

Mallu Sravana Sandhya
Azure Databricks Coding Assignment
06/01/24

1. Create a cluster & Attach the notebook to the cluster and run all commands in the notebook & creates a DataFrame from a Databricks dataset & Create a Visualizations in Databricks notebooks & Rename, duplicate, or remove a visualization or data profile.

First we have to sign in to the azure portal and go Azure data bricks Service as shown in below and click on create to create an azure databricks workspace

The screenshot shows the Azure Databricks workspace list page. The page title is "Azure Databricks" and the user is "IHHT (ihhtl@microsoft.com)". The page has a "Create" button and a "Manage view" button. Below these, there are filters: "Subscription equals all", "Resource group equals all", and "Location equals all". The page shows 2 records. The table has columns: Name, Type, Resource group, Location, and Subscription. The records are:

Name	Type	Resource group	Location	Subscription
kaishashdb	Azure Databricks Service	rg-azuser926_mml.local-M0g0t	Central India	Azure subscription 1
sarfaraz934	Azure Databricks Service	Az500test	East US	Azure subscription 1

Give the values in the fields of resource group, workspace name and manageresouce group name as shown below then click on review+create

The screenshot shows the "Create an Azure Databricks workspace" form. The form has the following fields:

- Subscription: Azure subscription 1
- Resource group: rg-azuser921_mml.local-yVpvZ
- Instance Details:
 - Workspace name: AzureDB-921
 - Region: Central India
 - Pricing Tier: Premium (+ Role-based access controls)
- Managed Resource Group name: AzureDB-921

At the bottom, there is a "Review + create" button and a "Next : Networking >" button.

Then click on create button

Home > Azure Databricks >

Create an Azure Databricks workspace

Validation Succeeded

Workspace name: AzureDB-921
Subscription: Azure subscription 1
Resource group: rg-azuser921_mml.local-yvpvZ
Region: Central India
Pricing Tier: premium
Managed Resource Group name: AzureDB-921

Networking

Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP): No
Deploy Azure Databricks workspace in your own Virtual Network (VNet): No

Encryption

Enable Infrastructure Encryption: No
Enable CMK for Managed Disks: No

Create < Previous Download a template for automation

We can see that deployment is in progress

Home >

rg-azuser921_mml.local-yvpvZ_AzureDB-921 | Overview

Deployment

Search

Delete Cancel Redeploy Download Refresh

Overview

Inputs

Outputs

Template

Deployment is in progress

Deployment name: rg-azuser921_mml.local-yvpvZ_AzureDB-921 Start time: 06/01/2024, 10:09:45
Subscription: Azure subscription 1 Correlation ID: 0848dcf8-f68c-4626-a051-4d8d33a19f...

Resource group: rg-azuser921_mml.local-yvpvZ

Deployment details

Resource	Type	Status	Operation details
AzureDB-921	Azure Databricks Service	Created	Operation details

Give feedback

Tell us about your experience with deployment

Microsoft Defender for Cloud

Secure your apps and infrastructure

Go to Microsoft Defender for Cloud >

Free Microsoft tutorials

Start learning today >

Work with an expert

Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support.

Find an Azure expert >

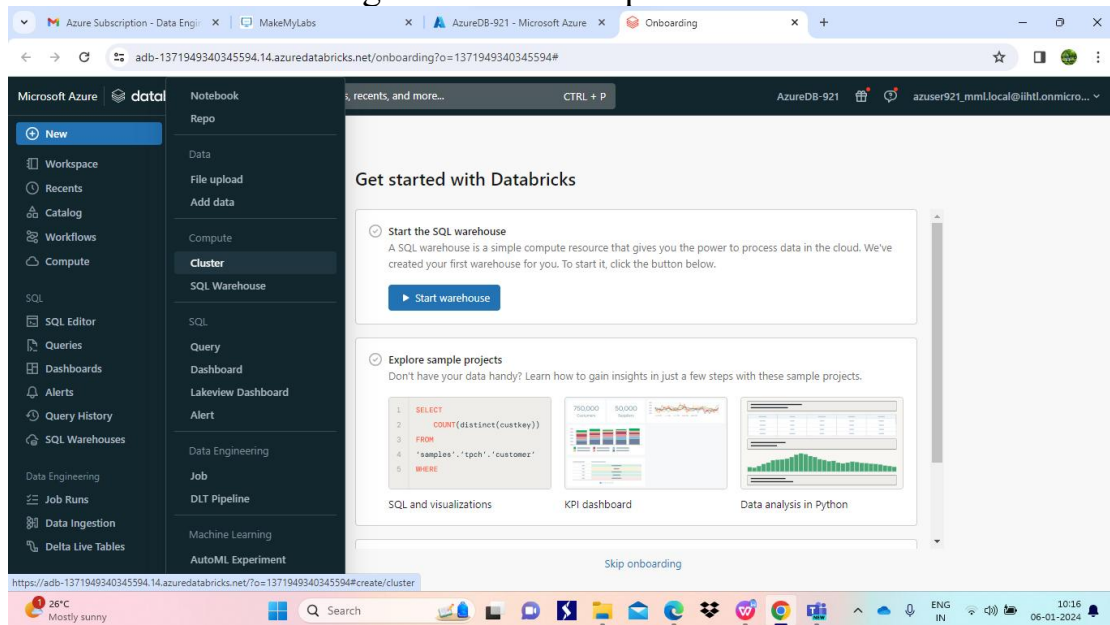
After completion of deployment of Azure databricks workspace, click on go to resource

The screenshot shows the Microsoft Azure portal interface. The top navigation bar includes the Microsoft Azure logo and a search bar. The main content area displays the deployment details for the resource 'rg-azuser921_mml.local-yVpvZ_AzureDB-921'. A green checkmark indicates that the deployment is complete. The deployment name is 'rg-azuser921_mml.local-yVpvZ_AzureDB-921', the subscription is 'Azure subscription 1', and the resource group is 'rg-azuser921_mml.local-yVpvZ'. The start time is '06/01/2024, 10:09:45'. A 'Go to resource' button is visible. On the right, there are links for 'Cost management', 'Microsoft Defender for Cloud', 'Free Microsoft tutorials', and 'Work with an expert'. The bottom status bar shows the weather as '26°C Mostly sunny' and the time as '10:15 06-01-2024'.

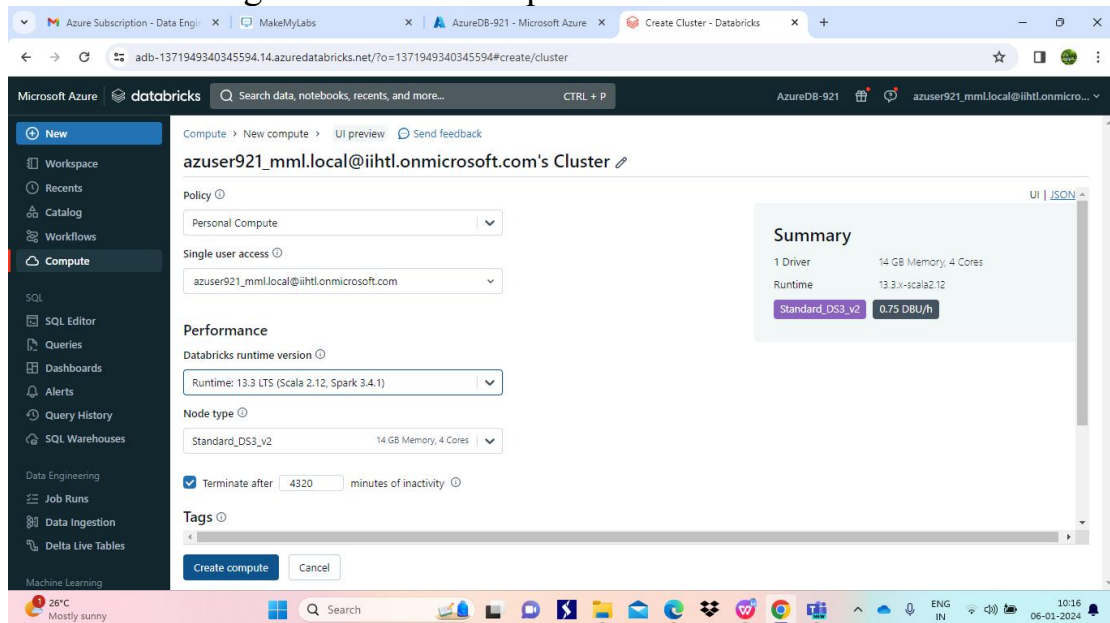
Now click on launch workspace

The screenshot shows the Microsoft Azure portal interface for the Azure Databricks workspace 'AzureDB-921'. The top navigation bar includes the Microsoft Azure logo and a search bar. The main content area displays the workspace details. The status is 'Active', the resource group is 'rg-azuser921_mml.local-yVpvZ', the location is 'Central India', the subscription is 'Azure subscription 1', and the subscription ID is '984f097c-963c-4eb6-a20d-839457ae9f08'. A 'Launch Workspace' button is visible. The bottom status bar shows the weather as '26°C Mostly sunny' and the time as '10:15 06-01-2024'.

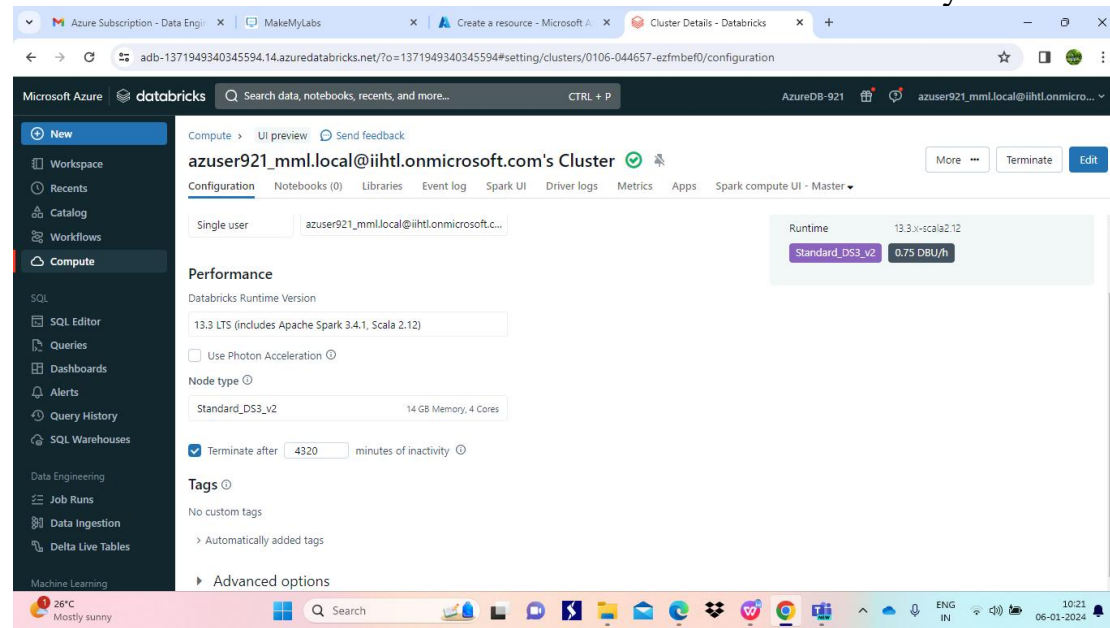
After launching of ADB workspace we can see as shown in the figure , now to create a cluster go to new on left panel and select cluster



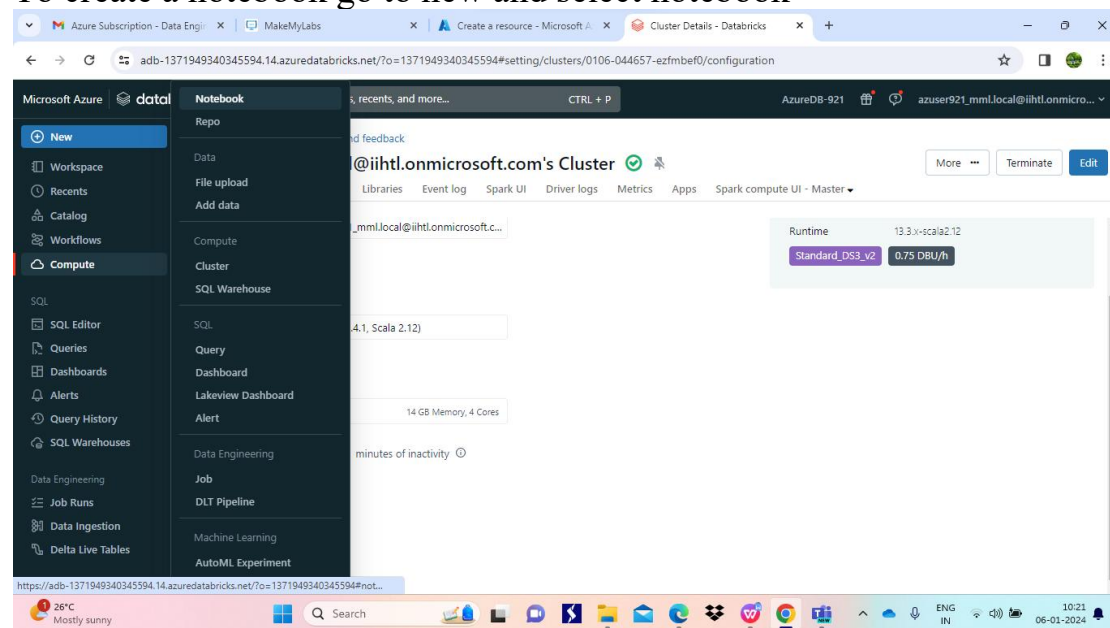
Now fill fields policy as personal compute and select databricks run version and give time to automatically terminate cluster after limited time of not using cluster which helps to save our credit in azure account



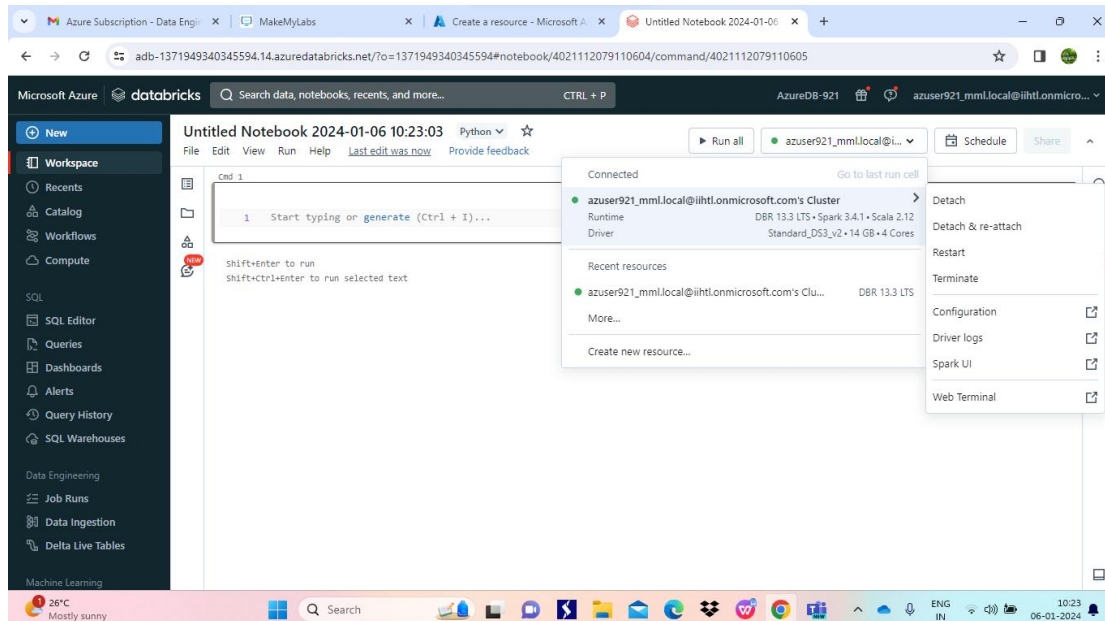
Then click on create and here the cluster is created successfully



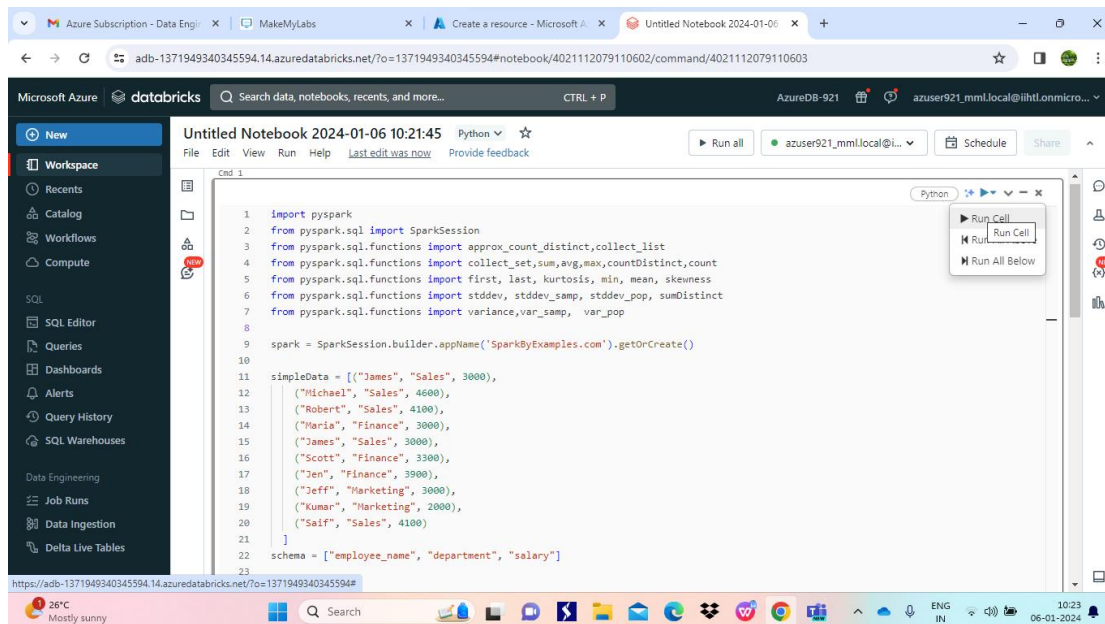
To attach the cluster to notebook we have to create a note book
To create a notebook go to new and select notebook



After creating a notebook we have to attach the cluster to the notebook in order to run the commands on notebook to attach cluster go to select cluster and click on dropdown and now select the cluster that we have created earlier



Now give some commands on notebook here I wrote a code to perform aggregate operations on table and click on run button to execute those commands



The screenshot shows a Databricks notebook interface with the following content:

```
51 | stddev_pop("salary").show(truncate=False)
52 | df.select(sum("salary")).show(truncate=False)
53 | df.select(sumDistinct("salary")).show(truncate=False)
54 | df.select(variance("salary"), var_samp("salary"), var_pop("salary")) \
55 | .show(truncate=False)
```

Below the code, the output of the last command is displayed as a table:

sum(salary)
765.9416862050705

The interface includes a sidebar with navigation options like Workspace, Recents, Catalog, Workflows, Compute, SQL, and more. The top bar shows the notebook title and various controls.

creating a DataFrame from a Databricks dataset create a new notebook and write the code as shown below and click on run we get the dataset in table

The screenshot shows a Databricks notebook interface with the following content:

```
1 | sparkDF=spark.read.csv("/databricks-datasets/bikeSharing/data-001/day.csv",header="true", inferSchema="true")
2 | display(sparkDF)
```

Below the code, the output is displayed as a table:

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp
1	2011-01-01	1	0	1	0	6	0	2	0.34416
2	2011-01-02	1	0	1	0	0	0	2	0.36347
3	2011-01-03	1	0	1	0	1	1	1	0.19636
4	2011-01-04	1	0	1	0	2	1	1	0.2
5	2011-01-05	1	0	1	0	3	1	1	0.22695
6	2011-01-06	1	0	1	0	4	1	1	0.20434
7	2011-01-07	1	0	1	0	5	1	2	0.19651

The interface includes a sidebar with navigation options like Workspace, Recents, Catalog, Workflows, Compute, SQL, and more. The top bar shows the notebook title and various controls.

To create a visualization click on the + symbol beside the table and select visualization as shown below

The screenshot shows the Databricks workspace with a notebook titled 'Untitled Notebook 2024-01-06 10:23:03'. The notebook contains a Python code cell that reads a CSV file from the Databricks datasets and displays the resulting DataFrame. The DataFrame has 731 rows and 14 columns. A visualization menu is open, showing options like 'Table', 'Data Profile', and 'Visualization'. The 'Visualization' option is selected, and a 'New result table: OFF' dropdown is visible.

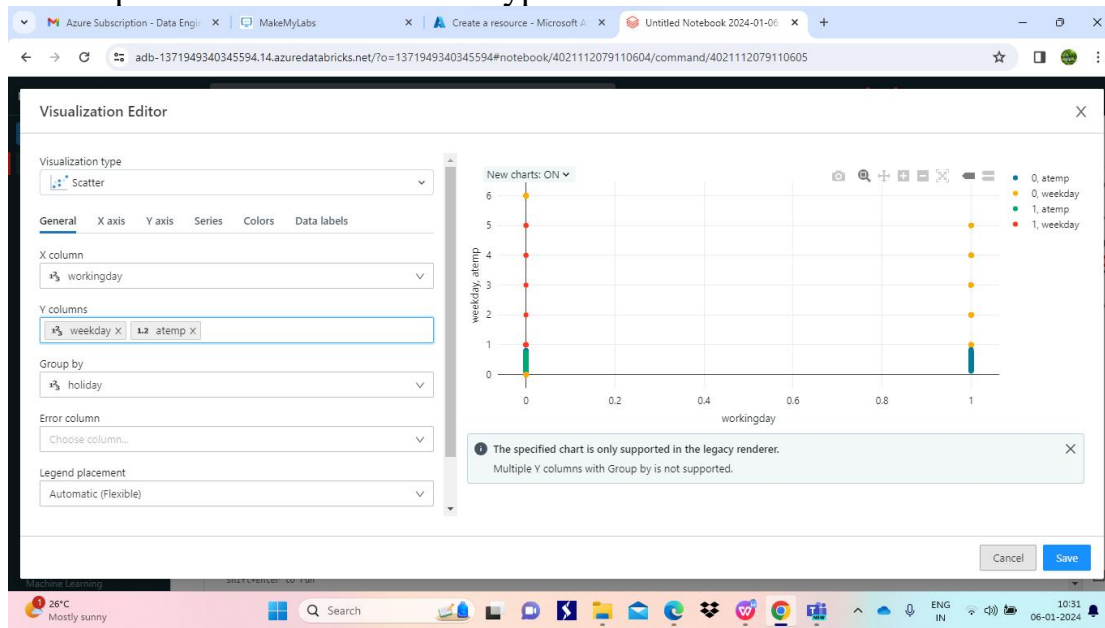
instant	ty	season	yr	mnth	holiday	weekday	workingday	weathersit	temp
1	1	2011-01-01	1	0	1	0	6	0	0.34416
2	2	2011-01-02	1	0	1	0	0	2	0.36347
3	3	2011-01-03	1	0	1	0	1	1	0.19636
4	4	2011-01-04	1	0	1	0	2	1	0.2
5	5	2011-01-05	1	0	1	0	3	1	0.22695
6	6	2011-01-06	1	0	1	0	4	1	0.20434
7	7	2011-01-07	1	0	1	0	5	2	0.19651

We get the vizualization editor here we can select any type of vizualization like bargraph, histogram, pichart, scatter, etc here in the below picture it shows histogram type of visualization
Here we have to select the values that should be given to X-axis and Y-axis

The screenshot shows the 'Visualization Editor' window in Databricks. The 'Visualization type' is set to 'Histogram'. The 'X column' is set to 'season' and the 'Number of bins' is set to 10. The histogram shows the distribution of the 'season' variable, with the Y-axis labeled 'COUNT' and the X-axis labeled 'season'. The histogram has four bars, each representing a different season value.

season	COUNT
1	150
2	150
3	150
4	150

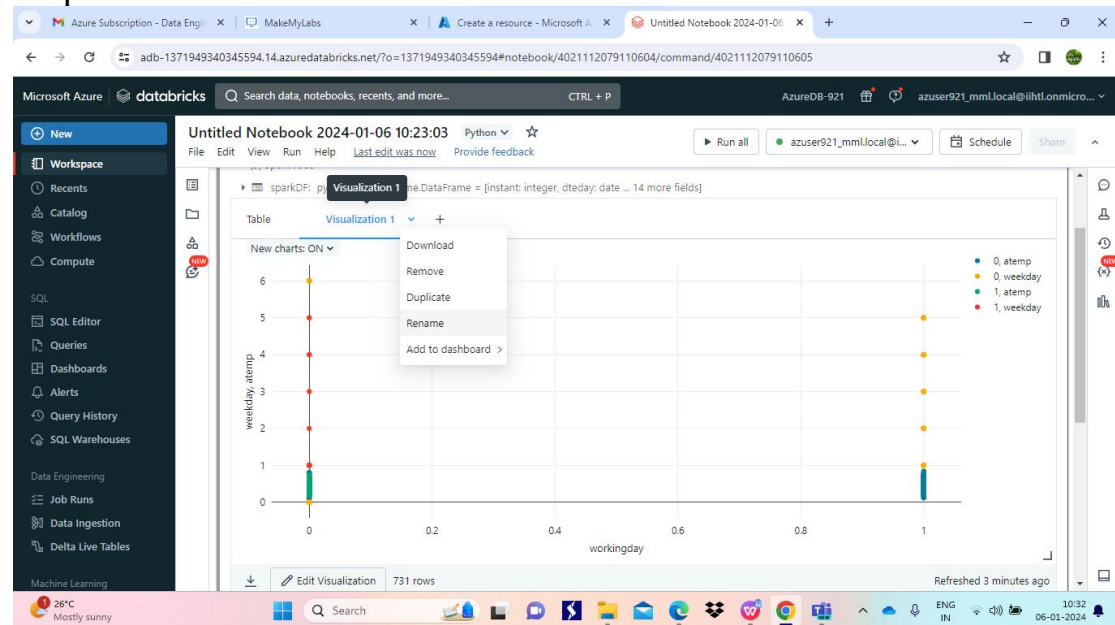
Below picture shows the scatter type visualization



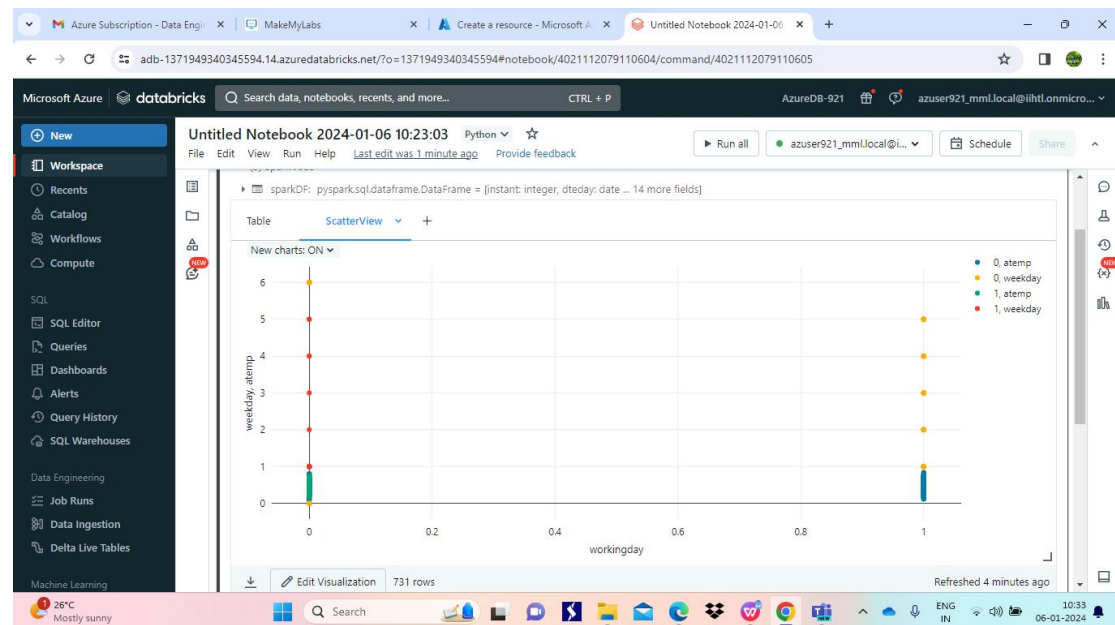
After editing the visualization we have to save the visualization picture to save click on the save button at the right bottom



Then we get the visualization on the notebook to rename it click on the dropdown beside the the visualization and select rename as shown below



We can edit the name of visualization here we renamed visualization1 to Scatterview



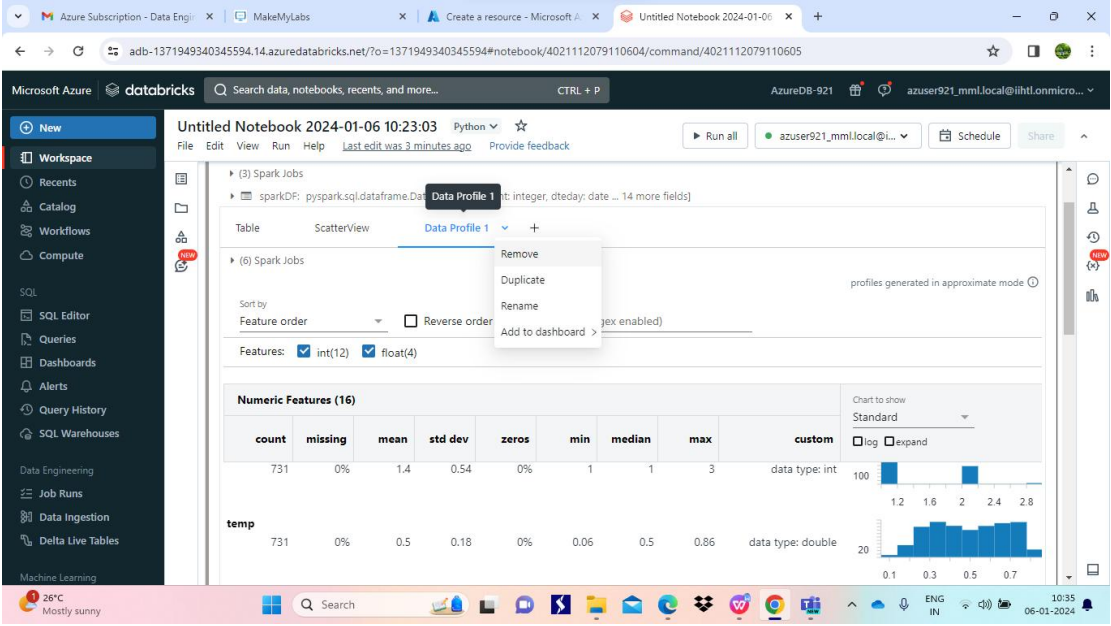
To create a duplicate dataprofile or visualization click on the dropdown beside the name and select duplicate

The screenshot shows the Databricks workspace interface. On the left is a sidebar with navigation options like Workspace, Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, and SQL Warehouses. The main area displays 'Untitled Notebook 2024-01-06 10:23:03'. A dropdown menu for 'Data Profile 1' is open, showing options: Remove, Duplicate, Rename, and Add to dashboard. The 'Duplicate' option is highlighted. Below the menu, the 'Data Profile 1' view is visible, showing a table of features and their statistics. The table has columns: count, missing, mean, std dev, zeros, min, median, max, and custom. The features listed are 'int(12)' and 'float(4)'. The 'Numeric Features (16)' section shows a table with columns: count, missing, mean, std dev, zeros, min, median, max, and custom. The features listed are 'instant' and 'dteday'. The 'instant' feature has a count of 731, missing 0%, mean 366, std dev 211.17, zeros 0%, min 1, median 366, max 731, and data type int. The 'dteday' feature has a count of 731, missing 0%, mean 1.338, std dev 18.2M, zeros 0%, min 1.298, median 1.338, max 1.368, and data type date. The 'Chart to show' dropdown is set to 'Standard'.

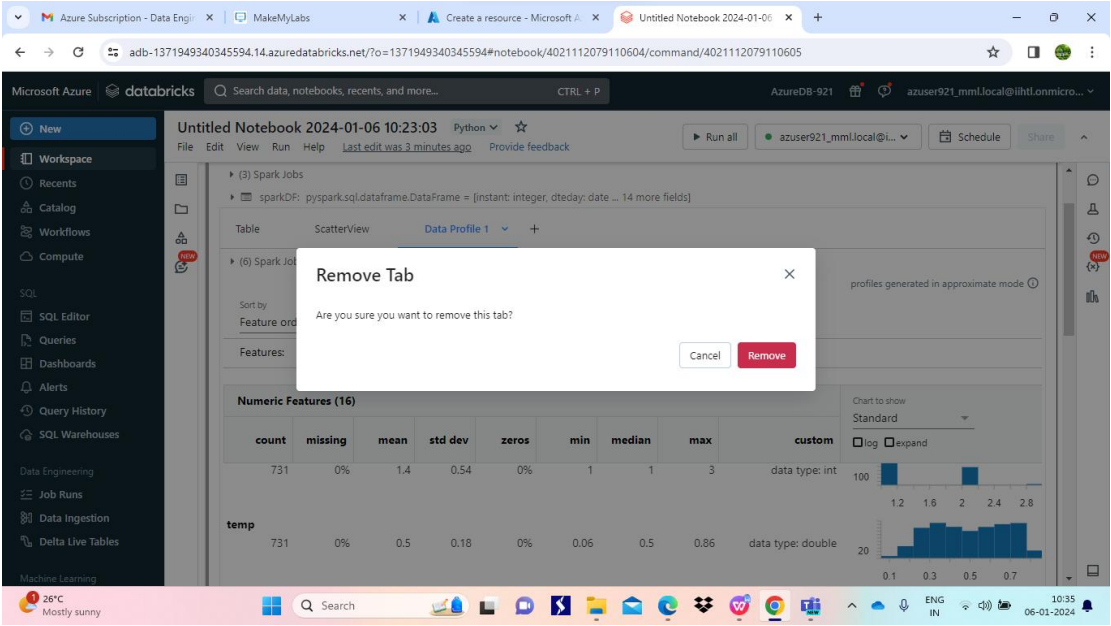
Here is the duplicate view of data profile

The screenshot shows the Databricks workspace interface. On the left is a sidebar with navigation options like Workspace, Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, and SQL Warehouses. The main area displays 'Untitled Notebook 2024-01-06 10:23:03'. The 'Data Profile 1 - Duplicate' view is visible, showing a table of features and their statistics. The table has columns: count, missing, mean, std dev, zeros, min, median, max, and custom. The features listed are 'instant' and 'dteday'. The 'instant' feature has a count of 731, missing 0%, mean 366, std dev 211.17, zeros 0%, min 1, median 366, max 731, and data type int. The 'dteday' feature has a count of 731, missing 0%, mean 1.338, std dev 18.2M, zeros 0%, min 1.298, median 1.338, max 1.368, and data type date. The 'Chart to show' dropdown is set to 'Standard'.

To remove the data profile or visualization click on the dropdown and select remove



Now we get popup box for confirmation of removing of dataprofile or visualization and click on remove



After removing data profile and visualization

The screenshot shows the Microsoft Azure Databricks interface. The top navigation bar includes 'Microsoft Azure', 'databricks', and a search bar. The left sidebar contains navigation options like 'New', 'Workspace', 'Recents', 'Catalog', 'Workflows', 'Compute', 'SQL', 'SQL Editor', 'Queries', 'Dashboards', 'Alerts', 'Query History', 'SQL Warehouses', 'Data Engineering', 'Job Runs', 'Data Ingestion', and 'Delta Live Tables'. The main area displays an 'Untitled Notebook' with a Python environment. The notebook content shows a Spark job that has executed, resulting in a table with 731 rows and 14 columns. The table columns are: instant, dteday, season, yr, mnth, holiday, weekday, workingday, weathersit, and temp. The table data is as follows:

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp
1	1	2011-01-01	1	0	1	0	6	0	2	0.34416
2	2	2011-01-02	1	0	1	0	0	0	2	0.36347
3	3	2011-01-03	1	0	1	0	1	1	1	0.19636
4	4	2011-01-04	1	0	1	0	2	1	1	0.2
5	5	2011-01-05	1	0	1	0	3	1	1	0.22695
6	6	2011-01-06	1	0	1	0	4	1	1	0.20434
7	7	2011-01-07	1	0	1	0	5	1	2	0.19654

The notebook also shows the command 'sparkDF: pyspark.sql.dataframe.DataFrame = [instant: integer, dteday: date ... 14 more fields]' and the execution time '7.17 seconds runtime'. The bottom status bar shows the temperature as 26°C and the date as 06-01-2024.

After completion of all the tasks that we have to perform in notebooks to run our queries terminate the cluster

The screenshot shows the Microsoft Azure Databricks interface with the 'Compute' tab selected. The top navigation bar includes 'Microsoft Azure', 'databricks', and a search bar. The left sidebar contains navigation options like 'New', 'Workspace', 'Recents', 'Catalog', 'Workflows', 'Compute', 'SQL', 'SQL Editor', 'Queries', 'Dashboards', 'Alerts', 'Query History', 'SQL Warehouses', 'Data Engineering', 'Job Runs', 'Data Ingestion', and 'Delta Live Tables'. The main area displays the 'Compute' tab with a list of clusters. The table columns are: State, Name, Policy, Runtime, Active m..., Active co..., Active D..., Source, Creator, and Notebooks. The table data is as follows:

State	Name	Policy	Runtime	Active m...	Active co...	Active D...	Source	Creator	Notebooks
Stopped	azuser921_mml.local@ihl.onmicro...	Personal Comp	13.3	-	-	-	UI	azuser921_mml...	-

The bottom status bar shows the temperature as 25°C and the date as 06-01-2024.

2. Explain the copy activity in Azure Data Factory

First to perform the copy activity we have to create Azure data factory and we have to launch ADF studio

- To do this we have to login to the Azure portal and click on create resource and select analytics after that we see data factory and click on it
- Now fill the fields select resource group and name of the resource and click on review+create
- Then we have review the details and then click on create
- After deployment of ADF resource go to the resource and launch the ADF studio
- Now we can see the ADF in new window here we have to give the properties to the ADF
- We go to source here we have create a connection from where we have to transfer or copy a file which can be a storage account or delta lake house gen2 or any storage where our file is placed select the file to be copied select duplicate in the type and test the connection and then click on create
- Click on next we move to destination similar to the source create a new connection give the destination path and test the connection and then click on create
- Click on next we get the summary of the ADF which shows the details of source and destination of file to be copied
- Now click on next to copy the file here we can see the deployment process first validating the run time environment then checking the registration of source and destination links
- Now it creates a pipeline to copy the file and provisioning of data
- After successful completion of copying of file it shows successful in green
- Now we can check the destination path to see whether the file is copied or not

To perform ADF copy activity we need storage accounts follow the below steps to create storage accounts

- ✓ In azure portal go to search bar and type storage accounts and then select it
- ✓ Now click on + create to create the storage account
- ✓ Fill the fields of resource group location storage account name and then click on review +create

- ✓ After reviewing click on create now the deployment is in progress after completion of deployment click on go to resource here we can find some option to create container to upload files/folder
- ✓ Now we can create a container and upload the files

To create a data lake storage gen -2 follow below steps

- ✓ In azure portal go to search bar and search for storage accounts classic and select it
- ✓ Now click on create to create a data lake storage gen-2
- ✓ Fill the fields of resource grp , account name etc
- ✓ Go to advanced and click on the check box hierarchical storage to allow delta lake storage gen 2
- ✓ Now click on review+create after reviewing click on create
- ✓ After completion of deployment of data lake storage gen2 click on go to resource and we can see option like upload file/folder, blob storage, rename ,etc