Engenharia de Dados com Hadoop e Spark

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

Bem-vindo(a)

Instrutores

Data Science Academy

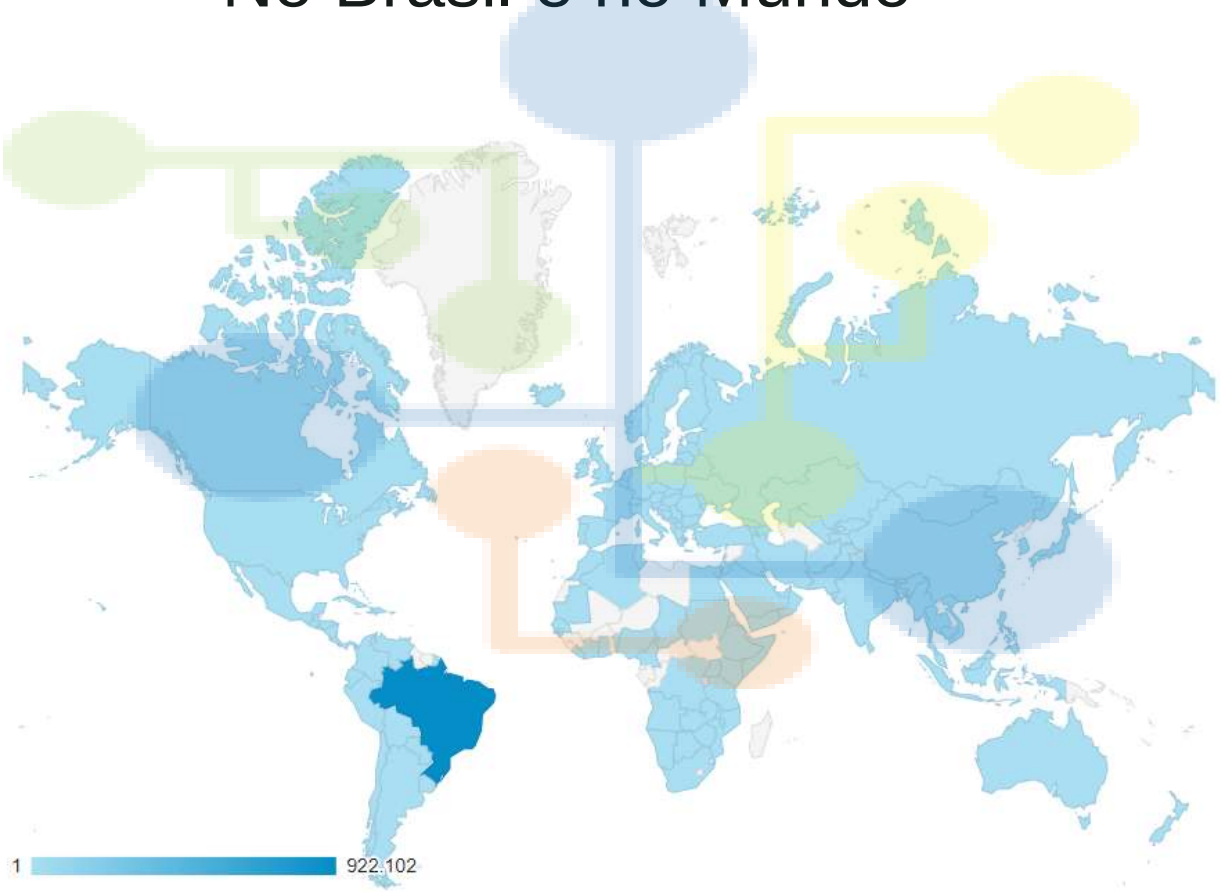
A Data Science Academy (DSA) é um portal de ensino online especializado em Big Data, Machine Learning, Inteligência Artificial, Desenvolvimento de Chatbots, Blockchain e tecnologias relacionadas.

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.

Instrutores

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

No Brasil e no Mundo



Engenharia de Dados com Hadoop e Spark

Engenharia de Dados com Hadoop e Spark

Engenharia de Dados com Hadoop e Spark

O que você vai aprender neste curso?

- Conceitos e definições de Big Data, Hadoop, Ecossistema Hadoop e Spark
- Como planejar, instalar e configurar um cluster Hadoop
- Como planejar, instalar e configurar o Ecossistema Hadoop (Hive, Hbase, Zookeeper, Flume, Oozie, Ambari, Sqoop, Spark e Storm)
- Configuração e utilização do HDFS e configurações avançadas do cluster Hadoop
- Administração e Manutenção do Hadoop e Spark

Engenharia de Dados com Hadoop e Spark

O que você vai aprender neste curso?

- Machine Learning com Apache Mahout
- Importação/exportação de dados e ETL com Sqoop
- Principais distribuições Hadoop do mercado: Cloudera e Hortonworks
- Infraestrutura de Big Data
- Análise de Big Data

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)

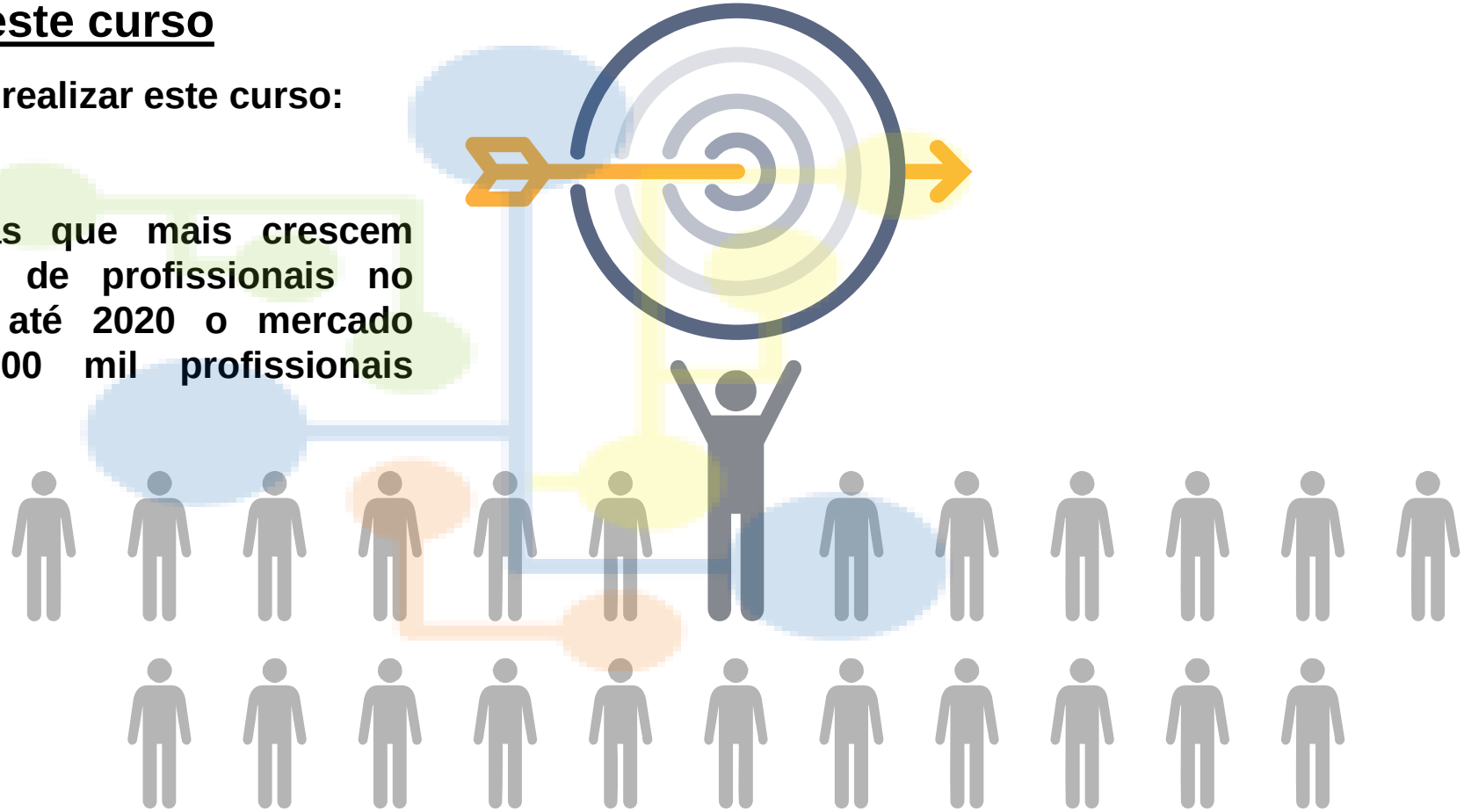
Muita vontade de aprender e entrar no mundo do Big Data (mandatório)

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.

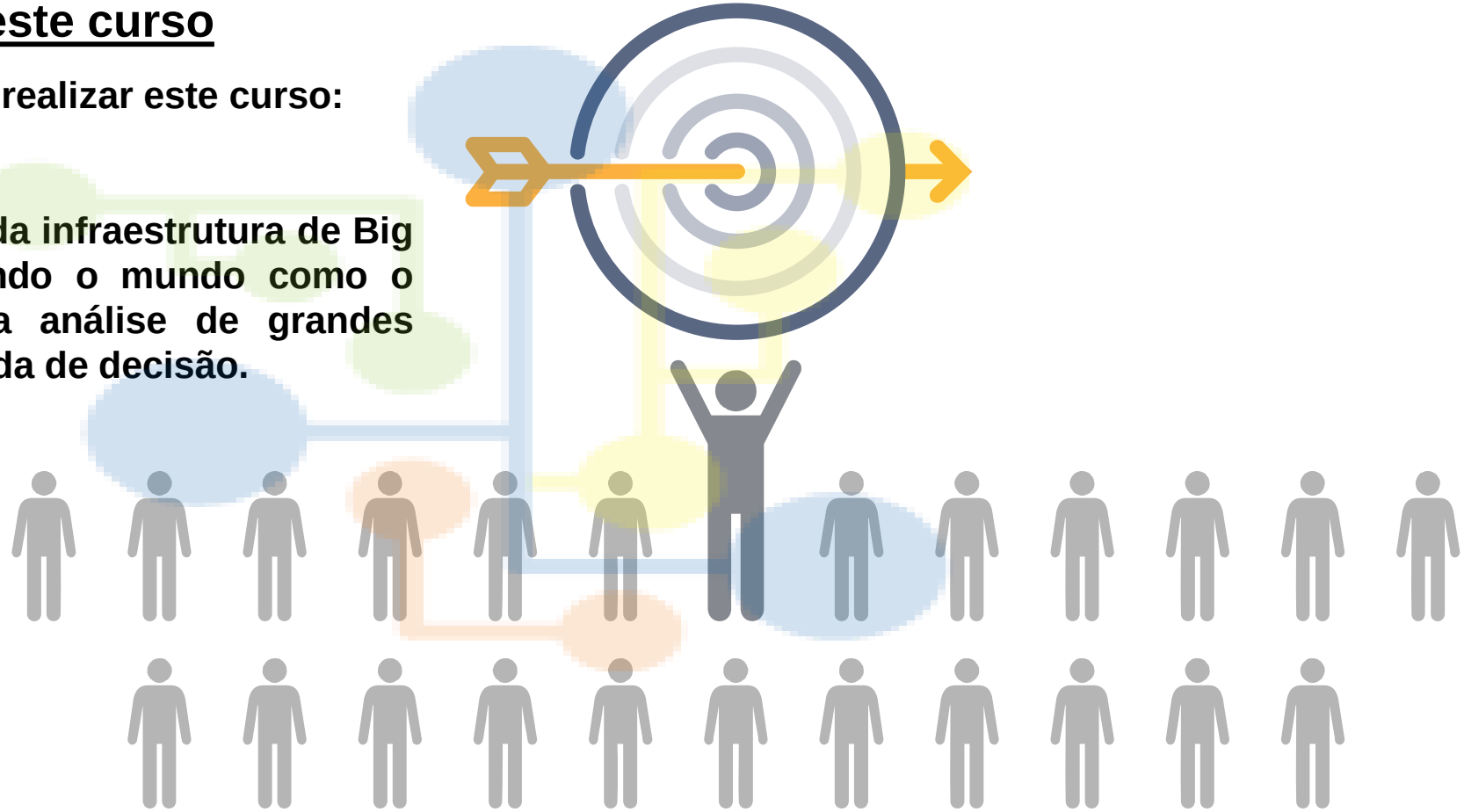


Engenharia de Dados com Hadoop e Spark

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.

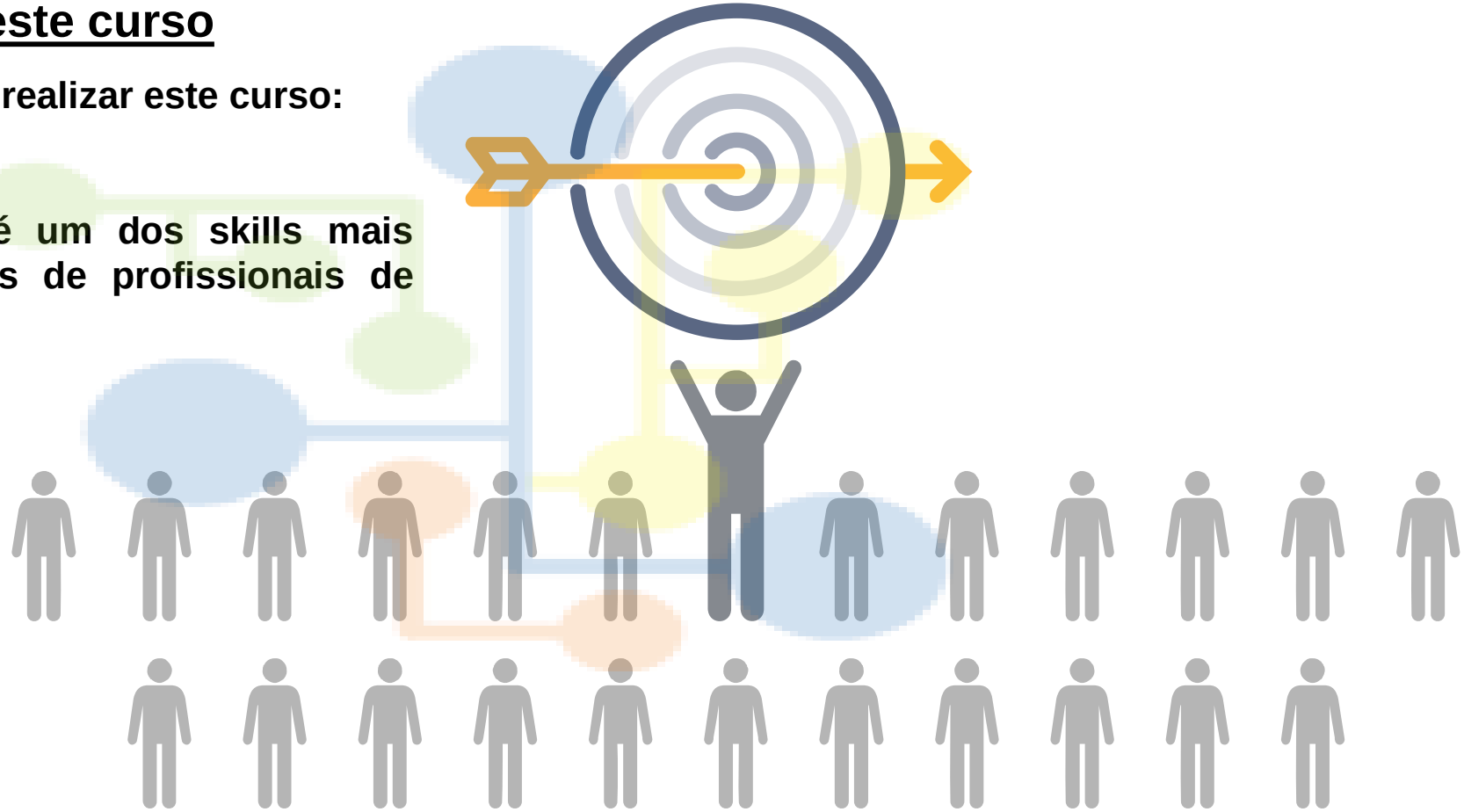


Engenharia de Dados com Hadoop e Spark

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.

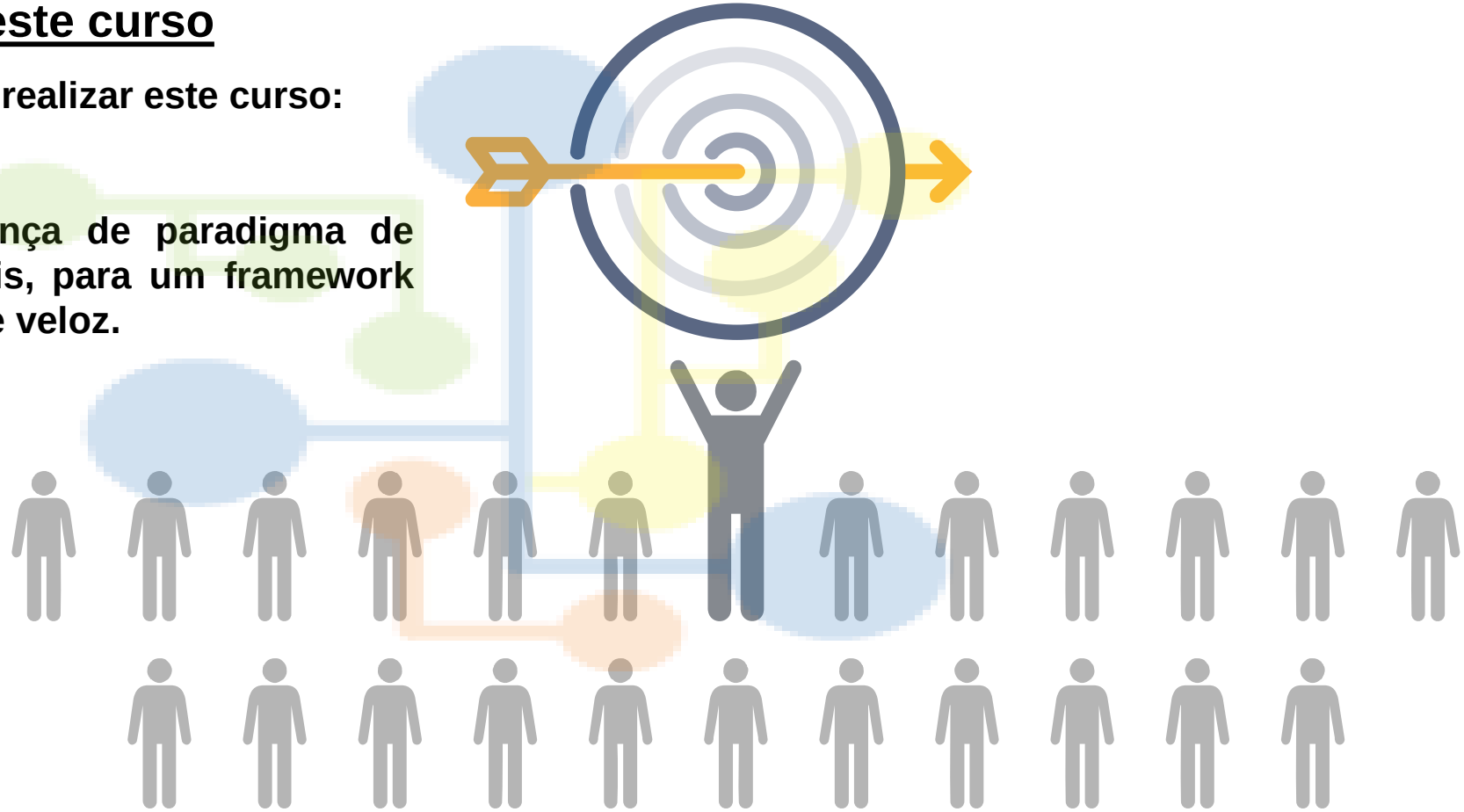


Engenharia de Dados com Hadoop e Spark

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

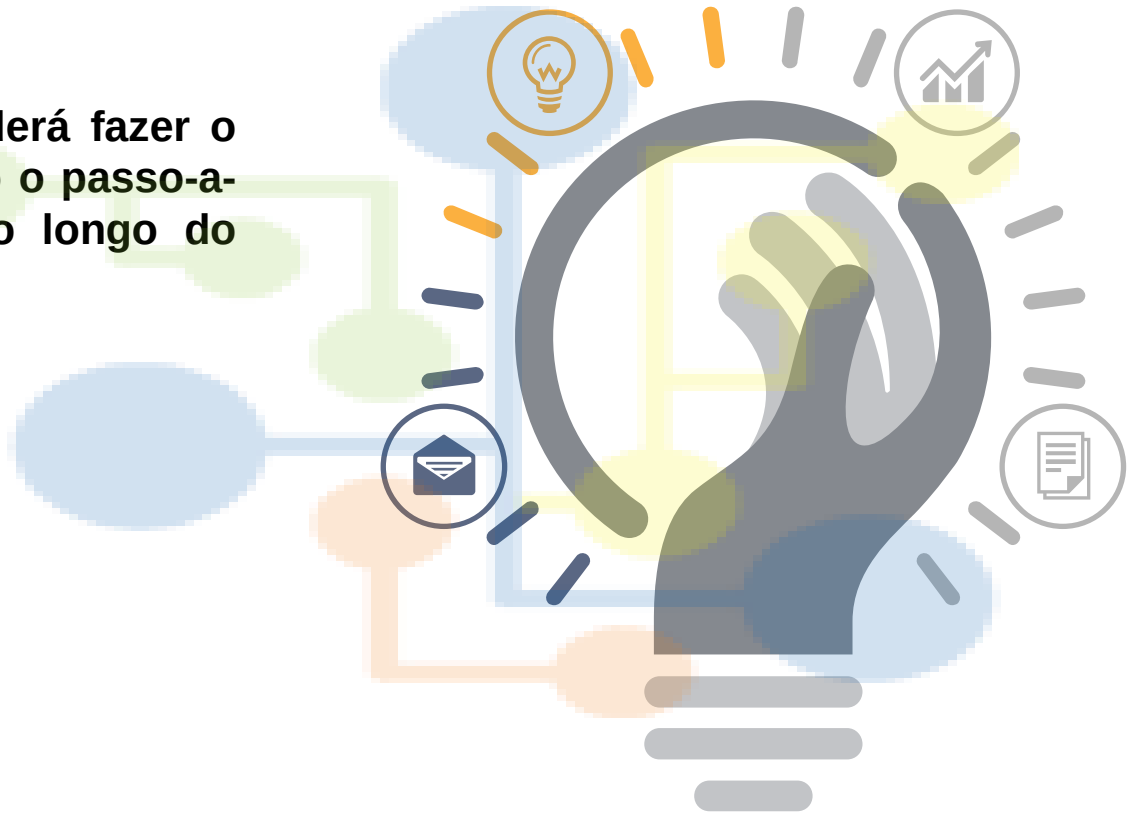
Estrutura do Curso

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.



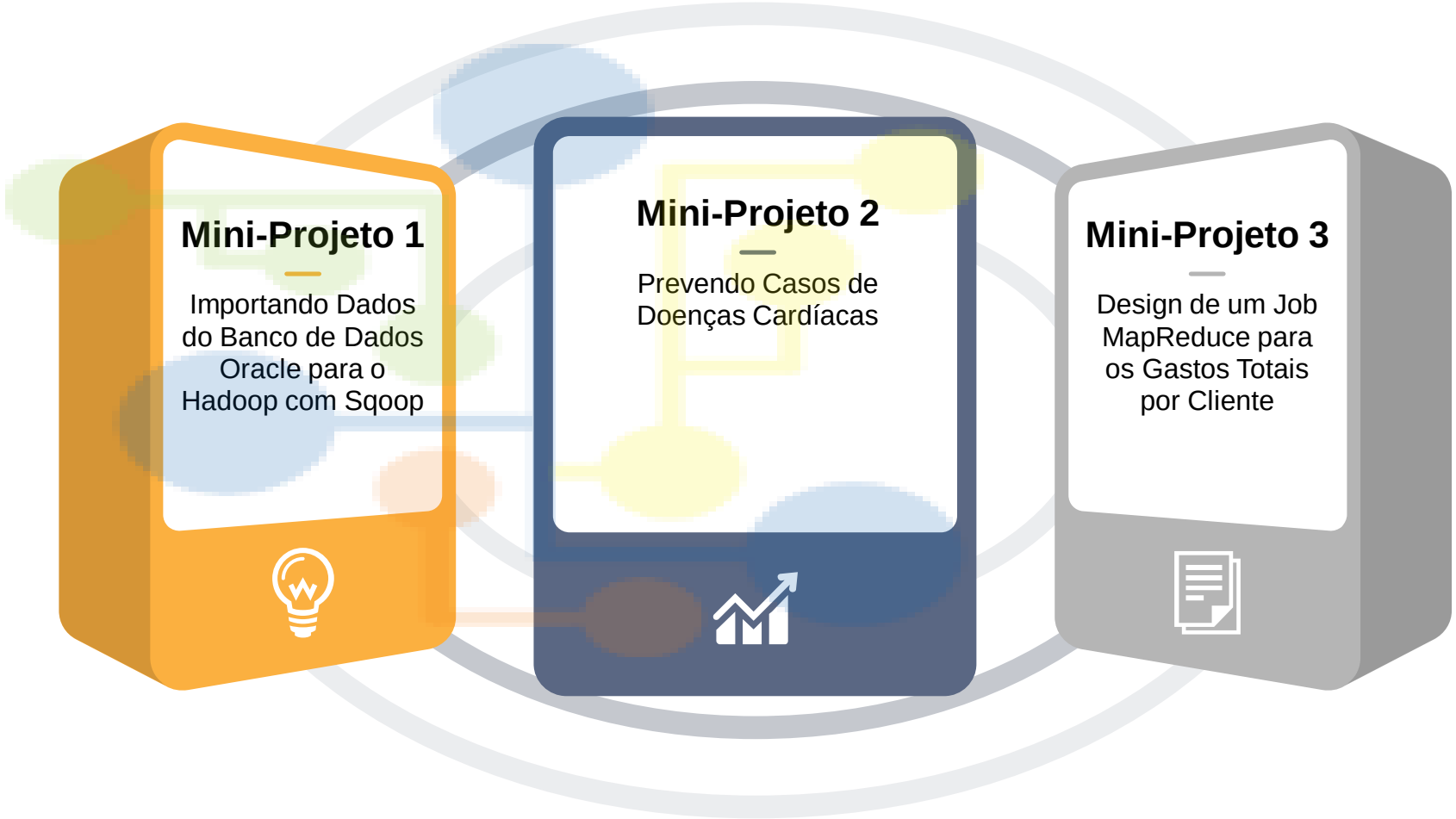
Engenharia de Dados com Hadoop e Spark

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

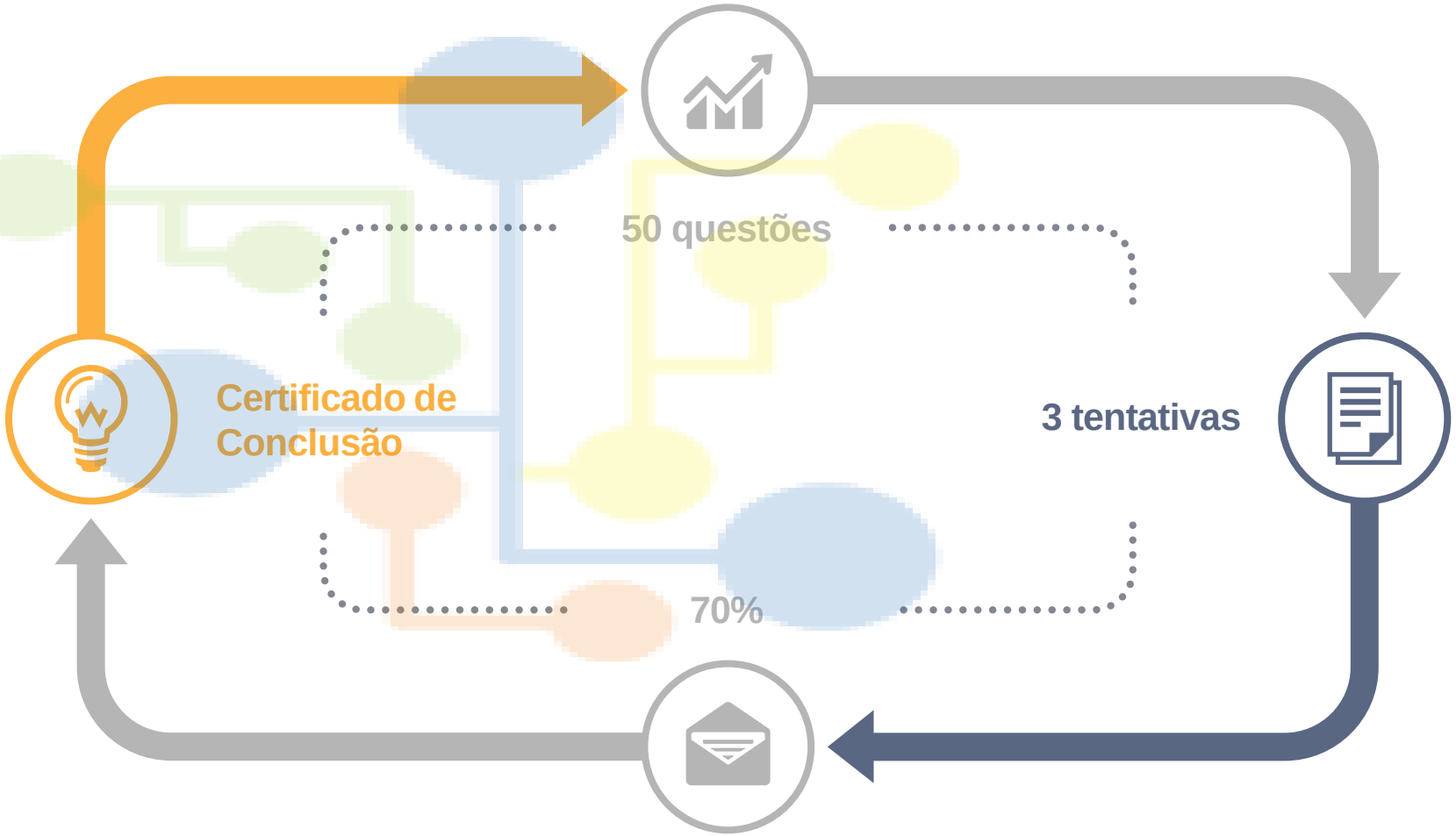


Engenharia de Dados com Hadoop e Spark

Projetos com Feedback

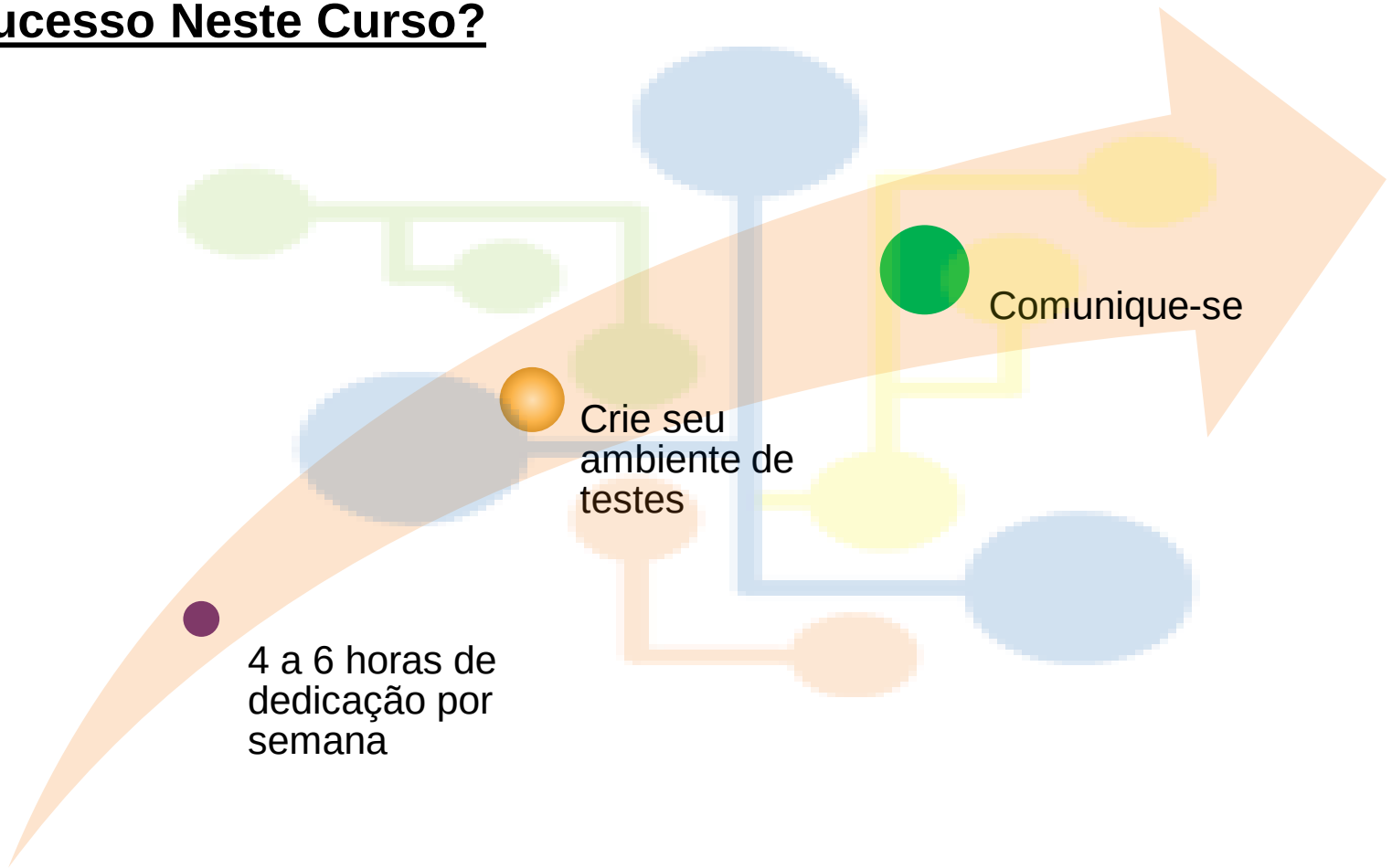
Engenharia de Dados com Hadoop e Spark

Avaliação Final



Engenharia de Dados com Hadoop e Spark

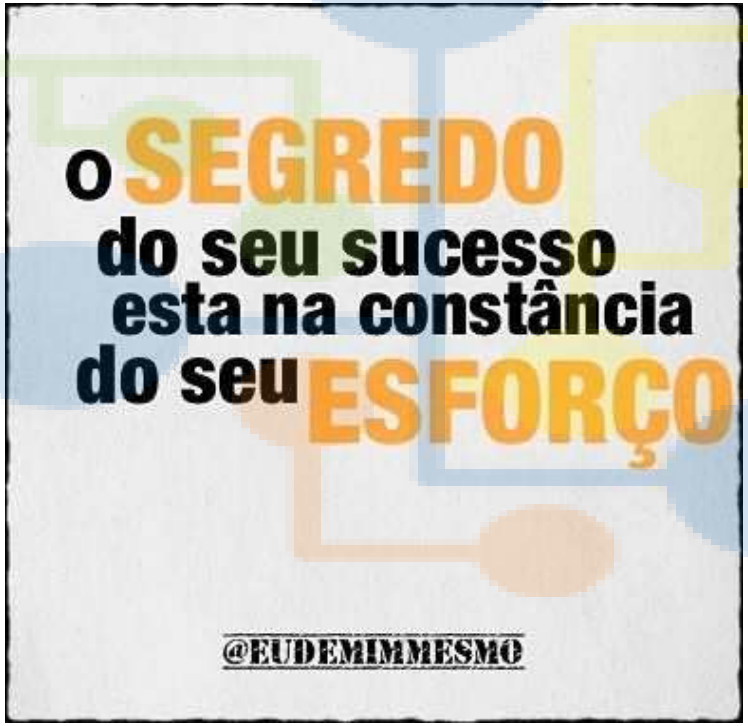
Como Obter Sucesso Neste Curso?



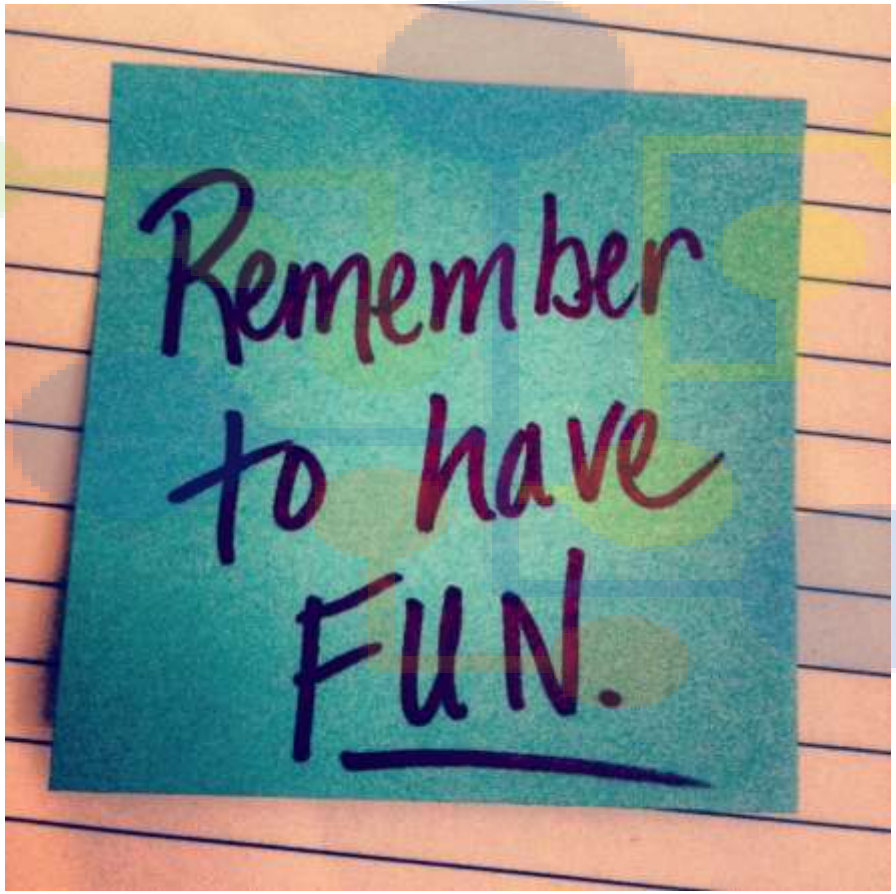
Engenharia de Dados com Hadoop e Spark

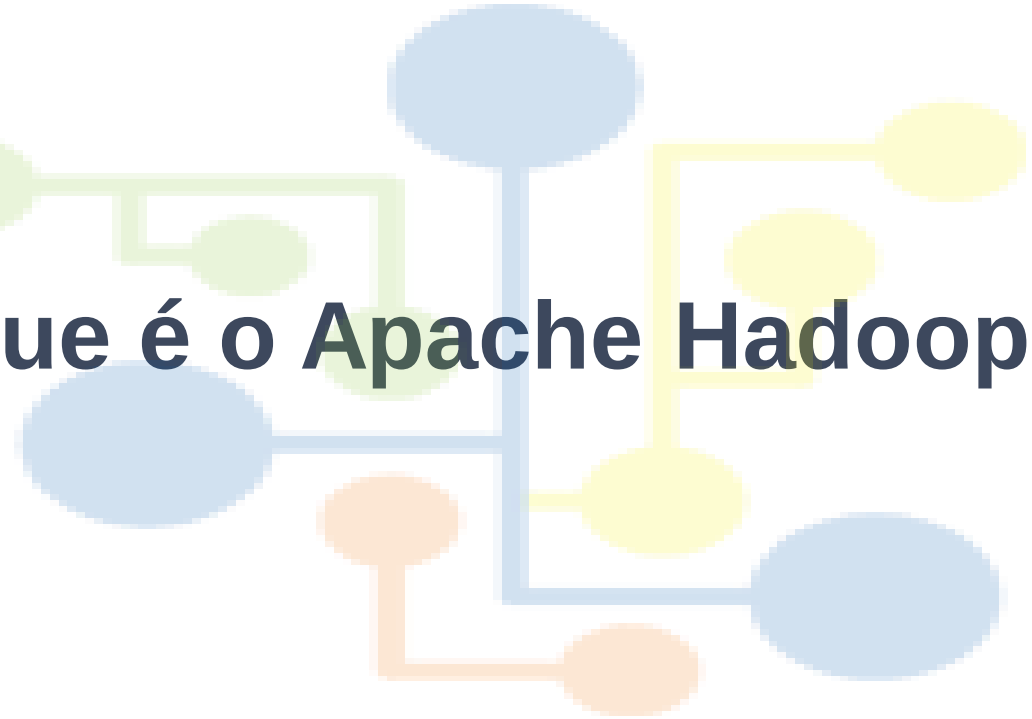
Engenharia de Dados com Hadoop e Spark

Engenharia de Dados com Hadoop e Spark



Engenharia de Dados com Hadoop e Spark





O que é o Apache Hadoop?

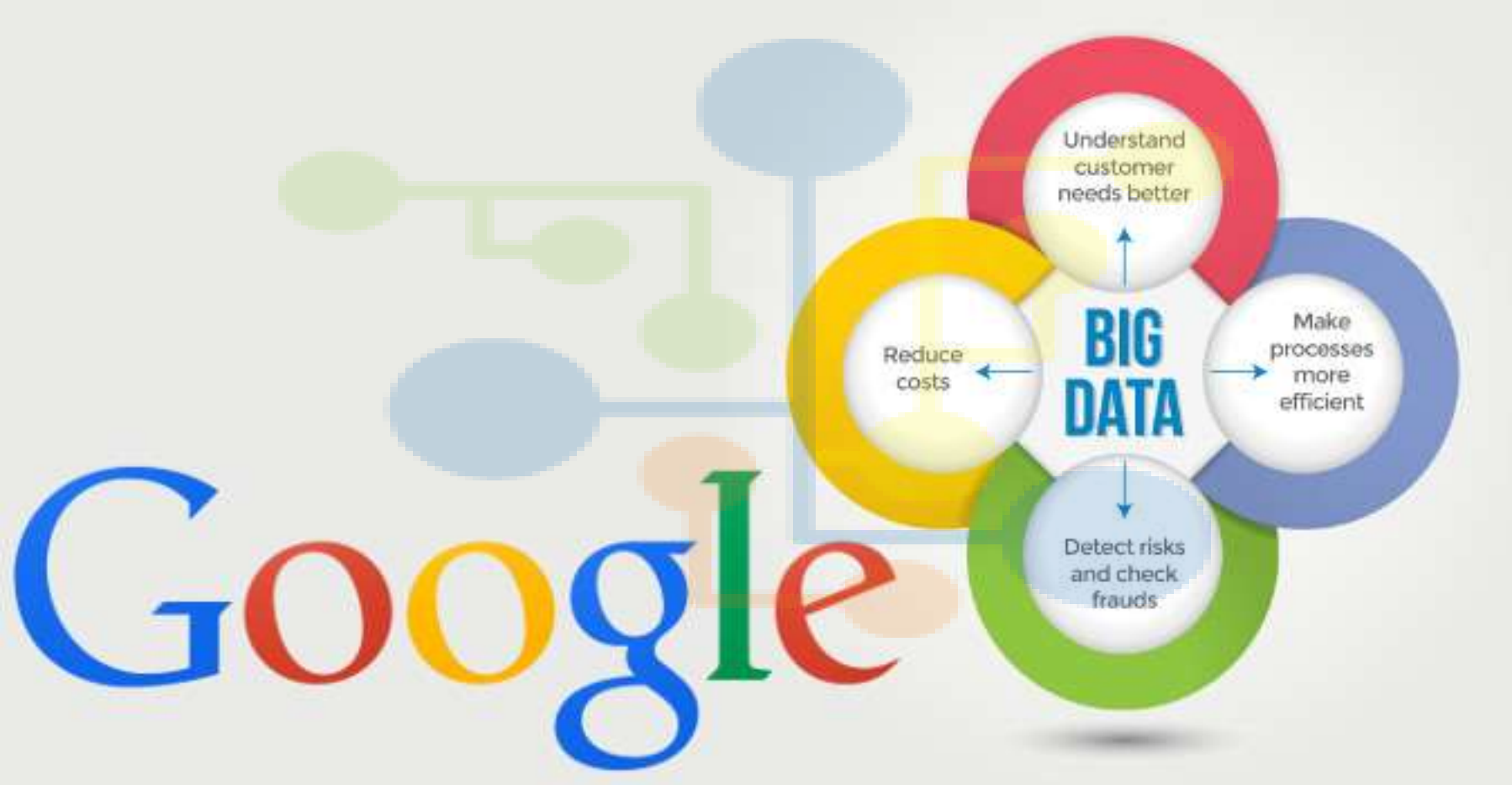
O que é o Apache Hadoop?

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.

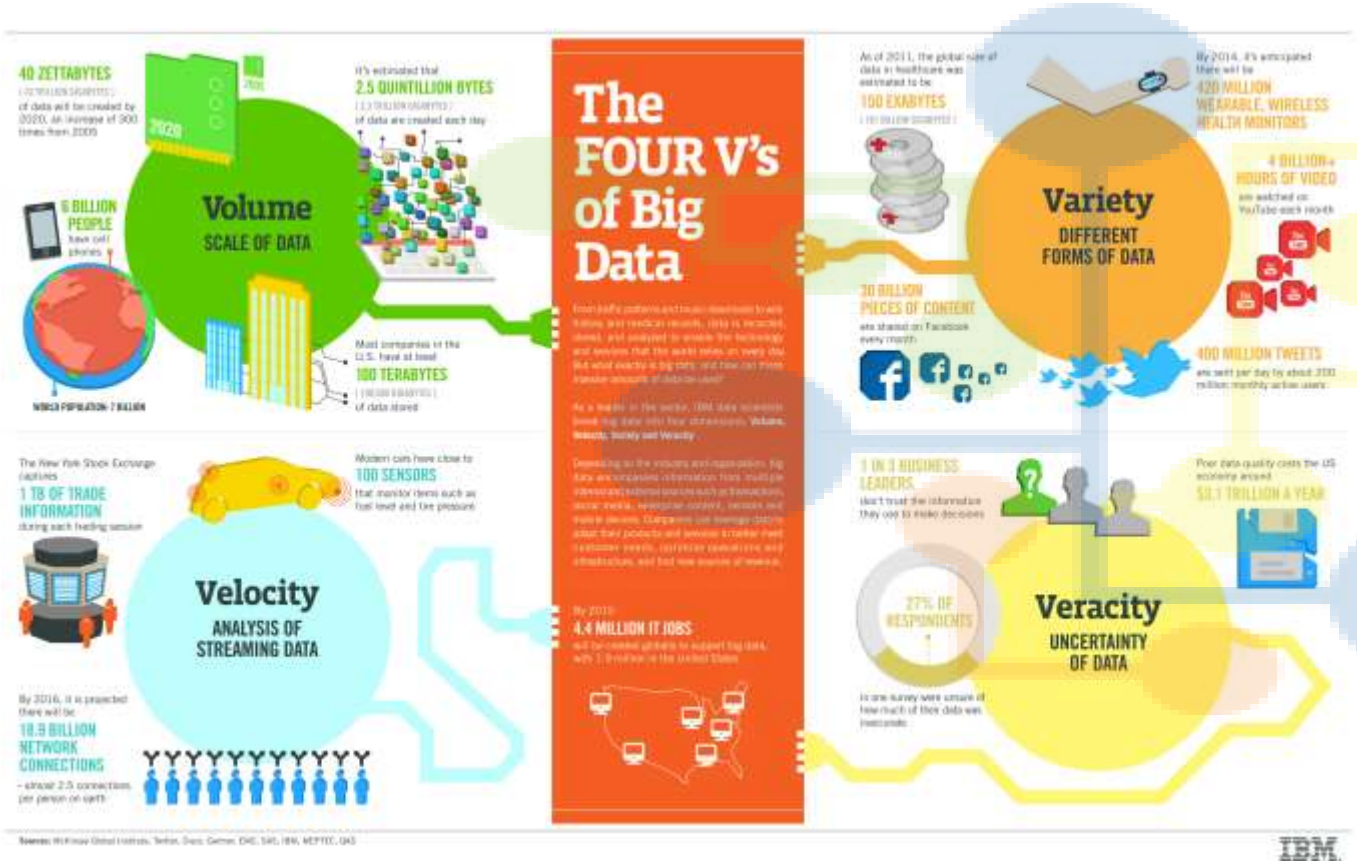
Sistemas corporativos, sistemas Web, mídias sociais, entre outros, produzem juntos um volume impressionante de dados, alcançando a dimensão de petabytes diários.

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

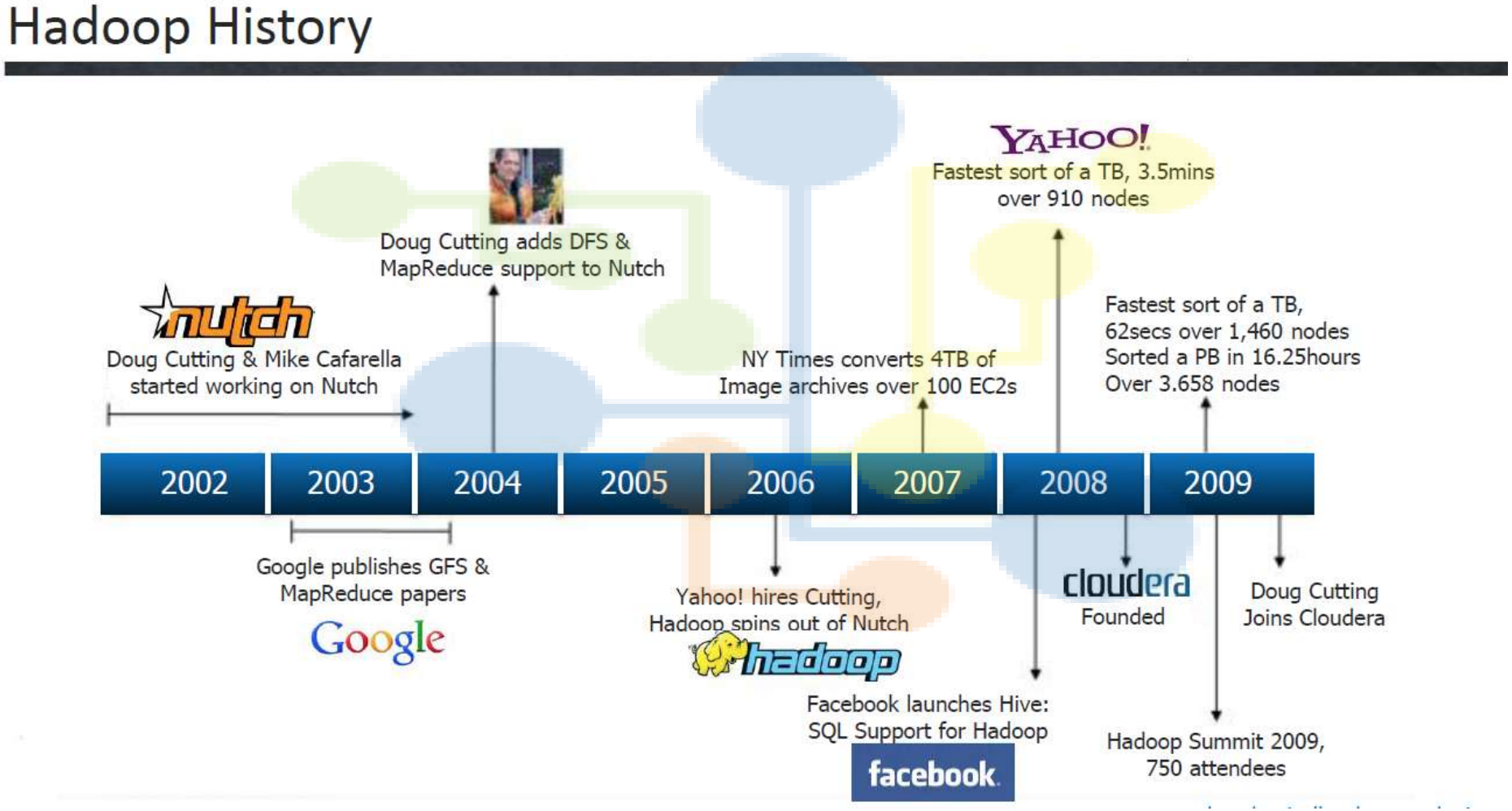
O que é o Apache Hadoop?

O que é o Apache Hadoop?

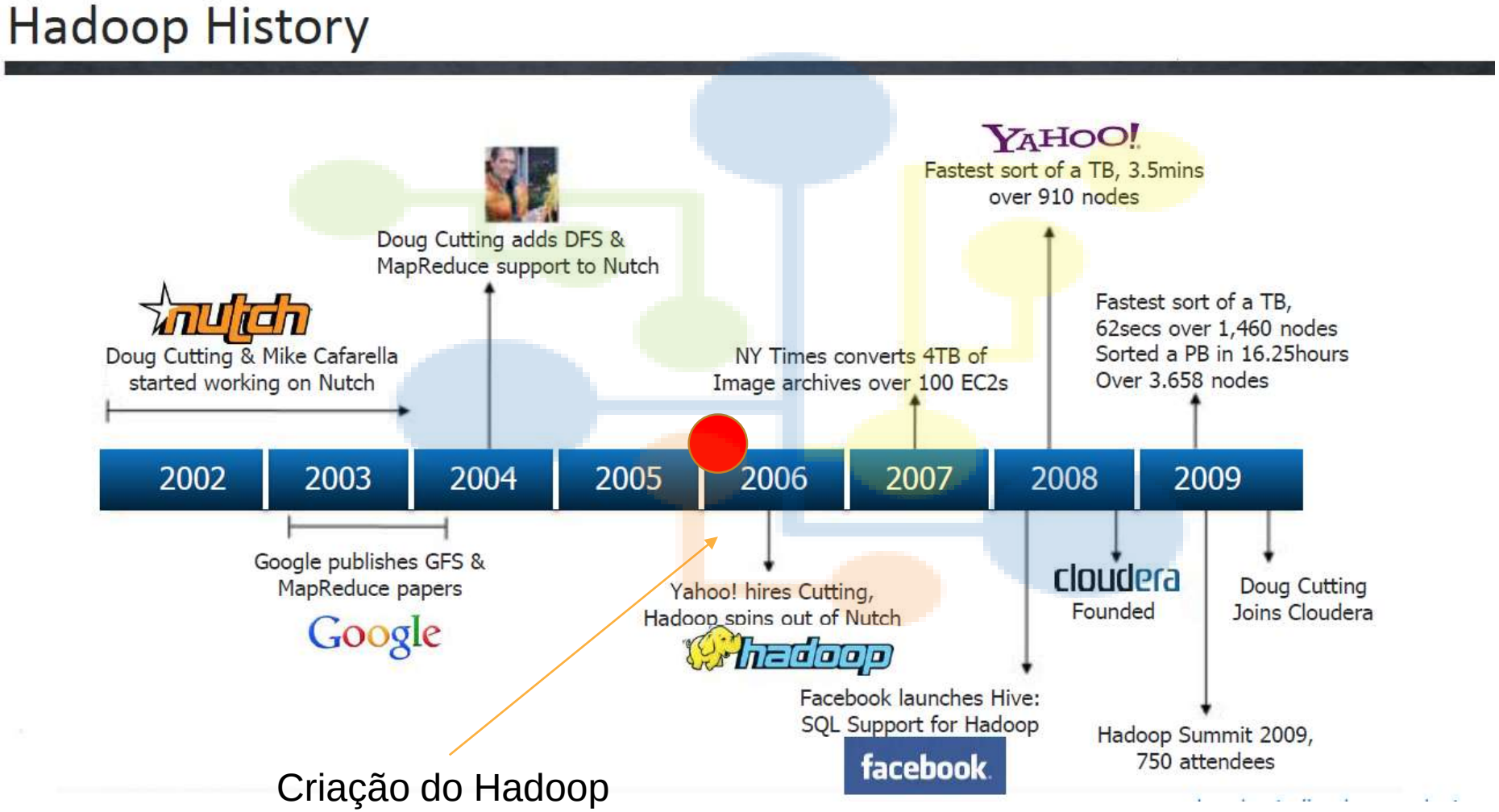
Computação Paralela

O que é o Apache Hadoop?

Uma Breve História do Apache Hadoop

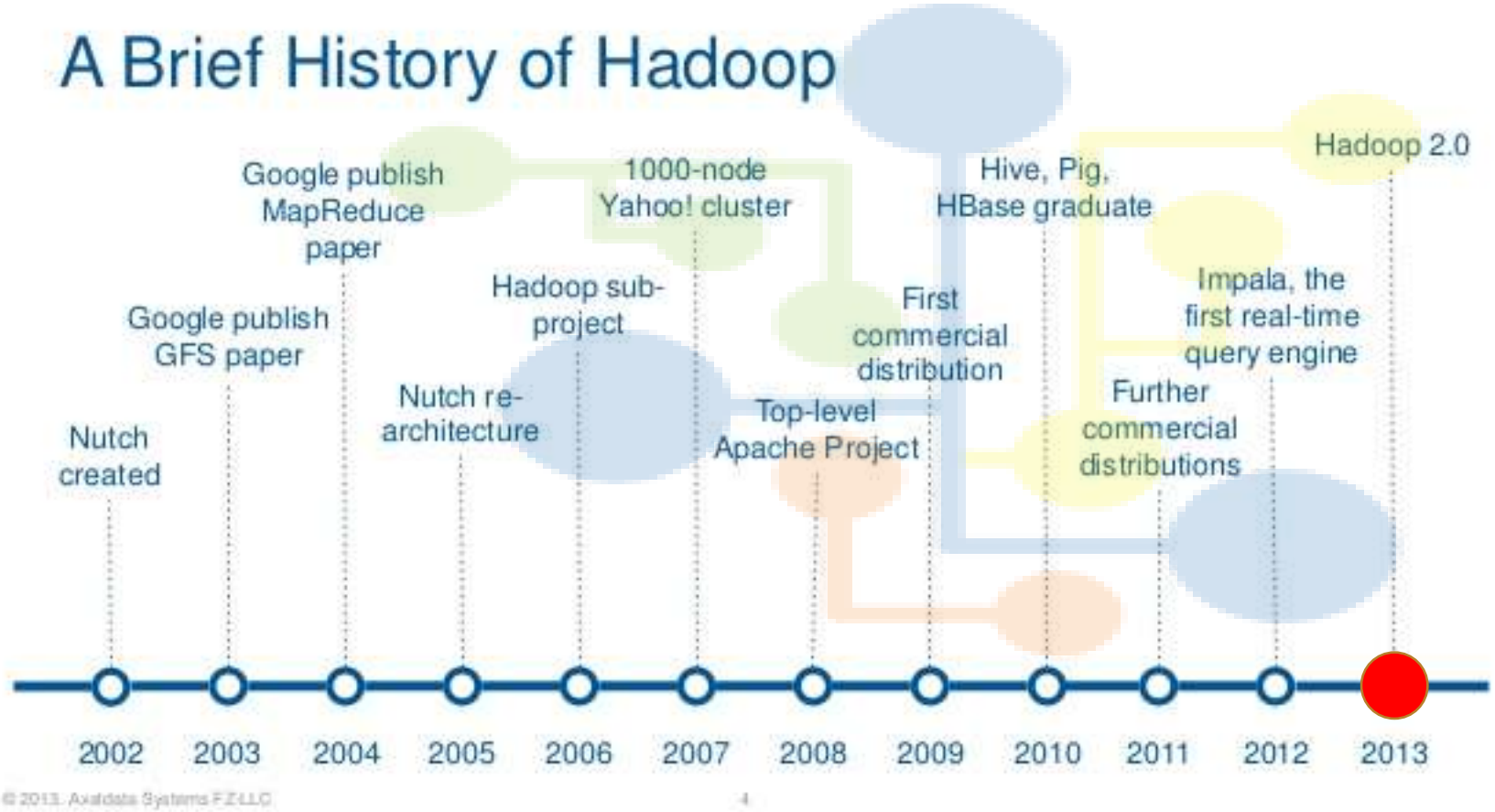


Uma Breve História do Apache Hadoop



Criação do Hadoop

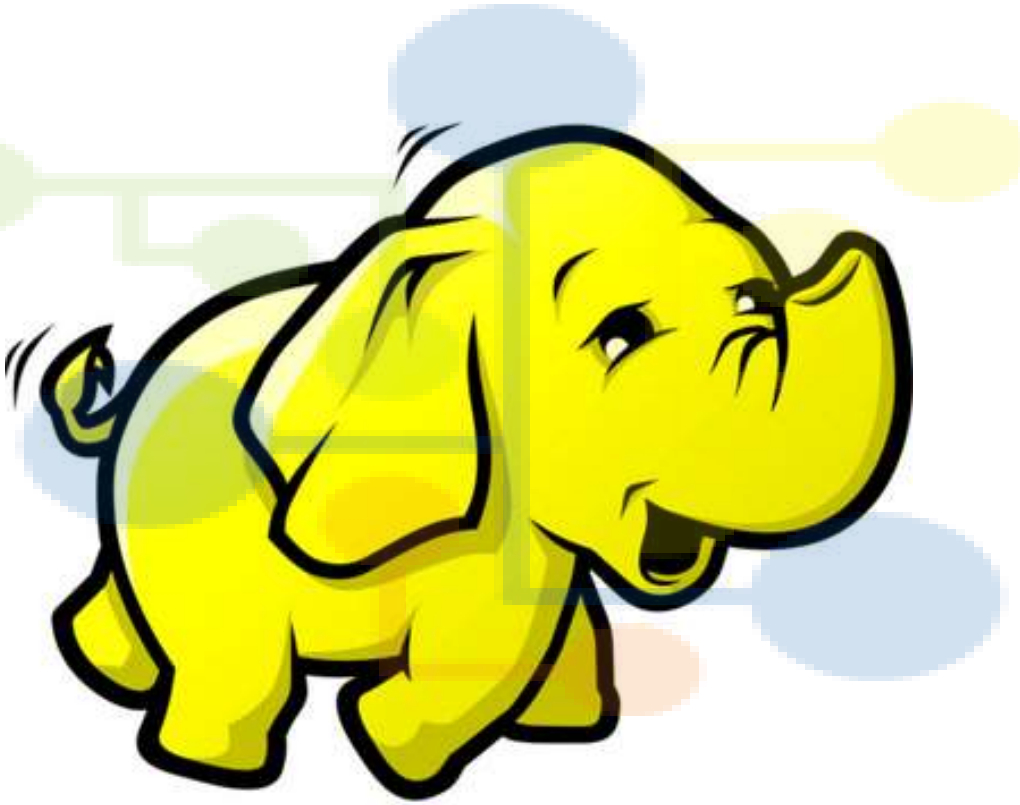
Uma Breve História do Apache Hadoop



2018
Versão 3.0

2019
Versão 3.2.x

O que é o Hadoop?



Quais os benefícios para as Empresas ao utilizar o Hadoop?

Benefícios do Hadoop

Benefícios do Hadoop

Open Source

Benefícios do Hadoop

Economia

Benefícios do Hadoop

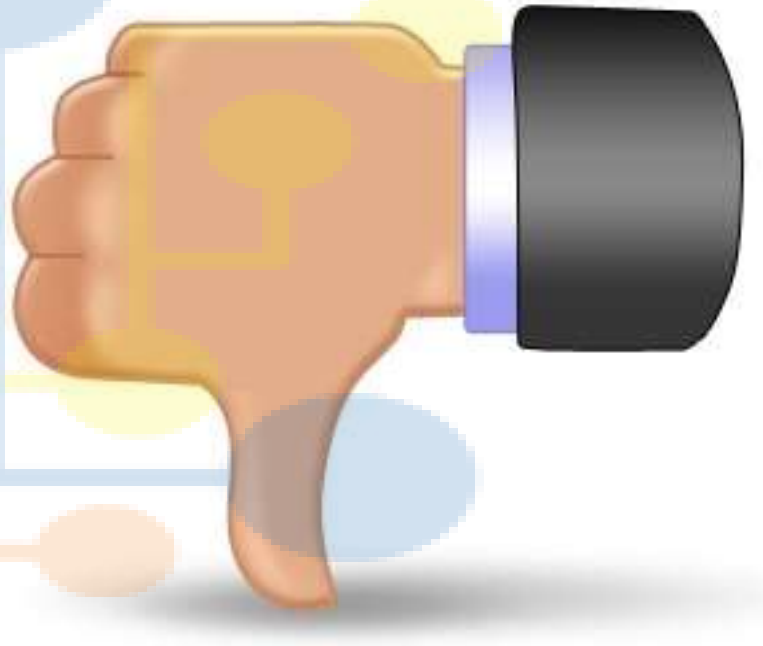
Escalabilidade

Benefícios do Hadoop

Robustez

Desvantagens do Hadoop

Desvantagens



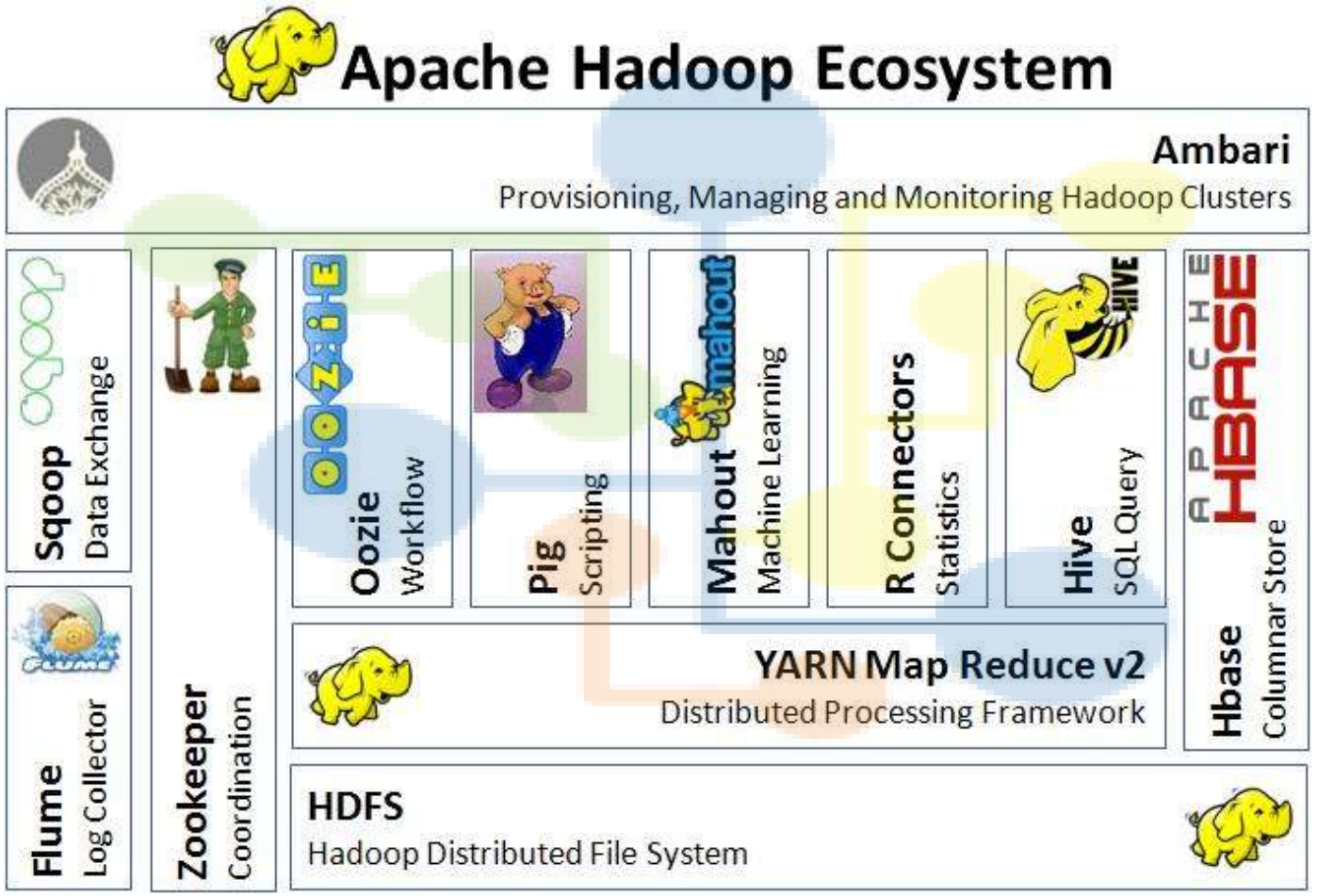
Desvantagens do Hadoop

Desvantagens

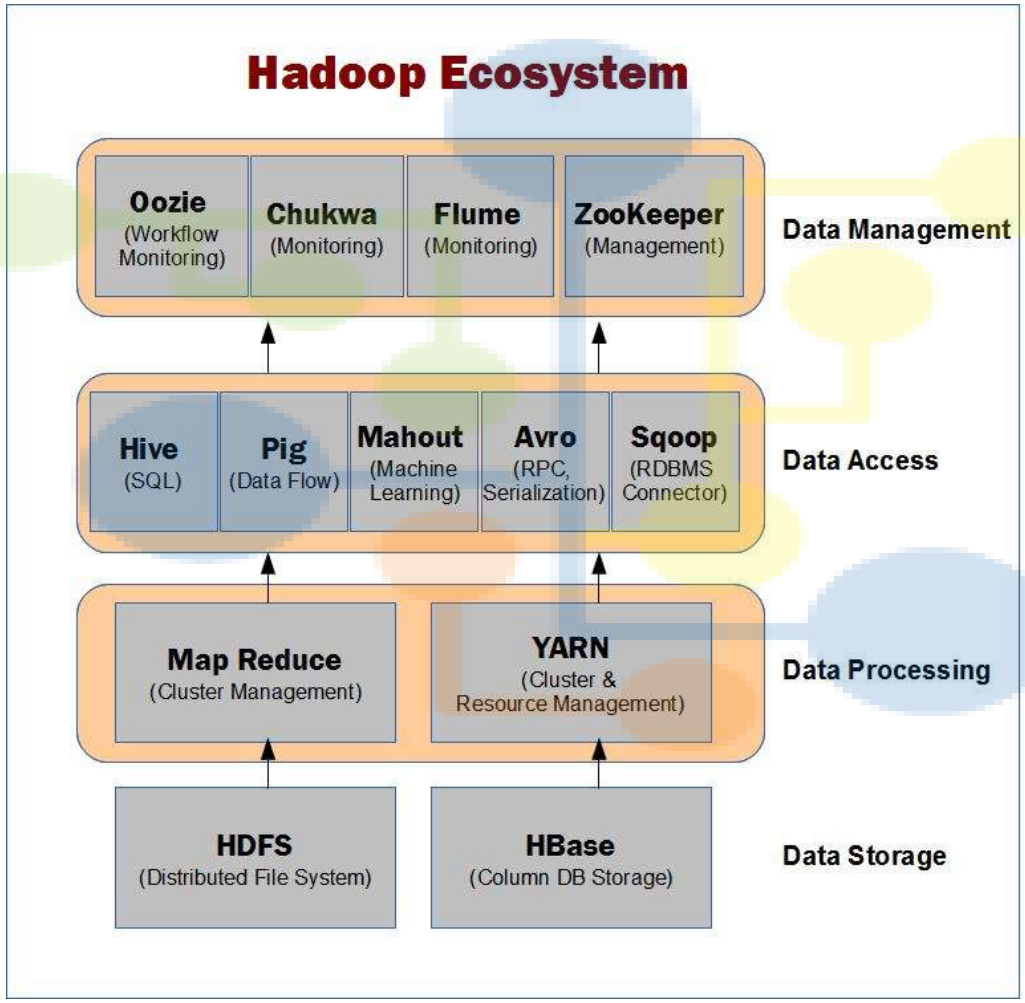
- Node Master Único
- Processamento de Arquivos Pequenos
- Muito Processamento em Poucos Dados

Ecossistema Hadoop

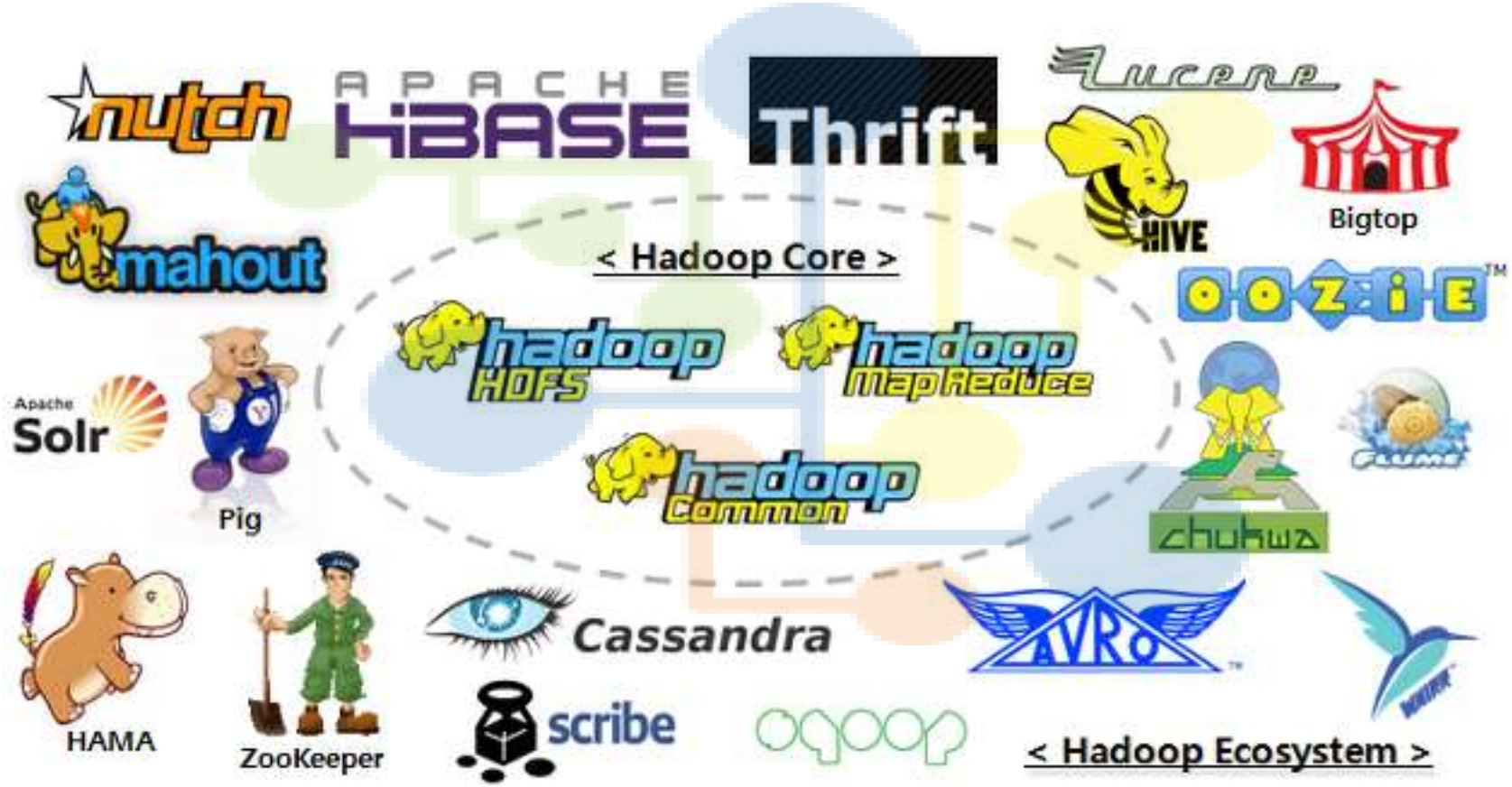
Ecossistema Hadoop



Ecossistema Hadoop



Ecossistema Hadoop

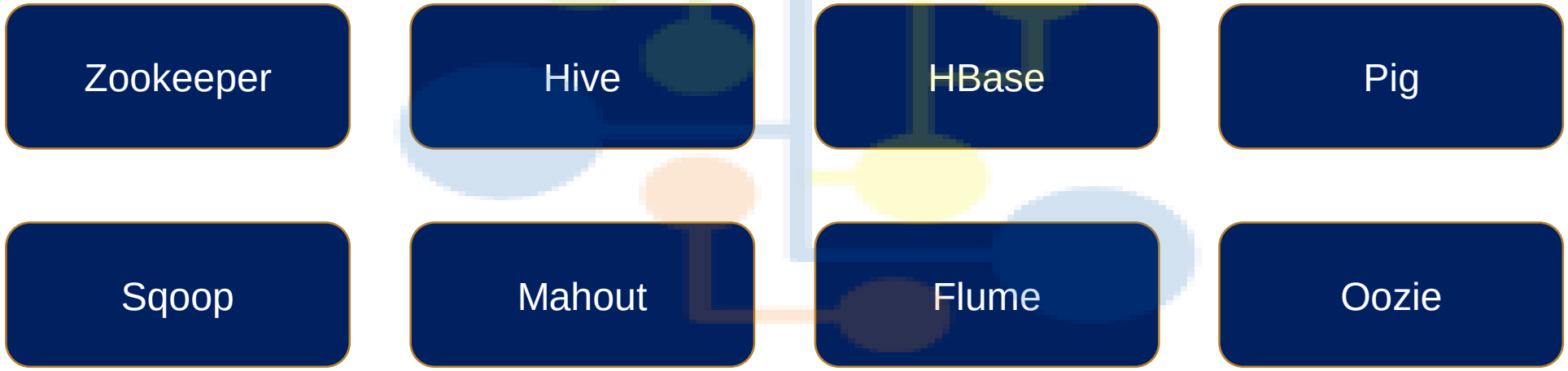


Ecossistema Hadoop

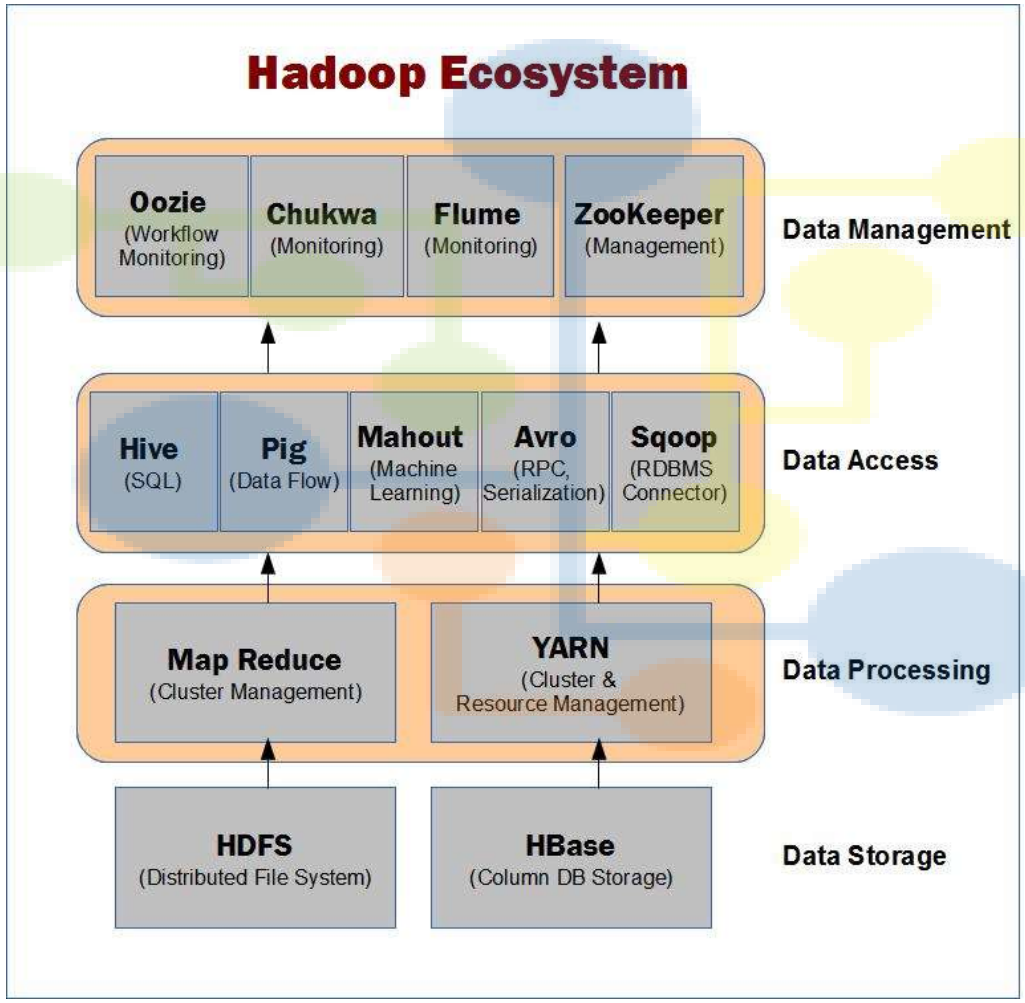
Projetos Principais do Ecossistema Hadoop

Ecossistema Hadoop

Outros Projetos Importantes

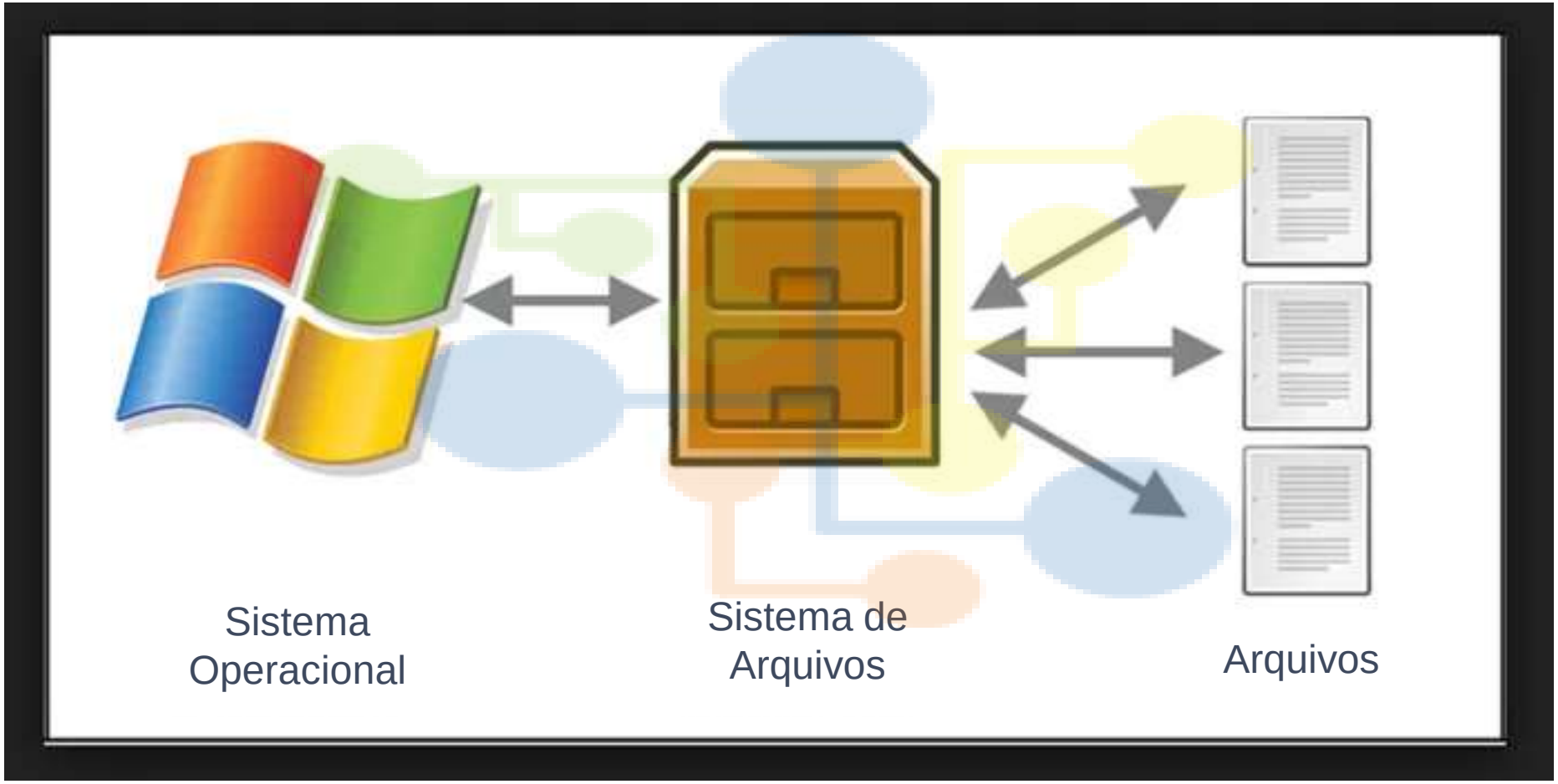


Ecossistema Hadoop



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
`ext2`	Sistema de arquivos padrão do Linux
`ext3`	Sistema de arquivos ext2 melhorado
`reiserfs`	Sistema de arquivos do tipo Journaling
`msdos`	Sistema de aruqivos FAT da Microsoft DOS
`vfat`	Sistema de arquivos FAT-32 do Microsoft Windows
`iso9660`	Sistema de arquivos do CD-ROM
`nfs`	Network File System. Usado par a montar dispositivos em computadores remotos.
`swap`	Sistema de arquivos de troca utilizando para memória virtual.
`proc`	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 `/proc`.

HDFS – Conceito e Importância

HDFS – Conceito e Importância

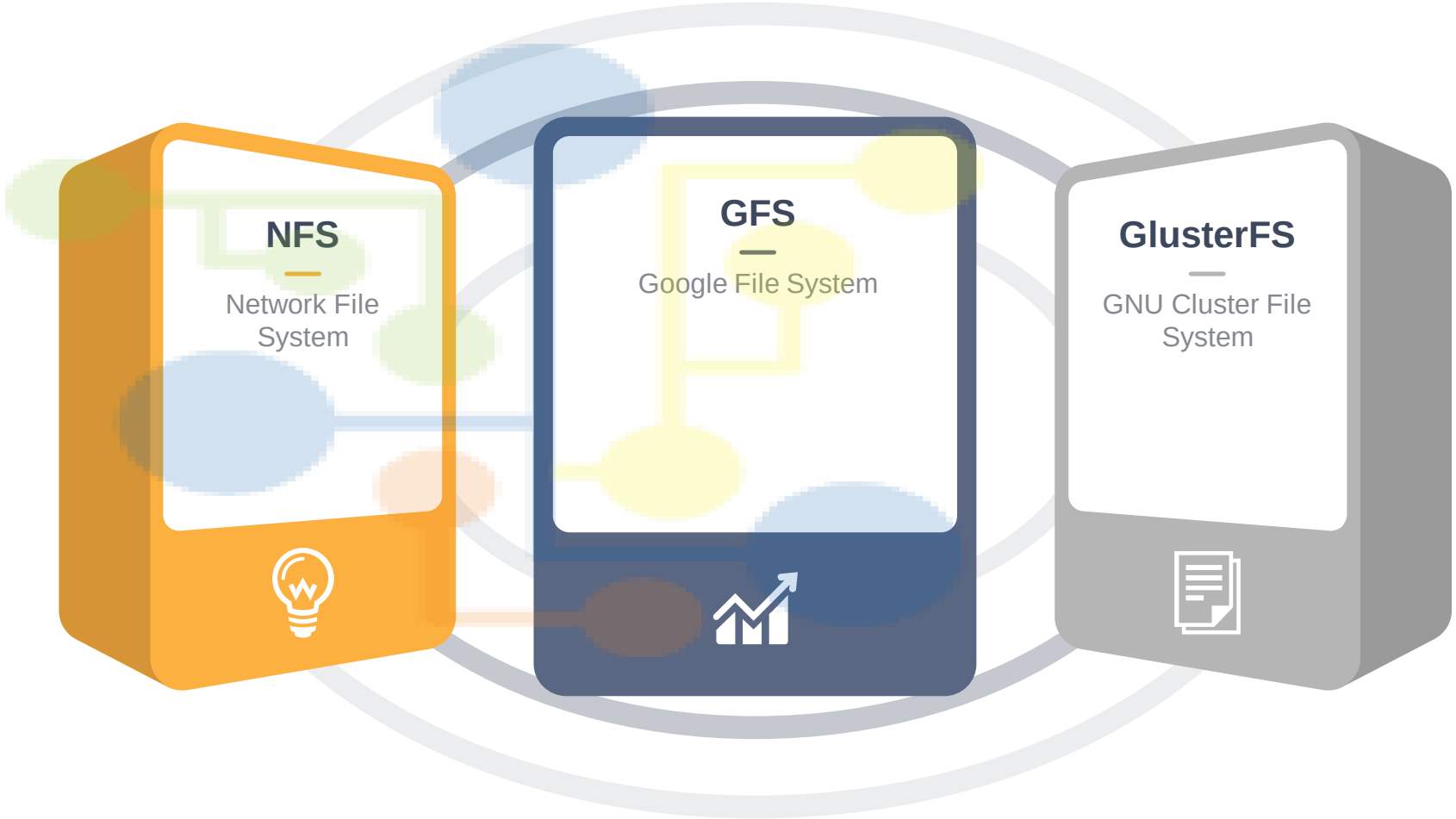
Sistema de Arquivos Distribuído

HDFS – Conceito e Importância

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

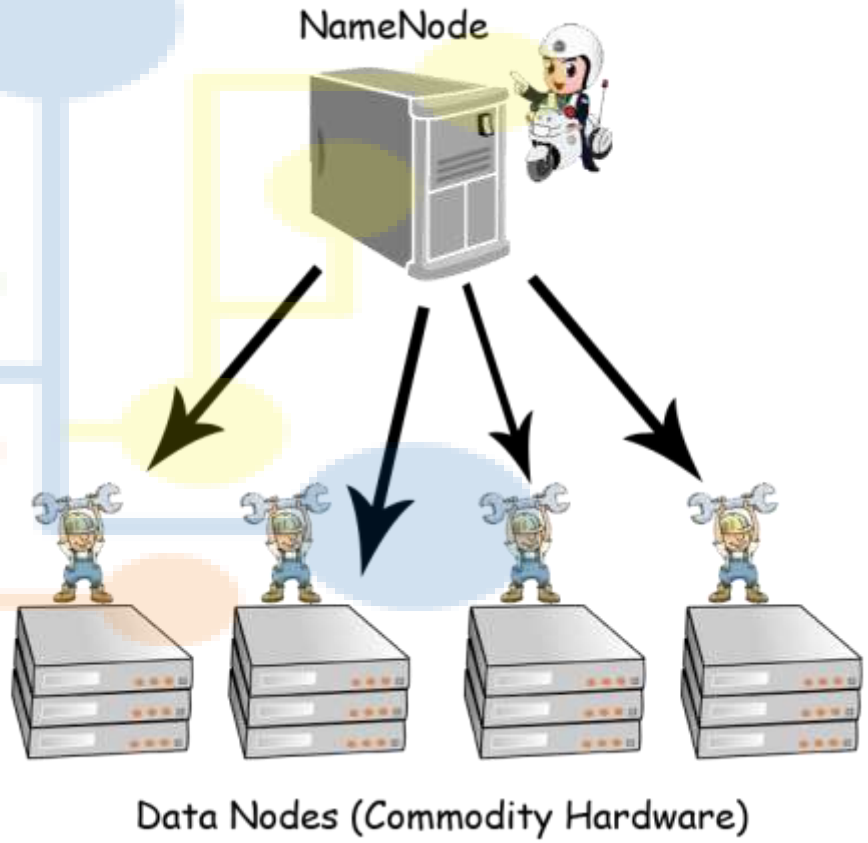
HDFS – Conceito e Importância

Outros Sistemas de Arquivos Distribuídos



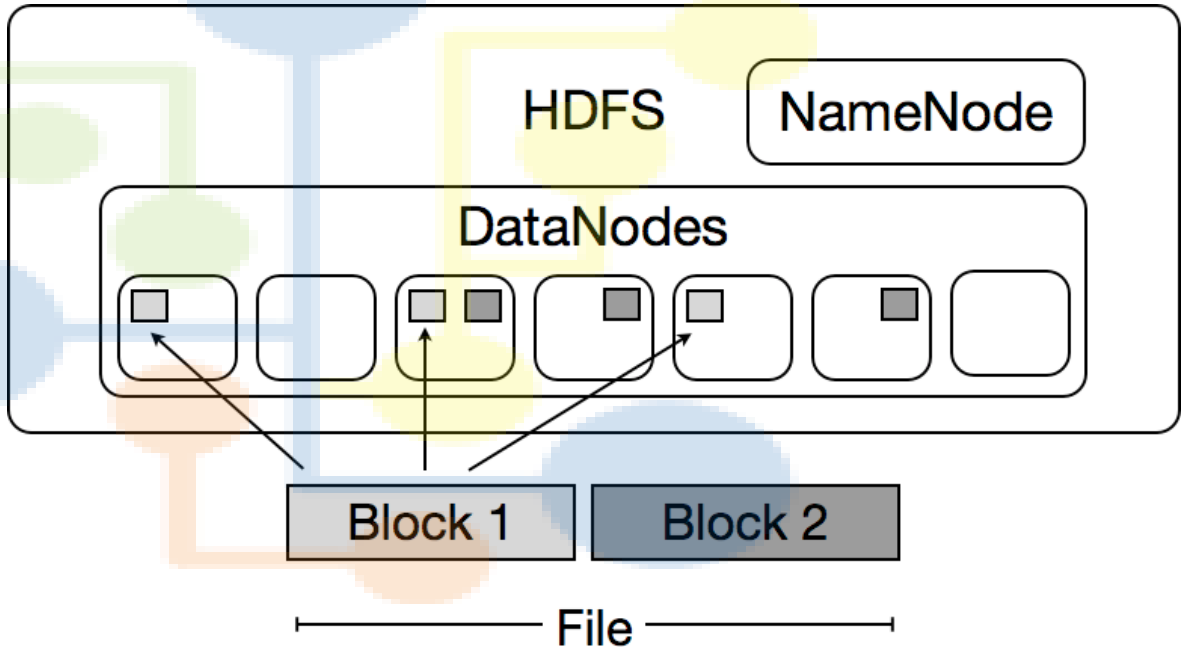
HDFS – Conceito e Importância

Hadoop Distributed File System



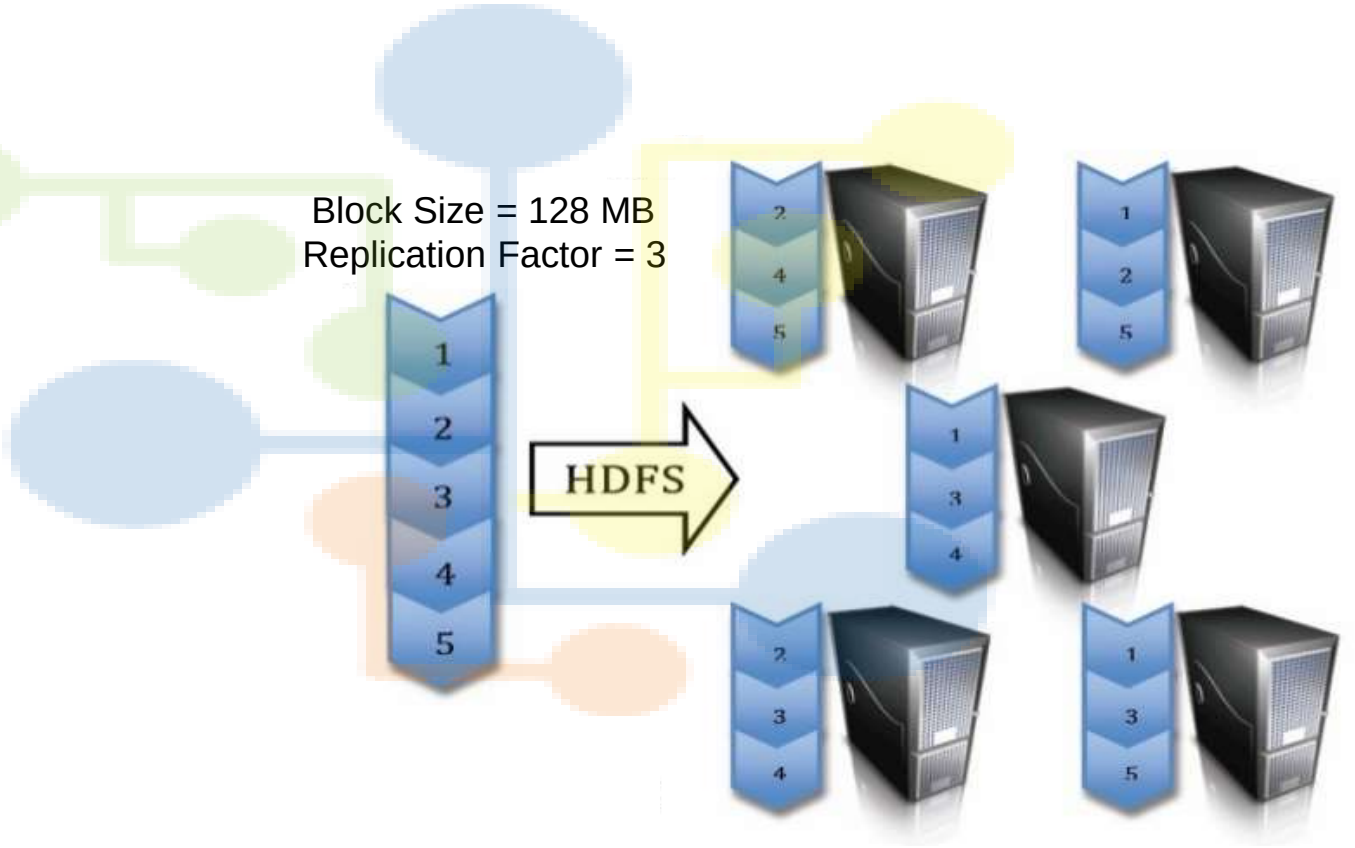
HDFS – Conceito e Importância

Hadoop Distributed File System



HDFS – Conceito e Importância

Hadoop Distributed File System



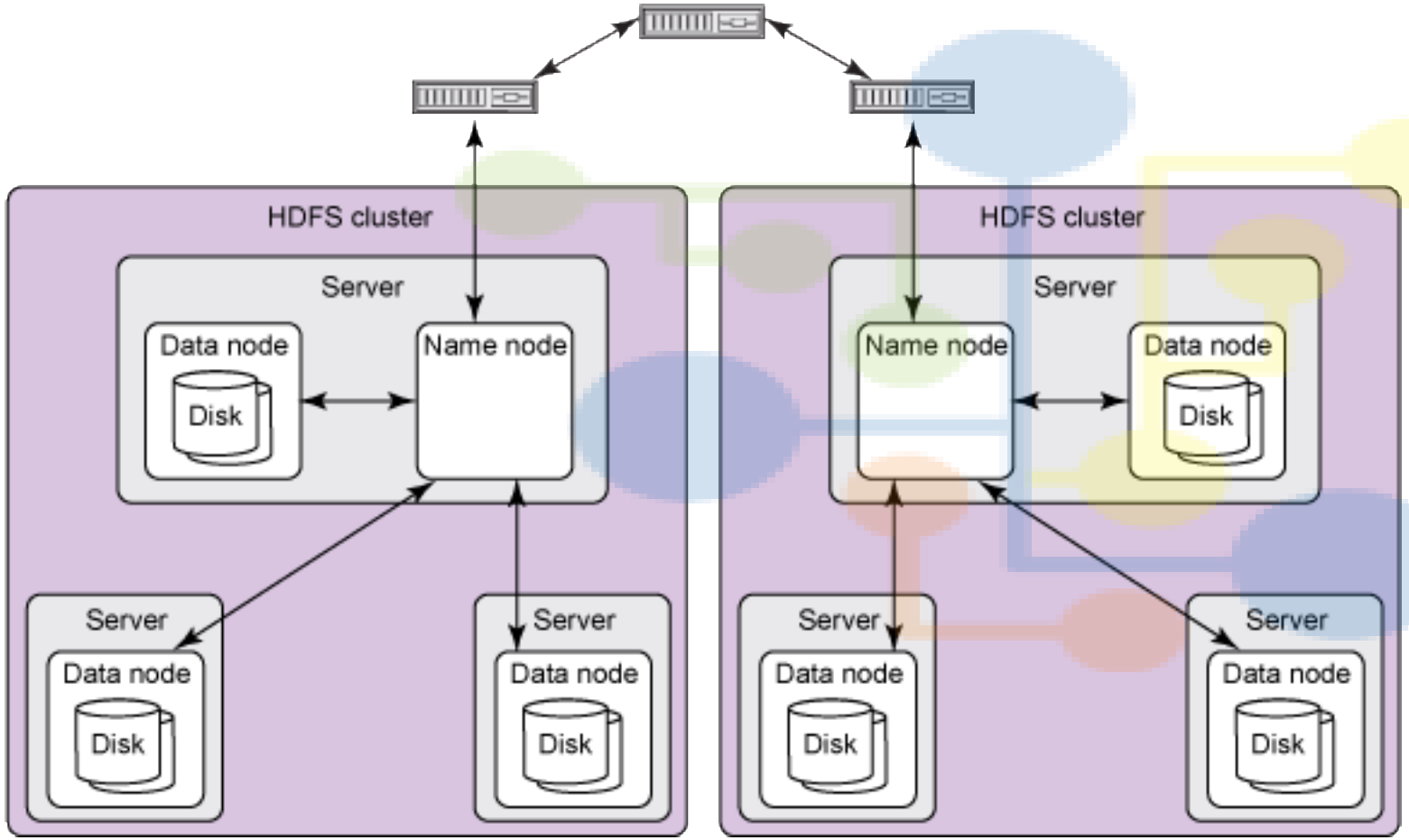
HDFS – Conceito e Importância

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.

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

HDFS (Hadoop Distributed File System)
Arquitetura

HDFS – Arquitetura

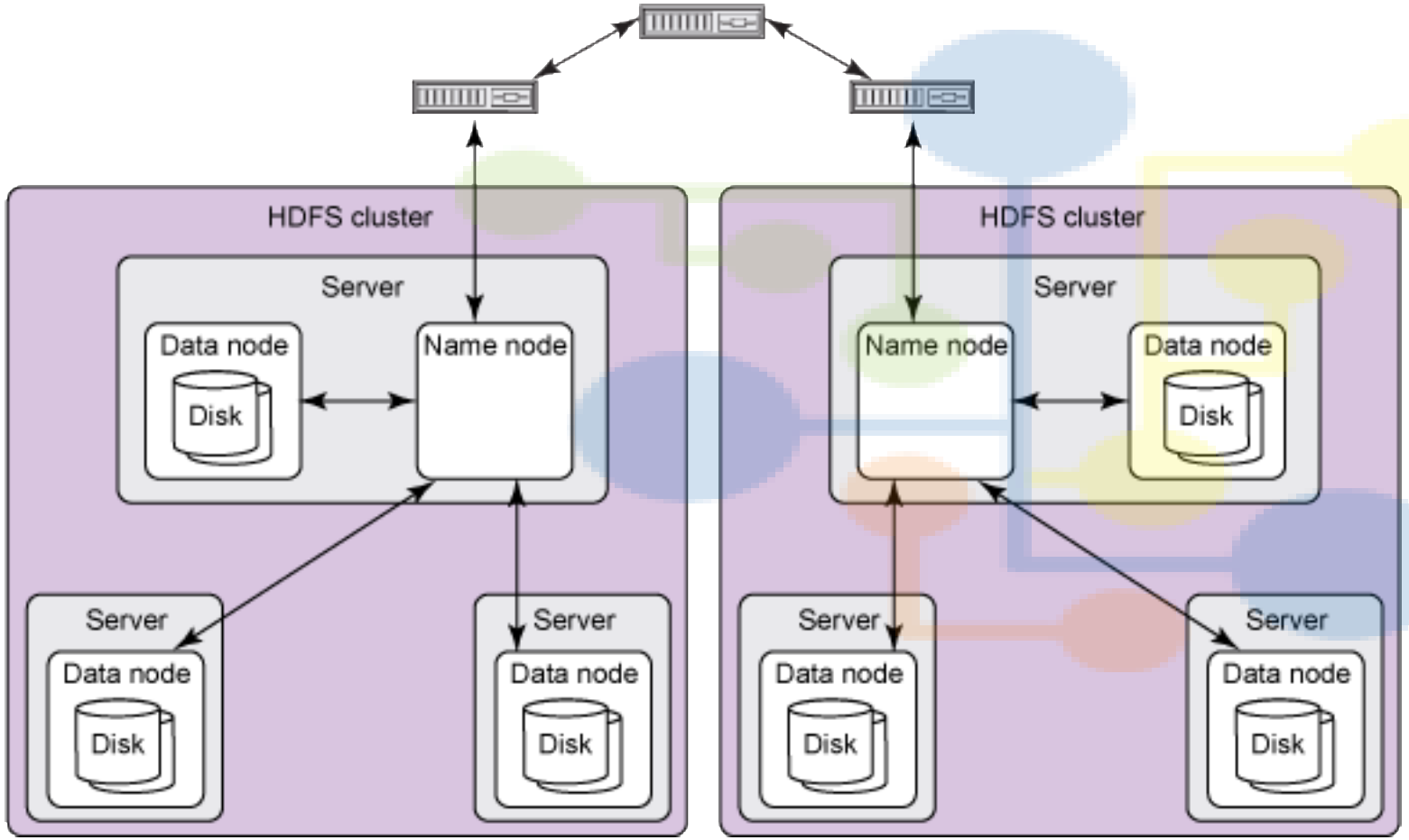


Arquitetura Master/Worker

HDFS – Arquitetura

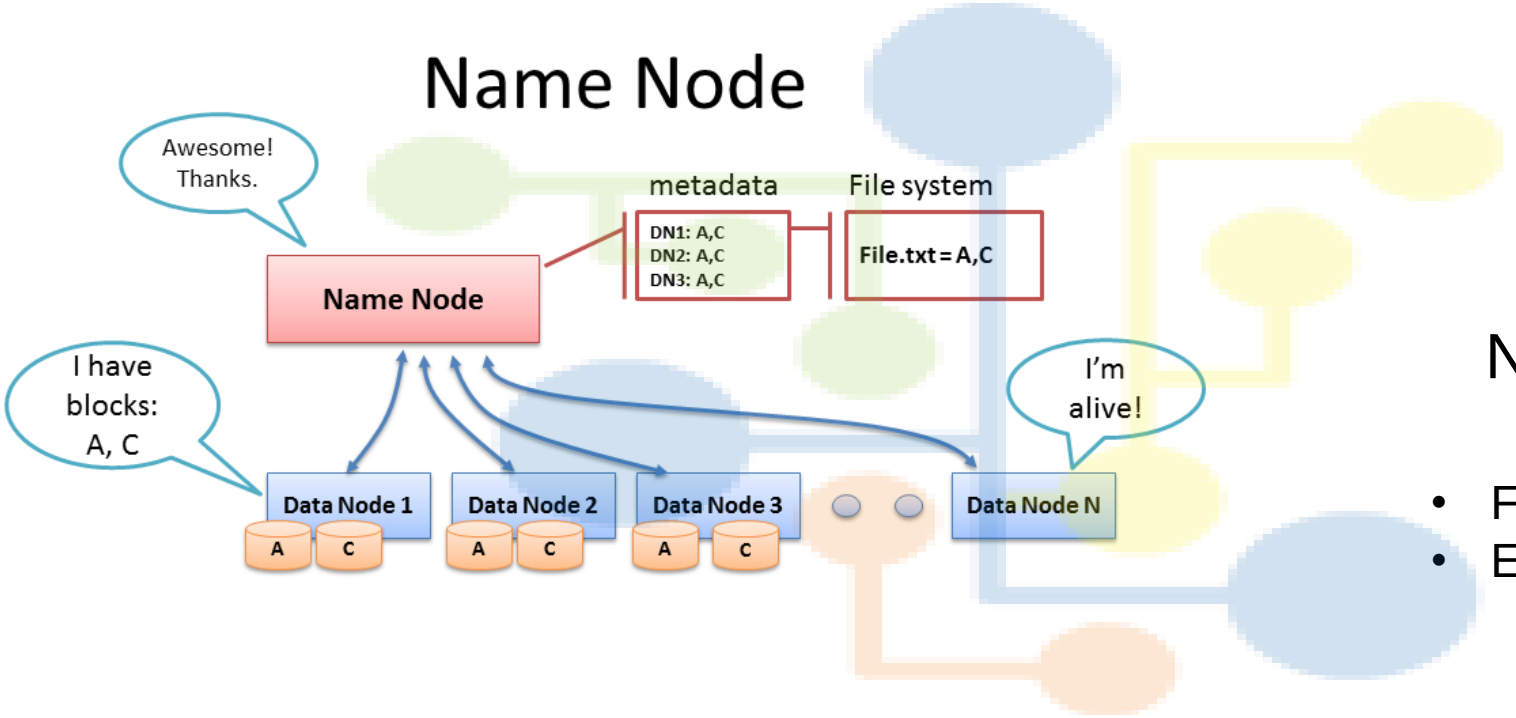
Arquitetura
Master/Worker

HDFS – Arquitetura



Arquitetura Master/Worker

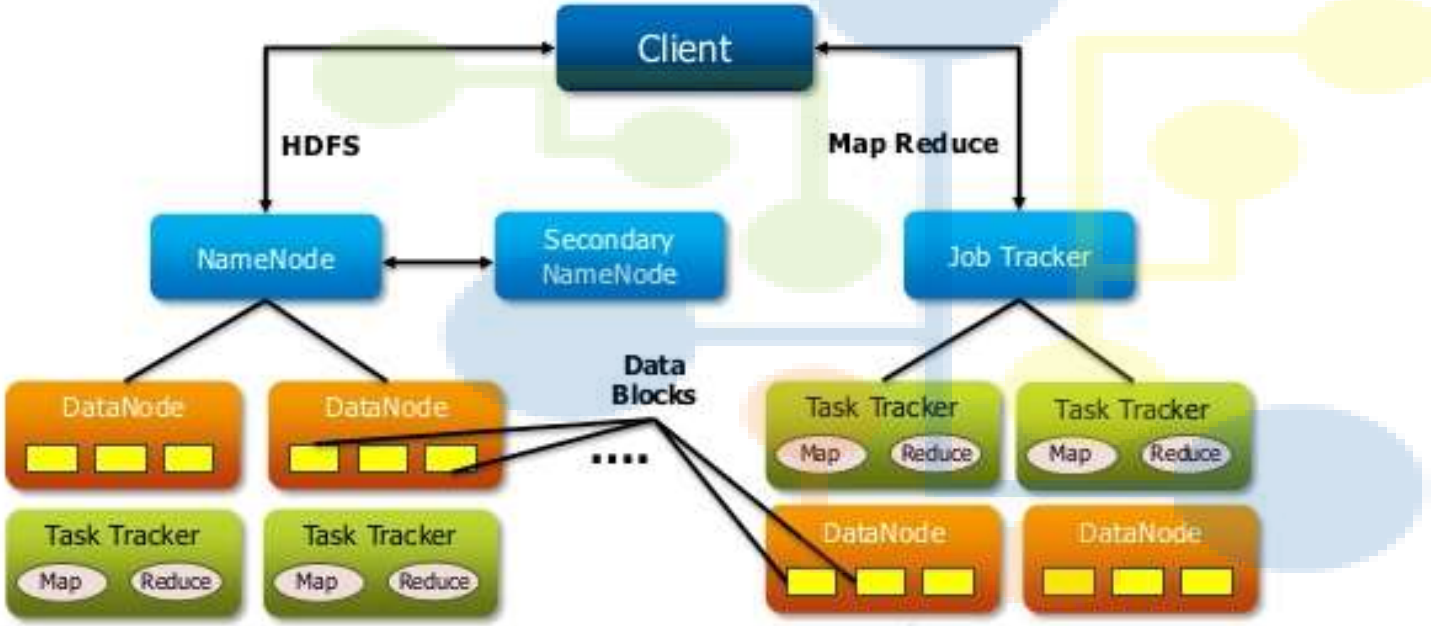
HDFS – Arquitetura



NameNode

- FsImage
- EditLog

HDFS – Arquitetura

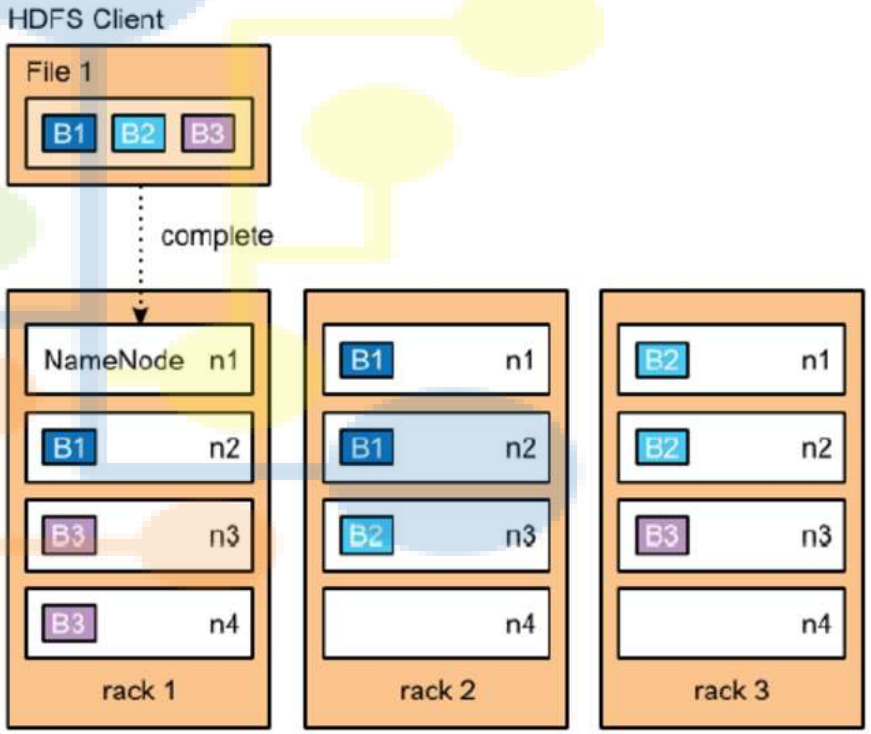


NameNode

- FsImage
- EditLog

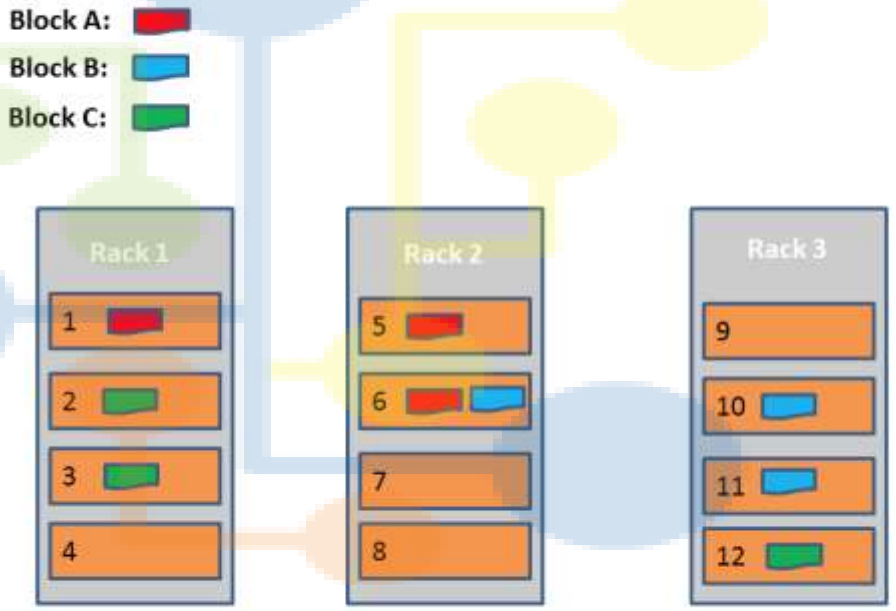
HDFS – Arquitetura

Replicação



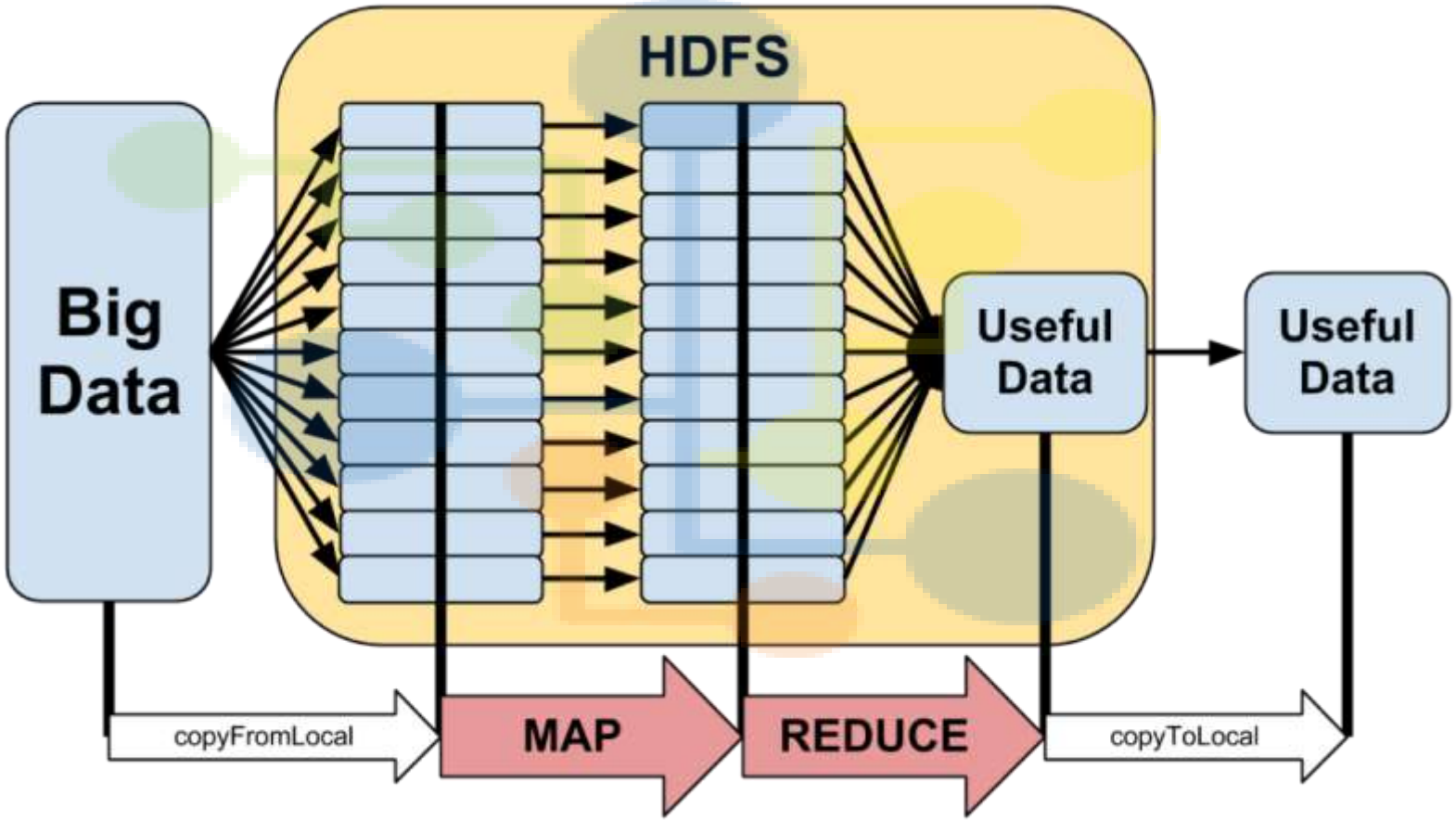
HDFS – Arquitetura

Replicação

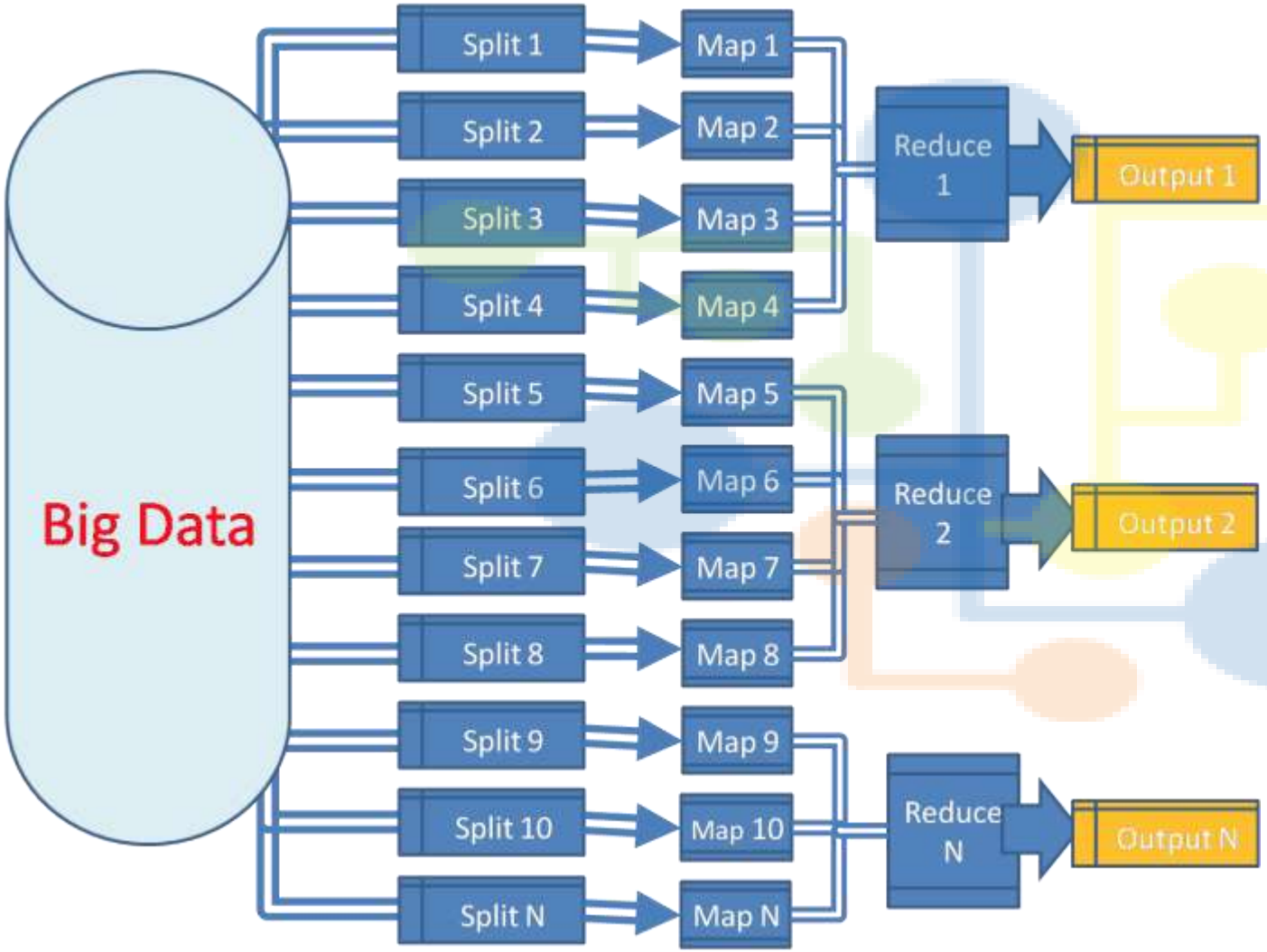


Definindo MapReduce

Definindo MapReduce

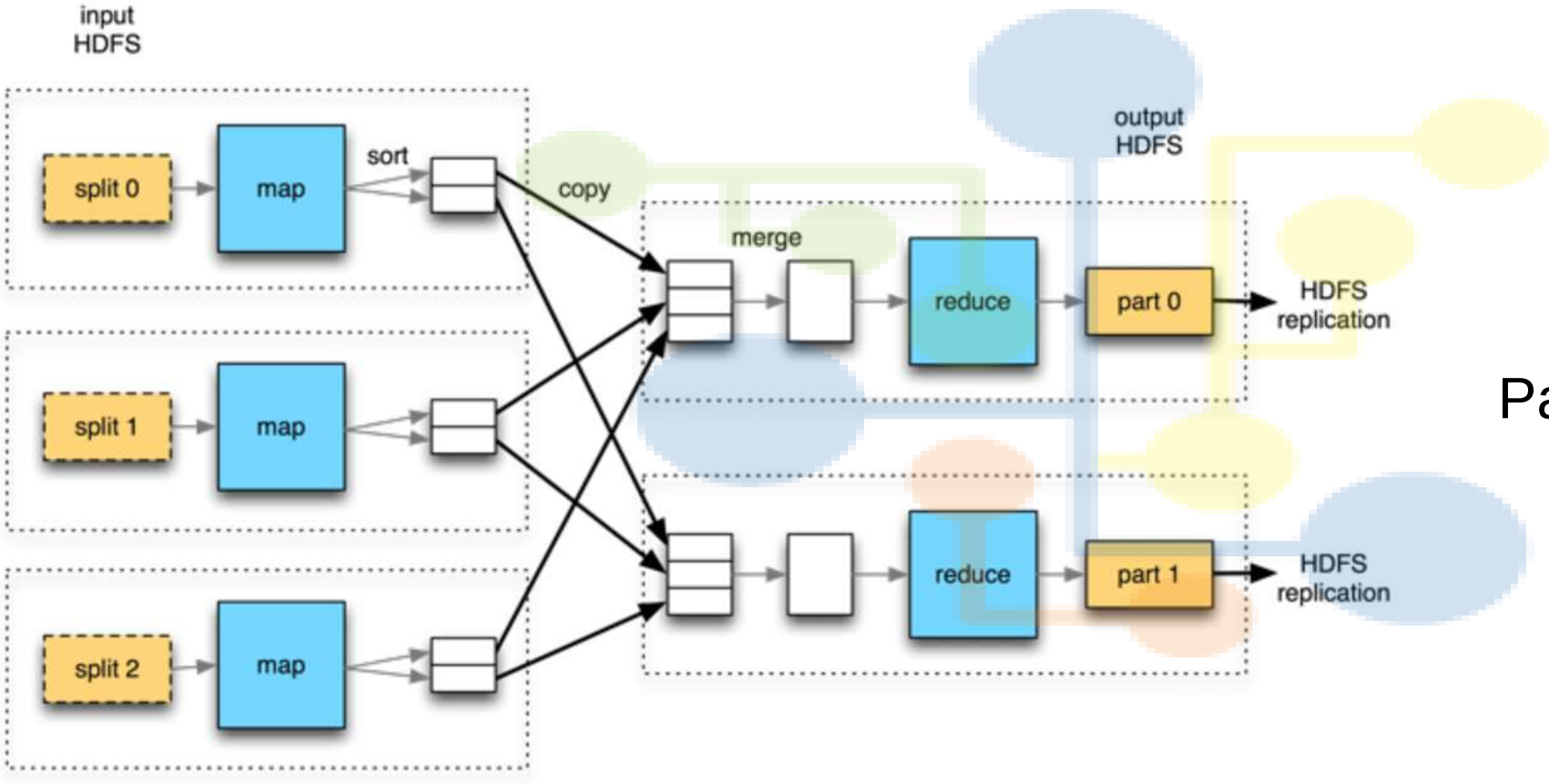


Definindo MapReduce



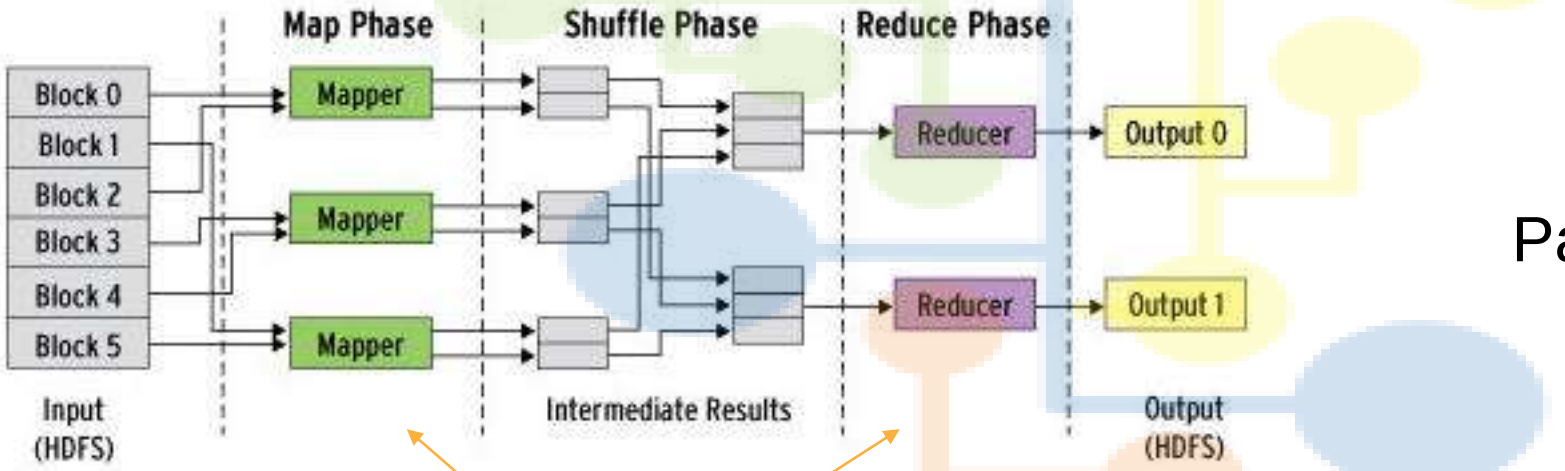
Processamento Paralelo e Distribuído

Definindo MapReduce



Processamento
Paralelo e Distribuído

Definindo MapReduce



Tarefas

Processamento
Paralelo e Distribuído

Hadoop
x
Bancos de Dados Relacionais

Hadoop x Bancos de Dados Relacionais

Bancos de Dados Relacionais

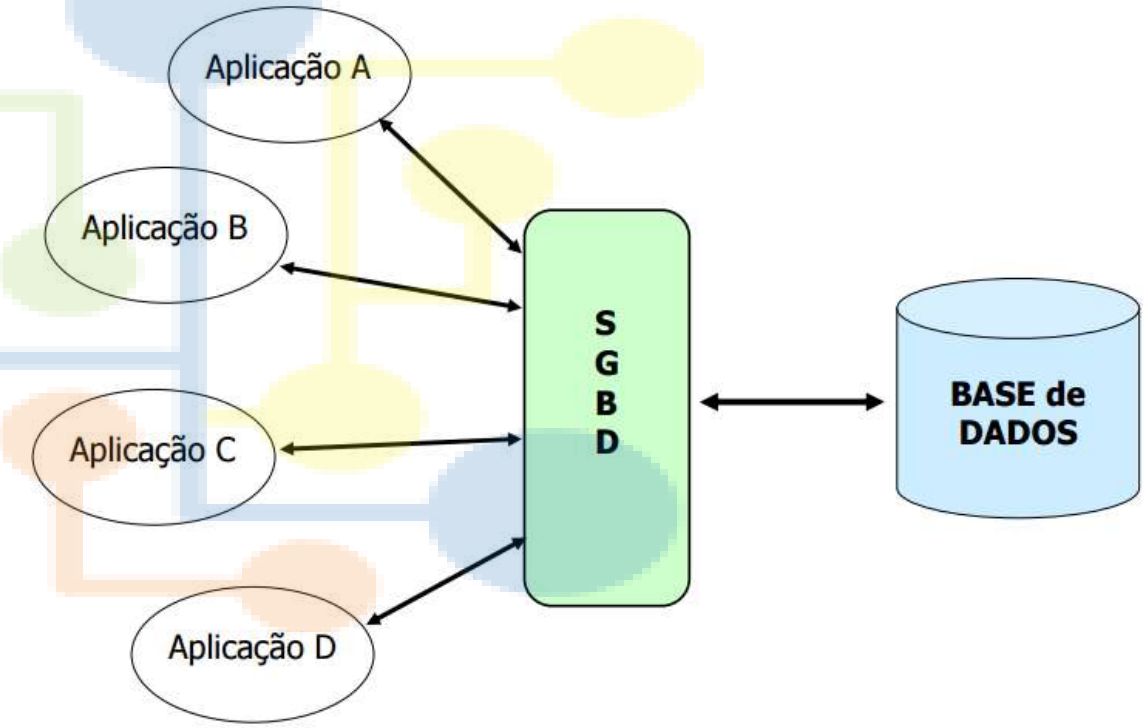
Hadoop x Bancos de Dados Relacionais

RDBMS
Relational Database
Management Systems

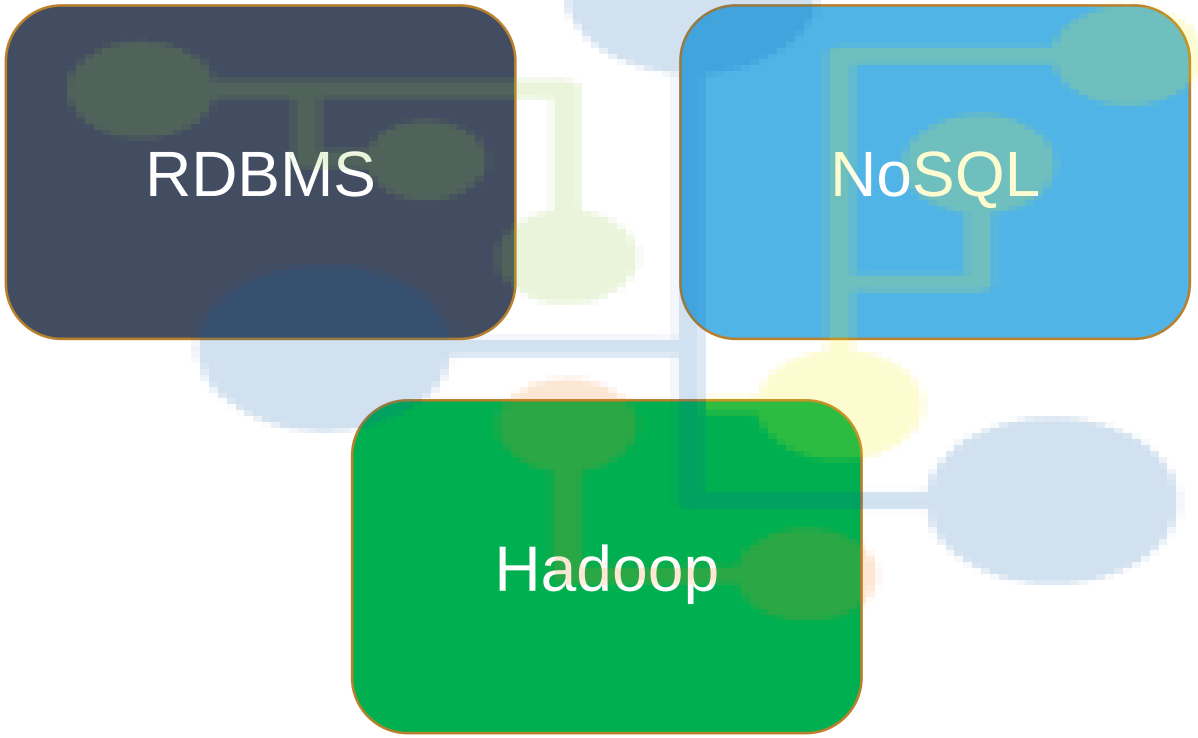
Hadoop x Bancos de Dados Relacionais

SGBD’s

Gerenciam um ou mais bancos de dados



Hadoop x Bancos de Dados Relacionais



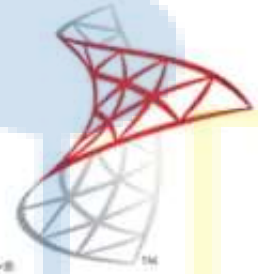
Hadoop x Bancos de Dados Relacionais

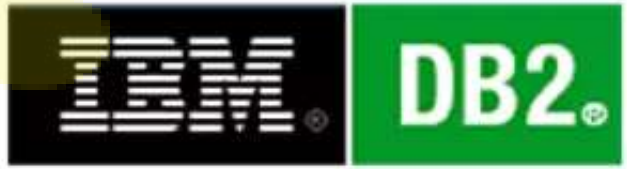
Hadoop x Bancos de Dados Relacionais

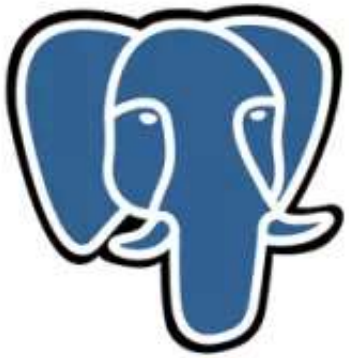
ACID

Atomicidade
Consistência
Isolamento
Durabilidade

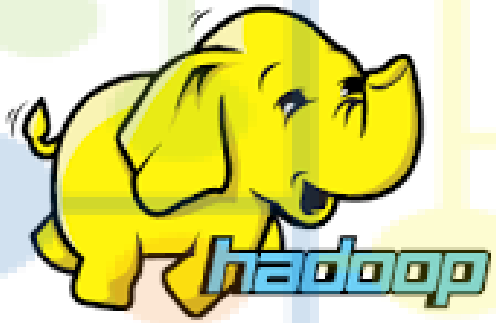
Hadoop x Bancos de Dados Relacionais





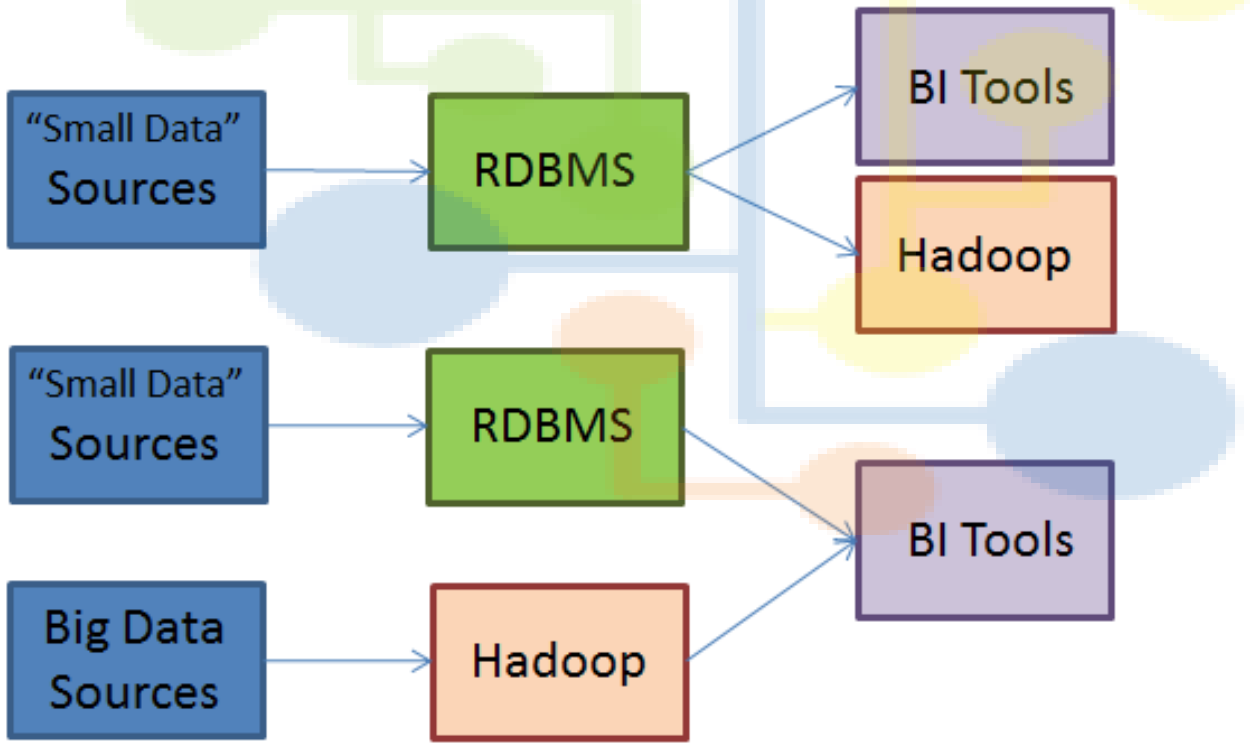


Hadoop x Bancos de Dados Relacionais



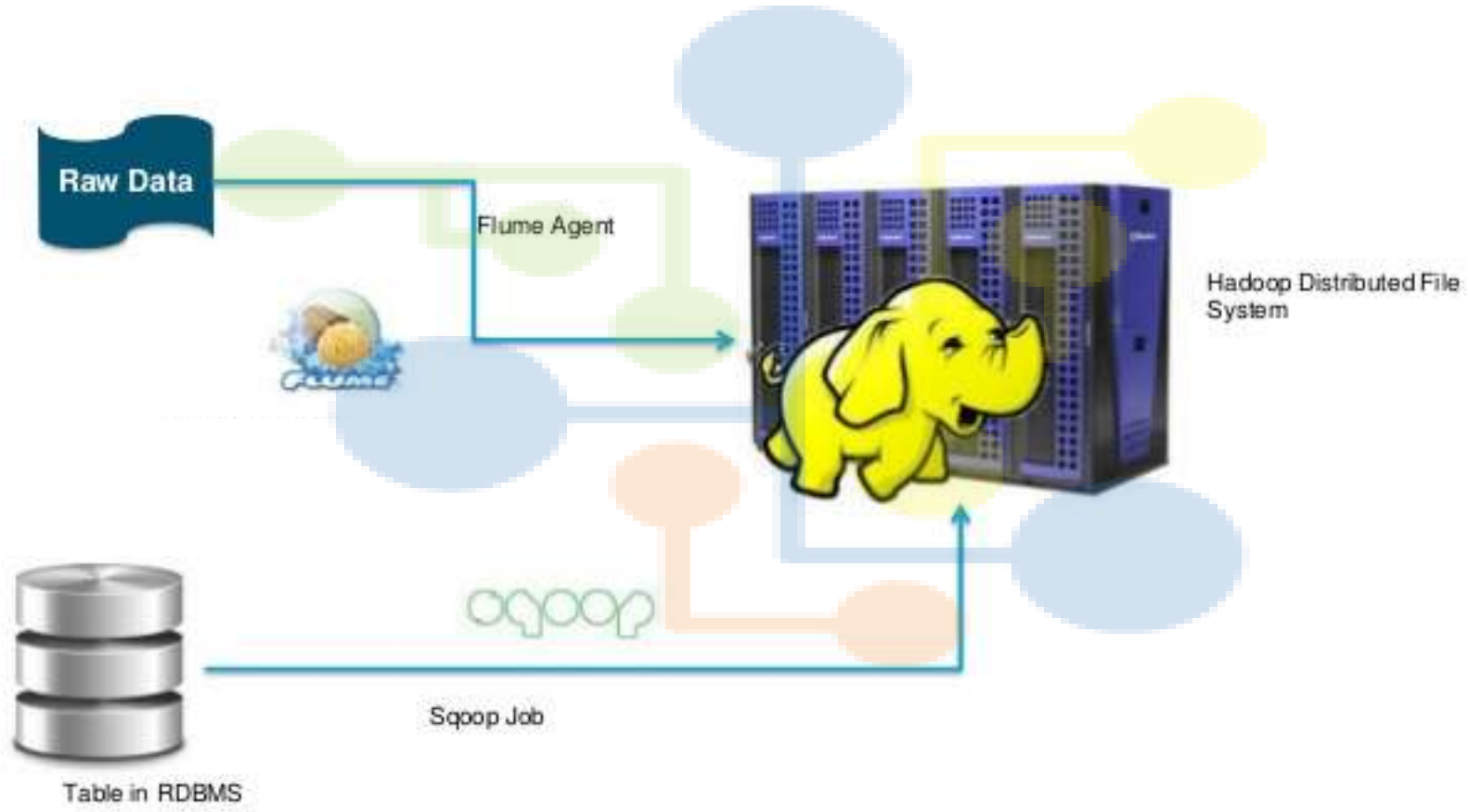
Hadoop x Bancos de Dados Relacionais

Hadoop → Grandes volumes de dados (estruturados ou não estruturados)
RDBMS → Dados transacionais (dados estruturados)



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