

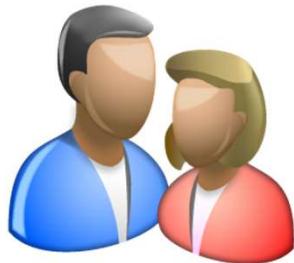
**FIAP GRADUAÇÃO**

# DATA SCIENCE

BIG DATA ARCHITECTURING & DATA INTEGRATION

Prof. Dr. Renê de Ávila Mendes

# Você



Fale um pouco sobre você:

- Sua faixa etária
- Trabalha ou não na área
- Objetivo com o curso





## Renê

- Graduação em Sistemas de Informação (2007)
- Mestrado em Engenharia Elétrica (2017) – Arquiteturas para Big Data
- Doutorado em E.E. e Computação (2021) – Complexidade de Dados
- Administrador de Sistemas no Instituto Presbiteriano Mackenzie
- Professor de matérias de Big Data (Mineração de Dados, Coleta de Dados e Arquitetura para Big Data)
- pf1752@fiap.com.br
- [www.linkedin.com/in/renemendes](https://www.linkedin.com/in/renemendes)

# Objetivos da disciplina

**DISCIPLINA:** Big Data Architecturing & Data Integration

**OBJETIVOS:** Entenda as principais **arquiteturas** para ingestão, processamento e análise de grandes volumes de dados. Conheça as principais **ferramentas** open-source de Big Data como Hadoop, MapReduce, Spark, Sqoop, NiFi, Flume, Kafka, Zookeeper, HBase, Hive e as **integre** com as ferramentas de **extração, transformação e carga** de dados em modelos dimensionais. Entenda conceitos sobre **computação paralela e distribuída**, aplicação do Hadoop e bases Apache e arquiteturas *serverless* e desacopladas. Veja como **visualizar os dados** estruturados ou não estruturados com ferramentas de Self-Service Business Intelligence como PowerBI, utilizando as melhores práticas de **visualização de dados**.

# **Assuntos – 1º Semestre**

- Introdução a Big Data
- Conceitos Computação Paralela e Distribuída, Lei de Moore, Sistema HDFS
- Aplicação Hadoop. Usos e Administração de Ambientes Hadoop
- Introdução a MapReduce
- Introdução à Integração de Dados
- Integração entre SQL e Hadoop - SQOOP
- Bases Apache
- Bases Apache - PIG
- Introdução da Data Streaming - FLUME
- Introdução a análise de dados com SPARK

PROJECT CHECK POINT  
CHALLENGE AND  
FEEDBACKS

**40%**

1 SEMANA DE FEEDBACK

GLOBAL SOLUTION

**60%**

1 SEMANA DE FEEDBACK

Primeiro Semestre = 40%

PROJECT CHECK POINT  
CHALLENGE AND  
FEEDBACKS

**40%**

1 SEMANA DE FEEDBACK

GLOBAL SOLUTION

**60%**

1 SEMANA DE FEEDBACK

Segundo Semestre = 60%

MODELO AVALIATIVO ANUAL

A Média Anual (MA), será obtida pela média das duas Médias Semestrais.

$$\text{MA} = (40\% \times \text{MS 1º semestre} + 60\% \times \text{MS 2º semestre})$$

Os critérios de aprovação baseiam-se na média semestral (para cursos semestrais) ou na média anual (para cursos anuais) obtida pelo aluno, conforme tabela abaixo:

MÉDIA FINAL (ANUAL OU SEMESTRAL)	SITUAÇÃO
0,0 a 3,9	Reprovado
4,0 a 5,9	Exame
6,0 a 10,0	Aprovado

CHALLENGE

# PROJECT-BASED LEARNING

APRENDIZADO PRÁTICO PARA SOLUÇÕES DE PROBLEMAS REAIS

## 1a. Série 2023

- TEMA EM PROCESSO DE DEFINIÇÃO
- LIVE PARA O KICK OFF EM  
MARÇO/2023



**Safra**



→ Mais sobre  
[Carreira](#)  
[Minhas Finanças](#)  
[Negócios](#)  
[Bancos](#)  
[Emprego](#)  
[Fintech](#)  
[Open Banking](#)  
[Tecnologia](#)  
[Vagas](#)

**Bradesco**

Engenheiros de softwares,  
engenheiros de dados, diversas  
posições em nuvem, cientista de  
dados, profissionais de data  
analytics, experiência do cliente

Mais de 1 mil vagas

**Santander**

Cientistas de dados,  
desenvolvedores, arquitetos (as)  
de softwares

Mais de 200 vagas

## Open Banking abre milhares de vagas em tecnologia, mas faltam profissionais

Executivos de diversas instituições contam os desafios de contratar profissionais  
diante da chegada do Open Banking e compartilham as vagas abertas

Por [Giovanna Sutto](#), [Sérgio Teixeira Jr.](#)  
12 ago 2021 08h00 - Atualizado 4 dias atrás

**Itaú**

Engenheiros de softwares,  
cientista de dados, profissionais  
de data analytics

Mais de 2 mil vagas

**Fonte:** <https://www.infomoney.com.br/carreira/open-banking-abre-milhares-de-vagas-em-tecnologia-mas-faltam-profissionais/>

**VC S/A**

ASSINE

DO MÊS | EDIÇÕES ANTERIORES | FINANÇAS PESSOAIS | CARREIRA | EMPREENDEDORISMO | ECONOMIA | DESE  
PESSOAL

CARREIRA

**Procuram-se profissionais de Tecnologia  
da Informação**

Entenda melhor a questão da falta de mão de obra e as possibilidades que o segmento oferece para quem deseja mergulhar nele.

Por **Monique Lima** Atualizado em 18 mar 2021, 10h07 - Publicado em 8 mar 2021, 06h00

**ESPECIALIDADES EM ALTA**

São muitas as áreas da TI, mas nem todas estão em ebulição.  
Veja o top 5 de profissões do setor em alta, segundo o LinkedIn:

ÁREA	CONHECIMENTOS PRIMORDIAIS
ENGENHEIRO DE CIBERSEGURANÇA	Docker Products; Ansible; DevOps; AWS; Kubernetes.
CIENTISTA DE DADOS	Machine Learning; Python; linguagem R.
ENGENHEIRO	Apache Spark; Apache Hadoop; Apache Hive; Python.
ESPECIALISTA EM INTELIGÊNCIA ARTIFICIAL	Machine Learning; deep learning; Python; ciência de dados.
PROGRAMADOR DE JAVASCRIPT	React.js; Node.js; AngularJS; Git; e MongoDB.

**Fonte:** <https://vocesa.abril.com.br/carreira/procuram-se-profissionais-de-tecnologia-da-informacao/>

# Mercado – Administrador de Redes e DBA (08/2021)

The image shows two separate LinkedIn search results pages. The top section is for 'Administrador de redes in Brazil' with 891 results, and the bottom section is for 'Dba in Brazil' with 243 results. Both sections include filters for 'Jobs', 'Date Posted', 'Experience Level', and 'Company'. Each result page has a search bar at the top, a job alert toggle switch, and a settings gear icon.

**Administrador de redes in Brazil**  
891 results

**Dba in Brazil**  
243 results

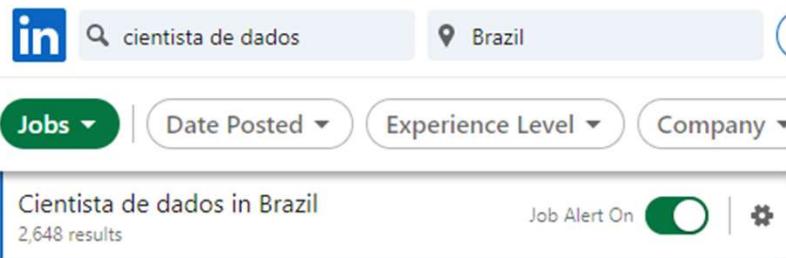
# Mercado – Administrador de Redes e DBA (08/2022)

The screenshot shows two LinkedIn search results pages. The top search is for "Administrador de redes" in Brazil, resulting in 405 jobs. The bottom search is for "DBA" in Brazil, resulting in 220 jobs. Both searches include filters for "Jobs", "Date Posted", "Experience Level", and "Company".

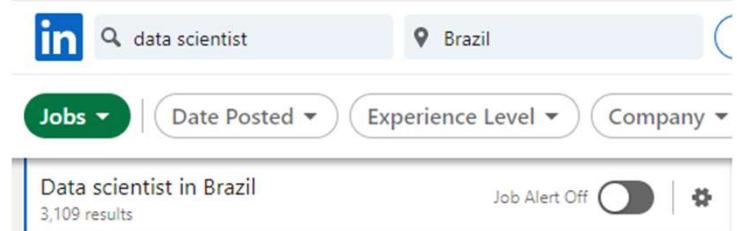
**Search 1: Administrador de redes in Brazil**  
405 results

**Search 2: DBA in Brazil**  
220 results

# Mercado – Cientista de Dados (08/2021)



LinkedIn search results for "cientista de dados" in Brazil. The search bar shows "cientista de dados" and "Brazil". Filter options include "Jobs", "Date Posted", "Experience Level", and "Company". The results section shows "Cientista de dados in Brazil" with "2,648 results". A "Job Alert On" toggle switch is shown as "On".



LinkedIn search results for "data scientist" in Brazil. The search bar shows "data scientist" and "Brazil". Filter options include "Jobs", "Date Posted", "Experience Level", and "Company". The results section shows "Data scientist in Brazil" with "3,109 results". A "Job Alert Off" toggle switch is shown as "Off".

# Mercado – Cientista de Dados (08/2022)

The image shows two separate LinkedIn search results for "data scientist" in Brazil. Both results are identical, displaying the same search interface and results count.

**Search Bar:** The search bar at the top contains the text "cientista de dados" and "Brazil".

**Filter Options:** Below the search bar are filter buttons for "Jobs", "Date Posted", "Experience Level", "Company", and "Job".

**Results Summary:** A blue header bar indicates "Cientista de dados in Brazil" and "14,362 results".

**Second Search Result:** Below the first result, another identical search interface is shown for "data scientist" in Brazil, with a results count of "14,492 results".

# Mercado – Engenheiro de Dados (08/2021)

The image displays two separate LinkedIn search results for "Engenheiro de dados" and "data engineer" in Brazil, illustrating the high demand for data engineers in the market.

**Top Search:** Engenheiro de dados in Brazil

- 1,740 results
- Job Alert On (switch is green)

**Bottom Search:** data engineer in Brazil

- 3,194 results
- Job Alert On (switch is green)

# Mercado – Engenheiro de Dados (08/2022)

The screenshot shows two LinkedIn search results for data-related roles in Brazil. The top search is for 'Engenheiro de dados' (2,165 results) and the bottom search is for 'data engineer' (3,863 results). Both searches include filters for 'Jobs', 'Date Posted', 'Experience Level', and 'Company'. Each search result has a 'Set alert' button.

Engenheiro de dados in Brazil  
2,165 results

Set alert

data engineer in Brazil  
3,863 results

Set alert



# Algumas definições

## **Data Analysis – Análise de Dados**

- A análise de dados é o processo de **examinar dados para encontrar fatos, relacionamentos, padrões, insights e/ou tendências.**
- O objetivo geral da análise de dados é **apoiar uma melhor tomada de decisão.**
- Um exemplo simples de análise de dados é a análise dos dados de vendas de sorvetes para determinar como o número de casquinhas de sorvete vendidas está relacionado à temperatura diária. Os resultados dessa análise apoiariam decisões relacionadas à quantidade de sorvete que uma loja deve pedir em relação às informações de previsão do tempo.
- A análise de dados ajuda a **estabelecer padrões e relacionamentos** entre os dados que estão sendo analisados.

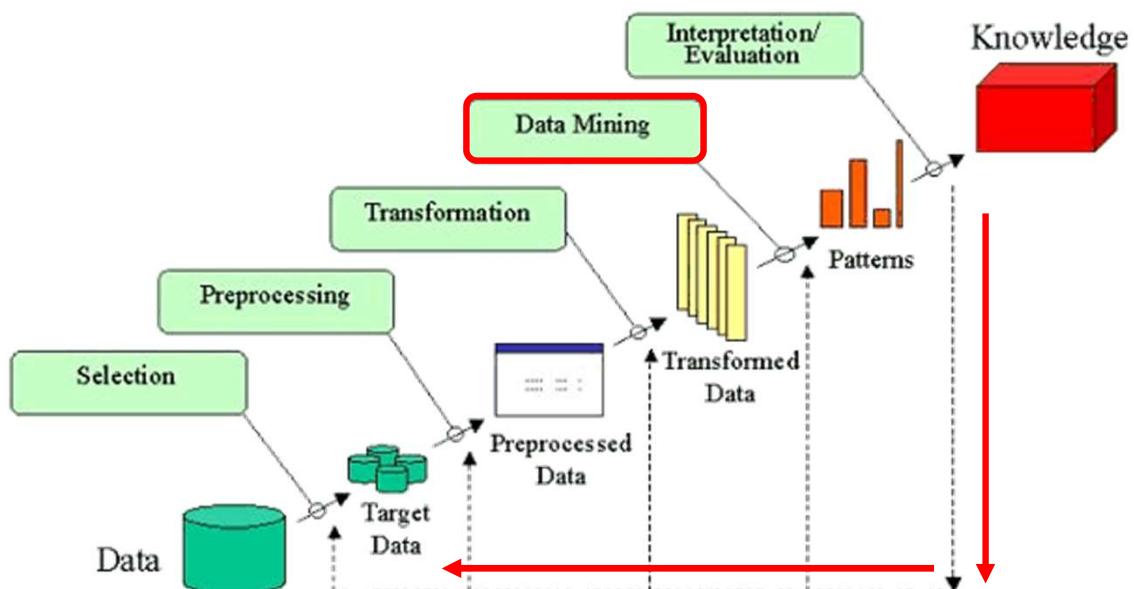
ERL, Thomas; KHATTAK, Wajid; BUHLER, Paul. **Big data fundamentals: concepts, drivers & techniques.** Prentice Hall Press, 2016.

## **Data Analytics – Analítica dos Dados**

- Analítica do dados é um **termo mais amplo** que abrange a análise de dados.
- A analítica dos dados é uma disciplina que inclui o **gerenciamento de todo o ciclo de vida dos dados**, que engloba a coleta, limpeza, organização, armazenamento, análise e controle de dados.
- O termo inclui o desenvolvimento de métodos de análise, técnicas científicas e ferramentas automatizadas.
- Em ambientes de Big Data, *data analytics* desenvolveu métodos que permitem que a análise de dados ocorra através do uso de estruturas e tecnologias distribuídas altamente escaláveis, capazes de analisar grandes volumes de dados de diferentes fontes.

ERL, Thomas; KHATTAK, Wajid; BUHLER, Paul. **Big data fundamentals: concepts, drivers & techniques**. Prentice Hall Press, 2016.

# Descoberta de Conhecimento em Bases de Dados

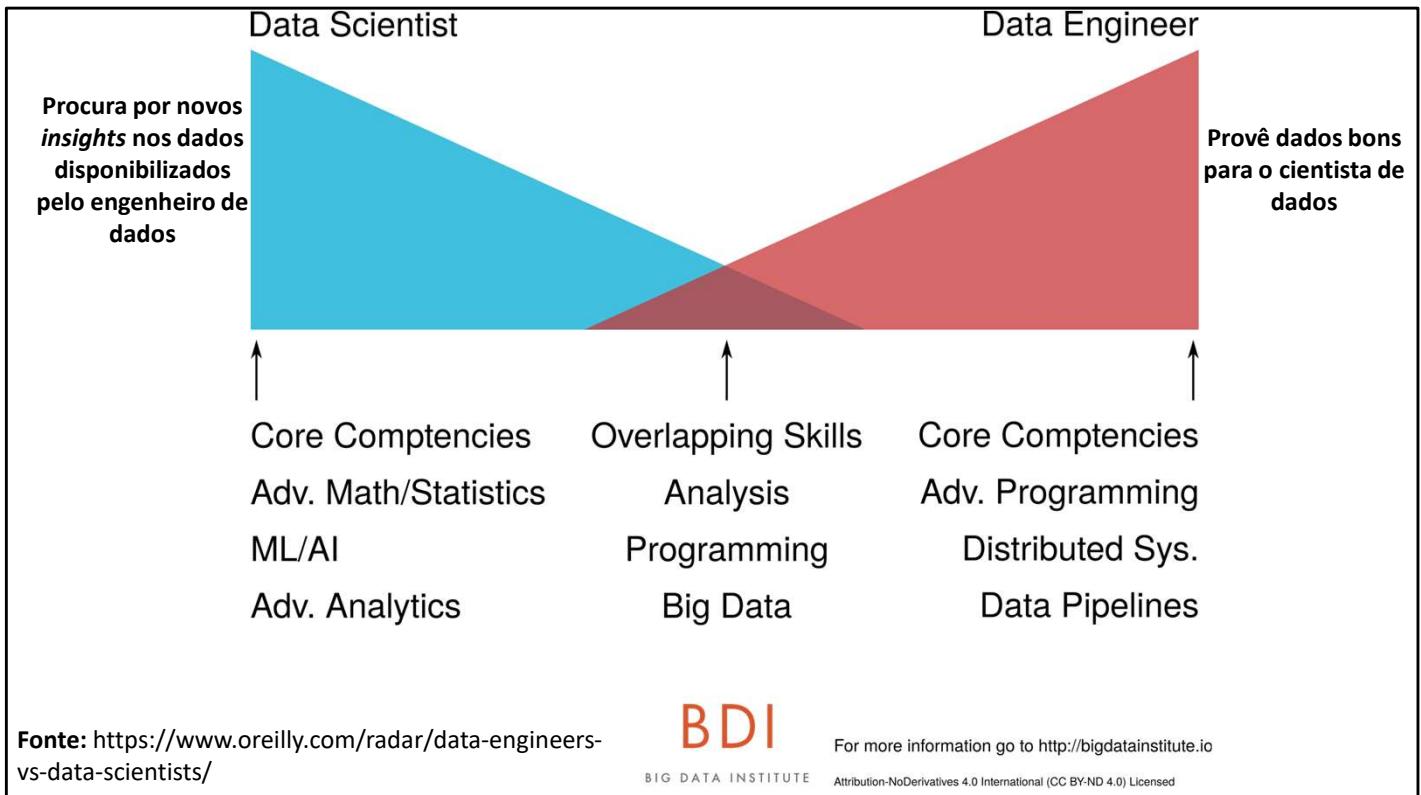


- FAYYAD, Usama; PAREKH, S. SHARKEY, Gregory; SIMHAPANI, Vaishali. From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, n. 3, p. 67, 1996.
- DM é parte de um processo de descoberta de conhecimento a partir dos dados.
  - Entre as tarefas de mineração está a classificação de dados
  - Os passos de **preparação, seleção e limpeza dos dados** e a incorporação de conhecimento prévio são essenciais para garantir a utilidade dos resultados

MAS O DESEMPENHO DA CLASSIFICAÇÃO É CONHECIDO APENAS DEPOIS DA EXECUÇÃO

Processo custoso

Se a classificação for ruim, será necessário reiniciar o processo



## Engenheiro de Dados

- Alguém que se especializou em **criar soluções de software** usando big data e para **big data**
- Conhece **programação (Java, Scala, Python)**, **sistemas distribuídos** e **ferramentas big data**
- Cria ***pipelines* de dados**

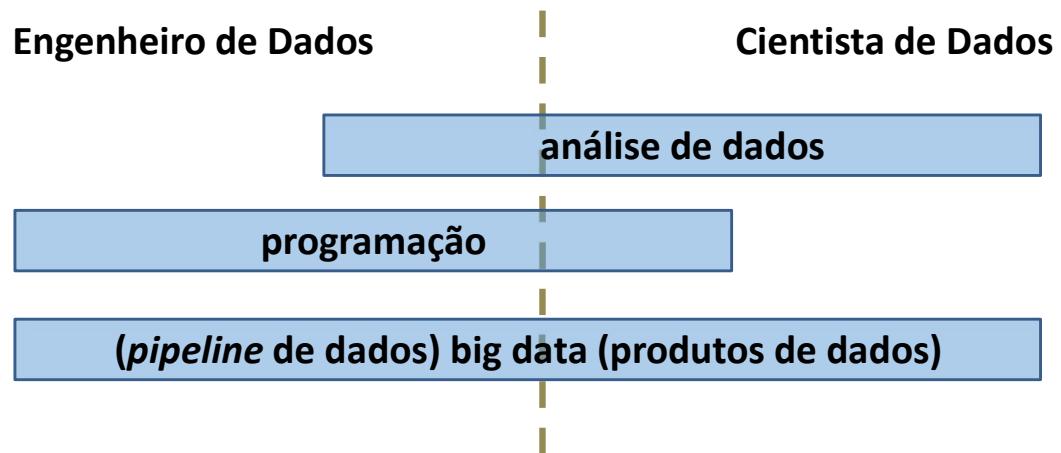
**Fonte:** <https://www.oreilly.com/radar/data-engineers-vs-data-scientists/>

## Cientista de Dados

- Alguém que adicionou aos seus conhecimentos em **matemática e estatística** o conhecimento em **programação** (?) para **analisar dados e criar modelos de dados**
- Conhece **matemática e estatística**, e usa recursos da **mineração de dados e de aprendizado de máquina** em **análises** de dados em um **domínio de negócio**

**Fonte:** <https://www.oreilly.com/radar/data-engineers-vs-data-scientists/>

# Competências comuns



Fonte: <https://www.oreilly.com/radar/data-engineers-vs-data-scientists/>



# **Big Data**

## **Caracterização e Definição**



# Quatro paradigmas da pesquisa científica

- 2 mil anos atrás
  - Ciência empírica, experimental
  - Descrição de fenômenos naturais
- Há algumas centenas de anos
  - Ramificação teórica, com leis
  - Usava modelos, generalizações
- Há umas poucas décadas
  - Ramificação computacional
  - Simulação de fenômenos complexos



HEY, A. J. et al. *The fourth paradigm: data-intensive scientific discovery*. [S.l.]: Microsoft research Redmond, WA, 2009.

# Quatro paradigmas da pesquisa científica

- Hoje: exploração de dados (E-Ciência)
  - Unifica a teoria, a experimentação e a simulação
  - Dados capturados por instrumentos ou gerados por um simulador
  - Processamento computacional
  - Análise científica de dados (estatística, data mining, analytics)



HEY, A. J. et al. *The fourth paradigm: data-intensive scientific discovery*. [S.l.]: Microsoft research Redmond, WA, 2009.



Dados gerados por pessoas em suas interações sociais através de aplicações da Internet, logs de computadores servidores e mesmo os dados de pesquisa científica, entre outras origens, são estimados em 2015 à quantidade de 30.000 gigabytes por segundo (MARZ; WARREN, 2015).



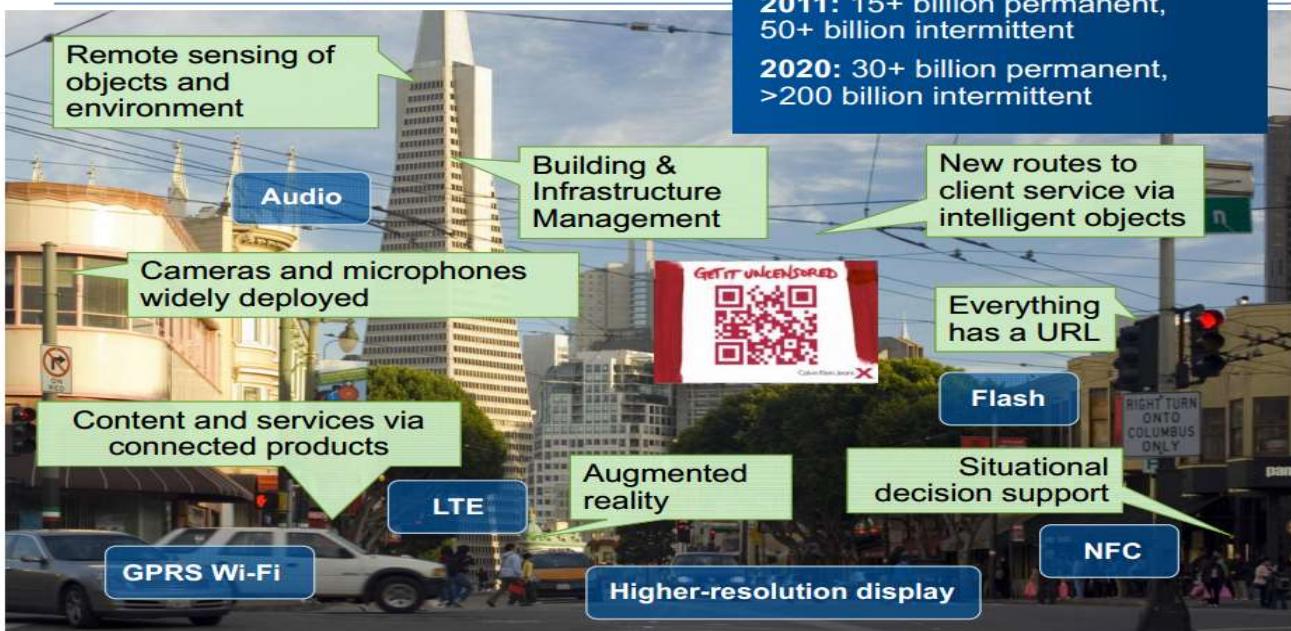
Dados gerados por pessoas em suas interações sociais através de aplicações da Internet, logs de computadores servidores e mesmo os dados de pesquisa científica, entre outras origens, são estimados em 2015 à quantidade de 30.000 gigabytes por segundo (MARZ; WARREN, 2015).

## The Internet of Things: It Is Already Here

Over 50% of Internet connections are things

2011: 15+ billion permanent,  
50+ billion intermittent

2020: 30+ billion permanent,  
>200 billion intermittent



Gartner - Top 10 Strategic Technology Trends for 2013

No entanto, o volume não é a única dimensão em que se nota crescimento: sensores, computadores e sistemas como redes sociais e lojas virtuais geram dados constantemente, os quais podem ser minerados e analisados em tempo real por aplicações de detecção de fraude e sistemas de recomendação. Dispositivos como sensores, gerenciadores, câmeras e microfones, eletrodomésticos conectados ou dispositivos de realidade aumentada, que podem ser categorizados pelo termo Internet das Coisas (IoT), geram dados não tratáveis por sistemas desenvolvidos para processar dados estruturados.

# E como estes dados podem ser usados ?



How data science and analytics can contribute to sustainable development



[www.unglobalpulse.org](http://www.unglobalpulse.org)

©UNGlobalPulse 2016

- ① **NO POVERTY**  
Spending patterns on mobile phone services can provide proxy indicators of income levels
- ② **CLEAN WATER AND SANITATION**  
Sensors connected to water pumps can track access to clean water
- ③ **REDUCED INEQUALITY**  
Speech-to-text analytics on local radio content can reveal discrimination concerns and support policy response
- ④ **LIFE BELOW WATER**  
Maritime vessel tracking data can reveal illegal, unregulated and unreported fishing activities
- ⑤ **ZERO HUNGER**  
Crowdsourcing or tracking of food prices listed online can help monitor food security in near real-time
- ⑥ **SUSTAINABLE CITIES AND COMMUNITIES**  
Smart metering allows utility companies to increase or restrict the flow of electricity, gas, or water to reduce waste and ensure adequate supply at peak periods
- ⑦ **PEACE, JUSTICE AND STRONG INSTITUTIONS**  
Sentiment analysis of social media can reveal public opinion on effective governance, public service delivery or human rights
- ⑧ **GOOD HEALTH AND WELL-BEING**  
Mapping the movement of mobile phone users can help predict the spread of infectious diseases
- ⑨ **RESPONSIBLE CONSUMPTION AND PRODUCTION**  
Online search patterns or e-commerce transactions can reveal the pace of transition to energy efficient products
- ⑩ **CLIMATE ACTION**  
Combining satellite imagery, crowd-sourced witness accounts and open data can help track deforestation
- ⑪ **INDUSTRY, INNOVATION AND INFRASTRUCTURE**  
Data from GPS devices can be used for traffic control and to improve public transport
- ⑫ **CLIMATE ACTION**  
Combining satellite imagery, crowd-sourced witness accounts and open data can help track deforestation
- ⑬ **PEACE, JUSTICE AND STRONG INSTITUTIONS**  
Sentiment analysis of social media can reveal public opinion on effective governance, public service delivery or human rights
- ⑭ **INDUSTRY, INNOVATION AND INFRASTRUCTURE**  
Data from GPS devices can be used for traffic control and to improve public transport
- ⑮ **CLIMATE ACTION**  
Combining satellite imagery, crowd-sourced witness accounts and open data can help track deforestation
- ⑯ **PEACE, JUSTICE AND STRONG INSTITUTIONS**  
Sentiment analysis of social media can reveal public opinion on effective governance, public service delivery or human rights
- ⑰ **PARTNERSHIPS FOR THE GOALS**  
Partnerships to enable the combining of statistics, mobile and internet data can provide a better and real-time understanding of today's hyper-connected world
- ⑱ **PARTNERSHIPS FOR THE GOALS**  
Partnerships to enable the combining of statistics, mobile and internet data can provide a better and real-time understanding of today's hyper-connected world

UN. United Nations Global Pulse - About. 2016. <http://www.unglobalpulse.org/about-new>

Tal abundância de dados tem motivado o desenvolvimento de aplicações que mineram e extraem informações de valor no seu contexto de pesquisa, a exemplo do Projeto Global Pulse, da Organização das Nações Unidas, que visa ao aproveitamento do Big Data como um bem público (UN, 2016).

# Afinal, o que é Big Data ?

- Dados com características como estas:
  - grandes (volume)
  - rápidos (velocidade)
  - difíceis de serem tratados pelos sistemas atuais (variedade)
- Dimensões
  - **Volume:** estima-se que o volume de dados disponível no universo digital seja de 2,7 zetabytes, ou  $2,7 \times 10^{21}$  bytes, ou ainda 2,7 trilhões de gigabytes;
  - **Variedade:** dados estruturados, semiestruturados e desestruturados;
  - **Velocidade:** fluxos de dados são gerados continuamente em grandes volumes;
  - **Variabilidade:** os dados podem assumir variedades de formas e interpretações diferentes;
  - **Valor:** possibilidade de extrair respostas aplicáveis ao negócio

MADDEN, S. From databases to big data. IEEE Internet Computing, IEEE, n. 3, p. 4–6, 2012.

Dados com características como estas, grandes (volume), rápidos (velocidade) e difíceis de serem tratados pelos sistemas atuais (variedade) definem o termo Big Data (MADDEN, 2012)

## História do termo “Big Data”

- 1998 - John Mashey
  - Cientista Chefe da empresa Silicon Graphics
  - “Big Data and the Next Wave of InfraStress”
- 1998 - Sholom M. Weiss e Nitin Indurkhyau
  - Ambiente acadêmico
  - “Predictive Data Mining: A Practical Guide”
- 2003 - Francis X. Diebold
  - “‘Big Data’ Dynamic Factor Models for Macroeconomic Measurement and Forecasting”

DIEBOLD, F. X. On the origin (s) and development of the term ‘big data’. PIER Working Paper, 2012

Aos problemas que lidam com dados que combinam estas características atribui-se o termo Big Data, termo cunhado por John Mashey, Cientista Chefe da empresa Silicon Graphics, em 1998, em seu trabalho intitulado “Big Data and the Next Wave of InfraStress” e utilizado academicamente em seu sentido atual inicialmente por Sholom M. Weiss e Nitin Indurkhyau, em 1998, em seu livro “Predictive Data Mining: A Practical Guide” e em 2003 por Francis X. Diebold em sua publicação intitulada “‘Big Data’ Dynamic Factor Models for Macroeconomic Measurement and Forecasting” (DIEBOLD, 2012).

# Big Data – Algumas aplicações

Tipo de negócio / setor	Exemplo dos tipos dados de entrada	Oportunidades de negócio
<b>1 - Bancos/crédito e Seguradoras</b>	<p>Histórico de transações.</p> <p>Ficha cadastral.</p> <p>Referências externas como o serviço de proteção ao crédito (SPC).</p> <p>Índices micro e macro econômicos.</p> <p>Dados geográficos e demográficos.</p>	<p>Aprovação de crédito.</p> <p>Flexibilização de taxas.</p> <p>Análise de mercado.</p> <p>Previsão de inadimplência.</p> <p>Detectação de fraudes.</p> <p>Identificação de novos nichos.</p> <p>Análise de risco de crédito.</p>
<b>2 - Segurança</b>	<p>Histórico de visitação.</p> <p>Ficha cadastral.</p> <p>Textos de notícias e de conteúdo da WEB.</p>	<p>Detectação de padrões de comportamentos físico ou digital que oferecem algum tipo de risco.</p>



Aquarela Advanced Analytics - <https://aquarela.com.br>

# Big Data – Algumas aplicações



<b>3 - Saúde</b>	Prontuário médico. Dados geográficos e demográficos. Sequenciamento de genomas.	Diagnóstico preditivo (previsão). Análise de dados genéticos. Descoberta de doenças e tratamentos. Mapa da saúde baseada em dados históricos. Efeitos adversos de medicamentos/tratamentos.
<b>4 - Óleo, gás e eletricidade</b>	Dados de sensores distribuídos.	Otimização dos recursos de produção Predição/detecção de falhas e fraudes.
<b>5 - Varejo</b>	Histórico de transações. Ficha cadastral. Percorso de compra em loja física e/ou virtual. Dados geográficos e demográficos. Dados de publicidade. Reclamações de clientes.	Aumento do faturamento pela otimização do mix de produto com base no padrão de comportamento durante de compra. Análise de faturamento (as-is, tendências), do alto volume de clientes e transações, crédito, perfil por regiões. Aumento de satisfação/fidelização dos clientes.

Aquarela Advanced Analytics - <https://aquarela.com.br>



# Big Data – Algumas aplicações

Otimização da produção em relação as vendas.  
Diminuição do tempo/quantidade de estocagem.  
Controle de qualidade.



PRODES 2009

## 6 - Produção

Dados dos sistema de gestão/produção ERPs.  
Dados de mercado.

Sugerir combinações ótimas de perfis de empresa, clientes, fornecedores para alavancagem de negócios.  
Identificação de oportunidades de sinergias.

## 7 - Organizações representativas

Ficha cadastral de clientes.  
Dados de eventos.  
Processo comercial em sistema de CRM.

Segmentação de mercado.  
Otimização a alocação dos recursos em publicidade.  
Descoberta de nichos de mercado.  
Desempenho de marca/produto.  
Identificação de tendências.

## 8 - Marketing

Índices micro e macroeconômicos.  
Pesquisa de mercado.  
Dados geográficos e demográficos.  
Conteúdo gerado pelos clientes. Dados dos concorrentes.

Aquarela Advanced Analytics - <https://aquarela.com.br>

# Big Data – Algumas aplicações

<b>9 - Educação</b>	Histórico escolar e frequências. Dados geográficos e demográficos.	Classificação de perfis de aprendizado com vistas a personalização da educação. Análise preditiva de evasão escolar.
<b>10 - Financeiro/Econômico</b>	Lista de ativos e seus valores. Histórico de transações. Índices micro e macroeconômicos.	Identificar o valor ótimo de compra de ativos complexos com muitas variáveis de análise (veículos, imóveis, ações e etc.). Determinação de tendências nos valores de ativos. Descoberta de oportunidades de arbitragem.
<b>11 - Logística</b>	Dados dos produtos. Rotas e pontos de entrega.	Otimização de fluxos de mercadorias. Otimização de estoques.

Aquarela Advanced Analytics - <https://aquarela.com.br>

# Big Data – Algumas aplicações

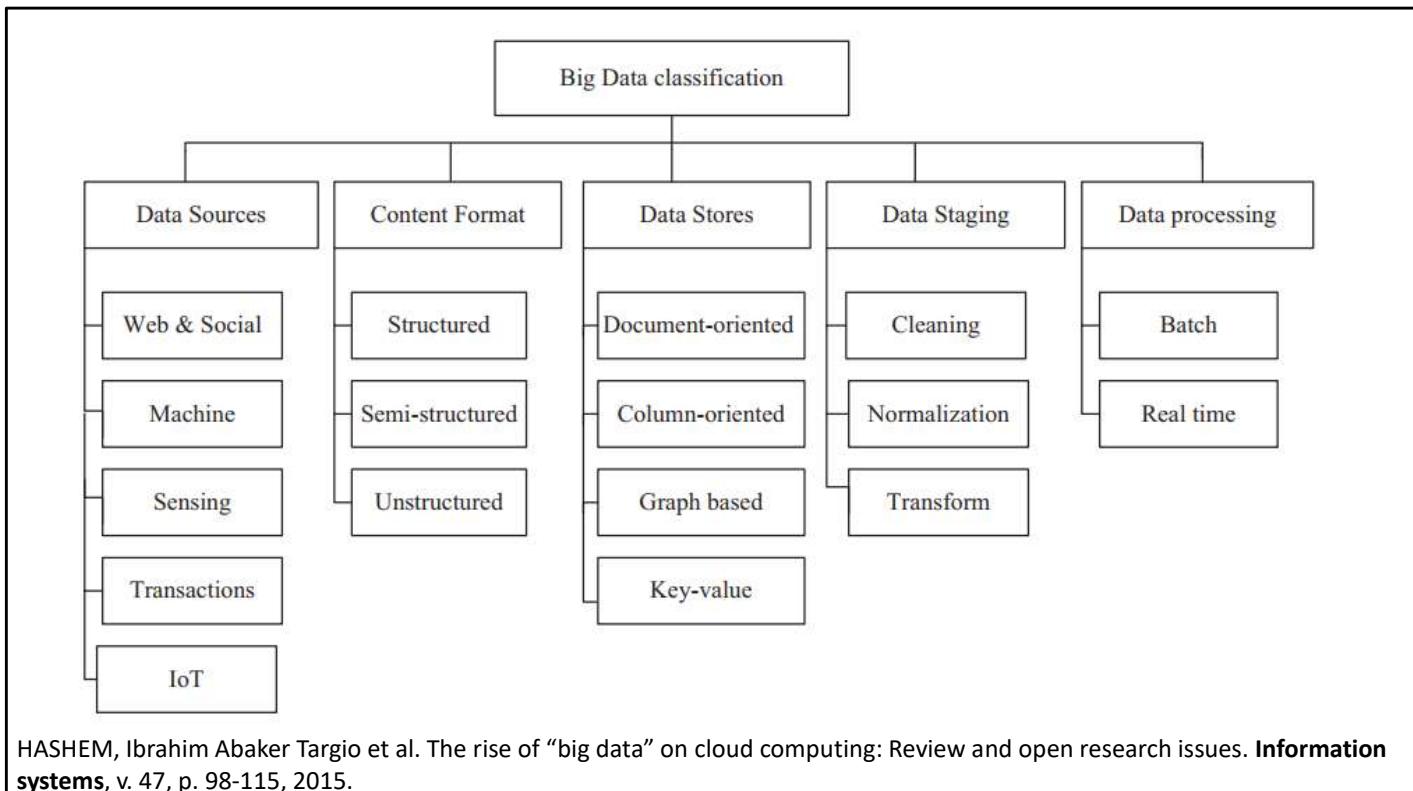
<b>12 - E-commerce</b>	Cadastro do cliente. Histórico das transações. Conteúdos gerados pelos clientes.	Aumento do faturamento por recomendações automáticas de produtos relacionados ao seu perfil e o perfil do produto. Aumento da satisfação/fidelização dos clientes.
<b>13 - Games, redes sociais e plataformas (freemium)</b>	Histórico de acessos. Cadastro dos usuários. Dados geográficos e demográficos.	Aumentar a taxa de conversão de usuários gratuitos para usuários pagantes pela detecção do comportamento e preferências dos usuários.
<b>14 - Recrutamento (RH)</b>	Cadastro das pessoas. Histórico profissional e currículos. Conexões em redes sociais.	Avaliação de perfil da pessoa para determinado cargo. Critérios para contratação, promoção e demissão. Melhor alocação dos recursos.

[www.aquare.la](http://www.aquare.la)

Aquarela Advanced Analytics - <https://aquare.la>

# **Big Data**

## **Uma classificação**



## DATA SOURCES

**Social media** - As mídias sociais são a fonte de informações geradas via URL para compartilhar ou trocar informações e idéias em comunidades e redes virtuais, como projetos colaborativos, blogs e microblogs, Facebook e Twitter.

**Machine-generated data** - Os dados da máquina são informações geradas automaticamente a partir de um hardware ou software, como computadores, dispositivos médicos ou outras máquinas, sem intervenção humana.

**Sensing** - Existem vários dispositivos sensores para medir quantidades físicas e transformá-las em sinais.

**Transactions** - Os dados da transação, como dados financeiros e de trabalho, compreendem um evento que envolve uma dimensão de tempo para descrever os dados.

**IoT** - A IoT representa um conjunto de objetos que são identificáveis exclusivamente como parte da Internet. Esses objetos incluem smartphones, câmeras digitais e tablets. Quando esses dispositivos se conectam pela Internet, eles permitem processos e serviços mais inteligentes que atendem às necessidades básicas, econômicas, ambientais e de saúde. Um grande número de dispositivos conectados à Internet oferece muitos tipos de serviços e produz grandes quantidades de dados e informações.

## CONTENT FORMAT

**Structured** - Os dados estruturados geralmente são gerenciados por SQL, uma

linguagem de programação criada para gerenciar e consultar dados nos SGBDR. Os dados estruturados são fáceis de inserir, consultar, armazenar e analisar. Exemplos de dados estruturados incluem números, palavras e datas.

**Semi-structured** - Dados semiestruturados são dados que não seguem um sistema de banco de dados convencional. Os dados semiestruturados podem estar na forma de dados estruturados que não são organizados em modelos de banco de dados relacional, como tabelas. Capturar dados semiestruturados para análise é diferente de capturar um formato de arquivo fixo. Portanto, a captura de dados semiestruturados requer o uso de regras complexas que decidem dinamicamente o próximo processo após a captura dos dados.

**Unstructured** - Dados não estruturados, como mensagens de texto, informações de localização, vídeos e dados de mídia social, são dados que não seguem um formato especificado. Considerando que o tamanho desse tipo de dados continua aumentando com o uso de smartphones, a necessidade de analisar e entender esses dados se tornou um desafio.

## DATA STORES

**Document-oriented** - Os repositórios de dados orientados a documentos são projetados principalmente para armazenar e recuperar coleções de documentos ou informações e suportar formatos complexos de dados em vários formatos padrão, como JSON, XML e binários (por exemplo, PDF e MS Word). Um armazenamento de dados orientado a documentos é semelhante a um registro ou linha em um banco de dados relacional, mas é mais flexível e pode recuperar documentos com base em seu conteúdo (por exemplo, MongoDB, SimpleDB e CouchDB).

**Column-oriented** - Um banco de dados orientado a colunas armazena seu conteúdo em colunas além de linhas, com valores de atributos pertencentes à mesma coluna armazenados de forma contígua. Um banco de dados orientado a colunas é diferente dos sistemas de banco de dados clássicos que armazenam linhas inteiras uma após a outra, como o BigTable.

**Graph database** - Um banco de dados de grafos, como o Neo4j, é projetado para armazenar e representar dados que utilizam um modelo de grafo com nós, arestas e propriedades relacionadas entre si por meio de relações.

**Key-value** - Chave-valor é um sistema alternativo de banco de dados relacional que armazena e acessa dados projetados para serem dimensionados para um tamanho muito grande. O Dynamo é um bom exemplo de um sistema de armazenamento de chave-valor altamente disponível; é usado pelo amazon.com em alguns de seus serviços. Outros exemplos de armazenamentos de chave-valor são Apache Hbase, Apache Cassandra e Voldemort. A Hbase usa o HDFS, uma versão de código aberto do BigTable do Google, construída em Cassandra. O Hbase armazena dados em tabelas, linhas e células. As linhas são classificadas por chave de linha e cada célula em uma tabela é especificada por uma chave de linha, uma chave de coluna e uma versão, com o conteúdo contido como uma matriz de bytes não interpretada.

## **DATA STAGING**

**Cleaning** - Limpeza é o processo de identificação de dados incompletos e sem sentido.

**Transform** - Transformação é o processo de transformar dados em um formato adequado para análise.

**Normalization** – Padronização dos valores em escalas semelhantes.

## **DATA PROCESSING**

**Batch** - Os sistemas baseados no MapReduce foram adotados por muitas organizações nos últimos anos para tarefas em lote de longa execução. Esse sistema permite o dimensionamento de aplicativos em grandes grupos de máquinas, incluindo milhares de nós.

**Real time** - Dados processados de forma continua, onde o fator restritivo é a latência.

# **Big Data Processamento**

# **KDD – KNOWLEDGE DISCOVERY IN DATABASES**

**Descoberta de Conhecimento em Bases de Dados**

## O QUE KDD É

O KDD é o processo não trivial de identificar padrões válidos, novos, potencialmente úteis e compreensíveis nos dados.

(Fayyad, Piatetsky-Shapiro, and Smyth 1996)

## O QUE KDD É

O KDD é o **processo não trivial** de identificar padrões válidos, novos, potencialmente úteis e compreensíveis nos dados.

(Fayyad, Piatetsky-Shapiro, and Smyth 1996)

Por não trivial, queremos dizer que alguma **pesquisa ou inferência** está envolvida; isto é, não é um cálculo direto de quantidades predefinidas, como calcular o valor médio de um conjunto de números.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

## O QUE KDD É

O KDD é o processo não trivial de identificar **padrões** válidos, novos, potencialmente úteis e compreensíveis nos dados.

(Fayyad, Piatetsky-Shapiro, and Smyth 1996)

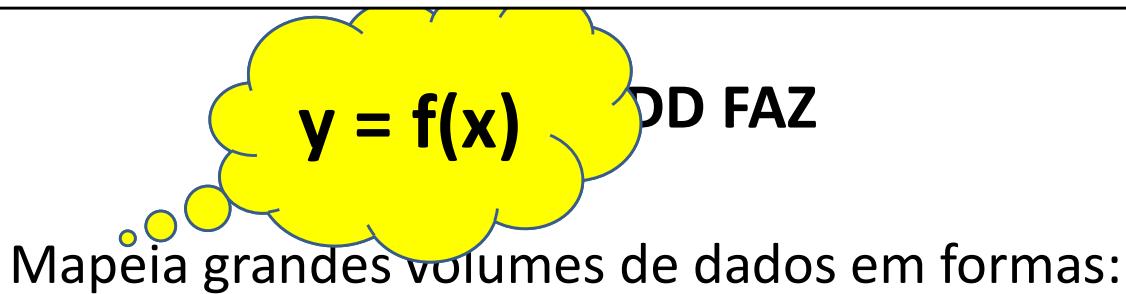
Aqui, dados são um conjunto de fatos (por exemplo, casos em um banco de dados) e padrão é uma **expressão em alguma linguagem** que descreve um subconjunto dos dados ou um modelo aplicável ao subconjunto.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

## O QUE KDD FAZ

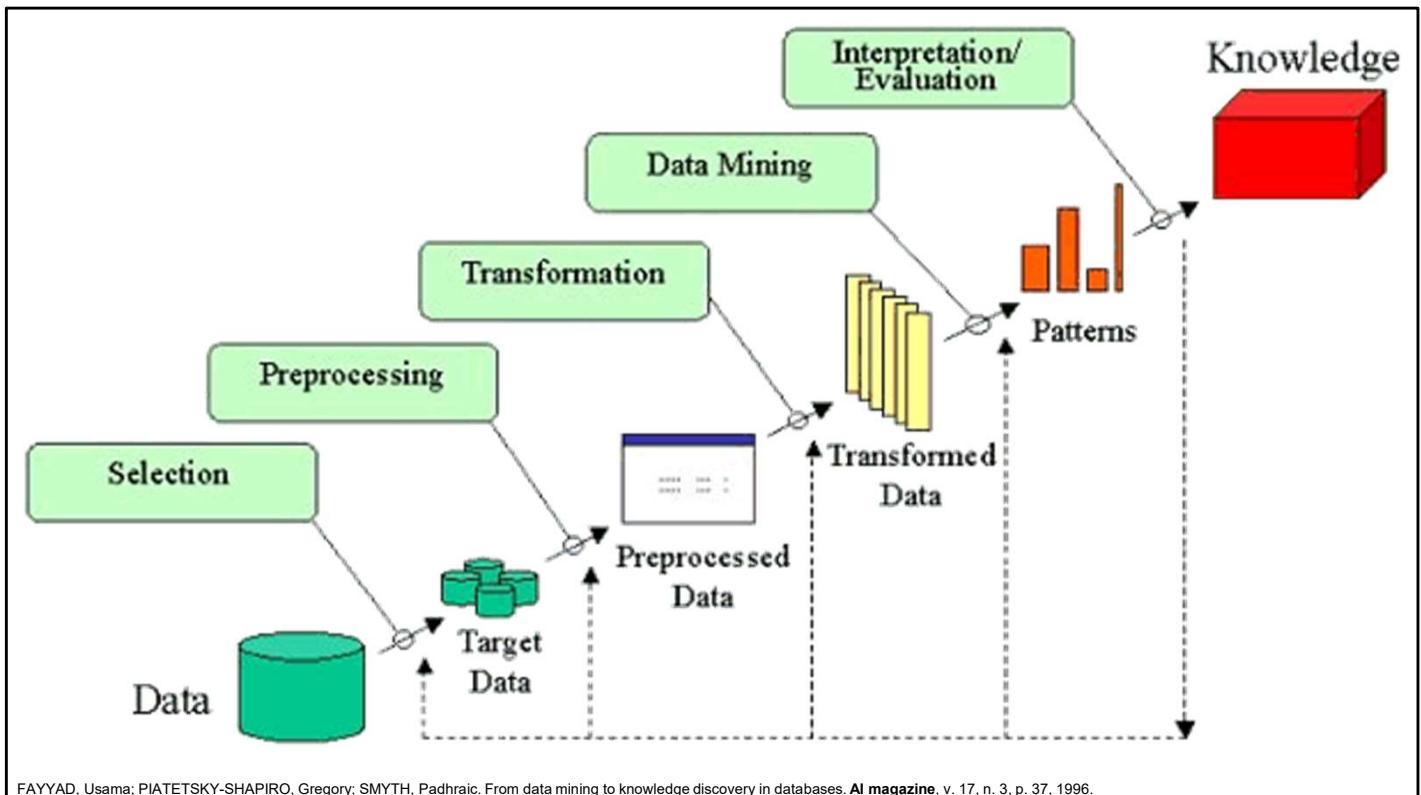
Mapeia grandes volumes de dados em formas:

- Mais compactas (análise descritiva)
- Mais abstratas (padrão/processo de geração dos dados)
- Mais úteis (modelos preditivos)



- Mais compactas (análise descritiva)
- Mais abstratas (padrão/processo de geração dos dados)
- Mais úteis (modelos preditivos)

Quando se pensa na palavra “mapear” deve-se pensar em uma função de mapeamento.



FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996.

O KDD é o processo não trivial de identificar **padrões** válidos, novos, potencialmente úteis e compreensíveis nos dados.

(Fayyad, Piatetsky-Shapiro, and Smyth 1996)

## **Tipos de processamento**

- Processamento em lote



## Tipos de processamento

- Processamento em lote



# **Tipos de processamento**

- Processamento em fluxo



## Tipos de processamento

- Processamento em fluxo



## Tipos de processamento

- Processamento em lote
  - Computa o **volume completo de dados** gerando visões a partir dele
  - Os volumes de dados são **grandes**
  - As **restrições de tempo** de resposta para a análise são menos exigentes
- Processamento em fluxo
  - Computa apenas o **volume de dados mais recente**
  - Requerido por aplicações de análise em **tempo real, ou quase real**
  - A variável tempo de resposta, ou latência, torna-se um **fator crítico** para o sucesso

## **Big Data – Por que uma arquitetura ?**

- Necessidade de inserir e pesquisar grandes volumes de dados com uma latência aceitável;

A escolha de arquitetura para sistemas Big Data surge da necessidade de inserir e pesquisar dados em grandes volumes e com uma latência aceitável para o contexto do sistema. Mesmo o escalamento dos tradicionais bancos de dados e até mesmo a inclusão de recursos de processamento assíncrono não são capazes de responder satisfatoriamente a problemas tais como corrupção de dados, tolerância a falhas e alta disponibilidade.

(MARZ; WARREN, 2015).

## **Big Data – Por que uma arquitetura ?**

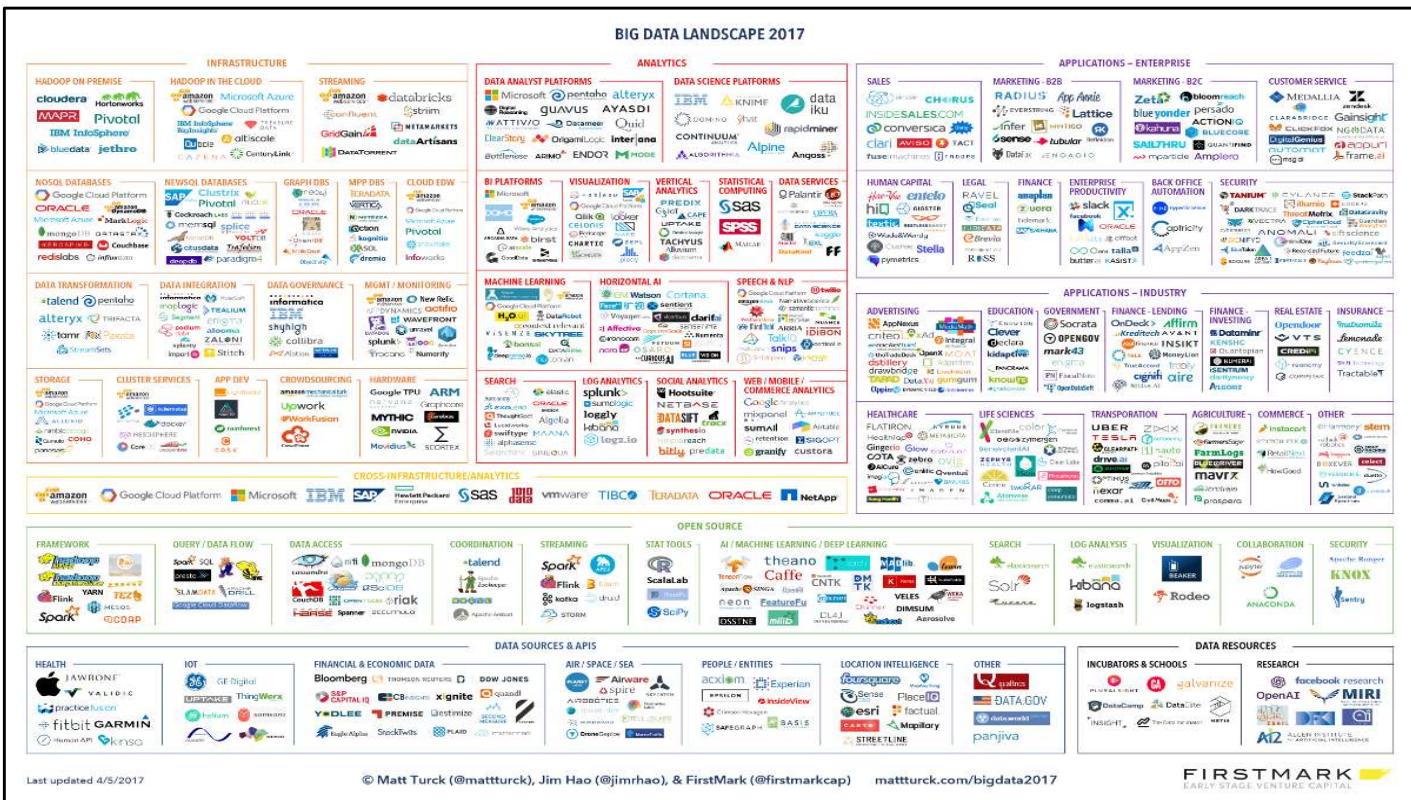
- Necessidade de inserir e pesquisar grandes volumes de dados com uma latência aceitável;
- O escalamento dos tradicionais bancos de dados e mesmo a inclusão de recursos de processamento assíncrono não respondem satisfatoriamente a problemas de corrupção de dados, tolerância a falhas e alta disponibilidade;

A escolha de arquitetura para sistemas Big Data surge da necessidade de inserir e pesquisar dados em grandes volumes e com uma latência aceitável para o contexto do sistema. Mesmo o escalamento dos tradicionais bancos de dados e até mesmo a inclusão de recursos de processamento assíncrono não são capazes de responder satisfatoriamente a problemas tais como corrupção de dados, tolerância a falhas e alta disponibilidade.

(MARZ; WARREN, 2015).

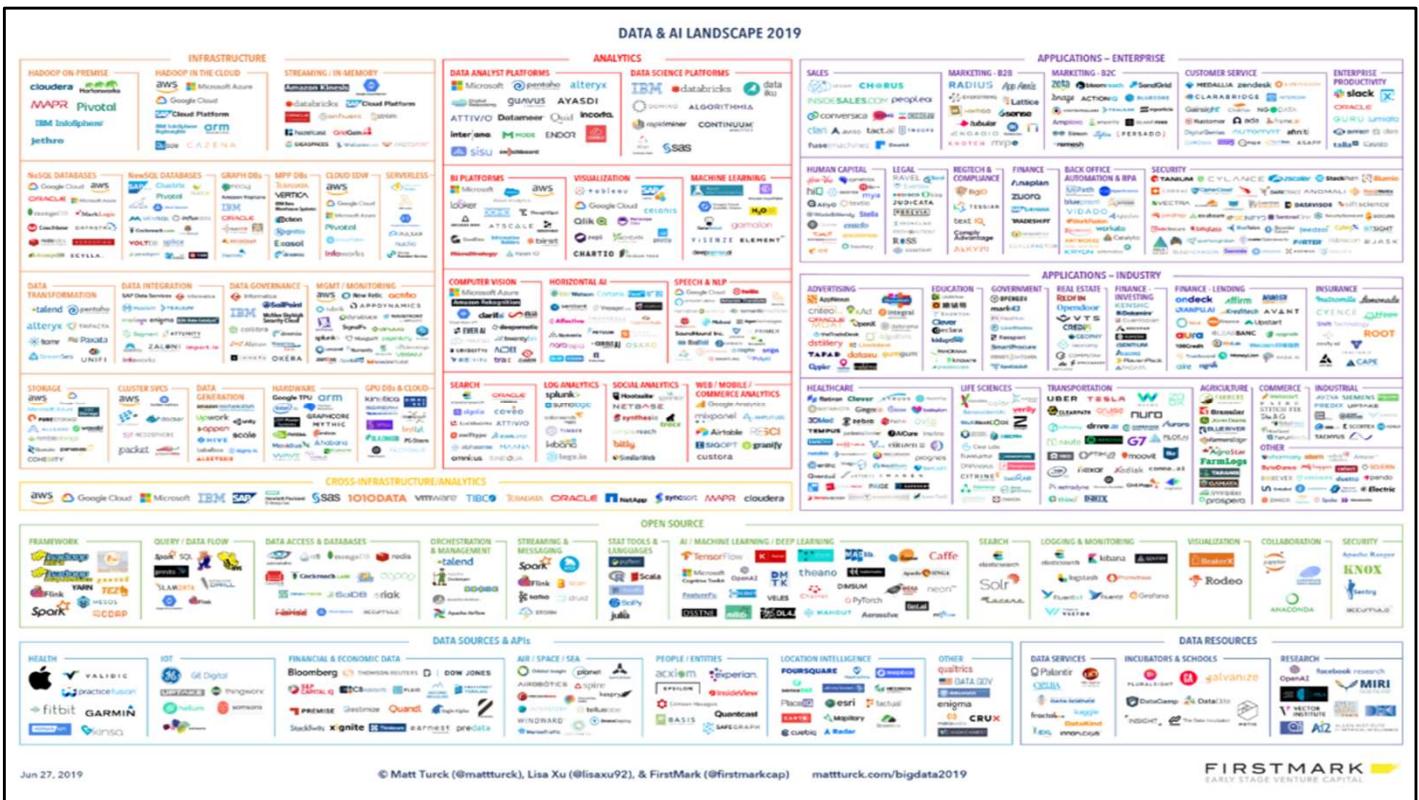
## **Big Data – Por que uma arquitetura ?**

- Necessidade de inserir e pesquisar grandes volumes de dados com uma latência aceitável;
- O escalamento dos tradicionais bancos de dados e mesmo a inclusão de recursos de processamento assíncrono não respondem satisfatoriamente a problemas de corrupção de dados, tolerância a falhas e alta disponibilidade;
- E ...

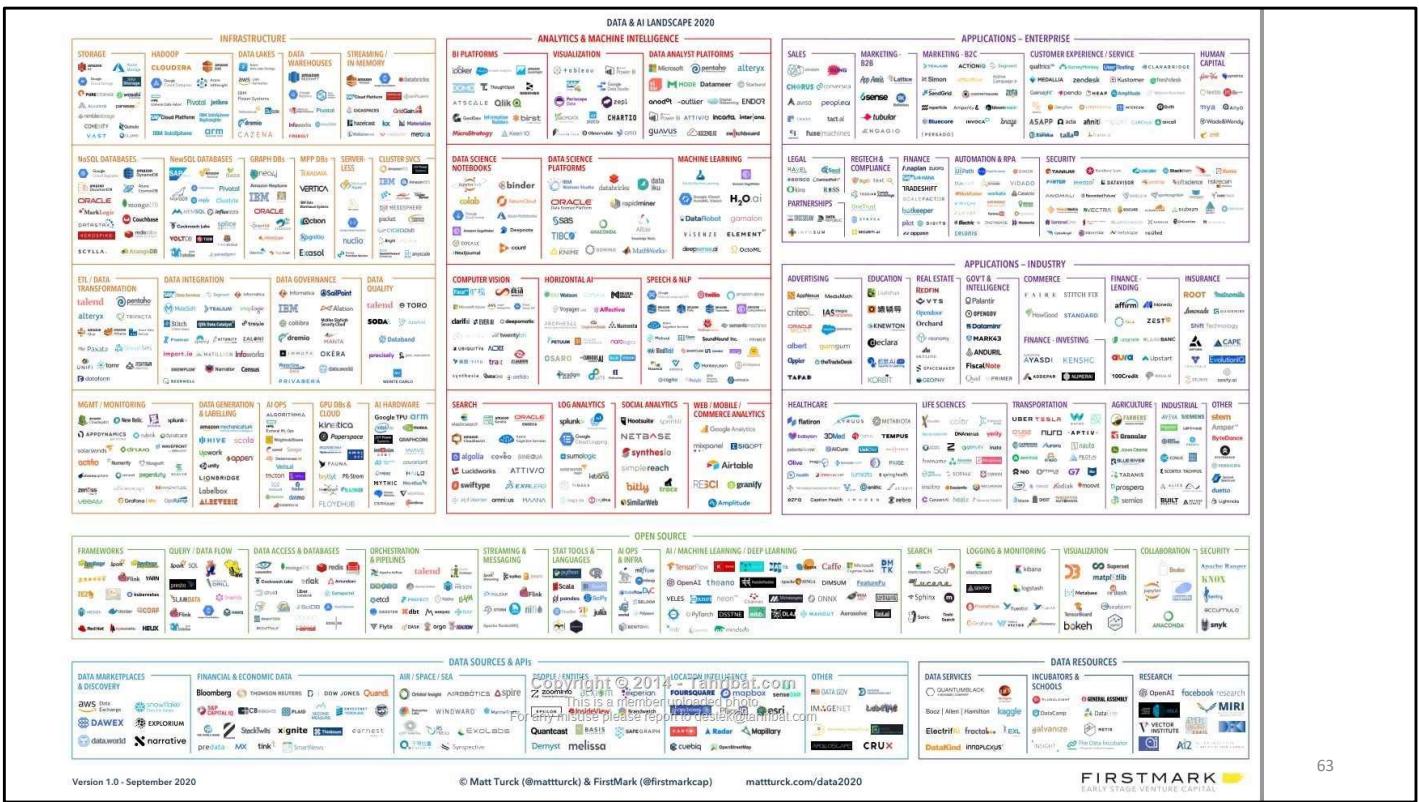


Um importante aspecto que diferencia o fenômeno Big Data do fenômeno conhecido como Business Intelligence é a oferta de ferramentas Open Source, muitas delas desenvolvidas e disponibilizadas por grandes empresas geradoras de dados, tais como Facebook, Yahoo!, Twiter e LinkedIn, tornando a implementação de soluções e as iniciativas analíticas em Big Data acessíveis (FAN; BIFET, 2013).





Fonte: <http://mattturck.com/>



## **Big Data – Principais arquiteturas**

- Kappa
- Lambda
- Liquid

## Referências e leituras recomendadas

TEJADA, Zoiner. **Mastering Azure Analytics: Architecting in the Cloud with Azure Data Lake, HDInsight, and Spark.**" O'Reilly Media, Inc., 2017.

<https://jornfranke.wordpress.com/2014/07/20/the-lambda-architecture-for-big-data-in-your-enterprise/>

<https://jornfranke.wordpress.com/2016/11/11/lambda-kappa-microservice-and-enterprise-architecture-for-big-data/>

MARZ, N.; WARREN, J. **Big Data: Principles and best practices of scalable realtime data systems.** [S.l.]: Manning Publications Co., 2015.

VANHOVE, T. et al. Managing the synchronization in the lambda architecture for optimized big data analysis. *IEICE Transactions on Communications*, The Institute of Electronics, Information and Communication Engineers, v. 99, n. 2, p. 297–306, 2016.

<https://blog.acolyer.org/2015/02/04/liquid-unifying-nearline-and-offline-big-data-integration/>

[http://lsds.doc.ic.ac.uk/sites/default/files/CIDR15\\_Paper25u.pdf](http://lsds.doc.ic.ac.uk/sites/default/files/CIDR15_Paper25u.pdf)



# Obrigado !

[renemendes@yahoo.com.br](mailto:renemendes@yahoo.com.br)

[www.linkedin.com/in/renemendes](https://www.linkedin.com/in/renemendes)