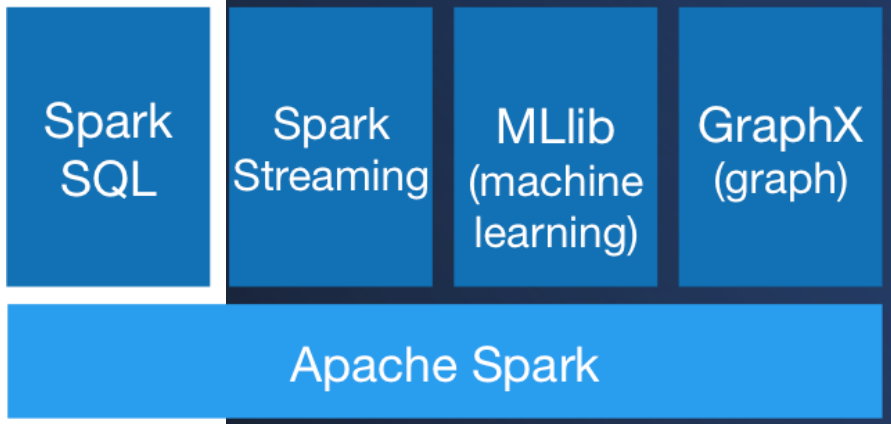
Arquitetura e Componentes

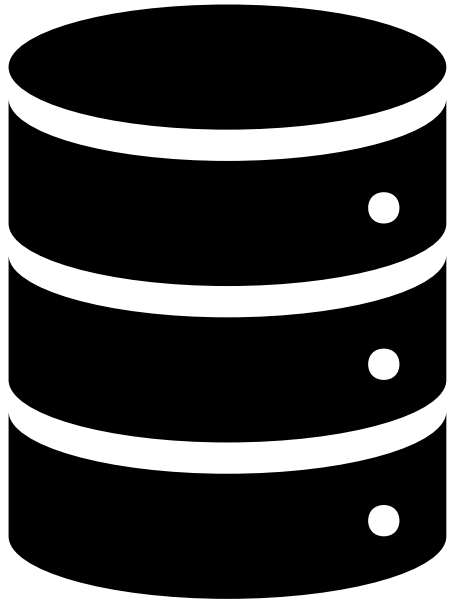
Componentes

- 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

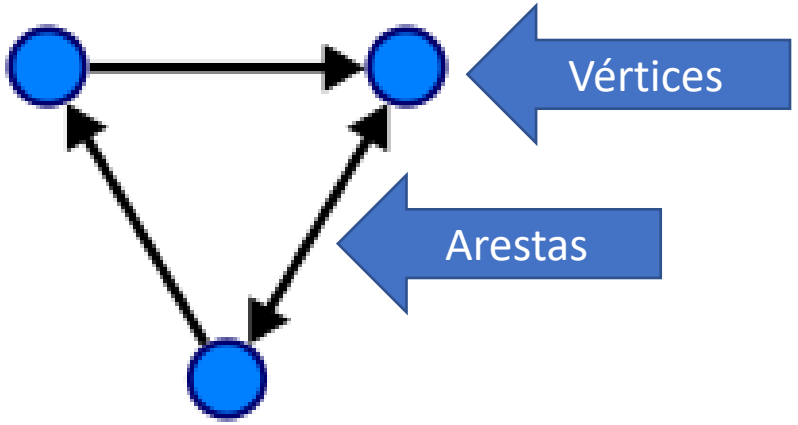


Streaming: Spark Structured Streaming

- Dados estruturados
- Novos registros adicionados ao final da tabela

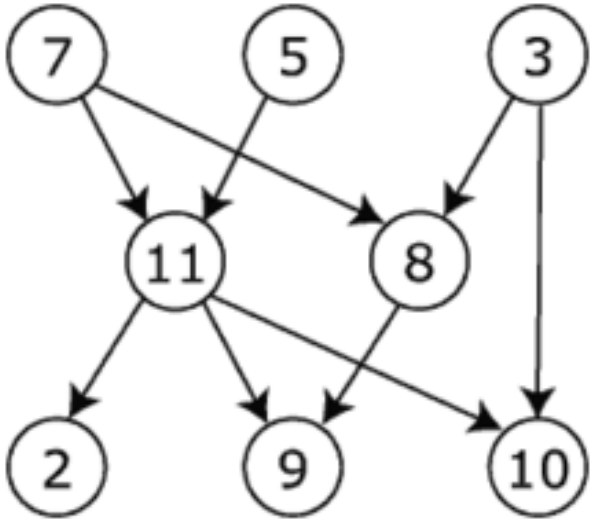
Grafos acíclicos dirigidos

- Spark Constrói Gráficos Acíclicos Dirigidos



Acíclicos Dirigidos

- Acíclicos: não tem ciclo
- Dirigidos: tem uma direção

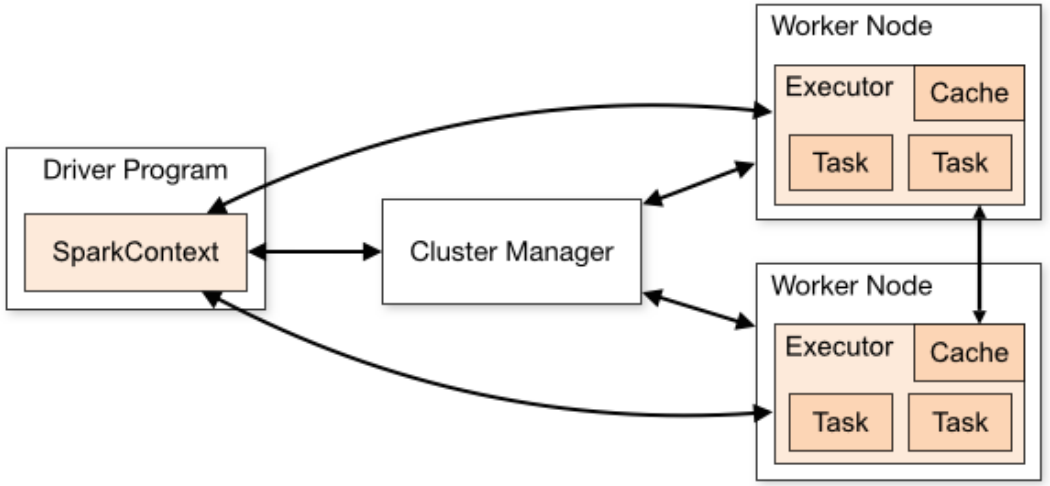


Tungsten

- Motor de Execução

Estrutura

- **Driver**: Inicializa SparkSession, solicita recursos computacionais do Cluster Manager, transforma as operações em DAGs, distribui estas pelos executers
- **Manager**: Gerencia os recursos do cluster. Quatro possíveis: built-in standalone, YARN, Mesos e Kubernetes
- **Executer**: roda em cada nó do cluster executando as tarefas

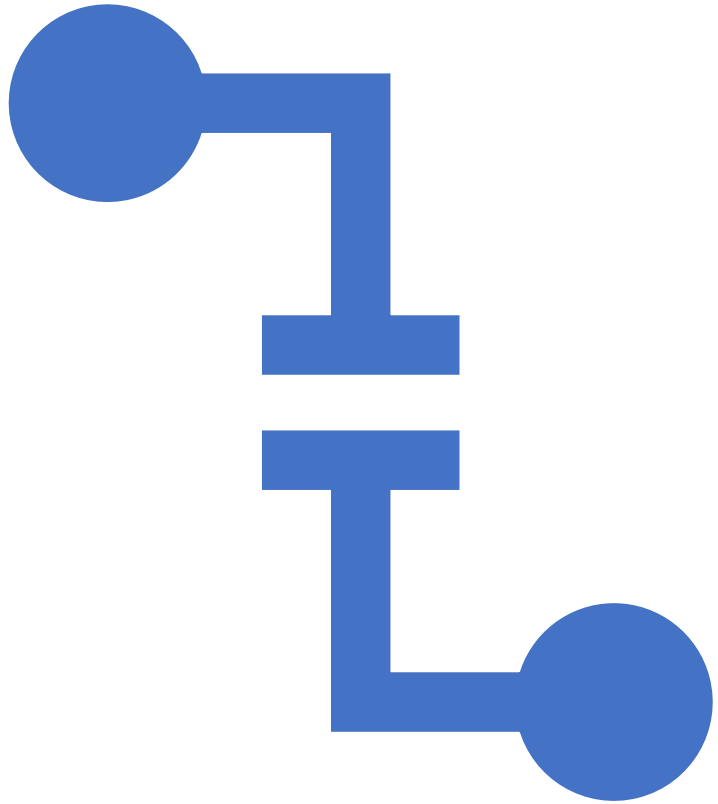


Elementos

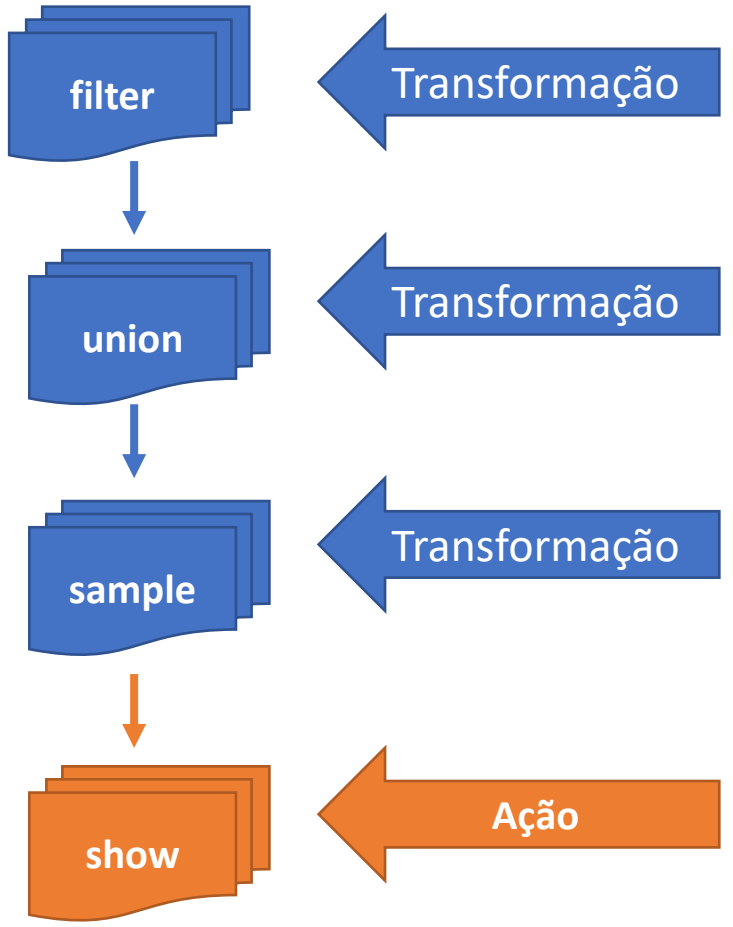
- SparkSession: Seção
- Aplication: Programa

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



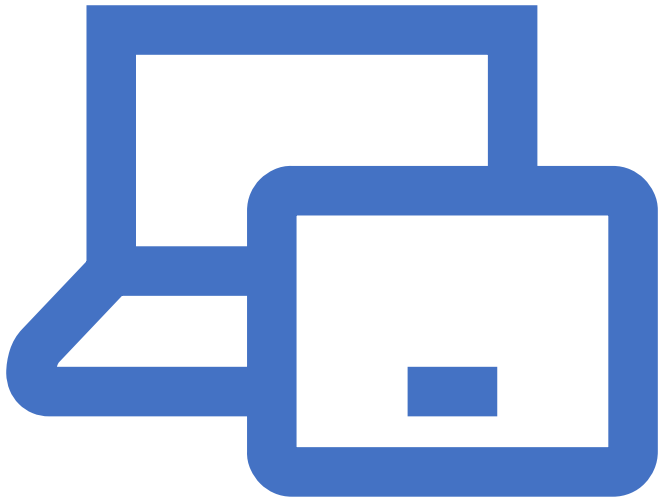
Transformações e Ações

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

Transformações: Narrow e Wide

- Os dados necessários estão em uma mesma partição
- Os dados necessários estão em mais de uma partição



Componentes

- Job: Tarefa
- Stage: Divisão do Job
- Task: Menor unidade de trabalho. Uma por núcleo e por partição