

# Megadados

Aula 1 – Apresentação do curso, introdução a sistemas de gerenciamento de bancos de dados

Fábio Ayres <[fabioja@insper.edu.br](mailto:fabioja@insper.edu.br)>



# Bem vindos!

Fábio Ayres

[fabioja@insper.edu.br](mailto:fabioja@insper.edu.br)

Aulas:

- Segundas, 7:30 – 9:30
- Quartas, 9:45 – 11:45

Atendimento:

- Terças, 10:15 – 11:45

# Objetivos de aprendizado

- Entender o que são megadados e quais os desafios inerentes a dados com esta escala, complexidade, e requisitos de performance
- Dado um problema, estabelecer uma estratégia de trabalho com megadados (integração, armazenamento, processamento, tomada de decisões)
- Projetar software analítico capaz de utilizar estratégias de computação distribuída para tratar de forma eficaz grandes volumes de dados
- Aplicar técnicas de recuperação de informação e mineração de dados.
- Descobrir e avaliar criticamente, de forma autônoma, tecnologias emergentes em big data.

# Estrutura do curso

- Parte I: Bancos de dados relacionais
  - Modelagem
  - SQL
  - Sistemas
- Parte II: Dados em larga escala
  - NoSQL
  - Processamento em lote: MapReduce e Spark
  - Máquinas de busca e recuperação de informação

# Instrumentos de avaliação

## Projetos:

- (APS1) Projeto 1: banco de dados relacional
- (APS2) Projeto 2: ETL com Spark
- Nota projetos:  $(APS1 + APS2) / 2$

## Provas:

- (P1) Avaliação intermediária
- (P2) Avaliação final
- Nota provas:  $(P1 + P2) / 2$

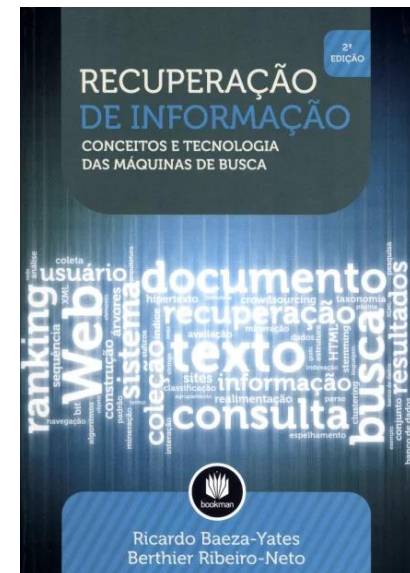
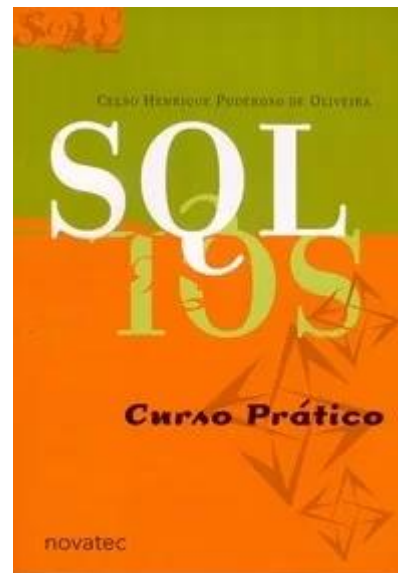
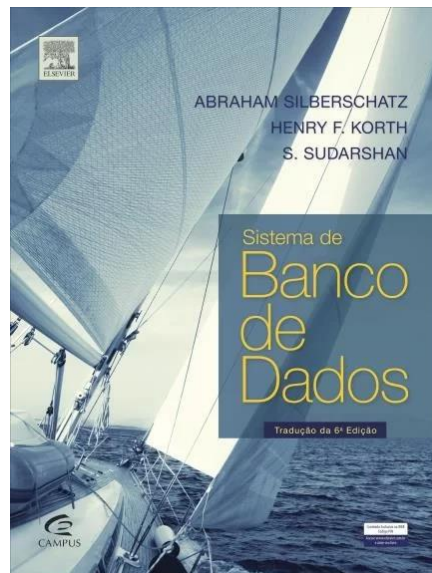
## Nota final:

- Se nota provas e nota projetos  $\geq 5$ : média provas e projetos
- Caso contrário:  $\min(\text{nota provas}, \text{nota projetos})$

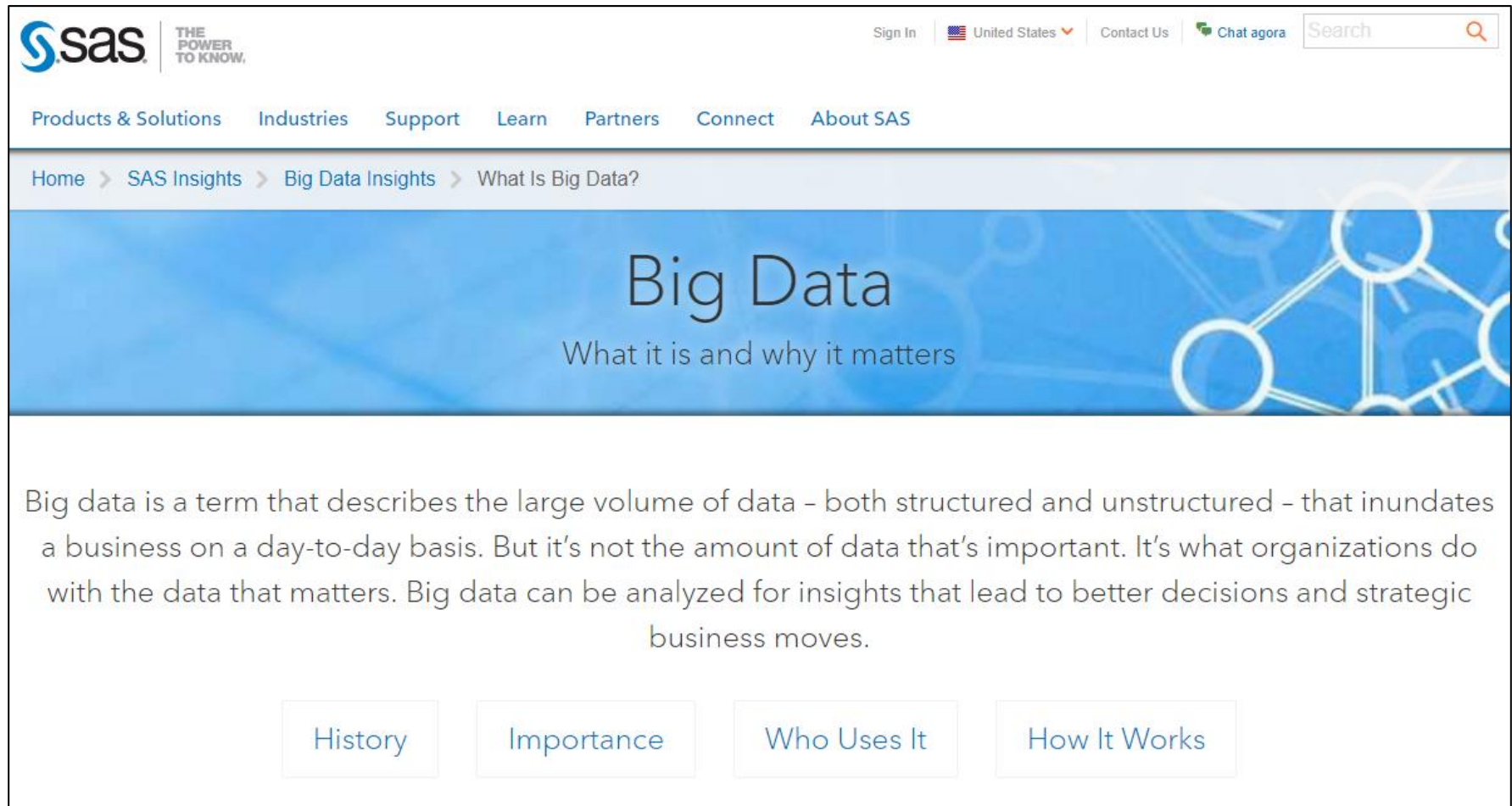
# BIBLIOGRAFIA

## BÁSICA

1	SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. <b>Sistema de Banco de Dados</b> . 6a Ed. Rio de Janeiro: Campus, 2012.
2	OLIVEIRA, C. H. P. <b>SQL: Curso Prático</b> . Novatec, 2002.
3	RIBEIRO-NETO, B.; BAEZA-YATES, R. <b>Recuperação de Informação - Conceitos e Tecnologia Das Máquinas de Busca</b> . 2ª Ed. 2013, Bookman



# O que é “Big Data”?



The screenshot shows the SAS website's 'Big Data' page. At the top is the SAS logo with the tagline 'THE POWER TO KNOW.' and navigation links for Sign In, United States, Contact Us, Chat agora, and a Search bar. Below this is a main navigation bar with links for Products & Solutions, Industries, Support, Learn, Partners, Connect, and About SAS. A breadcrumb trail reads: Home > SAS Insights > Big Data Insights > What Is Big Data?. The main content area has a blue header with the title 'Big Data' and subtitle 'What it is and why it matters', accompanied by a network diagram. The text below explains that big data is a term for large volumes of structured and unstructured data that inundates businesses daily, emphasizing that it's not just the volume but the insights derived that matter. At the bottom, four buttons are provided: History, Importance, Who Uses It, and How It Works.

**SAS** | THE POWER TO KNOW.

Sign In | United States | Contact Us | Chat agora | Search

Products & Solutions | Industries | Support | Learn | Partners | Connect | About SAS

Home > SAS Insights > Big Data Insights > What Is Big Data?

## Big Data

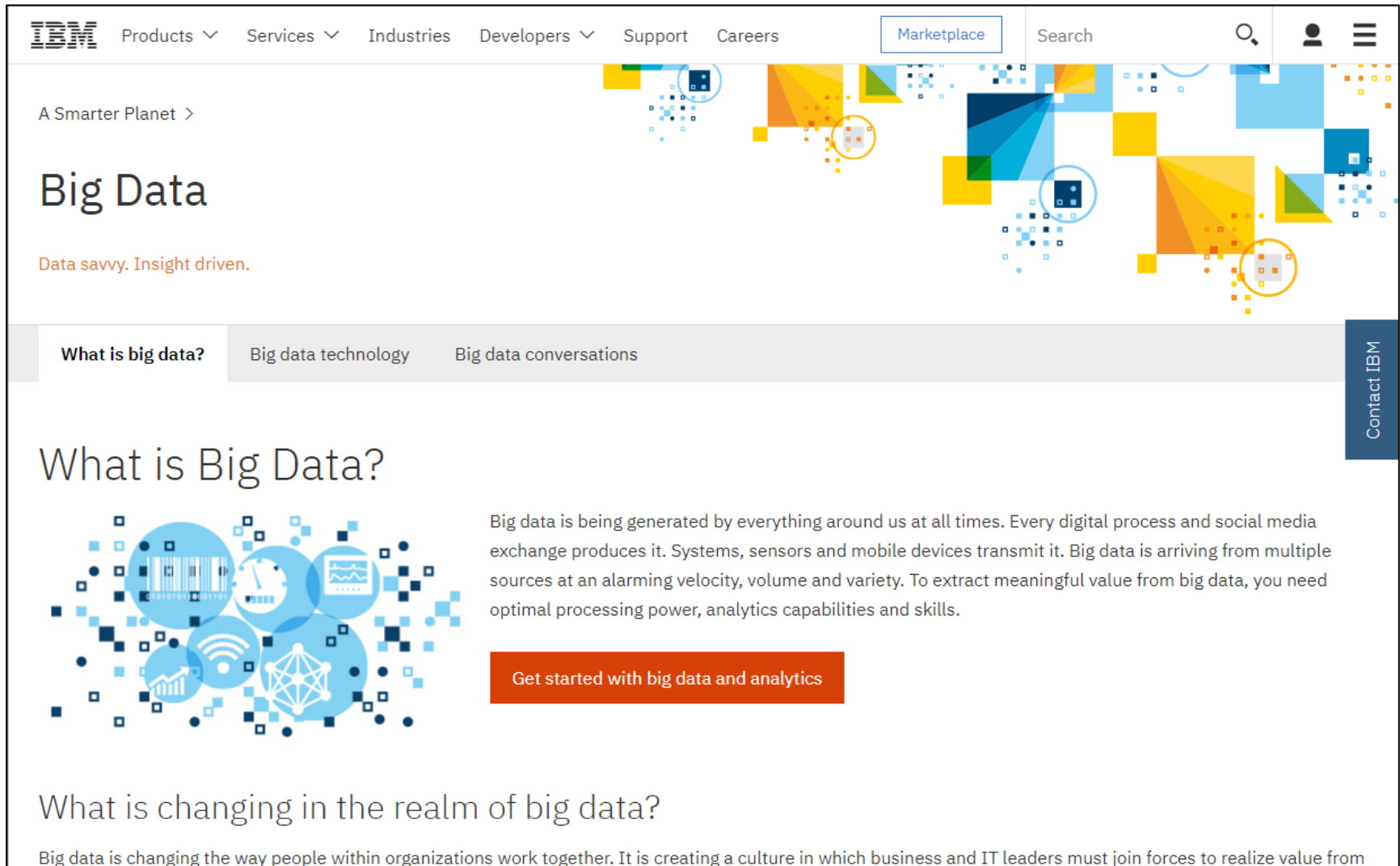
What it is and why it matters

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

History | Importance | Who Uses It | How It Works



# Big Data??



The screenshot shows the IBM Big Data website. At the top is the IBM logo and a navigation bar with links for Products, Services, Industries, Developers, Support, and Careers. A 'Marketplace' button and a search bar are also present. Below the navigation bar is a large graphic with colorful squares and circles. The main heading is 'Big Data' with the tagline 'Data savvy. Insight driven.' Below this is a horizontal menu with 'What is big data?' (selected), 'Big data technology', and 'Big data conversations'. The 'What is Big Data?' section features a graphic of various data-related icons (barcode, clock, line graph, Wi-Fi, network) and a text block explaining that big data is generated by everything around us at all times. A red button labeled 'Get started with big data and analytics' is positioned below the text. At the bottom, the section 'What is changing in the realm of big data?' is partially visible, with text stating that big data is changing the way people within organizations work together.

IBM Products Services Industries Developers Support Careers Marketplace Search


A Smarter Planet >

## Big Data

Data savvy. Insight driven.

What is big data? Big data technology Big data conversations

### What is Big Data?



Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety. To extract meaningful value from big data, you need optimal processing power, analytics capabilities and skills.

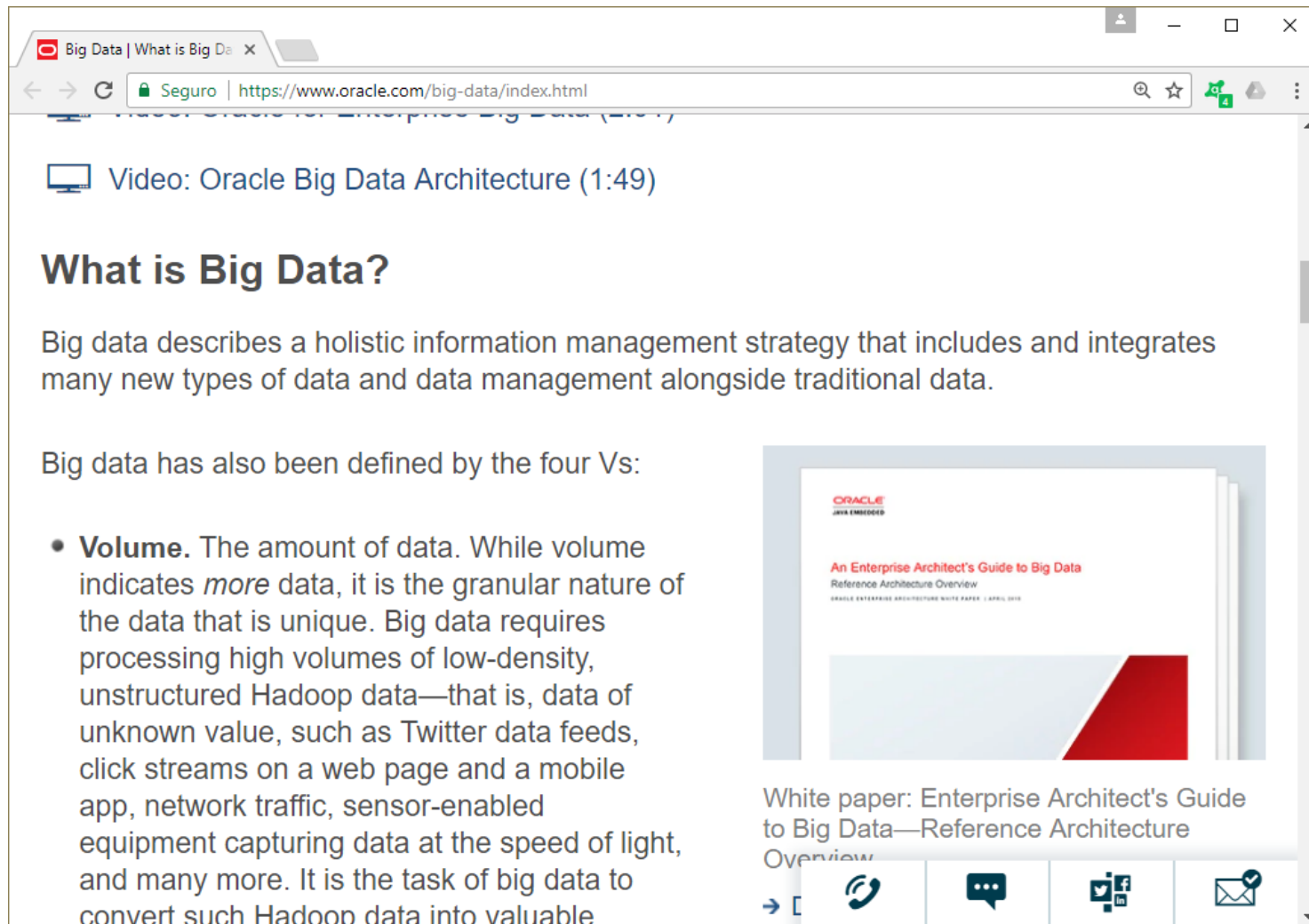
[Get started with big data and analytics](#)

### What is changing in the realm of big data?

Big data is changing the way people within organizations work together. It is creating a culture in which business and IT leaders must join forces to realize value from



# Big Data???

A screenshot of a web browser displaying the Oracle Big Data website. The browser's address bar shows the URL "https://www.oracle.com/big-data/index.html". The page content includes a video player for "Video: Oracle Big Data Architecture (1:49)", a section titled "What is Big Data?", and a definition of big data as a holistic information management strategy. Below this, it states that big data is defined by the four Vs, with the first bullet point for "Volume" describing the amount and granular nature of data. To the right, there is a thumbnail for a white paper titled "An Enterprise Architect's Guide to Big Data Reference Architecture Overview".

Big Data | What is Big Data

Seguro | <https://www.oracle.com/big-data/index.html>

Video: Oracle Big Data Architecture (1:49)

## What is Big Data?

Big data describes a holistic information management strategy that includes and integrates many new types of data and data management alongside traditional data.

Big data has also been defined by the four Vs:

- **Volume.** The amount of data. While volume indicates *more* data, it is the granular nature of the data that is unique. Big data requires processing high volumes of low-density, unstructured Hadoop data—that is, data of unknown value, such as Twitter data feeds, click streams on a web page and a mobile app, network traffic, sensor-enabled equipment capturing data at the speed of light, and many more. It is the task of big data to convert such Hadoop data into valuable

White paper: Enterprise Architect's Guide to Big Data—Reference Architecture Overview


# Big Data????

W Big data - Wikipedia

← → ↻ Seguro | [https://en.wikipedia.org/wiki/Big\\_data](https://en.wikipedia.org/wiki/Big_data) ☆ 🔍 🔒 ⋮

Not logged in [Talk](#) [Contributions](#) [Create account](#) [Log in](#)

Article [Talk](#) [Read](#) [View source](#) [View history](#)



**WIKIPEDIA**  
The Free Encyclopedia

[Main page](#)  
[Contents](#)  
[Featured content](#)  
[Current events](#)  
[Random article](#)  
[Donate to Wikipedia](#)  
[Wikipedia store](#)

[Interaction](#)  
[Help](#)  
[About Wikipedia](#)  
[Community portal](#)  
[Recent changes](#)  
[Contact page](#)

[Tools](#)  
[What links here](#)  
[Related changes](#)  
[Upload file](#)  
[Special pages](#)  
[Permanent link](#)  
[Page information](#)

## Big data

From Wikipedia, the free encyclopedia

*This article is about large collections of data. For the band, see [Big Data \(band\)](#).*

**Big data** is a term for **data sets** that are so large or complex that traditional **data processing application software** is inadequate to deal with them. Big data challenges include **capturing data**, **data storage**, **data analysis**, **search**, **sharing**, **transfer**, **visualization**, **querying**, **updating** and **information privacy**.

Lately, the term "big data" tends to refer to the use of **predictive analytics**, **user behavior analytics**, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem."<sup>[2]</sup> Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on."<sup>[3]</sup> Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-

**Global Information Storage Capacity**  
in optimally compressed bytes

Year	Storage Type	Capacity (Exabytes)	% of Total
1986	ANALOG	2.6	100%
1986	DIGITAL	0.02	1%
1993	ANALOG	3	3%
1993	DIGITAL	0.02	1%
2000	ANALOG	25	25%
2000	DIGITAL	25	25%
2002	ANALOG	50	50%
2002	DIGITAL	50	50%
2007	ANALOG	19	94%
2007	DIGITAL	280	94%

**ANALOG STORAGE (1986-2007):**

- Paper, film, audiotape and vinyl: 8%
- Analog videotape (S-VHS, Hi8, Hi8i): 94%
- Portable media, flash drives: 2%
- Portable hard disks: 2.4%
- CDs and record disks: 6.8%

**DIGITAL STORAGE (2000-2007):**

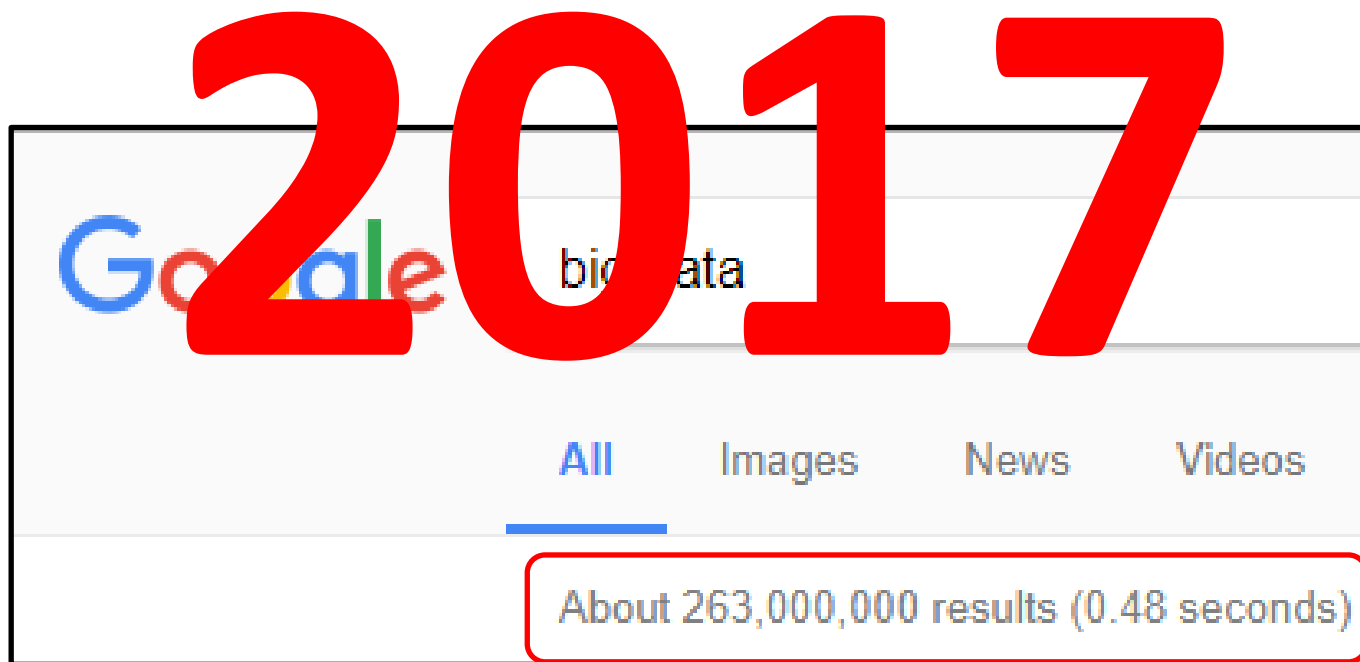
- Computer servers and mainframes: 8.9%
- Digital tape: 11.8%
- DVD/Blu-ray: 22.8%
- PC hard disks: 44.5%
- 12.8 billion gigabytes
- Other: < 1% (incl. chip cards, memory cards, floppy disks, mobile phones, PDA, cameras, camcorders, videogames)

Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60-65. <http://www.martin-hilbert.net/WorldInfoCapacity.html>

Growth of and digitization of global information-storage capacity<sup>[1]</sup>

Insper

# Big Data!



Olha o big data aqui!

Big Data!

# 2019



Olha o big data aqui!

# Atividade

Em grupos de 2-3 pessoas, discutam e respondam por escrito as seguintes perguntas:

- O que é big data?
- O que faz um cientista de dados?
- O que faz um engenheiro de dados?
- Quanto ganha um engenheiro de dados? Como você explica isso?

**10 min**

# Atividade

Quão “big” é “big data”? Troque de grupo e respondam por escrito as seguintes perguntas:

- Qual o tamanho da web? Qual a sua definição de “web”?
- O que mais existe, em big data, além da web? Liste 3 fontes de dados (atuais ou futuras) que você acha que geram “big data”. Para cada uma estime o volume de dados gerado (por unidade de tempo) e armazenado.
- O CommonCrawl está disponível no AWS S3, gratuitamente. O que é o CommonCrawl? Qual o tamanho? Quanto tempo levaria para baixar tudo, e como fazê-lo?

**10 min**

# Atividade

Muitos dados, mas e para processar tudo isso? Troque de grupo e respondam por escrito:

- Qual a máquina mais poderosa, em termos de CPU, RAM e disco, que está disponível na Amazon Web Services? Quanto custa? (Vamos ignorar máquinas com GPUs por agora, deixa “Supercomputação” e “Machine Learning” lidarem com isso.) Qual o preço de uma máquina “mais em conta” na Amazon?
- Como vocês acham que podemos fazer para processar 3 bilhões de páginas em poucas horas? Esboce um sistema computacional para fazer isso: indique o hardware completo, e a arquitetura de software. Quanto custaria na AWS? Quanto custaria ter esse hardware todo e desenvolver o software inteiro por conta própria?

**10 min**



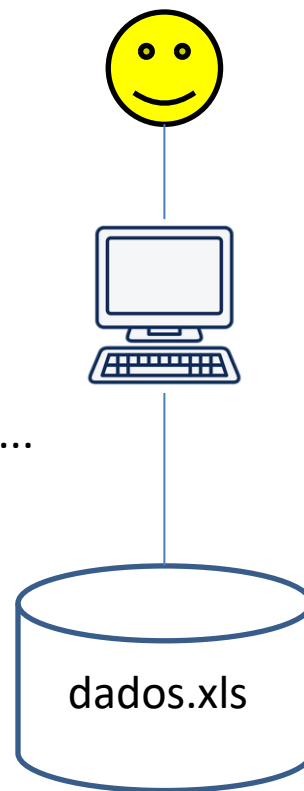
# **Sistemas de gerenciamento de bancos de dados**

# Banco de dados: porque?

- Você não precisa sempre de um banco de dados para armazenar dados! Você pode usar:
  - Um guardanapo de papel!
  - Um arquivo de texto no seu laptop!
  - Uma planilha Excel!
  - etc...
- Quando será que precisamos de um banco de dados?

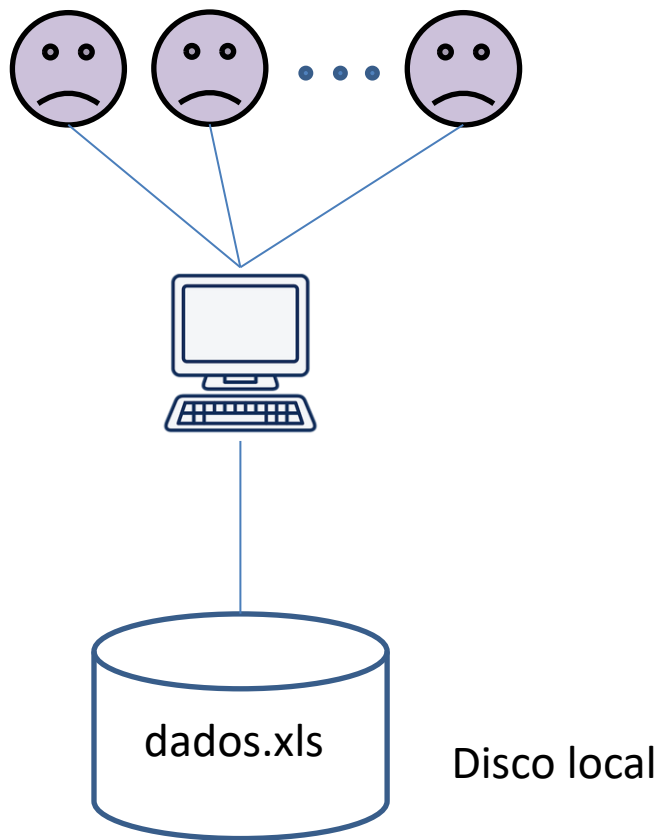
# Exemplo

Só um usuário, arquivo pequeno...  
Não precisa de banco de dados!

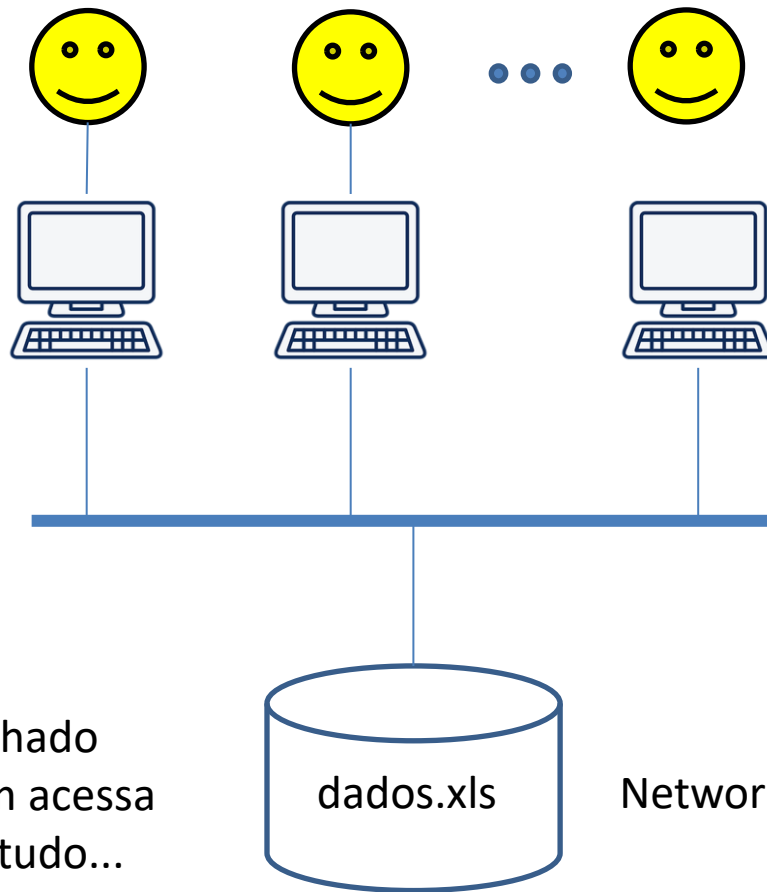


# Exemplo

Muitos usuários formando  
fila para conseguir acessar o  
terminal!

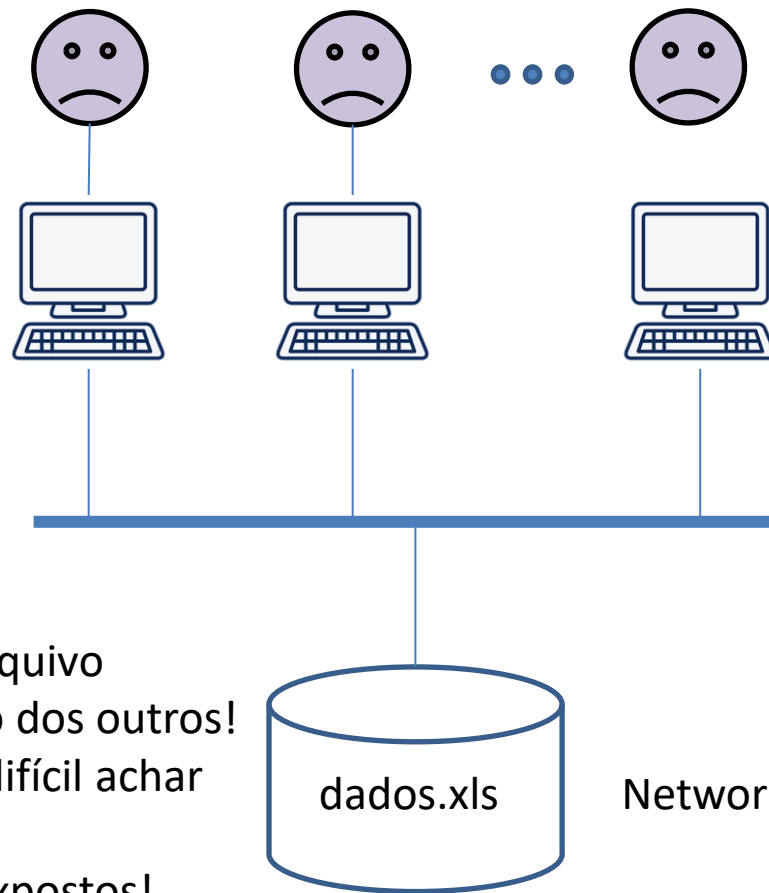


# Exemplo



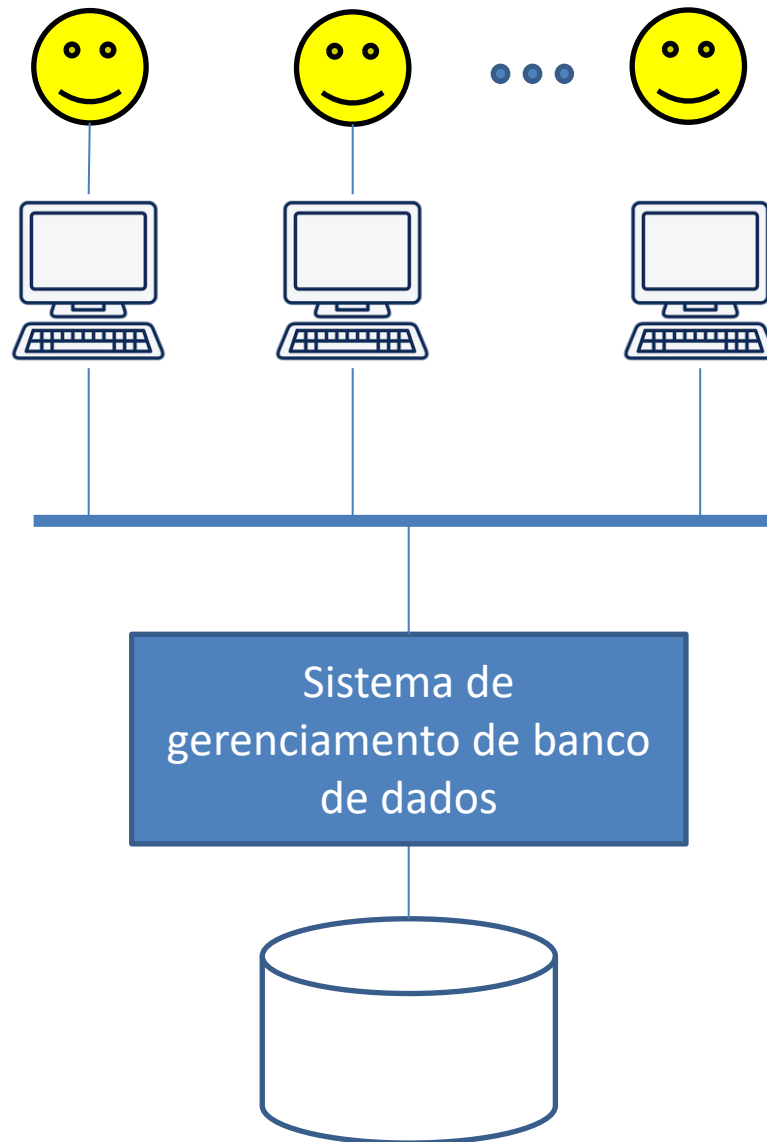
Arquivo está compartilhado na rede agora, cada um acessa com seu terminal. Contudo...

# Exemplo



- Não dá para trabalhar simultaneamente no arquivo sem estragar o trabalho dos outros!
- Dados muito grandes, difícil achar resultados
- Dados sigilosos estão expostos!

# Exemplo





# Motivos para ter um sistema de gerenciamento de banco de dados

- Tamanho

Pode não caber na RAM!

- Velocidade

Existem maneiras eficientes de armazenar e recuperar dados!

- Conveniência

O SGBD já vem com mecanismos sofisticados de consulta!

- Precisão

Um bom projeto evita redundâncias gerencia transações e mantém consistência!

- Proteção

Controle de acesso e registro de atividade!

- Robustez

Proteção contra falhas!

# Aplicações

- Vendas e estoque
- Recursos humanos e informações empresariais
- Dados científicos
- Informações geográficas
- Multimídia
- Jogos online
- Etc!



Sistema de informação

Aplicação/Serviço de consulta



Sistema de Gerenciamento de Banco de Dados

Processamento de consultas

Acesso aos dados



Armazenamento

# Tipos de banco de dados

- Relacional (também chamado de bancos de dados SQL): representa os dados usando o modelo relacional, onde dados são representados através de tabelas bidimensionais.
  - Este é o modelo mais usado em bancos de dados atualmente

# Tipos de banco de dados

- NoSQL: bancos de dados não-relacionais, dentre os quais destacam-se:
  - Key-value stores (e.g. Redis)
  - Document stores (e.g. MongoDB)
  - Column-oriented (e.g. Cassandra)

(Artigo interessante:

<http://www.dataversity.net/review-pros-cons-different-databases-relational-versus-non-relational/>)

# Tipos de banco de dados

- NewSQL: Nova geração de bancos de dados que mesclam as vantagens de alguns tipos de bancos NoSQL (como escalabilidade e disponibilidade) com garantias de consistência transacional do SQL.
  - Exemplo: Google Spanner

(Artigo interessante:

<http://www.odbms.org/blog/2018/03/on-rdbms-nosql-and-newsql-databases-interview-with-john-ryan/>)

# Para a próxima aula

- Definir grupos para o primeiro projeto

## Instalar

- Anaconda ou alguma versão de Python 3 com Jupyter Notebook
- MySQL Community Server
- MySQL Workbench



# Insper

[www.insper.edu.br](http://www.insper.edu.br)