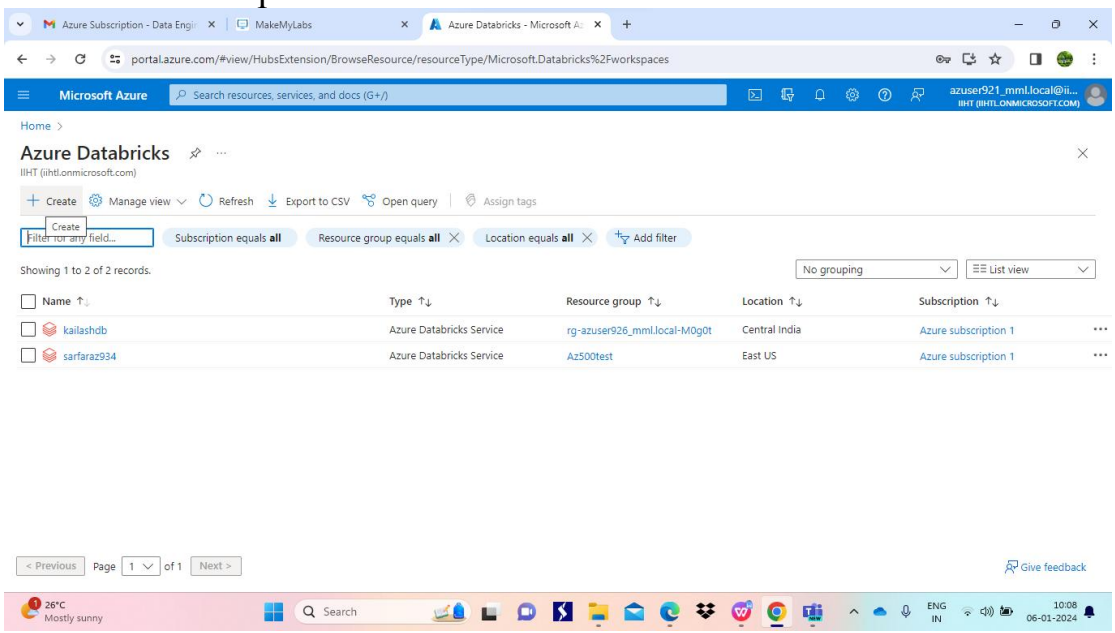
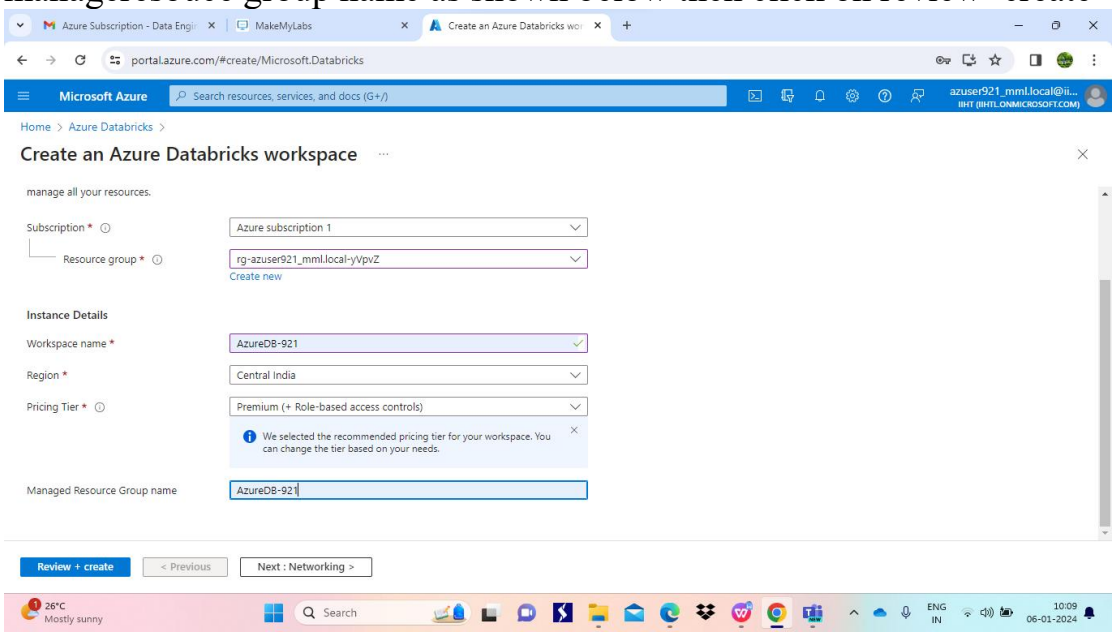
Mallu Sravana Sandhya
Azure Databricks Coding Assignment
06/01/24

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.

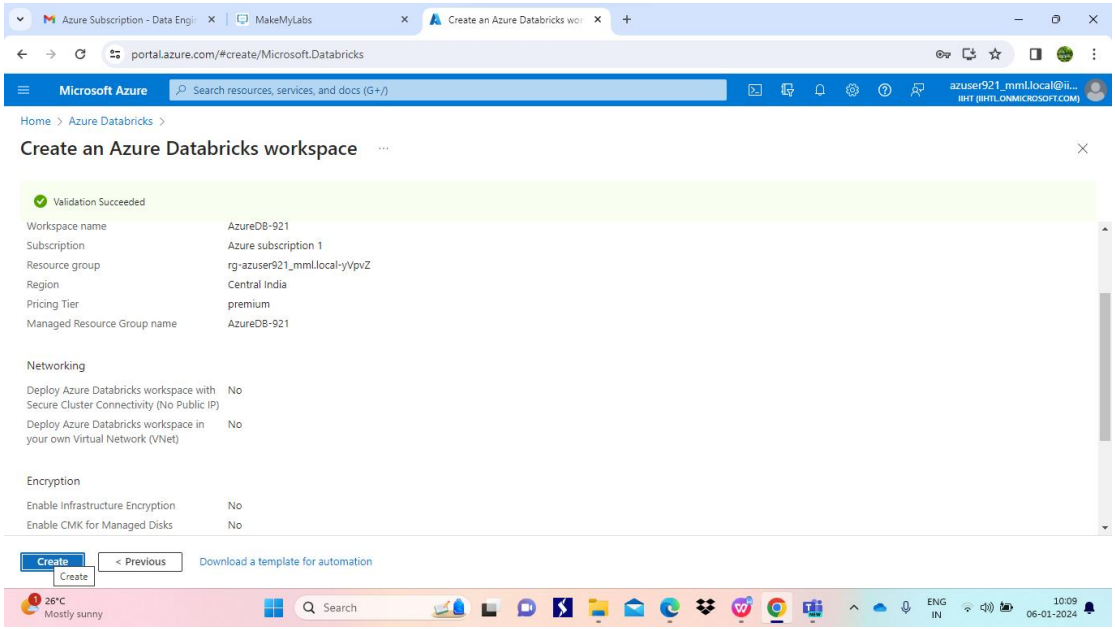
First we have to sign in to the azure portal and go Azure data bricks Service as shown in below and click on create to create an azure databricks workspace



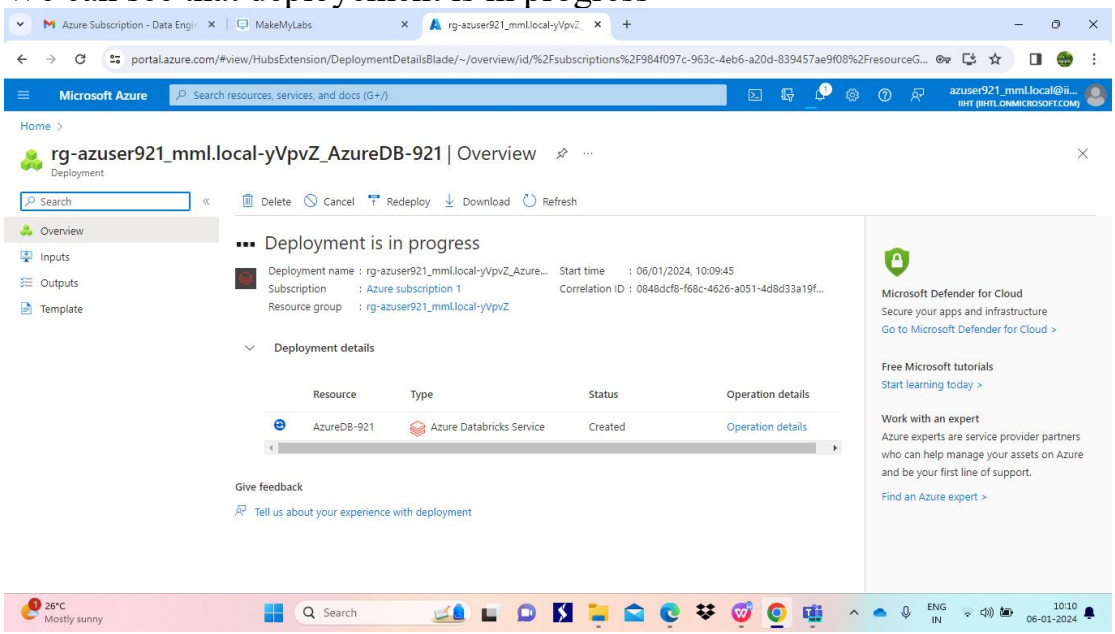
Give the values in the fields of resource group, workspace name and manageresouce group name as shown below then click on review+create



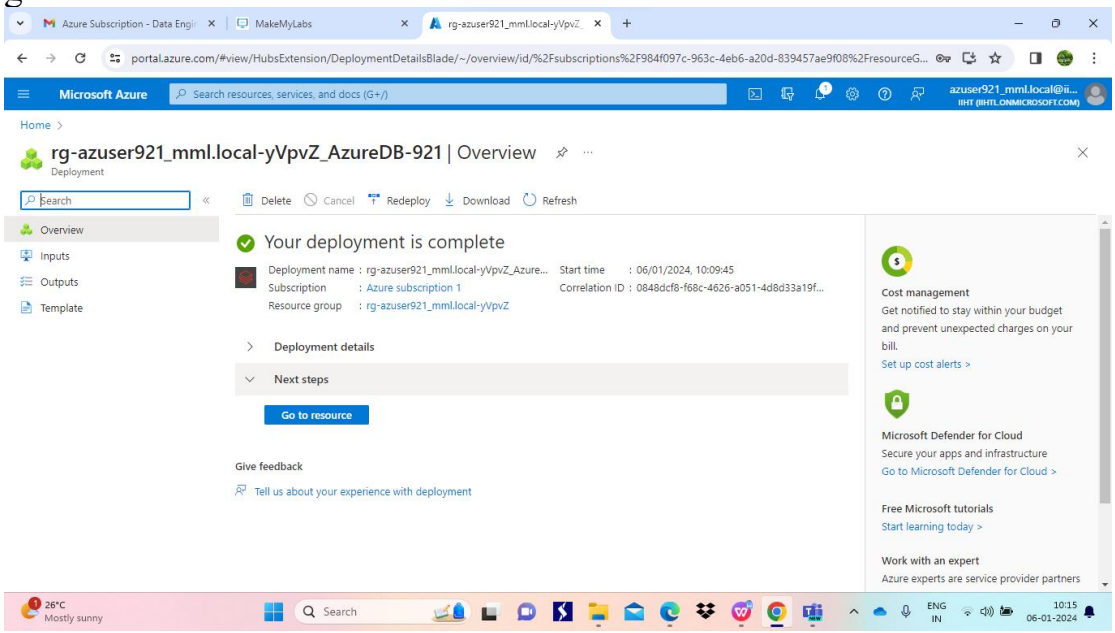
Then click on create button



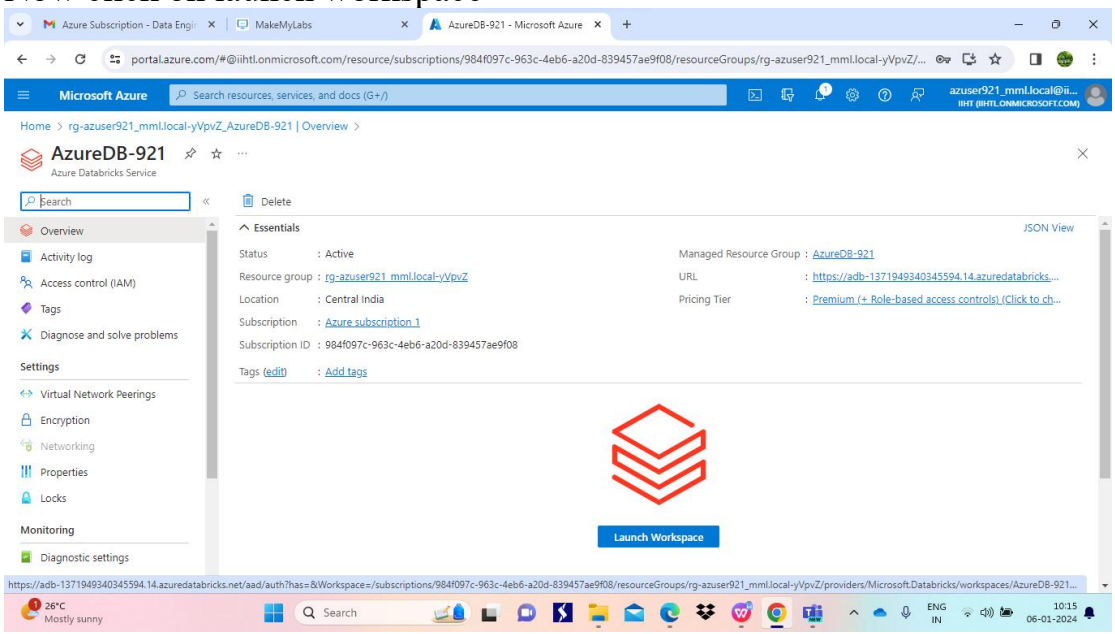
We can see that deployement is in progress



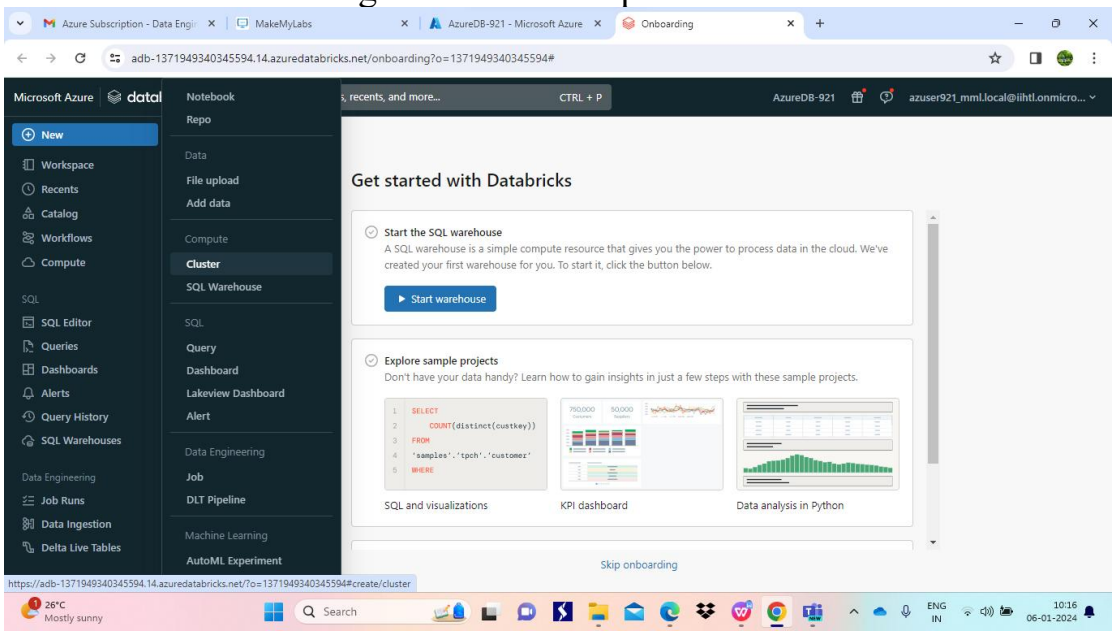
After completion of deployment of Azure databricks workspace, click on go to reource



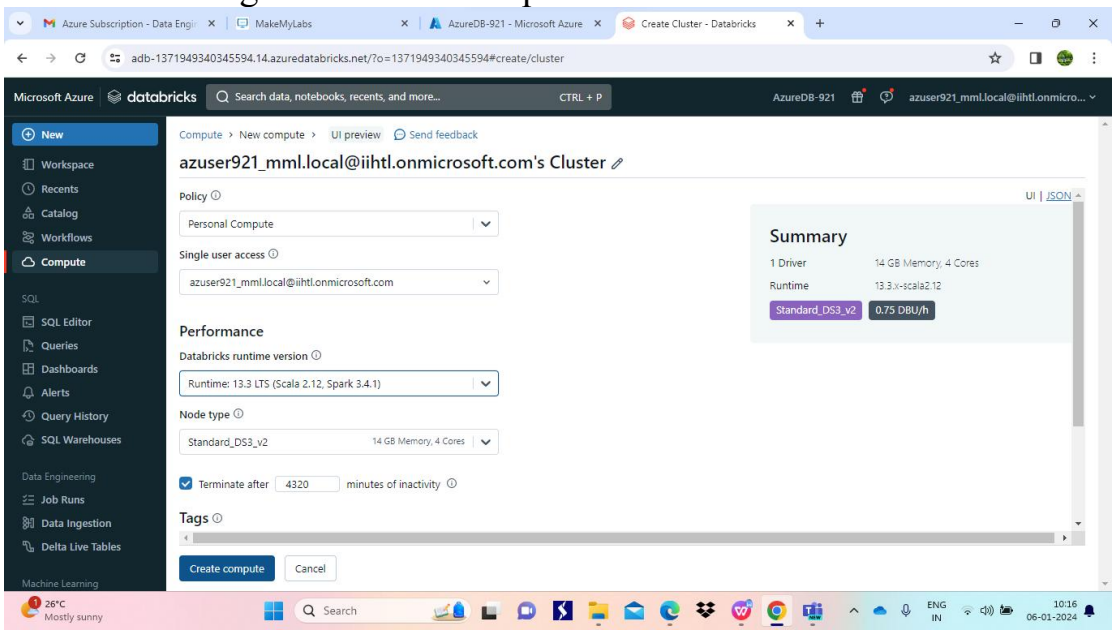
Now click on launch workspace



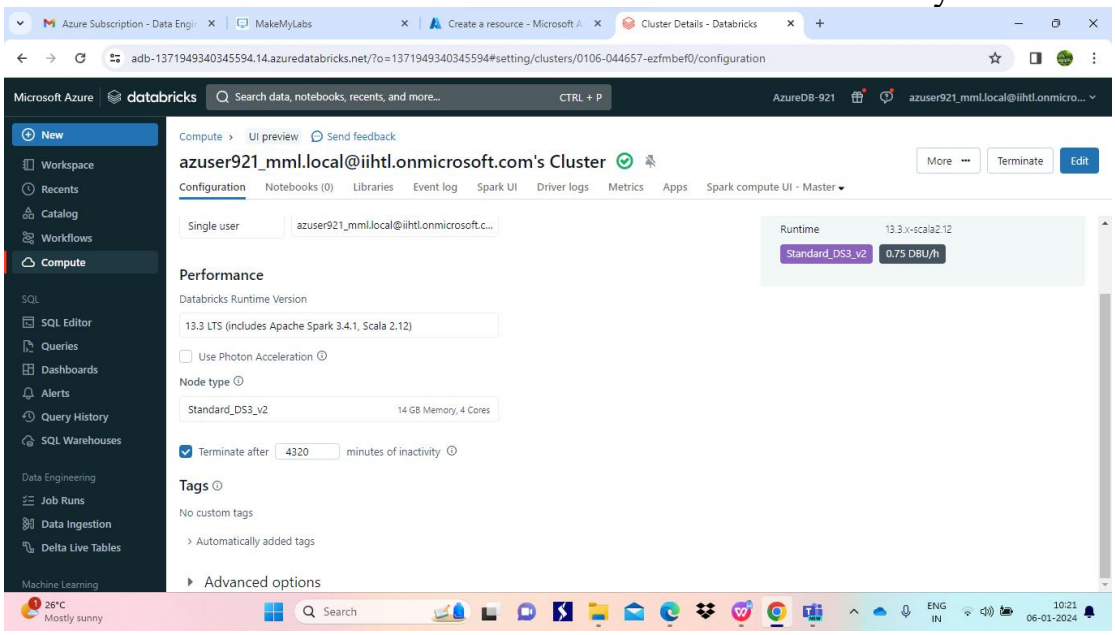
After launching of ADB workspace we can see as shown in the figure , now to create a cluster go to new on left panel and select cluster



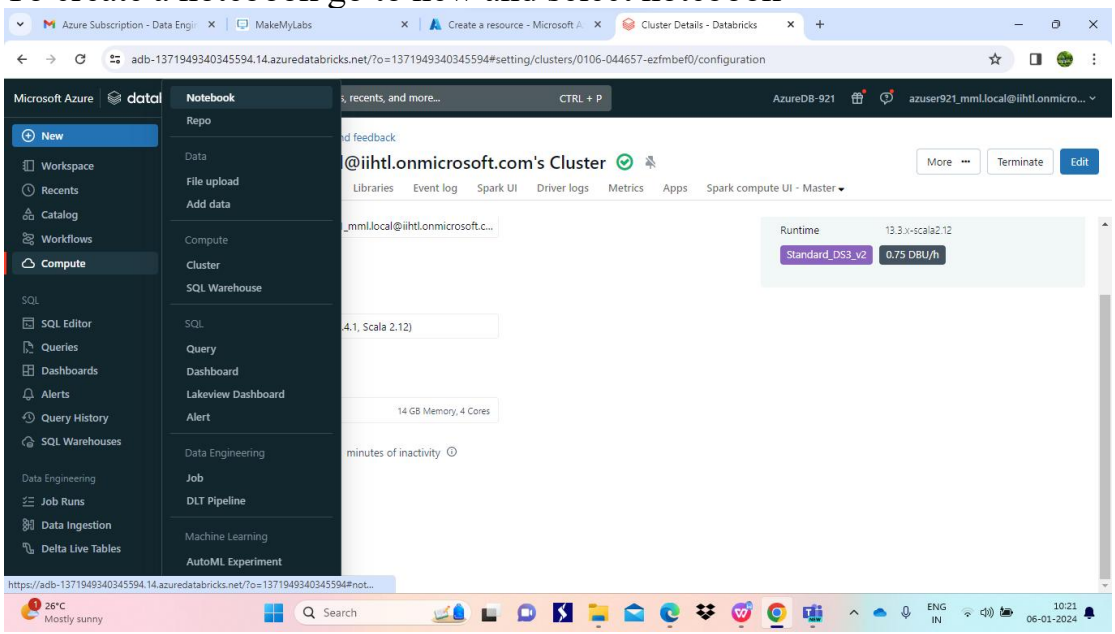
Now fill fields policy as personal compute and select databricks run version and give time to automativcally terminate cluster after limited time of not using cluster which helps to save our credit in azure account



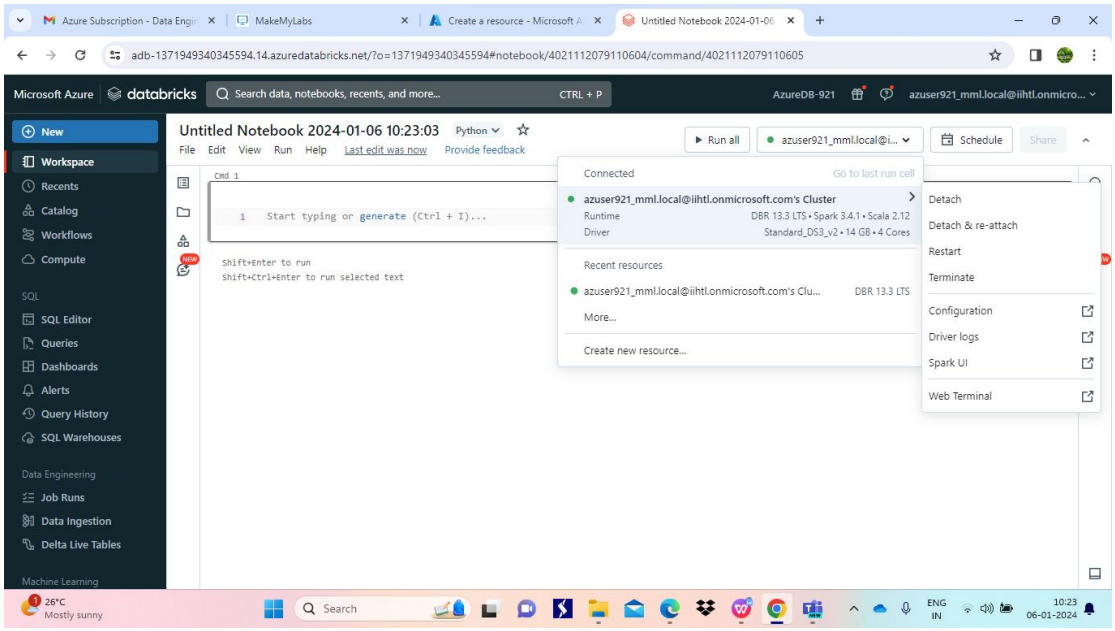
Then click on create and here the cluster is created successfully



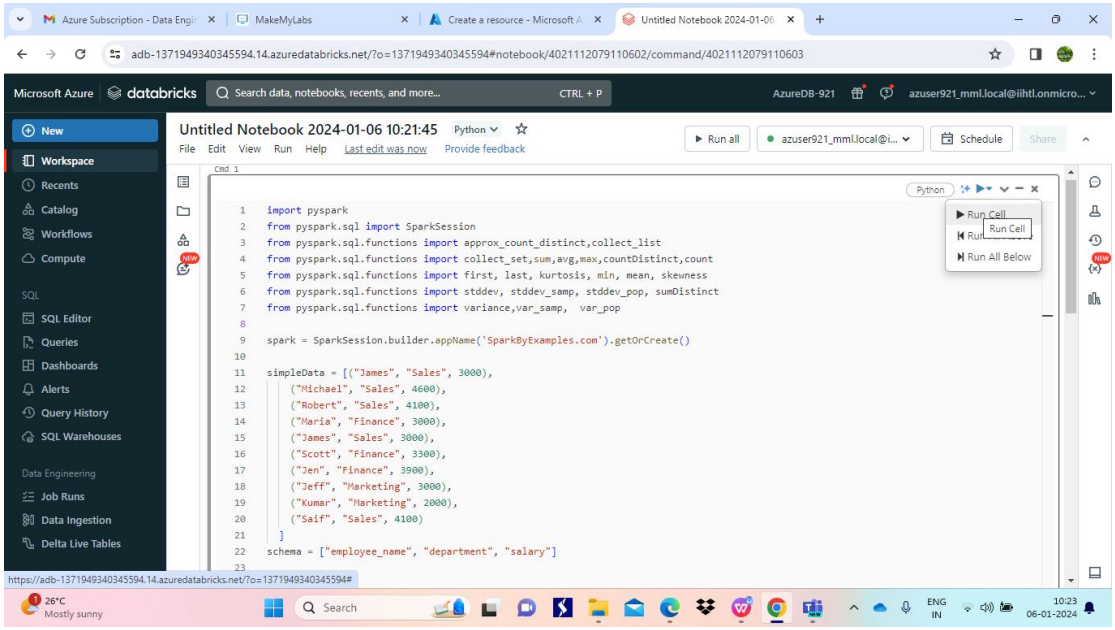
To attach the cluster to notebook we have to create a note book
To create a notebook go to new and select notebook

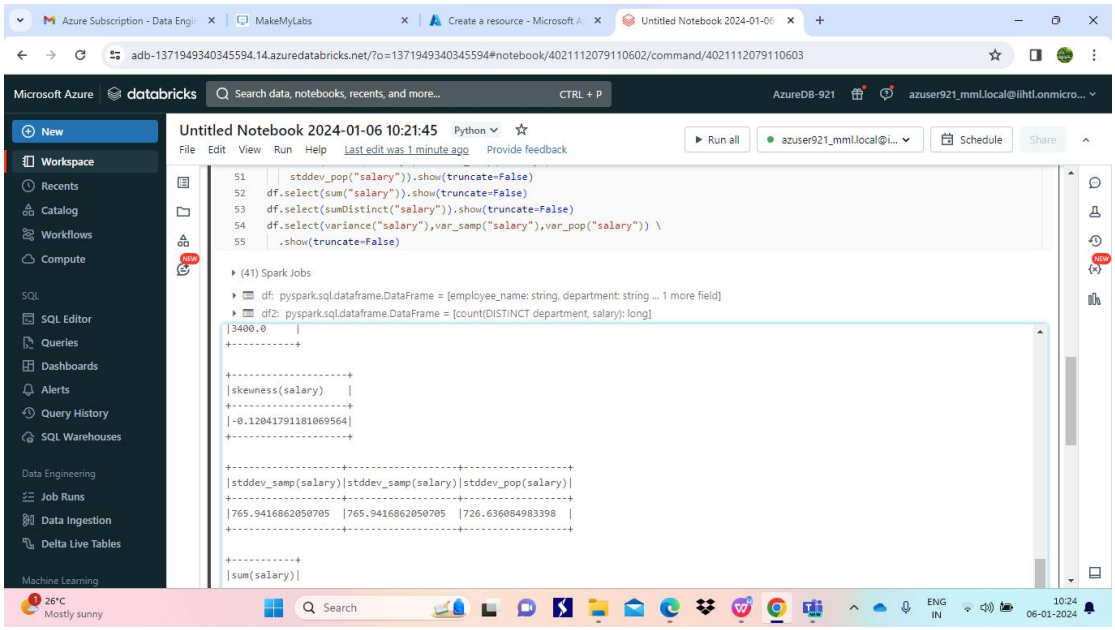


After creating a notebook we have to attach the cluster to the notebook in order to run the commands on notebook to attach cluster go to select cluster and click on dropdown and now select the cluster that we have created earlier

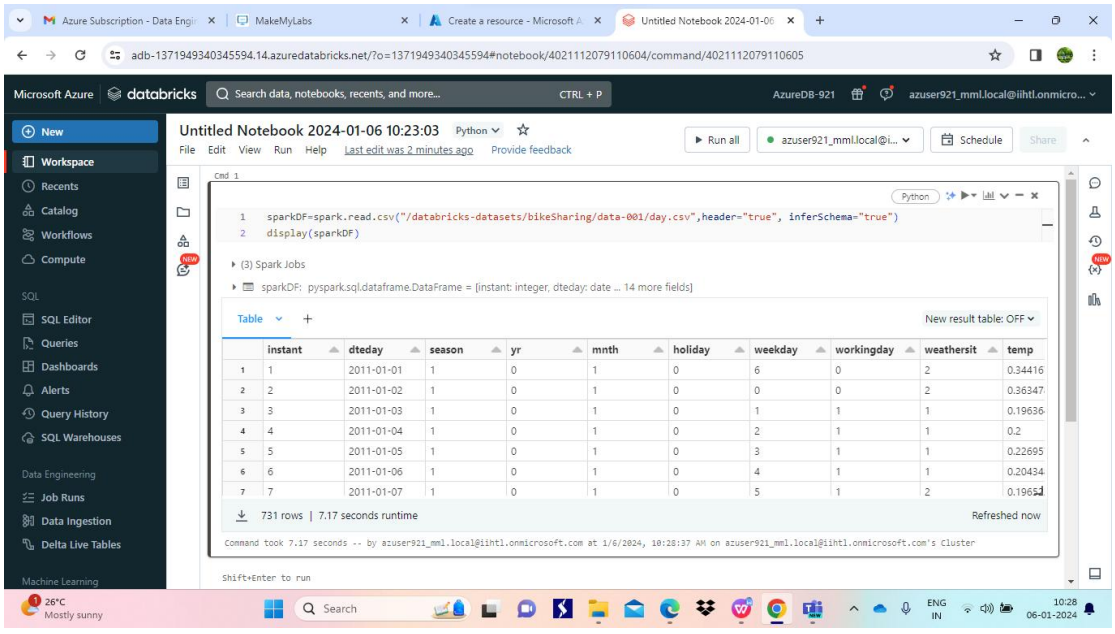


Now give some commands on notebook here I wrote a code to perform aggregate operations on table and click on run button to execute those commands

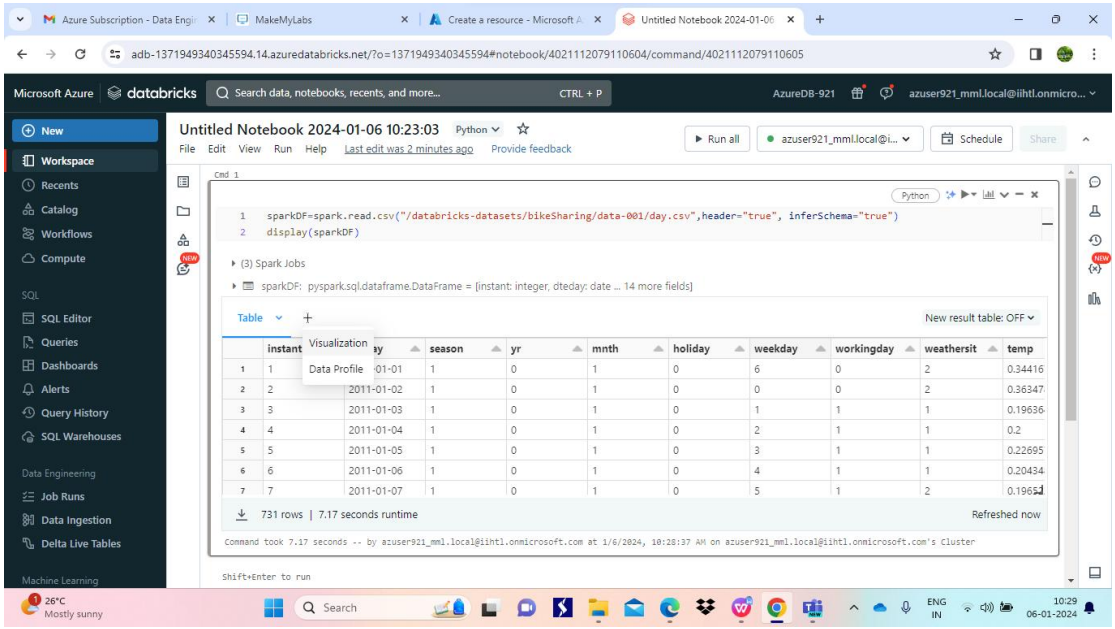




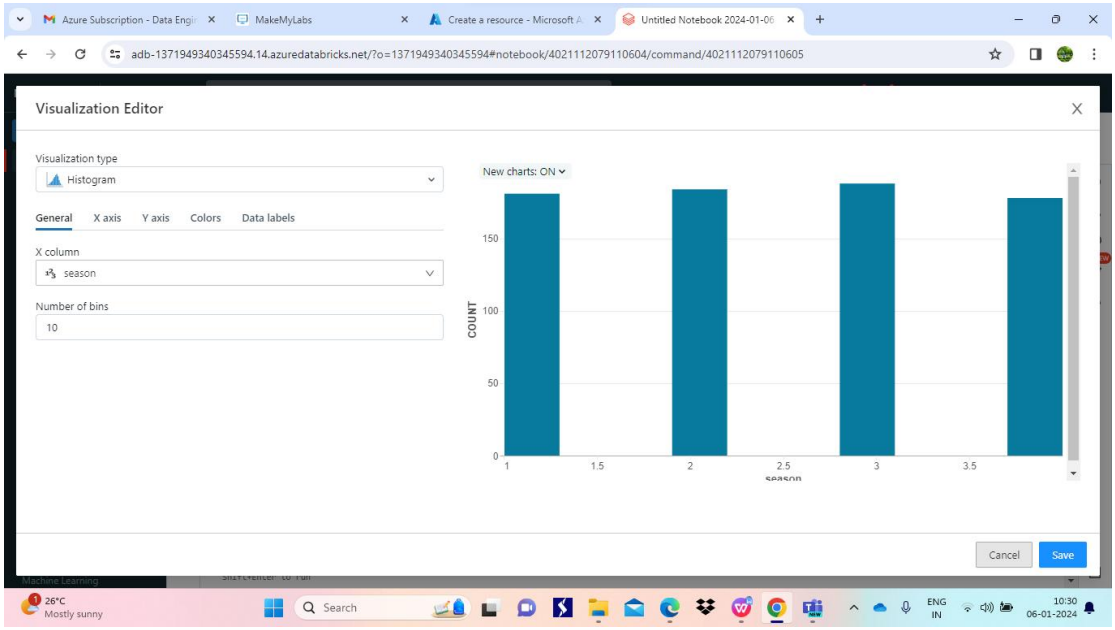
creating a DataFrame from a Databricks dataset create a new notebook and write the code as shown below and click on run we get the dataset in table



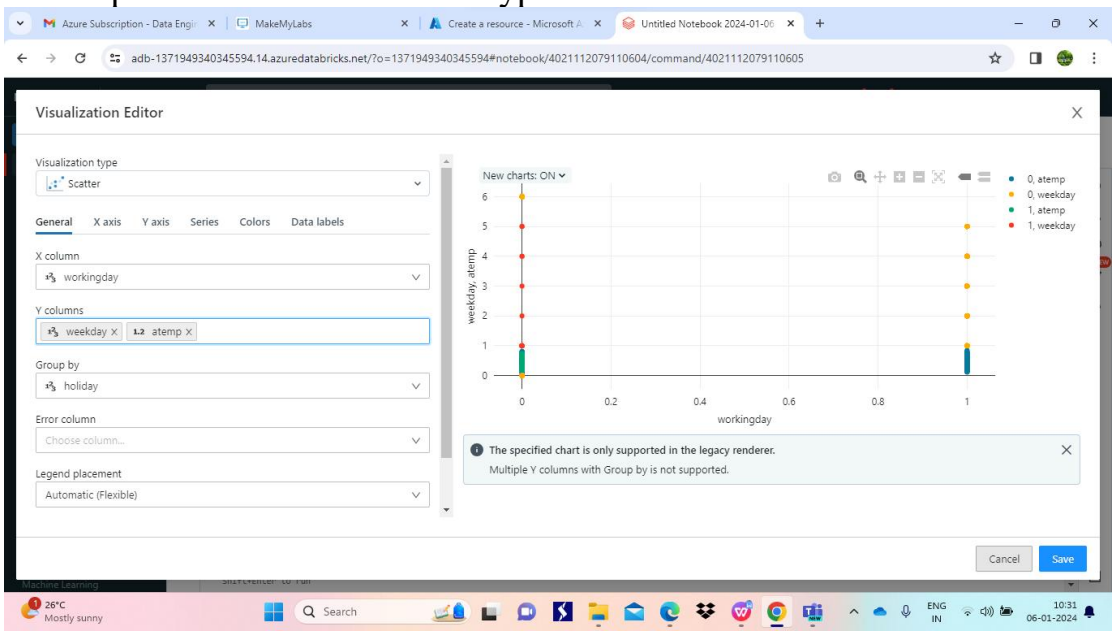
To create a visualization click on the + symbol beside the table and select visualization as shown below



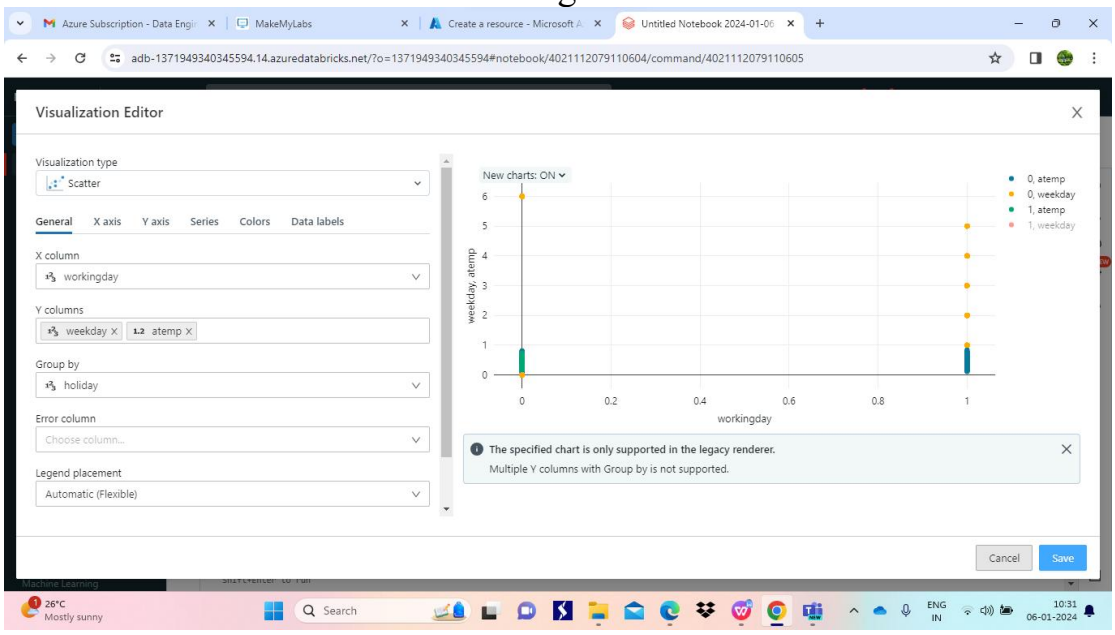
We get the vizualization editor here we can select any type of vizualization like bargraph,histogram, pichart, scatter, etc here in the below picture it shows histogram type of visualization
Here we have to select the values that should be given to X-axis and Y-axis



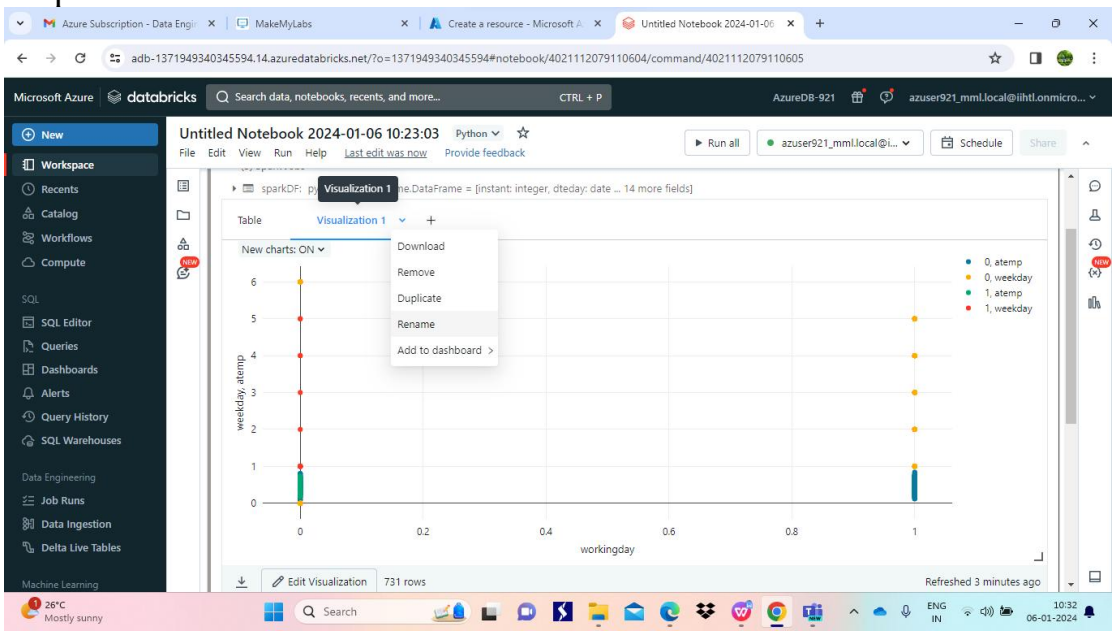
Below picture shows the scatter type visualization



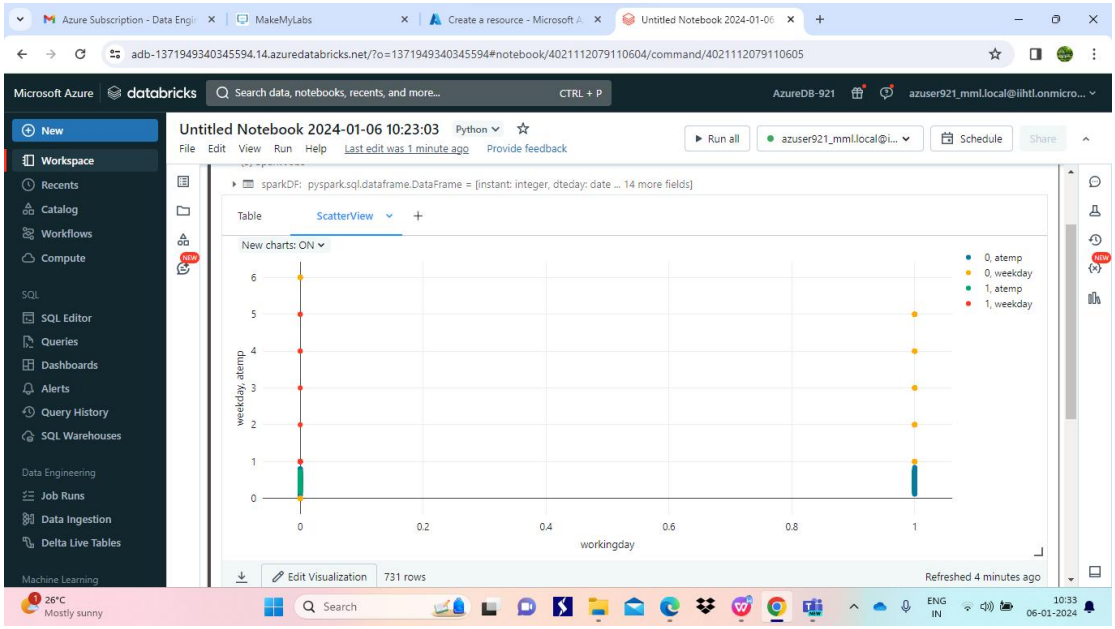
After editing the visualization we have to save the visualization picture to save click on the save button at the right bottom



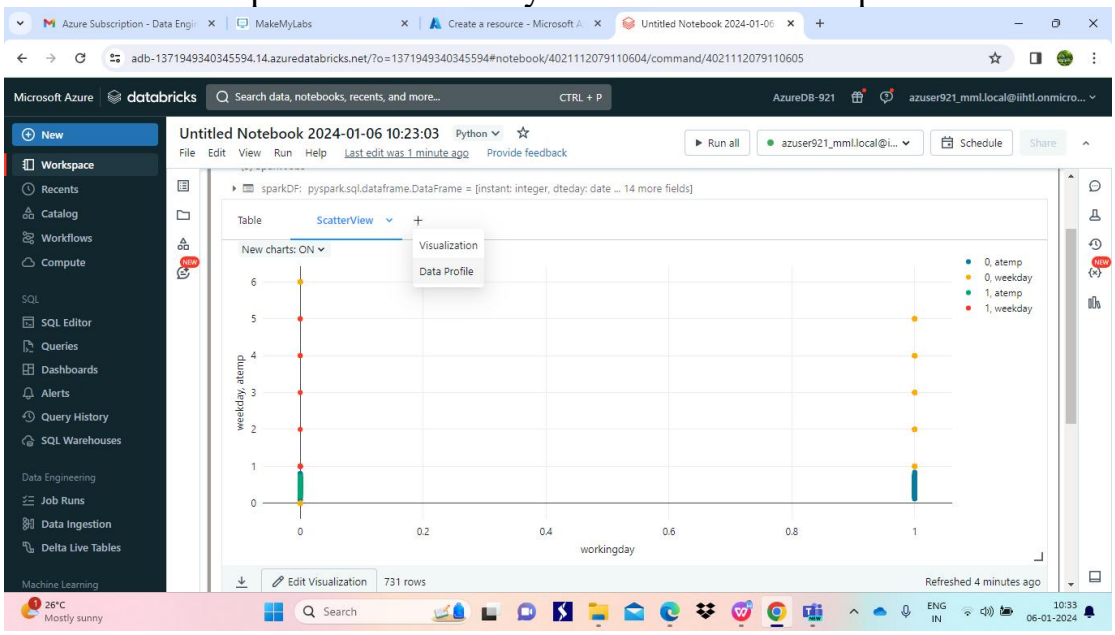
Then we get the visualization on the notebook to rename it click on the dropdown beside the the visualization and select rename as shown below



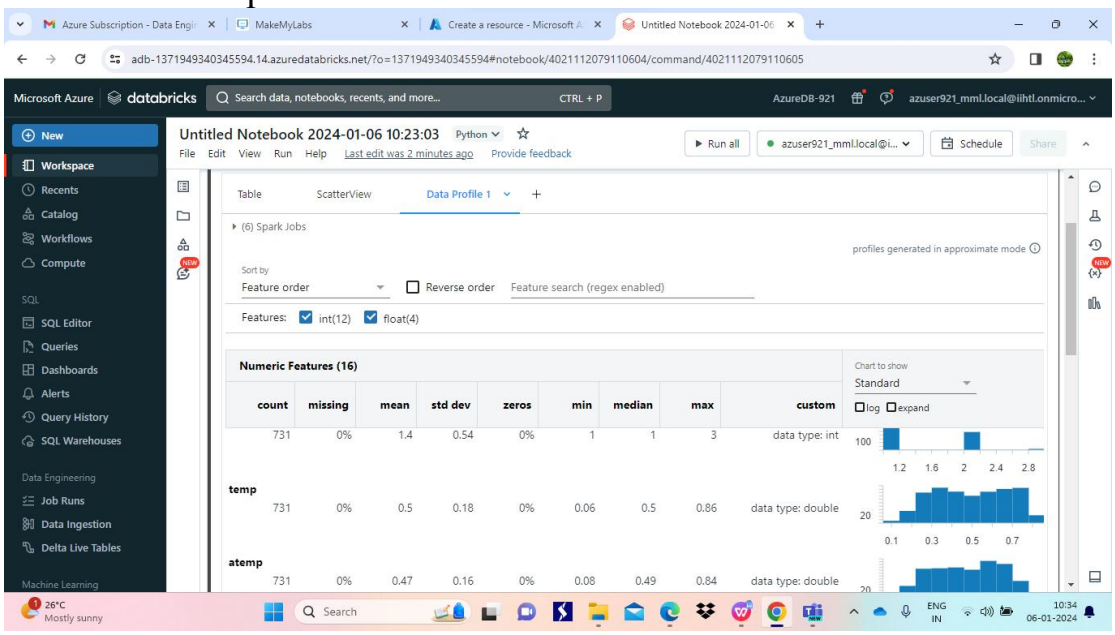
We can edit the name of visualization here we renamed visualization1 to Scatterview



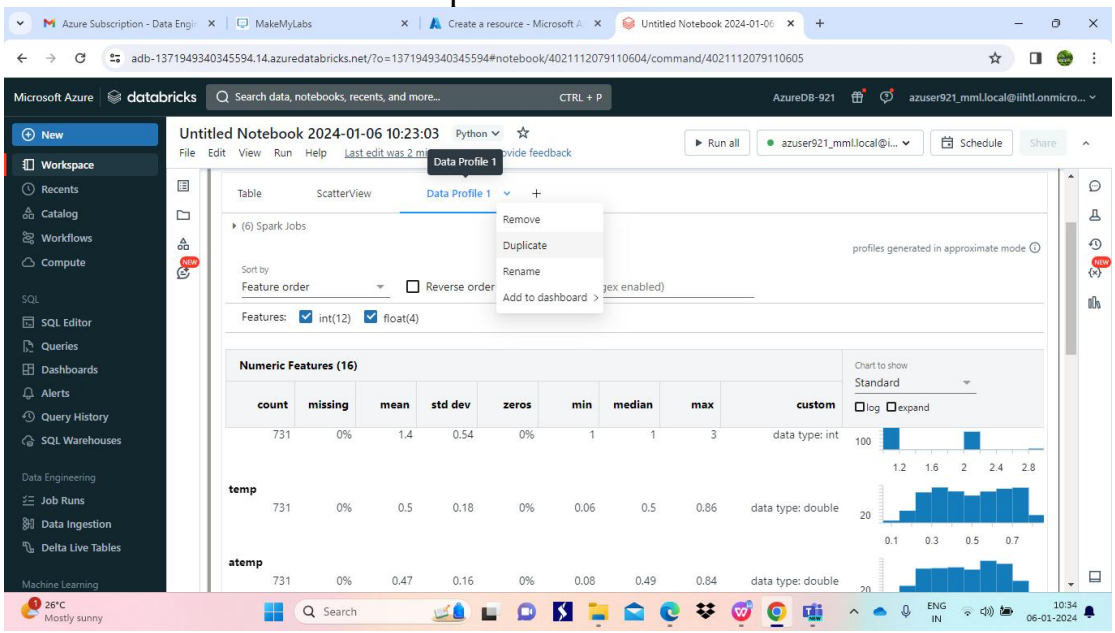
To create a data profile click on+ symbol and select data profile



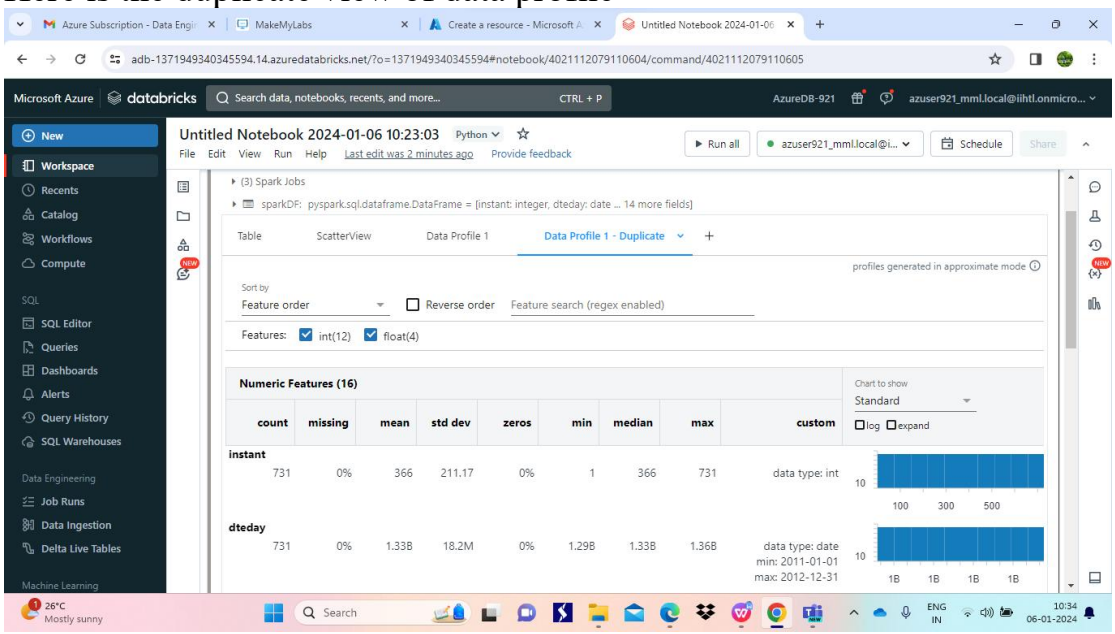
Here is the data profile of the taken dataset from databricks



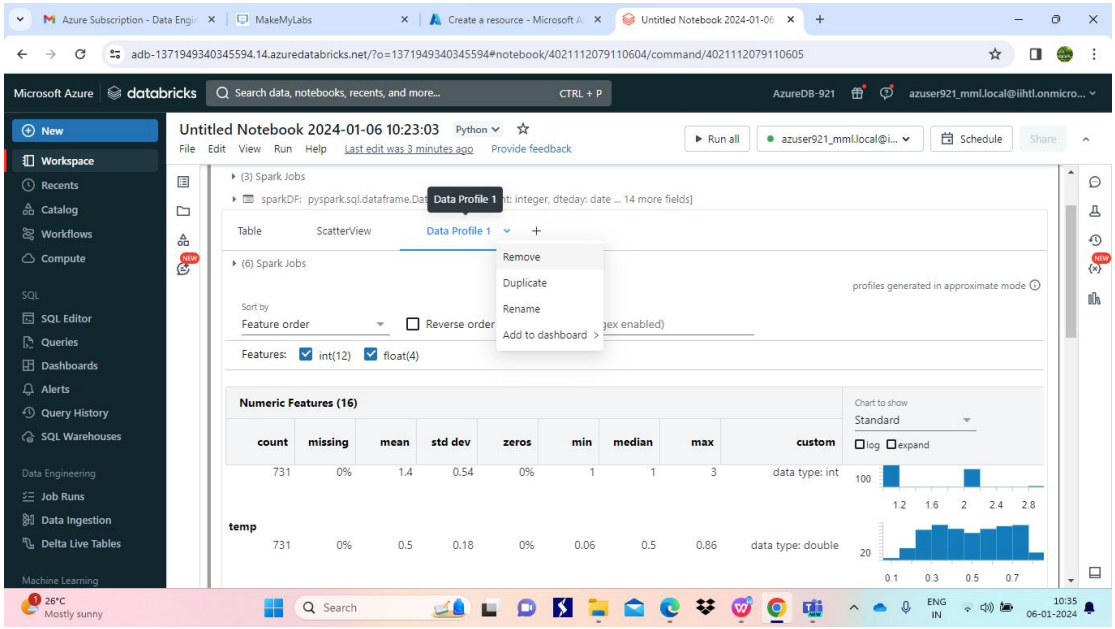
To create a duplicate dataprofile or visualization click on the dropdown beside the name and select duplicate



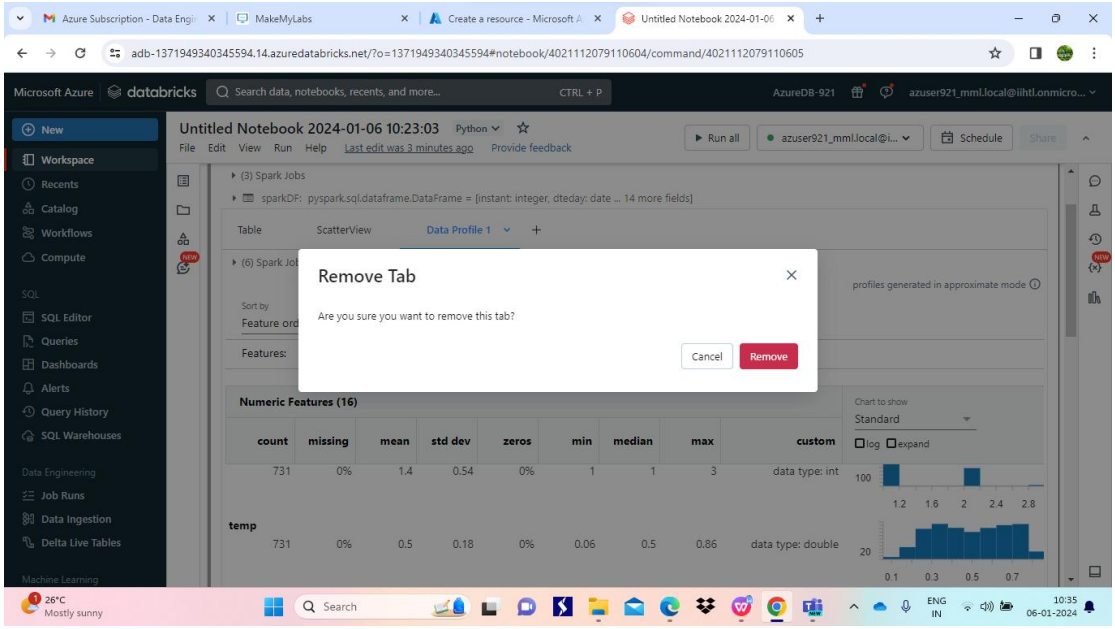
Here is the duplicate view of data profile



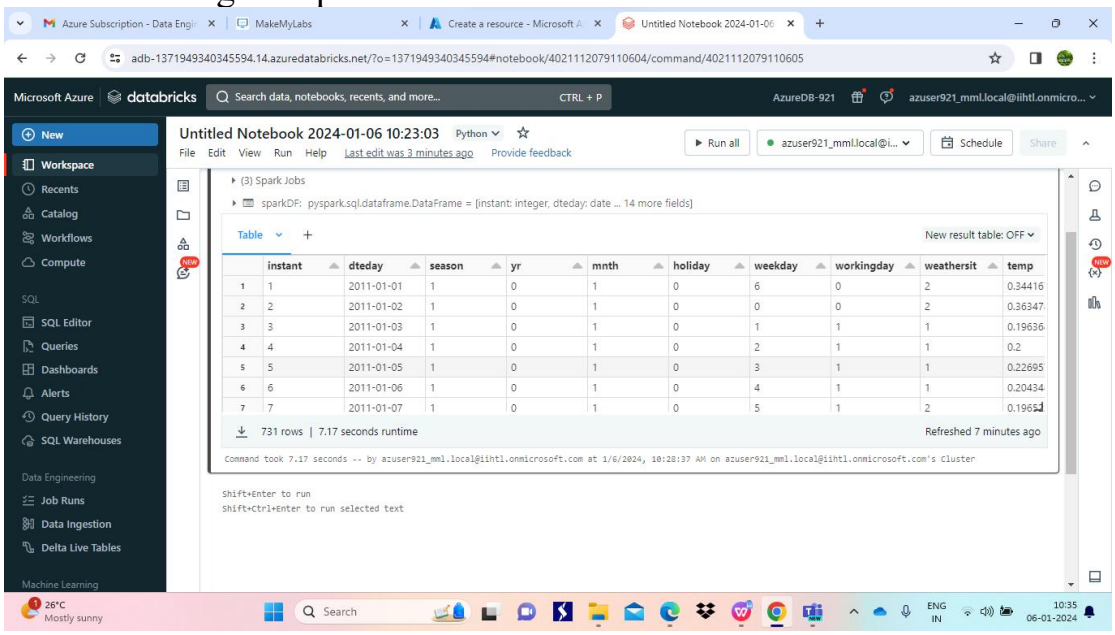
To remove the data profile or visualization click on the dropdown and select remove



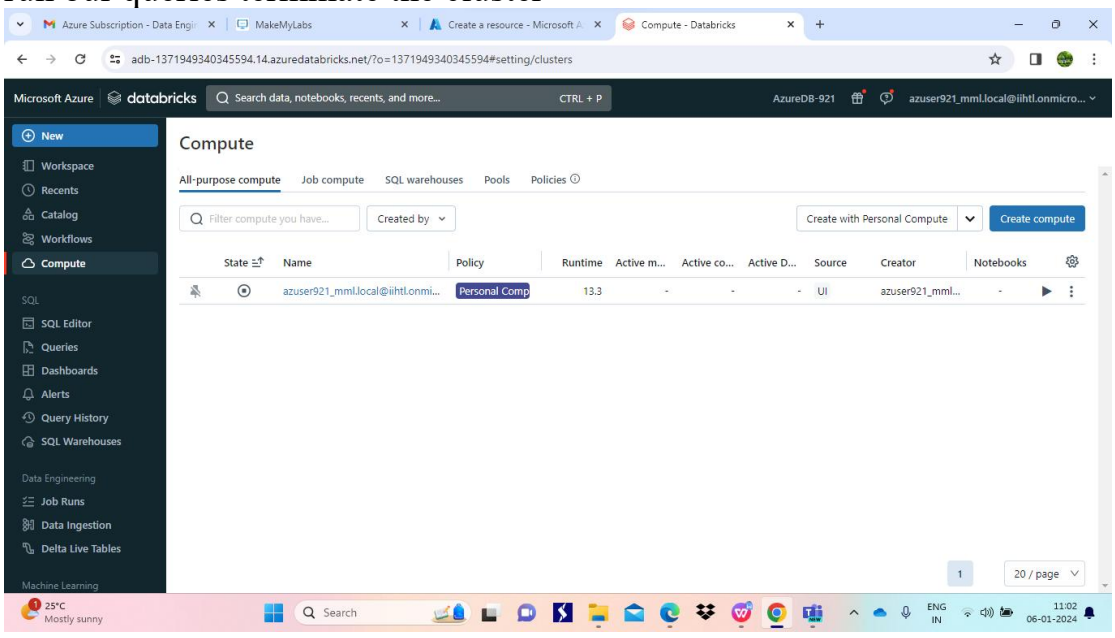
Now we get popup box for confirmation of removing of dataprofile or visualization and click on
remove



After remvoing data profile and visualization



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



2. **Expalin 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 deployement of ADF resouce 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 souce and destination of file to be copied
- Now click on next to copy the file here we can see the deployement process first validating the run time environment then checking the registration of souce 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 resoucre group location storage account name and then click on review +create

- ✓ After reviewing click on create now the deployement is in progress after completion of deployement 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 dala lake storage gen-2
- ✓ Fill the fileds of resource grp , account name etc
- ✓ Go to advanced and click on the check box hierarchial storage to allow delta lake storage gen 2
- ✓ Now click on review+create after reviewing click on create
- ✓ After completion of deployement of data lake storage gen2 click on go to resoucre and we can see option slike upload file/folder,blob storage,rename ,,etc