Data Engineering

Filipe Nascimento
fgvn@cesar.school

C.E.S.A.R

C . E . S . A . R

Pessoas impulsionando inovação.

Inovação impulsionando negócios.

Everton Dias
etgdb@cesar.org.br

Janaína Branco
jcb@cesar.org.br

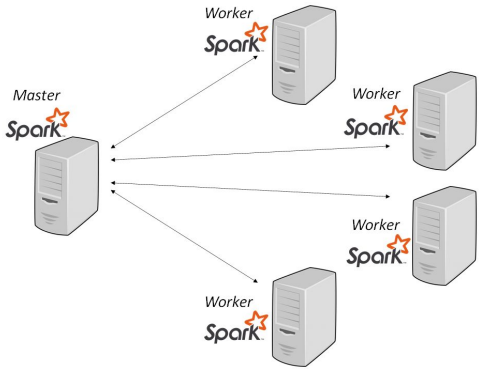
APACHE SPARK

Nesta fase abordaremos alguns conceitos ligados ao Apache Spark, processamento distribuído, big data e outros.

CONCEITOS

Framework que fornece uma plataforma analítica **unificada** para processamento de dados em **larga escala**.

Desenvolvido na Universidade da Califórnia e posteriormente repassado para a Apache Foundation o framework Spark provê uma série de recursos para processamento em clusters utilizando paralelismo e tolerância a falhas.

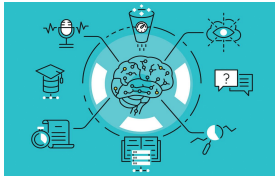


Usos do Spark

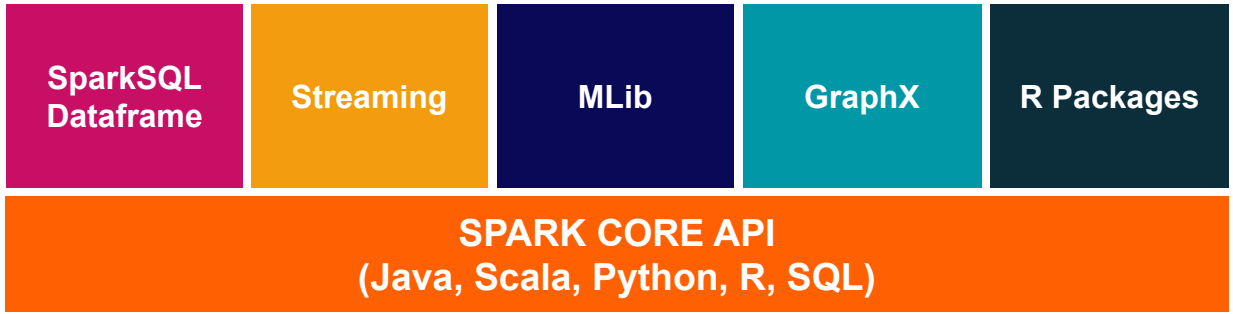
Preparar dados para análise

Analisar dados em tempo real

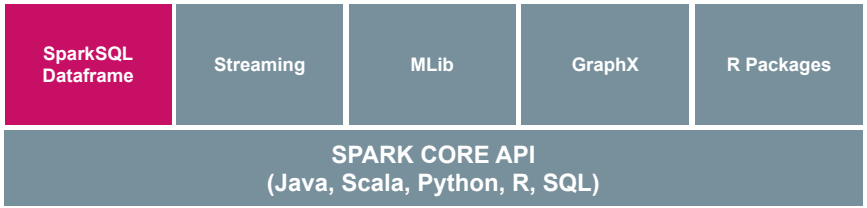
Criação de modelos de Machine Learning



PLATAFORMA UNIFICADA SPARK



PLATAFORMA UNIFICADA SPARK



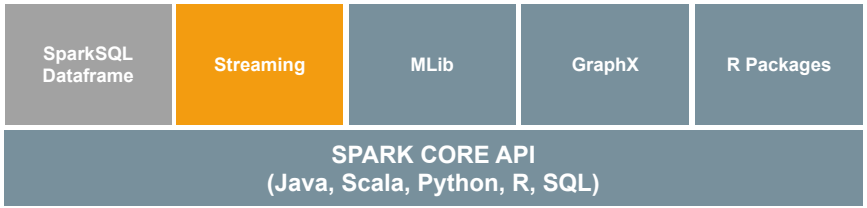
Spark SQL & DataFrame : trabalha com workloads de dados e ETL

- Hive
- CSV
- JSon
- RDBMS
- XML
- Parquet
- Cassandra
- RDDs

SparkSQL Dataframe

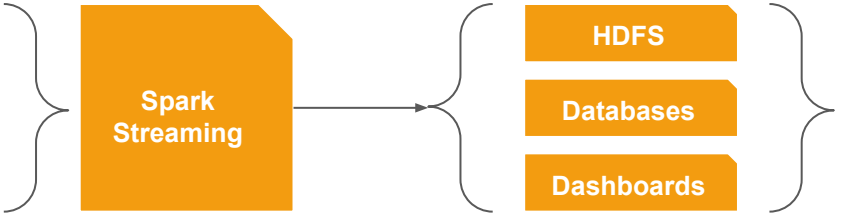
DataFrame			
	COl1	Col2	Col3
Row 1			
Row 1			

PLATAFORMA UNIFICADA SPARK

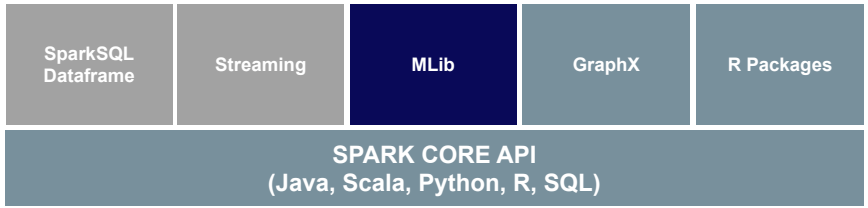


Streaming : trabalha com processamento em tempo real

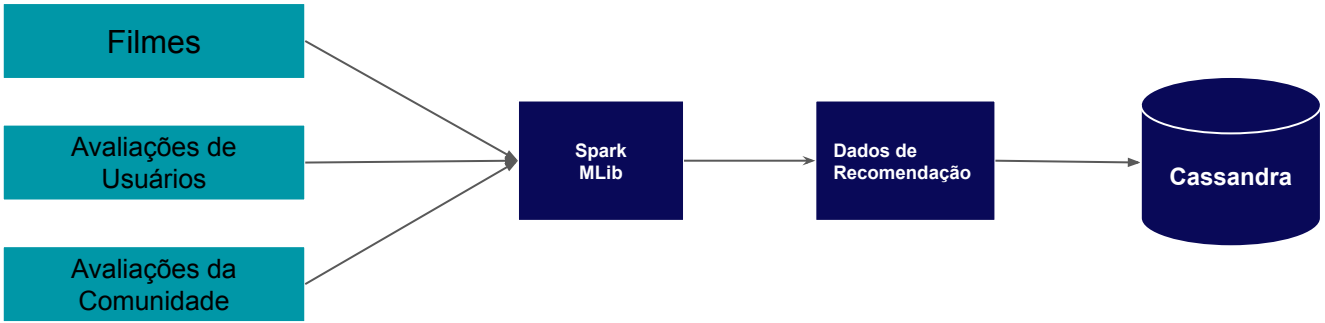
- Kafka
- Flume
- HDFS/S3
- Kinesis
- Twitter



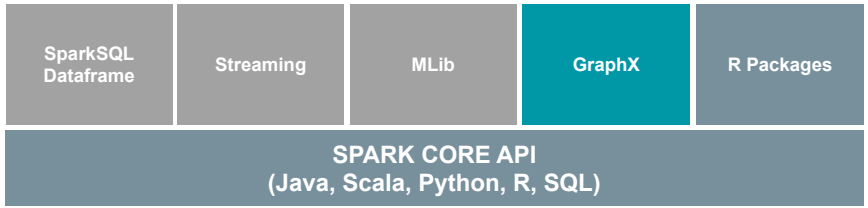
PLATAFORMA UNIFICADA SPARK



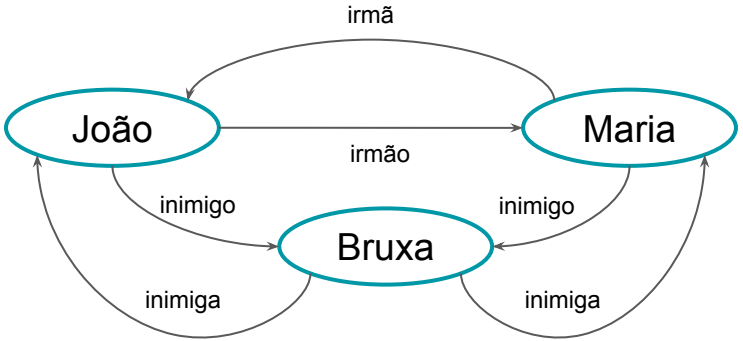
Spark MLib: Trabalha com pipelines de machine learning.



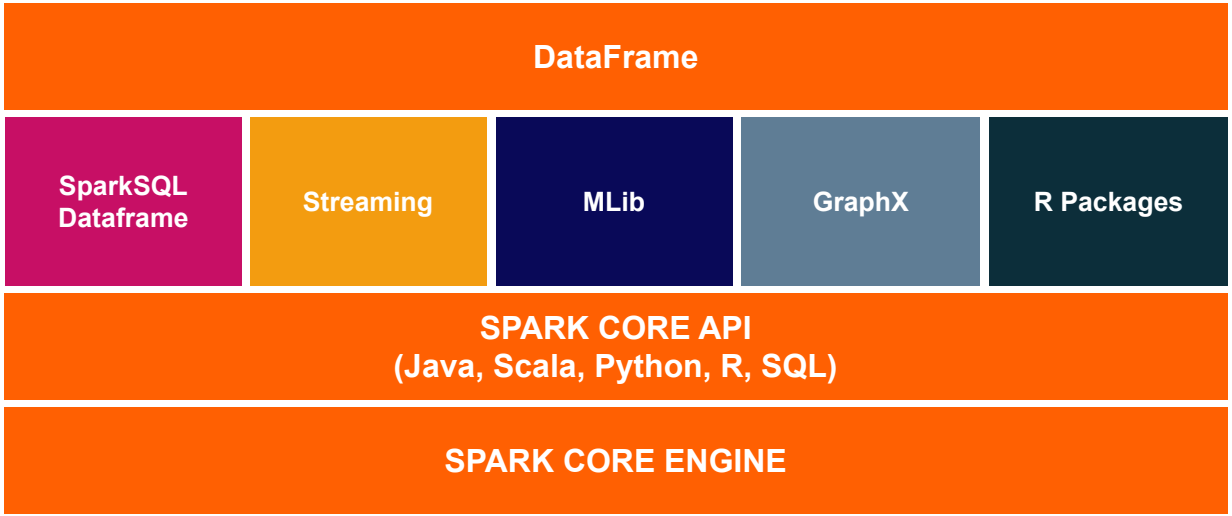
PLATAFORMA UNIFICADA SPARK



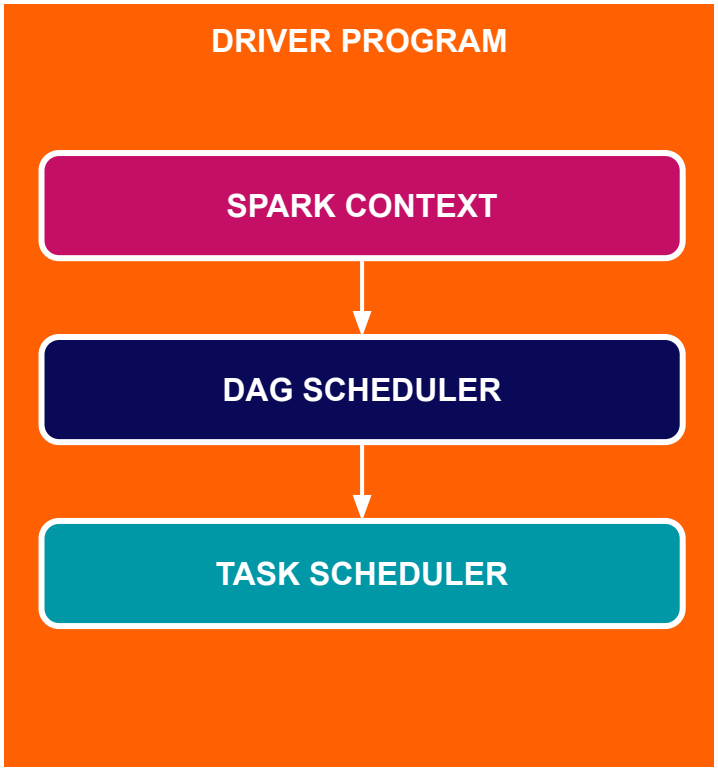
Spark GraphX : Segmento de recursos trabalhar com processamento de grafos.



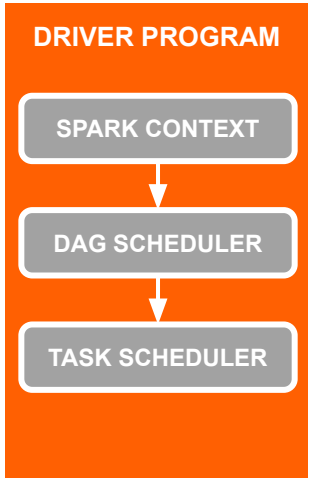
PLATAFORMA UNIFICADA SPARK



COMPONENTES DO SPARK

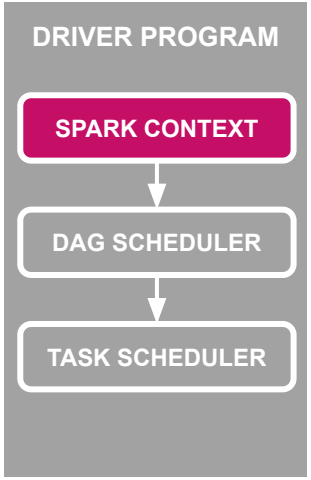


COMPONENTES DO SPARK



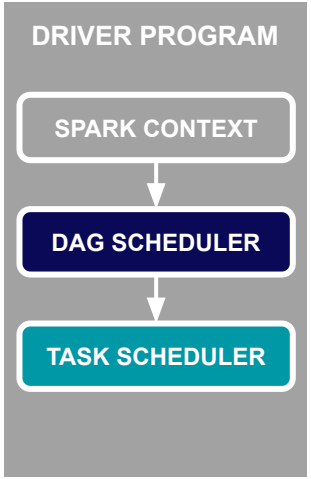
O ***driver program*** é o componente responsável pela orquestração do programa Spark, tendo uma posição como uma função principal contendo toda a lógica de execução do código, possui uma quantidade de memória alocada, necessitando uma atenção para evitar estouro de memória.

COMPONENTES DO SPARK



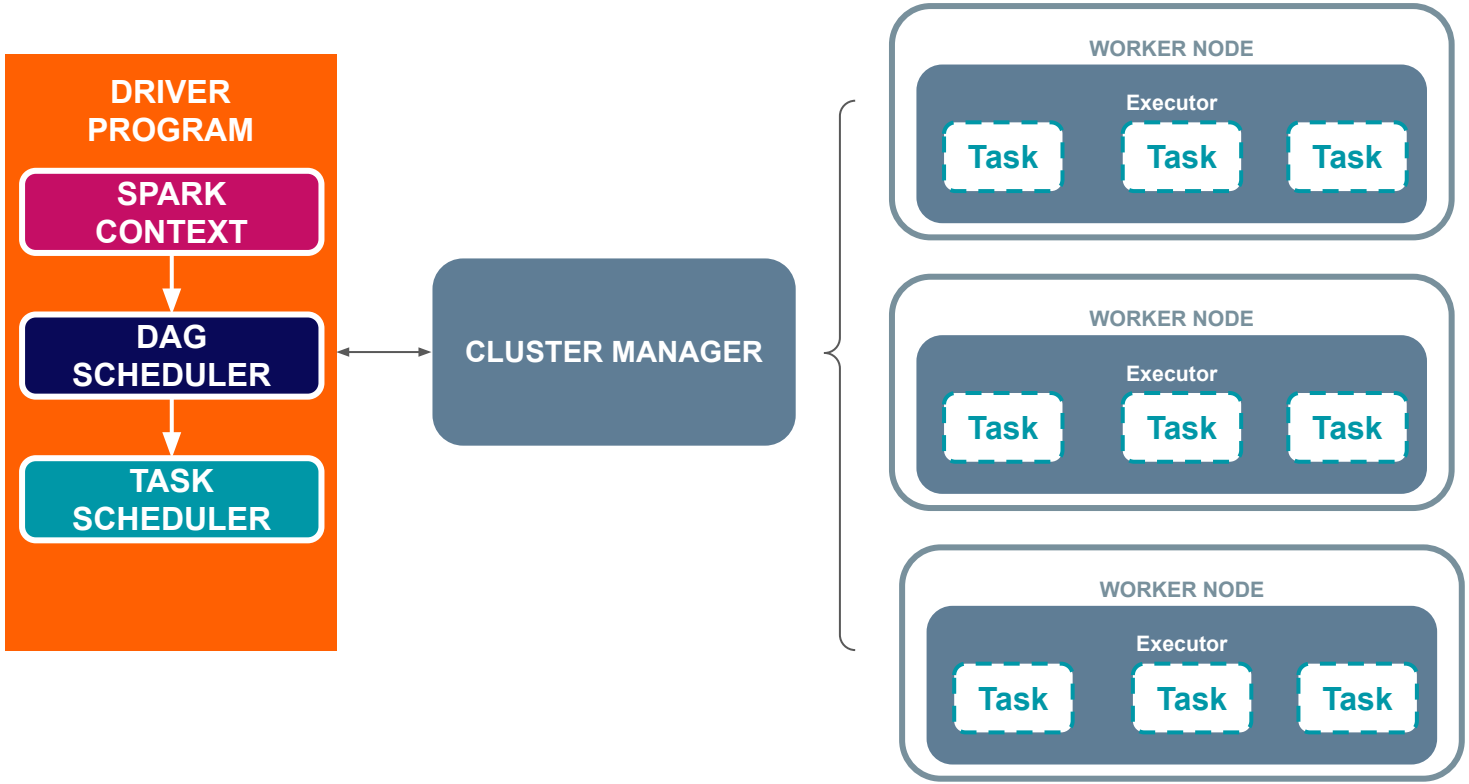
O ***spark context*** é o componente que faz a mediação entre o driver program e os executores, no spark context fazemos a configuração da quantidade de memória dos executores e número de cores dos executores.

COMPONENTES DO SPARK

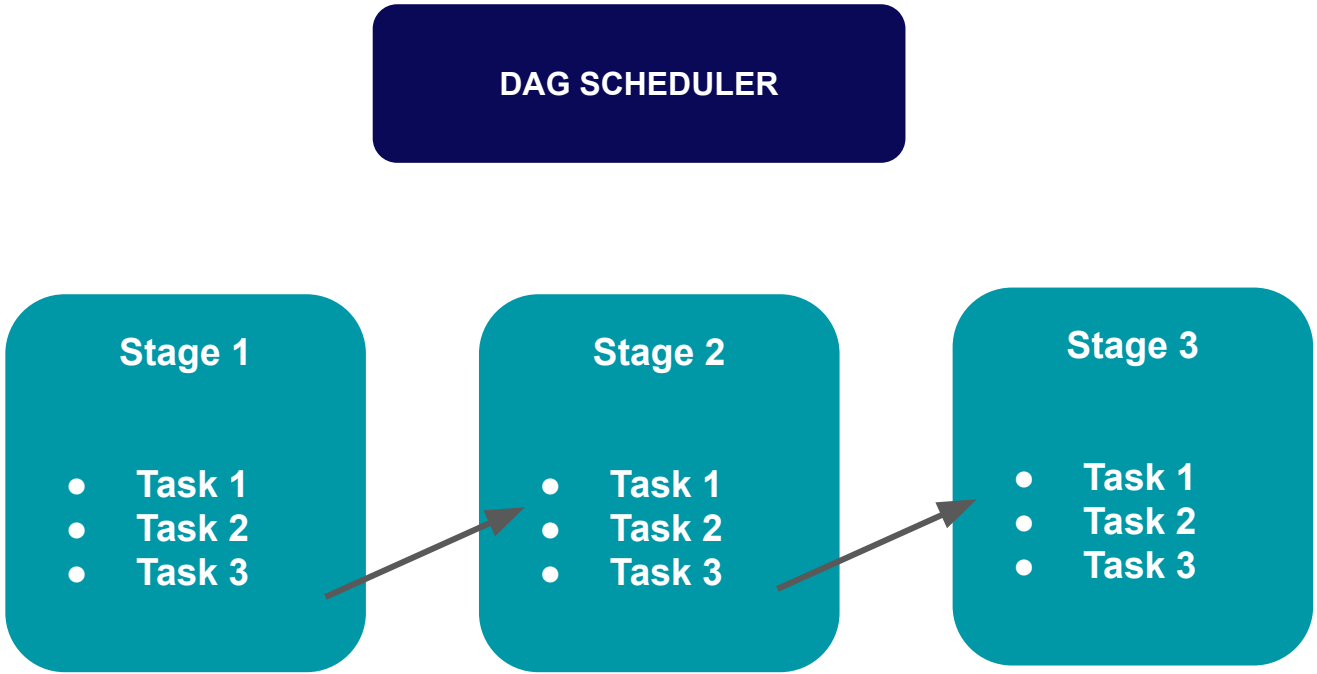


Os planos lógicos de execução possuem uma grande similaridade aos planos lógicos de execução dos bancos de dados, eles registram um passo a passo para operação nos dados, com base nesse plano lógico ele irá criar um dos componentes mais importantes do spark o - **DAG - Dynamically Acyclic Graph** - que irá manter as etapas para execução das nossas transformações em stages e com base nelas irá criar um **plano físico de execução**.

COMPONENTES DO SPARK

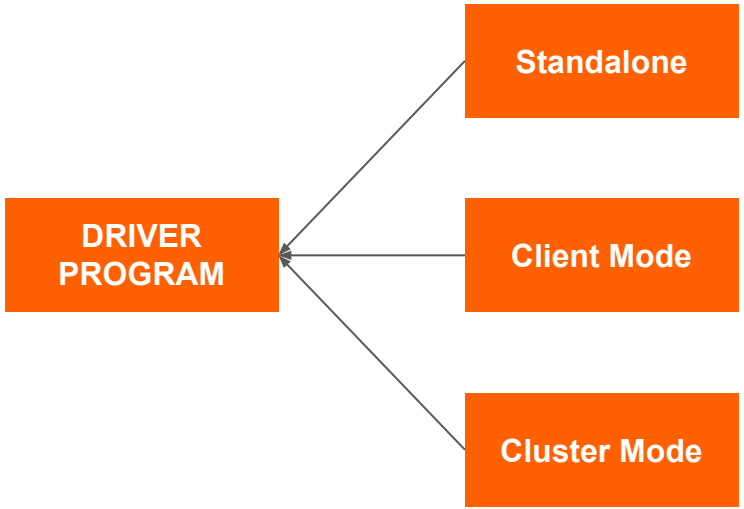


COMPONENTES DO SPARK



MODOS DE EXECUÇÃO DO SPARK

Os modos de execução do Spark irão diferir onde o driver program do spark irá rodar.



Execução local, sem utilização de clusters onde você pode distribuir os jobs entre os cores de sua máquina, porém a atuação é local apenas.

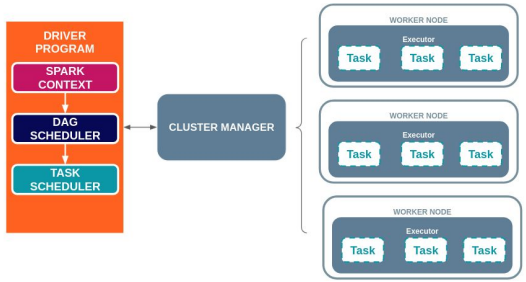
O driver program irá executar na mesma máquina onde foi iniciado o programa.

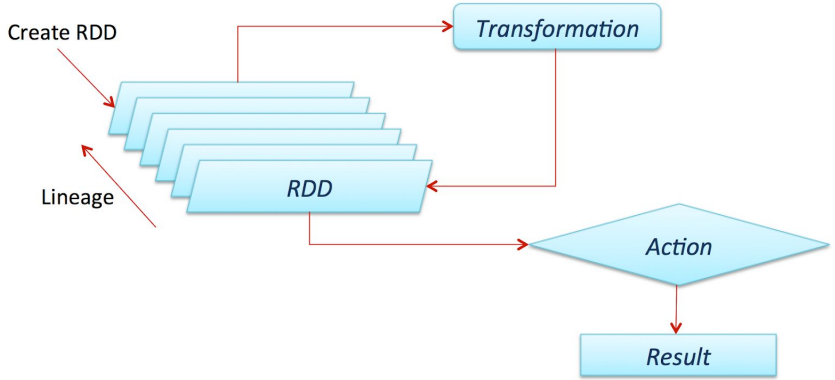
O driver program irá executar em qualquer uma das máquinas do cluster.

COMPONENTES DO SPARK

RDD - Resilient Distributed Dataset

- Imutáveis
- Transformations e Actions
- Lazy Evaluation
- Conhecer a lista de partitions
- Saber gerenciar as dependências de cálculo (DAG)





SPARK E HADOOP

Hadoop é uma estrutura de software open-source para armazenar dados e executar aplicações em clusters de hardwares comuns. Ele fornece armazenamento massivo para qualquer tipo de dado, grande poder de processamento e a capacidade de lidar quase ilimitadamente com tarefas e trabalhos ocorrendo ao mesmo tempo.

SPARK E HADOOP

- Armazena resultados parciais e finais em disco;
- Map Reduce;

- Armazena resultados parciais em memória e apenas os finais em disco;
- Map Reduce e outras funções de transformações de dados;

Planos de Execução e RDDs

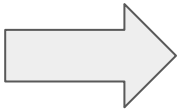
Alunos Cesar

Part #1
Recife

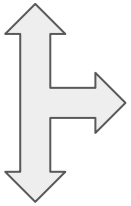
Part #2
Agreste

Part #3
Sertão

Narrow Dependency



Wide Dependency



Planos de Execução e RDDs

Alunos Cesar

alunos.map(...)

Part #1
Recife

Idade

Part #2
Agreste

Idade

Part #3
Sertão

Idade

Planos de Execução e RDDs

Alunos Cesar

alunos.map(...).filter(...)

Part #1
Recife

Idade

>=18

Part #2
Agreste

Idade

>=18

Part #3
Sertão

Idade

>=18

Planos de Execução e RDDs

Alunos Cesar

alunos.map(...).filter(...).reducebykey(...)

Part #1
Recife

Idade

>=18

18-25

Part #2
Agreste

Idade

>=18

25-35

Part #3
Sertão

Idade

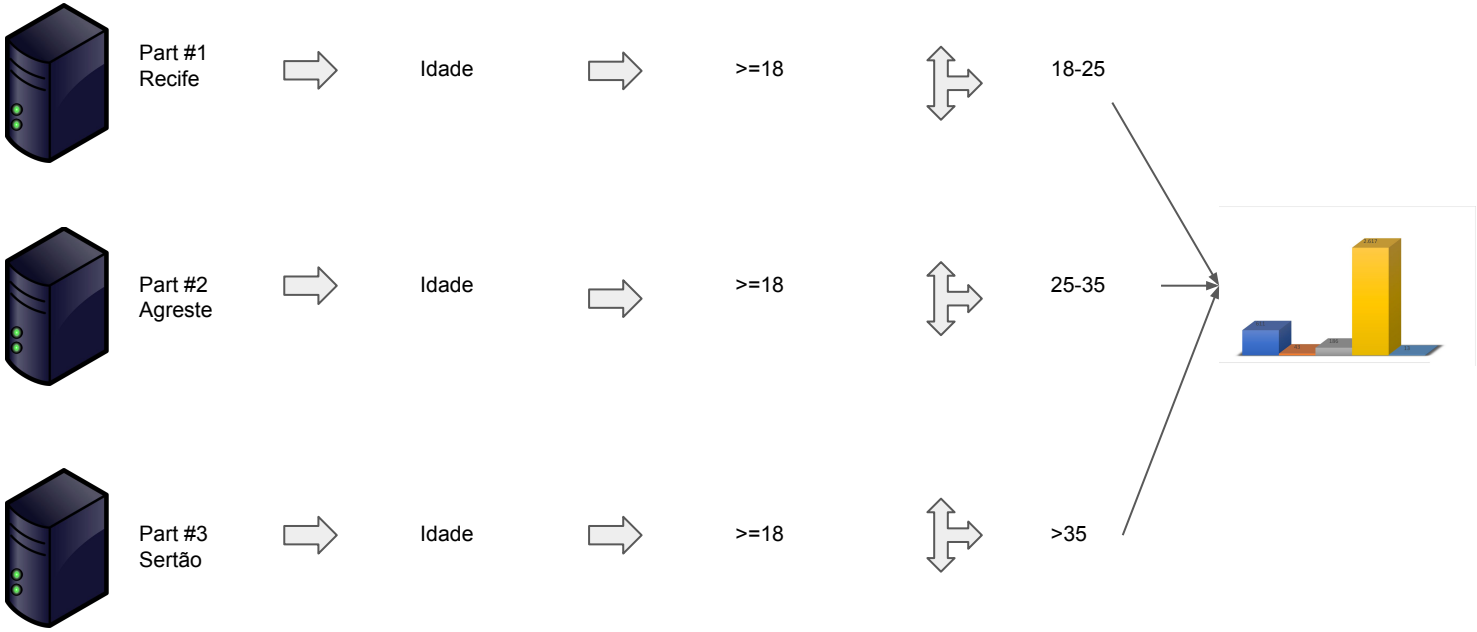
>=18

>35

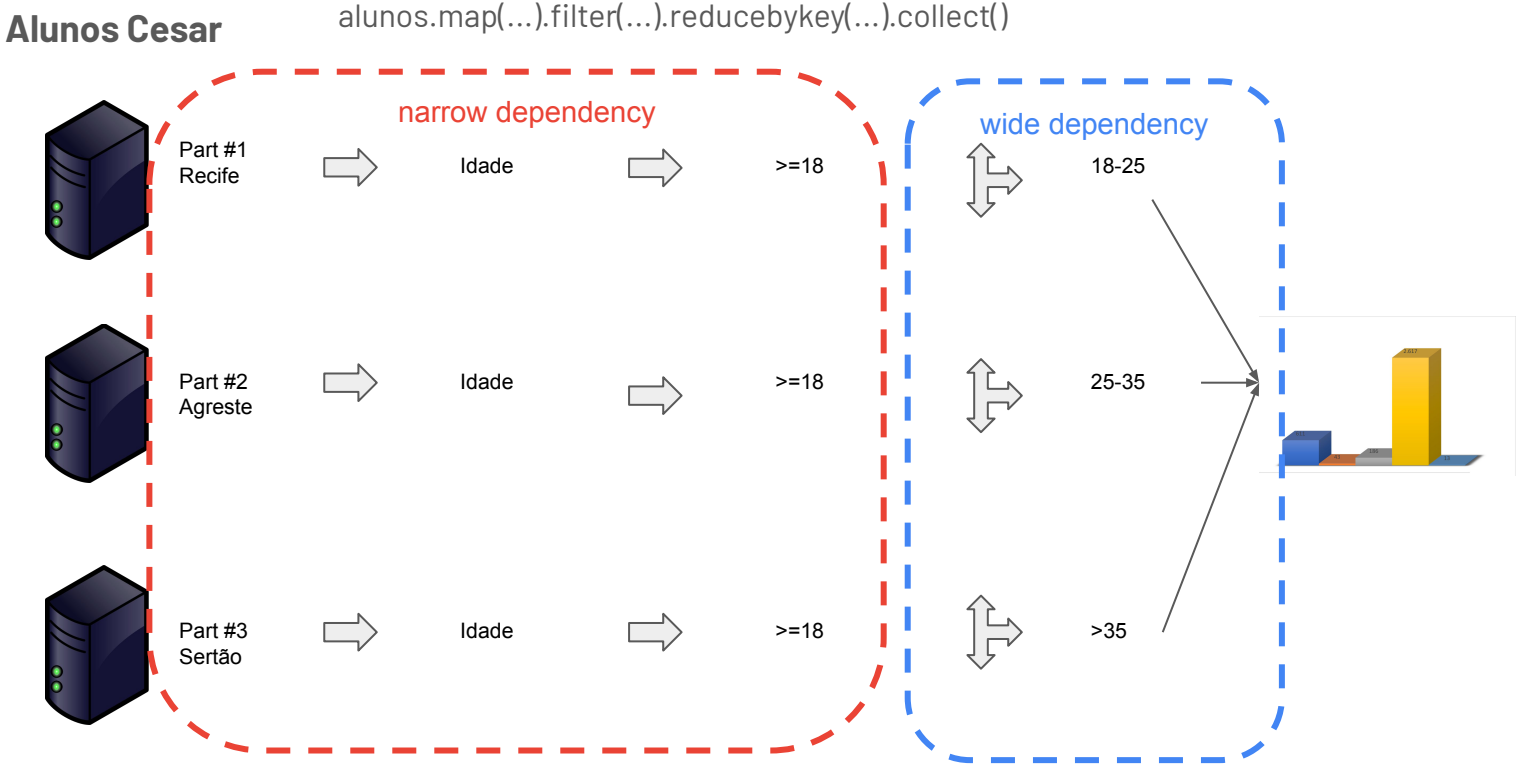
Planos de Execução e RDDs

Alunos Cesar

alunos.map(...).filter(...).reducebykey(...).collect()



Planos de Execução e RDDs



USO DE MEMÓRIA NO SPARK

MEMORY_ONLY	RDDs SALVOS EM MEMÓRIA, PARTE EXCEDENTE SENDO RECALCULADA QUANDO NECESSÁRIO
DISK_ONLY	TODOS DADOS DE RDDs SALVOS EM DISCO
MEMORY_AND_DISK	SALVA TODOS RDDs NA MEMÓRIA E O EXCEDENTE EM DISCO

PRÁTICA

AMANHÃ TEM MAIS SPARK!