

# COMPUTAÇÃO EM NUVEM E PROCESSAMENTO MASSIVO DE DADOS

---

Conceitos, tecnologias e aplicações

Jaqueline Joice Brito

Slides em colaboração com Lucas de Carvalho Scabora

# Sumário

- **Computação em Nuvem**
  - Definição
  - Modelos de Serviços
  - Prós e Contras
  - Aplicações
- **Processamento Massivo de Dados**
  - Hadoop e Spark
  - Exemplo: processamento de junção estrela

# O que é Computação em Nuvem?

- Computação em nuvem é um modelo que possibilita acesso a um conjunto de **recursos computacionais compartilhados e interligados** via rede



# O que é Computação em Nuvem?

A nuvem é uma **metáfora** para a Internet ou infraestrutura de comunicação entre os componentes arquiteturais

Abstração que oculta a complexidade de infraestrutura



# O que é Computação em Nuvem?

Em resumo, é uma plataforma que provê serviços sob-demanda, que está **sempre disponível**, em **qualquer lugar e a qualquer hora**



# O que é Computação em Nuvem?

## Definição segundo NIST

“Computação em nuvem é um modelo para permitir **acesso ubíquo, conveniente e sob demanda** via rede a um **agrupamento compartilhado e configurável de recursos computacionais** (por exemplo, redes, servidores, equipamentos de armazenamento, aplicações e serviços), **que pode ser rapidamente fornecido e liberado** com esforços mínimos de gerenciamento ou interação com o provedor de serviços.”

# O que é Computação em Nuvem?

## Principais Características

- Serviço sob demanda
  - Alocação dinâmica de recursos
- Acesso via rede
- Compartilhamento de recursos
- Elasticidade
  - Sensação de capacidade infinita de recursos
- Serviço mensurável
  - Modelo pay-as-you-go

# O que é Computação em Nuvem?

	Tecnologia	Modelo de Negócio
Mainframe	Computação Centralizada	Alto custo de hardware e software
Cliente/Servidor	Computação Distribuída	Licença para SO e aplicativos
Computação em Nuvem	Grandes data centers	Custo proporcional ao uso



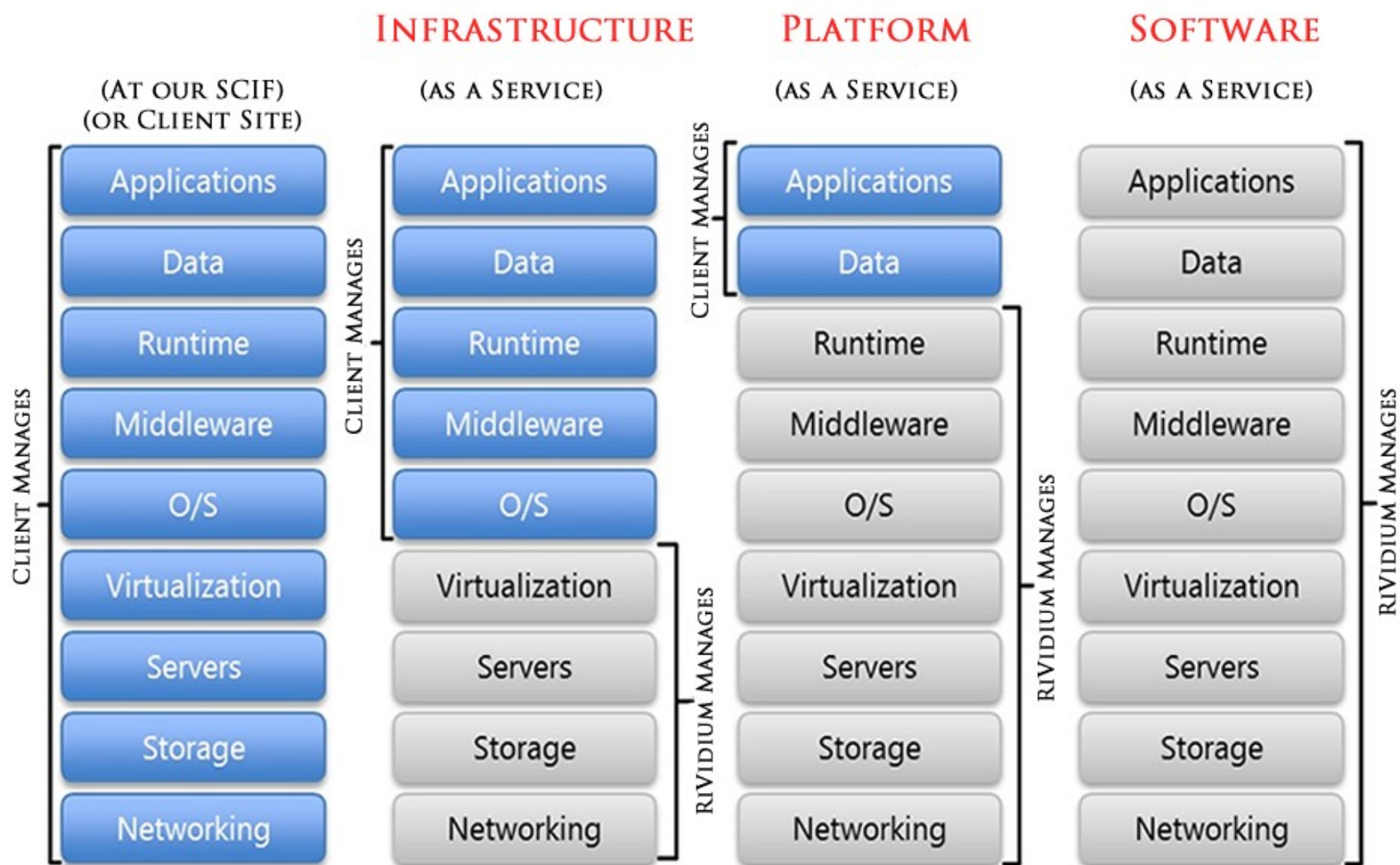
# O que é Computação em Nuvem?

## Modelos de implantação

- Privada
- Comunitária
- Pública
- Híbrida



# Modelos de Serviços



# Modelos de Serviços

## Termos Emergentes

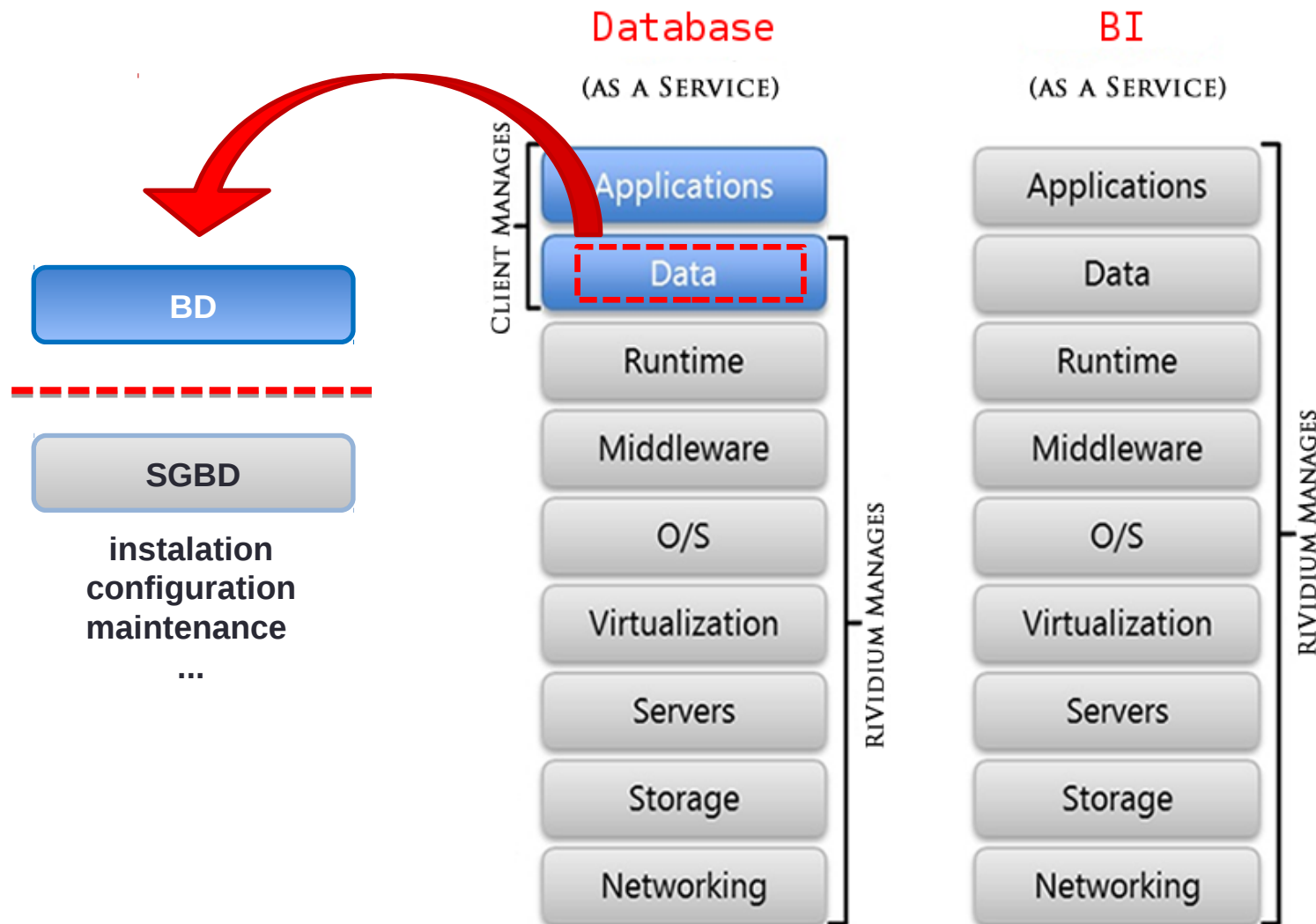
### Database as a Service (DBaaS)

O provedor de serviço tem a responsabilidade de instalar e dar manutenção ao banco de dados. Assim, o usuário, que contratou o serviço, apenas paga um valor proporcional ao uso.

### Business Intelligence as a Service (BlaaS)







Também chamado de Software como um serviço de business intelligence (SaaS BI), consiste em instalar e gerenciar aplicações de BI na nuvem.

# Modelos de Serviços



# Modelos de Serviços

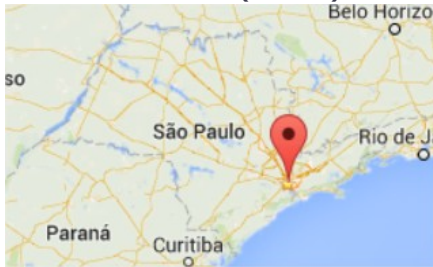
## Exemplos

Software as a Service (SaaS)	Platform as a Service (PaaS)	Infrastructure as a Service (IaaS)
		
		

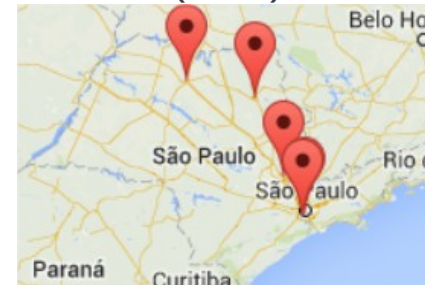
# Tipos de Armazenamento

## Exemplo da Microsoft Azure

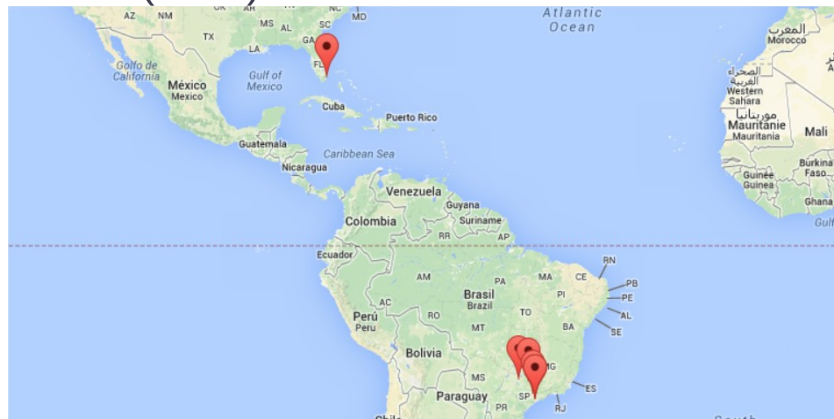
ARMAZENAMENTO  
COM REDUNDÂNCIA  
LOCAL (LRS)



ARMAZENAMENTO COM  
REDUNDÂNCIA DE ZONA  
(ZRS)



ARMAZENAMENTO  
COM REDUNDÂNCIA  
GEOGRÁFICA (GRS)



e

ARMAZENAMENTO COM  
REDUNDÂNCIA GEOGRÁFICA NO  
ACESSO DE LEITURA (RA-GRS)

# Prós e Contras

## Principais Vantagens

### Redução de investimento em TI (hardware e software)

- Envolvendo custo de manutenção, de pessoal, de espaço físico e de energia

### Confiabilidade

- Replicação dos dados
- Disponibilidade das aplicações

# Prós e Contras

## Principais Vantagens

- Não há contratos complexos e duradouros de prestação de serviço
  - Cobra-se apenas pelos recursos multiplicados pelo tempo de uso
- Escala por demanda
  - Capacidade virtualmente infinitos
- Abstração da tecnologia sendo utilizada



# Prós e Contras

## Principais Desafios

Segurança e confidencialidade dos dados

Gerenciamento dos dados

Disponibilidade

Integração de serviços

Necessidade de uma banda maior de internet

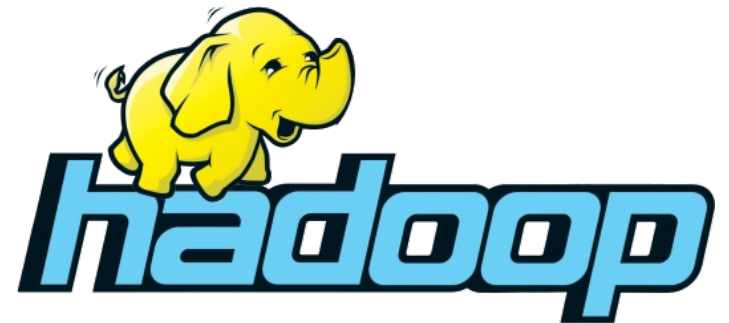
# Aplicações

## Usos mais comuns da computação em nuvem

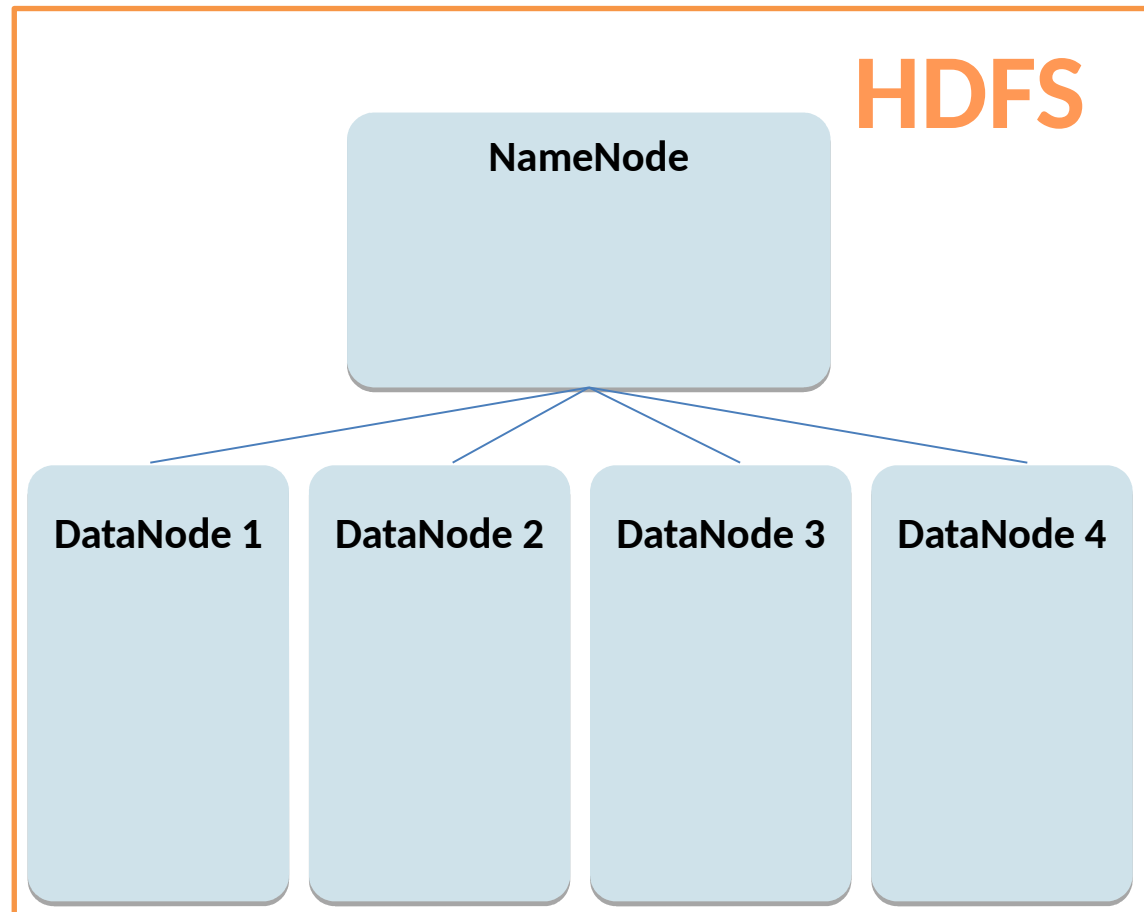
- › IaaS e PaaS
- › Armazenamento de dados
- › Ambientes de testes e desenvolvimento
- › **Processamento e análise de Big Data**
- › Backup

# Hadoop

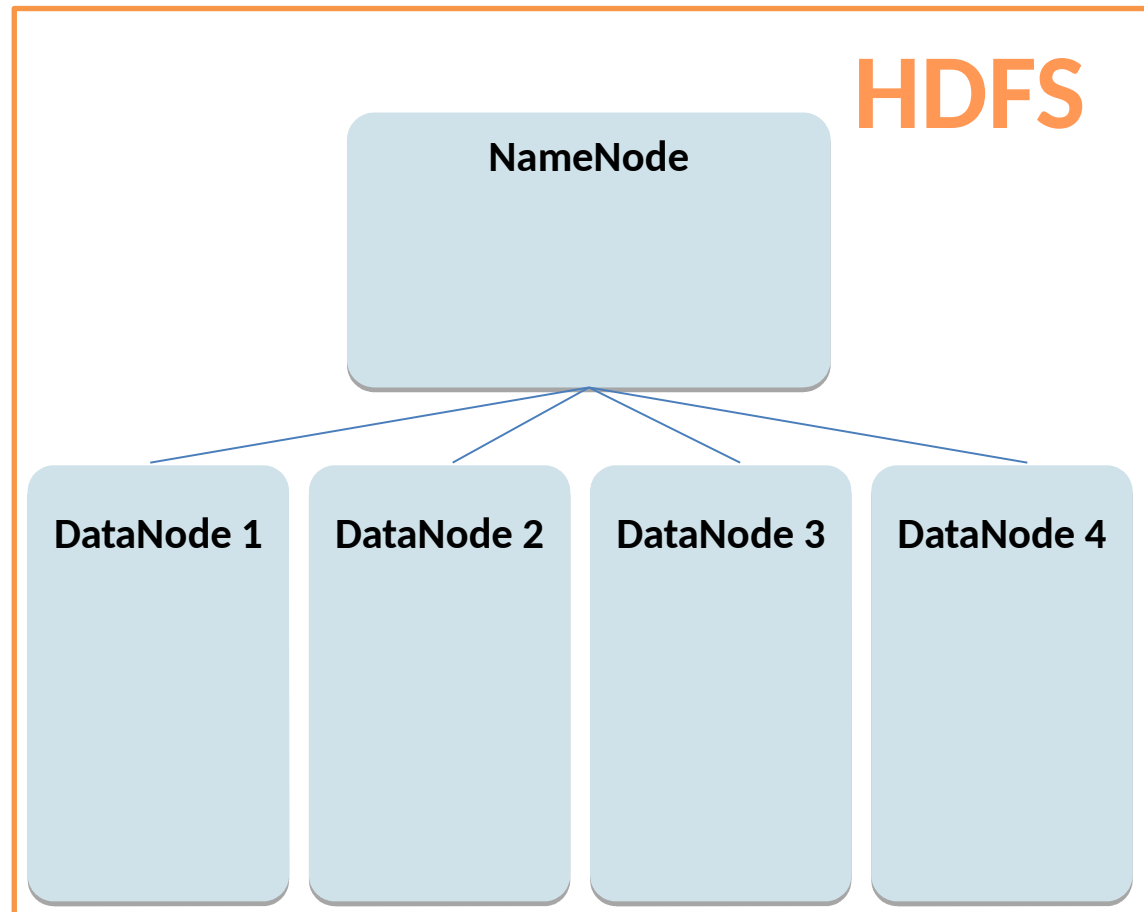
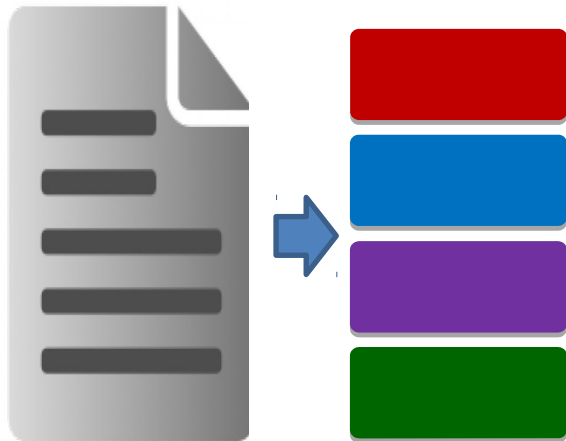
- Framework de processamento paralelo de dados em larga escala
- Altamente escalável
- Tolerante a falhas
- Disponível
- Principais componentes
  - Armazenamento
    - HDFS – Hadoop Distributed File System
  - Processamento
    - MapReduce



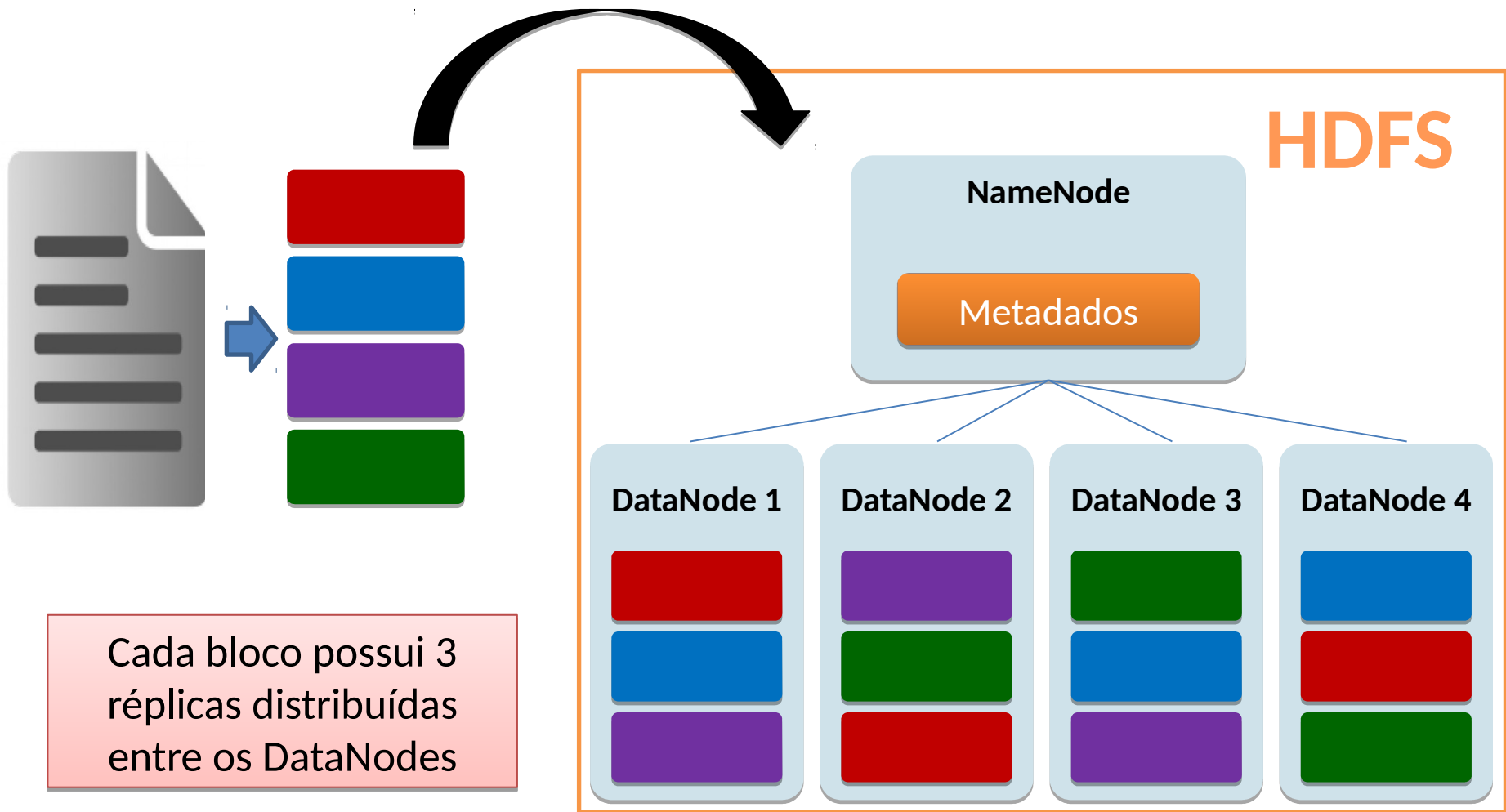
# HDFS – Hadoop Distributed File System



# HDFS – Hadoop Distributed File System



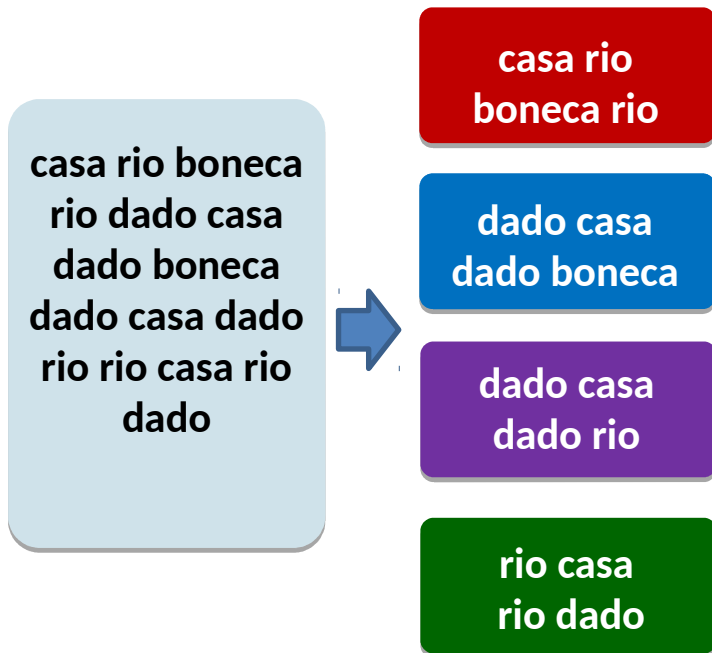
# HDFS – Hadoop Distributed File System



# MapReduce

**casa rio boneca  
rio dado casa  
dado boneca  
dado casa dado  
rio rio casa rio  
dado**

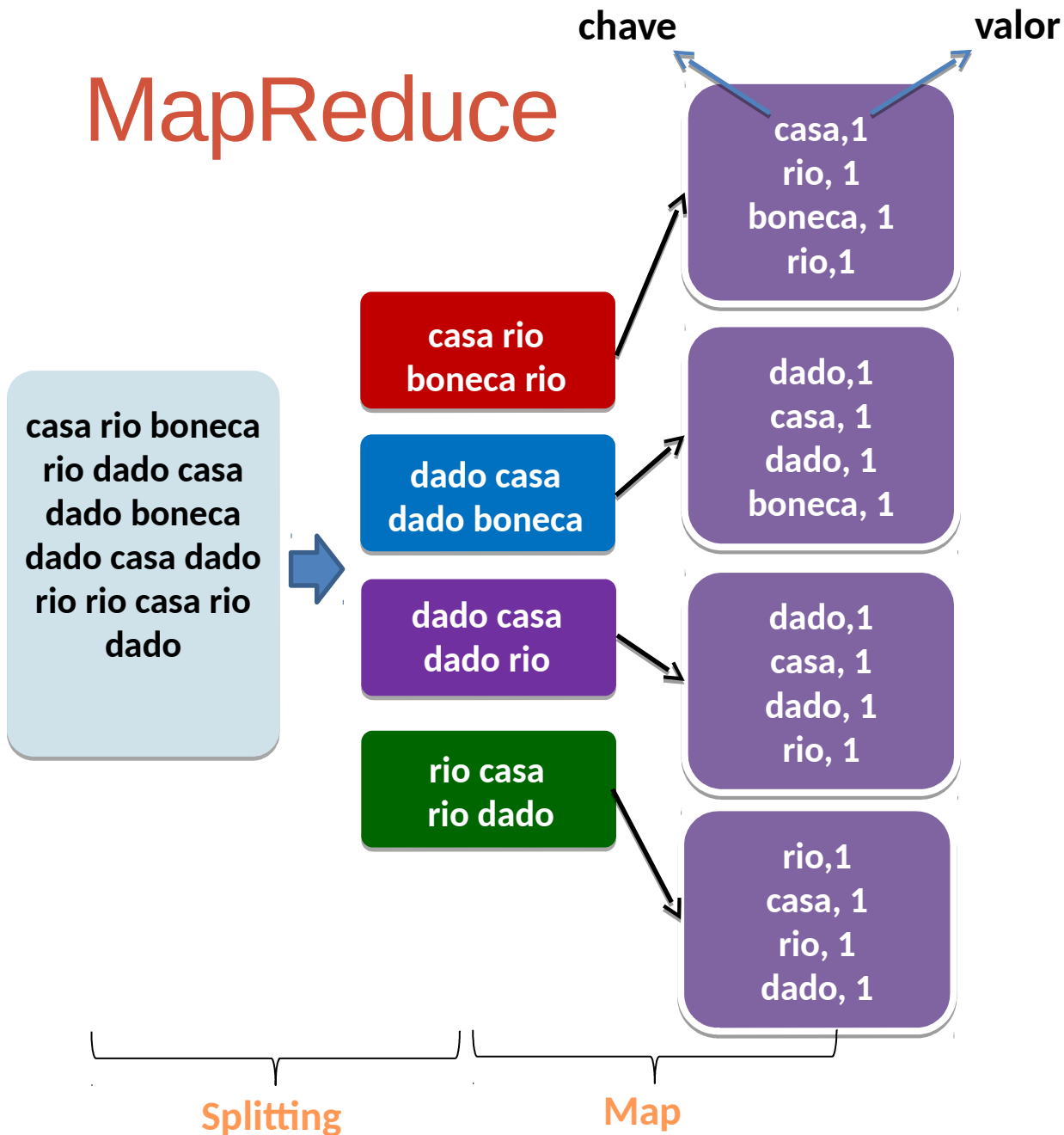
# MapReduce



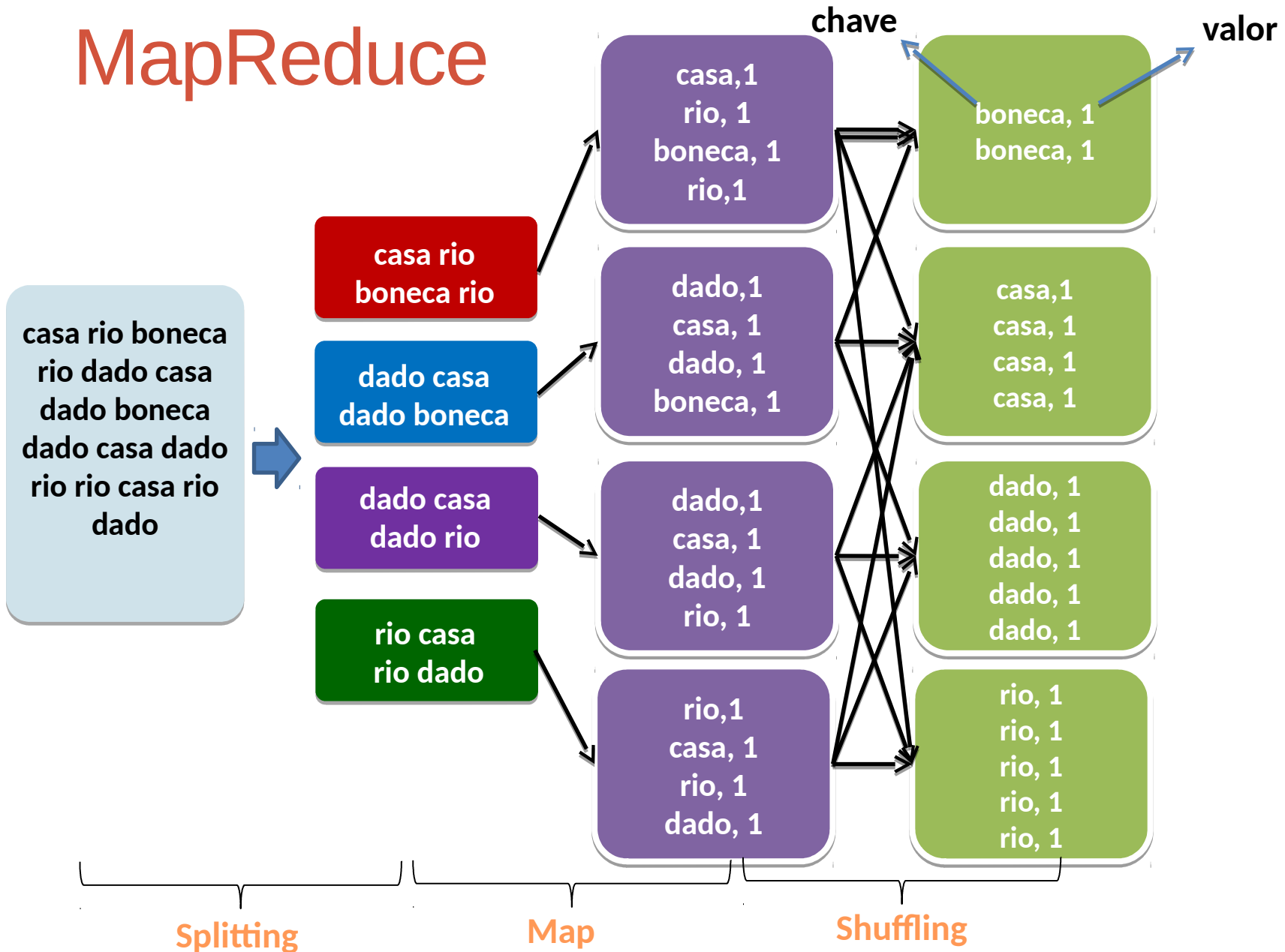
Splitting



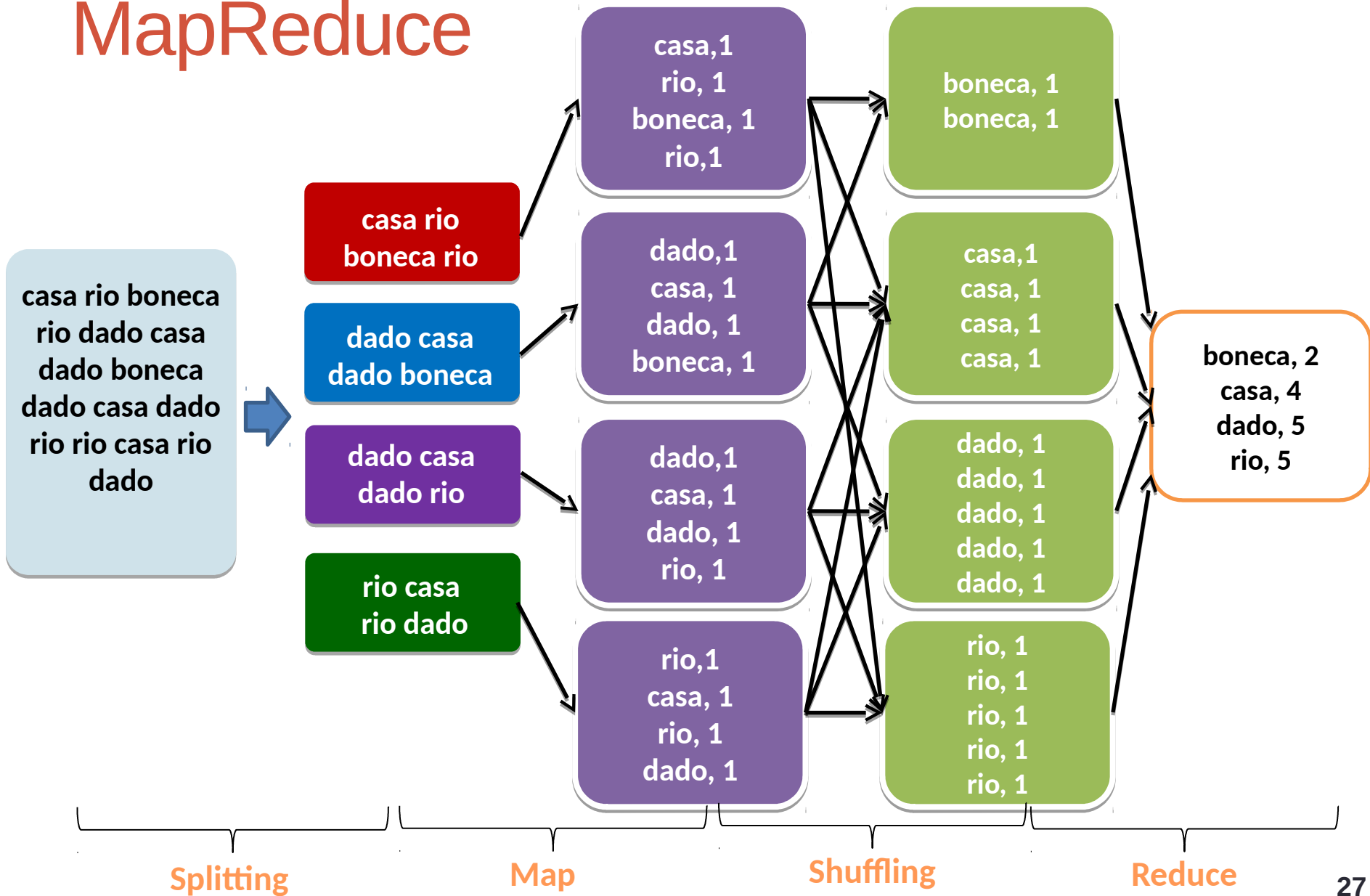
# MapReduce



# MapReduce



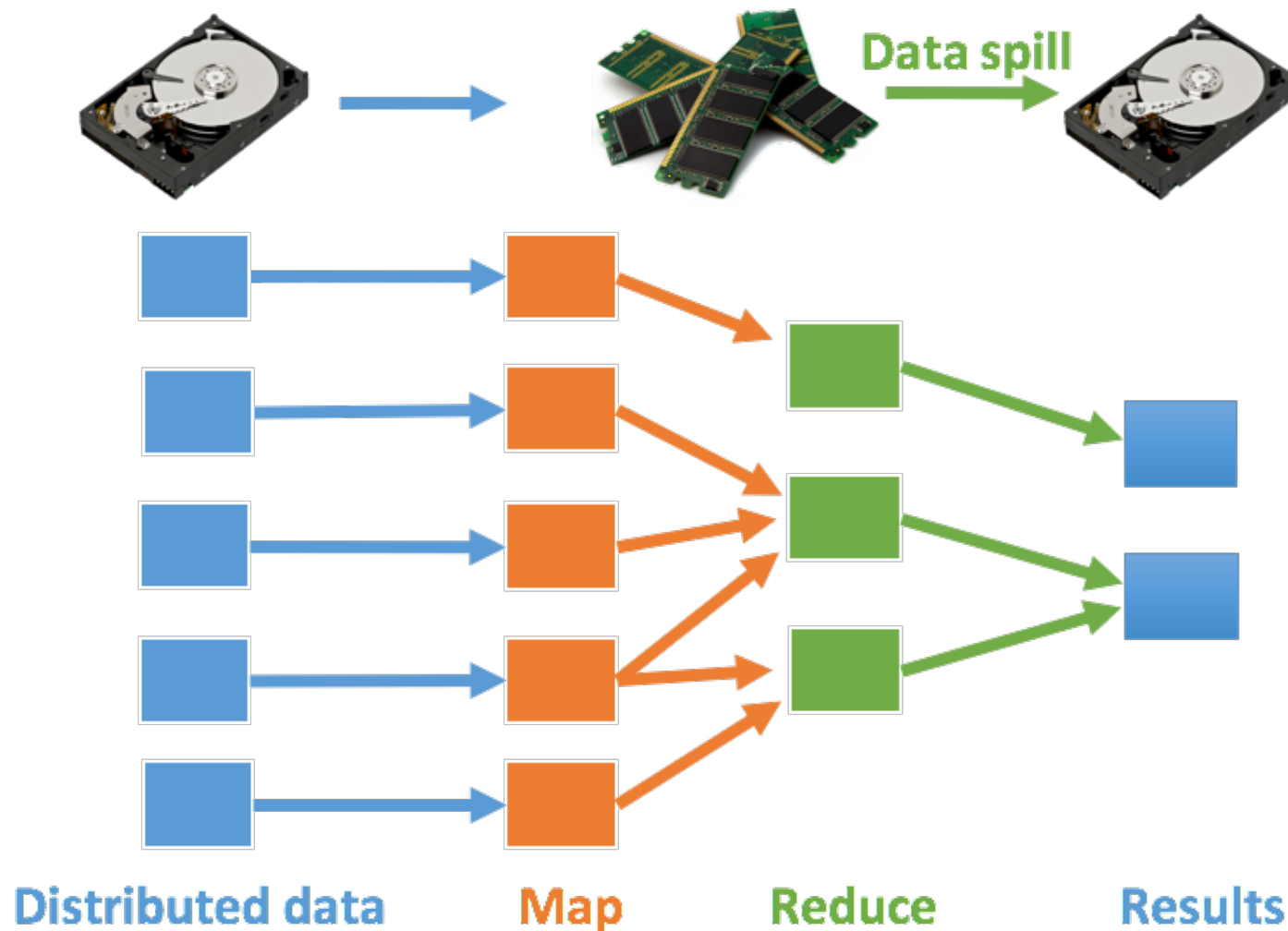
# MapReduce



# Spark

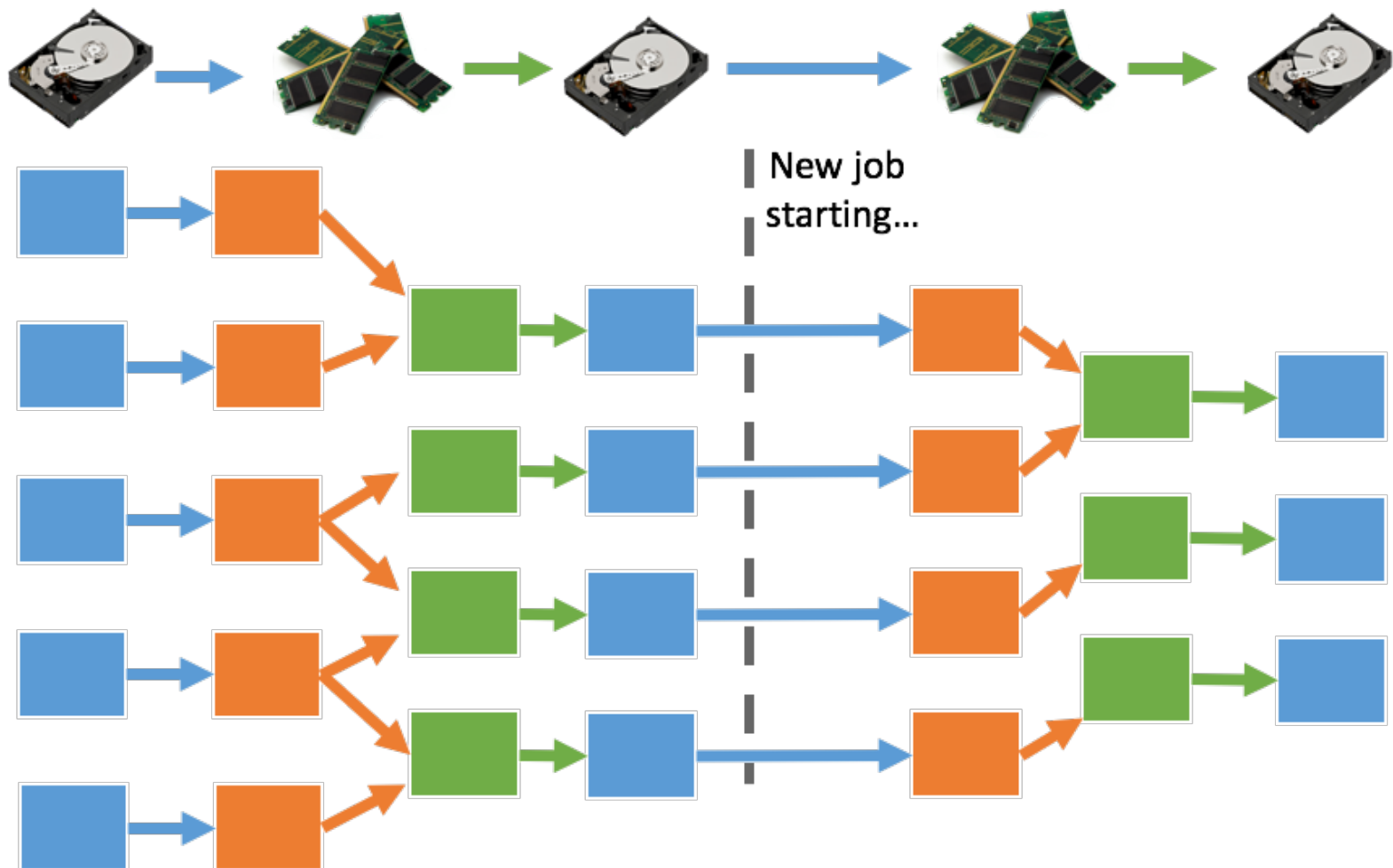
- Framework de processamento distribuído
- Computação em memória
- Baseado em RDDs – Resilient Distributed Datasets
  - Estruturas de dados em paralelo
  - Tolerante a falhas
  - Persistência de resultados intermediários em memória primária
  - Manipulação otimizada por meio de um conjunto operadores

# Fluxo dos Dados no MapReduce



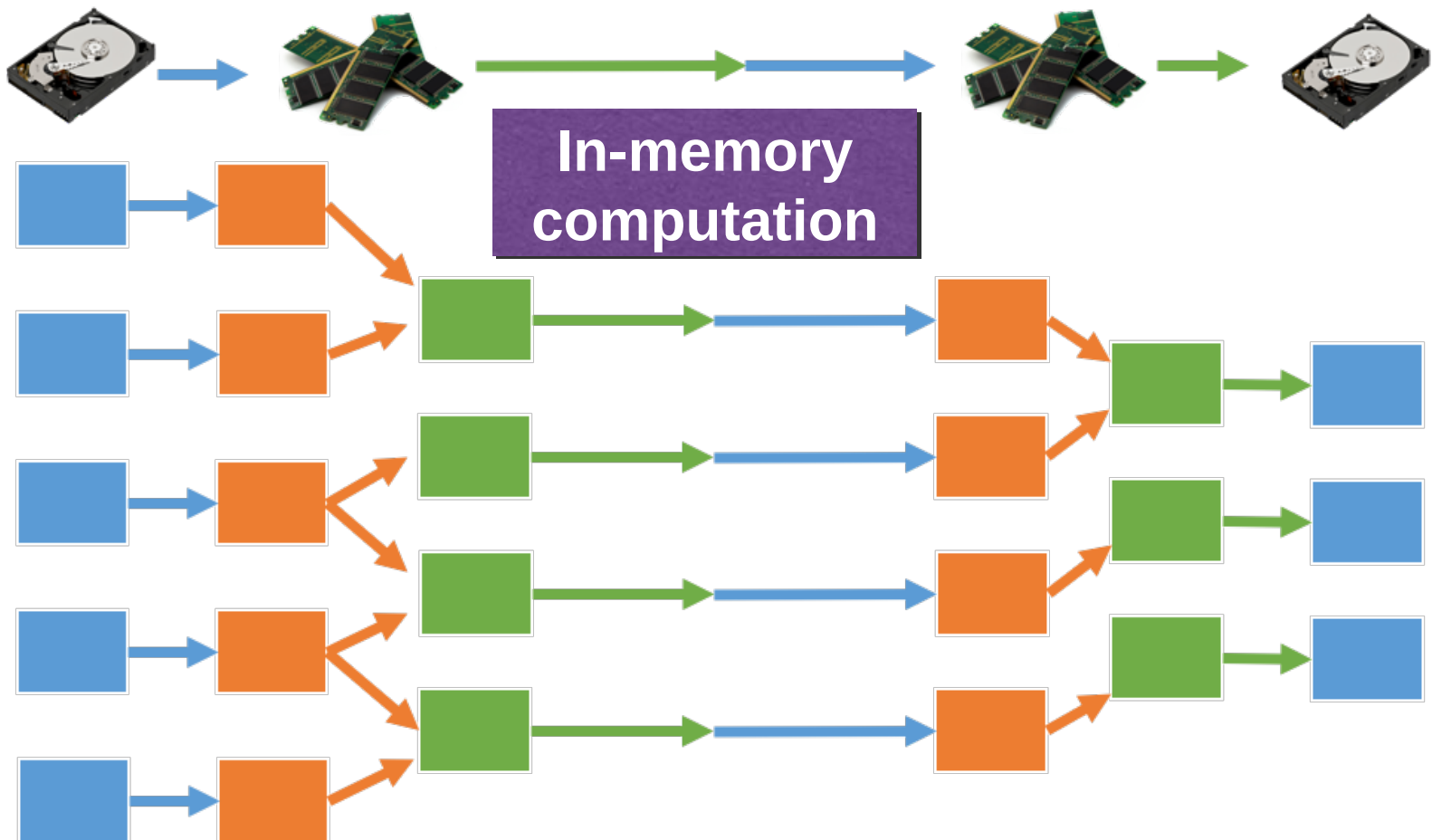
# Fluxo dos Dados no MapReduce

$$F \bowtie D_1 \bowtie D_2 \dots$$



# Fluxo dos Dados no Spark

$$F \bowtie D_1 \bowtie D_2 \dots$$



# Evolução: SQL on Hadoop

Criação de uma camada para prover a utilização de SQL

Motivação

- SQL é amplamente utilizados há muitos anos

Dificuldade em escrever programas em  
MapReduce/Spark

- SQL é traduzido para funções MapReduce e RDDs
  - HiveQL
  - Spark SQL e Data Frames



# Oportunidades

## Data Scientist

- Alguns conhecimentos necessários
  - Saber **programação**
  - Ser capaz de criar **modelos estatísticos**
  - Compreender as **diferentes plataformas de Big Data**
- Usualmente esse profissional é formado em Estatística, Matemática, Física ou Ciências da Computação



# Referências

<http://www.sciencedaily.com/releases/2013/05/130522085217.htm>

[http://www-3.unipv.it/ingegneria/copisteria\\_virtuale/motta/doss/13-PV-DoES-18-BigData-v3.pdf](http://www-3.unipv.it/ingegneria/copisteria_virtuale/motta/doss/13-PV-DoES-18-BigData-v3.pdf)

<http://hadoop.apache.org/>

<http://spark.apache.org/>

<http://azure.microsoft.com/en-us/pricing/details/storage/>

<http://thoughtsoncloud.com/2014/02/top-7-most-common-uses-of-cloud-computing/>

<http://pt.slideshare.net/laodias/os-cinco-vs-do-big-data>

<http://www.microsoft.com/en-us/news/presskits/cloud/docs/the-economics-of-the-cloud.pdf>

<http://www.cloudproviderusa.com/the-difference-between-public-cloud-private-cloud/>

<http://observatorio.inweb.org.br/dengueapp?next=/dengueapp/relatorio>

<https://www.youtube.com/watch?v=jyx8iP5tfCI&feature=youtu.be>