I. Spark properties

- Xử lý dữ liệu: Spark xử lý dữ liệu theo lô và thời gian thực
- Tính tương thích: Có thể tích hợp với tất cả các nguồn dữ liệu và định dạng tệp được hỗ trợ bởi cụm Hadoop.
- Hỗ trợ ngôn ngữ: hỗ trợ Java, Scala, Python và R.
- Phân tích thời gian thực:
  . Apache Spark có thể xử lý dữ liệu thời gian thực tức là dữ liệu đến từ các luồng sự kiện thời gian thực với tốc độ hàng triệu sự kiện mỗi giây. Ví dụ: Data Twitter chẳng hạn hoặc lượt chia sẻ, đăng bài trên Facebook. Sức mạnh Spark là khả năng xử lý luồng trực tiếp hiệu quả.
  . Apache Spark có thể được sử dụng để xử lý phát hiện gian lận trong khi thực hiện các giao dịch ngân hàng. Đó là bởi vì, tất cả các khoản thanh toán trực tuyến được thực hiện trong thời gian thực và chúng ta cần ngừng giao dịch gian lận trong khi quá trình thanh toán đang diễn ra.

  - Mục tiêu sử dụng:

  a. Xử lý dữ liệu nhanh và tương tác

  b. Xử lý đồ thị

  c. Công việc lặp đi lặp lại

  d. Xử lý thời gian thực

  e. joining Dataset

  f. Machine Learning

  g. Apache Spark là Framework thực thi dữ liệu dựa trên Hadoop HDFS. Apache Spark không thay thế cho Hadoop nhưng nó là một framework ứng dụng. Apache Spark tuy ra đời sau nhưng được nhiều người biết đến hơn Apache Hadoop vì khả năng xử lý hàng loạt và thời gian thực.

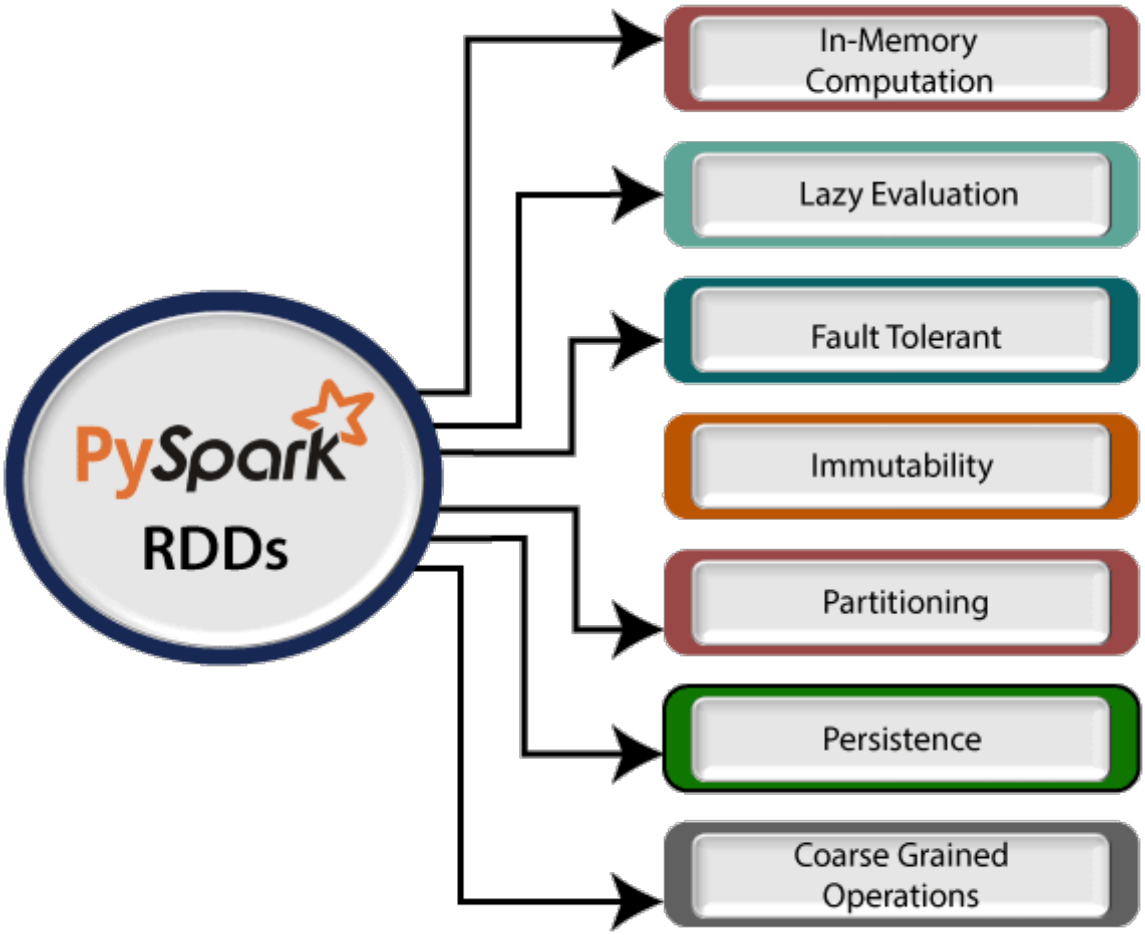
II. Spark RDD

1. Spark RDD là gì?

RDD là viết tắt của ''*Resilient Distribution Datasets*" (bộ dữ liệu phân phối khả năng phục hồi).

Distributed Datasets (RDDs). RDDs hỗ trợ hai kiểu thao tác thao tác: transformations và action. Thao tác chuyển đổi(tranformation) tạo ra dataset từ dữ liệu có sẵn. Thao tác actions trả về giá trị cho chương trình điều khiển (driver program) sau khi thực hiện tính toán trên dataset.

*Các đặc tính của RDD:



Tính toán trong bộ nhớ:
PySpark cung cấp khả năng tính toán trong bộ nhớ. Kết quả được tính toán và lưu trữ trong bộ nhớ phân tán (RAM) thay vì bộ nhớ ổn định (đĩa) và cung cấp tính toán rất nhanh.

Tiến hóa lười biếng:
Chuyển đổi trong PySpark RDDs là lười biếng. Nó không tính toán kết quả ngay lập tức có nghĩa là việc thực thi không bắt đầu cho đến khi một hành động được kích hoạt. Khi chúng ta gọi một số hoạt động trong RDD để chuyển đổi, nó không thực thi ngay lập tức. Lazy Evolution đóng một vai trò quan trọng trong việc tiết kiệm chi phí tính toán. Nó cung cấp sự tối ưu hóa bằng cách giảm số lượng truy vấn.

Khả năng chịu lỗi:
RDD theo dõi thông tin dòng dữ liệu để tự động tạo lại dữ liệu bị mất. Nếu lỗi xảy ra trong bất kỳ phân vùng RDD nào, thì phân vùng đó có thể được tính toán lại từ tập dữ liệu đầu vào chịu lỗi ban đầu để tạo nó.

Bất biến:
Dữ liệu đã tạo có thể được truy xuất bất cứ lúc nào nhưng không thể thay đổi giá trị của nó. RDD chỉ có thể được tạo thông qua các phép toán xác định.

Phân vùng:
RDD là tập hợp các mục dữ liệu khác nhau có kích thước rất lớn. Do kích thước của nó, chúng không thể vừa với một nút duy nhất và phải được phân vùng trên nhiều nút khác nhau.

Sự bền bỉ:
Đó là một kỹ thuật tối ưu hóa mà chúng ta có thể lưu kết quả đánh giá RDD. Nó lưu trữ kết quả trung gian để chúng ta có thể sử dụng thêm nếu cần. Nó làm giảm độ phức tạp của tính toán.

Hoạt động thu được thô:
Hoạt động chi tiết thô có nghĩa là chúng ta có thể chuyển đổi toàn bộ tập dữ liệu nhưng không biến đổi phần tử riêng lẻ trên tập dữ liệu. Mặt khác, chi tiết tốt có nghĩa là chúng ta có thể chuyển đổi từng phần tử trên tập dữ liệu.

2. Mô phỏng RDD

III. Spark DataFrame
1. Spark DataFrame là gì?

Spark DataFrame là phiên bản Spark 1.3. Nó là một tập hợp phân phối dữ liệu được sắp xếp vào các cột được đặt tên. Khái niệm khôn ngoan, nó bằng với bảng trong cơ sở dữ liệu quan hệ hoặc khung dữ liệu trong Python. Chúng ta có thể tạo DataFrame bằng cách sử dụng:

- Tệp dữ liệu có cấu trúc
- Bàn trong tổ ong
- Cơ sở dữ liệu bên ngoài
- Sử dụng RDD hiện có

2. Code ví dụ
    a. Cách tạo DataFrame

```
from pyspark.sql import *

Employee = Row("firstName", "lastName", "email", "salary")

employee1 = Employee('Basher', 'armbrust', 'bash@edureka.co', 100000)
employee2 = Employee('Daniel', 'meng', 'daniel@stanford.edu', 120000 )
employee3 = Employee('Muriel', None, 'muriel@waterloo.edu', 140000 )
employee4 = Employee('Rachel', 'wendell', 'rach_3@edureka.co', 160000 )
employee5 = Employee('Zach', 'galifianakis', 'zach_g@edureka.co', 160000 )

print(Employee[0])

print(employee3)

department1 = Row(id='123456', name='HR')
department2 = Row(id='789012', name='OPS')
department3 = Row(id='345678', name='FN')
department4 = Row(id='901234', name='DEV')
```

```
firstName
Row(firstName='Muriel', lastName=None, email='muriel@waterloo.edu', salary=140000)
```

```
departmentWithEmployees1 = Row(department=department1, employees=[employee1, employee2, employee5])
departmentWithEmployees2 = Row(department=department2, employees=[employee3, employee4])
departmentWithEmployees3 = Row(department=department3, employees=[employee1, employee4, employee3])
departmentWithEmployees4 = Row(department=department4, employees=[employee2, employee3])
```

+ Code + Text

```
departmentsWithEmployees_Seq = [departmentWithEmployees1, departmentWithEmployees2]
dframe = spark.createDataFrame(departmentsWithEmployees_Seq)
display(dframe)
dframe.show()
```

```
DataFrame[department: struct<id:string,name:string>, employees: array<struct<firstName:string,lastNam
+-------------+--------------------+
|   department|           employees|
+-------------+--------------------+
| [123456, HR]|[[Basher, armbrus...|
|[789012, OPS]|[[Muriel,, muriel...|
+-------------+--------------------+
```

b. Lấy dữ liệu từ csv vào DataFrame

```
fifa_df = spark.read.csv("path-of-file/fifa_players.csv",

fifa_df.show()
```

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

```
fifa_df.filter((fifa_df.Position=='C') && (fifa_df.Event=="G4
```

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

```
fifa_df.filter(fifa_df.MatchID=='1096').show()

fifa_df.filter(fifa_df.MatchID=='1096').count()  //to get
```

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

```
fifa_df.orderBy(fifa_df.MatchID).show()
```

```
+-------+-------+-------------+--------------------+-------+------------------+--------+---------+
|RoundID|MatchID|Team Initials|         Coach Name|Line-up|      Player Name|Position|    Event|
+-------+-------+-------------+--------------------+-------+------------------+--------+---------+
|    323|     25|          BRA|LAZARONI Sebastia...|      S|          TAFFAREL|      GK|     null|
|    323|     25|          BRA|LAZARONI Sebastia...|      S|      MAURO GALVAO|    null|Y50' O83'|
|    323|     25|          ARG|BILARDO Carlos (ARG)|      S|  Sergio GOYCOCHEA|      GK|     Y87'|
|    323|     25|          BRA|LAZARONI Sebastia...|      S|          JORGINHO|    null|     null|
|    323|     25|          ARG|BILARDO Carlos (ARG)|      S|     Jose BASUALDO|    null|     null|
|    323|     25|          BRA|LAZARONI Sebastia...|      S|     RICARDO GOMES|       C|     R85'|
|    323|     25|          ARG|BILARDO Carlos (ARG)|      S|  Jorge BURRUCHAGA|    null|     null|
|    323|     25|          BRA|LAZARONI Sebastia...|      S|             DUNGA|    null|     null|
|    323|     25|          ARG|BILARDO Carlos (ARG)|      S|  Claudio CANIGGIA|    null|     G81'|
|    323|     25|          BRA|LAZARONI Sebastia...|      S|            ALEMAO|    null|     O83'|
|    323|     25|          ARG|BILARDO Carlos (ARG)|      S|    Diego MARADONA|       C|     null|
|    323|     25|          BRA|LAZARONI Sebastia...|      S|            BRANCO|    null|     null|
|    323|     25|          ARG|BILARDO Carlos (ARG)|      S|    Ricardo GIUSTI|    null|     Y28'|
|    323|     25|          BRA|LAZARONI Sebastia...|      S|             VALDO|    null|     null|
|    323|     25|          ARG|BILARDO Carlos (ARG)|      S|      Pedro MONZON|    null|     Y27'|
|    323|     25|          BRA|LAZARONI Sebastia...|      S|            CARECA|    null|     null|
|    323|     25|          ARG|BILARDO Carlos (ARG)|      S|Julio OLARTICOECHEA|    null|     null|
|    323|     25|          BRA|LAZARONI Sebastia...|      S|            MULLER|    null|     null|
|    323|     25|          ARG|BILARDO Carlos (ARG)|      S|     Oscar RUGGERI|    null|     null|
|    323|     25|          BRA|LAZARONI Sebastia...|      S|     RICARDO ROCHA|    null|     Y40'|
+-------+-------+-------------+--------------------+-------+------------------+--------+---------+
```

IV. Nguồn tham khảo:

https://helpex.vn/article/huong-dan-pyspark-dataframe-gioi-thieu-ve-dataframes-5c6b21e6ae03f628d053c29e