Custom Training

Day 19

**Databricks**

It is a managed service. Provides platform for spark. Acts as a spark engine.

Whenever we request any cluster that is provided by Azure.

Catalog: Data where we can view the created tables.

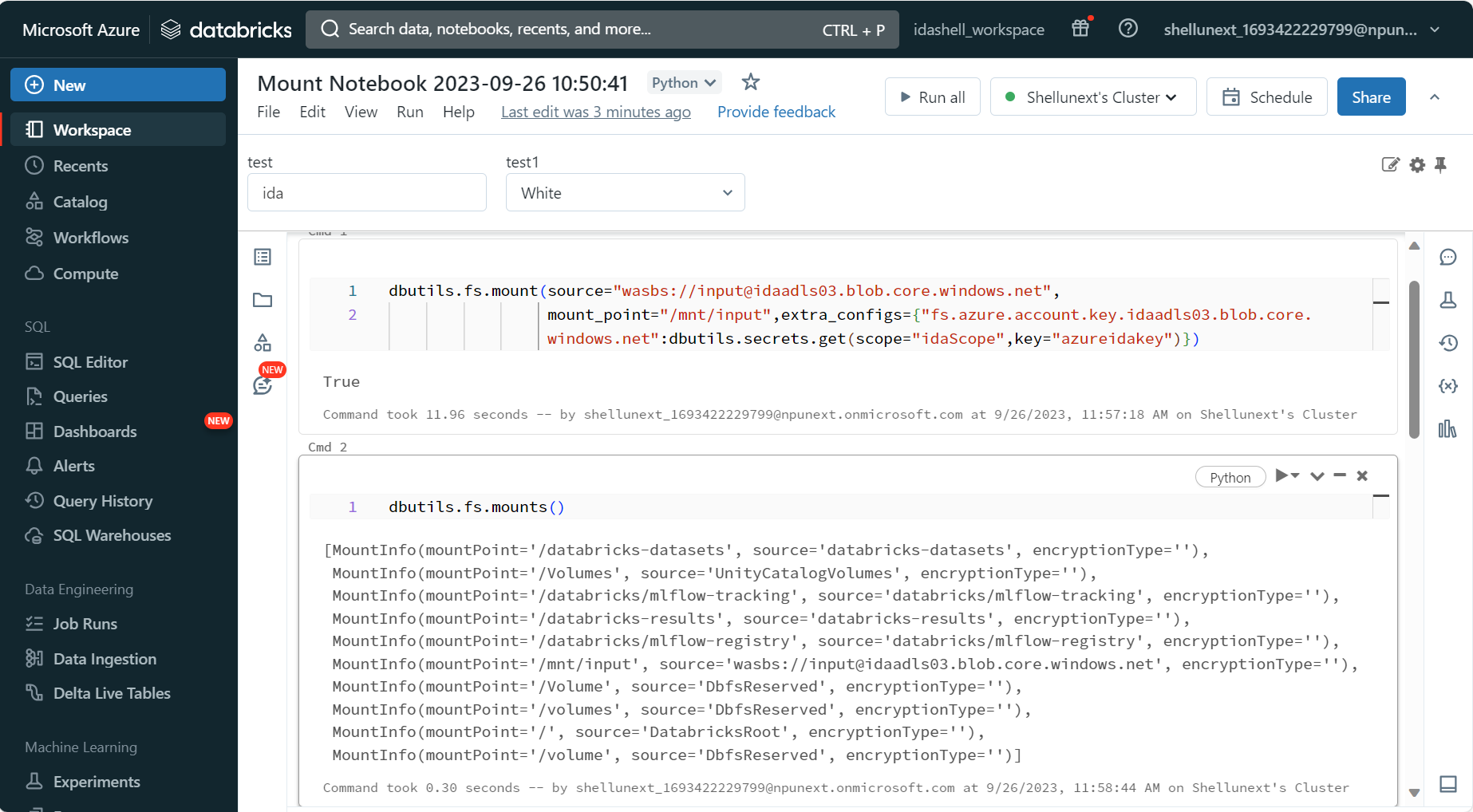
Cluster creation: VMs get created at the backend.

ALL-PURPOSE COMPUTE: To run a notebook we need this. But for scheduling purposes, it needs to run all the time, which is not recommended.

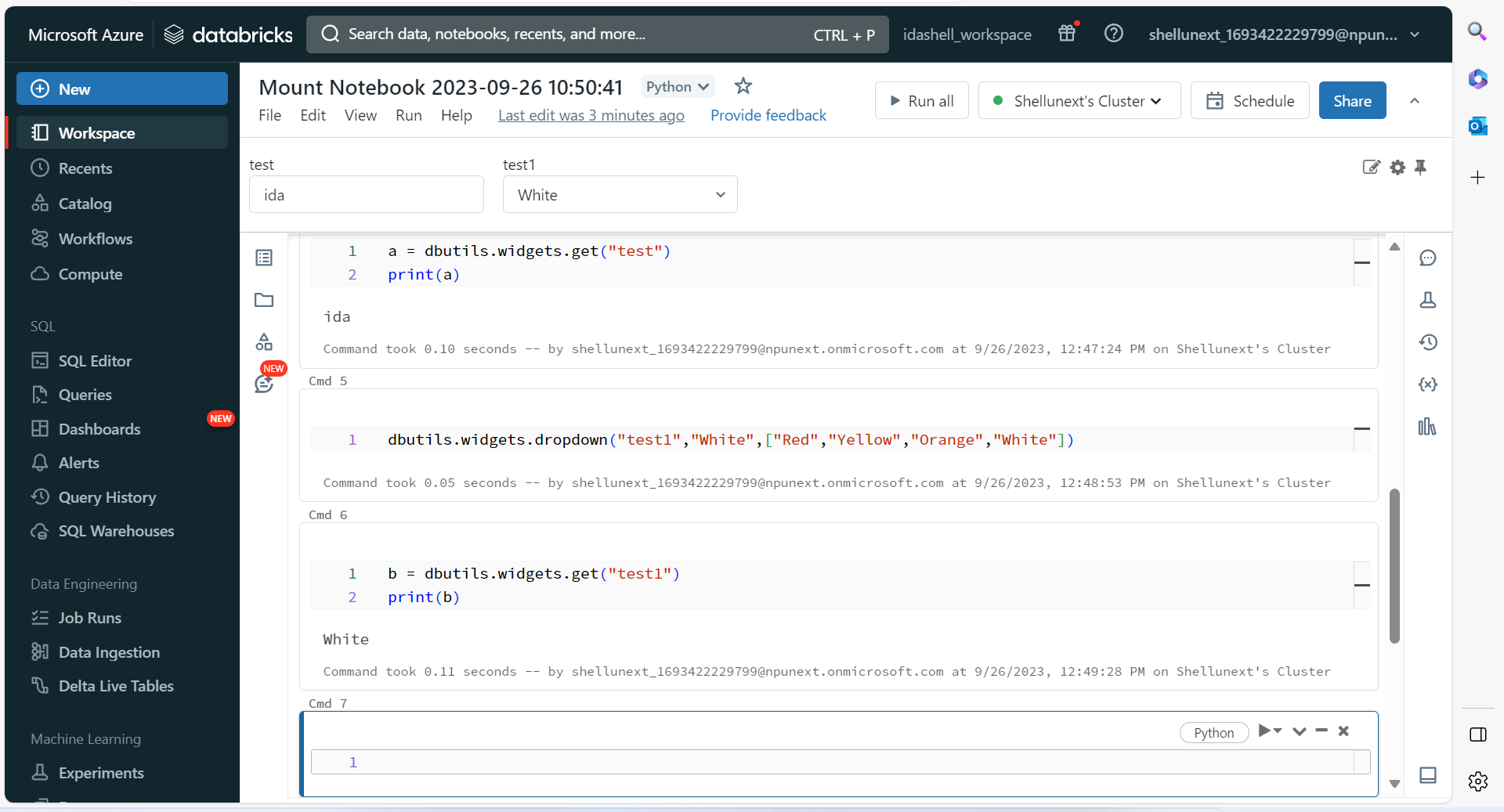
JOB COMPUTE: The cluster starts only at the schedule time and terminates even if the job fails.

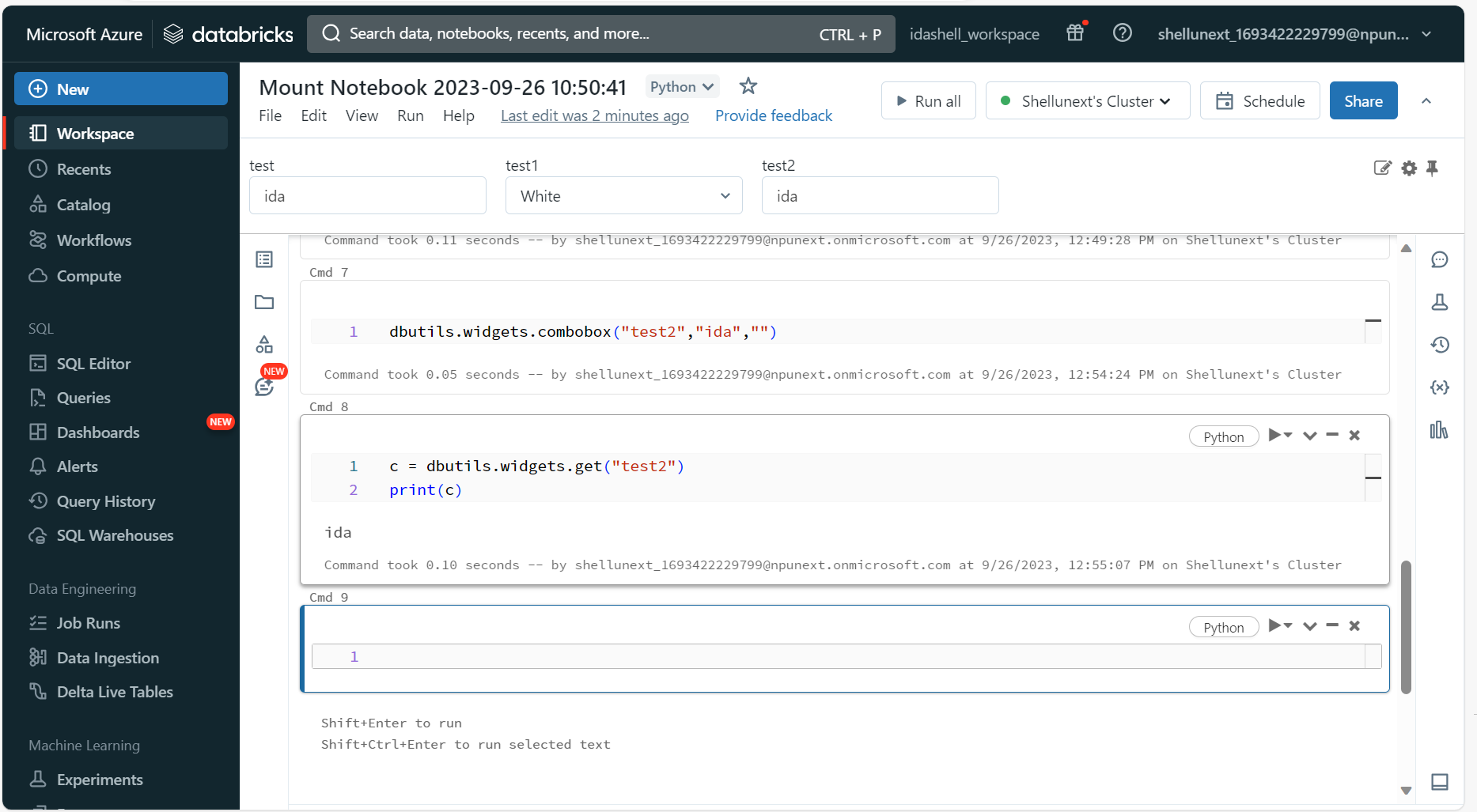
POOLS: We create VMs and then it reduces the cluster start time.

Mount: To access the data stored in our storage account/adls gen2 we need to mount it in our databricks. The ideal way is to do it through Azure Key Vault.

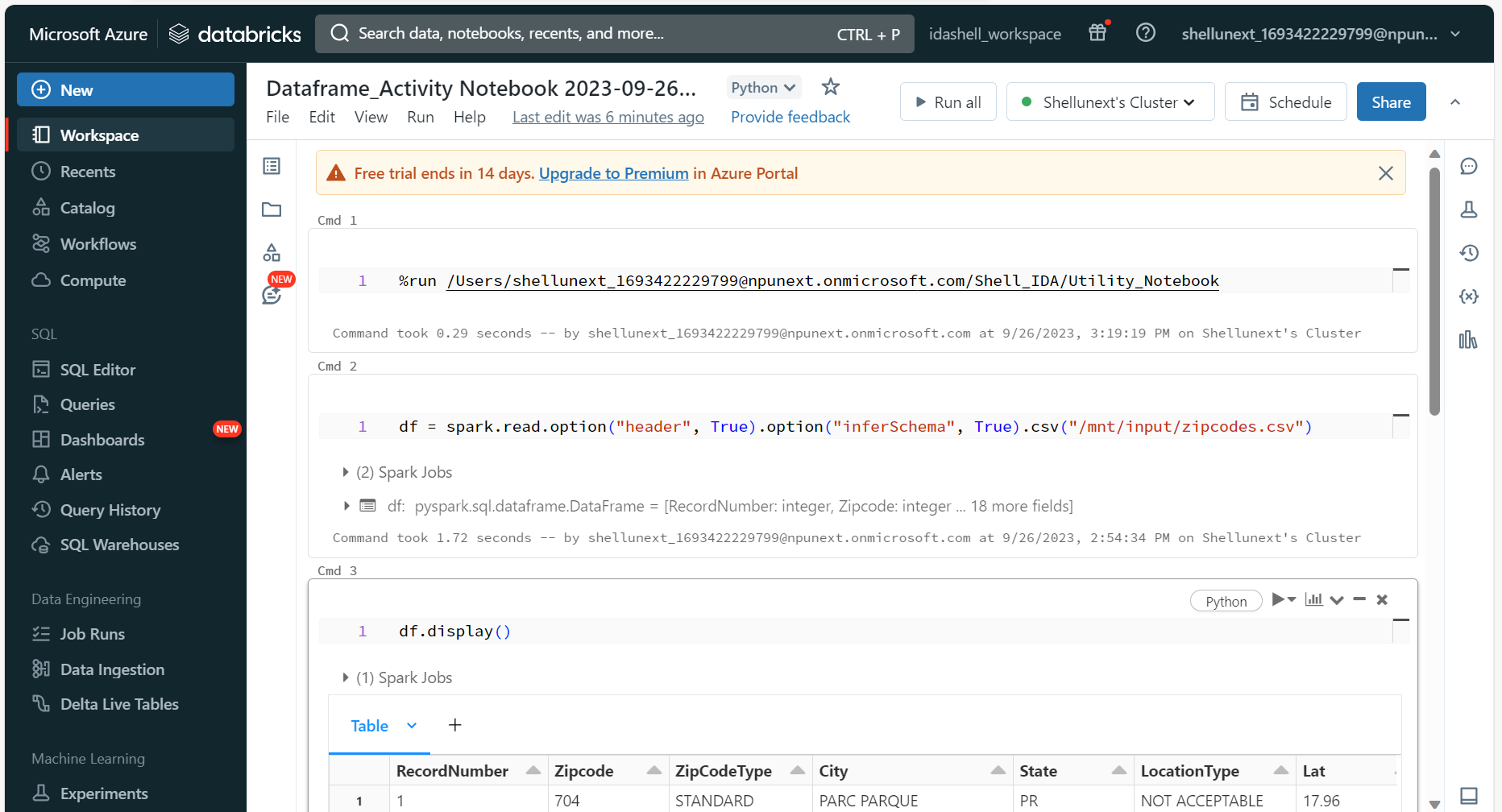


Widgets: To create parameters in the notebook.

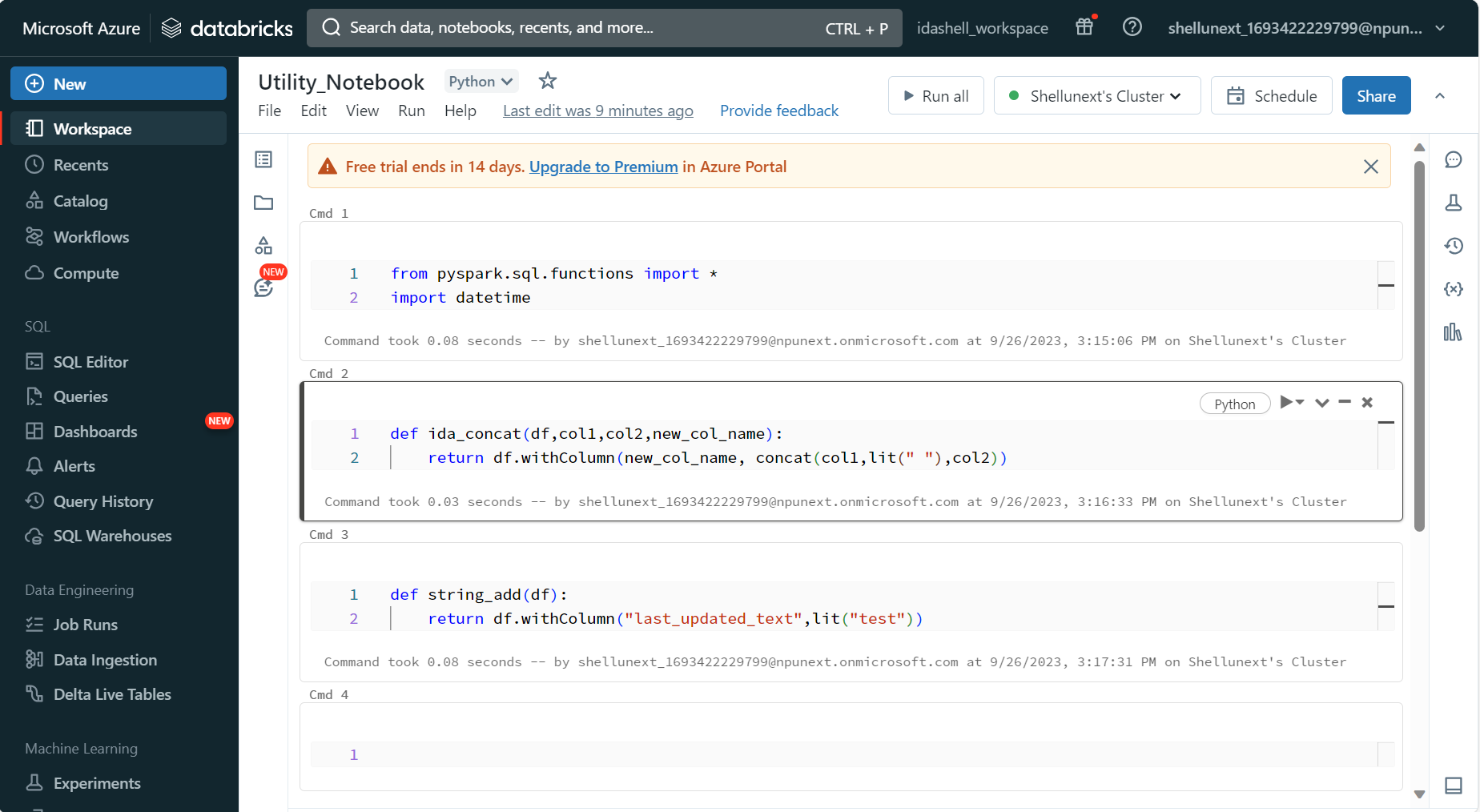


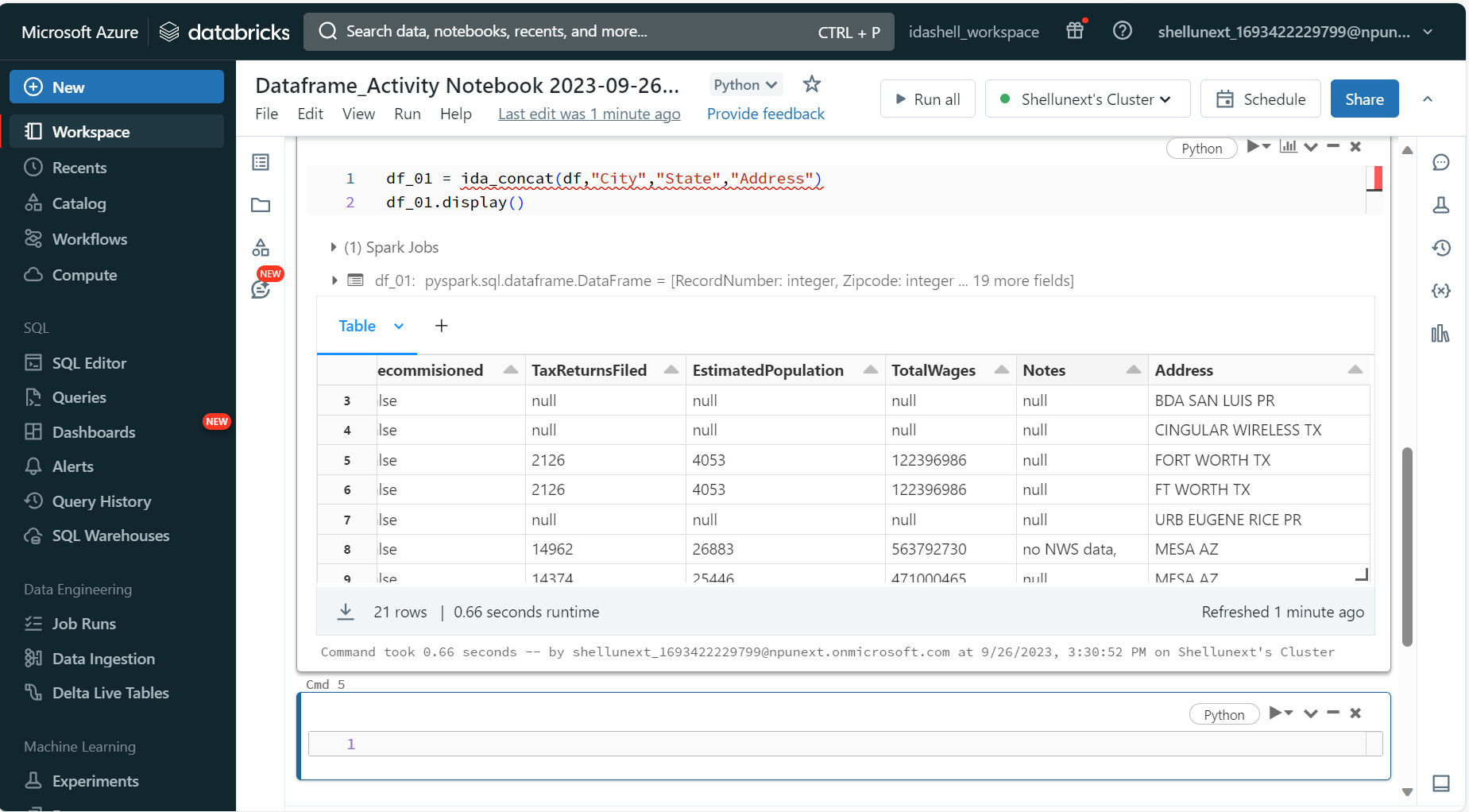


Dataframe Activity:



Utility Notebook:





Create python functions and register it for spark optimization, without registering also it will work but optimization won’t be there:



Structured Streaming:

Used to process the data immediately after it gets pushed into the source location.

