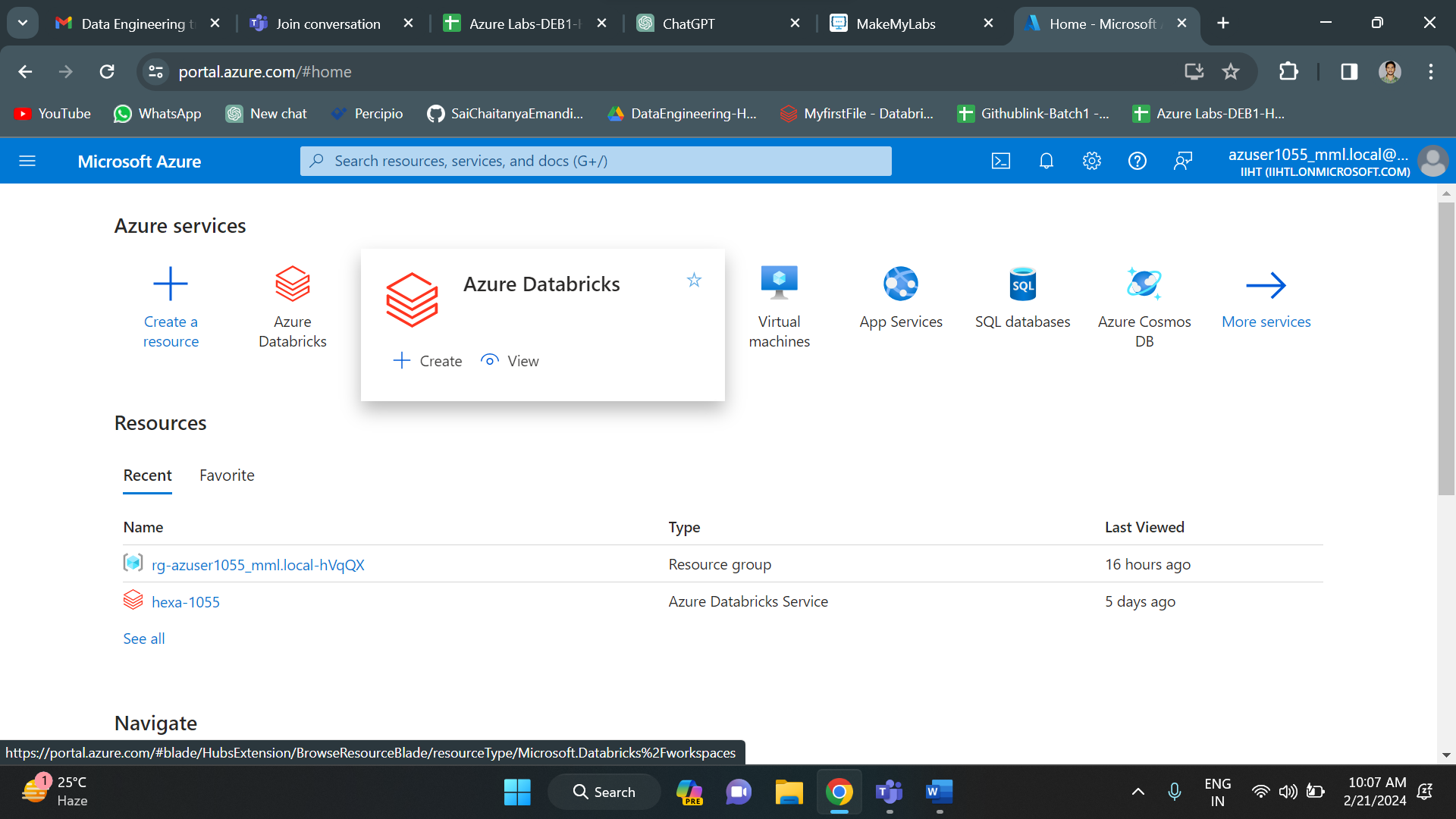
**Azure databricks coding challenge Q-1**

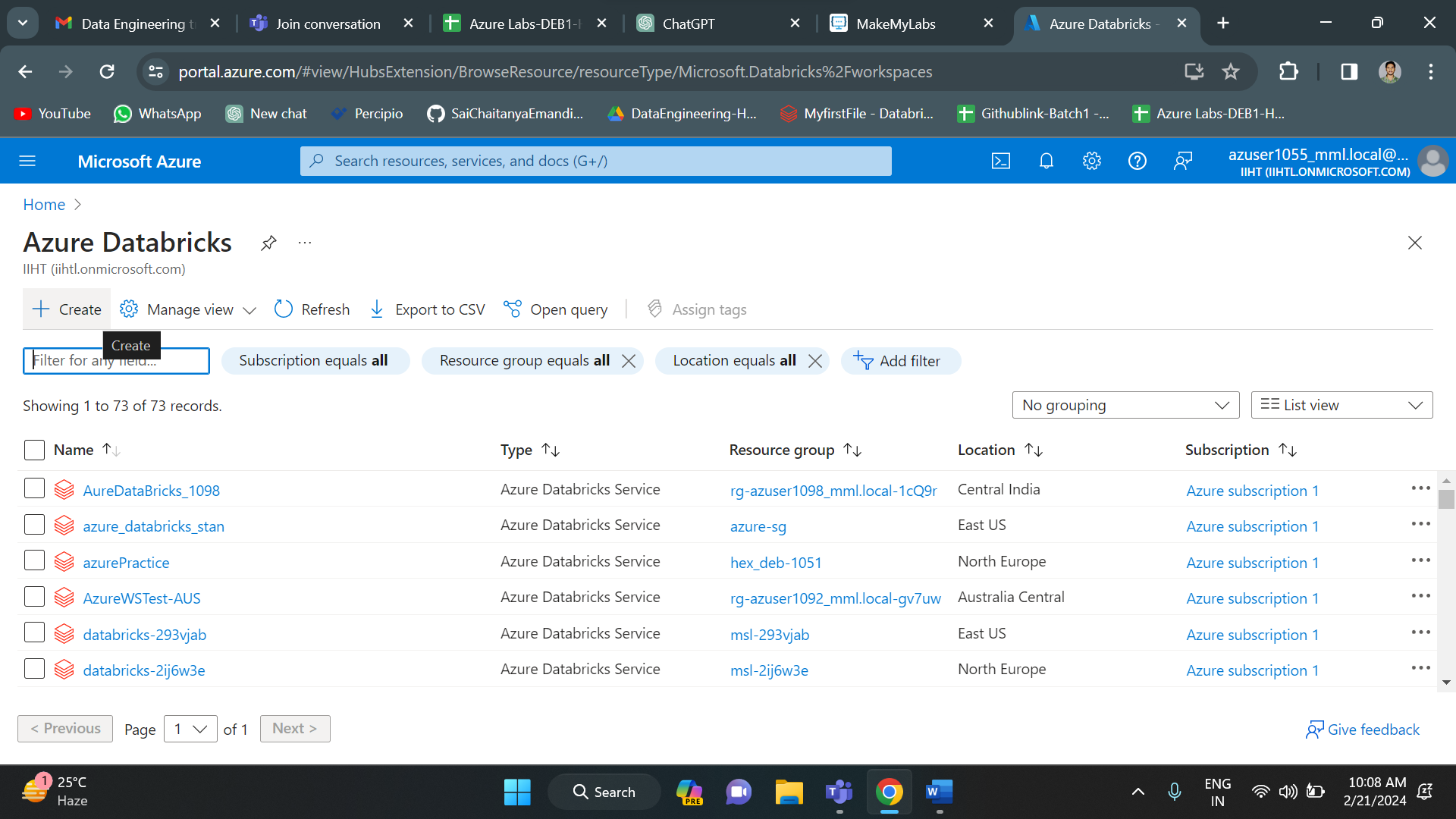
1. **EXPLORATORY DATA ANALYSIS AND VISUALIZATION IN DATABRICKS**

Azure databricks provides various tools to support the exploratory data analysis and visualization.

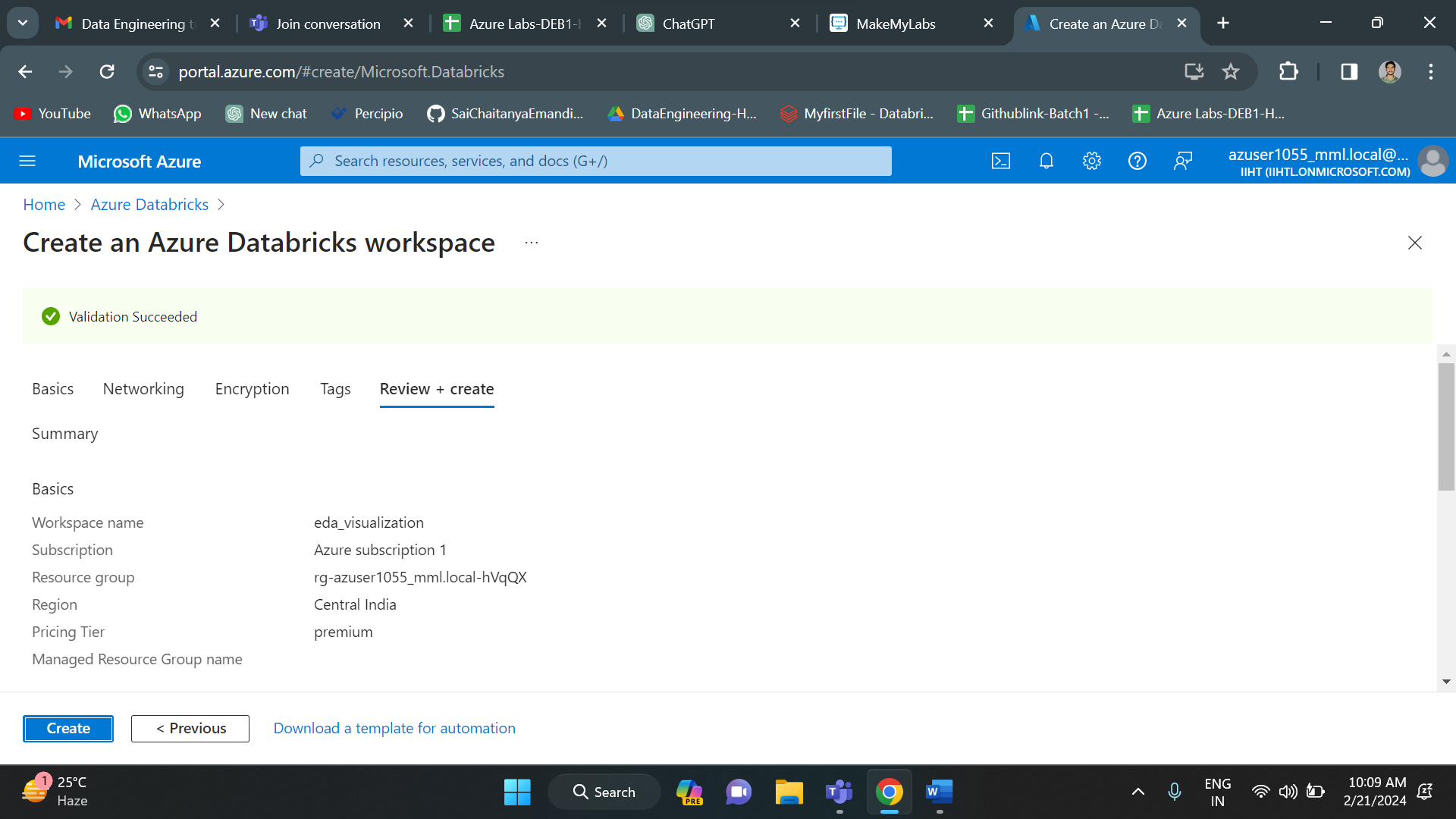
Firstly, login into azure portal and select Azure databricks.



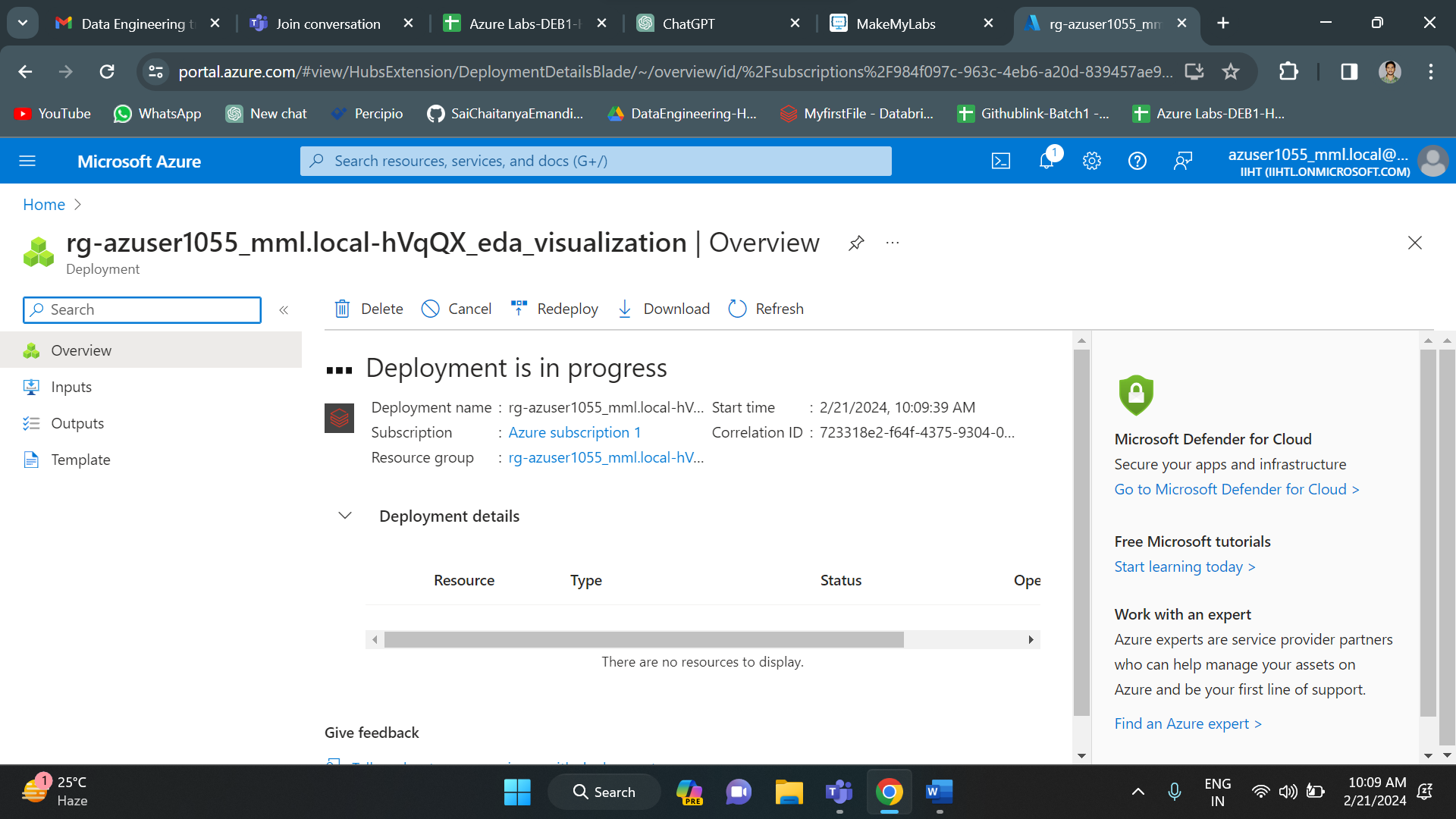
After selecting the azure databricks option, the below screen appears. In that you can see a create option to create azure databricks. Click on that.



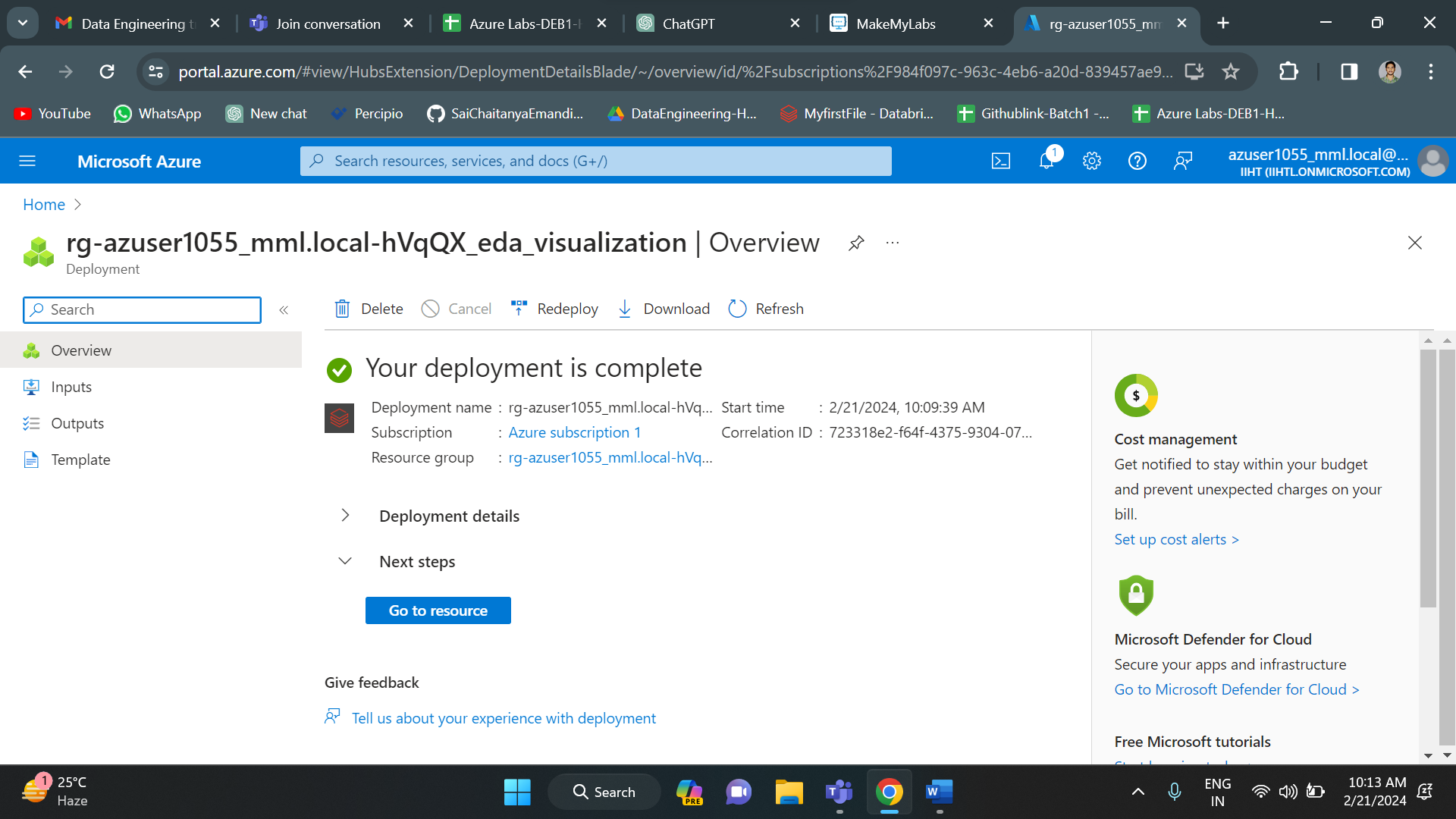
After clicking on create, it will ask you to fill some details like resource group, name etc. Then it will validate the data. After validation click on create option below to create data azure databricks.



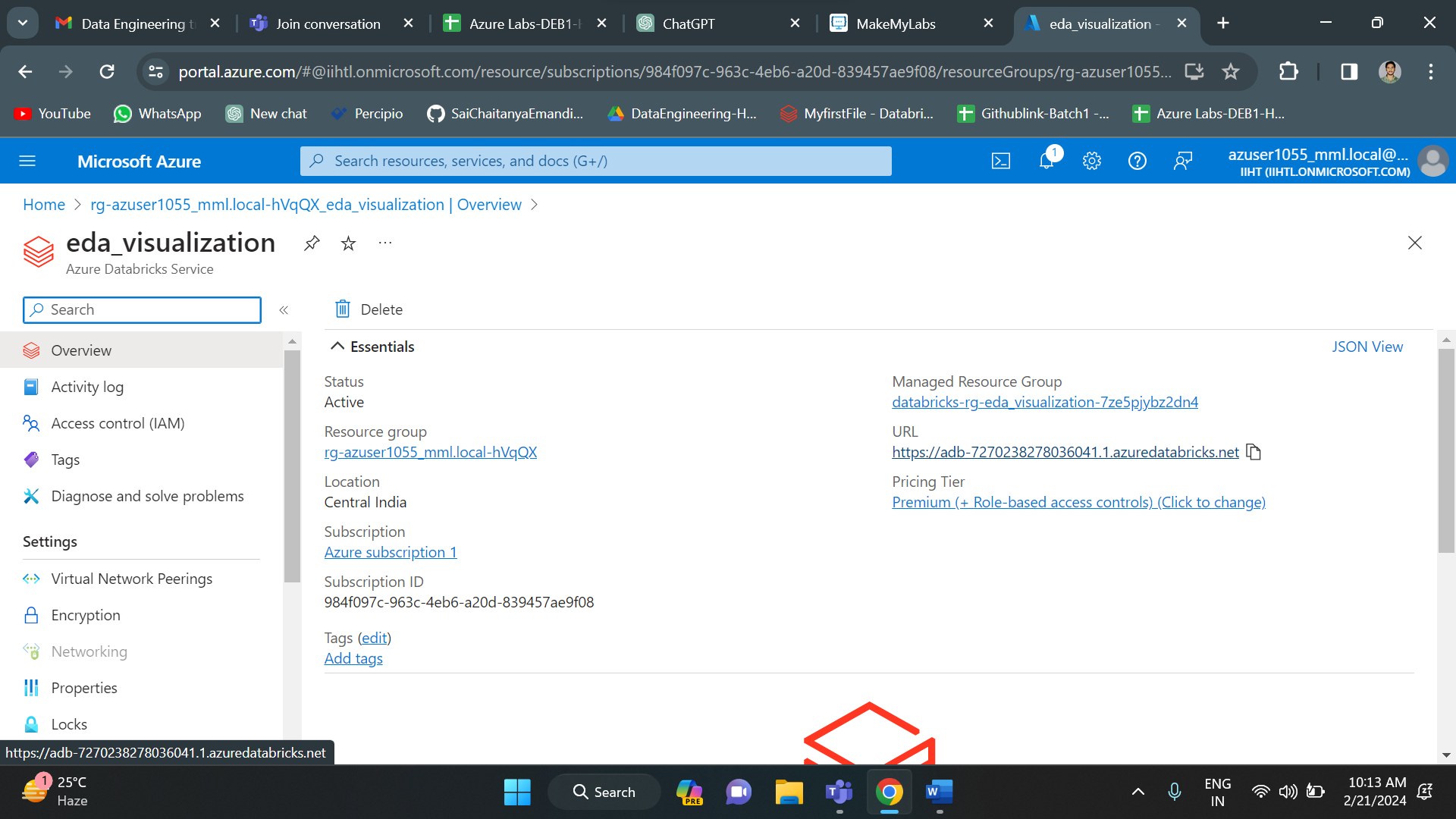
After clicking on create your deployment will be on progress.



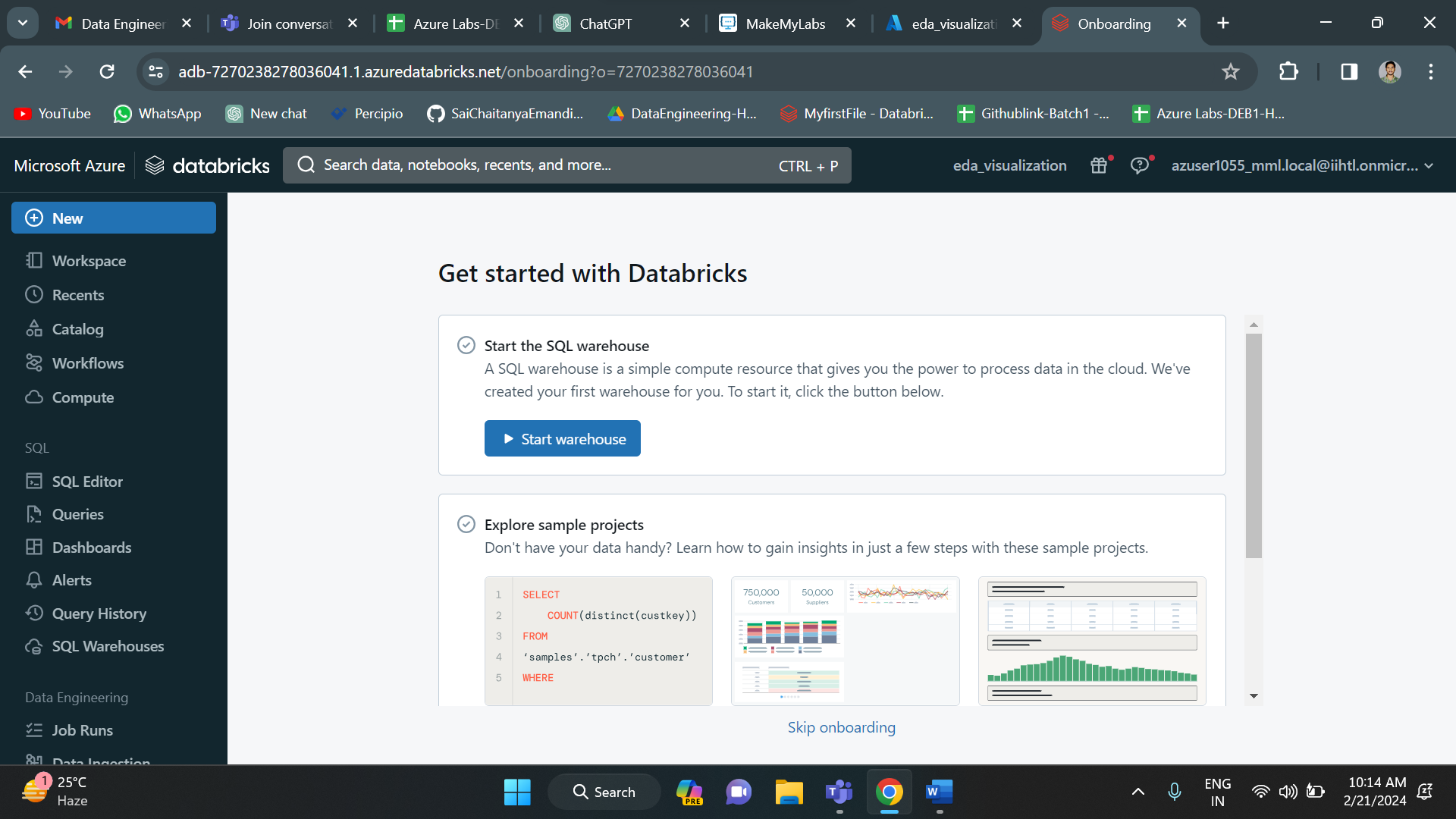
Once your deployment gets completed, click on go to resource to access the azure databricks.



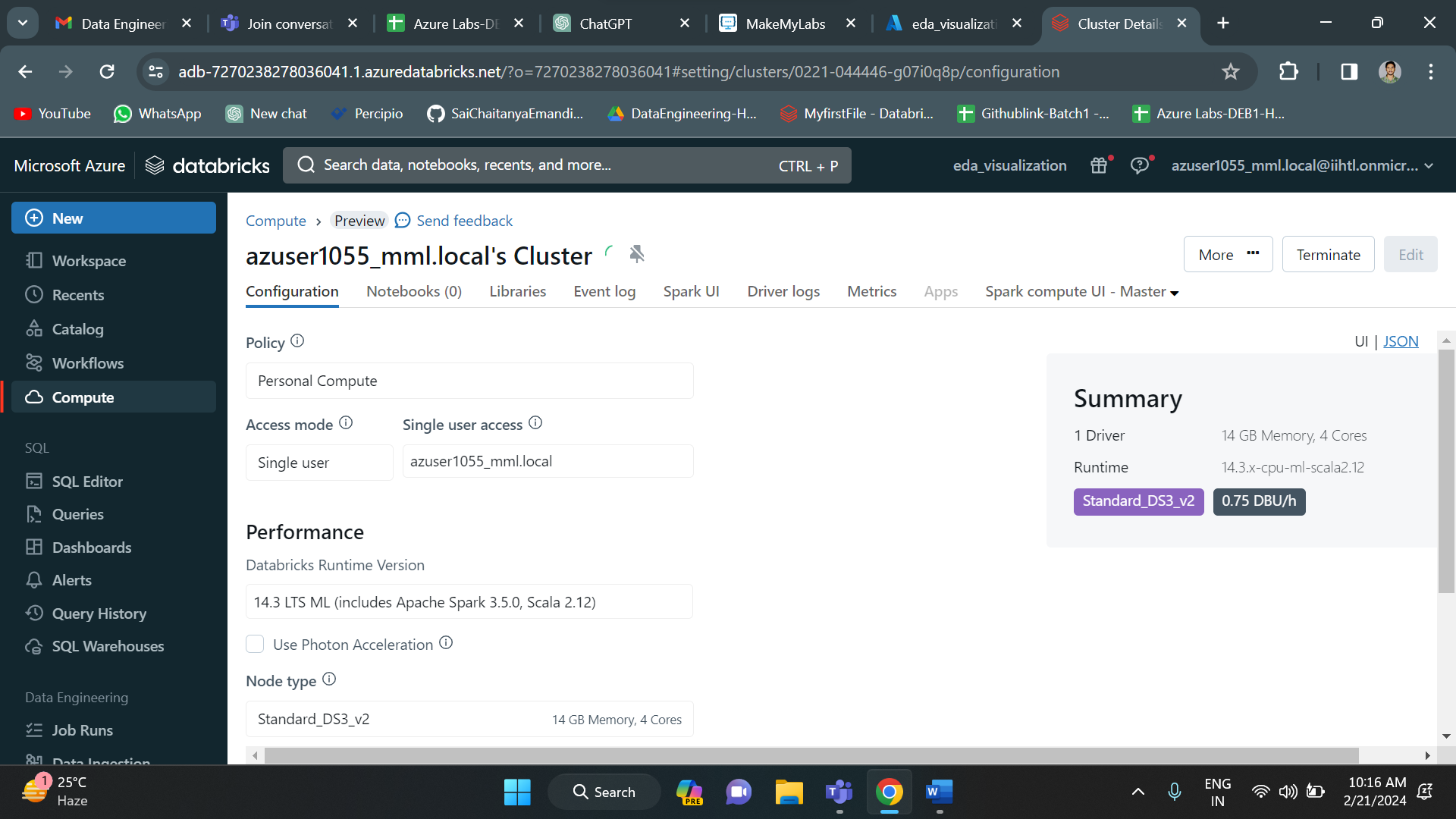
Below, picture represents the overview of azure databricks with a name called “eda\_visualization”. Copy paste the URL to go to the databricks platform.



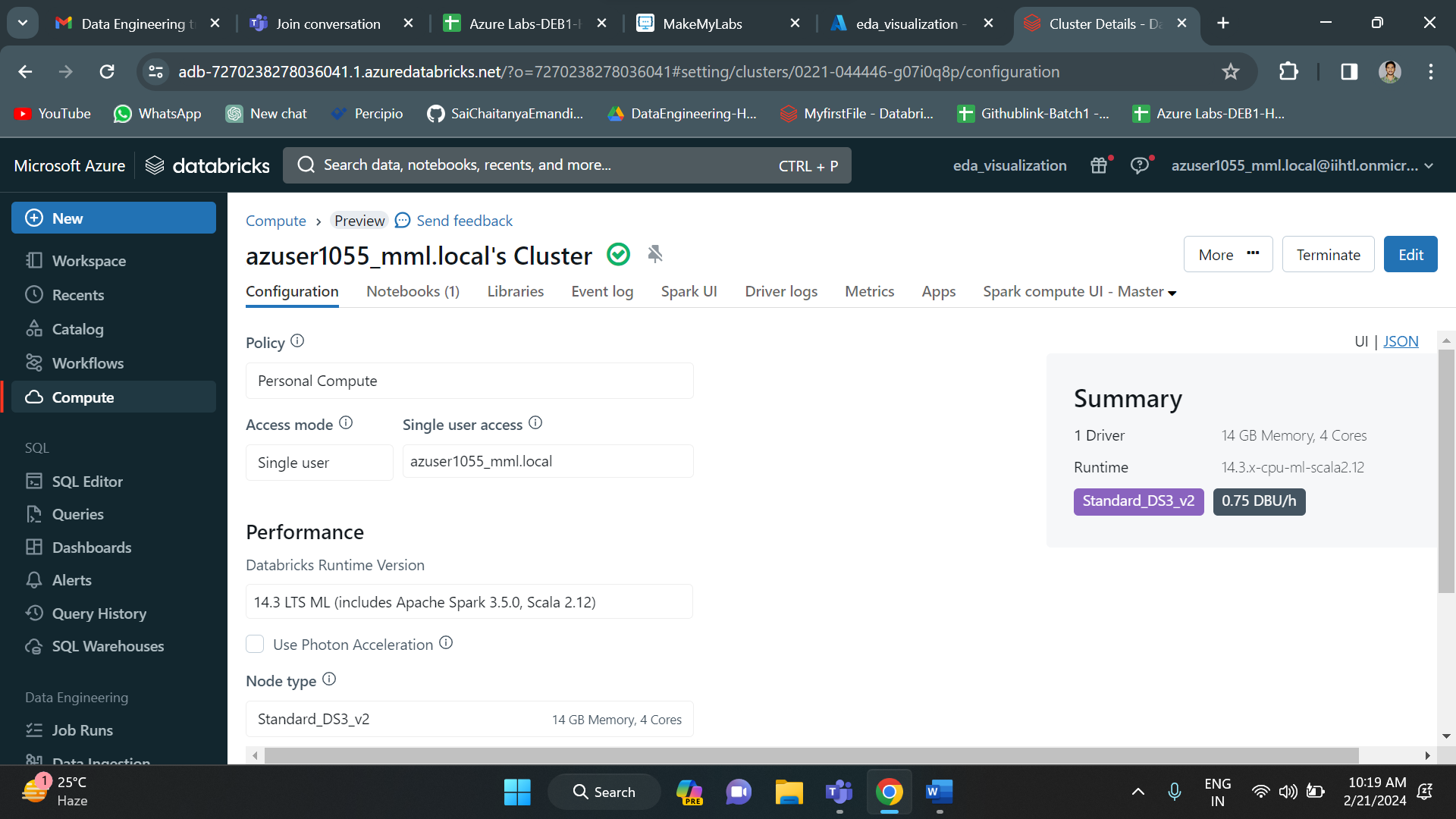
Below is the azure databricks homepage.



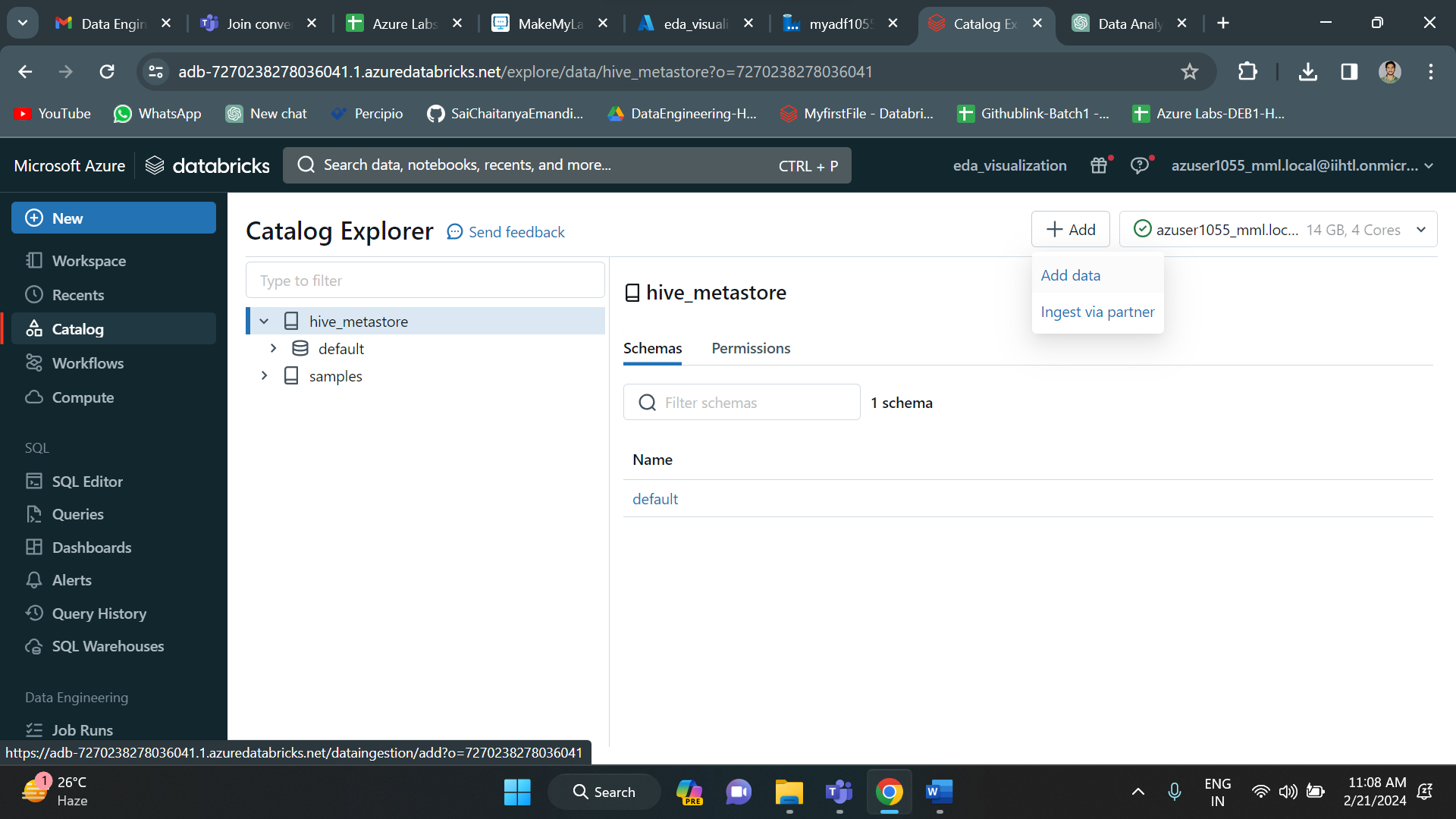
Now go to compute to create a cluster. Fill all the details and click on create to create a new cluster.



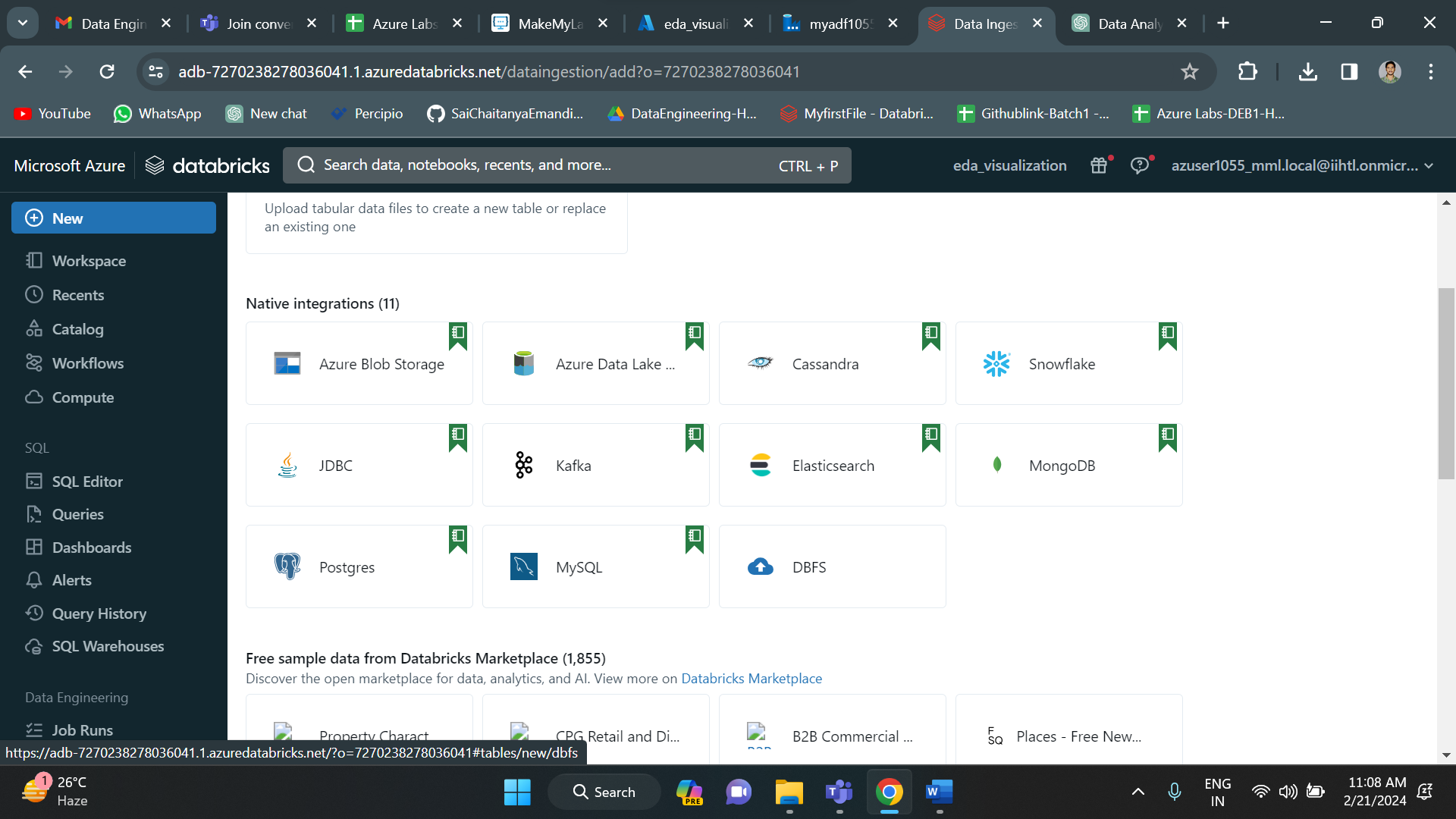
Now the cluster has been created.



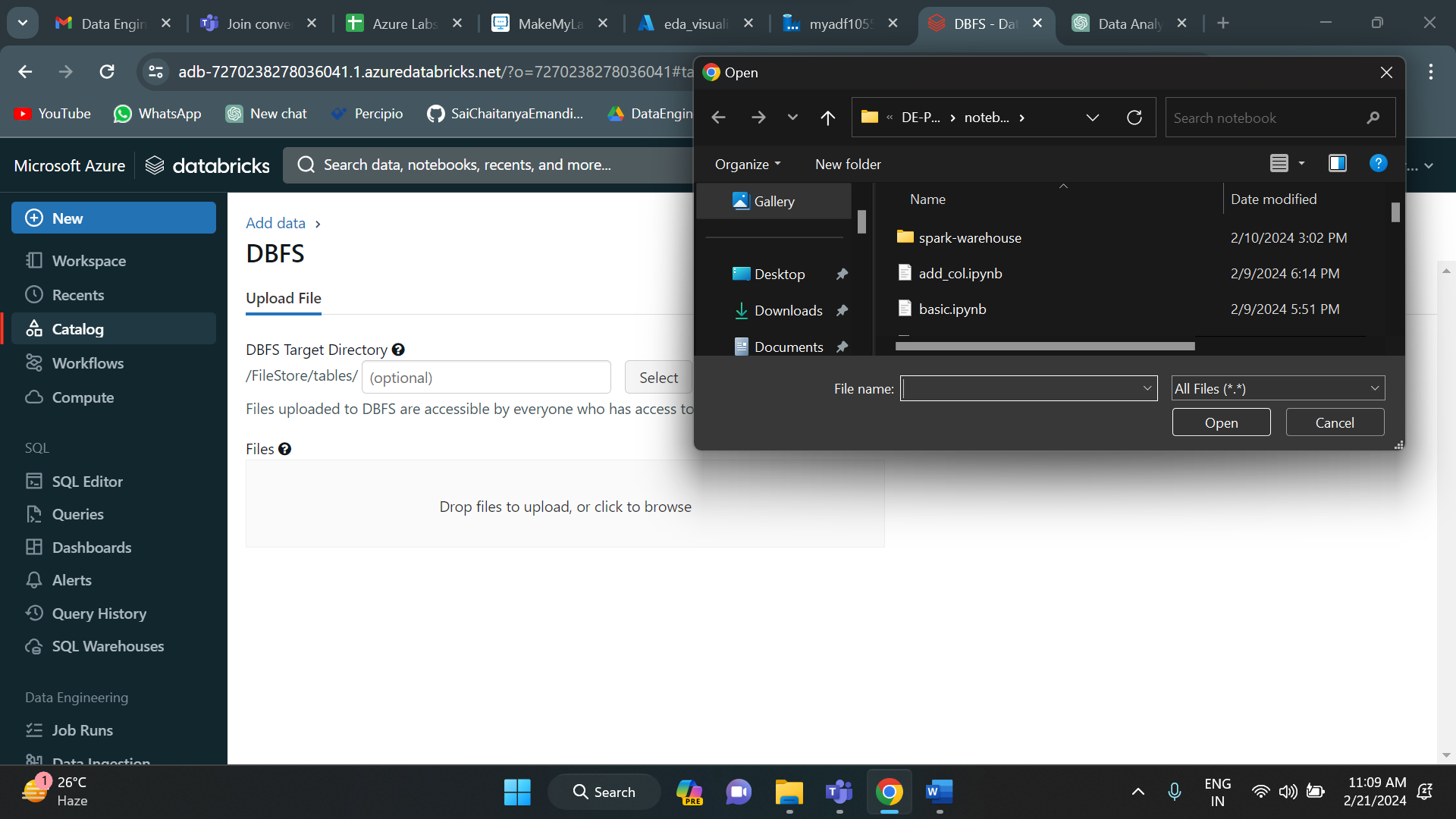
Next, we should upload data. For this got to catalog and then click on add-to-add data.



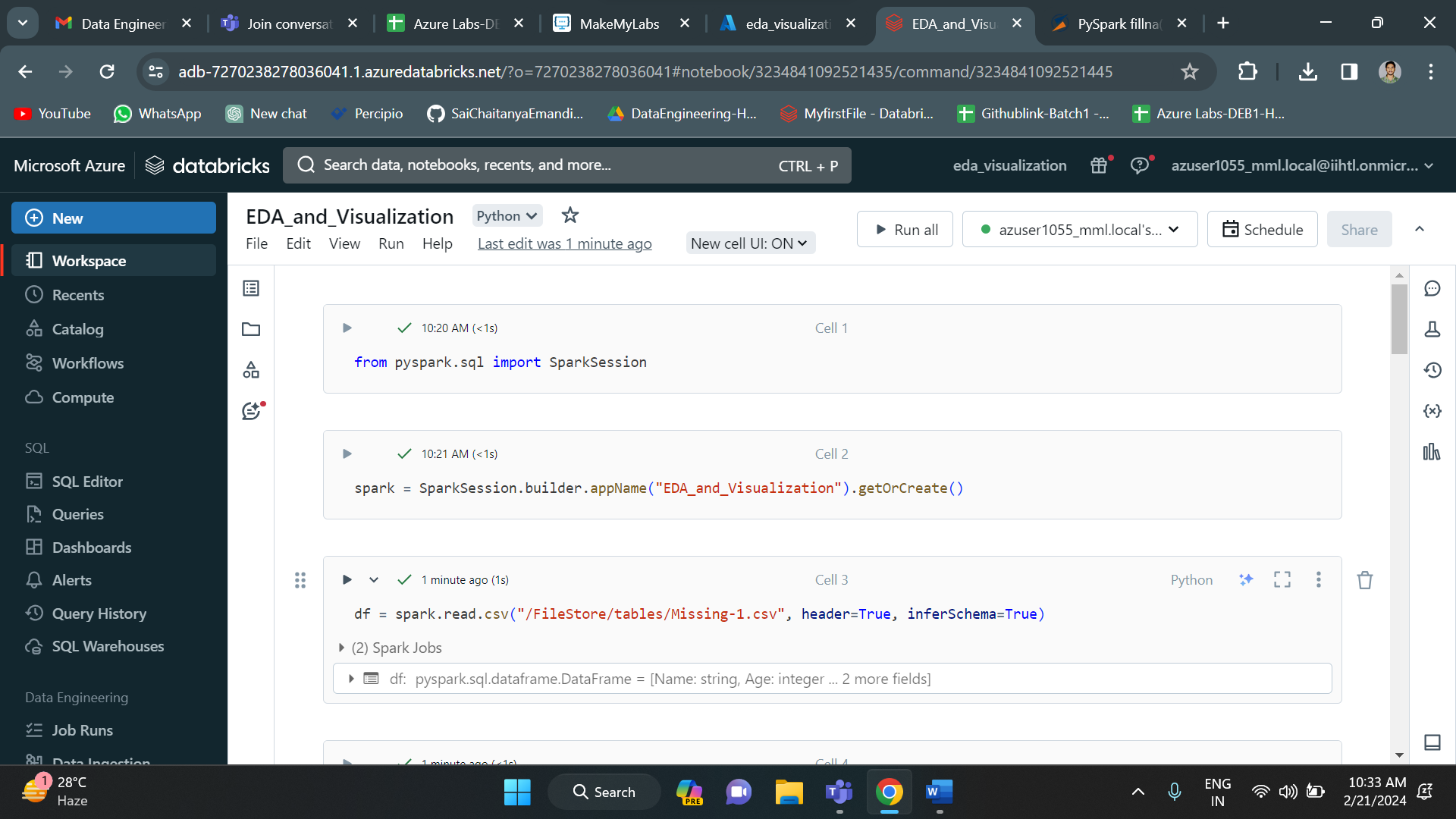
Select the DBFS option.

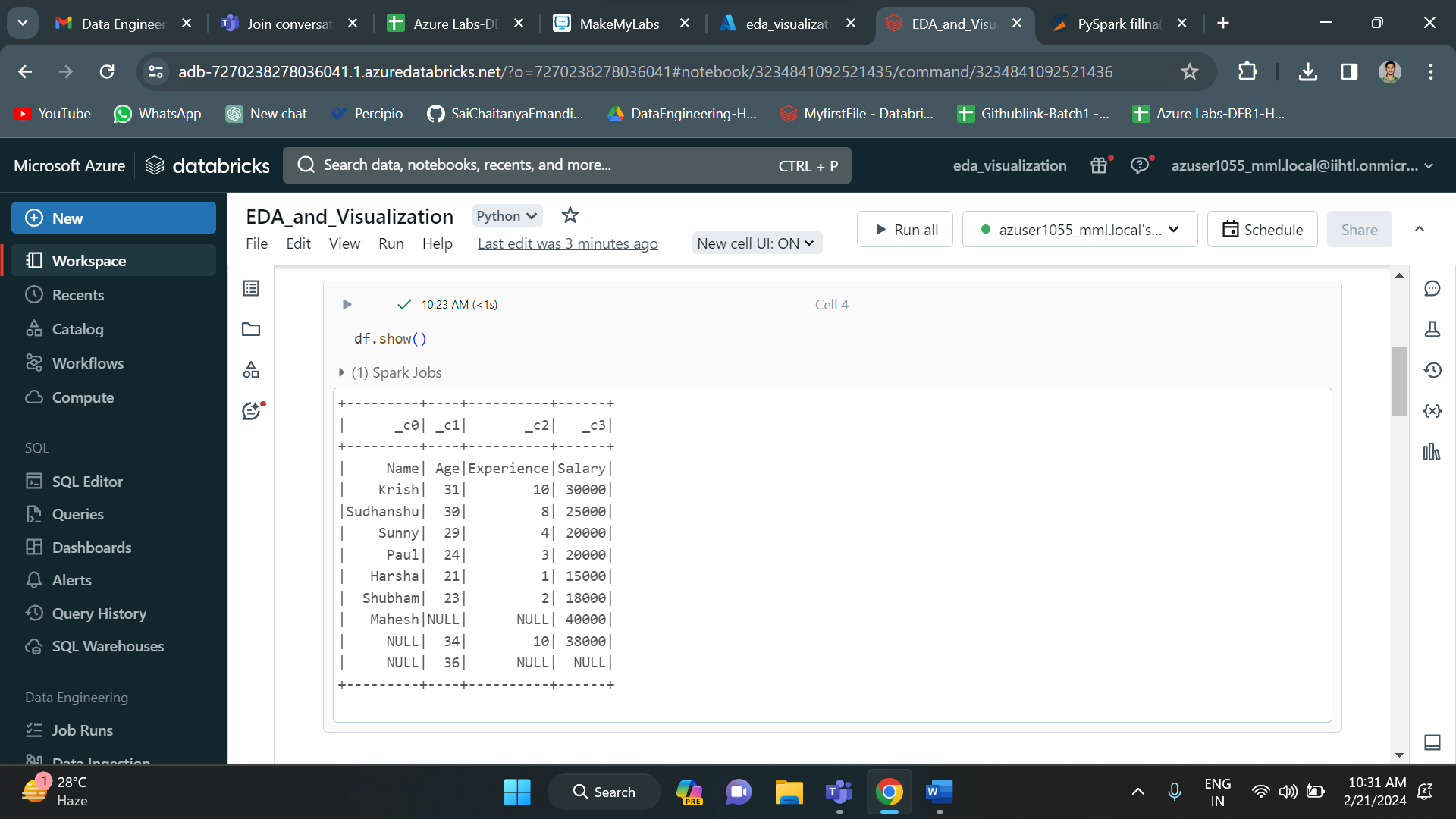


Now click on “drop files to upload or click to browse” button to add files and then press ok. The file gets added. I have added file called “Missing.csv”.

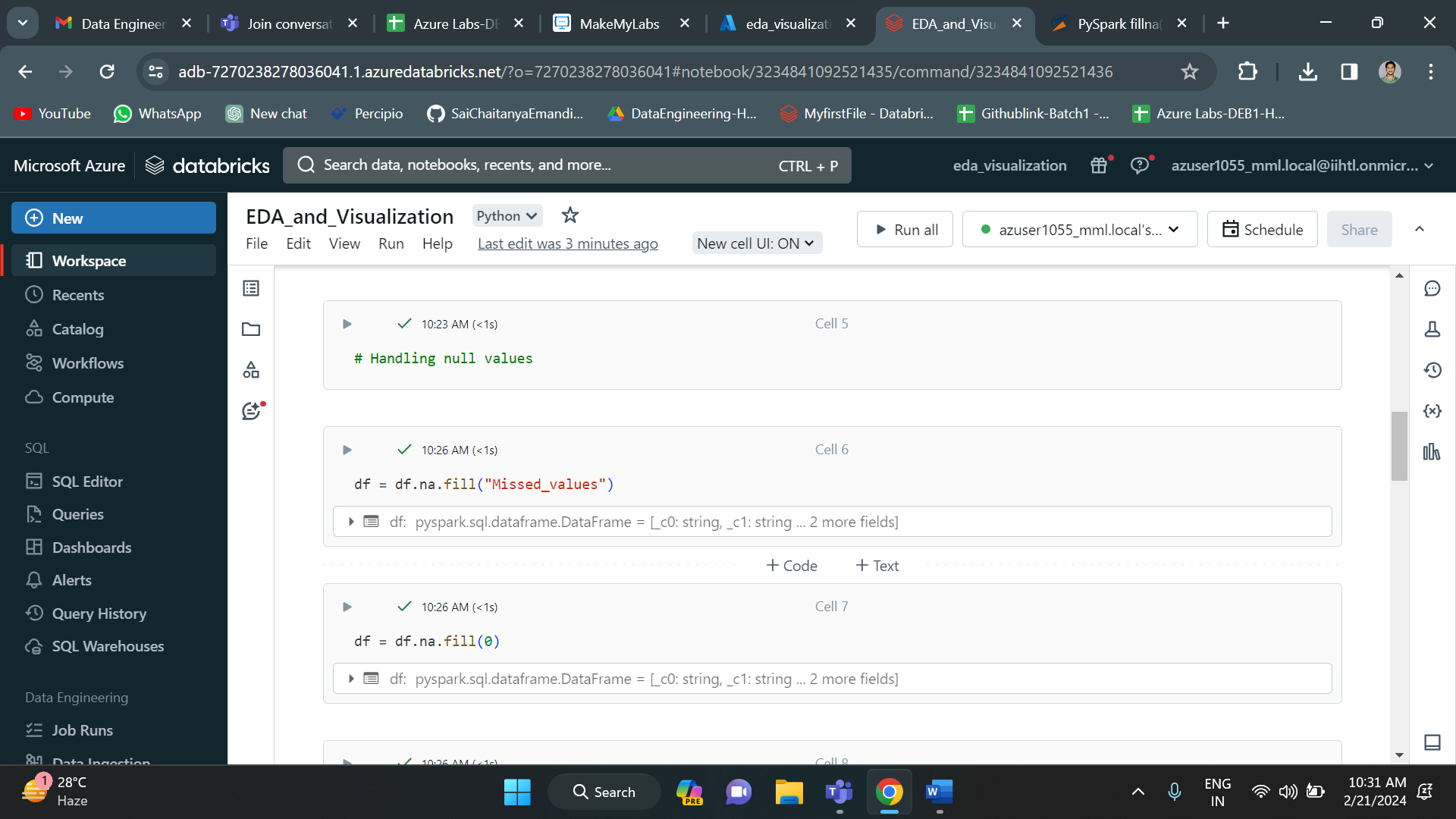


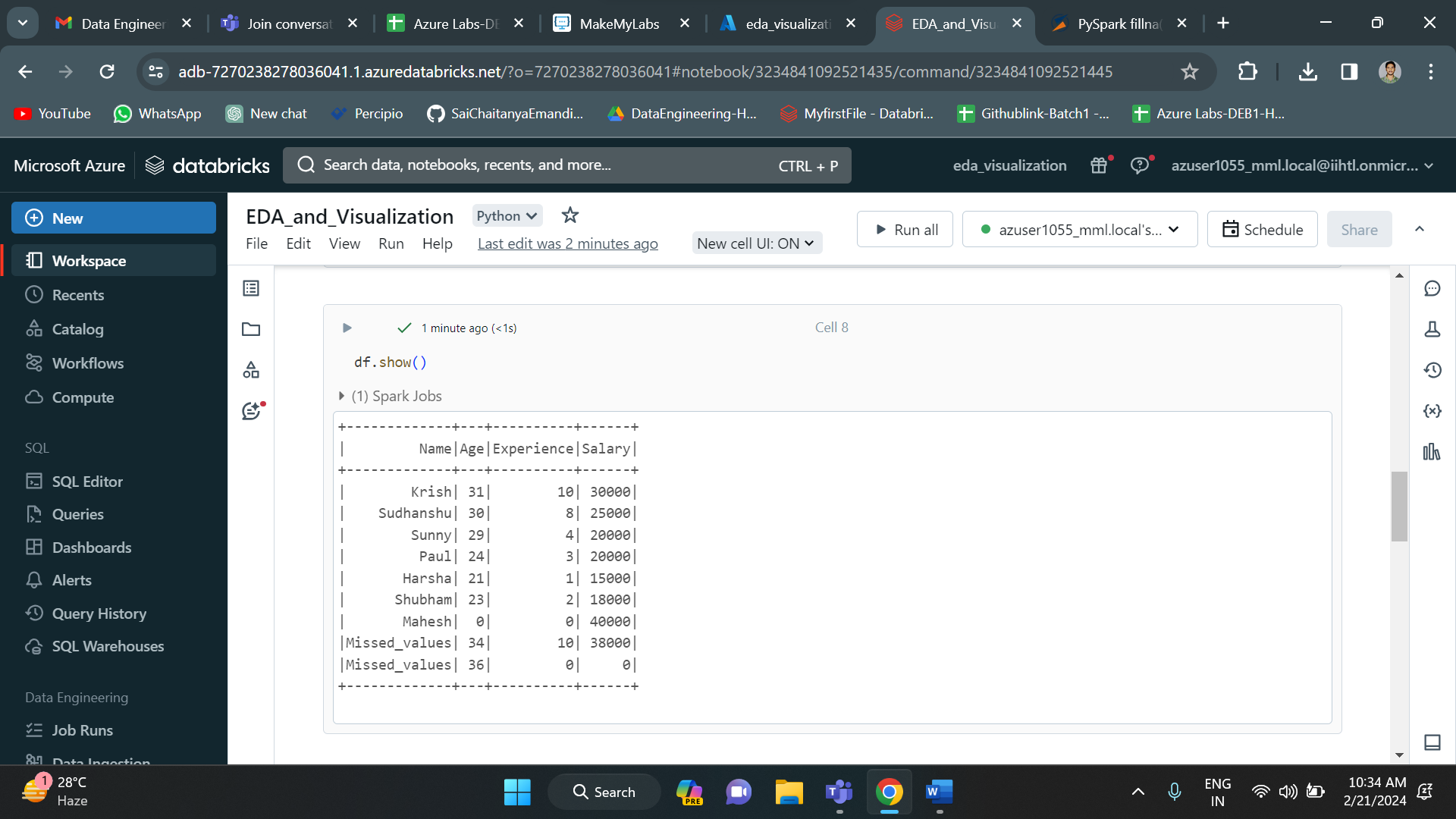
After adding the data, go to new click on notebook to get a notebook. Now you write the code inside it. I have imported required modules and created a spark session. And then I have read the csv file which I have uploaded earlier and done data cleaning process.





Since there are some null values in the dataset, I have removed them used na function in pyspark.

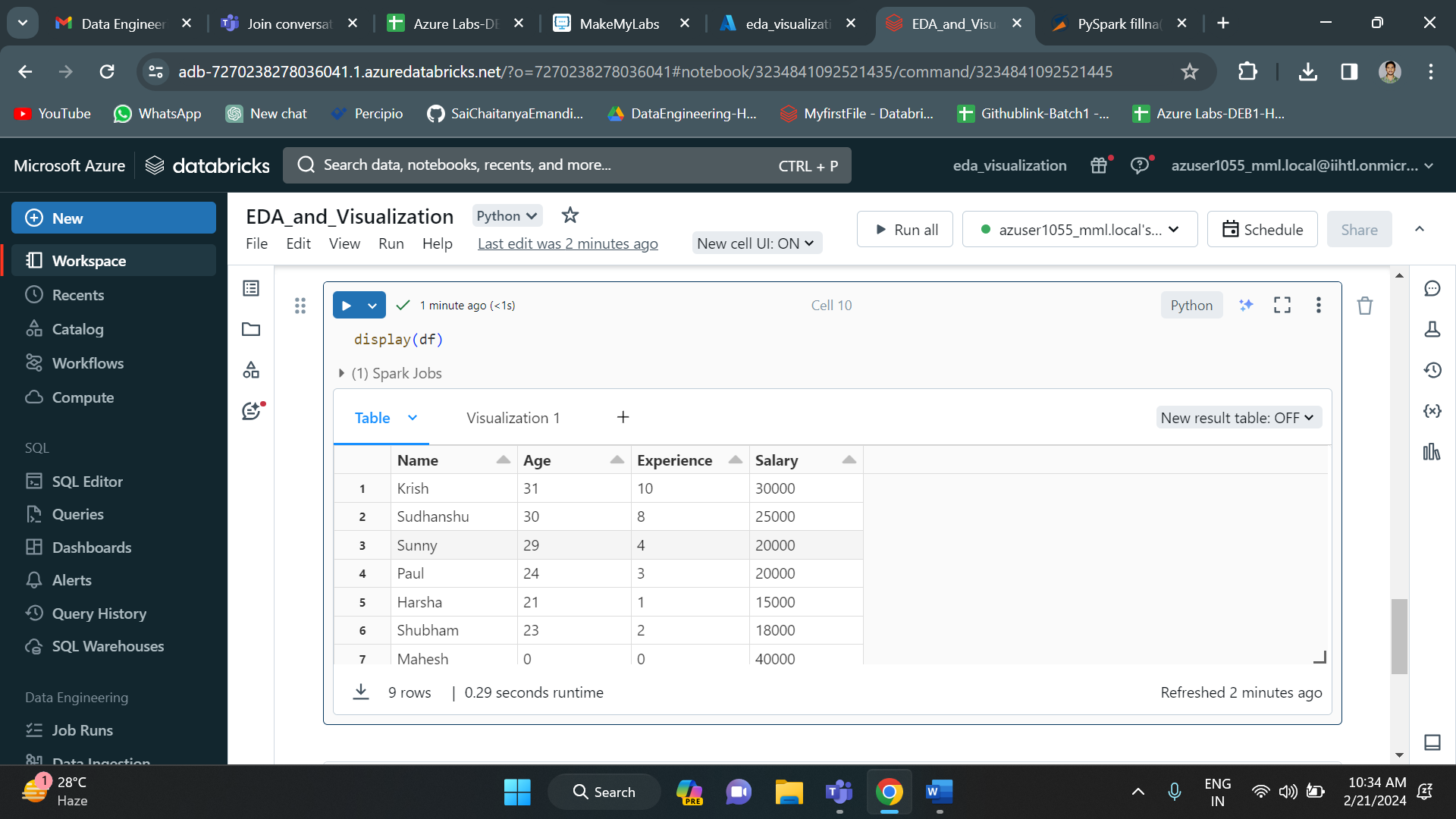




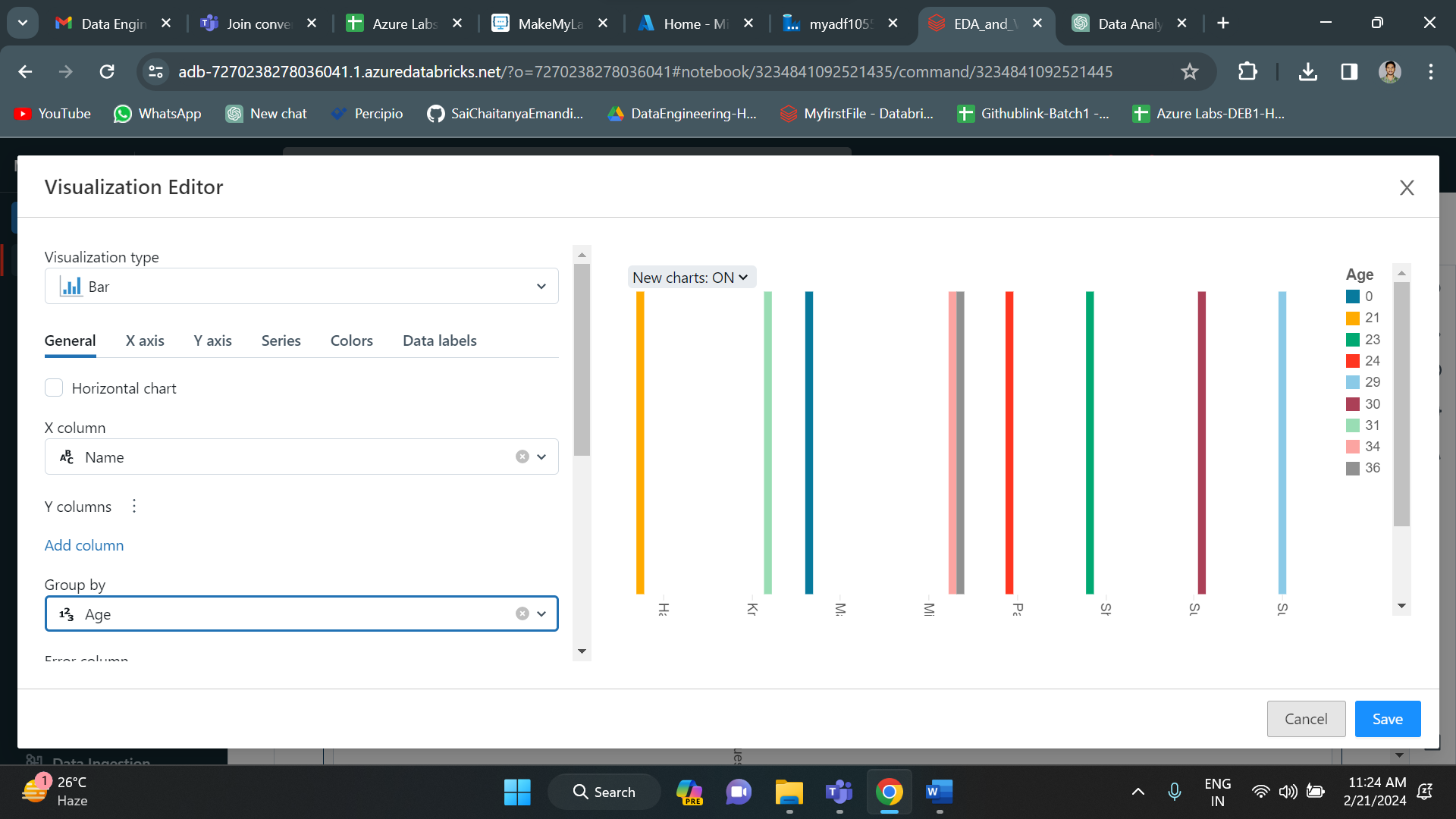


With this the data exploratory part is completed. I have loaded the data, applied cleaning process and transformed the data. Now it is ready to perform visualization.

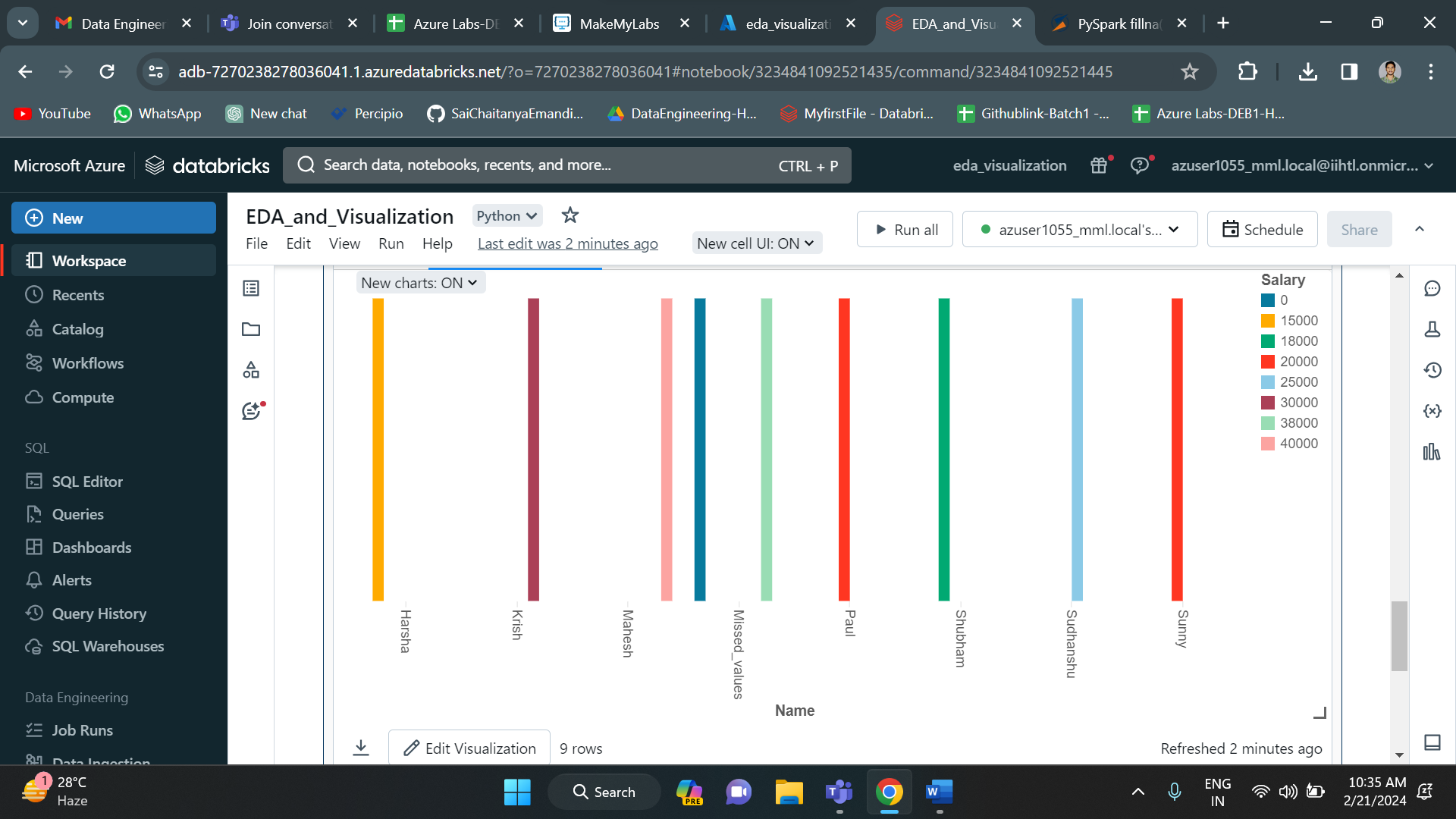
In order to perform visualization, first display the data using display function.



Then press on “+” icon and select visualization give all the column details.



Click on save to save the visualization.



Finally the visualization is successfully completed.