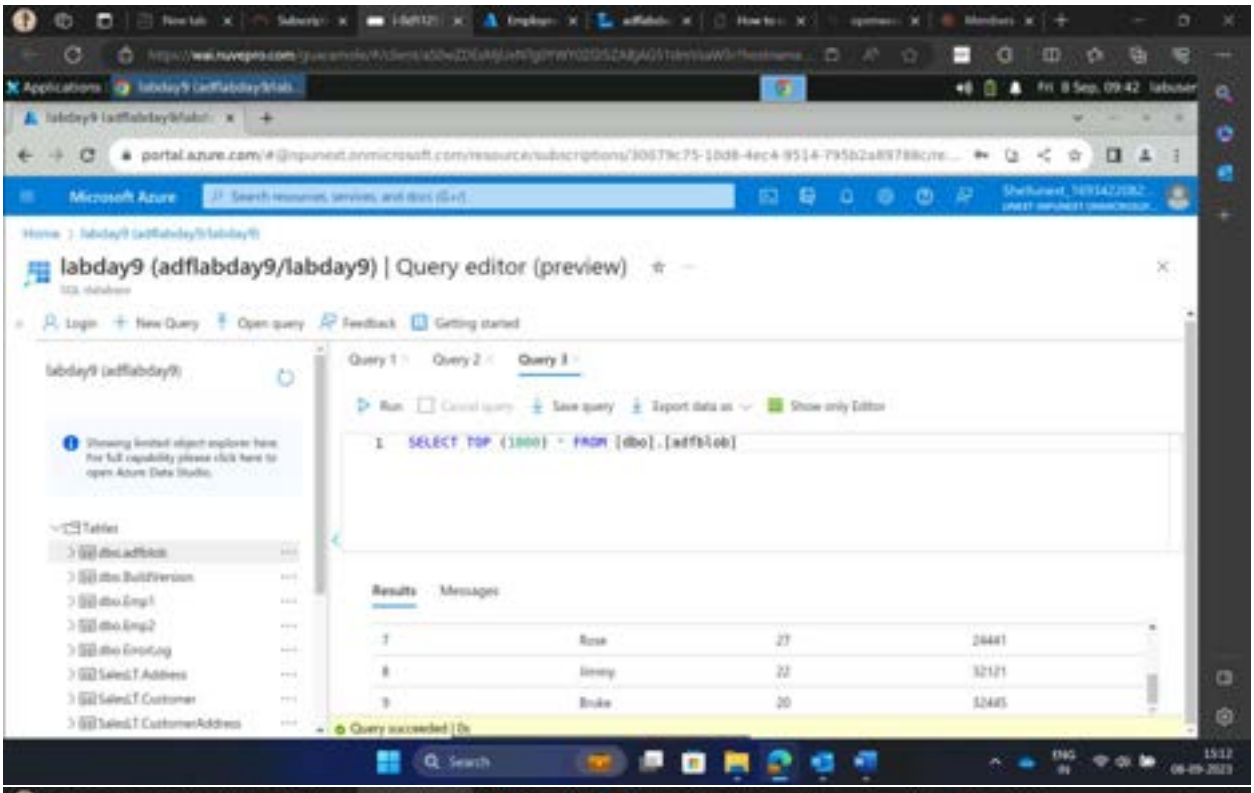

Microsoft Azure
Data Factory
Activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Debug
Add trigger
Copy data
Copy data1
Parameters
Variables
Settings
Output
Pipeline run ID:
Pipeline status
Succeeded
All status
Activity name
Activity status
Run start
Duration
Integration runtime
User properties
Copy data1
Succeeded

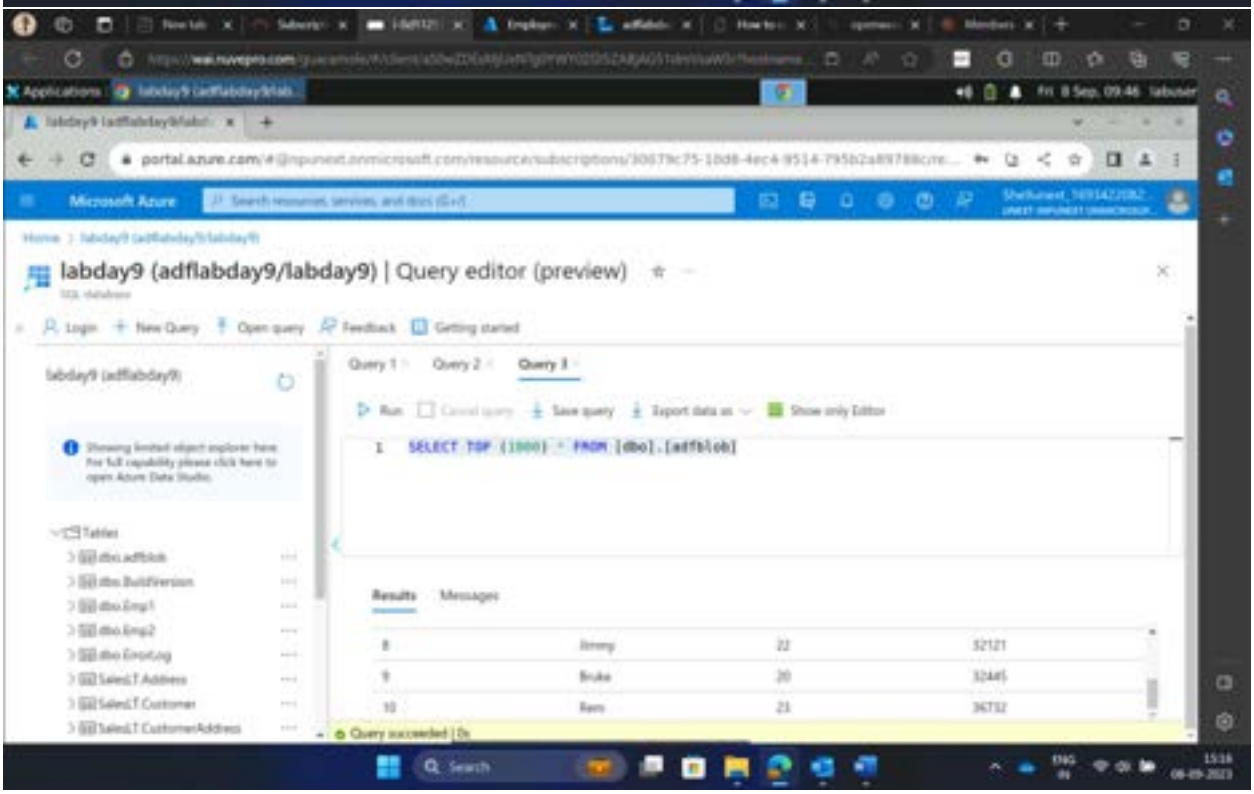

Microsoft Azure
data_f35c7587-ebca-47de-98e4-5956caf8e20b_94f9c6f7-dc20-4af6-8de0-965eee3f4778.txt
Blob
Save
Discard
Download
Refresh
Delete
Overview
Versions
Snapshots
Edit
Generate SAS
base,visibility,dt,timezone,id,name,cod
"stations",6000,1694156140,3600,2643743,"London",200
Text
Preview

CDC

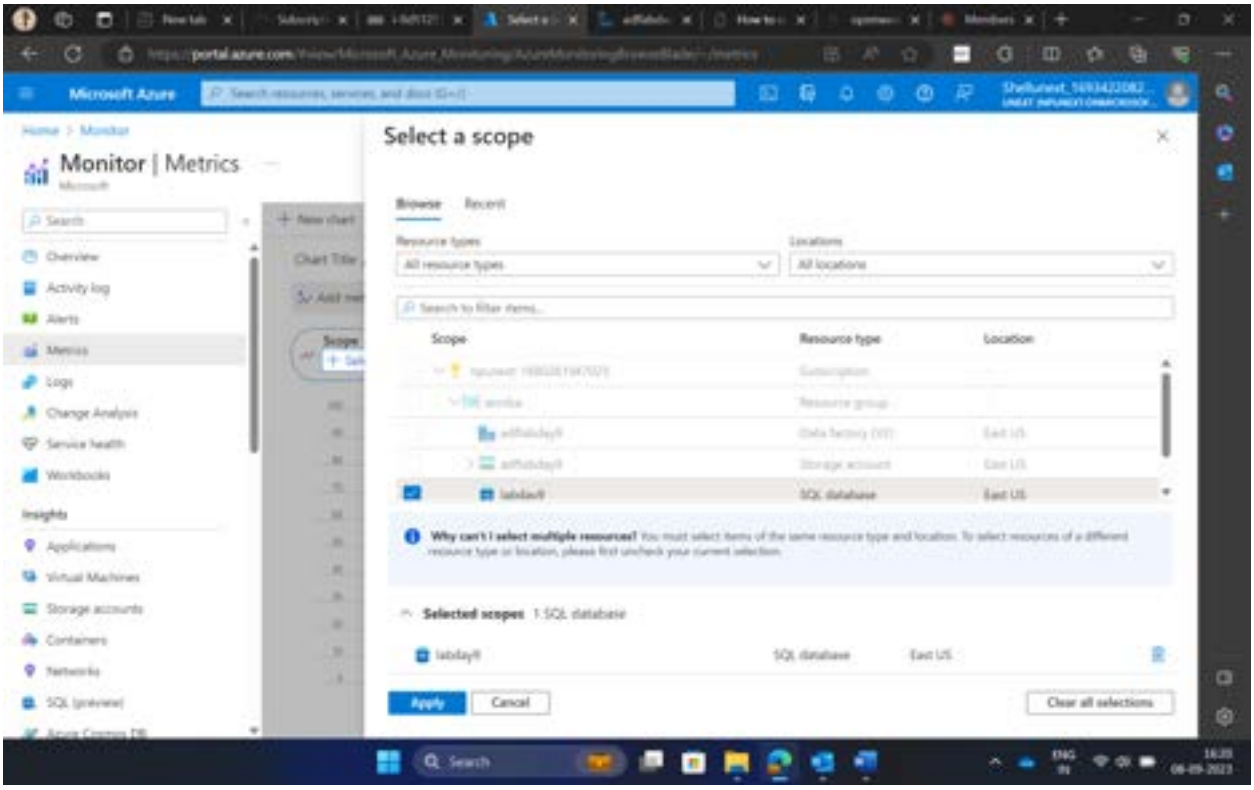

Microsoft Azure
Data Factory
Factory Resources
Pipelines
pipeline1
pipeline2
Change Data Capture (preview)
Datasets
Data flows
Power Query
Activities
Choose Your Sources
Change Data Capture (CDC) automatically detects data changes at the source and then sends the updated data to the destination.
CDC name *
adfcdc1
Source type *
DelimitedText
Source linked service *
AzureBlobStorage1
Source settings
Folder paths
adfblob
Continue
Cancel

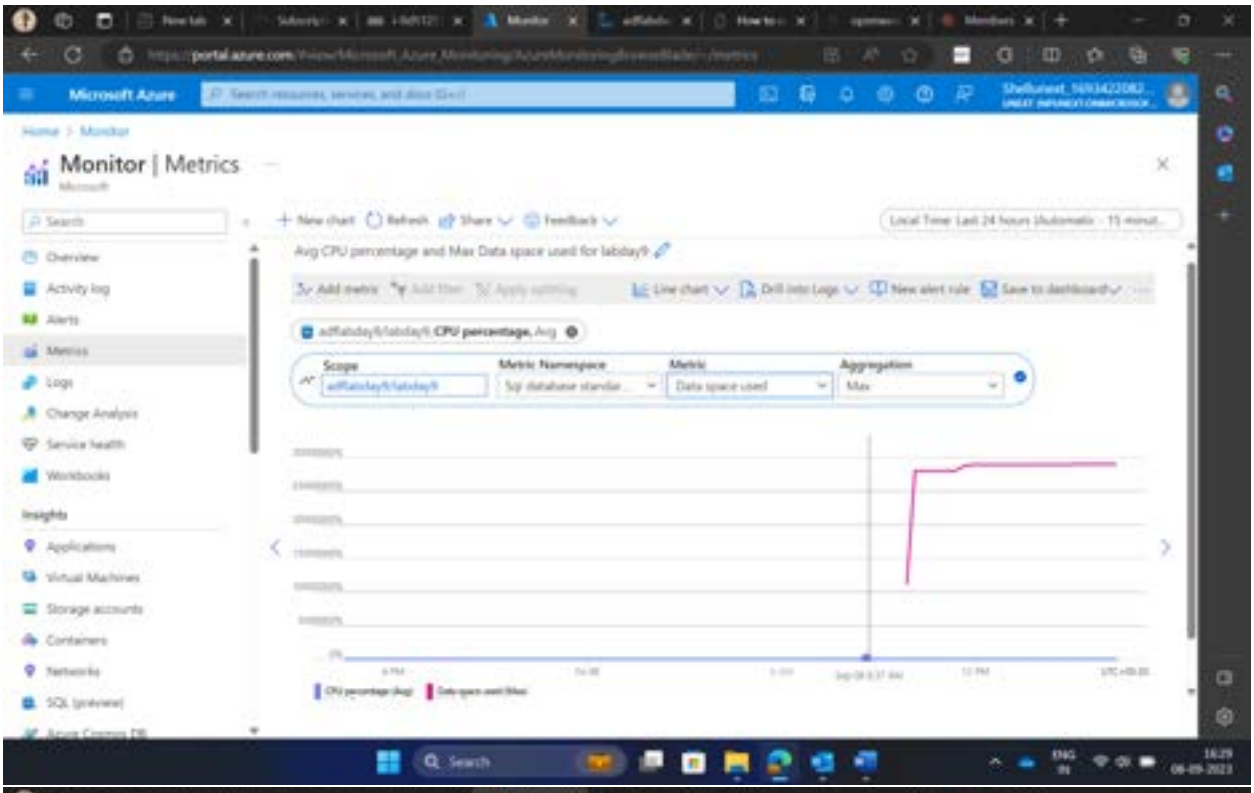

Microsoft Azure
Data Factory
Preview experience
Off
Data flow debug
New mapping
Delete
Source Table
Target Table
Columns mapped
Auto map
Source
Target
Source linked service
Type
Source tables

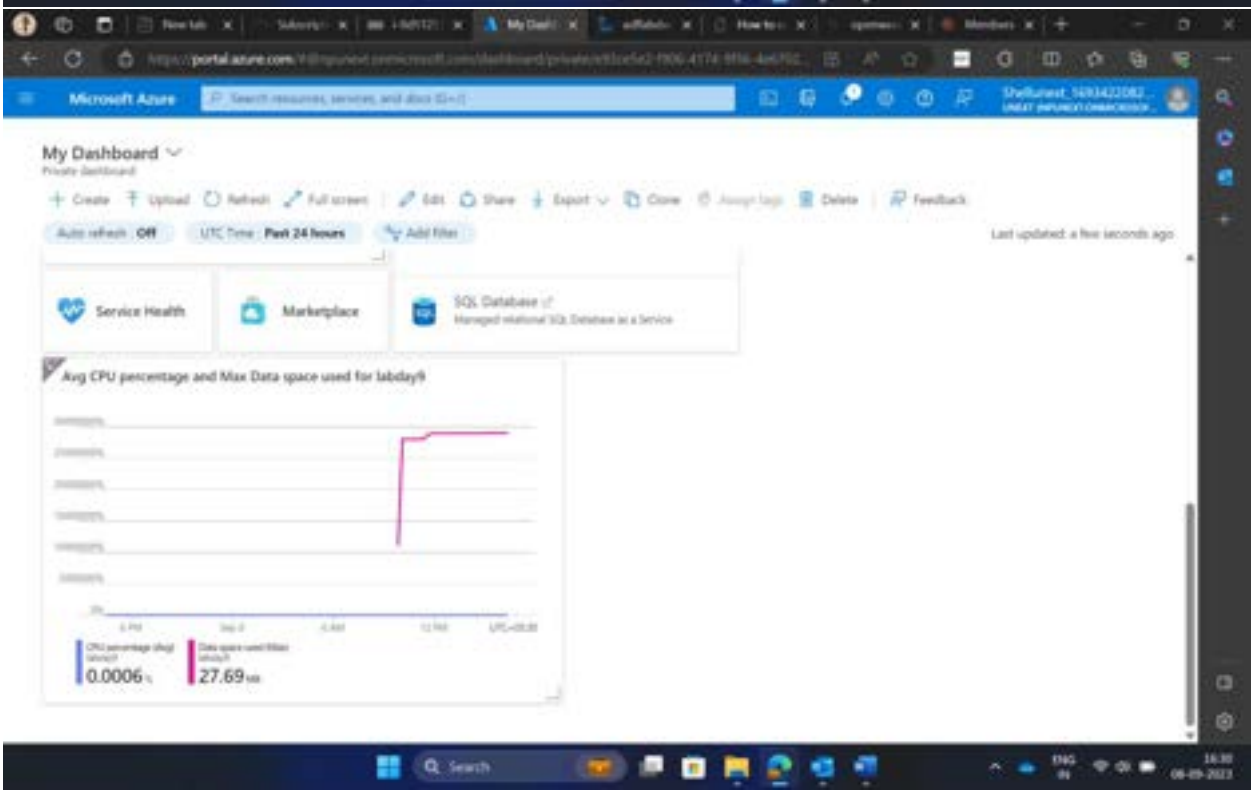

labday9 (adflabday9/labday9) | Query editor (preview)
SELECT TOP (1000) * FROM [dbo].[adfblob]
Results
Messages
7 Rose 27 26441
8 Jimmy 22 32121
9 Bruke 20 32445

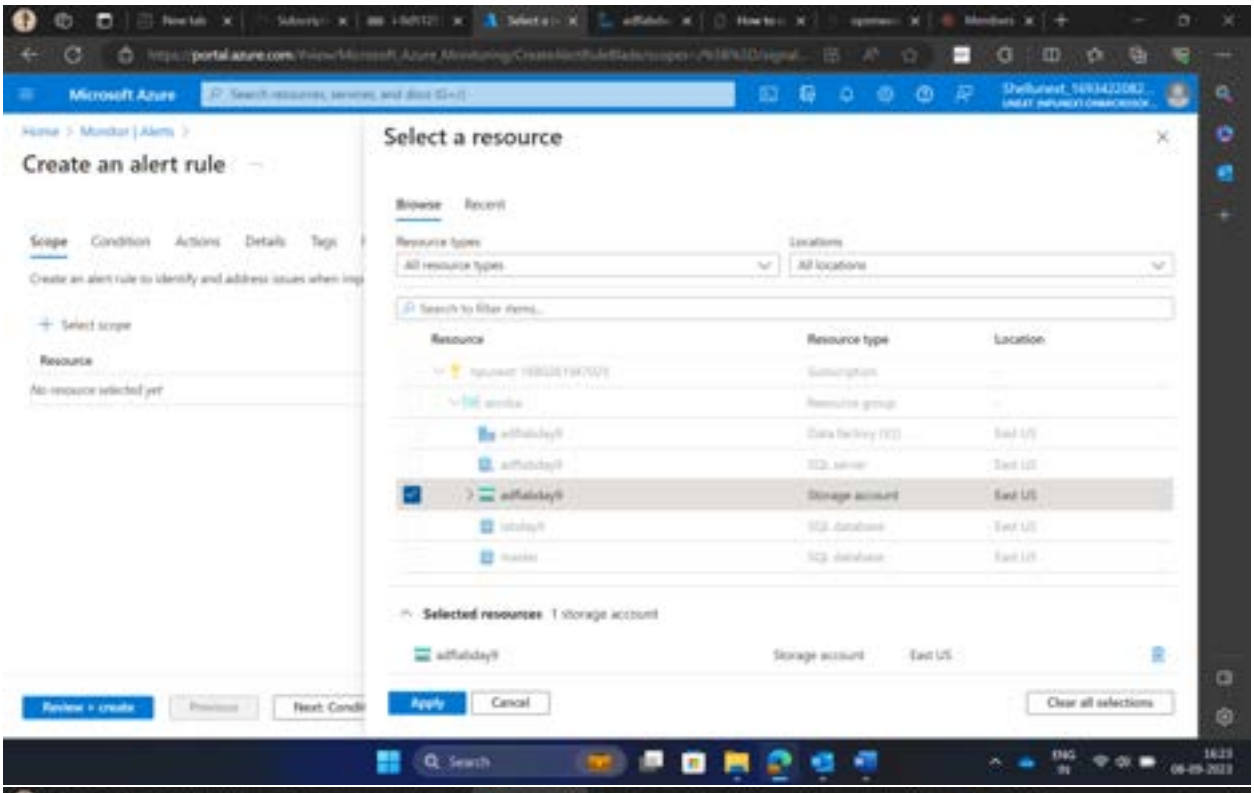

labday9 (adflabday9/labday9) | Query editor (preview)
SELECT TOP (1000) * FROM [dbo].[adfblob]
Results
Messages
8 Jimmy 22 32121
9 Bruke 20 32445
10 Rem 23 36732

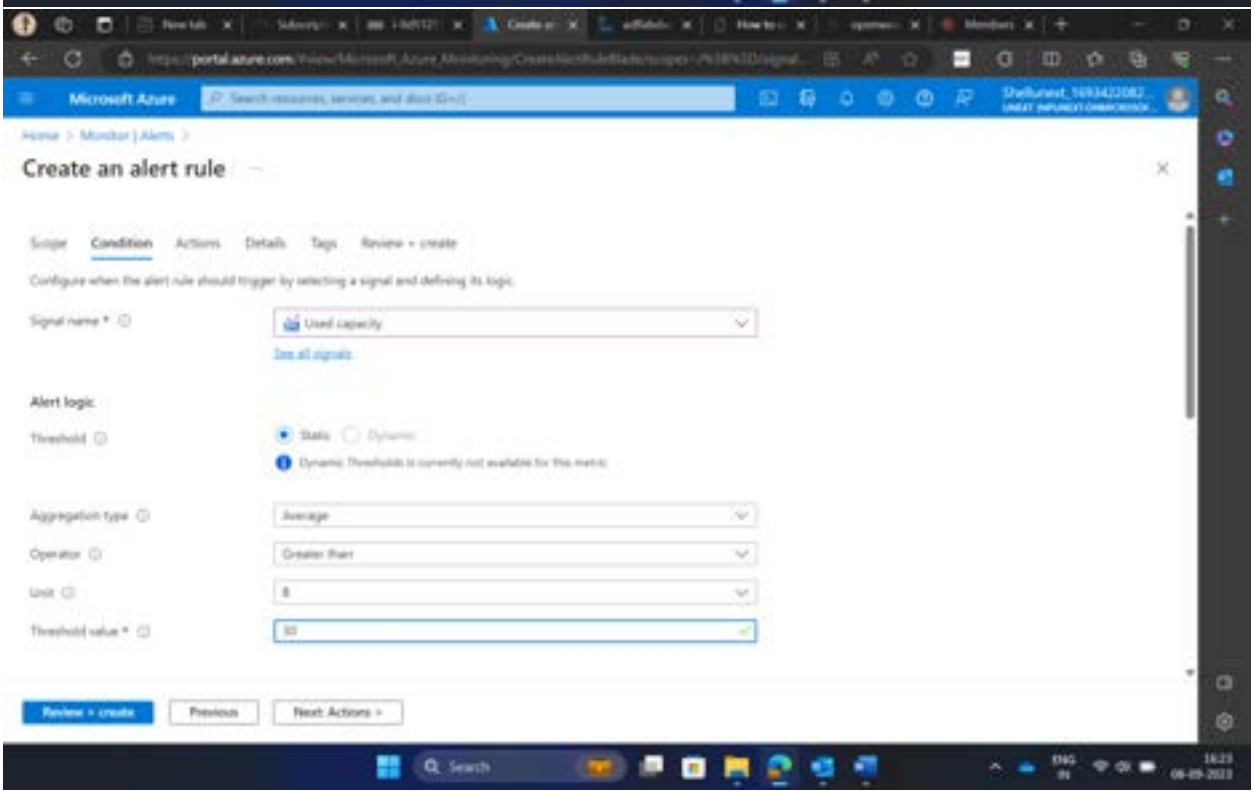
Monitoring and troubleshooting

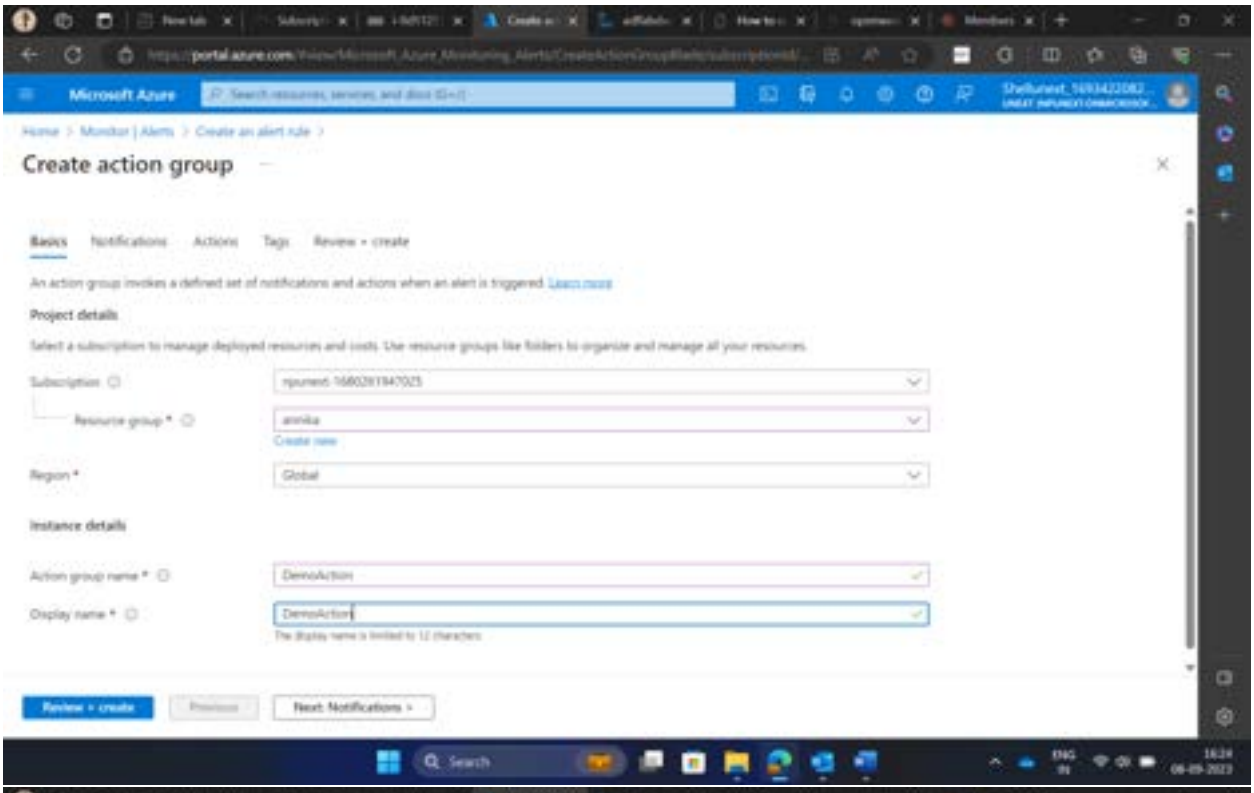


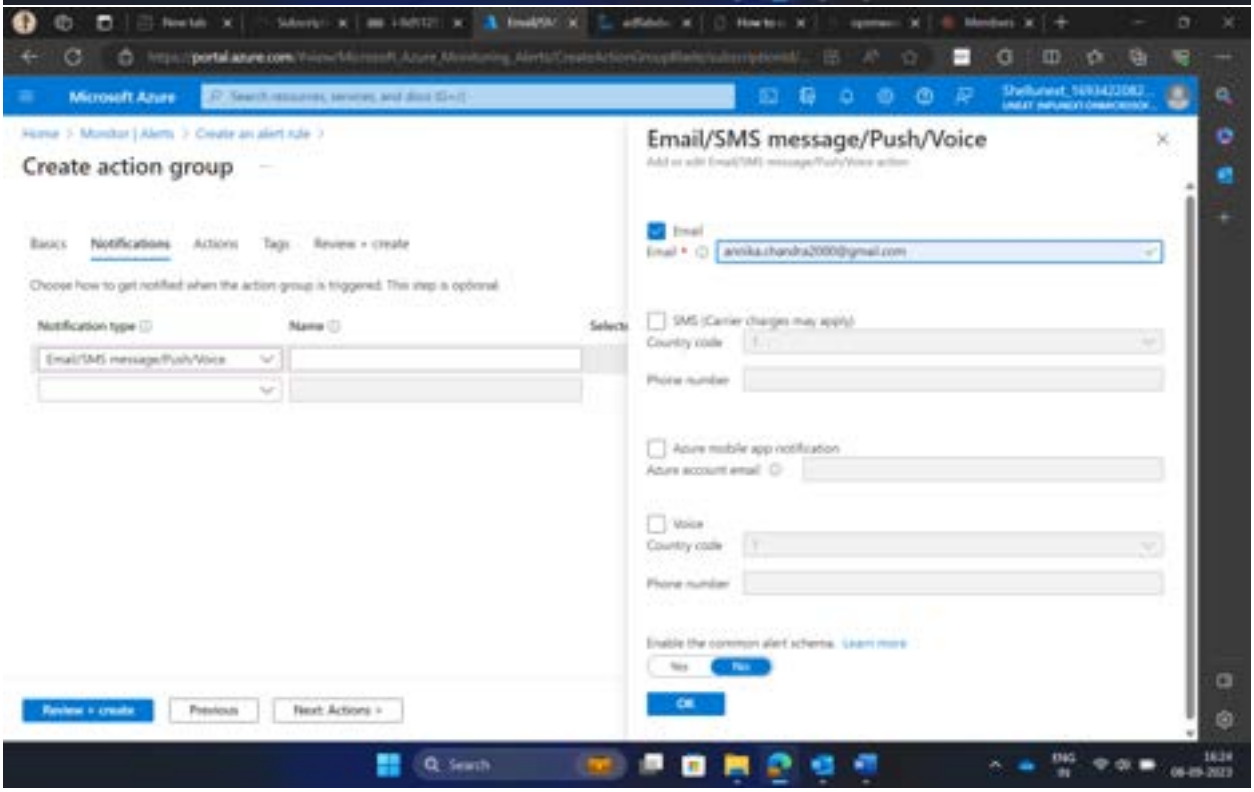

Microsoft Azure
Monitor | Metrics
Avg CPU percentage and Max Data space used for labday9
Scope
Metric Namespace
Metric
Aggregation
Data space used
Max


Microsoft Azure
My Dashboard
Service Health
Marketplace
SQL Database
Avg CPU percentage and Max Data space used for labday9
0.0006
27.69

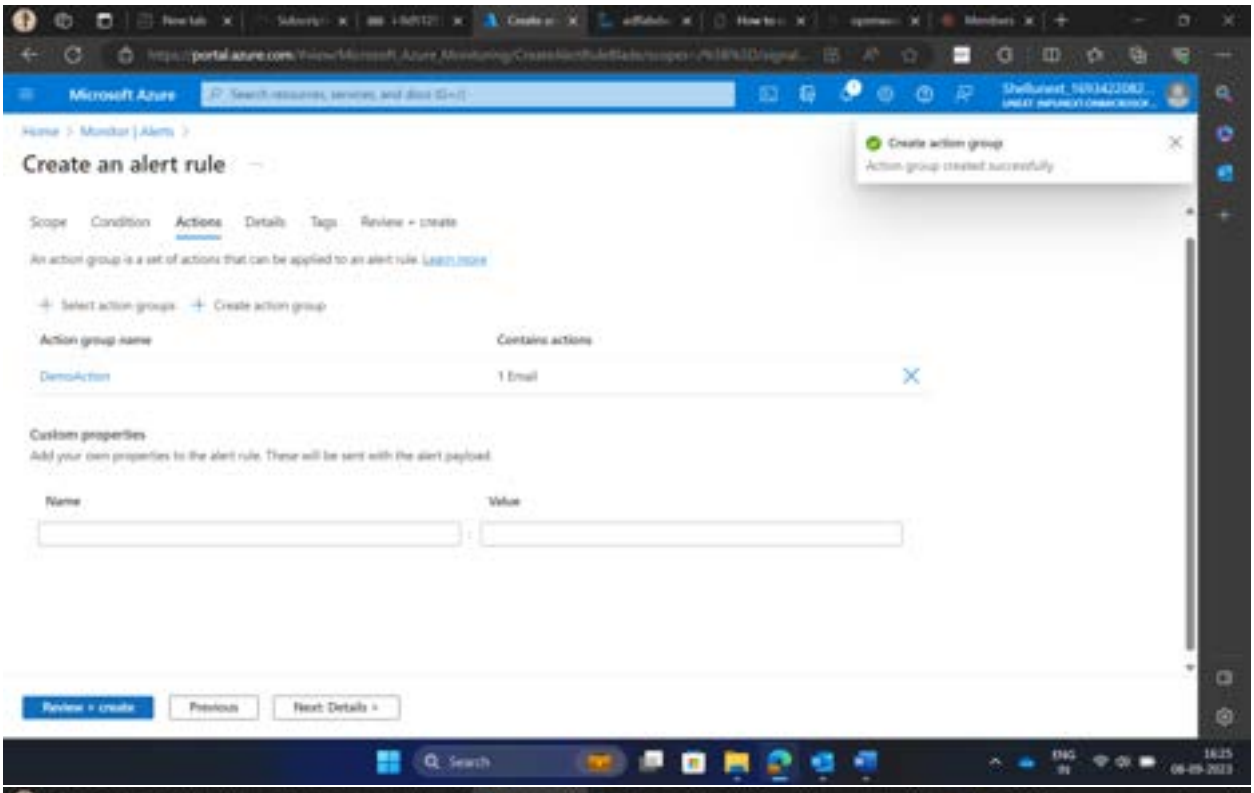

Microsoft Azure
Create an alert rule
Scope
Condition
Actions
Details
Tags
Select scope
Resource
No resource selected yet
Select a resource
Browse
Recent
Resource types
All resource types
Locations
All locations
Resource
Resource type
Location
adfdatalday9
Storage account
East US
Selected resources 1 storage account
adfdatalday9
Storage account
East US
Apply
Cancel
Clear all selections
Review + create
Previous
Next: Condit

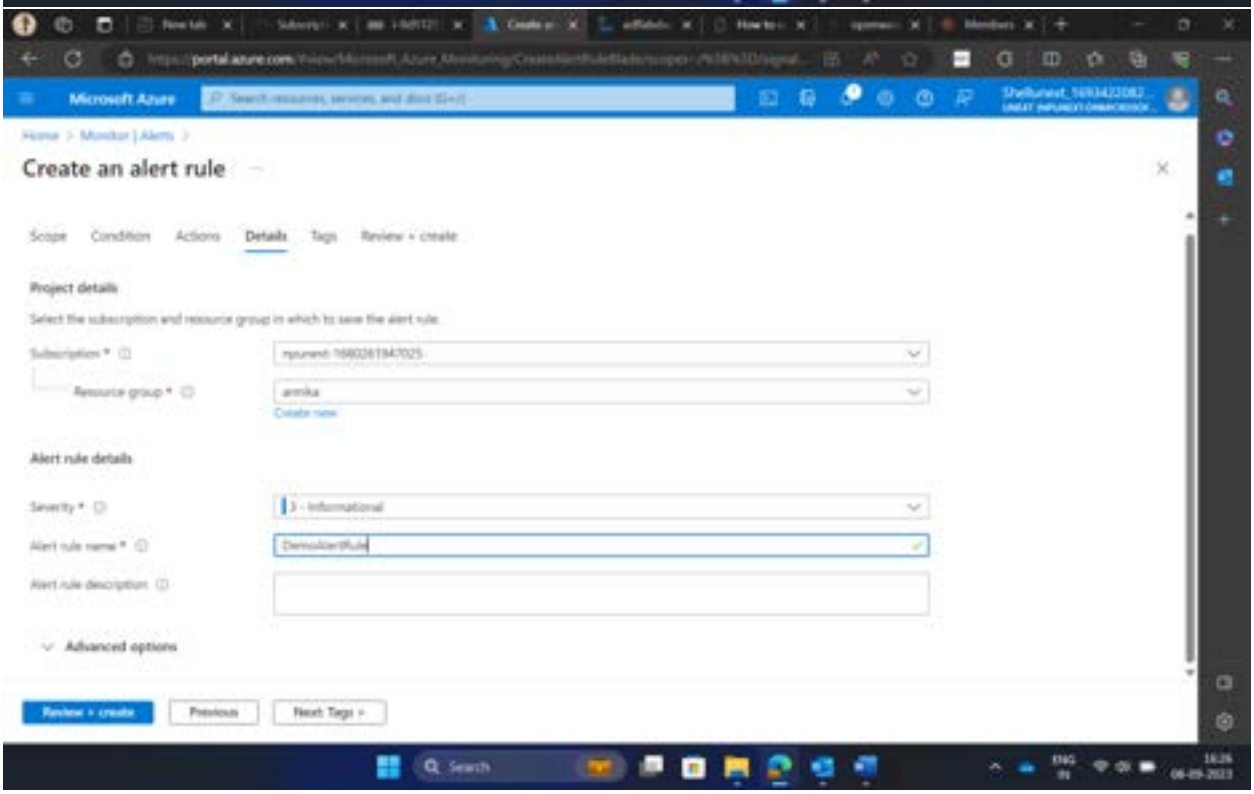

Microsoft Azure
Create an alert rule
Scope
Condition
Actions
Details
Tags
Review + create
Configure when the alert rule should trigger by selecting a signal and defining its logic.
Signal name *
Used capacity
See all signals
Alert logic
Threshold
Static
Dynamic
Dynamic Thresholds is currently not available for this metric.
Aggregation type
Average
Operator
Greater than
Unit
B
Threshold value *
50
Review + create
Previous
Next: Actions >

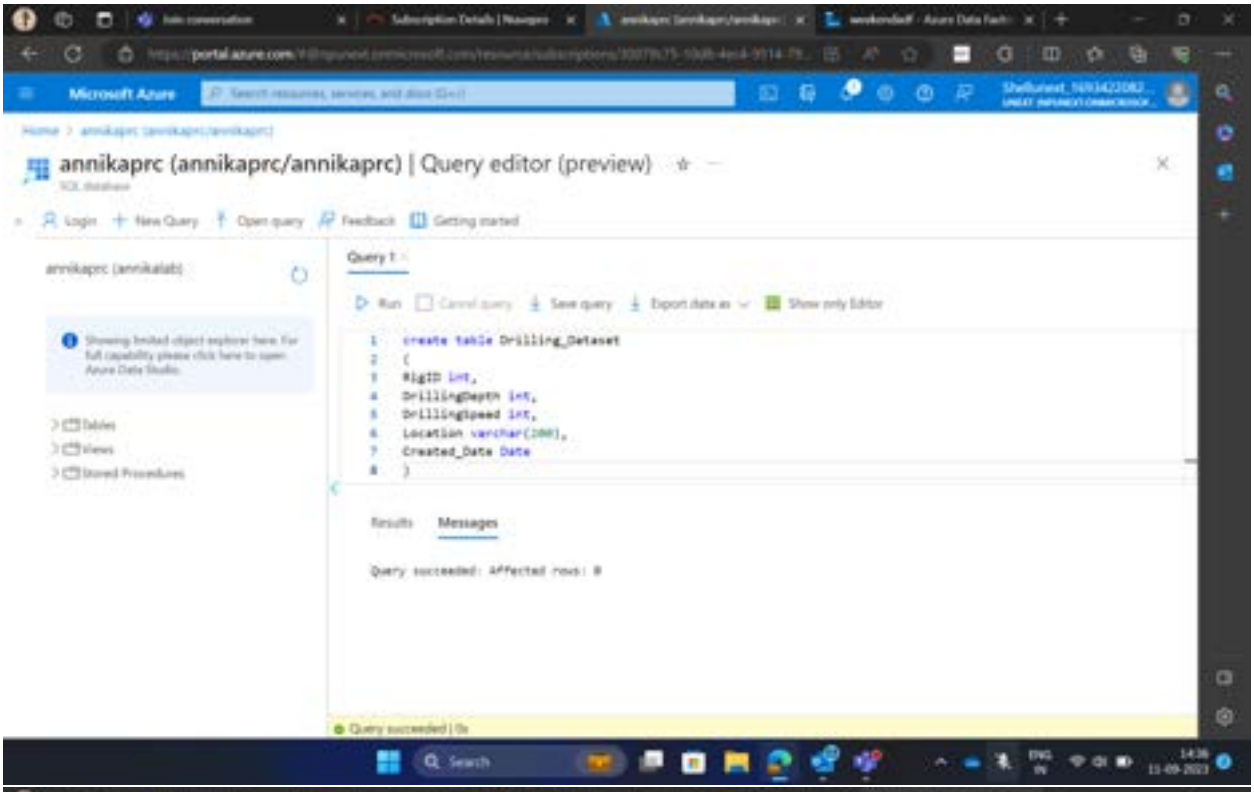

Microsoft Azure
Home > Monitor | Alerts > Create an alert rule >
Create action group
Basics
Notifications
Actions
Tags
Review + create
An action group invokes a defined set of notifications and actions when an alert is triggered. Learn more
Project details
Select a subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.
Subscription
Resource group *
annika
Create new
Region *
Global
Instance details
Action group name *
DemoAction
Display name *
DemoAction
The display name is limited to 12 characters
Review + create
Previous
Next: Notifications >

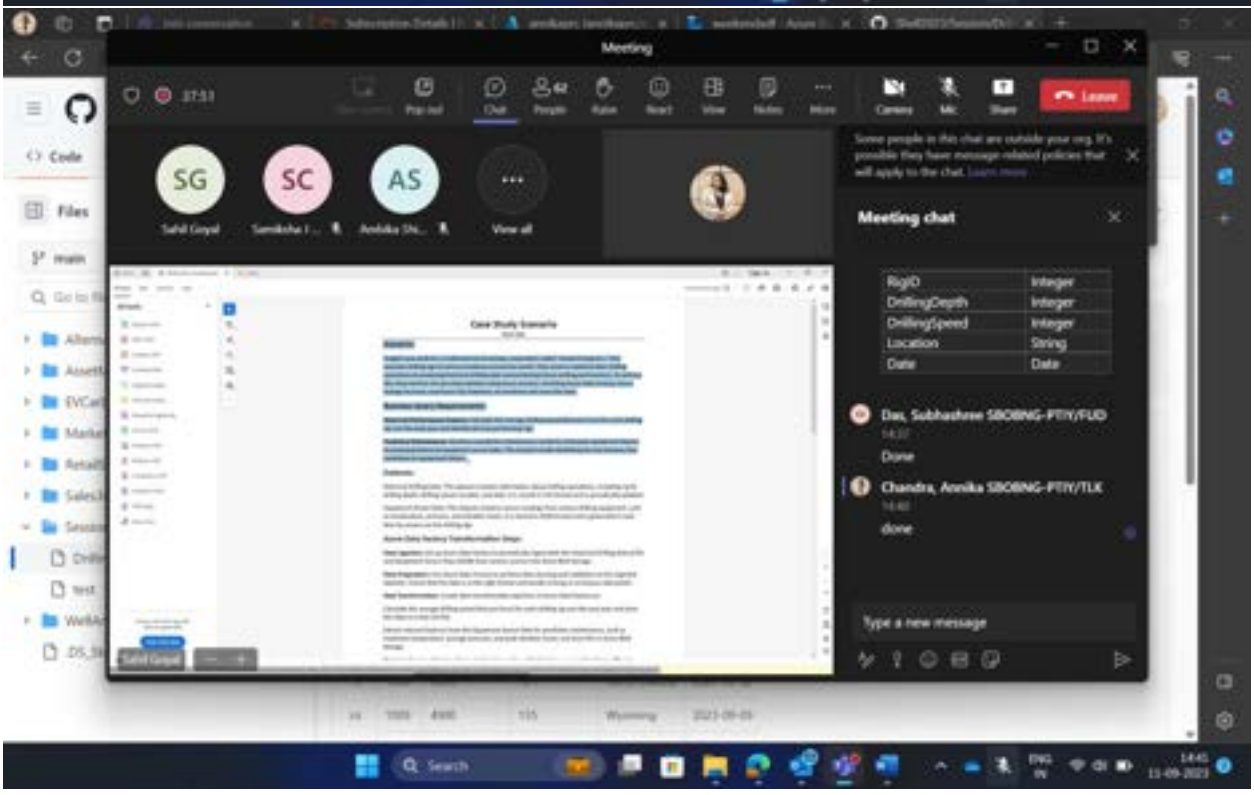

Microsoft Azure
Home > Monitor | Alerts > Create an alert rule >
Create action group
Basics
Notifications
Actions
Tags
Review + create
Choose how to get notified when the action group is triggered. This step is optional.
Notification type
Name
Email/SMS message/Push/Voice
Review + create
Previous
Next: Actions >

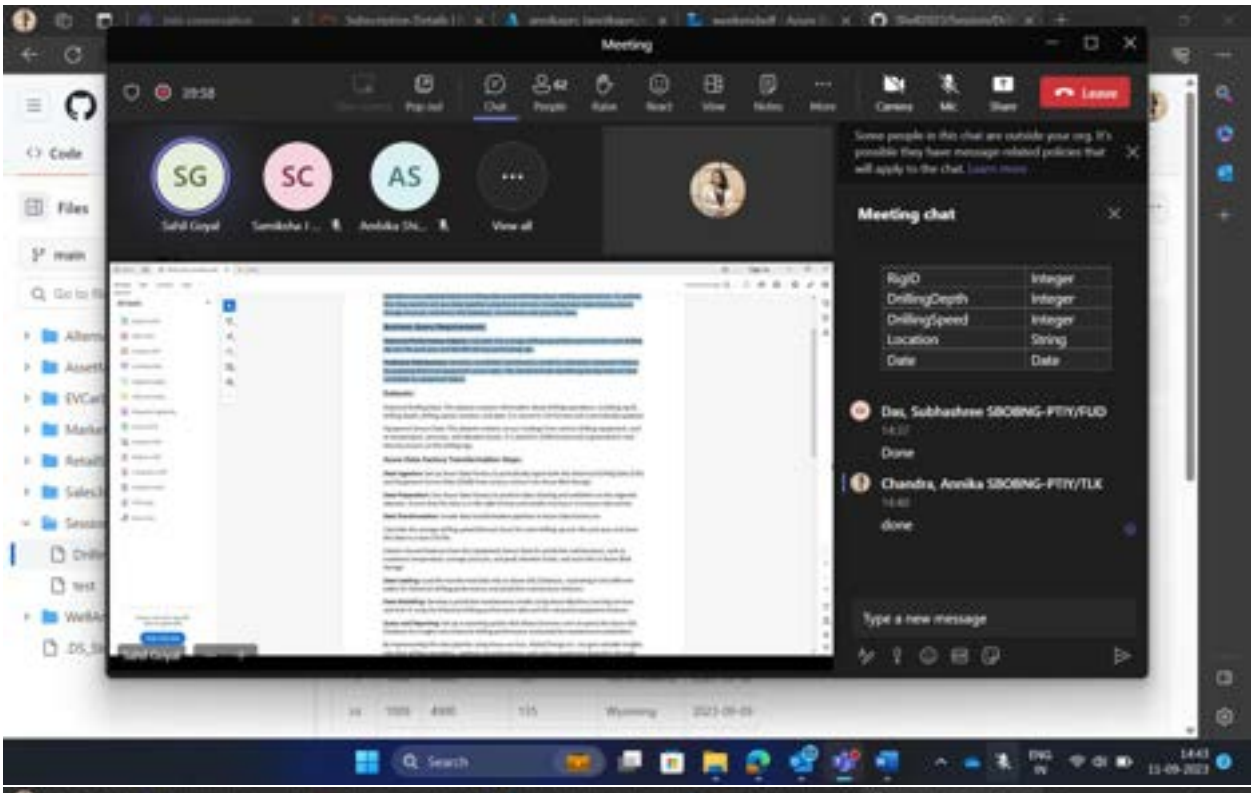
Email/SMS message/Push/Voice
Email
Email *
annika.chandra2000@gmail.com
SMS (Carrier charges may apply)
Country code
Phone number
Azure mobile app notification
Azure account email
Voice
Country code
Phone number
Enable the common alert schema. Learn more
Yes
No
OK

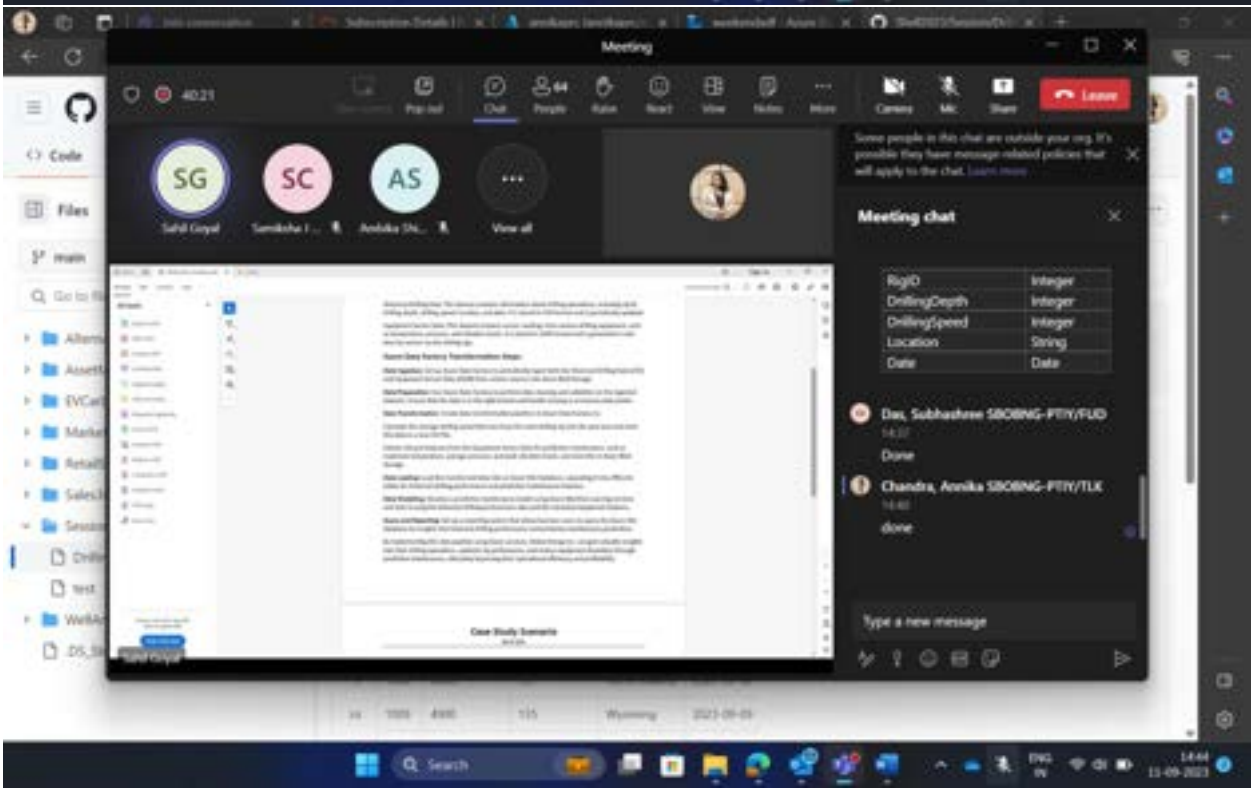

Microsoft Azure
Create action group
Action group created successfully
Create an alert rule
Scope
Condition
Actions
Details
Tags
Review + create
An action group is a set of actions that can be applied to an alert rule. Learn more
Select action groups
Create action group
Action group name
Contains actions
DemoAction
1 Email
Custom properties
Add your own properties to the alert rule. These will be sent with the alert payload.
Name
Value
Review + create
Previous
Next: Details >

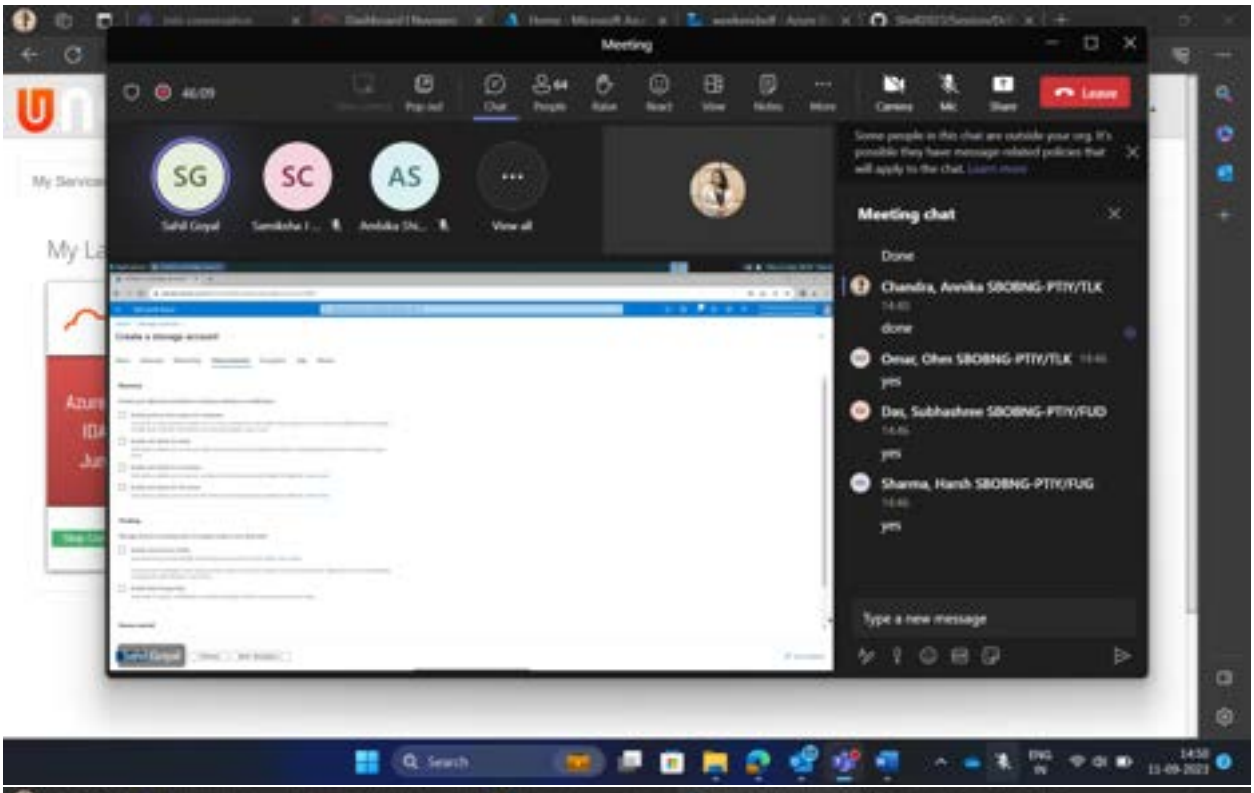

Microsoft Azure
Create an alert rule
Scope
Condition
Actions
Details
Tags
Review + create
Project details
Select the subscription and resource group in which to save the alert rule.
Subscription *
Resource group *
anmka
Create new
Alert rule details
Severity *
3 - Informational
Alert rule name *
Alert rule description
Advanced options
Review + create
Previous
Next: Tags >

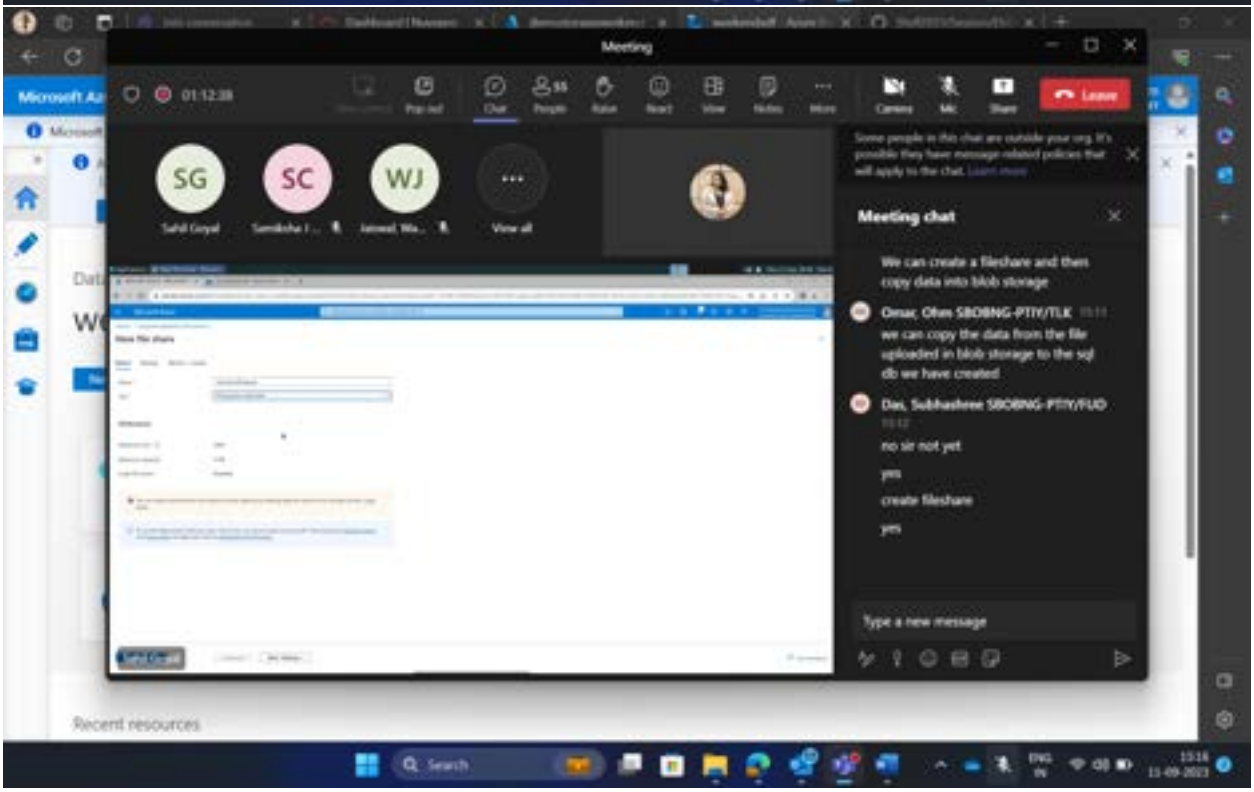

Microsoft Azure
annikaprc (annikaprc/annikaprc) | Query editor (preview)
Login
New Query
Open query
Feedback
Getting started
annikaprc (annikalab)
Query 1
Run
Cancel query
Save query
Export data as
Show only Editor
create table Drilling_Dataset
(
RigID int,
DrillingDepth int,
DrillingSpeed int,
Location varchar(200),
Created_Date Date
)
Results
Messages
Query succeeded: Affected rows: 0
Query succeeded | 0s


Meeting
Leave
SG
SC
AS
View all
Meeting chat
RigID
Integer
DrillingDepth
Integer
DrillingSpeed
Integer
Location
String
Date
Date
Das, Subhashree SBOBNG-PTIY/FUD
Done
Chandra, Annika SBOBNG-PTIY/TLK
done
Type a new message

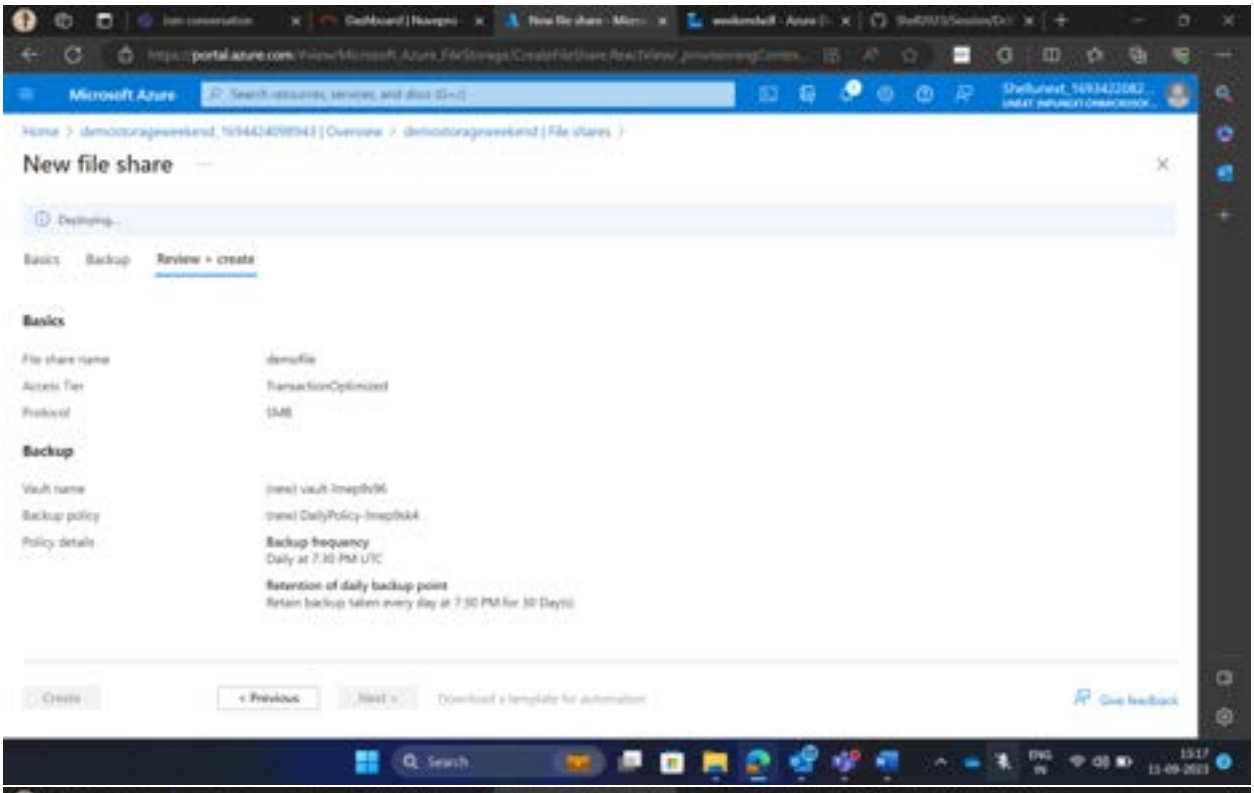

Meeting
Meeting chat
RigID Integer
DrillingDepth Integer
DrillingSpeed Integer
Location String
Date Date
Das, Subhashree SBOBNG-FTIY/FUD
Done
Chandra, Annika SBOBNG-FTIY/TLK
done
Type a new message

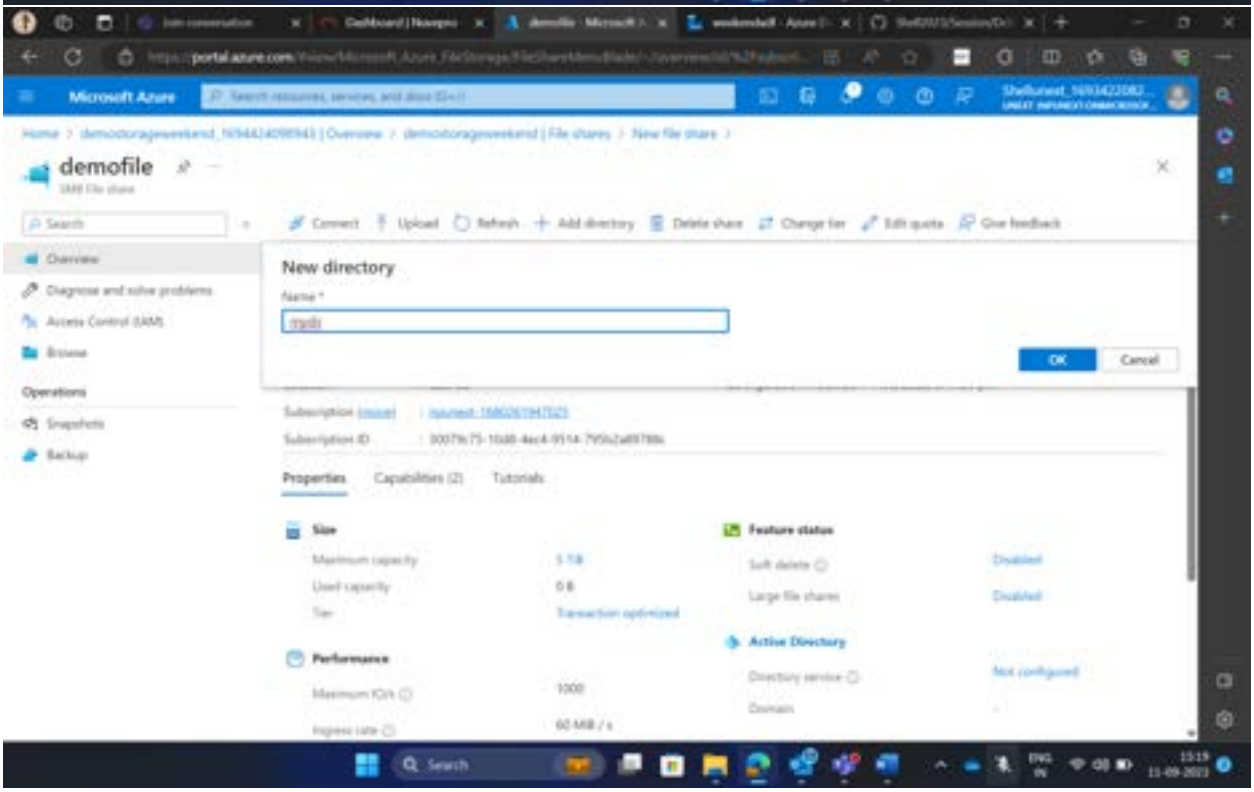

Meeting
Meeting chat
RigID Integer
DrillingDepth Integer
DrillingSpeed Integer
Location String
Date Date
Das, Subhashree SBOBNG-FTIY/FUD
Done
Chandra, Annika SBOBNG-FTIY/TLK
done
Type a new message

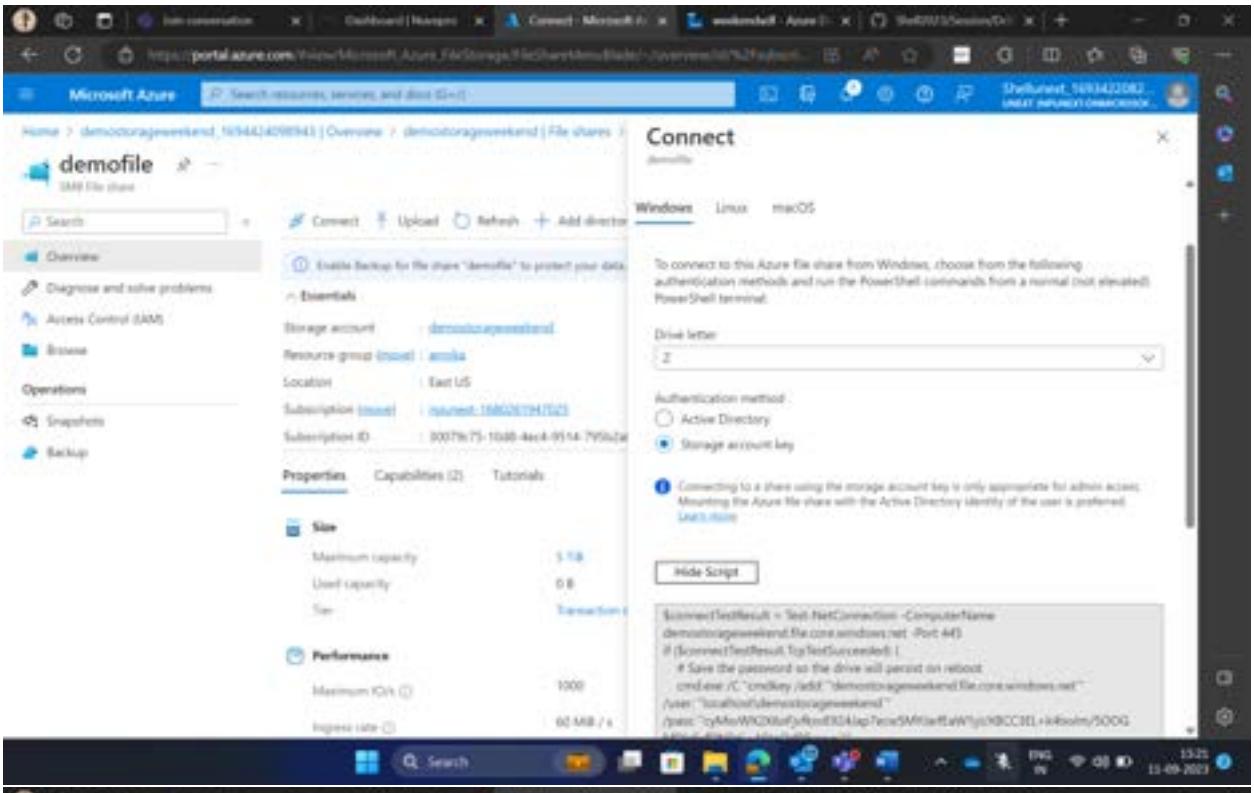


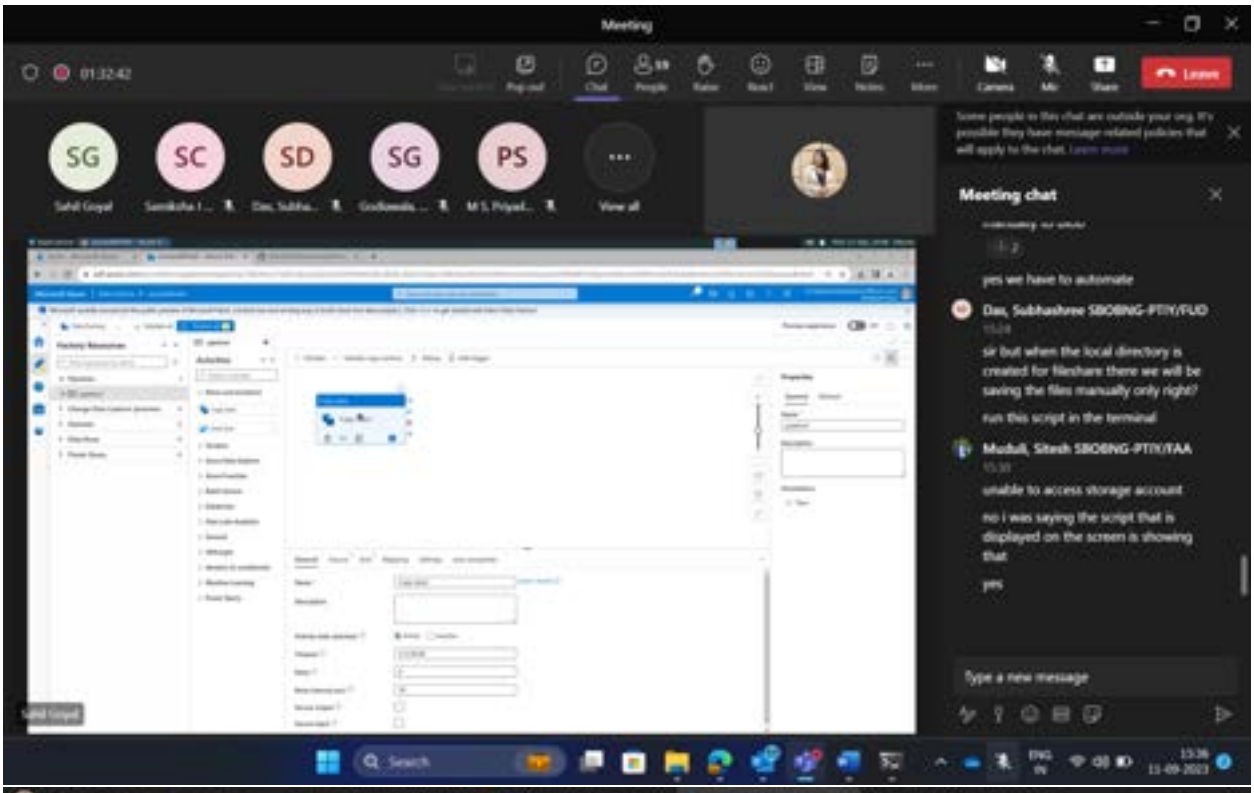


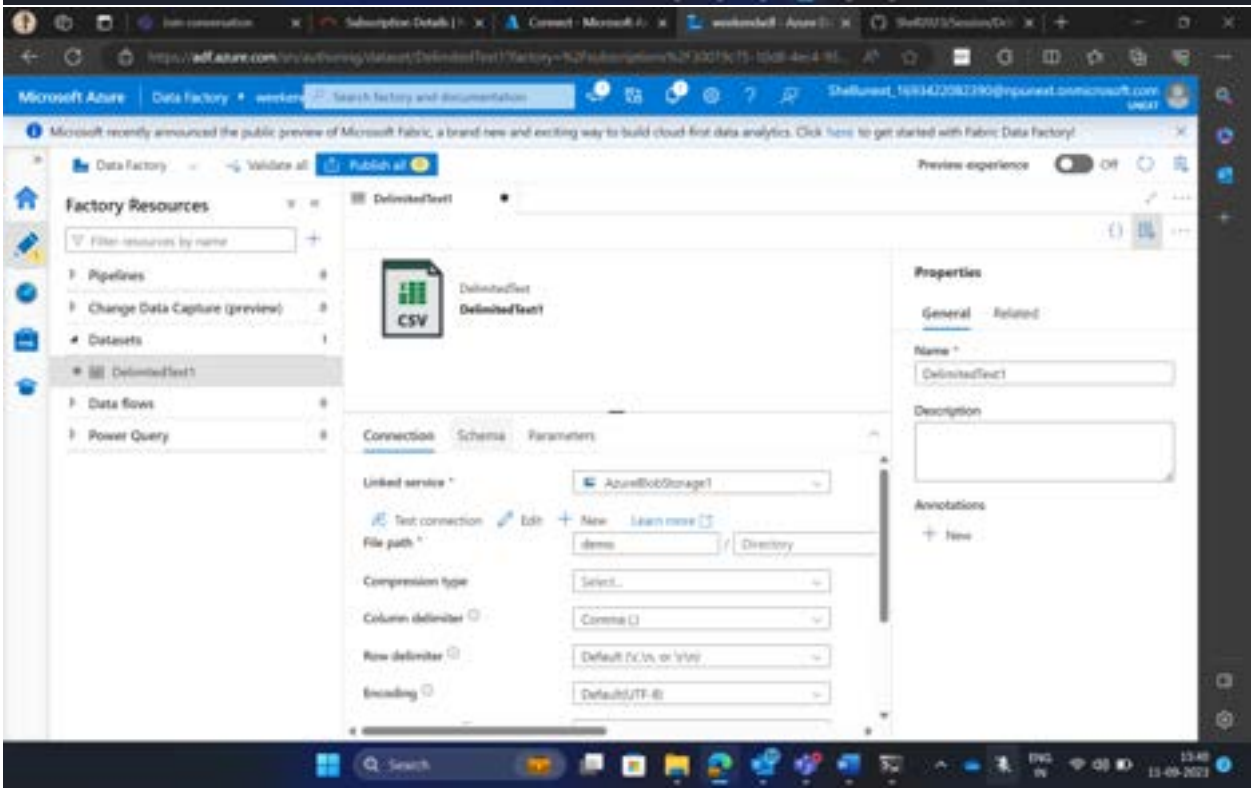
For uploading file automatically for tht we create file share in storage acc

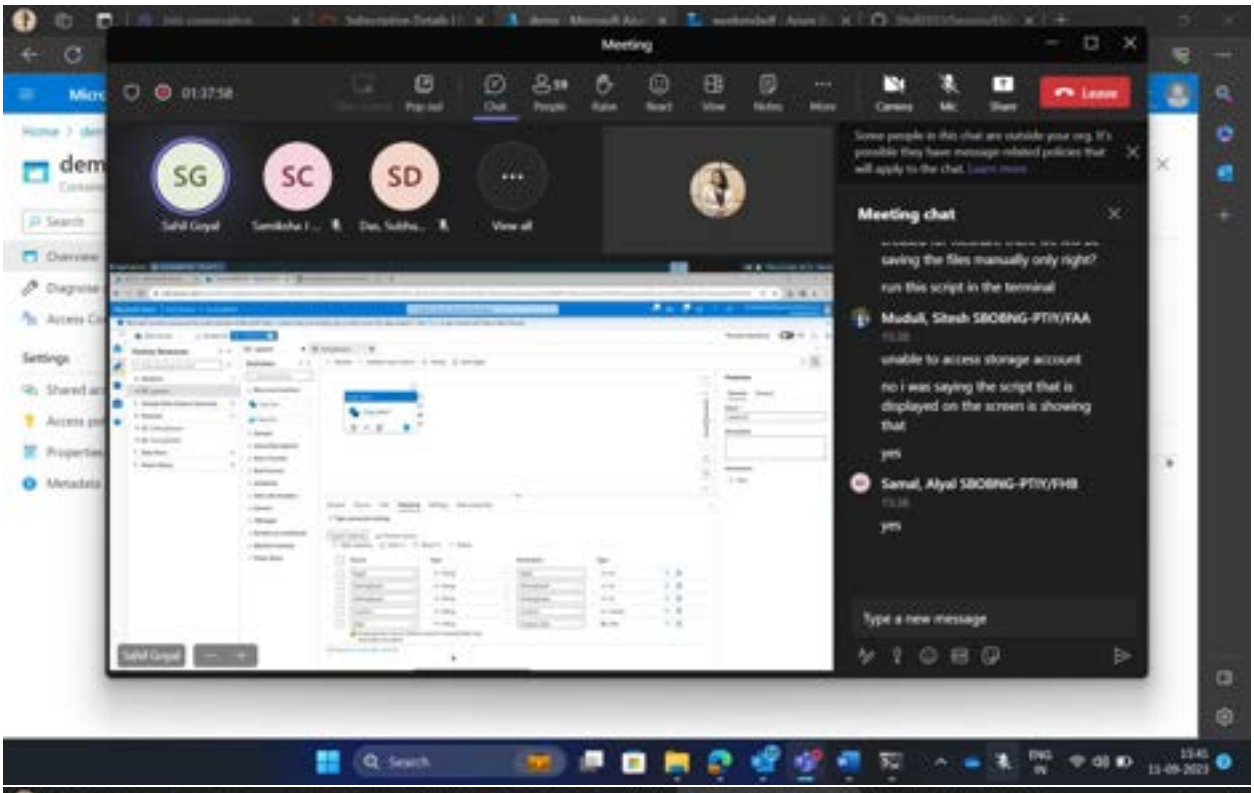

Microsoft Azure
New file share
Basics
Backup
Review + create
Basics
File share name
demofile
Access Tier
TransactionOptimized
Protocol
SMB
Backup
Vault name
Backup policy
Policy details
Backup frequency
Daily at 7:30 PM UTC
Retention of daily backup point
Create
< Previous
Next >
Give feedback

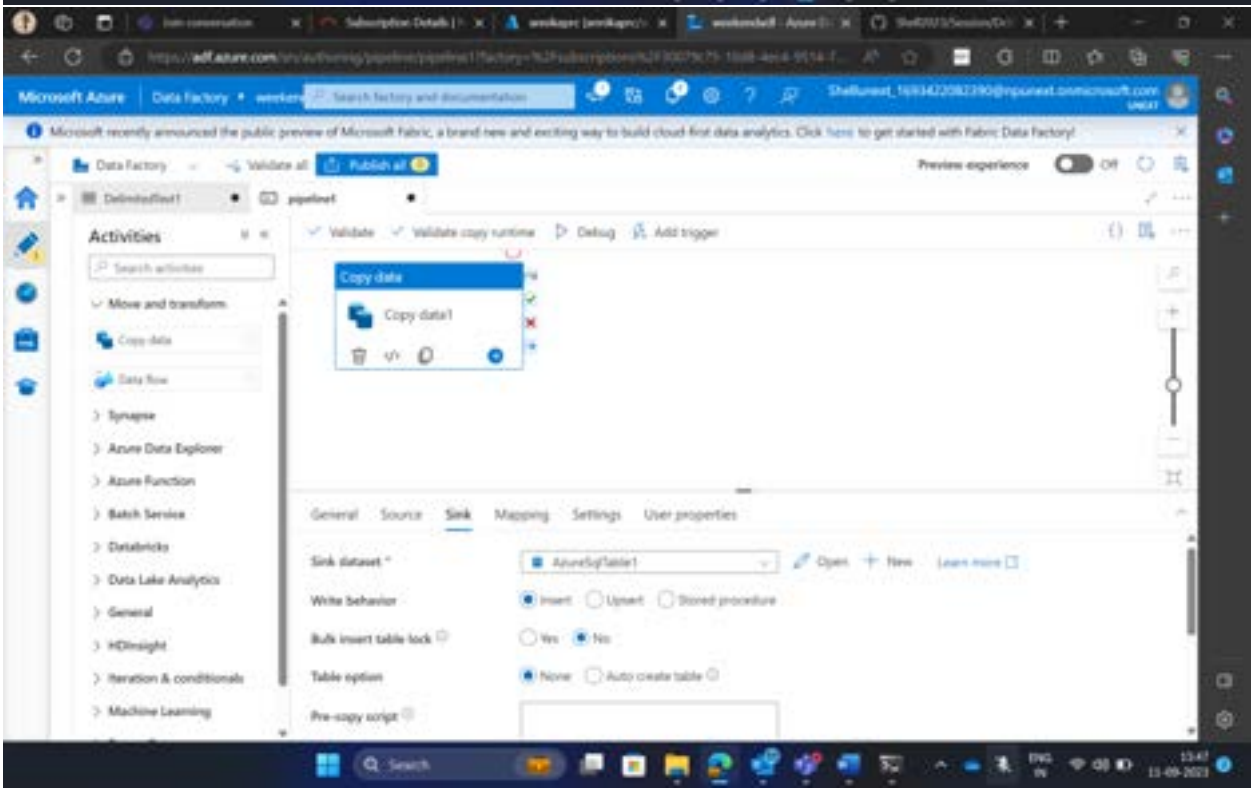

Microsoft Azure
demofile
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Browse
Operations
Snapshots
Backup
Connect
Upload
Refresh
Add directory
Delete share
Change tier
Edit quota
Give feedback
New directory
Name *
OK
Cancel
Properties
Capabilities (2)
Tutorials
Size
Maximum capacity
Used capacity
Tier
Performance
Feature status
Soft delete
Large file shares
Disabled
Active Directory
Directory service
Not configured
Domain

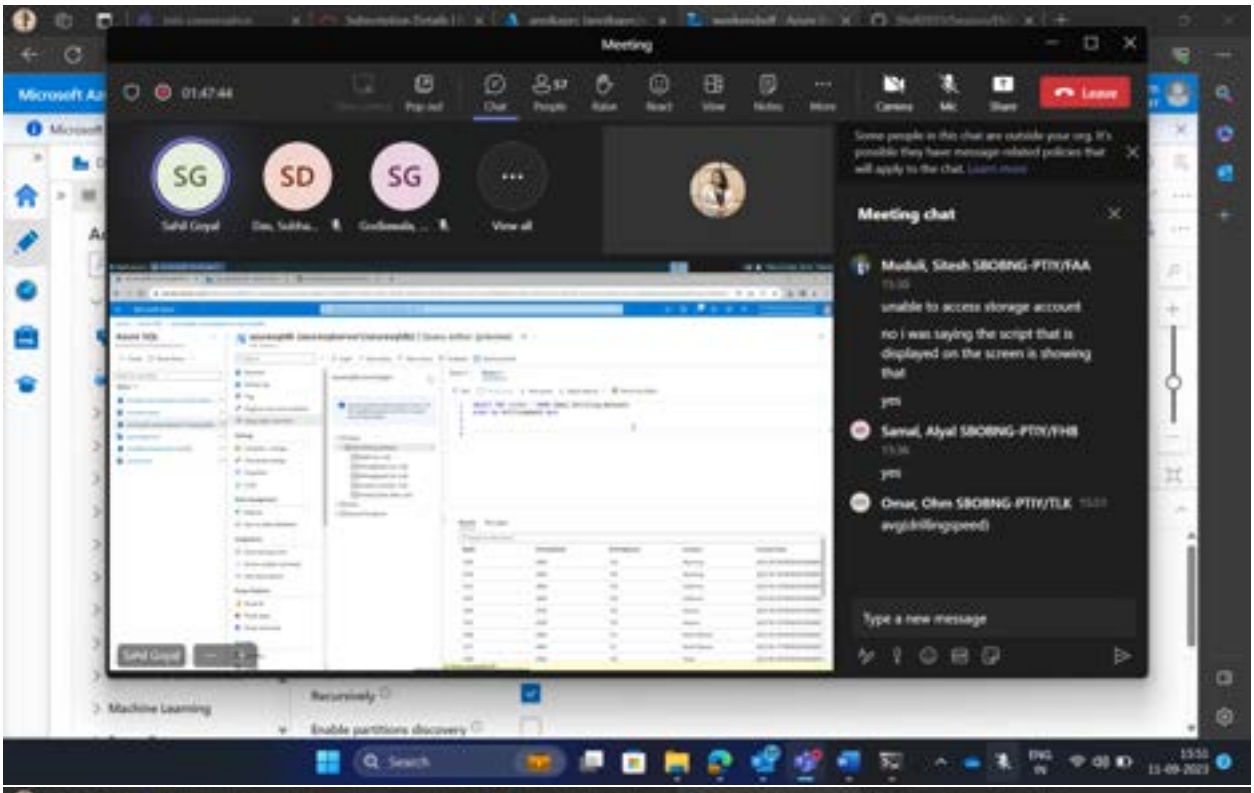

Microsoft Azure
demofile
Connect
Windows
Linux
macOS
To connect to this Azure file share from Windows, choose from the following authentication methods and run the PowerShell commands from a normal (not elevated) PowerShell terminal.
Drive letter
Z
Authentication method
Active Directory
Storage account key
Connecting to a share using the storage account key is only appropriate for admin access. Mounting the Azure file share with the Active Directory identity of the user is preferred.
Learn more
Hide Script
$connectTestResult = Test-NetConnection -ComputerName demostorageweekend.file.core.windows.net -Port 445
if ($connectTestResult.TcpTestSucceeded) {
# Save the password so the drive will persist on reboot
cmd.exe /C "cmdkey /add:`"demostorageweekend.file.core.windows.net`" /user:`"localhost\demostorageweekend`" /pass:`"cyMnvWX2KKnFjvRnnE924ap7ecxeSMYJarffLaW1yUX8CC3EL+k4tsoHs/5OOGMDKtkKfDN9H/+A0zoDdRQw==`""
Mount the drive
New-PSDrive -Name Z -PSProvider FileSystem -Root "\\demostorageweekend.file.core.windows.net\demofile" -Persist
} else {
Write-Error -Message "Unable to reach the Azure storage account via port 445. Check to make sure your organization or ISP is not blocking port 445, or use Azure P2S VPN, Azure S2S VPN, or Express Route to tunnel SMB traffic over a different port."
}
This script will check to see if this storage account is accessible via TCP port 445, which is the port SMB uses. If port 445 is available, your Azure file share will be persistently mounted. Your organization or internet service provider (ISP) may block port 445, however you may use Azure Point-to-Site (P2S) VPN, Azure Site-to-Site (S2S) VPN, or ExpressRoute to tunnel SMB traffic to your Azure file share over a different port.

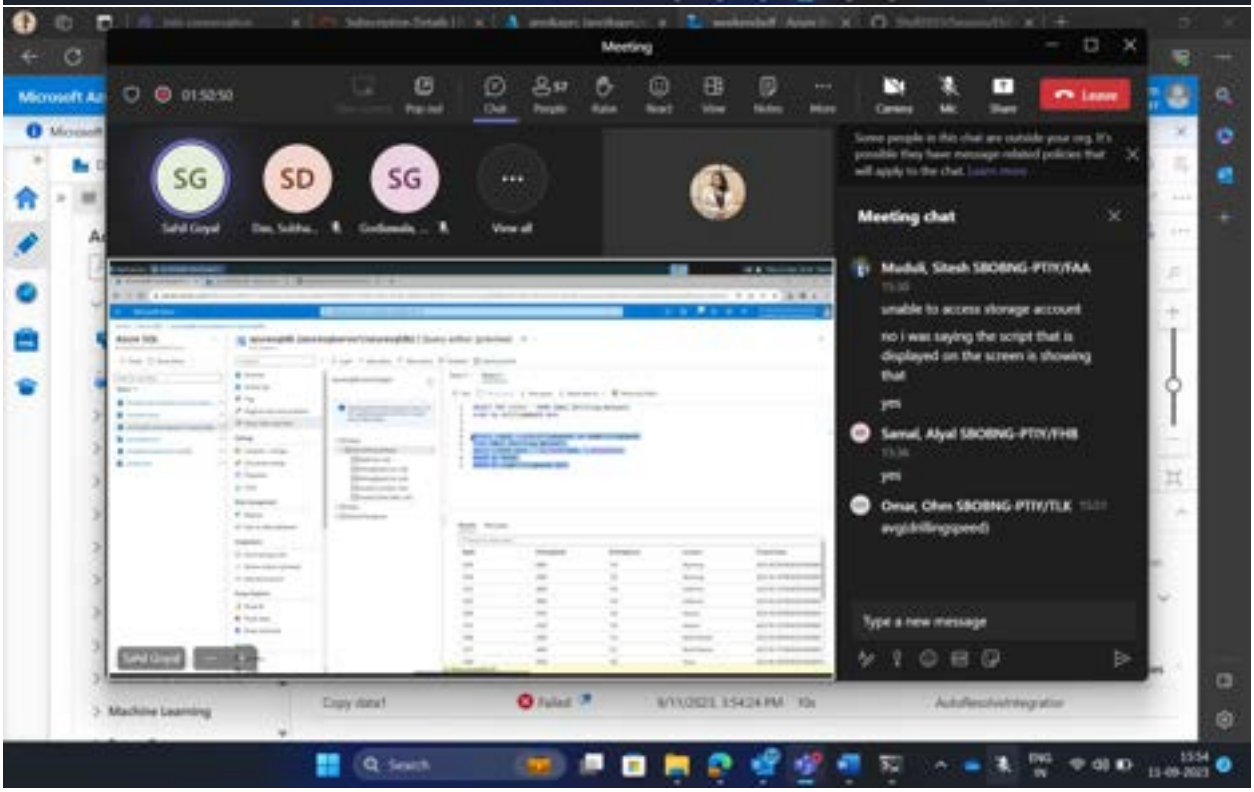

Meeting
Meeting chat
yes we have to automate
Das, Subhashree SBOBNG-PTIY/FUD
sir but when the local directory is created for fileshare there we will be saving the files manually only right?
run this script in the terminal
Mudull, Sitesh SBOBNG-PTIY/FAA
unable to access storage account
no i was saying the script that is displayed on the screen is showing that
yes
Type a new message

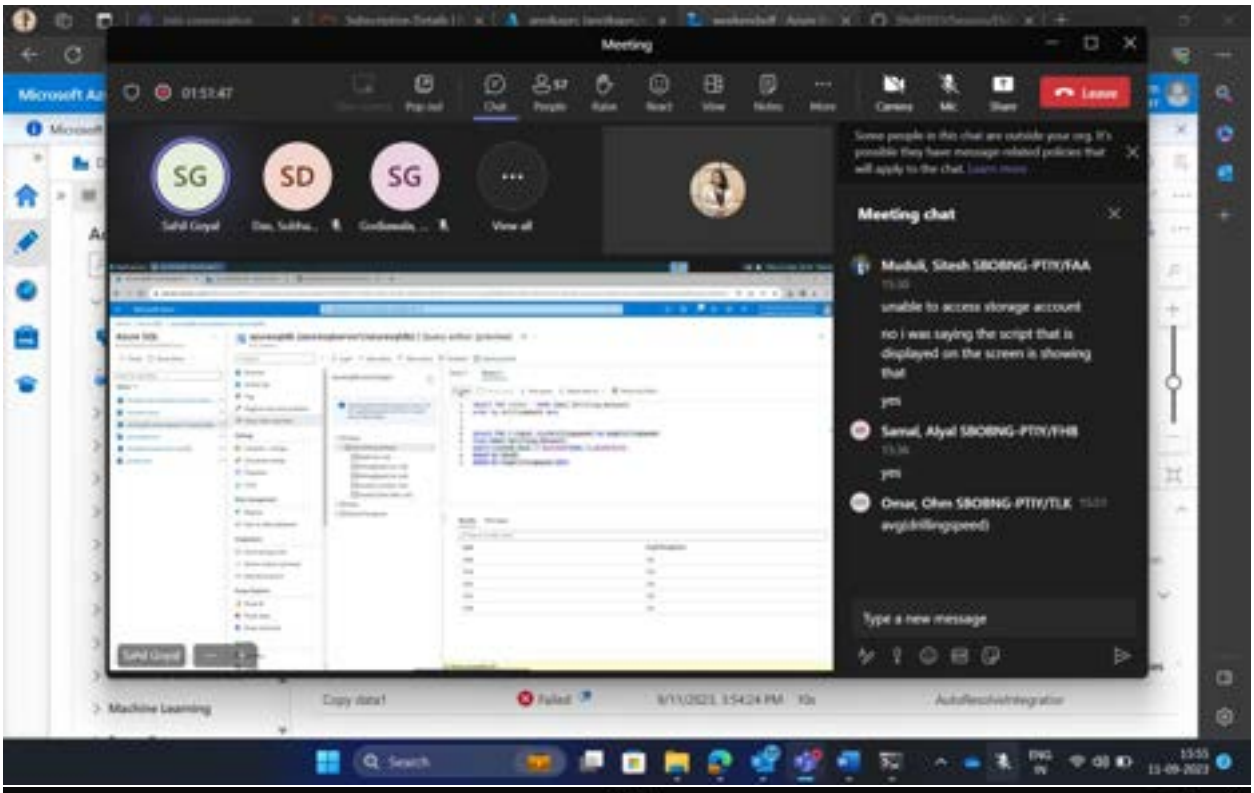

Microsoft Azure
Data Factory
Factory Resources
Pipelines
Change Data Capture (preview)
Datasets
Data flows
Power Query
Properties
General
Connection
Schema
Parameters
Linked service
Test connection
Edit
New
File path
Compression type
Column delimiter
Row delimiter
Encoding

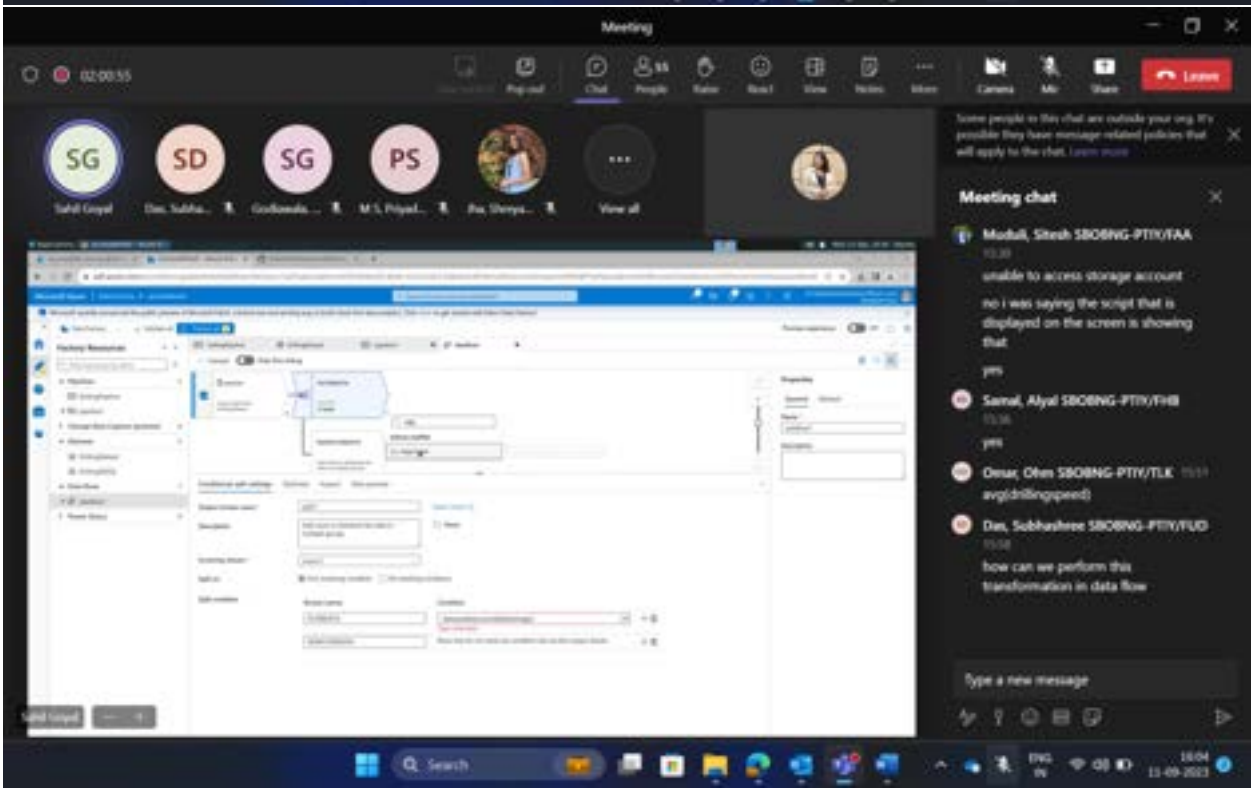

Meeting
Meeting chat
saving the files manually only right?
run this script in the terminal
Mudull, Sitesh SBOBNG-PTIY/FAA
unable to access storage account
no i was saying the script that is displayed on the screen is showing that
yes
Samal, Alyal SBOBNG-PTIY/FHB
yes
Type a new message

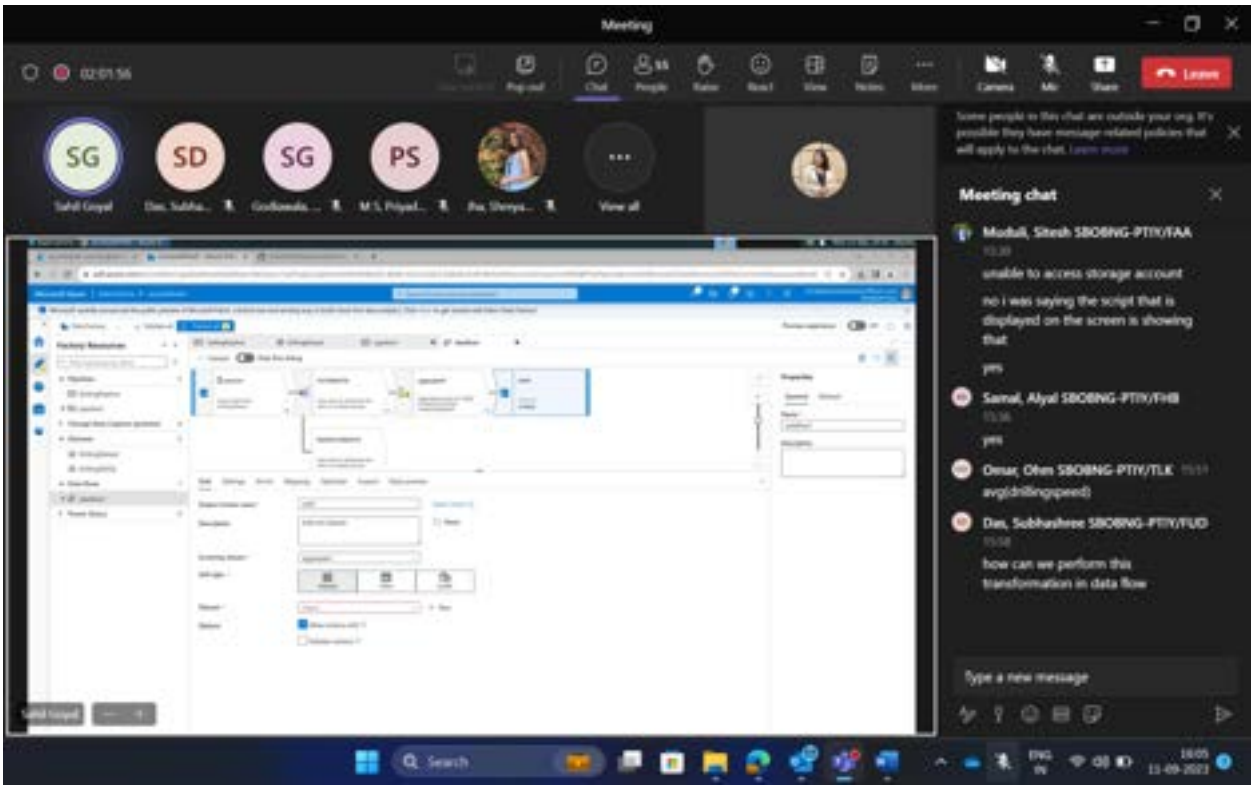

Microsoft Azure
Data Factory
Activities
Copy data
Copy data1
General
Source
Sink
Mapping
Settings
User properties

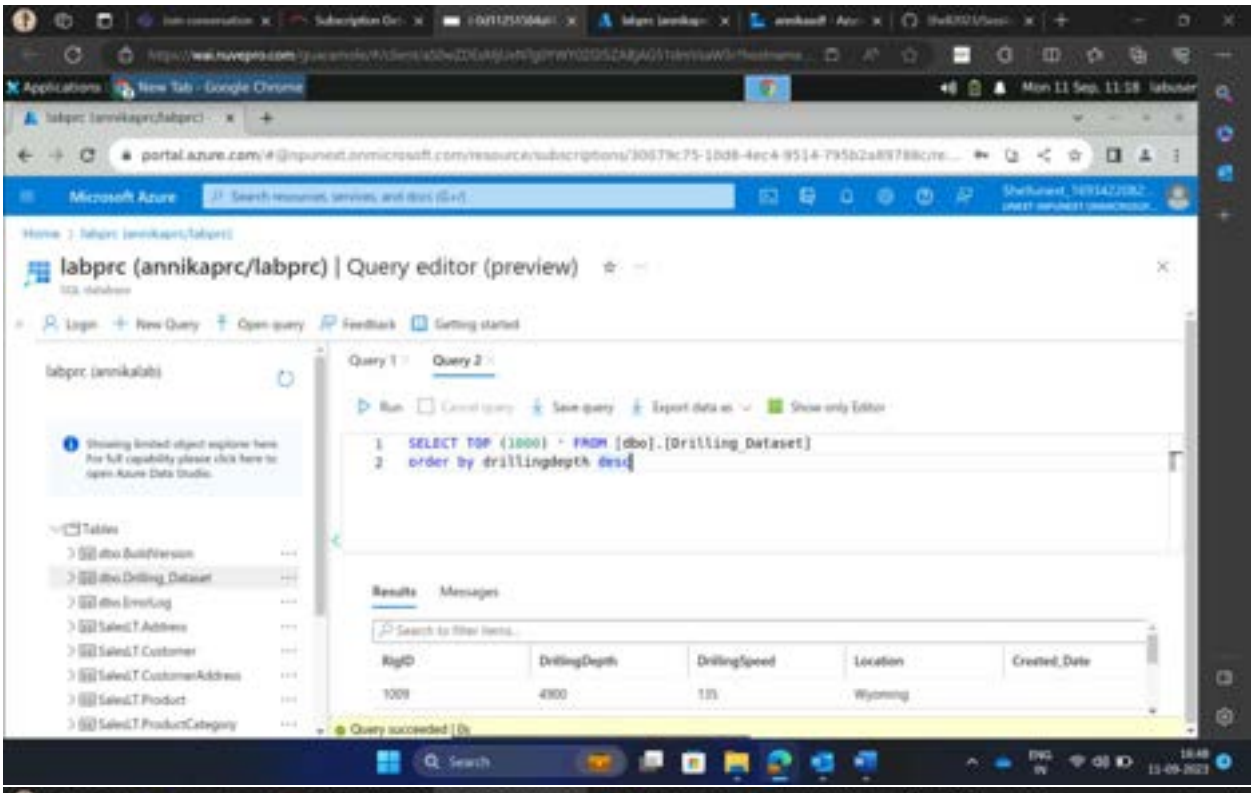

Meeting
01:47:44
Leave
SG
SD
SG
Meeting chat
Mudull, Sitesh SBOBNG-PTIY/FAA
unable to access storage account
no i was saying the script that is displayed on the screen is showing that
yes
Samal, Alyal SBOBNG-PTIY/FHB
yes
Omar, Ohm SBOBNG-PTIY/TLK
avgdrillingspeed)
Type a new message
Machine Learning
Recursively
Enable partitions discovery

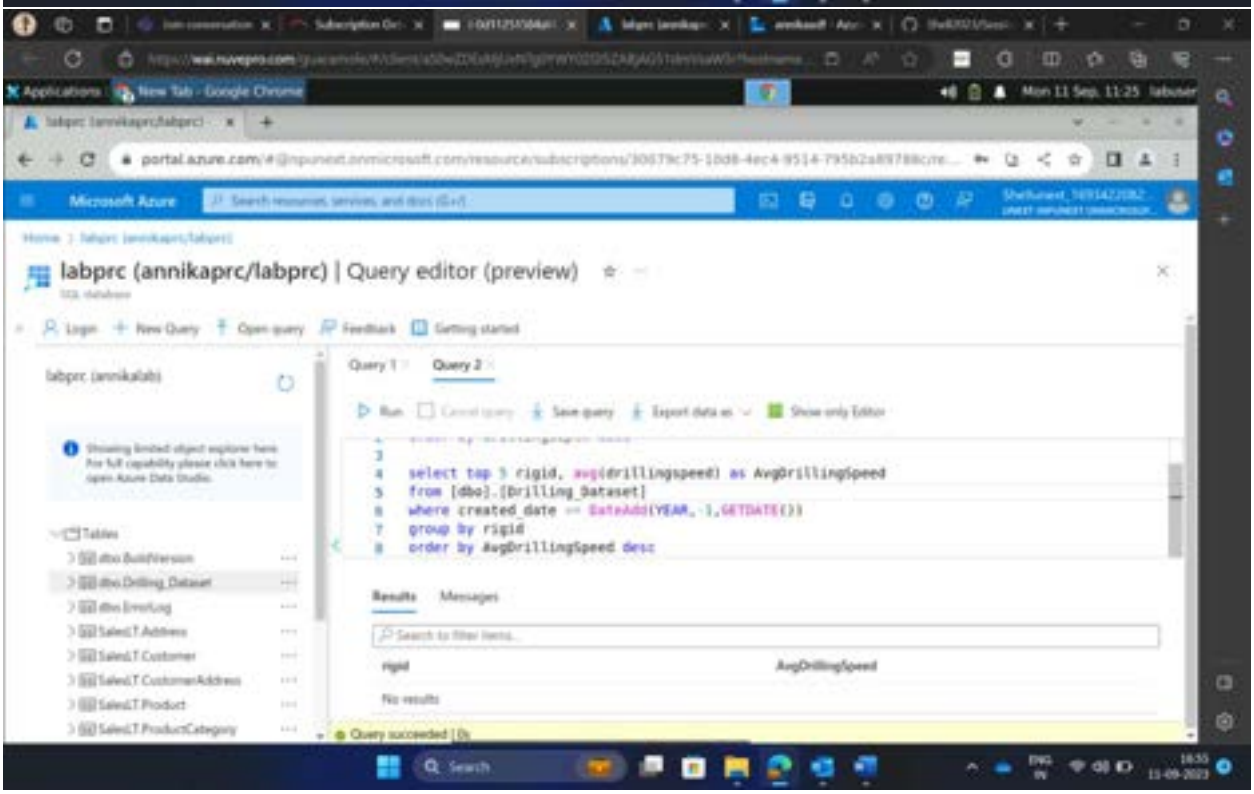

Meeting
01:50:50
Leave
SG
SD
SG
Meeting chat
Mudull, Sitesh SBOBNG-PTIY/FAA
unable to access storage account
no i was saying the script that is displayed on the screen is showing that
yes
Samal, Alyal SBOBNG-PTIY/FHB
yes
Omar, Ohm SBOBNG-PTIY/TLK
avgdrillingspeed)
Type a new message
Machine Learning
Copy data1
Failed


Meeting
Meeting chat
Modali, Sitesh SBOBNG-PTIY/FAA
unable to access storage account
no i was saying the script that is displayed on the screen is showing that
yes
Samal, Alyal SBOBNG-PTIY/FHB
yes
Onur, Ohm SBOBNG-PTIY/TLK
avg(drillingspeed)
Type a new message


Meeting
Meeting chat
Modali, Sitesh SBOBNG-PTIY/FAA
unable to access storage account
no i was saying the script that is displayed on the screen is showing that
yes
Samal, Alyal SBOBNG-PTIY/FHB
yes
Onur, Ohm SBOBNG-PTIY/TLK
avg(drillingspeed)
Das, Subhashree SBOBNG-PTIY/FUD
how can we perform this transformation in data flow
Type a new message

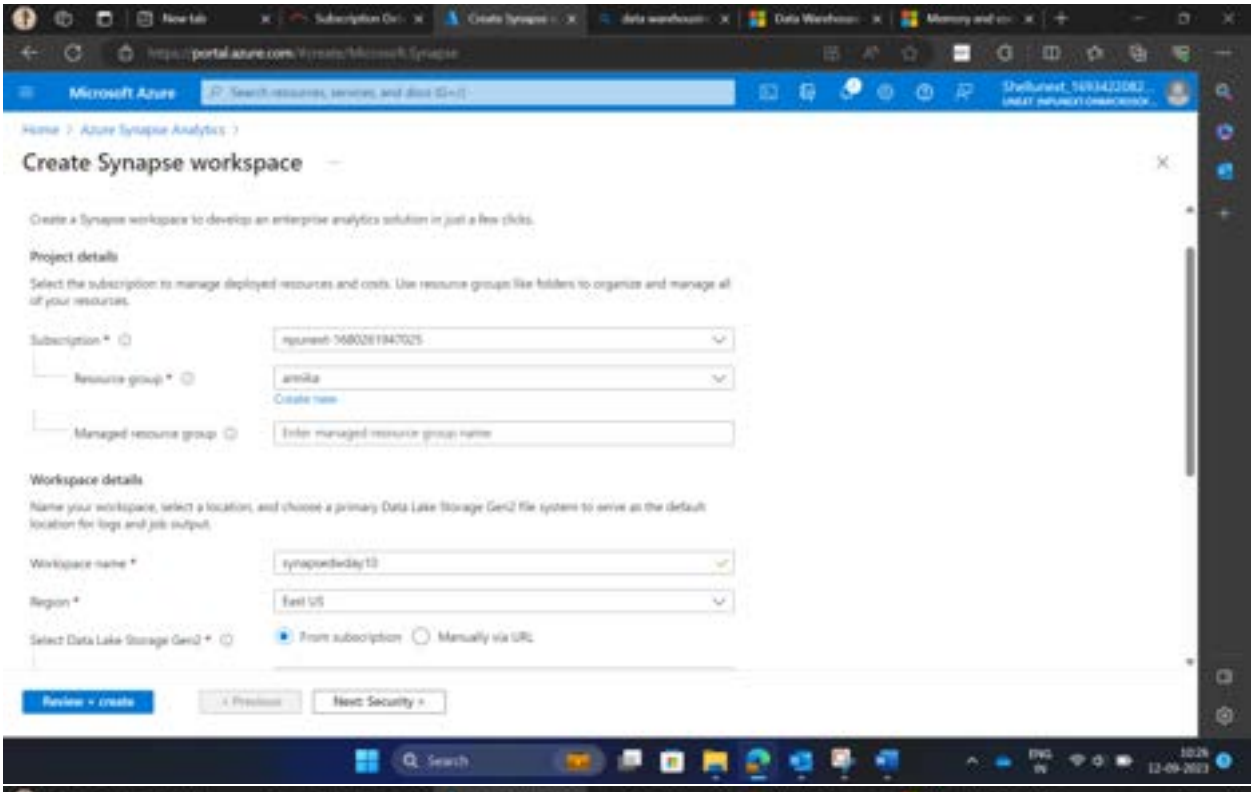

Meeting
Leave
SG
SD
SG
PS
View all
Meeting chat
Mudali, Sitesh SBOBNG-PTIY/TAA
unable to access storage account
no i was saying the script that is displayed on the screen is showing that
yes
Samal, Alyal SBOBNG-PTIY/FHB
yes
Omar, Ohm SBOBNG-PTIY/TLK
avg(drillingspeed)
Das, Subhashree SBOBNG-PTIY/FUD
how can we perform this transformation in data flow
Type a new message
Search

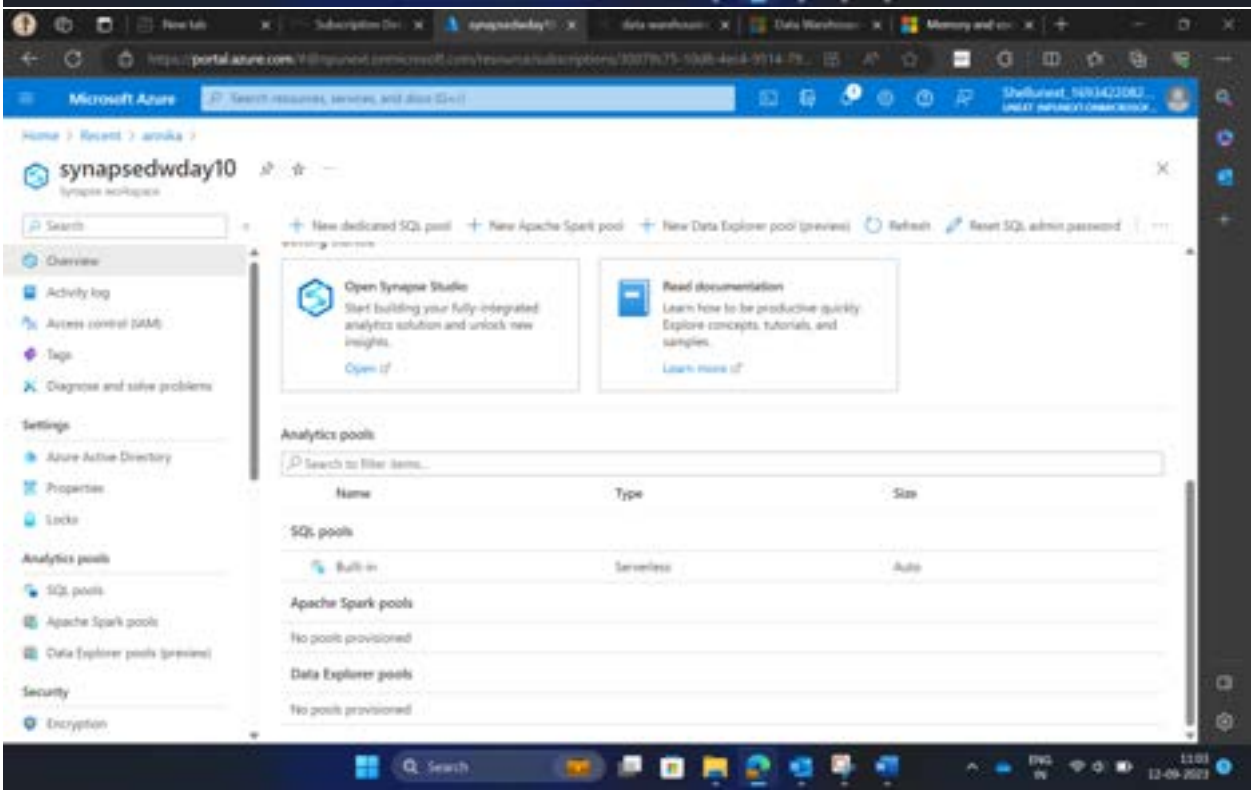

labprc (annikaprc/labprc) | Query editor (preview)
SELECT TOP (1000) * FROM [dbo].[Drilling_Dataset]
order by drillingdepth desc
RigID
DrillingDepth
DrillingSpeed
Location
Created_Date
1009
4900
135
Wyoming


labprc (annikaprc/labprc) | Query editor (preview)
select top 5 rigid, avg(drillingspeed) as AvgDrillingSpeed
from [dbo].[Drilling_Dataset]
where created_date >= DateAdd(YEAR, -1,GETDATE())
group by rigid
order by AvgDrillingSpeed desc
rigid
AvgDrillingSpeed
No results

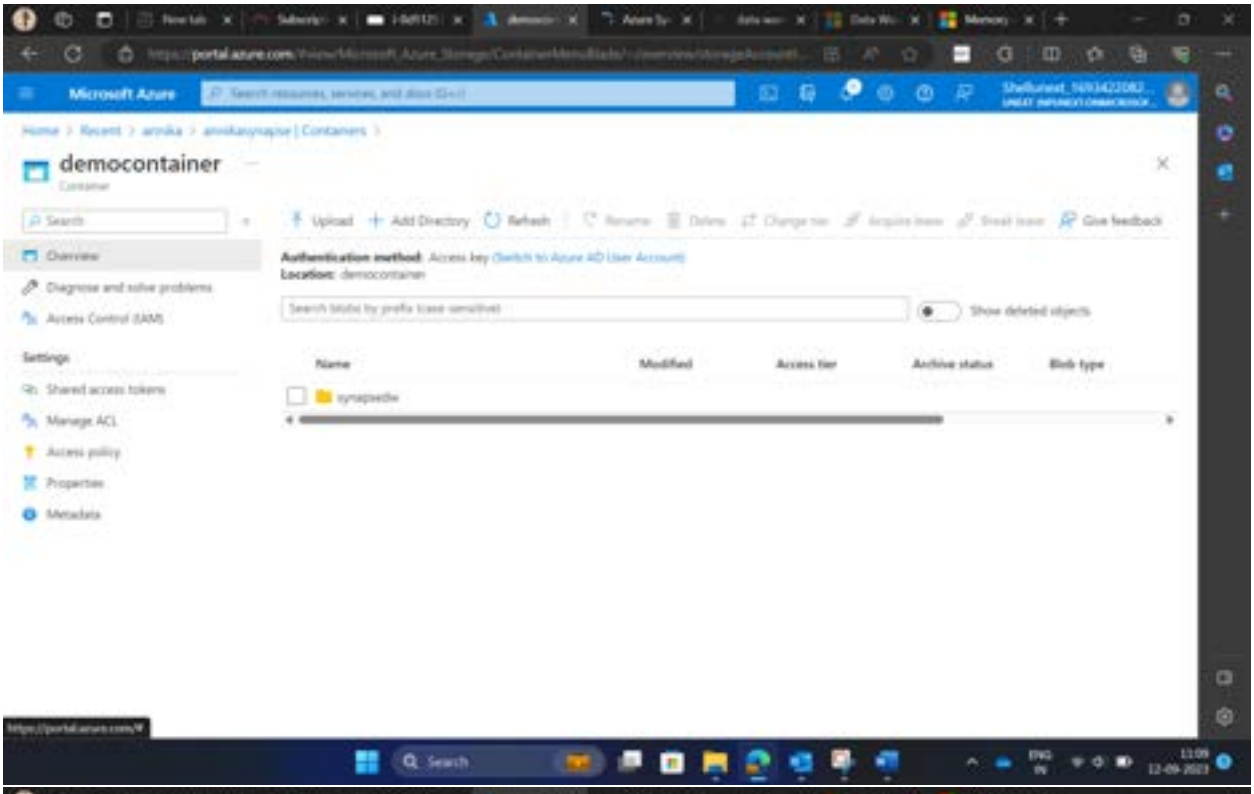
Azure Synapse Analytics 12.09.23 day 10

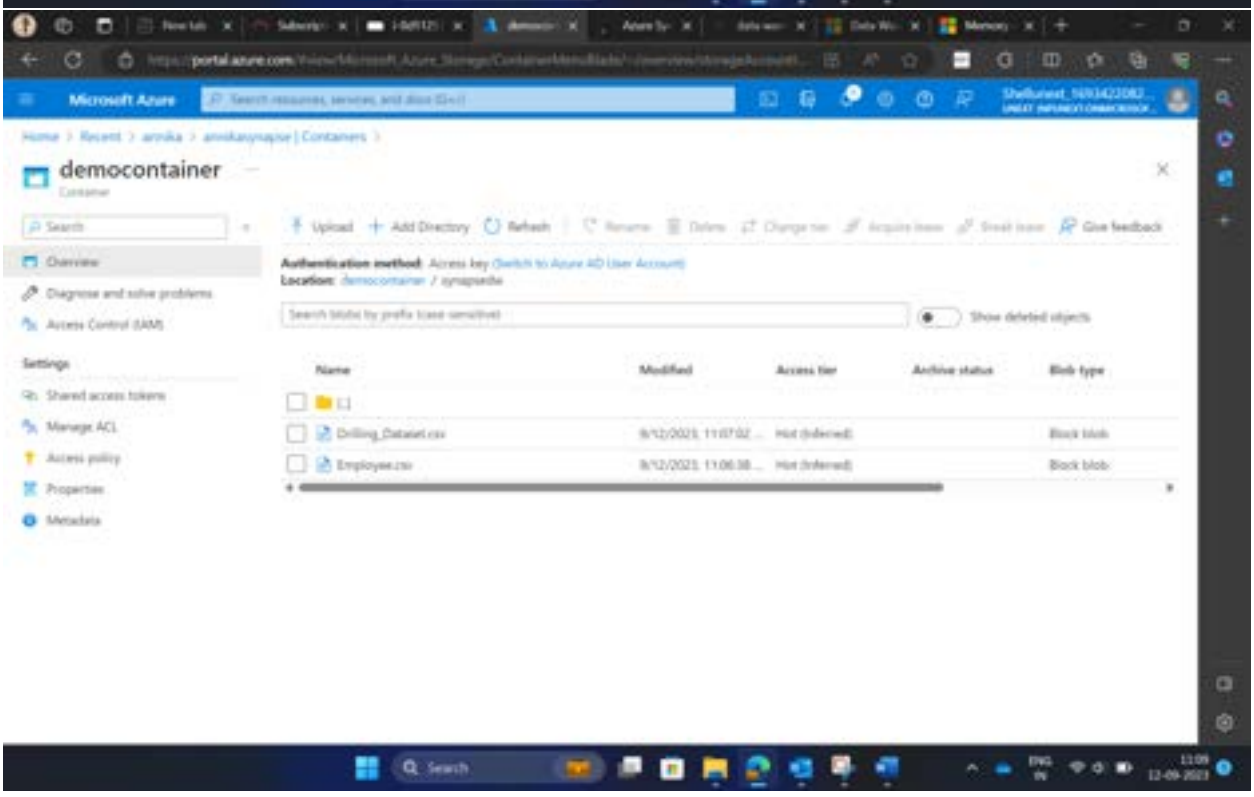
[Memory and concurrency limits - Azure Synapse Analytics | Microsoft Learn]

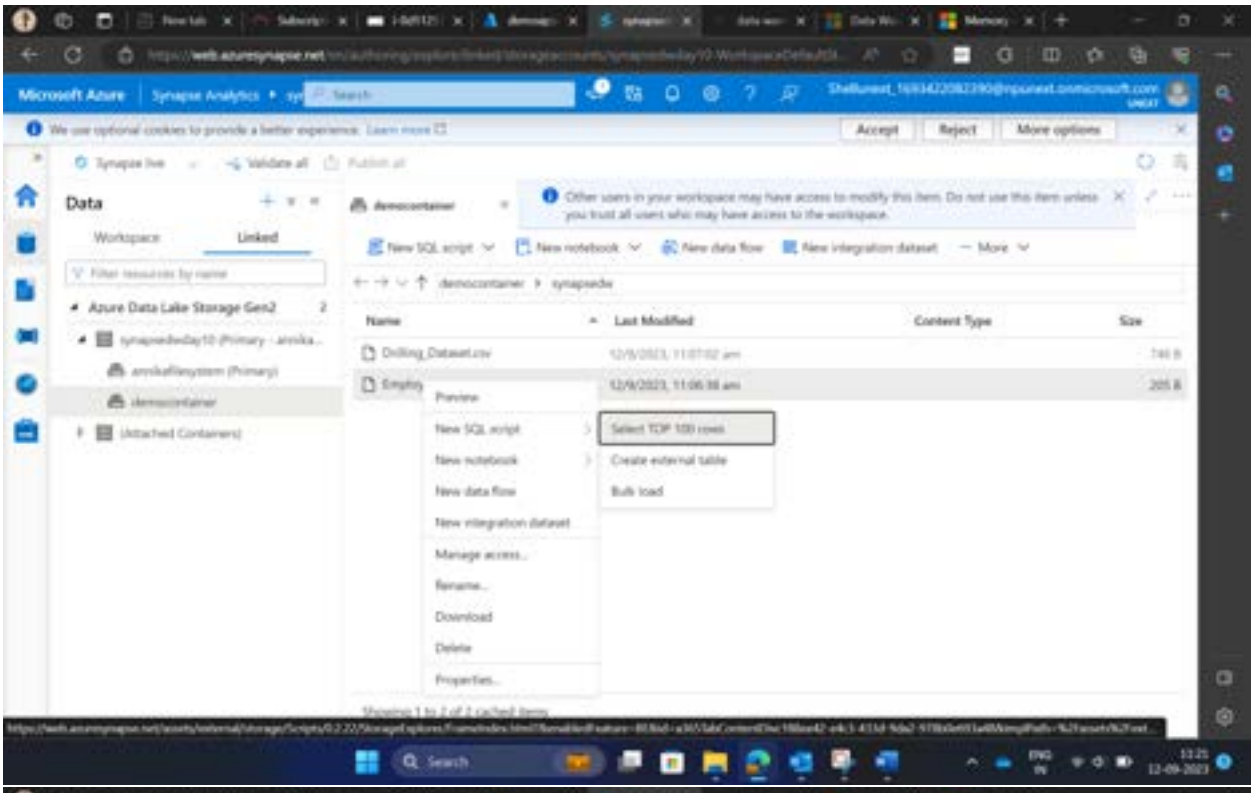


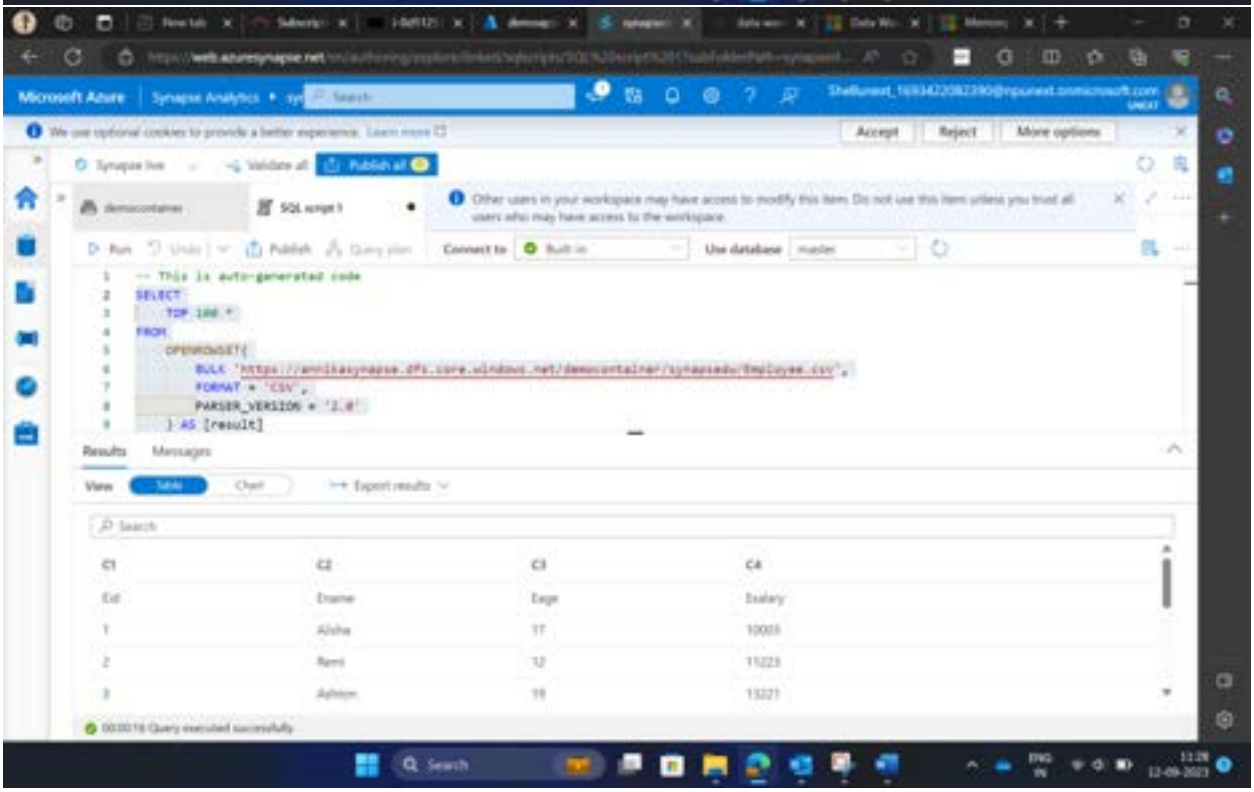


Serverless pool is created for us automatically

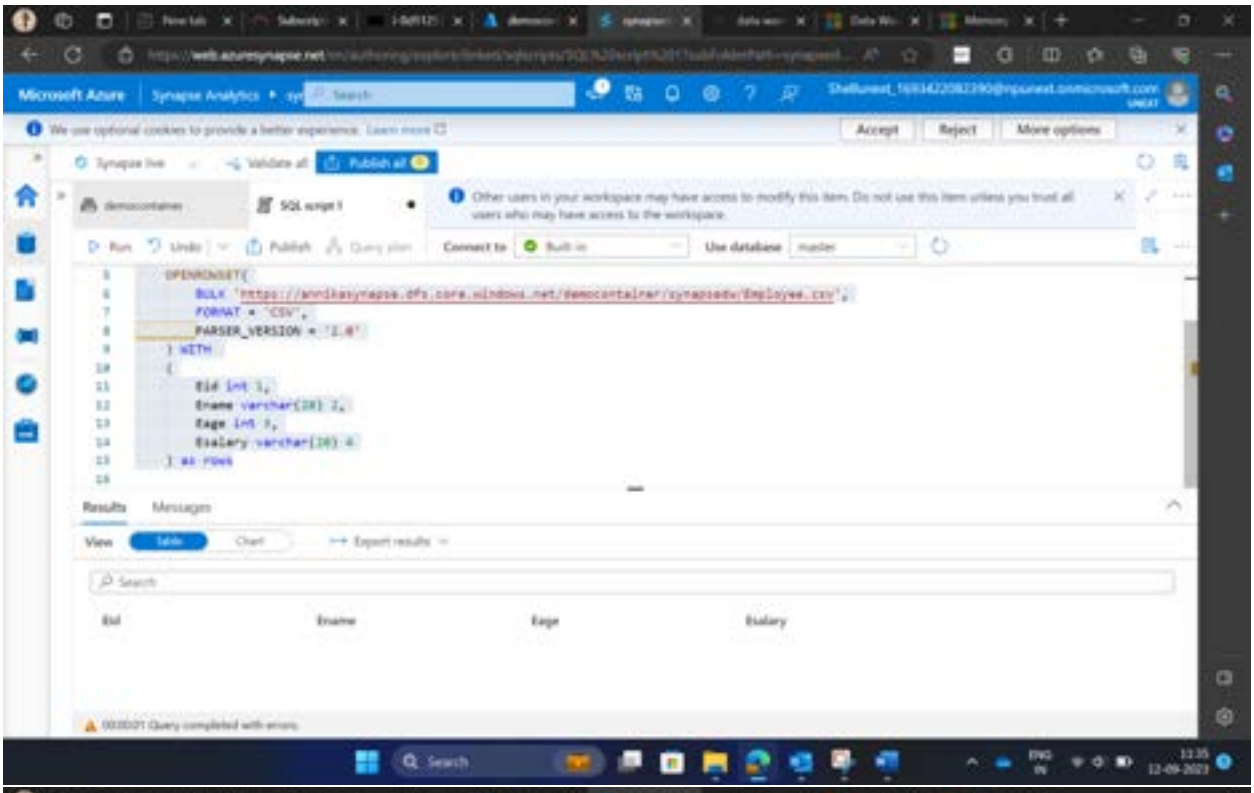


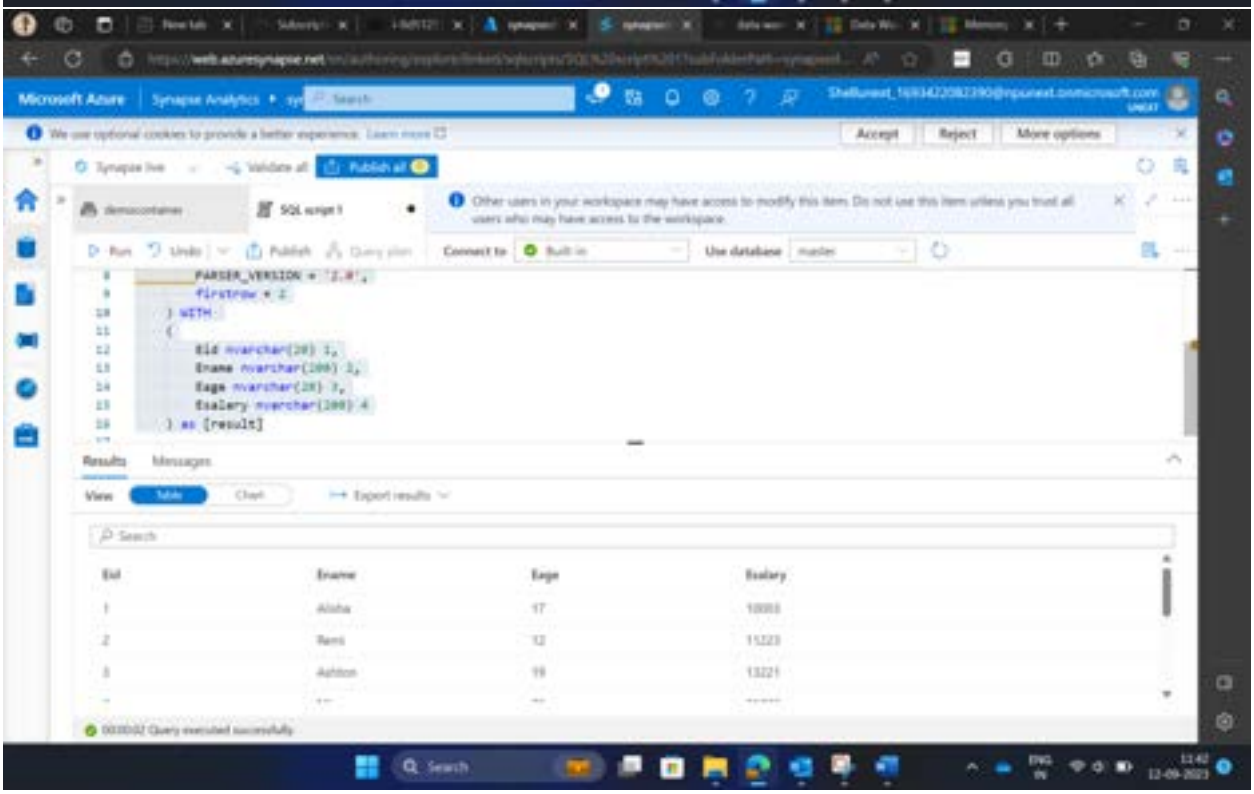


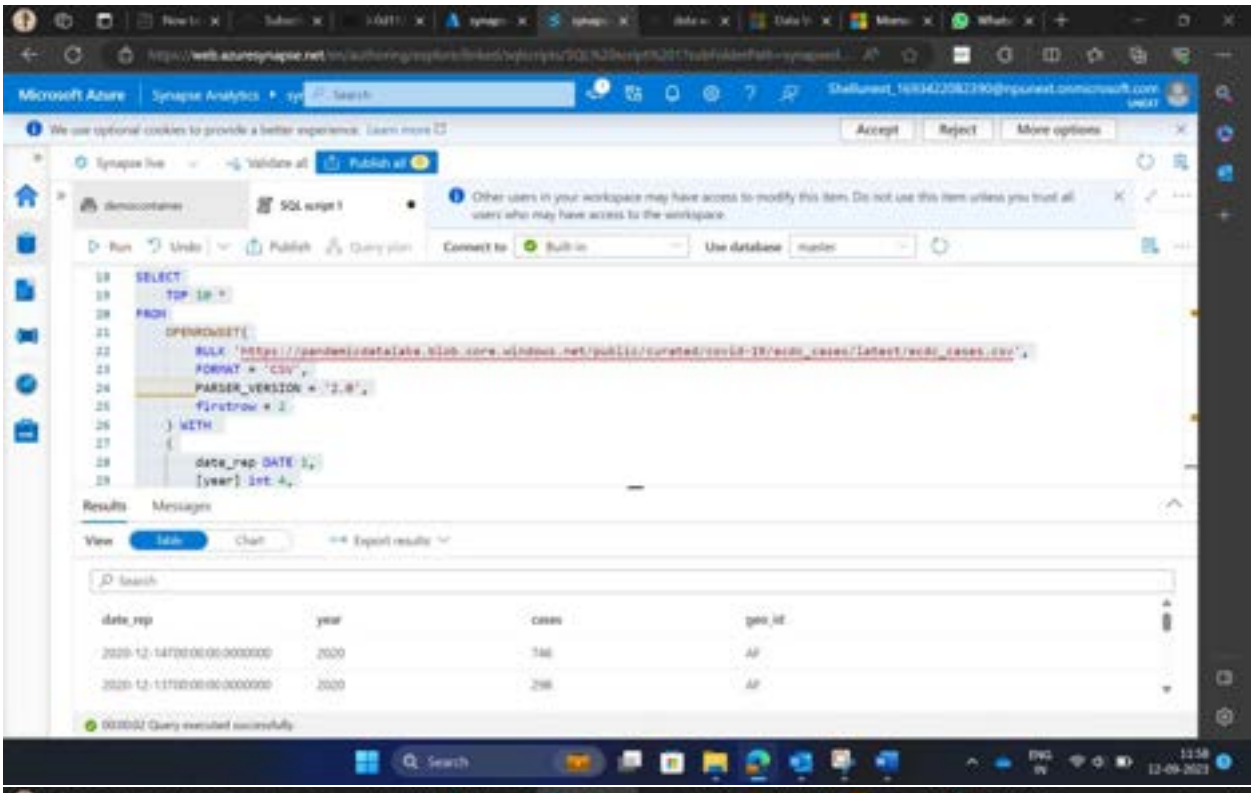


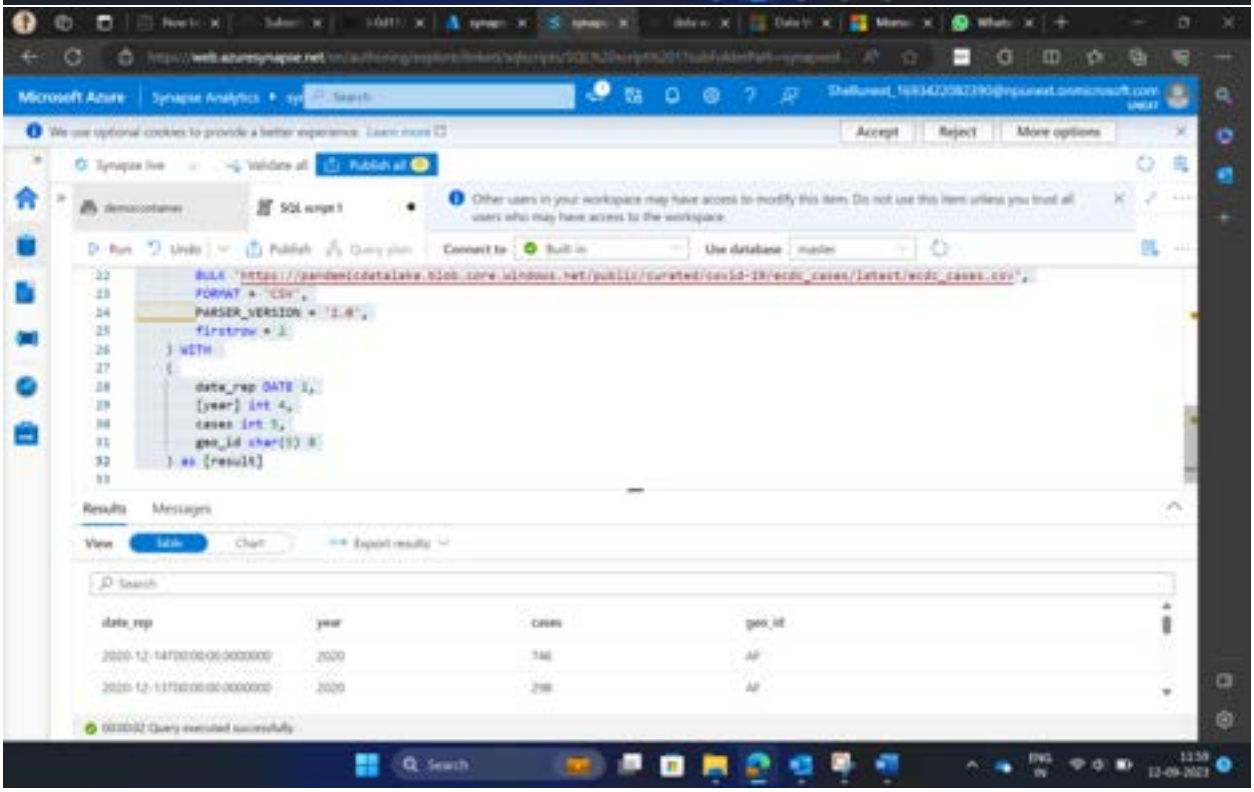


If row headings are not present in csv file

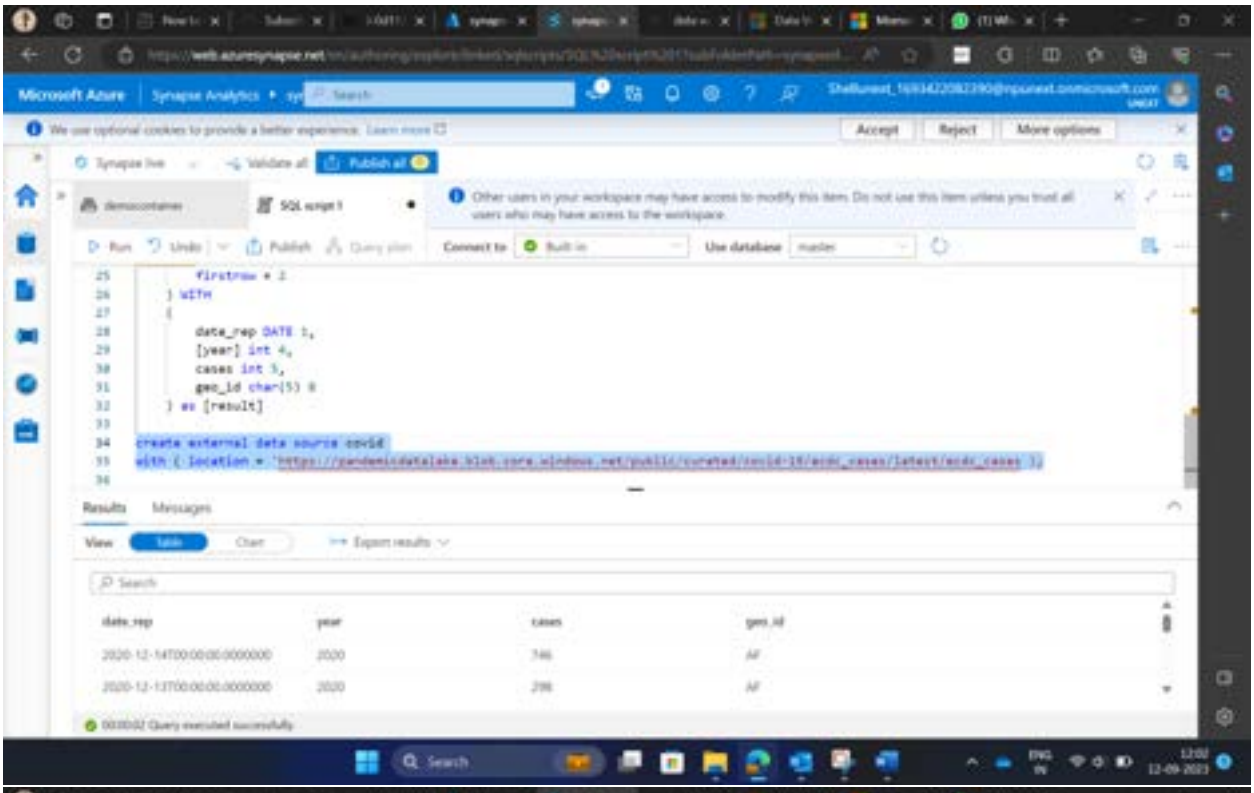

Microsoft Azure
Synapse Analytics
SQL script 1
Run
Undo
Publish
Connect to
Built-in
Use database
master
OPENROWSET(
BULK 'https://anrikasynapse.dfs.core.windows.net/democontainer/synapsedw/Employee.csv',
FORMAT = 'CSV',
PARSER_VERSION = '2.0'
) WITH
(
Eid int 1,
Ename varchar(20) 2,
Eage int 3,
Esalary varchar(20) 4
) as rows
Results
Messages
View
Table
Chart
Export results
Search
Eid
Ename
Eage
Esalary

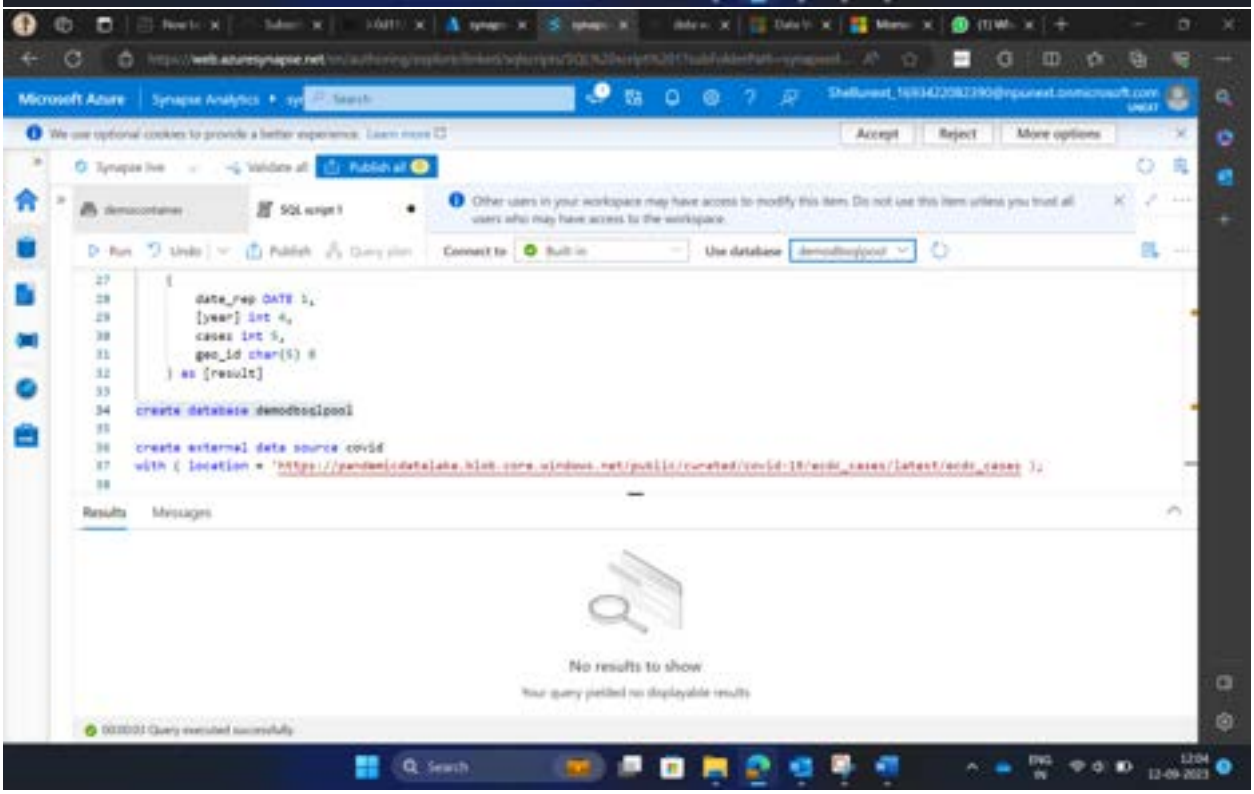

Microsoft Azure
Synapse Analytics
SQL script 1
Run
Undo
Publish
Connect to
Built-in
Use database
master
PARSER_VERSION = '2.0',
firstrow = 2
) WITH
(
Eid nvarchar(20) 1,
Ename nvarchar(100) 2,
Eage nvarchar(20) 3,
Esalary nvarchar(100) 4
) as [result]
Results
Messages
View
Table
Chart
Export results
Search
Eid
Ename
Eage
Esalary
1
Alisha
17
2
Rami
12
3
Ashton
19

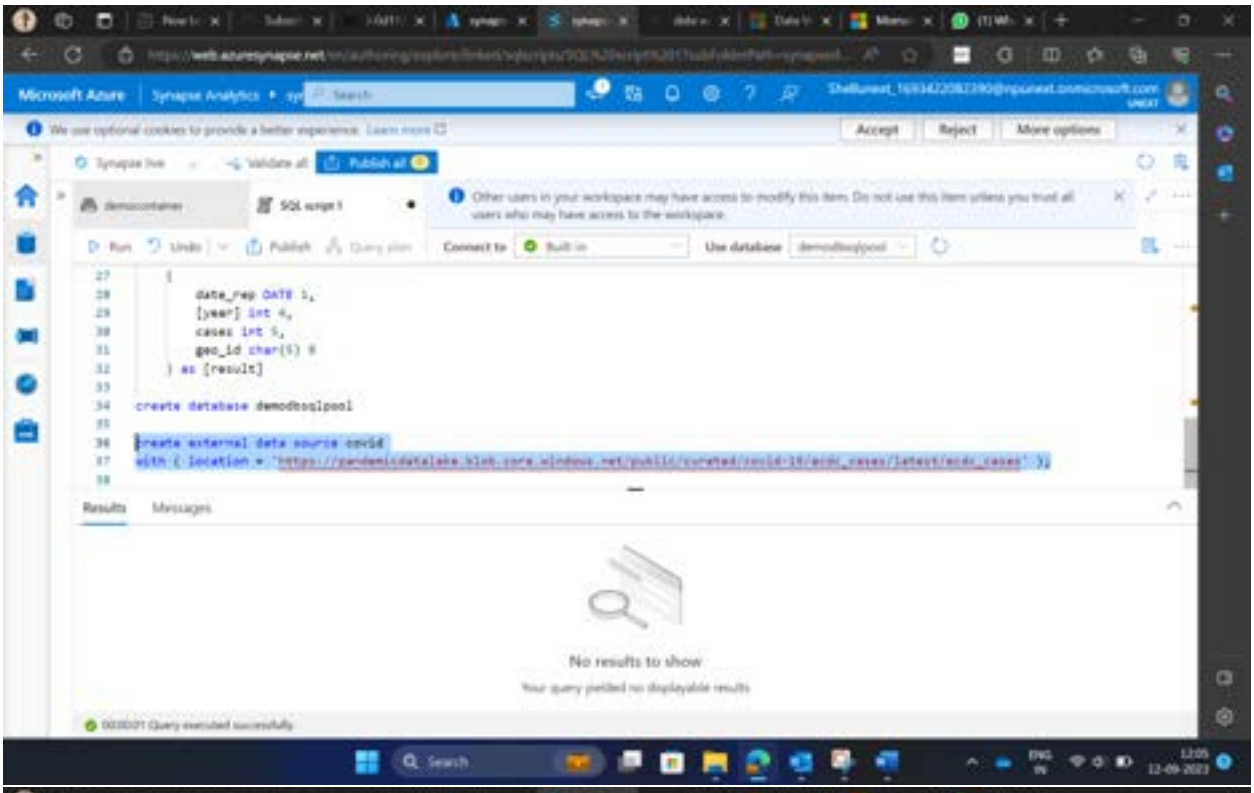


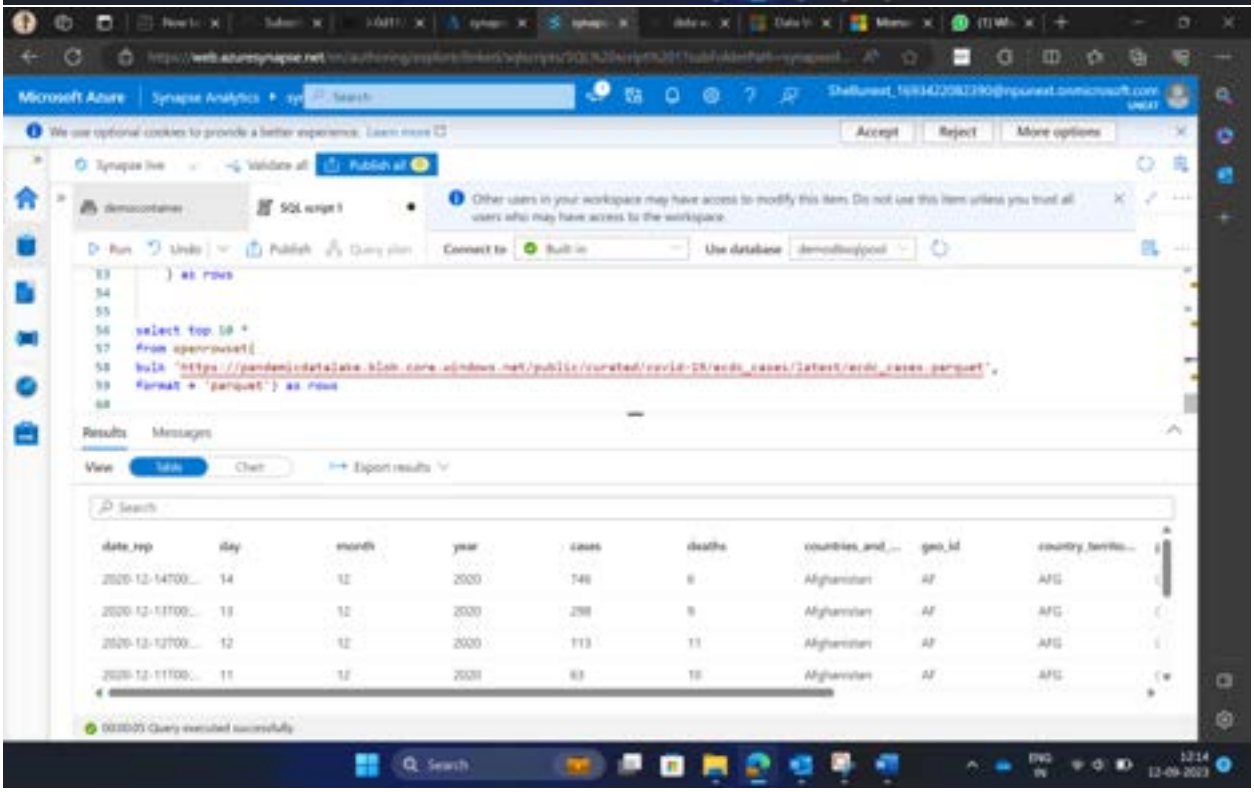


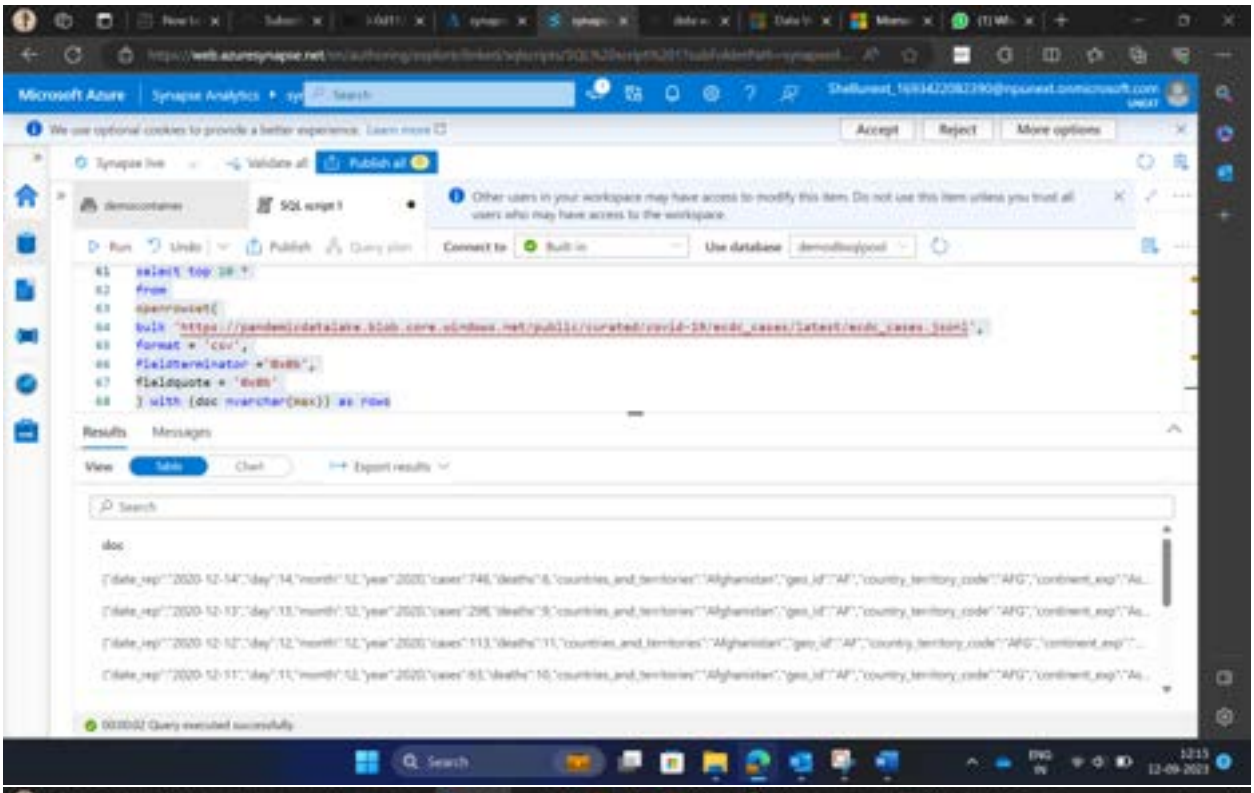
To create external source link

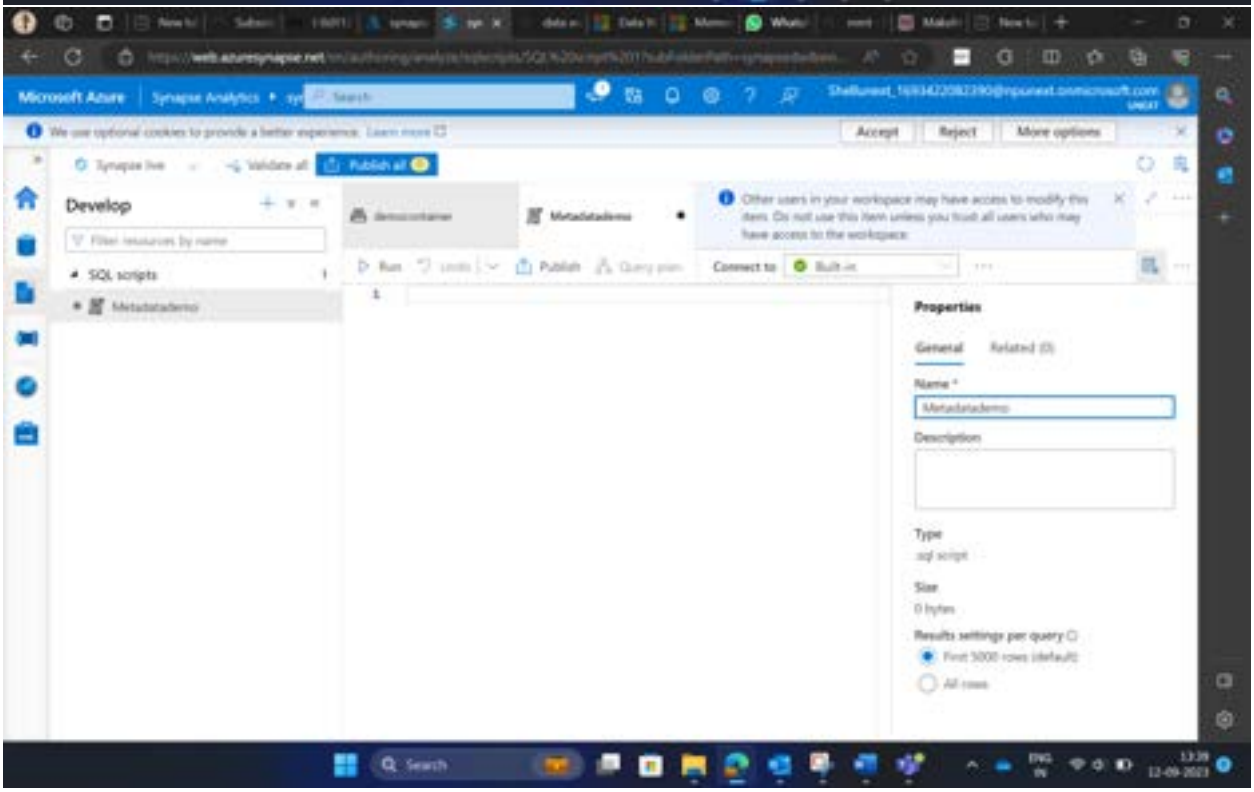

Microsoft Azure
Synapse Analytics
Accept
Reject
More options
Publish all
SQL script 1
Connect to
Use database
master
firstrow = 2
) WITH
(
date_rep DATE 1,
[year] int 4,
cases int 5,
geo_id char(5) 8
) as [result]
create external data source covid
with (location = 'https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/ecdc_cases/latest/ecdc_cases');
Results
Messages
View
Table
Chart
Export results
Search
date_rep
year
cases
geo_id
2020-12-14T00:00:00.0000000
2020
746
AF
2020-12-13T00:00:00.0000000
2020
298
AF
00:00:02 Query executed successfully.

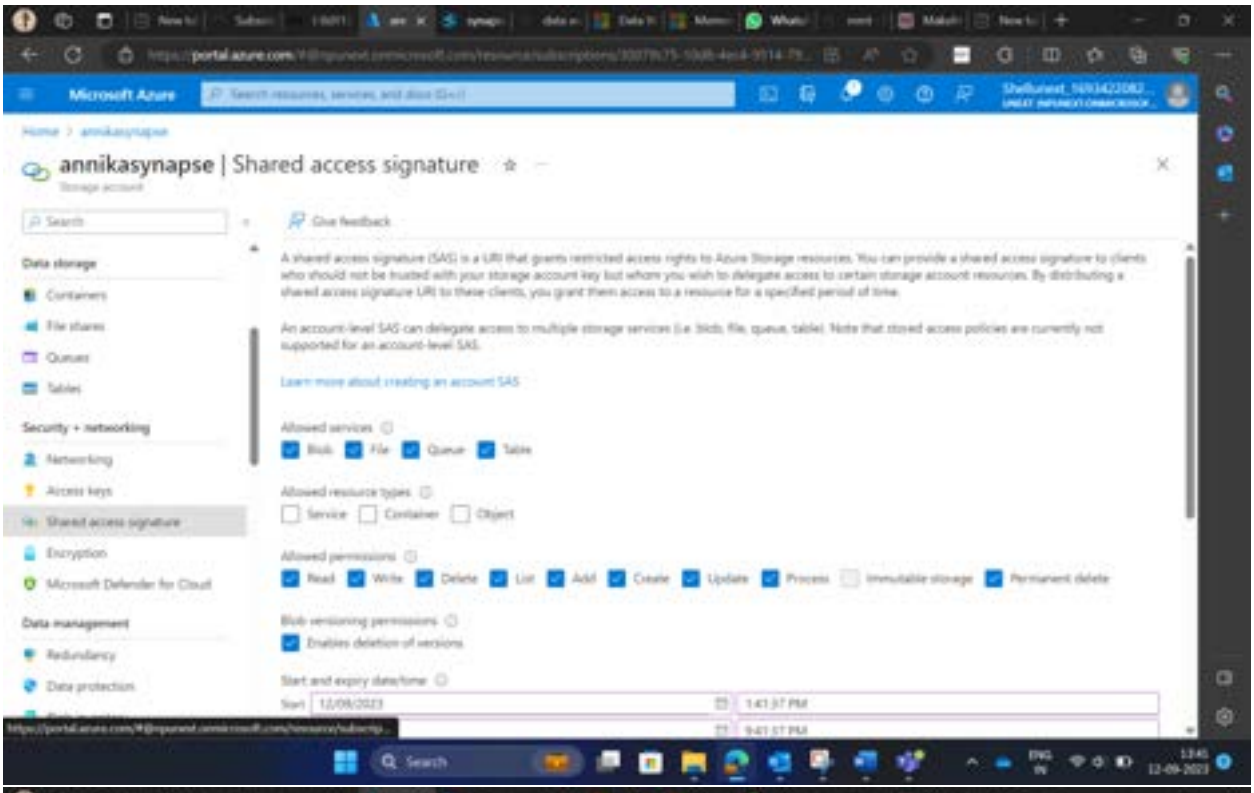

Microsoft Azure
Synapse Analytics
Accept
Reject
More options
Publish all
SQL script 1
Connect to
Use database
demodbsqlpool
(
date_rep DATE 1,
[year] int 4,
cases int 5,
geo_id char(5) 8
) as [result]
create database demodbsqlpool
create external data source covid
with (location = 'https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/ecdc_cases/latest/ecdc_cases');
Results
Messages
No results to show
Your query yielded no displayable results
00:00:03 Query executed successfully.

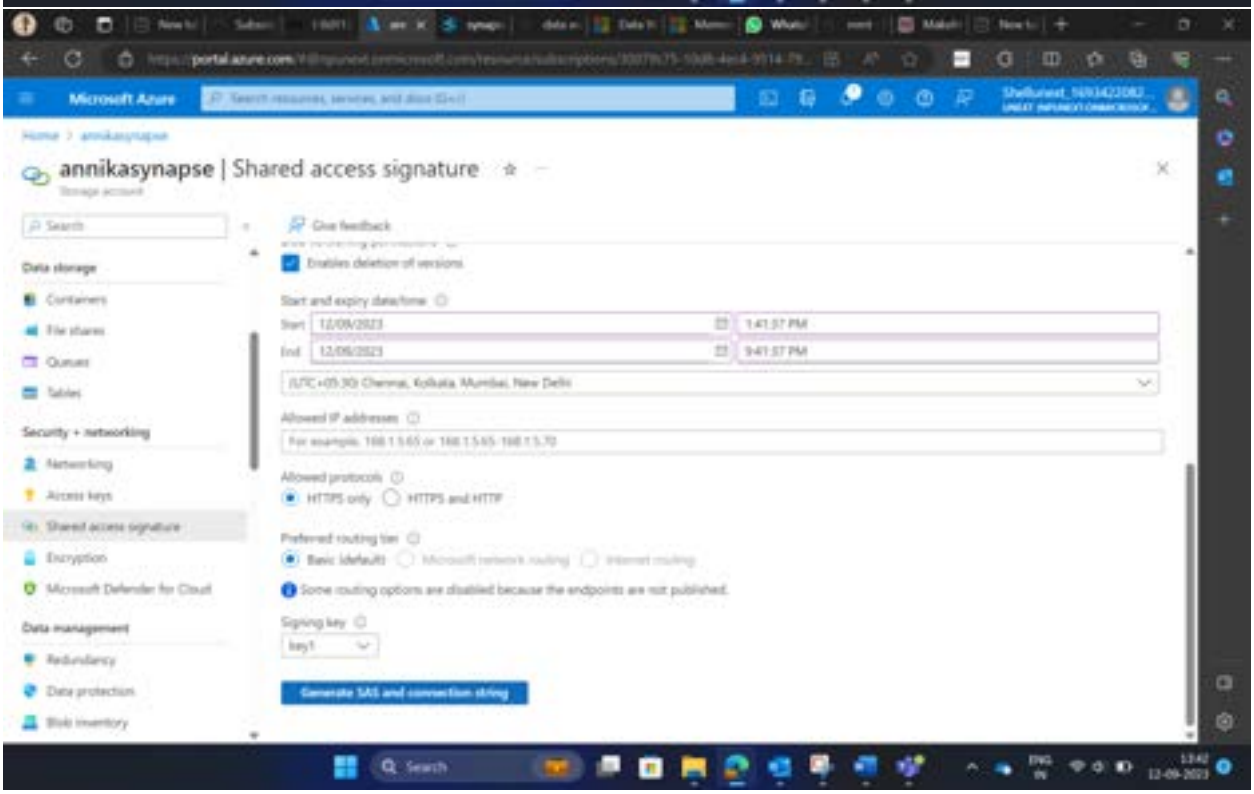

Microsoft Azure
Synapse Analytics
Accept
Reject
More options
SQL script 1
Connect to
Use database
Results
Messages
No results to show

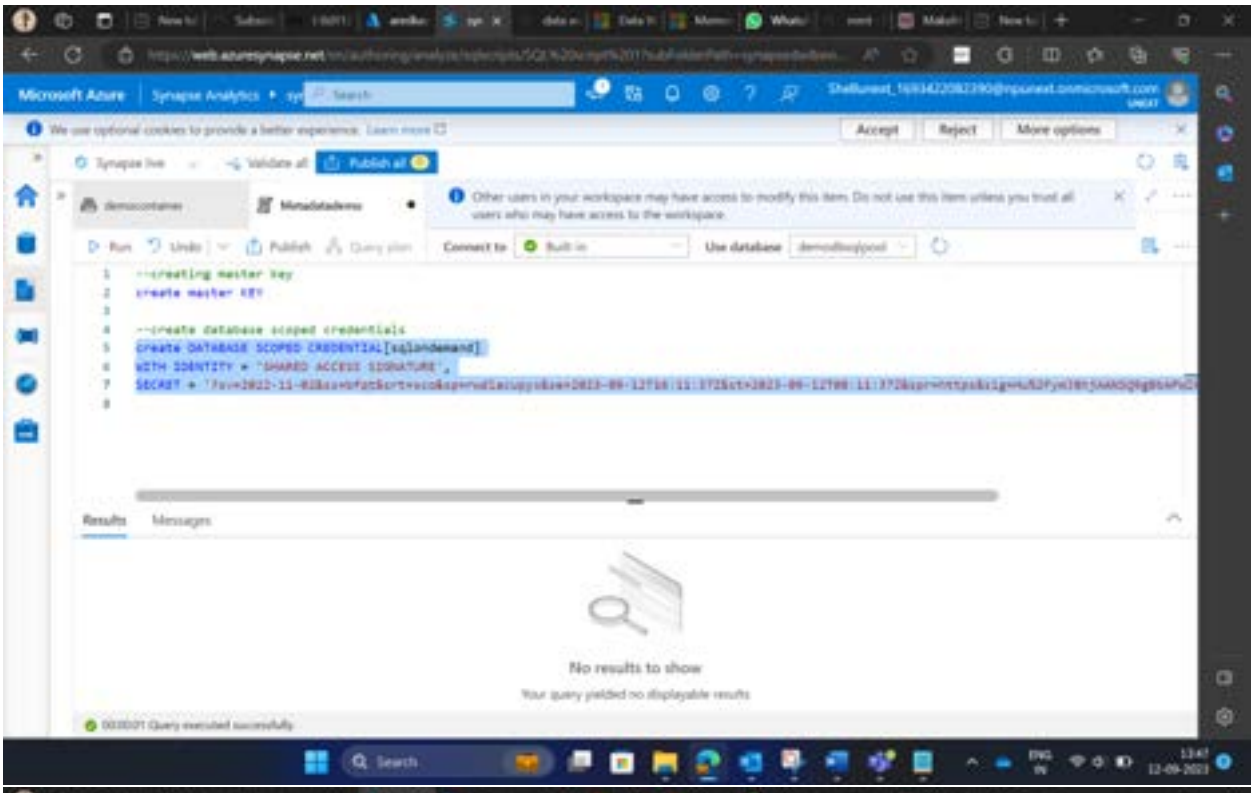

Microsoft Azure
Synapse Analytics
Accept
Reject
More options
SQL script 1
Connect to
Use database
Results
Messages
View
Table
Chart
Export results

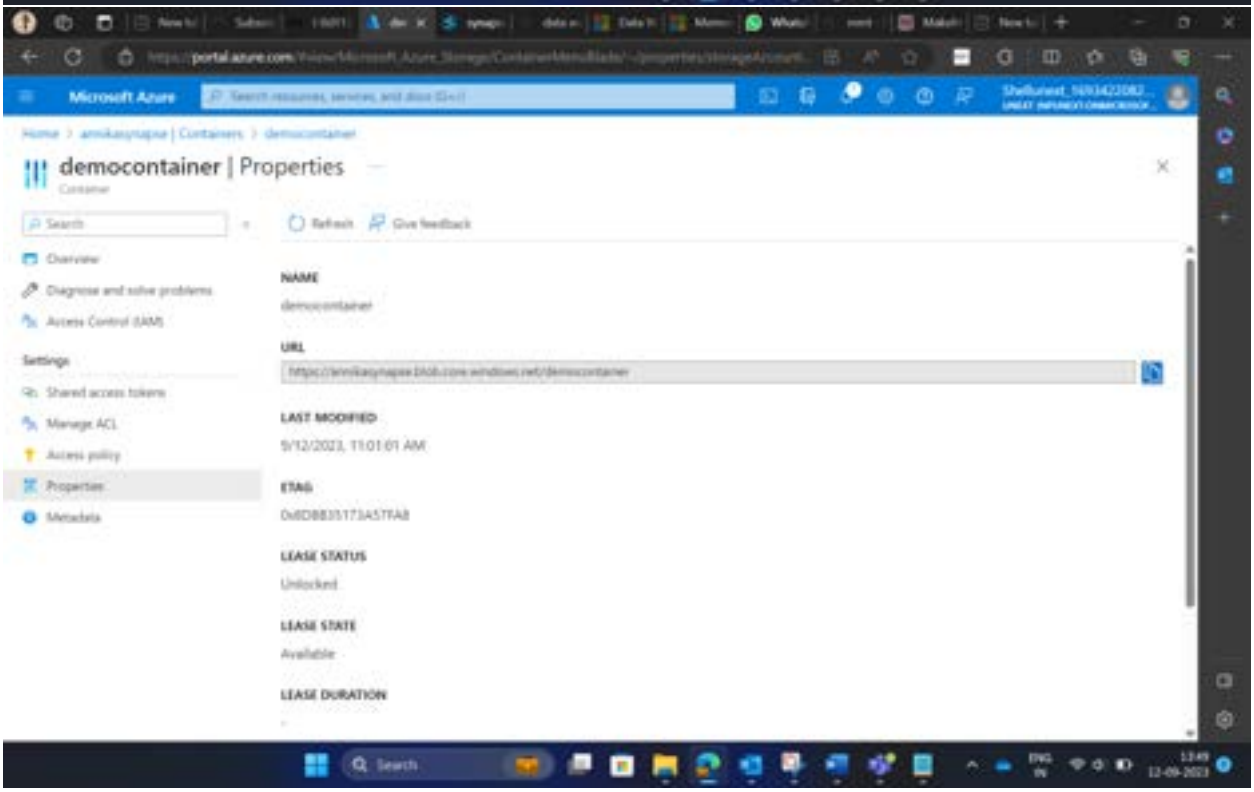

Microsoft Azure
Synapse Analytics
SQL script 1
Connect to
Use database
Results
Messages
doc

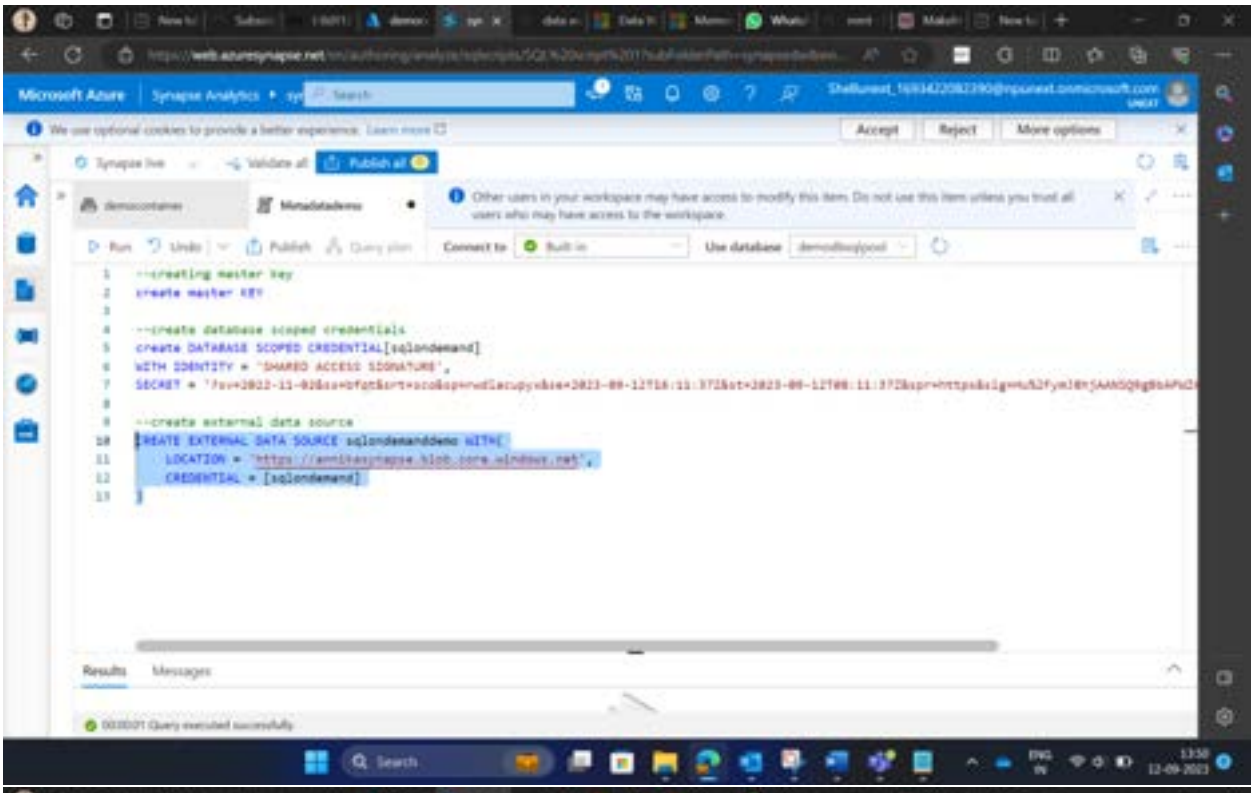

Microsoft Azure
Synapse Analytics
Develop
SQL scripts
Metadatademo
Properties
General
Name *
Description
Type
Size
Results settings per query

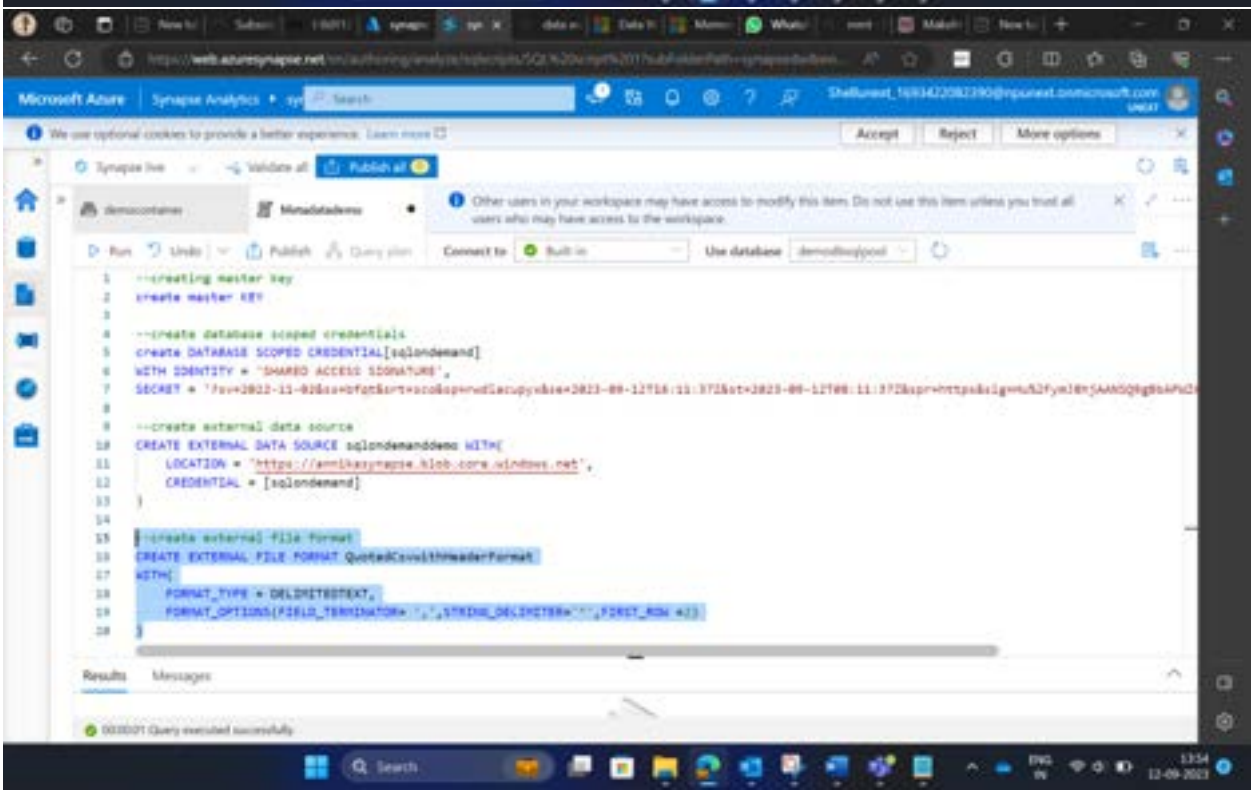

Microsoft Azure
annikasynapse | Shared access signature
Storage account
Search
Give feedback
Data storage
Containers
File shares
Queues
Tables
Security + networking
Networking
Access keys
Shared access signature
Encryption
Microsoft Defender for Cloud
Data management
Redundancy
Data protection
A shared access signature (SAS) is a URI that grants restricted access rights to Azure Storage resources. You can provide a shared access signature to clients who should not be trusted with your storage account key but whom you wish to delegate access to certain storage account resources. By distributing a shared access signature URI to these clients, you grant them access to a resource for a specified period of time.
An account-level SAS can delegate access to multiple storage services (i.e. blob, file, queue, table). Note that stored access policies are currently not supported for an account-level SAS.
Learn more about creating an account SAS
Allowed services
Blob
File
Queue
Table
Allowed resource types
Service
Container
Object
Allowed permissions
Read
Write
Delete
List
Add
Create
Update
Process
Immutable storage
Permanent delete
Blob versioning permissions
Enables deletion of versions
Start and expiry date/time
Start
12/09/2023
1:41:37 PM
9:41:37 PM

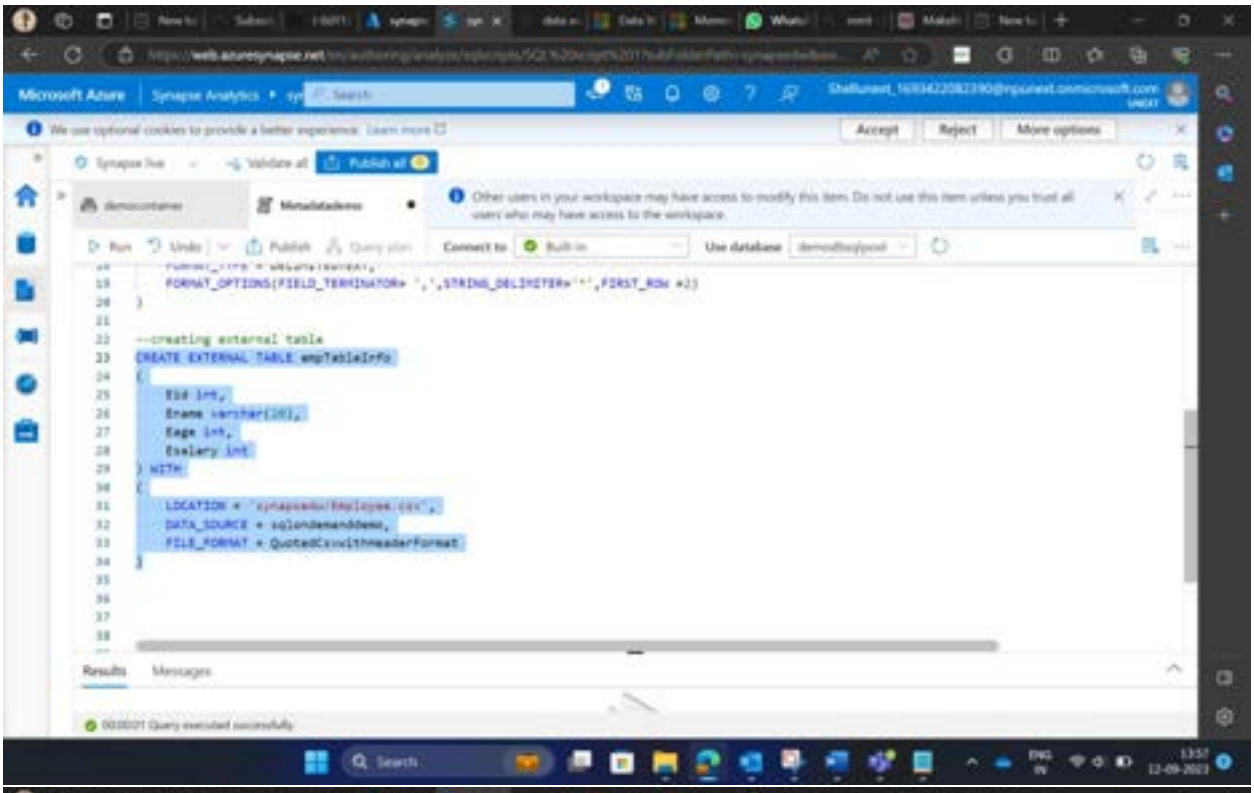

Microsoft Azure
annikasynapse | Shared access signature
Storage account
Search
Give feedback
Data storage
Containers
File shares
Queues
Tables
Security + networking
Networking
Access keys
Shared access signature
Encryption
Microsoft Defender for Cloud
Data management
Redundancy
Data protection
Blob inventory
Enables deletion of versions
Start and expiry date/time
Start
12/09/2023
1:41:37 PM
End
12/09/2023
9:41:37 PM
(UTC+05:30) Chennai, Kolkata, Mumbai, New Delhi
Allowed IP addresses
For example, 168.1.5.65 or 168.1.5.65-168.1.5.70
Allowed protocols
HTTPS only
HTTPS and HTTP
Preferred routing tier
Basic (default)
Some routing options are disabled because the endpoints are not published.
Signing key
key1
Generate SAS and connection string

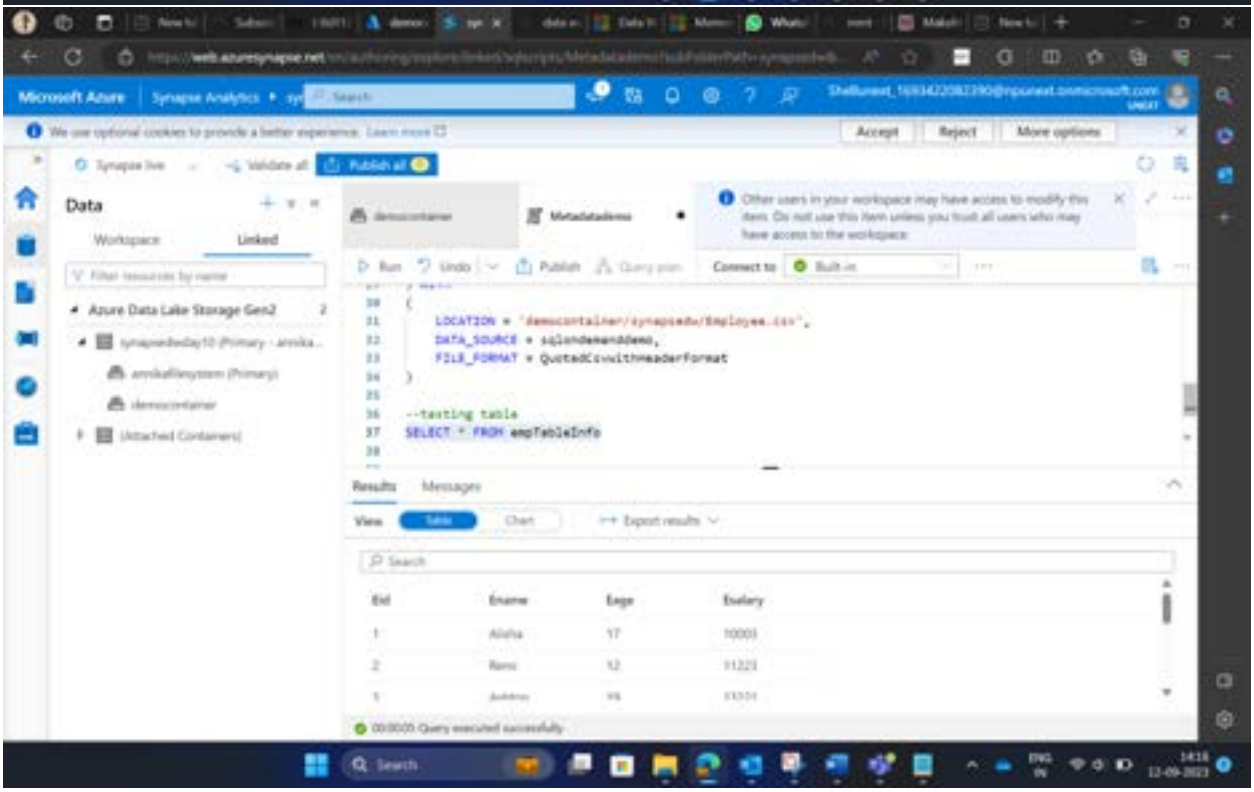

Microsoft Azure
Synapse Analytics
Accept
Reject
More options
Publish all
democontainer
Metadatademo
Connect to
Use database
--creating master key
create master KEY
--create database scoped credentials
CREATE DATABASE SCOPED CREDENTIAL [sqlondemand]
WITH IDENTITY = 'SHARED ACCESS SIGNATURE',
SECRET = '
Results
Messages
No results to show
Your query yielded no displayable results

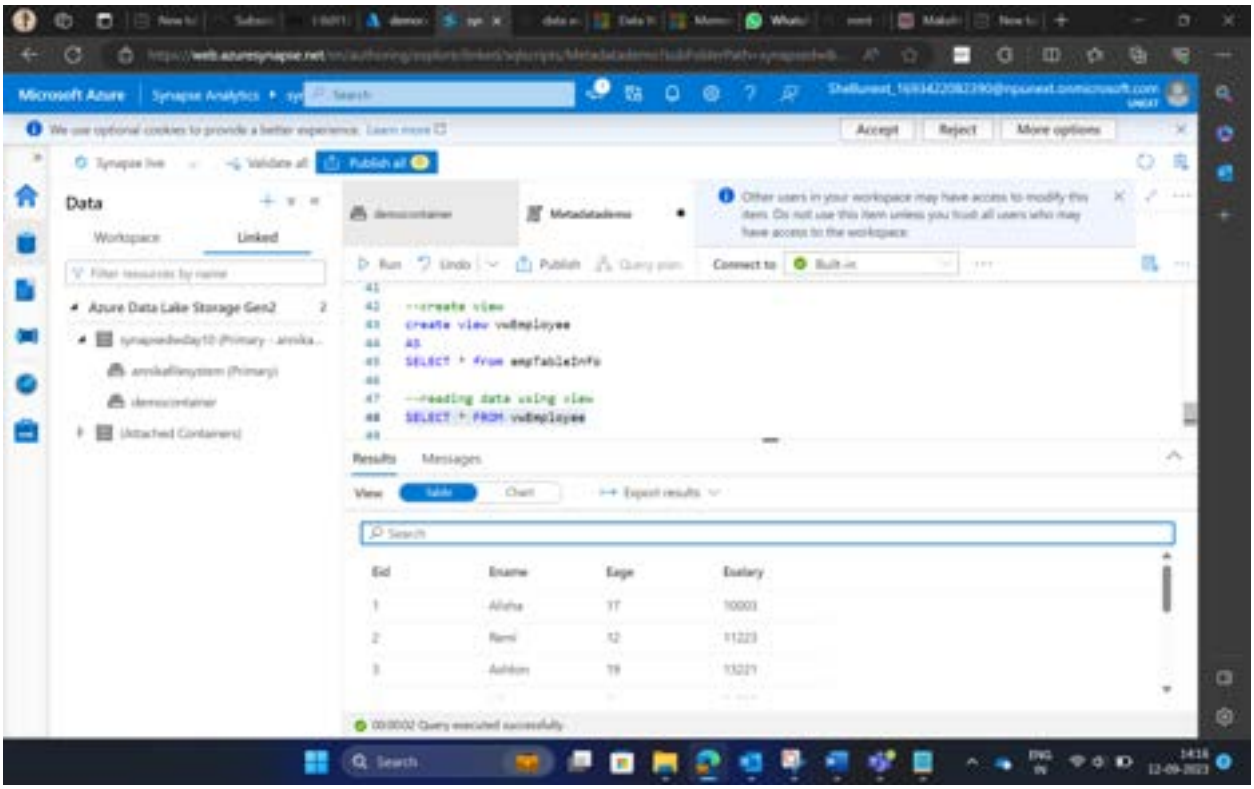

Microsoft Azure
democontainer | Properties
Container
Search
Refresh
Give feedback
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
NAME
democontainer
URL
LAST MODIFIED
9/12/2023, 11:01:01 AM
ETAG
LEASE STATUS
Unlocked
LEASE STATE
Available
LEASE DURATION


Microsoft Azure
Synapse Analytics
Accept
Reject
More options
Publish all
Run
Undo
Publish
Connect to
Use database
--creating master key
create master KEY
--create database scoped credentials
create DATABASE SCOPED CREDENTIAL[sqlondemand]
WITH IDENTITY = 'SHARED ACCESS SIGNATURE',
--create external data source
CREATE EXTERNAL DATA SOURCE sqlondemanddemo WITH(
LOCATION = 'https://amitkasynapse.blob.core.windows.net',
CREDENTIAL = [sqlondemand]
)
Results
Messages


Microsoft Azure
Synapse Analytics
Accept
Reject
More options
Publish all
Run
Undo
Publish
Connect to
Use database
--creating master key
create master KEY
--create database scoped credentials
create DATABASE SCOPED CREDENTIAL[sqlondemand]
WITH IDENTITY = 'SHARED ACCESS SIGNATURE',
--create external data source
CREATE EXTERNAL DATA SOURCE sqlondemanddemo WITH(
LOCATION = 'https://amitkasynapse.blob.core.windows.net',
CREDENTIAL = [sqlondemand]
)
--create external file format
CREATE EXTERNAL FILE FORMAT QuotedCsvWithHeaderFormat
WITH(
FORMAT_TYPE = DELIMITEDTEXT,
Results
Messages

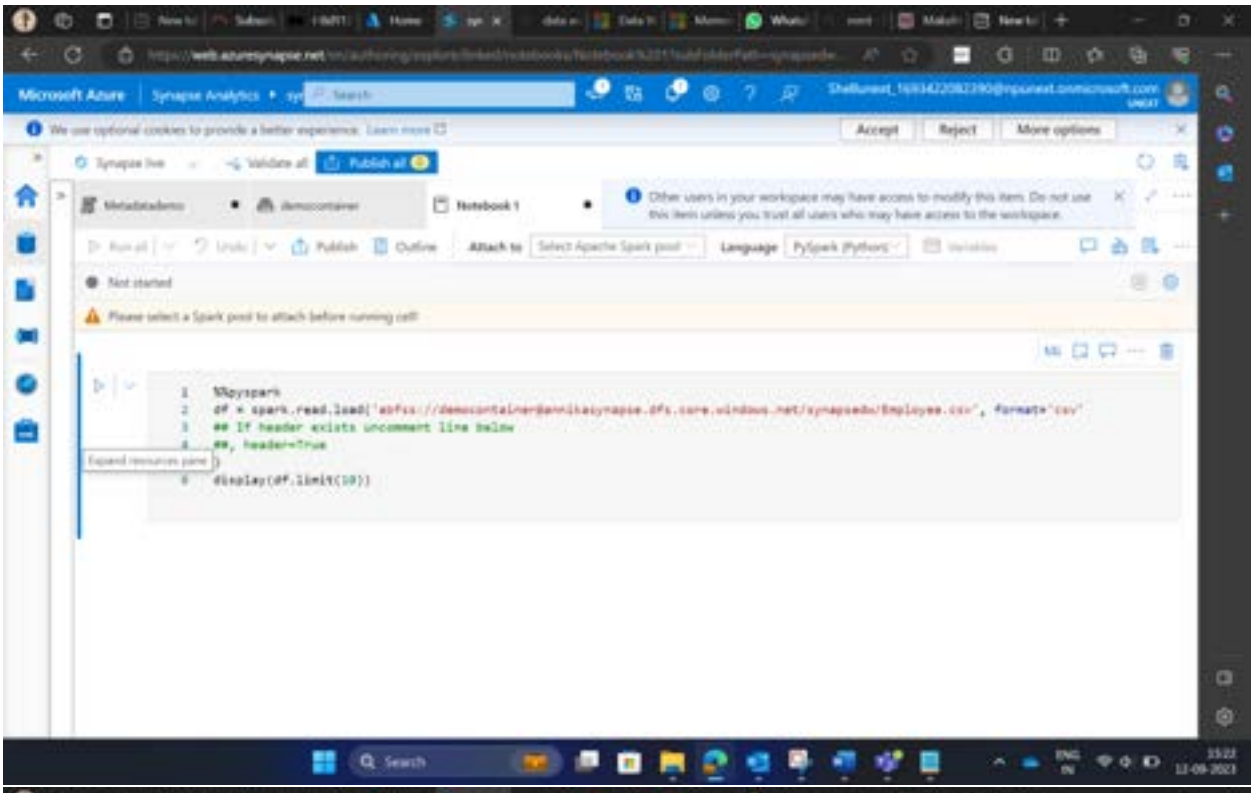

Microsoft Azure
Synapse Analytics
democontainer
Metadatademo
Connect to
Use database
FORMAT_OPTIONS(FIELD_TERMINATOR= ',',STRING_DELIMITER='"',FIRST_ROW =2)
)
--creating external table
CREATE EXTERNAL TABLE empTableInfo
(
Eid int,
Ename varchar(10),
Eage int,
Esalary int
) WITH
(
LOCATION = 'synapsedu/Employee.csv',
DATA_SOURCE = sqlondemanddemo,
FILE_FORMAT = QuotedCsvWithHeaderFormat
)
Results
Messages

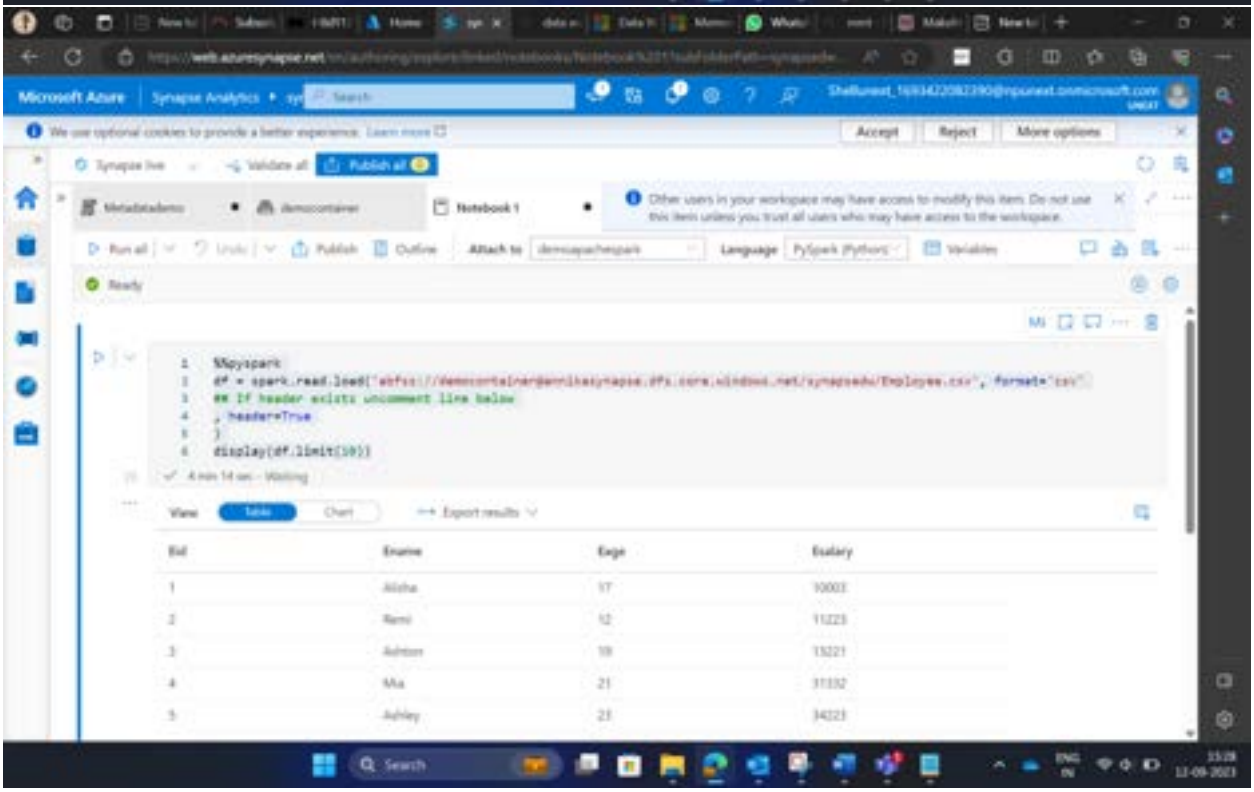

Microsoft Azure
Synapse Analytics
Data
Workspace
Linked
Azure Data Lake Storage Gen2
democontainer
Metadatademo
Connect to
(
LOCATION = 'democontainer/synapsedu/Employee.csv',
DATA_SOURCE = sqlondemanddemo,
FILE_FORMAT = QuotedCsvWithHeaderFormat
)
--testing table
SELECT * FROM empTableInfo
Results
Messages
View
Table
Chart
Export results
Eid
Ename
Eage
Esalary
1
Alisha
17
10000
2
Renu
12
11223



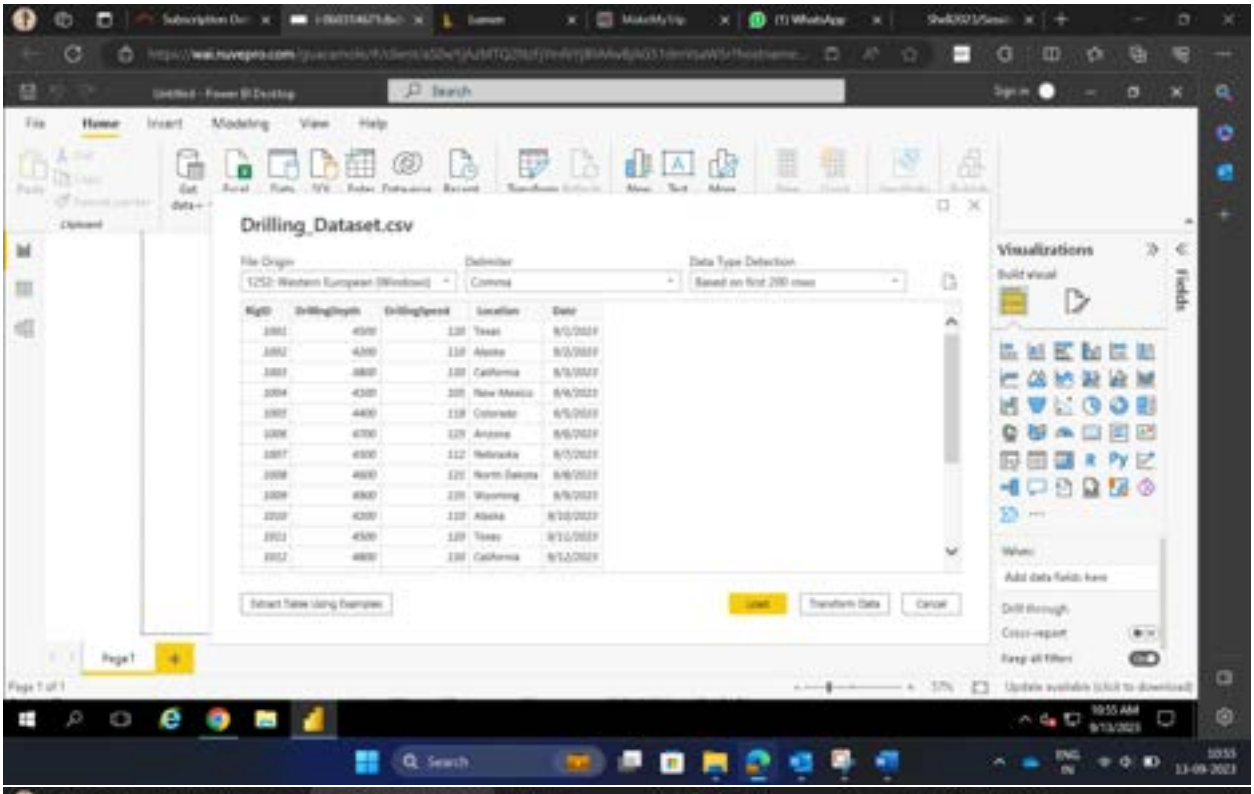
```
--creating master key
create master KEY

--create database scoped credentials
create DATABASE SCOPED CREDENTIAL[sqlondemand]
WITH IDENTITY = 'SHARED ACCESS SIGNATURE',
SECRET = '?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2023-09-
12T16:11:37Z&st=2023-09-
12T08:11:37Z&spr=https&sig=Hu%2FymJ8hjAANSQ9gBbAPWZKQ6HkHfKhTmv4MtgvLzUA%3D'

--create external data source
CREATE EXTERNAL DATA SOURCE sqlondemanddemo WITH(
    LOCATION = 'https://annikasynapse.blob.core.windows.net',
    CREDENTIAL = [sqlondemand]
)

--create external file format
CREATE EXTERNAL FILE FORMAT QuotedCsvwithHeaderFormat
WITH(
    FORMAT_TYPE = DELIMITEDTEXT,
    FORMAT_OPTIONS(FIELD_TERMINATOR= ',',STRING_DELIMITER='"',FIRST_ROW =2)
)
drop EXTERNAL table emptableinfo
--creating external table
```

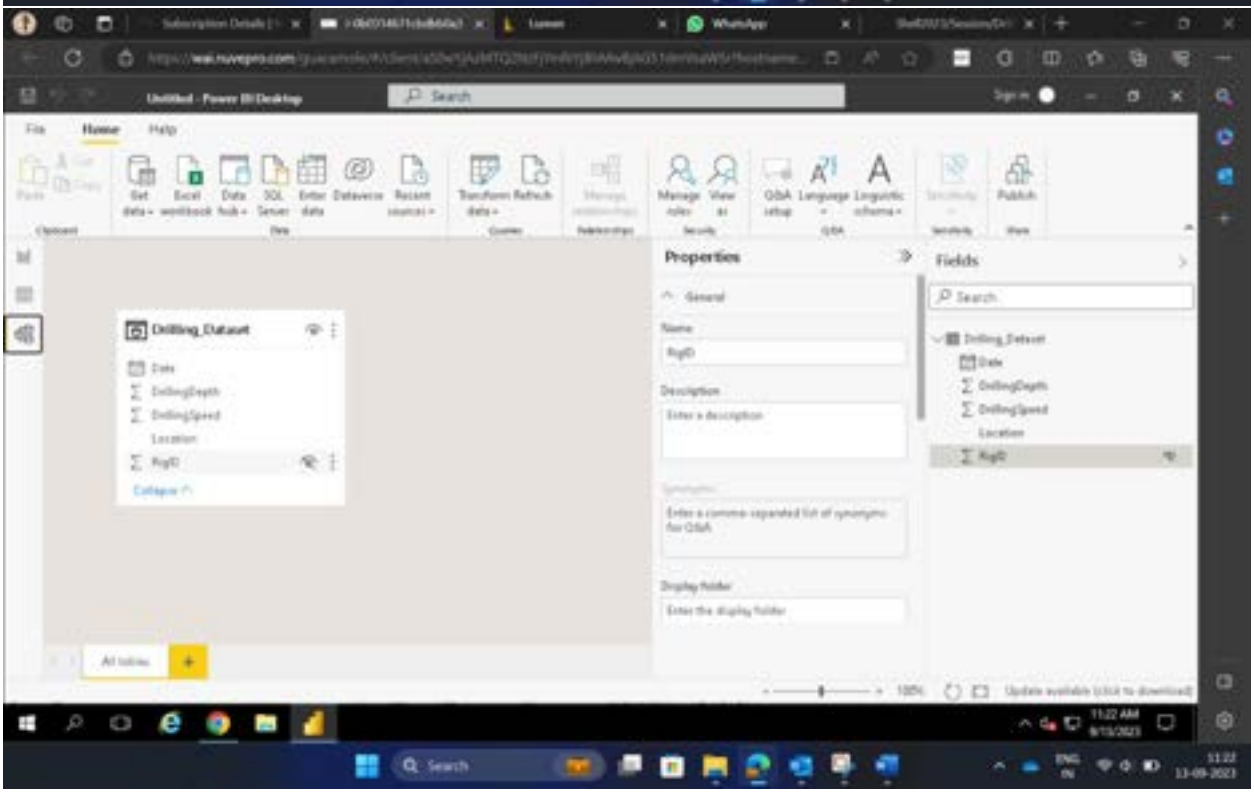
```sql
CREATE EXTERNAL TABLE empTableInfo
(
    Eid int,
    Ename varchar(20),
    Eage int,
    Esalary int
) WITH
(
    LOCATION = 'democontainer/synapsedw/Employee.csv',
    DATA_SOURCE = sqlondemanddemo,
    FILE_FORMAT = QuotedCsvwithHeaderFormat
)

--testing table
SELECT * FROM empTableInfo


drop EXTERNAL TABLE empTableInfo

--create view
create view vwEmployee
AS
SELECT * from empTableInfo

--reading data using view
SELECT * FROM vwEmployee
```

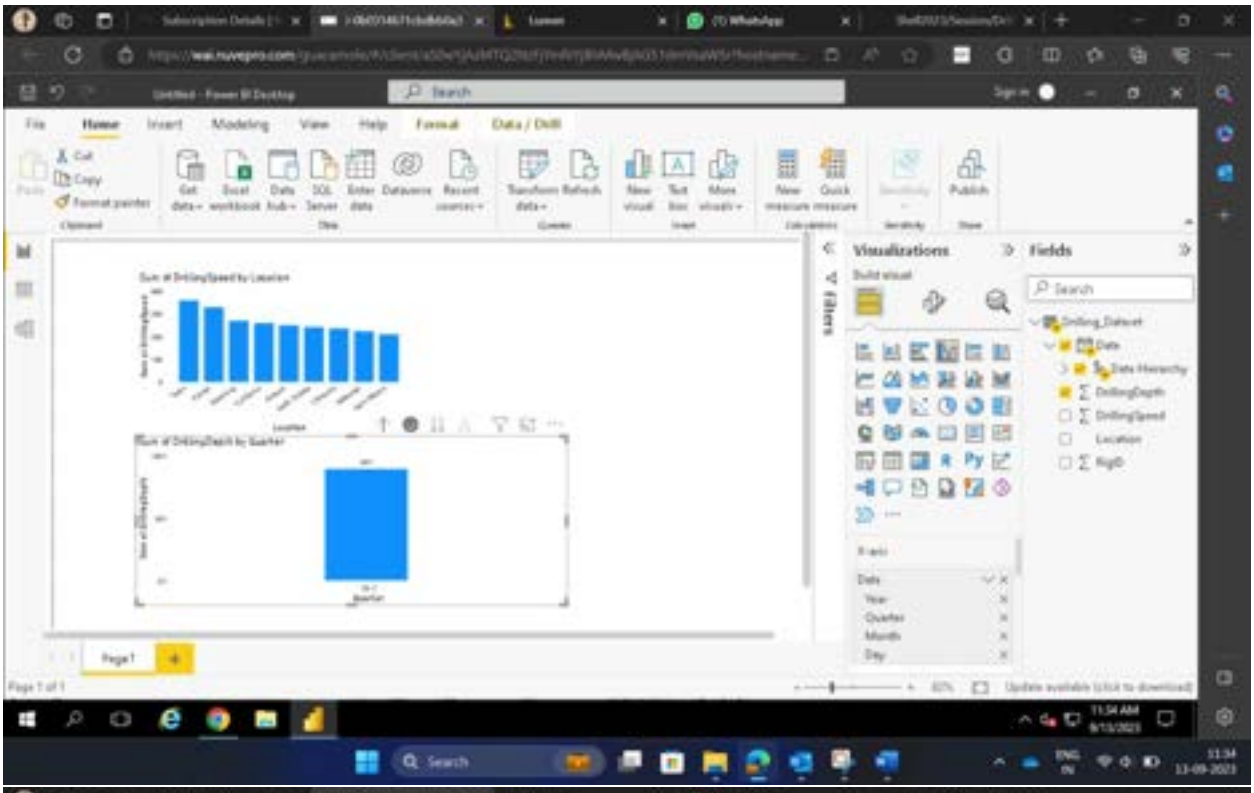

Microsoft Azure
Synapse Analytics
Notebook 1
Not started
Please select a Spark pool to attach before running cell
%%pyspark
df = spark.read.load('abfss://democontainer@anrikasynapse.dfs.core.windows.net/synapsedu/Employee.csv', format='csv'
If header exists uncomment line below
##, header=True
)
display(df.limit(10))

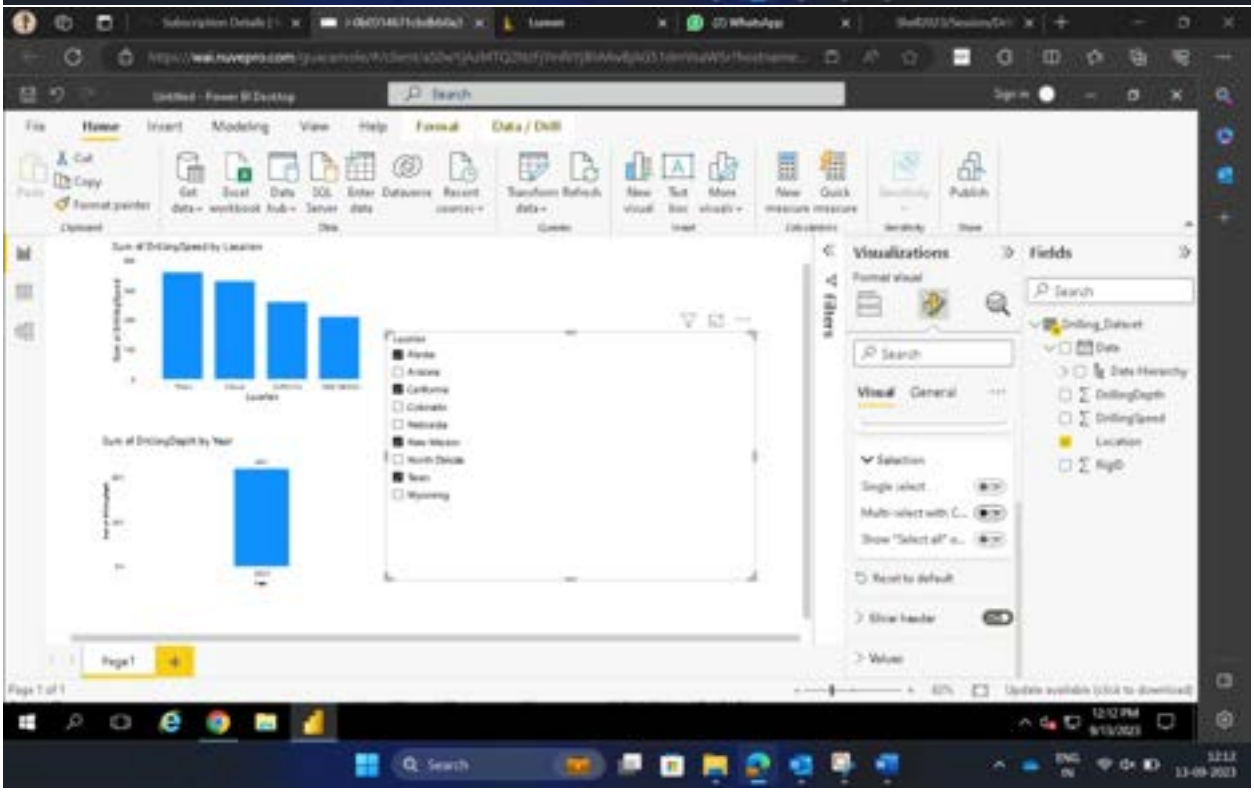

Microsoft Azure
Synapse Analytics
Notebook 1
Ready
%%pyspark
df = spark.read.load('abfss://democontainer@anrikasynapse.dfs.core.windows.net/synapsedu/Employee.csv', format='csv'
If header exists uncomment line below
, header=True
)
display(df.limit(10))
View
Table
Chart
Export results
Eid
Ename
Eage
Esalary
1
Alisha
17
10003
2
Remi
12
11223
3
Ashton
19
13221
4
Mia
21
31332
5
Ashley
23
34223

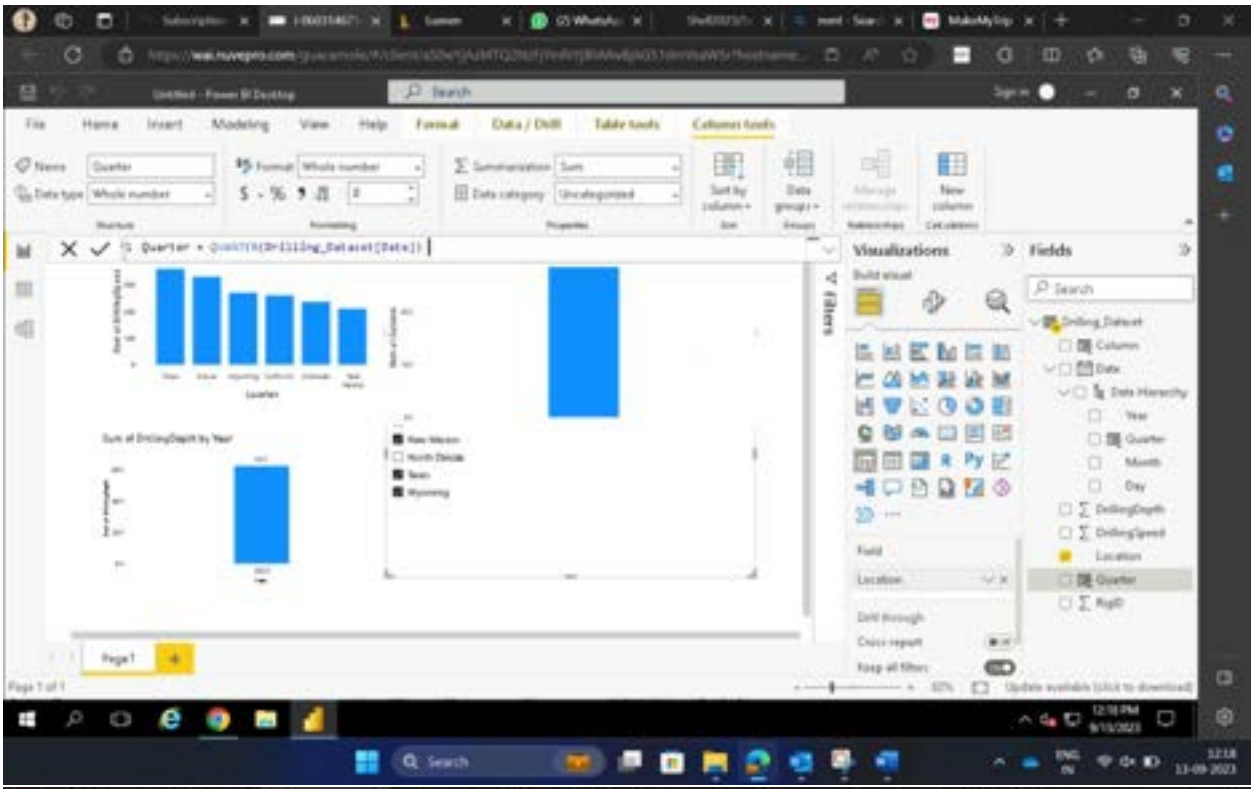
Power bi

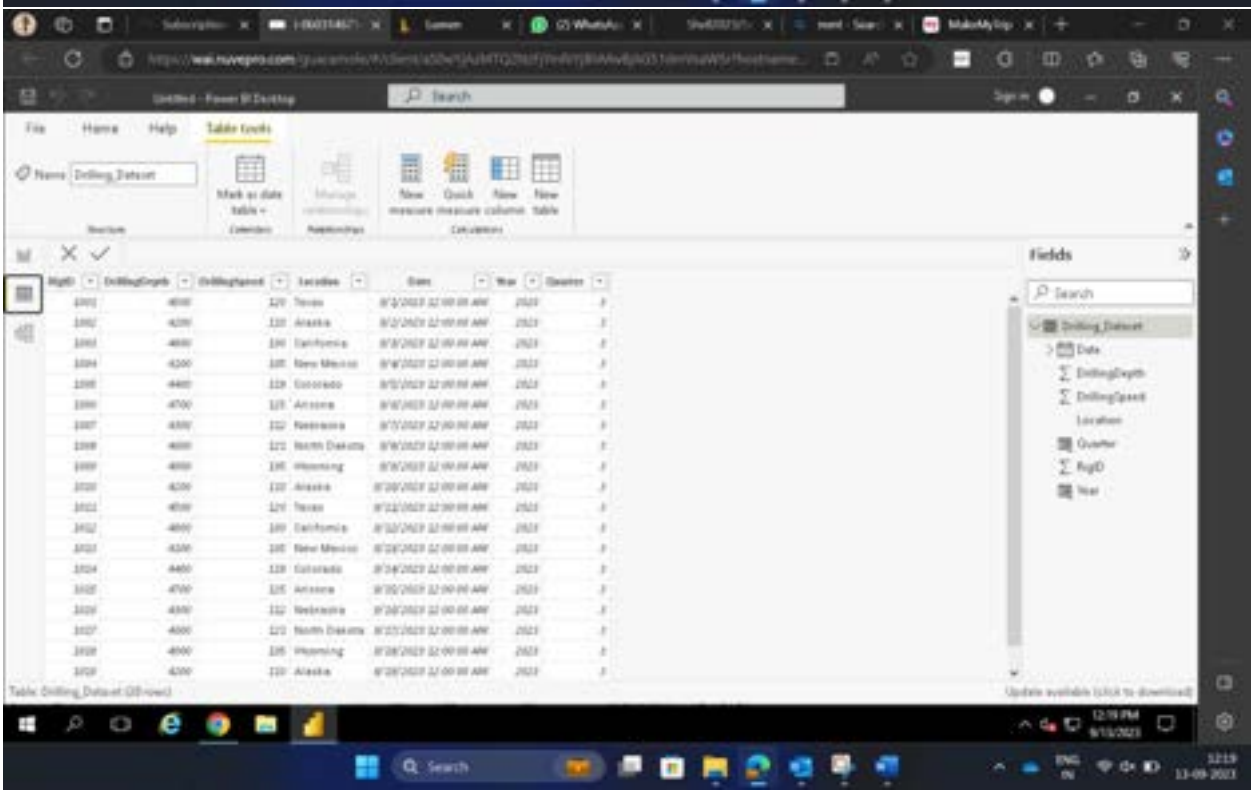

Drilling_Dataset.csv
Visualizations
Fields
Page 1

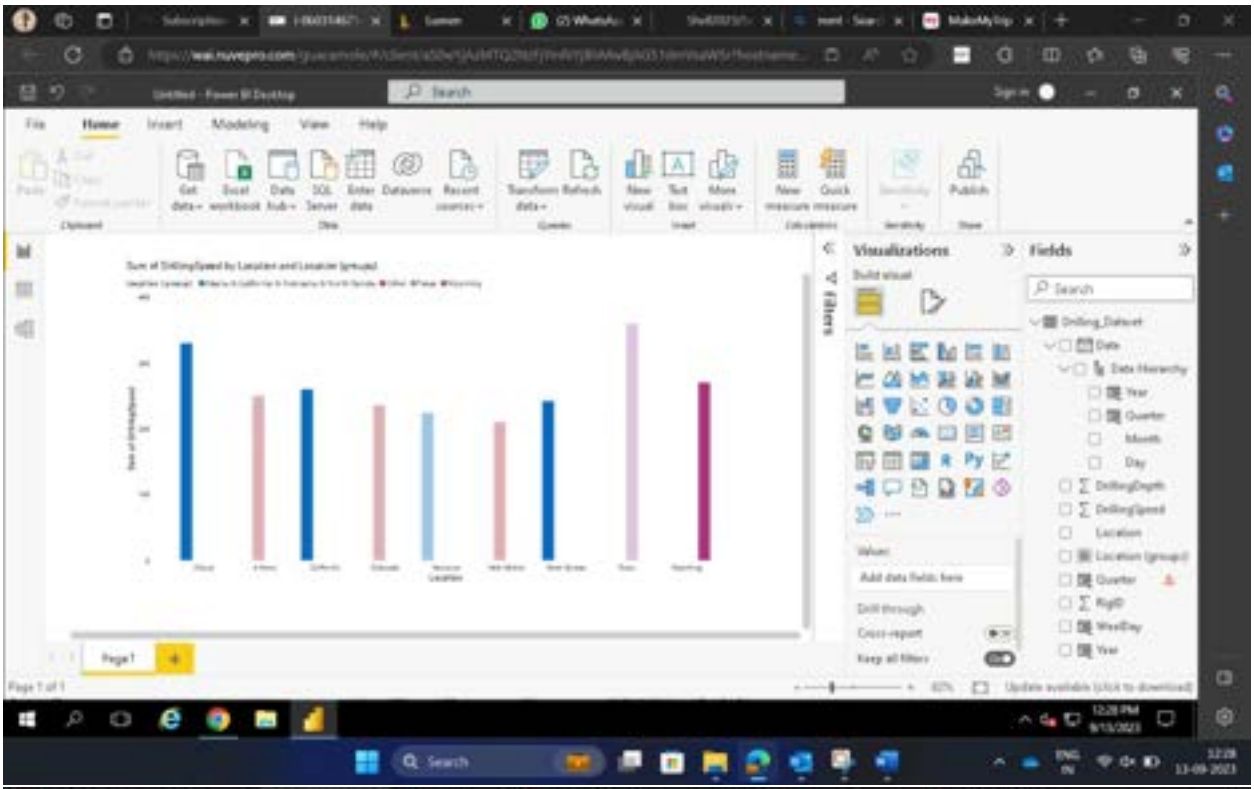

Drilling_Dataset
Properties
Fields
General
Name
Description
Display folder
All tables

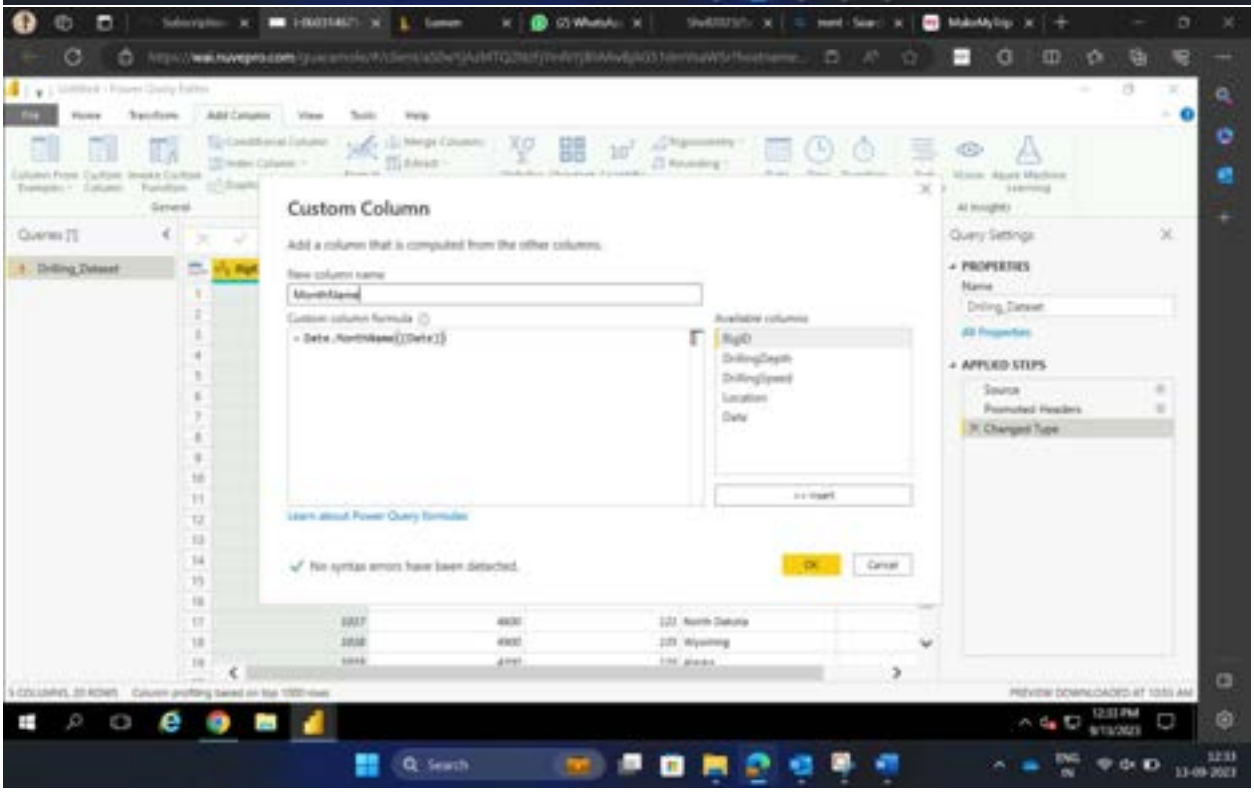


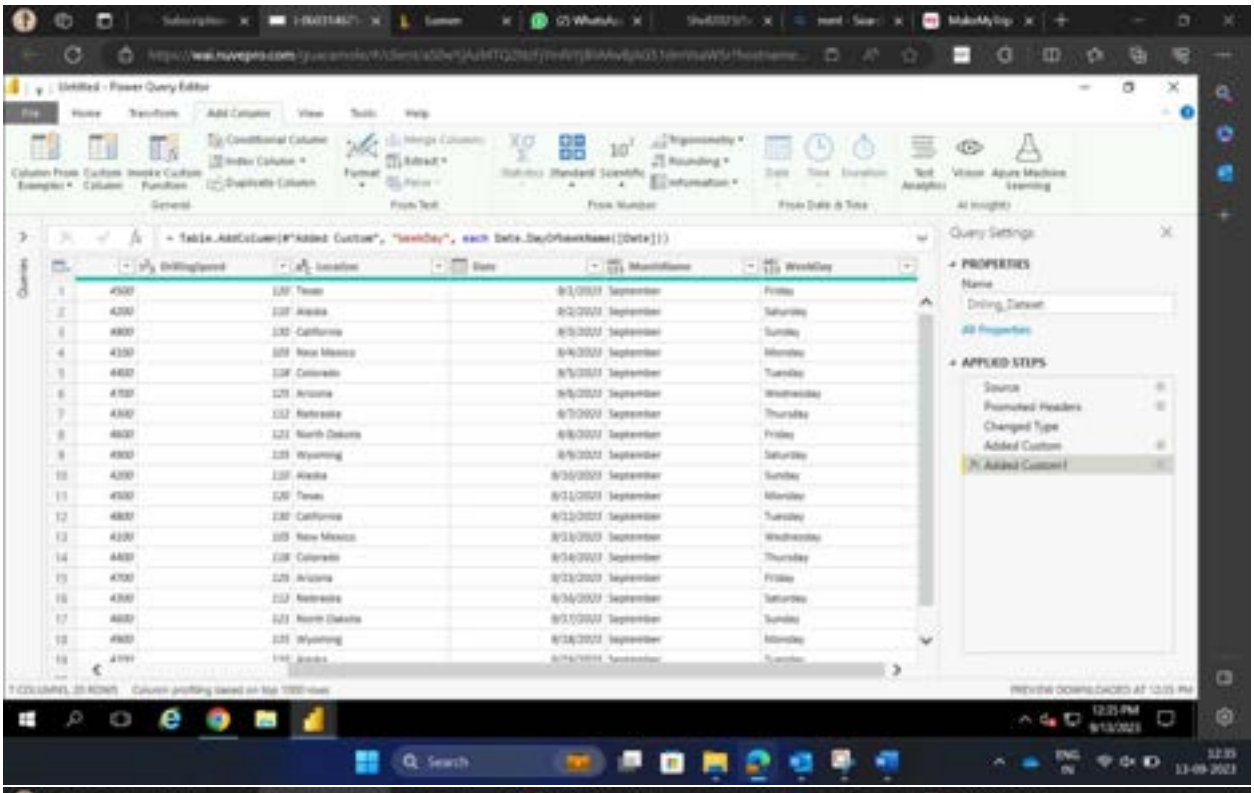


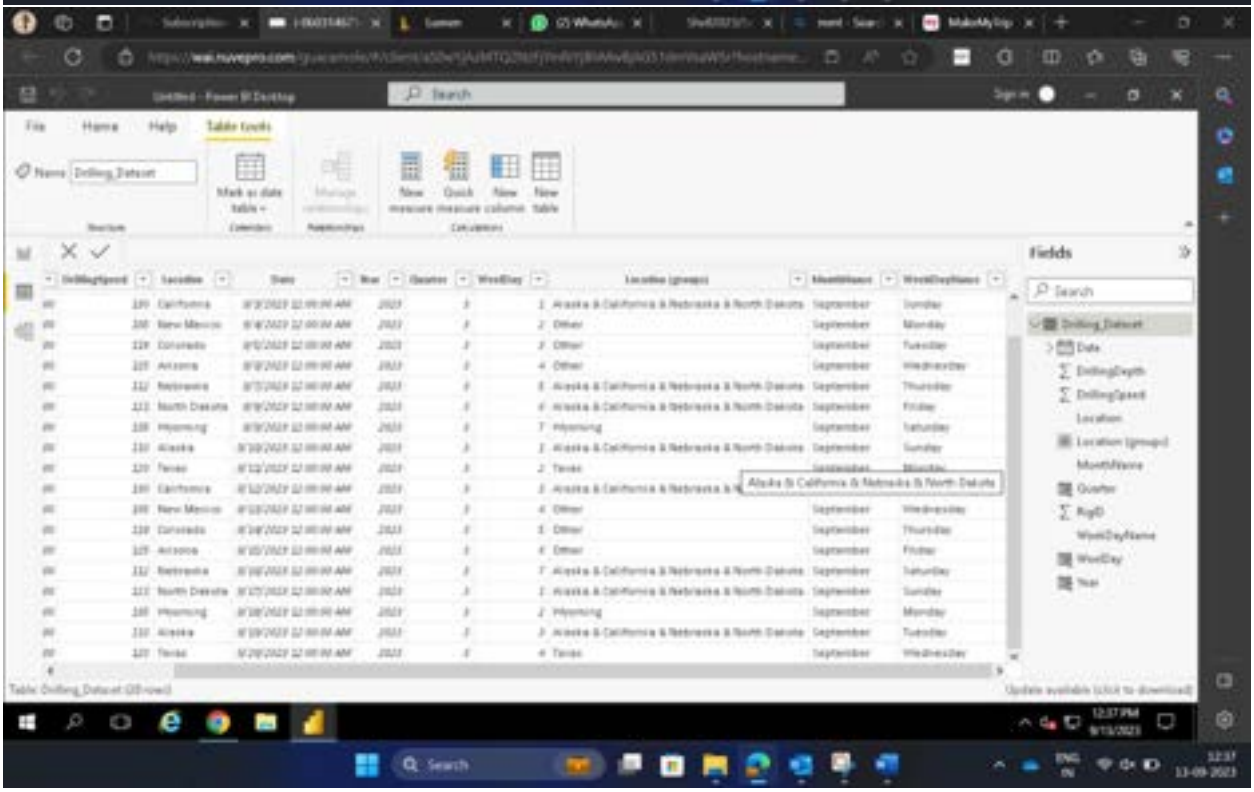


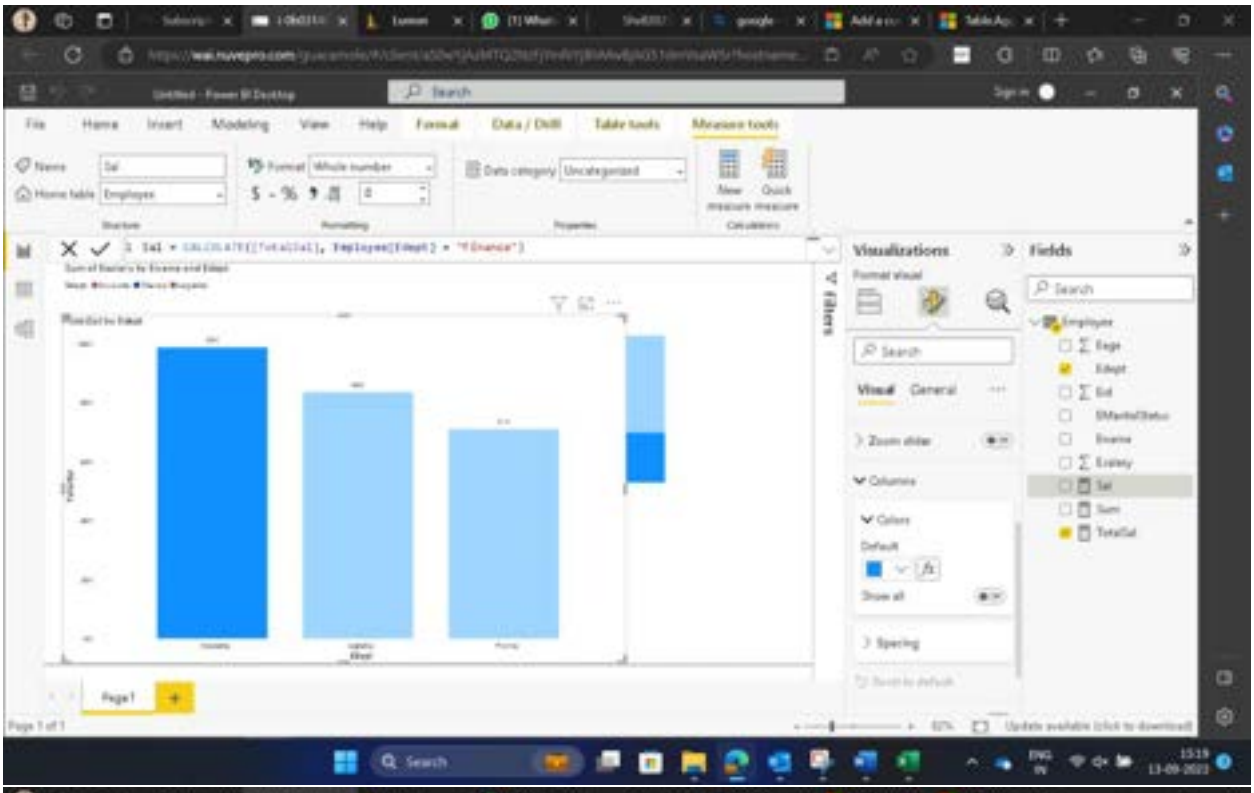


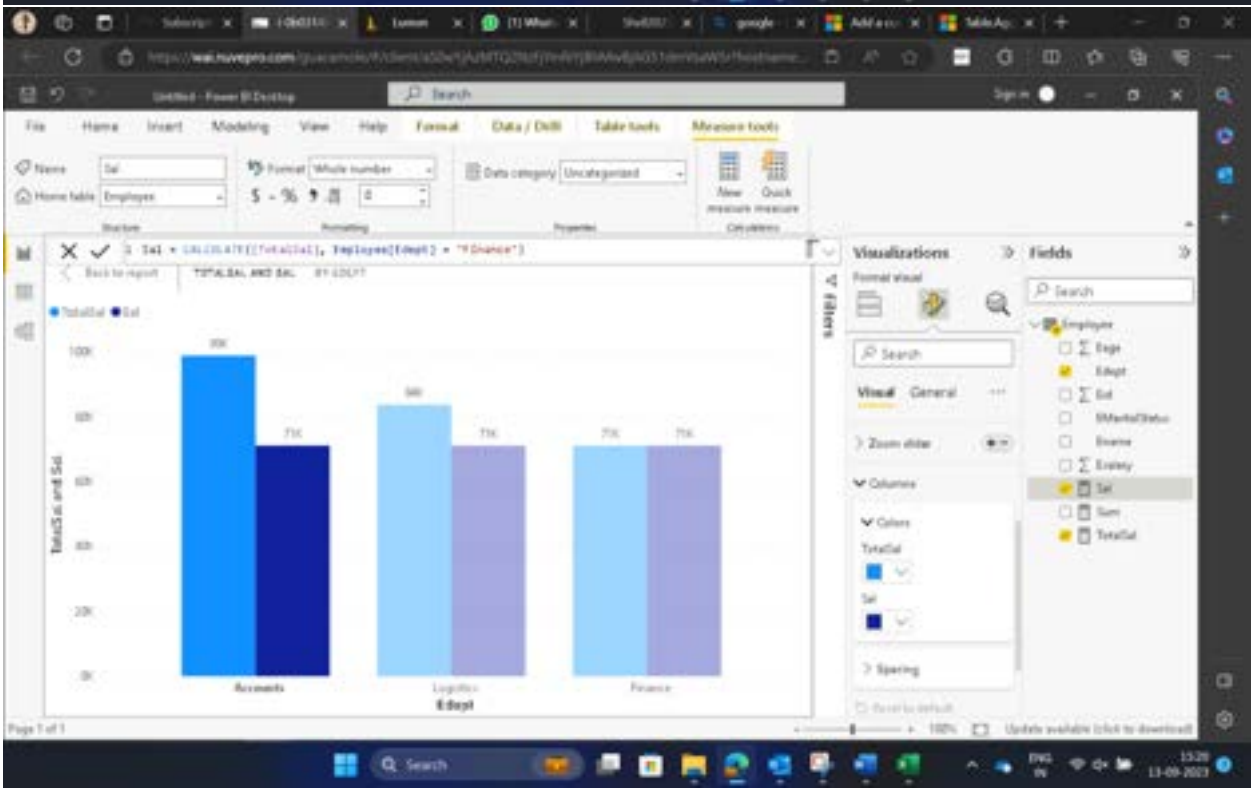

Visualizations
Fields
Page 1

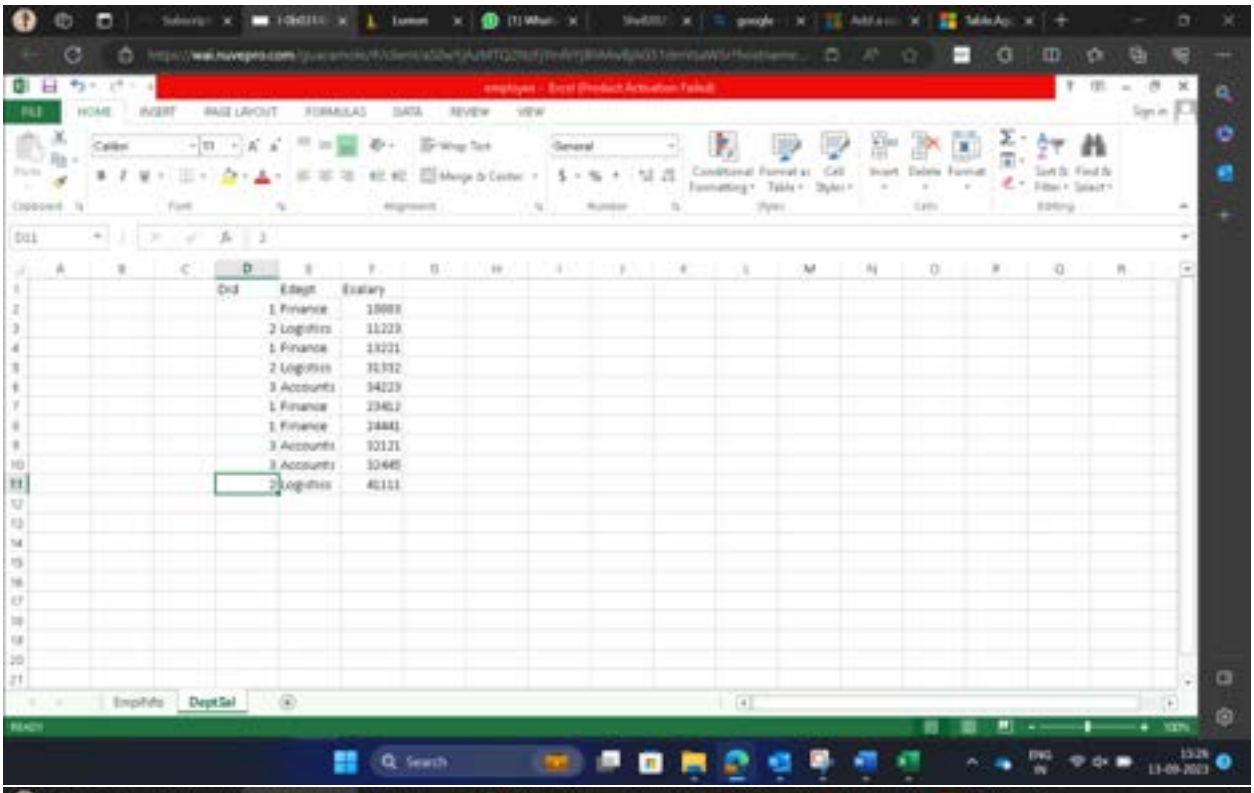

Custom Column
Add a column that is computed from the other columns.
New column name
MonthName
Custom column formula
= Date.MonthName([Date])
Available columns
RigID
DrillingDepth
DrillingSpeed
Location
Date
Learn about Power Query formulas
No syntax errors have been detected.
OK
Cancel
Query Settings
PROPERTIES
Name
Drilling_Dataset
APPLIED STEPS
Source
Promoted Headers
Changed Type

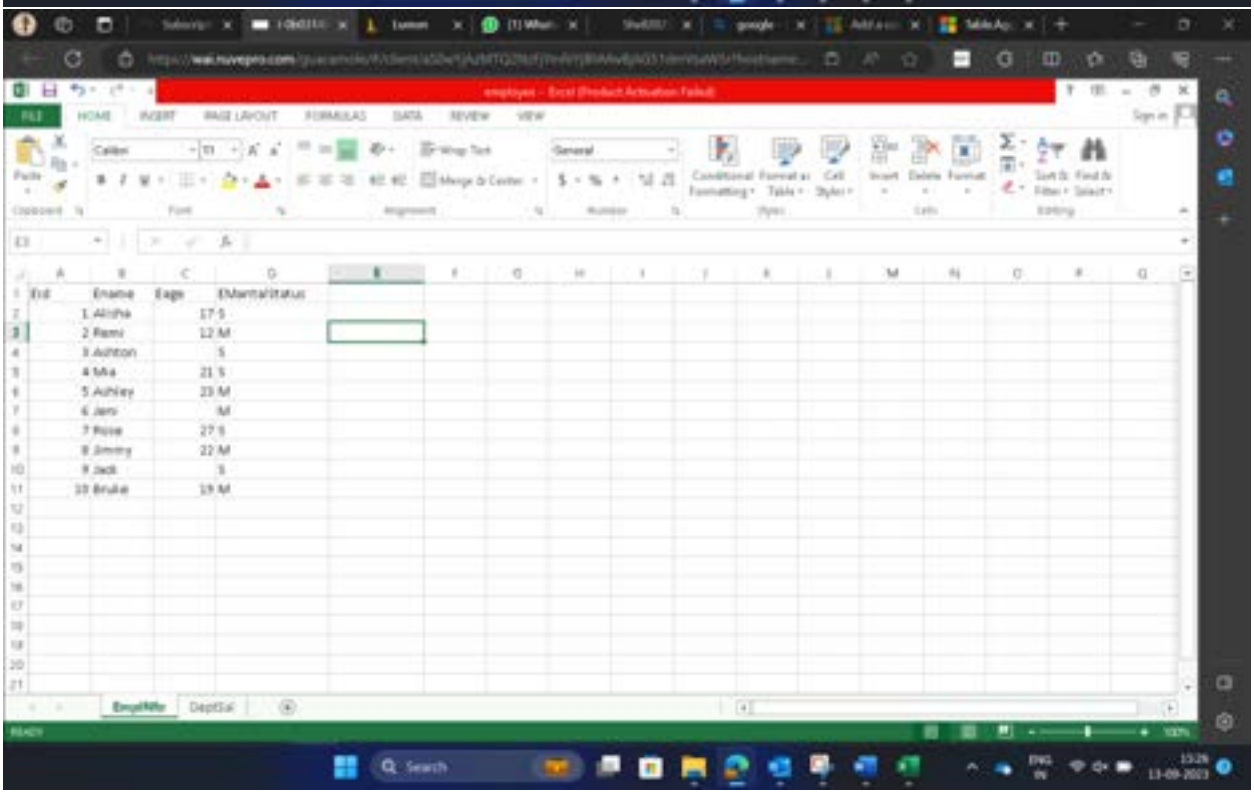

Untitled - Power Query Editor
Query Settings
PROPERTIES
Name
Drilling_Dataset
APPLIED STEPS
Source
Promoted Headers
Changed Type
Added Custom
Added Custom1

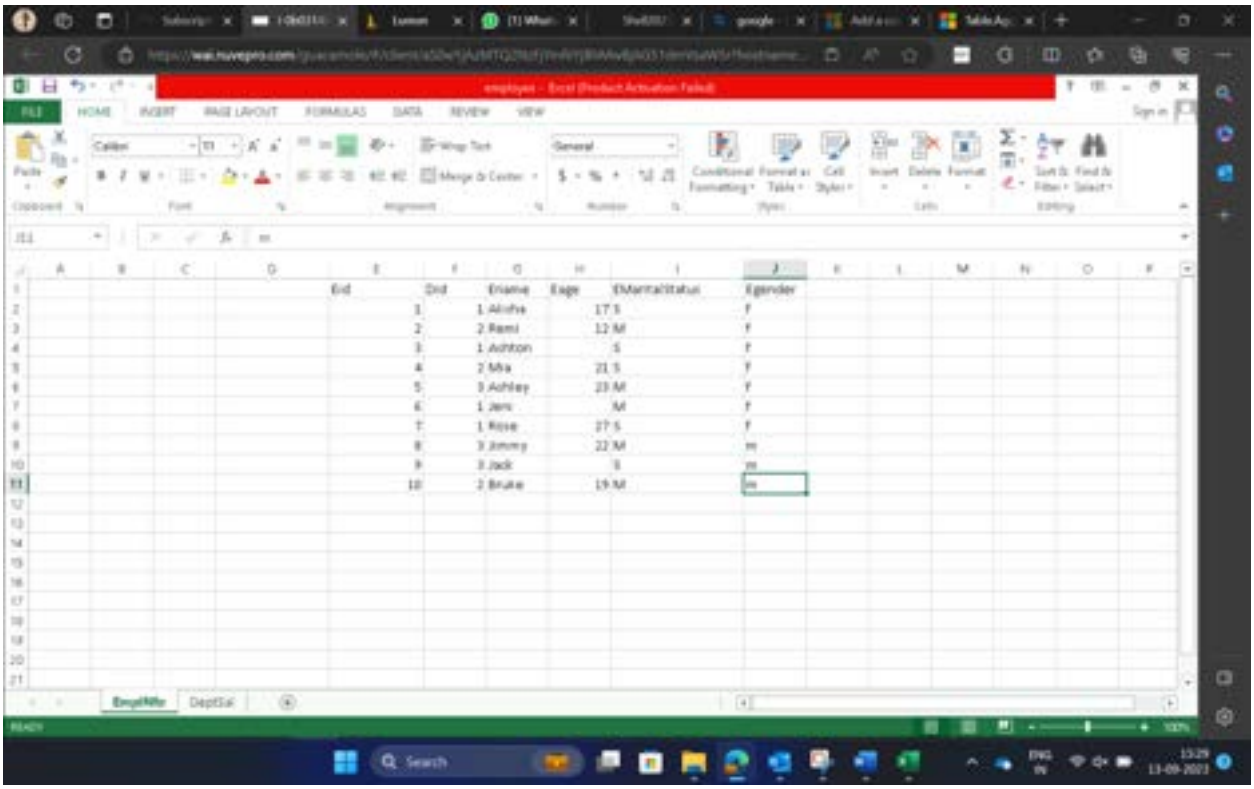

Untitled - Power BI Desktop
File
Home
Help
Table tools
Drilling_Dataset
Fields
Search
Drilling_Dataset
Date
DrillingDepth
DrillingSpeed
Location
Location (groups)
MonthName
Quarter
RigID
WeekDayName
WeekDay
Year
Alaska & California & Nebraska & North Dakota

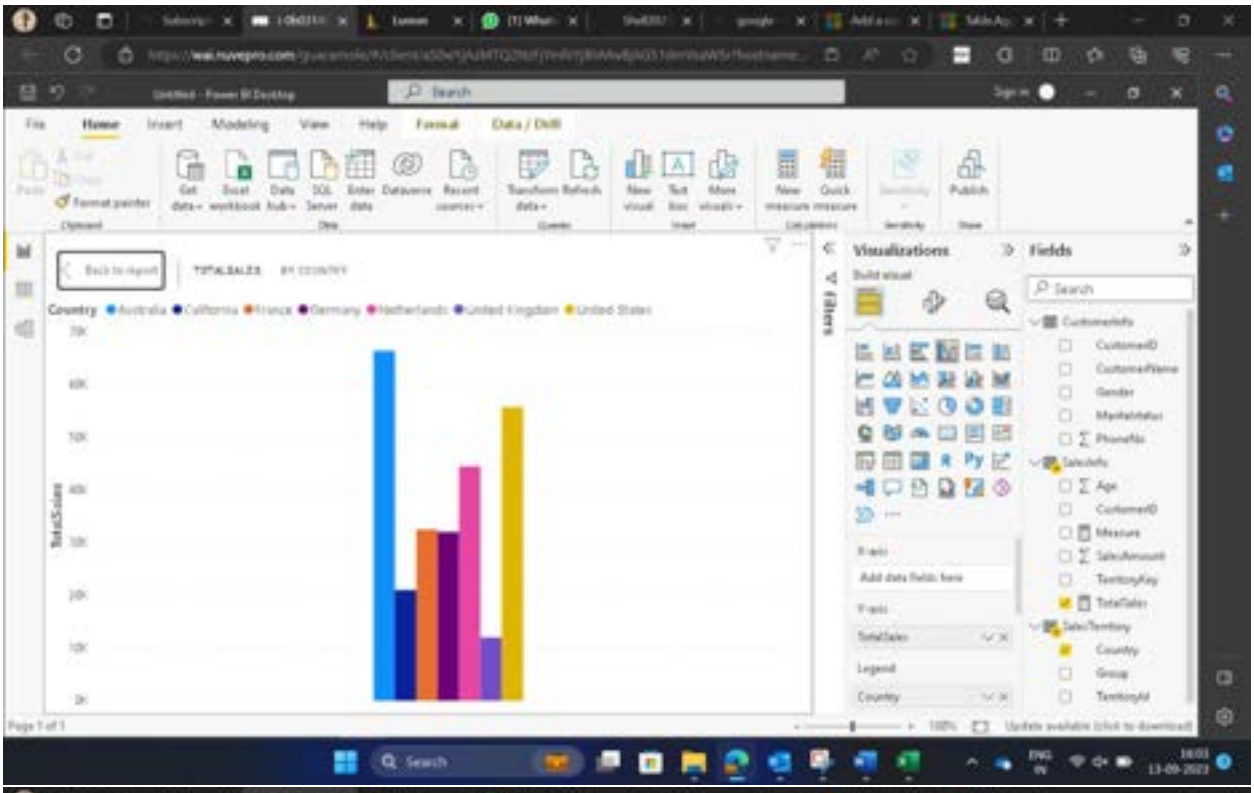


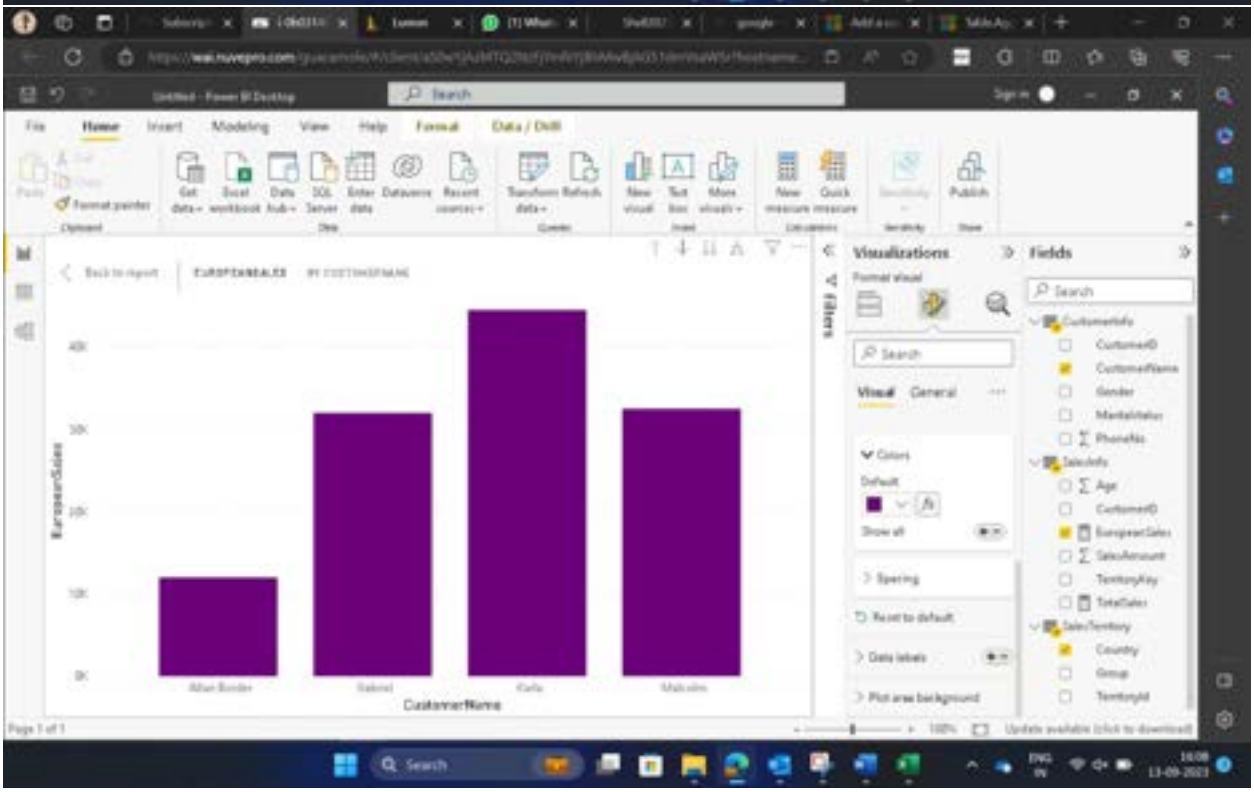


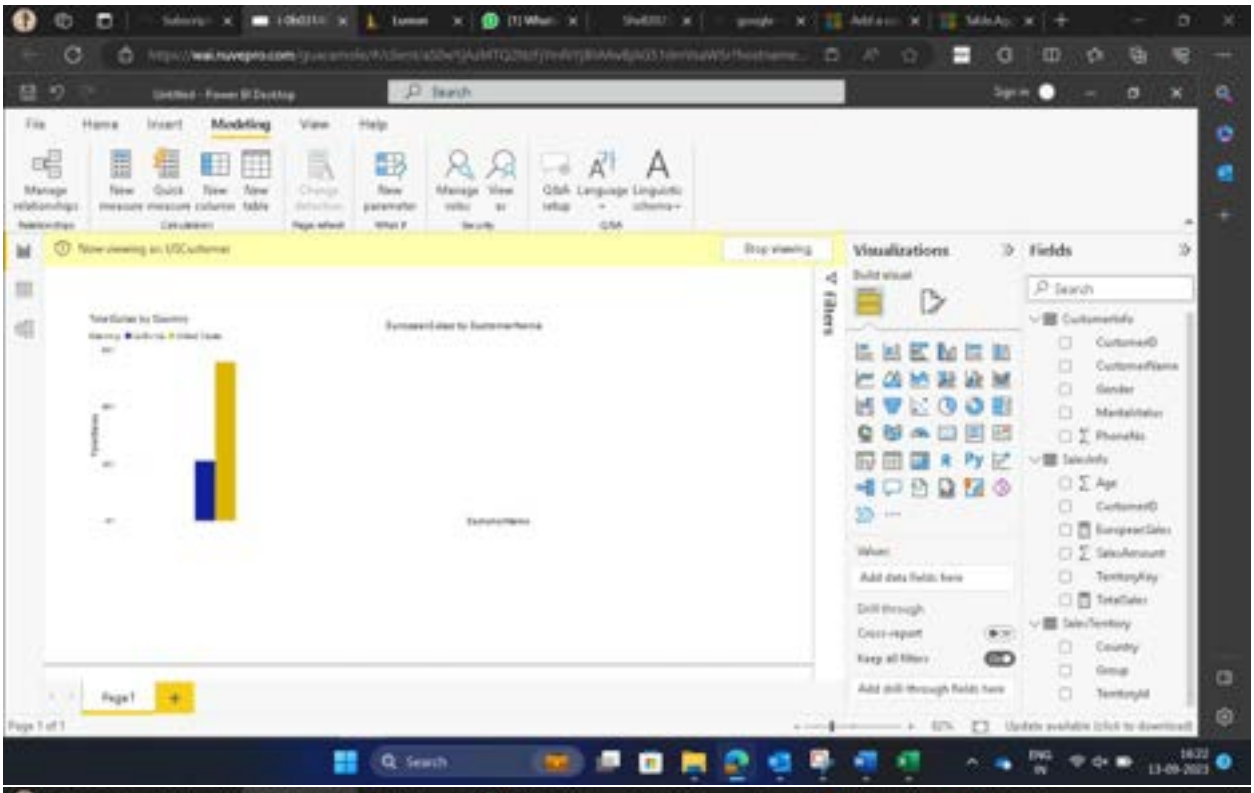

Did Edept Esalary
1 Finance 18883
2 Logistics 11223
1 Finance 13221
2 Logistics 31332
3 Accounts 34223
1 Finance 23452
1 Finance 24441
3 Accounts 32121
3 Accounts 32445
2 Logistics 45111
EmpInfo
DeptSal

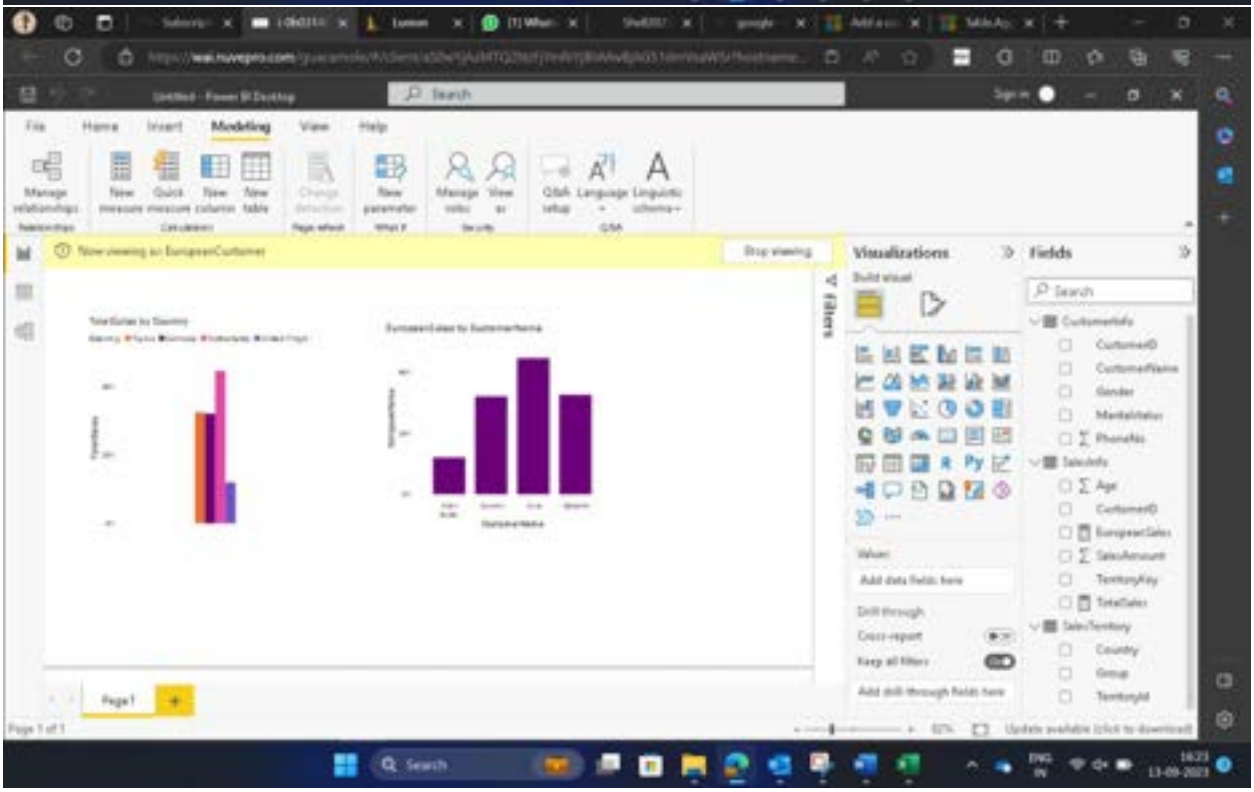

Eid Ename Eage EMaritalStatus
1 Alisha 17 S
2 Remi 12 M
3 Ashton S
4 Mia 21 S
5 Ashley 23 M
6 Jeni M
7 Rose 27 S
8 Jimmy 22 M
9 Jack S
10 Bruke 19 M
EmpInfo
DeptSal

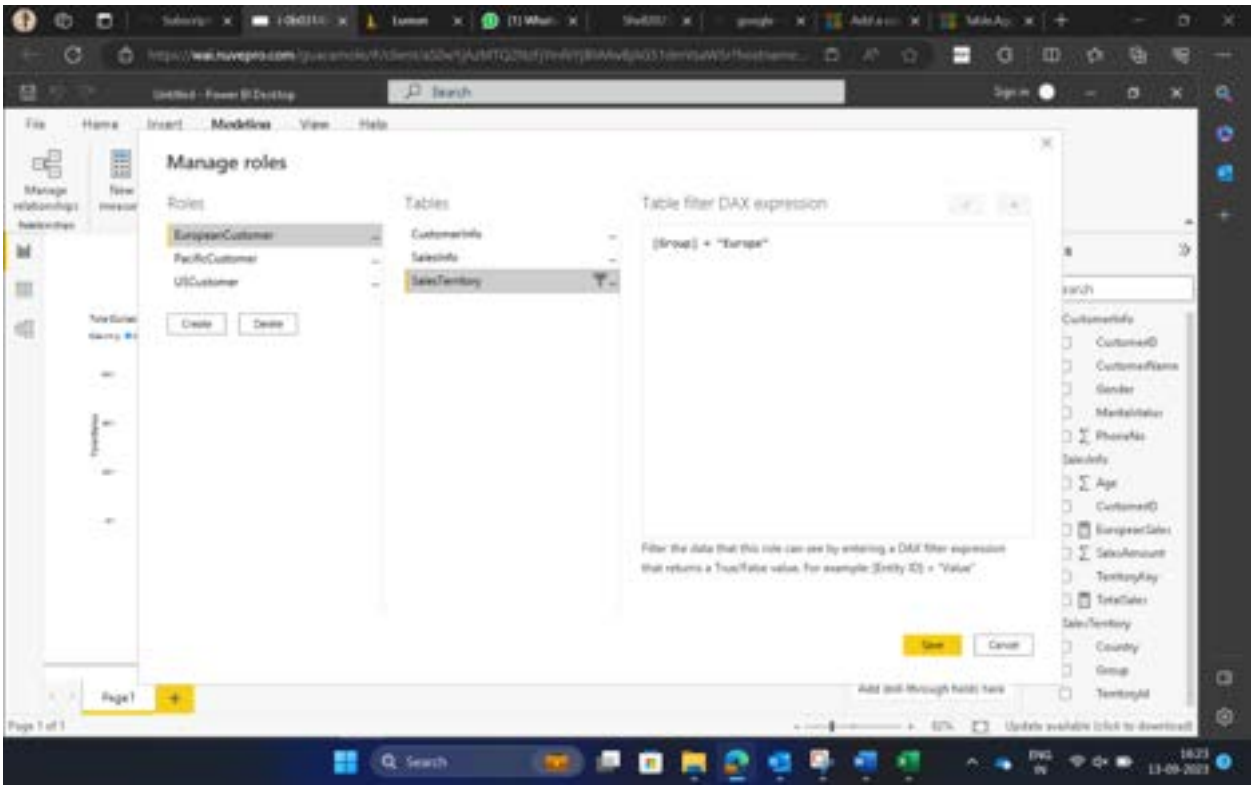

Eid Did Ename Eage EMaritalStatus Egender
1 1 Alisha 17 S f
2 2 Rami 12 M f
3 1 Ashton S f
4 2 Mia 21 S f
5 3 Ashley 23 M f
6 1 Jani M f
7 1 Rose 27 S f
8 3 Jimmy 22 M m
9 3 Jack S m
10 2 Bruke 19 M m
EmpWhr
DeptSal


Visualizations
Fields
Country
TotalSales

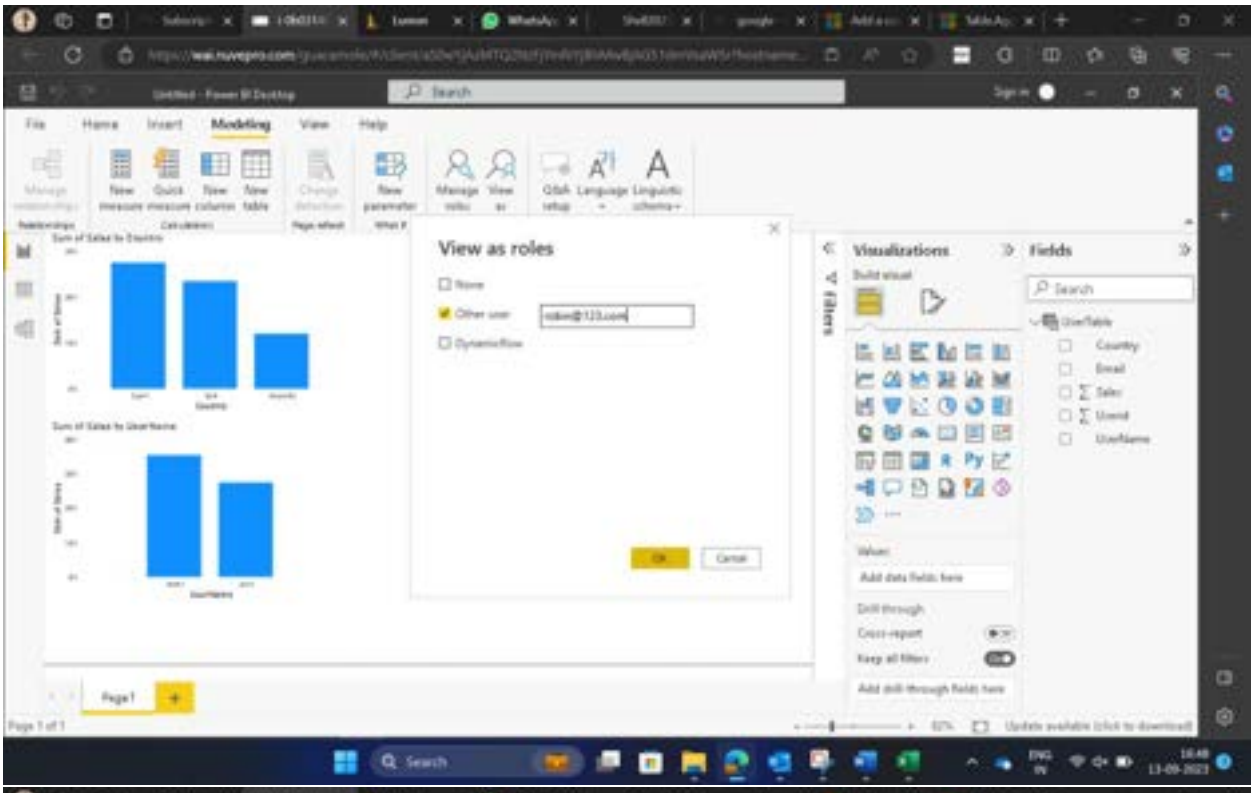

Visualizations
Fields
EuropeanSales
CustomerName

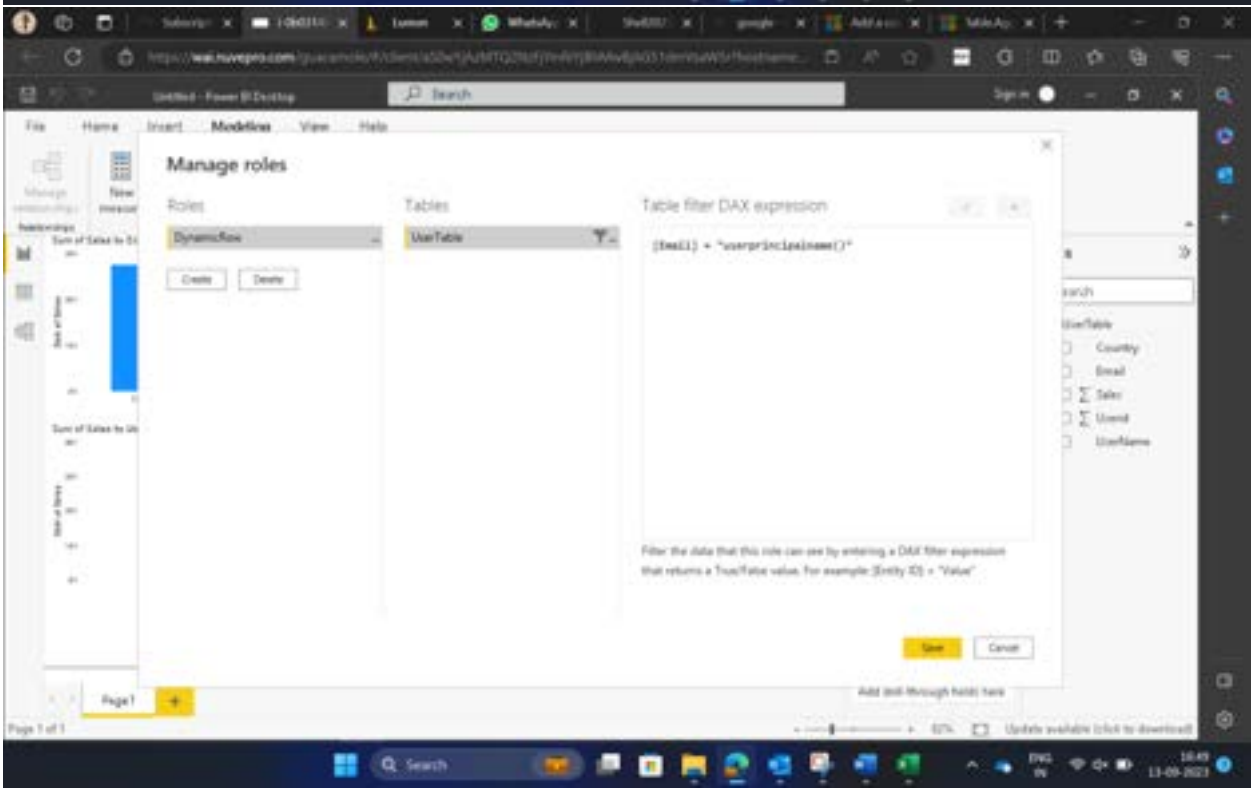







```
UserTable = DATATABLE(

   "Userid", INTEGER, "UserName", STRING , "Country", STRING, "Email", STRING, "Sales", INTEGER,
   {
      {1, "Robin", "Australia", "robin@123.com", 12000},
      {1, "Robin", "USA", "robin@123.com",23454},
      {2, "John", "Spain", "john@abc.com", 12500},
      {2, "John", "Spain", "john@abc.com", 15000}
   }


)
```


View as roles
None
Other user
DynamicRow
OK
Cancel
Visualizations
Fields
Page 1


Manage roles
Roles
Tables
Table filter DAX expression
DynamicRow
UseTable
[Email] = "userprincipalname()"
Create
Delete
Save
Cancel
Page 1

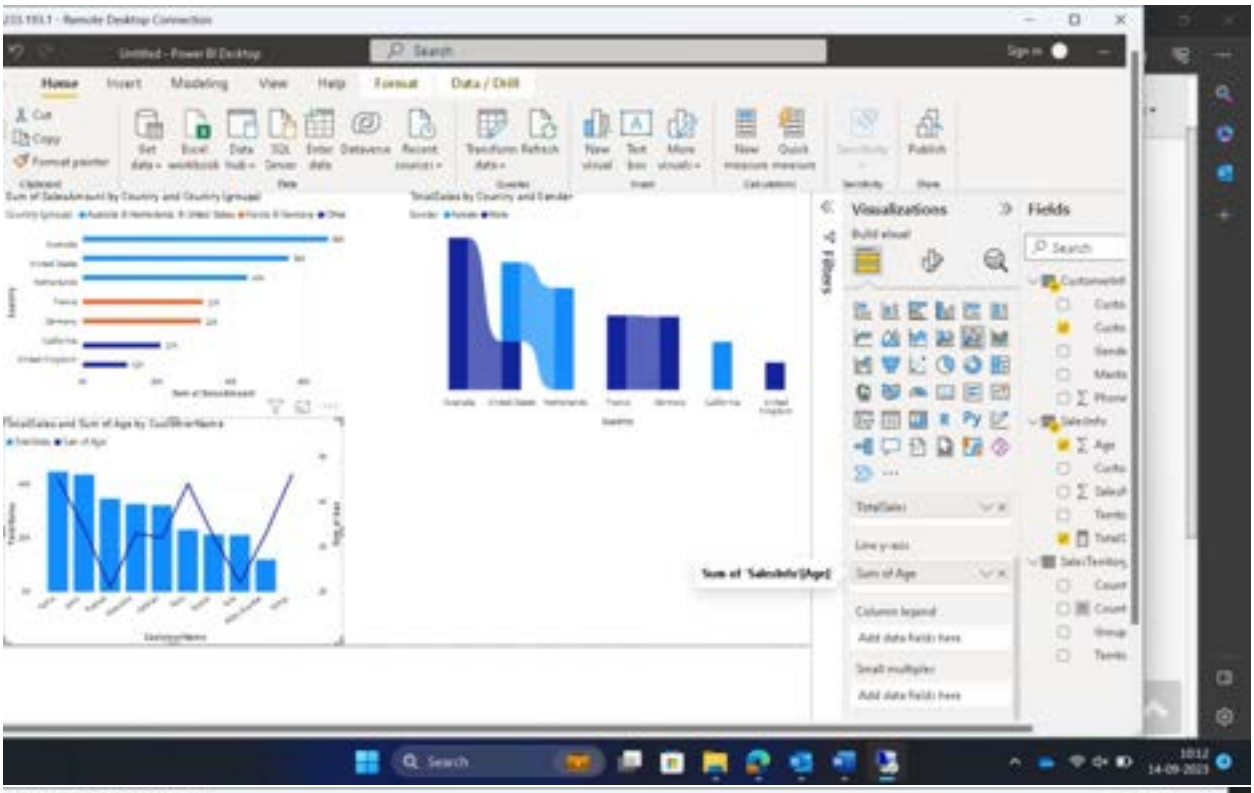
Day 12 14.09.23

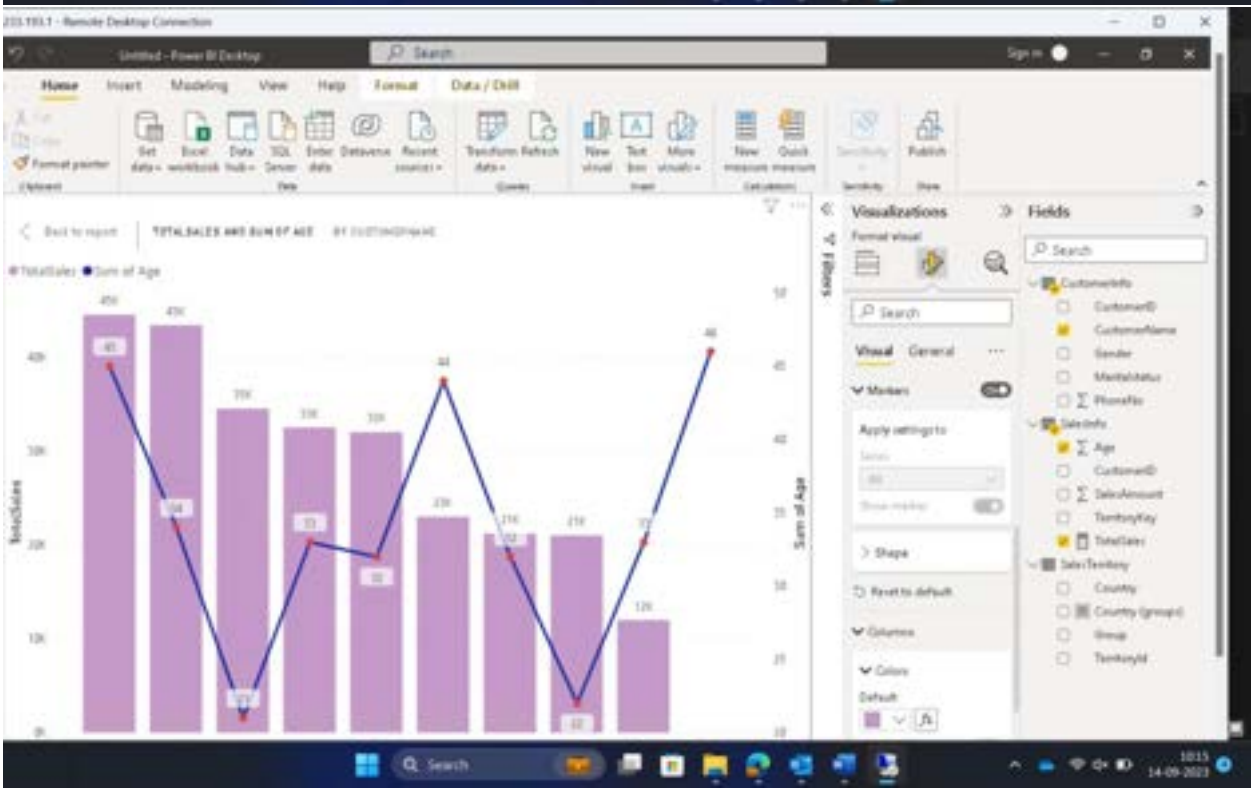
1.bar chart

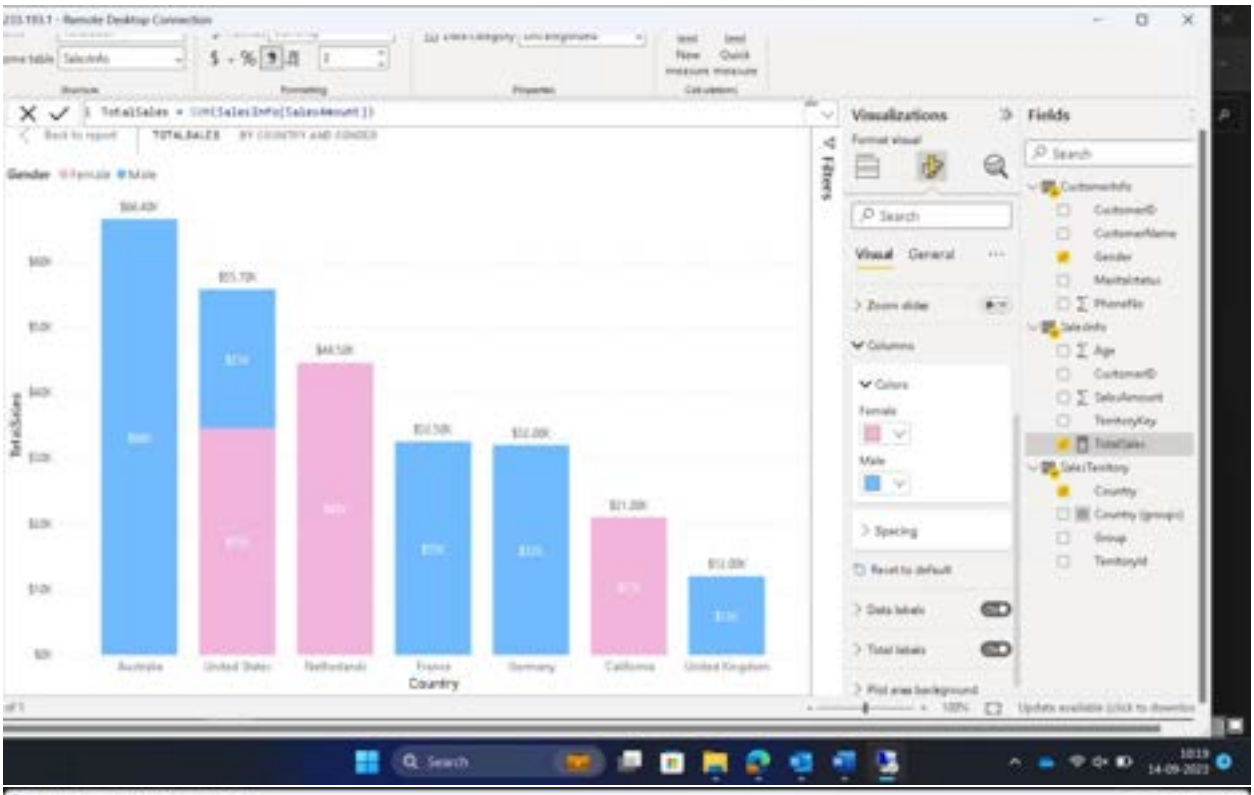
2. grouping data

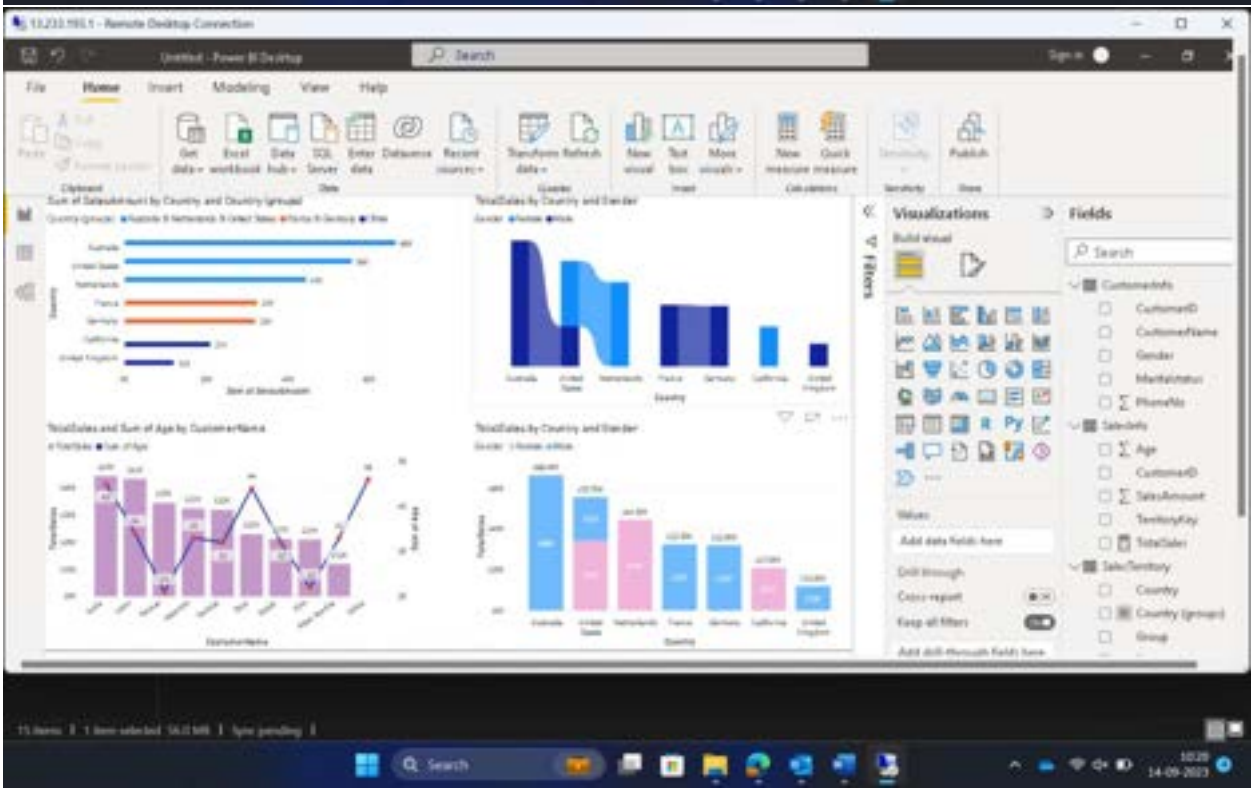
3. line and clustered chart

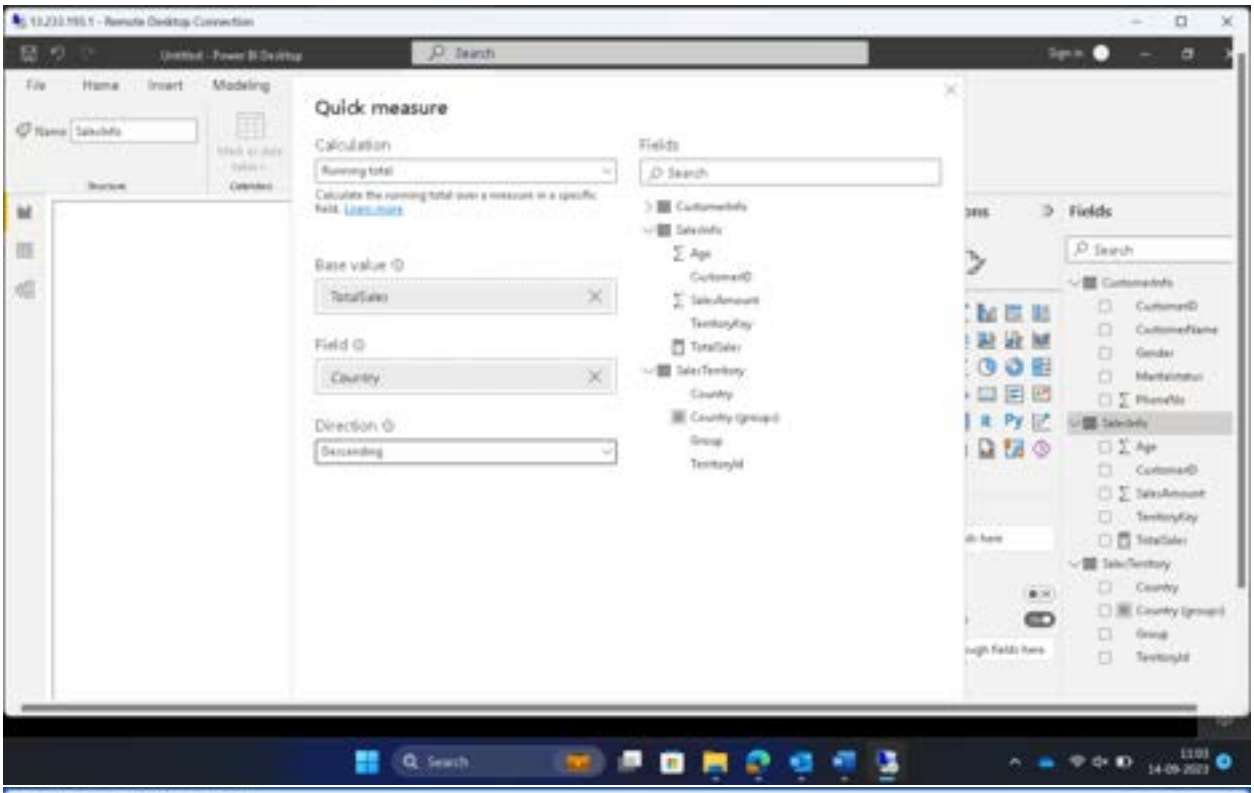
4. ribbon chart

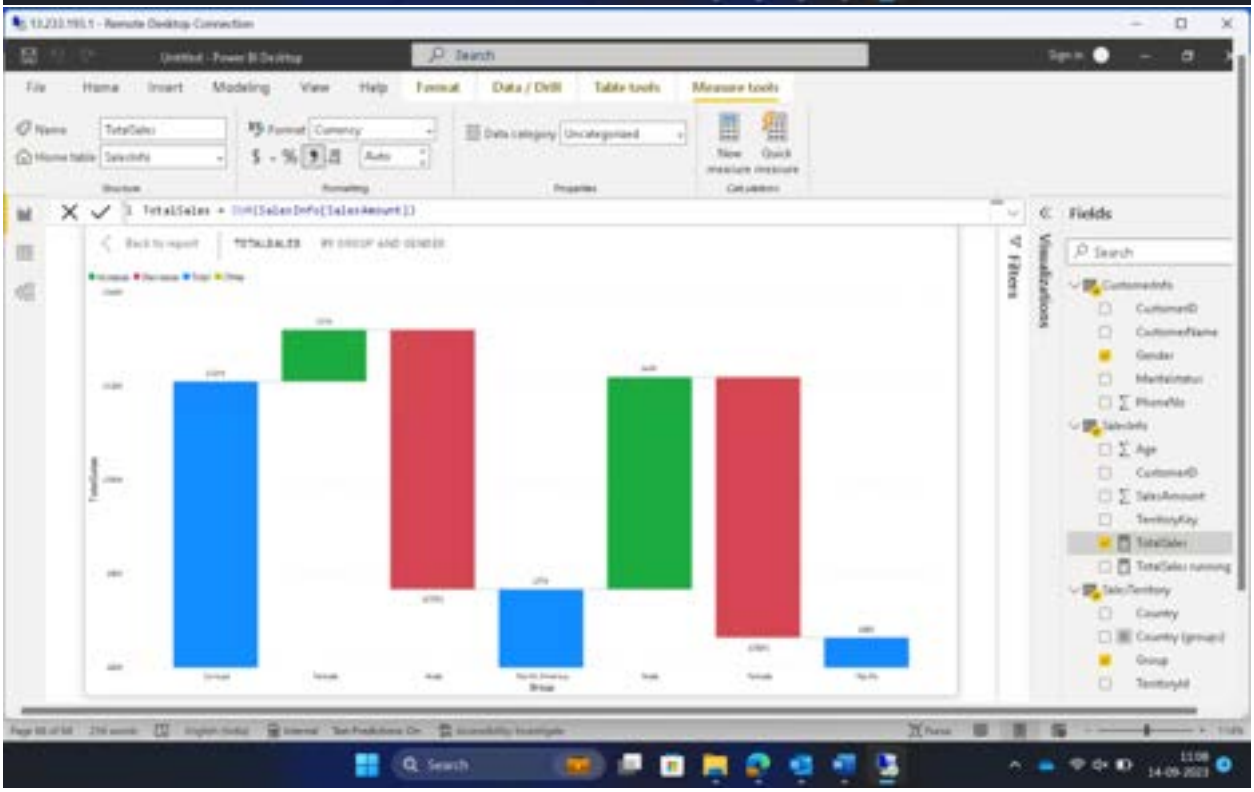

Untitled - Power BI Desktop
Visualizations
Fields
Sum of SalesInfo (Age)

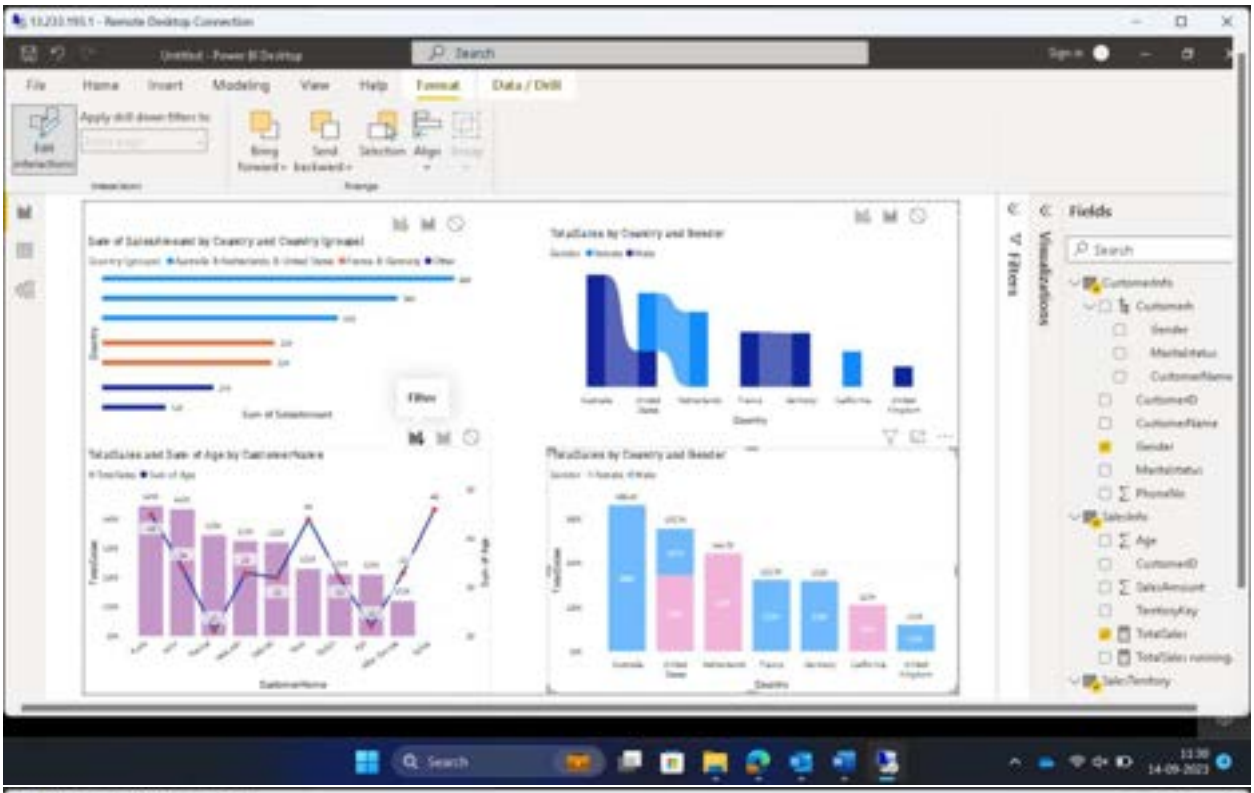

Untitled - Power BI Desktop
Visualizations
Fields
TOTALSALES AND SUM OF AGE

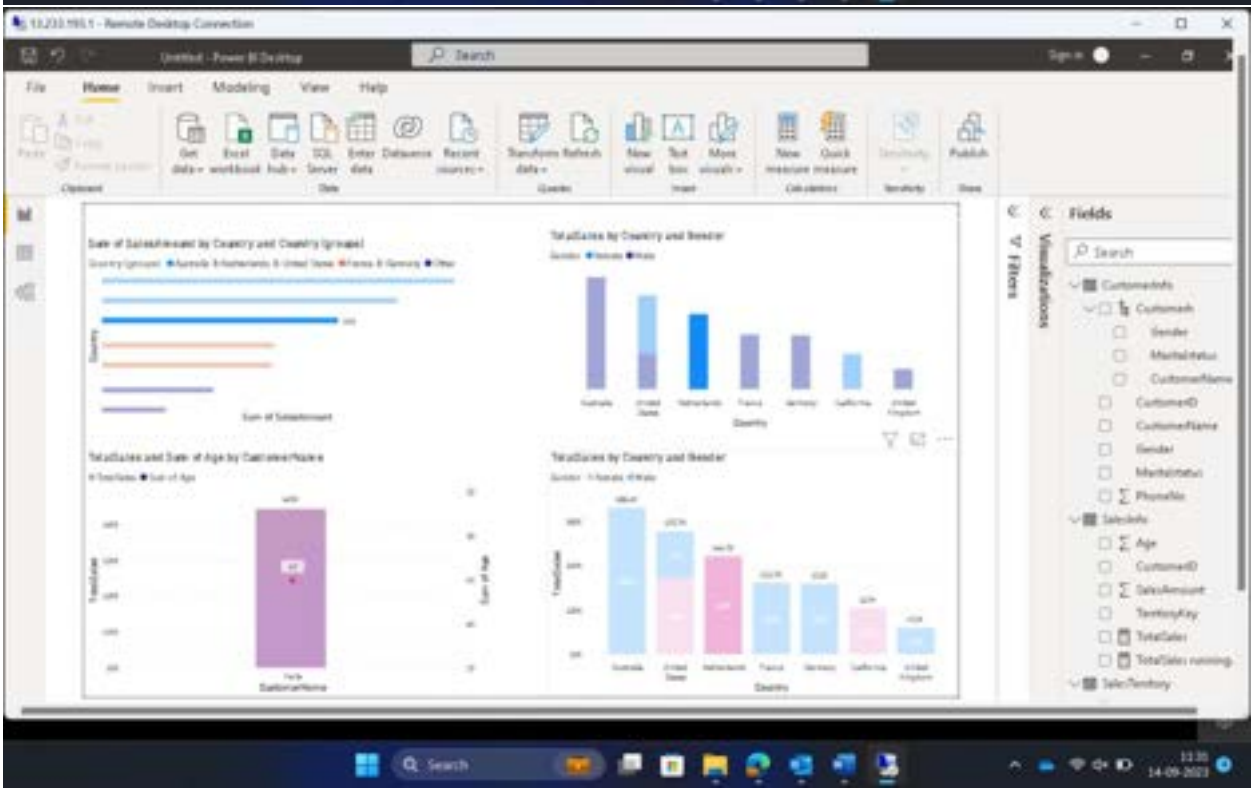

Visualizations
Fields
Gender
Country
Australia
United States
Netherlands
France
Germany
California
United Kingdom

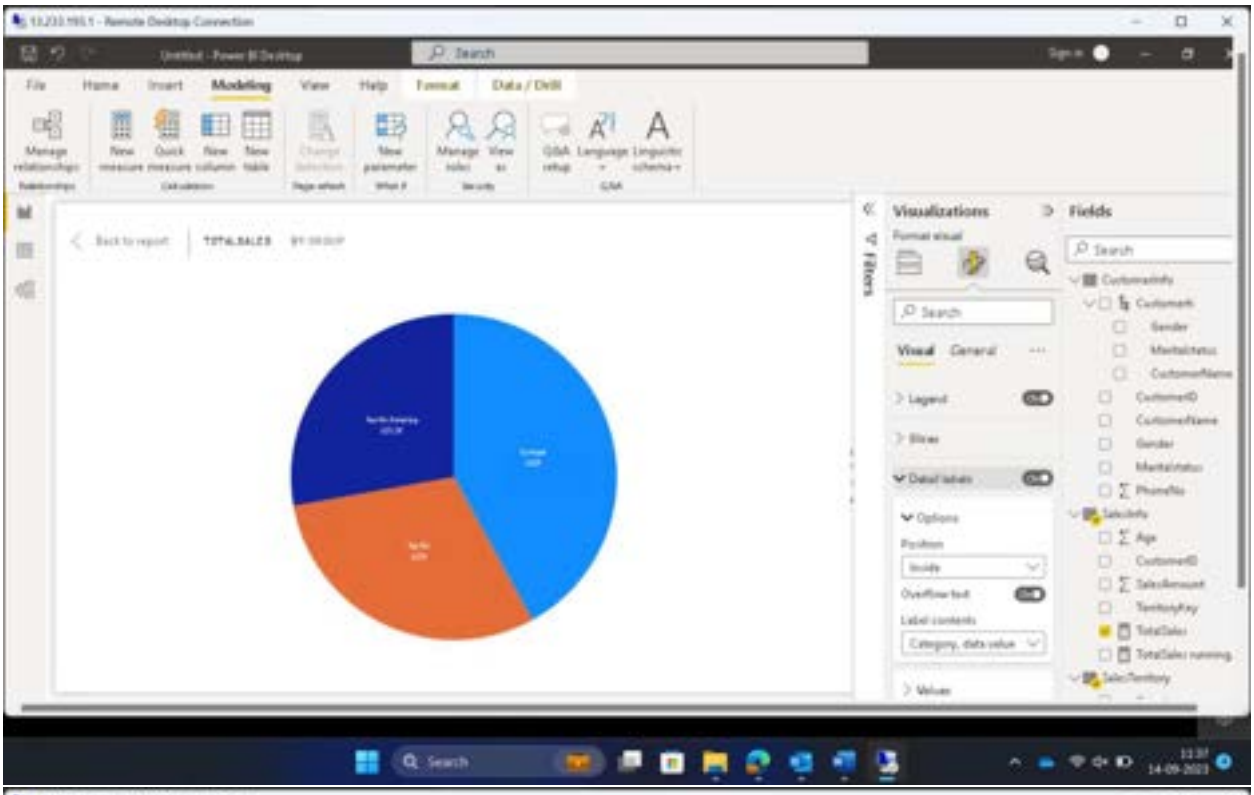

Visualizations
Fields
Build visual

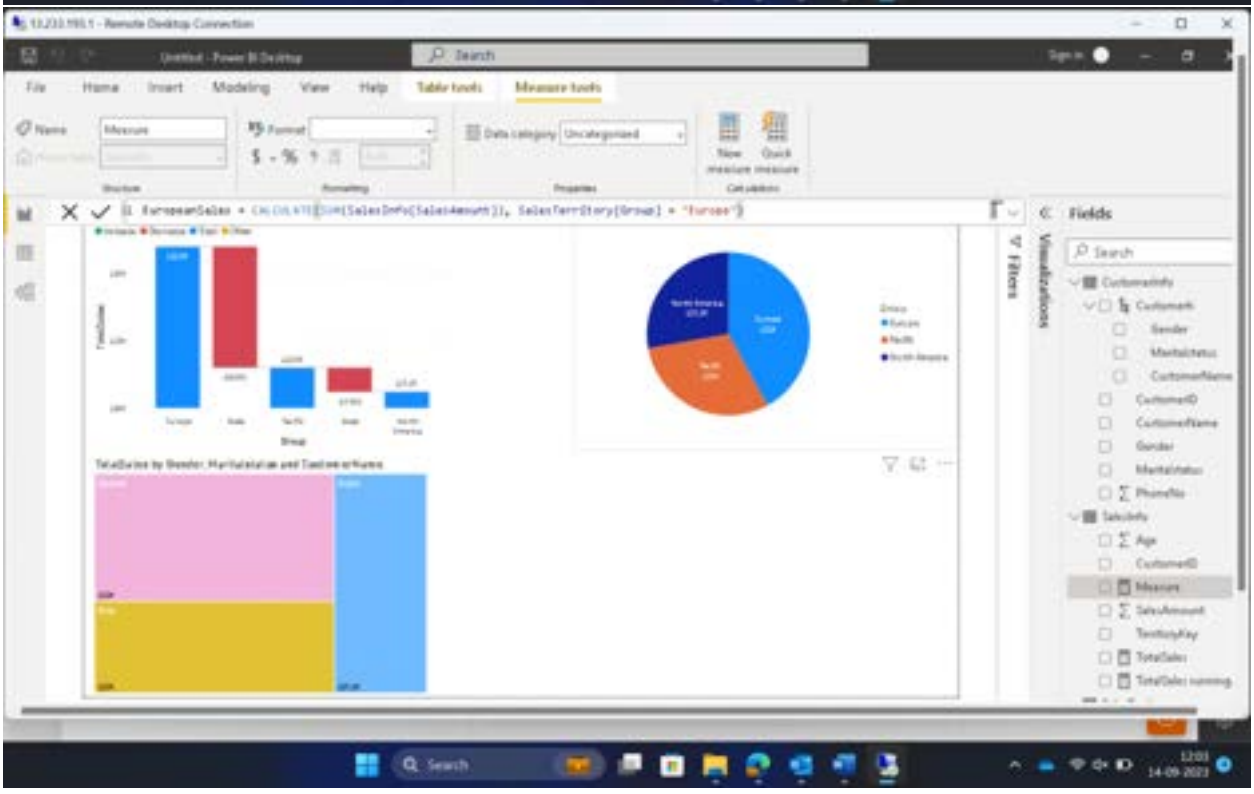

Quick measure
Calculation
Running total
Calculate the running total over a measure in a specific field. Learn more
Base value
TotalSales
Field
Country
Direction
Descending
Fields
Search
CustomerInfo
SalesInfo
Age
CustomerID
SalesAmount
TerritoryKey
TotalSales
SalesTerritory
Country
Country (group)
Group
TerritoryId

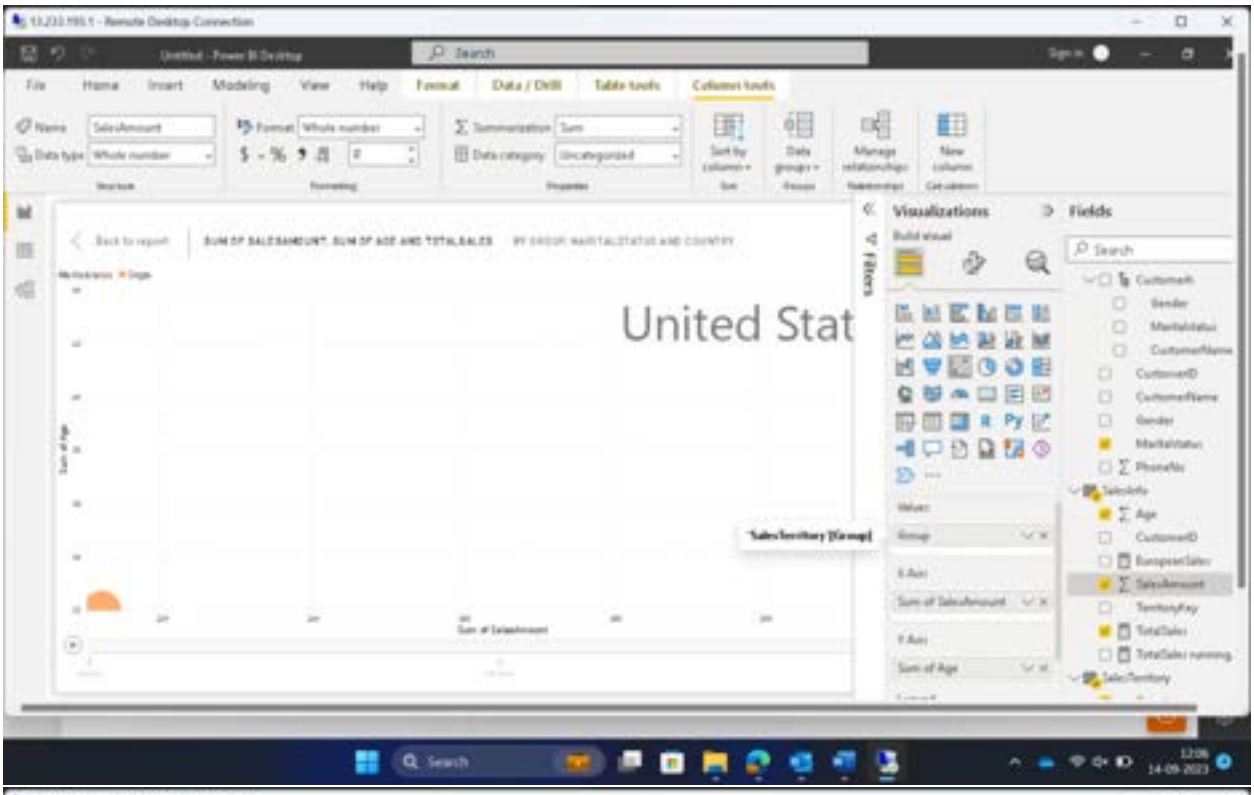

File
Home
Insert
Modeling
View
Help
Format
Data / Drill
Table tools
Measure tools
TotalSales = SUM(SalesInfo[SalesAmount])
Fields
Search
Visualizations
Filters

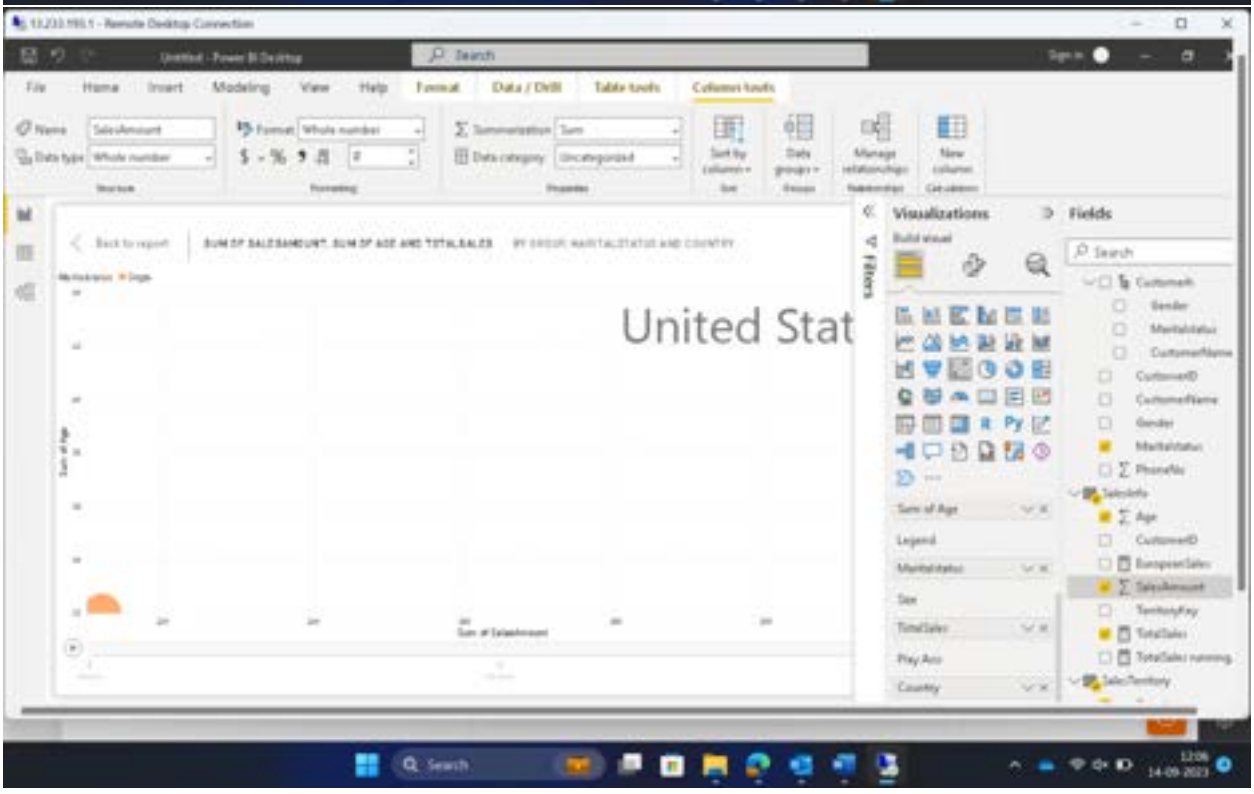


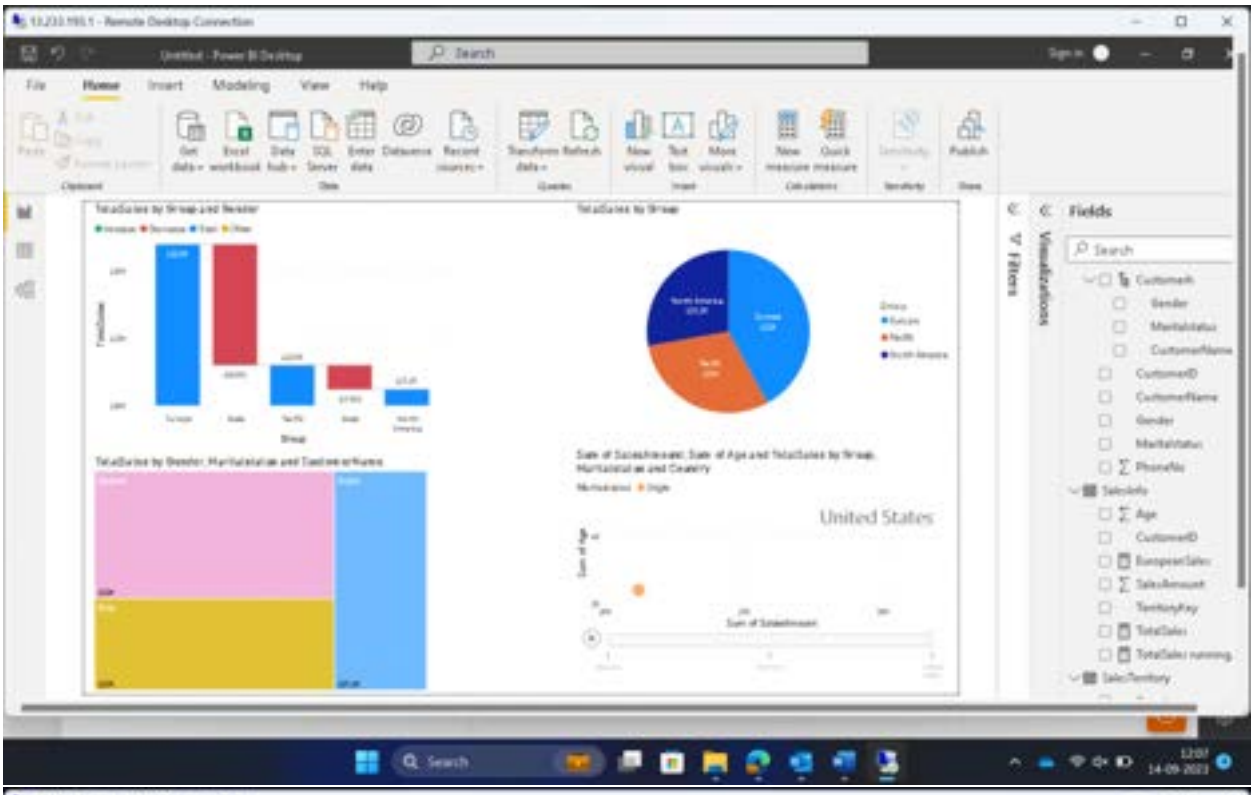


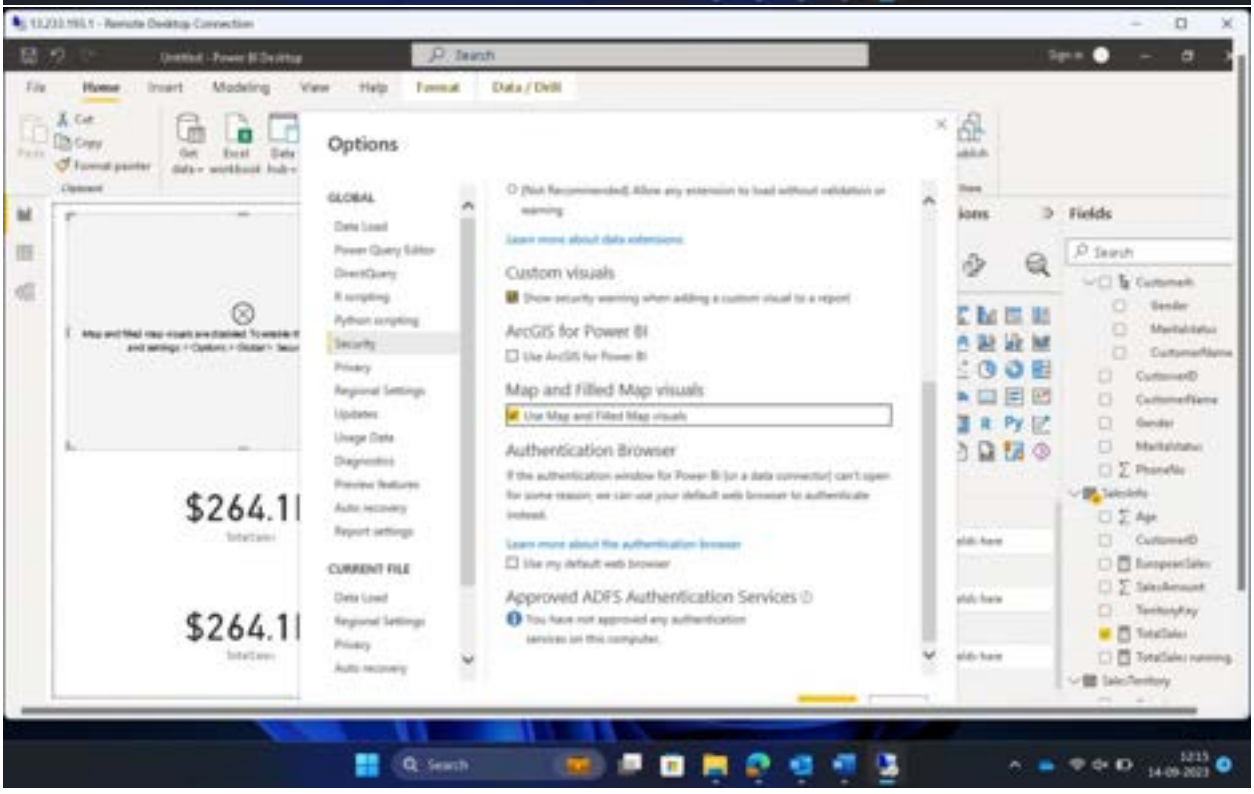

File
Home
Insert
Modeling
View
Help
Format
Data / Drill
Visualizations
Fields
Visual
General
Legend
Bars
Data labels
Options
Position
Overflow text
Label contents
Values

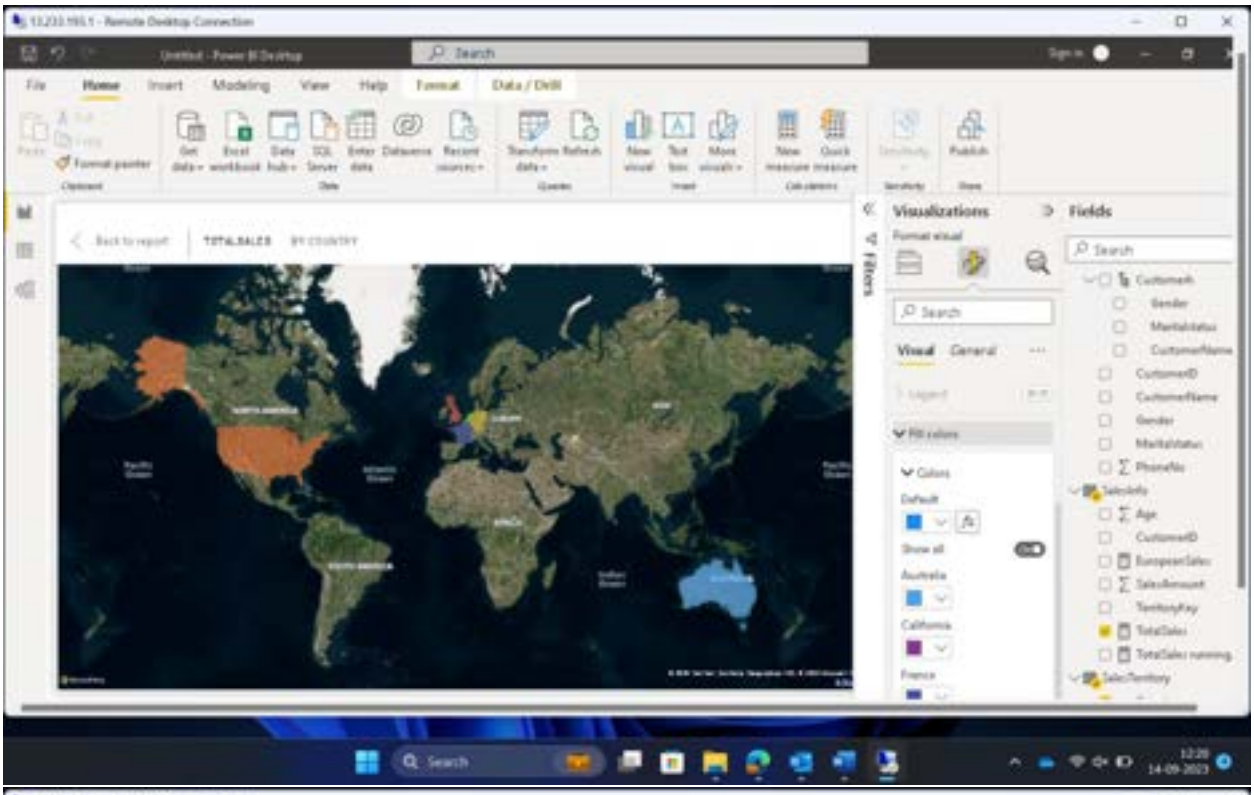

File
Home
Insert
Modeling
View
Help
Table tools
Measure tools
Fields
Visualizations
Filters

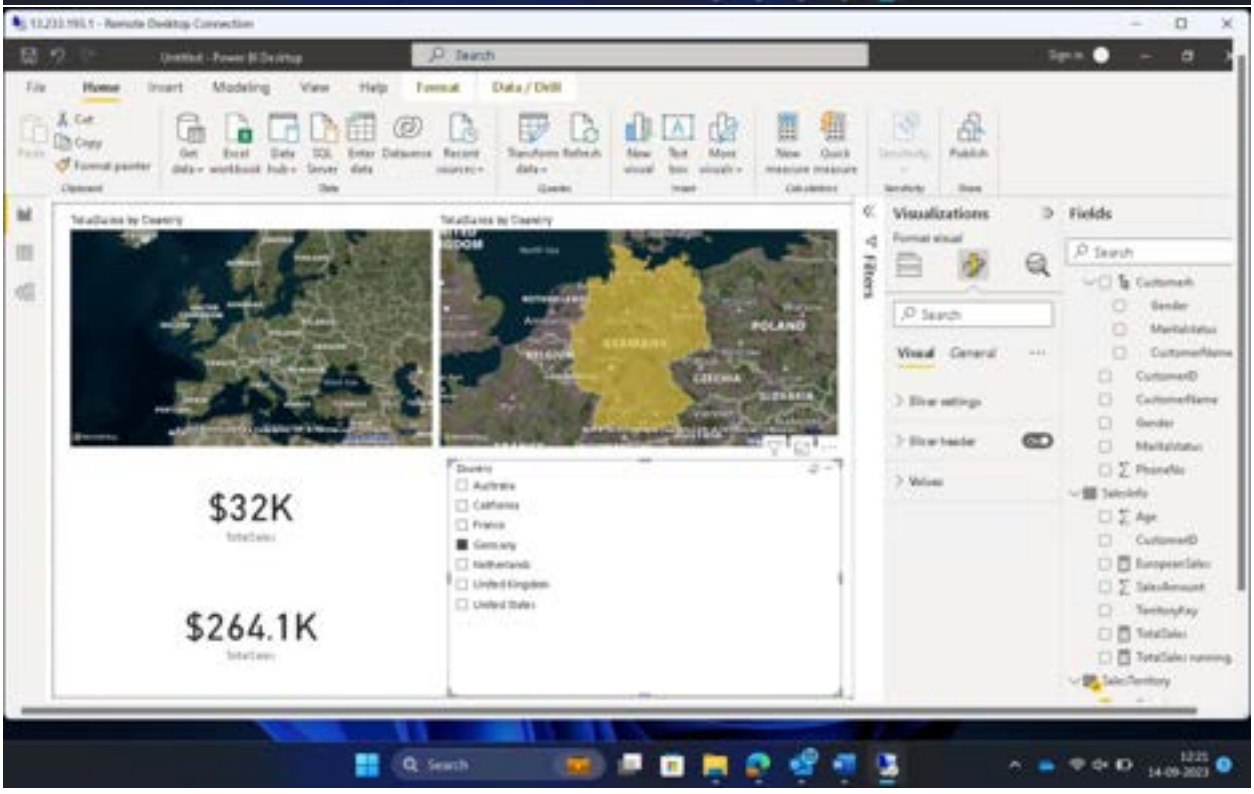

File
Home
Insert
Modeling
View
Help
Format
Data / Drill
Table tools
Visualizations
Fields
United Stat

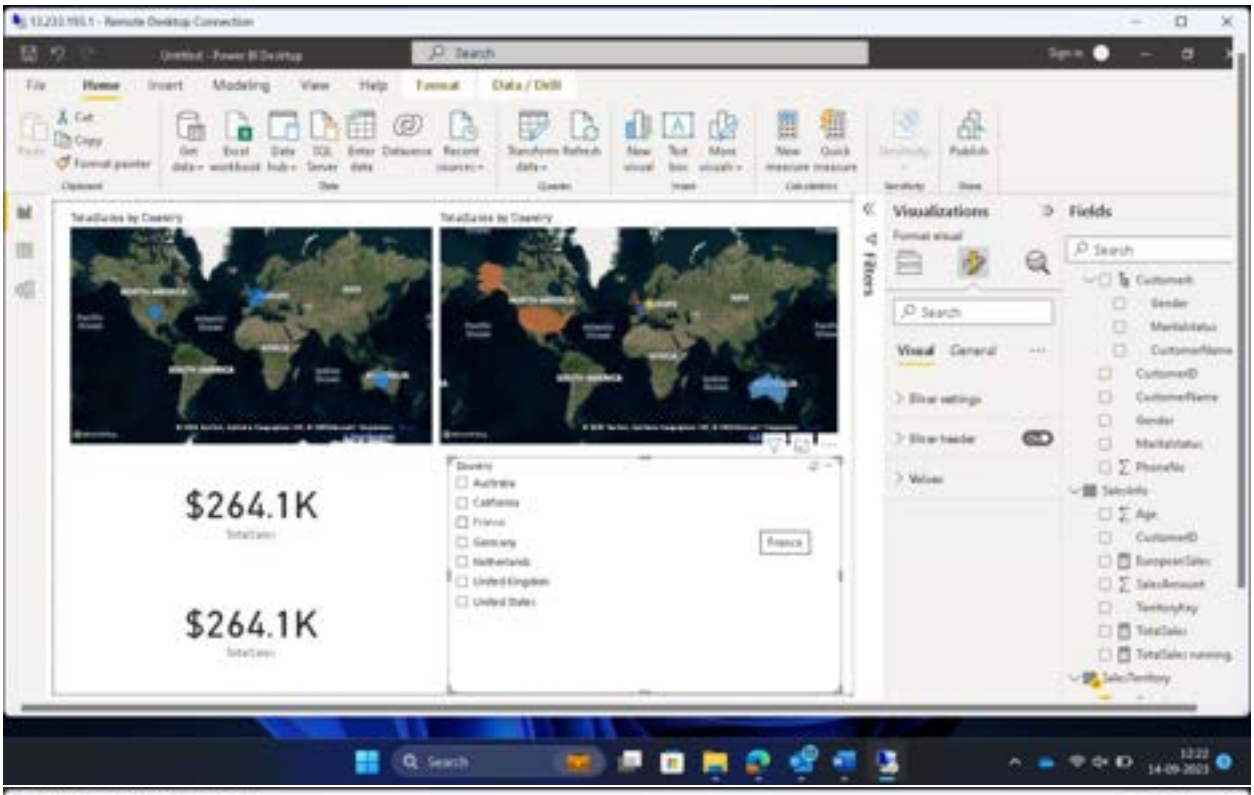

File
Home
Insert
Modeling
View
Help
Format
Data / Drill
Table tools
Visualizations
Fields
United Stat

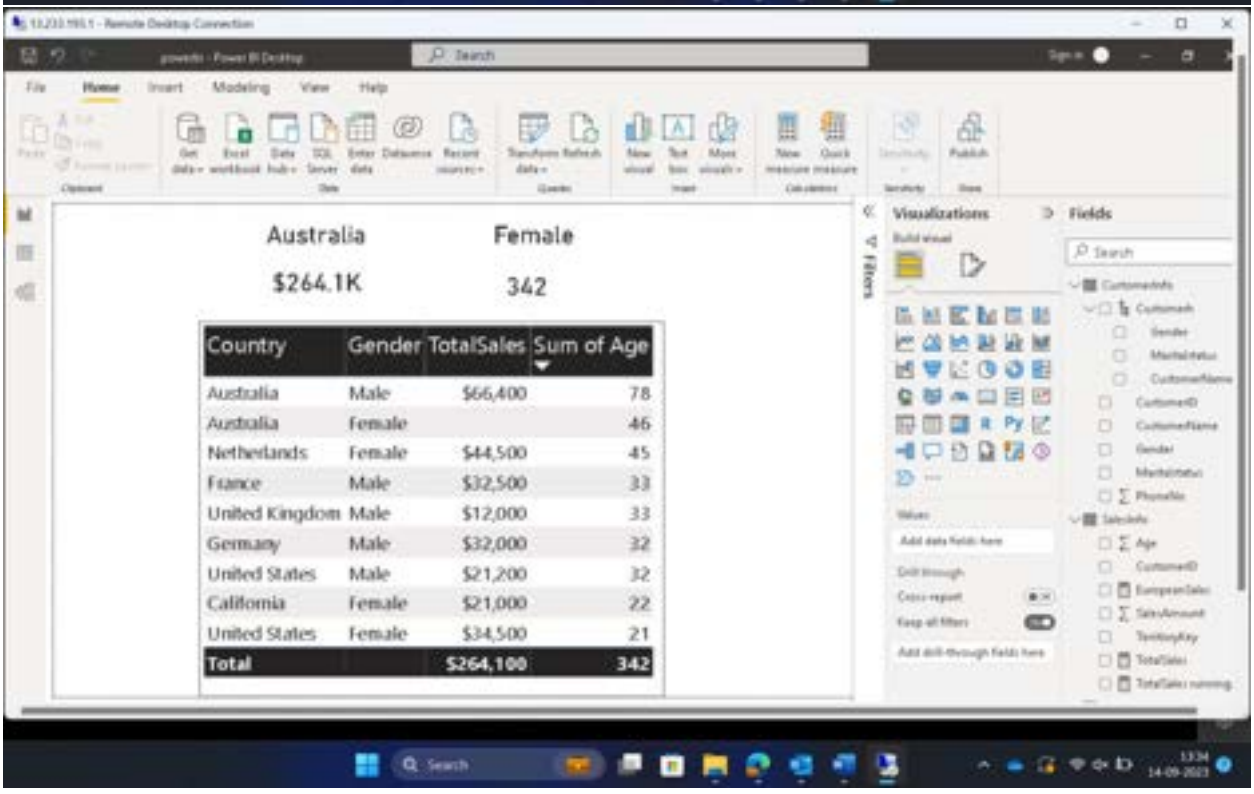

File
Home
Insert
Modeling
View
Help
Fields
United States


File
Home
Insert
Modeling
View
Help
Format
Data / Drill
Options
GLOBAL
Data Load
Power Query Editor
DirectQuery
R scripting
Python scripting
Security
Privacy
Regional Settings
Updates
Usage Data
Diagnostics
Preview features
Auto recovery
Report settings
CURRENT FILE
Data Load
Regional Settings
Privacy
Auto recovery
Custom visuals
Show security warning when adding a custom visual to a report
ArcGIS for Power BI
Use ArcGIS for Power BI
Map and Filled Map visuals
Use Map and Filled Map visuals
Authentication Browser
Use my default web browser
Approved ADFS Authentication Services
$264.1
$264.1
Fields

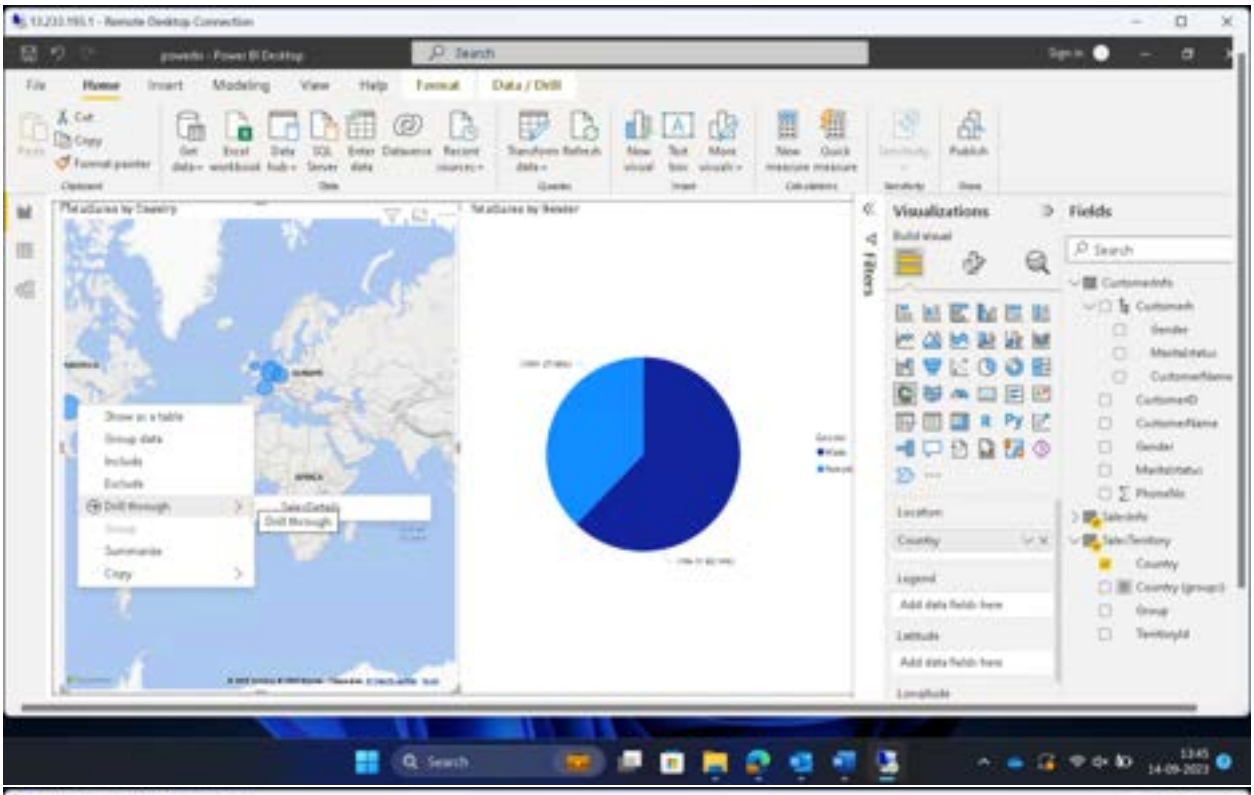

Visualizations
Fields
Filters
Visual
General

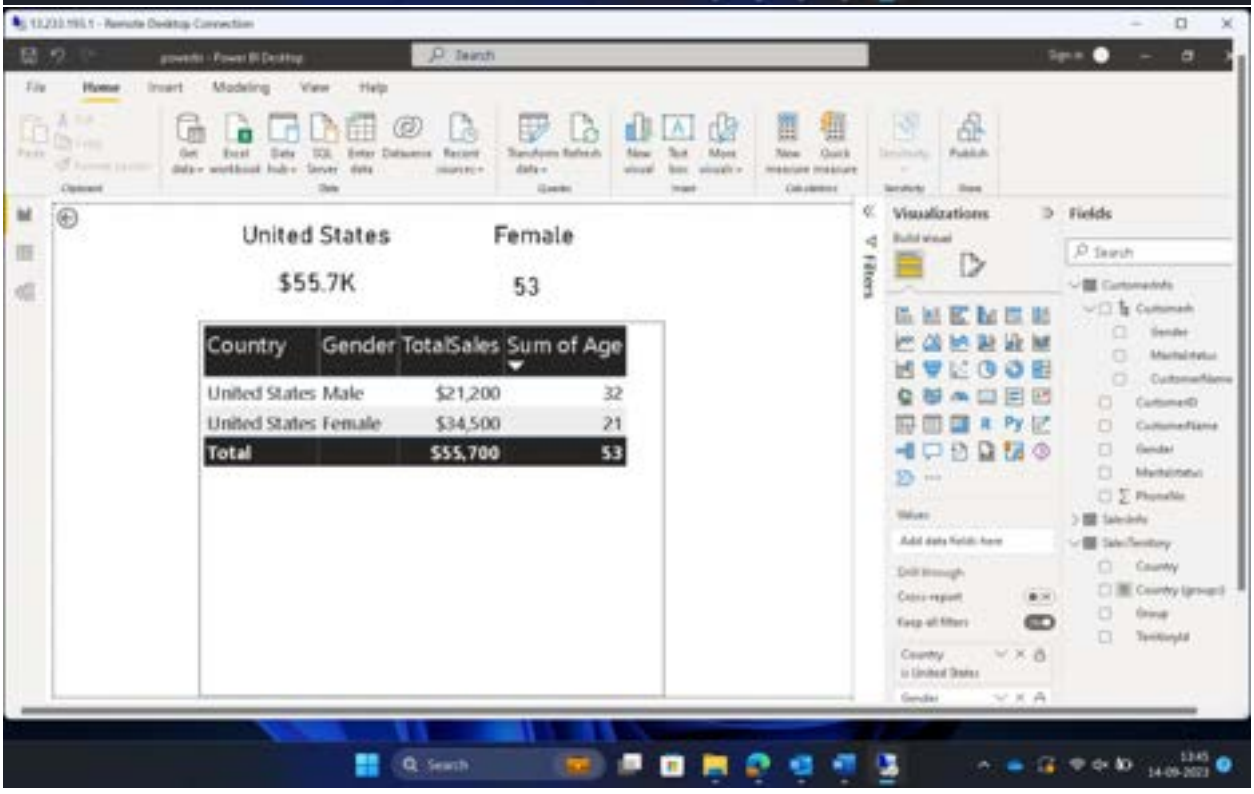

$32K
$264.1K
POLAND
Visualizations
Fields
Filters
Visual
General

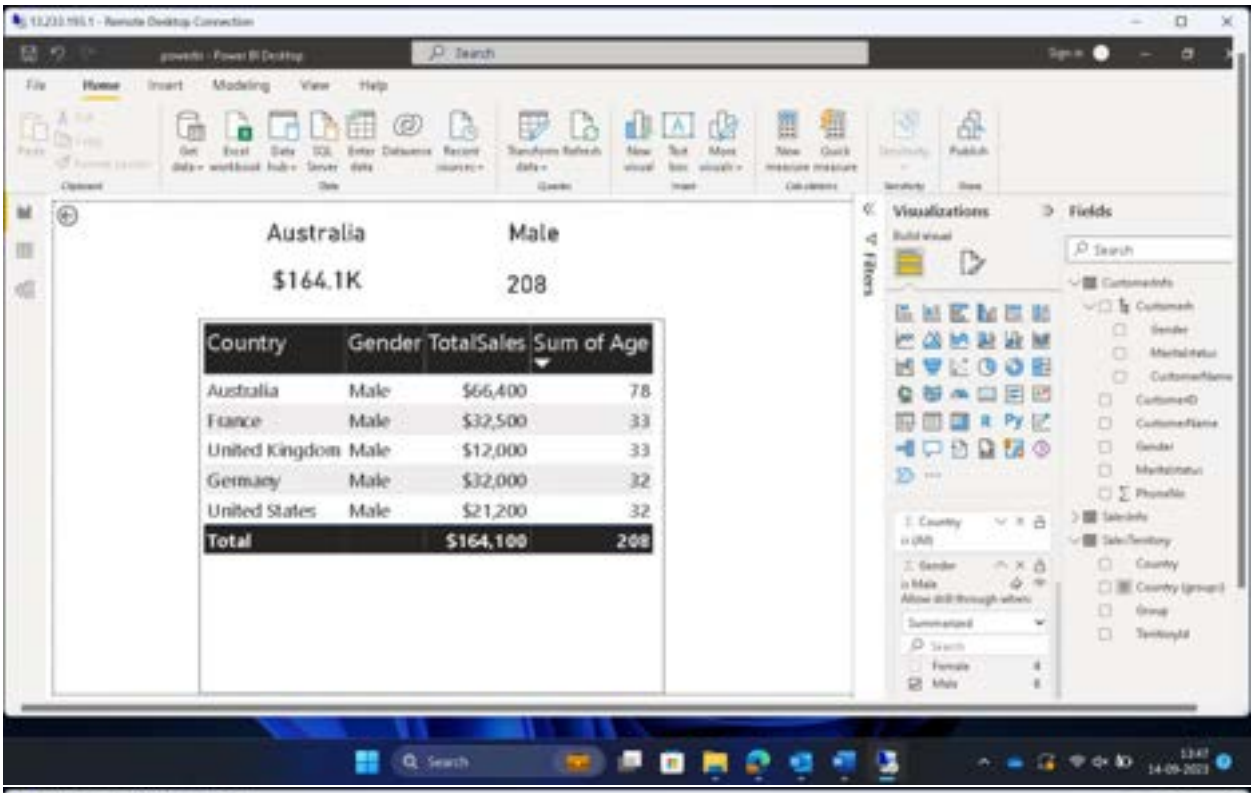


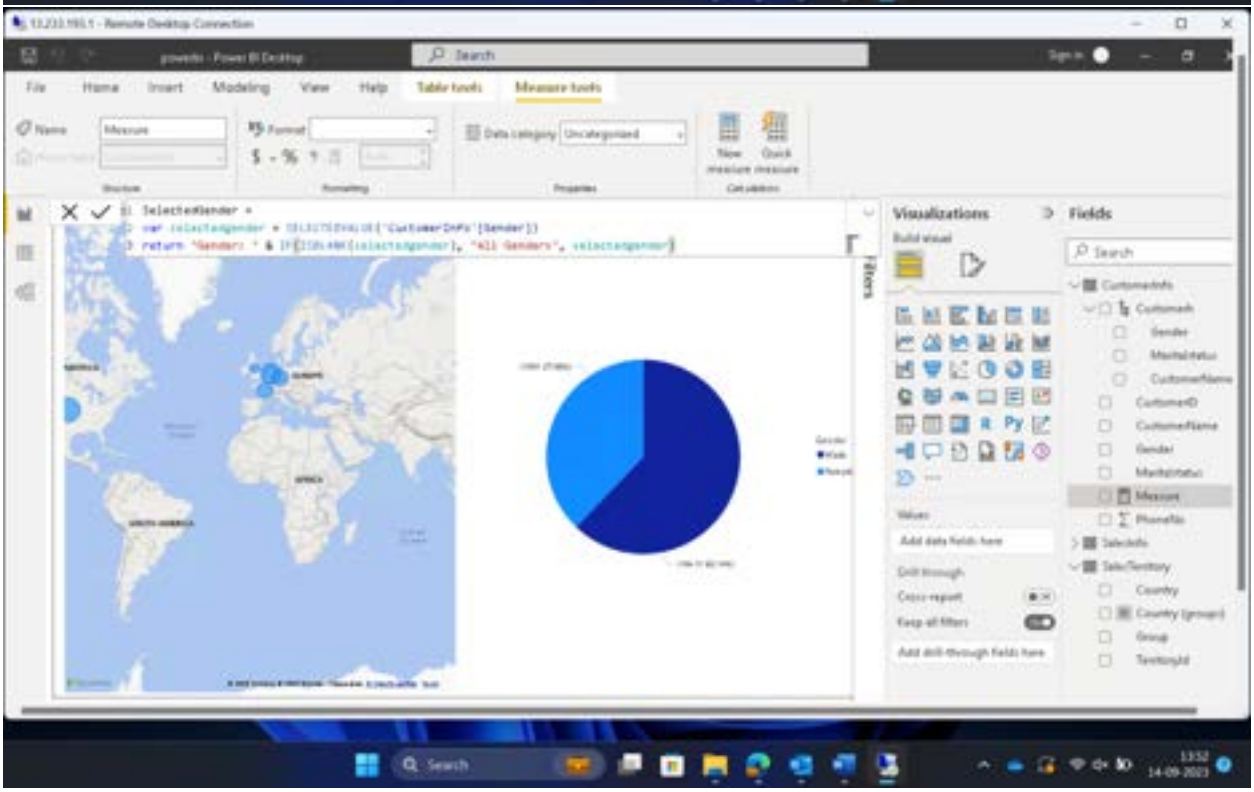


Country	Gender	TotalSales	Sum of Age
Australia	Male	$66,400	78
Australia	Female		46
Netherlands	Female	$44,500	45
France	Male	$32,500	33
United Kingdom	Male	$12,000	33
Germany	Male	$32,000	32
United States	Male	$21,200	32
California	Female	$21,000	22
United States	Female	$34,500	21
Total		**$264,100**	**342**

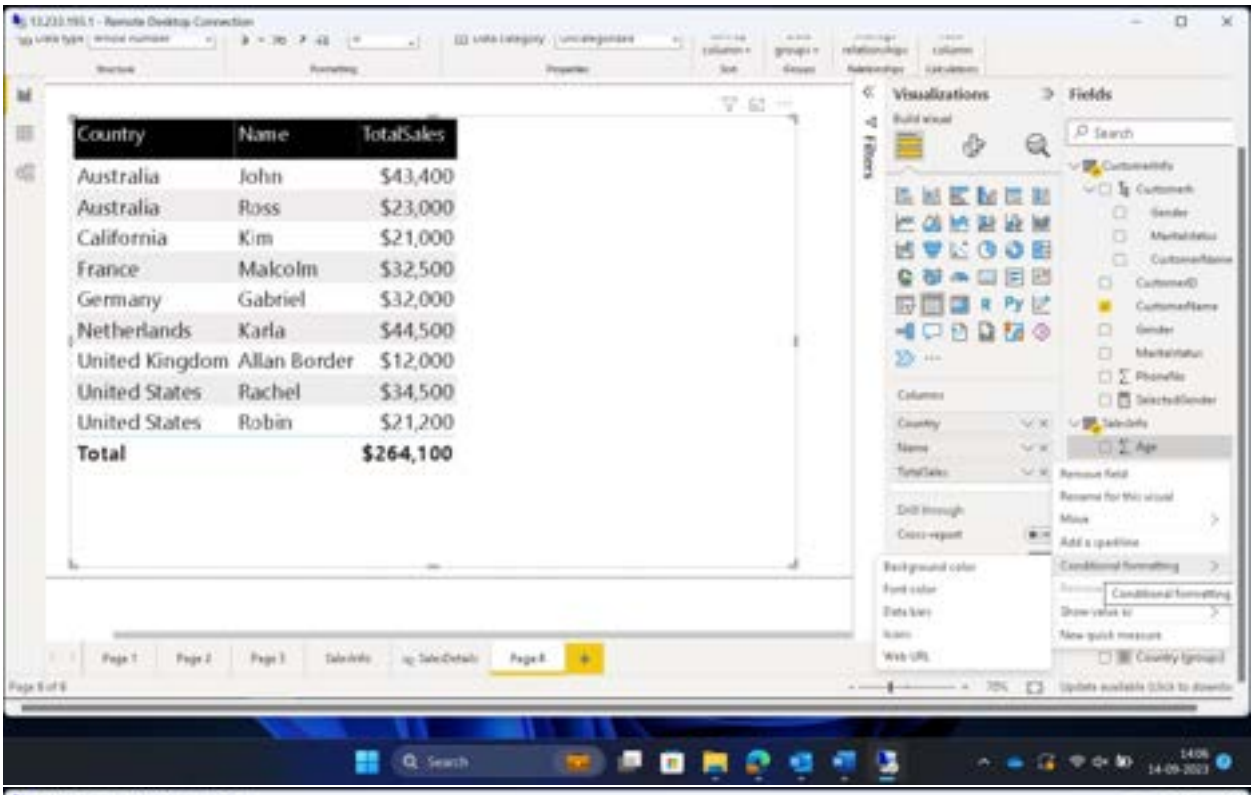


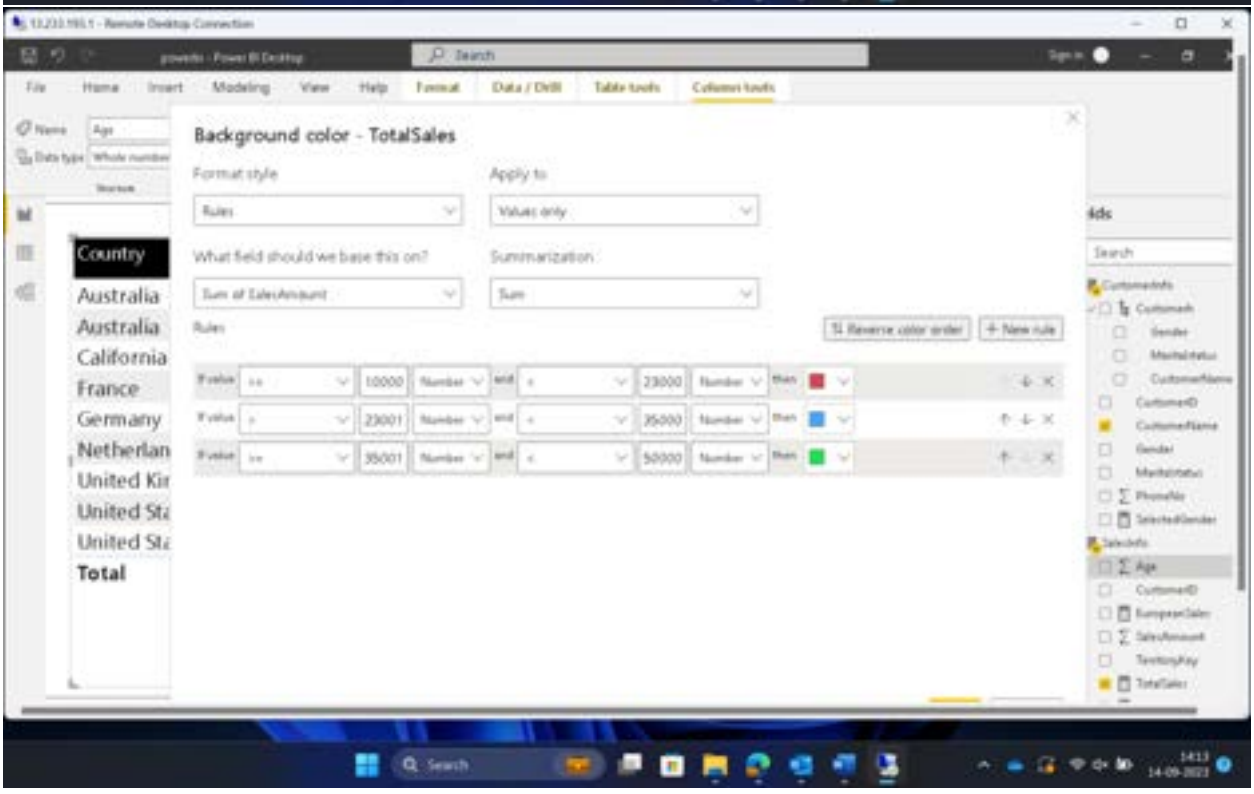

United States
Female
$55.7K
53
Country Gender TotalSales Sum of Age
United States Male $21,200 32
United States Female $34,500 21
Total $55,700 53

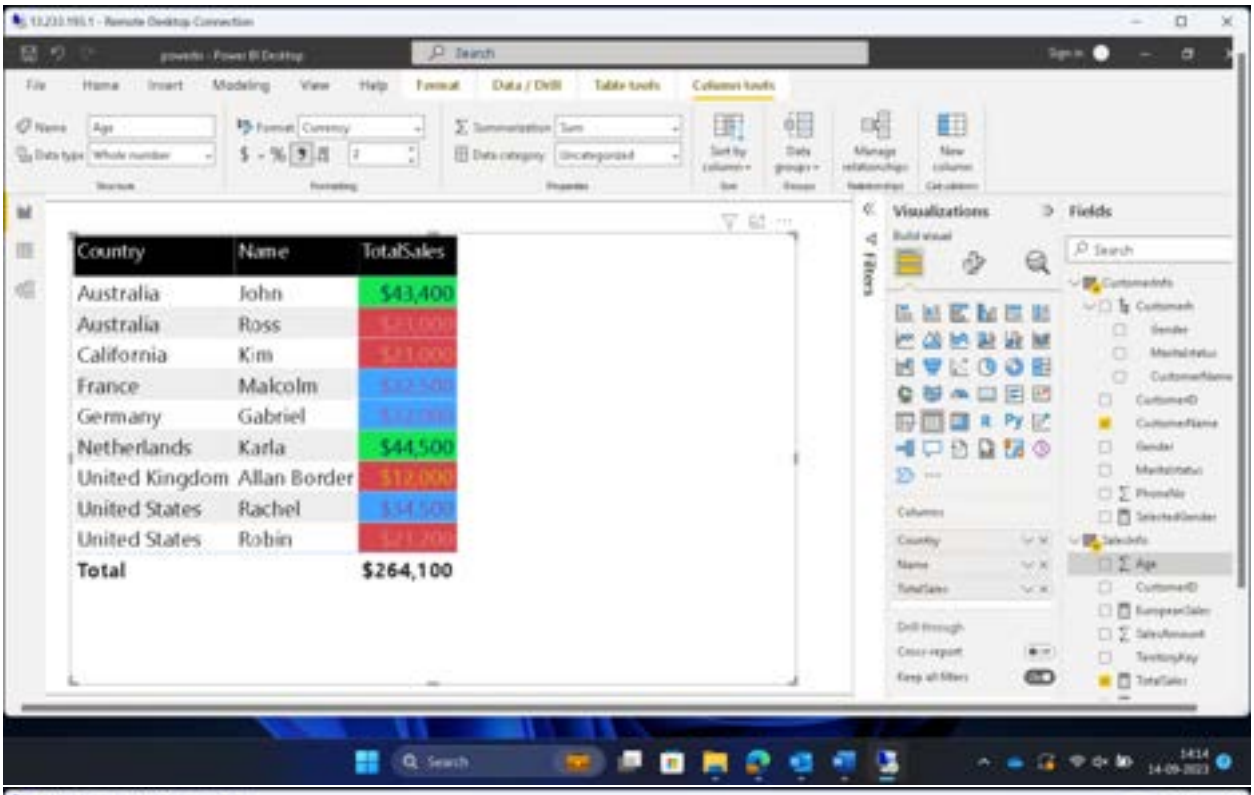


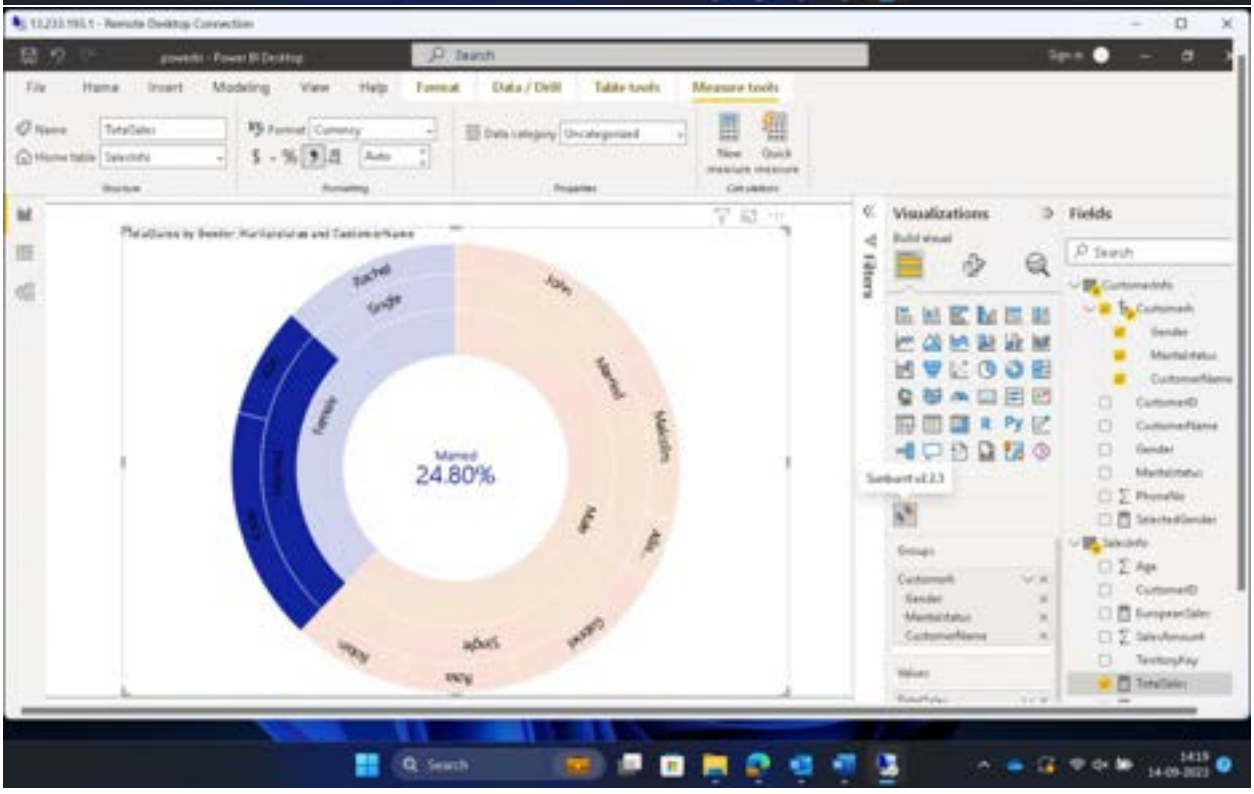


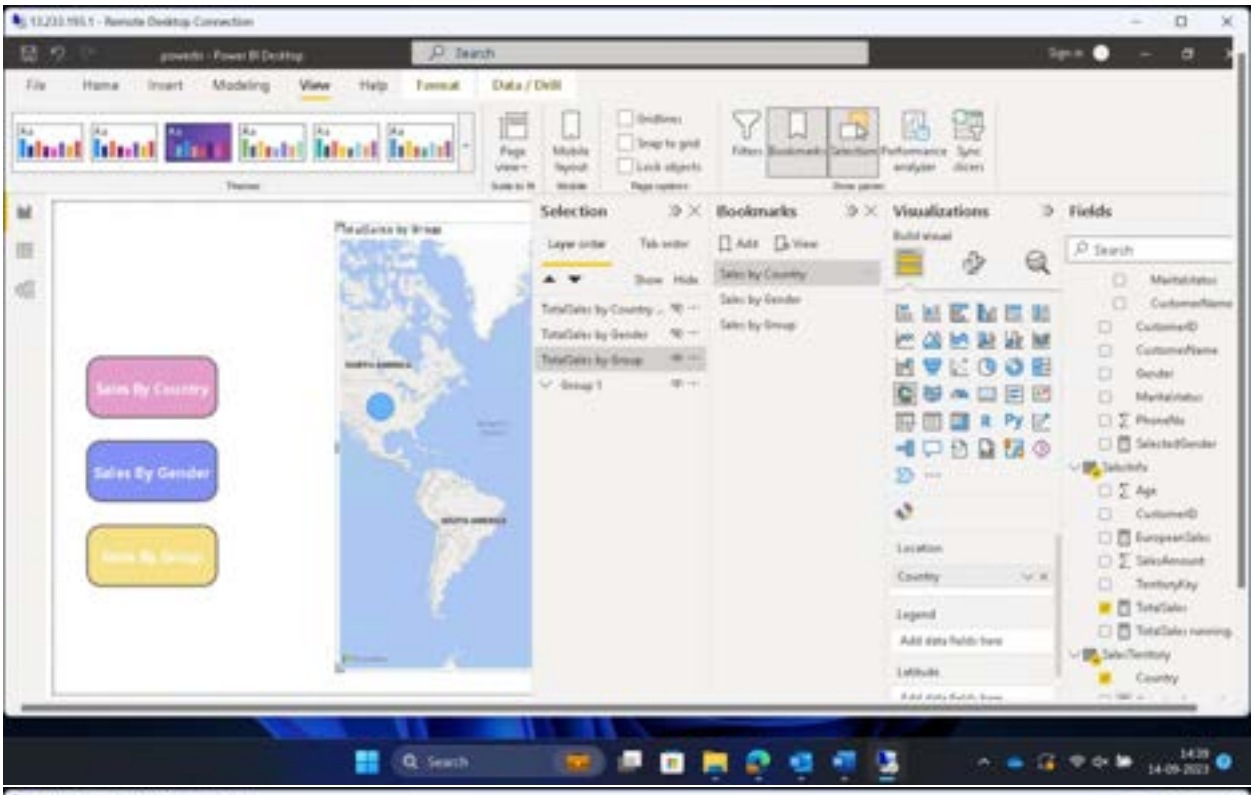

```
SelectedGender =
var selectedgender = SELECTEDVALUE('CustomerInfo'[Gender])
return "Gender: " & IF(ISBLANK(selectedgender), "All Genders", selectedgender)
```

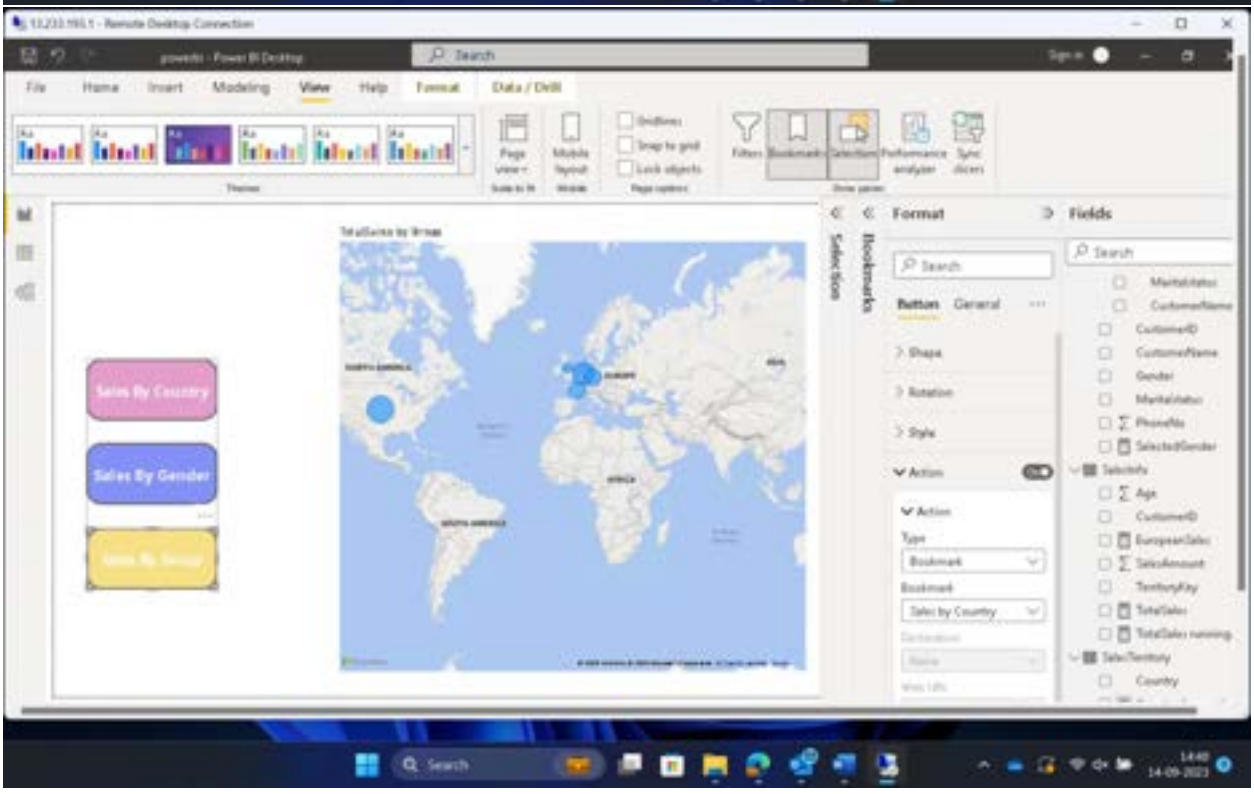

Country
Name
TotalSales
Australia
John
$43,400
Australia
Ross
$23,000
California
Kim
$21,000
France
Malcolm
$32,500
Germany
Gabriel
$32,000
Netherlands
Karla
$44,500
United Kingdom
Allan Border
$12,000
United States
Rachel
$34,500
United States
Robin
$21,200
Total
$264,100
Visualizations
Fields

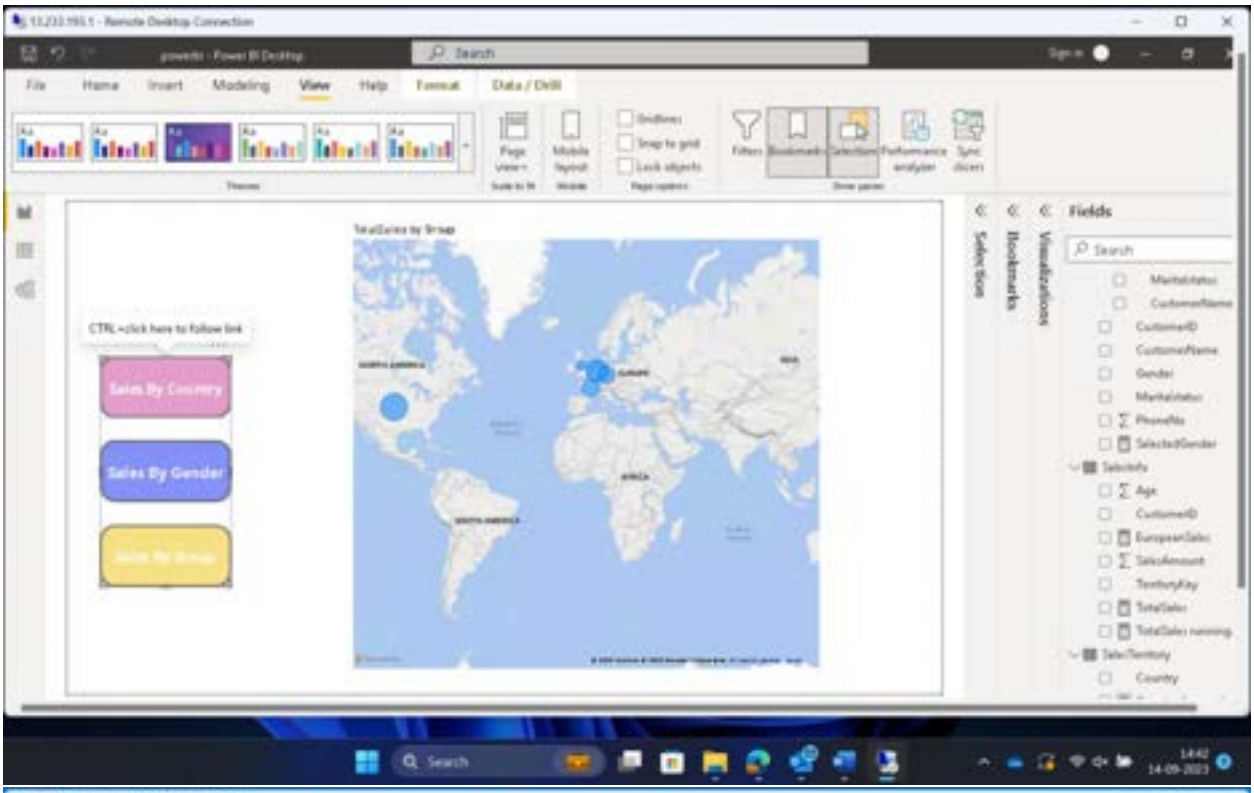

Background color - TotalSales
Format style
Apply to
Rules
Values only
What field should we base this on?
Summarization
Sum of SalesAmount
Sum
Rules
Reverse color order
New rule
Country
Total

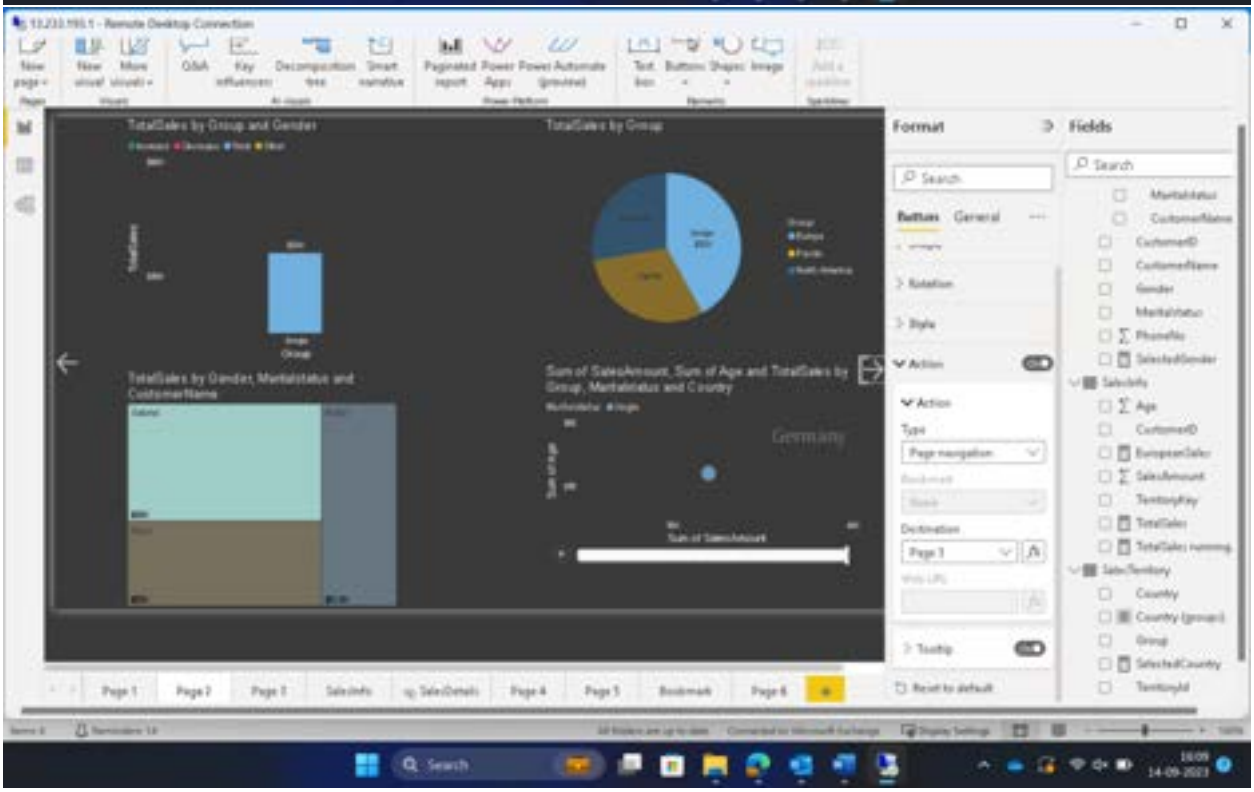

Country
Name
TotalSales
Australia
John
$43,400
Australia
Ross
California
Kim
France
Malcolm
Germany
Gabriel
Netherlands
Karla
$44,500
United Kingdom
Allan Border
$12,000
United States
Rachel
United States
Robin
Total
$264,100
Visualizations
Fields


Married
24.80%
Visualizations
Fields


Sales By Country
Sales By Gender
Selection
Bookmarks
Visualizations
Fields


Sales By Country
Sales By Gender
Format
Fields






```
DateRange =
DATESBETWEEN(FiscalCaldender[Date],"1-1-2023","5-4-2023")
```

```
FiscalCaldender = CALENDARAUTO(3)
```

```
Date = CALENDAR(MIN(Drilling_Dataset[Date]), MAX(Drilling_Dataset[Date]))
```

19.09.23

In [3]:
```
def my_function(*products):
    print("The costly product in the collection is: " + products[3])
my_function("Mouse","Keyboard","Laptop","Speaker","Projector")
```

```
The costly product in the collection is: Speaker
```

In [7]:
```
def my_function(**product):
    print("the hardware product name is:" + product["brandname"])

my_function(productname=  "mouse", brandname="Logitech")
```

```
the hardware product name is:Logitech
```

In [8]:
```
def my_function(**product):
    print("the hardware product name is:" + product["productname"])

my_function(productname=  "mouse", brandname="Logitech")
```

```
the hardware product name is:mouse
```

In [10]:
```
#keyword only arguments

def nameAge(name,age):
    print("Hi, I am:" ,name)
    print("Hi, My age is:" ,age)

#following the order of arguments
nameAge(name="annika",age=22)

#changing the order of arguments
nameAge(age=22,name="annika")
```

```
Hi, I am: annika
Hi, My age is: 22
Hi, I am: annika
Hi, My age is: 22
```

In [11]:
```
#positional parameter demo

def nameAge(name,age):
    print("Hi, I am:" ,name)
    print("Hi, My age is:" ,age)

print("Case-1:")
nameAge("Remi",20)
print("Case-2:")
nameAge(20,"Rem")
```

```
Case-1:
Hi, I am: Remi
Hi, My age is: 20
Case-2:
Hi, I am: 20
Hi, My age is: Rem
```

In [14]:
```
def minus(firstnum,secondnum):
    return firstnum-secondnum

firstnum,secondnum=20,10
result1=minus(firstnum,secondnum)
print("Used positional args",result1)
result2=minus(secondnum,firstnum)
print("Used positional args",result2)
```

```
Used positional args 10
Used positional args -10
```

In [15]:
```
class Person:
    def __init__(self,name,age):
        self.name=name
        self.age=age
p1= Person("Annika",22)

print(p1.name)
print(p1.age)
```

```
Annika
22
```

In [18]:
```
class Person:
    def __init__(self,name,age):
        self.name=name
        self.age=age

    def __str__(self):
        return f"{self.name}({self.age})" #f= float

p1= Person("Annika",22)

print(p1)
```

```
Annika(22)
```

In [23]:
```
class Person:
    def __init__(self,name,age):
        self.name=name
        self.age=age

    def myfunc(self):
        print("Hello my name is " + self.name + " and age is "+ str(self.age))

p1= Person("Annika",22)

p1.myfunc()
```

```
Hello my name is Annika and age is 22
```

In [24]:
```
class Person:
    def __init__(mysillyobject,name,age):
        mysillyobject.name=name
        mysillyobject.age=age

    def myfunc(abc):
        print("Hello my name is " + abc.name + " and age is "+ str(abc.age))

p1= Person("Annika",22)

p1.myfunc()
```

```
Hello my name is Annika and age is 22
```

In [27]:
```
#using zip() function
a= ("John","Oliver","Malcolm","Berry")
b=("Doe","Queen","Merlin","Remi")
result = zip(a,b)
tuple(result)
```

Out[27]:
```
(('John', 'Doe'),
 ('Oliver', 'Queen'),
 ('Malcolm', 'Merlin'),
 ('Berry', 'Remi'))
```

In [30]:
```
#enumerate
l1 =["eat","sleep","repeat"]
s1= "geek"

#creating enumerate objects
obj1=enumerate(l1)
obj2=enumerate(s1)

print("Return type:" , type(obj1))
print(list(enumerate(l1)))

#changing start index to 2 from 0
print(list(enumerate(s1,2)))
```

```
Return type: <class 'enumerate'>
[(0, 'eat'), (1, 'sleep'), (2, 'repeat')]
[(2, 'g'), (3, 'e'), (4, 'e'), (5, 'k')]
```

Day 15 20.09.23

In [5]:
```
#string formatting
a = int(input("Enter first number"))
b = int(input("Enter second number"))

#addnum = a+b
print (f"The addition of {a} and {b} is: ", a+b)
```

```
Enter first number3
Enter second number6
The addition of 3 and 6 is:  9
```

In [6]:
```
print (f"The difference of {a} and {b} is: ", a-b)
```

```
The difference of 3 and 6 is:  -3
```

In [ ]:

In [8]:
```
#if condition
#min age for voting

age = int(input("Enter the age: "))

if age>=18:
    print("Eligible to vote")
else:
    print("Not eligible to vote! ")
    print(f"Eligible to vote after {18-age} years")
```

```
Enter the age: 16
Not eligible to vote!
Eligible to vote after 2 years
```

In [9]:
```
#raw string
print("C:/programfiles/backup/newfolder/abcd/xyz")
```

```
C:/programfiles/backup/newfolder/abcd/xyz
```

In [10]:
```
print("C:\Users\Annika.Chandra\OneDrive - Shell\Desktop\Shell bootcamp\Customer.xlsx")
#in \t \b \a \e they are keyboard tabs
```

```
  Cell In[10], line 1
    print("C:\Users\Annika.Chandra\OneDrive - Shell\Desktop\Shell bootcamp\Customer.xlsx")
                                                                                        ^
SyntaxError: (unicode error) 'unicodeescape' codec can't decode bytes in position 2-3: truncated \UXXXXXXXX escape
```

In [11]:
```
#printing the string as it is

print(r"C:\Users\Annika.Chandra\OneDrive - Shell\Desktop\Shell bootcamp\Customer.xlsx")
```

```
C:\Users\Annika.Chandra\OneDrive - Shell\Desktop\Shell bootcamp\Customer.xlsx
```

In [21]:
```
s.add(15)
s #only possible to print directly in jupyter notebook
```

Out[21]:
```
{15, 23, 54, 65, 77, 87}
```

In [22]:
```
print(s) #everywhere we need to write print statement
```

```
{65, 54, 23, 87, 77, 15}
```

In [25]:
```
s.discard(23)
s
```

Out[25]:
```
{15, 54, 65, 77, 87}
```

In [ ]:

In [25]:
```
s.discard(23)
s
```

Out[25]:
```
{15, 54, 65, 77, 87}
```

In [26]:
```
s.remove(54)
s
```

Out[26]:
```
{15, 65, 77, 87}
```

In [27]:
```
s.pop() #pop only deletes the first element
s
```

Out[27]:
```
{15, 77, 87}
```

In [28]:
```
s1 = {32,5,33,10}

s | s1 #union
```

Out[28]:
```
{5, 10, 15, 32, 33, 77, 87}
```

In [29]:
```
s & s1 #intersection
```

Out[29]:
```
set()
```

In [30]:
```
s - s1 #difference   s - (set1&set2)
```

Out[30]:
```
{15, 77, 87}
```

In [31]:
```
s ^ s1 #symmetric difference (union of uncommon elements)
```

Out[31]:
```
{5, 10, 15, 32, 33, 77, 87}
```

In [36]:
```
#dictionary

d1 = {5: "C", 3:"JAVA" , 10:"CPP", 7:"Python", 9:"Scala"}
type(d1)
```

Out[36]:
```
dict
```

In [37]:
```
d1
```

Out[37]:
```
{5: 'C', 3: 'JAVA', 10: 'CPP', 7: 'Python', 9: 'Scala'}
```

In [38]:
```
d1[3]
```

Out[38]:
```
'JAVA'
```

In [41]:
```
d1[10]
```

Out[41]:
```
'CPP'
```

In [42]:
```
d1[1] = "Javacript"
d1
```

Out[42]:
```
{5: 'C', 3: 'JAVA', 10: 'CPP', 7: 'Python', 9: 'Scala', 1: 'Javacript'}
```

In [43]:
```
d1.get(7)
```

Out[43]:
```
'Python'
```

In [45]:
```
d1.get(6,"key not found") #get() is better than using [] while accessing any element
```

Out[45]:
```
'key not found'
```

In [56]:
```
editors={'localhost': 'Jupyter Notebook',
         'offline' : ['Notepad++', 'Pycharm', 'VSCode' , 'Spyder' , 'Atom'],
          'online' : {'google' : 'colaboratory', 'aws':'Sagemaker', 'azure':'azure ml studio'}}
editors
```

Out[56]:
```
{'localhost': 'Jupyter Notebook',
 'offline': ['Notepad++', 'Pycharm', 'VSCode', 'Spyder', 'Atom'],
 'online': {'google': 'colaboratory',
  'aws': 'Sagemaker',
  'azure': 'azure ml studio'}}
```

In [57]:
```
editors['online']
```

Out[57]:
```
{'google': 'colaboratory', 'aws': 'Sagemaker', 'azure': 'azure ml studio'}
```

In [58]:
```
editors['online']['aws']
```

Out[58]:
```
'Sagemaker'
```

In [59]:
```
editors.get('online')
```

Out[59]:
```
{'google': 'colaboratory', 'aws': 'Sagemaker', 'azure': 'azure ml studio'}
```

In [62]:
```
list1 = ['p1','p2','p3','p4']
list2 = ['pizza','burger','pasta','donut']

#creating dictionary using zip() on lists , can also work with tuple
d2 = dict(zip(list1,list2))
d2
```

Out[62]:
```
{'p1': 'pizza', 'p2': 'burger', 'p3': 'pasta', 'p4': 'donut'}
```

In [4]:
```
%%writefile p1.py

editors={'localhost': 'Jupyter Notebook',
         'offline' : ['Notepad++', 'Pycharm', 'VSCode' , 'Spyder' , 'Atom'],
          'online' : {'google' : 'colaboratory', 'aws':'Sagemaker', 'azure':'azure ml studio'}}
editors
editors['online']
editors['online']['aws']
editors.get('online')
```

```
Writing p1.py
```

In [ ]:
```
%%writing abc.java #any type of file can be created
```

In [9]:
```
#0 dimensions to 32 dimensions
```

In [10]:
```
a = np.array([1,2,3,4]) #vector #single dimensional
a
```

Out[10]:
```
array([1, 2, 3, 4])
```

In [11]:
```
type(a)
```

Out[11]:
```
numpy.ndarray
```

In [12]:
```
a.ndim
```

Out[12]:
```
1
```

In [14]:
```
a = np.array([[1,2,3,4],[2,4,6,3]]) #matrix #two dimensional
a
```

Out[14]:
```
array([[1, 2, 3, 4],
       [2, 4, 6, 3]])
```

In [15]:
```
a.shape
```

Out[15]:
```
(2, 4)
```

In [16]:
```
#for image processing we use 3 dimensional
```

In [17]:
```
a.dtype
```

Out[17]:
```
dtype('int64')
```

In [20]:
```
a = np.array([1,4.4,5,'dx',3])
a.dtype #unicode charater strings
```

Out[20]:
```
dtype('<U32')
```

In [22]:
```
a = np.array(['a','a','b','b','c','c','c'])
a.dtype
```

Out[22]:
```
dtype('<U1')
```

In [25]:
```
a = np.array([20,54,77,22])
np.sort(a)[::-1]
```

Out[25]:
```
array([77, 54, 22, 20])
```

In [27]:
```
a1 =np.array([2.3,2.2,6.1])
a2 = a1.astype(int)
a2
```

Out[27]:
```
array([2, 2, 6])
```

In [28]:
```
np.sin(a2)
```

Out[28]:
```
array([ 0.90929743,  0.90929743, -0.2794155 ])
```

In [29]:
```
np.cos(a2)
```

Out[29]:
```
array([-0.41614684, -0.41614684,  0.96017029])
```

In [30]:
```
np.tan(a2)
```

Out[30]:
```
array([-2.18503986, -2.18503986, -0.29100619])
```

In [30]:
```
np.tan(a2)
```

Out[30]:
```
array([-2.18503986, -2.18503986, -0.29100619])
```

In [31]:
```
np.round(np.tan(a2), 2)
```

Out[31]:
```
array([-2.19, -2.19, -0.29])
```

In [32]:
```
np.round(np.cbrt(a),2) #cube root of all elements
```

Out[32]:
```
array([2.71, 3.78, 4.25, 2.8 ])
```

In [34]:
```
np.round(np.sqrt(a),3)
```

Out[34]:
```
array([4.472, 7.348, 8.775, 4.69 ])
```

In [ ]:
```
np.power
```

```
In [43]: #np.arange() - return evenly spaced values within a given interval

         a = np.arange(1,11,1) #arange(start,stop,step)
         a
Out[43]: array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10])

In [44]: a = np.arange(11) #arange(start,stop,step)
         a
Out[44]: array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10])

In [45]: a = np.arange(0,11) #arange(start,stop,step)
         a
Out[45]: array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10])

In [46]: a = np.arange(1,11,2)
         a
Out[46]: array([1, 3, 5, 7, 9])

In [47]: np.abs(a) # gives absolute values
Out[47]: array([1, 3, 5, 7, 9])
```

```
In [47]: np.abs(a) # gives absolute values
Out[47]: array([1, 3, 5, 7, 9])

In [48]: np.exp(a) #exponential values
Out[48]: array([2.71828183e+00, 2.00855369e+01, 1.48413159e+02, 1.09663316e+03,
                8.10308393e+03])

In [49]: np.pi
Out[49]: 3.141592653589793

In [50]: np.arange(0,11,1)
Out[50]: array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10])

In [52]: #np.linspace   return evenly spaced numbers over a specifical interval

         np.linspace(0,11,10)
Out[52]: array([ 0.        ,  1.22222222,  2.44444444,  3.66666667,  4.88888889,
                 6.11111111,  7.33333333,  8.55555556,  9.77777778, 11.        ])
```

```
In [54]: a = np.linspace(10,50,5)
         a
Out[54]: array([10., 20., 30., 40., 50.])

In [55]: np.unique(a)
Out[55]: array([10., 20., 30., 40., 50.])

In [56]: np.unique(a,return_counts=True)
Out[56]: (array([10., 20., 30., 40., 50.]), array([1, 1, 1, 1, 1]))
```

```
In [61]: np.zeros(10)
Out[61]: array([0., 0., 0., 0., 0., 0., 0., 0., 0., 0.])

In [62]: #core python - str ("aa")
         #pandas - object
```

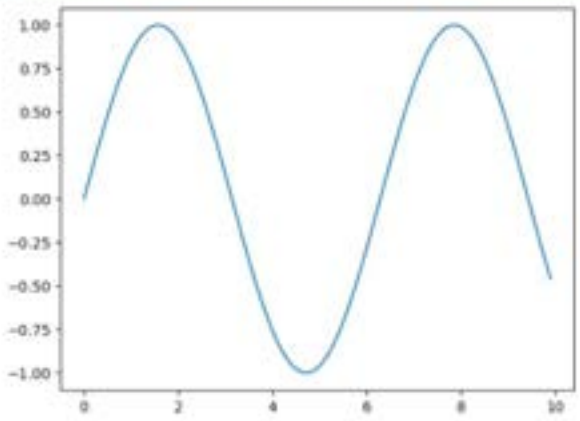
```
In [64]: np.random.randint(0,10,10) #(start,stop,number of random values) #otp generator
Out[64]: array([8, 8, 1, 9, 7, 9, 5, 8, 9, 9])
```

```
In [66]: np.random.choice([10,5,0], size=[3,3])
Out[66]: array([[10, 10,  5],
                [ 0, 10,  5],
                [ 0, 10, 10]])
```

In [70]:
```
np.random.seed(1234) #initialize a particular value so that everytime i run a randint()
#we will get the same values everytime
np.random.randint(0,10,5)
```

Out[70]:
```
array([3, 6, 5, 4, 8])
```

In [86]:
```
y = np.sin(x)
plt.plot(x,y)
plt.show()
```



In [79]:
```
x = np.arange(0,10,0.1)
x
```

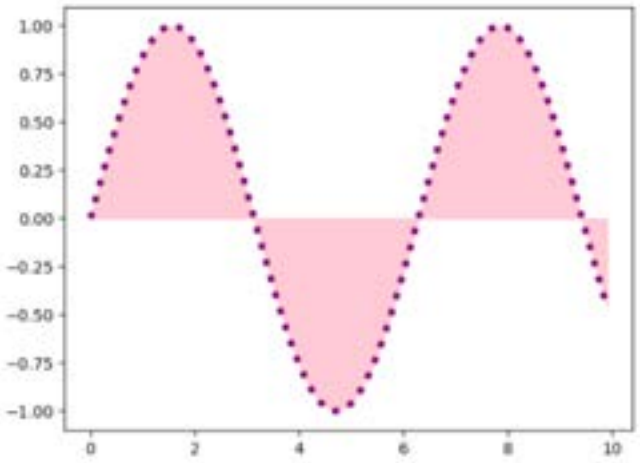
Out[79]:
```
array([0. , 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. , 1.1, 1.2,
       1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2. , 2.1, 2.2, 2.3, 2.4, 2.5,
       2.6, 2.7, 2.8, 2.9, 3. , 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8,
       3.9, 4. , 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 5. , 5.1,
       5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 6. , 6.1, 6.2, 6.3, 6.4,
       6.5, 6.6, 6.7, 6.8, 6.9, 7. , 7.1, 7.2, 7.3, 7.4, 7.5, 7.6, 7.7,
       7.8, 7.9, 8. , 8.1, 8.2, 8.3, 8.4, 8.5, 8.6, 8.7, 8.8, 8.9, 9. ,
       9.1, 9.2, 9.3, 9.4, 9.5, 9.6, 9.7, 9.8, 9.9])
```

In [86]:
```
y = np.sin(x)
plt.plot(x,y)
plt.show()
```

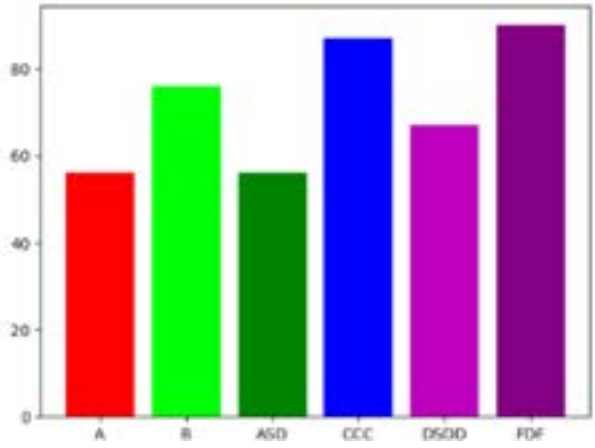
In [74]:
```
a = np.array([[[23,43,222]]])
plt.imshow(a)
```

Out[74]:
```
<matplotlib.image.AxesImage at 0x7f7c7bf1831b>
```

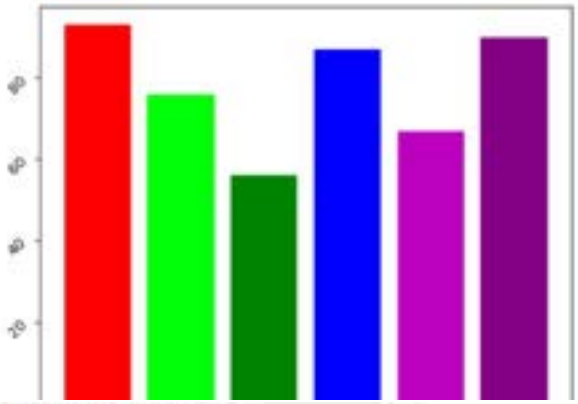
In [94]:
```
y = np.sin(x)
plt.plot(x,y,color= 'purple',linewidth =4,linestyle='dotted')
plt.fill_between(x,y,color='pink',alpha=0.9)
plt.show()
```



In [101]:
```
x = ['A','B','ASD','CCC','DSDD','FDF']
y = [56,76,56,87,67,90] #weight in kgs

colors12 = ['r', '#00ff00', 'g', 'b', 'm', 'purple']
plt.bar(x,y,color=colors12)

plt.show()
```



In [183]:
```
x = ['A','B','ASD','CCC','OSDD','FDF']
y = [93,76,56,87,67,90] #weight in kgs

colors12 = ['r', '#00ff00', 'g', 'b', 'm', 'purple']
plt.bar(x,y,color=colors12)

plt.xticks(rotation=90)
plt.yticks(rotation=45)

plt.show()
```



In [10]:
```
#series 1 dimensional
#data frame 2 dimensional
```

In [14]:
```
import numpy as np
#s1 = [100,200,300,400,500,600]
s1 = np.array([100,200,300,400,500,600])
sr1 = pd.Series(s1)
sr1
```

Out[14]:
```
0    100
1    200
2    300
3    400
4    500
5    600
dtype: int64
```

In [13]:
```
type(sr1)
```

Out[13]:
```
pandas.core.series.Series
```

In [18]:
```
d = {'col1': [100,200,300,400,500,600], 'col2': [100,200,300,400,234,500]}
s2 = pd.Series(d)
s2
```

Out[18]:
```
col1    [100, 200, 300, 400, 500, 600]
col2    [100, 200, 300, 400, 234, 500]
dtype: object
```

In [18]:
```
d = {'col1': [100,200,300,400,500,600], 'col2': [100,200,300,400,234,500]}
s2 = pd.Series(d)
s2
```

Out[18]:
```
col1    [100, 200, 300, 400, 500, 600]
col2    [100, 200, 300, 400, 234, 500]
dtype: object
```

In [19]:
```
df1 = pd.DataFrame(d)
df1
```

Out[19]:

	col1	col2
0	100	100
1	200	200
2	300	300
3	400	400
4	500	234
5	600	500

In [22]:
```
df = pd.read_csv("/home/labuser/Downloads/insurance.csv")
df
```

Out[22]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 7 columns

In [24]:
```
df.size
```

Out[24]: 9366

In [25]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

In [ ]:

In [26]:
```
df.describe()
```

Out[26]:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

In [28]:
```
df.describe(include='all').T #transpose
```

Out[28]:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
age	1338.0	NaN	NaN	NaN	39.207025	14.04996	18.0	27.0	39.0	51.0	64.0
sex	1338	2	male	676	NaN	NaN	NaN	NaN	NaN	NaN	NaN
bmi	1338.0	NaN	NaN	NaN	30.663397	6.098187	15.96	26.29625	30.4	34.69375	53.13
children	1338.0	NaN	NaN	NaN	1.094918	1.205493	0.0	0.0	1.0	2.0	5.0
smoker	1338	2	no	1064	NaN	NaN	NaN	NaN	NaN	NaN	NaN
region	1338	4	southeast	364	NaN	NaN	NaN	NaN	NaN	NaN	NaN
charges	1338.0	NaN	NaN	NaN	13270.422265	12110.011237	1121.8739	4740.28715	9382.033	16639.912515	63770.42801

In [ ]:

In [32]:
```
#df['age']

#df.age

df.iloc[:,0:1]
```

Out[32]:

	age
0	19
1	18
2	28
3	33
4	32
...	...
1333	50
1334	18
1335	18
1336	21
1337	61

In [34]:
```
df.iloc[:,[0]]
```

Out[34]:

	age
0	19
1	18
2	28
3	33
4	32
...	...
1333	50
1334	18
1335	18
1336	21
1337	61

1338 rows × 1 columns

In [37]:
```
df.iloc[:,[2,5,6,0]]
```

Out[37]:

	bmi	region	charges	age
0	27.900	southwest	16884.92400	19
1	33.770	southeast	1725.55230	18
2	33.000	southeast	4449.46200	28
3	22.705	northwest	21984.47061	33
4	28.880	northwest	3866.85520	32
...	...	...	...	...
1333	30.970	northwest	10600.54830	50
1334	31.920	northeast	2205.98080	18
1335	36.850	southeast	1629.83350	18
1336	25.800	southwest	2007.94500	21
1337	29.070	northwest	29141.36030	61

1338 rows × 4 columns

In [38]:
```
df.iloc[38:42,:]
```

Out[38]:

	age	sex	bmi	children	smoker	region	charges
38	22	male	35.600	0	yes	southwest	35585.57600
39	18	female	26.315	0	no	northeast	2198.18985
40	19	female	28.600	5	no	southwest	4687.79700
41	63	male	28.310	0	no	northwest	13770.09790

In [39]:
```
df.iloc[[100,200,300],:]
```

Out[39]:

	age	sex	bmi	children	smoker	region	charges

In [40]:
```
x = df.iloc[[100,200,300],:]
x
```

Out[40]:

	age	sex	bmi	children	smoker	region	charges
100	41	female	31.60	0	no	southwest	6186.1270
200	19	female	32.11	0	no	northwest	2130.6759
300	36	male	27.55	3	no	northeast	6746.7425

In [50]:
```
#df.drop(columns='smoker','axis=1',inplace=True)
x = df.drop(columns='smoker',axis=1)
x
```

Out[50]:

	age	sex	bmi	children	region	charges
0	19	female	27.900	0	southwest	16884.92400
1	18	male	33.770	1	southeast	1725.55230
2	28	male	33.000	3	southeast	4449.46200
3	33	male	22.705	0	northwest	21984.47061
4	32	male	28.880	0	northwest	3866.85520
...	...	...	...	...	...	...
1333	50	male	30.970	3	northwest	10600.54830
1334	18	female	31.920	0	northeast	2205.98080
1335	18	female	36.850	0	southeast	1629.83350
1336	21	female	25.800	0	southwest	2007.94500
1337	61	female	29.070	0	northwest	29141.36030

1338 rows × 6 columns

In [52]:
```
df['region'].unique()
```

Out[52]:
```
array(['southwest', 'southeast', 'northwest', 'northeast'], dtype=object)
```

In [52]:
```
df['region'].unique()
```

Out[52]:
```
array(['southwest', 'southeast', 'northwest', 'northeast'], dtype=object)
```

In [53]:
```
df['region'].nunique()
```

Out[53]:
```
4
```

In [54]:
```
df['region'].value_counts()
```

Out[54]:
```
southeast    364
southwest    325
northwest    325
northeast    324
Name: region, dtype: int64
```

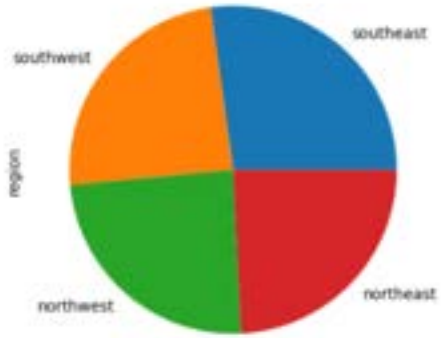
In [ ]:

```
In [55]: df['region'].value_counts().plot(kind='bar')
```

```
Out[55]: <Axes: >
```

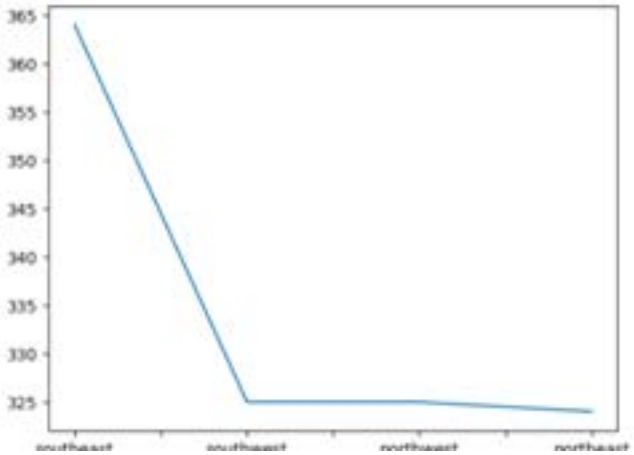

```
In [60]: df['region'].value_counts().plot(kind='pie')
```

```
Out[60]: <Axes: ylabel='region'>
```




```
In [56]: df['region'].value_counts().plot(kind='line')
```

```
Out[56]: <Axes: >
```



In [63]:
```
df.select_dtypes(include=['int64','float64'])
```

Out[63]:

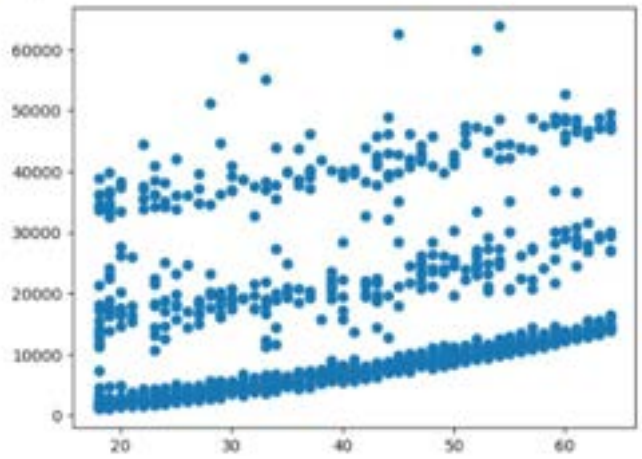
	age	bmi	children	charges
0	19	27.900	0	16884.92400
1	18	33.770	1	1725.55230
2	28	33.000	3	4449.46200
3	33	22.705	0	21984.47061
4	32	28.880	0	3866.85520
...	...	...	...	...
1333	50	30.970	3	10600.54830
1334	18	31.920	0	2205.98080
1335	18	36.850	0	1629.83350
1336	21	25.800	0	2007.94500
...	...	...	...	...

In [65]:
```
import matplotlib.pyplot as plt

plt.plot(df['age'],df['charges'])
```

Out[65]: [<matplotlib.lines.Line2D at 0x7fa802b9f4d0>]

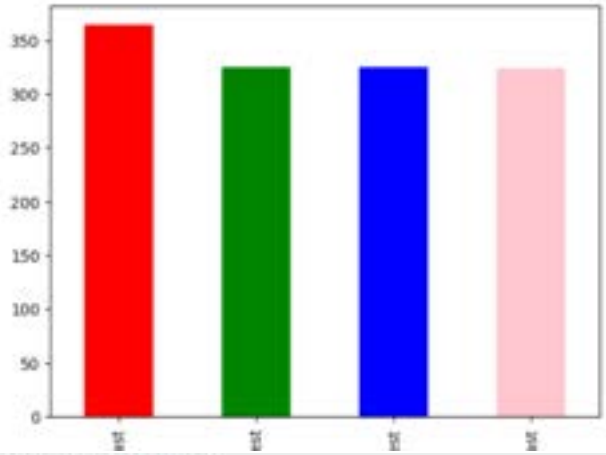
In [67]:
```
plt.scatter(df['age'],df['charges'])
```

Out[67]: <matplotlib.collections.PathCollection at 0x7fa8003f8690>



In [68]:
```
df['region'].value_counts().plot(kind='bar',color=['red','green','blue','pink'])
```

Out[68]:
```
<Axes: >
```



In [69]:
```
plt.hist(df['charges'])
```

Out[69]:
```
(array([536., 398., 129.,  86.,  35.,  59.,  57.,  32.,   2.,   4.]),
 array([ 1121.8739  ,  7386.729311, 13651.584722, 19916.440133,
        26181.295544, 32446.150955, 38711.006366, 44975.861777,
        51240.717188, 57505.572599, 63770.42801 ]),
 <BarContainer object of 10 artists>)
```

In [72]:
```
#filtering of data look for charges only for southwest region

df[df['region'] == "southwest"]
```

Out[72]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.9	0	yes	southwest	16884.92400
12	23	male	34.4	0	no	southwest	1826.84300
15	19	male	24.6	1	no	southwest	1837.23700
18	56	male	40.3	0	no	southwest	10602.38500
19	30	male	35.3	0	yes	southwest	36837.46700
...	...	...	...	...	...	...	...
1316	19	female	20.6	0	no	southwest	1731.67700
1329	52	male	38.6	2	no	southwest	10325.20600
1331	23	female	33.4	0	no	southwest	10795.93733
1332	52	female	44.7	3	no	southwest	11411.68500
1336	21	female	25.8	0	no	southwest	2007.94500

325 rows × 7 columns

In [74]:
```
df[(df['region'] == "northeast") & (df['smoker'] =="no")]
```

Out[74]:

	age	sex	bmi	children	smoker	region	charges
8	37	male	29.830	2	no	northeast	6406.41070
10	25	male	26.220	0	no	northeast	2721.32080
16	52	female	30.780	1	no	northeast	10797.33620
17	23	male	23.845	0	no	northeast	2395.17155
20	60	female	36.005	0	no	northeast	13228.84695
...	...	...	...	...	...	...	...
1318	35	male	39.710	4	no	northeast	19496.71917
1325	61	male	33.535	0	no	northeast	13143.33665
1326	42	female	32.870	0	no	northeast	7050.02130
1328	23	female	24.225	2	no	northeast	22395.74424
1334	18	female	31.920	0	no	northeast	2205.98080

257 rows × 7 columns

In [75]:
```
df[(df['region'] == "northeast") & (df['smoker'] =="no")&(df['sex'] == "female")]
```

Out[75]:

	age	sex	bmi	children	smoker	region	charges
16	52	female	30.780	1	no	northeast	10797.33620
20	60	female	36.005	0	no	northeast	13228.84695
26	63	female	23.085	0	no	northeast	14451.83515
31	18	female	26.315	0	no	northeast	2198.18985
49	24	female	26.600	0	no	northeast	3046.06200
...	...	...	...	...	...	...	...
1286	28	female	17.290	0	no	northeast	3732.62510
1290	38	female	19.950	2	no	northeast	7133.90250
1326	42	female	32.870	0	no	northeast	7050.02130
1328	23	female	24.225	2	no	northeast	22395.74424
1334	18	female	31.920	0	no	northeast	2205.98080

132 rows × 7 columns

In [76]:
```
df['smoker'] = df['smoker'].replace(to_replace=["yes","no"],value=[1,0])
df['smoker']
```

Out[76]:
```
0       1
1       0
2       0
3       0
4       0
       ..
1333    0
1334    0
1335    0
1336    0
1337    1
Name: smoker, Length: 1338, dtype: int64
```

In [77]:
```
len(df[(df['region'] == "northeast") & (df['smoker'] ==0)&(df['sex'] == "female")])
```

Out[77]: 132

In [78]:
```
#sorting value by region
df.sort_values(by='region')
```

Out[79]:

	age	sex	bmi	children	smoker	region	charges
668	62	male	32.015	0	1	northeast	45710.20785
319	32	male	37.335	1	0	northeast	4667.60765
844	53	male	30.496	0	0	northeast	10072.05505
317	54	male	32.775	0	0	northeast	10435.06525
315	52	male	33.250	0	0	northeast	9722.76950
...	...	...	...	...	...	...	...
290	28	female	33.400	0	0	southwest	3172.01800
888	22	male	39.500	0	0	southwest	1682.59700
294	25	male	26.800	3	0	southwest	3806.12700
918	61	female	28.200	0	0	southwest	13041.92100

In [81]:
```
df.sort_values(by='children',ascending=False)
```

Out[81]:

	age	sex	bmi	children	smoker	region	charges
1272	43	male	25.520	5	0	southeast	14478.33015
1130	39	female	23.870	5	0	southeast	8582.30230
1116	41	male	29.640	5	0	northeast	9222.40260
568	49	female	31.900	5	0	southwest	11552.90400
1245	28	male	24.300	5	0	southwest	5615.36900
...	...	...	...	...	...	...	...
618	19	female	33.110	0	1	southeast	34439.85590
619	55	female	37.100	0	0	southwest	10713.64400
623	18	male	33.535	0	1	northeast	34617.84065
624	59	male	28.785	0	0	northwest	12129.61415

In [83]:
```
df.groupby('region').mean()
```

```
/tmp/ipykernel_6447/1373985188.py:1: FutureWarning: The default value of numeric_only in DataFrameGroupBy.mean is d
eprecated. In a future version, numeric_only will default to False. Either specify numeric_only or select only colu
mns which should be valid for the function.
  df.groupby('region').mean()
```

Out[83]:

region	age	bmi	children	smoker	charges
northeast	39.268519	29.173503	1.046296	0.206790	13406.384516
northwest	39.196923	29.199785	1.147692	0.178462	12417.575374
southeast	38.939560	33.355989	1.049451	0.250000	14735.411438
southwest	39.455385	30.596615	1.141538	0.178462	12346.937377

In [84]:
```
df.groupby(['region','sex']).mean()
```

Out[84]:

region	sex	age	bmi	children	smoker	charges
northeast	**female**	39.639752	29.324317	1.006211	0.180124	12953.203151
	male	38.901840	29.024540	1.085890	0.233129	13854.005374
northwest	**female**	39.591463	29.277857	1.115854	0.176829	12479.870397
	male	38.795031	29.120155	1.180124	0.180124	12354.119575
southeast	**female**	39.108571	32.671257	1.051429	0.205714	13489.689243
	male	38.783069	33.990000	1.047619	0.291005	15879.617173
southwest	**female**	39.703704	30.060494	1.123457	0.129630	11274.411264
	male	39.208589	31.129448	1.159509	0.226994	13412.883576

```
In [85]: df.groupby(['region','sex']).aggregate({'charges' : ['min','max','count','mean']})
```

Out[85]:

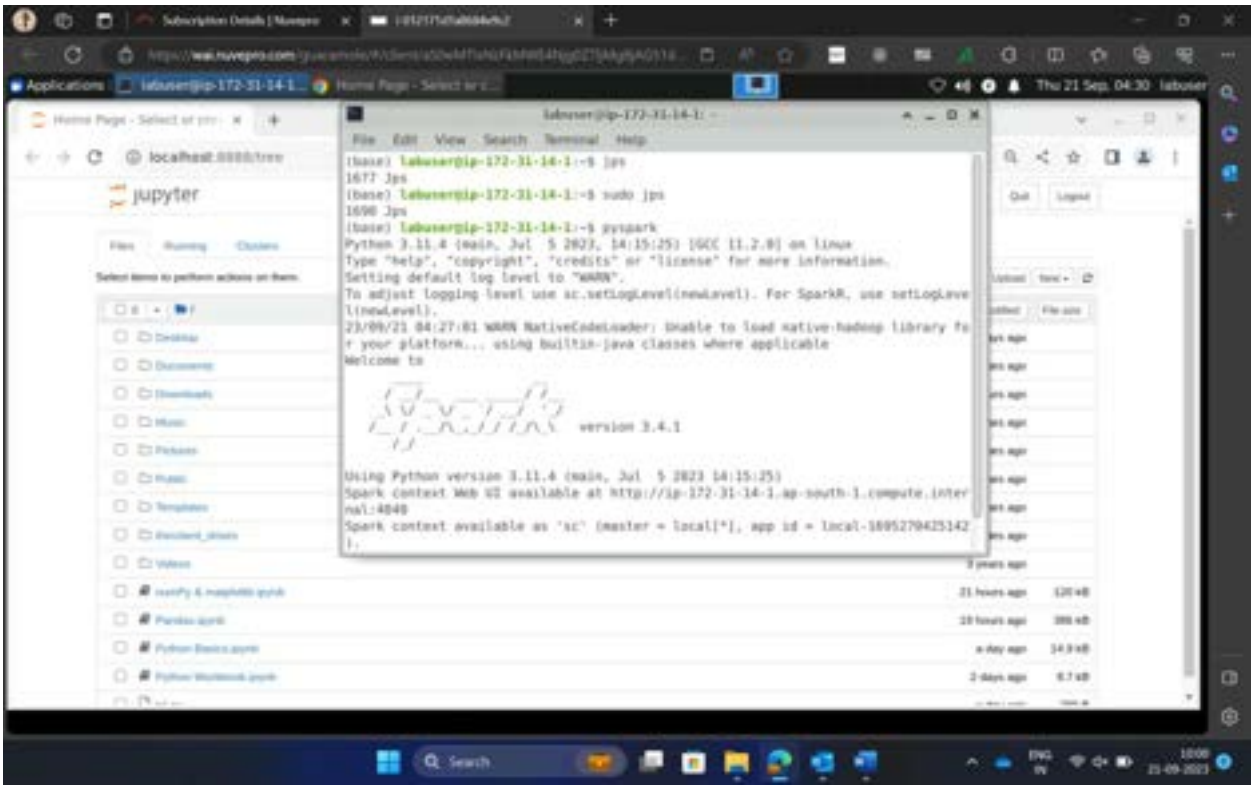
		charges			
		min	max	count	mean
region	sex				
northeast	female	2196.47320	58571.07448	161	12953.203151
	male	1694.79640	48549.17835	163	13854.005374
northwest	female	2117.33885	55135.40209	164	12479.870397
	male	1621.34020	60021.39897	161	12354.119575
southeast	female	1607.51010	63770.42801	175	13499.669243
	male	1121.87390	62592.87309	189	15679.617173
southwest	female	1727.78500	48824.45000	162	11274.411264
	male	1241.56500	52590.82939	163	13412.883576

```
In [86]: df.groupby(['region','sex']).aggregate({'charges' : ['min','max','count','mean'],'bmi':['mean']})
```

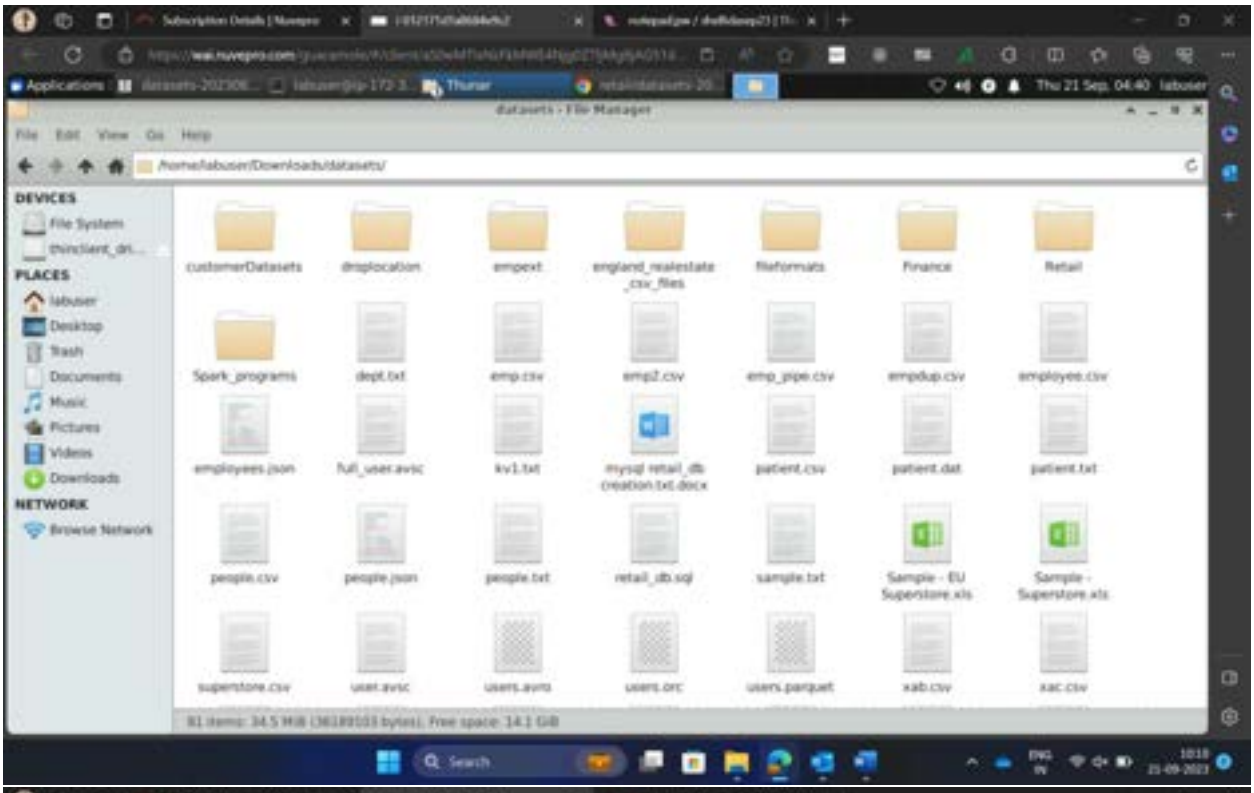
Out[86]:

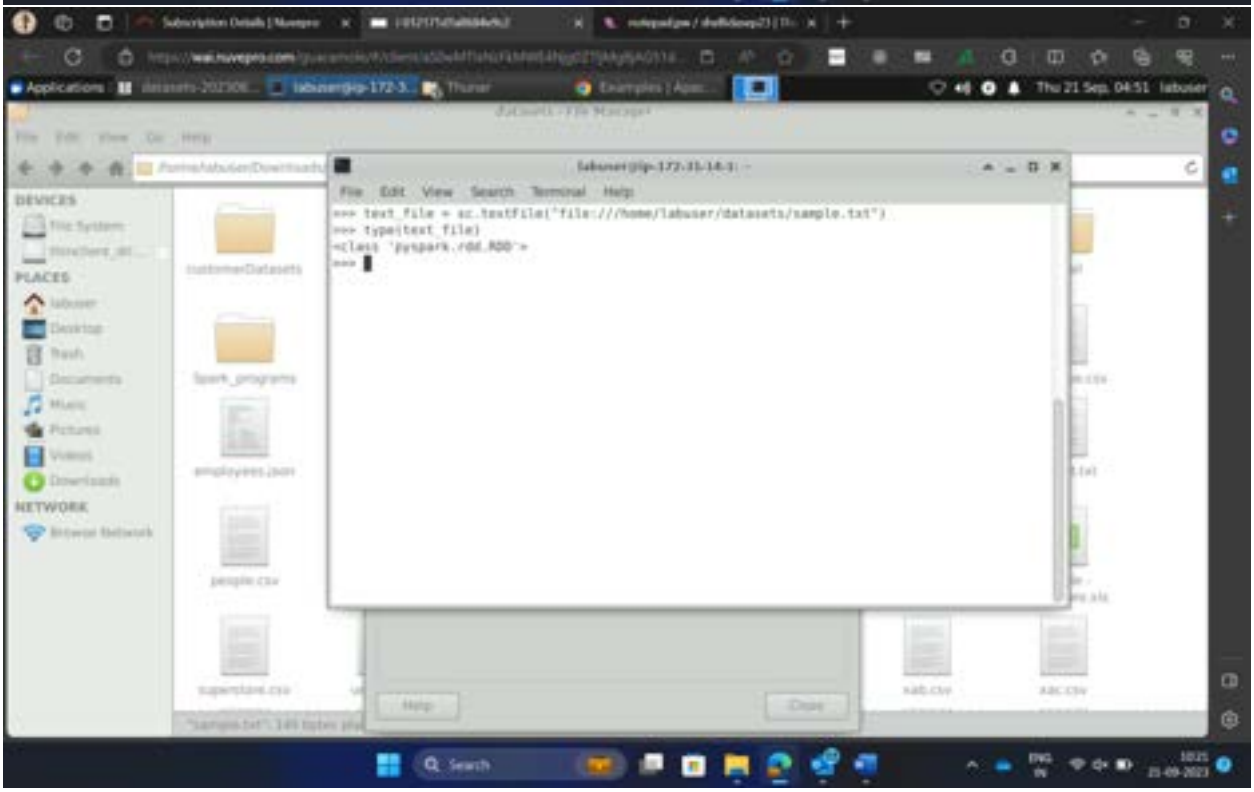
		charges				bmi
		min	max	count	mean	mean
region	sex					
northeast	female	2196.47320	58571.07448	161	12953.203151	29.324317
	male	1694.79640	48549.17835	163	13854.005374	29.024540
northwest	female	2117.33885	55135.40209	164	12479.870397	29.277957
	male	1621.34020	60021.39897	161	12354.119575	29.120155
southeast	female	1607.51010	63770.42801	175	13499.669243	32.671257
	male	1121.87390	62592.87309	189	15679.617173	33.990000
southwest	female	1727.78500	48824.45000	162	11274.411264	30.060494
	male	1241.56500	52590.82939	163	13412.883576	31.129448

Day 16



[notepad.pw / shellidasep23 | The napkin of the internet.]

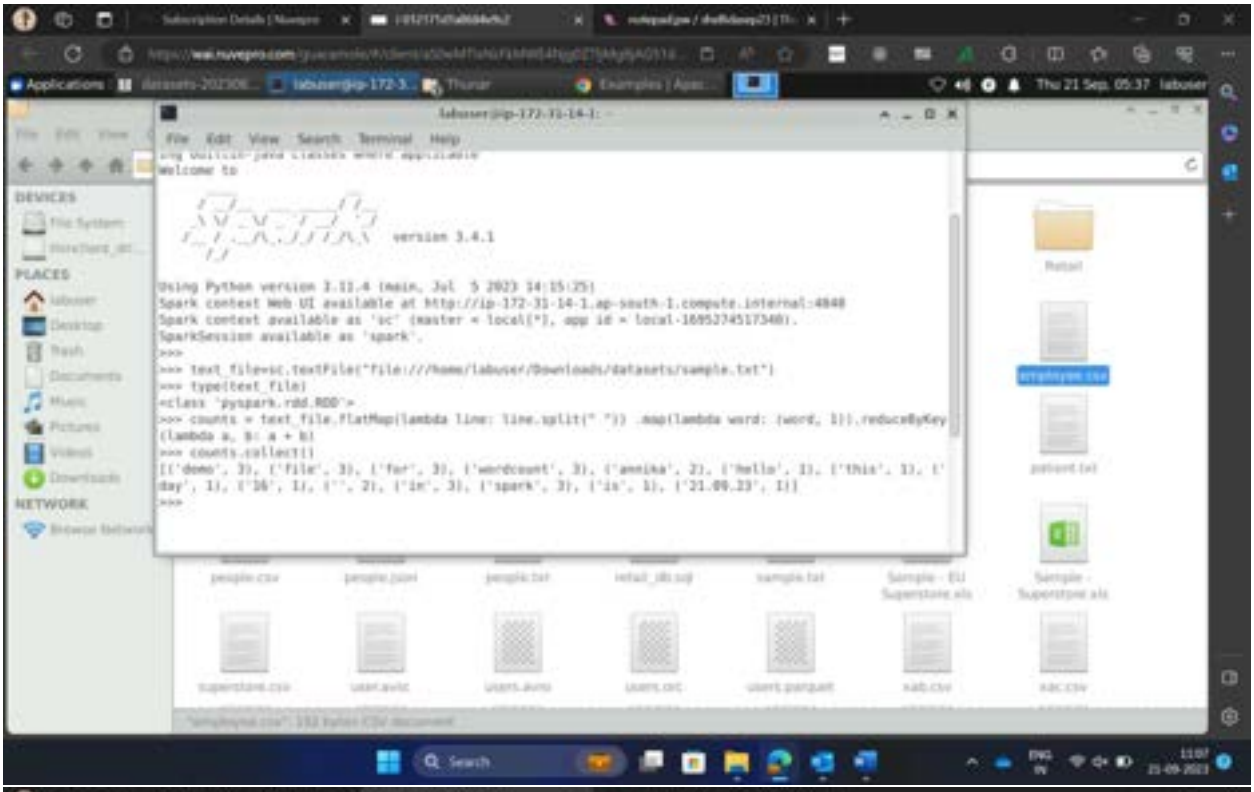


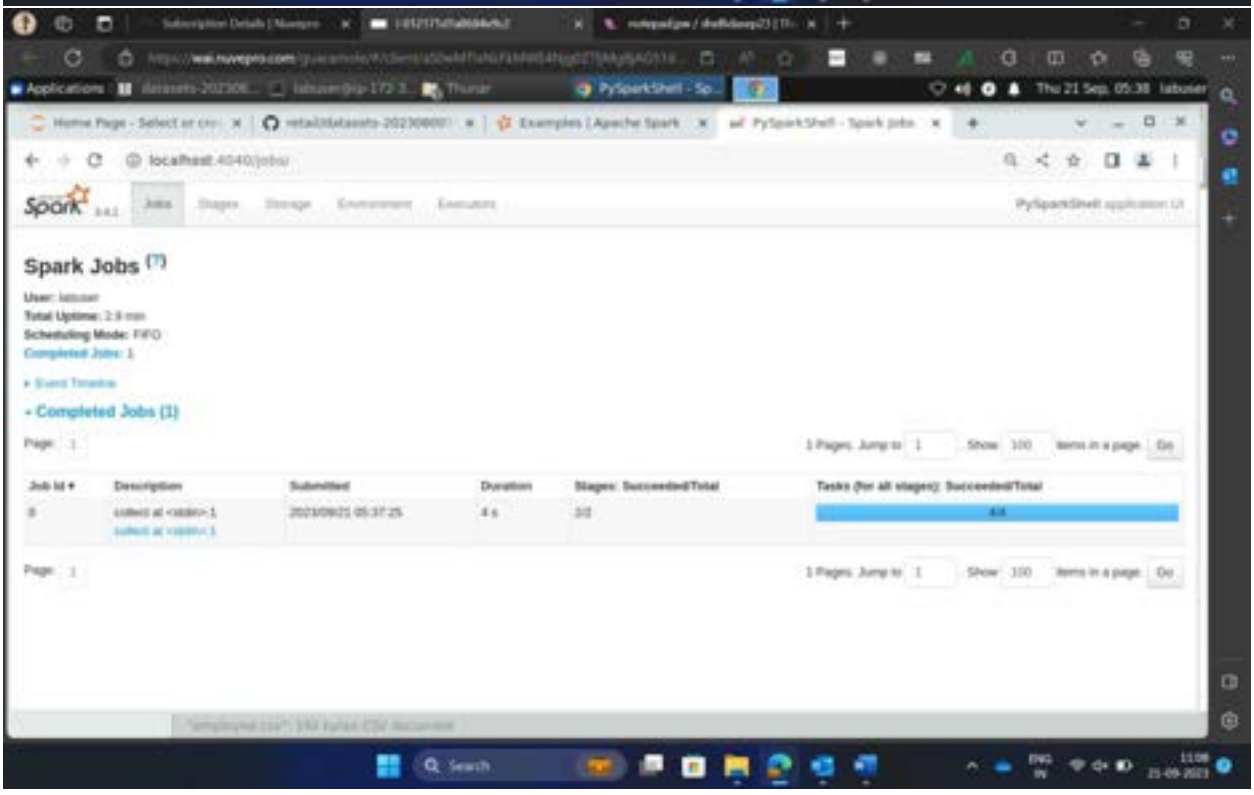


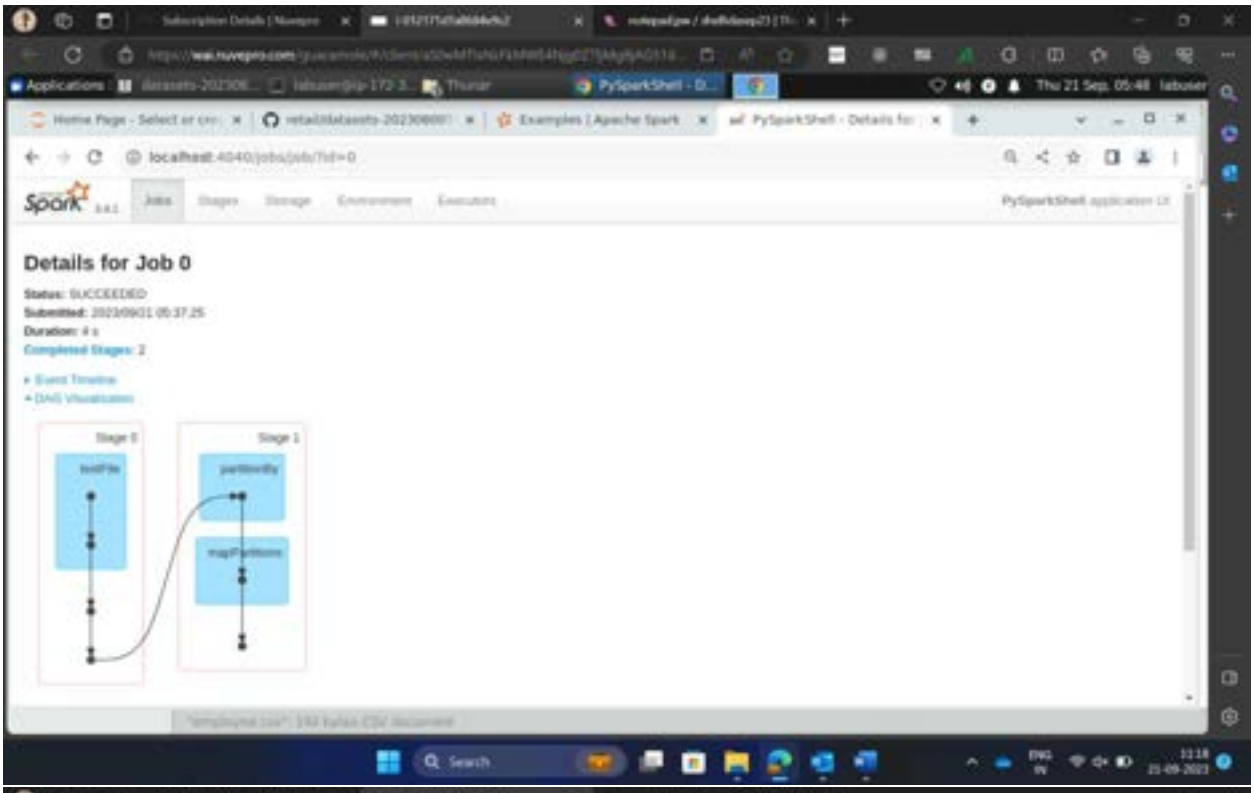
counts = text_file.flatMap(lambda line: line.split(" ")) .map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)

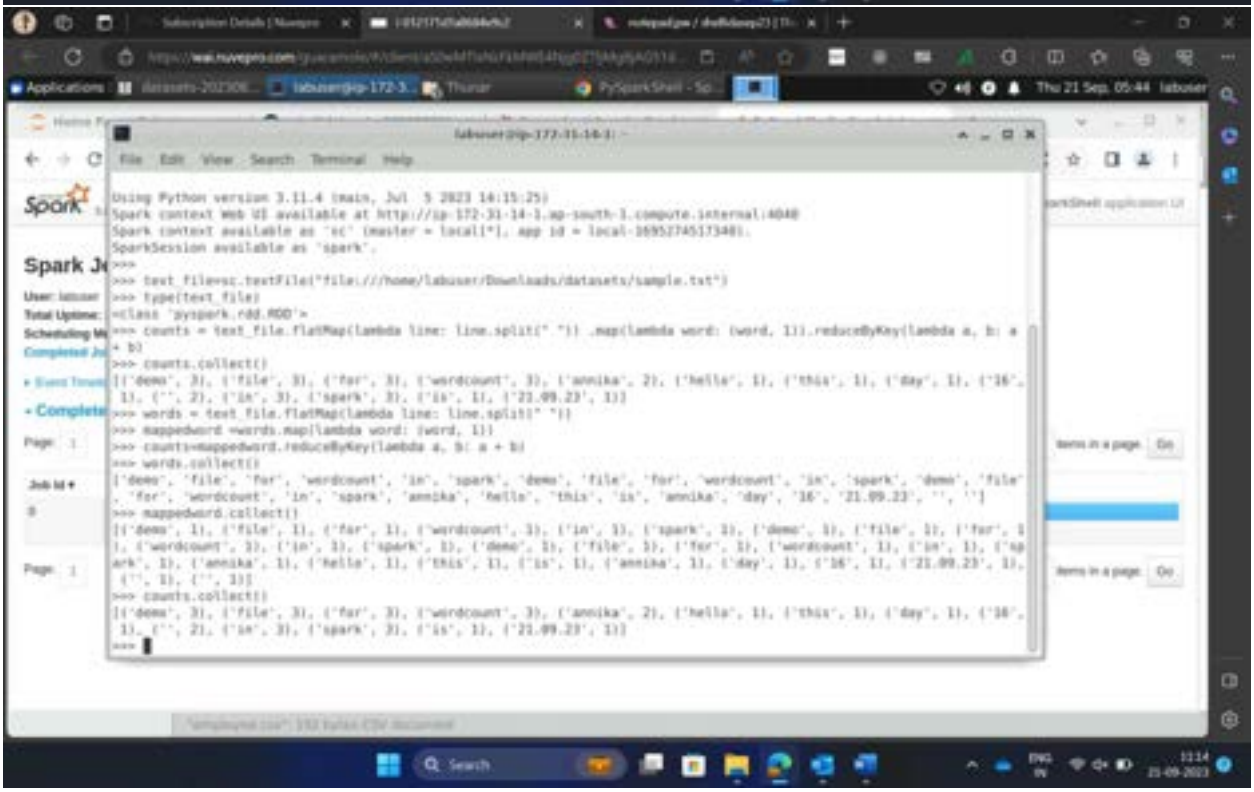
counts = text_file.flatMap(lambda line: line.split(" ")) .map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)

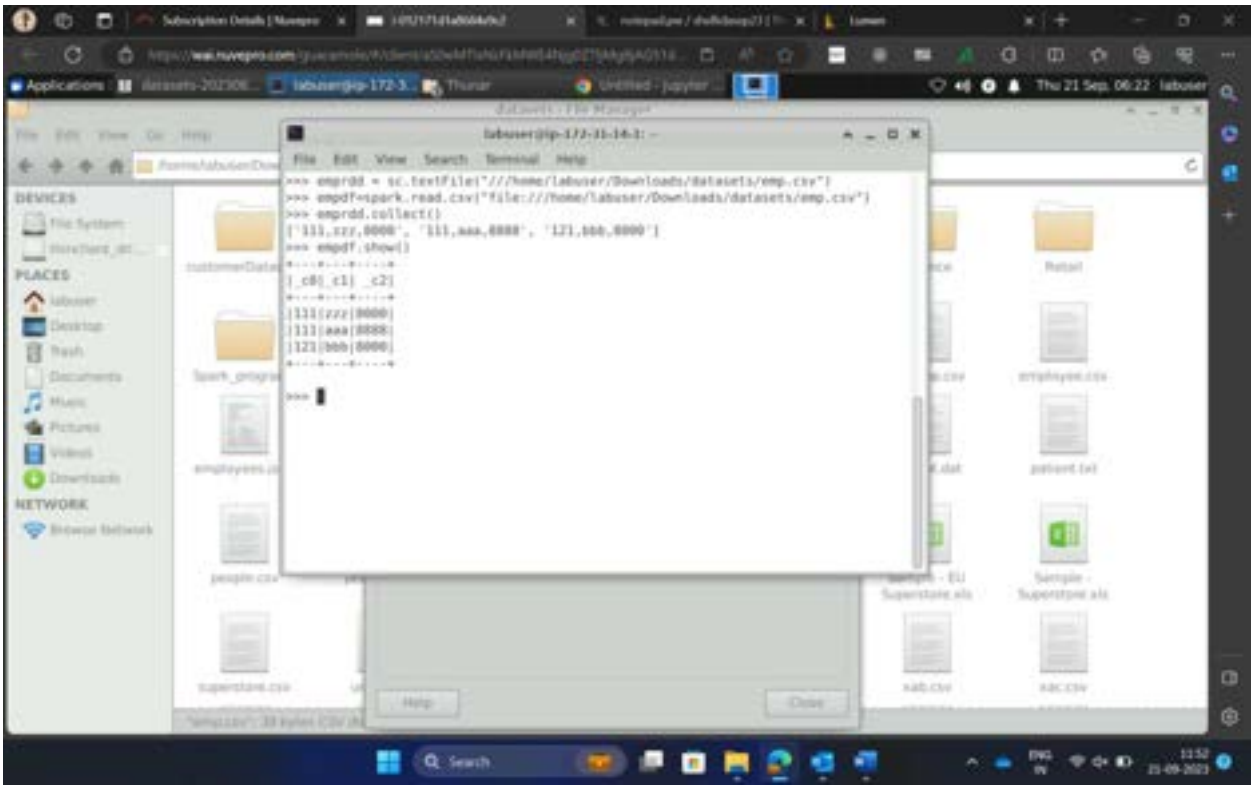
counts.collect()






Details for Job 0
Status: SUCCEEDED
Submitted: 2023/09/01 05:37:25
Duration: 4 s
Completed Stages: 2
Stage 0
textFile
Stage 1
partitionBy
mapPartitions


Using Python version 3.11.4 (main, Jul 5 2023 14:15:25)
Spark context Web UI available at http://ip-172-31-14-1.ap-south-1.compute.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1695274517340).
SparkSession available as 'spark'.
>>> test_file=sc.textFile("file:///home/labuser/Downloads/datasets/sample.txt")
>>> type(test_file)
<class 'pyspark.rdd.RDD'>
>>> counts = test_file.flatMap(lambda line: line.split(" ")) .map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
>>> counts.collect()
[('demo', 3), ('file', 3), ('for', 3), ('wordcount', 3), ('annika', 2), ('hello', 1), ('this', 1), ('day', 1), ('16', 1), ('', 2), ('in', 3), ('spark', 3), ('is', 1), ('21.09.23', 1)]
>>> words = test_file.flatMap(lambda line: line.split(" "))
>>> mappedword =words.map(lambda word: (word, 1))
>>> counts=mappedword.reduceByKey(lambda a, b: a + b)
>>> words.collect()
['demo', 'file', 'for', 'wordcount', 'in', 'spark', 'demo', 'file', 'for', 'wordcount', 'in', 'spark', 'demo', 'file', 'for', 'wordcount', 'in', 'spark', 'annika', 'hello', 'this', 'is', 'annika', 'day', '16', '21.09.23', '', '']
>>> mappedword.collect()
[('demo', 1), ('file', 1), ('for', 1), ('wordcount', 1), ('in', 1), ('spark', 1), ('demo', 1), ('file', 1), ('for', 1), ('wordcount', 1), ('in', 1), ('spark', 1), ('demo', 1), ('file', 1), ('for', 1), ('wordcount', 1), ('in', 1), ('spark', 1), ('annika', 1), ('hello', 1), ('this', 1), ('is', 1), ('annika', 1), ('day', 1), ('16', 1), ('21.09.23', 1), ('', 1), ('', 1)]
>>> counts.collect()
[('demo', 3), ('file', 3), ('for', 3), ('wordcount', 3), ('annika', 2), ('hello', 1), ('this', 1), ('day', 1), ('16', 1), ('', 2), ('in', 3), ('spark', 3), ('is', 1), ('21.09.23', 1)]
>>>




```
In [2]: import findspark

In [3]: findspark.init()

In [4]: findspark.find()
Out[4]: '/opt/anaconda3/lib/python3.11/site-packages/pyspark'

In [5]: from pyspark.sql import SparkSession

In [6]: spark = SparkSession \
            .builder \
            .appName("Python Spark SQL basic example") \
            .config("spark.some.config.option", "some-value") \
            .getOrCreate()

Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/09/21 06:35:48 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
23/09/21 06:35:50 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.

In [9]: empdf=spark.read.csv("file:///home/labuser/Downloads/datasets/emp.csv")
```

```
In [9]: empdf=spark.read.csv("file:///home/labuser/Downloads/datasets/emp.csv")

In [10]: empdf.show()
+---+---+----+
|_c0|_c1| _c2|
+---+---+----+
|111|zzz|8000|
|111|aaa|8888|
|121|bbb|8000|
+---+---+----+
```

In [11]:
```
salesdf=spark.read.option("header","true").csv("/home/labuser/Downloads/datasets/superstore.csv")
```

In [12]:
```
salesdf.show()
```

```
+-----+--------------+----------+----------+--------------+----------+---------------+-----------+-------------+-
-------------+----------+----------+-----+--------+----------+--------------+---------+-----------+---------+
-----------+---------------+--------+--------+----------+--------------+-------------+
|   ID|       OrderID| OrderDate|  ShipDate|      ShipMode|CustomerID|   CustomerName|    Segment|         City|
State|      Country|PostalCode|Market|      Region|       ProductID|      Category|Sub-Category|          ProductNa
me|          Sales|Quantity|Discount|    Profit|ShippingCost|OrderPriority|
+-----+--------------+----------+----------+--------------+----------+---------------+-----------+-------------+-
-------------+----------+----------+-----+--------+----------+--------------+---------+-----------+---------+
-----------+---------------+--------+--------+----------+--------------+-------------+
|32298| CA-2012-124891|31/07/2012|31/07/2012|      Same Day|  RH-19495|     Rick Hansen|   Consumer|New York City|
New York|United States|     10024|    US|      East| TEC-AC-10003033|    Technology| Accessories|Plantronics CS5
10...|        2309.65|       7|       0|  762.1845|     933.57 |     Critical|
|26341|  IN-2013-77878|05/02/2013|07/02/2013|  Second Class|  JR-16210|   Justin Ritter|  Corporate|   Wollongong|N
ew South Wales|    Australia|      null|  APAC|     Oceania| FUR-CH-10003950|      Furniture|      Chairs|Novimex E
xecutive...|            Black|3709.395|       9|       0.1|    -288.765|       923.63 |
|25338|  IN-2013-71249|17/10/2013|18/10/2013|   First Class|  CR-12730|    Craig Reiter|   Consumer|     Brisbane|
Queensland|    Australia|      null|  APAC|     Oceania| TEC-PH-10004664|    Technology|      Phones|   Nokia Smar
t Phone|  with Caller ID|5175.171|       9|       0.1|     919.971|      915.49 |
|13524|ES-2013-1579342|28/01/2013|30/01/2013|   First Class|  KM-16375|Katherine Murray|Home Office|       Berlin|
Berlin|      Germany|      null|    EU|     Central| TEC-PH-10004583|    Technology|      Phones|Motorola Smart Ph
one|       Cordless| 2892.51|       5|       0.1|      -96.54|      910.16 |
|47221|   SG-2013-4320|05/11/2013|06/11/2013|      Same Day|  RH-9495|     Rick Hansen|   Consumer|        Dakar|
```

In [13]:
```
employeedf =spark.read.option("header","true").csv("/home/labuser/Downloads/datasets/employee.csv")
```

In [14]:
```
employeedf.show()
```

```
+-----+-----+----+
|empno|ename| sal|
+-----+-----+----+
|  111|  zzz|8000|
|  111|  aaa|8888|
|  121|  bbb|8000|
| NULL|  ccc|9000|
|false|  ddd|6000|
|  555|  eee|7000|
|  765| null|null|
|  666|  fff|8090|
|  aaa| null|null|
| True| null|null|
| true| null|null|
|    a|    b|   c|
|    x| NULL|NULL|
| 1000| null|null|
|  222| NULL|NULL|
+-----+-----+----+
```

In [23]:
```
from pyspark.sql.types import StructType,StructField, StringType,IntegerType,DoubleType
```

In [24]:
```
#if we want any exact datatype for the data
EmpSchema = StructType([
    StructField('Empno', IntegerType(), True),
    StructField('Empname', StringType(), True),
     StructField('salary', DoubleType(), True)
])
```

In [25]:
```
employeedf =spark.read.option("header","true").schema(EmpSchema).csv("/home/labuser/Downloads/datasets/employee.csv"
```

In [29]:
```
employeedf.printSchema()
```

```
root
 |-- Empno: integer (nullable = true)
 |-- Empname: string (nullable = true)
 |-- salary: double (nullable = true)
```

In [ ]:

```
In [38]: #3 modes to correct data
         #1. Permissive
         #2.malformed
         #it will drop all bad rows
         employeedf_malformed=spark.read.option("header","true").option("mode","dropmalformed").schema(EmpSchema).csv("/home/
```

```
In [39]: employeedf_malformed.show()
```

```
+-----+-------+------+
|Empno|Empname|salary|
+-----+-------+------+
|  111|    zzz|8000.0|
|  111|    aaa|8088.0|
|  121|    bbb|8000.0|
|  555|    eee|7000.0|
|  666|    fff|8090.0|
+-----+-------+------+
```

jupyter PySpark (Dataframe Operations) Last Checkpoint: 3 hours ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

```
 Schema: Empno, Empname, salary
Expected: Empname but found: ename
CSV file: file:///home/labuser/Downloads/datasets/employee.csv
```

```
In [42]: #2. mode - failfast (purely for sensitive data)
         employeedf_failfast=spark.read.option("header","true").option("mode","failfast").schema(EmpSchema).csv("/home/labuse
```

```
In [43]: employeedf_failfast.show()
```

```
23/09/21 09:15:17 WARN CSVHeaderChecker: CSV header does not conform to the schema.
 Header: empno, ename, sal
 Schema: Empno, Empname, salary
Expected: Empname but found: ename
CSV file: file:///home/labuser/Downloads/datasets/employee.csv
23/09/21 09:15:17 ERROR Executor: Exception in task 0.0 in stage 20.0 (TID 22)
org.apache.spark.SparkException: [MALFORMED_RECORD_IN_PARSING] Malformed records are detected in record parsing:
[null,ccc,9000.0].
Parse Mode: FAILFAST. To process malformed records as null result, try setting the option 'mode' as 'PERMISSIVE'.
        at org.apache.spark.sql.errors.QueryExecutionErrors$.malformedRecordsDetectedInRecordParsingError(QueryEx
ecutionErrors.scala:1764)
        at org.apache.spark.sql.catalyst.util.FailureSafeParser.parse(FailureSafeParser.scala:69)
        at org.apache.spark.sql.catalyst.csv.UnivocityParser$.$anonfun$parseIterator$2(UnivocityParser.scala:456)
        at scala.collection.Iterator$$anon$11.nextCur(Iterator.scala:486)
        at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:492)
        at scala.collection.Iterator$$anon$10.hasNext(Iterator.scala:460)
        at org.apache.spark.sql.execution.datasources.FileScanRDD$$anon$1.hasNext(FileScanRDD.scala:125)
```

```
In [44]: salesdf.count() #to see no of records in a dataframe
Out[44]: 51290

In [45]: salesdf.head() #to see first 5 records
Out[45]: Row(ID=32298, OrderID='CA-2012-124891', OrderDate='31/07/2012', ShipDate='31/07/2012', ShipMode='Same Day', Custome
         rID='RH-19495', CustomerName='Rick Hansen', Segment='Consumer', City='New York City', State='New York', Country='Un
         ited States', PostalCode=10024, Market='US', Region='East', ProductID='TEC-AC-10003033', Category='Technology', Sub
         -Category='Accessories', ProductName='Plantronics CS510 - Over-the-Head monaural Wireless Headset System', Sales='2
         309.65', Quantity='7', Discount='0', Profit='762.1845', ShippingCost=' 933.57 ', OrderPriority='Critical')

In [46]: #to see the no of partitions in a dataframe'
         salesdf.rdd.getNumPartitions()
Out[46]: 2
```


```
In [46]: #to see the no of partitions in a dataframe'
         salesdf.rdd.getNumPartitions()
Out[46]: 2
```

In [47]:
```
#
salesdf.show(2)
```

```
+-----+--------------+----------+----------+------------+----------+-------------+---------+-------------+---------
------+------------+----------+------+-------+--------------+----------+-----------+------------------+-------
+--------+--------+--------+-----------+-------------+
|   ID|       OrderID| OrderDate|  ShipDate|    ShipMode|CustomerID| CustomerName|  Segment|         City|
State|      Country|PostalCode|Market| Region|     ProductID|  Category|Sub-Category|       ProductName|  Sales|
Quantity|Discount|  Profit|ShippingCost|OrderPriority|
+-----+--------------+----------+----------+------------+----------+-------------+---------+-------------+---------
------+------------+----------+------+-------+--------------+----------+-----------+------------------+-------
+--------+--------+--------+-----------+-------------+
|32298|CA-2012-124891|31/07/2012|31/07/2012|    Same Day|  RH-19495|  Rick Hansen| Consumer|New York City|       Ne
w York|United States|     10024|    US|   East|TEC-AC-10003033|Technology| Accessories|Plantronics CS510...|2309.65
|       7|       0|762.1845|     933.57 |     Critical|
|26341| IN-2013-77878|05/02/2013|07/02/2013|Second Class|  JR-16210|Justin Ritter|Corporate|   Wollongong|New South
Wales|    Australia|      null|  APAC|Oceania|FUR-CH-10003950| Furniture|      Chairs|Novimex Executive...|  Black|
3709.395|       9|     0.1|    -288.765|      923.63 |
+-----+--------------+----------+----------+------------+----------+-------------+---------+-------------+---------
------+------------+----------+------+-------+--------------+----------+-----------+------------------+-------
+--------+--------+--------+-----------+-------------+
only showing top 2 rows
```

In [48]:
```
#to see only the first row of the dataframe
salesdf.first()
```

In [48]:
```
#to see only the first row of the dataframe
salesdf.first()
```

Out[48]:
```
Row(ID=32298, OrderID='CA-2012-124891', OrderDate='31/07/2012', ShipDate='31/07/2012', ShipMode='Same Day', Custome
rID='RH-19495', CustomerName='Rick Hansen', Segment='Consumer', City='New York City', State='New York', Country='Un
ited States', PostalCode=10024, Market='US', Region='East', ProductID='TEC-AC-10003033', Category='Technology', Sub
-Category='Accessories', ProductName='Plantronics CS510 - Over-the-Head monaural Wireless Headset System', Sales='2
309.65', Quantity='7', Discount='0', Profit='762.1845', ShippingCost=' 933.57 ', OrderPriority='Critical')
```

In [49]:
```
#taking 2 records
salesdf.take(2)
```

Out[49]:
```
[Row(ID=32298, OrderID='CA-2012-124891', OrderDate='31/07/2012', ShipDate='31/07/2012', ShipMode='Same Day', Custom
erID='RH-19495', CustomerName='Rick Hansen', Segment='Consumer', City='New York City', State='New York', Country='U
nited States', PostalCode=10024, Market='US', Region='East', ProductID='TEC-AC-10003033', Category='Technology', Su
b-Category='Accessories', ProductName='Plantronics CS510 - Over-the-Head monaural Wireless Headset System', Sales
='2309.65', Quantity='7', Discount='0', Profit='762.1845', ShippingCost=' 933.57 ', OrderPriority='Critical'),
 Row(ID=26341, OrderID='IN-2013-77878', OrderDate='05/02/2013', ShipDate='07/02/2013', ShipMode='Second Class', Cus
tomerID='JR-16210', CustomerName='Justin Ritter', Segment='Corporate', City='Wollongong', State='New South Wales',
Country='Australia', PostalCode=None, Market='APAC', Region='Oceania', ProductID='FUR-CH-10003950', Category='Furni
ture', Sub-Category='Chairs', ProductName='Novimex Executive Leather Armchair', Sales=' Black', Quantity='3709.39
5', Discount='9', Profit='0.1', ShippingCost='-288.765', OrderPriority=' 923.63 ')]
```

In [ ]:

In [50]:
```
#creating subsets of the available data
#select is used for cols like in sql
salesdf.select("country","state").show()
```

```
+-------------+---------------+
|      country|          state|
+-------------+---------------+
|United States|       New York|
|    Australia|New South Wales|
|    Australia|     Queensland|
|      Germany|         Berlin|
|      Senegal|          Dakar|
|    Australia|New South Wales|
|  New Zealand|     Wellington|
|  New Zealand|       Waikato|
|United States|     California|
|United States| North Carolina|
|United States|       Virginia|
|  Afghanistan|          Kabul|
| Saudi Arabia|          Jizan|
|       Brazil|         Parana|
|        China|   Heilongjiang|
|       France| Ile-de-France|
|United States|       Kentucky|
|        Italy|        Tuscany|
|    Australia|     Queensland|
|     Tanzania|         Kigoma|
+-------------+---------------+
only showing top 20 rows
```

In [51]:
```
salesdf.select("country","state").filter("country='India'").show()
```

```
+-------+--------------+
|country|         state|
+-------+--------------+
|  India|       Gujarat|
|  India|       Haryana|
|  India|        Kerala|
|  India|     Jharkhand|
|  India|Madhya Pradesh|
|  India|         Delhi|
|  India|   Uttarakhand|
|  India|   West Bengal|
|  India|         Delhi|
|  India|       Gujarat|
|  India|         Bihar|
|  India|       Haryana|
|  India|     Jharkhand|
|  India|        Kerala|
|  India|Madhya Pradesh|
|  India|         Bihar|
|  India|        Kerala|
|  India| Uttar Pradesh|
|  India|   Uttarakhand|
|  India|  Chhattisgarh|
+-------+--------------+
```

In [54]:
```
#to get only the distinct data
salesdf.select("country","state").filter("country='India'").distinct().show()
```

```
+-------+-----------------+
|country|            state|
+-------+-----------------+
|  India|      Maharashtra|
|  India|    Uttar Pradesh|
|  India|      Uttarakhand|
|  India|           Kerala|
|  India|     Chhattisgarh|
|  India|       Tamil Nadu|
|  India|            Assam|
|  India|        Telangana|
|  India|       Chandigarh|
|  India|Jammu and Kashmir|
|  India|           Manipur|
|  India|            Odisha|
|  India|   Andhra Pradesh|
|  India|           Haryana|
|  India|           Tripura|
|  India|            Bihar|
|  India|        Rajasthan|
|  India|           Punjab|
|  India|          Gujarat|
|  India|      West Bengal|
+-------+-----------------+
only showing top 20 rows
```

In [56]:
```
#dropDuplicates() can also be used
salesdf.select("country","state").filter("country='India'").dropDuplicates().show()
```

```
+-------+-----------------+
|country|            state|
+-------+-----------------+
|  India|      Maharashtra|
|  India|    Uttar Pradesh|
|  India|      Uttarakhand|
|  India|           Kerala|
|  India|     Chhattisgarh|
|  India|       Tamil Nadu|
|  India|            Assam|
|  India|        Telangana|
|  India|       Chandigarh|
|  India|Jammu and Kashmir|
|  India|           Manipur|
|  India|            Odisha|
|  India|   Andhra Pradesh|
|  India|           Haryana|
|  India|           Tripura|
|  India|            Bihar|
|  India|        Rajasthan|
|  India|           Punjab|
|  India|          Gujarat|
|  India|      West Bengal|
+-------+-----------------+
only showing top 20 rows
```

In [58]:
```
#3 dimensions and 1 measure
salesdf.select("country","state","category","profit").filter("country='India'").dropDuplicates().show()
```

```
+-------+----------------+---------------+------+
|country|           state|       category|profit|
+-------+----------------+---------------+------+
|  India|          Kerala|Office Supplies|     0|
|  India|         Haryana|Office Supplies|     0|
|  India|    Chhattisgarh|Office Supplies|     0|
|  India|  Madhya Pradesh|     Technology|     0|
|  India|       Telangana|     Technology|     0|
|  India|       Rajasthan|     Technology|     0|
|  India|           Delhi|      Furniture|     0|
|  India|Jammu and Kashmir|      Furniture|     0|
|  India|           Assam|Office Supplies|     0|
|  India|          Punjab|Office Supplies|     0|
|  India|          Odisha|      Furniture|     0|
|  India|      Tamil Nadu|Office Supplies|     0|
|  India|     Uttarakhand|Office Supplies|     0|
|  India|      Puducherry|Office Supplies|     0|
|  India|         Gujarat|Office Supplies|     0|
|  India|         Manipur|      Furniture|     0|
|  India|         Tripura|      Furniture|     0|
|  India|     Maharashtra|     Technology|   0.5|
|  India|           Assam|      Furniture|     0|
|  India|     Maharashtra|     Technology|     0|
+-------+----------------+---------------+------+
only showing top 20 rows
```

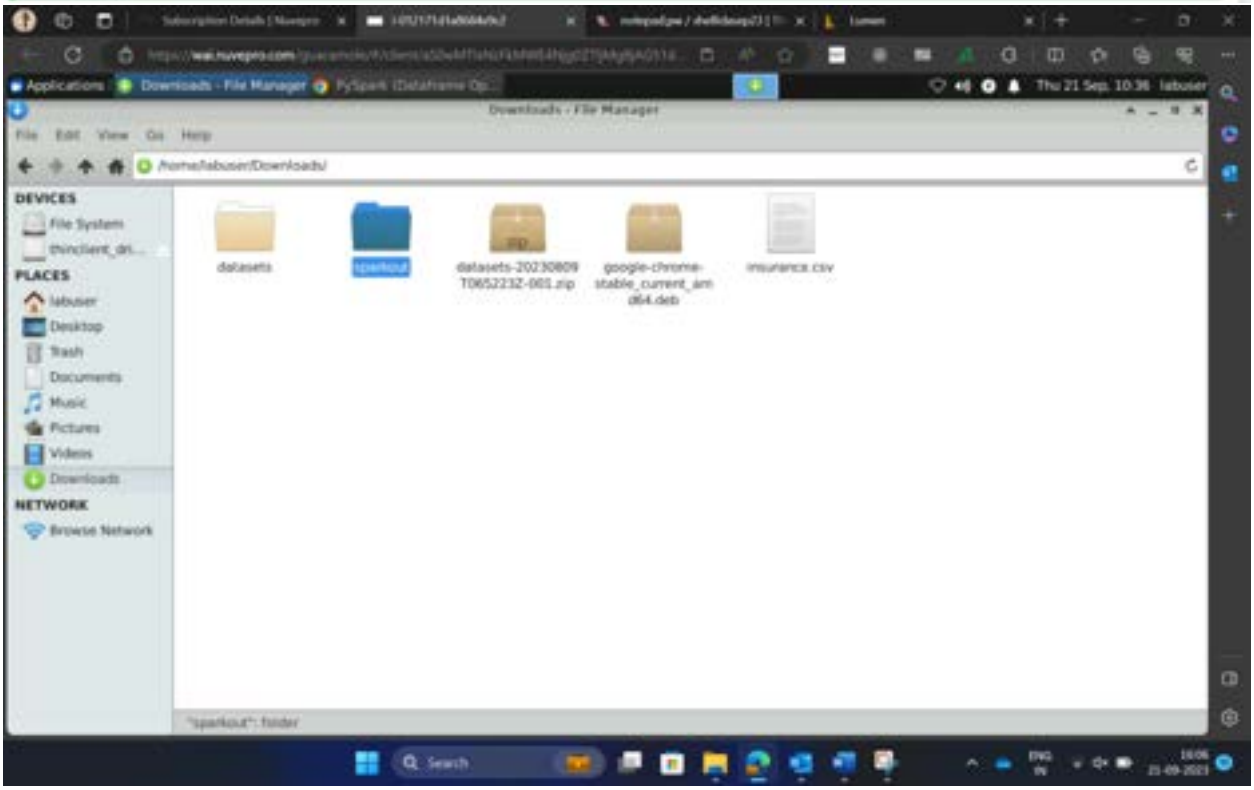
In [62]:
```
#3 dimensions and 1 measure
sales_india_df = salesdf.select("country","state","category","profit").filter("country='India'").dropDuplicates()
```

In [63]:
```
sales_india_df
```

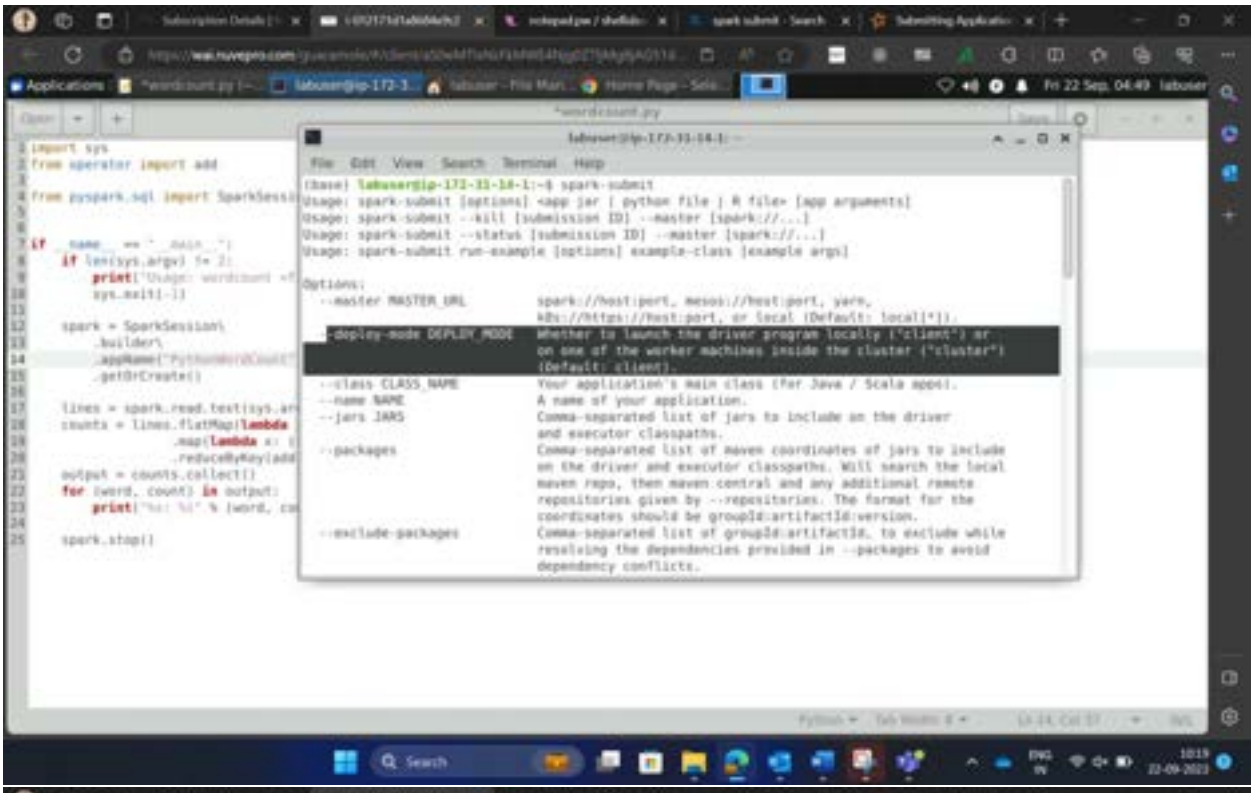
Out[63]:
```
DataFrame[country: string, state: string, category: string, profit: string]
```

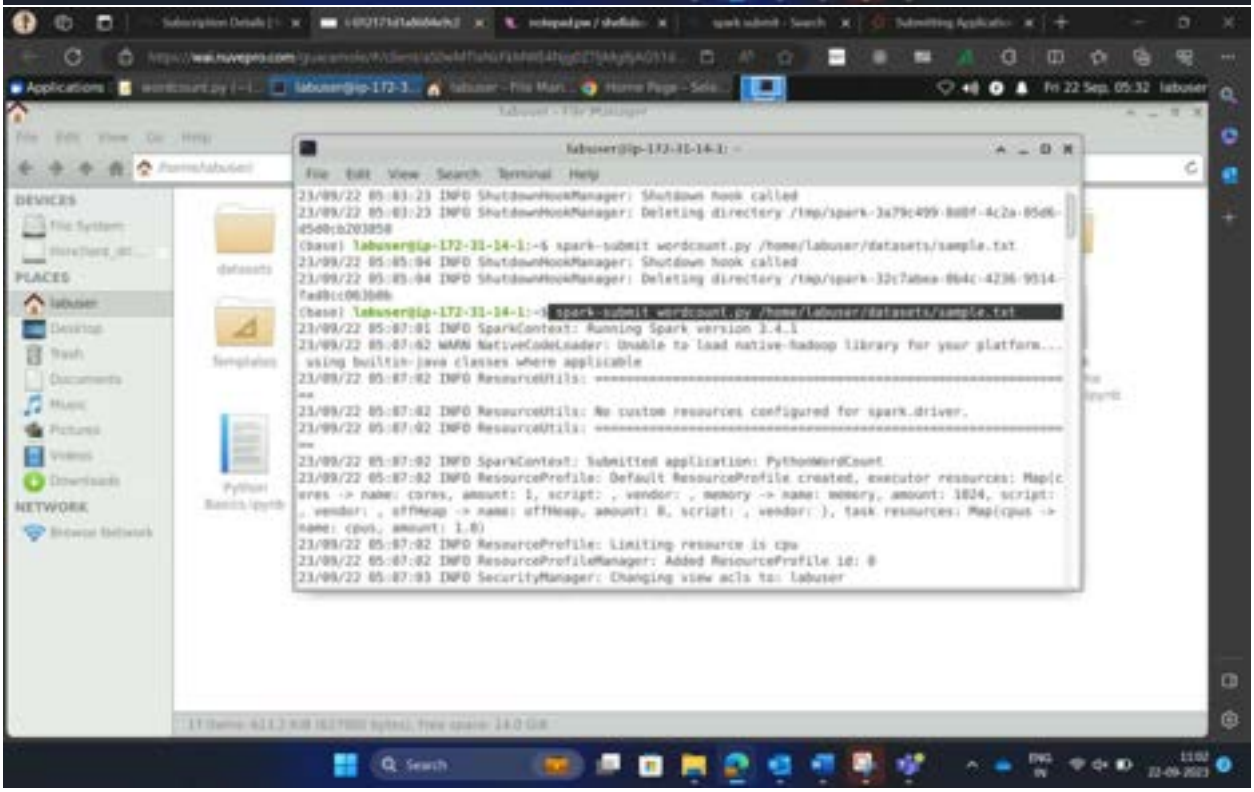
In [ ]:

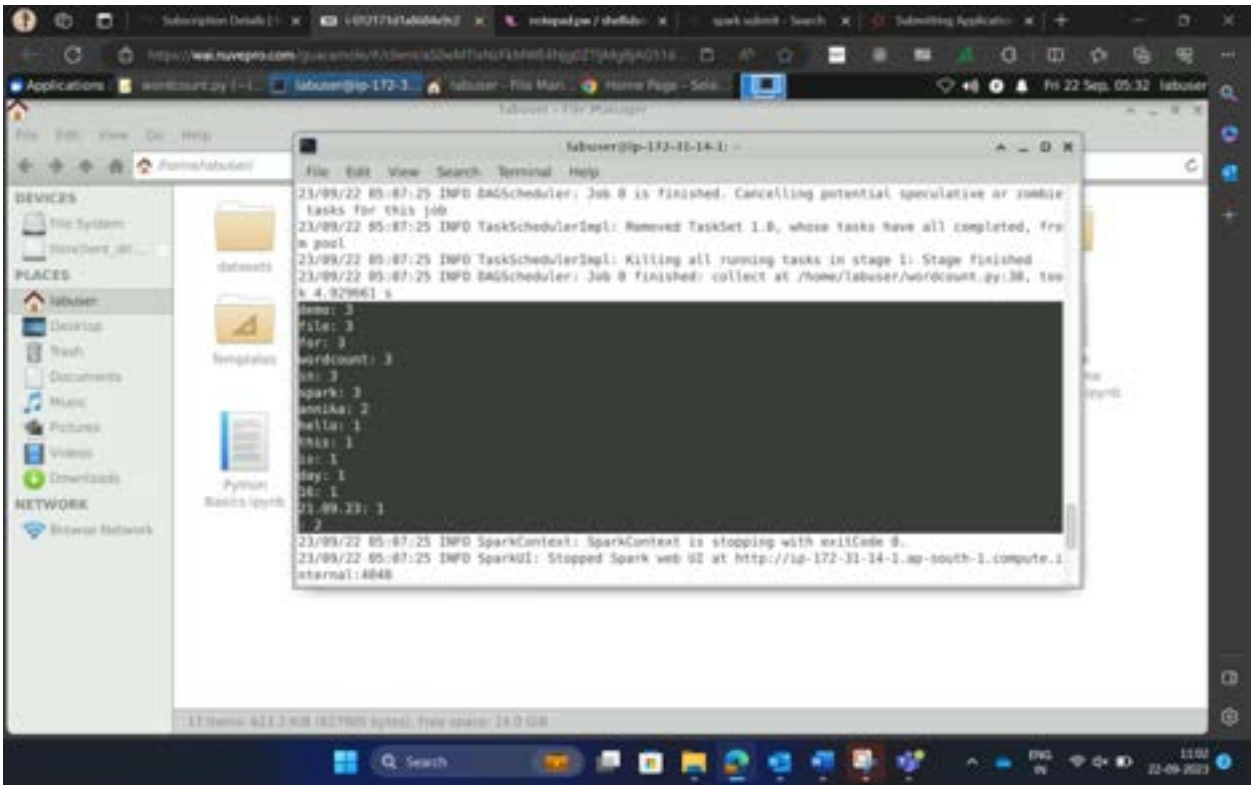
```
In [66]: sales_india_df.count()
Out[66]: 75

In [67]: #to check the activity in the localhost
         spark.sparkContext.uiWebUrl
Out[67]: 'http://ip-172-31-14-1.ap-south-1.compute.internal:4040'

In [68]: #to write into the file after the following transformations

         sales_india_df.write.format("json").save("/home/labuser/Downloads/sparkout/sales_india")

In [ ]:
```



Day 17 (22.09.23)


labuser@ip-172-31-14-1: ~
File Edit View Search Terminal Help
(base) labuser@ip-172-31-14-1:~$ spark-submit
Usage: spark-submit [options] <app jar | python file | R file> [app arguments]
Usage: spark-submit --kill [submission ID] --master [spark://...]
Usage: spark-submit --status [submission ID] --master [spark://...]
Usage: spark-submit run-example [options] example-class [example args]
Options:
--master MASTER_URL spark://host:port, mesos://host:port, yarn,
k8s://https://host:port, or local (Default: local[*]).
--deploy-mode DEPLOY_MODE Whether to launch the driver program locally ("client") or
on one of the worker machines inside the cluster ("cluster")
(Default: client).
--class CLASS_NAME Your application's main class (for Java / Scala apps).
--name NAME A name of your application.
--jars JARS Comma-separated list of jars to include on the driver
and executor classpaths.
--packages Comma-separated list of maven coordinates of jars to include
on the driver and executor classpaths. Will search the local
maven repo, then maven central and any additional remote
repositories given by --repositories. The format for the
coordinates should be groupId:artifactId:version.
--exclude-packages Comma-separated list of groupId:artifactId, to exclude while
resolving the dependencies provided in --packages to avoid
dependency conflicts.

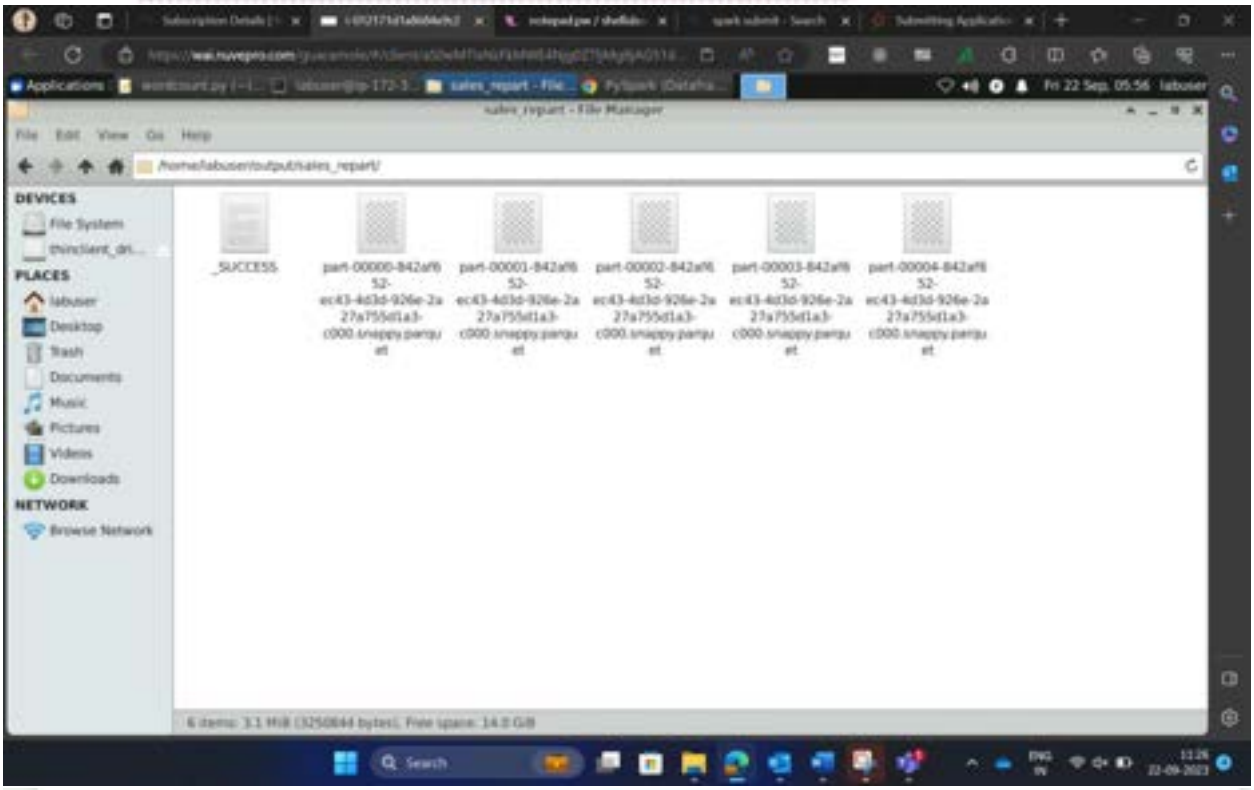

labuser@ip-172-31-14-1: ~
File Edit View Search Terminal Help
23/09/22 05:03:23 INFO ShutdownHookManager: Shutdown hook called
23/09/22 05:03:23 INFO ShutdownHookManager: Deleting directory /tmp/spark-3a79c499-808f-4c2a-85d6-d5d0cb203858
(base) labuser@ip-172-31-14-1:~$ spark-submit wordcount.py /home/labuser/datasets/sample.txt
23/09/22 05:05:04 INFO ShutdownHookManager: Shutdown hook called
23/09/22 05:05:04 INFO ShutdownHookManager: Deleting directory /tmp/spark-32c7abea-8b4c-4236-9514-fad0cc063d6b
(base) labuser@ip-172-31-14-1:~$ spark-submit wordcount.py /home/labuser/datasets/sample.txt
23/09/22 05:07:01 INFO SparkContext: Running Spark version 3.4.1
23/09/22 05:07:02 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/09/22 05:07:02 INFO ResourceUtils: ==========
23/09/22 05:07:02 INFO ResourceUtils: No custom resources configured for spark.driver.
23/09/22 05:07:02 INFO ResourceUtils: ==========
23/09/22 05:07:02 INFO SparkContext: Submitted application: PythonWordCount
23/09/22 05:07:02 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor:), task resources: Map(cpus -> name: cpus, amount: 1.0)
23/09/22 05:07:02 INFO ResourceProfile: Limiting resource is cpu
23/09/22 05:07:02 INFO ResourceProfileManager: Added ResourceProfile id: 0
23/09/22 05:07:03 INFO SecurityManager: Changing view acls to: labuser




```
In [35]: #to see the no of partitions in a dataframe'
         salesdf.rdd.getNumPartitions()

Out[35]: 2

In [48]: #repartition
         salesdf_repart=salesdf.repartition(5)

In [49]: salesdf_repart.rdd.getNumPartitions()
         [Stage 42:==============================>                          (1 + 1) / 2]

Out[49]: 5

In [51]: salesdf.count()

Out[51]: 51290

In [50]: salesdf_repart.count()


Out[50]: 51290

In [36]: #
         salesdf.show(2)
```

```
In [53]: #writing data as a parquet file
salesdf_repart.write.parquet("/home/labuser/output/sales_repart")
```

```
[Stage 55:==============================>                    (1 + 1) / 2]
```

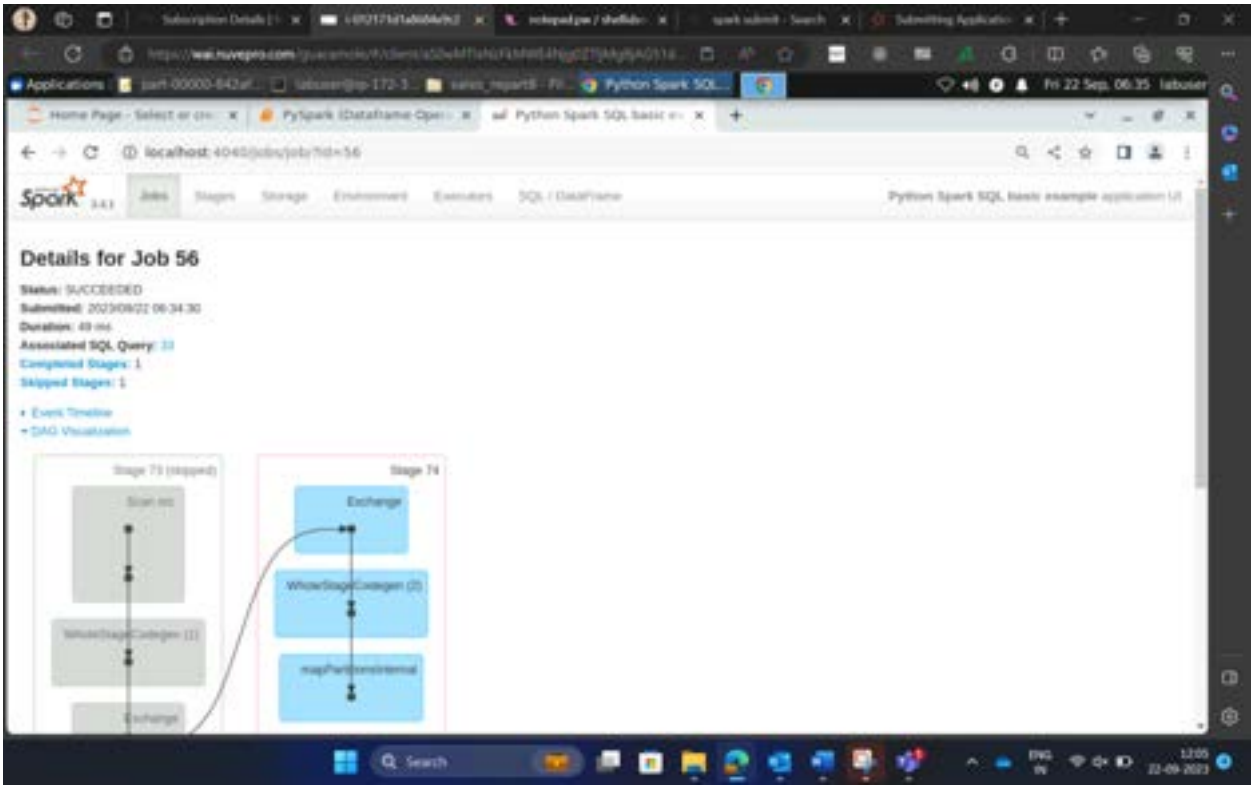



```
In [55]: salesdf.repartition(8).write.orc("/home/labuser/output/sales_repart8")

In [56]: #to create the parquet file
sales_parquet = spark.read.format("parquet").load("/home/labuser/output/sales_repart")

In [58]: sales_parquet_1 = spark.read.format("parquet").load("/home/labuser/output/sales_repart/part-00000-842af652-ec43-4d3c

In [59]: sales_parquet_1.count()
Out[59]: 10259
```

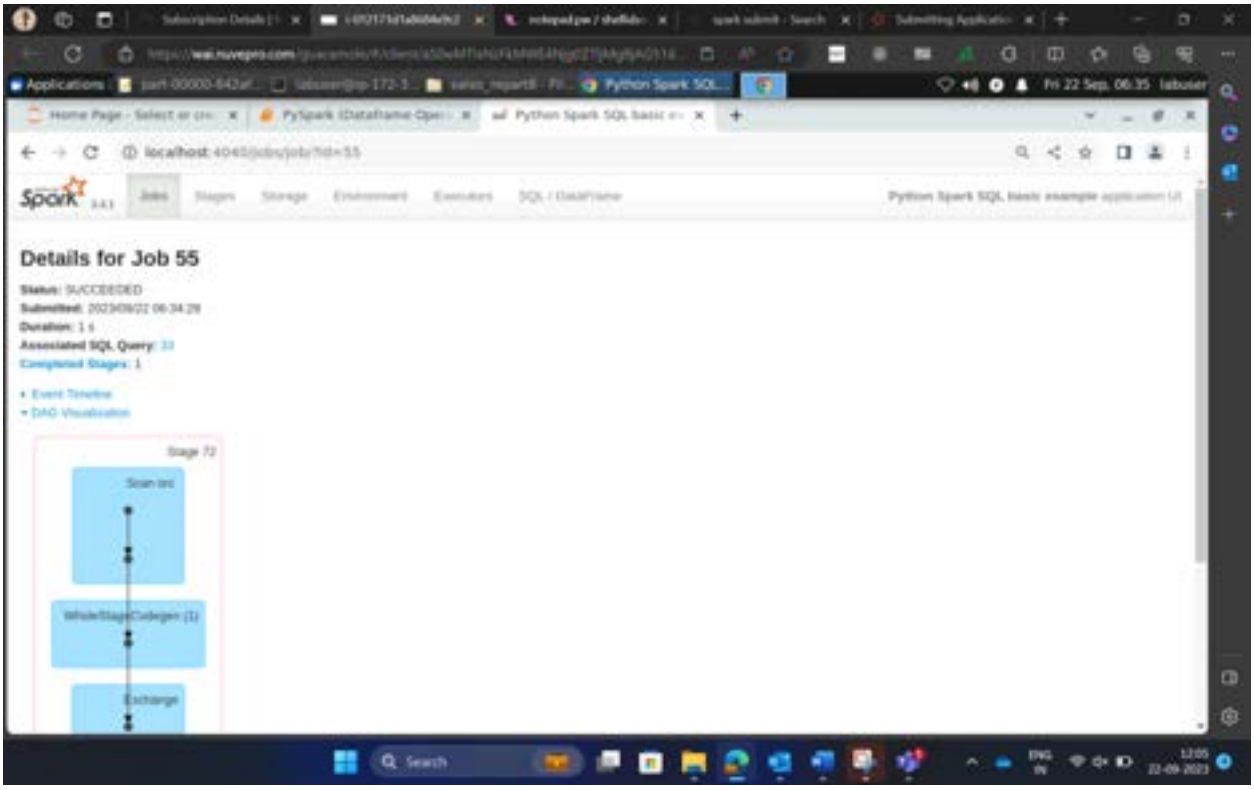
```
In [56]: #to create the parquet file
         sales_parquet = spark.read.format("parquet").load("/home/labuser/output/sales_repart")

In [58]: sales_parquet_1 = spark.read.format("parquet").load("/home/labuser/output/sales_repart/part-00000-842af052-ec43-4d3

In [59]: sales_parquet_1.count()
Out[59]: 10259

In [63]: sales_orc=spark.read.format("orc").load("/home/labuser/output/sales_repart8/")

In [66]: sales_orc.count()
Out[66]: 51290

In [64]: sales_orc.rdd.getNumPartitions()
Out[64]: 2

In [65]: sales_coal_df = sales_orc.coalesce(1)
```

```
In [64]: sales_orc.rdd.getNumPartitions()
Out[64]: 2

In [65]: sales_coal_df = sales_orc.coalesce(1)

In [67]: sales_coal_df.rdd.getNumPartitions()
Out[67]: 1

In [69]: sales_rep = sales_orc.repartition(1)

In [70]: sales_rep.rdd.getNumPartitions()
         [Stage 79:==============================>                        (1 + 1) / 2]
Out[70]: 1
```

For repartition 2 stages – full shuffle - > across partition



Coalesce – only one stage- decrease the partition - > no shuffle



SQL OPERATIONS

In [73]:
```
#to perform sql operations
#creating temp table for the dataframe

salesdf.createOrReplaceTempView("sales")
```

In [74]:
```
spark.sql("show tables").show()
```

```
+---------+---------+-----------+
|namespace|tableName|isTemporary|
+---------+---------+-----------+
|         |    sales|       true|
+---------+---------+-----------+
```

In [76]:
```
spark.sql("select country, state, profit from sales where country = 'Australia'").show()
```

```
+---------+----------------+------+
|  country|           state|profit|
+---------+----------------+------+
|Australia|  New South Wales|   0.1|
|Australia|      Queensland|   0.1|
|Australia|  New South Wales|   0.1|
|Australia|      Queensland|   0.1|
|Australia|      Queensland|   0.1|
|Australia|      Queensland|   0.1|
|Australia|Western Australia|   0.1|
|Australia|        Victoria|   0.1|
|Australia|  New South Wales|   0.1|
|Australia|  New South Wales|     0|
|Australia|  New South Wales|   0.1|
|Australia|Western Australia|   0.1|
|Australia|      Queensland|   0.1|
|Australia|  New South Wales|   0.1|
|Australia|  South Australia|   0.1|
|Australia|  South Australia|   0.1|
|Australia|Western Australia|   0.1|
|Australia|      Queensland|   0.1|
|Australia|Western Australia|   0.1|
|Australia|  South Australia|   0.1|
+---------+----------------+------+
only showing top 20 rows
```

In [77]:
```
spark.sql("select country, sum(profit) from sales group by country").show()
```

```
[Stage 79:>                                                         (0 + 2) / 2]
```

```
+-----------+------------------+
|    country|       sum(profit)|
+-----------+------------------+
|       Chad|               0.0|
|     Russia|               0.0|
|   Paraguay|             0.802|
|      Yemen|20.999999999999996|
|    Senegal|               0.0|
|     Sweden|103.10000000000001|
|Philippines|            235.55|
|    Eritrea|               0.0|
|   Djibouti|               0.0|
|   Malaysia|               0.0|
|  Singapore|               0.0|
|     Turkey| 826.8000000000085|
|       Iraq|               0.0|
|    Germany| 117.7999999999997|
|Afghanistan|               0.0|
|   Cambodia|               0.0|
|     Rwanda|               0.0|
|     Jordan|               0.0|
|      Sudan|               0.0|
|     France|204.34999999999903|
+-----------+------------------+
only showing top 20 rows
```

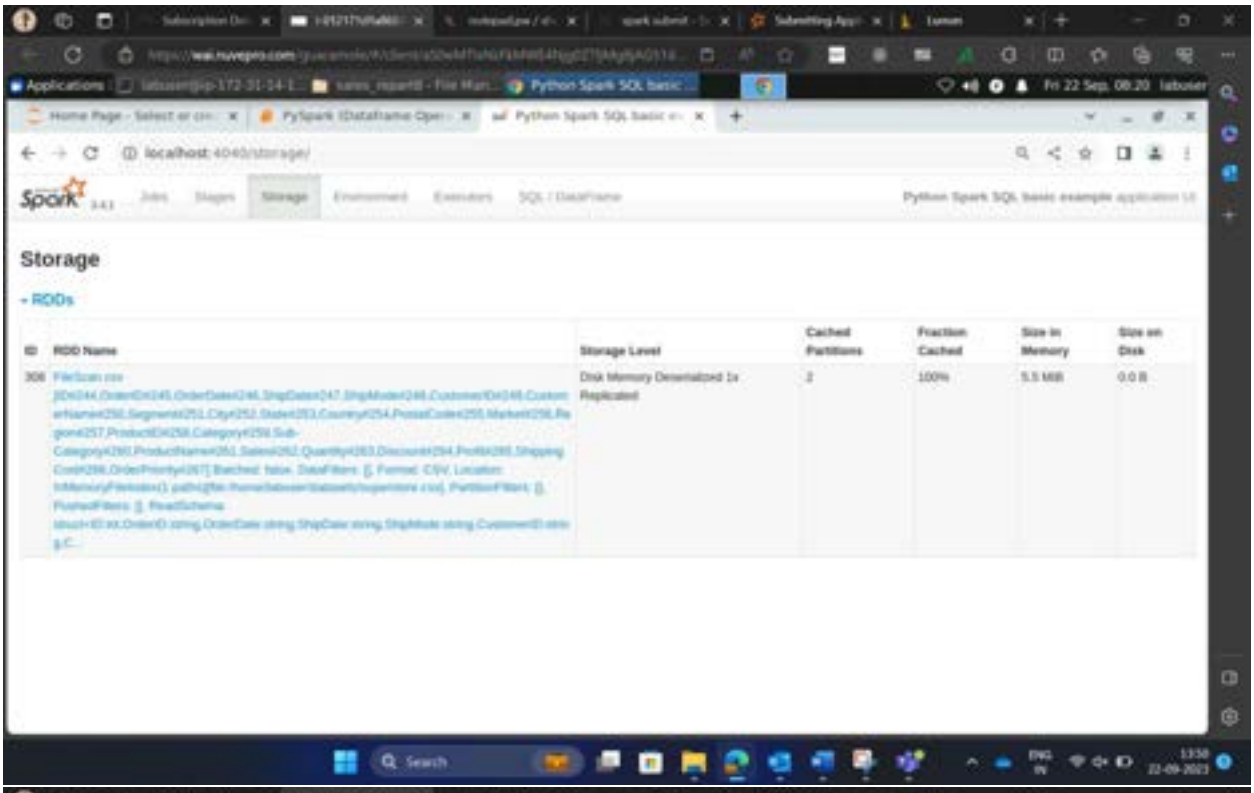
```
In [78]: #creating a new database
         spark.sql("create database shellida")

Out[78]: DataFrame[]

In [79]: #to see the database
         spark.sql("show databases").show()

         +---------+
         |namespace|
         +---------+
         |  default|
         | shellida|
         +---------+

In [80]: #to use a particular database
         spark.sql("use shellida")

Out[80]: DataFrame[]
```

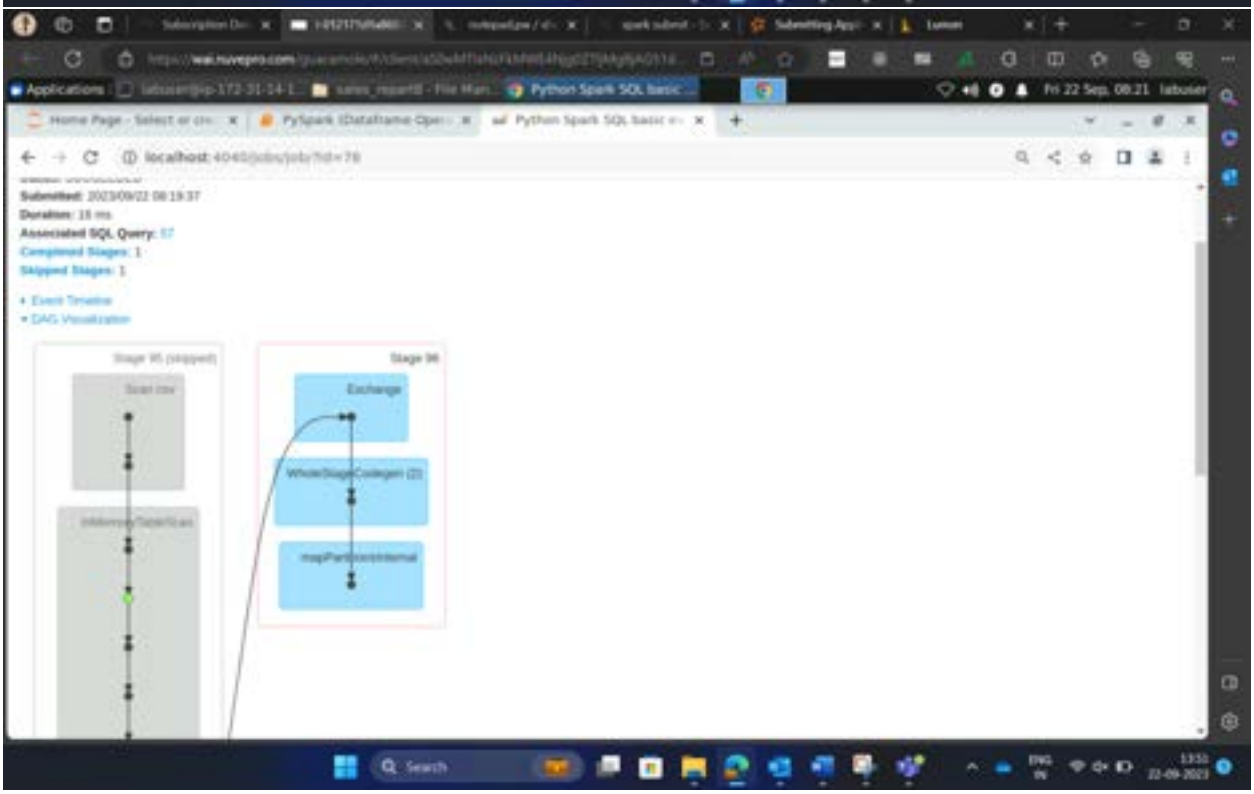
```
In [80]: #to use a particular database
         spark.sql("use shellida")

Out[80]: DataFrame[]

In [81]: #to write the data into a permanent table
         salesdf.write.saveAsTable("shellida.sales_perm")

In [82]: spark.sql("show tables").show()

         +---------+----------+-----------+
         |namespace| tableName|isTemporary|
         +---------+----------+-----------+
         | shellida|sales_perm|      false|
         |         |     sales|       true|
         +---------+----------+-----------+
```

Spark caching

```
In [86]: #caching

         empdf = spark.read.csv("/home/labuser/datasets/emp.csv")

In [87]: empdf.show()

         +---+---+----+
         |_c0|_c1| _c2|
         +---+---+----+
         |111|zzz|8000|
         |111|aaa|8888|
         |121|bbb|8000|
         +---+---+----+
```

```
In [89]: salesdf.cache()

Out[89]: DataFrame[ID: int, OrderID: string, OrderDate: string, ShipDate: string, ShipMode: string, CustomerID: string, Cust
         omerName: string, Segment: string, City: string, State: string, Country: string, PostalCode: int, Market: string, R
         egion: string, ProductID: string, Category: string, Sub-Category: string, ProductName: string, Sales: string, Quant
         ity: string, Discount: string, Profit: string, ShippingCost: string, OrderPriority: string]

In [90]: salesdf.count()

Out[90]: 51290

In [ ]:
```


Python Spark SQL basic ...
Fri 22 Sep, 08:20 labuser
localhost:4040/storage/
Spark 3.4.1
Jobs
Stages
Storage
Environment
Executors
SQL / DataFrame
Python Spark SQL basic example application UI
Storage
- RDDs
ID
RDD Name
Storage Level
Cached Partitions
Fraction Cached
Size in Memory
Size on Disk
308
FileScan csv [ID#244,OrderID#245,OrderDate#246,ShipDate#247,ShipMode#248,CustomerID#249,CustomerName#250,Segment#251,City#252,State#253,Country#254,PostalCode#255,Market#256,Region#257,ProductID#258,Category#259,Sub-Category#260,ProductName#261,Sales#262,Quantity#263,Discount#264,Profit#265,ShippingCost#266,OrderPriority#267] Batched: false, DataFilters: [], Format: CSV, Location: InMemoryFileIndex(1 paths)[file:/home/labuser/datasets/superstore.csv], PartitionFilters: [], PushedFilters: [], ReadSchema: struct<ID:int,OrderID:string,OrderDate:string,ShipDate:string,ShipMode:string,CustomerID:string,C...
Disk Memory Deserialized 1x Replicated
2
100%
5.5 MiB
0.0 B
Search
ENG IN
13:50 22-09-2023


Python Spark SQL basic ...
Fri 22 Sep, 08:21 labuser
localhost:4040/jobs/job/?id=78
Submitted: 2023/09/22 08:19:37
Duration: 15 ms
Associated SQL Query: 57
Completed Stages: 1
Skipped Stages: 1
Event Timeline
DAG Visualization
Stage 96
Exchange
WholeStageCodegen (2)
Search
ENG IN
13:51 22-09-2023

In [91]:
```
empdf.show() #source location was changed and the file was not cached thts why error
```

```
    170     except Py4JJavaError as e:
    171         converted = convert_exception(e.java_exception)

File /opt/anaconda3/lib/python3.11/site-packages/py4j/protocol.py:326, in get_return_value(answer, gateway_clien
t, target_id, name)
    324 value = OUTPUT_CONVERTER[type](answer[2:], gateway_client)
    325 if answer[1] == REFERENCE_TYPE:
--> 326     raise Py4JJavaError(
    327         "An error occurred while calling {0}{1}{2}.\n".
    328         format(target_id, ".", name), value)
    329 else:
    330     raise Py4JError(
    331         "An error occurred while calling {0}{1}{2}. Trace:\n{3}\n".
    332         format(target_id, ".", name, value))

Py4JJavaError: An error occurred while calling o270.showString.
: org.apache.spark.SparkException: Job aborted due to stage failure: Task 0 in stage 97.0 failed 1 times, most re
cent failure: Lost task 0.0 in stage 97.0 (TID 112) (ip-172-31-14-1.ap-south-1.compute.internal executor driver):
```

In [92]:
```
salesdf.show() # location was changed but the output is coming from the cache memory
```

```
+-----+---------------+----------+----------+-------------+----------+---------------+----------+-------------+-
-------------+-------------+----------+------+-----------+---------------+--------------+-----------+---------
----------+---------------+--------+--------+----------+-----------+-------------+
|   ID|        OrderID| OrderDate|  ShipDate|     ShipMode|CustomerID|   CustomerName|   Segment|         City|
State|      Country|PostalCode|Market|     Region|     ProductID|      Category|Sub-Category|         ProductNa
me|          Sales|Quantity|Discount|    Profit|ShippingCost|OrderPriority|
+-----+---------------+----------+----------+-------------+----------+---------------+----------+-------------+-
-------------+-------------+----------+------+-----------+---------------+--------------+-----------+---------
----------+---------------+--------+--------+----------+-----------+-------------+
|32298| CA-2012-124891|31/07/2012|31/07/2012|     Same Day|  RH-19495|    Rick Hansen|  Consumer|New York City|
New York|United States|     10024|    US|       East| TEC-AC-10003833|    Technology| Accessories|Plantronics CS5
10...|        2309.65|       7|       0|  762.1845|     933.57 |     Critical|
|26341|  IN-2013-77878|05/02/2013|07/02/2013|  Second Class|  JR-16210|  Justin Ritter|  Corporate|   Wollongong|N
ew South Wales|    Australia|      null|  APAC|     Oceania| FUR-CH-10003950|     Furniture|      Chairs|Novimex E
xecutive...|          Black|3709.395|       9|       0.1|   -288.765|      923.63 |
|25330|  IN-2013-71249|17/10/2013|18/10/2013|   First Class|  CR-12730|    Craig Reiter|  Consumer|     Brisbane|
```

In [94]:
```
salesdf.unpersist() #after running unpersist() the cache gets deleted from the storage in the webui
```

Out[94]:
```
DataFrame[ID: int, OrderID: string, OrderDate: string, ShipDate: string, ShipMode: string, CustomerID: string, Cust
omerName: string, Segment: string, City: string, State: string, Country: string, PostalCode: int, Market: string, R
egion: string, ProductID: string, Category: string, Sub-Category: string, ProductName: string, Sales: string, Quant
ity: string, Discount: string, Profit: string, ShippingCost: string, OrderPriority: string]
```

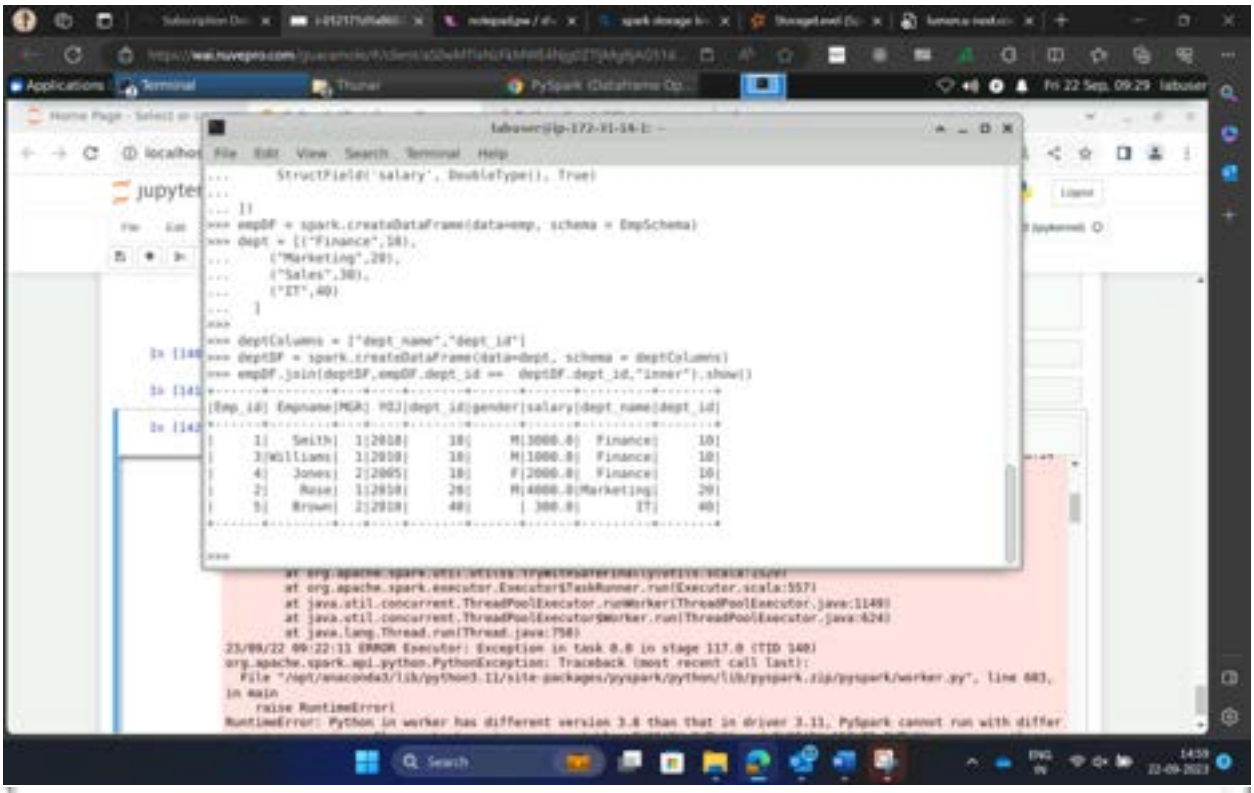
```
In [103]: salesdf.unpersist()

Out[103]: DataFrame[ID: int, OrderID: string, OrderDate: string, ShipDate: string, ShipMode: string, CustomerID: string, Cust
          omerName: string, Segment: string, City: string, State: string, Country: string, PostalCode: int, Market: string, R
          egion: string, ProductID: string, Category: string, Sub-Category: string, ProductName: string, Sales: string, Quant
          ity: string, Discount: string, Profit: string, ShippingCost: string, OrderPriority: string]

In [104]: from pyspark.storagelevel import *

In [105]: salesdf.persist(StorageLevel.MEMORY_AND_DISK_2)

Out[105]: DataFrame[ID: int, OrderID: string, OrderDate: string, ShipDate: string, ShipMode: string, CustomerID: string, Cust
          omerName: string, Segment: string, City: string, State: string, Country: string, PostalCode: int, Market: string, R
          egion: string, ProductID: string, Category: string, Sub-Category: string, ProductName: string, Sales: string, Quant
          ity: string, Discount: string, Profit: string, ShippingCost: string, OrderPriority: string]

In [106]: salesdf.show()

          [Stage 106:>                                                        (0 + 1) / 1]

          +-----+---------------+----------+----------+--------------+----------+----------------+-----------+-------------+-
          ----------------+--------------+-----------+------+------------+----------+---------------+-----------+-------------+---------
```

Spark 3.4.1 Jobs Stages Storage Environment Executors SQL / DataFrame Python Spark SQL basic example application UI

Storage

- RDDs

ID	RDD Name	Storage Level	Cached Partitions	Fraction Cached	Size in Memory	Size on Disk
360	FileScan csv [ID#244,OrderID#245,OrderDate#246,ShipDate#247,ShipMode#248,CustomerID#249,CustomerName#250,Segment#251,City#252,State#253,Country#254,PostalCode#255,Market#256,Region#257,ProductID#258,Category#259,Sub-Category#260,ProductName#261,Sales#262,Quantity#263,Discount#264,Profit#265,ShippingCost#266,OrderPriority#267] Batched: false, DataFilters: [], Format: CSV, Location: InMemoryFileIndex[], paths=[file:/home/labuser/datasets/superstore.csv], PartitionFilters: [], PushedFilters: [], ReadSchema: struct<ID:int,OrderID:string,OrderDate:string,ShipDate:string,ShipMode:string,CustomerID:string,C...	Disk Memory Serialized 2x Replicated	1	50%	2.6 MiB	0.0 B



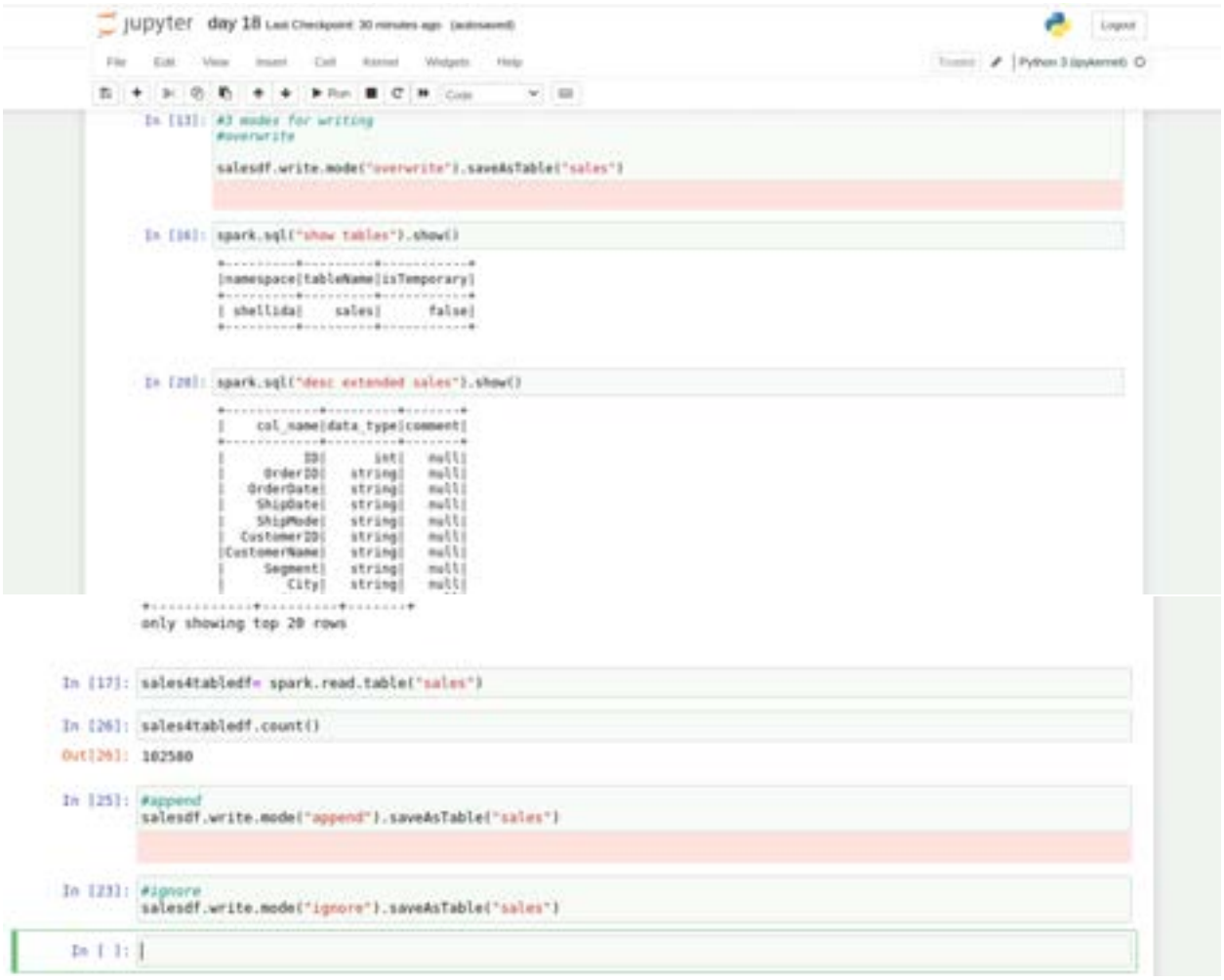

```
>>> empDF.join(deptDF,empDF.dept_id ==  deptDF.dept_id,"full").show()
+------+--------+----+----+-------+------+------+---------+-------+
|Emp_id| Empname| MGR| YOJ|dept_id|gender|salary|dept_name|dept_id|
+------+--------+----+----+-------+------+------+---------+-------+
|     1|   Smith|   1|2018|     10|     M|3000.0|  Finance|     10|
|     3|Williams|   1|2010|     10|     M|1000.0|  Finance|     10|
|     4|   Jones|   2|2005|     10|     F|2000.0|  Finance|     10|
|     2|    Rose|   1|2010|     20|     M|4000.0|Marketing|     20|
|  null|    null|null|null|   null|  null|  null|    Sales|     30|
|     5|   Brown|   2|2010|     40|      | 300.0|       IT|     40|
|     6|   Brown|   2|2010|     50|      |2000.0|     null|   null|
+------+--------+----+----+-------+------+------+---------+-------+

>>>
```

```
>>> empDF.join(deptDF,empDF.dept_id ==  deptDF.dept_id,"right").show()
+------+--------+----+----+-------+------+------+---------+-------+
|Emp_id| Empname| MGR| YOJ|dept_id|gender|salary|dept_name|dept_id|
+------+--------+----+----+-------+------+------+---------+-------+
|     4|   Jones|   2|2005|     10|     F|2000.0|  Finance|     10|
|     3|Williams|   1|2010|     10|     M|1000.0|  Finance|     10|
|     1|   Smith|   1|2018|     10|     M|3000.0|  Finance|     10|
|     2|    Rose|   1|2010|     20|     M|4000.0|Marketing|     20|
|  null|    null|null|null|   null|  null|  null|    Sales|     30|
|     5|   Brown|   2|2010|     40|      | 300.0|       IT|     40|
+------+--------+----+----+-------+------+------+---------+-------+

>>>
```

```
>>> empDF.join(deptDF,empDF.dept_id ==  deptDF.dept_id,"leftanti").show()
+------+--------+---+----+--------+------+------+
|Emp_id|Empname|MGR| YOJ|dept_id|gender|salary|
+------+--------+---+----+--------+------+------+
|     6|  Brown|  2|2010|     50|      |2000.0|
+------+--------+---+----+--------+------+------+
```

```
>>> empDF.join(deptDF,empDF.dept_id ==  deptDF.dept_id,"leftsemi").show()
+------+---------+---+----+--------+------+------+
|Emp_id| Empname|MGR| YOJ|dept_id|gender|salary|
+------+---------+---+----+--------+------+------+
|     1|   Smith|  1|2018|     10|     M|3000.0|
|     3|Williams|  1|2010|     10|     M|1000.0|
|     4|   Jones|  2|2005|     10|     F|2000.0|
|     2|    Rose|  1|2010|     20|     M|4000.0|
|     5|   Brown|  2|2010|     40|      | 300.0|
+------+---------+---+----+--------+------+------+
```

Day 18



```
In [27]: sales_US=salesdf.select("country","state","Profit").filter("country='United States'").show()

+-------------+--------------+----------+
|      country|         state|    Profit|
+-------------+--------------+----------+
|United States|      New York|  762.1845|
|United States|    California|  1906.485|
|United States|North Carolina|-1862.3124|
|United States|      Virginia|    83.281|
|United States|      Kentucky|  517.4793|
|United States|      Illinois|   341.994|
|United States|    California|  363.9048|
|United States|         Texas|   -350.49|
|United States|    California|  135.4068|
|United States|      New York| 1371.9804|
|United States|North Carolina|-3839.9904|
|United States|     Minnesota| 4630.4755|
|United States|      New York|  1143.891|
|United States|    California|   839.986|
|United States|    Washington| 1480.4671|
|United States|       Florida|  327.5922|
|United States|      New York|  2229.024|
|United States|      Virginia|  694.5015|
|United States|       Georgia|  3177.475|
|United States|      Michigan| 2504.2216|
+-------------+--------------+----------+
only showing top 20 rows
```

```
In [29]: from pyspark.sql.functions import col, sum

         #col function
         #transformations -> groupby, agg, cast,alias, filter

         sales_US =salesdf.groupBy("Country","state").agg(sum(col("Profit").cast('int')).alias("totSales")).filter("country='
         [Stage 15:>                                                 (0 + 2) / 2]

+-------------+--------------+--------+
|      Country|         state|totSales|
+-------------+--------------+--------+
|United States|      Virginia|   17154|
|United States|South Carolina|    1774|
|United States| West Virginia|     186|
|United States|         Maine|     417|
|United States|       Montana|    1852|
|United States|       Alabama|    4645|
|United States|       Wyoming|     100|
|United States|         Idaho|     695|
|United States|    New Jersey|    6916|
|United States|      Arkansas|    3437|
|United States|      Michigan|   21869|
|United States|      Illinois|   -9298|
|United States|          Ohio|  -14490|
|United States|      New York|   65943|
|United States|      Missouri|    6208|
|United States|   Mississippi|    2445|
```

from pyspark.sql.functions import col, sum

#col function

#transformations -> groupby, agg, cast,alias, filter

sales_US
=salesdf.groupBy("Country","state").agg(sum(col("Profit").cast('int')).alias("totSales")).filter("country='United States'").show()

```
In [30]: salesdf.select("Region").where("Country='India'").orderBy("region",ascending=True).show()
```

```
+------------+
|      Region|
+------------+
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
|Central Asia|
+------------+
only showing top 20 rows
```

```
In [33]: salesdf.select("Region","country").sort(col("region").desc()).orderBy("country",ascending=False).filter("region='Cen
```

```
+------------+-----------+
|      Region|    country|
+------------+-----------+
|Central Asia|      Nepal|
|Central Asia|Afghanistan|
|Central Asia| Bangladesh|
|Central Asia|  Sri Lanka|
|Central Asia|   Pakistan|
|Central Asia|      India|
+------------+-----------+
```

```
In [34]: spark.sql("select country,region,avg(profit) from sales group by country,region").filter("Country='India'").orderBy(
```

```
+-------+------------+--------------------+
|country|      region|         avg(profit)|
+-------+------------+--------------------+
|  India|Central Asia|0.017664807459007974|
+-------+------------+--------------------+
```

```
In [35]: # to get tyhe average sales
from pyspark.sql.functions import col, sum, avg

sales_US =salesdf.groupBy("Country","state").agg(sum(col("Profit").cast('int')).alias("totSales"),avg(col("Profit").
```

```
+--------------+--------------+--------+-------------------+
|       Country|         state|totSales|           AvgSales|
+--------------+--------------+--------+-------------------+
|United States|      Virginia|   17154|  76.58035714285714|
|United States|South Carolina|    1774|  42.23809523809524|
|United States| West Virginia|     186|               46.5|
|United States|         Maine|     417|             52.125|
|United States|       Montana|    1852| 123.46666666666667|
|United States|       Alabama|    4645|  76.14754098360656|
|United States|       Wyoming|     100|              100.0|
|United States|         Idaho|     695| 33.095238095238095|
|United States|    New Jersey|    6916|   53.6124031007752|
|United States|      Arkansas|    3457|  57.28333333333333|
|United States|      Michigan|   21869|   86.0984251968504|
|United States|      Illinois|   -9298|-18.898373983739837|
|United States|          Ohio|  -14490|-30.895522388059703|
|United States|      New York|   65943|  58.77272727272727|
|United States|      Missouri|    6200|  94.06060606060606|
|United States|   Mississippi|    2445| 46.132075471698116|
|United States|      Nebraska|    1421|  37.39473684210526|
|United States|     Minnesota|    9820| 110.33707865168539|
|United States|    Washington|   29900|  59.20792079207921|
|United States|        Kansas|     794| 33.083333333333336|
+--------------+--------------+--------+-------------------+
only showing top 20 rows
```

from pyspark.sql.functions import col, sum

```
sales_US
=salesdf.groupBy("Country","state").agg(sum(col("Profit").cast('int')).alias("totSales")).filter("country='United States'").show()


salesdf.select("Region","country").sort(col("region").desc()).orderBy("country",ascending=False).filter("region='Central Asia'").distinct().show()


 spark.sql("select country,region,avg(profit) from sales group by country,region").filter("Country='India'").orderBy("region").show()
```

In [54]:
```
sales_withcolumn = salesdf.withColumn("SalesLevel", when(col("profit")>=10000,"High").when(col("profit")<10000,"Aver
```

In [55]:
```
sales_withcolumn.select("country","state","profit","salesLevel").show()
```

```
+-------------+--------------+----------+----------+
|      country|         state|    profit|salesLevel|
+-------------+--------------+----------+----------+
|United States|      New York|  762.1845|   Average|
|    Australia|New South Wales|       0.1|   Average|
|    Australia|    Queensland|       0.1|   Average|
|      Germany|        Berlin|       0.1|   Average|
|      Senegal|         Dakar|         0|   Average|
|    Australia|New South Wales|       0.1|   Average|
|  New Zealand|     Wellington|         0|   Average|
|  New Zealand|       Waikato|         0|   Average|
|United States|    California|  1906.485|   Average|
|United States| North Carolina|-1862.3124|   Average|
|United States|      Virginia|    83.281|   Average|
|  Afghanistan|         Kabul|         0|   Average|
| Saudi Arabia|         Jizan|         0|   Average|
|       Brazil|        Parana|         0|   Average|
|        China|  Heilongjiang|         0|   Average|
|       France| Ile-de-France|       0.1|   Average|
|United States|      Kentucky|  517.4793|   Average|
|        Italy|       Tuscany|         0|   Average|
|    Australia|    Queensland|       0.1|   Average|
```

In [57]:
```
from pyspark.sql.functions import lit
sales_US.withColumn("CountryCode",lit("US")).show()
```

```
+-------------+--------------+----------+------------------+-----------+
|      Country|         state|TotalSales|          AvgSales|CountryCode|
+-------------+--------------+----------+------------------+-----------+
|United States|      Virginia|     17154|  76.58035714285714|         US|
|United States|South Carolina|      1774|  42.23809523809524|         US|
|United States| West Virginia|       186|              46.5|         US|
|United States|         Maine|       417|            52.125|         US|
|United States|       Montana|      1852|123.46666666666667|         US|
|United States|       Alabama|      4645|  76.14754098360656|         US|
|United States|       Wyoming|       100|             100.0|         US|
|United States|         Idaho|       695| 33.095238095238095|         US|
|United States|    New Jersey|      6916|   53.6124031007752|         US|
|United States|      Arkansas|      3437|  57.28333333333333|         US|
|United States|      Michigan|     21869|   86.0984251968504|         US|
|United States|      Illinois|     -9298|-18.898373983739837|         US|
|United States|          Ohio|    -14490|-30.895522388059703|         US|
|United States|      New York|     65943|  58.77272727272727|         US|
|United States|      Missouri|      6208|  94.06060606060606|         US|
|United States|   Mississippi|      2445| 46.132075471698116|         US|
|United States|      Nebraska|      1421|  37.39473684210526|         US|
|United States|     Minnesota|      9820| 110.33707865168539|         US|
|United States|    Washington|     29900|  59.20792079207921|         US|
|United States|        Kansas|       794| 33.083333333333336|         US|
```

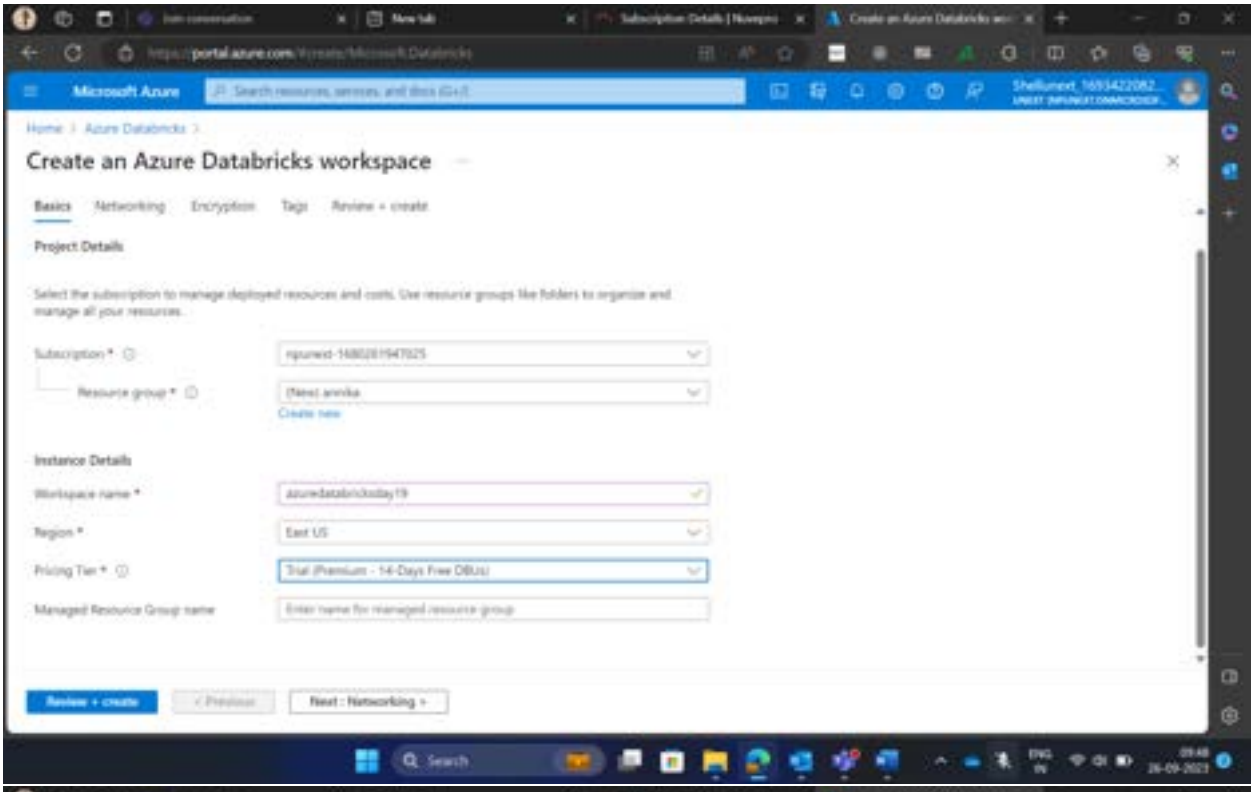
In [65]:
```
from pyspark.sql.types import StructType,StructField, StringType,IntegerType,DoubleType

EmpSchema = StructType([
    StructField('Empno', IntegerType(), True),
    StructField('Empname', StringType(), True),
     StructField('salary', DoubleType(), True)
])
```

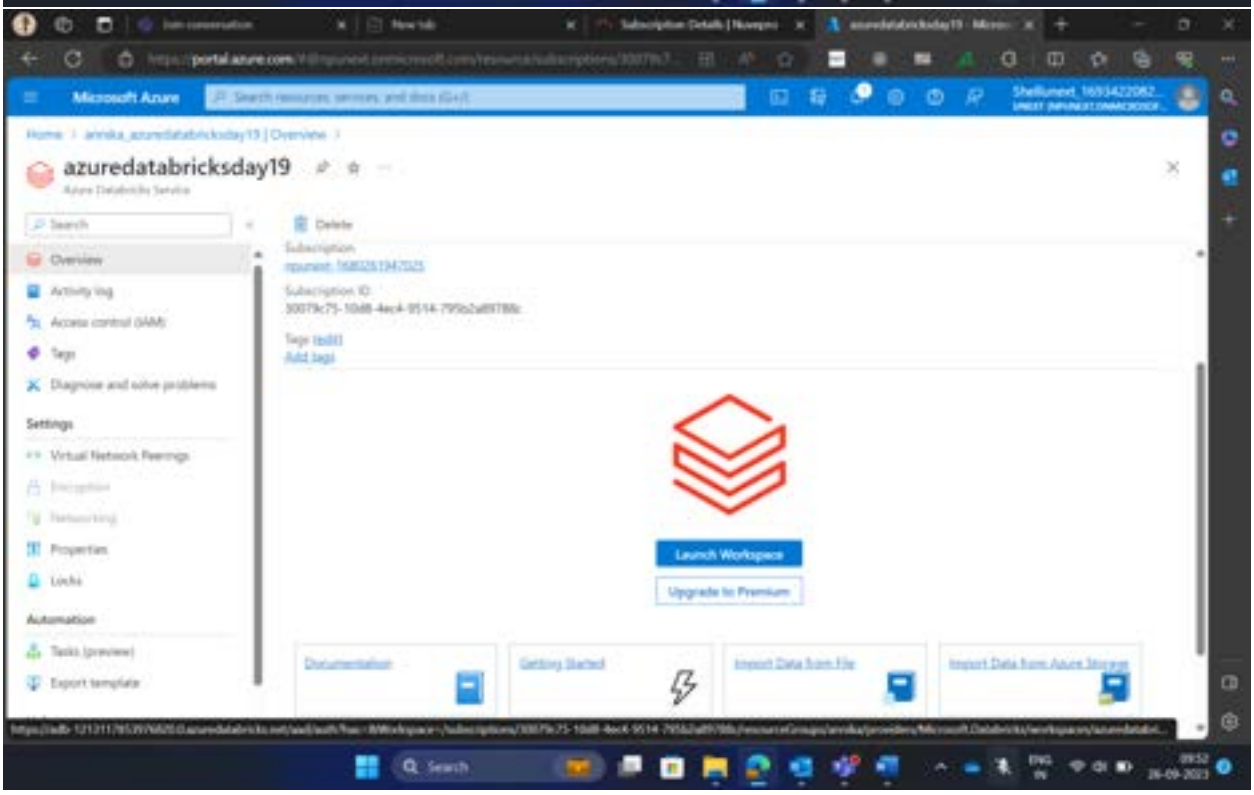
In [66]:
```
employeedf=spark.read.option("header","true").schema(EmpSchema).csv("/home/labuser/datasets/employee.csv")
```

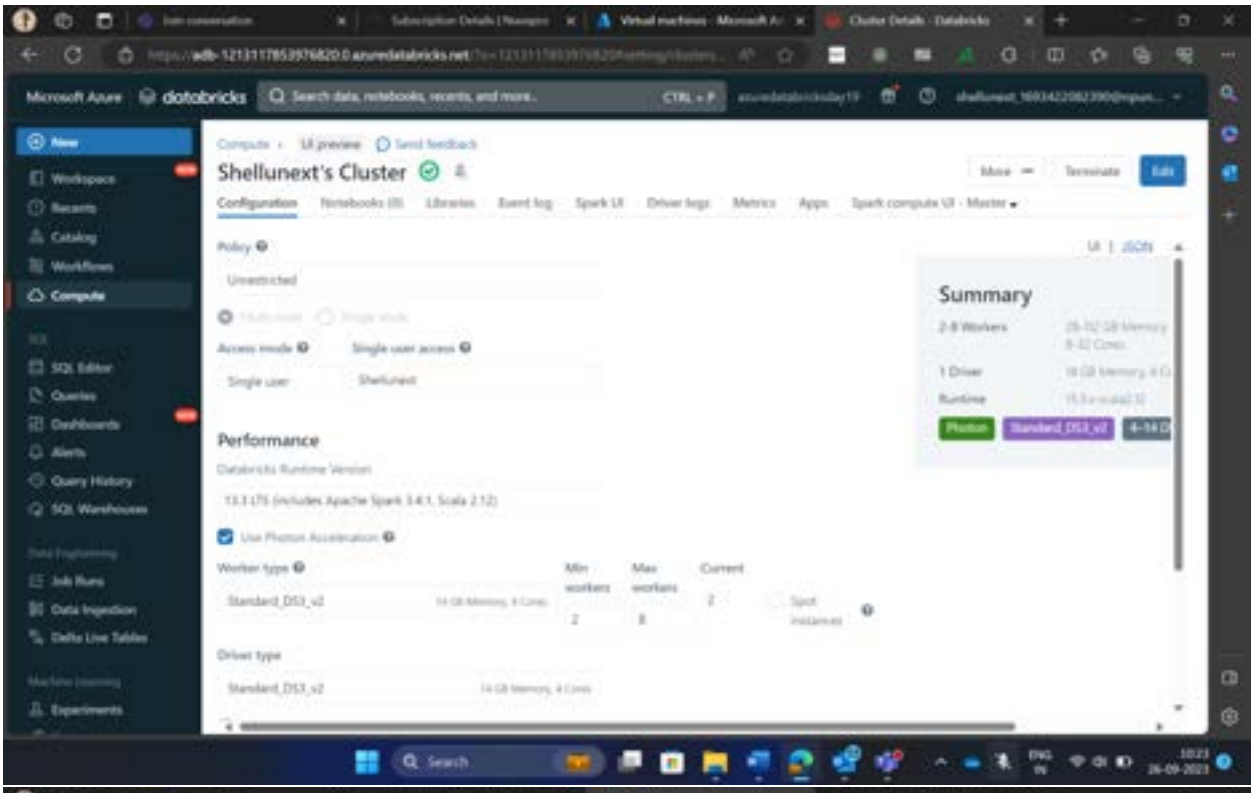
In [67]:
```
employeedf.dropna().show()
```

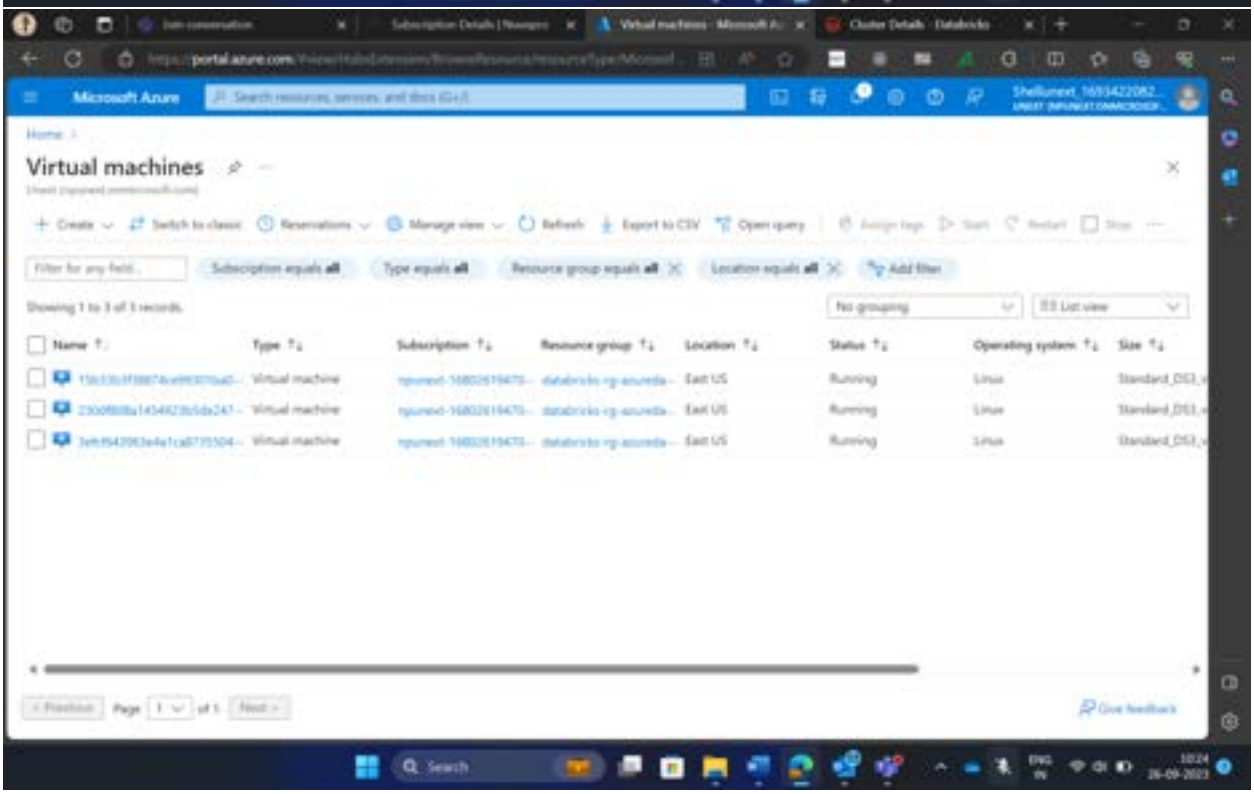
```
+-----+-------+------+
|Empno|Empname|salary|
+-----+-------+------+
|  111|    zzz|8000.0|
|  111|    aaa|8888.0|
|  121|    bbb|8000.0|
|  555|    eee|7000.0|
|  666|    fff|8890.0|
+-----+-------+------+
```

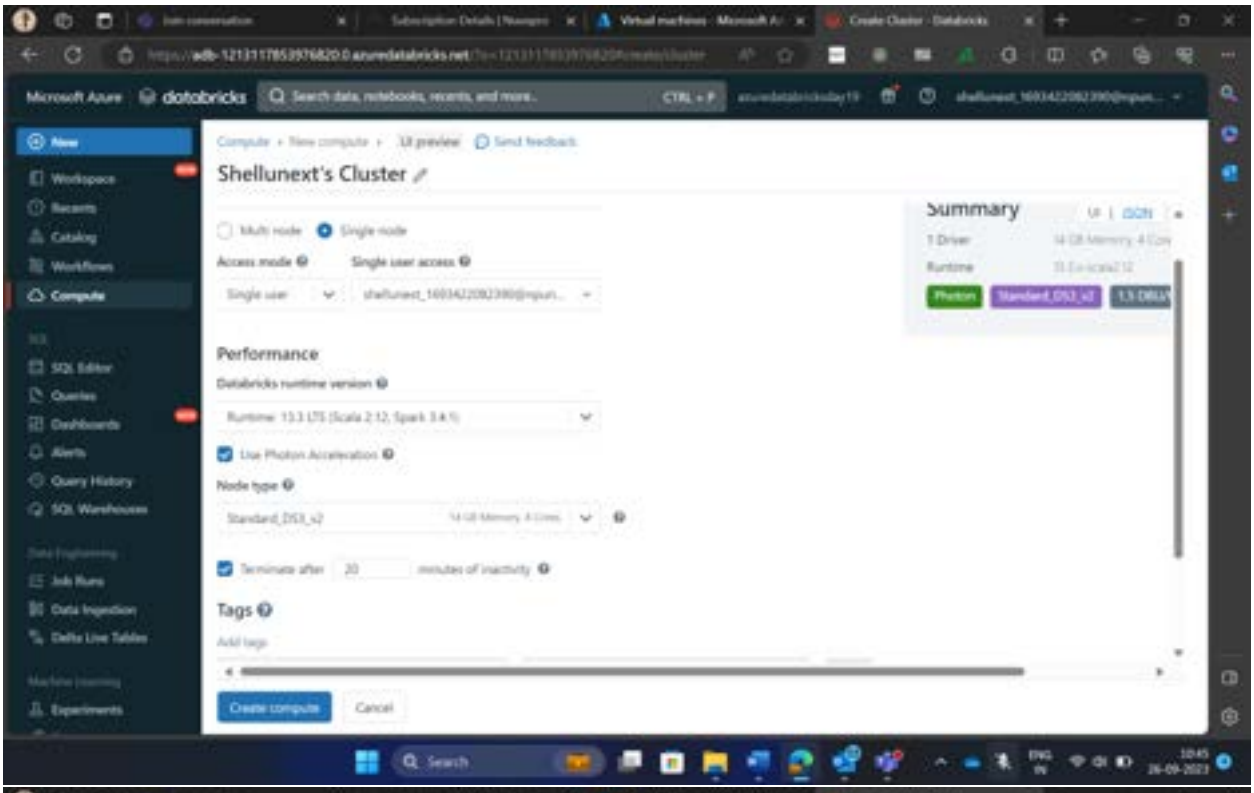
Day 19 (26.09.23)

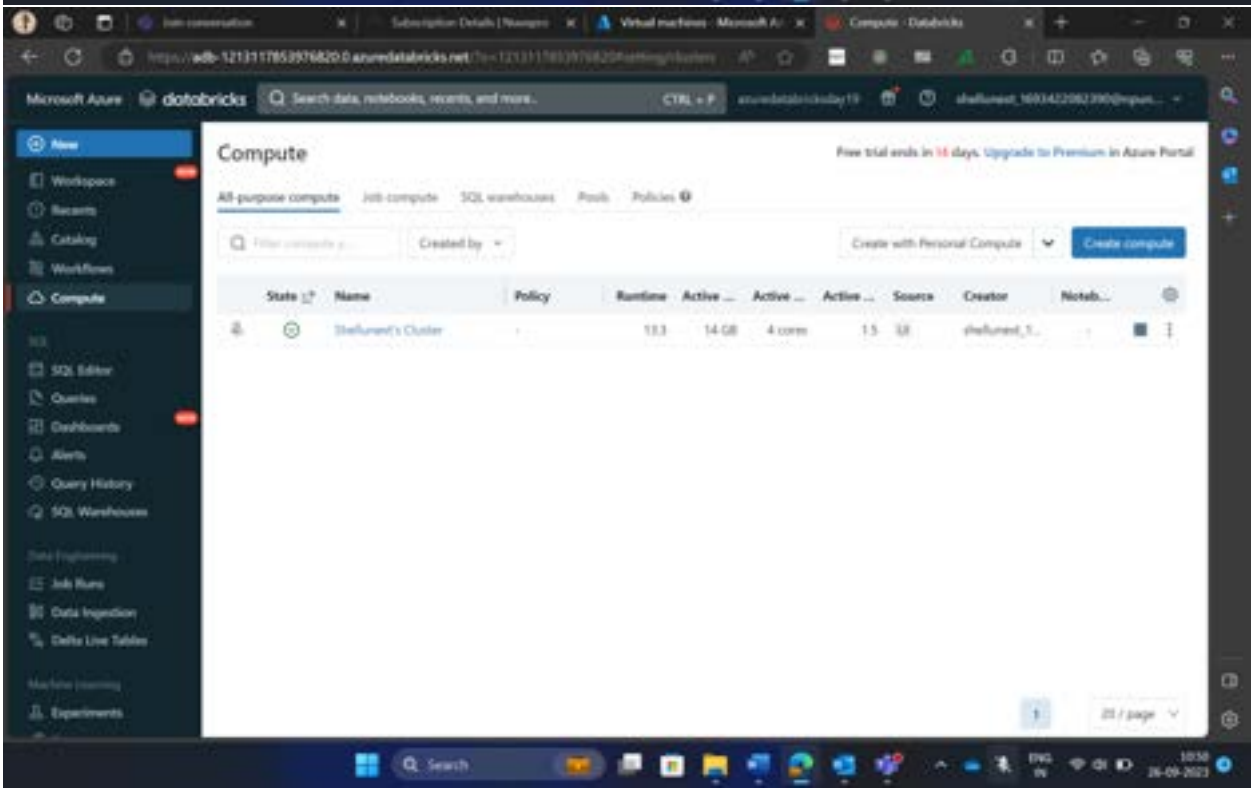


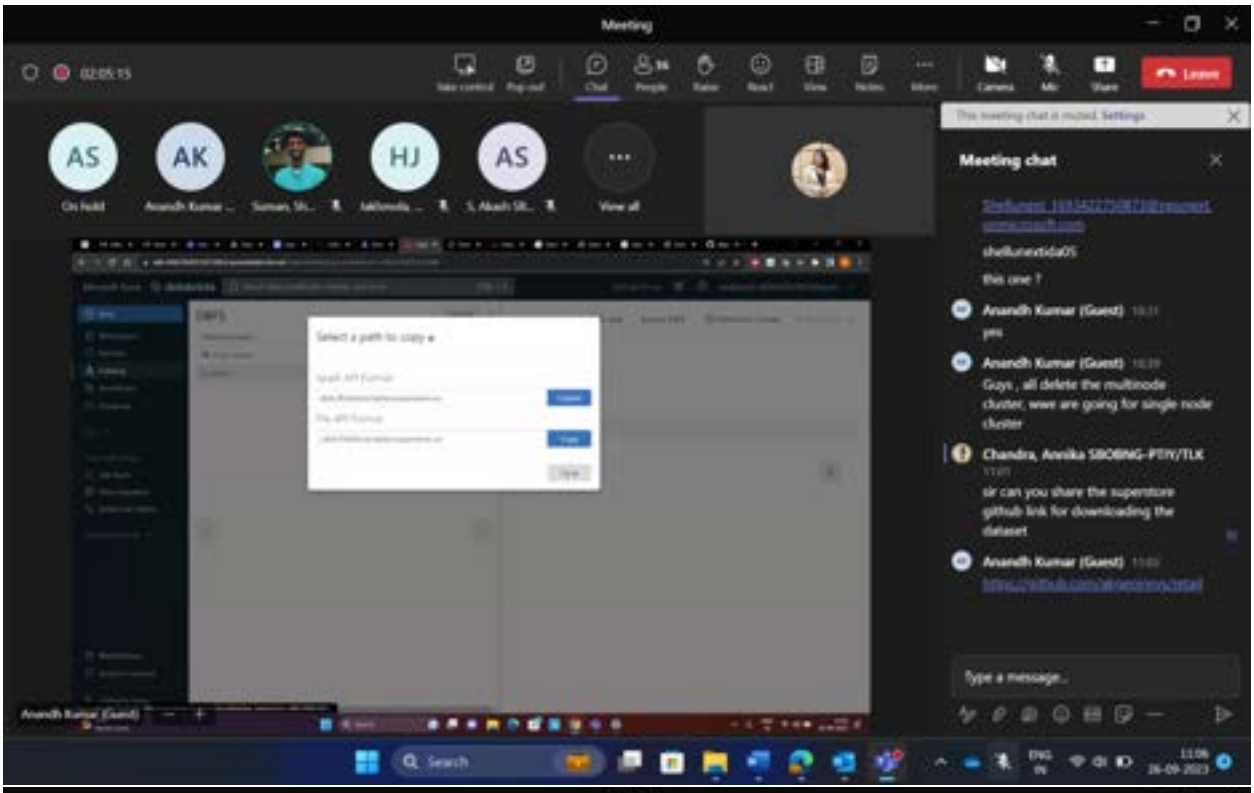


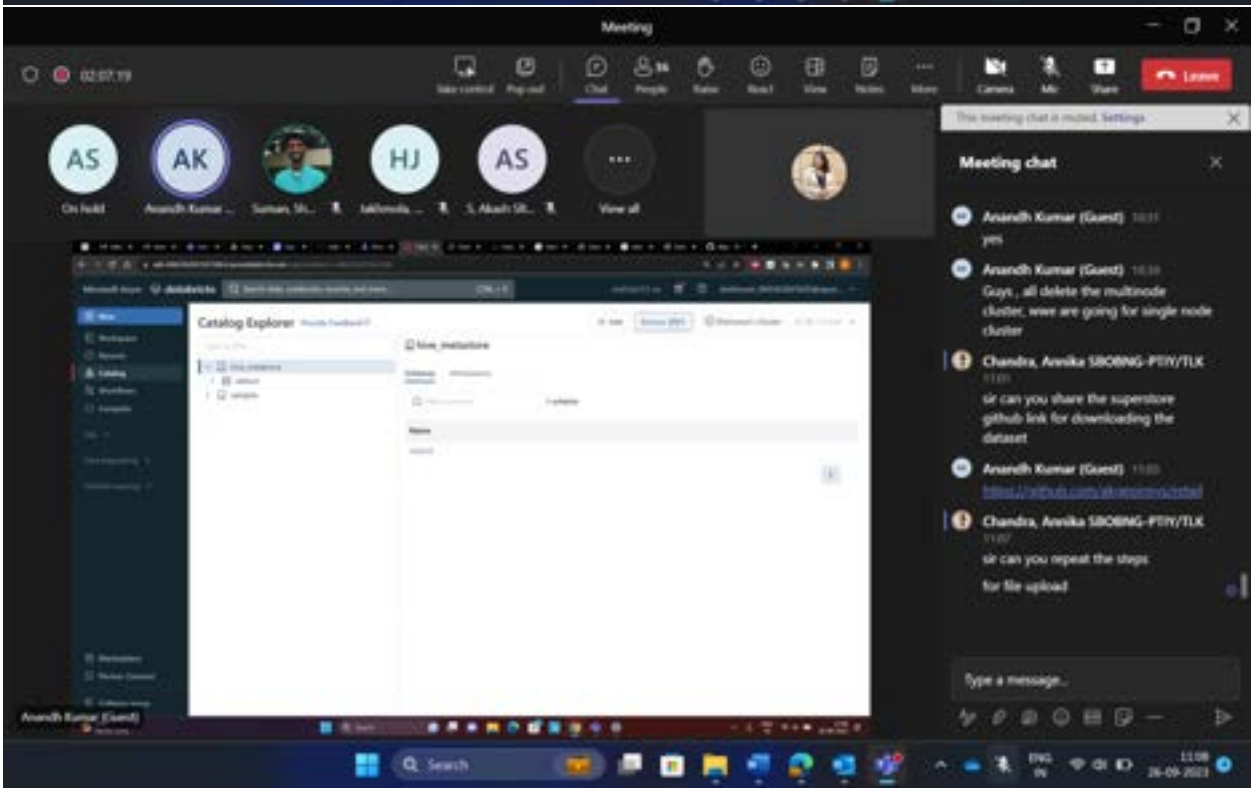

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
New
Workspace
Recents
Catalog
Workflows
Compute
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Job Runs
Data Ingestion
Delta Live Tables
Experiments
Compute
UI preview
Send feedback
Shellunext's Cluster
More
Terminate
Edit
Configuration
Notebooks (0)
Libraries
Event log
Spark UI
Driver logs
Metrics
Apps
Spark compute UI - Master
Policy
Unrestricted
Access mode
Single user access
Single user
Shellunext
Performance
Databricks Runtime Version
13.3 LTS (includes Apache Spark 3.4.1, Scala 2.12)
Use Photon Acceleration
Worker type
Standard_DS3_v2
Min workers
Max workers
Current
2
8
2
Spot instances
Driver type
Standard_DS3_v2
Summary
2-8 Workers
1 Driver
Runtime
Photon
Standard_DS3_v2
4-56
Search
ENG
IN
10:23
26-09-2023

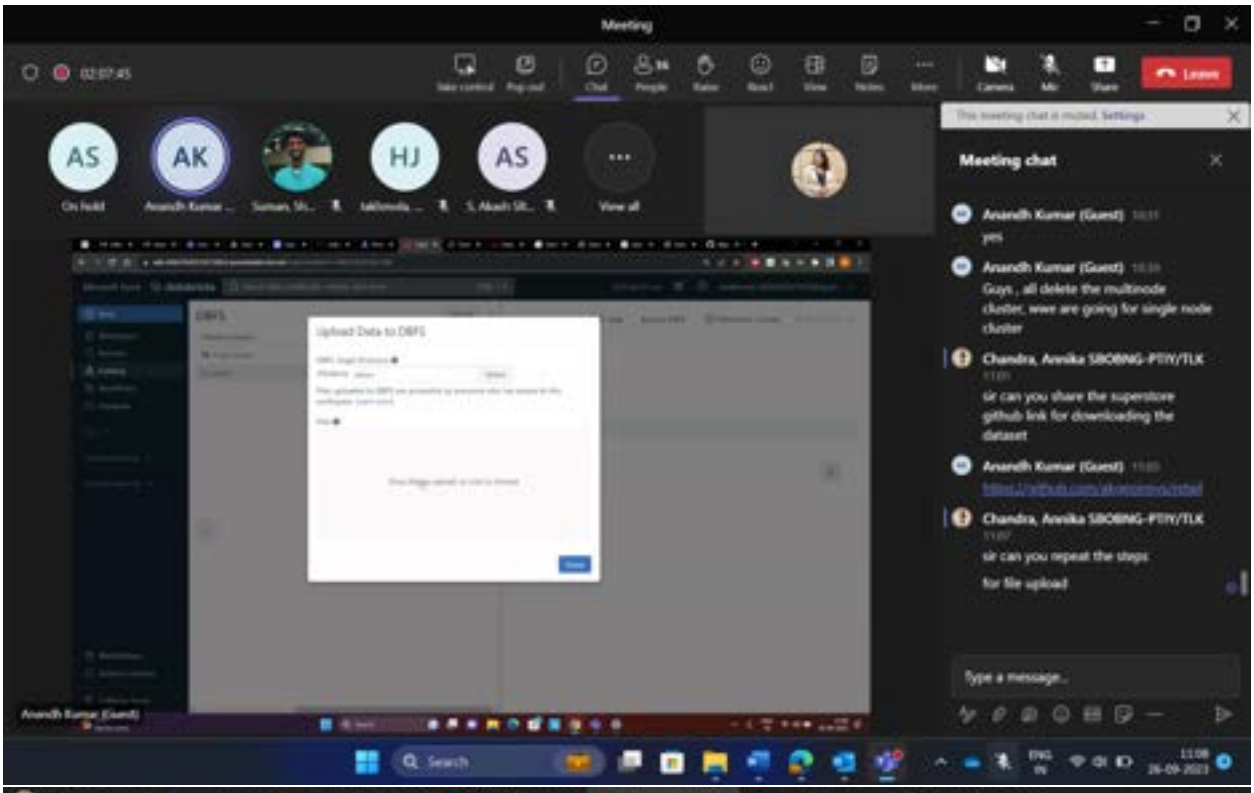

Microsoft Azure
Search resources, services, and docs (G+/)
Home
Virtual machines
Create
Switch to classic
Reservations
Manage view
Refresh
Export to CSV
Open query
Start
Restart
Stop
Filter for any field...
Subscription equals all
Type equals all
Resource group equals all
Location equals all
Add filter
Showing 1 to 3 of 3 records.
No grouping
List view
Name
Type
Subscription
Resource group
Location
Status
Operating system
Size
Virtual machine
East US
Running
Linux
Page
of 1
Next
Give feedback
Search
ENG
IN
10:24
26-09-2023

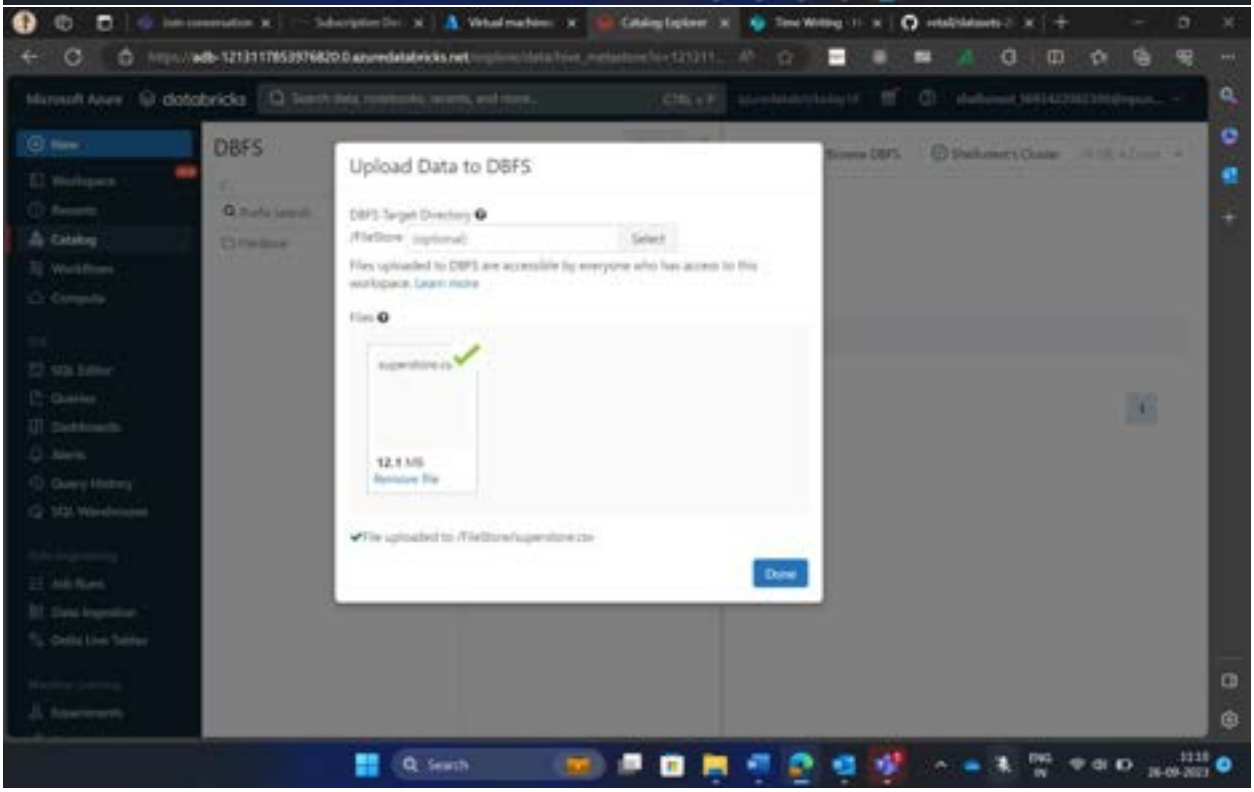

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
Shellunext's Cluster
Summary
Performance
Tags
Create compute
Cancel

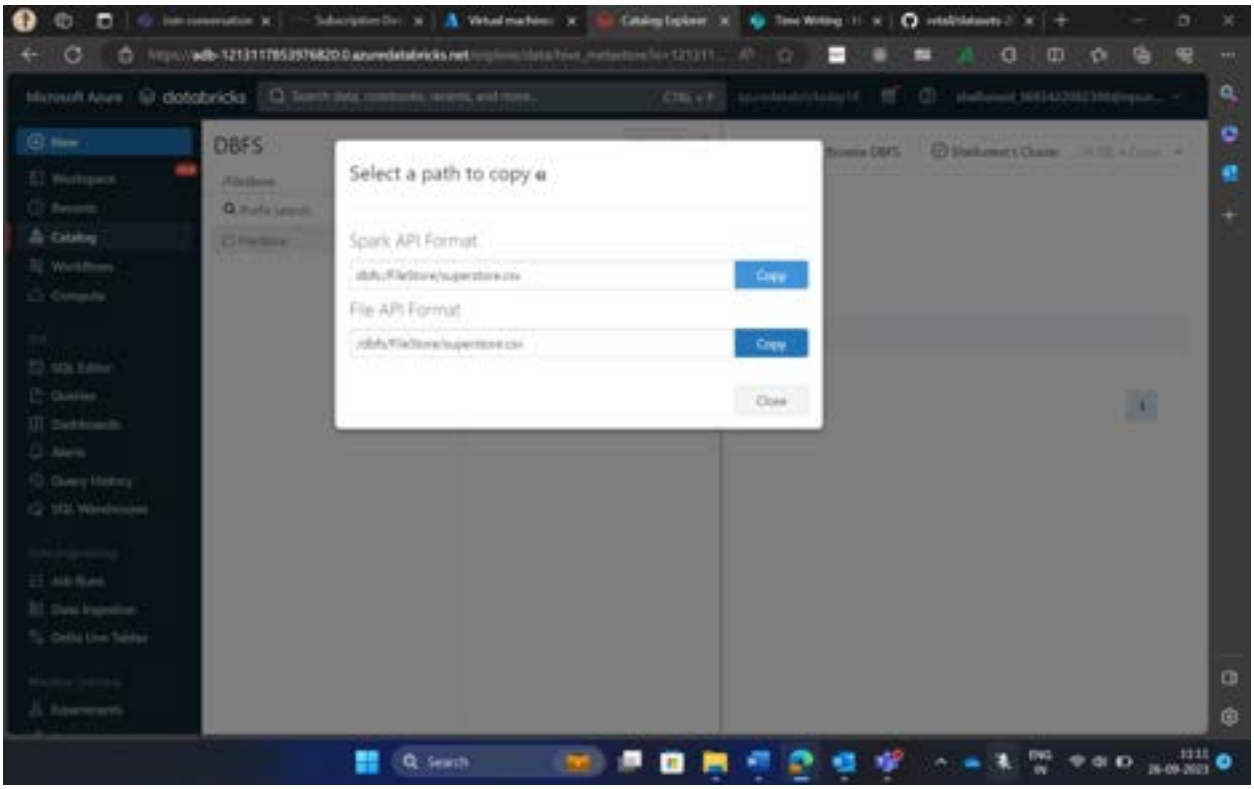

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
Compute
Free trial ends in 14 days. Upgrade to Premium in Azure Portal
All-purpose compute
Job compute
SQL warehouses
Pools
Policies
Create compute


Meeting
Meeting chat
Anandh Kumar (Guest)
yes
Anandh Kumar (Guest)
Guys , all delete the multinode cluster, wwe are going for single node cluster
Chandra, Annika SBOBNG-PTIY/TLK
sir can you share the superstore github link for downloading the dataset
Anandh Kumar (Guest)
Type a message...

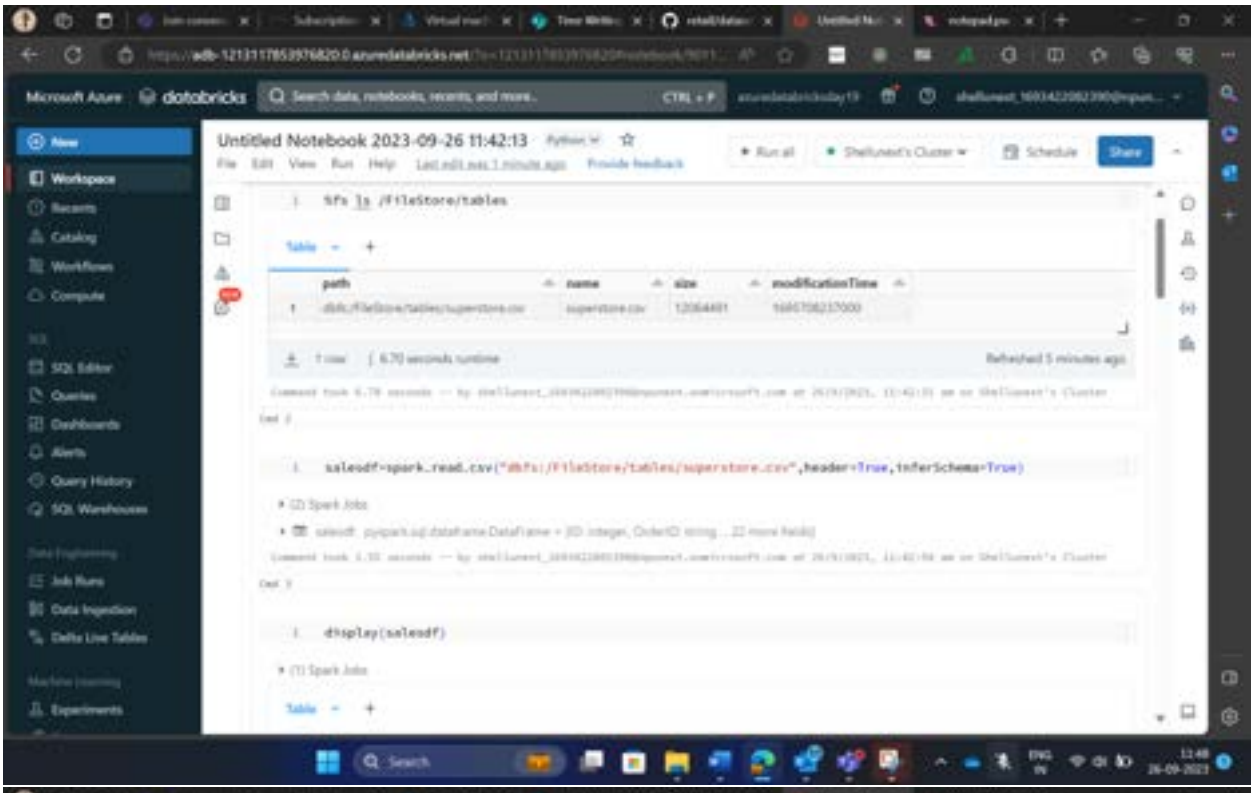

Meeting
Meeting chat
Anandh Kumar (Guest)
yes
Anandh Kumar (Guest)
Guys , all delete the multinode cluster, wwe are going for single node cluster
Chandra, Annika SBOBNG-PTIY/TLK
sir can you share the superstore github link for downloading the dataset
Anandh Kumar (Guest)
Chandra, Annika SBOBNG-PTIY/TLK
sir can you repeat the steps
for file upload
Type a message...

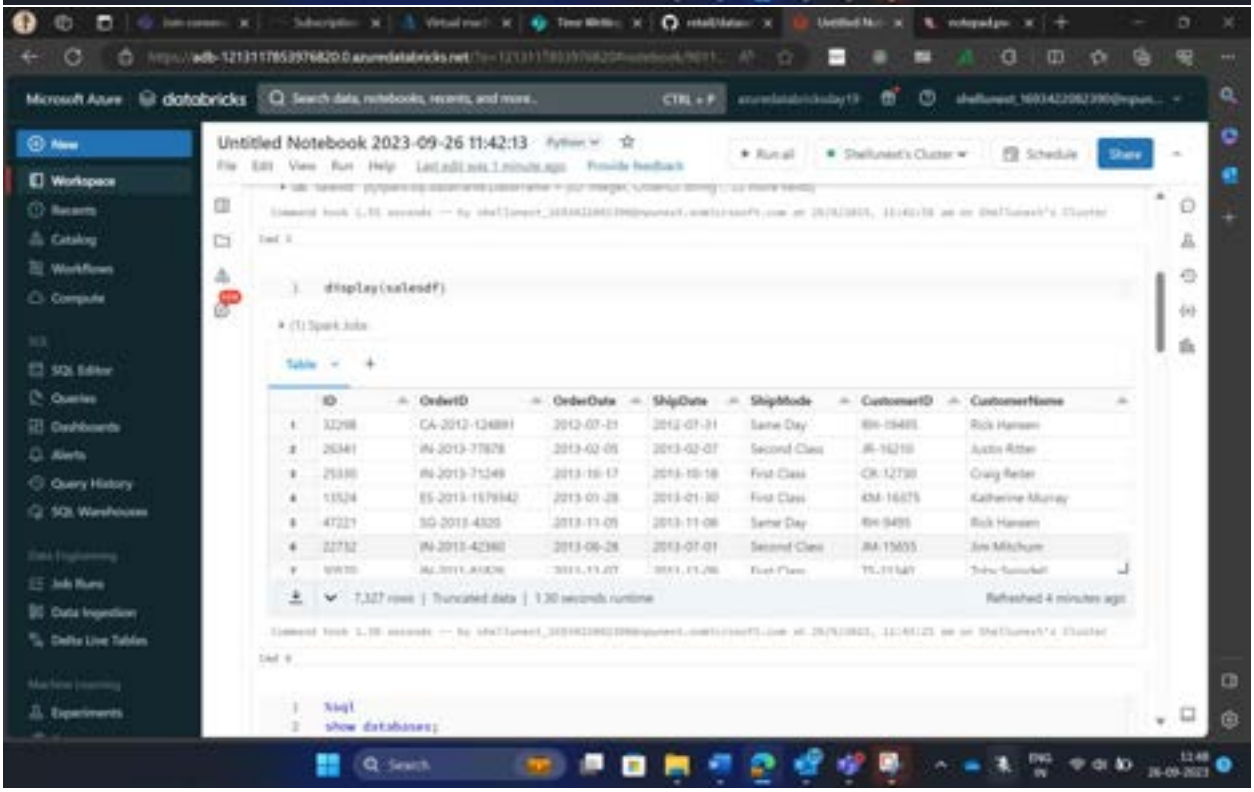

Meeting
Meeting chat
Anandh Kumar (Guest)
yes
Anandh Kumar (Guest)
Guys , all delete the multinode cluster, wwe are going for single node cluster
Chandra, Anvika SBOBNG-PTIY/TLK
sir can you share the superstore github link for downloading the dataset
Anandh Kumar (Guest)
Chandra, Anvika SBOBNG-PTIY/TLK
sir can you repeat the steps
for file upload
Type a message...
Upload Data to DBFS

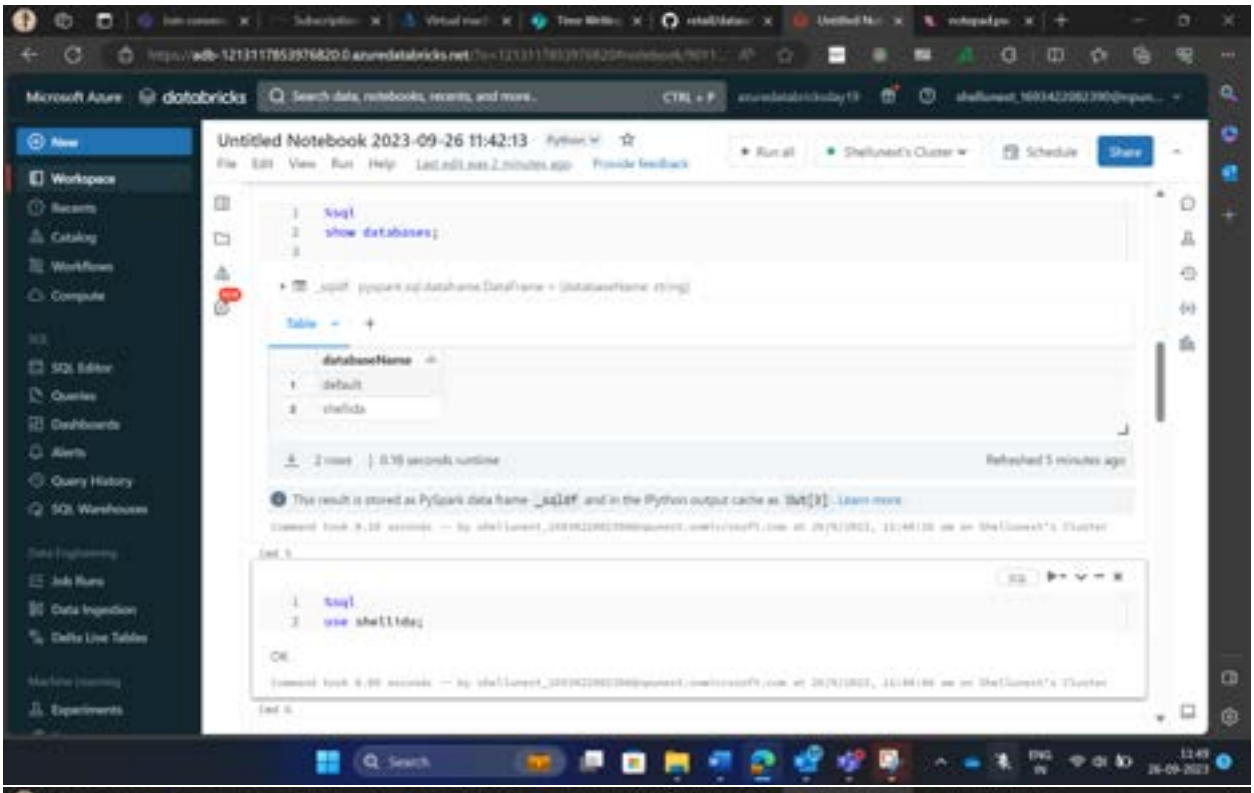

Microsoft Azure
databricks
DBFS
Upload Data to DBFS
DBFS Target Directory
/FileStore (optional)
Select
Files uploaded to DBFS are accessible by everyone who has access to this workspace. Learn more
Files
superstore.csv
12.1 MB
Remove file
File uploaded to /FileStore/superstore.csv
Done

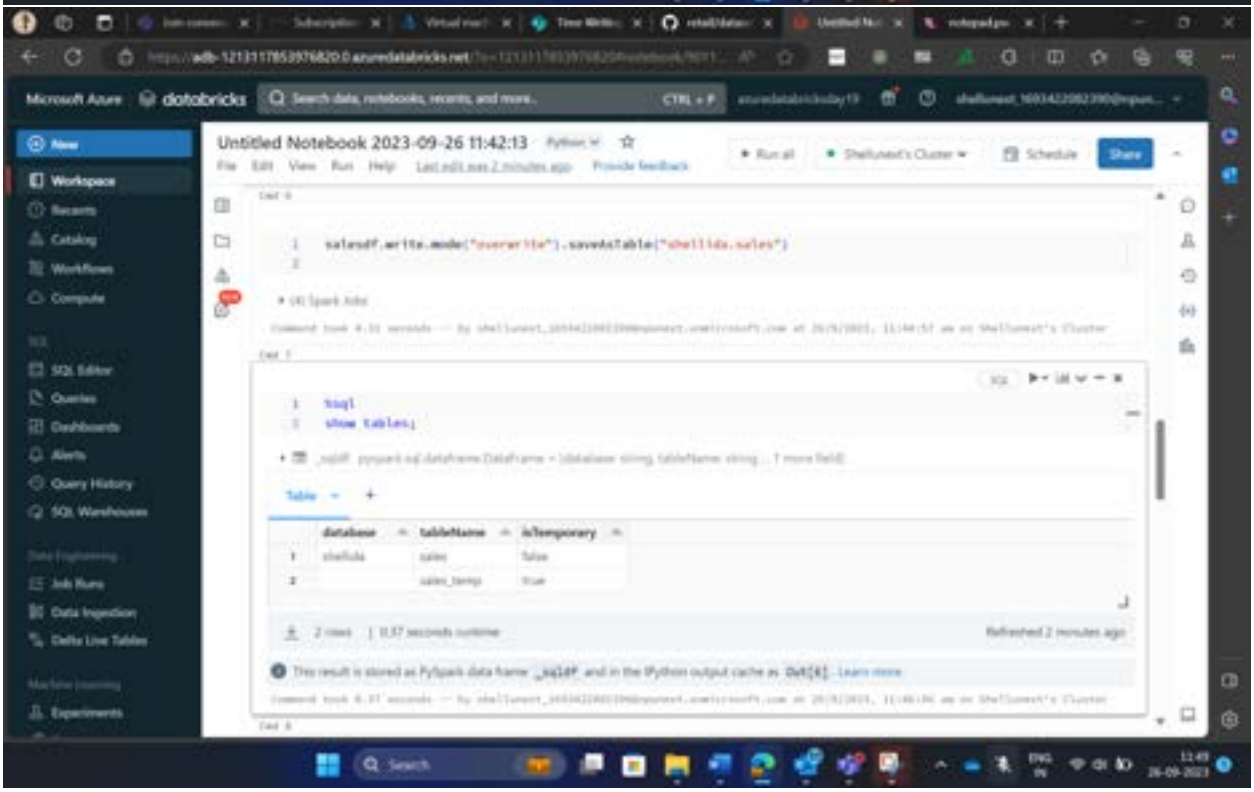


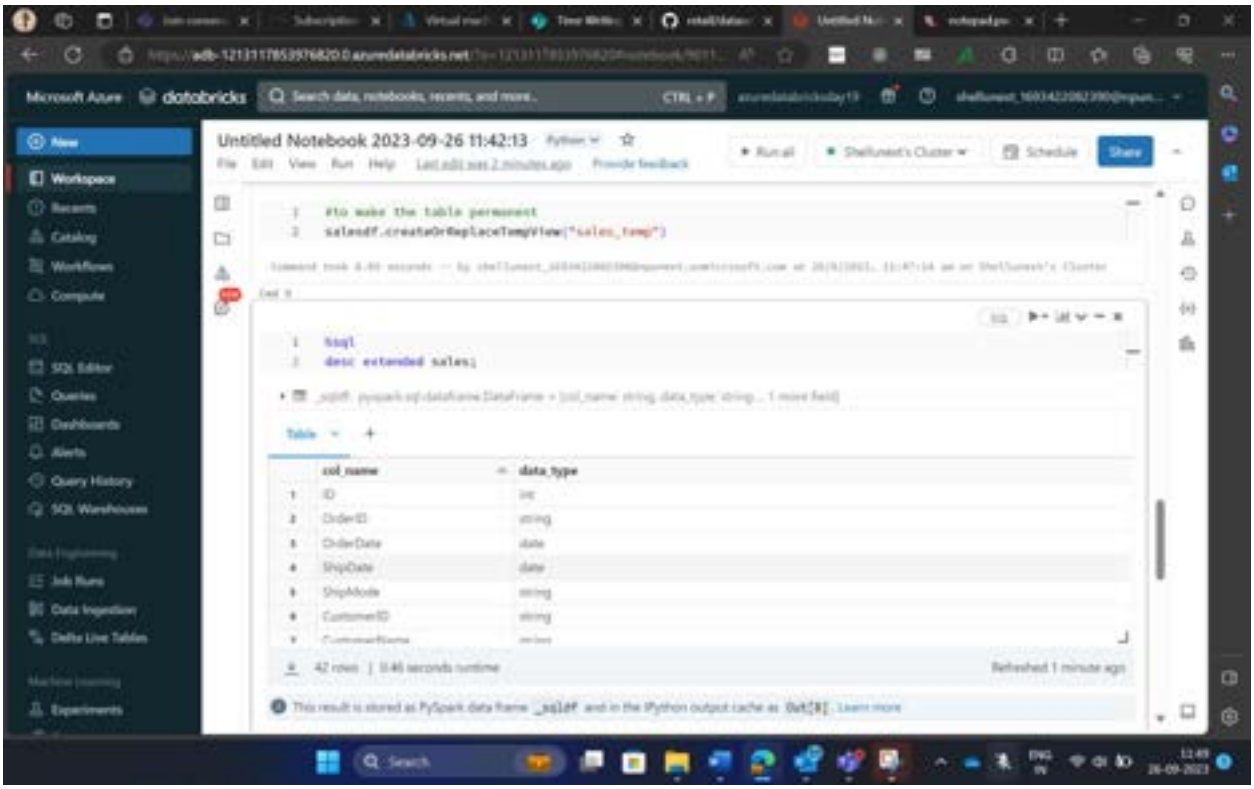
dbfs:/FileStore/superstore.csv


Microsoft Azure
databricks
Untitled Notebook 2023-09-26 11:42:13
%fs ls /FileStore/tables
path
name
size
modificationTime
saleodf=spark.read.csv("dbfs:/FileStore/tables/superstore.csv",header=True,inferSchema=True)
display(saleodf)

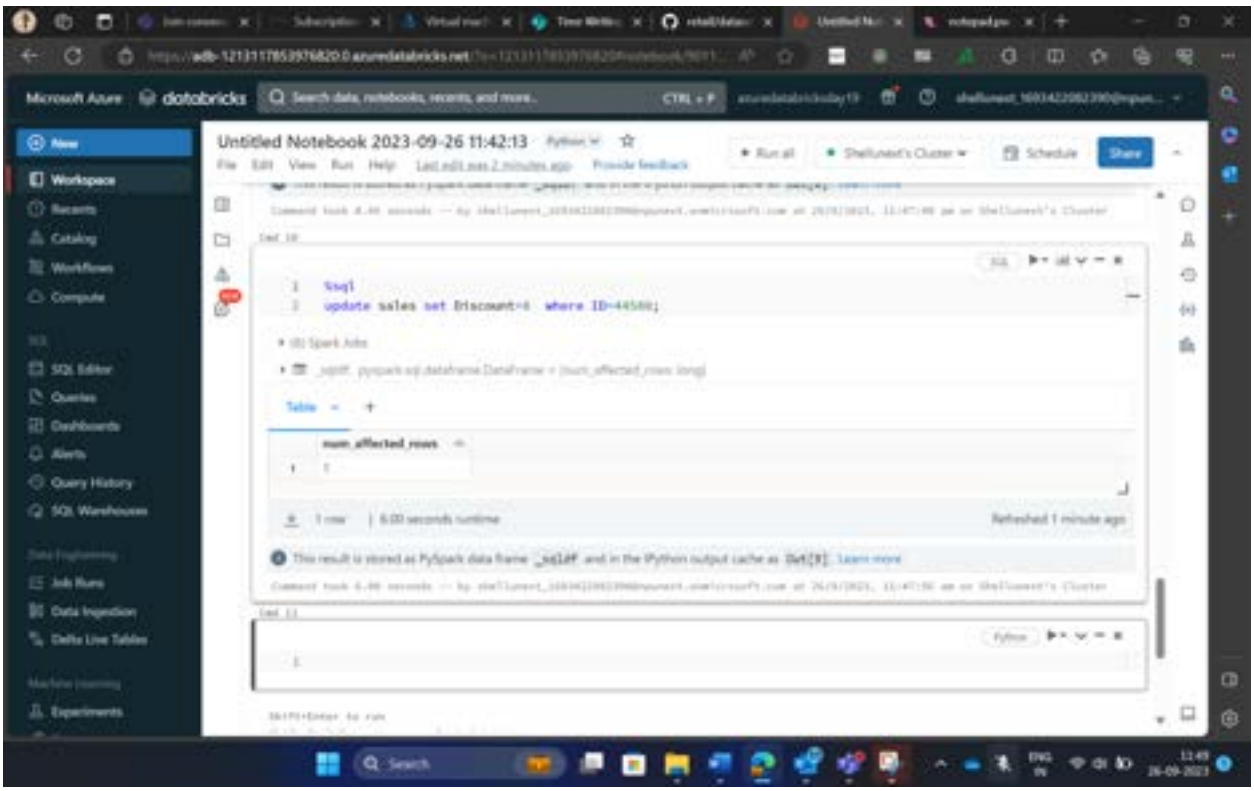

Microsoft Azure
databricks
Untitled Notebook 2023-09-26 11:42:13
display(saleodf)
ID
OrderID
OrderDate
ShipDate
ShipMode
CustomerID
CustomerName
%sql
show databases;

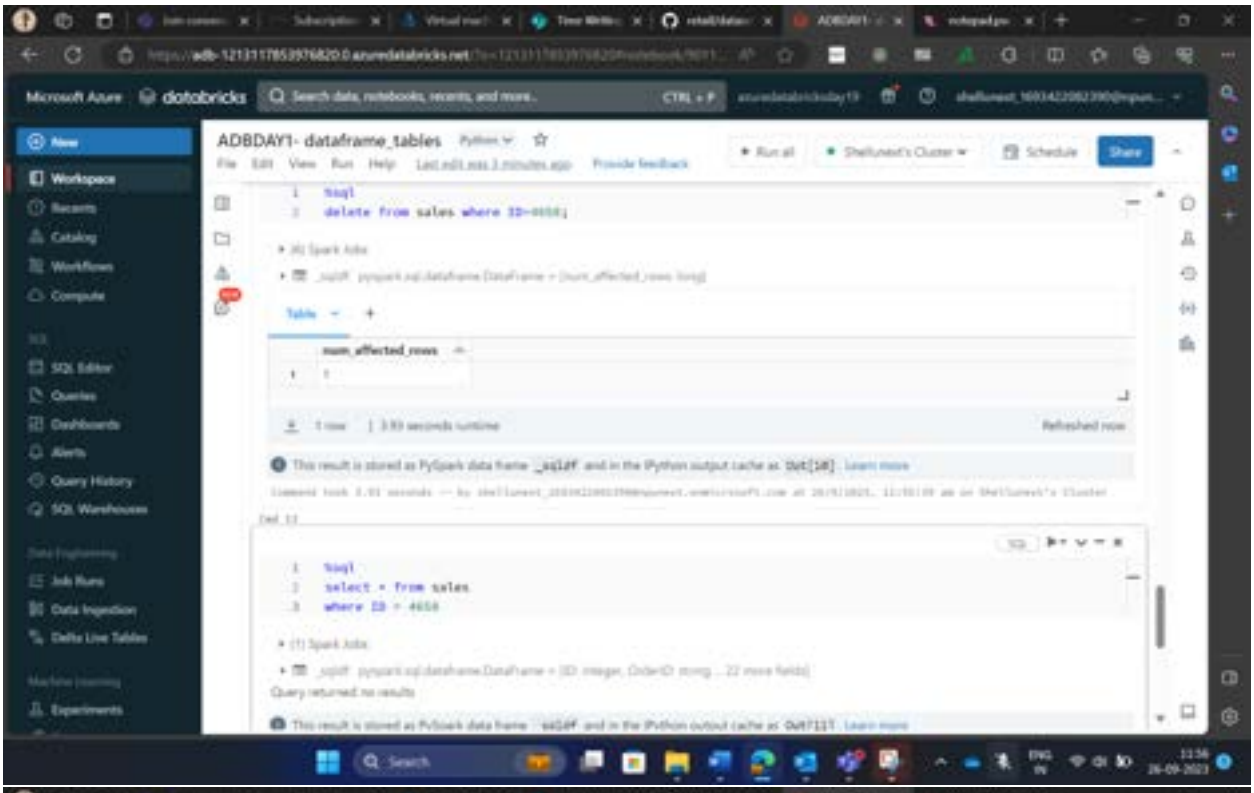

Microsoft Azure
databricks
Untitled Notebook 2023-09-26 11:42:13
Run all
Schedule
Share
%sql
show databases;
databaseName
default
shellida
2 rows | 0.18 seconds runtime
Refreshed 5 minutes ago
%sql
use shellida;
OK

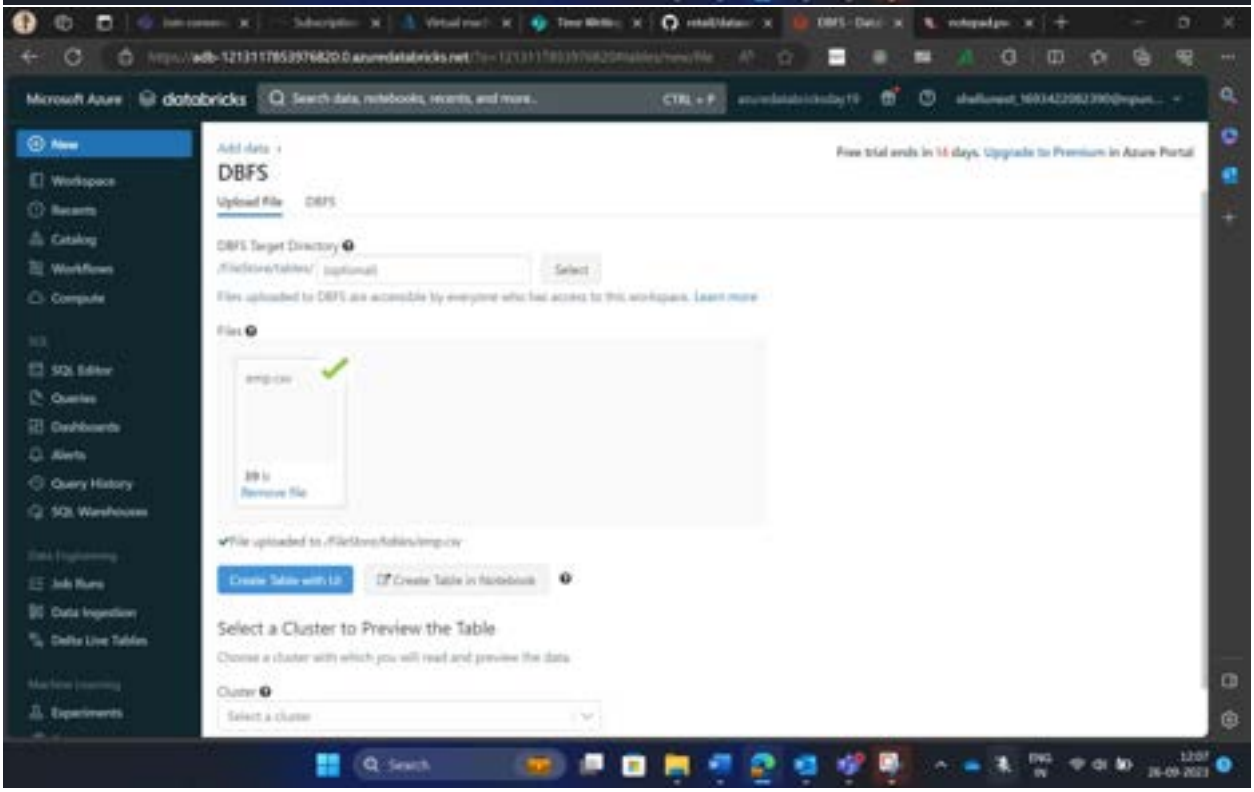

Microsoft Azure
databricks
Untitled Notebook 2023-09-26 11:42:13
Run all
Schedule
Share
salesdf.write.mode("overwrite").saveAsTable("shellida.sales")
%sql
show tables;
database
tableName
isTemporary
shellida
sales
false
sales_temp
true
2 rows | 0.37 seconds runtime
Refreshed 2 minutes ago

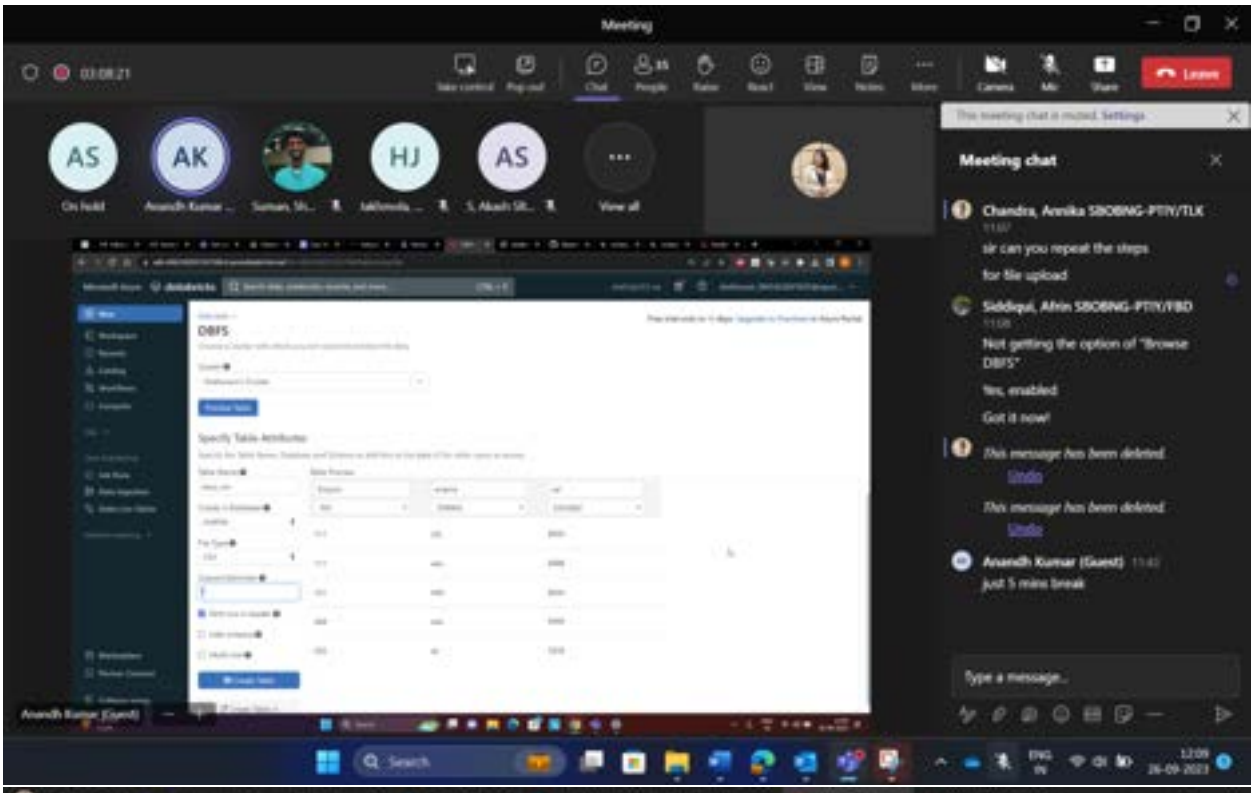

Microsoft Azure
databricks
Untitled Notebook 2023-09-26 11:42:13
#to make the table permanent
salesdf.createOrReplaceTempView("sales_temp")
%sql
desc extended sales;
col_name
data_type
ID
OrderID
OrderDate
ShipDate
ShipMode
CustomerID
42 rows | 0.46 seconds runtime
Refreshed 1 minute ago

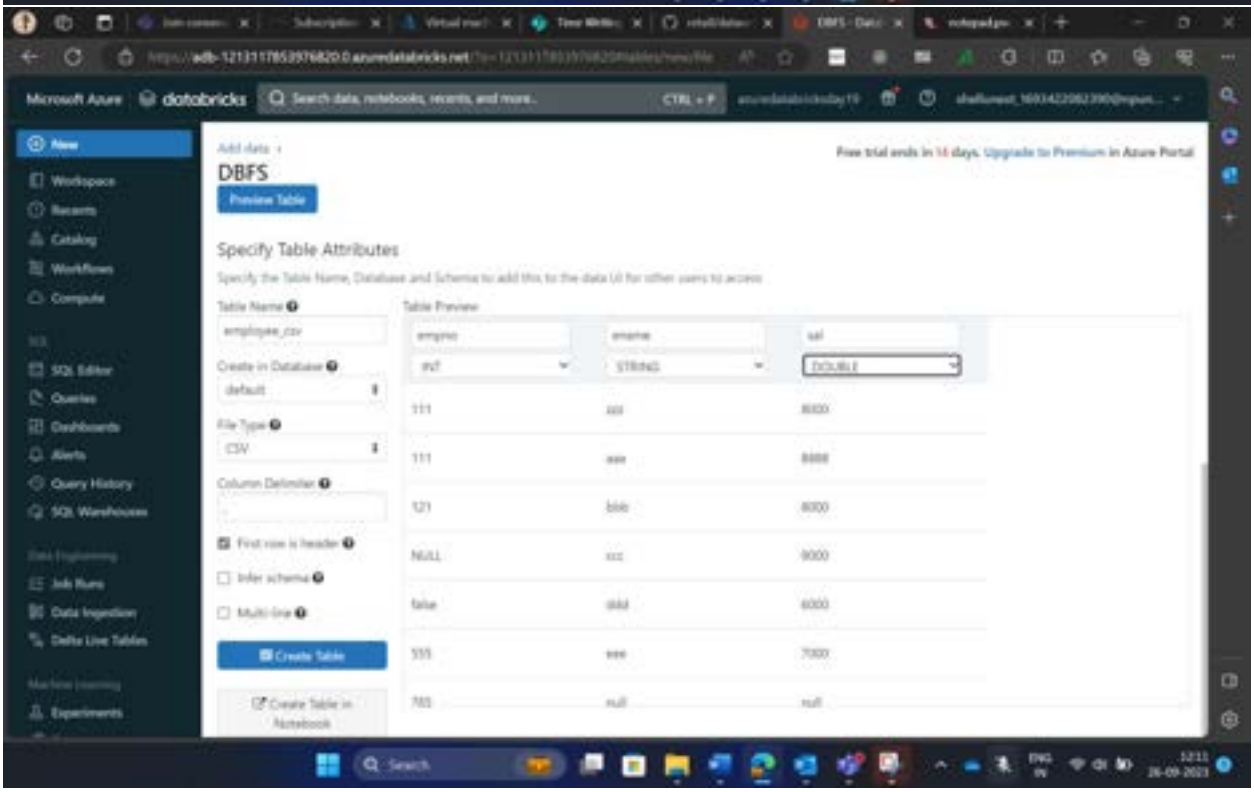
v

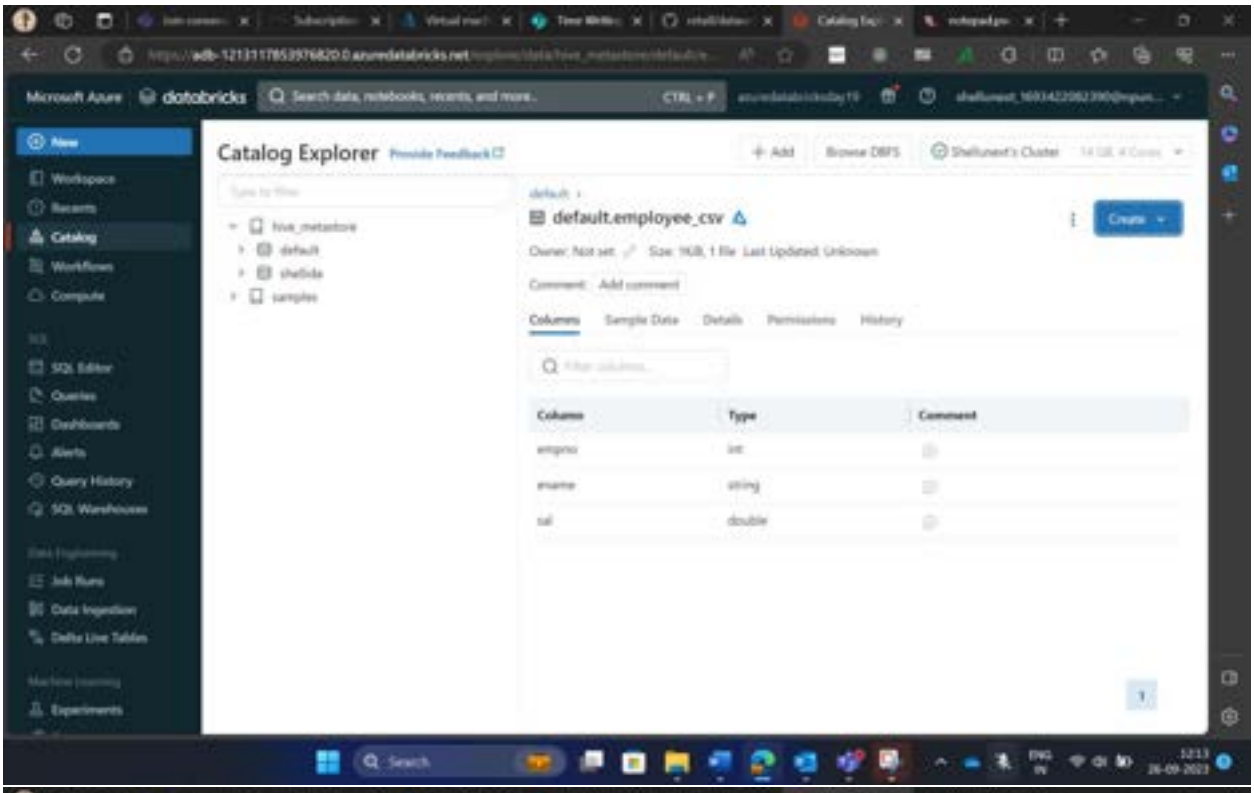

Microsoft Azure
databricks
Untitled Notebook 2023-09-26 11:42:13
%sql
update sales set Discount=0 where ID=44500;
num_affected_rows
1 row | 6.00 seconds runtime
Refreshed 1 minute ago

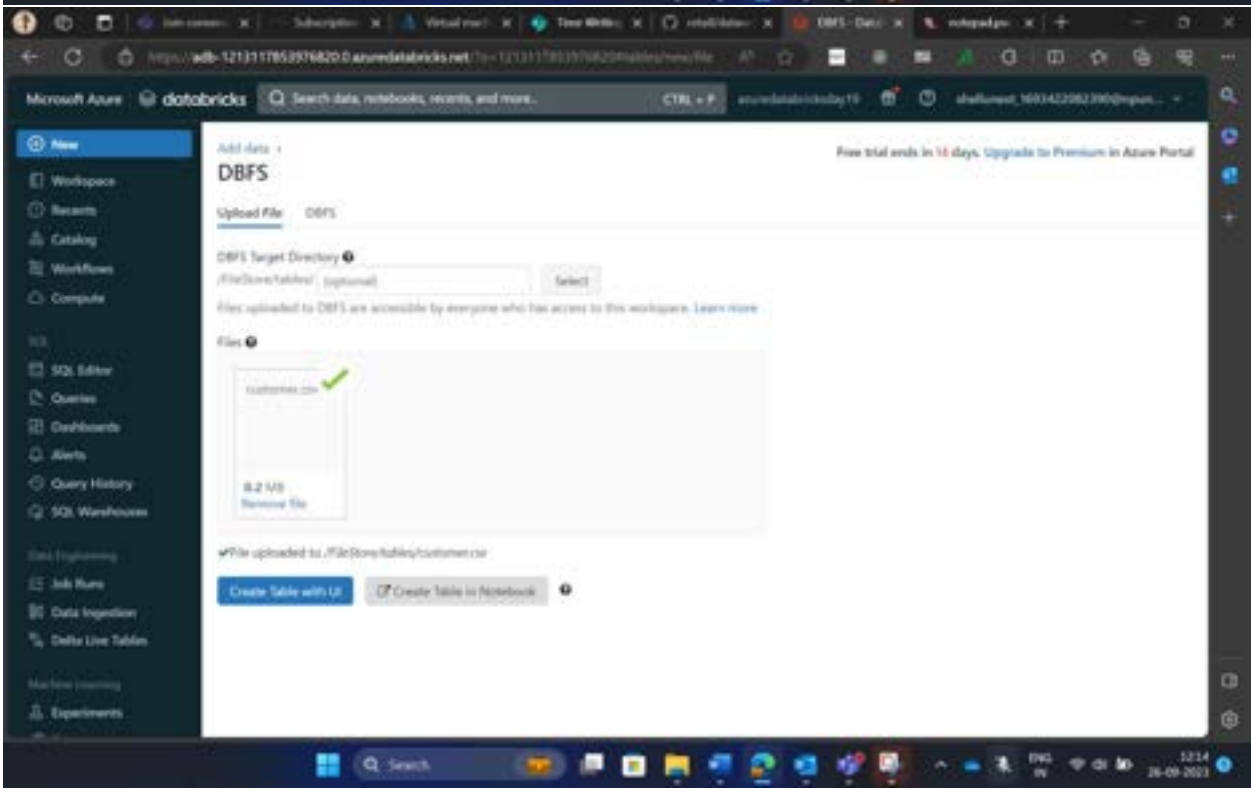

Microsoft Azure
databricks
ADBDAY1- dataframe_tables
Run all
Schedule
Share
num_affected_rows
Query returned no results

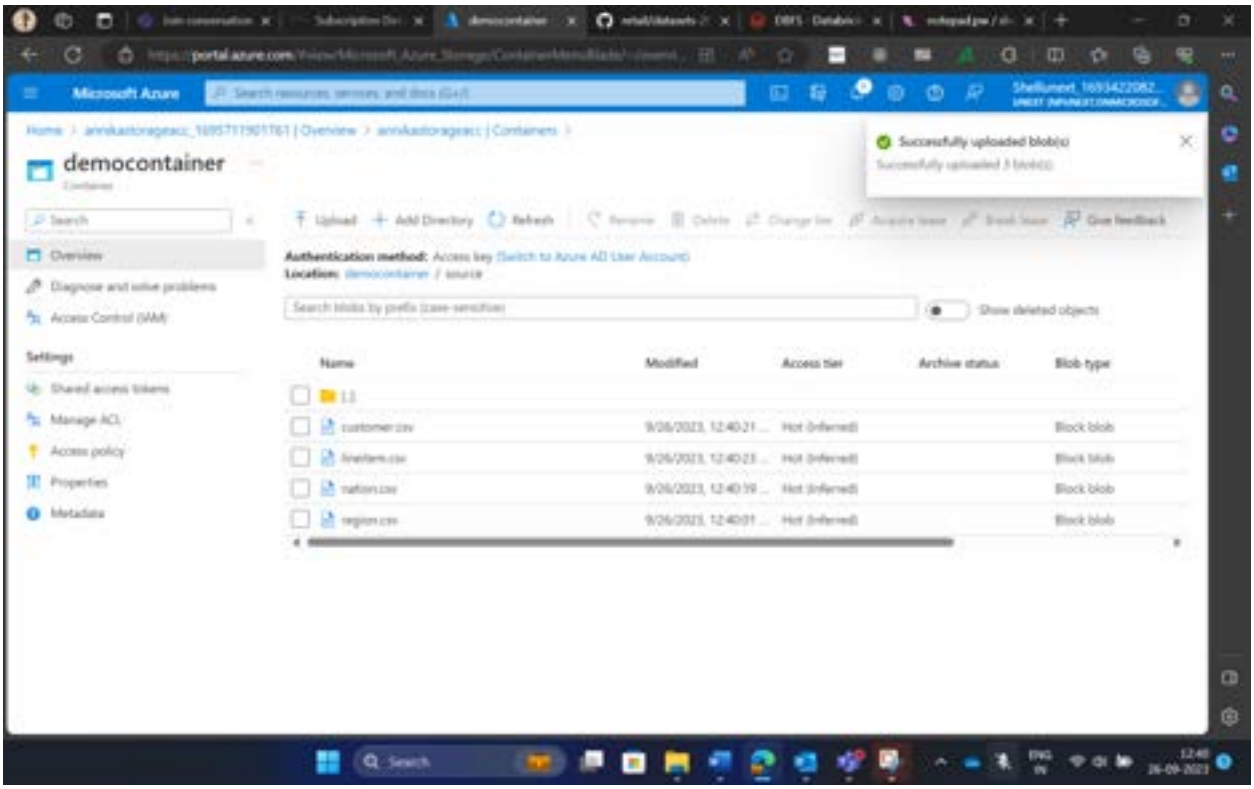

Microsoft Azure
databricks
DBFS
Upload File
DBFS Target Directory
Select
Create Table with UI
Create Table in Notebook
Select a Cluster to Preview the Table
Cluster

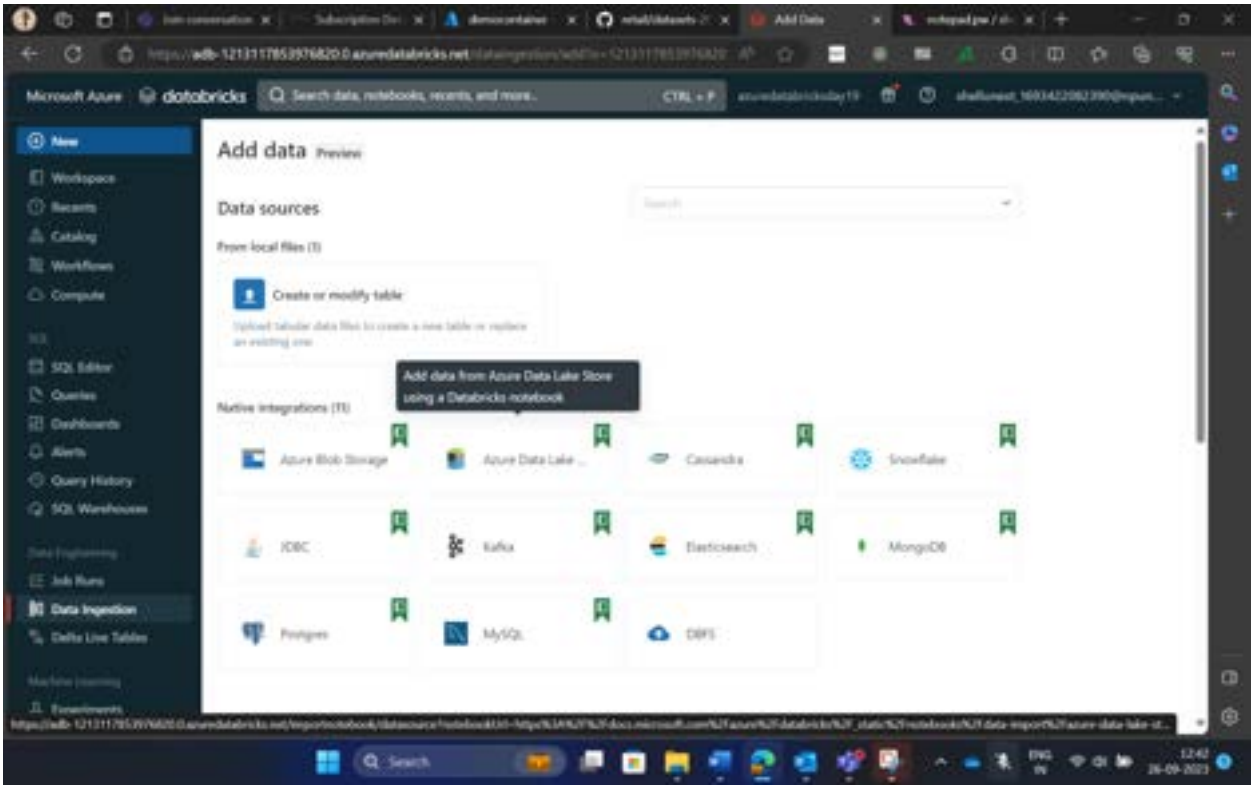

Meeting
Meeting chat
Chandra, Annika SBOBNG-PTIY/TLK
sir can you repeat the steps
for file upload
Siddiqui, Afrin SBOBNG-PTIY/FBD
Not getting the option of "Browse DBFS"
Yes, enabled
Got it now!
This message has been deleted
Undo
This message has been deleted
Undo
Anandh Kumar (Guest)
just 5 mins break
Type a message...
DBFS


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
Free trial ends in 14 days. Upgrade to Premium in Azure Portal
DBFS
Preview Table
Specify Table Attributes
Table Name
employee_csv
Create in Database
default
File Type
CSV
Column Delimiter
First row is header
Infer schema
Multi-line
Create Table
Create Table in Notebook
Table Preview

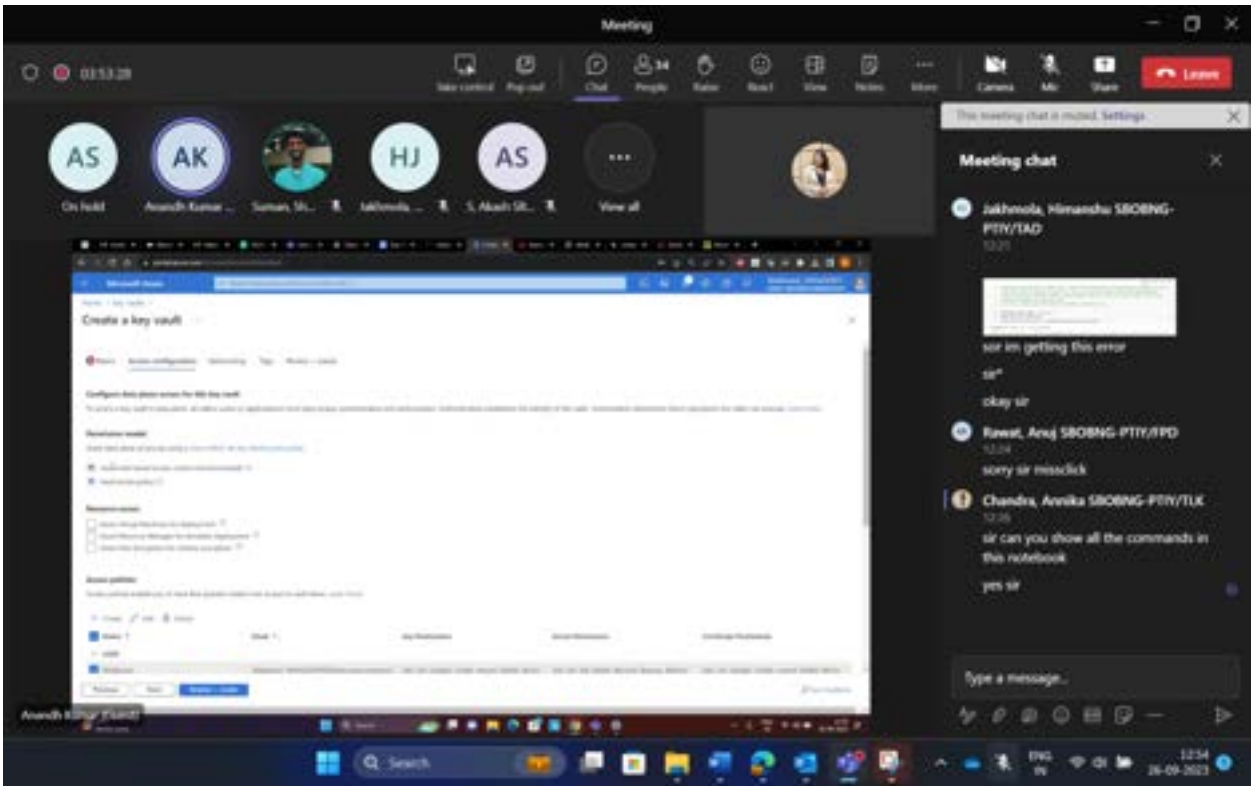

Microsoft Azure
databricks
Catalog Explorer
default.employee_csv
Columns
Sample Data
Details
Permissions
History
Column
Type
Comment
empno
int
ename
string
sal
double

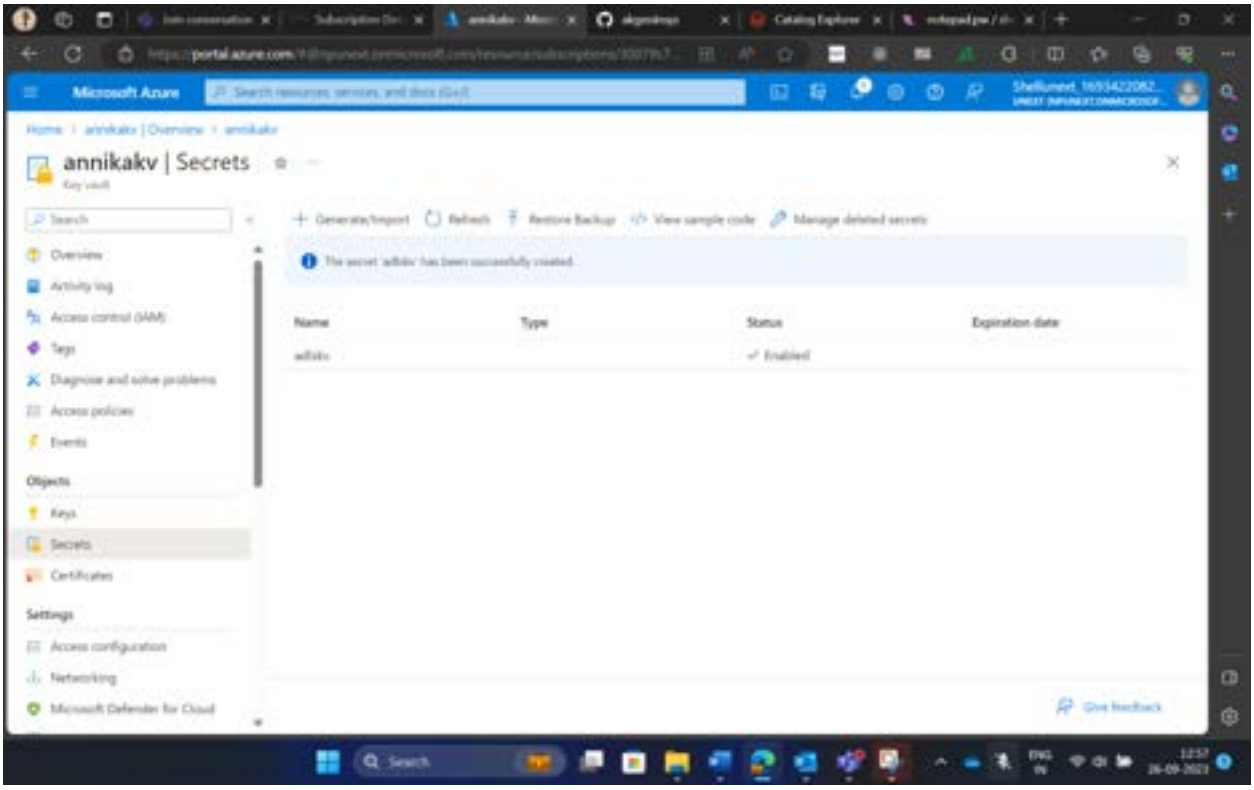

Microsoft Azure
databricks
DBFS
Upload File
DBFS
DBFS Target Directory
Select
Create Table with UI
Create Table in Notebook


democontainer
Successfully uploaded blob(s)
Upload
Add Directory
Refresh
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Name
Modified
Access tier
Archive status
Blob type


Add data
Preview
Data sources
From local files (1)
Create or modify table
Native integrations (11)
Add data from Azure Data Lake Store using a Databricks notebook
Azure Blob Storage
Azure Data Lake ...
Cassandra
Snowflake
JDBC
Kafka
Elasticsearch
MongoDB
Postgres
MySQL
DBFS

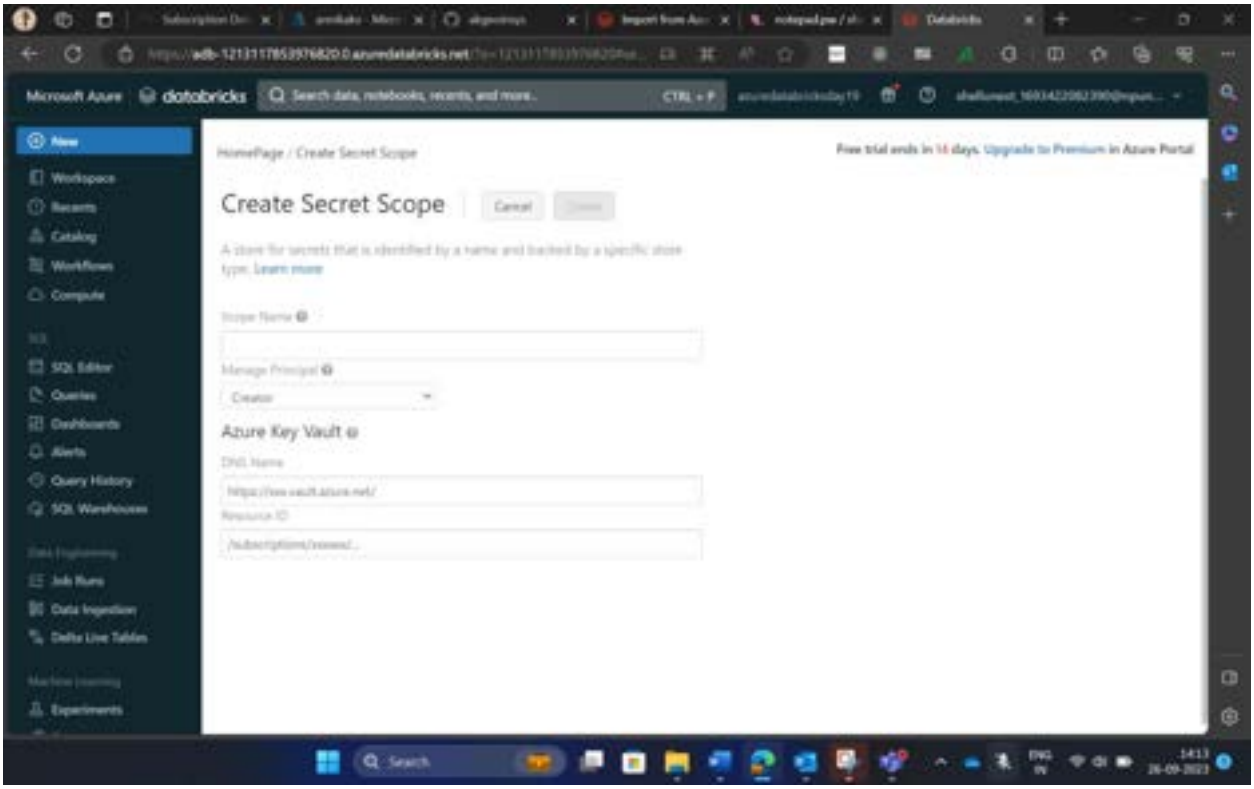
XDmK6b6u/sxVQuV5/KHgC/Gwm4tG6DCVIj7Swnl0IXdRkktcdCmsHTtx7FH9QpdsDiQaZIIOHvSC+AStk5PjfA==

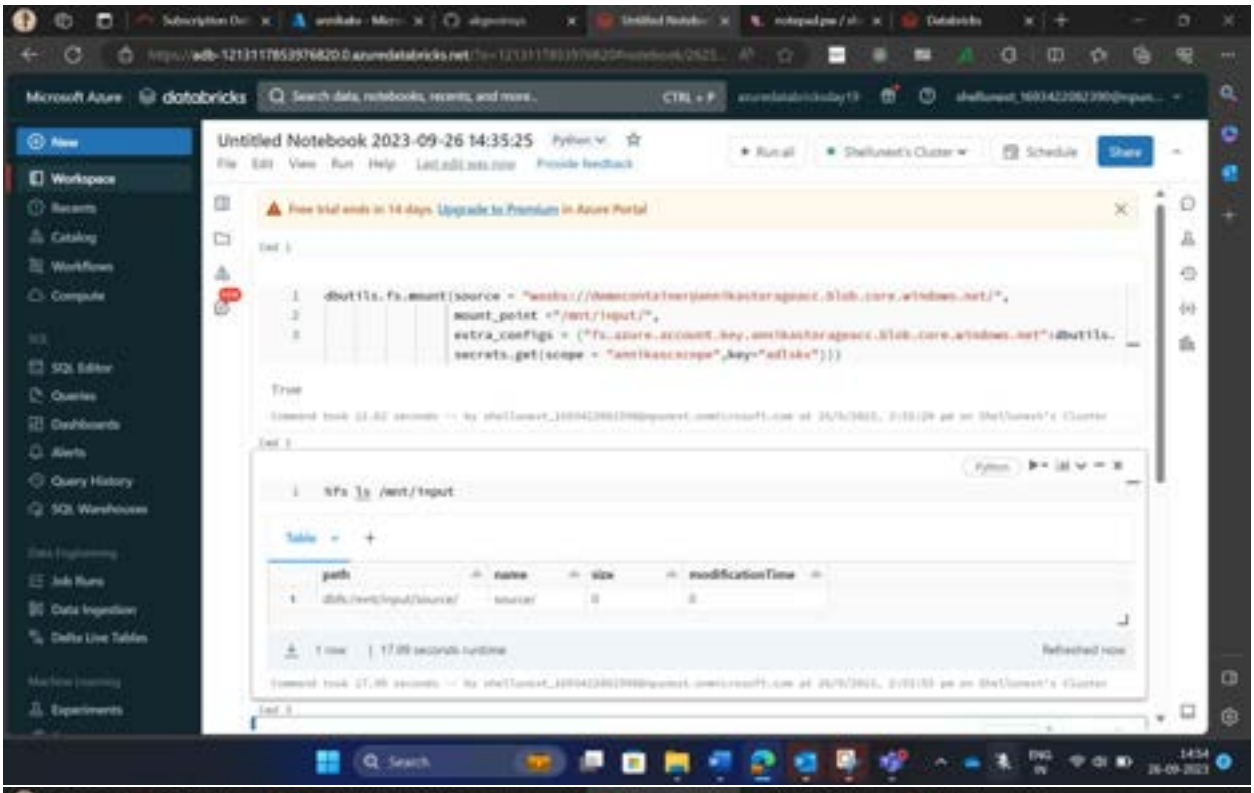


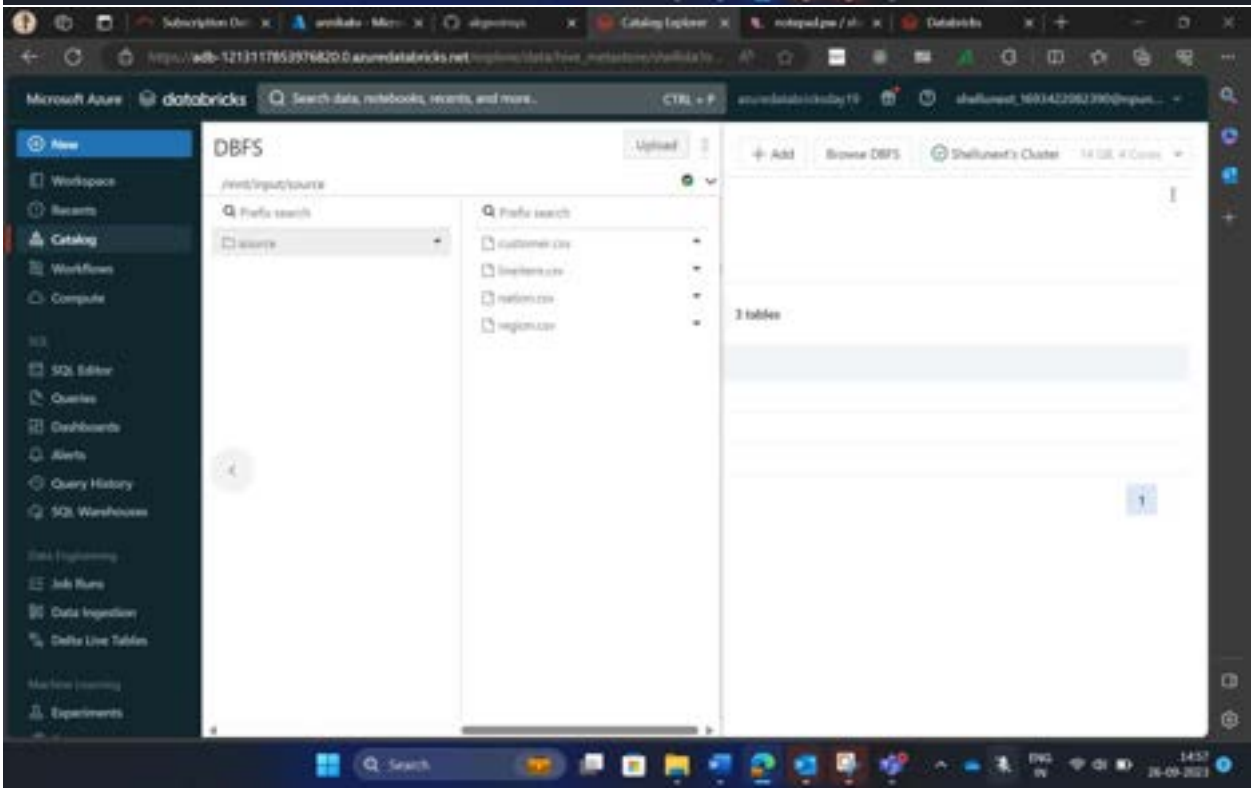


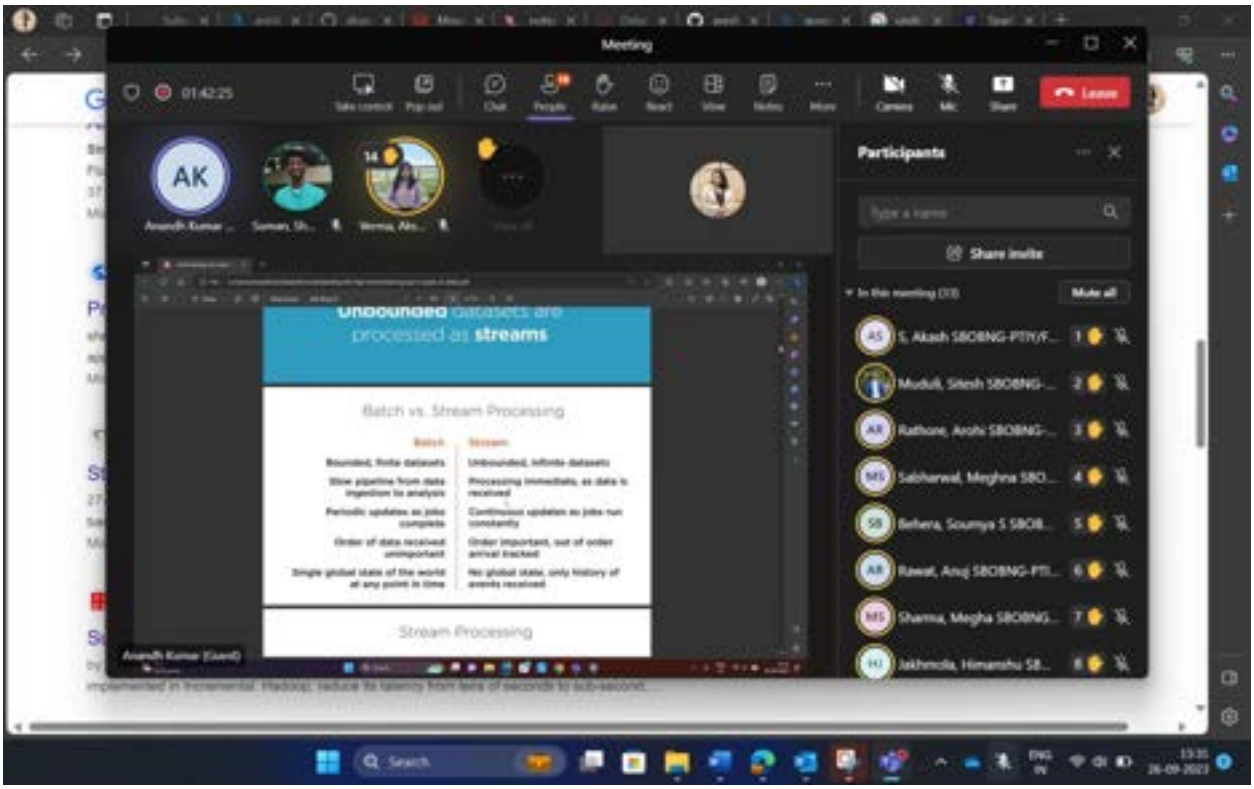
To create new scope

[Databricks (azuredatabricks.net)]

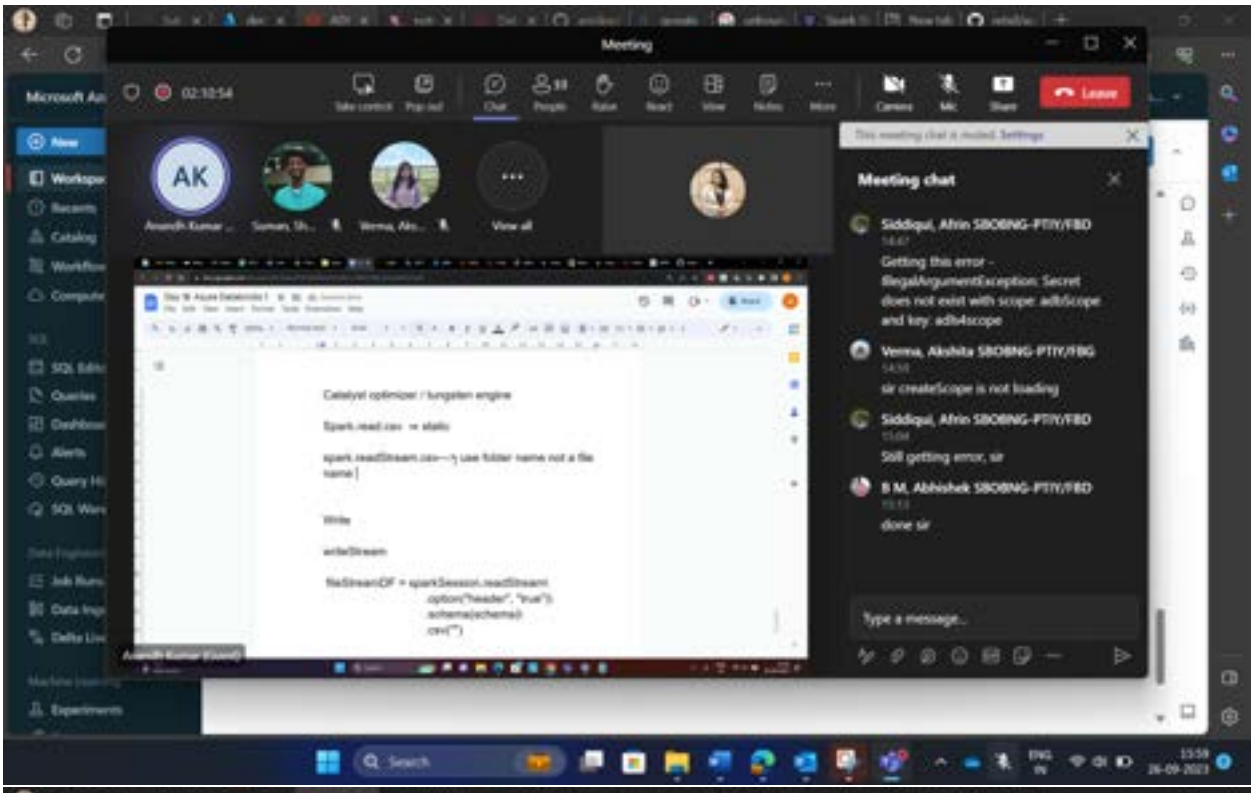


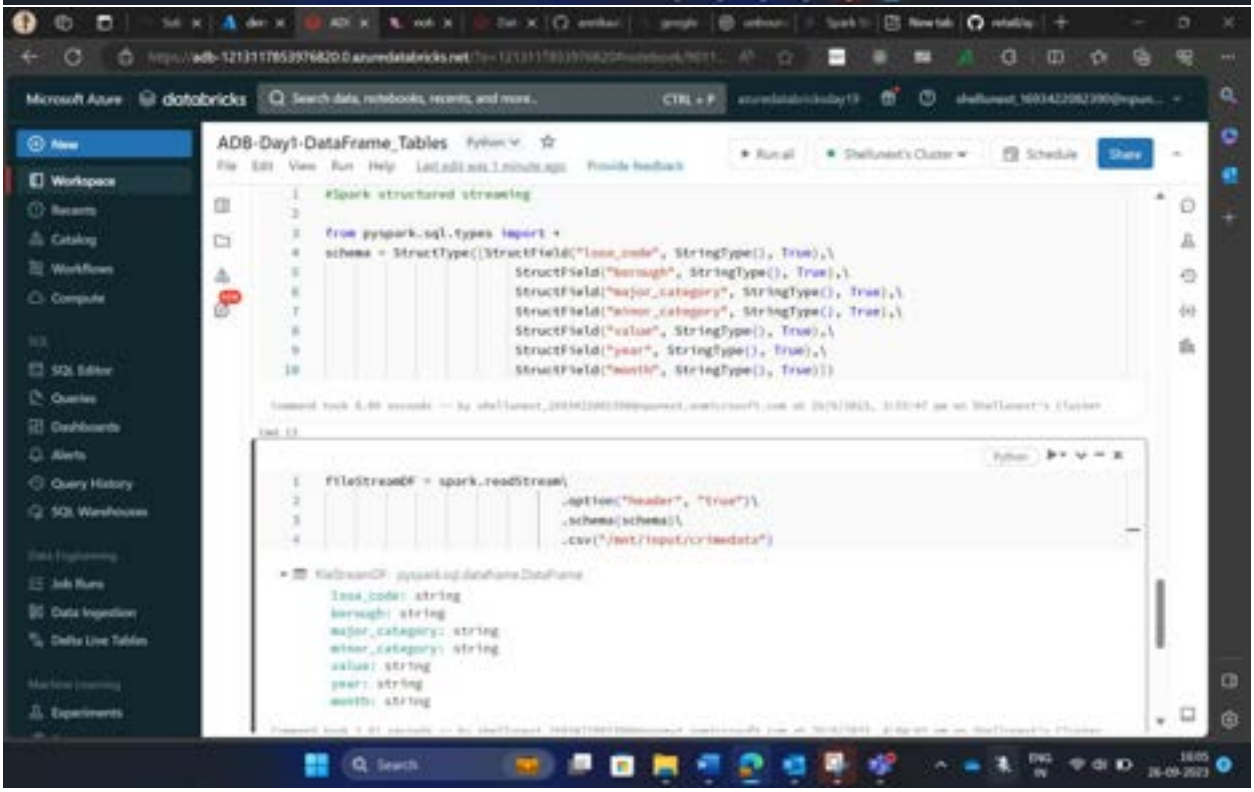

Microsoft Azure
databricks
Untitled Notebook 2023-09-26 14:35:25
dbutils.fs.mount(source = "wasbs://democontainerjanvikastorageacc.blob.core.windows.net/",
mount_point ="/mnt/input/",
extra_configs = {"fs.azure.account.key.janvikastorageacc.blob.core.windows.net":dbutils.
secrets.get(scope = "janvikascope",key="adlsky")})
True
%fs ls /mnt/input
path
name
size
modificationTime

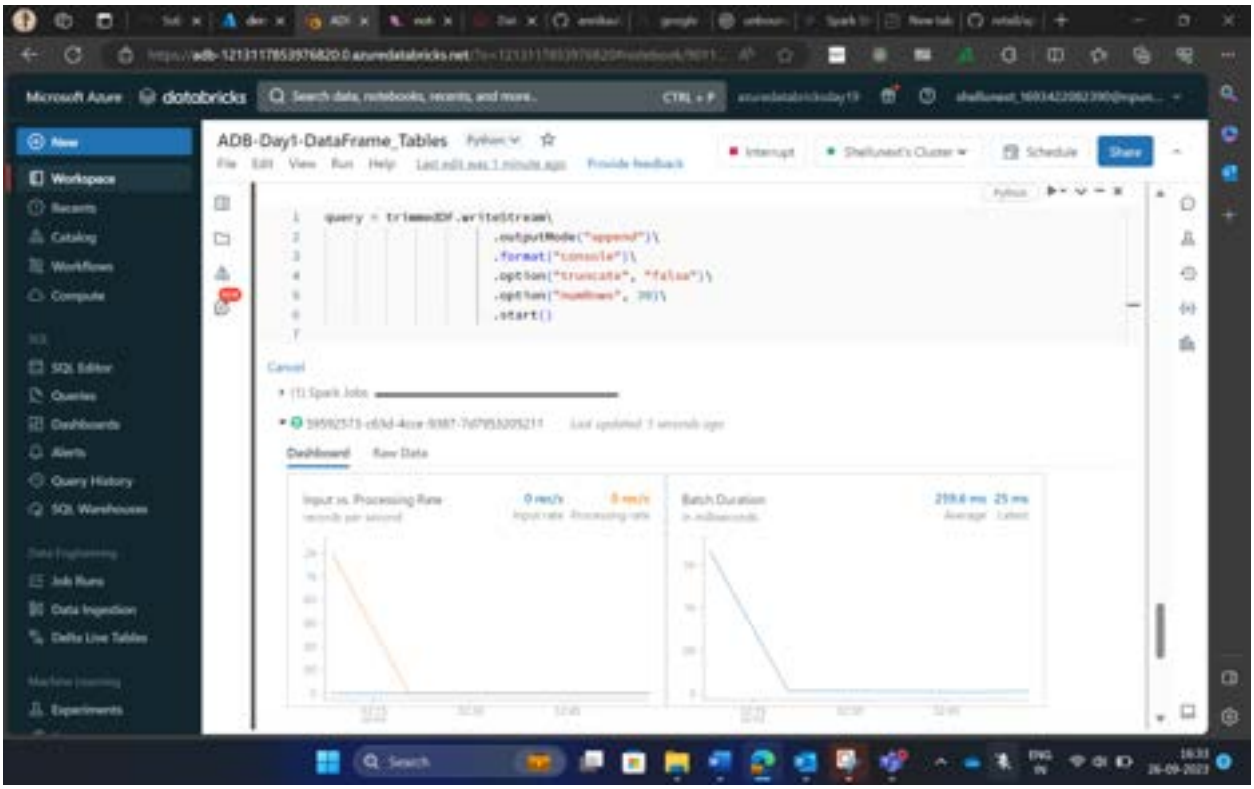

Microsoft Azure
databricks
DBFS
Upload
Add
Browse DBFS
3 tables



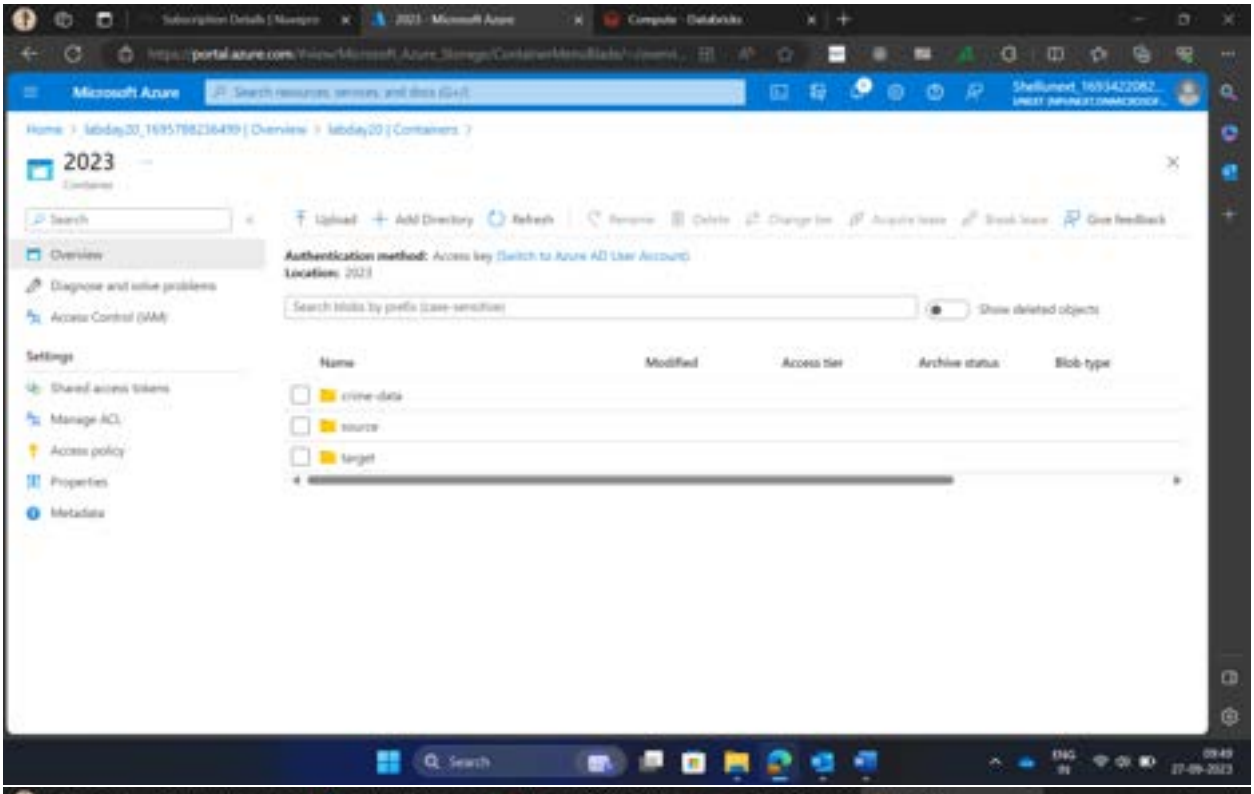
Spark streaming user defined schema

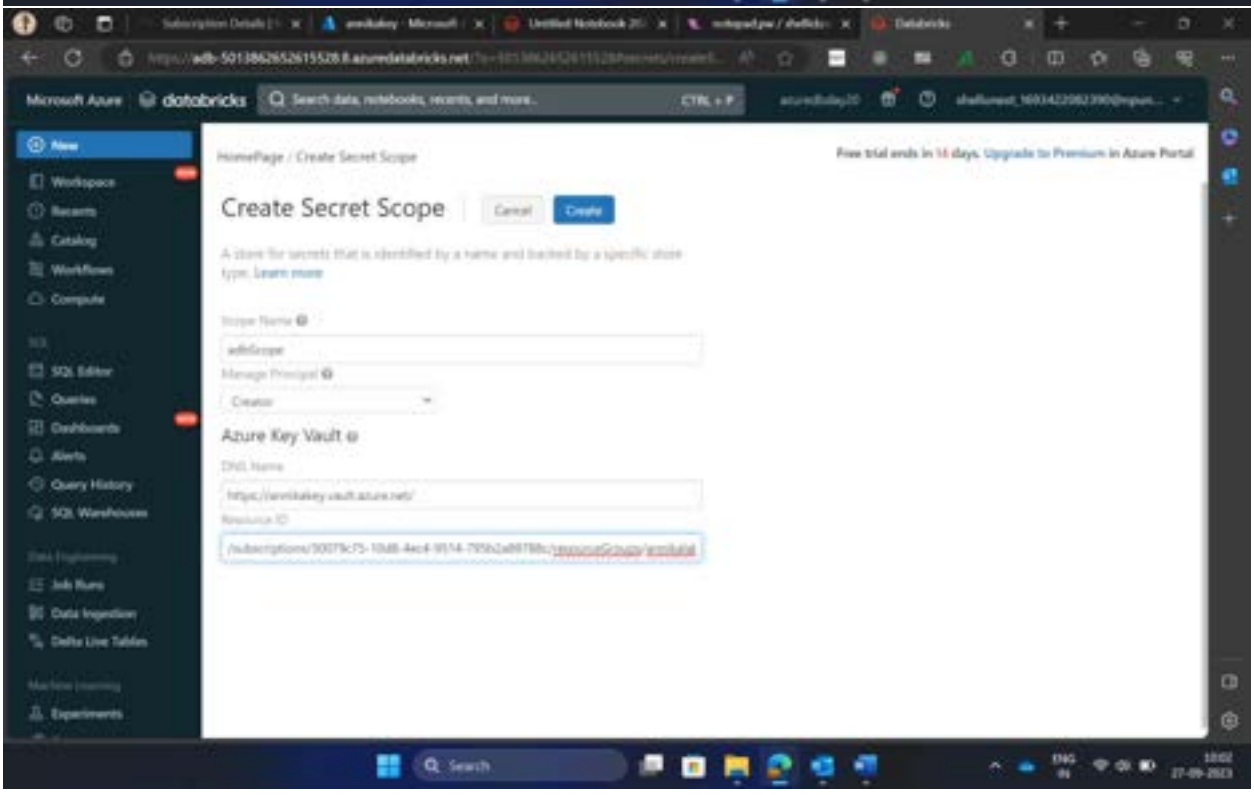

Meeting
Meeting chat
Siddiqui, Afrin SBOBNG-PTIY/FBD
Getting this error - IllegalArgumentException: Secret does not exist with scope: adlsScope and key: adls4scope
Verma, Akshita SBOBNG-PTIY/FBG
sir createScope is not loading
Siddiqui, Afrin SBOBNG-PTIY/FBD
Still getting error, sir
B M, Abhishek SBOBNG-PTIY/FBD
done sir
Type a message..

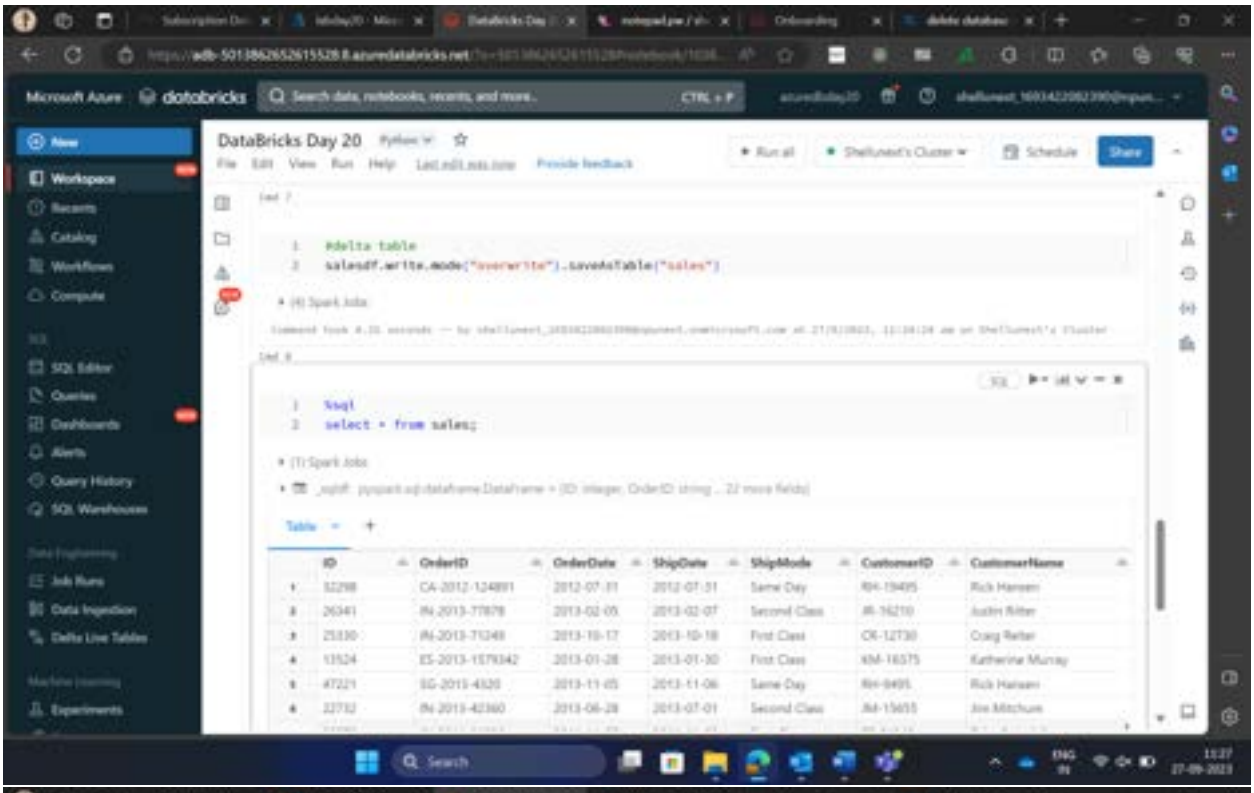

ADB-Day1-DataFrame_Tables
#Spark structured streaming
from pyspark.sql.types import *
schema = StructType([StructField("loss_code", StringType(), True),\
StructField("borough", StringType(), True),\
StructField("major_category", StringType(), True),\
StructField("minor_category", StringType(), True),\
StructField("value", StringType(), True),\
StructField("year", StringType(), True),\
StructField("month", StringType(), True)])
fileStreamDF = spark.readStream\
.option("header", "true")\
.schema(schema)\
.csv("/mnt/input/crimedata")
loss_code: string
borough: string
major_category: string
minor_category: string
value: string
year: string
month: string

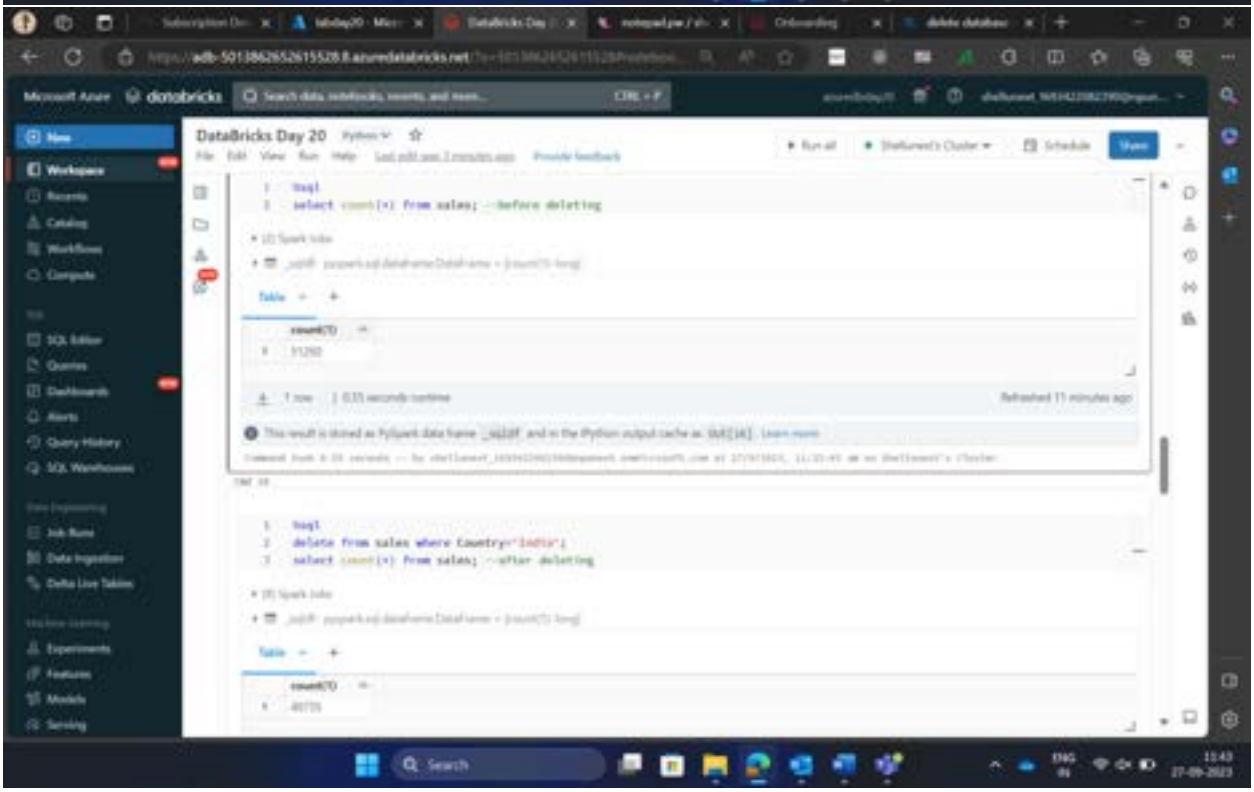

Microsoft Azure
databricks
ADB-Day1-DataFrame_Tables
Interrupt
Schedule
Share
query = trimmedDF.writeStream\
.outputMode("append")\
.format("console")\
.option("truncate", "false")\
.option("numRows", 20)\
.start()
Cancel
Dashboard
Raw Data
Input vs. Processing Rate
Batch Duration

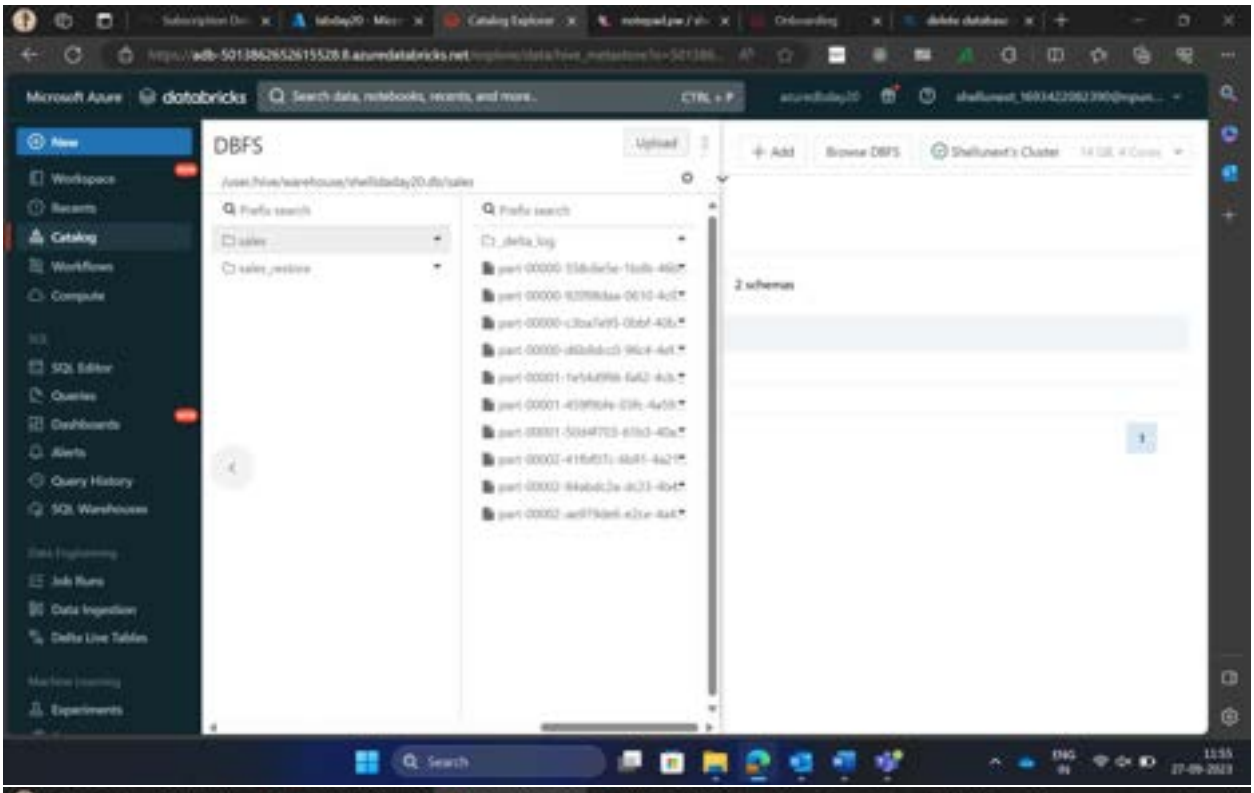
Day 20 27.09.23

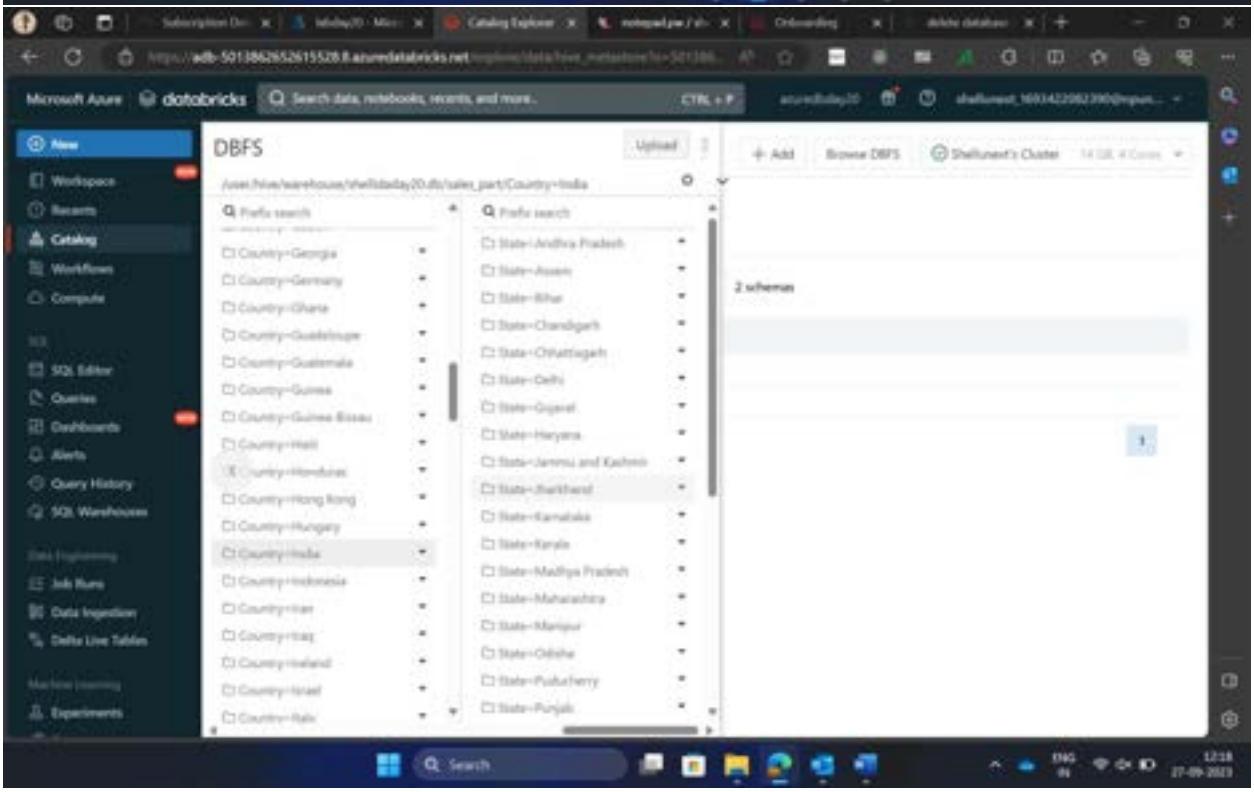


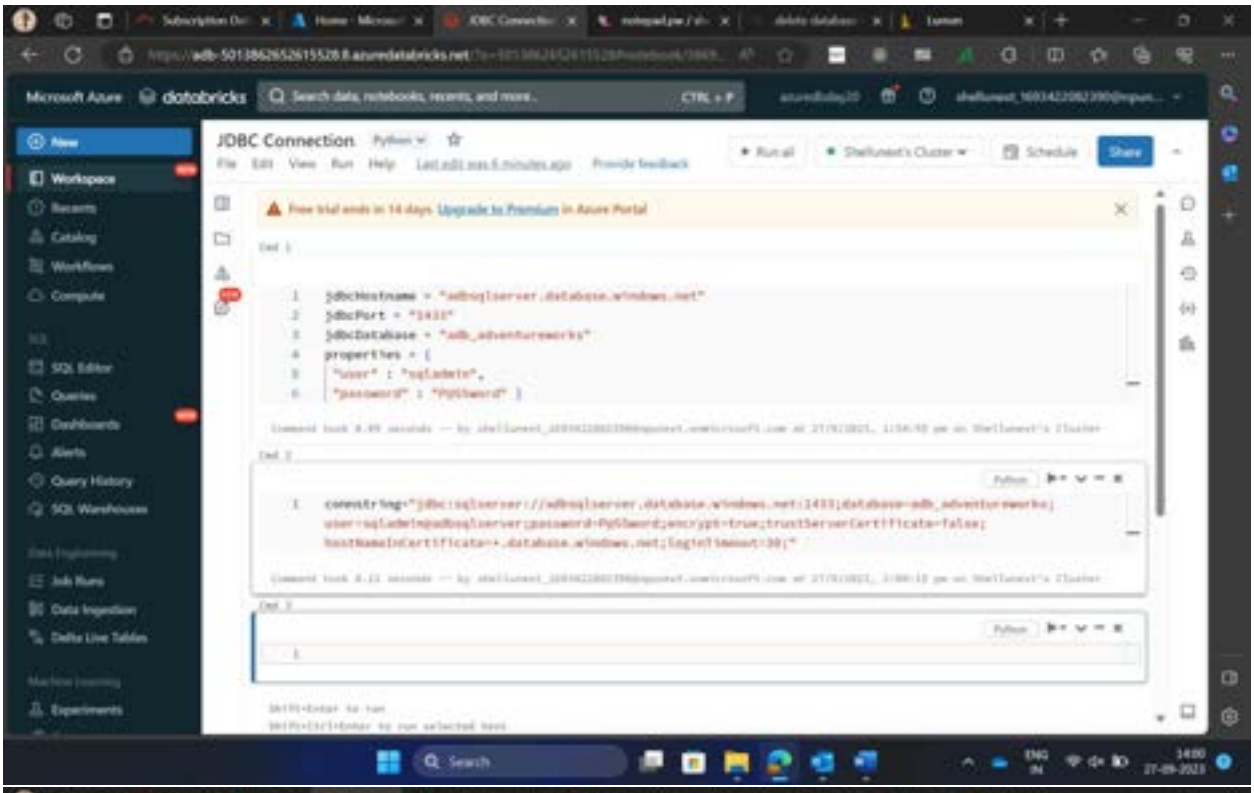


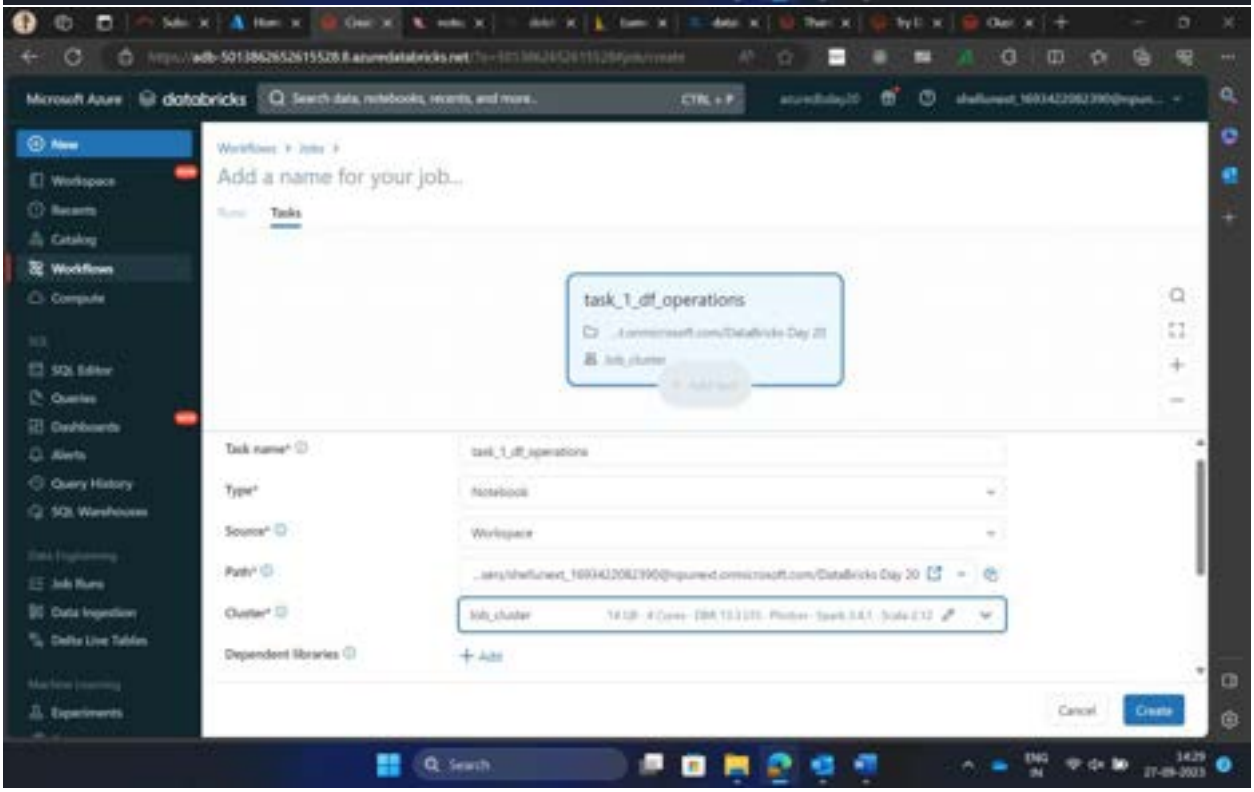

DataBricks Day 20
Share
%sql
select * from sales;

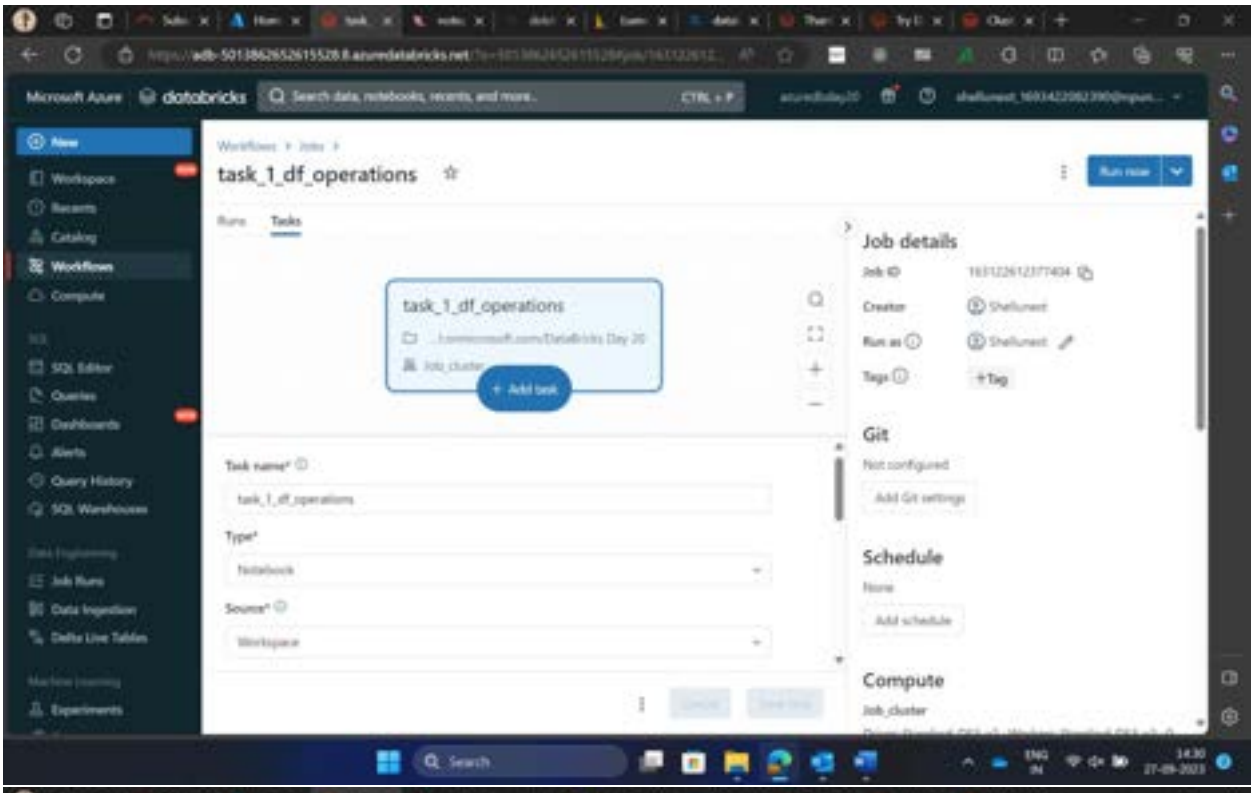

DataBricks Day 20
Share
%sql
select count(*) from sales; --before deleting
%sql
delete from sales where Country='India';
select count(*) from sales; --after deleting

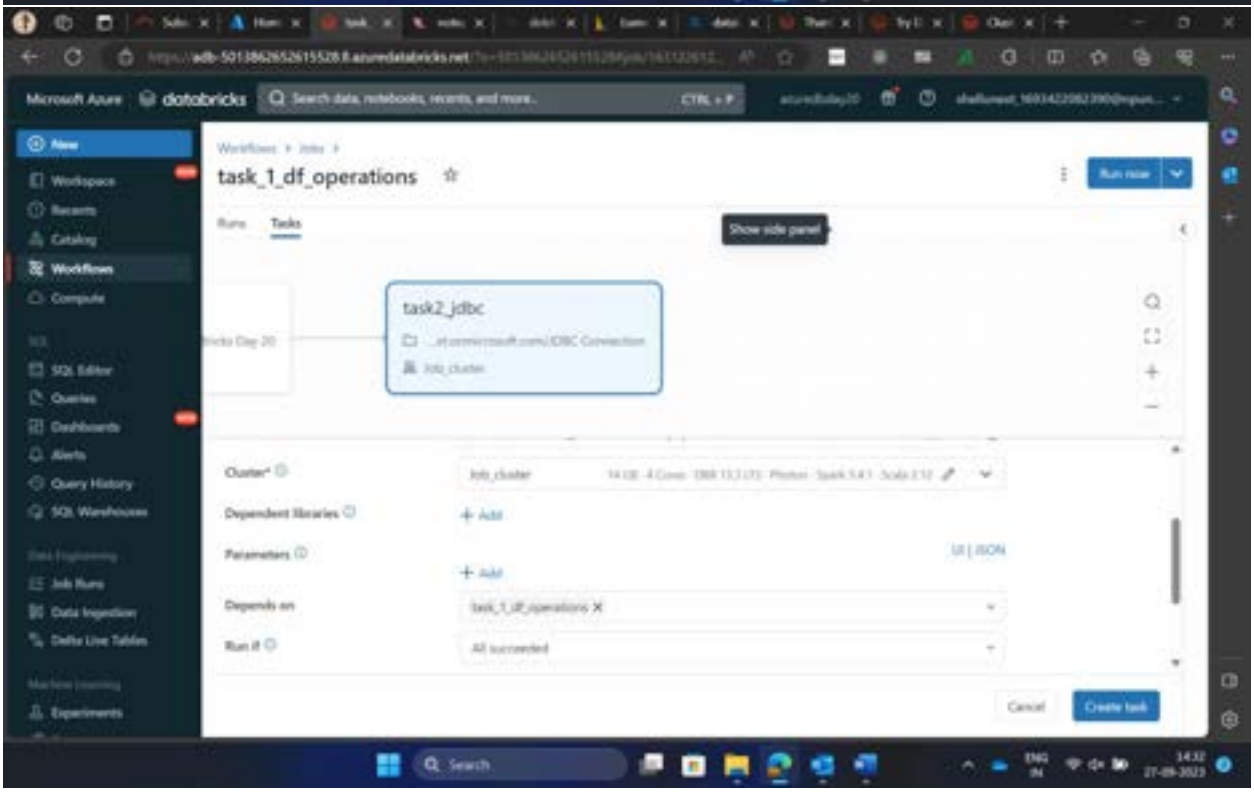

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
DBFS
Upload
Add
Browse DBFS
2 schemas

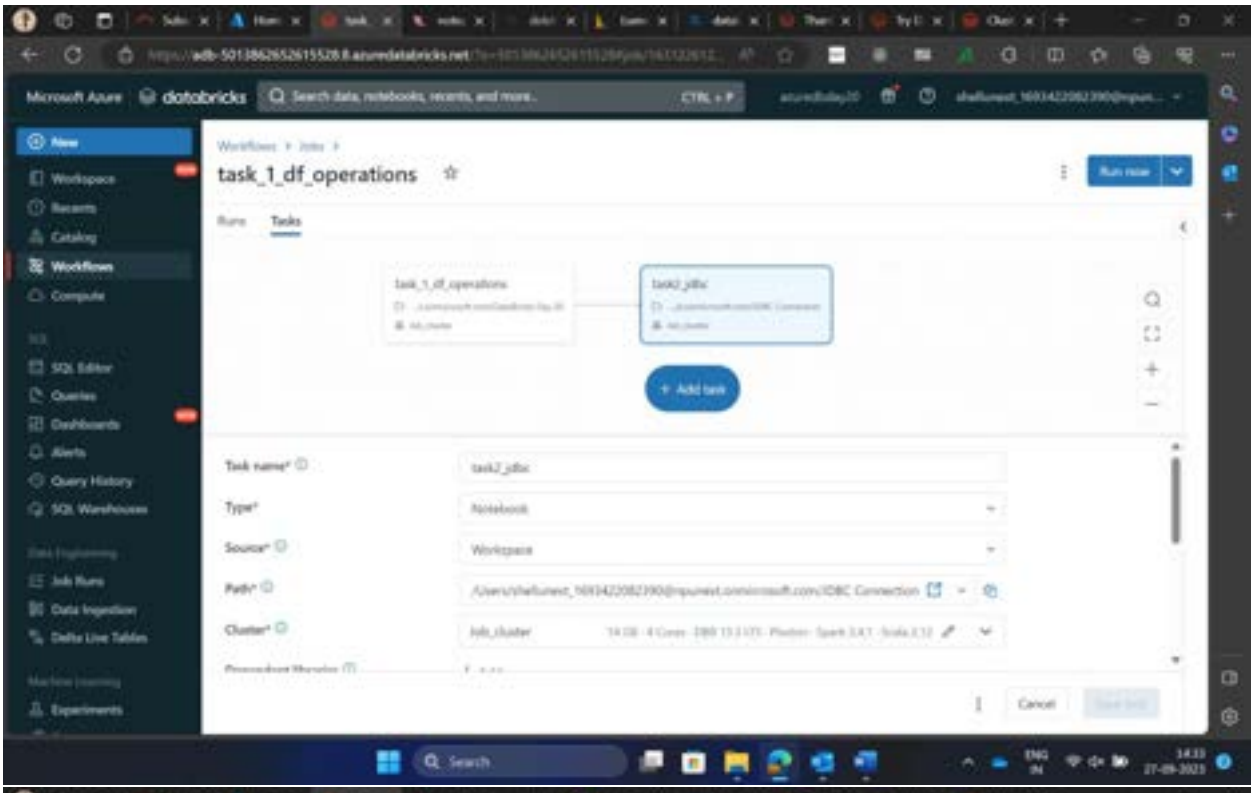

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
DBFS
Upload
Add
Browse DBFS
2 schemas

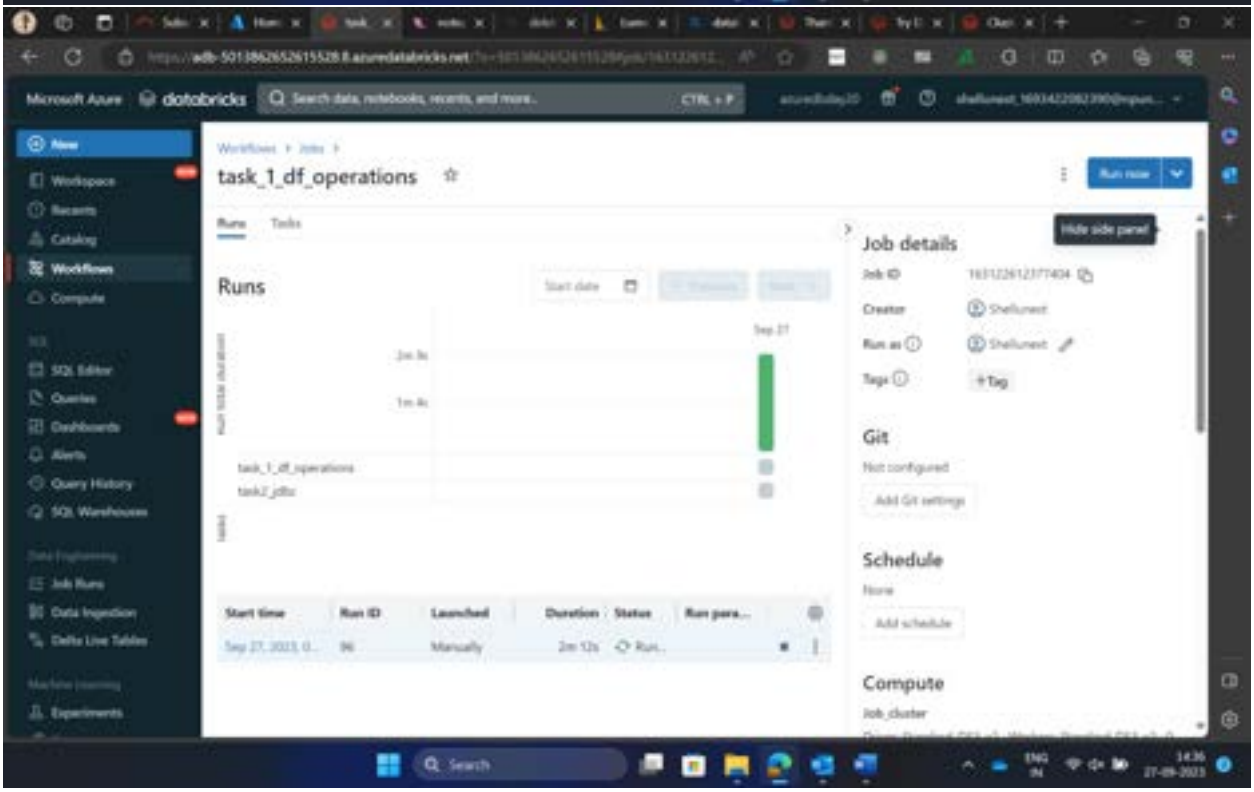

Microsoft Azure
databricks
JDBC Connection
Workspace
Recents
Catalog
Workflows
Compute
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Job Runs
Data Ingestion
Delta Live Tables
Experiments
jdbcHostname = "adbsqlserver.database.windows.net"
jdbcPort = "1433"
jdbcDatabase = "adb_adventureworks"
properties = {
"user" : "sqladmin",
"password" : "P@55word" }
connstring="jdbc:sqlserver://adbsqlserver.database.windows.net:1433;database=adb_adventureworks;
user=sqladmin@adbsqlserver;password=P@55word;encrypt=true;trustServerCertificate=false;
hostNameInCertificate=*.database.windows.net;loginTimeout=30;"

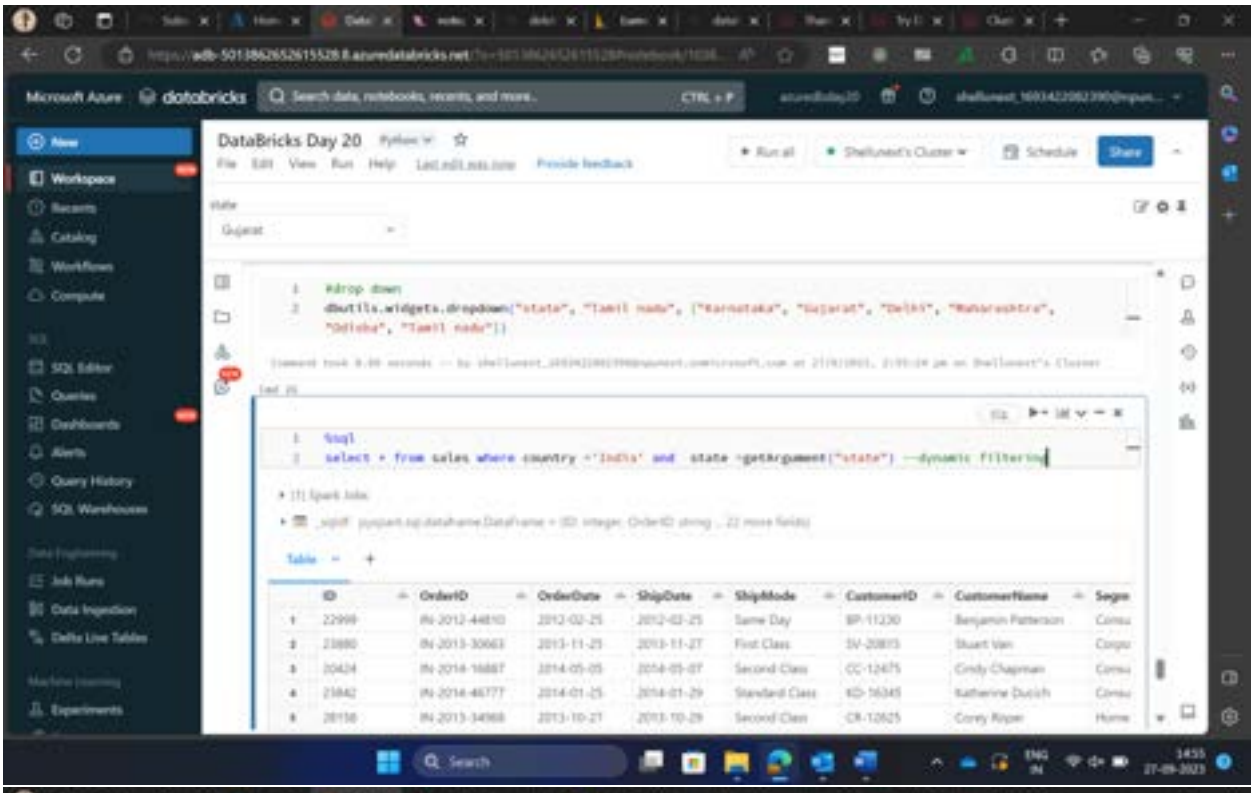

Microsoft Azure
databricks
Add a name for your job...
Tasks
task_1_df_operations
Task name*
Type*
Notebook
Source*
Workspace
Path*
Cluster*
Dependent libraries
+ Add
Cancel
Create

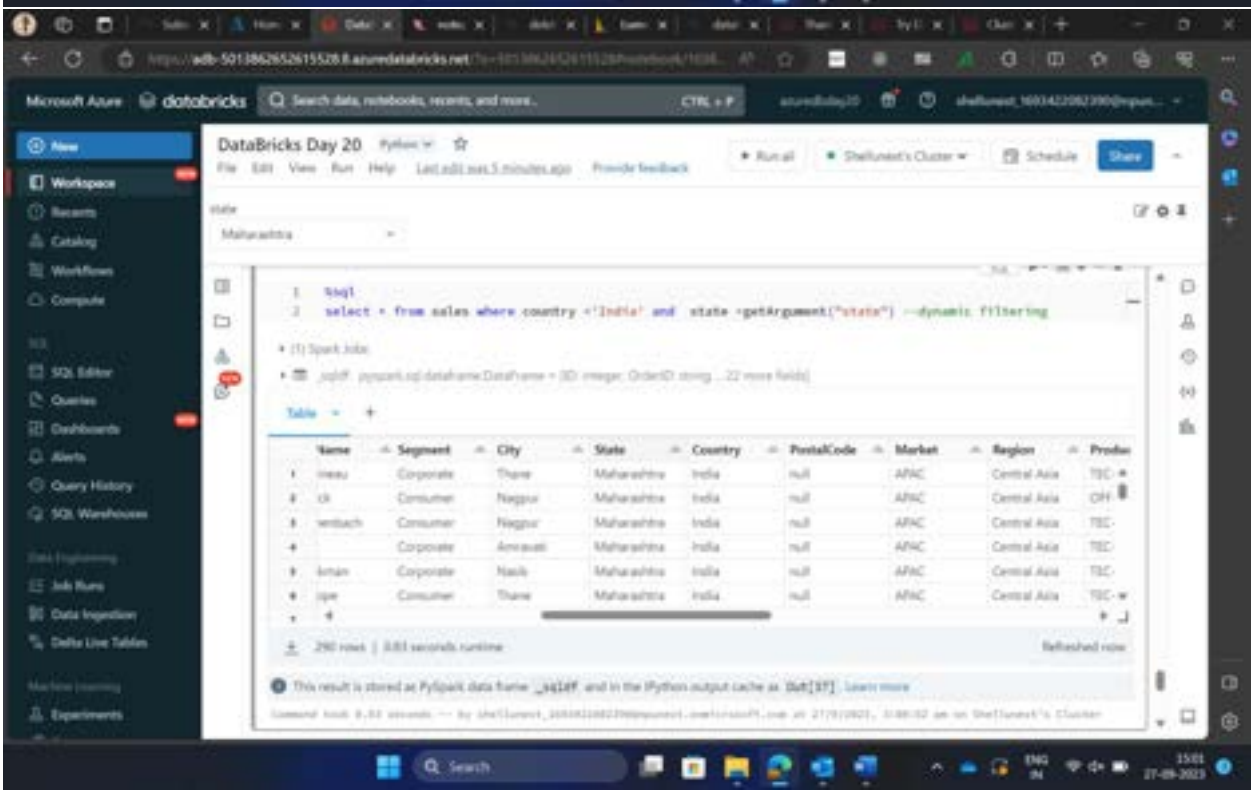

Microsoft Azure
databricks
task_1_df_operations
Runs
Tasks
task_1_df_operations
+ Add task
Job details
Job ID
Creator
Run as
Tags
+ Tag
Git
Not configured
Add Git settings
Schedule
None
Add schedule
Compute
Job_cluster
Task name*
task_1_df_operations
Type*
Notebook
Source*
Workspace

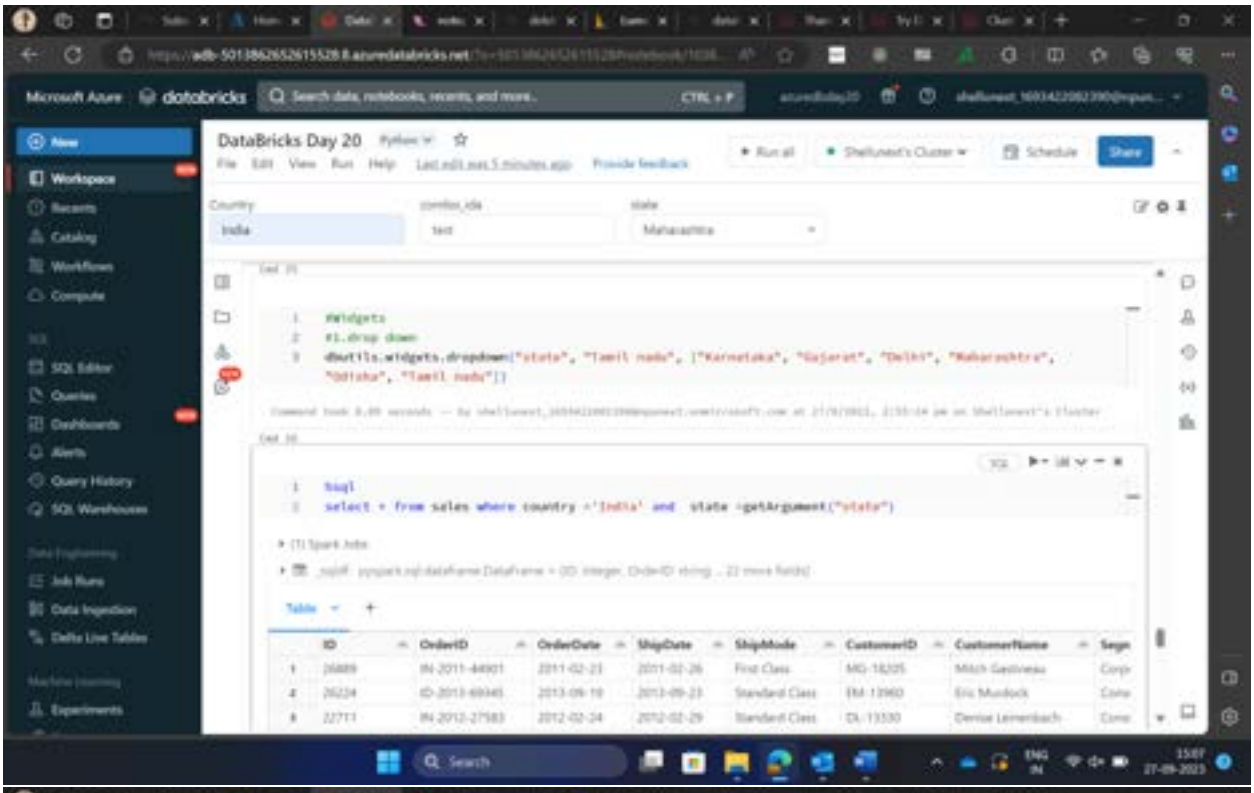

Microsoft Azure
databricks
task_1_df_operations
Runs
Tasks
Show side panel
task2_jdbc
Job_cluster
Cluster*
Dependent libraries
+ Add
Parameters
+ Add
Depends on
task_1_df_operations
Run if
All succeeded
Cancel
Create task

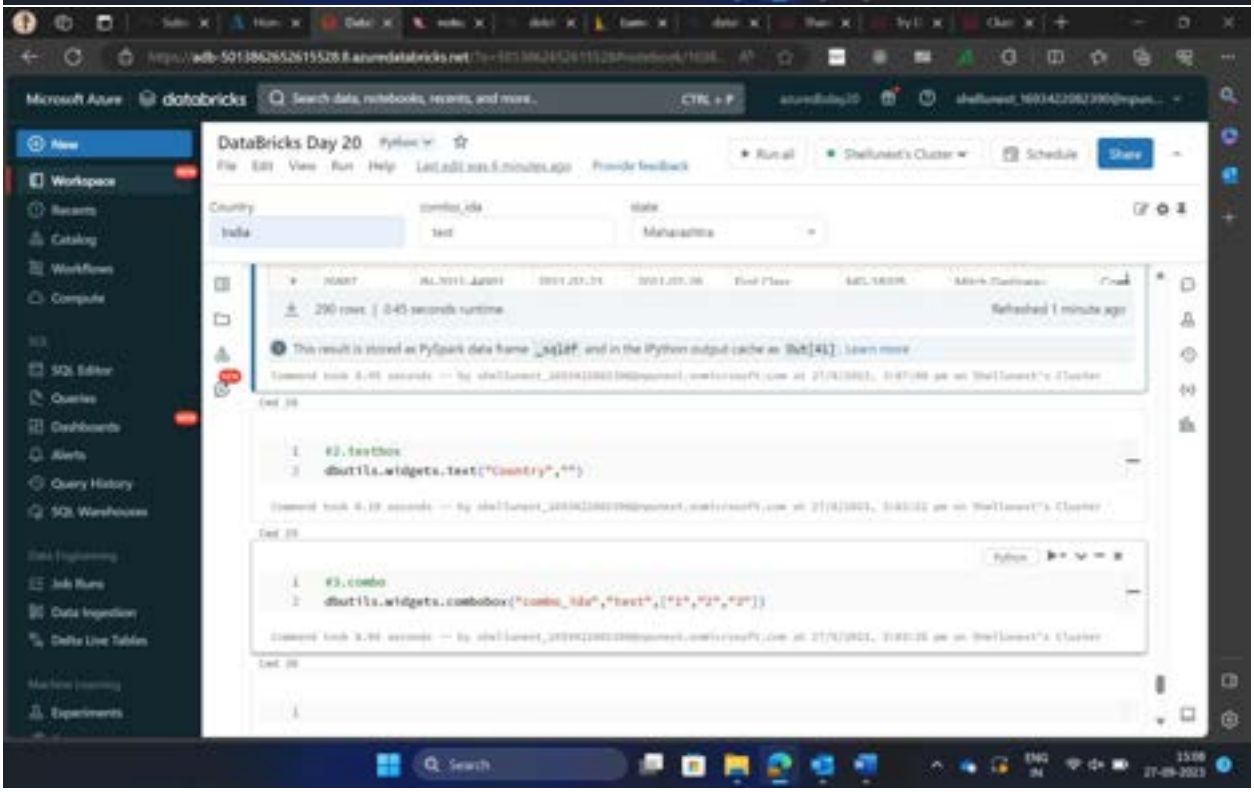

Microsoft Azure
databricks
task_1_df_operations
Run now
Runs
Tasks
+ Add task
Task name*
task2_jdbc
Type*
Notebook
Source*
Workspace
Path*
Cluster*
Cancel


Microsoft Azure
databricks
task_1_df_operations
Run now
Runs
Tasks
Runs
Job details
Hide side panel
Job ID
Creator
Run as
Tags
+Tag
Git
Not configured
Add Git settings
Schedule
None
Add schedule
Compute
Job_cluster
Start time
Run ID
Launched
Duration
Status
Run para...

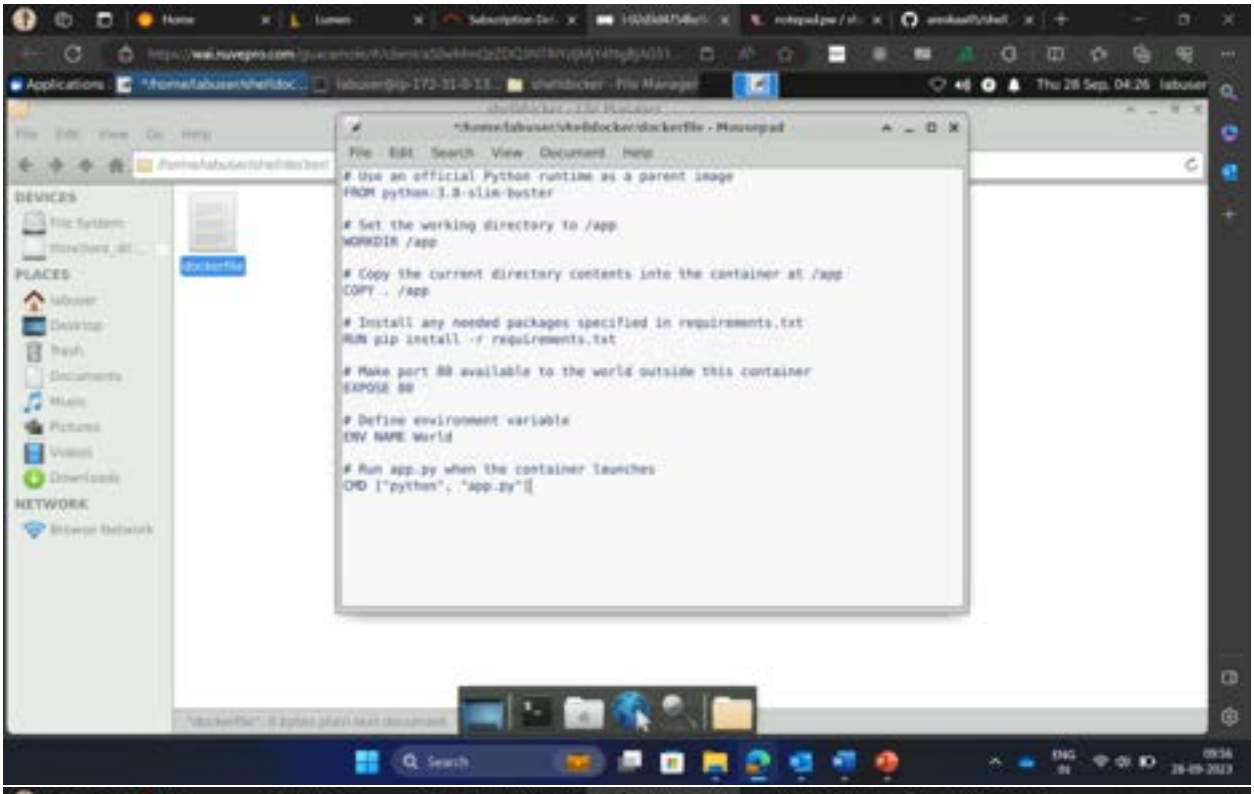

Microsoft Azure
databricks
DataBricks Day 20
Gujarat
#drop down
dbutils.widgets.dropdown("state", "Tamil nadu", ["Karnataka", "Gujarat", "Delhi", "Maharashtra", "Odisha", "Tamil nadu"])
%sql
select * from sales where country ='India' and state =getArgument("state") --dynamic filtering
Table
ID
OrderID
OrderDate
ShipDate
ShipMode
CustomerID
CustomerName
Segm

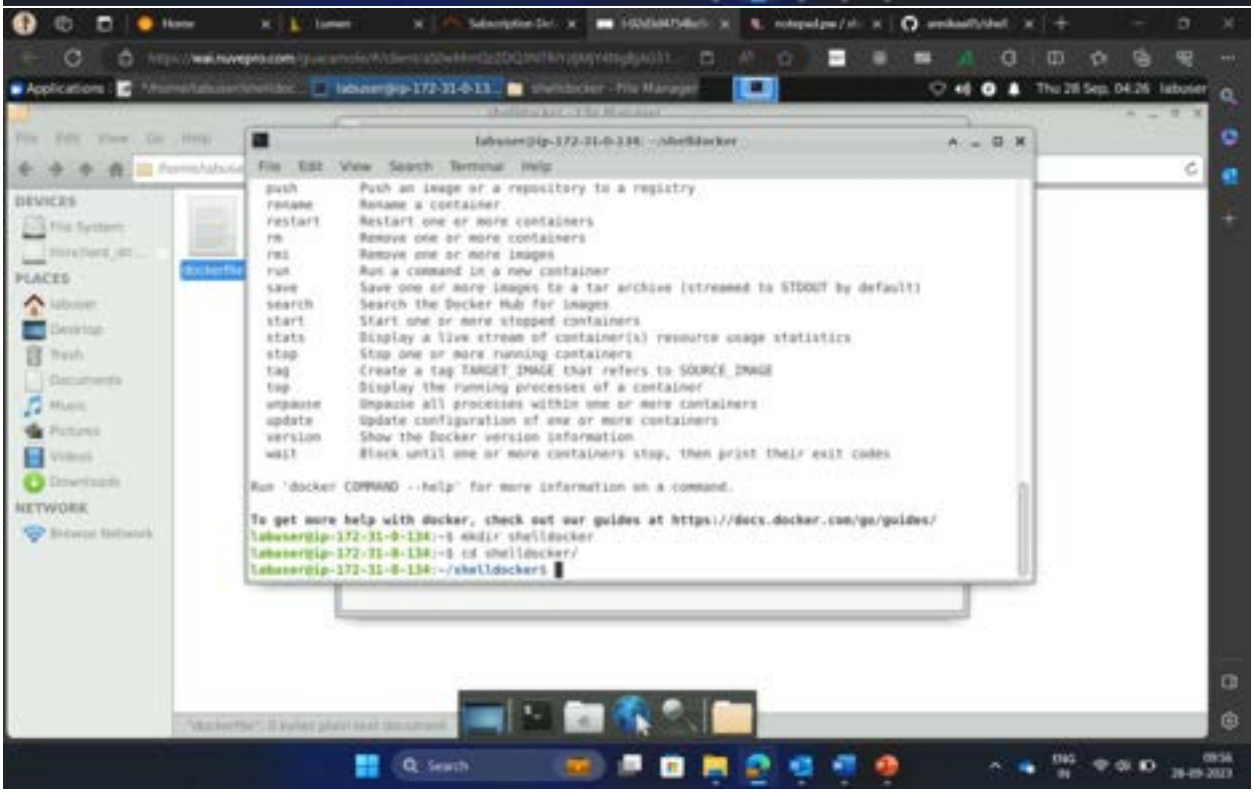

Microsoft Azure
databricks
DataBricks Day 20
Maharashtra
%sql
select * from sales where country ='India' and state =getArgument("state") --dynamic filtering
Table
Name
Segment
City
State
Country
PostalCode
Market
Region
Produ
290 rows | 0.83 seconds runtime
Refreshed now
This result is stored as PySpark data frame _sqldf and in the IPython output cache as Out[37]

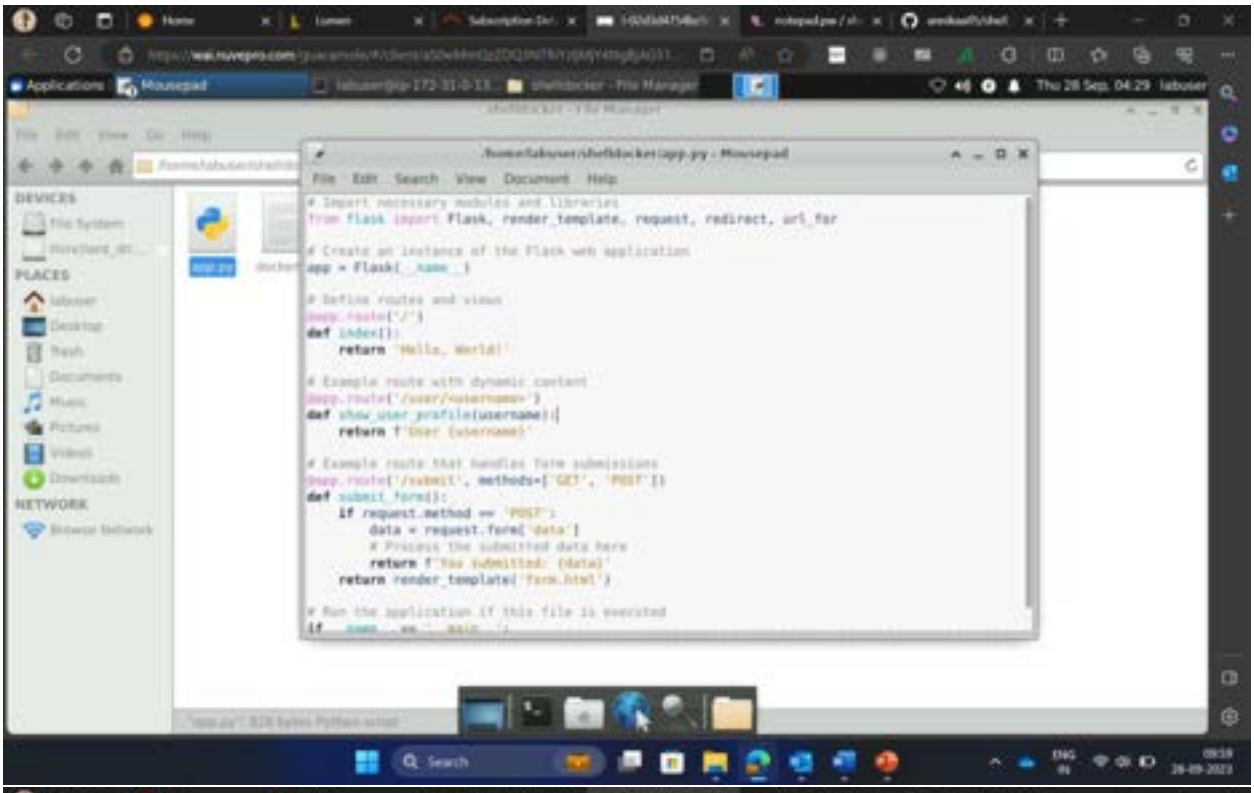

DataBricks Day 20
Country
combo_ida
state
India
text
Maharashtra
#Widgets
#1.drop down
dbutils.widgets.dropdown("state", "Tamil nadu", ["Karnataka", "Gujarat", "Delhi", "Maharashtra", "Odisha", "Tamil nadu"])
%sql
select * from sales where country ='India' and state =getArgument("state")
Table
ID
OrderID
OrderDate
ShipDate
ShipMode
CustomerID
CustomerName
Segm

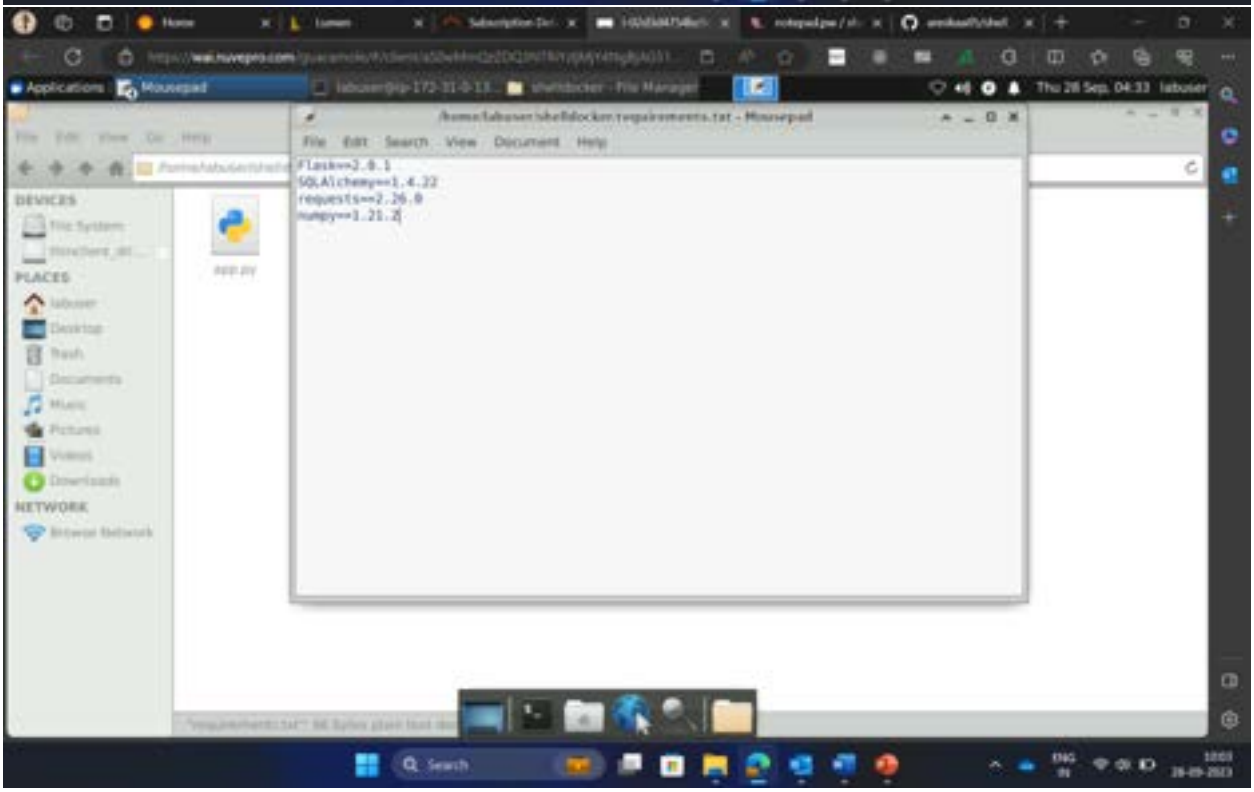

DataBricks Day 20
Country
combo_ida
state
India
text
Maharashtra
290 rows | 0.45 seconds runtime
Refreshed 1 minute ago
This result is stored as PySpark data frame _sqldf, and in the IPython output cache as Out[41]. Learn more
#2.textbox
dbutils.widgets.text("Country","")
#3.combo
dbutils.widgets.combobox("combo_ida","test",["1","2","3"])

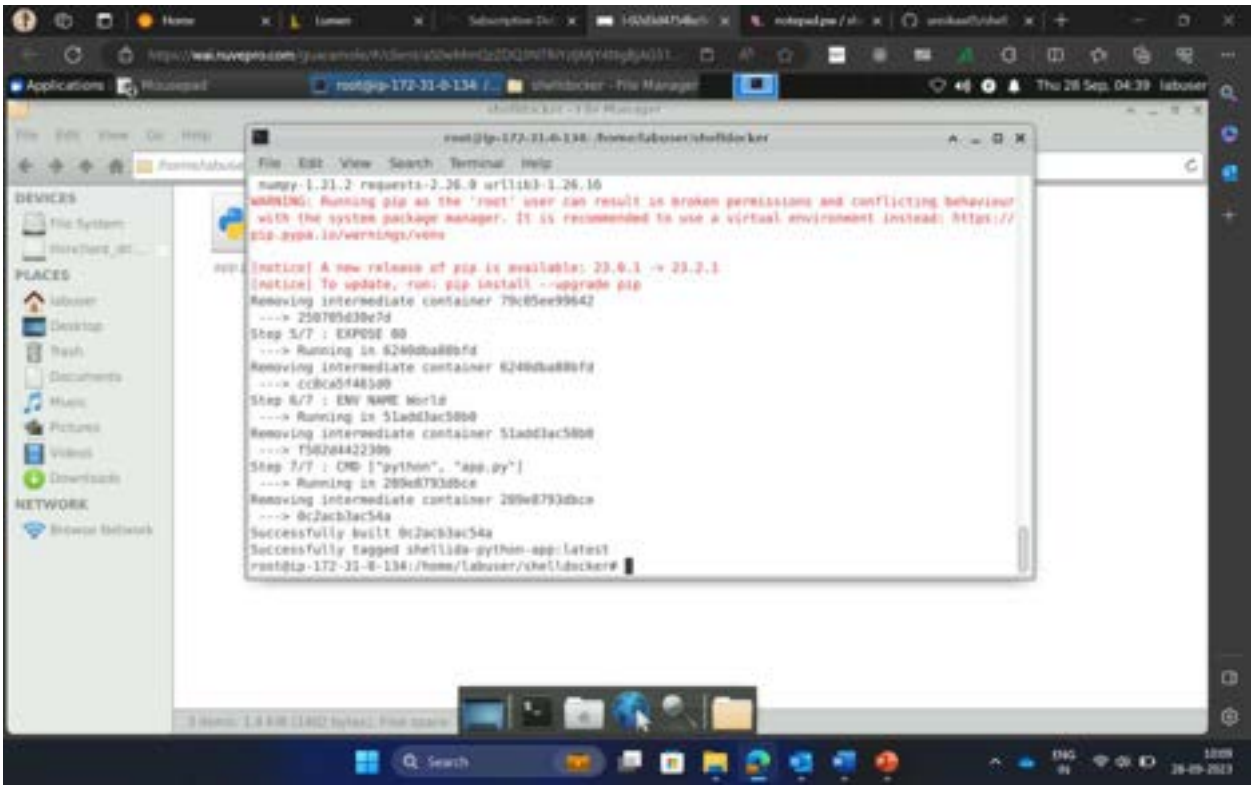
Day 21 28.09.2023






/home/labuser/shelldocker/app.py - Mousepad
File Edit Search View Document Help
# Import necessary modules and libraries
from flask import Flask, render_template, request, redirect, url_for
# Create an instance of the Flask web application
app = Flask(__name__)
# Define routes and views
@app.route('/')
def index():
return 'Hello, World!'
Example route with dynamic content
@app.route('/user/<username>')
def show_user_profile(username):
return f'User {username}'
Example route that handles form submissions
@app.route('/submit', methods=['GET', 'POST'])
def submit_form():
if request.method == 'POST':
data = request.form['data']
# Process the submitted data here
return f'You submitted: {data}'
return render_template('form.html')
# Run the application if this file is executed
if __name__ == '__main__':


/home/labuser/shelldocker/requirements.txt - Mousepad
File Edit Search View Document Help
Flask==2.0.1
SQLAlchemy==1.4.22
requests==2.26.0
numpy==1.21.2



docker build -t shellida-python-app .

#access denied

sudo su

docker build -t shellida-python-app .

docker images

docker run shellida-python-app

#to install azure-cli

curl -sL https://aka.ms/InstallAzureCLIDeb | sudo bash

#to check the azure cli

az

#to login azure account

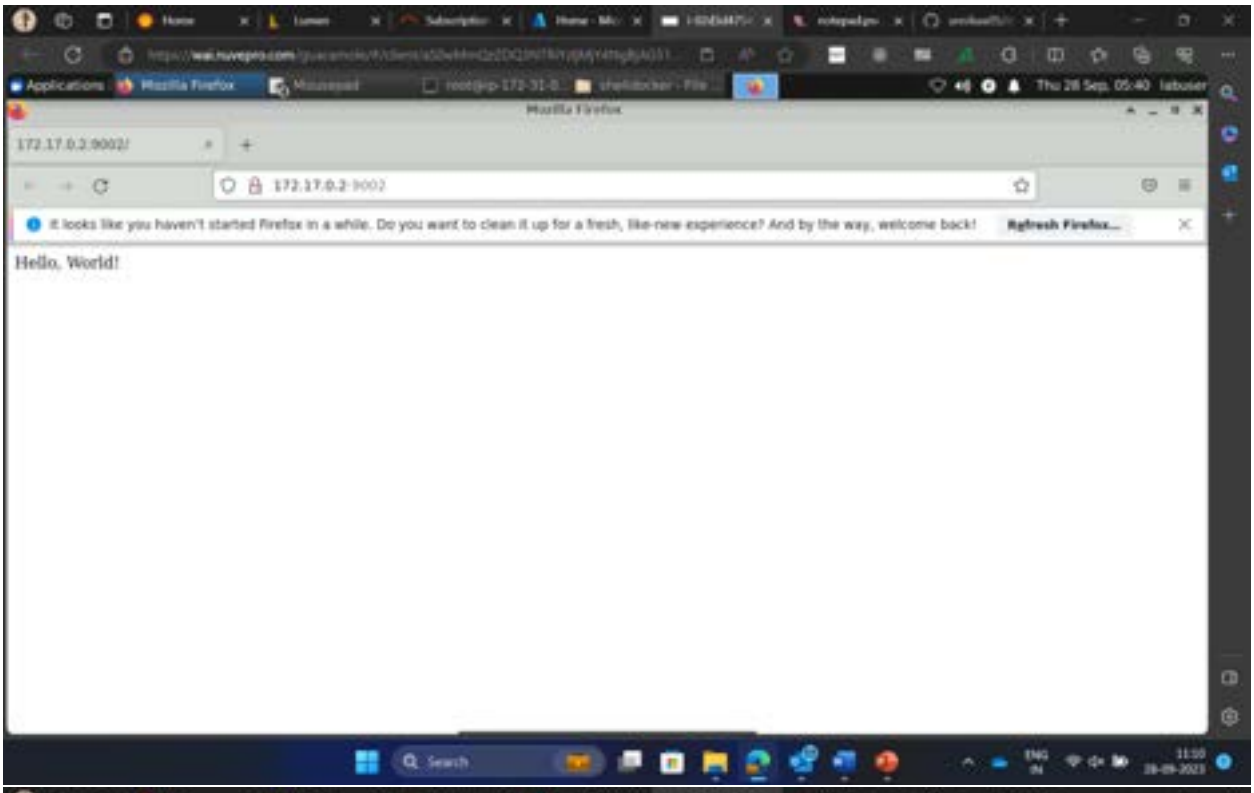
az login   -u username -p

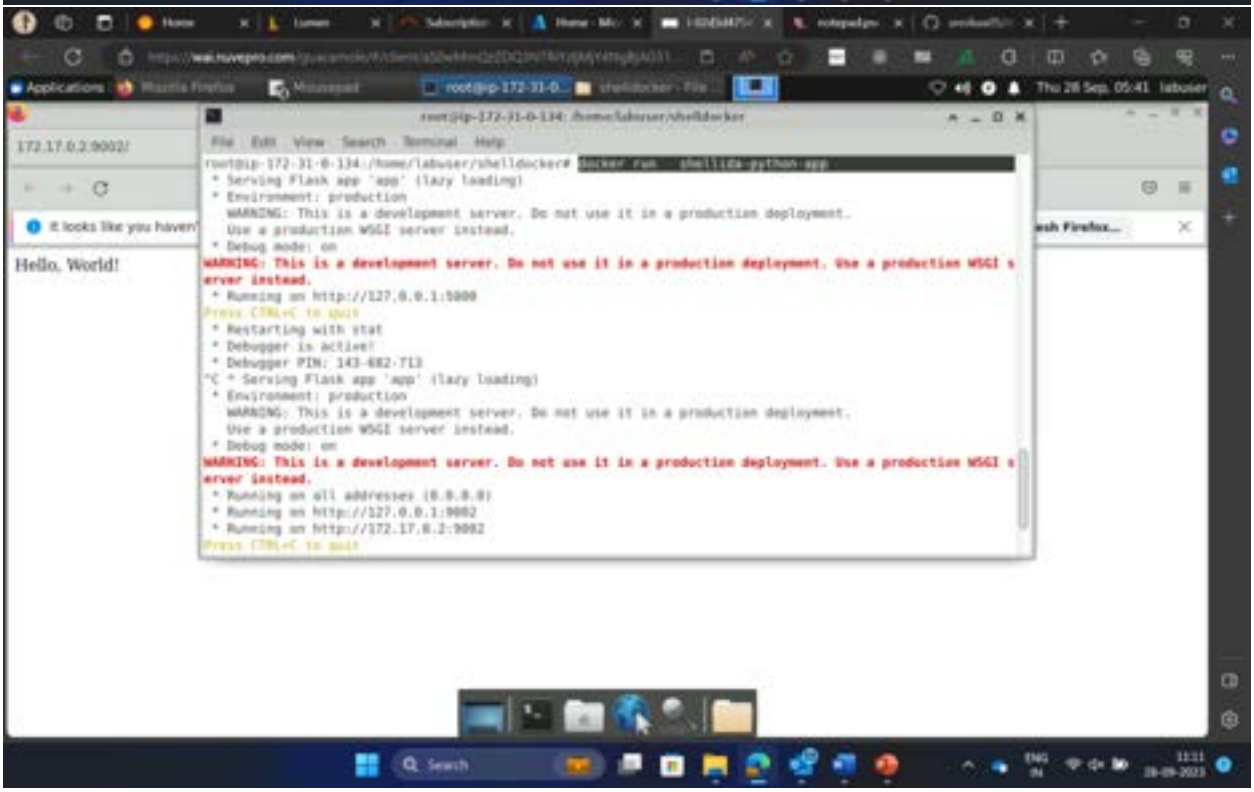
az login -u Shellunext_1693422082390@npunext.onmicrosoft.com -p "c2-677a62aN24)"

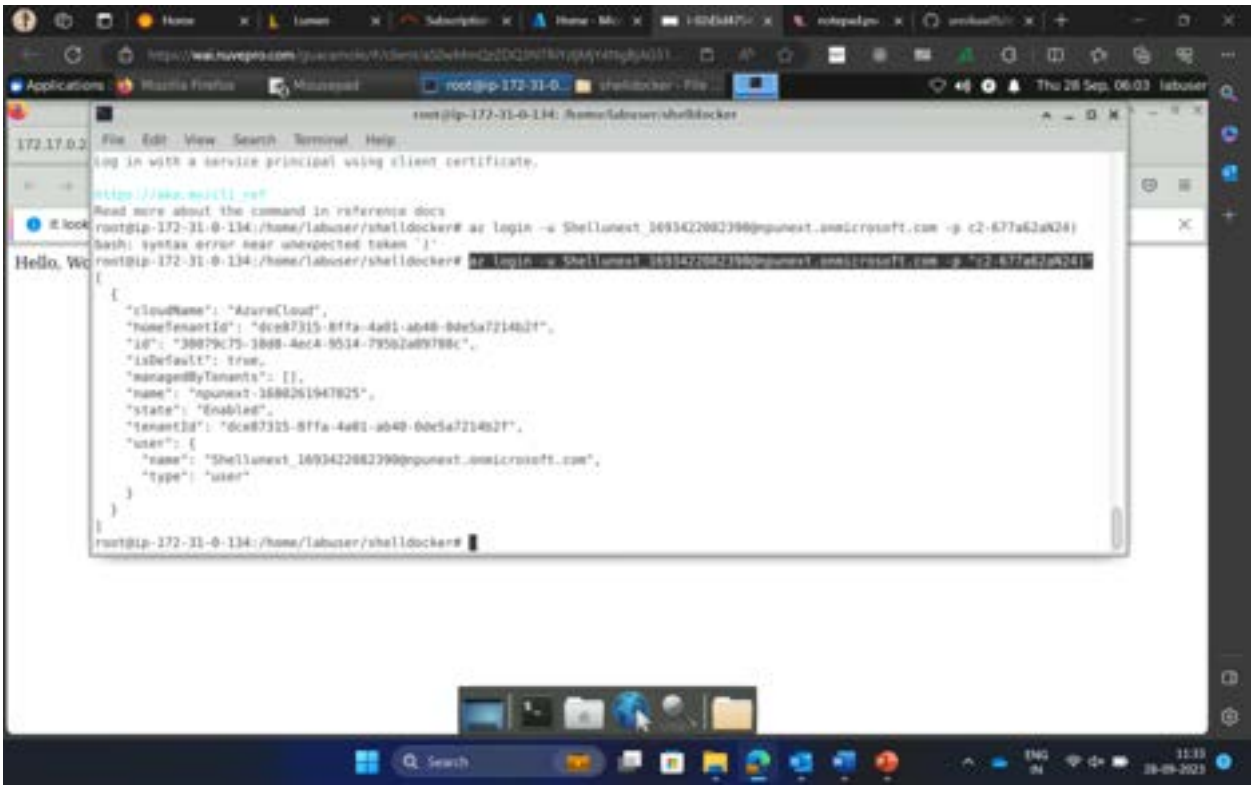
az acr login --name azcontainerday21(container name)

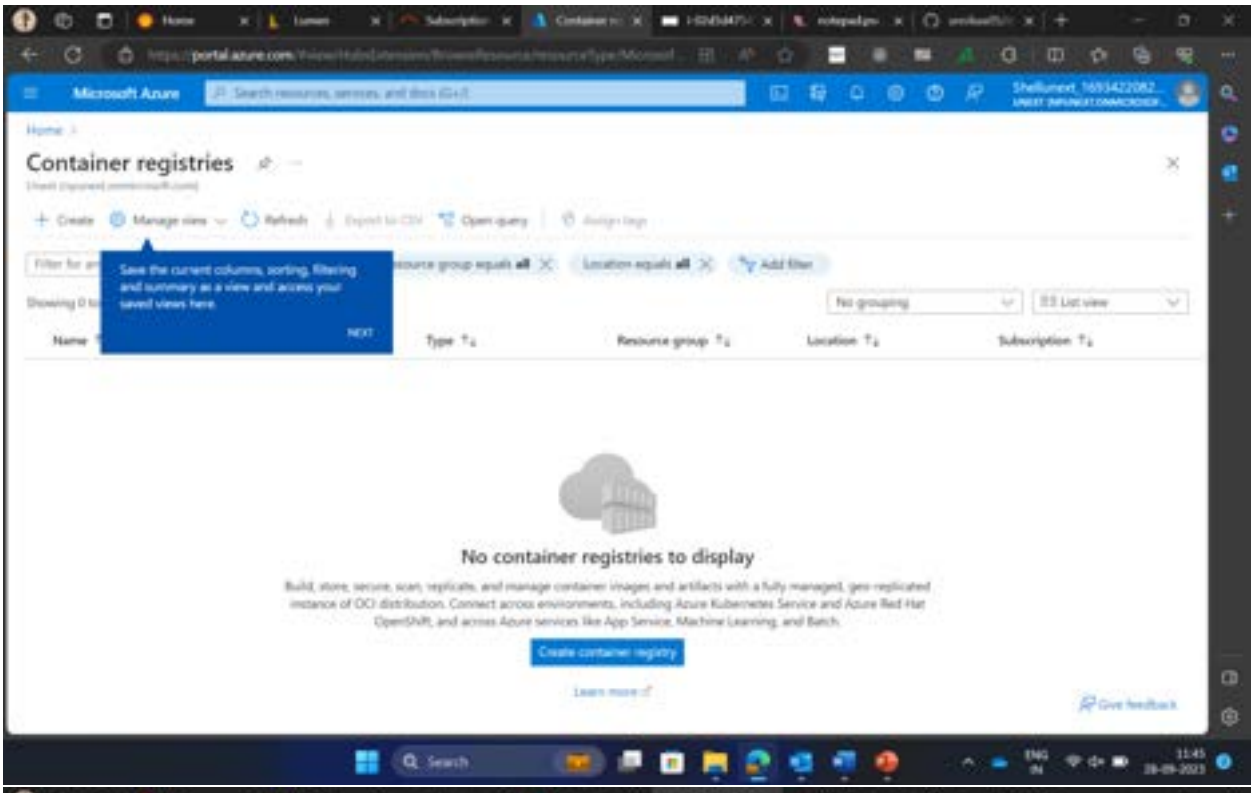
docker tag shellida-python-app azcontainerday21.azurecr.io/ida-python-app:v1

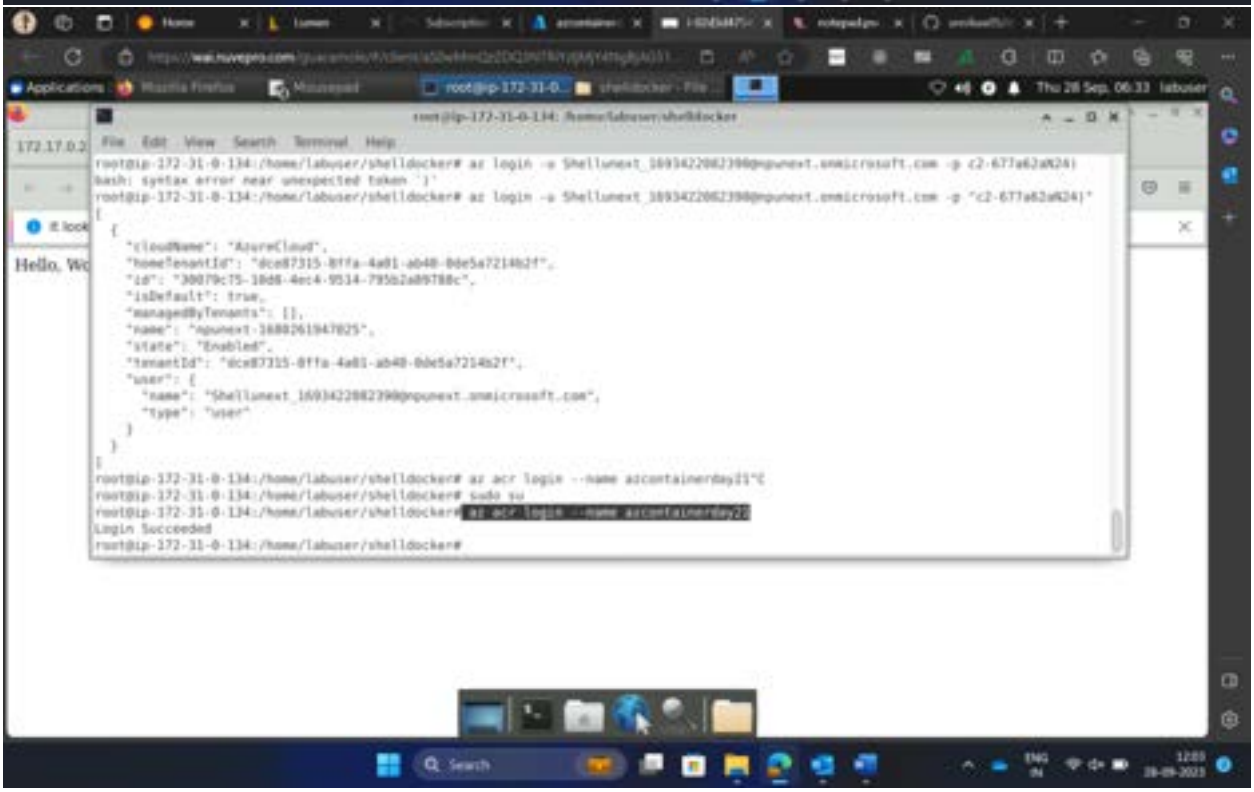
docker push azcontainerday21.azurecr.io/ida-python-app:v1

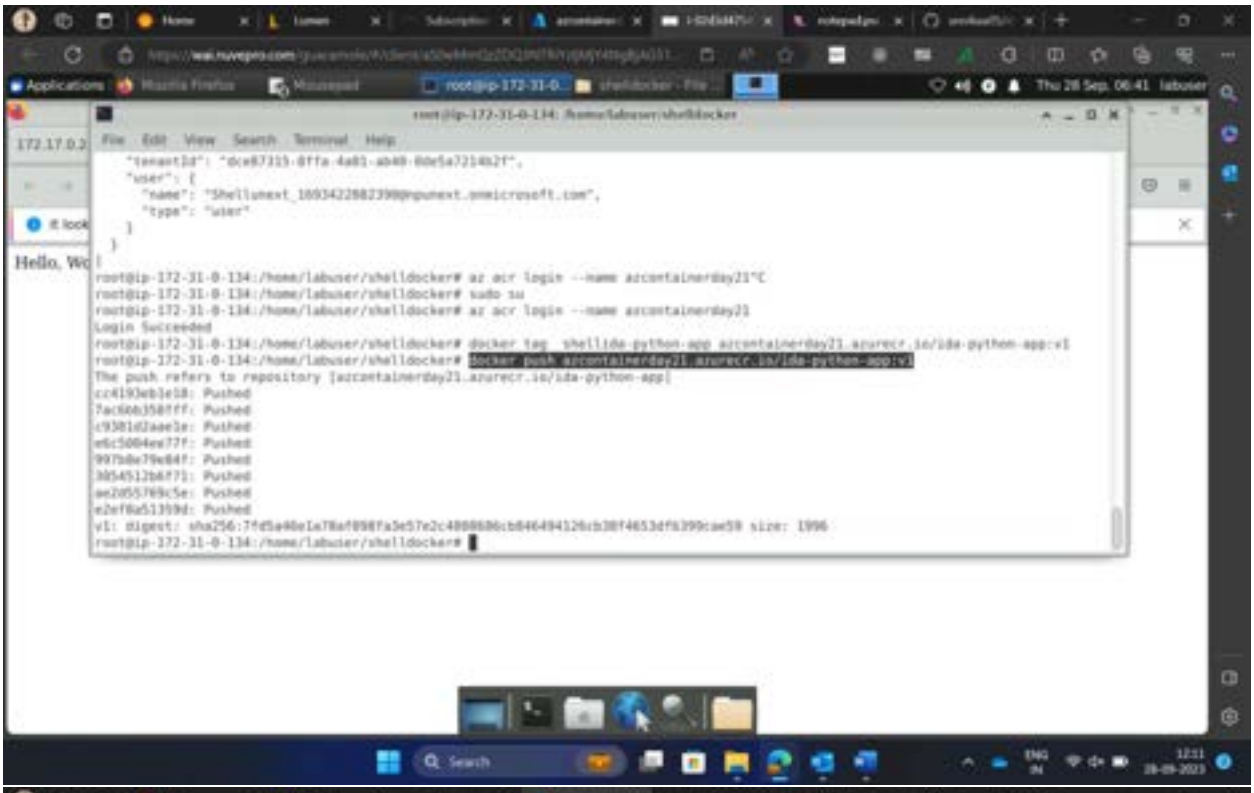

Mozilla Firefox
172.17.0.2:9002/
172.17.0.2:9002
It looks like you haven't started Firefox in a while. Do you want to clean it up for a fresh, like-new experience? And by the way, welcome back!
Refresh Firefox...
Hello, World!

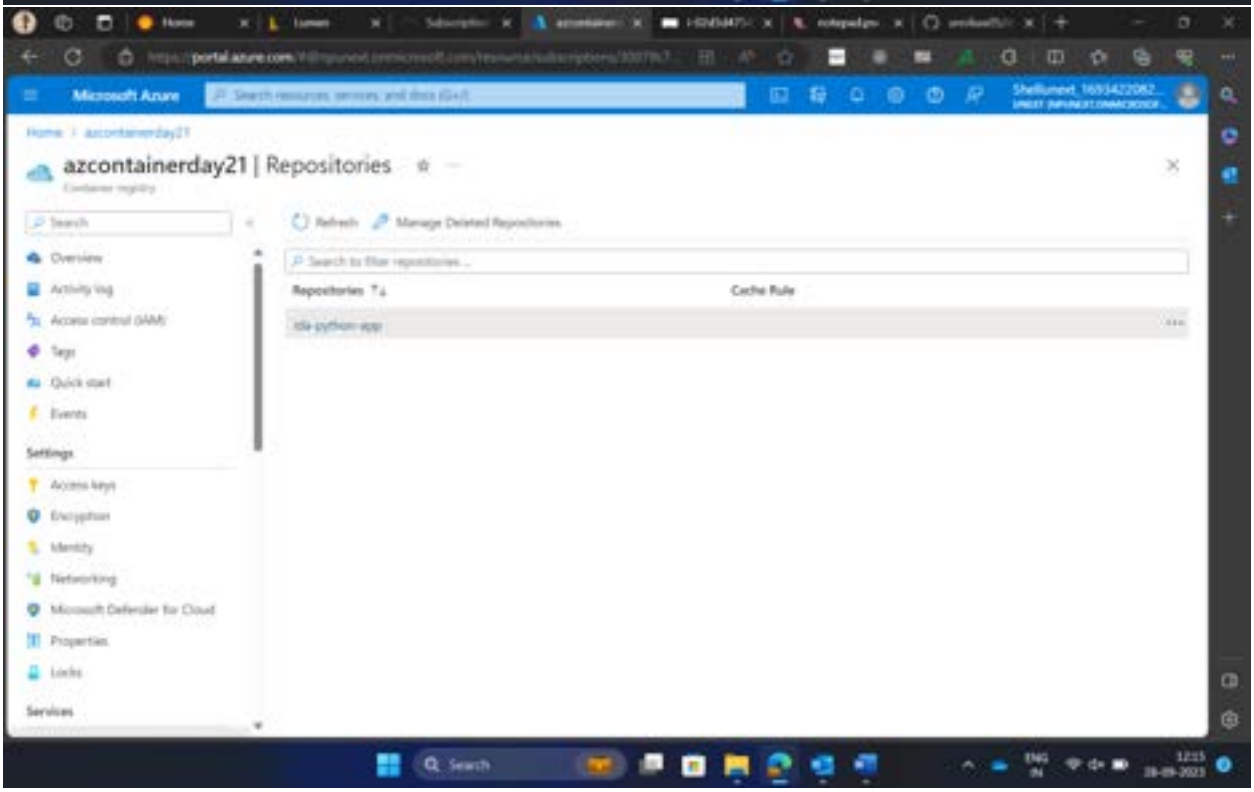

root@ip-172-31-0-134: /home/labuser/shelldocker
File Edit View Search Terminal Help
root@ip-172-31-0-134:/home/labuser/shelldocker# docker run -d shellida-python-app
* Serving Flask app 'app' (lazy loading)
* Environment: production
WARNING: This is a development server. Do not use it in a production deployment.
Use a production WSGI server instead.
* Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on http://127.0.0.1:5000
Press CTRL+C to quit
* Restarting with stat
* Debugger is active!
* Debugger PIN: 143-682-713
^C * Serving Flask app 'app' (lazy loading)
* Environment: production
WARNING: This is a development server. Do not use it in a production deployment.
Use a production WSGI server instead.
* Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on all addresses (0.0.0.0)
* Running on http://127.0.0.1:9002
* Running on http://172.17.0.2:9002
Press CTRL+C to quit
Hello, World!

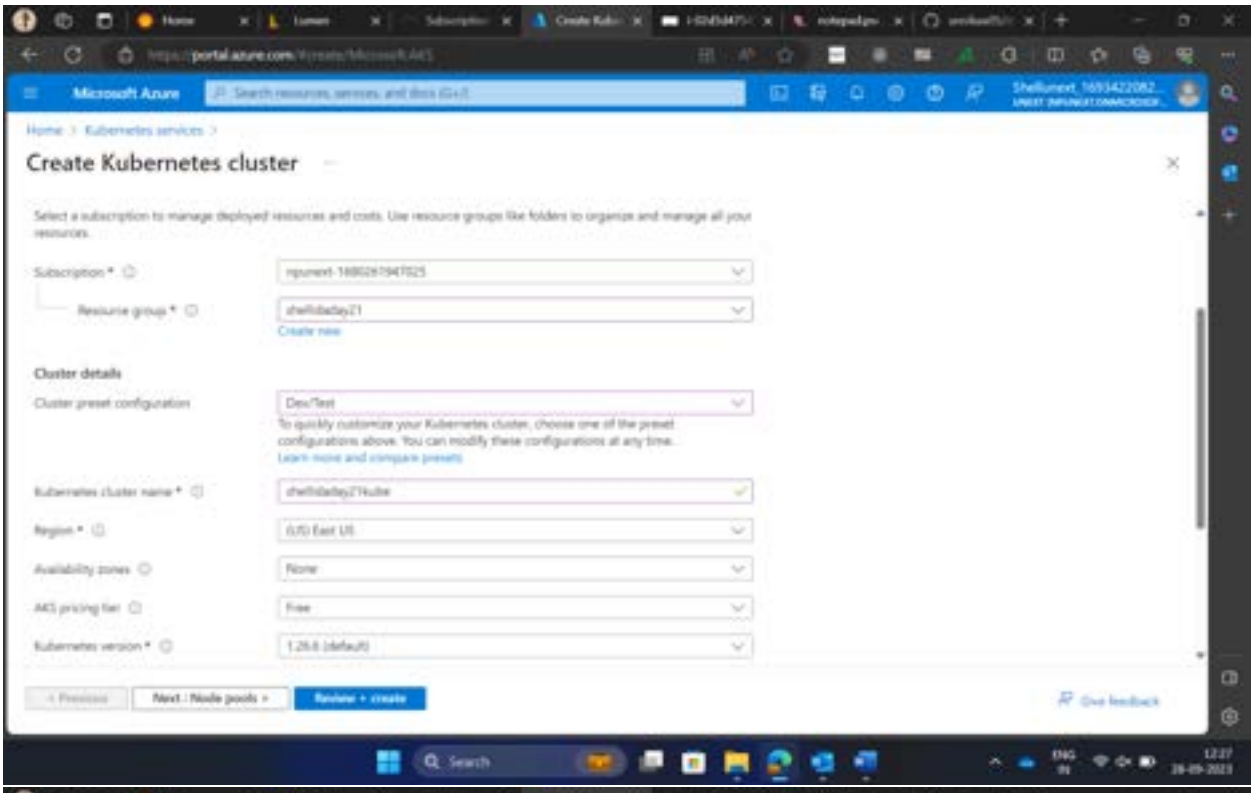

Applications
Mozilla Firefox
Thu 28 Sep, 06:03
labuser
root@ip-172-31-0-134: /home/labuser/shelldocker
File Edit View Search Terminal Help
Log in with a service principal using client certificate.
Read more about the command in reference docs
root@ip-172-31-0-134:/home/labuser/shelldocker# az login -u Shellunext_1693422082390@npunext.onmicrosoft.com -p c2-677a62aN24)
bash: syntax error near unexpected token `)'
root@ip-172-31-0-134:/home/labuser/shelldocker# az login -u Shellunext_1693422082390@npunext.onmicrosoft.com -p 'c2-677a62aN24)'
[
{
"cloudName": "AzureCloud",
"homeTenantId": "dce87315-8ffa-4a01-ab40-0de5a7214b2f",
"id": "30079c75-10d8-4ec4-9514-795b2a09788c",
"isDefault": true,
"managedByTenants": [],
"name": "npunext-1680261947825",
"state": "Enabled",
"tenantId": "dce87315-8ffa-4a01-ab40-0de5a7214b2f",
"user": {
"name": "Shellunext_1693422082390@npunext.onmicrosoft.com",
"type": "user"
}
}
]
root@ip-172-31-0-134:/home/labuser/shelldocker#
Search
ENG
IN
11:33
28-09-2023

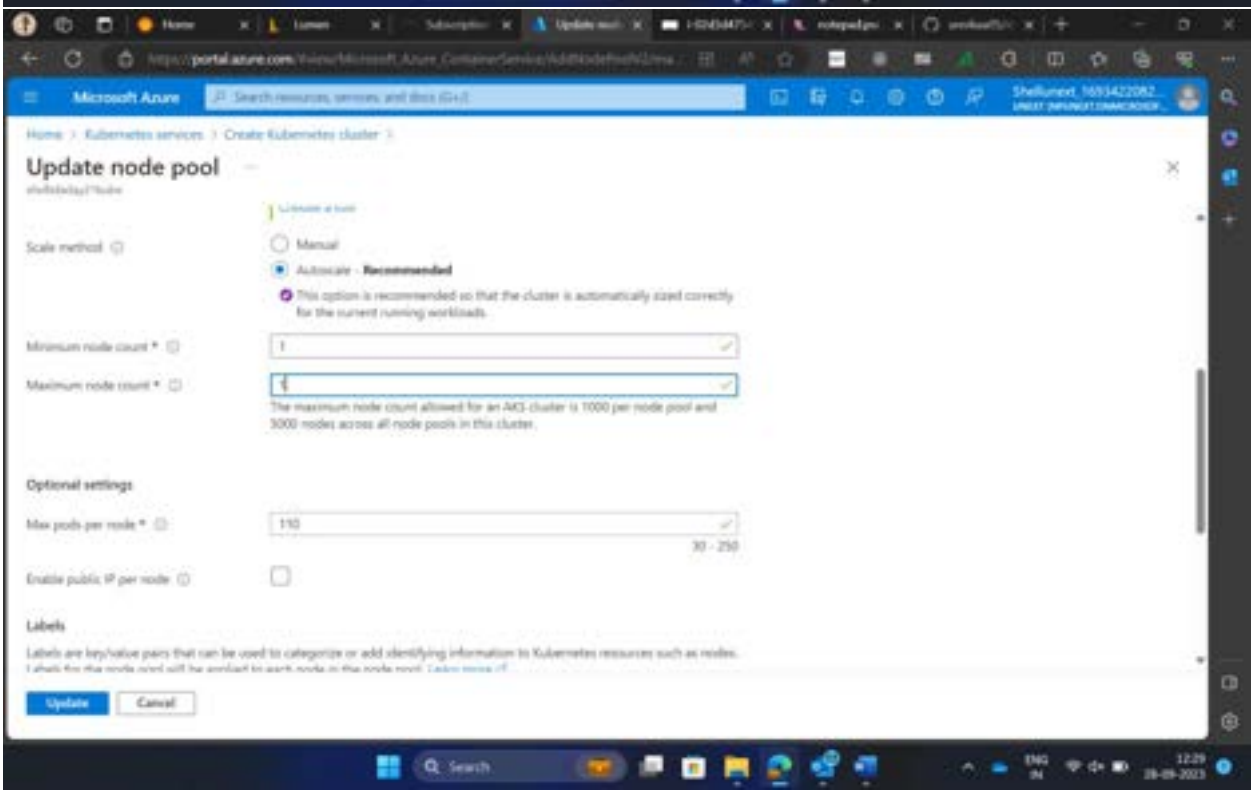

Microsoft Azure
Container registries
Create
Manage view
Refresh
Open query
Save the current columns, sorting, filtering and summary as a view and access your saved views here.
NEXT
Name
Type
Resource group
Location
Subscription
No container registries to display
Build, store, secure, scan, replicate, and manage container images and artifacts with a fully managed, geo-replicated instance of OCI distribution. Connect across environments, including Azure Kubernetes Service and Azure Red Hat OpenShift, and across Azure services like App Service, Machine Learning, and Batch.
Create container registry
Learn more
Give feedback

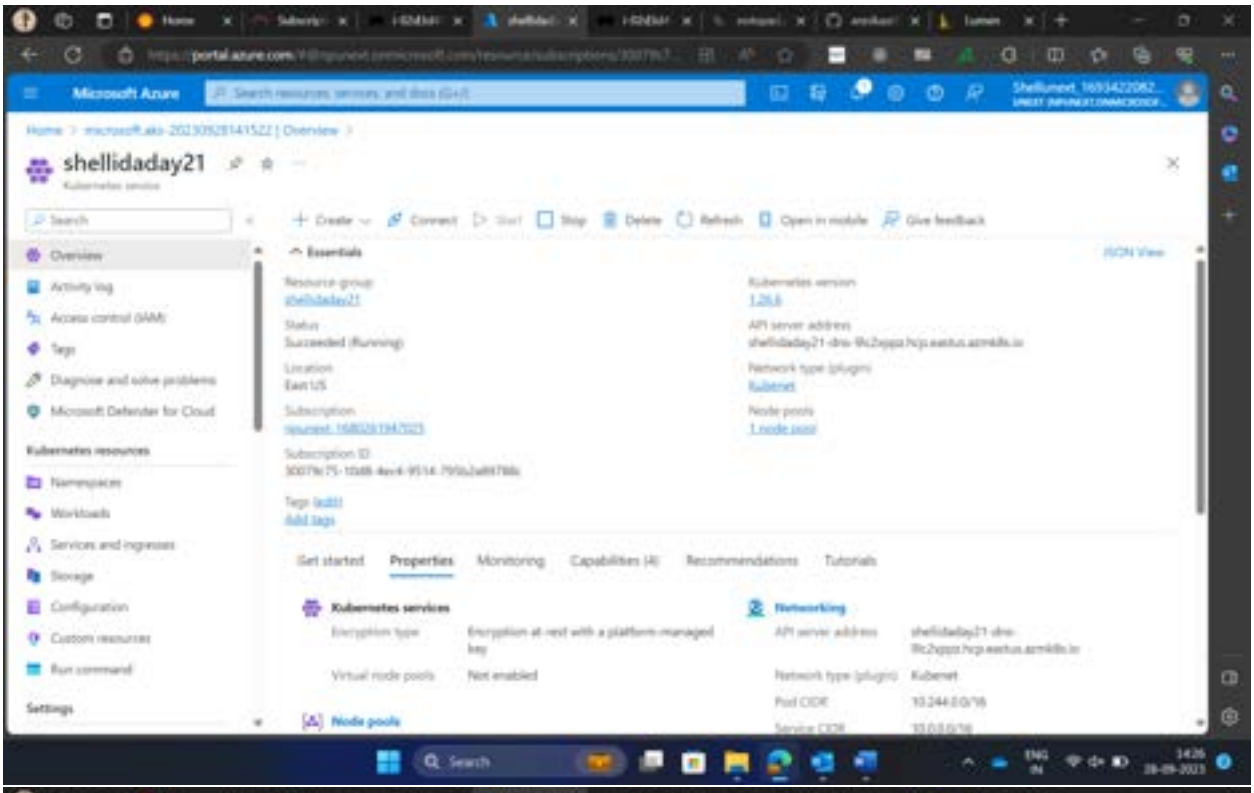

root@ip-172-31-0-134: /home/labuser/shelldocker
File Edit View Search Terminal Help
root@ip-172-31-0-134:/home/labuser/shelldocker# az login -u Shellunext_1693422082398@npunext.onmicrosoft.com -p c2-677a62a824)
bash: syntax error near unexpected token `)'
root@ip-172-31-0-134:/home/labuser/shelldocker# az login -u Shellunext_1693422082398@npunext.onmicrosoft.com -p "c2-677a62a824)"
[
{
"cloudName": "AzureCloud",
"homeTenantId": "dce87315-0ffa-4a01-ab40-0de5a7214b2f",
"id": "30079c75-10d8-4ec4-9514-785b2a897b8c",
"isDefault": true,
"managedByTenants": [],
"name": "npunext-1688261947825",
"state": "Enabled",
"tenantId": "dce87315-0ffa-4a01-ab40-0de5a7214b2f",
"user": {
"name": "Shellunext_1693422082398@npunext.onmicrosoft.com",
"type": "user"
}
}
]
root@ip-172-31-0-134:/home/labuser/shelldocker# az acr login --name azcontainerday21^C
root@ip-172-31-0-134:/home/labuser/shelldocker# sudo su
root@ip-172-31-0-134:/home/labuser/shelldocker# az acr login --name azcontainerday21
Login Succeeded
root@ip-172-31-0-134:/home/labuser/shelldocker#

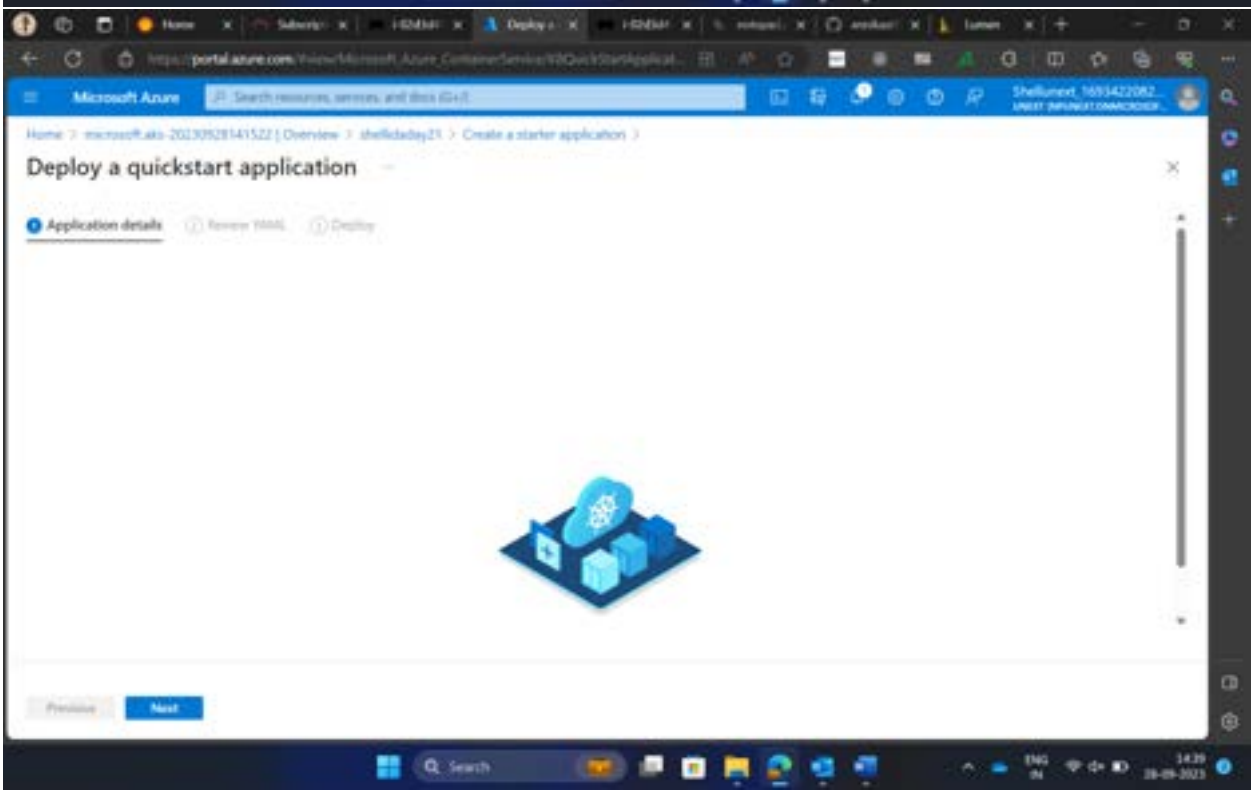

root@ip-172-31-0-134: /home/labuser/shelldocker
File Edit View Search Terminal Help
"tenantId": "dce87315-0ffa-4a81-ab40-80e5a7214b2f",
"user": {
"name": "Shellunext_1693422882390@nunext.onmicrosoft.com",
"type": "user"
}
}
root@ip-172-31-0-134:/home/labuser/shelldocker# az acr login --name azcontainerday21^C
root@ip-172-31-0-134:/home/labuser/shelldocker# sudo su
root@ip-172-31-0-134:/home/labuser/shelldocker# az acr login --name azcontainerday21
Login Succeeded
root@ip-172-31-0-134:/home/labuser/shelldocker# docker tag shellida-python-app azcontainerday21.azurecr.io/ida-python-app:v1
root@ip-172-31-0-134:/home/labuser/shelldocker# docker push azcontainerday21.azurecr.io/ida-python-app:v1
The push refers to repository [azcontainerday21.azurecr.io/ida-python-app]
cc4190eb1e18: Pushed
7ac6bb358fff: Pushed
c9381d2aae3e: Pushed
e6c5004ee77f: Pushed
997b8e79e84f: Pushed
3054512b6f71: Pushed
ae2d557b9c5e: Pushed
e2ef8a51359d: Pushed
v1: digest: sha256:7fd5a40e1a78af898fa3e57e2c4888686cb846494126cb38f4653df6399cae59 size: 1996
root@ip-172-31-0-134:/home/labuser/shelldocker#

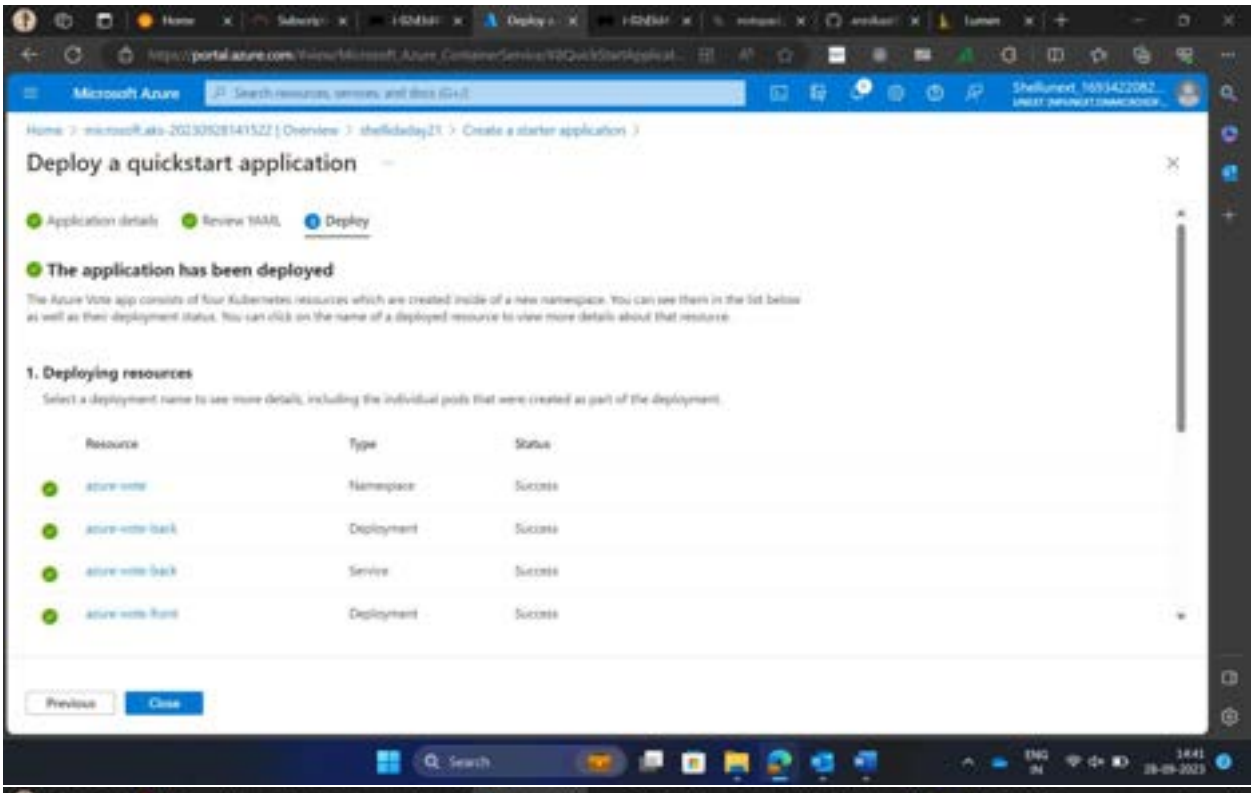

Microsoft Azure
azcontainerday21 | Repositories
Refresh
Manage Deleted Repositories
Overview
Activity log
Access control (IAM)
Tags
Quick start
Events
Settings
Access keys
Encryption
Identity
Networking
Microsoft Defender for Cloud
Properties
Locks
Services
Repositories
Cache Rule
ida-python-app

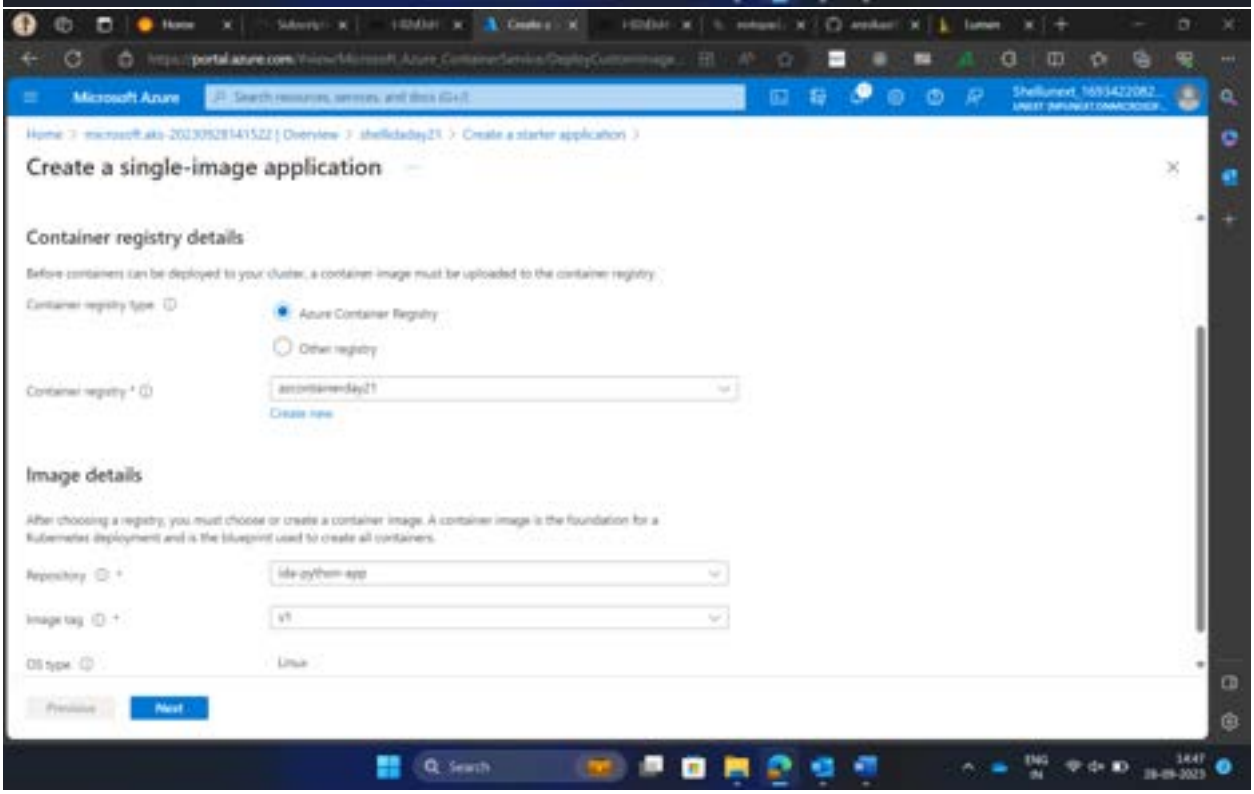

Microsoft Azure
Create Kubernetes cluster
Select a subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.
Subscription *
Resource group *
Create new
Cluster details
Cluster preset configuration
Dev/Test
To quickly customize your Kubernetes cluster, choose one of the preset configurations above. You can modify these configurations at any time.
Learn more and compare presets
Kubernetes cluster name *
Region *
Availability zones
None
AKS pricing tier
Free
Kubernetes version *
1.26.6 (default)
Next : Node pools >
Review + create
Give feedback

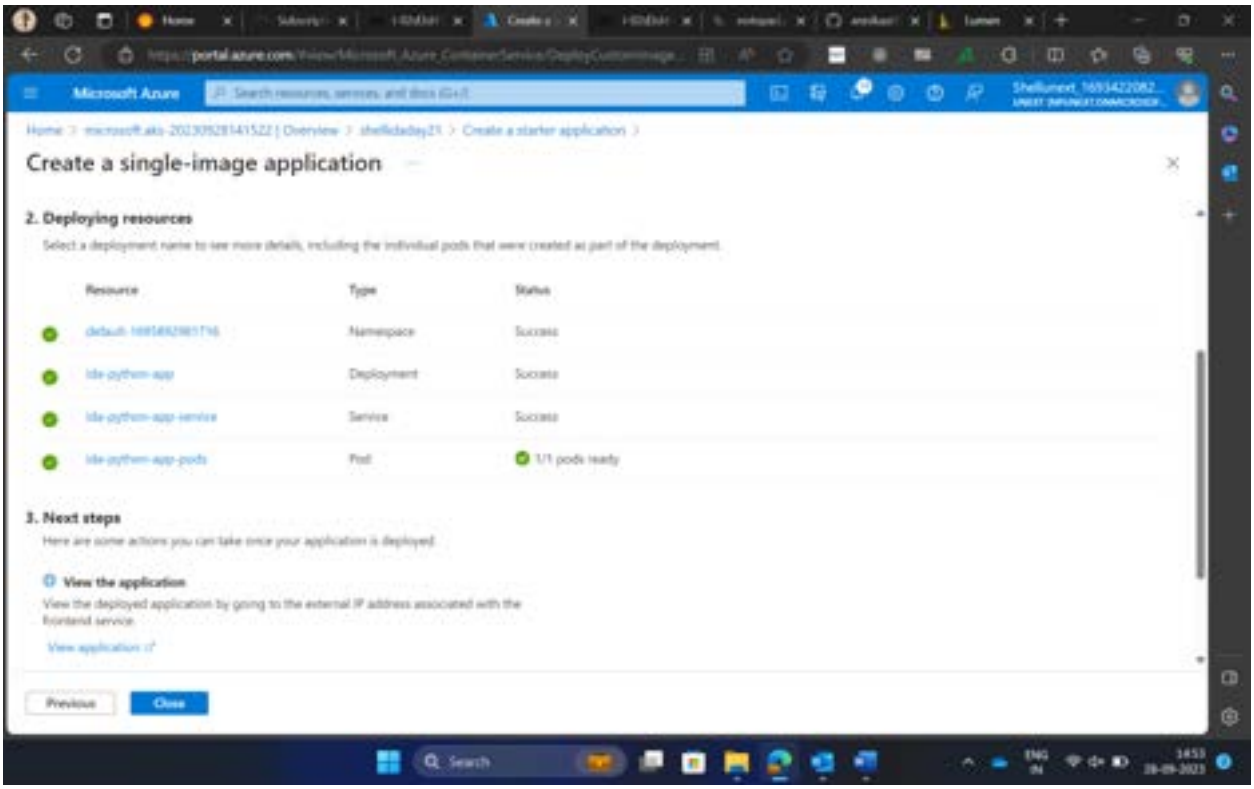

Microsoft Azure
Home > Kubernetes services > Create Kubernetes cluster >
Update node pool
Scale method
Manual
Autoscale - Recommended
This option is recommended so that the cluster is automatically sized correctly for the current running workloads.
Minimum node count *
1
Maximum node count *
The maximum node count allowed for an AKS cluster is 1000 per node pool and 5000 nodes across all node pools in this cluster.
Optional settings
Max pods per node *
110
30 - 250
Enable public IP per node
Labels
Update
Cancel


Microsoft Azure
shellidaday21
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Microsoft Defender for Cloud
Kubernetes resources
Namespaces
Workloads
Services and ingresses
Storage
Configuration
Custom resources
Run command
Settings
Create
Connect
Start
Stop
Delete
Refresh
Open in mobile
Give feedback
Essentials
JSON View
Resource group
shellidaday21
Status
Succeeded (Running)
Location
East US
Subscription
Subscription ID
Tags (edit)
Add tags
Kubernetes version
1.26.6
API server address
Network type (plugin)
Kubenet
Node pools
1 node pool
Get started
Properties
Monitoring
Capabilities (4)
Recommendations
Tutorials
Kubernetes services
Encryption type
Encryption at rest with a platform-managed key
Virtual node pools
Not enabled
Networking
API server address
Network type (plugin)
Kubenet
Pod CIDR
10.244.0.0/16
Node pools

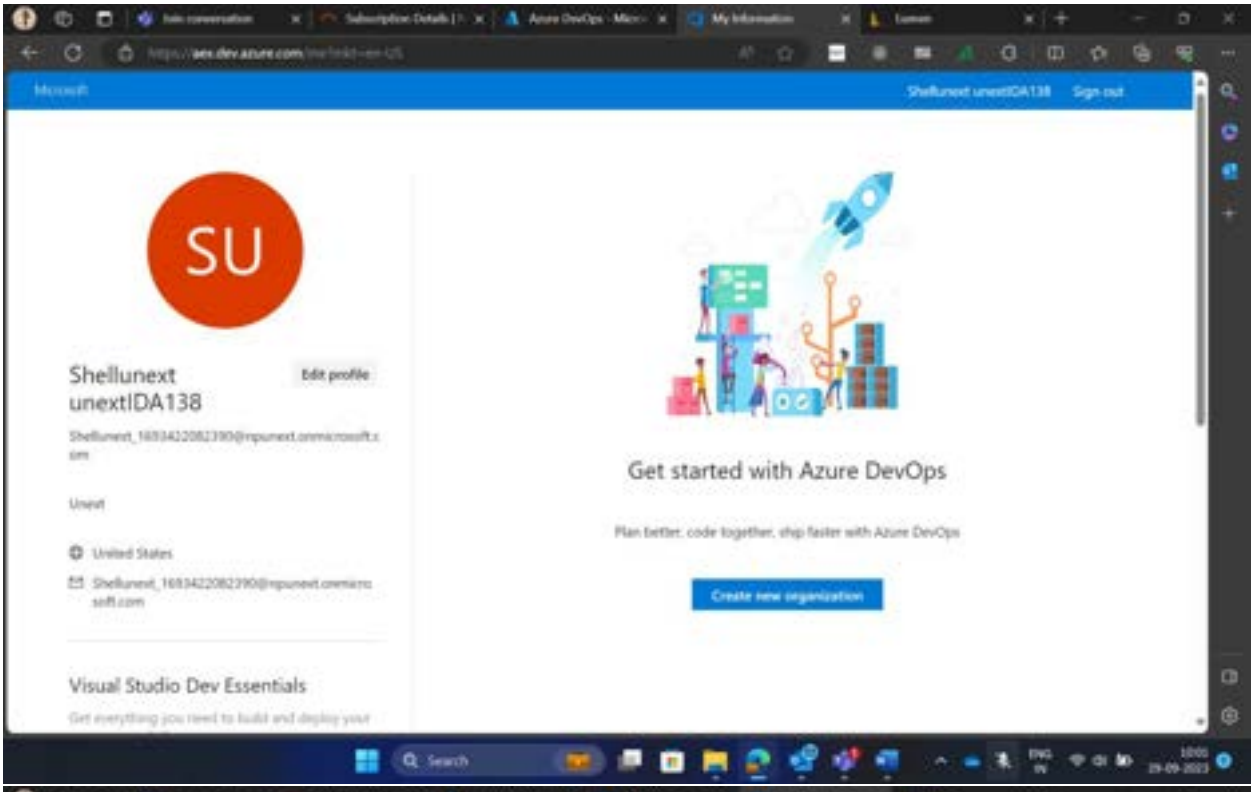

Microsoft Azure
Deploy a quickstart application
Application details
Review YAML
Deploy
Previous
Next

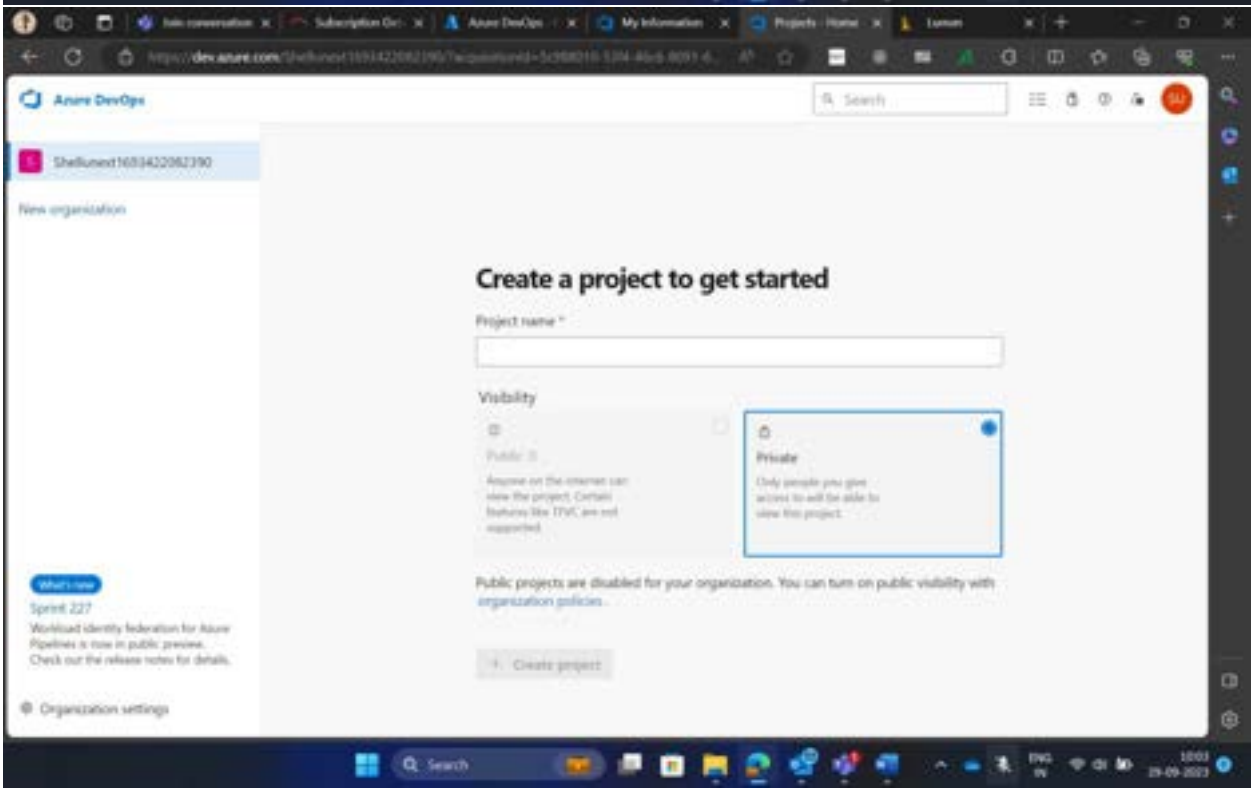

Microsoft Azure
Deploy a quickstart application
Application details
Review YAML
Deploy
The application has been deployed
1. Deploying resources
Resource
Type
Status
azure-vote
Namespace
Success
azure-vote-back
Deployment
Success
azure-vote-back
Service
Success
azure-vote-front
Deployment
Success
Previous
Close

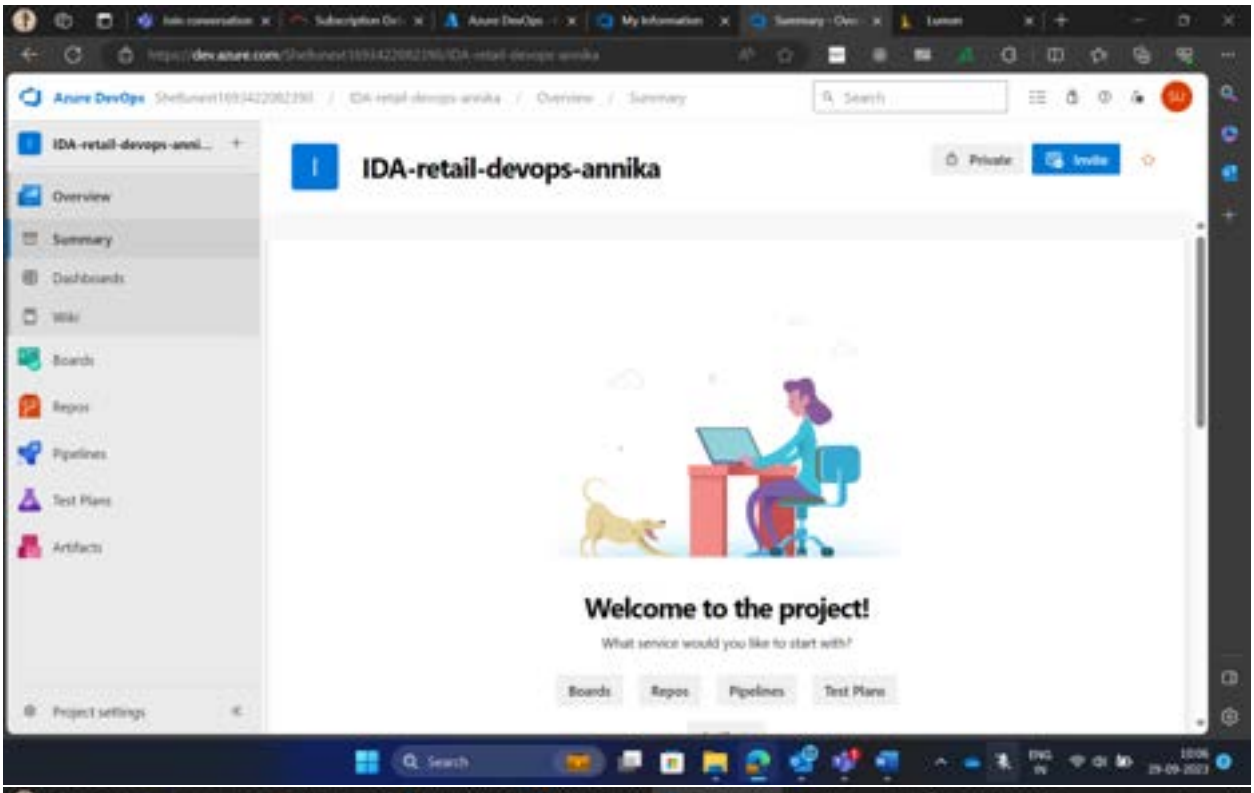

Microsoft Azure
Create a single-image application
Container registry details
Before containers can be deployed to your cluster, a container image must be uploaded to the container registry.
Azure Container Registry
Other registry
Create new
Image details
After choosing a registry, you must choose or create a container image. A container image is the foundation for a Kubernetes deployment and is the blueprint used to create all containers.
Linux
Next

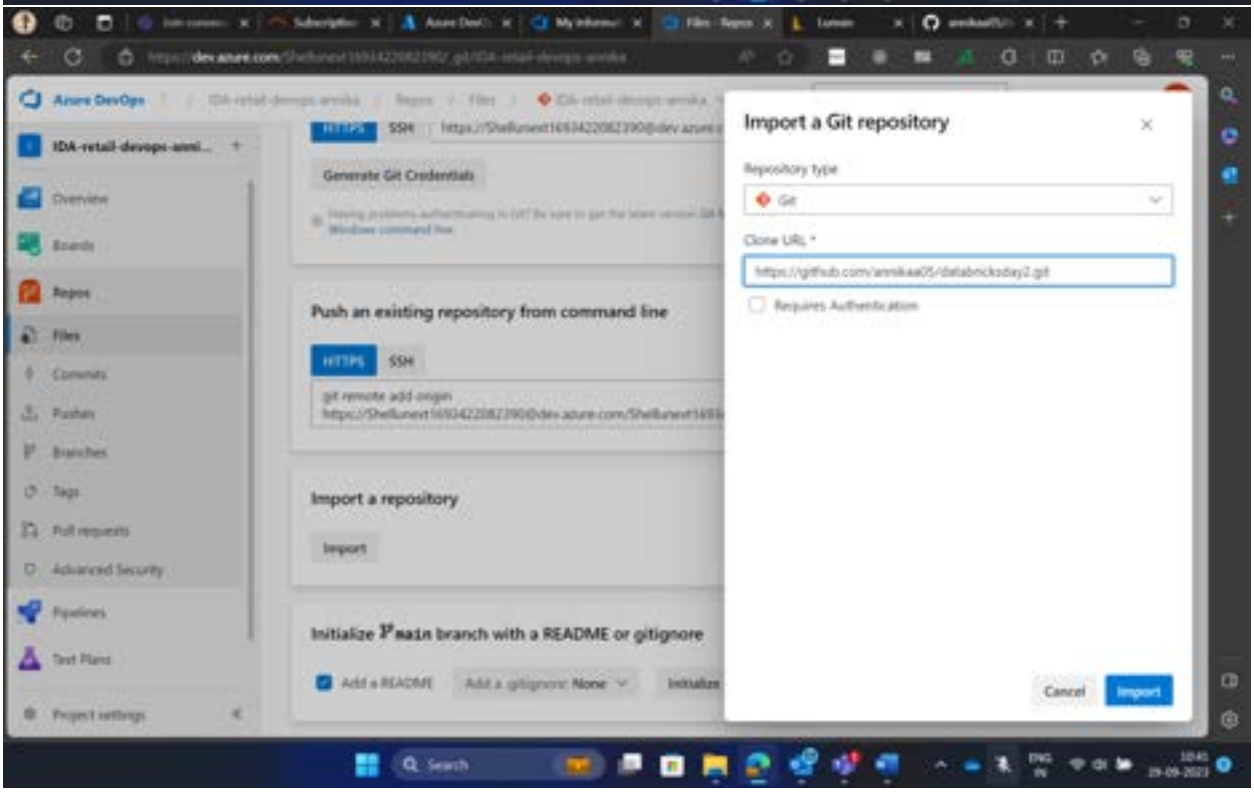

Microsoft Azure
Create a single-image application
2. Deploying resources
Select a deployment name to see more details, including the individual pods that were created as part of the deployment.
Resource
Type
Status
default-1695492981716
Namespace
Success
ide-python-app
Deployment
Success
ide-python-app-service
Service
Success
ide-python-app-pods
Pod
1/1 pods ready
3. Next steps
Here are some actions you can take once your application is deployed.
View the application
View the deployed application by going to the external IP address associated with the frontend service.
View application
Previous
Close

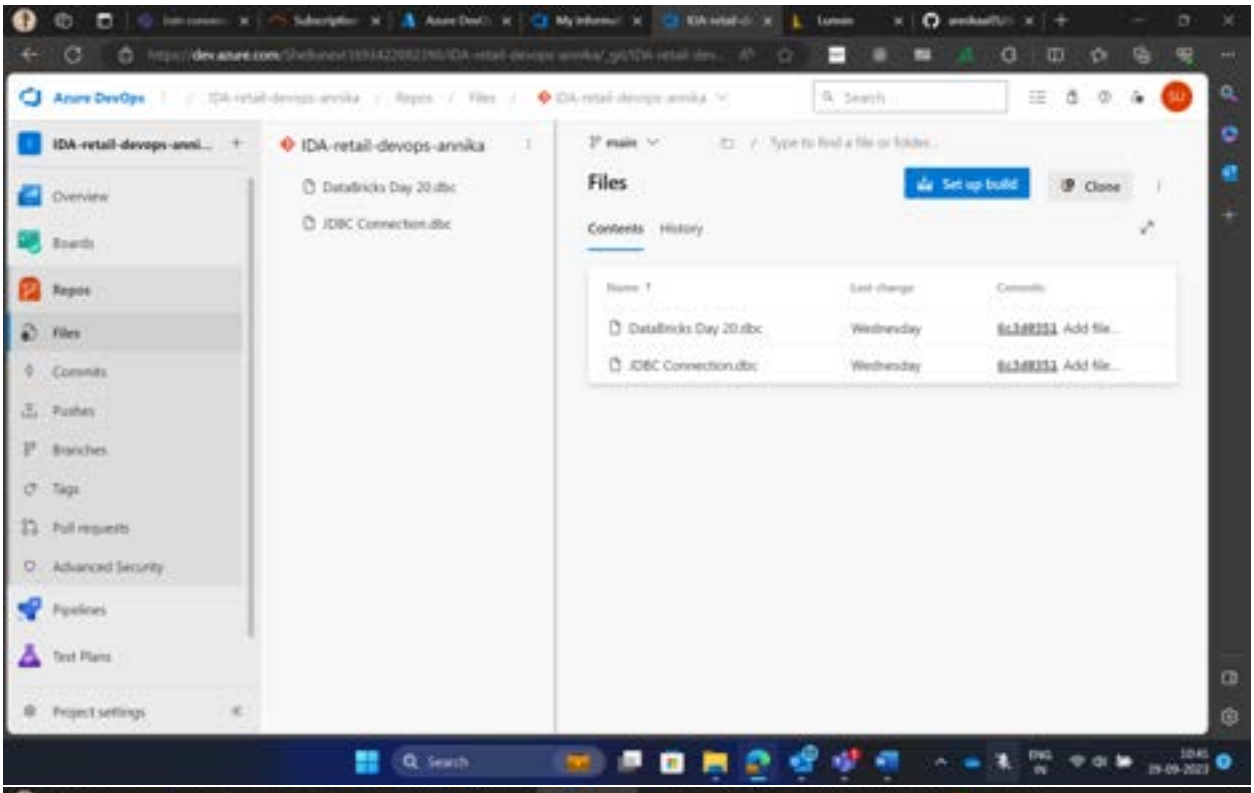
Day 22

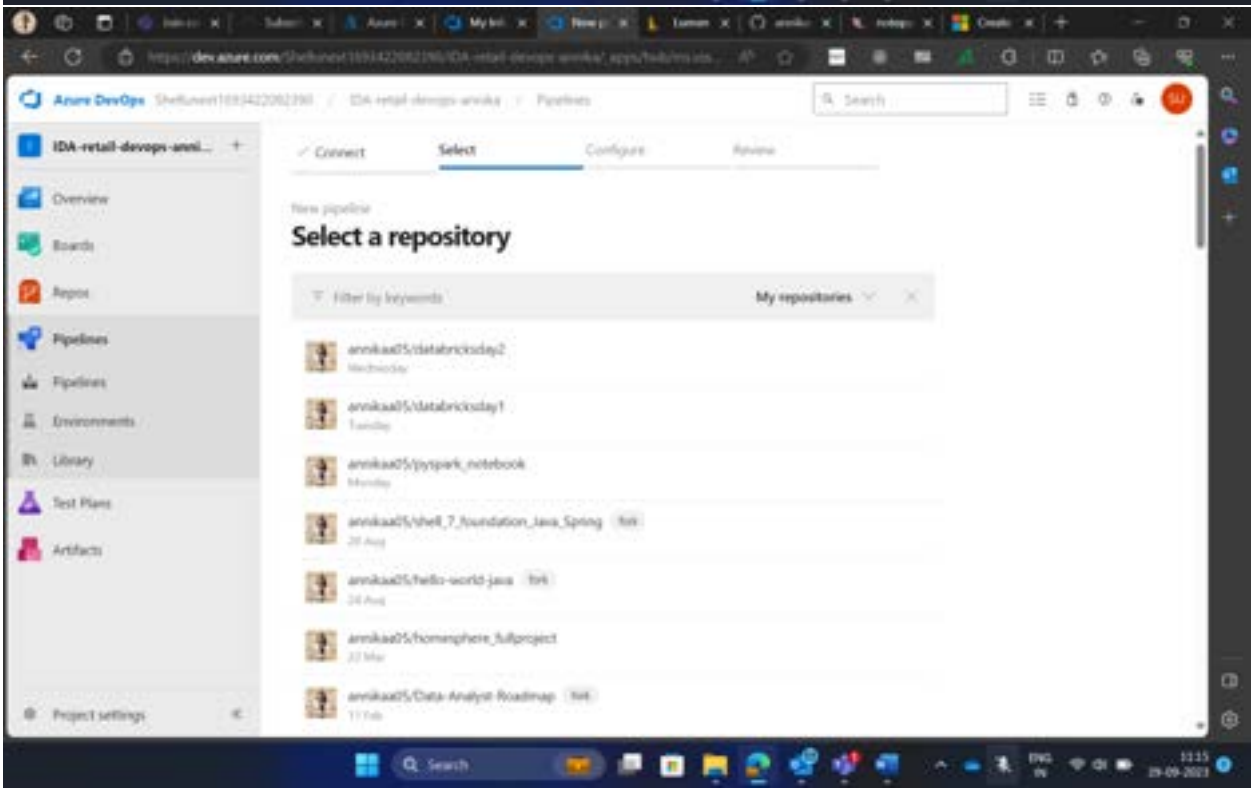

Microsoft
Shellunext unextIDA138 Sign out
SU
Shellunext unextIDA138
Edit profile
Unext
United States
Visual Studio Dev Essentials
Get started with Azure DevOps
Plan better, code together, ship faster with Azure DevOps
Create new organization

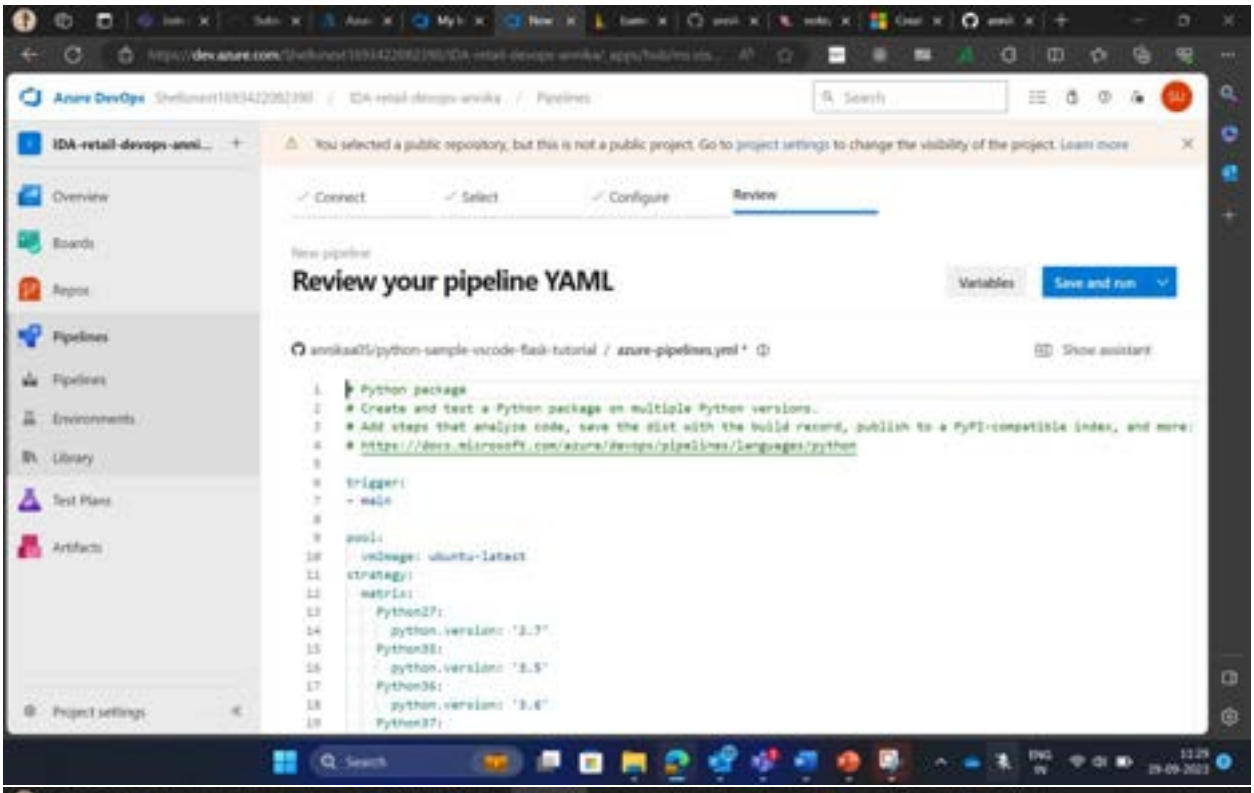

Azure DevOps
Shellunext1693422082390
New organization
Create a project to get started
Project name *
Visibility
Private
Public projects are disabled for your organization. You can turn on public visibility with organization policies.
Create project
What's new
Sprint 227
Organization settings

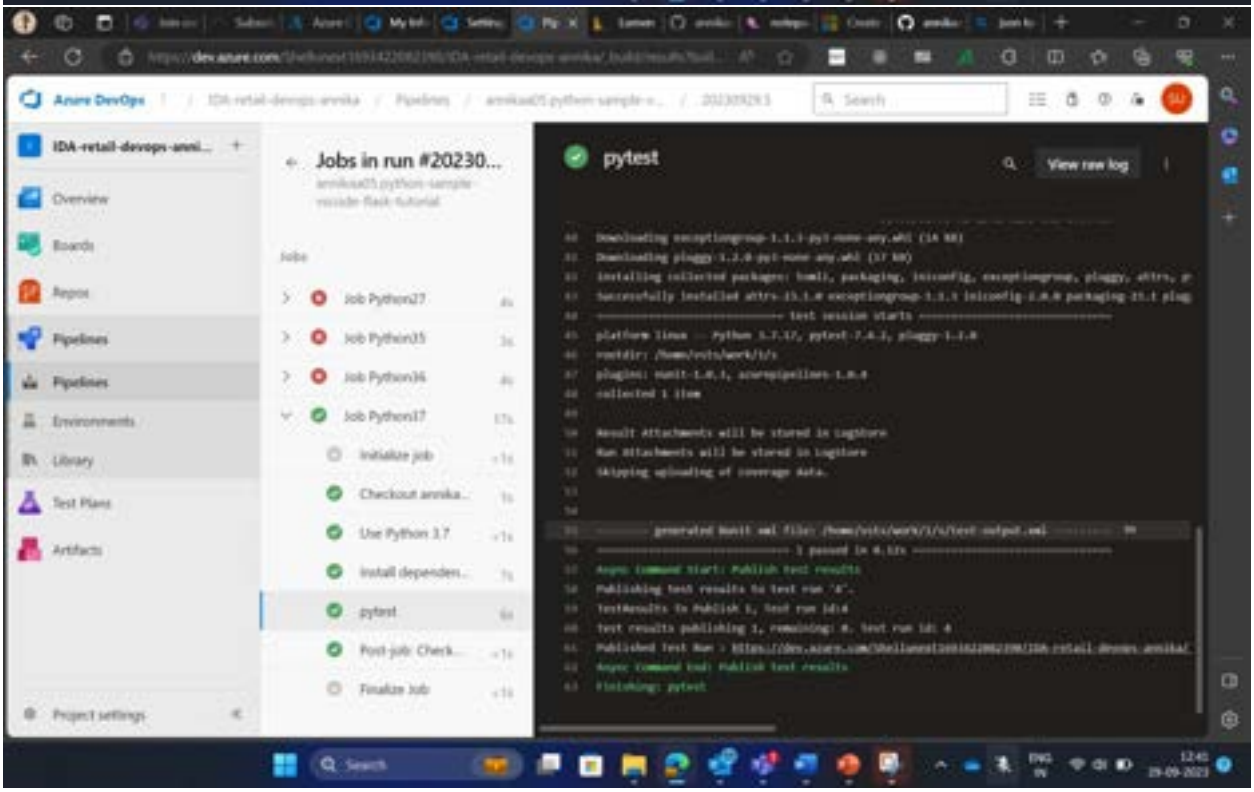

Azure DevOps
IDA-retail-devops-annika
Overview
Summary
Dashboards
Wiki
Boards
Repos
Pipelines
Test Plans
Artifacts
Project settings
Private
Invite
Welcome to the project!
What service would you like to start with?
Boards
Repos
Pipelines
Test Plans

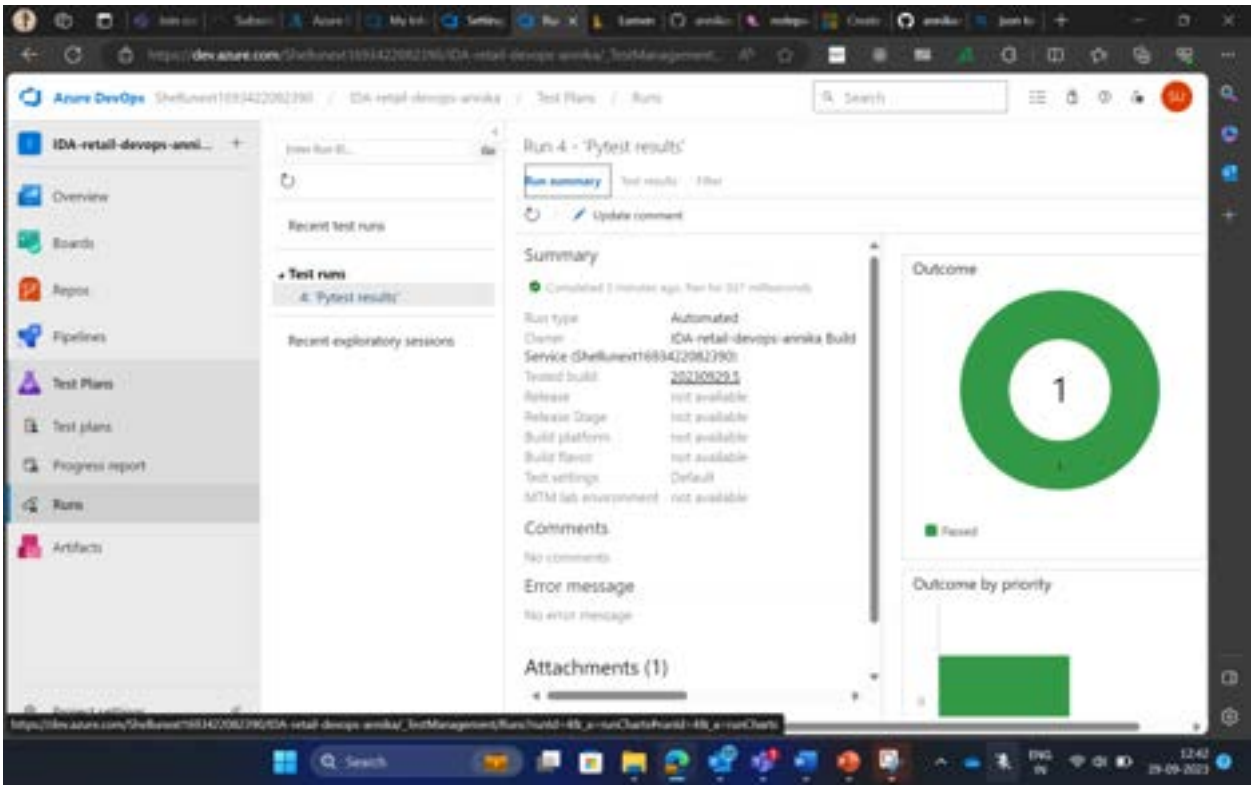

Azure DevOps
Generate Git Credentials
Push an existing repository from command line
HTTPS
SSH
Import a repository
Import
Initialize main branch with a README or gitignore
Add a README
Import a Git repository
Repository type
Git
Clone URL *
Requires Authentication
Cancel
Import


Azure DevOps
IDA-retail-devops-anni...
Overview
Boards
Repos
Files
Commits
Pushes
Branches
Tags
Pull requests
Advanced Security
Pipelines
Test Plans
Project settings
IDA-retail-devops-annika
Databricks Day 20.dbc
JDBC Connection.dbc
main
Files
Set up build
Clone
Contents
History
Name
Last change
Commits
Databricks Day 20.dbc
Wednesday
6c1489351 Add file...
JDBC Connection.dbc
Wednesday
6c1489351 Add file...


Azure DevOps
IDA-retail-devops-anni...
Overview
Boards
Repos
Pipelines
Pipelines
Environments
Library
Test Plans
Artifacts
Project settings
Connect
Select
Configure
Review
New pipeline
Select a repository
Filter by keywords
My repositories
annikaa05/databricksday2
annikaa05/databricksday1
annikaa05/pyspark_notebook
annikaa05/shell_7_foundation_Java_Spring
fork
annikaa05/hello-world-java
fork
annikaa05/homesphere_fullproject
annikaa05/Data-Analyst-Roadmap
fork


Azure DevOps
IDA-retail-devops-anni...
Overview
Boards
Repos
Pipelines
Pipelines
Environments
Library
Test Plans
Artifacts
Project settings
You selected a public repository, but this is not a public project. Go to project settings to change the visibility of the project. Learn more
Connect
Select
Configure
Review
Review your pipeline YAML
Variables
Save and run
annikaa05/python-sample-vscode-flask-tutorial / azure-pipelines.yml *
Show assistant
Python package
Create and test a Python package on multiple Python versions.
Add steps that analyze code, save the dist with the build record, publish to a PyPI-compatible index, and more:
https://docs.microsoft.com/azure/devops/pipelines/languages/python
trigger:
- main
pool:
vmImage: ubuntu-latest
strategy:
matrix:
Python27:
python.version: '2.7'
Python35:
python.version: '3.5'
Python36:
python.version: '3.6'
Python37:


Azure DevOps
IDA-retail-devops-anni...
Overview
Boards
Repos
Pipelines
Pipelines
Environments
Library
Test Plans
Artifacts
Project settings
Jobs in run #20230...
Jobs
Job Python27
Job Python35
Job Python36
Job Python37
Initialize job
Checkout annika...
Use Python 3.7
Install dependen...
pytest
Post-job: Check...
Finalize Job
pytest
View raw log
Downloading exceptiongroup-1.1.1-py3-none-any.whl (14 kB)
Downloading pluggy-1.2.0-py3-none-any.whl (17 kB)
Installing collected packages: toml, packaging, iniconfig, exceptiongroup, pluggy, attrs, p
Successfully installed attrs-23.1.0 exceptiongroup-1.1.1 iniconfig-2.0.0 packaging-23.1 plug
test session starts
platform linux -- Python 3.7.17, pytest-7.4.2, pluggy-1.2.0
rootdir: /home/vsts/work/1/s
plugins: xunit-1.0.1, azurepipelines-1.0.4
collected 1 item
Result Attachments will be stored in logStore
Run Attachments will be stored in logStore
Skipping uploading of coverage data.
generated Xunit xml file: /home/vsts/work/1/s/test-output.xml
1 passed in 0.12s
Async Command Start: Publish test results
Publishing test results to test run '4'.
TestResults to Publish 1, test run id:4
test results publishing 1, remaining: 0. test run id: 4
Published Test Run : https://dev.azure.com/Shellunext1693422082390/IDA-retail-devops-annikaa
Async Command End: Publish test results
Finishing: pytest


Azure DevOps
IDA-retail-devops-anni...
Overview
Boards
Repos
Pipelines
Test Plans
Test plans
Progress report
Runs
Artifacts
Recent test runs
Test runs
4 'Pytest results'
Recent exploratory sessions
Run 4 - 'Pytest results'
Run summary
Update comment
Summary
Run type
Automated
Owner
IDA-retail-devops-anniks Build Service (Shelluser0169342082390)
Tested build
20230529.5
Release
not available
Release Stage
not available
Build platform
not available
Build flavor
not available
Test settings
Default
MTM lab environment
not available
Comments
No comments
Error message
No error message
Attachments (1)
Outcome
1
Passed
Outcome by priority

Boards:
Work items :

Epic: High-level features or objectives for your PySpark application.
Feature: Major functionalities or components within an Epic.

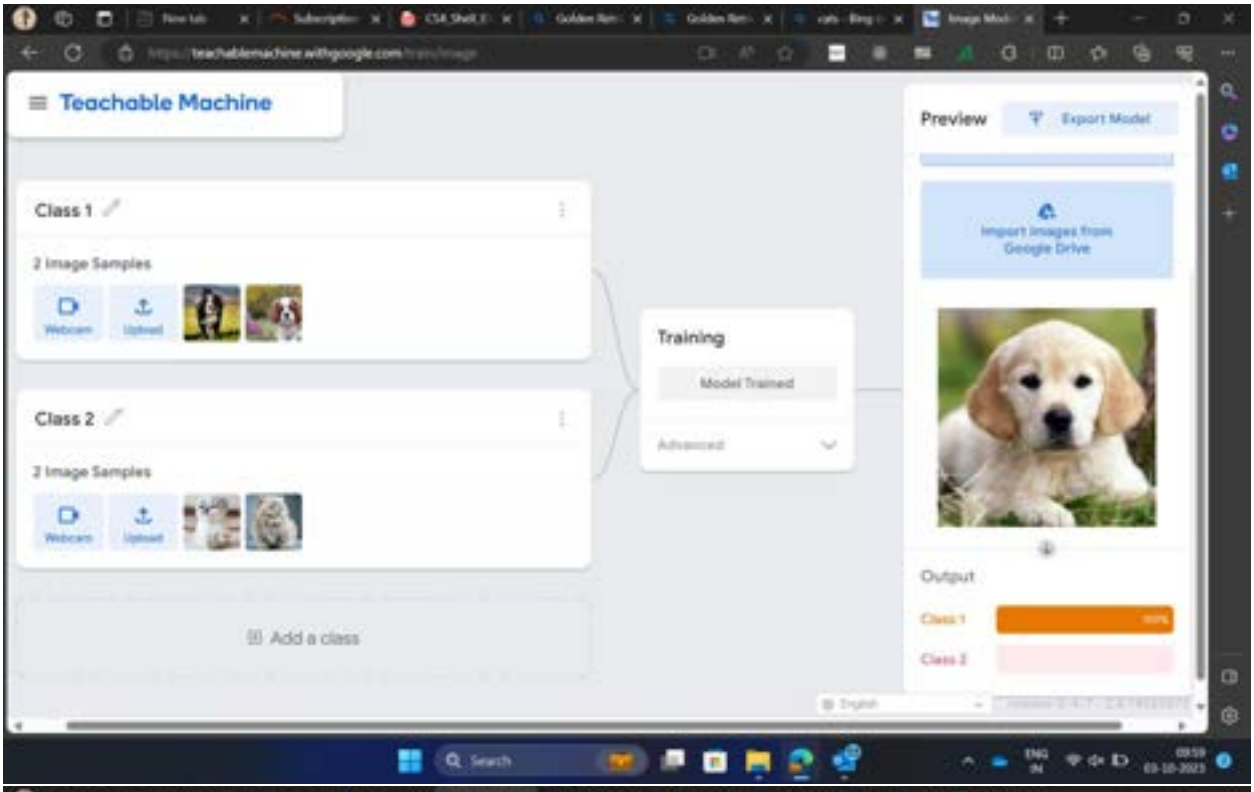
User Story: User-centric descriptions of functionality.
Task: Individual development tasks related to User Stories.

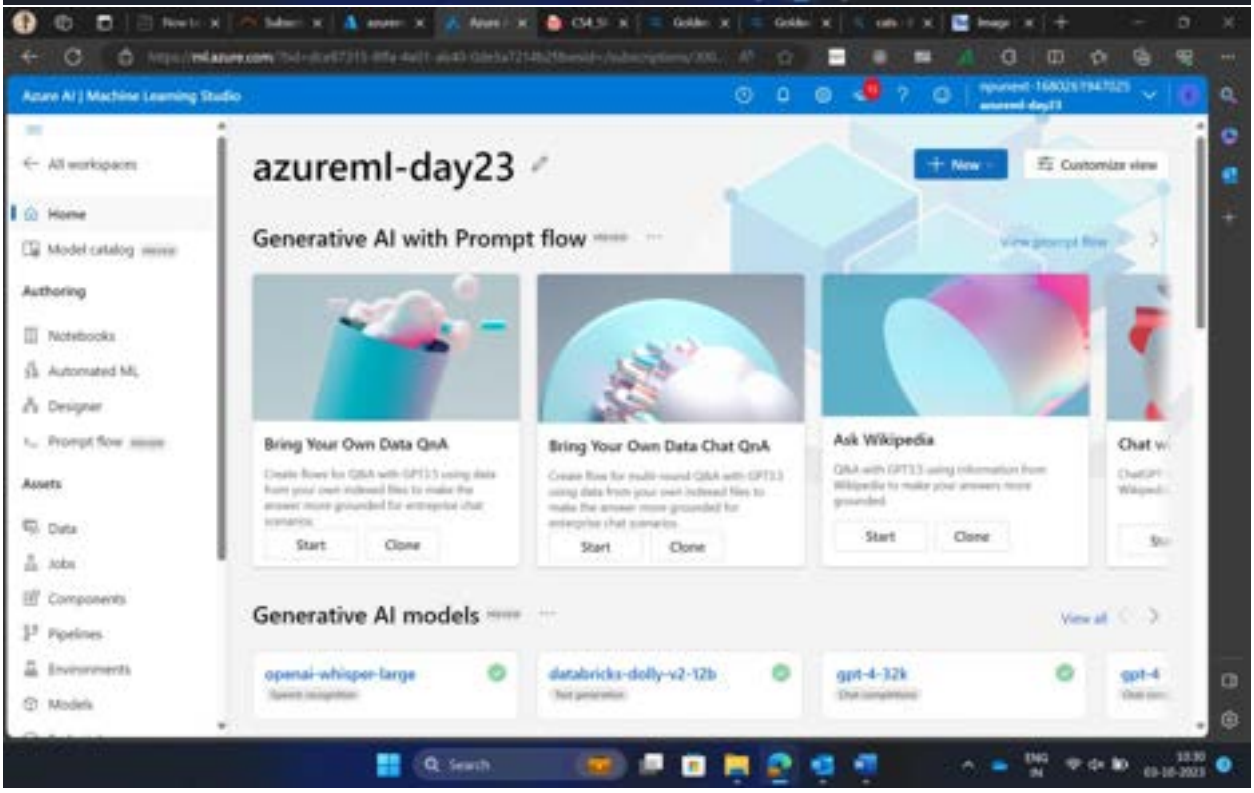
Work items :

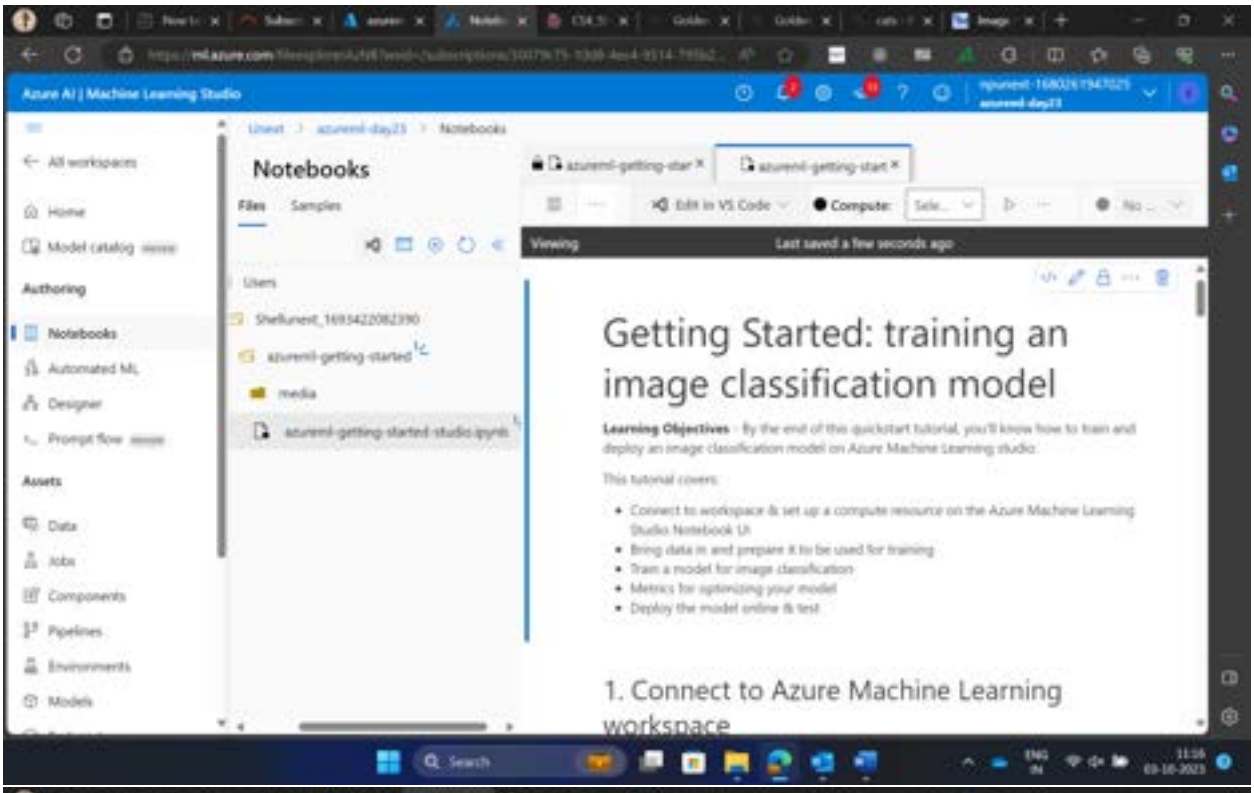
Epic: High-level features or objectives for your PySpark application.
Feature: Major functionalities or components within an Epic.
Data ingestion, cleansing , scrubbing , transformation , analysis

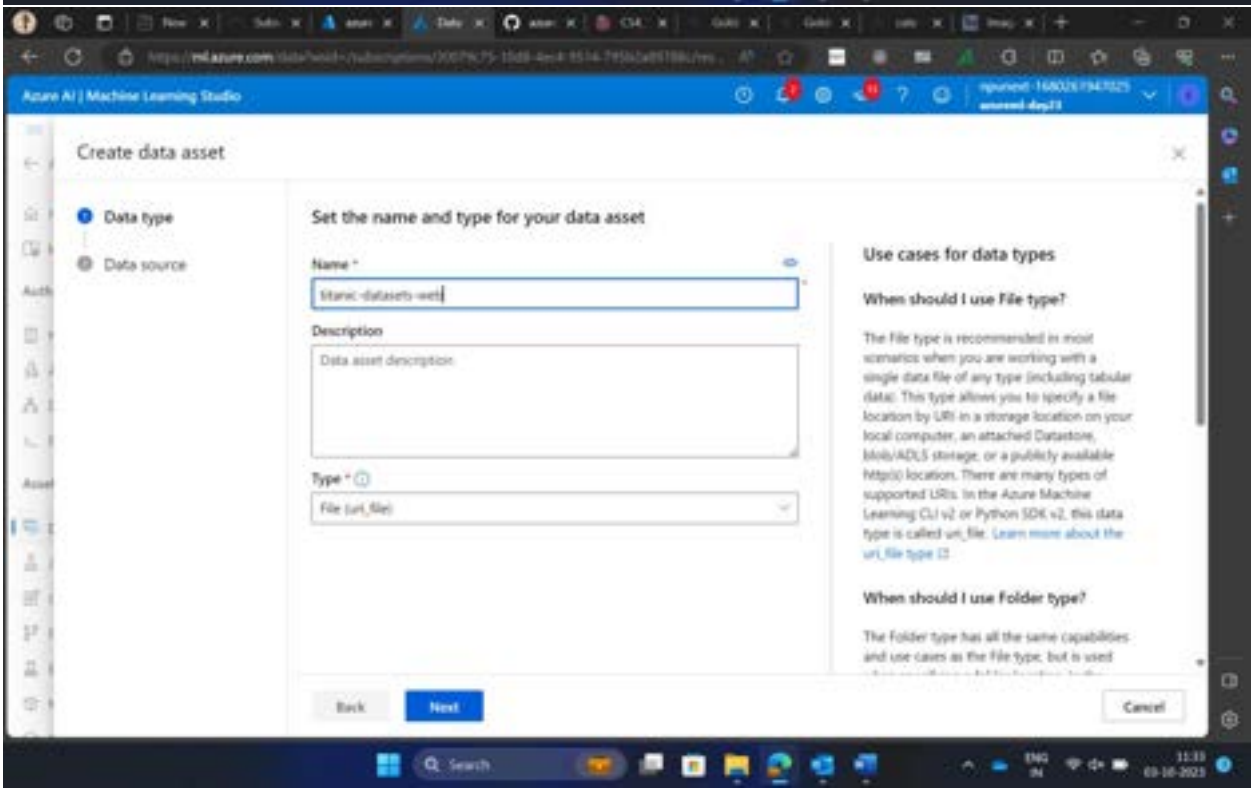
User Story: User-centric descriptions of functionality.

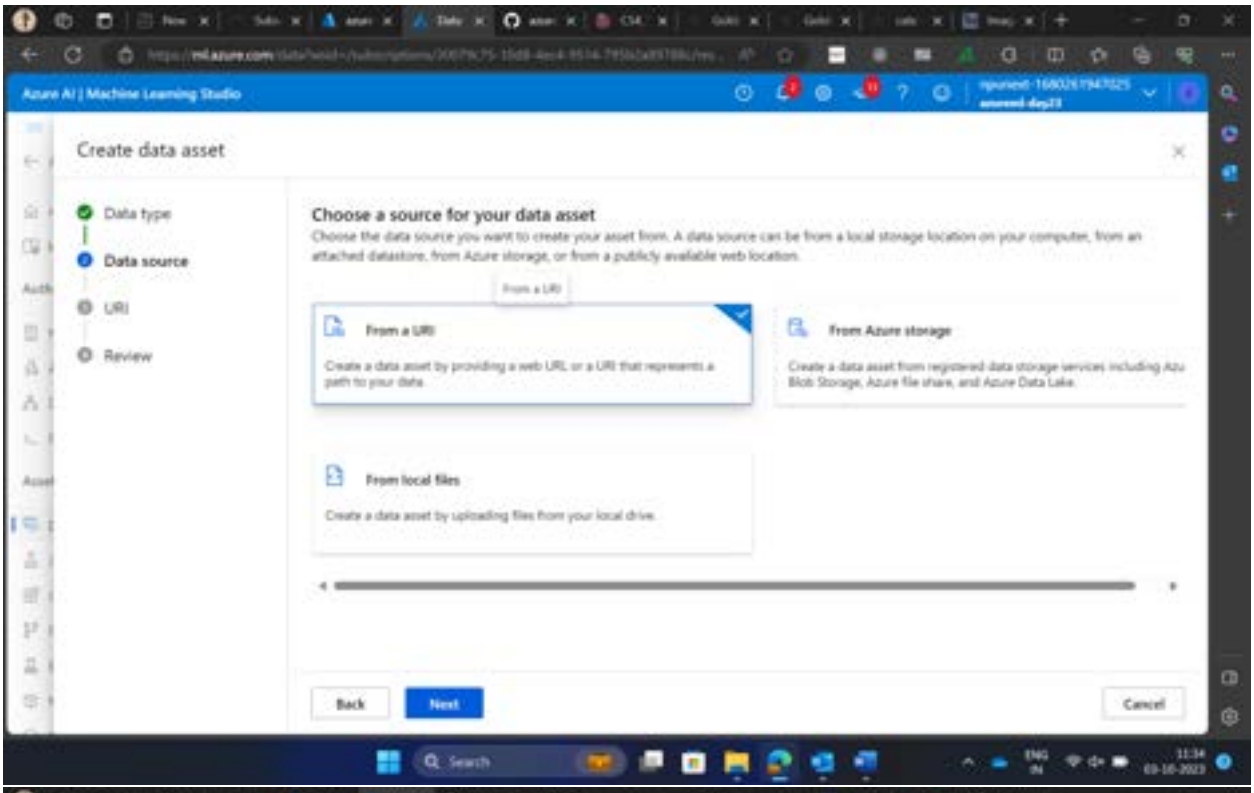
Day 23 03.10.23

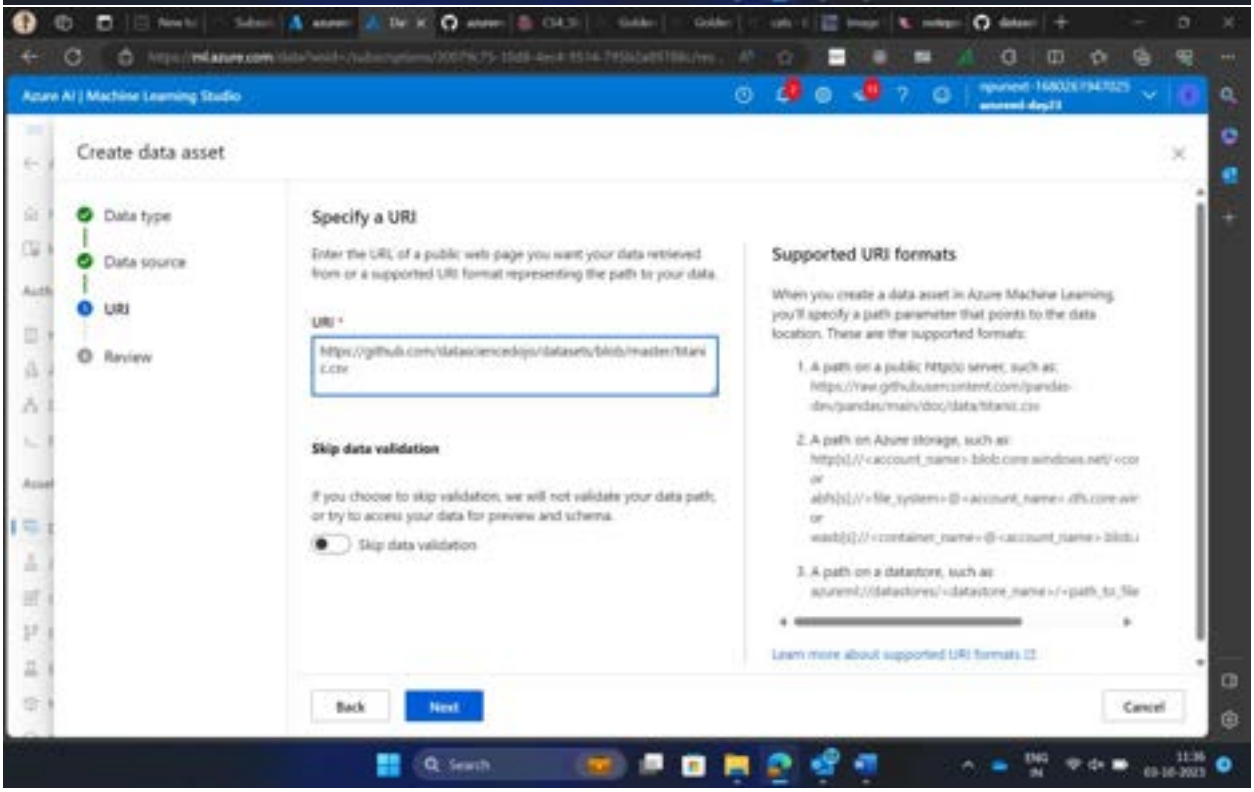


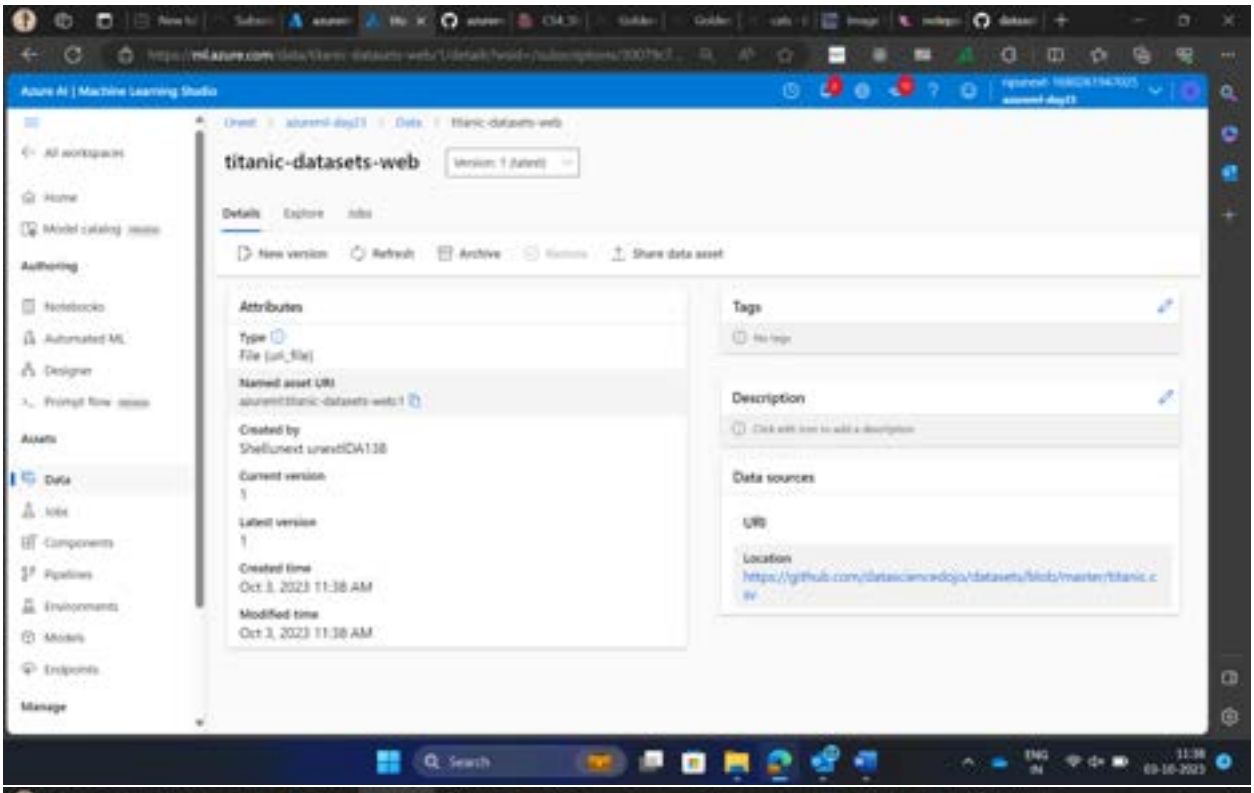


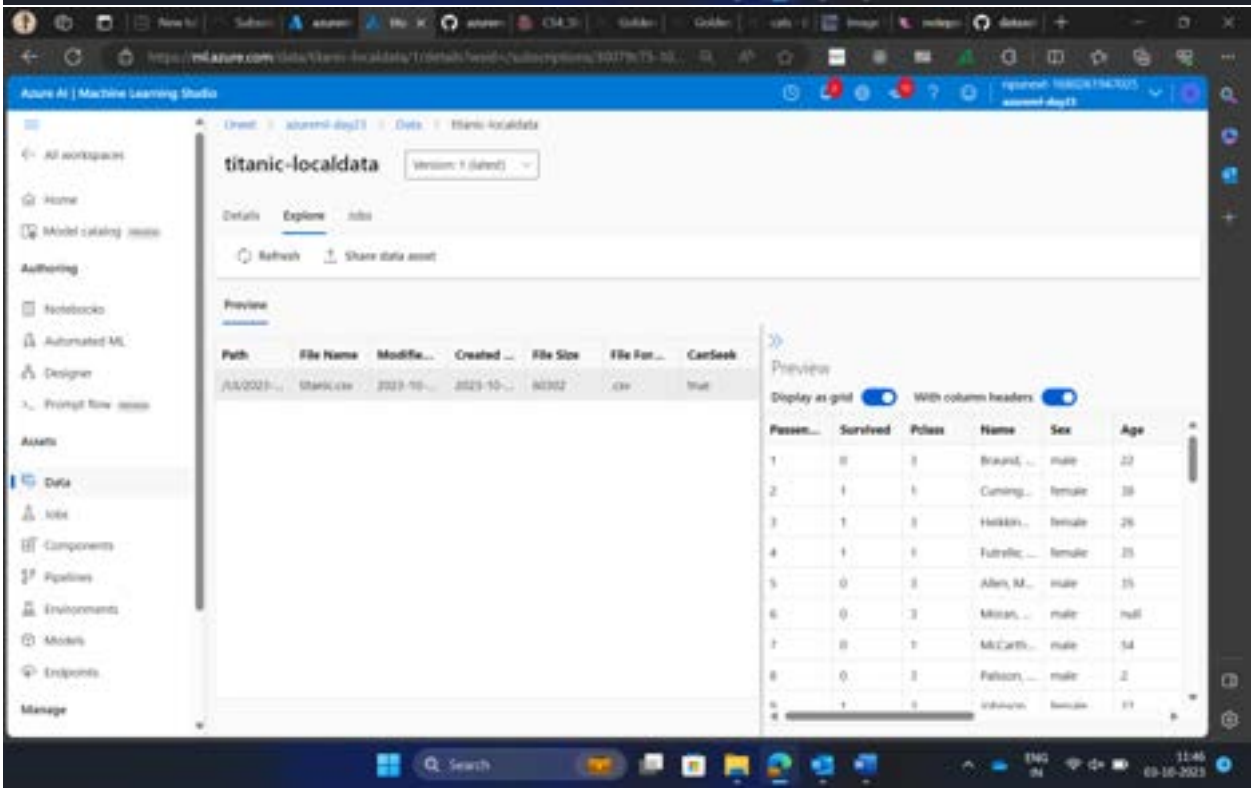

Azure AI | Machine Learning Studio
Notebooks
Files
Samples
azureml-getting-start
Edit in VS Code
Compute:
Viewing
Last saved a few seconds ago
Getting Started: training an image classification model
Learning Objectives - By the end of this quickstart tutorial, you'll know how to train and deploy an image classification model on Azure Machine Learning studio.
This tutorial covers
Connect to workspace & set up a compute resource on the Azure Machine Learning Studio Notebook UI
Bring data in and prepare it to be used for training
Train a model for image classification
Metrics for optimizing your model
Deploy the model online & test
1. Connect to Azure Machine Learning workspace

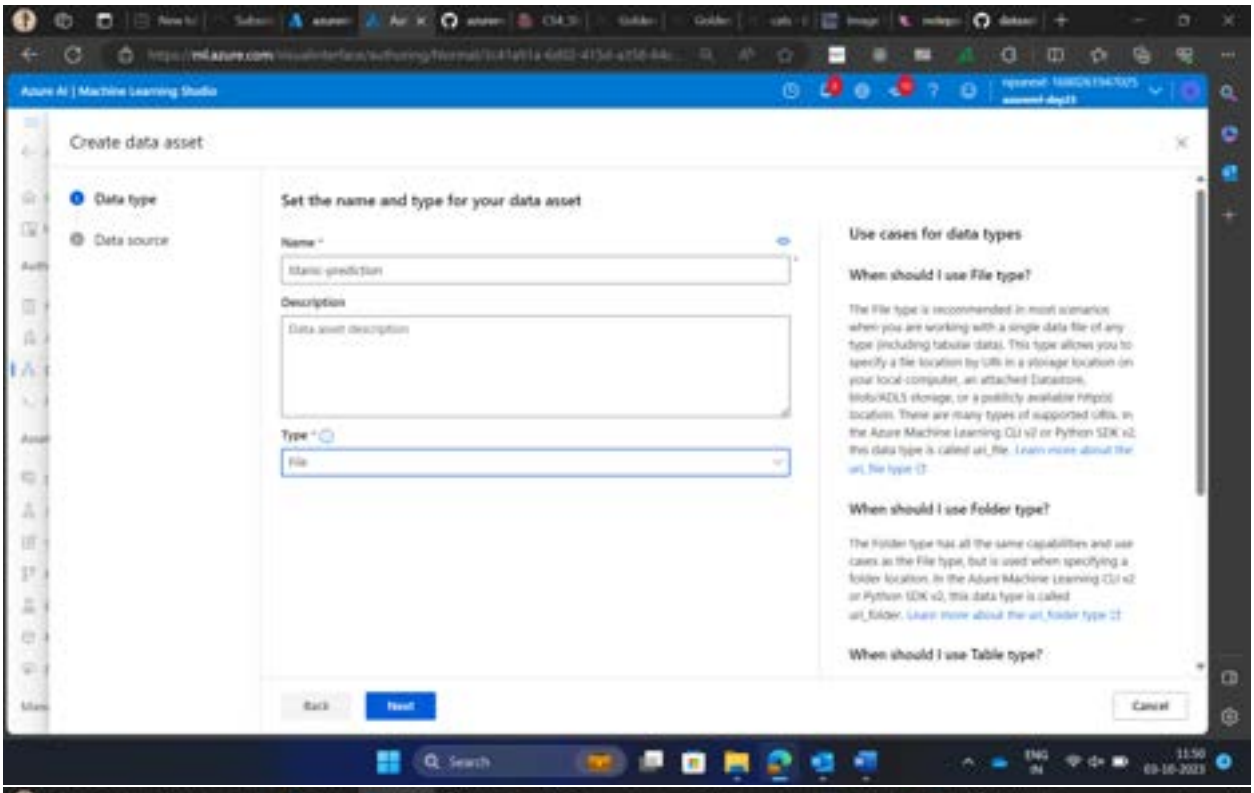

Azure AI | Machine Learning Studio
Create data asset
Data type
Data source
Set the name and type for your data asset
Name *
titanic-datasets-web
Description
Data asset description
Type *
File (uri_file)
Use cases for data types
When should I use File type?
When should I use Folder type?
Back
Next
Cancel

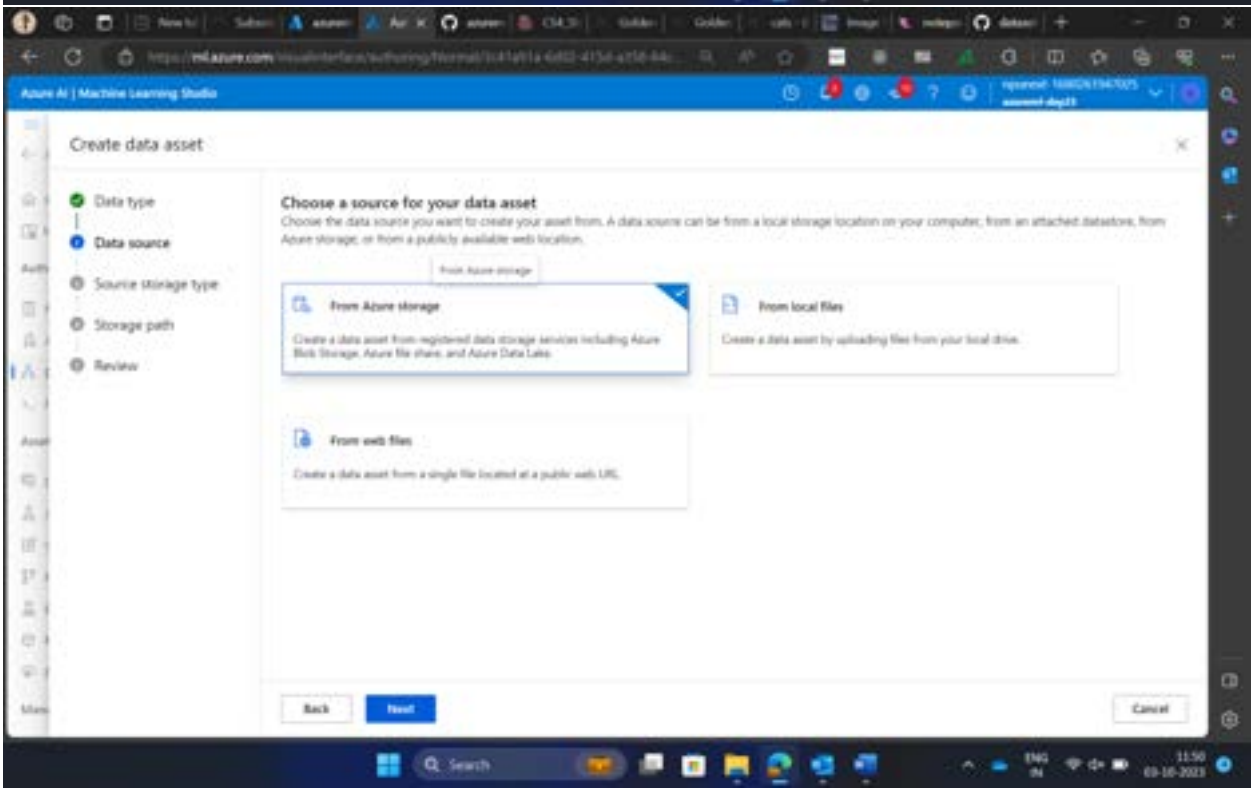

Azure AI | Machine Learning Studio
Create data asset
Data type
Data source
URI
Review
Choose a source for your data asset
Choose the data source you want to create your asset from. A data source can be from a local storage location on your computer, from an attached datastore, from Azure storage, or from a publicly available web location.
From a URI
Create a data asset by providing a web URL or a URI that represents a path to your data
From Azure storage
From local files
Create a data asset by uploading files from your local drive
Back
Next
Cancel

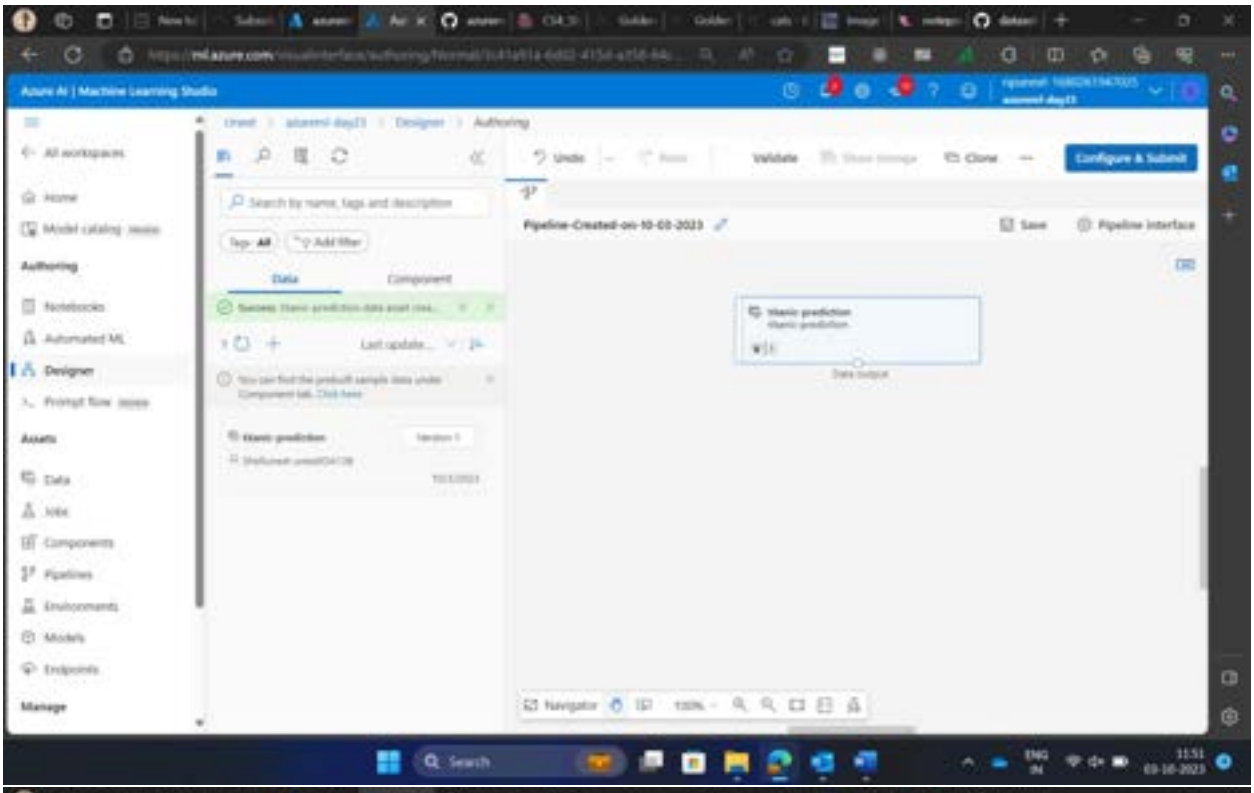

Azure AI | Machine Learning Studio
Create data asset
Data type
Data source
URI
Review
Specify a URI
Enter the URL of a public web page you want your data retrieved from or a supported URI format representing the path to your data.
URI *
https://github.com/datasciencedojo/datasets/blob/master/titanic.csv
Skip data validation
If you choose to skip validation, we will not validate your data path, or try to access your data for preview and schema.
Skip data validation
Supported URI formats
When you create a data asset in Azure Machine Learning, you'll specify a path parameter that points to the data location. These are the supported formats:
1. A path on a public http(s) server, such as:
https://raw.githubusercontent.com/pandas-dev/pandas/main/doc/data/titanic.csv
2. A path on Azure storage, such as:
3. A path on a datastore, such as:
Learn more about supported URI formats
Back
Next
Cancel

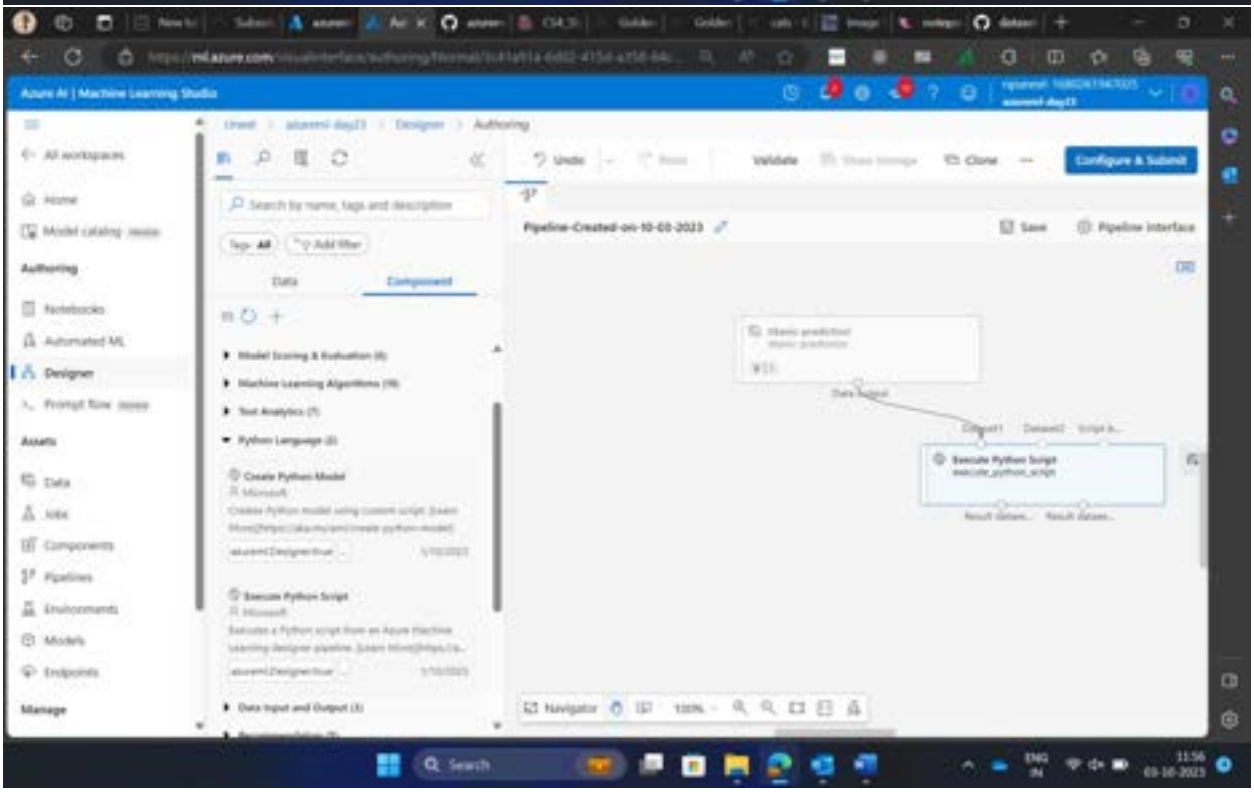

Azure AI | Machine Learning Studio
titanic-datasets-web
Details
Explore
Jobs
New version
Refresh
Archive
Share data asset
Attributes
Type
File (uri_file)
Named asset URI
Created by
Current version
1
Latest version
1
Created time
Oct 3, 2023 11:38 AM
Modified time
Oct 3, 2023 11:38 AM
Tags
Description
Data sources
URI
Location
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Manage
Authoring
Notebooks
Automated ML
Designer
Assets

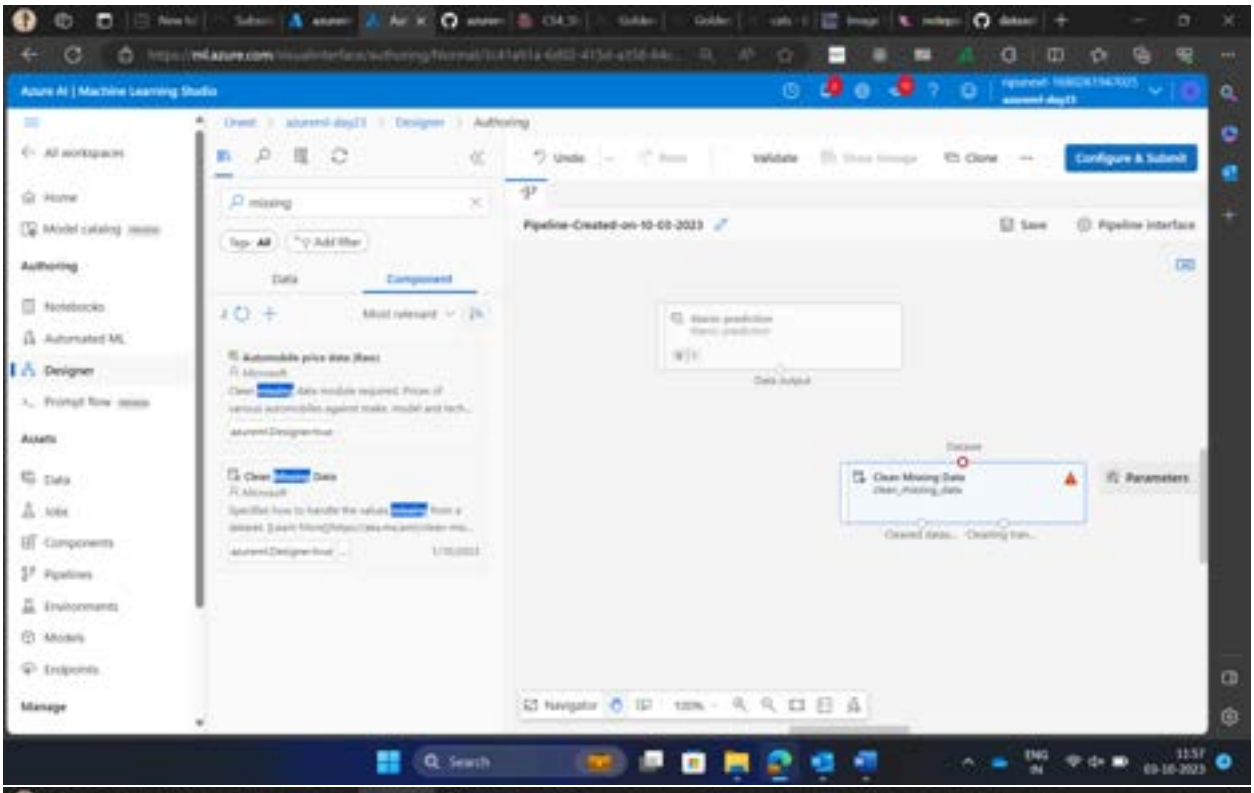

Azure AI | Machine Learning Studio
titanic-localdata
Details
Explore
Jobs
Refresh
Share data asset
Preview
Path
File Name
Modifie...
Created ...
File Size
File For...
CanSeek
Previews
Display as grid
With column headers
Passen...
Survived
Pclass
Name
Sex
Age
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Manage
Authoring
Notebooks
Automated ML
Designer
Assets

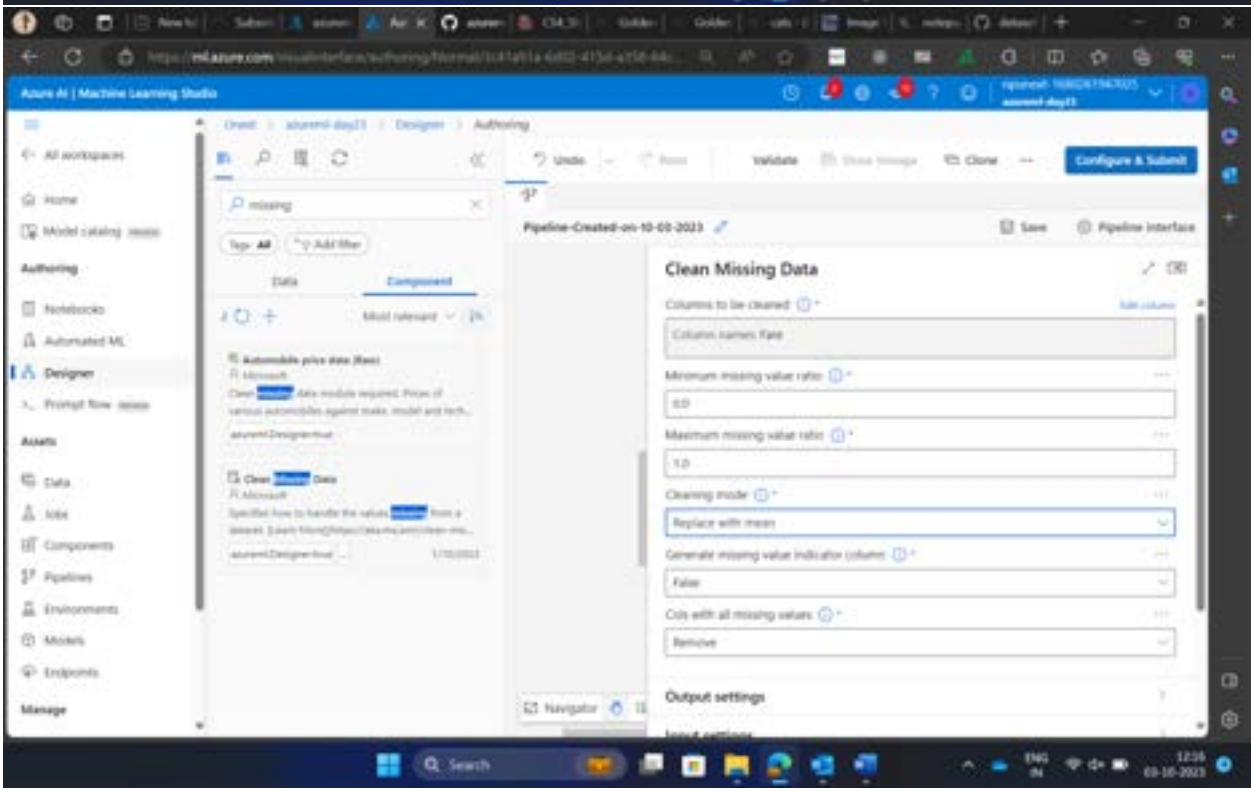

Azure AI | Machine Learning Studio
Create data asset
Data type
Data source
Set the name and type for your data asset
Name *
titanic-prediction
Description
Data asset description
Type *
File
Use cases for data types
When should I use File type?
When should I use Folder type?
When should I use Table type?
Back
Next
Cancel

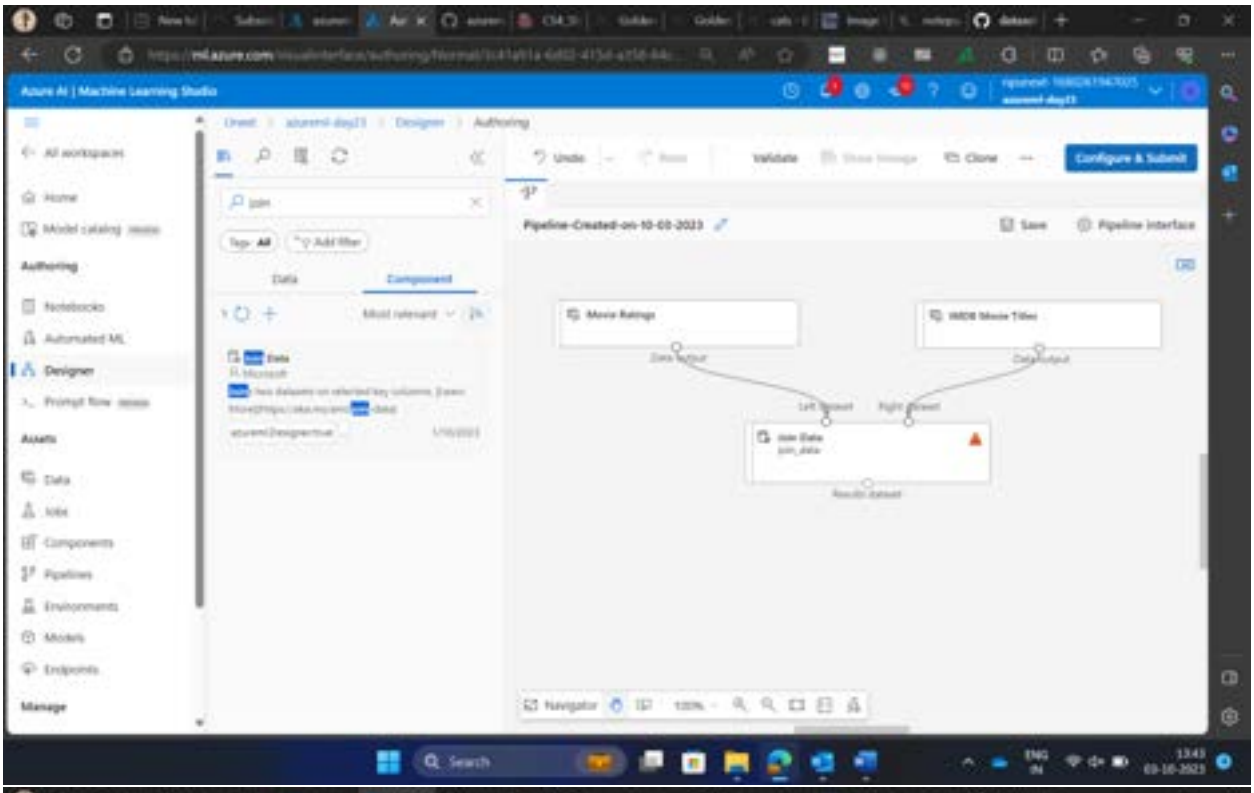

Azure AI | Machine Learning Studio
Create data asset
Data type
Data source
Source storage type
Storage path
Review
Choose a source for your data asset
Choose the data source you want to create your asset from. A data source can be from a local storage location on your computer, from an attached datastore, from Azure storage, or from a publicly available web location.
From Azure storage
Create a data asset from registered data storage services including Azure Blob Storage, Azure file share, and Azure Data Lake.
From local files
Create a data asset by uploading files from your local drive.
From web files
Create a data asset from a single file located at a public web URL.
Back
Next
Cancel

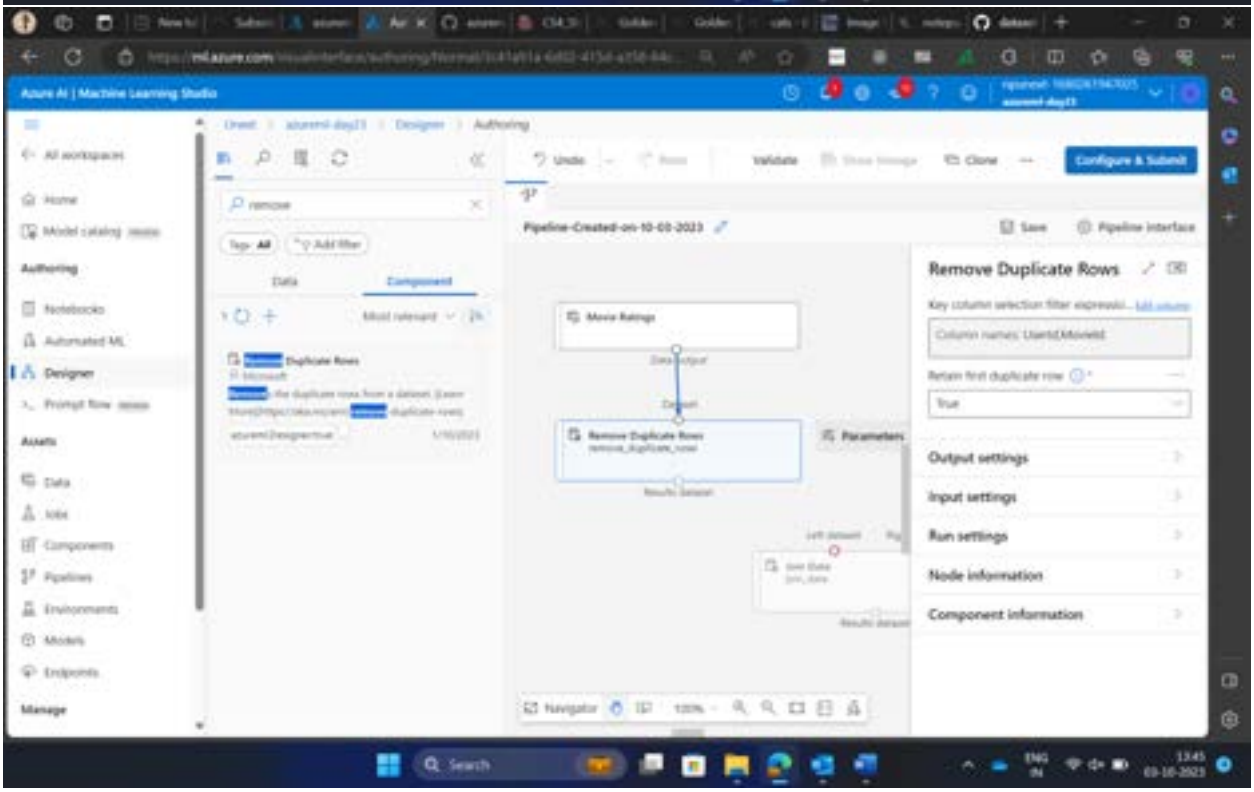

Azure AI | Machine Learning Studio
Configure & Submit
Pipeline-Created-on-10-03-2023
Save
Pipeline interface
Validate
Clone
Undo
Designer
Authoring
Data
Component

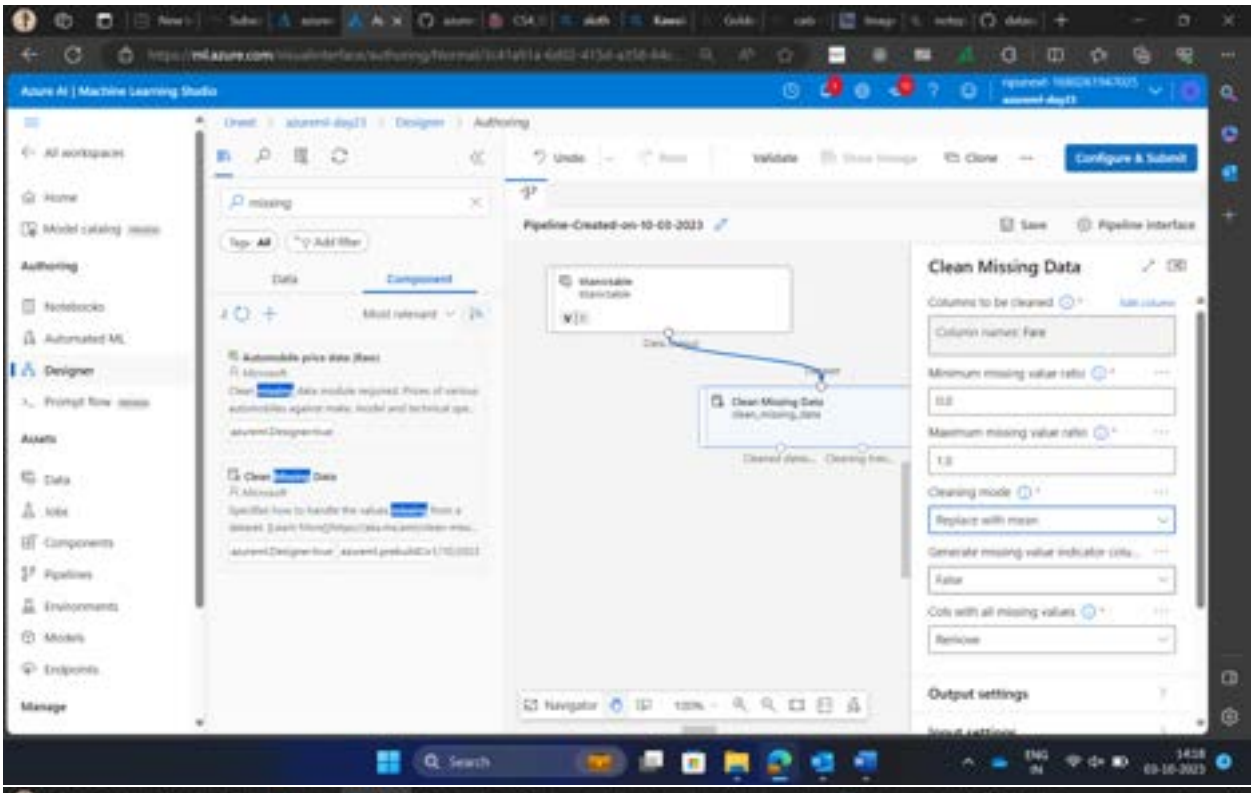

Azure AI | Machine Learning Studio
Configure & Submit
Pipeline-Created-on-10-03-2023
Save
Pipeline interface
Validate
Clone
Undo
Designer
Authoring
Data
Component
Execute Python Script
Create Python Model

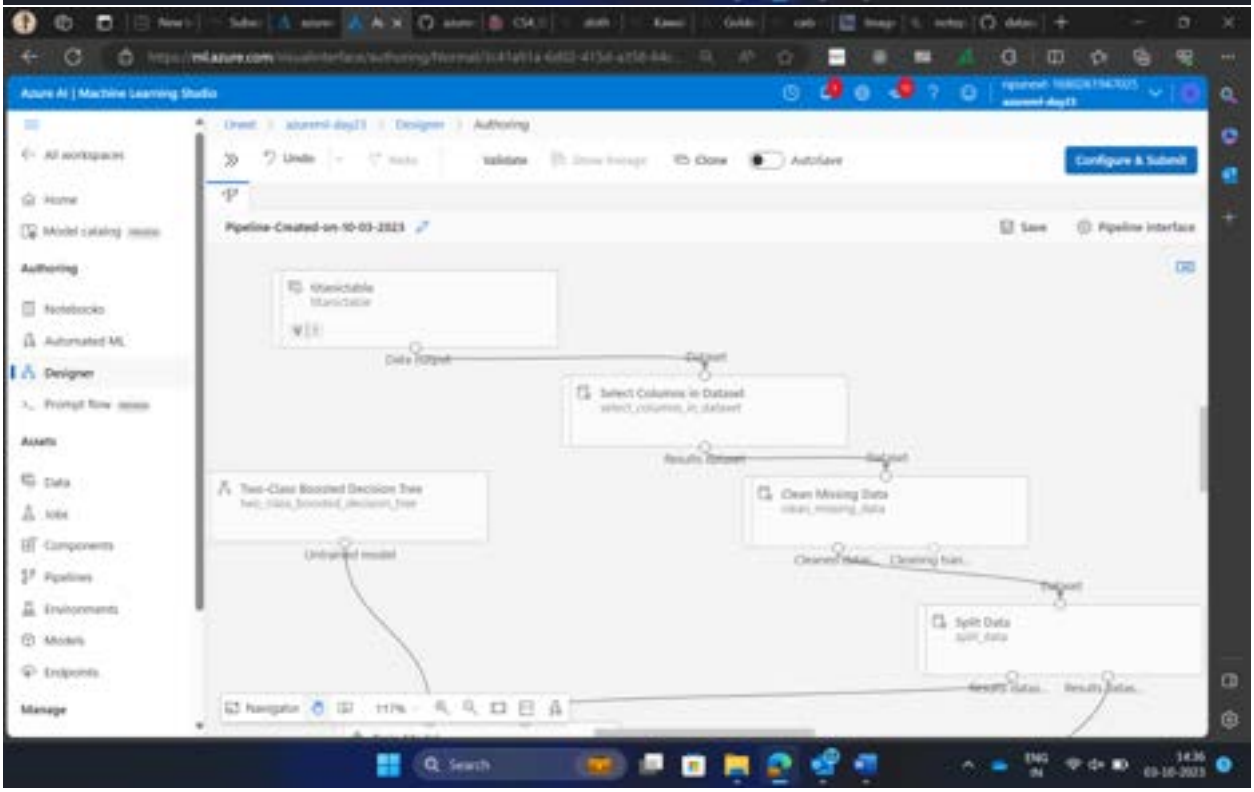

Azure AI | Machine Learning Studio
Configure & Submit
Pipeline-Created-on-10-03-2023
Clean Missing Data
Parameters

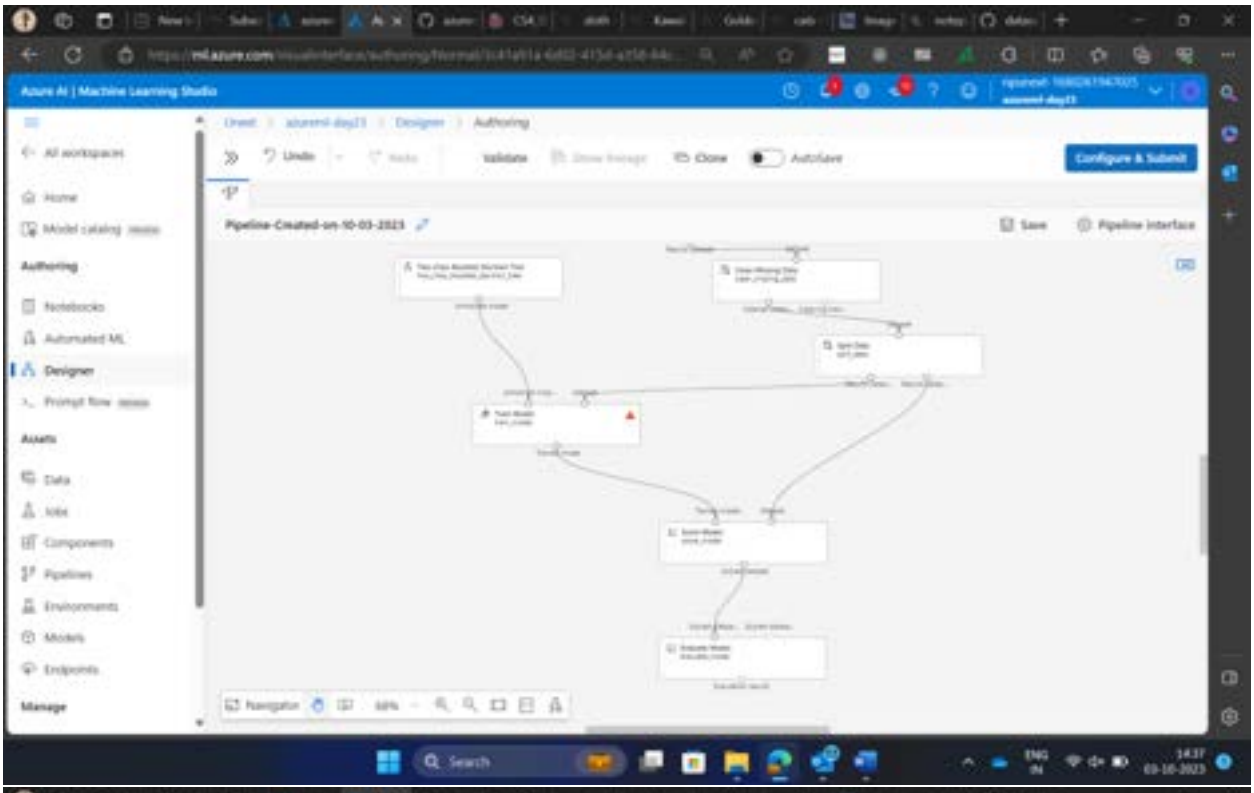

Azure AI | Machine Learning Studio
Configure & Submit
Pipeline-Created-on-10-03-2023
Clean Missing Data
Columns to be cleaned
Minimum missing value ratio
Maximum missing value ratio
Cleaning mode
Replace with mean
Generate missing value indicator column
False
Cols with all missing values
Remove
Output settings

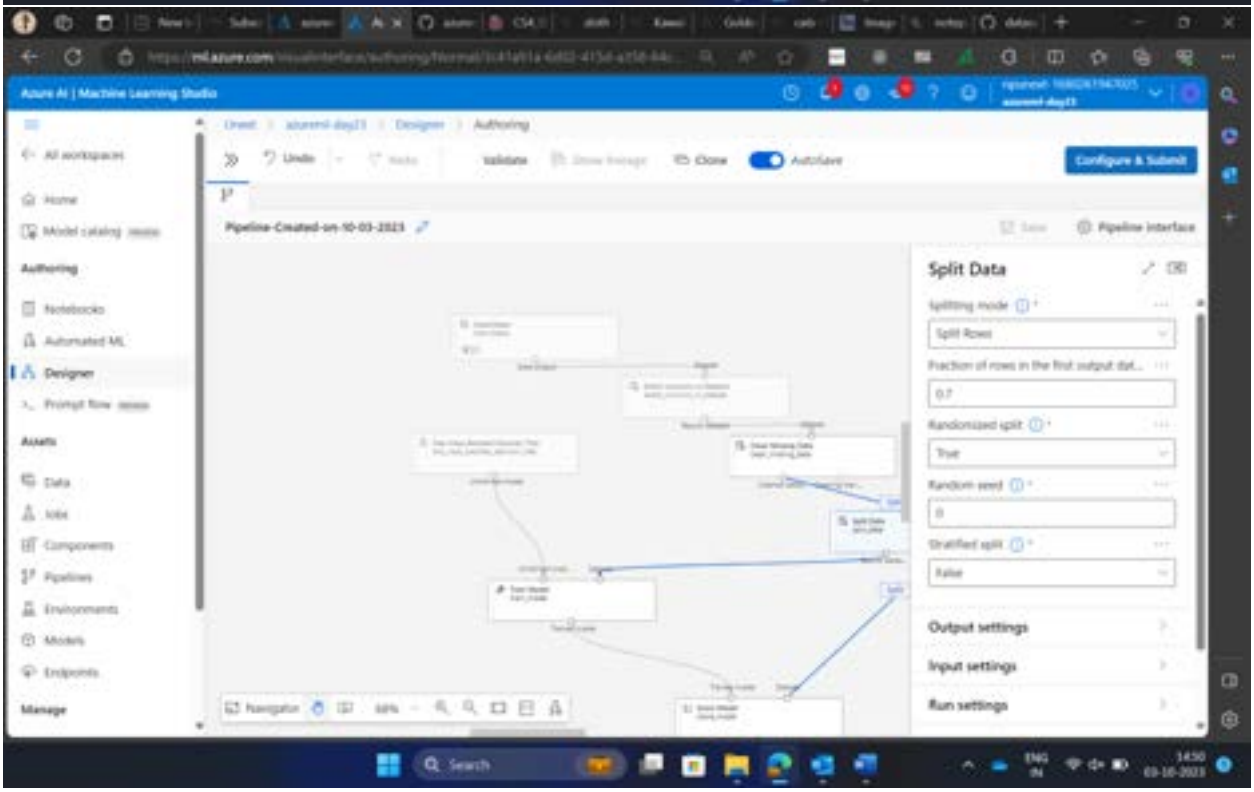

Azure AI | Machine Learning Studio
Configure & Submit
Pipeline-Created-on-10-03-2023
Movie Ratings
IMDB Movie Titles
Join Data

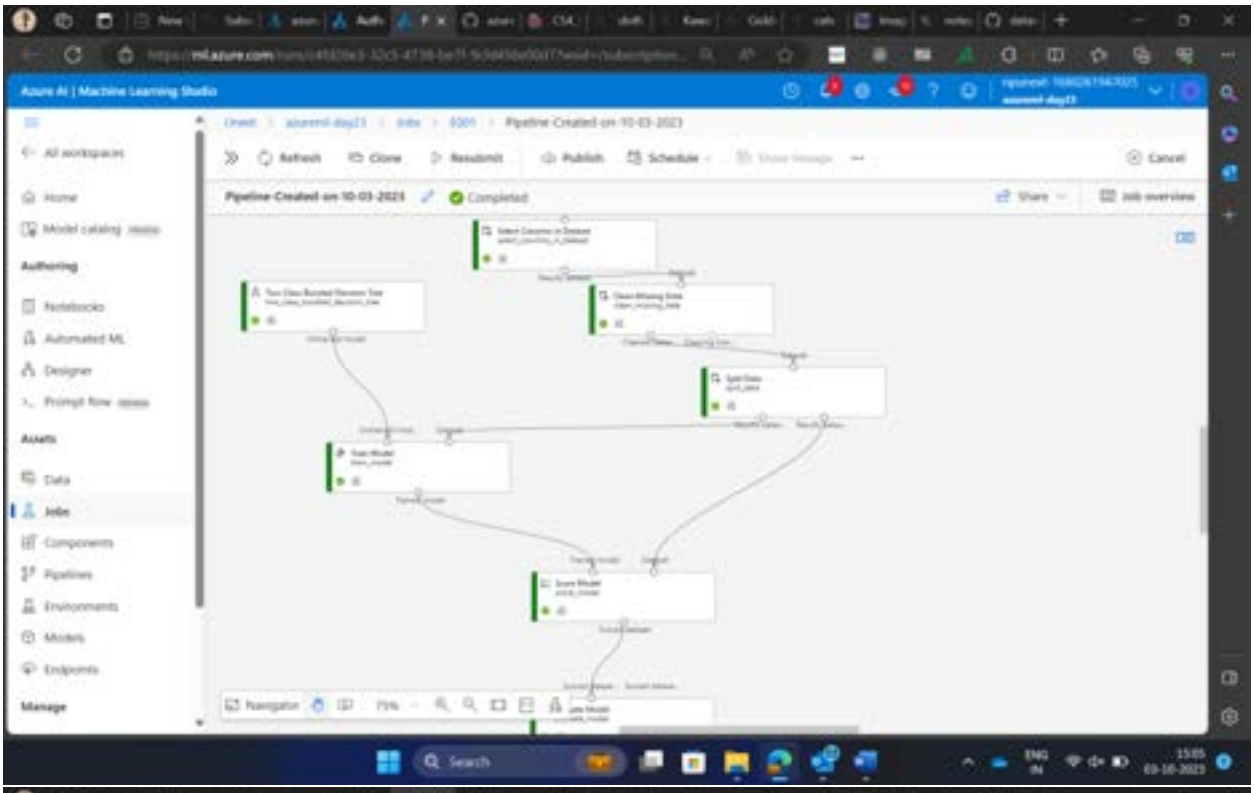

Azure AI | Machine Learning Studio
Configure & Submit
Pipeline-Created-on-10-03-2023
Movie Ratings
Remove Duplicate Rows
Remove Duplicate Rows
Output settings
Input settings
Run settings
Node information
Component information

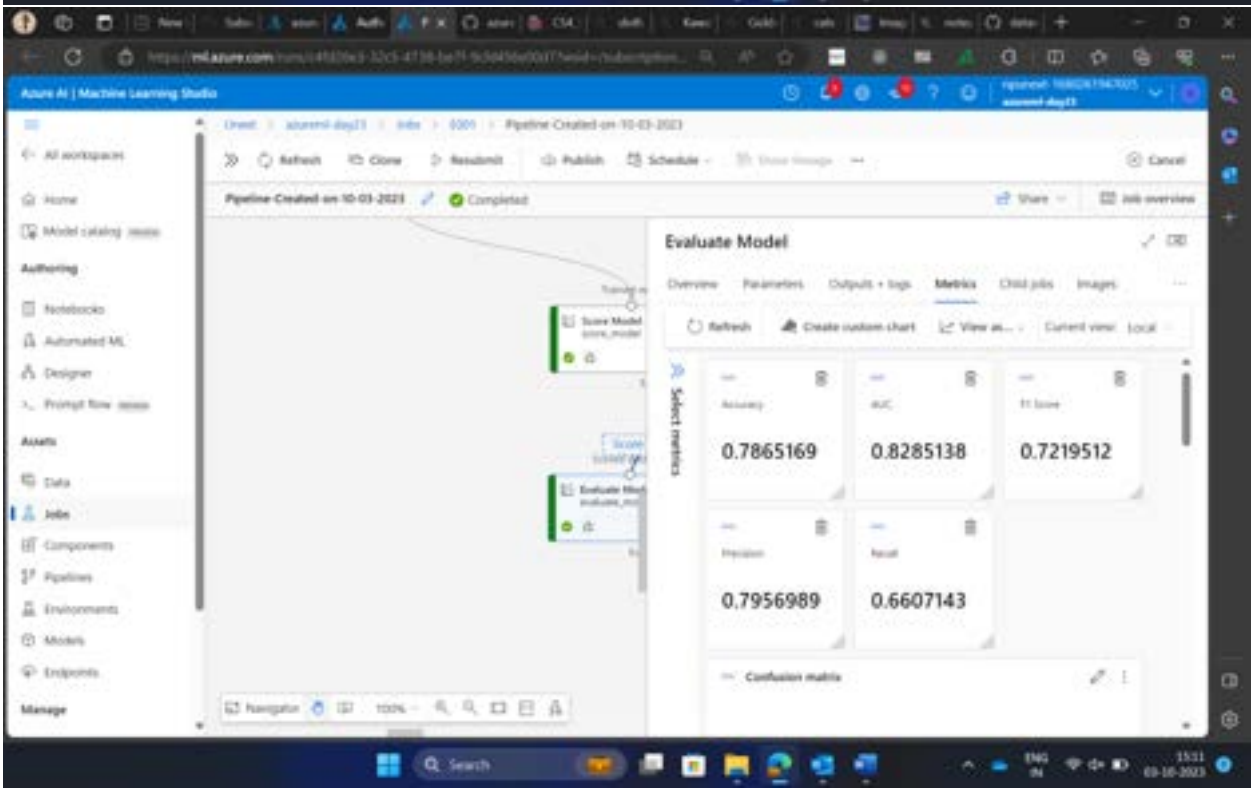

Azure AI | Machine Learning Studio
Configure & Submit
Pipeline-Created-on-10-03-2023
Clean Missing Data
Columns to be cleaned
Minimum missing value ratio
Maximum missing value ratio
Cleaning mode
Replace with mean
Generate missing value indicator colu...
False
Cols with all missing values
Remove
Output settings

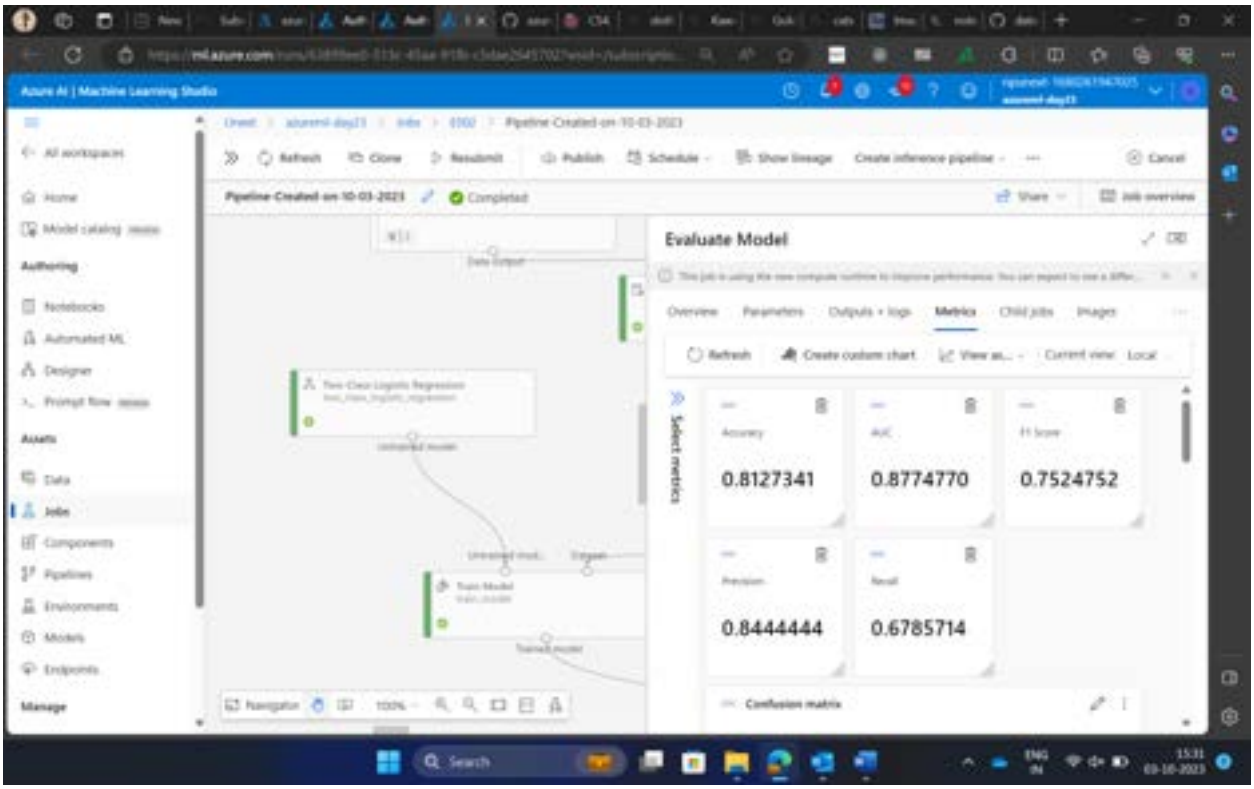

Azure AI | Machine Learning Studio
Configure & Submit
Pipeline-Created-on-10-03-2023
Select Columns in Dataset
Two-Class Boosted Decision Tree
Clean Missing Data
Split Data


Azure AI | Machine Learning Studio
Configure & Submit
Pipeline-Created-on-10-03-2023
Authoring
Designer
Assets
Manage

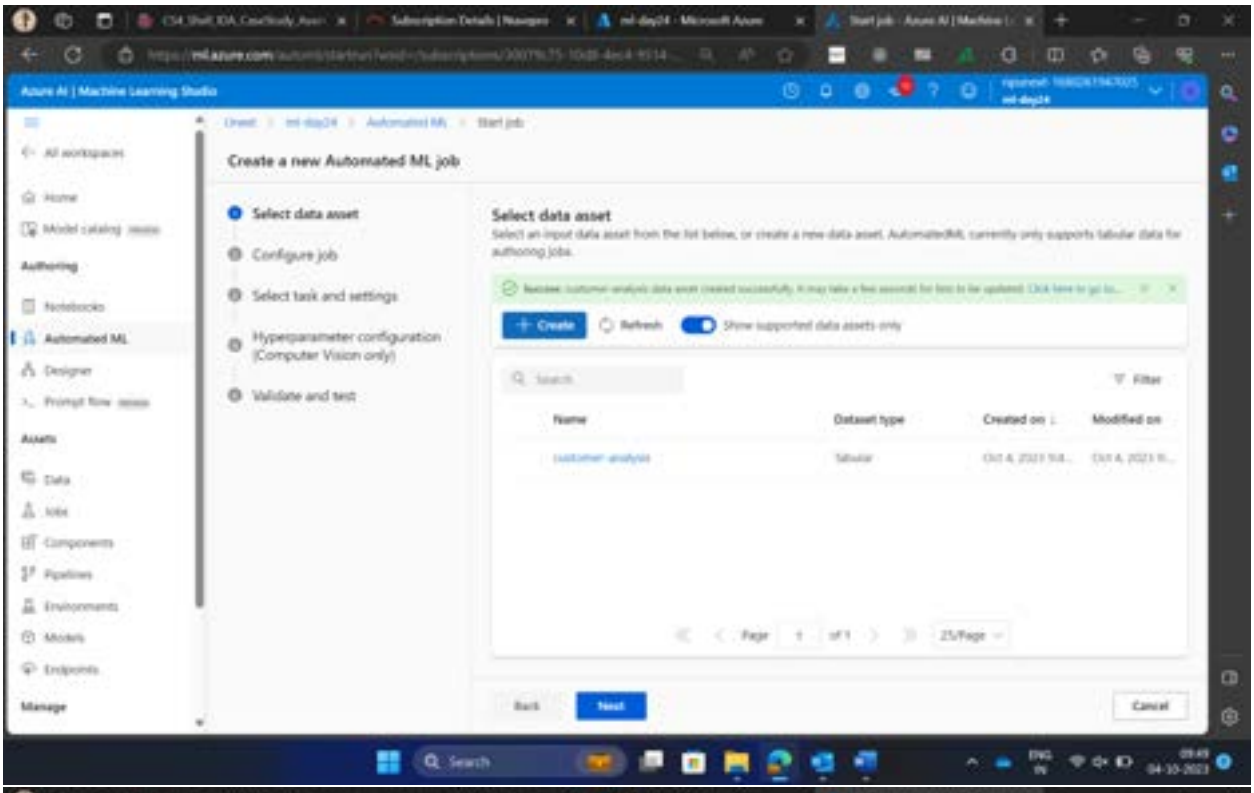

Azure AI | Machine Learning Studio
Configure & Submit
Pipeline-Created-on-10-03-2023
Split Data
Output settings
Input settings
Run settings
Authoring
Designer
Assets
Manage

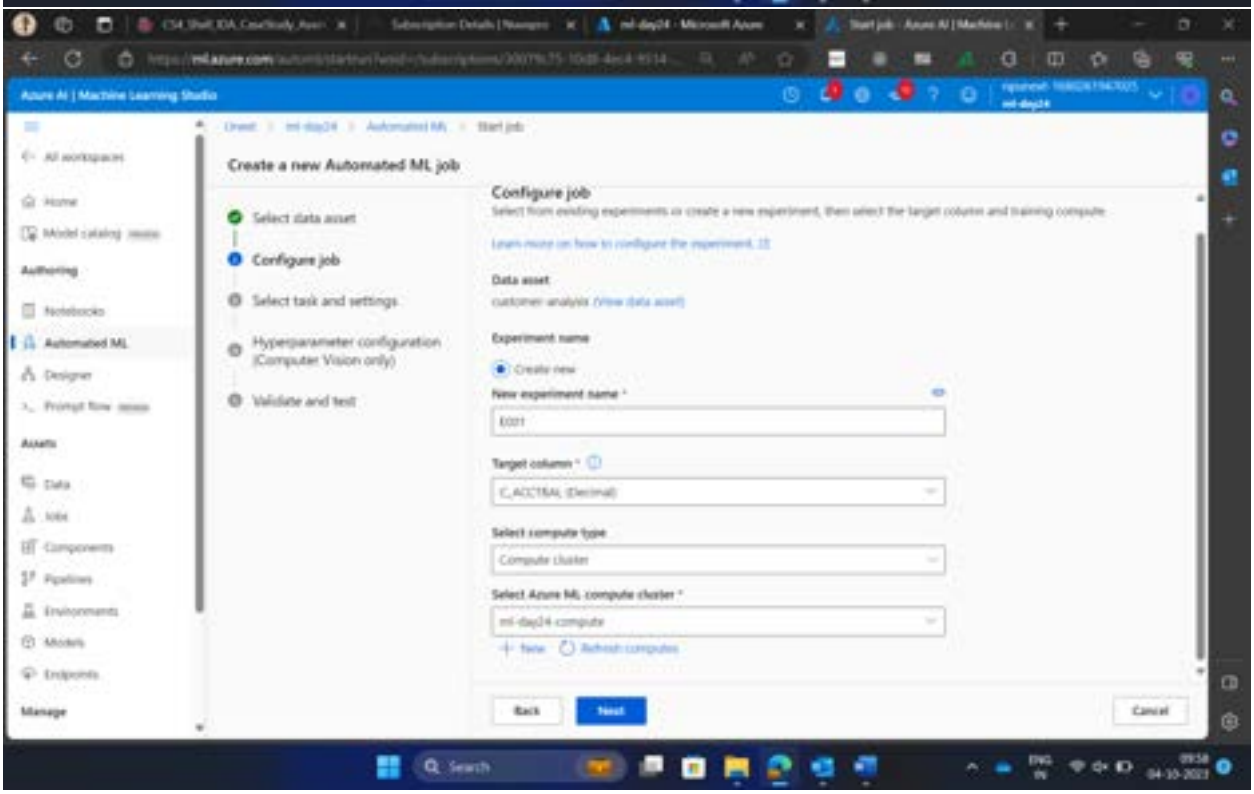

Azure AI | Machine Learning Studio
Pipeline-Created-on-10-03-2023
Completed
Refresh
Clone
Resubmit
Publish
Schedule
Cancel
Share
Job overview
All workspaces
Home
Model catalog
Authoring
Notebooks
Automated ML
Designer
Prompt flow
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Manage

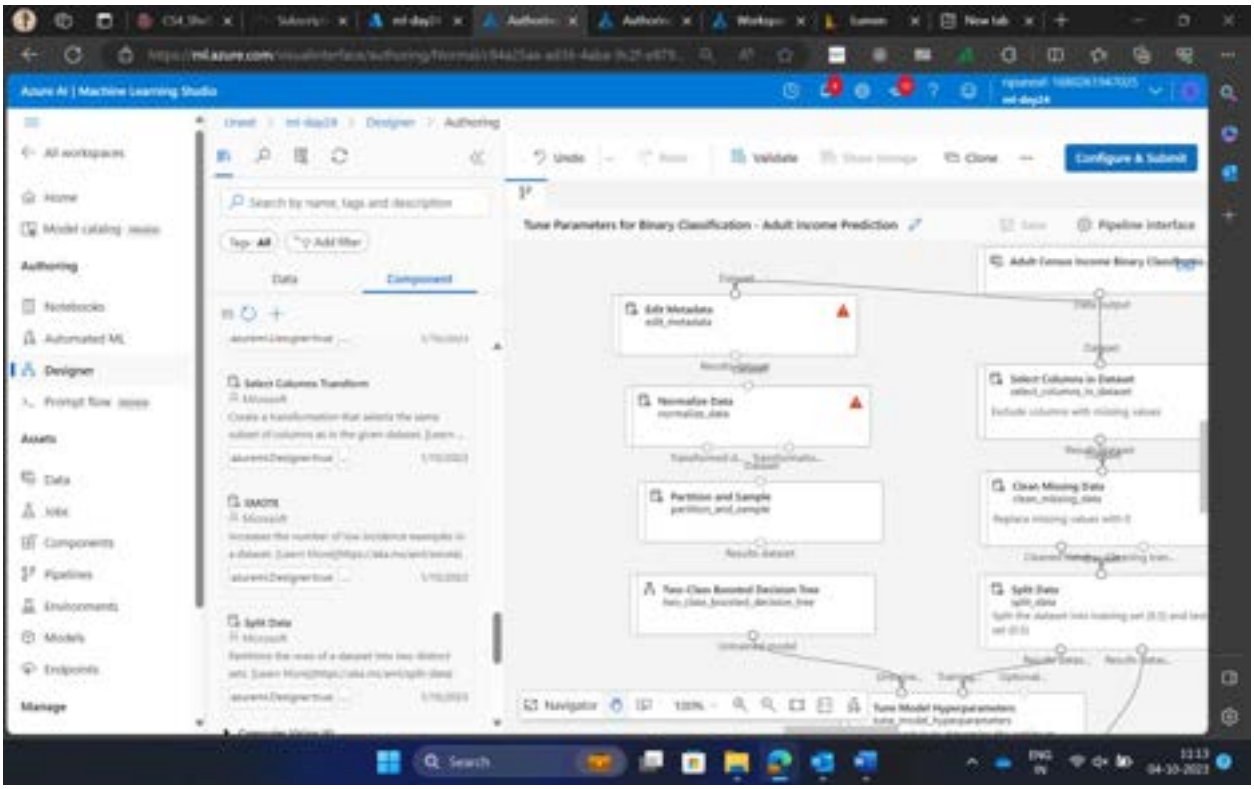

Azure AI | Machine Learning Studio
Pipeline-Created-on-10-03-2023
Completed
Evaluate Model
Overview
Parameters
Outputs + logs
Metrics
Child jobs
Images
Refresh
Create custom chart
View as...
Current view:
Local
Select metrics
Accuracy
0.7865169
AUC
0.8285138
F1 Score
0.7219512
Precision
0.7956989
Recall
0.6607143
Confusion matrix
All workspaces
Home
Model catalog
Authoring
Notebooks
Automated ML
Designer
Prompt flow
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Manage


Azure AI | Machine Learning Studio
Pipeline Created on 10-03-2023
Completed
Evaluate Model
Overview
Parameters
Outputs + logs
Metrics
Child jobs
Images
Refresh
Create custom chart
Select metrics
Accuracy
0.8127341
AUC
0.8774770
F1 Score
0.7524752
Precision
0.8444444
Recall
0.6785714
Confusion matrix

Day 24


Azure AI | Machine Learning Studio
Create a new Automated ML job
Select data asset
Configure job
Select task and settings
Hyperparameter configuration (Computer Vision only)
Validate and test
Select data asset
Select an input data asset from the list below, or create a new data asset. AutomatedML currently only supports tabular data for authoring jobs.
+ Create
Refresh
Show supported data assets only
Name
Dataset type
Created on
Modified on
customer_analysis
Tabular
Back
Next
Cancel


Azure AI | Machine Learning Studio
Create a new Automated ML job
Select data asset
Configure job
Select task and settings
Hyperparameter configuration (Computer Vision only)
Validate and test
Configure job
Select from existing experiments or create a new experiment, then select the target column and training compute.
Data asset
customer_analysis (View data asset)
Experiment name
Create new
New experiment name *
Target column *
Select compute type
Compute cluster
Select Azure ML compute cluster *
+ New
Refresh computes
Back
Next
Cancel


Azure AI | Machine Learning Studio
All workspaces
Home
Model catalog
Authoring
Notebooks
Automated ML
Designer
Prompt flow
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Manage
Undo
Validate
Clone
Configure & Submit
Tune Parameters for Binary Classification - Adult Income Prediction
Pipeline interface
Edit Metadata
Normalize Data
Partition and Sample
Two-Class Boosted Decision Tree
Select Columns in Dataset
Clean Missing Data
Split Data
Tune Model Hyperparameters
Navigator