DataFrames and Spark SQL hands-on exercise

In [1]:
```
import findspark
findspark.init()

import pandas as pd
import pyspark
```

In [3]:
```
from pyspark.sql import SparkSession

spark = SparkSession.builder\
        .master("local[*]")\
        .appName('PySpark_Df')\
        .getOrCreate()
```

In [5]:
```
fifa_df = spark.read.csv (r"C:\Users\nobel\OneDrive\Desktop\Spark and data bricks\WorldCupPlayers.csv",
                          inferSchema = True,
                          header = True)
fifa_df.show ()
```

```
+-------+-------+-------------+------------------+-------+------------+-----------------+--------+---------+
|RoundID|MatchID|Team Initials|         Coach Name|Line-up|Shirt Number|      Player Name|Position|    Event|
+-------+-------+-------------+------------------+-------+------------+-----------------+--------+---------+
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|      Alex THEPOT|      GK|     null|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|  Oscar BONFIGLIO|      GK|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0| Marcel LANGILLER|    null|     G40'|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|     Juan CARRENO|    null|     G70'|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Ernest LIBERATI|    null|     null|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|     Rafael GARZA|       C|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Andre MASCHINOT|    null|G43' G87'|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|    Hilario LOPEZ|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Etienne MATTLER|    null|     null|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|   Dionisio MEJIA|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|     Marcel PINEL|    null|     null|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|     Felipe ROSAS|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Alex VILLAPLANE|       C|     null|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|     Manuel ROSAS|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|   Lucien LAURENT|    null|     G19'|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|        Jose RUIZ|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|   Marcel CAPELLE|    null|     null|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|  Alfredo SANCHEZ|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|Augustin CHANTREL|    null|     null|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|   Efrain AMEZCUA|    null|     null|
+-------+-------+-------------+------------------+-------+------------+-----------------+--------+---------+
only showing top 20 rows
```

In [6]:
```
fifa_df.printSchema()
```

```
root
 |-- RoundID: integer (nullable = true)
 |-- MatchID: integer (nullable = true)
 |-- Team Initials: string (nullable = true)
 |-- Coach Name: string (nullable = true)
 |-- Line-up: string (nullable = true)
 |-- Shirt Number: integer (nullable = true)
 |-- Player Name: string (nullable = true)
 |-- Position: string (nullable = true)
 |-- Event: string (nullable = true)
```

In [7]:
```
fifa_df.count()
```

Out[7]:
```
37784
```

In [8]:
```
fifa_df.describe('Position').show()
```

```
+-------+--------+
|summary|Position|
+-------+--------+
|  count|    4143|
|   mean|    null|
| stddev|    null|
|    min|       C|
|    max|     GKC|
+-------+--------+
```

In [9]:
```
fifa_df.select('Player Name','Coach Name').distinct().show()
```

```
+--------------------+--------------------+
|         Player Name|          Coach Name|
+--------------------+--------------------+
|    Arturo FERNANDEZ| BRU Francisco (ESP)|
|Cayetano CARRERAS...|DURAND LAGUNA Jos...|
|  Ernesto MASCHERONI|SUPPICI Alberto (...|
|          Aziz FAHMY|   McREA James (SCO)|
|        Gyula POLGAR|    NADAS Odon (HUN)|
|  Ernesto ALBARRACIN|PASCUCCI Felipe (...|
| Armando CASTELLAZZI|POZZO Vittorio (ITA)|
|     Jaroslav BOUCEK|   PETRU Karel (TCH)|
|           Erwin NYC|  KALUZA Jozef (POL)|
|     Stanislaw BARAN|  KALUZA Jozef (POL)|
|     Fernando ROLDAN|BUCCIARDI Arturo ...|
|            Joe MACA|  JEFFREY Bill (SCO)|
|               INDIO|  MOREIRA Zeze (BRA)|
|      Rene DEREUDDRE|PIBAROT Pierre (FRA)|
|    Anton MALATINSKY|    CEJP Josef (TCH)|
|    Alberto MARIOTTI|LORENZO Juan Carl...|
|  Alfredo DI STEFANO|HERRERA Helenio (...|
|             FIDELIS| FEOLA Vicente (BRA)|
|     Stoyan YORDANOV|BOZHKOV Stefan (BUL)|
|      Wim RIJSBERGEN| MICHELS Rinus (NED)|
+--------------------+--------------------+
only showing top 20 rows
```

In [10]:
```
fifa_df.filter(fifa_df.MatchID =='1096').count()
```

Out[10]:
```
33
```

In [11]:
```
fifa_df.filter((fifa_df.Position == 'C') & (fifa_df.Event=="G40'")).show()
```

```
+-------+-------+-------------+--------------------+-------+------------+---------------+--------+-----+
|RoundID|MatchID|Team Initials|          Coach Name|Line-up|Shirt Number|    Player Name|Position|Event|
+-------+-------+-------------+--------------------+-------+------------+---------------+--------+-----+
|    201|   1089|          PAR|DURAND LAGUNA Jos...|      S|           0|Luis VARGAS PENA|       C| G40'|
|    429|   1175|          HUN|  DIETZ Karoly (HUN)|      S|           0|  Gyorgy SAROSI|       C| G40'|
+-------+-------+-------------+--------------------+-------+------------+---------------+--------+-----+
```

In [12]:
```
fifa_df.createOrReplaceTempView("temp_table")

spark.sql("select * from temp_table where MatchID >= 20").show()
```

```
+-------+-------+-------------+------------------+-------+------------+-----------------+--------+---------+
|RoundID|MatchID|Team Initials|         Coach Name|Line-up|Shirt Number|      Player Name|Position|    Event|
+-------+-------+-------------+------------------+-------+------------+-----------------+--------+---------+
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|      Alex THEPOT|      GK|     null|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|  Oscar BONFIGLIO|      GK|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0| Marcel LANGILLER|    null|     G40'|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|     Juan CARRENO|    null|     G70'|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Ernest LIBERATI|    null|     null|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|     Rafael GARZA|       C|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Andre MASCHINOT|    null|G43' G87'|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|    Hilario LOPEZ|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Etienne MATTLER|    null|     null|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|   Dionisio MEJIA|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|     Marcel PINEL|    null|     null|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|     Felipe ROSAS|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Alex VILLAPLANE|       C|     null|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|     Manuel ROSAS|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|   Lucien LAURENT|    null|     G19'|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|        Jose RUIZ|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|   Marcel CAPELLE|    null|     null|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|  Alfredo SANCHEZ|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|Augustin CHANTREL|    null|     null|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|   Efrain AMEZCUA|    null|     null|
+-------+-------+-------------+------------------+-------+------------+-----------------+--------+---------+
only showing top 20 rows
```

In [ ]:
```
```