Spark SQL

March 28, 2020

[1]:

```
import os
os.environ['PYSPARK_PYTHON'] = '/usr/bin/python3'

from pyspark.sql import SparkSession

sc = SparkSession.builder.master('spark://172.18.0.7:7077').config('spark.
 ↪executor.memory', '1g').getOrCreate()
```

[3]:

```
df = sc.read.option('delimiter', ';').option('header', 'true').
 ↪option('inferschema', 'true')\
    .csv('hdfs://172.18.0.9:9000/base_parcial.csv')
```

[7]:

```
df.createOrReplaceTempView('comercio')
```

[18]:

```
df.printSchema()
df.head(1)
```

```
root
 |-- country_or_area: string (nullable = true)
 |-- year: integer (nullable = true)
 |-- comm_code: integer (nullable = true)
 |-- commodity: string (nullable = true)
 |-- flow: string (nullable = true)
 |-- trade_usd: long (nullable = true)
 |-- weight_kg: long (nullable = true)
 |-- quantity_name: string (nullable = true)
 |-- quantity: decimal(20,0) (nullable = true)
 |-- category: string (nullable = true)
```

[18]:

```
[Row(country_or_area='Afghanistan', year=2016, comm_code=10410,
commodity='Sheep, live', flow='Export', trade_usd=6088, weight_kg=2339,
quantity_name='Number of items', quantity=Decimal('51'),
category='01_live_animals')]
```

[8]:

```
sqlQuantidade = sc.sql('select country_or_area, count(*) as qty from comercio␣
 ↪group by country_or_area')
```

```
sqlQuantidade.show()
```

```
+------------------+-----+
|   country_or_area|  qty|
+------------------+-----+
|     Côte d'Ivoire|25925|
|              Chad|  497|
|   Rep. of Moldova| 6871|
|          Anguilla| 3804|
|          Paraguay| 8313|
|             Yemen| 2599|
|State of Palestine| 1973|
|           Senegal| 4671|
|        Cabo Verde|19966|
|            Sweden| 9652|
|          Kiribati| 3085|
|         Fmr Sudan| 9692|
|         Cook Isds| 1487|
|            Guyana|14859|
|           Eritrea|  754|
|       Philippines| 8496|
|          Djibouti|  557|
|             Tonga| 1288|
|          Malaysia|24683|
|         Singapore|10536|
+------------------+-----+
only showing top 20 rows
```

[9]:
```
sqlQuantidade.rdd.count()
```

[9]:
```
207
```

1 Pais com maior quantidade de transações comerciais efetuadas

[36]:
```
sqlEx = sc.sql('select country_or_area, count(*) as qty from comercio group by␣
 ↪country_or_area order by qty desc limit 5')
sqlEx.show()
```

```
+--------------------+-----+
|     country_or_area|  qty|
+--------------------+-----+
|           Australia|89487|
|              Canada|69468|
|             Austria|58683|
|          Argentina|57729|
|China, Hong Kong SAR|55535|
+--------------------+-----+
```

2 Mercadoria com a maior quantidade de transações comerciais no Brasil

[35]:

```
sqlEx = sc.sql("select commodity, count(*) as qty from comercio where␣
 ↪country_or_area = 'Brazil' group by commodity order by qty desc limit 5")
sqlEx.show()
```

```
+--------------------+---+
|           commodity|qty|
+--------------------+---+
|Industrial fatty ...| 84|
|Cigarette or pipe...| 69|
|Synthetic organic...| 69|
|Polymer based pai...| 68|
|Washing and clean...| 68|
+--------------------+---+
```

3 Quantidade de transações financeiras realiazadas por ano

[22]:

```
sqlEx = sc.sql("select year, count(*) as qty from comercio group by year order␣
 ↪by year")
sqlEx.show()
```

```
+----+------+
|year|   qty|
+----+------+
|1988| 11210|
|1989| 23181|
|1990| 26372|
|1991| 30632|
|1992| 44143|
|1993| 56180|
|1994| 74074|
|1995| 86789|
|1996| 93028|
|1997|102880|
|1998|106086|
|1999|113039|
|2000|125212|
|2001|126207|
|2002|126231|
|2003|130248|
|2004|131457|
|2005|133867|
|2006|135996|
|2007|135855|
+----+------+
only showing top 20 rows
```

4 Mercadoria com maior quantidade de transações financeiras

[34]:
```
sqlEx = sc.sql("select commodity, count(*) as qty from comercio group by␣
 ↪commodity order by qty desc limit 5")
sqlEx.show()
```

```
+--------------------+----+
|           commodity| qty|
+--------------------+----+
|Food preparations…|8048|
|Sugar confectione…|7764|
|Sauces nes, mixed…|7693|
|Cigarettes contai…|7620|
|Communion wafers,…|7530|
+--------------------+----+
```

5 Mercadoria com maoir quantidade de transações financeiras em 2016

[33]:
```
sqlEx = sc.sql("select commodity, count(*) as qty from comercio where year =␣
 ↪2016 group by commodity order by qty desc limit 5")
sqlEx.show()
```

```
+--------------------+---+
|           commodity|qty|
+--------------------+---+
|Food preparations…|278|
|Cigarettes contai…|272|
|Communion wafers,…|269|
|Sauces nes, mixed…|268|
|Sweet biscuits, w…|267|
+--------------------+---+
```

6 Mercadoria com maior quantidade de transações financeiras em 2016, no Brasil

[32]:
```
sqlEx = sc.sql("select commodity, count(*) as qty from comercio where year =␣
 ↪2016 and country_or_area = 'Brazil' group by commodity order by qty desc␣
 ↪limit 5")
sqlEx.show()
```

```
+--------------------+---+
|           commodity|qty|
+--------------------+---+
|Sunflower or saff…|  2|
|Caviar and caviar…|  2|
|           Soaps nes|  2|
| Cheese, blue-veined|  2|
|Hair, human, unwo…|  2|
+--------------------+---+
```

7 Mercadoria com maior total de peso, de acordo com totas transações comerciais

[31]:
```
sqlEx = sc.sql("select commodity, sum(weight_kg) as qty from comercio group by␣
 ↪commodity order by qty desc limit 5")
sqlEx.show()
```

```
+--------------------+--------------+
|           commodity|           qty|
+--------------------+--------------+
|Petroleum oils, o…|46002412921988|
|Iron ore, concent…|34878419167261|
|Ice, snow and pot…|25759966772196|
|Bituminous coal, …|21959118134095|
|Oils petroleum, b…|19216117424056|
+--------------------+--------------+
```

8 Mercadoria com maior total de peso, de acordo com todas as transações comerciais, separadas de acordo com o ano

[37]:

```
sqlEx = sc.sql("select year, commodity, sum(weight_kg) as qty from comercio␣
 ↪group by year, commodity order by year, qty desc")
sqlEx.show()
```

```
+----+--------------------+------------+
|year|           commodity|         qty|
+----+--------------------+------------+
|1988|Iron ore, concent…|288037119271|
|1988|Petroleum oils, o…|275362576231|
|1988|Bituminous coal, …|207345820868|
|1988|Oils petroleum, b…| 46409294336|
|1988|Iron ore, concent…| 35704526099|
|1988|Natural gas in ga…| 33085326564|
|1988|Natural gas, liqu…| 31032184409|
|1988|Wheat except duru…| 25521701421|
|1988|Maize except seed…| 24239136321|
|1988|Light petroleum d…| 19452652381|
|1988|Pebbles, gravel, …| 18926080388|
|1988|Fuel oils nes, he…| 18494640517|
|1988|Coal except anthr…| 13937858572|
|1988|Natural sands nes…| 10798045384|
|1988|Gas oils - bunker…| 10253434963|
|1988|Salt (sodium chlo…|  9629692808|
|1988|Coke, semi-coke o…|  9470623632|
|1988|  Propane, liquefied|  8761684270|
|1988|Aluminium oxide, …|  8355422784|
|1988|          Soya beans|  7996459678|
+----+--------------------+------------+
only showing top 20 rows
```

9 Média de peso por mercadoria, separada de acordo com o ano

[38]:

```
sqlEx = sc.sql("select commodity, year, avg(weight_kg) as qty from comercio␣
 ↪group by year, commodity order by year, qty desc")
sqlEx.show()
```

```
+--------------------+----+--------------------+
|           commodity|year|                 qty|
+--------------------+----+--------------------+
|Bituminous coal, …|1988|3.455763681133333…|
|Petroleum oils, o…|1988|3.059584180344444…|
|Iron ore, concent…|1988| 1.80023199544375E10|
|Oils petroleum, b…|1988|     1.1602323584E10|
|Natural gas in ga…|1988| 4.726475223428572E9|
|Natural gas, liqu…|1988| 4.433169201285714E9|
|Aluminium oxide, …|1988|       4.177711392E9|
|Pebbles, gravel, …|1988|3.1543467313333335E9|
|Iron ore, concent…|1988| 2.550323292785714E9|
|Light petroleum d…|1988|      1.9452652381E9|
|Fuel oils nes, he…|1988|      1.8494640517E9|
|Natural sands nes…|1988|1.7996742306666667E9|
|Coal except anthr…|1988|      1.3937858572E9|
|Salt (sodium chlo…|1988| 1.375670401142857E9|
|Wheat except duru…|1988|1.3432474432105262E9|
|Maize except seed…|1988|     1.21195681605E9|
|Gas oils - bunker…|1988| 9.321304511818181E8|
|          Soya beans|1988|       7.996459678E8|
|Coke, semi-coke o…|1988|        7.89218636E8|
|Petroleum coke, n…|1988| 7.729372491428572E8|
+--------------------+----+--------------------+
only showing top 20 rows
```

10 Média de peso por mercadoria comercializadas no Brasil, separadas por ano

[30]:

```
sqlEx = sc.sql("select commodity, year, avg(weight_kg) as qty from comercio␣
 ↪where country_or_area = 'Brazil' group by year, commodity order by year")
sqlEx.show()
```

```
+--------------------+----+-----------+
|           commodity|year|        qty|
+--------------------+----+-----------+
|Tuna nes, fresh o…|1989|     4437.0|
|Bovine cuts bonel…|1989|6.5151152E7|
|Frozen vegetable …|1989|       48.0|
|Emery & natural a…|1989|    11699.0|
|Antibiotics nes, …|1989|   113801.5|
|Natural graphite,…|1989|    23339.0|
|Pigments and prep…|1989|    27804.5|
|Garlic, fresh or …|1989|  6331098.5|
|Ground-nuts shell…|1989|  1821731.5|
|Granules, chippin…|1989|        0.0|
|Zinc oxide and pe…|1989|   690314.0|
|Vat dyes and prep…|1989|    62856.0|
|Veg nes, mixes, p…|1989|   141288.0|
|Copper oxides and…|1989|   106948.5|
|Artificial and pr…|1989|   251041.5|
|Heavy water (deut…|1989|        0.0|
|Sulphites of meta…|1989|    59245.0|
|Pistachios, fresh…|1989|    22933.0|
|      Yeasts, active|1989|  1011594.5|
|Fructose, syrup >…|1989|      148.0|
+--------------------+----+-----------+
only showing top 20 rows
```