RDD, DataFrame e DataSet

RDD – Resilient Distributed Datasets (RDD)

- Estrutura básica de baixo nível
- Dados “imutáveis”, distribuídos pelo cluster
- Em memória
- Pode ser persistindo em disco
- Tolerante a falha
- Operações sobre um RDD criam um novo RDD

RDD

Estrutura de baixo nível

Complexo e verboso

Otimização difícil pelo Spark

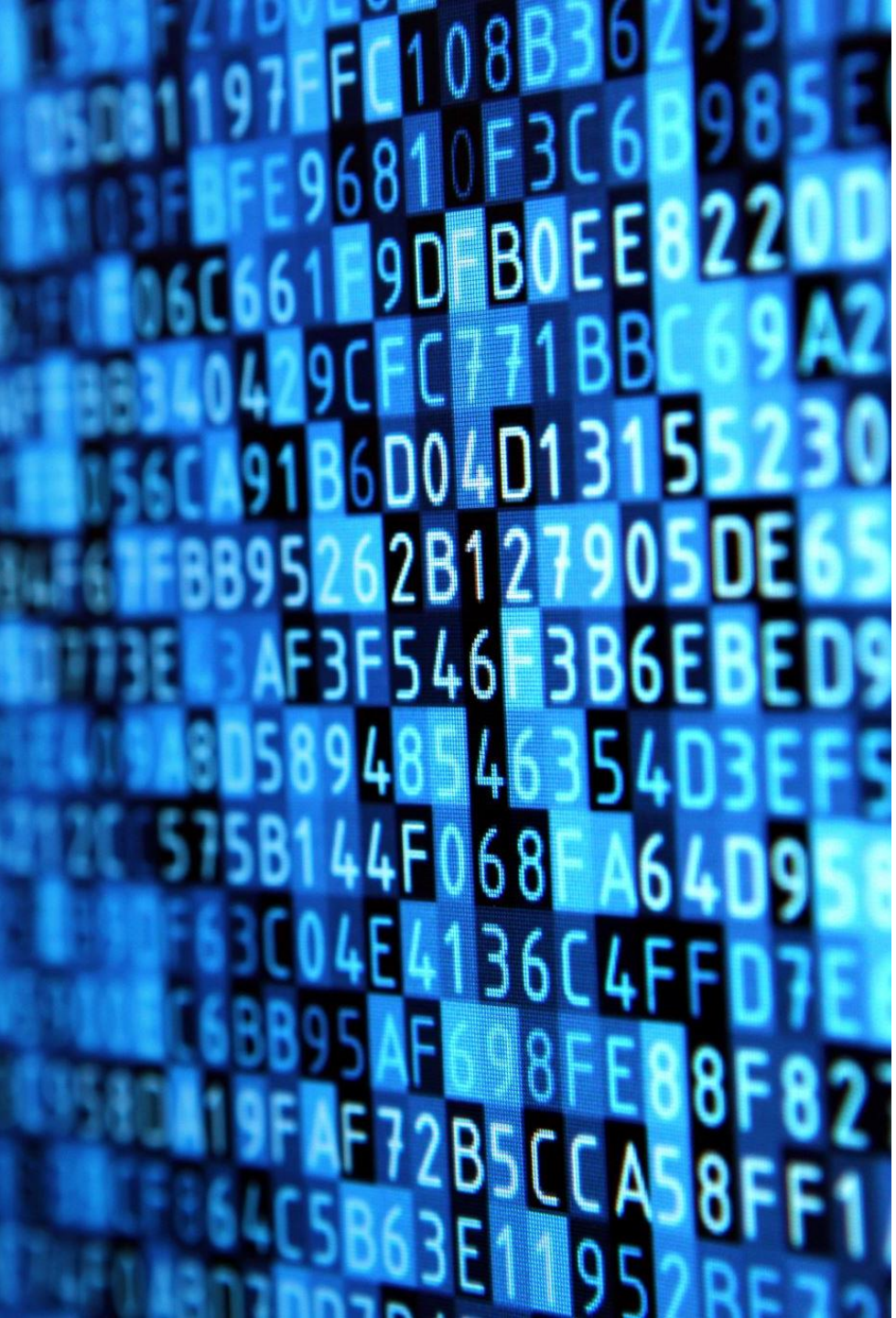
Transformações
map
filter
flatMap
mapPartitions
mapPartitionsWithIndex
sample
union
intersection
distinct
groupByKey
reduceByKey
aggregateByKey
sortByKey
join
cogroup
cartesian
pipe
coalesce
repartition
repartitionAndSortWithinPartitions

Ações
reduce
collect
count
first
take
takeSample
takeOrdered
saveAsTextFile
saveAsSequenceFile
saveAsObjectFile
countByKey
foreach

Dataset e DataFrame

SEMELHANTE A UMA TABELA DE BANCO DE DADOS

COMPATÍVEL COM OBJETOS DATAFRAME DO R E PYTHON



Dataset

- Disponíveis em Java e Scala
- Não disponíveis em R e Python

No curso

- Vamos estudar RDD, porém:
- Prioridade será o DataFrame