IE University

Sunday, 25th of October 2020

Spark – MBD Term II by

Professor Raul Marin Perez

Individual Assignment

*Football Player Analysis* by

**Albert Sehnaoui**

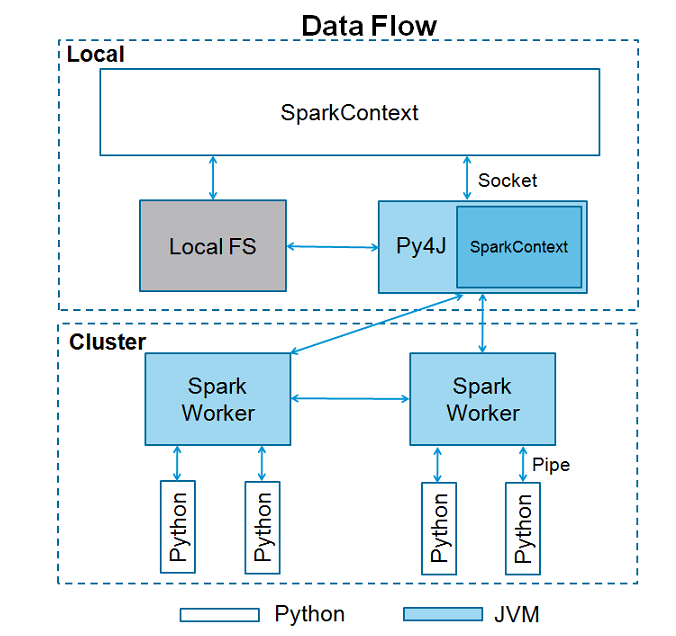


# *Purpose of this analysis:*

In this analysis, I will learn about spark and use some functions I’ve learnt during the course. I have chosen Football Players dataset from ‘Kaggle’ as a dataset. I will analyze players’ ability and team state using Apache Spark. We will dive deep in the understanding of players’ abilities, details, and general performance information.

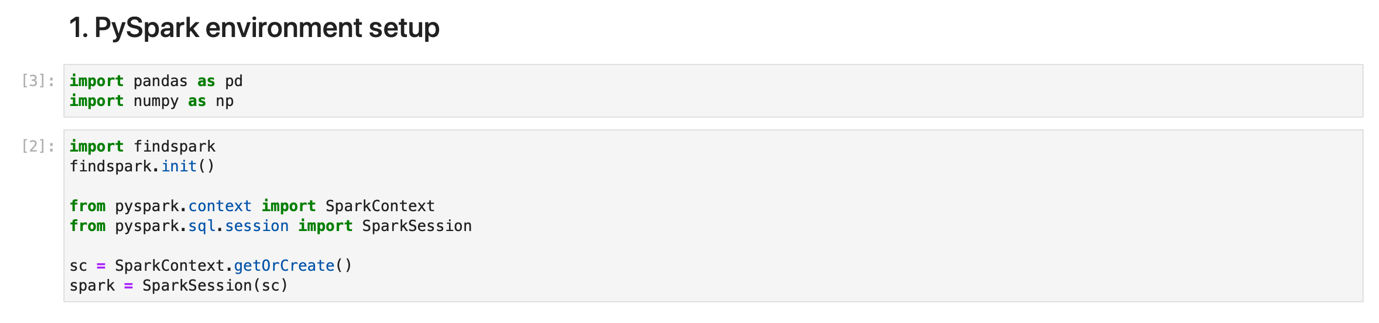
*Outline:*

1. PySpark **Environment Setup**
2. Data source and **Spark Data Abstraction** (DataFrame) **set up**
3. Data set **Metadata Analysis**:
   1. Display Schema and Size of the DataFrame
   2. Get multiple Random Samples from the data set
   3. Get specific One Row from the data set
   4. Columns/fields categorization
4. Columns groups **Basic Profiling** to better understand our data set:
   1. Player General Information columns basic profiling
   2. Argentina Team members
5. Answer some **Business Questions** to improve service:
   1. Sum of IntCaps and IntGoals per Age
   2. Pick Top 10 players with good tackling
   3. Categorize Strikers



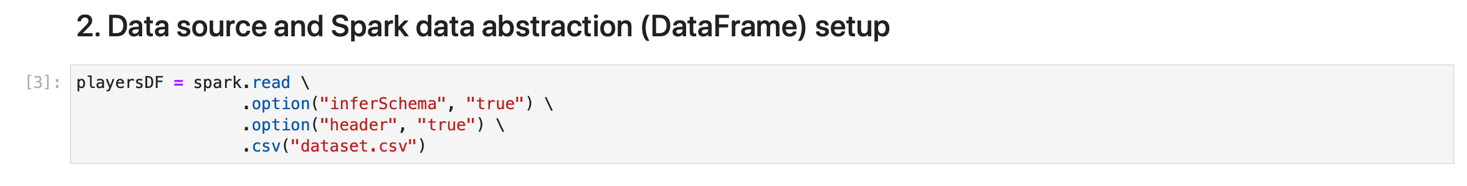
1. PySpark Environment Set-Up:

In this phase, we established our environment through the usage at first of Oracle VM launching it to be able to use Spark properly. Then, in Jupyter notebook with Python, we import *findspark, SparkContext, and SparkSession.* We have to initialize the findspark using findspark.init() function. Then create Spark Context and Session to use spark in python. And it is as easy as this:

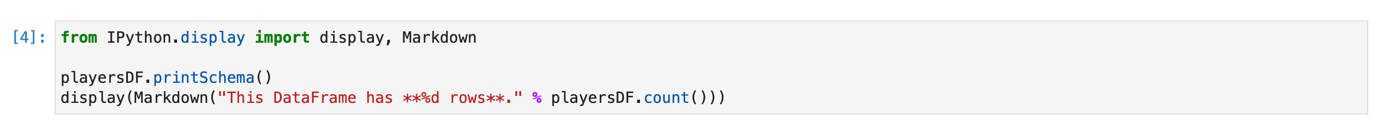


1. Data Source and Spark Data Abstraction:

In this phase, we very simply imported our dataset that we obtained from Kaggle *“Football Manager Data”*. With a *Spark.read function:*



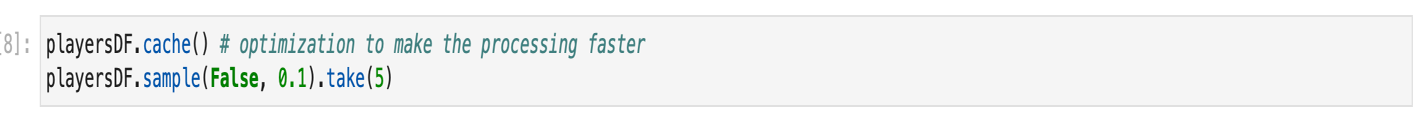
1. Data set metadata analysis:
   1. Display Schema and Size:

In this step, we dive deeper in the skeleton of our dataset, by obtaining the number of rows. We have exactly 159,541 rows. And we understand the type of data we are working with: *Integers, Floats, and Objects.*

And we can get a count of the data frame rows, count () function is used:

* 1. Get multiple Random Samples from the data set:

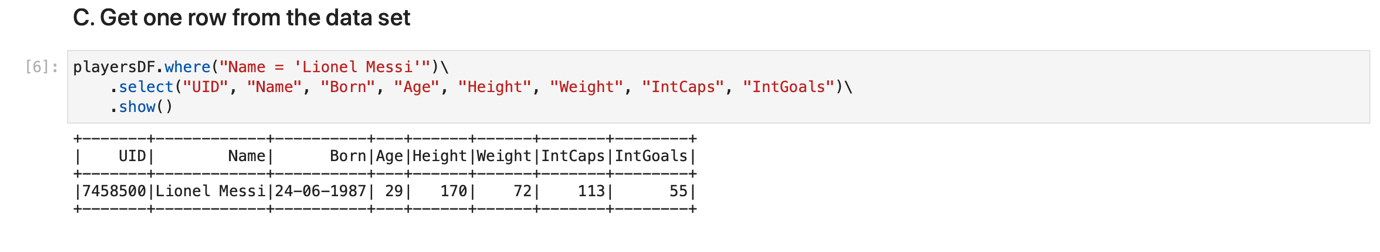
In order to get multiple random samples from the dataset, we have to call cache() function first. This function is for optimization to make the processing faster. After that, call sample().take() to take number of samples from the data set.



In this step, we developed the random samples of our data and attributed each one a value. We ‘digitalized’ our data.

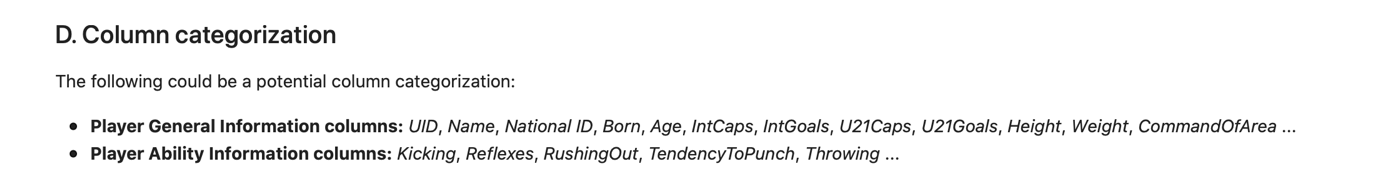
* 1. Get specific One Row from the data set:

Here I get player info for ‘Lionel Messi’ from the data set. We will use where () function to query certain rows and select function to get specific columns. And selected the column we wanted to display about Lionel Messi. The specific information we wanted to obtain from him with select (), and to display it in a table through .show ().



* 1. Columns/Fields Categorization:

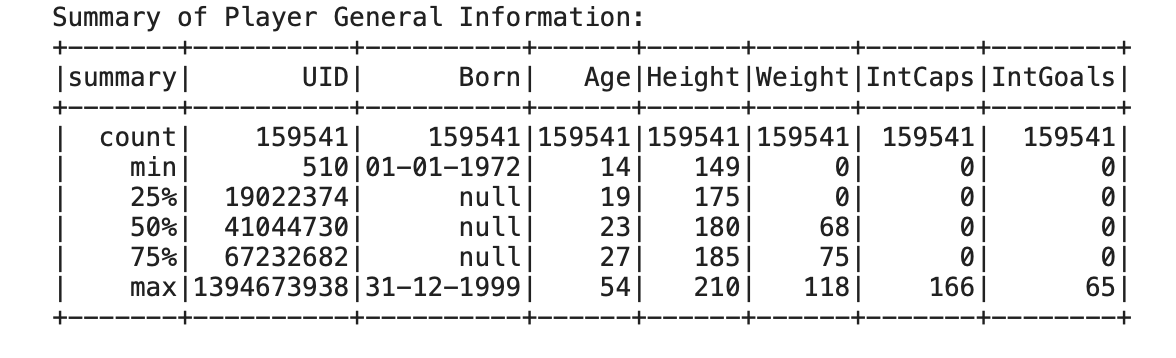
The following is a columns Categorization, defining the columns we will take into consideration and the division. And some detailed insights about the players metrics and focus points.



1. Columns groups Basic Profiling to better understand our data set:
   1. Player General Information columns basic profiling:

In this middle-step from our whole report and notebook, we display general information about our players and the diversity, distribution, and qualifications they show. We obtained six output tables in this step and gained insights from them.

Through select () and summary () functions, we obtained a table that delivered the count, min, max and the quantiles (25-50-75%).



Next, we scouted our data table for null values in order to treat them and we didn’t obtain any null-value:

