Custom Learnings

Day 19

Azure 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.

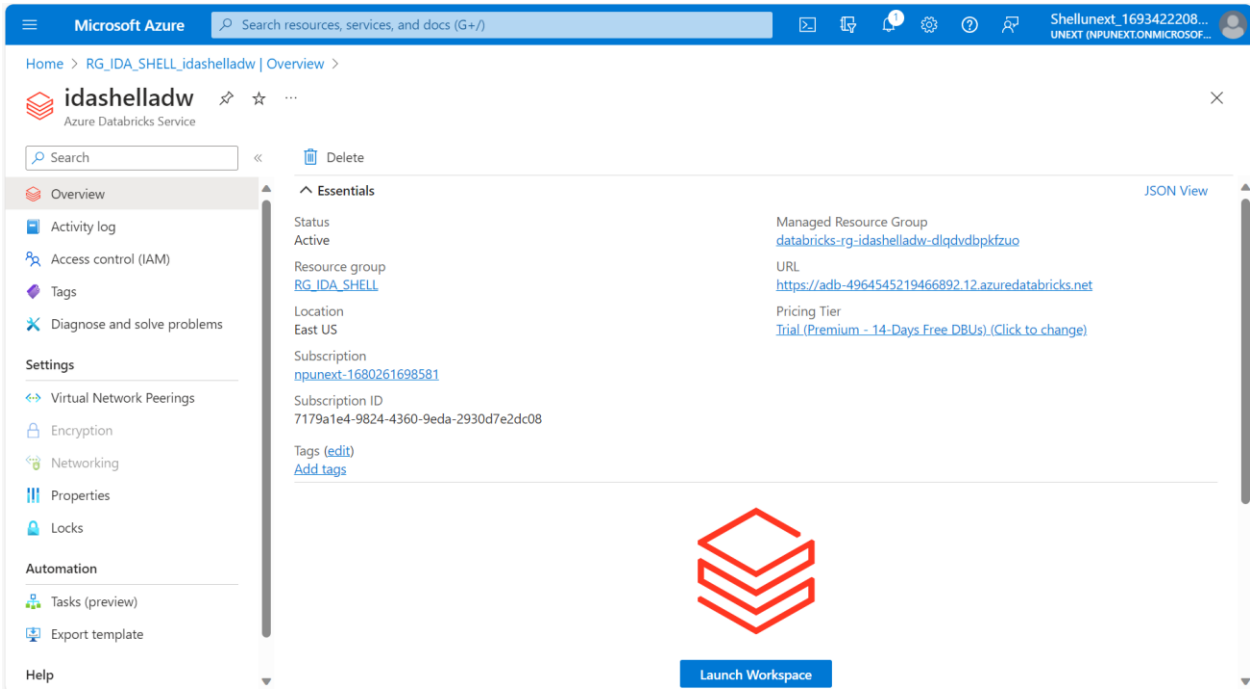
ALL-PURPOSE COMPUTE: To run a notebook we need this. But for scheduling purposes, it needs to run all the time, which is not recommended.

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

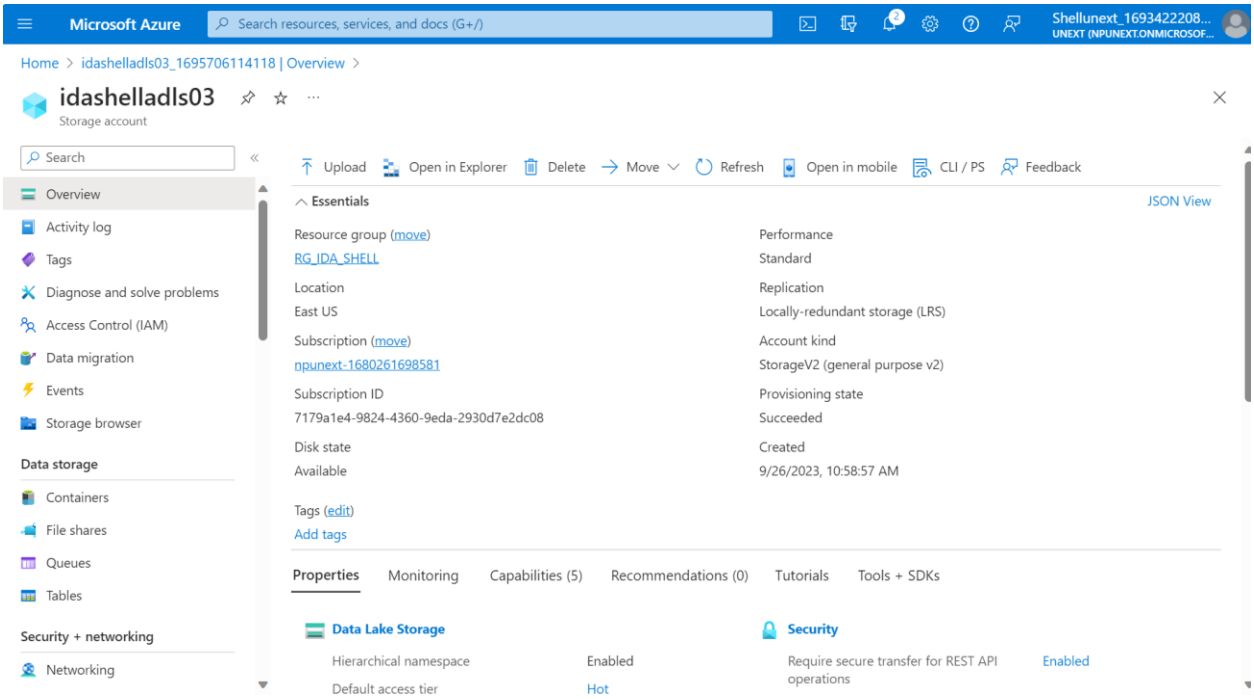
POOLS: We create VMs and then it reduces the cluster start time.

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.

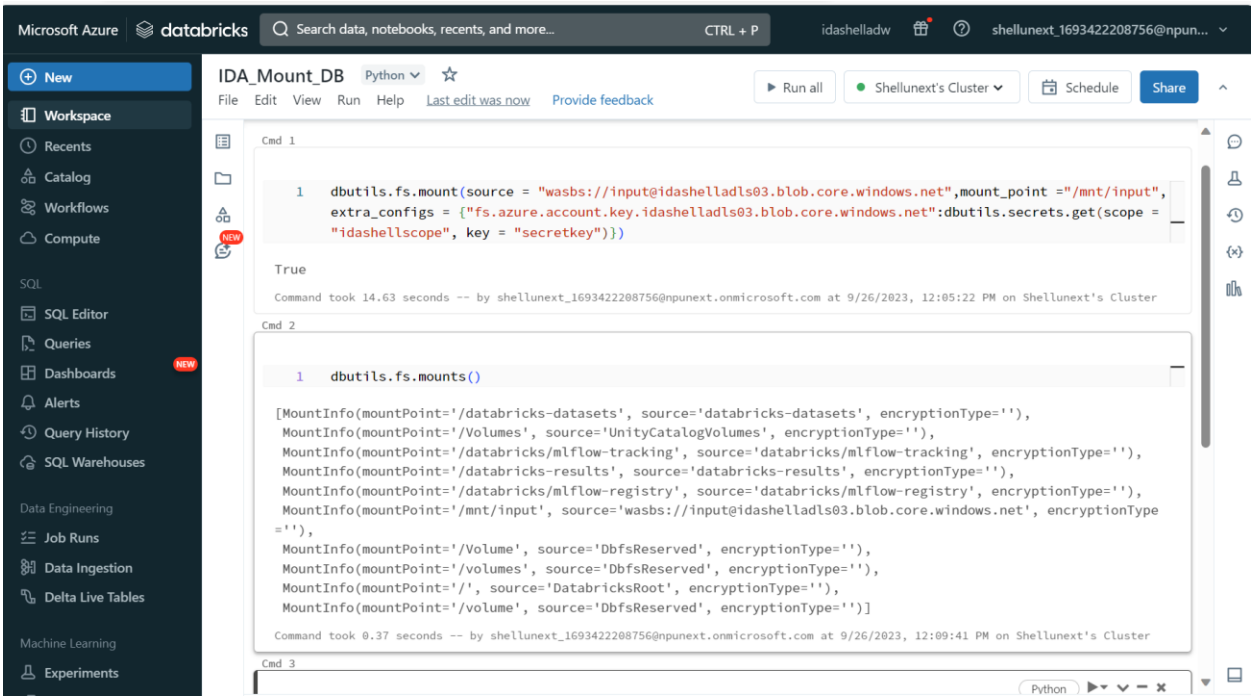
Creation of azure databricks:



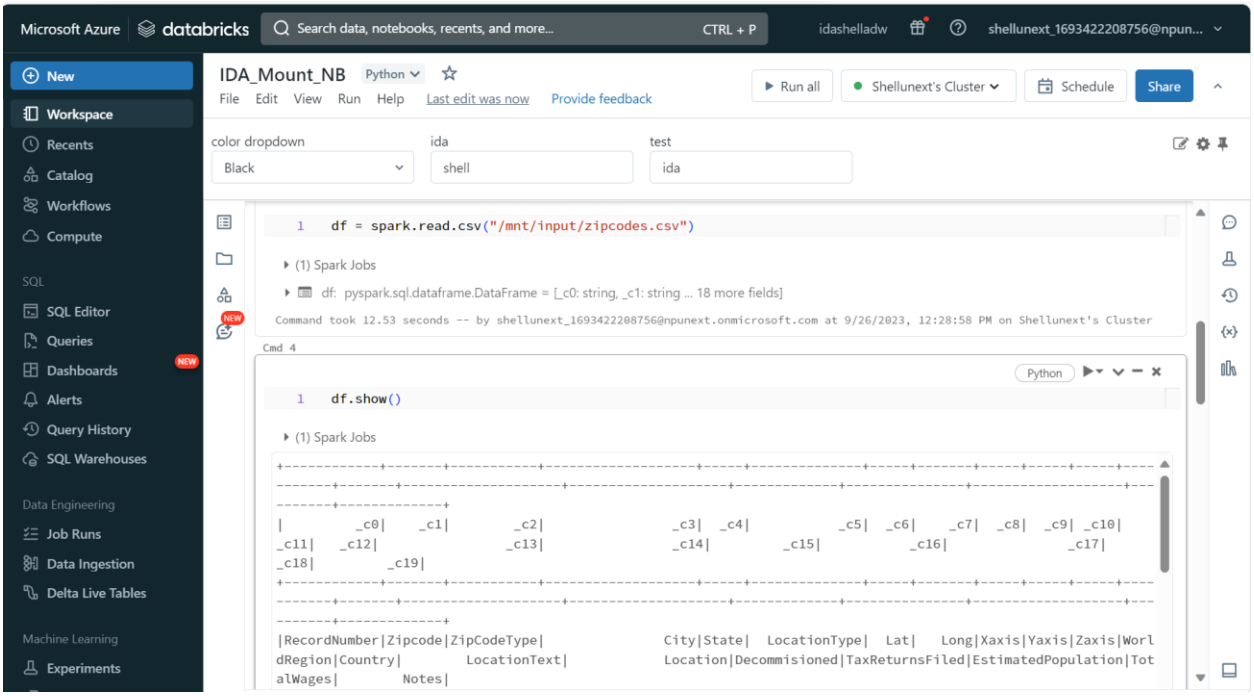
Creation of storage account:



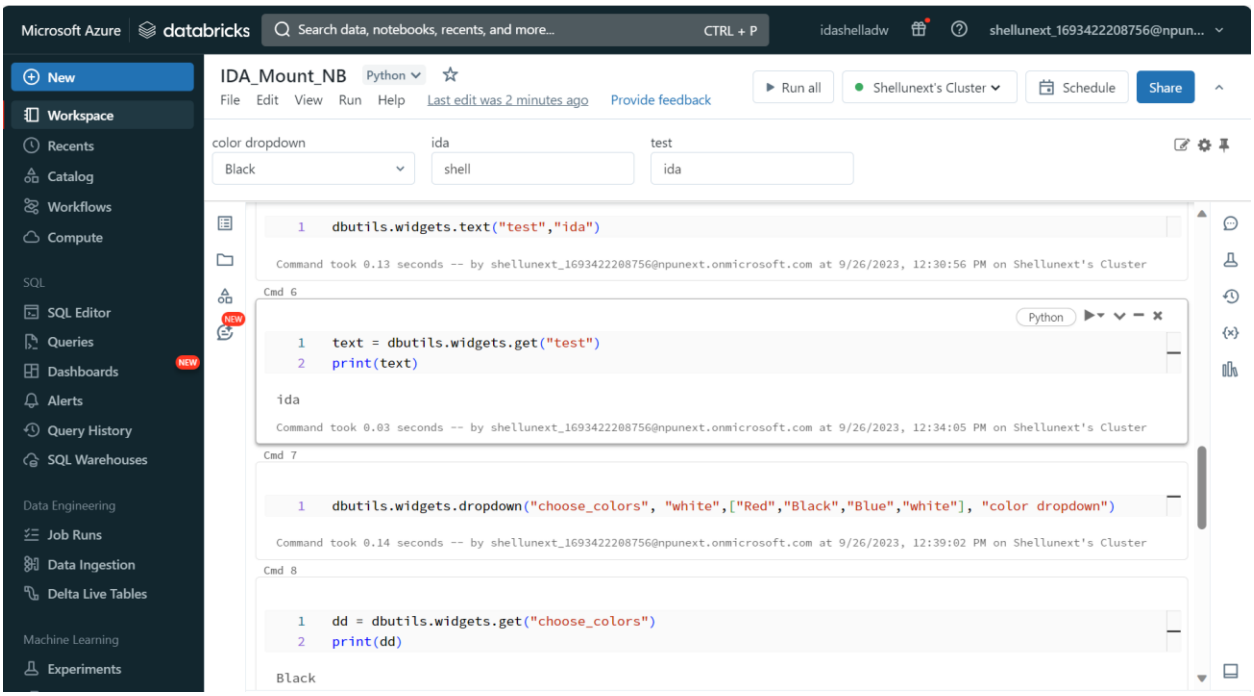
Mounting:

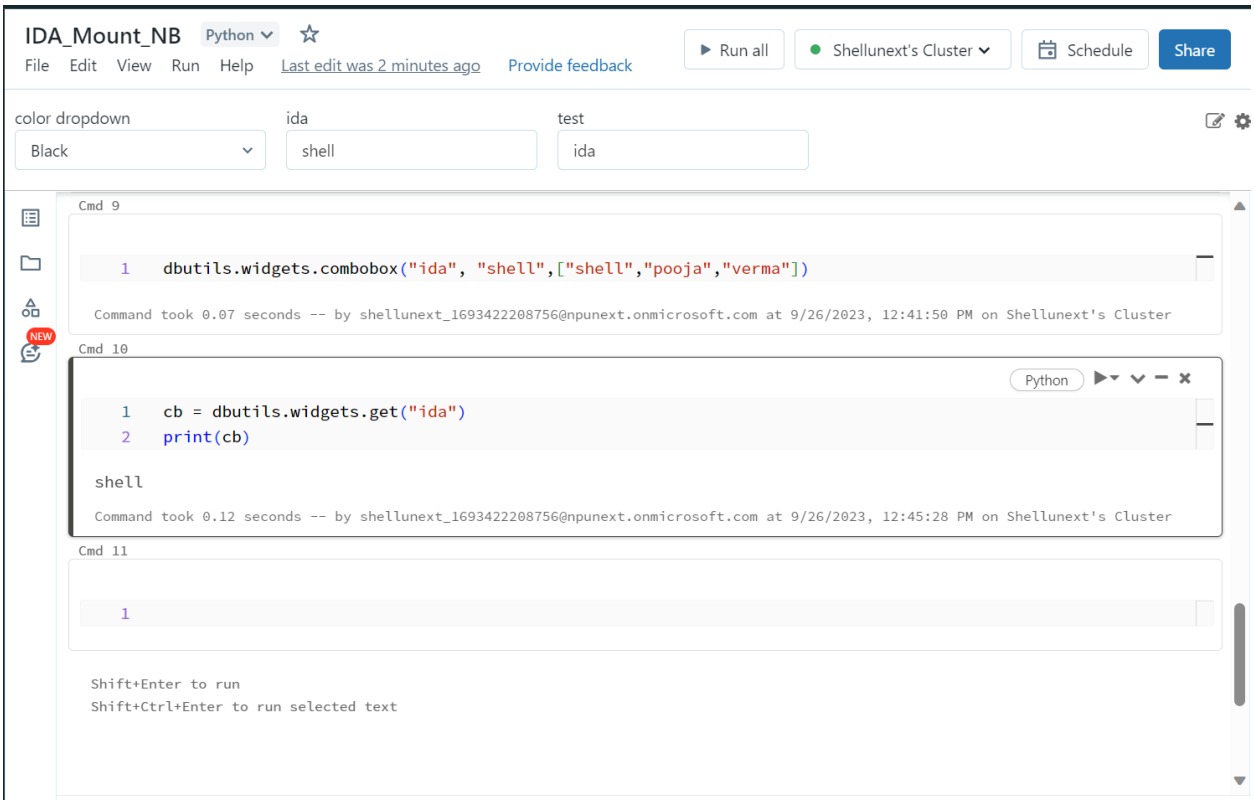


Dataframe

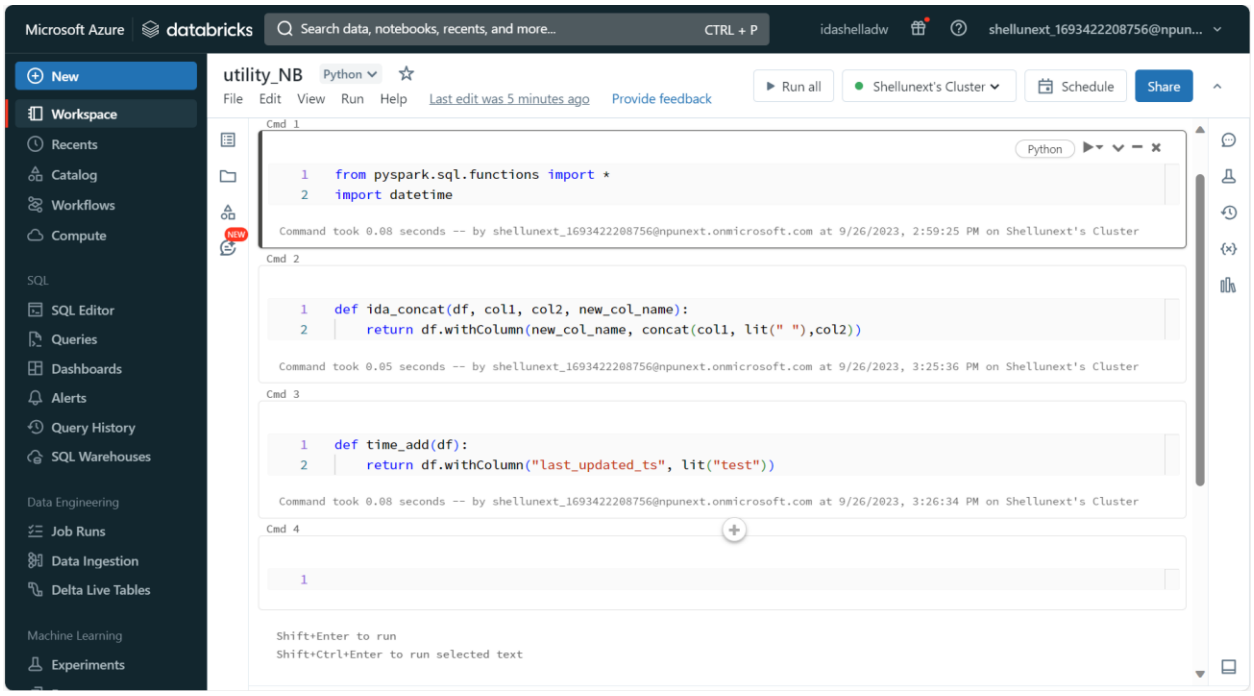


Widgets: To create parameters in the notebook.

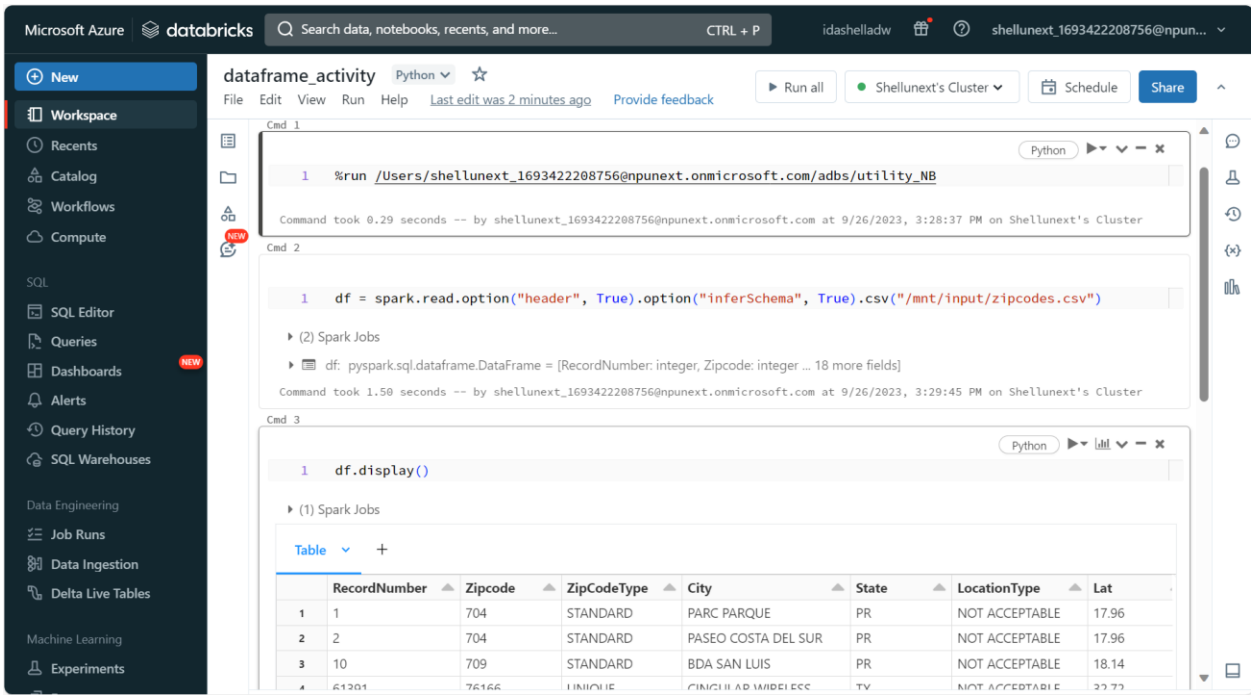




Utility NoteBook:

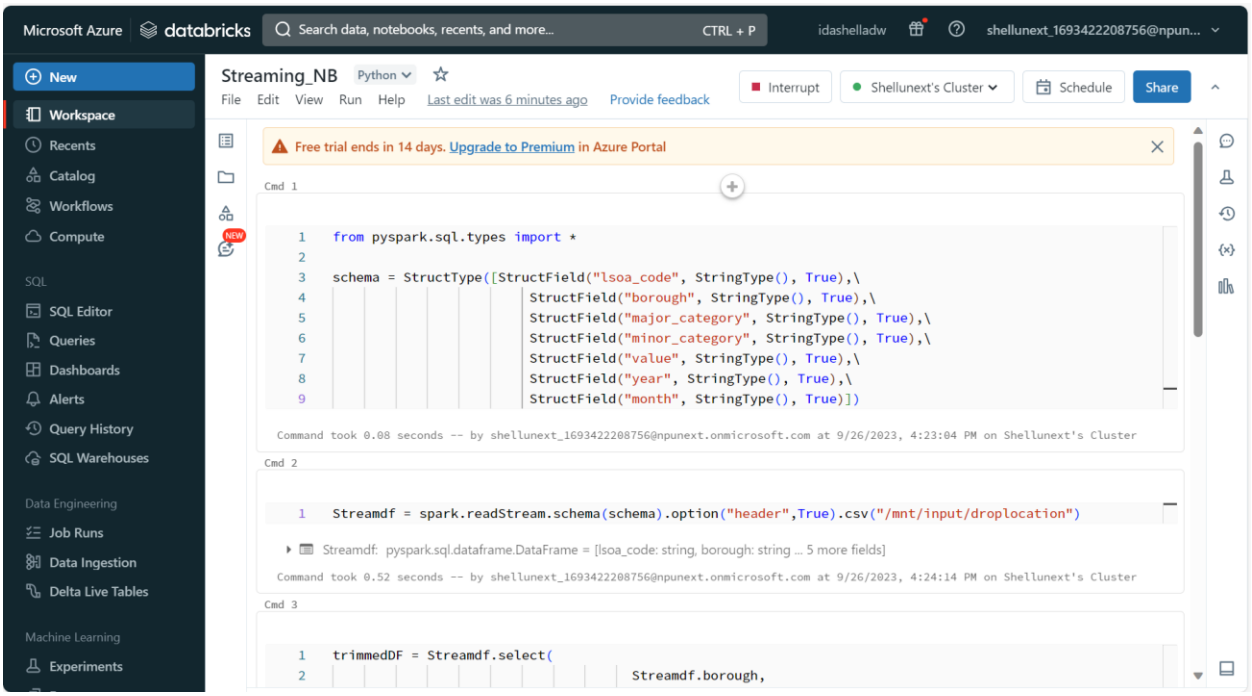


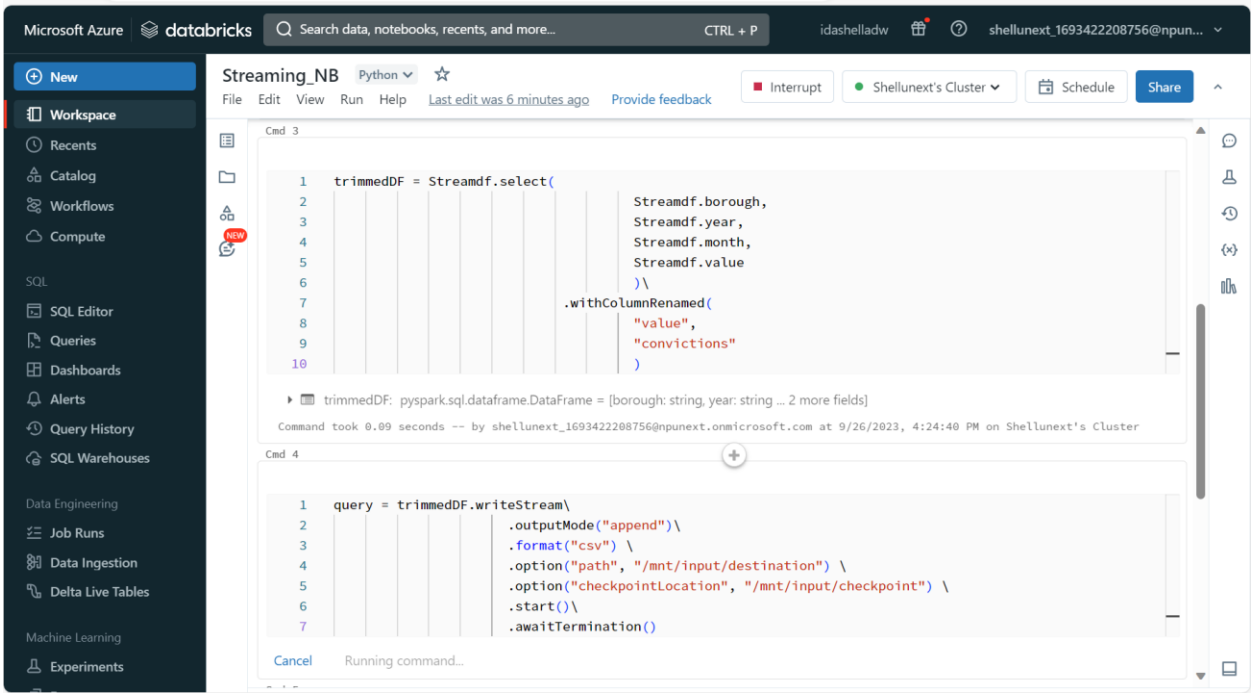
DataFrame Activity:

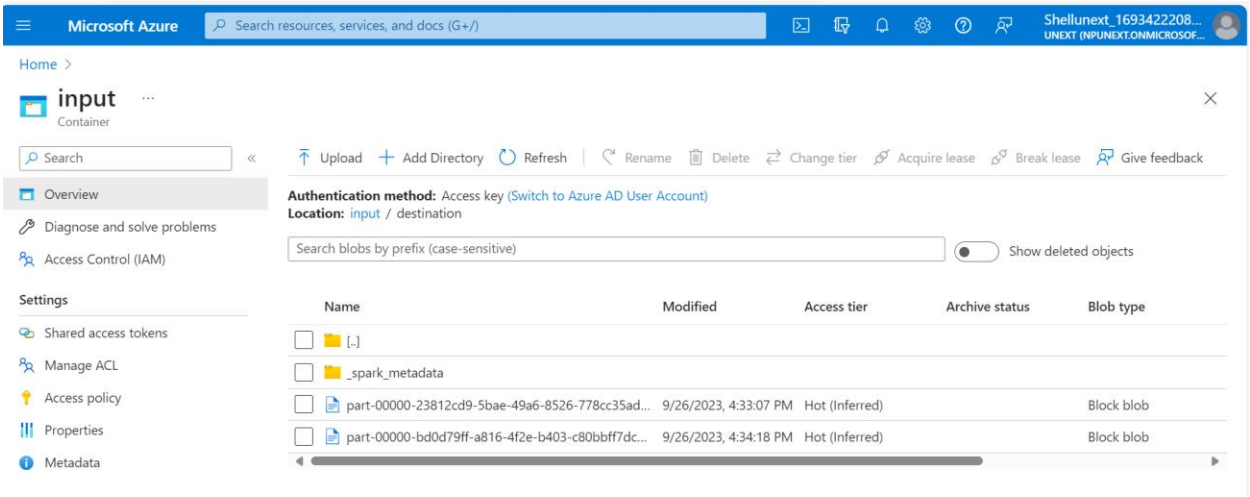


Structured Streaming:

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




Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
idashelladw
shellunext_1693422208756@npun...
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Experiments
Streaming_NB
Python
File Edit View Run Help
Last edit was 6 minutes ago
Provide feedback
Interrupt
Shellunext's Cluster
Schedule
Share
Cmd 3
trimmedDF = Streamdf.select(
Streamdf.borough,
Streamdf.year,
Streamdf.month,
Streamdf.value
)\
.withColumnRenamed(
"value",
"convictions"
)
trimmedDF: pyspark.sql.dataframe.DataFrame = [borough: string, year: string ... 2 more fields]
Command took 0.09 seconds -- by shellunext_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 4:24:40 PM on Shellunext's Cluster
Cmd 4
query = trimmedDF.writeStream\
.outputMode("append")\
.format("csv") \
.option("path", "/mnt/input/destination") \
.option("checkpointLocation", "/mnt/input/checkpoint") \
.start()\
.awaitTermination()
Cancel
Running command...


Microsoft Azure
Search resources, services, and docs (G+/)
Shellunext_1693422208...
UNEXT (NPUNEXT.ONMICROSOF...
Home
input
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Authentication method: Access key (Switch to Azure AD User Account)
Location: input / destination
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
[..]
_spark_metadata
part-00000-23812cd9-5bae-49a6-8526-778cc35ad...
9/26/2023, 4:33:07 PM
Hot (Inferred)
Block blob
part-00000-bd0d79ff-a816-4f2e-b403-c80bbff7dc...
9/26/2023, 4:34:18 PM
Hot (Inferred)
Block blob