Spark-SQL

November 25, 2024

[4]:
```
# Import SparkSession
from pyspark.sql import SparkSession

# Create SparkSession
spark = SparkSession.builder.master("local[1]").appName("SparkByExamples.com").
 ↪getOrCreate()

#Read CSV file into table
df = spark.read.option("header",True).csv("/Users/admin/Jupyter Examples/
 ↪simple-zipcodes.csv")
df.printSchema()
df.show()
```

```
root
 |-- RecordNumber: string (nullable = true)
 |-- Country: string (nullable = true)
 |-- City: string (nullable = true)
 |-- Zipcode: string (nullable = true)
 |-- State: string (nullable = true)

+------------+-------+--------------------+-------+-----+
|RecordNumber|Country|                City|Zipcode|State|
+------------+-------+--------------------+-------+-----+
|           1|     US|         PARC PARQUE|    704|   PR|
|           2|     US|PASEO COSTA DEL SUR|    704|   PR|
|          10|     US|        BDA SAN LUIS|    709|   PR|
|       49347|     US|                HOLT|  32564|   FL|
|       49348|     US|           HOMOSASSA|  34487|   FL|
|       61391|     US|  CINGULAR WIRELESS|  76166|   TX|
|       61392|     US|          FORT WORTH|  76177|   TX|
|       61393|     US|            FT WORTH|  76177|   TX|
|       54356|     US|         SPRUCE PINE|  35585|   AL|
|       76511|     US|            ASH HILL|  27007|   NC|
|           4|     US|    URB EUGENE RICE|    704|   PR|
|       39827|     US|                MESA|  85209|   AZ|
|       39828|     US|                MESA|  85210|   AZ|
|       49345|     US|            HILLIARD|  32046|   FL|
|       49346|     US|              HOLDER|  34445|   FL|
```

```
|           3|     US|      SECT LANAUSSE|    704|   PR|
|       54354|     US|      SPRING GARDEN|  36275|   AL|
|       54355|     US|        SPRINGVILLE|  35146|   AL|
|       76512|     US|           ASHEBORO|  27203|   NC|
|       76513|     US|           ASHEBORO|  27204|   NC|
+------------+-------+--------------------+-------+-----+
```

[6]:
```
# Create temporary table
spark.read.option("header",True).csv("/Users/admin/Jupyter Examples/
 ↪simple-zipcodes.csv").createOrReplaceTempView("Zipcodes")
```

[7]:
```
# DataFrame API Select query
df.select("country","city","zipcode","state").show(5)
```

```
+-------+--------------------+-------+-----+
|country|                city|zipcode|state|
+-------+--------------------+-------+-----+
|     US|         PARC PARQUE|    704|   PR|
|     US|PASEO COSTA DEL SUR|    704|   PR|
|     US|        BDA SAN LUIS|    709|   PR|
|     US|                HOLT|  32564|   FL|
|     US|           HOMOSASSA|  34487|   FL|
+-------+--------------------+-------+-----+
only showing top 5 rows
```

[12]:
```
# DataFrame API where()
df.select("country","city","zipcode","state").where("state == 'TX'").show()
```

```
+-------+-----------------+-------+-----+
|country|             city|zipcode|state|
+-------+-----------------+-------+-----+
|     US|CINGULAR WIRELESS|  76166|   TX|
|     US|       FORT WORTH|  76177|   TX|
|     US|         FT WORTH|  76177|   TX|
+-------+-----------------+-------+-----+
```

[15]:
```
# In spark you can use like this
result = spark.sql(""" SELECT  country, city, zipcode, state FROM ZIPCODES␣
 ↪WHERE state = 'AZ' """)
result.show()
```

```
+-------+----+-------+-----+
|country|city|zipcode|state|
+-------+----+-------+-----+
|     US|MESA|  85209|   AZ|
|     US|MESA|  85210|   AZ|
```

```
+-------+----+-------+-----+
```

[16]:
```
# SQL GROUP BY clause
result = spark.sql(""" SELECT state, count(*) as count FROM ZIPCODES GROUP BY␣
 ↪state""")
result.show()
```

```
+-----+-----+
|state|count|
+-----+-----+
|   AZ|    2|
|   NC|    3|
|   AL|    3|
|   TX|    3|
|   FL|    4|
|   PR|    5|
+-----+-----+
```

[17]:
```
# Create a temporary table for population
spark.read.option("header",True).csv("/Users/admin/Jupyter Examples/
 ↪state-population.csv").createOrReplaceTempView("Populations")
```

[18]:
```
result = spark.sql(""" SELECT * FROM POPULATIONS """)
result.show()
```

```
+-----+----------+
|State|population|
+-----+----------+
|   PR|        23|
|   FL|       456|
|   TX|      1000|
|   AZ|        78|
|   AL|        21|
|   NC|        40|
+-----+----------+
```

[19]:
```
# Inner Join Operation
result = spark.sql(""" SELECT * FROM ZIPCODES Z, POPULATIONS P WHERE Z.STATE=P.
 ↪STATE""")
result.show()
```

```
+------------+-------+------------------+-------+-----+-----+----------+
|RecordNumber|Country|              City|Zipcode|State|State|population|
+------------+-------+------------------+-------+-----+-----+----------+
|           1|     US|       PARC PARQUE|    704|   PR|   PR|        23|
|           2|     US|PASEO COSTA DEL SUR|    704|   PR|   PR|        23|
```

```
|          10|     US|       BDA SAN LUIS|    709|   PR|   PR|        23|
|       49347|     US|              HOLT|  32564|   FL|   FL|       456|
|       49348|     US|         HOMOSASSA|  34487|   FL|   FL|       456|
|       61391|     US| CINGULAR WIRELESS|  76166|   TX|   TX|      1000|
|       61392|     US|        FORT WORTH|  76177|   TX|   TX|      1000|
|       61393|     US|          FT WORTH|  76177|   TX|   TX|      1000|
|       54356|     US|       SPRUCE PINE|  35585|   AL|   AL|        21|
|       76511|     US|          ASH HILL|  27007|   NC|   NC|        40|
|           4|     US|   URB EUGENE RICE|    704|   PR|   PR|        23|
|       39827|     US|              MESA|  85209|   AZ|   AZ|        78|
|       39828|     US|              MESA|  85210|   AZ|   AZ|        78|
|       49345|     US|          HILLIARD|  32046|   FL|   FL|       456|
|       49346|     US|            HOLDER|  34445|   FL|   FL|       456|
|           3|     US|     SECT LANAUSSE|    704|   PR|   PR|        23|
|       54354|     US|     SPRING GARDEN|  36275|   AL|   AL|        21|
|       54355|     US|       SPRINGVILLE|  35146|   AL|   AL|        21|
|       76512|     US|          ASHEBORO|  27203|   NC|   NC|        40|
|       76513|     US|          ASHEBORO|  27204|   NC|   NC|        40|
+------------+-------+-------------------+-------+-----+-----+----------+
```

[22]:
```
case class Home(city: String, size: Int, lotSize: Int, bedrooms: Int, bathrooms:
 ↪ Int, price: Int)
val homes = List(Home("San Francisco", 1500, 4000, 3, 2, 1500000),
Home("Palo Alto", 1800, 3000, 4, 2, 1800000),
Home("Mountain View", 2000, 4000, 4, 2, 1500000),
Home("Sunnyvale", 2400, 5000, 4, 3, 1600000),
Home("San Jose", 3000, 6000, 4, 3, 1400000),
Home("Fremont", 3000, 7000, 4, 3, 1500000),
Home("Pleasanton", 3300, 8000, 4, 3, 1400000),
Home("Berkeley", 1400, 3000, 3, 3, 1100000),
Home("Oakland", 2200, 6000, 4, 3, 1100000),
Home("Emeryville", 2500, 5000, 4, 3, 1200000))
```

```
  Cell In[22], line 1
    case class Home(city: String, size: Int, lotSize: Int, bedrooms: Int,␣
 ↪bathrooms: Int, price: Int)
         ^
SyntaxError: invalid syntax
```

[ ]: