02.Arquitetura e Componentes

quinta-feira, 21 de setembro de 2023 14:28

Componentes

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

Spark SQL

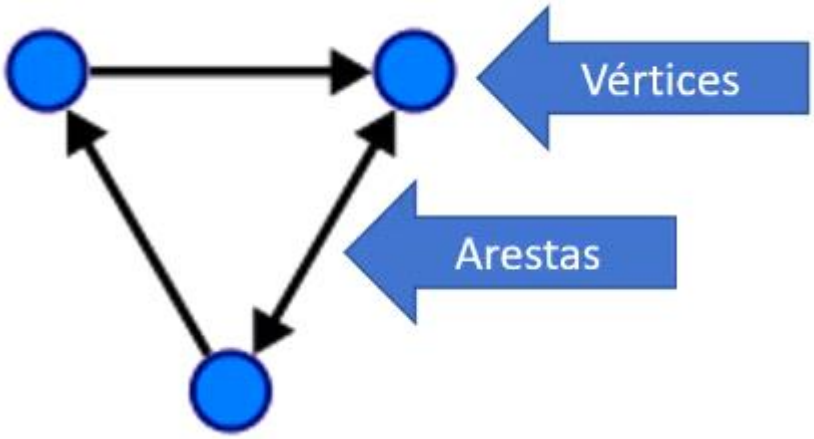
- Permite ler dados tabulares de várias fontes (CSV, Json, Parquet, ORC etc),
- Pode usar sintaxe SQL

Spark Streaming

- Dados Estruturados
- Detecta e processa novos dados a medidas que são adicionados

Grafos acíclicos dirigidos

- Spark constroi gráficos acíclicos dirigidos



Tungsten

- Motor de execução do Spark

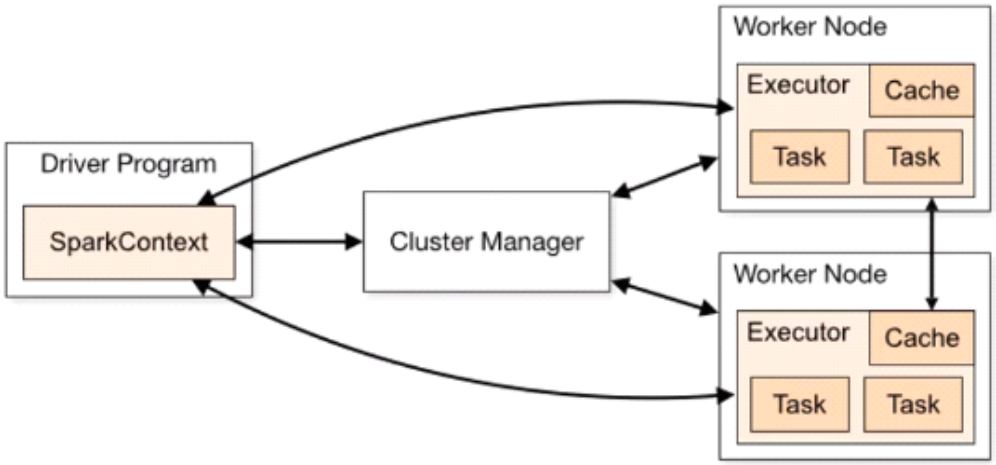
Estrutura do Spark

Driver: Inicializa SparkSession, solicita recursos computacionais do cluster manager, transforma as operações em DAGs, distribui estas pelos executers.

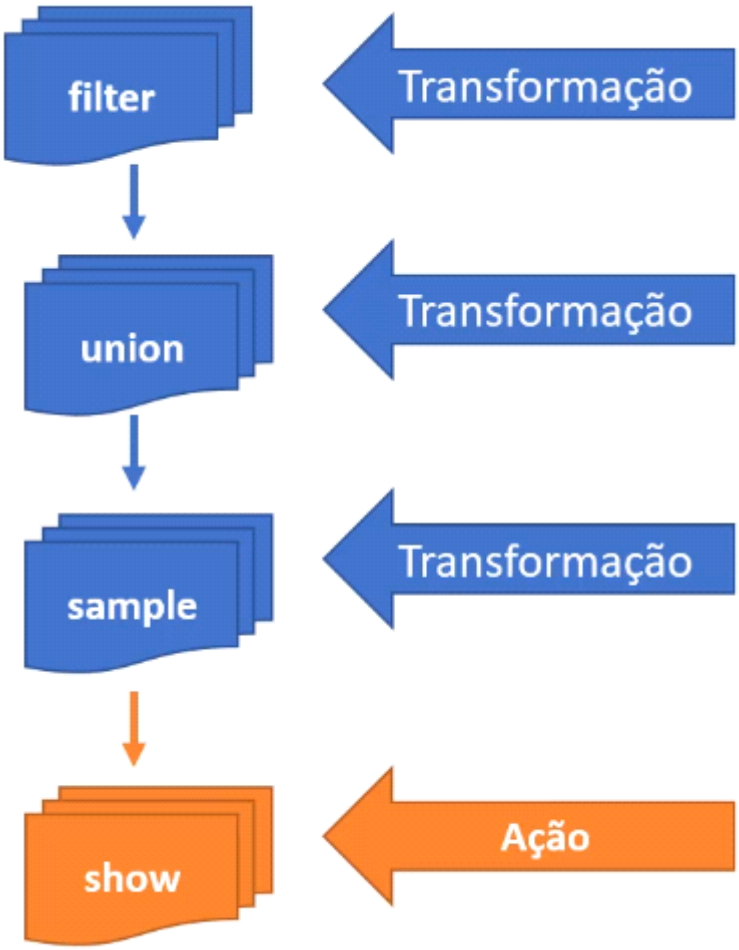
Manager: Gerencia os recursos do Cluster.

Quatro possíveis: Built-in standalone, YARN, Mesos e Kubernetes

Executer: Roda em cada nó do cluster executando as tarefas



- 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



Transformações e Ações

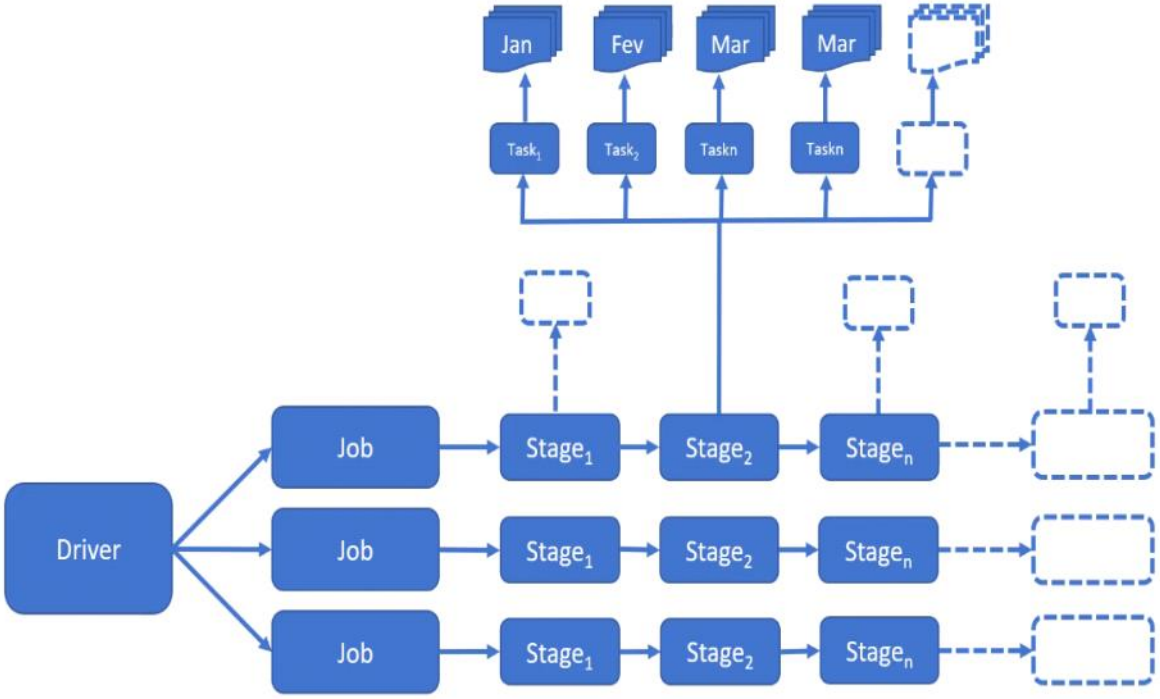
Transformações	Ações
map	reduce
filter	collect
flatMap	count
mapPartitions	first
mapPartitionsWithIndex	take
sample	takeSample
union	takeOrdered
intersection	saveAsTextFile
distinct	saveAsSequenceFile
groupByKey	saveAsObjectFile
reduceByKey	countByKey
aggregateByKey	foreach
sortByKey	
join	
cogroup	
cartesian	
pipe	
coalesce	
repartition	
repartitionAndSortWithinPartitions	

Existem 2 tipos de transformações:

Narrow - Dados estão em 1 mesma partição
Wide - Dados estão em mais de 1 partição

Como os dados Spark são processados?

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


Jan
Fev
Mar
Mar
Task1
Task2
Taskn
Taskn
Driver
Job
Job
Job
Stage1
Stage2
Stagen
Stage1
Stage2
Stagen
Stage1
Stage2
Stagen