Tarea DataFrames PySpark

PROCESAMIENTO DE GRANDES BASES DE DATOS

PÉREZ ROSAS LUIS ALFREDO | 325057368 | 2025-B | GRUPO: 2

Importación de librerias requeridas

In [1]:

```
import findspark
findspark.init()

import pandas as pd
import pyspark
```

Creación de sesion de Spark

In [2]:

```
from pyspark.sql import SparkSession

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

Lectura de archivo .csv, volcado de datos en dataframe y muestra rapida de datos

In [4]:

```
fifa_df = spark.read.csv ("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|   EPIFANIO
SAGUN|      GK|     NULL|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|   ERIBERTO
ROSAS|      GK|     NULL|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|    MARIANA
GONZE|    NULL|     G40'|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|     JUANA A
NDANA|    NULL|     G70'|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0| ERIBERT WIL
LIAMS|    NULL|     NULL|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|     Rafael
GARZA|       C|     NULL|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Andre MASC
HINOT|    NULL|G43' G87'|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|    Hilario
LOPEZ|    NULL|     NULL|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Etienne MA
TTLER|    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 VILLA
PLANE|       C|     NULL|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|     Manuel
ROSAS|    NULL|     NULL|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|   Lucien LA
URENT|    NULL|     G19'|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|        Jose
RUIZ|    NULL|     NULL|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|   Marcel CA
PELLE|    NULL|     NULL|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|  Alfredo SA
NCHEZ|    NULL|     NULL|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|Augustin CHA
NTREL|    NULL|     NULL|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|   Efrain AM
EZCUA|    NULL|     NULL|
+-------+-------+-------------+-------------------+-------+------------+------------
-----+--------+---------+
only showing top 20 rows
```

Muestra de estructura / esquema del dataframe creado

In [5]:

```
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)
```

Conteo de los registros existentes en el dataframe

In [6]:

```
fifa_df.count()
```

Out[6]:

```
37784
```

Estadisticos base de la columna Player Name; al ser una variable cadena no es posible obtener algunos estadisticos

In [9]:

```
fifa_df.describe('Player Name').show()
```

```
+-------+-----------+
|summary|Player Name|
+-------+-----------+
|  count|      37784|
|   mean|       NULL|
| stddev|       NULL|
|    min|   ?URI?I?|
|    max|        �ZIL|
+-------+-----------+
```

Relacion agrupada de Jugadores y Numero de playera

In [10]:

```
fifa_df.select('Player Name','Shirt Number').distinct().show()
```

```
+-------------------+------------+
|        Player Name|Shirt Number|
+-------------------+------------+
|     Marcel CAPELLE|           0|
|     Demetrio NEYRA|           0|
|       Rudolf RAFTL|           0|
|   Sigmund HARINGER|           0|
|           BALTAZAR|           0|
|       Oscar MIGUEZ|           0|
|       Willy KERNEN|          14|
|        Kurt HAMRIN|           7|
|    Rudolf SZANWALD|           1|
|       Kurt SCHMIED|          12|
|      Bobby TRAINOR|          22|
|      Domingo PEREZ|           7|
|     Honorino LANDA|           9|
|     Jimmy ARMFIELD|           2|
|    Anibal TARABINI|          22|
|     Andrej KVASNAK|           6|
|     Staffan TAPPER|          14|
|      Enrique WOLFF|          20|
|Ramon Maria CALDERE|          18|
|             MORATO|          13|
+-------------------+------------+
only showing top 20 rows
```

Filtro de datos de dataframe, cuenta el numero de registros donde RoundID es igual a 201

In [17]:
```
fifa_df.filter(fifa_df.RoundID =='201').count()
```

Out[17]:
```
566
```

Filtro de datos donde posicion es GK y MatchID es 1096

In [28]:
```
fifa_df.filter((fifa_df.Position == 'GK') & (fifa_df.MatchID=="1096")).show()
```

```
+-------+-------+-------------+-------------------+-------+------------+------------
--+--------+-----+
|RoundID|MatchID|Team Initials|         Coach Name|Line-up|Shirt Number|   Player Na
me|Position|Event|
+-------+-------+-------------+-------------------+-------+------------+------------
--+--------+-----+
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|EPIFANIO SAG
UN|      GK| NULL|
|    201|   1096|          MEX|   LUQUE Juan (MEX)|      S|           0|ERIBERTO ROS
AS|      GK| NULL|
+-------+-------+-------------+-------------------+-------+------------+------------
--+--------+-----+
```

Creación de una tabla / vista temporal tomando como fuente de datos el dataframe fifa_df y posterior consulta de tipo SQL donde MatchID es mayor o igual 2000

In [34]:

```
fifa_df.createOrReplaceTempView("temp_table")

spark.sql("select * from temp_table where MatchID >= 2000").show()
```

```
+-------+-------+-------------+-------------------+-------+------------+------------
--------+--------+---------+
|RoundID|MatchID|Team Initials|         Coach Name|Line-up|Shirt Number|         Pla
yer Name|Position|    Event|
+-------+-------+-------------+-------------------+-------+------------+------------
--------+--------+---------+
|    262|   2003|          FRG|SCHOEN Helmut (FRG)|      S|           1|          Se
pp MAIER|      GK|     NULL|
|    262|   2003|          CHI|  ALAMOS Luis (CHI)|      S|           1|   Leopoldo 
VALLEJOS|      GK|     NULL|
|    262|   2003|          FRG|SCHOEN Helmut (FRG)|      S|           2|         Ber
ti VOGTS|    NULL|     NULL|
|    262|   2003|          CHI|  ALAMOS Luis (CHI)|      S|           2|      Roland
o GARCIA|    NULL|      Y1'|
|    262|   2003|          FRG|SCHOEN Helmut (FRG)|      S|           3|       Paul 
BREITNER|    NULL|     G18'|
|    262|   2003|          CHI|  ALAMOS Luis (CHI)|      S|           3|    Alberto 
QUINTANO|    NULL|     NULL|
|    262|   2003|          FRG|SCHOEN Helmut (FRG)|      S|           4|Hans Georg S
CHWAR...|    NULL|     NULL|
|    262|   2003|          CHI|  ALAMOS Luis (CHI)|      S|           4|       Anton
io ARIAS|    NULL|     NULL|
|    262|   2003|          FRG|SCHOEN Helmut (FRG)|      S|           5|   Franz BEC
KENBAUER|       C|     NULL|
|    262|   2003|          CHI|  ALAMOS Luis (CHI)|      S|           5|      Elias 
FIGUEROA|    NULL|     NULL|
|    262|   2003|          FRG|SCHOEN Helmut (FRG)|      S|           8|      Bernd 
CULLMANN|    NULL|     NULL|
|    262|   2003|          CHI|  ALAMOS Luis (CHI)|      S|           6|      Juan R
ODRIGUEZ|    NULL|     O83'|
|    262|   2003|          FRG|SCHOEN Helmut (FRG)|      S|           9|   Juergen G
RABOWSKI|    NULL|     NULL|
|    262|   2003|          CHI|  ALAMOS Luis (CHI)|      S|           7|      Carlos
CASZELY|    NULL|Y13' R67'|
|    262|   2003|          FRG|SCHOEN Helmut (FRG)|      S|          11|       Jupp 
HEYNCKES|    NULL|     NULL|
|    262|   2003|          CHI|  ALAMOS Luis (CHI)|      S|           8|    Francisc
o VALDES|       C|     O76'|
|    262|   2003|          FRG|SCHOEN Helmut (FRG)|      S|          12|    Wolfgang
OVERATH|    NULL|     O75'|
|    262|   2003|          CHI|  ALAMOS Luis (CHI)|      S|           9|      Sergio
AHUMADA|    NULL|     NULL|
|    262|   2003|          FRG|SCHOEN Helmut (FRG)|      S|          13|        Gerd
MUELLER|    NULL|     NULL|
|    262|   2003|          CHI|  ALAMOS Luis (CHI)|      S|          10|      Carlos
REINOSO|    NULL|      Y1'|
+-------+-------+-------------+-------------------+-------+------------+------------
--------+--------+---------+
only showing top 20 rows
```