

Coding Challenge | Azure Databricks

Divya Sree Murali

10/12/2024

a. Create a cluster & Attach the notebook to the cluster and run all commands in the notebook & creates a DataFrame from a Databricks dataset & Create a Visualizations in Databricks notebooks

1.Data Visualization

1.1.Running sql command and viewing visualization for the results



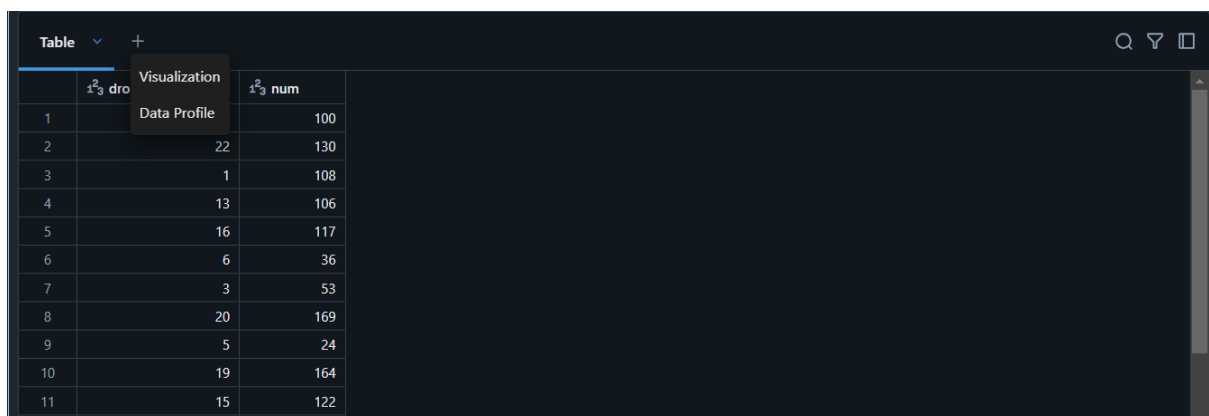
```
%sql
USE CATALOG SAMPLES;
SELECT
  hour(tpep_dropoff_datetime) as drop1,
  COUNT(*) AS new
FROM samples.nyctaxi.trips
where pickup_zip in ('10001','10003','10004')
group by 1
```

(2) Spark Jobs

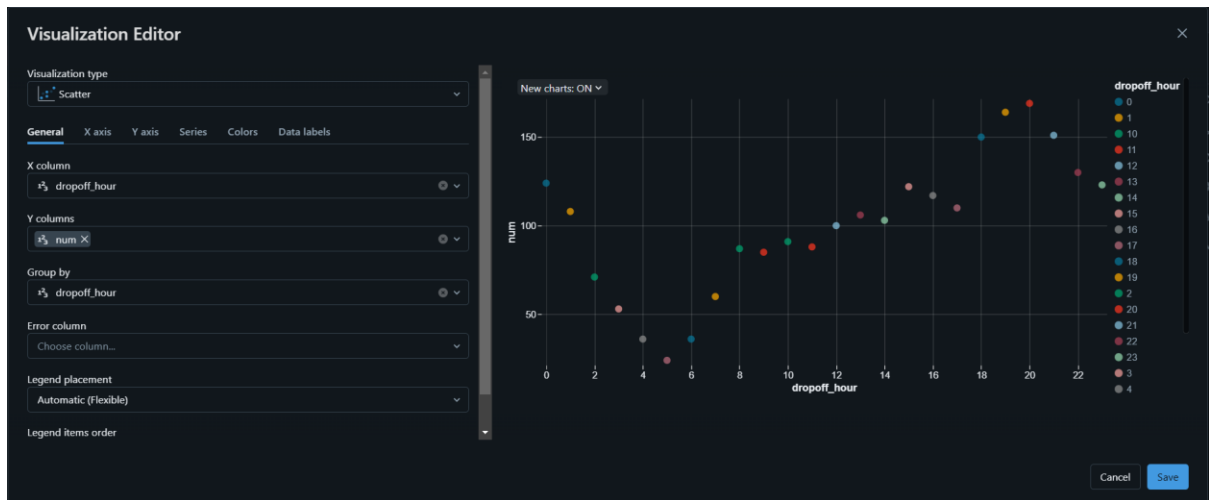
_sqlidf: pyspark.sql.dataframe.DataFrame = [drop1: integer, new: long]

Table Visualization 1 Visualization 2 Visualization 3 +

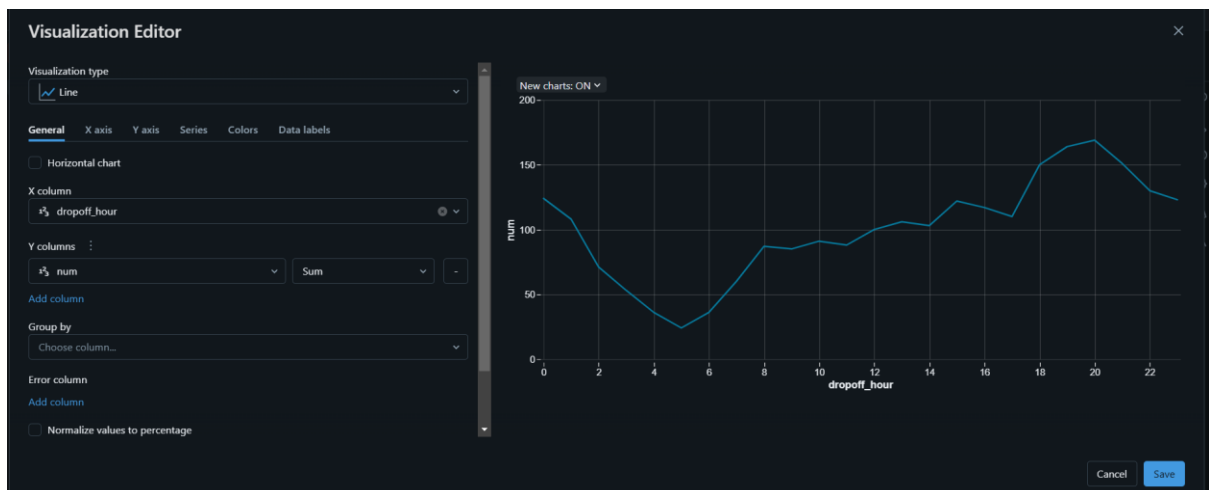
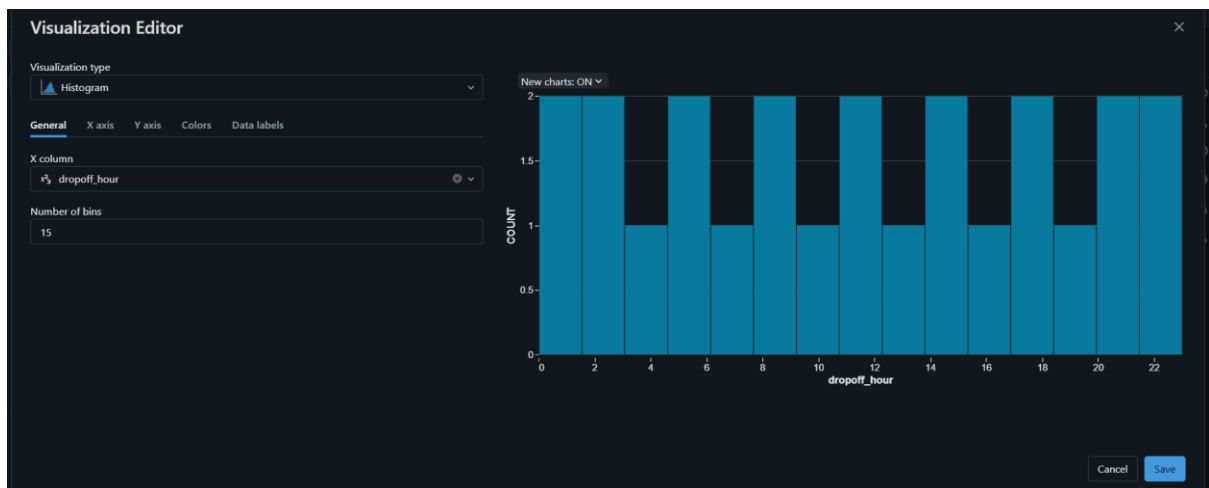
Select visualization to get the editor board



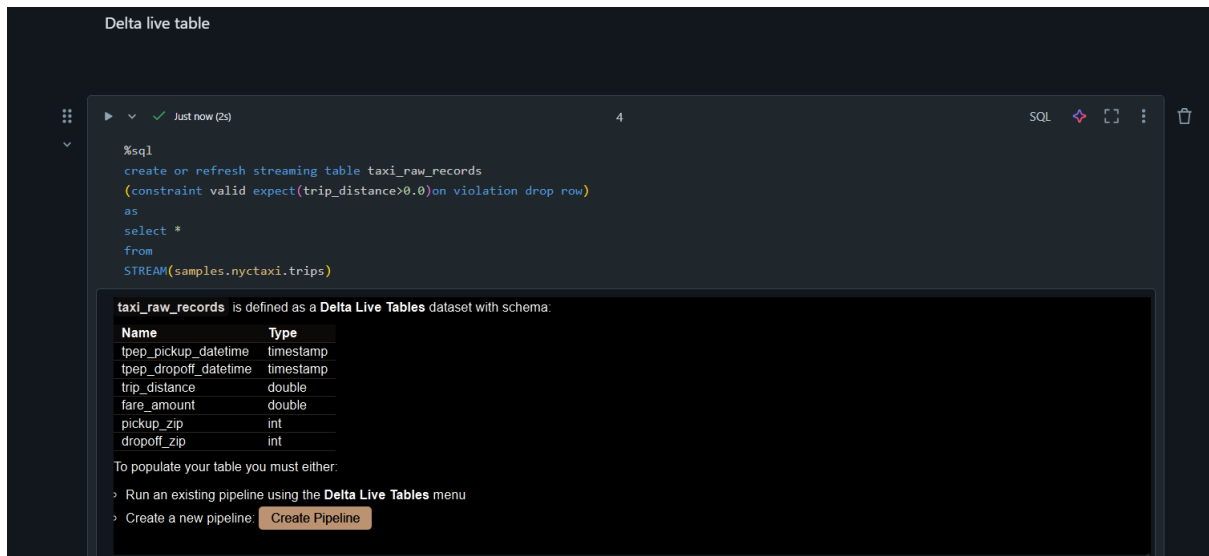
	drop1	new
1		100
2	22	130
3	1	108
4	13	106
5	16	117
6	6	36
7	3	53
8	20	169
9	5	24
10	19	164
11	15	122



Change the visualization type and the dashboards can be saved for future requirements.

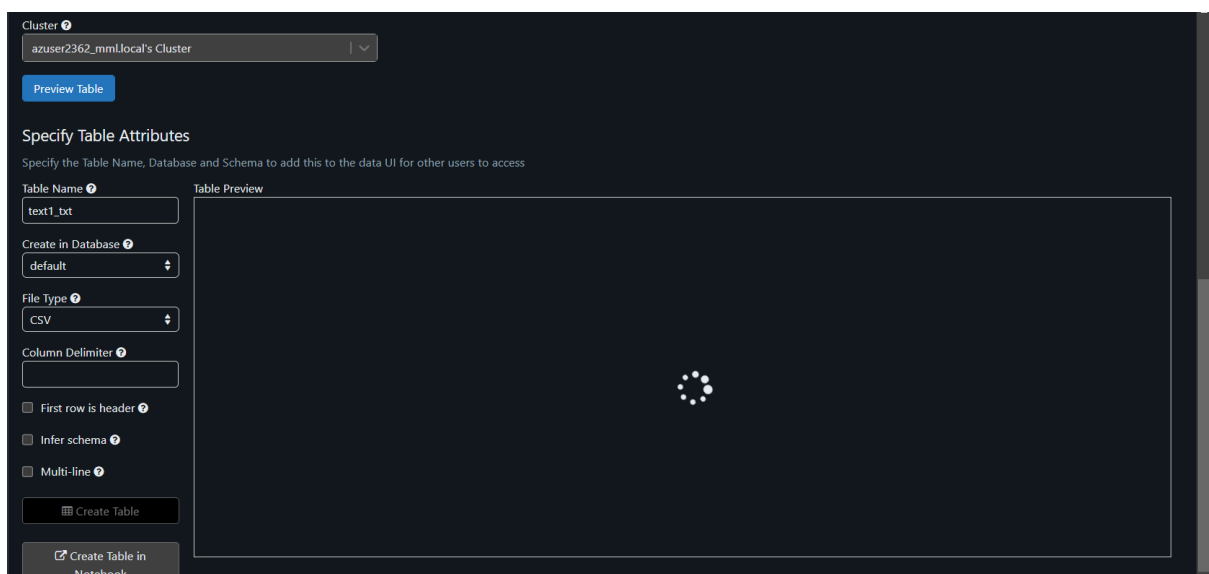
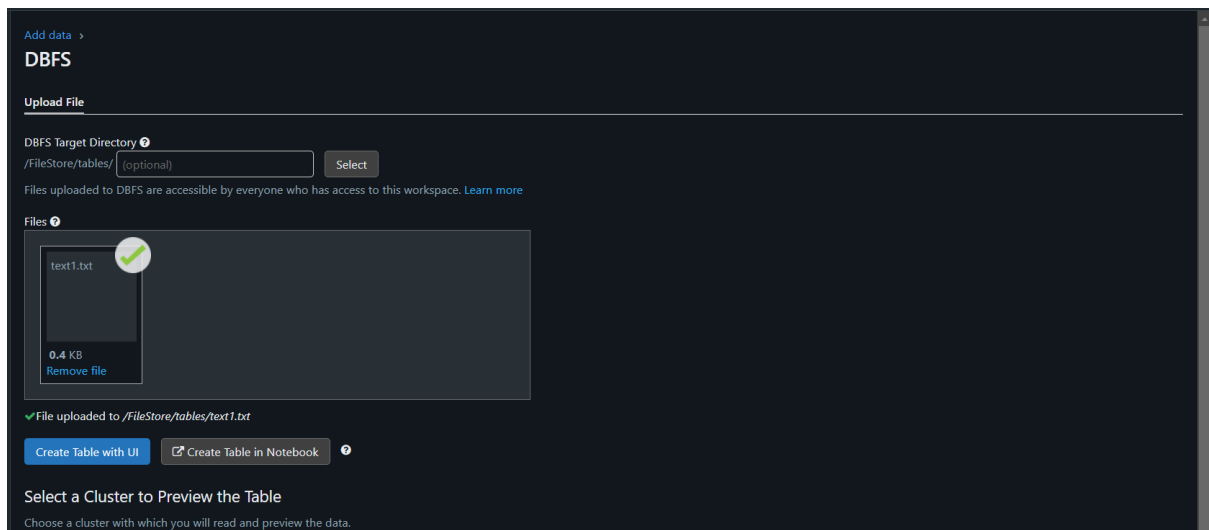


2.Creating Delta Live tables



3. Writing into delta tables

3.1. Uploading sample data into catalog



3.2. Reading the csv file and writing it in delta format

Writing data into delta tables

```
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName("Delta").getOrCreate()
```

```
s1=spark.read.option("inferSchema",True).option("header",True).csv("/FileStore/tables/new1.txt")
s1.write.format("delta").mode("overwrite").save("/FileStore/tables/delta_train/")
```

(8) Spark Jobs

s1: pyspark.sql.dataframe.DataFrame = [RowNumber: integer, CustomerId: integer ... 11 more fields]

Displaying the log details of the above code

In 1 minute (1s)

```
display(dbutils.fs.ls("/FileStore/tables/delta_train/"))
```

(2) Spark Jobs

	path	name
1	dbfs:/FileStore/tables/delta_train/_delta_log/	_delta_log/
2	dbfs:/FileStore/tables/delta_train/part-00000-9612b4b2-1b8d-41a2-826e-10dbcd6ab9cd-c000.snappy.parquet	part-00000-9612b4b2-1b8d-41a2-826e-10dbcd6ab9cd-c000.snappy.parquet

2 rows | 0.87 seconds runtime

Refreshed now

3.4. Reading content from delta format and performing manipulations in it

```
data=spark.read.format("delta").load("dbfs:/user/hive/warehouse/export") #Reading table in delta format
data.show(10)
```

(1) Spark Jobs

data: pyspark.sql.dataframe.DataFrame = [id: long, firstName: string ... 6 more fields]

	id	firstName	middleName	lastName	gender	birthDate	ssn	salary
2	An	Amira	Cowper	F	1992-02-08 05:00:00	978-97-8086	40203	
19	Adelle	Kathryn	Grigoriev	F	1978-11-14 05:00:00	923-23-5984	60600	
40	Alethea	Dolly	Brickdale	F	1960-06-26 04:00:00	908-46-4236	89151	
61	Adelia	Gita	Vassel	F	1990-10-24 04:00:00	947-17-3832	63563	
63	Anjanette	Clelia	Hicks	F	1973-10-26 04:00:00	981-32-5795	61310	
75	Alica	Elfrieda	Mousdall	F	1953-12-21 05:00:00	954-16-6401	96490	
90	Alice	Marisol	Novill	F	1992-05-21 04:00:00	924-10-7563	34053	
100	Annalisa	Nova	Patesel	F	1974-01-25 04:00:00	922-68-2542	66481	
105	Annie	Jeanne	Beecker	F	1961-12-09 05:00:00	903-15-7880	63493	
110	Alayna	Donnetta	Passby	F	1985-11-09 05:00:00	937-24-5796	76967	

only showing top 10 rows

3.5. Getting the results where last name of the user ends with 'r'

Just now (1s) 10

```
from pyspark.sql.functions import *
dfresult=data.select("*").filter(col("lastName").endsWith("r"))
display(dfresult)
```

▶ (1) Spark Jobs

dfresult: pyspark.sql.dataframe.DataFrame = [id: long, firstName: string ... 6 more fields]

	id	firstName	middleName	lastName	gender	birthDate	ssn	salary
2	105	Annie	Jeanne	Beecker	F	1961-12-09T05:00:00.000+00:...	903-15-7880	63493
3	293	Alida	Peggy	Bracher	F	1972-09-18T04:00:00.000+00:...	953-82-2303	107303
4	296	Ann	Vinita	Roscher	F	1986-05-31T04:00:00.000+00:...	936-40-3049	36909
5	348	Andra	Ami	Player	F	1984-09-10T04:00:00.000+00:...	986-41-7846	97363
6	640	Angelic	Rossana	Raiker	F	1969-06-07T04:00:00.000+00:...	985-74-9894	30965
7	646	Alecia	Latrice	Dener	F	1973-04-02T05:00:00.000+00:...	926-24-5548	64624
8	835	Antoinette	Era	Warriner	F	1982-07-08T04:00:00.000+00:...	954-43-1908	69049
9	2	An	Amira	Cowper	F	1992-02-08T05:00:00.000+00:...	978-97-8086	40203
10	105	Annie	Jeanne	Beecker	F	1961-12-09T05:00:00.000+00:...	903-15-7880	63493
11	293	Alida	Peggy	Bracher	F	1972-09-18T04:00:00.000+00:...	953-82-2303	107303
12	296	Ann	Vinita	Roscher	F	1986-05-31T04:00:00.000+00:...	936-40-3049	36909
13	348	Andra	Ami	Player	F	1984-09-10T04:00:00.000+00:...	986-41-7846	97363
14	640	Angelic	Rossana	Raiker	F	1969-06-07T04:00:00.000+00:...	985-74-9894	30965

3.6.Listing all the datasets in filestore

In 1 minute (8s) 12 Python

```
%fs
ls /databricks-datasets #listing all the datasets in filestore
```

	path	name	size	modificationTime
1	dbfs:/databricks-datasets/COVID/	COVID/	0	1733830997723
2	dbfs:/databricks-datasets/README.md	README.md	976	1532502320000
3	dbfs:/databricks-datasets/Rdatasets/	Rdatasets/	0	1733830997723
4	dbfs:/databricks-datasets/SPARK_README.md	SPARK_README.md	3359	1516124912000
5	dbfs:/databricks-datasets/adult/	adult/	0	1733830997723
6	dbfs:/databricks-datasets/airlines/	airlines/	0	1733830997723
7	dbfs:/databricks-datasets/amazon/	amazon/	0	1733830997723
8	dbfs:/databricks-datasets/asa/	asa/	0	1733830997723
9	dbfs:/databricks-datasets/atlas_higgs/	atlas_higgs/	0	1733830997723
10	dbfs:/databricks-datasets/bikeSharing/	bikeSharing/	0	1733830997723
11	dbfs:/databricks-datasets/cctvVideos/	cctvVideos/	0	1733830997723

3.7.Listing secret scopes

In 2 minutes (<1s) 13

```
dbutils.secrets.listScopes()
#Listing the secret scope and Used to securely store and retrieve secrets, like credentials and API keys.
```

[]

3.8. Creating text widget and getting input values

▶ ✓ In 2 minutes (<1s) 14

```
dbutils.widgets.text("input", "default_value", "Enter a value")#Creating a text widget
dbutils.widgets.get("input")
```

'default_value'

Enter a value
input
10

3.9. Listing the particular dataset

▶ ✓ In 2 minutes (<1s) 15

```
display(dbutils.fs.ls("dbfs:/databricks-datasets/flights/"))
```

▶ (2) Spark Jobs

Table + 🔍 🏠

	^A _C path	^A _C name	¹ ₃ size	¹ ₃ modificationTime
1	dbfs:/databricks-datasets/flights/README.md	README.md	412	1516326253000
2	dbfs:/databricks-datasets/flights/airport-codes-na.txt	airport-codes-na.t...	11411	1516326394000
3	dbfs:/databricks-datasets/flights/departuredelays.csv	departuredelays.csv	33396236	1516326403000

⬇ 3 rows | 0.48 seconds runtime Refreshed in 2 minutes

3.10. Getting all the folder names

▶ ✓ In 2 minutes (1s) 16

```
for foldername in dbutils.fs.ls("databricks-datasets/"):
    print(foldername.name)
```

nyctaxi-with-zipcodes/
online_retail/
overlap-join/
power-plant/
retail-org/
rwe/
sai-summit-2019-sf/
sample_logs/
samples/
sfo_customer_survey/
sms_spam_collection/
songs/
structured-streaming/
timeseries/
...

3.11.Running other notebooks using dbutils

✓ In 2 minutes (11s)

17

`dbutils.notebook.run("/Workspace/Users/azuser2362_mm1.local@techademy.com/new",10)`

Notebook Workflows

Start time	End time	Notebook path	Duration	Status	Error code	Run parameters	
Dec 10, 2024, 05:27 PM	Dec 10, 2024, 05:27 PM	...chademy.com/new	8s	✓ Succeeded			