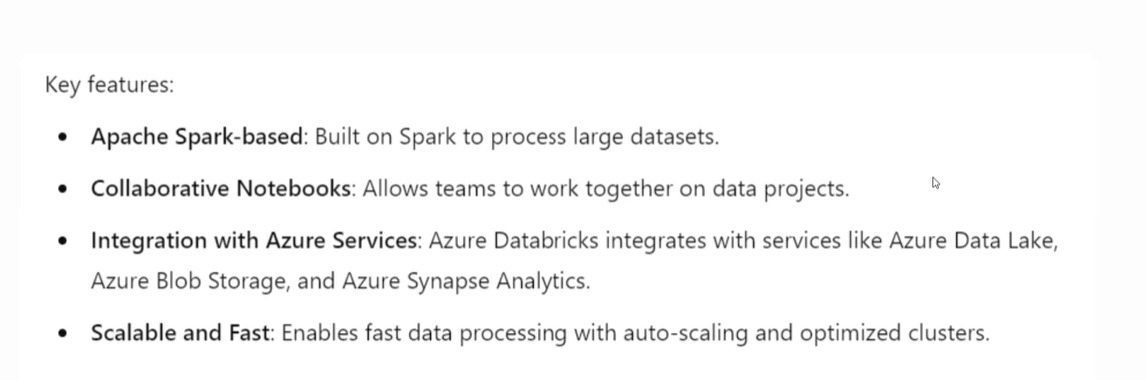
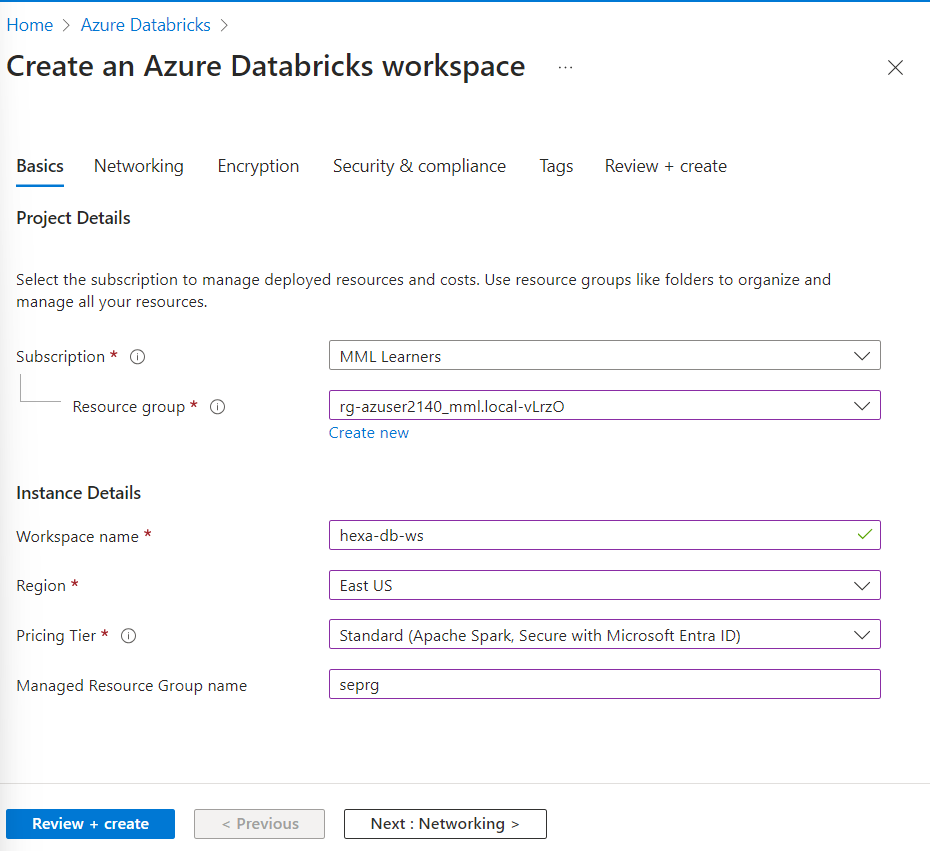
**AZURE DATA BRICKS**

**Day 1 (11-09-2024)**





Subscription: MML Learners(company access)

Resource group: default one

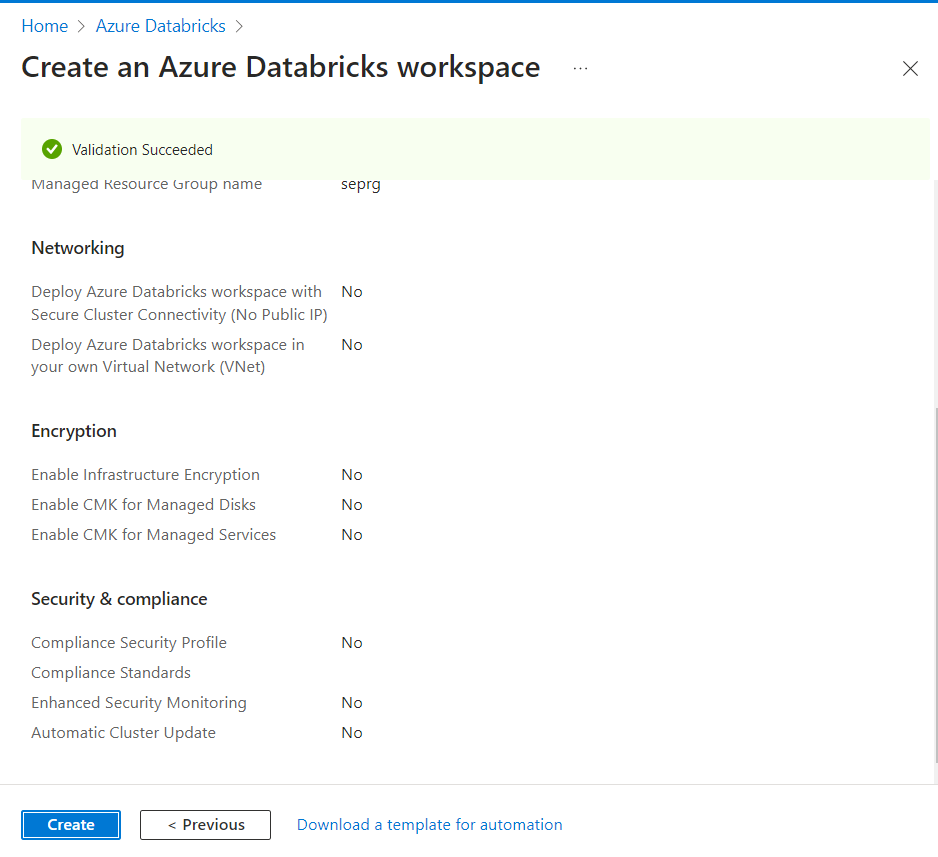
Workspace name: preferred name for that brick

Region: should be based on the near location, now it should be default due to access

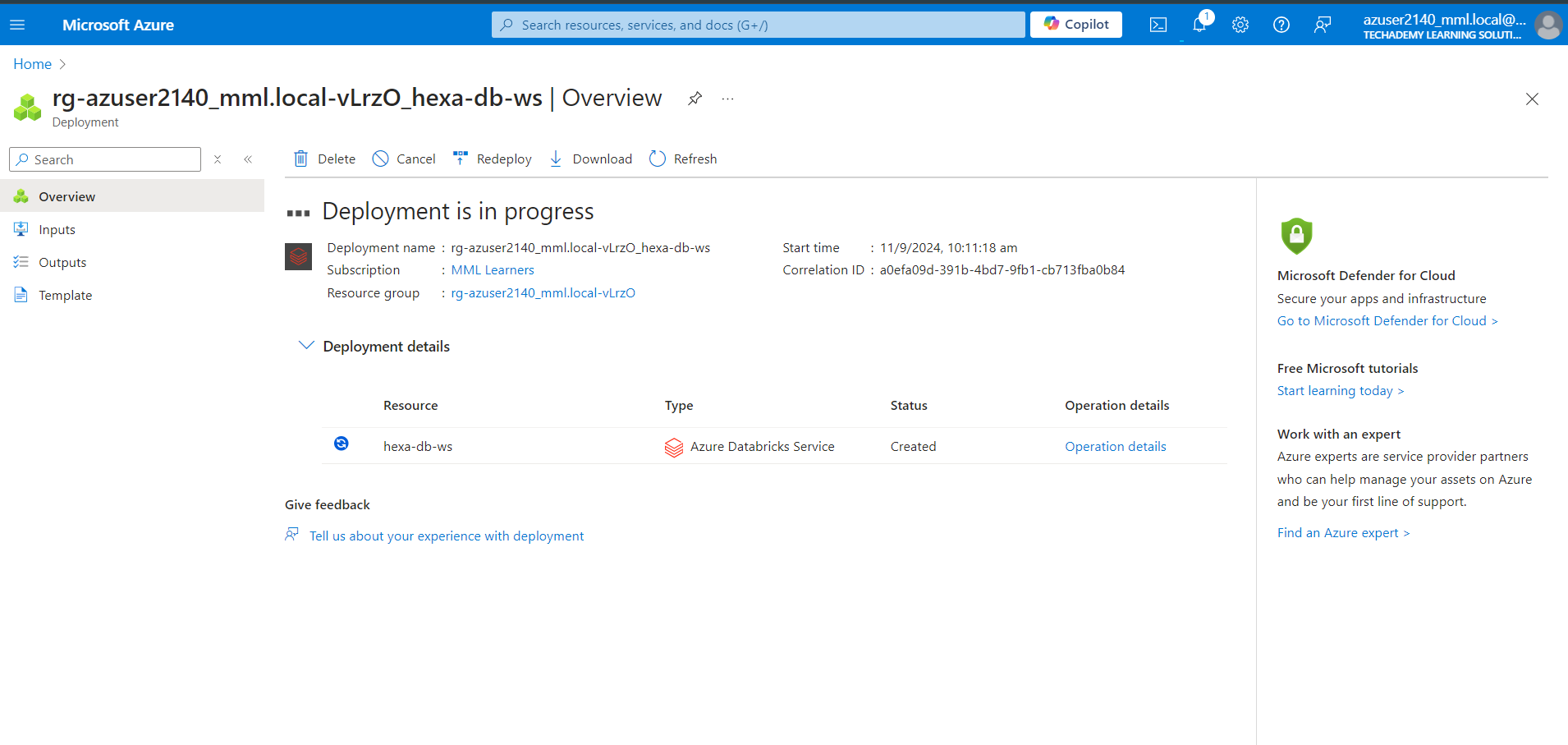
Pricing Tier: based on access by company

Managed Resource Group name: September rg

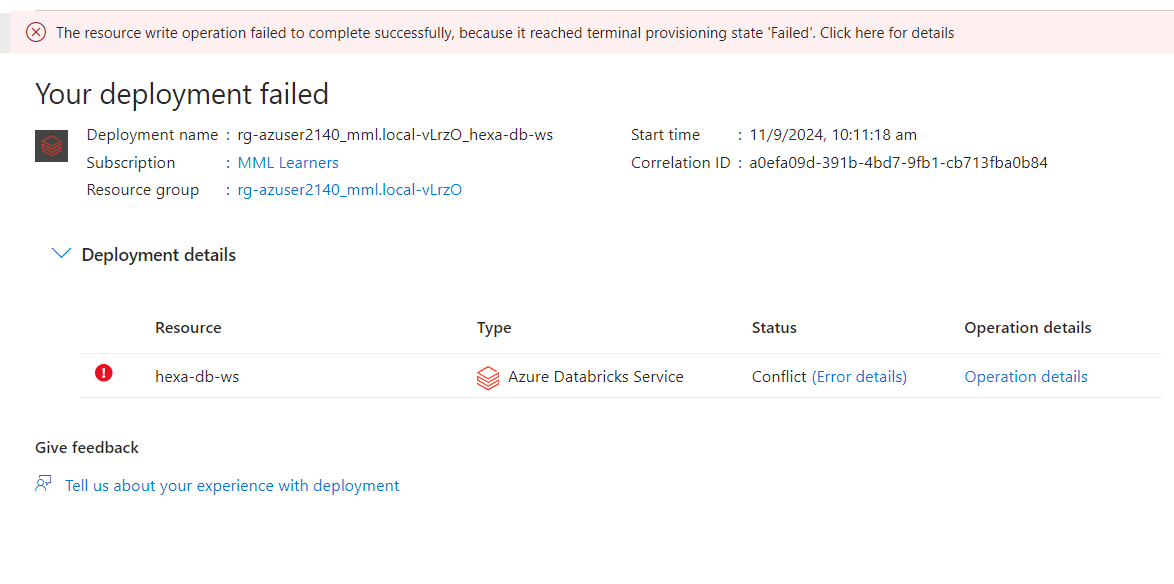
**Click review and create,once validation succeeded you are good to go**



Once u click on create 🡪 deployment will be in progress

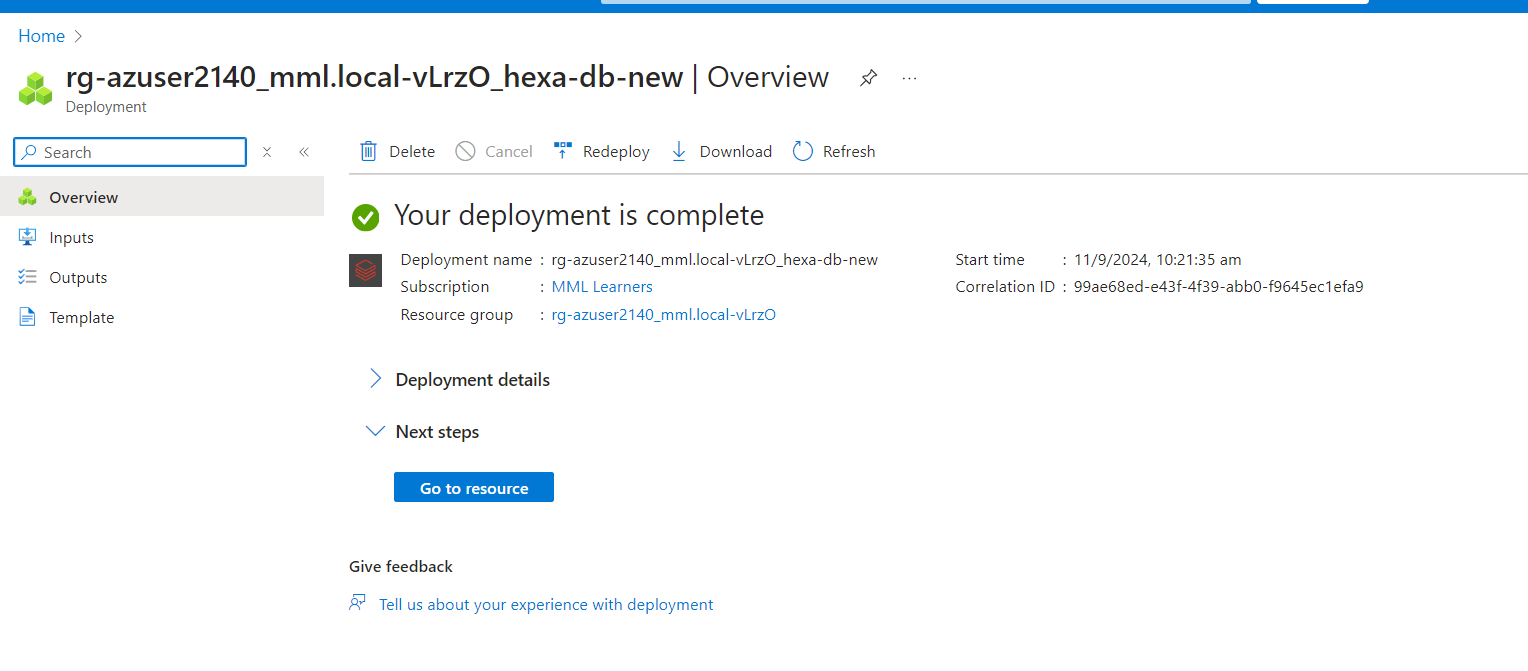


Deploymemt failed

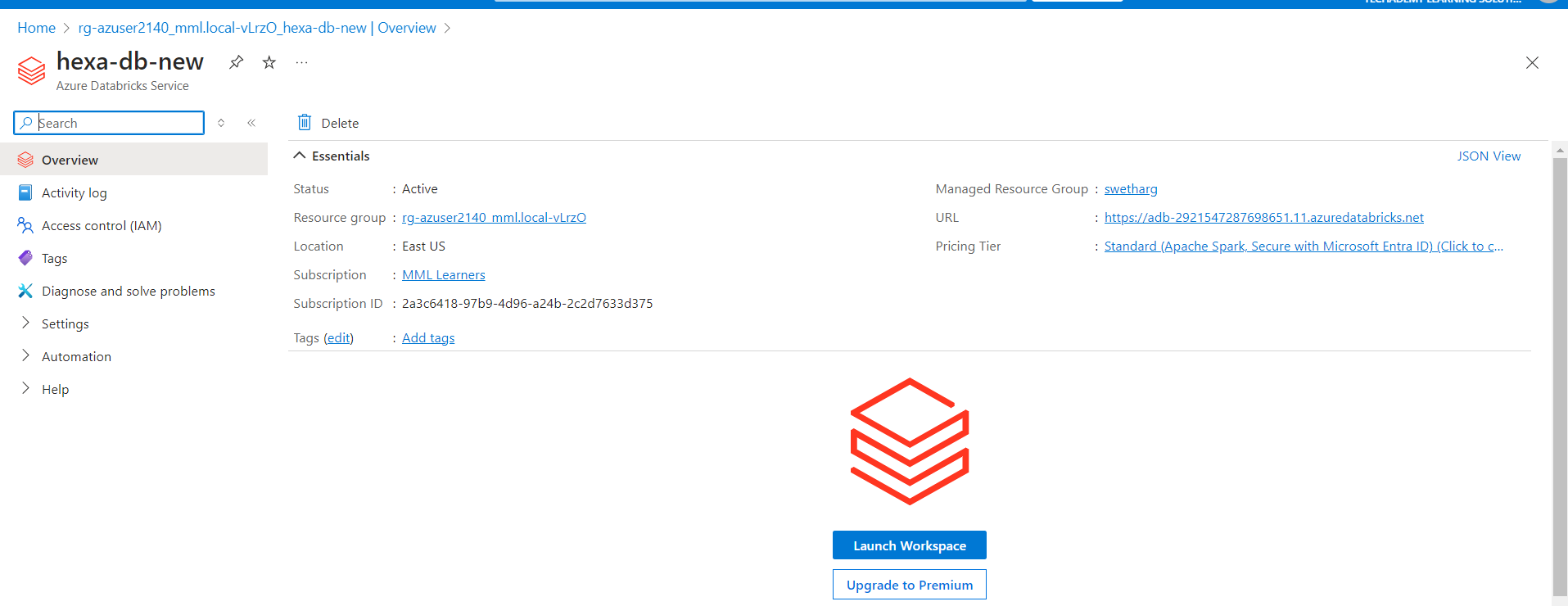


I solved the error by repeating the same process one more time and u need to change managed Resource Group name too so that deployment will be completed  
  
soln:

Repeat the process

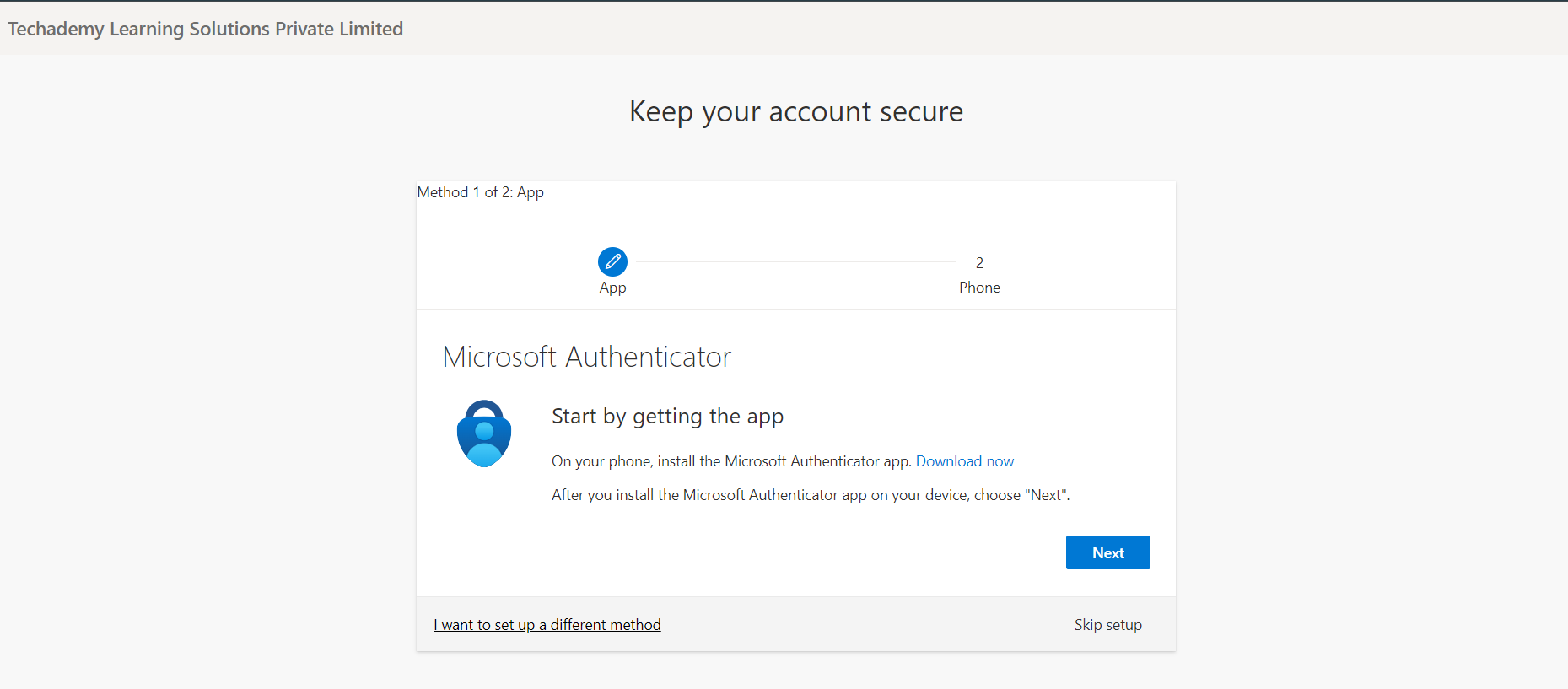


Click go to resource



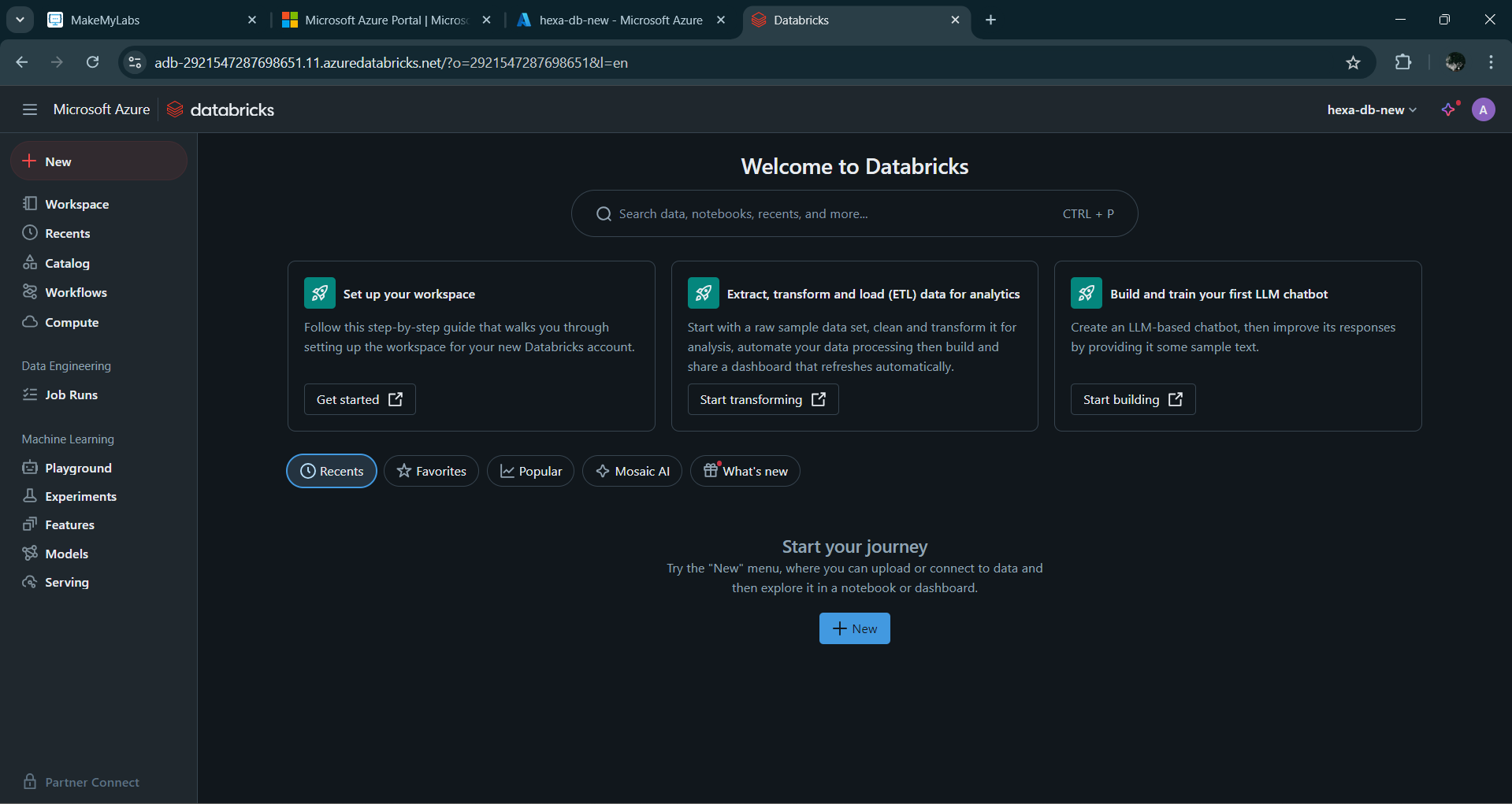
Now launch workspace

After entering pass key



Skip setup

Your workspace will be created



Compute is a virtual machine where we will work with our data so that we are creating our machine by giving specification like how much ram and all

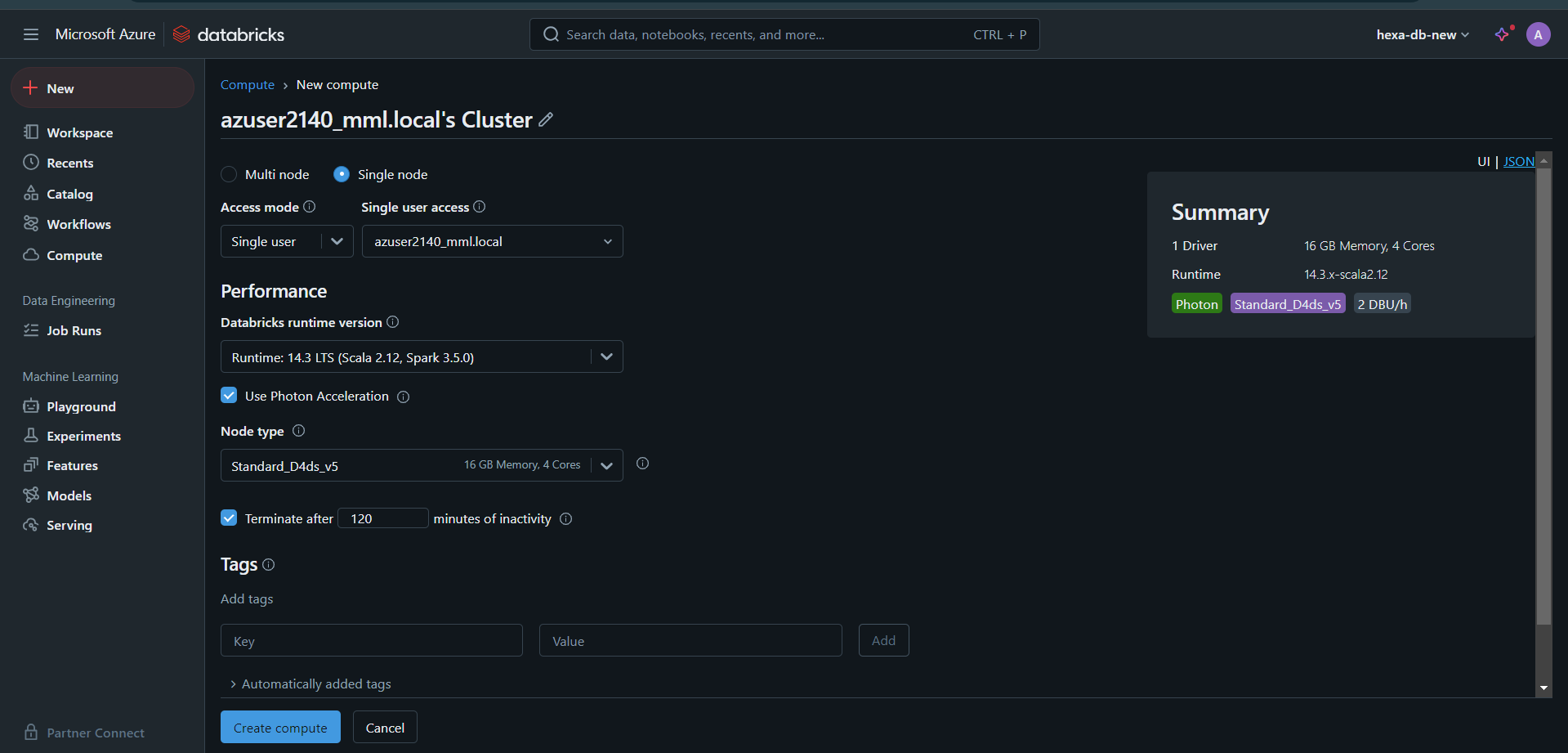
Eg:

Colab 🡪they are providing a environment for running code with pyspark environment ..so apart from colab we need some environment to work apart from google so they choose google over azure

Colab 🡪using free services

Azure 🡪hexaware will pay money

Inside workspace 🡪 compute 🡪 create compute



Single node 🡪 one person

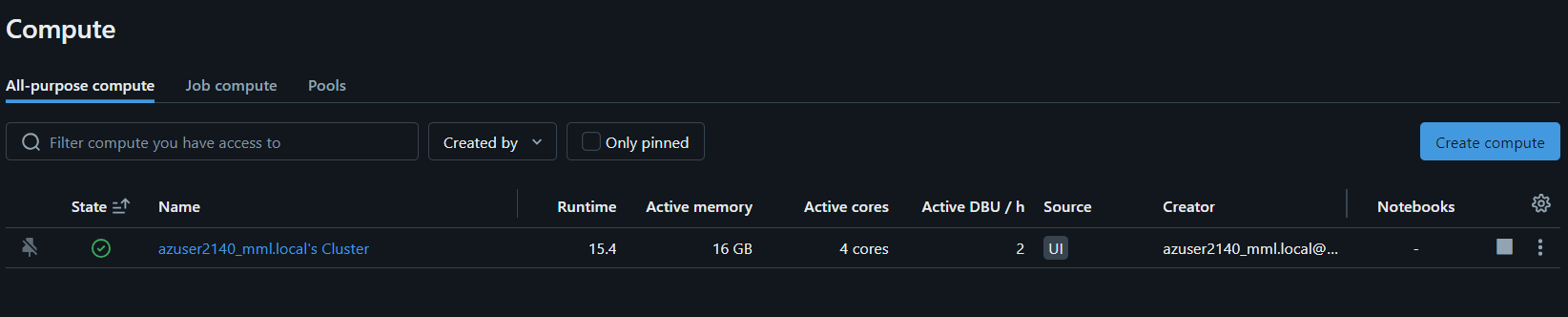
Multi node 🡪 shared workspace

User access <-> node type – leave it as default

Nothing to be filled in the region of tags

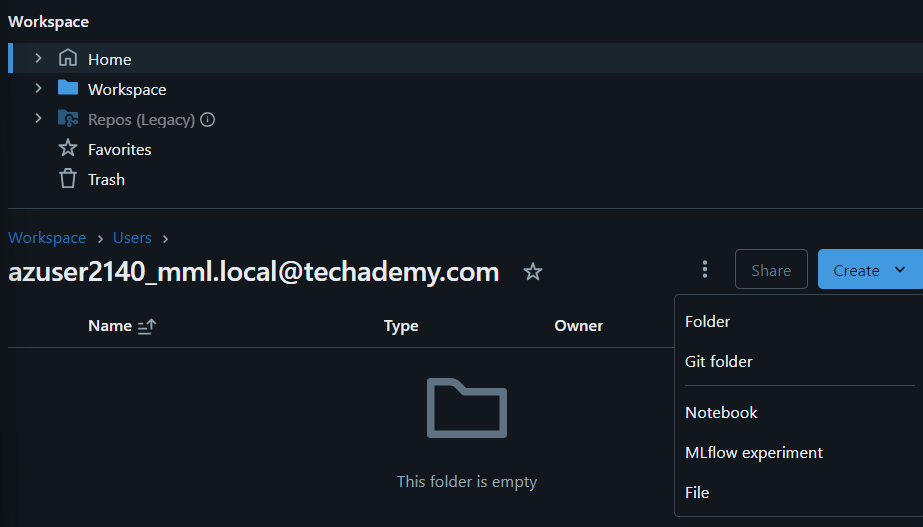
Then click create compute

It may take some time

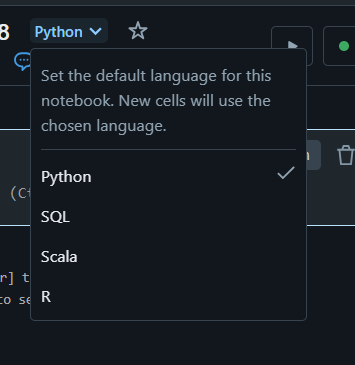


After that u need to create a notebook for that

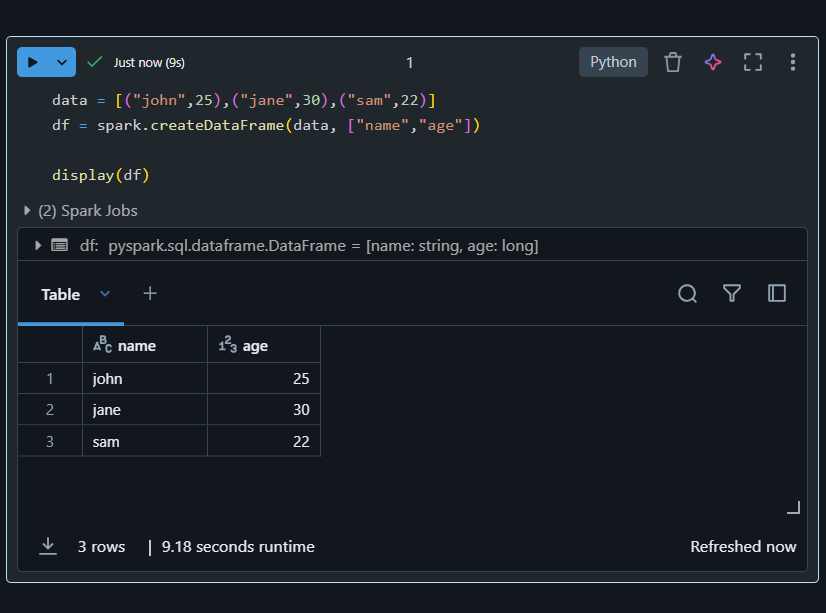
Go to workspace 🡪 create 🡪notebook



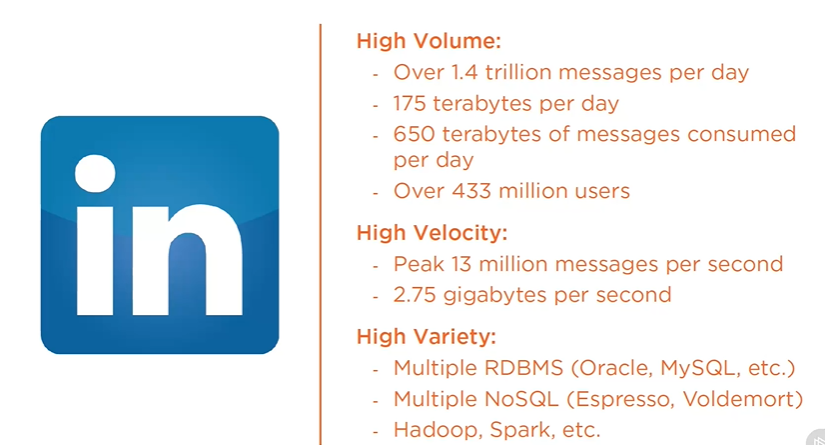
It’s upto u to choose language

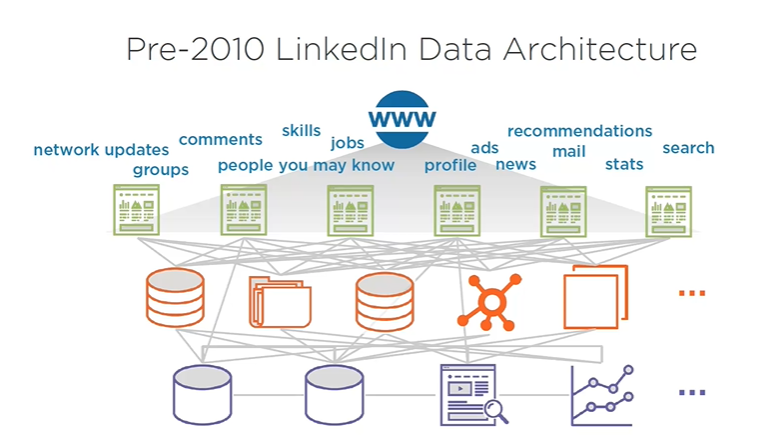


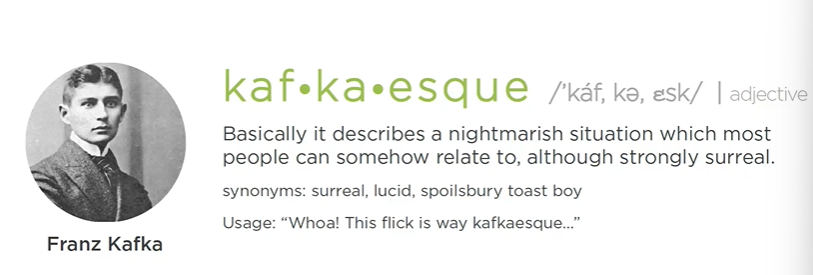
Now u can run code as per requirments

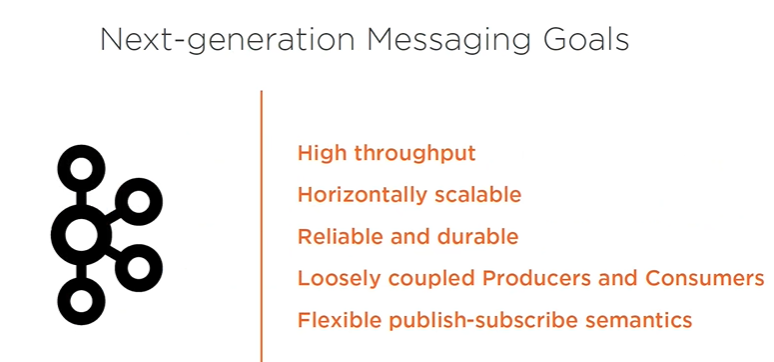


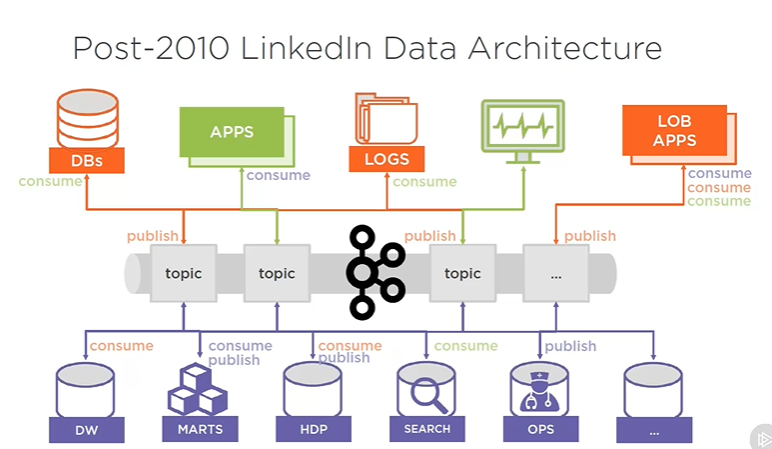
Workspace more similar with google colab



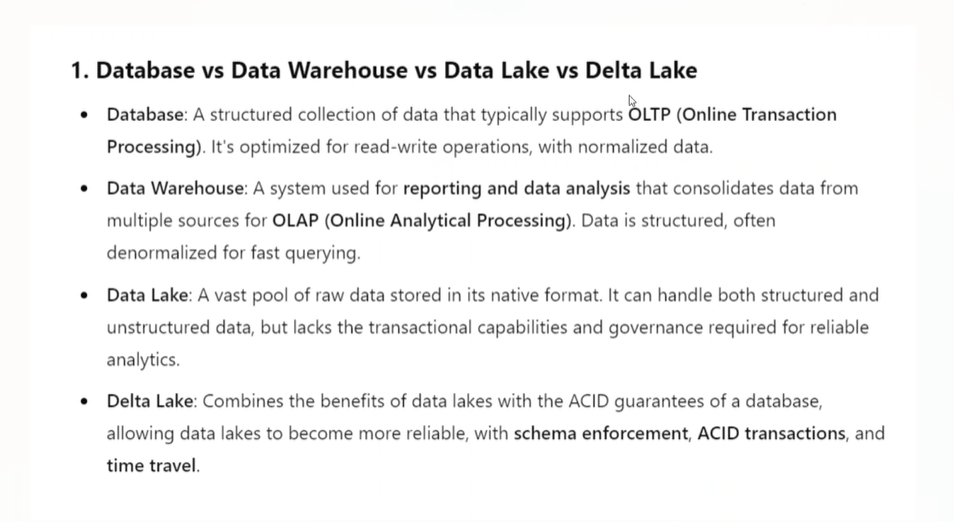


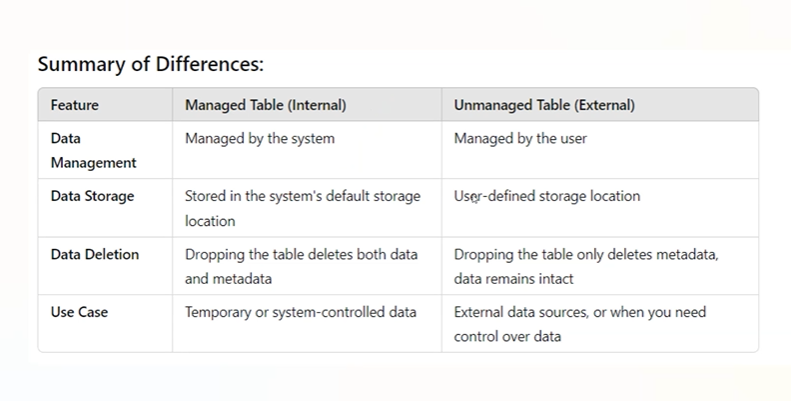


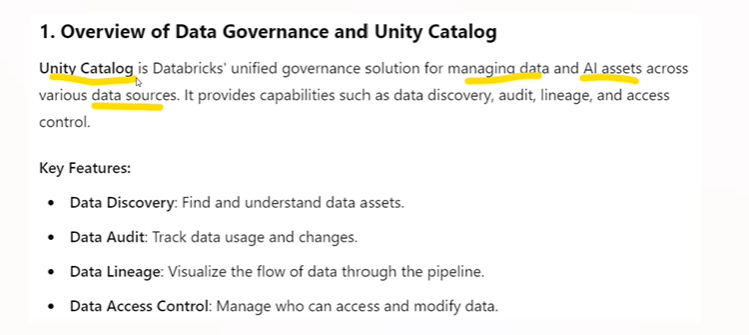


kafka way of handling data|^

Day 2 (12-09-24)







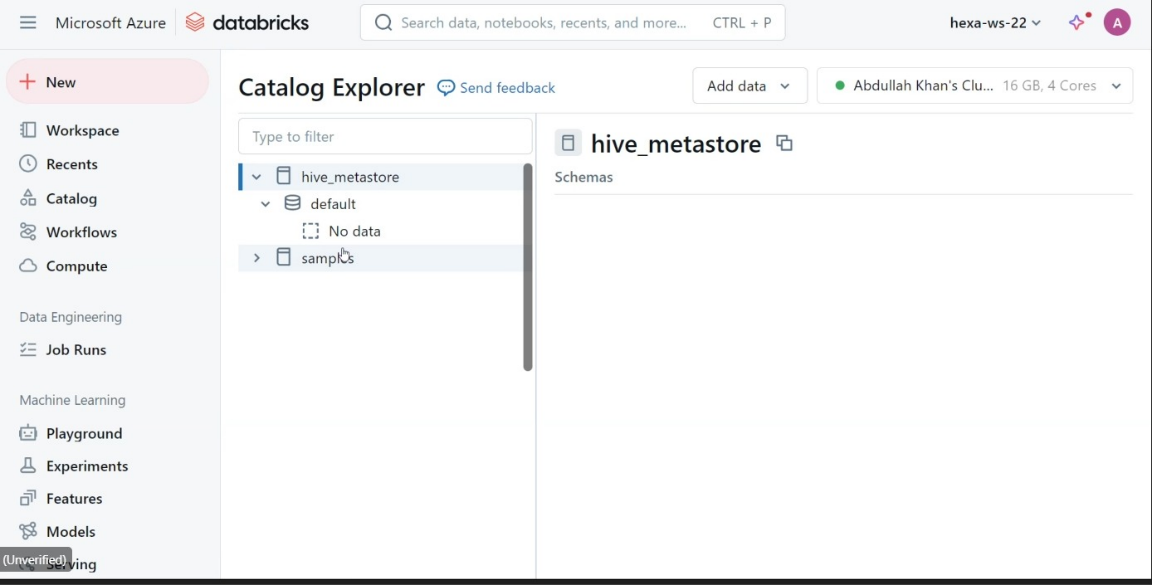
Catalog processing

Catalog – handle huge data

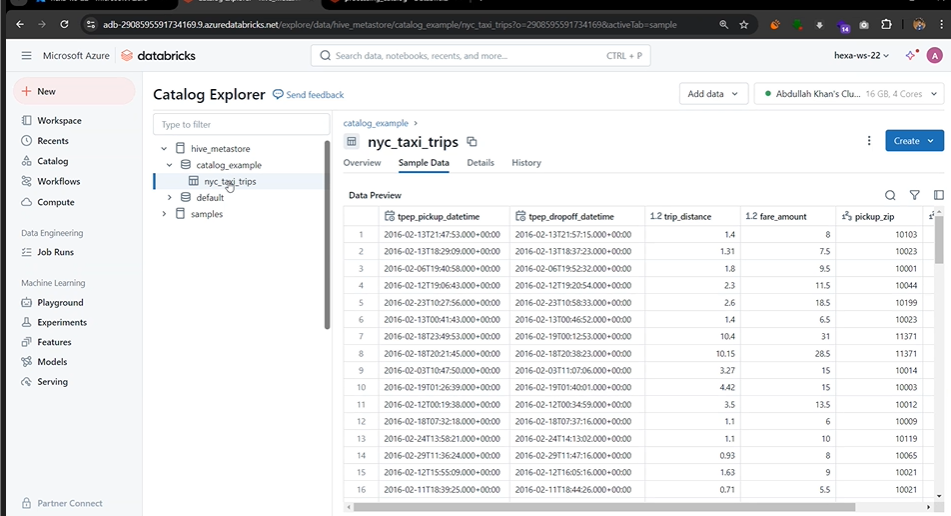
Here we cannot upload huge gb datasets so they provide some sample data to work on it like nyctaxi etc..,

Location of sample datasets

Azure 🡪 catalog 🡪 hive\_metastore 🡪 samples







<https://drive.google.com/file/d/1w5tFx3xR7mWnnx0mdhIqGvVccIroPxdG/view?usp=drive_link>

<https://drive.google.com/drive/folders/1LR00tJ_7Nay7PReqtpKe1AIMKRicxTpY>

unity catalog |^

----------------------------------------------

connecting notebook in data bricks

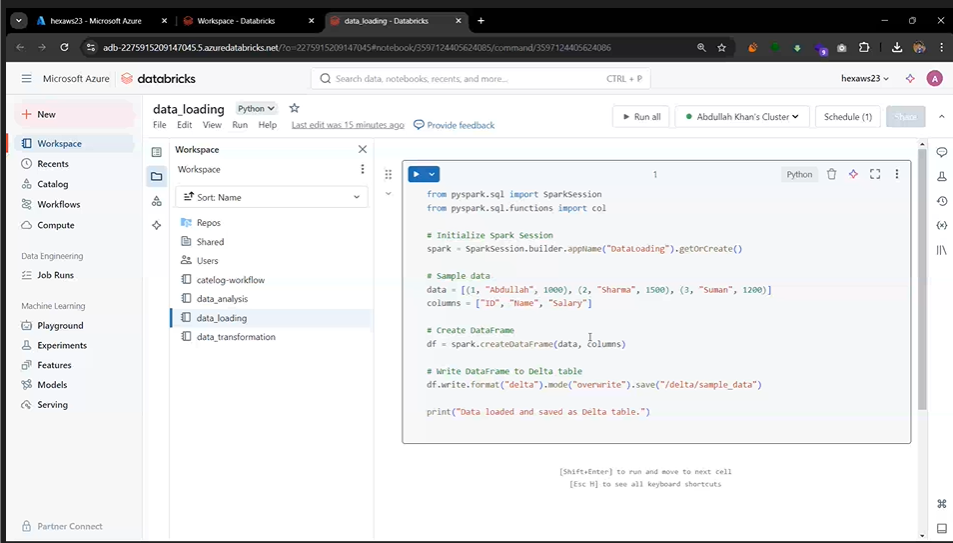
data\_loading

data\_transformation

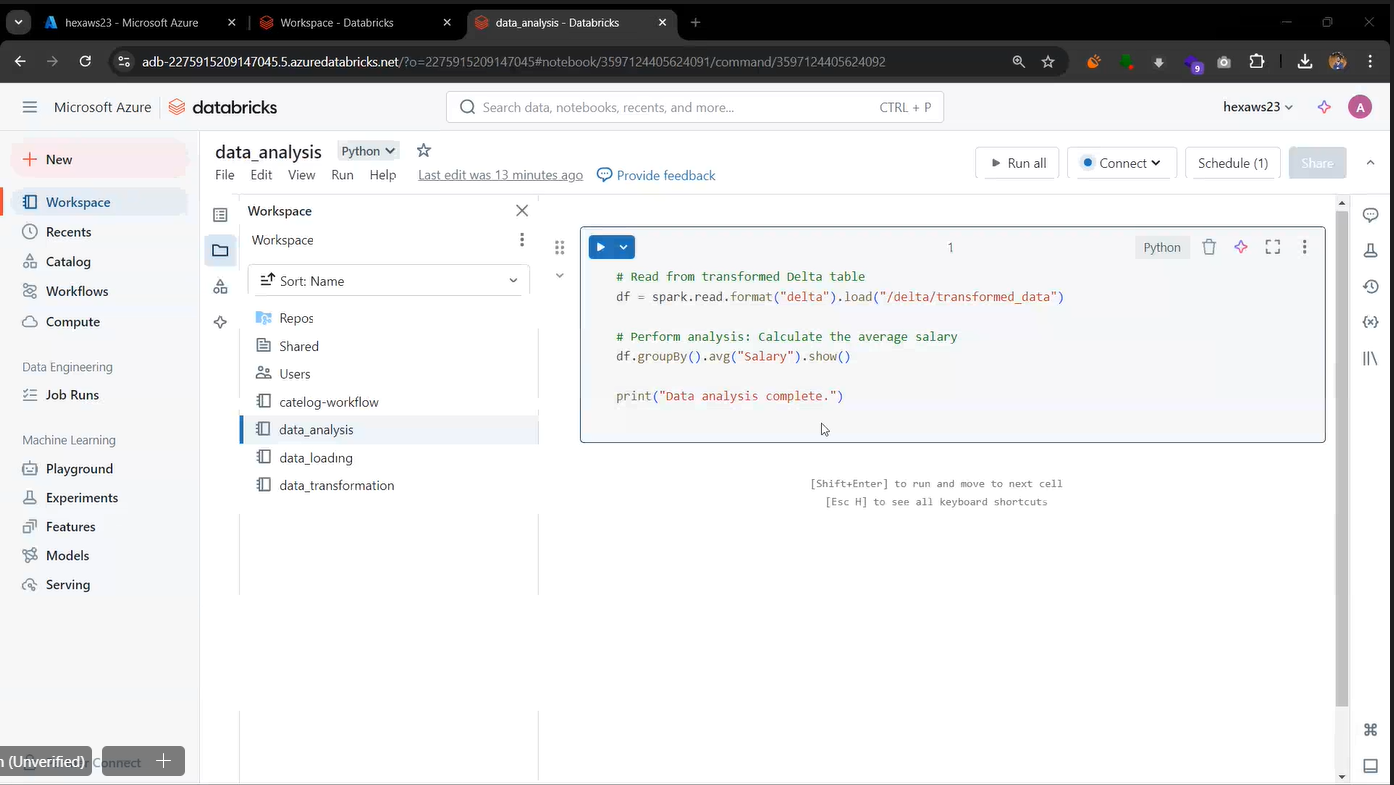
data analysis

each notebook is dependent on other notebook

data loading --> data transformation --> data analysis



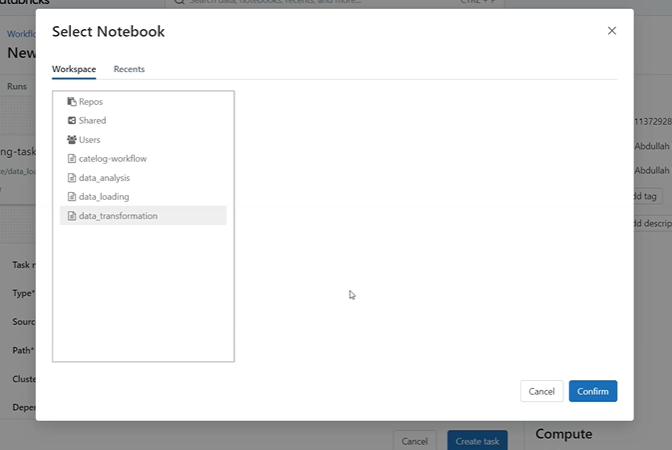




Pipelines 🡪 connecting one tasks to another

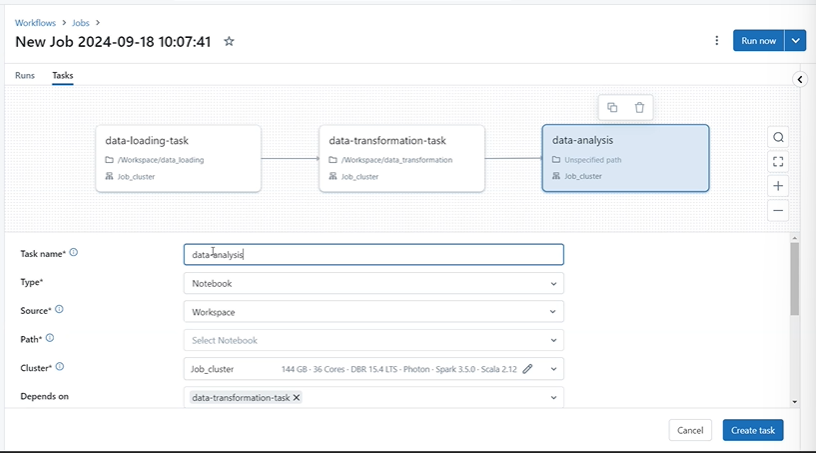
creating jobs

workflows 🡪jobs

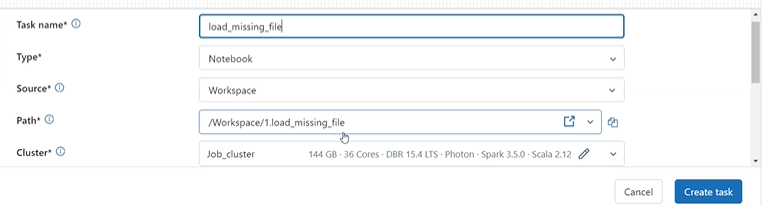




Connect 3 notebooks

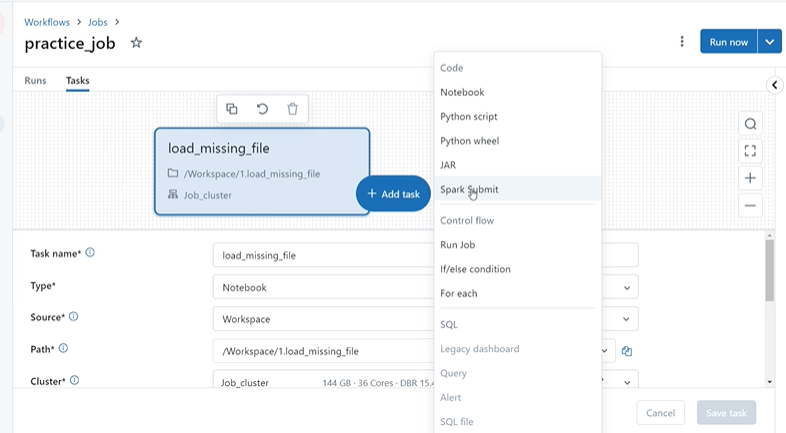


Parameters while creating jobs

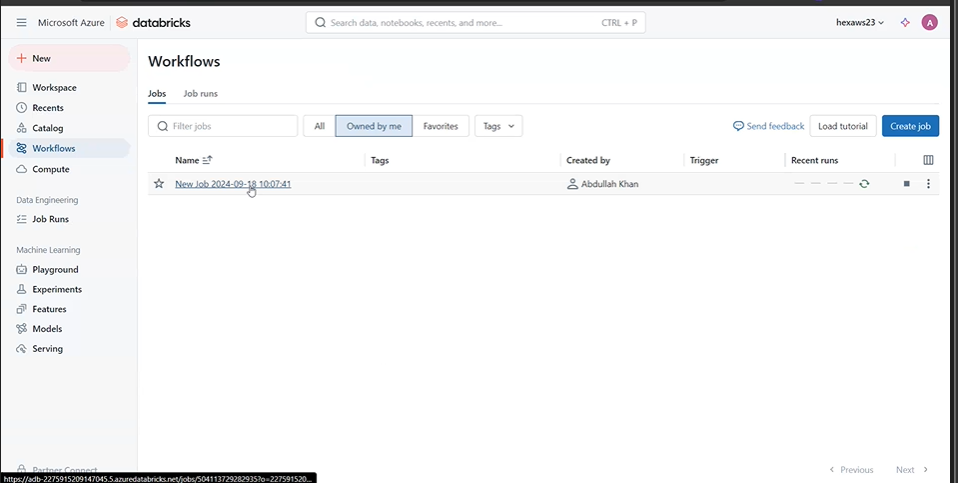


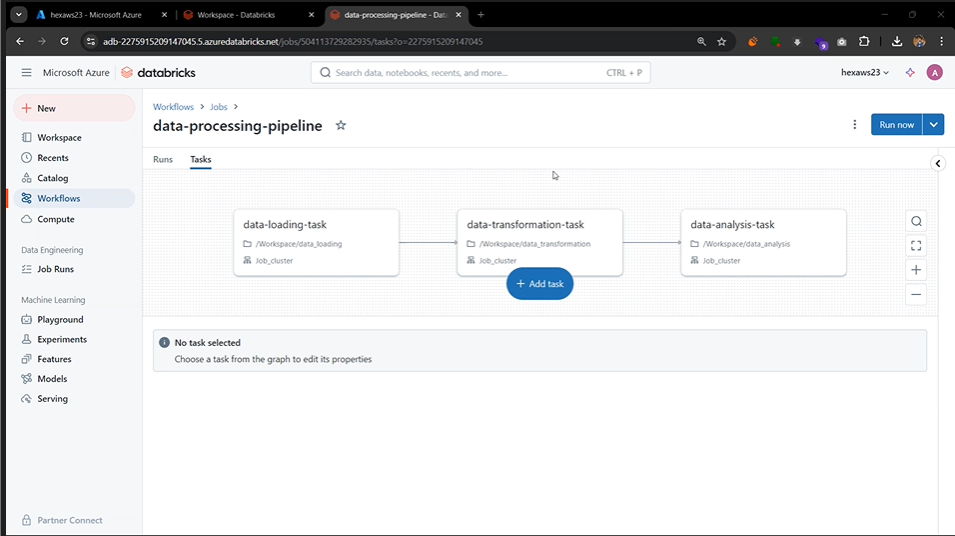


We can select what we want to link like notebook,script etc…,

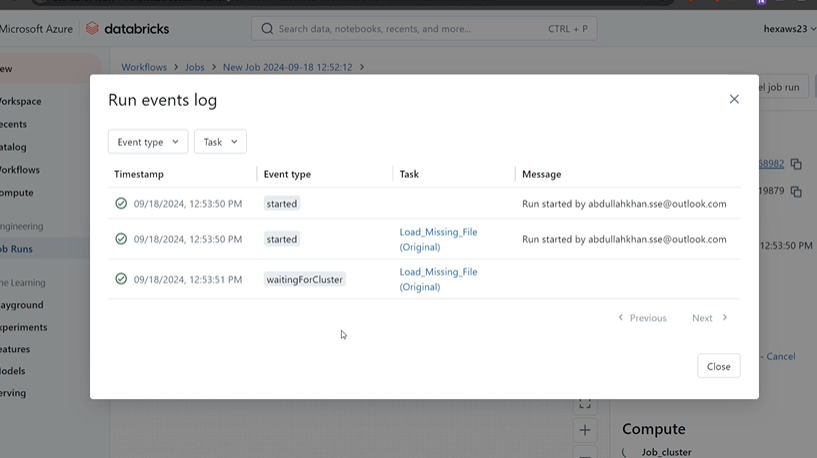


Select first task 🡪Run now

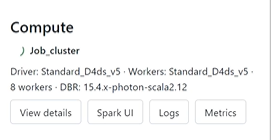




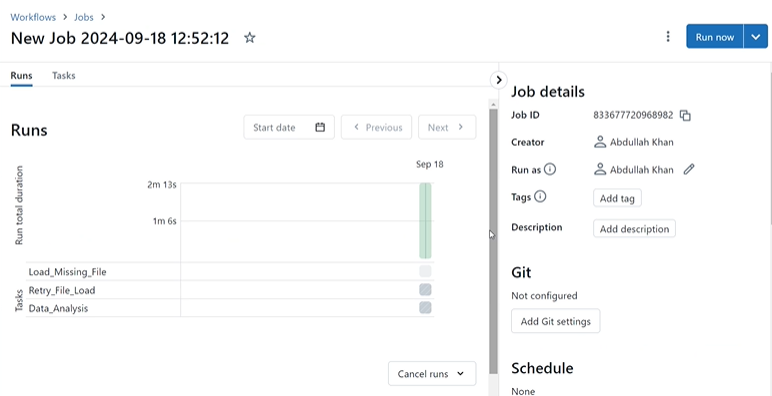
We can also check the log details too



Once it’s completed the jobs are created



Output:



Azure quota exception

Error occurred because of limit

Available limit : 20 , required limit : 32

If you want you can proceed with fixing it but it takes some available credits amount

