databricksPysparks Basics : column manipulation and Dataframe creation

(http://databricks.com/spark/)

```
import numpy as np
from pyspark.sql import SparkSession
from pyspark import SparkContext
from pyspark.sql.functions import *

#----------------------------------#

sc = SparkContext.getOrCreate()
spark =  SparkSession(sc)



###  Spark Create DataFrame with Examples.


#1. Spark Create DataFrame with Examples

columns  =  ('languages','users')
data = (('java','2000'),('python','30000'))
rdd  = sc.parallelize(data)

df_rdd = rdd.toDF()
df_rdd.printSchema()
```

```
root
 |-- _1: string (nullable = true)
 |-- _2: string (nullable = true)
```

```
# 2. Create DataFrame from Dictionary

data_map = { 'language':['python','java','c#'],
         'user':['30k','20k','10k'],
        'spped': ['2x','4x','8x']}


map_1 = [(k,)+(v,) for k,v in data_map.items()]


df = spark.createDataFrame(map_1,['key','val'])

df.show()
```

```
+--------+------------------+
|     key|               val|
+--------+------------------+
|language|[python, java, c#]|
|    user|   [30k, 20k, 10k]|
|   spped|      [2x, 4x, 8x]|
+--------+------------------+
```

```
from databricks import koalas as ks
df = ks.DataFrame(data_map)
sdf  = df.to_spark()
sdf.show()
```

```
+--------+----+-----+
|language|user|spped|
+--------+----+-----+
|  python| 30k|   2x|
|    java| 20k|   4x|
|      c#| 10k|   8x|
+--------+----+-----+
```

```
#3. Using createDataFrame() with the Row type

from pyspark.sql.types import StructType,StructField,IntegerType,StringType

data = [(1,'12102021','13102021'),
       (2,'11112021','21112021'),
       (3,'2102021','15102021')]
schema = StructType([
    StructField("id", IntegerType(), True),
    StructField("created_at", StringType(), True),
    StructField("updated_at", StringType(), True)
])

df2  = spark.createDataFrame(data= data,schema  = schema)
df2.show()
```

```
+---+----------+----------+
| id|created_at|updated_at|
+---+----------+----------+
|  1|  12102021|  13102021|
|  2|  11112021|  21112021|
|  3|   2102021|  15102021|
+---+----------+----------+
```

```
#4.  Create Spark DataFrame from CSV


filename  ='/FileStore/tables/Bowler_data.csv'

df  = spark.read.csv(filename)
df.show()
```

```
+----+-----+----+----+----+----+-----+----+--------------+-------------+------
-----+----------+--------------+---------+
| _c0|  _c1| _c2| _c3| _c4| _c5|  _c6| _c7|           _c8|          _c9|
_c10|      _c11|          _c12|     _c13|
+----+-----+----+----+----+----+-----+----+--------------+-------------+------
-----+----------+--------------+---------+
|null|Overs|Mdns|Runs|Wkts|Econ|  Ave|  SR|    Opposition|       Ground| Start
Date|  Match_ID|        Bowler|Player_ID|
|   1|  8.0|   0|  57|   0|7.12|    -|   -|       v India|       Nagpur|18 Dec
2009|ODI # 2933|Suranga Lakmal|    49619|
|   2| 10.0|   0|  55|   2|5.50|27.50|30.0|       v India|      Kolkata|24 Dec
2009|ODI # 2935|Suranga Lakmal|    49619|
|   3|    -|   -|   -|   -|   -|    -|   -|       v India|        Delhi|27 Dec
```

```
2009|ODI # 2936|Suranga Lakmal|    49619|
|   4|  9.0|   1|  63|   2|7.00|31.50|27.0|  v Bangladesh|        Dhaka| 4 Jan
2010|ODI # 2937|Suranga Lakmal|    49619|
|   5|  8.0|   1|  48|   0|6.00|    -|   -|       v India|        Dhaka| 5 Jan
2010|ODI # 2938|Suranga Lakmal|    49619|
|   6| 10.0|   0|  75|   0|7.50|    -|   -|       v India|        Dhaka|10 Jan
2010|ODI # 2941|Suranga Lakmal|    49619|
|   7|  7.0|   0|  52|   2|7.42|26.00|21.0|     v England|     The Oval|28 Jun
```

```
'''Other fomat supported
Xml
json
text
tsv
avro
parquet
 HBase
jdbc
hive
'''
```

```
Out[106]: 'Other fomat supported \nXml\njson\ntext\ntsv\navro\nparquet\n HBase
\njdbc\nhive\n'
```

```
#B. Spark withColumnRenamed to Rename Column
```

```
filename  ='/FileStore/tables/Bowler_data.csv'

df  = spark.read.load(filename,format ='csv',header="true")
df.show()
```

```
+---+-----+----+----+----+----+-----+----+-------------+-------------+-------
----+----------+--------------+---------+
|_c0|Overs|Mdns|Runs|Wkts|Econ|  Ave|  SR|    Opposition|       Ground| Start
Date|  Match_ID|        Bowler|Player_ID|
+---+-----+----+----+----+----+-----+----+-------------+-------------+-------
----+----------+--------------+---------+
|  1|  8.0|   0|  57|   0|7.12|    -|   -|       v India|       Nagpur|18 Dec
2009|ODI # 2933|Suranga Lakmal|    49619|
|  2| 10.0|   0|  55|   2|5.50|27.50|30.0|       v India|      Kolkata|24 Dec
2009|ODI # 2935|Suranga Lakmal|    49619|
|  3|    -|   -|   -|   -|   -|    -|   -|       v India|        Delhi|27 Dec
```

```
2009|ODI # 2936|Suranga Lakmal|    49619|
|  4|  9.0|   1|  63|   2|7.00|31.50|27.0|  v Bangladesh|        Dhaka| 4 Jan
2010|ODI # 2937|Suranga Lakmal|    49619|
|  5|  8.0|   1|  48|   0|6.00|    -|   -|       v India|        Dhaka| 5 Jan
2010|ODI # 2938|Suranga Lakmal|    49619|
|  6| 10.0|   0|  75|   0|7.50|    -|   -|       v India|        Dhaka|10 Jan
2010|ODI # 2941|Suranga Lakmal|    49619|
|  7|  7.0|   0|  52|   2|7.42|26.00|21.0|     v England|     The Oval|28 Jun
2011|ODI # 3165|Suranga Lakmal|    49619|
```

```
#rename column

df  = df.withColumnRenamed("Econ","Economy")



# multiple columns

df  = df.withColumnRenamed("Mdns","Maidens") \
        .withColumnRenamed('Opposition','Opponent')

df.show()
```

```
+---+-----+-------+----+----+-------+-----+----+--------------+-------------+-
----------+----------+--------------+---------+
|_c0|Overs|Maidens|Runs|Wkts|Economy|  Ave|  SR|      Opponent|       Ground|
Start Date|  Match_ID|        Bowler|Player_ID|
+---+-----+-------+----+----+-------+-----+----+--------------+-------------+-
----------+----------+--------------+---------+
|  1|  8.0|      0|  57|   0|   7.12|    -|   -|       v India|       Nagpur|1
8 Dec 2009|ODI # 2933|Suranga Lakmal|    49619|
|  2| 10.0|      0|  55|   2|   5.50|27.50|30.0|       v India|      Kolkata|2
4 Dec 2009|ODI # 2935|Suranga Lakmal|    49619|
|  3|    -|      -|   -|   -|      -|    -|   -|       v India|        Delhi|2
7 Dec 2009|ODI # 2936|Suranga Lakmal|    49619|
|  4|  9.0|      1|  63|   2|   7.00|31.50|27.0|  v Bangladesh|        Dhaka|
4 Jan 2010|ODI # 2937|Suranga Lakmal|    49619|
|  5|  8.0|      1|  48|   0|   6.00|    -|   -|       v India|        Dhaka|
5 Jan 2010|ODI # 2938|Suranga Lakmal|    49619|
|  6| 10.0|      0|  75|   0|   7.50|    -|   -|       v India|        Dhaka|1
0 Jan 2010|ODI # 2941|Suranga Lakmal|    49619|
|  7|  7.0|      0|  52|   2|   7.42|26.00|21.0|     v England|     The Oval|2
8 Jun 2011|ODI # 3165|Suranga Lakmal|    49619|
|  8|  7.5|      0|  43|   3|   5.48|14.33|15.6|     v England|        Leeds|
```

```
df = df.withColumn('venue',col('Ground'))
df.show()
```

```
+---+-----+-------+----+----+-------+-----+----+-------------+-------------+-
---------+----------+-------------+---------+-------------+
|_c0|Overs|Maidens|Runs|Wkts|Economy|  Ave|  SR|      Opponent|       Ground|
Start Date|  Match_ID|       Bowler|Player_ID|        venue|
+---+-----+-------+----+----+-------+-----+----+-------------+-------------+-
---------+----------+-------------+---------+-------------+
|  1|  8.0|      0|  57|   0|   7.12|    -|   -|       v India|       Nagpur|1
8 Dec 2009|ODI # 2933|Suranga Lakmal|    49619|       Nagpur|
|  2| 10.0|      0|  55|   2|   5.50|27.50|30.0|       v India|      Kolkata|2
4 Dec 2009|ODI # 2935|Suranga Lakmal|    49619|      Kolkata|
|  3|    -|      -|   -|   -|      -|    -|   -|       v India|        Delhi|2
7 Dec 2009|ODI # 2936|Suranga Lakmal|    49619|        Delhi|
|  4|  9.0|      1|  63|   2|   7.00|31.50|27.0|  v Bangladesh|        Dhaka|
4 Jan 2010|ODI # 2937|Suranga Lakmal|    49619|        Dhaka|
|  5|  8.0|      1|  48|   0|   6.00|    -|   -|       v India|        Dhaka|
5 Jan 2010|ODI # 2938|Suranga Lakmal|    49619|        Dhaka|
|  6| 10.0|      0|  75|   0|   7.50|    -|   -|       v India|        Dhaka|1
0 Jan 2010|ODI # 2941|Suranga Lakmal|    49619|        Dhaka|
|  7|  7.0|      0|  52|   2|   7.42|26.00|21.0|     v England|     The Oval|2
8 Jun 2011|ODI # 3165|Suranga Lakmal|    49619|     The Oval|
|  8|  7.5|      0|  43|   3|   5.48|14.33|15.6|     v England|        Leeds|
```

```
df  = df.drop('Ground')
df = df.withColumn('Ground',col('Venue'))
df.show()
```

```
+---+-----+-------+----+----+-------+-----+----+-------------+-----------+---
-------+--------------+---------+-------------+-------------+
|_c0|Overs|Maidens|Runs|Wkts|Economy|  Ave|  SR|      Opponent| Start Date|  M
atch_ID|        Bowler|Player_ID|        venue|       Ground|
+---+-----+-------+----+----+-------+-----+----+-------------+-----------+---
-------+--------------+---------+-------------+-------------+
|  1|  8.0|      0|  57|   0|   7.12|    -|   -|       v India|18 Dec 2009|ODI
# 2933|Suranga Lakmal|    49619|       Nagpur|       Nagpur|
|  2| 10.0|      0|  55|   2|   5.50|27.50|30.0|       v India|24 Dec 2009|ODI
# 2935|Suranga Lakmal|    49619|      Kolkata|      Kolkata|
|  3|    -|      -|   -|   -|      -|    -|   -|       v India|27 Dec 2009|ODI
# 2936|Suranga Lakmal|    49619|        Delhi|        Delhi|
|  4|  9.0|      1|  63|   2|   7.00|31.50|27.0|  v Bangladesh| 4 Jan 2010|ODI
# 2937|Suranga Lakmal|    49619|        Dhaka|        Dhaka|
|  5|  8.0|      1|  48|   0|   6.00|    -|   -|       v India| 5 Jan 2010|ODI
# 2938|Suranga Lakmal|    49619|        Dhaka|        Dhaka|
|  6| 10.0|      0|  75|   0|   7.50|    -|   -|       v India|10 Jan 2010|ODI
# 2941|Suranga Lakmal|    49619|        Dhaka|        Dhaka|
```

```
|  7|  7.0|      0|  52|   2|   7.42|26.00|21.0|     v England|28 Jun 2011|ODI
# 3165|Suranga Lakmal|    49619|     The Oval|     The Oval|
|  8|  7.5|      0|  43|   3|   5.48|14.33|15.6|     v England| 1 Jul 2011|ODI
```

```
df1= df.select(df["Match_ID"].alias("match_id"),col("*"))
df1.show()
```

```
+----------+---+-----+-------+----+----+-------+-----+----+--------------+----
-------+----------+--------------+---------+-------------+-------------+
|  match_id|_c0|Overs|Maidens|Runs|Wkts|Economy|  Ave|  SR|      Opponent| Sta
rt Date|  Match_ID|        Bowler|Player_ID|        venue|       Ground|
+----------+---+-----+-------+----+----+-------+-----+----+--------------+----
-------+----------+--------------+---------+-------------+-------------+
|ODI # 2933|  1|  8.0|      0|  57|   0|   7.12|    -|   -|       v India|18 D
ec 2009|ODI # 2933|Suranga Lakmal|    49619|       Nagpur|       Nagpur|
|ODI # 2935|  2| 10.0|      0|  55|   2|   5.50|27.50|30.0|       v India|24 D
ec 2009|ODI # 2935|Suranga Lakmal|    49619|      Kolkata|      Kolkata|
|ODI # 2936|  3|    -|      -|   -|   -|      -|    -|   -|       v India|27 D
ec 2009|ODI # 2936|Suranga Lakmal|    49619|        Delhi|        Delhi|
|ODI # 2937|  4|  9.0|      1|  63|   2|   7.00|31.50|27.0|  v Bangladesh| 4 J
an 2010|ODI # 2937|Suranga Lakmal|    49619|        Dhaka|        Dhaka|
|ODI # 2938|  5|  8.0|      1|  48|   0|   6.00|    -|   -|       v India| 5 J
an 2010|ODI # 2938|Suranga Lakmal|    49619|        Dhaka|        Dhaka|
|ODI # 2941|  6| 10.0|      0|  75|   0|   7.50|    -|   -|       v India|10 J
an 2010|ODI # 2941|Suranga Lakmal|    49619|        Dhaka|        Dhaka|
|ODI # 3165|  7|  7.0|      0|  52|   2|   7.42|26.00|21.0|     v England|28 J
un 2011|ODI # 3165|Suranga Lakmal|    49619|     The Oval|     The Oval|
|ODI # 3167|  8|  7.5|      0|  43|   3|   5.48|14.33|15.6|     v England| 1 J
```

```
df_test = spark.createDataFrame([("x", 1), ("y", 2)],
                                ["col_1", "col_2"])


# Approach - 1 : using withColumnRenamed function.
df_test.withColumnRenamed("col_1", "col_3").show()
```

```
+-----+-----+
|col_3|col_2|
+-----+-----+
|    x|    1|
|    y|    2|
+-----+-----+
```

```
# Approach - 2 : using alias function.
df_test.select(df_test["col_1"].alias("col5"), "col_2").show()
```

```
+----+-----+
|col5|col_2|
+----+-----+
|   x|    1|
|   y|    2|
+----+-----+
```

```
# Approach - 3 : using selectExpr function.
df_test.selectExpr("col_1 as col_3", "col_2").show()
```

```
+-----+-----+
|col_3|col_2|
+-----+-----+
|    x|    1|
|    y|    2|
+-----+-----+
```

```
+-----+-----+
|col_3|col_2|
+-----+-----+
|    x|    1|
|    y|    2|
+-----+-----+
```