Low Level API | High Level API

RDD, | Dataframe, Dataset

1.x | 2.x - 2.1, 2.4, 2.x
1.5/1.6

Not used

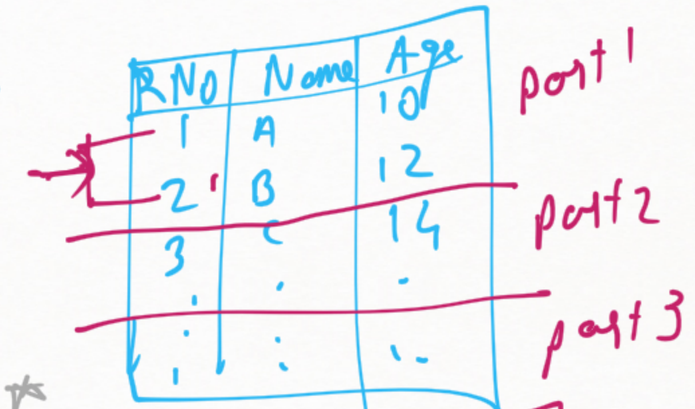
- Legacy Code
- Tight Control

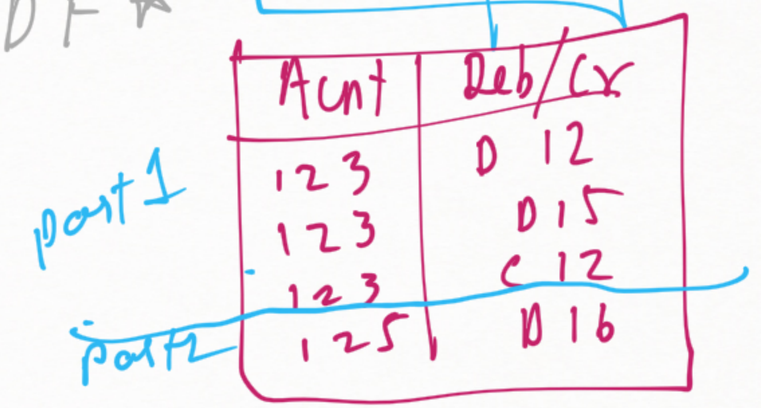
(RDD)

* DF → RDD (Tight Control) → DF *

Bank

- DF & Dataset is on top of RDD





RDD ⇒ Resilient distributed Dataset

Resilient → Recovery quickly

distributed → spread across nodes

→ Immutable

Val

→ Partitioned

Var ⇒ var a = 5, a = 6 ✓ { • Mutable • can change content }

Val ⇒ val a = 5, a = 6 ✗ { • Immutable • can not change content }

Spark

→ 1.x / Not much used 2.x, 3.x

→ Programmer has to take care of Optimizat^n

Optimizat^n

RDD A RDD B

join (2)

filter (1)

Transformation :
* Transform / Modify
* filter (where)

Lazy.
* Transformation will not be executed until you execute action

Action :
* Causes execution of Transformation
* Save to a file / Display on Terminal.