Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Engenharia de Dados com Hadoop e Spark

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Bem-vindo(a)

Instrutores

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

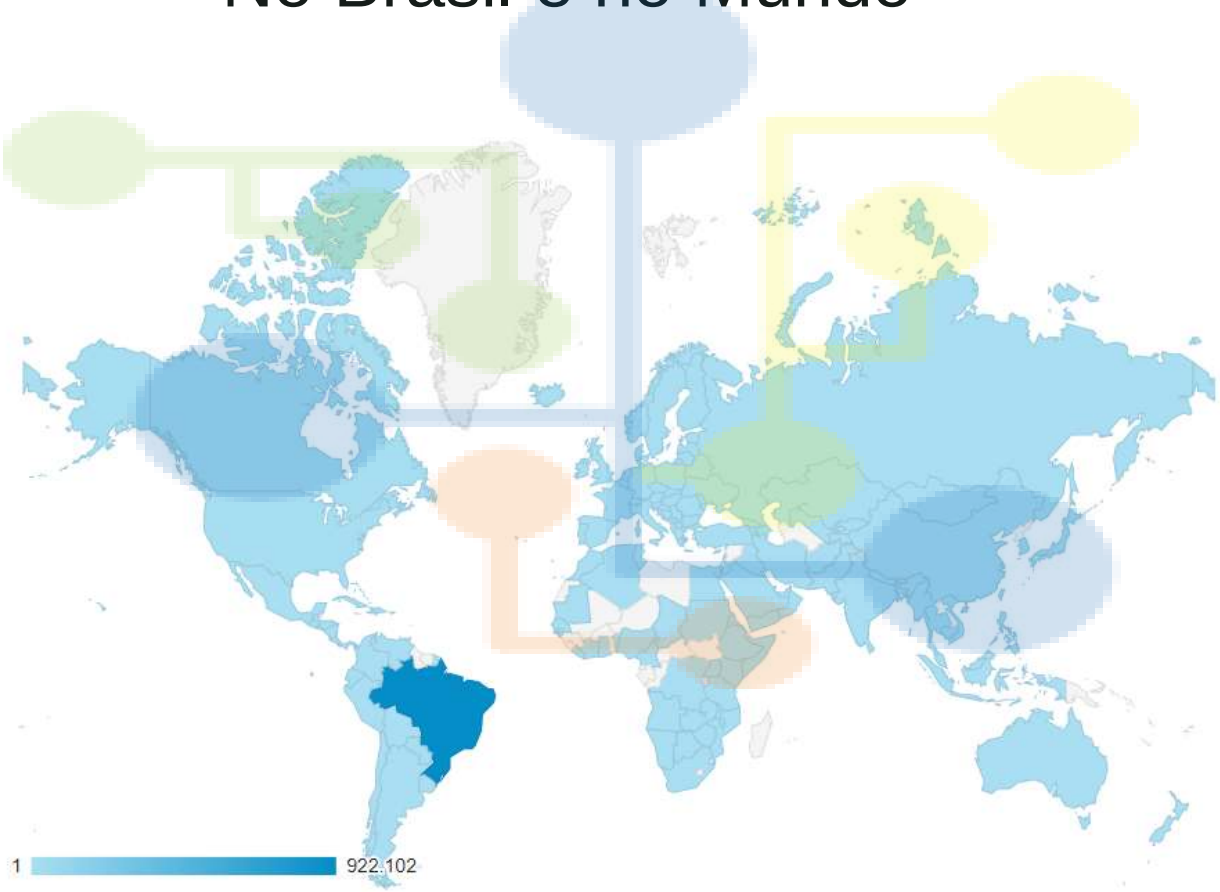
Data Science Academy

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.

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

No Brasil e no Mundo



Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Engenharia de Dados com Hadoop e Spark

Engenharia de Dados com Hadoop e Spark

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

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

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

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

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

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

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

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)

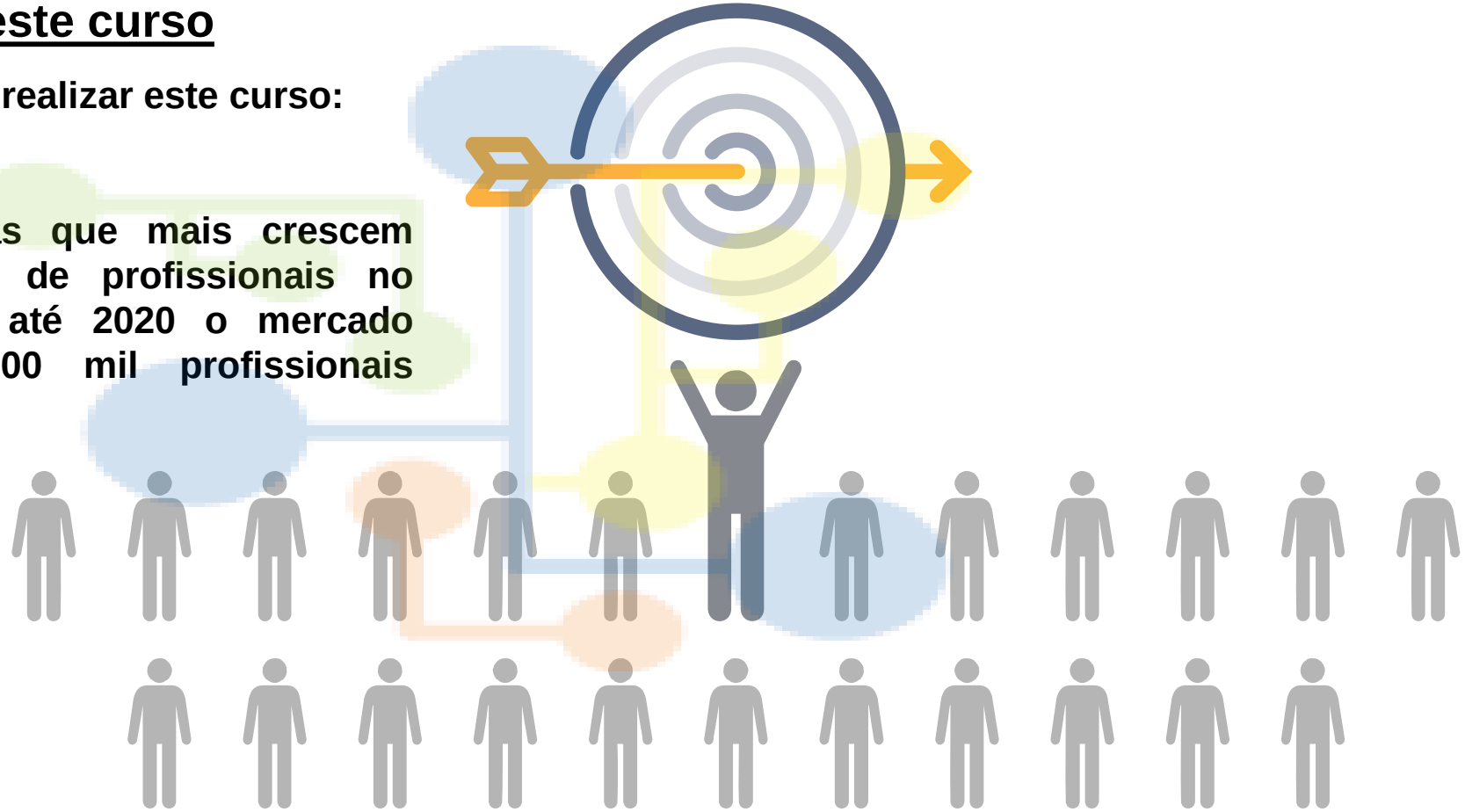
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

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.



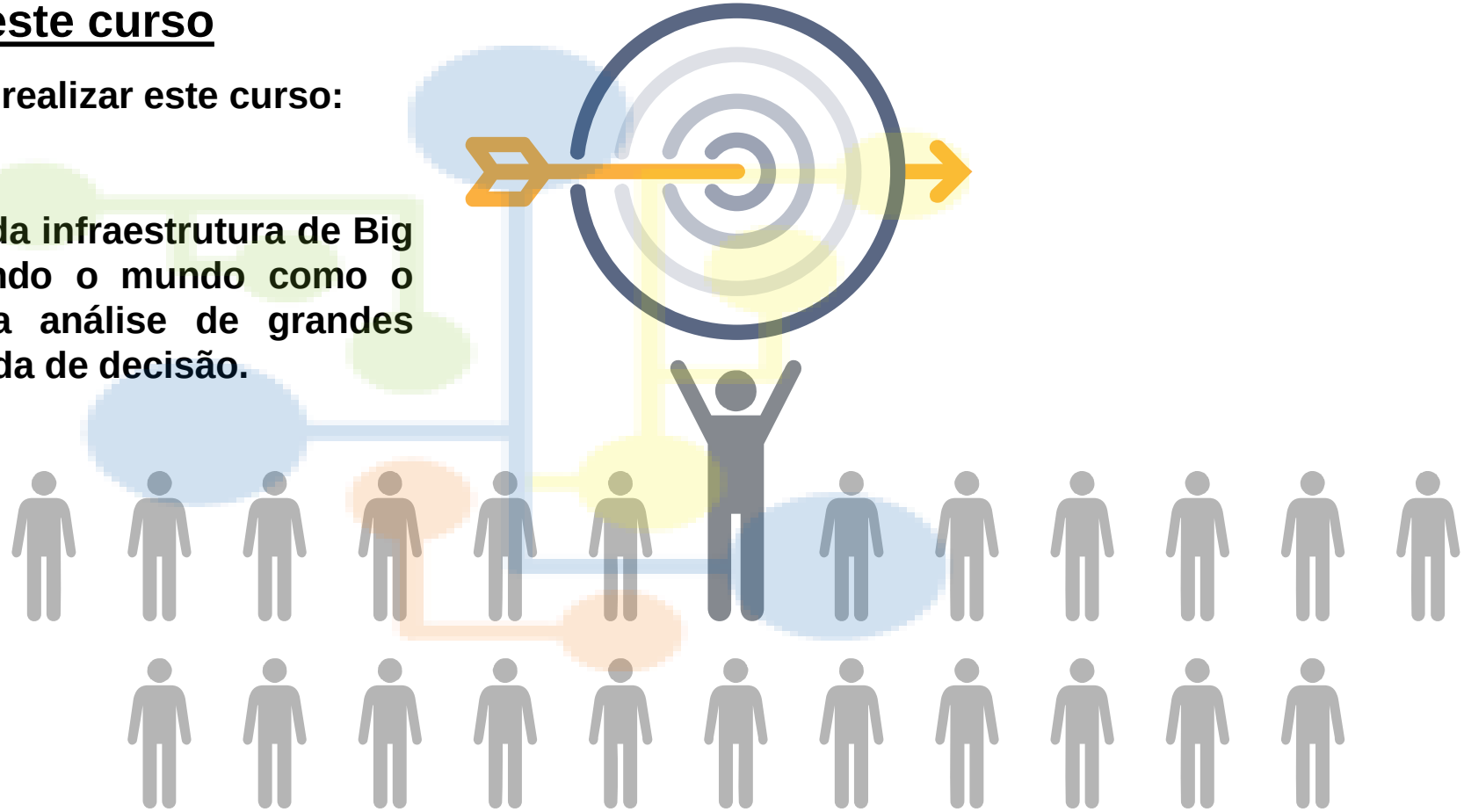
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

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.



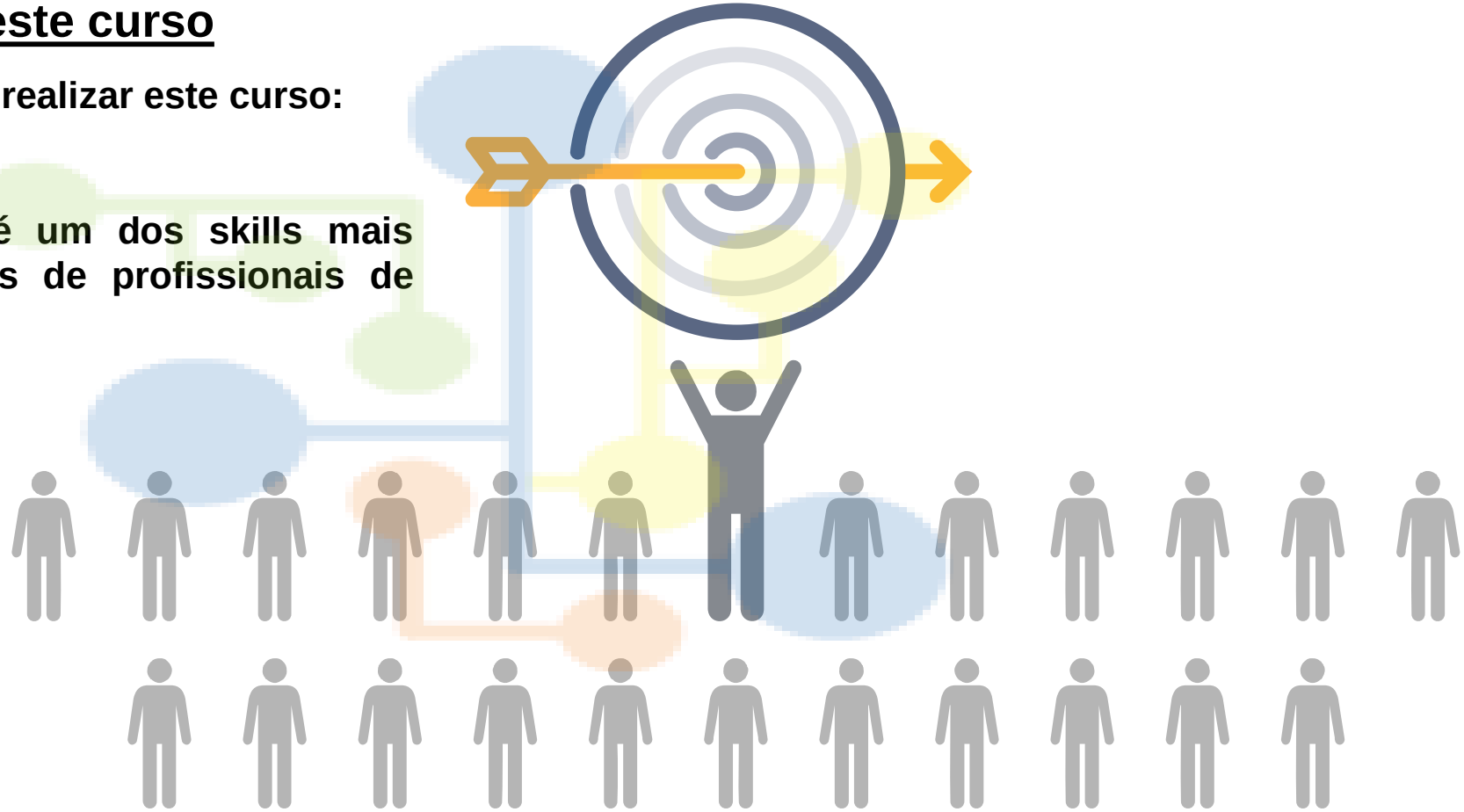
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

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.



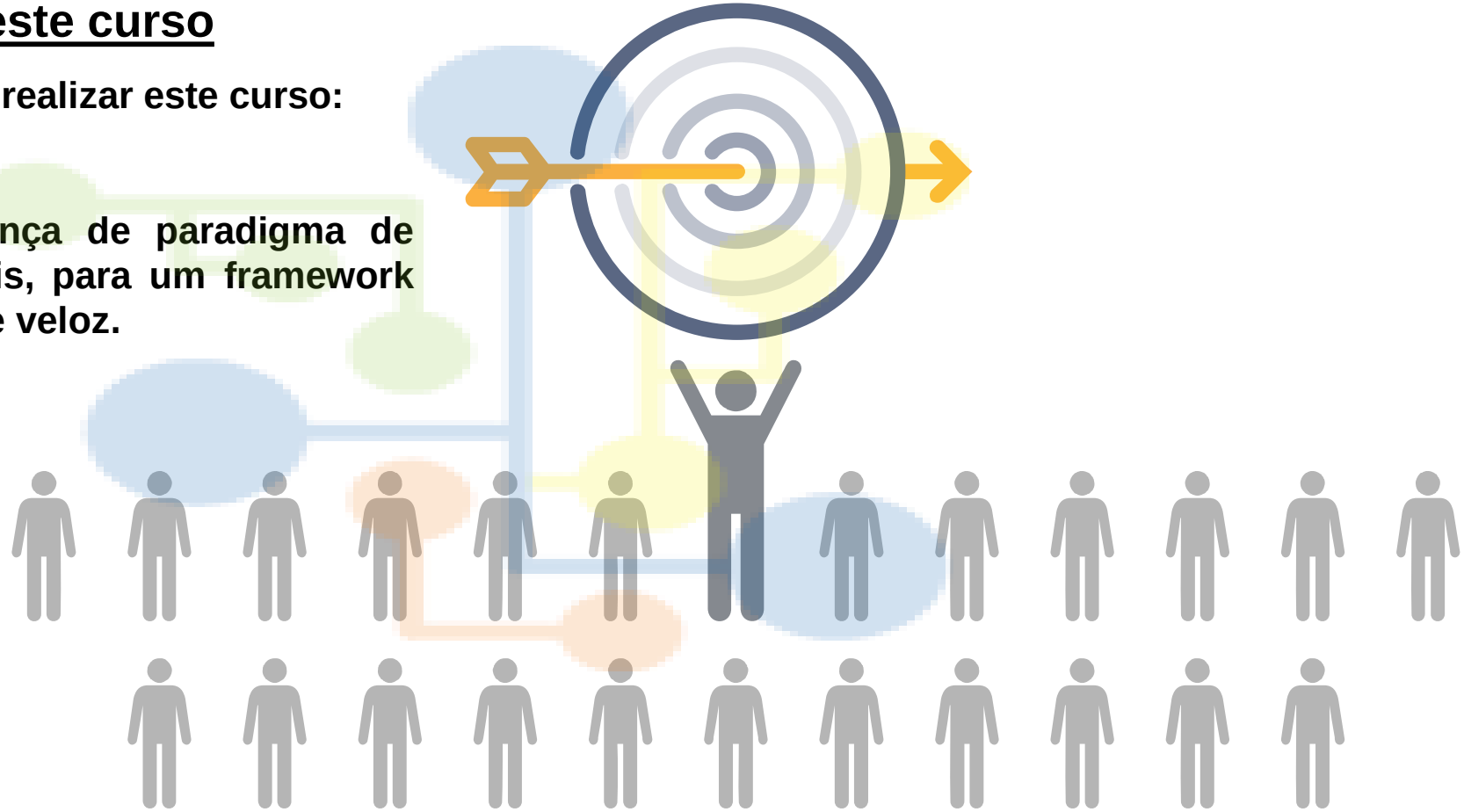
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

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.



Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Engenharia de Dados com Hadoop e Spark

Estrutura do Curso

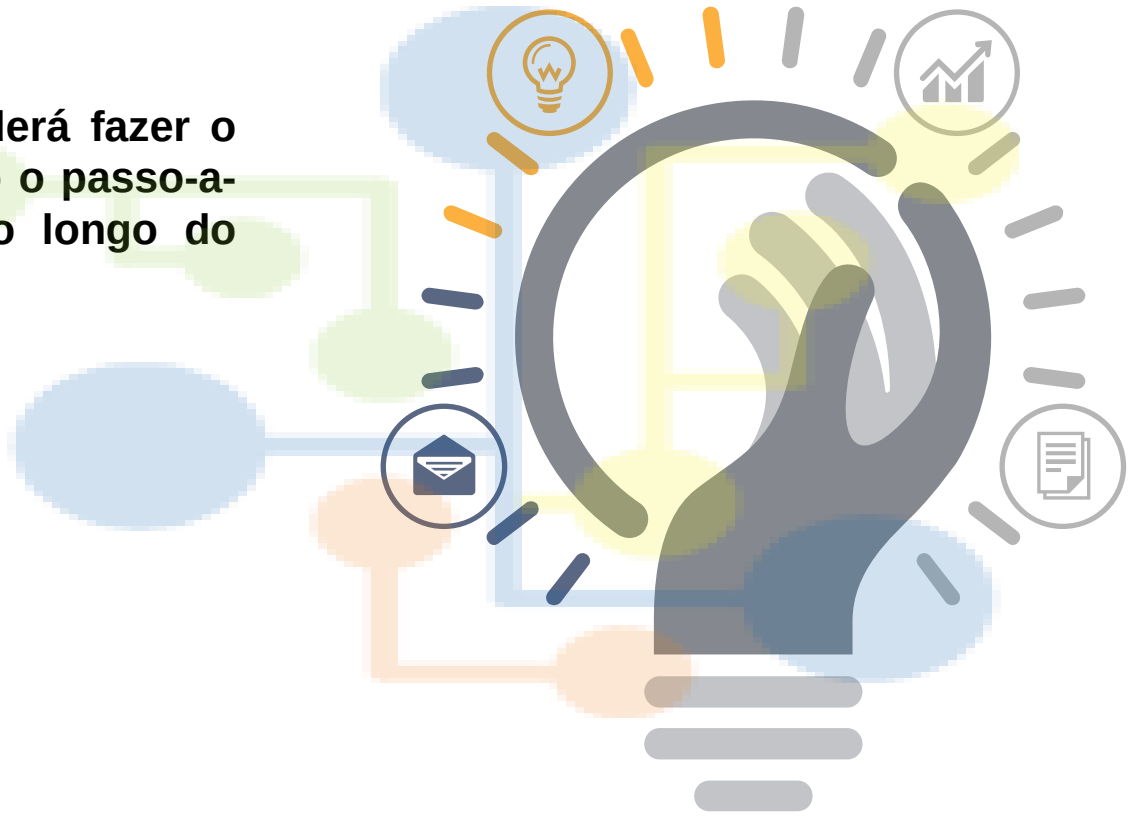
Para tornar sua experiência de aprendizagem ainda mais completa, você terá quizzes e labs ao longo de todos os capítulos.

Engenharia de Dados com Hadoop e Spark

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

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.



Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Engenharia de Dados com Hadoop e Spark

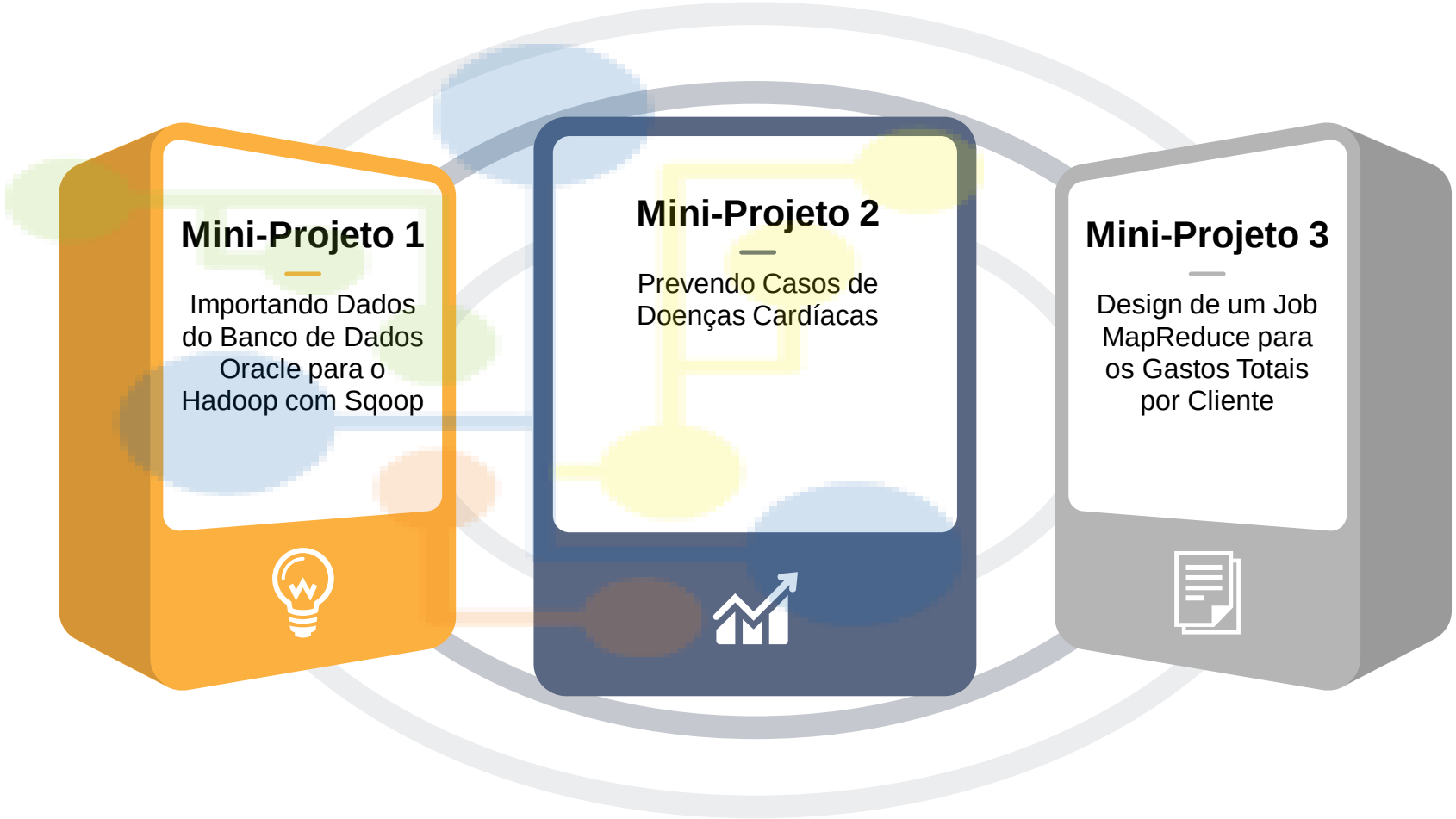
Estrutura do Curso

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

Engenharia de Dados com Hadoop e Spark

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Mini-Projetos



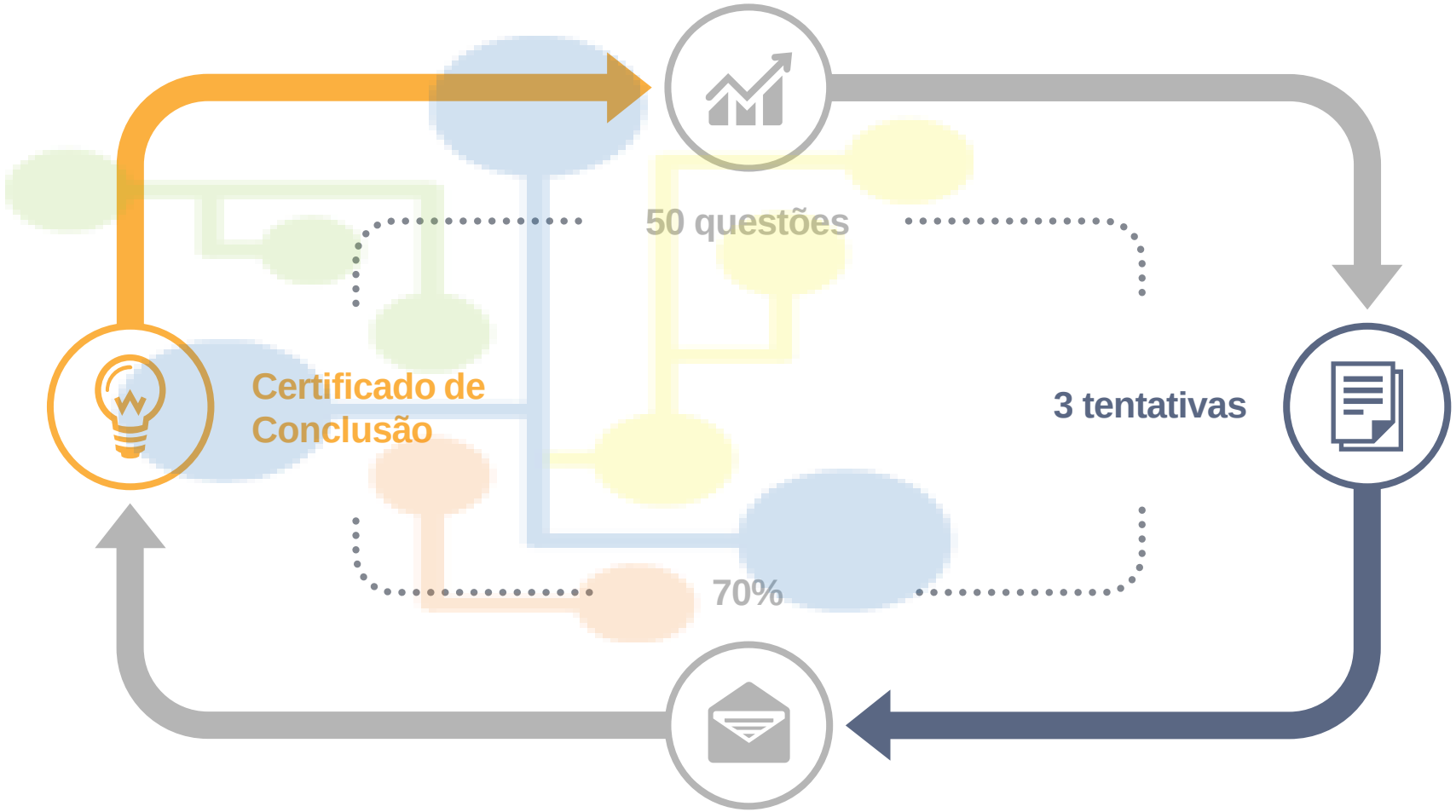
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Projetos com Feedback

Engenharia de Dados com Hadoop e Spark

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

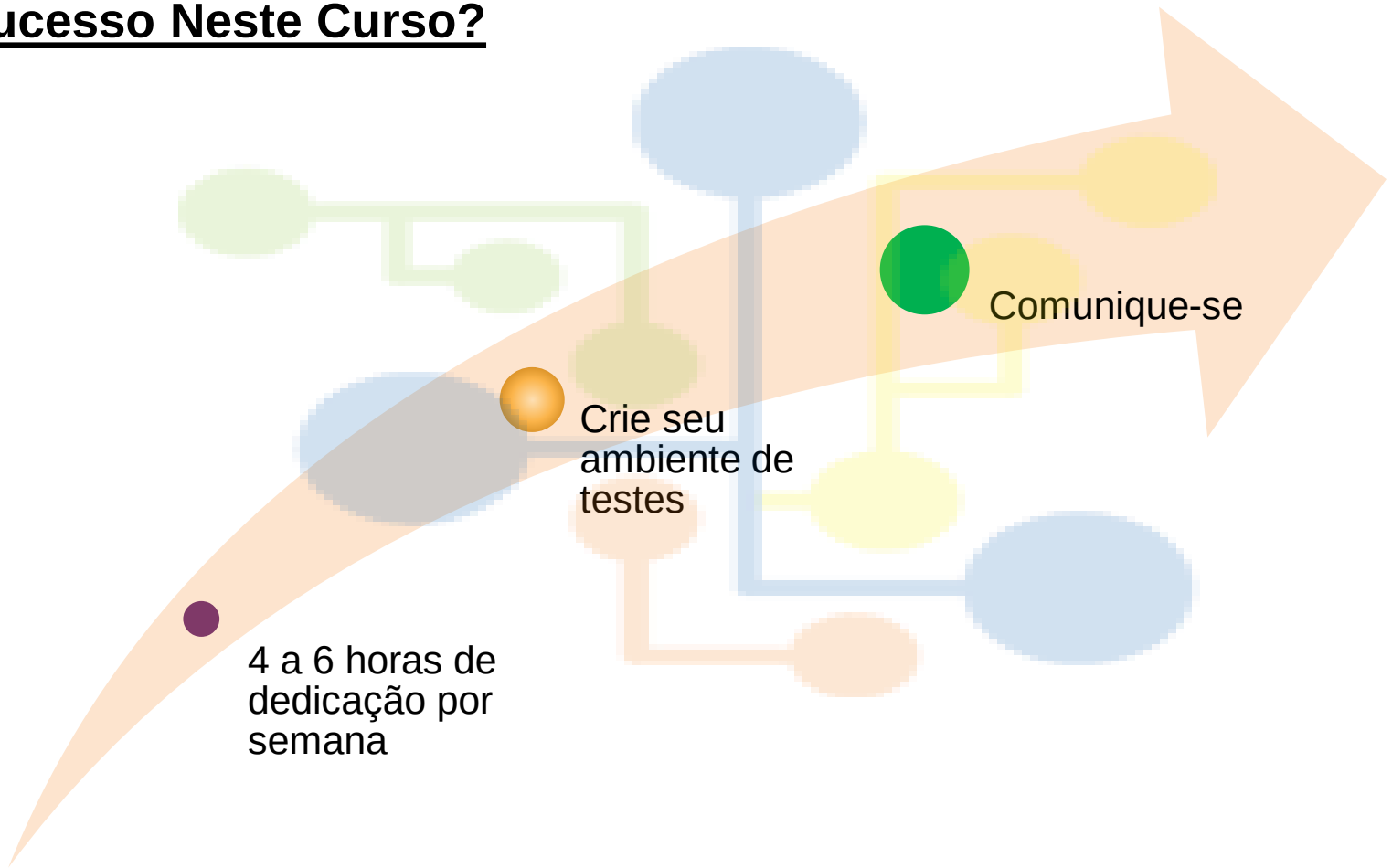
Avaliação Final



Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Engenharia de Dados com Hadoop e Spark

Como Obter Sucesso Neste Curso?



Engenharia de Dados com Hadoop e Spark

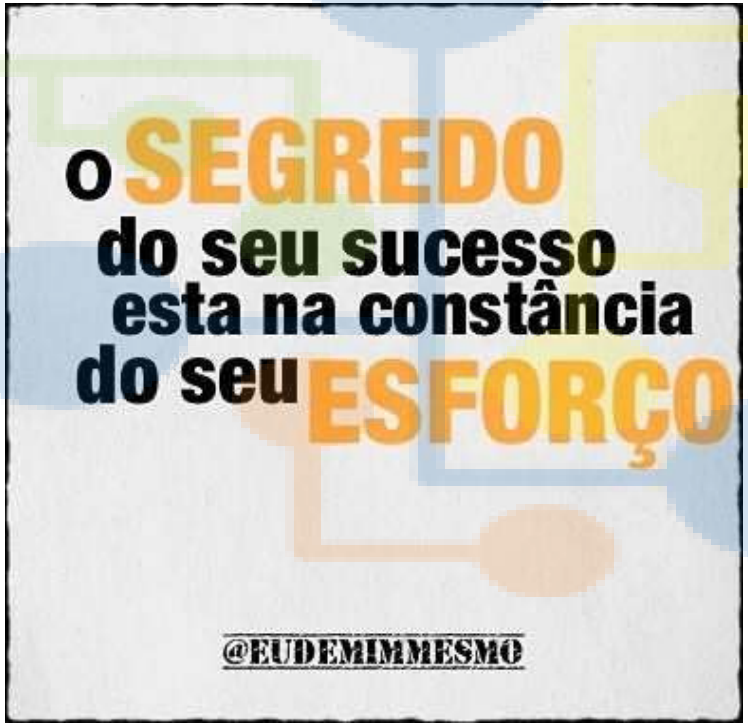
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Engenharia de Dados com Hadoop e Spark

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

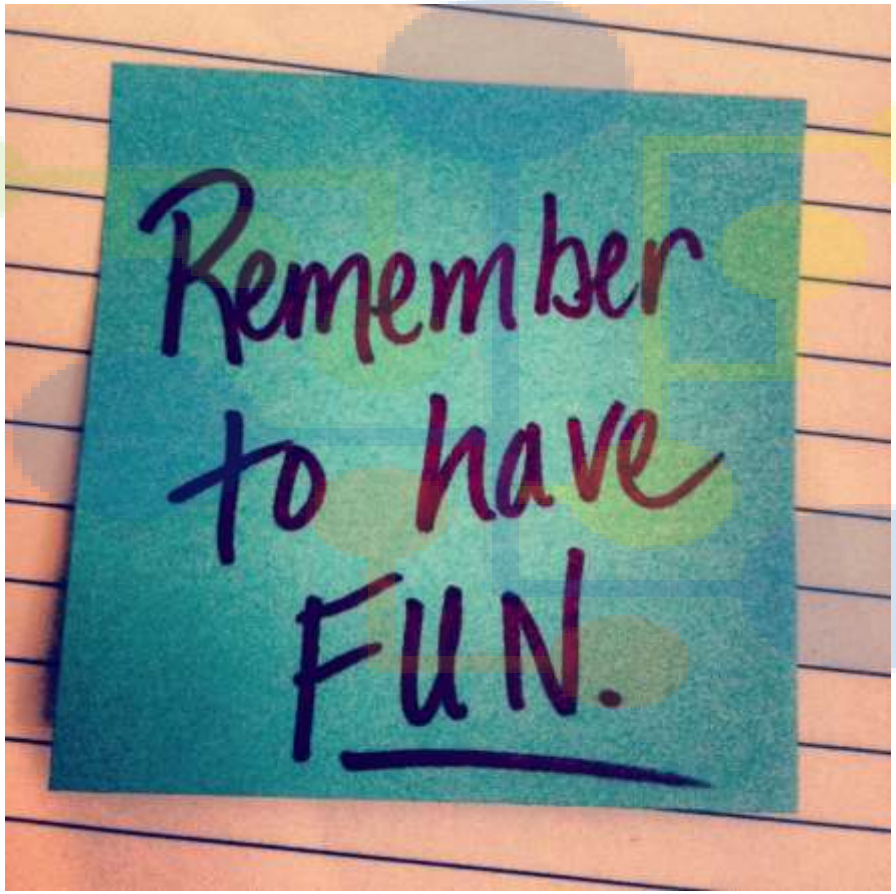
Engenharia de Dados com Hadoop e Spark

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



Engenharia de Dados com Hadoop e Spark

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



O que é o Apache Hadoop?

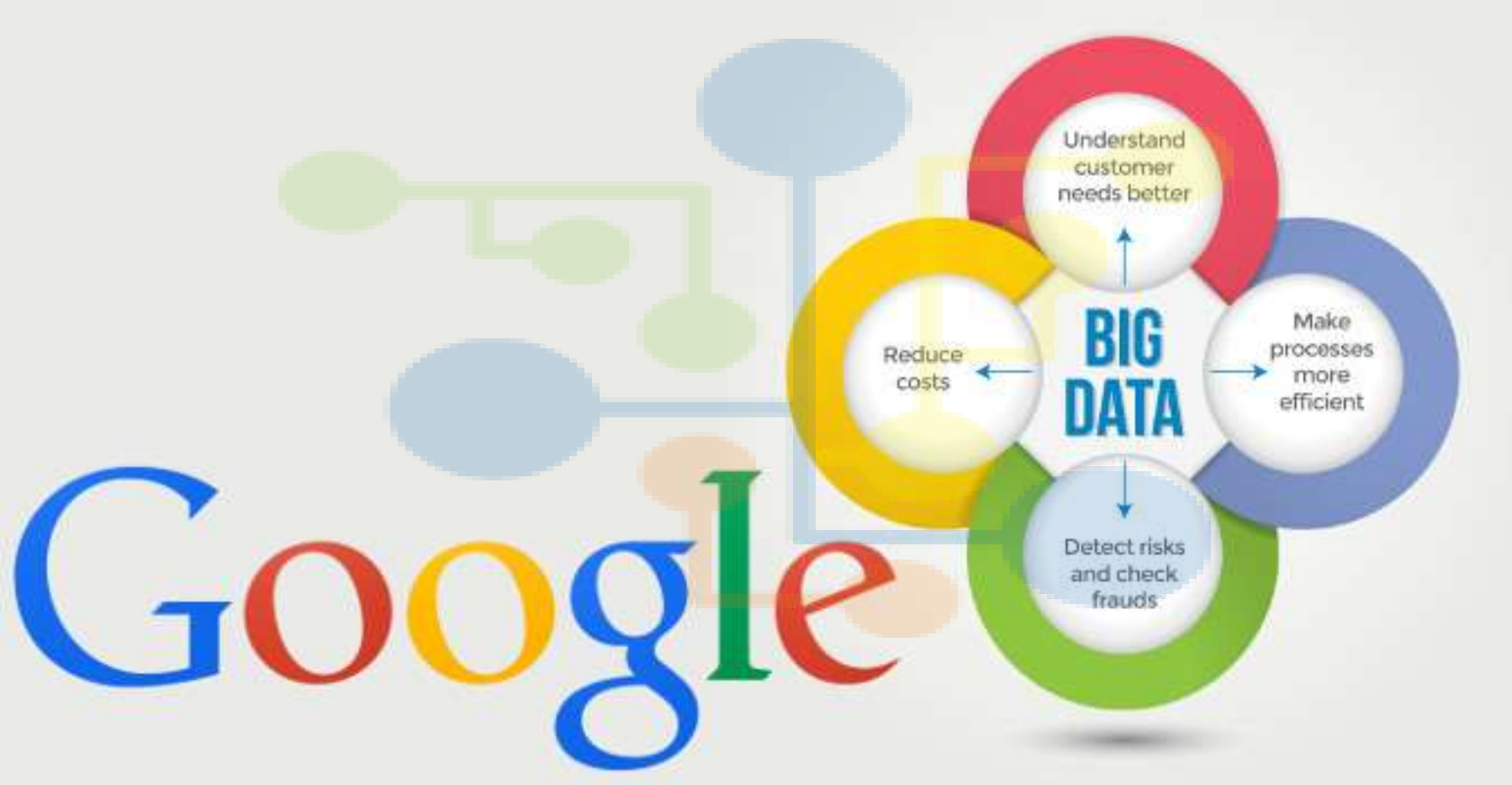
O que é o Apache Hadoop?

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.

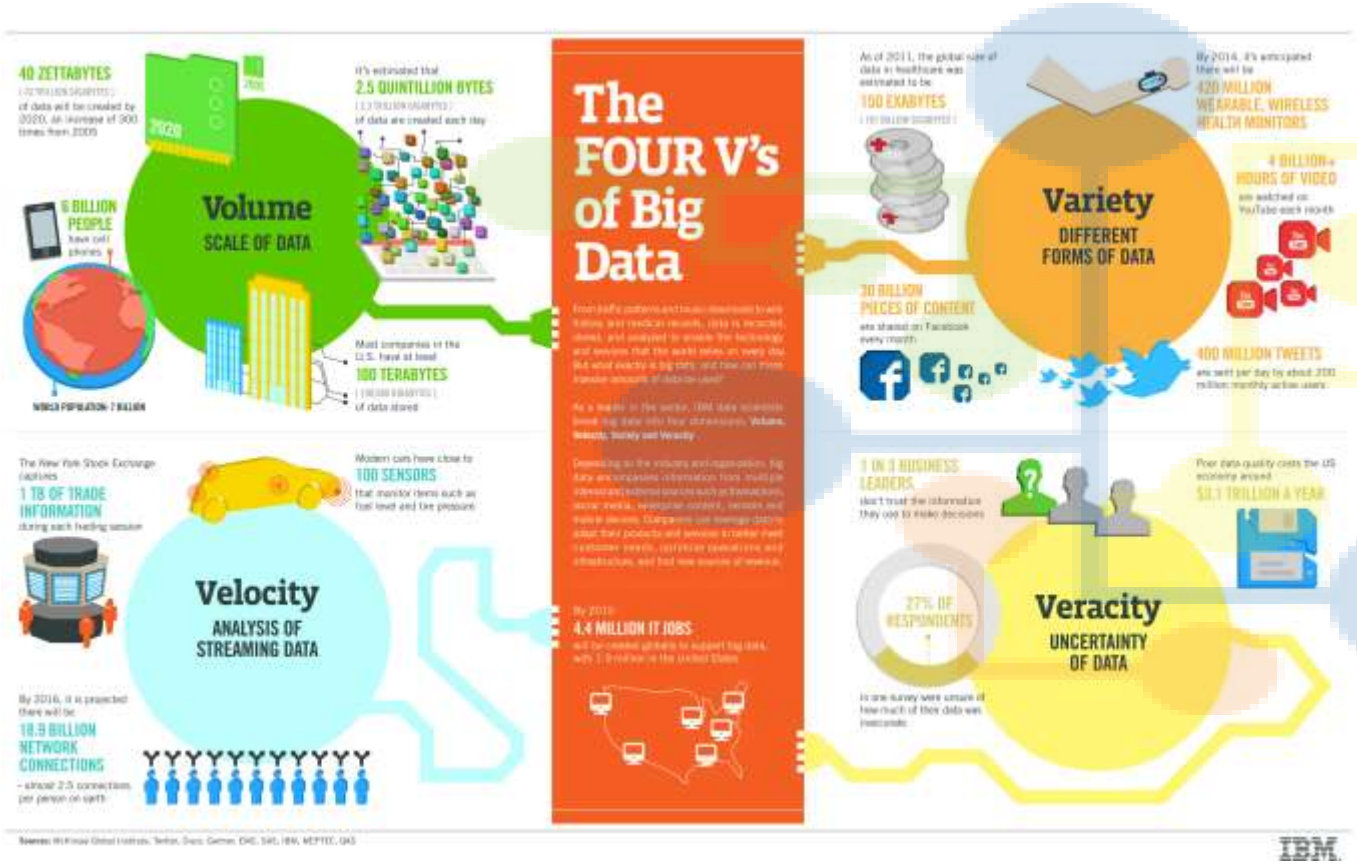
O que é o Apache Hadoop?



O que é o Apache Hadoop?

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

O que é o Apache Hadoop?



Os 4 V's do Big Data:

- Volume
- Variedade
- Velocidade
- Veracidade

O que é o Apache Hadoop?

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

O que é o Apache Hadoop?

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

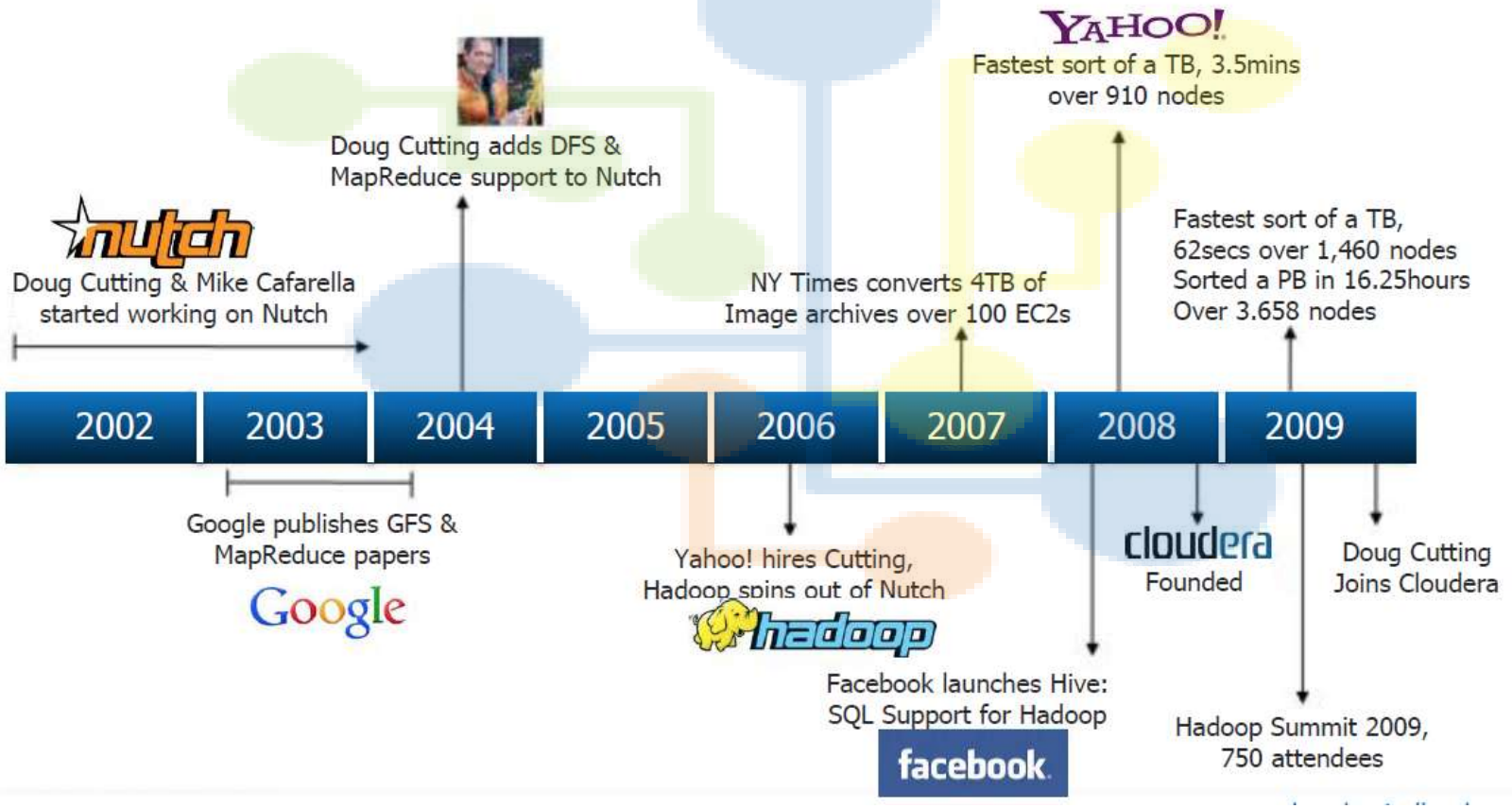
Computação Paralela

O que é o Apache Hadoop?

Uma Breve História do Apache Hadoop

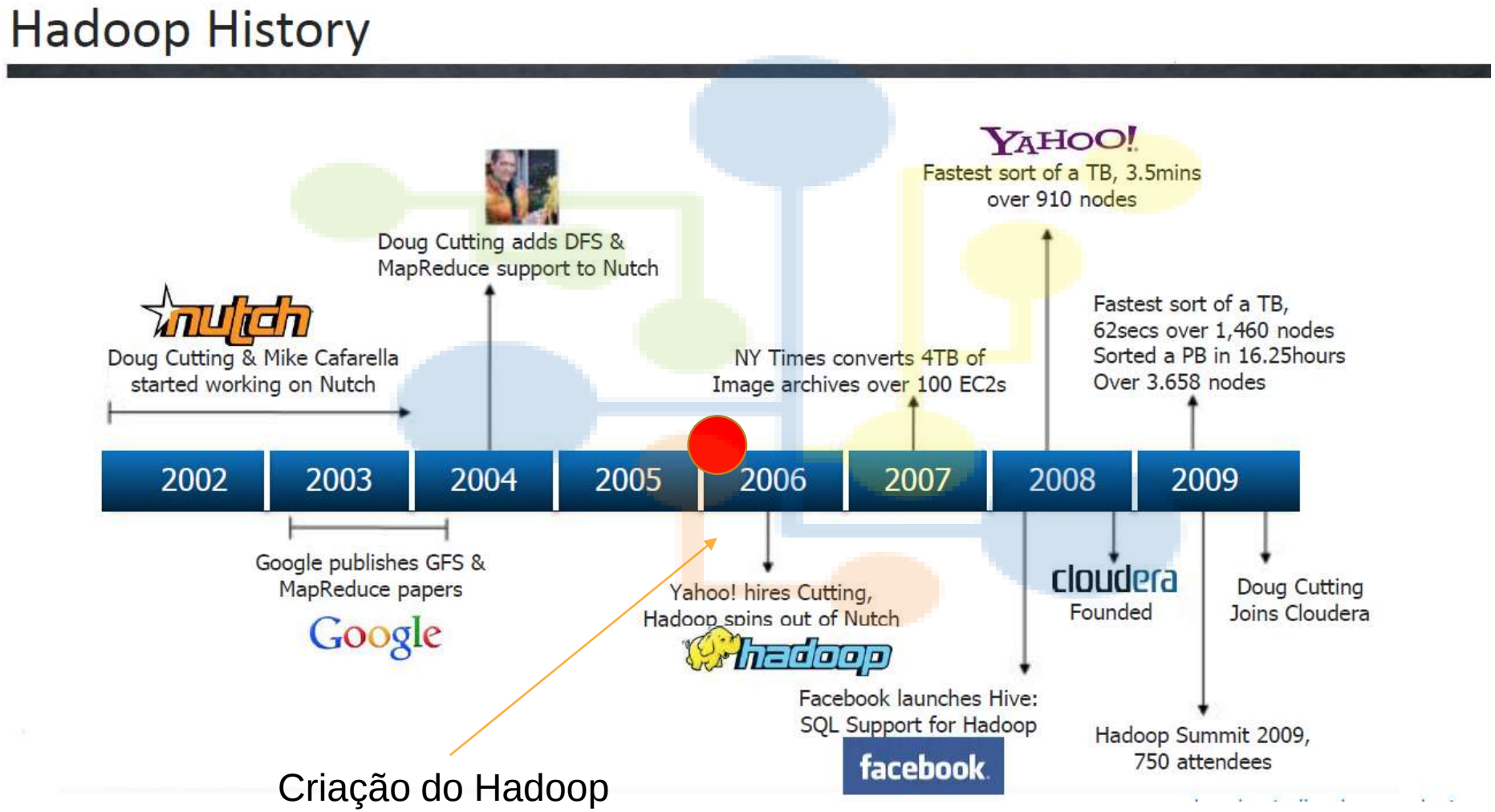
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Hadoop History



Uma Breve História do Apache Hadoop

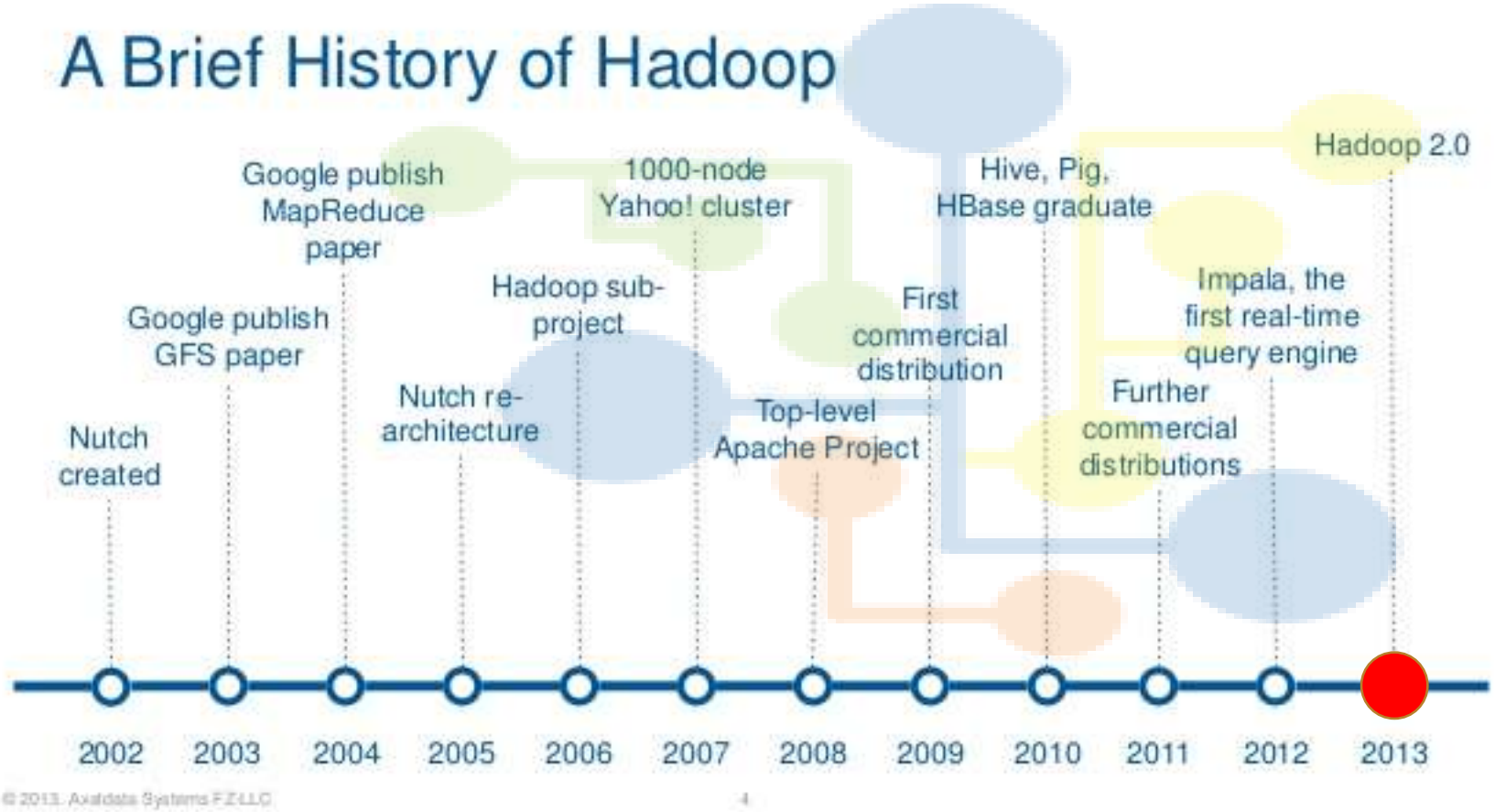
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



Criação do Hadoop

Uma Breve História do Apache Hadoop

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

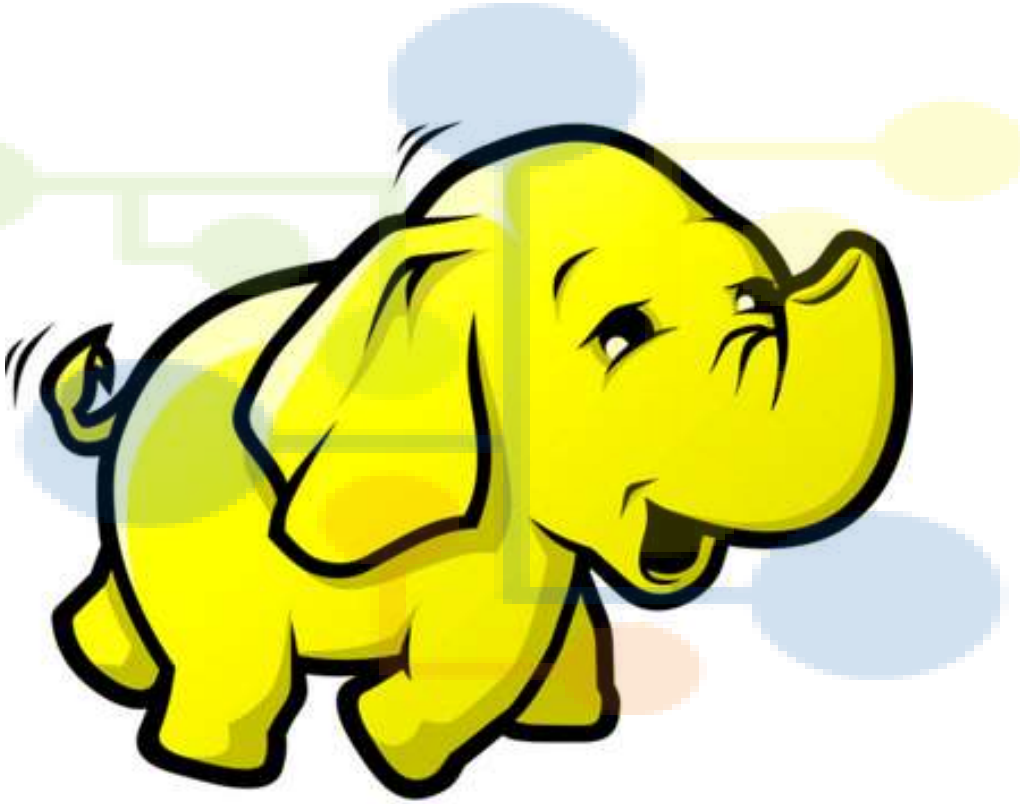


2018
Versão 3.0

2019
Versão 3.2.x

O que é o Hadoop?

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



Quais os benefícios para as Empresas ao utilizar o Hadoop?

Benefícios do Hadoop

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Benefícios do Hadoop

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Open Source

Benefícios do Hadoop

Economia

Benefícios do Hadoop

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Escalabilidade

Benefícios do Hadoop

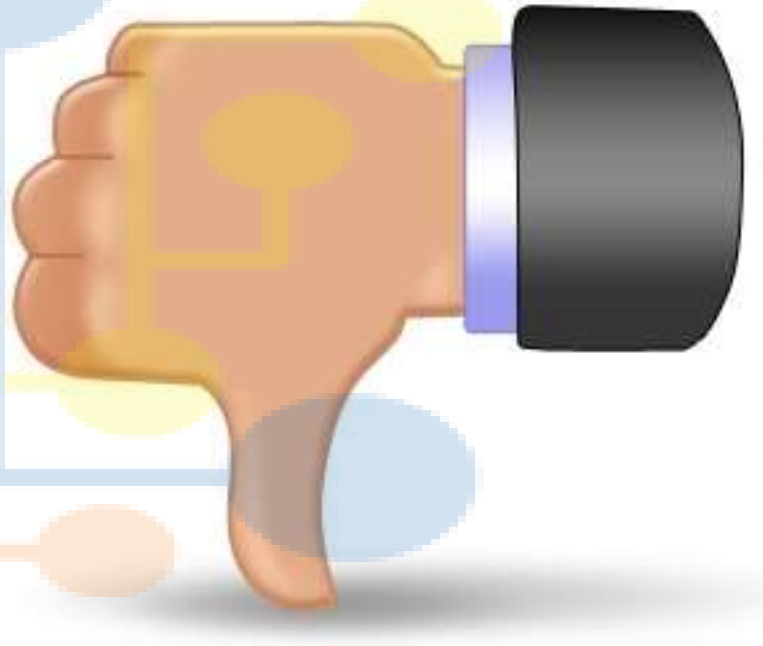
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Robustez

Desvantagens do Hadoop

Data Science Academy Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Desvantagens



Desvantagens do Hadoop

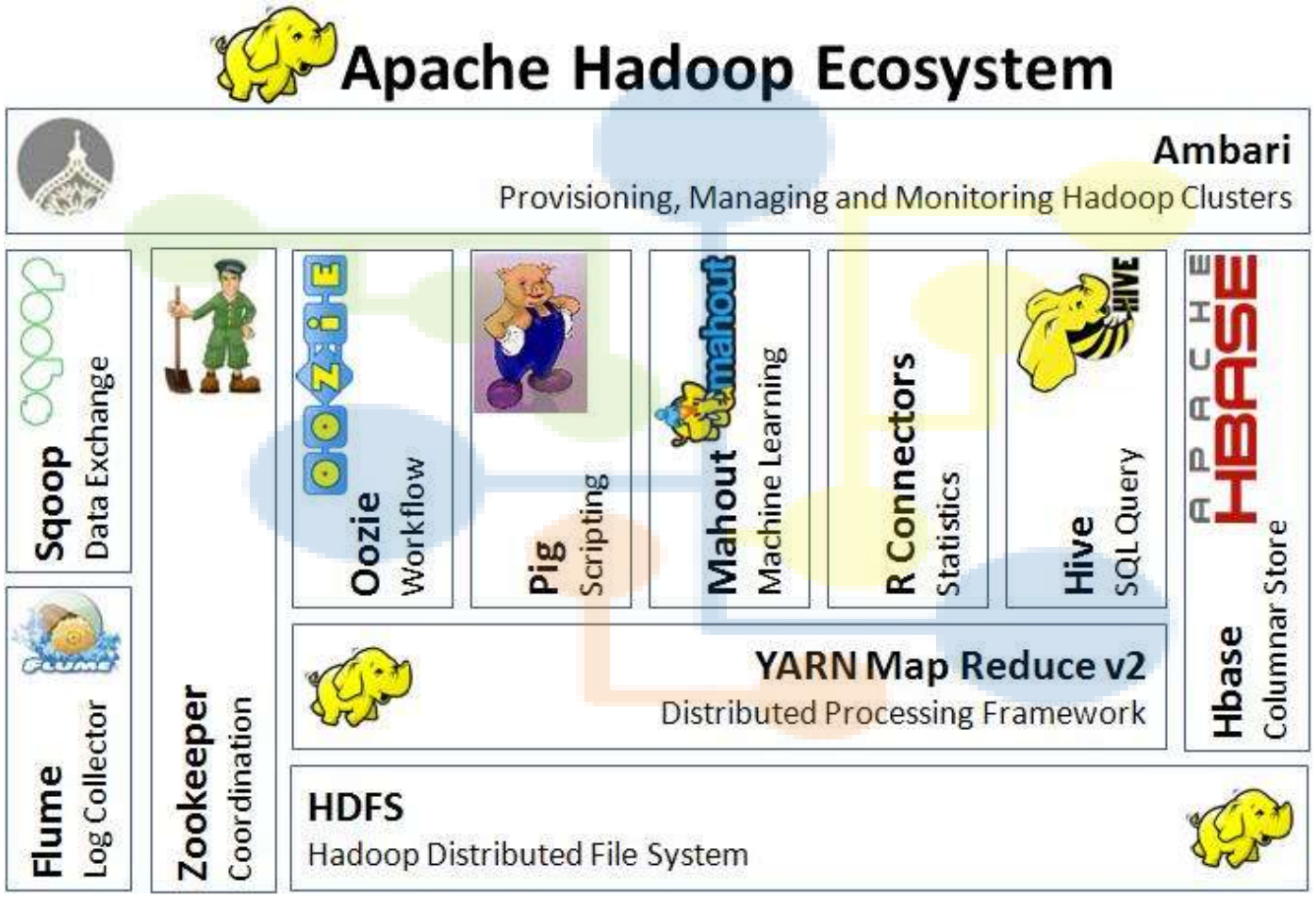
Data Science Academy Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Desvantagens

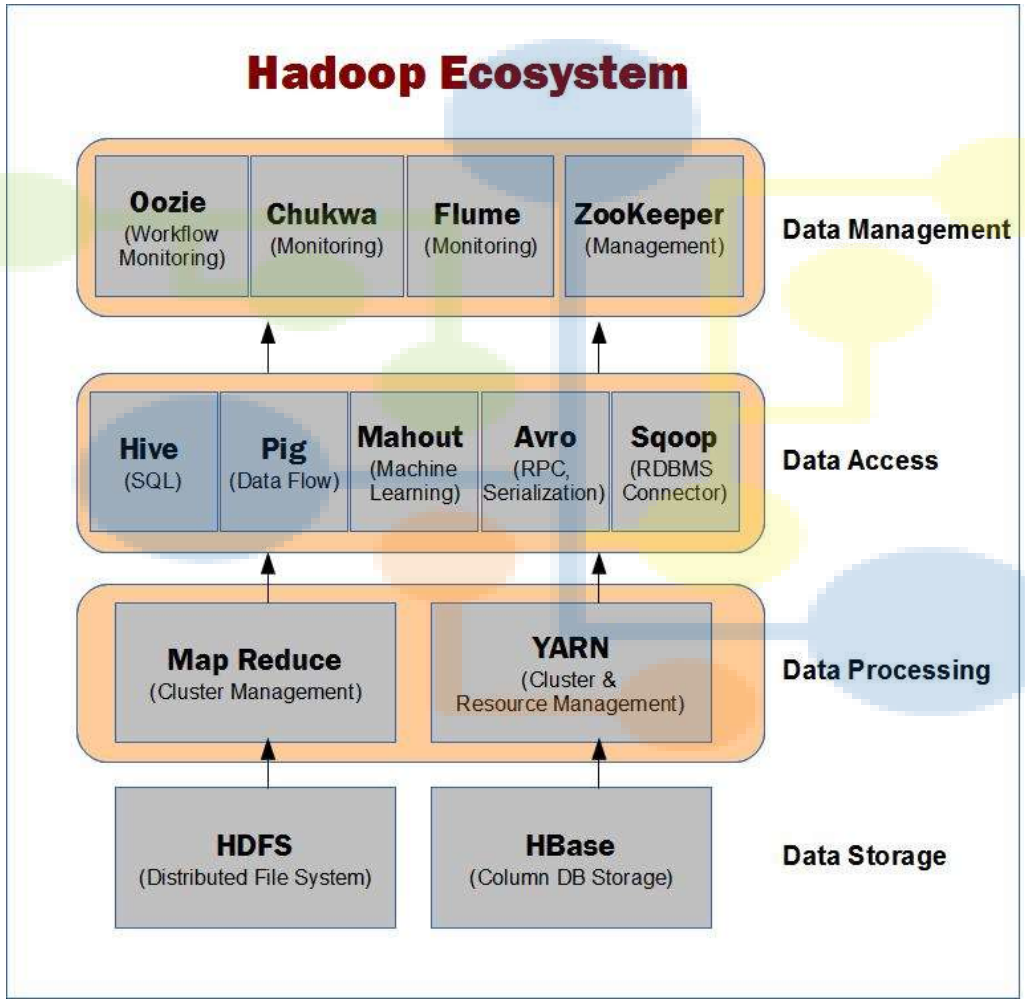
- Node Master Único
- Processamento de Arquivos Pequenos
- Muito Processamento em Poucos Dados

Ecossistema Hadoop

Ecossistema Hadoop

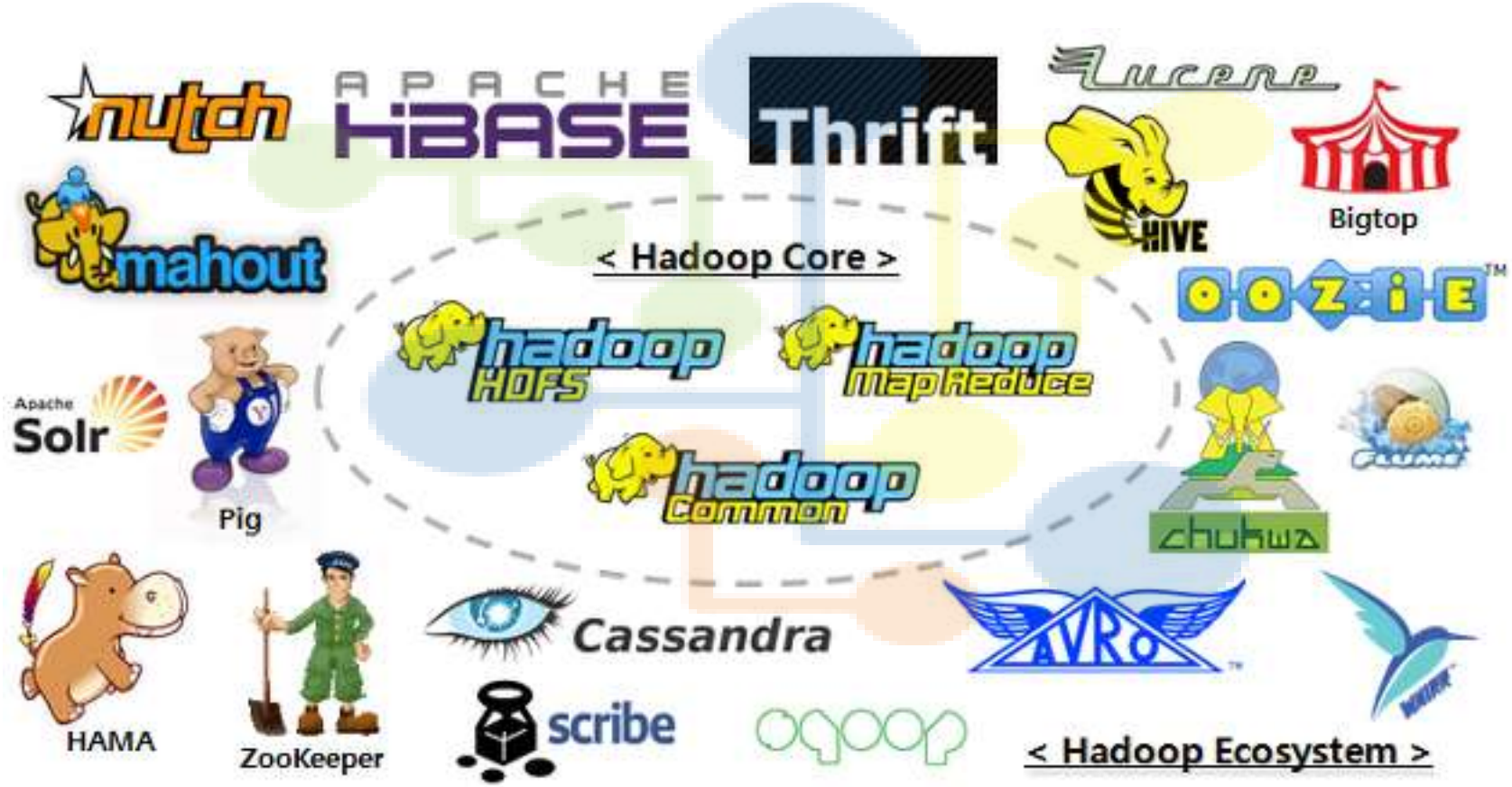


Ecossistema Hadoop



Ecossistema Hadoop

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



Ecossistema Hadoop

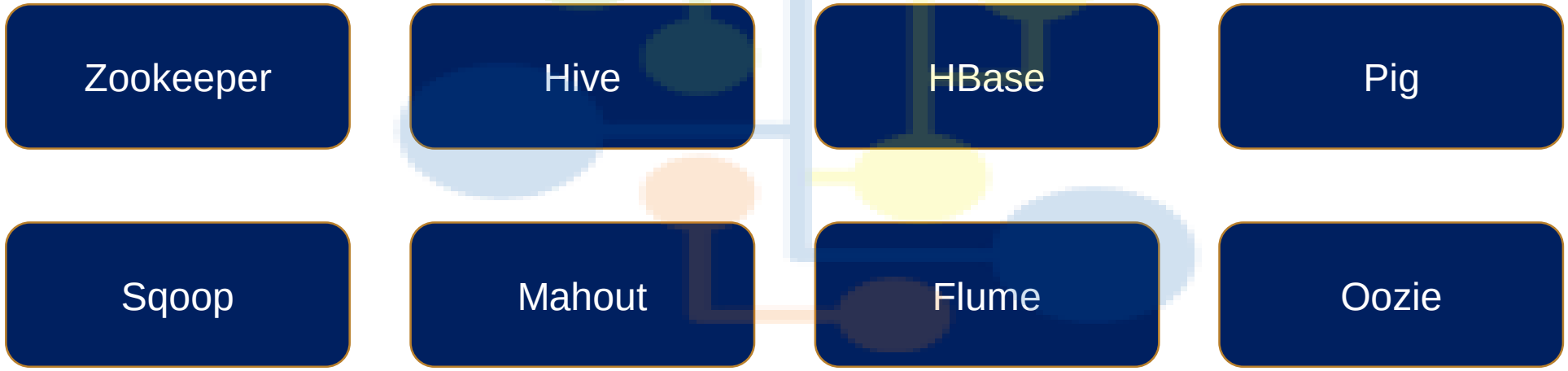
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Projetos Principais do Ecossistema Hadoop

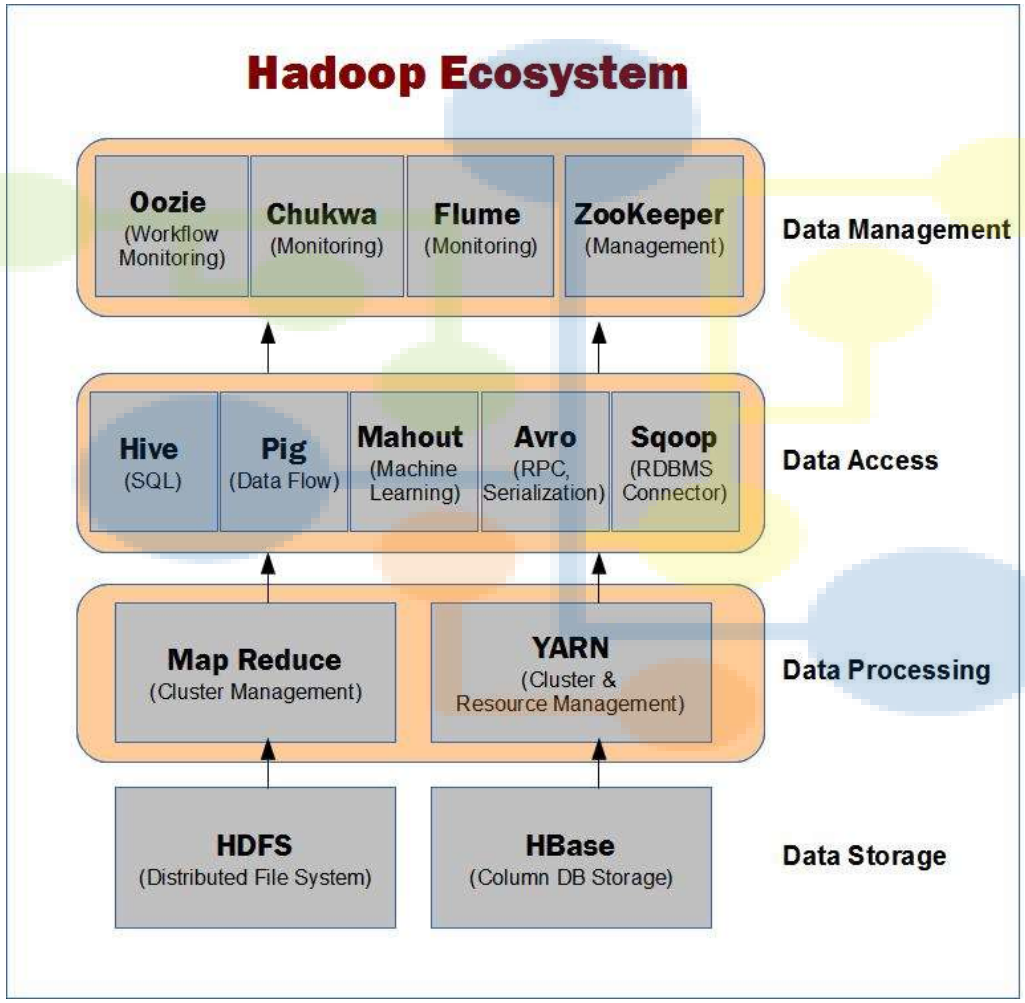
Ecossistema Hadoop

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Outros Projetos Importantes



Ecossistema Hadoop

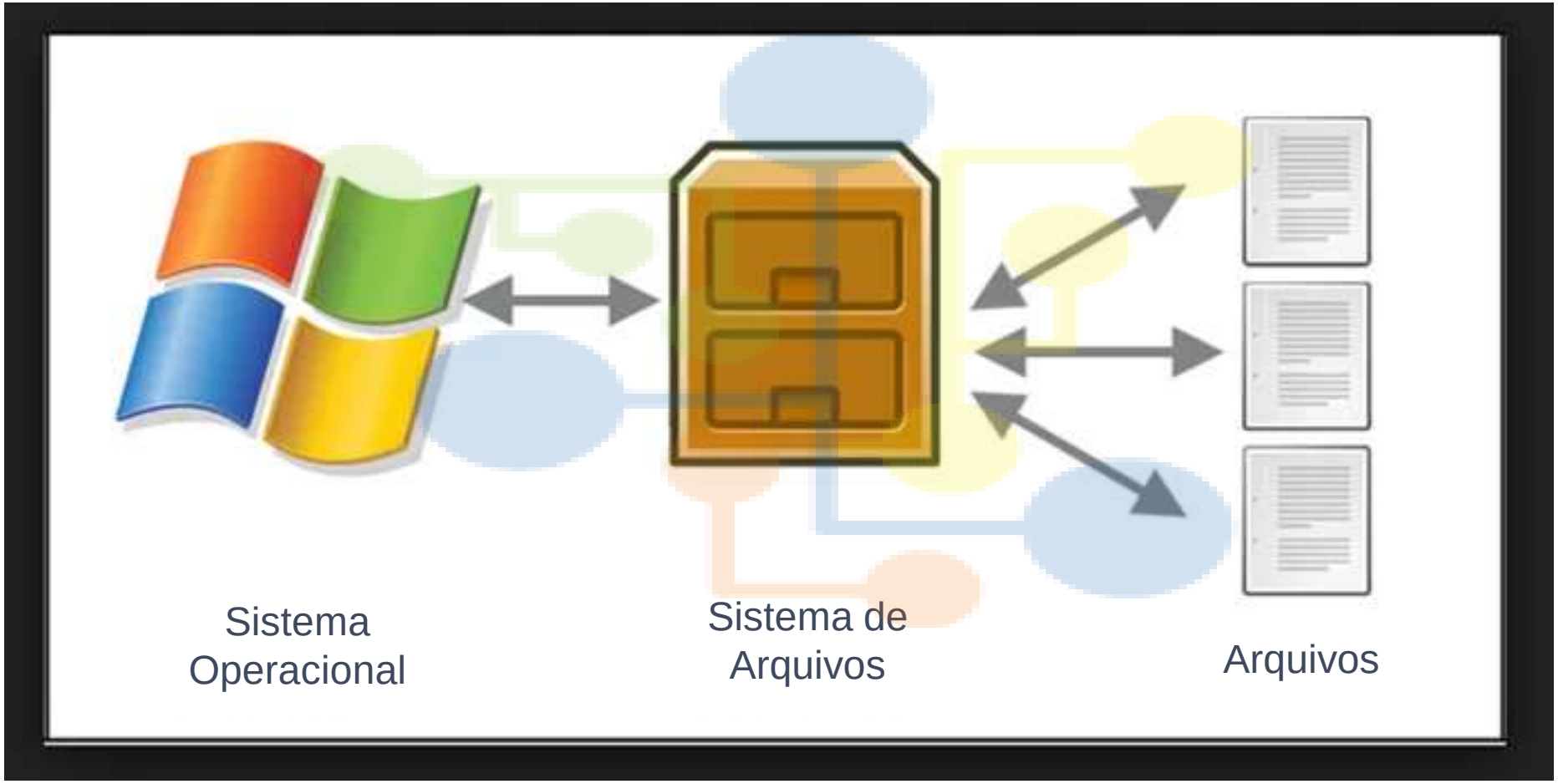


Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

HDFS (Hadoop Distributed File System)
Conceito e Importância

HDFS – Conceito e Importância

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



HDFS – Conceito e Importância

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

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

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

HDFS – Conceito e Importância

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Sistema de Arquivos Distribuído

HDFS – Conceito e Importância

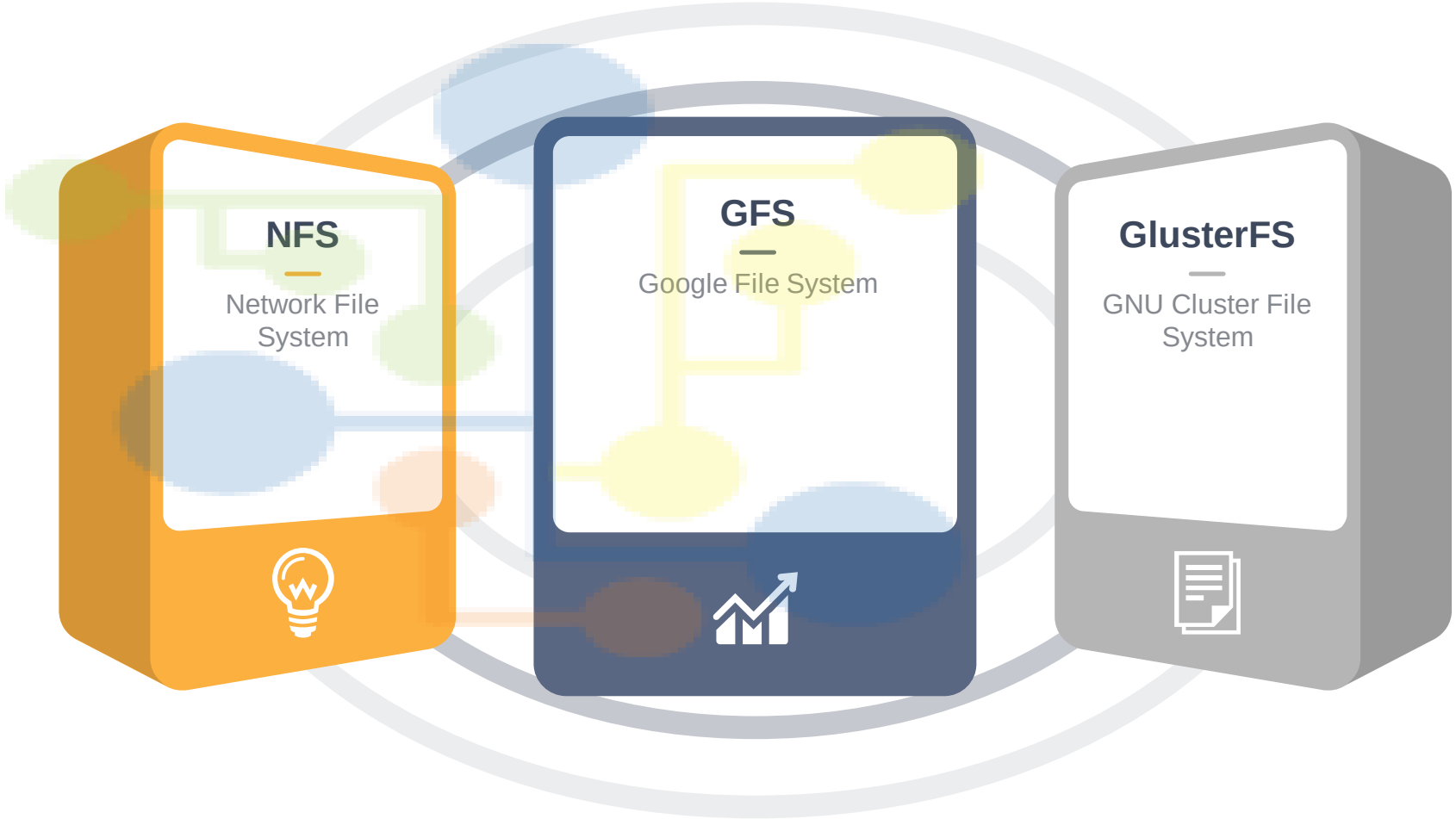
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

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

HDFS – Conceito e Importância

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

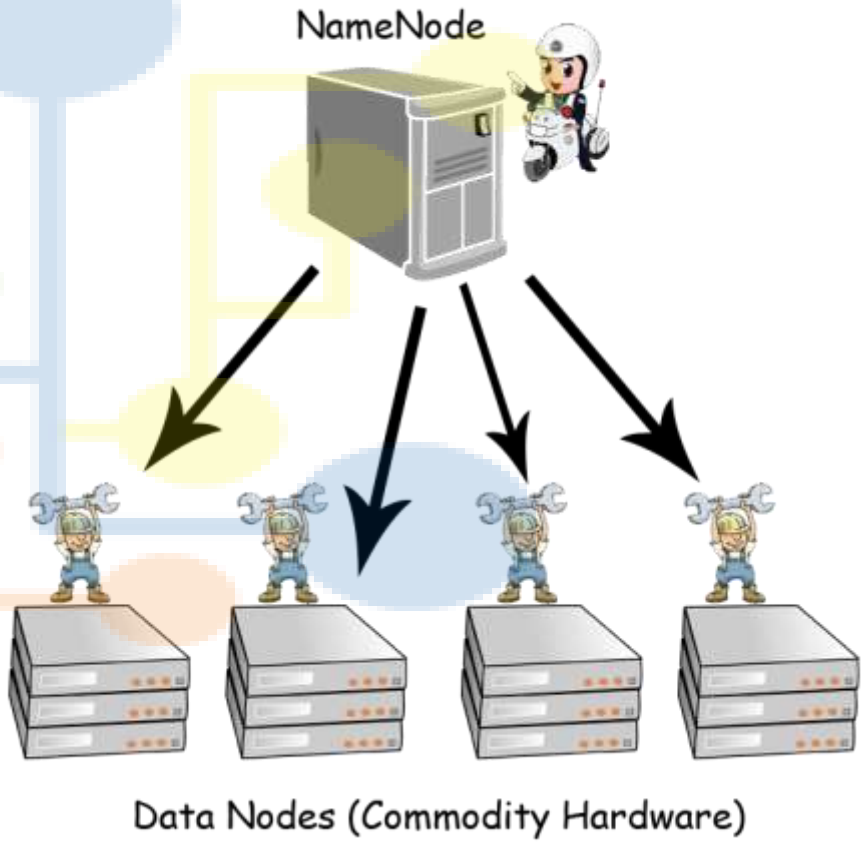
Outros Sistemas de Arquivos Distribuídos



HDFS – Conceito e Importância

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

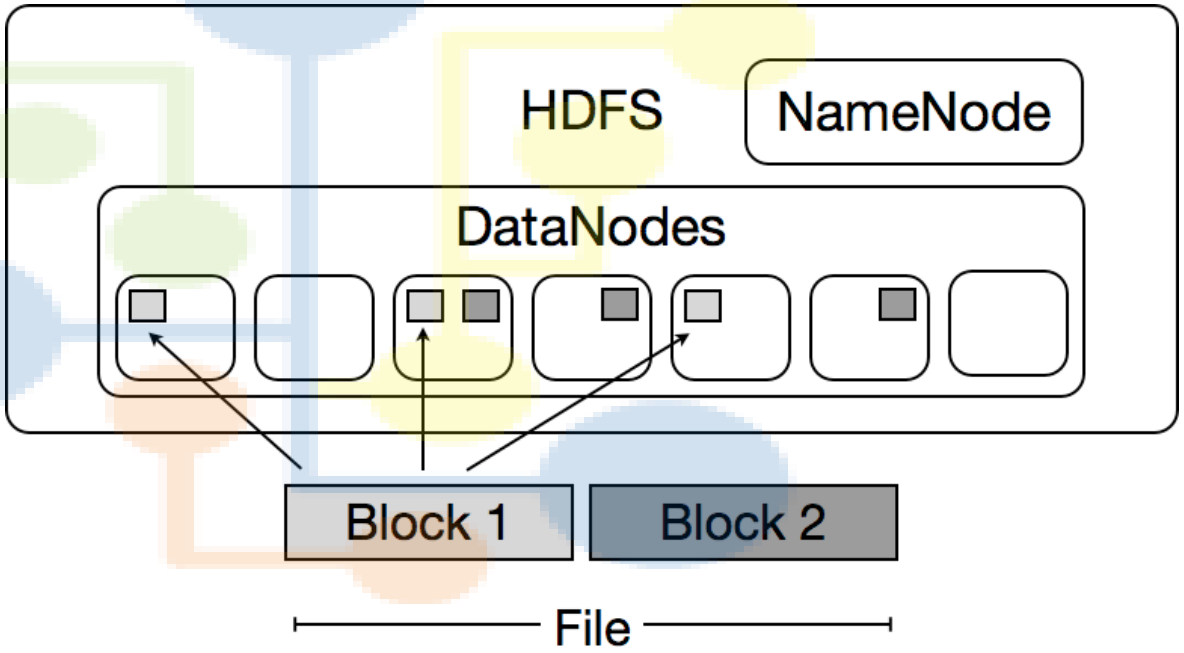
Hadoop Distributed File System



HDFS – Conceito e Importância

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

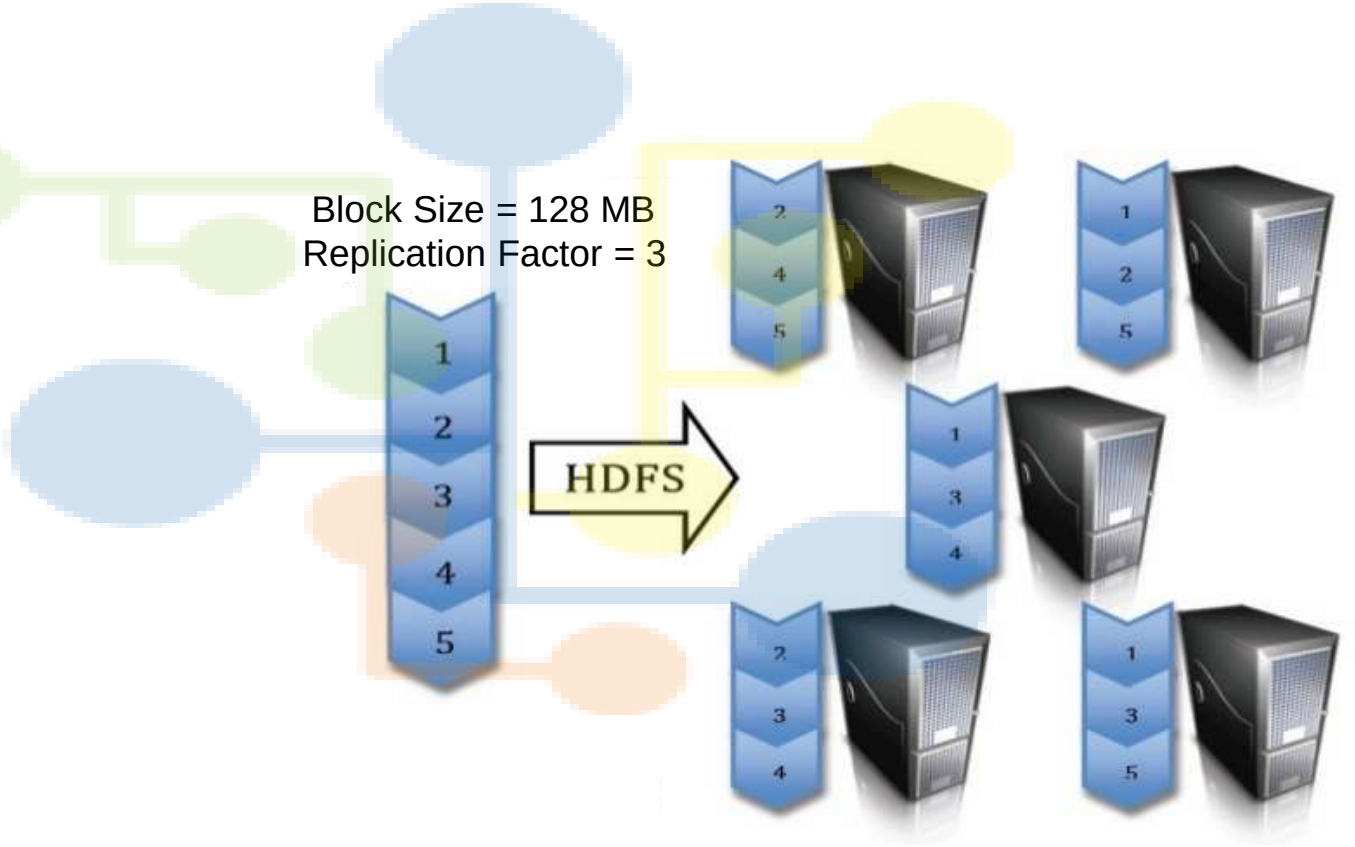
Hadoop Distributed File System



HDFS – Conceito e Importância

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Hadoop Distributed File System



HDFS – Conceito e Importância

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

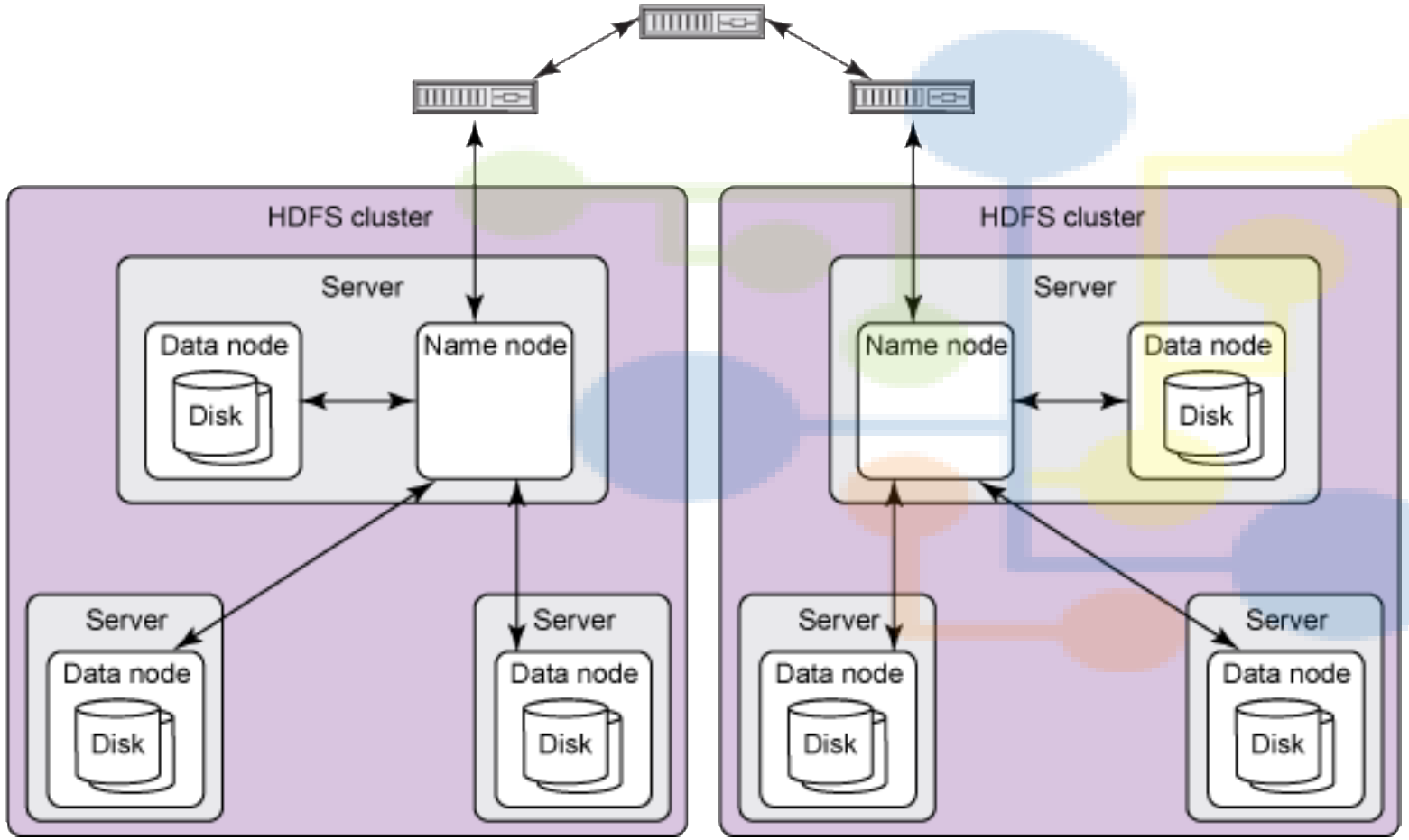
O HDFS foi criado para resolver ”Big Problems” e por isso seu funcionamento e arquitetura são próprios para se trabalhar com grandes arquivos de dados e distribuir esses arquivos em blocos ao longo de um cluster de computadores, para que possam ser processados em paralelo.

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

HDFS (Hadoop Distributed File System)
Arquitetura

HDFS – Arquitetura

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



Arquitetura Master/Worker

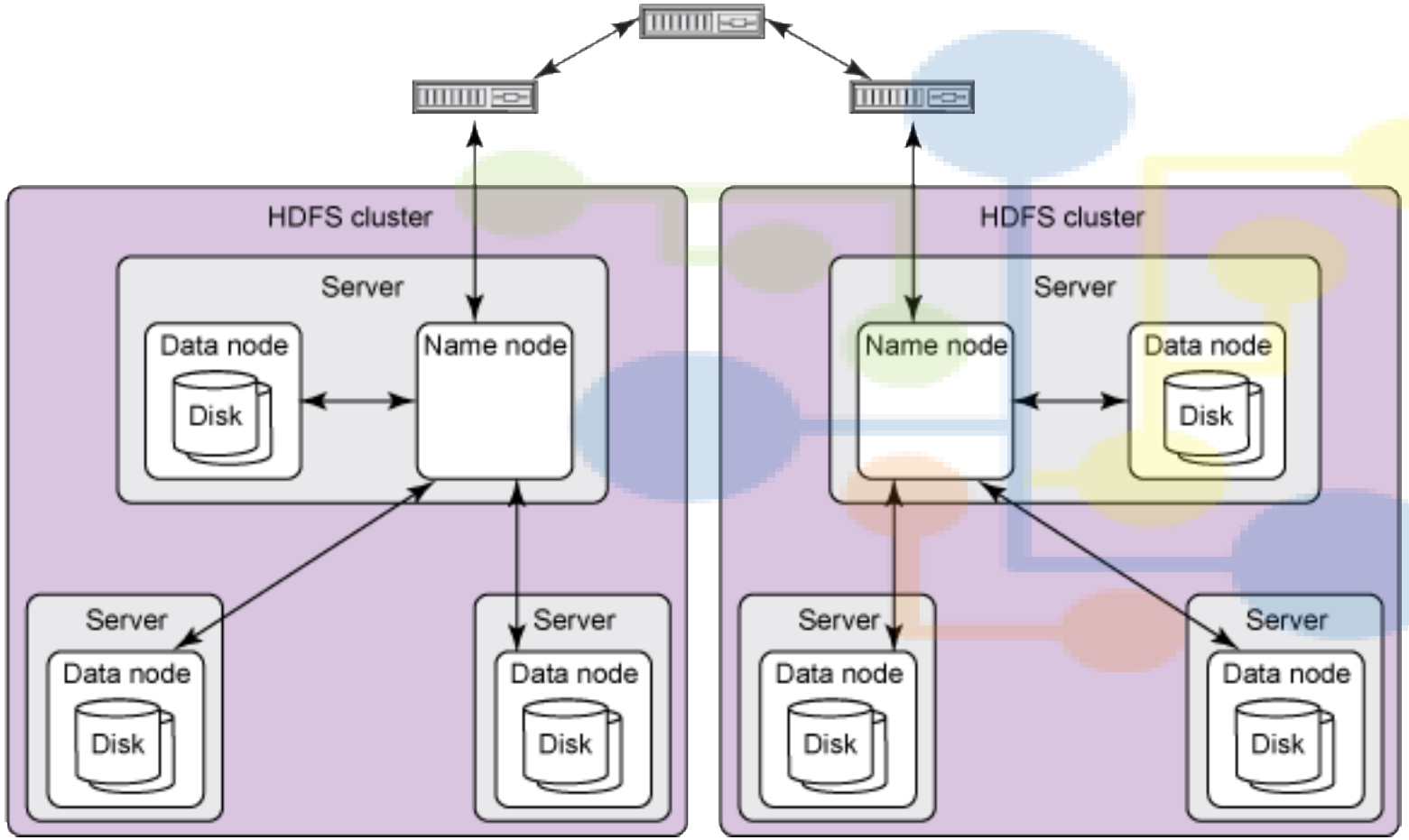
HDFS – Arquitetura

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Arquitetura
Master/Worker

HDFS – Arquitetura

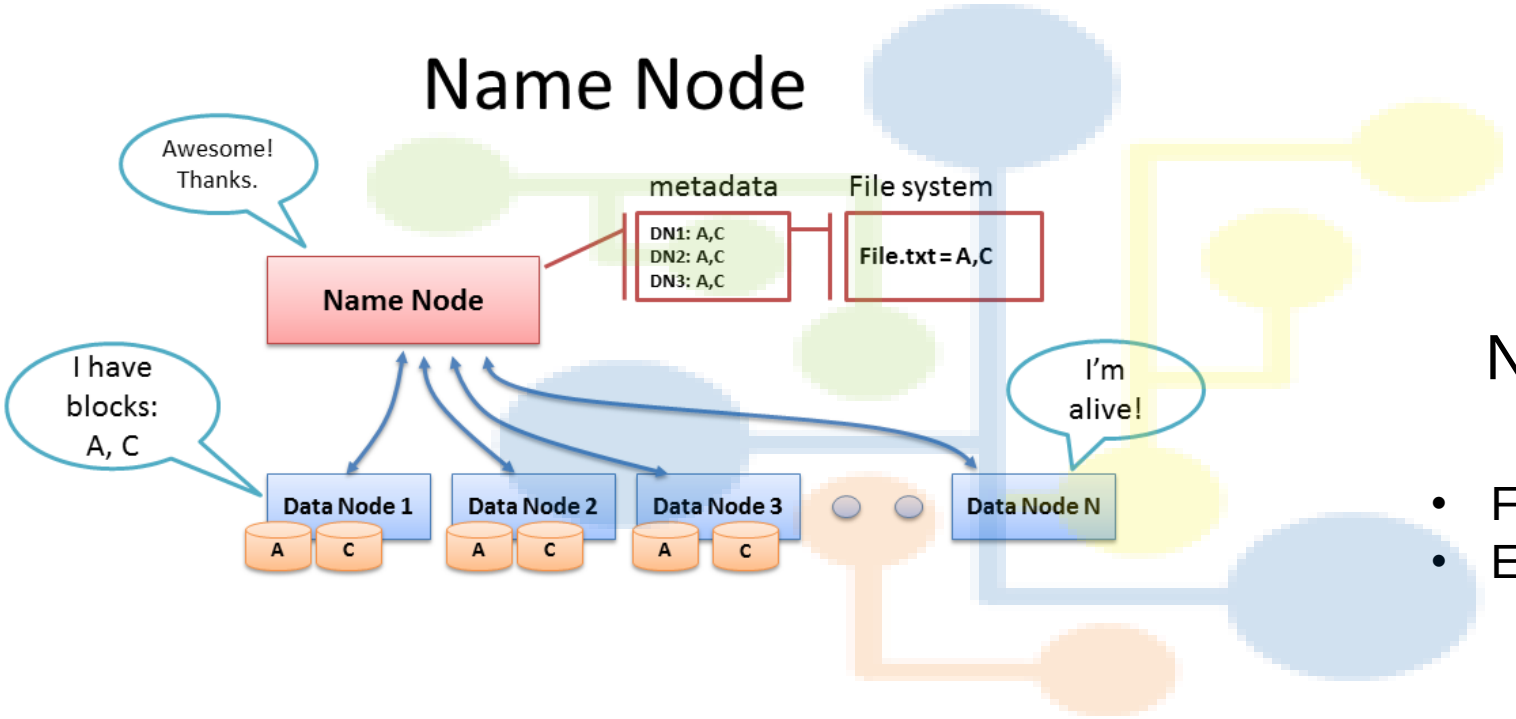
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



Arquitetura Master/Worker

HDFS – Arquitetura

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

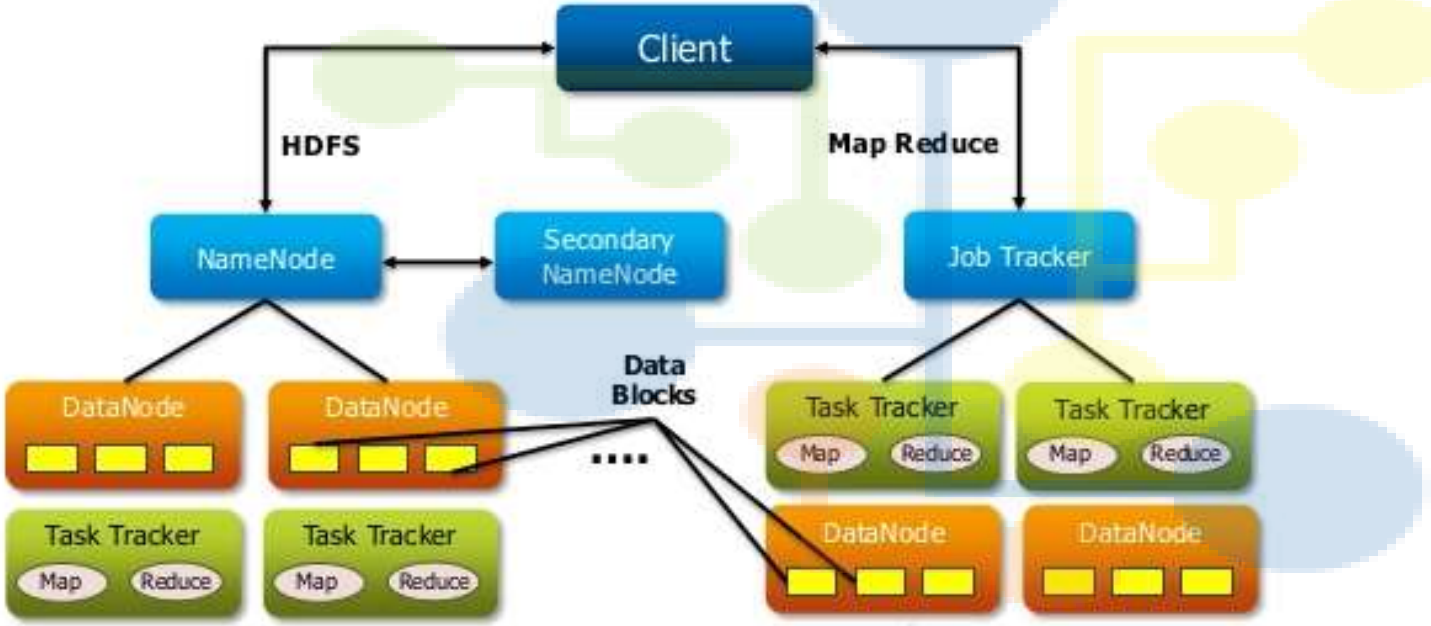


NameNode

- FsImage
- EditLog

HDFS – Arquitetura

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

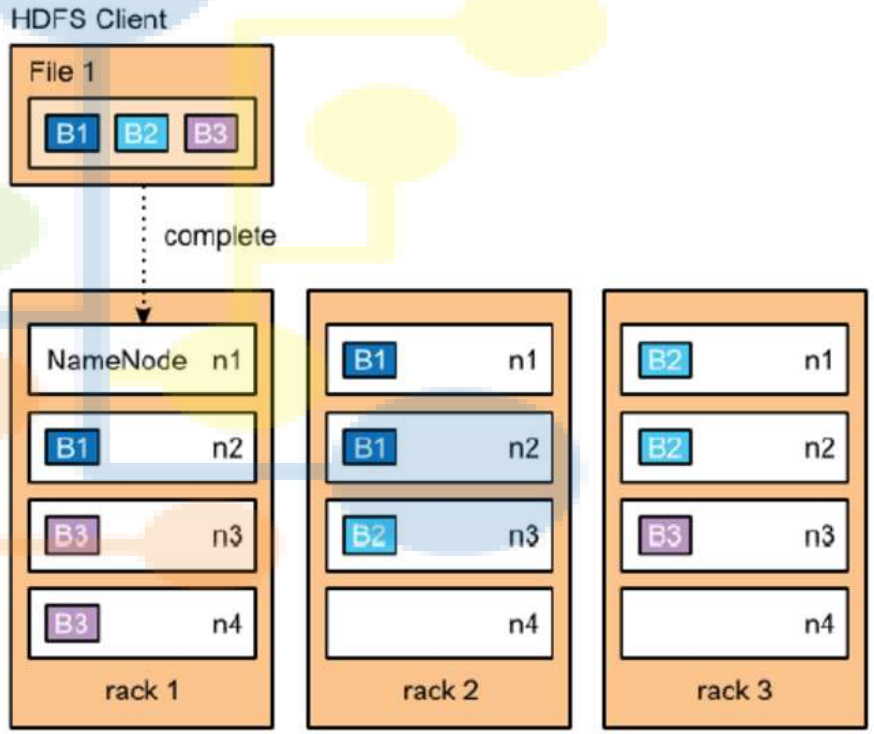


NameNode

- FsImage
- EditLog

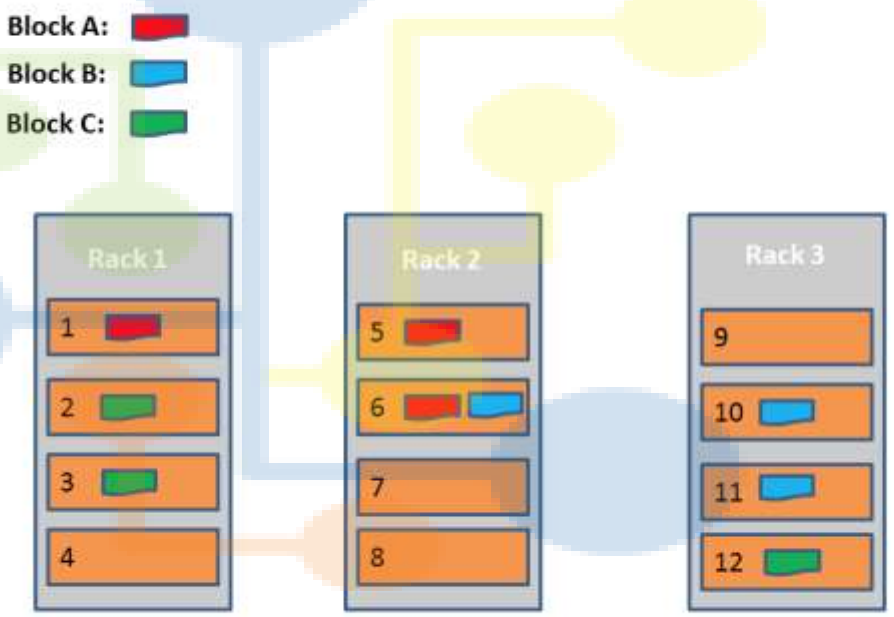
HDFS – Arquitetura

Replicação



HDFS – Arquitetura

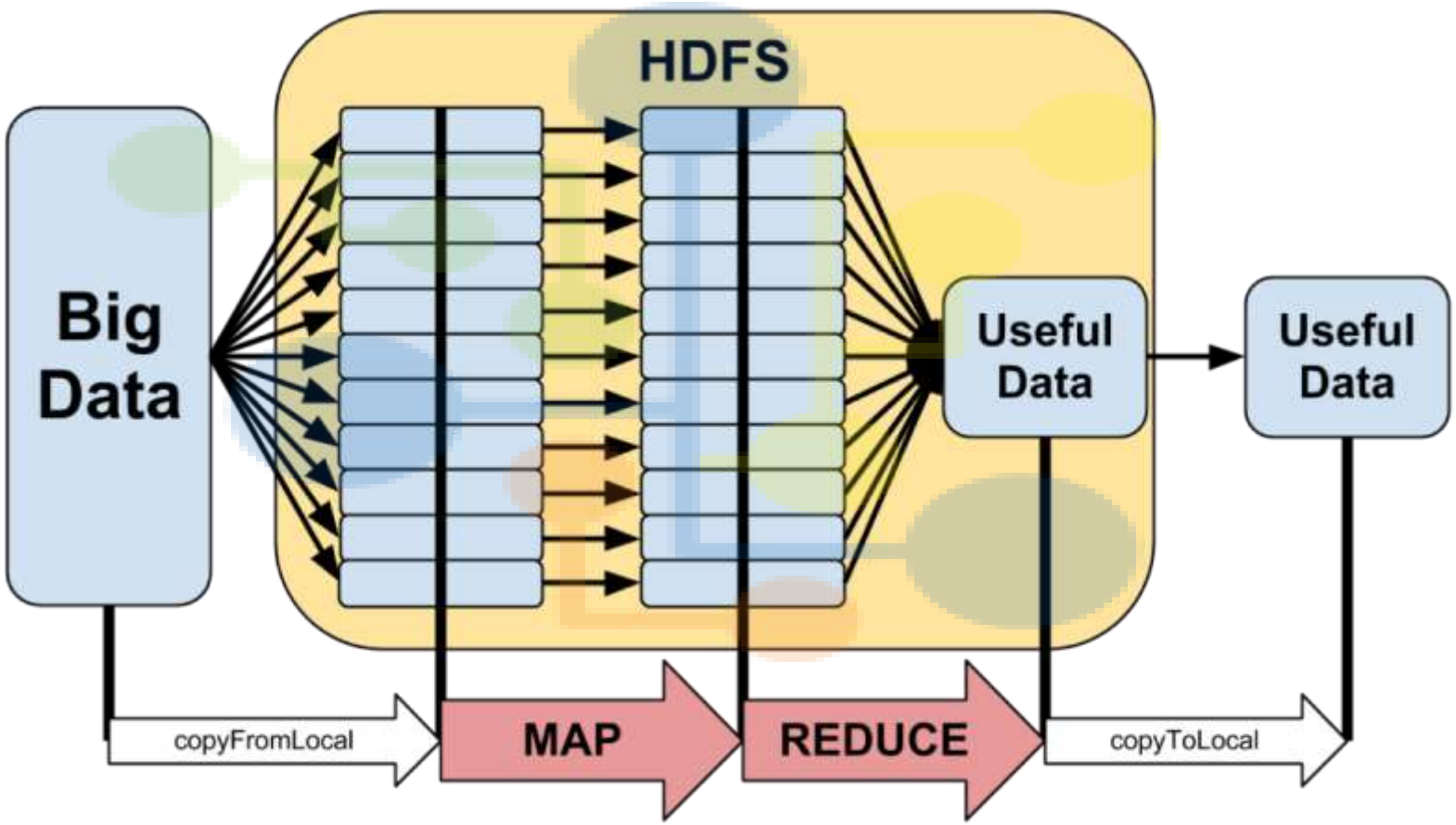
Replicação



Definindo MapReduce

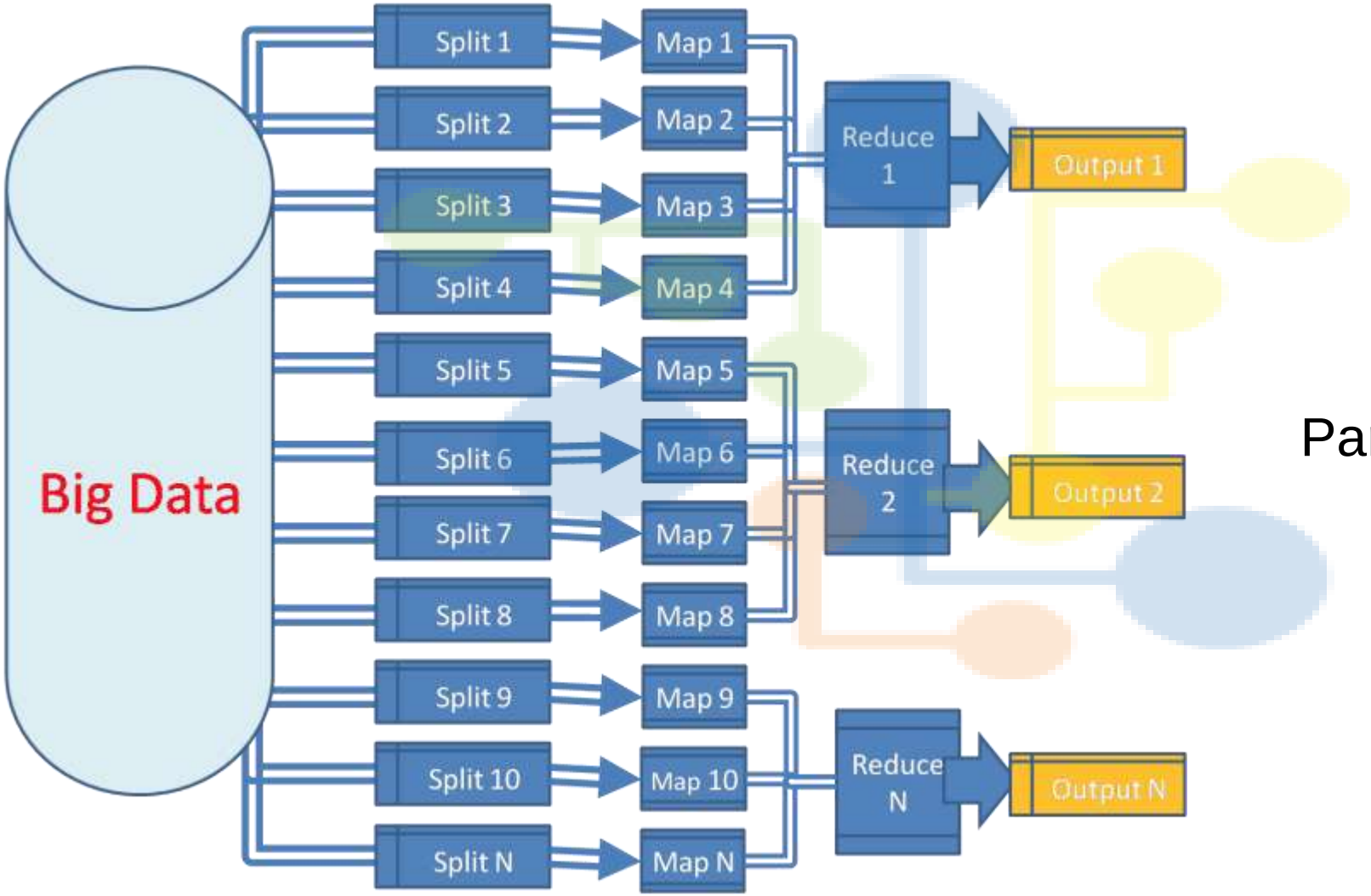
Definindo MapReduce

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



Definindo MapReduce

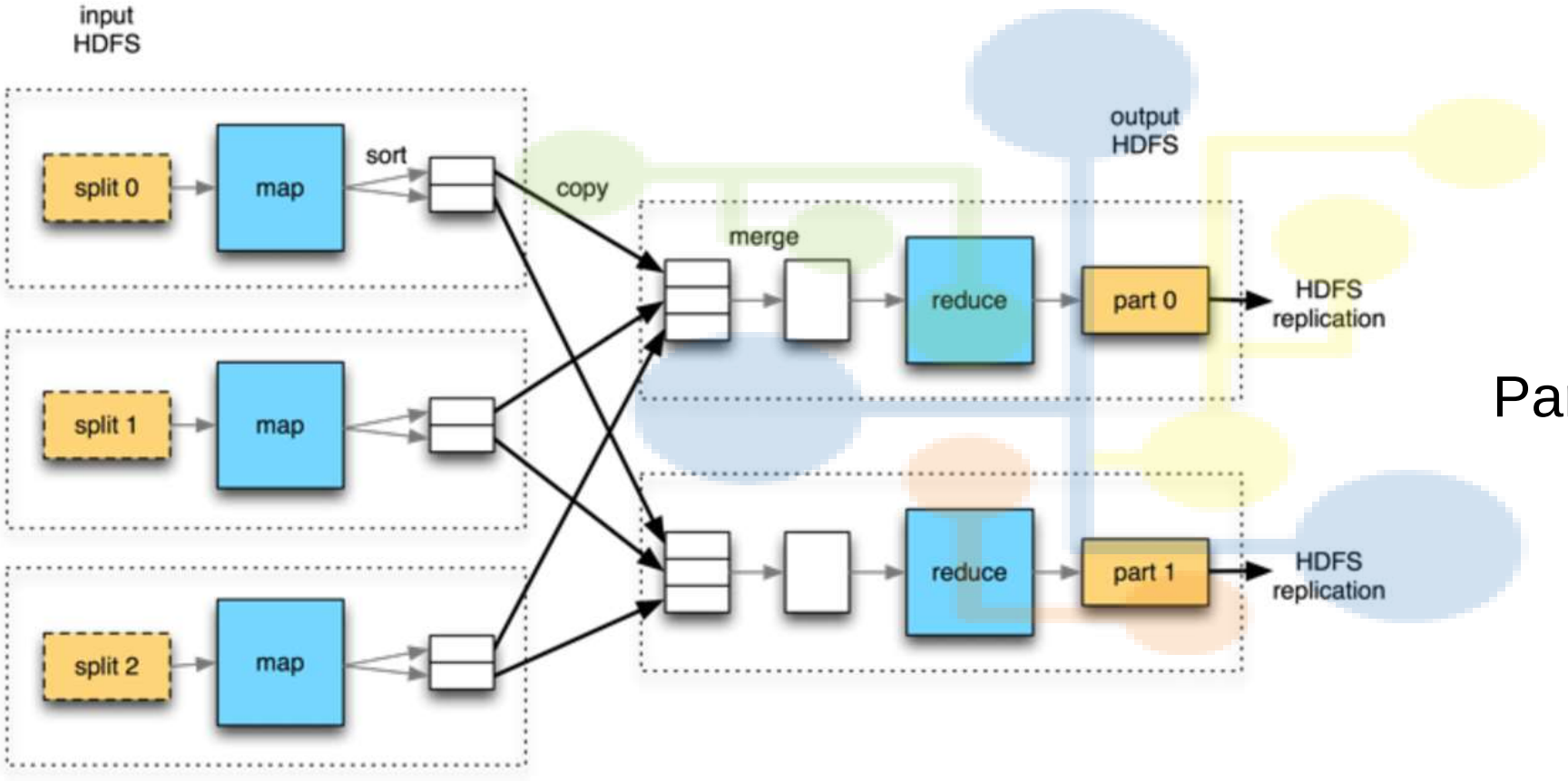
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



Processamento
Paralelo e Distribuído

Definindo MapReduce

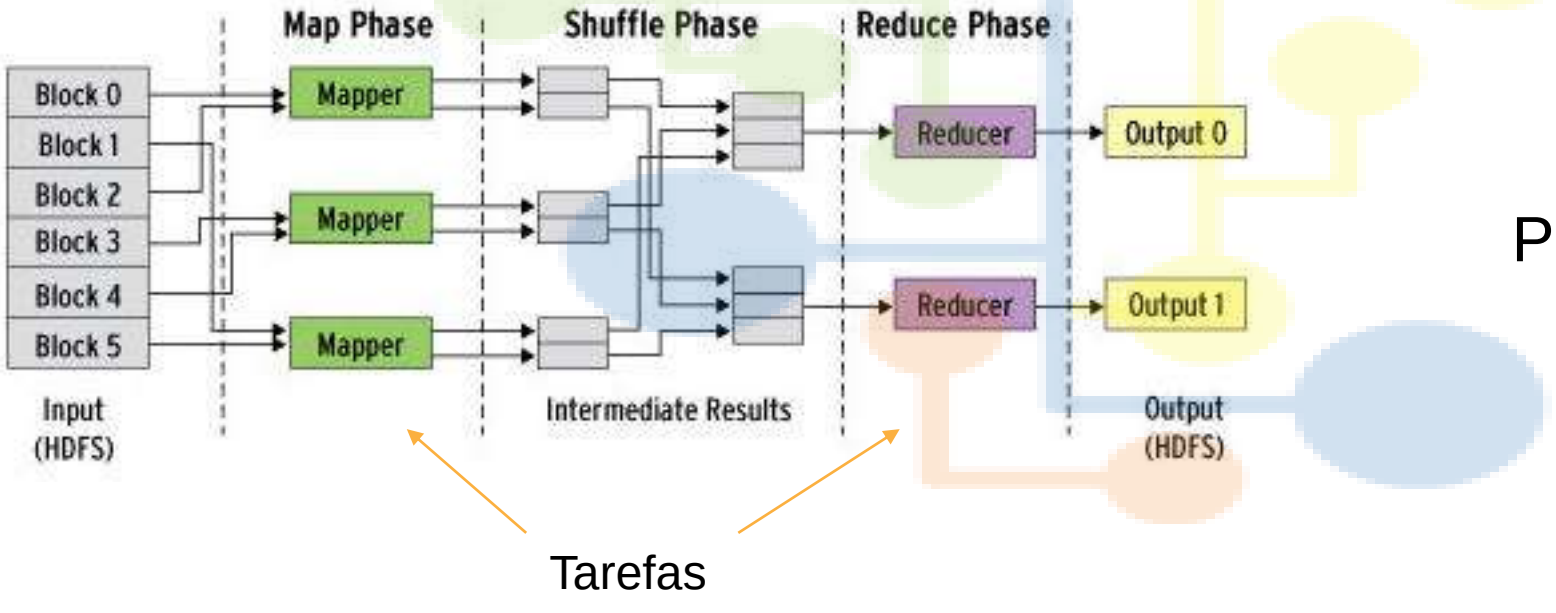
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



Processamento
Paralelo e Distribuído

Definindo MapReduce

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



Processamento
Paralelo e Distribuído

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Hadoop
x
Bancos de Dados Relacionais

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Bancos de Dados Relacionais

Hadoop x Bancos de Dados Relacionais

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

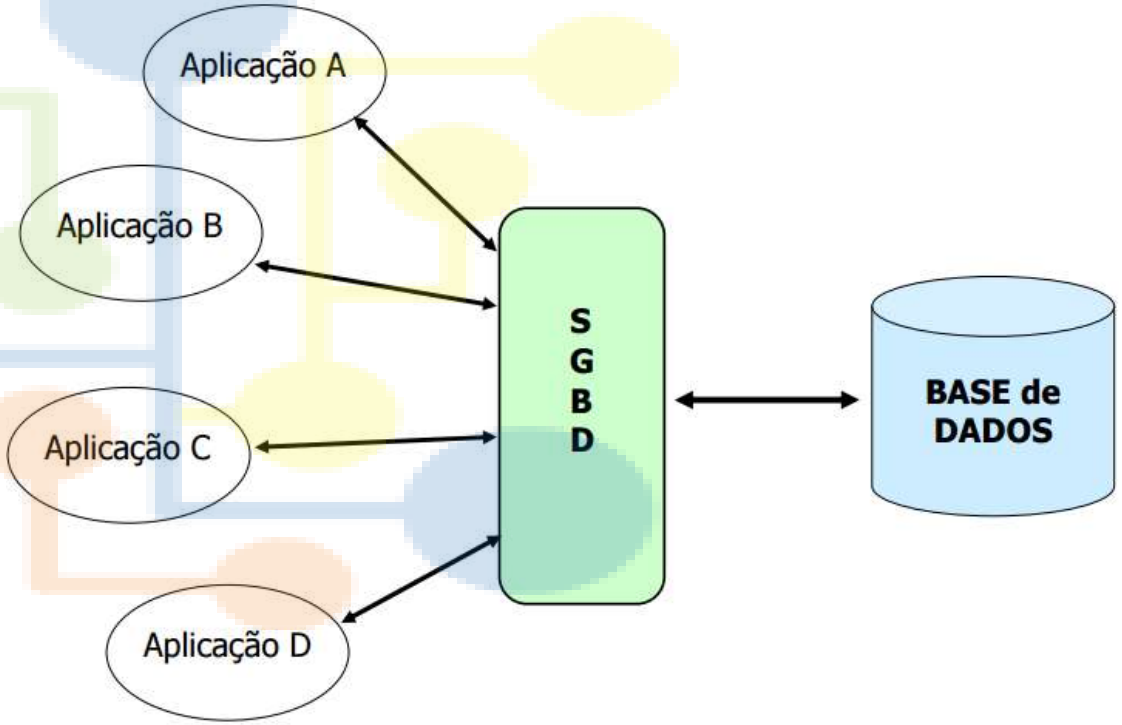
RDBMS
Relational Database
Management Systems

Hadoop x Bancos de Dados Relacionais

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

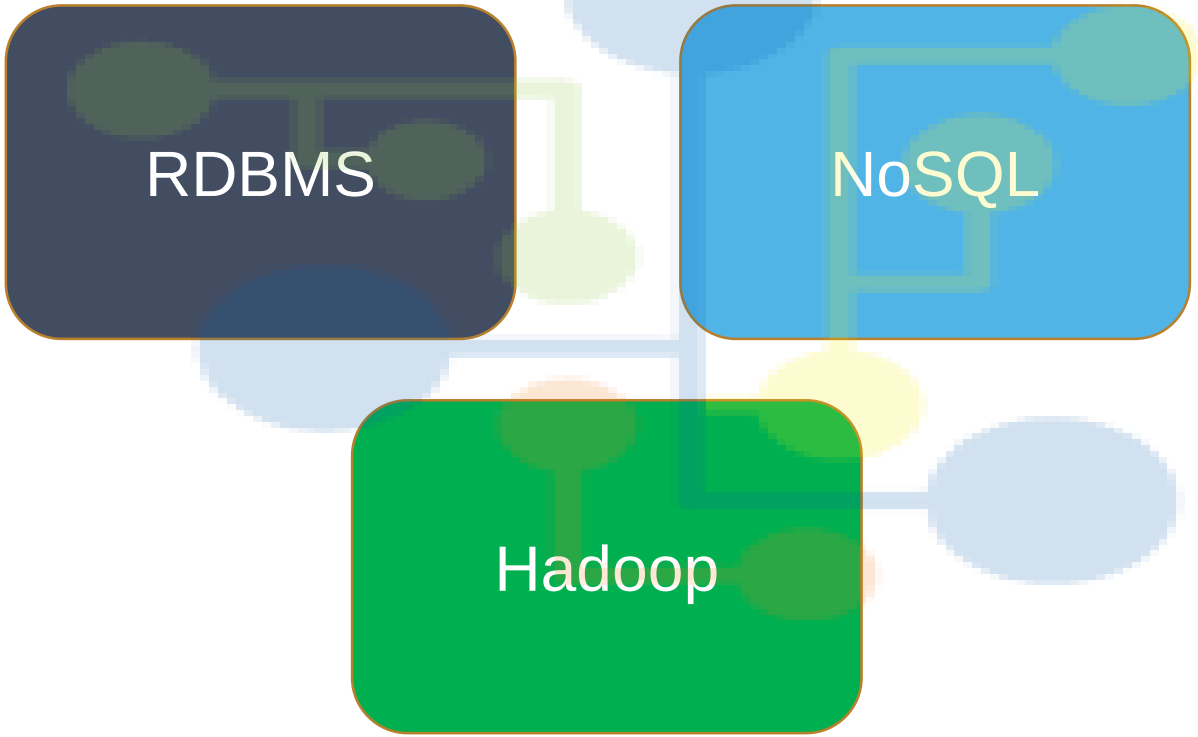
SGBD’s

Gerenciam um ou mais bancos de dados



Hadoop x Bancos de Dados Relacionais

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



Hadoop x Bancos de Dados Relacionais

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

SQL

Hadoop x Bancos de Dados Relacionais

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

ACID

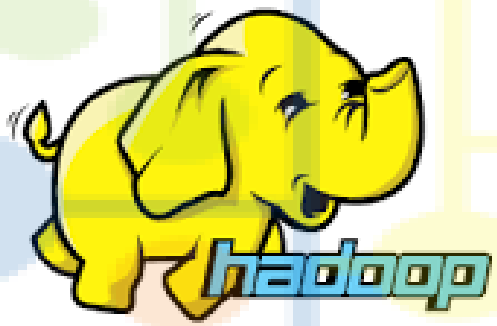
Atomicidade
Consistência
Isolamento
Durabilidade

Hadoop x Bancos de Dados Relacionais

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Hadoop x Bancos de Dados Relacionais

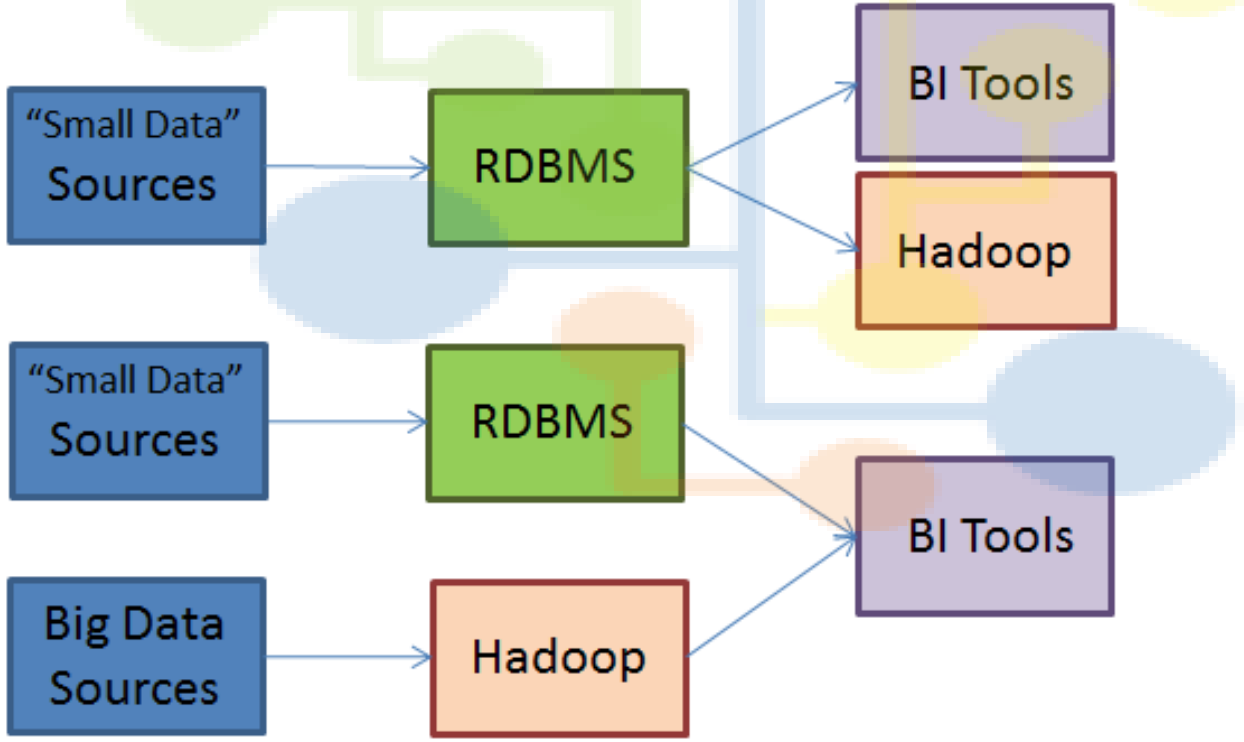
Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



Hadoop x Bancos de Dados Relacionais

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

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

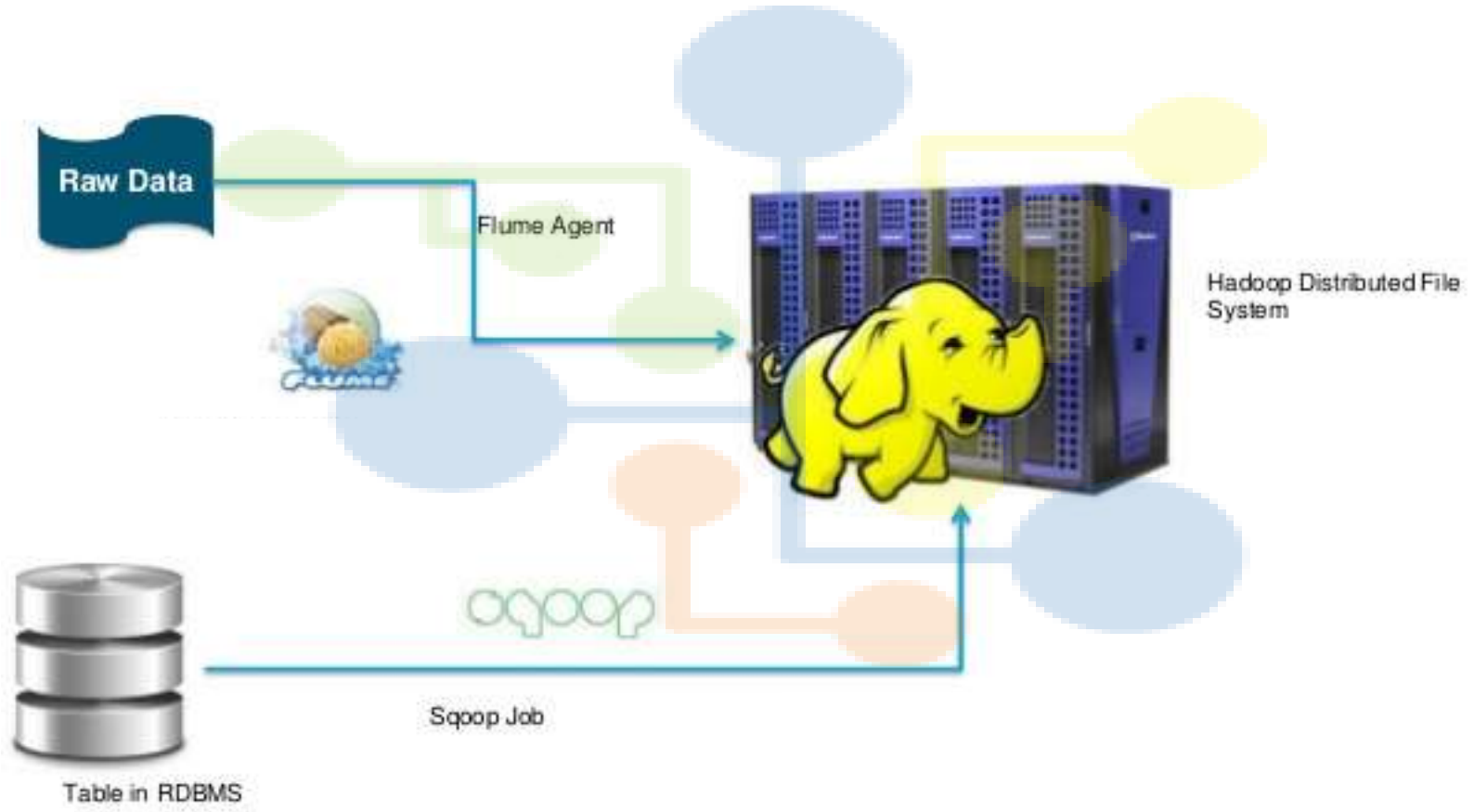


Hadoop x Bancos de Dados Relacionais

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

Hadoop x Bancos de Dados Relacionais

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e



Hadoop x Bancos de Dados Relacionais

Data Science Academy ericgpti@gmail.com 5b1700f85e4cde813d8b459e

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