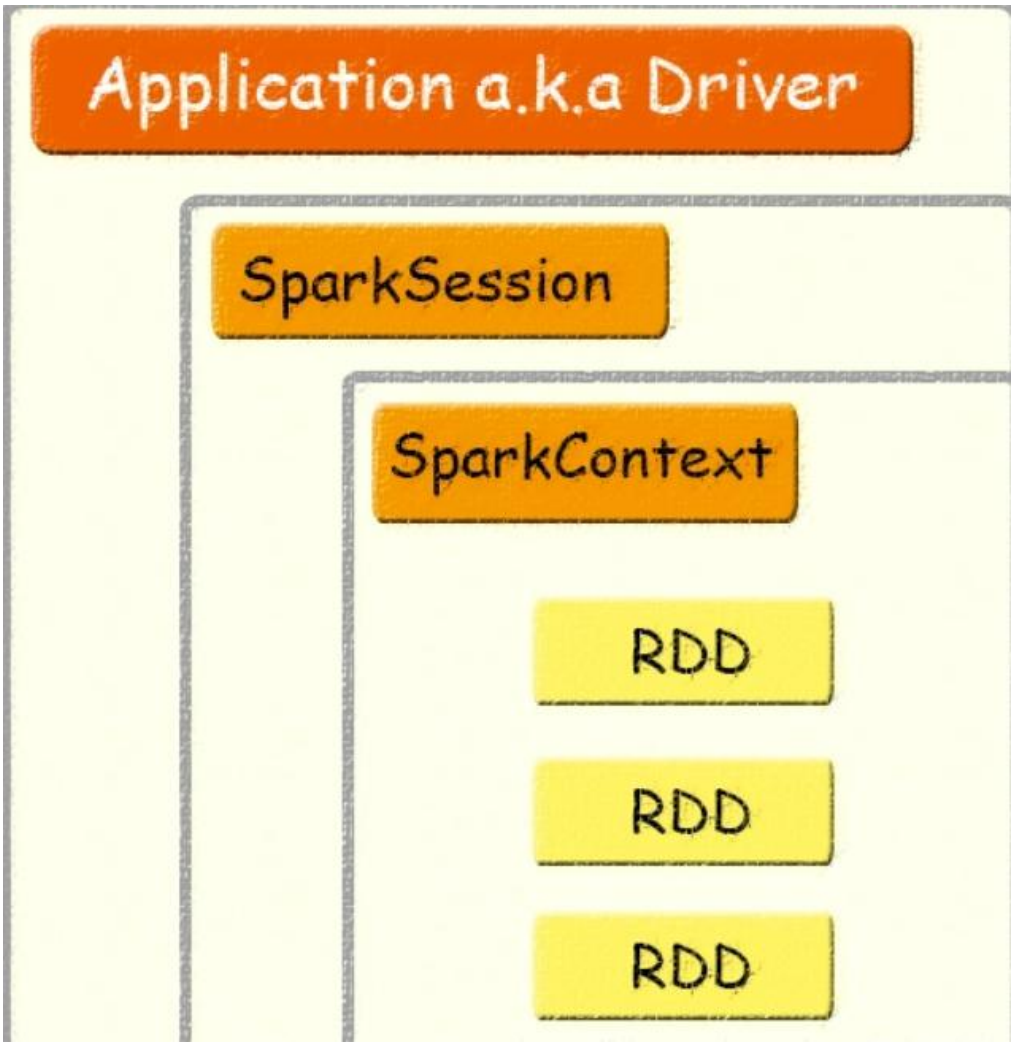
Spark - SparkSession

Spark session: Đại diện cho khả năng tương tác với executors trong 1 chương trình. Spark session chính là entry point của mọi chương trình Spark. Từ SparkSession, có thể tạo RDD/ DataFrame/ DataSet, thực thi SQL... từ đó thực thi tính toán phân tán.



Spark Session bao gồm tất cả các API context có sẵn như:

- Spark Context
- SQL Context
- Streaming Context
- Hive Context

Khởi tạo SparkSession trong Spark-shell

Mặc định, Spark shell cung cấp một `spark` object, có thể sử dụng để khởi tạo các biến:

```
scala> val sqlcontext = spark.sqlContext
```

Khởi tạo SparkSession trong chương trình Scala

Để khởi tạo SparkSession trong Scala hoặc Python, bạn cần sử dụng phương thức builder() và getOrCreate():

```
val spark = SparkSession.builder()
      .master("local[1]")
      .appName("Laptrinh.vn")
      .getOrCreate();
```

`master()` - Nếu bạn đang chạy trên cluster, bạn cần sử dụng master name như một đối số của master(), thông thường nó là `yarn` hoặc `mesos`

- Sử dụng `local[x]` khi chạy ở chế độ standalone, x (số int > 0) - là số partition được tạo khi sử dụng RDD, DataFrame, and Dataset. Lý tưởng nhất, x là số CPU core bạn có.

`appName()` - Tên của application.

`getOrCreate()` - Return a SparkSession object nếu đã tồn tại, ngược lại sẽ tạo mới.

SparkSesssion Method

STT	Method	Mô tả
1	version	Return Spark version
2	builder	Khởi tạo new SparkSession
3	createDataFrame	Khởi tạo DataFrame từ collection và RDD
4	createDataset	Khởi tạo Dataset từ DataFrame, Collection và RDD
5	emptyDataFrame	Khởi tạo một empty DataFrame
6	emptyDataset	Khởi tạo một empty Dataset
7	getActiveSession	Return một active SparkSession
8	implicits	Truy cập một nested Scala object

9	read	Trả về một thể hiện của lớp DataFrameReader, được sử dụng để đọc các bản ghi từ csv, parquet, avro và các định dạng tệp khác vào DataFrame.
10	readStream	Trả về một thể hiện của lớp DataStreamReader, nó được sử dụng để đọc dữ liệu truyền trực tuyến, có thể được sử dụng để đọc dữ liệu truyền trực tuyến vào DataFrame.
11	sparkContext	Trả về một SparkContext.
12	sql	Trả về DataFrame sau khi thực thi SQL được đề cập.
13	sqlContext	Trả về SQLContext.
14	stop	Dừng SparkContext hiện tại.
15	table	Trả về DataFrame của một bảng hoặc dạng xem.
16	udf	Tạo một UDF Spark để sử dụng nó trên DataFrame, Dataset và SQL.

Revision #1
Created 13 February 2021 05:34:16 by Laptrinh.vn
Updated 13 February 2021 06:04:37 by Laptrinh.vn