sparksessionpractice

```
%pyspark                                                          FINISHED
# Numero de muertes a nivel mundial
```

Took 0 sec. Last updated by anonymous at May 16 2020, 8:49:41 PM.

```
%pyspark                                                          FINISHED

from pyspark.sql import SparkSession
```

Took 0 sec. Last updated by anonymous at May 16 2020, 8:45:17 PM.

```
%pyspark                                                          FINISHED
spark = SparkSession.builder.appName('Basics').getOrCreate()
```

Took 0 sec. Last updated by anonymous at May 16 2020, 8:45:21 PM.

```
%pyspark                                                          FINISHED

SensorDataRDD = sc.textFile('s3://aws-logs-061895363576-us-east-1/mundial/time_series_covid19_deat

SensorDataCount = SensorDataRDD.count()
SensorDataCount
```

267

Took 7 sec. Last updated by anonymous at May 16 2020, 8:45:51 PM.

```
%pyspark                                                          FINISHED

df = spark.read.csv('s3://aws-logs-061895363576-us-east-1/mundial/time_series_covid19_deaths_globa
```

Took 1 sec. Last updated by anonymous at May 16 2020, 8:46:03 PM.

```
%pyspark                                                          FINISHED

df.columns
```
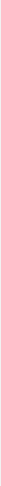
['Province/State', 'Country/Region', 'Lat', 'Long', '1/22/20', '1/23/20', '1/24/20', '1/25/20', '1/
26/20', '1/27/20', '1/28/20', '1/29/20', '1/30/20', '1/31/20', '2/1/20', '2/2/20', '2/3/20', '2/4/2
0', '2/5/20', '2/6/20', '2/7/20', '2/8/20', '2/9/20', '2/10/20', '2/11/20', '2/12/20', '2/13/20',
'2/14/20', '2/15/20', '2/16/20', '2/17/20', '2/18/20', '2/19/20', '2/20/20', '2/21/20', '2/22/20',
'2/23/20', '2/24/20', '2/25/20', '2/26/20', '2/27/20', '2/28/20', '2/29/20', '3/1/20', '3/2/20',
'3/3/20', '3/4/20', '3/5/20', '3/6/20', '3/7/20', '3/8/20', '3/9/20', '3/10/20', '3/11/20', '3/12/2
0', '3/13/20', '3/14/20', '3/15/20', '3/16/20', '3/17/20', '3/18/20', '3/19/20', '3/20/20', '3/21/2
0', '3/22/20', '3/23/20', '3/24/20', '3/25/20', '3/26/20', '3/27/20', '3/28/20', '3/29/20', '3/30/2
0', '3/31/20', '4/1/20', '4/2/20', '4/3/20', '4/4/20', '4/5/20', '4/6/20', '4/7/20', '4/8/20', '4/
9/20', '4/10/20', '4/11/20', '4/12/20', '4/13/20', '4/14/20', '4/15/20', '4/16/20', '4/17/20', '4/1
8/20', '4/19/20', '4/20/20', '4/21/20', '4/22/20', '4/23/20', '4/24/20', '4/25/20', '4/26/20', '4/2
7/20', '4/28/20', '4/29/20', '4/30/20', '5/1/20', '5/2/20', '5/3/20', '5/4/20', '5/5/20', '5/6/20',
'5/7/20', '5/8/20', '5/9/20', '5/10/20', '5/11/20', '5/12/20', '5/13/20', '5/14/20', '5/15/20']

Took 0 sec. Last updated by anonymous at May 16 2020, 8:46:07 PM.

```
%pyspark
df.printSchema()
```
FINISHED

sparksessionpractice

```
root
 |-- Province/State: string (nullable = true)
 |-- Country/Region: string (nullable = true)
 |-- Lat: string (nullable = true)
 |-- Long: string (nullable = true)
 |-- 1/22/20: string (nullable = true)
 |-- 1/23/20: string (nullable = true)
 |-- 1/24/20: string (nullable = true)
 |-- 1/25/20: string (nullable = true)
 |-- 1/26/20: string (nullable = true)
 |-- 1/27/20: string (nullable = true)
 |-- 1/28/20: string (nullable = true)
 |-- 1/29/20: string (nullable = true)
 |-- 1/30/20: string (nullable = true)
 |-- 1/31/20: string (nullable = true)
 |-- 2/1/20: string (nullable = true)
 |-- 2/2/20: string (nullable = true)
 |-- 2/3/20: string (nullable = true)
```

Took 0 sec. Last updated by anonymous at May 16 2020, 8:46:13 PM.

```
%pyspark

df.head(5)
```
FINISHED

```
[Row(Province/State=None, Country/Region='Afghanistan', Lat='33.0', Long='65.0', 1/22/20='0', 1/
23/20='0', 1/24/20='0', 1/25/20='0', 1/26/20='0', 1/27/20='0', 1/28/20='0', 1/29/20='0', 1/30/20
='0', 1/31/20='0', 2/1/20='0', 2/2/20='0', 2/3/20='0', 2/4/20='0', 2/5/20='0', 2/6/20='0', 2/7/2
0='0', 2/8/20='0', 2/9/20='0', 2/10/20='0', 2/11/20='0', 2/12/20='0', 2/13/20='0', 2/14/20='0',
2/15/20='0', 2/16/20='0', 2/17/20='0', 2/18/20='0', 2/19/20='0', 2/20/20='0', 2/21/20='0', 2/22/
20='0', 2/23/20='0', 2/24/20='0', 2/25/20='0', 2/26/20='0', 2/27/20='0', 2/28/20='0', 2/29/20
='0', 3/1/20='0', 3/2/20='0', 3/3/20='0', 3/4/20='0', 3/5/20='0', 3/6/20='0', 3/7/20='0', 3/8/20
='0', 3/9/20='0', 3/10/20='0', 3/11/20='0', 3/12/20='0', 3/13/20='0', 3/14/20='0', 3/15/20='0',
3/16/20='0', 3/17/20='0', 3/18/20='0', 3/19/20='0', 3/20/20='0', 3/21/20='0', 3/22/20='1', 3/23/
20='1', 3/24/20='1', 3/25/20='2', 3/26/20='4', 3/27/20='4', 3/28/20='4', 3/29/20='4', 3/30/20
='4', 3/31/20='4', 4/1/20='4', 4/2/20='6', 4/3/20='6', 4/4/20='7', 4/5/20='7', 4/6/20='11', 4/7/
20='14', 4/8/20='14', 4/9/20='15', 4/10/20='15', 4/11/20='18', 4/12/20='18', 4/13/20='21', 4/14/
20='23', 4/15/20='25', 4/16/20='30', 4/17/20='30', 4/18/20='30', 4/19/20='33', 4/20/20='36', 4/2
1/20='36', 4/22/20='40', 4/23/20='42', 4/24/20='43', 4/25/20='47', 4/26/20='50', 4/27/20='57',
4/28/20='58', 4/29/20='60', 4/30/20='64', 5/1/20='68', 5/2/20='72', 5/3/20='85', 5/4/20='90', 5/
5/20='95', 5/6/20='104', 5/7/20='106', 5/8/20='109', 5/9/20='115', 5/10/20='120', 5/11/20='122',
5/12/20='127', 5/13/20='132', 5/14/20='136', 5/15/20='153'), Row(Province/State=None, Country/Re
gion='Albania', Lat='41.1533', Long='20.1683', 1/22/20='0', 1/23/20='0', 1/24/20='0', 1/25/20
```

Took 1 sec. Last updated by anonymous at May 16 2020, 8:46:18 PM.

```
%pyspark

df.describe().show()
```
FINISHED

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

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

sparksessionpractice

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

Took 6 sec. Last updated by anonymous at May 16 2020, 8:46:31 PM.

%pyspark   FINISHED

```
df.orderBy(df['5/15/20'].desc()).show()
```

```
4/20|3/5/20|3/6/20|3/7/20|3/8/20|3/9/20|3/10/20|3/11/20|3/12/20|3/13/20|3/14/20|3/15/20|3/16/20|
3/17/20|3/18/20|3/19/20|3/20/20|3/21/20|3/22/20|3/23/20|3/24/20|3/25/20|3/26/20|3/27/20|3/28/20|
3/29/20|3/30/20|3/31/20|4/1/20|4/2/20|4/3/20|4/4/20|4/5/20|4/6/20|4/7/20|4/8/20|4/9/20|4/10/20|
4/11/20|4/12/20|4/13/20|4/14/20|4/15/20|4/16/20|4/17/20|4/18/20|4/19/20|4/20/20|4/21/20|4/22/20|
4/23/20|4/24/20|4/25/20|4/26/20|4/27/20|4/28/20|4/29/20|4/30/20|5/1/20|5/2/20|5/3/20|5/4/20|5/5/
20|5/6/20|5/7/20|5/8/20|5/9/20|5/10/20|5/11/20|5/12/20|5/13/20|5/14/20|5/15/20|
+-----------------+------------------+-------+-------+-------+-------+-------+-------
+-------+-------+-------+-------+-------+-------+-------+-------+-------+------+-----
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
------+-------+-------+------+-------+-------+-------+-------+-------+-------+------+---
---+-------+------+-------+-------+-------+-------+-------+-------+-------+------+-------+-
------+-------+------+-------+-------+-------+-------+-------+-------+-------+-------+-
-----+-------+------+-------+-------+-------+-------+-------+-------+-------+------+---
----+-------+------+-------+-------+-------+-------+-------+-------+-------+------+-------+-
-----+-------+------+-------+-------+-------+-------+-------+-------+-------+------+-
+------+------+------+------+-------+-------+-------+-------+-------+-------+
|                 null|    North Macedonia| 41.6086| 21.7453|      0|      0|      0|      0|      0
|        0|       0|      0|      0|      0|      0|      0|      0|      0|      0|      0|      0|      0
```

Took 0 sec. Last updated by anonymous at May 16 2020, 8:47:02 PM.

%pyspark   FINISHED
```
from pyspark.sql.functions import mean
```

Took 0 sec. Last updated by anonymous at May 16 2020, 8:49:11 PM.

%pyspark   FINISHED
```
df.select(mean("5/15/20")).show()
# Promedio de muertes el 15 de mayo
```

```
+-----------------+
|     avg(5/15/20)|
+-----------------+
|1156.639097744361|
+-----------------+
```

Took 7 sec. Last updated by anonymous at May 16 2020, 8:49:30 PM.

%pyspark   READY