

# Engenharia de Dados Big Data

## Aula 1 - Introdução ao Big Data

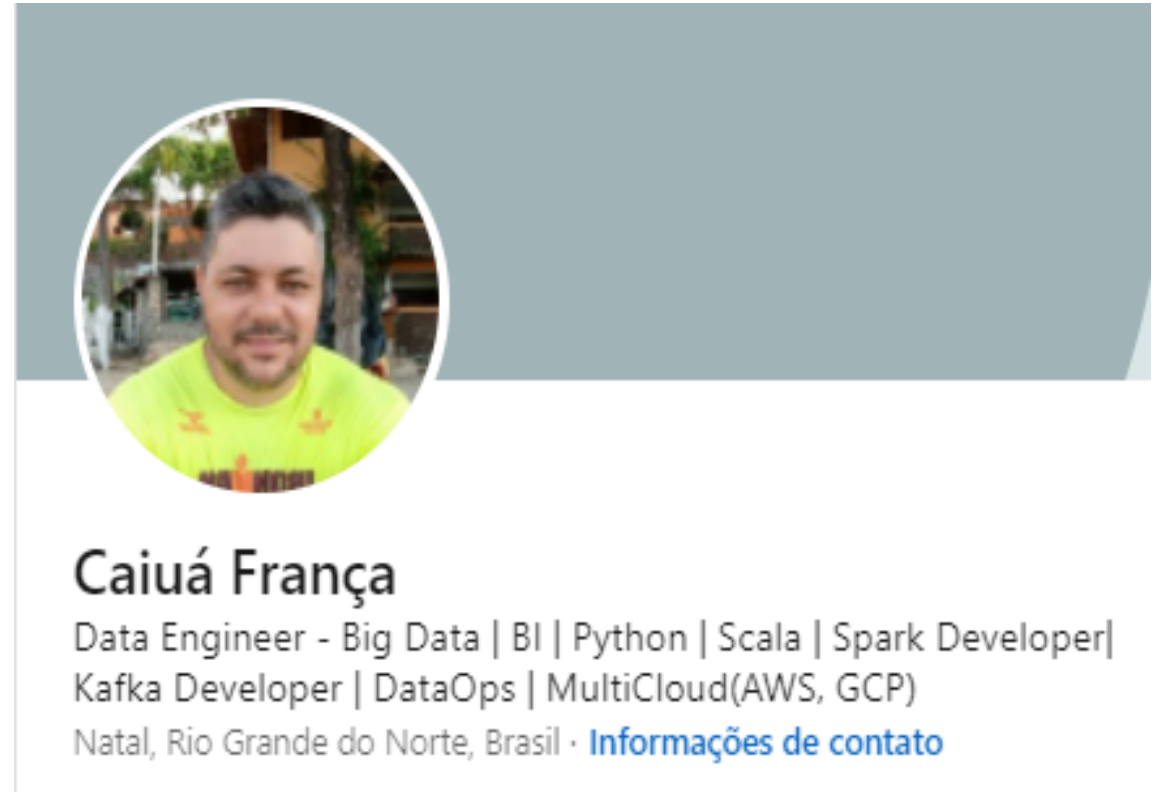


# Sobre mim:

Eu sou Caiuá França.

- Eng. de Dados Big Data Sr.
- 1 Ano de Indra (Minsait)
- Bacharel em Sistemas de Informação
- Pos Graduação em BI e Big Data

E-mail: [cgomesf@minsait.com](mailto:cgomesf@minsait.com)



# Índice

|                    |    |
|--------------------|----|
| O que é Big Data   | 4  |
| Mercado Atual      | 7  |
| Arquitetura Hadoop | 11 |
| HDFS               | 16 |
| Map Reduce         | 18 |
| Yarn               | 20 |
| Hadoop Common      | 20 |

# O que é Big Data?

São dados com maior volume, velocidade e variedade, que softwares tradicionais de processamento não conseguem gerenciar. (oracle.com)

São informações de alto volume, velocidade e variedade que exigem formas inovadoras e econômicas de processamento e permitem uma visão aprimorada, para tomada de decisões e automação de processos. (gartner.com)

# 1

# 5V's do Big Data



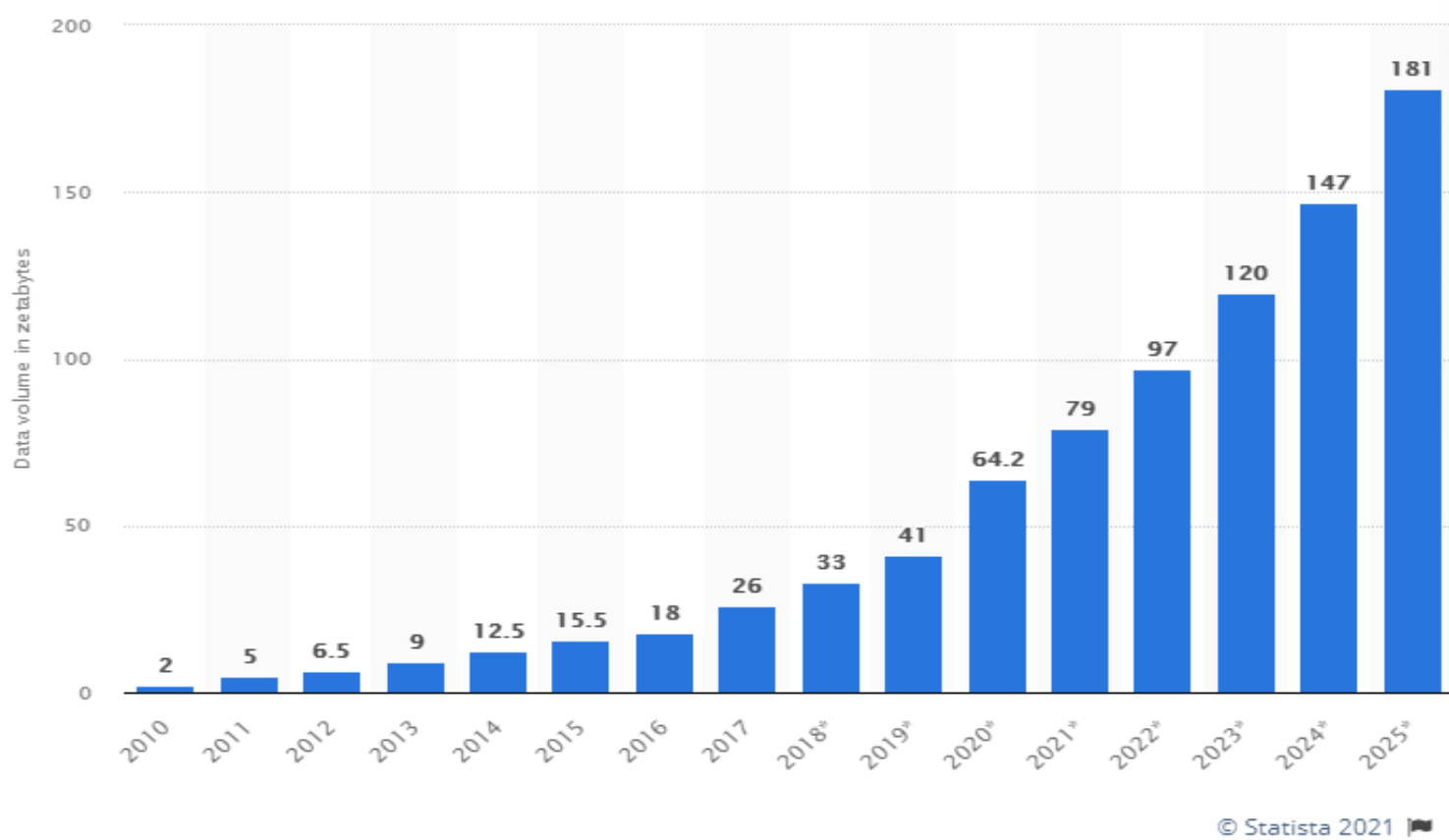
# Origem do Hadoop



# Mercado Atual

indra

Previsão de crescimento dos dados no planeta até 2025



2

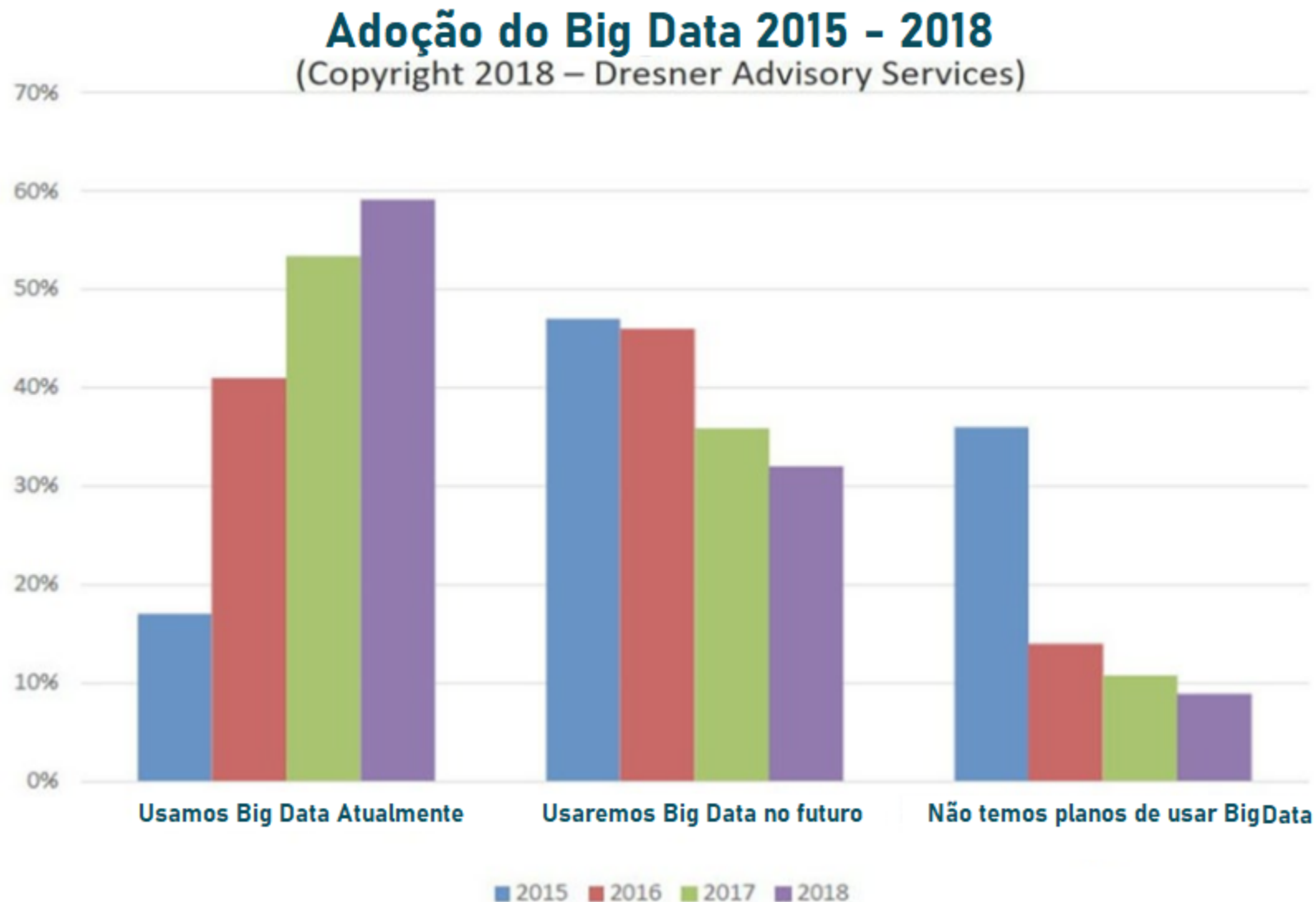
# Mercado Atual

- Google gera 100 PetaBytes de dados por dia.
- Facebook gera 30 ou mais PetaBytes de dados por dia.
- Twitter gera 100 TeraBytes por dia.
- Spotify gera 64 TeraBytes por dia
- eBay gera 100 PetaBytes por dia.
- Até 2007 tínhamos gerado 300 EB de dados, hoje já excedemos a casa dos 4.000 EB,
- Em 2020 geramos mais 50.000 EB ou 50 ZetaBytes.
- O Google é a maior empresa de Big Data do mundo, processando 3,5 bilhões de solicitações diárias,
- gerando e armazenando 10 ExaByte de dados.
- Noventa (90%) de todos os dados do mundo foram produzidos nos últimos 2 anos.

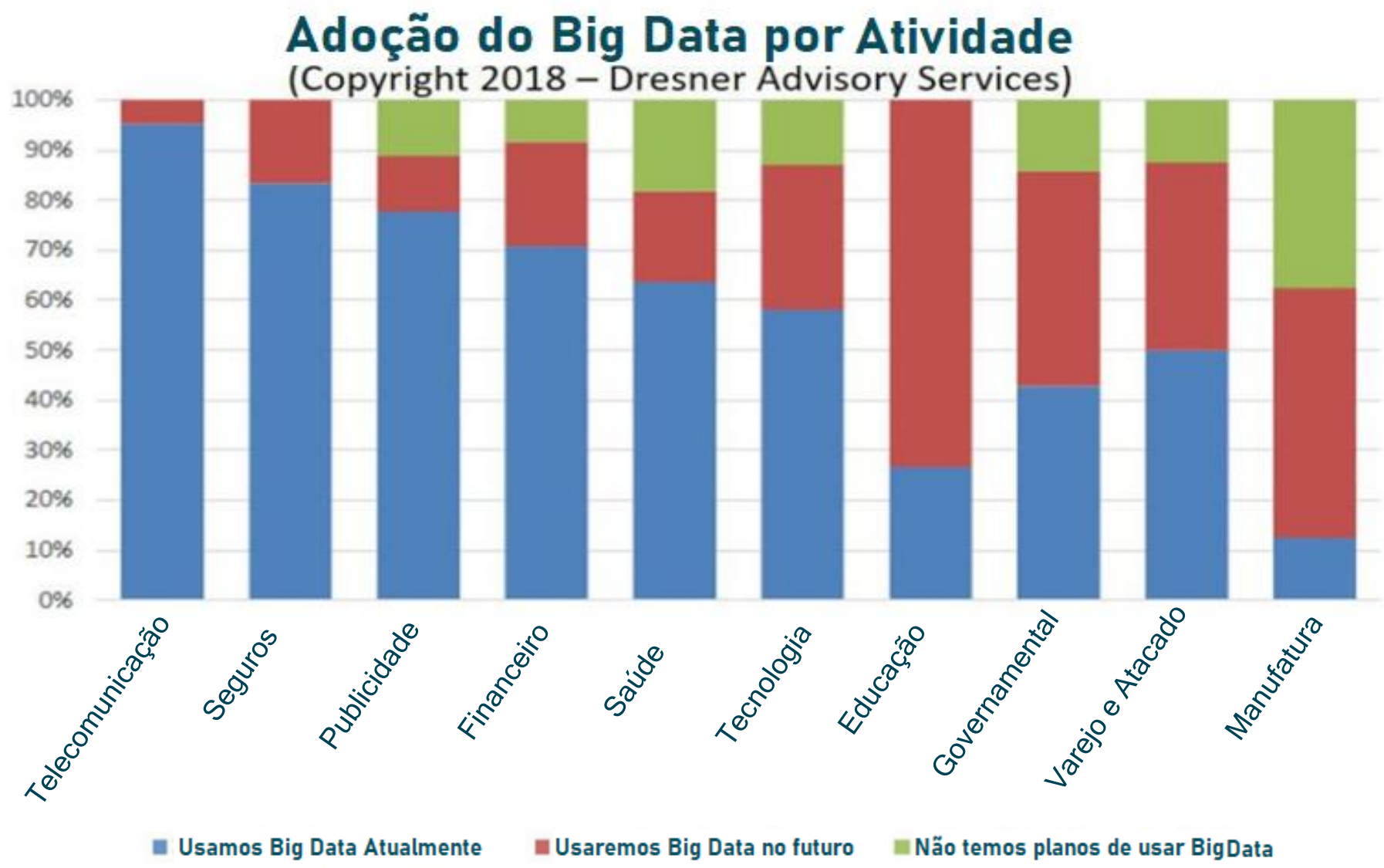
Big Data para executivos de Mercado, Thomas H. Davenport.



# Mercado Atual



# Mercado Atual



# Arquitetura Hadoop

- Criadores:



Doug Cutting



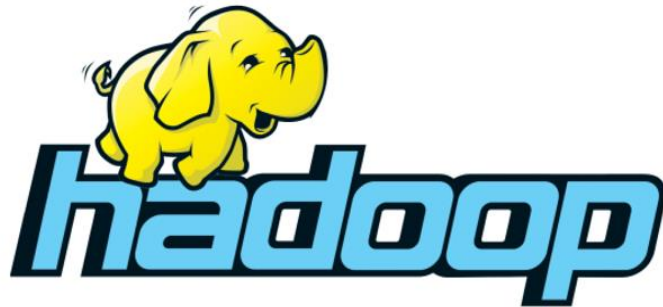
Mike Cafarella

3

# Arquitetura Hadoop

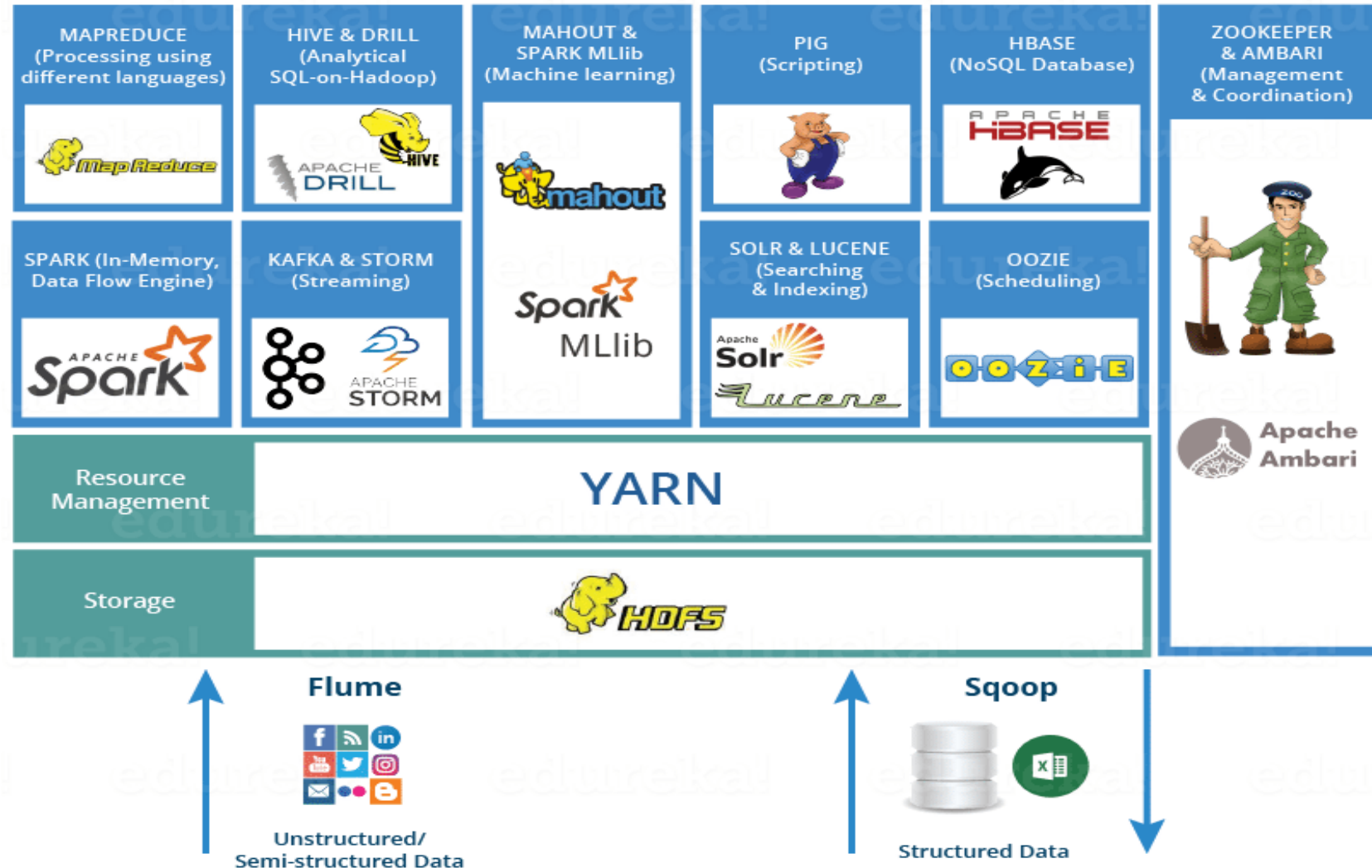
## O que é Hadoop?

- Plataforma que fornece infraestrutura econômica e escalável
- Processamento em lote para grandes quantidades de dados
- Armazenamento e Processamento distribuído
- Deu origem ao ecossistema Big Data
- 4 Módulos: HDFS, Mapreduce, Hadoop Common e Yarn



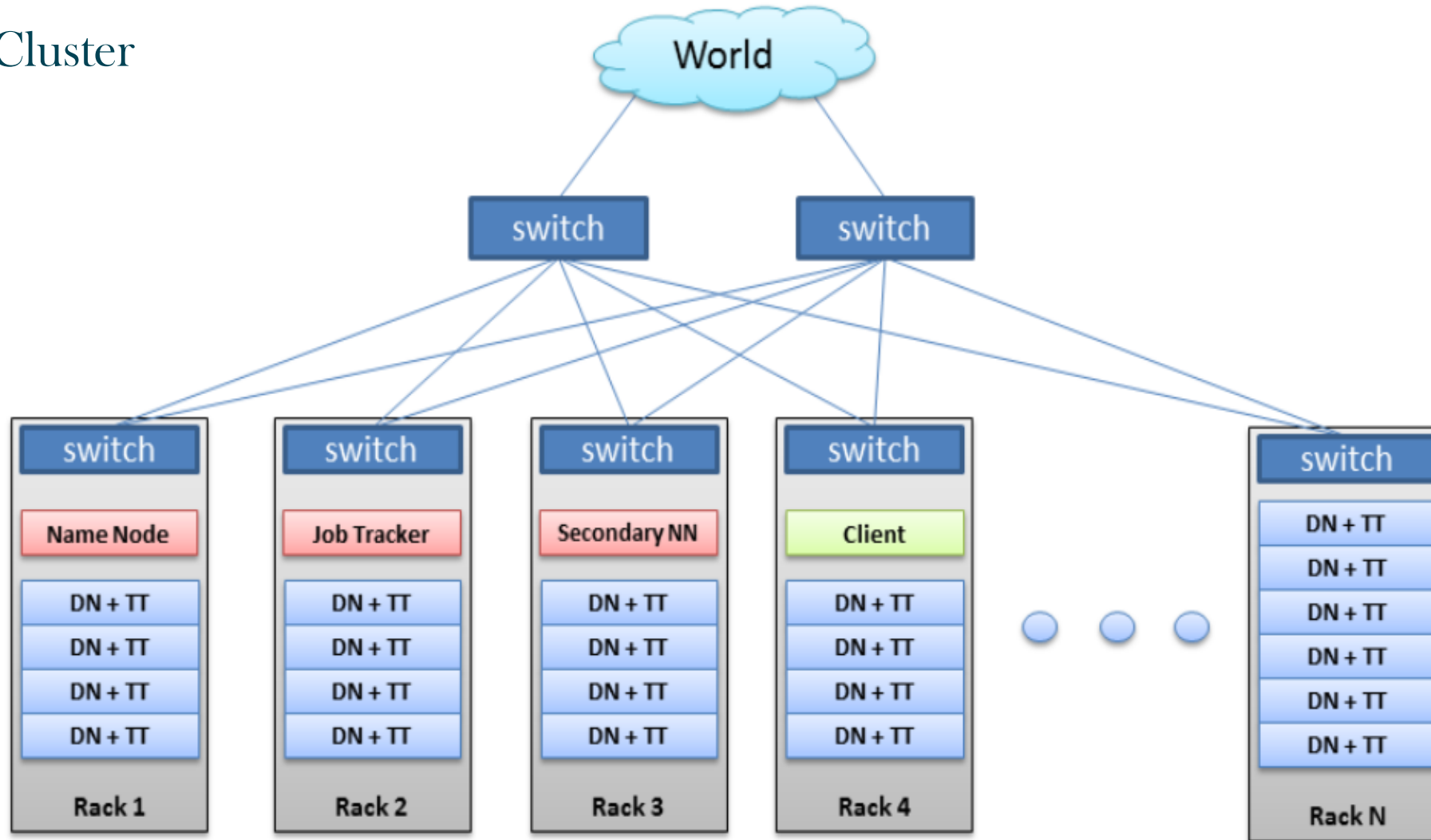
# Arquitetura Hadoop

## Ecossistema Hadoop



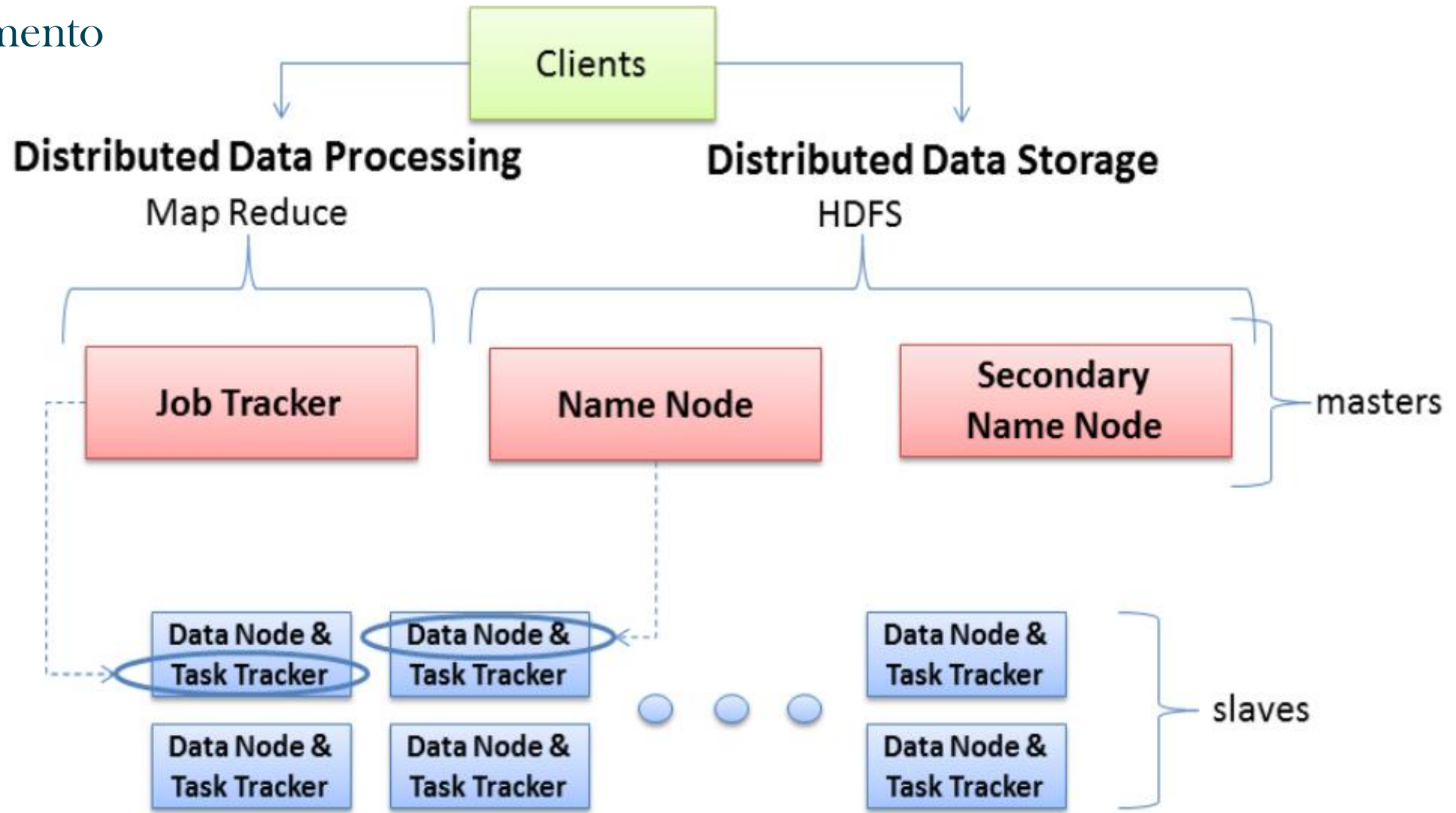
# Arquitetura Hadoop

## Hadoop Cluster



# Arquitetura Hadoop

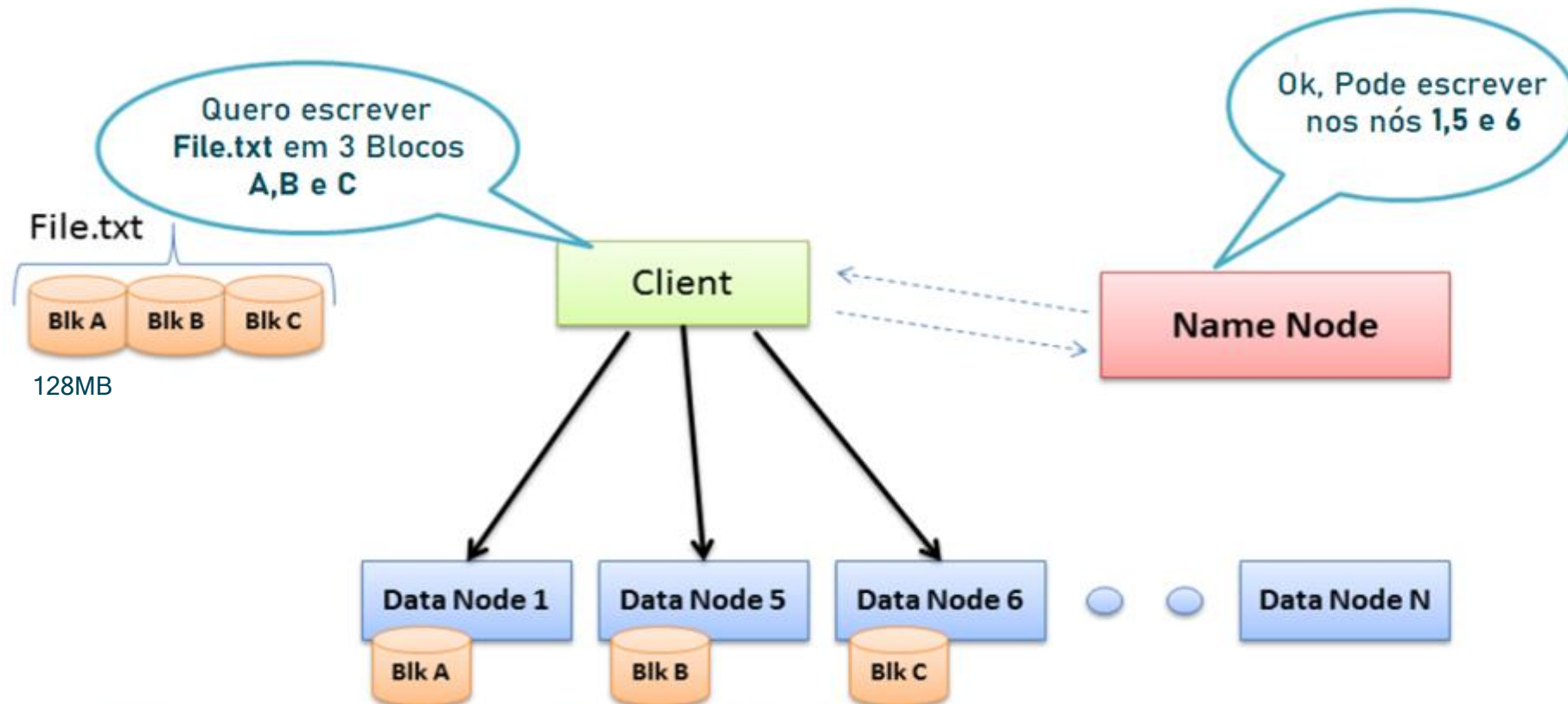
## Funcionamento





# Arquitetura Hadoop

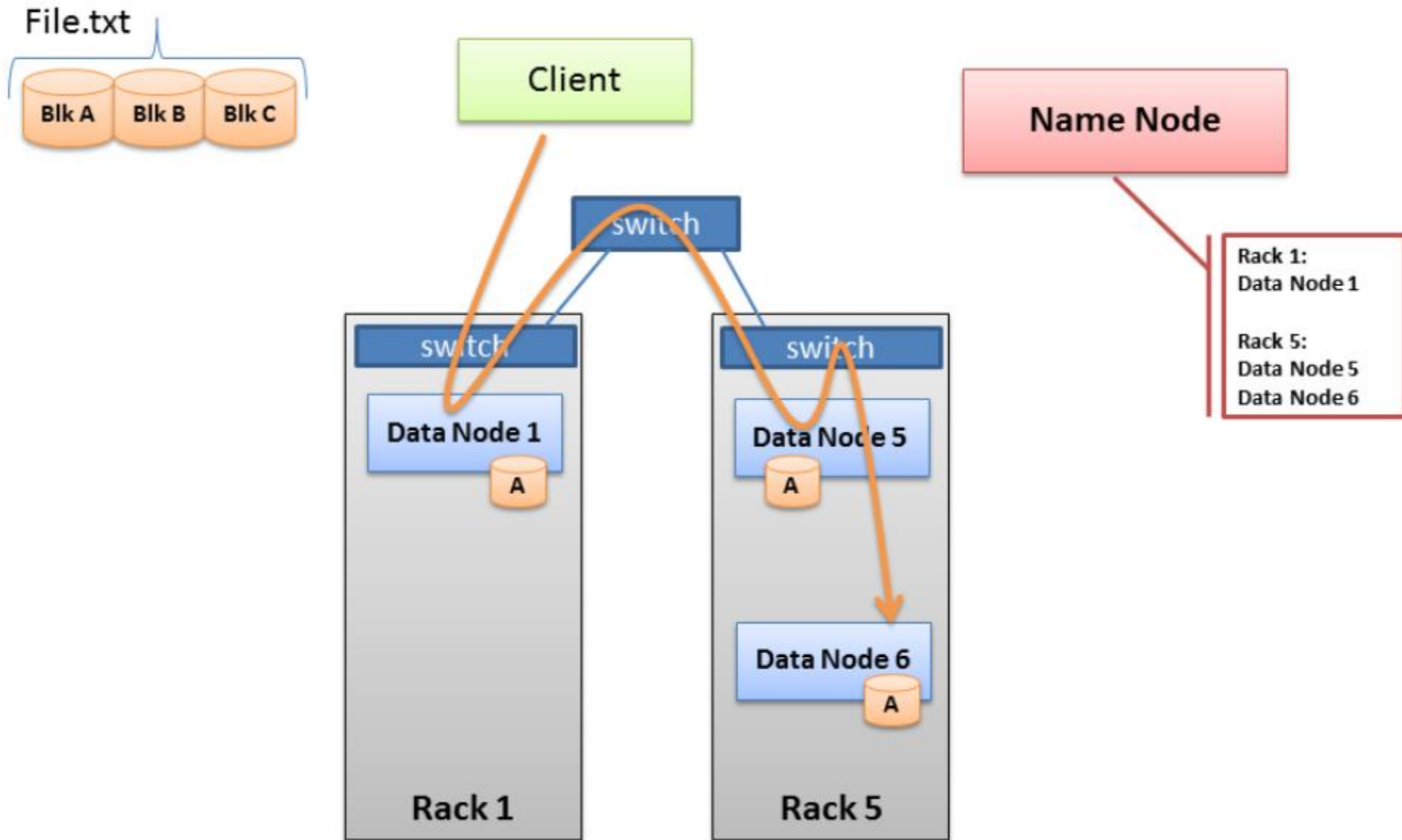
## HDFS



- Cliente consulta Name Node
- Cliente escreve no Data Node
- Data Node replica o bloco
- Ciclo se repete para o próximo bloco

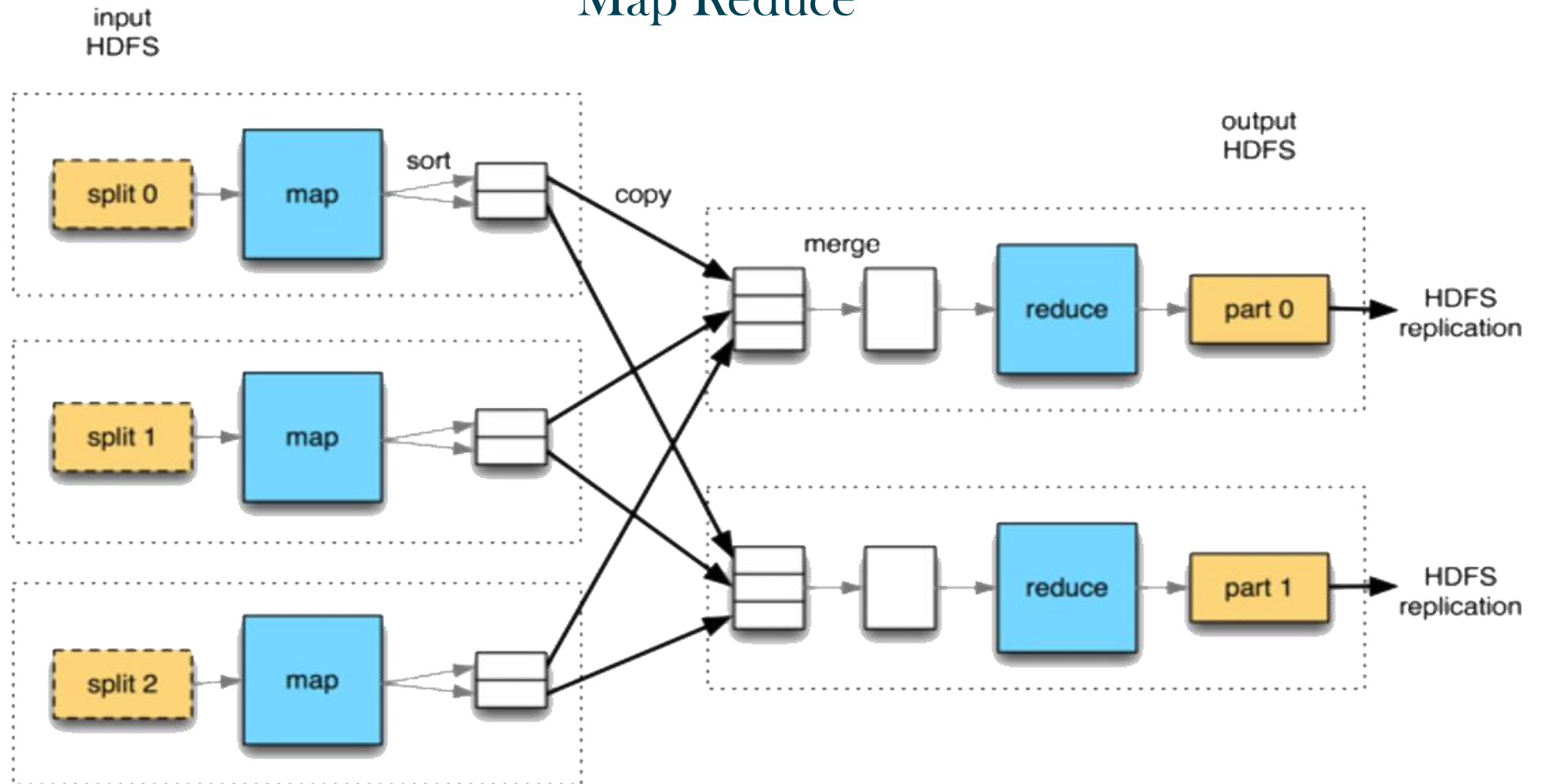


# Arquitetura Hadoop



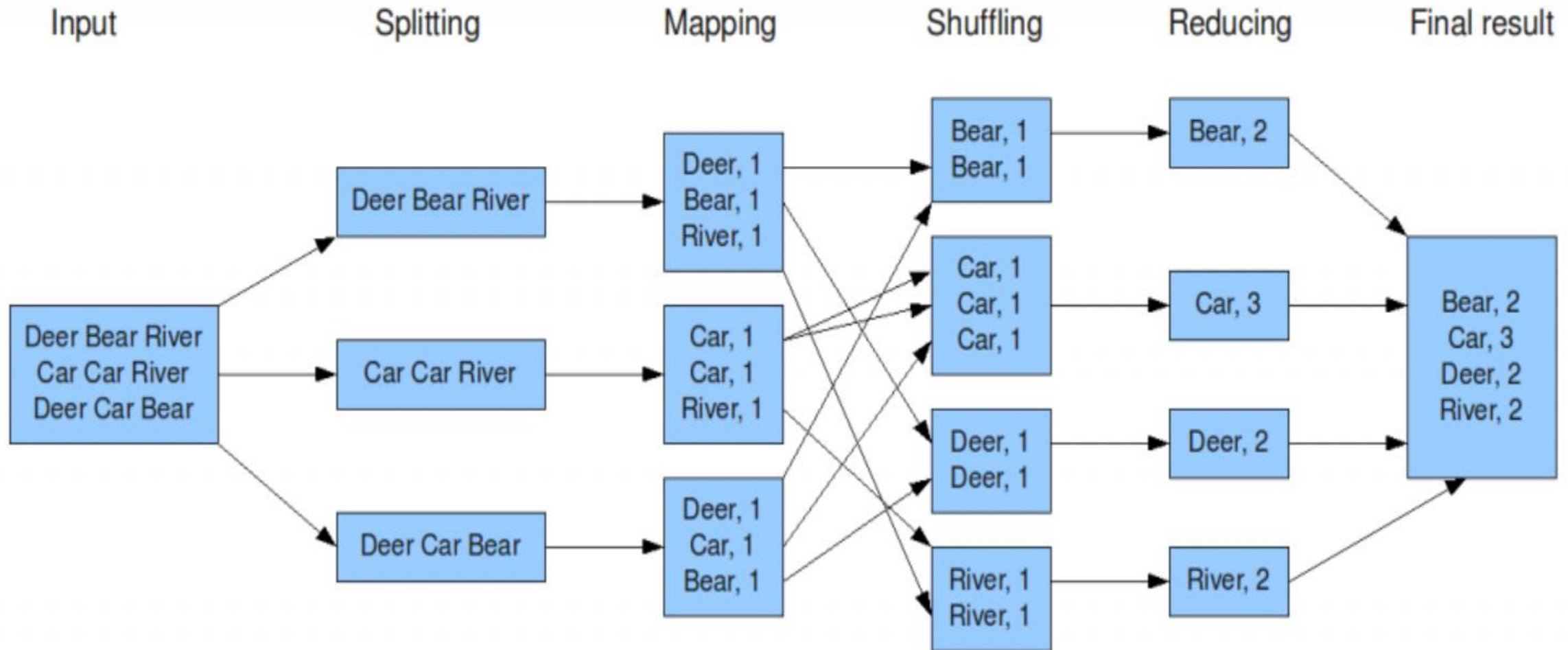
# Arquitetura Hadoop

## Map Reduce



# Arquitetura Hadoop

## Map Reduce



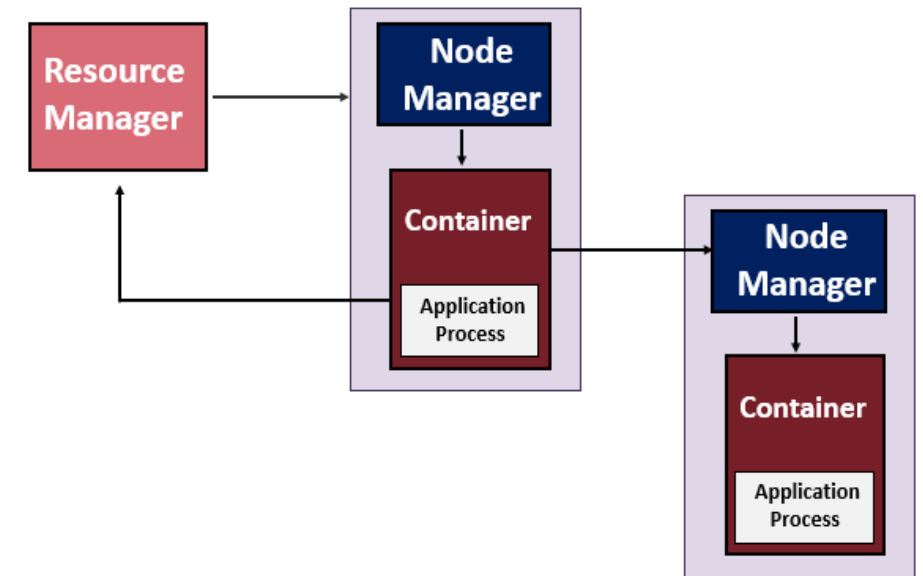
# Arquitetura Hadoop

## Yarn

Considere o YARN como o cérebro do ecossistema Hadoop. Executa todas as atividades de processamento, alocando recursos e agendando tarefas.

**ResourceManager e NodeManager.**

- **ResourceManager** é um nó principal ele recebe as solicitações de processamento e, em seguida, passa as partes das solicitações para os **NodeManagers** correspondentes, onde o processamento real ocorre.
- **NodeManagers** são instalados em cada **DataNode**. É responsável pela execução da tarefa em cada **DataNode** único.



## Hadoop Common

Contém as bibliotecas Java e arquivos comuns que são necessários para todos os módulos Hadoop.



# Migração HDP => CDP

## MIGRAÇÃO CLUSTER HADOOP - HDP 2.6.4 para CDP 7.1.6

### O QUE É?

Devido ao “End Of Life” e término de suporte ao produto HDP 2.6.4 (Hortonworks Data Platform), o cluster Hadoop está em migração para nova infraestrutura com o produto **CDP 7.1.6 (Cloudera Data Platform)**.



### QUAIS SÃO OS PRINCIPAIS BENEFÍCIOS?



Entre os principais benefícios estão a atualização tecnológica do ambiente com a versão atualizada do produto, suporte renovado pelo fornecedor e atualização de hardware dos servidores que compõem o ambiente .

**+450**

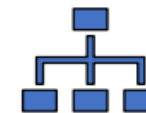


Servidores compõem o  
Hadoop

**25<sub>Pb</sub>**



Capacidade do  
Datalake



Maior cluster da  
América Latina

**indra**  
At the core