

Microsoft Azure | Overview

Your deployment is complete

Deployment name: annika\_1693809509828 Start time: 9/4/2023, 12:08:41 PM  
Subscription: mytenant-1683261916 Correlation ID: c46f3118-0c11-475c-847e-64740ffec2  
Resource group: annika

Deployment details

Next steps

Get to resource

Give feedback

Help us about your experience with deployment

Code 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 | Create an alert rule - Microsoft

Create an alert rule

Signal name: Ingress

Alert logic

Threshold: Static

We currently support alert rules with Dynamic Threshold condition as a single criteria

Aggregation type: Total

Operator: Greater than

Unit: %

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

The screenshot shows two Microsoft Azure browser windows side-by-side.

**Top Window: Create an alert rule**

This window is titled "Create an alert rule". It displays a warning message: "The following fields require attention: Alert rule name".

**Project details:**  
Subscription: annika (selected)  
Resource group: annika (selected)

**Alert rule details:**  
Severity: Informational  
Alert rule name: alertrule1  
Alert rule description: This is alert 1

**Bottom Window: annika | Metrics**

This window shows the Metrics blade for the "annika" storage account. The left sidebar includes "Resource sharing (CORE)", "Advisor recommendations", "Endpoints", and "Logs". Under "Monitoring", "Metrics" is selected, showing a chart for "Sum Ingress and Sum Egress for annika". The chart displays two data series: "Ingress (Sum)" and "Egress (Sum)". The "Ingress (Sum)" value is 1.1 GB and the "Egress (Sum)" value is 1.7 GB. The chart has a timestamp range from 2023-07-20T00:00 to 2023-07-20T15:00 UTC+01:00.

A success message is displayed in the top right corner: "Alert rule created. Alert rule alertrule1 successfully created. It might take a few minutes for changes to be shown." The URL for this page is https://portal.azure.com/#resource/resourceGroups/annika/providers/Microsoft.Storage/storageAccounts/annika/metricsMetrics.

Alert rules - Microsoft Azure

Microsoft Azure | Search resources, services, and docs (S+)

Home > amika;1613809109818 | Overview > amika | Alerts >

### Alert rules

+ Create Columns Refresh Export to CSV Open query ⌂ Details ⌂ Details ⌂ Details

Search Subscription: spawat-1486261916031 ⌂ More ⌂ No grouping

Name	Condition	Severity	Target scope	Target resource type	Signal type	Status
alarm1	Ingress > 10 and T more ...	Informational	amika	Storage account	Metric	Enabled

Showing 1 - 1 of 1 results.

containerlab1 - Microsoft Azure

https://portal.azure.com/#view/Microsoft\_Azure\_Storage\_ContainerMetricsBlade/Overview?SubscriptionId=spawat-1486261916031&ResourceGroup=amika&ContainerName=containerlab1&\_\_v=1

Microsoft Azure | Search resources, services, and docs (S+)

Home > amika;1613809109818 | Overview > amika | Container >

### containerlab1

Container

Upload Change access level Refresh Delete Change key Acquire key Create lease View snapshots

Successfully uploaded blob(s)  
Successfully uploaded 1 blob(s).

Authentication method: Access key (Switch to Azure AD User Account)

Location: containerlab1

Search blobs by prefix (case-sensitive): Show deleted blobs

Add filter

Name	Modified	Access tier	Archive status	Block type
agile-methodology_695x280.webp	9/4/2023, 12:18:26 PM	Hot (oldest)		Block blob
waterfall-method.webp	9/4/2023, 12:18:19 PM	Hot (oldest)		Block blob

Microsoft Azure (Search resources, services, and more) ShellUser\_1883422072\_1960

## alertrule1

Metric alert rule

Overview

Activity log

Access control (IAM)

Tags

Settings

Links

Automation

Tasks (preview)

Export template

New Support Request

Resource group (root) : annika

Location (Region) : Global

Subscription (root) : npunet\_INDIAFYME2

Subscription ID : 2e24ff00-8ab6-4bb6-9166-8738f72d79dd

Severity : 1 - Informational

Description : This is alert 1

Tag Info Add tags

Scope

Resource : annika

Hierarchy : npunet\_H883261916031 > annika

Conditions

Name	Time series monitored	Estimated monthly cost
alartrule1	1	0.00

JSON View

This screenshot shows the 'alertrule1' metric alert rule configuration in the Microsoft Azure portal. The alert is set to trigger at level 1 (Informational) and has a single condition monitoring the 'alartrule1' time series. The alert was created on 5/4/2023 at 12:28 PM and is currently in a 'Reed' state. The affected resource is 'annika'. The alert is triggered by a single data point shown in a line chart.

Microsoft Azure (Search resources, services, and more) ShellUser\_1883422072\_1960

## annika | Alerts

Storage account

Endpoints

Logs

Monitoring

Insights

Alerts

Metrics

Workbooks

Diagnostic settings

Logs

Monitoring (classic)

Metrics (classic)

Diagnostic settings (classic)

Usage Metrics

Automation

Tasks (preview)

Cloud Shell

Modify results

Search

alartrule1

Total alerts 1

Name : alertrule1

Severity : 1 - Informational

Fixed time : 5/4/2023, 12:28 PM

Affected resource : annika

Hierarchy : npunet\_H883261916031 > annika

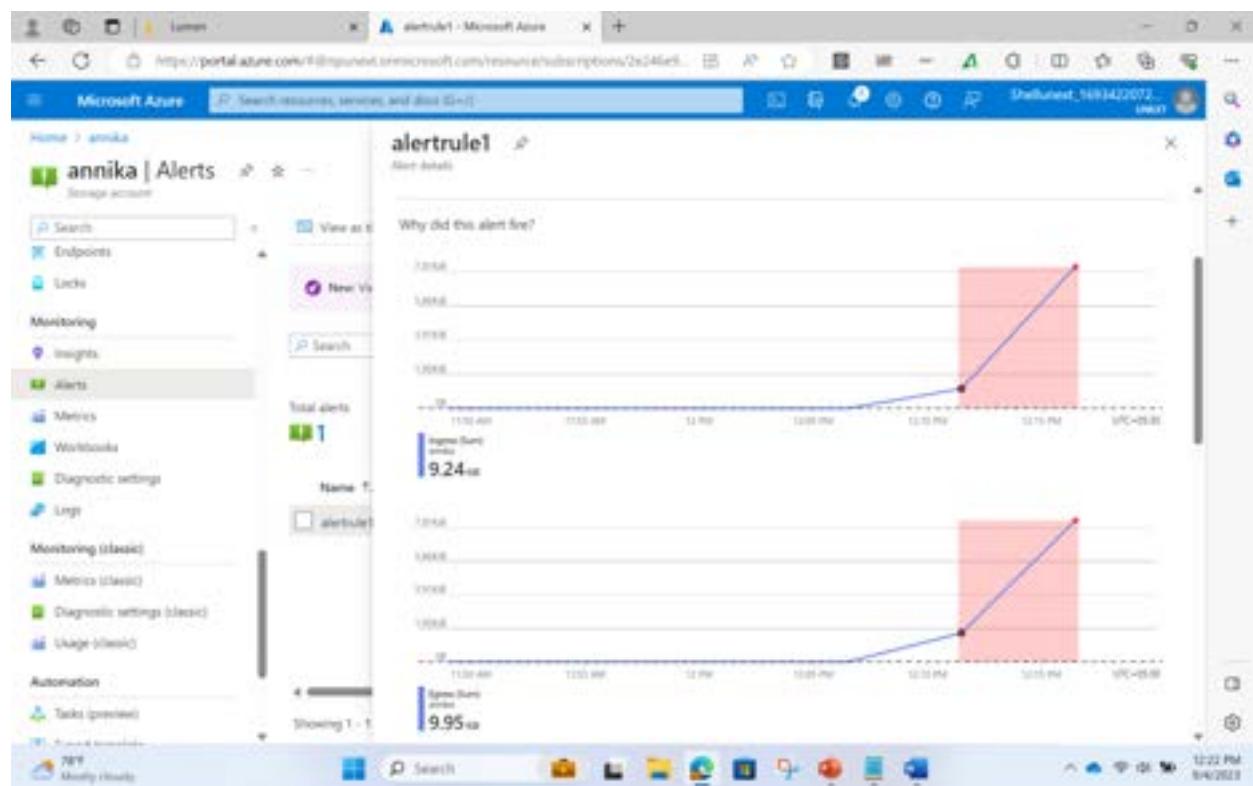
User response : New

Alert condition : Reed

Change user response

Why did this alert fire?

This screenshot shows the details of the 'alertrule1' alert triggered on the 'annika' storage account. The alert was triggered at 12:28 PM on 5/4/2023. The alert condition is 'Reed' and the user response is 'New'. The alert is triggered by a single data point shown in a line chart. The chart shows a value of 9.24 at 11:00 AM, which then rises sharply to approximately 10,000 by 12:28 PM.



The screenshot shows two identical views of the Microsoft Azure Storage account settings for the storage account 'annikastorage1'. Both views are titled 'annikastorage1 | Static website' and are under the 'Storage accounts' section.

The left sidebar lists several options: Access keys, Shared access signature, Encryption, Microsoft Defender for Cloud, Data management (with sub-options: Redundancy, Data protection, Object replication, Blob inventory), Static website (which is selected and highlighted in grey), Lifecycle management, and Azure search.

The main pane displays the 'Static website' configuration. It includes a note about hosting static websites on the blob service, mentioning that server-side scripting is not supported. It also notes that data is replicated asynchronously from primary to secondary regions, so files at the secondary endpoint may not be immediately available or in sync with files at the primary endpoint. A link to learn more is provided.

The 'Static website' section has a 'Disabled' button (which is blue) and an 'Enabled' button (which is purple). Below these buttons are two input fields: 'Index document name' set to 'index.html' and 'Error document path' set to 'error.html'.

At the bottom of the main pane, there is a note stating: 'An Azure Storage container has been created to host your static website.' Below this note is a 'Search' bar containing the URL 'https://annikastorage1.z13.web.core.windows.net/'.

The status bar at the bottom of the browser window shows the date '04-09-2023' and time '13:53'.

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

The screenshot shows the Microsoft Azure Storage Container Overview page for a container named '\$web'. The left sidebar includes links for Overview, Diagnosis and solve problems, Access Control (IAM), Settings (Shared access tokens, Access policy, Properties, Metadata, Editor (preview)), and Container (Search, Upload, Change access level, Refresh, Delete, Change key, Acquire key, Create lease, View properties). The main area displays a table of files:

Name	Modified	Access tier	Archive status	Blob type
script.html	9/4/2023, 1:55:26 PM	Hot (Inferred)		Block blob
index.html	9/4/2023, 1:55:24 PM	Hot (Inferred)		Block blob

A success message at the top right says 'Successfully uploaded 1 file'.

The screenshot shows a web browser displaying a website at <https://amikastorage123.web.core.windows.net>. The page title is 'Welcome to ABC Corporates' and the subtext is 'We are at your service'.

**Deployment**

**Overview**

Your deployment is complete

Deployment name: azuredatalakelab3\_1... Start time: 9/6/2023 2:03:21 PM  
Subscription: hyunjeon\_1000361947023 Correlation ID: 17ed823b-4657-4124-a1eb-2e028  
Resource group: arminka

Deployment details

Next steps

Get to resource

Give feedback

Help 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.

**azuredatalakelab3**

Storage account

Upload Open in Explorer Delete Move Refresh Open in mobile CPU / PS Feedback

Resource group (read-only): arminka Location: East US Primary/Secondary Location: Primary: East US, Secondary: West US Subscription (read-only): hyunjeon\_1000361947023 Subscriptions ID: 30079c75-10d8-4e44-9514-790a2ab8f8dc Disk state: Primary Available, Secondary Available

Tags (edit) Add tags

Properties Monitoring Capabilities (0) Recommendations (0) Tutorials: Tools + SDKs

Data Lake Storage

Setting	Value
Hierarchical namespace	Enabled
Default access tier	Hot
Blob anonymous access	Disabled
Blob soft delete	Enabled (7 days)
Container soft delete	Disabled
Versioning	Disabled

Security

Setting	Value
Require secure transfer for REST API operations	Enabled
Storage account key access	Enabled
Minimum TLS version	Version 1.2
Infrastructure encryption	Disabled

Networking

14:04 04-09-2023

The screenshot shows two views of the Azure Cosmos DB Data Explorer interface.

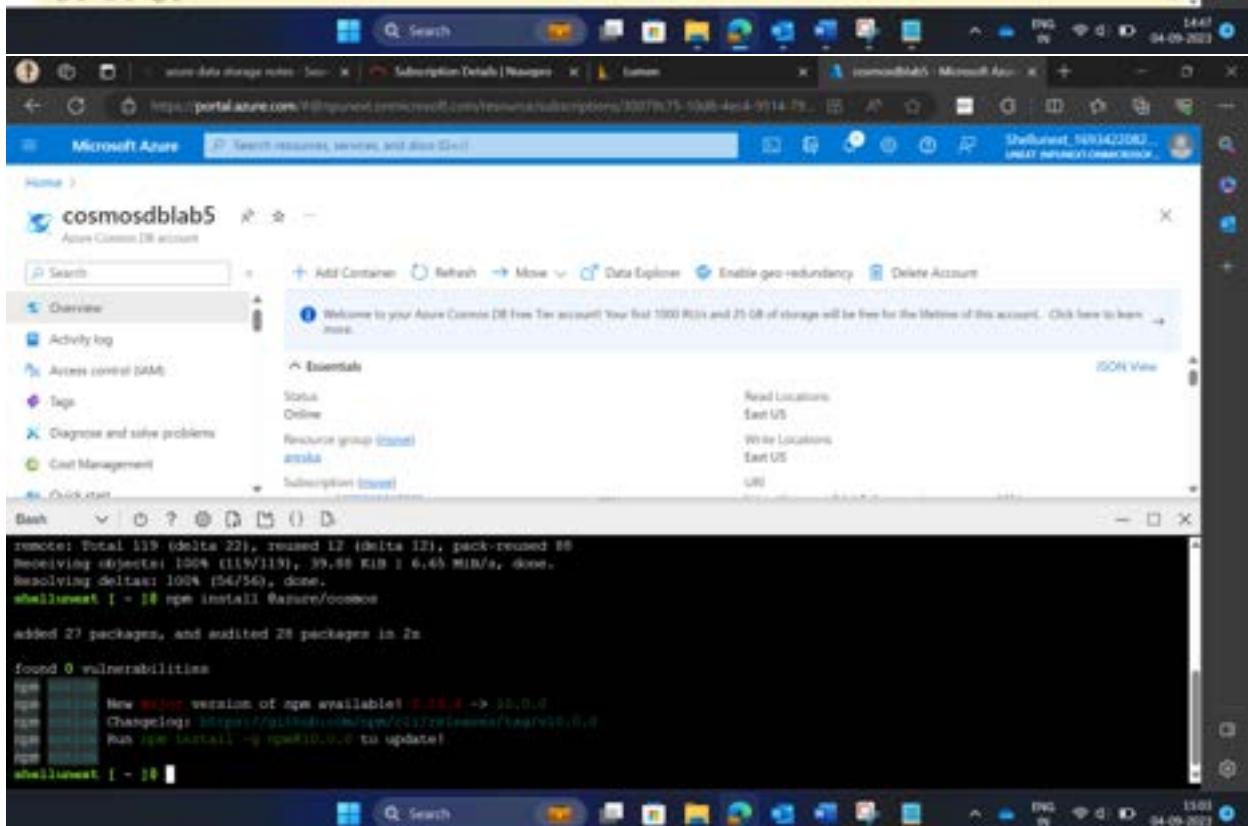
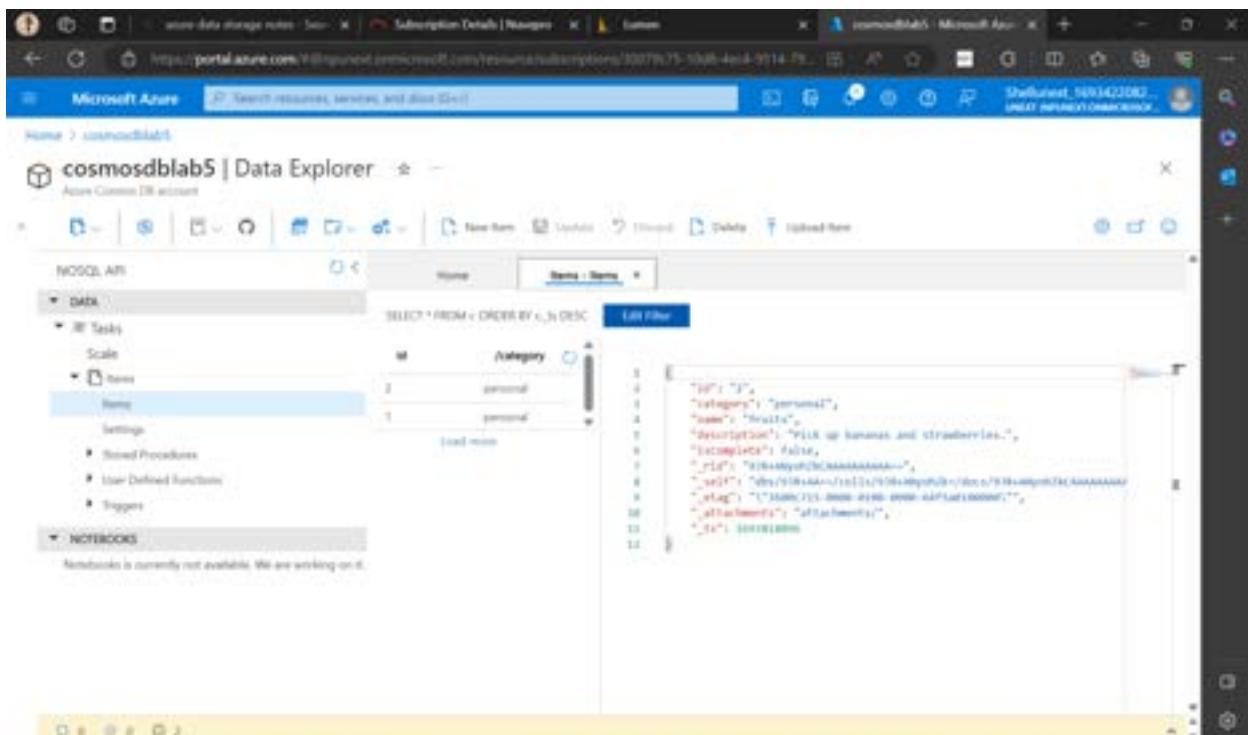
**Top View:** The main area displays a "Welcome to Azure Cosmos DB" message: "Globally distributed, multi-model database service for any scale". Below this are three cards: "Launch quick start", "New Container", and "Connect". The "New Container" card is highlighted with a blue circle. The left sidebar shows the "Data Explorer" tab is selected, along with other options like Overview, Activity log, and Data.

**Bottom View:** This view shows the Data Explorer interface with a query editor. The query is:

```
SELECT * FROM c
```

The results pane displays a single document:

```
[{"id": "1", "category": "personal", "name": "groceries", "description": "Pick up apples and strawberries.", "Incomplete": false}
```



[Subscription Details | Overview](https://portal.azure.com/#blade/HubsBlade/resourceType=Microsoft.DocumentDB%2faccounts/resourceId=/subscriptions/00079b75-10d8-4e14-9914-79...)

<https://cosmosdblab5.documents.azure.com:443/>

## cosmosdblab5 | Keys

Read-write Keys    Read-only Keys

PRIMARY KEY

SECONDARY KEY

PRIMARY CONNECTION STRING

SECONDARY CONNECTION STRING

## cosmosdblab5 | Metrics

Avg Server Side Latency for cosmosdblab5

Local Time 9/4 3:18 AM - 9/5 3:18 AM (Automated)

cosmosdblab5.Server Side Latency.Avg

1.27 ms

DAY 7

[Subscription Details \(Overview\)](https://portal.azure.com/#resource/Microsoft_SQLDatabase) [Configure Microsoft Azure](https://portal.azure.com/#blade/Configure_Microsoft_Azure/Configure_Microsoft_SQLDatabase)

Microsoft Azure [Search resources, services, and docs \(Ctrl+F\)](#) [Feedback](#) [Logout](#)

Home > Select SQL deployment option >

## 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: [Compare DTU options](#)

**\$ (Basic)**

Data max size (GB):  [Z](#)

[Apply](#)

[Subscription Details \(Overview\)](https://portal.azure.com/#resource/Microsoft_SQLDatabase) [Configure Microsoft Azure](https://portal.azure.com/#blade/Configure_Microsoft_SQLDatabase)

Microsoft Azure [Search resources, services, and docs \(Ctrl+F\)](#) [Feedback](#) [Logout](#)

Home > Select SQL deployment option >

## Create SQL Database

[Basic](#) [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\_CI\_AS. [Learn more](#)

Collation:

[Review + create](#) [< Previous](#) [Next: Tags >](#)

The screenshot displays two separate Azure resource creation pages side-by-side.

**Create SQL Database** (Top Window):

- Resource Group:** annikaday7db
- Database name:** annikaday7db
- Server:** (new) annikadatabox (East US)
- Want to use SQL elastic pool?**: No
- Workload environment:** Development
- Compute + storage:** Basic (2 GB storage)

**Create Data Factory** (Bottom Window):

- Subscription:** 1480261947025
- Resource group:** annikaday7
- Instance details:**
  - Name:** annikadaf
  - Region:** East US
  - Version:** V2

A deployment status message is visible on the right side of the Data Factory window:  
Deployment succeeded.  
Deployment: Microsoft SQL Database NewDatabase NewService (V2).  
In resource group 'annikaday7' was successful.

<https://portal.azure.com/#create/microsoft.rdbaservicestore/vmwithsqlservergen1>

Microsoft Azure  Subscription Details (Manage) Create a virtual machine - More +

Create a virtual machine

SQL Server settings

Basics Disks Networking Management Monitoring Advanced SQL Server settings Tags Review + create

**Security & Networking**

SQL connectivity: Public (Internet)

Port: 5433

**SQL Authentication**

SQL Authentication: Enable

Login name: annikavmsql7

Password: [REDACTED]

Azure Key Vault integration: Disable

**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

<https://portal.azure.com/#create/microsoft.rdbaservicestore/vmwithsqlservergen1>

Microsoft Azure  Subscription Details (Manage) Create a virtual machine - More +

Create a virtual machine

Monitoring

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 | [awkaad](#) | Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand-new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Preview experience  Off

Home Author Monitor Manage Learning Center

Data Factory Validate all [Previous](#)

Integration runtimes

The Integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environment. Learn more [\[?\]](#)

+ New Refresh

Filter by name

Showing 1 - 1 of 1 items:

Name	Type	Sub-type	Status	Related	Region	Version
AutoResolveIntegrationR...	Azure	Public	Running	0	Auto Resolve	—

Search

ENGLISH (IN) 06-09-2023

Microsoft Azure | Data Factory | [awkaad](#) | Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand-new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Integration runtime setup

Integration Runtime is the native compute used to execute or dispatch activities. Choose what integration runtime to create based on required capabilities. Learn more [\[?\]](#)

+ New Refresh

Filter by name

Showing 1 - 1 of 1 items

Name
AutoResolveIntegrat...

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

Search

ENGLISH (IN) 06-09-2023

Microsoft Azure | Data Factory | [anikash](#) | Search factory and documentation

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 skip 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.

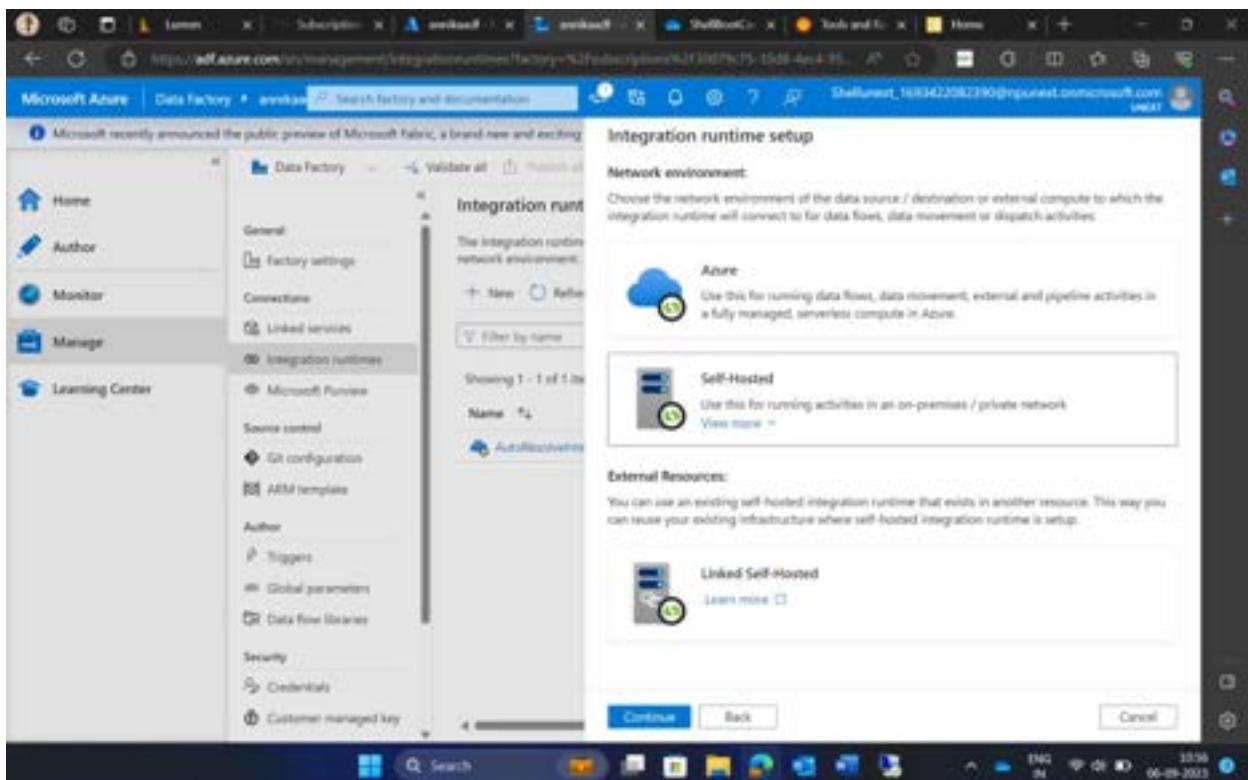
External Resources:

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 set up.

Linked Self-Hosted

Learn more ▾

Continue Back Cancel



Microsoft Azure | Data Factory | [anikash](#) | Search factory and documentation

Integration runtime setup

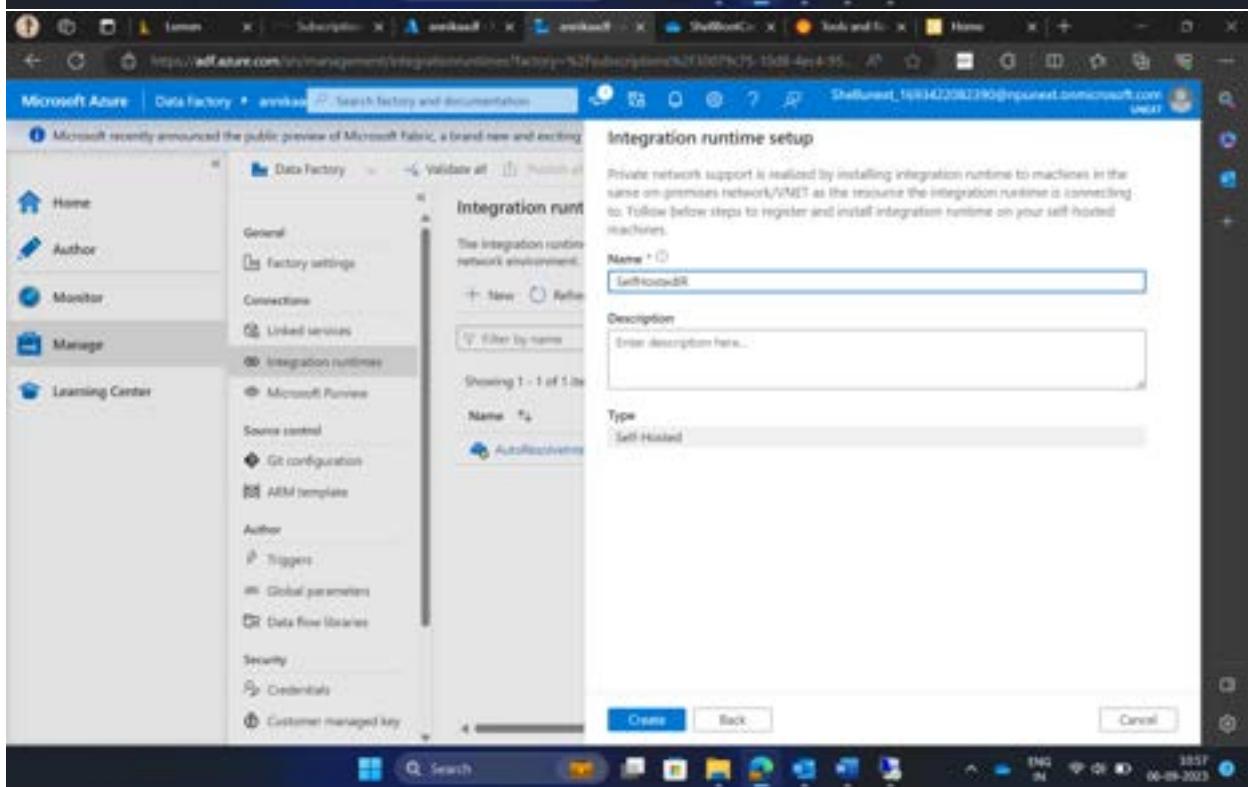
Private network support is realized by installing integration runtime to machines in the same on-premises Network/VMN as the resource the integration runtime is connecting to. Follow below steps to register and install integration runtime on your self-hosted machines.

Name

Description

Type

Create Back Cancel



The screenshot shows the Microsoft Edge browser with the URL <https://adf.azure.com/#/management/integrationruntimes>. The page title is "Welcome to Microsoft Edge". The main content area displays the "Integration runtimes" blade for a Data Factory named "amikadff". The blade includes a summary section, a table listing two runtimes (AutoResolveIntegrationRuntime and SelfHostedIR), and a "New" button to add more.

The screenshot shows the Microsoft Edge browser with the same URL as the previous screenshot. The main content area displays the "Linked services" blade for the same Data Factory "amikadff". This blade shows a table with one entry ("FileServer1") and a "New" button to add more.

The image shows a Windows desktop environment with two Microsoft Azure Data Factory management interfaces running in separate windows.

**Top Window:** Microsoft Integration Runtime Configuration Manager

- Header:** 2025-104183 - Remote Desktop Connection, annikaadff - Microsoft Azure, annikaadff - Azure Data Factory, Download Microsoft Integration Runtime, Welcome to Microsoft Edge.
- Content:** Home, Settings, Diagnostics, Update, Help. A green checkmark indicates "Self-hosted node is connected to the cloud service".
  - Data Factory:** annikaadff
  - Integration Runtime:** SelfHostedIR
  - Node:** annikavmday7

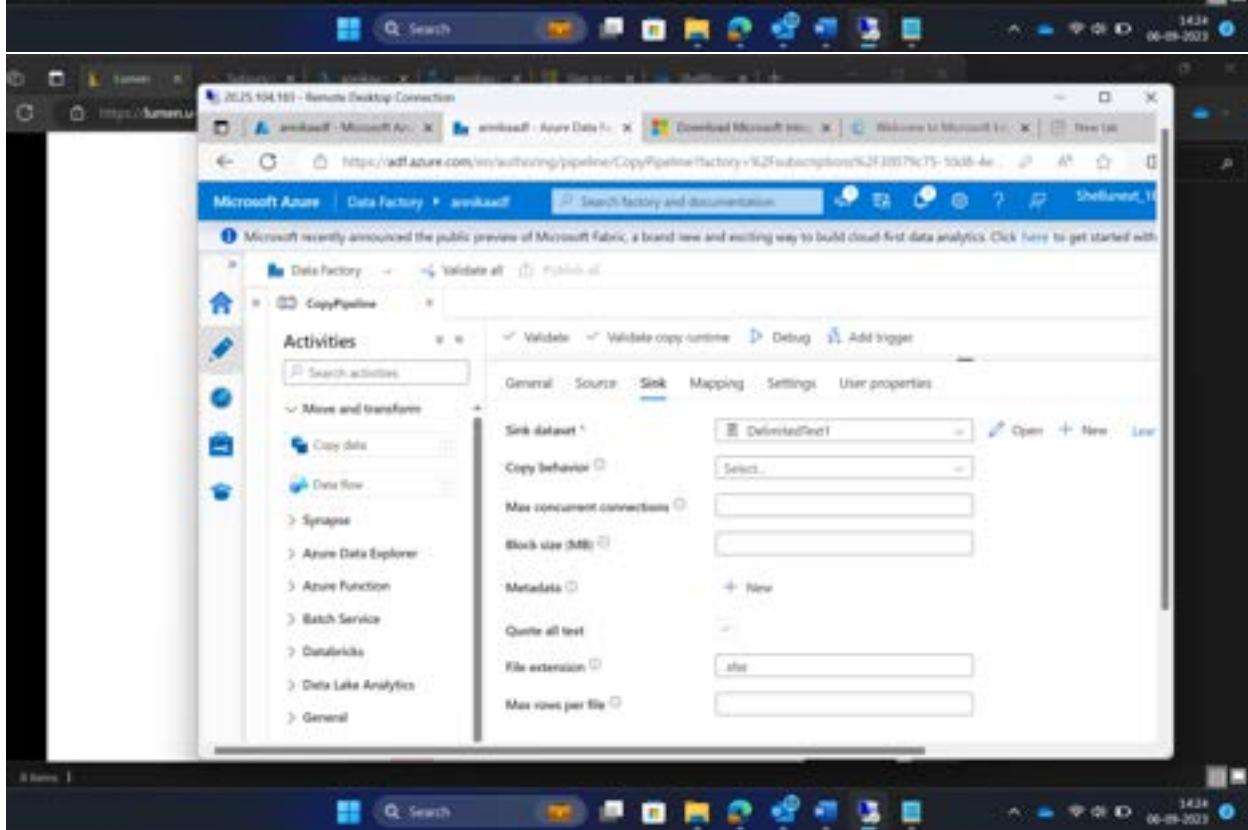
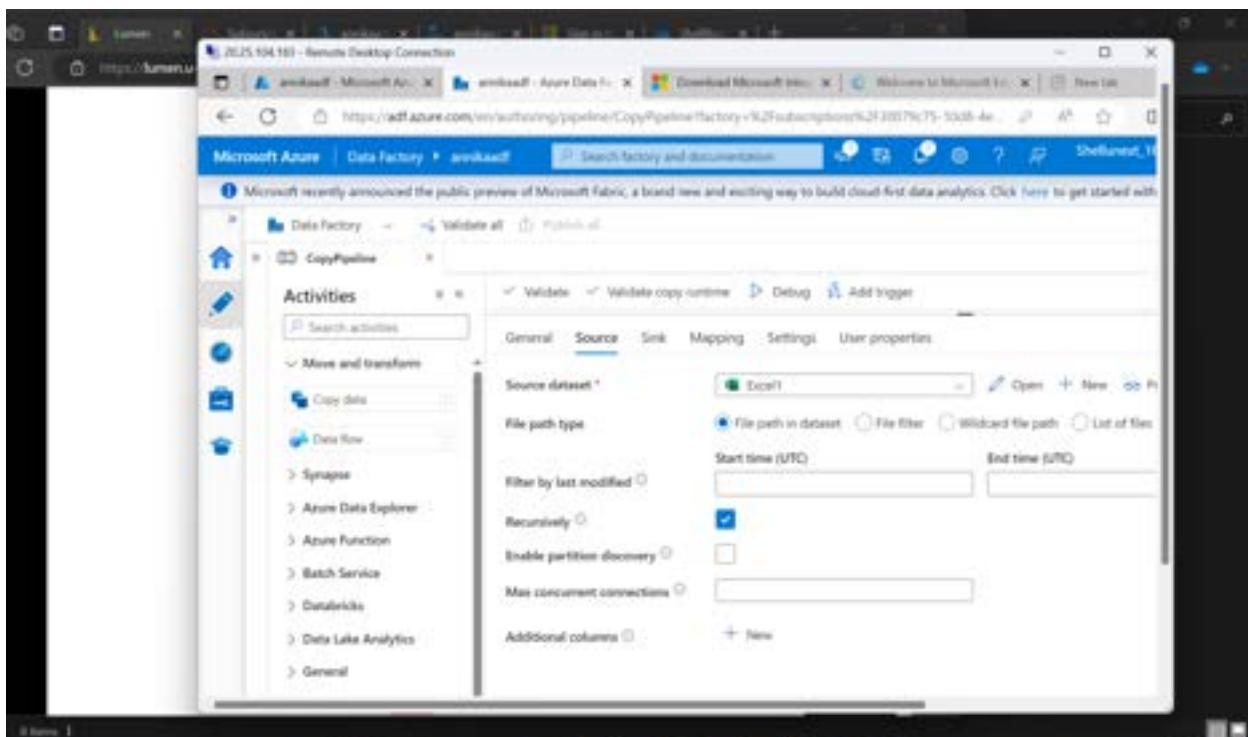
Buttons: Stop Service, Generate Backup, Import Backup.

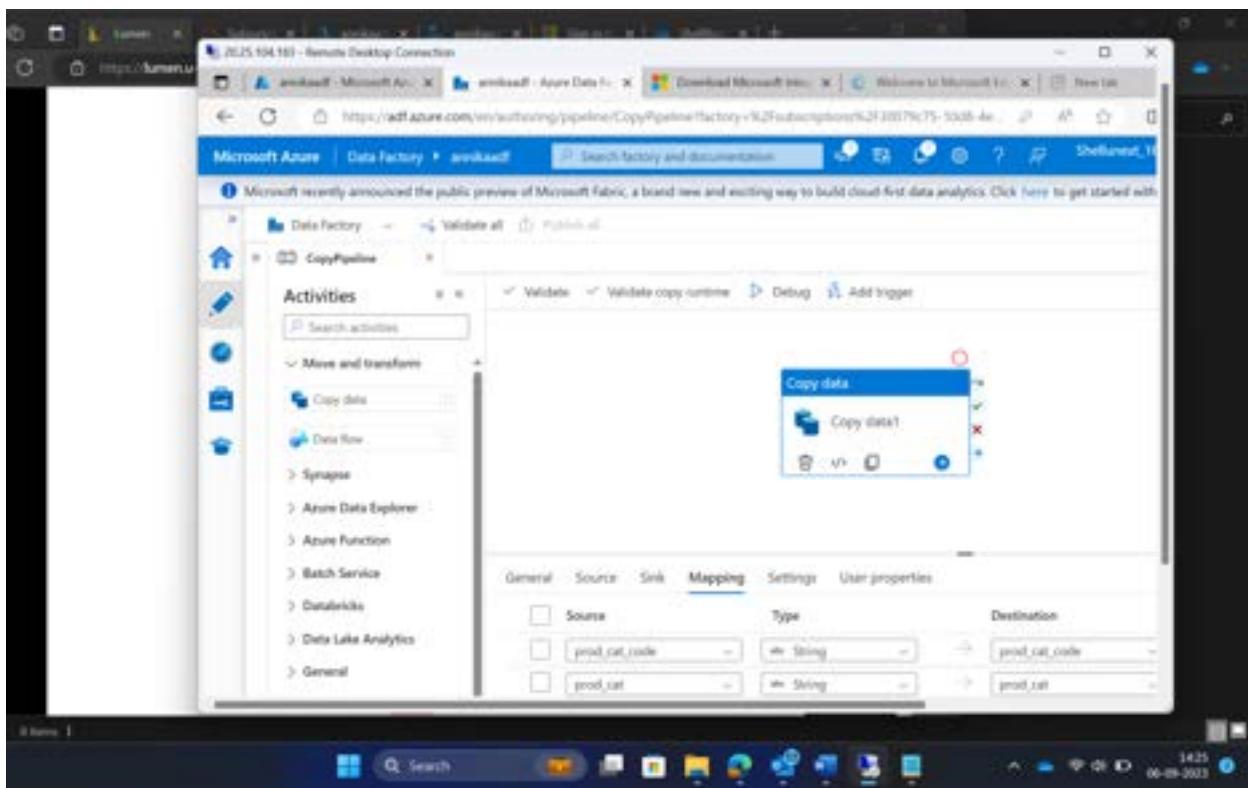
Message: Connected to the cloud service (Data Factory V2).

**Bottom Window:** Microsoft Azure | Data Factory | annikaadff

- Header:** 2025-104183 - Remote Desktop Connection, annikaadff - Microsoft Azure, annikaadff - Azure Data Factory, Download Microsoft Integration Runtime, Welcome to Microsoft Edge.
- Content:** Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Microsoft Fabric.
  - Left Sidebar:** General, Factory settings, Connections, Linked services, Integration runtimes, Microsoft Purview, Source control, Git configuration, ARM template, Author, Triggers, Global parameters, Data flow libraries.
  - Right Content:** Linked services. Subtitle: Linked service defines the connection information to a data store or compute. Learn more. New. Filter by name: Any. Showing 1 - 2 of 2 items.

Name	Type	Related
AzureBlockBlobStorage1	Azure Block Storage	0
FileServer1	File system	0





## Lab 1 Azure Data Factory

The image displays two separate Microsoft Edge browser windows running on a Windows 10 desktop. Both windows have their taskbar icons visible at the bottom.

**Top Window (Container Storage):**

- Address bar: https://portal.azure.com/#view/microsoft\_azure\_storage/containersblade/-/overview/storageaccount/democontainer
- Title: Microsoft Azure - democontainer - Containers
- Content: Shows the Azure Storage Container 'democontainer' with an 'Overview' tab selected. It lists a single blob named 'inputImg.tif'. Other tabs include 'Search', 'Upload', 'Change access level', 'Refresh', 'Delete', 'Change key', 'Access level', and 'Arch.'
- Left sidebar: Includes 'Search resources, services, and docs (S4L)', 'Home > annikaday7 | Containers > democontainer', 'Overview', 'Diagnose and solve problems', 'Access Control (IAM)', 'Settings' (with 'Shared access tokens', 'Access policy', 'Properties', and 'Metadata' options), and an 'Add filter' button.

**Bottom Window (Database Management):**

- Address bar: https://portal.azure.com/#@inputmp-test/annikadabwork/annikaday7lab
- Title: Microsoft Azure - annikadabwork (annikadabwork/annikaday7lab) | Query editor (preview)
- Content: Shows the Azure Data Studio 'Query editor (preview)' interface. A table named 'annikaday7lab (sqlabday7)' is listed under 'Tables'. The 'Query 1' pane contains T-SQL code for creating a table:

```
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
```

- Left sidebar: Includes 'Search resources, services, and docs (S4L)', 'Home > annikaday7lab (annikadabwork/annikaday7lab)', 'Overview', 'Activity log', 'Tags', 'Diagnose and solve problems', 'Query editor (preview)' (selected), 'Settings' (with 'Compute + storage', 'Connection strings', 'Properties', 'Locks', 'Data management', and 'Replica' options), and a 'Create a...' button.

The screenshot shows the Microsoft Azure Data Factory interface. On the left, a vertical navigation pane lists steps: Properties, Source, Dataset, Configuration, Destination, Settings, and Review and finish. The 'Source' step is currently selected. The main panel displays the 'Preview data' section. It shows a preview of data from a linked service named 'Azure Blob Storage'. The object being previewed is 'democontainer/inputEmp.txt'. The 'Preview' tab is active, showing two rows of data: John Doe and Jane Doe. The 'Schema' tab is also visible. The status bar at the bottom indicates '2 items'.

The screenshot shows the Microsoft Azure Data Factory interface. The vertical navigation pane on the left is identical to the previous screenshot, with 'Source' selected. The main panel now displays the 'Destination data store' configuration step. It asks to specify the destination data store for the copy task. Under 'Destination type', 'All' is selected. Under 'Connection', a 'Select...' button is shown. To the right, a 'New connection' dialog is open. It shows 'Azure SQL Database' as the selected type. Under 'Account selection method', 'From Azure subscription' is selected. An 'Azure subscription' dropdown shows 'rgnuser1-1680261947025 (0007c75-10d8-4ec8-9514-795b...' as the selected item. The 'Server name' field contains 'arnikadatadb'. The 'Database name' field contains 'arnikadatadb'. The 'Authentication type' is set to 'SQL authentication'. The 'User name' field contains 'sqladmin7'. The 'Password' field is filled with '\*\*\*\*\*'. A 'Save' button is at the bottom of the dialog. The status bar at the bottom indicates '2 items'.

The screenshot shows two consecutive steps in the Microsoft Azure Data Factory Copy Data tool:

**Step 1: Destination data store**

The left sidebar shows the navigation path: Microsoft Azure | Data Factory > arekawaf. The main panel is titled "Copy Data tool" and displays the "Destination" step. It asks to specify the destination data store for the copy task, using an existing connection or creating a new one. The "Destination type" dropdown is set to "All". A connection named "AzureSqlDatabase1" is selected. The "Source" dropdown is set to "inputDmp" and the "Target" dropdown is set to "dbo". A link "Use existing table" is present.

**Step 2: Column mapping**

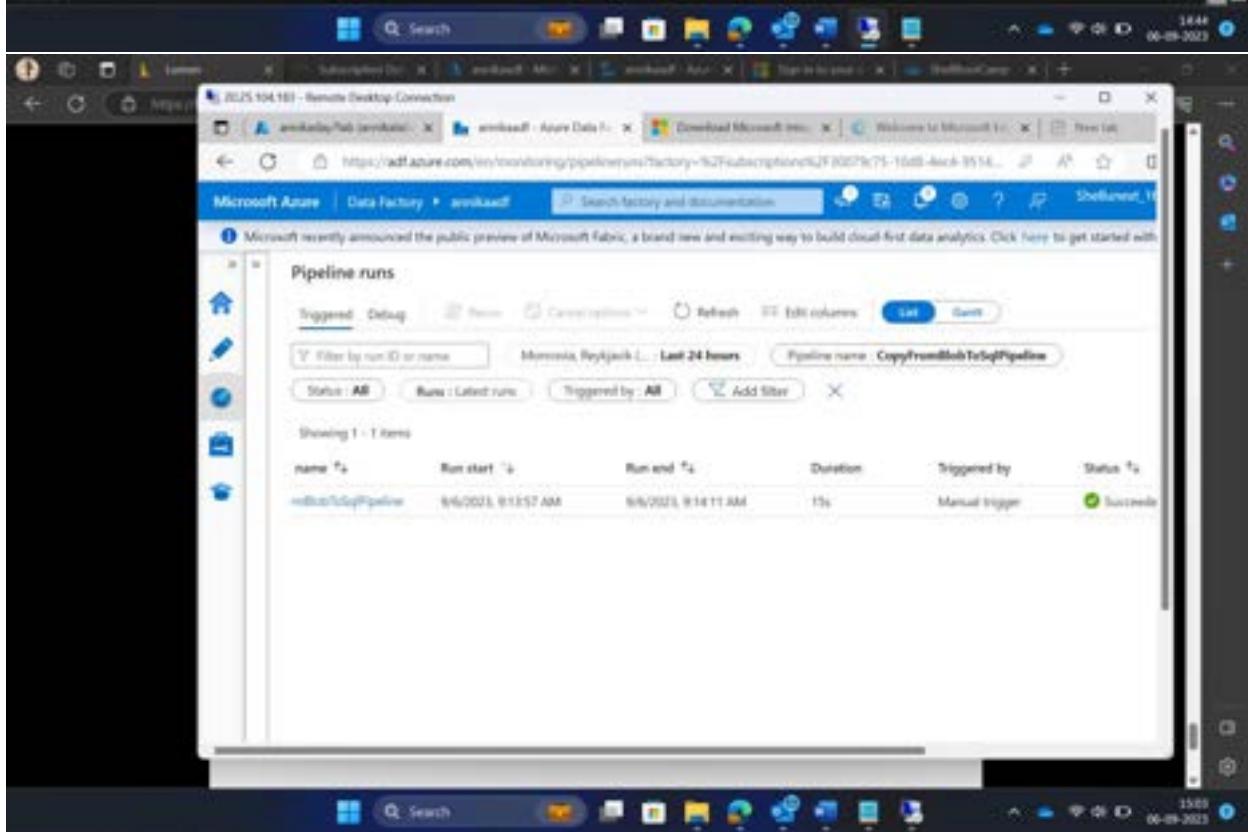
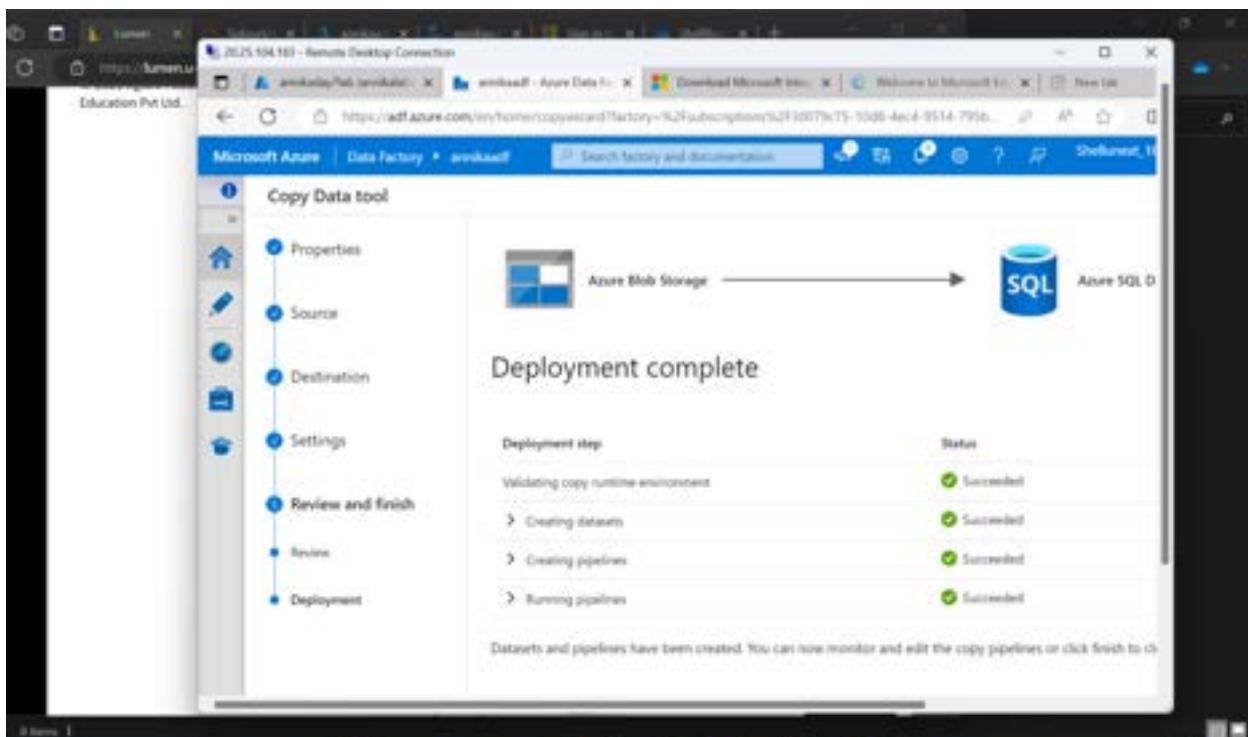
The left sidebar shows the navigation path: Microsoft Azure | Data Factory > arekawaf. The main panel is titled "Copy Data tool" and displays the "Column mapping" step. It shows "Table mappings (1)" for "Azure Web Storage File" as the source and "dbo" as the target. Under "Column mappings", there is a table with three rows:

Type conversion settings
+ New mapping
Clear
Reset
Delete

The table lists three columns:

Source	Type
KeyName	= String
LastName	= String

Below the table, under "Azure SQL Database sink properties", there is a section for "Pre-copy script" which is currently empty.



The screenshot displays two windows of the Microsoft Azure Data Factory interface.

**Top Window:** Shows the pipeline editor for the 'CopyFromBlobToSqlPipeline'. The left sidebar lists 'Factory Resources' including Pipelines, Datasets, Data Flows, and Power Query. The main area shows the pipeline structure with a 'Get Metadata' activity followed by a 'Copy data' activity. The 'Get Metadata' activity is selected, showing its configuration details: General tab with Name 'Get Metadata1' and Description 'Get Metadata1'; Settings tab; and User properties tab. The 'Copy data' activity is also visible.

**Bottom Window:** Shows the pipeline run history for the same pipeline. The left sidebar lists various service activities. The main area displays the pipeline run ID: '1b0cd3233-4a4-4b4d-a02e-c700d6d008c'. It shows a table of activities with their status, type, start time, and duration. Two activities are listed: 'Copy\_mds' (Status: Succeeded, Type: Copy data, Start: 9/6/2023, 10:28:26 AM, Duration: 10s) and 'GetInformation' (Status: Succeeded, Type: Get Metadata, Start: 9/6/2023, 10:28:26 AM, Duration: 2s).

The screenshot shows the Azure Data Factory pipeline details for a run ID of 1bcd3233-4aa4-4b4d-a02a-c702d5d008c. The pipeline status is green. The pipeline has two activities:

Activity name	Activity status	Activity type	Run start	Durati
Copy_m55	Succeeded	Copy data	9/6/2023, 10:28:26 AM	10s
GetInformation	Succeeded	Get Metadata	9/6/2023, 10:28:26 AM	2s

The screenshot shows the output details for the Copy\_m55 activity. The activity type is Copy data, and it succeeded. The output details include:

```
    "DataRead": 38,
    "DataWritten": 28,
    "FileRead": 1,
    "InafkaReadConnections": 1,
    "InafkaPeakConnections": 2,
    "Download": 2,
    "RowsCopied": 2,
    "CopyDuration": 8.
```

The pipeline status is green. The pipeline has two activities:

Activity name	Activity status	Activity type	Run start	Durati
Copy_m55	Succeeded	Copy data	9/6/2023, 10:28:26 AM	10s
GetInformation	Succeeded	Get Metadata	9/6/2023, 10:28:26 AM	2s

Day 8

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

Microsoft Azure | Data Factory | [Create](#) | [Search factory and documentation](#)

Preview experience

1 Microsoft recently announced the public preview of Microsoft Fabric, a brand-new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Data Factory > [New resource](#) [Search](#)

Factory Resources

- Pipelines
- Change Data Capture (preview)
- Datasets
- AzureSqlTable1**
- Data flows
- Power Query

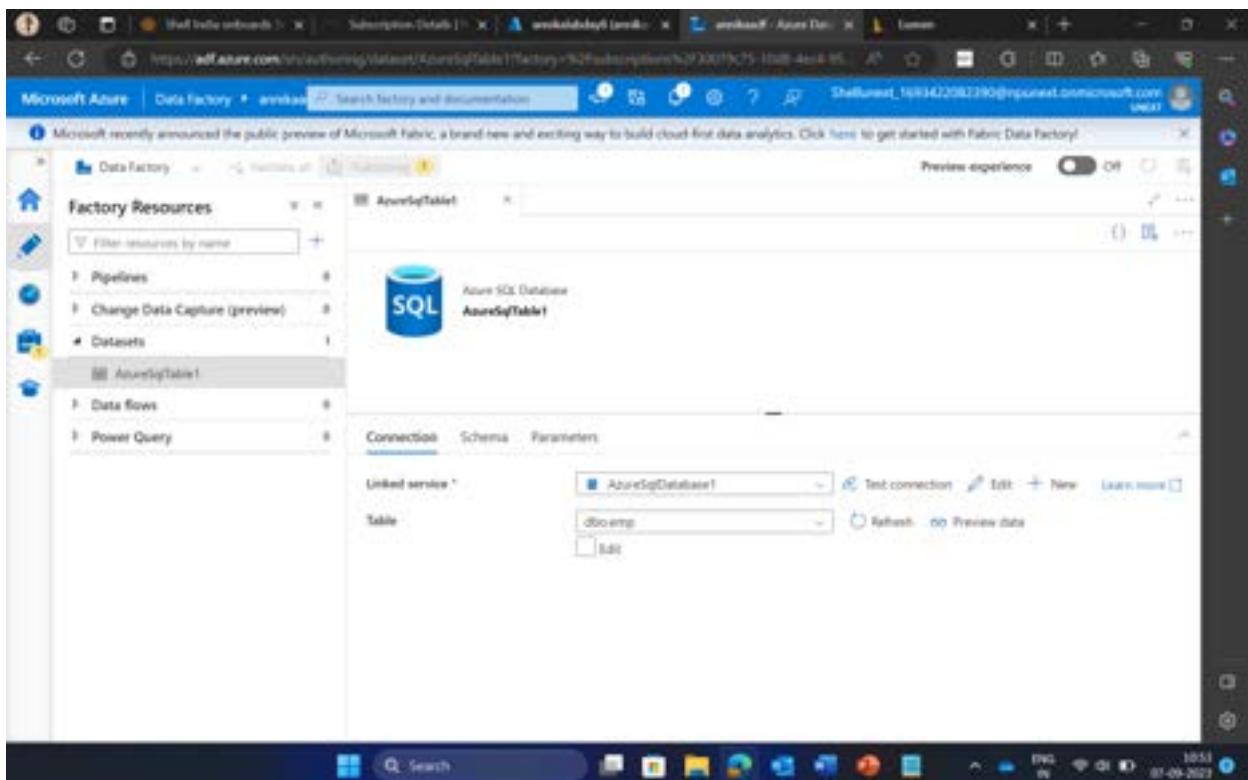
AzureSqlTable1

Azure SQL Database

Connection Schema Parameters

Untitled service  AzureSqlDatabase1  Test connection  Edit  New  Learn more

Table  dbo.emp  Refresh  Preview data  Edit



Microsoft Azure | Data Factory | [Create](#) | [Search factory and documentation](#)

Preview experience

1 Microsoft recently announced the public preview of Microsoft Fabric, a brand-new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Data Factory > [New resource](#) [Search](#)

Factory Resources

- Pipelines
- pipeline1**
- Change Data Capture (preview)
- Datasets
- AzureSqlTable1**
- Data flows
- Power Query

pipeline1

Validate  Debug  Add trigger

Activities

- Append variable
- Delete
- Execute Pipeline
- Execute OData package
- Fail
- Get Metadata
- Lookup
- Stored procedure
- Script
- Set variable
- Validation

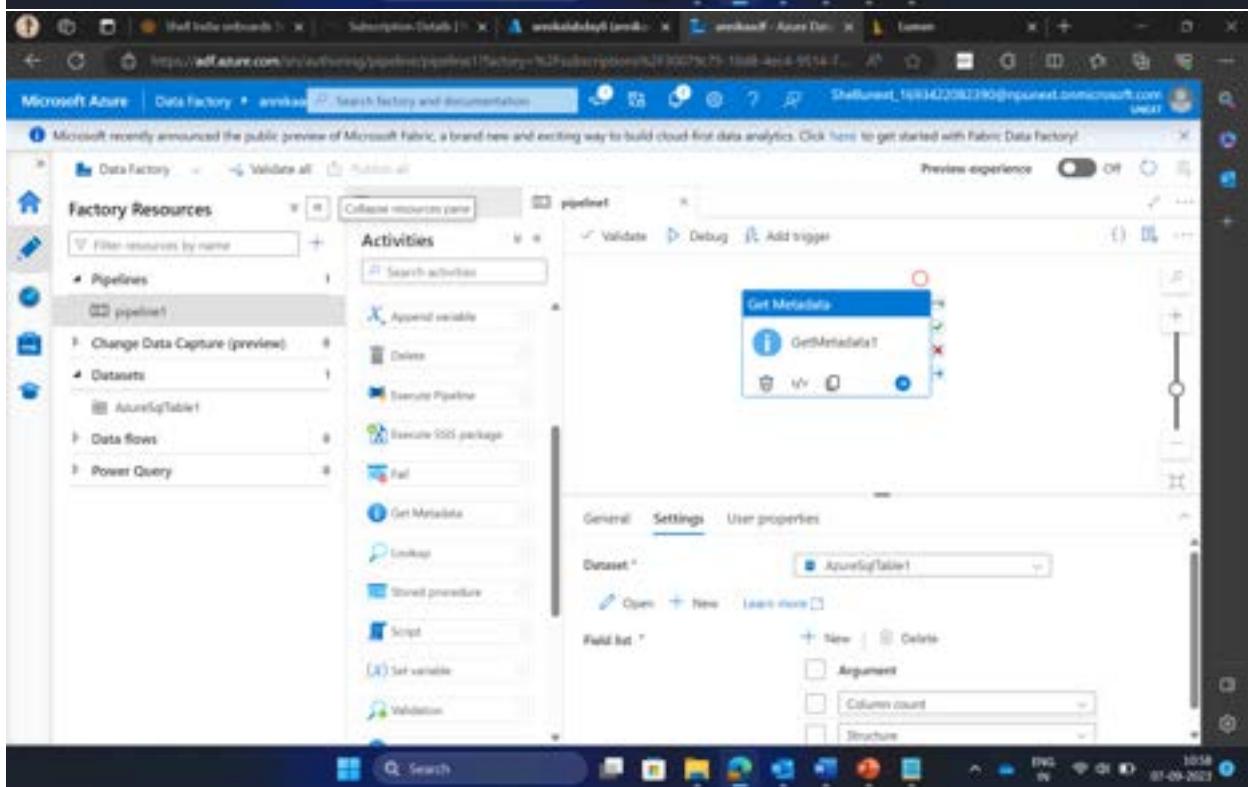
Get Metadata

GetMetadata1

General Settings User properties

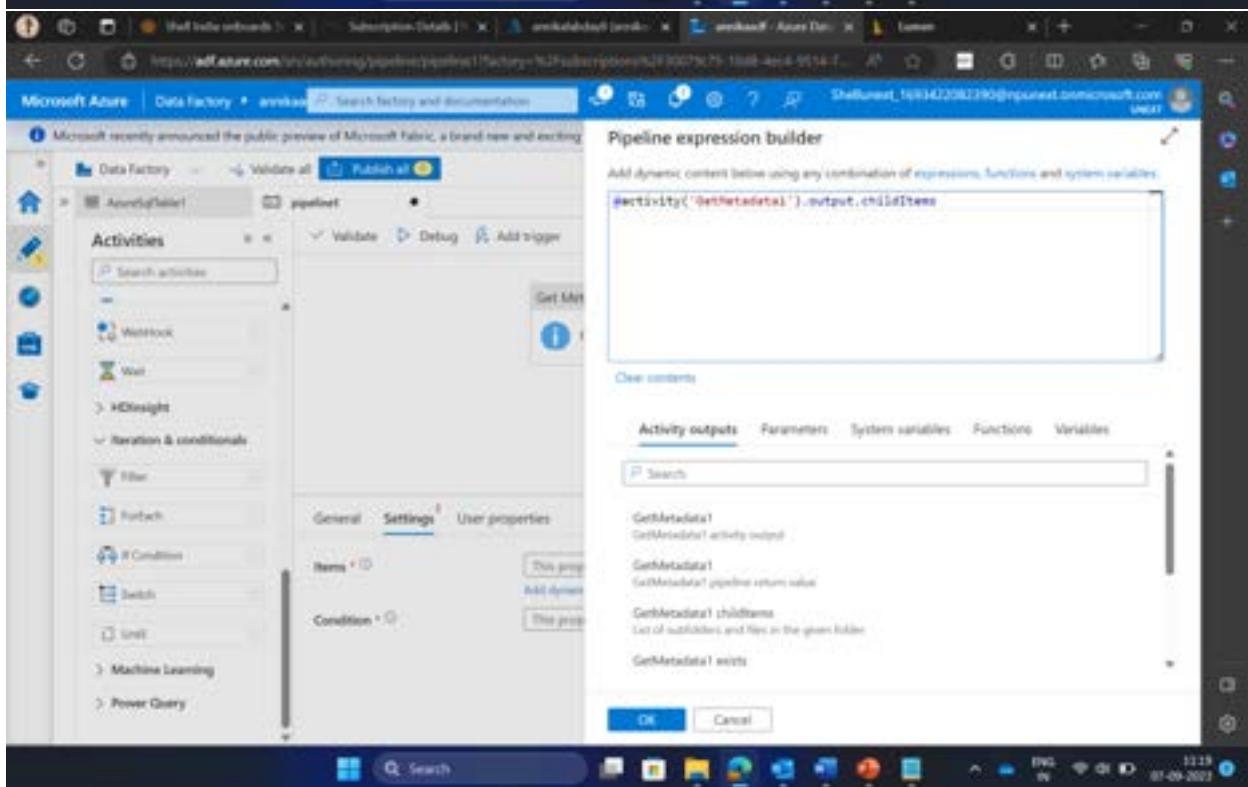
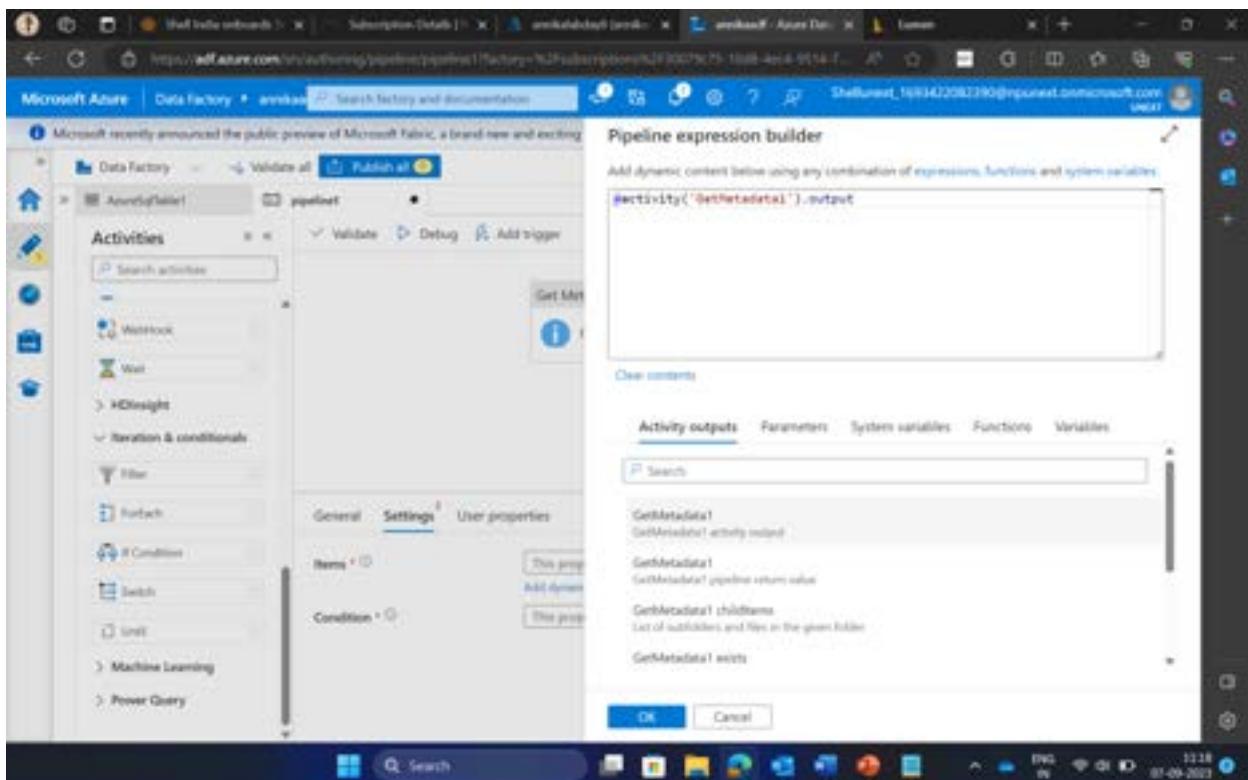
Dataset  AzureSqlTable1  Open  New  Learn more

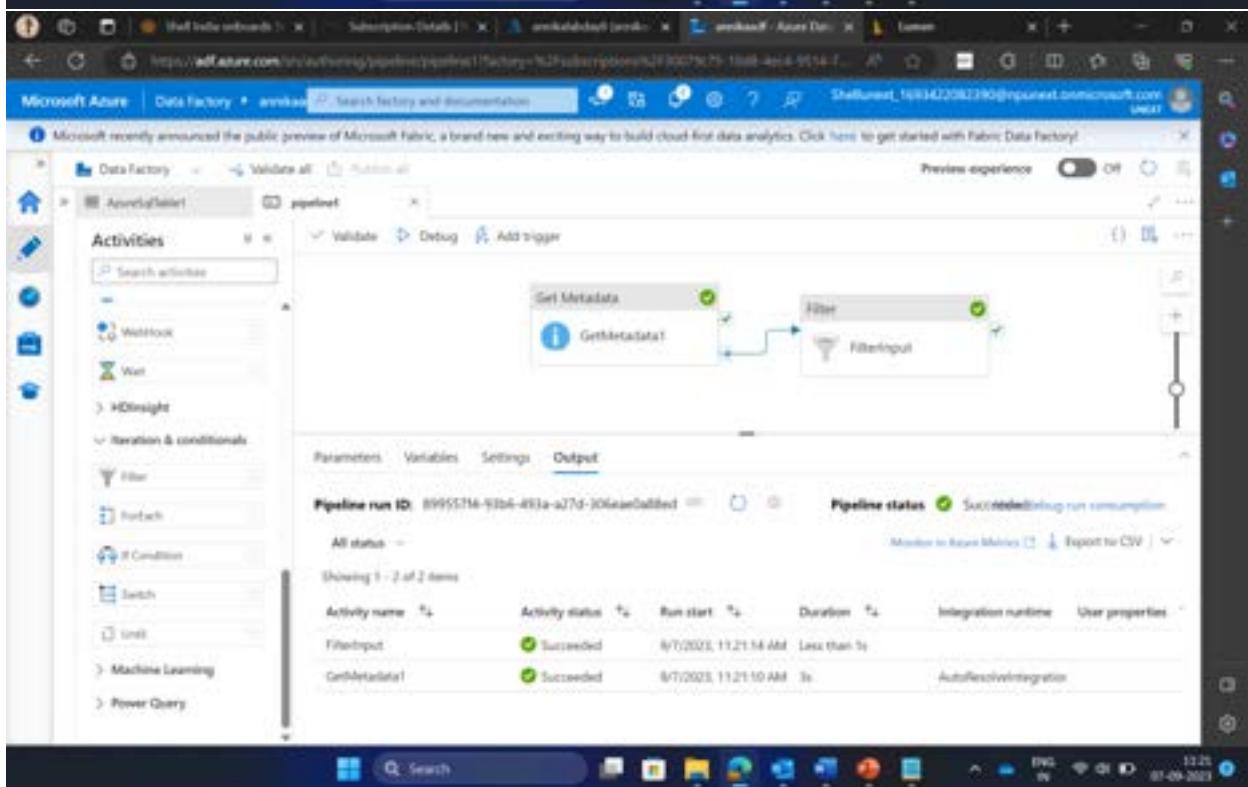
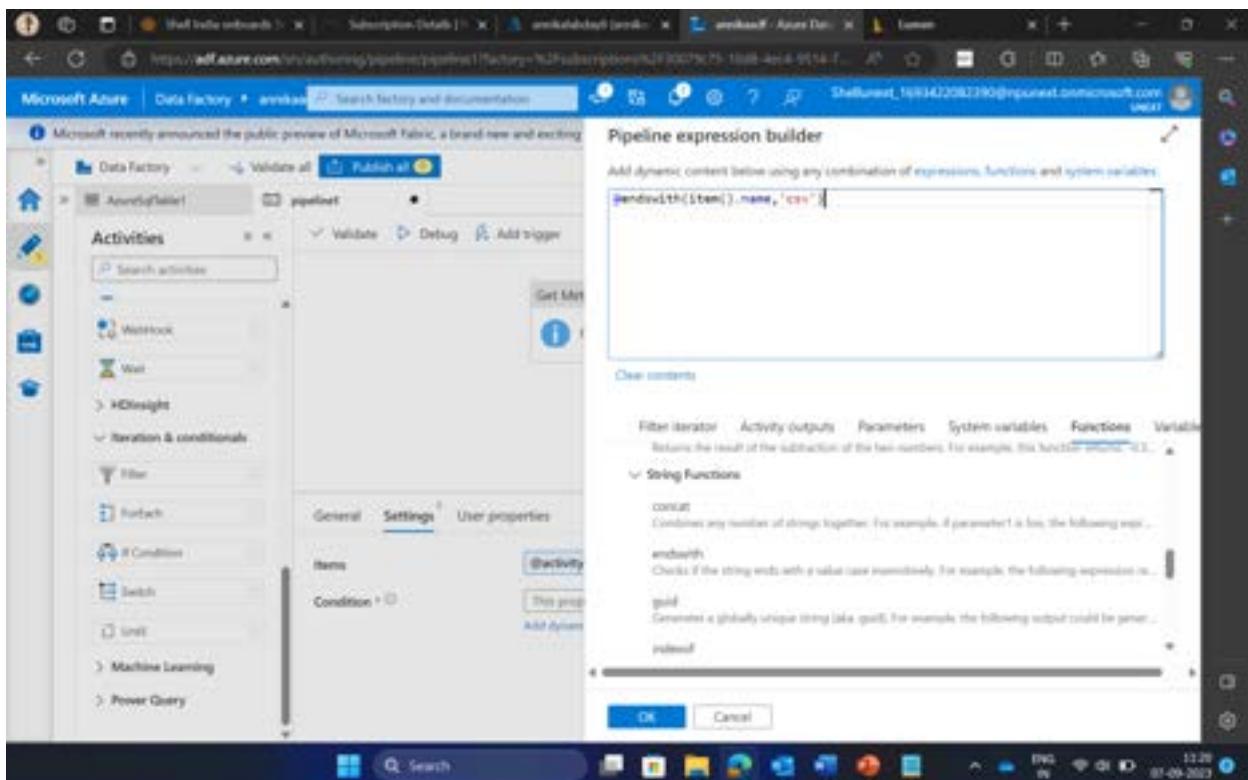
Field list  + New  Delete Argument  Column count  Structure

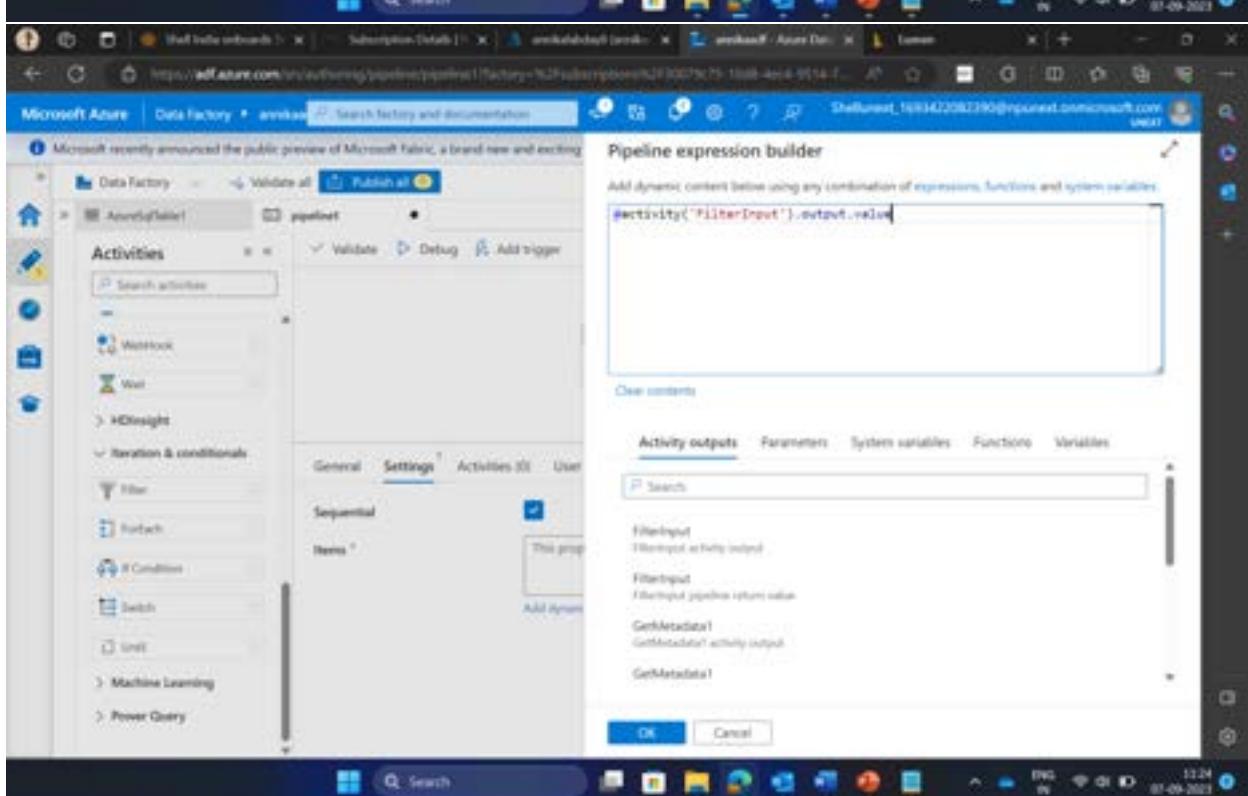
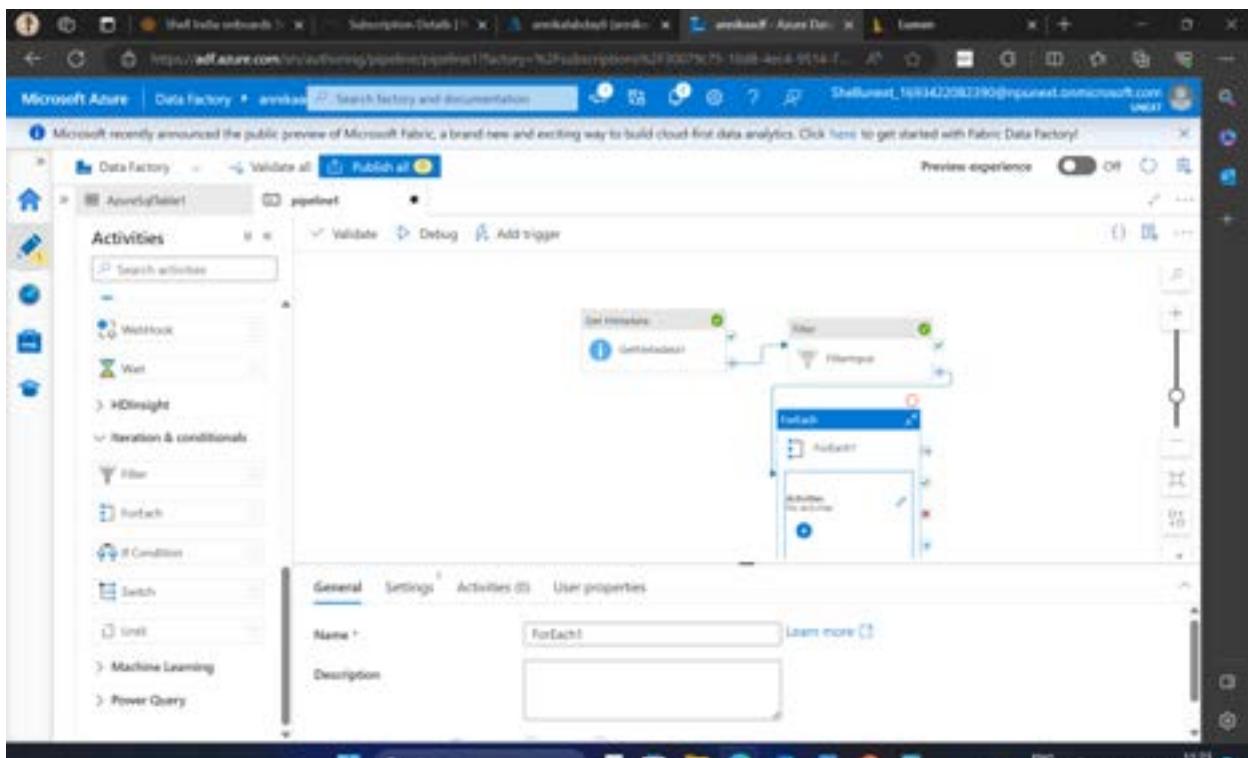


The screenshot shows the Microsoft Azure Data Factory pipeline editor interface. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (selected), 'Change Data Capture (preview)', 'Datasets' (including 'AzureSqlTable' and 'DelimitedText'), 'Data flows', and 'Power Query'. The main workspace displays a pipeline named 'pipeline1' with a single activity: 'Get Metadata' (activity ID: GetMetadata1). The 'Output' tab for this activity shows the following JSON output:

```
{ "items": [ { "name": "File1.csv", "type": "File" }, { "name": "File2.csv", "type": "File" }, { "name": "File3.csv", "type": "File" } ], "effectiveIntegrationRuntime": "AutoResolveIntegrationRuntime (East US)", "executionDuration": 0, "availabilityQueue": 0, "integrationRuntimeQueue": 0, "billingReference": { "activityType": "PipelineActivity", "billableDuration": 0, "meterType": "AzureB", "duration": 0.01666666666666666, "unit": "Hour" } }
```







Microsoft Azure | Data Factory | pipeline1 | Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand-new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!

Data Factory > Validate all Publish

AnonDataFactory pipeline1

Activities

Validate All Debug Add trigger

Get Metadata

GetFileSize

General Settings User properties

Name: GetFileSize

Description:

Activity state (preview): Active

Timeout: 0:12:00:00

Retry: 0

Retry interval (sec): 30

Search

Microsoft Azure | Data Factory | pipeline1 | Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand-new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!

Data Factory > Validate all Publish

AnonDataFactory pipeline1

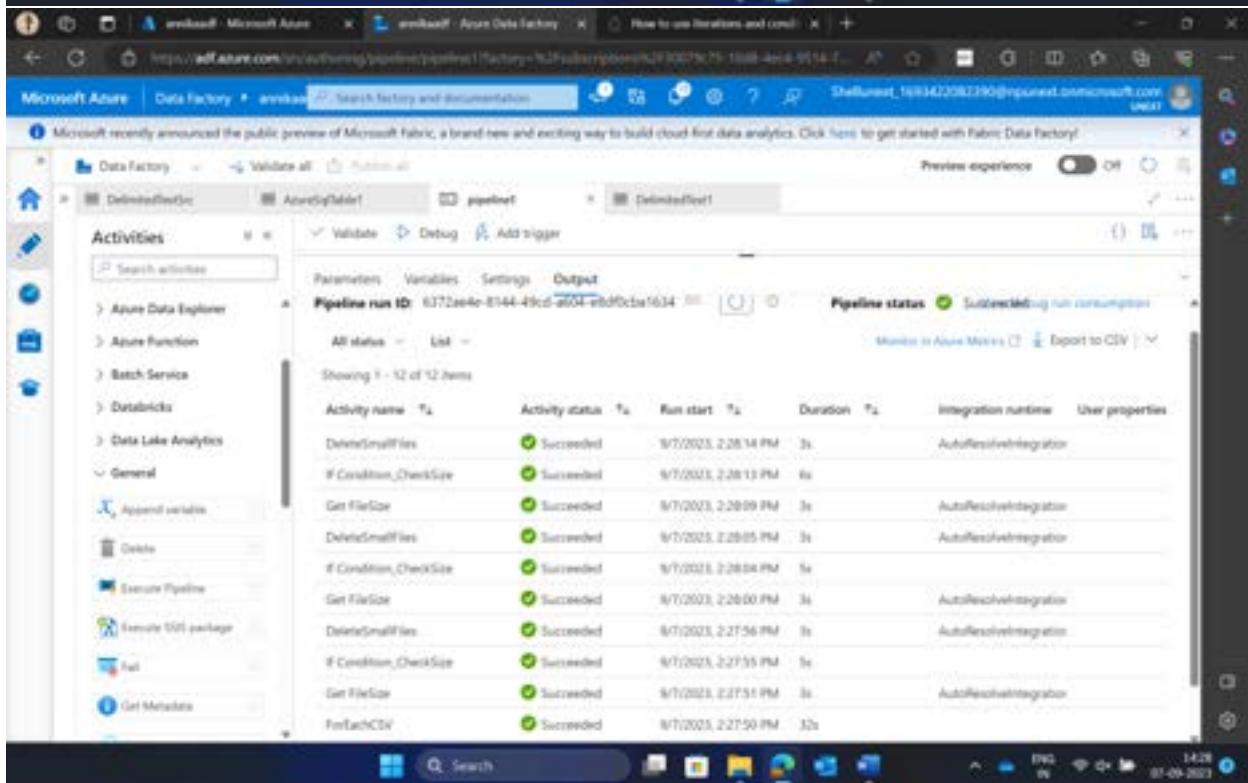
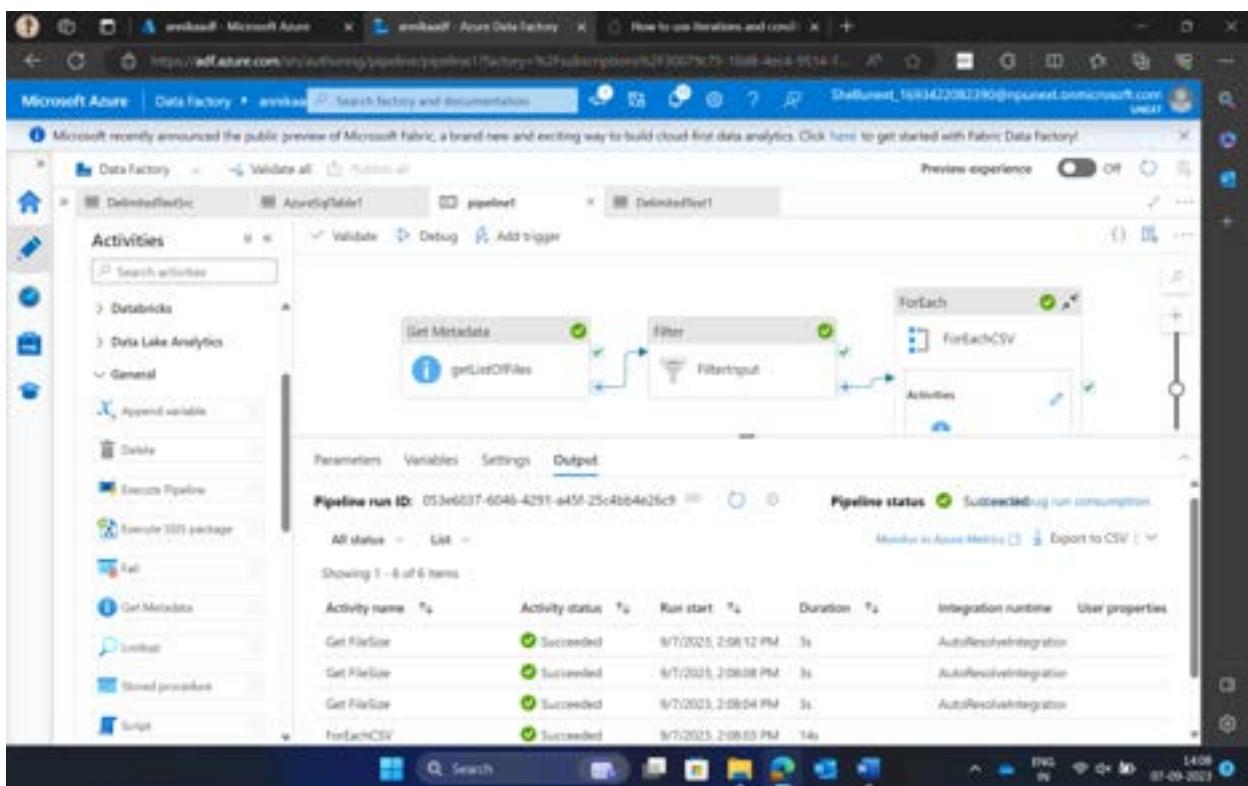
DelimitedText1

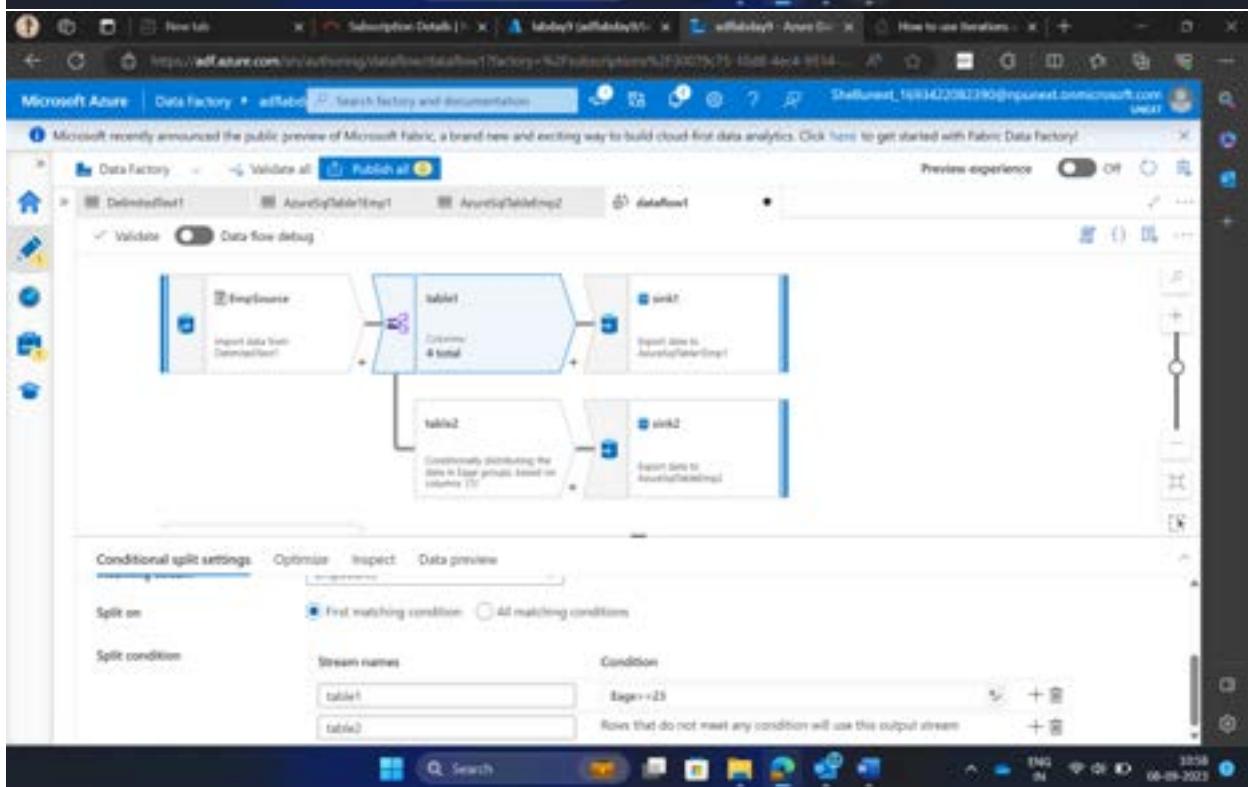
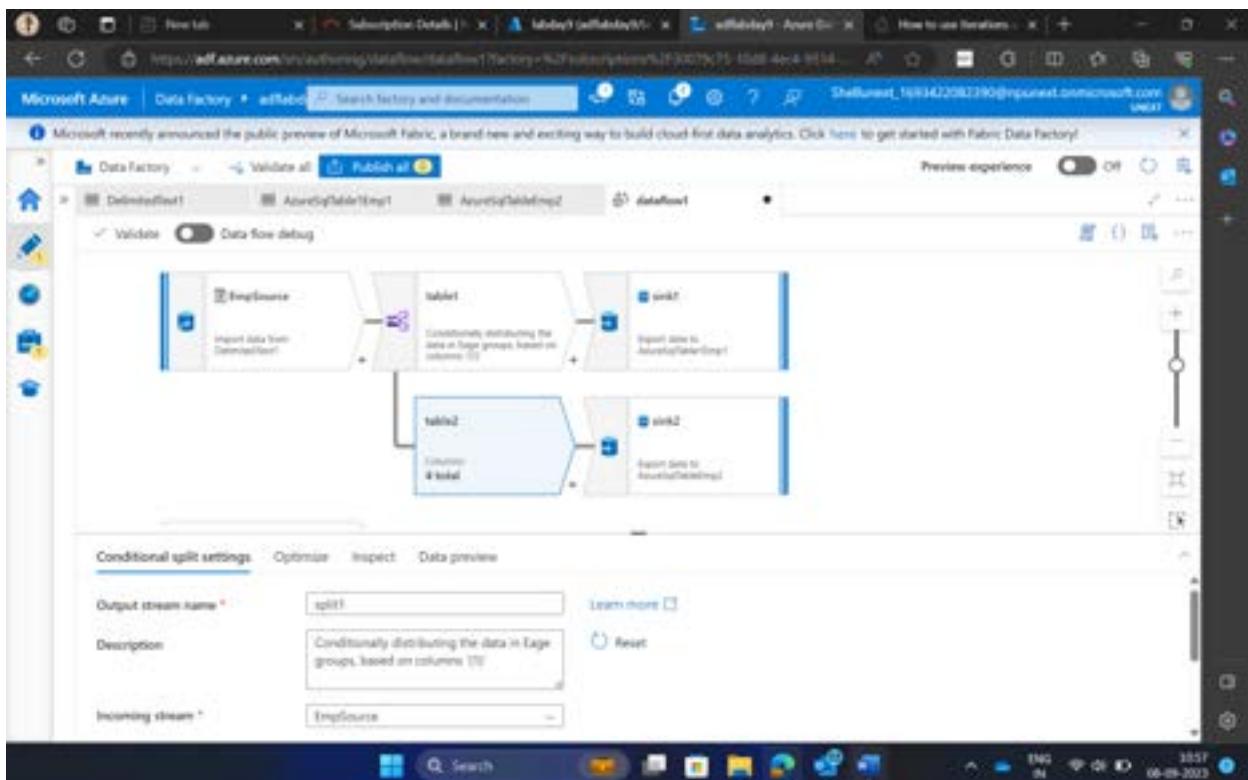
Connection Schema Parameters

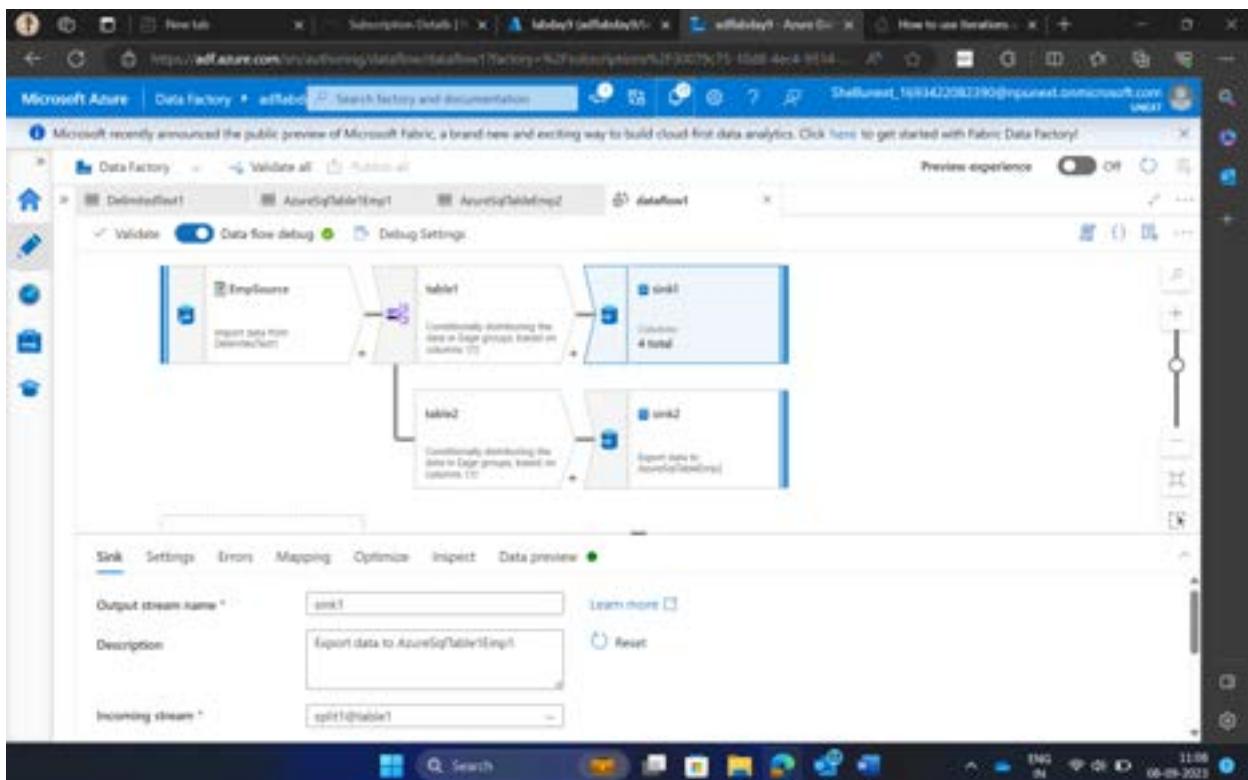
+ New Delete

Name: filename Type: String Default value: csv

Search







The screenshot shows the Microsoft Azure Data Factory Data Flow interface with the 'Data preview' tab selected. The preview area displays the following data:

	ID	Name	Age	Salary
1	Ashley	23	34223	
2	Irene	24	23412	
3	Rose	27	24481	

Microsoft Azure | Data Factory | adfRebo | Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand-new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Data Factory > Validate all Publish all

DefinableDev1 AzureSqlTableSink1 AzureSqlTableSink2 pipeline

Activities

Move and transform

Copy data

Data flow

Synapse

Azure Data Explorer

Azure Function

Batch Service

Databricks

Data Lake Analytics

General

HDInsight

Iteration & conditionals

Machine Learning

Data Flow

Data Row1

Parameters Variables Settings Output

Pipeline run ID: 6999431d-cff9-4a07-8ba1-d118c981eebb Pipeline status: Succeeded

All status Monitor In-Pause Metrics Export to CSV

Showing 1 - 1 of 1 items

Activity name	Activity status	Run start	Duration	Integration runtime	User properties
Data Row1	Succeeded	8/6/2023, 11:27:11 AM	52s	AutoResolveIntegration	

Search

The screenshot shows the Microsoft Azure Data Factory pipeline editor interface. On the left, there's a sidebar with various activity types like Move and transform, Copy data, Data flow, Synapse, etc. The main workspace shows a single Data Flow activity named 'Data Row1'. Below the workspace, the 'Output' tab is selected, displaying a table of pipeline run details. The table has columns for Activity name, Activity status, Run start, Duration, Integration runtime, and User properties. One row is shown: 'Data Row1' with status 'Succeeded', run start at '8/6/2023, 11:27:11 AM', duration '52s', and integration runtime 'AutoResolveIntegration'. The pipeline run ID is 6999431d-cff9-4a07-8ba1-d118c981eebb, and the pipeline status is 'Succeeded'.

The screenshot shows the Microsoft Azure Data Factory pipeline editor. A 'Copy data' activity is selected in the center workspace. The left sidebar lists various activity types: Move and transform, Synapse, Azure Data Explorer, Azure Function, Batch Service, Databricks, Data Lake Analytics, General, HDInsight, Iteration & conditionals, Machine Learning, and Power Query. The 'Output' tab is active for the selected 'Copy data' activity. Pipeline run details show a successful run ID: dMa70a1-c6d0-449f-9471-a574a96a4c7. The pipeline status is green, indicating success. Below the pipeline run details, a table displays the activity run status, showing one item named 'Copy data' with a status of 'Succeeded'. The bottom section of the editor shows a preview of the data being copied, displaying a JSON-like structure with fields: base, visibility, dt, timezone\_id, name, cod, stations, and a timestamp.

Microsoft Azure | Data Factory | adffaboy | Search factory and documentation

Microsoft recently announced the public preview of Microsoft Fabric, a brand-new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Activities

Copy data

Parameters Variables Settings Output

Pipeline run ID: dMa70a1-c6d0-449f-9471-a574a96a4c7

Pipeline status: Succeeded (debug run consumption)

Activity name	Activity status	Run start	Duration	Integration runtime	User properties
Copy data	Succeeded	9/6/2023, 12:31:53 PM	13s	AutoResolveIntegrationRuntime	

data\_f35c7587-ebca-47de-98e4-5956caf8e20b\_94f9c6f7-dc20-4af6-8de0-965eee3f4778.txt

base,visibility,dt,timezone\_id,name,cod  
"stations",8888,1694156140,3600,2643743,"London",200

Test Preview

## CDC

The screenshot displays two windows from the Microsoft Azure Data Factory interface.

**Top Window: Choose Your Sources**

- Left Panel:** Shows the "Factory Resources" navigation pane with sections: Pipelines, Datasets, Data flows, and Power Query. Under Pipelines, "pipeline1" and "pipeline2" are listed, with "pipeline2" currently selected.
- Right Panel:** A configuration dialog for "Change Data Capture (CDC)".
  - CDC name:** adfdc1
  - Source type:** DelimitedText
  - Source linked service:** AzureBlobStorage1
  - Source settings:** (button)
  - Folder paths:** adfdc1

**Bottom Window: Pipeline Configuration**

- Left Panel:** Shows the "pipeline1" pipeline configuration with a "Data flow" step selected.
- Right Panel:** A "Data flow" configuration screen.
  - Source:** AzureSqlDatabase1
    - Source table:** t1
    - Mapping:** New mapping
  - Target:** Target table t2
    - Target table:** dtbs.tmp2
    - Columns mapped:** Auto map

The image displays two nearly identical screenshots of the Microsoft Azure Query editor (preview) interface, stacked vertically. Both screenshots show a query results page for the database 'labday9'.

**Query Editor Interface:**

- Header:** Microsoft Azure, portal.azure.com, labday9 (adflabday9/labday9), Home > labday9 (adflabday9/labday9).
- Toolbar:** Login, New Query, Open query, Feedback, Getting started.
- Object Explorer:** Shows the database 'labday9' and its tables: adftbl01, adftbl02, adftbl03, adftbl04, adftbl05, SalesLT.Address, SalesLT.Customer, and SalesLT.CustomerAddress.
- Query Editor:** Three tabs: Query 1, Query 2, and Query 3. The Query 3 tab is active, containing the following SQL code:

```
1 SELECT TOP (1000) * FROM [dbo].[adftbl01]
```
- Results:** Tab labeled 'Results'. The output shows a table with four columns: Row, Name, Age, and BirthDate. The data is as follows:

Row	Name	Age	BirthDate
1	Ram	23	28/08/1991
2	Somy	22	31/12/1992
3	Bruke	20	32/08/1994
4			
5			
6			
7			
8			
9			
10	Ram	23	28/08/1991

**Message Bar:** 'Query succeeded [0s]'.

**System Status:** Shows a taskbar at the bottom with icons for search, file, and system status.

## Monitoring and troubleshooting

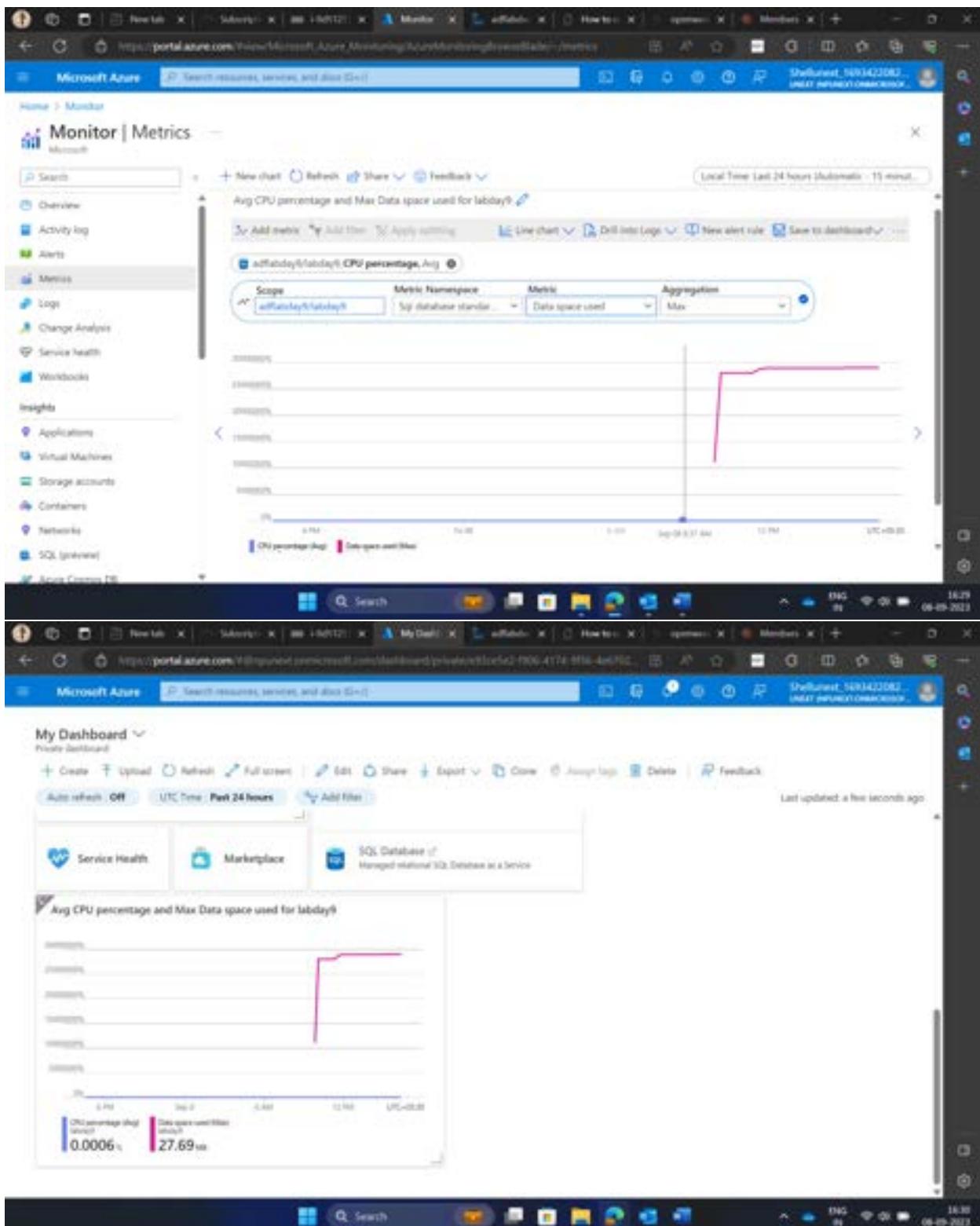
The screenshot shows the Microsoft Azure portal interface. The left sidebar navigation bar includes links for Home, Monitor, Overview, Activity log, Alerts, Metrics (which is currently selected), Log, Change Analysis, Service health, and Workbooks. Below this is an Insights section with links for Applications, Virtual Machines, Storage accounts, Containers, Networks, SQL (preview), and Azure Cosmos DB.

The main content area displays a 'Select a scope' dialog. At the top, there are tabs for 'Browse' (selected) and 'Recent'. Under 'Resource types', 'All resource types' is chosen. Under 'Locations', 'All locations' is chosen. A search bar labeled 'Search to filter items...' is present. The 'Scope' table lists resources:

Scope	Resource type	Location
subscription (100020194707)	Subscription	
+ service	Resource group	
adfsdev01	Data factory (V1)	East US
adfsdev02	Storage account	East US
+ labday01	SQL database	East US

A tooltip message at the bottom left of the dialog reads: 'Why can't I select multiple resources? You must select items of the same resource type and location. To select resources of a different resource type or location, please first uncheck your current selection.'

The 'Selected scopes' section shows '1 SQL database': 'labday01' (SQL database, East US). At the bottom of the dialog are 'Apply' and 'Cancel' buttons, and a 'Clear all selections' link.



Microsoft Azure | portal.azure.com | https://portal.azure.com/#view/Microsoft\_Azure\_Monitoring/CreateAlertRuleScope/01BFD0/SignalR

## Select a resource

Create an alert rule to identify and address issues when important metrics exceed specific thresholds.

**Scope**

+ Select scope

**Resource**

No resource selected yet.

**Browse** Recent

Resource types All resource types Locations All locations

Search to filter items...

Resource	Resource type	Location
affilateay9	Storage account	East US
affilateay9	Data factory (V2)	East US
affilateay9	SQL Server	East US
affilateay9	SQL database	East US
affilateay9	SQL database	East US

Selected resources 1 storage account

affilateay9 Storage account East US

Review + create Previous Next: Condition Apply Cancel Clear all selections

Microsoft Azure | portal.azure.com | https://portal.azure.com/#view/Microsoft\_Azure\_Monitoring/CreateAlertRuleScope/01BFD0/SignalR

## Create an alert rule

Configure when the alert rule should trigger by selecting a signal and defining its logic.

**Condition**

Signal name \* Used capacity

Show all signals

Alert logic:

Threshold  Static  Dynamic  
Dynamic Thresholds is currently not available for this metric.

Aggregation type: Average

Operator: Greater Than

Unit: %

Threshold value \* 100

Review + create Previous Next: Actions >

[https://portal.azure.com/#view/Microsoft\\_Azure\\_Monitoring\\_Alerts/CreateActionGroupBladeSubscription](https://portal.azure.com/#view/Microsoft_Azure_Monitoring_Alerts/CreateActionGroupBladeSubscription)

## 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:  [Create new](#)

Region:

**Instance details**

Action group name:  Display name:   
The display name is limited to 12 characters.

[Review + create](#) [Previous](#) [Next: Notifications >](#)

[https://portal.azure.com/#view/Microsoft\\_Azure\\_Monitoring\\_Alerts/CreateActionGroupNotificationsBladeSubscription](https://portal.azure.com/#view/Microsoft_Azure_Monitoring_Alerts/CreateActionGroupNotificationsBladeSubscription)

## 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:  Select:

Email  
Email:

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](#)

[Review + create](#) [Previous](#) [Next: Actions >](#)

Microsoft Azure | https://portal.azure.com/#view/Microsoft\_Azure\_Monitoring/CreateAlertRuleScope/01890D91

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 group + Create action group

Action group name Contains actions

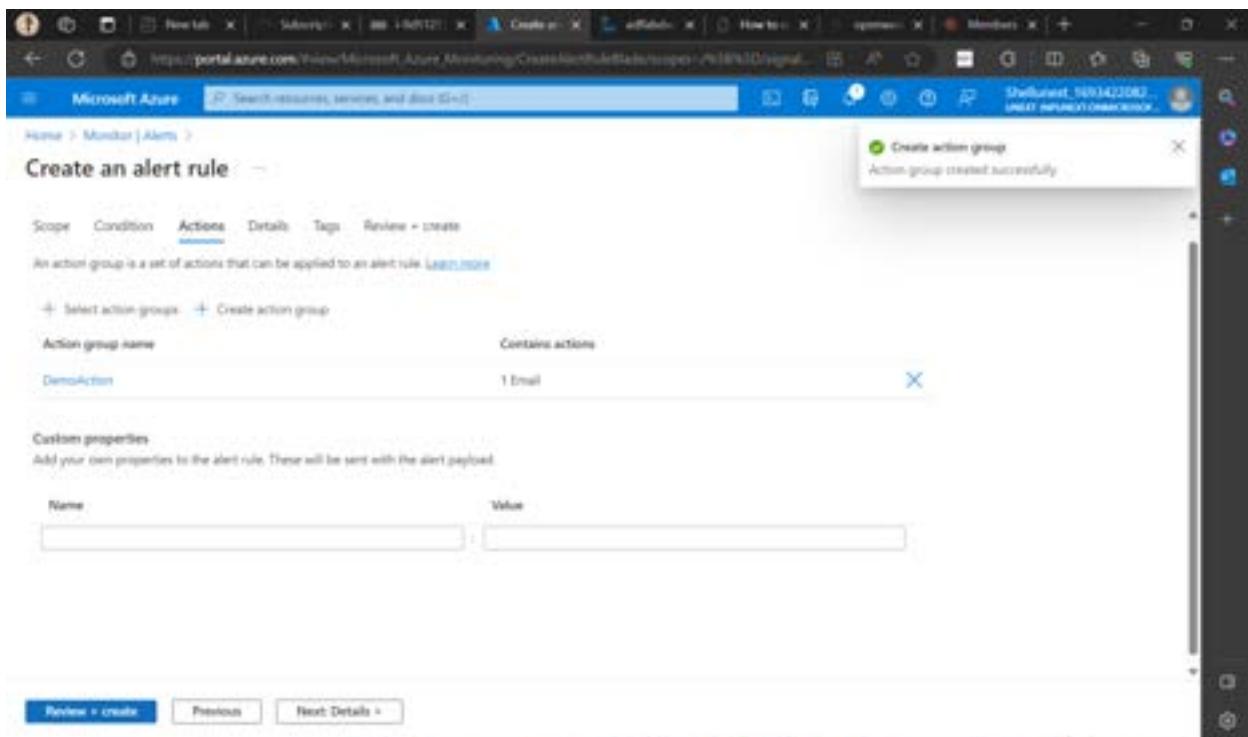
DemoAction 1 Email X

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 | https://portal.azure.com/#view/Microsoft\_Azure\_Monitoring/CreateAlertRuleScope/01890D91

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 \* (1) rpuen01:1680341947025

Resource group \* (1) amika Create new

Alert rule details

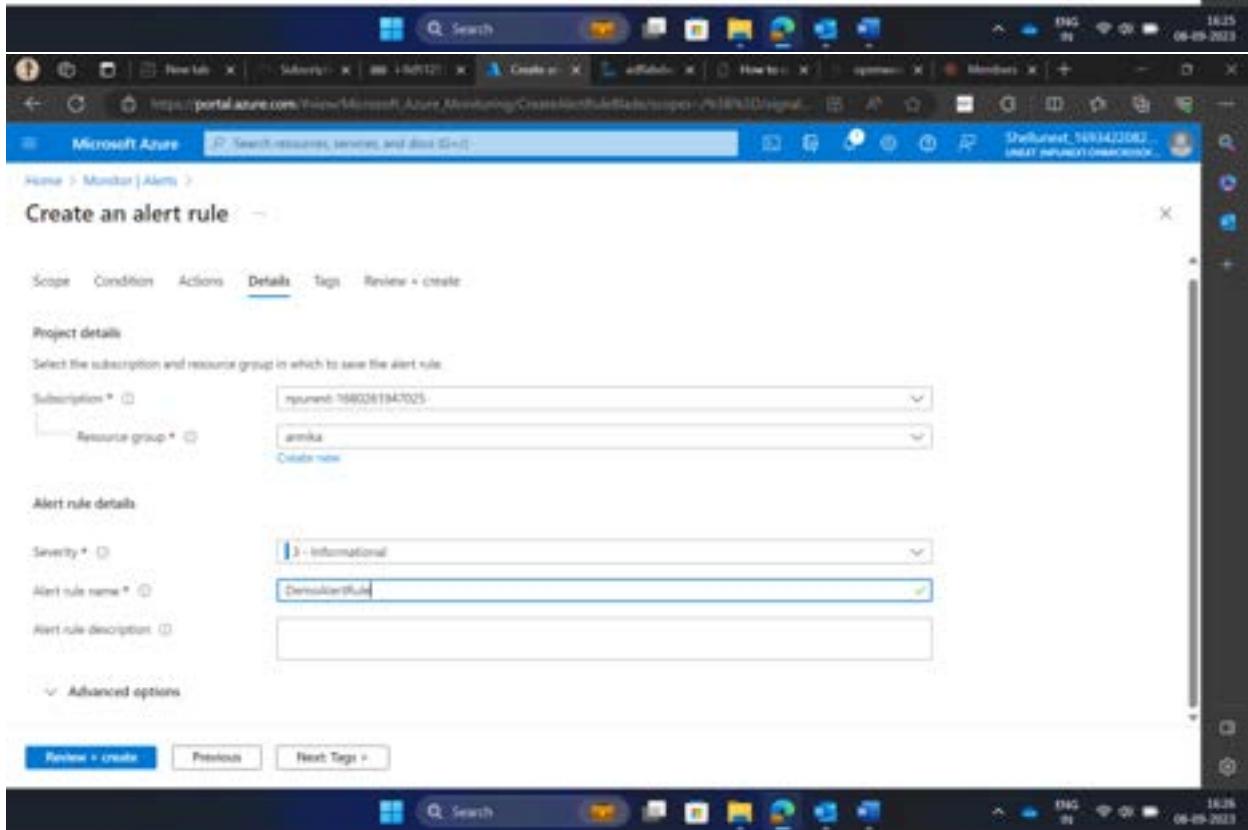
Severity \* (1) 3 - Informational

Alert rule name \* (1) DemoAlertRule

Alert rule description (0)

Advanced options

Review + create Previous Next: Tags >



The screenshot displays two windows side-by-side. The left window is a Microsoft Azure Data Studio instance titled "annikaprc (annikaprc/annikaprc) | Query editor (preview)". It shows a T-SQL script for creating a table named "Drilling\_Dataset". The right window is a Microsoft Teams "Meeting" interface.

**Query Editor (Left Window):**

```
1: create table Drilling_Dataset
2: (
3:     RigID int,
4:     DrillingDepth int,
5:     DrillingSpeed int,
6:     Location varchar(200),
7:     Created_Date Date
8: )
```

The results of the query execution are shown below:

Query succeeded: Affected rows: 8

**Meeting Chat (Right Window):**

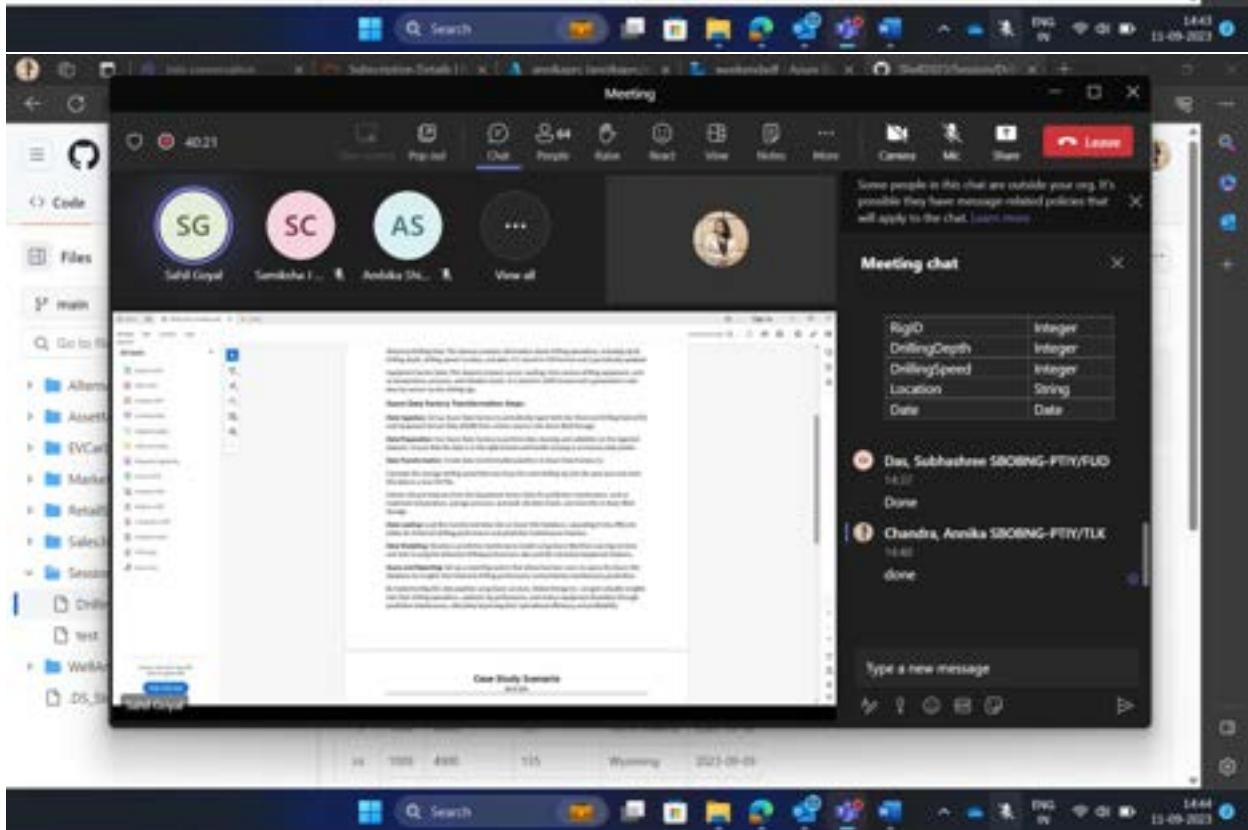
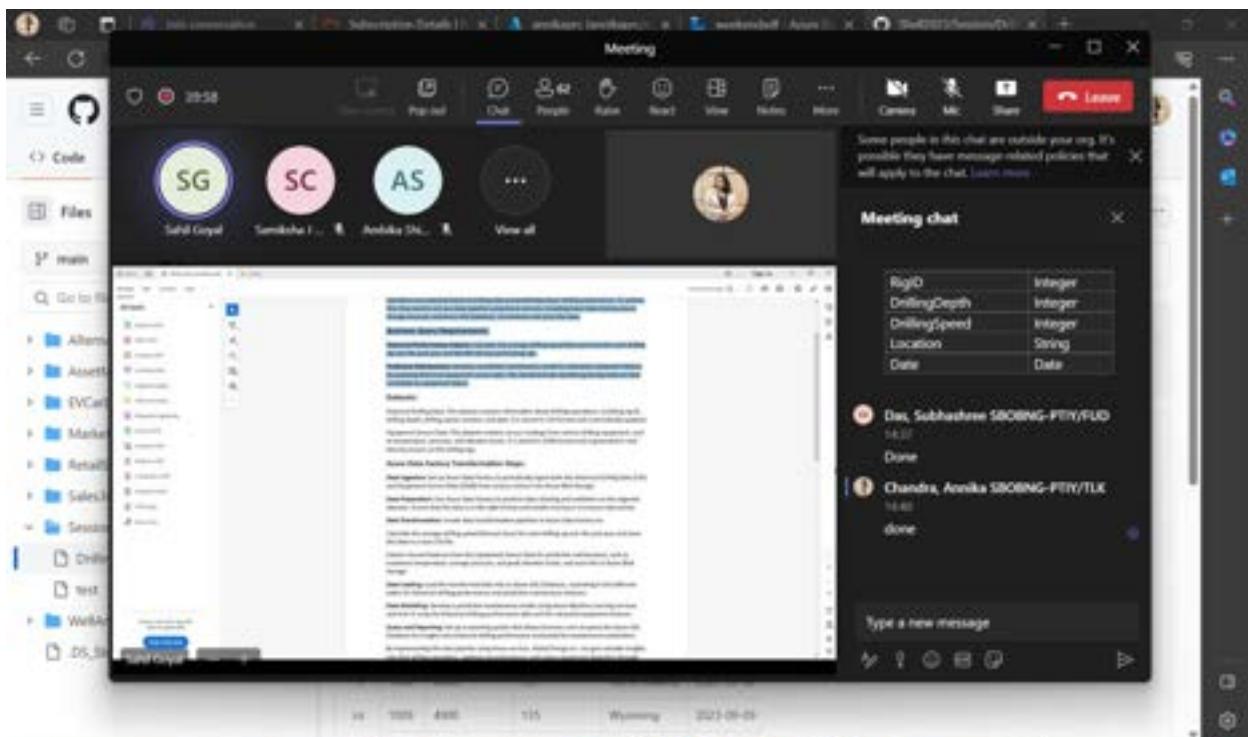
Meeting chat window showing a table definition and a list of messages:

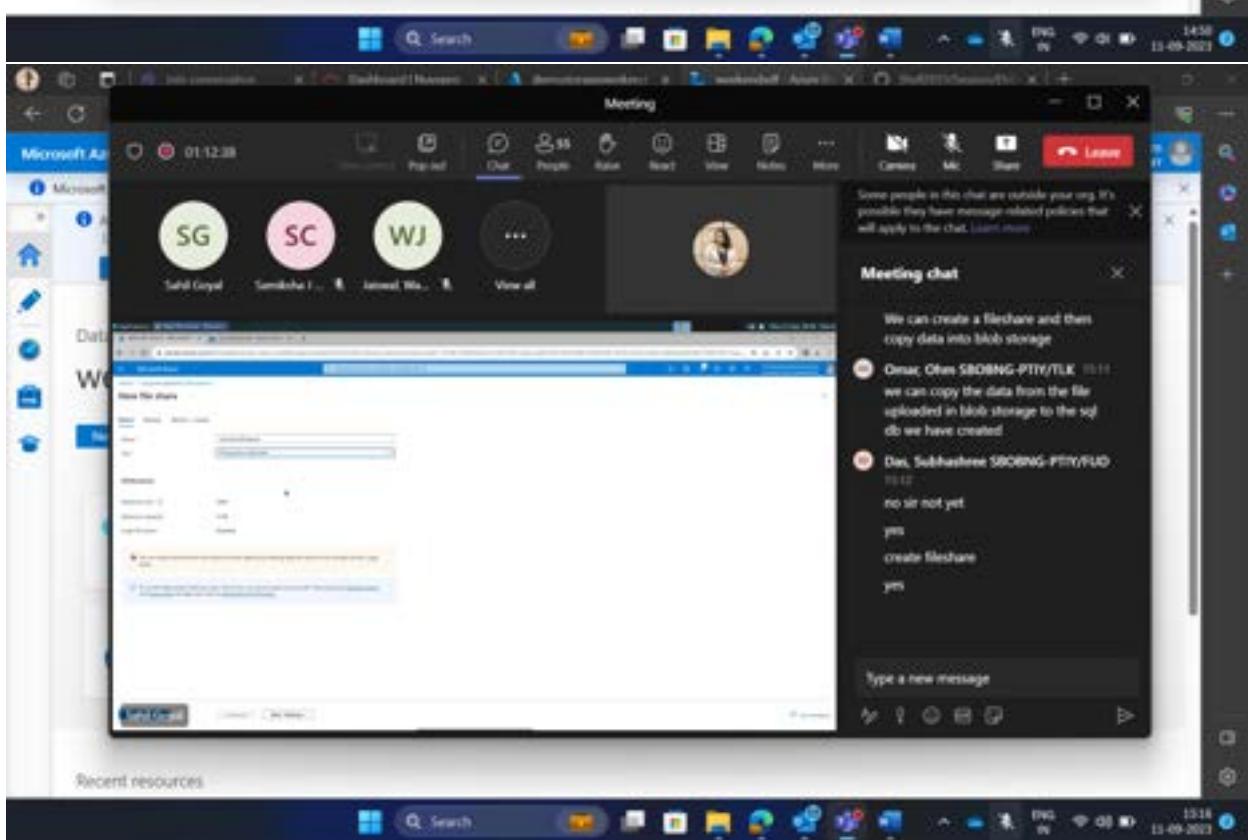
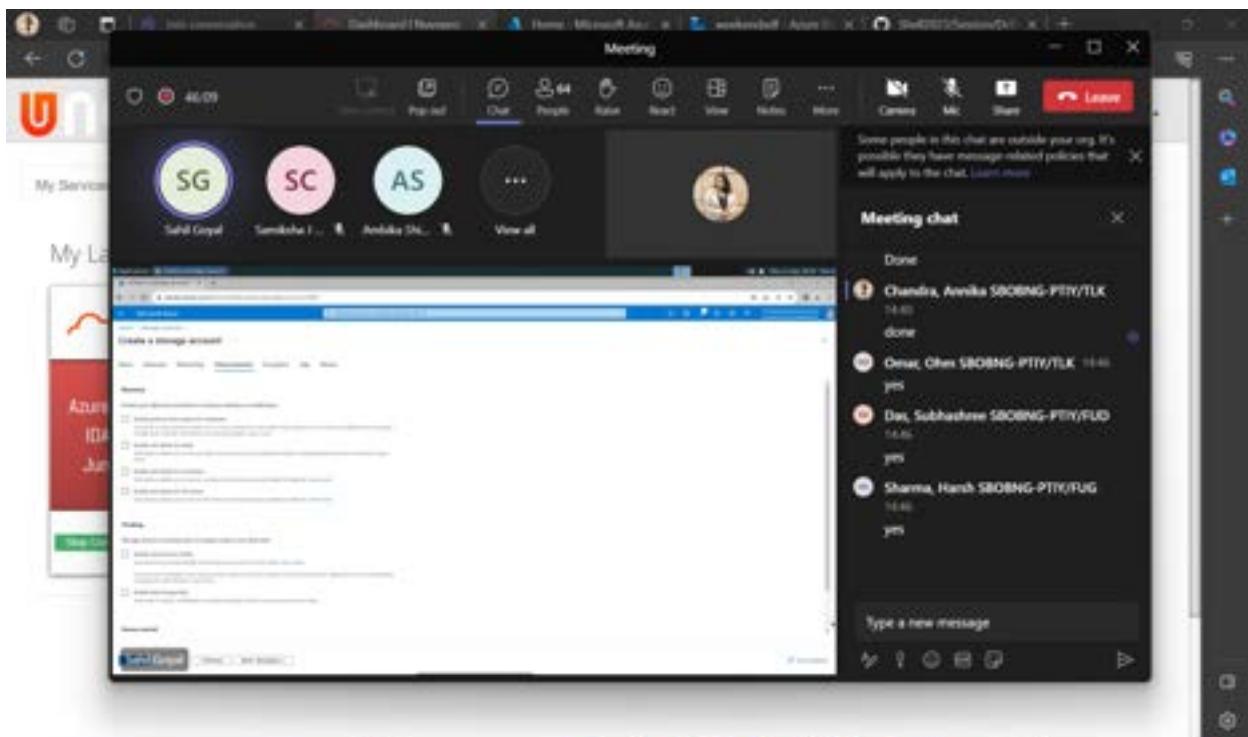
Column	Type
RigID	Integer
DrillingDepth	Integer
DrillingSpeed	Integer
Location	String
Date	Date

Messages in the chat:

- Das, Subhashree 580BNG-PRY/FUD 14:07 Done
- Chandra, Annika 580BNG-PRY/TLX 14:07 done

A "Type a new message" input field is visible at the bottom of the chat window.





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

Microsoft Azure | Search resources, services, and docs (en-US)

Home > demotoragewinkel\_109404098943 | Overview > demotoragewinkel | File shares > New file share

New file share

Deploying...

Basics Backup Review + create

**Basics**

File share name: demofile  
Access Tier: TransactionOptimized  
Protocol: SMB

**Backup**

Vault name: (new) vault-implpbk6  
Backup policy: (new) DailyPolicy-Implpbk6  
Policy details: Backup frequency: Daily at 7:30 PM UTC  
Retention of daily backup point: Retain backup taken every day at 7:30 PM for 30 days

**Next** < Previous Download a template for automation Give feedback

Microsoft Azure | Search resources, services, and docs (en-US)

Home > demotoragewinkel\_109404098943 | Overview > demotoragewinkel | File shares > New file share > demofile

SMB File share

Search Connect Upload Refresh Add directory Delete share Change tier Edit quota Give feedback

**Overview**

Diagnose and solve problems Access Control (IAM) Browse Operations Snapshots Backup

**New directory**

Name: mydir

Subscription: (new) Irausset-109404098943/23 Subscription ID: 3007675-1094-4ec4-9514-795a7a68788c

**Properties**

Size: Maximum capacity: 5 TB Used capacity: 0 B Size: Transaction optimized Feature status: Soft delete: Disabled Large file shares: Disabled

**Performance**

Maximum IO/s: 1000 Ingress rate: 60 MB/s Active Directory: Directory service: Not configured Domain:

**Microsoft Azure** | Search resources, services, and docs (Ctrl+F)

Home > demotoragewinkel (109402409843) | Overview > demotoragewinkel | File shares

**demofile** - SMB File share

Search

Overview

- Diagnose and solve problems
- Access Control (IAM)
- Browse
- Operations
- Snapshots
- Backup

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

```
$connectTestResult = Test-NetConnection -ComputerName demotoragewinkel.file.core.windows.net -Port 445
if ($connectTestResult.TcpTestSucceeded) {
    # Save the password so the drive will persist on reboot
    cmdkey /C "demofile\demofile" /user:"$username"
    /pass:"$password"
    /store:$demotoragewinkel
}
$script:demotoragewinkel = New-PSDrive -Name Z -PSProvider FileSystem -Root "\\demotoragewinkel.file.core.windows.net\demofile" -Persist
```

**Connect**

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

```
$connectTestResult = Test-NetConnection -ComputerName demotoragewinkel.file.core.windows.net -Port 445
if ($connectTestResult.TcpTestSucceeded) {
    # Save the password so the drive will persist on reboot
    cmdkey /C "demofile\demofile" /user:"$username"
    /pass:"$password"
    /store:$demotoragewinkel
}
$script:demotoragewinkel = New-PSDrive -Name Z -PSProvider FileSystem -Root "\\demotoragewinkel.file.core.windows.net\demofile" -Persist
```

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

01:32:42 Chat People Notes More Camera Me Share Leave

SG SC SD SG PS ... View all

Sahil Goyal Smitika I... Das, Subha... Godavari... M. Priya... Some people in this chat are outside your org. It's possible they have message-related policies that will apply to the chat. Learn more

Meeting chat

1.2 yes we have to automate

Das, Subha 10:00 sir but when the local directory is created for Nasshare there we will be saving the files manually only right? run this script in the terminal

Mudit, Shekhar 10:00 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

15:36 13-09-2023

Join conversation Subscription Details Connect - Microsoft Edge workcontext - Azure Data Factory [https://datafactory.azure.com/] 18:00 13-09-2023

Microsoft Azure Data Factory - datasets Search factory and documentation

1 Microsoft recently announced the public preview of Microsoft Fabric, a brand-new and exciting way to build cloud-first data analytics. Click here to get started with Fabric: Data Factory!

Preview experience: OFF

Factory Resources

Filter resources by name

Pipelines 0 Change Data Capture (preview) 0 Datasets 1 DelimitedText1

DelimitedText1

Properties

General Related

Name: DelimitedText1

Description:

Annotations

+ New

Connection Schema Parameters

Linked service: AzureBlobStorage1

File path: demo

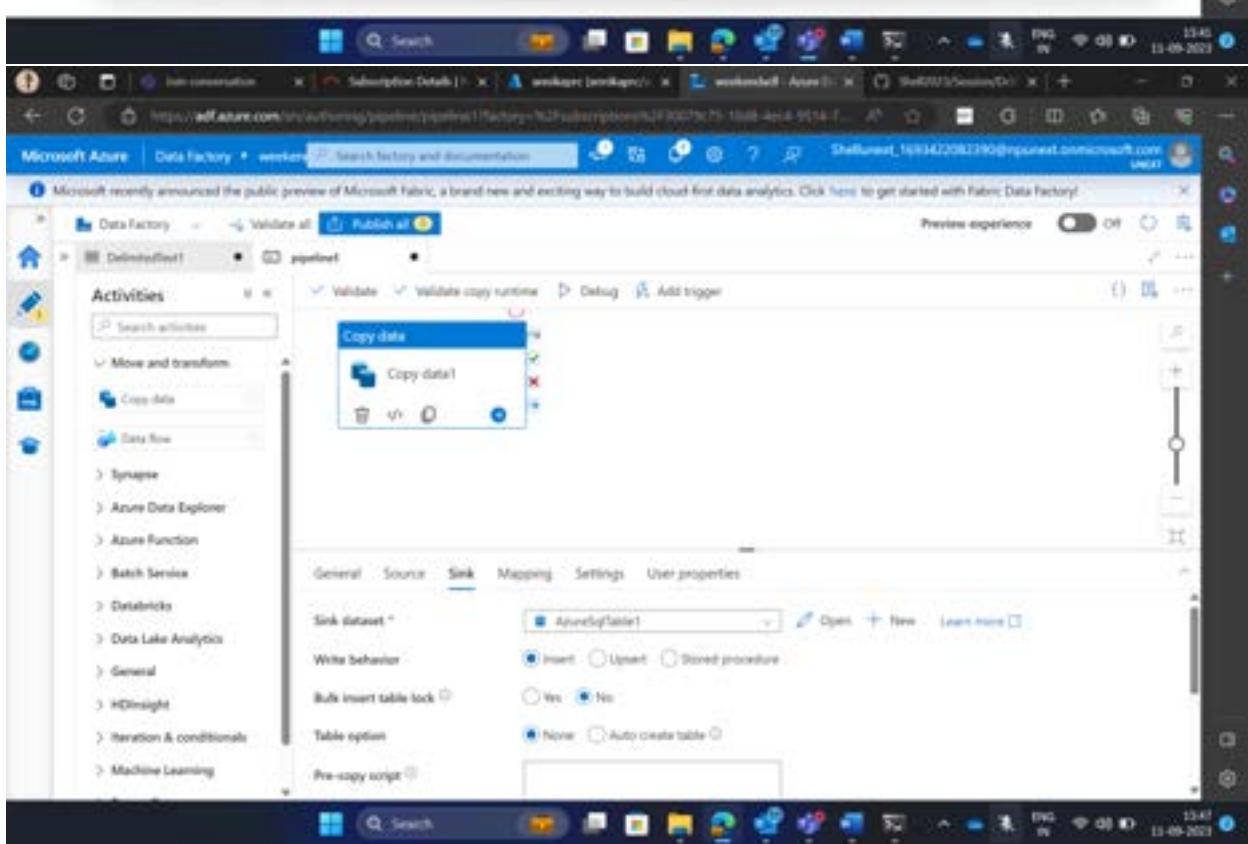
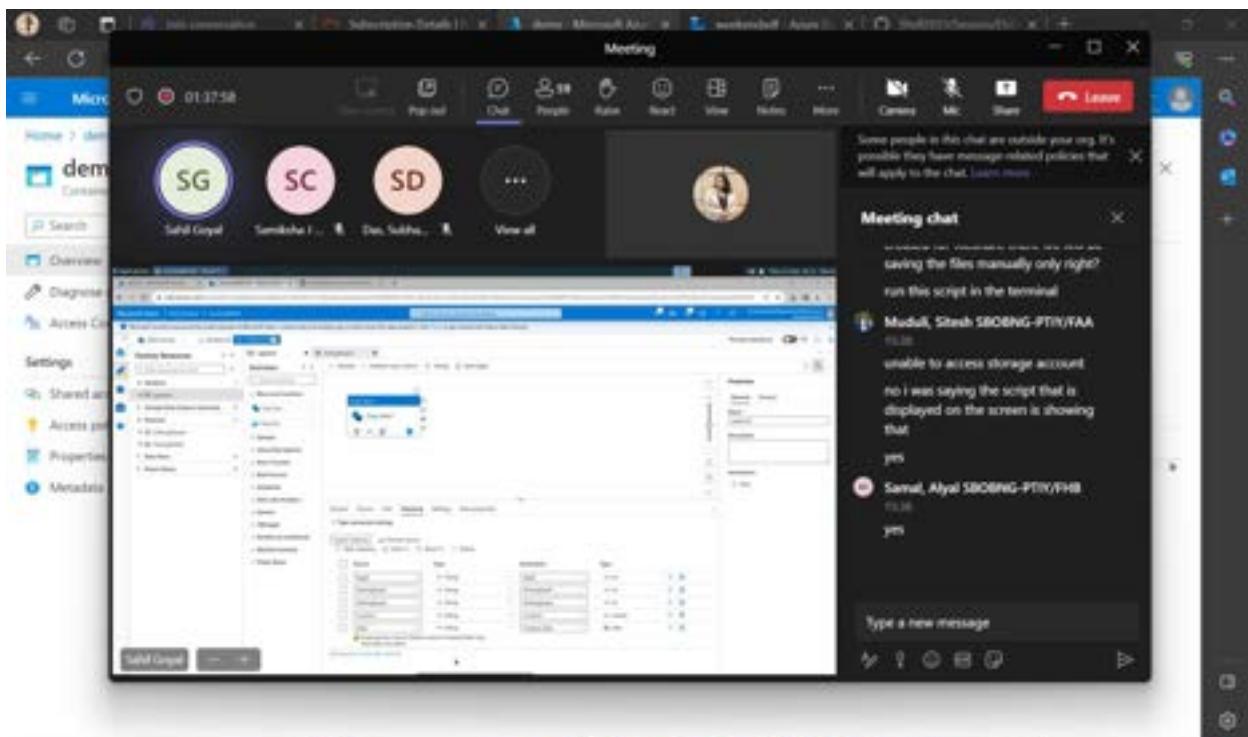
Compression type: Select...

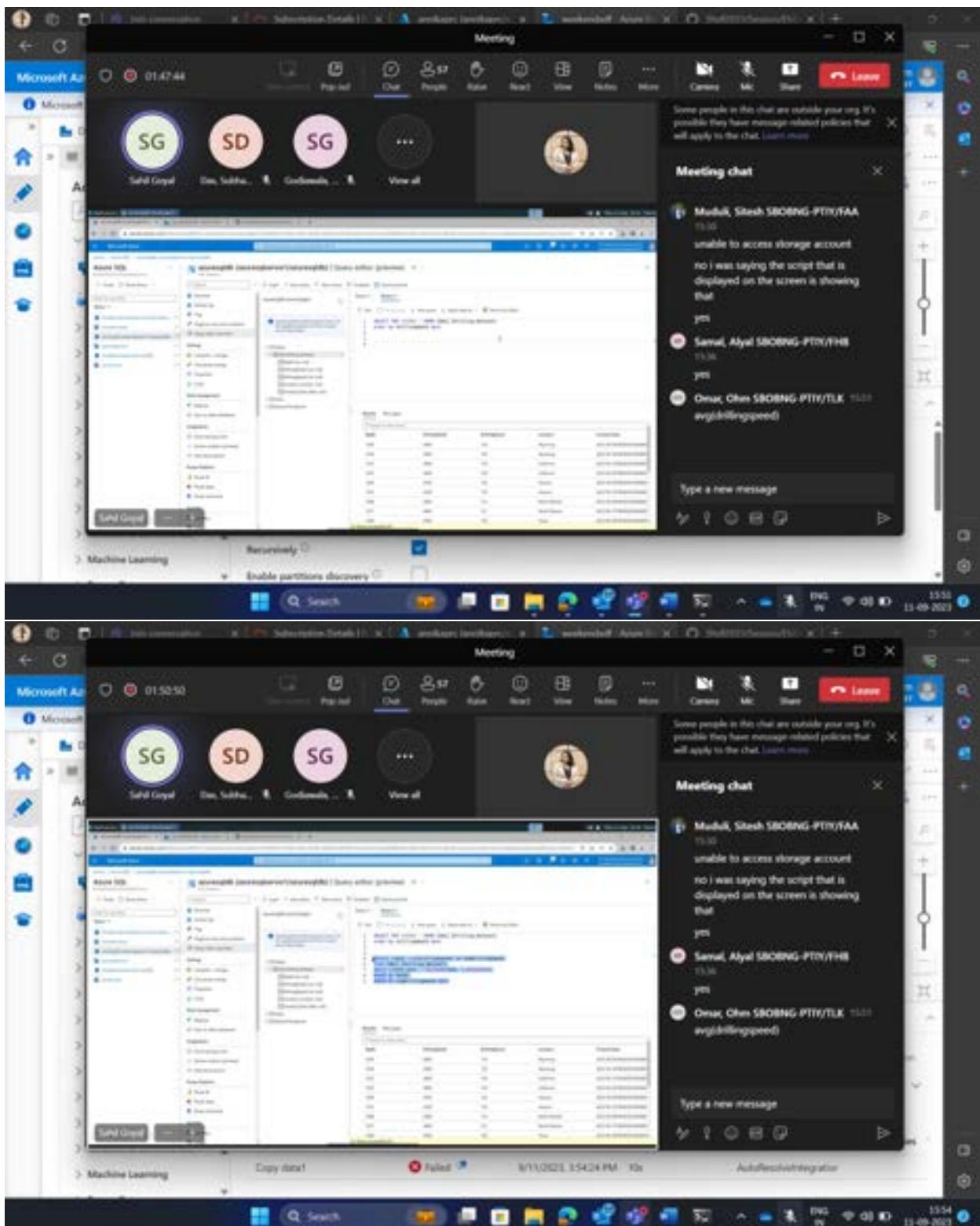
Column delimiter: Comma (,)

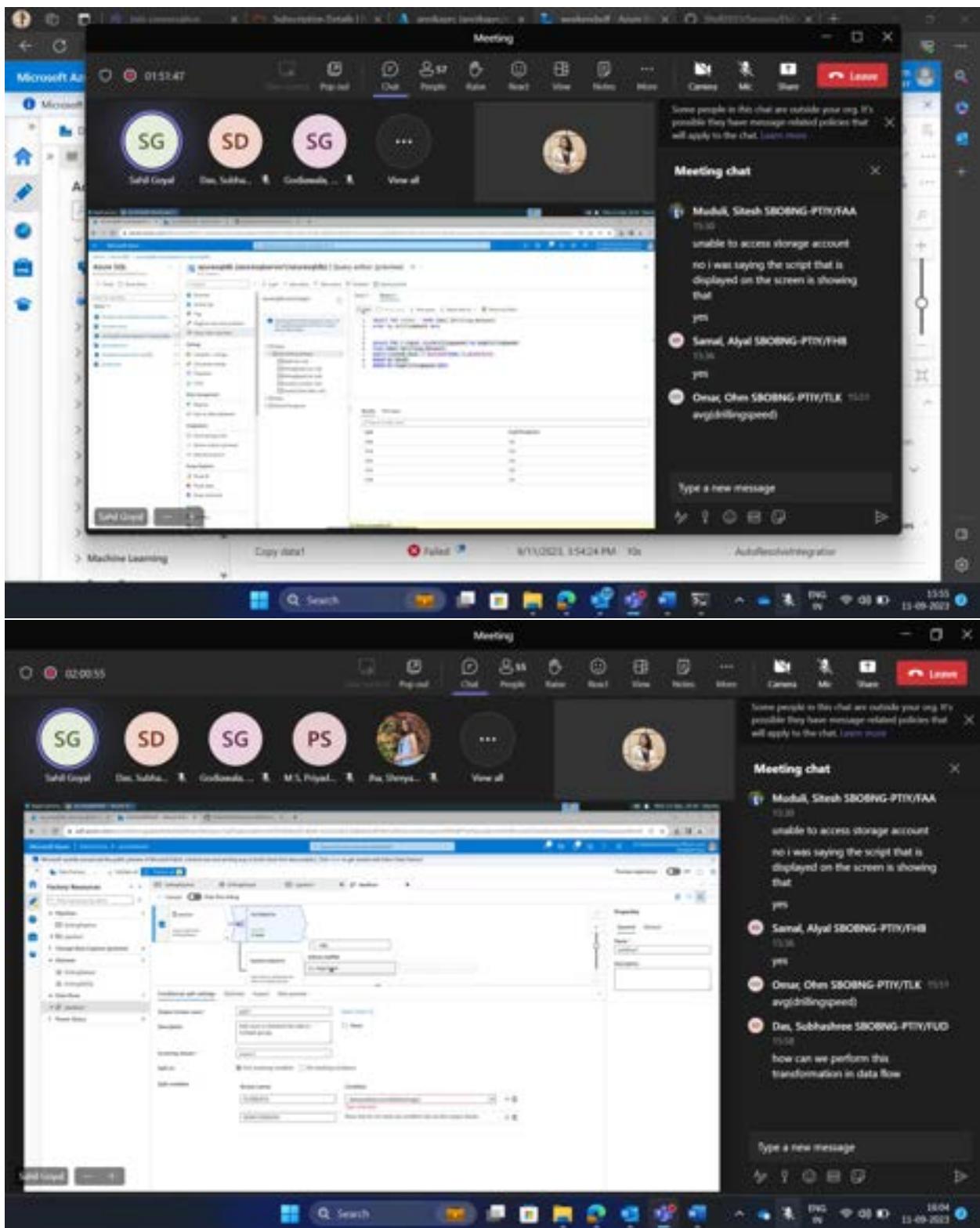
Row delimiter: Default (%0D or %0A)

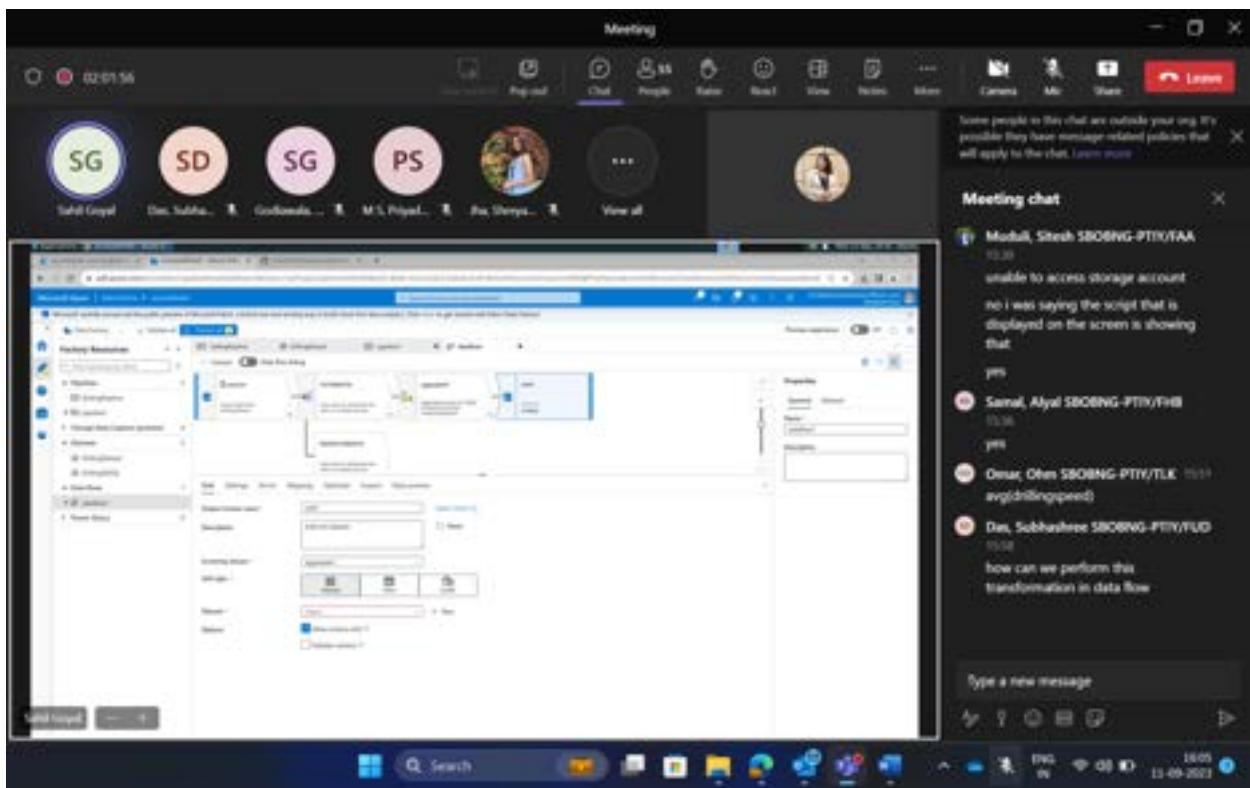
Encoding: Default(UTF-8)

15:40 13-09-2023









The screenshot shows two separate sessions of the Microsoft Azure Query editor (preview) running on a Windows 10 desktop. Both sessions are connected to the same Azure subscription and database context: 'labprc (annikaprc/labprc) | Query editor (preview)'.

**Session 1 (Top):**

- Query 1:** A simple SELECT query retrieves the top 1000 rows from the 'Drilling\_Dataset' table, ordered by 'drillingdepth'. The results show a single row with RigidID 1009, DrillingDepth 4800, DrillingSpeed 135, Location Wyoming, and Created Date 2023-09-11.
- Session 2 (Bottom):** A more complex query calculates the average drilling speed for each rig ID over the last year. It uses a subquery to find the date one year ago from today, then groups by rig ID and orders by the average speed in descending order.

**Session 1 Results:**

Rigid	DrillingDepth	DrillingSpeed	Location	Created Date
1009	4800	135	Wyoming	2023-09-11

**Session 2 Results:**

Rigid	AvgDrillingSpeed
No results	

## Azure Synapse Analytics 12.09.23 day 10

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

The screenshot shows two consecutive screenshots of the Microsoft Azure portal interface.

**Screenshot 1: Create Synapse workspace**

This screen is titled "Create Synapse workspace". It has a sub-header "Create a Synapse workspace to develop an enterprise analytics solution in just a few clicks." Below this, there's a "Project details" section where the user selects a subscription ("Subscription" dropdown set to "rgsynapse-7480261947025") and a resource group ("Resource group" dropdown set to "arnika"). There is also a "Managed resource group" section with a text input field "Enter managed resource group name".

Below the project details is a "Workspace details" section. It includes fields for "Workspace name" (set to "synapsesedwday10"), "Region" (set to "East US"), and "Select Data Lake Storage Gen2" (radio button selected for "From subscription").

At the bottom of this screen are three buttons: "Review + create", "Next: Preview", and "Next: Security".

**Screenshot 2: synapsesedwday10 - Synapse workspace**

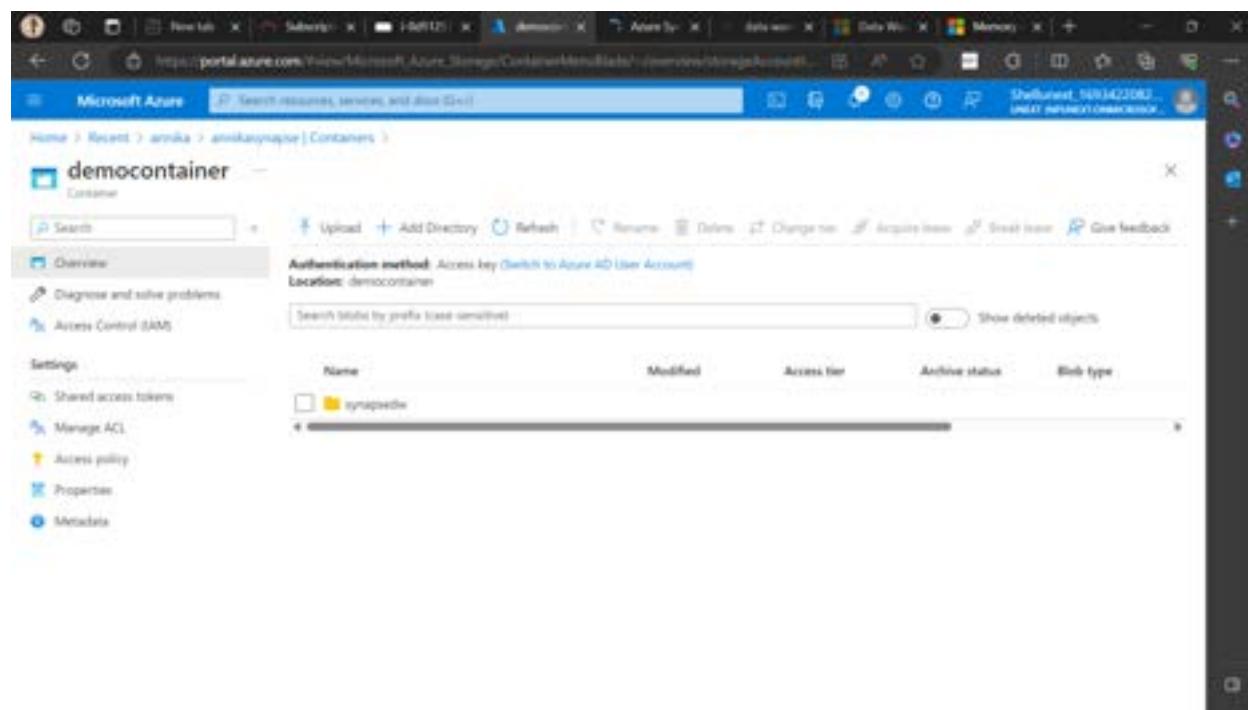
This screen shows the dashboard for the newly created workspace "synapsesedwday10". The left sidebar contains navigation links: Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Settings (Azure Active Directory, Properties, Locks), Analytics pools (SQL pools, Apache Spark pools, Data Explorer pools (preview)), Security, and Encryption.

The main content area features two cards: "Open Synapse Studio" (with a "Start building your fully-integrated analytics solution and unlock new insights." message) and "Read documentation" (with a "Learn how to be productive quickly. Explore concepts, tutorials, and samples." message).

Below these cards is a section titled "Analytics pools:" with a search bar "Search to filter items...". A table lists the provisioned pools:

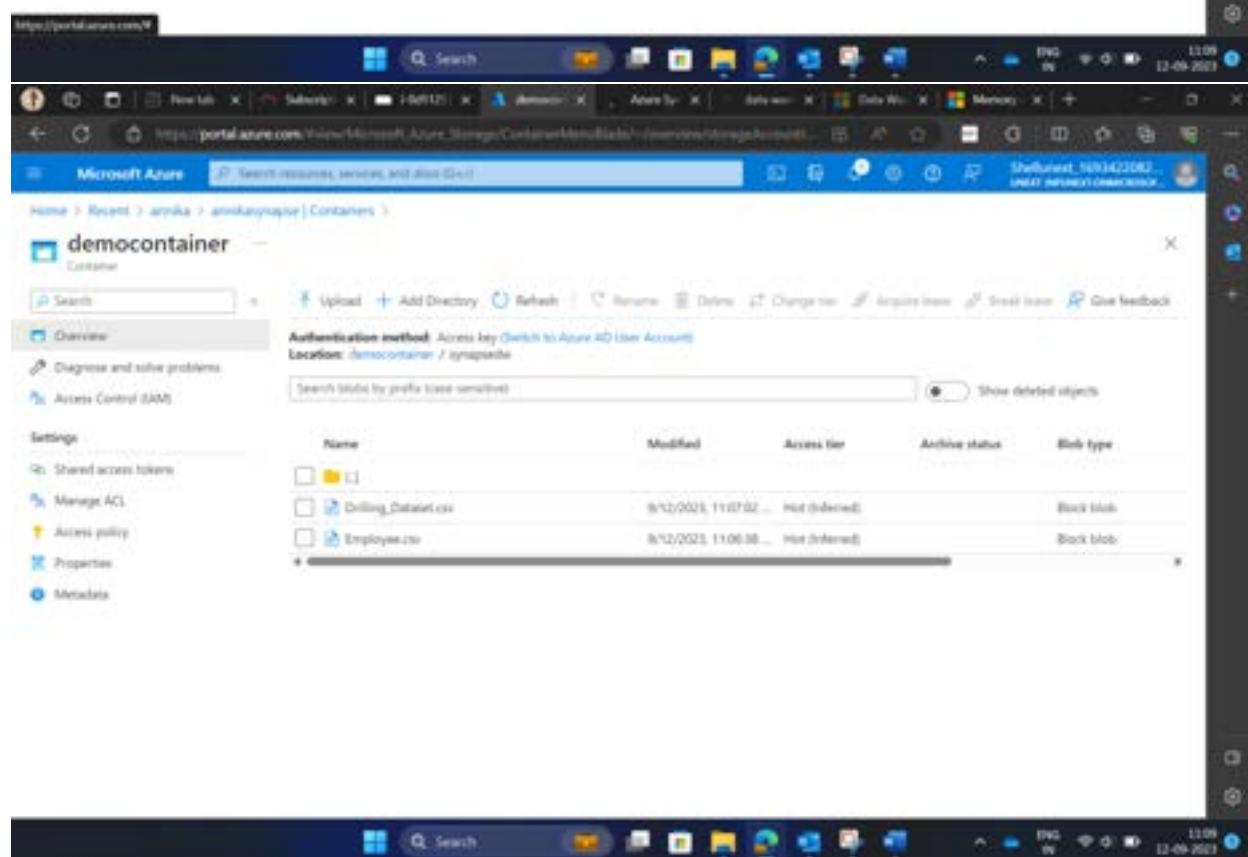
Name	Type	Size
Built-in	Serverless	Auto
Apache Spark pools	No pools provisioned	
Data Explorer pools	No pools provisioned	

Serverless pool is created for us automatically



The screenshot shows the Microsoft Azure Storage Container Overview page for a container named 'democontainer'. The container has one blob named 'synapse'. The blob's properties are as follows:

Name	Modified	Access tier	Archive status	Block type
synapse	8/12/2023, 11:07:02 ...	Hot (selected)		

The screenshot shows the Microsoft Azure Storage Container Overview page for a container named 'democontainer'. The container now contains three blobs: 'synapse', 'Drilling\_Database.csv', and 'Employee.csv'. The blobs' properties are as follows:

Name	Modified	Access tier	Archive status	Block type
synapse	8/12/2023, 11:07:02 ...	Hot (selected)		
Drilling_Database.csv	8/12/2023, 11:06:38 ...	Hot (selected)		Block blob
Employee.csv	8/12/2023, 11:06:38 ...	Hot (selected)		Block blob

The screenshot shows the Microsoft Azure Synapse Analytics Data workspace interface. On the left, there's a navigation pane with icons for Home, Storage, Databricks, Synapse, Data Wiz, Data Flow, and Monitor. The main area is titled 'Data' and has tabs for 'Workspace' and 'Linked'. Under 'Linked', it shows 'Azure Data Lake Storage Gen2' and 'synapsesday10 (Primary - amrik...)' which contains 'amrikasystem (Primary)' and 'democontainer'. A context menu is open over a dataset named 'Empty'. The menu includes options like 'New SQL script', 'New notebook', 'New data flow', 'New integration dataset', 'More...', 'Manage access...', 'Rename...', 'Download', 'Delete', and 'Properties...'. A tooltip says 'Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.' Below the menu, a SQL script editor window is visible with the following T-SQL code:

```
1 -- This is auto-generated code.
2
3 SELECT
4     [TDF_100]
5 FROM
6     OPENROWSET(
7         BULK 'https://amrikasynapse.dfs.core.windows.net/democontainer/synapsesday10/Employee.csv',
8         FORMAT = 'CSV',
9         PARSE_TYPE = 'T-SQL'
10    ) AS [result]
```

The results pane shows a table with four columns: C1, C2, C3, and C4. The data is as follows:

C1	C2	C3	C4
Ed	Diane	Edge	Expiry
1	Aisha	17	10000
2	Romi	12	15023
3	Aishan	19	13227

At the bottom of the results pane, a message says '00:00:18 Query executed successfully.'

If row headings are not present in csv file

The screenshot displays two separate sessions within the Microsoft Azure Synapse Analytics workspace, both titled "democontainer" and "SQL script 1".

**Session 1 (Top):**

```
OPENROWSET(
    BULK 'https://anindikaysynapsedfs.core.windows.net/democontainer/synapsedw/Employees.csv',
    FORMAT = 'CSV',
    PARSE_VERSION = '2.8'
) WITH (
    Id INT 1,
    Ename VARCHAR(20) 2,
    Eage INT 3,
    Esalary VARCHAR(10) 4
) AS [Rows]
```

**Session 2 (Bottom):**

```
PARSE_VERSION = '2.8'
FIRSTROW + 2
) WITH (
    Id VARCHAR(20) 1,
    Ename VARCHAR(20) 2,
    Eage VARCHAR(20) 3,
    Esalary VARCHAR(20) 4
) AS [result]
```

Both sessions show a results grid with columns: Id, Ename, Eage, and Esalary. The data for Session 1 includes rows for Id 1 (Aloha, 17, 1000), Id 2 (Rani, 12, 1522), and Id 3 (Aishan, 19, 1322). The data for Session 2 includes rows for Id 1 (Aloha, 17, 1000), Id 2 (Rani, 12, 1522), and Id 3 (Aishan, 19, 1322).

The image shows two side-by-side screenshots of the Microsoft Azure Synapse Analytics workspace interface. Both screenshots display a SQL script editor window with a query being run against a database named 'master'.

**Top Screenshot:**

- Script Content:**

```
18 SELECT
19     TOP 10 *
20 FROM
21     OPENROWSET(
22         BULK 'https://pandemicdatalake.blob.core.windows.net/public/certified/covid-19/covid_cases/latest/covid_cases.csv',
23         FORMAT = 'CSV',
24         PARSE_VERSION = '1.8',
25         firstrow = 2
26     ) WITH
27     (
28         date_rep DATE,
29         [year] INT,
30         cases INT,
31         geo_id CHAR(3)
32     ) AS [result]
```
- Results:** A table showing two rows of data from the COVID-19 dataset.

date_rep	year	cases	geo_id
2020-12-14T00:00:00Z	2020	146	AF
2020-12-11T00:00:00Z	2020	298	AF

**Bottom Screenshot:**

- Script Content:** The same SQL script as the top screenshot.
- Results:** A table showing the same two rows of data from the COVID-19 dataset.

To create external source link

Microsoft Azure | Synapse Analytics | https:// Search | [Accept](#) | [Reject](#) | [More options](#)

Synapse live | Validate all | Publish all |

democontainer | SQL script | Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Run Undo Publish Query plan Connect to Built-in use database master

```
25     FirstName nvarchar(2);
26   } WITH
27   (
28     date_rept DATE 1,
29     [year] int 4,
30     cases int 5,
31     geo_id char(5) 8
32   ) AS [result]
33
34   create external data source covid
35   with (location = 'https://pandemicdata.blob.core.windows.net/public/curated/covid-19/covid\_cases/latest/covid\_cases' );
36
```

Results Messages

View Table Chart Export results ↗

Search

date_rept	year	cases	geo_id
2020-12-14T00:00:0000000	2020	746	AF
2020-12-13T00:00:0000000	2020	298	AF

00:00:02|Query executed successfully|

Microsoft Azure | Synapse Analytics | https:// Search | [Accept](#) | [Reject](#) | [More options](#)

Synapse live | Validate all | Publish all |

democontainer | SQL script | Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Run Undo Publish Query plan Connect to democnspool use database democnspool

```
27   )
28   date_rept DATE 1,
29   [year] int 4,
30   cases int 5,
31   geo_id char(5) 8
32 ) AS [result]
33
34 create database democnspool
35
36 creates external data source covid
37 with (location = 'https://pandemicdata.blob.core.windows.net/public/curated/covid-19/covid\_cases/latest/covid\_cases' );
38
```

Results Messages

No results to show  
Your query printed no displayable results

00:00:01|Query executed successfully|

Microsoft Azure | Synapse Analytics | [Search](#)

We use optional cookies to provide a better experience. Learn more [Accept](#) [Reject](#) [More options](#)

Synapse live | Validate all | Publish all

democontainer SQL script

Run Undo Publish Query plan Connect to Built-in use database demodbsqlpool

```

27   |
28   date_rep DATE 1,
29   [year] INT 4,
30   cases INT 5,
31   geo_id CHAR(5) 6
32 ) AS [result]
33
34 CREATE DATABASE demodbsqlpool
35
36 CREATE EXTERNAL DATA SOURCE covid
37 WITH (LOCATION = 'https://pandemicdata.blob.core.windows.net/public/covid-19/covid_cases/latest/covid_cases')
38

```

Results Messages

No results to show.  
Your query yielded no displayable results.

00:00:01 Query executed successfully.

Microsoft Azure | Synapse Analytics | [Search](#)

We use optional cookies to provide a better experience. Learn more [Accept](#) [Reject](#) [More options](#)

Synapse live | Validate all | Publish all

democontainer SQL script

Run Undo Publish Query plan Connect to Built-in use database demodbsqlpool

```

33 ) AS rows
34
35
36 SELECT TOP 10 *
37 FROM openrowset(
38   BULK 'https://pandemicdata.blob.core.windows.net/public/covid-19/covid_cases/latest/covid_cases.parquet',
39   FORMAT = 'PARQUET') AS rows
40

```

Results Messages

View [Table](#) [Chart](#) [Export results](#)

date_rep	day	month	year	cases	deaths	countries_and_t...	geo_id	country territor...
2020-12-14T00...	14	12	2020	746	6	Afghanistan	AF	AFG
2020-12-13T00...	13	12	2020	298	9	Afghanistan	AF	AFG
2020-12-12T00...	12	12	2020	113	11	Afghanistan	AF	AFG
2020-12-11T00...	11	12	2020	63	10	Afghanistan	AF	AFG

00:00:05 Query executed successfully.

Microsoft Azure | Synapse Analytics | https://Search

We use optional cookies to provide a better experience. Learn more [?]

Accept Reject More options

Synapse live Validate all Publish [ ]

democontainer SQL script

Run Undo Publish Query plan Connect to Built-in use database democloudpool

```
41 select top 20 *
42 from
43 openrowset(
44 bulk 'https://raw.githubusercontent.com/wdowd/covid-19-repo/cases/latest/covid_cases.csv',
45 format = 'csv',
46 Fieldterminator = ',',
47 Fieldquote = '\"'
48 ) with (doc nvarchar(max)) as doc
```

Results Messages

View Table Chart Export results

Search

doc

[redacted]

000002 Query executed successfully.

Microsoft Azure | Synapse Analytics | https://Search

We use optional cookies to provide a better experience. Learn more [?]

Accept Reject More options

Synapse live Validate all Publish [ ]

Develop

Filter resources by name

SQL scripts Metadata demo

Run Undo Publish Query plan Connect to Built-in

Properties

General Related (0)

Name

Description

Type sql script

Size 0 bytes

Results settings per query

First 5000 rows (default)

All rows

<https://portal.azure.com/#@annikasynapse/resource/subscriptions/0007b75-10d8-4e14-9f14-75>

Microsoft Azure | Shared access signature

Storage account

Search  Give feedback

A shared access signature (SAS) is a URL 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 (or whom you wish to delegate access to certain storage account resources). By distributing a shared access signature (URL) 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

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

blobInventory

https://portal.azure.com/#@annikasynapse/resource/subscriptions/0007b75-10d8-4e14-9f14-75

Search

Start  End

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 datetime: Start  End

Allowed IP addresses:

Allowed protocols:  HTTPS only  HTTPS and HTTP

Preferred routing tier:  Basic (default)  Microsoft network routing  Internet routing

Some routing options are disabled because the endpoints are not published.

Signing key:

Generate SAS and connection string

Search

12:08 13/08/2023

Microsoft Azure | Shared access signature

Storage account

Search  Give feedback

blobInventory

https://portal.azure.com/#@annikasynapse/resource/subscriptions/0007b75-10d8-4e14-9f14-75

Search

Start  End

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 datetime: Start  End

Allowed IP addresses:

Allowed protocols:  HTTPS only  HTTPS and HTTP

Preferred routing tier:  Basic (default)  Microsoft network routing  Internet routing

Some routing options are disabled because the endpoints are not published.

Signing key:

Generate SAS and connection string

Search

12:08 13/08/2023

Microsoft Azure | Synapse Analytics | https://synapseanalytics.azuredatabricks.net/ | ShellUser\_1003422082390@ipuser@microsoft.com | user

We use optional cookies to provide a better experience. Learn more | Accept | Reject | More options

Synapse live | Validate all | Publish all |

demodatastore | Metadata store

Run Undo Publish Query plan Connect to Built-in use database demodatastore

```
--Creating master key  
create master KEY  
  
--Create database scoped credential  
CREATE DATABASE SCOPED CREDENTIAL [sqladminemand]  
WITH SECRET = 'SHARED-ACCESS-SIGNATURE'  
SECRET = 'T7v+19z2-11-R0BaH0fz8OrtwcokspnLkijpukie+2023-09-12T18:11:37Zct>2023-09-12T00:11:37Z&pr=https://1g4u50Py0Rt5Aa0Dg8uAPC'
```

Results Messages

No results to show  
Your query yielded no displayable results

0000001 Query executed successfully.

Microsoft Azure | Search resources, services, and more (Q+)

democontainer | Properties

Container

Search Refresh Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

NAME: democontainer

URL: https://amikasyntapse.D4B.UK.westus2.azuredatabricks.net/demodatastore

LAST MODIFIED: 9/12/2023, 11:01:01 AM

ETAG: 0x00000000000000000000000000000000

LEASE STATUS: Unlocked

LEASE STATE: Available

LEASE DURATION:

```
--creating master key
CREATE MASTER KEY

--create database scoped credentials
CREATE DATABASE SCOPED CREDENTIAL[sqolondemand]
WITH IDENTITY = 'SHARED ACCESS SIGNATURE',
SECRET = '7sv+282-11-816aaH0f&ur7wcolspnrdLcLcyp4d8e=2023-09-12T18:11:37Z&t=2023-09-12T08:11:37Z&p=https://sqlofyml8rjAAQ0rg8APD'

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

```

```
--creating master key
CREATE MASTER KEY

--create database scoped credentials
CREATE DATABASE SCOPED CREDENTIAL[sqolondemand]
WITH IDENTITY = 'SHARED ACCESS SIGNATURE',
SECRET = '7sv+282-11-816aaH0f&ur7wcolspnrdLcLcyp4d8e=2023-09-12T18:11:37Z&t=2023-09-12T08:11:37Z&p=https://sqlofyml8rjAAQ0rg8APD'

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

--create external file format
CREATE EXTERNAL FILE FORMAT QuotedCsvWithHeaderFormat
WITH(
FORMAT_TYPE = DELIMITEDTEXT,
FORMAT_OPTIONS(FIELD_TERMINATOR = ',', FIRST_ROW = 4))

```

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. In the center, there is a query editor window titled 'democontainer' under the 'Metadata' tab. The code in the editor is:

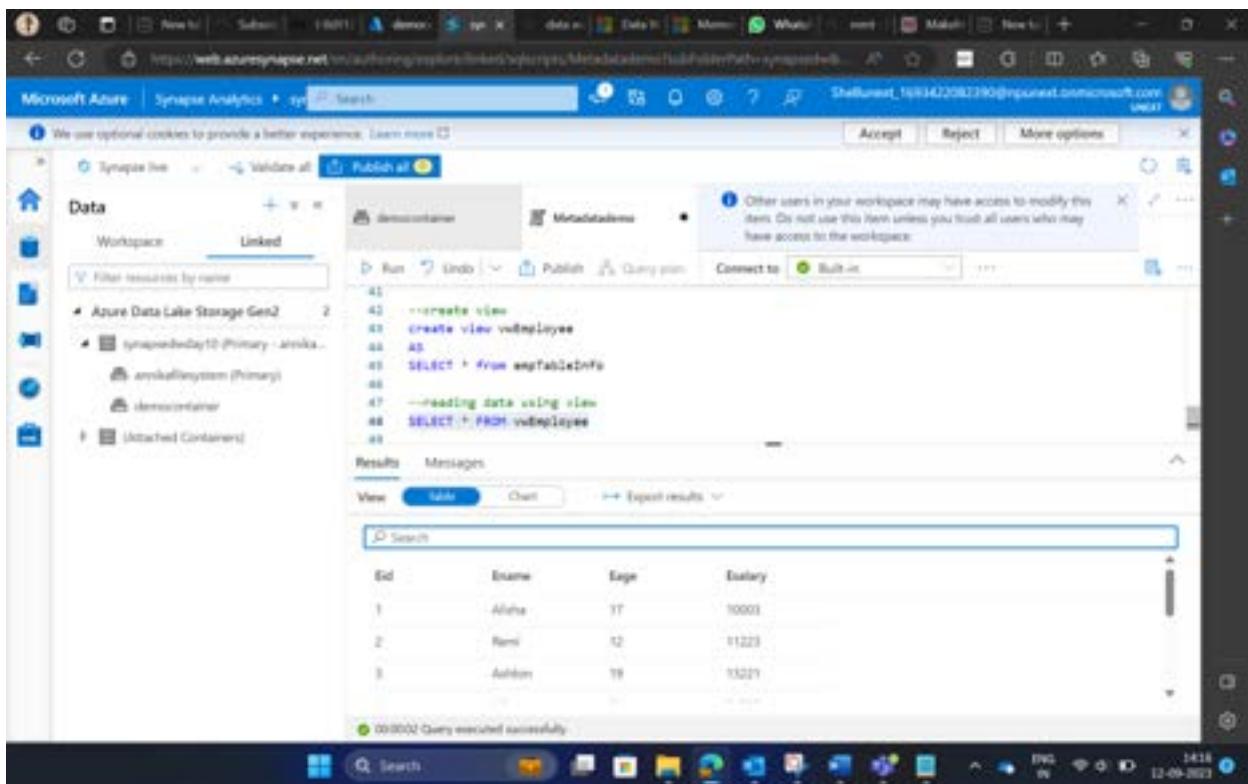
```
1 --Creating external table
2 CREATE EXTERNAL TABLE empTableInfo
3 (
4     Eid int,
5     Ename varchar(10),
6     Eage int,
7     Esalary int
8 ) WITH
9 (
10     LOCATION = 'synapssed10/Employee.csv',
11     DATA_SOURCE = sqldemandeds10,
12     FILE_FORMAT = QuotedCSVwithHeaderFormat
13 )
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
```

At the bottom of the editor, the status bar shows '00:00:01 Query executed successfully.'

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. On the left, there is a sidebar titled 'Data' with a 'Linked' tab selected. Under 'Linked', it lists 'Azure Data Lake Storage Gen2' and 'synapssed10 (Primary - armika...)' which contains 'armikaflysystem (Primary)' and 'democontainer'. The main area shows the same query editor as the first screenshot, but now displaying results. The results table is:

Eid	Ename	Eage	Esalary
1	Alisa	17	10000
2	Renu	12	11220
3	Aishwarya	19	11030

At the bottom of the editor, the status bar shows '00:00:01 Query executed successfully.'



```
--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=rw&lacupyx&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
```

```
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
```

The screenshot shows a Microsoft Azure Synapse Analytics notebook titled "Notebook 1". The code cell contains the following PySpark code:

```
1 Myspark
2 df = spark.read.load('abfss://democontainer@synapse-dfs.core.windows.net/synapseeu/Employee.csv', 'format'='csv')
3 # If header exists uncomment line below
4 # header=True
5
6 display(df.limit(10))
```

The screenshot shows the same Microsoft Azure Synapse Analytics notebook after the code has been run. The status bar indicates "Ready". The results section displays the following table:

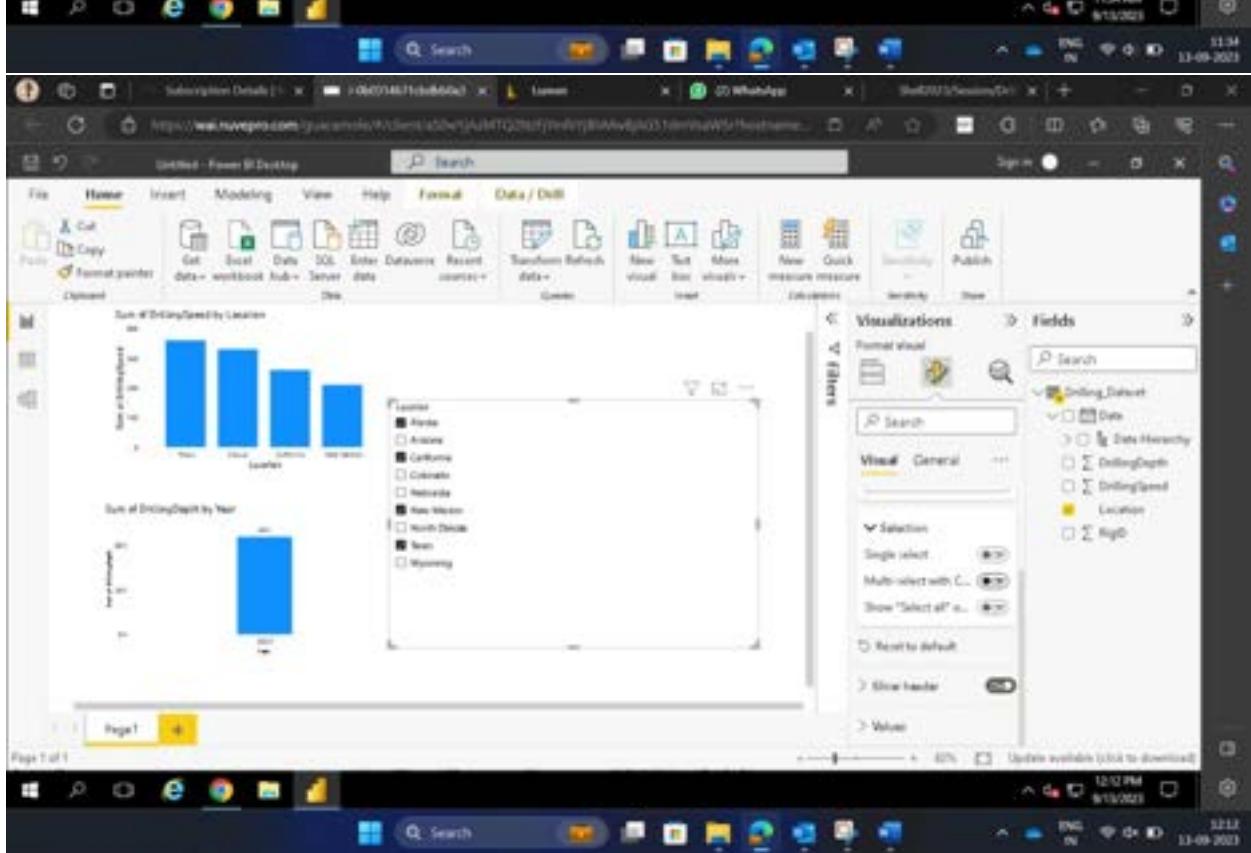
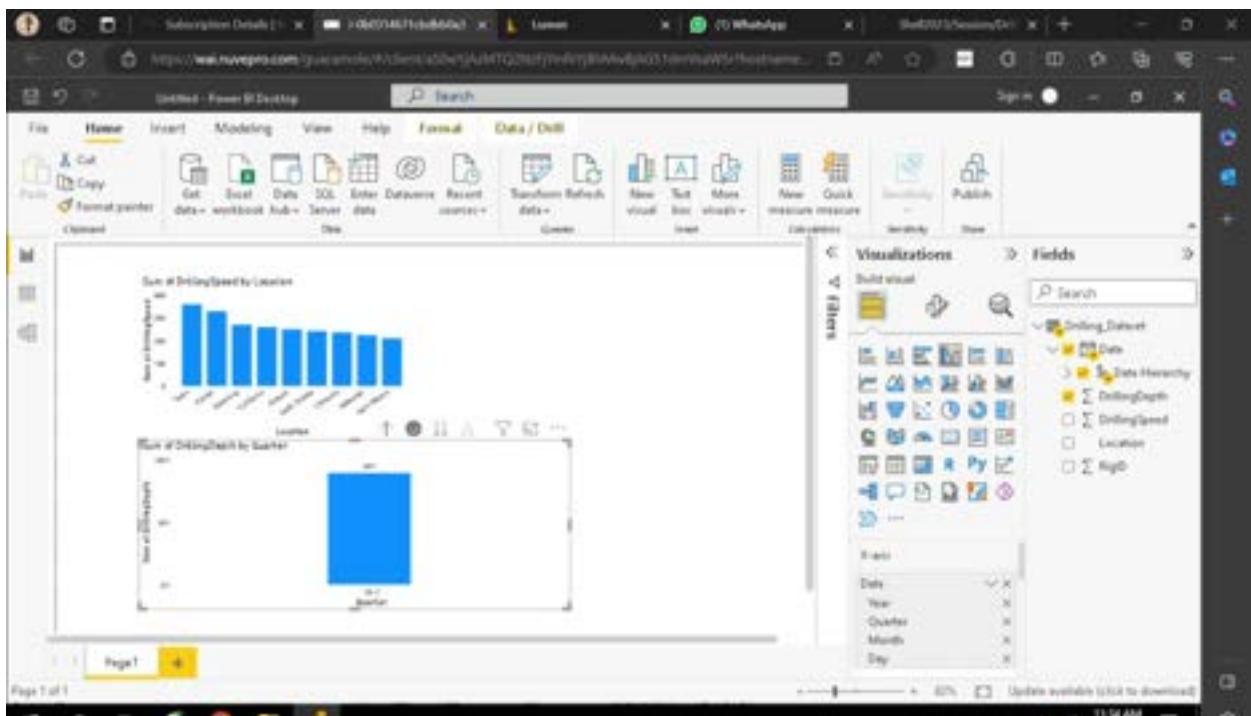
Id	Ename	Esal	Esalary
1	Aisha	17	10000
2	Renu	12	11000
3	Adrian	19	13221
4	Mia	21	11132
5	Ashley	21	14023

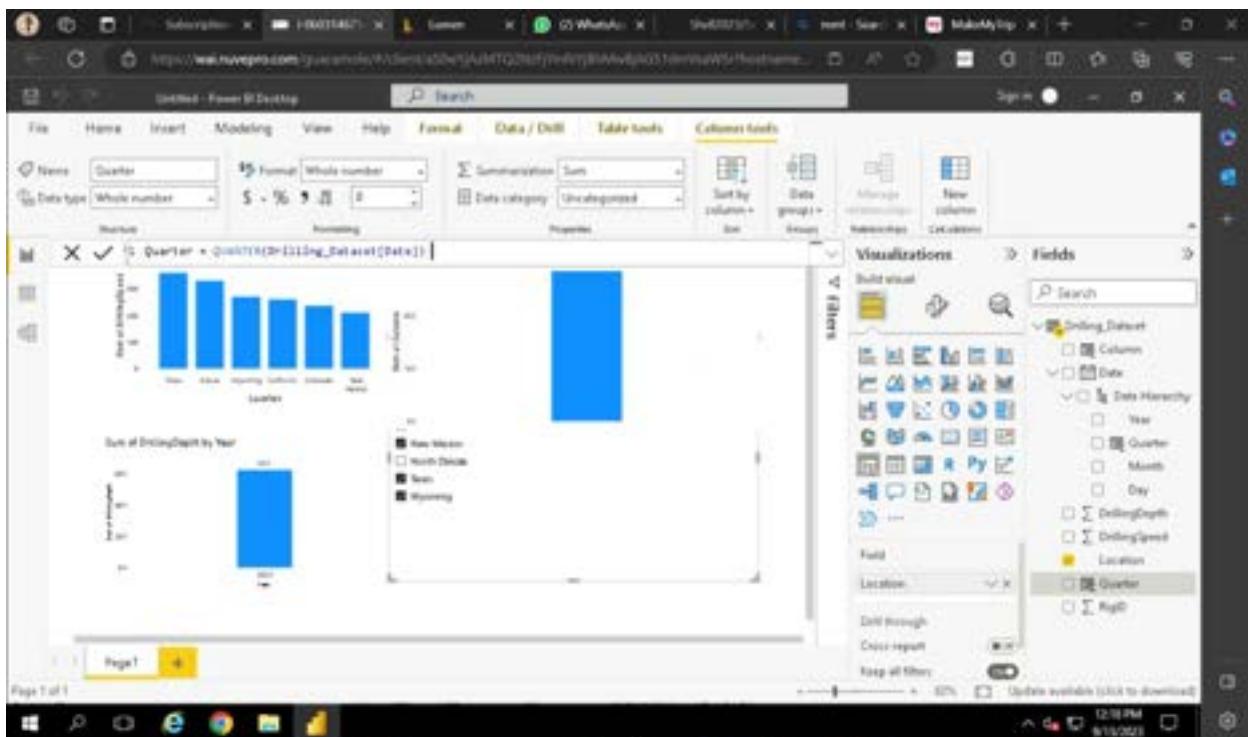
## Power bi

The screenshot shows two windows of the Power BI Desktop application.

**Top Window:** This window displays the 'Drilling\_Dataset.csv' import dialog. The 'File Origin' dropdown is set to '1252 - Western European (Windows)'. The 'Delimiter' dropdown is set to 'Comma'. The 'Data Type Detection' dropdown is set to 'Based on first 200 rows'. The main area shows a preview of the data with columns: RigID, DrillingDepth, DrillingSpeed, Location, Date. Below the preview is a 'Select Table Using Examples' button, followed by 'Load', 'Transform Data', and 'Cancel' buttons. A progress bar at the bottom indicates 'Page 1 of 1'.

**Bottom Window:** This window shows the 'Drilling\_Dataset' dataset properties. The 'Name' field is set to 'RigID'. The 'Description' field is empty. The 'Fields' pane on the right lists the columns: Drilling\_Dataset, Date, DrillingDepth, DrillingSpeed, Location, and RigID. The 'RigID' column is currently selected.





Untitled - Power BI Desktop

File Home Help Table tools

Name: DrillingDataset  
Mark as data table + Relationships Calculations

Fields

DrillingDataset

Drill through  
Cross-report  
Keep all filters  
Update available (Click to download)

ID	DrillingDepth	Location	Date	Year	Quarter
1001	4000	129 Texas	8/1/2023 12:00:00 AM	2023	3
1002	4000	130 Alaska	8/1/2023 12:00:00 AM	2023	3
1003	4000	131 California	8/1/2023 12:00:00 AM	2023	3
1004	4000	132 New Mexico	8/1/2023 12:00:00 AM	2023	3
1005	4000	133 Colorado	8/1/2023 12:00:00 AM	2023	3
1006	4000	134 Arizona	8/1/2023 12:00:00 AM	2023	3
1007	4000	135 Nebraska	8/1/2023 12:00:00 AM	2023	3
1008	4000	136 North Dakota	8/1/2023 12:00:00 AM	2023	3
1009	4000	137 Wyoming	8/1/2023 12:00:00 AM	2023	3
1010	4000	138 Arkansas	8/1/2023 12:00:00 AM	2023	3
1011	4000	129 Texas	8/2/2023 12:00:00 AM	2023	3
1012	4000	130 California	8/2/2023 12:00:00 AM	2023	3
1013	4000	131 New Mexico	8/2/2023 12:00:00 AM	2023	3
1014	4000	132 Colorado	8/2/2023 12:00:00 AM	2023	3
1015	4000	133 Arizona	8/2/2023 12:00:00 AM	2023	3
1016	4000	134 Nebraska	8/2/2023 12:00:00 AM	2023	3
1017	4000	135 North Dakota	8/2/2023 12:00:00 AM	2023	3
1018	4000	136 Wyoming	8/2/2023 12:00:00 AM	2023	3
1019	4000	137 Alaska	8/2/2023 12:00:00 AM	2023	3

Table: DrillingDataset (38 rows)

Screenshot of Power BI Desktop showing a bar chart and the Power Query Editor.

**Power BI Desktop:**

- Visualizations pane:** Shows the visual hierarchy for "Sum of DrillingSpeed by Location and Location Grouped".
- Fields pane:** Lists fields from the "Drilling\_Dataset":
  - Date Hierarchy
    - Year
    - Quarter
    - Month
    - Day
  - DrillingDepth
  - DrillingSpeed
  - Location
  - Location (grouped)
  - Quarter
  - Right
  - WeekDay
  - Year

**Power Query Editor:**

- Custom Column dialog:** Creating a column named "MonthName" with the formula `= Date.MonthName([Date])`.
- Available columns:** Shows "RigID", "DrillingDepth", "DrillingSpeed", "Location", and "Date".
- Properties pane:** Shows "Name: Drilling\_Dataset" and "Changed Type".

Row	Date	DrillingSpeed	Location
1	2022-01-01	4000	100 - North Dakota
2	2022-01-01	4000	100 - Wyoming
3	2022-01-01	4000	100 - Alaska

Untitled - Power Query Editor

File Home Transform Add Column View Tools Help

General

Format Text Number Date Time Duration Text Analytics Visual Azure Machine Learning

Query Settings

**PROPERTIES**

Name: Drilling\_Dataset

All Properties

**APPLIED STEPS**

Source Promoted Headers Changed Type Added Custom [Added Custom]

Table: #Added\_CustDate, "NewDay", each Data.DayOrWeekname[Date]]

Drilling\_Dataset

1 4500 120 Texas 8/3/2023 September Friday  
 2 4500 120 Alaska 8/5/2023 September Saturday  
 3 4500 120 California 8/6/2023 September Sunday  
 4 4500 120 New Mexico 8/6/2023 September Monday  
 5 4500 120 Colorado 8/7/2023 September Tuesday  
 6 4500 120 Arizona 8/8/2023 September Wednesday  
 7 4500 120 Nebraska 8/9/2023 September Thursday  
 8 4500 120 North Dakota 8/10/2023 September Friday  
 9 4500 120 Wyoming 8/11/2023 September Saturday  
 10 4500 120 Florida 8/12/2023 September Sunday  
 11 4500 120 Texas 8/13/2023 September Monday  
 12 4500 120 California 8/14/2023 September Tuesday  
 13 4500 120 New Mexico 8/15/2023 September Wednesday  
 14 4500 120 Colorado 8/16/2023 September Thursday  
 15 4500 120 Arizona 8/17/2023 September Friday  
 16 4500 120 Nebraska 8/18/2023 September Saturday  
 17 4500 120 North Dakota 8/19/2023 September Sunday  
 18 4500 120 Wyoming 8/20/2023 September Monday  
 19 4500 120 Alaska 8/21/2023 September Tuesday

Untitled - Power BI Desktop

File Home Help Table tools

Name: Drilling\_Dataset

Mark as date table +

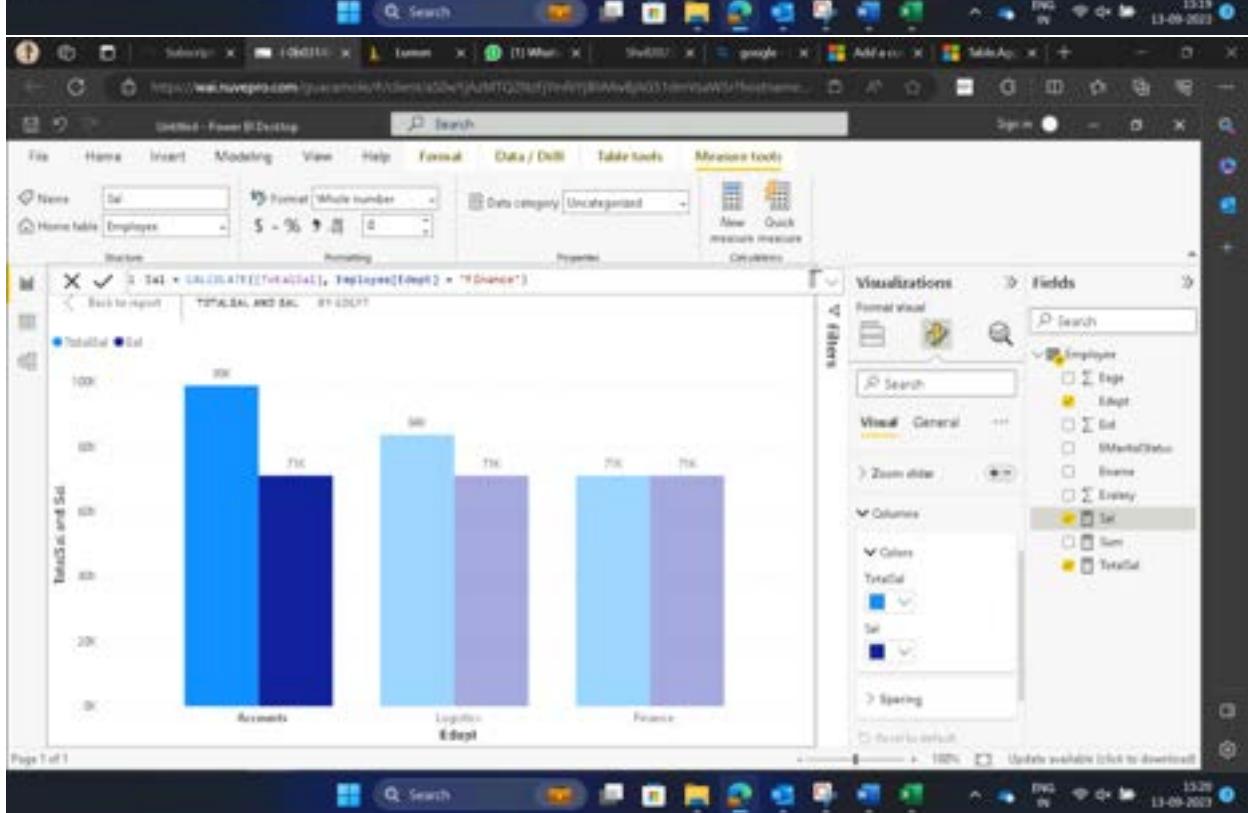
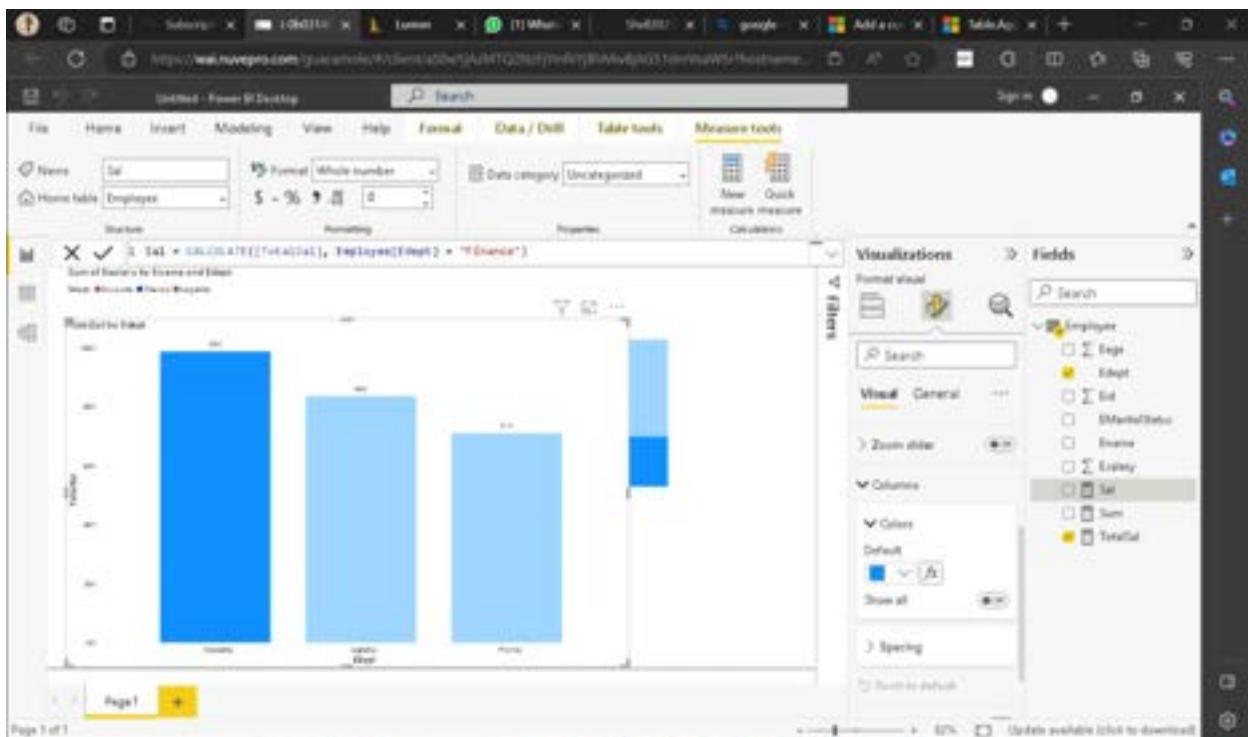
Manage relationships

New Quick measure column table Calculations

Fields

Drilling\_Dataset

1 120 California 8/3/2023 02:00:00 AM 2023 3 1 Alaska & California & Nebraska & North Dakota September Sunday  
 2 120 New Mexico 8/4/2023 02:00:00 AM 2023 3 2 Other September Monday  
 3 120 Colorado 8/5/2023 02:00:00 AM 2023 3 3 Other September Tuesday  
 4 120 Arizona 8/6/2023 02:00:00 AM 2023 3 4 Other September Wednesday  
 5 120 Nebraska 8/7/2023 02:00:00 AM 2023 3 5 Alaska & California & Nebraska & North Dakota September Thursday  
 6 120 North Dakota 8/8/2023 02:00:00 AM 2023 3 6 Alaska & California & Nebraska & North Dakota September Friday  
 7 120 Wyoming 8/9/2023 02:00:00 AM 2023 3 7 Wyoming September Saturday  
 8 120 Florida 8/10/2023 02:00:00 AM 2023 3 8 Alaska & California & Nebraska & North Dakota September Sunday  
 9 120 California 8/11/2023 02:00:00 AM 2023 3 9 Alaska & California & Nebraska & North Dakota September Monday  
 10 120 New Mexico 8/12/2023 02:00:00 AM 2023 3 10 Other September Tuesday  
 11 120 Colorado 8/13/2023 02:00:00 AM 2023 3 11 Other September Wednesday  
 12 120 Arizona 8/14/2023 02:00:00 AM 2023 3 12 Other September Thursday  
 13 120 Nebraska 8/15/2023 02:00:00 AM 2023 3 13 Alaska & California & Nebraska & North Dakota September Friday  
 14 120 North Dakota 8/16/2023 02:00:00 AM 2023 3 14 Alaska & California & Nebraska & North Dakota September Saturday  
 15 120 Wyoming 8/17/2023 02:00:00 AM 2023 3 15 Wyoming September Sunday  
 16 120 Florida 8/18/2023 02:00:00 AM 2023 3 16 Texas September Monday  
 17 120 California 8/19/2023 02:00:00 AM 2023 3 17 Alaska & California & Nebraska & North Dakota September Tuesday  
 18 120 New Mexico 8/20/2023 02:00:00 AM 2023 3 18 Other September Wednesday  
 19 120 Colorado 8/21/2023 02:00:00 AM 2023 3 19 Other September Thursday  
 20 120 Arizona 8/22/2023 02:00:00 AM 2023 3 20 Other September Friday  
 21 120 Nebraska 8/23/2023 02:00:00 AM 2023 3 21 Alaska & California & Nebraska & North Dakota September Saturday  
 22 120 North Dakota 8/24/2023 02:00:00 AM 2023 3 23 Wyoming September Sunday  
 23 120 Wyoming 8/25/2023 02:00:00 AM 2023 3 24 Texas September Monday  
 24 120 Florida 8/26/2023 02:00:00 AM 2023 3 25 Other September Tuesday



The image shows two Microsoft Excel spreadsheets side-by-side, both titled "DeptSal" and "EmpInfo".

**DeptSal Table:**

ID	Dept	Salary
1	Finance	10000
2	Logistics	11223
3	Finance	13231
4	Logistics	151512
5	Accounts	14223
6	Finance	19402
7	Logistics	24441
8	Accounts	21211
9	Accounts	10485
10	Logistics	41111

**EmpInfo Table:**

ID	Name	Age	EMail	Status
1	Aishika	17	S	S
2	Remi	12	M	
3	Avantika	5		
4	Maia	21	S	
5	Ashley	23	M	
6	Jani	18	M	
7	Rose	27	S	
8	Jimmy	22	M	
9	Jack	3	S	
10	Bruke	19	M	

Employee - Excel [Read-Only] [Autosave]

File Insert Page Layout Formulas Data Review View

Cells General Conditional Formatting Cell Styles Insert Delete Format Sort & Filter Go To... Editing

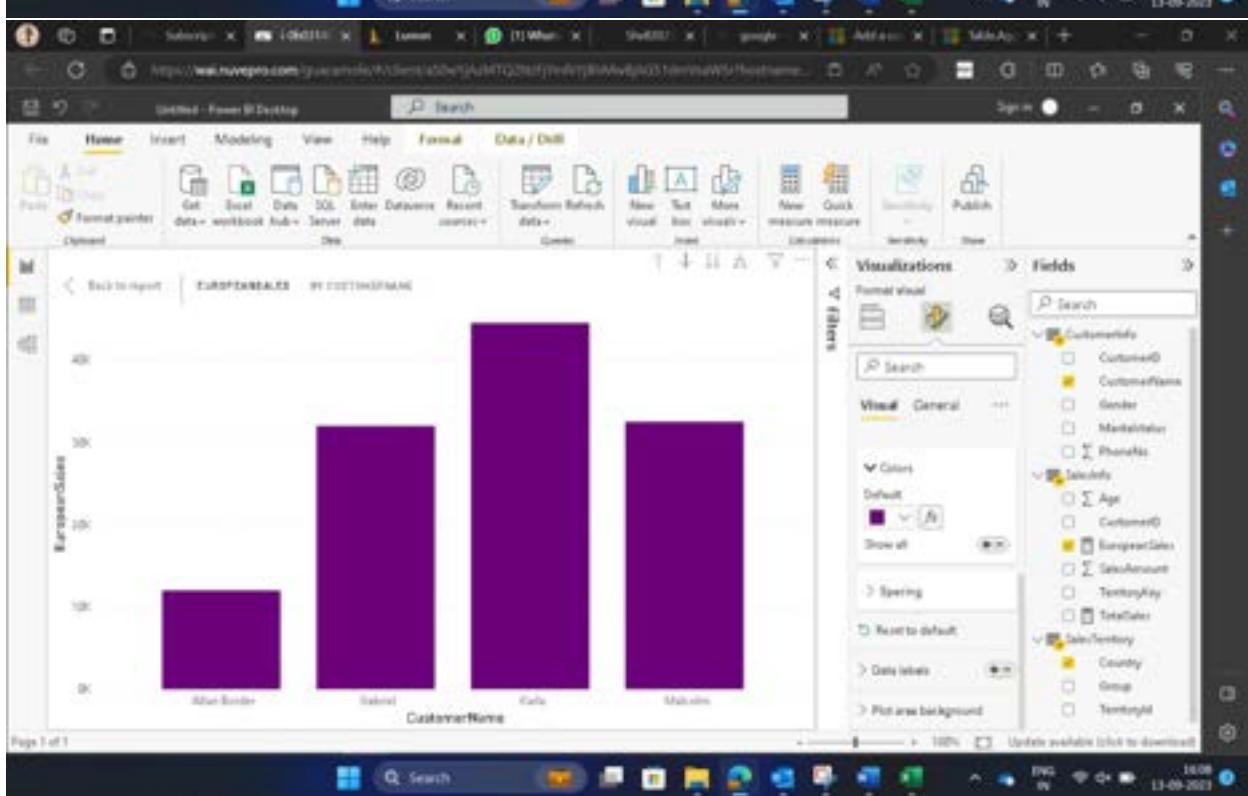
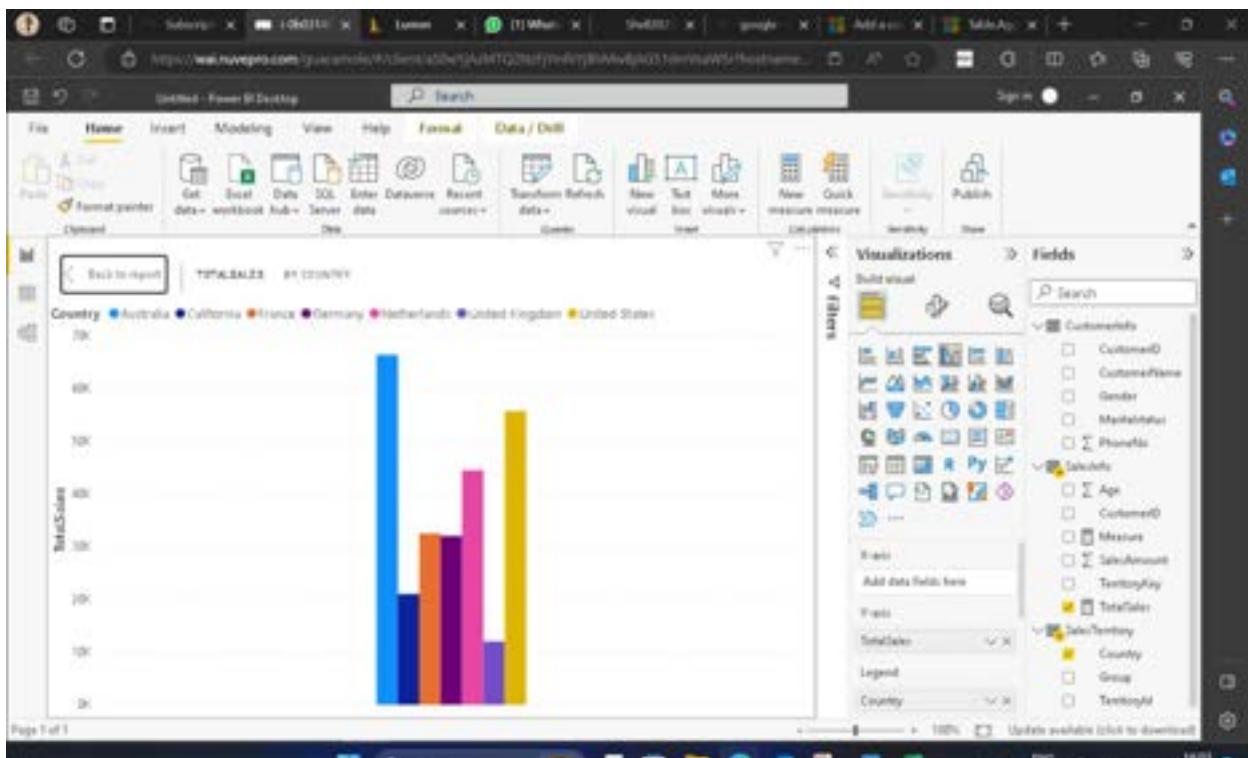
Font Alignment Number Style Cell Type Date Time

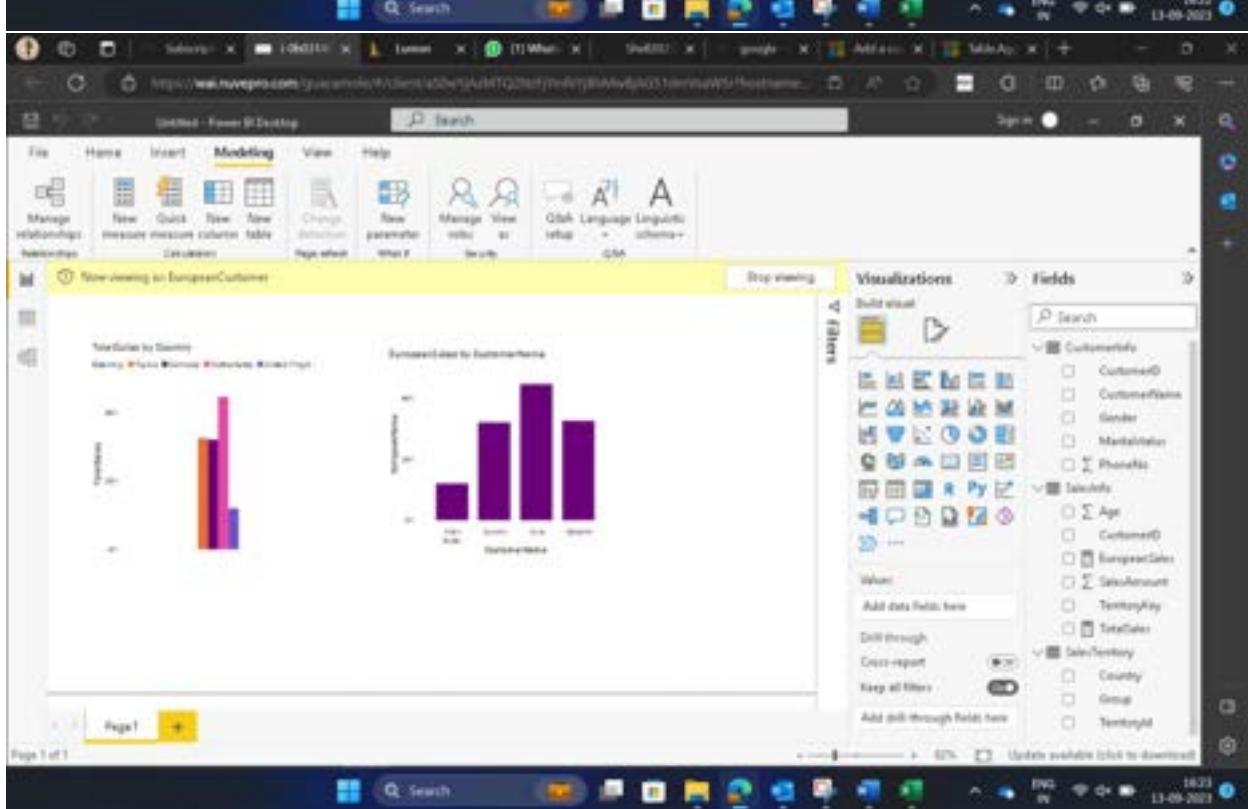
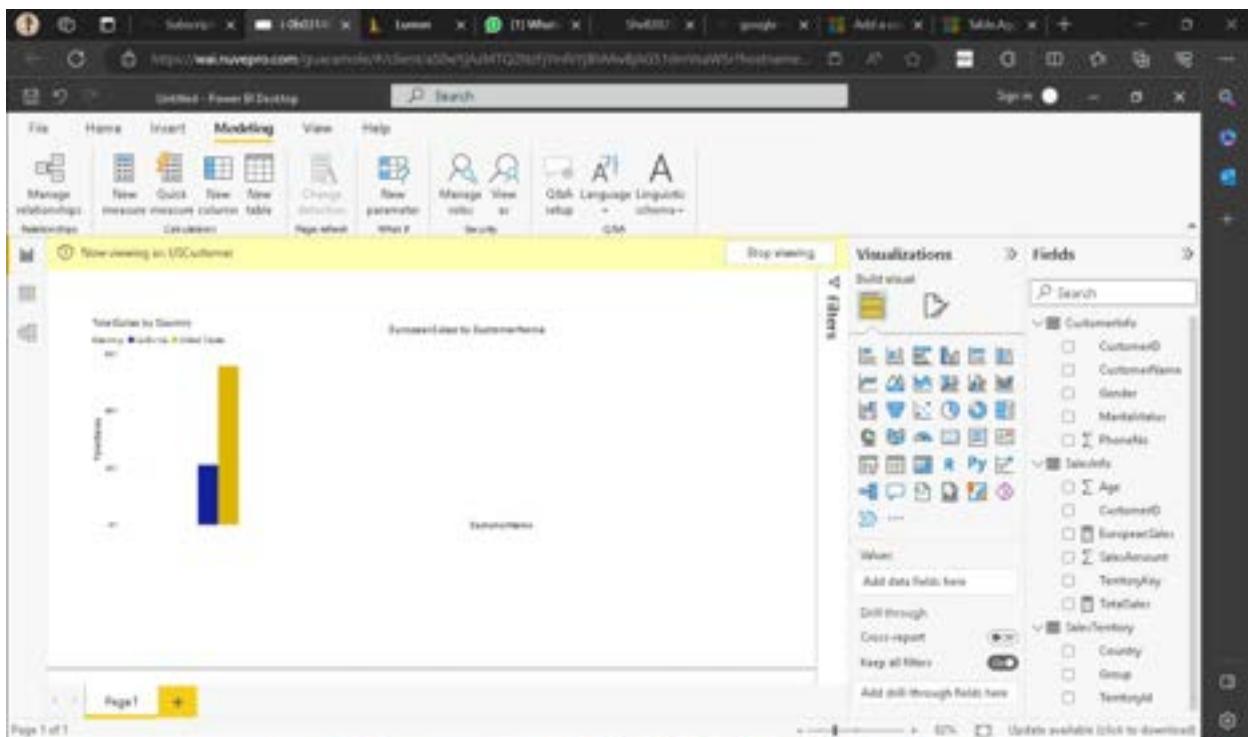
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	Id	First Name	Last Name	Age	EMailAddress	Gender										
1	1	Aliya	17	17	f											
2	2	Rami	12	12	m											
3	3	Ashton	5	5	f											
4	4	Z Ma	21	21	f											
5	5	Ashley	23	23	m											
6	6	Jens	4	4	f											
7	7	L Rose	27	27	f											
8	8	Z Jimmy	22	22	m											
9	9	Zack	8	8	m											
10	10	Z Brooke	19	19	m											

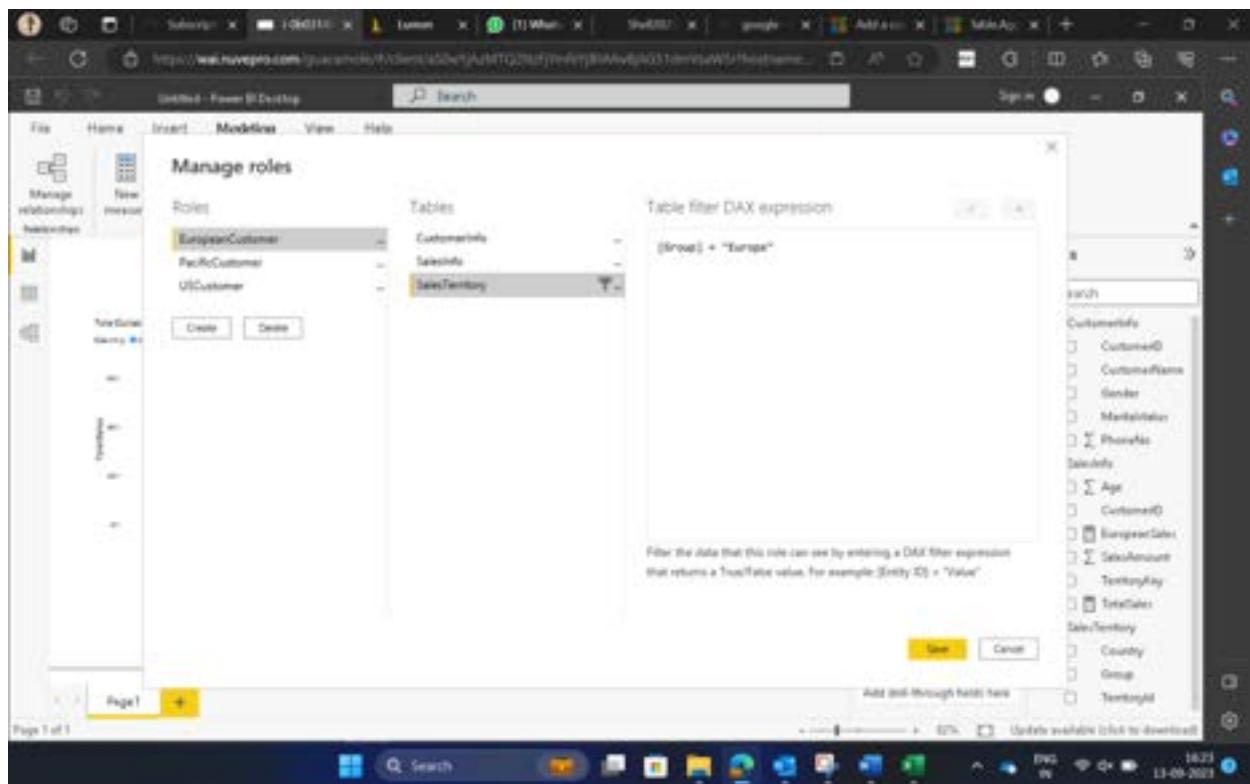
EmpFilter DeptSel

Search

13-09-2021 13:29 100%







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}  
}  
)
```

Untitled - Power BI Desktop

File Home Insert Modeling View Help

New Quick View New measure New column Table Calculations Change detection New parameter What If Manage roles Manage View AI GQL Language Linguistic schema

Sum of Sales by Country

Sum of Sales by UserName

View as roles

None  
 Other user:   
 DynamicRow

Visualizations Fields

UserTable

- Country
- Email
- $\Sigma$  Sales
- $\Sigma$  User
- UserName

Value

Add data field here

Drill through

Cross-report

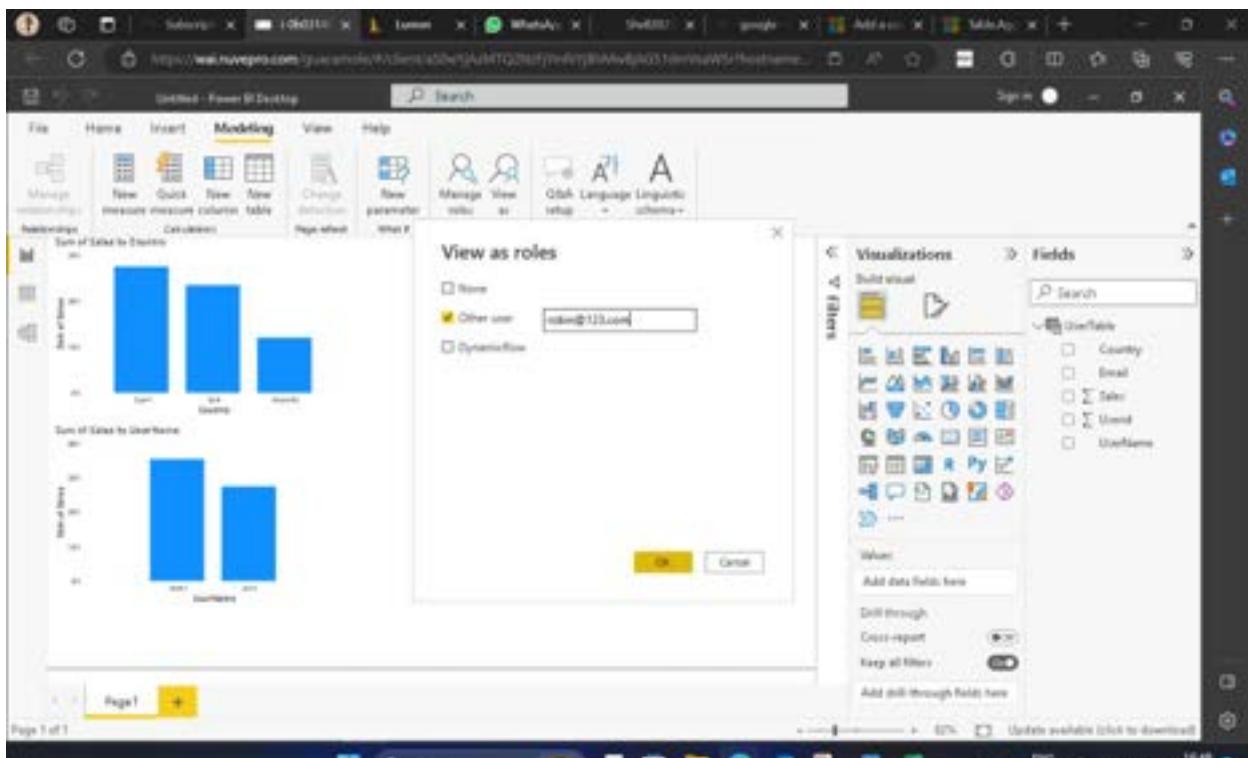
Keep all filters

Add drill-through fields here

Page 1 of 1

Search

16:49 13-09-2021



Untitled - Power BI Desktop

File Home Insert Modeling View Help

New Quick View New measure New column Table Calculations Change detection New parameter What If Manage roles Manage View AI GQL Language Linguistic schema

Sum of Sales by Country

Sum of Sales by UserName

Manage roles

Roles: **DynamicRow** Tables: **UserTable**

Table filter DAX expression:

```
(EntityID = "userprincipalname")
```

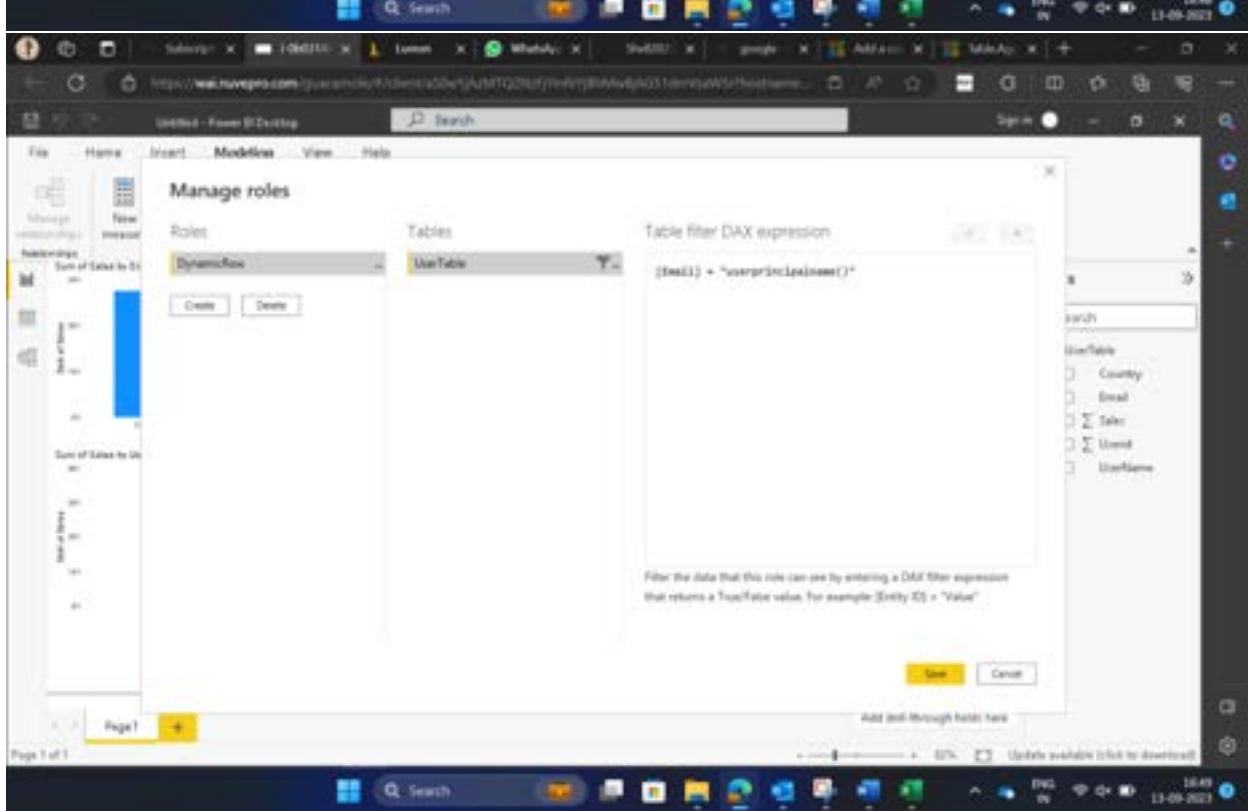
Filter the data that this role can see by entering a DAX filter expression that returns a True/False value. For example: (Entity ID) = "Value"

Add drill-through fields here

Page 1 of 1

Search

16:49 13-09-2021



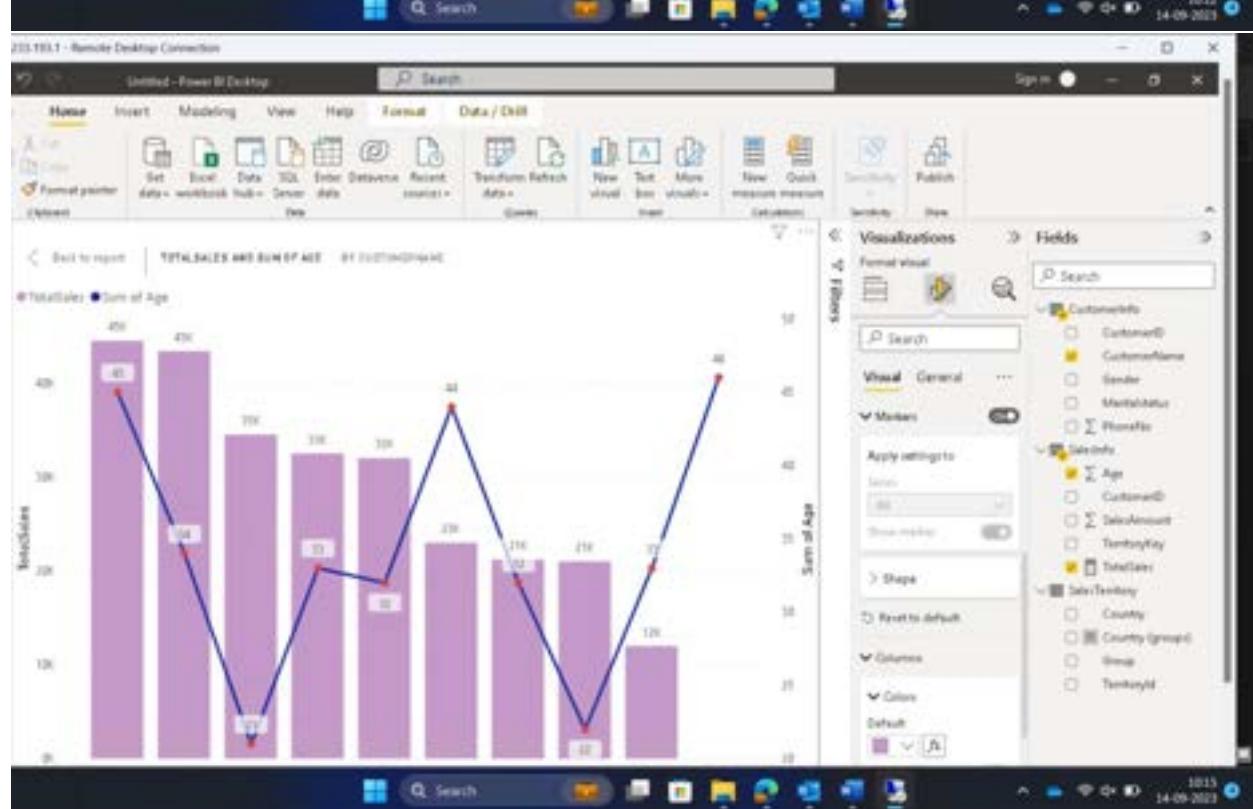
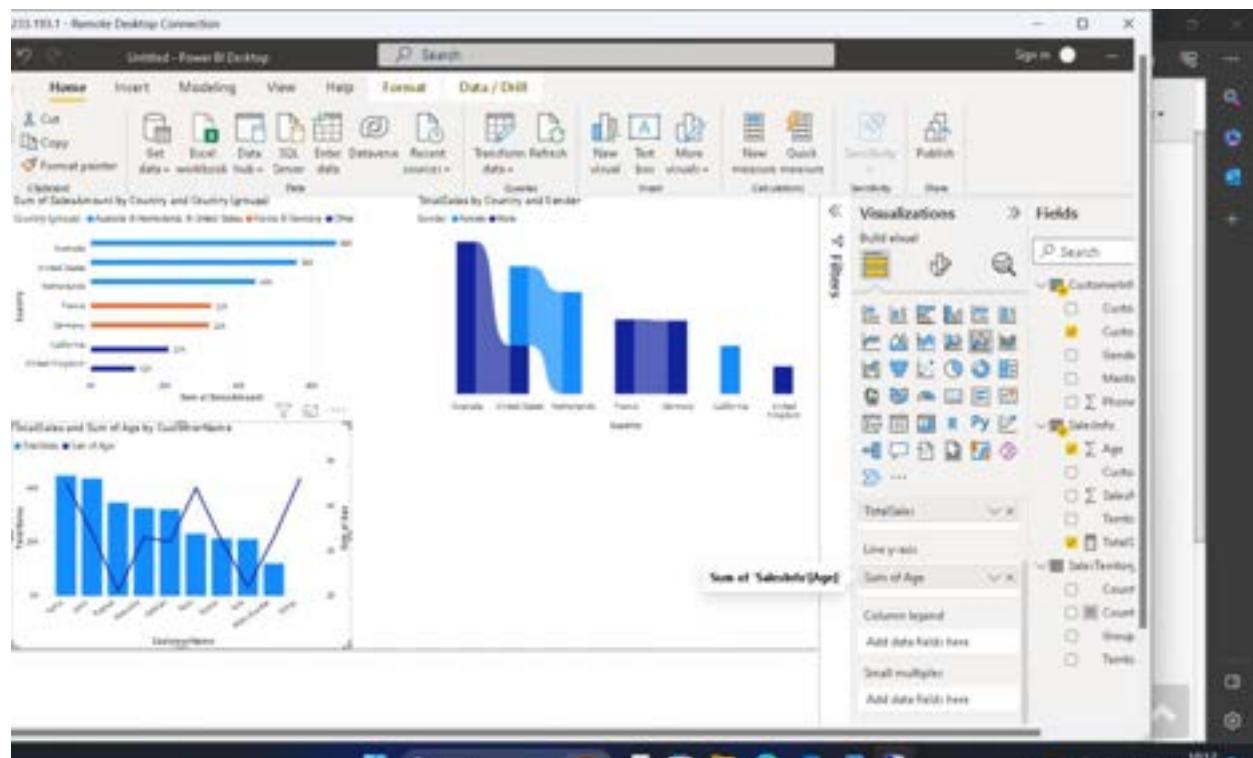
Day 12 14.09.23

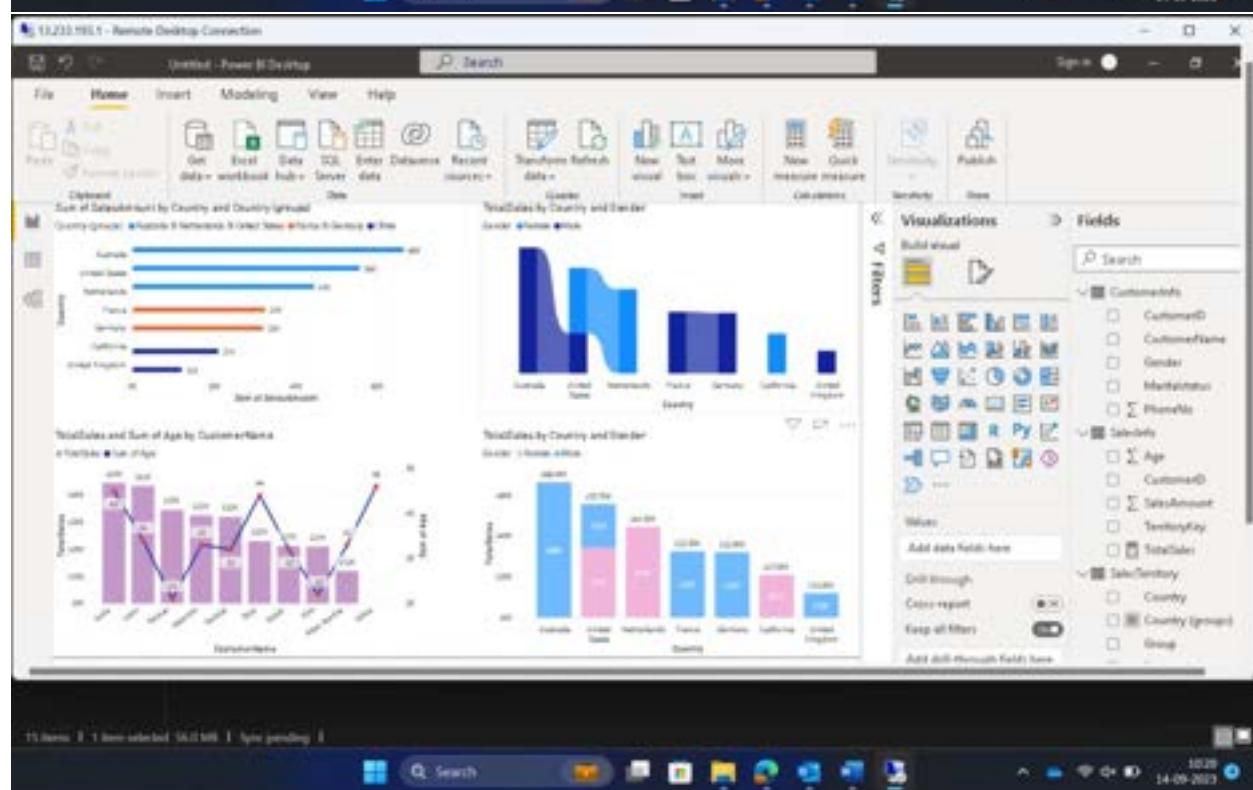
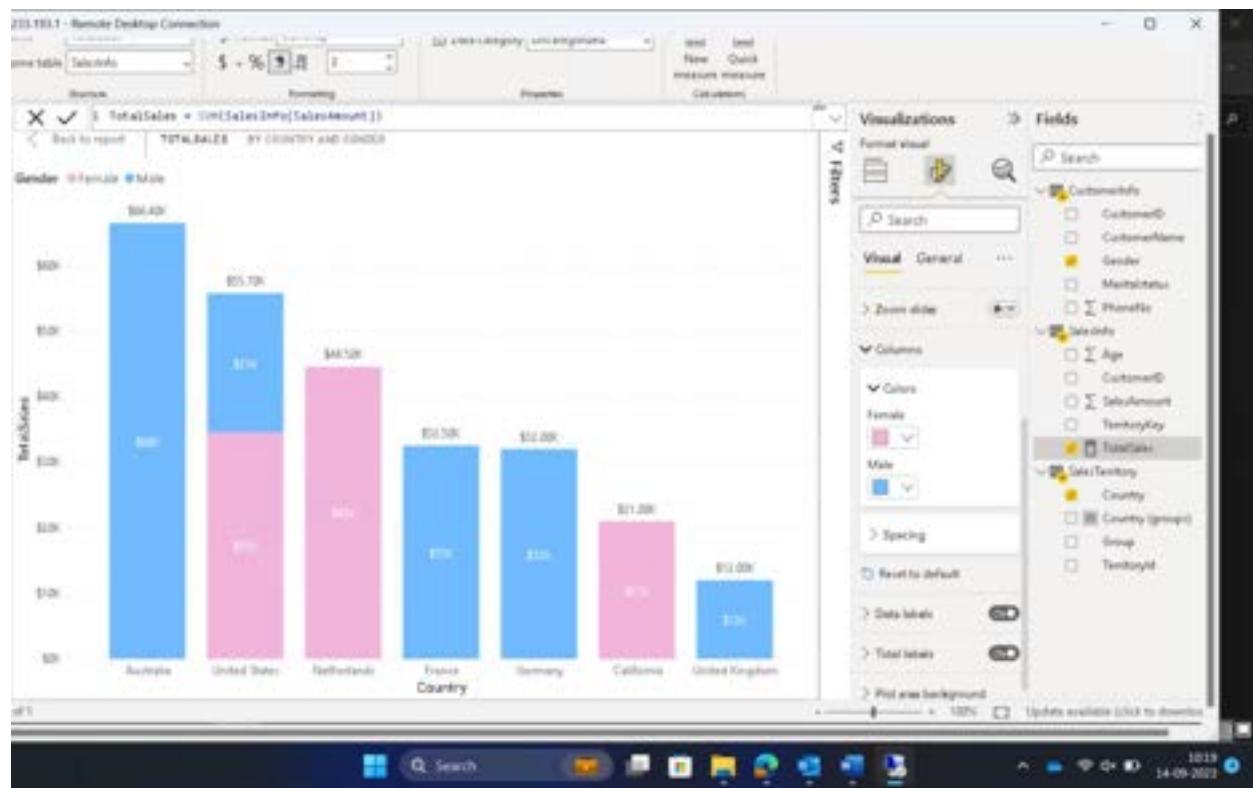
1. bar chart

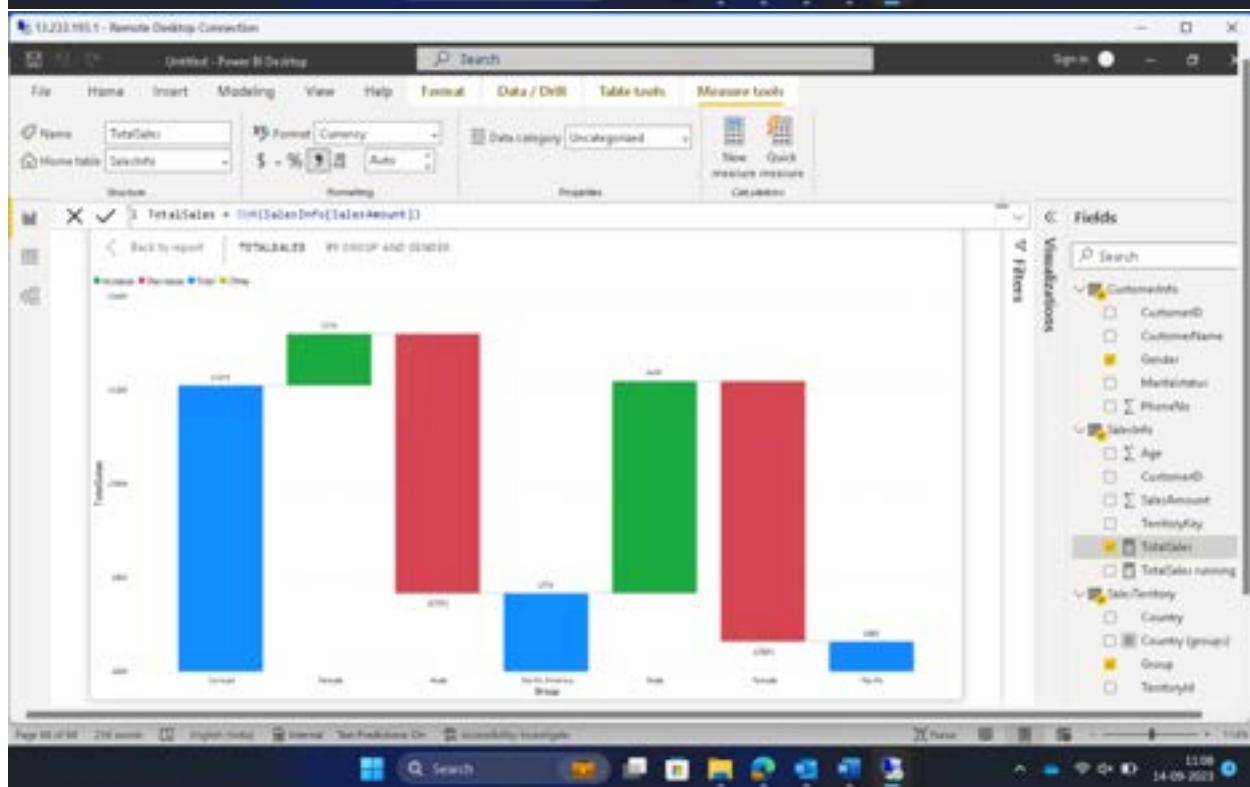
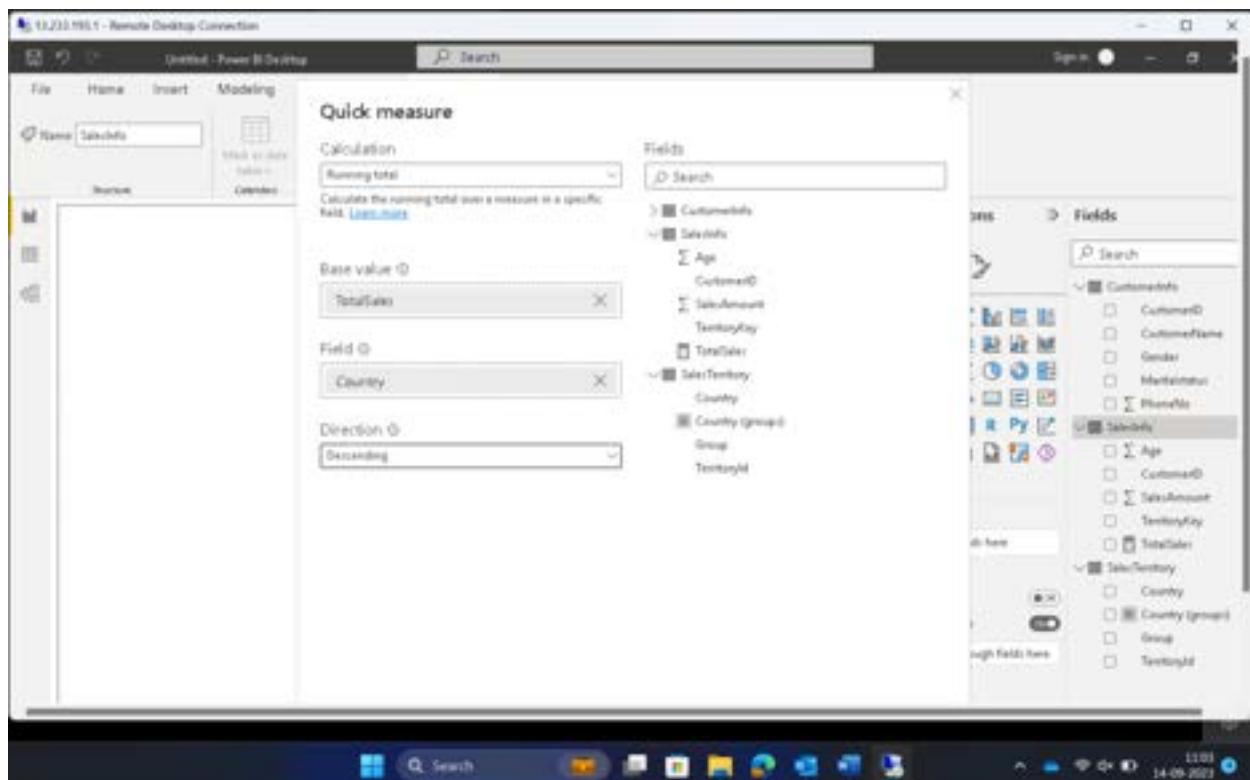
2. grouping data

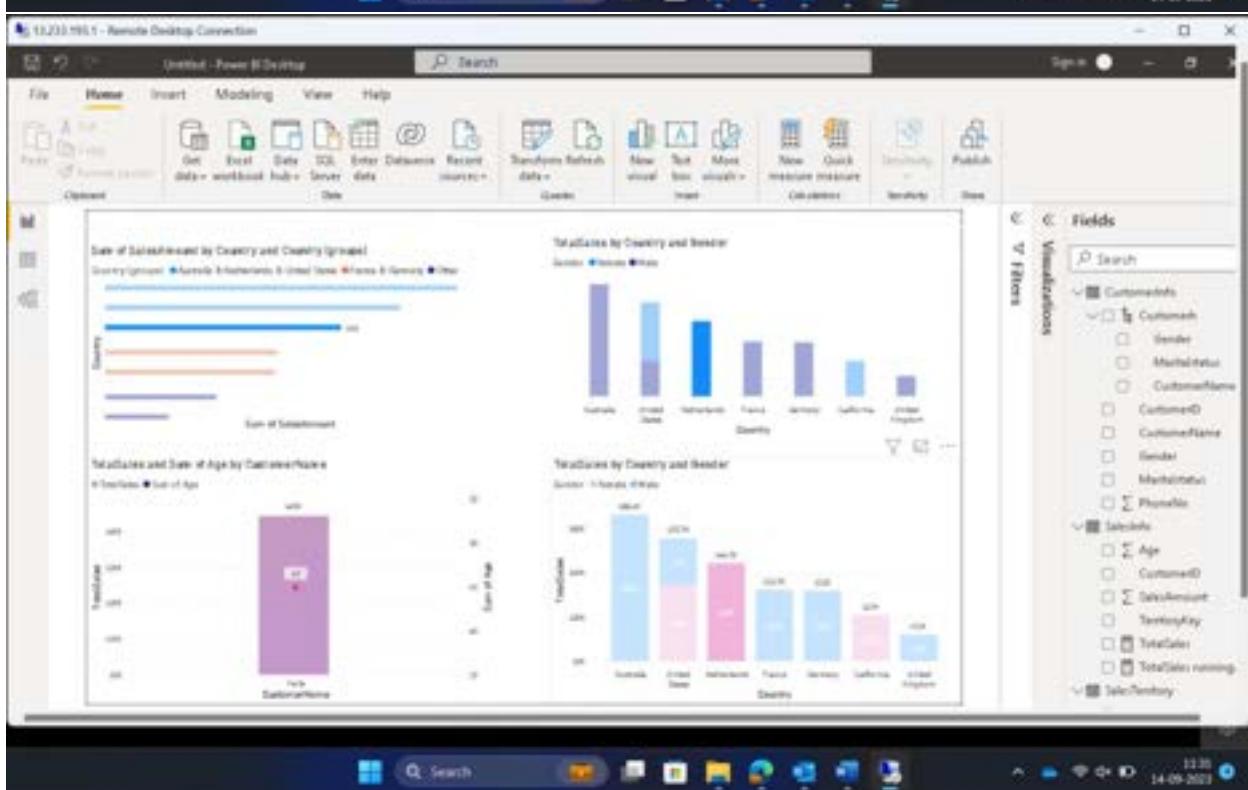
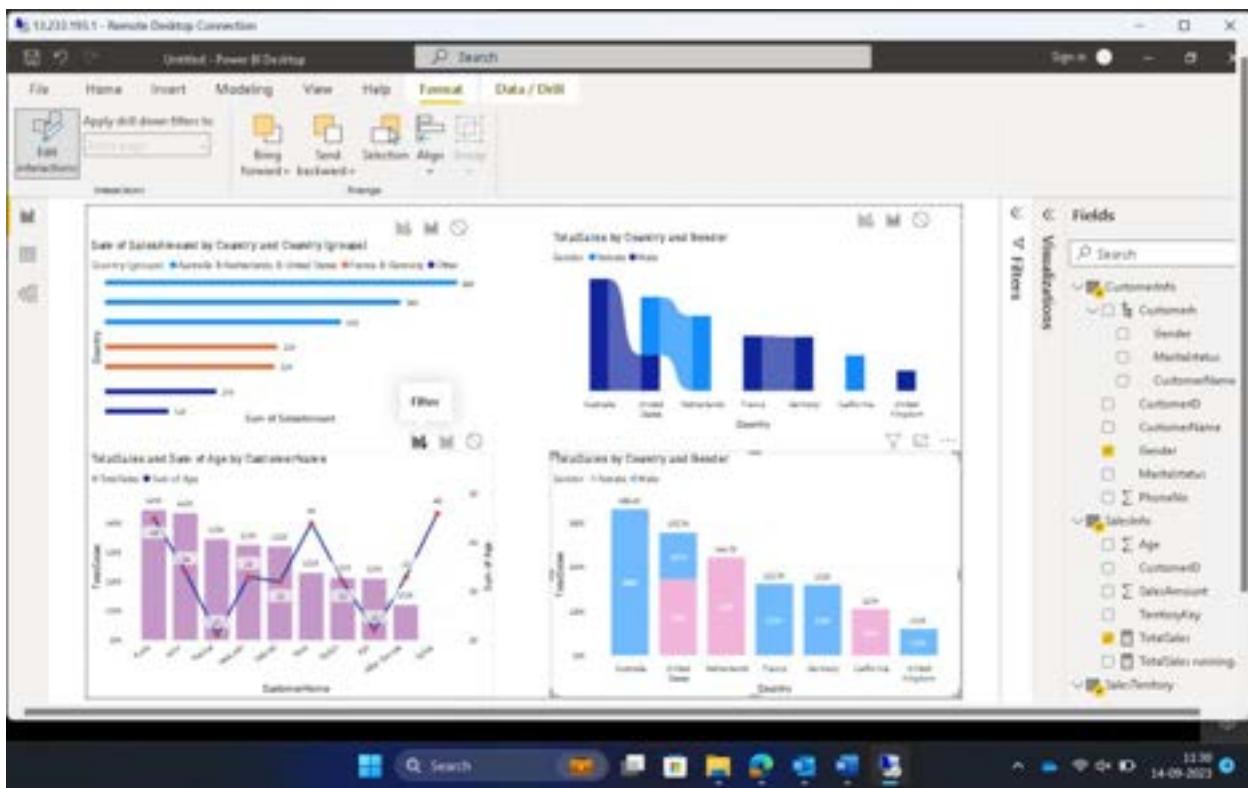
3. line and clustered chart

4. ribbon chart









The screenshot shows a Microsoft Power BI Desktop interface. The ribbon at the top has tabs: File, Home, Insert, Modeling (which is selected), View, Help, Format, and Data / Drill. Below the ribbon are several icons for managing relationships, creating measures, and working with tables and columns. A search bar is located above the main content area. The main area displays a pie chart titled "TOTAL SALES" from 01-09-2013. The chart is divided into three segments: North America (30%), Europe (30%), and Asia Pacific (40%). To the right of the chart is a "Visualizations" pane with a "Format visual" tab and a "Fields" pane. The "Fields" pane lists various data items under "Visuals": CustomerKey, CustomerName, CustomerID, Gender, MaritalStatus, PhoneNo, Security, and TotalSales. It also lists "Options" such as Position (Inside), OverflowText, and Label contents (Category, data value). The "Values" section includes SalesTerritory and SalesTerritoryKey.

The screenshot shows the Microsoft Power BI Desktop interface with three visualizations:

- Bar Chart:** Titled "EuropeanSales", it displays SalesAmount by Region. The Y-axis ranges from 0 to 1000. The data is as follows:

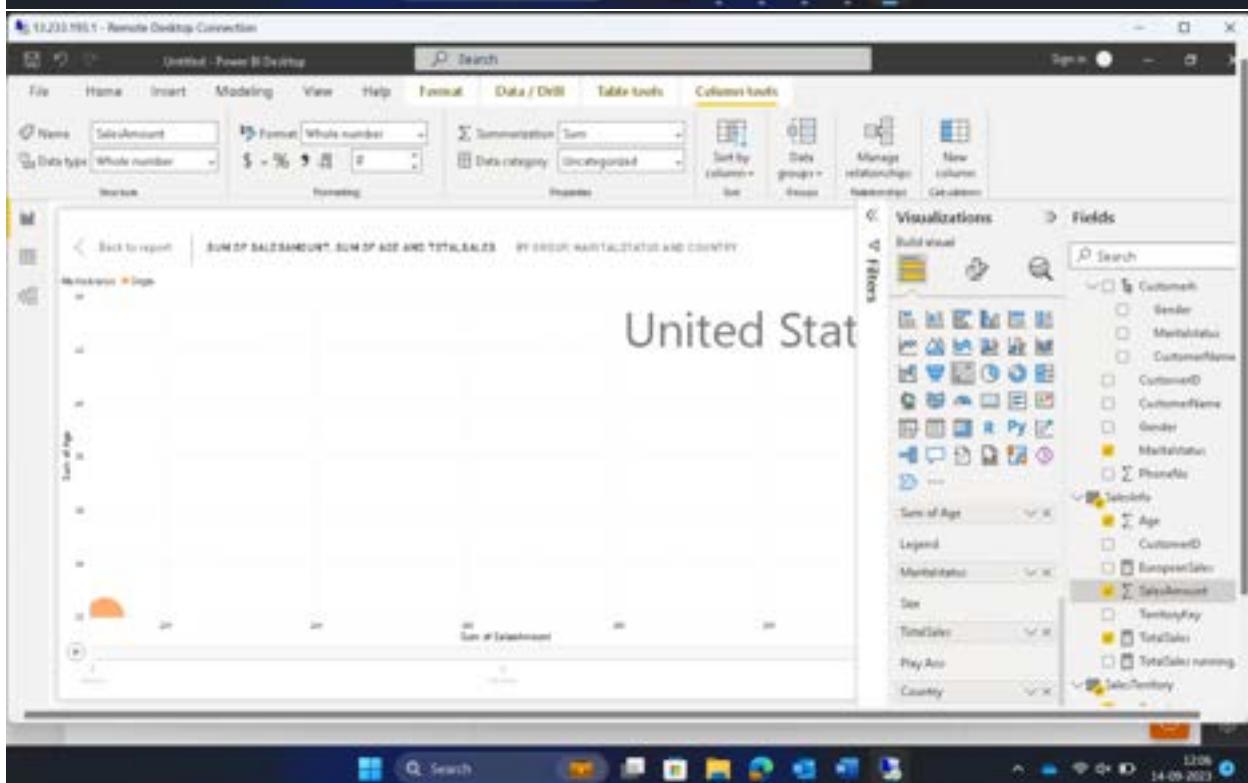
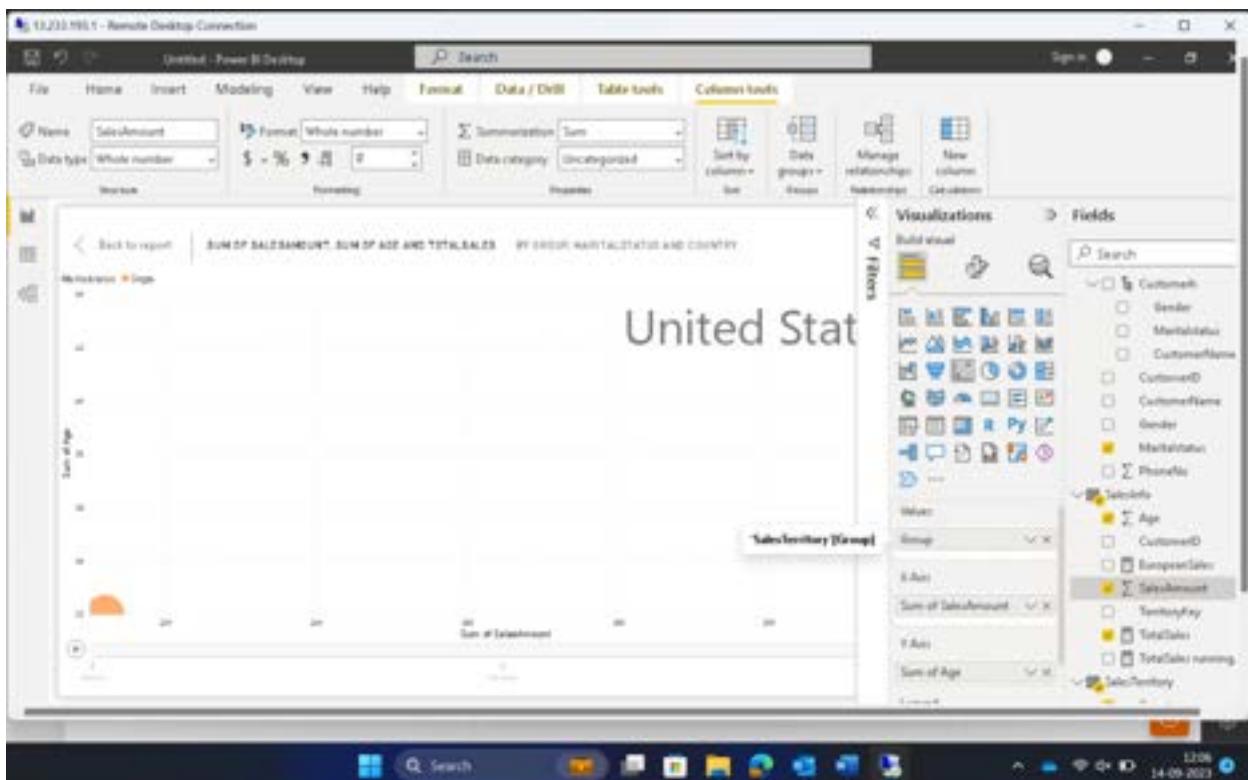
Region	SalesAmount
Europe	1000
America	800
Asia	400
Japan	300
Australia	200

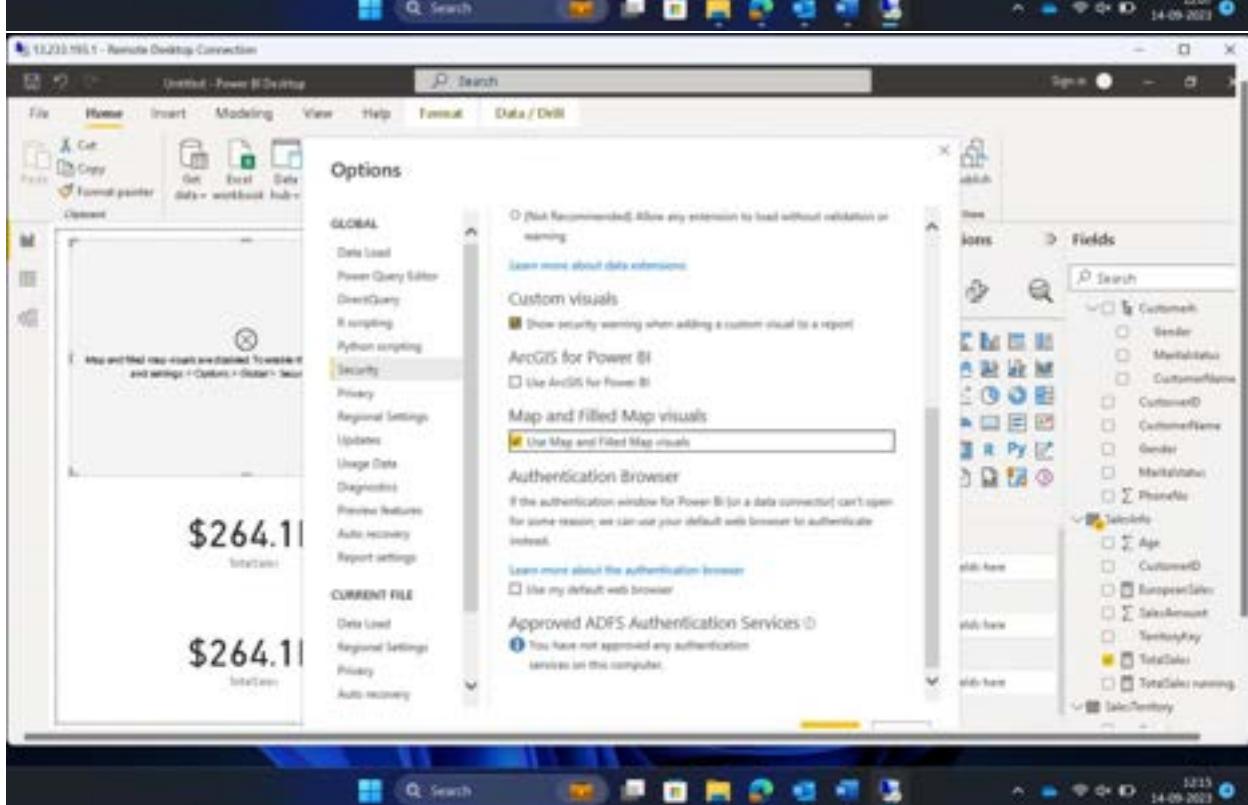
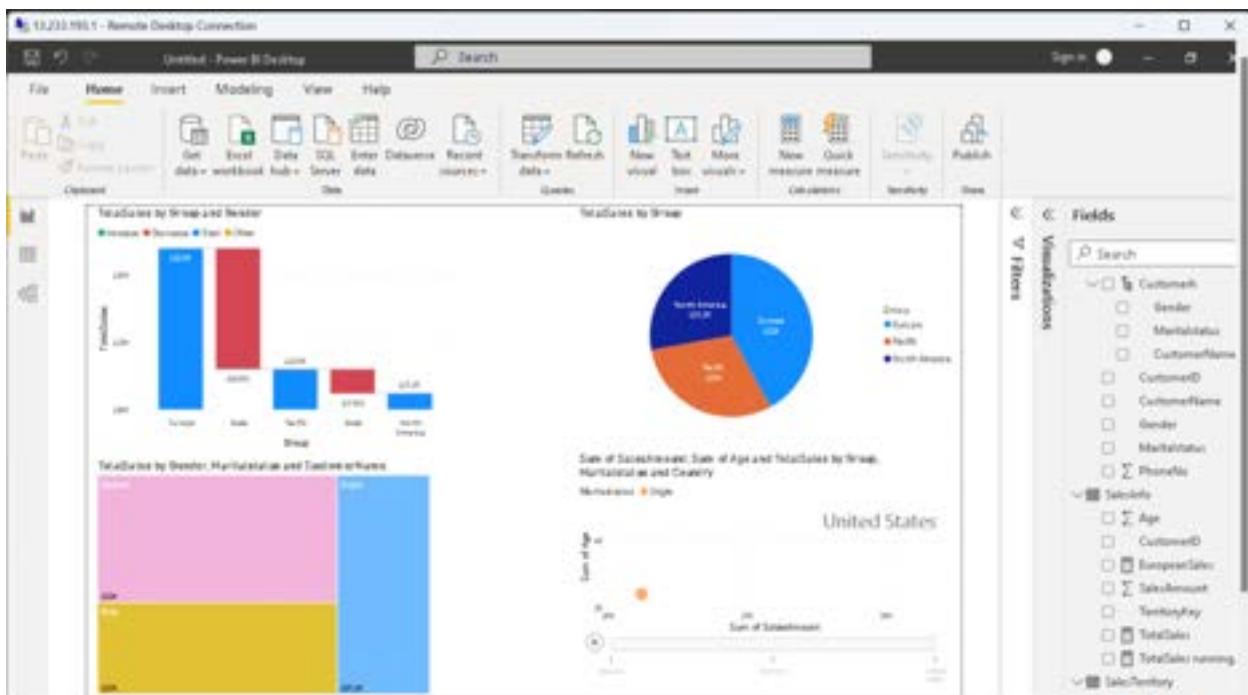
- Pie Chart:** Titled "TotalSales by Gender", it shows the distribution of TotalSales by Gender. The data is as follows:

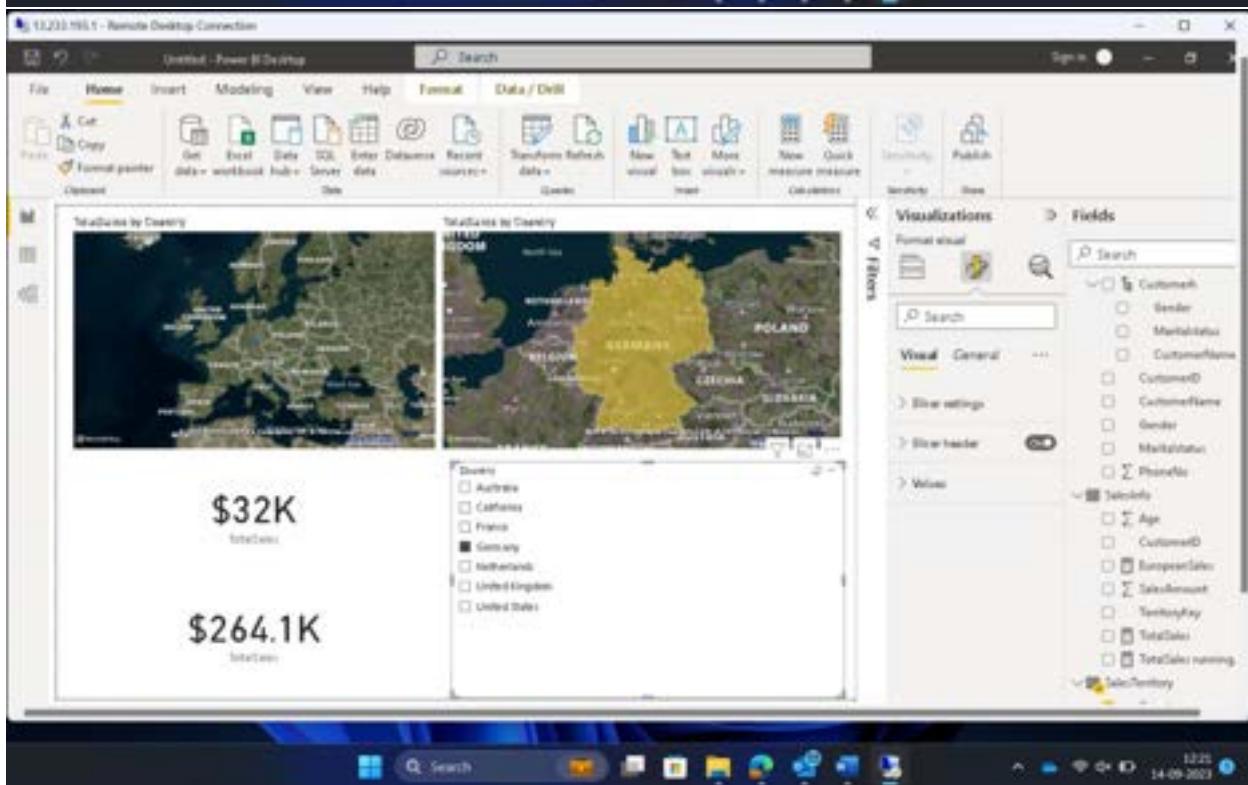
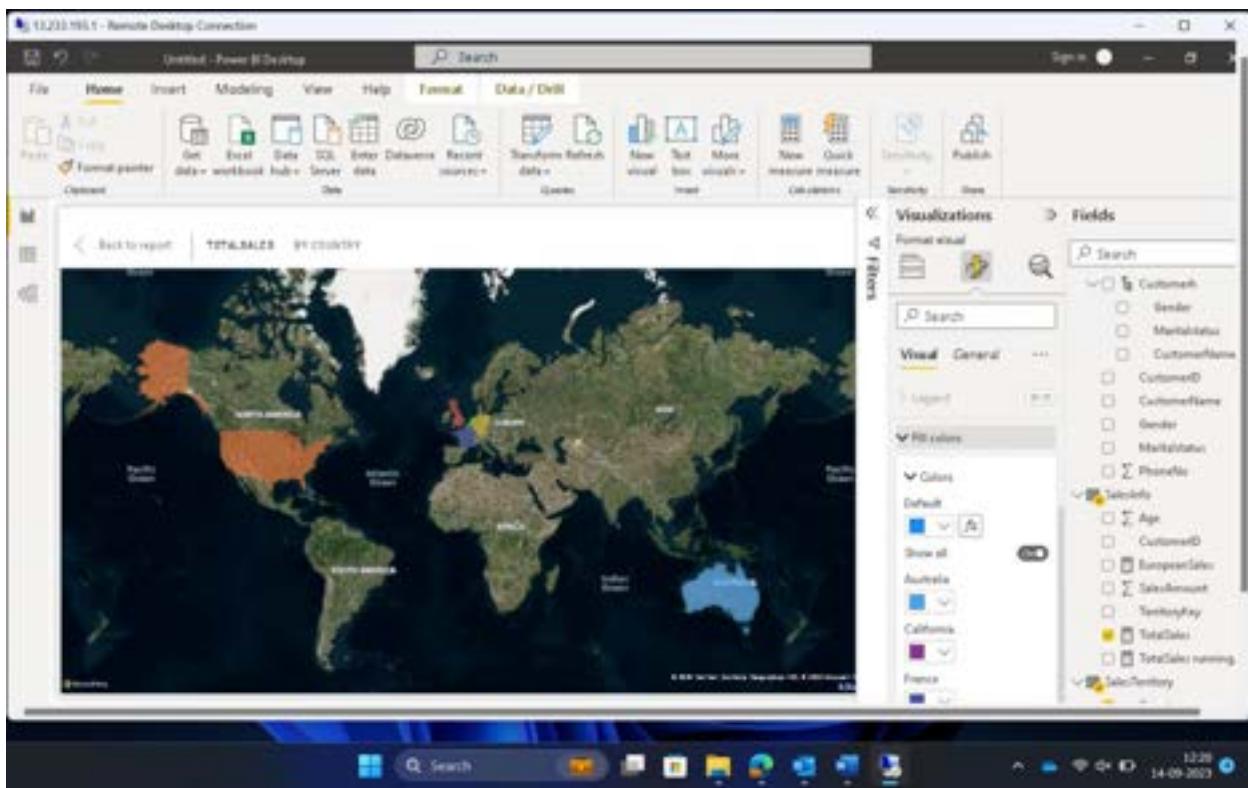
Gender	TotalSales
Male	500
Female	400

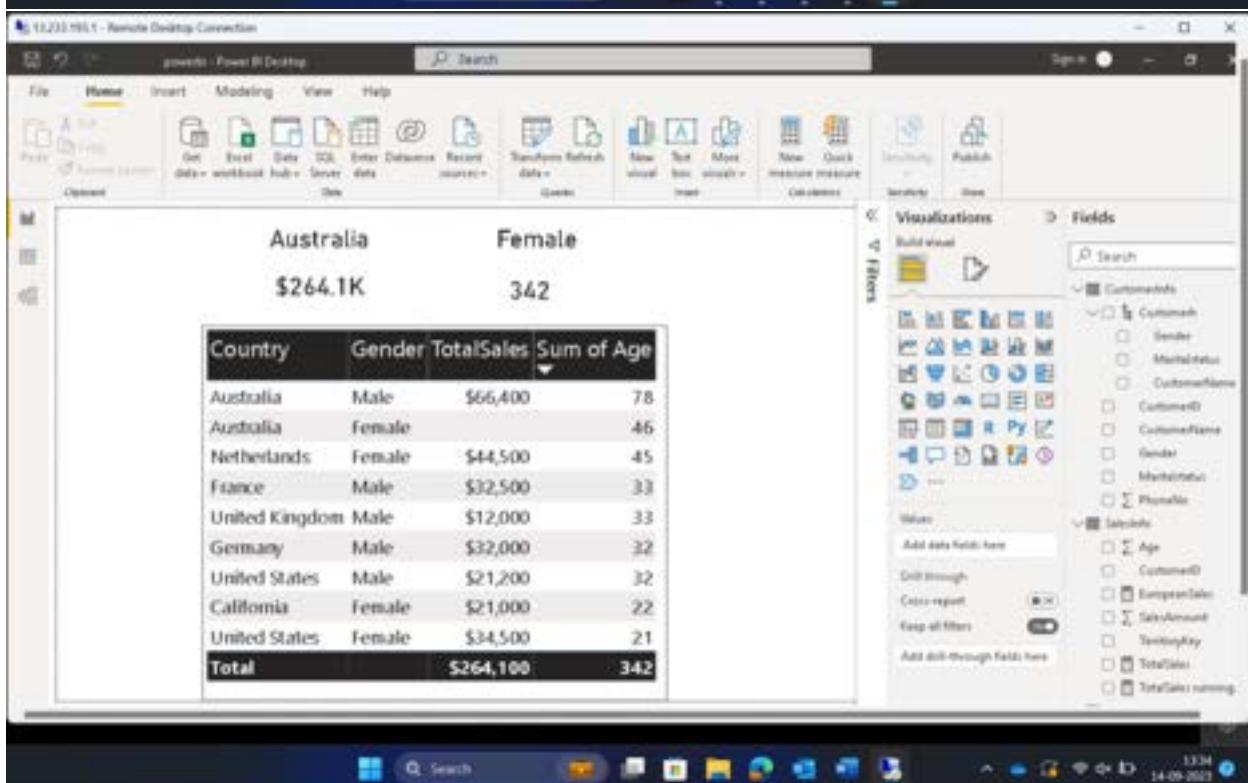
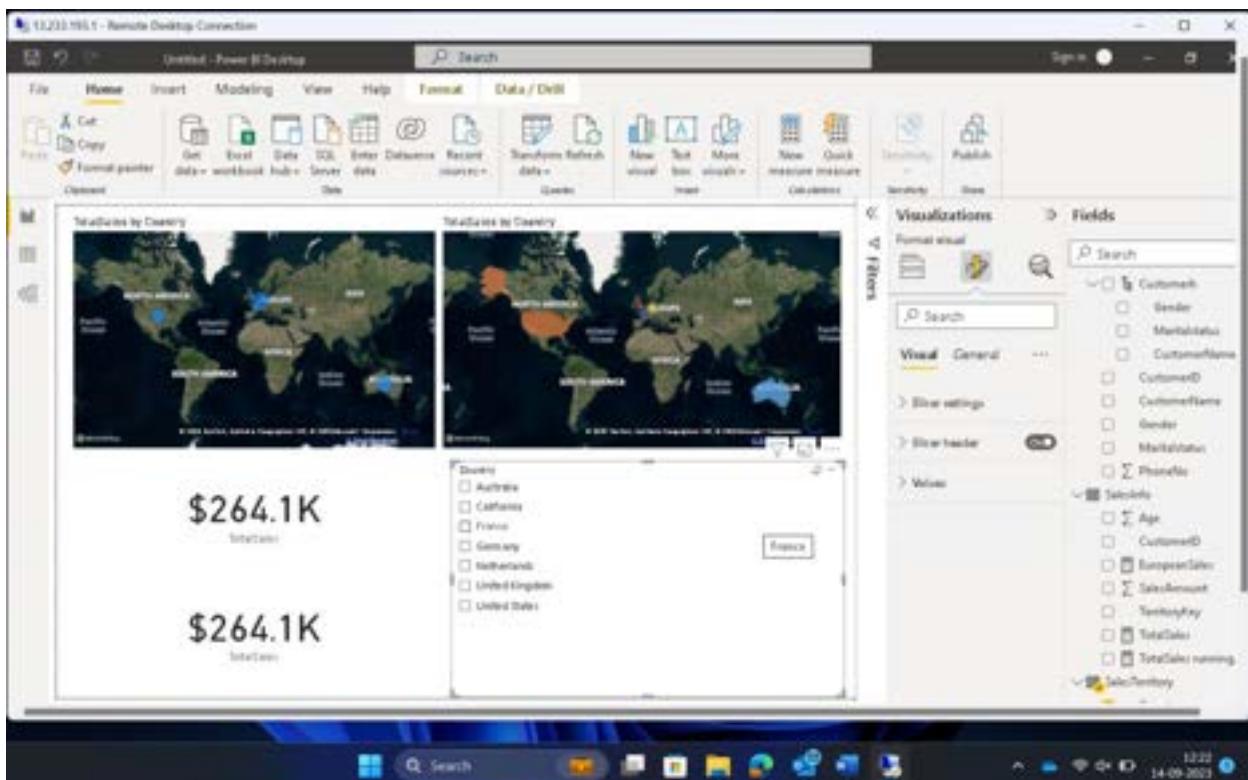
- Stacked Bar Chart:** Titled "TotalSales by Gender, MaritalStatus and TotalSalesRank", it shows the breakdown of TotalSales by Gender and MaritalStatus. The data is as follows:

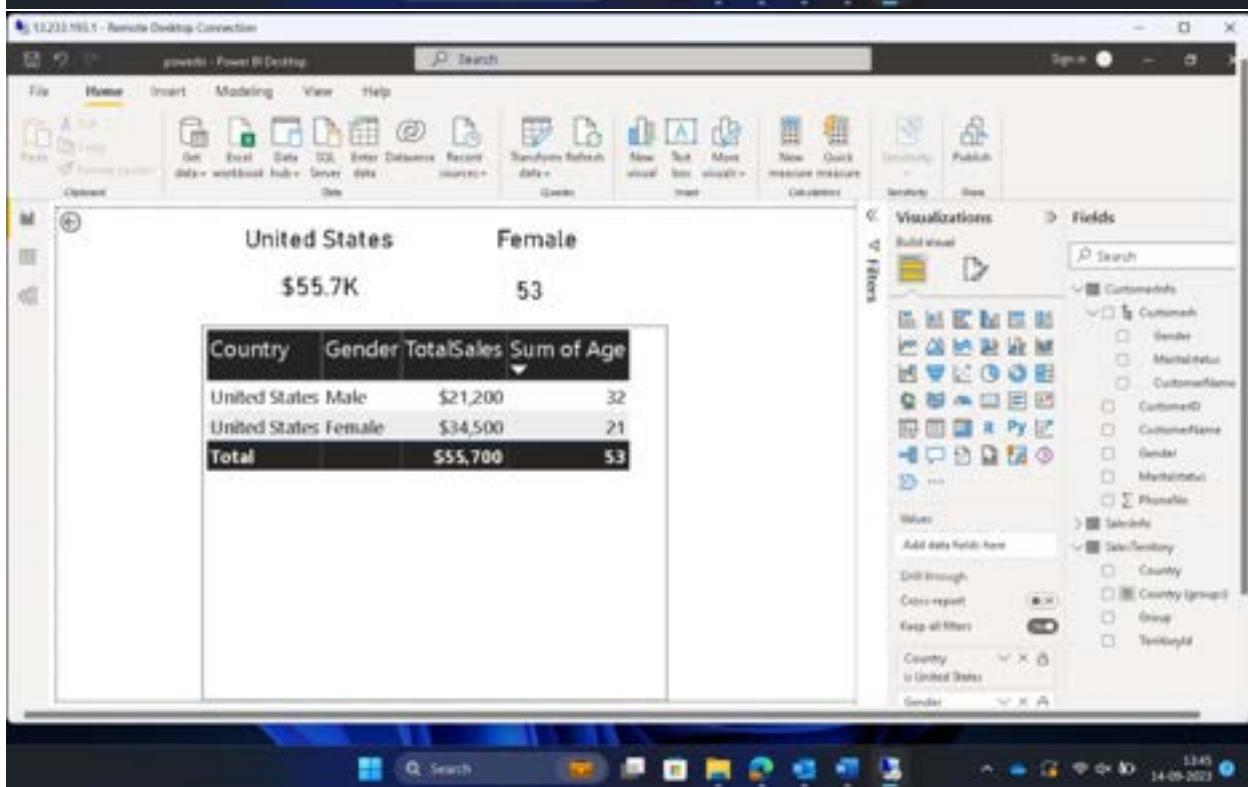
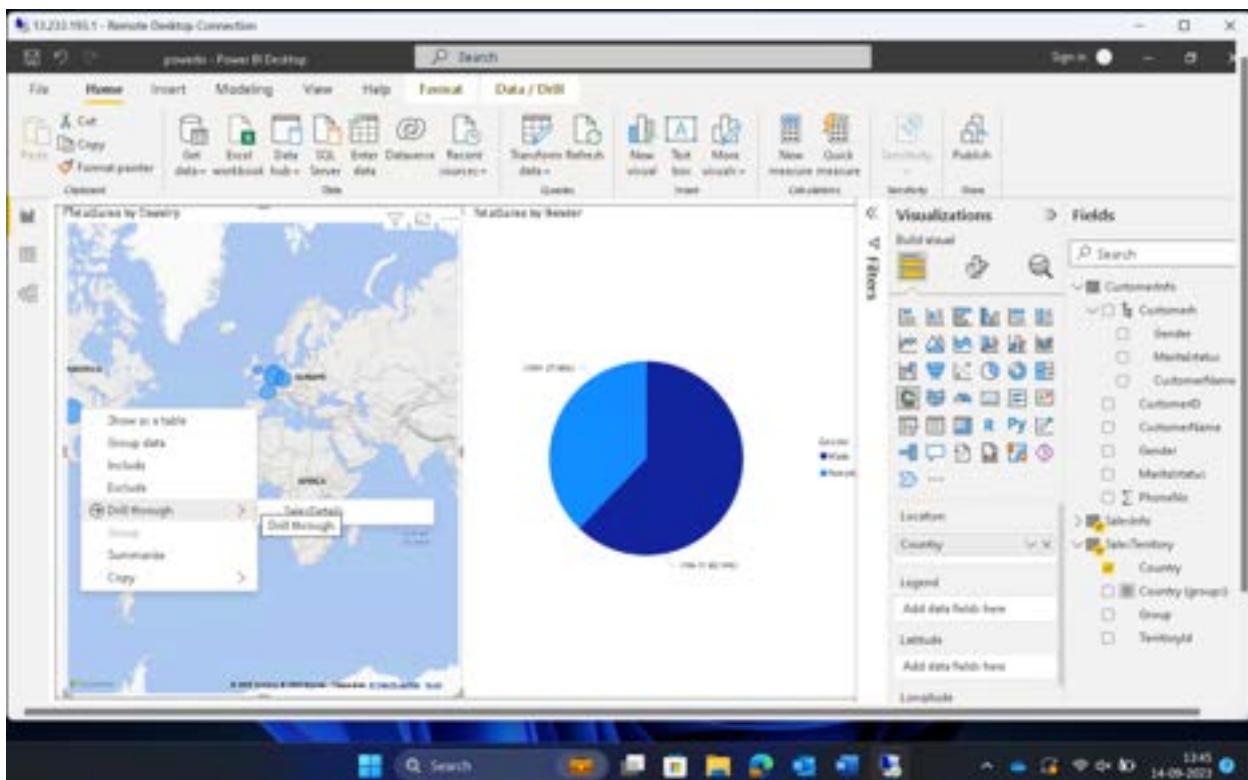
MaritalStatus	Gender	TotalSales
Married	Male	300
Married	Female	200
Single	Male	100
Single	Female	100











132.0.190.1 - Remote Desktop Connection

powerbi - Power BI Desktop

File Home Insert Modeling View Help

Get data + workload Data DAX Enter Database Recent sources Sandboxed Refresh data... Queries New visual Test back-end More visual... New Quick measure measure Calculations Options

Visualizations Fields

Australia Male

\$164.1K 208

Country	Gender	TotalSales	Sum of Age
Australia	Male	\$66,400	78
France	Male	\$32,500	33
United Kingdom	Male	\$12,000	33
Germany	Male	\$32,000	32
United States	Male	\$21,200	32
<b>Total</b>		<b>\$164,100</b>	<b>208</b>

132.0.190.1 - Remote Desktop Connection

powerbi - Power BI Desktop

File Home Insert Modeling View Help Table tools Measure tools

Name: Measure Format: \$ - % +

Data category: Uncategorized

New Quick measure measure Calculations

Visualizations Fields

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

```
SelectedGender =
```

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

112331901 - Remote Desktop Connection

100% 800x600 - white number

00 Data Category: ungrouped

Columns: 3

Rows: 10

Relationships: Customer

Columns: 3

Visualizations: Fields

Build visual

Search: Customer

Customer

- CustomerID
- CustomerName
- Gender
- MaritalStatus
- CustomerName
- CustomerID
- CustomerName
- Gender
- MaritalStatus
- PhoneNo
- SelectedGender

Age

Remove field

Reorder for this visual

Mode

Add a quickline

Conditional formatting

Show value to

New quick measure

Web URL

Background color

Font color

Data key

Icons

Web URL

Update available (click to download)

Page 8 of 8

Page 1 | Page 2 | Page 3 | TableInfo | [Selected](#) | Page 8 | [+/-](#)

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
<b>Total</b>		<b>\$264,100</b>

112331901 - Remote Desktop Connection

powerbi - Power BI Desktop

File Home Insert Modeling View Help Format Data / Drill Table tools Column tools

Search

Background color - TotalSales

Format style: Apply to: Rules Values only

What field should we base this on? Summarization: Sum

Rules: Reverse color order + New rule

Value: < 10000 Number and > 23000 Number Then: Red

Value: > 23000 Number and < 25000 Number Then: Blue

Value: > 25000 Number and < 35000 Number Then: Green

Customer

- CustomerID
- CustomerName
- Gender
- MaritalStatus
- CustomerName
- CustomerID
- CustomerName
- Gender
- MaritalStatus
- PhoneNo
- SelectedGender

Age

Reverse color order

+ New rule

Search

Customer

- CustomerID
- CustomerName
- Gender
- MaritalStatus
- CustomerName
- CustomerID
- CustomerName
- Gender
- MaritalStatus
- PhoneNo
- SelectedGender

Age

CustomerID

CustomerName

Gender

MaritalStatus

PhoneNo

SelectedGender

TotalSales

14:08 14-09-2021

14:13 14-09-2021

132.0.190.1 - Remote Desktop Connection

powertbl - Power BI Desktop

File Home Insert Modeling View Help Format Data / Drill Table tools Column tools

Name: Age Data type: Whole number Format: Currency \$ - % ⚡ Sum Data category: Uncategorized Sort by column: Name Data group: None Manage relationships New column Generate

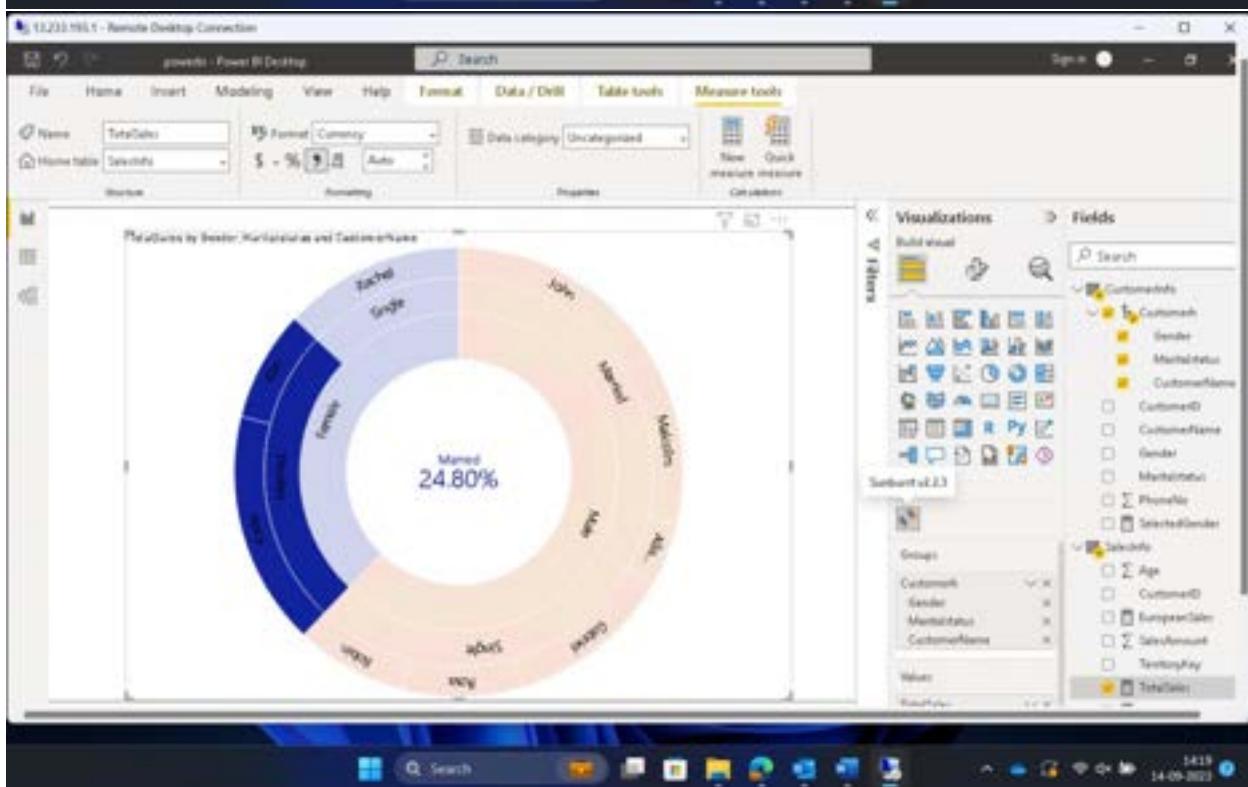
Visualizations Fields

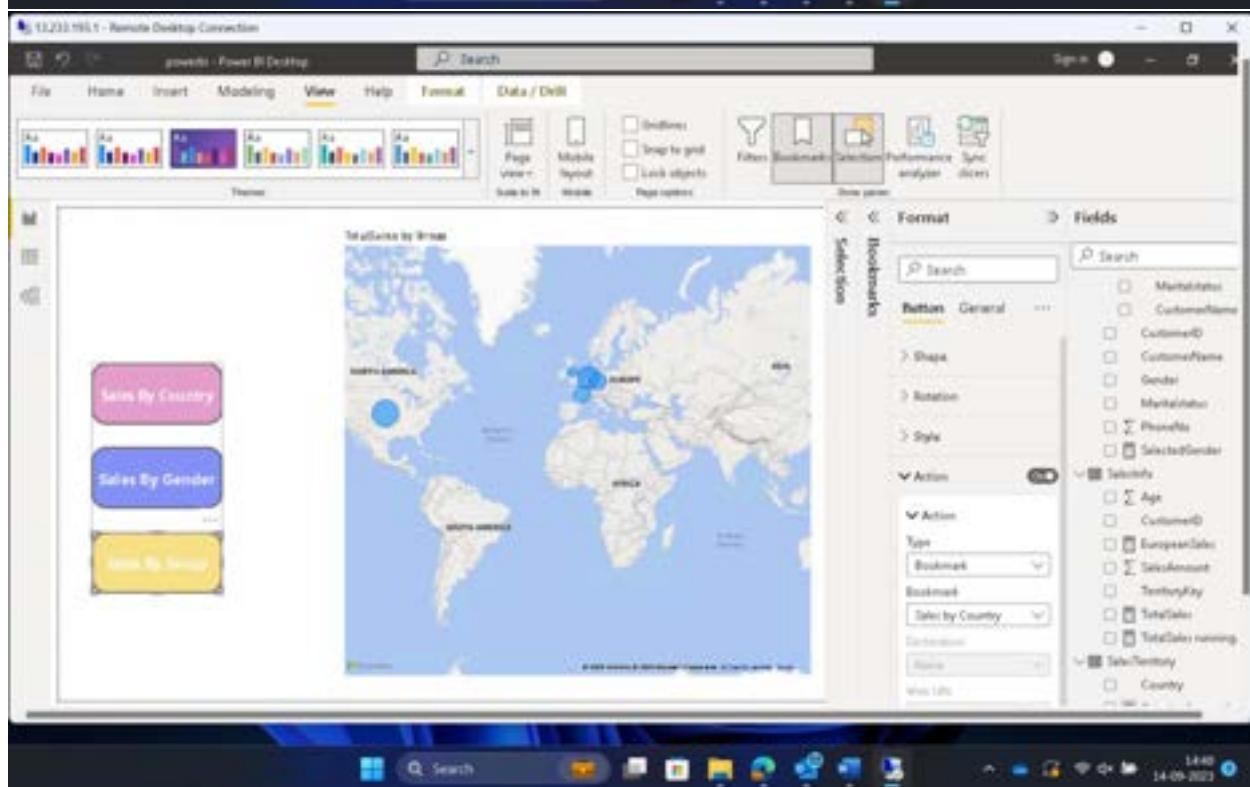
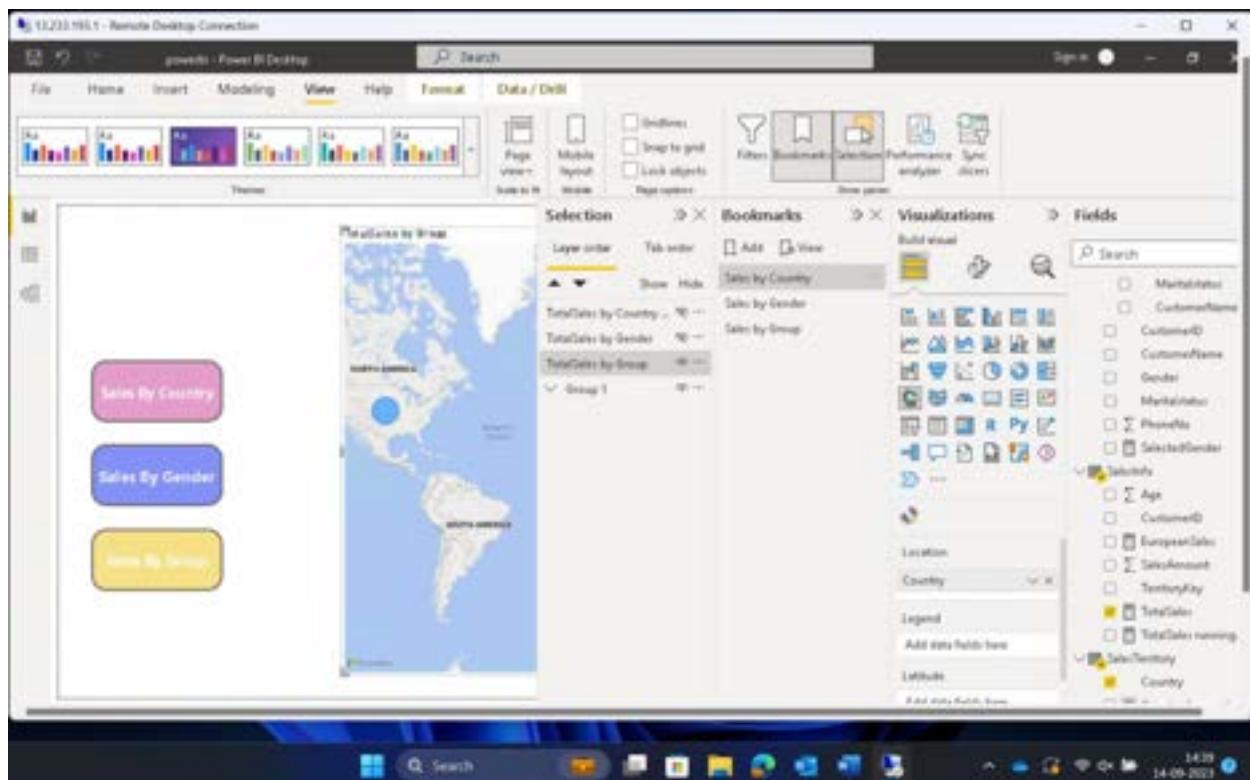
Build visual Search

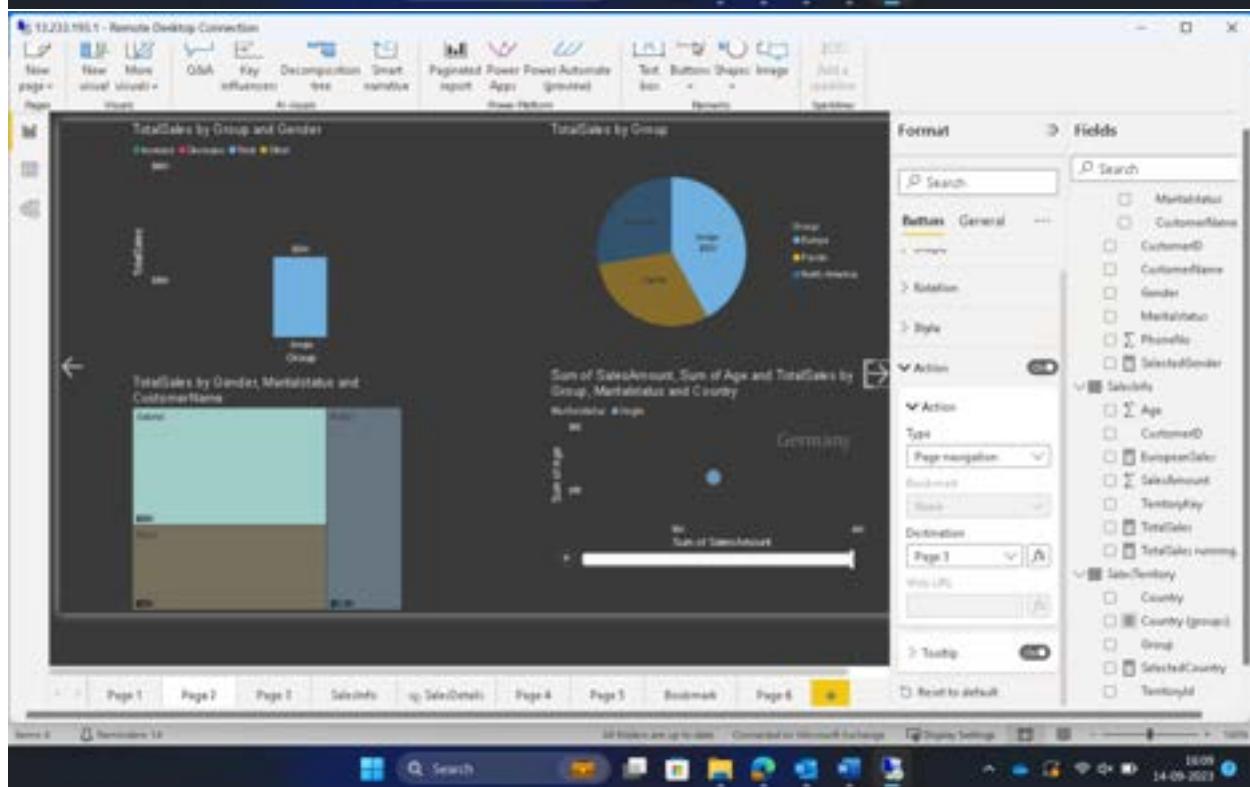
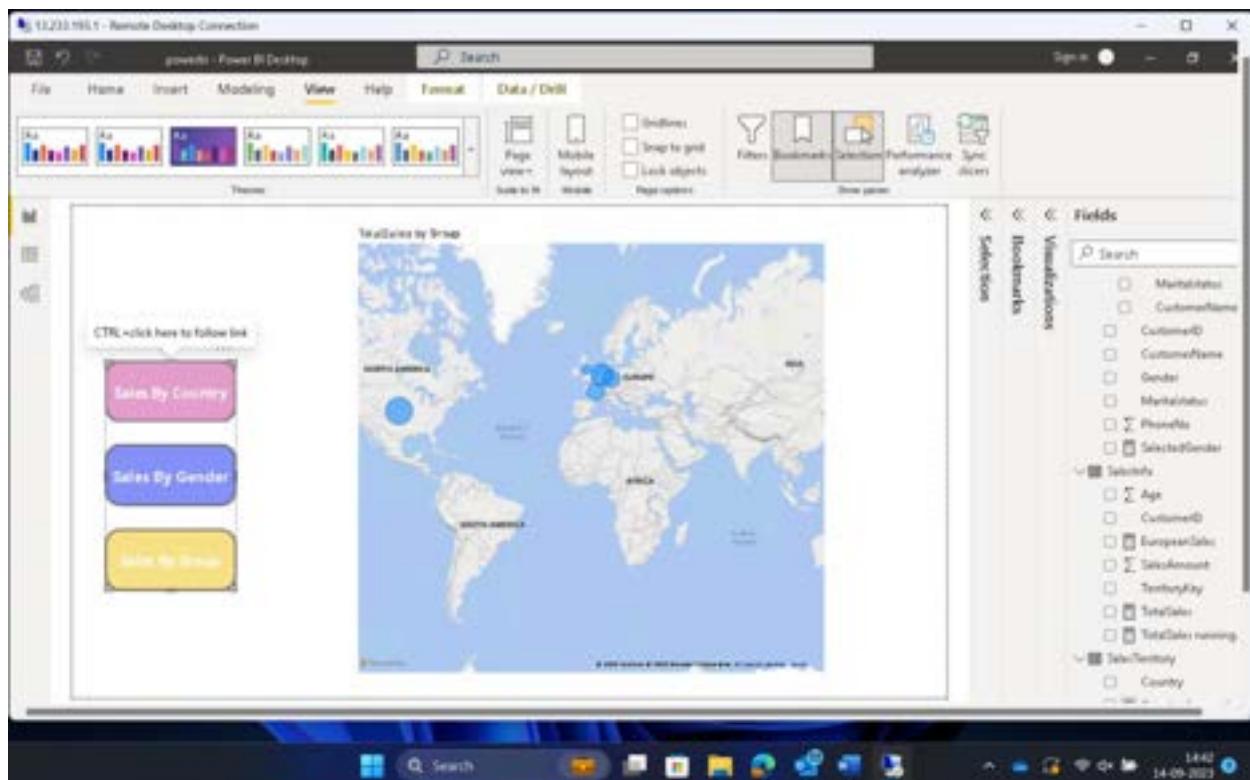
Country Name TotalSales

Australia	John	\$43,400
Australia	Ross	\$1,000
California	Kim	\$2,100
France	Malcolm	\$22,500
Germany	Gabriel	\$12,000
Netherlands	Karla	\$44,500
United Kingdom	Allan	\$12,000
United States	Rachel	\$34,500
United States	Robin	\$21,200
<b>Total</b>		<b>\$264,100</b>

Drill through Cross-report Keep all filters







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 [28]: #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 [7]: 
```

```
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 Jupyter it is like 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 \U0000000X 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)
#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] = "Javascript"
d1

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

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 [50]: 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 pi.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 pi.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)
          a

Out[11]: numpy.ndarray

In [12]: a.ndim
          a

Out[12]: 1

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

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

In [15]: a.shape
          a

Out[15]: (2, 4)

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

In [17]: a.dtype
          a.dtype

Out[17]: dtype('int64')

In [20]: a = np.array(['a','a','b','b','c','c','c'])
          a.dtype #unicode character strings

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

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

Out[22]: dtype('c8')

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)
          a2

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

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

Out[29]: array([-0.41614684, -0.41614684,  0.96017829])

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

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

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

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

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

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

In [32]: np.round(np.cbrt(a),3) #cube root of all elements
          a

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

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

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

In [ ]: np.power

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel) 0

```
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(1) #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 [48]: np.abs(a) # gives absolute values
Out[48]: array([1, 3, 5, 7, 9])

In [49]: np.exp(a) #exponential values
Out[49]: array([2.71828183e+00, 2.00055369e+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
Out[62]: 'aa'

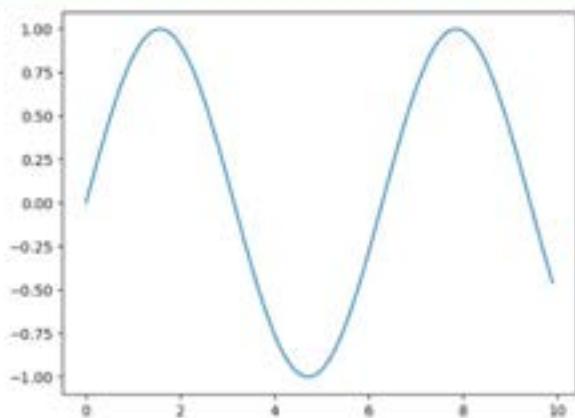
In [64]: np.random.randint(0,10,10) #(start,stop,number of random values) #otp generator
Out[64]: array([8, 8, 1, 9, 7, 9, 5, 8, 9, 9])

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()
```

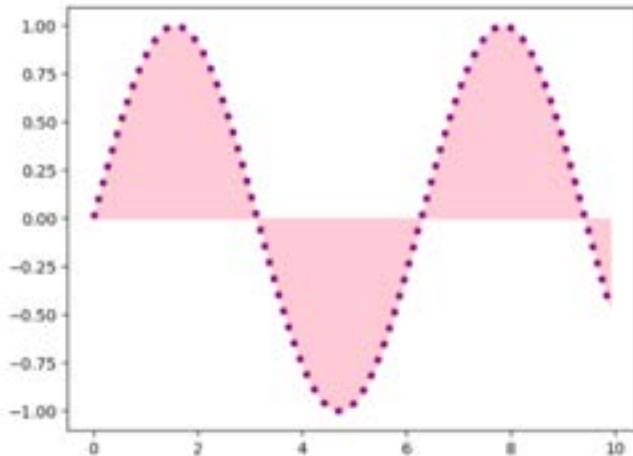


```
In [74]: a = np.array([[23,43,222]])
plt.imshow(a)

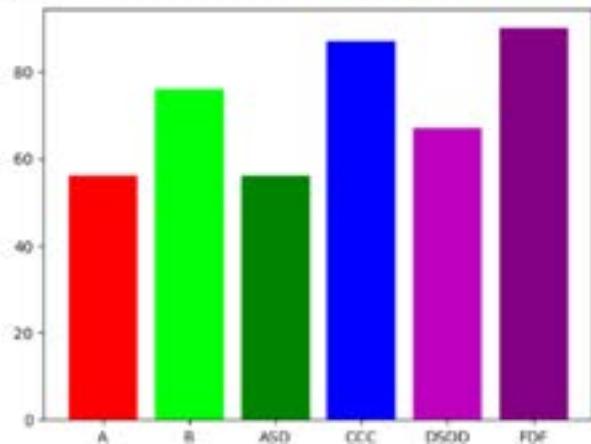
Out[74]: <matplotlib.image.AxesImage at 0x7f7c7bf18310>
```



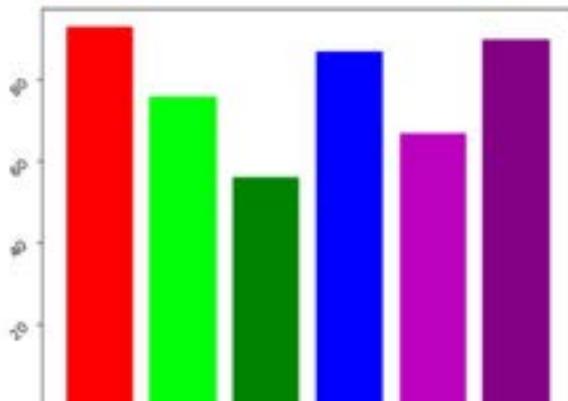
```
In [94]: y = np.sin(x)
plt.plot(x,y,color='purple',linewidth=4,linestyle='dotted')
plt.fill_between(x,y,color='pink',alpha=0.9)
plt.show()
```



```
In [95]: x = ['A','B','ASD','CCC','DSDD','FDF']
y = [56,76,56,87,67,98] #weight in kgs
colors12 = ['r', '#00ffff', 'g', 'b', 'm', 'purple']
plt.bar(x,y,color=colors12)
plt.show()
```



```
In [103]: x = ['A','B','ASD','CCC','DSDD','FDF']
y = [93,76,56,87,67,99] #weight in kgs
colors12 = ['r', '#00ffff', 'g', 'b', 'n', 'purple']
plt.bar(x,y,color=colors12)
plt.xticks(rotation=90)
plt.yticks(rotation=45)
plt.show()
```



```
In [104]: #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 [15]: type(sr1)
```

```
Out[15]: pandas.core.series.Series
```

```
In [16]: d = {'col1': [100,200,300,400,500,600], 'col2': [100,200,300,400,234,500]}
s2 = pd.Series(d)
s2
```

```
Out[16]: col1    [100, 200, 300, 400, 500, 600]
col2    [100, 200, 300, 400, 234, 500]
dtype: object
```

```
In [17]: d = {'col1': [100,200,300,400,500,600], 'col2': [100,200,300,400,234,500]}
s2 = pd.Series(d)
s2
```

```
Out[17]: 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	3684.32400
1	18	male	33.770	1	no	southeast	1725.55200
2	28	male	33.000	3	no	southeast	4493.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3666.89520
..	..	..	..	..	..	..	..
1333	50	male	30.870	3	no	northwest	19600.54800
1334	18	female	31.920	0	no	northwest	2205.58000
1335	18	female	34.850	0	no	southeast	1829.83368
1336	21	female	25.800	0	no	southwest	2007.94600
1337	62	female	29.070	0	yes	northwest	29143.36030

1338 rows × 7 columns

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

```
Out[24]: 9366
```

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

```
Out[25]:
```

```
<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 [26]:
```

```
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	14.049990	8.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287350
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	18639.512515
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	14.049990	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	8.098187	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	southwest	364	NaN	NaN	NaN	NaN	NaN	NaN	NaN
charges	1338.0	NaN	NaN	NaN	13270.422265	12110.011237	1121.8739	4740.28715	9382.033	18639.512515	63770.42801

```
In [29]:
```

```
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[:,[1,5,6,8]]
```

```
Out[37]:
```

```
   bmi  region  charges  age
```

```
0 27.900 southwest 36884.92400 19
```

```
1 33.770 southeast 1725.55230 18
```

```
2 33.000 southwest 4449.46200 28
```

```
3 22.705 northwest 21364.47061 33
```

```
4 28.880 northwest 3868.85520 32
```

```
.. .. .. ..
```

```
1333 30.970 northwest 10600.54830 50
```

```
1334 31.820 northeast 2205.38080 18
```

```
1335 36.850 southeast 1629.83350 18
```

```
1336 29.800 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	25.600	0	yes	southwest	35585.57600
39	18	female	26.115	0	no	northeast	2198.18985
40	19	female	28.600	5	no	southwest	4687.79000
41	63	male	28.310	0	no	northwest	13770.09790
42	28	male	36.400	1	yes	southwest	51184.55814
43	19	male	25.425	0	no	northwest	1625.43375
44	62	female	32.365	3	no	southwest	15612.18325
45	26	male	26.800	0	no	southwest	2362.30000
46	35	male	34.670	1	yes	northeast	39774.27630
47	60	male	31.900	0	yes	southwest	48173.36100
48	24	female	28.600	0	no	northwest	3046.06200
49	31	female	26.620	2	no	southwest	4949.75870

```
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	22.11	0	no	northwest	2130.6758
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	southwest	1725.55230
2	28	male	33.000	3	southwest	4449.46200
3	33	male	22.705	0	northwest	23984.47061
4	32	male	26.890	0	northwest	3866.85520
...	...	...	...	...	...	...
1333	50	male	30.970	3	northwest	10600.54800
1334	18	female	31.920	0	northwest	2205.38080
1335	19	female	36.850	0	southwest	1629.83350
1336	21	female	25.800	0	southwest	2007.34500
1337	61	female	29.070	0	northwest	29141.38030

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'].unique()
```

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

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

```
Out[54]:
```

southeast	364
southwest	325
northwest	325
northeast	324

Name: region, dtype: int64

```
In [ ]:
```

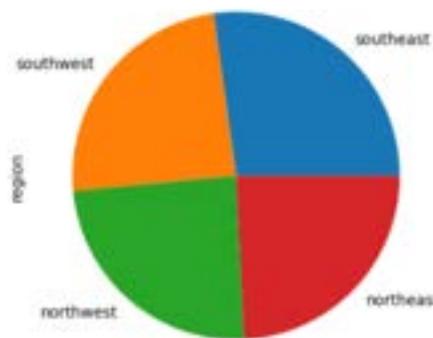
```
In [55]: df['region'].value_counts().plot(kind='bar')
```

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



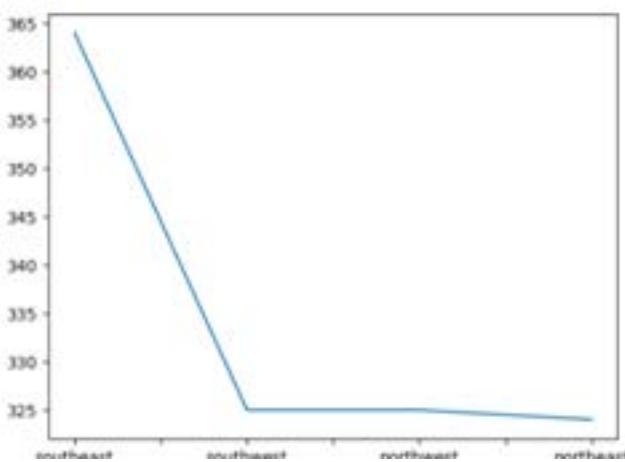
```
In [60]: df['region'].value_counts().plot(kind='pie')
```

```
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	16684.92400
1	18	33.770	1	1725.55230
2	26	33.000	3	4449.46200
3	33	22.705	0	21984.47061
4	32	26.880	0	3866.85820
..	..	..	..	..
1333	52	30.970	3	10600.54830
1334	18	31.920	0	2295.98080
1335	18	36.860	0	1629.83350
1336	21	25.800	0	2007.94900
1337	21	26.800	0	1629.83350

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

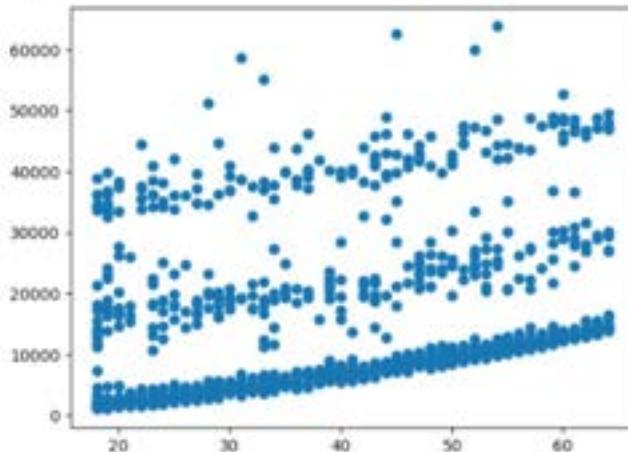
```
plt.plot(df['age'],df['charges'])
```

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



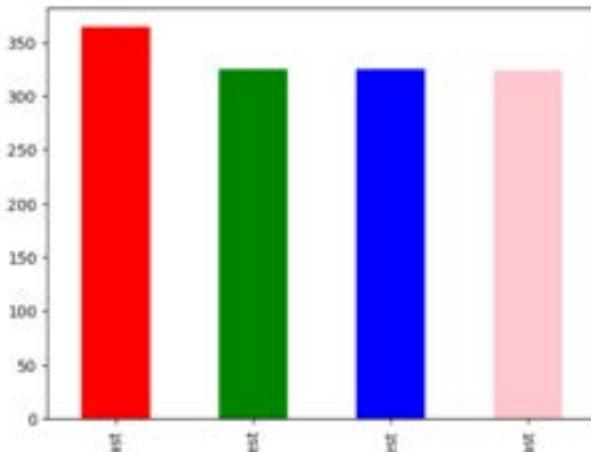
```
In [67]: plt.scatter(df['age'],df['charges'])
```

```
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., 298., 129., 86., 35., 59., 57., 32., 2., 4.]),  
 array([ 1121.8739,  7386.729311, 13651.584722, 19916.448133,  
 26181.295544, 32446.150955, 38711.006366, 44975.861777,  
 51240.717188, 57505.572999, 63770.42801]),  
<matplotlib.container object at 0x000000000000000>)
```



```
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	1829.84300
15	19	male	24.8	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
...	...	...	...	...	...	...	...
1326	19	female	20.6	0	no	southwest	1731.87700
1329	52	male	38.6	2	no	southwest	10325.20600
1331	23	female	33.4	0	no	southwest	10795.93700
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
...	...	...	...	...	...	...	...
1328	36	male	39.710	4	no	northeast	19496.71917
1325	63	male	33.535	0	no	northeast	13143.33665
1326	42	female	32.870	0	no	northeast	7950.02130
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	23.885	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	3732.62550
1290	38	female	19.950	2	no	northeast	7133.90250
1326	42	female	32.870	0	no	northeast	7950.02130
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	48210.20795
319	32	male	37.335	1	0	northeast	4667.60745
844	53	male	30.495	0	0	northeast	10372.05505
317	54	male	32.775	0	0	northeast	15435.08625
315	62	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	1682.58700
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	southwest	14479.33015
1130	39	female	23.870	5	0	southwest	8582.30230
1118	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	5615.36000
—	—	—	—	—	—	—	—
618	29	female	33.120	0	1	southwest	34439.85000
619	55	female	37.100	0	0	southwest	10713.64400
623	18	male	33.520	0	1	northeast	34817.84068
624	59	male	28.785	0	0	northeast	12129.61415

```
In [82]: df.groupby('region').mean()
```

/tmp/ipykernel\_6447/1373985188.py:1: FutureWarning: The default value of numeric\_only in DataFrameGroupBy.mean is deprecated. In a future version, numeric\_only will default to False. Either specify numeric\_only or select only columns which should be valid for the function.  
df.groupby('region').mean()

```
Out[82]:
```

	age	bmi	children	smoker	charges
region					
northeast	39.268519	29.172603	1.046296	0.206790	13406.386518
northwest	39.198823	29.399785	1.147682	0.178462	12417.575374
southeast	38.939640	33.355689	1.049451	0.250000	14735.411438
southwest	39.455385	30.596615	1.141538	0.178462	12346.937377

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

```
Out[83]:
```

	age	bmi	children	smoker	charges	
region	sex					
northeast	female	39.639752	29.324317	1.006211	0.180124	12953.203151
	male	38.901840	29.024540	1.085890	0.233129	13854.006374
northwest	female	39.591463	29.277957	1.115854	0.176829	12479.870397
	male	38.799031	29.120155	1.180124	0.180124	12394.119575
southeast	female	39.108871	32.671257	1.051429	0.205714	13489.680243
	male	38.783069	33.990000	1.047623	0.291005	15879.817173
southwest	female	39.703704	30.060494	1.129457	0.129630	11274.411264
	male	39.208589	31.129448	1.159529	0.220994	13412.883576

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

```
Out[85]:
```

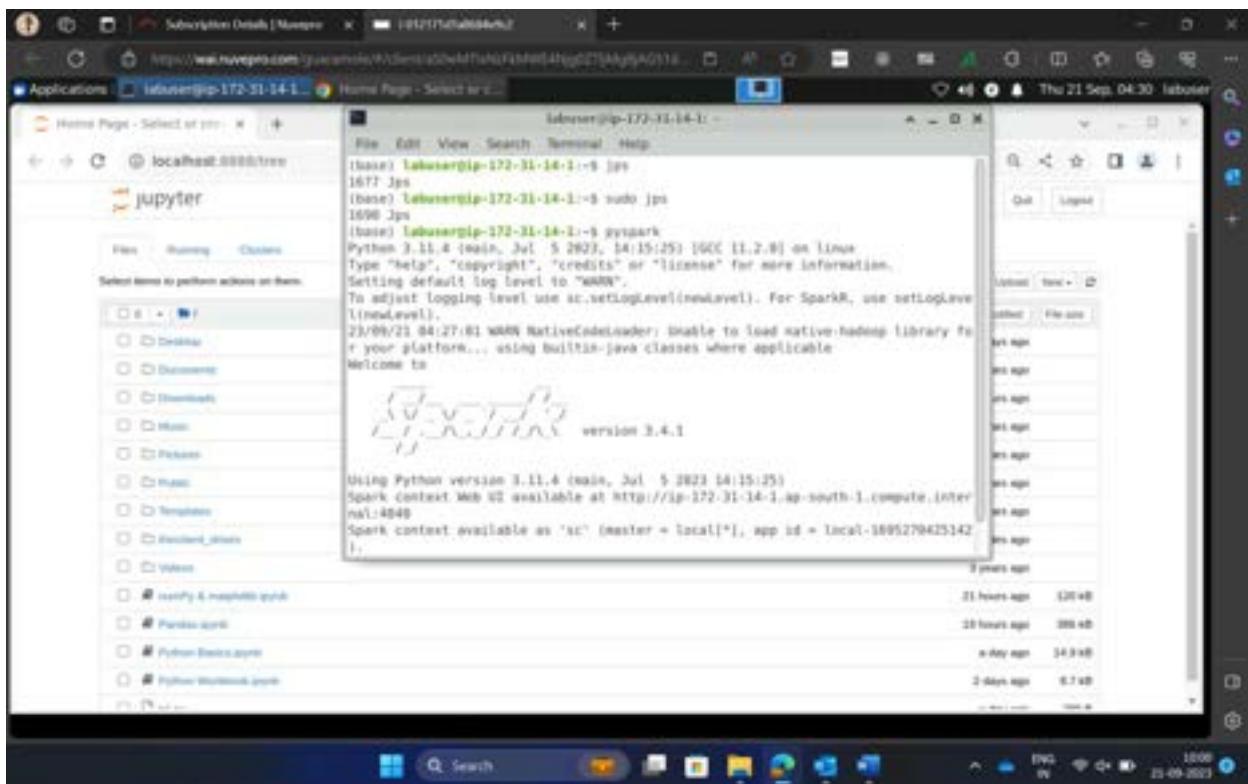
		charges			
		min	max	count	mean
region	sex				
northeast	female	2196.47320	58571.07448	161	12953.203151
	male	1094.79640	48549.17835	163	13854.005374
northwest	female	2117.33885	55135.40209	164	12479.870397
	male	1621.34020	60021.39887	161	12354.129575
southeast	female	1607.51010	63770.42981	175	13499.669243
	male	1121.87390	62592.87309	189	15878.617173
southwest	female	1727.78600	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']})
```

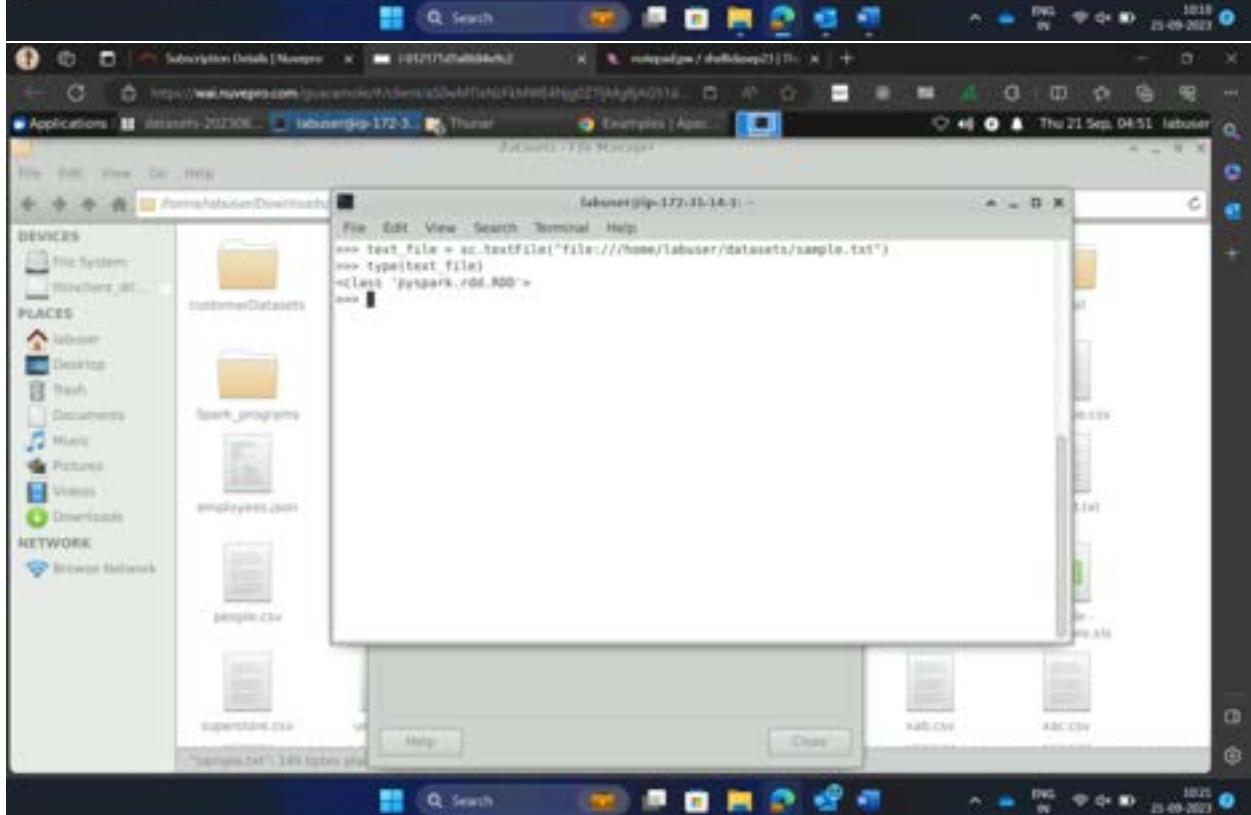
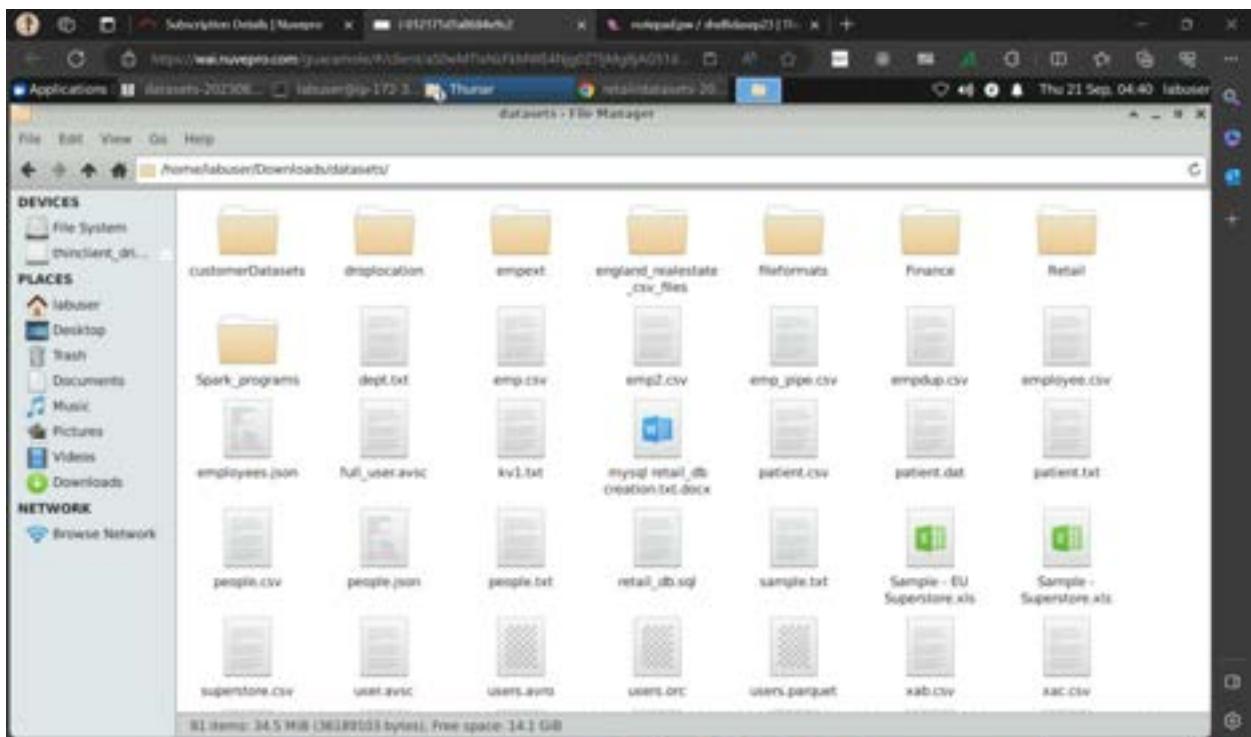
```
Out[86]:
```

		charges				bmi
		min	max	count	mean	mean
region	sex					
northeast	female	2196.47320	58571.07448	161	12953.203051	29.324327
	male	1094.79640	48549.17835	163	13854.005374	29.036540
northwest	female	2117.33885	55135.40209	164	12479.870397	29.277987
	male	1621.34020	60021.39887	161	12354.129575	29.120155
southeast	female	1607.51010	63770.42981	175	13499.669243	32.671257
	male	1121.87390	62592.87309	189	15878.617173	33.990000
southwest	female	1727.78600	48824.45000	162	11274.411264	30.060484
	male	1241.56500	52590.82939	163	13412.883576	31.129448

Day 16



[notepad.pw](http://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()

```

```

Welcome to
version 3.4.1

Using Python version 3.11.4 (main, Jul 5 2023 14:15:29)
Spark context Web UI available at http://ip-172-31-14-1.compute.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1695274517348).
SparkSession available as 'spark'.
...
>>> text_file = textFile("file:///home/labuser/Downloads/datasets/sample.txt")
>>> type(text_file)
<class 'pyspark.rdd.RDD'>
>>> counts = text_file.flatMap(lambda line: line.split(" ")).map(lambda word: (word, 1)).reduceByKey(
lambda a, b: a + b)
>>> counts.collect()
[('done', 3), ('file', 3), ('for', 3), ('wordcount', 3), ('annika', 2), ('Hello', 1), ('this', 1), ('day', 1), ('is', 1), ('in', 1), ('spark', 1), ('a', 1)]
>>>

```

File Edit View Search Terminal Help

using [PySpark](#) for this basic word application

Devices

- File System
- Network

Places

- Labuser
- Desktop
- Trash
- Documents
- Musics
- Pictures
- Videos
- Downloads

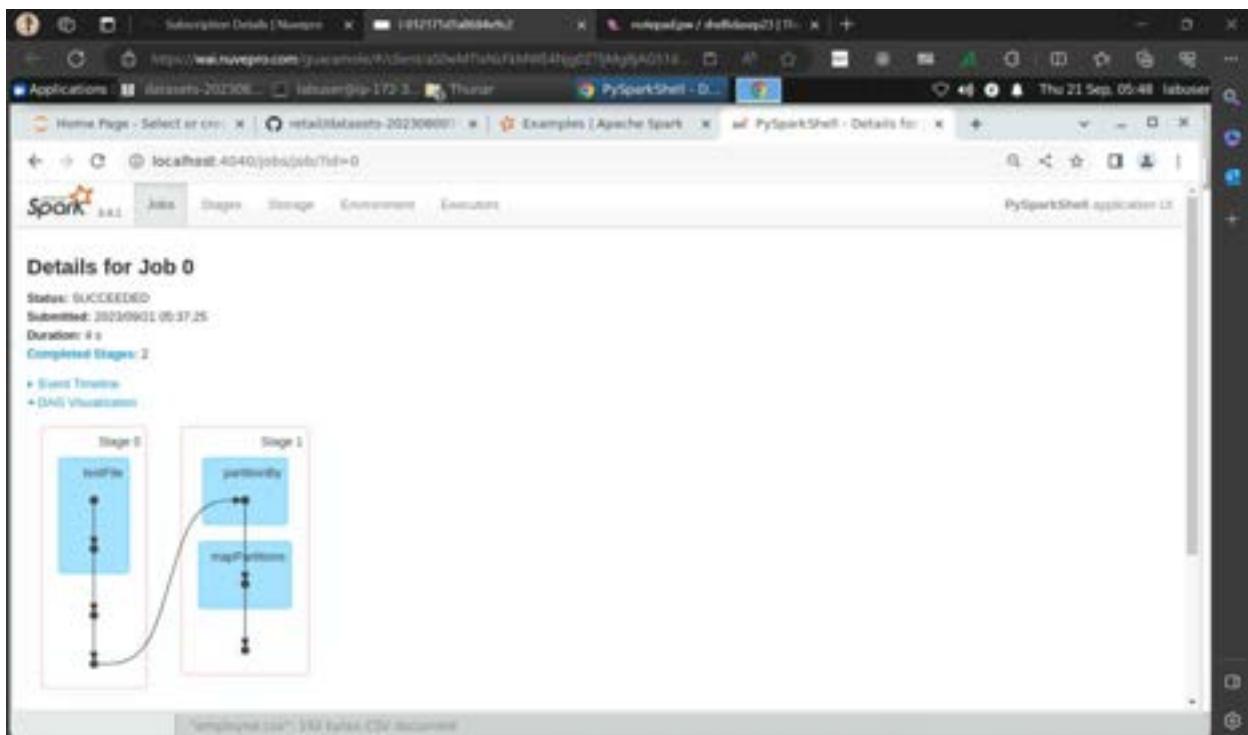
Network

Browse network

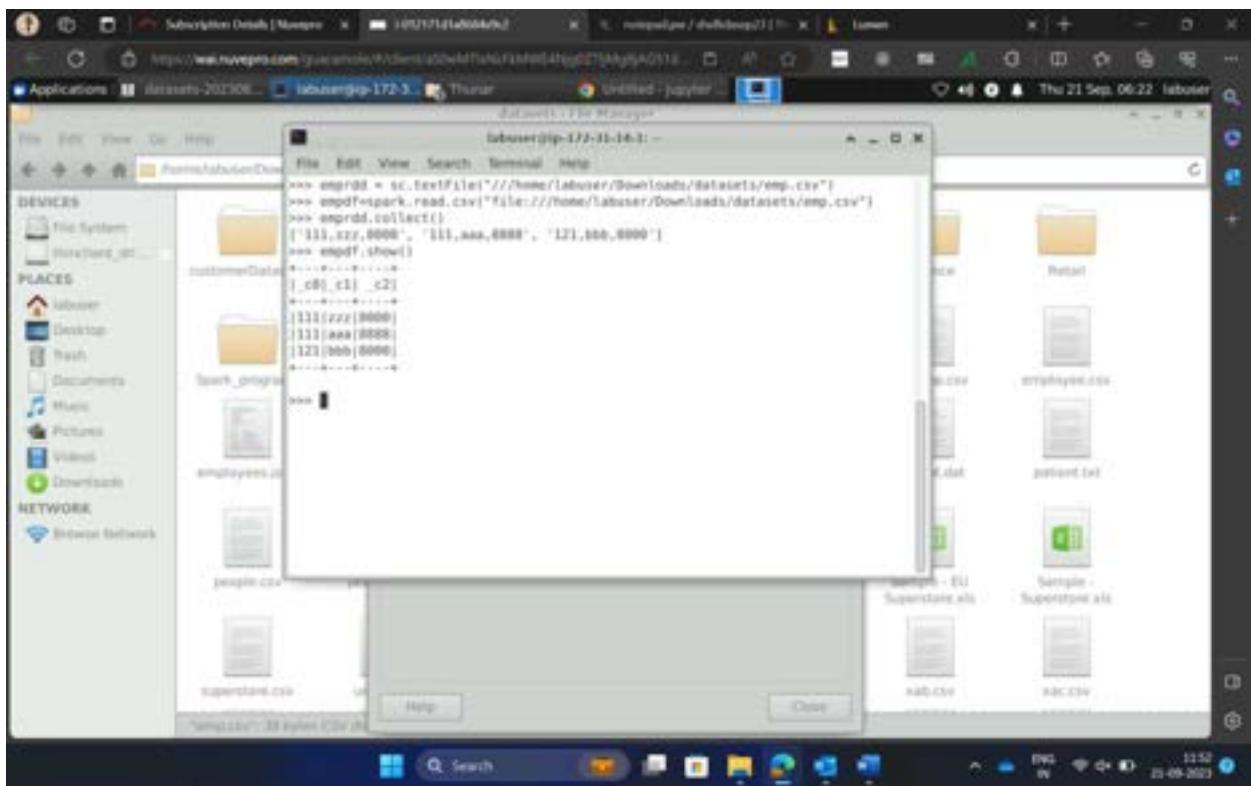
python.csv people.json people.txt reddit\_dbs.json sample.txt Sample - EU Superstore.xls Sample - Superstore.xls

superstore.csv user\_avsc users.avsc users.json users.parquet user.csv SAC.csv

Job ID	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	counts at <stdin>	2023/09/21 05:37:25	4 s	3/3	6/6



```
Using Python version 3.11.4 (main, Jul  3 2023 16:15:25)
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-1695274513348).
SparkSession available as 'spark'.
>>> 
>>> text_file = sc.textFile("file:///home/labuser/Downloads/datasets/sample.txt")
>>> type(text_file)
<class 'pyspark.rdd.RDD'>
>>> counts = text_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), ('is', 3), ('spark', 3), ('in', 2), ('21.09.23', 1)),]
>>> words = text_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', 'this', 'day', '16', '21.09.23', '.', '.')]
>>> mappedword.collect()
[('demo', 1), ('file', 1), ('for', 1), ('wordcount', 3), ('in', 3), ('spark', 3), ('this', 1), ('day', 1), ('16', 1), ('.', 2), ('.', 1), ('spark', 3), ('21.09.23', 1), ('.', 1), ('wordcount', 3), ('16', 1), ('.', 1), ('spark', 3), ('.', 1), ('annika', 2), ('hello', 1), ('this', 1), ('is', 1), ('annika', 1), ('day', 1), ('16', 1), ('21.09.23', 1), ('.', 1), ('.', 1), ('.', 1)]
>>> counts.collect()
[(('demo', 3), ('file', 3), ('for', 3), ('wordcount', 3), ('annika', 2), ('hello', 1), ('this', 1), ('day', 1), ('16', 1), ('.', 2), ('is', 3), ('spark', 3), ('in', 2), ('21.09.23', 1))]
```



The screenshot shows a Jupyter Notebook interface with several code cells and their corresponding outputs. The cells include:

- 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)`.  
23/05/21 06:25:48 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jar  
vs classes where applicable  
23/05/21 06:25:50 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
- In [8]: `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|8888|
|111|aaa|8888|
|121|bbb|8888|
+---+---+---+

```
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| ProductCategory| 
| State| Country|PostalCode|Market| Region| ProductID| Category|Sub-Category| ProductName|
| me| Sales|Quantity|Discount| Profit|ShippingCost|OrderPriority|
+-----+
| 32298| CA-2012-124891|21/07/2012|31/07/2012| Same Day| RH-39495| 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| JM-16210| Justin Ritter| Corporate| Wollongong| 
| New South Wales| Australia| null| APAC| Oceania| FUR-CH-10003058| Furniture| Chairs|Novimex Executive...| 
| 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| 
| Queensland| Australia| null| APAC| Oceania| TEC-PH-10004664| Technology| Phones| Nokia Smart Phone| 
| with Caller ID|5175.371| 9| 0.1| 919.971| 915.49| 
| 13524|ES-2013-1579342|28/01/2013|28/01/2013| First Class| KM-16375|Katherine Murray|Home Office| Berlin| 
| Berlin| Germany| null| EU| Central| TEC-PH-10004583| Technology| Phones|Motorola Smart Phone| 
| 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| 222|6000|
| 111| aaa|8888|
| 121| bbb|8000|
| NULL| ccc|9000|
| false| ddd|6000|
| 555| eee|7000|
| 765| null|null|
| 666| fff|8890|
| aaa| null|null|
| True| null|null|
| true| null|null|
| a| b| c|
| x| NULL|NULL|
| 1800| 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('ename', StringType(), True),
    StructField('salary', DoubleType(), True)
])

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

In [26]: employeedf.printSchema()
root
 |-- empno: integer (nullable = true)
 |-- ename: string (nullable = true)
 |-- salary: double (nullable = true)
```

In [27]:

```
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/labuser/Downloads/datasets/employee.csv")

```

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

	(Empno Emplname salary)
1	111 zzz 8000.0
2	111 aaa 8000.0
3	121 bbb 8000.0
4	555 eee 7000.0
5	666 fff 8000.0

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

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (system) O

```
In [40]: Schema: Empno, Emplname, salary
Expected: Emplname but found: emname
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/labuser/Downloads/datasets/employee.csv")
```

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

```
23/09/21 09:15:17 WARN CSVHeaderChecker: CSV header does not conform to the schema.
  Header: empno, emname, sal
  Schema: Empno, Emplname, salary
Expected: Emplname but found: emname
CSV file: file:///home/labuser/Downloads/datasets/employee.csv
23/09/21 09:15:17 ERROR Executor: Exception in task 0.0 in stage 26.0 (TID 220)
org.apache.spark.SparkException: IMALFORMED_RECORD_IN_PARSING Malformed records are detected in record parsing
  (null,ccc,8000.0).
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(QueryExecutionErrors.scala:1766)
  at org.apache.spark.sql.catalyst.util.FailureSafeParser.parse(FailureSafeParser.scala:69)
  at org.apache.spark.sql.catalyst.csv.UnivocityParser$1.nextToken$1$nextCurIterator(UnivocityParser.scala:456)
  at scala.collection.Iterator$anon$1$1.nextToken$1$nextCurIterator$1(Iterator.scala:486)
  at scala.collection.Iterator$anon$1$1.nextToken$1$nextCurIterator$1(Iterator.scala:492)
  at scala.collection.Iterator$anon$1$1.nextToken$1$nextCurIterator$1(Iterator.scala:460)
  at org.apache.spark.sql.execution.datasources.FileScanRDD$ScanObject$1.nextToken(FileScanRDD.scala:125)
```

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

```
Out[44]: 51290
```

```
In [45]: salesdf.head(5) #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', CustomerID='RH-19495', CustomerName='Rick Hansen', Segment='Consumer', City='New York City', State='New York', Country='United States', PostalCode=10024, Market='US', Region='East', ProductID='TEC-AC-10003033', Category='Technology', SubCategory='Accessories', ProductName='Plantronics CS510 - Over-the-Head monaural Wireless Headset System', Sales=2399.65, 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| New York|United States| 10024| US| East|TEC-AC-10003033|Technology| Accessories|Plantronics CS510...|2309.65| | | | | |
| 7| 0| 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]: RowID=32298, OrderID='CA-2012-124891', OrderDate='31/07/2012', ShipDate='31/07/2012', ShipMode='Same Day', CustomerID='RH-19495', CustomerName='Rick Hansen', Segment='Consumer', City='New York City', State='New York', Country='United 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='2309.65', Quantity='7', Discount='0', Profit='762.1845', ShippingCost=' 933.57 ', OrderPriority='Critical')
```

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

```
Out[49]: [RowID=32298, OrderID='CA-2012-124891', OrderDate='31/07/2012', ShipDate='31/07/2012', ShipMode='Same Day', CustomerID='RH-19495', CustomerName='Rick Hansen', Segment='Consumer', City='New York City', State='New York', Country='United 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='2309.65', Quantity='7', Discount='0', Profit='762.1845', ShippingCost=' 933.57 ', OrderPriority='Critical'),
RowID=26341, OrderID='IN-2013-77878', OrderDate='05/02/2013', ShipDate='07/02/2013', ShipMode='Second Class', CustomerID='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='Furniture', Sub-Category='Chairs', ProductName='Novimex Executive Leather Armchair', Sales=' Black', Quantity='3709.395', Discount='9', Profit='0.1', ShippingCost='-288.765', OrderPriority=' 923.63 ')]
```

```
In [50]:
```

```
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  8.5
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
```

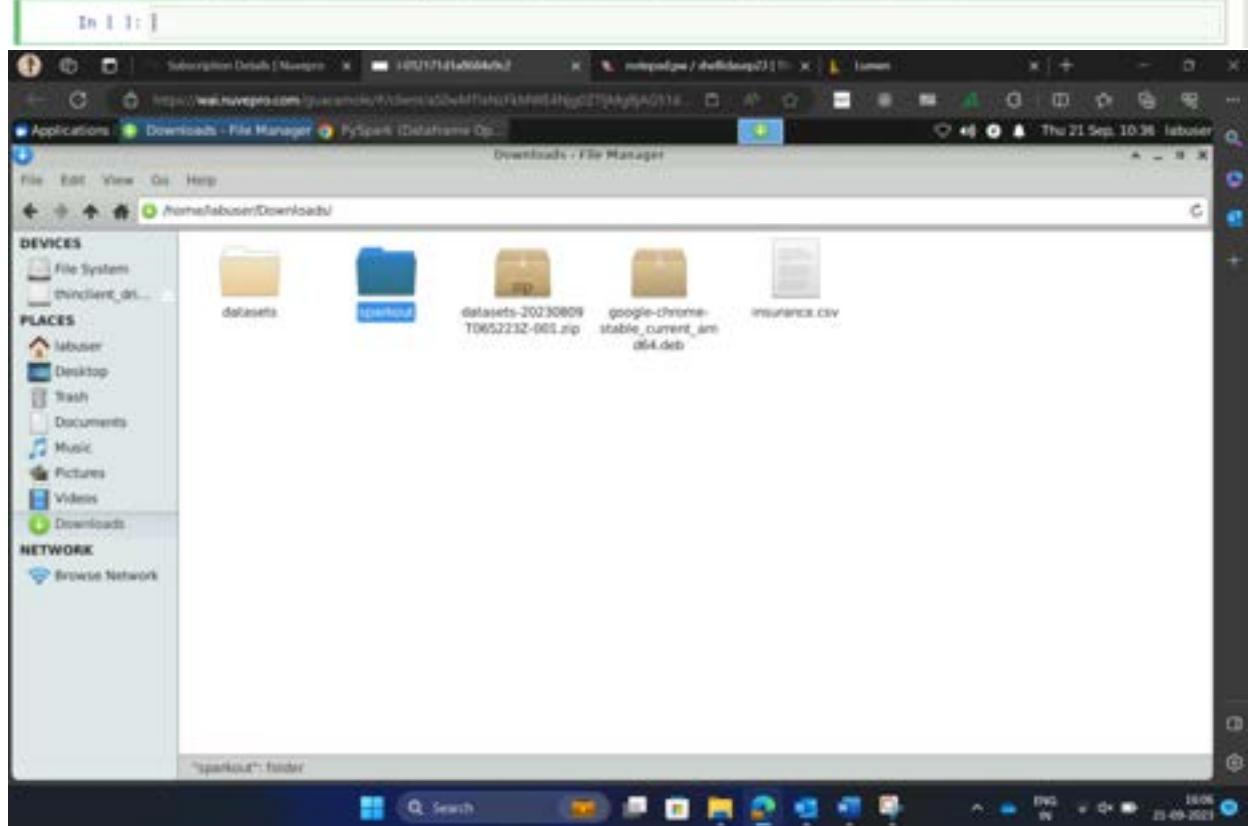
```
Out[63]: DataFrame{country: string, state: string, category: string, profit: string}
```

```
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:4840'

In [68]: #to write into the file after the following transformations
sales_india_df.write.format("json").save("/home/labuser/Downloads/sparkout/sales_india")
```



Day 17 (22.09.23)

```
Subscription Details | X https://wwwspark.com/guviemail/client/submit/PySpark/PySpark0118 | Applications | wordcount.py | labuser@ip-172-31-14-1 | labuser - File Manager Home Page - Sola... | spark submit | Search | Submitting Application | + | - | X | ... |
```

File Edit View Search Terminal Help

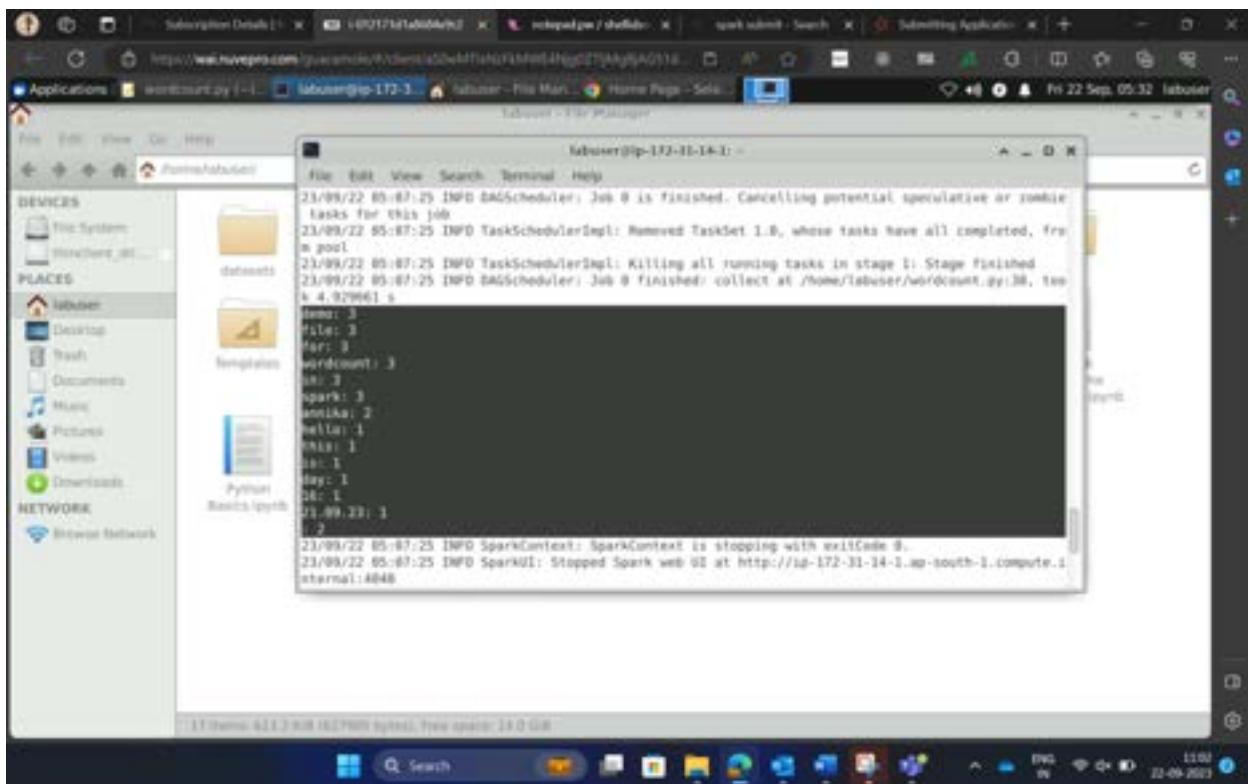
```
(base) labuser@ip-172-31-14-1:~$ spark-submit
Usage: spark-submit [options] <app jar | python file> [<app arguments>]
Usage: spark-submit --kill [submission ID] --master [spark://...]
Usage: spark-submit --status [submission ID] --master [spark://...]
Usage: spark-submit run-example [options] example-class [example args]

Options:
  --master MASTER_URL          spark://host:port, mesos://host:port, yarn,
                                k8s://https://host:port, or local (Default: local[1])
  --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 spec).
  --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 repos, 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.
```

```
Subscription Details | X https://wwwspark.com/guviemail/client/submit/PySpark/PySpark0118 | Applications | wordcount.py | labuser@ip-172-31-14-1 | labuser - File Manager Home Page - Sola... | spark submit | Search | Submitting Application | + | - | X | ... |
```

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-3a79c499-8bf8-4c2a-85d6-4500e20305b8
(base) labuser@ip-172-31-14-1:~$ spark-submit wordcount.py /home/labuser/datasets/sample.txt
23/09/22 05:03:04 INFO ShutdownHookManager: Shutdown hook called
23/09/22 05:03:04 INFO ShutdownHookManager: Deleting directory /tmp/spark-32c7abec-864c-4236-9514-fad91c06306b
(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 built-in 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<String -> Map<String, ExecutorResource>@63333333, task resources: Map<String -> Map<String, TaskResource>@63333333
23/09/22 05:07:02 INFO ResourceProfile: Executor resources: Map<String -> ExecutorResource@63333333
23/09/22 05:07:02 INFO ResourceProfile: Task resources: Map<String -> TaskResource@63333333
23/09/22 05:07:02 INFO ResourceProfileManager: Limiting resource is cpu
23/09/22 05:07:02 INFO ResourceProfileManager: Added ResourceProfile@8: 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 [56]: #salesdf.show(2)
```

```
In [55]: #writing data as a parquet file
salesdf_repart.write.parquet("/home/labuser/output/sales_report")
[Stage 55:----- (1 + 1) / 2]
```

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

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

```
In [58]: sales_parquet_1 = spark.read.format("parquet").load("/home/labuser/output/sales_report/part-00000-842af62-ec43-4d3c-27a755d1a3-000.snapy.parquet")
```

```
In [59]: sales_parquet_1.count()
Out[59]: 18259
```

```

In [64]: #to create the parquet file
sales_parquet = spark.read.format("parquet").load("/home/labuser/output/sales_report")

In [65]: sales_parquet_1 = spark.read.format("parquet").load("/home/labuser/output/sales_report/part-00000-842af652-ec43-4d30-9e0c-1a2a2f3a2300.parquet")
+-----+
| (1, 1) |
+-----+
| count |
+-----+
| 10259 |
+-----+  

In [66]: sales_parquet_1.count()

Out[66]: 10259

In [67]: sales_orc=spark.read.format("orc").load("/home/labuser/output/sales_report/*")

In [68]: sales_orc.count()

Out[68]: 51290

In [69]: sales_orc.rdd.getNumPartitions()

Out[69]: 2

In [70]: sales_coal_df = sales_orc.coalesce(1)

+-----+
| count |
+-----+
| 51290 |
+-----+  

In [71]: sales_orc.rdd.getNumPartitions()

Out[71]: 2

In [72]: sales_coal_df = sales_orc.coalesce(1)

+-----+
| count |
+-----+
| 51290 |
+-----+  

In [73]: sales_coal_df.rdd.getNumPartitions()

Out[73]: 1

In [74]: sales_rep = sales_orc.repartition(1)

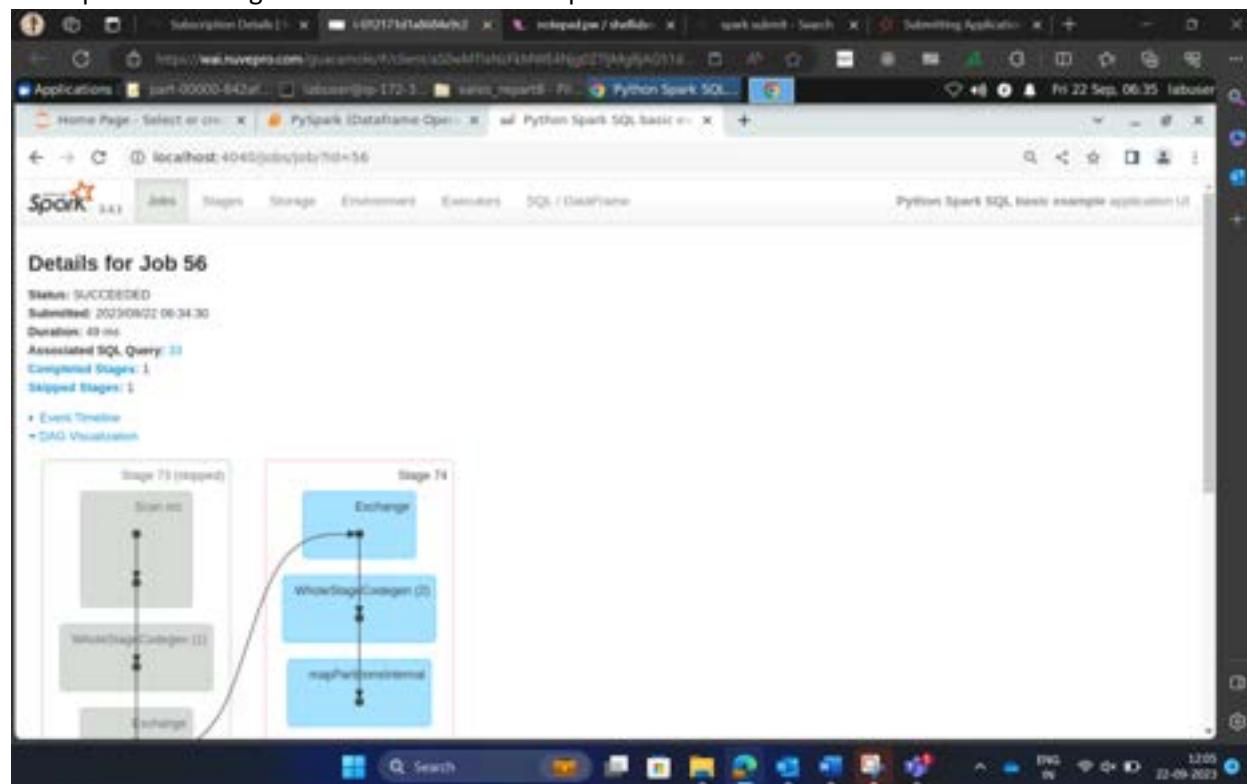
+-----+
| repartition |
+-----+
| 1 |
+-----+  

In [75]: sales_rep.rdd.getNumPartitions()

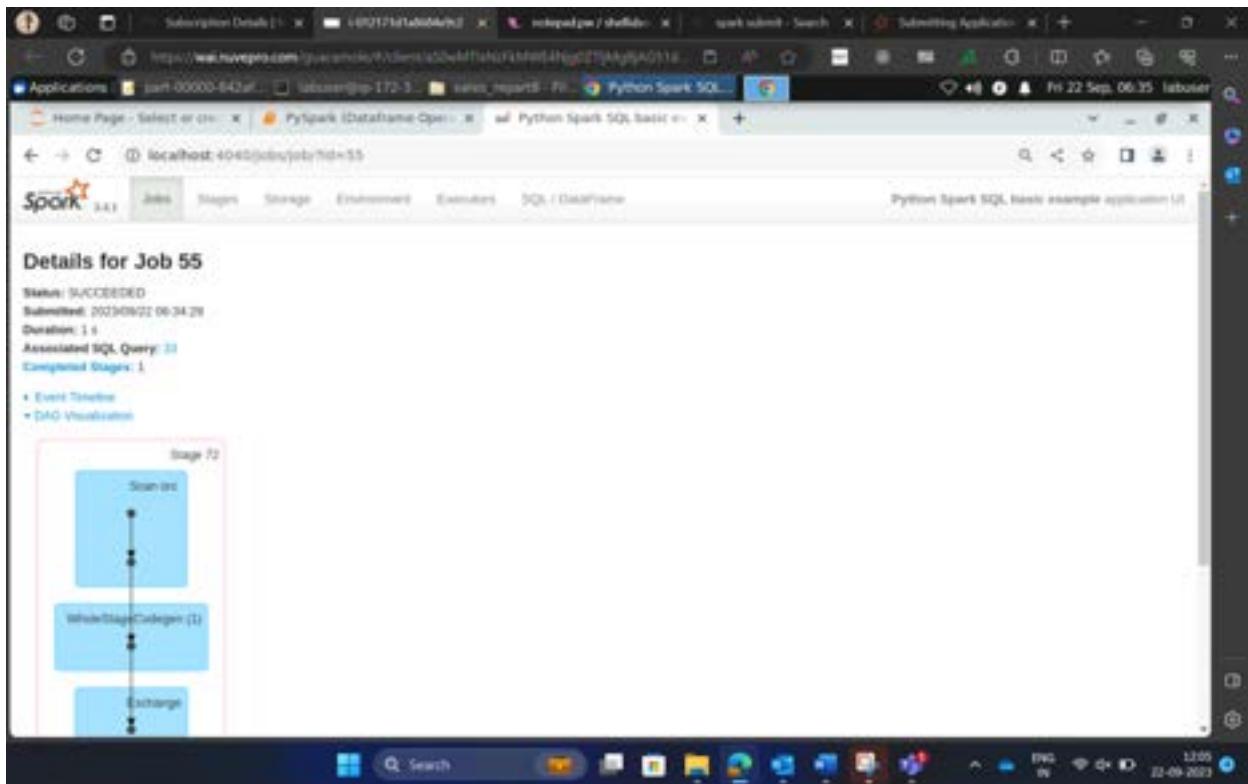
Out[75]: 1

```

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	New South Wales	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) / 23

+	-	-	-	-
*	-----*	-----*	-----*	*
	country	sum(profit)	-----	*
	-----	-----	-----	*
Chad	0.0			
Russia	0.0			
Paraguay	0.002			
Yemen	20.99999999999996			
Senegal	0.0			
Sweden	103.18000000000003			
Philippines	235.55			
Eritrea	0.0			
Djibouti	0.0			
Malaysia	0.0			
Singapore	0.0			
Turkey	826.800000000005			
Iraq	0.0			
Germany	117.799999999997			
Afghanistan	0.0			
Cambodia	0.0			
Rwanda	0.0			
Jordan	0.0			
Sudan	0.0			
France	204.3499999999903			

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_pern")

In [82]: spark.sql("show tables").show()
+-----+-----+
|namespace| tableName|isTemporary|
+-----+-----+
| shellida|sales_pern|    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|
|113|aaa|8000|
|121|bbb|8000|
+---+---+


In [88]: salesdf.cache()
Out[88]: DataFrame[ID: int, OrderID: string, OrderDate: string, ShipDate: string, ShipMode: string, CustomerID: string, CustomerName: string, Segment: string, City: string, State: string, Country: string, PostalCode: int, Market: string, Region: string, ProductID: string, Category: string, Sub-Category: string, ProductName: string, Sales: string, Quantity: string, Discount: string, Profit: string, ShippingCost: string, OrderPriority: string]

In [90]: salesdf.count()
Out[90]: 51290

In [11]: |
```

Applications | labuser@ip-172-31-14-1: ~ | kudu\_inserts - File Manager | Python Spark SQL basic example application 1.0 | Fri 22 Sep, 08:20 labuser

localhost:4040/storage/

**Storage**

+ RDDs

ID	RDD Name	Storage Level	Cached Partitions	Fraction Cached	Size in Memory	Size on Disk
306	FileScan(rdd_0)	Disk, Memory, Deserialized, In Replicated	2	100%	5.5 MB	0.0 B

Submitted: 2023/09/02 08:19:37  
Duration: 15 ms  
Associated SQL Query: [#7](#)  
Completed Stages: 1  
Staged Stages: 1

+ Event Timeline  
+ DAG Visualization

```

graph TD
    subgraph Stage0 [Stage 0 (prepared)]
        S0_Scan[Scan rdd]
        S0_Unlambda[UnlambdaToPairFunc]
        S0_Scan --> S0_Unlambda
    end
    subgraph Stage1 [Stage 1]
        S1_Exchange[Exchange]
        S1_WSCG[WholeStageCodegen (2)]
        S1_MP[mapPartitionsInternal]
        S1_Exchange --- S1_WSCG
        S1_WSCG --- S1_MP
    end
    S0_Scan --> S1_Exchange
  
```

```
In [91]: espdf.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_client,
   1, target_id, name)
   2 value = OUTPUT_CONVERTER[type](answer[2:], gateway_client)
   3 if answer[1] == REFERENCE_TYPE:
--> 326     raise Py4JJavaError(
   327         "An error occurred while calling %s(%s).%s.\n%s",
   328         format(target_id, ".", name), value)
   329 else:
   330     raise Py4JError(
   331         "An error occurred while calling %s(%s[%s]). Trace:\n%s\n%s",
   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 recent 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
Sale	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-2013-77870	05/02/2013	07/02/2013	Second Class	JR-16210	Justin Ritter	Corporate	Wollongong N
new South Wales	Australia	null	APAC	Oceania	FUR-CH-10003950	Furniture	Chairs	Movianex E
executive...	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
Queensland	Australia	null	APAC	Oceania	FRU-AU-10003950	Books	Books	Booktopia
10...	2309.65	7	0	762.1845	933.57	Critical		

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

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

```
In [183]: salesdf.unpersist()
Out[183]: DataFrame[OrderID: int, OrderID: string, OrderDate: string, ShipDate: string, ShipMode: string, CustomerID: string, CustomerName: string, Segment: string, City: string, State: string, Country: string, PostalCode: int, Market: string, Region: string, ProductID: string, Category: string, Sub-Category: string, ProductName: string, Sales: string, Quantity: string, Discount: string, Profit: string, ShippingCost: string, OrderPriority: string]

In [184]: from pyspark.storagelevel import *

In [185]: salesdf.persist(StorageLevel.MEMORY_AND_DISK_2)
Out[185]: DataFrame[OrderID: int, OrderID: string, OrderDate: string, ShipDate: string, ShipMode: string, CustomerID: string, CustomerName: string, Segment: string, City: string, State: string, Country: string, PostalCode: int, Market: string, Region: string, ProductID: string, Category: string, Sub-Category: string, ProductName: string, Sales: string, Quantity: string, Discount: string, Profit: string, ShippingCost: string, OrderPriority: string]

In [186]: salesdf.show()
[Stage 106: 0 + 1] / 1
```

Spark 2.2.0

Jobs Stages Storage Environment Executors SQL / DataFrame Python Spark SQL basic example application 1.0

### Storage

- RDDs

ID	RDD Name	Storage Level	Cached Partitions	Fraction Cached	Size in Memory	Size on Disk
300	FileScan inv	Disk Memory Serialized 2x Replicated	1	50%	2.6 MB	0.0 B
	[OrderID, 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: None, DataFilters: [], Format: CSV, Location: InMemoryFileIndex[paths@file:///home/hadoop/tables/experiments.csv], PartitionFilters: [], PushedFilters: [], ReadSchema: struct<of int> OrderID:int, OrderDate:string, ShipDate:string, ShipMode:string, CustomerID:int, CustomerName:string, Segment:string, City:string, State:string, Country:string, PostalCode:int, Market:string, Region:string, ProductID:string, Category:string, Sub-Category:string, ProductName:string, Sales:string, Quantit...					

Subscription Details | F1217Results | https://www.pyspark.com | spark storage | Storage View | Remote History | + | - | X | ...

Applications Terminal PySpark (DataFrames API) Home Page - Select or open a notebook | Localhost | LabVIEW (ip-172-21-18-1) | File Edit View Search Terminal Help

jupyter

```

In [1]: empDF = spark.createDataFrame(data=emp, schema = EmpSchema)
dept = [{"Finance":10}, {"Marketing":20}, {"Sales":30}, {"IT":40}]

In [2]: deptColumns = ["dept_name","dept_id"]
deptDF = spark.createDataFrame(data=dept, schema = deptColumns)
empDF.join(deptDF, empDF.dept_id == deptDF.dept_id, "inner").show()

In [3]: empDF.join(deptDF, empDF.dept_id == deptDF.dept_id, "inner").select("Emp_id", "Empname", "MGR", "Y0J", "dept_id", "gender", "salary", "dept_name", "dept_id").show()

In [4]: empDF.join(deptDF, empDF.dept_id == deptDF.dept_id, "full").show()

In [5]: empDF.join(deptDF, empDF.dept_id == deptDF.dept_id, "right").show()

```

23/09/22 09:22:13 ERROR Executor: Exception in task 0.0 in stage 117.0 (TID 148)
org.apache.spark.api.python.PythonException: Traceback (most recent call last):
 File "/opt/miniconda3/lib/python3.11/site-packages/pyspark/python/lib/pyspark\_worker.py", line 483, in main
 raise RuntimeError
RuntimeError: Python in worker has different version 3.8 than that in driver 3.11. PySpark cannot run with different Python versions.

23/09/22 09:22:13 INFO Executor: Exception in task 0.0 in stage 117.0 (TID 148)
at org.apache.spark.executor.Executor\$TaskRunner.run(Executor.scala:357)
at java.util.concurrent.ThreadPoolExecutor\$Worker.runTask(ThreadPoolExecutor.java:1149)
at java.util.concurrent.ThreadPoolExecutor\$Worker.run(ThreadPoolExecutor.java:624)
at java.lang.Thread.run(Thread.java:798)

In [1]: empDF = spark.createDataFrame(data=emp, schema = EmpSchema)
dept = [{"Finance":10}, {"Marketing":20}, {"Sales":30}, {"IT":40}]

In [2]: deptColumns = ["dept\_name","dept\_id"]
deptDF = spark.createDataFrame(data=dept, schema = deptColumns)
empDF.join(deptDF, empDF.dept\_id == deptDF.dept\_id, "inner").show()

In [3]: empDF.join(deptDF, empDF.dept\_id == deptDF.dept\_id, "inner").select("Emp\_id", "Empname", "MGR", "Y0J", "dept\_id", "gender", "salary", "dept\_name", "dept\_id").show()

In [4]: empDF.join(deptDF, empDF.dept\_id == deptDF.dept\_id, "full").show()

In [5]: empDF.join(deptDF, empDF.dept\_id == deptDF.dept\_id, "right").show()

```
>>> 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| 1986.485|  
|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| 135.4066|  
|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| 2584.2236|  
+-----+-----+-----+  
only showing top 28 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='U  
nited States'").  
+-----+-----+-----+  
| 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| 693|  
|United States| New Jersey| 6916|  
|United States| Arkansas| 3437|  
|United States| Michigan| 21869|  
|United States| Illinois| -9298|  
|United States| Ohio| -14498|  
|United States| New York| 65943|  
|United States| Missouri| 6208|  
|United States| Mississippi| 2445|  
|United States| Alaska| 9491|
```

```
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='U  
nited States'").show()
```

```
In [33]: salesdf.select("Region","country").sort(col("region").desc()).orderBy("country",ascending=False).filter("region='Central Asia'")  
+-----+  
| 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("region")  
+-----+-----+-----+  
|country| region| avg(profit)|  
+-----+-----+-----+  
| India|Central Asia|0.017664887459987974|  
+-----+-----+
```

```
In [35]: # to get type 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|76.58835714285714|
|United States|South Carolina|1774|42.23809523809524|
|United States|West Virginia|106|46.5|
|United States|Maine|417|52.125|
|United States|Montana|1852|123.46666666666667|
|United States|Alabama|4645|76.14754098369856|
|United States|Wyoming|100|100.0|
|United States|Idaho|695|33.095238095238095|
|United States|New Jersey|6918|53.6124031907752|
|United States|Arkansas|3437|57.28333333333333|
|United States|Michigan|21869|86.0984251968584|
|United States|Illinois|-9290|-18.898373983739837|
|United States|Ohio|-14490|-38.895522388859703|
|United States|New York|65943|58.77272727272727|
|United States|Missouri|6208|94.06666666666666|
|United States|Mississippi|2445|46.132075471698116|
|United States|Nebraska|1421|37.39473684219526|
|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,"Average"))

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| 1996.485| 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| 76.58835714285714| 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| 76.14754698360856| US|
|United States| Wyoming| 100| 100.0| US|
|United States| Idaho| 695| 33.095238095238095| US|
|United States| New Jersey| 6916| 53.6124831807752| 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| -30.89522388059703| US|
|United States| New York| 65943| 58.77272727272727| US|
|United States| Missouri| 6268| 94.06060606060606| US|
|United States| Mississippi| 2445| 46.132975471698116| US|
|United States| Nebraska| 1421| 37.39473684218526| US|
|United States| Minnesota| 9820| 110.33707865168529| US|
|United States| Washington| 29960| 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|6000.0|
| 111|   aaa|6888.0|
| 121|   bbb|6000.0|
| 555|   eee|7000.0|
| 666|   fff|6890.0|
+-----+
```

Day 19 ( 26.09.23)

The screenshot shows two consecutive pages from the Microsoft Azure portal.

**Create an Azure Databricks workspace** (Top Window):

- Project Details:** Subscription: 'annika-168026947025', Resource group: 'Databricks' (Create new).
- Instance Details:** Workspace name: 'azuredatabricksday19', Region: 'East US', Pricing Tier: 'Trial (Premium - 14-Days Free DBU)'. Managed Resource Group name: 'Enter name for managed resource group'.
- Buttons at the bottom: 'Review + create', 'Previous', 'Next : Networking >'.

**azuredatabricksday19 - Overview** (Bottom Window):

- Overview:** Subscription: 'annika-168026947025', Subscription ID: '30079c75-1048-4ec4-9514-795ca2ff878b', Tags: 'tag111' (Add tags).
- Settings:** Virtual Network, Encryption, Networking, Properties, Locks.
- Automation:** Tasks (preview), Export template.
- Actions:** Delete, Launch Workspace, Upgrade to Premium.
- Quick Actions:** Documentation, Getting Started, Import Data from File, Import Data from Azure Storage.

Microsoft Azure Microsoft Azure databricks

Subscription Details | Resource Groups Virtual machines - Microsoft Azure Cluster Details - Databricks

https://adb-121311785397682.azuredatabricks.net/?i=123337883978291&sf=Hiring%20Users... https://shellunext\_1693422082390@inputs... shellunext\_1693422082390@inputs...

Search data, notebooks, recent, and more... CTRL + F

New Workspace Recents Catalog Workflows Compute UI preview Send feedback

Shellunext's Cluster

Configuration Notebooks (0) Libraries Event log Spark UI Driver logs Metrics Apps Spark compute UI - Master

Move Terminate Edit

Policy Unrestricted

Multi-node Single node

Access mode Single user access Single user Shellunext

Summary

2-8 Workers (8-32 GB Memory) 8-32 Cores 1 Driver (8 GB Memory, 4 Cores) Runtime (15.5 minutes) 12

Standard\_DS3\_v2 Standard\_DS3\_v2 4-16GB

UI 150%

Performance

Databricks Runtime Version: 13.1 LTS (includes Apache Spark 3.4.1, Scala 2.12)

Use Photon Acceleration:

Worker type: Standard\_DS3\_v2 Min workers: 2 Max workers: 8 Spot instances: 0

Driver type: Standard\_DS3\_v2 (16 GB Memory, 4 Cores)

Microsoft Azure Microsoft Azure portal.azure.com

Subscription Details | Resource Groups Virtual machines - Microsoft Azure Cluster Details - Databricks

https://portal.azure.com/#view/HubsExtension/BrowseResource/resourceType/Microsoft...

Search resources, services, and discs (Q+R)

Home Virtual machines

Create Switch to classic Reservations Manage view Refresh Export to CSV Open query Assign tags Start Restart Stop

Filter for any field... Subscription equals all Type equals all Resource group equals all Location equals all Add filter

No grouping List view

Showing 1 to 3 of 3 records.

Name	Type	Subscription	Resource group	Location	Status	Operating system	Size
19c030f108c74e98370a4...	Virtual machine	rgunrest-148032119475...	databricks-rg-accelera...	East US	Running	Linux	Standard_DS3_v2
2300980ba1434023b5d247...	Virtual machine	rgunrest-148032119475...	databricks-rg-accelera...	East US	Running	Linux	Standard_DS3_v2
3ef6542903e4a1ca873504...	Virtual machine	rgunrest-148032119475...	databricks-rg-accelera...	East US	Running	Linux	Standard_DS3_v2

Page 1 of 1 Next >

Give feedback

Microsoft Azure databricks Search data, notebooks, recent, and more... CTRL + F Send feedback

https://adb-1213117853976820.azuredatabricks.net/0c1233117853976820/clusters/shellunext\_1693422962390@inputs...

**Shellunext's Cluster**

Multi-node  Single-node

Access mode:  Single user access  Single user

Single user: shellunext\_1693422962390@inputs...

**summary**

1 Driver	14 GB Memory	4 Cores
Runtime	23.0s	100%
Photon	Standard_DS3_v2	1.3 DBUs

**Performance**

Databricks runtime version: Runtime: 13.3 LTS (Scala 2.12, Spark 3.4.1)

Use Photon Acceleration

Node type: Standard\_DS3\_v2 (14 GB Memory, 4 Cores)

Terminate after: 20 minutes of inactivity

**Tags**

Add tag:

**Create compute** **Cancel**

10:45 26-09-2023

Microsoft Azure databricks Search data, notebooks, recent, and more... CTRL + F Send feedback

https://adb-1213117853976820.azuredatabricks.net/0c1233117853976820/clusters/shellunext\_1693422962390@inputs...

**Compute**

All-purpose compute Job compute SQL warehouses Pools Policies 0

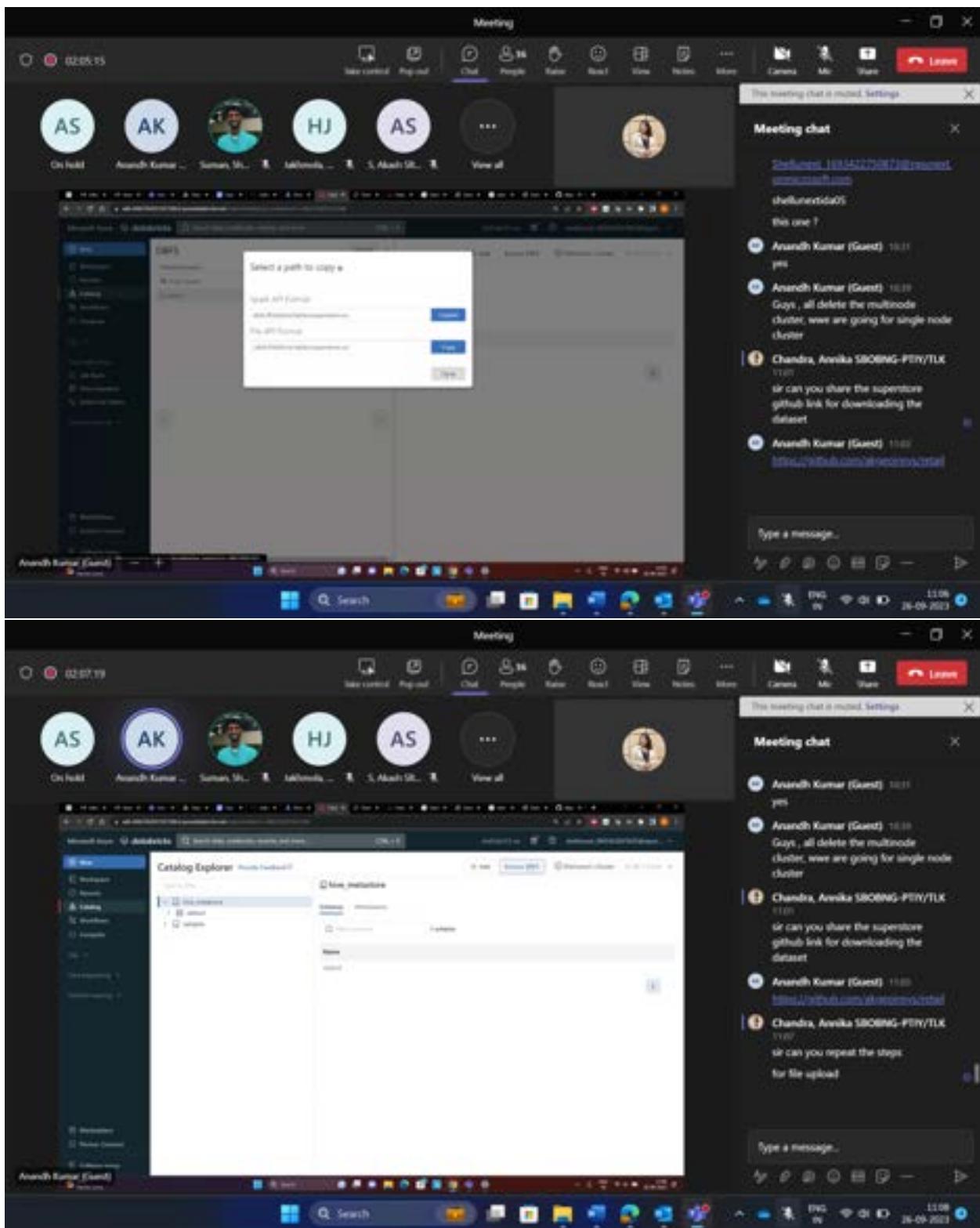
Free trial ends in 16 days. Upgrade to Premium in Azure Portal

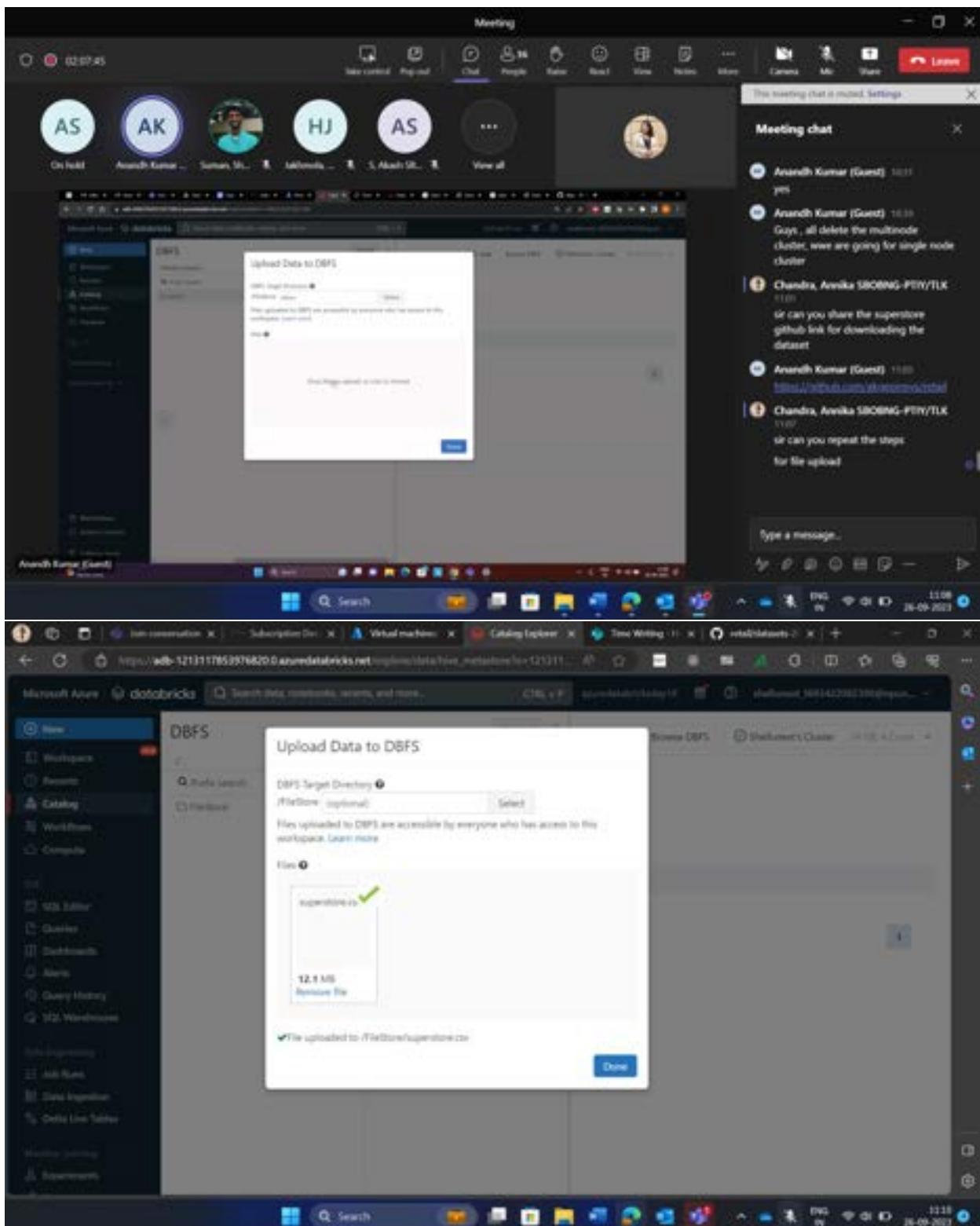
State	Name	Policy	Runtime	Active ...	Active ...	Active ...	Source	Creator	Notes...
Running	Shellunext's Cluster	-	13.3	14:08	4 cores	13	18	shellunext_1...	

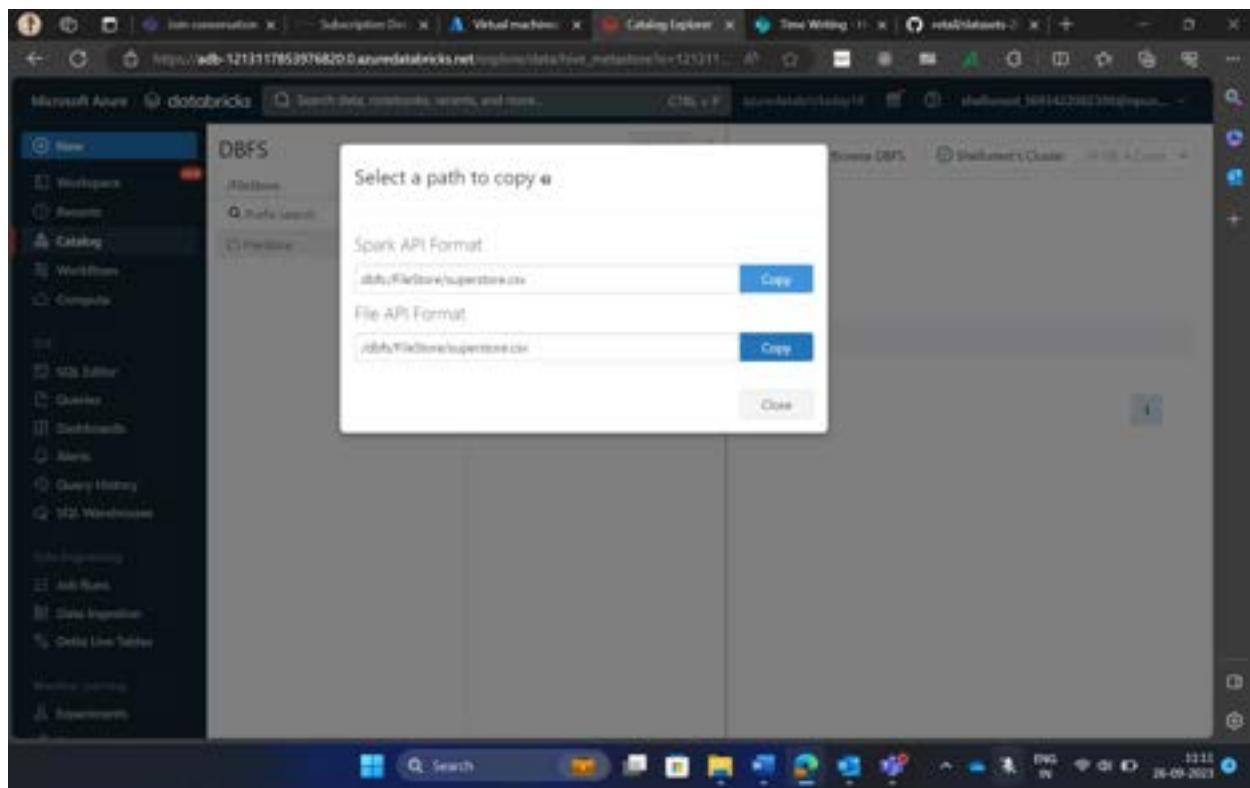
**Create with Personal Compute** **Create compute**

23 / page

10:45 26-09-2023







dbfs:/FileStore/superstore.csv

Microsoft Azure Databricks

Untitled Notebook 2023-09-26 11:42:13 Python 3

File Edit View Run Help LastEdited: 1 minute ago Provide feedback

Run all ShellCluster's Cluster Schedule Share

SPFs 3s /FileStore/tables

Table +

path	name	size	modificationTime
dbfs:/FileStore/tables/superstore.csv	superstore.csv	12064401	1986706217000

1 row | 6.70 seconds runtime Refreshed 5 minutes ago

Comment took 6.70 seconds — by shellUser\_1693422962390@sparkuser.azuredatabricks.net at 2023/09/26, 11:42:13 am on ShellCluster's Cluster

Cell 1

```
1. salesDF=spark.read.csv("dbfs:/FileStore/tables/superstore.csv",header=True,inferSchema=True)
```

\* (1) Spark Jobs

```
1. salesDF = spark.read.csv("dbfs:/FileStore/tables/superstore.csv", header=True, inferSchema=True)
```

Comment took 1.03 seconds — by shellUser\_1693422962390@sparkuser.azuredatabricks.net at 2023/09/26, 11:42:14 am on ShellCluster's Cluster

Cell 2

```
1. display(salesDF)
```

\* (1) Spark Jobs

Table +

Microsoft Azure Databricks

Untitled Notebook 2023-09-26 11:42:13 Python 3

File Edit View Run Help LastEdited: 1 minute ago Provide feedback

Run all ShellCluster's Cluster Schedule Share

1 row | 1.10 seconds Refreshed 4 minutes ago

Comment took 1.10 seconds — by shellUser\_1693422962390@sparkuser.azuredatabricks.net at 2023/09/26, 11:42:13 am on ShellCluster's Cluster

Cell 1

```
1. display(salesDF)
```

\* (1) Spark Jobs

Table +

ID	OrderID	OrderDate	ShipDate	ShipMode	CustomerID	CustomerName
1	22108	CA-2012-124891	2012-07-11	Same Day	IR-15405	Rick Hansen
2	26341	IN-2013-77878	2013-02-05	Second Class	IR-16210	Justo Ritter
3	25330	IN-2013-71249	2013-10-17	First Class	CR-12730	Craig Reiter
4	11524	ES-2013-167942	2013-01-28	First Class	EM-16275	Katherine Murray
5	47221	SG-2013-4325	2013-11-05	Same Day	IR-9495	Rick Hansen
6	22732	IN-2013-42380	2013-08-26	Second Class	IR-15655	Jim Mitchell
7	98700	MU-2011-4XQW	2011-11-27	Fast Lane	TL-71160	Tony Soccetti

7,327 rows | Truncated data | 1.30 seconds runtime Refreshed 4 minutes ago

Comment took 1.30 seconds — by shellUser\_1693422962390@sparkuser.azuredatabricks.net at 2023/09/26, 11:42:13 am on ShellCluster's Cluster

Cell 2

```
1. MySQL
```

```
2. show databases;
```

The screenshot shows two consecutive screenshots of a Microsoft Azure Databricks notebook titled "Untitled Notebook 2023-09-26 11:42:13".

**Screenshot 1:**

- SQL Editor:

```
1: mysql
2: show databases;
```
- Output:

```
+---+-----+
| databaseName |
+---+-----+
| default      |
| shell0       |
+---+
```

2 rows | 0.19 seconds runtime  
Refreshed 5 minutes ago

This result is stored as PySpark data frame `_sql1DF` and in the Python output cache as `Df[1]`. Learn more

**Screenshot 2:**

- SQL Editor:

```
1: mysql
2: use shell0;
```
- Output:

```
+---+-----+
| databaseName |
+---+-----+
| shell0      |
+---+
```

2 rows | 0.19 seconds runtime  
Refreshed 5 minutes ago

**Screenshot 3:**

- SQL Editor:

```
1: salesDF.write.mode("overwrite").saveAsTable("shell0.sales")
```
- Output:

```
+---+-----+
| databaseName |
+---+-----+
| shell0      |
+---+
```

4 rows | 0.11 seconds runtime  
Refreshed 2 minutes ago

This result is stored as PySpark data frame `_sql1DF` and in the Python output cache as `Df[2]`. Learn more

Microsoft Azure Databricks

Untitled Notebook 2023-09-26 11:42:13 - Python 3

```
1 #to make the table permanent
2 salesDF.createOrReplaceTempView("sales_temp")
```

Comment task 4.00 seconds -- by shellUser\_1693425982390@spark-vm-01.azuredatabricks.net at 2023/09/26, 11:47:14 am on ShellUser's Cluster

Cell 9

```
1 %sql
2 desc extended sales;
```

100% Spark Jobs

```
_sqldf: pySpark sql DataFrame Dataframe = (col_name: string, data_type: string, ... 1 more field)
```

Table

col_name	= data_type
1 ID	int
2 OrderID	string
3 OrderDate	date
4 ShipDate	date
5 ShipMode	string
6 CustomerID	string
7 CustomerName	string

42 rows | 0.46 seconds runtime

Refreshed 1 minute ago

This result is stored as PySpark data frame `_sqldf` and in the Python output cache as `Det[2]`. Learn more

12:49 26-09-2023

V Microsoft Azure Databricks

Untitled Notebook 2023-09-26 11:42:13 - Python 3

```
1 #to make the table permanent
2 salesDF.createOrReplaceTempView("sales_temp")
```

Comment task 4.00 seconds -- by shellUser\_1693425982390@spark-vm-01.azuredatabricks.net at 2023/09/26, 11:47:14 am on ShellUser's Cluster

Cell 10

```
1 %sql
2 update sales set Discount=8 where ID=44580;
```

100% Spark Jobs

```
_sqldf: pySpark sql DataFrame Dataframe = (num_affected_rows: long)
```

Table

num_affected_rows
1

1 row | 6.00 seconds runtime

Refreshed 1 minute ago

This result is stored as PySpark data frame `_sqldf` and in the Python output cache as `Det[3]`. Learn more

Comment task 4.00 seconds -- by shellUser\_1693425982390@spark-vm-01.azuredatabricks.net at 2023/09/26, 11:47:14 am on ShellUser's Cluster

Cell 11

```
1
```

Shift+Enter to run

12:49 26-09-2023

Microsoft Azure | databricks | Search data, notebooks, recent, and more... CTRL + P anndatabricksday19 ShellUser's Cluster Schedule Share

ADBDAY1- dataframe\_tables Python

File Edit View Run Help Last edit 3 minutes ago Provide feedback Run all ShellUser's Cluster Schedule Share

1. `sql`  
2. `delete from sales where ID=4658;`

+ (1) Spark Jobs  
+ (1) \_sql: pySparkSQLDatabase: DataFrame [ID: integer, OrderID: string ... 22 more fields]

Table +  
`sum_affected_rows =`  
1  
1 row | 3.89 seconds runtime Refreshed now

This result is stored as PySpark DataFrame `_sqlDF`, and in the Python output cache at `out[18]`. Learn more

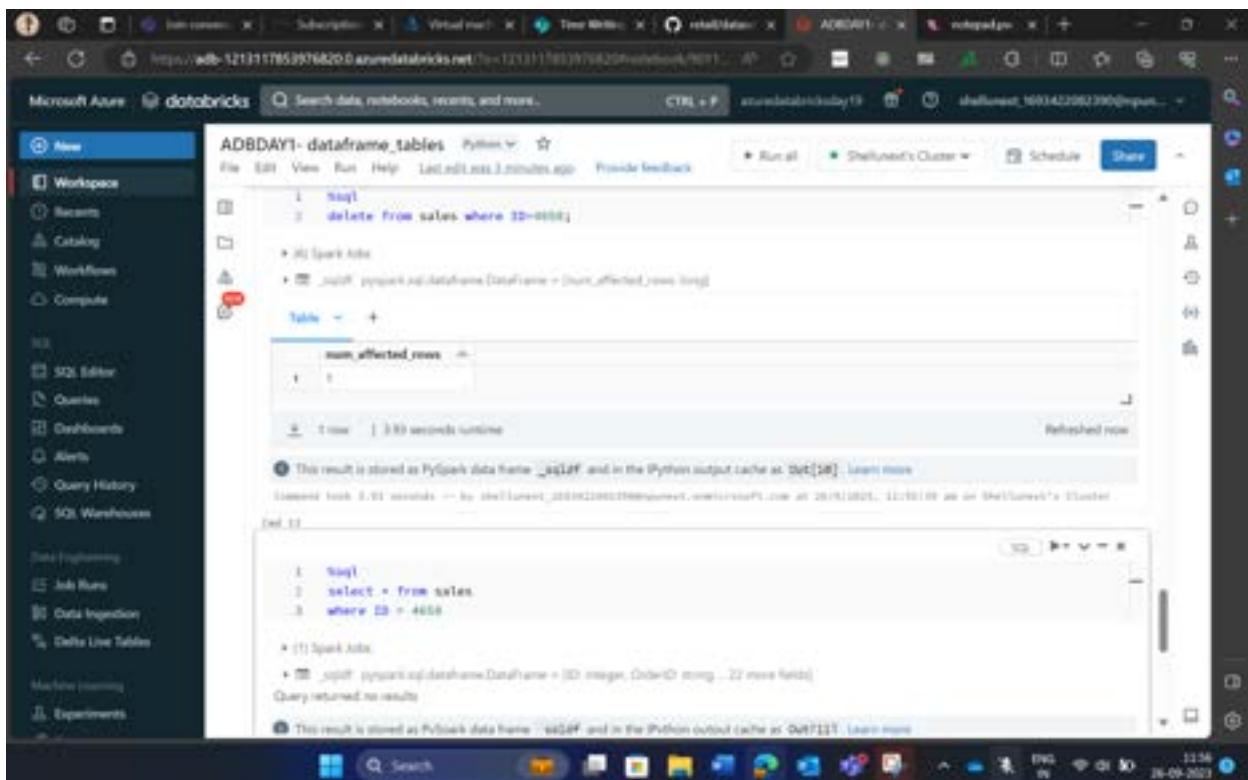
Comment took 3.89 seconds — by shellUser, 2023-09-26T10:00:00Z, anndatabricks.com at 2023/09/26, 11:10:09 am on ShellUser's Cluster

Det. 11

1. `sql`  
2. `select * from sales`  
3. `where ID = 4658;`

+ (1) Spark Jobs  
+ (1) \_sql: pySparkSQLDatabase.DataFrame [ID: integer, OrderID: string ... 22 more fields]  
Query returned no results

This result is stored as PySpark DataFrame `_sqlDF`, and in the Python output cache at `out[19]`. Learn more



Microsoft Azure | databricks | Search data, notebooks, recent, and more... CTRL + P anndatabricksday19 ShellUser's Cluster Schedule Share

DBFS

Upload file DBFS

DBFS Target Directory  Select

Free trial ends in 16 days. Upgrade to Premium in Azure Portal

File  `emp.csv` ✓  
39 B Remove file

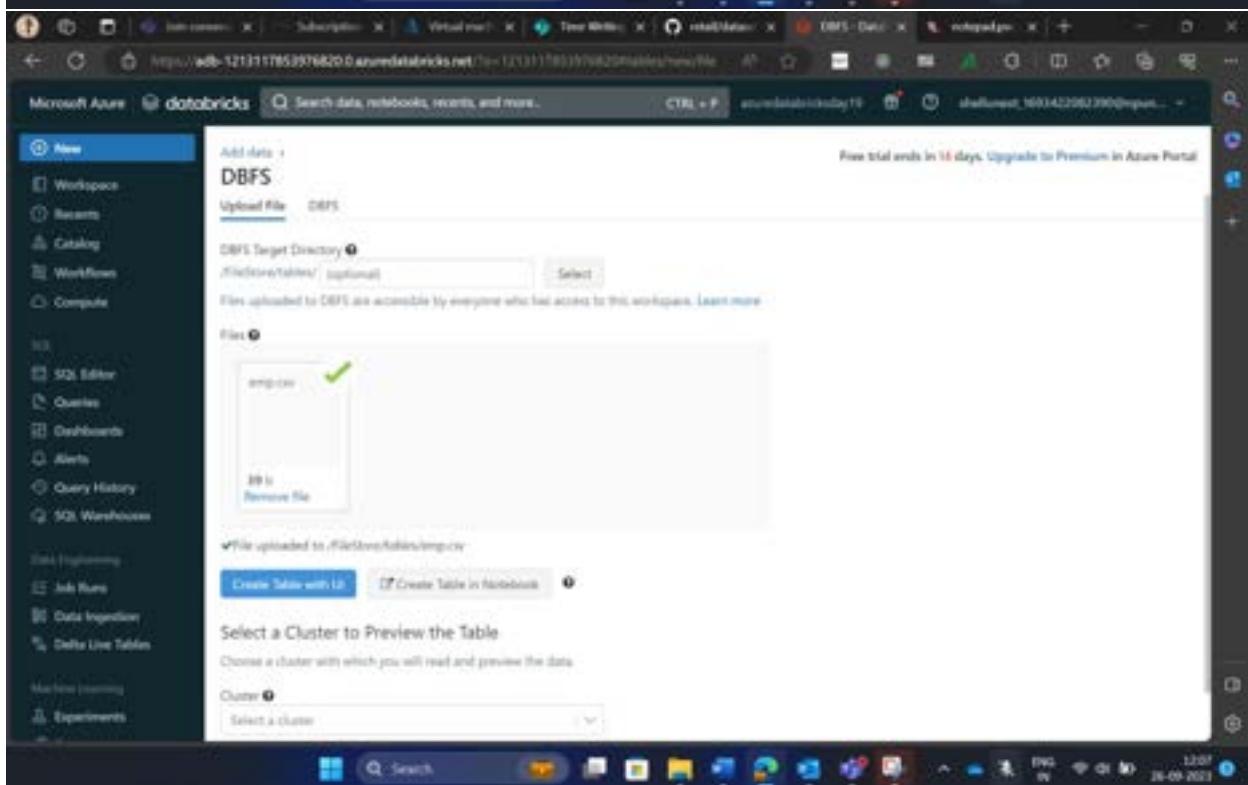
✓ File uploaded to /FileStore/Tables/emp.csv

Create Table with UI Create Table in notebook

Select a Cluster to Preview the Table

Choose a cluster with which you will read and preview the data.

Cluster



Meeting

03:08:21

Slate control | Payload | Chat | People | Raise | Read | View | Notes | More | Camera | Me | Share | Leave

Meeting chat

Chandra, Annika SBOBNG-PTV/TLK 11:07  
can you repeat the steps  
for file upload

Siddiqui, Arin SBOBNG-PTV/PBD 11:08  
Not getting the option of "Browse DBFS"  
Yes, enabled  
Got it now!

This message has been deleted 11:09

This message has been deleted 11:09

Anand Kumar (Guest) 11:10  
just 5 min break

Type a message...

DBFS

Create a folder with the same name or choose an existing one.

Specify Table Attributes

Add data →

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

Search

Microsoft Azure Microsoft databricks Search data, notebooks, recent, and more... CTRL + F https://adb-1213117853976820.0.azuredatabricks.net/.../catalog/default?\_ga=2.1422962390.1941214019.1663264114-1000000000.1663264114 Catalog Explorer Browse DBFS Shellunes's Cluster 14 GB, 4 Cores

New Workspace Recent Catalog Workflow Compute

SQL Editor Queries Dashboards Alerts Query History SQL Warehouses

Data Engineering Job Runs Data Ingestion Delta Live Tables

Machine Learning Experiments

Catalog Explorer [Provide Feedback](#)

Type to filter:

- default
- default.employee\_csv
- default
- default
- samples

Owner: Not set Size: 1KB, 1 file Last Updated: Unknown

Comment: Add comment

Columns Sample Data Details Permissions History

Column	Type	Comment
empno	int	
ename	string	
sal	double	

Create

Microsoft Azure Microsoft databricks Search data, notebooks, recent, and more... CTRL + F https://adb-1213117853976820.0.azuredatabricks.net/.../dbfs?\_ga=2.1422962390.1941214019.1663264114-1000000000.1663264114 DBFS

New Workspace Recent Catalog Workflow Compute

SQL Editor Queries Dashboards Alerts Query History SQL Warehouses

Data Engineering Job Runs Data Ingestion Delta Live Tables

Machine Learning Experiments

Add data: DBFS

Upload File DBFS

DBFS Target Directory [/FileStore/tabels/](#) [optional] Select

Free trial ends in 16 days. Upgrade to Premium in Azure Portal

File [customer.csv](#) ✓  
8.2 MB Remove file

\*File uploaded to /FileStore/tabels/customer.csv

[Create Table with UI](#) [Create Table in Notebook](#)

Microsoft Azure | Search resources, services, and data (G+)

Home > annikastorageacc\_1095711901161 | Overview > annikastorageacc | Containers > democontainer

democontainer

Successfully uploaded blob(s)  
Successfully uploaded 3 blob(s)

Search

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

Authentication method: Access key (Switch to Azure AD User Account)

Location: democontainer / source

Search blobs by prefix (case-sensitive)

Name Modified Access tier Archive status Blob type

1.j	9/26/2023, 12:40:21...	Hot (Inferred)		Block blob
customer.csv	9/26/2023, 12:40:21...	Hot (Inferred)		Block blob
AnimeType.csv	9/26/2023, 12:40:21...	Hot (Inferred)		Block blob
nation.csv	9/26/2023, 12:40:19...	Hot (Inferred)		Block blob
region.csv	9/26/2023, 12:40:01...	Hot (Inferred)		Block blob

12:40 26-09-2023

This screenshot shows the Microsoft Azure Storage Container Overview page for the 'democontainer' in the 'annikastorageacc' account. It displays a list of five blobs: '1.j', 'customer.csv', 'AnimeType.csv', 'nation.csv', and 'region.csv'. Each blob is listed with its name, modified date (9/26/2023), access tier (Hot), archive status (not applicable for blobs), and blob type (Block blob). A success message at the top right indicates that three blobs were uploaded.

Microsoft Azure | @ databricks | Search data, notebooks, recent, and more... CTRL + F

annikadatabricks@1213117853976290 | Overview | Databricks | Data Ingestion | Data Import | Data Lake | Data Pipelines | Data Engineering | Machine Learning | Experiments

New

Workspace

Recent

Catalog

Workflow

Compute

SQL

SQL Editor

Queries

Dashboards

Alerts

Query History

SQL Warehouses

Data Ingestion

Delta Live Tables

Machine Learning

Experiments

Add data

Data sources

Create or modify table

From local files (T)

Add data from Azure Data Lake Store using a Databricks notebook

Native integrations (T)

Azure Blob Storage

Azure Data Lake ...

Cassandra

Snowflake

IDB

Kafka

Elasticsearch

MongoDB

PostgreSQL

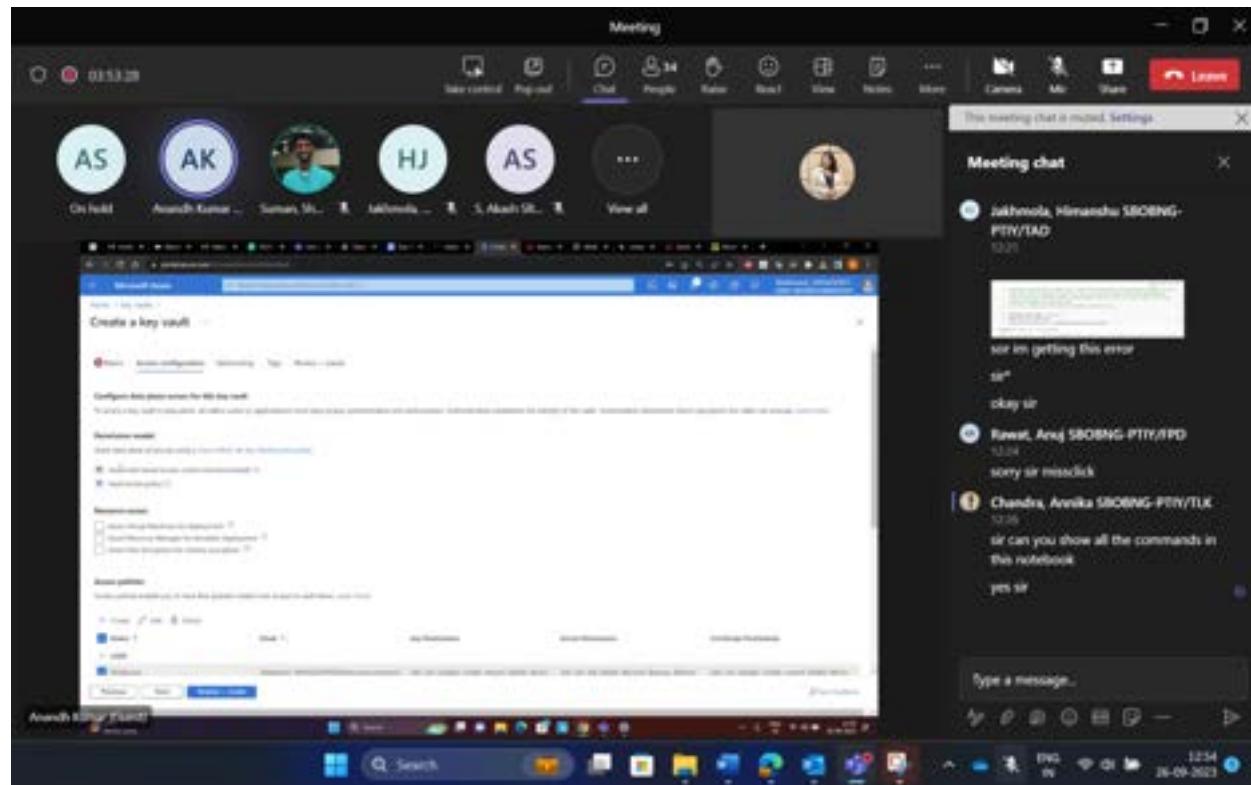
MySQL

DBS

12:40 26-09-2023

This screenshot shows the Databricks Data Ingestion interface. On the left, there's a sidebar with various navigation options like Workspace, Recent, Catalog, etc. The main area is titled 'Add data' and shows 'Data sources'. It includes sections for 'Create or modify table' (with a note about creating a new table or replacing an existing one) and 'Native integrations (T)'. Below these are icons for Azure Blob Storage, Azure Data Lake, Cassandra, Snowflake, IDB, Kafka, Elasticsearch, MongoDB, PostgreSQL, MySQL, and DBS. The interface is designed to facilitate importing data from various cloud storage and database systems into Databricks.

XDmK6b6u/sxVQuV5/KHgC/Gwm4tG6DCVi7Swnl0IXdRkktcdCmsHTtx7FH9QpdsDiQaZIOHvSC+ASTk5Pjf  
A==



The screenshot shows the Microsoft Azure portal with the URL 'https://portal.azure.com/#blade/HubsBlade/resourceType/keyVaults/resourceName/annikaku/secrets'. The page title is 'annikaku | Secrets'. The left sidebar includes sections for Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Access policies, Events, Keys, Secrets (selected), and Certificates. The main content area shows a table for secrets:

Name	Type	Status	Expiration date
adilic		✓ Enabled	

A message at the top of the table says: 'The secret 'adilic' has been successfully created.'

To create new scope

### Databricks (azuredatabricks.net)

The screenshot shows the Databricks Azure portal interface. On the left, there is a sidebar with various navigation options such as Workspace, Databricks, Catalog, Workflow, Compute, SQL Editor, Queries, Databases, Alerts, Query History, and others. The main area is titled "Create Secret Scope". It contains fields for "Scope Name" (with "Creator" selected as the "Manage Principal"), "Azure Key Vault" (with "https://new-vault.azure.net/" in the "DNS Name" field and "/subscriptions/xxxxxx" in the "Resource ID" field), and a "Cancel" button.

The screenshot shows a Microsoft Azure Databricks notebook titled "Untitled Notebook 2023-09-26 14:35:25". The left sidebar displays the "Catalog" section. The main area contains the following code:

```
dbutils.fs.mount(source = "wasbs://databricks@azuredatalakestorage1.blob.core.windows.net/",  
                  mount_point = "/mnt/input/",  
                  extra_configs = {"fs.azure.account.key": azuredatalakestorage1.blob.core.windows.net": dbutils.secrets.get(scope = "AzureDataLakeStorage1", key="adlsKey")})
```

The "Run Cell" button is highlighted. Below the code, the output shows the command took 21.62 seconds to run on "Shellunet\_1693422962390" and lists the mounted path:

path	name	size	modificationTime
/mnt/input/source/	source/	0	0

The "Refreshed now" timestamp is 26/09/2023.

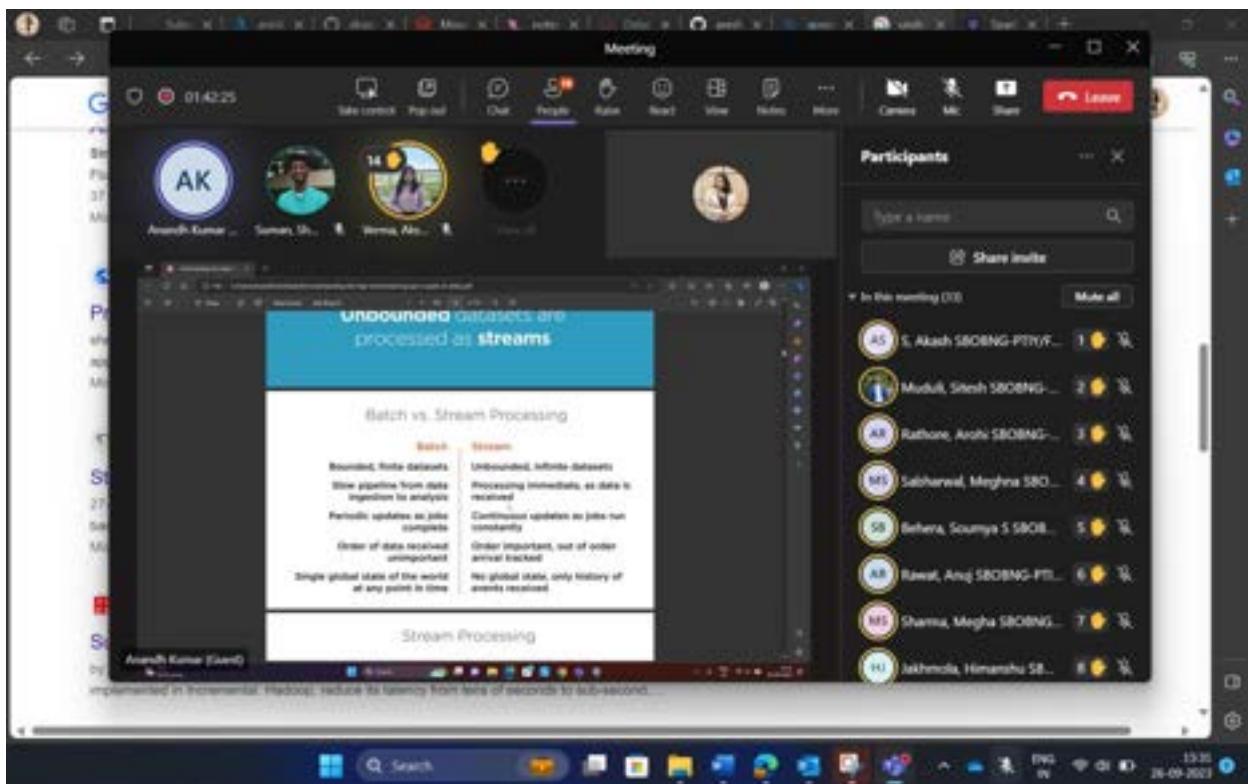
The screenshot shows the Microsoft Azure Databricks Catalog Explorer. The left sidebar displays the "Catalog" section. The main area shows the mounted DBFS directory:

Path: /mnt/input/source

3 tables

customer.csv  
dimension.csv  
nation.csv  
region.csv

The "Upload" button is visible at the top right of the catalog view.



## Spark streaming user defined schema

The image shows two overlapping windows. The top window is a Microsoft Teams meeting titled 'Meeting' with participants Amanish Kumar, Somers, Sh., Venna, Akh., and Siddiqui, Afroz. The bottom window is a Databricks notebook titled 'ADB-Day1-DataFrame\_Tables' running on a cluster named 'ShellCluster'. The notebook contains the following code:

```
1 #spark structured streaming
2
3 from pyspark.sql import *
4 schema = StructType([StructField("issue_code", StringType(), True), \
5                     StructField("borough", StringType(), True), \
6                     StructField("major_category", StringType(), True), \
7                     StructField("minor_category", StringType(), True), \
8                     StructField("value", StringType(), True), \
9                     StructField("year", StringType(), True), \
10                    StructField("month", StringType(2, True))])
11
12 fileStreamDF = spark.readStream \
13     .option("header", "true") \
14     .schema(schema) \
15     .csv("/mnt/ADBDAY1/inputdata")
```

Microsoft Azure: databricks

ADB-Day1-DataFrame\_Tables Python V. ⚡

File Edit View Run Help Last Edit 1 min ago Provide feedback Interrupt Shellinski's Cluster Schedule Share

New Workspace Recent Catalog Workflows Compute

SQL Editor Queries Dashboards Alerts Query History QD Warehouses

Note Programming Job Runs Data Ingestion Delta Live Tables

Machine Learning Experiments

ADB-Day1-DataFrame\_Tables

```
1 query = trimmedDF.writeStream
  .outputMode("append")
  .format("console")
  .option("truncate", "false")
  .option("maxLines", 100)
  .start()
```

Cancel

11 Spark Jobs

19592575-cfbd-4cc8-9887-7d795305211 Last updated 3 seconds ago

Dashboard Raw Data

Inputs: Processing Rate records per second

0 m/s 0 m/s

Batch Duration in milliseconds

219.6 ms 25 ms Average Latency

This screenshot shows a Databricks notebook interface. On the left is a sidebar with navigation links like Workspace, Recent, Catalog, etc. The main area has a code editor with a single line of Scala code for a stream processing job. Below the code is a 'Cancel' button. A 'Spark Jobs' section shows one active job. At the bottom, there are two charts: 'Inputs: Processing Rate' (records per second) and 'Batch Duration' (in milliseconds). The processing rate chart shows a sharp drop from approximately 200 to 0 records per second over time. The batch duration chart shows a constant value of 25 ms.

Day 20 27.09.23

The image shows two side-by-side screenshots of Microsoft Azure interfaces.

**Top Screenshot (Azure Storage Container):** This screenshot shows the Azure Storage Container interface. The URL is [https://portal.azure.com/#view/Microsoft\\_Azure\\_Storage/ContainerMenuBlade/Overview](https://portal.azure.com/#view/Microsoft_Azure_Storage/ContainerMenuBlade/Overview). The container name is '2023'. The table lists three blob items: 'crime-data', 'source', and 'target'. The 'source' item has a yellow warning icon.

Name	Modified	Access tier	Archive status	Blob type
crime-data				
source				
target				

**Bottom Screenshot (Databricks Secret Scope):** This screenshot shows the 'Create Secret Scope' dialog in Databricks. The URL is <https://adb-5013842652615528.8.azuredatabricks.net/>. The scope name is 'adbscope'. The 'Azure Key Vault' section shows the 'Resource ID' field containing the value: /subscriptions/50079c75-10d8-4cc4-95f4-75052a8f1fb6/resourceGroups/adbdatalake

The screenshot shows the Microsoft Azure Databricks Day 20 workspace. On the left, the sidebar includes sections for Workspace, Recent, Catalog, Workflow, Compute, SQL Editor, Queries, Dashboards, Alerts, Query History, and SQL Workhouse. The main area has tabs for 'Python' and 'SQL'. The 'Python' tab contains code for creating a delta table:

```
1 delta table
2 salesdf.write.mode("overwrite").saveTable("sales")
```

The 'SQL' tab contains a query:

```
1 MySQL
2 select * from sales;
```

Below the SQL tab is a table titled 'sales' with columns: ID, OrderID, OrderDate, ShipDate, ShipMode, CustomerID, and CustomerName. The data is as follows:

	ID	OrderID	OrderDate	ShipDate	ShipMode	CustomerID	CustomerName
1	12298	CA-2012-124891	2012-07-01	2012-07-31	Same Day	884-19405	Rick Hansen
2	20341	IN-2013-77878	2013-02-05	2013-02-07	Second Class	884-16210	Justin Miller
3	25130	IN-2013-71048	2013-08-17	2013-10-18	First Class	OK-12730	Craig Retter
4	11524	ES-2013-1570342	2013-01-28	2013-01-30	First Class	NM-16375	Katherine Murray
5	47221	SG-2013-4326	2013-11-05	2013-11-08	Same Day	884-0491	Rob Hansen
6	12732	IN-2013-42360	2013-06-28	2013-07-01	Second Class	884-15655	Jim Mitchell

```
1: #sql
2: select count(*) from sales; --before deleting
3: 
4: #SQL Jobs
5: +----+-----+
6: | jobID | 
7: +----+-----+
8: | 1 | 
9: +----+-----+
10: 
11: count(*) |
12: 51250
13: 
14: 7 rows | 0.37 seconds runtime
15: 
16: This result is stored as Pyspark DataFrame _sqlDF and in the Python output cache as: sql[18]. Learn more
17: Committed from 8:30 seconds ago by the Cluster_1899412900@lensparent, committed 2023-09-27T07:30:11Z on BellHorn's Cluster
18: 
19: 
20: #sql
21: delete from sales where Country='India';
22: select count(*) from sales; --after deleting
23: 
24: #SQL Jobs
25: +----+-----+
26: | jobID | 
27: +----+-----+
28: | 2 | 
29: +----+-----+
30: 
31: count(*) |
32: 49770
33: 
34: 7 rows | 0.37 seconds runtime
```

The screenshot shows the Microsoft Azure Databricks interface. The left sidebar is the navigation menu with options like Workspace, Catalog, Workflow, Compute, etc. The main area is titled "DBFS" and shows the path "/user/hive/warehouse/shell0day20.dbs/sales". A search bar at the top right says "Search data, notebooks, recent, and more...". Below it, there's a "Prefix search" dropdown set to "sales" and a "Profile search" dropdown. A list of partitions is displayed under "sales\_log":

- part-00000-11ab0ebe-16bf-46f7-a5d2-4c5...
- part-00000-92298ddaa0010-4cf7...
- part-00000-c3ba7ef91-0bbf-40e...
- part-00000-d03abcc3-96c4-4e1...
- part-00001-fef44998-54f2-41c...
- part-00001-439ff504-079-4e1...
- part-00001-35a44703-61b3-40e...
- part-00002-416a531-60f1-4e1...
- part-00002-94abdc2a-4c33-404...
- part-00002-ae9f9061-e2ce-4a1...

A message "2 schemas" is visible on the right.

This screenshot shows the same Databricks interface as the first one, but the path in the main area is "/user/hive/warehouse/shell0day20.dbs/sales\_part/Country+India". The "Prefix search" dropdown is now set to "Country+India". The list of partitions under "Country+India" includes:

- Country-Georgia
- Country-Germany
- Country-Ghana
- Country-Guadeloupe
- Country-Guatemala
- Country-Guinea
- Country-Haiti
- Country-Honduras
- Country-Hong Kong
- Country-Hungary
- Country-India
- Country-Indonesia
- Country-Iran
- Country-Iraq
- Country-Ireland
- Country-Israel
- Country-Japan

A message "2 schemas" is visible on the right.

Microsoft Azure | databricks | Search data, notebooks, recent, and more... CTRL + F

JDBC Connection Python ▾

File Edit View Run Help Last edited 5 minutes ago Provide feedback Run all ShellCluster10 Schedule Share

Free trial ends in 14 days Upgrade to Premium in Azure Portal

Untitled

```
jdbcHostname = "adbsqldataserver.database.windows.net"
jdbcPort = "1433"
jdbcDatabase = "adw_adventureworks"
properties = {
    "user": "sa@adwadventureworks",
    "password": "P@ssw0rd"}
```

Comment task 8.00 seconds — by shellcluster\_1000422062390@runmed.microsoft.com on 27/09/2023, 11:56:40 pm in ShellCluster10 Cluster

Untitled

```
connect("jdbc:sqlserver://adbsqldataserver.database.windows.net:1433;database=adw_adventureworks;
username=sa@adwadventureworks;password=P@ssw0rd;encrypt=true;trustServerCertificate=false;
hostNameInCertificate=adbsqldataserver.database.windows.net;loginTimeout=30);"
```

Comment task 8.00 seconds — by shellcluster\_1000422062390@runmed.microsoft.com on 27/09/2023, 11:56:40 pm in ShellCluster10 Cluster

Untitled

Untitled Editor for run  
switch to lineage to see selected runs

Q Search

ENGLISH IN 30 27-09-2023

Microsoft Azure | databricks | Search data, notebooks, recent, and more... CTRL + F

Workflow Yesterdays Add a name for your job...

Run Tasks

task\_1\_dfl\_operations

... from microsoft.com/Databricks Day 20  
Job cluster

Task name\* task\_1\_dfl\_operations

Type\* Notebook

Source\* Workspace

Path\* ... from microsoft.com/Databricks Day 20

Cluster\* Job cluster 14.0GB 4 Cores 128GB Python - Spark 3.3.1 Scala 3.2

Dependent Libraries + Add

Create Cancel

Q Search

ENGLISH IN 30 27-09-2023

The screenshot shows the Microsoft Azure Databricks interface. The left sidebar is titled 'New' and includes options like Workspace, Recents, Catalog, Workflows, Compute, SQL Editor, Queries, Dashboards, Alerts, Query History, and SQL Warehouses. The 'Workflows' section is currently selected. The main area displays a workflow named 'task\_1\_df\_operations'. A card for this task shows it's associated with 'https://anonymouseffector.com/Databricks Day 20' and 'Job\_cluster'. A blue button '+ Add task' is visible. To the right, the 'Job details' panel is open, showing the Job ID (101122612377404), Creator (Shelluser), Run as (Shell user), and Tags (+ Tag). Below that is the 'Git' section, which is 'Not configured' with a link to 'Add Git settings'. The 'Schedule' section shows 'None' and a 'Add schedule' button. At the bottom, the 'Compute' section lists 'Job\_cluster'. The top navigation bar shows the URL https://adb-50138626526155288.azuredatabricks.net/ and the search bar 'Search data, notebooks, records, and more...'. The status bar at the bottom indicates the browser is up-to-date.

The screenshot shows the Microsoft Azure Databricks interface. The left sidebar is titled "Microsoft Azure" and includes sections for "New", "Workspace", "Recent", "Catalog", "Workflow", "Compute", "SQL", "SQL Editor", "Queries", "Dashboards", "Alerts", "Query History", "SQL Workspaces", "Data Engineering", "Job Runs", "Data Ingestion", "Delta Live Tables", "Machine Learning", and "Experiments". The "Workflow" section is currently selected.

The main area displays a workflow named "task\_1\_df\_operations". A new task, "task2\_jdbc", is being created. The task configuration includes:

- Cluster**: Set to "Ints\_cluster" (Type: Databricks Cluster, Version: 14.0, Status: Active, Last Run: 14 hours ago, Spark Version: 3.4.1, Node Count: 1).
- Dependent Libraries**: "+ Add"
- Parameters**: "+ Add"
- Depends on**: "task\_1\_df\_operations" (with a delete icon)
- Run if**: "All succeeded"

A tooltip for "task2\_jdbc" indicates it is "Blocks Day 20" and lists its dependencies: "...etconnectivity.com/JDBC Connection" and "Ints\_cluster".

At the bottom right, there are "Cancel" and "Create task" buttons.

Microsoft Azure Databricks

Workflow: task\_1

task\_1\_df\_operations

Run Tasks

task\_1\_of\_operations

task\_2\_of\_operations

Add task

Task name\*: task\_2\_of\_operations

Type\*: Notebook

Source\*: Workspace

Path\*: /Users/shellunest\_169342082390@ipunrest.onmicrosoft.com/IDBC Connection

Cluster\*: job\_cluster

Processor Workers (T)

Cancel Save

This screenshot shows the Databricks Workflow interface. On the left, there's a sidebar with various options like Workspace, Catalog, and Machine Learning. The main area shows a workflow named 'task\_1'. A new task, 'task\_2\_of\_operations', is being added. The task configuration includes a 'Type' of 'Notebook' from 'Workspace', a path in 'IDBC Connection', and a 'Cluster' of 'job\_cluster'. A 'Processor Workers' dropdown is also present.

Microsoft Azure Databricks

Workflow: task\_1

task\_1\_df\_operations

Runs Tasks

task\_1\_of\_operations

task\_2\_of\_operations

Start date: Sep 27

Job details

Job ID: 163122612777434

Creator: Shellunest

Run as: Shellunest

Tags: +Tag

Git

Not configured

Add Git settings

Schedule

None

Add schedule

Compute

job\_cluster

Databricks Runtime 8.2, Workspace, Python 3, Scala 2.12

This screenshot shows the completed workflow 'task\_1'. It lists two tasks: 'task\_1\_of\_operations' and 'task\_2\_of\_operations'. The 'Runs' tab shows the execution history for both tasks, with a green bar indicating successful completion. To the right, detailed information about the workflow is provided, including 'Job details' (Job ID, Creator, Run as), 'Git' (not configured), 'Schedule' (none), and 'Compute' (job\_cluster). The bottom status bar indicates the runtime version is 8.2.

Microsoft Azure Databricks

DataBrics Day 20 Python

File Edit View Run Help LastEdit 5 minutes ago Provide feedback

Run all ShellCluster's Cluster Schedule Share

State: Gujarat

1. #drop down  
2. dbutils.widgets.dropdown("state", "Tamil nadu", ["Karnataka", "Gujarat", "Delhi", "Maharashtra", "Telangana", "Tamil nadu"])

Command took 0.00 seconds --- by ShellCluster\_202308270900@mskcontrol.conferenceP.com at 21/08/2023, 11:59:04 am on ShellCluster's Cluster

Last 5

1. #sql  
2. select \* from sales where country = 'India' and state = getArgument("state") --dynamic filtering

+ 11 Spark Jobs

+ 100+ pySparkSQL DataFrame Dataframe = (20 integer, OrderID string, ... 22 more fields)

Table +

ID	OrderID	OrderDate	ShipDate	ShipMode	CustomerID	CustomerName	Segment
1	22999	2012-02-25	2012-02-25	Same Day	BP-11230	Benjamin Patterson	Consumer
2	23880	2013-11-25	2013-11-27	First Class	SU-20803	Bruce Van	Consumer
3	20424	2014-05-05	2014-05-07	Second-Class	OC-12475	Grody Chapman	Consumer
4	21842	2014-01-25	2014-01-29	Standard Class	KD-16345	Katherine Dutch	Consumer
5	20150	2013-10-08	2013-10-21	Second Class	CR-12625	Corey Krieger	Home

Q Search

1453 27-08-2023

Microsoft Azure Databricks

DataBrics Day 20 Python

File Edit View Run Help LastEdit 5 minutes ago Provide feedback

Run all ShellCluster's Cluster Schedule Share

State: Maharashtra

1. #sql  
2. select \* from sales where country = 'India' and state = getArgument("state") --dynamic filtering

+ 11 Spark Jobs

+ 100+ pySparkSQL DataFrame Dataframe = (20 integer, OrderID string, ... 22 more fields)

Table +

Name	Segment	City	State	Country	PostalCode	Market	Region	Product
1	Tele	Corporate	Thane	Maharashtra	India	null	APAC	Central Asia
2	CI	Consumer	Nagpur	Maharashtra	India	null	APAC	Central Asia
3	ventech	Consumer	Nagpur	Maharashtra	India	null	APAC	Central Asia
4	Corporate	Amaravati	Maharashtra	India	null	APAC	Central Asia	TBC
5	Arman	Corporate	Nashik	Maharashtra	India	null	APAC	Central Asia
6	ope	Consumer	Thane	Maharashtra	India	null	APAC	Central Asia
7	4							

290 rows | 0.01 seconds runtime

Refreshed now

This result is stored as PySpark DataFrame \_q41EF and in the Python output cache as: Dat[3P]. Learn more

Comment took 0.00 seconds --- by ShellCluster\_202308270900@mskcontrol.conferenceP.com at 21/08/2023, 11:59:02 am on ShellCluster's Cluster

Q Search

1501 27-08-2023

DataBricks Day 20 Python

File Edit View Run Help Last edit was 5 minutes ago Provide feedback Run all Shellunde's Cluster Schedule Share

Country: `country_id` state: `state`

India: `text` Maharashtra

Code 1 Widgets 2 `#!drop_down` 3 `databricks.widgets.dropdown("state", "Tamil nadu", ["Karnataka", "Gujarat", "Delhi", "Rajasthan", "Odisha", "Tamil nadu"])`

Command took 0.00 seconds — by shellunde, 2023-09-28T00:00:00Z on 21/09/2023, 11:33:14 pm on Shellunde's cluster

Code 1 SQL 2 `select * from sales where country = 'India' and state = getArgument("state")`

(1) Spark Job

+ `sql://pySparkSQL:databricksDataFrame[DataFrame <= 22 rows, 7 columns]`

Table +

ID	OrderID	OrderDate	ShipDate	ShipMode	CustomerID	CustomerName	Segment
1	10689	2011-04-01	2011-04-26	First Class	MD-10205	Mich. Gourmet	Corp
2	10234	2013-09-05	2013-09-23	Standard Class	EM-19960	Sinc. Mudrock	Corpo
3	10771	2012-02-24	2012-02-29	Standard Class	DI-13330	Denise Lernerbach	Corpo

Microsoft Azure datarocks Search data, notebooks, recent, and more... CTR + P azuredatabricks20 shelluser\_1693422062900@inputs... Q

DataBricks Day 20 Python star

File Edit View Run Help Last edit was 6 minutes ago Provide feedback Run cell Shell user's Cluster Schedule Share

Country: `combo_id` state: `state`

India: `text` Maharashtra: `dropdown`

296 rows | 0:45 seconds runtime Refreshed 1 minute ago

This result is stored as PySpark data frame `_sqlDF`, and in the Python output cache as `lit[41]`. Learn more.

Command took 0.01 seconds — by shelluser\_1693422062900@inputs... on 21/8/2021, 10:47:00 pm on Shell user's Cluster

Cell 20

```
1 #3, textbox
2 dbutils.widgets.text("Country", "")
```

Command took 0.19 seconds — by shelluser\_1693422062900@inputs... on 21/8/2021, 10:47:11 pm on Shell user's Cluster

Cell 21

```
1 #3, dropdown
2 dbutils.widgets.dropdown("combo_id", "text", ["1", "2", "3"])
```

Command took 0.01 seconds — by shelluser\_1693422062900@inputs... on 21/8/2021, 10:47:15 pm on Shell user's Cluster

Cell 22

```
1
```

Day 21 28.09.2023

The screenshot displays a Linux desktop environment with two terminal windows open in a window manager.

**Top Terminal Window:**

```
File Edit Search View Document Help
# Use an official Python runtime as a parent image
FROM python:3.8-slim-buster

# Set the working directory to /app
WORKDIR /app

# Copy the current directory contents into the container at /app
COPY . /app

# Install any needed packages specified in requirements.txt
RUN pip install -r requirements.txt

# Make port 80 available to the world outside this container
EXPOSE 80

# Define environment variable
ENV NAME World

# Run app.py when the container launches
CMD ["python", "app.py"]
```

**Bottom Terminal Window:**

```
File Edit View Search Terminal Help
push Push an image or a repository to a registry.
rename Rename a container.
restart Restart one or more containers.
rm Remove one or more containers.
rmi Remove one or more images.
run Run a command in a new container.
save Save one or more images to a tar archive (streamed to STDOUT by default).
search Search the Docker Hub for images.
start Start one or more stopped containers.
stats Display a live stream of container(s) resource usage statistics.
stop Stop one or more running containers.
tag Create a tag TARGET_IMAGE that refers to SOURCE_IMAGE.
top Display the running processes of a container.
update Update all processes within one or more containers.
version Show the Docker version information.
wait Block until one or more containers stop, then print their exit codes.

Run 'docker COMMAND --help' for more information on a command.

To get more help with Docker, check out our guides at https://docs.docker.com/guide/
labuser@ip-172-31-0-134:~$ docker shell
labuser@ip-172-31-0-134:~$ cd shell
labuser@ip-172-31-0-134:~/shell$
```

```
# 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__)

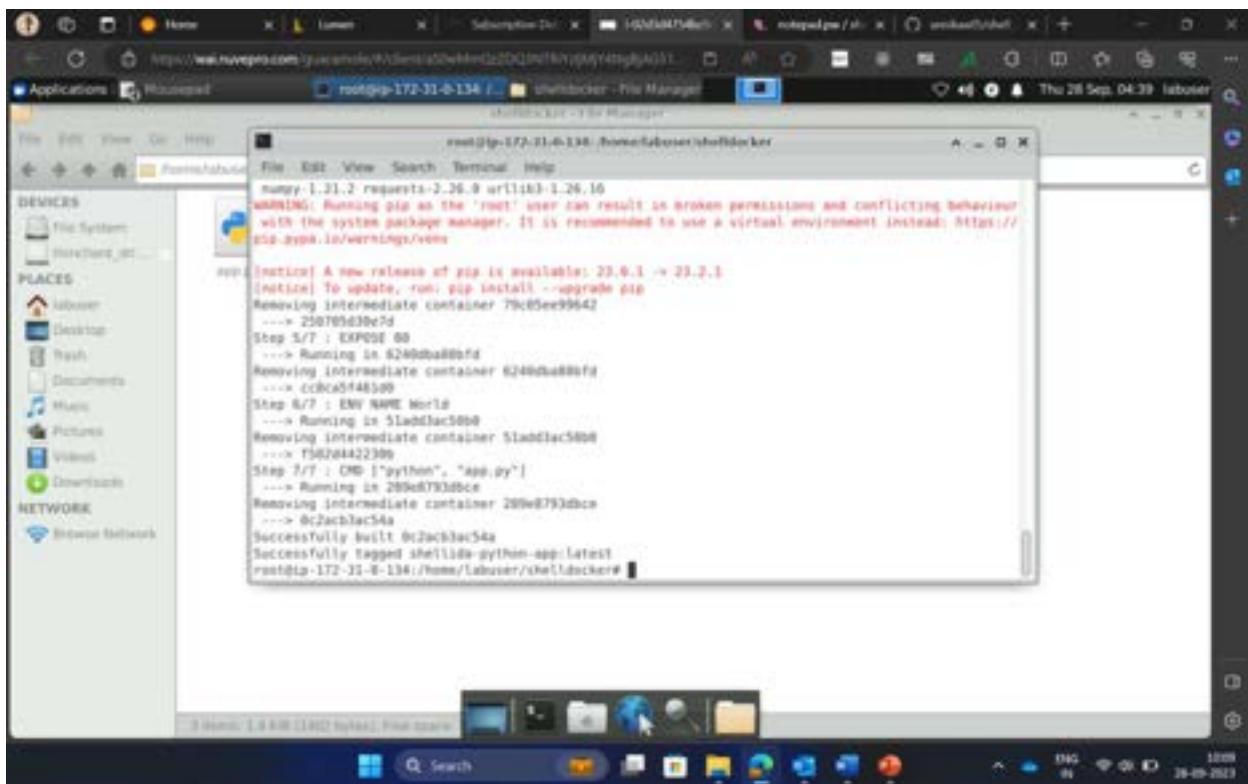
# Define routes and views
@app.route('/')
def index():
    return "Hello, World!"

# Example route with dynamic content
@app.route('/user/')
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"New submitted: {data}"
    return render_template('form.html')

# Run the application if this file is executed
if __name__ == '__main__':
    app.run()
```

```
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 .
```

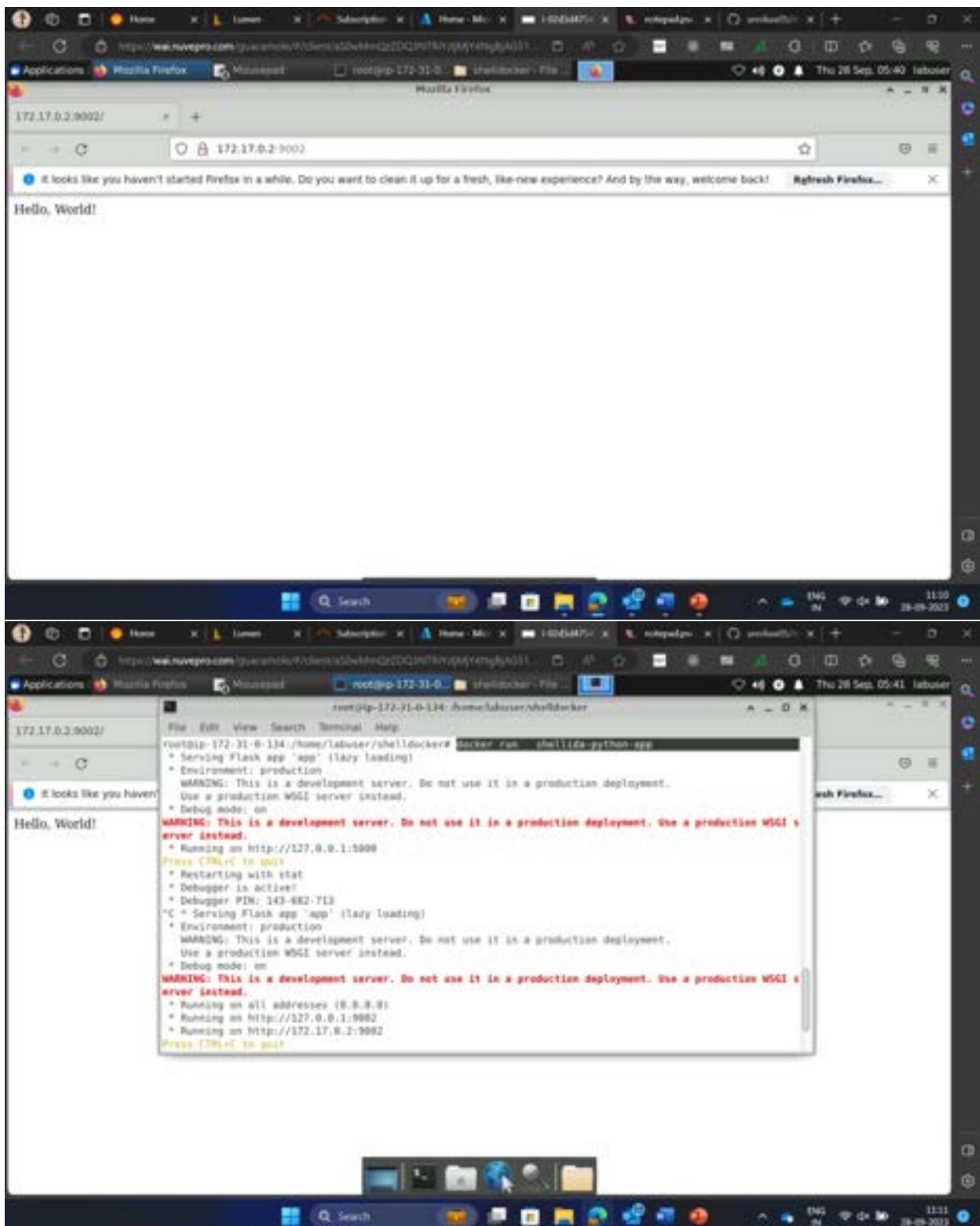
```
docker images
```

```
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
```



A screenshot of a Linux desktop environment, likely Kali Linux, showing a terminal window and a file manager.

The terminal window (root@ip-172-31-0-134:/home/labuser/shelllocker) displays the following command and its output:

```
root@ip-172-31-0-134:/home/labuser/shelllocker# az login --service-principal -u Shellnest_1693422082399@puneext.onmicrosoft.com -p c2-677a63a0241
az: syntax error near unexpected token `'
root@ip-172-31-0-134:/home/labuser/shelllocker# az login --service-principal -u Shellnest_1693422082399@puneext.onmicrosoft.com -p "c2-677a63a0241"
{
  "cloudName": "AzureCloud",
  "homeTenantId": "dc8f7315-8ffa-4a81-ab40-00e5a7214b21",
  "id": "39879c75-3888-4ec4-9514-79502a89798c",
  "isDefault": true,
  "managedByTenants": [],
  "name": "puneext-1693422082399",
  "state": "Enabled",
  "tenantId": "dc8f7315-8ffa-4a81-ab40-00e5a7214b21",
  "user": {
    "name": "Shellnest_1693422082399@puneext.onmicrosoft.com",
    "type": "user"
  }
}
root@ip-172-31-0-134:/home/labuser/shelllocker#
```

The desktop interface includes a taskbar with icons for various applications like a browser, file manager, and terminal, along with system status indicators at the bottom.

The image shows two screenshots of a Linux terminal session on a Windows host.

**Microsoft Azure Container Registries:**

- The top screenshot shows the Azure portal interface for "Container registries".
- The URL is <https://portal.azure.com/#view/Hubs/resources?resourceType=Microsoft.ContainerRegistry>.
- The search bar contains "Container registries".
- Filter options include "source group equals all", "location equals all", and "Add view".
- Columns: Name, Type, Resource group, Location, Subscription.
- A message says "No container registries to display".
- Description: "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."
- Buttons: "Create container registry", "Learn more", "Give feedback".

**Terminal Session:**

- The bottom screenshot shows a terminal window titled "root@ip-172-31-0-134: ~\$".
- The title bar also shows "shellcontainer" and "shellocker".
- The terminal output is:

```
File Edit View Search Terminal Help
root@ip-172-31-0-134:/home/labuser/shellocker# az login --shell -n ShellNext_3893422062398@openext.microsoft.com -p "c2-677a62a2d4"
bash: syntax error near unexpected token `)'
root@ip-172-31-0-134:/home/labuser/shellocker# az login --shell -n ShellNext_3893422062398@openext.microsoft.com -p "c2-677a62a2d4"
[{"cloudName": "AzureCloud",
 "homeTenantId": "dc8f7315-8ff1-4a81-ab48-8de5a7214b2f",
 "id": "30079c75-1868-4ec4-9514-795b2a89788c",
 "isDefault": true,
 "managedByTenants": [],
 "name": "openext_3893422062398",
 "state": "Enabled",
 "tenantId": "dc8f7315-8ff1-4a81-ab48-8de5a7214b2f",
 "user": {
 "name": "ShellNext_3893422062398@openext.microsoft.com",
 "type": "user"
 }
}
root@ip-172-31-0-134:/home/labuser/shellocker# az acr login --name accountname21
root@ip-172-31-0-134:/home/labuser/shellocker# sudo su
root@ip-172-31-0-134:/home/labuser/shellocker# az acr login --name accountname20
Login Succeeded
root@ip-172-31-0-134:/home/labuser/shelllocker#
```

```
root@ip-172-31-0-134:~# az acr login --name azcontainerdaily21
root@ip-172-31-0-134:~# sudo su
root@ip-172-31-0-134:~# /home/labuser/shelldocker# az acr login --name azcontainerdaily21
Login Succeeded
root@ip-172-31-0-134:~# docker tag shellida-python-app azcontainerdaily21.azurecr.io/ida-python-app:v1
root@ip-172-31-0-134:~# docker push azcontainerdaily21.azurecr.io/ida-python-app:v1
The push refers to repository [azcontainerdaily21.azurecr.io/ida-python-app]
c04193ebe58: Pushed
7ac88d358ff: Pushed
c938102aae3e: Pushed
efc5004ee77f: Pushed
9979a279e8ff: Pushed
bf54512b477f: Pushed
ae2055789c5e: Pushed
e2ef8a51399d: Pushed
v1: digest: sha256:7fd5a40e1a7af098fa3e57e2c4800880eb84649412b1530f4653dfe399cae58 size: 1996
root@ip-172-31-0-134:~# /home/labuser/shelldocker#
```

The screenshot shows the Microsoft Azure portal interface. The top navigation bar includes 'Home', 'Subscriptions', 'Access control (IAM)', 'Logs', 'Metrics', 'Events', 'Properties', 'Locks', and 'Services'. The main area displays the 'azcontainerdaily21' container registry. On the left, a sidebar lists 'Overview', 'Activity log', 'Access control (IAM)', 'Tags', 'Quick start', 'Events', 'Settings' (with sub-options like 'Access keys', 'Encryption', 'Identity', 'Networking', 'Microsoft Defender for Cloud', 'Properties', 'Locks', and 'Services'), and 'Services'. The right pane shows the 'Repositories' tab, which lists a single repository named 'ida-python-app'. A search bar at the top of the repository list contains the text 'Search for this repository...'. Below the repository list, there is a 'Cache Rule' section.

Microsoft Azure | Search resources, services, and docs (EN-US) ShellSession\_1691422082

Home > Kubernetes services > 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:

AKS pricing tier:

Kubernetes version\*:

[Previous](#) [Next: Node pools](#) [Review + create](#) [Give feedback](#)

Microsoft Azure | Search resources, services, and docs (EN-US) ShellSession\_1691422082

Home > Kubernetes services > Create Kubernetes cluster > Update node pool

Update node pool

Scale method:  Autoscale - Recommended  
This option is recommended so that the cluster is automatically sized correctly for the current running workload.

Minimum node count\*:

Maximum node count\*:   
The maximum node count allowed for an AKS cluster is 1000 per node pool and 3000 nodes across all node pools in this cluster.

Optional settings

Max pods per node\*:  30 - 250

Enable public IP per node:

Labels

Labels are key/value pairs that can be used to categorize or add identifying information to Kubernetes resources such as nodes.  
I usually like to associate my node pool with the account I want to use in the node pool. [Learn more](#)

[Update](#) [Cancel](#)

Microsoft Azure | Search resources, services, and data (F4) | ShellSession\_1693422082

Home > microsoftaks-20230928141522 | Overview > shelliddaday21

Kubernetes service

shelliddaday21

Search Create Connect Start Stop Delete Refresh Open in mobile Give feedback

Overview Essentials

Resource group: shelliddaday21

Status: Succeeded (Running)

Location: East US

Subscription: Microsoft Cloud (US)

Kubernetes version: 1.26.0

API server address: shelliddaday21-ebs-9c2992.hcp.eastus.azmk8s.io

Network type (plugin): Kubelet

Node pools: 1 node pool

Activity log Access control (IAM) Tags Diagnose and solve problems Microsoft Defender for Cloud

Tags (0) Add tags

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: shelliddaday21-ebs-9c2992.hcp.eastus.azmk8s.io

Network type (plugin): Kubelet

Port CIDR: 10.244.0.0/16

Service CIDR: 10.0.0.0/16

Namespaces Workloads Services and ingress Storage Configuration Custom resources Run command Settings

Node pools

Search

Microsoft Azure | Search resources, services, and data (F4) | ShellSession\_1693422082

Home > microsoftaks-20230928141522 | Overview > shelliddaday21 > Create a starter application

Deploy a quickstart application

Application details Review YAML Deploy

Next

Previous

Search

[https://portal.azure.com/#view/Microsoft\\_Azure\\_ConsoleServices/QuickstartAppDeployment](https://portal.azure.com/#view/Microsoft_Azure_ConsoleServices/QuickstartAppDeployment)

## Deploy a quickstart application

**Application details** **Review YAML** **Deploy**

**The application has been deployed**

The Azure VNet app consists of four Kubernetes resources which are created inside of a new namespace. You can see them in the list below, as well as their deployment status. You can click on the name of a deployed resource to view more details about that resource.

**1. 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
azure-vnet	Namespace	Success
azure-vnet-back	Deployment	Success
azure-vnet-front	Service	Success
azure-vnet-front	Deployment	Success

**Previous** **Close**

[https://portal.azure.com/#view/Microsoft\\_Azure\\_ConsoleServices/CreateCustomImage](https://portal.azure.com/#view/Microsoft_Azure_ConsoleServices/CreateCustomImage)

## 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.

**Container registry type:**  Azure Container Registry  Other registry

**Container 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.

**Repository:**

**Image tag:**

**OS type:**

**Previous** **Next**

[Home](#) > [microservices-20230528t141522](#) | [Overview](#) > [�elliadaily21](#) > [Create a starter application](#)

## 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-1681402160176	Namespace	Success
lola-python-app	Deployment	Success
lola-python-app-service	Service	Success
lola-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

The image shows two screenshots of the Azure DevOps interface. The top screenshot is the 'My Information' page for the user 'Shellunext' (represented by a red circle with 'SU'). It displays basic profile information: name, email, location ('United States'), and a large 'Get started with Azure DevOps' section featuring a rocket launch icon. The bottom screenshot is the 'Create a project to get started' page. It prompts the user to enter a 'Project name' and choose a 'Visibility' level ('Public' or 'Private'). A note states that public projects are disabled. At the bottom, there is a 'Create project' button.

Shellunext (SU)

Shellunext

unextIDA138

Shellunext\_1693422082390@npunext.onmicrosoft.com

United States

Get started with Azure DevOps

Plan better, code together, ship faster with Azure DevOps

Create new organization

Visual Studio Dev Essentials

Get everything you need to build and deploy your

Shellunext (SU)

New organization

Create a project to get started

Project name:

Visibility

Public

Private

Only people you give access to will be able to view this project.

Public projects are disabled for your organization. You can turn on public visibility with organization policies.

Create project

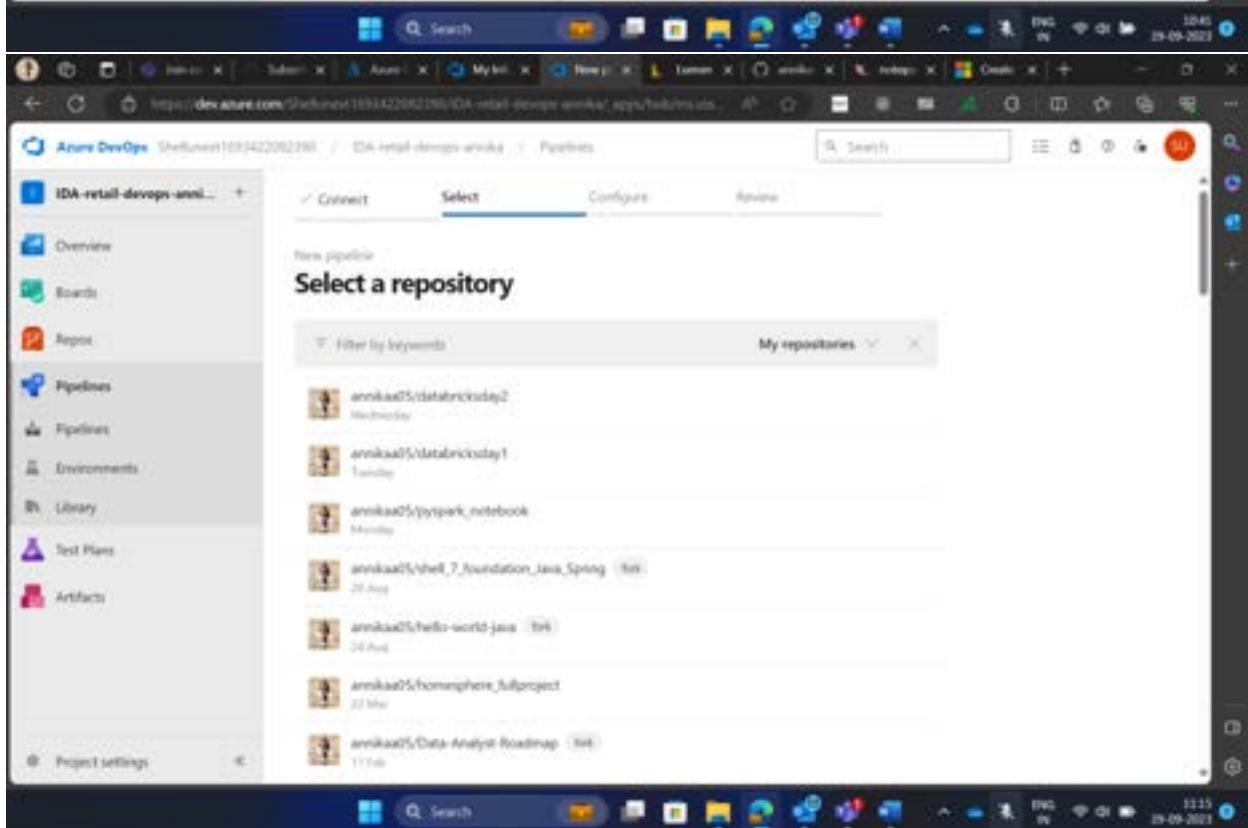
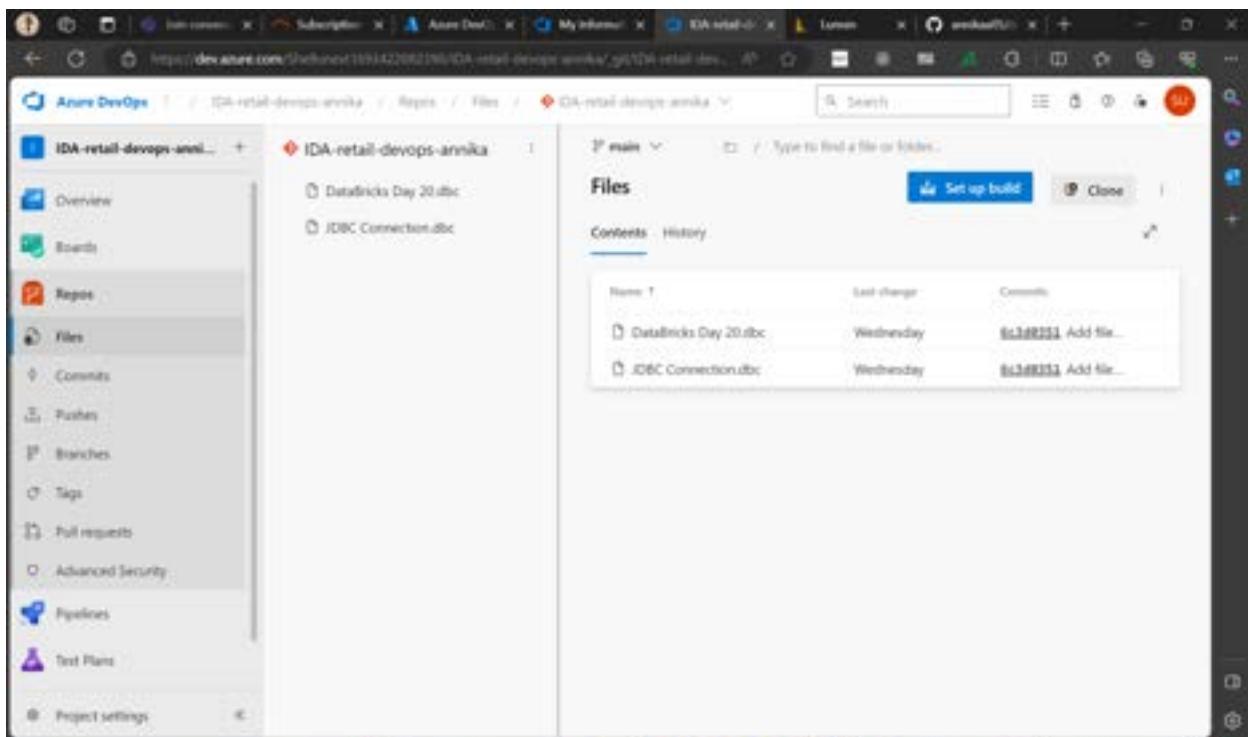
What's new

Sprint 227

Workload identity federation for Azure Pipelines is now in public preview. Check out the release notes for details.

Organization settings

The screenshot shows two windows of the Azure DevOps interface. The top window is the project overview for 'IDA-retail-devops-annika'. It features a sidebar with links like Overview, Summary, Dashboards, Wiki, Boards, Report, Pipelines, Test Plans, and Artifacts. The main area has a cartoon illustration of a person at a desk with a cat, and a 'Welcome to the project!' message. Below it are tabs for Boards, Repos, Pipelines, and Test Plans. The bottom window is a modal titled 'Import a Git repository'. It contains fields for 'Repository type' (set to 'Git'), 'Clone URL' (containing the GitHub URL 'https://github.com/annikaa05/databricksday2.git'), and a checkbox for 'Requires Authentication'. There are also sections for 'Generate Git Credentials', 'Push an existing repository from command line', 'Import a repository', and 'Initialize main branch with a README or gitignore'. Buttons for 'Cancel' and 'Import' are at the bottom right.



Anne DevOps ShellSession10342082390 IDA-retail-devops-anne... / Pipelines

Overview Boards Report 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

New pipeline Review your pipeline YAML

Variables Save and run

```
annika@172-16-1-11:~/python-sample-vscode-fail-tutorial$ azur...  
  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/pipelines/languages/python  
  
  trigger:  
    - none  
  
  pool:  
    vmImage: ubuntu-latest  
  strategy:  
    matrix:  
      Python27:  
        python.version: '2.7'  
      Python38:  
        python.version: '3.8'  
      Python39:  
        python.version: '3.9'  
      Python310:  
        python.version: '3.10'  
  steps:  
    - task: PythonTool@0  
      inputs:  
        pythonVersion: 'Python 3.7'  
        script: 'echo "Hello, world!"'
```

Anne DevOps IDA-retail-devops-anne... / Pipelines / annika05/python-sample-v... / 203309293

Overview Boards Report Pipelines Pipelines Environments Library Test Plans Artifacts Project settings

Jobs in run #203309293 annika05/python-sample-v... / vscode-fail-tutorial

Job Job Python27 Job Python38 Job Python39 Job Python310 Initialize job Checkout annika05 Use Python 3.7 install dependencies Python Post-job: Check... Finalize Job

pytest

View new log

```
annika@172-16-1-11:~/python-sample-vscode-fail-tutorial$ azur...  
  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/pipelines/languages/python  
  
  trigger:  
    - none  
  
  pool:  
    vmImage: ubuntu-latest  
  strategy:  
    matrix:  
      Python27:  
        python.version: '2.7'  
      Python38:  
        python.version: '3.8'  
      Python39:  
        python.version: '3.9'  
      Python310:  
        python.version: '3.10'  
  steps:  
    - task: PythonTool@0  
      inputs:  
        pythonVersion: 'Python 3.7'  
        script: 'echo "Hello, world!"'
```

Anne DevOps ShellServer1693422082390 / IDA-retail-devops-annika / Test Plans / Runs

Run 4 - 'Pytest results'

Run summary Test results Filter

Recent test runs

+ Test runs a. 'Pytest results'

Recent exploratory sessions

Summary

Completed 2 minutes ago. Took for 517 milliseconds.

Run type: Automated  
Owner: IDA-retail-devops-annika build Service (ShellServer1693422082390)

Tested build: 2023102913  
Release: Not available  
Release Stage: Not available  
Build platform: Not available  
Build flavor: Not available  
Test settings: Default  
MTM lab assignment: Not available

Comments: No comments.

Error message: No error message.

Attachments (1)

Outcome

Passed

Outcome by priority

Attachments (1)

https://dev.azure.com/ShellServer1693422082390/IDA-retail-devops-annika/\_TestManagement/Runs/RunId-48\_a-runChart?field=48\_a-runCharts

Boards:

Work items :

Epic: High-level features or objectives for your PySpark application.

Feature: Major functionalities or components within an Epic.

I

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

The image shows two screenshots of AI development platforms side-by-side.

**Top Screenshot: Teachable Machine**

This interface allows users to train machine learning models for image classification. It features two main sections for training data:

- Class 1:** Contains 2 image samples of a dog. Buttons for "Webcam" and "Upload" are visible.
- Class 2:** Contains 2 image samples of a cat. Buttons for "Webcam" and "Upload" are visible.

A central "Training" section displays the status "Model Trained". Below it is an "Advanced" dropdown menu. To the right, a preview window shows a puppy image with a confidence bar indicating a high probability for "Class 1" (orange bar) and a low probability for "Class 2" (pink bar). A "Preview" tab is selected at the top right.

**Bottom Screenshot: Azure AI Machine Learning Studio**

This screenshot shows the "Generative AI with Prompt flow" workspace titled "azureml-day23".

The left sidebar includes navigation links such as "All workspaces", "Home", "Model catalog", "Authoring" (with "Notebooks", "Automated ML", "Designer", and "Prompt flow" sub-options), "Assets" (with "Data", "Jobs", "Components", "Pipelines", "Environments", and "Models" sub-options), and "Customize view".

The main workspace displays several cards for generating AI models:

- Bring Your Own Data QnA:** Create flows for QnA with GPT3.5 using data from your own dataset files to make the answers more grounded for enterprise chat scenarios. Buttons: "Start" and "Clone".
- Bring Your Own Data Chat QnA:** Create flows for multi-round QnA with GPT3.5 using data from your own dataset files to make the answers more grounded for enterprise chat scenarios. Buttons: "Start" and "Clone".
- Ask Wikipedia:** QnA with GPT3.5 using information from Wikipedia to make your answers more grounded. Buttons: "Start" and "Clone".
- Chat with GPT:** Chat with GPT. Buttons: "Start" and "Clone".

Below these cards, a "Generative AI models" section lists four available models:

- openai-whisper-large**: Status: "OpenAI accepted".
- databricks-dolly-v2-12b**: Status: "Not generated".
- gpt-4-32k**: Status: "Out of memory".
- gpt-4**: Status: "Out of memory".

The bottom of the screen shows a Windows taskbar with various pinned icons and system status indicators.

The screenshot shows the Azure Machine Learning Studio interface. The left sidebar has sections for 'All workspaces', 'Home', 'Model catalog', 'Authoring' (with 'Notebooks' selected), 'Automated ML', 'Designer', and 'Prompt flow'. The 'Assets' section includes 'Data', 'Jobs', 'Components', 'Pipelines', 'Environments', and 'Models'. The main area shows a 'Notebooks' list under 'azureml-day23 / Notebooks'. A notebook titled 'azureml-getting-started' is selected, showing its code content: `# azuresdkforpy - get started with Python notebooks`. The right side features a large preview pane with the title 'Getting Started: training an image classification model' and a 'Learning Objectives' section listing tasks like connecting to workspace, preparing data, training a model, and deploying it.

## Getting Started: training an image classification model

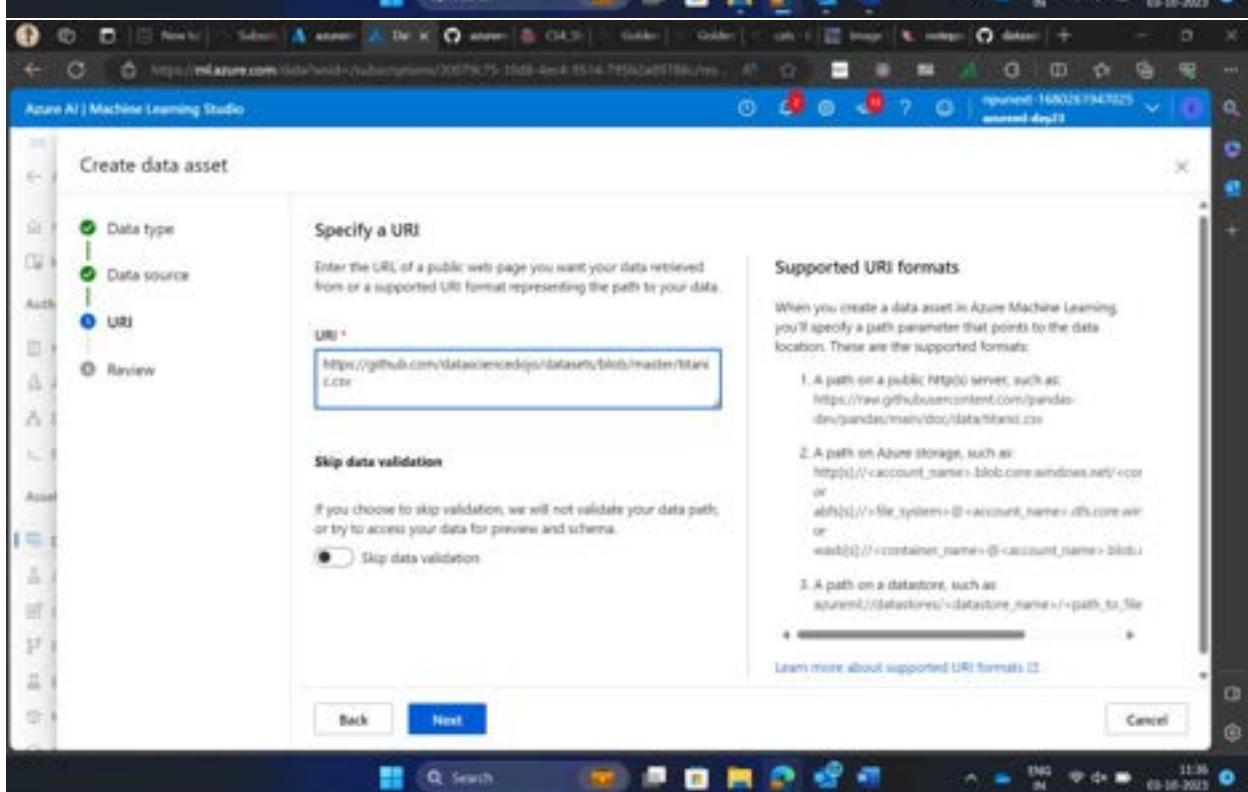
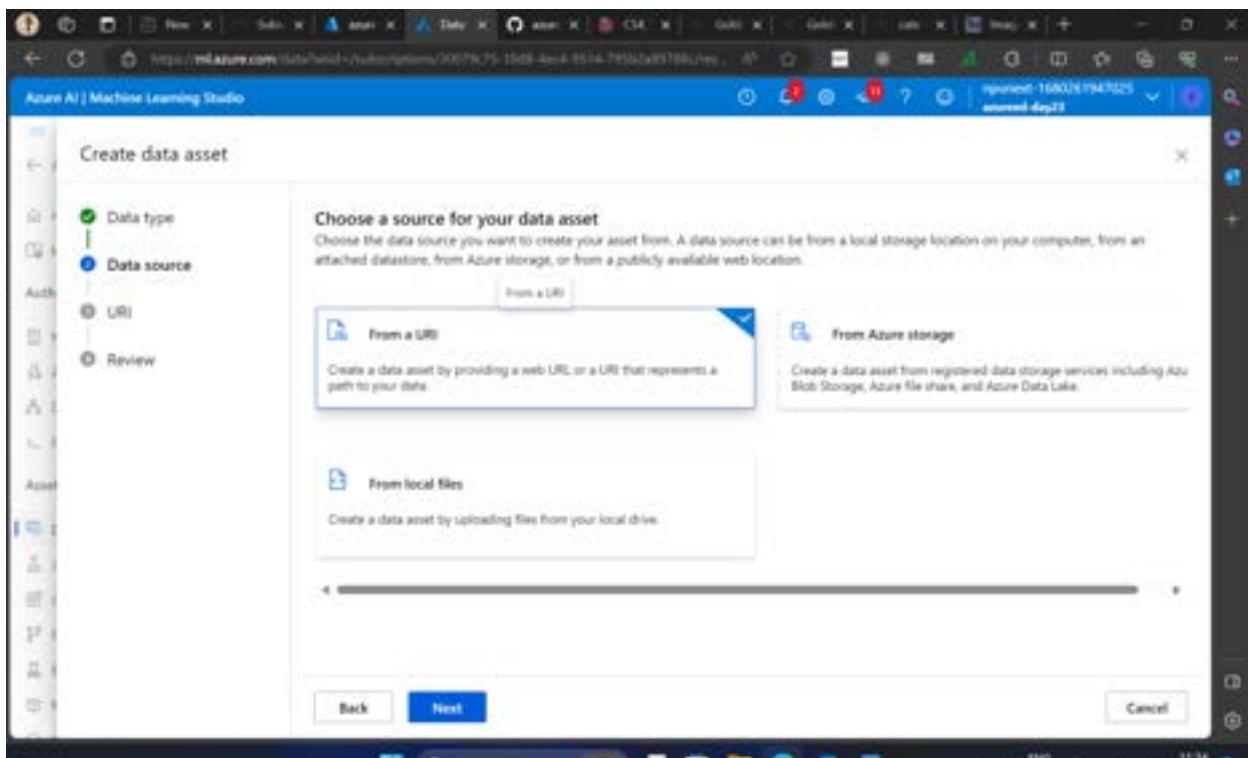
**Learning Objectives** · By the end of this quickstart tutorial, you'll know how to train and deploy an image classification model on Azure Machine Learning studio.

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

The screenshot shows the 'Create data asset' interface in the Azure AI Machine Learning Studio. On the left, a sidebar lists 'Assets' and other categories. The main area has a title 'Create data asset' and a navigation bar with 'Data type' (selected), 'Data source', 'Next', and 'Cancel' buttons. The 'Data type' section contains fields for 'Name' (set to 'Titanic-datasets-well'), 'Description' (empty), and 'Type' (set to 'File (uri\_file)'). To the right, a sidebar titled 'Use cases for data types' provides information about file and folder types, mentioning that the file type is recommended for most scenarios involving a single data file or tabular data. It also notes that the folder type is used for specific use cases.



The screenshot displays two separate windows of the Azure AI Machine Learning Studio interface, both showing the "Data" section.

**Top Window (titanic-datasets-web):**

- Attributes:**
  - Type: File (url\_file)
  - Named asset URI: [azuredl/titanic-datasets-web:1](#)
  - Created by: ShellUser: userID1018
  - Current version: 1
  - Lastest version: 1
  - Created time: Oct 3, 2023 11:38 AM
  - Modified time: Oct 3, 2023 11:38 AM
- Tags:** No tags
- Description:** Click with icon to add a description
- Data sources:** URL: <https://github.com/datascan/codjo/datasets/blobs/master/titanic.csv>

**Bottom Window (titanic-locodata):**

- Preview:**
  - Path: /AZ2023/.../Statistics
  - File Name: Statistics
  - Modified: 2023-10-03
  - Created: 2023-10-03
  - File Size: 80KB
  - File Format: CSV
  - CanSeek: True
- Preview Data:**

PassengerId	Survived	Pclass	Name	Sex	Age
1	0	3	Braund, Mr. Owen Harris	male	22
2	1	1	Collett, Mrs. J. Bruce	female	38
3	1	3	Heikkinen, Laina	female	26
4	1	1	Futrelle, Mrs. Jacques Heath	female	35
5	0	3	Allen, Mr. William Henry	male	35
6	0	3	Mitchell, Mr. John	male	NA
7	0	1	McCarthy, Mrs. Benjamin	male	54
8	0	3	Peterson, Master. Gustav	male	2
9	1	3	Johnson, Mrs. Oscar Wenzel	female	31

**Create data asset**

Data type      Set the name and type for your data asset

Name:

Description:

Type:  File

**Use cases for data types**

The File type is recommended in most scenarios when you are working with a single data file of any type (including tabular data). This type allows you to specify a file location by URL in a storage location on your local computer, an attached datastore, Blob/ADLS storage, or a publicly available HTTP/HTTPS location. There are many types of supported files in the Azure Machine Learning CLI v2 or Python SDK v2. This data type is called `uri_file`. Learn more about the `uri_file` type.

**When should I use File type?**

The File type has all the same capabilities and use cases as the Folder type, but is used when specifying a folder location. In the Azure Machine Learning CLI v2 or Python SDK v2, this data type is called `uri_folder`. Learn more about the `uri_folder` type.

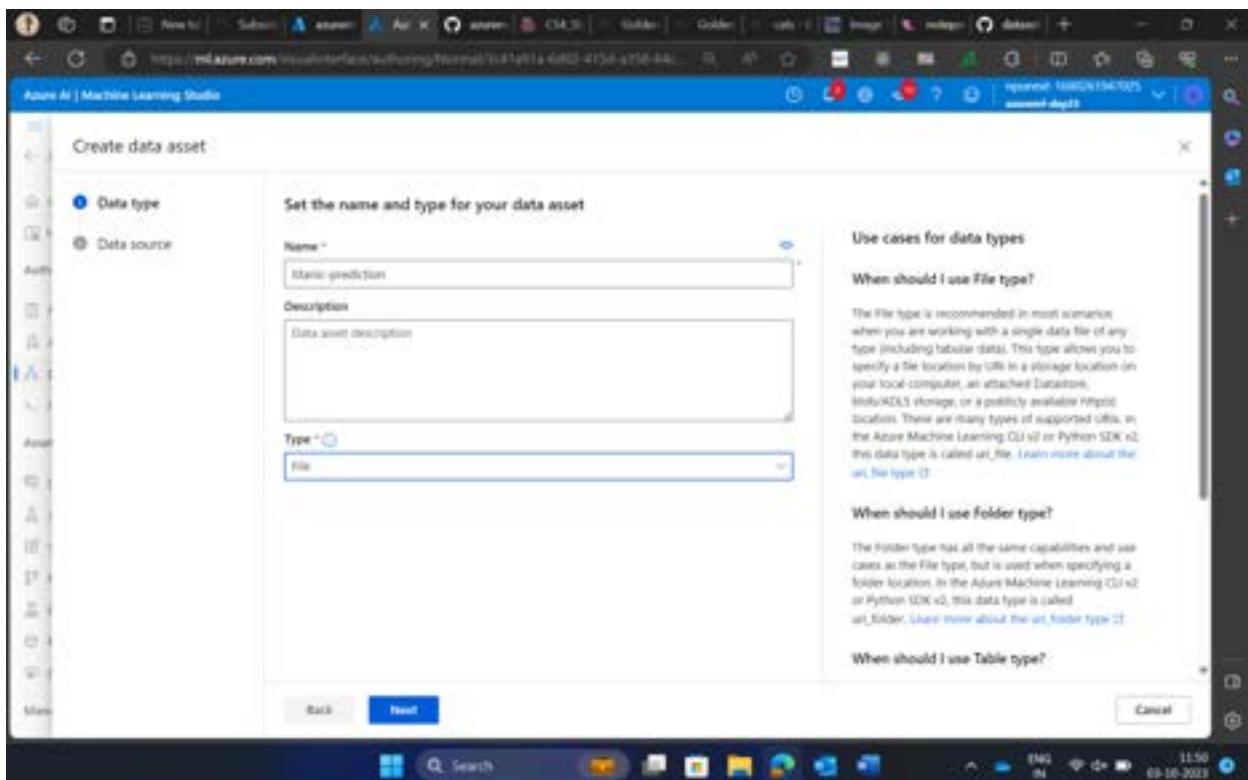
**When should I use Folder type?**

The Folder type has all the same capabilities and use cases as the File type, but is used when specifying a folder location. In the Azure Machine Learning CLI v2 or Python SDK v2, this data type is called `uri_folder`. Learn more about the `uri_folder` type.

**When should I use Table type?**

**Cancel**

**Back** **Next**



**Create data asset**

Data type      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**

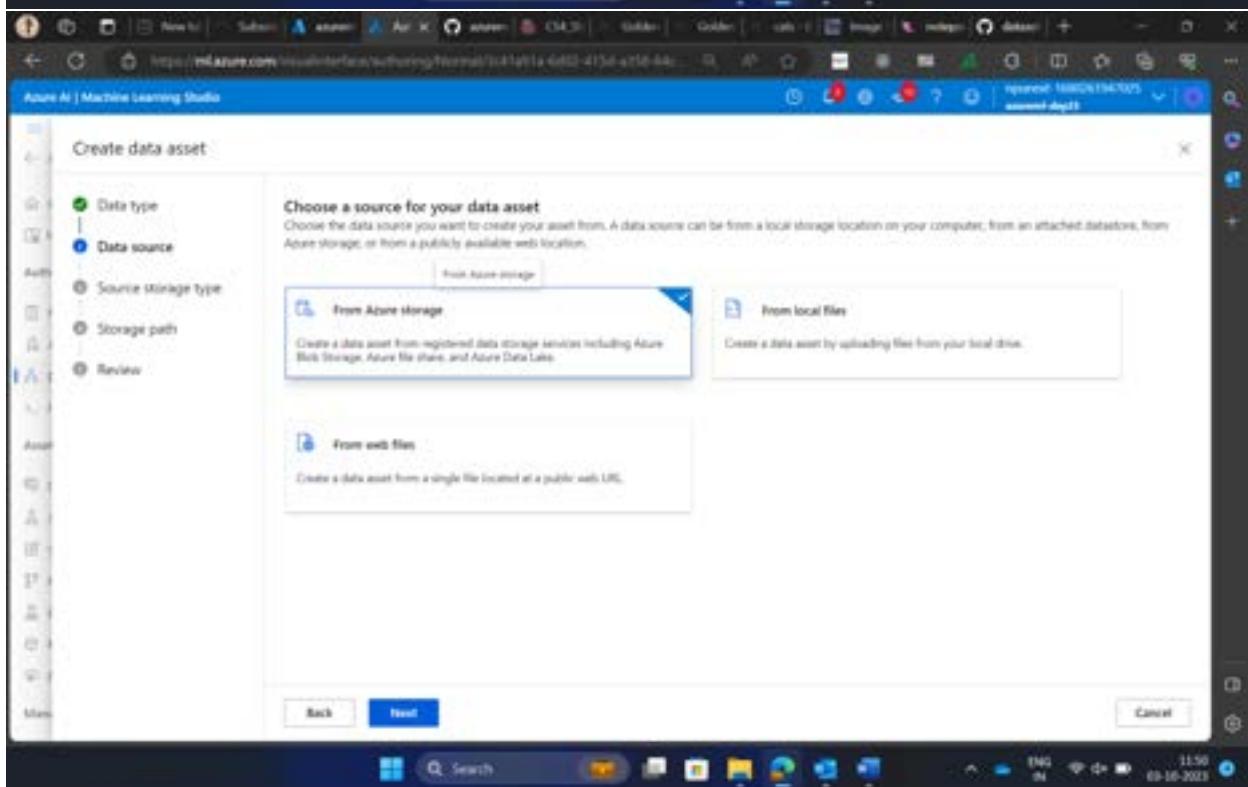
Create a data asset from registered data storage services including Azure Blob Storage, Azure File shares, and Azure Data Lake.

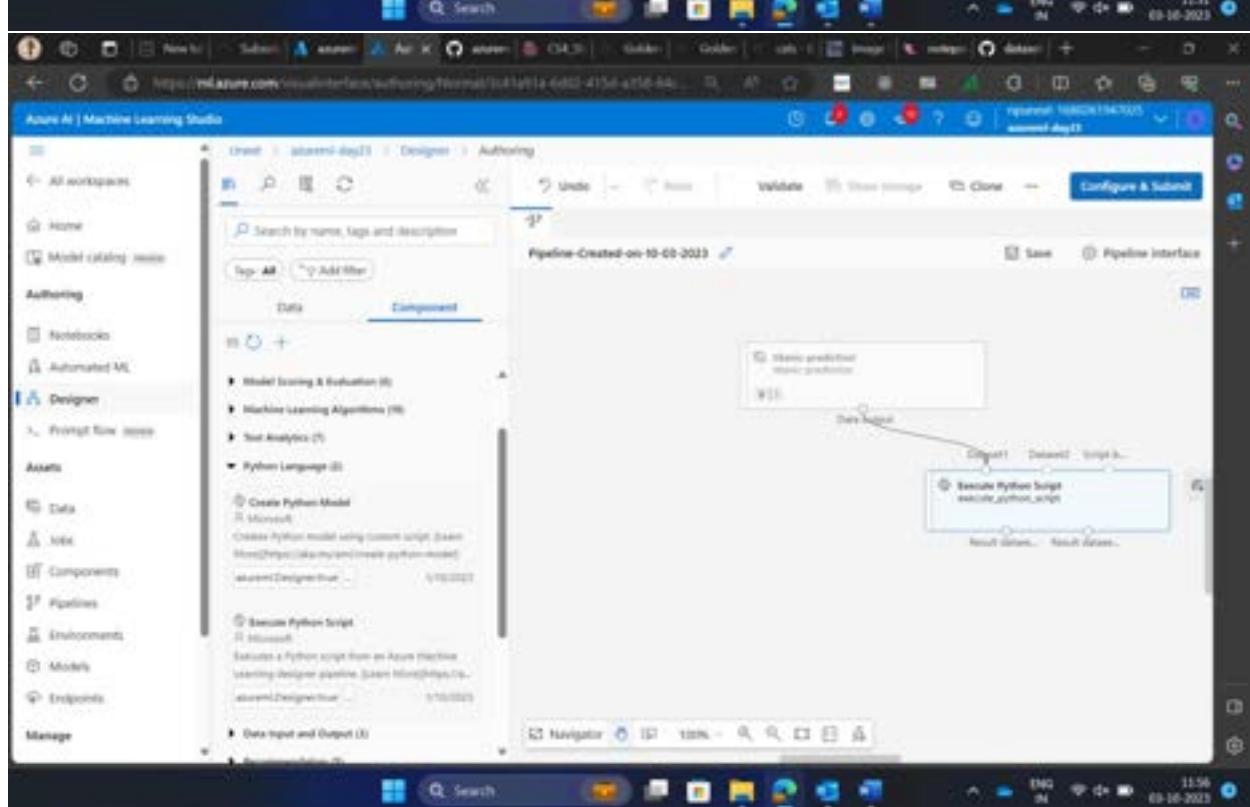
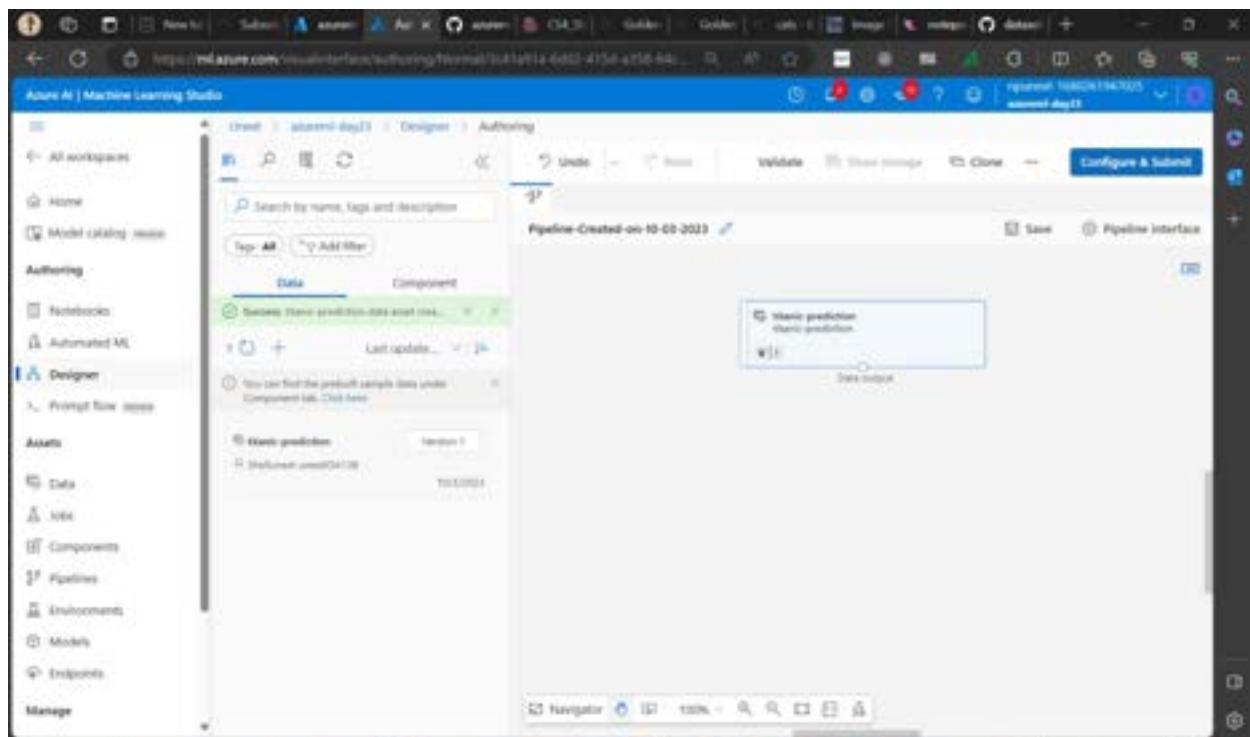
**From local files**

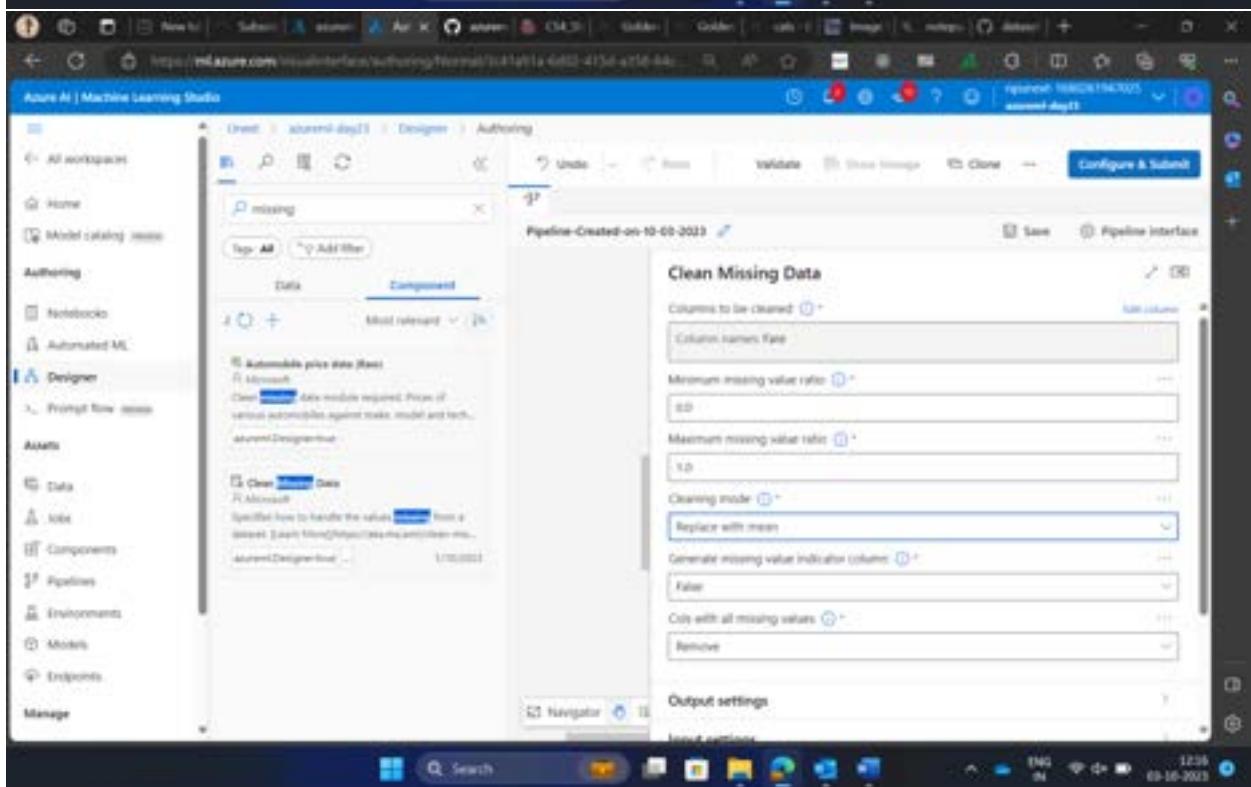
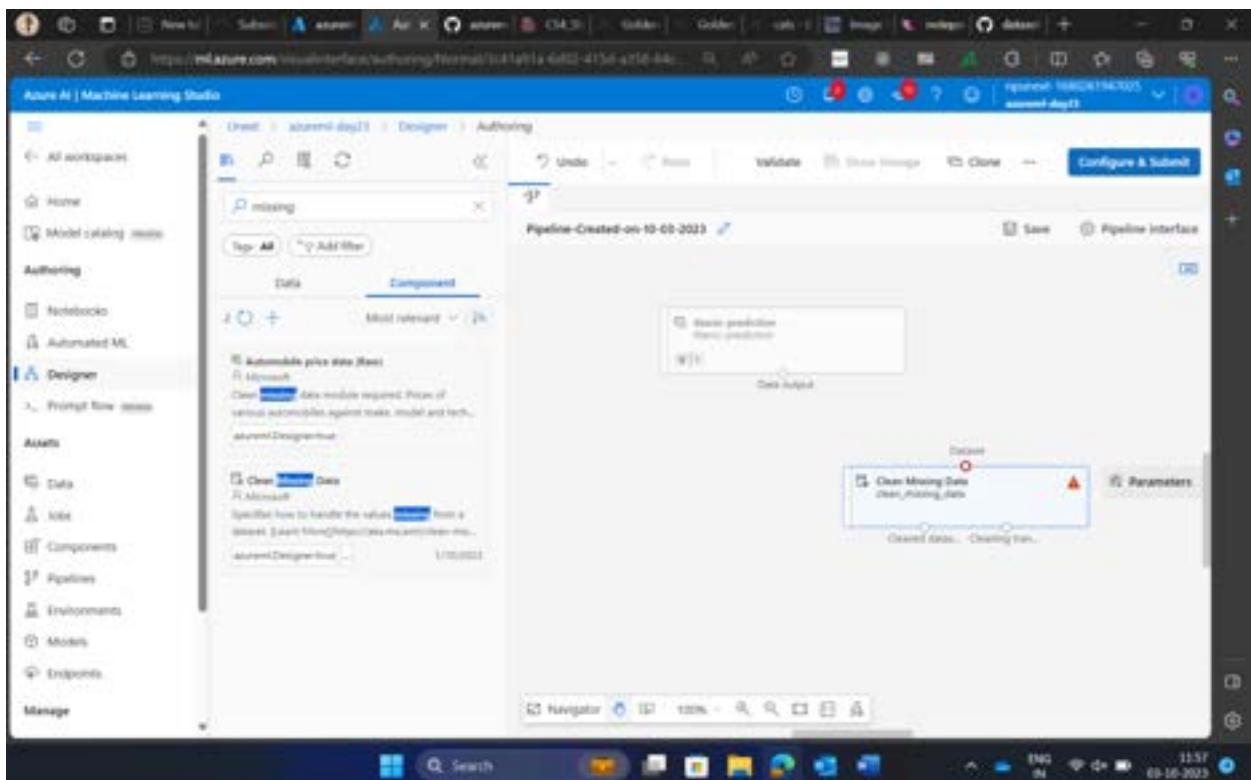
Create a data asset by uploading files from your local drive.

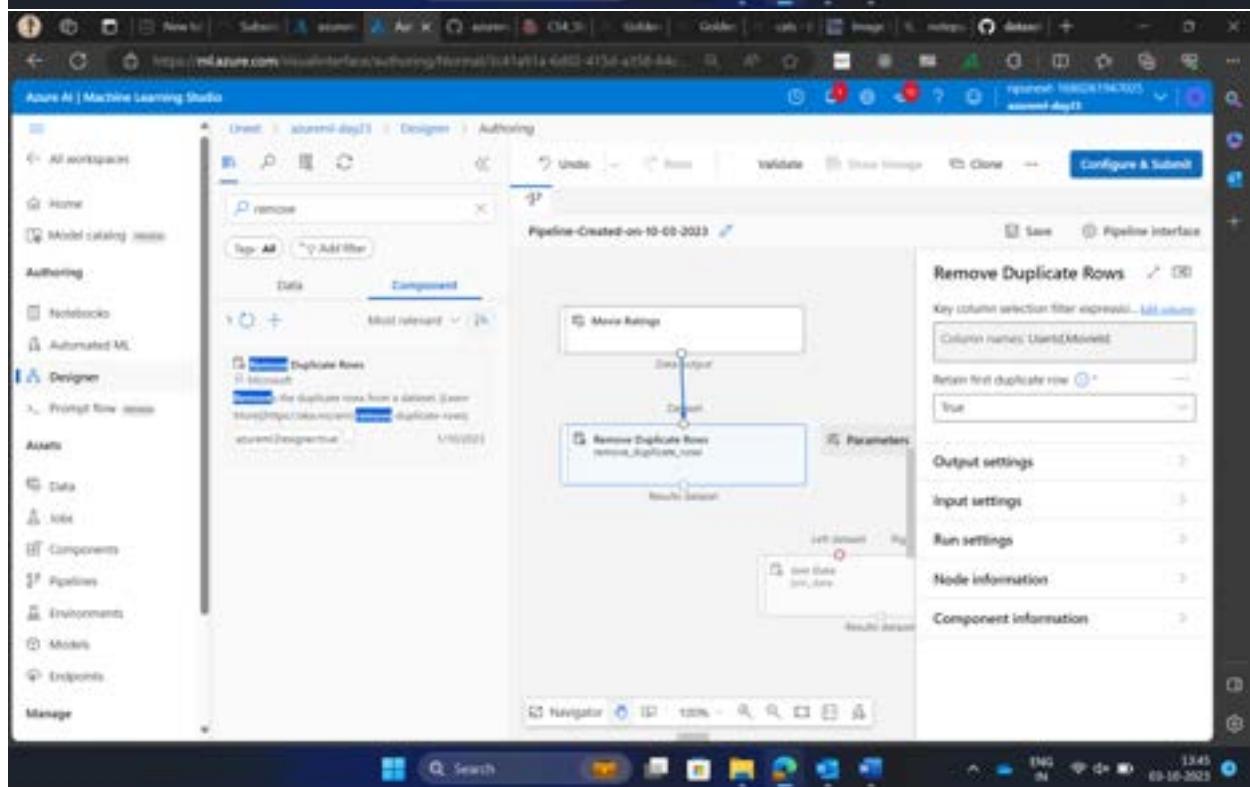
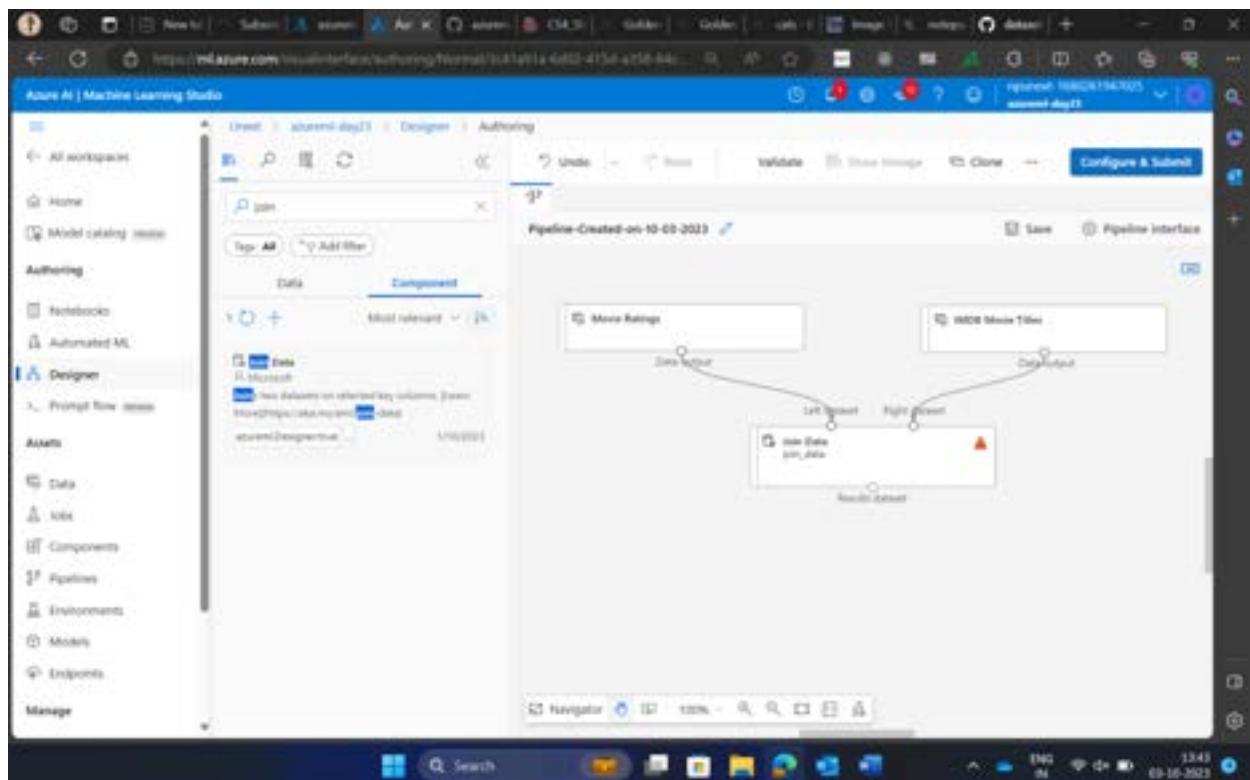
**From web files**

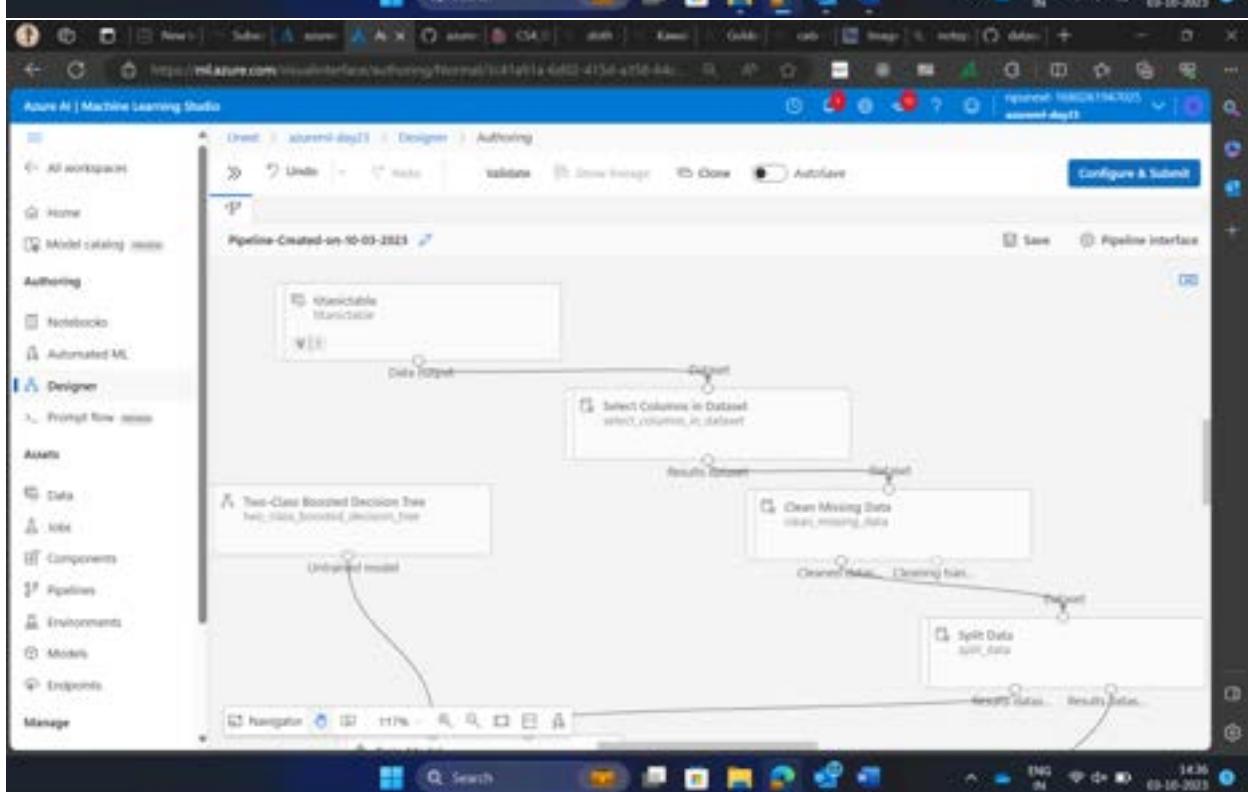
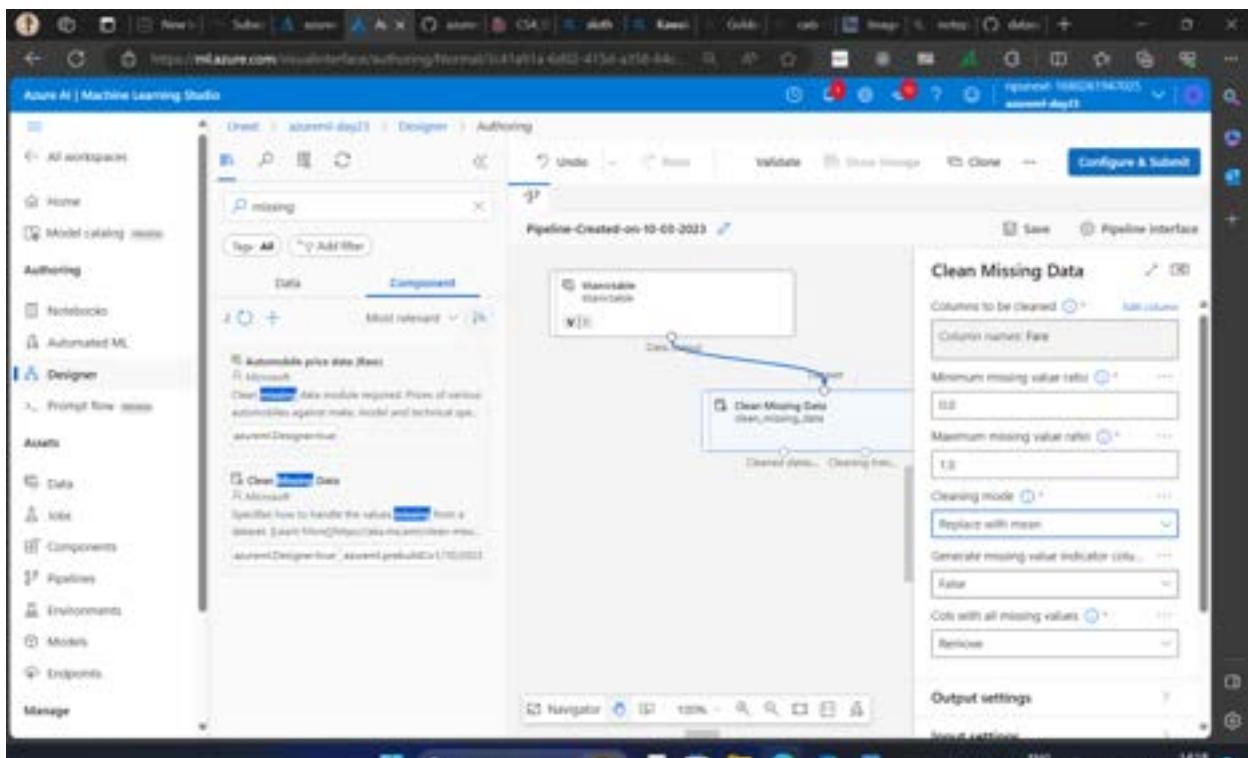
Create a data asset from a single file located at a public web URL.

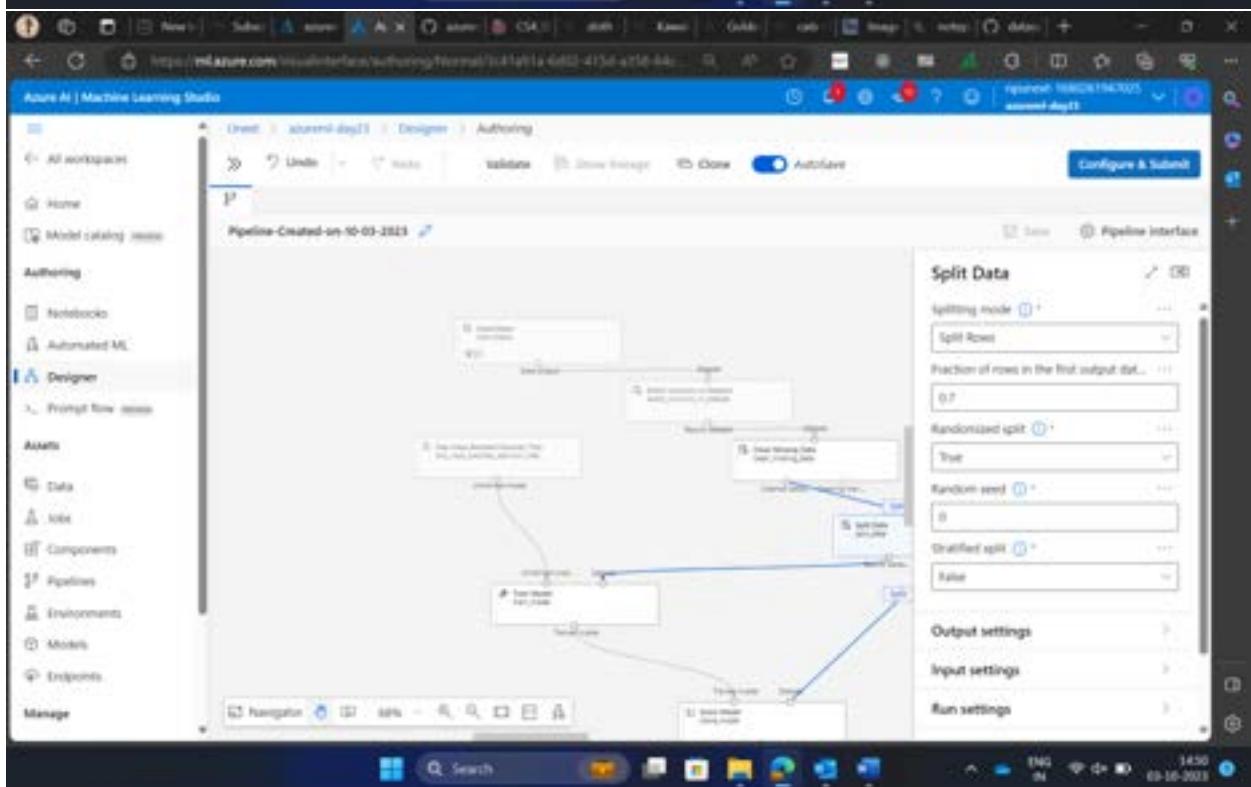
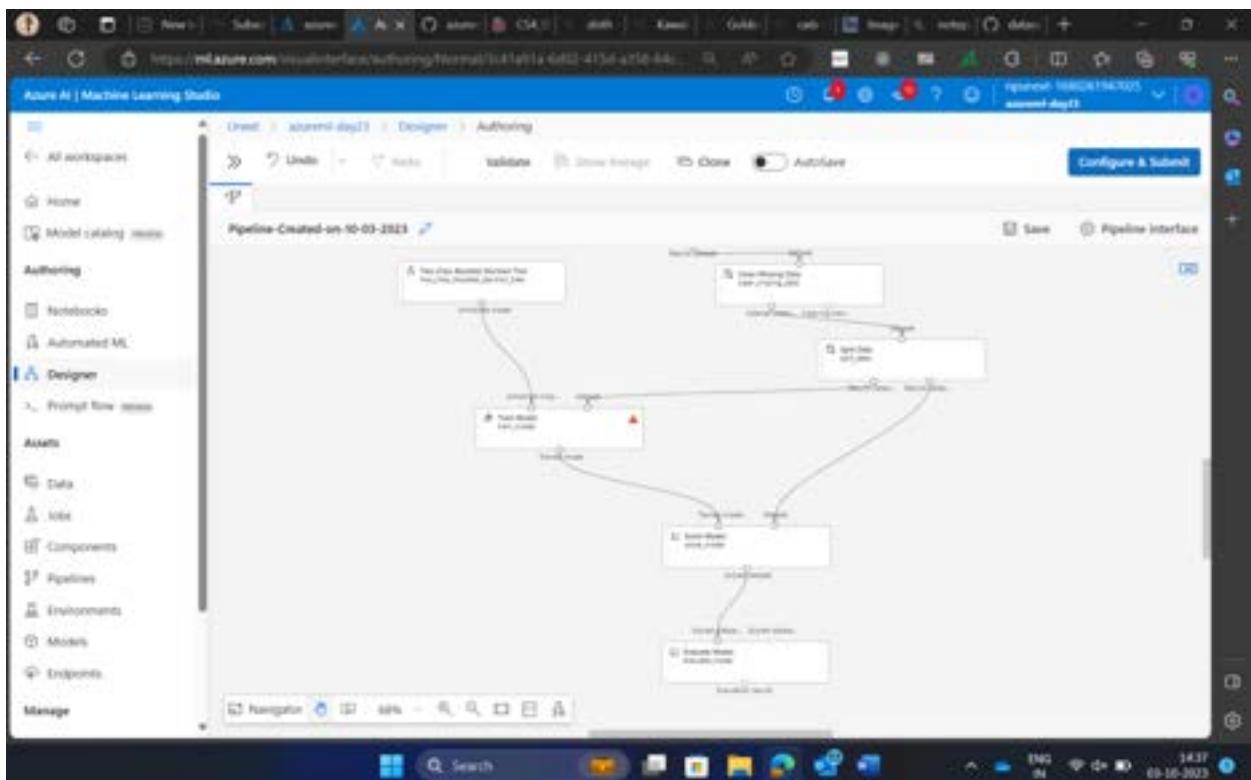


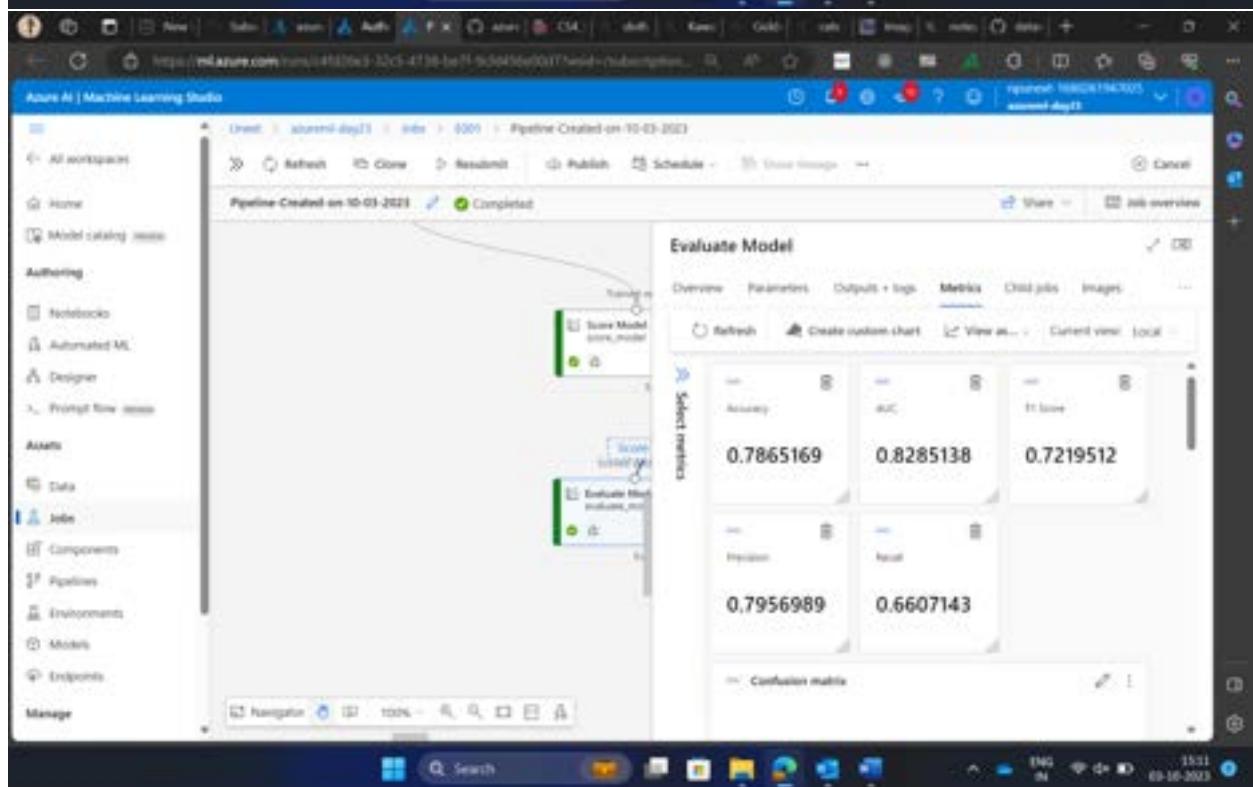
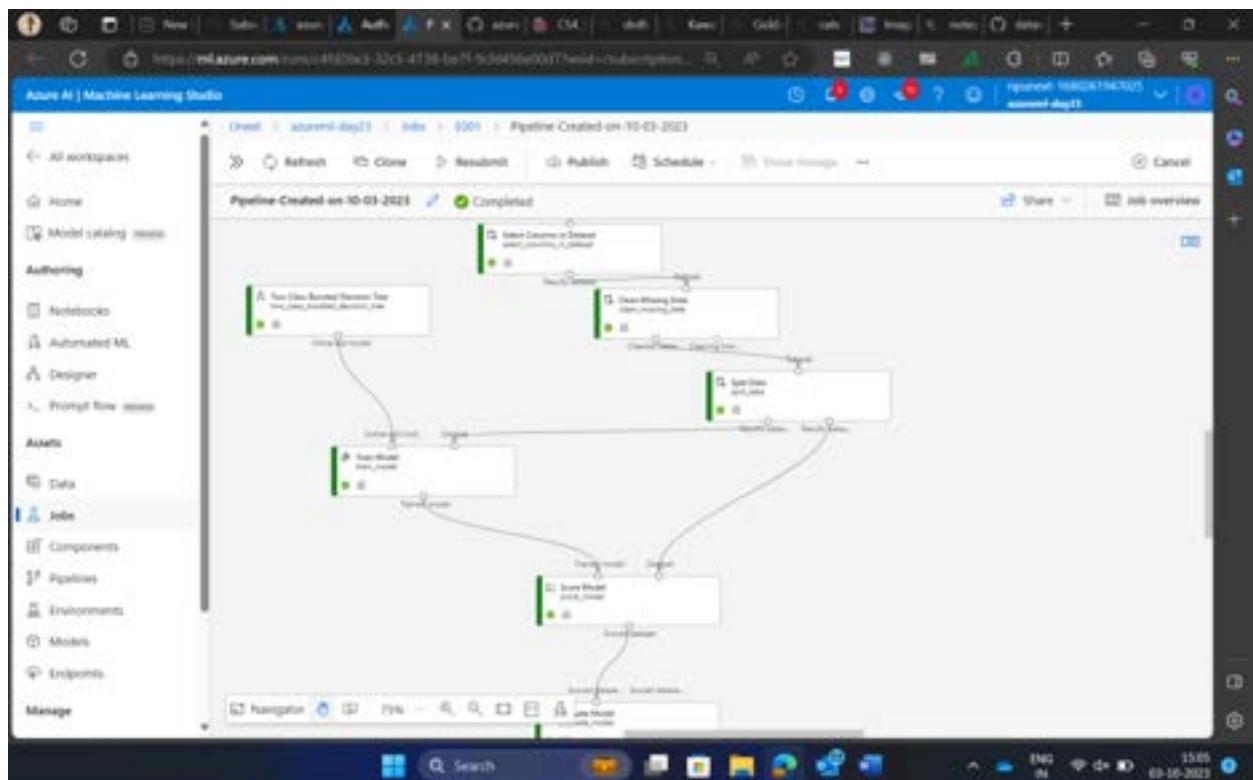












Azure AI | Machine Learning Studio

https://ml.azure.com/runs/01000.../ai/mlops/2545702?view=script

Pipeline Created on 10-03-2023

Completed

Evaluate Model

This job is using the new compute instance to improve performance. This can impact to use a higher number of cores.

Overview Parameters Outputs + logs Metrics Child jobs Images

Refresh Create custom chart View as... Current view Local

Select metric

	Accuracy	AUC	F1 score
Test	0.8127341	0.8774770	0.7524752
Training	0.8444444	0.6785714	

Confusion matrix

```
graph LR; TestData[Test Data] --> TestModel[Test Model]; TestData --> TrainedModel[Trained model]; TestModel --> Accuracy1[Accuracy]; TestModel --> AUC1[AUC]; TestModel --> F1Score1[F1 score]; TrainedModel --> Accuracy2[Accuracy]; TrainedModel --> AUC2[AUC]; TrainedModel --> F1Score2[F1 score]
```

Day 24

Screenshot of the Azure AI | Machine Learning Studio interface showing the "Create a new Automated ML job" wizard. The current step is "Select data asset". A success message indicates that the "customer-analysis" data asset was created successfully. The data assets list shows one entry:

Name	Dataset type	Created on	Modified on
customer-analysis	Tabular	04/10/2023 10:18...	04/10/2023 10:18...

Screenshot of the Azure AI | Machine Learning Studio interface showing the "Create a new Automated ML job" wizard. The current step is "Configure job". The configuration details are as follows:

- Data asset: customer-analysis (view data asset)
- Experiment name: E001
- Target column: C\_ACCTBAL (Decimal)
- Select compute type: Compute cluster
- Select Azure ML compute cluster: ml-dsp24 compute

