PySpark Kết Hợp Với DataFrame

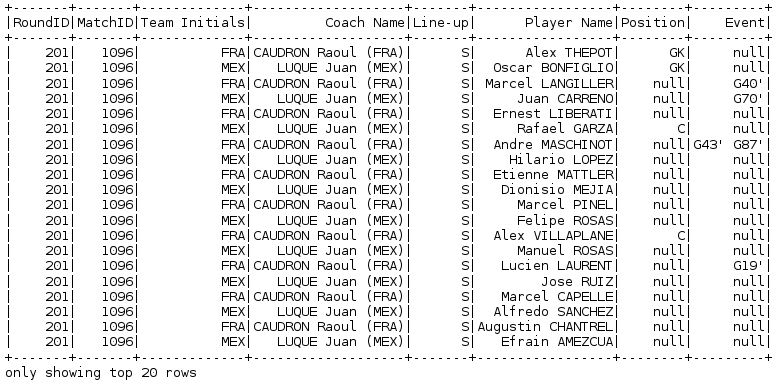
DataFrame trong PySpark được tạo bởi nhiều cách như là có thể được tạo từ các file CSV, JSON, XML,..

Các chức năng cơ bản của DataFrame:

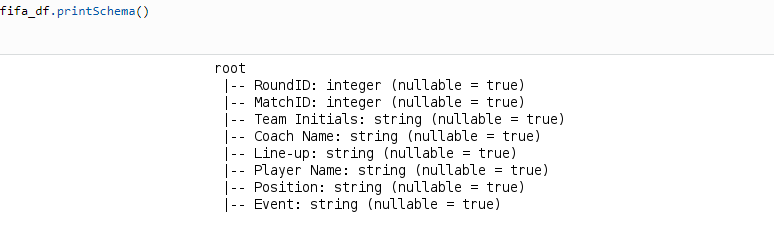
* Đọc file (file csv)



Và đây là kết quả sau khi đọc:

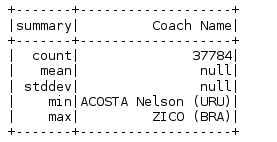


* Schema của DataFrame: Khi chúng ta muốn xem cấu trúc của DataFrame thì ta sử dụng hàm printSchema(). Hàm này sẽ giúp chúng ta biết được các cột dữ liệu là gì và có thể biết được cột đó có thể có giá trị null không.



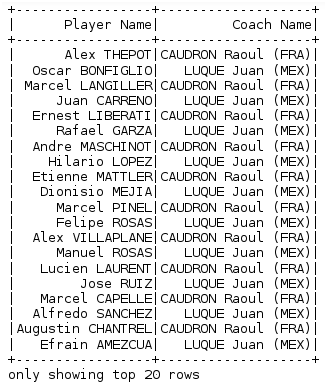
* Miêu tả cụ thể một cột: Nếu bạn muốn thông tin của một cột trong DataFrame thì bạn sẽ dùng hàm describe().





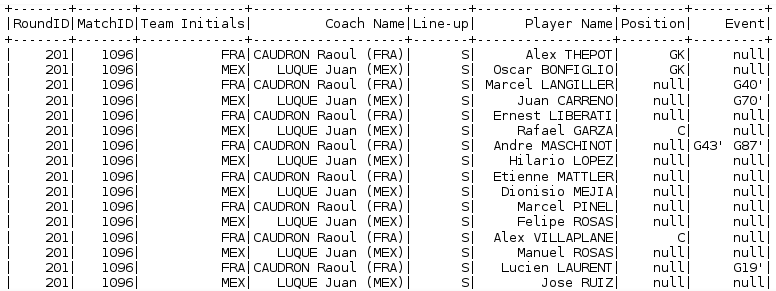
* In ra nhiều cột: Nếu bạn muốn in ra một số cột bất kỳ thì ta sẽ dùng hàm select().





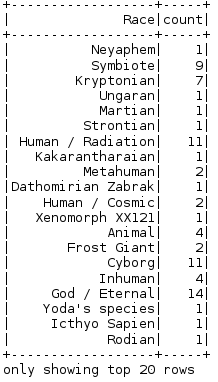
* Filtering Data: Nếu bạn muốn filter data trong DataFrame thì ta sử dụng hàm filter(). Nó sẽ giúp ta filter dữ liệu theo điều kiện mà ta muốn.





* Group Data: Khi bạn muốn nhóm dữ liệu trên một cột nhật định. Dưới đây sẽ là ví dụ về việc group data dựa theo cột Race và ta sẽ đếm có bao nhiêu giá trị trong 1 group.





**Link Tham Khảo:**

https://dzone.com/articles/pyspark-dataframe-tutorial-introduction-to-datafra