Querying Data with DataFrames (Part 1)

Justin Pihony

@JustinPihony | justin-pihony.blogspot.com

Course Overview

DataFrames

Datasets

Spark Streaming

Optimizing Towards Fast Data

Module Overview

DataFrames

- Operations
- Functions
- Flattening

Spark SQL: The Future of Spark

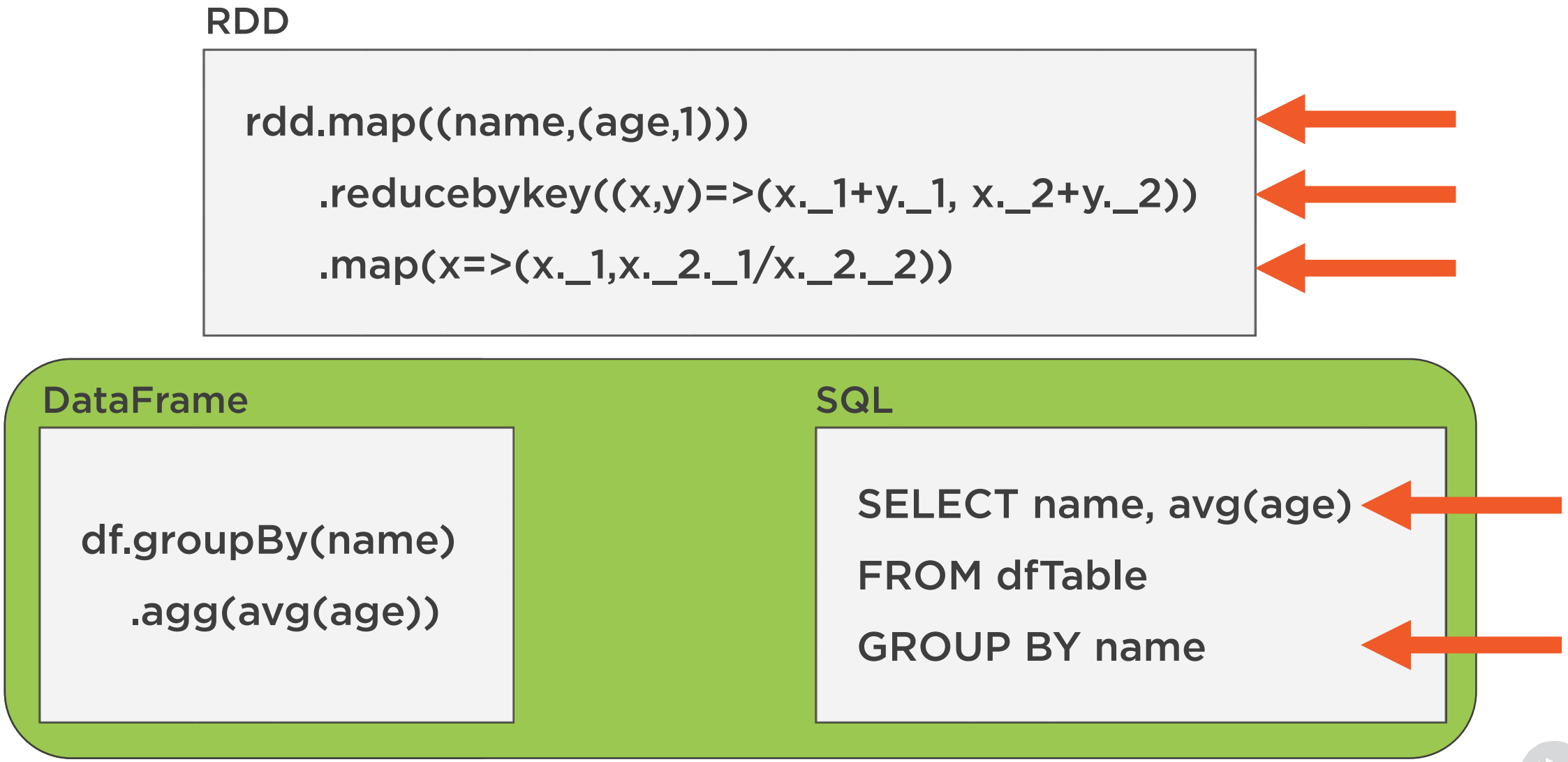
DataFrames

DataFrame

A table, or two-dimensional array-like structure, in which each column contains measurements on one variable, and each row contains one case.

https://stackoverflow.com/questions/31508083/difference-between-dataframe-and-rdd-in-spark

Codability++


RDD
rdd.map((name,(age,1)))
.reducebykey((x,y)=>(x._1+y._1, x._2+y._2))
.map(x=>(x._1,x._2._1/x._2._2))
DataFrame
df.groupBy(name)
.agg(avg(age))
SQL
SELECT name, avg(age)
FROM dfTable
GROUP BY name

Catalyst Optimizer

“

...the future of Spark performance, with more efficient storage options, advanced optimizers, and direct operations on serialized data.

-High Performance Spark by Holden Karau & Rachel Warren

Optimizations

```
(x: String) => x.endsWith("post")
```

```
class generated {
  def apply(String): Boolean
}
```

```
col("x").endsWith("post")
```

```
EndsWith(x, Lit("post"))
```

Faster, Smaller, Smarter

Windows Note

```
ıp\hive
)d -R 777 \tmp\hive
```

ly/2twRUVA

RDD to DataFrames Made Easy

RDD-like DataFrame Methods

- Actions
  - count
  - first/head/take(AsList)
- Transformations
  - except/intersect/union(All)
  - sort(WithinPartitions)/orderBy
  - sample/randomSplit
- Operations
  - map(Partitions)/flatMap/foreach(Partition)
- checkpoint (2.1)

Querying with SQL

SQL in 2.0

SQL 2003

Subquery support

Native SQL parser

Introduction to SQL

https://www.pluralsight.com/courses/introduction-to-sql

Summary

The DataFrame API

Functions

Exploding arrays