



Engenharia de Dados com Hadoop e Spark



Bem-vindo(a)





Usando MapReduce em Grandes Volumes de Dados

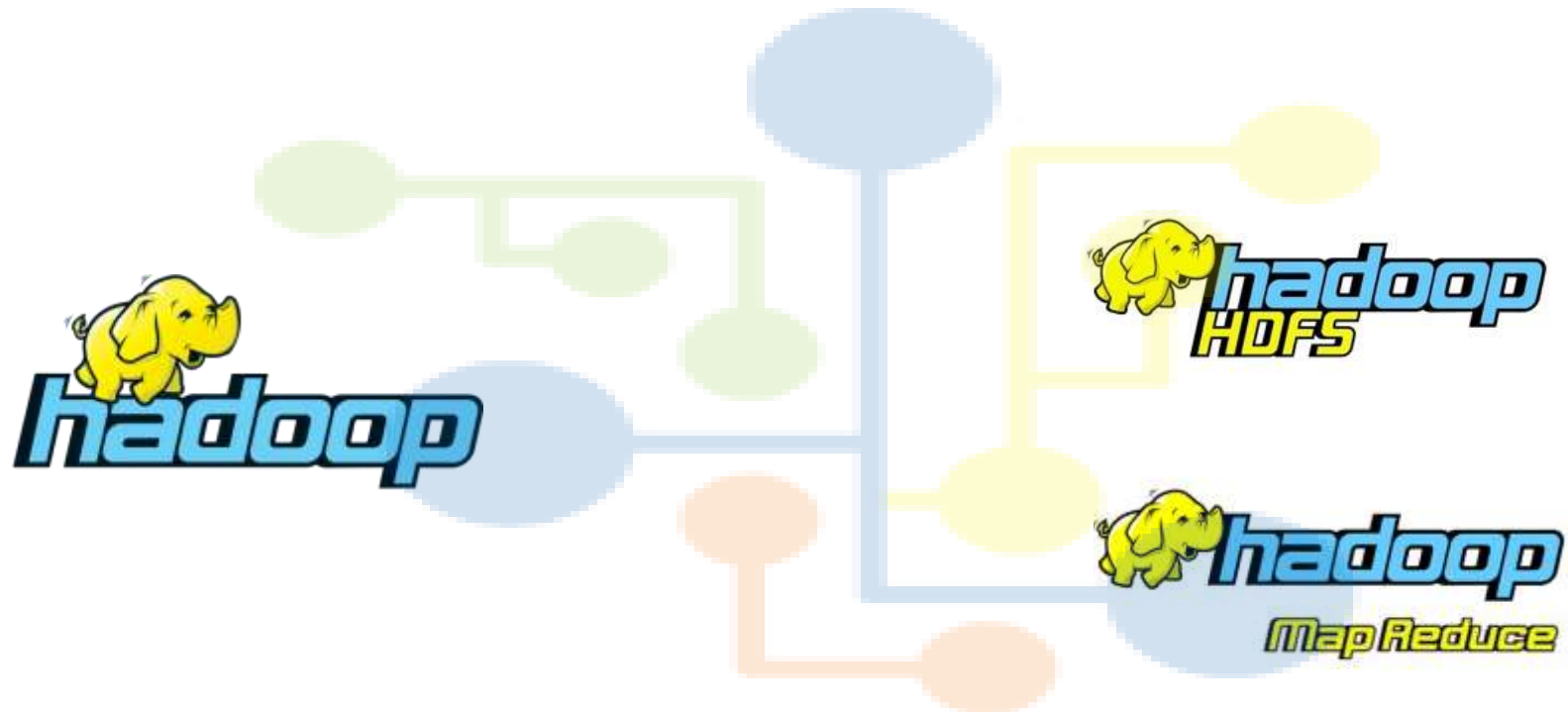
A decorative background graphic consisting of a network of nodes and lines. The nodes are colored in shades of blue, green, yellow, and orange, and are connected by lines of the same colors, forming a complex, interconnected structure.

MapReduce



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d





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

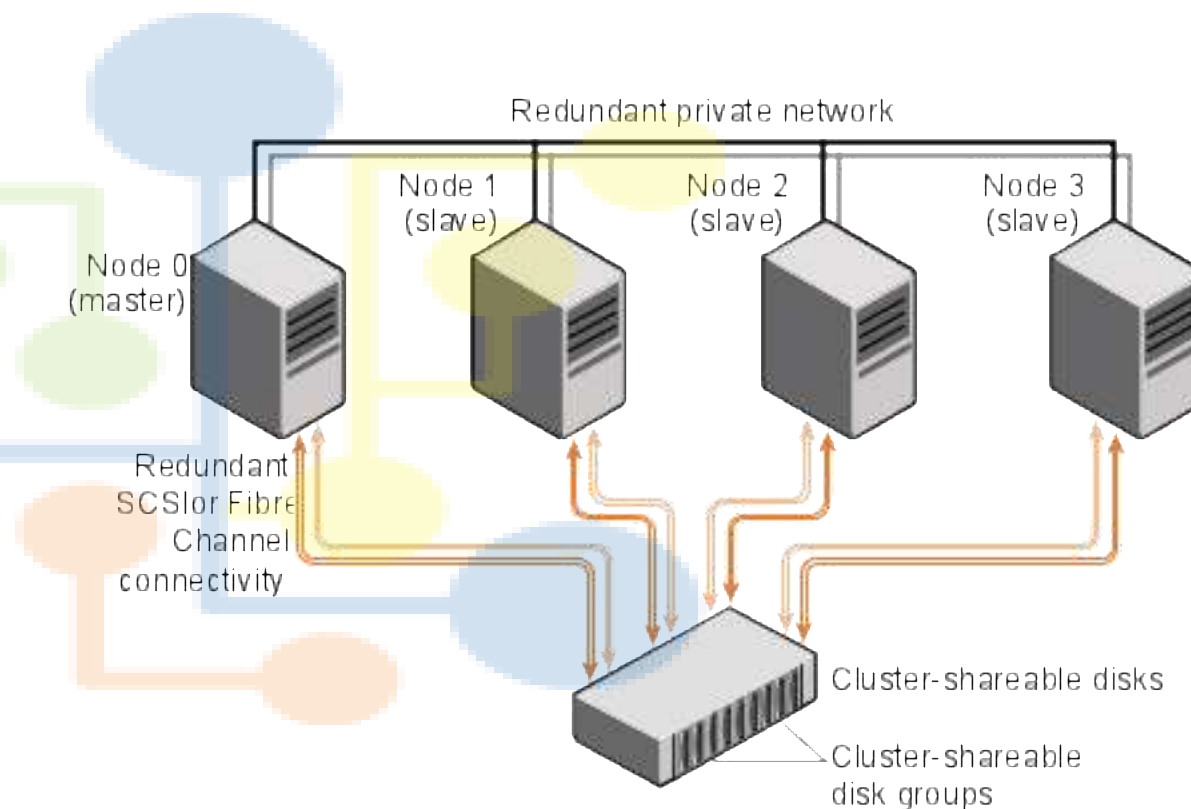


Computação Distribuída



Computação Distribuída

Sistema de Processamento
Distribuído e Paralelo





Computação Distribuída

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



Computação Distribuída



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



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

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



Computação Distribuída

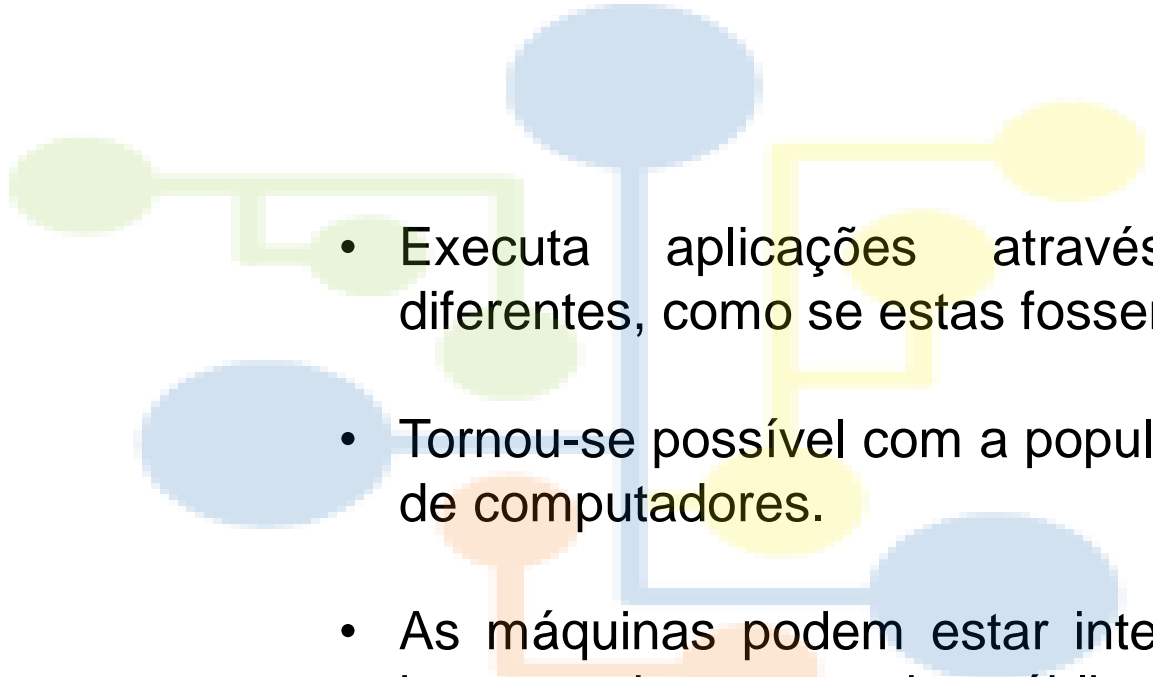
A computação distribuída consiste na utilização de um conjunto de máquinas conectadas por uma rede de comunicação, atuando como um único sistema.





Computação Distribuída

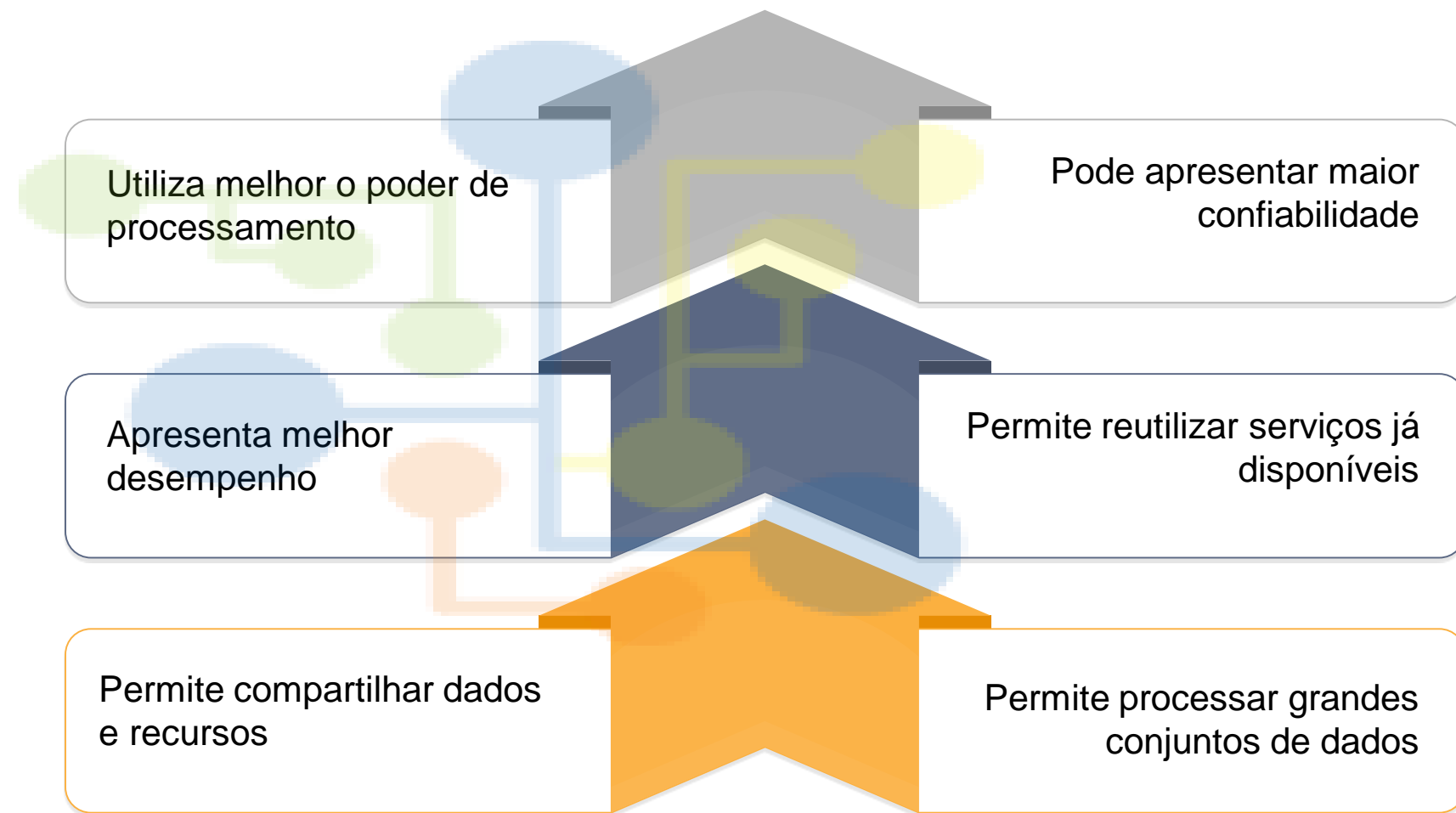
Computação Distribuída

- 
- A diagram illustrating a distributed network. It features several circular nodes of different colors (blue, green, yellow, orange) connected by lines. The nodes are arranged in a non-hierarchical, interconnected manner, representing a distributed system. Some nodes are larger than others, and the connections form a complex web.
- 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.



Computação Distribuída

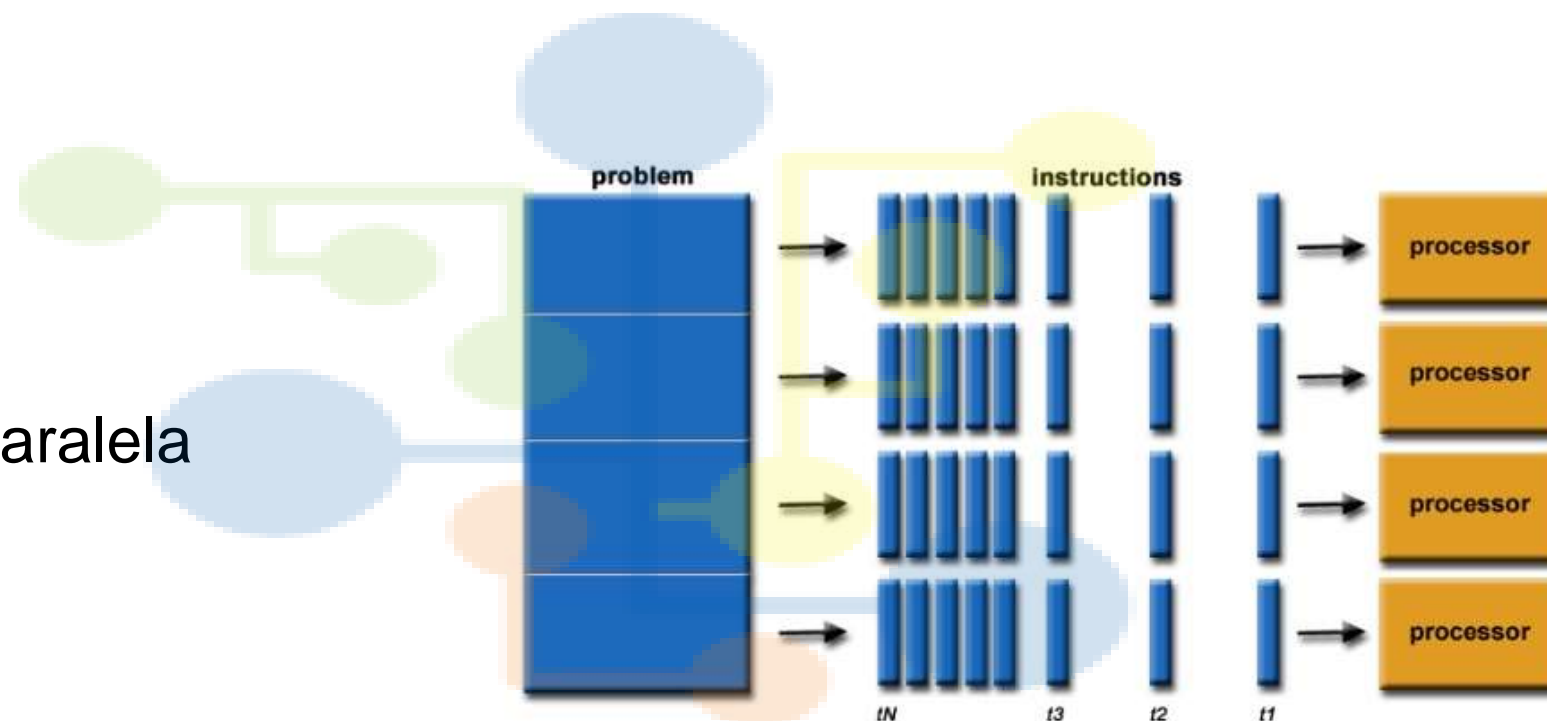
Computação Distribuída Vantagens





Computação Distribuída

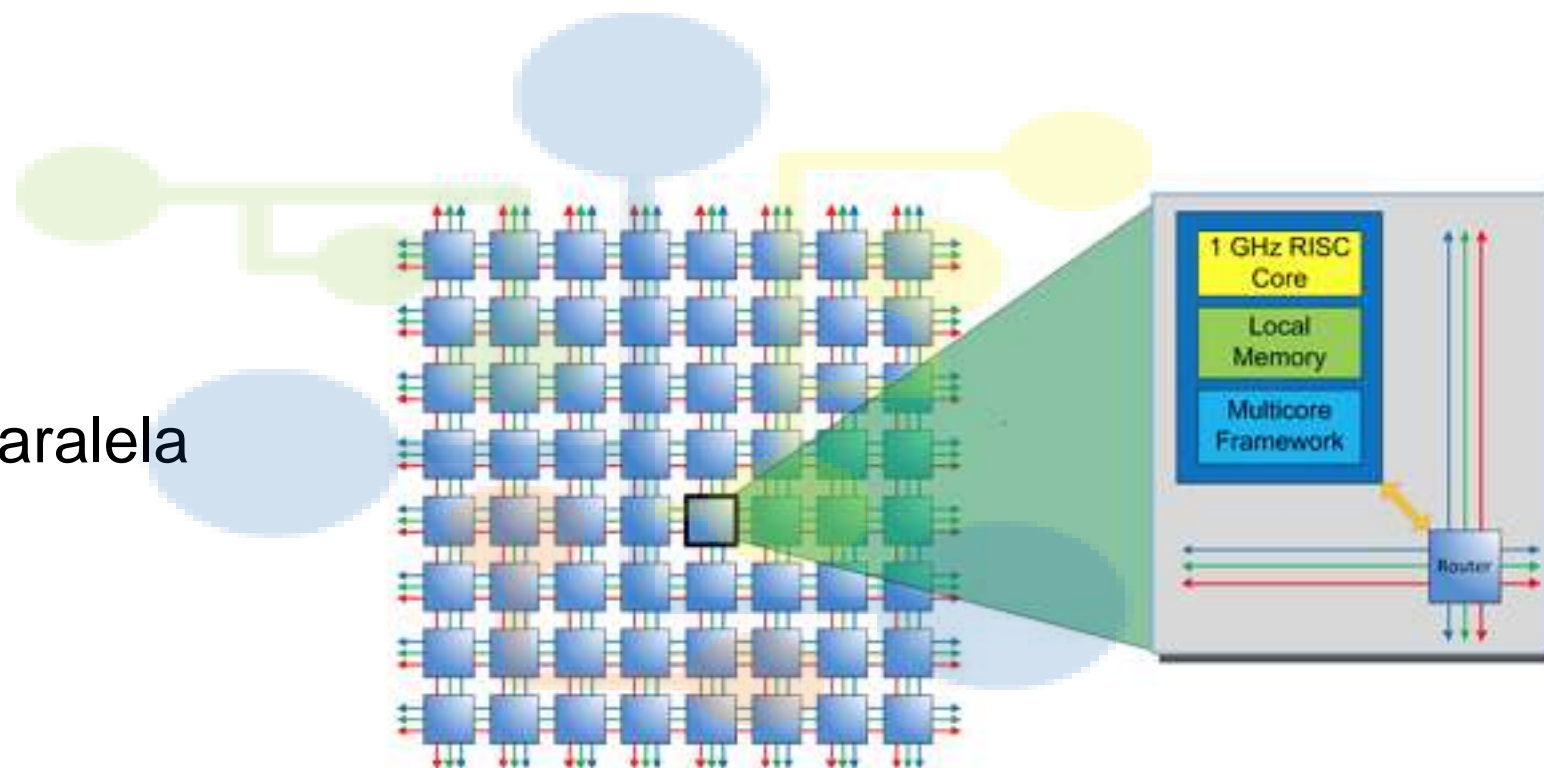
Computação Paralela





Computação Distribuída

Computação Paralela





Computação Distribuída

Programação Paralela em GPU



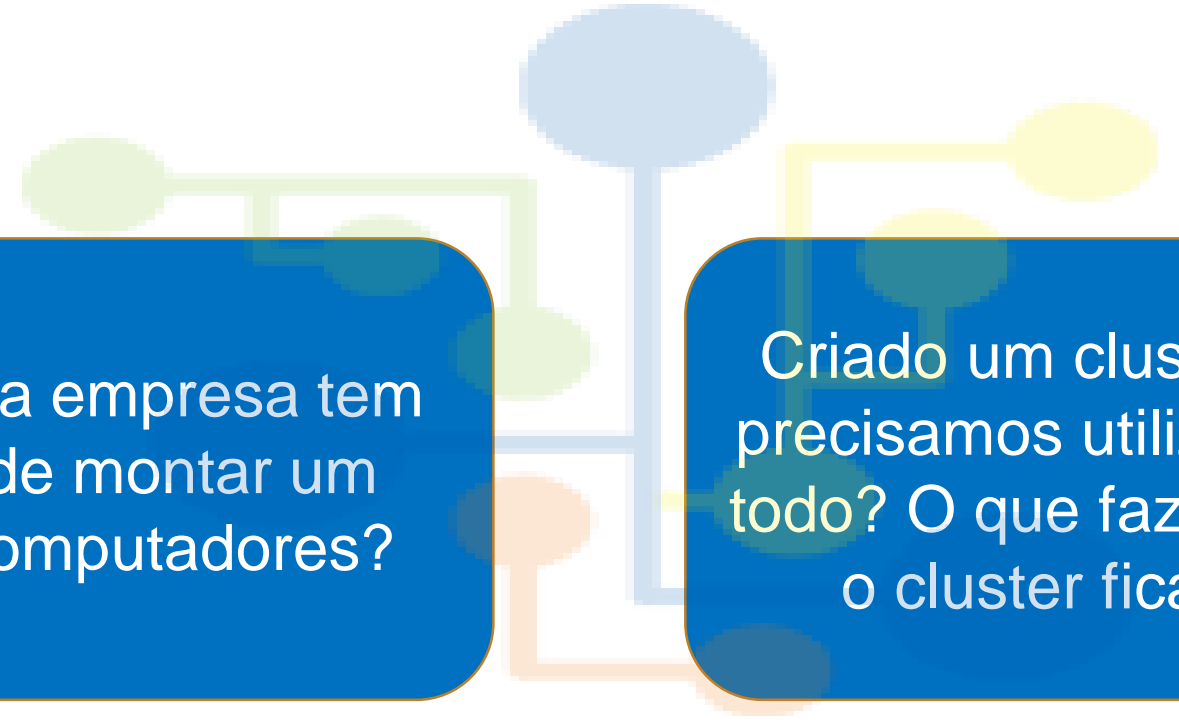


Computação Distribuída Cloud Computing

A faint, stylized network diagram in the background, consisting of several circular nodes connected by lines. The nodes are colored in shades of blue, green, yellow, and orange, and the lines are thin and light-colored, creating a subtle pattern behind the main text.



Computação Distribuída - Cloud Computing

A faint, stylized diagram in the background shows a network of nodes (circles) connected by lines, representing a distributed system or cloud architecture. The nodes are in various colors (blue, green, yellow, orange) and are interconnected in a non-linear fashion.

Será que uma empresa tem condições de montar um cluster de computadores?

Criado um cluster, será que precisamos utilizá-lo o tempo todo? O que fazemos quando o cluster ficar ocioso?



Computação Distribuída - Cloud Computing





Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

Computação Distribuída - Cloud Computing





Computação Distribuída - Cloud Computing

Infraestrutura como um serviço
(IaaS)





Computação Distribuída - Cloud Computing

Plataforma como um serviço
(PaaS)





Computação Distribuída - Cloud Computing

Software como um serviço
(SaaS)

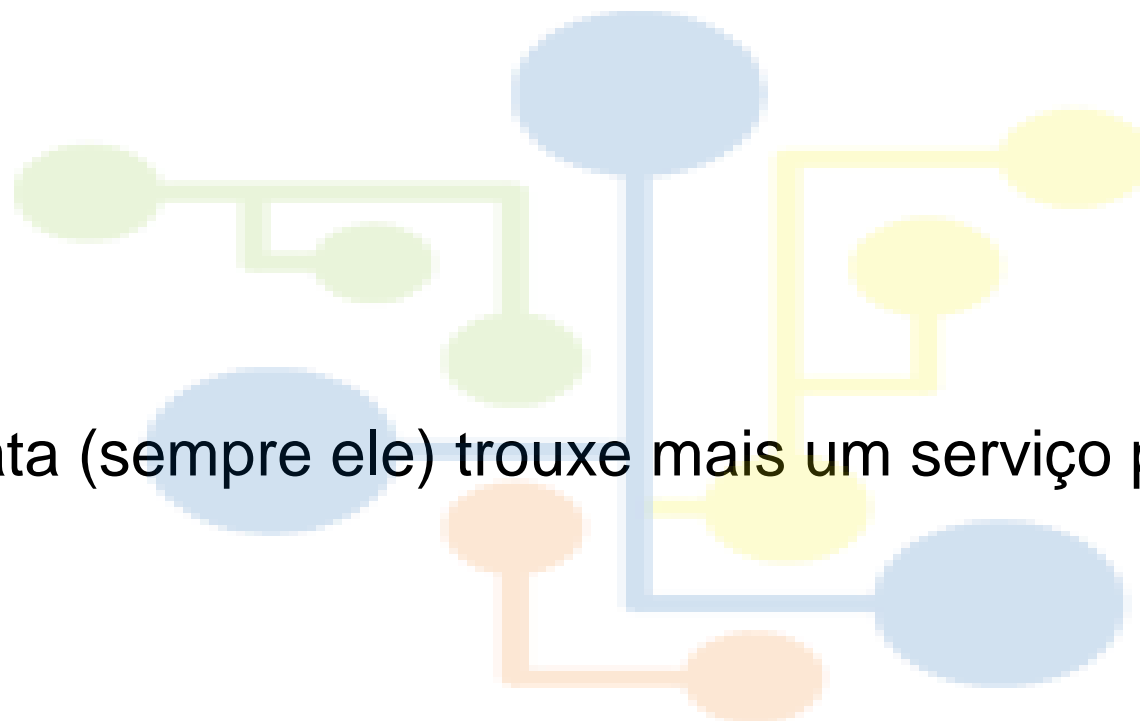
Pay as you go





Computação Distribuída - Cloud Computing

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





Computação Distribuída - Cloud Computing

Big Data como um serviço
(BDaaS)

Pay as you go





Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

Computação Distribuída - Cloud Computing





Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

Computação Distribuída - Cloud Computing

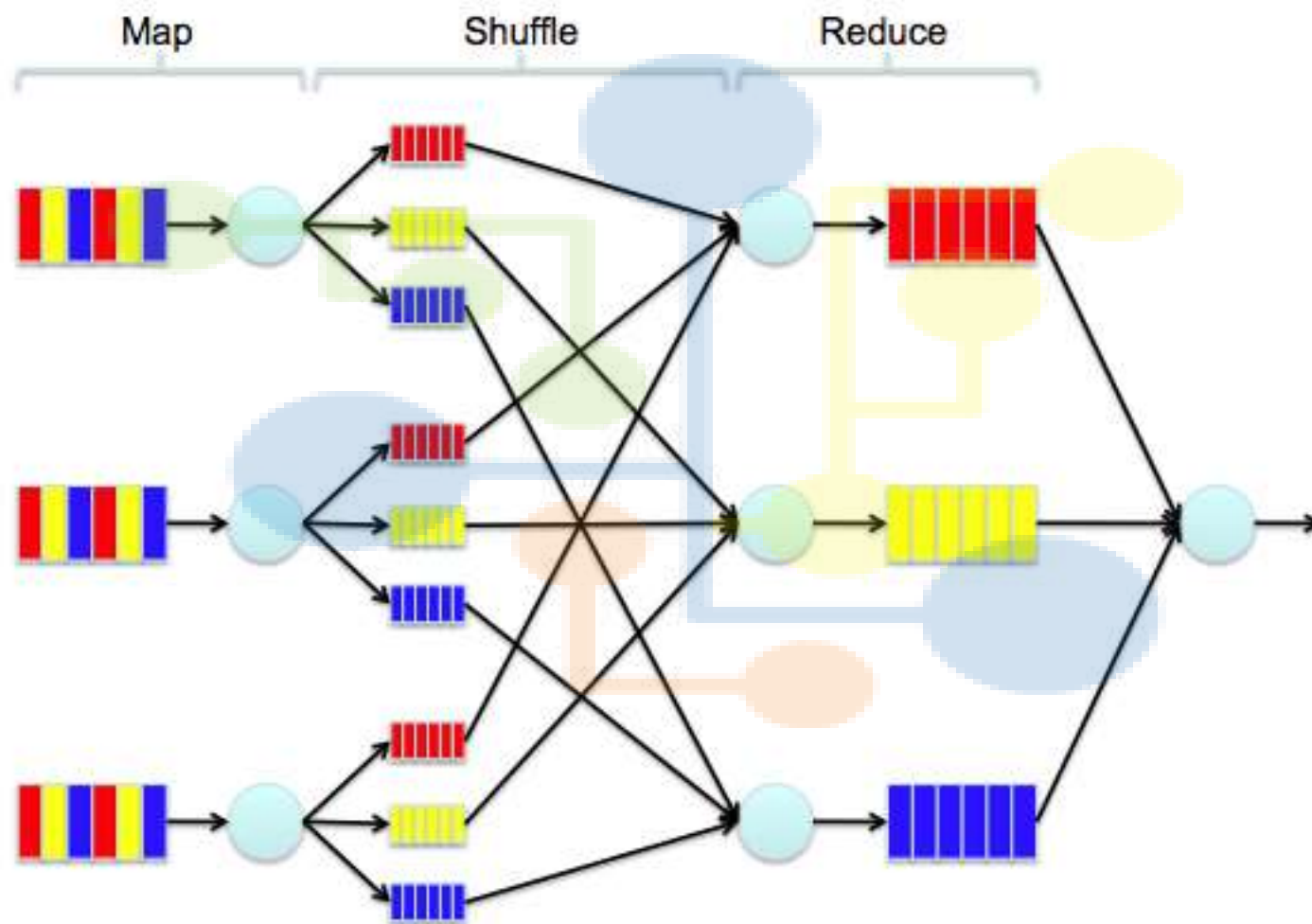




O Modelo de Programação MapReduce



O Modelo de Programação MapReduce





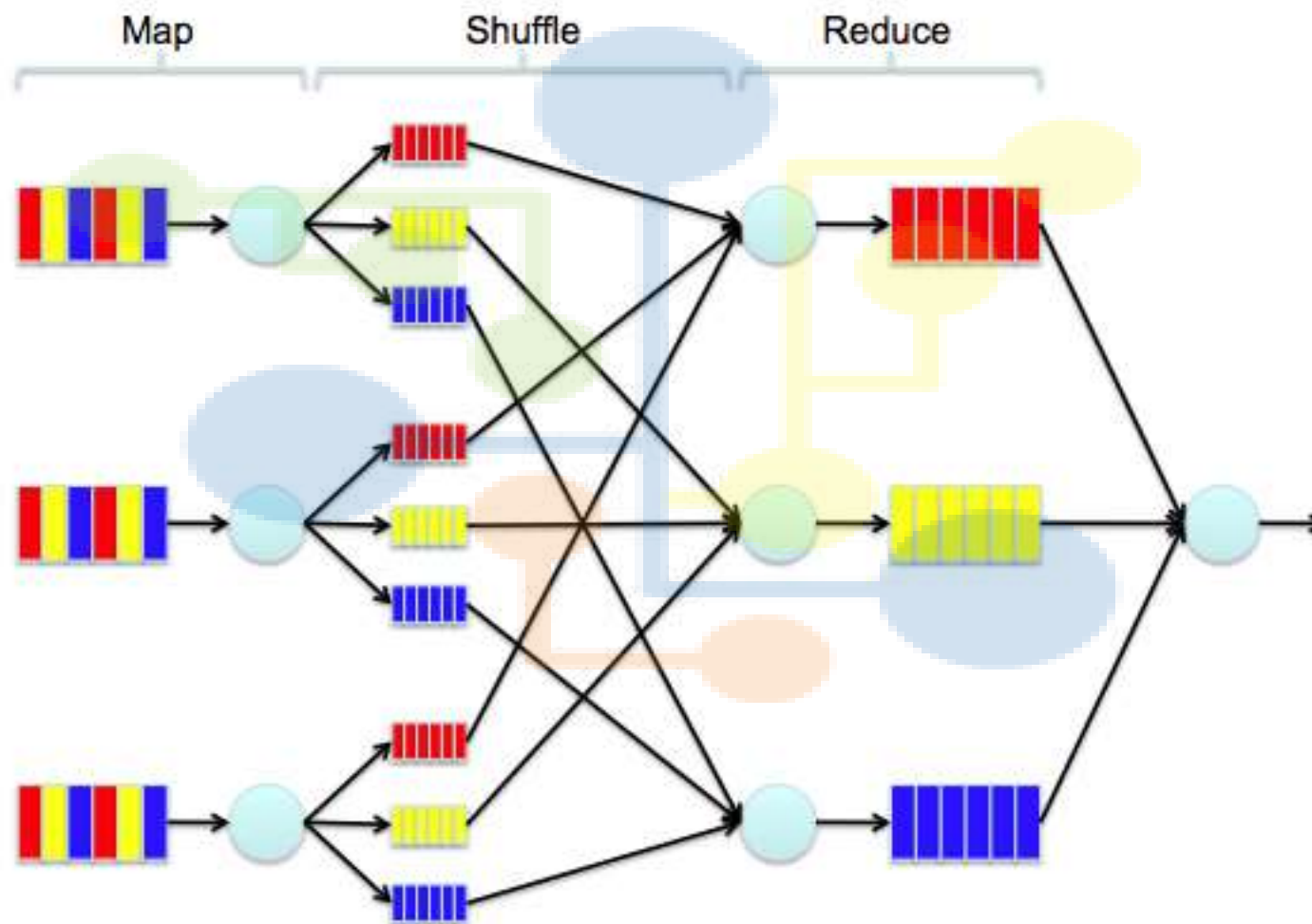
O Modelo de Programação MapReduce

Como exatamente funciona o modelo MapReduce?



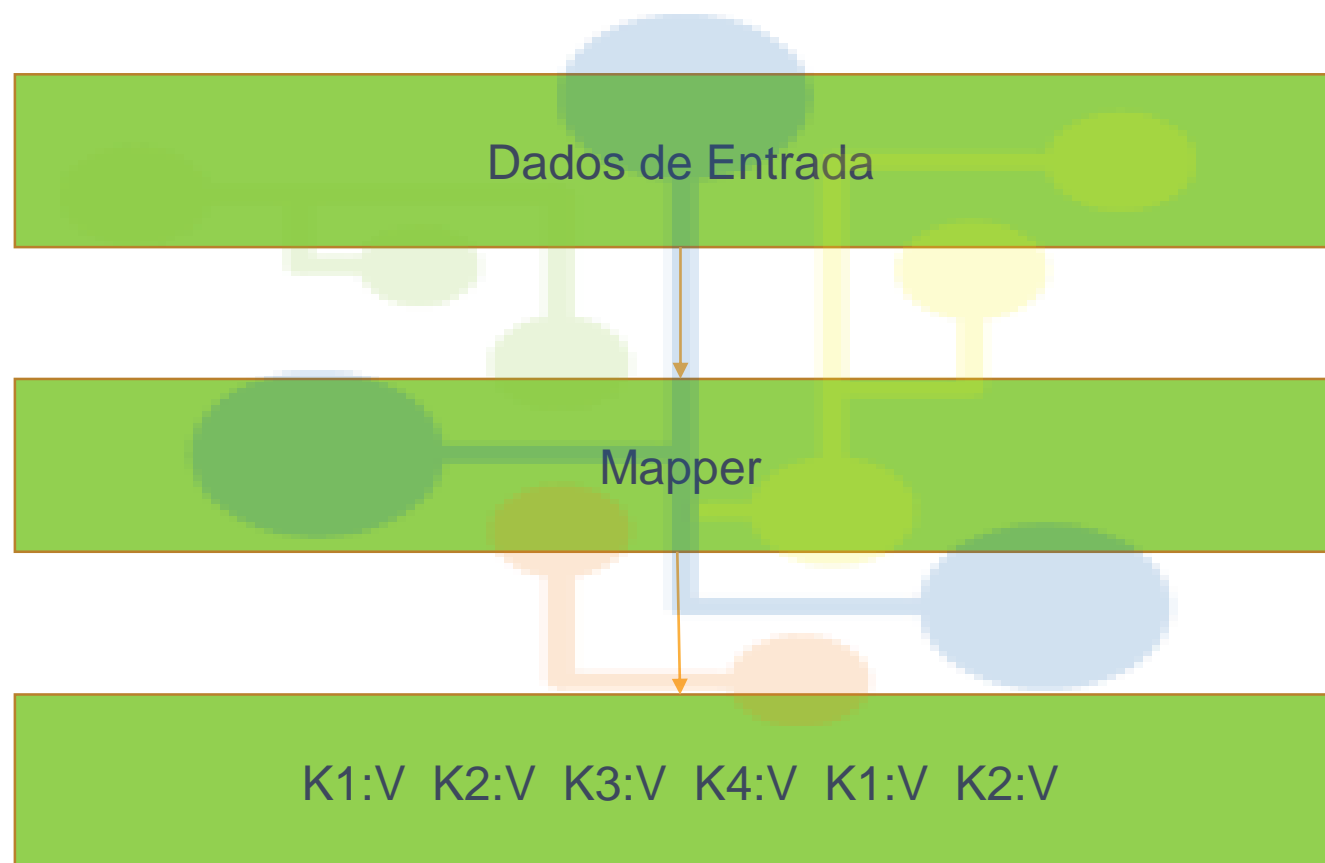


O Modelo de Programação MapReduce





O Modelo de Programação MapReduce





O Modelo de Programação MapReduce

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

Você, Cientista de Dados!



O Modelo de Programação MapReduce

Feito o mapeamento, o Shuffle agrupa todos os pares de chave/valor e entrega para a etapa de redução.

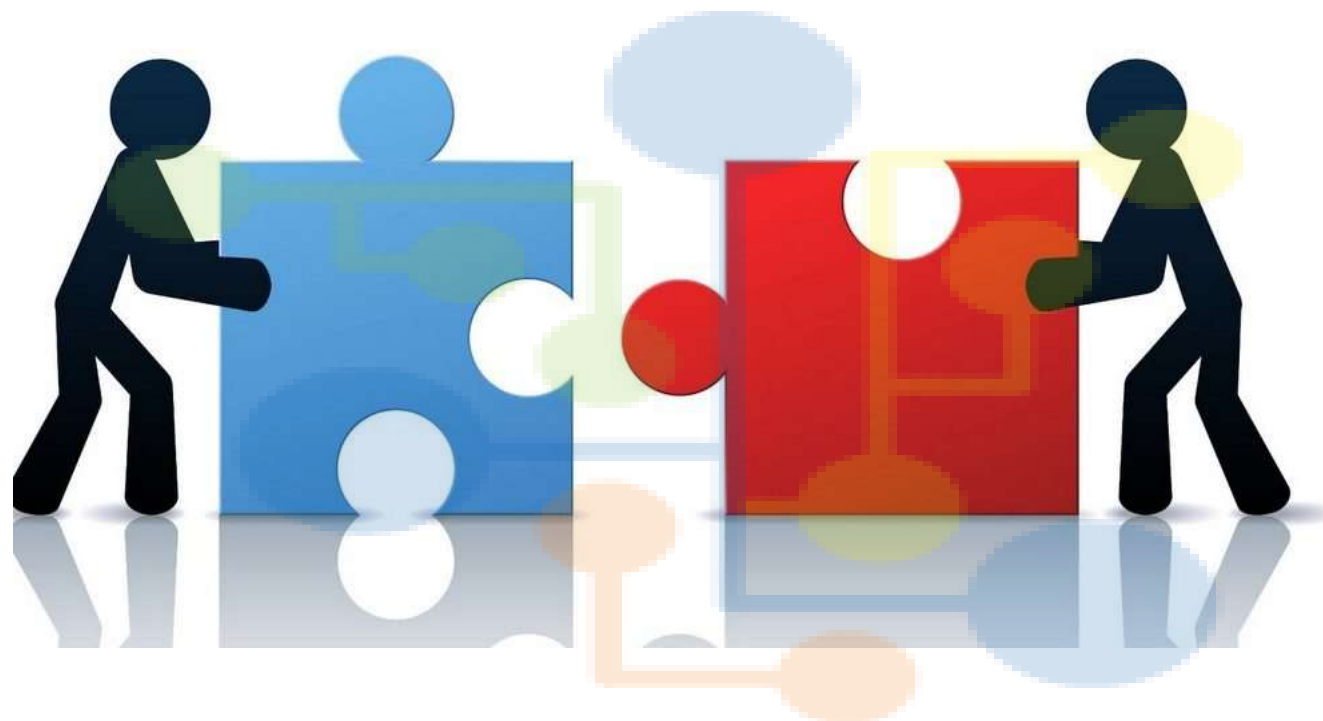
A redução, por exemplo, pode retornar as chaves e o total de suas ocorrências, reduzindo assim os dados à informação que você precisa:

A diagram illustrating a MapReduce job. It shows several nodes (blue, green, yellow, orange) connected by lines, representing the flow of data. A green box at the bottom represents a reducer's output.

K1:3
K2:8
K3:5



O Modelo de Programação MapReduce



Exemplo



O Modelo de Programação MapReduce

Dataset MovieLens (u.data)

Quantos filmes cada pessoa assistiu?

| 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 |

Mapper

Key : Value

userID : movieID

241 : 198

197 : 302

197 : 378

186 : 153

165 : 349

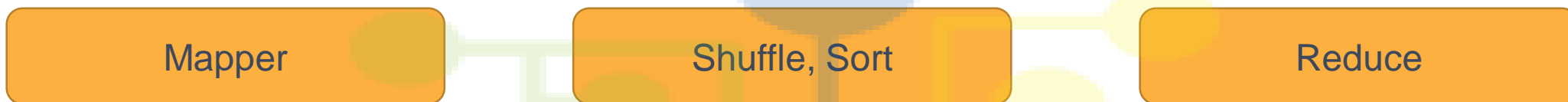
187 : 472

187 : 267



O Modelo de Programação MapReduce

Quantos filmes cada pessoa assistiu?



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

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

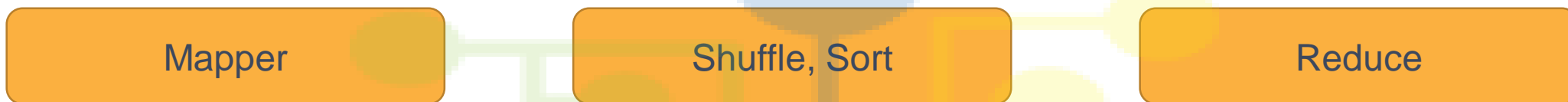
O Modelo de Programação MapReduce





O Modelo de Programação MapReduce

Quantos filmes cada pessoa assistiu?



Key : Value

userID : movieID

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



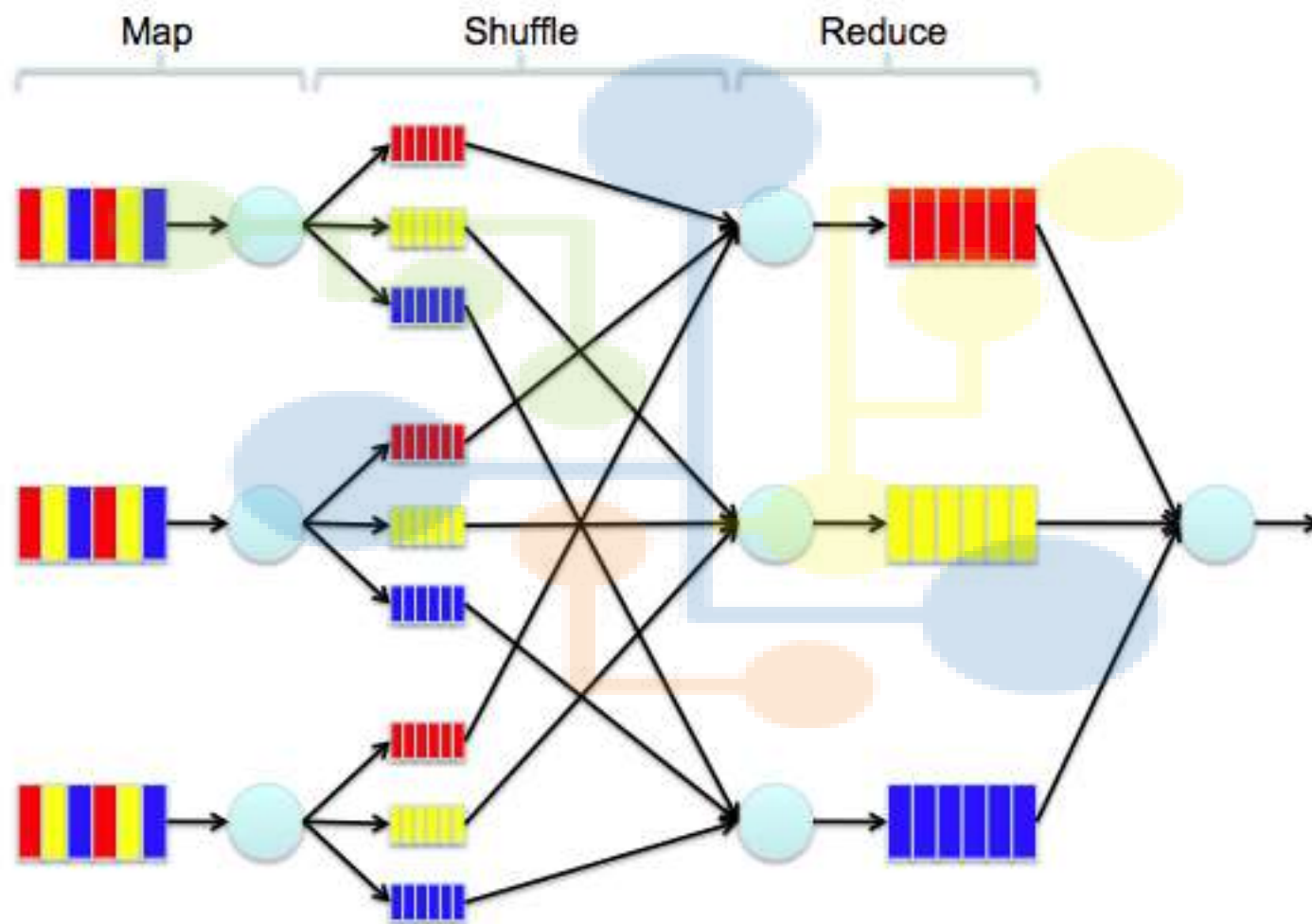
241 : 198
197 : 302, 378
186 : 153
165 : 349
187 : 472, 267



241 : 1
197 : 2
186 : 1
165 : 1
187 : 2

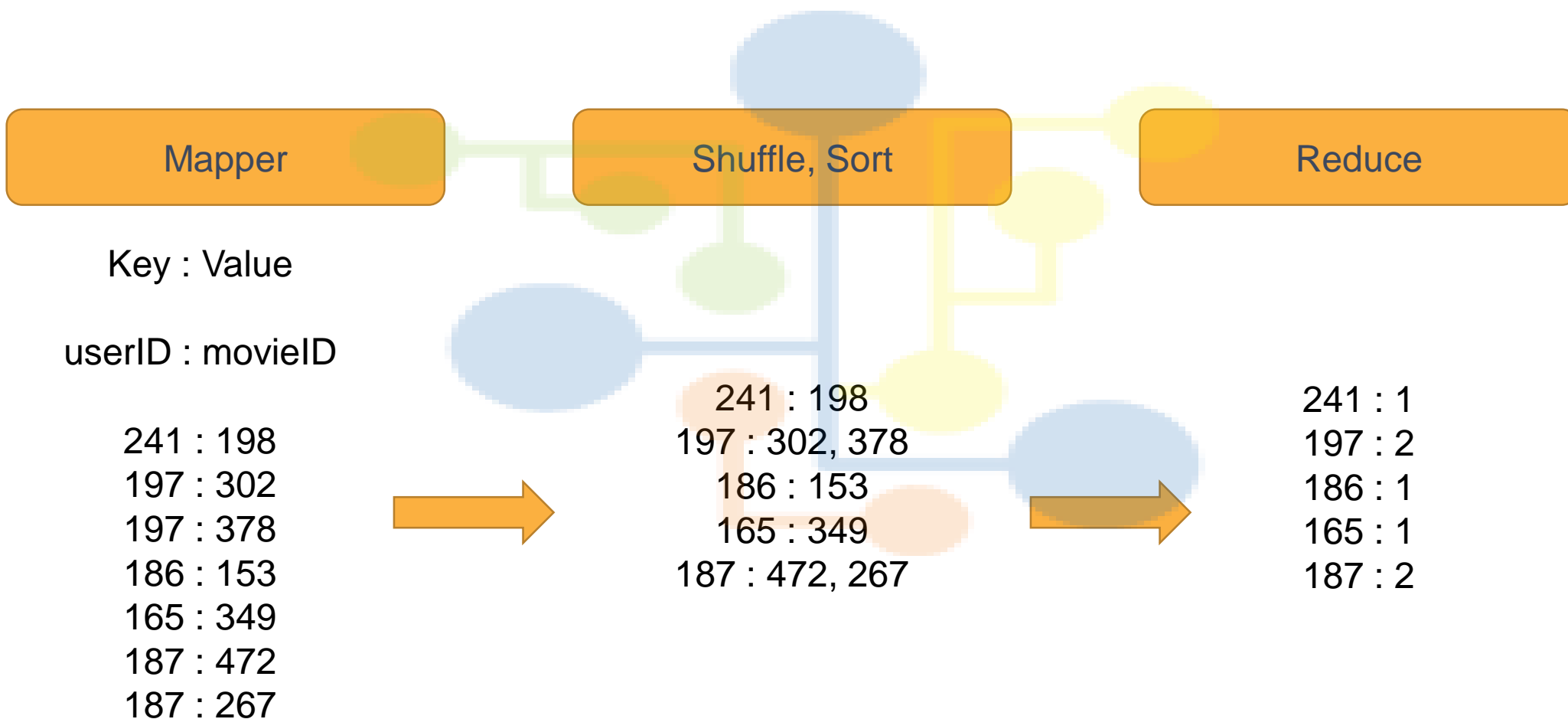


O Modelo de Programação MapReduce





O Modelo de Programação MapReduce

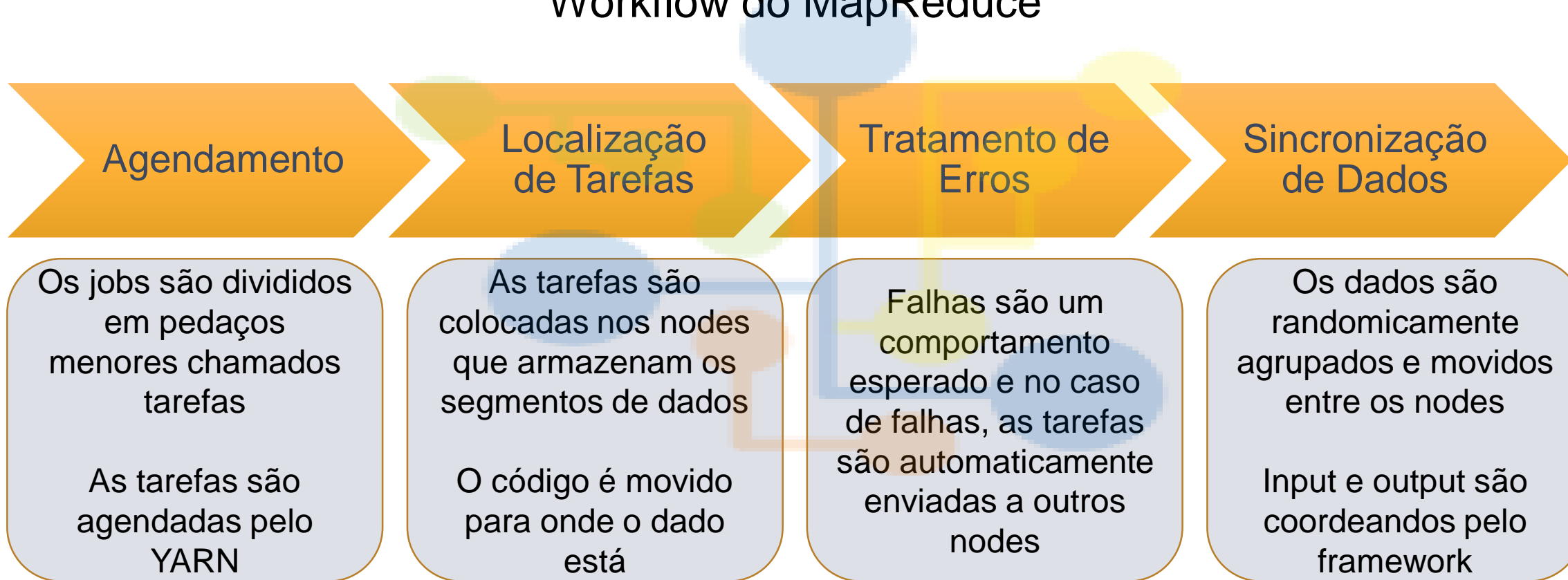




Como o MapReduce Utiliza a Computação Distribuída



Workflow do MapReduce



MapReduce



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

{Muito Satisfeito, Satisfeito, Pouco Satisfeito, Insatisfeito}



MapReduce



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

Input

Split

Mapping

Shuffle

Reduce

Output

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
Satisfeito,
Satisfeito,
Pouco
Satisfeito,
Insatisfeito

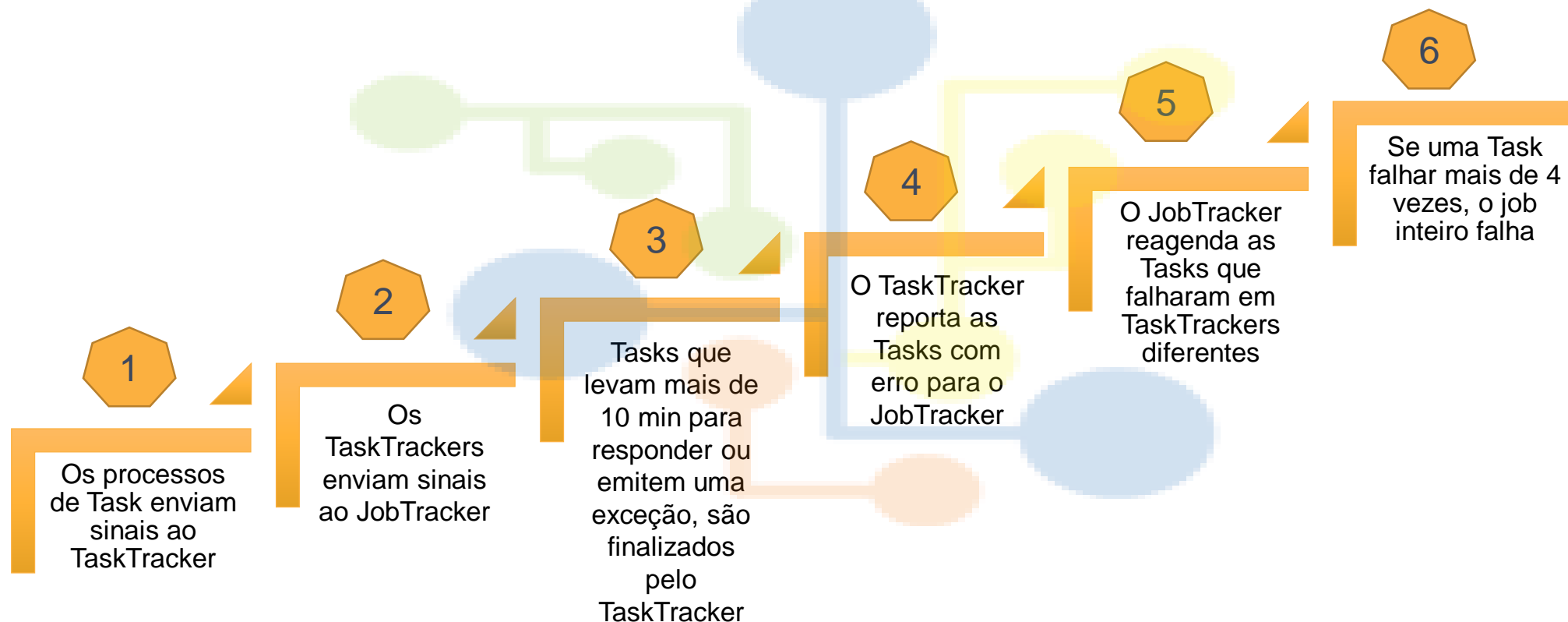
Muito, 1
Satisfeito, 3
Pouco, 1
Insatisfeito, 1

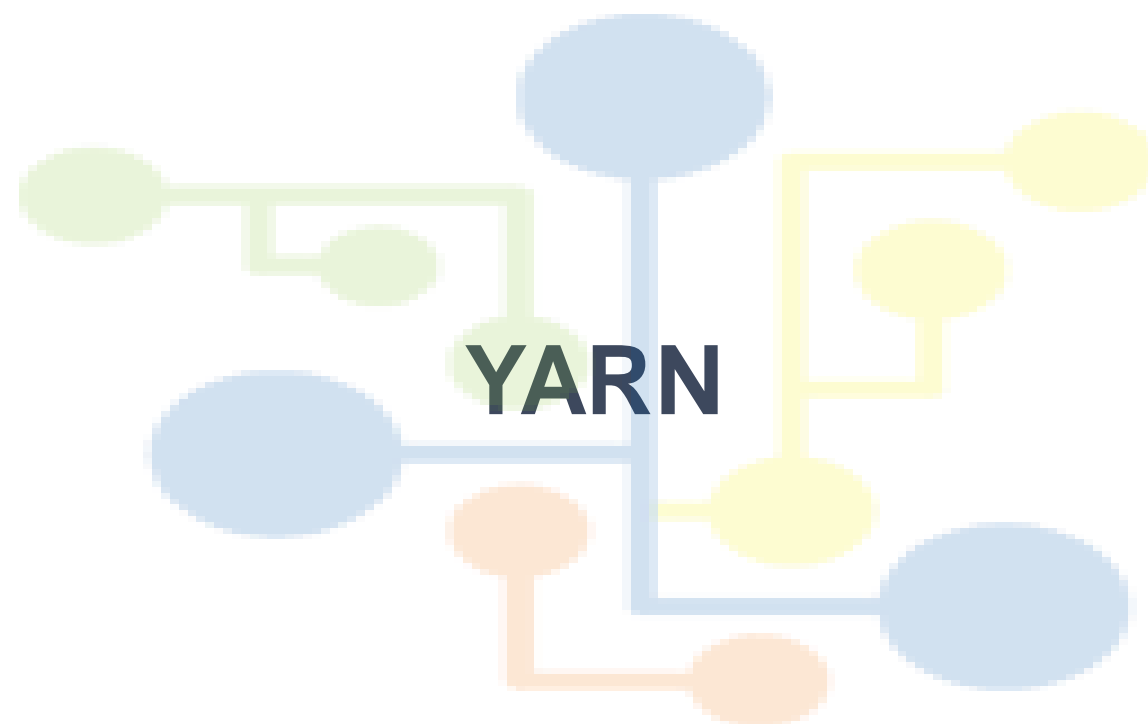
Características do MapReduce

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

Processo de Recuperação a Falhas do MapReduce





Apache YARN - “Yet Another Resource Negotiator” é a camada de gerenciamento de recursos do Hadoop.



O YARN foi introduzido no Hadoop 2.x e permite diferentes mecanismos de processamento de dados, como processamento de grafos, processamento interativo, processamento de fluxo e processamento em lote para executar e processar dados armazenados no HDFS.



Além do gerenciamento de recursos, o Yarn também é usado para agendamento de tarefas (jobs). O YARN amplia o poder do Hadoop para outras tecnologias, para que possam aproveitar as vantagens do HDFS (sistema de armazenamento mais confiável e popular do planeta) e do cluster de baixo custo.



O Apache YARN também é considerado como o sistema operacional de dados do Hadoop. A arquitetura do YARN fornece uma plataforma de processamento de dados de uso geral que não se limita apenas ao MapReduce.





Obrigado
