

Pyspark Exercises

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

The screenshot shows the Databricks workspace interface. At the top, it displays 'Workspace > Users >' followed by the email address 'jemimajayaraman@gmail.com'. Below this is a 'Create' button. A table lists a single folder entry:

Name	Type	Owner	Created at
databrickstutorial	Folder	jemima aj	Aug 21, 202...

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

The screenshot shows the Databricks Catalog interface. On the left, there's a sidebar with 'New', 'Workspace', 'Recents', 'Search', and 'Catalog' (which is selected). The main area is titled 'Data' and contains two sections: 'Databases' and 'Tables'. Both sections have a note: 'You need to create a cluster to access tables'. In the 'Tables' section, there is one entry: 'jilu.default.big_mart_sales'.

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

/FileStore/tables/BigMart_Sales.csv

4. Data reading

The screenshot shows a Jupyter Notebook cell. The code is:

```
dbutils.fs.ls('/FileStore/tables')
Out[9]: [FileInfo(path='dbfs:/FileStore/tables/BigMart_Sales.csv', name='BigMart_Sales.csv', size=869537, modificationTime=1755800992000)]
```

Below the cell, the output is shown:

```
df = (spark.read.format('csv').option('Inferschema',True).option('header',True).load('/FileStore/tables/BigMart_Sales.csv'))
(2) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [Item_Identifier: string, Item_Weight: double ... 10 more fields]
```

5. Display the data

2 minutes ago (1s) 4 Python

```
df.display()
```

(1) Spark Jobs

Table +

	A ^B _C Item_Identifier	1.2 Item_Weight	A ^B _C Item_Fat_Content	1.2 Item_Visibility	A ^B _C Item_Type
1	FDA15	9.3	Low Fat	0.016047301	Dairy

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

/FileStore/tables/drivers.json

Just now (1s) 2 Python

```
df_json = spark.read.format('json').option('Inferschema',True)\n    .option('header',True)\n    .option('multiline',False).load('/FileStore/tables/drivers.json')
```

1 minute ago (1s) 3 Python

```
df_json.display()
```

(1) Spark Jobs

Table +

	A ^B _C code	A ^B _C dob	1 ² ₃ driverId	A ^B _C driverRef	name
1	HAM	1985-01-07		1 hamilton	> {"forename": "Lewis", "surname": "Hamilton"}
2	HEI	1977-05-10		2 heidfeld	> {"forename": "Nick", "surname": "Heidfeld"}
3	ROS	1985-06-27		3 rosberg	> {"forename": "Nico", "surname": "Rosberg"}

7. Define Schema

SCHEMA DEFINITION

Just now (<1s) 9

```
df.printSchema()
```

```
root
 |-- Item_Identifier: string (nullable = true)
 |-- Item_Weight: double (nullable = true)
 |-- Item_Fat_Content: string (nullable = true)
 |-- Item_Visibility: double (nullable = true)
 |-- Item_Type: string (nullable = true)
 |-- Item_MRP: double (nullable = true)
 |-- Outlet_Identifier: string (nullable = true)
 |-- Outlet_Establishment_Year: integer (nullable = true)
```

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 +

	A _B C Item_Identifier	A _B C Item_Weight	A _B C Item_Fat_Content	1.2 Item_Visibility	A _B C 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_struct_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 InferSchema

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)
```