




## Aula 1 - Big Data

---

Lucio Monteiro



Big Data

---

## Big Data

“Data is the new science. Big Data holds the answers.” - Patrick P. Gelsinger, Forbes



# Big Data

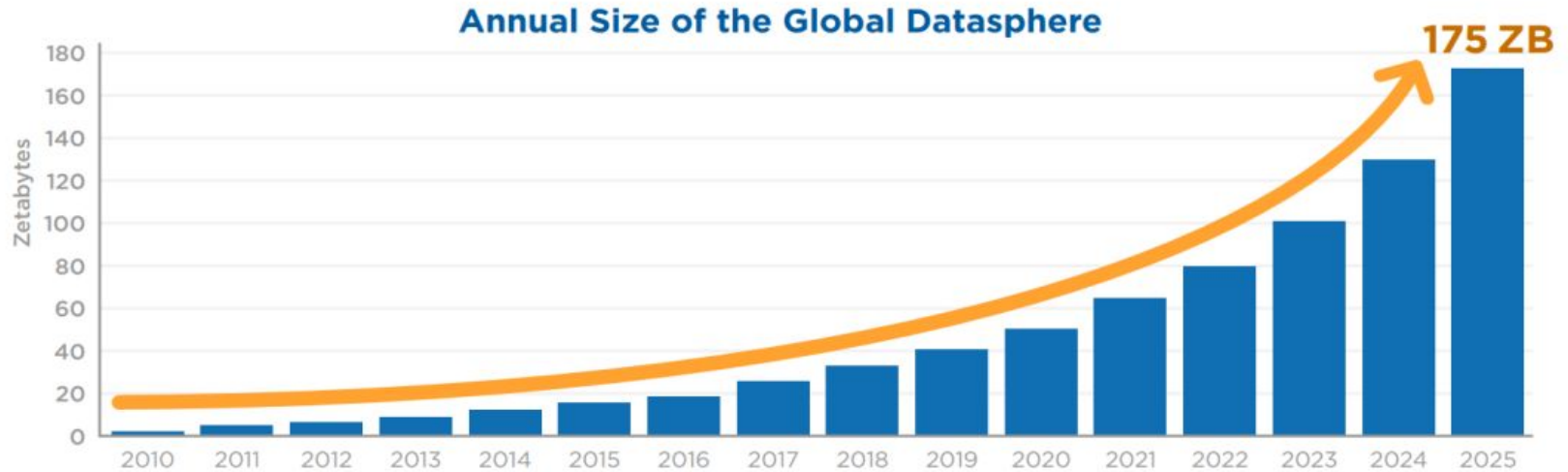
O que é Big Data?

**Big Data** foi definido em um artigo de Douglas Laney, da Gartner, como:

"...ativos de informações de alto volume, alta velocidade e/ou alta variedade que exigem formas inovadoras e econômicas de processamento de informações que permitem uma visão aprimorada, tomada de decisões e automação de processos." Partindo desta definição somos capazes de identificar três Vs: Volume, Velocidade e Variedade.

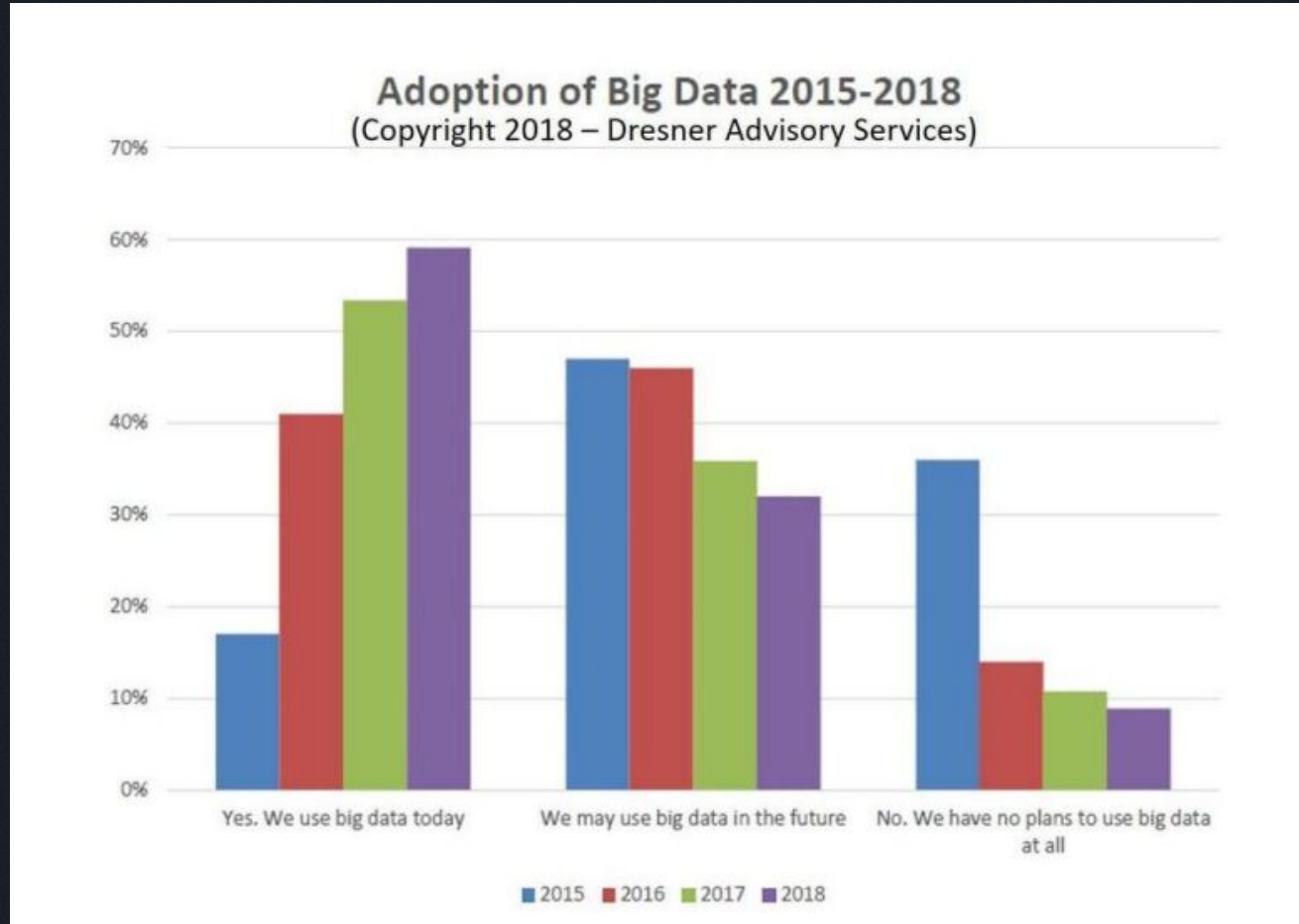
Fonte: [Gartner](#)

## Crescimento dos Dados



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

- Facebook - recebe 600 terabytes de dados por dia
- Youtube - recebe 300 horas de vídeo por minuto
- Google - 40.000 consultas de pesquisa por segundo



## 3 Vs

Partindo da definição apresentada acima, podemos notar três "Vs" comumente relacionados ao tema: Volume, Velocidade, Variedade

### Volume

Talvez seja o "V" mais lembrado desta lista e sua presença aqui é facilmente justificada: a quantidade de dados sendo produzidos, e seus tamanhos, vêm crescendo exponencialmente. Big data tem o papel de nos possibilitar usar essa gigantesca massa.



# Volume





## Velocidade

Trata-se não somente da **velocidade** com que estes dados vêm sendo gerados (mensagens em redes sociais, transações de cartões de crédito), mas também do seu ritmo não necessariamente constante que cria uma necessidade de receber e manipulá-los em momentos de pico.

## Variedade

Tradicionalmente profissionais de dados eram responsáveis por armazenar, tratar, manipular e analisar dados ditos como estruturados. Todavia, com a infinidade de dispositivos capazes de produzir e coletar informações, considerando também as diferentes formas de comunicação humana (texto, áudio, imagem), arquivos considerados relevantes à uma organização encontra-se cada vez mais em **diferentes formatos e extensões**, estruturados ou não.

## Variedade - diferentes tipos de dados

### **Estruturado**

Dados com comprimento e tipo pré-definidos agrupados em linhas e colunas (tabelar), como tabelas de bancos de dados relacionais.

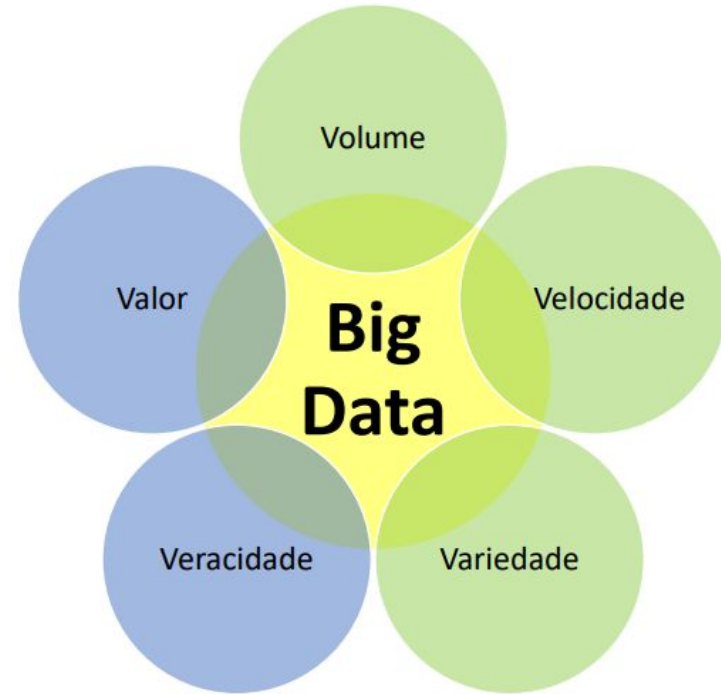
### **Semi-estruturado**

Dados que não possuem um comprimento ou tipo definido, mas tem formato padronizado, tais quais? arquivos xml, json, avro e parquet.

### **Não-estruturado**

Dados que não possuem uma estrutura ou formato padronizados, tais quais? vídeos, imagens, textos etc.

1. Volume
2. Velocidade
3. Variedade
4. Veracidade
5. Valor
6. Variabilidade
7. Validade
8. Vulnerabilidade
9. Volatilidade
10. Visualização



### Veracidade

Refere-se a qualidade dos dados que estão sendo analisados. Dados de alta veracidade tendem a ter mais valor a ser extraído se comparado a dados com baixa veracidade.

### Valor

Diz respeito ao valor que os dados geram para os usuários e para os negócios



# Exemplos de Implementação de Big Data

## Campanhas de marketing

A partir de mídias sociais é possível colher informações sobre a percepção que consumidores possuem sobre o negócio (análise de sentimento) e dessa forma direcionar de forma mais assertiva campanhas de marketing, por exemplo. Big Data passa a ser facilmente uma realidade para este cenário pelo volume de usuários em redes sociais, consequentemente de dados gerados, e variedade comumente semi ou não estruturada, como textos e vídeos.

## IoT (Internet of Things)

Dispositivos IoT, como sensores em máquinas industriais, turbinas de aviões ou estufas em plantações são capazes de gerar dados a respeito de temperatura, pressão, e muitos outros medidores relevantes para o desempenho do negócio. Em geral, a velocidade com que estes dados são gerados é alta e em logs, os caracterizando como semiestruturados, portanto.

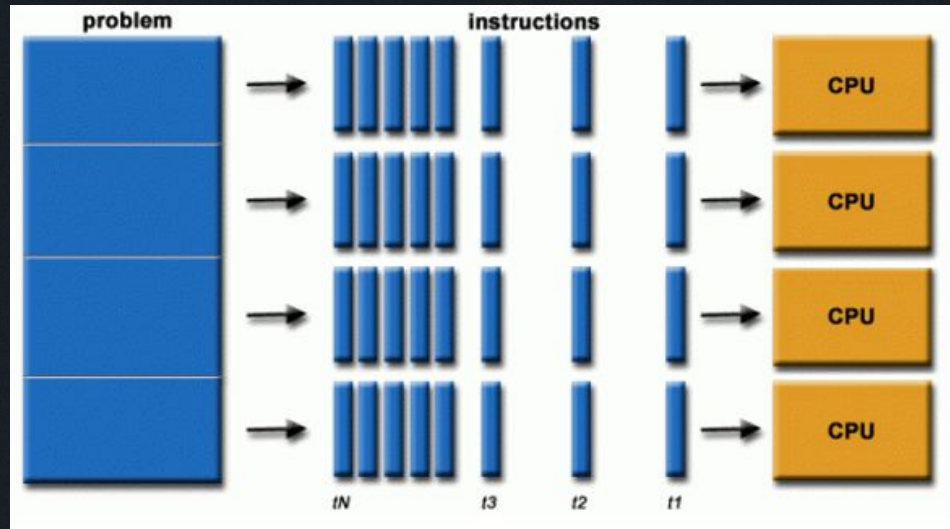
# Computação Paralela e Computação Distribuída

Assumindo então esta nova realidade onde a volumetria de dados processados é gigantesca e a velocidade com que são gerados e precisam estar disponíveis para tomada de decisão é altíssima, como ter poder computacional que atenda essa demanda?

## Computação Paralela

Este tipo de computação se caracteriza pelo uso simultâneo de várias CPUs para realizar trabalhos computacionais.

Utilizando essa técnica, é possível ultrapassar as limitações tecnológicas de uma máquina comum, aumentando sua velocidade e poder de processamento.

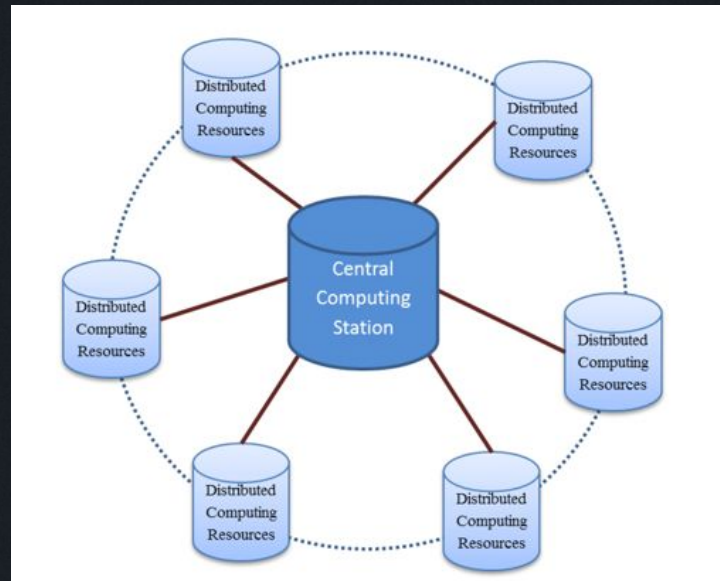


# Computação Paralela e Computação Distribuída

## Computação Distribuída

Este tipo de computação se caracteriza pela presença de uma coleção de computadores autônomos interligados através de uma rede de computadores e equipados com software que permita o compartilhamento dos recursos do sistema, tais quais hardware, software e dados.

## Cluster

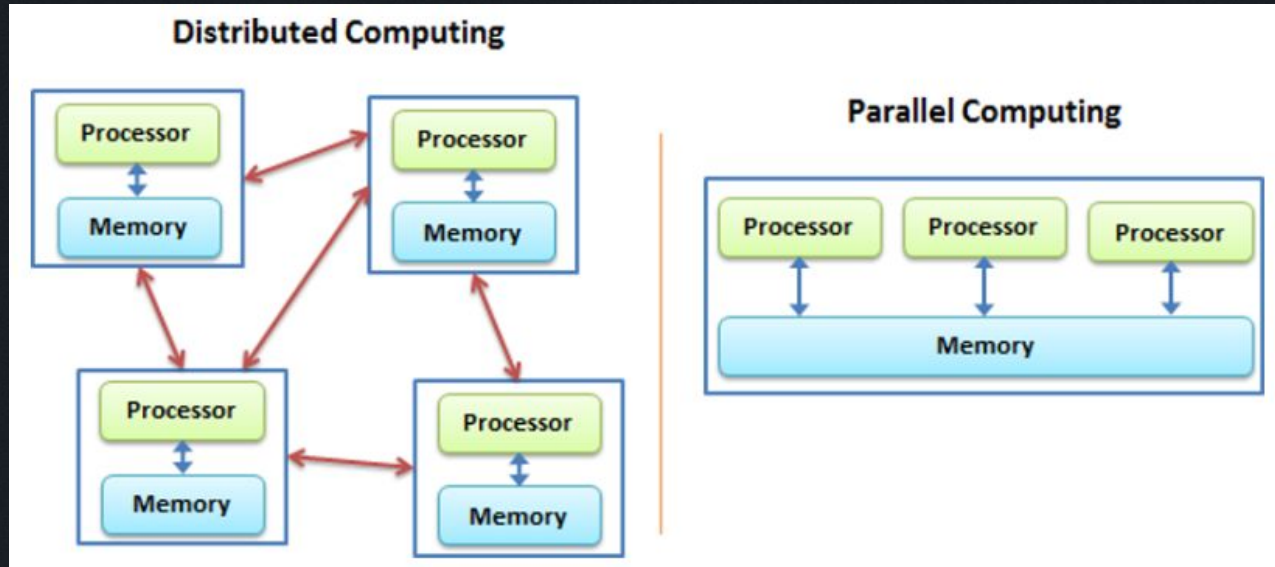




## Computação Paralela x Computação Distribuída

Perceba na imagem abaixo como através de computação paralela podemos dividir tarefas entre diferentes CPUs para que possam ser executadas simultaneamente, entretanto, ainda partilham a mesma memória e estão sob controle do mesmo sistema operacional.

Já quando utilizamos computação distribuída temos inúmeras máquinas independentes em termos de funcionamento que, através de rede e software específicos, se comunicam e executam partes menores de uma tarefa maior simultaneamente.



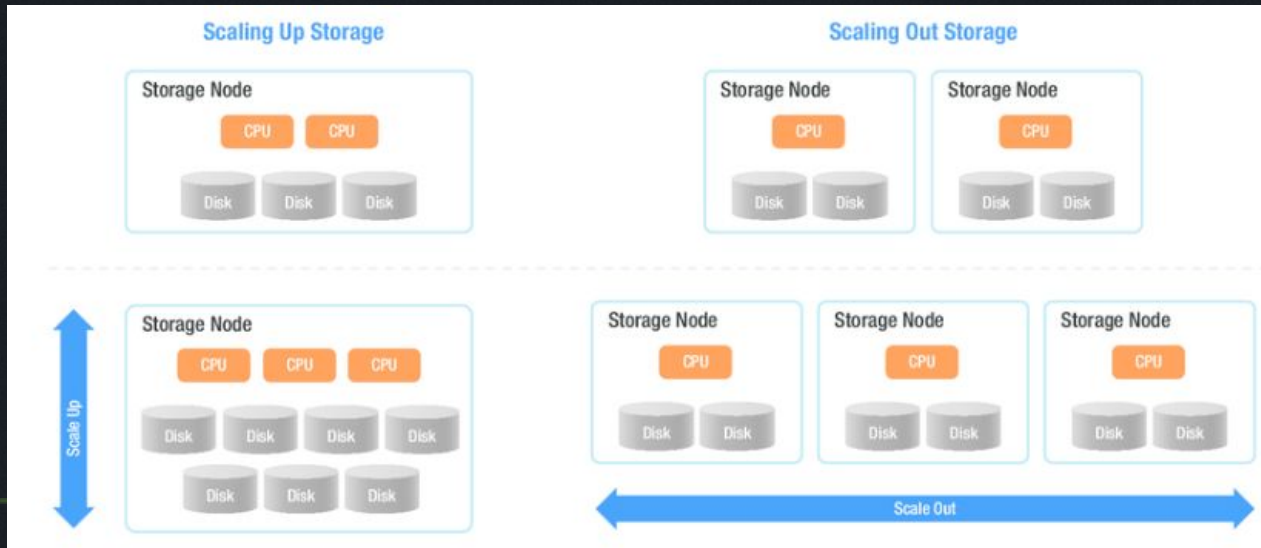
# Escalonamento Vertical x Escalonamento Horizontal

## Escalonamento vertical

Escalonamento vertical, ou scale up, tem relação direta com computação paralela, uma vez que para aumentar o poder processamento visa aumento de CPUs e de memória da máquina.

## Escalonamento horizontal

Escalonamento horizontal, ou scale out, tem relação direta com computação distribuída, uma vez que para aumentar o poder de processamento visa acrescentar mais máquinas ao cluster.



## Conclusões

Entendemos então que ambas as estratégias são de fato capazes de aumentar o poder de processamento computacional e são efetivas dentro de seus propósitos. Todavia, em cenário de Big Data, a computação se mostra muito adequada pois esbarra em poucos limites, partindo do pressuposto que é sempre possível aumentar a quantidade de nós de um cluster.



# Indicações e Bibliografias

[O que é Big Data e para que\(m\) serve](#)

[Big Bets on Big Data](#)

[Qual a diferença de Analista, Cientista e Engenheiro de Dados?](#)

[Big Data - Gartner](#)

[Domo](#)

[Paralelismo em Computadores com Tecnologia Multicore](#)

[Diferenças entre computação distribuída e computação paralela](#)

[Computação distribuída: introdução](#)

Obrig.ada