



# Engenharia de Dados com Hadoop e Spark

---



# Bem-vindo(a)





Data Science Academy

**A Data Science Academy (DSA) é um portal de ensino online especializado em Big Data, Machine Learning, Inteligência Artificial, Desenvolvimento de Chatbots, Blockchain e tecnologias relacionadas.**

**Nosso objetivo é fornecer aos alunos conteúdo de alto nível por meio do uso de computador, tablet ou smartphone, em qualquer lugar, a qualquer hora, 100% online e 100% em português.**

## No Brasil e no Mundo





# Engenharia de Dados com Hadoop e Spark



# Engenharia de Dados com Hadoop e Spark

Cluster  
Hadoop

Capítulos  
2, 3 e 4

Armazenamento  
de Dados  
Capítulos  
5, 6 e 7

Machine  
Learning

Capítulo  
8

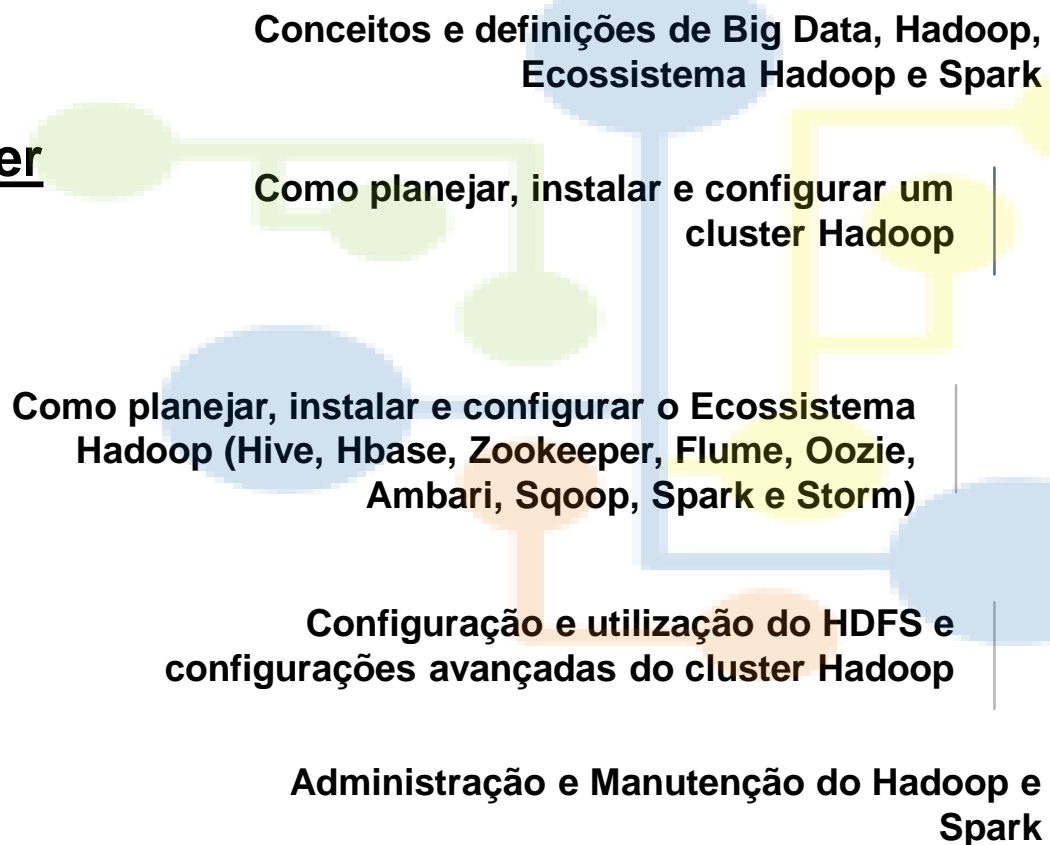
Hadoop e  
Spark

Capítulo  
9



# Engenharia de Dados com Hadoop e Spark

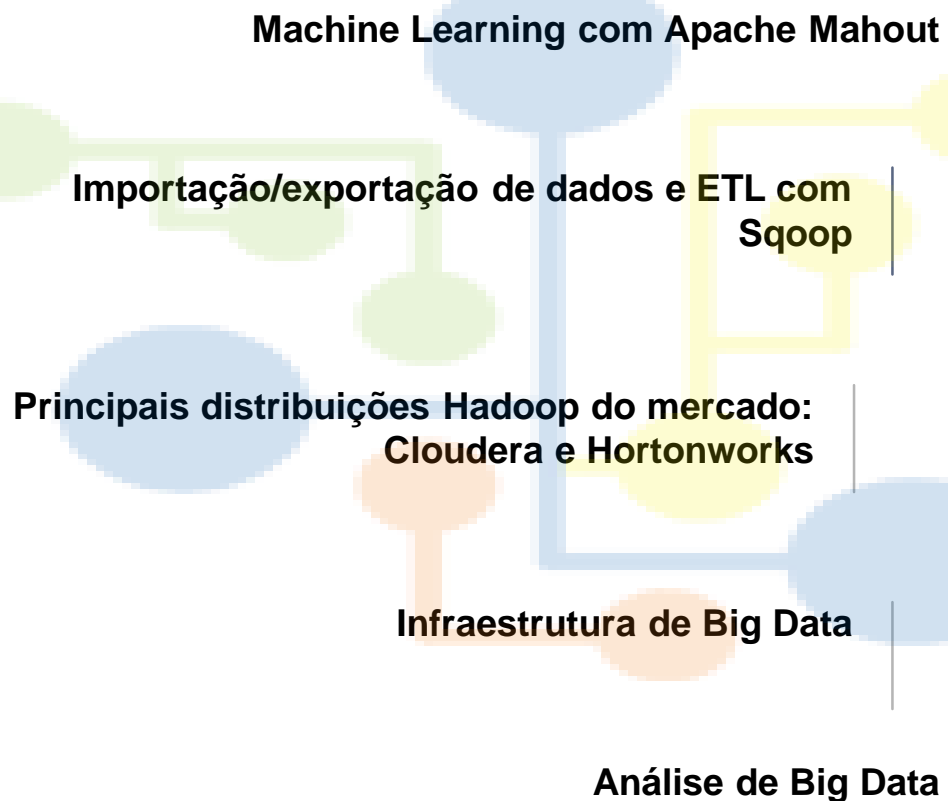
## O que você vai aprender neste curso?





# Engenharia de Dados com Hadoop e Spark

O que você vai aprender neste curso?







# Engenharia de Dados com Hadoop e Spark

E quais são os pré-requisitos?

**Curso Big Data  
Fundamentos 2.0**



**Conhecimentos  
básicos de  
sistema  
operacional Linux  
(desejável)**



**Conhecimentos  
básicos de  
linguagem de  
programação  
(desejável)**



**Muita vontade de  
aprender e entrar  
no mundo do Big  
Data  
(mandatório)**



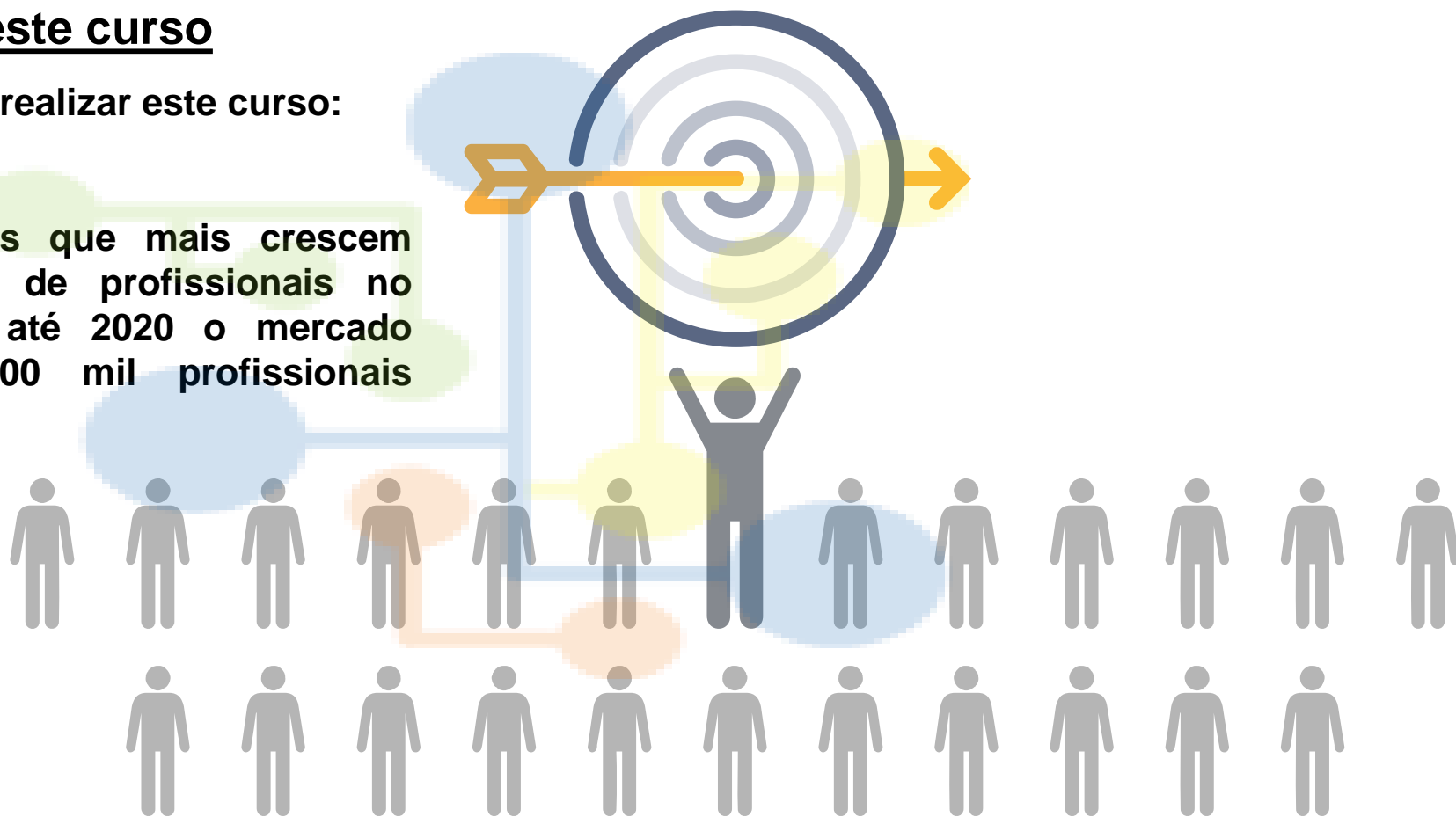


# Engenharia de Dados com Hadoop e Spark

## Benefícios em realizar este curso

São muitos os benefícios em realizar este curso:

Big Data é uma das áreas que mais crescem atualmente. Há um déficit de profissionais no mercado e estima-se que até 2020 o mercado precisará de mais de 200 mil profissionais habilitados em Big Data.



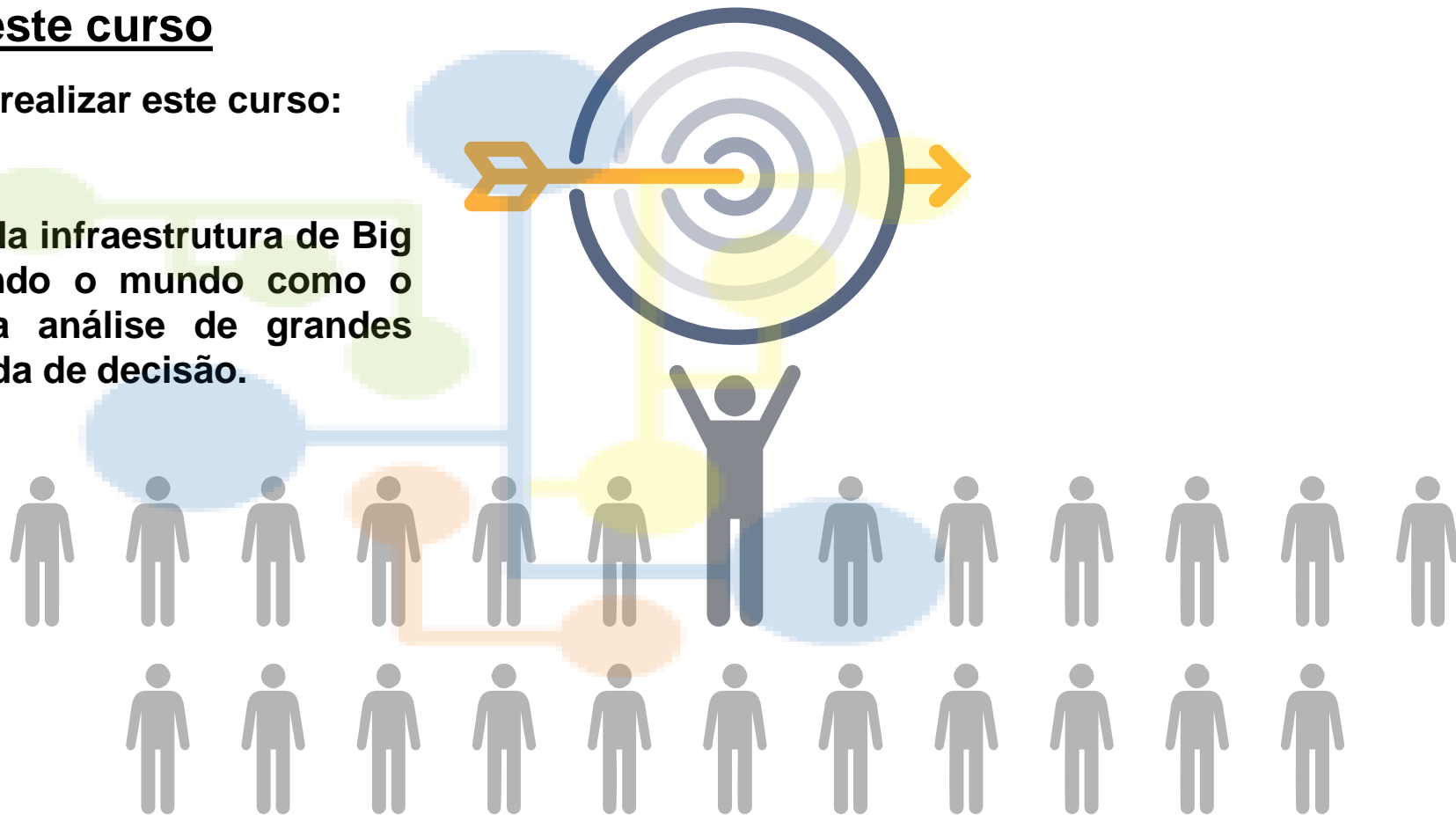


# Engenharia de Dados com Hadoop e Spark

## Benefícios em realizar este curso

São muitos os benefícios em realizar este curso:

Hadoop é a tecnologia base da infraestrutura de Big Data, que está revolucionando o mundo como o conhecemos. Ele permite a análise de grandes volumes de dados para tomada de decisão.



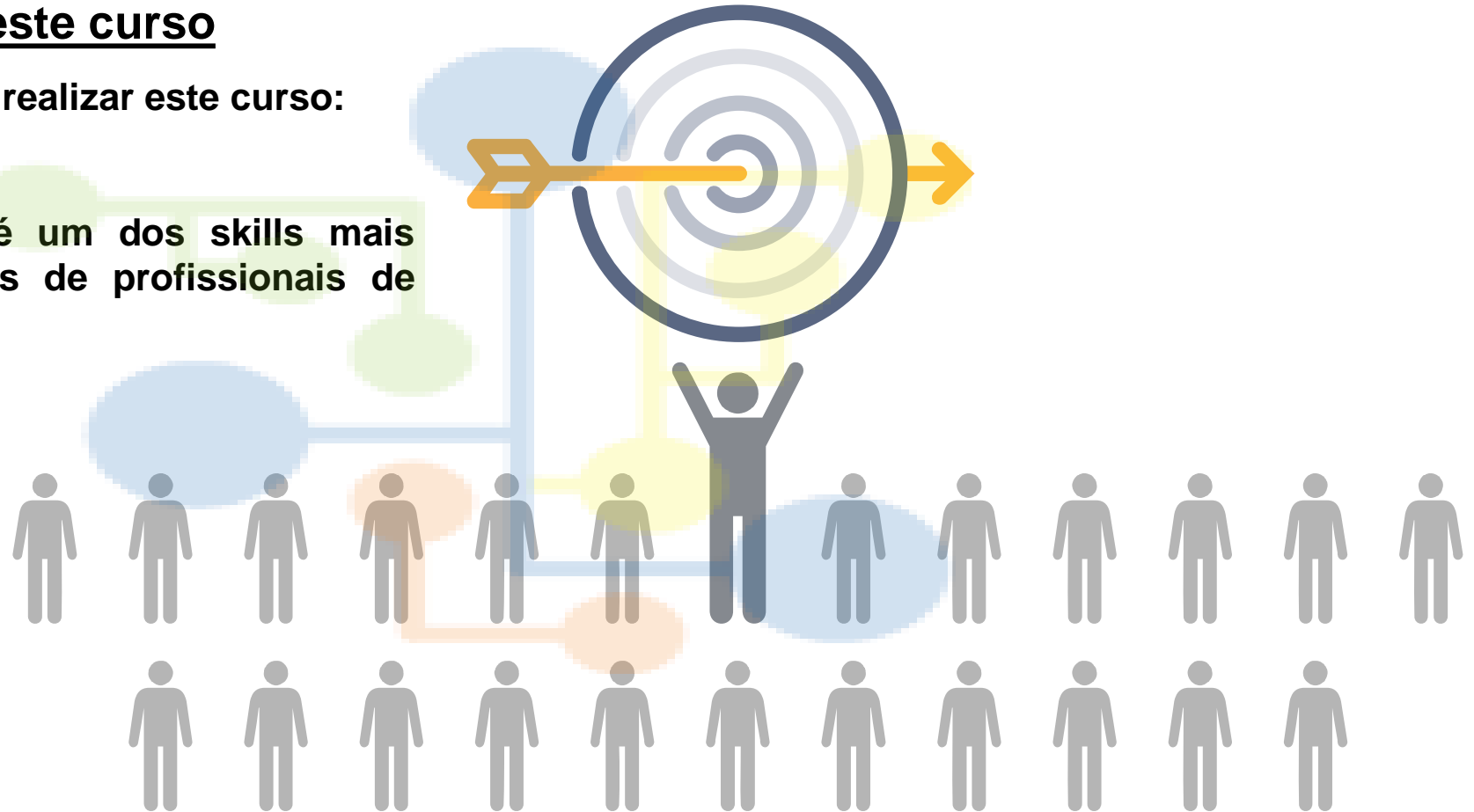


# Engenharia de Dados com Hadoop e Spark

## Benefícios em realizar este curso

São muitos os benefícios em realizar este curso:

Conhecimento de Hadoop é um dos skills mais procurados por recrutadores de profissionais de Big Data.



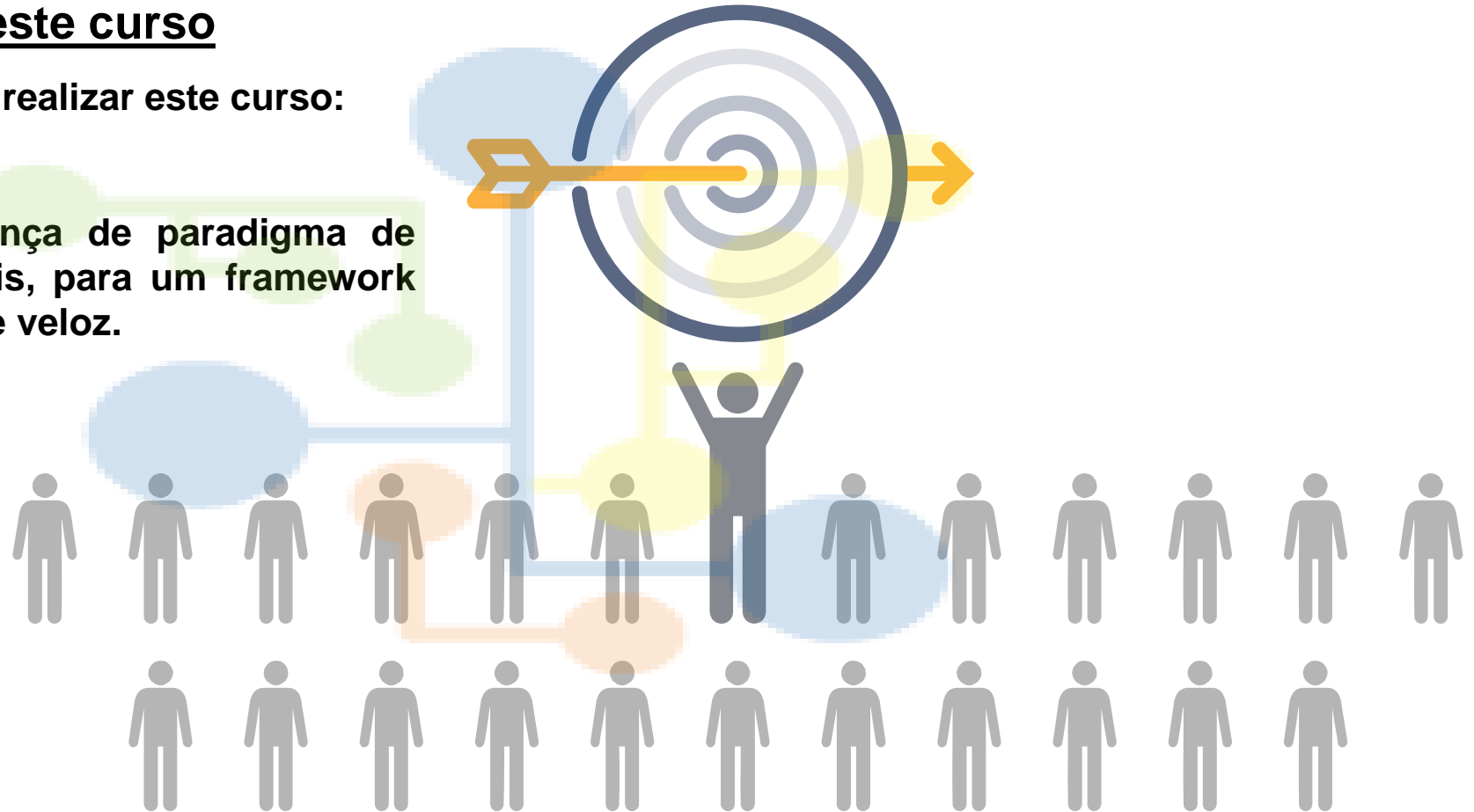


# Engenharia de Dados com Hadoop e Spark

## Benefícios em realizar este curso

São muitos os benefícios em realizar este curso:

O Hadoop permite a mudança de paradigma de bancos de dados tradicionais, para um framework de dados versátil, adaptável e veloz.





# Engenharia de Dados com Hadoop e Spark

## Estrutura do Curso

Para tornar sua experiência de aprendizagem ainda mais completa, você terá quizzes e labs ao longo de todos os capítulos.





# Engenharia de Dados com Hadoop e Spark

## Estrutura do Curso

Você também terá acesso e poderá fazer o download dos e-books com todo o passo-a-passo de cada lab realizado ao longo do curso.





# Engenharia de Dados com Hadoop e Spark

## Estrutura do Curso

Fique tranquilo se você não possui experiência em sistema operacional Linux. Tudo será explicado passo a passo.

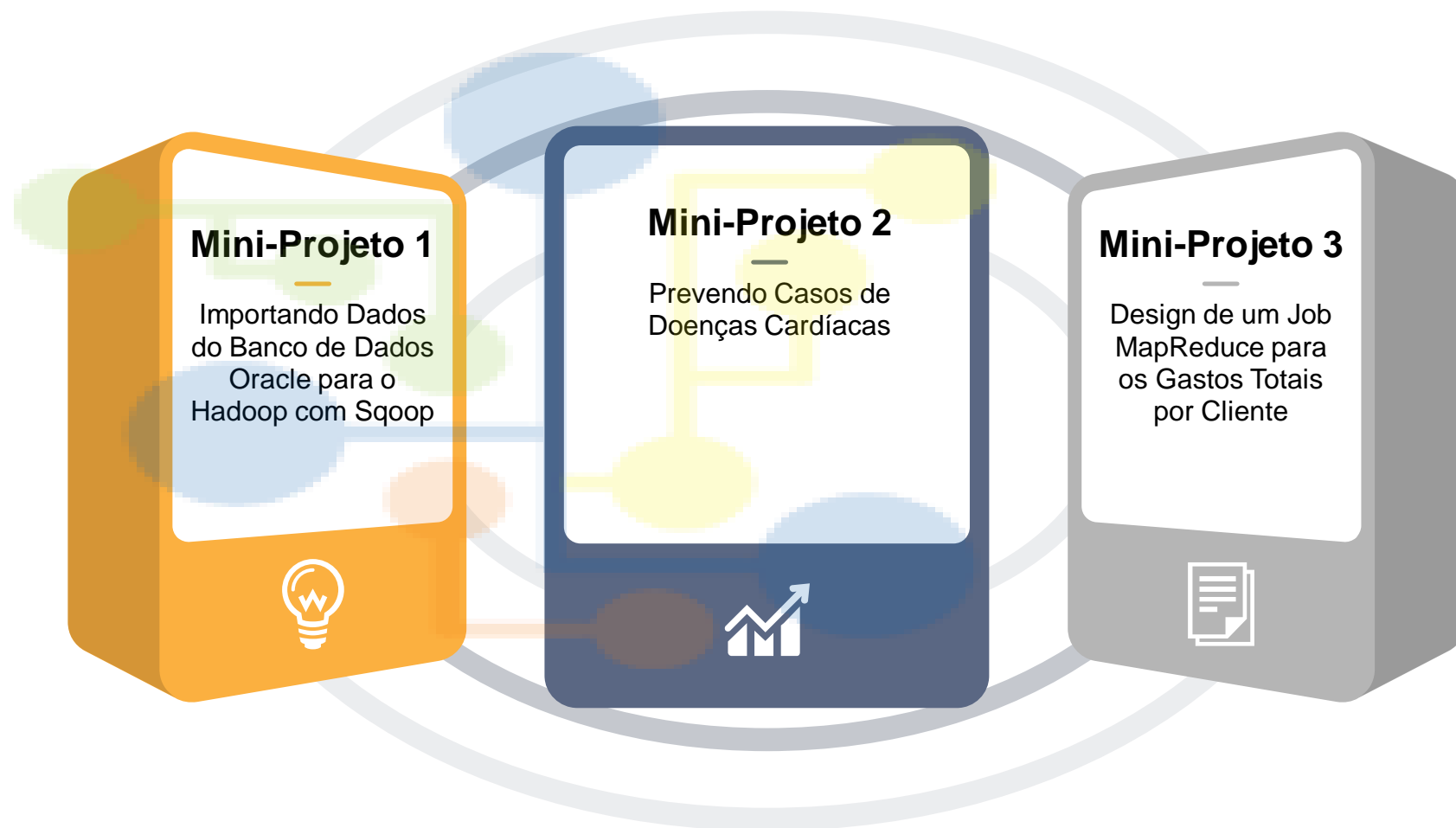






# Engenharia de Dados com Hadoop e Spark

## Mini-Projetos





# Engenharia de Dados com Hadoop e Spark

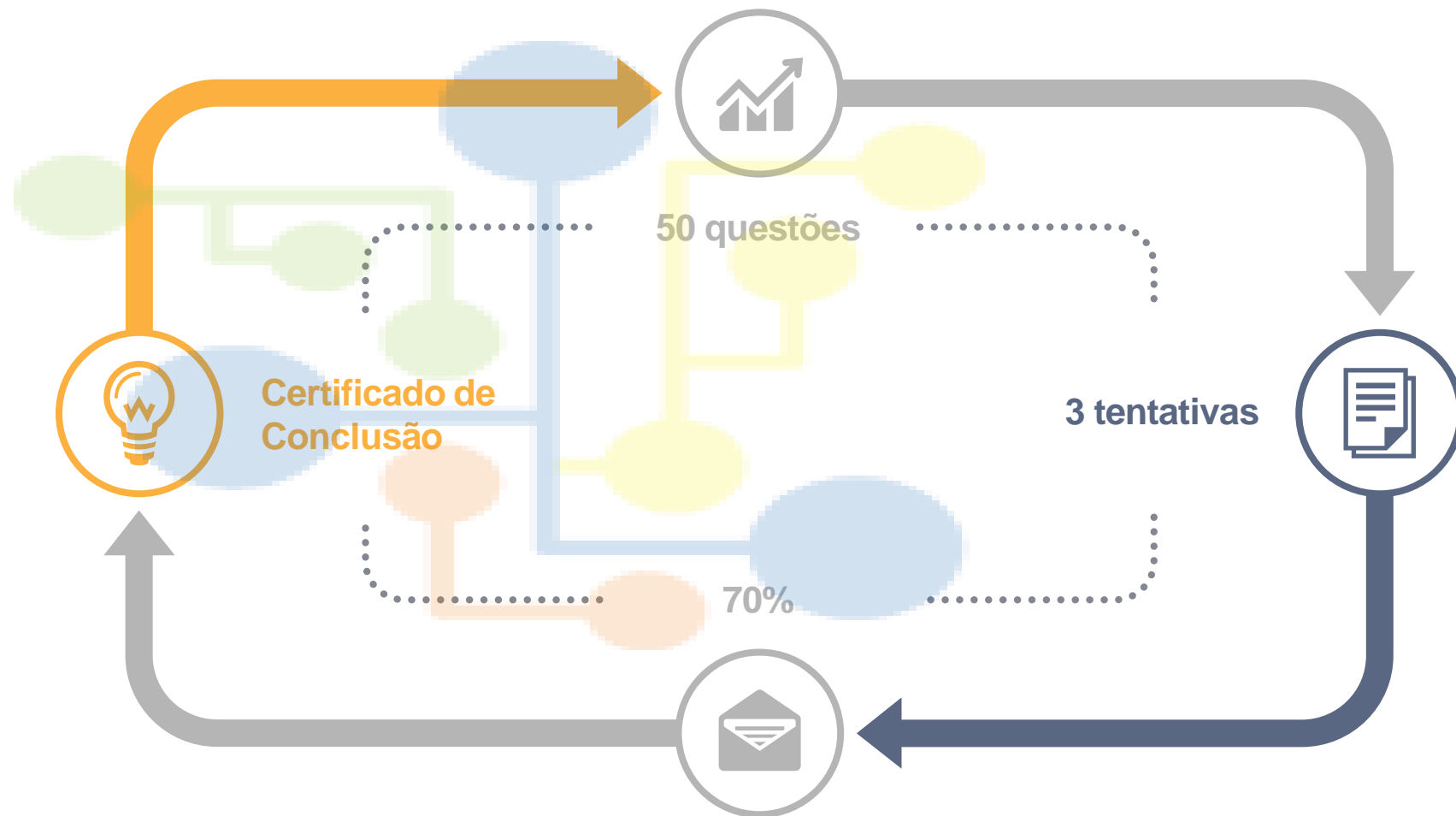
## Projetos com Feedback





# Engenharia de Dados com Hadoop e Spark

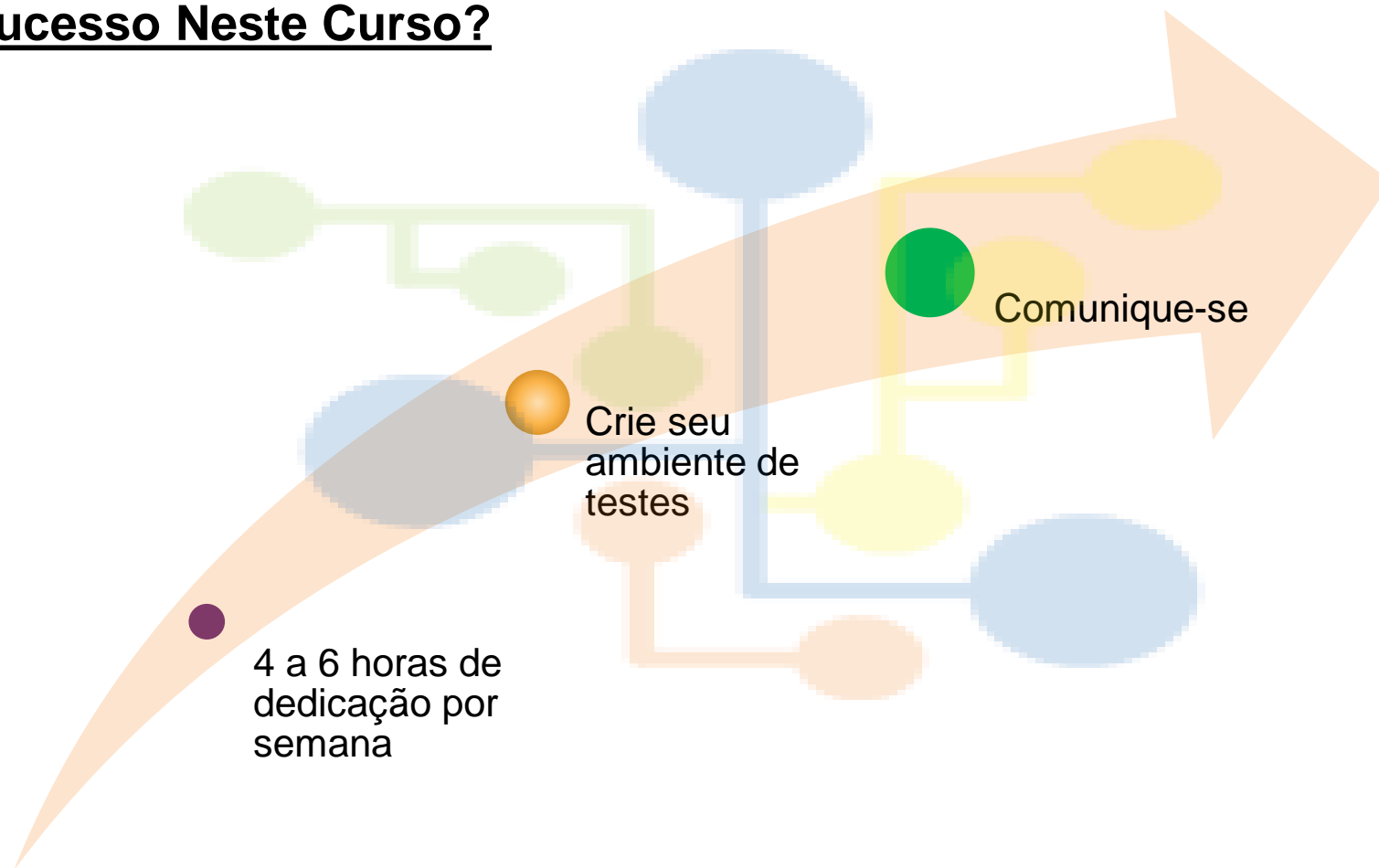
## Avaliação Final





# Engenharia de Dados com Hadoop e Spark

## Como Obter Sucesso Neste Curso?





Data Science  
Academy

Data Science Academy marcelo\_eidi12@hotmail.com 5d5c42d55e4cde68f38b457d

# Engenharia de Dados com Hadoop e Spark





Data Science  
Academy

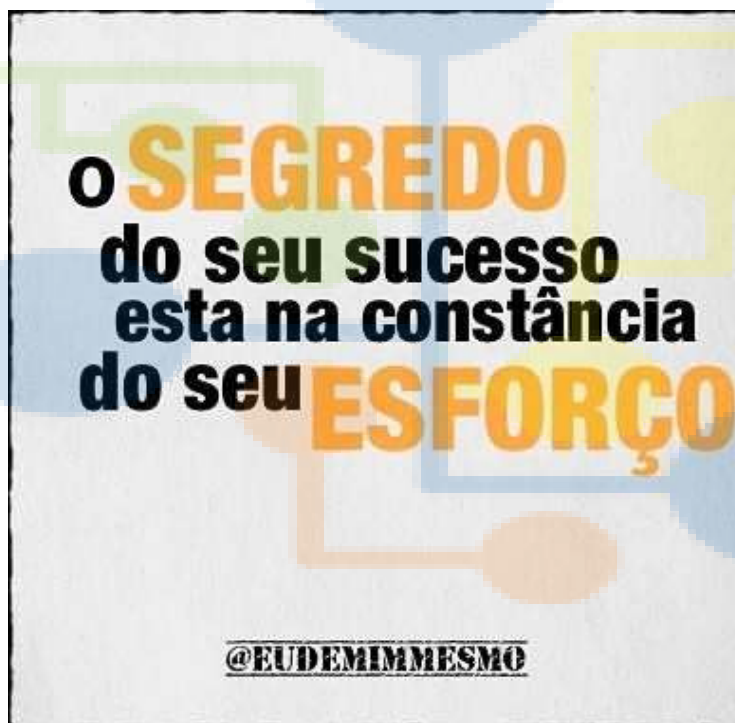
Data Science Academy marcelo\_eidi12@hotmail.com 5d5c42d55e4cde68f38b457d

# Engenharia de Dados com Hadoop e Spark



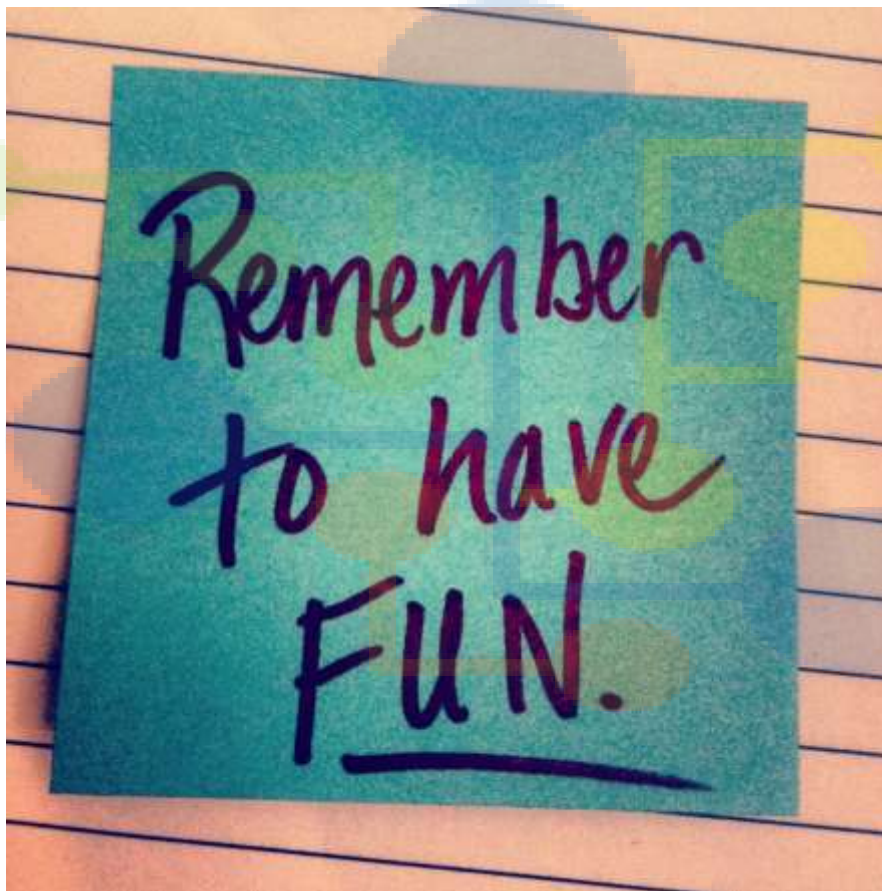


# Engenharia de Dados com Hadoop e Spark





# Engenharia de Dados com Hadoop e Spark





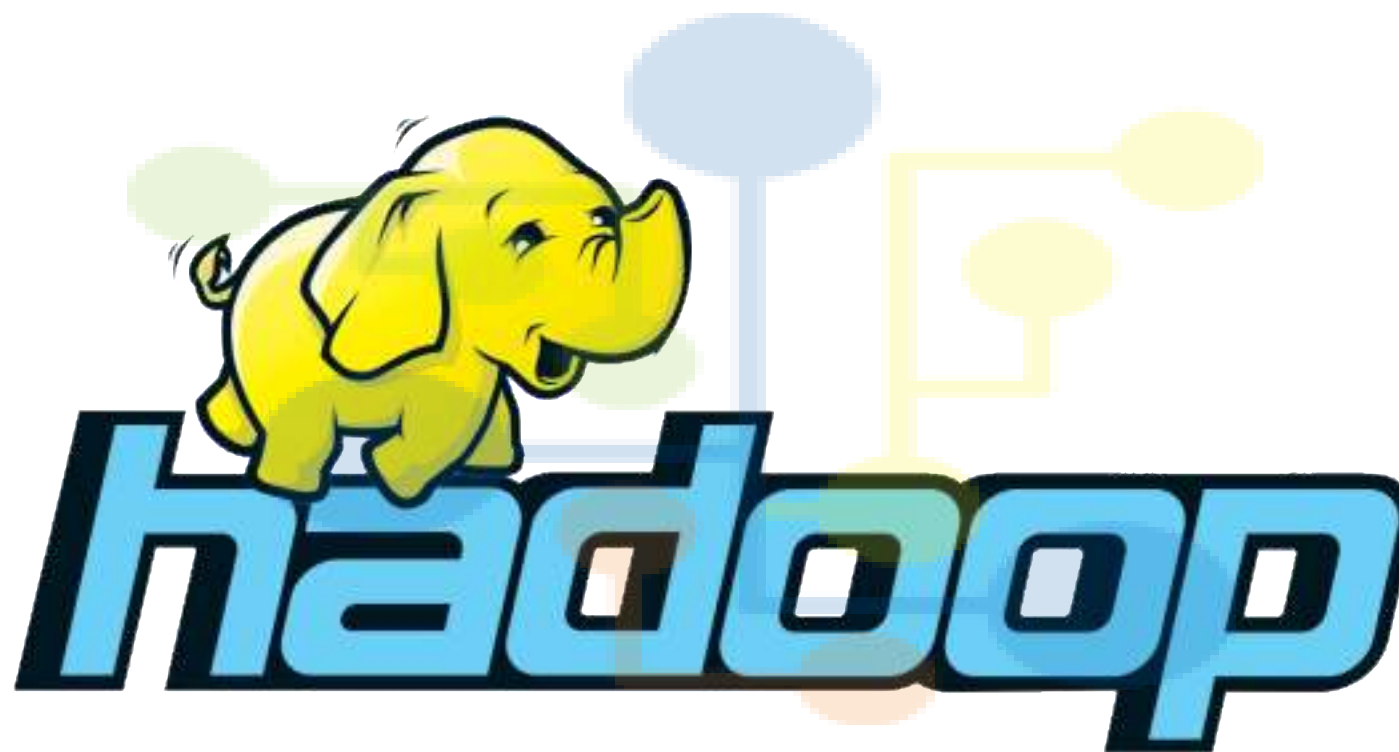


# O que é o Apache Hadoop?





# O que é o Apache Hadoop?





# O que é o Apache Hadoop?

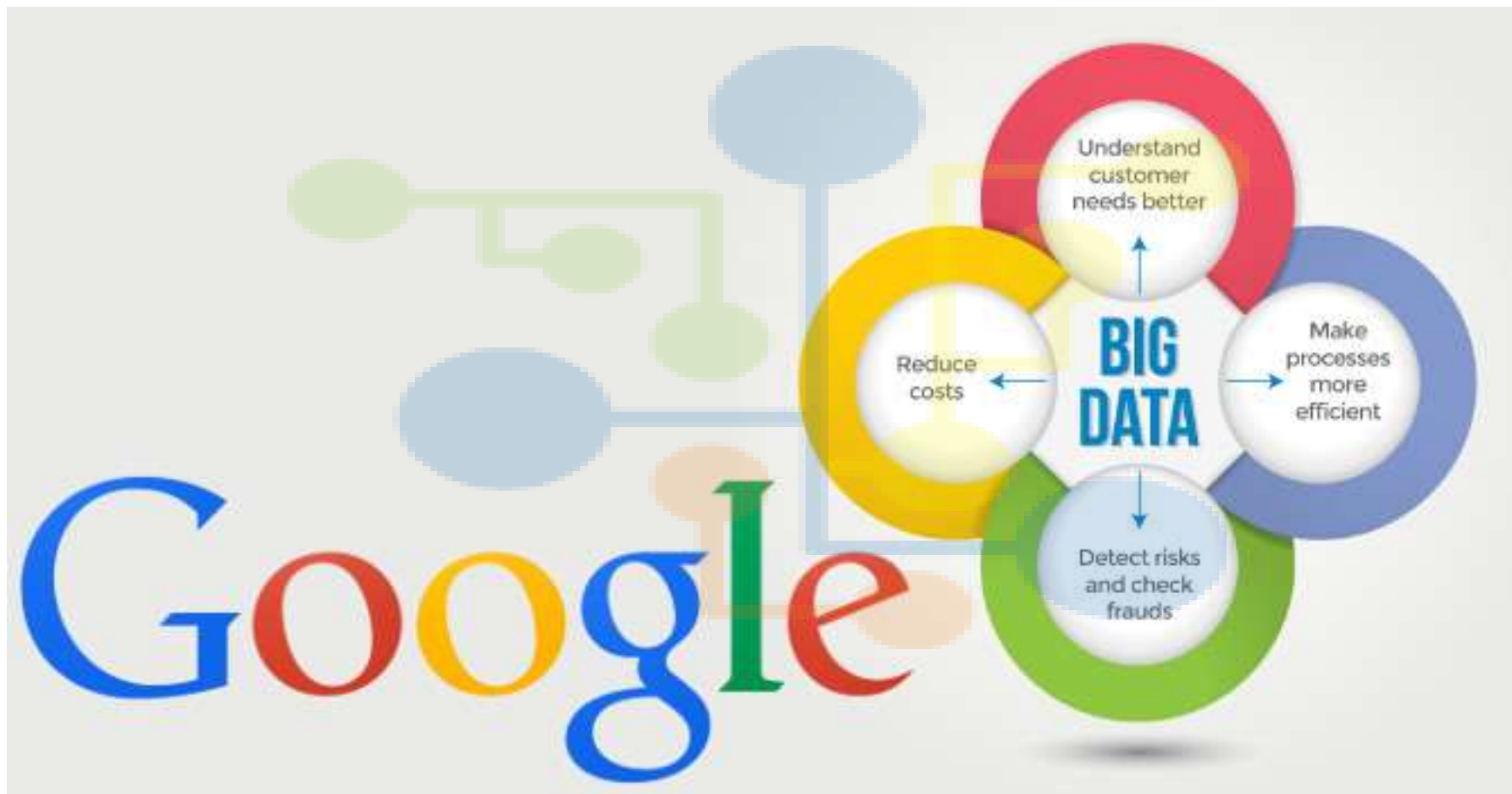
Um dos grandes desafios computacionais da atualidade é armazenar, manipular e analisar, de forma inteligente, a grande quantidade de dados existente.

Sistemas corporativos, sistemas Web, mídias sociais, entre outros, produzem juntos um volume impressionante de dados, alcançando a dimensão de petabytes diários.





# O que é o Apache Hadoop?



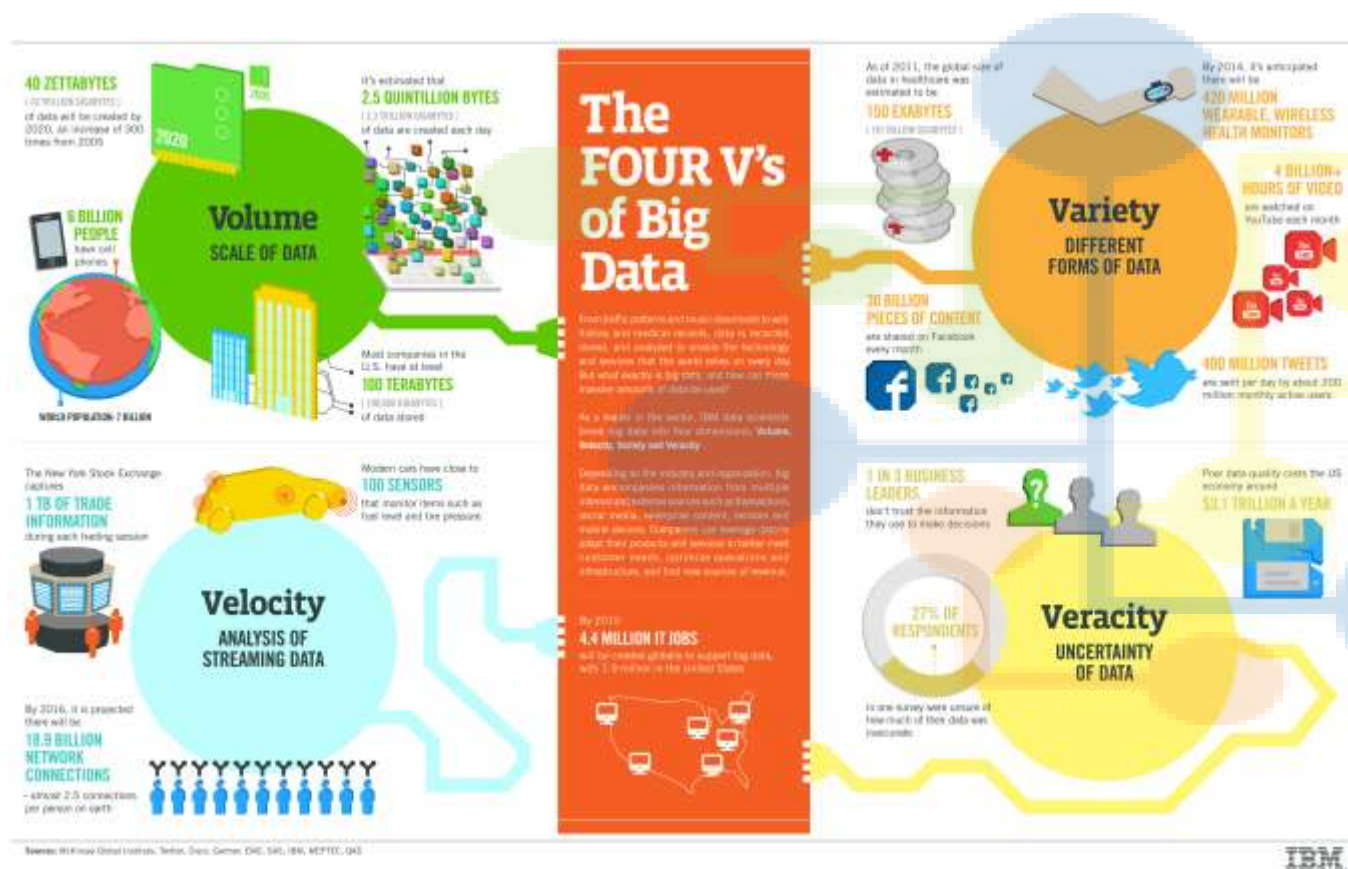


# O que é o Apache Hadoop?





# O que é o Apache Hadoop?



Os 4 V's do Big Data:

- Volume
- Variedade
- Velocidade
- Veracidade





Data Science  
Academy

Data Science Academy marcelo\_eidi12@hotmail.com 5d5c42d55e4cde68f38b457d

# O que é o Apache Hadoop?





# O que é o Apache Hadoop?

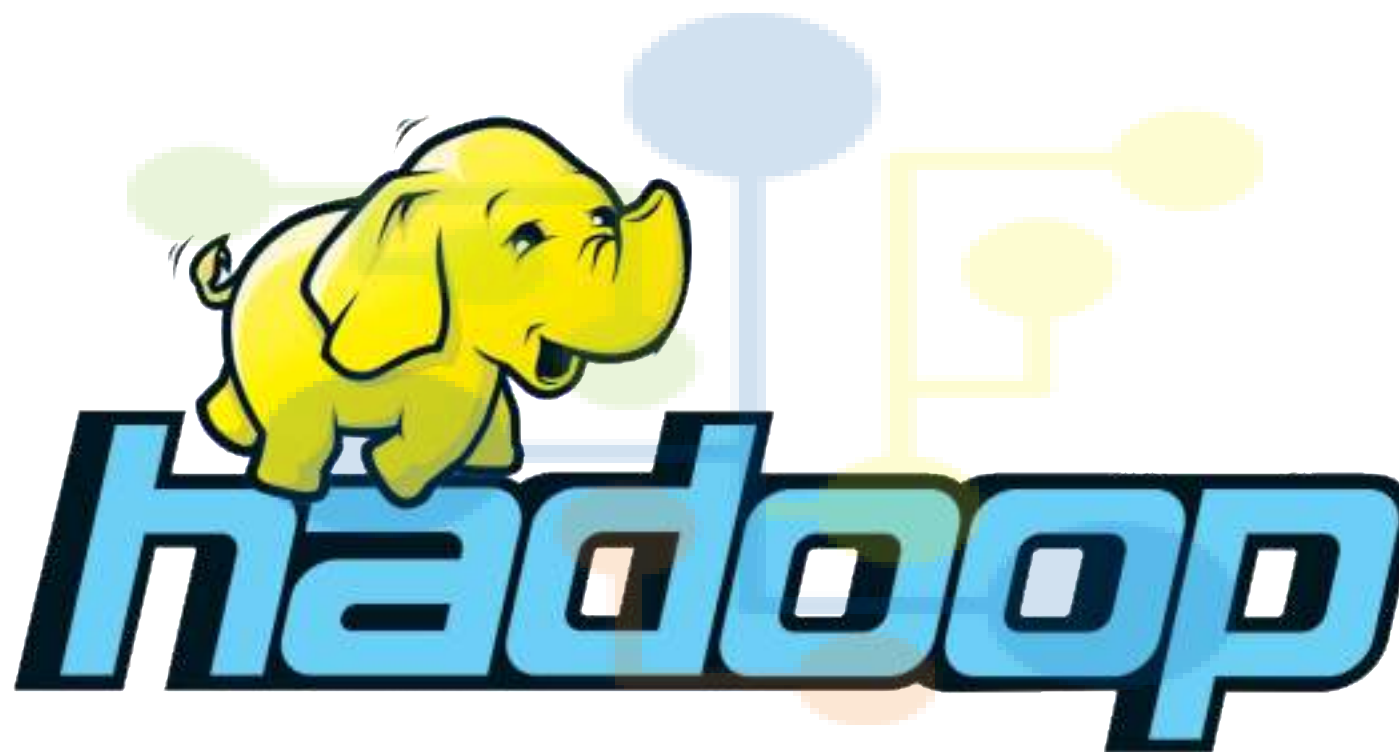
Computação Paralela







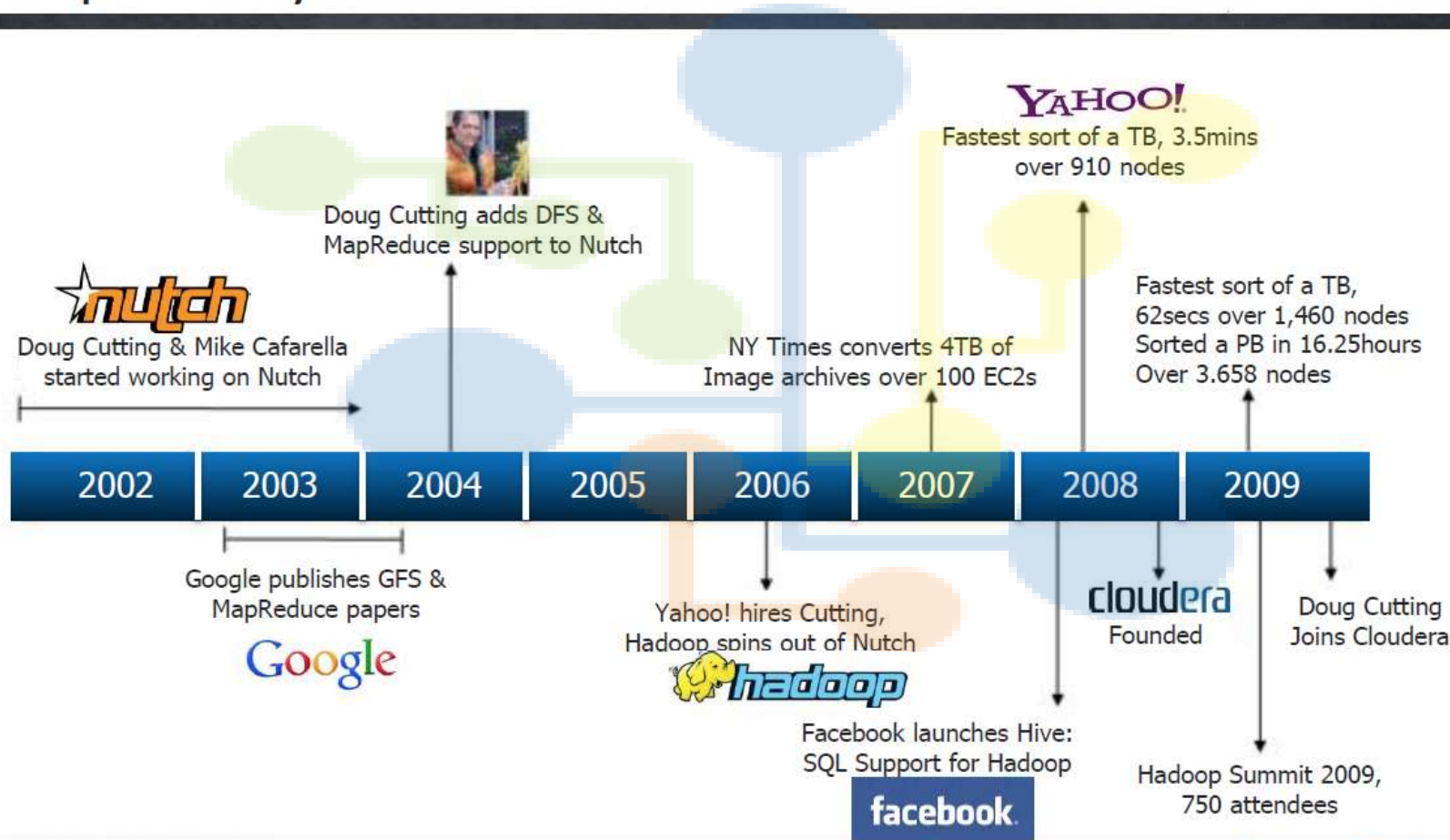
# O que é o Apache Hadoop?





# Uma Breve História do Apache Hadoop

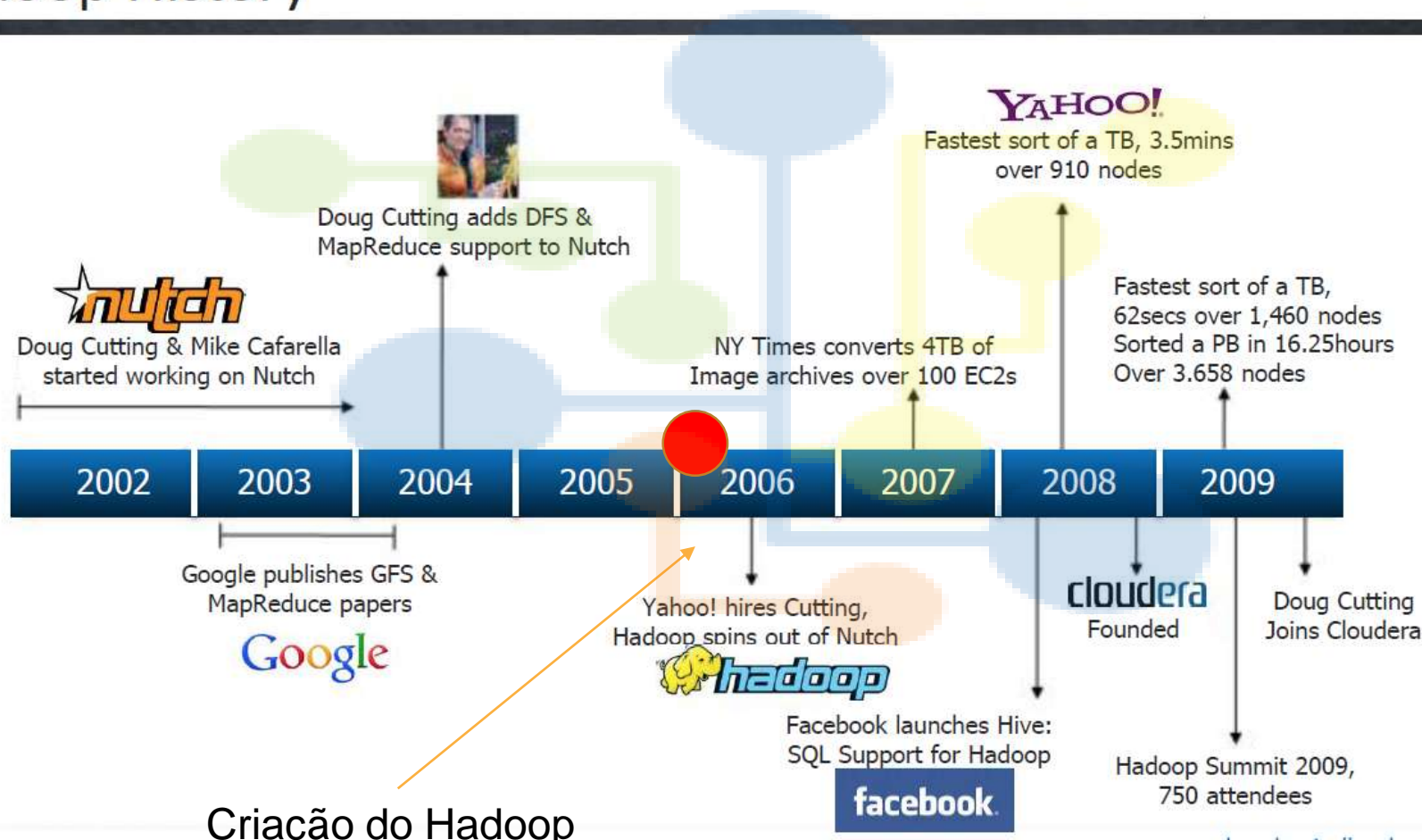
## Hadoop History





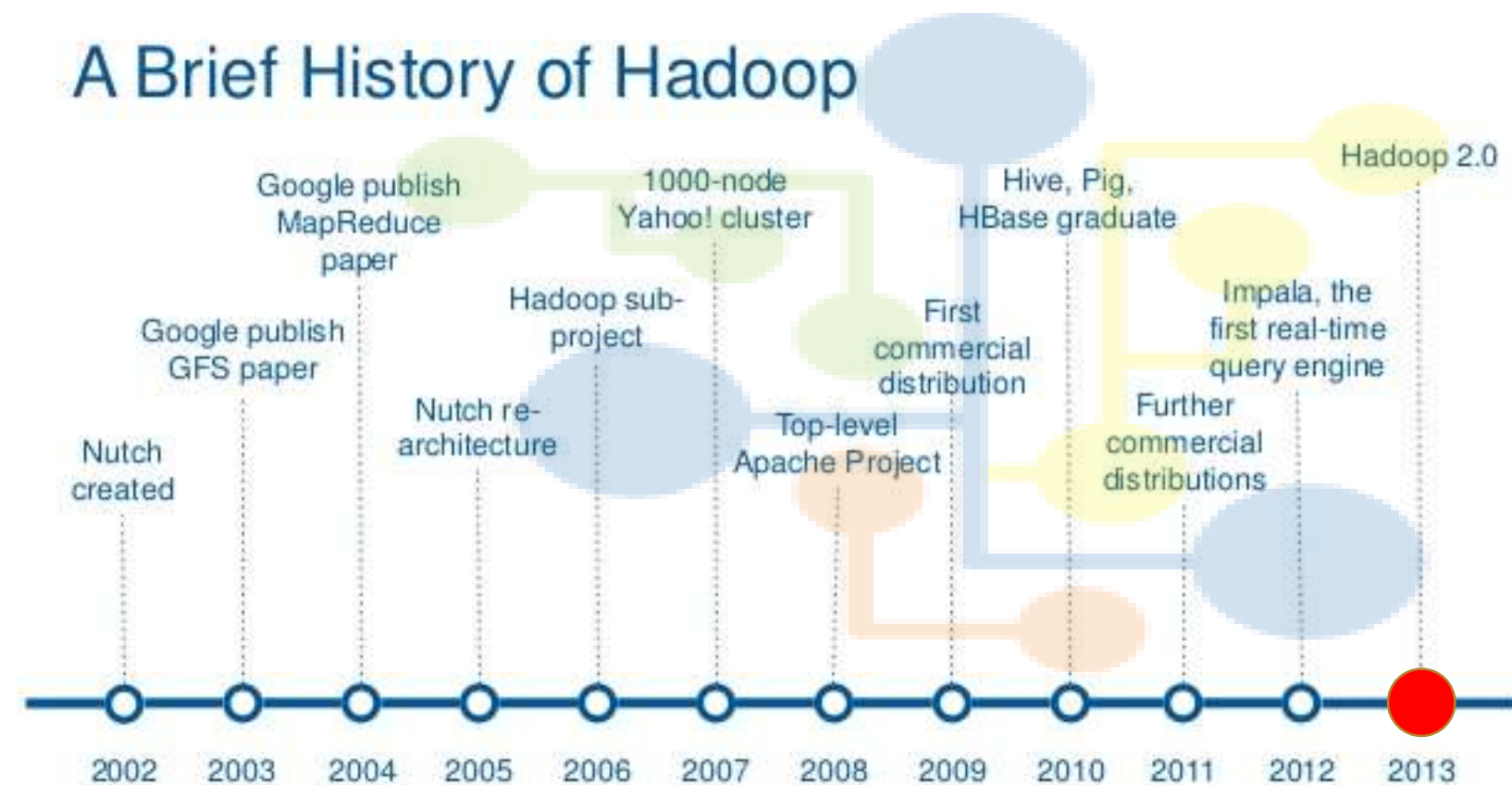
# Uma Breve História do Apache Hadoop

## Hadoop History





# Uma Breve História do Apache Hadoop

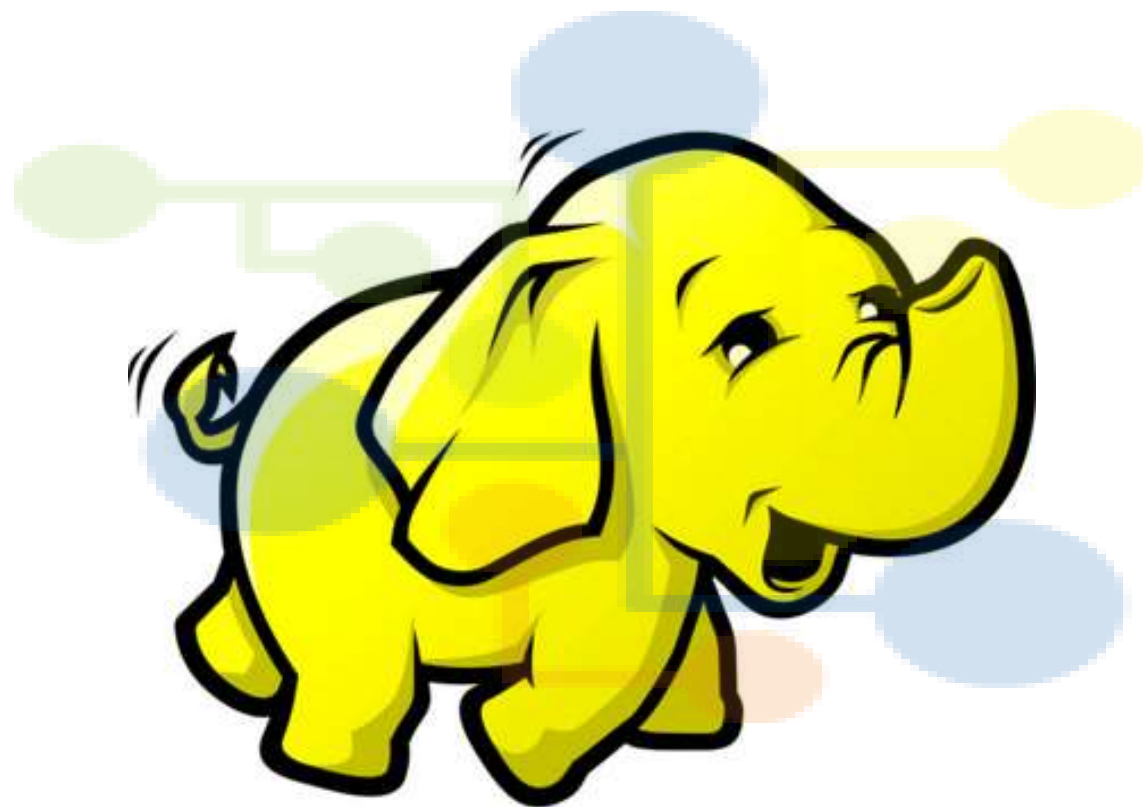


2018  
Versão 3.0

2019  
Versão 3.2.x



# O que é o Hadoop?





# Quais os benefícios para as Empresas ao utilizar o Hadoop?

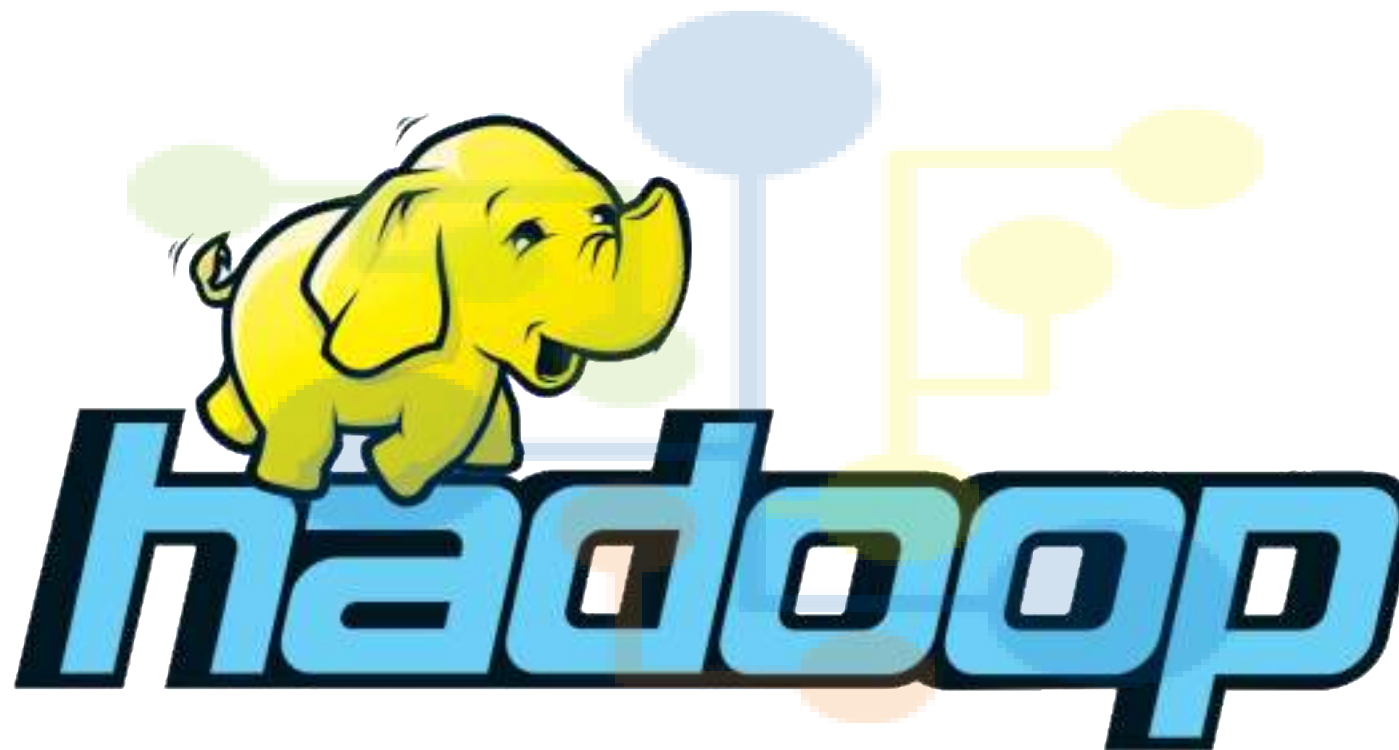




Data Science  
Academy

Data Science Academy marcelo\_eidi12@hotmail.com 5d5c42d55e4cde68f38b457d

# Benefícios do Hadoop



# Benefícios do Hadoop

# Open Source







# Benefícios do Hadoop



Economia



# Benefícios do Hadoop



Escalabilidade



# Benefícios do Hadoop

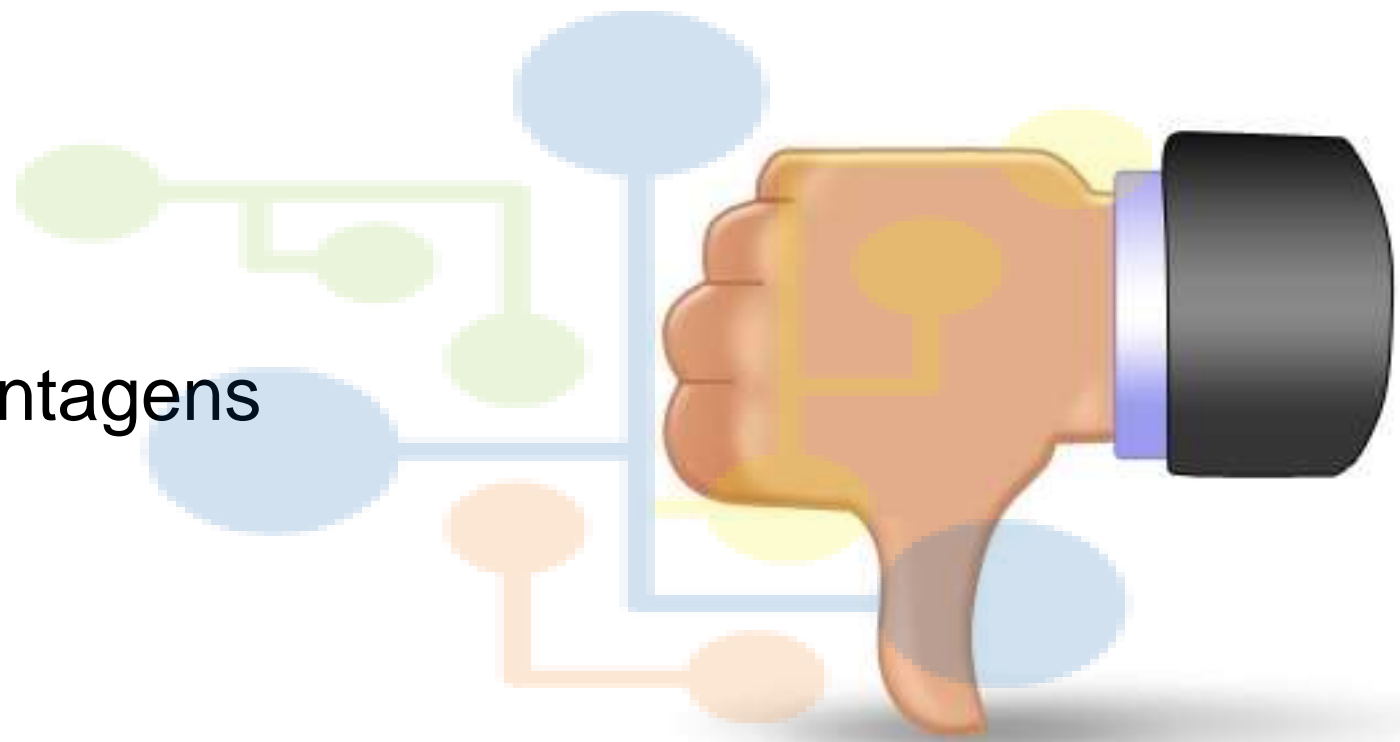
Robustez





# Desvantagens do Hadoop

Desvantagens

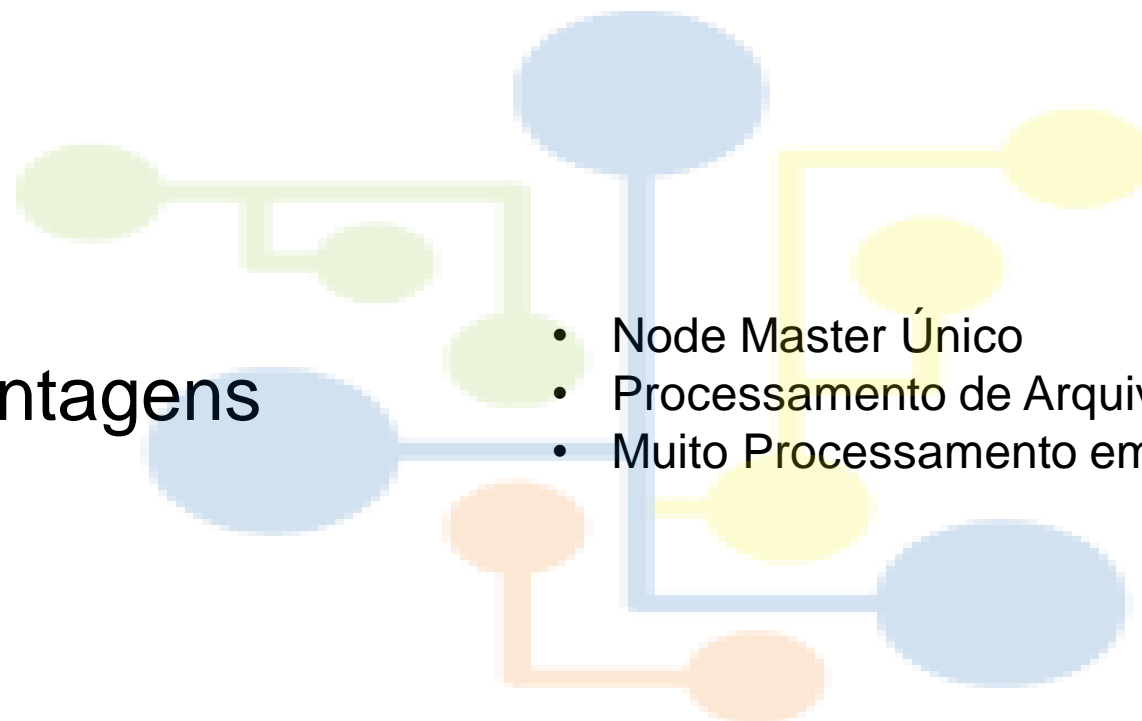




# Desvantagens do Hadoop

## Desvantagens

- Node Master Único
- Processamento de Arquivos Pequenos
- Muito Processamento em Poucos Dados





# **Ecosystema Hadoop**



# Ecosystem Hadoop

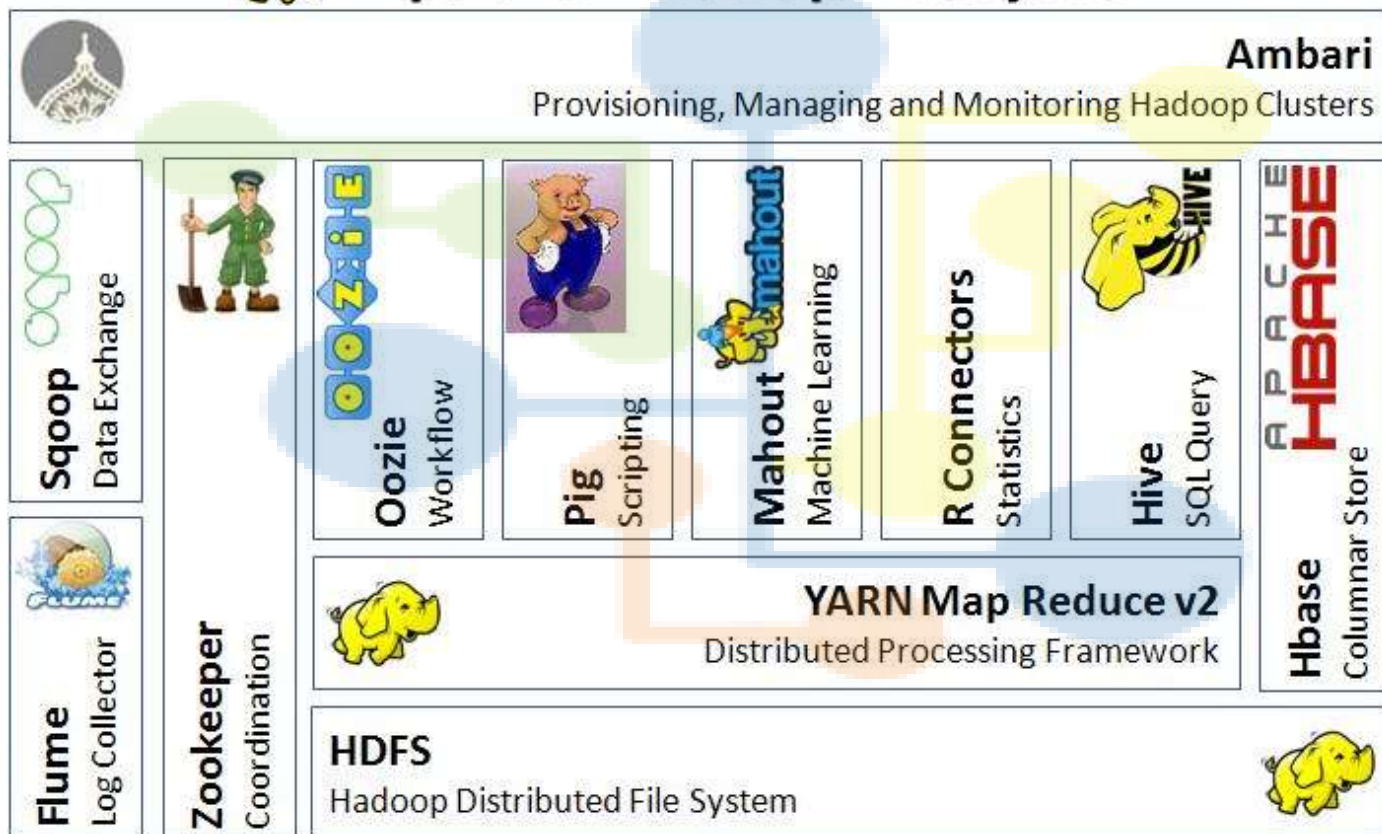


Data Science  
Academy

Data Science Academy marcelo\_eidi12@hotmail.com 5d5c42d55e4cde68f38b457d

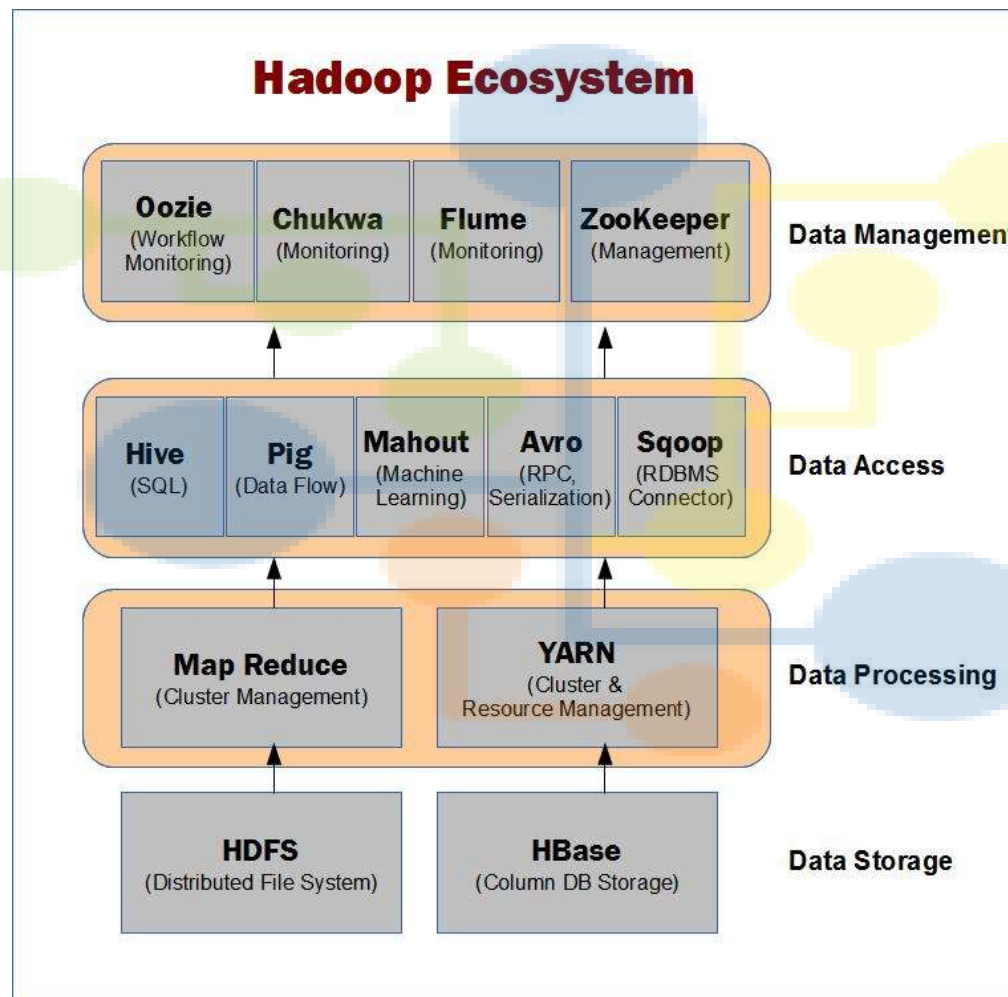


## Apache Hadoop Ecosystem





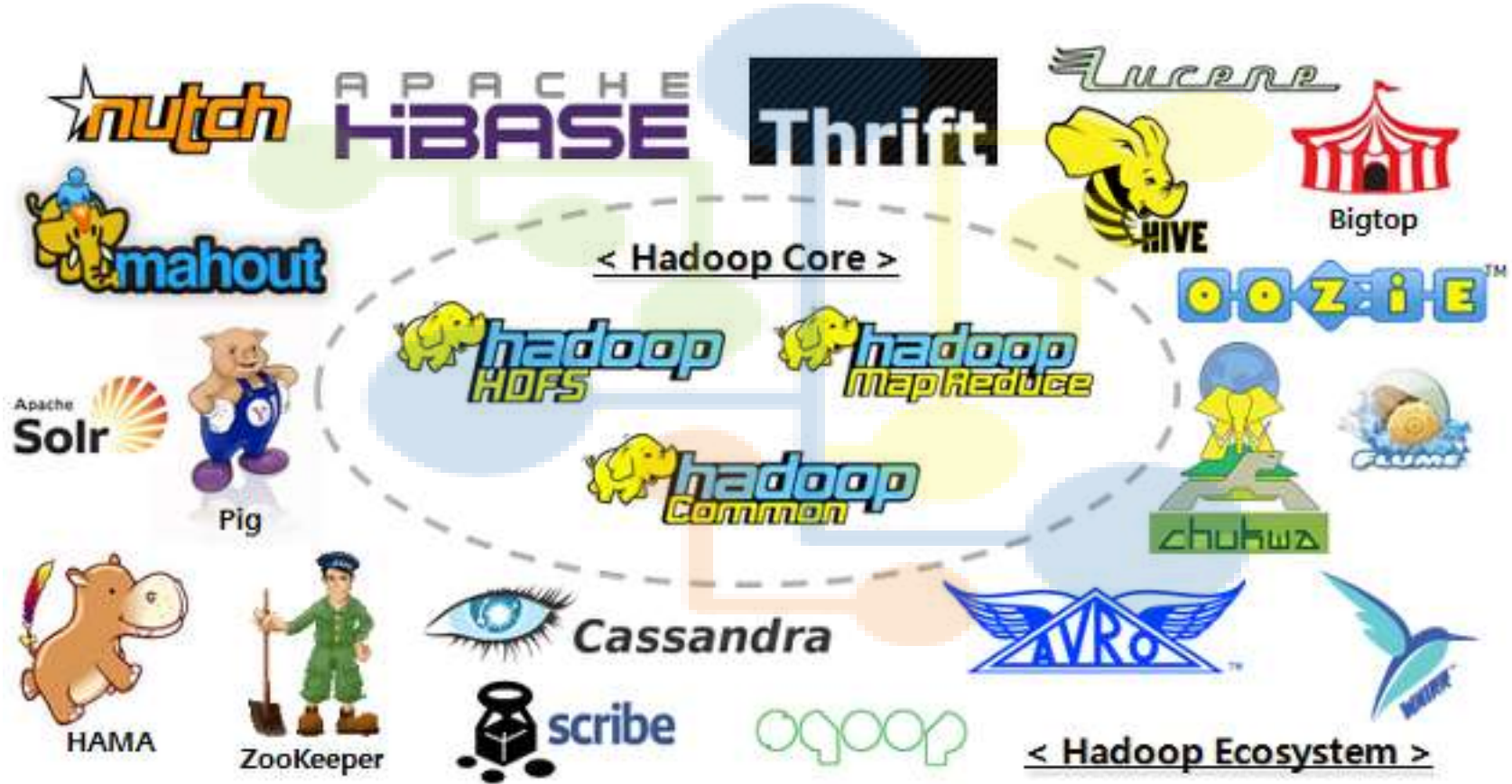
# Ecossistema Hadoop







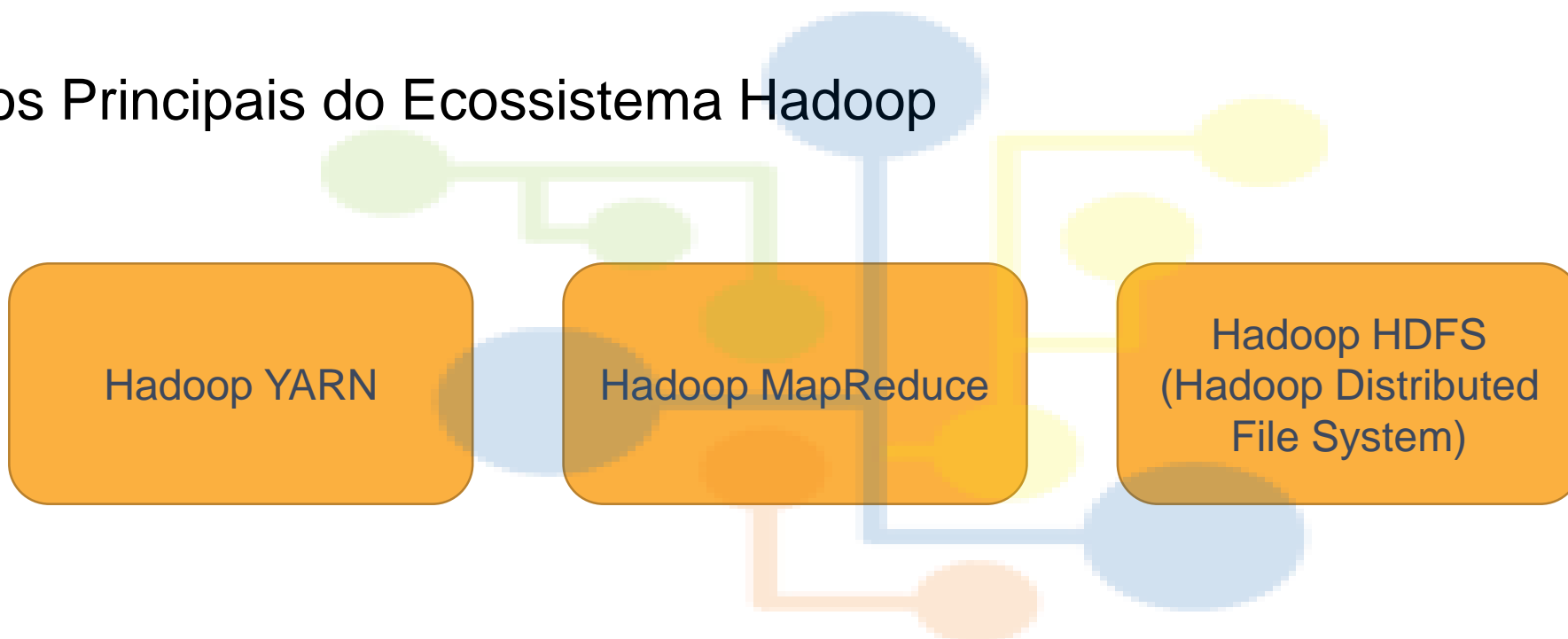
# Ecosystem Hadoop





# Ecossistema Hadoop

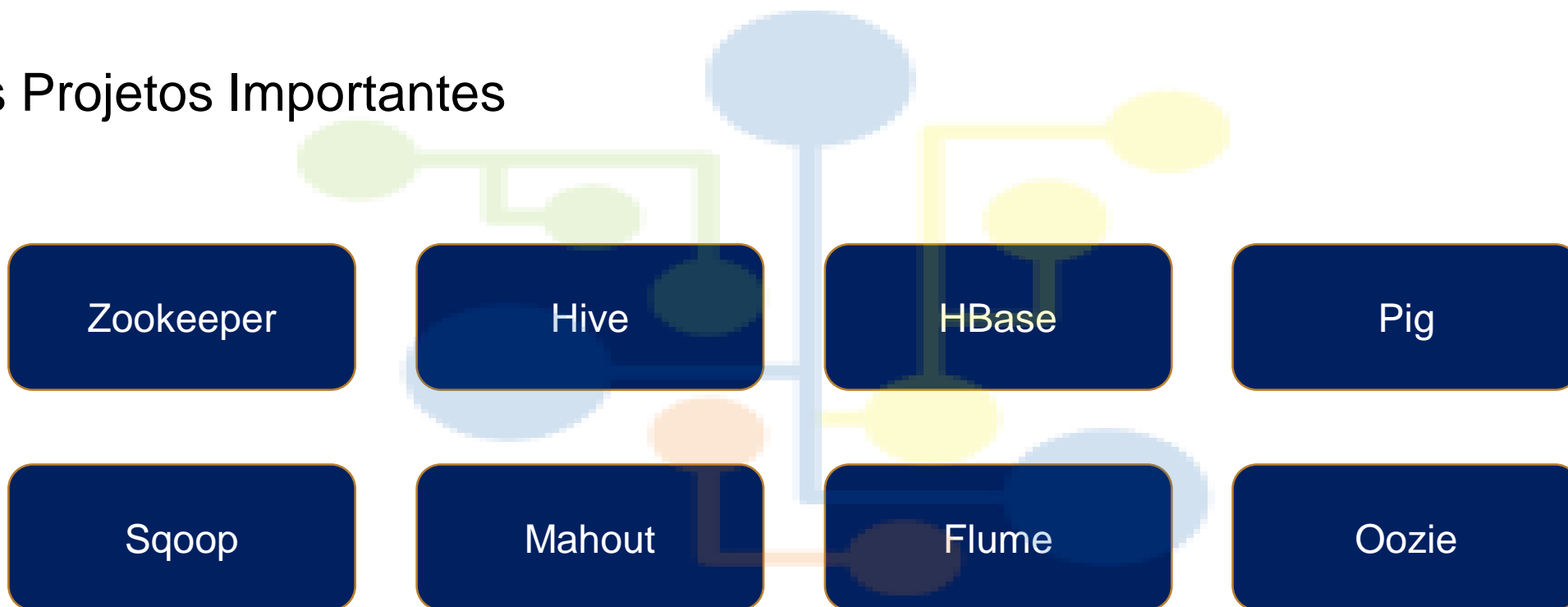
## Projetos Principais do Ecossistema Hadoop





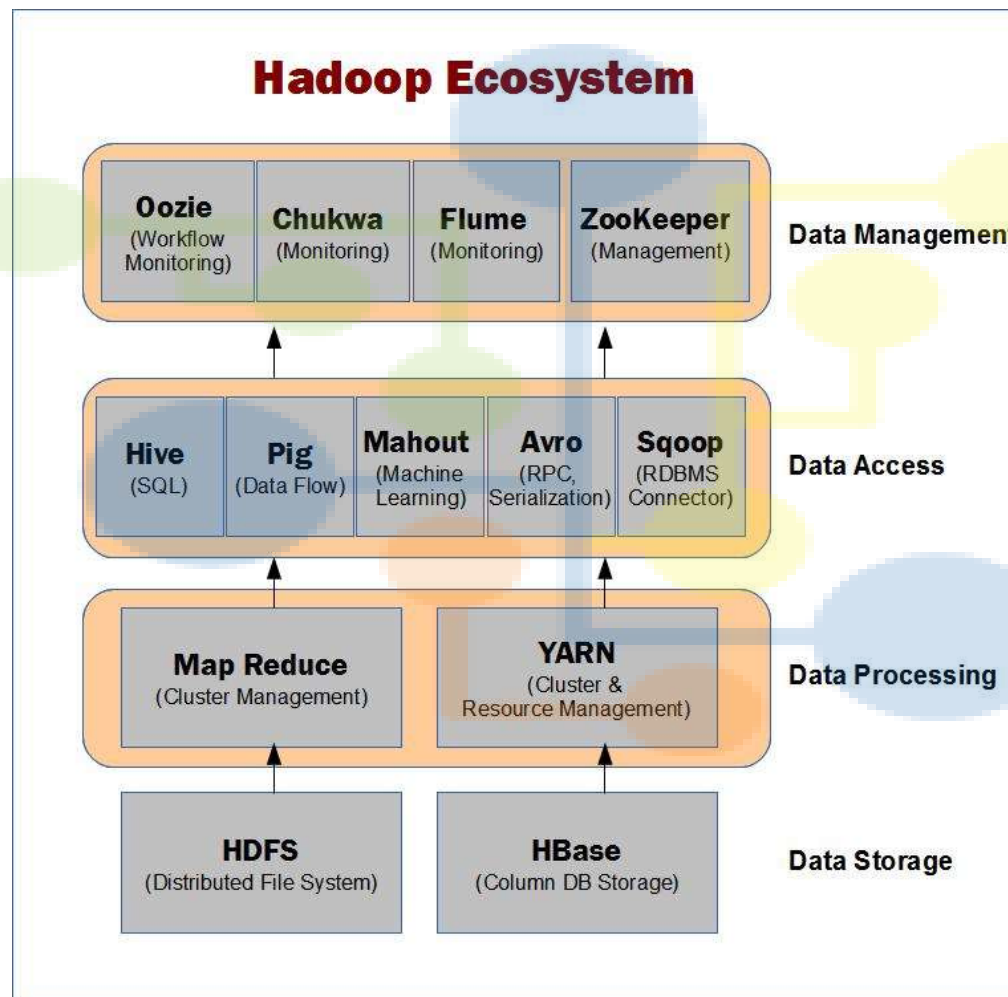
# Ecosystem Hadoop

## Outros Projetos Importantes





# Ecossistema Hadoop

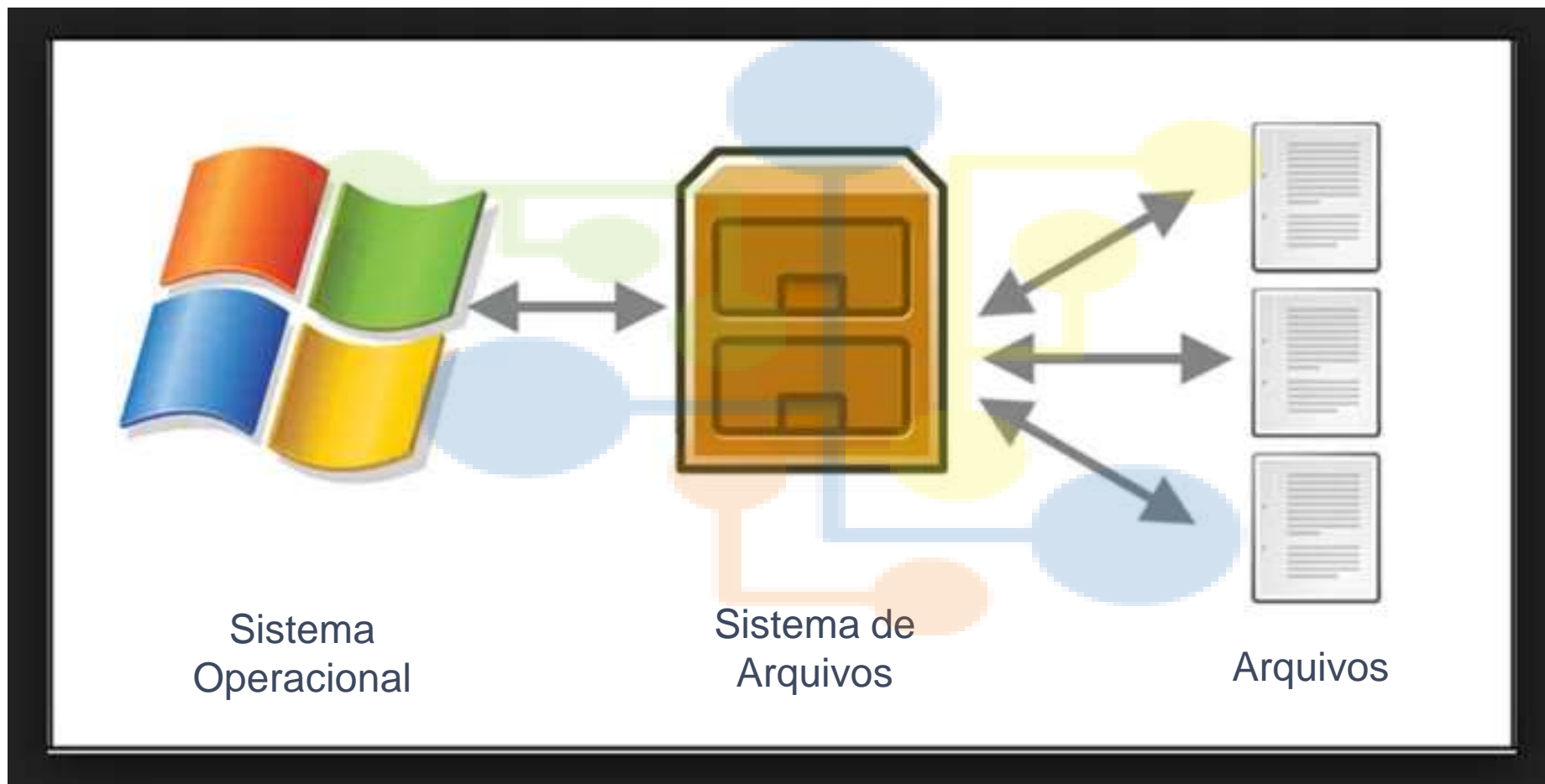




# HDFS (Hadoop Distributed File System) Conceito e Importância



# HDFS – Conceito e Importância





# HDFS – Conceito e Importância

Os tipos de Sistemas de Arquivos são:

Tipo	Descrição
<b>ext2</b>	Sistema de arquivos padrão do Linux
<b>ext3</b>	Sistema de arquivos ext2 melhorado
<b>reiserfs</b>	Sistema de arquivos do tipo Journaling
<b>msdos</b>	Sistema de arquivos FAT da Microsoft DOS
<b>vfat</b>	Sistema de arquivos FAT-32 do Microsoft Windows
<b>iso9660</b>	Sistema de arquivos do CD-ROM
<b>nfs</b>	Network File System. Usado para montar dispositivos em computadores remotos.
<b>swap</b>	Sistema de arquivos de troca utilizado para memória virtual.
<b>proc</b>	Uma janela especial dentro do Kernel do Linux. Utilizada pelos usuários, programas e utilitários para escrever ou ler parâmetros do Kernel. Geralmente montado no diretório <code>/proc</code> .



Data Science  
Academy

Data Science Academy marcelo\_eidi12@hotmail.com 5d5c42d55e4cde68f38b457d

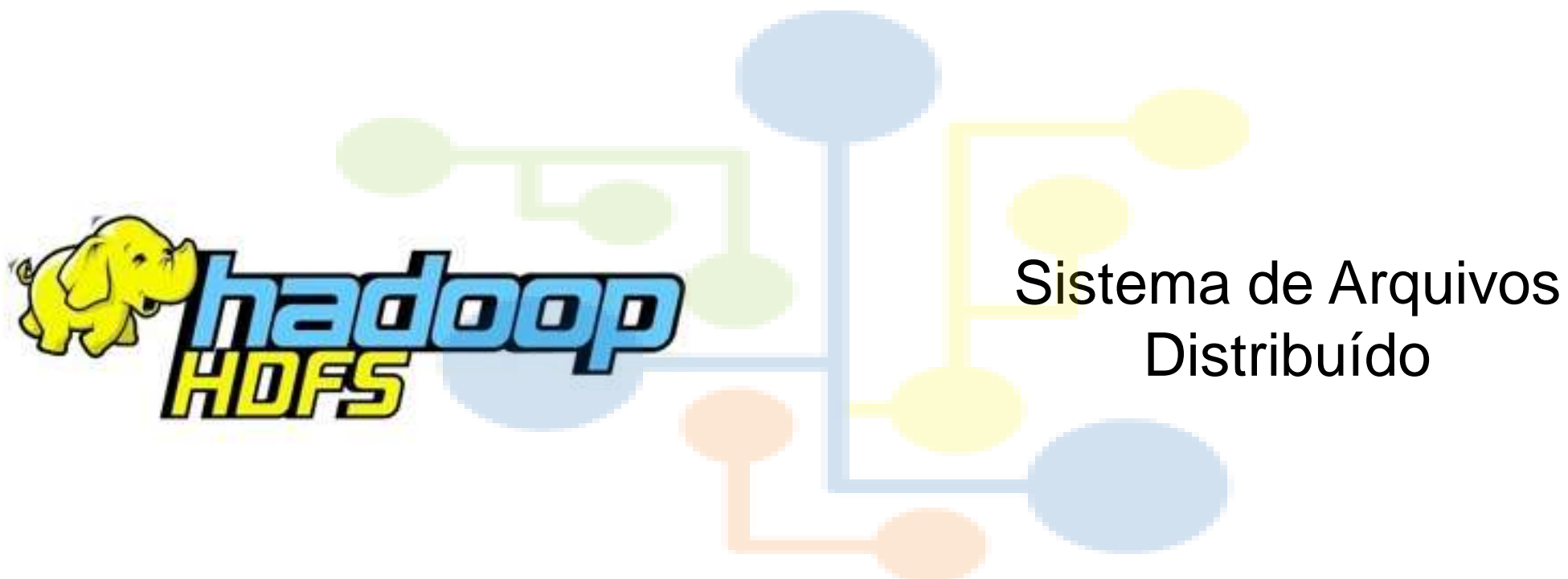
# HDFS – Conceito e Importância







# HDFS – Conceito e Importância





# HDFS – Conceito e Importância

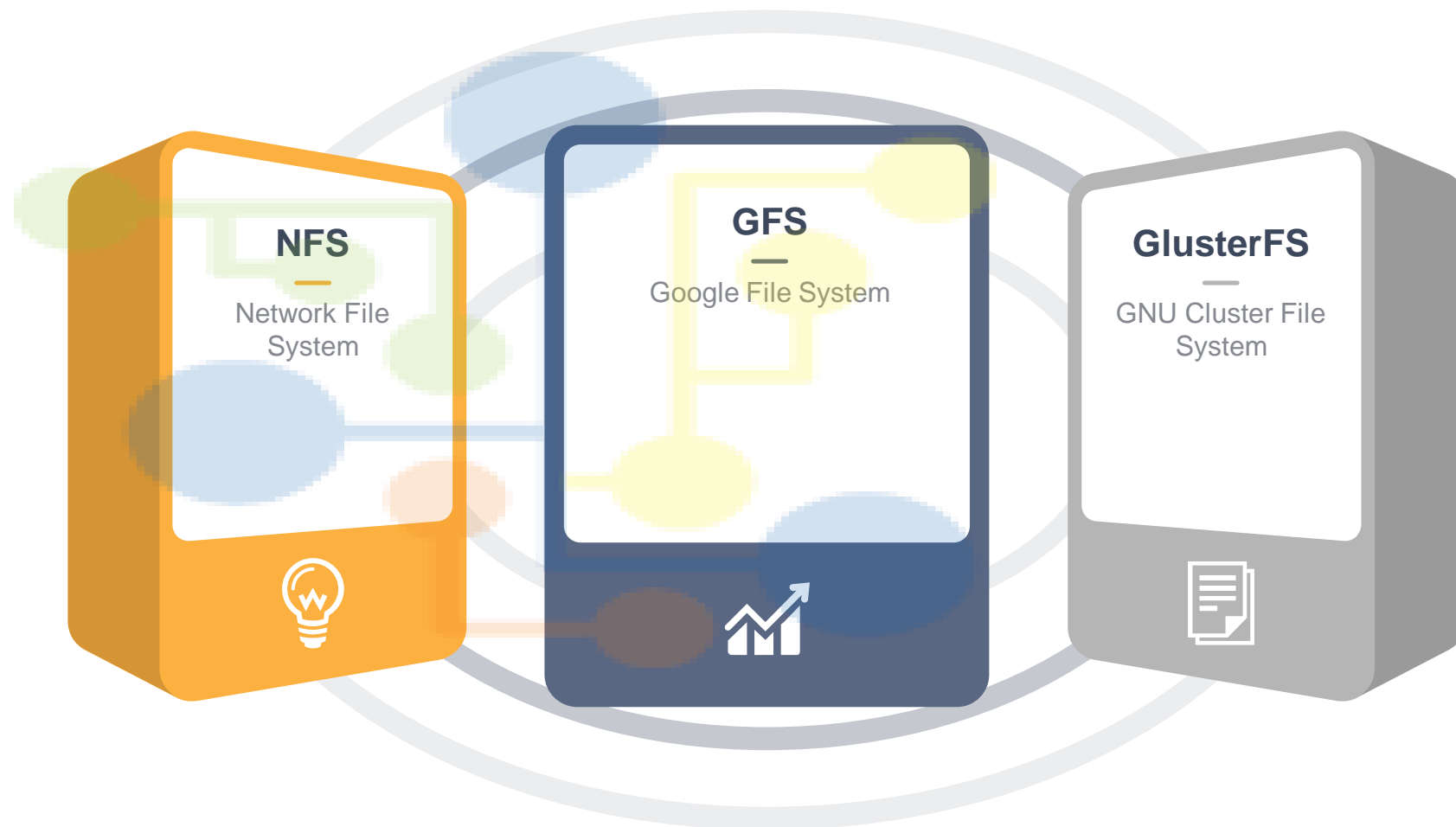


- Tolerância a Falhas
- Integridade
- Segurança
- Desempenho
- Consistência



# HDFS – Conceito e Importância

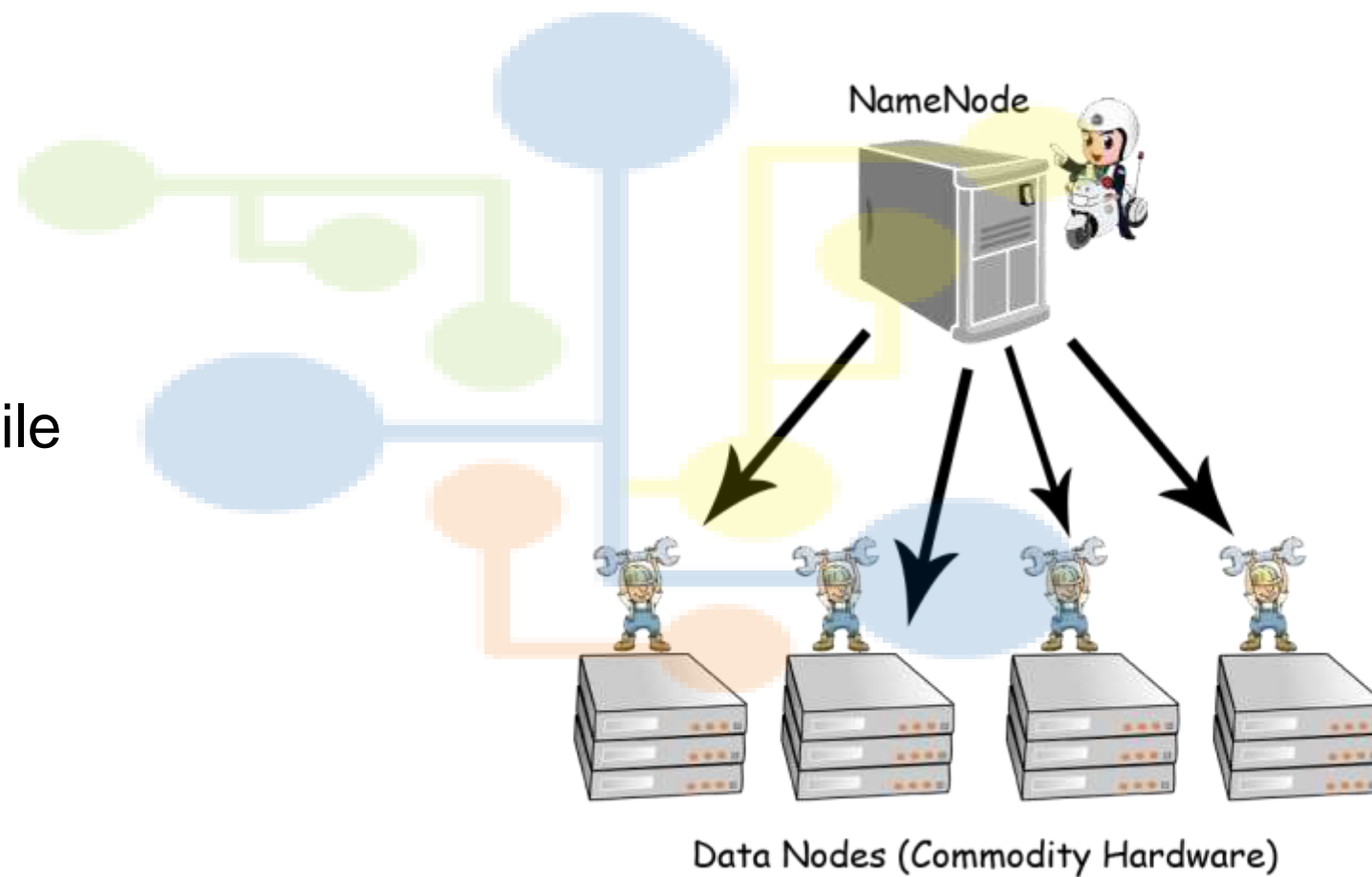
Outros Sistemas  
de Arquivos  
Distribuídos





# HDFS – Conceito e Importância

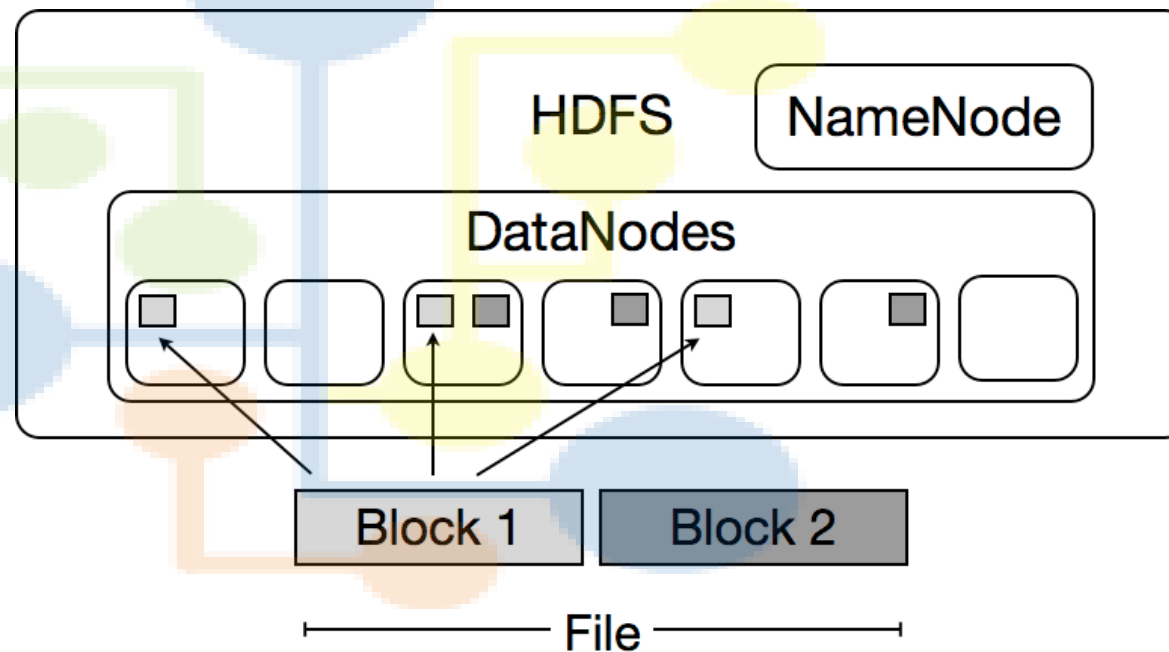
Hadoop  
Distributed File  
System





# HDFS – Conceito e Importância

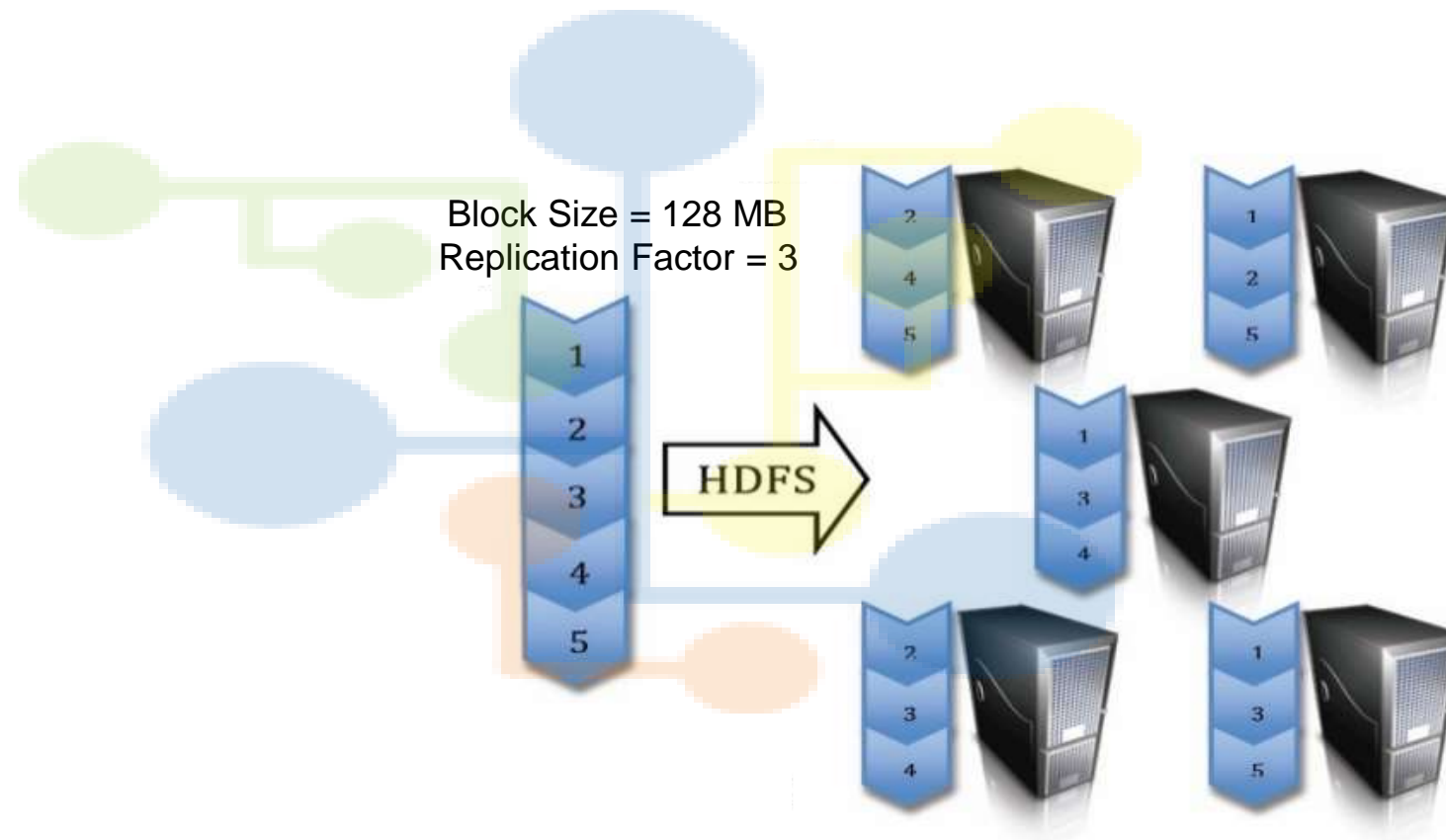
Hadoop  
Distributed File  
System





# HDFS – Conceito e Importância

Hadoop  
Distributed File  
System





# HDFS – Conceito e Importância

**O HDFS foi criado para resolver "Big Problems" e por isso seu funcionamento e arquitetura são próprios para se trabalhar com grandes arquivos de dados e distribuir esses arquivos em blocos ao longo de um cluster de computadores, para que possam ser processados em paralelo.**



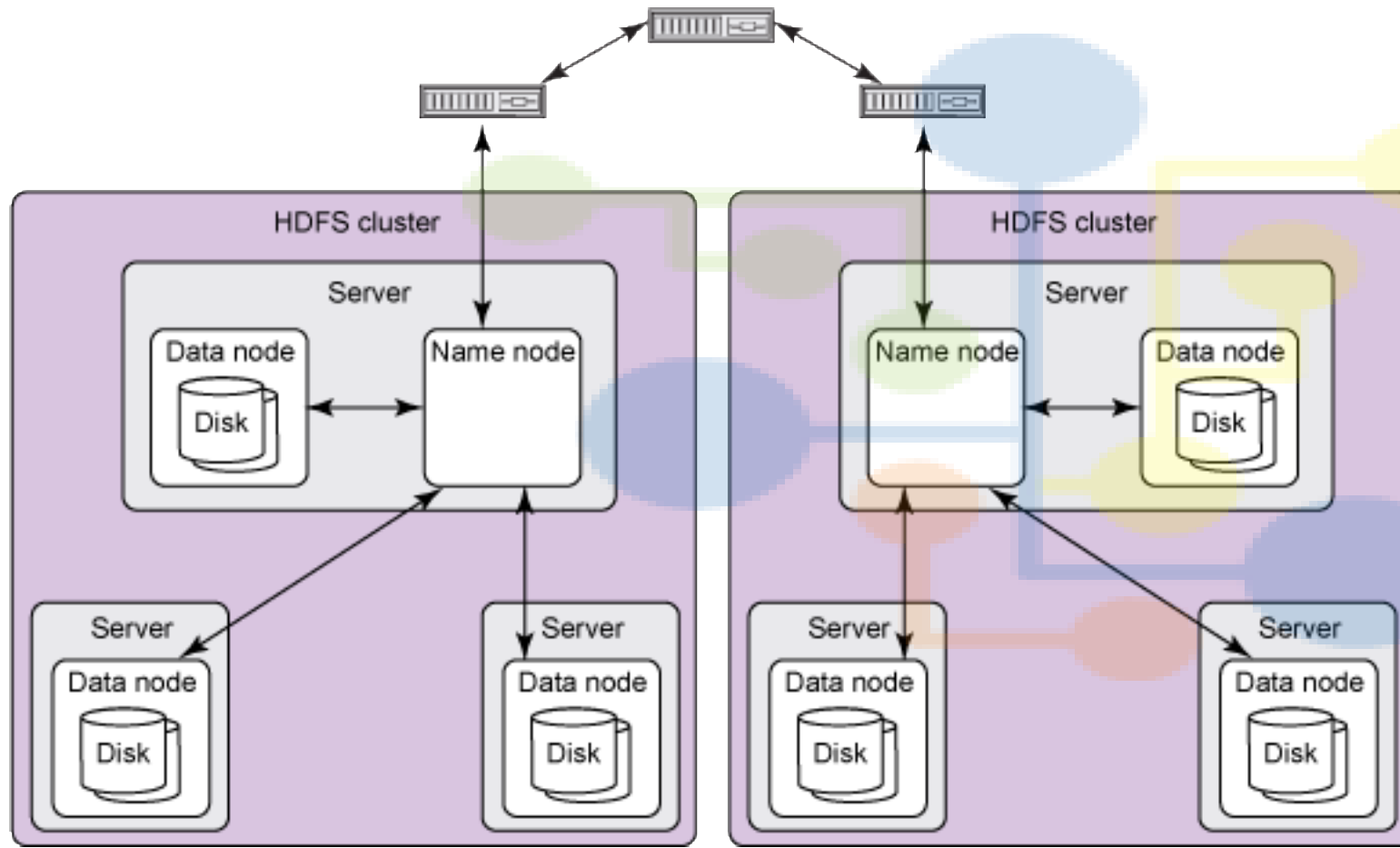
# HDFS (Hadoop Distributed File System) Arquitetura

A faint, stylized diagram of the HDFS architecture is visible in the background. It shows a central blue node connected to several other nodes (blue, yellow, and orange) in a distributed manner, representing the Master-Slave architecture.





# HDFS – Arquitetura



Arquitetura  
Master/Worker



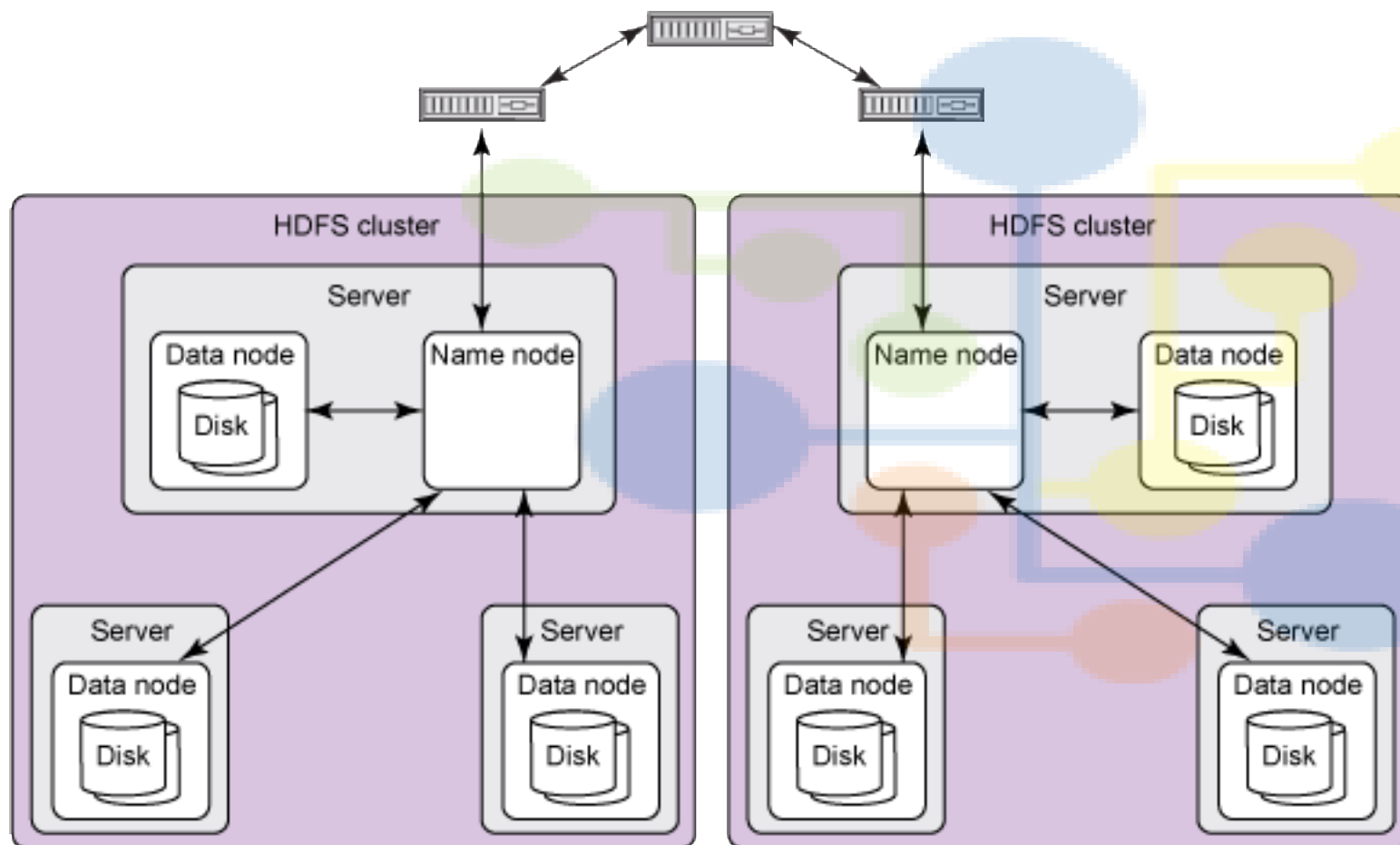
# HDFS – Arquitetura



Arquitetura  
Master/Worker



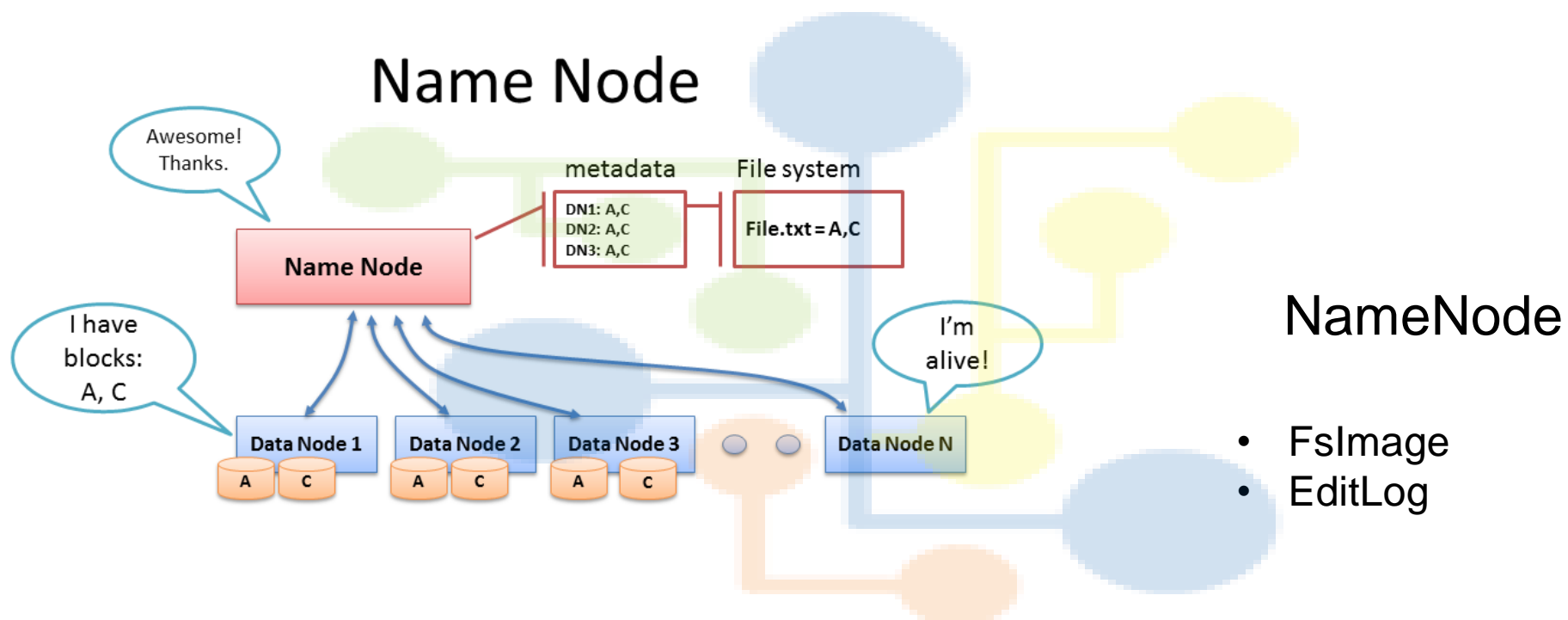
# HDFS – Arquitetura



Arquitetura  
Master/Worker



# HDFS – Arquitetura

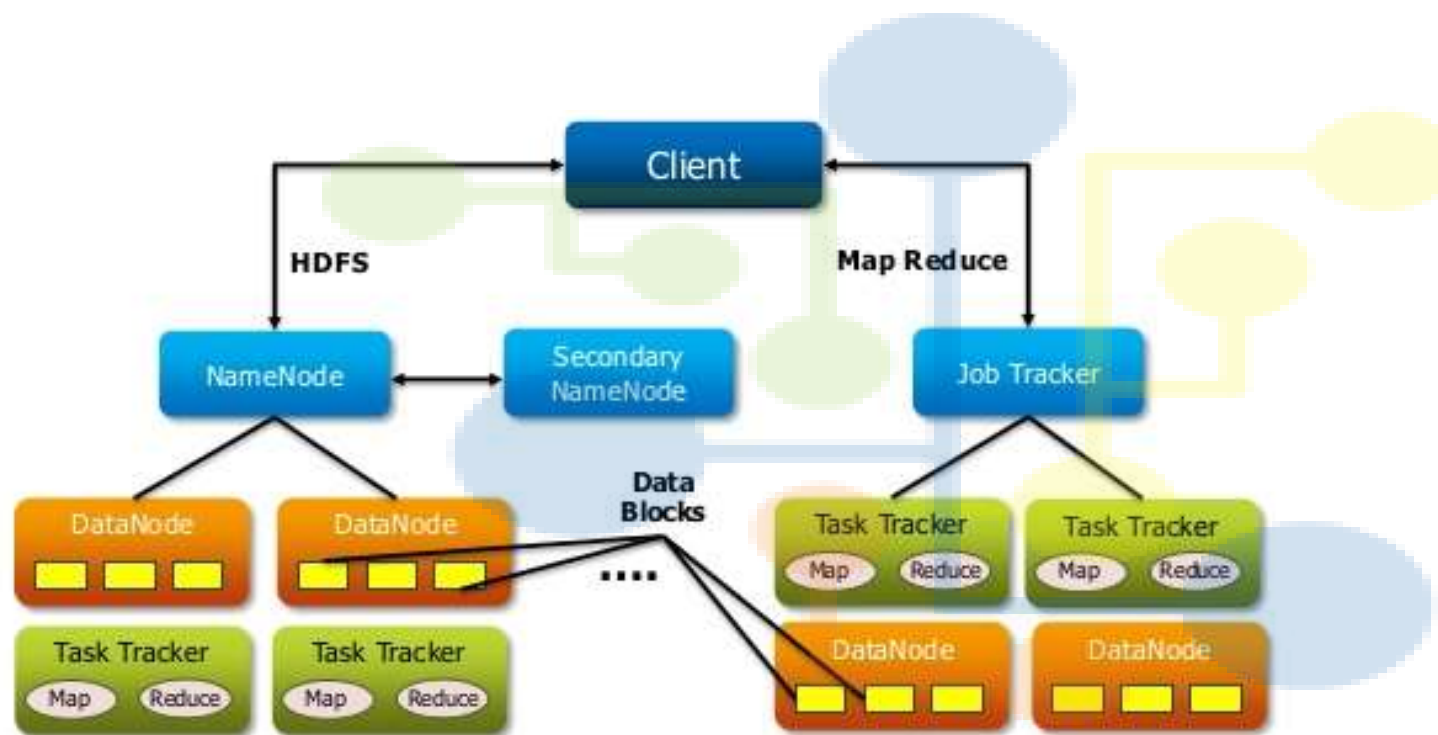


# HDFS – Arquitetura



Data Science  
Academy

Data Science Academy marcelo\_eidi12@hotmail.com 5d5c42d55e4cde68f38b457d



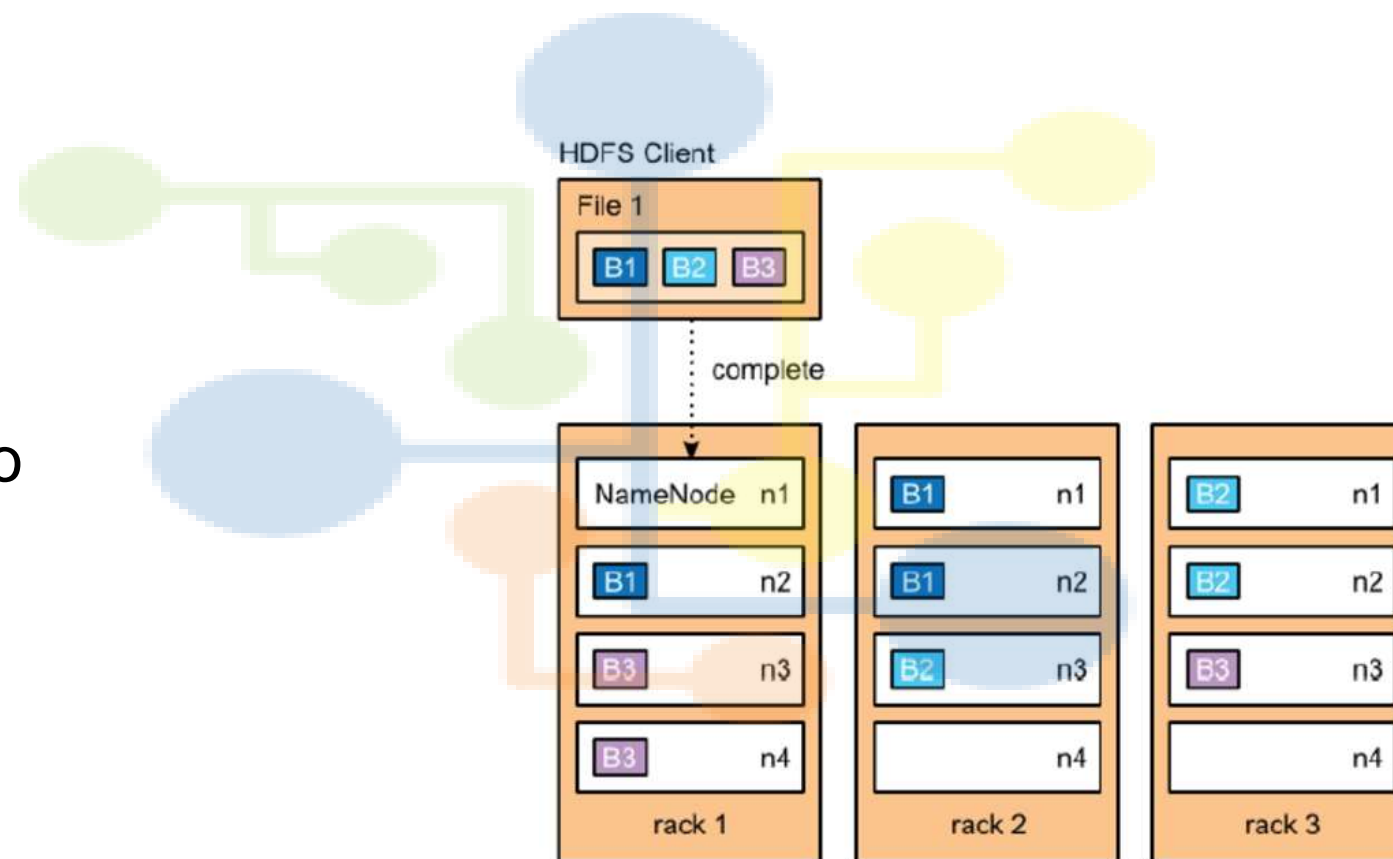
NameNode

- FsImage
- EditLog



# HDFS – Arquitetura

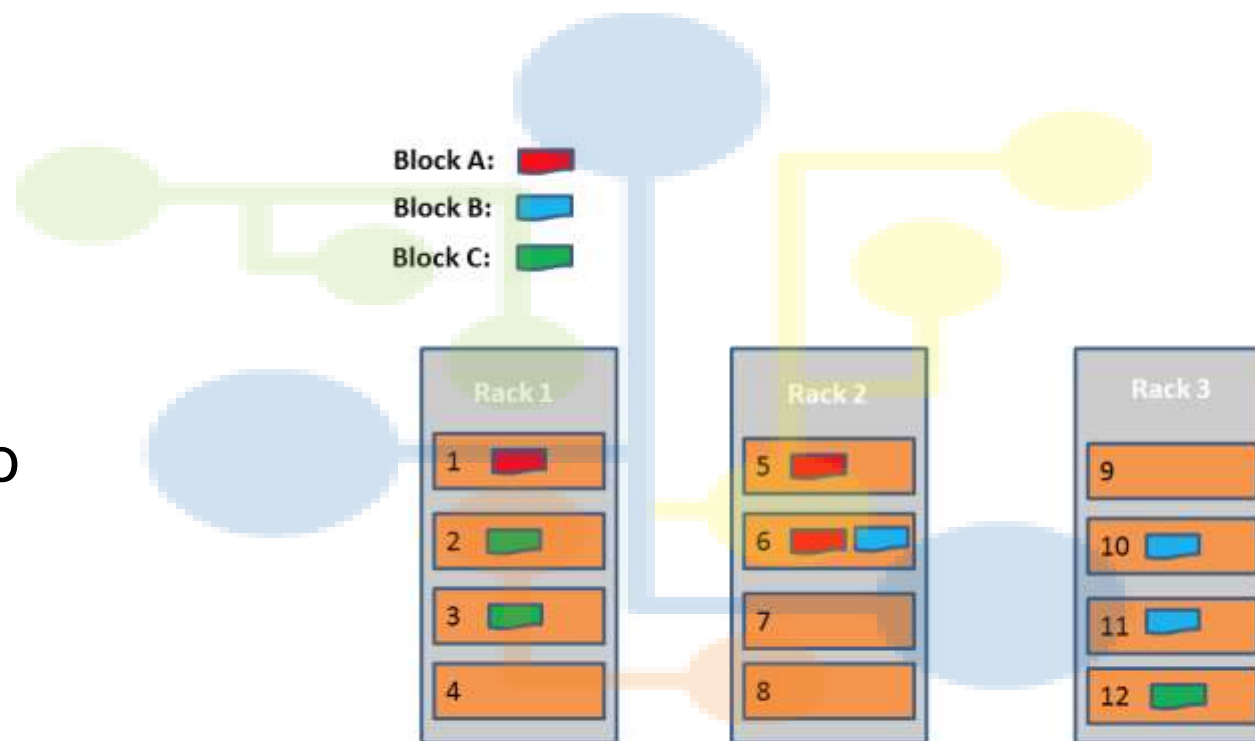
Replicação





# HDFS – Arquitetura

Replicação





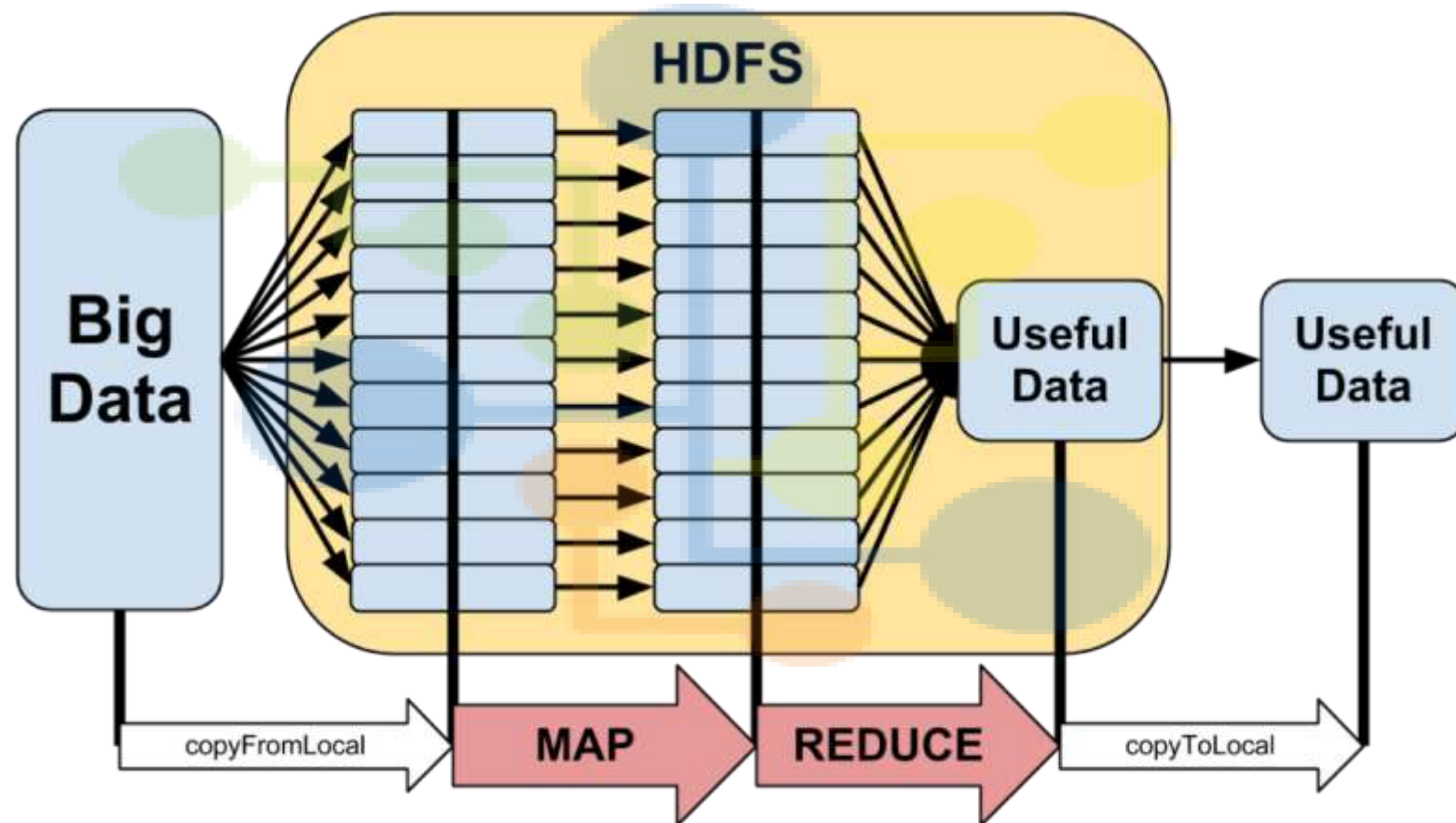
# Definindo MapReduce





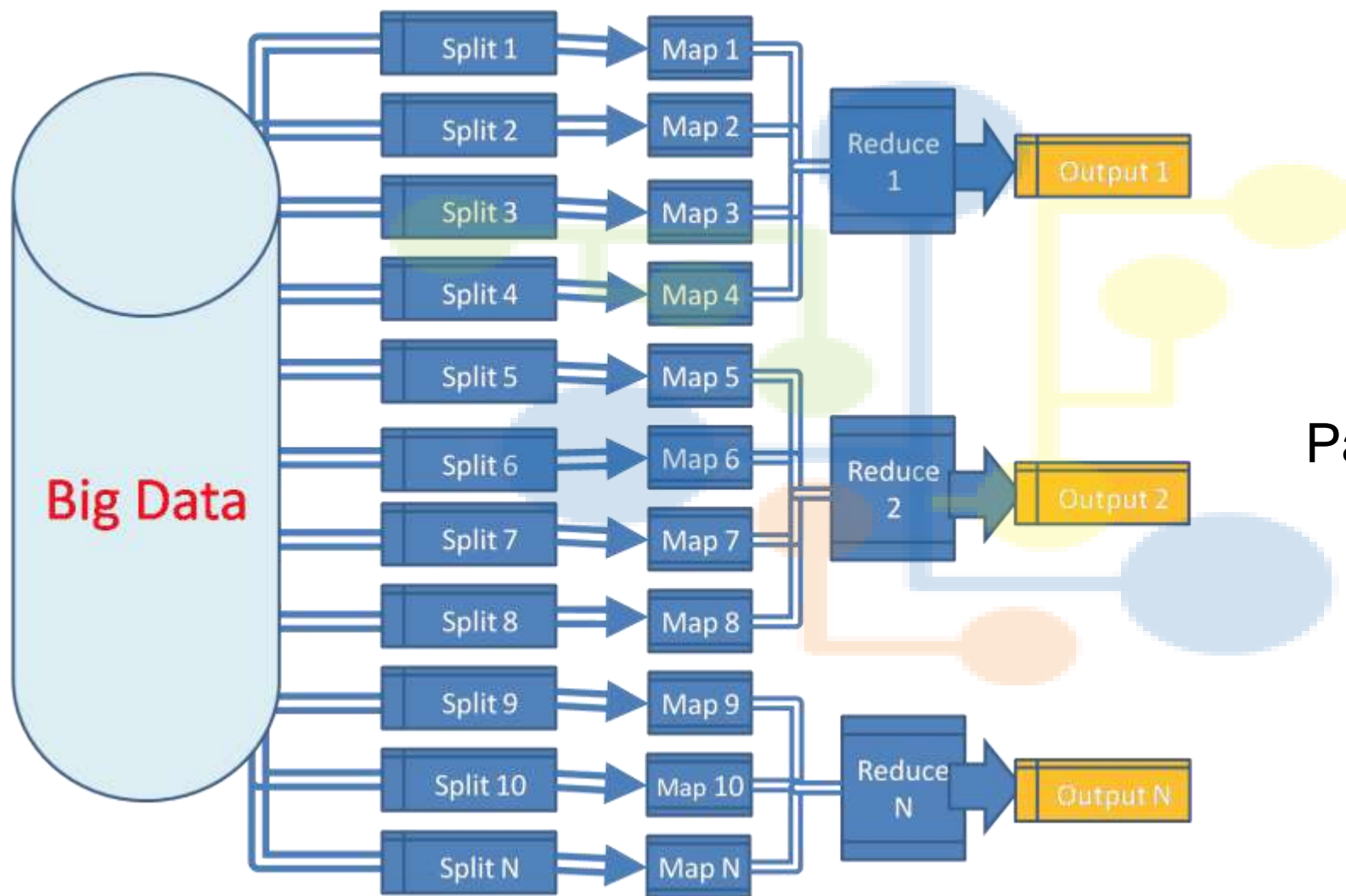


# Definindo MapReduce





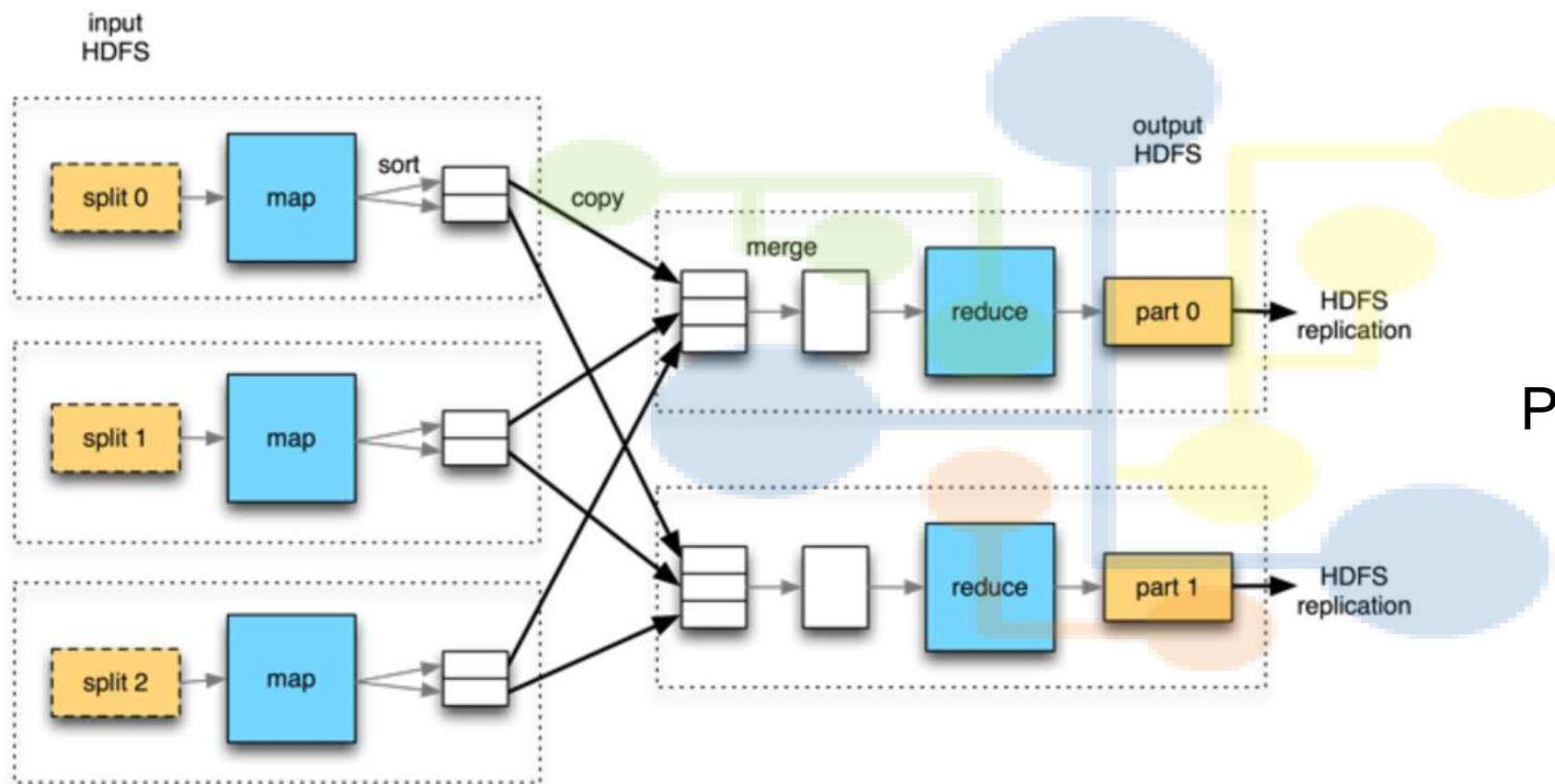
# Definindo MapReduce



Processamento  
Paralelo e Distribuído



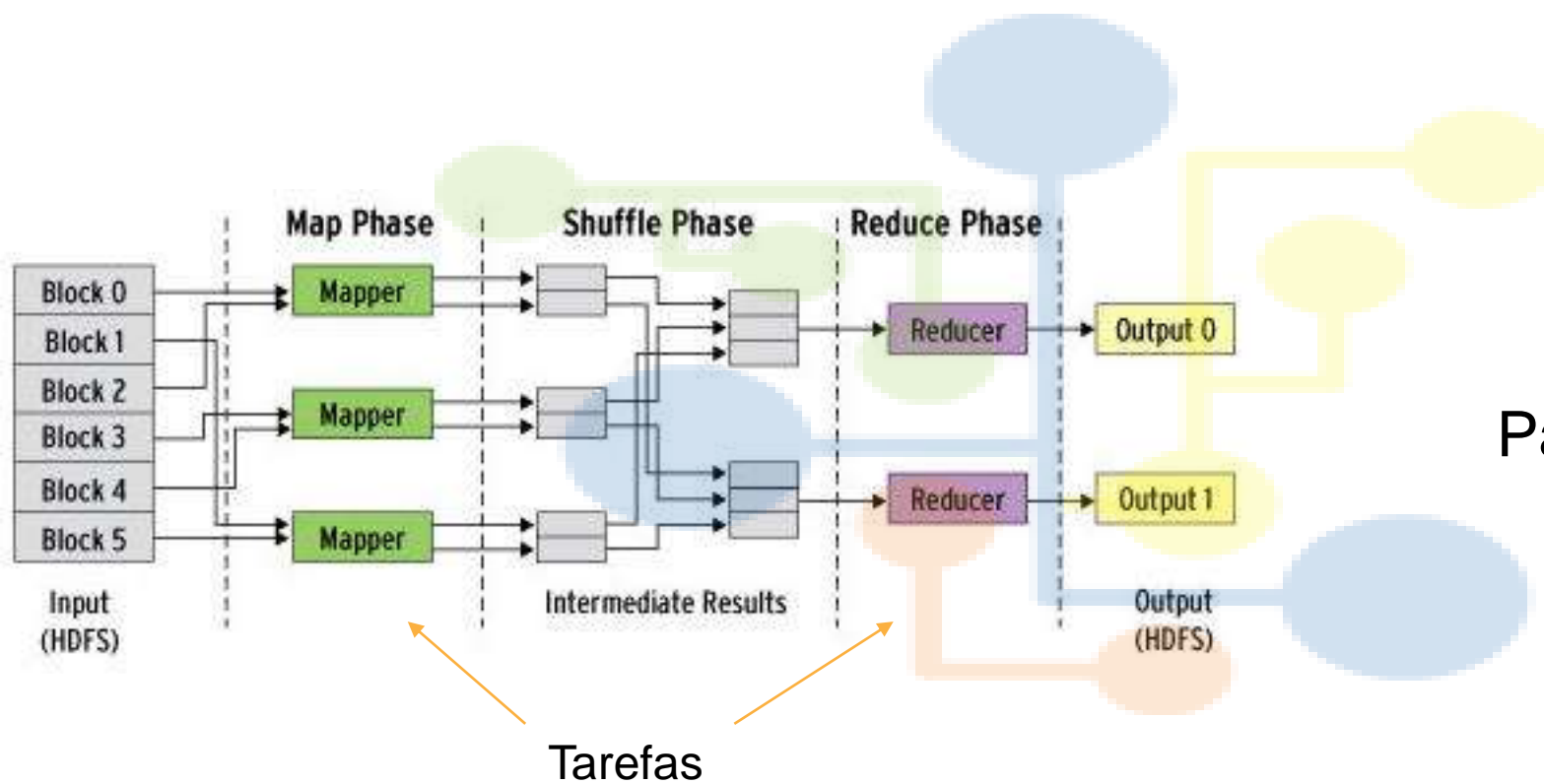
# Definindo MapReduce



Processamento  
Paralelo e Distribuído



# Definindo MapReduce



Processamento  
Paralelo e Distribuído



# Hadoop x Bancos de Dados Relacionais

A faint, stylized diagram in the background, consisting of various colored circles (blue, green, yellow, orange) connected by lines, resembling a network or data flow.



# Hadoop x Bancos de Dados Relacionais



Bancos de Dados Relacionais

A faint, stylized diagram of a relational database structure is visible in the background. It consists of several circular nodes connected by lines. The nodes are colored in light blue, light green, light yellow, and light orange. The connections are represented by lines of the same colors, forming a network that suggests data relationships between different tables or components of a database.



# Hadoop x Bancos de Dados Relacionais

RDBMS  
Relational Database  
Management Systems

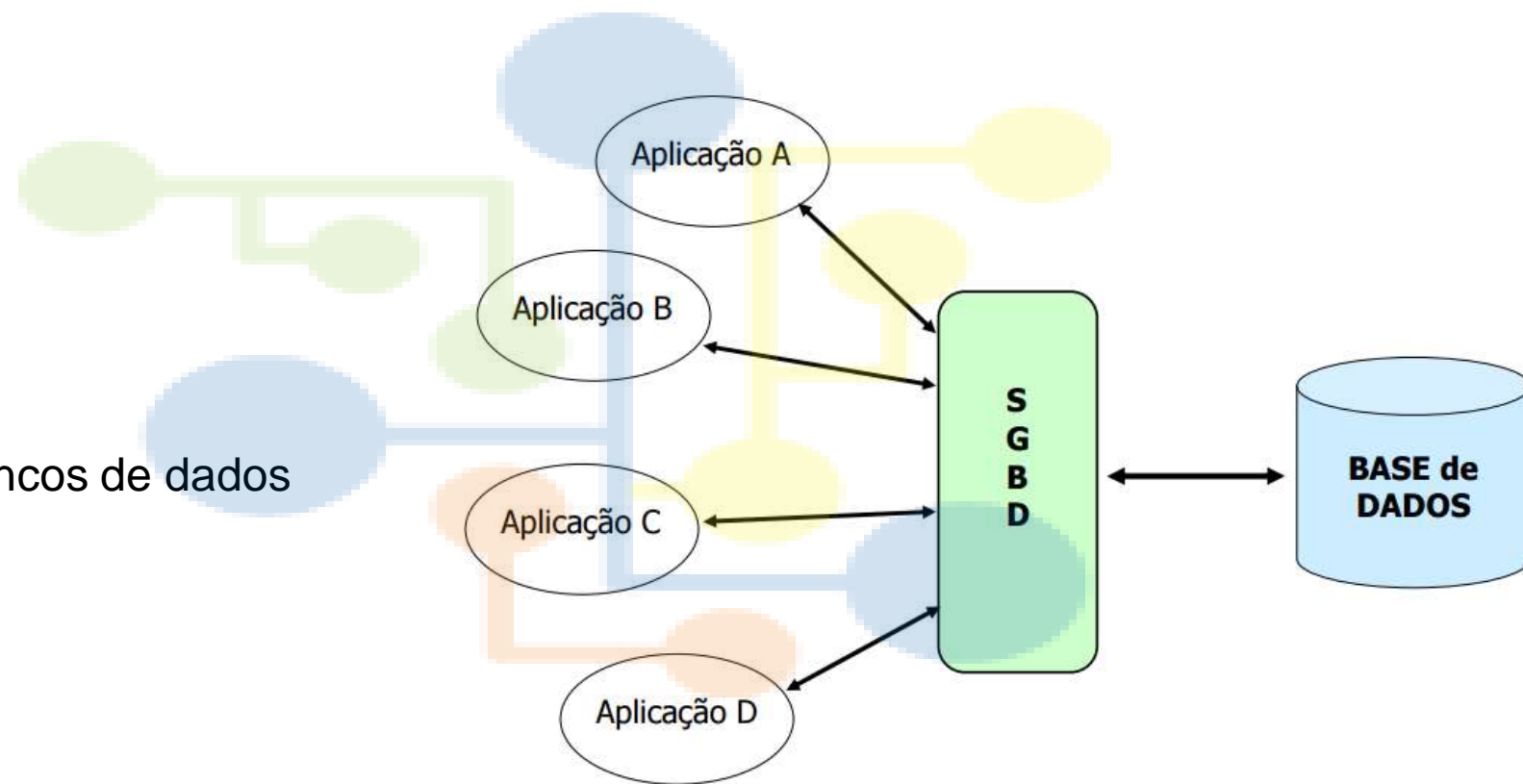




# Hadoop x Bancos de Dados Relacionais

SGBD's

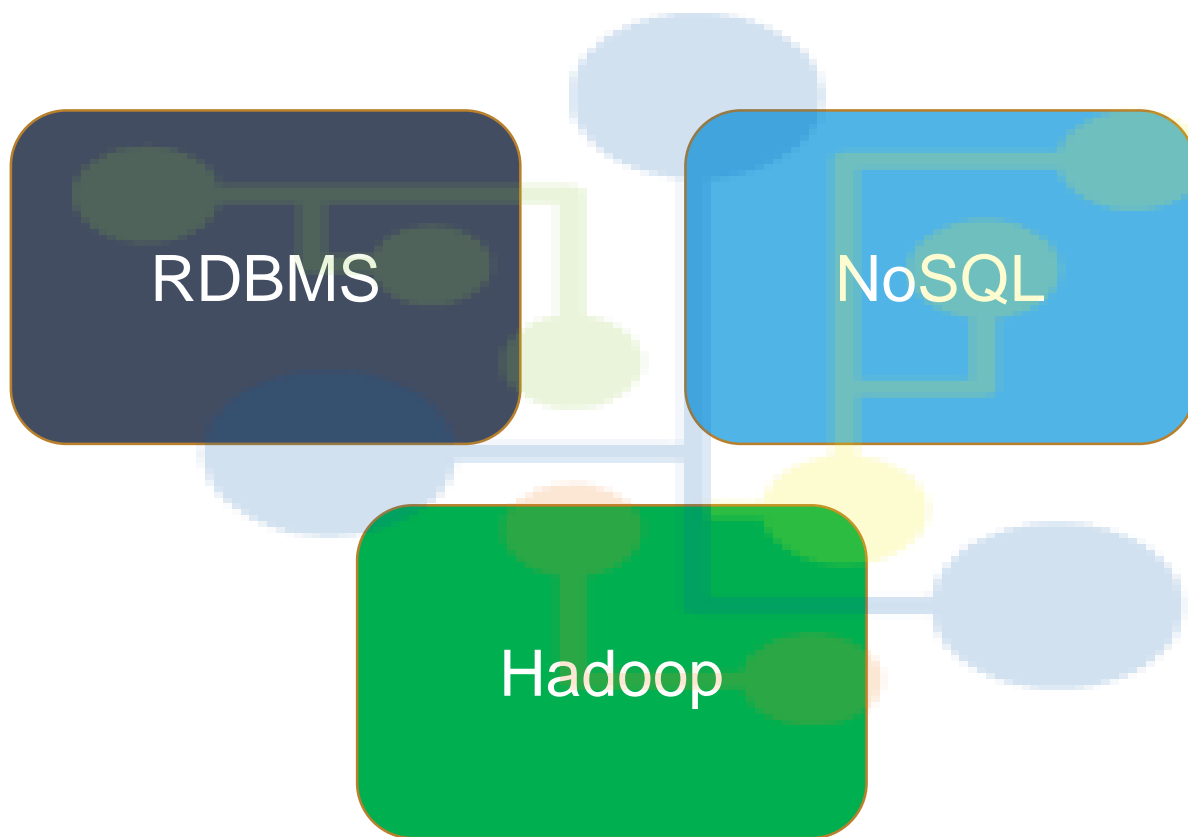
Gerenciam um ou mais bancos de dados







# Hadoop x Bancos de Dados Relacionais





Data Science  
Academy

Data Science Academy marcelo\_eidi12@hotmail.com 5d5c42d55e4cde68f38b457d

# Hadoop x Bancos de Dados Relacionais





# Hadoop x Bancos de Dados Relacionais





# Hadoop x Bancos de Dados Relacionais

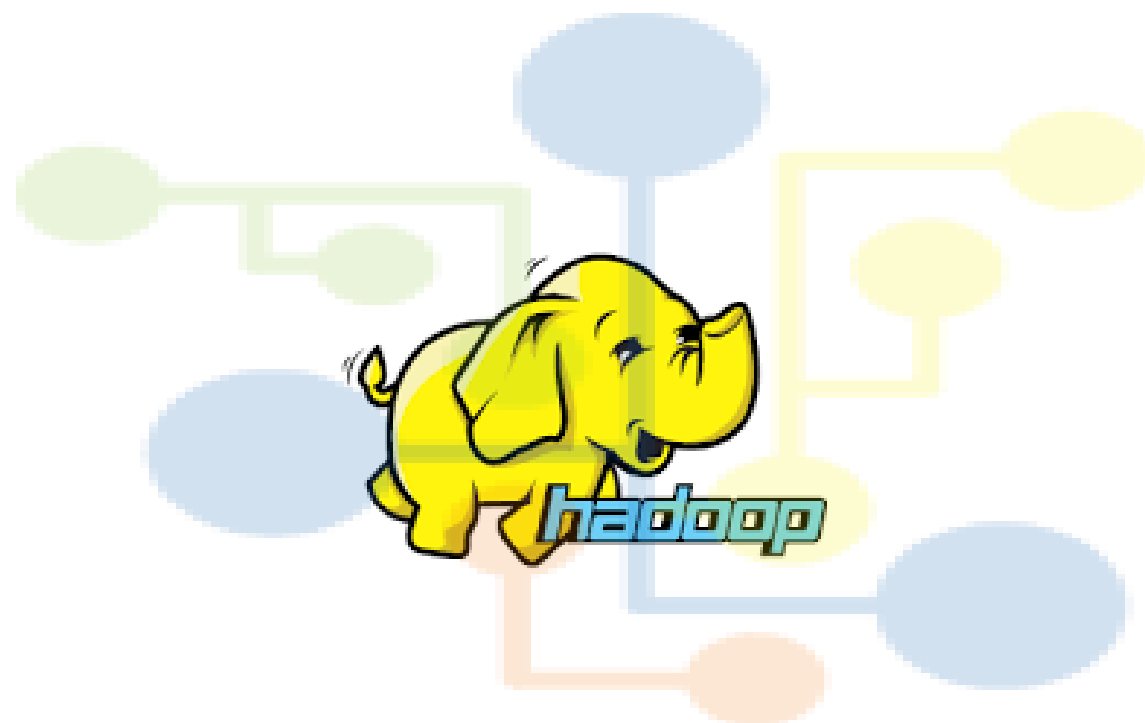




Data Science  
Academy

Data Science Academy marcelo\_eidi12@hotmail.com 5d5c42d55e4cde68f38b457d

# Hadoop x Bancos de Dados Relacionais

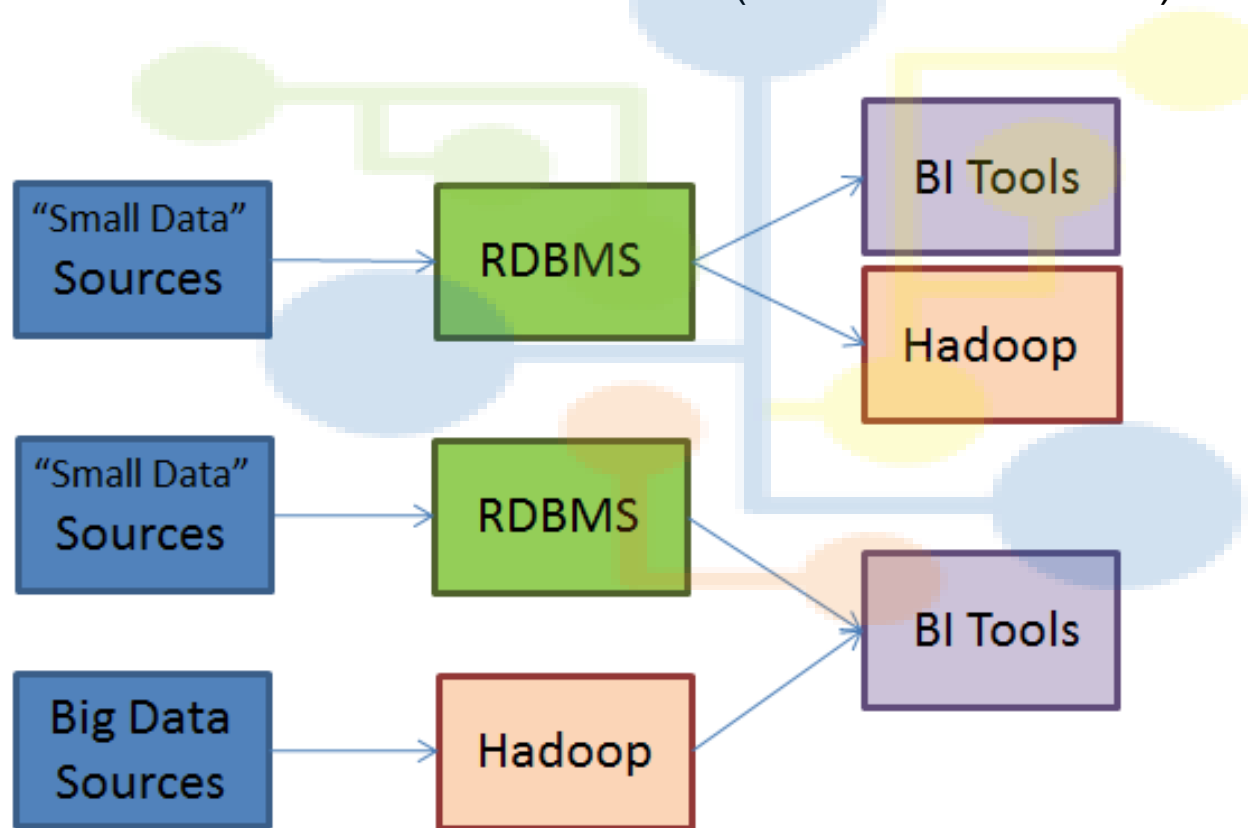




# Hadoop x Bancos de Dados Relacionais

Hadoop → Grandes volumes de dados (estruturados ou não estruturados)

RDBMS → Dados transacionais (dados estruturados)





Data Science  
Academy

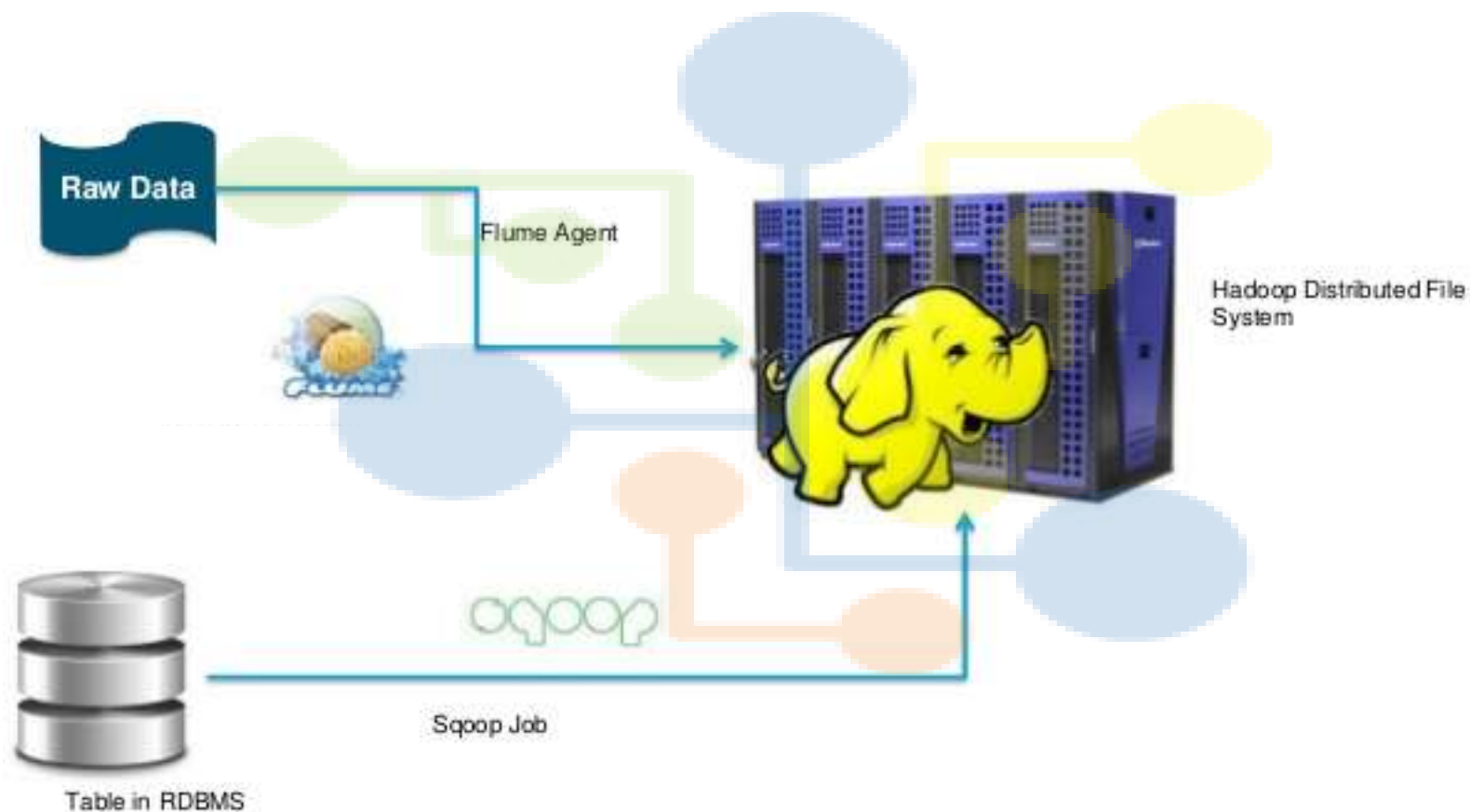
Data Science Academy marcelo\_eidi12@hotmail.com 5d5c42d55e4cde68f38b457d

# Hadoop x Bancos de Dados Relacionais





# Hadoop x Bancos de Dados Relacionais







# Hadoop x Bancos de Dados Relacionais

Hadoop processa dados em batch. Consequentemente, ele não deve ser usado para processar dados transacionais. Mas o Hadoop pode resolver muitos outros tipos de problemas relacionados ao Big Data.



# Obrigado

---