

## 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.

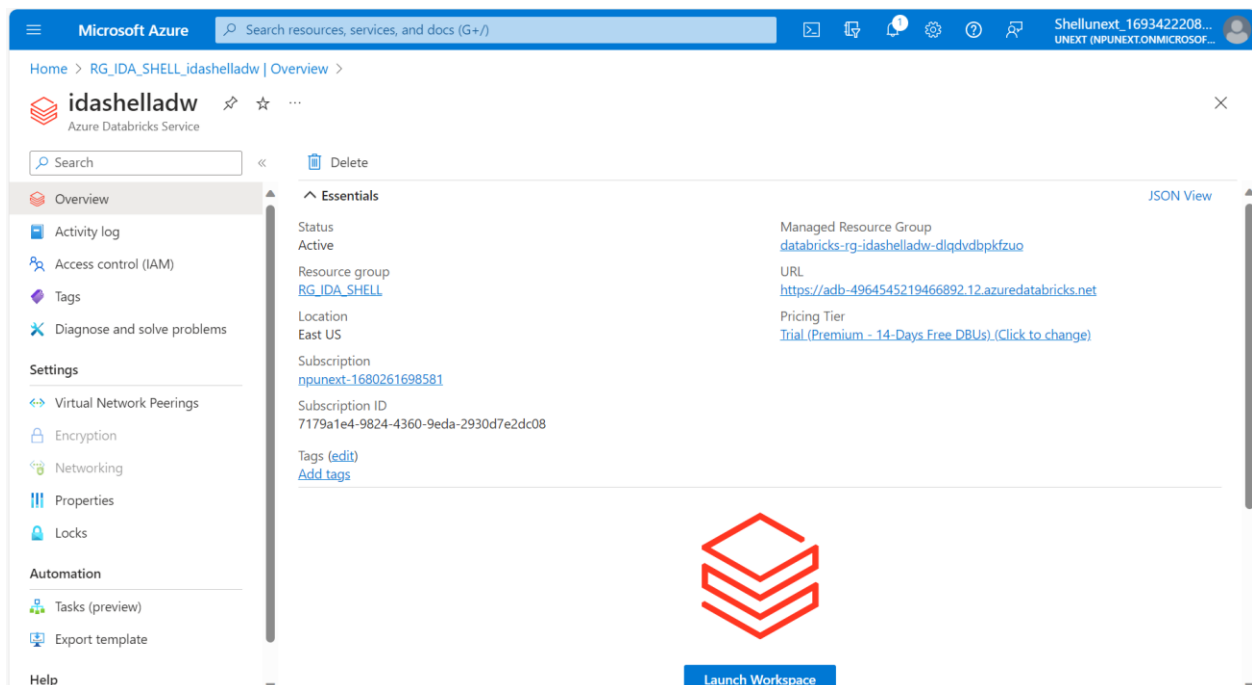
ALL-PURPOSE COMPUTE: To run a notebook we need this. But for scheduling purposes, it needs to run all the time, which is not recommended.

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:

The screenshot shows the Microsoft Azure portal interface for a storage account named **idashelladls03**. The left sidebar contains navigation options like Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, Events, Storage browser, Data storage, Containers, File shares, Queues, Tables, Security + networking, and Networking. The main content area displays the account's Essentials, Properties, Monitoring, Capabilities (5), Recommendations (0), Tutorials, and Tools + SDKs. The Essentials section lists details such as Resource group (RG\_IDA\_SHELL), Location (East US), Subscription (npunext-1680261698581), Subscription ID (7179a1e4-9824-4360-9eda-2930d7e2dc08), Disk state (Available), Performance (Standard), Replication (Locally-redundant storage (LRS)), Account kind (StorageV2 (general purpose v2)), Provisioning state (Succeeded), and Created date (9/26/2023, 10:58:57 AM). The Properties section shows Data Lake Storage settings (Hierarchical namespace: Enabled, Default access tier: Hot) and Security settings (Require secure transfer for REST API operations: Enabled).

## Mounting:

The screenshot shows the Databricks IDE interface. The left sidebar contains navigation options like 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, and Experiments. The main content area displays a Python script in a workspace named **IDA\_Mount\_DB**. The script defines a function `dbutils.fs.mount` to mount an Azure Storage account. The command is executed, and the output shows the successful mounting of the storage account. The command took 14.63 seconds to execute.

```
1 dbutils.fs.mount(source = "wasbs://input@idashelladls03.blob.core.windows.net", mount_point = "/mnt/input",  
extra_configs = {"fs.azure.account.key.idashelladls03.blob.core.windows.net": dbutils.secrets.get(scope =  
"idashellscope", key = "secretkey")})
```

True

Command took 14.63 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 12:05:22 PM on Shellunext's Cluster

```
1 dbutils.fs.mounts()
```

[MountInfo(mountPoint='/databricks-datasets', source='databricks-datasets', encryptionType=''),  
MountInfo(mountPoint='/Volumes', source='UnityCatalogVolumes', encryptionType=''),  
MountInfo(mountPoint='/databricks/mlflow-tracking', source='databricks/mlflow-tracking', encryptionType=''),  
MountInfo(mountPoint='/databricks-results', source='databricks-results', encryptionType=''),  
MountInfo(mountPoint='/databricks/mlflow-registry', source='databricks/mlflow-registry', encryptionType=''),  
MountInfo(mountPoint='/mnt/input', source='wasbs://input@idashelladls03.blob.core.windows.net', encryptionType  
=''),  
MountInfo(mountPoint='/Volume', source='DbfsReserved', encryptionType=''),  
MountInfo(mountPoint='/volumes', source='DbfsReserved', encryptionType=''),  
MountInfo(mountPoint='/', source='DatabricksRoot', encryptionType=''),  
MountInfo(mountPoint='/volume', source='DbfsReserved', encryptionType='')]

Command took 0.37 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 12:09:41 PM on Shellunext's Cluster

## Dataframe

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

IDA\_Mount\_NB Python File Edit View Run Help Last edit was now Provide feedback Run all Shellunext's Cluster Schedule Share

color dropdown ida test  
Black shell ida

```
1 df = spark.read.csv("/mnt/input/zipcodes.csv")
```

(1) Spark Jobs  
df: pyspark.sql.dataframe.DataFrame = [c0: string, c1: string ... 18 more fields]  
Command took 12.53 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 12:28:58 PM on Shellunext's Cluster

Cmd 4 Python

```
1 df.show()
```

(1) Spark Jobs

_c0	_c1	_c2	_c3	_c4	_c5	_c6	_c7	_c8	_c9	_c10	_c11	_c12	_c13	_c14	_c15	_c16	_c17
_c11	_c12	_c13	_c14	_c15	_c16	_c17											
_c18	_c19																

|RecordNumber|Zipcode|ZipCodeType|City|State|LocationType|Lat|Long|Xaxis|Yaxis|Zaxis|WorldRegion|Country|LocationText|Location|Decommissioned|TaxReturnsFiled|EstimatedPopulation|TotalWages|Notes|

Widgets: To create parameters in the notebook.

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

IDA\_Mount\_NB Python File Edit View Run Help Last edit was 2 minutes ago Provide feedback Run all Shellunext's Cluster Schedule Share

color dropdown ida test  
Black shell ida

```
1 dbutils.widgets.text("test", "ida")
```

Command took 0.13 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 12:30:56 PM on Shellunext's Cluster

Cmd 6 Python

```
1 text = dbutils.widgets.get("test")
2 print(text)
```

ida  
Command took 0.03 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 12:34:05 PM on Shellunext's Cluster

Cmd 7

```
1 dbutils.widgets.dropdown("choose_colors", "white", ["Red", "Black", "Blue", "white"], "color dropdown")
```

Command took 0.14 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 12:39:02 PM on Shellunext's Cluster

Cmd 8

```
1 dd = dbutils.widgets.get("choose_colors")
2 print(dd)
```

Black

IDA\_Mount\_NB Python ☆

File Edit View Run Help Last edit was 2 minutes ago Provide feedback Run all Shellunext's Cluster Schedule Share

color dropdown Black ida test

Cmd 9

```
1 dbutils.widgets.combobox("ida", "shell", ["shell", "pooja", "verma"])
```

Command took 0.07 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 12:41:50 PM on Shellunext's Cluster

Cmd 10

```
1 cb = dbutils.widgets.get("ida")
2 print(cb)
```

shell

Command took 0.12 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 12:45:28 PM on Shellunext's Cluster

Cmd 11

```
1
```

Shift+Enter to run  
Shift+Ctrl+Enter to run selected text

## Utility NoteBook:

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P idashelladv 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

utility\_NB Python ☆

File Edit View Run Help Last edit was 5 minutes ago Provide feedback Run all Shellunext's Cluster Schedule Share ^

Cmd 1

```
1 from pyspark.sql.functions import *
2 import datetime
```

Command took 0.08 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 2:59:25 PM on Shellunext's Cluster

Cmd 2

```
1 def ida_concat(df, col1, col2, new_col_name):
2     return df.withColumn(new_col_name, concat(col1, lit(" "), col2))
```

Command took 0.05 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 3:25:36 PM on Shellunext's Cluster

Cmd 3

```
1 def time_add(df):
2     return df.withColumn("last_updated_ts", lit("test"))
```

Command took 0.08 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 3:26:34 PM on Shellunext's Cluster

Cmd 4

```
1
```

Shift+Enter to run  
Shift+Ctrl+Enter to run selected text

## DataFrame Activity:

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

### dataframe\_activity Python

File Edit View Run Help Last edit was 2 minutes ago Provide feedback Run all Shellunext's Cluster Schedule Share

Cmd 1

```
%run /Users/shellunext_1693422208756@npunext.onmicrosoft.com/adbs/utility_NB
```

Command took 0.29 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 3:28:37 PM on Shellunext's Cluster

Cmd 2

```
1 df = spark.read.option("header", True).option("inferSchema", True).csv("/mnt/input/zipcodes.csv")
```

(2) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [RecordNumber: integer, Zipcode: integer ... 18 more fields]

Command took 1.50 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 3:29:45 PM on Shellunext's Cluster

Cmd 3

```
1 df.display()
```

(1) Spark Jobs

Table +

	RecordNumber	Zipcode	ZipCodeType	City	State	LocationType	Lat
1	1	704	STANDARD	PARC PARQUE	PR	NOT ACCEPTABLE	17.96
2	2	704	STANDARD	PASEO COSTA DEL SUR	PR	NOT ACCEPTABLE	17.96
3	10	709	STANDARD	BDA SAN LUIS	PR	NOT ACCEPTABLE	18.14
4	61201	76166	UNIQUE	CINQUE AD WIDEFESS	TX	NOT ACCEPTABLE	22.72

## 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

Free trial ends in 14 days. Upgrade to Premium in Azure Portal

Cmd 1

```
1 from pyspark.sql.types import *
2
3 schema = StructType([StructField("lsoa_code", StringType(), True),\
4                               StructField("borough", StringType(), True),\
5                               StructField("major_category", StringType(), True),\
6                               StructField("minor_category", StringType(), True),\
7                               StructField("value", StringType(), True),\
8                               StructField("year", StringType(), True),\
9                               StructField("month", StringType(), True)])
```

Command took 0.08 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 4:23:04 PM on Shellunext's Cluster

Cmd 2

```
1 Streamdf = spark.readStream.schema(schema).option("header", True).csv("/mnt/input/droplocation")
```

Streamdf: pyspark.sql.dataframe.DataFrame = [lsoa\_code: string, borough: string ... 5 more fields]

Command took 0.52 seconds -- by shellunext\_1693422208756@npunext.onmicrosoft.com at 9/26/2023, 4:24:14 PM on Shellunext's Cluster

Cmd 3

```
1 trimmedDF = Streamdf.select(
2
```

Streamdf.borough,

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

Streaming\_NB Python File Edit View Run Help Last edit was 6 minutes ago Provide feedback Interrupt Shellunext's Cluster Schedule Share

Cmd 3

```
1 trimmedDF = Streamdf.select(
2     Streamdf.borough,
3     Streamdf.year,
4     Streamdf.month,
5     Streamdf.value
6 )
7
8 .withColumnRenamed(
9     "value",
10    "convictions"
11 )
```

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

```
1 query = trimmedDF.writeStream\
2     .outputMode("append")\
3     .format("csv") \
4     .option("path", "/mnt/input/destination") \
5     .option("checkpointLocation", "/mnt/input/checkpoint") \
6     .start()\
7     .awaitTermination()
```

Cancel Running command...

Microsoft Azure Search resources, services, and docs (G+/J) Shellunext\_1693422208... UNEXT (NPUNEXT.ONMICROSOF...)

Home >

input Container

Search Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease Give feedback

Overview Diagnose and solve problems Access Control (IAM)

Settings Shared access tokens Manage ACL Access policy Properties Metadata

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
<input type="checkbox"/>	[-]				
<input type="checkbox"/>	[-]				
<input type="checkbox"/>	part-00000-23812cd9-5bae-49a6-8526-778cc35ad...	9/26/2023, 4:33:07 PM	Hot (Inferred)		Block blob
<input type="checkbox"/>	part-00000-bd0d79ff-a816-4f2e-b403-c80bbff7dc...	9/26/2023, 4:34:18 PM	Hot (Inferred)		Block blob