sparksessionpractice

FINISHED

```
%pyspark
# Número de casos positivos a nivel mundial
```

Took 0 sec. Last updated by anonymous at May 16 2020, 8:57:01 PM.

FINISHED

```
%pyspark
from pyspark.sql import SparkSession
```

Took 0 sec. Last updated by anonymous at May 16 2020, 8:57:07 PM.

FINISHED

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

Took 0 sec. Last updated by anonymous at May 16 2020, 8:57:10 PM.

FINISHED

```
%pyspark

SensorDataRDD = sc.textFile('s3://aws-logs-061895363576-us-east-1/mundial/time_series_covid19_co
SensorDataCount = SensorDataRDD.count()
SensorDataCount
```

267

Took 7 sec. Last updated by anonymous at May 16 2020, 8:57:21 PM.

FINISHED

```
%pyspark
df = spark.read.csv('s3://aws-logs-061895363576-us-east-1/mundial/time_series_covid19_confirmed_
```

Took 0 sec. Last updated by anonymous at May 16 2020, 8:57:36 PM.

FINISHED

```
%pyspark
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, 8:57:39 PM.

FINISHED

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

```
|-- 1/31/20: string (nullable = true)
|-- 2/1/20: string (nullable = true)
|-- 2/2/20: string (nullable = true)
|-- 2/3/20: string (nullable = true)
|-- 2/4/20: string (nullable = true)
|-- 2/5/20: string (nullable = true)
|-- 2/6/20: string (nullable = true)
|-- 2/7/20: string (nullable = true)
|-- 2/8/20: string (nullable = true)
|-- 2/9/20: string (nullable = true)
|-- 2/10/20: string (nullable = true)
|-- 2/11/20: string (nullable = true)
|-- 2/12/20: string (nullable = true)
|-- 2/13/20: string (nullable = true)
|-- 2/14/20: string (nullable = true)
|-- 2/15/20: string (nullable = true)
|-- 2/16/20: string (nullable = true)
|-- 2/17/20: string (nullable = true)
```

Took 0 sec. Last updated by anonymous at May 16 2020, 8:57:45 PM.

%pyspark   FINISHED

```
df.head(5)
```

```
[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='1', 2/25/20='1', 2/26/20='1', 2/27/20='1', 2/28/20='1', 2/29/20
='1', 3/1/20='1', 3/2/20='1', 3/3/20='1', 3/4/20='1', 3/5/20='1', 3/6/20='1', 3/7/20='1', 3/8/20
='4', 3/9/20='4', 3/10/20='5', 3/11/20='7', 3/12/20='7', 3/13/20='7', 3/14/20='11', 3/15/20='1
6', 3/16/20='21', 3/17/20='22', 3/18/20='22', 3/19/20='22', 3/20/20='24', 3/21/20='24', 3/22/20
='40', 3/23/20='40', 3/24/20='74', 3/25/20='84', 3/26/20='94', 3/27/20='110', 3/28/20='110', 3/2
9/20='120', 3/30/20='170', 3/31/20='174', 4/1/20='237', 4/2/20='273', 4/3/20='281', 4/4/20='29
9', 4/5/20='349', 4/6/20='367', 4/7/20='423', 4/8/20='444', 4/9/20='484', 4/10/20='521', 4/11/20
='555', 4/12/20='607', 4/13/20='665', 4/14/20='714', 4/15/20='784', 4/16/20='840', 4/17/20='90
6', 4/18/20='933', 4/19/20='996', 4/20/20='1026', 4/21/20='1092', 4/22/20='1176', 4/23/20='127
9', 4/24/20='1351', 4/25/20='1463', 4/26/20='1531', 4/27/20='1703', 4/28/20='1828', 4/29/20='193
9', 4/30/20='2171', 5/1/20='2335', 5/2/20='2469', 5/3/20='2704', 5/4/20='2894', 5/5/20='3224',
5/6/20='3392', 5/7/20='3563', 5/8/20='3778', 5/9/20='4033', 5/10/20='4402', 5/11/20='4687', 5/1
2/20='4963', 5/13/20='5226', 5/14/20='5639', 5/15/20='6053'), Row(Province/State=None, Country/R
```

Took 1 sec. Last updated by anonymous at May 16 2020, 8:57:51 PM.

%pyspark   FINISHED

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

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

```
--------------+-----------------+-----------------+-----------------+-----------------+----
--------------+-----------------+-----------------+-----------------+-----------------+----
--------------+-----------------+-----------------+-----------------+-----------------+-----
-------------+
|summary|Province/State|Country/Region|              Lat|             Long|         1/22/20|
1/23/20|         1/24/20|         1/25/20|         1/26/20|         1/27/20|           1/2
8/20|         1/29/20|         1/30/20|         1/31/20|           2/1/20|           2/2/
```

sparksessionpractice

Took 6 sec. Last updated by anonymous at May 16 2020, 8:58:01 PM.

%pyspark   FINISHED

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

```
+--------------+--------------+-----------------+--------+-------+-------+-------+-------+----
---+-------+-------+-------+-------+-------+-------+------+-------+-------+------+-----+---
---+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-----+----
--+-------+-------+-------+-------+-------+-------+-------+-------+-------+------+------+
------+------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-----
--+------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-----+-----
--+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+------+---
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+------+-----
+------+-------+-------+-------+-------+-------+-------+-------+-------+------+------+----
---+-------+------+-------+-------+-------+-------+-------+-------+------+------+
|Province/State|Country/Region|              Lat|            Long|1/22/20|1/23/20|1/24/20|1/25/20|1/2
6/20|1/27/20|1/28/20|1/29/20|1/30/20|1/31/20|2/1/20|2/2/20|2/3/20|2/4/20|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|2/2
0/20|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|3/3/2
0|3/4/20|3/5/20|3/6/20|3/7/20|3/8/20|3/9/20|3/10/20|3/11/20|3/12/20|3/13/20|3/14/20|3/15/20|3/1
6/20|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|3/2
8/20|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|4/1
0/20|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|4/2
```

Took 1 sec. Last updated by anonymous at May 16 2020, 8:58:16 PM.

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

Took 0 sec. Last updated by anonymous at May 16 2020, 8:58:26 PM.

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

```
+------------------+
|      avg(5/15/20)|
+------------------+
|17076.492481203008|
+------------------+
```

Took 0 sec. Last updated by anonymous at May 16 2020, 8:58:50 PM.

%pyspark   READY