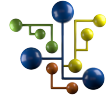


Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde819d18b450

# Big Data Real-Time Analytics com Python e Spark



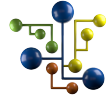


Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

**Seja muito bem-vindo(a)!**

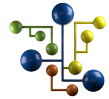


Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

Este é o Segundo Curso da  
Formação Cientista de Dados



Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

## Estrutura do Curso

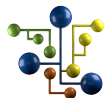
Parte 1  
Análise de  
Dados com  
Python

Parte 2  
Estatística

Parte 3  
Machine  
Learning

Parte 4  
Spark





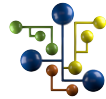
Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

## O que não veremos neste curso?

**Conteúdo Básico de Python**  
**Configuração do Cluster Spark**



Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

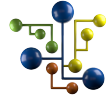
## Quais ferramentas iremos usar?

**Anaconda Python**

**Jupyter Notebook**

**Apache Spark**

**Ambiente em Cloud da Databricks**

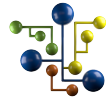


Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

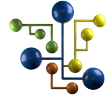




# Big Data Real-Time Analytics com Python e Spark

**Cada mini-projeto será apresentado com sua especificação, documentação e scripts.**





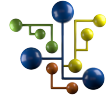
Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

**Avaliação Final**

**50 questões – 3 tentativas – 70%**



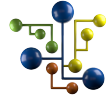
Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

**Bonus**

**Usando Linguagem R e Spark**



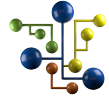
Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

**Pré-requisito**

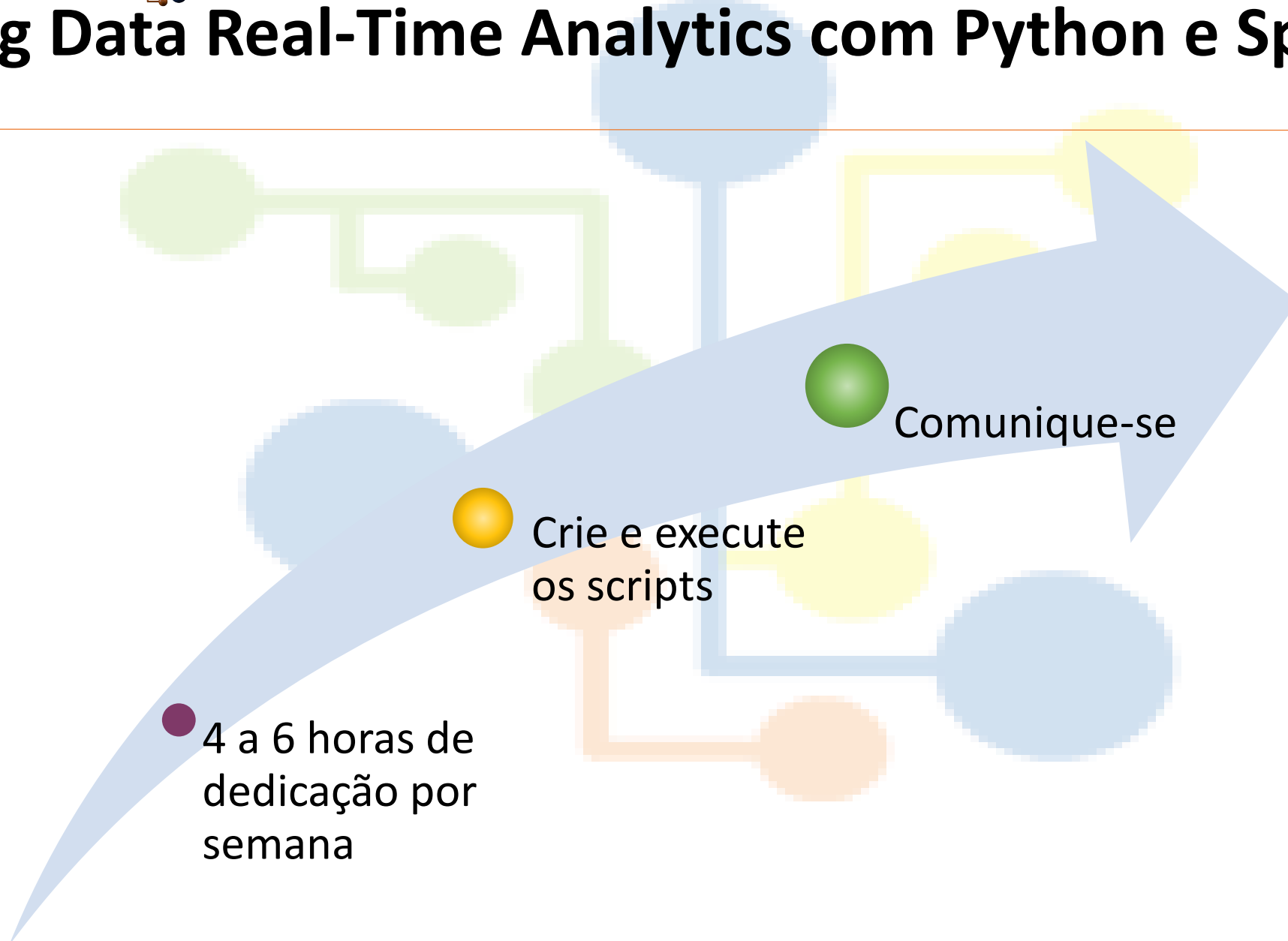
**Python Fundamentos Para Análise de Dados**

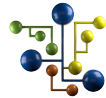


Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark





# Big Data Real-Time Analytics com Python e Spark

## Objetivos ao final deste curso:

- Desenvolver habilidades de processamento e análise de dados em tempo real.
- Aprender técnicas de Machine Learning e Processamento de Dados.
- Compreender os conceitos do ciclo de vida de projetos de Big Data Analytics.
- Aplicar o conhecimento deste curso em casos reais do dia a dia.
- Compreender a função da Estatística no processo de Data Science.

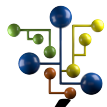


# O que é o Apache Spark?

O que é o Apache Spark?







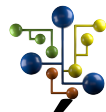
Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# O que é o Apache Spark?

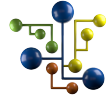


Spark



# O que é o Apache Spark?

Spark é atualmente o projeto open-source mais ativo ligado a Big Data. Embora ele seja considerado o sucessor do Hadoop/MapReduce, veremos que não é bem assim e as duas tecnologias podem ser usadas em conjunto. Mas o Spark realmente oferece vantagens.

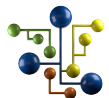


Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

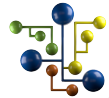
## Principais Benefícios do Apache Spark



# Principais Benefícios do Apache Spark

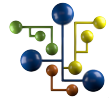
**O Spark oferece 3 benefícios principais:**

- **Fácil de usar**
- **Veloz**
- **Engine de uso geral**



# Principais Benefícios do Apache Spark

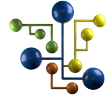
**Apache Spark é uma plataforma de computação em cluster (conjunto de computadores), criado para ser veloz e de uso geral, sendo ideal para processamento iterativo e processamento de streaming de dados (fluxo contínuo de dados).**



# Principais Benefícios do Apache Spark

**Spark realiza a computação em memória (o que ajuda a explicar sua velocidade), mas também é eficiente quando executa aplicações em disco.**





Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

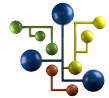
## Principais Benefícios do Apache Spark



# Principais Benefícios do Apache Spark

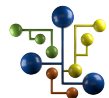
**O Spark oferece 3 benefícios principais:**

- **Fácil de usar**
- **Veloz**
- **Engine de uso geral**



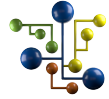
# Principais Benefícios do Apache Spark

**Apache Spark é uma plataforma de computação em cluster (conjunto de computadores), criado para ser veloz e de uso geral, sendo ideal para processamento iterativo e processamento de streaming de dados (fluxo contínuo de dados).**



# Principais Benefícios do Apache Spark

**Spark realiza a computação em memória (o que ajuda a explicar sua velocidade), mas também é eficiente quando executa aplicações em disco.**

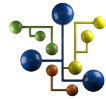


Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

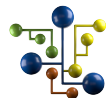
Por que Python e Spark?



# Por que Python e Spark?

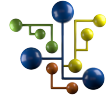
Embora o Spark tenha sido desenvolvido em Java e Scala, as habilidades da linguagem Python para manipulação e análise de dados, além de Machine Learning, a tornam a ferramenta ideal para trabalhar com Spark!





# Por que Python e Spark?

Caso queira aprender Spark com Scala, temos o curso Machine Learning com Linguagem Scala e Spark aqui na DSA.



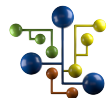
Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

Apache Spark Workflow

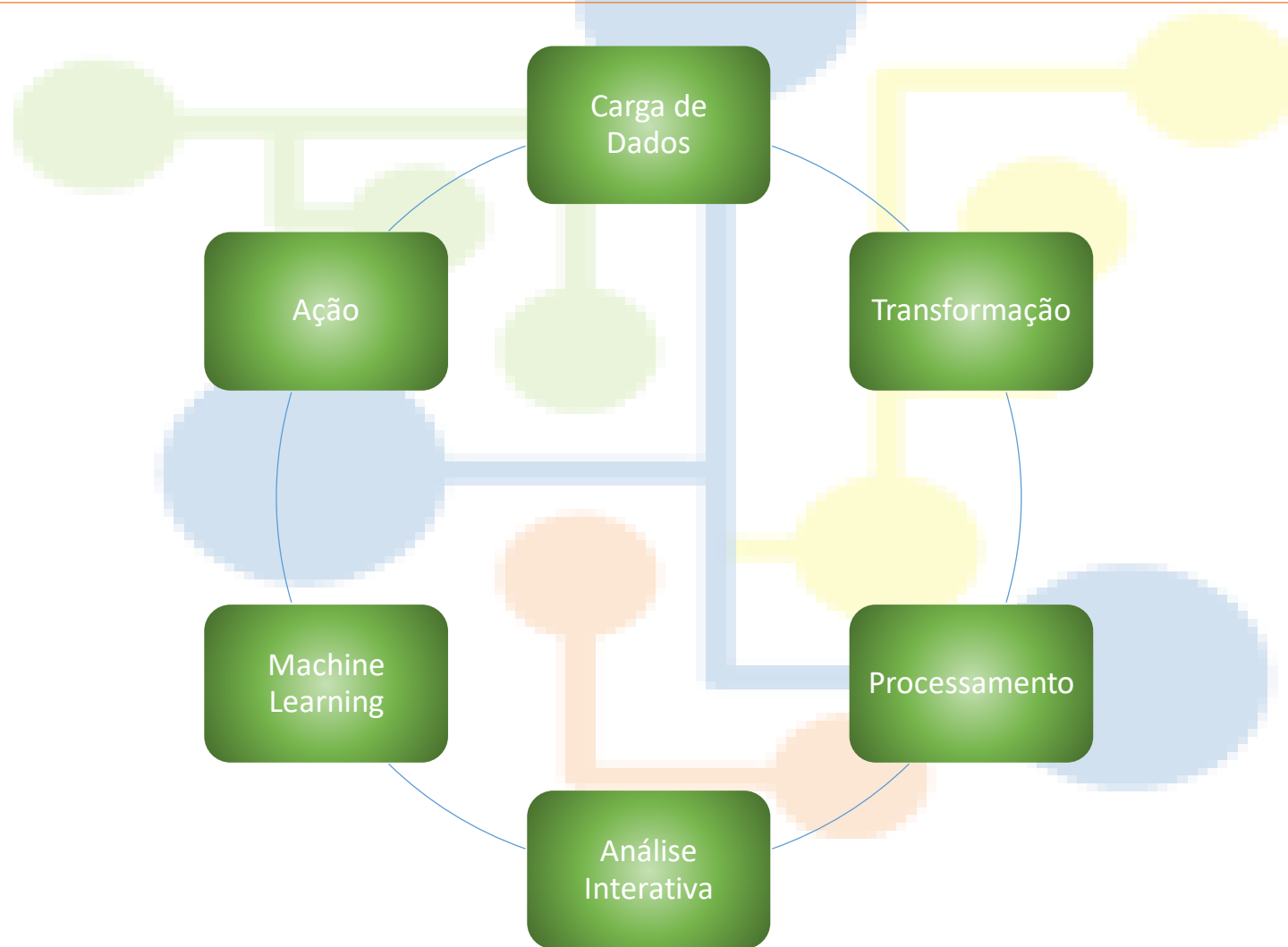


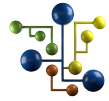


Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Apache Spark Workflow



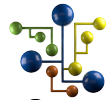


Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

Real-Time Analytics e  
Computação Distribuída

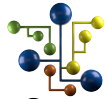


Data Science  
Academy

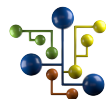
Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Real-Time Analytics e Computação Distribuída









Data Science  
Academy

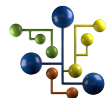
Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Real-Time Analytics e Computação Distribuída

## Big Data

Desafio 1  
Coletar os dados

Desafio 2  
Analisar em Tempo Real



# Real-Time Analytics e Computação Distribuída

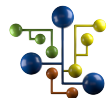
Big Data é definido pelos seus V's.

Volume

Variedade

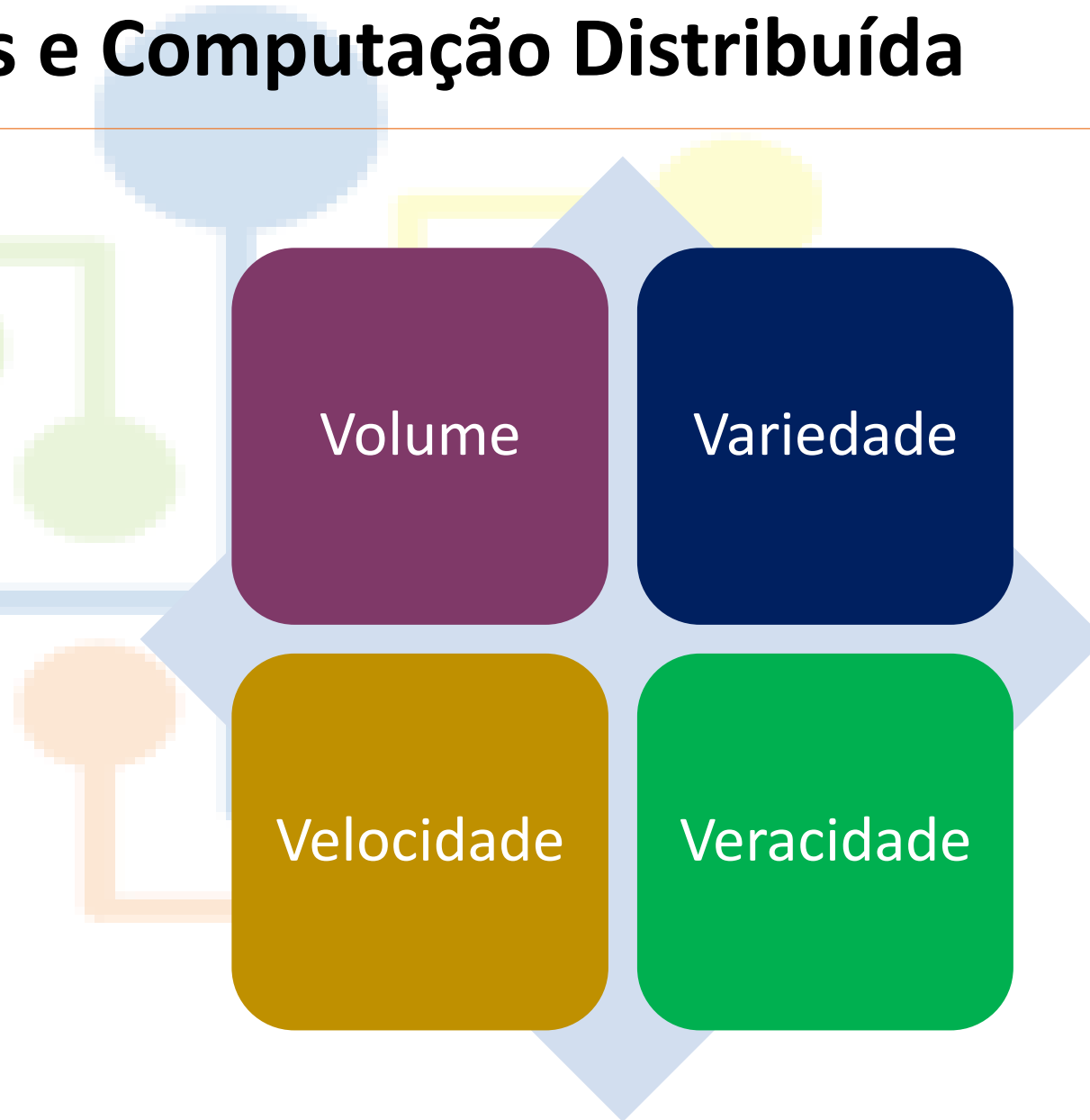
Velocidade

Veracidade



# Real-Time Analytics e Computação Distribuída

Neste curso vamos focar no V de velocidade do Big Data, ou seja, velocidade com que os dados são gerados.





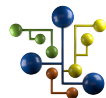
Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

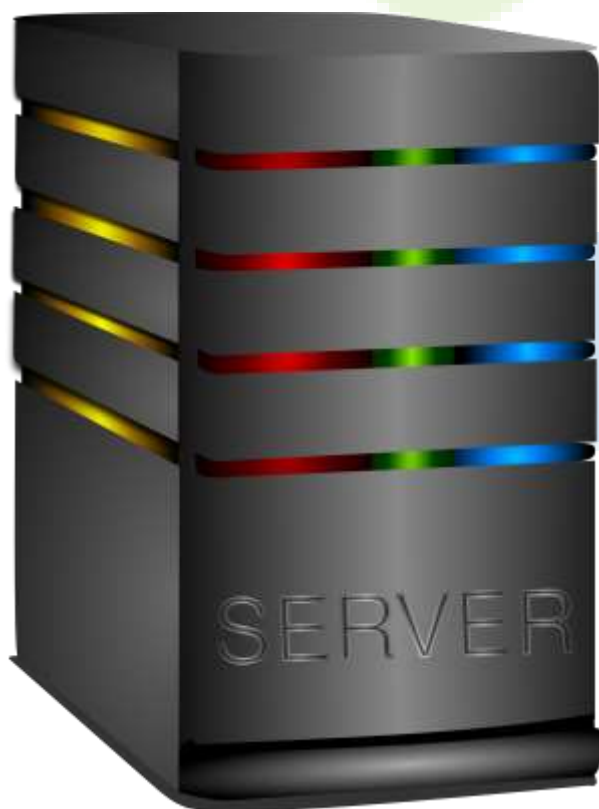
# Real-Time Analytics e Computação Distribuída

A large, faint background diagram showing a complex network of nodes and connections. The nodes are colored in shades of blue, green, yellow, and orange, and are connected by lines of the same colors, forming a web-like structure.

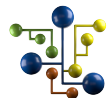
## De onde vem o Big Data?



# Real-Time Analytics e Computação Distribuída



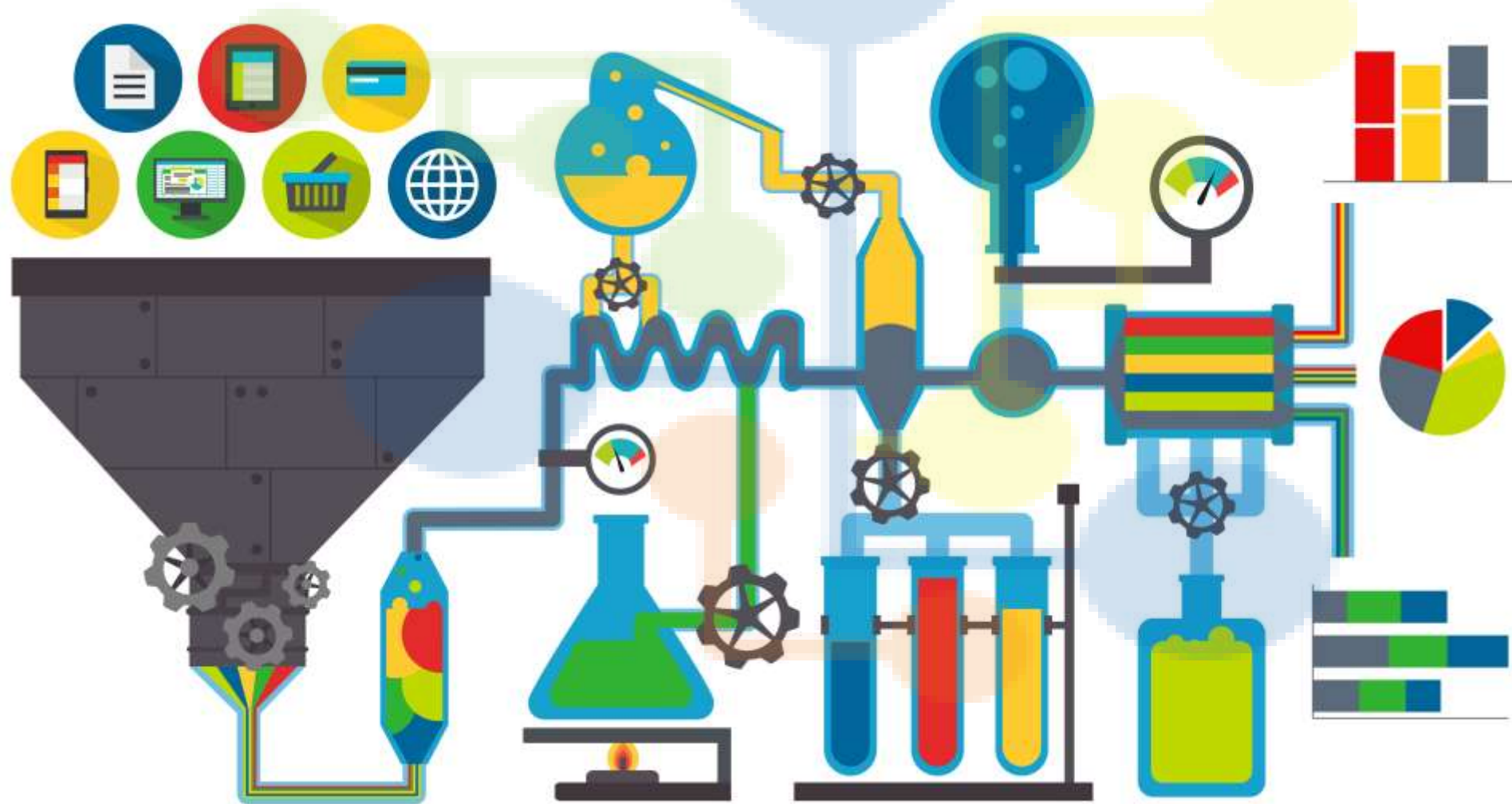
Até alguns anos atrás, a única forma de analisar todo este conjunto de dados seria através de soluções analíticas que eram executadas em apenas um computador com grande capacidade computacional, um servidor!

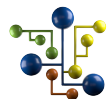


Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Real-Time Analytics e Computação Distribuída





Data Science  
Academy

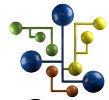
Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Real-Time Analytics e Computação Distribuída



**60 TB/dia**

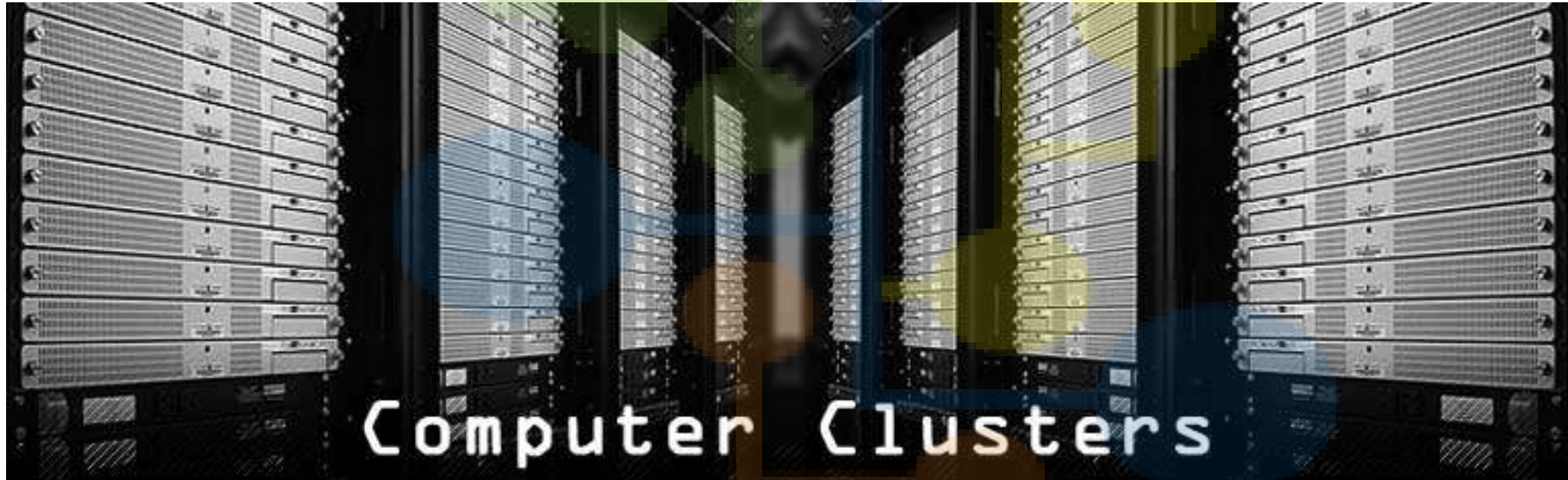




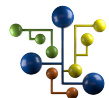
Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Real-Time Analytics e Computação Distribuída







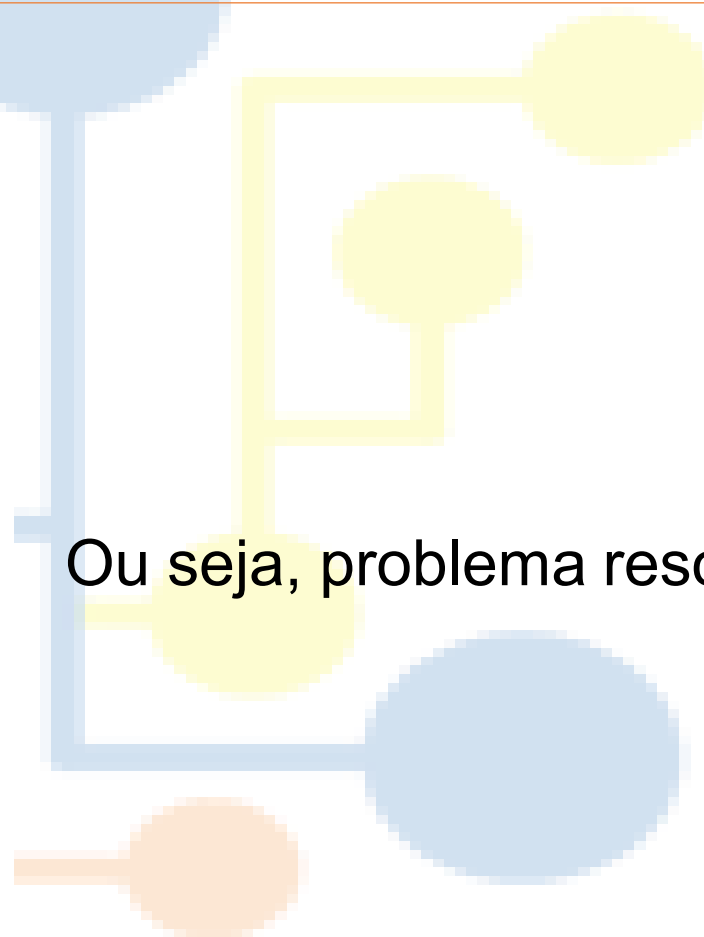
Data Science  
Academy

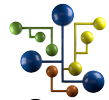
Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Real-Time Analytics e Computação Distribuída



Ou seja, problema resolvido???



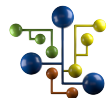


Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Real-Time Analytics e Computação Distribuída





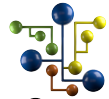
Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Real-Time Analytics e Computação Distribuída

## Computação Distribuída



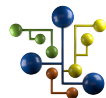


Data Science  
Academy

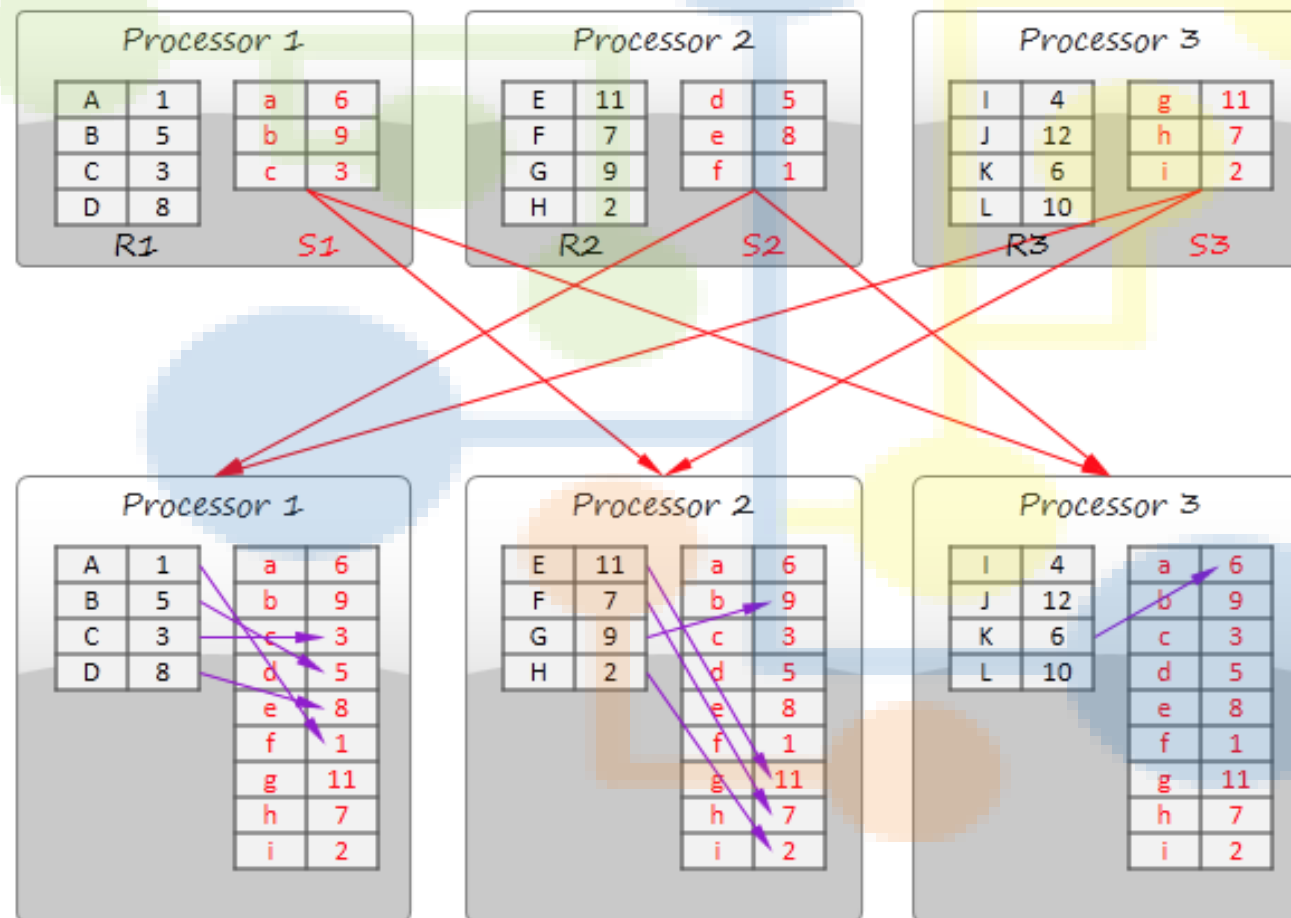
Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Real-Time Analytics e Computação Distribuída

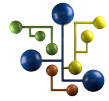
**Hadoop e Spark foram pensados  
para Computação Distribuída!**



# Real-Time Analytics e Computação Distribuída



→ Replication



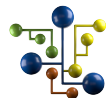
Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

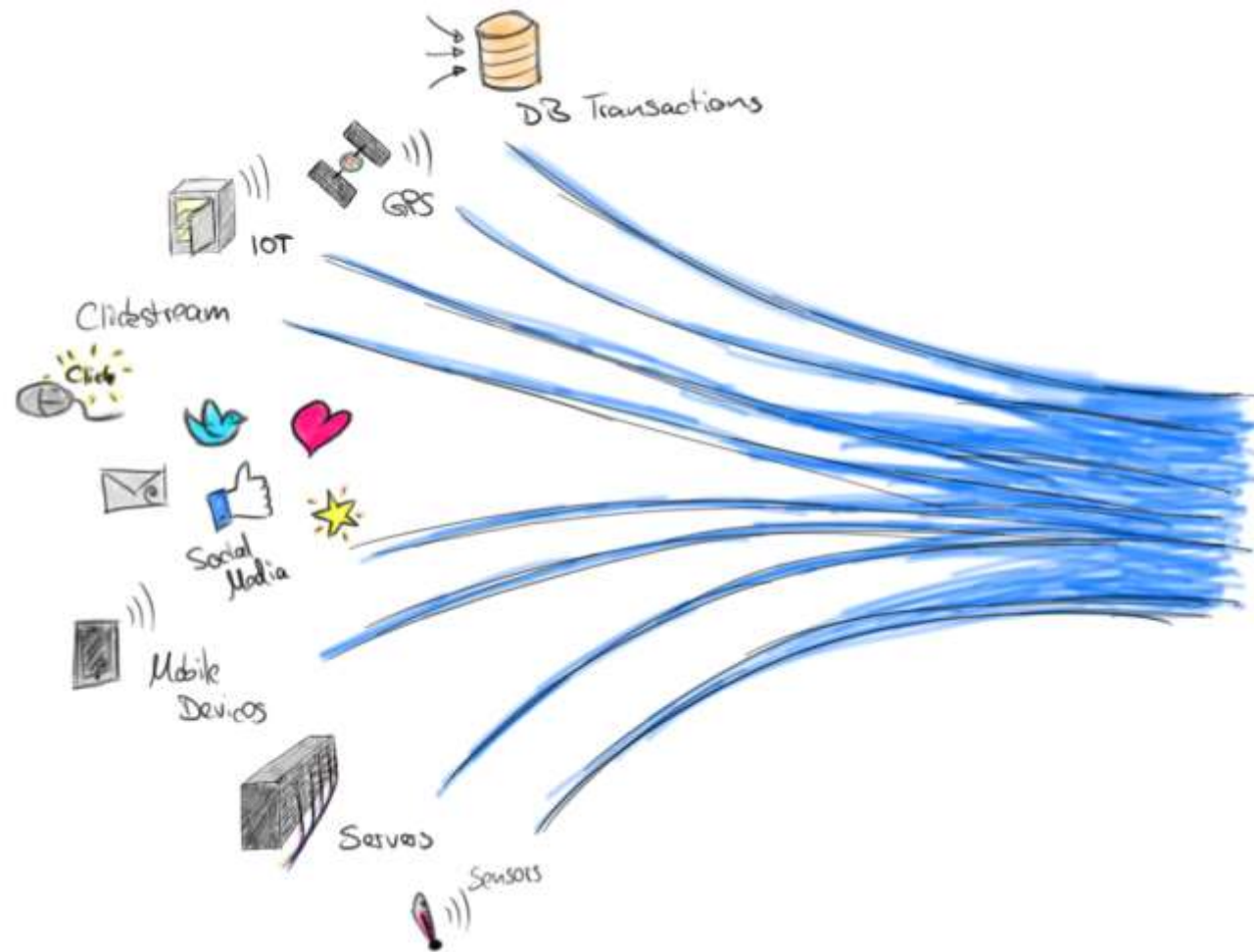
O Que é Streaming de Dados?

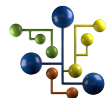




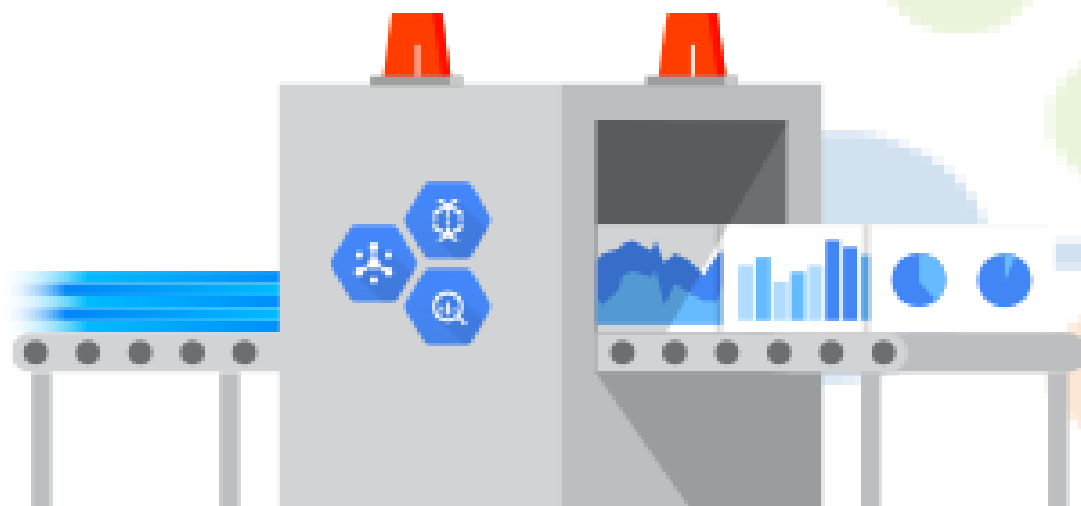
# O Que é Streaming de Dados?

Dados em streaming são dados gerados continuamente por milhares de fontes de dados, que geralmente enviam os registros de dados simultaneamente, em tamanhos pequenos (na ordem dos kilobytes).



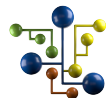


# O Que é Streaming de Dados?

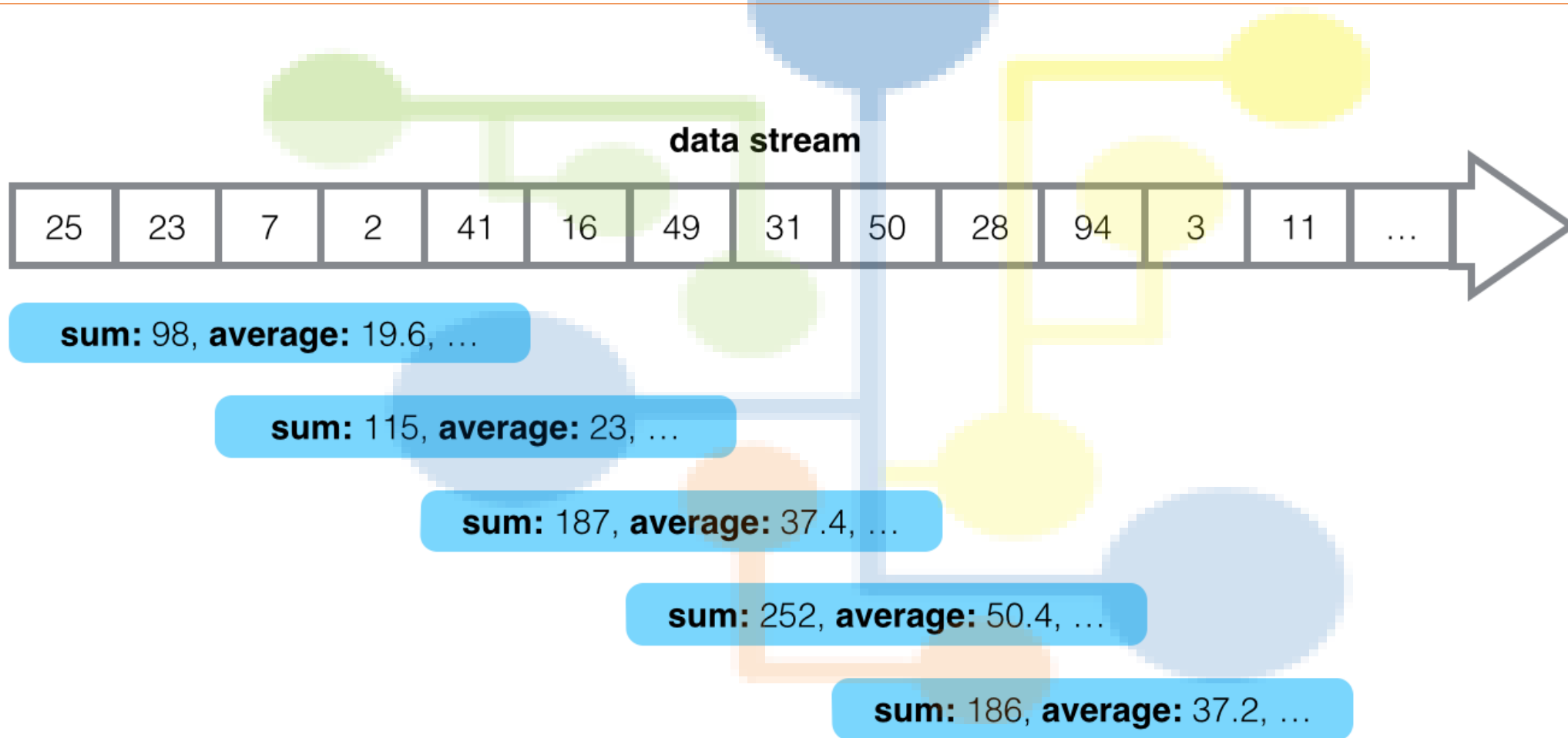


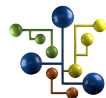
Os dados em streaming devem ser processados de maneira sequencial e incremental por registro e usados para uma ampla variedade de análises de dados, como correlações, agregações, filtragem e amostragem.



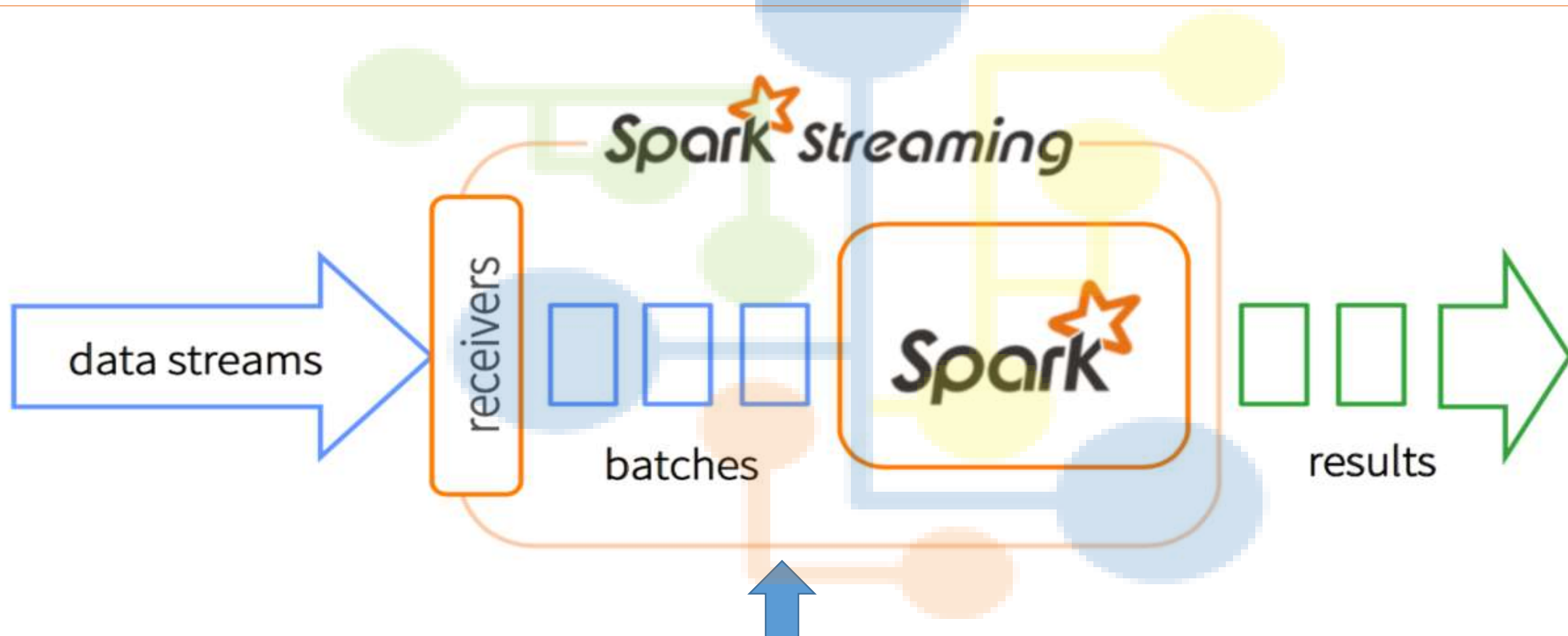


# O Que é Streaming de Dados?

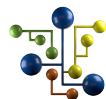




# O Que é Streaming de Dados?



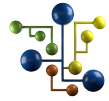
Analisar os dados enquanto eles estão sendo gerados!



# O Que é Streaming de Dados?

## Possíveis Fontes para o Spark Streaming:

- Sensores em veículos
- Monitoramento de cotação de ações na bolsa de valores
- Arquivos texto (no momento em que eles são gerados)
- Redes Sociais (Facebook, Twitter, Instagram)
- Dados de dispositivos móveis
- Cliques em web sites
- Apache Kafka
- Apache Flume



Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

Processamento de Lotes (Batch)

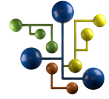
X

Processamento de Streaming de Dados



# Processamento de Lotes (Batch) x Processamento de Streaming de Dados

	Batch	Streams
Escopo de Dados	Consultas ou processamento de todos ou da maioria dos dados no conjunto de dados.	Consultas ou processamento de dados dentro de um período rotacional, ou apenas do registro de dados mais recente.
Tamanho dos Dados	Grandes lotes de dados.	Registros individuais ou microlotes compostos de alguns registros.
Desempenho	Latências em minutos ou horas.	Exige latência na ordem dos segundos ou milissegundos.
Análise	Análises de dados mais complexas.	Análises de dados menos complexas.



Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Big Data Real-Time Analytics com Python e Spark

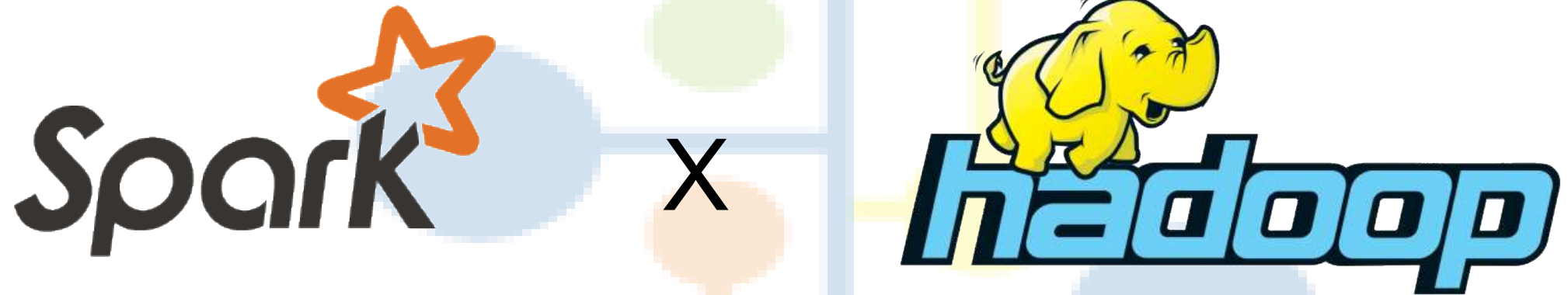
Apache Spark  
X  
Apache Hadoop

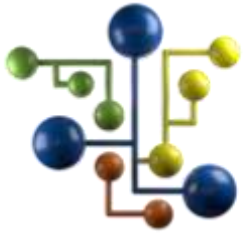


Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Spark x Hadoop



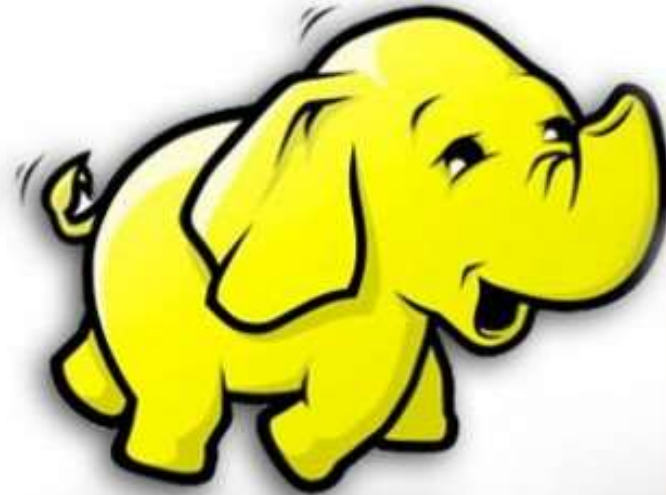


Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Spark x Hadoop

## Hadoop







Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Spark x Hadoop

**Mas e quando o volume de dados  
não for tão grande?**



Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Spark x Hadoop

**E se o volume de dados estiver em streaming,  
ou seja, fluxo contínuo de dados?**



Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Spark x Hadoop

O Apache Spark foi a primeira plataforma de Big Data a integrar processamento de dados em batch, streaming e computação distribuída em um único framework.



# Spark x Hadoop

Hadoop	Spark
Armazenamento distribuído + Computação distribuída	Somente computação distribuída
Framework MapReduce	Computação genérica
Normalmente processa dados em disco (HDFS)	Em disco / Em memória
Não é ideal para trabalho iterativo	Excelente para trabalhos iterativos (Machine Learning)
Processo batch	Até 10x mais rápido para dados em disco Até 100x mais rápido para dados em memória
Basicamente Java	Suporta Java, Python, Scala e R
Não possui um shell unificado	Shell para exploração ad-hoc



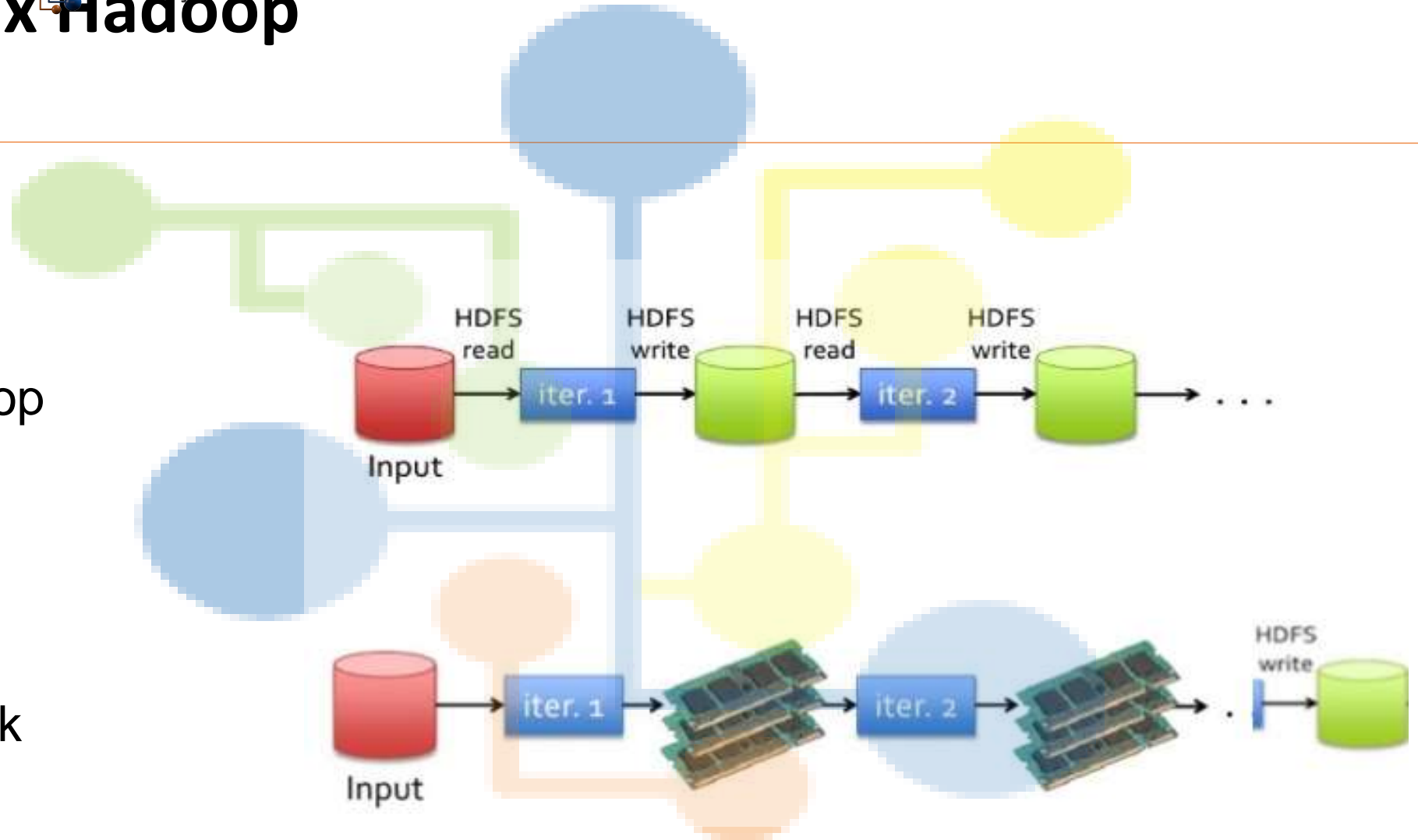
Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Spark x Hadoop

Hadoop

Spark





Data Science  
Academy

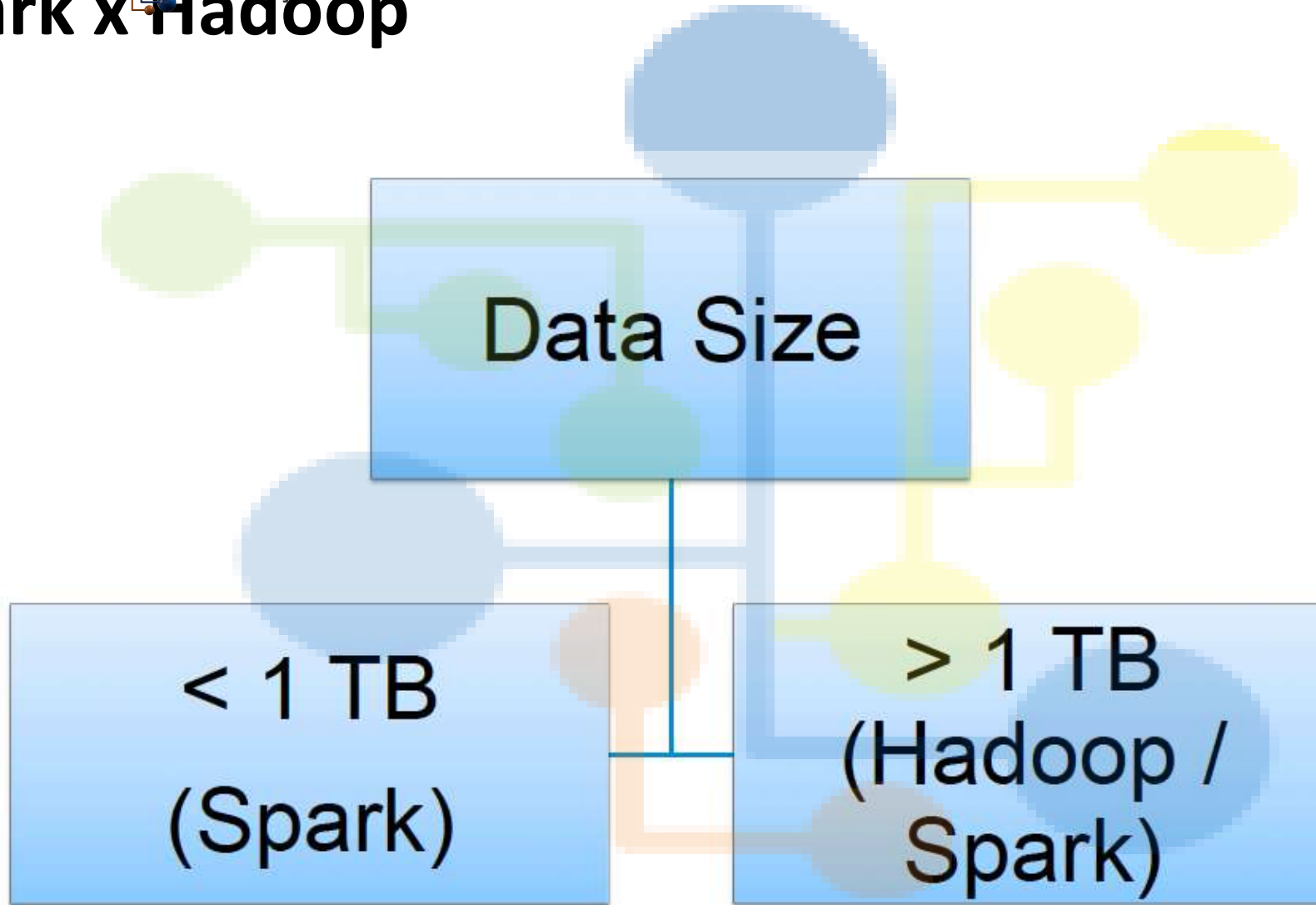
Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e

# Spark x Hadoop

**O Spark é muito bom quando os dados podem ser processados em memória.  
Mas e quando não podem?**



# Spark x Hadoop





# Spark x Hadoop

	Hadoop	Spark
Processamento batch	Hadoop MapReduce (Java, Pig, Hive)	Spark RDD (Java, Python, Scala, R)
Query SQL	Hadoop: Hive	Spark SQL
Processamento Stream / Processamento em Tempo Real	Storm, Kafka	Spark Streaming
Machine Learning	Mahout	Spark ML Lib
Algoritmos iterativos	Lento	Muito rápido (em memória)
Workflow ETL	Pig, Flume	Pig com Spark ou Mix de Spark SQL e programação RDD
Volume de Dados	Volume gigante (Petabytes)	Volume médio (Gigabytes / Terabytes)





Data Science  
Academy

Data Science Academy ericgpt@gmail.com 5b1700f85e4cde813d8b459e



Tenha uma Excelente Jornada de Aprendizagem.

Muito Obrigado por Participar!

Equipe Data Science Academy