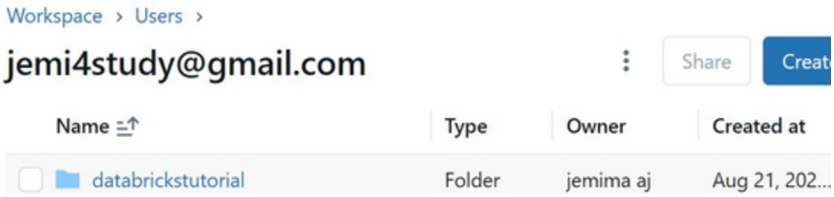
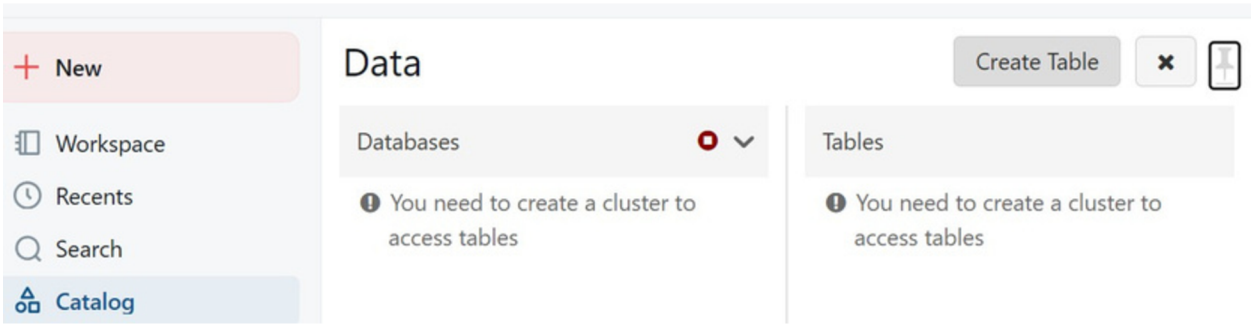
Pyspark Exercises

1. Create a databricks community edition account
2. Create databricks folder under Worspace>>users>youremail



3. Create a table under catalog and load the bigmarket.sales file

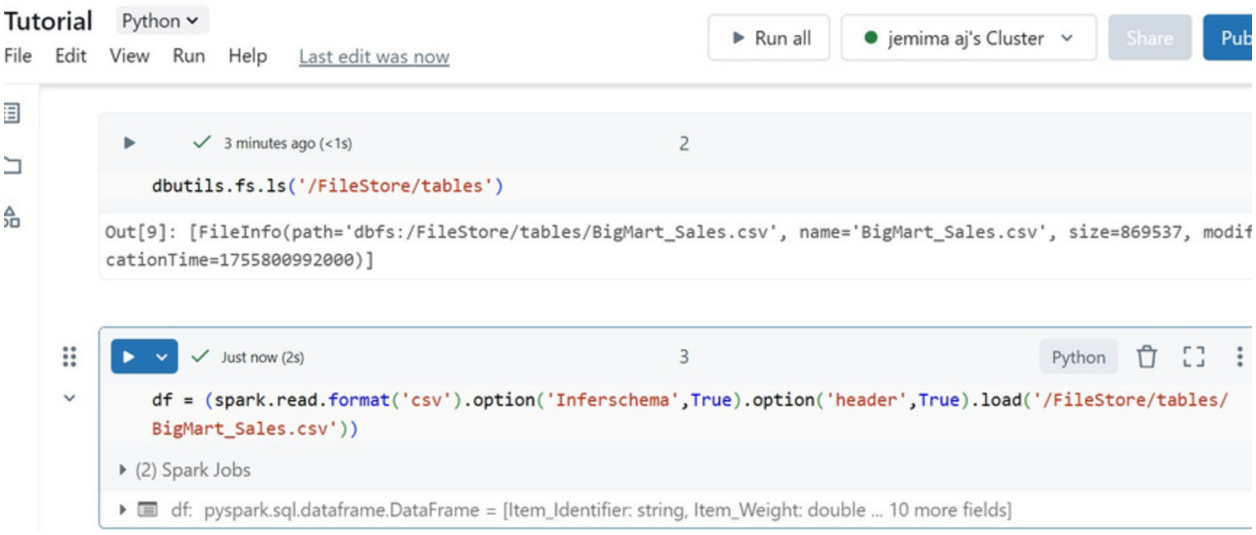


jilu.default.big_mart_sales

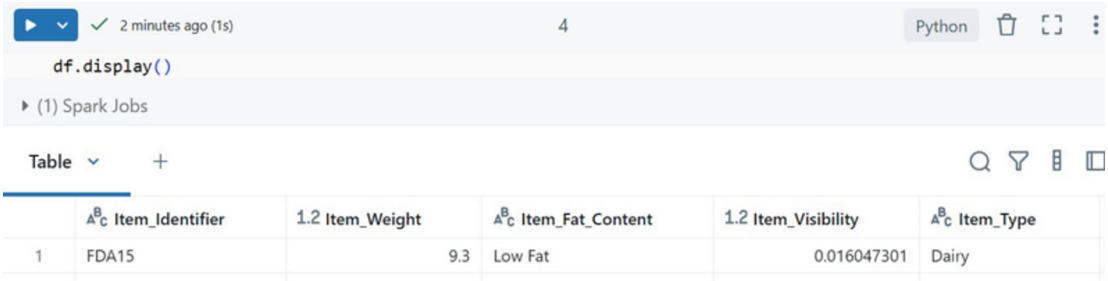
/Volumes/jemi/default/jilu/BigMart Sales.csv

/FileStore/tables/BigMart_Sales.csv

4. Data reading

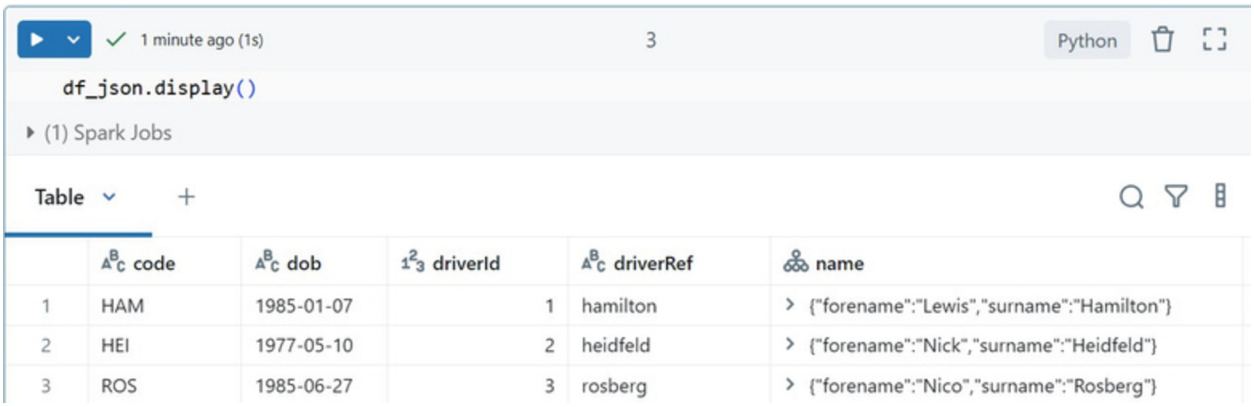


5. Display the data



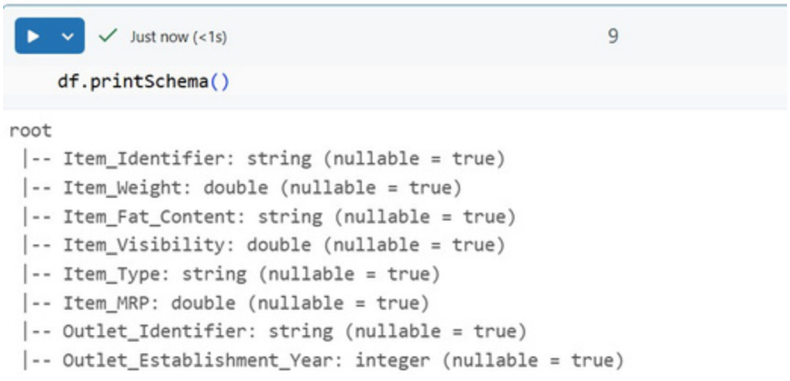
6. Upload the .json file and try to read it

/FileStore/tables/drivers.json



7. Define Schema

SCHEMA DEFINITION



Just now (<1s) 12

```
from pyspark.sql import SparkSession

# Define schema using DDL
my_ddl_schema = """
Item_Identifier STRING,
Item_Weight STRING,
Item_Fat_Content STRING,
Item_Visibility DOUBLE,
Item_Type STRING,
Item_MRP DOUBLE,
Outlet_Identifier STRING,
Outlet_Establishment_Year INT,
Outlet_Size STRING,
```

2 minutes ago (1s) 13

```
df = spark.read.format('csv') \
    .schema(my_ddl_schema) \
    .option('header', True) \
    .load('/FileStore/tables/BigMart_Sales.csv')
```

```
df.display(5)
```

(1) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [Item_Identifier: string, Item_Weight: string ... 10 more fields]

Table

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type
1	FDA15	9.3	Low Fat	0.016047301	Dairy
2	DRC01	5.92	Regular	0.019278216	Soft Drinks
3	FDN15	17.5	Low Fat	0.016760075	Meat

8. Struct type schema

Just now (<1s)

```
from pyspark.sql.types import *
from pyspark.sql.functions import *
```

```
my_strct_schema = StructType([
    StructField('Item_Identifier',StringType(),True),
    StructField('Item_Weight',StringType(),True),
    StructField('Item_Fat_Content',StringType(),True),
    StructField('Item_Visibility',StringType(),True),
    StructField('Item_MRP',StringType(),True),
    StructField('Outlet_Identifier',StringType(),True),
    StructField('Outlet_Establishment_Year',StringType(),True),
    StructField('Outlet_Size',StringType(),True),
    StructField('Outlet_Location_Type',StringType(),True),
    StructField('Outlet_Type',StringType(),True),
    StructField('Item_Outlet_Sales',StringType(),True)
```

9.Reset using Inferschma

Just now (<1s) 17

```
df = spark.read.format('csv') \
    .schema(my_struct_schema) \
    .option('header', True) \
    .load('/FileStore/tables/BigMart_Sales.csv')
```

```
df.printSchema()
```

```
root
 |-- Item_Identifier: string (nullable = true)
 |-- Item_Weight: string (nullable = true)
 |-- Item_Fat_Content: string (nullable = true)
```