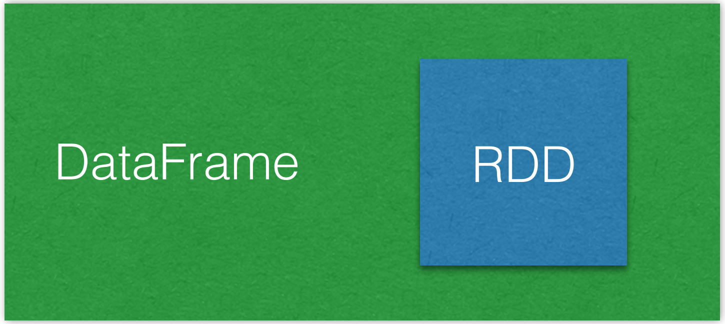
In this Video you will learn…

ApacheSparkSQL

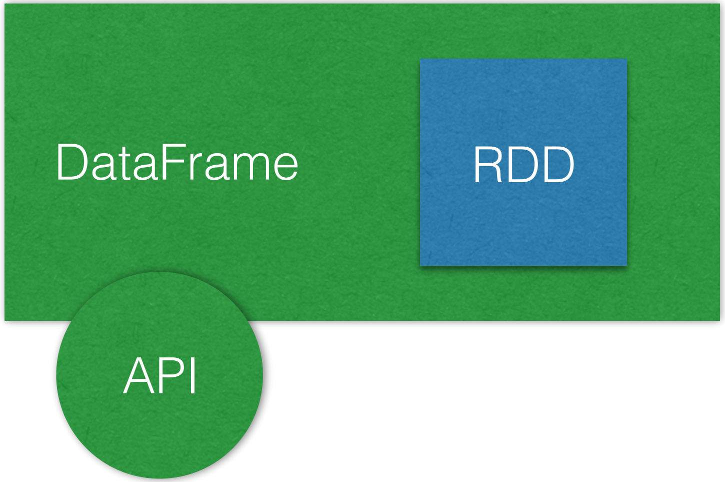
ApacheSparkSQL

RDD

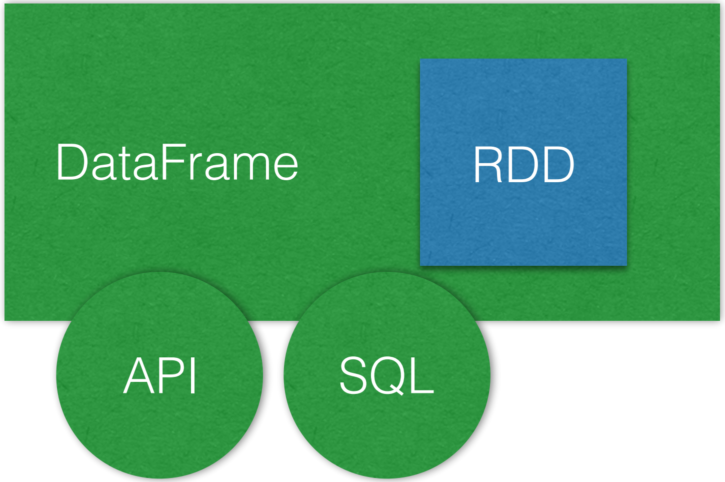
ApacheSparkSQL



ApacheSparkSQL



ApacheSparkSQL



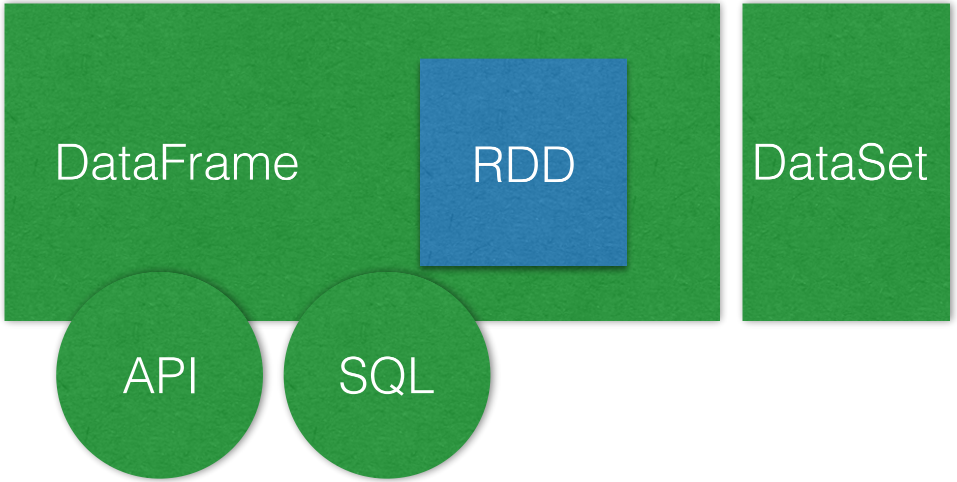
DataFrame

RDD

API

SQL

ApacheSparkSQL



Schemas

- RDDs are schema less (schema on read)
- DataFrames have a schema
    - lazy, inferred
    - explicitly defined

The “Catalyst”

- Creates “logical execution plan” (LEP) from SQL
- Optimises LEP to “physical execution plans” (PEPs)
- based on statistics chooses best PEP to execute
- similar to cost based optimisers in RDBMs

Project Tungsten

- Java Virtual Machine (JVM) is an art piece
- General purpose byte code execution engine
- JVM objects & Garbage Collection (GC) overhead
    - 4 byte string is 48 byte on the JVM
    - GC optimises on object life time estimation
    - Spark knows this better than JVM

Project Tungsten

- L1/L2/L3 Cache friendly data structures
- Code generation to remove
  - boxing of primitive types
  - polymorphic function dispatching

Summary

- ApacheSpark supports SQL via data frame API
- Internally still RDDs are used
- Makes writing ApacheSpark jobs easier
- Performance benefits through Catalyst & Tungsten

The next module covers…

End to End Scenario