databricks2023-02-26 - Supermarket data

(https://databricks.com)

Overview

This notebook will show you how to create and query a table or DataFrame that you uploaded to DBFS. DBFS (https://docs.databricks.com/user-guide/dbfs-databricks-file-system.html) is a Databricks File System that allows you to store data for querying inside of Databricks. This notebook assumes that you have a file already inside of DBFS that you would like to read from.

This notebook is written in **Python** so the default cell type is Python. However, you can use different languages by using the `%LANGUAGE` syntax. Python, Scala, SQL, and R are all supported.

```
# File location and type
file_location = "/FileStore/tables/order_products.csv"
file_type = "csv"

# The applied options are for CSV files. For other file types, these will be
ignored.
order_products = spark.read.csv(file_location,header=True,inferSchema=True)


file_location = "/FileStore/tables/products.csv"
products = spark.read.csv(file_location,header=True,inferSchema=True)


file_location = "/FileStore/tables/orders_small_version.csv"
orders_small_version =
spark.read.csv(file_location,header=True,inferSchema=True)


file_location = "/FileStore/tables/departments.csv"
departments = spark.read.csv(file_location,header=True,inferSchema=True)


file_location = "/FileStore/tables/aisles.csv"
aisles = spark.read.csv(file_location,header=True,inferSchema=True)


order_products.printSchema()

root
 |-- order_id: integer (nullable = true)
 |-- product_id: integer (nullable = true)
```

```
 |-- add_to_cart_order: integer (nullable = true)
 |-- reordered: integer (nullable = true)
```

```
products.printSchema()
```

```
root
 |-- product_id: integer (nullable = true)
 |-- product_name: string (nullable = true)
 |-- aisle_id: string (nullable = true)
 |-- department_id: string (nullable = true)
```

```
orders_small_version.printSchema()
```

```
root
 |-- order_id: integer (nullable = true)
 |-- user_id: integer (nullable = true)
 |-- eval_set: string (nullable = true)
 |-- order_number: integer (nullable = true)
 |-- order_dow: integer (nullable = true)
 |-- order_hour_of_day: integer (nullable = true)
 |-- days_since_prior_order: double (nullable = true)
```

```
departments.printSchema()
```

```
root
 |-- department_id: integer (nullable = true)
 |-- department: string (nullable = true)
```

```
aisles.printSchema()
```

```
root
 |-- aisle_id: integer (nullable = true)
 |-- aisle: string (nullable = true)
```

```
order_products.show(5)
```

```
+--------+----------+-----------------+---------+
|order_id|product_id|add_to_cart_order|reordered|
+--------+----------+-----------------+---------+
|       1|     49302|                1|        1|
|       1|     11109|                2|        1|
|       1|     10246|                3|        0|
|       1|     49683|                4|        0|
|       1|     43633|                5|        1|
+--------+----------+-----------------+---------+
```

```
only showing top 5 rows


products.show(5)

+----------+--------------------+--------+-------------+
|product_id|        product_name|aisle_id|department_id|
+----------+--------------------+--------+-------------+
|         1|Chocolate Sandwic...|      61|           19|
|         2|    All-Seasons Salt|     104|           13|
|         3|Robust Golden Uns...|      94|            7|
|         4|Smart Ones Classi...|      38|            1|
|         5|Green Chile Anyti...|       5|           13|
+----------+--------------------+--------+-------------+
only showing top 5 rows


orders_small_version.show(5)

+--------+-------+--------+------------+---------+-----------------+--------
--------------+
|order_id|user_id|eval_set|order_number|order_dow|order_hour_of_day|days_sin
ce_prior_order|
+--------+-------+--------+------------+---------+-----------------+--------
--------------+
| 2539329|      1|   prior|           1|        2|                8|
null|
| 2398795|      1|   prior|           2|        3|                7|
15.0|
|  473747|      1|   prior|           3|        3|               12|
21.0|
| 2254736|      1|   prior|           4|        4|                7|
29.0|
|  431534|      1|   prior|           5|        4|               15|
28.0|
+--------+-------+--------+------------+---------+-----------------+--------
--------------+
only showing top 5 rows


departments.show(5)

+-------------+----------+
|department_id|department|
+-------------+----------+
|            1|    frozen|
|            2|     other|
|            3|    bakery|
|            4|   produce|
|            5|   alcohol|
```

```
+-------------+----------+
only showing top 5 rows


aisles.show(5)

+--------+--------------------+
|aisle_id|               aisle|
+--------+--------------------+
|       1|prepared soups sa...|
|       2|   specialty cheeses|
|       3| energy granola bars|
|       4|       instant foods|
|       5|marinades meat pr...|
+--------+--------------------+
only showing top 5 rows


temp_table_name = "order_products"
order_products.createOrReplaceTempView(temp_table_name)

temp_table_name = "products"
products.createOrReplaceTempView(temp_table_name)

temp_table_name = "orders_small_version"
orders_small_version.createOrReplaceTempView(temp_table_name)

temp_table_name = "departments"
departments.createOrReplaceTempView(temp_table_name)

temp_table_name = "aisles"
aisles.createOrReplaceTempView(temp_table_name)
```

Table

	product_id	product_name	reordered_sum
1	24852	Banana	16557
2	13176	Bag of Organic Bananas	13362
3	21137	Organic Strawberries	8603
4	21903	Organic Baby Spinach	8055
5	47766	Organic Avocado	6226
6	47209	Organic Hass Avocado	6042
7	47626	Large Lemon	5923

10 rows