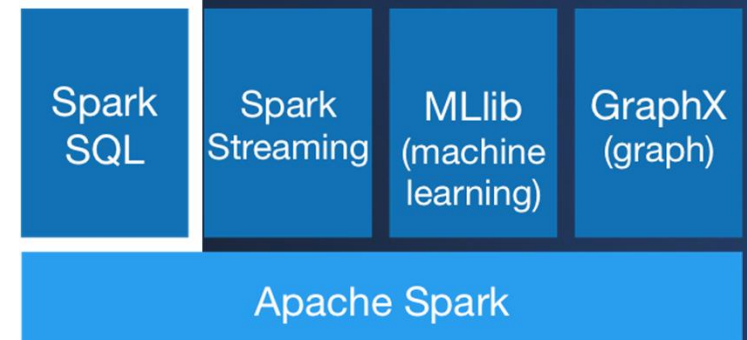


Arquitetura e Componentes



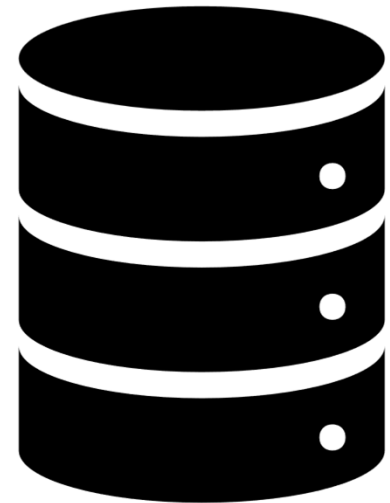
Componentes

- Machine Learning (Mlib)
- SQL (Spark SQL)
- Processamento em Streaming
- Processamento de Grafos (GraphX)



Spark SQL

- Permite ler dados tabulares de várias fontes (CSV, Json, Parquet, ORC etc)
- Pode usar sintaxe SQL



Streaming: Spark Structured Streaming

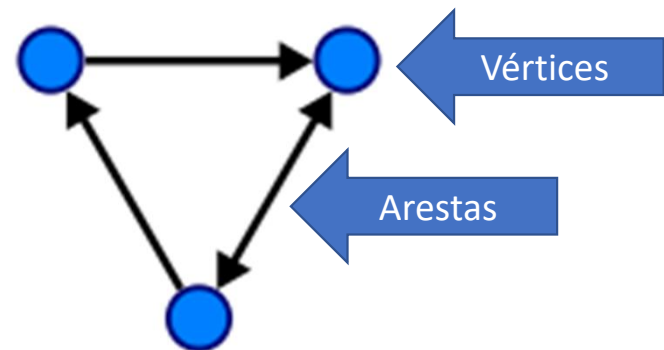
- Dados estruturados





Grafos acíclicos dirigidos

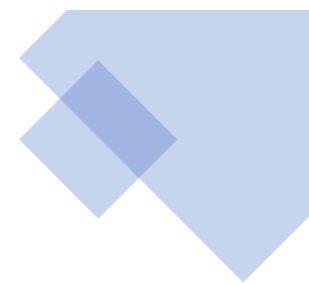
- Spark Constrói Gráficos Acíclicos Dirigidos



A decorative graphic in the top-left corner consisting of a large yellow triangle pointing downwards and a smaller yellow diamond overlapping its bottom-left corner.

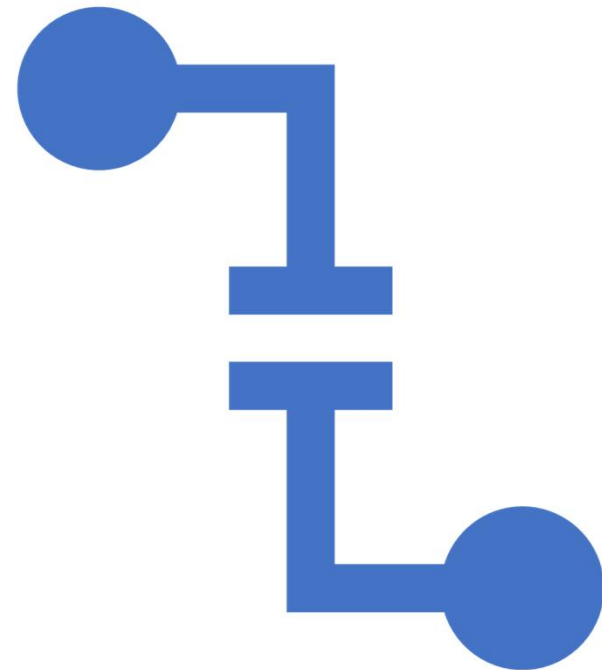
Elementos

- SparkSession: Seção
- Application: Programa

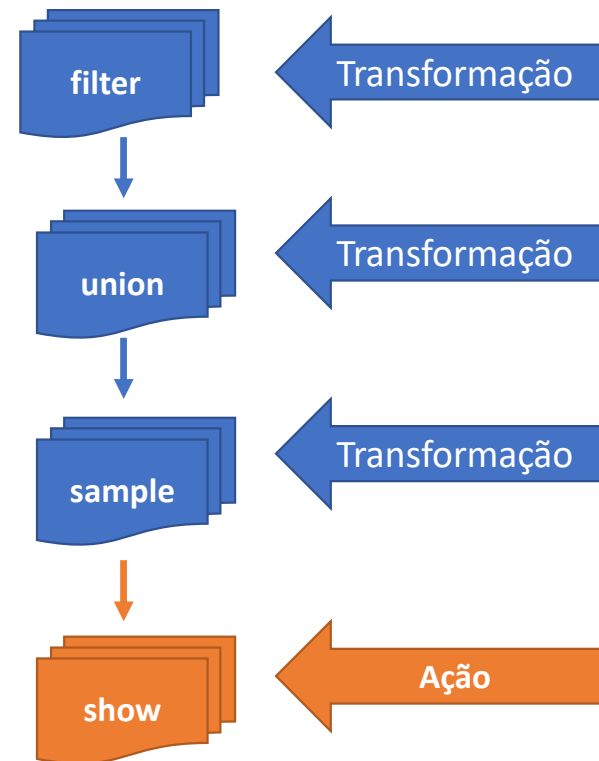


Transformações e Ações

- Um data frame é imutável: traz tolerância a falha
- Uma transformação gera um novo data frame.
- O processamento de transformação de fato só ocorre quando há uma Ação: Lazy Evaluation



Lazy Evaluation





Transformações e Ações

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