Engenharia de Dados com Hadoop e Spark

Data Science Academy

Data Science Academy

Bem-vindo

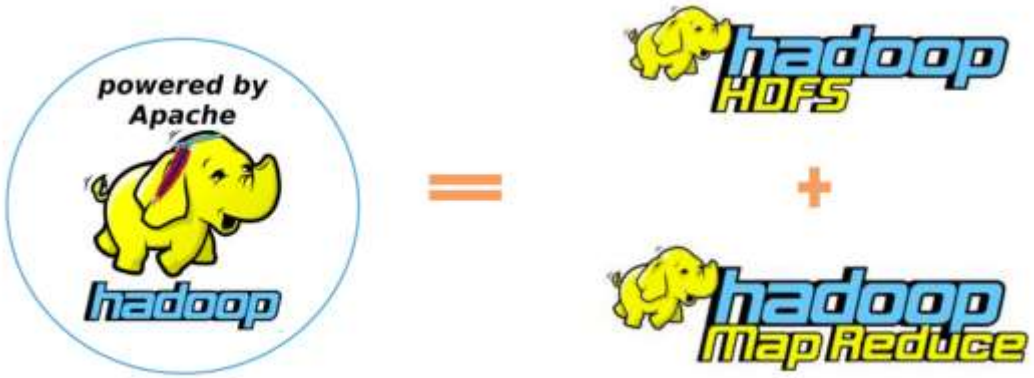
Data Science Academy

Usando MapReduce em Grandes Volumes de Dados

Data Science Academy

MapReduce

Hadoop



MapReduce

O que vamos estudar neste capítulo?

- Computação Distribuída
- Funcionamento do MapReduce
- Processamento de Dados Armazenados no HDFS
- Processamento de Big Data
- Criação e Monitoramento de Jobs MapReduce
- Processamento de Jobs MapReduce em Nuvem, com o Serviço AWS da Amazon

Data Science Academy

MapReduce

O que exatamente é processamento MapReduce?

Programação de Computadores

Python Fundamentos para Análise de Dados
Big Data Real-Time Analytics com Python e Spark

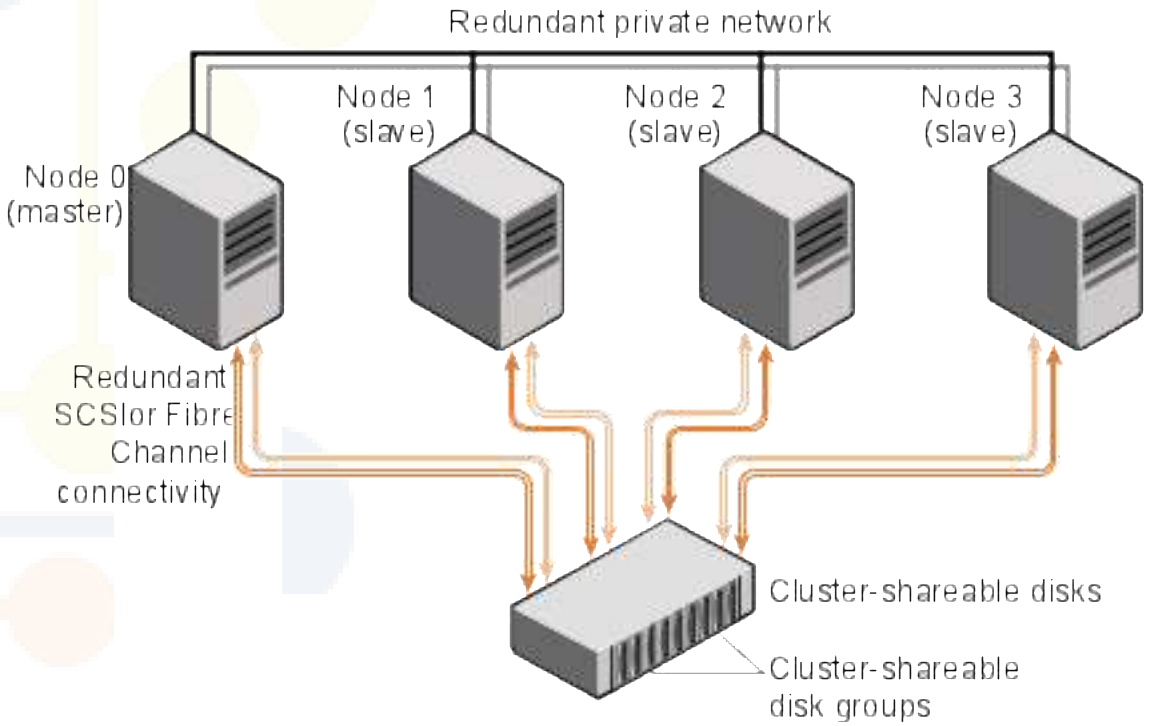
Data Science Academy

Computação Distribuída

Data Science Academy

Computação Distribuída

Sistema de Processamento Distribuído e Paralelo



Data Science Academy

Computação Distribuída

Uma tarefa qualquer, pode ser dividida em várias subtarefas e então executadas em paralelo.

Data Science Academy

Computação Distribuída

- Pesquisas científicas
- Previsões climáticas
- Descoberta de novas partículas
- Controle de epidemias
- Armazenamento e Processamento de Big Data

Computação Distribuída

Computação Distribuída

Sistemas Computacionais estão cada vez mais elaborados e complexos

Grande parte das máquinas interligadas por redes de computadores

Computação Distribuída

Sistemas Distribuídos
Maior poder de processamento
Maior carga, maior número de usuários
Melhor tempo de resposta
Maior confiabilidade

Data Science Academy

Computação Distribuída

A computação distribuída consiste na utilização de um conjunto de máquinas conectadas por uma rede de comunicação, atuando como um único sistema

Data Science Academy

Computação Distribuída

Computação Distribuída

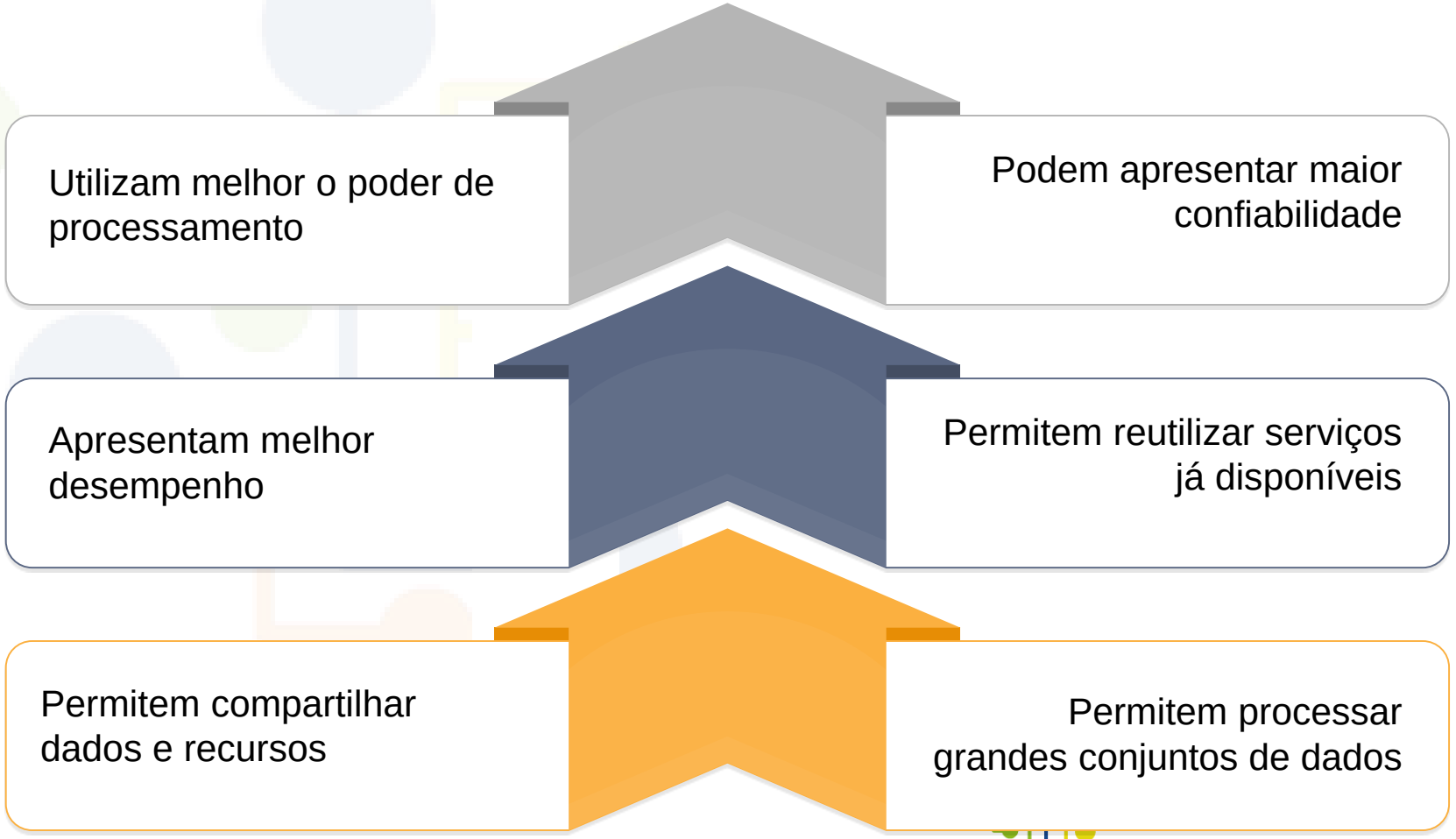
- Executa aplicações através de máquinas diferentes, como se estas fossem uma só
- Tornou-se possível com a popularização das redes de computadores
- As máquinas podem estar interligadas por redes intranets, internet, redes públicas e privadas

Data Science Academy

Computação Distribuída

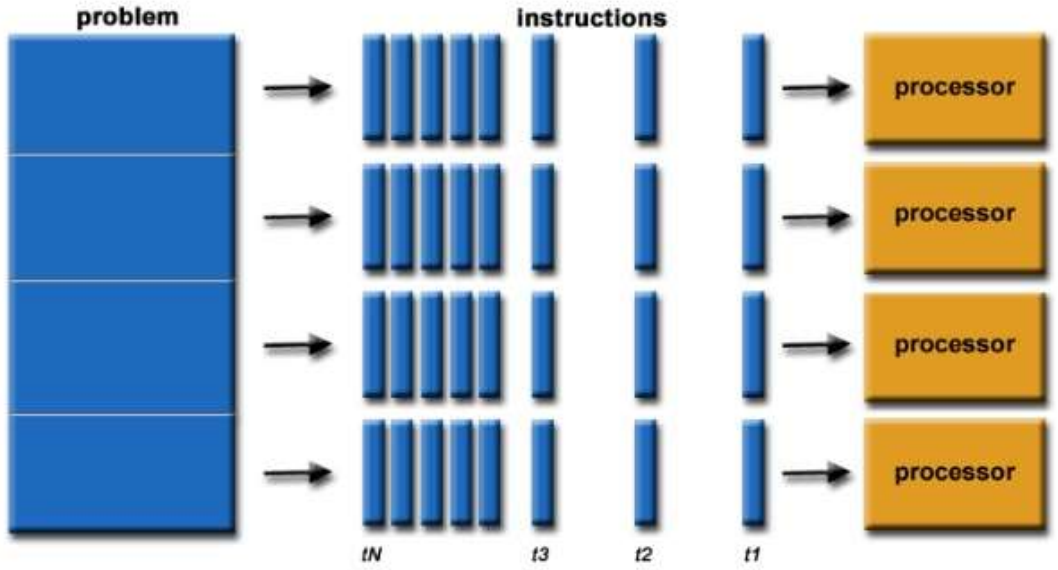
Computação Distribuída

Vantagens



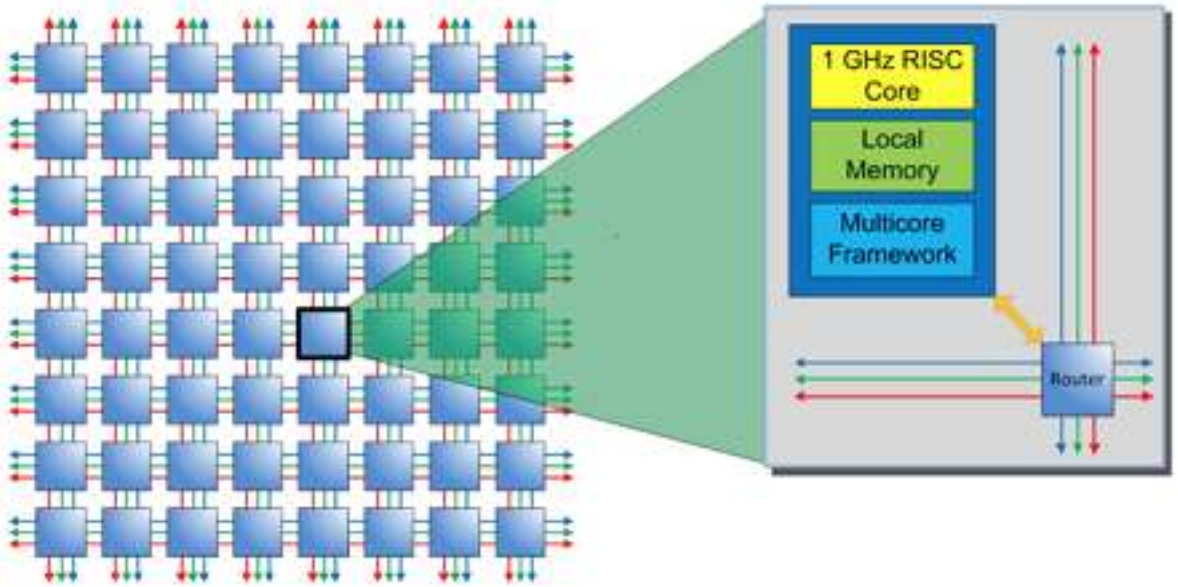
Computação Distribuída

Computação Paralela



Computação Distribuída

Computação Paralela



Computação Distribuída
Cloud Computing

Data Science Academy

Computação Distribuída

Computação Distribuída

Computação Distribuída

Computação Distribuída

Infraestrutura como um serviço (IaaS)

Computação Distribuída

Plataforma como um serviço (PaaS)

Computação Distribuída

Software como um serviço (SaaS)

Computação Distribuída

Mas o Big Data (sempre ele) trouxe mais um serviço para a nuvem

Computação Distribuída

Big Data como um serviço (BDaaS)

Computação Distribuída

Big Data como um serviço (BDaaS)

Computação Distribuída

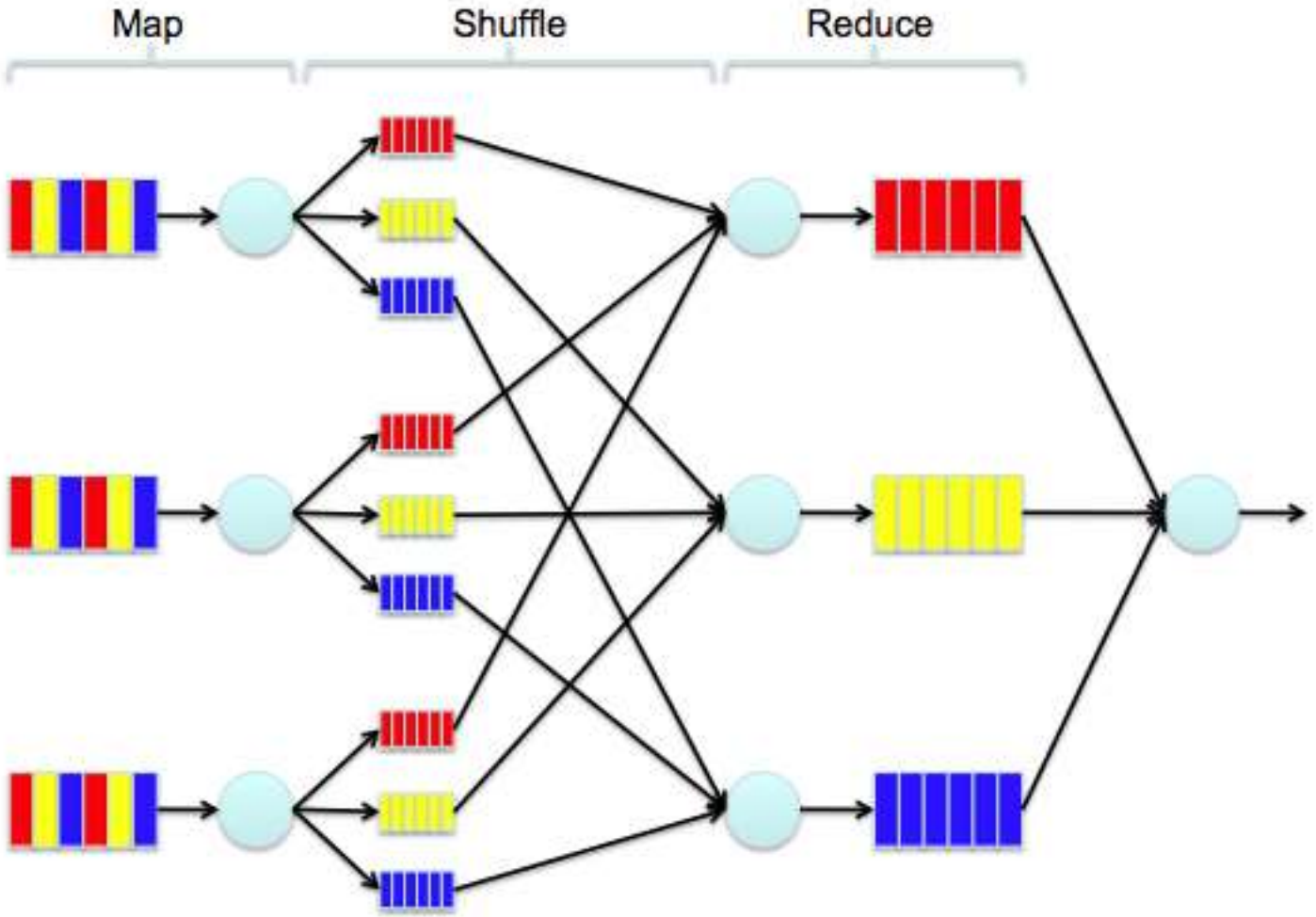
Computação Distribuída

Data Science Academy

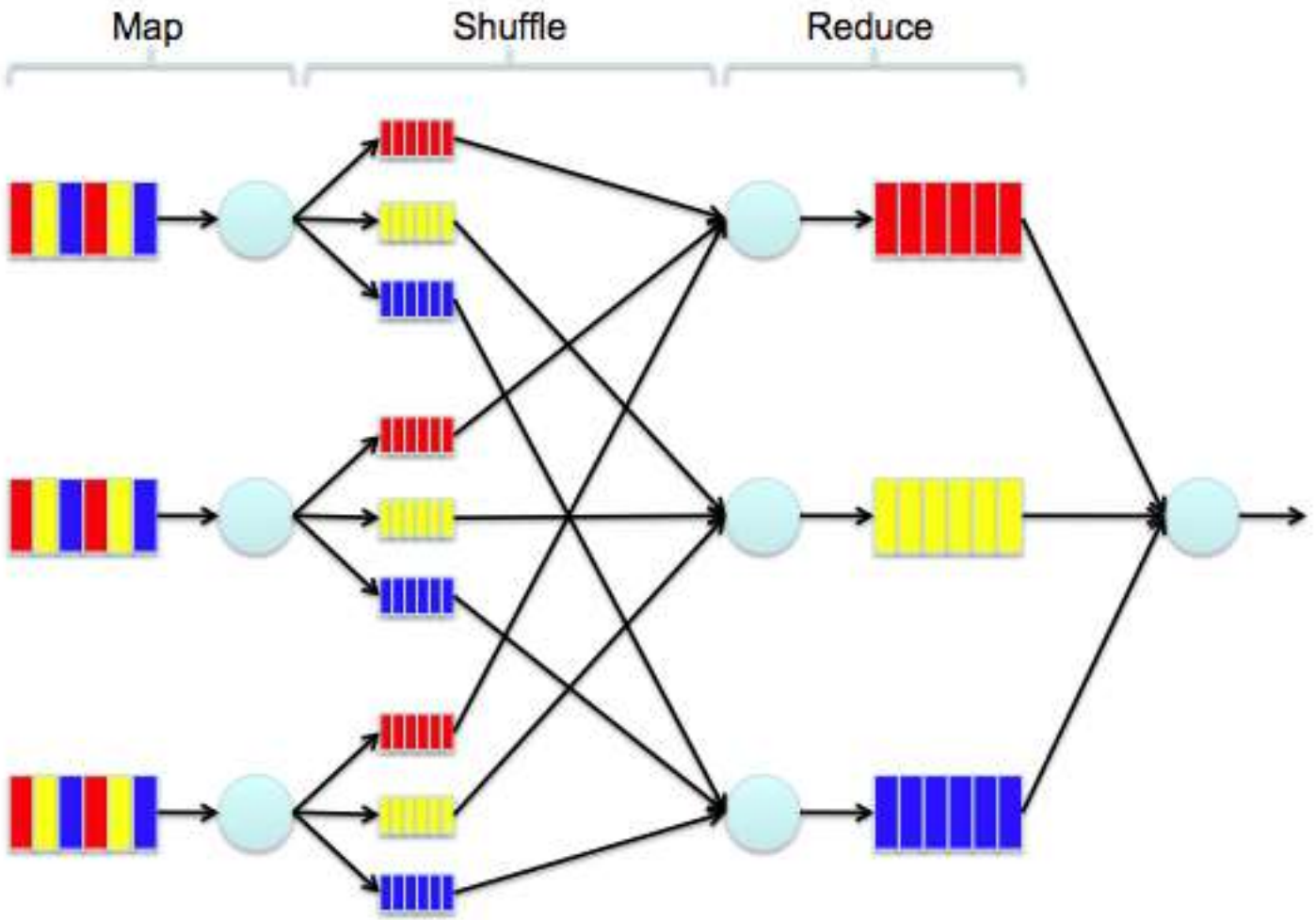
O Modelo de Programação MapReduce

Data Science Academy

MapReduce



MapReduce

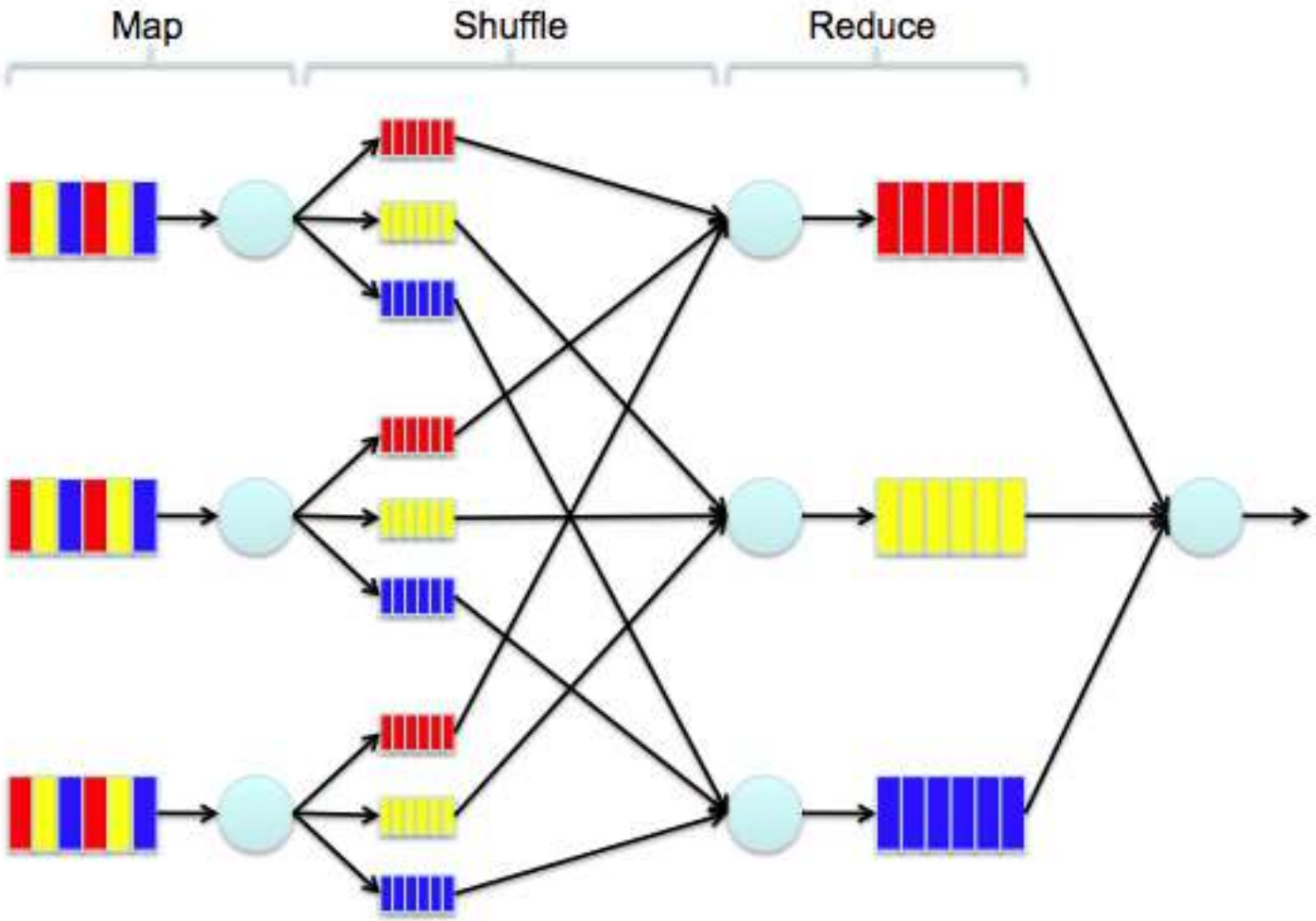


MapReduce

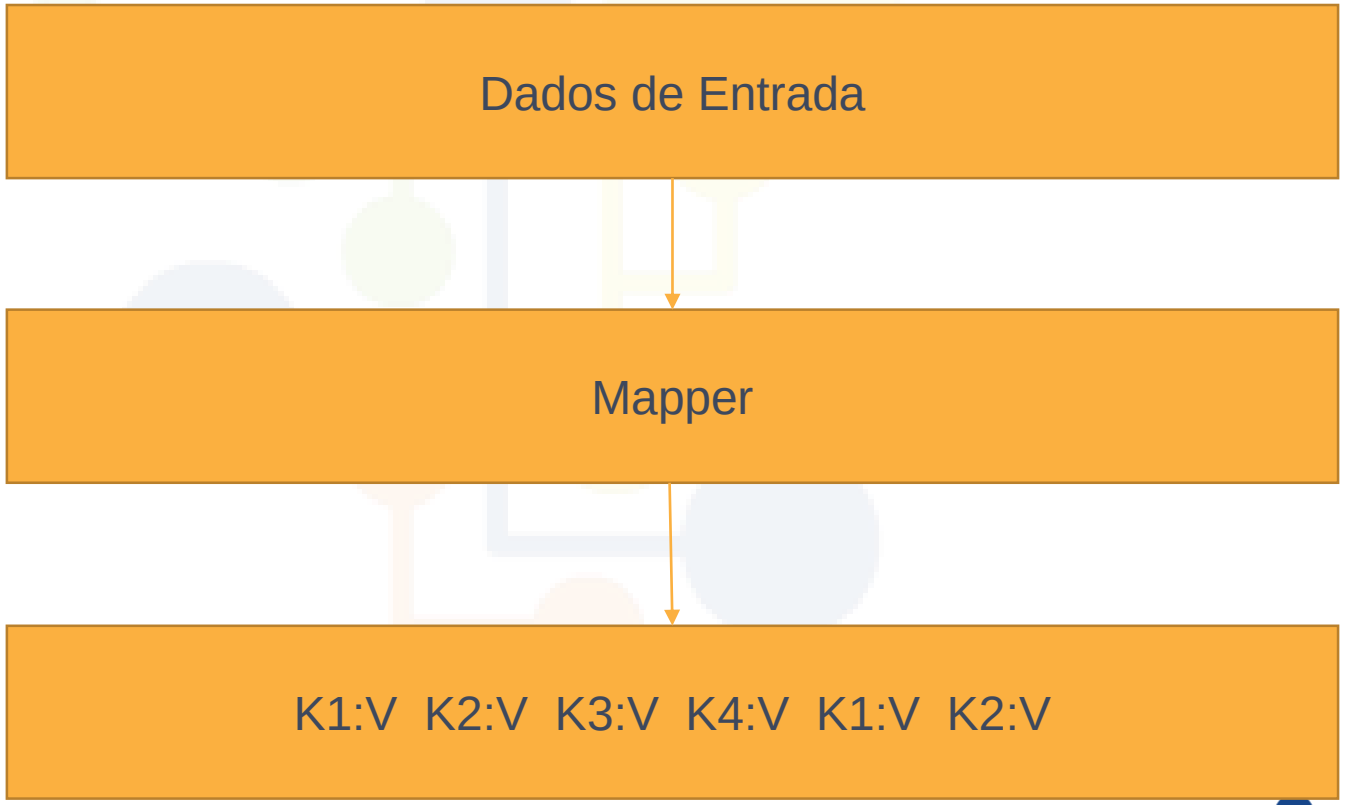
Como exatamente funciona o modelo MapReduce?

Data Science Academy

MapReduce



MapReduce



MapReduce

Quem define o que será a chave e o que será o valor?

Você, Cientista de Dados!

Data Science Academy

MapReduce

Dataset MovieLens (u.data)

userID	movieID	rating	timestamp
241	198	2	981769876
197	302	3	781769876
197	378	4	751769876
186	153	4	721769876
165	349	3	741769876
187	472	1	681769876
187	267	2	581769876

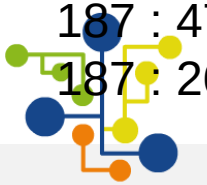
Quantos filmes cada pessoa assistiu?

Mapper

Key : Value

userID : movieID

241 : 198
197 : 302
197 : 378
186 : 153
165 : 349
187 : 472
187 : 267



MapReduce

Quantos filmes cada pessoa assistiu?

Mapper	Shuffle, Sort	Reduce
Key : Value userID : movieID 241 : 198 197 : 302 197 : 378 186 : 153 165 : 349 187 : 472 187 : 267	241 : 198 197 : 302, 378 186 : 153 165 : 349 187 : 472, 267	241 : 1 197 : 2 186 : 1 165 : 1 187 : 2

Data Science Academy

MapReduce

Data Science Academy

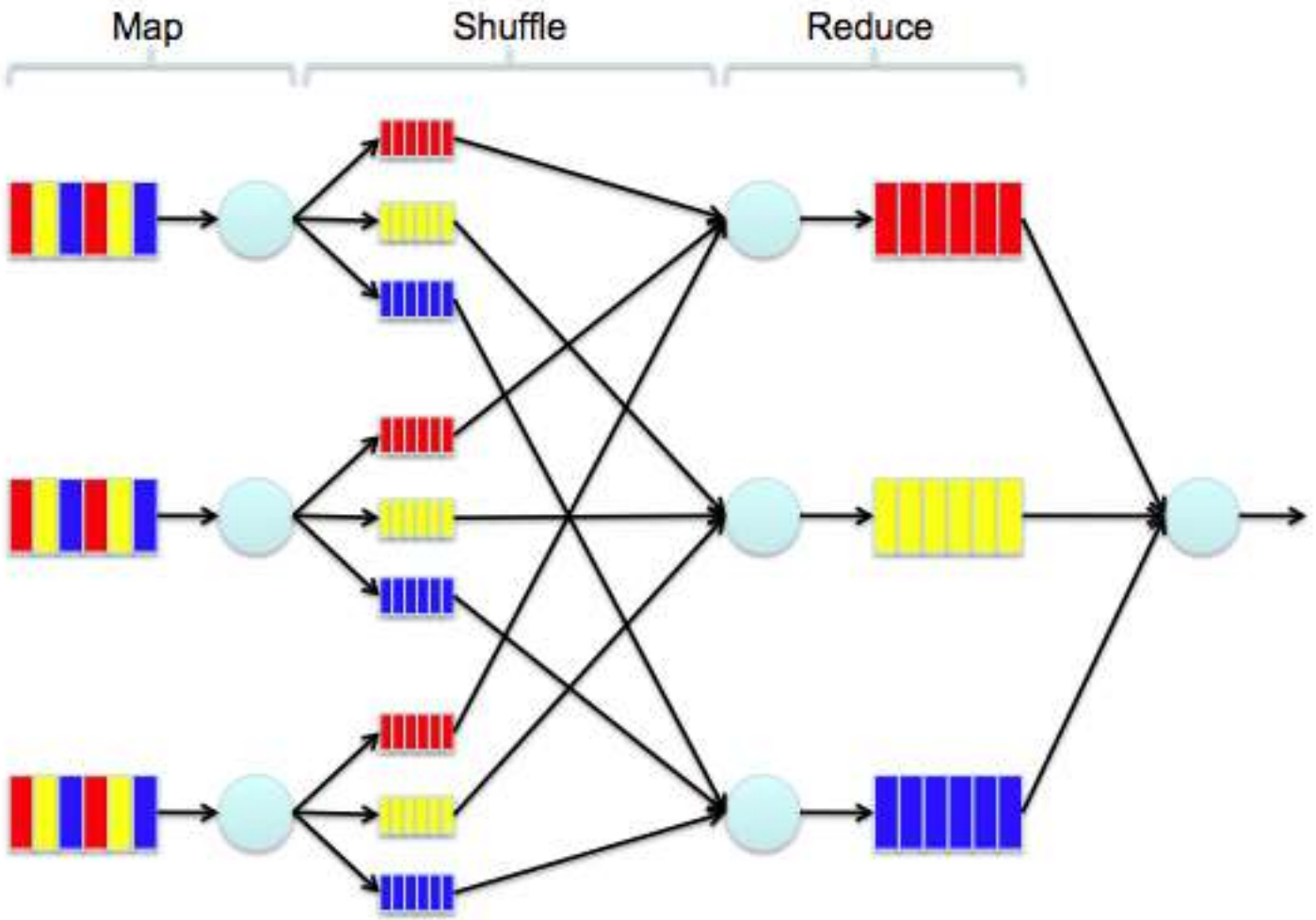
MapReduce

Quantos filmes cada pessoa assistiu?

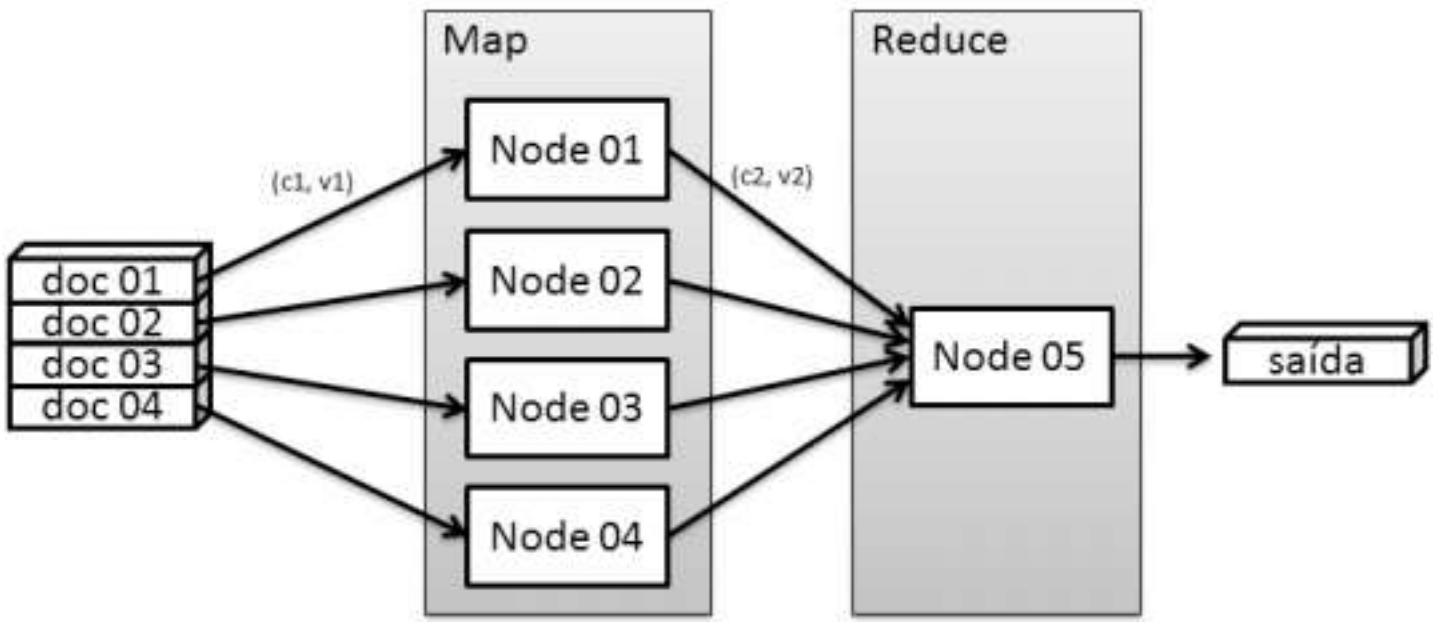
Mapper	Shuffle, Sort	Reduce
Key : Value		
userID : movieID		
241 : 198	241 : 198	241 : 1
197 : 302	197 : 302, 378	197 : 2
197 : 378	186 : 153	186 : 1
186 : 153	165 : 349	165 : 1
165 : 349	187 : 472, 267	187 : 2
187 : 472		
187 : 267		

Data Science Academy

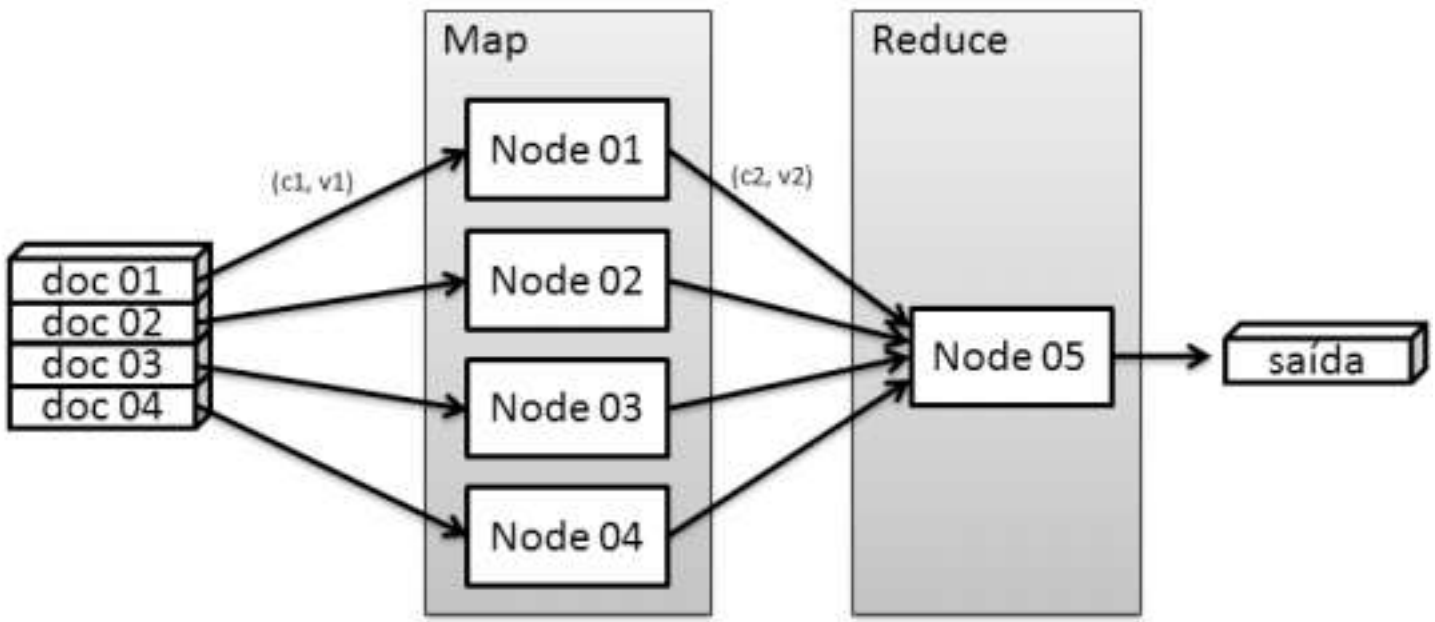
MapReduce



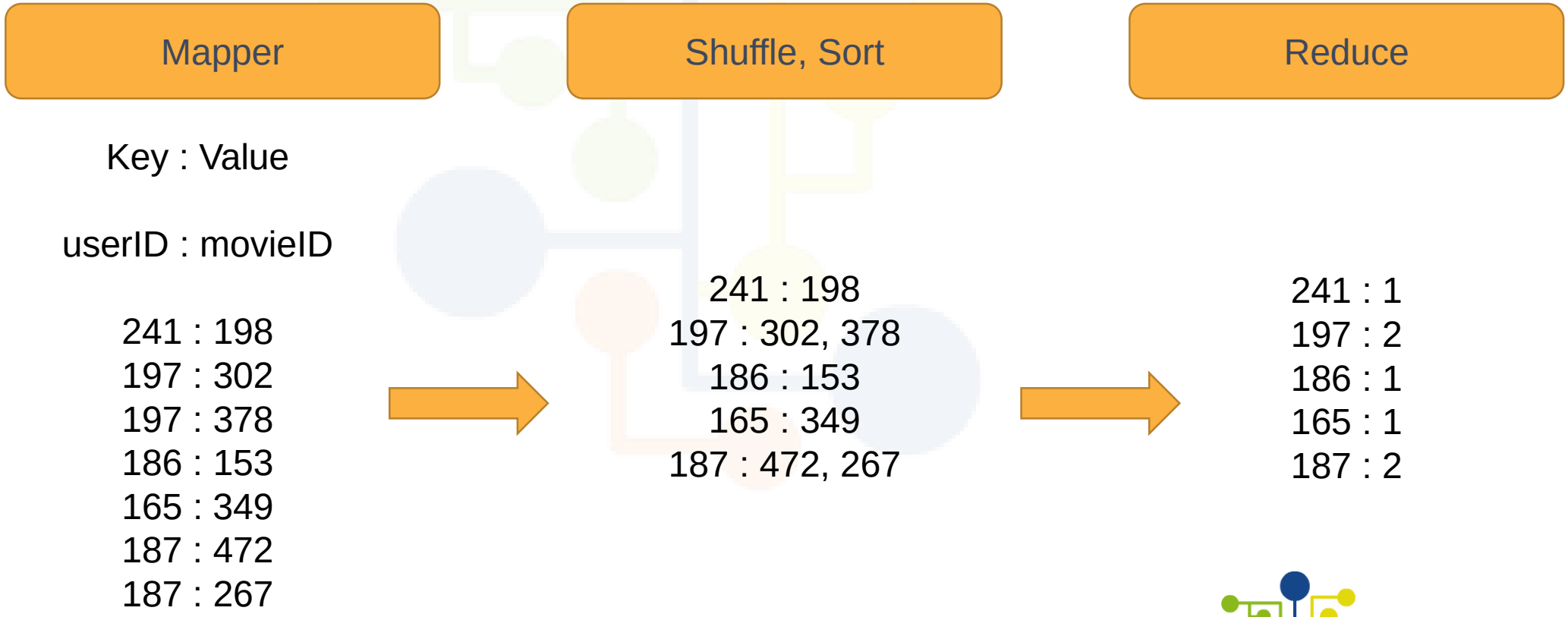
MapReduce


Map
Reduce
(c1, v1)
(c2, v2)
doc 01
doc 02
doc 03
doc 04
Node 01
Node 02
Node 03
Node 04
Node 05
saída

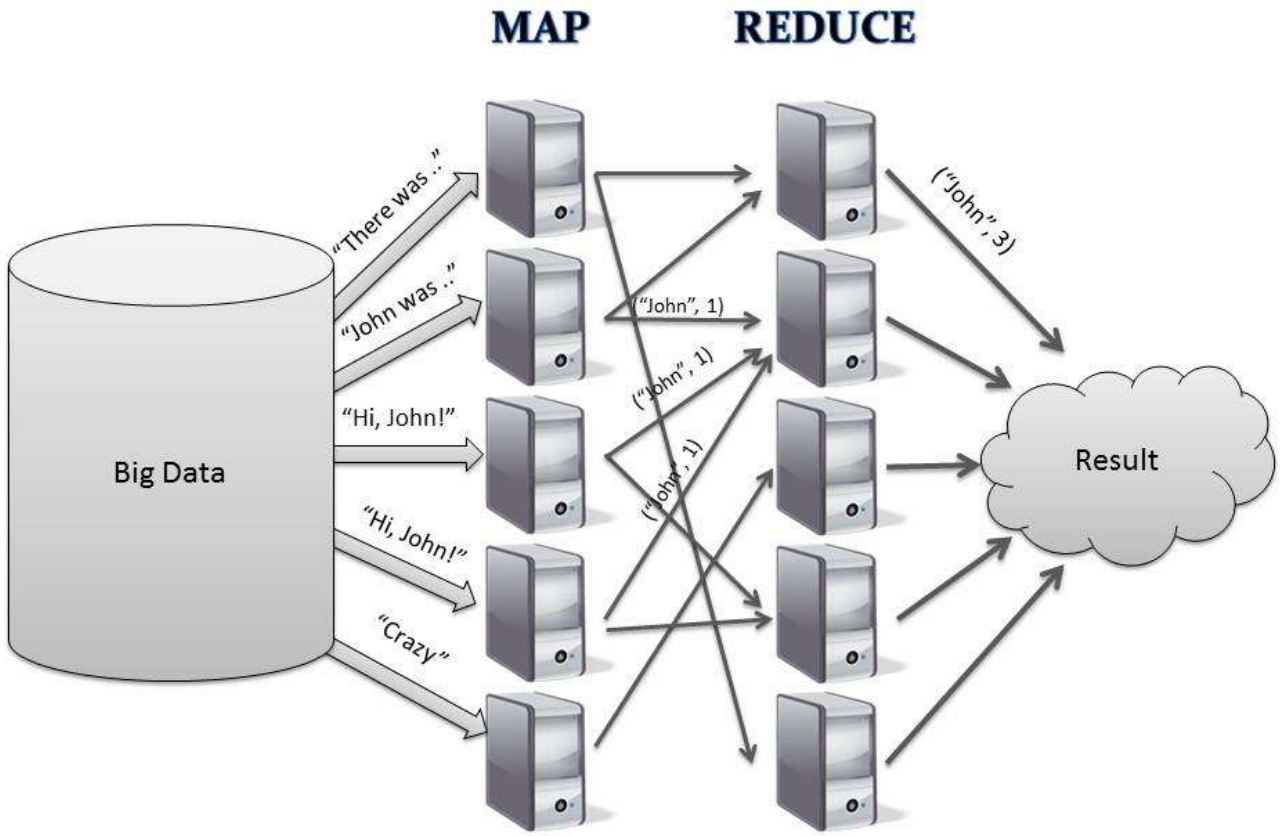
MapReduce


Map
Reduce
(c1, v1)
(c2, v2)
doc 01
doc 02
doc 03
doc 04
Node 01
Node 02
Node 03
Node 04
Node 05
saída

MapReduce


Mapper
Shuffle, Sort
Reduce
Key : Value
userID : movieID
241 : 198
197 : 302
197 : 378
186 : 153
165 : 349
187 : 472
187 : 267
241 : 198
197 : 302, 378
186 : 153
165 : 349
187 : 472, 267
241 : 1
197 : 2
186 : 1
165 : 1
187 : 2

MapReduce



Data Science Academy

Como o MapReduce Utiliza a Computação Distribuída

Data Science Academy

MapReduce

A operação do MapReduce inclui

Comunicação eficiente entre os nodes

Tratamento de erros e problemas de performance

Computação paralela através do cluster

Funções de mapeamento e redução

MapReduce

Workflow do MapReduce

Agendamento	Localização de Tarefas	Tratamento de Erros	Sincronização de Dados
Os jobs são divididos em pedaços menores chamados tarefas As tarefas são agendadas pelo TaskTracker	As tarefas são colocadas nos nodes que armazenam os segmentos de dados O código é movido para onde o dado está	Falhas são um comportamento esperado e no caso de falhas, as tarefas são automaticamente enviadas a outros nodes	Os dados são randomicamente agrupados e movidos entre os nodes Input e output são coordeandos pelo framework

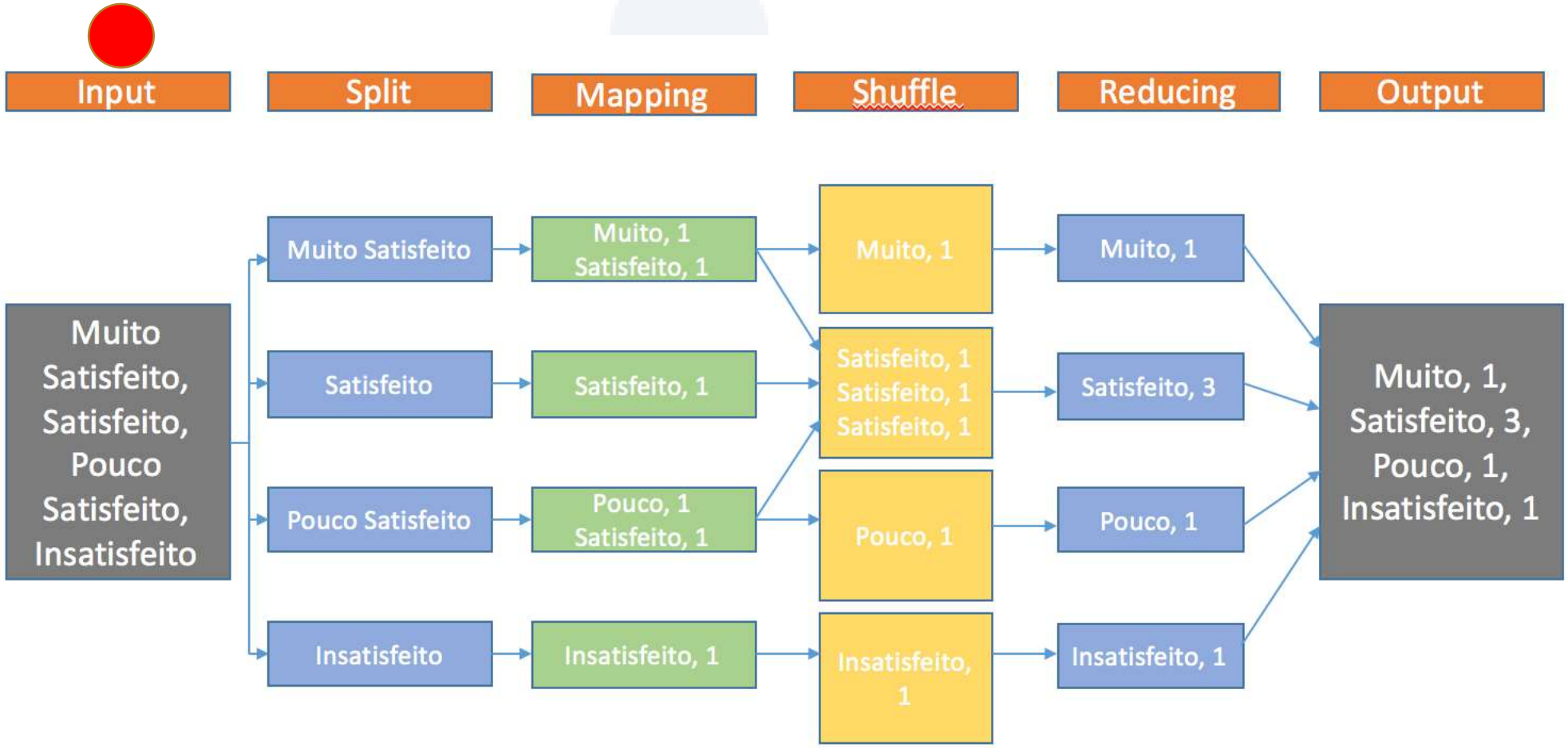
Data Science Academy

MapReduce

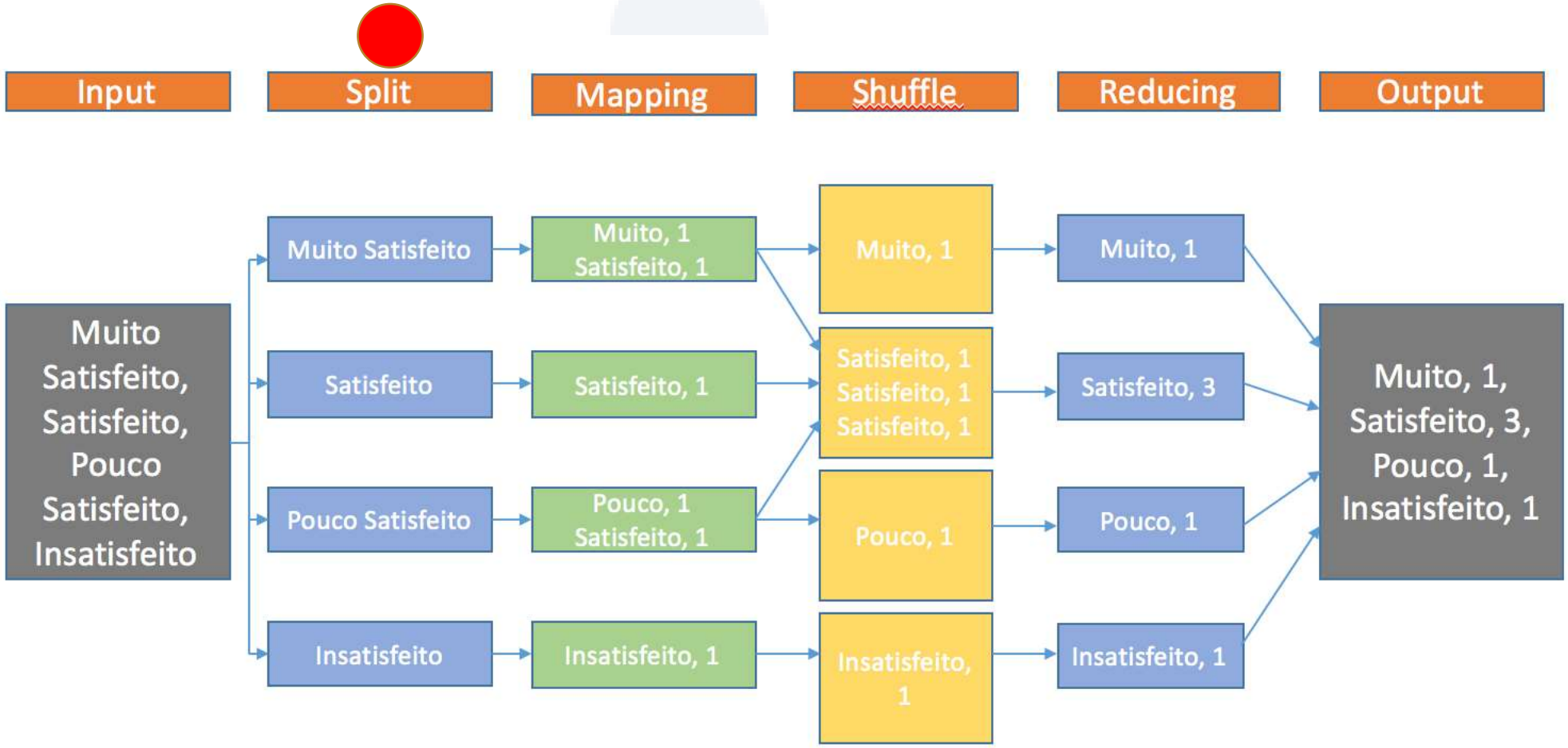
{Muito Satisfeito, Satisfeito, Pouco Satisfeito, Insatisfeito}

Data Science Academy

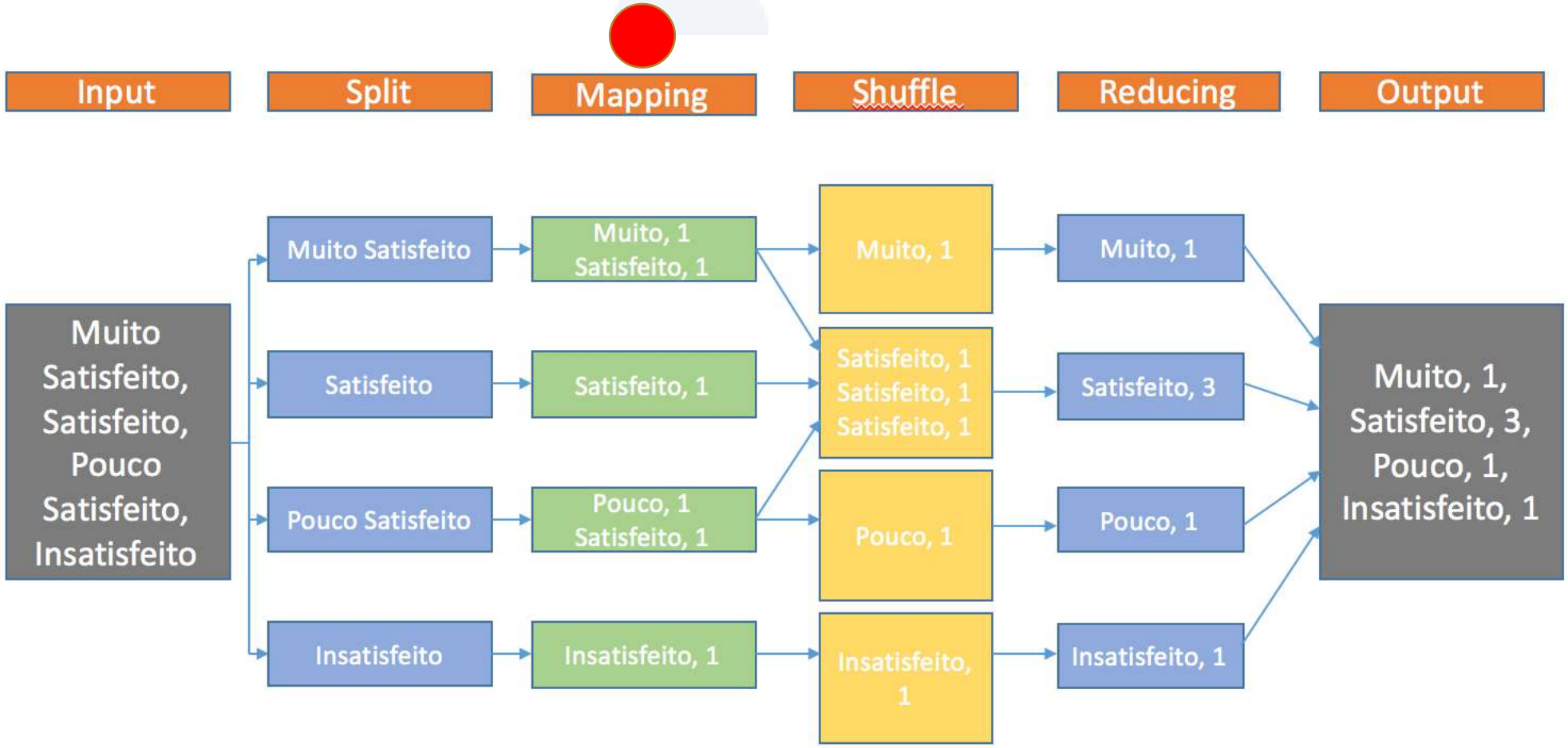
MapReduce


Input
Split
Mapping
Shuffle
Reducing
Output
Muito Satisfeito, Satisfeito, Pouco Satisfeito, Insatisfeito
Muito Satisfeito
Satisfeito
Pouco Satisfeito
Insatisfeito
Muito, 1 Satisfeito, 1
Satisfeito, 1
Pouco, 1 Satisfeito, 1
Insatisfeito, 1
Muito, 1
Satisfeito, 1 Satisfeito, 1 Satisfeito, 1
Pouco, 1
Insatisfeito, 1
Muito, 1
Satisfeito, 3
Pouco, 1
Insatisfeito, 1
Muito, 1, Satisfeito, 3, Pouco, 1, Insatisfeito, 1

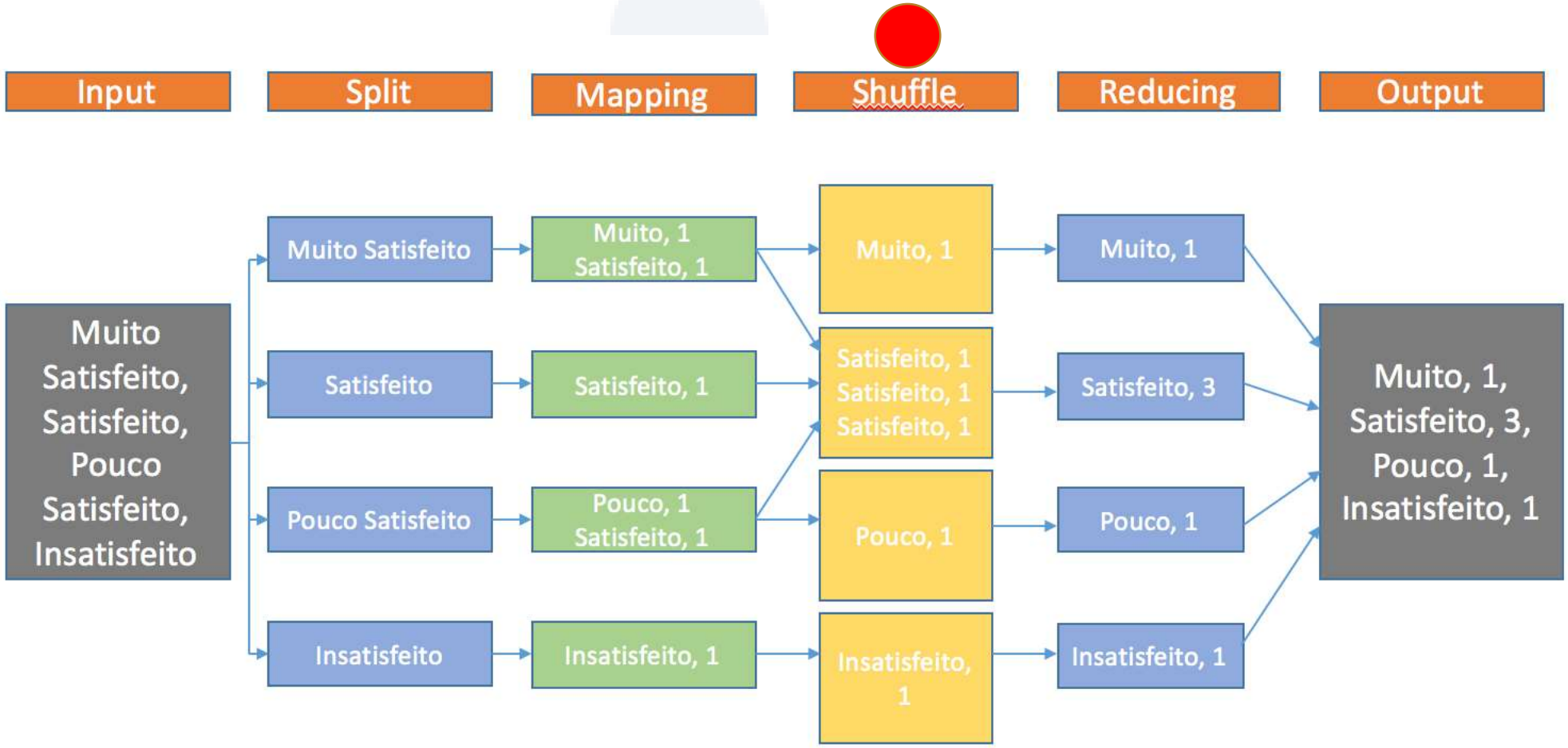
MapReduce


Input
Split
Mapping
Shuffle
Reducing
Output
Muito Satisfeito, Satisfeito, Pouco Satisfeito, Insatisfeito
Muito Satisfeito
Satisfeito
Pouco Satisfeito
Insatisfeito
Muito, 1 Satisfeito, 1
Satisfeito, 1
Pouco, 1 Satisfeito, 1
Insatisfeito, 1
Muito, 1
Satisfeito, 1 Satisfeito, 1 Satisfeito, 1
Pouco, 1
Insatisfeito, 1
Muito, 1
Satisfeito, 3
Pouco, 1
Insatisfeito, 1
Muito, 1, Satisfeito, 3, Pouco, 1, Insatisfeito, 1

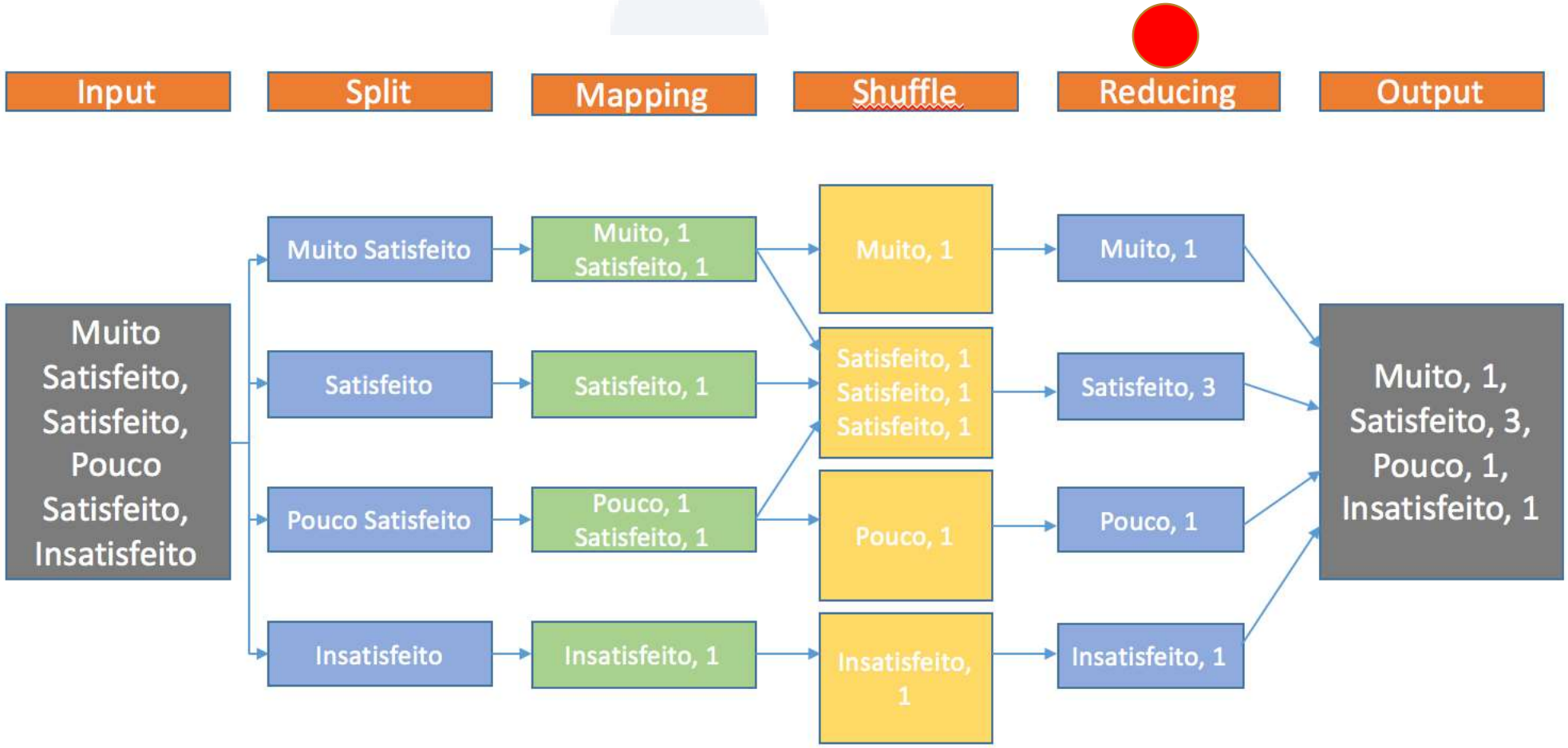
MapReduce


Input
Split
Mapping
Shuffle
Reducing
Output
Muito Satisfeito, Satisfeito, Pouco Satisfeito, Insatisfeito
Muito Satisfeito
Satisfeito
Pouco Satisfeito
Insatisfeito
Muito, 1 Satisfeito, 1
Satisfeito, 1
Pouco, 1 Satisfeito, 1
Insatisfeito, 1
Muito, 1
Satisfeito, 1 Satisfeito, 1 Satisfeito, 1
Pouco, 1
Insatisfeito, 1
Muito, 1
Satisfeito, 3
Pouco, 1
Insatisfeito, 1
Muito, 1, Satisfeito, 3, Pouco, 1, Insatisfeito, 1

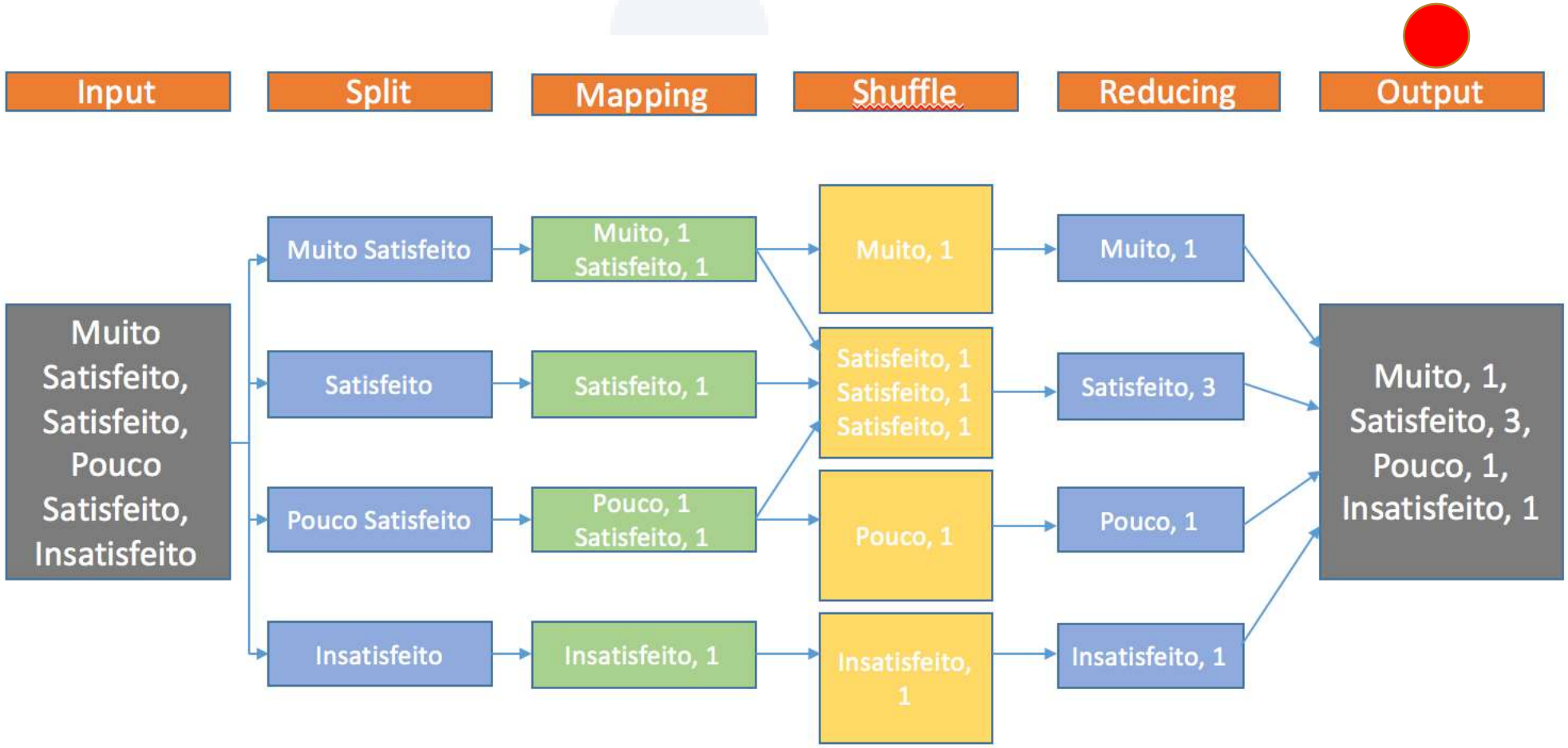
MapReduce


Input
Split
Mapping
Shuffle
Reducing
Output
Muito Satisfeito, Satisfeito, Pouco Satisfeito, Insatisfeito
Muito Satisfeito
Satisfeito
Pouco Satisfeito
Insatisfeito
Muito, 1 Satisfeito, 1
Satisfeito, 1
Pouco, 1 Satisfeito, 1
Insatisfeito, 1
Muito, 1
Satisfeito, 1 Satisfeito, 1 Satisfeito, 1
Pouco, 1
Insatisfeito, 1
Muito, 1
Satisfeito, 3
Pouco, 1
Insatisfeito, 1
Muito, 1, Satisfeito, 3, Pouco, 1, Insatisfeito, 1

MapReduce


Input
Split
Mapping
Shuffle
Reducing
Output
Muito Satisfeito, Satisfeito, Pouco Satisfeito, Insatisfeito
Muito Satisfeito
Satisfeito
Pouco Satisfeito
Insatisfeito
Muito, 1 Satisfeito, 1
Satisfeito, 1
Pouco, 1 Satisfeito, 1
Insatisfeito, 1
Muito, 1
Satisfeito, 1 Satisfeito, 1 Satisfeito, 1
Pouco, 1
Insatisfeito, 1
Muito, 1
Satisfeito, 3
Pouco, 1
Insatisfeito, 1
Muito, 1, Satisfeito, 3, Pouco, 1, Insatisfeito, 1

MapReduce


Input
Split
Mapping
Shuffle
Reducing
Output
Muito Satisfeito, Satisfeito, Pouco Satisfeito, Insatisfeito
Muito Satisfeito
Satisfeito
Pouco Satisfeito
Insatisfeito
Muito, 1 Satisfeito, 1
Satisfeito, 1
Pouco, 1 Satisfeito, 1
Insatisfeito, 1
Muito, 1
Satisfeito, 1 Satisfeito, 1 Satisfeito, 1
Pouco, 1
Insatisfeito, 1
Muito, 1
Satisfeito, 3
Pouco, 1
Insatisfeito, 1
Muito, 1, Satisfeito, 3, Pouco, 1, Insatisfeito, 1

MapReduce

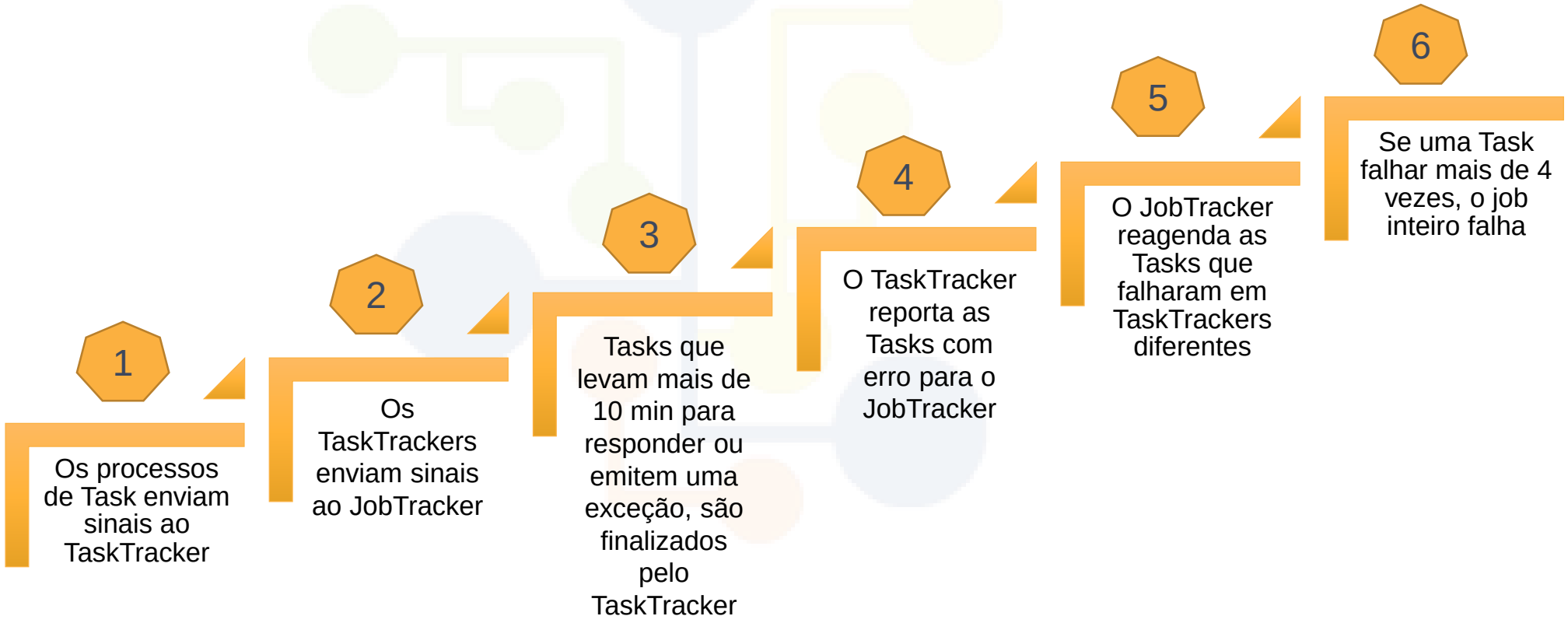
Características do MapReduce

Algumas das principais características do MapReduce:

- Consegue trabalhar com grandes volumes de dados
- Funciona bem com o conceito WORM (Write Once and Read Many)
- Permite paralelismo
- As operações são realizadas próximas dos dados
- Hardware e storage de baixo custo podem ser usados
- O runtime fica responsável por dividir e mover os dados para as operações

MapReduce

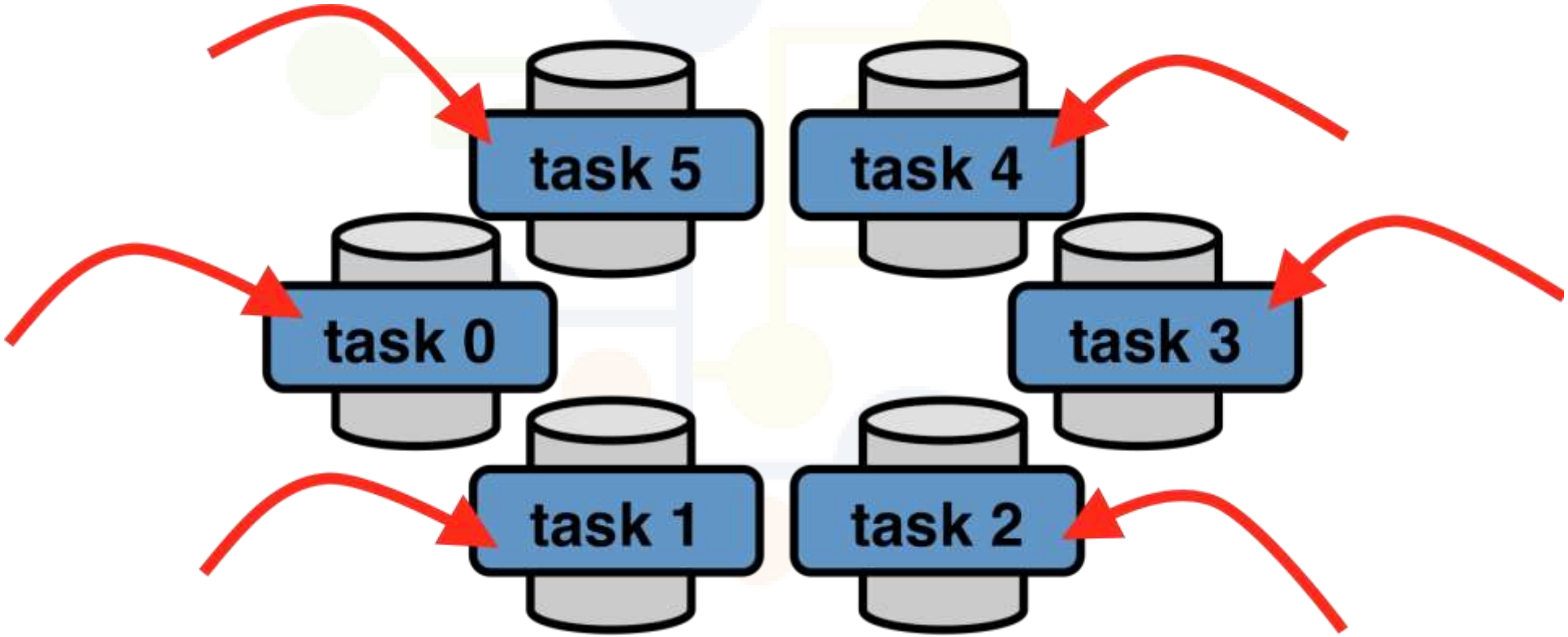
Processo de Recuperação a Falhas do MapReduce



YARN

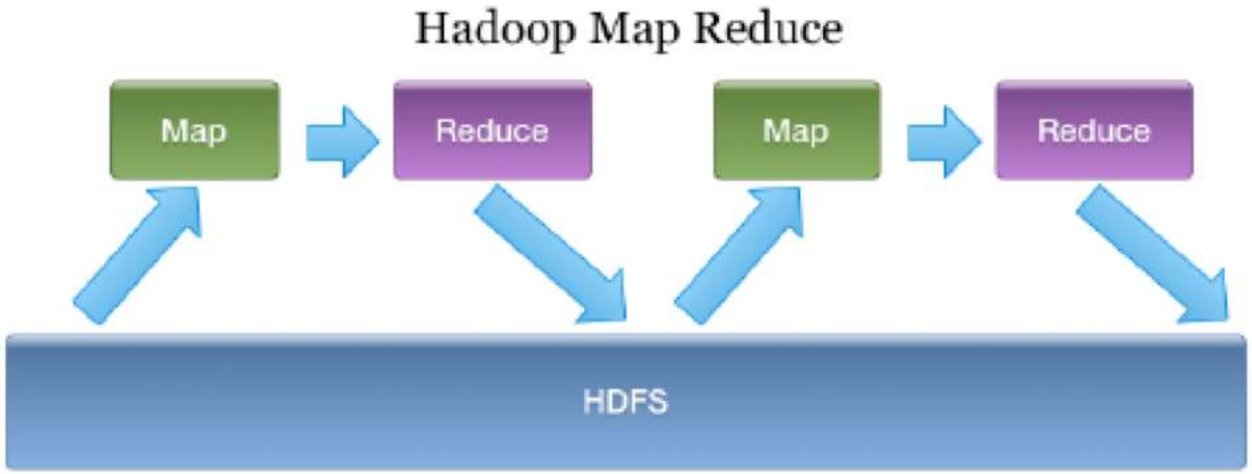
Data Science Academy

YARN

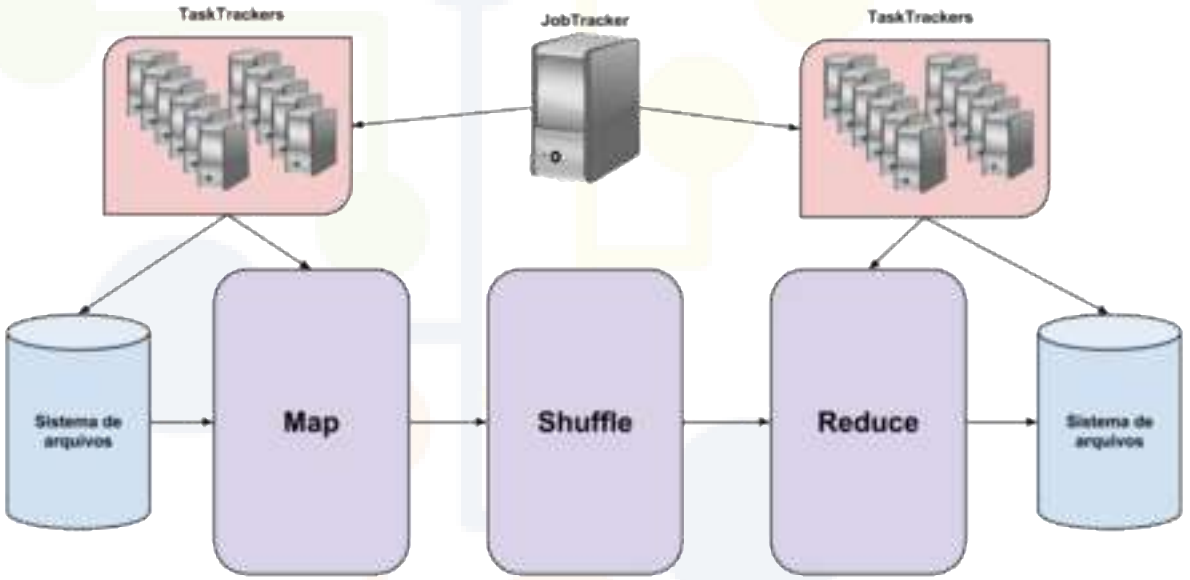


Data Science Academy

YARN



YARN


TaskTrackers
JobTracker
TaskTrackers
Sistema de arquivos
Map
Shuffle
Reduce
Sistema de arquivos

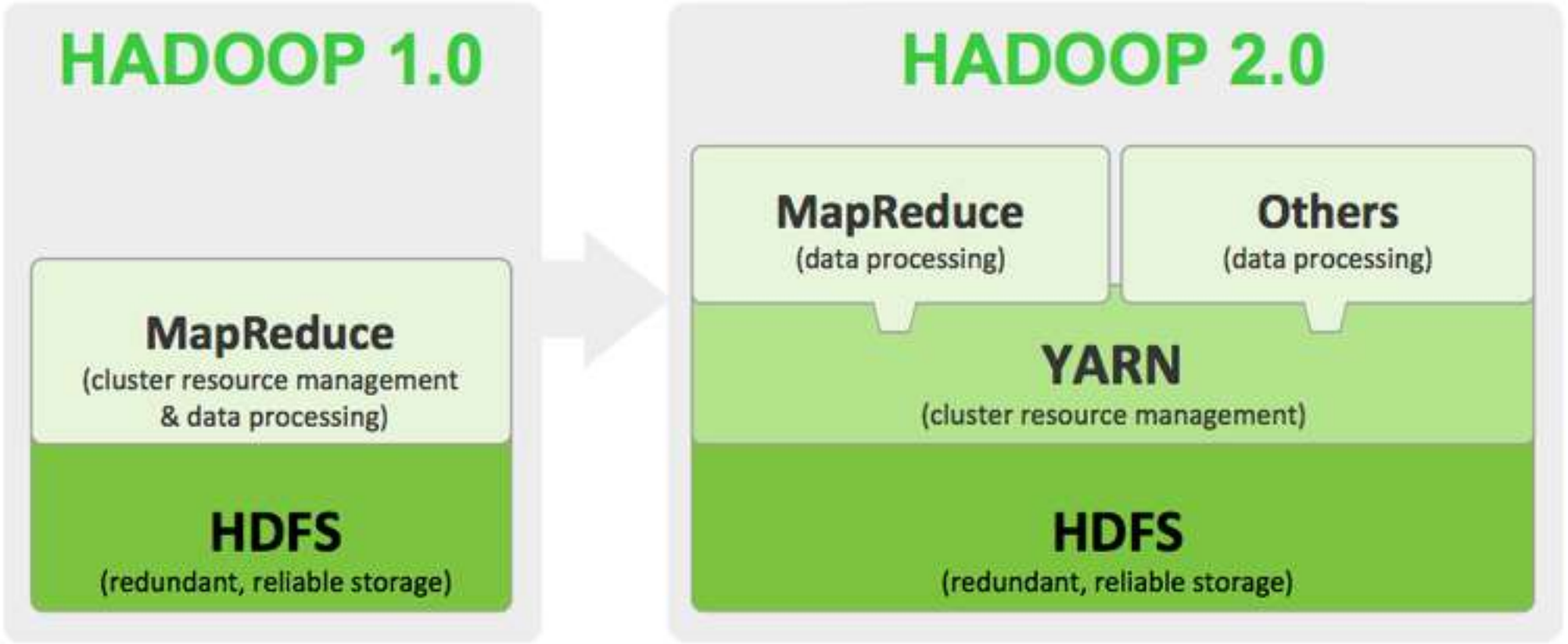
YARN

YARN

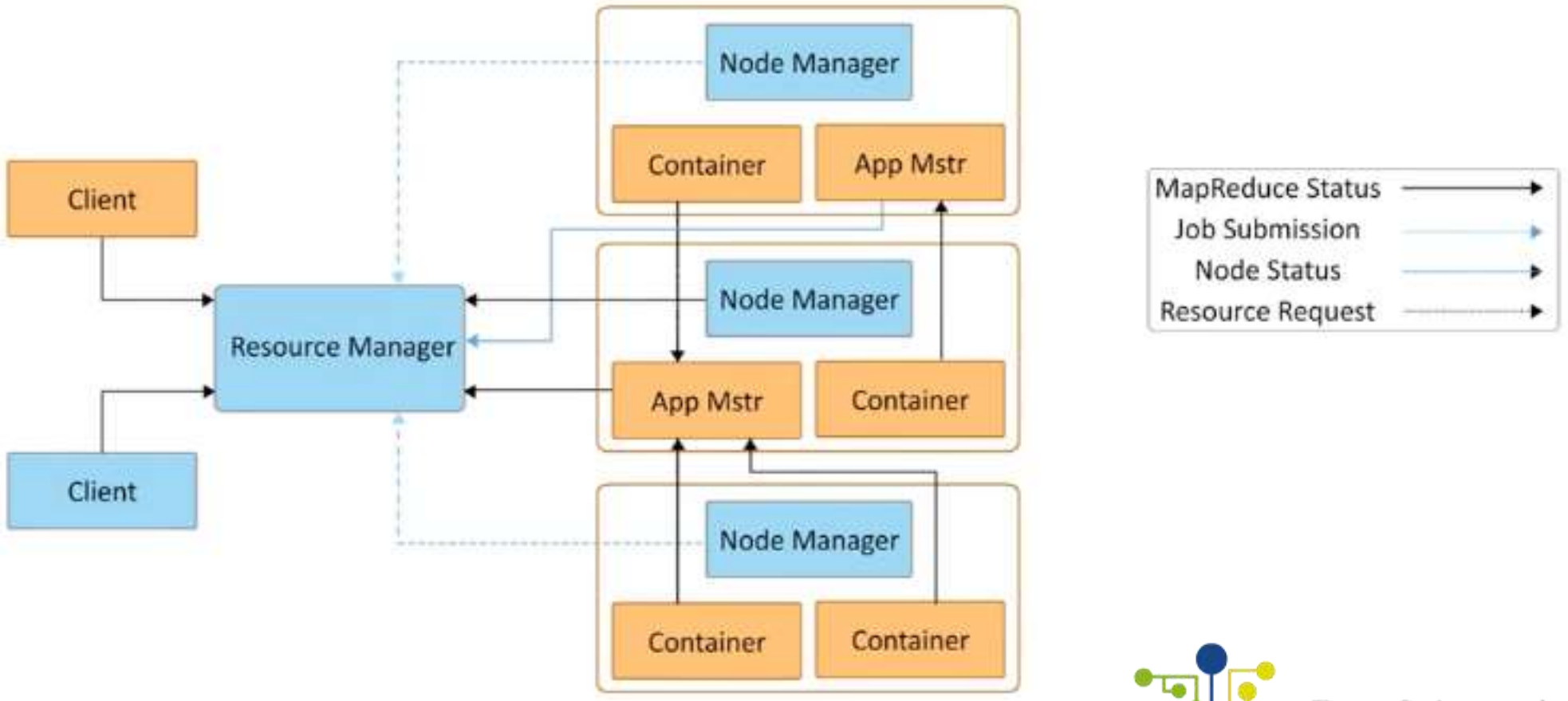
Mas uma nova ferramenta, chamada ”**Y**et **A**nother **R**esource **N**egotiatior” (YARN) foi introduzida com o Hadoop 2.0 e a gestão de recursos foi transferida do MapReduce para o YARN

Data Science Academy

YARN


HADOOP 1.0
MapReduce
(cluster resource management & data processing)
HDFS
(redundant, reliable storage)
HADOOP 2.0
MapReduce
(data processing)
Others
(data processing)
YARN
(cluster resource management)
HDFS
(redundant, reliable storage)

YARN


Client
Client
Resource Manager
Node Manager
Container
App Mstr
Node Manager
App Mstr
Container
Node Manager
Container
Container
MapReduce Status
Job Submission
Node Status
Resource Request

YARN

YARN

A necessidade do YARN surgiu a partir de problemas com a gestão de recursos em versões anteriores do MapReduce:

- A utilização do cluster era baixa durante as tarefas de MapReduce
- Recursos não eram compartilhados
- Haviam problemas com a preferência de tarefas a serem executadas
- Havia apenas 1 JobTracker por cluster

YARN

Benefícios de utilização do YARN

- Os recursos são aplicados de forma mais eficiente e apenas quando necessário
- Aplicação MapReduce e Não-MapReduce podem ser executadas no mesmo cluster
- O conceito de Application Master foi introduzido e as responsabilidades do JobTracker redesenhadas

YARN

YARN Daemons

Resource Manager	Application Master
Executa no node Master	Executa nos nodes slaves
Agendador global de recursos	Comunica com o gerenciador de recursos

Data Science Academy

Obrigado

Data Science Academy

www.datascienceacademy.com.br